



THE UNIVERSITY *of* EDINBURGH

This thesis has been submitted in fulfilment of the requirements for a postgraduate degree (e.g. PhD, MPhil, DClinPsychol) at the University of Edinburgh. Please note the following terms and conditions of use:

This work is protected by copyright and other intellectual property rights, which are retained by the thesis author, unless otherwise stated.

A copy can be downloaded for personal non-commercial research or study, without prior permission or charge.

This thesis cannot be reproduced or quoted extensively from without first obtaining permission in writing from the author.

The content must not be changed in any way or sold commercially in any format or medium without the formal permission of the author.

When referring to this work, full bibliographic details including the author, title, awarding institution and date of the thesis must be given.

An investigation into methods of predicting income from credit card holders using panel data

Denys Osipenko

A thesis submitted in fulfilment of the requirements for the degree of
Doctor of Philosophy in Management Science and Business Economics



THE UNIVERSITY
of EDINBURGH

2018

Statement of Originality

This thesis has been composed by myself and contains no material that has been accepted for the award of any other degree at any university.

A part of this thesis has been published in

Osipenko, D. & Crook, J. (2015a). Credit Card Holders' Behavior Modeling: Transition Probability Prediction with Multinomial and Conditional Logistic Regression in SAS/STAT®, *SAS Forum, 2015*.
<http://support.sas.com/resources/papers/proceedings15/3217-2015.pdf>

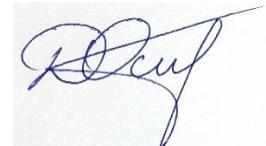
Osipenko, D. & Crook, J. (2015b). The Comparative Analysis of Predictive Models for Credit Limit Utilization Rate with SAS/STAT®, *SAS Forum, 2015*.

<http://support.sas.com/resources/papers/proceedings15/3328-2015.pdf>

Permission to include text from that paper has been gained from the publisher and the authors.

To the best of my knowledge and belief this thesis contains no other material previously published by any other person except where due acknowledgment has been made.

Denys Osipenko



Apr 2017

Abstract

A credit card as a banking product has a dual nature both as a convenient loan and a payment tool. Credit card profitability prediction is a complex problem because of the variety of the card holders' behaviour patterns, a fluctuating balance, and different sources of interest and transactional income. The state of a credit card account depends on the type of card usage and payments delinquency, and can be defined as inactive, transactor, revolver, delinquent, and default. The proposed credit cards profit prediction model consists of four stages: i) utilisation rate and interest rate income prediction, ii) non-interest rate income prediction, iii) account state prediction with conditional transition probabilities, and iv) the aggregation of the partial models into total income estimation.

This thesis describes an approach to credit card account-level profitability prediction based on multistate and multistage conditional probabilities models with different types of income and compares methods for the most accurate predictions. We use application, behavioural, card state, and macroeconomic characteristics as predictors.

This thesis contains nine chapters: Introduction, Literature Review, six chapters giving descriptions of the data, methodologies and discussions of the results of the empirical investigation, and Conclusion.

Introduction gives the key points and main aims of the current research and describes the general schema of the total income prediction model. Literature Review proposes a systematic analysis of academic work on loan profit modelling and highlights the gaps in the application of profit scoring to credit cards income prediction. Chapter 3 describes the data sample and gives the overview of characteristics.

Chapter 4 is dedicated to the prediction of the credit limit utilisation and contains the comparative analysis of the predictive accuracy of different regression models. We apply five methods such as i) linear regression, ii) fractional regression, iii) beta-regression, iv) beta-transformation, and v) weighted logistic regression with data binary transformation for utilisation rate prediction for one- and two-stage models.

Chapters 5 and 6 are dedicated to modelling the transition probabilities between credit card states. Chapter 5 describes the general model setups, model building methodology such as transition probability prediction with conditional binary logistic, ordinal, and

multinomial regressions, the data sample description, the univariate analysis of predictors. Chapter 6 discusses regression estimation results for all types of regression and a comparative analysis of the models.

Chapter 7 describes an approach to the non-interest rate income prediction and contains a comparative analysis of panel data regression techniques such as pooled and four random effect methods. We consider two sources of non-interest income generation: i) interchange fees and foreign exchange fees from transactions via point-of-sales (POS) and ii) ATM fees from cash withdrawals. We compare the predictive accuracy of a one-stage approach, which means the usage of a single linear model for the income amount estimation, and a two-stage approach, which means that the income amount conditional on the probability of POS and ATM transaction.

Chapter 8 aggregates the results from the partial models into a single model for total income estimation. We assume that a credit card account does not have a single particular state and a single behavioural type in the future, but has a chance to move to any of possible states. The income prediction model is selected according to these states, and the transition probabilities are used as weights for the particular interest rate and non-interest rate income prediction models.

Conclusion highlights the contributions of this research. We propose an innovative methodological approach for credit card income prediction as a system of models, which considers the estimation of the income from different sources and then aggregates the income estimations weighted by the states transition probabilities. The results of comparative analysis of regression methods for: i) utilization rate of credit limit and ii) non-interest income prediction, iii) the use of panel data with pooled and random effect for profit scoring, and iv) account level non-binary target transition probabilities estimation for credit cards can be used as benchmarks for further research and fill the gaps of empirical investigations in the literature. The estimation of the transition probability between states at the account level helps to avoid the memorylessness property of the Markov Chains approach. We have investigated the significance of predictors for models of this type. The proposed modelling approach can be applied for the development of business strategies such as credit limit management, customer segmentation by the profitability and behavioural type.

Acknowledgments

I dedicate my PhD thesis to my Mom and Dad, who supported, understood, missed me, and believed in me all this time.

I am grateful to my friends and loved one at home, who waited for me and pushed me. I will not forget my colleagues and classmates at the University of Edinburgh. And I will never forget the view from the window of my room on the roofs of Edinburgh, Firth of Forth, and Arthur's Seat.

I appreciate the help from all the members of the Credit Research Centre including Professor Thomas Archibald, Professor Jake Ansell, and Dr. Galina Andreeva.

I express my deep gratitude to Professor Lyn Thomas for the inspiration.

And I say many thanks for the invaluable help, advice, patience, and support to the person without whom this work would not be possible – my Supervisor, Professor Jonathan Crook.

Table of Contents

Abstract	v
Acknowledgments.....	vii
List of Tables.....	xiv
List of Figures	xix
List of Abbreviations.....	xxiii
1 Chapter One. Introduction	1
1.1 Credit Scoring Overview	1
1.2 Credit Cards and scoring	6
1.3 Profit Scoring	7
1.4 General Model description	11
1.5 Aims of research and research questions.....	16
1.6 Thesis Contributions.....	18
1.7 Thesis Structure	20
2 Chapter Two. Literature review.....	23
2.1 Introduction	23
2.2 Credit Scoring.....	23
2.2.1 The probability of Default modelling	23
2.2.2 EaD modelling	27
2.2.3 LGD modelling	30
2.3 Generic methods for Profit scoring	33
2.3.1 Profit modelling	33
2.3.2 Efficient cut-offs	35
2.3.3 Optimal Credit Limit Policy.....	36
2.3.4 Game theory.....	37
2.3.5 Use of Survival Models.....	37
2.3.6 Measures of Profit	40
2.4 Use of transition probabilities for profit scoring	44
2.5 Credit card usage and credit limit utilisation rate modelling	49
2.6 Credit card total and transactional income modelling	58
2.7 Gaps in the literature	60
2.8 Conclusion.....	62
3 Chapter Three. Data description and variables	65
3.1 Introduction	65

3.2	Panel Data overview and observation and performance windows	65
3.3	Independent Variables	74
3.4	Dependent Variables.....	84
3.5	Data sampling	85
3.6	Conclusion	87
4	Chapter Four. A comparative analysis of predictive models for credit limit utilisation rate.....	89
4.1	Introduction	89
4.2	Utilisation rate model and methods	91
4.2.1	Linear model approach.....	91
4.2.2	Beta-regression approach	93
4.2.3	Beta-transformation plus OLS	95
4.2.4	Utilisation rate Modelling with Fractional logit transformation	95
4.2.5	A weighted Logistic regression with binary transformation approach	96
4.2.6	Two-stage model	98
4.3	Portfolio overview	99
4.3.1	Portfolio vintage analysis and the utilisation rate distribution	99
4.3.2	Macroeconomic environment.....	102
4.4	The utilisation rate modelling conception, cross-sectional analysis and model segments	106
4.4.1	The utilisation rate and credit limit	106
4.4.2	Segments of the model	108
4.4.3	Cross-sectional analysis of the utilisation rate: characteristics	110
4.5	Model estimation and results	118
4.5.1	Parameter estimates for One-stage model.....	118
4.5.2	Assessment of the models' accuracy.....	123
4.5.3	The estimation with a lagged endogenous variable	131
4.5.4	Two-stage model summary for the 6-months utilisation rate prediction	135
4.6	Conclusion	138
5	Chapter Four. Credit Card Holders' States Transition Probability: Model description	141
5.1	Introduction	141
5.2	Segmentation	142
5.2.1	General description of credit card states	142

5.2.2	Reasons for transitions between states	146
5.2.3	<i>Graphical presentation of transactor and revolver</i>	150
5.3	Model Description	152
5.3.1	Credit card states and sources of income	152
5.3.2	The estimation of the required number of models	158
5.4	Transition states model.....	160
5.4.1	General model description	160
5.4.2	Model 1 – Decision tree of the conditional logistic regressions with a binary target	164
5.4.3	Model 2 – Ordinal logistic regression with the non-binary target	168
5.4.4	Model 3 – Multinomial logistic regression with the non-binary target	169
5.5	Univariate analysis and variables selection.....	171
5.6	Empirical transition matrices.....	184
5.6.1	Initial empirical transition matrix and states stability	184
5.6.2	Testing for Markovity	186
5.7	Conclusion.....	187
6	Chapter Six. Credit Card Holders' States Transition Probability: Modelling Results	189
6.1	Introduction	189
6.2	Modelling Results - Multinomial vs ordinal logistic regression	190
6.2.1	The results of the coefficients estimation	190
6.2.2	Ordinal and Multinomial Logistic Regression Validation	203
6.3	The multistage binary logistic regression model.....	204
6.3.1	Regression coefficients estimation.....	204
6.3.2	Validation Results for binary logistic regression models	218
6.4	New Definition of States and an Introduction of Additional States	219
6.5	Multinomial regression coefficients estimations results	226
6.5.1	Model t+1 estimations.....	226
6.5.2	Model t+3	235
6.5.3	Multinomial regression models validation results	237
6.6	Conclusion.....	241
7	Chapter Seven. Transactional Income Prediction.....	245
7.1	Introduction	245
7.2	Panel models.....	248

7.2.1	Panel models description.....	248
7.2.2	R-Squared for Panel data.....	253
7.2.3	Methods for Random-Effects Models Estimation.....	255
7.3	Transactional income modelling conception	259
7.3.1	Direct estimation	260
7.3.2	Two-stage model – indirect estimation	261
7.3.3	Income as a proportion of the credit limit – indirect estimation	262
7.3.4	Data description and covariates selection	263
7.4	Non-interest income modelling results.....	270
7.4.1	The probability of transaction	270
7.4.2	Distributions for POS and ATM income.....	275
7.4.3	Explanatory Variables Distribution.....	278
7.5	POS income estimation	282
7.5.1	Comparative analysis of the regression coefficients for pooled and random effect estimation methods	282
7.5.2	Comparative analysis of the goodness-of-fit of the pooled and random effect models for POS income	293
7.6	Estimation of income from ATM cash withdrawals	298
7.7	The comparative analysis of pooled and random effect regression coefficient estimates for POS and ATM transactions income	299
7.8	Summary of the non-interest income functions performance	304
7.9	Conclusion	306
8	Chapter 8. Total income prediction with an aggregated model	309
8.1	Overview	309
8.2	An approach to the total income calculation	311
8.2.1	The relationship between an account state and income	311
8.2.2	Income calculation for states	316
8.2.3	Total income calculation	320
8.3	Total Income distributions and calculation results	326
8.3.1	The distribution of target variables	326
8.3.2	Data used in the Direct Estimation Method	332
8.4	Direct Total income prediction with a linear model.....	334
8.5	Comparative analysis of the validation results of aggregated and direct estimation models for the total income.....	338
8.6	Expected Loss Modelling	350

8.7	Total Profit and Profitability	352
8.8	Scenario analysis and business contribution examples	356
8.8.1	Impact of the different behavioural scenarios on the total income from the credit card	356
8.8.2	Examples of the business implementation of the model for credit card income prediction	359
8.9	Conclusion.....	361
9	Chapter Nine. Conclusion.....	365
9.1	Summary	365
9.2	Contributions	370
9.2.1	Academic contribution	370
9.2.2	Practical impact.....	374
9.3	Limitations and further research.....	378
9.3.1	Limitations	378
9.3.2	<i>Further research</i>	380
	Reference.....	383
	Appendix 1. The polled and random-effect methods for ATM income	393
	Appendix 2. Classification of some profit modelling literature sources.....	397

List of Tables

Table 1.1 Confusion matrix.....	5
Table 3.1 Total number of observations (accounts) for each month including inactivated and having less than 12-month history	69
Table 3.2 Number of accounts by Month on Book for a portfolio with 12 months or more behavioural history.....	73
Table 3.3 List of the original raw data, behavioural, application and macroeconomic characteristics	75
Table 3.4 Month numbers in behavioural characteristics	79
Table 3.5 Descriptive statistics of Independent variables	79
Table 3.6 Descriptive statistics of Dependent variables	85
Table 3.7 Number of observations for the behavioural sample by month (for accounts with 12 or more observations in a sample after July 2010)	87
Table 4.1 The binary target transformation for weighted logistic regression	97
Table 4.2 Correlation between Macro Indicators and Utilisation Rate	105
Table 4.3 The definition of 3 types of models depending on MOB.....	109
Table 4.4 Expected effects of the variables for inclusion in a model	114
Table 4.5 The number of observations and average utilisation rate values (target) for three segments: No Changed Limit, Changed Limit, and Application (MOB 1-5). .	119
Table 4.6 Comparative analysis of OLS parameters estimation for three segments	122
Table 4.7 Summary validation of the regression methods for three utilisation rate segments	125
Table 4.8 Descriptive statistic for predicted distributions for the utilisation rate for 1-6 months for Limit NO Change Model	129
Table 4.9 Outcome Distributions for five prediction methods for the utilisation rate for 1-6 months for Limit NO Change Model.....	129
Table 4.10 Predictive accuracy for monthly utilisation rate model: Limit No Change model.....	131
Table 4.11 Estimated coefficients for OLS pooled and Arellano-Bond method for 1 month utilisation rate.....	134
Table 4.12 The fitting accuracy of Pooled OLS and Arellano-Bond method estimation	134

Table 4.13 Estimated coefficients for logistic regression for utilisation rate equal to 0 and 1	136
Table 4.14 Comparative analysis of fitting accuracy for two-stage models for the 6-months period.....	138
Table 5.1 Account state definition and related assessments	156
Table 5.2 Multinomial logistic regression models covering	159
Table 5.3. Multi-stage logistic regression models covering	159
Table 5.4 Selected covariates and expectations for their impact on the probability of transition.....	178
Table 5.5 Number of transitions from state S_i at t to state S_j in t+1.....	185
Table 5.6 Proportion of cases in state S_i at t to state S_j in t+1	185
Table 5.7 Stability of states.....	186
Table 6.1 Multinomial logistic regression estimation results	193
Table 6.2 Ordinal regression coefficient Estimations (part 1 – for Non-active and Transactors).....	199
Table 6.3 Ordinal regression coefficient Estimations (part 2 – for Revolvers, Delinquent 1, and Delinquent 2)	201
Table 6.4 Comparative Analysis of Ordinal and Multinomial Regression Models Validation – Test Sample	203
Table 6.5 Number of observations in Non-active state.....	206
Table 6.6 Conditional binary logistic regression Estimations for Non-Active state	206
Table 6.7 Number of observations in Transactor state	209
Table 6.8 Conditional binary logistic regression Estimations for Transactor state .	209
Table 6.9 Number of observations in Revolver state.....	212
Table 6.10 Conditional binary logistic regression Estimations for Revolver state..	212
Table 6.11 Number of observations in Delinquent 1 state.....	214
Table 6.12 Conditional binary logistic regression Estimations for Delinquency 1 month state	214
Table 6.13 Number of observations in Delinquent 2 state.....	216
Table 6.14 Conditional binary logistic regression Estimations for Delinquency 2 month state	216
Table 6.15 Gini and KS values for binary logistic regression (test sample).....	218

Table 6.16 Definitions for new full set of states (as of the end of month).....	221
Table 6.17 Possible transitions from current state to states for N steps (full set of states)	223
Table 6.18 Full states empirical transition matrix for t+1.....	224
Table 6.19 Full states empirical transition matrix for t+6.....	225
Table 6.20 Target frequencies for t+1, non-active state.....	227
Table 6.21 Multinomial regression estimations for t+1, from non-active state	227
Table 6.22 Target frequencies for t+1, from transactor state	228
Table 6.23 Multinomial regression estimations for t+1, from transactor state	228
Table 6.24 Target frequencies for t+1, revolver state	230
Table 6.25 Multinomial regression estimations for t+1, from revolver state.....	230
Table 6.26 Target frequencies for t+1, from revolver paid state.....	232
Table 6.27 Multinomial regression estimations for t+1, from revolver paid state... ..	232
Table 6.28 Target frequencies for t+1, from delinquent 1 state	233
Table 6.29 Multinomial regression estimations for t+1, from delinquents 1 state... ..	233
Table 6.30 Coefficients estimations for Revolver state transition for t+3 states prediction.....	236
Table 6.31 KS and Gini coefficients for development and validation sample target t+1	238
Table 6.32 KS and Gini coefficients for development and validation sample target t+3	239
Table 6.33 KS and Gini coefficients for development and validation sample target t+6	240
Table 7.1 Selected covariates and expectations for the impact on transactional income	264
Table 7.2The distribution of binary target for the probability of transaction	270
Table 7.3 The logistic regression coefficient estimation for POS and ATM transaction probability during 6 months	271
Table 7.4 AUC and Gini for POS and ATM development and validation samples ..	275
Table 7.5 Distribution characteristics of the target – POS amount for 6 months	276
Table 7.6 Distribution quantiles of the target – POS amount for 6 months	276
Table 7.7 Distribution characteristics of the target – ATM amount for 6 months ...	277

Table 7.8 Distribution quantiles of the target – ATM amount for 6 months	278
Table 7.9 POS Income 6 months – Linear Regression Estimation Results	284
Table 7.10 Assessing the fit of One-stage 6 months income model for full (positive POS transaction and zero income) data sample	294
Table 7.11 Assessing the fit of POS income model - second stage: conditional on positive POS transaction (POS Sum for 6 month > 0).....	294
Table 7.12 Assessing the Fit of Two-stage 6 months income model result – Option 1: Non-zero income condition is $\text{Pr}(\text{POS}) > 0.5$	297
Table 7.13 Assessing the Fit of Two-stage 6 months income model result – Option 2: POS Sum X $\text{Pr}(\text{POS} > 0)$	297
Table 7.14 Assessing the fit of One-stage ATM model for full (positive ATM transaction and zero income) data sample	298
Table 7.15 Assessing the fit of ATM income model - second stage: conditional on positive POS transaction (POS Sum 6 month > 0)	299
Table 7.16 Comparative analysis of pooled and random effect regression coefficient estimations for POS and ATM income	302
Table 7.17 Comparative analysis of pooled and random effect regression coefficient estimations for POS and ATM income (continue).....	303
Table 7.18 Comparative analysis of pooled and random effect regression coefficient estimations for POS and ATM income (continue).....	304
Table 7.19 Performance quality of the income prediction models	305
Table 8.1 Account state definition and related assessments	313
Table 8.2 Sources of income for each state.....	314
Table 8.3 Matrix of models for the set of account states	315
Table 8.4 Statistics: Distribution of Total income over 6 months, by account states at time t	328
Table 8.5 Quartiles of the total income for 6 months' distribution by account states at time t	329
Table 8.6 Share of Transactional and Interest Income received during the month of the current account state.....	330
Table 8.7 Share of Sum of Transactional and Sum of Interest Income received for six months from the initial account state	331

Table 8.8 The example of accounts data sample with calculations	333
Table 8.9 OLS Estimation coefficients for Direct Total income	336
Table 8.10 Descriptive statistics for Total income for 6 months	340
Table 8.11 Descriptive statistics for Total income for 6 months: quantiles.....	340
Table 8.12 Descriptive statistics for total income for the first month.....	341
Table 8.13 Descriptive statistics for total income for first month: quartiles.....	341
Table 8.14 Validation results of aggregated and direct for the total income prediction	345
Table 8.15 Confusion matrix of number of observed vs. predicted values for total income t+1 month.....	349
Table 8.16 Confusion matrix of number of observed vs. predicted values for total income for 6-month period.....	349
Table 8.17 Total profit and profitability for portfolio for all time	354
Table 8.18 Profit and profitability for portfolio for one month fixed observation point of the total data sample.....	355
Table 8.19 Total profit and profitability predictive accuracy	355
Table 8.20 An example of impact of different scenarios on the total income	358
Table 8.21 An example of the credit limit management matrix based on the profitability and the probability of default (is not related to thesis data sample)....	360

List of Figures

Figure 1.1. Calculation of Gini index and area under ROC curve	3
Figure 1.2. Calculation of Kolmogorov-Smirnov test value.....	5
Figure 1.3. General model schema.....	12
Figure 3.1 An example of the 6-months observation and 6-months performance window for 12-months loan issuance period	70
Figure 3.2 An example of usage of cohorts of the 6-months observation and 6-months performance period for one loan issued (activated) in July 2010	72
Figure 3.3 Number of observations (cases) by Month on Book	73
Figure 3.4 The dynamic of macroeconomic indicators: unemployment rate, Local currency to Euro exchange rate, and CPI till June 2012	84
Figure 4.1 Two-stage model tree	98
Figure 4.2 The vintages of the credit limit utilisation rate: by Month on Book	99
Figure 4.3 The utilisation rate vintage: by the month of activation	100
Figure 4.4 The utilisation rate distribution for active accounts	101
Figure 4.5 Macroeconomic indicators dynamics and the utilisation rate.....	103
Figure 4.6 Logarithms of some macroeconomic indicators and the utilisation rate dynamics	104
Figure 4.7 The constant utilisation rate for an increased credit limit.....	106
Figure 4.8 The average outstanding balance and credit limit distribution by the utilisation rate.....	107
Figure 4.9 Average Utilisation rate by Month on Book.....	109
Figure 4.10 The utilisation rate (avg_UT), average balance, and average credit limit by gender.....	111
Figure 4.11 The utilisation rate by age group - time variation	111
Figure 4.12 The utilisation rate, average balance, and credit limit by education	112
Figure 4.13 The utilisation rate by education - time variation	112
Figure 4.14 The utilisation rate, average balance, and credit limit by employment status	113
Figure 4.15 Beta Transformation + OLS distribution versus observed	127
Figure 4.16 Fractional regression distribution versus observed.....	128
Figure 5.1 Transition between main states over a month.....	145

Figure 5.2 Transition between all states over a month.....	145
Figure 5.3 Example of the active revolver outstanding balance by months	151
Figure 5.4 Example of the non-active revolver outstanding balance by months	151
Figure 5.5 Example of the transactor outstanding balance by days	152
Figure 5.6 Example of the revolve - cash-user outstanding balance by months.....	152
Figure 5.7 Multistage schema of the conditional logistic regression models	164
Figure 5.8 An example of the univariate analysis of the credit score by the utilization rate.....	171
Figure 5.9 Univariate distribution from all states S_t to state S_{t+1} for an average purchase transaction to average balance for a month by the equal number of observations in the range.....	173
Figure 5.10 Univariate distribution from all states S_t to state S_{t+1} for average purchase transaction to average monthly balance (without revolvers)	173
Figure 5.11 Univariate distribution from all states S_t to state S_{t+1} for an average purchase transaction to average monthly balance by equal ranges	174
Figure 5.12 Univariate distribution from some state S_t to state S_{t+1} for the ration of an average purchase transaction to average monthly balance by the equal number of observations in the range.....	175
Figure 5.13 Univariate distribution from All S_t to S_{t+1} : Consecutive Months in Current State.....	176
Figure 5.14 Univariate distribution from S_t to S_{t+1} by Consecutive Months in Current State.....	177
Figure 7.1 ROC Curves for POS and ATM probability of transaction for 6 months	274
Figure 7.2 Distribution of the target – POS amount for 6 months	275
Figure 7.3 Distribution of the target – ATM amount for 6 months	277
Figure 7.4 Dependence of Average POS income for 6 months on the Logarithm of Sum of Purchases to Sum of Payments for one month	279
Figure 7.5 Dependence of Average POS income for 6 months on Average Number of Debit Transactions for 1-3 months.....	279
Figure 7.6 Dependence of Average POS income for 6 months on the Credit Limit Utilization rate for 6 months	279

Figure 7.7 Dependence of Average POS income for 6 months on Age	280
Figure 7.8 Dependence of Average ATM income for 6 months on the Logarithm of Sum of Purchases to Sum of Payments for one month	280
Figure 7.9 Dependence of Average ATM income for 6 months on Average Number of Debit Transactions for 1-3 months.....	281
Figure 7.10 Dependence of Average ATM income for 6 months on the Credit Limit Utilization rate for 6 months	281
Figure 7.11 Dependence of Average ATM income for 6 months on Age	281
Figure 7.12 The distributions of observed vs. predicted POS 6 months income values for the range (0,10].....	296
Figure 8.1 Density distribution of total income over six months (total_income6) ..	326
Figure 8.2 Density distribution of the transactional income over 6 months (trans_inc6)	327
Figure 8.3 Density distribution of interest income over 6 months (interest_inc6) ..	327
Figure 8.4 Share of Transactional and Interest Income received during the month of an account state	330
Figure 8.5 Share of Sum of Interest and Transactional Income over six months from the initial account state.....	331
Figure 8.6 Total income observed and total income predicted with TID for six-month period (density histogram cut for high frequent values).	342
Figure 8.7 Total income observed and total income predicted with TID for six month period density histogram for all values.	343
Figure 8.8 Total income observed and total income predicted with TID for +1 month period density histogram.	343

List of Abbreviations

ATM	Automated teller machine	MAE	Mean Absolute Error
CCF	Credit Conversion Factor	MOB	Months on Book
CPI	Consumer price index	MDP	Markov Decision Process
DPD	Days Past Due	NL	Nerlove's Method
EaD	Exposure at Default	OLS	Ordinary Least Squares
EOP	End of Period	POS	Point of Sales
EUR	Euro	PD	Probability of Default
FB	Fuller and Battese Method	RMSE	Root-mean-square error
GDP	Gross Domestic Product	TMM	Transition matrix model
GLM	Generalised Linear Model	WH	Wallace and Hussain Method
LEQ	Loan Equivalent	WK	Wansbeek and Kapteyn Method
LGD	Loss Given Default	UT	Utilisation rate

1 Chapter One. Introduction

1.1 Credit Scoring Overview

Credit scoring methodology has been used by financial institutions to make lending decisions since the 1940s. Originally, this methodology entailed finding the best possible binary classification of loans into two groups: ‘Good’ and ‘Bad’ accounts using a set of factors. The relevant target groups are often defined on the basis of account delinquencies, such as some missing payments or some months in delinquency. The original aim of the credit scoring business approach was the discrimination of loans or loan applications by the probability of default, or the probability of being ‘Bad’ in a given period of loan life. The lenders possess a tool which estimates the probability of a negative event and makes a decision according on their risk appetites or other business targets such as the bad rate of the portfolio, marginal loan probability of default, application flow acceptance rate, and/or profitability.

Since the implementation of the credit scoring model in relevant industries (for loan granting purposes), the practice has spread to other areas of lending and credit life cycle. As a result, we face such types of credit risk scoring as behavioural, collection, fraud, and others. Behavioural scoring is used for assessing existing loans and making a decision on cross-selling, prolongation, credit limit changes, etc. Collection scoring is a subtype of behavioural scoring, usually employed for delinquent or defaulted loans. It is used for a decision of what type of actions should be applied for the best loan recovery or restructuring. Fraud scoring is being implemented for the estimation of the probability of fraudulent actions from the customer’s side, and others. Consequently, the ‘Bad’ case definition has been transformed from the number of days past due to, for example, repayment of the debt amount in arrears after some period or action, or fraud identifier, etc. The probability prediction tasks in risk management have spread to the rates and amount estimation as Loss Given Default, which is the share of the loan balance lost in case of default, and Exposure at Default – the outstanding balance under risk at the time of default. Moreover, credit scoring methodology has transferred outside of the risk assessment area and started to be used

for marketing and customer relationship management purposes such as response scoring, attrition scoring, usage scoring, etc. Response scoring is the estimation of the probability of response for some action such as mail, call, etc. Attrition scoring is the estimation of the probability of early payment and account foreclosure. Propensity scoring is a customer's likelihood of using a credit facility. Usage scoring is the prediction of the credit line use or the utilisation rate for the credit limit (Thomas, 2000; Anderson, 2007; Mays, 2001).

However, the important business task of the decision-making process is the prediction and maximisation of profit. Profit maximisation problems are often reduced to optimal cut-off strategy, thus, finding a balance between expected income and expected loss, and borrowers' utility and acceptance (Crook et al., 2007). However, this is a simplification because fixed and variable costs are omitted. Usually, the profit estimation is made at the portfolio level. However, credit scoring may contribute elements towards making a profit prediction at the account level.

In traditional credit scoring one question to estimate a classifier to predict, *ex ante* the probability that an applicant or a current customer will repay their loan as scheduled. The aim is to predict as accurately as possible.

The logistic regression model has been the most popular technique and an industrial standard for credit scoring development for many years (Thomas et al., 2002). Many machine learning techniques have also been tested for credit scoring tasks and have shown good predictive power results (for example, Lotterman et al. (2015) for comparison of the accuracy of different techniques). However, the scorecards built with logistic regression remain the most popular tool for credit risk assessment, mainly because of easily interpretable results and a transparent relationship between predictors and outcome.

The performance of a classifier is measured as its ability to discriminate *ex ante* between defaulting (bad) and non-defaulting (good) borrowers. The actual discriminatory power of a model can be reviewed only *ex post* once the evidence becomes available.

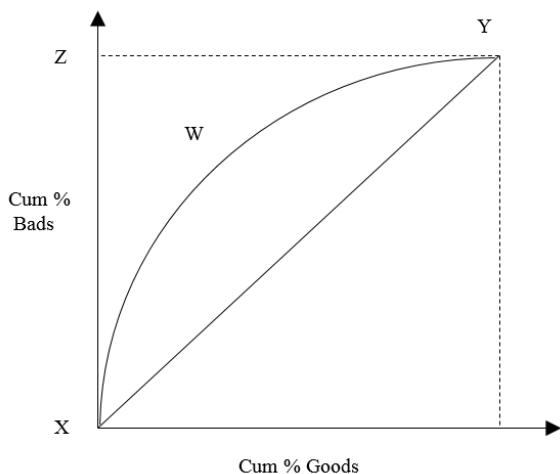
Statistical tests of the discriminatory ability of a classifier should be performed based on a test sample. There are several tests used for such purposes, but the most commonly

used, which are within the scope of this thesis, are: i) Gini coefficient and Area Under Curve (AUC), and ii) the Kolmogorov-Smirnov (K-S) test (Thomas et al., 2002; Siddiqi, 2006), and iii) percentage correctly classified.

Gini index and Area Under the ROC Curve (AUC)

The Gini coefficient and the area under the Receiver Operating Curve (AUC) are measures discrimination achieved by scoring model across all values of a cut-off score. Let $F(S|G)$ and $F(S|B)$ denote the cumulative distribution of score for the goods and bards respectively. That is $F(S|G) = \Pr(\text{Score} \leq S|G)$ and $F(S|B) = \Pr(\text{Score} \leq S|B)$. If we plot these values against each other, we gain Figure 1.1. This is the Receiver Operating Characteristic Curve (ROC).

Figure 1.1. Calculation of Gini index and area under ROC curve



If the scoring model discriminated perfectly, then these would be a score such that all the Bads would have a score, S_c , below this and all the Goods would have a score above this. In Figure 1.1 for score below S_c , $\Pr(\text{Score} \leq S_c|G) = 0$ and $\Pr(\text{Score} \leq S_c|B) = 1$ and for scores above S_c $\Pr(\text{Score} \leq S_c|G) = 1$ and $\Pr(\text{Score} \leq S_c|B) = 0$. This is represented by the line XYZ. If the discrimination was no better than random the plot would have the line XY. The further the plot of $F(S|B)$ against $F(S|G)$ is from the XY line and towards the XZY line the better the discrimination. The area under the ROC is a measure of discrimination.

The Gini coefficient is defined as

$$\text{Gini} = 2\text{AUC} - 1 \quad (1.1)$$

and so is monotonic in AUC.

Kolmogorov-Smirnov test

The Kolmogorov–Smirnov test (KS test) is a nonparametric test for the equality of continuous, one-dimensional distributions. It is used for a comparison the equality or diversity of two samples of observations. The Kolmogorov–Smirnov statistic is calculated as a distance between the empirical cumulative distribution functions of two samples. The null distribution of this statistic is calculated under the null hypothesis that the samples are drawn from the same distribution (in the two-sample case). The two-sample K–S test is a general nonparametric method for comparing two samples.

The KS statistic is defined as the maximum difference between the cumulative percentage of ‘goods’ and the cumulative percentage of ‘bads’ as score increases. This statistic can be used to compare different models built on the same sample. The Kolmogorov-Smirnov test as a statistical test can be used to compare the results of the model on the development sample and a test sample. The KS Test consists of comparing the maximum difference between the defaulters and the non-defaulters’ distribution, as well as comparing the value to a set of critical values that depend on the size of the sample *and the required significance level of the test. The KS statistic is defined as*

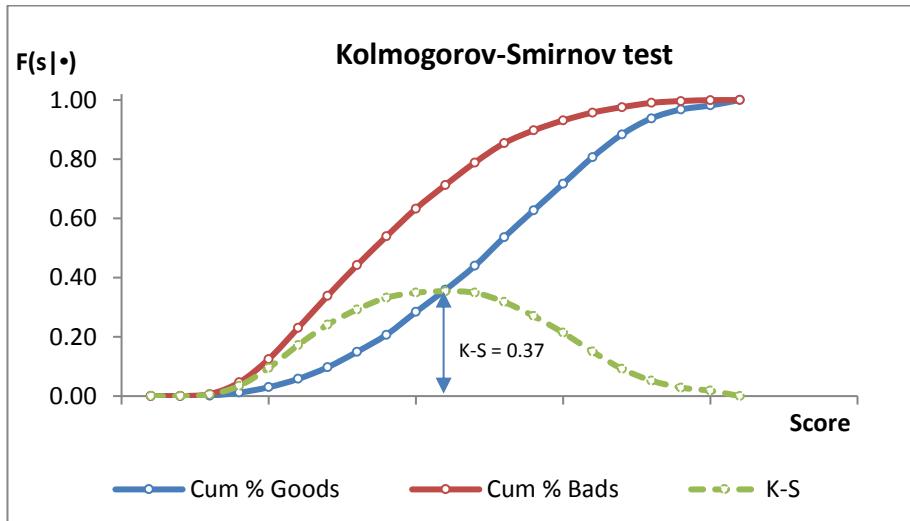
$$D_n = \sup |F_n(x) - F(x)| = \sup |F(s|B) - F(s|G)| \quad (1.2)$$

where $\sup x$ is the supremum of the set of distances, and

$F_n(x)$ denotes the empirical cumulative distribution function of the basic sample, or $F(s|B)$ denotes the probability that a ‘bad’ case has a score less than s , and $F(x)$ denotes the second empirical cumulative distribution function, or $F(s|G)$ denotes the probability that a ‘good’ case has a score less than s . These two cumulative distributions are plotted against score in Figure 1.2.

In practice, the statistic requires a relatively large number of data points to properly reject the null hypothesis. K-S statistic is the maximum value of the distance between the cumulative good and cumulative bad curves.

Figure 1.2. Calculation of Kolmogorov-Smirnov test value



In Figure 1.2 the K-S curve (dotted line) shows the distance between the cumulative distribution of ‘goods’ (blue line at the bottom) and the cumulative distribution of ‘bads’ (red line at the top). The cumulative distribution value 1 related to the maximum (or minimum) available score. The maximum distance between the cumulative distributions is the Kolmogorov-Smirnov test value and for example in Figure 1.3. is equal to 0.37. Low KS test values are below 0.2 may indicate poor discrimination power of a scoring model.

Percentage correctly classified

A popular method for measuring the discrimination of a scoring model is the percentage correctly classified cases, or $1 - \text{error rate}$.

Consider Table 1.1 which is called a Confusion matrix.

Table 1.1 Confusion matrix

		Observed	
		Good	Bad
Predicted	Good	n_{GG}	n_{BG}
	Bad	n_{GB}	n_{BB}

The percentage of good cases wrongly classified as bad is calculated as

$$G^W = n_{BG}/(n_{GG} + n_{BG})$$

and the percentage of good cases correctly classified as good is calculated as

$$G^T = n_{GG}/(n_{GG} + n_{BG})$$

where n_{BG} is the number of bad cases classified as goods (wrongly classified), n_{GG} is the number of good cases classified as goods (correctly classified).

The percentage of bad cases classified as good is calculated as

$$B^W = n_{GB}/(n_{GB} + n_{BB})$$

and the percentage of bad cases correctly classified as bad is calculated as

$$B^T = n_{BB}/(n_{GB} + n_{BB})$$

where n_{GB} is the number of good cases classified as bads (wrongly classified), n_{BB} is the number of bad cases classified as bads (correctly classified). The cases are classified to the good or bad segment with use a cut-off. The cut-off can be set up to keep the same bad rate (the proportion of good and bad cases) in the tested sample as in the training one.

The percentage of correctly classified (PCC) is calculated as

$$PCC = (n_{GG} + n_{BB})/(n_{GG} + n_{GB} + n_{BG} + n_{BB})$$

A weakness of the percentage correctly classified (PCC) is that it is depending on the selected cut-off. The percentage correctly classified depends on the training sample bad rate. Percentage correctly classified is not used as an accuracy measure in the scope of this research because it depends on the selected cut off and we do not use any cut offs in this research.

It can be shown that measured discrimination derived from a training sample are overoptimistic and an independent test sample should be used (Thomas et al., 2017).

1.2 Credit Cards and scoring

Credit scoring has been widely used, in particular, due to the mass issuance of credit cards, starting in the 1960s. Today, credit cards are some of the most popular non-cash instruments in the World, and the second most popular non-cash instrument in the United States. The main reason for the growing popularity is that ‘credit cards offer

unique benefits to consumers and merchants and profit opportunities to banks' (Chakravorti, 2003).

A credit card is a banking product that has a dual nature. It is a convenient loan similar both to a cash loan and to finance a purchase. On the other hand, it is a payment tool for making purchases with the use of card systems. Loan and payment features in the same financial instrument make the task of prediction of credit card profitability more complex than for standard loans. Moreover, a credit card has a fluctuating balance, and credit card profitability prediction is a complex problem, given the differing behaviours of relevant credit card holders, and different sources of interest and transactional income. The use of traditional techniques gives satisfactory empirical results. However, a lot of the industrial models are simplified and apply a lot of assumptions and simplifications, which reduce the accuracy of prediction.

A credit card has a fluctuating balance, and its accurate forecast is a current problem of credit risk management, liquidity risk, business strategies, customer segmentation and other tasks of bank management. Modelling default needs to consider this dual nature of revolving products both as a loan and as a payment tool.

The state of a credit card account depends on the type of card usage and payments delinquency. The state of a credit card account can be defined as inactive, transactor, revolver, delinquent, or default; and credit card income prediction needs to consider the features of each state in individual models. The estimation of the transition probability matrix between states at the account level helps to avoid the memorylessness property of the Markov Chains approach, which is often used for a transition probabilities estimation at the pooled level. The proposed model for the credit cards net income prediction consists of four stages: account state prediction with conditional transition probabilities, the outstanding balance and interest income estimation, non-interest income estimation, and total net income estimation.

1.3 Profit Scoring

At account level the literature uses a range of different empirical definitions of profit. Thus Serrano-Cinca and Gutierrez-Niet (2016) define profitability to be an internal rate of return where the numerator of each term is merely interest received by the lender. So et al. (2014) define profit as the present value of expected future income,

variable costs apart from the cost of funds represented by the discount rate are omitted as are fixed cost. Verbaken et al. (2014) use, in effect, interest income tier no variable or fixed costs. Stewart (2010) uses interest received less interchange fees plus other fees received.

In principle expected profit can be thought of as the present value of expected net cash inflows, each of which is expected income less all expected costs of granting a loan. Credit scores are, in practice and in the literature, the probability that an account will move into a default state anytime within a fixed time horizon. Typically, this horizon is 12 or 18 or 24 months. Using information about amount outstanding at the time of default and the proportion of the balance lost in the event of default, together with the probability of default enables a prediction of expected income and expected loss over the time horizon of 12, 18, or 24 months to be made. However, this is not the total profit even in this window since other fixed and variable costs also have to be subtracted.

In this thesis we consider the expected income and expected losses of principal over a limited time horizon, but not over the entire life of an account and we do not consider other fixed and variable costs. For this reason, we denote expected income less expected losses as expected net income. The aim of this thesis is to develop models to predict, at the time of a loan application, this net income of an account. Notice however, that much of the literature would define what we call ‘net income’, as profit.

The idea of ‘profit scoring’ developed from the credit default scoring because of tasks faced by bank management. One of the main tasks of credit portfolio management is ‘profit’ maximisation. For example, Thomas et al. 2002 demonstrated that credit limit management can be applied for profit maximising decisions based on default scores. The optimal cut-off setting approach in a portfolio level model can be based on the revenue and costs equation (Thomas, 2007). Anderson (2007) investigates the credit risk management cycle and the credit scoring role in wider credit processes, and proposes profit-based cut-off setting, based on a profit modelling approach.

Stewart (2010) identifies four complexities in profit scoring or profit prediction for credit cards. The first problem is the variety of definitions of ‘profit’, and lack of a single definition and understanding by financial institutions of what is ‘profit’ at an

account level. The second problem lies in the correlation between risk and profit, as high-risk loans often have also high profitability. In this case, a profit scoring model does not predict a strong contribution to the lender in the decision-making process. The third problem is the shape of profit distributions. Inactive cards do not make a profit and low-activity cards make profit close to zero. Credit cards in a write-off state usually have outstanding balance close to the credit limit, and consequently negative profit equal to charge-off amount. So, the profit distribution across accounts has a gap between zero and a concentration of the values where there are amounts charged off. On the other hand, the profit distribution generally has a long smooth tail whose profit is positive. In the case of a credit card portfolio it is expected that a low number of cards would generate high profit with high concentration of profit values is close to zero. Thus, the profit distribution does not have the normal distribution but rather has an exponential one. The fourth problem is the population instability, which has more significant impact on revenue and churn than on charge-off prediction. Moreover, profit models are sensitive to changes in pricing and fees.

‘Profit’ scoring has a number of implementations at a portfolio level, for example, as a Markov Decision Process (So and Thomas , 2008; So and Thomas, 2011). At an account level it may be predicted as a survival model (Andreeva et. Al, 2007, Ma et.al, 2010), the Heckman two-stage procedure (Banasik and Crook, 2001), logistic and linear regression for the utility function (Finlay, 2008), genetic algorithms (Finlay, 2010), logistic regression for spend (Stewart, 2010) etc.

One can follow different methods to predict profit. One can predict income and deduct charge-off losses, so profit can be both positive or negative, and is calculated as a sum of inflows from interest rates, fees, commissions etc. and outflows (losses) from loans charged-off. In this case, the final profit amount from total loan activity can be estimated with a single model. If we consider the prediction of revenue (income) for prediction, we use inflows from interest rates, fees, commissions, etc. only for the definition of profit, but charge-offs (losses) can be deducted as a separate step for the final profit amount estimation at an account level. In this thesis, we use the second approach. We build a model for the revenue amount only, but do not consider losses from charge-offs and other costs. Instead we deduct from the income, estimated

Expected Losses according to the Basel concept that is the product of the estimated values of the probability of default, loss given default, and exposure at default.

There is almost no literature that describes the profit prediction or profit scoring methodology that lenders usually use. However, casual empiricism indicates that traditionally, banks use the probability of default (PD) prediction scorecards for decision making and provisioning purposes as well as loss given default (LGD) and exposure at default (EaD) estimation models. Regulators also use the same type of risk assessment models and indicators for capital adequacy control as, for example, in the Basel requirements. Expected Loss and Economic Capital indicators are based on PD, LGD, and EaD estimations. These indicators are important for banking system stability control, for provisioning purposes and capital allocation for expected and unexpected losses. However, for commercial banks and other financial lending institutions, the main aim of which is the profit generation, the revenue and profit prediction can be used with the expected losses estimation.

Generally, from the accounting point of view, in anyone time period profit is an economic value, which reflects the financial result of commercial activity as income (revenue) less costs (expenditure). In this case, in practice profit scoring at the account level is typically used to predict the total result for a loan from revenues generated by the paid interest and losses from defaults, or non-payments. Thus, the profit scoring model consists at least of two parts: income and loss. However, often for the profit scoring purposes, the profit is defined as the accumulated cash flow from the account for some period. In this case, the expected loss component is not considered, or included in the total cash flow. The commercial banks and financial institutions for strategy in the lending process may put profit optimisation at the forefront. In the case of satisfaction of regulatory requirements, profit management using profit scoring can be used instead of traditional strategies, built on credit risk scoring.

This thesis presents an approach to credit card account-level profitability or rather *net income prediction*, based on multistate and multistage conditional probabilities models with different types of income, and compares methods for the most accurate prediction. We use application, behavioural, card state dynamics, and macroeconomic characteristics, and their combinations as predictors.

1.4 General Model description

The primary purpose of this project is to design a general model for predicting total net income over a period of credit card activity, that combines various net income sources and behavioural types of cardholders, and to assess the predictive accuracy of various regression methods applied for each sub-model with panel data. The logical schema of the total net income model is shown in Figure 1.3, and the general methodology is presented below.

In our approach expected net income is computed as expected interest income plus expected non-interest transactional income less expected losses.

The computation can be summarized as follows:

Net income = total transactional income + total income from interest – expected losses

Total transactional income = Transactional income from Point-of-Sale (POS) + transactional income from Automated Teller Machine (ATM)

Transactional income from POS = Probability of POS transaction x income from POS transaction

Transactional income from ATM = Probability of ATM transaction x income from ATM transaction

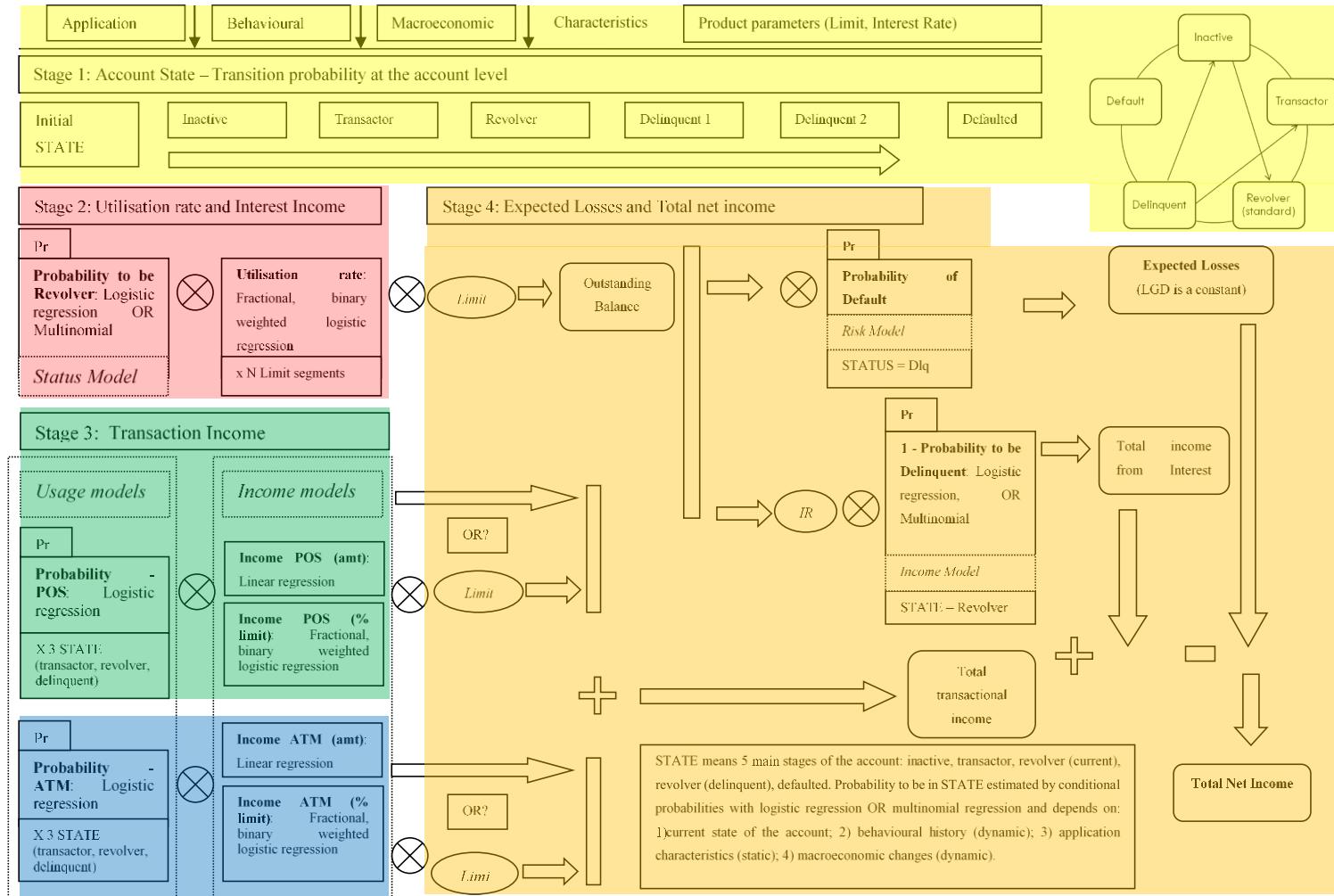
Total Income from Interest = Outstanding Balance x Interest rate,

where

Outstanding Balance = Probability of being a revolver x Utilisation rate x Credit Limit

Expected Losses = Probability of being a revolver x Utilisation rate x Credit Limit x Probability of Default

Figure 1.3. General model schema



At its highest level, the model consists of five stages: i) prediction of account (consumer) behavioural state with conditional transition probabilities, ii) estimation of outstanding balance and interest income, iii) estimation of non-interest, or transactional, income, iv) estimation of expected loss, and v) total net income estimation (see Figure 1.3).

The model input consists of two types of factors: characteristics (or predictors), and constants. Characteristics come from three sources: i) loan applications at account level from a bank's application processing system; ii) core or accounting banking systems, aggregated into the data warehouse at the account level, and iii) the national statistical bureau. Application data consists of consumer socio-demographic characteristics such as age, gender, education, marital status, residence, a region of residence, number of family members, and others, as well as economic factors such as monthly income, sources of income, and spousal income. Behavioural data contains dynamic (historical) information about consumer transactions, balances, and delinquencies, such as outstanding balance at the end of the month; average monthly outstanding balance, number of debit and credit transactions, average, maximum and minimum transaction amount per month, days past due counter, and arrears amount at the end of the month. Finally, data from the national statistical bureau contains macroeconomic indicators, such as GDP, CPI, unemployment rate, and foreign currency exchange rates. All of the required behavioural and macroeconomic data was collected on a monthly basis.

The model also uses constants, which originate from the application and behavioural data, and primarily describe product characteristics such as interest rate, loan amount, and loan term. However, because the current research focuses on credit cards (credit lines), credit limits can vary over time. Thus, the credit limit value is a predictor, not a constant, in the forecasting equations. A description of the data appears in Chapter 3.

The first stage of the net income prediction model is the prediction of the account behavioural state via transition probabilities between states at the account level. The procedures are shaded in font yellow in the Figure 1.3. Traditionally, a transition matrix is used to calculate the transition probabilities at the portfolio/pool level. However, in this case, the number of states for transition is greater than two; for

example, a revolver can be a transactor or stay a revolver, become delinquent, or become inactive. The number of transition probabilities is $N-1$, where N is the number of possible states for transition. Thus, in this research, two multi-target response models—conditional logistic regression and multinomial logistic regression—are used to predict the probability of transition. Description of the methods is continued in Chapter 5, and estimation results are in Chapter 6. The model scheme contains only main states. The revolver repaid state and delinquent 2 state are used in the net income estimation and in the scheme they are considered as the revolver state and delinquent state respectively, but have different sets of transitions and the probabilities of transitions from and to other states.

To make the predictions we include application, behavioural, loan, and macroeconomic characteristics and use an account level panel data set. The output consists of the estimated probabilities of transitions between possible states, for each account, at each available point in time.

The second stage (shaded red in Figure 1.3) involves estimating outstanding balance and interest income. The estimated outstanding balance is calculated as the product of credit limit and estimated utilisation rate. The outstanding balance at the end of a month (or due date) can be positive for revolver consumers only. Thus, the probability of being a revolver is the result of the first stage transition estimation. The estimated utilisation rate is used to predict outstanding balance. Because the utilisation rate is bounded between 0 and 1, it is necessary to apply this estimation technique to keep the outcome in this range.

The second stage inputs are the application, behavioural, loan, and macroeconomic characteristics of the account level panel data set, and the account states as the result of the first stage (as the second stage is calculated solely for revolvers and early delinquent states). The output of Stage 2 consists of the estimated utilisation rate for outstanding balance prediction, which is used for interest income and estimated loss.

The utilisation rate model is described in Chapter 4 and the probability of being a revolver model is described in Chapter 5 and 6.

The third stage involves non-interest income estimation from two sources: interchange fees (from point-of-sales (POS) purchases), and ATM cash withdrawal fees. There are

two approaches to the modelling here: i) direct prediction of income amount, and ii) indirect estimation of percentage of credit limit. For interest rate income, the second approach is used; however, in this stage, direct prediction of the transaction income amount is used. Estimation of transaction income depends on the probability of the customer POS transaction shaded green in Figure 1.3 or ATM withdrawal shaded blue in the corresponding period. The probability of transaction type is estimated using a binary logistic regression for each possible transaction status: transactor, revolver and delinquent. Third stage inputs are the same as those in the second stage, and stage output consists of the income amount from POS and ATM transactions. The relevant models are described in Chapter 7.

The fourth stage is the estimation of expected losses. The computation is shaded orange in Figure 1.3. Expected losses equal the probability of being a revolver times the utilisation rate times limit times probability of default. This is effectively the probability of default times loss given default times exposure of default. The probability of default is a transition probability to the default state as modelled in Chapters 5 and 6. Loss given default is treated as a constant. Exposure at default depends on the expected outstanding balance at the point of default. We use the credit limit utilisation rate models to predict the outstanding balance. These models are explained in Chapter 4 and the transition models of Chapters 5 and 6 are used to predict the probability an account is a revolver.

In the fifth and final stage, total net income is calculated as the difference between the sum of interest and transactional income, and the estimated expected loss. (See Chapter 8 for a detailed description of this stage).

The set of income amount prediction models covers all possible states of a given account. The relevant income model is included in the total income amount calculation, depending on the credit card state at the prediction point $t+n$, where t is the observation point in time, n is a type of step—for example, months—from the observation point to the prediction point, or prediction horizon. The future state is defined via the probability of transition from the state at the observation point to the state at the prediction point $t+n$. Therefore, rather than assuming that an account will reach a certain state at time $t+n$, we hypothesise that an account will be in all possible

states $s_{t+n} \in S_{t+n} | s_t$ at time $t+n$, with the relevant probability of transition from state s_t to state s_{t+n} in n steps (months), where $S_{t+n} | s_t$ is the set of possible states of transitions for n steps from the state at time t , depending on the current state s_t .

Prediction of income amount from interest rate and transactions at the different account states, and prediction of the transition probability between states, are independent parts of the total income calculation process. Income amounts from various sources and the transition probabilities of account states are aggregated at the final stage.

1.5 Aims of research and research questions

The main aims of this research are as follows:

- i. to create an approach for prediction of total income and partial incomes for credit card accounts, considering the existing gaps in the literature;
- ii. to test empirically the fitting accuracy of regression methods for the prediction of the credit limit utilisation rate, multi-target probability of transition between the account states, and the transactional profit amount with a credit cards panel data sample;
- iii. to test empirically the goodness-of-fit of the credit card total income and profit prediction aggregated model using a panel data sample.

We have the following question for research according to the structure of the Chapters.

Firstly, which *estimation technique* gives most accurate prediction of utilisation rate at account level: i) linear regression (OLS), ii) fractional regression, iii) Beta-regression (non-linear), iv) Beta-transformation + OLS/GLM, or v) Weighted logistic regression with data binary transformation (Chapter 4)? The problem is that the utilisation rate distribution is U-shaped and bounded between zero and one. Linear regressions can give biased results at the tails of the distribution for the lowest and the highest utilisation rates. However, there is little literature that models credit limit utilisation rate.

Secondly, which approach to the prediction of *multistate transition probability* at the account level gives the most accurate results: multinomial regression, ordinal regression, or decision trees with conditional binary logistic regressions (Chapter 5 and 6)? The problem is that in the case of more than two possible outcomes for the

transition probability between account states, it is necessary either to build the multistage conditional binary probabilities model or to use a multiresponse regression such as multinomial or ordinal logistic regression. However, the empirical knowledge of the application of this method does not extend to the credit cardholders' behaviour.

Thirdly, which *modelling approach* gives the most accurate prediction of the *total income amount*: i) a direct method: total income amount prediction with a single regression, or ii) an indirect method: the simple or weighted sum of the partial income predictions (Chapter 8)? The problem is that direct prediction of the total income amounts can give lower predictive accuracy than the prediction with the partial models, which describe the various sources of the total income by its origin.

Fourthly, what changes in *the significance level of each characteristic* take place in the case of *the random effects* in comparison with pooled model for panel data when modelling transition income? What method for panel data regression analysis gives the highest predictive accuracy? (Chapter 7). The problem is that a cross-sectional analysis does not discover the time variance of predictors. Credit card usage and profitability depends on the behaviour type of the cardholder, but changes in a customer's behaviour can be specific and systematic, i.e. caused by personal factors or by macroeconomic or bank policy changes. The use of random effects in a panel model can reveal the significance level of independent covariates. However, random-effect regression methods usually show poor predictive accuracy in comparison with pooled regression.

Fifthly, what are the goodness-of-fit values for the total income and net income predictions and according to the model for their components predictions: the interest income, the transactional income, the utilisation rate, and the probability of transition for the given data sample? (Chapters 4,6,7,8 and Conclusion). The problem is that there are few benchmarks for the accuracy for such types of parameters prediction.

Finally, what are the most significant explanatory variables for the parameters related to the credit card account profit predictions? (Chapters 3,5,6,7 and Conclusion). The main problem is that literature on profit modelling often contains a short list of the explanatory variables, and the predictors are mainly the application characteristics.

There is a lack of description of behavioural variables and their impact on the income for various types of credit card holders' behaviour in academic papers.

1.6 Thesis Contributions

This research offers both theoretical and empirical contributions. Firstly, we proposed an approach for the ex ante prediction of net income for different behavioural types of credit card applicants. We propose an approach for credit card income prediction as a system of models, which considers the estimation of income from different sources and then aggregates the income predictions weighted by the behavioural states transition probabilities. The main differences between the proposed model and existing models include: i) the estimation of individual components of net income, ii) their aggregation, and iii) consideration of the transition of account behavioural states in the total net income estimation.

First, the usage of credit cards is a topic discussed in many papers (for example, Crook et al., 1992; Banasik and Crook, 2001; Hand and Till, 2003). However, there is lack of research on the prediction of the credit limit utilisation rate. We implemented some methods already used for proportions prediction such as Loss Given Default (Yao et al., 2014; Arsova et al., 2011) and applied them to *the utilisation rate for the first time* (for example, Agarwal et al., 2006).

Second, predictive models for risk and profit parameters can be built for a credit card portfolio at the pool level with, for example, a Markov Chain (So and Thomas, 2011). However, significant differences between credit card usage types can decrease the predictive accuracy of the models, because the different forms of credit card usage have individual behavioural drivers for risk, utilisation, purchases, and profit (So et al., 2014; Tan and Yen, 2010). We tested *multitarget logistic regression* models for the probability of transition between states. We have found only Volker (1982) related to the use of *multinomial logistic regression* for the modelling of bankcard utilisation at the account level, and So and Thomas (2014) use *multinomial logistic regression* for the prediction of the transition between inactive, closed, and active account segments, and *cumulative (or ordinal) logistic regression* for the prediction the probability of transition between credit score bands. Kim, Y. and Sohn, S.Y. (2008) dedicated to the estimation of transition probabilities of credit ratings with use of

random effects multinomial regression model. This the first time that the multinomial logistic regression and multistage binary logistic regression have been used for the prediction of transition probabilities between income-based credit card states at account level.

Third, Baltagi et al. (2002) tested different random-effect variance component methods for panel data as proposed by Fuller and Battese (1974), Wansbeek and Kapteyn (1989), Wallace and Hussain (1969), and Nerlove (1971) with Monte-Carlo simulated data sample. However, our research tests the predictive accuracy of these random effect methods for panel data for transactional income for the first time.

Fourth, we have found some empirical goodness-of-fit values for a given data sample for total income and net income predictions and according to the model so also for their component prediction models of the interest income, the transactional income, the utilisation rate, and the probability of transition for the given data sample. The models, which cover the scope of income related predictions mentioned above, are estimated for the first time for a single data sample and in the scope of a single aggregated model for the total net income prediction.

Fifth, we selected the most significant explanatory variables for the parameters related to the credit cards profit prediction. We found few papers which describe customer related behavioural covariates in this context. Behavioural and loan characteristics used by Crook and Leow (2014), who used payment amount, proportion of credit drawn, indicator for improvement in state from 3 months previous, lagged 3 months. Nie et al. (2010) for panel data clustering use behavioural characteristics such as Trade times via ATM and via POS. Ju et al. (2015) described a behavioural credit scoring model with time-dependent covariates for stress testing. We used similar application and macroeconomic covariates for the net income and transition probabilities prediction. However, we have significantly expanded the list of behavioural variables compared with the literature and tried to explore how these variables drive net income.

The proposed modelling approach can be applied for the development of business strategies such as credit limit management, customer segmentation by the profitability and behavioural type, and others.

Hereinafter, the terms net income and total income are used to indicate the same.

1.7 Thesis Structure

This thesis contains eight chapters including: Introduction, Literature Review, five substantive chapters describing the methodologies and discussions of the results of the empirical investigation, and a Conclusion.

This introduction gives the key points and main aims of the current research and describes the general schema of the total net income prediction model. The total model is split into: i) the interest rate income prediction; ii) the non-interest rate income prediction; iii) the transition probabilities between states prediction; and iv) the aggregation of the partial models into total income estimation.

Chapter 2, the literature review, systematically reviews relevant literature in the area of loan profit modelling and highlights the gaps in the application of profit scoring to credit cards income prediction.

Chapter 3 gives the overview of the data set and data samples for empirical investigation, provides with brief exploratory data analysis, and describes characteristics, which are used as covariates and dependent variables.

Chapter 4 is dedicated to the prediction of the credit limit utilisation rate for credit card holders and contains the empirical investigation with the comparative analysis of the predictive accuracy of different regression models. We apply five methods for utilisation rate prediction including: i) linear regression (OLS); ii) fractional regression (quasi-likelihood); iii) beta-regression (non-linear); iv) beta-transformation + OLS/GLM; v) weighted logistic regression with data binary transformation. The fractional regression approach and quite a recent approach of weighted logistic regression with binary transformation have shown more accurate prediction in comparison with other traditional methods.

Chapters 5 and 6 are dedicated to modelling the transition probabilities between credit card states and contains the comparative analysis of several regression models for the multistate non-binary target prediction. The transition probabilities topic is split into two Chapters – methodology and empirical investigation. Chapter four describes the general model setups, model building methodology such as transition probability prediction with conditional binary logistic, ordinal and multinomial regressions. Chapter five discusses the data sample description, regression estimation results for all

types of regression and a comparative analysis of the models. At the final stage of Chapter five, we propose an updated set of the credit card account states, discuss the results of the multinomial model estimations, and the final model choice for the net income prediction. The binary conditional logistic regression gives a more accurate prediction but is sensitive to the order of transitions. Thus, the multinomial regression has been selected for modelling as an approach, which gives more balanced estimations and is easier to implement.

Chapter 7 describes an approach to the non-interest rate, or transactional, income prediction and contains a comparative analysis of panel data regression techniques. We consider two sources of non-interest income generation: i) interchange fees and foreign exchange fees from transactions via point-of-sales (POS); and ii) ATM fees from cash withdrawals. Two periods for the predicted income are used in the models: one month and six months. We compare the predictive accuracy of a one-stage approach, which means the usage of a single linear model for the income amount estimation, and a two-stage approach, which means that the income amount estimation is conditional on the probability of POS and ATM transaction. Finally, we test different definitions for the target in regression models as the direct estimation of the income amount and indirect estimation as the share of the credit limit and the outstanding balance.

Chapter 8 aggregates the results from the partial models into a single model for total net income estimation. We assume that a credit card account holder does not have a single particular state and a single behavioural type in the future but has a chance of transition to any of several possible states. The income prediction model is selected according to these states, and the transition probabilities are used as weights for the particular interest rate and non-interest rate income prediction models. The results are the comparative analysis of goodness-of-fit of the total income direct estimation, the simple sum of partial models for interest and transactional income, the sum of state income partial models weighted by the individual transition probabilities and the probabilities from the transition matrix. To finalise the chapter, we give the estimation of the total profit for the credit portfolio based on the estimated income and loss.

The Conclusion contains the summary of the empirical results and highlights the contributions of this research.

2 Chapter Two. Literature review

2.1 Introduction

This chapter gives an overview of literature in credit risk and profit scoring, credit cards usage, and transition probabilities modelling, which contributes to forming the theoretical foundation of this study. Credit Scoring (Section 2.2) gives the basis for Profit Scoring development (Section 2.3). Markov Chains and transition probabilities (Section 2.4) are widely used for credit risk and profit prediction. Modelling of credit cards usage and credit limit utilisation rate (Section 2.5) applies to interest income prediction. Interest and Transactional income modelling literature is discussed in Section 2.6. Section 2.7 outlines gaps in the literature and concludes.

2.2 Credit Scoring

2.2.1 The probability of Default modelling

Credit scoring as the quantitative methodology of credit risk assessment arose from classification problems like quadratic discriminant analysis. British statistician, biologist and geneticist Fisher (1936) introduced the idea of discriminating between groups in a population. Durand (1941) provided the first research in the application of this technique to credit risk assessment. The problem was set as a binary task to classify credit applications as good or bad ones. This idea remains the current approach without significant conceptual changes. Further approaches to lending decisions have mostly been investigated as a binary classification problem with a well-described mathematical background such as Greene (1992), Crook et al. (1992b), and Thomas et al. (2017). In spite of a number of articles and papers, for example, Myers and Forgy (1963), Lewis (1992) and Rosenberg and Gleit (1994), Greene (1992), Hand (1997), Crook et al. (2007), credit scoring does not have a single established methodology and standard rules. It stays mainly as a system of various approaches to the credit risk assessment. Some internal working papers of consultancy companies such as Fair Isaac and Experian contain standardised industrial approaches to the development and validation of credit scorecards.

Thomas et al. (2007) give an overview of the history of credit scoring and the evolution of the modelling methods. Recently, the academic framing of this topic was

summarised by Thomas, Edelman and Crook (2017), and Anderson (2007). Thomas et al. (2017) gives a systematic overview of credit scoring and discuss as statistical and non-statistical methods in scoring systems, the development process as logistic regression, decision trees, application and implementation processes of scorecards. It also describes the behavioural scoring based on Markov chains and profit scoring. Siddiqi (2005) and Mays (2004) have described the common procedures for scorecard development and implementation in a more practical and business-oriented sense, without strict mathematical expositions, but their works became popular introductory books in the credit scoring area for bankers. Anderson (2007) discussed all theoretical and practical aspects of credit scoring development, implementation, and usage processes, the regulatory environment and credit risk management cycle. There are several papers which contain a review of current research topics in consumer credit risk assessment, such as Crook, Edelman and Thomas (2007) or credit scoring development stages, such as Thomas (2000). The results of these authors provide the basis for much of the following credit scoring researches and academic programmes.

Hand (1997) presented a review of statistical methods traditionally used for credit scoring such as discriminant analysis, linear and logistic regression, mathematical programming, neural networks, and time-varying methods. He highlighted the nearest neighbour methods as one of the best for prediction, but mentioned that the method selection depends on the problem. Hand (1997) gave the set of characteristics, which are the most common for application scoring, such as the time at present address, home status, customer's annual income, type of bank account, age, marital status, time with employer, but he did not mention predictors for behavioural models. However, it was notified that improvement could be made rather not in statistical methods, but in finding new predictors, which reflect different sources of data, and further development can be concentrated in the selection of the new areas of credit scoring application such as pricing models, fraud detection, and profitability scoring.

Myers and Forgy (1963) used discriminant analysis for the development of a scoring system to predict credit risk for conditional sales contracts on mobile homes. They selected four weighting systems: conventional discriminant analysis, stepwise regression, equal weights for all predictive variables, and discriminant analysis weights based upon selected subsamples of cases. The predictors included traditional

characteristics such as age, marital state, number of dependents, total monthly income, existence of home and work phone, time at present job as well as the characteristics, which are specific for the product, for example, auto year, length of trailer, the purpose of buying trailer for pleasure, new parking address, and mailing address different than new parking address. They assessed the effectiveness of the scoring system with two dimensions: correlation of actual and predicted scores and a number of bad accounts which would be eliminated at the cost of the indicated number of good accounts. The best results, except simply equal weights, were obtained from discriminant analysis weights based upon selected subsamples of cases, which was developed to improve discrimination power at the lower score levels.

The discriminant analysis of dichotomous variables can be performed with loglinear models (Lachenbruch and Goldstein, 1979), which allow one to use goodness of fit statistics for model building and variables selection using discrete variables, and to use information about the ordered structure of categorial variables.

Wiginton (1980) compared a logit model and a discriminant model. He used two subsets of variables: demographic, for example, number of dependants, homeowner or renting, pleasure or business use of vehicle, and economic, for example, industry class of the place of employment, occupation type, number of year with employer. Wiginton (1980) used the number of correctly classified good and bad cases as the performance measure for comparative analysis. He concluded that the maximum likelihood estimation of the parameters of the logit function can be more applicable for credit scoring tasks than the linear discriminant model because the factors used in scoring usually are discrete and qualitative.

Lessmann et al. (2015), as development of Crook et al. (2007), found that partitioning and neural networks give the best classification accuracy. However linear regression and logistic regression show similar results and classification accuracy depends not only on the predictive techniques but on the quality of the data sample.

Logistic regression is one of the most common techniques for the credit scoring tasks (Thomas, 2000). The logit of the probability of an event is related to a linear equation of explanatory variables as follows:

$$\text{Log}\left(\frac{P_{gi}}{1 - p_{gi}}\right) = \beta_0 + \boldsymbol{\beta}^T \mathbf{x}_i$$

where p_{gi} is the probability of being good for i case, \mathbf{x}_i is a vector of covariates, and $\boldsymbol{\beta}$ is a vector of parameters for estimation.

The maximum likelihood approach is used to estimate the parameters of the logistic regression and the likelihood function can be defined as follows (Greene, 2002):

$$\ln L(\boldsymbol{\theta} | \mathbf{y}) = \sum_{i=1}^n \ln f(y_i | \boldsymbol{\theta}),$$

where $L(\boldsymbol{\theta})$ is the likelihood function, evaluated at an unknown parameter vector $\boldsymbol{\theta}$; $f(y_i | \boldsymbol{\theta})$ is the probability density function for a random variable y_i .

The probability of being good for case i for given parameters is calculated as follows:

$$p_{gi} = \frac{e^{\boldsymbol{\beta}^T \mathbf{x}_i}}{1 + e^{\boldsymbol{\beta}^T \mathbf{x}_i}}$$

Generally, the credit score, which is built with logistic regression, can be defined as follows (Thomas and Malik, 2010):

$$s(\mathbf{x}) = \log\left(\frac{p(G | \mathbf{x})}{p(B | \mathbf{x})}\right) \Leftrightarrow p(G | \mathbf{x}) = \frac{1}{1 + e^{-s(\mathbf{x})}},$$

where $s(\mathbf{x})$ is a score depending on the vector of characteristics \mathbf{x} and $p(G | \mathbf{x})$ the probability of being good conditional on covariates \mathbf{x} .

However, in recent times machine learning techniques are becoming more popular for credit scoring. Thus, machine learning individual classifier methods such as neural networks and Support Vector Machines (Crook et al., 2007), and homogenous ensemble methods such as gradient boosting and random forest, and heterogeneous ensemble methods such as probabilistic model for classifier competence and stacking (Lessmann et al., 2015) have demonstrated higher predictive accuracy than traditional regression methods with logistic regression.

The traditional background of credit scoring is the classification of the applications into likely good or bad and the prediction of the probability of default. However, risk

management requires more advanced estimations. Basel (2004; 2011) proposed the estimation of the expected losses as the product of the Probability of Default (PD), Loss Given Default (LGD), and Exposure at Default (EaD), both at the portfolio/pool level and at the account level depending on the Basel accord implementation stage as follows

$$EL = PD \times LGD \times EaD.$$

Thus, papers which discuss the LGD and EaD risk parameters estimation at the account level, also are related to the credit scoring topic.

Many techniques for default probabilities assessment have been developed, such as linear and logistic regression, decision trees (chi-squared automatic interaction detection, or CHAID), neural networks, genetic algorithms, expert systems, survival analysis. (Thomas et al., 2002). For instance, Desai et al. (1997) apply neural network and genetic algorithms for credit scoring. However, the logistic regression is the most popular method. The standard logistic regression with the binary outcome is used to assess the probability of default of the loan in full lifecycle.

2.2.2 EaD modelling

The Exposure at Default (EaD) parameter is an important component of the Expected Loss estimation. For standard loans it can be quite a trivial task of the estimation of the outstanding balance according to the payment schedule and possible prepayments and arrears may impact on the outstanding balance volatility. However, for credit cards the prediction of the outstanding balance at the time of default can be a complex task because a cardholder theoretically may have any outstanding balance in the range between zero and the credit limit.

Initially, Exposure at Default requires the estimation of the outstanding balance. Kim and DeVaney (2001) predicted the amount of outstanding balances among credit card revolvers. They applied the Ordinary Least Squares (OLS) method for amount prediction with the Heckman procedure to correct for sampling biases. It was found that the sets of characteristics related to the probability to have an outstanding balance and to the amount of outstanding balance are different. The result seems expected and logical because higher income and desire for a higher standard of living lead to an increased outstanding balance, but this is not necessary for the probability of credit

card use. Kim and DeVaney (2001) used the characteristics similar to those used for the probability of default estimation in credit scoring such as age, marital status, education, income, the number of credit cards. However, they also used unusual predictors, which are related to the customer expectations and behaviour type, such as income expectation, positive attitude toward the use of credit for vacation, positive attitude toward the use of credit for luxury goods, loan payment habit (behind schedule or missed payment). These variables were used as logical binary variables and have significant parameters in regression model.

Qi (2009) performed an empirical investigation of Loan Equivalent Exposure (LEQ) Factor, the Credit Conversion Factor (CCF), and the Exposure at Default Factor(EADF) to estimate the Exposure at Default (EaD) parameter at the account level. The covariates used in regression are mainly behavioural variables: amount values, for example, Monthly payment amount on the account 12 months prior to default, Highest balance ever attained on the account up to 12 months prior to default; ratios, for example, Account credit limit utilization rate ($Bal_Amt0/Cred_Amt0$) 12 months prior to default, Aggregate balance to credit limit ratio for open bankcard tradelines (12 months prior to default); numbers (or counts), for example, Number of inquiries within 6 months (observed 12 months prior to default); and dummy variables, for example, Dummy variable for accounts that were closed 12 months prior to default, Dummy variable for accounts with utilization rate $>95\%$ 12 months prior to default. Qi (2009) used two methods of variable selection: include all variables in the model and use a stepwise procedure to select the most significant predictors. The separate models were developed for accounts, which had a current state 12 month prior to default, and which were delinquent. High utilisation rate, account age and balance became the main LEQ driver for both segments. However, other characteristics were occurred significant either for current or for delinquent accounts only. Thus Qi showed that it makes sense to apply different models and sets of predictors for different account states.

Jacobs (2010) empirically tested the LEQ Factor, CCF, and EADF with S&P and Moodys Rated Defaulted Borrowers Revolving Lines of Credits, so the corporate data was used. He found that the EADF factor had shown the highest predictive accuracy,

the CCF factor – the worst one, and the LEQ factor had predictive accuracy between EADF and CCF.

Leow and Crook (2016) investigate a mixture models for credit cards EaD estimation and predict the outstanding balance not only at the default point, but for defaulters and non-defaulters. The expected balance $B_{it\tau}$ for account i at time τ is defined as sum of three products of the probability $P(\bullet)$ the outstanding balance will be below, equal, or above the credit limit $L_{it\tau}$. They estimated balance for these cases at the default time as follows:

$$\begin{aligned} E(B_{it\tau} | d_i = 1) &= (P(B_{it\tau} > L_{it\tau} | d_i = 1) \times E(B_{it\tau} | B_{it\tau} > L_{it\tau}, d_i = 1)) \\ &+ (P(B_{it\tau} = L_{it\tau} | d_i = 1) \times E(B_{it\tau} | B_{it\tau} = L_{it\tau}, d_i = 1)) \\ &+ (P(B_{it\tau} < L_{it\tau} | d_i = 1) \times E(B_{it\tau} | B_{it\tau} < L_{it\tau}, d_i = 1)). \end{aligned}$$

They found that mixture models give more accurate estimations in comparison with other methods such as Loan Equivalent Exposure (LEQ) Factor, the Credit Conversion Factor (CCF), and the Exposure at Default Factor(EADF). The estimation of the credit limit and balance are both important for EaD prediction and their panel model with random effects had a R-squared 0.9 for limit and 0.47 for balance estimation. However, some empirical findings show that explanatory variables such as time in address and time with the bank, which usually are significant for the probability of default estimation, can be insignificant for limit and balance modelling.

A mixed set of application (time-independent) variables \mathbf{x}_i , behavioural (time-varying) variables \mathbf{y}_{it} , and macroeconomic factors \mathbf{z}_t can be used in a single model as a linear regression equation. Leow and Crook, 2016) use a mix of application, behavioural, and macroeconomic variables in a single regression model. They selected a quite traditional set of application variables such as age, income, employment group, time with bank, and behavioural variables lagged 12 months for credit cards EaD prediction: average transaction value, number of cash withdrawals, amount of cash withdrawals, credit limit, rate of total jumps, proportion of months in arrears, repayment amount, and outstanding balance, which have demonstrated good significance.

2.2.3 LGD modelling

Crook and Bellotti (2009) use application variables such as age, time with bank, time at address, income, number of cards, demographic group; account variables specific for recoveries modelling such as months on book at default points, balance at default; and macroeconomic variables such as bank interest rate, unemployment rate, and earnings growth (log), and interactions between application and macroeconomic variables for LGD prediction. They tested several regression models such as Ordinary Least Square(OLS), Tobit, Fractional logit transformation, beta distribution transformation. The best predictive accuracy was surprisingly given by an OLS model with interaction of application and macroeconomic covariates.

LGD is bounded between 0 and 1 and requires appropriate methods to keep the predicted value in this range. One of the techniques that has been used is fractional logit regression proposed by Papke & Wooldridge (1996). The Bernoulli log-likelihood function is given by

$$l_i(\mathbf{b}) = y_i \log[G(\mathbf{x}_i \mathbf{b})] + (1 - y_i) \log[1 - G(\mathbf{x}_i \mathbf{b})],$$

where $G(\bullet)$ is a known function satisfying $0 < G(z) < 1$ for all $z \in \mathbb{R}$. $G(\bullet)$ is chosen to be a cumulative distribution function, typically logistic function.

The quasi-likelihood estimator of β is obtained from the maximization of

$$\max_{\mathbf{b}} \sum_{i=1}^N l_i(\mathbf{b})$$

Crook and Bellotti (2009) apply the Fractional logit transformation for the Loss Given Default parameter modelling:

$$T_{RR} = \log(RR) - \log(1 - RR),$$

where RR is recovery rate and $RR = 1 - LGD$.

The use of the weighted logistic regression with binary transformation of the data sample (Arsova et al., 2011; Barkel and Siddiqi, 2012) is a relatively innovative approach. The logit function is bounded between 0 and 1 and usually applied for the prediction of probability. To apply logistic regression, which uses the binary distributed target, a variable need to be transformed from continuous to binary form.

LGD can be considered as the probability to use the credit limit by 100%. For example, the rate 75% can be presented as 75% probability of full use of the credit limit and 25% probability of zero use of the credit limit. This approach is used by Barkel and Siddiqi (2012) for Loss Given Default prediction. Each observation is presented as two observations (or two rows) with the same set of predictors according to the good/bad or 0/1 definition, which is used in logistic regression. The outcome with target 1 corresponds to the rate r , which determines the weight equal to rate r . The outcome with target 0 corresponds to the rate $1-r$, which determines the weight equal to $1-r$. The logistic regression probability of event is the rate estimation.

Stoyanov (2009) presented the approaches to the LGD account level prediction. He compares, in particular, such methods as a binary transformation of the LGD using uniform random numbers and a binary transformation of the LGD using manual cut-offs, and recommends for usage the stepwise logistic regression model with uniform random numbers transformation of LGD to a binary variable. However, the empirical results have been provided in the subsequent investigations. Arsova et al. (2011) applied both direct approaches to the LGD modelling such as OLS regression, beta regression and fractional regression and indirect approaches such as logistic regression with binary transformation of LGD by random number, logistic regression with binary transformation of LGD by weights, and also multi-stage models like Ordinal Logistic Regression with nested Linear Regression. The best fitting accuracy for a validation sample for credit cards has been obtained with Fractional Logit and Binary Logistic Regression with weighting transformation – R-squared 0.148 for both methods. However, these results show poor predictive accuracy and do not perform significantly better than other tested models.

Generally, logistic regression matches the log of the probability odds by a linear combination of the characteristic variables as

$$\text{logit}(p_i) = \ln\left(\frac{P_i}{1-p_i}\right) = \beta_0 + \boldsymbol{\beta} \cdot \mathbf{x}_i^T,$$

where

p_i is the probability of particular outcome, β_0 and $\boldsymbol{\beta}$ are regression coefficients, \mathbf{x} are predictors.

$P_i = E(Y_i | \mathbf{x}_i, \beta) = \Pr(Y_i = 1 | \mathbf{x}_i, \beta)$ is the probability of the event for observation i .

This approach can be interpreted as the following: the LGD equal to 10% is the same as from 100 accounts with 10 accounts having LGD equal to 100% and 90 accounts having LGD equal to 0%.

Some regression methods can be used for Loss Given Default prediction with some modifications as well as new machine learning techniques (Yao et al., 2014; 2015). Yao et al. (2015) compare several regression methods such as Linear Regression, Fractional response regression, linear regression with beta-transformation, Two-stage models, Least-squared Support Vector Regression, Least Squared Support Vector Regression with a Logistic Transformation, and other methods for LGD prediction at the account level. Methods based on Support Vector Regression (SVR) have demonstrated the highest fitting accuracy with R-squared around 0.6 while Linear Regression and OLS with beta-transformation have shown R-squares equal to 0.28 and 0.1 for the combined results of a segmented model. Fractional regression has the highest result among methods, which are not related to SVR, with R-square around 0.34. However, these models contain recovery, accounting, and macroeconomic variables and do not related to retail lending.

Lotterman et al. (2012) compared both linear methods such as Ordinary Least Squares (OLS), Beta-regression, Beta-transformation plus OLS, Box-Cox transformation, and non-linear methods such as Regression trees, Least squares support vector machine, Artificial neural networks for LGD with loan-level data from financial institutions. They also apply combined two-stage (mixture) modelling approaches such as logistic regression for the estimation of the probability of LGD ending up at the peaks [LGD=0] at the first stage, linear regression for estimation of LGD values for $LGD > 0$ at the second stage, and to make the final prediction a weighted average LGD in the peak and the estimate produced by the second-stage model with weighted given by their respective probabilities. Lotterman et al. (2012) gained empirical results that Artificial neural networks and mixture modelling approach OLS+Non-linear models demonstrate the highest predictive accuracy.

2.3 Generic methods for Profit scoring

Since the 1990s the idea of using credit scoring methods for profitability prediction has developed as an evolution of existing approaches and techniques. Thomas (2000) provided an initial overview of the literature on profit scoring. He showed the prerequisites for the evolution of risk credit scoring and defined four groups of approaches to profit scoring.

The first approach uses the existing scorecards, which estimate default, usage, acceptance, and attrition. The profit for groups of the population is modelled as depending on the scores given from these measures. Oliver (1993) is one of the founders of this approach who proposed to use decision rules based on a transaction profit score and a default score. Li and Hand (1997) proposed to use indirect variables for outcome estimation instead of direct modelling of profit and losses.

The second approach mentioned by Thomas (2000) is the regression approach of credit scoring to profit as a linear function of the categorical application from covariates (see, e.g. Lai & Ying, 1994) with censored data. However, this methodology has not been developed further.

The third approach is based on Markov chains stochastic models applied as one of the methodologies for behavioural default scoring. This approach is used as one of the basic and successful techniques for profit scoring modelling (see Cyert et al., 1962; Thomas, 1994; So and Thomas, 2011).

The fourth approach to profit scoring is survival analysis, which estimates the reliability of machines and people. Narain (1992) and Banasik et al. (1999) were the first researchers to propose applying this technique to credit scoring. They considered along with borrower time to default the early payoff aspect. Such incorporation of default risk and attrition components in one model is of particular importance for profit scoring. Stepanova and Thomas (2001) introduced significant contribution to survival analysis usage for credit scoring (see section 2.3.5. Use of Survival Models).

2.3.1 Profit modelling

One of the first stages of the research in profitability scoring is the development of a profitability scorecard, which means the function or model of the set of profit parameters, such as revenue (or usage), response (or cross-sales), retention (decision

to stay or to go) and risk (or default), named as the matrix approach according to Anderson (2007). Finlay (2009) proposed a similar approach as the combination of individual components of the profit and single aggregated models of the profit contribution. Thomas (2007) investigated the use of the Markov chains for the analysis of the customer behaviour. Despite a number of advantages, this Markov decision model of profitability has a limitation conditioned by homogeneity and applicability on a portfolio level.

Profitability has different driving parameters. For standard loans with an instalments schedule, for example, the most obvious characteristics are attrition (or retention, business component) and expected losses (risk component) that cause the deviation of the actual profitability level from the planned one. Expected losses according to Basel II (2004) consist of three main components: the probability of default, loss given default and exposure at default.

For profit modelling, Finlay (2008) proposed replacing the binary classification models with continuous models that considered the applicability of such models across different types of credit product. This helps to reflect the various sources of the revenue at different points in the customer relationship. Finlay (2009) subsequently developed this idea and applied genetic algorithms to consumer behaviour modelling and, even though the prediction results were appropriate, they did not significantly exceed the traditional industry regression model results. He identified weaknesses of traditional classification approaches to credit scoring such as logistic regression. The use of binary logic about the positive contribution of the profit can lead to misspecification of the problem in case of good accounts generating a loss and bad accounts generating a profit despite their possible repayment classification (Finlay, 2010). A contribution from each account i is defined as the difference between return (fixed proportion of gross payment N) and loss (fixed proportion of outstanding balance K)

$$C_i = \alpha N_i - \beta K_i = R_i - L_i$$

The results showed that scores generated from modelling individual components of profit contribution, such as default probability, bad debt and revenue, outperform traditional classification models used to produce estimates of the likelihood of default.

Stewart (2010) investigated the difficulties of building and implementing profit models for credit cards in comparison with cost or charge-off models and proposed the methodology for modelling credit cards outcome. His profit-based scoring system consists of two models: a cost model and a revenue model. He has noted that direct modelling of profit is useless because of the strong correlation between profit and charge-off. To avoid the correlation, Stewart (2010) proposed to segment accounts by risk to control charge-off before profit model building. Also, he criticised the approaches such as neural networks and Markov chains because of their high-cost of implementation and difficulties with updates. At the same time, his dual score-based charge-off and spend model is easy to calibrate and for tracking, both of which are important for practical usage in flexible and fluctuating economic environments. The important thing in the context of my research is that modelling issues in this work are considered for credit card portfolios.

Bailey (2004) paid attention to the marketing essence of credit cards management strategies and divided models into three categories: risk, reward and retention. He also remarked that ‘credit risk is a driver of profitability’ (Bailey, 2004). Bailey (2004) proposed a potential strategic segments table for different marketing activities. There are three dimensions: risk, retention and revolution in the classification approach.

The profit maximising decisions and default-based scores described by Thomas (2002) has become one of the key points in optimal dynamic modelling. Later Thomas (2009) explained an approach for optimal cutoff setting in a portfolio level model with predetermined equity capital. Anderson (2007) investigated the credit risk management cycle and the credit scoring role in the whole credit process, and he proposed a profit-based cutoff setting, based on a profit modelling approach.

2.3.2 Efficient cut-offs

Oliver and Wells (2001) proposed one of the examples of efficient cutoff policies in the context of the account acquisition problem. They build a cutoff strategy on expected losses vs expected profit curve instead of the standard bad rate vs acceptance rate. As a result, they came to the problem of maximising expected profit subject to a lower bound constraint on expected volume. They showed the difference between a traditional risk management policy where the efficient decision represents an odds cut-

off that is higher than the profit-maximizing one and is a reason to have a decreased credit risk level and decreased volume of a portfolio. Their new approach where optimal decisions can lead to low risk on individual accounts, or to higher risk account to increase market share, both satisfying profit-maximisation criteria.

Lieli and White (2010) proposed an approach to the construction of credit rules based on profitability maximisation principles with the use of credit scoring. They look at the construction of the optimal loan approval rule in the context of a binary decision/binary outcome problem and estimate it with the use of a context-specific cutoff for credit scores or, in other words, the conditional probability of compliance/default. The resulting decision rule in the case of conditional probabilities should in theory lead to more profitable lending decisions that can be obtained if the model were estimated with a maximum likelihood method.

2.3.3 Optimal Credit Limit Policy

An optimal credit limit control, side-by-side with pricing and cutoff strategies, is one of the principal conditions of efficient, up-to-date credit risk management. The existing credit limit control approaches (Thomas, 2007) are based on the probability of default. The use of the expected losses to determine the credit limit strategy is efficient from the credit risk point of view. However, the use of the combination of profitability assessment and default probabilities in one dynamic optimisation model can give an advantage for efficient decision-making on the general entity level.

One approach to setting credit limit and cut-offs is to choose them to maximise profit. This became the evolutional stage of profit scoring origination.

Trench et al. (2003) proposed a model named PORTICO (portfolio control and optimisation). They used a Markov decision process (MDP) functional for particular time horizons – optimal discounted Net Present Value in state s and time t :

$$V_t(s) = \max_{a \in A_s} \left\{ r(s_a) + \beta \sum_{j \in S} p(j | s_a) V_{t+1}(j) \right\},$$

where A_s is the set of actions available at state s ; $r(s_a)$ is net cash flow; β is a one-period discount factor; and $p(j | s_a)$ is the transition probability of moving from state j to state s_a .

Trench et al. (2003) pay attention to whether the model satisfies the Markovian assumption to use the MDP approach. This means that the transitions from a state at time t to any state at the next period must depend on the current state only, or to be path independent. To decrease the likelihood of violation of the path independence assumption, they propose to identify variables that carried some history. ‘For example, most card issuers segment their customers as revolvers (those who carry balances), transactors (those who pay off the whole amount every month), and inactives (those who are not using the card)’ (Trench et al., 2003, p.12).

Trench et al. (2003) was one of the first papers in the series of papers in the field of the use of Markov chains for profit scoring. These papers have two distinctive features: i) transition probabilities prediction at the pool and account level, and ii) an account states definition based on behavioural score and/or customer segmentation, which contains some history. Some of these papers are discussed in the next section.

2.3.4 Game theory

Profit scoring can be based on game theory. Keeney and Oliver (2005) discuss the iso-curves idea as an extension of the win-win products approach. They propose a model that is a game between a customer and a bank. A consumer has the objective to maximise this utility, which depends on the loan amount the customer tries to optimise, and on the interest rate the customer tries to minimise. A lender has the objective to maximise profit and market share measured by the number of customers. Increasing loan amount causes profit to grow but also bads so there is a significant rise in default risk. However, it is possible to find an optimal point on the iso-preference curve to maximise the expected profit. An optimal offer for a consumer is selected from a set of offers for the constant utility of the business, which depends on market share and profit combination. This work purposes the idea that profit can be controlled by choosing the loan offer for a given probability of acceptance. This approach contributes to a business point of view by showing how a lender can select amount and interest rate taking into account the applicants likely reaction.

2.3.5 Use of Survival Models

Stepanova and Thomas (2001) introduced the techniques of proportional hazards regression for behavioural scoring models. The scores based on the Proportional

Hazards Analysis Behaviour have shown better performance than the traditional logistic regression scores for long term loans. They improved the performance of the survival-analysis models for both profit (or early repayment) and default prediction with three extensions of Cox's proportional hazards model and applied these techniques to personal loans data. They propose a new method for coarse classing with Cox's proportional hazard model, which avoids a strict time horizon for the good/bad event definition as, for instance, logistic regression requires.

Later Andreeva et al. (2005, 2007) provided the specialised research into the application of this approach for profit scoring. Andreeva et al. (2005) used the proportional hazard model. One of the applications of scoring in profitability modelling was the empirical investigation of the relationship between the net present value from a revolving credit account and time to default and second purchase. The research is based on a measure of profit which may be used together with traditional measures of predictive accuracies, such as bad rate (Andreeva et al., 2007). The model uses the probability of surviving until time T before the event such as default and a second purchase occurs. The hazard function is based on the proportional hazard function

$$h(t, \mathbf{x}) = h_0(t)e^{\beta^T \mathbf{x}},$$

where h_0 is a baseline hazard, and β is a vector of parameters to be estimated.

It assumes that covariates acted multiplicatively on time and the baseline hazard and defined as

$$h(t, \mathbf{x}) = h_0(te^{\beta^T \mathbf{x}})e^{\beta^T \mathbf{x}}$$

The distinguishing characteristic of this investigation is the application of survival analysis for profitability modelling. The profit has been predicted as a function of default and purchase propensity, where the time to second purchase was presented as credit card usage estimation. The final prediction of the net revenue uses OLS regression with the debt amount at the first purchase and survival probabilities to a second purchase and default in 25 months as predictors. Andreeva et al. (2007) demonstrated that the use of a second purchase survival analysis could improve the predictive accuracy of profit of non-default credit card accounts.

Total expected profit can be generated by different income drivers and calculated as the sum of the estimated profit from each driver. For instance, Ma et al. (2010) estimated total profit as conditional and unconditional and calculated the conditionally expected profit as the sum of four sources. These are the expected scheduled monthly payment from non-defaulted and non-repaid early customers, the expected early paid balance, the expected recovery amount from customers who defaulted but did not pay back early before time of default, and the expected receipt from insurance premia.

They applied logistic regression for the prediction of probability that a customer accepts a loan offer from a lender conditional on interest rate, loan amount, loan term, and applicant characteristics, which are standard for application scoring such as age, time with bank, occupation group code, time at address, number of dependants, monthly income, number of cards as well as bureau characteristics.

Ma et al. (2010) used survival probabilities of non-default and non-early repayment at time t for case i as follows:

$$S_{i,t}(t, \beta, \mathbf{x}) = S_0(t) e^{\beta^T \mathbf{x}_i},$$

where

$S_{i,t}$ is survival probability of event: non-default and non-early repayment (according to index),

$$S_0(t) = \exp\left(-\int_0^t h(t) dt\right),$$

$h_i(t, \beta, \mathbf{x}) = h_0(t)f(\beta, \mathbf{x})$ is a hazard function,

$h_0(t)$ is baseline hazard function,

\mathbf{x} is a vector of applicants characteristics, β is a vector of estimated coefficients.

The expected profit equation is written as follows:

$$\begin{aligned} E_{t=c}(\pi|\mathbf{x}) &= E_{t=c}(\pi|a|\mathbf{x}) \times E_{t=c}(p(a)|\mathbf{x}) \\ &\quad + E_{t=c}(\pi|\bar{a}|\mathbf{x}) \times (1 - E_{t=c}(p(a)|\mathbf{x})) \end{aligned}$$

Where

$a(\bar{a})$ are that the potential borrower accepts (rejects) the offer

π is the present value of the profits at point t equal to c.

All revenues are discounted at the opportunity cost of funds. The innovative element is that a probability of acceptance function is integrated within an expected profit function.

The empirical research was based on estimated of a Cox's proportional hazard model where effects of the interest rate and insurance were the most significant both for survival time until default and for early payment models. The loan amount was also a high-level predictive factor for early repayment function, but it was not significant individually for the default model. The results confirmed that the market share and the profit are interdependent. The optimal interest rate can be chosen from the iso-profits and iso-preference contours for a certain loan amount depending on the lender's strategy. Moreover, different market segments such as internet and non-internet can have separate iso-profit and iso-preference contours, and consequently, this can lead to different policy decisions in risk-based pricing.

Shuai and Shi (2009) provided an empirical investigation of factors that have an impact on credit card profit. They analysed the functional structure of revenue and cost and concluded that credit card profit depends on spending structure and mainly overdraft balance.

2.3.6 Measures of Profit

The measure of income and profit can be different. It can be an absolute value such as the present value of net revenue at the end of the month at the time of default and the second purchase (Andreeva et al., 2007), net revenue minus bad debt (Finlay, 2010). It can be relative values such as the relative profit measure as the customer lifetime value divided by the outstanding debt (Sánchez Barrios et al., 2014), the internal rate of return (IRR) of each loan, which is lender's effective interest rate (Serrano-Cinca & Gutiérrez-Nieto, 2016), or the expected maximum profit measure (Verbaken et al., 2014).

Also, it can be the estimation of the probability of credit offered take-up (Ma et al., 2010; So et al., 2014). Sánchez Barrios et al. (2014, p.443) defines that 'Profit scoring is directly related to the concept of customer lifetime value (CLV), which can be quantified as the net present value of the discounted cash flows generated by customers.'

Sánchez Barrios et al. (2014) use logistic regression for prediction of both the probability of default and the probability of repurchase. The probability of repurchase is in some way a development of the previous paper by Andreeva et al. (2007), which was dedicated to the repurchase (second loan) prediction and used a survival analysis for profit estimation. Sánchez Barrios et al. (2014) apply several measures for profit. The first measure is cumulative earnings before interests, taxes, depreciation, and amortisations (EBITACUM) and computed as the cumulative operational profit from interest payments and commissions per customer after deducting variable and fixed costs. The second measure for profit is the cumulative return on assets (ROACUM), where profit is defined as follows

$$ROAcum_t = EBITAcum_t / FBdef_t ,$$

where $FBdef_t$ is deflated final balance at time t.

$EBITAcum$ is used for a monetary model, and $ROAcum$ is used for the so-called relative model.

For default and repurchase models Sánchez Barrios et al. (2014) use the following set of covariates: age, location, contract, job, marital status, poor strata, education level, duration first loan, years at address, number of dependents, type of activity, type of product (or first product purchased), and credit limit usage. Only credit limit usage can be attributed to behavioural characteristics, and type of the first product and duration of the first loan can be attributed to behavioural characteristics similar to credit bureau features, which are used in application scoring models. However, the previous and current financial behaviour of a customer is not considered.

Sánchez Barrios et al. (2014) estimate direct and indirect models. For direct models, they use $EBITAcim$ and $ROAcum$ as target variables. For an indirect model they built default and repurchase predictive models for 12 months and 30 months and then use the models' outputs as covariates in Ordinary Least Squares Regressions for $EBITAcum_t$ and $ROAcum$. They use error rate as the accuracy of prediction measure, and both monetary and relative direct models outperform an indirect model.

Sánchez Barrios et al. (2014) have found that customers with low credit limit are more profitable than customers with higher credit limits most likely because of the high probability of repurchase. This type of customer behaviour is similar to the credit card

holders behaviour such as transactors and revolvers with regular spending transactions. Although the probability of default can be predicted for a 12-month period, as it traditionally is in industry, the probability of repurchase requires the longer term estimation for customer lifetime value. The use of the indirect model can be useful for a deep insight into the profit origination and the correlation between profit and explanatory variables.

The results of Sánchez Barrios et al. (2014) have become a basis for the further development of the time-to-profit models for credit cards in Sánchez Barrios et al. (2016). Time-to-profit can be measured with the cumulative profit per customer as follows:

$$EBITAcum_t = \sum_{k=1}^t \frac{EBITA_Z \times df_z}{(1+r)^y} \times (1+r)^{t-k}$$

where t is the number of months for an observation; df_z is a monthly inflation factor; and r is the cost for the company to invest funds in the credit programme.

Survival analysis and a discrete hazard model are used to predict the point in time when a customer becomes profitable for the first time since the first purchase. Different specifications of baseline hazard in a proportional odds model give various classification accuracies; however, it is quite high with AUC values between 0.83 and 0.97. A survival model was compared with OLS models from Sánchez Barrios et al. (2014) for their impact on portfolio profits and returns for different acceptance rates. Survival models have demonstrated better results for portfolio returns, but OLS models give improved portfolio profit. Sánchez Barrios et al. (2016) investigate the impact of application covariates (variables, which are time-independent and have values at the observation point) on the outcome, but do not investigate an impact of behavioural variables, which are time-varying and reflect the behaviour of an account.

Along with revenue or income-based and repayment-based measures, the difference between income and losses can be used as a profit scoring measure (for example, Thomas, 2007; Banasik & Crook, 2010; Ma et al., 2010; Verbaken et al., 2014).

Verbaken et al. (2014) propose to use the expected maximum profit measure for model performance assessment along with AUC and accuracy. This considers the use of an

optimal cut off and more profitable scoring model. The average classification profit per borrower is calculated as follows

$$P(t; b_0, c_1, c^*) = (b_0 - c^*)\pi_0 F_0(t) - (c_1 + c^*)\pi_1 F_1(t),$$

where b_0 is benefit, c_1 is cost, c^* is a cost to the company for an individual case, t is cut-off, s is score, F_0 is the cumulative density function for cases, where $s < t$, and F_1 is the cumulative density function for so-called non-cases, where $s \geq t$;

π_0 is prior probability of class 0, and π_1 is prior probability of class 1.

The expected maximum profit is defined as follows

$$EMP = \int_{b_0} \int_{c_1} P(T(\theta); b_0, c_1, c^*) \cdot h(b_0, c_1) dc_1 db_0,$$

where $h(b_0, c_1)$ is the joint probability density of the classification costs.

Verbaken et al. (2014) apply logistic regression and neural networks for the binary credit scoring models. The cut-off based on the model with EMP performance measure has demonstrated the best accuracy and a high improvement in total profit in comparison with approaches, based on No model, accuracy-based and AUC-based set up of cut-offs.

A comparison of algorithms for credit and profit scoring demonstrate that the most accurate classifier can be obtained with use of the scorecard, which is not the most profitable (Lessmann et al., 2015). The scorecard profitability is estimated by examining classification errors costs as a weighted sum of the false positive rate (FPR) and the false negative rate (FNR), weighted with costs for decisioning $C(-|+)$:

$$C(s) = C(+|-) \times FPR + C(+|-) \times FNR$$

However, the weakness of this approach, which tries to consider scorecard profitability with classification error, is that both the false positive rate and the false negative rate depend on the threshold. Lessmann et al. (2015) use Bayes optimal threshold with *a priori* probability of the bad debt and apply a set of cost ratios which consider that to grant a loan to a bad customer is worse than reject in a good customer. They demonstrated that the most accurate classifier will not necessarily increase the benefits of the usage of the model in case of high cost of the risk misclassification.

Profit scoring is applied not only for bank lending but for quite modern peer-to-peer (P2P) lending and various measures are used. Profit scoring can use Internal rate of Return (IIR) as a profitability measure, which has come from investment analysis and is calculated as a discount rate that brings to zero the net present value of net cash inflows. In the loan market, IIR is the lender's effective interest rate.

Serrano-Cinca and Gutiérrez-Nieto (2016) investigate an indicator, which can be applied as an alternative measure of the probability of default estimation for a decision-making system used in peer-to-peer (P2P) lending. They found that some variables that are significant for the probability of default prediction can be insignificant for the profit prediction, such as loan purpose. Some of the loan purposes have different significance and tendencies (signs) for the probability of default and profitability predictions. For example, credit cards have the probability of default of 9.29% and 6.27% IRR, but car loans have shown slightly lower PD, but significantly lower profitability. Serrano-Cinca and Gutiérrez-Nieto (2016) selected the borrower's rate of interest, borrower's indebtedness, and loan purpose as a factor, which explains the profitability of the loan.

The linear multivariate regression has shown low fitting accuracy with R-squared 0.015 for a model with excluded interest rate and built with all other variables such as loan purpose, risk grade, borrower characteristics, credit history and borrower indebtedness. On the other hand, the model with an only single interest rate as a covariate shows high R-squared around 0.49, and fitting accuracy does not grow after another covariates inclusion.

However, Internal Rate of Return was used by Serrano-Cinca and Gutiérrez-Nieto for the P2P market where individual lenders provide loans to individual borrowers, and this simple measure might not be suitable for more complicated bank lending models. Generally, Serrano-Cinca and Gutiérrez-Nieto (2016) analysed few characteristics and did not provide convincing evidence of the significantly different impact of the same covariates for default and profit models.

2.4 Use of transition probabilities for profit scoring

Transition matrices are widely used in credit risk modelling for rating migration prediction such as in CreditMetrics (Gupton, Finger & Bhatia, 1997) and

CreditPortfolioView by McKinsey (Wilson, 1998). One of the first applications of a Markov Decision Process (MDP) to loan repayment was by Bierman and Hausman (1970). Frydman et al. (1985) investigated the mover-stayer model for credit behaviour. The core element of the MDP use in credit modelling is a state of the account. The state can be defined as a range of the behaviour score of account or the number of periods in arrears (Thomas, 2007).

Let's give a brief description of Markov Chains. Let X_0, X_1, \dots, X_I be random variables which take values in one of the I states.

The process is a Markov chain if

$$\Pr\{X_{n+1} = j | X_0 = k_0, X_1 = k_1, \dots, X_{n-1} = k_{n-1}, X_n = i\} = \Pr\{X_{n+1} = j | X_n = i\} \quad (2.1)$$

for $\forall n$ and i, j where $1 \leq i, j \leq I$.

The right-hand side of equation (2.1) gives a set of transition probabilities and is represented as $p_n(i, j)$. The sum of all transition probabilities from state i to all possible states j is equal to one $\sum_j p_n(i, j) = 1$. The transition probabilities matrix is defined as $(P_n)(i, j) = p_n(i, j)$.

The Markov process is stationary if $p_n(i, j) = p(i, j)$ for all n , i , and j . For the stationary process, the n -stage transition probability is calculated as $P\{X_n = j | X_0 = i\} = P^n(i, j)$ (Thomas, 2007).

Thomas and So (2011) proposed a model for credit limit policies to maximise customer lifetime value. The key part of this investigation is their use of Markov Decision Processes (MDP) to set up optimal credit limit strategy depending on the customer's behaviour, as well as his or her revenue and risk estimations. An MDP is used to estimate a lifetime dynamic parameter. They use the discounted cost optimality equation where λ is a discount factor between time periods. 'This leads to the following optimality equation for $V_t(l, i)$ which is the maximum expected profit over the next t periods that can be obtained from an account which is currently in behaviour state i , and with a credit limit of l ' (Thomas and So, 2011, p.124):

$$V_t(l, i) = \max_{l' \geq l} \left\{ \sum_{i'} p(i'|l, i) [r(i'|l, i) + \lambda V_{t-1}(l', i')] \right\}$$

The formula maximises the reward over the next t periods in the case of the credit limit changes to l' at the end of the current period for an account with behavioural score state i . The $p(i'|l, i)$ is the probability of changes of behavioural score i to i' . In that case, the reward to the lender from the credit card of such a transition is $r(i'|l, i)$ and the reward on the remaining $t-1$ periods if the behavioural score changes to i' is $V_{t-1}(l', i')$. λ is the discount factor to deal with the fact that the reward in the last $t-1$ periods actually occur one period after these used in calculating $V_{t-1}(l', i')$. The optimality principle says that the optimal decision l' , is the one that maximises this sum of the future period gains, and credit limits in these periods are stable or grow up.

Finally, Thomas and So (2011) obtain transition matrices and reward functions, and as a result the set of optimal policies for a particular data sample. The strength of the proposed approach is that credit card customer states are considered as transactor and revolver. One weakness of this approach is the use of a discount factor, which has an expert basis, not statistical. Another weak point is that the optimisation considers increases or a stable limit only. It looks reasonable to decrease the limit for a customer with a high value of the probability of default (hereinafter PD) and high potential Exposure at Default (hereinafter EaD), but with high relative profitability.

Malik and Thomas (2012) applied the cumulative logistic regression model for prediction of transition between the behavioural score ranges and the default state in a transition matrix model. They built a prediction for multi-stage period (to 4 periods) with second order Markov chain. The macroeconomic variables such as changes in Gross Domestic Product (GDP), Consumer price index (CPI), and Unemployment rate, and age of loan are used for explaining the non-stationary chain.

It is necessary to mention some optimisation tasks with the use of Markov Decision Process. Thus Ching et al. (2004) solved the recursive relation problem of customer lifetime value to find a strategy to conduct or not to perform the limit increase. A Markov chain is used for the maximisation of total expected revenue, and obtained from a stochastic dynamic program with t months remaining for a customer in State i :

$$v_i(t) = \max_{j=1 \dots M} \left(c_i^{(j)} - d_j + \alpha \sum_{k=0}^{N-1} p_{ik}^{(j)} v_k(t-1) \right),$$

where N the total number of states (indexed by $i = 0, 1, \dots, N-1$),

t is the number of months in the planning horizon (indexed by $t=1 \dots, T$),

d_j the resources required for carrying out promotion plan j in each period,

$c_i^{(j)}$ is the revenue obtained from a customer in State i with the j^{th} promotion plan in each period;

$p_{ik}^{(j)}$ - the transition probability for the customer to move from State i to State k under the j^{th} promotion plan in each period.

The customers are classified with four states: low volume user (State 1), medium volume user (State 2), high volume user (State 3), and State 0 if a customer was inactive during the period of observation. Markov chains give an advantage of considering the transition between states for the indefinite horizon.

Linear programming techniques are used for optimisation task solving. However, all abovementioned papers use homogeneous Markov chains and this is a significant limitation. Lando and Skodeberg (2002) investigate non-homogeneous Markov chains for credit rating transitions. The MDP homogeneous requirement problem can partially be avoided, for instance, with use of customer behaviour pattern segmentation (Till & Hand, 2003).

Crook and Leow (2014) use intensity transition models, based on survival analysis with time-dependent variables, to predict the delinquency state of credit card loans. They applied a non-homogenous Markov chain with four states: up-to-date, one month in arrears, two months in arrears, and default. The transitions between states depend on the individual application and behavioural characteristics of the debtor.

An intensity model is defined as follows:

$$\alpha_{hji}(\tau) = Y_{hi}(\tau) \alpha_{hj0}(\tau) \exp \left\{ \beta_h^T Z_i(\tau) \right\}$$

where ' $Y_{hi}(\tau)$ is an indicator for whether individual i was in state h at time τ , α_{hj0} is the baseline transition intensity for state h to state j and β_{hj} is a vector of unknown regression coefficients for the m covariates' Crook and Leow (2014, p.687).

They found that different algorithms and cut-off values may give good prediction accuracy at the pool level but deteriorate the prediction accuracy at the account level. However, they used mainly application covariates such as age, employment, income, time at address, and few behavioural characteristics lagged three months: logarithm of the credit limit, the logarithm of the payment amount, the proportion of credit drawn, an indicator for improvement in state from 3 months previous.

Binary logistic regression is a particular case of the multinomial logistic regression model which compares only one dichotomy. Multinomial logistic regression, as well as ordered logistic regression, refers to multiple choice models. The probability to make a choice j for Y_i from the set of possible choices J can be written for the probability as follows (Greene, 2002):

$$Prob(Y_i = j|x_i) = \frac{e^{\beta_j' x_i}}{\sum_{k=1}^J e^{\beta_k' x_i}}$$

where x_i is a vector of characteristics.

One of the applications of multinomial regression for credit card usage states modelling has been proposed by Volker (1982). He defined four types of card usage (hold bankcard, use credit, use regularly, and use moderately) and compared how the same set of predictors (age, professional skills, marital status, region of residence etc.) impacts on the customer probability to obtain one of the mentioned statuses.

The model of credit card usage types prediction is defined as

$$\ln\left(\frac{P_{ij}}{P_{i1}}\right) = \beta_j X_i$$

'where P_{ij} is the probability of individual i selecting alternative j , normalization is to the first alternative, X is a vector of explanatory variables, and B_j is a set of alternative-specific coefficients' (Volker, 1982). So the model is presented as a prediction of the probability of each type of card usage considering the probability of usage for a non-holding (no active) bank card.

We have found that there is a lack of literature about transition probabilities estimated for the credit cards, especially, at the account level. The credit states are often defined as inactive, current, delinquent, and default. Some papers (for example, Lando and Skødeberg, 2002; So and Thomas, 2011; Malik and Thomas, 2012; Leow and Crook, 2014) discuss the transitions between states, which are based on the risk or loss definitions, but mainly not on the income or profit definitions.

2.5 Credit card usage and credit limit utilisation rate modelling

Initial research into customer behaviour consists of papers dedicated to the economic organisation of households (e.g. Awh et al., 1974; Bryant, 1990) and has become the basis for further research on credit products usage and risk modelling. The first set of investigations in the area of credit cards usage paid attention to the consumer credit demand (White, 1976; Dunkelberg & Stafford, 1971; Duca & Whitesell, 1995) and to the probability of credit card use (White, 1975; Dunkelberg & Smiley, 1975; Heck, 1987).

Heck (1987) investigated credit cards use and, in particular, the probability of households using major types of credit cards and reasons for the utilisation behaviours of different types of credit card. Equations were estimated for five products: gasoline, bank, retail store, general purpose and special account credit cards, and checked against a set of hypotheses about the impact of households' characteristics on card utilisation. He found that there were a number of socio-economic characteristics, such as customer income, age, education level, occupation, number of family members, presence of child under six years old, race, village residence and a number of social-psychological characteristics (e.g. customer expectation about future income and the country's financial conditions) that affect the probability of usage. It is presented by the probit equation as follows:

$$P(Y) = \frac{1}{2\pi} \int_{-\infty}^{USEG_i} \exp\left(\frac{-USEG_i^2}{2}\right) dUSEG_i,$$

where $P(Y)$ is the probability of using a gasoline card,

$USEG = \beta_0 + \beta_1(X_1) + \dots + \beta_n(X_n)$ is a linear combination of usage function for the probit equation and X_i is an independent variable or characteristic of a household used in the research model.

The probit equation and the probit estimation produced efficient results to demonstrate the variation of the probability of card usage between different credit card types. Thus, the set of explanatory characteristics varies depending on card type, and the same consumer can demonstrate different behaviour depending on which type of card is used. This investigation concentrated only on the probability of use, but not the depth of usage, use rates, intensity or other qualities. Moreover, there is no mention of the utilisation as a percentage of credit limits. However, notions of the probability of credit card usage, activation, reactivation and other actions have become one of the key issues in current research.

Crook et al., 1992 investigated the predictors of whether a customer will or will not use a credit card and found they are similar to risk level predictors, such as age, applicant's income, age with the bank, residential status, etc. The possibility to discriminate between users and non-users is highly important for the development of the efficient customer usage motivation strategy and charging policy.

Each nominal value was replaced by its weight of evidence measure derived as $X(i)$ for each i nominal value:

$$X_j(i) = \ln \frac{u_i}{v_i} + \ln \frac{V_T}{U_T}$$

where u_i and v_i are the number of users and non-users in value i , and U and V are the total number of users and non-users in the sample. The linear discriminant analysis was used to identify the standardised coefficients of the function, and the Partial-F statistics were used to indicate the significance of the covariates. Also, the role of the regional location (postcode characteristics) was discovered, but the results were not sufficiently robust to include postcodes in the model because of high sampling variance and the high chance of the 'artificially significant role' in comparison with other variables.

There is a lack of fundamental work dedicated to the prediction of a credit card utilisation rate in the academic literature. Kim and DeVaney (2001) provided a comprehensive investigation of the revolver customer's outstanding balance. The important issue is the approach to the revolver customer definition and the testing of hypotheses about customer characteristics and the chance of being a revolver. According to the authors, 'credit card payment behaviour and the outstanding credit card balance are affected by consumption needs, current resources (the budget constraint) and future resources, the interest rate, and the consumer's preferences' (Kim & DeVaney, 2001, p.68).

A credit card is both a payment tool and a convenient source of credit (Garman & Forgue, 1997). Kim & DeVaney (2001, p.67) refer to the definition of revolving credit card users from Bird, Hagstrom & Wild (1997) that the revolver is a credit card holder 'who have an outstanding credit card balance after the last monthly payment'.

Another definition of revolver and transactor customer proposed by Hsieh (2004, p.626): 'Revolver users always carry a credit card balance, rolling over part of the bill to the next month, instead of paying off the balance in full each month' and 'Transactor users pay in full on or before the due date of the interest-free credit period and do not incur any interest payments or finance charges'.

Credit card utilisation was also been used as the proxy for credit money demand. Thus Cohen-Cole (2011), in his investigation of the presence of racial disparities in the supply of credit cards, has used the utilisation of credit as the specification of demand. Dunn and Kerr (2002) investigated the impact of a customer's search cost and high rejection of applications on large balances of consumers. They found that search cost is no longer the main factor in the credit market. They propose a probit model for the prediction of the customer creditworthiness and utility and use it for estimation of the consumers' propensity, which is measured as the probability of applicant rejection.

Splitting by customer usage can also be applied to improve the predictive accuracy of the scoring model. Banasik et al., (2001) assumed that predictive characteristics of customers' low and high credit card usage could differ from those for a default model. The Heckman two-stage procedure was applied to predict desired credit cards usage. The set of predictors that explained whether the desired credit line is more likely to be

constrained or not corresponds with the set of usual risk drivers. For example, it is expected that a customer has a higher level of risk if he or she has a substantial number of children, is in the military or works in the private sector, or rents an apartment. The same factors were used in the proposed model to estimate the probability of credit limit constraint.

This investigation provides an important result for current credit cards usage research. The lender applies risk scoring-based credit limit strategy. The ‘good’ customers have the desired limit and use a credit card as expected to satisfy their needs. The ‘bad’ customers have a limit which is below their desired balance and probably will have higher credit line usage (or utilisation rate). This approach shows that credit cards usage estimation may have an impact on a risk-based decision-making process because it can be more beneficial for a lender to accept more risky customers, but with higher card usage. Moreover, the question remains about the primary nature of the risk that impacts usage or the usage that impacts risk.

Agarwal et al. (2006) give a similar point of view on the dependence between credit risk and credit line utilisation with two hypotheses. The first is that higher credit risk customers will have lower initial credit utilisation. The second is that changes in customer credit risk have an impact on the credit line utilisation. They investigated mortgage loans in this work. The following set of parameters were used for the utilisation rate prediction: FICO score, original loan-to-value (OLTV), debt-to-income, fraud indicator, prior delinquency, prior bankruptcy and prior foreclosure. Moreover, FICO score changes were included in the model as a factor.

The borrower i ’s credit line utilization U_i was predicted as the regression equation estimated with the least squares approach

$$U_i = \beta_0 + \beta_1 OLT\!V_i + \beta_2 r_i + \beta_3 FICO_i + \sum_{t=1}^6 \delta_k State_{ki} + \alpha \lambda_i + \varepsilon_i$$

where $OLT\!V_i$ is the original loan-to-value, r_i is the current mortgage interest rate, $FICO_i$ is the borrower’s credit score at origination, and $State_i$ is dummy variables for the borrower’s location. The inverse Mills ratio λ is inserted to correct the sample biases.

A hazard function was estimated with a multinomial logit framework and each initially observed credit line is non-delinquent. Higher credit risk is associated with higher utilisation and less prepayment. The postcodes are used to identify changes in local house prices, and macroeconomic variables like unemployment are included to connect local economic risk factors and state dummy variables. The empirical results demonstrate that a decline in credit quality (or increase in risk level) is the cause of the higher increase in the utilisation rate and the decrease in the probability of prepayment. However, customers with low credit scores have initially a low utilisation rate in comparison with high credit score customers because customers with high scores usually rationally manage their debts.

Credit card limit utilisation modelling can be performed as a direct prediction for the outstanding balance amount. For example, Kim and DeVaney (2001) used the Heckman selection model in a two-step procedure: estimate the likelihood of having an outstanding balance and predict the outstanding balance amount. They found that ‘Education, income, real assets, credit card interest rate, number of credit cards, the credit limit, a positive attitude to credit, and behind schedule payments were positively related to the outstanding credit card balance’. They applied this investigation for revolver customers only, but not for so-called convenient users.

The splitting of credit cards holders up into the revolvers and transactors has become widespread in lenders’ business strategies and risk management tasks, and it is reflected in the most up-to-date literature. Thus, the segregation of consumers who use their credit card as a convenient payment tool and customers who use their credit card like a credit instrument has become one of the analytical problems in credit cards modelling (Cheu & Loke, 2010). Investigation of socio-demographic factors and use of the client usage type behaviour models in application processes, like an industrial standard application scoring use, can help the lender to diversify their risks and increase profitability because of more accurate customer management and credit products targeting.

Customers, who always pay back the full amount of the balance each month, or transactors, are less risky for the bank, but this also means they are less profitable. In the traditional scoring models, the aim is to estimate the probability of default for the

separation of good and bad cases. However, only revolvers can generate a probability of default and so give the bad cases for development samples. However, the transactors *a priori* generate only good cases. So risk estimation when including transactors can be biased.

So et al. (2014) proposed to create a risk scoring model based on revolving cases to get more accurate estimations. The model was built in four stages: standard scorecard, transactor/revolver scorecard, a Good/Bad scorecard restricted to revolvers and the risk assessment system based on the transactor/revolver scorecard. The model estimated the chance that if the customer is a transactor that is ‘who pays off the balance for at least 12 months before the sampling time’ with the use of logistic regression. The probability of the customer to be good (non-defaulted) was calculated as the sum of two components: the probability of being a transactor who has *a priori* non-defaulted, and the product of the probability to be a revolver and to be good:

$$P(G|x) = P(T|x) + P(R|x)P(G|x, R)$$

where

$P(T|x)$ is the probability of a customer being a transactor for (s)he application characteristics \mathbf{x}

$P(R|x)$ is the probability of a customer being a revolver for (s)he application characteristics \mathbf{x}

$P(G|x, R)$ is the probability of a customer being Good conditional on having application characteristics \mathbf{x} and on the customer having a revolver state.

Also it can be presented in the format of logistic function probability:

$$P(G|x) = \frac{1}{1 + e^{-s_t(\mathbf{x})}} + \frac{1}{1 + e^{s_t(\mathbf{x})}} + \frac{1}{1 + e^{-s_R(\mathbf{x})}}$$

where

$-s_t(\mathbf{x})$ is a log odds score which determines the chance that an applicant being a transactor in terms of her/his application characteristics \mathbf{x} ;

$s_t(\mathbf{x})$ is the chance that an applicant is a non-transactor, that is a revolver;

$-s_R(\mathbf{x})$ is a log odds score which determines the chance that a revolver customer being Good.

The expected profit after N periods is defined as a function of interest rate r and the stationary hazard rate p of being good:

$$e(r, p) = P \left((m-1) + \frac{(1+r)^{N-1} p^N}{(1+r_F)^N} + \frac{(1-l_D)(1+r)^{N-1}(1-p^N)}{(1+r_F)^N} \right)$$

After some transformations with transactor/revolver probability score, the expected profit per customer is calculated as the integral of the product of three components: $e(r, p)$, the expected profit depending on interest rate r and the hazard rate to be good, p ; $q(r, p)$, the probability that the customer will take a card; and $f(p, t)$, the density function of p and transactors score t :

$$E(r) = \int_{-\infty}^{\infty} \int_0^1 e(r, p) q(r, p) f(p, t) dp dt$$

The innovation proposed by the So et al. (2014) credit card profitability model is to consider the length of time the cost of a purchase stays on balance.

A credit card's limit may be highly volatile over time, higher than the household income volatility, and this motivates householders to hold high-interest rate debts due to the chance of losing their source of fund (Fulford, 2015). Fulford (2015) sets credit limits in proportion to income and describes the model as an MDP with a “permanent” (random walk) component, which determines the actual limit when an individual can borrow, and with a transitory component, which determines whether an individual can borrow at all as follows:

$$B_{it} = D_{it} e^{X_{it}\beta_{it}} M_{it} W_{it}$$

where

D_{it} determines whether person i can borrow at time t and is equal to 0 or to 1;

X_{it} are covariates such as age, date, geographical location, and credit risk; and

M_{it} and W_{it} are innovations specific to i , defined as $m_{it} = \ln M_{it}$ and $w_{it} = \ln W_{it}$.

The permanent component is defined as $m_{it} = m_{it-1} + v_{it}$, and v_{it} and w_{it} are independent and distributed across individuals and over time. Thus Fulford (2015) uses panel data for the credit limit estimation and considers cross-sectional and time-varying components. The usage of panel data and random effects is discussed in Section 2.7.

The credit limit is estimated conditional on being able to borrow as a linear regression with logarithm transformation

$$\ln \hat{B}_{it} = \rho f_{it} + X_{it}\beta + b_{it}$$

where f_{it} is the ratio of a number of the reported accounts to the number of opened accounts, and X_{it} is a vector of covariates. After this, the model estimates the transitory and permanent variances faced by individuals from the residual b_{it} .

Then Fulford (2015) solves the problem of finding the credit limit for maximizing the difference between consumption paid for with cash and consumption paid for with debt. This problem solves for fixed and stochastic credit limit using a Lagrange approach with a set of constraints. The basic optimization model is following:

$$\max_{\{c_t^w, c_t^b\}_{t=0}^{\infty}} E_0 \left[\sum_{t=0}^{\infty} \beta^t u(c_t^w + c_t^b) \right]$$

where $u(c_t^w + c_t^b)$ is utility function of cash and debt consumption respectively for period t .

This considers financial uncertainty of households, their income expectations and explains the way that credit limits affect the household savings, borrowings, and consumption. Fulford (2015) uses a narrow set of covariates for the credit limit prediction such as age, size of household, income, home ownership. However, he also uses questions to households about the distribution of money spending and borrowing as covariates which is rare in research because it requires special preliminary data gathering.

Fulford and Schuh (2017) continue to investigate consumption over the customer life-cycle and pay attention to credit card utilisation as a ratio of the debt and the credit card limit. They predict the utilisation rate with an autoregressive model AR(1) and

try to maximise a customer's utility for the remaining life from age t given current resources and expected future income.

$$\max_{\{X_s, \pi_s\}_{s=t}^T} \left\{ E \left[\sum_{s=t}^T \beta^{s-t} u(C_s) + \beta^{T+1} S(A_T) \right] \right\}$$

where $u(C_s)$ is period utility from consumption C_s ; A_t the sum of assets left at the end of the previous period. The decision at t depends on customer expectations and utility to be at ages $s \geq t$.

Fulford and Schuh (2017) show that credit limit changes can be very significant over a life-cycle, but credit limit utilisation is stable over a business-cycle and for different ages with slight utilisation rate declines over the life-cycle. The US customers long period data has shown that the credit limit utilisation rate is much less volatile than credit or debt. The linear models for credit limit utilisation rate prediction were based on credit bureau (Equifax) data and written as follows

$$v_{it} = \theta_t + \theta_a + \alpha_i + \beta v_{it-1} + \epsilon_{it}$$

where $v_{it} = D_{it}/B_{it}$ is the credit utilisation given the credit limit B_{it} and the current debt D_{it} , conditional on the credit limit $B_{it} > 0$, and age θ_a and quarter θ_t effects that allow utilization to vary by age and time.

Models have demonstrated moderate predictive accuracy with R-squared between 0.43 and 0.75, and included individual fixed effects, quarter effects, and age effects.

Some models (as, for example, Kim and DeVaney, 2001) included the credit limit as a variable, but the forecasted value is an absolute amount. In our current investigation, we concentrate on the utilisation rate, that is, the percentage of the allowed credit level that is used by the customer. In our opinion, the utilisation rate approach can give an adequate customer behaviour estimation in the sense of consumption habits and customer demand for money.

We have found many papers, which investigate the probability of usage and type of usage of credit cards. Also there is a lot of papers, which discuss the prediction of the outstanding balance for current and defaulted states. However, we have found that there is a lack of papers dedicated to the prediction of the credit limit utilisation rate

as a ratio of the outstanding balance to the credit limit. In our opinion, the use of the utilisation rate can help to get more accurate estimations of the outstanding balance for the exposure at default and income interest prediction because the average outstanding balance can change in proportion to the credit limit and cardholders may spend not exact amount of money, but a part of the credit limit.

2.6 Credit card total and transactional income modelling

There are few papers which are dedicated to credit cards transactional income prediction. Most related papers discuss the probability of credit card usage (for example, Crook et al., 1992a) or the probability of a card spending transaction (for example, Andreeva et al., 2005).

So et al. (2014) mentioned the two main revenue sources from a credit card: the interest charged on the card balance, and the merchant service charge, or interchange fees. However, they predict the total profit from the credit card with Good-Bad scorecards considering the cardholder's transactor or revolver state. They found that the usage of transactor and revolver states for the prediction of total profitability gives more exact results than without this type of models' segmentation.

Lucas (2001) defined the following sources of credit card income:

- i. Interest on accounts that revolve or do not pay the balance in full
- ii. An 'interchange' fee (or merchant fee)
- iii. A period fee, e.g. annual fee, monthly fee, processing fee
- iv. Additional service payments (e.g. insurance payments).

Some literature discusses the socially optimal and profit-maximizing interchange fee (Chakravorti, 2003). The interchange fees may vary for different card issuers and merchants. So, whether or not the customer decides to buy some goods from a specific merchant may depend on the volume of interchange fees. If some merchants set up different cash and card prices this may impact on the customer's behaviour.

The problem of a consumer's choice between Point-of-sale (POS) and Automated teller machine (ATM) transactions depends on a set of parameters such as the cost of using cash, the POS coverage and can be explained, for example, by a Nash game (Markose and Loke, 2003). Humphrey, Kim and Vale (2001) estimate the consumer's

demand for payment choices: checks, ATMs and POS. They use an indirect utility function to model separability between demand on POS and ATM payment instruments.

The ATM fees may be subdivided into various types depending on category, frequency, source, and generating subject (Hayashi, Sullivan, and Weiner, 2003). Credit Card Interchange Fee Rates in European countries are around 1-1.5%, in US – around 2% (Hayashi, 2010).

Generally, non-interest income from fee-based activities can be more volatile than interest income because of credit risk and interest rate fluctuations (DeYoung and Roland, 2001). Total non-interest income and financial performance from fee-based activities can be estimated with regression models at the bank level. DeYoung and Rice (2004) build linear regression models to investigate the impact of bank characteristics, market conditions, and technological developments covariates on the non-interest income indicators such as the ratio of noninterest income to assets and the six-year average of return-on-equity. The covariates used were bank level indicators of bank activities, for example, bank financial performance, core deposits-to-assets, loans-to-assets, natural logarithm of bank assets to bank accounts; number of automated teller machines, number of cashless transactions per capita, the dollar amount of mutual funds, and dollar amount of mortgage-backed-securities per capita; and also state a dummy and time dummy variables to control unspecified cross-sectional and intertemporal sources of variation in non-interest income. DeYoung and Rice (2004) found that non-interest income had become more important than it was earlier, and banks should concentrate on fees-based services, which bring higher profit and improve the risk-return tradeoff at the beginning of the period of investigation, but lead to a worsening of the risk-return tradeoff in the recent years.

There are some customer behaviour modelling approaches to modelling non-interest income that are alternatives to the traditional econometric techniques. For example, Hsieh and Chu (2009) apply a self-organising map (SOM) neural network and a decision tree for the task of credit card customer behaviour analysis. They use these techniques to divide customers into homogeneous groups and to identify whether a customer is a revolver user, a transactor user, or a convenience user. The repayment

behaviour and credit card usage is used for customer profiling. Hsieh and Chu (2009) applied CHAID, classification and regression trees (CART) for customer value discrete-valued target functions approximation. The result of Hsieh and Chu's (2009) analysis was the definition of a customer value for different marketing strategies. The customers were divided into three profitable groups such as convenient users, revolvers, and transactors according to their behaviour and characteristics. These cluster profiles were used for the prediction of customer's revenue and personal bankruptcy. Similar research problems such as customer segmentation by profitable groups according to their behaviour and characteristics, and their solution with the use of data mining methods, can be also found in other works, such as Farajian and Mohammadi (2010), Au and Chan (2003).

There are few papers on the prediction of transactional, or non-interest, income and the aggregated total amount of the credit card income, or profit, which is generated from the number of different sources.

2.7 Gaps in the literature

The published papers are mainly dedicated to the probabilities of card usage and repurchase, debt amount prediction, the transition between risk states, and optimisation tasks based on the risk-revenue approach. However, there is a lack of papers for the prediction of the total income amount from credit card activities. The majority of existing models investigate one part of the profit, generated by a credit card, or give a general prediction of the total income.

An account during the life cycle can be in various states, such as a revolver and transactor as well as inactive or in the default state. The definitions for account segments vary between different sources. For transactor and revolver states some researcher consider customer purchases and payments, and others – a term of positive or zero outstanding balance. Despite the potential ability to give a strict and specific ‘revolver-transactor’ definition, it remains uncertain about the revenue from credit card usage.

In some papers (for example, Greene, 1992; Crook et. al., 1992a; Hand, 2001; Banasik et.al., 2001), the predictors are constant over time and do not depend on time-varying variables such as macroeconomic factors, internal bank changes, and customer

behavioural characteristics. However, the set of characteristics with time series as panel data can have a dramatic impact on regression results. So there is a lack of panel data methods and their tests on empirical data in the area of income modelling.

Existing models of the dynamic behaviour of the consumer based on the Markov Decision Process and Markov Chains have a significant limitation because they have the time homogeneity requirement. This assumption strictly reduces the predictive power and stability of MDP models, and the previous history of the behaviour of each account is not considered. However, empirical research demonstrates that credit card usage has a high-level dependence on the previous history of the customer's behaviour pattern.

In some literature the models, which are built for bad debt/loan default/credit risk prediction, are used for profit or income prediction and profit or income-based decision making. However, risk drivers can be different from the variables which explain the loan profitability (Serrano-Cinca & Gutiérrez-Nieto, 2016). Thus there is a need to develop income prediction models and profit scoring models, which differ from risk scoring models and are intended to fill this gap between explanatory risk variables and income as a dependent target to be predicted.

So et al. (2014) bring out a set of problems with the definition of transactors. The core concerns for a cardholder to be assigned as a transactor are: i) the length of appropriate period, or the number of months, of paying off the full balance, ii) consideration of missing payments, notably, for short-term delinquencies such as a payment in several days after the due date or errors in the repayment amount. The definition of the fixed period for the transactor state requires analysis and explanation why some period is selected and imposes restrictions on other states and target periods used in models with this definition of the transactor state. Thus the fixed period for the transactor state makes the models sensitive to the individual behaviour of a customer and create a problem as to define other states. For example, if the transactor definition is 'six months with full paying off balance' the question is what state should be assigned to a customer who has five months of full repayment history and has not used a card for one of six months. According to a strict six consecutive full repayment months definition such a customer is not a transactor, as well as (s)he is not a revolver or an

inactive customer for six months period. One of the solutions for this problem can be to avoid the strict definition of state for some period but the use of monthly defined states. Thus, a customer may have an individual state for each month.

Many papers investigate the probability of transition to delinquent and default states, or between risk grades (levels), for example, Thomas and So (2011), Crook and Leow (2014), but there is a lack of papers which also investigate transition between different types of states, based on the source of income definition.

In the literature related to credit risk and profit modelling researchers mainly concentrate on: i) how a single or several risk or profit drivers explain the target variable, ii) which method such as regression analysis or machine learning technique gives higher predictive accuracy. However, there is a lack of papers which test how a set of many various time-independent covariates (application variables) and time-dependent covariates (behavioural variables) impacts on the target variables such as utilisation rate, transition probability, usage probability, and income amount. Moreover, we have not found papers, which cover the full aspects and drivers of the profit or income modelling from credit card usage in a single model.

There also is a lack of empirical investigations into profit scoring or income scoring for the credit cards. Thus, the questions about model validation parameters benchmarks and sets of the significant explanatory characteristics are live issues. There is also a lack of paper that estimate credit cards income from a range of the different types of income sources.

2.8 Conclusion

This Chapter discusses the papers, which have become a theoretical basis for this thesis and revealed gaps in the literature. *First*, we made a brief overview of traditional credit scoring methods used for the estimation of the Probability of Default. Also, we considered some papers in the estimation of Exposure at Default and Loss Given Default parameters. *Second*, we discussed the methods for Profit scoring such as profit modelling, efficient cut-offs, optimal credit limit policy, game theory, and use of survival methods. The important section is an overview of the measures of profit because the findings from discussed papers can be used as targets for our models for income prediction. *Third*, we discussed the use of transition probabilities for profit

scoring, but have found that the most states, which are used for the transition probabilities, are related to risk scoring and do not consider income-based states. *Fourth*, we reviewed papers on credit card usage and credit limit utilisation rate modelling, and have found a lack of papers, which give an empirical investigation of the credit limit utilisation rate as a ratio of the outstanding balance to the credit limit. *Fifth*, we reviewed papers on credit card total and transactional income modelling and found few papers on the prediction of transactional, or non-interest, income and the aggregated total amount of the credit card income, or profit from various sources. *Finally*, we discussed the gaps in the literature related to profit scoring and income prediction. In this research, we will try to fill some of the theoretical and practical gaps in the literature.

3 Chapter Three. Data description and variables

3.1 Introduction

This Chapter introduces the data sample and variables for this research. Section 3.2 describes data sources, the data set as panel data, discusses vintages, periods of data cohorts, observation and performance windows. Section 3.3 gives a list of raw and derivative variables, computation methods and descriptive statistics of independent variables. Section 3.4 gives a list of dependent variables and their descriptive statistics. Section 3.5 describes the data sampling and how the initial data set has been split into development and validation parts for modelling and testing purposes. Section 3.6 concludes the data gathering and sampling issues.

3.2 Panel Data overview and observation and performance windows

The data consists of account level observation from a bank in East Europe. The data set consists of four types of variables. First, monthly indicators of credit card behaviour at the account level such as the outstanding balance, days past due, arrears amount – from an accounting banking data warehouse. Second, credit card transactions, aggregated on a monthly basis such as by taking the sum, minimum, maximum and average of spending amounts (or debit transactions) and loan payment (or credit transaction), and transactional profit – from the data providers transactional banking warehouse. Third, cardholders' applications, which contains socio-demographic and financial characteristics – from the data providers banking application processing system, and iv) macroeconomic indicators – from the National statistical bureau. The empirical investigation is based on the anonymized data sample of credit card accounts, which contains application and behavioural data.

Application data contains a consumer's socio-demographic characteristics such as age, gender, education, marital status, residence, region of residence, number of family members, and economic factors such as monthly income, sources of income, spouse income at the time of application. Behavioural data is originated from the data warehouse and contains dynamic information about consumer transactions, balances and delinquencies, such as outstanding balance at the end of a month and the average monthly outstanding balance, number of debit and credit transactions, average, maximum and minimum transaction amount per month, a days past due counter,

arrears amount at the end of month etc. National statistical bureau data contains macroeconomic indicators such as Gross Domestic Product (GDP), Consumer Price Index (CPI), unemployment rate, foreign currency exchange rates, and average salary. These macroeconomic variables have been selected because we expect the highest impact of these five variables on customer spending and payment behaviour among the majority of macroeconomic indexes, and because these variables have easier explanation and direct relation to consumer credit consumption than others. All required behavioural and macroeconomic data is collected monthly or extrapolated on a monthly basis. The model also uses covariates, which are originated from the application and behavioural data and describe the financial product characteristics such as the interest rate and loan amount, or credit limit. These variables may be either time-varying or constant for an account over time. For this research, the credit limit value is not a constant and is time-varying, and the interest rate is a constant.

The panel data sample has two and a half years of payment history – time series of 30 cohorts. The data sample contains 21,750 accounts and a total of 439,450 observations for all periods. However, data is not available for all accounts for the full period because the loan activation period was started after the first behavioural date as of July 2010.

The period of active lending is January 2008 - October 2008 and January 2010 – December 2012. Because of the financial crisis active lending was stopped by the Bank in the period from November 2008 to December 2009. The credit cards, which were granted in the crisis period, were mostly direct debit credit cards and issued for low risk existing clients. Thus, the period of possible modelling with enough behavioural data and various customers' behaviour starts in July 2010. However, to avoid losing valuable information from the pre-crisis lending, the applications for 2008 are also used in the modelling. The available period for the behaviour investigation is from July 2010 to Dec 2012. However, the lending period is in this behavioural time range, and new accounts come into the credit portfolio during the period of the investigation. This fact has both negative and positive sides. The negative issue is that the client population structure and volume of the portfolio varies over time. On the other hand, the positive issue is that all vintages are available for modelling and models will not be biased because of only old vintage data.

For modelling purposes, the data set has been cleaned and transformed. The following accounts have been removed from the data sample:

- ✓ Accounts without activation for the period of investigation (never active)
- ✓ Accounts with activation for the period of the investigation but with the insufficient amounts on balance for the whole period (less than the equivalent of 5 GBP or 70 local currency units).
- ✓ Accounts without application information because it is not possible to match application data
- ✓ Accounts with missed or incorrect application information – the application data is blank or contain unreliable values
- ✓ Accounts with unreliable behavioural data such as illogical dependencies between fields, the inconsistency of the data in time, missing months.

From the total set of 21,750 accounts, we have selected the accounts, which i) have been activated before or during the observation period after July 2010, but not later than 12 months before the last available date in the data sample (Dec 2012), ii) have at least 12 months of behavioural history. For an active account, we understand an account with a positive outstanding balance due to the debit transaction (spending or purchases). An inactive account is an account with the outstanding balance equal to zero. The activated account is an account with at least one debit (purchase) transaction. An activated account can be active or inactive at a time point depending on the outstanding balance value. Because of credit card specific characteristics, which consider the possibility of the occurrence of debt and its payment within a month, the outstanding balance is calculated as a monthly average. Thus, an account, which was activated in July 2010, but paid back in August 2010 and had not been closed during at least 12 months, is activated, inactive for the period after July 2010, and is presented in the data sample. An account, which was activated in July 2010, but paid back in August 2010 and had been closed for 12 months, is activated and closed, so, it is excluded from the data sample. An account which has been activated before the July 2010 but has zero outstanding balance after July 2010 and did not have any transactions, is excluded from the data sample.

The total number of observations in the data sample increases from 508 in July 2010 to a maximum of 21,389 in December 2011 and then slight decreases to 19,313 in the last observation month in December 2012 (see Table 3.1). In case of exclusion of inactivated accounts and accounts, which have less than 12-month history, the number of observations for each month has significantly decreased. The first month, for which 12 months behavioural history is available from the earliest month possible which of July 2010, is June 2011. However, at this point, we have 2,544 accounts, which satisfy the requirements of activity and 12 months history, and there is a maximum number of such accounts equal to 14,029 in Jan 2011.

Thus, we can use June 2011 as the first point for modelling and totally have 19 cohorts (from June 2011 to Dec 2012), which have at least 12-month history. We can split the observation period and performance period into the equal parts of six months each. So, the 6-month observation window can be used for computing the predictors, and a 6-months performance window can be used for the target variable. Being consistent across all cases we will use a 6-month observation window for behavioural characteristics, even if data over a longer period is available for some cases.

Thus, December 2010 is the first observation point because of a requirement for a 6-month history for behavioural characteristics. June 2011 is the end of the first 6-months performance window (performance point). For one-month performance window we also use the first performance point as June 2011 and the observation point May 2011 to be consistent with the condition that all accounts must be active at least 12 months.

Table 3.1 Total number of observations (accounts) for each month including inactivated and having less than 12-month history

Month	Total number of observations in a sample	Total number of behaviour for 12m+ observations
201007	508	
201008	511	
201009	641	
201010	1,315	
201011	2,697	
201012	4,083	
201101	5,648	
201102	8,480	
201103	10,529	
201104	12,511	
201105	14,523	
201106	15,101	2,544
201107	17,009	4,042
201108	18,161	6,601
201109	18,795	8,423
201110	19,570	10,141
201111	20,469	11,885
201112	21,389	12,382
201201	21,366	14,029
201202	21,331	14,025
201203	21,262	14,010
201204	21,175	14,006
201205	21,032	13,964
201206	20,874	13,897
201207	20,654	13,717
201208	20,431	13,504
201209	20,177	13,260
201210	19,857	12,942
201211	19,536	12,614
201212	19,313	12,398

The data sample is panel data. It contains characteristics for the accounts on a monthly basis for a period of two and half years. Thus, the data sample merges cross-sectional data and time-series data for each observation.

The data sample has 30 periods (July 2010 – Dec 2012) or cohorts of behavioural characteristics such as outstanding balance, spending and payment transactions, arrears amount, and days past due counter. We have data for new loans granted in the period from July 2010 to Dec 2011. For the training sample, we use the period from

July 2010 as the first application date to the June 2012 as the last available application date for 6-month performance window. The following six months from July to Dec 2012 is the performance period. Accounts started after July 2012 are excluded from the data sample. Each account has at least 12 months from the date of activation till December 2012.

Figure 3.1 shows all possible options of usage for each vintage and how many times an account can take part in the data sample. For instance, the account started in July 2010 has 30 months in July 2010 – December 2012. In case the observation window is six months and performance window is six months this account can be used 30-12=18 times in the data sample to present the account dynamic for all possible periods.

An example of observation periods for the calculation of predictors and the performance periods for the outcome is show in Figure 3.2. It contains as an example an account started in July of 2010. The first observation point is the end of December 2010 (light grey bar) for loans activated in July of 2010, and the performance window (6-12 months of history – dark grey bar) for this vintage is January-June 2011.

Figure 3.1 An example of the 6-months observation and 6-months performance window for 12-months loan issuance period

T	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24
Month(M)	7	8	9	10	11	12	1	2	3	4	5	6	7	8	9	10	11	12	1	2	3	4	5	6
Year(Y)	2010						2011						2012											
Vintage	Y	M	Observations period						Performance															
1	2010	7	1	2	3	4	5	6	7	8	9	10	11	12										
2		8	1	2	3	4	5	6	7	8	9	10	11	12										
3		9	1	2	3	4	5	6	7	8	9	10	11	12										
4		10	1	2	3	4	5	6	7	8	9	10	11	12										
5		11	1	2	3	4	5	6	7	8	9	10	11	12										
6		12	1	2	3	4	5	6	7	8	9	10	11	12										
7	2011	1		1	2	3	4	5	6	7	8	9	10	11	12									
8		2		1	2	3	4	5	6	7	8	9	10	11	12									
9		3		1	2	3	4	5	6	7	8	9	10	11	12									
10		4		1	2	3	4	5	6	7	8	9	10	11	12									
11		5		1	2	3	4	5	6	7	8	9	10	11	12									
12	2011	6		1	2	3	4	5	6	7	8	9	10	11	12									
13		7			1	2	3	4	5	6	7	8	9	10	11	12								
14		8			1	2	3	4	5	6	7	8	9	10	11	12								
15		9			1	2	3	4	5	6	7	8	9	10	11	12								
16		10			1	2	3	4	5	6	7	8	9	10	11	12								
17		11			1	2	3	4	5	6	7	8	9	10	11	12								
18		12			1	2	3	4	5	6	7	8	9	10	11	12								

Consider new loans granted in the period from July 2010 to Dec 2011. For the development sample, we use the period from July 2010 as the first application date to the June 2012 as the last available date, that is Dec 2012 minus six months (due to

applied performance period). Figure 3.2 shows all possible options of usage for each vintage and how many times an account can take part in the data sample.

Loans, activated in July 2010 has a MOB equal to 1 in July 2010, MOB equal 2 in August 2010, etc. The loans activated from July 2010 till December 2011 have enough periods to be in both development and validation samples and to be used for the calculation of behavioural characteristics for the observation window and 6-months in performance window.

For accounts opened in month t we have 6 months of behavioural variable values followed by a further 6 months of performance date (see observation point in Figure 3.2). A further 6 months of data enable validation. This is possible for all cases that opened till observation point 18. Six observation points from 14 to 19 are selected for out-of-time validation.

For whole period, presented in the data sample, the account activated in July 2010 can give 18 observation points to the Dec 2011 with the later performance point June 2012 to keep the 6 months performance period for development sample. Six periods (cohorts) of observation point after Dec 2011 and performance point after June 2012 can be used both for out-of-time validation and for development sample in case we avoid out-of-time validation (see section 3.5).

Figure 3.2 An example of usage of cohorts of the 6-months observation and 6-months performance period for one loan issued (activated) in July 2010

t	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	
Month	7	8	9	10	11	12	1	2	3	4	5	6	7	8	9	10	11	12	1	2	3	4	5	6	7	8	9	10	11	12	
Year	2010						2011						2012																		
Obs.point	Observations period	Performance																													
1	1 2 3 4 5 6	7 8 9	10 11 12	13 14 15	16 17 18	19 20 21	22 23 24	25 26 27	28 29 30																						
2	1 2 3 4 5 6	7 8 9	10 11 12	13 14 15	16 17 18	19 20 21	22 23 24	25 26 27	28 29 30																						
3	1 2 3 4 5 6	7 8 9	10 11 12	13 14 15	16 17 18	19 20 21	22 23 24	25 26 27	28 29 30																						
4	1 2 3 4 5 6	7 8 9	10 11 12	13 14 15	16 17 18	19 20 21	22 23 24	25 26 27	28 29 30																						
5	1 2 3 4 5 6	7 8 9	10 11 12	13 14 15	16 17 18	19 20 21	22 23 24	25 26 27	28 29 30																						
6	1 2 3 4 5 6	7 8 9	10 11 12	13 14 15	16 17 18	19 20 21	22 23 24	25 26 27	28 29 30																						
7	1 2 3 4 5 6	7 8 9	10 11 12	13 14 15	16 17 18	19 20 21	22 23 24	25 26 27	28 29 30																						
8	1 2 3 4 5 6	7 8 9	10 11 12	13 14 15	16 17 18	19 20 21	22 23 24	25 26 27	28 29 30																						
9	1 2 3 4 5 6	7 8 9	10 11 12	13 14 15	16 17 18	19 20 21	22 23 24	25 26 27	28 29 30																						
10	1 2 3 4 5 6	7 8 9	10 11 12	13 14 15	16 17 18	19 20 21	22 23 24	25 26 27	28 29 30																						
11	1 2 3 4 5 6	7 8 9	10 11 12	13 14 15	16 17 18	19 20 21	22 23 24	25 26 27	28 29 30																						
12	1 2 3 4 5 6	7 8 9	10 11 12	13 14 15	16 17 18	19 20 21	22 23 24	25 26 27	28 29 30																						
13	1 2 3 4 5 6	7 8 9	10 11 12	13 14 15	16 17 18	19 20 21	22 23 24	25 26 27	28 29 30																						
14	1 2 3 4 5 6	7 8 9	10 11 12	13 14 15	16 17 18	19 20 21	22 23 24	25 26 27	28 29 30																						
15	1 2 3 4 5 6	7 8 9	10 11 12	13 14 15	16 17 18	19 20 21	22 23 24	25 26 27	28 29 30																						
16	1 2 3 4 5 6	7 8 9	10 11 12	13 14 15	16 17 18	19 20 21	22 23 24	25 26 27	28 29 30																						
17	1 2 3 4 5 6	7 8 9	10 11 12	13 14 15	16 17 18	19 20 21	22 23 24	25 26 27	28 29 30																						
18	1 2 3 4 5 6	7 8 9	10 11 12	13 14 15	16 17 18	19 20 21	22 23 24	25 26 27	28 29 30																						
19	1 2 3 4 5 6	7 8 9	10 11 12	13 14 15	16 17 18	19 20 21	22 23 24	25 26 27	28 29 30																						

Observation for testing

Because we use a 6-month prediction horizon, the training period contains dates from July 2010 to June 2012 and the performance period from July 2010 to Dec 2012. All characteristics are recorded monthly. We use cross-sectional and time analysis. For cross-sectional regression analysis, we use pooled data. We introduce the assumption that each month gives an independent observation. So if one account has T months of history, this means that we have T observations. Depending on the model and type of analysis a different number of cases can be created. For a 1-month prediction with current month predictors only we have T-1 cases. For 1-month prediction with a 6-month observation window, we have T-6 cases. For 6-month prediction with a 6-month observation window, we have T-11 cases. One account generates the number of cases equal to a number of months minus required periods for behaviour characteristics calculation and performance window. We use accounts with a minimum history period of 12 months: 6 months for prediction and 6 months for behavioural characteristics calculations.

Two approaches are used in the investigation of the panel data in this paper: pooled and random effects models. In the case of pooled data, we exclude the time component, and each time observation for the same account is included in the data sample as a new

independent observation. Also, we assume that the correlation between any cases is close to zero and is not taken into account. In the case of the random effect approach, we consider each account in time, and all related observations (values of the characteristic for period) are considered as observations for a single case (or account). All chapters, except Chapter 7, use the pooled data, so it is rather observation level model because each observation, which is obtained from an account, is considered as a separate account at each observation point. The number of observation points depends on the periods, which included in the data sample for development and validation.

We have a stable population of the MOB 18, and after this term, the number of accounts with longer history decreases (see Figure 3.3). It happens because of the growth in the new loans at the period which is presented in the data sample. Only a small number of accounts (less than 100) have MOB higher than 27.

Figure 3.3 Number of observations (cases) by Month on Book

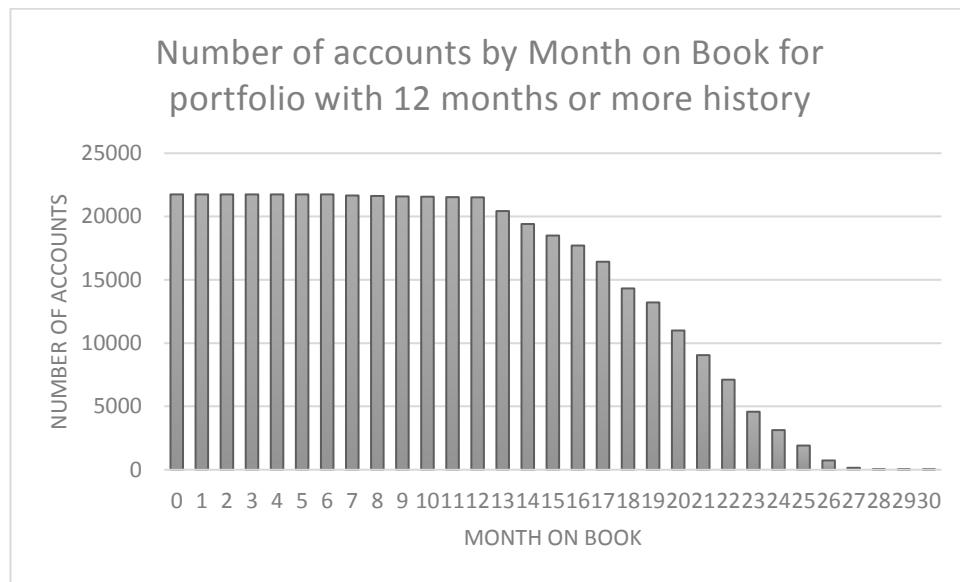


Table 3.2 represents a graphical information about the distribution of accounts by Month on Book from the Figure 3.3. The total number of accounts is 21748 and it decreases to 11 at 30 Month on Book.

Table 3.2 Number of accounts by Month on Book for a portfolio with 12 months or more

behavioural history

Month on Book	Number of accounts	Month on Book	Number of accounts
0	21748	16	17708
1	21748	17	16430
2	21748	18	14321
3	21748	19	13214
4	21748	20	10997
5	21747	21	9051
6	21747	22	7112
7	21660	23	4581
8	21625	24	3129
9	21587	25	1909
10	21565	26	732
11	21533	27	152
12	21515	28	29
13	20432	29	18
14	19410	30	11
15	18498		

We have selected a 6-months observation window and 6-months maximum performance window. However, depending on the objectives of the analysis the performance windows can be reduced to three, two, or one month(s).

3.3 Independent Variables

Behavioural characteristics reflect the various aspects of customer and account dynamics such as the history of spending, payments, balance states, arrears etc. Instead of raw data different derivative characteristics can be created and used as predictors. We classify them as structure ratios, dynamics ratios, counters, maximum, minimum and average values.

We have developed behavioural characteristics, based on the monthly raw data such as the maximum, minimum and average outstanding balance, credit and debit turnovers, number of debit and credit transactions, and days past the due counter. Behavioural characteristics describe the changes of factors – dynamics (for example, the ratio of the outstanding balance in June to the outstanding balance in May), and relations between factors (for example, the ratio of the average debit transaction to the outstanding balance for the last six month). We give a list of behavioural

characteristics for selected predictors, which have been included in models after testing their significance in models.

Table 3.3 contains the name and description of the characteristics of the data sample, used in this research. Column ‘Formula’ contains the equation for the calculation of the behavioural variables.

Table 3.3 List of the original raw data, behavioural, application and macroeconomic characteristics

Characteristic Name	Description	Formula
<i>Raw data</i>		
Month	The month of the data	
Balance_EOP (OB_eop)	The outstanding balance at the end of period (EOP)	
Avg_Balance (Avg_OB)	Average Outstanding balance over a month	
Min_Balance (Min_OB)	Minimum Outstanding balance over a month	
Max_Balance (Max_OB)	Maximum Outstanding balance over a month	
Sum_crd	Credit transactions (payments) amount over a month	
Sum_deb	Debit transactions (purchases) amount over a month	
Num_cred	Number of Credit transactions (payments) over a month	
Num_deb	Number of Debit transactions (purchases) over a month	
Avg_crd	Average credit transaction amount over a month	
Avg_deb	Average debit transaction amount over a month	
Min_crd	Minimum credit transaction amount over a month	
Min_deb	Minimum debit transaction amount over a month	
Max_crd	Maximum credit transaction amount over a month	
Max_deb	Maximum debit transaction amount over a month	
Amount Due EOP	Amount to pay at the due date at the end of the period	
DPD	Days past due counter at the end of period (month)	
<i>Behavioural Characteristics</i>		
MOB	Month on Book	
Limit	Credit limit at the observation point	
UT	Utilisation rate at observation point	$= \text{Avg_OB}_1 / \text{Limit}_1$
UT16	Utilisation rate for the last 6 months	$= \sum_{t=1}^6 \text{Avg_OB}_t / \sum_{t=1}^6 \text{Limit}_t$
b_AvgOB1_to_MaxOB1_ln	The logarithm of average OB to maximum OB for the last month	$= \ln(\text{Avg_OB}_1 / \text{Max_OB}_1)$
b_TRmax_deb1_To_Limit_ln	The logarithm of the maximum debit transaction (purchases) for the last month to the credit limit	$= \ln(\text{Max_deb}_1 / \text{Limit}_1)$
b_TRAvg_deb1_to_avgOB1_ln	Logarithm of average debit transaction to average OB for the last month	$= \ln(\text{Avg_deb}_1 / \text{Avg_OB}_1)$

Table 3.3 List of the original raw data, behavioural, application and macroeconomic characteristics

Characteristic Name	Description	Formula
b_TRsum_deb1_to_TRsum_crd1_ln	Logarithm of sum debit transaction to sum of credit transactions (payments) for the last month	$= \ln(Sum_deb_1 / Sum_crd_1)$
b_AvgOB16_to_MaxOB16_ln	The logarithm of average OB to maximum OB for the last six months	$= \ln\left(\frac{\sum_{t=1}^6 Avg_OB_t}{6} / \max_{t=1...6} Max_OB_t\right)$
b_TRmax_deb16_To_Limit_ln	The logarithm of the maximum debit transaction (purchases) for the last six months to the credit limit	$= \ln\left(\max_{t=1...6} deb / \frac{\sum_{t=1}^6 Limit_t}{6}\right)$
b_TRavg_deb16_to_avgOB16_ln	Logarithm of average debit transaction to average OB for the last six months	$= \ln\left(\frac{\sum_{t=1}^6 Avg_deb_t}{6} / \frac{\sum_{t=1}^6 Avg_OB_t}{6}\right)$
b_TRsum_deb16_to_TRsum_crd16_ln	Logarithm of sum debit transaction to sum of credit transactions (payments) for the last six months	$= \ln\left(\frac{\sum_{t=1}^6 sum_deb_t}{\sum_{t=1}^6 sum_crd_t}\right)$
b_UT1_to_AvgUT16ln	The logarithm of the utilisation rate at the observation point divided by the average utilisation rate for the last six months	$= \ln\left(\frac{Avg_OB_1}{Limit_1} / \frac{\sum_{t=1}^6 Avg_OB_t}{\sum_{t=1}^6 Limit_t}\right)$
b_UT1to2ln	The logarithm of the utilisation rate at observation point to the previous months	$= \ln\left(\frac{Avg_OB_1}{Limit_1} / \frac{Avg_OB_2}{Limit_2}\right)$
b_UT1to6ln	The logarithm of the utilisation rate at observation point to the utilisation rate six months before	$= \ln\left(\frac{Avg_OB_1}{Limit_1} / \frac{Avg_OB_6}{Limit_6}\right)$
b_NumDeb13to46ln	The logarithm of the number of debit transactions for the last three months to the number of debit transactions for months 4-6 before the observation point	$= \ln\left(\frac{\sum_{t=1}^3 num_deb_t}{\sum_{t=4}^6 num_crd_t}\right)$
b_inactive13	Binary indicator whether the account was inactive last three months	$\sum_{t=1}^3 Avg_OB_t = 0$
b_avgNumDeb16	Average number of debit transactions for the last six months	$= \frac{\sum_{t=1}^6 NumDeb_t}{6}$
b_OB_avg_to_eop1ln	The logarithm of the average outstanding balance for the last month to the outstanding balance at the end of the period at the observation point	$= \ln\left(\frac{Avg_OB_1}{OB_eop_1}\right)$
b_DelBucket16	Maximum bucket of delinquency for the last six months	
b_pos_flag_0	A binary indicator of Point-of-sales (POS) transaction at the observation point month	
b_pos_flag_13	A binary indicator of Point-of-sales (POS) transaction for the last three months	
b_atm_flag_0	A binary indicator of ATM cash withdrawals at the observation point month	
b_atm_flag_13	A binary indicator of ATM cash withdrawals for the last three months	
b_pos_flag_used46vs13	A binary indicator of POS transactions 4-6 months before and no transactions for the last three months	
b_pos_flag_use13vs46	A binary indicator of POS transactions for the last three months and no transactions for 4-6 months before	
b_atm_flag_used46vs13	A binary indicator of ATM transactions 4-6 months before and no transactions for the last three months	

Table 3.3 List of the original raw data, behavioural, application and macroeconomic characteristics

Characteristic Name	Description	Formula
b_atm_flag_use13vs46	A binary indicator of ATM transactions for the last three months and no transactions for 4-6 months before	
b_pos_use_only_flag_13	A binary indicator of POS transactions only for the last three months	
no_dpd	A binary indicator of no delinquency at the observation point	
max_dpd_60	Binary indicator if the Maximum number of Days Past Due was 60 or more for life-time	
<i>Application characteristics</i>		
AgeGRP1	Age less than 25	
AgeGRP3	Age more than 35	
customer_income_ln	The logarithm of the ratio of the customer monthly income to the average monthly income in a portfolio of N customers for T periods	$\ln\left(\frac{\text{customer_income}_t}{\sum_{t=1}^T \sum_{n=1}^N \text{customer_income}_{nt}}\right) / NT$
Edu_High	Education: higher	
Edu_Special	Education: Special	
Edu_TwoDegree	Education: two degrees/ PhD	
Marital_Civ	Marital status: Civil Marriage	
Marital_Div	Marital status: Divorced	
Marital_Sin	Marital status: Single	
Marital_Wid	Marital status: Widow	
position_Man	Employment status: Manager	
position_Oth	Employment status: Others	
position_Tech	Employment status: Technical Staff	
position_Top	Employment status: Top Manager	
sec_Agricult	Sector of Industry: Agriculture	
sec_Constr	Sector of Industry: Construction	
sec_Energy	Sector of Industry: Energy	
sec_Fin	Sector of Industry: Finance	
sec_Industry	Sector of Industry: Heavy Industry	
sec_Manufact	Sector of Industry: Manufacture	
sec_Mining	Sector of Industry: Mining	
sec_Service	Sector of Industry: Service	
sec_Trade	Sector of Industry: Trade	
sec_Trans	Sector of Industry: Transport	
car_Own	Car owner: Owner	
car_coOwn	Car owner: Co-Owner	
real_Own	Real estate: Owner	
real_coOwn	Real estate: Co-Owner	
reg_ctr_Y	The region of living: Capital or region centre	
reg_ctr_N	The region of living: Province	
child_1	Number of children: 0	
child_2	Number of children: 1-2	
child_3	Number of children: 3 or more	
<i>Macroeconomic Characteristics</i>		
Unempl_Inyoy	The logarithm of the unemployment rate change year on year	$= \ln\left(\frac{\text{Unempl}_t}{\text{Unempl}_{t-12}}\right)$

Table 3.3 List of the original raw data, behavioural, application and macroeconomic characteristics

Characteristic Name	Description	Formula
LCY_EURRate_Lnmom	The logarithm of the exchange rate of local currency to Euro at the observation point in comparison with the previous month	$= \ln\left(\frac{LCY/EUR_t}{LCY/EUR_{t-1}}\right)$
LCY_EURRate_Lnyoy	The logarithm of the exchange rate of local currency to Euro at the observation point in comparison with the same period of the previous year	$= \ln\left(\frac{LCY/EUR_t}{LCY/EUR_{t-12}}\right)$
CPI_Lnqoq	The logarithm of the current Consumer Price Index at the observation point to the previous quarter CPI	$= \ln\left(\frac{CPI_t}{CPI_{t-3}}\right)$
SalaryYear_Lnyoy	The logarithm of the Average Salary at the observation point in comparison with the same period of the previous year	$= \ln\left(\frac{Salary_t}{Salary_{t-12}}\right)$
<i>Limit Characteristics</i>		
l_ch1_ln	Limit change month ago	$= \ln(Limit_t/Limit_{t-1})$
l_ch6_ln	limit change 6 month ago	$= \ln(Limit_t/Limit_{t-6})$
<i>State characteristics</i>		
d_State_2_NA	Dummy variables for state N month ago (in the example, 2) if the current is 1 for non-active (NA), transactor (Tr), revolver (Re), repaid (RP), delinquent in the 1 st bucket (D1), delinquent in the 2 nd bucket (D2), and defaulted (Df).	
d_State_2_Tr		
d_State_2_Re		
d_State_2_D1		
d_State_2_D2		
d_State_2_Df		
s_cons	Consecutive number of months in the current state (without changes)	
s_month_since_NA (Tr, Re, D1, D2)	Number of months since the respective state NA, Tr, Re, D1, D2. If currently, the account is related, the value of the characteristic is 0. Default (Df) is absorbing state and could be once only, so we do not use it for factors.	
s_times	Number of times the account has been in the certain state (NA, TR, RE, RP, D1, D2, Df)	

Behavioural characteristics were created from the original raw data. Index definition in characteristic formulas means the month' number calculated backwards (see Table 3.4). For example, Month 1 is the current month at the observation point in time. So, month 2 means the previous month (or -1 month), etc. For behavioural characteristics calculation for some period, we use the form of the variable name as 16 the first number '1' is the number of the first and the second number '6' is the number of the last month in the period. For example, June is the current month at the observation point, or month number 1.

Table 3.4 Month numbers in behavioural characteristics

Month name	Jan	Feb	Mar	Apr	May	Jun
Month Number	6	5	4	3	2	1

Thus, if the observation point is June, for the characteristic ‘average outstanding balance for the last six months’ we use notation AvgBalance_16, which means the average outstanding balance for Jan-Jun. The notation AvgBalance_13 means the average outstanding balance for April – June. This type of codification is used for all behavioural variables.

In Table 3.5 we give the descriptive statistics of Independent variables: mean, standard deviation, minimum, and maximum values. The statistics are computed for the selected behavioural data sample of 218,384 observations from July 2011 till Dec 2012 based on accounts with at least 12-months history after activation. The detailed univariate analysis of the most important characteristics for each model or target is presented in related chapters.

Aggregated behavioural variables, computed with the original behavioural characteristics as ratios, are continuous variables and transformed by taking the natural logarithm for the linearization and adjustment of the ratio values below one (between zero and one) and above one, which can get very high values because of division by values close to zero. The statistics for one-month independent variables like b_TRmax_deb1_To_Limit_In are given for one-month lag only. For two and more month lags the distributions are close to the one-month lag because they are generally the same variables, but with various monthly shifts. For example, b_TRmax_deb1_To_Limit_In value for July 2011 observation point is the same as b_TRmax_deb2_To_Limit_In for August 2011 observation point.

Table 3.5 Descriptive statistics of Independent variables

Variable	Mean	Std Dev	Minimum	Maximum
<i>Behavioural variables</i>				
mob	13.51793	4.827455	6	25
Limit	6577.42	3718.08	1000	25000
UT0_1	0.593668	0.377371	0	1
b_UT1to2ln	-0.08124	2.246459	-11.5129	9.21034
b_UT1to6ln	0.091862	3.73345	-11.5129	9.21034
avg_balance	3662.56	3173.67	0	24844.67

Table 3.5 Descriptive statistics of Independent variables

Variable	Mean	Std Dev	Minimum	Maximum
avg_deb_amt	164.8021	298.969	0	24000
sum_crd_amt	497.9807	969.8792	0	72756.08
sum_deb_amt	520.488	1013.27	0	73092.23
max_deb_amt	431.6089	914.1299	0	72369.45
min_deb_amt	81.79572	259.4212	0	24000
b_AvgOB1_to_MaxOB1_ln	-0.15095	0.373702	-11.5129	0
b_TRmax_deb16_To_Limit_ln	-2.72296	2.638661	-14.5087	3.847278
b_TRmax_deb16_To_avgOB16_ln	-1.04142	1.817086	-11.5129	16.11488
b_TRavg_deb16_to_avgOB16_ln	-2.67046	1.518592	-11.5129	12.93683
b_TRsum_deb16_to_avgOB16_ln	-0.05021	1.431545	-11.5129	16.11488
b_TRsum_deb16_to_TRsum_crd16_ln	0.125644	1.309218	-11.7382	11.3306
b_TRmax_deb1_To_Limit_ln	0.063842	0.206466	0	33.97551
b_TRavg_deb1_to_avgOB1_ln	-1.6822	4.25762	-11.3721	13.50133
b_TRsum_deb1_to_TRsum_crd1_ln	1.430965	3.818373	-12.1626	10.82287
b_NumDeb13to46ln	-0.09114	2.562781	-11.5129	9.21034
b_avgNumDeb13	2.690313	2.578128	0	182
b_OB13_to_OB46ln	-0.03284	2.684412	-14.862	12.95496
b_OB1_to_OB2_ln	-0.07668	2.188966	-14.039	13.8779
b_OB_avg_to_eop1ln	0.297785	1.66411	-3.13549	10.77501
b_pos_flag_use13vs46	0.107476	0.309718	0	1
b_atm_flag_use13vs46	0.101662	0.302204	0	1
b_pos_use_only_flag_13	0.097123	0.296125	0	1
b_atm_use_only_flag_13	0.25789	0.437474	0	1
b_TRsum_crd1_to_OB1_ln	-2.39728	2.679099	-11.5834	12.11063
b_payment_lt_5p_1	0.517574	0.499692	0	1
b_maxminOB_limit_1_ln	-2.60605	1.572124	-14.6856	6.866695
b_OBbias_1_ln	0.092261	0.969539	-3.2581	3.258097
b_maxminOB_avgOB_1_ln	-1.84667	1.534619	-13.9991	5.899721
b_TRsum_deb1_to_2_ln	-0.1426	2.499068	-13.4455	13.364
b_TRsum_crd1_to_2_ln	-0.15321	3.317569	-12.9002	13.07044
l_ch1_ln	0.013527	0.090737	-0.60488	2.807054
l_ch1_flag	0.03325	0.179289	0	1
l_ch6_flag	0.096408	0.29515	0	1
<i>Application variables</i>				
age	37.13331	10.10096	21	59
customer_income_ln	-0.28765	0.474374	-1.02962	1.966113
Edu_High	0.463549	0.498671	0	1
Edu_Secondary	0.179225	0.383542	0	1
Edu_Special	0.33926	0.473459	0	1
Edu_TwoDegree	0.017966	0.132827	0	1
Marital_Civ	0.050738	0.219462	0	1
Marital_Div	0.106825	0.308891	0	1

Table 3.5 Descriptive statistics of Independent variables

Variable	Mean	Std Dev	Minimum	Maximum
Marital_Mar	0.602627	0.489356	0	1
Marital_Sin	0.210572	0.407716	0	1
Marital_Wid	0.029239	0.168476	0	1
position_Empl	0.495625	0.499982	0	1
position_Man	0.111583	0.314853	0	1
position_Oth	0.138067	0.344971	0	1
position_Tech	0.230231	0.420982	0	1
position_Top	0.024495	0.154579	0	1
sec_Agricult	0.033332	0.179502	0	1
sec_Constr	0.017169	0.129901	0	1
sec_Energy	0.046799	0.211209	0	1
sec_Fin	0.082694	0.27542	0	1
sec_Gov	0.409957	0.491827	0	1
sec_Industry	0.010208	0.100516	0	1
sec_Manufact	0.019368	0.137815	0	1
sec_Mining	0.046963	0.211561	0	1
sec_Service	0.226676	0.418682	0	1
sec_Trade	0.088946	0.284666	0	1
sec_Trans	0.017888	0.132546	0	1
car_Own	0.174873	0.379859	0	1
car_coOwn	0.071777	0.258118	0	1
car_no	0.753351	0.431062	0	1
real_Own	0.451348	0.497629	0	1
real_coOwn	0.275815	0.446925	0	1
real_no	0.272837	0.445419	0	1
reg_ctr_Y	0.269004	0.443443	0	1
reg_ctr_N	0.659661	0.473824	0	1
reg_ctr_cap	0.071335	0.257384	0	1
child_1	0.377144	0.484673	0	1
child_2	0.293491	0.459134	0	1
child_3	0.032417	0.177104	0	1
child_0	0.298693	0.457686	0	1
<i>Macroeconomic variables</i>				
Unempl_lnyoy	-0.05452	0.033011	-0.12731	-0.025
UAH_EURRate_lnmom	0.004638	0.022414	-0.03483	0.05573
UAH_EURRate_lnyoy	-0.01059	0.089315	-0.17062	0.14323
CPI_Inqoq	0.008691	0.014782	-0.01609	0.03657
SalaryYear_lnyoy	0.104684	0.038408	0.01882	0.16382
<i>State variables</i>				
s_cons	5.016049	1.778304	1	6
s_month_since_NA_full	5.7805	2.477922	0	7
s_month_since_Tr_full	6.765453	1.087746	0	7

Table 3.5 Descriptive statistics of Independent variables

Variable	Mean	Std Dev	Minimum	Maximum
s_month_since_Re_full	0.88386	2.102951	0	7
s_month_since_RP_full	6.332922	1.76262	0	7
s_month_since_D1_full	6.655942	1.367952	0	7
s_month_since_D2_full	6.916604	0.673915	0	7
s_times_NA	0.76431	1.718128	0	6
s_times_TR	0.067602	0.328905	0	6
s_times_RE	4.832531	1.987947	0	6
s_times_RP	0.15937	0.416118	0	3
s_times_D1	0.103201	0.456344	0	6
s_times_D2	0.019614	0.16089	0	4
d_StateFull_1_NA	0.124135	0.329736	0	1
d_StateFull_1_Tr	0.010649	0.102645	0	1
d_StateFull_1_Re	0.801384	0.398959	0	1
d_StateFull_1_RP	0.02689	0.161761	0	1
d_StateFull_1_D1	0.020566	0.141925	0	1
d_StateFull_1_D2	0.004025	0.063313	0	1
d_StateFull_1_Df	0.012352	0.110451	0	1

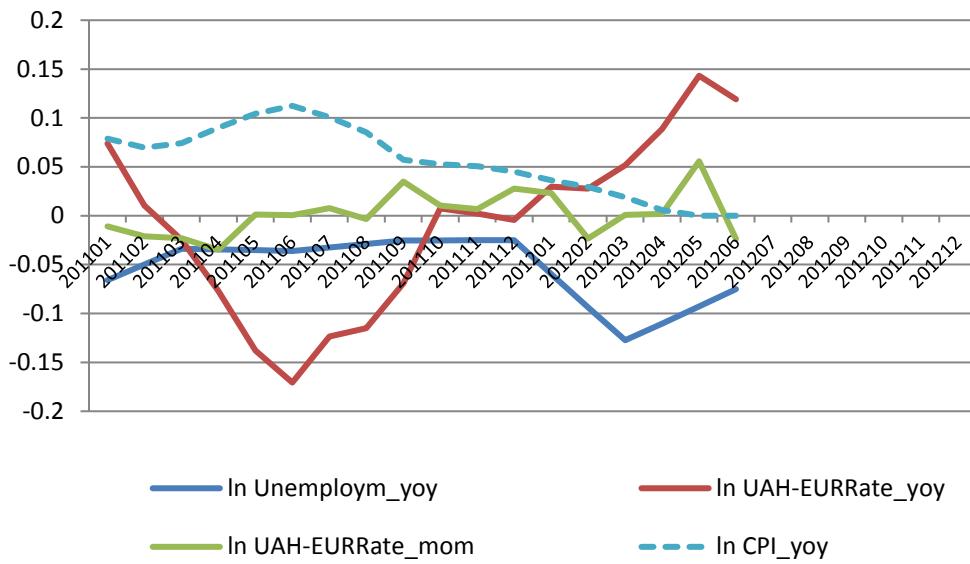
The most independent variables are used as covariates in models. We explain the reasons for the inclusion of variable in models in each related chapter. The explanation includes the expectation of the impact of each covariate on the target variable, which is used in the related model. However, data sets may vary for different models. Firstly, time aggregated behavioural variables, which related to average values for some period, were replaced by original variables (for example, the outstanding balance, purchase transaction amounts) for panel data models to avoid high multicollinearity. In this case, the problem is caused by overlapping of monthly variables such as average outstanding balance for the last six months before the observation date. It will be slightly changed with a one-month shift and this variable for month t will be highly correlated with variable value at month $t-1$. So we used for panel data variables or rates related to one month only with some time lags.

Secondly, specific variables are introduced into the thesis in the later chapter, are not used in preceding chapters. For example, the utilisation rate models do not contain covariates related to the state because state variables are introduced in a later chapter about transition probabilities.

Thirdly, we excluded highly correlated variables, which reflect the same behaviour of the account or can be computed as an expression of other variables. Finally, we have excluded variables, which have found in preliminary analysis had a very low correlation with dependent, or target, variables.

Macroeconomic indicators describe the dynamic of the country economy and are available from the official government sources. Figure 3.4 contains some key macroeconomic indicator changes for one year and one month. The indicators are unemployment rate, local currency to euro exchange rate, and consumer price index till the maximum available observation point June 2012. Visual analysis of the macroeconomic indicators demonstrates volatility and non-stable trends. For instance, CPI changes were stable in 2011 and negative values indicate that the price index slightly decreased for this period compared with previous year values. A significant decrease can be seen in April-May 2011. UAH/EUR exchange rate shows how many Euros one local currency costs. Thus, a higher UAH-EUR value means more expensive and stronger local currency. Since July 2011 the year on year change in local currency showed stable increases in value that can be explained by a deep fall the year before. However, monthly changes in the UAH-EUR exchange rate showed a minor fluctuations without a strong trend.

Figure 3.4 The dynamic of macroeconomic indicators: unemployment rate, Local currency to Euro exchange rate, and CPI till June 2012



3.4 Dependent Variables

We use three types of dependent variables: i) the income amount from a transaction and the interest income amount, ii) the utilisation rate computed as the ratio of the outstanding balance to the credit limit, and iii) the delinquency state of an account at time t . For transactional income, we have two original dependent variables: income from Point-of-Sales transaction (Target_POS), generated from the interchange fees, and cash withdrawal transaction (Target_ATM), generated by ATM fees. Dependent, or target, characteristics such as the utilisation rate and state are generated from the behavioural characteristics.

The dependent variable for the utilisation rate is computed as the outstanding balance divided by the credit limit. It is used for the interest income prediction. The interest income is computed as the average outstanding balance for a period multiplied by the annual percentage rate (APR) for this period and multiplied by the appropriate number of months.

In Table 3.6 we give the descriptive statistics for dependent variables: mean, standard deviation, minimum, and maximum values. We give statistics for six months lags. The descriptive statistics values show that the target for different lags are quite stable with

some increase in means and standard deviation for transactional income, and slight decrease in the same statistics for utilisation rate.

Table 3.6 Descriptive statistics of Dependent variables

Variable	Mean	Std Dev	Minimum	Maximum
Target_POS_m1	0.668559	3.308274	0	69.9997
Target_ATM_m1	12.4506	27.88999	0	249.6
Target_POS_m2	0.683924	3.350361	0	69.9997
Target_ATM_m2	13.39609	29.87108	0	249.6
Target_POS_m3	0.69252	3.372742	0	69.9997
Target_ATM_m3	14.07918	31.27919	0	249.6
Target_POS_m4	0.701165	3.411877	0	69.9997
Target_ATM_m4	14.50021	32.18445	0	249.6
Target_POS_m5	0.711091	3.437341	0	69.9997
Target_ATM_m5	14.78888	32.90481	0	249.6
Target_POS_m6	0.727094	3.490426	0	69.9997
Target_ATM_m6	14.97395	33.51906	0	249.6
Target_UT_plus1	0.585749	0.37672	0	1
Target_UT_plus2	0.577909	0.376447	0	1
Target_UT_plus3	0.5707	0.376723	0	1
Target_UT_plus4	0.563043	0.377512	0	1
Target_UT_plus5	0.554737	0.37856	0	1
Target_UT_plus6	0.545119	0.379772	0	1

Account state variables can be used both as independent and as a dependent variables. In the first case, they are used as covariates in regression models. In the second case, they are targets for the prediction of the probability of transition. The distributions of state variables for 1-month lag are given in Table 3.3. Detailed distributions of the account states are given in Chapter 6.

3.5 Data sampling

We split the data into 3 data samples: development, validation out-of-sample, and validation out-of-time. The development (or training) sample and validation out-of-sample data set are split up in the proportion of 80/20 with a univariate random distribution. We ran the univariate random number generator at the account (not observation) level and assigned all accounts, which gained the random value below 0.8, to the development sample; otherwise, all accounts, which gained the random value equal or above 0.8, were assigned to validation out-of-sample data set.

In this research, we have panel data, but in many models, we used as pooled data. Only models from Chapter 7 investigate panel models, so they consider time variance for an account, presented in the panel data. So, each account gives for each sample as many observations as possible according to the observation and performance windows definitions for each vintage. For example, an account with 18-month history has 12 months for observation and performance windows, and these windows can be shifted 6 times. So in total an account with 18-time cohorts gives 7 observations for data sample: dated as 12, 13, 14, 15, 16, 17, and 18 months of history. However, we split the data sample at the account level, not at the observation level. So, the same account can be either in the development or the validation sample, but not in both.

Also, for some models, we use Validation out-of-time sample. This contains observations with a performance period from the last 6 months of the data set – from July 2012 to Dec 2012, and observation window – from January 2012 to June 2012.

Table 3.7 shows the number of observations in each data sample account with 12-month or longer history by the last months of the performance window. In total the development sample contains 98,624 observations, the validation out-of-sample data contains 41,325 observations, and the validation out-of-time data contains 78,435 observations. However, some models can use constrained data sub-sets only. For example, the models for transactional income from point-of-sales and cash withdrawals use only accounts which had the appropriate type of transactions. The data samples for models for transition probabilities from certain states contain only observations each of which is in the appropriate state at the observation point.

Table 3.7 Number of observations for the behavioural sample by month (for accounts with 12 or more observations in a sample after July 2010)

Month	Development sample	Validation Out-of-sample	Validation Out-of-time	Total number with behaviour for 12m+ observations
201106	1,791	753		2,544
201107	2,850	1,192		4,042
201108	4,631	1,970		6,601
201109	5,936	2,487		8,423
201110	7,160	2,981		10,141
201111	8,374	3,511		11,885
201112	8,734	3,648		12,382
201201	9,889	4,140		14,029
201202	9,887	4,138		14,025
201203	9,874	4,136		14,010
201204	9,871	4,135		14,006
201205	9,839	4,125		13,964
201206	9,788	4,109		13,897
201207			13,717	13,717
201208			13,504	13,504
201209			13,260	13,260
201210			12,942	12,942
201211			12,614	12,614
201212			12,398	12,398
Total	98,624	41,325	78,435	218,384

3.6 Conclusion

This Chapter described the data sources, gave the list of dependent and independent variables used for further modelling, and discussed data sampling. We have a panel data of credit cards with application and behavioural variables, collected for a two and a half year period, and use one and a half years to compute behavioural and target variables for observation and performance windows. Also, we added several macroeconomic indexes to test their impact on the predicted credit cards income, utilisation and state variables, but the time horizon is quite short to make reliable conclusions about the impact of the macro variables on the investigated behaviour of customers. We have 50 behavioural, 40 application (including binary dummy variables) variables, 5 macroeconomic variables, and 4 state variables. However, we use more than 20,000 accounts, which give around 218 thousand observations, and it is expected that the estimations and models obtained with this number of observations

will be reliable. The development and validation samples defined in this chapter are used for all models in the next chapters. Detailed descriptions of the distributions of the most important dependent and independent variables are given in related chapters.

4 Chapter Four. A comparative analysis of predictive models for credit limit utilisation rate

4.1 Introduction

A credit card is a banking product, which has a dual nature both as providing a convenient loan and providing a payment tool. This makes the task of profitability prediction for this product more complicated than for standard loans. Moreover, a credit card has a fluctuating balance, and its accurate forecast is an actual problem of credit risk management, liquidity risk, business strategies, customer segmentation and other strategies of bank management. The use of traditional techniques gives acceptable empirical results. However, a majority of the industrial models are simplified and apply many assumptions. For example, these assumptions include data stationarity, an unchanged impact of the macroeconomic parameters on micro-level characteristics for long periods, ideal shapes of distributions, which are fitted to satisfy the modelling method. A lot of these assumptions and limitations are caused by applied modelling techniques and are made to simplify the economic entity to balance modelling costs, reduce the time for calculations and gain sufficient prediction accuracy. The difficulties with accuracy of credit card usage models can be caused by more complex customer behaviour – ‘wishing to use’ and ‘how to use’ - in comparison with customer risk model – ‘must pay’. Credit products, especially credit cards, are sensitive both to systematic factors such as macroeconomic trends and cycles and to individual factors such as behavioural patterns of a customer, for example, the spending desire, life circumstances, and personal financial literacy.

This research proposes and investigates a set of approaches to the prediction of the credit limit utilisation rate. We define a credit limit utilisation rate as the outstanding balance divided by the credit limit. A credit card holder’s behaviour pattern changes over time, and these changes have an impact not only on the risk but also on the income generated by the card.

The usage of credit cards is a topic discussed in many papers (for example, Crook et al., 1992; Banasik and Crook, 2001; Hand and Till, 2003) along with credit cards outstanding balance (Kim and DeVaney, 2001; Tan et al., 2011; Leow and Crook, 2016). However, there is a lack of research on the prediction of the credit limit

utilisation rate. We implemented some methods that have been used for the prediction of proportions such as Loss Given Default (Arsova et al., 2011; Barkel and Siddiqi, 2012; Yao et al., 2014) and applied them to *the utilisation rate* (for example, Agarwal et al., 2006). Also we used the credit card usage approaches (Crook et al., 1992; Banasik and Crook, 2001) as the probability of full use or no use of credit card in the two-stage model.

We test a one-stage and a two-stage utilisation rate model. A one-stage model is a direct estimation of the utilisation rate. A two-stage model is an alternative method of estimation: at the first stage we calculate the probability of full utilisation and zero utilisation, and at the second stage calculate the utilisation rate for the range between zero and one. We use cross-sectional socio-demographic, time-varying behavioural, and time-varying macroeconomic characteristics as predictors. We expect that the use of approaches like weighted logistic regression with a binary transformed sample can give more accurate prediction than linear regression.

We compare the following methods for credit limit utilisation rate prediction: i) five direct estimation techniques such as ordinary linear regression, beta regression, beta transformation plus general linear models (GLM), fractional regression (quasi-likelihood estimation), and weighted logistic regression for binary transformed data, ii) two-stage models with the same direct estimation methods and the logistic regression at the first stage for the estimation of the probability of full use and no use.

The Chapter has the following structure. Section 2 describes definitions and the modelling concept of the utilisation rate, and discusses the methods for the utilisation rate prediction, which are selected for this research. We present a comparative analysis of a set of methods for the utilisation rate prediction with a given data sample. These are i) five direct estimation techniques such as ordinary linear regression, beta regression, beta transformation plus general linear models (GLM), fractional regression (quasi-likelihood estimation), and weighted logistic regression for binary transformed data, ii) two-stage models with the same direct estimation methods and the logistic regression at the first stage for the estimation of the probability of full use and no use.

Section 3 gives a brief portfolio overview of vintage analysis and macroeconomic environment. This section discovers the empirical background for the utilisation rate prediction instead of the direct prediction of the credit card outstanding balance. Section 4 describes the model setups, contains the univariate analysis and distributions of the dependent and explanatory variables.

Section 5 contains the results of the estimation of direct and indirect models, built with five approaches, and discusses the comparative analysis of the models' validation. Finally, section 6 concludes the results of the investigation and suggests the most accurate approach for the prediction of utilisation rate.

4.2 Utilisation rate model and methods

We try to predict the utilisation rate at the account level. We use pooled panel data, and each account has monthly cohorts, which is dependent on the vintage. This means that each month represents a single observation for an account. We take as many observation points for each account, as possible with limitations for the observation window - the period for behavioural characteristics calculation (at least six months for some variables), and for the performance window - the period for the target variable (at least six months).

Each observation is a single case. We apply five methods for proportions modelling: linear regression (OLS), fractional regression (quasi-likelihood), beta-regression (non-linear), beta-transformation + OLS/GLM, and weighted logistic regression with binary data transformation.

4.2.1 Linear model approach

Let's define the utilisation rate as outstanding balance (OB) and credit limit (L) ratio

$$UR = \frac{OB}{L} \quad (4.1)$$

The utilisation rate is defined as a ratio of the sum of the outstanding balance amounts for a certain period and average credit limit L for the same period:

$$UR_{t0,t1} = \frac{\sum_{t \in [t0,t1]} OB_t}{\text{Avg}(L_t)} \quad (4.2)$$

The utilisation rate $UT_{i,t}$ depends on behavioural, application, and macroeconomic characteristics, and also on the previous period's utilisation rate with time lags:

$$UT_{i,t+T} = \sum_{l=0}^{L_{max}} \phi_l UT_{i(t-l)} + \sum_{b=1}^{B_{cnt}} \beta_b B_{bit} + \sum_{a=1}^{A_{cnt}} \alpha_a A_{ai} + \sum_{m=1}^{M_{cnt}} \gamma_m M_{mt} \quad (4.3)$$

where

$\phi, \alpha, \beta, \gamma$ are regression coefficients (slopes) to be estimated;

B_{bit} is a behavioural characteristic b for case i in period t – time-varying;

A_{ai} is application characteristic a for case i – time-invariant;

M_{mt} is macroeconomic factor m in period t – time-varying;

l is time lag;

T is the period of prediction in months (can be 1 month or more);

$UT_{i(t-l)}$ is the utilisation rate for case i for time $t-l$ (current time minus l months) from L_{max} time lags, used as a predictor.

The time parameter t is used to identify the point in time for behavioural characteristics' lag calculation, if t is fixed in time and cross-sectional analysis is used. So it does not matter what period t the observations relate to. Panel data is investigated in Chapter 6. The cross-sectional model is given by equation (4.3).

The model can also be written in vector format such as:

$$UT_{it+T} = \phi UT_{i,t-L} + \alpha^T A + \beta^T B + \gamma^T M \quad (4.4)$$

where ϕ, α^T, β^T and γ^T are vectors of parameters to be estimated;

A is a vector of application characteristics (time-invariate);

B is a vector of behavioural characteristics (time-varying);

M is a vector of macroeconomic characteristics (time-varying);

L is time lag for the utilisation rate.

The linear regression is applied for proportions modelling (Belotti and Crook, 2012; Arsova et al., 2011) and is written in the general form as

$$\begin{aligned}y_i &= \beta_0 + \beta^T \mathbf{x}_i + \varepsilon_i \\ \varepsilon_i &\sim N(0, \sigma^2)\end{aligned}\tag{4.5}$$

where \mathbf{x}_i is a vector of explanatory variables, β is a vector of unknown coefficients of regression, ε_i are unobserved scalar random errors, which has a standard normal distribution.

The unknown parameters can be estimated with, for instance, ordinary least squares (OLS), which minimises the sum of the squares of the differences between values, predicted by a linear function, and observed values (Greene, 2002).

One of the weaknesses of the linear regression application for Utilisation rate modelling is the unlimited predicted range of the variable. It can be fixed with a conditional function like the following

$$f(x) = \begin{cases} 0, & f(x) < 0 \\ f(x), & f(x) \in [0;1] \\ 1, & f(x) > 1 \end{cases}\tag{4.6}$$

This approach can cause a high concentration of the rate values on the bounds 0 and 1. However, this approach is easy to use and it can be applied for quick preliminary analysis of correlations and general trends.

4.2.2 Beta-regression approach

One of the ways to set bounds for a target variable is to apply a transformation of the empirical distribution to a theoretical one with appropriate limits. A beta distribution can be applied to match the distribution shape with bounds, in our case they are 0 and one.

A beta-distribution has one important feature: it can be bounded between two values and parameterised by two positive parameters, which define the shape of the distribution. The empirical probability density function of the utilisation rate distribution is U-shape. The parameters α and β should be set up to match the density function shape to minimise the residuals between the empirical and the theoretical distribution.

The beta distribution probability density function is

$$f(x) = \frac{1}{B(\alpha, \beta)} x^{\alpha-1} (1-x)^{\beta-1} \quad (4.7)$$

where $\alpha, \beta > 0$ and $B(\alpha, \beta)$ is beta-function.

The parameters α and β are set to match the theoretical distribution as closely as possible to empirical one. A beta distribution function can be represented using a Gamma function as

$$\text{Beta}(y, \alpha, \beta) = \frac{\Gamma(\alpha + \beta)}{\Gamma(\alpha)\Gamma(\beta)} y^{\alpha-1} (1-y)^{\beta-1} \quad (4.8)$$

where Γ is Gamma function.

The parameters α and β can be calculated as a function of the distribution mean and standard deviation:

$$\begin{aligned} \text{Alpha} &= \left[\mu^2 \times \frac{(1-\mu)}{\sigma^2} \right] - \mu \\ \text{Beta} &= \text{Alpha} \times \left(\frac{1}{\mu} - 1 \right) \end{aligned} \quad (4.9)$$

Because the outcome is in the range between 0 and 1 the logistic transformation is used to find the dependences between predictors the $\mathbf{x}(a)$ and dependent variables:

$$\mu(a) = L(x(a)^T \beta) = \frac{e^{x(a)^T \beta}}{1+e^{x(a)^T \beta}} \quad (4.10)$$

The maximum likelihood estimation method is applied to find the regression coefficients β

$$l(\mathbf{b}, \phi) = \sum_{a \in A} \ln \text{Beta}(y(a), L(x(a)^T \beta), \phi) \quad (4.11)$$

The inverse beta-transformation with cumulative distribution function is applied to find the real rate value according to the estimated one. After this transformation the logistic regression is applied to estimate dependence between the beta-function and predictors, and then the inverse transformation from the cumulative probability function is applied to get the utilisation rate. The Beta distribution is not designed for values x , which are exactly equal to 0 or exactly equal to 1.

4.2.3 Beta-transformation plus OLS

The algorithm is the following. First, find the beta-distribution coefficients (alpha and beta) to fit the development sample distribution using non-linear regression procedures. Second, replace the value of the real target variable by those from the ideal beta-distributed. Thirdly, find the appropriate normal distributed value. Then, run OLS or a Generalized Linear Mixed Model to find regression coefficients. Then perform the inverse transformation for a normal distribution and then for the inverse for the beta regression. To obtain a prediction it is necessary to transform linear regression results with normal and then beta regression with constant alpha and beta coefficients found at the first stage.

4.2.4 Utilisation rate Modelling with Fractional logit transformation

The utilisation rate is bounded between 0 and 1 and requires appropriate methods to keep the predicted value in this range. One of the techniques is fractional logit regression proposed by Papke & Wooldridge (1996). The Bernoulli log-likelihood function is given by

$$l_i(\mathbf{b}) = y_i \log[G(\mathbf{x}_i \mathbf{b})] + (1 - y_i) \log[1 - G(\mathbf{x}_i \mathbf{b})] \quad (4.12)$$

The quasi-likelihood estimator of β is obtained from the maximization of

$$\max_{\mathbf{b}} \sum_{i=1}^N l_i(\mathbf{b}) \quad (4.13)$$

Crook and Belotti (2009) apply the Fractional logit transformation for Loss Given Default parameter modelling:

$$T_{LGD} = \log(LGD) - \log(1 - LGD) \quad (4.14)$$

where LGD is Loss Given Default.

The LGD parameter has the same features as the utilisation rate, which yield the U-shape and bimodal distribution. Thus, similar techniques can be applied for the modelling and parameter estimations. The general utilisation rate equation can be written as (4.3).

The utilisation rate UT is transformed to UT_{TR} for regression estimation as follows

$$UT_{TR} = \ln(UT) - \ln(1 - UT). \quad (4.15)$$

The inverse transformation of the predicted value is the following:

$$UT = \frac{\exp(UT_{TR})}{1 + \exp(UT_{TR})} \quad (4.16)$$

Quasi-likelihood methods are used to estimate the parameters in the model.

The SAS procedure GLIMMIX is used for the regression coefficients estimation. Procedure GLIMMIX performs estimation and statistical inference for generalized linear mixed models (GLMMs). A generalized linear mixed model is a statistical model that extends the class of generalized linear models (GLMs) by incorporating normally distributed random effects.

4.2.5 A weighted Logistic regression with binary transformation approach

A relatively innovative approach is the use of weighted logistic regression with binary transformation of the data sample (Arsova et al., 2011; Barkel and Siddiqi, 2012). The logit function is bounded between 0 and 1 and usually applied for the prediction of probability. To apply logistic regression, which uses a binary distribution of the target proportion, a variable needs to be transformed from a continuous to a binary form. To do this we consider utilisation rate as the probability to use the credit limit by 100%. For example, the utilisation rate 75% can be presented as a 75% probability of full use of the credit limit and 25% probability of zero use of the credit limit. This approach is used by Barkel and Siddiqi (2012) for Loss Given Default prediction. Each observation is presented as two observations (or two rows) with the same set of predictors according to the good/bad or 0/1 definition, which is used in logistic regression. The outcome with target 1 corresponds to the rate r , which determines the weight equal to rate r . The outcome with target 0 corresponds to the rate $1-r$, which determines the weight equal to $1-r$. The logistic regression probability of an event is the utilisation rate estimation.

The data sample is transformed according to the scheme from Table 4.1:

Table 4.1 The binary target transformation for weighted logistic regression

Utilisation rate	Binary recovery – target	Weight
1	1	1
0	0	1
R, 0<r<1	1 0	r 1-r

This set of methods is used to model LGD at account level. Stoyanov (2009) investigated, in particular, the following approaches to LGD account level modelling as a binary transformation of the LGD using uniform random numbers, and a binary transformation of the LGD using a manual cut-off. Arsova et al. (2011) applied both direct approaches to the LGD modelling such as OLS regression, beta regression and fractional regression and indirect approaches such as logistic regression with binary transformation of LGD by random number, logistic regression with binary transformation of LGD by weights, and also multi-stage models like Ordinal Logistic Regression with nested Linear Regression.

This study uses weighted logistic regression with a binary transformed sample for rate estimations. Logistic regression matches the log of the probability odds to a linear combination of the characteristic variables as

$$\text{logit}(p_i) = \ln\left(\frac{p_i}{1-p_i}\right) = \beta_0 + \boldsymbol{\beta} \cdot \mathbf{x}_i^T, \quad (4.17)$$

where

p_i is the probability of a particular outcome, β_0 and $\boldsymbol{\beta}$ are regression coefficients, \mathbf{x} are predictors.

$P_i = E(Y_i | \mathbf{x}_i, \boldsymbol{\beta}) = \Pr(Y_i = 1 | \mathbf{x}_i, \boldsymbol{\beta})$ is the probability of the event for observation i .

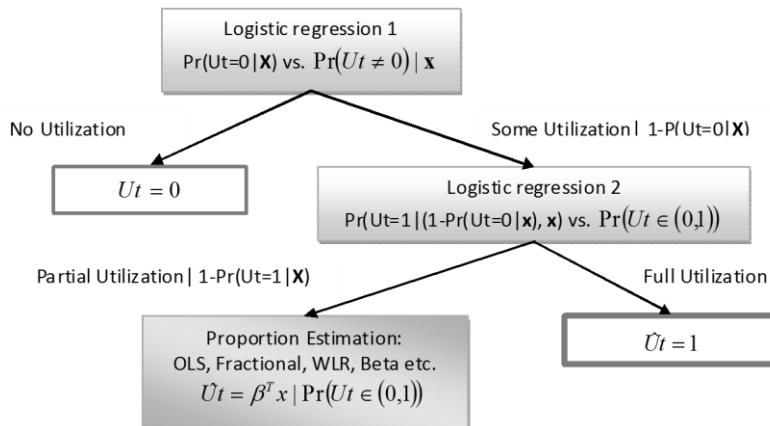
In general, this approach can be interpreted as the following: the utilisation equal to 10% is the same as from accounts 10 accounts having utilisation equal 100% and 90 accounts have utilisation equal to 0%.

4.2.6 Two-stage model

In a two-stage model in Figure 4.1 the first stage as the modelling of the probability to get a limiting value such as 0 or 1, is calculated, and then the proportion estimation in the interval (0;1) is applied.

At the first stage the probability that an account has zero utilisation ($\Pr(Ut=0)$) and then that an account has full utilisation ($\Pr(Ut=1)$) in the performance period is calculated with binary logistic regression. At the second stage the proportion between 0 and 1 excluding 0 and 1 values is calculated following a one-stage direct estimation methodology.

Figure 4.1 Two-stage model tree



The two-stage model utilisation rate is calculated with the following formula:

$$U_t = (1 - \Pr(U_t = 0))(\Pr(U_t = 1) + (1 - \Pr(U_t = 1)) \cdot E(U_t | U_t \neq 0, U_t \neq 1)) \quad (4.18)$$

where

$\Pr(U_t=0)$ and $\Pr(U_t=1)$ are the probabilities that the utilisation rate is equal to 0 or 1 respectively.

$E(U_t | U_t \neq 0, U_t \neq 1)$ is the utilisation rate estimation for the utilisation rates not equal to zero and not equal to one.

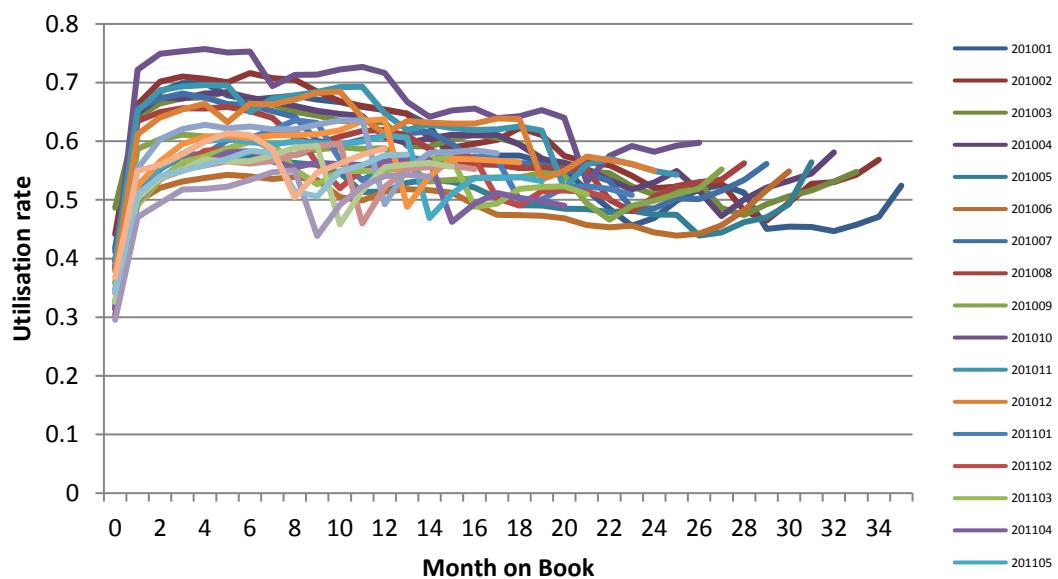
4.3 Portfolio overview

4.3.1 Portfolio vintage analysis and the utilisation rate distribution

The portfolio analysis in this chapter contains a review of the portfolio dynamics of the utilisation rate and outcome characteristic distributions. The vintage analysis demonstrates some portfolio characteristics for different vintages (activation date) such as the month of application and month of credit line activation. The characteristic which is investigated in the vintage analysis in this chapter is the utilisation rate, and the time variable is the month on book (MOB).

In Figure 4.2, which presents the vintages of the credit limit utilisation rate by Month on Book, each curve is related to a certain month of credit card activation, or vintage for the period from January of 2010 (marked as '201001') to May of 2011 (marked as '201105'). All curves of the utilisation rates by the month on book have a similar share to each other, a dramatic rise in the first and second month, then the values level-off for 3-5 months and a further slight decrease in the credit limit utilisation during the 12-18 month period. After a one-and-a-half-year period the utilisation rate of the majority of vintages converge to the same range of values between 45% and 60% with some insignificant fluctuations. However, the range of the utilisation rate at the peak is considerable, the low value of approximately 50% and the highest value almost 75%.

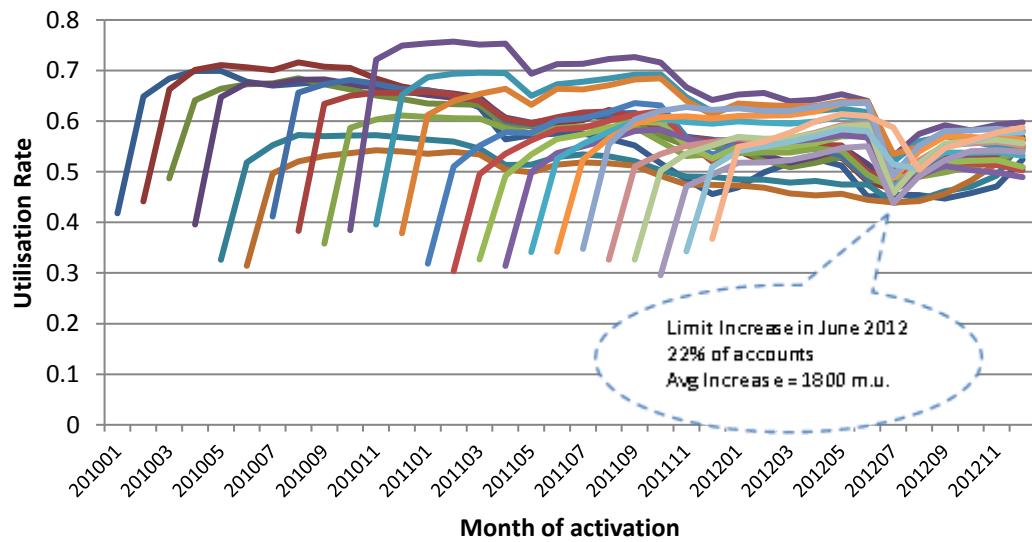
Figure 4.2 The vintages of the credit limit utilisation rate: by Month on Book



Thus if long-term fluctuations can be explained by some external factors and by unobserved individual factors, the initial increase in the utilisation rate is associated with certain specific to vintage conditions such as special terms, usage motivation, specific credit limits and credit policy rules for a certain period. Therefore, the vintage, or the period of the credit card activation, is one of the key factors which need to be considered in the modelling process.

The comparative analysis of the shape of the utilisation rate curve for the monthly vintages for the period 2010-2011 by month of balance, or vintage analysis, has shown that consumer behaviour is different depending on the month of the loan (credit card activation). However, some systematic similarities can be identified as it has been shown in Figure 4.3, which present the same data as Figure 4.2, but vintage curves are built by real months of activation instead of the months on book. So in Figure 4.3 each utilisation rate vintage curve starts at the point related to the month of activation. We can see that the utilisation rate for all portfolio drops in June 2012. This total shift is caused by the credit limit increase procedure for a significant number of accounts from all vintages or months of activation.

Figure 4.3 The utilisation rate vintage: by the month of activation

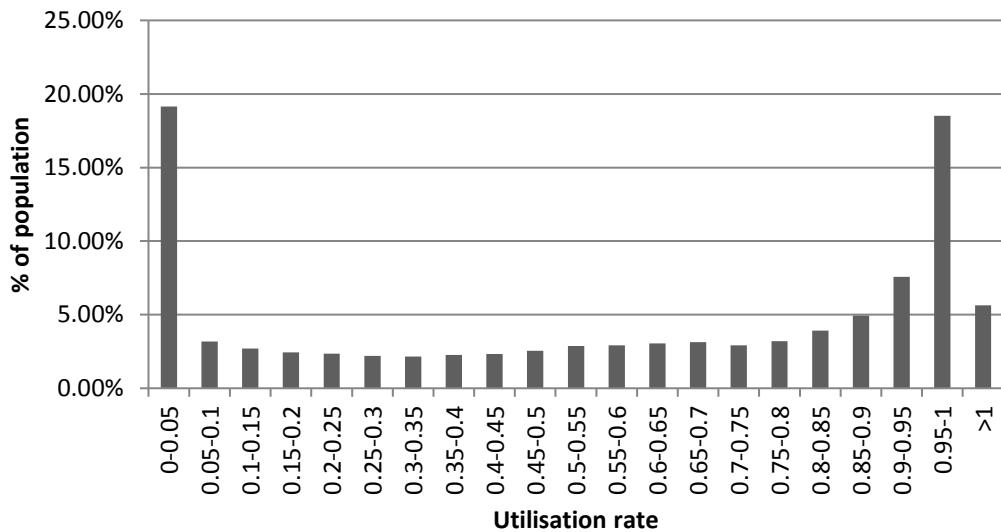


We have investigated a dynamic model of the utilisation rate, but it is also necessary to know the distribution of this characteristic, not only the mean values. A random variable includes investigating the probability density function (pdf). The shape of the distribution and general descriptive statistics give information for further analysis and

modelling steps such as characteristic behaviour, data quality, methods and techniques which can be applied for prediction. The majority of random variables in social and economic processes have a normal distribution and rarely have lognormal, poison and uniform distributions.

The utilisation rate density has a U-shape distribution as can be seen from the empirical data investigation. Figure 4.4 demonstrates the utilisation rate distribution for active accounts for the full data sample. The distribution is based on the credit limit utilisation history of active accounts and each case (or observation) presents an individual utilisation rate of the account at certain time.

Figure 4.4 The utilisation rate distribution for active accounts



As it can be seen from Figure 4.4 a little less than 20% of the observations have a utilisation rate from 0% to 5% and the same approximately 20% of observations have a utilisation rate from 95% to 100%. Other cases are distributed almost uniformly from 5% to 80%, and the slight increase in the observations can be observed after 85% rate value. However, in 5% of cases, the utilisation rate exceeds one hundred per cent. This indicates that the use of loan funds is higher than the set credit limit. This can be explained by i) credit overlimit of the outstanding balance; ii) technical accrual of the interest rates, fees and commissions on the principal account. For further analysis these cases are replaced by 100% utilisation to avoid misinterpretation of the utilisation rate and according to the business logic. However, the set of cases can be allocated as a separate category to investigate the reasons and drivers of the credit overlimit.

4.3.2 Macroeconomic environment

Macroeconomic indicators describe the dynamic of the Ukrainian economy and are available from the official government sources. Figure 4.5 contains some key macroeconomic indicator changes for one year and one month. The indicators are the unemployment rate, a local currency to euro exchange rate, and consumer price index.

We have selected these macroeconomic variables because we believe that they have a systemic impact on the usage of credit cards and utilisation rate.

We expect that the macro unemployment rate will be positively correlated with utilisation rate because a higher level of unemployment may cause the increase in demand for money. The year on year change of the unemployment rate has been selected because it is quite an inert index in the National statistical bureau and does not reflect rapid changes in the macroeconomy. The increase in Consumer Price Index means the rise in average prices of a basket of consumer goods and services. Thus, the usage of credit limit may also be increased in this case. The increase in local currency to Euro exchange rate (LCY/EUR exchange rate) may cause purchasing power loss because of growth of prices: directly for export goods and indirectly for local goods in proportion to the export component. Thus, the utilisation rate may have a positive correlation to the foreign currency exchange rate. We selected yearly and monthly changes of the foreign currency rate for testing in the model.

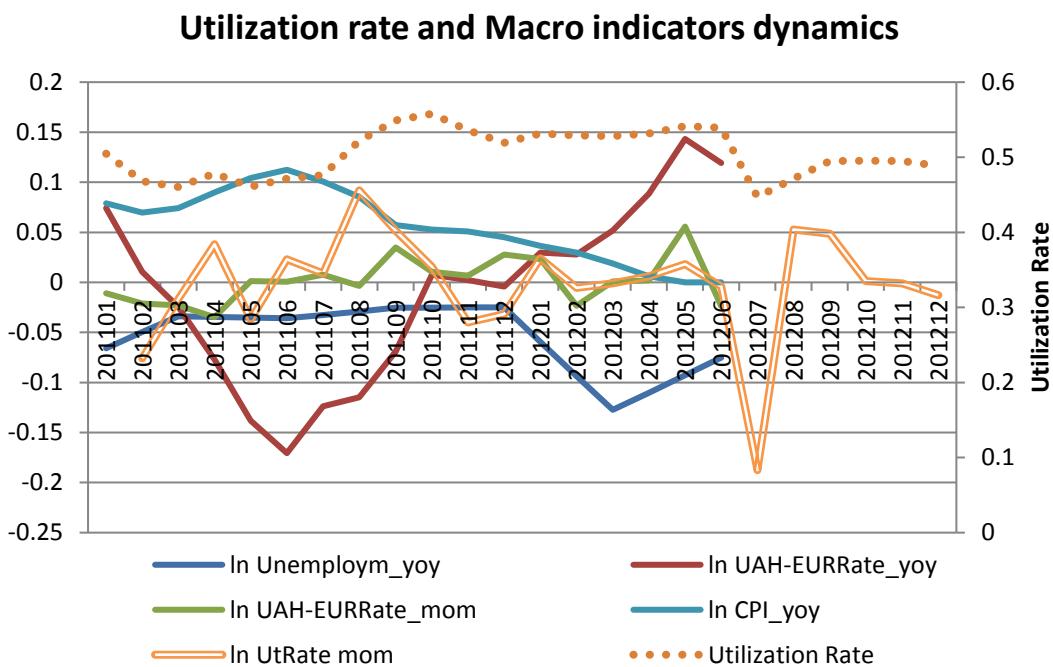
As can be seen from the Figure 4.5 the macroeconomic indicators demonstrate volatility and non-stable trends. For instance, CPI changes were stable in 2011 and negative values indicate that the price index slightly decreased for this period compared with previous year values. A significant decrease can be seen in April-May 2011. The UAH/EUR exchange rate shows how many Euros one local currency costs. Thus higher UAH-EUR value means a more expensive and stronger local currency. Since July 2011 the year on year change in local currency showed stable increases in value what can be explained by a deep fall the year before. However, monthly changes in the UAH-EUR exchange rate showed minor fluctuations without a strong trend.

We assume that in the short-term period the macroeconomic fluctuations have a more significant impact on the consumer behaviour than absolute indicators values. We expect that the changes in the customer behaviour is a reaction to the changes in the

environment, which can be revealed in a monthly period. Thus, we try to find correlations between the utilisation rate, which reflects the customer behaviour, and changes in macroeconomic variables, but not between the utilisation rate and macroeconomic trends.

We use the logarithm of the ratio of the current value of the indicator and the previous value of the indicator with some time lag. The logarithm is used for linearization or alignment of increase and decrease changes for unbiased scale for inclusion in linear models, which we use in this thesis. For example, for two times increase of indicator value the ratio will have a value equal to 2, and for two times the decrease of indicator value the ratio will have a value equal to 0.5. In comparison with ratio value, which reflect no change in indicator and equal to 1, we have a non-linear relationship as -0.5 and 1 for the same absolute value of changes as two times. To avoid this, we use logarithm, which give -0.69314 for two times decrease in indicator and 0.69314 for two times increase in indicator value.

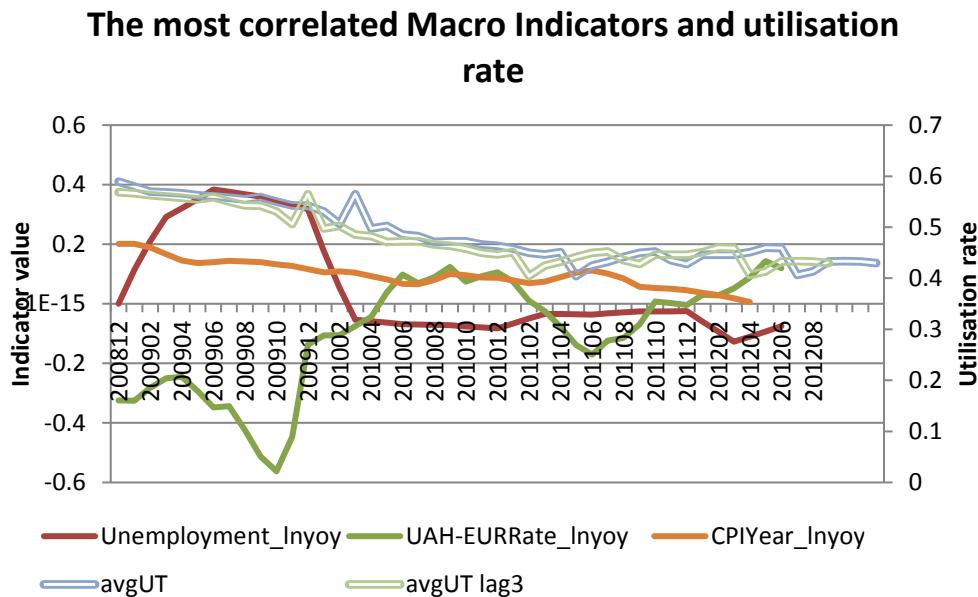
Figure 4.5 Macroeconomic indicators dynamics and the utilisation rate



The correlation between macro indicator differences, lagged 3 months, and the credit card average utilisation rate can be discovered visually. Figure 4.6 demonstrates the dynamics for 4 years period of the average utilisation rate – current and with a 3

months lag – and the following macroeconomic indicators: unemployment rate, the local currency exchange rate to EUR, and CPI.

Figure 4.6 Logarithms of some macroeconomic indicators and the utilisation rate dynamics



The period of 2009 – the first part of 2010 indicates dramatic changes in macroeconomic indexes after the recession. This single but not systematic action coupled with a change in internal banking strategies can have a significant impact on the general predictive model and may bias estimations for the stable period. Thus to avoid the impact of fluctuations associated with the crisis we consider only the period after July 2010 and before the maximum data sample date - December 2012.

Table 4.2 shows the Pearson correlation coefficients between utilisation rate value (UT), the utilisation rate changes with a 3 month lag (ln UT), logarithm of the utilisation rate changes (ln UT lag3) and some selected macroeconomic characteristics. The latter include unemployment rate, logarithm of the ratio of the unemployment rates current value and one year ago value (Unemployment_ln yoy), logarithm of local currency to Euro exchange rate changes for one year and one month (UAH-EURRate_Inyoy and UAH-EURRate_In mom respectively), logarithm of foreign direct investment year on year differences (FDI_Inyoy) and consumer price index changes for one year (CPIYear_Ln yoy). The most significant Pearson correlation has been found between utilisation rate with 3 months lag (UT lag3) and the following macroeconomic indicators: logarithm of the unemployment rate year on year

(Unemployment_Ln yoy), local currency to foreign currency exchange rate year on year (UAH-EURRate_Lnyoy) and CPI year on year changes (CPIYear_Lnyoy). A high positive correlation of 0.768 can be identified between the utilisation rate and the unemployed rate yearly changes. The 3-month lagged utilisation rate variable, lagged 3 months, has an even higher correlation (0.82) with the utilisation rate. The logarithm of CPI yearly changes index (CPIYear_Lnyoy) also has a Pearson correlation of 0.8, and this means that a general increase in the consumer prices is associated with more active credit card usage. However, despite the high correlation, it is necessary to note that this result is observed at pool level for average values and does not mean that utilisation rate of individuals will have a high correlation with the macro variables.

Table 4.2 Correlation between Macro Indicators and Utilisation Rate

	Unemployment Rate	Unemployment _Ln yoy	UAH- EURRate_Ln yoy	UAH- EURRate _Ln mom	CPIYear_ Ln yoy
UT	0.4049	0.7683	-0.7099	-0.2999	0.8016
In UT	-0.1134	-0.0409	0.0233	-0.0379	-0.053
UT lag3	0.3896	0.8227	-0.7640	-0.3017	0.8339
In UT lag3	-0.0498	0.0193	0.0351	-0.0085	0.1017

There are two options as to how to use the macroeconomic variables in the predictive model. The first option is to include them directly in the regression equations. For the panel data set macroeconomic variables may play the same role as a predictor such as account level behavioural variables, but having the same impact on all observations at each period. The second option allows that macroeconomic changes have a different impact on customer segments. In this case, the model might have two stages. The first stage is the pool level prediction for selected segments. The second stage is the account level prediction which uses an individual model for each segment. This approach uses a hypothesis that the same customer behaviour for the same macroeconomic conditions has a different outcome for different customer segments. We use the first type of the model in this research – macroeconomic characteristics are included in the list of all characteristics and used in the regression equation as other predictors.

We have a 30 month period of credit card portfolio behavioural characteristics. Macroeconomic data exists for a longer period but we can apply it for a 30 month period of portfolio data. Two and a half years is quite a short period for macroeconomic

model building, especially for analysis of the impact of macroeconomic cycles on customer behaviour. However, some changes in macroeconomic indicators might have a systematic impact on customer behaviour at the portfolio level, and we can consider them as covariates in the regression model.

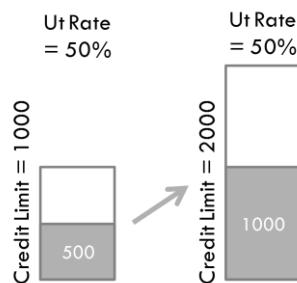
4.4 The utilisation rate modelling conception, cross-sectional analysis and model segments

4.4.1 The utilisation rate and credit limit

The credit limit utilisation rate depends on the four sets of factors: i) socio-demographic and financial customer characteristics from the application form, ii) customer behavioural characteristics such as debt service and credit usage, iii) credit product parameters such as credit limit, iv) macroeconomic factors such as unemployment rate, CPI, etc.

If the credit limit is a constant, the utilisation rate completely depends on the outstanding balance (Figure 4.7). However, in the case of an increase or decrease of the credit limit, the possible changes in the utilisation rate can be various. The credit limit depends on credit policy rules and is defined according to the customer risk profile. The same behavioural customer segments have various outstanding balances correlated particularly with the credit limit. Thus a customer segment does not have a typical outstanding balance, but it may have a typical utilisation rate as a proportion of the credit limit.

Figure 4.7 The constant utilisation rate for an increased credit limit

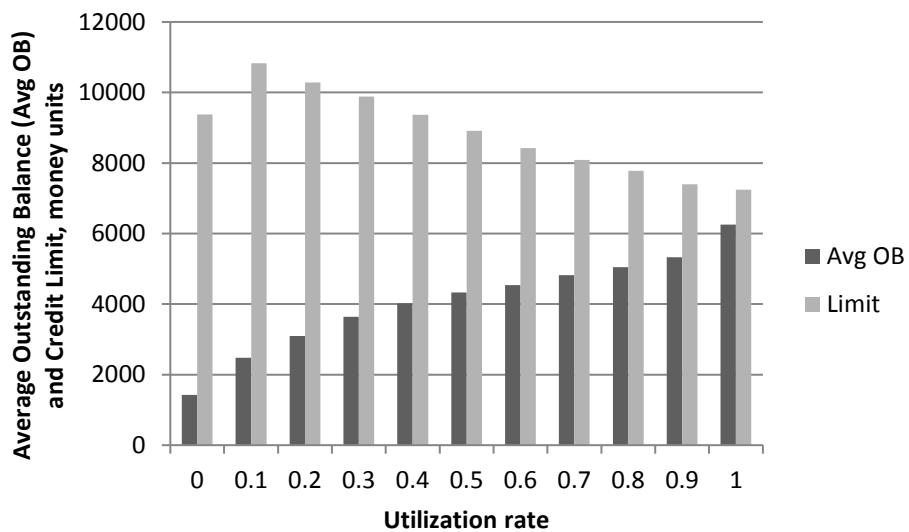


The prediction of utilisation rate instead of the outstanding balance amount is used to avoid possible changes in predicted behaviour caused by the bank credit policy and customer expectations. Firstly, different start terms such as a bank's credit policy and

changes in the product's parameter may affect the initial credit limit for the same category of customers. Secondly, credit card customer behaviour may be influenced mainly by the available balance as part of the credit limit and then by the amount of the available outstanding balance. The outstanding balance may be caused by the psychological factor such as 'I have to spend a share of the money, which are available for spending' or credit limit, and by the initial credit limit, which often depends on customer wealth and risk estimation and can be correlated with the client's appetite for spending. In this case, the prediction results will be explained mainly by the credit limit because the considered banking product is used to cover the short-term customer spending.

The utilisation rate values depend directly on the outstanding balance and inversely on credit limit as it can be seen in Figure 4.8. Thus it is necessary to be careful with interpretation and understanding of the utilisation rate because if one of the numerator or denominator is not a constant, the utilisation rate may have high values. For example both a low limit and low outstanding balance value and at the same time a high limit and high outstanding balance value could give the same utilisation rate.

Figure 4.8 The average outstanding balance and credit limit distribution by the utilisation rate



The utilisation rate has a direct positive correlation with the outstanding balance. Because the credit limit is the denominator of the utilisation rate ratio a credit limit decrease results in an increase in the utilisation rate. Thus the customer with the same behavioural pattern can have different utilisation rate when the outstanding balance is

the same because of the credit limit changes. The customer behaviour can be changed because of the limit changes. This is the reason why we split up the model into two segments for i) credits with unchanged limits and ii) credits with the limits which have been changed.

4.4.2 Segments of the model

The behavioural characteristics are based on a certain history horizon at the account level, for instance, the average outstanding balance for the last 6-month period, maximum debit turnover for the last 3 months, maximum delinquency for the whole period of the account history. It is possible to calculate such behavioural characteristics for the accounts with a sufficient period after activation and continuous credit limit usage. The number of Months on Book (MOB) should be no less than the required period for the characteristic, for example, for the Average Debit Turnover to the credit limit for the last 6 months the account must have MOB more or equal to 6. The application of behavioural characteristics over a long period gives a more complete picture of a customer's behaviour, but it is not possible to use them for a new portfolio. Customers at the first month on book have application characteristics only, but do not have behaviour for the current account at all. Thus for accounts with low MOB only behavioural factors over a short period can be used for modelling. To avoid the use of null values for the factors which are impossible to calculate properly in the single model, a separate model for the short period will be used.

Moreover, accounts at the early stages have generally different behavioural patterns. For instance, Figure 4.9 shows rapid growth of the utilisation rate during the first three months after credit card activation, continuous damped growth until the sixth month with the maximum plateau in months 7-9, followed by a slight decrease in the utilisation with convergence to around 50%. It can be seen that the general behavioural patterns during the initial period of card usage are different from the mature period, and the trends of the dependent variable and behavioural factors at the early and late stages can be different. Thus because of the difference in the behavioural characteristics calculation and the differences in the utilisation rate behaviour at the early and late stages of credit history it is rational to allocate a separate model for the

low MOB period compared with a long MOB period. In our case two periods have been chosen: MOB from 1 to 5 month and MOB more than 6 months.

Figure 4.9 Average Utilisation rate by Month on Book

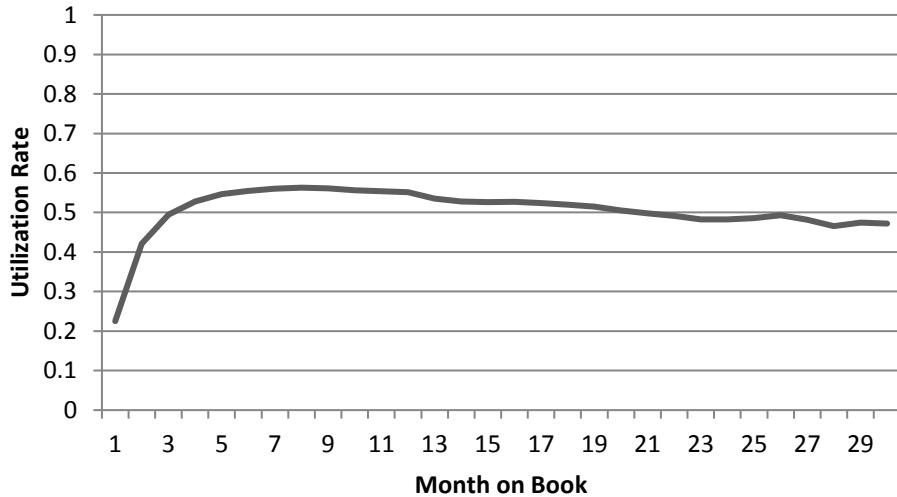


Table 4.3 demonstrates the definition of three segments of the models with application, with behavioural variables without changes in the limit, and with behavioural variables with changes in the limit. The application model segment is defined as a segment for accounts with months on balance less than five. For accounts with MOB equal to 6 or more at the observation point, the limit changes feature is used for the segment definition.

Table 4.3 The definition of 3 types of models depending on MOB

1	2	3	4	5	6	7	8	9	10	11	12	13	...	24	25	26	27	28	29	30		
Model APP: MOB 1-5					Model BEH NL: MOB 6+ and Limit NO Change															Performance		
					Model BEH CL: MOB 6+ and Limit Changed																	

For the first five months on the book, the model called the Application model (Model APP) is applied. This model contains application and short-term behavioural characteristics. Two long-term behavioural models with a non-changed limit (BEH NL) and secondly with a changed limit (BEH CL) are applied for the next 18 months (from month 6 to month 24). This period from 6th to 24th month on book is used as a performance window for the development and validation with appropriate lag for any previous observation window. For example, a loan activated in July 2011 has 6 months

of history from July till December and a 6 month performance period from January to June 2012.

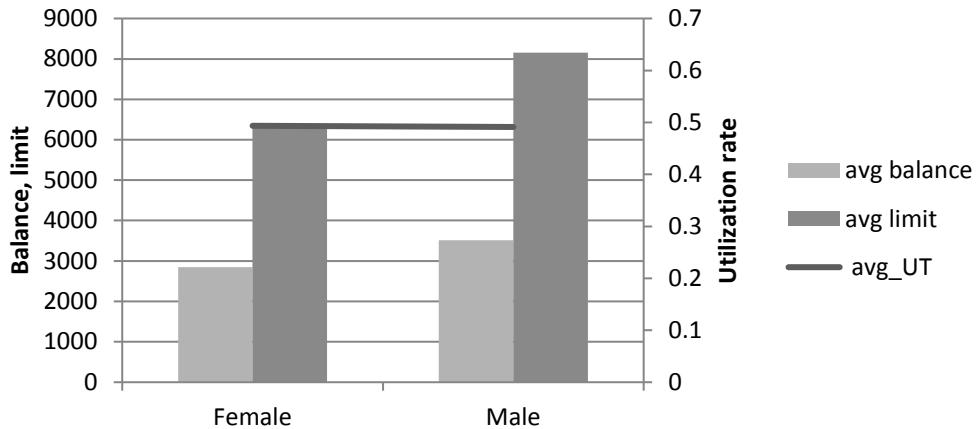
Thus we have three segments for modelling: MOB 6 and more with no limit change, MOB 6 and more with limit change and MOB 1-5. The difference between the segments is in the sets of explanatory variables and observation points of the accounts, but the regression techniques are the same for all segments.

4.4.3 Cross-sectional analysis of the utilisation rate: characteristics

We investigated a set of application characteristics as predictors for the utilisation rate outcome. The key characteristics are presented in details and more are included in the regression coefficients estimation. We compare the average outstanding balance (avg_balance), credit limit (avg_limit), and the utilisation rate (avg_UT) distribution by predictive characteristics. We use the visual analysis of whether the average utilisation rate values for characteristic bands differ significantly, or if there are any tendencies in the utilisation rate. If the utilisation rate values are approximately equal (the difference is less than 5%), the characteristic is excluded from the set of covariates at the preliminary stage of analysis as insignificant. However, we especially tried to include into the model as much characteristics as possible to see the signs and significance of the estimates. The insignificant characteristics can be identified with high p-level values (more than 0.05).

Some factors are not predictive, such as a gender. The utilisation rate for gender does not differ significantly, but the average balance for a male is higher than for female, most likely because of higher limits (see Figure 4.10).

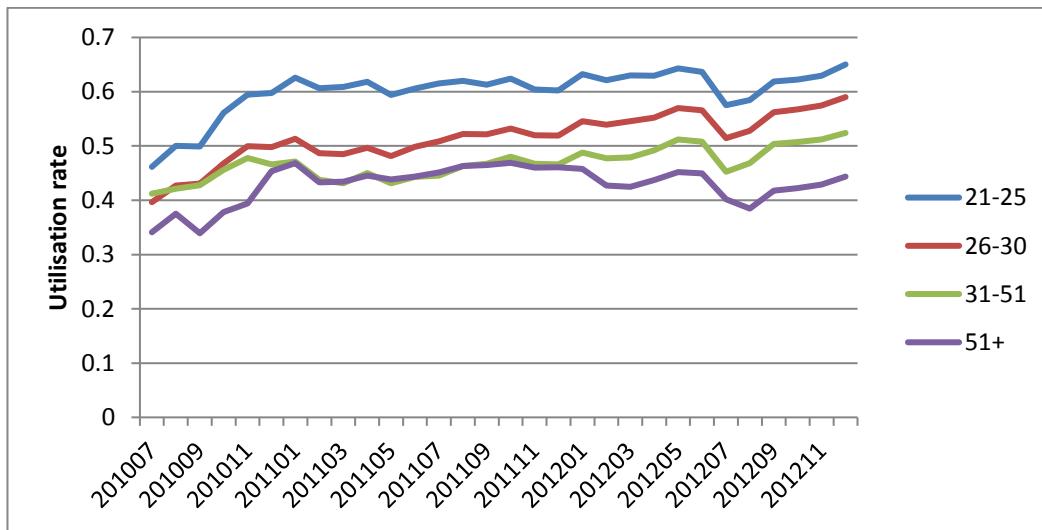
Figure 4.10 The utilisation rate (avg_UT), average balance, and average credit limit by gender



However, the cross-sectional analysis does not take into account a time component and possible changes in the relation between regressors and outcome, but uses average values (means) if usage differs over time.

The utilisation rate for age has time variation (see Figure 4.11). Under the assumption that limit increases have been made for all age groups in the equal proportion the time variation can be explained by the fact that the different age groups have an individual reaction to macroeconomic changes. Young people have a higher utilisation rate than older people mainly because of low credit limits.

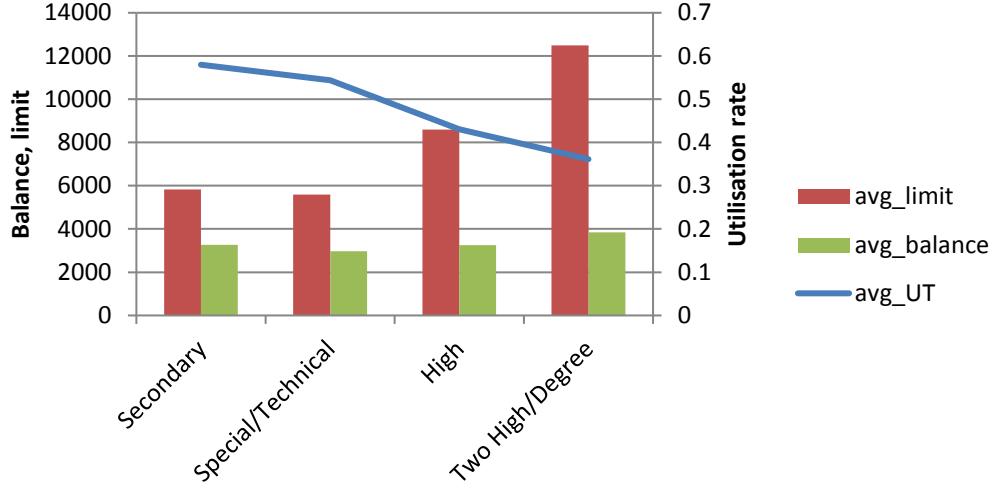
Figure 4.11 The utilisation rate by age group - time variation



The cross-sectional analysis of education demonstrates that different education categories have various utilisation rates. Customers with secondary education have the highest utilisation rate of around 60%, while two degree and doctorate customers have

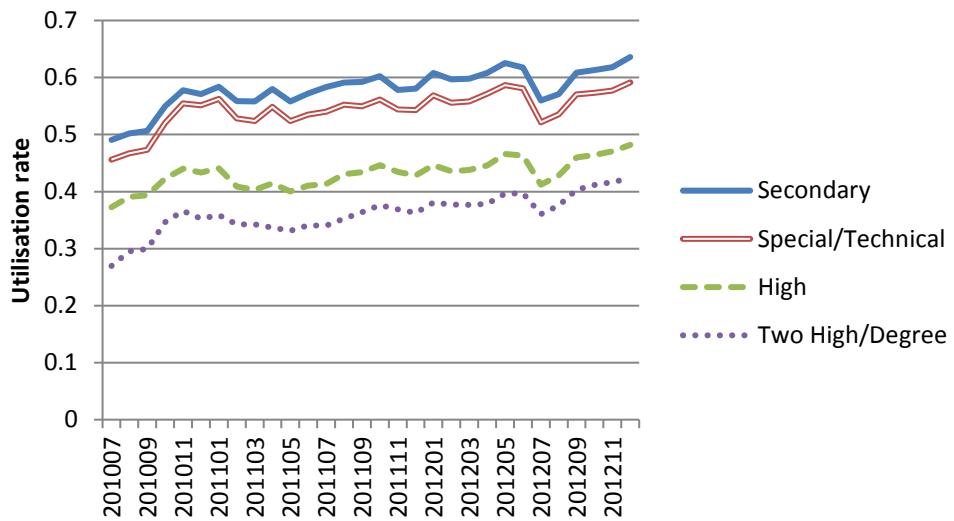
only 35% credit limit utilisation. However, the outstanding balances do not differ significantly, and the higher credit limits for customers explain the utilisation rate variation with Higher and Two high/Doctorate education. The analysis of the utilisation rate by education over time demonstrates simultaneous changes in the utilisation rate for all type of education. Thus the education covariate is stable over time.

Figure 4.12 The utilisation rate, average balance, and credit limit by education



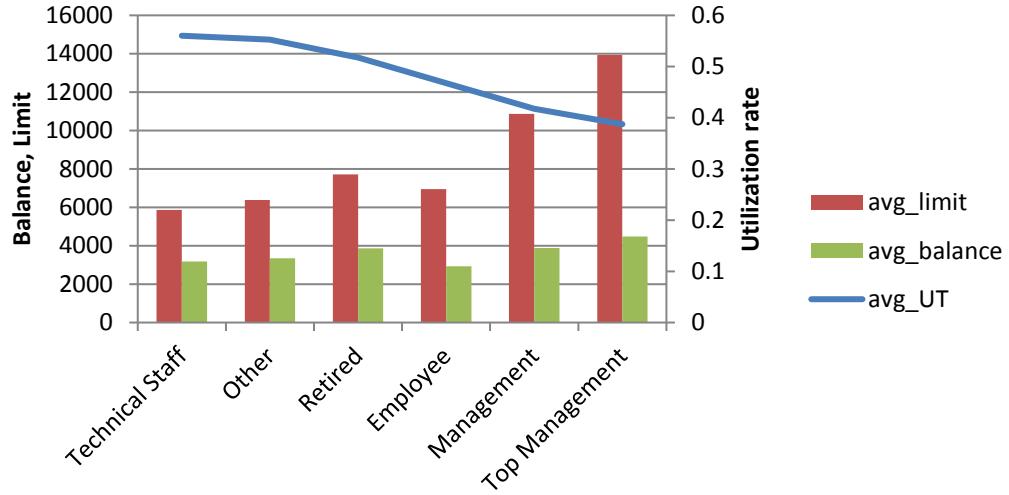
There is no time variation for education factor in the sense that the utilisation rate for all education types changes simultaneously (see Figure 4.13). So the education is a stable parameter and all education groups have the same average utilisation rate changes.

Figure 4.13 The utilisation rate by education - time variation



Employment status also has little variation over time for any specific borrower. Education and employment status can be significant application characteristics for utilisation rate prediction. For example, between positions the utilisation rate varies from 0.6 for technical staff to 0.4 for top management (see Figure 4.14).

Figure 4.14 The utilisation rate, average balance, and credit limit by employment status



Thus, cross-sectional analysis has demonstrated that the utilisation rate varies depending on the characteristic values, so can be discriminated and used in the predictive model.

We expect the following effect from the inclusion of the variables as a predictor into the model (Table 4.4).

Table 4.4 Expected effects of the variables for inclusion in a model

Variable name	Description	The reason for the inclusion in a model
Mob	Month on Book	We expect a sharp increase in utilisation rate for several months after the activation of the credit card and a slight decrease in utilisation rate after this period
Limit	Credit limit at the observation point	The higher credit limit might slightly decrease the utilisation rate, but generally should be an insignificant parameter
UT	Utilisation rate at the observation point	Utilisation rate might stay at the same level, decrease or increase depending on the behavioural type of the client. For a customer who actively use a credit limit, we expect the same level of utilisation rate, for an inactive customer or who has become an inactive recently - an increase in utilisation rate, for a delinquent - a decrease or the same level in case of transition to default state
b_AvgOB16_to_MaxOB16_In	The logarithm of average OB to maximum OB for the last six months	If the average outstanding balance for a period is less than maximum one for the same period, this means some variation in the outstanding balance for this period. If the average outstanding balance close to maximum one, this means that the outstanding balance is stable and indicator value is close to 1. We expect higher utilisation for stable outstanding balance.
b_TRmax_deb16_To_Limit_In	The logarithm of the maximum debit transaction (purchases) for the last six months to the credit limit	The higher value of the ration of the maximum purchase amount for six months period to credit limit might cause a higher utilisation rate in future because a customer tends to spend
b_TRavg_deb16_to_avgOB16_In	The logarithm of the average debit transaction to average OB for the last six months	The higher average purchase amount for a period to average outstanding balance might cause lower utilisation rate in future because a customer might tend to spending, but also he or she pays back significant debt amount and decrease the outstanding balance and, consequently, the utilisation rate value.
b_TRsum_deb16_to_TRsum_cr16_In	The logarithm of sum debit transaction to the sum of credit transactions (payments) for the last six months	The higher value of the ratio of the average purchase amount for six months period to average payment amount might cause higher utilisation rate in future because if a customer spends more than pays back, the outstanding balance and utilisation rate will grow
b_UT1_to_AvgU16In	The logarithm of the utilisation rate at the observation point divided by the average utilisation rate for the last six months	In case of decrease of current utilisation rate in comparison with previous average utilisation rate for six months, utilisation rate in the next period might also decrease, and increase in the opposite case.
b_UT1to2In	The logarithm of the utilisation rate at observation point to the previous months	In case of decrease of current utilisation rate in comparison with the previous month, utilisation rate in the next period might also decrease, and increase in the opposite case.
b_UT1to6In	The logarithm of the utilisation rate at observation point to the	In case of decrease of current utilisation rate in comparison with utilisation rate six months ago,

Table 4.4 Expected effects of the variables for inclusion in a model

Variable name	Description	The reason for the inclusion in a model
	utilisation rate six months before	utilisation rate in the next period might also decrease, and increase in the opposite case.
b_NumDeb13to46ln	The logarithm of the number of debit transactions for the last three months to the number of debit transactions for months 4-6 before the observation point	The increase of the number of spending transactions for the last three months in comparison with the number of spending transactions for 4-6 months before the observation point might increase the utilisation rate because of more active spending transactions.
b_inactive13	Binary indicator whether the account was inactive last three months	If an account was inactive for the last three month, we expect a low utilisation rate
b_avgNumDeb16	Average number of debit transactions for the last six months	It is expected that the higher number of spending transactions causes higher utilisation rate
b_OB_avg_to_eop1ln	The logarithm of the average outstanding balance for the last month to the outstanding balance at the end of the period at the observation point	If the outstanding balance at the end of the month is higher than average balance for this month, we expect the increase in the balance in the next month
b_DelBucket16	Maximum bucket of delinquency for the last six months	Higher delinquency buckets might cause a higher utilisation rate because customers usually go to default with higher outstanding balance than non-delinquent balance
b_pos_flag_0	A binary indicator of Point-of-sales (POS) transaction at the observation point month	POS transactions increase the utilisation rate
b_pos_flag_13	A binary indicator of Point-of-sales (POS) transaction for the last six month	POS transactions increase the utilisation rate
b_atm_flag_0	A binary indicator of ATM cash withdrawals at the observation point month	ATM transactions increase the utilisation rate
b_atm_flag_13	A binary indicator of ATM cash withdrawals for the last six month	ATM transactions increase the utilisation rate
b_pos_flag_use d46vs13	A binary indicator of POS transactions 4-6 months before and no transactions for the last three month	Non-use the credit card for POS transactions after a period of usage might decrease the credit limit utilisation rate
b_pos_flag_use 13vs46	A binary indicator of POS transactions for the last three month and no transactions for 4-6 months before	The start of use the credit card for POS transactions might increase the credit limit utilisation rate
b_atm_flag_use d46vs13	A binary indicator of ATM transactions 4-6 months before and no transactions for the last three month	Non-use the credit card for ATM transactions after a period of usage might decrease the credit limit utilisation rate

Table 4.4 Expected effects of the variables for inclusion in a model

Variable name	Description	The reason for the inclusion in a model
b_atm_flag_use_13vs46	A binary indicator of ATM transactions for the last three month and no transactions for 4-6 months before	The start of use the credit card for ATM transactions might increase the credit limit utilisation rate
b_pos_use_only_flag_13	A binary indicator of POS transactions only for the last three month	If a customer uses a credit card for POS transactions only, this might increase the utilisation rate in comparison with the inactive state, but unclear higher or lower in comparison with ATM transactions
no_dpd	A binary indicator of no delinquency at the observation point	An impact of the non-delinquency state on the utilisation rate is questionable. Generally, delinquent accounts might have a higher utilisation rate.
max_dpd_60	Binary indicator if the Maximum number of Days Past Due was 60 or more for life-time	An impact of previous default state on the utilisation rate is questionable. Generally, delinquent accounts might have a higher utilisation rate.
AgeGRP1	Age less than 25	We expect that young customers have a higher utilisation rate because of low limits and active credit usage
AgeGRP3	Age more than 35	We expect that older customers have lower utilisation rate because of high limits and nonactive credit usage
customer_income_ln	The logarithm of the ratio of the customer monthly income to the average monthly income in a portfolio	We expect that customers with higher income have a lower utilisation rate because of high limits. However, active usage of credit card is questionable.
Edu_	Education	We expect that customers with special education have a higher utilisation rate than customers with higher education because of low limits and need for credit
Marital_	Marital status	We expect that married customers might have a lower utilisation rate than other categories
position_	Employment status	We expect that customers with technical staff position have higher utilisation rate because of low limits and need for credit, and top managers can use a credit card more actively, but they have low utilisation rate because of high credit limits
sec_	Sector of Industry	We expect that customers with agriculture and construction industry have higher utilisation rate because of low limits and need for credit
Car _	Car owner	Car owners usually spend more money, but they probably have higher credit limits than a customer without a car. So we expect a lower utilisation rate for this category. Car co-owners utilisation rate is questionable and might be similar to car owner category.
real_	Real estate	Real Estate owners probably spend more money, but they probably have higher credit limits than a customer without Real Estate. So we expect a

Table 4.4 Expected effects of the variables for inclusion in a model

Variable name	Description	The reason for the inclusion in a model
reg_ctr_	Region of living	lower utilisation rate for this category. Real Estate co-owners utilisation rate is questionable and might be similar to car owner category.
child_	Number of children	We expect that customers from capital and region centres use credit cards more actively than from province. We expect that customers from province might have high utilisation rates because of low credit limits
Unempl_Inoy	The logarithm of the unemployment rate change year on year	We expect that customers without children have a lower utilisation rate than customers who have children. We expect that customers with many children have a high utilisation rate because of the need for credit
UAH_EURRate_l nmom	The logarithm of the exchange rate of local currency to Euro at the observation point in comparison with the previous month	We expect that the macro unemployment rate will be positively correlated with utilisation rate because a higher level of unemployment might cause the increase in demand for money.
UAH_EURRate_l nyoy	The logarithm of the exchange rate of local currency to Euro at the observation point in comparison with the same period of the previous year	The increase in local currency to Euro exchange rate (LCY/EUR exchange rate) might cause purchasing power loss because of growth of prices: directly for export goods and indirectly for local goods in proportion to the export component. Thus, the utilisation rate might have a positive correlation to the foreign currency exchange rate.
CPI_Inqoq	The logarithm of the current Consumer Price Index at the observation point to the previous quarter CPI	The increase in local currency to Euro exchange rate (LCY/EUR exchange rate) might cause purchasing power loss because of growth of prices: directly for export goods and indirectly for local goods in proportion to the export component. Thus, the utilisation rate might have a positive correlation to the foreign currency exchange rate.
SalaryYear_Inyo y	The logarithm of the Average Salary at the observation point in comparison with the same period of the previous year	The increase in Consumer Price Index means the rise in average prices of a basket of consumer goods and services. Thus, the usage of credit limit can also be increased in this case.
I_ch1_In	Limit change month ago	The increase in Salary might decrease the utilisation rate expected value because of more available money. However, the customers appetite for spending might also increase so that the utilisation rate might stay the same as before the increase of the salary. On the other hand, we expect that a decrease in salary will cause an increase in utilisation rate.
I_ch6_In	limit change six months ago	The increase of credit limit decrease the utilisation rate in a short period, but it might recover to the same level as it was before
		The increase of credit limit decrease the utilisation rate in a short period, but it might recover to the same level as it was before

4.5 Model estimation and results

A one stage model means one equation to direct estimate the proportion is applied. In section 4.5.1 we consider general results of the utilisation models application. In the section 4.5.2 we describe and discuss estimation results for five models for three segments – MOB 6 and more with no limit change, MOB 6 and more with limit change and MOB 1-5. Section 4.5.3 contains the summary and comparative analysis of the models using the validation results. We investigate five direct estimation techniques such as ordinary linear regression, beta regression, beta transformation plus general linear models (GLM), fractional regression (quasi-likelihood estimation), and weighted logistic regression for binary transformed data.

4.5.1 Parameter estimates for One-stage model

In our model the credit limit utilisation rate depends on four sets of factors: i) socio-demographic and financial customer characteristics from the application form, ii) customer behavioural characteristics such as debt service and credit usage, iii) credit product parameters such as credit limit, iv) macroeconomic parameters such as unemployment rate, CPI, etc.

We use a maximum performance window of six months for utilisation rate prediction for the following reasons: i) the period of investigation available from the given data set is two and a half years, ii) the behavioural characteristics are calculated for a maximum six-month period, and iii) we use the business logic to build relatively short-term, but accurate predictive models. The prediction of the utilisation rate for short periods such as one month, three months, and six months is justifies because it may help to built strategy for immediate actions, for example, to reduce the credit limit to decrease the exposure at default for high-risk clients, or to increase the credit limit to motivate a customer to engage in more spending transactions and to avoid clients' attrition.

We applied five methods for direct proportion prediction for three segments of models.

We split up the common model into three segments: MOB 1-5 (APP), MOB 6+ non-changed limit (BEH NL) and MOB6+ changed limit (BEH CL). Current distributions present BEH NL model. Five methods give distributions for APP and BEH CL with

similar shapes and parameters. Thus they can be applied for all months on book and the segment with changes in credit limit.

We selected for regression analysis accounts with at least 12 months history. Two segments: Limit No Change (BEH NL) and Limit Changed (BEH CL) contain the observations for accounts at sixth or more Months on Book. An account goes to the BEH NL segment when no credit limit changes were applied for a whole available period of investigation. We have 106,158 observations for accounts in Limit No Change segment. An account goes to the BEH CL segment if no credit limit changes were applied at any point of the period of investigation. Totally we have 96,488 observations for accounts in Limit Changed segment. Application (APP) segment contains totally 14, 374 observations for the same accounts, but at the 5th Months on Book. This period is not enough to create behavioural characteristics, and we use mainly application characteristics to predict the utilisation rate. We split the total population of accounts into train and test samples in proportion 70/30 (Table 4.5).

Table 4.5 The number of observations and average utilisation rate values (target) for three segments: No Changed Limit, Changed Limit, and Application (MOB 1-5)

Segment	Train	Test	Total	Target_UT							
				plus1M	plus2M	plus3M	plus4M	plus5M	plus6M	6 months	
Limit No Change	BEH NL	74,311	31,847	106,158	0.53	0.53	0.52	0.52	0.51	0.51	0.52
Limit Changed	BEH CL	67,542	28,946	96,488	0.64	0.63	0.62	0.61	0.60	0.60	0.62
Application	APP	10,062	4,312	14,374	0.24	0.48	0.56	0.59	0.61	0.61	0.52
Total		151,914	65,106	217,020							

The modelling segments have different average target values of utilisation rate as shown in Table 4.5. Average target utilisation rate for a 6-month period is 0.52 for BEH NL, 0.62 for BEH CL, and 0.52 for APP segments. So, if to predict average values of the utilisation rate for an account at 5th Month on Book and to the same account after 6 MOB, the average 6-month utilisation rate should equal. However, for the segment with an increased credit limit (we do not have cases in the data sample with credit limit decrease) the target utilisation rate for 6 months is higher by 0.1. The target utilisation rate for the 1-month period is almost the same as the utilisation rate for a 6-month period for the BEH NL and BEH CL segments – 0.53 and 0.64 respectively. However, the target value of utilisation rate for 1-month period for the APP segment is significantly lower and at 0.24, but it decreases to 0.61 for utilisation rate target value at the 6th month. This corresponds to Figure 4.9, which show the

utilisation rate increase to 12 Month on Book and slight steady decrease after the 13th Month on Book.

Models for APP, BEH NL and BEH CL used different sets of characteristics. The application parameters are used for all segments, but long-term behavioural predictors are not used for APP, and changes in limit are used for BEH CL only. We provide OLS method estimations here for the comparative analysis of the model segments in Table 4.6.

The same characteristics have different estimated parameters and t-values for APP, BEH NL and BEH CL. For example, the utilisation rate at the observation point is less significant for an account with changes in credit limit (BEH CL) than for an account without changes in credit limit (BEH NL) – t-values 104 versus 170 – but the current utilisation rate predictor exceeds the significance of other characteristics. MOB is less significant for the segment with changes in credit limit, than for other loans. MOB has a negative trend and can be a valuable parameter for the first months of the life-cycle.

Credit limit changes (*l_ch6_ln*) have an impact on the utilisation rate. The parameter estimation equal to 0.096 for Logarithm of credit limit change six months ago is significant and means that increase in the credit limit two times increases the utilisation rate by 6.2% ($\ln(2)*0.096$). So, the limit increase generally motivates credit cardholders to spend money, but not too much.

The application segment means that an account contains less than 6 months of history. The usage of behavioural characteristics is limited, but possible for characteristics, which i) include the period of observation of less than 6 months, for example, change of the utilisation rate for the previous month or current state of the account, and ii) include a 6 month period, but can be calculated without mistakes in logic and their values make sense for shorter periods, for example, the logarithm of maximum spending transaction amount for the last 6 months to the average credit limit (*b_TRmax_deb16_To_Limit_ln*) or the logarithm of ratio of the sum of spending transactions to the sum of debt repayment transactions (*b_TRsum_deb16_to_TRsum_crd16_ln*).

The use of the ratio of the current utilisation rate to the utilisation rate six months ago (*b_UT1to6ln*) for a period of less than 6 months does not make sense. The ratio of

average outstanding balance to maximum outstanding balance for the last 6 months (*b_AvgOB16_to_MaxOB16_ln*) is excluded from the set of APP model covariates, because of a low variance of values – at the initial stage average outstanding balance is close to the maximum of one. However, this covariate is significant for behavioural segments and has positive signs around 0.04 for BEH NL and BEH CL – the average outstanding balance close to the maximum outstanding balance results in higher utilisation in the next 6 months.

Macroeconomic characteristics reflect the systematic factor because they affect all accounts equally at any time point. Because of a short period of observation (two and half years) we were not able to build a macroeconomic model to predict changes in the utilisation rate caused by economic cycles. Instead of this we included some macroeconomic indicators, which reflect the changes in the economic environment and slightly shift the utilisation rate for all population depending on the time point. Changes in the unemployment level and CPI have a positive, but non-significant impact on the utilisation rate. The change of the logarithm of ratio of unemployment rate at the observation point to the unemployment rate one year ago by 1 unit (or the change in this ratio 2.72 times) will result the change of the utilisation rate by 0.26, 0.52, and 0.14 for BEH NL, BEH CL, and APP segments respectively. The yearly change of CPI has larger effect and has opposite signs for behavioural and application models – 0.62, 1.46, and -0.6 for BEH NL, BEH CL, and APP segments respectively. This means that after 6th month on book CPI has an expected positive correlation with utilisation rate – an increase in CPI corresponds to purchasing power loss and this may cause an increase of consumer demand for borrowed fund. However, in the first 5 months after credit card activation an increase in the credit limit utilisation rate is observed for the period of the increase in customers purchasing power.

Table 4.6 Comparative analysis of OLS parameters estimation for three segments

Characteristic	MOB 6+ - Limit NO Change					MOB 6+ - Limit Changed					MOB 1-5					
	Parameter Estimate	Standard error	t Value	Pr > t	Parameter Estimate	Standard error	t Value	Pr > t	Parameter Estimate	Standard error	t Value	Pr > t	Parameter Estimate	Standard error	t Value	Pr > t
Intercept	0.19868	0.00823	24.14	<.0001	0.14837	0.01252	11.85	<.0001	0.2773	0.0156	17.78	<.0001				
<i>Account info</i>																
mob	-0.00328	0.0001604	-20.44	<.0001	-0.00188	0.0002648	-7.09	<.0001								
limit	1.59E-07	1.33E-07	1.19	0.2345	-2.7E-06	2.31E-07	-11.67	<.0001	-2.4E-06	2.93E-07	-8.2	<.0001				
UT	0.53061	0.00312	170.19	<.0001	0.51333	0.00492	104.28	<.0001	0.43759	0.00581	75.3	<.0001				
avg_balance	2.07E-06	2.37E-07	8.73	<.0001	3.84E-06	3.57E-07	10.74	<.0001	0.0000032	5.61E-07	5.7	<.0001				
<i>Behavioural - dynamic</i>																
b_AvgOB16_to_MaxOB16_In	0.04088	0.00134	30.59	<.0001	0.04039	0.00233	17.34	<.0001								
b_Trmax_deb16_To_Limit_In	0.00699	0.00049122	14.22	<.0001	0.01649	0.0008275	19.93	<.0001	-0.00552	0.00045667	-12.09	<.0001				
b_Travg_deb16_to_avgOB16_In	-0.01841	0.00068774	-26.76	<.0001	-0.03006	0.00125	-24.08	<.0001	-0.01085	0.00105	-10.31	<.0001				
b_Tsum_deb16_to_Tsum_crd16_In	0.01087	0.00061894	17.55	<.0001	0.01241	0.00119	10.46	<.0001	0.02698	0.00074873	36.04	<.0001				
b_UT1_to_AvgUT16ln	-0.00282	0.00040175	-7.01	<.0001	-0.00195	0.0006926	-2.82	0.0048	0.00912	0.00068481	13.32	<.0001				
b_UT1to2ln	0.00178	0.00030936	5.76	<.0001	0.0009643	0.0005288	1.82	0.0682	-0.00734	0.00021089	-34.79	<.0001				
b_UT1to6ln	-0.0048	0.00025297	-18.99	<.0001	-0.00647	0.0004217	-15.34	<.0001								
b_NumDeb13to46ln	0.00545	0.00033788	16.14	<.0001	0.00937	0.0006162	15.21	<.0001								
b_inactive13	0.0824	0.000464	17.77	<.0001	0.16916	0.00812	20.84	<.0001								
b_avgNumDeb16	0.0001071	0.00004173	2.57	0.0103	-0.00149	0.0003031	-4.91	<.0001	0.00679	0.00034267	19.83	<.0001				
b_OB_avg_to_eop1ln	-0.00132	0.00033009	-4	<.0001	-0.000629	0.00052	-1.21	0.2262	-0.00927	0.00064119	-14.45	<.0001				
b_DelBucket16	0.02571	0.00235	10.96	<.0001	0.03251	0.00441	7.36	<.0001	0.02082	0.00821	2.54	0.0112				
b_pos_flag_0	0.01847	0.00174	10.63	<.0001	0.01697	0.00244	6.94	<.0001	0.021	0.00263	7.99	<.0001				
b_pos_flag_13	0.03935	0.00213	18.51	<.0001	0.04299	0.003	14.32	<.0001								
b_atm_flag_0	0.053	0.00152	34.91	<.0001	0.05427	0.00213	25.44	<.0001	0.09135	0.00279	32.7	<.0001				
b_atm_flag_13	0.05824	0.00246	23.7	<.0001	0.03996	0.00355	11.25	<.0001								
b_pos_flag_used46vs13	0.02783	0.00019	14.64	<.0001	0.02612	0.0027	9.68	<.0001								
b_pos_flag_use13vs46	-0.02589	0.00203	-12.78	<.0001	-0.03041	0.00286	-10.61	<.0001								
b_atm_flag_used46vs13	0.01634	0.0022	7.43	<.0001	0.00351	0.00329	1.07	0.2861								
b_atm_flag_use13vs46	-0.02562	0.00214	-11.96	<.0001	-0.01679	0.00306	-5.48	<.0001								
b_pos_use_only_flag_13	0.01185	0.00275	4.31	<.0001	0.00375	0.00397	0.95	0.3441	-0.03498	0.0047	-7.45	<.0001				
no_dpd	-0.00524	0.00413	-1.27	0.2039	0.00506	0.00694	0.73	0.4657								
max_dpd_60	0.02693	0.00698	3.86	0.0001	0.03216	0.01178	2.73	0.0063								
<i>Application- static</i>																
AgeGRP1	0.01494	0.00189	7.9	<.0001	0.01191	0.0025	4.77	<.0001	0.02992	0.00353	8.47	<.0001				
AgeGRP3	-0.00166	0.00172	-0.97	0.3323	0.00783	0.00234	3.34	0.0008	-0.01675	0.0032	-5.23	<.0001				
customer_income_In	-0.03324	0.00165	-20.1	<.0001	-0.02355	0.00258	-9.11	<.0001	-0.04541	0.00327	-13.87	<.0001				
Edu_High	-0.02094	0.00175	-11.94	<.0001	-0.0196	0.00249	-7.86	<.0001	-0.04343	0.00326	-13.33	<.0001				
Edu_Special	-0.00198	0.00169	-1.17	0.2415	-0.00226	0.00243	-0.93	0.3518	-0.01095	0.00314	-3.48	0.0005				
Edu_TwoDegree	-0.01803	0.00389	-4.63	<.0001	-0.01047	0.00539	-1.94	0.052	-0.0494	0.00731	-6.76	<.0001				
Marital_Civ	0.00494	0.00268	1.84	0.0653	0.00403	0.0038	1.06	0.2898	0.03416	0.00501	6.82	<.0001				
Marital_Div	0.00377	0.0019	1.98	0.0478	0.00829	0.00279	2.97	0.003	0.02359	0.00352	6.71	<.0001				
Marital_Sin	0.00625	0.00201	3.11	0.0019	0.00631	0.00282	2.24	0.0253	0.019	0.00376	5.06	<.0001				
Marital_Wid	0.02133	0.00349	6.12	<.0001	0.03034	0.00613	4.95	<.0001	0.03118	0.00642	4.85	<.0001				
position_Man	0.01174	0.00183	6.43	<.0001	0.01479	0.00272	5.45	<.0001	0.01727	0.00342	5.06	<.0001				
position_Oth	0.00891	0.00181	4.91	<.0001	0.00718	0.00257	2.79	0.0052	0.00821	0.00337	2.43	0.015				
position_Tech	0.00673	0.0017	3.96	<.0001	0.01045	0.00245	4.26	<.0001	0.01417	0.00315	4.49	<.0001				
position_Top	0.00989	0.00326	2.94	0.0033	0.01065	0.00568	1.87	0.0611	0.00769	0.00629	1.22	0.2211				
sec_Agricult	0.00428	0.00327	1.31	0.1905	0.01956	0.00485	4.03	<.0001	-0.0196	0.00608	-3.22	0.0013				
sec_Constr	-0.00428	0.00437	-0.98	0.3279	-0.0252	0.00651	-3.87	0.0001	-0.01599	0.00802	-2	0.046				
sec_Energy	-0.00213	0.00281	-0.76	0.4483	-0.00362	0.00401	-0.9	0.3664	-0.00917	0.00522	-1.76	0.0792				
sec_Fin	-0.02624	0.00211	-12.45	<.0001	-0.03388	0.00316	-10.72	<.0001	-0.04477	0.00388	-11.55	<.0001				
sec_Industry	0.00344	0.00552	0.62	0.5328	0.00516	0.008	0.64	0.5195	-0.000521	0.01043	-0.05	0.9602				
sec_Manufact	0.00805	0.00443	1.87	0.0613	0.00217	0.00622	0.35	0.7266	-0.02594	0.00785	-3.3	0.001				
sec_Mining	0.00362	0.00299	1.21	0.2259	0.00344	0.00436	0.79	0.4303	0.01126	0.00563	2	0.0456				
sec_Service	-0.00867	0.00158	-5.49	<.0001	-0.01918	0.00232	-8.27	<.0001	-0.01405	0.00296	-4.75	<.0001				
sec_Trade	-0.00604	0.00212	-2.84	0.0045	-0.00638	0.003	-2.13	0.0332	-0.00529	0.00395	-1.34	0.18				
sec_Trans	-0.00676	0.00414	-1.63	0.1026	-0.01358	0.00608	-2.23	0.0256	-0.01666	0.00761	-2.19	0.0286				
car_Own	-0.01099	0.0015	-7.34	<.0001	-0.00857	0.00219	-3.91	<.0001	-0.01934	0.00279	-6.93	<.0001				
car_coOwn	0.00212	0.00228	0.93	0.3517	0.00259	0.00332	0.78	0.4353	-0.00269	0.00425	-0.63	0.5277				
real_Own	0.0005113	0.00145	0.35	0.7248	0.00244	0.00205	1.19	0.2341	-0.00849	0.00269	-3.15	0.0016				
real_coOwn	-0.00257	0.00154	-1.68	0.0939	-0.00185	0.00214	-0.87	0.3866	0.00291	0.00286	1.02	0.3097				
reg_ctr_Y	-0.00803	0.00223	-3.6	0.0003	-0.01008	0.00316	-3.19	0.0014	-0.00798	0.00256	-3.12	0.0018				
reg_ctr_N	-0.00653	0.00224	-2.92	0.0035	-0.00913	0.00318	-2.87	0.0041	0.00625	0.00419	1.49	0.1358				
child_1	0.01064	0.0019	5.61	<.0001	0.01041	0.00266	3.91	<.0001	0.0208	0.00356	5.84	<.0001				
child_2	0.0065	0.00108	6.02	<.0001	0.00441	0.00157	2.8	0.0051	0.01498	0.00203	7.39	<.0001				
child_3	0.03175	0.00362	8.78	<.0001	0.02034	0.00564	3.61	0.0003	0.03036	0.00672	4.52	<.0001				
<i>Macroeconomic - dynamic</i>																
Unempl_Inyoy_6	0.26643	0.02401	11.1	<.0001	0.52425	0.03853	13.61	<.0001	0.57361	0.14279	4.02	<.0001				
UAH_EURRate_Inmom_6	-0.07515	0.02953	-2.54	0.0109	-0.09468	0.04392	-2.16	0.0311								

In Table 4.6, almost all characteristics, included in the regression model for testing their signs and significance, have a p-level value close to zero. This means that these characteristics can be kept in the model and they do not deteriorate the model accuracy because of high intercorrelation. Only behavioural characteristic No DPD with p-levels equal to 0.2 and 0.46 for BEH NL and BEH CL, some application dummy characteristics such as ‘Real estate owner’ and ‘Car co-owner’ with p-level around 0.3 and 0.56 respectively, and macroeconomic variable ‘Logarithm of Yearly Salary changes’ with p-level equal to 0.88 for APP segment can be excluded from the list of covariates.

4.5.2 Assessment of the models' accuracy

We assess the predictive accuracy of the model with training and testing samples. The factors chosen for the model validation are R square, mean absolute error (MAE), and root-mean-square error (RMSE).

The coefficient of determination R-square shows the proportion of the variance of the dependent variable which is explained by predictors and is used as a measure of Goodness-of-Fit of the linear model.

The standard R-square is defined as

$$R^2 = \frac{SS_{reg}}{SS_{tot}} = 1 - \frac{SS_{res}}{SS_{tot}} \quad (4.19)$$

where

$SS_{tot} = \sum_i (y_i - \bar{y})^2$ is the total sum of squares,

$SS_{reg} = \sum_i (\hat{y}_i - \bar{y})^2$ is the explained sum of squares.

$SS_{res} = \sum_i (y_i - \hat{y}_i)^2$ is the residual sum of squares.

The R-squared value closer to 1 shows higher fitting accuracy. For the comparison of the model we also use the error measures.

Mean absolute error (MAE) is an average of the absolute errors and calculated as follows:

$$MAE = \frac{1}{n} \sum_{i=1}^n |\hat{y}_i - y_i| = \frac{1}{n} \sum_{i=1}^n |e_i|, \quad (4.20)$$

where \hat{y}_t is the predicted value and y_t is the observed value.

Root-mean-square error (RMSE) is the standard deviation of the difference between predicted and observed values and calculated as follows:

$$RMSE = \sqrt{\frac{\sum_{t=1}^n (\hat{y}_t - y_t)^2}{n}}, \quad (4.21)$$

where \hat{y}_t is the predicted value and y_t is the observed value.

We use five regression methods. OLS uses a direct linear relationship between the predicted utilisation rate and the linear equation of covariates and the fitting accuracy can be assessed with R-squared directly. Fractional regression, Beta regression, and Beta transformation plus OLS have a transformed dependent variable, but the same linear equation in the right-hand side as the OLS method. Thus R-squared can be applied to the fitting accuracy assessment in a linear equation with the transformed outcome. Weighted Logistic regression with binary transformation uses maximum likelihood estimation for the binomial outcome, but it is applied for the prediction of the rates. We have a linear relationship between covariates and the dependent variable, transformed with logit. Thus, R-squared also can be applied for the assessment of the fitting accuracy between observed and predicted values.

The results of the validation of the regression methods for No Limit Change (BEH NL), Limit Changed (BEH CL) and Months n Book less than 6 (APP) segments for the one-stage model are presented in Table 4.7. We assess the predictive accuracy of the model with R-squared, MAE, and RMSE, and R-squared is selected as the indicator for the model for different segments comparison because it is possible to use an absolute value of the coefficient determination. For MAE and RMSE we can compare models inside one segment only, because they indicate an absolute, but not a relative error, which depends on the data sample.

Table 4.7 shows that the highest values of R-squared for all 3 types of the models is for MOB 6 and more with no limit change 0.5522; for MOB 6 or more with changes in limit it is 0.5066 and for MOB 1-5 it is 0.4535 and in all segments is given by the weighted logistic regression approach. However, the fractional regression approach gives an R-square value almost the same as the weighted logistic method. Other

methods such as OLS, Beta regression (non-linear mixed procedure) and beta-transformation + OLS have shown weaker results. The lowest coefficient of determination value was given by the Beta-regression plus OLS approach. However, the Beta-regression plus OLS approach has the smallest MAE values (0.1779, 0.1831 and 0.2051 for MOB 6+ no limit change, MOB 6+ with changes in limit and MOB 1-5 respectively). For comparison, the weighted logistic regression has MAE values 0.1922, 01941 and 0.2169 for the same types of models. OLS gives results which are not noticeably worse than fractional regression, but the OLS method can give results out of the defined range of permissible values.

Table 4.7 Summary validation of the regression methods for three utilisation rate segments

One-Stage Model	Method	Training Sample			Test Sample			
		R2	MAE	RMSE	R2	MAE	RMSE	
Month on Book 6 or more	OLS	0.5498	0.1930	0.2537	0.5498	0.1930	0.2537	
	Fractional (Quasi-Likelihood)	0.5502	0.1922	0.2544	0.5509	0.1919	0.2534	
	Limit NO Change	Beta regression (nlmixed)	0.5341	0.2076	0.2589	0.5344	0.2071	0.2580
		Beta transformation + OLS	0.4698	0.1779	0.2761	0.4707	0.1781	0.2751
		Weighted Logistic Regression	0.5522	0.1921	0.2538	0.5533	0.1917	0.2527
	OLS	0.5010	0.1967	0.2552	0.5064	0.1955	0.2527	
	Fractional (Quasi-Likelihood)	0.5040	0.1950	0.2544	0.5099	0.1937	0.2518	
	Limit Changed	Beta regression (nlmixed)	0.4877	0.2080	0.2586	0.4911	0.2071	0.2566
		Beta transformation + OLS	0.4246	0.1831	0.2740	0.4350	0.1810	0.2704
		Weighted Logistic Regression	0.5066	0.1941	0.2538	0.5136	0.1926	0.2509
Month on Book 1-5	OLS	0.4481	0.2200	0.2820	0.4474	0.2180	0.2802	
	Fractional (Quasi-Likelihood)	0.4513	0.2171	0.2812	0.4494	0.2154	0.2796	
	Beta regression (nlmixed)	0.4075	0.2431	0.2922	0.4085	0.2400	0.2898	
		Beta transformation + OLS	0.3324	0.2051	0.3102	0.3287	0.2048	0.3088
		Weighted Logistic Regression	0.4535	0.2169	0.2806	0.4547	0.2146	0.2783

The comparative analysis of results Table 4.7 given from training sample and testing sample has demonstrated that the best predictive models for the direct estimation of utilisation rate for all segments (BEH NL, BEH CL, and APP) are fractional regression and weighted logistic regression with binary transformation of the data sample because of the highest R-squared and lowest RMSE indicators. Although the Beta regression plus OLS method has the lowest MAE indicator, this method has not been selected because of with high RMSE, which is more sensitive to outliers than MAE due to the usage of the squares of errors.

There are few studies that model utilisation rate. Fulford and Schuh (2017) gained an R-squared around 0.74 for the utilisation rate model built on Equifax/NY Fed CCP data with use of time series technique. Arsova et al. (2011) and Loterman et al. (2012)

have shown the same best methods for Loss Given Default prediction. Thus our utilisation rate modelling results are consistent with those obtained from comparing similar algorithms for modelling a different proportion. Some papers related to credit card usage and debt prediction show lower results for the coefficient of determination. For example, these R-squared values were found for Estimated Expenditure R^2 of 0.097, [Banasik & Crook, 2001], credit card usage 0.3919 (linear bivariate correlation as alternative calculation of R-squared), [Kim & De Vaney, 2001], Outstanding Credit Card Balances 0.30, [Tan et al., 2011], Card Debt 0.10, [Cohen-Cole, 2001]. Loss Given Default prediction with the same models is noticeably more accurate with R^2 of 0.443 Qi & Zhao (2011), 0.8 – 0.4. Yao et al. (2014).

The statistics for assessment of the predictive accuracy R-squared, MAE, and RMSE might give the different results in comparison with the fitting of the observed and predicted target distributions. For example, the visual analysis shows that the Beta Transformation + OLS method has the most closely fitted distribution predicted values of utilisation rate for 1 month (M1 in figure) period (blue colour) to the distribution of observed values (red colour) among all methods for No Change Limit segment (Figure 4.15). The distribution of predicted values has an insignificant shift to the left side of the distribution of observed values around the utilisation values, but for low and moderate utilisation values the shapes of observed and predicted distribution are close to each other. Two curves show the smoothed distribution of observed and predicted values, presented with the histogram. Thin lines show the spline approximation of the distribution function for both observed and predicted values.

Figure 4.15 Beta Transformation + OLS distribution versus observed

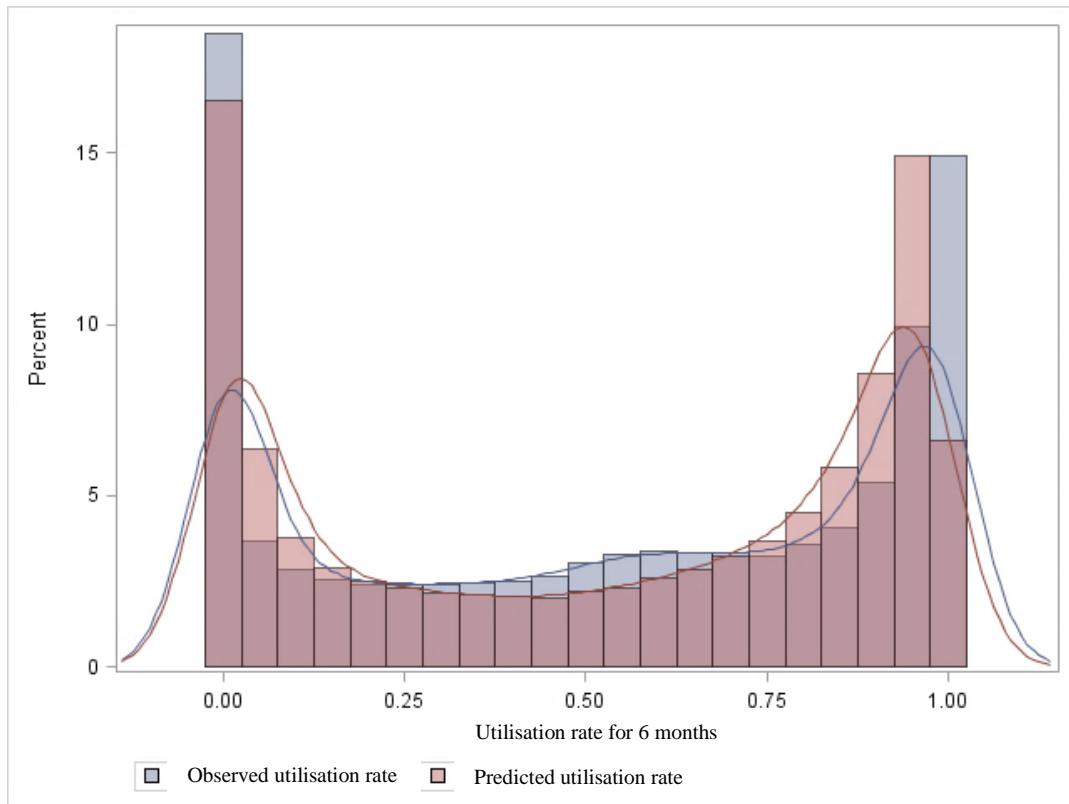
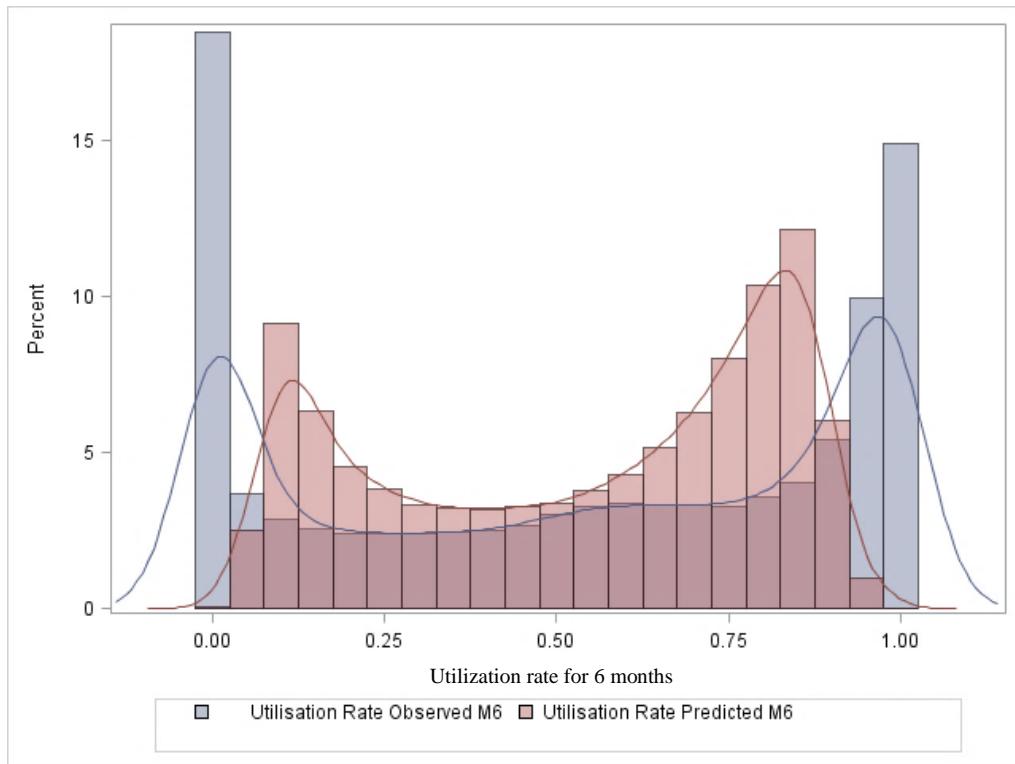


Figure 4.16 shows the distributions of utilisation rate: observed (blue colour) and predicted using fractional regression (red colour). Fractional regression with the highest predictive accuracy has low fitting between observed and predicted values distributions according to results of the visual analysis, especially, in the area of zero and one values, no utilised card and full utilisation. We have few cases for full utilisation rate and peak in the area of the utilisation rate around 0.8-0.85, so in this case, the fractional regression generally underestimated the utilisation rate. Thus, the fine fitting of the shape of the original distribution may not necessarily be the indicator of fine predictive accuracy of the model.

Figure 4.16 Fractional regression distribution versus observed



The distributions of observed and predicted values for each model can be compared with descriptive statistics (Table 4.8). We compare distributions by mean, standard deviation, skewness, the coefficient of variation, variance, kurtosis, and standard error mean.

As can be seen, the mean values are around the 0.53 value for all approaches excluding beta regression, which has underestimated mean value. On the other hand, Beta transformation with OLS shows the highest standard deviation and variation. Standard deviation has higher values around 0.37 for Beta+OLS approach, which is close to observed distribution. Other methods have standard deviation around 0.25-0.28, so the predicted values are more concentrated around the mean, than for Beta+OLS method. OLS, Fractional regression, and Weighted logistic regression with binary transformation have the most skewed distribution of predicted values in comparison with other methods (skewness = -0.36, -0.34, and -0.35 respectively). Beta regression and Beta transformation with OLS demonstrate coefficients of skewness close to the observed distribution: -0.26 and -0.23 versus the original of -0.2. Variance and Kurtosis coefficients of the distribution predicted with Beta-transformation + OLS method values are also close to the observed distribution than for other methods.

Table 4.8 Descriptive statistic for predicted distributions for the utilisation rate for 1-6 months for Limit NO Change Model

Statistic	Observed	OLS	Fractional	Beta regression	Beta+OLS	Weighted Logistic Regression
Mean	0.53520	0.53520	0.53893	0.50482	0.53220	0.53516
Std Deviation	0.37922	0.28105	0.28297	0.25019	0.37851	0.28243
Skewness	-0.20935	-0.36165	-0.34751	-0.26532	-0.23848	-0.35561
Coeff Variation	70.855	52.5123	52.5070	49.5595	71.1215	52.7744
Variance	0.14381	0.07899	0.08007	0.06259	0.14327	0.07976
Kurtosis	-1.52198	-1.19212	-1.37168	-1.35845	-1.59755	-1.35032

Table 4.9 presents the distributions of observed and predicted values for the average utilisation rate for 6 months period for Limit No Change Model. The values for observed and Fractional regression are used in Figure 4.16. The closest median value of the predicted distribution to the observed distribution is obtained with a simple OLS method (0.598 versus 0.592). No method has a shape which is close to observed distribution for the high utilised segment. The 95% quantile of values have a value of 1, which means that at least 5% of the population has full utilisation for the 6 months period. Estimated models demonstrate lower utilisation (around 0.9-0.95) for this quantile. The same situation is observed for zero utilisation – really 10% of observations have no utilisation for 6 months period, but the estimated models predict some utilisation (around 0.1) for this quantile.

Table 4.9 Outcome Distributions for five prediction methods for the utilisation rate for 1-6 months for Limit NO Change Model

Quantile	Observed	OLS	Fractional	Beta regression	Beta+OLS	Weighted Logistic Regression
100% Max	1	1.1187	0.9821	0.9495	1.0000	0.9829
99%	1	0.9440	0.9251	0.8713	0.9962	0.9188
95%	1	0.8960	0.8845	0.8287	0.9789	0.8804
90%	0.993	0.8647	0.8626	0.8053	0.9663	0.8581
75% Q3	0.924	0.7843	0.7987	0.7346	0.9079	0.7940
50% Median	0.592	0.5985	0.6153	0.5543	0.6285	0.6103
25% Q1	0.124	0.2753	0.2551	0.2608	0.0997	0.2568
10%	0.00	0.1177	0.1152	0.1333	0.0057	0.1103
5%	0.00	0.0770	0.0904	0.1143	0.0025	0.0850
1%	0.00	-0.0045	0.0562	0.0872	0.0007	0.0446
0% Min	0.00	-0.3780	0.0055	0.0232	0.0000	0.0018

The predicted values of utilisation rate using OLS have a U-shape frequency distribution with some values below zero and higher than 1 (see Table 4.9). To avoid the predicted values out of the 0,1 interval without use of a conditional function like if

a value less than 0 let it be 0, we apply other approaches which do give predicted outcome values in the range of 0 and 1.

Four other approaches have outcome values strictly between 0 and 1. The U-shape distribution corresponds with original utilisation rate distribution in the data sample. Beta regression (NLMIX) distribution is similar to fractional one but the left peak of low utilisation is higher than the area of high utilisation. This means that the prediction can be underestimated. The U-shape of the beta-transformation approximated with OLS corresponds closely to the observed distribution. The validation results of Beta-transformation+OLS are the weakest among all models (Table 4.7). An approach with weighted logistic regression also has a right peak not at the high bound, but in 0.85-0.9 area. The predicted outcome can be underestimated for high utilisation values, but the statistical test sample results have shown this method to give the best values. Weighted logistic and Fractional response regressions give similar distributions and higher predictive power in comparison with the other approaches.

We have selected the fractional regression method for further implementation in the total income model because it has given one of the highest predictive accuracies in the test samples and it is easy to compute. For the final income model, we need to estimate monthly utilisation rate for months $t+1, t+2, \dots, t+n$, where t is the observation point. In Table 4.10 we present the results of predictive accuracy for monthly utilisation rate models for the period from 1st to 6th month and the average utilisation rate for six months. The highest accuracy comes from a prediction for the first month after the observation point – R-squared is around 0.9. The lowest accuracy occurs for the sixth month after the observation point, which has been used for the testing of methods – R-squared is 0.5509 for the test sample. However, the average utilisation rate prediction over months 1 to 6 gives the fitting accuracy as for the 2nd month and R-squared around 0.79.

Table 4.10 Predictive accuracy for monthly utilisation rate model: Limit No Change model

Model Name	Training Sample			Test Sample		
	R^2	MAE	RMSE	R^2	MAE	RMSE
UT - Month 1	0.9055	0.0696	0.1156	0.9076	0.0696	0.1146
UT - Month 2	0.7918	0.1101	0.1714	0.7931	0.1106	0.1714
UT - Month 3	0.7182	0.1358	0.1995	0.7164	0.1369	0.2007
UT - Month 4	0.6638	0.1539	0.2183	0.6618	0.1551	0.2196
UT - Month 5	0.6205	0.1681	0.2326	0.6189	0.1692	0.2337
UT - Month 6	0.5502	0.1922	0.2544	0.5509	0.1919	0.2534
UT - Months 1-6	0.7934	0.1135	0.1614	0.7917	0.1146	0.1626

For the Limit Change and MOB 1-5 models the distributions and proportions have similar tendencies and differ by scales only, so we will not describe the detail of the distributions of predicted values to avoid repetition.

The analysis demonstrates that we have discrepancies between the results of methods for assessment of the predictive accuracy of the model such as statistical coefficients R-squared, MAE, and RMSE, and on the other hand, assessment with a comparison of descriptive statistics of observed and predicted distributions of the outcome. We believe that the statistics of fitting accuracy between observed and predicted outcome values are a more important indicator because we need to find the close relationship between predictors and predicted outcome, but not to fit the outcome distributions only.

4.5.3 The estimation with a lagged endogenous variable

The variables in the given regression model for the prediction of credit limit utilisation rate can be considered as a mixture of endogenous and exogenous. Application, behavioural, and macroeconomic predictors (or independent variables) are exogenous variables because their values are not dependent on the predicted variable – utilisation rate. In some sense, the behavioural variables, which are derivative from the initial (raw) variables, can be endogenous. For example, the changes in the outstanding balance depend on the outstanding balance and both variables are used as predictors. The current outstanding balance also can depend on the spending and payment transactions, so there are intercorrelations between predictors of the model. The utilisation rate is calculated as the ratio of the outstanding balance and credit limit, but

because of changes in the credit limit, is not necessarily fully correlated with the outstanding balance. We predict the utilisation rate and assume that it depends on some covariates, which are not dependent on the predicted variable. In short, we assume zero correlation between each covariate and the error. However, we believe the utilisation rate in period t depends on the utilisation rate in the previous periods. Thus, the utilisation rate in the previous periods is lagged endogenous variables. These variables can be estimated with a pooled regression approach, as, has been done in Section 4.5.1. However, an OLS estimate is inconsistent when a model has a lagged endogenous variable. But, estimates, which have been designed for use when we have a lagged endogenous variable, do yield consistent estimates. For example, Arellano and Bond (1991) proposed to use all previous values of the lagged exogenous variable to predict the present, instead of using a single value of its previous value or a fixed time series observation window.

Let's introduce the equation for dependent variable y_{it} based on panel data as follows:

$$y_{it} = \sum_{l=1}^{maxlag} \phi_l y_{i(t-l)} + \sum_{k=1}^K \beta_k x_{itk} + \gamma_i + \alpha_t + \varepsilon_{it} \quad (4.22)$$

where

γ_i is a cross-sectional fixed effect,

α_t is a time series fixed effect

x_{itk} is endogenous characteristics,

β_k is a coefficient of regression for endogenous characteristic x_k ,

ϕ_l is a coefficient of regression for lagged characteristics y_{t-l} .

The formula 3.22 is similar to formula 3.3, which also uses the lagged utilisation rate values to predict the utilisation rate. Dynamic Panel Estimators use lagged exogenous variables as y_{t-1} , y_{t-2} , etc. But in the Arellano and Bond (1991) approach y_{i1} is used to predict y_{i2} ; y_{i1} and y_{i2} to predict y_{i3} ; y_{i1} , y_{i2} , and y_{i3} to predict y_{i4} ; and so on.

The simple case of an autoregression in a panel setting (with only individual effects) is:

$$y_{it} = \phi y_{i(t-1)} + \gamma_i + \varepsilon_{it},$$

where ε_{it} is an estimation error.

Then the difference in the preceding relationship is as follows:

$$\Delta y_{it} = \phi \Delta y_{i(t-l)} + v_{it}$$

where $v_{it} = \varepsilon_{it} - \varepsilon_{it-l}$.

An efficient estimation is possible by using additional lags of the dependent variable as instruments. For example, both $y_{i,t-2}$ and $y_{i,t-3}$ might be used as instruments. The highest number of instruments for the dependent variable can be obtained at time, which is close to the final time period T.

$$\widehat{\beta}_{AB} = \left[\left(\sum_{i=1}^N \widetilde{\mathbf{X}}_i' \mathbf{Z}_i \right) \mathbf{W}_N \left(\sum_{i=1}^N \mathbf{Z}_i' \widetilde{\mathbf{X}}_i \right) \right]^{-1} \left(\sum_{i=1}^N \widetilde{\mathbf{X}}_i' \mathbf{Z}_i \right) \mathbf{W}_N \left(\sum_{i=1}^N \mathbf{Z}_i' \widetilde{\mathbf{y}}_i \right)$$

where $\widetilde{\mathbf{X}}_t$ is a $(T-2) \times (K+1)$ matrix with t th row $(\Delta y_{i,t-1}, \Delta \mathbf{x}'_{it})$, $t = 3, \dots, T$, $\widetilde{\mathbf{y}}_t$ is a $(T-2) \times 1$ vector with t th row Δy_{it} , and Z_i is a $(T-2) \times r$ matrix of instruments

$$\mathbf{Z}_i = \begin{bmatrix} \mathbf{z}'_{i3} & \mathbf{0} & \cdots & \mathbf{0} \\ \mathbf{0} & \mathbf{z}'_{i4} & & \vdots \\ \vdots & & \ddots & \mathbf{0} \\ \mathbf{0} & \cdots & \mathbf{0} & \mathbf{z}'_{iT} \end{bmatrix}$$

We used the same data sample as in section 4.5.1. Parameter estimates for One-stage model section and estimated coefficients using the Arellano and Bond approach with SAS procedure PANEL. The results are provided for No Limit Changed Segment both Pooled and Arellano-Bond Methods for 1-month prediction (Table 4.11). We used only behavioural time-varying variables and excluded application time-constant variables.

The utilisation rate at the month of observation (UT_1), which is lagged for the training sample for 1 month from the target utilisation rate, is the most significant covariate for the OLS pooled method with t-value = 572.63. However, for the Arellano-Bond method, the lagged utilisation rate does not have any advantages in comparison with other covariates (with t-value = 3.7 only), with a coefficient value 0.4 instead of 0.8 for OLS pooled method.

Table 4.11 Estimated coefficients for OLS pooled and Arellano-Bond method for 1 month utilisation rate

Variable	Pooled OLS				Arellano-Bond			
	Estimate	Standard Error	t Value	Pr > t	Estimate	Standard Error	t Value	Pr > t
Intercept	0.0218	0.0036	6.0100	<.0001				
mob	-0.0019	0.0001	-21.3700	<.0001	-0.0096	0.0017	-5.5100	<.0001
limit_6	0.0000	0.0000	.	.	0.0000	0.0000	1.1600	0.2471
avg_balance	0.0000	0.0000	.	.	0.0000	0.0000	3.9600	<.0001
UT_1	0.8222	0.0014	572.6300	<.0001	0.4003	0.1091	3.6700	0.0002
b_UT1_to_AvgUT16ln	0.0008	0.0002	4.2400	<.0001	0.0649	0.0169	3.8500	0.0001
b_UT1to2ln	0.0004	0.0002	2.2900	0.0221	0.0077	0.0075	1.0300	0.3013
b_UT1to6ln	-0.0023	0.0001	-17.1600	<.0001	-0.0214	0.0072	-2.9800	0.0029
b_AvgOB16_to_MaxOB16_ln	-0.0024	0.0007	-3.4200	0.0006	0.0525	0.0502	1.0500	0.2955
b_TRmax_deb16_To_Limit_ln	0.0048	0.0002	27.5400	<.0001	-0.0059	0.0130	-0.4600	0.6486
b_TRavg_deb16_to_avgOB16_ln	0.0007	0.0003	1.9500	0.0512	0.0442	0.0379	1.1700	0.2434
b_TRsum_deb16_to_TRsum_crd16_ln	0.0074	0.0003	23.7000	<.0001	-0.0391	0.0247	-1.5800	0.1135
b_NumDeb13to46ln	0.0053	0.0002	23.0600	<.0001	-0.0372	0.0129	-2.8800	0.0040
b_avgNumDeb13	0.0008	0.0001	5.4800	<.0001	0.0357	0.0131	2.7200	0.0064
b_OB13_to_OB46ln	-0.0034	0.0003	-13.6500	<.0001	0.0347	0.0096	3.6100	0.0003
b_OB_avg_to_eop1ln	-0.0126	0.0002	-74.7800	<.0001	-0.1179	0.0142	-8.3300	<.0001
b_pos_flag_1	0.0002	0.0008	0.2100	0.8350	-0.1745	0.0275	-6.3500	<.0001
b_pos_flag_13	0.0160	0.0010	15.7100	<.0001	0.0438	0.0590	0.7400	0.4581
b_atm_flag_1	0.0438	0.0007	60.1100	<.0001	-0.2099	0.0195	-10.7400	<.0001
b_atm_flag_13	-0.0007	0.0012	-0.5600	0.5745	-0.0729	0.1515	-0.4800	0.6303
b_pos_flag_used46vs13	0.0067	0.0009	7.5200	<.0001	0.1688	0.0535	3.1500	0.0016
b_pos_flag_use13vs46	-0.0125	0.0010	-12.9600	<.0001	0.0258	0.0908	0.2800	0.7764
b_atm_flag_used46vs13	-0.0077	0.0010	-7.3700	<.0001	0.1092	0.1335	0.8200	0.4132
b_atm_flag_use13vs46	0.0020	0.0010	1.9600	0.0499	-0.0867	0.0930	-0.9300	0.3510
b_pos_use_only_flag_13	-0.0006	0.0013	-0.4400	0.6610	-0.3093	0.1318	-2.3500	0.0189
b_TRsum_crd13_to_OB13_ln	-0.0031	0.0002	-12.9000	<.0001	0.0381	0.0218	1.7500	0.0803
b_payment_lt_5p_13	0.0249	0.0008	32.6600	<.0001	0.1279	0.0733	1.7400	0.0812
l_ch1_ln	-0.0022	0.0050	-0.4300	0.6658	-0.1188	0.0356	-3.3400	0.0008
l_ch1_flag	0.0396	0.0026	15.5300	<.0001	0.0658	0.0163	4.0400	<.0001
l_ch6_flag	0.0112	0.0010	11.1900	<.0001	0.0023	0.0075	0.3100	0.7562
Unempl_Inyoy_6	-0.5382	0.0120	-44.9700	<.0001	-0.1837	0.0825	-2.2300	0.0260
UAH_EURRate_Inmom_6	0.6832	0.0145	47.0900	<.0001	0.6493	0.0447	14.5400	<.0001
UAH_EURRate_Inyoy_6	-0.2845	0.0088	-32.3600	<.0001	-0.1071	0.0369	-2.9100	0.0037
CPI_Inqoq_6	-0.0502	0.0233	-2.1600	0.0311	-0.8964	0.1151	-7.7900	<.0001
SalaryYear_Inyoy_6	0.2733	0.0198	13.8100	<.0001	-0.2272	0.0779	-2.9200	0.0036

Table 4.12 shows that the Arellano-Bond method has significantly weaker fitting accuracy in comparison with the pooled OLS method. R-squared for the test sample is only 0.23 for Arellano-Bond method versus 0.88 for Pooled OLS. MAE and RMSE values are significantly higher for Arellano-Bond estimation.

Table 4.12 The fitting accuracy of Pooled OLS and Arellano-Bond method estimation

Model for UT m+1	Train			Test		
	R^2	MAE	RMSE	R^2	MAE	RMSE
OLS Pooled	0.8985	0.0685	0.1294	0.8795	0.0895	0.1514
Arellano-Bond	0.2637	0.2617	0.3486	0.2381	0.2895	0.3716

The use of methods developed for a model with lagged endogenous variables appears to get weaker fitting accuracy than pooled methods, which do not consider the intercorrelation between endogenous variables.

4.5.4 Two-stage model summary for the 6-months utilisation rate prediction

The two-stage model consists of two parts: the probability of zero utilisation and the probability of full utilisation. It uses logistic regression and the proportion estimation with the use of the set of the same methods as for the one-stage model.

We use logistic regression models for the prediction of the probability of the utilisation rate being 0 and 1. We demonstrate the model for the probability of utilisation rate values for 6 months period. The set of covariates is the same as for the utilisation rate value prediction as was used in the linear regressions (Table 4.13) because we believe that covariates, which explain the utilisation rate, also are good for the prediction of the probability of use and active usage is correlated with high utilisation rate. The probabilities modelled are $\text{Pr}(\text{UT}=0)$ versus $\text{Pr}(\text{UT}>0)$ and $\text{Pr}(\text{UT}=1)$ versus $\text{Pr}(\text{UT}<1)$.

Table 4.13 shows the results. The coefficients are mainly significant apart from the ratio of the spending transactions to payment transactions for the last 6 month ($b_{\text{TRsum_deb16_to_TRsum_crd16}}$), possibly because of correlation with other behavioural characteristics, which discover transactions and balance changes for 6 months. Some coefficients have opposite signs, for example, average utilisation rate for the last 6 month for the probability of the utilisation rate being zero is positive and for the probability being one is negative. So, it is logical that the high previous utilisation causes a higher probability of full utilisation. Credit cardholders, who use the credit card mainly for Point-of-Sales transactions, have a higher probability to get zero utilisation in 6 months ($b_{\text{pos_use_only_flag_13}} = -0.1345$). On the other hand, credit cardholders, who use the credit card mainly for ATM cash withdrawals, have a higher probability to get full utilisation in 6 months ($b_{\text{atm_use_only_flag_13}} = -0.8294$). According to results, an increase in CPI may cause an increase of the probability of full utilisation of credit card ($\text{CPIYear_lnyoy_6} = 6.8829$), but a decrease in the probability of zero utilisation ($\text{CPIYear_lnyoy_6} = -3.2825$).

Table 4.13 Estimated coefficients for logistic regression for utilisation rate equal to 0 and 1

Model	Pr (UT = 0)			Pr (UT = 1)		
	Parameter	Estimate	Standard error	Pr > ChiSq	Estimate	Standard error
Intercept	-0.6627	0.220700	0.0027	5.6316	0.185900	<.0001
mob	-0.0393	0.001380	<.0001	0.0396	0.002030	<.0001
limit_6	-0.00001	0.000003	0.0012	0.000172	0.000014	<.0001
avg_balance_6	0.000106	0.000006	<.0001	-0.00023	0.000014	<.0001
b_AvgOB13_TO_MaxOB13	1.7832	0.032800	<.0001	2.6545	0.124800	<.0001
b_TRmax_deb16_To_Limit	-0.2305	0.018000	<.0001	-0.2499	0.024700	<.0001
b_TRsum_deb16_to_avg	-0.0015	0.000499	0.0026	0.0162	0.005840	0.0055
b_TRsum_deb16_to_Trsum_crd16	0.000012	0.000019	0.5297	-0.00000859	0.000109	0.9371
b_Avg_UT16	1.4957	0.038300	<.0001	-7.2261	0.119100	<.0001
b_UT1_to_AvgUT16	0.1178	0.006360	<.0001	-0.7594	0.034500	<.0001
b_OB_avg_to_eop1	-0.00002	0.000009	0.0319	-0.00004	0.000015	0.0044
b_inactive13	0.11	0.025000	<.0001	-3.9989	0.168000	<.0001
b_fullpaid1	-0.5401	0.021400	<.0001	-0.4998	0.118300	<.0001
b_avgNumDeb16	0.0966	0.004290	<.0001	-0.0006	0.001040	0.5636
b_max_dpd16	0.00517	0.000926	<.0001	-0.00171	0.000225	<.0001
no_dpd	0.2342	0.202400	0.2472	2.1901	0.068200	<.0001
max_dpd_30	0.2697	0.199900	0.1772	1.8441	0.067600	<.0001
max_dpd_60	0.2023	0.226900	0.3727	0.9979	0.071200	<.0001
b_pos_flag_0	0.3719	0.023000	<.0001	-0.4087	0.019700	<.0001
b_pos_flag_13	0.6741	0.029000	<.0001	-1.1132	0.047300	<.0001
b_atm_flag_0	0.4401	0.018900	<.0001	-0.4117	0.018600	<.0001
b_pos_flag_use13vs46	-0.1753	0.024500	<.0001	0.2957	0.027700	<.0001
b_atm_flag_use13vs46	-0.1681	0.021300	<.0001	0.5827	0.039300	<.0001
b_pos_use_only_flag_13	-0.1345	0.026300	<.0001	0.6223	0.037100	<.0001
b_atm_use_only_flag_13	0.5797	0.021000	<.0001	-0.8294	0.046700	<.0001
AgeGRP1	0.036	0.019400	0.0638	-0.0809	0.022000	0.0002
AgeGRP3	0.0114	0.016600	0.4907	0.3172	0.021800	<.0001
customer_income_In	-0.0959	0.020300	<.0001	0.2974	0.029000	<.0001
Edu_High	-0.0708	0.018900	0.0002	0.2482	0.021300	<.0001
Edu_Special	0.0237	0.018700	0.2042	0.0696	0.019200	0.0003
Edu_TwoDegree	0.00344	0.038900	0.9295	0.1455	0.066300	0.0282
Marital_Civ	-0.0541	0.030300	0.0736	-0.1103	0.032400	0.0007
Marital_Div	-0.00026	0.018800	0.9889	-0.2041	0.023900	<.0001
Marital_Sin	-0.0799	0.017100	<.0001	-0.141	0.019400	<.0001
Marital_Wid	-0.0085	0.034000	0.8023	0.0667	0.049400	0.1764
position_Man	-0.0318	0.018400	0.0847	-0.0854	0.029100	0.0033
position_Oth	0.0399	0.018600	0.0318	-0.0904	0.023000	<.0001
position_Tech	0.0455	0.017900	0.0109	-0.1447	0.020300	<.0001
position_Top	-0.0181	0.033800	0.5916	-0.0468	0.068900	0.4969
sec_Agricult	-0.00769	0.032400	0.8125	-0.0451	0.041900	0.282
sec_Constr	0.1343	0.046900	0.0042	0.08	0.056000	0.1536
sec_Energy	-0.0844	0.027800	0.0024	-0.1953	0.036800	<.0001
sec_Fin	0.12	0.019500	<.0001	0.2024	0.036100	<.0001
sec_Industry	-0.0834	0.054500	0.1258	-0.3229	0.061500	<.0001
sec_Manufact	0.1818	0.048400	0.0002	-0.3358	0.047000	<.0001
sec_Mining	-0.0319	0.031700	0.3139	-0.2339	0.035100	<.0001
sec_Service	0.0311	0.016000	0.0526	-0.1861	0.020900	<.0001
sec_Trade	0.1324	0.024500	<.0001	-0.1032	0.026300	<.0001
sec_Trans	0.00584	0.043000	0.892	0.0174	0.053400	0.7439
Unempl_Inyoy_6	-1.6733	0.233500	<.0001	-1.9065	0.303800	<.0001
UAH_EURRate_Inyoy_6	-1.8034	0.125200	<.0001	1.089	0.196900	<.0001
CPIYear_Inyoy_6	-3.2825	0.312800	<.0001	6.8829	0.481300	<.0001

To assess the predictive accuracy of the logistic models, we use three indicators: the Kolmogorov–Smirnov test, Gini index, and ROC. For the utilisation rate we use the same statistics as for one-stage model (R-square, MAE, and RMSE). We provide with results of the predictive accuracy of the two-stage model for the segment MOB 6+ with No changes in the limit in Table 4.14.

Generally, two-stage models show better fitting accuracy and prediction results for development and testing samples than one-stage, or direct, models (Belotti and Crook, 2012). However, for our results the differences in prediction errors are insignificant. For example, for Limit No Change model for OLS method for one-stage and two-stage approaches R-squared = 0.5498 and 0.5536, MAE = 0.193 and 0.191 respectively. However, if we compare the Stage 2 model with one-stage direct estimation it can be seen that the one-stage model gives better results. For example, one-stage and stage 2 of two-stage model R-squared = 0.5498 versus 0.4235, and MAE= 0.193 versus 0.195 respectively. However, this difference is compensated high prediction performance at the Stage 1 model – logistic regression, which has KS = 0.6262 and 0.5931, Gini = 0.7479 and 0.7243 for the probability of the utilisation rate being equal to zero and the utilisation rate being equal to 1 respectively. Thus the predictive modes for the utilisation rate between 0 and 1 show weaker fitting accuracy than the models, which includes 0 and 1 bounds. On the other hand, the use of the two-stage model, which predicts the probability to be in boundary states, gives slightly better results than the one-stage model.

Table 4.14 Comparative analysis of fitting accuracy for two-stage models for the 6-months period

Month on Book Changes	Limit	Stage	Method	Two-stage model			One-stage model (for comparison)					
				Development Sample			Validation Out-of-sample			Development Sample		
MOB 6 or more	Limit NO change	Stage 1	Probability	KS	Gini	ROC	KS	Gini	ROC			
			Pr(UT=0)	Logistic Regression	0.6262	0.7479	0.8739	0.6331	0.7547	0.8774		
		Stage 2	Proportion Estimation	R2	MAE	RMSE	R2	MAE	RMSE			
			OLS	0.4310	0.1948	0.2462	0.4235	0.1950	0.2462			
		0<UT<1	Fractional(Quasi-Likelihood)	0.4309	0.1946	0.2463	0.4235	0.1950	0.2462			
			Beta regression (nlmixed)	0.4183	0.2102	0.2506	0.4108	0.2104	0.2507			
			Beta transformation + OLS	0.3680	0.1802	0.2673	0.3618	0.1809	0.2673			
		Two-stage	Weighted Logistic Regression	0.4325	0.1945	0.2457	0.4253	0.1948	0.2456			
			Aggregate	R2	MAE	RMSE	R2	MAE	RMSE	R2	MAE	RMSE
			OLS	0.5534	0.1913	0.2535	0.5536	0.1910	0.2526	0.5498	0.1930	0.2537
			Fractional(Quasi-Likelihood)	0.5527	0.1915	0.2536	0.5529	0.1912	0.2528	0.5502	0.1922	0.2544
			Beta regression (nlmixed)	0.5366	0.2068	0.2581	0.5364	0.2063	0.2574	0.5341	0.2076	0.2589
		0<= UT <=1	Beta transformation + OLS	0.4720	0.1773	0.2754	0.4724	0.1774	0.2745	0.4698	0.1779	0.2761
			Weighted Logistic Regression	0.5548	0.1914	0.2531	0.5553	0.1910	0.2521	0.5522	0.1921	0.2538

The two-stage models for other two segments: Limit Changed (BEH CL) and Month on Book less than 6 (APP) have shown approximately the same results for models quality assessment and validation. The difference is in the relative scale only. For example, the APP model has shown lower KS and Gini parameters (~0.30 and ~0.40 respectively) than the BEH NL model, but it is normal and expected for the application scoring.

4.6 Conclusion

The aim of this Chapter is to find a more accurate method for the prediction of credit limit utilisation rate. We applied some methods already used to predict proportions such as Loss Given Default, and we compare the results with that published in the literature.

The general modelling trends for rates, discussed in the literature, are confirmed for the utilisation rate too. Because the utilisation rate has an outcome value bounded between 0 and 1, specific methods should be applied to predict these values, for example, beta-regression, fractional regression, or weighted logistic regression with binary transformation. Linear regression and other unbounded outcome methods can be applied but can give results out of the outcome area.

We applied five one stage methods and two-stage methods. The one stage methods were OLS, fractional regression (quasi-likelihood), beta-regression (non-linear), beta-transformation + OLS/GLM, and weighted logistic regression with data binary transformation.

The best validation result has been obtained from fractional regression and weighted logistic regression with data binary transformation in section 4.5.2. However, OLS results are only marginally different and predict a similar distribution. The Beta transformation method predicts the most similar distribution shape but has the worst predictive accuracy results. Two-stage models show the slightly better result in section 4.5.4 for all five approaches than the one-stage model. The probabilities estimation models for the utilisation rate bound values 0 and 1 and have high performance results for credit risk behavioural models.

We also segmented our population and used three separate groups of models for customers with less than 6-months on balance, a customer with 6 or more months on balance and no changes in the limit, and a customer with 6 or more months on book and an increased limit. These three segments of the sample have different sets of covariates. For example, additional limit changes parameters or a limited number of behavioural characteristics for MOB less than 6 accounts.

Generally, the best results were given by the two-stage model with fractional regression and weighted logistic regression with binary transformation. Models for changed limit are slightly more accurate than MOB 6+ without limit change rather because of additional parameters, models for MOB less than 6 show weaker predictive power because of the short behavioural history and these models are mainly based on application data.

We provide with a list of the most significant explanatory characteristics for the utilisation rate prediction with related positive or negative correlation between characteristics.

Behavioural characteristics such as the credit limit utilisation rate at the observation point month, logarithm of the sum of purchase transactions to the sum of payments in month 1, purchase amount for the last 6 months, ATM cash withdrawals flag, credit limit change have *a positive correlation* with the utilisation rate, but credit limit, month on book, logarithm of the average purchase transaction amount to the average outstanding balance for six months, loan payment amount for last 3 months *have a negative correlation* with the utilisation rate.

Application characteristics such as age less than 25-year-old, number of children more than one have *a positive correlation* with the utilisation rate, but the logarithm of the customer income to the average income, higher education, position top manager, sector of work finance *have a negative correlation* with the utilisation rate.

State characteristic a number of times in inactive state for the last year has a negative correlation with the utilisation rate.

Macroeconomic characteristic CPI change for the last quarter has a positive correlation and the exchange rate of the local currency to EUR for the last year has a negative correlation with the utilisation rate.

Because of the usage of lagged endogenous variables as predictors, we tested a method for simple dynamic panel estimation – the Arellano-Bond (1991) method. However, it did not yield prediction that were as accurate as we have got with use of other methods such as OLS, fractional regression, which do not use lagged endogenous variables.

The next stage of the investigation for the utilisation rate modelling can be dedicated to the use of other methods for a prediction like CHAID, SVM etc. and based on LGD modelling experience can give even higher performance results than regression. Credit limit utilisation rate depends on the customer behavioural pattern and revolvers, and transactors can have different utilisation rates. This assumption has been investigated indirectly as behavioural characteristics reflect behavioural patterns but more accurate estimation of the utilisation rate can require models segmentation by the behavioural pattern.

There is lack of research on the prediction of the *credit limit utilisation rate*. We implemented some methods already used for proportions prediction such as Loss Given Default (Yao *et al.*, 2014; Arsova *et al.*, 2011) and applied them to *the utilisation rate* (for example, Agarwal *et al.*, 2006). Also we used the credit card usage approaches (Crook *et al.*, 1992; Banasik and Crook, 2001) as the probability of full use or no use of a credit card in a two-stage model. We also tested the Arellano and Bond (1991) method for estimation of the utilisation rate with lagged endogenous variables. We extended Fulford and Schuh (2017) by application of the new methods for the prediction of the utilisation rate.

5 Chapter Four. Credit Card Holders' States Transition Probability: Model description

5.1 Introduction

Credit card profitability prediction is a complex problem because of the variety of the cardholders' behaviour patterns and different sources of the interest and transactional income. Each consumer account can move to a number of states such as a 'transactor', a 'revolver', and a 'delinquent' and a model for generated income prediction is required. For more accurate modelling of credit cardholder behaviour, the dual nature of revolving products both as a standard loan and as a payment tool need to be considered. Thus, scoring models should be split up according to customer behaviour segment and source of income for the bank. The behavioural state of the credit card account depends on the type of card usage and payments delinquency. Thus, the following credit card states can be defined: inactive, transactor, revolver, delinquent, default. This definition of states can be applied both to a credit card account and to the credit cardholder. In the scope of this work, the state of a customer and state of an account is considered as the same concept.

The estimation of the transition probability matrix between states at the account level helps to avoid the memorylessness property of the Markov Chains approach. The proposed credit cards profit prediction model consists of five stages: account or consumer status prediction with conditional transition probabilities, outstanding balance and interest income estimation, non-interest income estimation, expected losses estimation, and profit estimation. The detailed investigation has shown that additional intermediate states such as paid revolver can give a more accurate estimation of the transition probabilities.

The main aim of this chapter is to give the definition of credit card states and describe the model and methods for the prediction of the transition probabilities between credit card states. The empirical investigation and results of the modelling are provided in the next Chapter 6.

This Chapter consists of 6 sections. Section 5.2 discusses segmentation, section 5.3 gives the general model setups and model building methodology, and section 5.4

describes methods for regression analysis such as transition probability prediction with conditional binary logistic, ordinal and multinomial regressions. Section 5.5 gives a brief univariate analysis of main characteristics used for transition probabilities prediction. Section 5.6 discusses some transition matrices issues and show why we avoid the use of Markov Chains in this research.

This Chapter fills the gap in the use of transition models at the account level in the application for the prediction of credit card income. The proposed model includes not only delinquent versus non-delinquent and active versus inactive states, but the extended set of states for income prediction tasks.

We developed a methodology for the total income prediction, which accumulates several individual models for various sources of revenue and explores the income in the dimension of credit card behavioural types. We have proposed the set of behavioural states for the more accurate credit card income prediction, which contains only profit related states.

5.2 Segmentation

5.2.1 General description of credit card states

More accurate modelling of account behaviour can be gained through segmentation. For instance, McKinsey (Fiorio et al., 2014) proposes credit card segmentation with three card types — rewards, low-rate and subprime, and notes that credit card holders are typically categorized as transactors, revolvers or subprime. In the terminology of McKinsey (Fiorio et al., 2014) the customers can be related to three segments: i) Prosperous and content, ii) Deal chasers, iii) Financially stressed. However, in the literature a more popular segmentation of credit cards is the diversification by the level of risk such as behavioural score (Malik & Thomas, 2007; So & Thomas, 2011), delinquency bucket, usage activity (Crook et.al, 1992) and usage type (So et. Al, 2014).

Predictive models for risk and usage can be built for a whole credit card portfolio at the pool level as a Markov Chain. However, significant differences between credit card usage types can decrease the predictive accuracy for a generic or a whole portfolio model because the different usage types have individual behavioural drivers for risk,

utilisation, purchases types, and profit. For example, a non-active credit card does not bring the risk of non-payment in the current period, but, it can bring relatively high default risk in case of rapid activation which can be caused for reasons such as the card being stolen and used by fraudsters or by hackers, acute shortage of money. An active credit card user who has a positive, outstanding balance at the end of a billing period is a standard cardholder who is interesting to a bank as a generator of income and loss. An early stage delinquent customer brings a higher risk of default than a current customer. However, if (s)he is a so-called ‘lazy payer’ or user who pays every month but with several days of delinquency, generally a lender can get more income and profit in comparison with a disciplined debtor because of the penalty for the late payment.

Thus, the probability of events such as non-payment and default depends on the previous behaviour, on the current state and the following states. The ‘following states’ mean the states which occur after the current time (the observation point) and before the time of the event (performance point). For example, if the event of default is defined as four consecutively missed monthly payments and we use a six-month prediction horizon an account can get to the state of default from the current state in several ways. However, we can predict only the probability of the next months states. So the predictive model for N steps forward from current time t can be built as i) a direct prediction for event occurrence after N steps or as ii) a probability of an event at stage N conditional on states $t, t+1, \dots, t+N-1$.

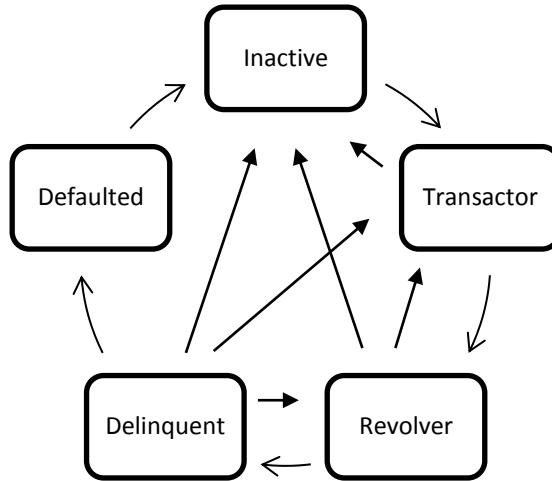
There are several definitions of segmentations and states in the literature and industry practice. The most popular types are delinquent and default for risk modelling purposes, and this is a topic of credit risk scoring investigations. For non-delinquent credit cards, the literature proposes different approaches to segmentation. Thomas and So (2011) used segmentation by the behavioural score, or risk of default, for the prediction of the profitability of credit cards.

At a high level, a credit card holder can be non-active, active, delinquent, or a defaulter. Active and non-delinquent credit cards holders can be split up into two groups: revolvers and transactors. A revolver is a user who carries a positive credit card balance and does not pay off the balance in full each month – a rollover. A transactor is a user

who pays in full on or before the due date of the interest-free credit period. Such a competent or a convenient user does not incur any interest payments or finance charges. At the more aggregated level the system of credit card main states can be described by the following set: inactive, transactor, revolver, delinquent and default. The account's state is predicted for the period $t+1$, where t is the observation point. Each account can move to a limited number of specific states only, depending on the current state. An inactive account in period t can transit to transactor or revolver in period $t+1$. A transactor can become a revolver or inactive. A revolver can become delinquent or a transactor or an inactive in the next period. Delinquency is a state, from which an account can transit to any state, including default. A cardholder can pay back the full arrears amount and become inactive. Depending on the zero or non-zero outstanding balance and on the availability of purchase transactions, an account can move to inactive, transactor or revolver state in the next month.

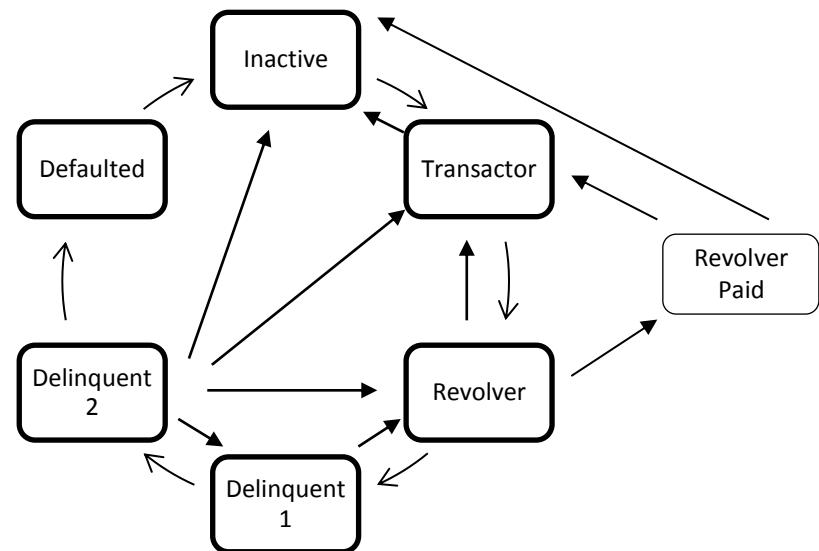
Figure 5.1 shows the possible transitions between account states over a month. The direction of the arrows shows the possible transitions. For example, a revolver can move to a transactor state, a transactor can move to a revolver state, but a transactor cannot jump to a delinquent state – cardholder must keep a positive outstanding balance at the end of period, so to become a revolver, and then go to the delinquency state in case of non-payment of obligatory payments. A cardholder in the default state can theoretically become inactive after repayment of total debt. However, in our research we consider default state as an absorbing state, so a cardholder with 61+ days past due is not able to use his/her credit card further and stay in a credit portfolio as a bad debt without write-off.

Figure 5.1 Transition between main states over a month



The set of states can be extended with additional states (see Figure 5.2). The delinquent state consists of two states: delinquent for 1 month (1-30 days past due) and delinquent for 2 months (31-60 days past due). A revolver can return to inactive or transactor state only via transitional ‘Revolver paid’ state. This state means that a cardholder has a positive outstanding balance at the beginning of the month, but has paid full debt amount and has a zero outstanding balance at the end of the month. A detailed discussion of this state is presented in Chapter 6.5.

Figure 5.2 Transition between all states over a month



The system of transitions between states describes the change in state from time t to time $t+1$. However, for $t+2$ and more periods in the future more transitions are possible for more combinations of states. For example, a transactor and even an inactive account can become delinquent in two months via a revolver state in $t+1$. Generally, the transition is possible from any state to any state for some number of steps.

A particular question is the definition of the defaults state. The default state is defined as a certain number of consecutive months in delinquency or, in other words, number of missing payments. This can be specified, for example, as three and more missing payment, i.e. as days past due 61 or more (DPD 61+). Generally, it is possible to choose between two options for the transition from the default state. The first option is that a borrower can pay back arrears amount and come back to any other state as inactive, current or delinquent. The second option is to consider that the default state is an absorbing state and if a customer has got to default (s)he is not able to move back and stays defaulted forever.

In the current investigation, we have chosen the second option. The default is the absorbing state. Later the expected losses estimation is corrected with Loss Given Default estimation. We allow an account to remain in a state without transitioning to another state for an unlimited period.

Over its lifetime an inactive account can move to any state in a limited number of steps. For the default state definition of 4 or more missed payments, it is possible to get from the inactive state to default in a minimum period of 5 months (inactive – revolver – delinquent 1 month – delinquent 2 months – default).

In this section we give a general overview and concept of credit card account states. The detailed definition of states and their connection with credit card income sources are discussed in Section 5.3.1.

5.2.2 Reasons for transitions between states

Initially, credit card holders' behaviour is conditioned by the reasons they get the credit card. The initial reasons can be classified as i) a conscious decision to get a revolving loan, ii) a conscious decision to get a payment tool with a credit limit (can be together with revolving loan), iii) additional banking product/cross-sell proposed to client and accepted by him/her without of previous conscious necessity. In first and second cases

it is expected that the customer will use a credit card for his or her own needs actively, rapidly, and continuously at least for a certain period to satisfy his or her needs. Generally, the usage period for an active credit card is assumed as endless or lifetime. These types of cardholders can have a number of usage scenarios, for instance: i) to make purchases and then pay back full debt amount within the interest rate accrual date (could be end of month or end of grace period), so-called transactors, ii) to make purchases and pay back by monthly payments without significant spending transactions, iii) continual revolving use of the credit limit with frequent purchases and payments but with a positive outstanding balance for a long period, so-called revolvers, iv) to make purchases to the credit limit but do not pay the amount due at the due date and become delinquent, and then, after several months in arrears, to become a defaulted customer.

These typical behavioural patterns can be expanded. For example, such a non-active revolver may use his credit card for purchases infrequently but has a positive outstanding balance. Each customer can have mixed behaviours and transfer between types. These transfers can be caused by: i) necessity of credit funds to address personal liquidity gaps, ii) active periods of spending, for example, systematic periods such as Christmas or a specific period before a wedding, iii) availability or lack of available funds to pay back all required payments (amounts due). This relates to rational customer behaviour. However, the behavioural type can also be caused by irrational reasons from the financial literacy point of view. We can mention several scenarios.

The first example is the following. A credit card holder decides not to pay back a loan for subjective reasons even though he or she has enough funds to cover the required payments. If we do not consider fraud cases, the deliberate non-paying can be caused by both i) irrational behaviour such as temporary mood depression periods, rapid non-loyalty to the bank, and ii) objective factors, which do not depend on the customer, such as a temporary stay in places where bank branches are not located. However, it also can be unintentional action caused by ordinary forgetfulness. We can investigate periodicity of this type of behaviour, but we do not have information on whether a customer has money and decided not to pay, except cases when we have access to customer current accounts. On the other hand, we can see the same type of behaviour

when a customer ‘can’ but ‘do not want’ to pay consciously because he or she has higher priority payments before the bank loan payment.

The second scenario of the irrational behaviour and financial incompetence is that a credit card holder has enough money to cover the outstanding balance, but he or she does not pay back the full amount and pays interest for a long period. This can be caused by the misunderstanding of a credit card holder how to minimize his or her spending and use a credit card rationally, not only as a credit product.

The third example is a customer who has decided not to use the credit card, but keeps it and does not close an account. A customer can simply forget about the credit card. However, it can be that he or she does not need credit funds currently, but relies on the credit card on a rainy day. For this type of a customer, who has got a credit card as a reserve source of funds or even as a gift, the usage type and time of the usage are less predictable in comparison with planned expenditures and cash flows of a customer, who uses a credit card as a payment tool. These credit limits can be rapidly activated after a long inactive period and then closed after several months of usage, or used in full in one month and then default.

So the reasons why a customer choose a particular state and moves to other states can be systematic – general for all customers, and specific – individual for each customer. For instance, the reasons for the credit cardholder state choice can be the following:

- i. for a transactor state – a customer has enough money to pay back the full amount during the first month of the grace period and makes purchases regularly;
- ii. for a revolver state – a customer makes purchases continually or sometimes, but does not have enough money to pay back full amount during the grace period;
- iii. for an inactive state – a customer does not need purchases, or does not have enough money to pay for a loan and understands this;
- iv. for a delinquent state – a customer has already made purchases, but does not have enough money to pay because of financial difficulties or other factors (we consider rational and objective factors; however, irrational and subjective

factors also can be considered as a reason for non-payment, and we consider them in the same model).

These types of behaviour are caused by individual reasons, and factors, which are unobserved directly in our investigation. Generally, data sets collected from customer application and credit card usage history do not contain information about a customer's private life, psychological characteristics, intentions and plans. In the current investigation, we have static socio-demographic and financial application characteristics and dynamic data at the account level about transactions, payments, outstanding balance and delinquencies. Thus we do not have predictors of customer-specific actions and irrational (from the financial literacy point of view) behaviour. We try to guess customer-specific behaviour i) directly – if the customer has had such kind of actions before, however, we do not know reasons for this behaviour, and ii) indirectly – finding a typical behavioural pattern and discovering indirect cross-factors which have a high correlation with these types of behaviour.

A borrower may also default due to hypothetic discounting. Here a borrower adopts a very high time preference rate between t and $t+1$ and borrows heavily. Then in $t+1$ and forwards the individual adopts a lower discount rate. However, the initial borrowing was more than the individual can repay with given income (Meier and Sprenger, 2007).

These cases are difficult for detection and prediction. They are the source of instability in general customer behavioural patterns and decrease the predictive power of a model trained on typical cases. However, these cases exist in real life, and we suppose they take a significant part of all customer behaviour cases, but it is difficult to confirm without special investigation. It is expected that the presence of noise in the transition model caused by unexpected jumps in an account states will reduce the predictive accuracy and stability of the model.

The most typical and stable cases are those that are inactive or revolvers. Transactors are rather unstable and transition between inactive and revolver states. Defaulter is an absorbing state and delinquency is a transitional state to default, but can also be unstable because there are usually a few cases which stay in delinquency for a long time.

The transition from the Transactor to the Revolver state can be caused by the following reasons: i) large debt amount which is higher than available part of the income to pay back the full debt amount, ii) customer income decrease, iii) a motivation from the bank to use credit limit and keep outstanding balance.

The transition from Inactive to Active state (Transactor or Revolver) can be caused by the following reasons: i) needs for money, ii) a motivation from the bank to use the card up to the credit limit

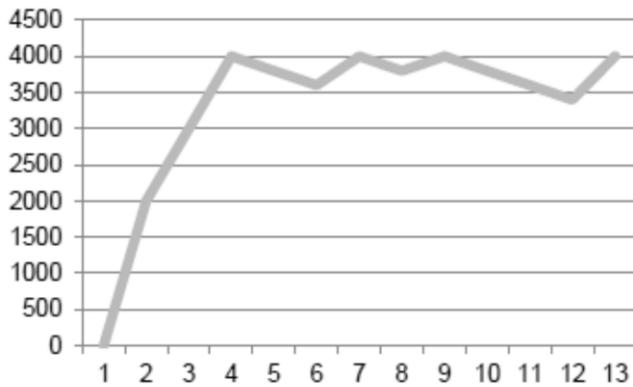
The transition from Revolver to Transactor state can be caused by the following reasons: i) debt amount which is less than available part of the income to pay back full debt amount, ii) customer income increases. The reasons for transition from Revolver to Revolver are the same as for Transactor to Revolver.

The transition from Revolver to Delinquent state can be caused by the following reasons: i) amount due (obligatory payment amount) is higher than available part of income, ii) customer is not able to pay for non-financial reasons such as an extended vacation without access to the bank, accidents.

5.2.3 Graphical presentation of transactor and revolver

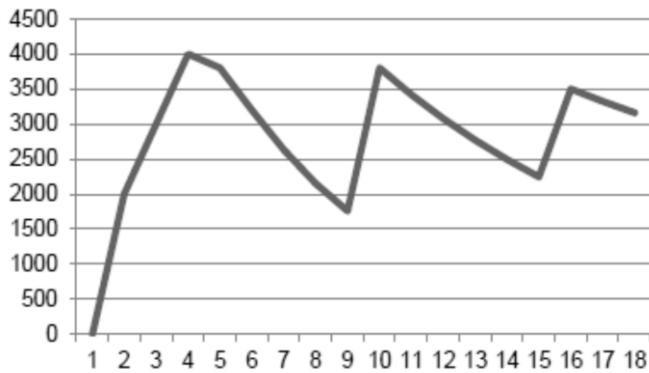
The state's definition can also be shown with the outstanding balance graphs distributed by the month since activation, or month on balance. A revolver can be presented as an active revolve and non-active one. Active revolver (see Figure 5.3) is a type of behaviour when a customer spends up to his/her credit limit, pays back a part of the amount due, and then makes spending again. In Figure 5.3 we have a hypothetical example of the dynamic of the credit card outstanding balance over 13 months, and the outstanding balance value after achieving some level, around 4000 money units, with lower volatile to balance, but the balance does not decrease or increase significantly.

Figure 5.3 Example of the active revolver outstanding balance by months



Non-active revolver (see Figure 5.4) is a type of behaviour when a customer spends up to his/her credit limit and pays back a part of the amount due over several months, and then starts spending again. The borrowing may be repeated.

Figure 5.4 Example of the non-active revolver outstanding balance by months



In our investigation, we do not split revolvers into these two categories. A revolver in our case can show any behaviour type conditional on the positive outstanding balance at the end of the period and no delinquency.

Transactor (see Figure 5.5) can have different values of the outstanding balance within a month but at the end of the period, the outstanding balance must be equal to zero. Figure 5.5 represents an example of the time part of the outstanding balance within a month on a daily basis, and we can see an example of the outstanding balance cycle for three months.

Figure 5.5 Example of the transactor outstanding balance by days

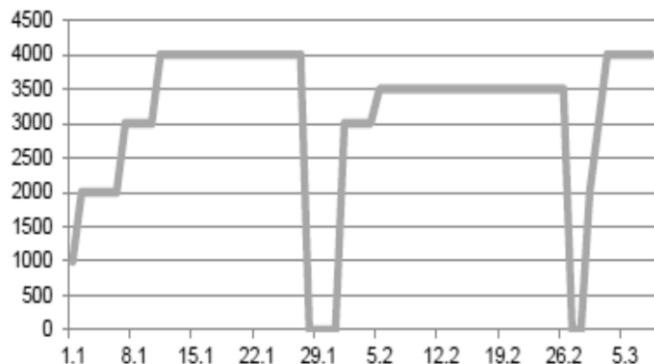
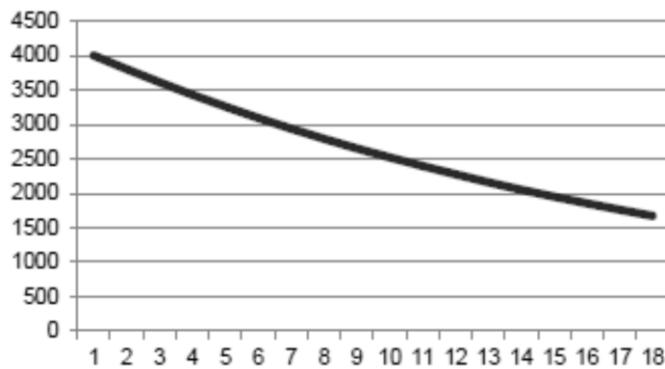


Figure 5.6 Example of the revolve - cash-user outstanding balance by months



One more type of the revolver is rather non-typical credit card user, and is a cash user. This client withdraws money once and then pays back monthly, but does not make any spending (see Figure 5.6). We did not separate this type of behaviour in the current research due to lack of cases with this type of behaviour.

5.3 Model Description

5.3.1 Credit card states and sources of income

The dual function of credit cards and segmentation by types of usage are investigated in a variety of papers. Credit card holders can be divided to those who use and those who do not use a card (Crook et al., 1992), and it is possible to define a specific set of characteristic values for credit card users. Credit card usage can be modelled as low or high with a separate model for each type, for instance, as a two-stage model, where the first stage estimates the desired usage, and at the second stage scoring models are applied for low and high usage prediction (Banasik et al., 2001). Other types of

customer segmentation are Convenient Users, who use credit cards as a payment tool and cardholders, who uses a credit card as a credit instrument – Revolving Users. These two segments have different characteristics to determine the user's type in some papers (Cheu, Loke, 2010; Tan, Steven, Yen, 2011) and also can have different factors affecting outstanding credit card balance (Kim, DeVaney, 2001). The credit card income estimation can correlate with different credit usage types and segments can have individual models. The account segmentation by the level of risk – the probability of default or credit score - is quite common, for example, credit rating transitions (Lando and Scudberg, 2002), segmentation in Markov chains for consumer credit behaviour and ratings (Malik and Thomas, 2012), transition probabilities by delinquency levels (Leow and Crook, 2014). We can see the development of risk modelling to income modelling as, for example, the use of behavioural score for transition matrices in income prediction tasks (So and Thomas, 2011). However, a segmentation not by risk, but by usage types for income modelling is poorly discussed in the literature. Some investigations show that the transactor/revolver split can give more accurate income estimates in comparison with the standard Good/Bad approach (So et al., 2014).

The states of the credit also can be classified into the mover-stayer dimensions. ‘Stayers are those who pay off their debt fully each month and so always remain in the highest good state. Movers are customers whose payment history are more varied, including partial and missing payments’ (Thomas et al., 2001, p.97). These definitions correspond with transactors for stayers and revolver and delinquent for movers in our research. Mover and stayer segmentation can be used, for example, for states in transition matrices (Hand & Till, 2003), where stayer is a customer who stays at the initial state and mover is a customer with transitions between states. It corresponds with stable and unstable states in our model.

Generally (Thomas et al., 2001; So and Thomas, 2011) risk management approaches define delinquent and non-delinquent account buckets as the following: current, up to 30 days past due (DPD) - Bucket 1, DPD 31-60 - Bucket 2, DPD 61-90 - Bucket 3, and default. Those accounts in the current state may differentiate by the level of risk, or score. The current state is often split up by ratings or risk level, in particular, with the use of behavioural score, for instance, ‘low risk’ with probability of default less

than some low threshold, ‘moderate risk’ with probabilities of default between low and high thresholds, and ‘high risk’ with probability of default more than some high threshold. The number of risk bands, or ratings, can vary but is usually from 5 to 10. Because the aim of our investigation is the income prediction of credit card usage, we propose to define the credit card statuses subject to the revenue source and the revenue availability.

The revenue is generated from two sources: i) interest revenue, and ii) transaction revenue. Interest revenue is generated as a percentage of active outstanding balance, usually for monthly accruals and is taken as an average for each month. Transaction revenue is generated from debit transactions as interchange fees from purchases and ATM usage fees from cash withdrawal.

The important point of the model building conception is what time horizon is used for prediction. In one month an account can jump from the current state t to any state, possible for transition from the current state. However, for longer period an account can move from the current state to any state. For example, an inactive account can become a default after 5 months, or transitions, with moving scenario as ‘inactive’ – ‘revolver’ – ‘delinquent month 1’ – ‘delinquent month 2’ – ‘delinquent month 3’ – ‘default’. So if we talk about Expected Loss or Revenue, it is necessary to define what period is used for prediction. Expected Loss estimation for inactive, transactor, revolver, delinquent month1 and delinquent month 2 for the month $t+1$ is not applicable. However, if we talk, for example, about a 6-month horizon any of the mentioned states is possible.

A revolver account generates both the interest rate and transaction revenues, and the shortest period it is able to generate losses if the default state is defined as 4 or more missed payments is between period 0 and period $t+4$. In this case a revolver from the current states must go forward in delinquency states as from 0 delinquency bucket to 1, from 1 to 2, from 2 to 3, and finally from 3 to 4. An inactive account does not generate an income because of it has no active outstanding balance and no debit transactions, and requires one more month in comparison with a revolver state to generate losses for period $t+5$ because it must move to the revolver or transactor states at the first stage. A transactor account generates transaction revenue only. A 1, 2, and

3 month delinquency account can generate transaction revenue and a penalty. However, the penalty is accrued in the current period but earned after the resumption of payments. A 1 month delinquency generates risk for $t+3$ because an account for transition from up to 30 days past due (1 missed payment) to 90 or more days past due (4 or more missed payments) must not pay at least 3 payments and move from the state 1 to 2, from 2 to 3 and finally from 3 to 4.

The probability of default can be estimated for any, state in two ways: first, the N -stage prediction model for $t+i$, where N is a number of transitions to the default state from the current state at time t , where i is the ordered number of the state between current and N , and second, the one-stage predictive model for $t+N$ horizon. However, we mentioned the minimum number of steps to get to the default state, and for longer periods it is possible a transition with different ways between possible states to get to the default state. For example, if we use a 6 month period for prediction, an account in 3-month delinquency state can recover to the revolver and then again go to delinquency and default in month 5.

The system of states definitions and related assessment is presented in Table 5.1. Note that the average outstanding balance (OB) means the average outstanding balance for the month calculated as the sum of daily balances divided by the number of days. For a transactor definition, we use OB_eop – the outstanding balance at the end of period (month). If a customer has ‘decided’ to be a transactor, (s)he must have a zero balance at the end of the period and positive debit transactions (purchases or money withdrawals) during the month. DPD means Days Past Due counter and equal to the number of days since the first due date when the required payment has been missed. Generally and in our observation data set, the due date is set up monthly. So the number of missing payments is equal to the number of months with non-payments. A number of days in a month is different, and for convenience, it is usually taken as 30. We define a 2 month of delinquency state as 1st month DPD between 1 and 30, 2nd month DPD between 31 and 60. After 61st day of delinquency, an account is counted as defaulted. The default state is absorbing. This means that an account is not able to return to any of the other states and stays in default. For modelling purposes we are able to use any default definition on the assumption of business logic and data sample limitations and characteristics. We decided to use 60+ DPD as our default definition, or 3 and more

delinquent payments, because of lack of cases for 61-90 DPD, or a 3 month delinquency state.

Table 5.1 Account state definition and related assessments

Account status	Symbol	Definition	Risk assessment	Revenue assessment	Note
closed	C	The account is closed or inactive for more than 6 months	No	No	Excluded from the analysis
inactive	NA	Average OB = 0 and Debit Turnover Amount = 0	No	No	Expected Loss (EL) can be estimated with state transitions
transactor	TR	OB_eop = 0 and Debit Turnover Amount (purchase) > 0	No	Debit Transactions Amount x Transaction Profit Rate	TR Profit Rate = (avg interchange rate + fees rate) EL – see inactive note
revolver (current)	RE	Average OB > 0 and DPD = 0	Behavioural (transition probability) score for current	Limit x Utilization Rate x Interest Rate + Debit Transactions Amount x Transaction Profit Rate	-
delinquent	DL	Average OB > 0 and (DPD > 0 and DPD <=60)	Behavioural (transition probability) score for delinquent	No	If credit card is not blocked, the transaction income exists
default	D	Average OB > 0 and DPD > 60	LGD	-	Recovery is not an income.

Accounts with status ‘Closed’ are excluded from the current research, so we investigate only active and inactive, but open accounts. An inactive is an account which has zero average outstanding balance in the month before the observation point and does not have any purchase transactions for the observation month.

A transactor is an account which has zero outstanding balance at the end of the month and positive debit turnover amount (purchases transactions) during the month. A transactor does not generate a risk in a sense a transactor account is not able to go to the delinquent state. A transactor generates transactional income, which can be computed as Debit Transactions Amount x Transaction Income Rate.

A revolver is an account which has a positive average outstanding balance for the month and zero days past due. A revolver generates a credit risk, and it can be estimated as a probability of transition to the delinquent state. A revolver generates transactional and interest income which can be computed as Limit x Utilization Rate x Interest Rate + Debit Transactions Amount x Transaction Income Rate.

A delinquent is an account which has positive average outstanding balance for the month and positive days past due. A delinquent generates a credit risk, and it can be estimated as a probability of transition to the default state. A delinquent account may generate a transactional income, and we assume that it is not able to generate interest income.

A defaulted account is an account with DPD counter more than 60 days and positive, outstanding balance. In our model, we assume for simplification that the Loss given default equal to 1, so the Expected Losses are equal to the outstanding balance at the time of default.

An account can move from its current state into any of $N-1$ other states when the number of time periods is large, where N is the number of states. The number of transition probabilities is $N-1$, where N is the number of states. For a standard scoring model such as the probability of default estimation, we need the model for only one probability. For example, the probability of moving to default state is p . Then the probability to stay in non-default state is $1-p$. However, in our model of the credit card holder's behaviour the number of states, which an account can move between, is more than two, for example, a revolver can move to transactor, delinquent, and inactive states, or stay a revolver. Thus it is necessary to estimate the set of $N_{s,t+1}-1$ transition probabilities p_j and

$$\sum_{j=1}^{N_{s,t+1}-1} p_j = 1,$$

where $N_{s,t+1}$ is the number of states, which account can move from the state s at time $t+1$.

An account in each state except inactive and defaulted can generate an income. However, the sources of income are different. For instance, a delinquent account can generate non-interest income due to interchange fees from merchants and penalty but does not generate an interest income because of non-paid debt. For a portfolio of active credit cardholders the main expected state is a revolver, and it is expected that the majority of observations will be in the revolver state.

For the segmentation of the current or revolver state by the level of risk, behavioural scoring can be used (Thomas et al., 2012). For transition probabilities estimation we consider that detailed segmentation, can give more accurate prediction results. However, for income forecast at the account level, we believe the 5 states is enough considering the business logic of the income accruals. However, we split the delinquent state into the first month of delinquency and the second month of delinquency for transition probabilities prediction to be consistent with migration matrices and keep the logic of moving an account from current to defaulted state. The second part of transition probabilities modelling shows that the separation of the additional segment from transactor state to revolver sub-state brings new logic into the transactor state definition (see Section 6.5 of Chapter 6).

5.3.2 The estimation of the required number of models

We defined that a credit card account can be at least in one of five states such as inactive, transactor, revolver, delinquent, and default. Transition to any states can be done from four states: inactive, transactor, revolver, and delinquent. Thus each state requires a model or a set of models for the estimation of the transition probabilities. There are two concepts of how many models we need. The first one is to build a single model for each state, which estimates a group of probabilities of transition to possible states. The second approach is to build a single model for the estimation of the probability of each possible transition. For the first approach, we can use, for example, the multinomial regression or ordinal logistic regression, which are discussed in sections 5.4.3 and 5.4.4. These methods require fewer models for computation (see Table 5.2) than, for example, multistage binary logistic regression, which give a separate estimation for each transition (see Table 5.3).

Table 5.2 Multinomial logistic regression models covering

Status	To					
From	Non Active	Transactor	Revolver	Delinquent 1	Delinquent 2	Defaulted
Inactive	Model NA			X	X	X
Transactor	Model Tr			X	X	X
Revolver	Model Re			X	X	
Delinquent 1	Model D1					X
Delinquent2	Model D2					
Default	X	X	X	X	X	X

A single multinomial regression model covers all possible transitions from an original state, which are shown with a shaded cell in Table 5.2. The model NA for the inactive state covers transitions to inactive, transactor, and revolver states. The model Tr for transactor state covers transitions to inactive, transactor, and revolver states. The model Re for revolver state covers transitions to inactive, transactor, revolver, and delinquent 1 (DPD 1-30) states. The model D1 for delinquent 1 state covers transitions to inactive, transactor, revolver, delinquent 1 and delinquent 2 (DPD 31-60) states. The model D2 for delinquent 2 state covers transitions to inactive, transactor, revolver, delinquent 1, and delinquent 2, and default states. Because the default is absorbing state and according to assumptions of our model it is impossible to return to any other state from the default state, the model for the default state is not applied.

Table 5.3. Multi-stage logistic regression models covering

Status	To					
From	Non Active	Transactor	Revolver	Delinquent 1	Delinquent 2	Defaulted
Non Active	Model_NA_NA	Model_NA_Tr	Model_NA_Re	X	X	X
Transactor	Model_Tr_NA	Model_Tr_Tr	Model_Tr_Re	X	X	X
Revolver	Model_Re_NA	Model_Re_Tr	Model_Re_Re	Model_Re_D1	X	X
Delinquent 1	Model_D1_NA	Model_D1_Tr	Model_D1_Re	Model_D1_D1	Model_D1_D2	X
Delinquent 2	Model_D2_NA	Model_D2_Tr	Model_D2_Re	Model_D2_D1	Model_D2_D2	Model_D2_Df
Defaulted	X	X	X	X	X	X

If we use an assumption that for each original state and each destination state the transition probabilities regression equation may have different slopes and trends for the predictors, then it is necessary to build N-1 destination models for each original

state, where N is a number of possible transitions from the current state. The required models are shown with shaded cells in Table 5.3. Non-shaded cells with models' names mean that the separate estimations are not required for these particular segments because of the full set of N transitions can be explained with N-1 models. Cells marked with 'X' indicate an impossible transition. For example, transactor can transit to be a non-active, a transactor, and a revolver, but cannot transit to the delinquent or defaulted states at the next period and does not need a prediction model as it is shown by cells with 'X' in Table 5.3.

Non-shaded cells with the name of the models show that the appropriate model can be built for the transition prediction, but also can be calculated from other probabilities with a full set of events formula. For example, the probability of a transition from the non-active state to the revolver state can be calculated one minus the probability of the transition to inactive minus probability of transition to transactor state:

$$P(s_{i,t+1} = RE | s_t = NA) = 1 - P(s_{i,t+1} = NA | s_t = NA) - P(s_{i,t+1} = TR | s_t = NA) \quad (5.1)$$

where

s_t is the current time point,

$s_{i,t}$ is the state of account i at time t ,

$P(s_{i,t+1} = RE | s_{i,t} = NA)$ is the probability of transition of account i , which is in the inactive state ($s_t = NA$) at time t to the revolver state ($s_{t+1} = RE$) at time $t+1$;

$P(s_{i,t+1} = NA | s_t = NA)$ is the probability that an account i stays inactive at time $t+1$;

$P(s_{i,t+1} = TR | s_t = NA)$ is the probability that an account i moves from the inactive state ($s_{t+1} = NA$) to the transactor state ($s_{t+1} = TR$).

We give a detailed description of the states and transition model in section 5.4.

5.4 Transition states model

5.4.1 General model description

A credit card account at time t can be in one of J_N possible states $S_t = \{J_1, J_2, \dots, J_j, \dots, J_N\}$, where N is a number of states.

We try to predict the probability of transition of account i from state $S_{i,t}$ at time t to state $S_{i,t+1}$ at time $t+1$. The state at time t is known. At time $t+1$ the account can be in any of a subset of possible states. The set of possible transitions is conditional on current state at time t :

$$\begin{aligned} S_{i,t+1} &= \{NA, Tr, Re \mid S_{i,t} = NA\} \\ S_{i,t+1} &= \{NA, Tr, Re \mid S_{i,t} = Tr\} \\ S_{i,t+1} &= \{NA, Tr, Re, D1 \mid S_{i,t} = Re\} \\ S_{i,t+1} &= \{Tr, Re, D1, D2 \mid S_{i,t} = D1\} \\ S_{i,t+1} &= \{Tr, Re, D1, D2, Df \mid S_{i,t} = D2\} \end{aligned} \quad (5.2)$$

The common transition prediction model estimates the probability of the event that the account will jump to state s depending on customer static predictors such as application characteristics, and on dynamic predictors such as account behavioural characteristics and on the history of previous states:

$$\Pr(S_{i,t+1} = \{J_1, J_2, \dots, J_j, \dots, J_N\} \mid S_{i,t} = j_{i,t}) = f(\mathbf{x}_i, \mathbf{x}_{i,t-T}, s_{i,t-T}) \quad (5.3)$$

\mathbf{x}_i – vector of predictors for an account i not varying in time (application characteristics)

$\mathbf{x}_{i,t}$ – vector of predictors for an account i varying in time (behavioural characteristics)

$S_{i,t}$ – state of the account i at time $t-T$, where T is time lag for predictors

We use several retrospective periods in one model to take into account possible changes in customer behaviour and use them as a set of predictors. However, this can cause multicollinearity and overlapping because we use panel data with 1 month step and predict for 1 month into the future. The following model includes 3 months history of behavioural characteristics \mathbf{x} and states as of one, two, and three months (at t , $t-1$, and $t-2$ respectively) before the predicted state at $t+1$:

$$\Pr(S_{i,t+1} = \{J_1, J_2, \dots, J_j, \dots, J_N\} \mid S_{i,t} = j_{i,t}) = f(\mathbf{x}_i, \mathbf{x}_{i,t}, \mathbf{x}_{i,t-1}, \mathbf{x}_{i,t-2}, s_{i,t}, s_{i,t-1}, s_{i,t-2}) \quad (5.4)$$

In this case, overlapping of behaviour period for predictors and multicollinearity can be observed. We explain this as follows. For example, we use monthly panel data as a development sample and take behavioural predictors for 3 months to build a prediction for the fourth month. Thus from 1 year we can take behavioural characteristics for the same account calculated for January, February, and March for April prediction, February, March, and April for May prediction, March, April, and May for June prediction etc. However, in this example, March behaviour is considered in the parameter estimation as the pool of three slices three times, and February has gotten into the pool two times. The new month, when the observation window slides at the next period, add only one-third of information to the aggregated characteristic such as the ratio of average outstanding balance for the last three month to average outstanding balance to months 4-6, the ratio of the sum of debit transactions to the sum of credit transaction for the last three month. For example, the first pool of months for the computation of behavioural characteristics is February, March, and April, and the second pool of months is March, April, and May. Predictors calculated as an average or the sum of three months characteristics will be highly correlated for close observation points and one characteristic from the next period (or slice) can be linearly predicted from the previous period (or slice). This means high multicollinearity.

To avoid multicollinearity, we try to use a single month's history for each model as follows:

$$\Pr(S_{i,t+1} = \{J_1, J_2, \dots, J_N\} | S_{i,t} = j_{i,t}) = f(\mathbf{x}_i, \mathbf{x}_{i,t}) \quad (5.5)$$

However, this model loses a part of important behavioural information about previous values of ϕ characteristic. Thus we try to compare two sets of the predictor. First, aggregated behavioural characteristics such as averages, maxima, ratios calculated on raw data for several months periods, for example, the ratio of the maximum purchase amount to the average outstanding balance for the last three months. Second, raw data such as outstanding balances, arrears amounts etc. are used directly and their ratios for a one month period, for example, purchase amount to the outstanding balance in month t .

The probability that account i will be at state S from the set of states J_1, \dots, J_N at time $t+1$ is conditional on the current state S_t and depends on vectors of static predictors \mathbf{x}_i

as application characteristics which are not time-varying and behavioural predictors x_{it} as outstanding balance, arrears amount etc. and/or their aggregates and derivative functions, which are time-varying.

In this research the transition models are applied for the prediction of income and losses for 1 month period and longer. The transition probabilities for $t+2, t+3, \dots, t+T$, where T is a maximum number of months used for the prediction of the state's transition, are explained with the same set of predictors and are also conditional on the current state S_t , but require an individual model for each prediction horizon such as follows:

$$\begin{aligned} \Pr(S_{i,t+2} = \{J_1, J_2, \dots, J_N\} | S_{i,t} = j_{i,t}) &= f_2(\mathbf{x}_i, \mathbf{x}_{i,t}, \mathbf{x}_{i,t-1}, \mathbf{x}_{i,t-2}, s_{i,t}, s_{i,t-1}, s_{i,t-2}) \\ \Pr(S_{i,t+3} = \{J_1, J_2, \dots, J_N\} | S_{i,t} = j_{i,t}) &= f_3(\mathbf{x}_i, \mathbf{x}_{i,t}, \mathbf{x}_{i,t-1}, \mathbf{x}_{i,t-2}, s_{i,t}, s_{i,t-1}, s_{i,t-2}) \\ \Pr(S_{i,t+T} = \{J_1, J_2, \dots, J_N\} | S_{i,t} = j_{i,t}) &= f_T(\mathbf{x}_i, \mathbf{x}_{i,t}, \mathbf{x}_{i,t-1}, \mathbf{x}_{i,t-2}, s_{i,t}, s_{i,t-1}, s_{i,t-2}) \end{aligned} \quad (5.6)$$

To avoid multicollinearity, we also try to build a model with predictors for only one-period t . The set of the models is as follows:

$$\begin{aligned} \Pr(S_{i,t+1} = \{J_1, J_2, \dots, J_N\} | S_{i,t} = j_{i,t}) &= f(\mathbf{x}_i, \mathbf{x}_{i,t}, s_{i,t}) \\ \Pr(S_{i,t+2} = \{J_1, J_2, \dots, J_N\} | S_{i,t} = j_{i,t}) &= f_2(\mathbf{x}_i, \mathbf{x}_{i,t}, s_{i,t}) \\ \Pr(S_{i,t+T} = \{J_1, J_2, \dots, J_N\} | S_{i,t} = j_{i,t}) &= f_T(\mathbf{x}_i, \mathbf{x}_{i,t}, s_{i,t}) \end{aligned} \quad (5.7)$$

In the section 5.3.1 of this chapter, we have introduced the set of states of an account transition. Using letters definitions, the system of transition probabilities models from the current state S_t to the state in the next month S_{t+1} , which cover the full set of transitions, can be written in general form as follows:

$$\begin{aligned} \Pr(S_{i,t+1} = \{NA, Tr, Re\} | S_{i,t} = NA) &= f_{NA}(\mathbf{x}_i; \mathbf{x}_{i,t}, \dots, \mathbf{x}_{i,t-T}; s_{i,t}, \dots, s_{i,t-T}) \\ \Pr(S_{i,t+1} = \{NA, Tr, Re\} | S_{i,t} = Tr) &= f_{Tr}(\mathbf{x}_i; \mathbf{x}_{i,t}, \dots, \mathbf{x}_{i,t-T}; s_{i,t}, \dots, s_{i,t-T}) \\ \Pr(S_{i,t+1} = \{NA, Tr, Re, Dl\} | S_{i,t} = Re) &= f_{RE}(\mathbf{x}_i; \mathbf{x}_{i,t}, \dots, \mathbf{x}_{i,t-T}; s_{i,t}, \dots, s_{i,t-T}) \\ \Pr(S_{i,t+1} = \{Tr, Re, Dl, D2\} | S_{i,t} = Dl) &= f_{Dl}(\mathbf{x}_i; \mathbf{x}_{i,t}, \dots, \mathbf{x}_{i,t-T}; s_{i,t}, \dots, s_{i,t-T}) \\ \Pr(S_{i,t+1} = \{Tr, Re, Dl, D2, Df\} | S_{i,t} = D2) &= f_{D2}(\mathbf{x}_i; \mathbf{x}_{i,t}, \dots, \mathbf{x}_{i,t-T}; s_{i,t}, \dots, s_{i,t-T}) \end{aligned} \quad (5.8)$$

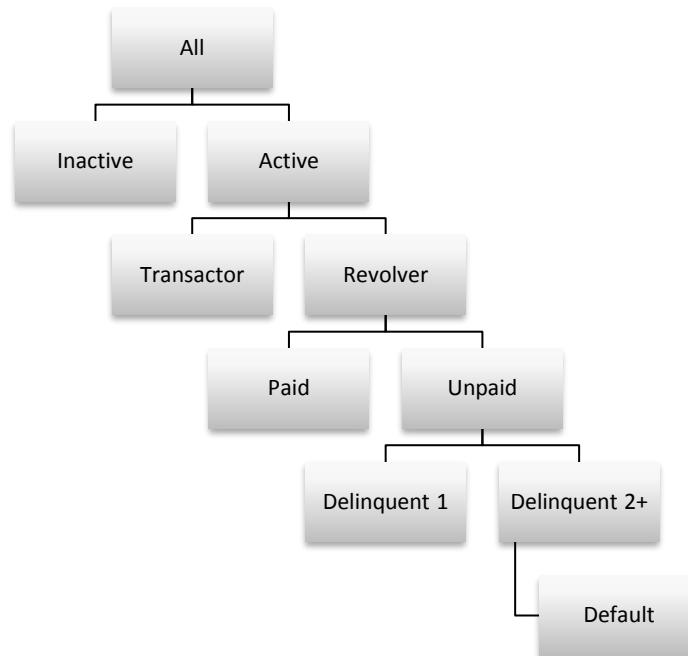
where $\mathbf{x}_i; \mathbf{x}_{i,t}, \dots, \mathbf{x}_{i,t-T}; s_{i,t}, \dots, s_{i,t-T}$ are the set of three groups of predictors: not depending on time-applications characteristics, time-varying – behavioural characteristics, and states;

T – is a maximum number of retrospective months used for predictors.

5.4.2 Model 1 – Decision tree of the conditional logistic regressions with a binary target

The problem of the transition probabilities for the set of states can be presented as a binary decision tree where a number of leaves are equal to the number of states N and number of transition models is N-1. The result of the regression analysis is a set of the conditional logistic regressions each with a binary target. The general model can be presented as a binary tree (see Figure 5.7).

Figure 5.7 Multistage schema of the conditional logistic regression models



At each stage, we predict the probability of transition $\Pr(s_{t+1} = \bullet)$ to one of the states at the next level conditional on the state at the high level. For a full description of all stages, we need six equations for non-active, transactor, revolver, delinquent 1, delinquent 2, and default states.

$$\Pr(s_{t+1} = NA) = \boldsymbol{\beta}_{NA}^T \mathbf{x} \quad (5.9)$$

$$\frac{\Pr(s_{t+1} = TR)}{1 - \Pr(s_{t+1} = NA)} = \beta_{TR}^T \mathbf{x} \quad (5.10)$$

$$\frac{\Pr(s_{t+1} = RE)}{(1 - \Pr(s_{t+1} = TR))(1 - \Pr(s_{t+1} = NA))} = \beta_{RE}^T \mathbf{x} \quad (5.11)$$

$$\frac{\Pr(s_{t+1} = RP | s_t \in RE)}{(1 - \Pr(s_{t+1} = TR))(1 - \Pr(s_{t+1} = NA))(1 - \Pr(s_{t+1} = RE))} = \beta_{RP}^T \mathbf{x} \quad (5.12)$$

$$\frac{\Pr(s_{t+1} = D1 | s_t \notin (NA, TR))}{(1 - \Pr(s_{t+1} = NA))(1 - \Pr(s_{t+1} = TR))(1 - \Pr(s_{t+1} = RE))} = \beta_{D1}^T \mathbf{x} \quad (5.13)$$

$$\frac{\Pr(s_{t+1} = D2 | s_t \notin (NA, TR))}{(1 - \Pr(s_{t+1} = NA))(1 - \Pr(s_{t+1} = TR))(1 - \Pr(s_{t+1} = RE))(1 - \Pr(s_{t+1} = D1))} = \beta_{D2}^T \mathbf{x} \quad (5.14)$$

$$\begin{aligned} & \frac{\Pr(s_{t+1} = DF | s_t \notin (NA, TR, RE, D1))}{(1 - \Pr(s_{t+1} = NA))(1 - \Pr(s_{t+1} = TR))(1 - \Pr(s_{t+1} = RE))(1 - \Pr(s_{t+1} = D1))(1 - \Pr(s_{t+1} = D2))} = \\ & = \beta_{DF}^T \mathbf{x} \end{aligned} \quad (5.15)$$

where

β is a vector of coefficients for the related state index for a transactor (TR), a revolver (RE), a revolver paid (RP), a delinquent 1-30 (D1), a delinquent 31-60 (D2), and a default (DF).

\mathbf{x} is a vector of covariates,

s_t is the account state at the observation point;

s_{t+1} is the account state at the performance point.

For the full set of states, we added the Revolver Paid state (5.12) to take into the account a transition from revolver state and do not mix it with real transactor state. It is necessary for further income prediction because transactor obligatory has a zero balance at the end of a period and a revolver may have a zero balance in the case (s)he pays back the full amount this month, but a transactor does not generate interest income, and the revolver does.

Current conditional probability equations are defined for a transition from non-active state to defaulted state as a possible evolution of an active credit account. Conditional probabilities of the transition to the following state are calculated depending on the current state. For example, the customer can move to revolver state from inactive, transactor, delinquent 1, and delinquent 2 states, except absorbing default state.

$$\Pr(s_{t+1} = TR) = \beta_{TR}^T \mathbf{x} \quad (5.16)$$

$$\Pr(s_{t+1} = RE) = \beta_{RE}^T \mathbf{x} \quad (5.17)$$

$$\Pr(s_{t+1} = RP | s_t \in RE) = \beta_{RP}^T \mathbf{x} \quad (5.18)$$

$$\Pr(s_{t+1} = D1 | s_t \notin \{NA, TR, RP\}) = \beta_{D1}^T \mathbf{x} \quad (5.19)$$

$$\Pr(s_{t+1} = D2 | s_t \notin \{NA, TR, RE, RP\}) = \beta_{D2}^T \mathbf{x} \quad (5.20)$$

$$\Pr(s_{t+1} = DF | s_t \notin \{NA, TR, RE, RP, D1\}) = \beta_{DF}^T \mathbf{x} \quad (5.21)$$

In this model, we use binary logistic regression conditional on the current state for multistage decision tree building. Generally, logistic regression fits the log of the probability odds to a linear combination of the characteristic variables as

$$\text{logit}(p_i) = \ln\left(\frac{p_i}{1 - p_i}\right) = \beta_0 + \boldsymbol{\beta} \cdot \mathbf{x}_i^T, \quad (5.22)$$

where p_i is the probability of an outcome, β_0 is an intercept and $\boldsymbol{\beta}$ is a column vector of regression coefficients, and \mathbf{x} is a column vector of predictors.

The probability of an event indicated by Y_i for the i^{th} observation is calculated as

$$P_i = E(Y_i | \mathbf{x}_i, \boldsymbol{\beta}) = \Pr(Y_i = 1 | \mathbf{x}_i, \boldsymbol{\beta}) \quad (5.23)$$

For the estimation of coefficients, we used the binary logistic regression stepwise method with significance levels for entering effects and for removing effects equal to 0.1. Generally the significance level value equal to 0.05 is recommended for usage. However, we applied a less strict condition to add more predictors because we want to test the impact of various covariates of the target variable.

We build a Multistage Binary Model with predictors based on several periods. For the current investigation, we have chosen three months for the behavioural characteristics calculation. As we try to predict the transition for the period of 1 month, we assume that the period for transition drivers would be enough for the trial model which we use for regression methods comparative analysis. However, for the final model, we will use a longer period of 6 months. Thus in the current model the state S at time t+1 depends on behavioural characteristics at time t, t-1, and t-2. Binary regression is used to predict the transition probabilities in multistate.

For example, we have 3 states: A, B, and C. We need to estimate the probability of transition to all states with a binary model. Thus we need to use two groups of target variables. For a full description we must use all possible states for transition at the current step, so we are not able to estimate the probability of A and B only, but must take into account C cases too. So we can merge B and C in a single state and get binary model with dichotomic outcome. At the first stage we run the model:

$$\Pr(A) = \ln\left(\frac{P(A)}{1 - P(A)}\right) = \ln\left(\frac{P(A)}{P(B) + P(C)}\right) = \mathbf{b} \cdot \mathbf{x}^T$$

For the first step we operate with two states: $A, (B \cup C)$

$$\Pr(Y_i = 0) = \frac{e^{\beta_0^T x_i}}{e^{\beta_0^T x_i} + e^{\beta_1^T x_i}}$$

$$\Pr(Y_i = 1) = \frac{e^{\beta_1^T x_i}}{e^{\beta_0^T x_i} + e^{\beta_1^T x_i}}$$

The probability of transition to the first state is predicted at stage 1. Then for the remaining states, which are not in state 1, the probability of transfer to state 2 is estimated. For 4 states we apply a 3 stage model. The order of stages can be selected in different ways depending on business logic. For example, ‘delinquent month 2’ state follows ‘delinquent months 1’ state if the customer has not paid the amount due. However, the opposite backward transition from Delinquency 2 to Delinquency 1 state is also possible if a customer with two delinquent payments pays back only one payment in arrears.

The basic model predicts the probability of transitions to the full set of possible states. Basic model means the N-1 order of a number of probabilities required for a full set of states description. If the set of states consists of 3 states A,B, and C it is necessary to have two probabilities estimations $\text{Pr}(A)$ and $\text{Pr}(B)$, and the probability for state C will be calculated as

$$\text{Pr}(Y = C) = 1 - (\text{Pr}(A) + \text{Pr}(B))$$

5.4.3 Model 2 – Ordinal logistic regression with the non-binary target

A dependent variable which has more than two discrete possible values can also be predicted with a single model with the non-binary target. Depending on whether the outcome categories are ordered or unordered the complicated procedure of multilevel decision tree building can be avoided with *ordered (ordinal) logistic regression* and multinomial logistic regression.

In the case of an ordered outcome variable, the ordered logistic regression can be applied, and the structural equation can be defined in general form as $Y_i^* = \beta^T X_i + \varepsilon_i$ with

$$Y_i = \begin{cases} 1 & \text{if } Y_i^* \leq \mu_0 \\ 2 & \text{if } \mu_0 < Y_i^* \leq \mu_1 \\ 3 & \text{if } \mu_1 < Y_i^* \leq \mu_2 \\ \dots \\ N & \text{if } \mu_N < Y_i^* \end{cases} \quad (5.24)$$

where

Y_i is the observed state for observation i where Y_i takes ordinal values as follows: state NA, state TR, state RE, state D1, state D2, and state DF;

Y_i^* is an unobserved dependent variable (the exact level of agreement with the statement proposed),

X_i is a vector of variables that explains the probability of the particular transition by case i ;

i as a case;

β is a vector of coefficients;

μ_k are the threshold parameters to be estimated along with β ;

and ε_i is a disturbance term that is assumed normally distributed.

The threshold parameters are given by

$$c_k(\mathbf{x}) = \ln \frac{P(Y \leq j | \mathbf{x})}{P(Y > j | \mathbf{x})} = \ln \frac{\varphi_0(\mathbf{x}) + \varphi_1(\mathbf{x}) + \dots + \varphi_j(\mathbf{x})}{\varphi_{j+1}(\mathbf{x}) + \varphi_{j+2}(\mathbf{x}) + \dots + \varphi_N(\mathbf{x})} = \tau_j - \mathbf{x}^T \boldsymbol{\beta} \quad (5.25)$$

where

τ_j are the cut points between the categories,

$\varphi_j(\mathbf{x})$ is the probability of being in class j given covariates \mathbf{x} .

The final parameter estimation is a system of equations:

$$\begin{aligned} \ln \frac{Pr(Y_i=1)}{Pr(Y_i=N)} &= \boldsymbol{\beta}_1^T \cdot \mathbf{X}_i \\ \ln \frac{Pr(Y_i=1)}{Pr(Y_i=N)} &= \boldsymbol{\beta}_1^T \cdot \mathbf{X}_i \\ \ln \frac{Pr(Y_i=N-1)}{Pr(Y_i=N)} &= \boldsymbol{\beta}_{N-1}^T \cdot \mathbf{X}_i \end{aligned} \quad (5.26)$$

where N is the number of possible values for the outcome.

5.4.4 Model 3 – Multinomial logistic regression with the non-binary target

In the case of an unordered response variable, multinomial logistic regression can be applied to target prediction. A single equation ordered logistic regression returns a value in the range between 0 and 1, and the response is assigned depending on the interval defined by thresholds, or cut-offs. In the case of multinomial logistic regression we need $N-1$ models, where N is a number of categories (or possible values of the target), to compare each category to a reference category. The multinomial logistic regression model compares a number of dichotomies. Binary logistic regression is a particular case of the multinomial logistic regression model which compares only one dichotomy.

The probability of default is estimated as follows:

$$Pr(Y_i = j) = \frac{e^{\boldsymbol{\beta}^T \mathbf{x}_{ij}}}{\sum_{j=1}^J e^{\boldsymbol{\beta}^T \mathbf{x}_{ij}}} \quad (5.27)$$

where j is a state, J is the number of states, $\boldsymbol{\beta}$ is coefficients of regression, x is characteristics for case i in j state.

One of the applications of multinomial regression for credit card usage states modelling has been proposed by Volker (1982). He defined four types of card usage (hold bankcard, use credit, use regularly, and use moderately) and compared how the same set of predictors (age, professional skills, marital status, region of residence etc.) impacts on the customer probability to obtain one of the mentioned statuses.

The model of credit card usage types is defined as

$$\ln \left(\frac{P_{ij}}{P_{i1}} \right) = \beta_i^T X_i \quad (5.28)$$

'where P_{ij} is the probability of individual i selecting alternative j , normalisation is to the first alternative, X is a vector of explanatory variables, and β_j is a set of alternative-specific coefficients' (Volker, 1982). So the model is presented as a prediction of the probability of each usage type compared with that for a non-holding bank card.

A hypothesis to be investigated is that the multistage binary regression predicts transition probabilities less accurately than ordinal and multinomial logistic regression for all sets of outcomes. However, conditional binary logistic regression can give stronger results for particular segments, for example, if estimated at the first level of the conditional tree. The estimation results for multinomial and ordinal regressions can predict less accurately than the binary logistic regression. However, if we consider many outcomes and need to build a multilevel decision tree, multinomial regression can give more confident estimates. The residuals (errors) on the tails of distribution for some untypical cases can be higher, but generally the model is less robust than a multistage binary logistic regression. The estimations for the multinomial logistic regression are given on the same data set simultaneously. On the other hand, the multistage binary logistic regression considers a certain order of the stages, which can be defined based on the business logic or statistical issues, and the set of conditional logistic regressions is built on the data subsamples, which are dependent on the order of stages.

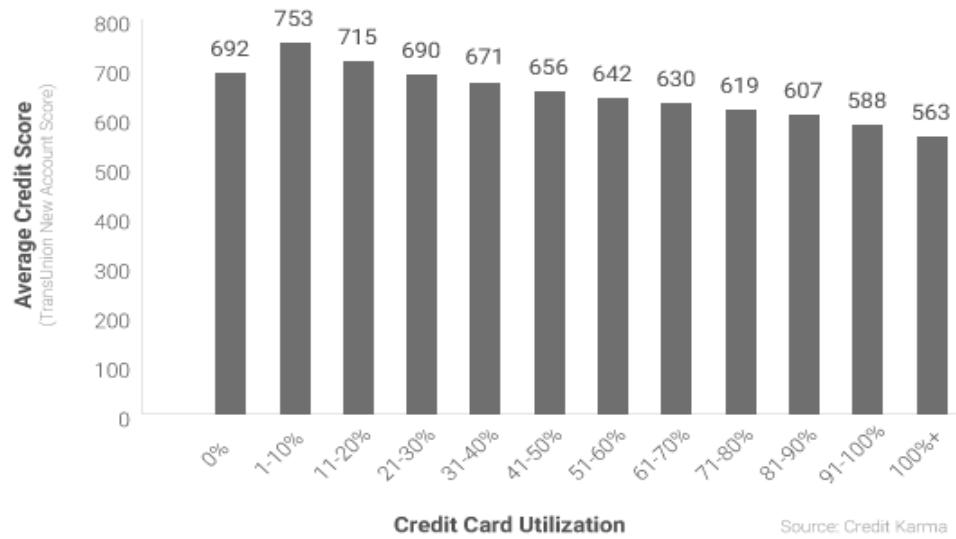
Ordinal logistic regression takes into account the order of outcomes and can give better predictive power. However, in cases where the outcomes are not strictly or logically ordered, as in the case of credit card transition states, the application of ordered regression looks quite artificial. For example, the transition of an account from inactive

to the delinquent state via revolver states, or moving between levels of delinquency is logical and can be represented in a certain order. On the other hand, the order of transition of an account from transactor to inactive or revolver state is questionable and depends on how the model builder defines the business logic of this transition. Ignoring the ordinality of the dependent variable leads to losses in efficiency. In this research we are checking whether ordinal regression gives better results than multinomial regression.

5.5 Univariate analysis and variables selection

We consider the distribution of the target by independent variables. We do for preliminary choosing of the set of variables for inclusion into the model. We find that credit card state is correlated with behavioural variables. For example, in Figure 5.8 we present a client rating, or the probability of default, which is correlated with the utilisation rate. The example is taken from the open source <https://www.creditkarma.com/article/CreditCardUtilizationAndScore>.

Figure 5.8 An example of the univariate analysis of the credit score by the utilization rate



A higher utilisation rate corresponds to a low credit score (and a high probability of default). However, for cases with no-utilised card limit (equal to zero), the average credit score is slightly lower than for the interval with the utilisation rate to 20%. The credit score for zero utilisation rate is at the average level of the credit limit utilisation.

For an understanding of dependencies between characteristic values and transition probabilities, we perform a univariate analysis of each factor. The analysis helps us to make a decision whether to include or not the factor into the model, to check business logic and test data inconsistencies.

Below we provide an example of some application, behavioural and states characteristics analysis. An example of a state transition distribution of a behavioural characteristic ‘Average debit (spending) transaction in the last month to average outstanding balance in the last month’ b_Tavg_deb1_to_avgOB1 – the ratio of average debit turnover (purchase transactions) in month 1 to average outstanding balance in month 1 by characteristic’s values band, is shown in Figure 5.9. The characteristic values on the z-axe of Figure 5.9 represent the low bound of the range. The ranges are defined to put the approximately equal number of observation in each band. Thus the last two columns relate to high values where the sum of the purchase amount for the month exceeds the average outstanding balance for the month. This can be explained as cases with low average outstanding balance in denominator what gives extremely high values of the ratio. For example, the customer has an active balance at the beginning of the period (month) say 10 money units, makes a purchase for a significant amount of 1000 money units on the 5th day, but paid all debt 1010 on the 6th day of the month. So the average balance is 35 and the average transaction amount is 1000. The characteristic value, in this case, is $1000/35 = 28$. Moreover, the last column with extreme values is corresponding with 0/0. This is typical to non-active and transactor accounts only.

Figure 5.9 Univariate distribution from all states S t to state S t+1 for an average purchase transaction to average balance for a month by the equal number of observations in the range

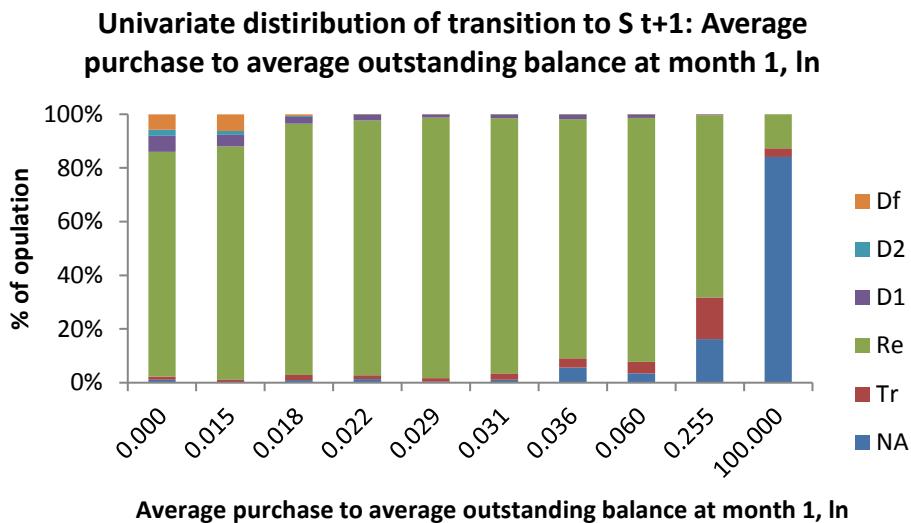
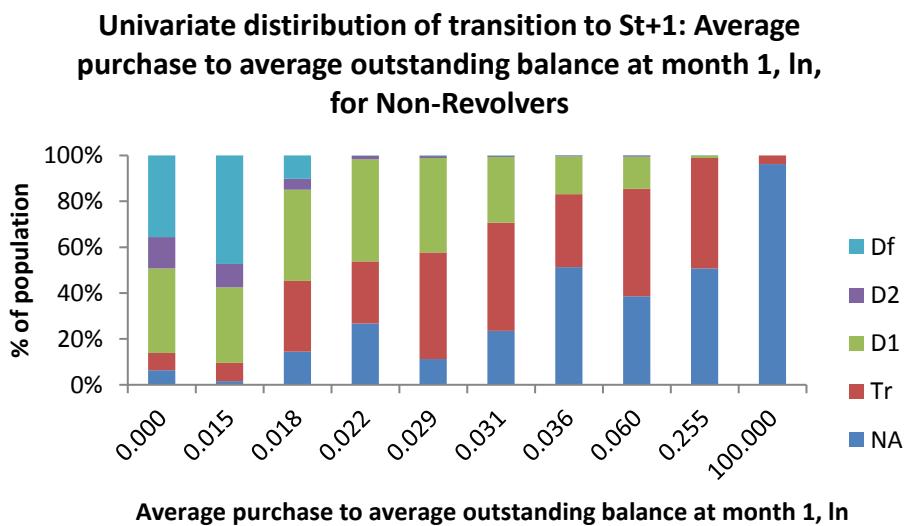


Figure 5.9 shows the transition from all possible states to all possible states. As it can be seen the transition to revolver state dominates on other states transitions. For characteristics values from 100 this means no balance and no transaction or transaction and paid balance on the same day. Generally, the dependence between the characteristic value and transition to the state does exist and is observed. If we look at non-revolvers only it is possible to see more detailed interactions as in Figure 5.10.

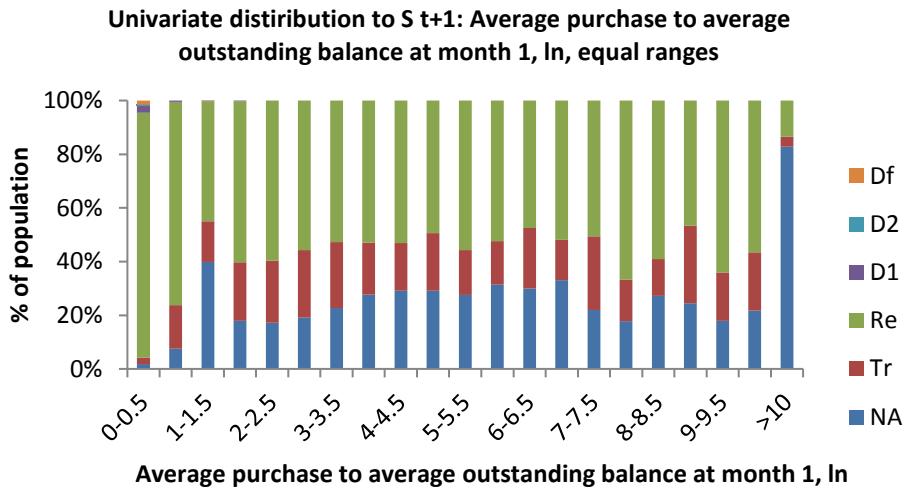
Figure 5.10 Univariate distribution from all states S t to state S t+1 for average purchase transaction to average monthly balance (without revolvers)



Thus the accounts tend to become non-active with higher values of average purchase to average outstanding balance in a current month. On the other hand, for defaulters and the delinquent, the low values are typical.

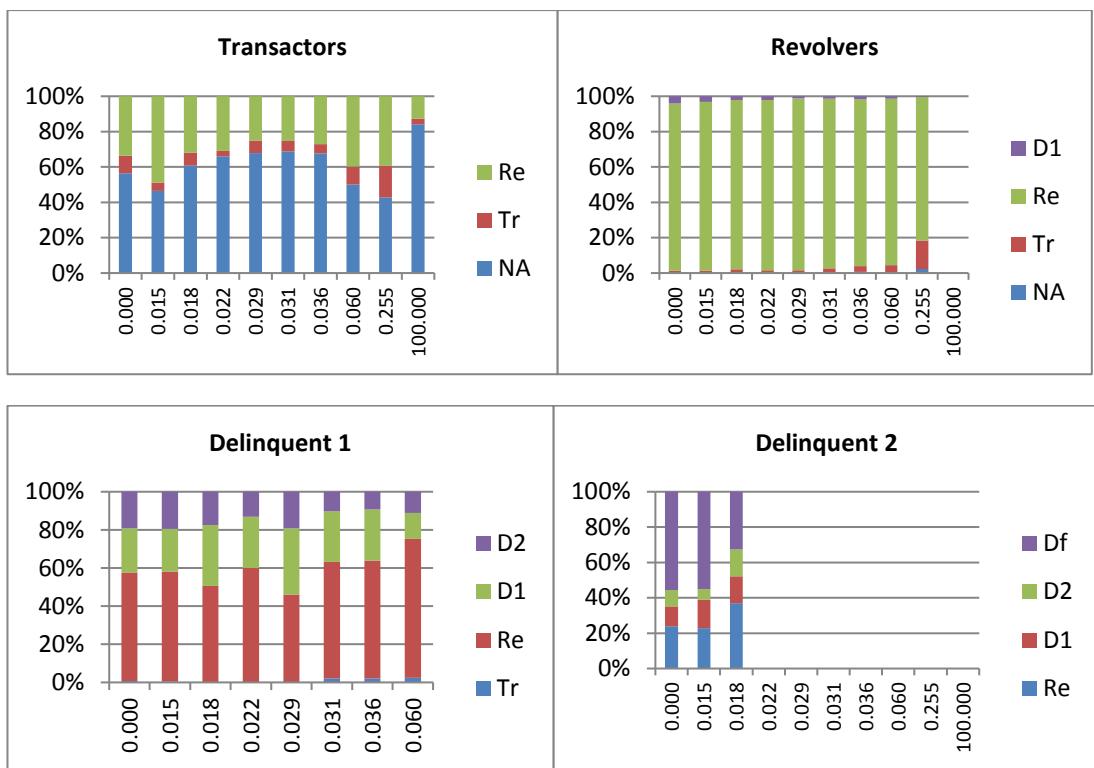
Another view of the characteristics can be obtained if we define the ranges with an equal step, but normalise to 100%. Then we get an unequal number of observations in each band as presented in Figure 5.11.

Figure 5.11 Univariate distribution from all states S_t to state S_{t+1} for an average purchase transaction to average monthly balance by equal ranges



It can be seen from the Figure 5.11 that Revolvers and Inactive states have changed from low (high) to high (low) values of transaction and balance rate, but in the middle of the range, the relations are not clearly defined. The average values in the interval between 1.5 and 9.5 do not vary significantly, so the use of dummy variables or some non-linear transformation may be unuseful. Almost all current delinquent transitions are concentrated in the lowest band because in the delinquency state a credit card is blocked except some exceptional cases. However, if we split it into segments by current states (state ‘from’) the picture is entirely different (see Figure 5.12).

Figure 5.12 Univariate distribution from some state S_t to state S_{t+1} for the ratio of an average purchase transaction to average monthly balance by the equal number of observations in the range



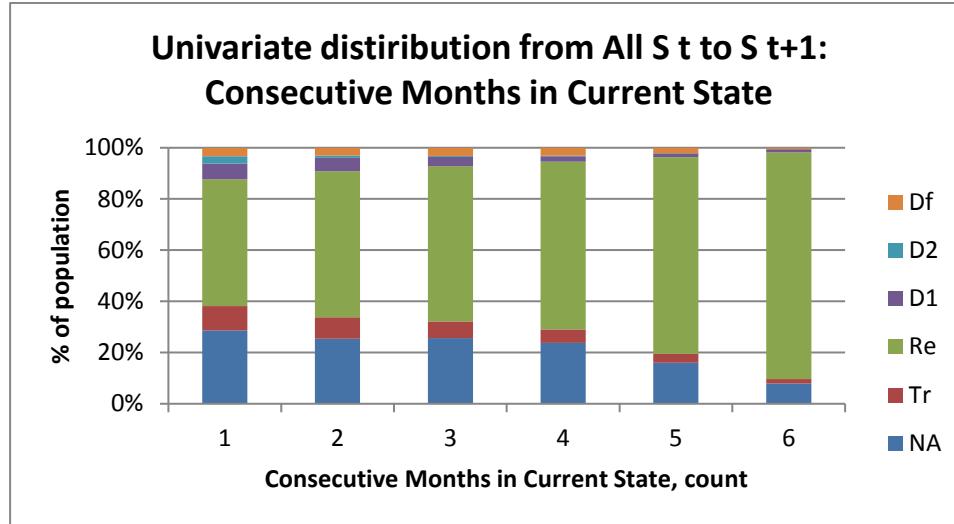
The distributions in Figure 5.12 for all four histograms each by the current state are quite different from the initial histogram in Figure 5.9 built for all current states. At the current state histogram for a revolver, the $t+1$ revolver state dominates. Transactors tend to go rather to non-active or revolver state and have a non-linear trend with some increase of NA for the high values of the characteristic. D1 has a little increase in the middle of characteristics values range for 0.2-0.3 and decrease for the high values. D2 has a slight trend for defaulters to decrease and revolvers to increase for higher values. The figure for the non-active current state does not exist because the rate of purchases to balance is not applicable to these cases.

The results of the univariate analysis for state variables have more expected (or monotonic) trends. For instance, a variable ‘Number of Consecutive Months in Current State’ show a pattern for the given sample and each segment transition. The value ‘6’ in diagrams for Number of Consecutive Months in Current State means 6 or more months in the the current state.

Figure 5.13 shows a normalised histogram of the distribution of transition from state S_t to state S_{t+1} . Generally, the customer has more chances to stay a revolver when the

number of months in revolver state decreases. However, for non-active as for other states, the trend is the opposite. For the transition from inactive state the Consecutive Months in Current State shows that the longer a customer stays non-active, the lower the chances to activate and become a revolver or a transactor.

Figure 5.13 Univariate distribution from All S_t to S_{t+1}: Consecutive Months in Current State



If we split the distribution of characteristic ‘Consecutive Months in Current State’ by S_t segments: inactive, transactor, revolver, delinquent 1, and delinquent 2 states, various distributions of transition to other possible states can be seen (Figure 5.14). For example, if an account is consequently 6 or more month in transactor state, it will stay as a transactor in the next month with 80% chance, and for revolvers this indicator even close to 100%. However, delinquent 1 accounts will rather go to revolver state. For Delinquent 2 accounts, we do not have enough observations with Consecutive Months in Current State more than 3, so Delinquent 2 is not a stable state.

Figure 5.14 Univariate distribution from S_t to S_{t+1} by Consecutive Months in Current State



Inactive and revolver states are stable in the sense that the longer they stay in the state, the higher the chances that an account will stay in the same state in the future. For transactors and delinquents, it is not a rule, and they have a tendency not to stay in the same state for a long time.

With the use of visual analysis, we have empirically shown that the behavioural characteristics have the various distribution of transitions to another possible state for one month. Thus they can be used for predictive analysis. A detailed description of characteristics is given in Chapter 3. The result of regression analysis, characteristics significance and trends are discussed in Chapter 6.

Selected covariates and expectations for their impact on the probability of transition is presented in Table 5.4.

Table 5.4 Selected covariates and expectations for their impact on the probability of transition

Variable name	Description	Expected impact on the probability of transition
AgeGRP1	Age less than 25	Younger customers are more active and have more chances for the transition to the revolver state, but also to the delinquent and default state because of a lack of money and low payment discipline
AgeGRP3	Age more than 35	Older customers might have more chances for the transition to the inactive state
avg_balance_1	Average outstanding balance for the last month	Higher current outstanding balance might increase chances of transition to revolver or delinquent state
b_atm_flag_1	A binary indicator of ATM cash withdrawals transaction at the observation point month	The use of credit card for ATM cash withdrawals increases chances to have a positive balance in the next period
b_atm_flag_use1 3vs46	A binary indicator of ATM transactions for the last three month and no transactions for 4-6 months before	The start of use the credit card for ATM transactions may increase the probability to have a positive balance in the next period and to go to revolver state
b_atm_use_only_flag	A binary indicator of ATM transactions only	If a customer uses a credit card for ATM cash withdrawals only, this might mean that he/she has low financial literacy. So such behaviour may increase chances to go to the delinquency and default states
b_avgNumDeb13	Average number of debit transactions for the last three months	The higher number of transactions decrease the chance of a transition to the inactive state
b_AvgOB1_to_M axOB1_ln	The logarithm of average OB to maximum OB for the last month	If the average outstanding balance for a period is less than maximum balance for the same period, this means some variation in the outstanding balance for this period. If the average outstanding balance is close to maximum one, this means that the outstanding balance is stable and its value is close to 1. We expect a high probability of revolver or delinquent states for higher values of this variable.
b_AvgOB16_to_MaxOB16	The logarithm of average OB to maximum OB for the last six months	If the average outstanding balance for six months period is less than maximum balance for the same period, this means some variation in the outstanding balance for this period. If the average outstanding balance is close to maximum one, this

Table 5.4 Selected covariates and expectations for their impact on the probability of transition

Variable name	Description	Expected impact on the probability of transition
b_max_dpd16	Maximum days past due for the last 6 months	means that the outstanding balance is stable and its value is close to 1. We expect a higher probability of revolver or delinquent states for higher values of this variable.
b_maxminOB_avg OB_1_ln	The logarithm of the ratio of the difference between maximum and minimum values of the outstanding balance to the average outstanding balance for the last month	We expect that an account which had a high number of days past due will have a higher chance of a transition to the delinquent and defaulted states
b_maxminOB_lim it_1_ln	The logarithm of the ratio of the difference between maximum and minimum values of the outstanding balance for the last month to the credit limit	High difference between the minimum and maximum balance divided by the average outstanding balance for period mean high volatility of the balance and is a sign of the active use of the credit card. We expect a high chance of a transition to the transactor or revolver state for values of this characteristic which exceed zero.
b_OB_avg_to_eo p1ln	The logarithm of the average outstanding balance for the last month to the outstanding balance at the end of the period at the observation point	High difference between the minimum and maximum balance divided by the credit limit mean high volatility of the balance and is a sign of the active use of the credit card. We expect a high chance of a transition to the transactor or revolver state for values of this characteristic which are close to zero. If the outstanding balance at the end of the month is higher than average balance for this months, we expect the increase of the chance to be a revolver in the next month
b_OB1_to_OB2_ln	The ratio of the outstanding balance in the last month to the previous month	If the outstanding balance at the end of the month is higher than average balance for this months, we expect the increase of the chance to be a revolver in the next month
b_OBbias_1_ln	The logarithm of the ratio of minimum outstanding balance to maximum balance for the period	In case of decrease of the outstanding balance in comparison with previous month value we expect a higher chance of a transition to the inactive state.
b_payment_lt_5p_1	The flag whether payment amount is lower than minimum obligatory payment (amount due, which is equal to 5% of the average outstanding balance)	We expect that the high volatility of the outstanding balance may increase the chance of a transition to revolver or transactor states
		We expect that if the payment is less than the amount due the probability of the delinquent and default states will increase

Table 5.4 Selected covariates and expectations for their impact on the probability of transition

Variable name	Description	Expected impact on the probability of transition
b_pos_flag_1	A binary indicator of Point-of-sales (POS) transaction at the observation point month	The use of credit card for POS transactions increases chances to have a positive balance in the next period and become a transactor or revolver
b_pos_flag_use13_vs46	A binary indicator of POS transactions for the last three month and no transactions for 4-6 months before	The start of use the credit card for POS transactions may increase the probability to have a positive balance in the next period and to go to revolver or transactor state
b_pos_use_only_flag_	A binary indicator of POS transactions only	If a customer uses a credit card for POS transactions only, this might mean that he/she has appropriate financial literacy. So such behaviour may decrease chances to go to the delinquency and default states, and increase chances to go to the transactor state
b_TRavg_deb1_to_26_ln	Average debit transactions for the last month to the average debit transaction for 2-6 months	In case the debit transaction amount for the last month is higher than for the previous months we can expect a high chance of a transition to revolver state
b_TRavg_deb1_to_avg0	Average debit transactions for the period to the average outstanding balance	The high amount of debit transactions for the period to the average outstanding balance may increase a chance of a transition to revolver state
b_TRavg_deb16_t_o_avg0	Average debit transactions to the average outstanding balance for the last 6 months	The high amount of debit transactions for the period to the average outstanding balance may increase a chance of a transition to revolver state
b_TRmax_deb_TRsum_deb_16	Maximum debit transaction to the sum of debit transactions for the last 6 months	The higher ratio of maximum spending amount to the sum of purchases for the period means high volatility of transactions and may increase the chance of a transition to the revolver or transactor state
b_TRmax_deb1_T_o_Limit	The maximum debit transaction for the last month to credit limit	The higher ratio of maximum spending amount to the credit limit for the period means high volatility of transactions and may increase the chance of a transition to the revolver or transactor state
b_TRsum_crd1_t_o_2_ln	Sum of credit transactions for the last month to the sum of credit transactions in the previous month	The low value of payment amount in the last month to the previous month value may increase the chance of a transition to inactive or delinquent states
b_TRsum_crd1_t_o_OB1	Sum of credit transactions for the last month to the outstanding balance	The low value of payment amount in the last month to the outstanding balance may

Table 5.4 Selected covariates and expectations for their impact on the probability of transition

Variable name	Description	Expected impact on the probability of transition
b_TRsum_crd13_to_46_In	average outstanding balance Credit payment in the last three months in comparison to 4-6 month	increase the chance of a transition to inactive or delinquent states The decrease of credit payment in the last three months in comparison to 4-6 month might mean both decreases in the outstanding balance and decrease in payment amounts. So this can be the reason for the transition to the inactive state or the delinquent state.
b_TRsum_crd13_to_OB	Sum of the payment amount for the three months period to the average outstanding balance	The low value of payment amount for the period to the outstanding balance may increase the chance of a transition to inactive or delinquent states
b_TRsum_deb1_t_o_2_In	The spending amount in the last month in comparison to the previous month	The increase in spending amount in the last month in comparison to the previous month may increase the chance of a transition to revolver or transactor states
b_TRsum_deb1_t_o_avgO	Sum of spending the amount in the last month in comparison to the outstanding balance	The increase in spending amount in the last month in comparison to the outstanding balance may increase the chance of a transition to revolver or transactor states
b_TRsum_deb1_t_o_TRsumcrd	Sum of spending transaction in the last month to the sum of payment	In case the sum of spending transaction exceeds the sum of payment, the chance of a transition to the revolver or delinquent state may increase. In case the sum of payment transaction exceeds the sum of spending transactions, the chance of a transition to the inactive or transactor state may increase.
b_UT13to46In	The utilisation rate in the last three months in comparison to previous 4-6 months	The increase in the utilisation rate in the last three months in comparison to previous 4-6 months may increase the probability of transition to the revolver or delinquent state
b_UT1to2In	The logarithm of the utilisation rate at observation point to the previous months	
d_State_2_D1	Delinquency 1-30 state in the previous month	If an account is a delinquent, it has high chances of transition to higher delinquency or revolver states in the next period
d_State_2_D2	Delinquency 31-60 state in the previous month	If an account is a delinquent 2 it has high chances of transition to default state in the next month
d_State_2_NA	The Inactive state in the previous month	If an account is inactive it has high chances to stay inactive in the next period

Table 5.4 Selected covariates and expectations for their impact on the probability of transition

Variable name	Description	Expected impact on the probability of transition
d_State_2_Re	The Revolver state in the previous month	If an account is a revolver, it has high chances to stay inactive in the next period
d_State_2_Tr	The Transactor state in the previous month	If an account is inactive, it has high chances to move to inactive or revolver states in the next period
MOB	Month on Book	We expect that higher month on book increases the probability of transition to the revolver and inactive states but decreases the chance of a transition to the transactor and delinquent states.
s Been D1	The flag that account has been in the Delinquent 1 state	If an account has been a delinquent this may increase the chance to be a delinquent in the next period
s Been Tr	The flag that account has been in the transactor state	If an account has been a transactor, this may increase the chance to be a transactor in the next period
s_cons	Number of consequence months in the current state	The longer staying in the current state may increase of the chance of staying in the current state in the next period
s_month_since_NA	Number of months since the inactive state	The longer period after the inactive state might decrease the chance of a transition to an inactive state
s_month_since_RP	Number of months since the revolver repaid the state	The longer period after the repayment state might decrease the chance of a transition to inactive and repayment state
s_times_RE	Number of times in the revolver state	The high number of the revolver states might increase the probability of transition to the revolver state
s_times_RP	Number of times in the revolver repaid the state	The high number of full repayments might decrease the probability of transition to the delinquency state
s_times_TR	Number of times in the transactor state	The high number of the transactor state might increase the probability of transition to the transactor state and decrease the probability to be inactive
sum_deb_num_1	Number of debit transactions for the last month	The high number of debit transactions might increase the chance of a transition to revolver or transactor state
customer_income_In	The logarithm of the ratio of the customer monthly income to the average monthly income in a portfolio	Customers with high income might have high chances of transition to transactor state because they use credit cards mainly as a payment instrument
Edu_	Education	Customers with higher education might have high chances of transition to transactor state because they are more qualified users of financial instruments.

Table 5.4 Selected covariates and expectations for their impact on the probability of transition

Variable name	Description	Expected impact on the probability of transition
Marital_	Marital status	Customers with secondary education might have high chances of transition to delinquent and default state because they have poor financial literacy and low income Single customers might have higher chances of transition to the delinquent and default state
position_	Employment status	Top managers might have high chances of transition to transactor state because they are more qualified users of financial instruments. Technical staff might have high chances of transition to delinquent and default state because they have poor financial literacy and low income
sec_	Sector of Industry	Customers from agriculture and construction industry might have higher chances of transition to the revolver and delinquent states
car_	Car owner	Car owners might have a higher chance of a transition to the revolver state because they probably have more spending
real_	Real estate	Customers who rent a flat might have a higher chance of a transition to the revolver and delinquent state because they probably have more spending
reg_ctr_	Region of living	A customer from capital and region centres might have higher chances of transition to transactor and revolver states because they have more possibilities for spending credit money
child_	Number of children	Customers with children have higher chances of transition to the transactor and revolver states because they have more spending
Unempl_Inyoy	The logarithm of the unemployment rate change year on year	The unemployment rate will be positively correlated with the probability of transition to revolver and delinquent states because a higher level of unemployment may cause the increase in demand for money.
UAH_EURRate_In mom	The logarithm of the exchange rate of local currency to Euro at the observation point in comparison with the previous month	The increase in local currency to Euro exchange rate (LCY/EUR exchange rate) might cause purchasing power loss because of growth of prices: directly for export goods and indirectly for local goods in proportion to the export component. Thus, the probability of transition to the revolver and delinquent states may have a positive

Table 5.4 Selected covariates and expectations for their impact on the probability of transition

Variable name	Description	Expected impact on the probability of transition
UAH_EURRate_In yoY	The logarithm of the exchange rate of local currency to Euro at the observation point in comparison with the same period of the previous year	correlation to the foreign currency exchange rate. The increase in local currency to Euro exchange rate (LCY/EUR exchange rate) might cause purchasing power loss because of growth of prices: directly for export goods and indirectly for local goods in proportion to the export component. Thus, the probability of transition to the revolver and delinquent states may have a positive correlation to the foreign currency exchange rate.
CPI_Inqoq	The logarithm of the current Consumer Price Index at the observation point to the previous quarter CPI	The increase in Consumer Price Index means the rise in average prices of a basket of consumer goods and services. Thus, the probability of transition to the revolver state can also be increased in this case.
SalaryYear_InyoY	The logarithm of the Average Salary at the observation point in comparison with the same period of the previous year	The increase in Salary may increase the probability of transition to the inactive state because of more available money. On the other hand, we expect that a decrease in salary will cause an increase in the probability of transition to delinquent and default states
I_ch1_In	Limit change month ago	The increase of credit limit might increase chances of transition to revolver and delinquent states
I_ch6_In	limit change 6 months ago	The increase of credit limit might increase chances of transition to revolver and delinquent states

5.6 Empirical transition matrices

5.6.1 Initial empirical transition matrix and states stability

Migration matrices and Markov Chains are popular techniques for risk, profit, and transition estimation at the pool level (Thomas and So, 2011; Malik and Thomas, 2012). In risk management, migration matrices are a widespread methodology for credit portfolio forecasting for the measurement of impairments using expected credit losses in IFRS methodology (for example, Abad and Suarez, 2017 – European Systemic Risk Board papers).

We need to understand the common portfolio distributions and tendencies for further account level modelling. Also, the predicted portfolio level aggregated from account

level predictions will be used for the validation and backtesting do general models. Transition matrix shows the percentage of cardholders moving between states from an initial state to a final one. In our case, we investigate movement between following account states: Non-Active, Transactor, Revolver, Delinquent, and Delinquent 2 from t to $t+1$. Table 5.5 presents the observed number of cases that moved between states in our sample period and Table 5.6 presents rates of moving between states from t to $t+1$.

The ‘Total’ column in Table 5.6 shows the initial share of states, the last row – the share after the transition. So the initial share of Revolvers is 81.37%. After transition this share slightly decreases to 80.94%. Thus the number of revolvers declines over the period. This occurs because of the absorbing state of default and the slight increase of non-active accounts (12.93% vs. 12.84%) in comparison with the initial state. Generally, 94.46% of Revolvers stay Revolvers. Thus we can classify states by stability.

Table 5.5 Number of transitions from state S_i at t to state S_j in $t+1$

		To (j)						Total
From (i)	NA	Tr	Re	D1	D2	Df		
NA	21,588	856	3,285					25,729
Tr	3,578	808	2,391					6,777
Re	753	5,079	154,023	3,201				163,056
D1		29	2,302	956	743			4,030
D2			191	105	65	426		787
Total	25,919	6772	162,192	4,262	808	426		200,379

Table 5.6 Proportion of cases in state S_i at t to state S_j in $t+1$

		To						Total
From	NA	Tr	Re	D1	D2	Df		
NA	83.91%	3.33%	12.77%	0.00%	0.00%	0.00%	12.84%	
Tr	52.80%	11.92%	35.28%	0.00%	0.00%	0.00%	3.38%	
Re	0.46%	3.11%	94.46%	1.96%	0.00%	0.00%	81.37%	
D1	0.00%	0.72%	57.12%	23.72%	18.44%	0.00%	2.01%	
D2	0.00%	0.00%	24.27%	13.34%	8.26%	54.13%	0.39%	
Total	12.93%	3.38%	80.94%	2.13%	0.40%	0.21%	100.00%	

A stable state means that accounts tend to stay in this state or move in this state. A non-stable state means that accounts tend to leave this state. In the migration matrix (Table 5.6) the non-active state has 83.91% of accounts which stay in the non-active

state for $t+1$, and 94.46% of revolvers also stay in current state. On the other hand, only 11.92% of transactors stay in the transactor state, and most accounts move to the non-active or revolver states. The same tendencies take place for delinquency states: 18.44% of delinquency cases move to default in the 2nd month, 23.72% of cases stay in the same state, but 57.12% return to the revolver state. For the transition from delinquency the dominated tendency is moving to the default (and absorbing) state.

Table 5.7 Stability of states

State	Stability
Non-active	Stable
Transactor	Non-stable
Revolver	Stable
Delinquent	Non-stable
Defaulted	Stable

Thus the transactors and delinquent accounts have a tendency to jump to any of the stable states, and a small proportion stays in the same non-stable state. To find factors for this behaviour can be the topic for further investigation. The mover-stayer concept (Hand and Till, 2003; Thomas and Ho, 2001) can be applied for the modelling of such types of states transitions.

5.6.2 Testing for Markovity

Markov chains and Markov Decision Processes have a broad application in the modelling of the dynamic behaviour of the consumer. The main feature of a Markov chain is that the future state of the process depends on the past only through the present state. Thus in first-order Markov Chains, the next state depends on current state only and not on the state at time $t-1$, $t-2$ and earlier.

We wish to test whether account transition is stationary and so whether we can make predictions using a zero-order, or random, Markov Chain. First, the transition matrices were built on the existing data set for a one-year period. The states are defined as inactive, transactor, revolver, delinquent 1, delinquent 2, and default as an absorbing state.

We need to check if the transitions between states from period $t-1$ to t and from t to $t+1$ are independent. To test of obtained transition matrices for Markovity we apply the Pearson goodness of fit statistic (Thomas et al., 2004).

The expected number of transitions (k,i,j) is

$$e(k, i, j) = N(k, i)p(i, j) = \frac{N(k, i)n(i, j)}{N(i)} \quad (5.29)$$

where k is previous state, i is current and j is the next state, $n(i,j)$ is the number of cases of transition from i to j , $p(i,j)$ is transition relative frequency.

We test the null hypothesis that the Markov Chain is the first-order against the alternative hypothesis that it is second order using the following Chi-square statistic:

$$\chi^2 = \sum_{i=1}^M \sum_{k=1}^M \sum_{j=1}^M \frac{[n(k, i, j) - e(k, i, j)]^2}{e(k, i, j)}. \quad (5.30)$$

We find the calculated χ^2 value is 37,695 for transition matrix in Table 5.6 while the critical value is 521.33, and so reject the null hypothesis. We also found that if we aggregate states in non-delinquency, delinquent, and default states we again reject the null hypothesis.

Thus the application of Markov Chains for the transition probabilities prediction at the pool level is questionable. Moreover, an approach for prediction of portfolio dynamics for six periods based on $t+1$ transition matrix multiplied five times has given results with very high dispersion and convergence of distribution of account states at further months is far from expected distribution.

5.7 Conclusion

In this chapter we discussed two possible ways to predict the transition probabilities: i) a pool level approach with the use of transition matrices and Markov Chains, and ii) an account level approach with the use of logistic regressions. Our data does not satisfy the criteria of Markovity and stationarity (Thomas et al., 2004)), so the first-order Markov Chain cannot be used. Thus we tried to use logistic regression for prediction for 1, 2, and more months and then join them for long period transition prediction. The results are discussed in the next Chapter 5.

Predictive models for risk and profit parameters can be built for a whole credit card portfolio at the pool level with, for example, a Markov Chain (So and Thomas, 2011).

However, significant differences between credit card usage types can decrease the predictive accuracy of the model, because the different forms of credit card usage have individual behavioural drivers for risk, utilisation, purchases, and profit (So et al., 2014, Tan and Yen, 2010).

Credit card holders' multistate transition probabilities modelling allows one to estimate future income as depending not only on the current state but also depending on the possible future states. It uses the transition probability as a weight for the expected income estimation. We describe the multistate model based on states, which are related to income, but not to risk or delinquency only.

We developed an approach for the total income prediction, which accumulates several individual models for various sources of revenue and explores the income in the dimension of credit card behavioural types. For example, Till and Hand (2003) and Leow and Crook (2014) use delinquency states, So and Thomas (2011) investigate transition to particular states such as Closed, Bad1, Bad2, Bad 3 + cycle, Inactive, and current with estimated ten levels of risk with behavioural scores. They use the transition probabilities between ten behavioural score bands. We proposed the set of behavioural states for the more accurate credit card income prediction, which contains only income related states, excluding related risk segmentation.

The results of the empirical analysis using multistage binary logistic regression models and multinomial and ordinal logistic regression models are presented and discussed in Chapter 6.

6 Chapter Six. Credit Card Holders' States Transition Probability: Modelling Results

6.1 Introduction

This chapter describes an empirical investigation of the comparative analysis of multinomial logistic regression, ordinal regression, and conditional probabilities model to see which algorithm gives the most accurate predictions of transition probabilities.

The central question in this chapter is: which approach to the prediction of multistate transition probabilities gives more accurate predictions: multinomial logistic regression or a decision tree with conditional binary logistic regressions. The first stage of the income estimation model (see Figure 1.3 from Chapter 1) is the determination of the account status via transition probabilities at the account level. Three approaches to predict the status have been investigated: i) conditional logistic regression, ii) multinomial logistic regression, iii) ordinal logistic regression.

The comparative analysis and validation results of multinomial, ordinal, and conditional binary logistic regression for the transition probabilities prediction can be used to make recommendations for further investigation.

Also, we propose to include a new state - revolver repaid, which is according to the definition proposed by Kallberg and Saunders (1983) the same state as 'true revolver'. The revolver repaid state is used for the identification of the transition of a revolver account to an inactive state. A customer who fully repays the debt amount at the end of the month cannot formally be allocated into any of the transactor, or inactive, or revolver states. This account generated interest income, and, possibly, transactional income, but, firstly, it does not generate a credit risk for the next month, and, secondly, the set of possible states for transition and the probabilities of transition are different from the revolver or inactive state. We believe that inclusion of this new state helps to make a more accurate prediction of transition and income.

Significant differences between credit card usage types can decrease the predictive accuracy of such models, because the different forms of credit card usage have individual behavioural drivers for risk, utilisation, purchases, and profit (So et al.,

2014; Tan and Yen, 2010). Our contribution is that we proposed and extended an approach of individual transition probabilities for each credit card holder depending on the individual behaviour instead of pool level probabilities of transition computed with the transition matrix.

This chapter discusses the coefficient estimation results for all types of regression, sets of predictors, and time horizons, and comparative analysis of the models. Section 6.2 describes modelling results for multinomial and ordinal logistic regression. Then we compare the results of the validation of multinomial and ordinal logistic regression and select the best method for transition probabilities prediction in Section 6.3. In Section 6.4 we describe multistage binary logistic regression and compare it with multinomial logistic regression approach, selected in the previous section. Section 6.5 introduces the updated set of the credit card account states. Section 6.6 discusses the results of the multinomial model estimations for the full set model, and the final model choice for the income prediction, and conclusions.

6.2 Modelling Results - Multinomial vs ordinal logistic regression

6.2.1 The results of the coefficients estimation

The primary task of this section is to estimate multinomial and ordinal logistic regressions with a set of the behavioural, application, macroeconomic, and account state predictors and to provide a comparative analysis of which type of model gives better prediction results for given data sample.

In this section, we predict the probability of transition from the state at current time t to the state at the next month $t+1$. So we test several regression methods and describe them in detail for a short performance window. Then we will choose the most accurate and reliable method according to results of validation and business logic, and apply it to more extended periods till $t+6$.

We use the characteristics described in Chapter 3 as covariates lagged for one, two, and three months for the models of this Chapter. There are four categories of variables: behavioural (computed, with prefix `b_`, and several original such as `avg_balance` – average outstanding balance, `dpd` – days past due, `sum_deb_num` – sum of debit transactions for month); state – the current state of the account and some behavioural

characteristics, which reflect changes of the account state; application (without prefix); and macroeconomic variables (Euro to Local currency exchange rate 1 month lagged, Euro to Local currency exchange rate 1 year lagged, unemployment rate change for 1 year, yearly CPI changes). See Table 3.1 for a full definition of all variables.

6.2.1.1 Multinomial logistic regression

First, we built the *multinomial logistic model*. The results of the estimation of multinomial logistic regression are presented in Table 6.1. The number of models is equal to the number of current (initial) states – five in our case. The state of an account at time t , from which we predict the transition, is represented in the top row with the heading ‘FROM’ and correspond to the appropriate model (one model for each state). The second row of the heading defines the state $t+1$ or state transition ‘TO’, which we try to predict. We present in the table the estimated coefficients and test the null hypothesis that the regression coefficient is equal to zero $\text{Pr} > \text{ChiSq}$.

The probability of default is estimated as follows:

$$Pr(Y_i = j) = \frac{e^{\beta' \mathbf{x}_{ij}}}{\sum_{j=1}^J e^{\beta' \mathbf{x}_{ij}}}$$

where j is a state, J is the number of states, β is coefficients of regression, x is characteristics value for i case in j state.

The transactor state is selected as the base state, and the probability of the transition to this state is defined as a one minus the sum of the probabilities of transition from state J_t to all possible states J_{t+1} except the transactor state $1 - \sum Pr(J_{t+1}|J_t)$.

In the data sample the event is marked as ‘1’ and non-event marked as ‘0’ for the event probability modelled as $\text{Target} = 0$. So, the lower values of the coefficients of regression correspond the higher probability of an event.

We included characteristics, which are used in Table 3.1, into the model and use logistic regression to show the results of inclusion of all factors in regression. We explain the reasons for inclusion of characteristics into the model in Section 5.5 of Chapter 5. We selected behavioural variables calculated for a single period, but not the aggregates for several months as it has been done for the utilisation rate model. In

Chapter 5 we explained that such variables selection is made to reduce the impact of multicollinearity, especially, for short-term (one-month) prediction horizon.

Table 6.1 Multinomial logistic regression estimation results

FROM	NA		Tr		Re			D1			D2			
	TO	NA	Re	NA	Re	NA	Re	D1	Re	D1	D2	D1	D2	Df
Characteristic	Estimate Pr > ChiS													
AgeGRP1	-0.1401	0.2754	0.00529	0.9696	-0.0649	0.6739	0.1653	0.2862	-0.276	0.0508	0.0431	0.4345	-0.035	0.6943
AgeGRP3	-0.0656	0.5703	-0.0331	0.7908	-0.2087	0.124	-0.1713	0.2094	-0.0109	0.9295	0.038	0.4357	-0.3453	0
avg_balance_1	-1.0877	0.0745	0.00511	0.9934	-0.0001	0.1388	-0.0001	0.0886	0.00003	0.6243	4.3E-05	0.0521	-6E-05	0.1856
avg_balance_2	-0.0003	0.0096	-0.0003	0.0096	2.5E-06	0.9728	0.00008	0.2721	1.2E-05	0.8799	5.8E-06	0.8413	3.7E-05	0.5876
avg_balance_3	9.6E-05	0.1894	0.00012	0.1078	2.2E-05	0.7021	-6E-05	0.339	-6E-06	0.925	2.1E-06	0.9279	0.00004	0.4117
b_atm_flag_0	-1.8317	0	-0.4584	0.021	-0.2975	0.0351	-0.1159	0.4037	0.554	0.0002	0.3541	0	0.5159	0
b_atm_flag_1	-0.35	0.0686	-0.1669	0.4086	0.0774	0.5802	0.1955	0.1582	0.2188	0.1222	0.3215	0	0.2437	0.0148
b_AvgOB1_to_MaxOB1_In	-0.0369	0.7329	0.0587	0.5736	-0.0918	0.2435	-2.9955	0.044	-0.3376	0.0526	-0.3035	0	1.2137	0.0041
b_AvgOB2_to_MaxOB2_In	0.2188	0.1429	0.2106	0.2747	-0.1817	0.2626	-0.1045	0.5293	-0.0289	0.8548	0.1264	0.0423	-0.3386	0.1443
b_AvgOB3_to_MaxOB3_In	-0.2635	0.2548	-0.2639	0.268	-0.1918	0.291	0.1185	0.5484	-0.1296	0.3381	0.1353	0.0242	-0.0414	0.8338
b_fullpaid1	0.2684	0.8777	0.3536	0.8429	0.4419	0.2072	0.519	0.1429	0	0	0	0	0	0
b_inactive2	0.4123	0.5309	-0.3916	0.577	-0.023	0.9758	-0.7323	0.3487	-0.7052	0.3692	-1.0136	0.0025	-2.4675	0.0853
b_inactive3	1.2187	0.0167	0.7085	0.197	0.3896	0.5829	0.831	0.2634	0.6817	0.3499	0.9021	0.002	3.9043	0.0219
b_maxminOB_avgOB_1_In	0.0733	0.7298	0.1868	0.3898	-1.6626	0.3264	-2.7864	0.0433	-0.1353	0.1473	-0.722	0	-1.2043	0
b_maxminOB_avgOB_2_In	0.1626	0.1634	0.1613	0.2193	-0.2185	0.0238	-0.1909	0.0542	-0.2076	0.0158	-0.1563	0	-0.5273	0
b_maxminOB_avgOB_3_In	-0.1188	0.281	-0.0973	0.3951	-0.0814	0.3955	0.0859	0.3929	-0.0614	0.4572	-0.012	0.7248	0.0312	0.7923
b_maxminOB_limit_1_In	0.00125	0.9738	-0.0488	0.2208	-0.0898	0.0366	-0.0217	0.6253	-0.0361	0.4676	0.3009	0	1.2029	0
b_maxminOB_limit_2_In	0.0292	0.3988	0.037	0.3203	0.0443	0.3698	0.0215	0.6688	0.00419	0.9353	-0.0126	0.5623	0.3303	0.0021
b_maxminOB_limit_3_In	0.00505	0.8872	-0.0168	0.6566	0.0562	0.2331	-0.0009	0.9858	-0.0678	0.1686	-0.1276	0	0.0559	0.5719
b_OB_avg_to_eop1In	-0.0313	0.8497	-0.0764	0.6496	-0.0243	0.8119	-0.0645	0.5364	-0.0345	0.2928	-0.2993	0	-1.686	0
b_OBBias_1_In	-0.0948	0.5972	-0.0614	0.7399	-0.2073	0.2417	-0.1974	0.2779	-0.1828	0.0022	0.081	0.0002	-0.0534	0.1229
b_OBBias_2_In	0.0771	0.3805	0.0387	0.7049	0.0306	0.6797	-0.0324	0.6697	0.1093	0.0463	0.2434	0.2394	-0.1469	0
b_OBBias_3_In	-0.0245	0.8093	-0.0541	0.6067	0.0864	0.2591	0.1294	0.1101	0.0361	0.4854	0.0195	0.3417	-0.0498	0.1296
b_payment_lt_5p_1	0.5741	0.4554	0.6278	0.4221	-0.7728	0.3219	-0.6128	0.4536	-0.0813	0.7011	-0.1712	0.006	-0.1589	0.0855
b_pos_flag_0	-1.7162	0	-0.6021	0.0013	-0.5502	0	-0.3985	0.0035	0.1095	0.4278	-0.0236	0.6458	0.4922	0
b_pos_flag_1	-0.5217	0.0028	-0.4588	0.013	-0.1066	0.4301	0.2704	0.0461	-0.2305	0.0889	0.014	0.7841	0.3768	0
b_TRavg_deb1_to_avgOB	-0.0032	0.9951	1.0052	0.0525	0.237	0.1162	0.0571	0.6797	-0.0448	0.8	-0.2567	0	-0.3326	0.0051
b_TRavg_deb2_to_avgOB	0.4652	0.0693	0.2523	0.3471	0.2386	0.0948	0.0135	0.9217	-0.0053	0.9701	0.0317	0.5278	-0.1889	0.12
b_TRavg_deb3_to_avgOB	0.3055	0.1234	0.1213	0.559	0.2306	0.106	-0.1216	0.3749	-0.0911	0.5976	-0.1923	0.0002	-0.3432	0.0015
b_TRmax_deb1_To_avgOB	1.325	0.5718	-1.5656	0.5052	-0.9042	0.0078	-0.7877	0.0176	0.696	0.0293	0.5049	0	0.7291	0
b_TRmax_deb1_To_Limit	-4.856	0	0.6618	0.1433	0.117	0.5699	0.2881	0.1095	-0.1145	-0.07	0.0948	0.1696	0.0025	-11.562
b_TRmax_deb2_To_avgOB	-1.3985	0.0118	-0.589	0.3088	-0.2418	0.4754	0.386	0.2474	-0.0266	0.9332	0.3251	0.005	0.6423	0.0006
b_TRmax_deb2_To_Limit	0.3656	0.4147	0.2313	0.6135	0.0228	0.9045	-0.0344	0.8431	-0.0193	0.8929	-0.0984	0.0112	-0.095	0.6321
b_TRmax_deb3_To_avgOB	-0.1781	0.6809	-0.4546	0.3084	-0.396	0.2322	-0.4472	0.1628	0.3337	0.2835	0.1548	0.1497	0.3934	0.0173
b_TRmax_deb3_To_Limit	-0.1415	0.6309	-0.1188	0.6906	-0.0426	0.8124	-0.204	0.2147	0.0711	0.536	-0.0608	0.1177	-0.4589	0.0906
b_TRsum_crd1_to_OB1	0.225	0	0.2076	0	0.3621	0.002	-0.0736	0.0506	-0.2829	0.0331	0.00358	0.8992	0.1589	0.0083
b_TRsum_deb1_to_avgOB	-1.5762	0.4881	0.3922	0.8631	0.1076	0.7743	0.7248	0.0484	-0.112	0.7502	-0.2786	0.024	-1.02	0
b_TRsum_deb1_to_Tsum	0.1312	0.0394	0.1363	0.039	0.4341	0.0005	0.0268	0.8265	-0.8244	0	0.0842	0.0158	0.706	0
b_TRsum_deb2_to_avgOB	0.8319	0.131	0.2948	0.6079	-0.124	0.7241	-0.4632	0.1831	0.0917	0.7778	-0.3092	0.0102	-0.571	0.0046
b_TRsum_deb2_to_Tsum	-0.0202	0.3984	0.0023	0.928	0.041	0.0381	0.0385	0.057	-0.0297	0.1216	0.00758	0.3655	0.2388	0
b_TRsum_deb3_to_avgOB	-0.2825	0.4943	0.2365	0.5775	0.0316	0.9212	0.4564	0.1422	-0.2272	0.4339	-0.0057	0.9545	-0.2914	0.0523
b_TRsum_deb3_to_Tsum	-0.0083	0.6723	-0.0186	0.3729	0.0249	0.2013	0.03	0.1323	0.029	0.1553	0.00012	0.9882	0.0348	0.0032

Table 6.1. Multinomial logistic regression estimation results (continued)

FROM	NA		Tr		Re			D1			D2																
TO	NA	Re	NA	Re	NA	Re	D1	Re	D1	D2	D1	D2	Df														
Characteristic	Estimate	Pr > ChiS	Estimate																								
car_coOwn	0.0299	0.8417	0.0773	0.6313	0.1151	0.4951	-0.003	0.9862	-0.0217	0.8954	0.0303	0.6344	0.0621	0.5819	2.6657	0.417	2.609	0.4273	2.381	0.4692	0.8606	0.246	1.1457	0.1741	0.4672	0.4724	
car_Own	-0.2328	0.0071	-0.2584	0.0062	0.1383	0.1912	0.0979	0.36	0.121	0.2366	-0.0765	0.0553	-0.1689	0.028	-0.2936	0.6434	-0.6762	0.293	-0.9772	0.1328	0.8641	0.226	1.2013	0.1532	0.7807	0.1738	
child_1	0.00368	0.9767	0.1232	0.3635	-0.1011	0.5041	0.0297	0.8457	-0.2935	0.0246	0.1426	0.0081	-0.0429	0.6315	-0.2001	0.7733	-0.2612	0.7091	0.1066	0.8797	0.3451	0.5078	-0.3706	0.5134	0.6709	0.0933	
child_2	-0.0484	0.489	-0.0308	0.6834	0.00101	0.9903	0.0224	0.7905	-0.1921	0.0089	0.026	0.3849	-0.1158	0.0292	0.4159	0.4063	0.2942	0.5598	0.4019	0.4285	0.1458	0.673	-0.0626	0.879	0.3527	0.1916	
child_3	-0.1743	0.4839	-0.0352	0.8954	-0.1401	0.601	0.077	0.7751	-0.4585	0.0883	0.063	0.5281	0.1018	0.5652	-2.1678	0.0516	-2.8868	0.0118	-2.189	0.0582	-1.3333	0.3921	-0.3832	0.7734	-0.0879	0.9238	
CPI_Inqoq	9.7349	0.0033	2.7421	0.4419	4.8214	0.2102	2.1729	0.5769	0.9741	0.7884	2.4505	0.08	3.8918	0.1223	-5.3645	0.8056	4.3879	0.8418	3.2276	0.8841	18.2184	0.3271	41.4062	0.05	23.5585	0.0937	
customer_income_In	-0.0231	0.7836	-0.0562	0.5395	0.1905	0.051	-0.0253	0.7968	-0.2435	0.0247	-0.0399	0.3332	0.1819	0.0312	-1.2381	0.0931	-1.2208	0.1016	-1.1533	0.1242	-0.1121	0.8688	-0.9751	0.2339	0.4935	0.3403	
d_State_2_D1	0	0	0	0	-99.871	0.4722	-144.7	0.298	-133.9	0.9115	-72.706	0.7501	-161.9	0.4782	7.8179	0.8619	-1.4902	0.9736	6.5014	0.885	6.4841	0.2461	-20.399	0.2498	0.836	0.8512	
d_State_2_D2	0	0	0	0	-40.933	0.748	-67.292	0.5958	-62.393	0	-41.997	0.9726	-89.778	0.9415	2.8945	0.9008	-1.5675	0.9463	2.6442	0.9094	3.4395	0.2643	-10.288	0.2568	1.8558	0.4486	
d_State_2_NA	0.9642	0.0291	0.2806	0.5459	-116.9	0.4502	-168.4	0.2778	-151.6	0.8999	-78.764	0.7346	-182.4	0.4324	0	0	0	0	0	0	0	0	0	0	0	0	0
d_State_2_Re	0	0	0	0	-117.9	0.4464	-169	0.2762	-152.7	0.8991	-78.841	0.7343	-183.9	0.4286	8.5625	0.8707	-4.3496	0.9342	4.549	0.9311	0	0	0	0	0	0	0
dpd_2	0.5559	0.9371	0.7189	0.9187	-1.7422	0.3352	-2.3914	0.1898	-1.7975	0.0427	-0.5547	0.3178	-1.9668	0.0004	0.1333	0.8563	-0.0118	0.9872	0.1175	0.8733	0.0894	0.3084	-0.3153	0.2747	0.0205	0.7752	
dpd_3	-0.1585	0.9931	-0.8506	0.9632	-0.5599	0.7371	-0.6033	0.7176	-0.2264	0.9492	0.2411	0.5437	-0.0156	0.969	0.0881	0.4574	0.0381	0.7591	0.0951	0.4227	0.00974	0.8557	-0.0123	0.9555	0.00088	0.9832	
Edu_High	-0.2605	0.0594	-0.2517	0.0888	0.1439	0.3389	-0.1483	0.3243	0.1559	0.2981	-0.1002	0.0575	-0.2857	0.0005	0.6947	0.2537	0.3916	0.5231	0.4634	0.4528	0.3481	0.483	0.447	0.4538	-0.2271	0.5404	
Edu_Special	-0.096	0.4921	-0.117	0.4335	0.0768	0.6133	-0.0726	0.634	0.1799	0.2388	-0.0649	0.2173	-0.1493	0.0591	0.7023	0.2558	0.54	0.3857	0.7396	0.2368	0.5276	0.2666	0.5912	0.256	0.1483	0.6547	
Edu_TwoDegree	-0.4174	0.0755	-0.375	0.1422	0.4218	0.137	0.1876	0.4972	0.1337	0.6241	-0.13	0.2142	-0.7077	0.0019	3.5065	0.4463	2.5078	0.587	2.5964	0.5743	-7.3757	0.6787	-8.3877	0.6926	-1.7527	0.148	
Intercept	33.4798	0.9896	116	0.964	288.6	0.3261	346.9	0.2352	194.2	0.8755	89.4211	0.638	212.1	0.2644	0.8195	0.988	-6.3708	0.907	-17.424	0.9884	-59.321	0.1382	-19.811	0.7089	-39.967	0.0869	
I_ch1_flag	0.6351	0.3469	0.2946	0.6796	-1.0417	0.0741	-0.5053	0.3709	0.313	0.542	-0.308	0.0938	-0.7341	0.0422	1.2056	0.8103	1.162	0.8187	1.5135	0.7659	-0.385	0.9468	36.7375	0.2351	3.6258	0.3165	
I_ch1_ln	-0.0827	0.0156	-0.8491	0.3477	0.9629	0.3323	0.6677	0.4923	-1.3775	0.1589	0.2442	0.4015	0.4317	0.6159	-13.374	0.256	-8.0253	0.509	-12.092	0.3226	4.854	0.8447	-172.4	0.301	-22.281	0.1763	
I_ChangeFlag	0.1413	0.4054	0.1013	0.5762	0.2537	0.1495	0.3392	0.0515	0.1751	0.2331	0.2313	0.02203	0.021	0.2959	0.7068	0.2839	0.7211	0.2113	0.7919	0.8379	0.2458	0.1849	0.8484	0.1268	0.8341		
Marital_Civ	0.2618	0.2407	0.3286	0.1608	-0.0186	0.9333	-0.0798	0.7198	0.1925	0.3299	-0.0463	0.555	0.0142	0.9094	-0.4081	0.7028	-0.4853	0.6522	0.0266	0.9803	-0.6593	0.2922	-0.7256	0.2849	-0.2997	0.5121	
Marital_Div	0.1093	0.4282	0.0815	0.5818	0.1325	0.4022	0.2102	0.1841	0.2091	0.1176	0.00715	0.8944	0.0485	0.6013	-0.3729	0.5845	-0.469	0.4967	-0.2439	0.7257	0.0209	0.9696	0.3245	0.585	-0.2845	0.5034	
Marital_Sin	-0.0147	0.9106	-0.1339	0.3448	-0.1124	0.4763	-0.1695	0.2842	-0.2053	0.162	0.00898	0.8779	0.0395	0.6706	-0.2791	0.6882	-0.4138	0.5549	-0.1343	0.8489	0.4932	0.3462	-0.4255	0.4593	0.4476	0.2683	
Marital_Wid	0.1452	0.5833	-0.0355	0.902	0.1943	0.5764	0.2399	0.4972	0.5533	0.0328	0.1819	0.124	0.2077	0.2913	0.1027	0.9616	0.0935	0.9651	-0.1131	0.9583	-2.2483	0.888	3.9136	0.3545	3.3911	0.4001	
mob	0.0333	0.0008	0.00131	0.9038	0.00068	0.9546	0.00483	0.6856	-0.0051	0.6618	-0.0038	0.3937	0.0305	0.0001	0.0421	0.5555	0.0401	0.5771	0.0249	0.7315	0.0226	0.6973	0.1358	0.0458	-0.0019	0.9674	
position_Man	0.0331	0.7688	0.0669	0.5829	-0.1598	0.2084	-0.0012	0.9922	0.0145	0.9069	-0.0135	0.7842	-0.0155	0.8671	-0.2944	0.6675	-0.2627	0.7043	-0.0648	0.9261	-0.2939	0.6020	-1.3203	0.0975	-0.0562	0.9031	
position_Oth	-0.0688	0.6015	0.085	0.5468	-0.1552	0.2708	-0.151	0.2915	-0.1945	0.1708	0.0505	0.3376	-0.0454	0.6031	0.2617	0.7192	0.0836	0.9096	0.5723	0.4381	-0.1104	0.8506	-0.4215	0.5301	0.2769	0.517	
position_Tech	-0.1826	0.1452	-0.0975	0.4712	0.2278	0.1282	0.0981	0.5204	-0.3158	0.0331	0.0234	0.6447	-0.0393	0.6263	0.9567	0.1867	0.9638	0.1859	1.3088	0.0734	-0.2207	0.6196	-0.5237	0.3167	0.00784	0.9819	
position_Top	-0.1295	0.453	-0.1334	0.4826	-0.0529	0.7998	0.2592	0.2069	0.1248	0.5477	-0.272	0.0009	-0.1058	0.5287	0.8127	0.6242	0.7915	0.6365	0.8241	0.6259	0.7232	0.655	-4.9273	0.768	0.00258	0.9985	
SalaryYear_Inyoy	3.9776	0.1443	2.1943	0.4565	2.6257	0.4202	5.5404	0.0937	-2.953	0.3385	-1.2255	0.3124	5.0524	0.0166	-11.541	0.5333	-9.6483	0.6053	0.4909	0.9792	3.7264	0.8019	-1.5802	0.9309	12.0248	0.2961	
sec_Agricult	0.0818	0.721	0.1372	0.5757	-0.0934	0.727	0.2218	0.4112	-0.2513	0.3171	-0.0563	0.54	-0.0125	0.9344	-0.2057	0.0137	-1.443	0.0837	-1.5995	0.0603	0.1916	0.8262	0.5034	0.6029	-0.2106	0.7659	
sec_Constr	-0.0053	0.986	0.07	0.8297	0.2369	0.5105	-0.0714	0.8492	-0.8891	0.0637	-0.0002	0.9987	0.1998	0.31	0.6052	0.7062	0.4622	0.7759	0.5904	2.309	0.0807	4.2817	0.0013	1.6982	0.1457		
sec_Energy	0.1082	0.5595	-0.0441	0.8267	0.067	0.756	0.0946	0.6713	-0.411	0.077	-0.1809	0.0186	-0.166	0.2425	0.5675	0.7685	0.5569	0.7733	0.3726	0.8476	1.4409	0.1554	2.0089	0.0798	1.3969	0.1223	
sec_Industry	-0.4463	0.2844	0.059	0.8935	-0.1073	0.8024	0.2143	0.6037	0.1011	0.8234	0.0371	0.826	0.1013	0.7195	-2.9446	0.0321	-2.6976	0.0519	-3.6861	0.0108	-5.3394	0.7938	-5.6243	0.7962	-0.2265	0.8562	
sec_Manufact	0.1441	0.6814	0.1061	0.7787	0.7302	0.1761	1.0287	0.0572	-0.7717	0.1028	-0.0399	0.7476	-0.0461	0.8187	-0.6997	0.6049	-0.9048	0.5117	-0.3341	0.8076	0.2655	0.7768	1.0736	0.3382	-0.726	0.3105	
sec_Mining	0.1796	0.4567	0.1052	0.6855	-0.0493	0.8541																					

There are no parameters for the transition to transactor state in each model because it is used as the base state, and estimates of the probability of an event for the transactor state can be computed according to the full probability formula as one minus sum of the probabilities of other estimated events.

The use of this technique shows that several estimated coefficients have a p-level value (PR > Chi-Squared) close to zero and are significant. Models of transition from the revolver state have the largest number of significant variables. For instance, the following behavioural characteristics have low p-values for all transition models: the use of POS transaction in the last month (b_pos_flag_1), the use of ATM withdrawals in the last month (b_atm_flag_1), the logarithm of ratio of payment sum to the outstanding balance (b_TRsum_crd1_to_OB1_ln), the utilisation rate at the last month (UT0_1). The utilisation rate is significant for the revolver state only, and previous months' utilisation is also insignificant according to the results. On the other hand, there are only a few application characteristics, which have p-values close to zero. It is important to observe that the same characteristic can be significant for one model and not significant for another, for example, the indicator of POS transaction is a significant covariate for the transition from transactor and revolver states. However, whilst it is significant for transitions from delinquent 1-30 to inactive or to revolver states and it is insignificant for transitions from delinquent 1-30 to delinquent 31-60 states (p-level 0.9) and transitions from delinquent 31-60 state to any other state (p-level values around 0.7 – 0.96). It is observed and seems logical that previous history of customer delinquency for t-1 and t-2 months (dpd2 and dpd3) is significant for predicting transitions from revolver to delinquent 1 state. This may reflect the scenario that the cardholder went to delinquency several months ago, paid back arrears amount in the current month and so has become a revolver, but it indicated some problems with debt servicing.

Indicators of full paid debt in the current month (b_fullpaid_1) and inactive state one and two months ago (b_inactive_2 and b_inactive_3) are significant for the inactive and transactor states, but not applicable (have all zero estimates) for the revolver and delinquency states.

As it can be seen from the Table 6.1 (continued), application characteristics are commonly insignificant. So, these covariates can be excluded from the model without significant risk of the deterioration in predictive accuracy.

According to the results, the changes in credit limit in the previous month *L_ch1_flag* are significant for transitions from the revolver state only. Other states are insensitive to the short time credit limit changes, and, for example, transactors have not been motivated to spend more money by a credit limit change and keep a positive outstanding balance at the end of the next month.

Higher Month on Book decreases the probability of transition to the Delinquency 1 state and increases the probability of staying in the same state for revolvers. Staying in the inactive state for longer decreases the probability to stay inactive in the next month. However, for other states MOB is an insignificant predictor.

From the set of macroeconomic covariates we can mention only the logarithm of yearly change in the unemployment rate as a significant characteristic for the transition from the revolver state to the delinquent state with a negative sign. This means an inverse correlation between increase of the unemployment rate and the probability to go to delinquency with 1-30 days past due state.

We test three sets of predictors, which are built with application, behavioural, state, and macroeconomic characteristics for understanding the sensitivity of the model accuracy (predictive power) to different types of transition drivers. The problem is that some behavioural characteristics are complicated for extraction and computation and can have very low marginal values to the predictive power of the model.

Despite of observed insignificance of the majority of covariates, the predictive power of a model can be high and a model can be acceptable for usage. In section 6.2.2 we discuss the results of validation of the multinomial logistic regression model for various states.

6.2.1.2 *Ordinal logistic regression model*

Secondly, we estimate *Ordinal logistic regression* model for each origin state in time t , which generates a single set of coefficient estimates for all destination states. The difference between the models for the predicted probabilities transitions to different states is due to the intercepts only. The results of estimation are presented in Table 6.2

and Table 6.3. Thus, multinomial regression gives individual slopes for each predictor, but ordinal regression has the same slopes between factors and outcome, and the difference between the transition probability, for example, from an inactive state to inactive or revolver state is defined by the intercepts only. Ordinal regression gives the cumulative probability of an event as the output. The use of the appropriate estimated intercept from the model for a selected state gives the estimate of the probability of transition to the required state:

$e^{\alpha_1}/(1 + e^{\alpha_1})$ is the probability of transition to Inactive state (no intercept),

$e^{\alpha_1+\beta_1}/(1 + e^{\alpha_1+\beta_1})$ is the probability of transition to Transactor state,

$e^{\alpha_1+\beta_2}/(1 + e^{\alpha_1+\beta_2})$ is the probability of transition to Revolver state,

$e^{\alpha_1+\beta_3}/(1 + e^{\alpha_1+\beta_3})$ is the probability of transition to Delinquent 1-30 state,

$e^{\alpha_1+\beta_4}/(1 + e^{\alpha_1+\beta_4})$ is the probability of transition to Delinquent 31-60 state,

$e^{\alpha_1+\beta_5}/(1 + e^{\alpha_1+\beta_5})$ is the probability of transition to the Default t state,

where α_1 is the sum of products of all regression coefficients and related characteristic values, and β_n is an intercept for the transition to related state (see above).

We can see that Chi-Squares are high and some of them are close to 1, for example, the average outstanding balance in the last month (avg_balance_1), number of days past due three months before observation point (dpd_3), the amount of maximum spending transaction to the credit limit in three months before the observation point (b_TRmax_deb3_To_Limit). This means that the coefficients estimations are insignificant and this may be caused by the correlation between predictors, the same as in the multinomial regression case.

The most significant characteristics vary depending on the model state. For example, the utilisation rate at the observation month (UT0_1) is not significant for the transition probability from Inactive, Transactor, and Delinquent 2 states, but is significant for transition probability from Revolver and Delinquent 1 states. Month on Book is significant for Inactive and Revolver states but is not significant for other states. The limit change (l_ch1_ln) is also significant for Inactive and Revolver states (Chi-

Squared close to zero), but according to results of the estimation has a slight impact for transitions from Transactor or Delinquent states.

The macroeconomic characteristics are mainly significant. The most influential covariates are quarterly changes in Consumer Product Index (CPI_Inqoq) and yearly changes in the average level of salary (SalaryYear_Inyoy) which have Chi-Squared values close to zero. Yearly changes in the unemployment rate (Unempl_Inyoy) and monthly changes in EUR exchange rate to local currency (LCY_EURRate_Inmom) are significant for the Revolver state only.

We also investigated how account states at time $t-1$ (previous month before observation point) impact on the transition probability (d_State_2_NA, d_State_2_Tr, d_State_2_Re, d_State_2_D1, and d_State_2_D2). According to the results of estimation, only delinquency states (d_State_2_D1 and d_State_2_D2) are significant characteristic for transition from the revolver state and the inactive state (d_State_2_NA) is significant for transition from inactive state. Other states have Chi-Squared values close to one for all transitions. This can be explained that only revolver state is stable and can be continued for an account for quite a long period. On the other hand, transactor and delinquent states are unstable because account tends to become inactive, or revolver, or go to default. It can happen but is not common behaviour, for example, to have 1-month delinquency and pay back only one payment each month.

Table 6.2 Ordinal regression coefficient Estimations (part 1 – for Non-active and Transactors)

Parameter	NA		Tr	
	Estimate	Pr > ChiSq	Estimate	Pr > ChiSq
Intercept 1	-86.2142	0.672	-72.5859	0.9862
Intercept 2	-85.921	0.6731	-72.0153	0.9863
Intercept 3				
Intercept 4				
Intercept 5				
UTO_1	0	.	-0.1941	0.6237
UTO_2	-0.0966	0.8264	0.4925	0.1568
UTO_3	-0.1497	0.5894	-0.7861	0.006
avg_balance_1	-0.9277	0.0052	0.000012	0.802
avg_balance_2	-3.02E-06	0.9574	-0.00006	0.1375
avg_balance_3	-9.04E-06	0.7951	0.000068	0.0424
dpd_2	-0.1573	0.2623	0.6544	0.2817
dpd_3	0.736	0.4211	0.0501	0.6458
b_AvgOB1_to_MaxOB1_ln	-0.0689	0.3614	1.0775	0.3216
b_AvgOB2_to_MaxOB2_ln	0.0526	0.6599	-0.0823	0.3105
b_AvgOB3_to_MaxOB3_ln	-0.0122	0.8872	-0.3072	0.0076
b_TRmax_deb1_To_Limit	-3.3346	<.0001	-0.0918	0.4659
b_TRmax_deb2_To_Limit	0.0442	0.7816	0.0399	0.7438
b_TRmax_deb3_To_Limit	0.00724	0.9631	0.1091	0.3448
b_TRmax_deb1_To_avgO	2.4498	0.0335	-0.0935	0.6242
b_TRmax_deb2_To_avgO	-0.8157	0.004	-0.5633	0.0053
b_TRmax_deb3_To_avgO	0.2455	0.2489	0.131	0.4791
b_TRAvg_deb1_to_avgO	-0.8892	0.0156	0.2054	0.0293
b_TRAvg_deb2_to_avgO	0.197	0.1473	0.2006	0.0232
b_TRAvg_deb3_to_avgO	0.1759	0.0919	0.284	0.0013
b_TRsum_deb1_to_avgO	-1.6143	0.1271	-0.5366	0.01
b_TRsum_deb2_to_avgO	0.5522	0.0386	0.3172	0.13
b_TRsum_deb3_to_avgO	-0.4905	0.0125	-0.4368	0.0128
b_TRsum_deb1_to_TRsum	-0.0172	0.6268	0.2977	<.0001
b_TRsum_deb2_to_TRsum	-0.0188	0.1126	0.000294	0.98
b_TRsum_deb3_to_TRsum	0.00799	0.4182	-0.00428	0.7167
sum_deb_num_6	0.0469	0.2865	-0.00439	0.7944
sum_deb_num_7	0.0413	0.1834	0.0225	0.1194
sum_deb_num_8	-0.00314	0.8833	0.0264	0.054
b_inactive2	0.708	0.0342	0.3859	0.375
b_inactive3	0.6078	0.0218	-0.382	0.3822
b_fullpaid1	-0.1356	0.8812	0.00402	0.9864
b_OB_avg_to_eop1ln	0.0565	0.5207	0.0304	0.6286
b_pos_flag_0	-1.0753	<.0001	-0.183	0.0267
b_pos_flag_1	-0.0954	0.3387	-0.3295	<.0001
b_atm_flag_0	-1.3525	<.0001	-0.2087	0.0151
b_atm_flag_1	-0.177	0.0971	-0.1591	0.0602
l_ChangeFlag	0.0474	0.54	-0.0991	0.2907
l_ch1_ln	-1.209	0.0092	0.2058	0.6914
l_ch1_flag	0.3393	0.2745	-0.3851	0.2318
Mob	0.0308	<.0001	-0.00166	0.8179
b_TRsum_crd1_to_OB1_ln	0.00557	0.8617	0.3145	<.0001
b_payment_lt_5p_1	0.0454	0.9144	-0.1379	0.7452
b_maxminOB_limit_1_1	0.0362	0.0646	-0.0674	0.006
b_maxminOB_limit_2_1	-0.00195	0.9107	0.0245	0.3944
b_maxminOB_limit_3_1	0.0154	0.3434	0.0553	0.0498
b_OBBias_1_ln	-0.025	0.7899	0.00782	0.9342
b_OBBias_2_ln	0.0472	0.4071	0.0593	0.0917
b_OBBias_3_ln	0.0235	0.5497	-0.0355	0.3924
b_maxminOB_avgOB_1_1	-0.1031	0.3795	1.2582	0.2269
b_maxminOB_avgOB_2_1	0.0309	0.6594	-0.0469	0.3384
b_maxminOB_avgOB_3_1	-0.019	0.6624	-0.1711	0.0018
AgeGRP1	-0.1342	0.0339	-0.2082	0.0199
AgeGRP3	-0.0317	0.5649	-0.0445	0.5782
customer_income_ln	0.0164	0.6979	0.1835	0.0024
Edu_High	-0.0324	0.6106	0.2518	0.0036
Edu_Special	0.014	0.8259	0.1253	0.1468
Edu_TwoDegree	-0.1003	0.4052	0.1831	0.2637
Marital_Civ	-0.0241	0.8028	0.0681	0.6024
Marital_Div	0.0329	0.6026	-0.0655	0.458

Parameter	NA		Tr	
	Estimate	Pr > ChiSq	Estimate	Pr > ChiSq
Marital_Sin	0.0892	0.1724	0.0477	0.6054
Marital_Wid	0.1576	0.2029	-0.0298	0.8761
position_Man	-0.021	0.704	-0.1384	0.0753
position_Oth	-0.1286	0.0358	-0.00491	0.9544
position_Tech	-0.0993	0.0993	0.1213	0.1462
position_Top	-0.0236	0.7974	-0.2725	0.0285
sec_Agricult	-0.0295	0.7763	-0.3014	0.0404
sec_Constr	-0.0781	0.5822	0.297	0.1581
sec_Energy	0.1314	0.1492	-0.0511	0.6858
sec_Fin	-0.0246	0.6612	-0.1792	0.032
sec_Industry	-0.4874	0.0153	-0.1947	0.4646
sec_Manufact	0.1037	0.5188	-0.2171	0.3253
sec_Mining	0.0931	0.3971	-0.2613	0.0766
sec_Service	0.019	0.7048	-0.1116	0.1143
sec_Trade	-0.065	0.3797	-0.1845	0.0629
sec_Trans	-0.0366	0.7936	-0.3308	0.0782
car_Own	-0.011	0.8028	0.0355	0.5729
car_coOwn	-0.0435	0.5335	0.0884	0.3895
real_Own	0.0279	0.555	0.0069	0.9186
real_coOwn	0.0591	0.2521	0.0149	0.8367
reg_ctr_Y	0.0883	0.2199	0.2789	0.0035
reg_ctr_N	0.1176	0.1045	0.2853	0.0033
child_1	-0.1007	0.0983	-0.0999	0.2517
child_2	-0.0229	0.5017	-0.0125	0.7958
child_3	-0.1624	0.1845	-0.1691	0.2862
Unempl_Inyoy	1.2888	0.0927	0.7713	0.4731
LCY_EURRate_Inmom	3.36	0.0007	2.1686	0.1172
LCY_EURRate_Inyoy	-0.5268	0.3717	1.0213	0.2354
CPI_Inqoq	7.1007	<.0001	2.8813	0.2103
SalaryYear_Inyoy	2.0588	0.1233	-2.2999	0.238
d_State_2_NA	0.7618	0.0005	63.8706	0.9879
d_State_2_Tr	0.1723	0.1506	63.788	0.9879
d_State_2_Re	0	.	63.4942	0.988
d_State_2_D1	0	.	57.0944	0.9892
d_State_2_D2	0	.	38.5436	0.9927

Table 6.3 Ordinal regression coefficient Estimations (part 2 – for Revolvers, Delinquent 1, and Delinquent 2)

Parameter	Re		D1		D2	
	Estimate	Pr > ChiSq	Estimate	Pr > ChiSq	Estimate	Pr > ChiSq
Intercept 1	-4.5794	<.0001				
Intercept 2	-2.2054	0.0205	-5.0833	0.0544		
Intercept 3	7.354	<.0001	2.0523	0.4361	86.0055	<.0001
Intercept 4			3.5372	0.1796	86.9383	<.0001
Intercept 5					87.4153	<.0001
UT0_1	-1.4508	<.0001	5.6878	0.0007	5.7243	0.4038
UT0_2	0.1578	0.4532	-3.7157	0.0445	6.3059	0.3752
UT0_3	-0.074	0.6419	-0.2389	0.7538	-5.7916	0.0974
avg_balance_1	0.000039	0.0324	-0.00009	0.4661	0.000849	0.4381
avg_balance_2	-0.00001	0.5684	0.000047	0.7555	-0.00127	0.292
avg_balance_3	-0.00002	0.2488	1.97E-06	0.9797	0.000457	0.1668
dpd_2	-0.0108	0.3166	-0.0189	0.5645	0.00939	0.8361
dpd_3	0.0135	0.1903	-0.0137	0.0003	0.00914	0.7594
b_AvgOB1_to_MaxOB1_1	0.3144	<.0001	-0.434	0.9016	16.6986	0.3523
b_AvgOB2_to_MaxOB2_1	0.0508	0.3662	0.2106	0.8311	2.6256	0.805
b_AvgOB3_to_MaxOB3_1	-0.0672	0.2044	0.8534	0.0226	-0.7406	0.7947
b_TRmax_deb1_To_Limi	-0.128	0.0023	0.9742	0.2917	-376	<.0001
b_TRmax_deb2_To_Limi	0.0496	0.1596	0.0179	0.9577	-22.7971	0.2048
b_TRmax_deb3_To_Limi	0.0684	0.0588	0.231	0.6411	-0.9902	0.3626
b_TRmax_deb1_To_avgO	-0.5102	<.0001	-0.2503	0.4381	14.069	0.0006
b_TRmax_deb2_To_avgO	-0.5585	<.0001	-0.2373	0.3148	1.1191	0.4558
b_TRmax_deb3_To_avgO	-0.2347	0.0018	0.06	0.7706	0.2212	0.6761
b_TRAvg_deb1_to_avgO	0.3213	<.0001	0.0864	0.7803	2.533	0.2777
b_TRAvg_deb2_to_avgO	0.0463	0.2674	-0.049	0.7952	1.1903	0.3219
b_TRAvg_deb3_to_avgO	0.1816	<.0001	0.2031	0.181	0.0245	0.9554
b_TRsum_deb1_to_avgO	0.8894	<.0001	-0.0185	0.962	-4.293	0.2399
b_TRsum_deb2_to_avgO	0.6227	<.0001	0.3596	0.1943	-1.4668	0.2994
b_TRsum_deb3_to_avgO	0.1269	0.0652	-0.1645	0.3679	0.3974	0.423
b_TRsum_deb1_to_TRsu	-0.7482	<.0001	-0.3062	0.0238	7.6448	<.0001
b_TRsum_deb2_to_TRsu	-0.1196	<.0001	-0.0701	<.0001	-0.1298	0.0062
b_TRsum_deb3_to_TRsu	-0.00445	0.4277	-0.00739	0.469	-0.0529	0.0163
sum_deb_num_6	0.00667	0.1816	0.025	0.652	1.0799	0.1902
sum_deb_num_7	0.00529	0.2476	-0.0044	0.8792	0.1364	0.6231
sum_deb_num_8	0.00384	0.4699	0.0231	0.3292	-0.0155	0.8266
b_inactive2	0.3954	0.2159	0	.	0	.
b_inactive3	-0.838	0.0021	-0.1872	0.897	0	.
b_fullpaid1	0	.	0	.	0	.
b_OB_avg_to_eop1ln	0.286	<.0001	-1.6873	0.5682	-20.5815	0.0341
b_pos_flag_0	-0.1616	<.0001	0.1144	0.814	-0.3648	0.8021
b_pos_flag_1	-0.2775	<.0001	0.1507	0.1557	0	.
b_atm_flag_0	-0.3015	<.0001	0.0218	0.8932	0.9199	0.611
b_atm_flag_1	-0.2671	<.0001	-0.2395	0.1048	0.0951	0.8446
l_ChangeFlag	-0.0912	0.0286	0.0687	0.5924	-0.00347	0.9927
l_ch1_ln	-0.7288	0.0023	1.2759	0.6175	1.511	0.8929
l_ch1_flag	0.377	0.0053	-0.0241	0.9706	-1.226	0.6232
mob	-0.0127	0.0003	0.00147	0.8904	0.00862	0.7615
b_TRsum_crd1_to_OB1_ln	-0.359	<.0001	-0.1011	0.503	8.9662	<.0001
b_payment_lt_5p_1	0.1244	0.0032	-0.9964	<.0001	-0.7117	0.191
b_maxminOB_limit_1_1	-0.266	<.0001	-4.3158	0.0005	16.2099	0.0449
b_maxminOB_limit_2_1	-0.0646	0.0016	1.405	0.2533	-15.1805	0.0695
b_maxminOB_limit_3_1	0.0495	0.0081	-0.1671	0.2929	3.4395	0.3137
b_OBBias_1_ln	0.0249	0.0956	0.1916	0.0025	0.4641	0.0596
b_OBBias_2_ln	0.0411	0.0056	-0.00489	0.9154	0.1763	0.3407
b_OBBias_3_ln	0.0237	0.1043	-0.032	0.4291	0.1855	0.1289
b_maxminOB_avgOB_1_1	0.3886	<.0001	4.5262	0.0002	-16.028	0.0494
b_maxminOB_avgOB_2_1	0.1573	<.0001	-1.2487	0.3228	15.2753	0.0705
b_maxminOB_avgOB_3_1	0.00363	0.9	0.1855	0.3016	-3.556	0.3167
AgeGRP1	-0.1391	0.0009	-0.0954	0.3621	-0.2101	0.4002
AgeGRP3	0.0651	0.0856	0.0258	0.818	-0.491	0.0864
customer_income_ln	-0.0327	0.3372	-0.0199	0.8763	-0.3405	0.3137
Edu_High	0.1257	0.0011	0.1852	0.0559	0.2481	0.3224
Edu_Special	0.0654	0.0791	0.00955	0.9159	0.1138	0.5996
Edu_TwoDegree	0.2189	0.0111	0.7978	0.0157	0.9718	0.2917

Parameter	Re		D1		D2	
	Estimate	Pr > ChiSq	Estimate	Pr > ChiSq	Estimate	Pr > ChiSq
Marital_Civ	0.0179	0.7591	-0.2192	0.1252	-0.0252	0.934
Marital_Div	-0.00366	0.9292	0.00319	0.979	0.2331	0.4119
Marital_Sin	-0.0522	0.2371	-0.0426	0.6898	-0.308	0.2487
Marital_Wid	-0.033	0.6906	0.0711	0.8011	-1.3468	0.2451
position_Man	-0.0274	0.49	-0.1358	0.2574	-0.0745	0.8068
position_Oth	-0.032	0.4144	-0.1031	0.3657	-0.1433	0.618
position_Tech	-0.0201	0.5917	-0.2064	0.0285	-0.0213	0.9255
position_Top	0.1822	0.0113	0.0288	0.9132	0.3237	0.692
sec_Agricult	-0.00667	0.9243	-0.267	0.1459	0.1043	0.8235
sec_Constr	-0.1382	0.1474	-0.1929	0.4186	-0.076	0.8858
sec_Energy	0.0542	0.3783	0.0963	0.6074	-0.583	0.2542
sec_Fin	0.0134	0.7612	-0.00956	0.9586	0.0959	0.8719
sec_Industry	-0.0626	0.6221	0.4449	0.2456	-0.4648	0.6668
sec_Manufact	-0.00326	0.9717	-0.2596	0.2944	0.7647	0.1119
sec_Mining	0.0357	0.5958	-0.1056	0.5453	-0.6445	0.1814
sec_Service	-0.0895	0.0084	-0.0692	0.4648	0.1423	0.5468
sec_Trade	-0.1724	0.0002	-0.1954	0.0779	0.3029	0.2709
sec_Trans	-0.0914	0.3241	0.1564	0.5022	-0.6809	0.3491
car_Own	0.0787	0.015	0.4955	<.0001	-0.1702	0.611
car_coOwn	-0.0388	0.4309	0.1459	0.3347	-0.1862	0.6285
real_Own	-0.00975	0.7576	0.00765	0.9316	0.0152	0.9453
real_coOwn	-0.0423	0.2037	0.0849	0.3219	-0.0155	0.9406
reg_ctr_Y	0.0434	0.3764	-0.1568	0.2355	0.1005	0.7778
reg_ctr_N	0.0441	0.3675	-0.2264	0.0812	0.0896	0.8003
child_1	-0.0711	0.0853	-0.1383	0.1955	-0.57	0.0319
child_2	-0.00763	0.7446	0.0398	0.5714	-0.2502	0.1537
child_3	-0.096	0.2226	0.4017	0.1393	0.1494	0.8145
Unempl_Inyoy	2.9852	<.0001	3.8952	0.0069	-8.4511	0.024
LCY_EURRate_Inmom	-2.3598	0.0003	-3.2493	0.0665	0.2764	0.946
LCY_EURRate_Inyoy	1.109	0.0066	1.847	0.1399	1.1928	0.7123
CPI_Inqoq	-4.4991	<.0001	-7.5881	0.0271	-15.7537	0.0955
SalaryYear_Inyoy	-2.4495	0.008	-6.5951	0.02	-15.1797	0.046
d_State_2_2_NA	-1.2641	0.2616	0	.	0	.
d_State_2_2_Tr	-1.1172	0.3114	0	.	0	.
d_State_2_2_Re	-1.5624	0.1559	0.6225	0.7947	0	.
d_State_2_2_D1	-3.4797	0.0005	-1.1011	0.592	0.931	0.7467
d_State_2_2_D2	-1.8875	0.0122	-1.1825	0.299	-0.5427	0.7345

The use of the estimation of ordinal logistic regression for analysis of signs and impacts of covariates on the outcome has a weak feature – the difference between the probability of transition to states are explained with intercepts only. Thus, for example, the chances of transition from inactive state to other states such as inactive, transactor, and revolver according to the model interpretation differ for the same value for all cases because of 3 different values of intercept (0, -86.2142, and -85.921). According to the ordinal logistic regression model, two cases with different behavioural characteristics have different probabilities of transition, for instance, from inactive to revolver state and from inactive to inactive state, but the difference between the probability of transition from inactive to the revolver and from inactive to inactive states be the same for both cases.

6.2.2 Ordinal and Multinomial Logistic Regression Validation

We test the predictive power of the multinomial and ordinal regression model with two characteristics: Kolmogorov-Smirnov and Gini coefficients. Table 6.4 presents the results of a comparative analysis between Ordinal and Multinomial Regression Models. The columns ‘Difference values’ are calculated as (Gini Multinomial – Gini Ordinal) and (KS Multinomial – KS ordinal) and Ratio values are calculated as (Gini Multinomial/Gini Ordinal) and (KS Multinomial/KS Ordinal). These states are used as a measure for comparison of KS and Gini coefficients for two regressions.

Table 6.4 Comparative Analysis of Ordinal and Multinomial Regression Models Validation – Test Sample

		Ordinal		Multinomial		Difference		Ratio	
		Gini	KS	Gini	KS	Gini	KS	Gini	KS
NA	Na	32.33%	25.72%	33.11%	26.05%	0.78%	0.33%	1.02	1.01
	Tr	32.76%	27.37%	46.20%	35.50%	13.45%	8.13%	1.41	1.30
	Re	29.68%	21.84%	29.99%	22.14%	0.32%	0.30%	1.01	1.01
Tr	Na	46.39%	34.79%	46.70%	34.95%	0.32%	0.17%	1.01	1.00
	Tr	20.00%	15.36%	45.82%	35.04%	25.82%	19.68%	2.29	2.28
	Re	36.54%	27.04%	38.16%	28.47%	1.62%	1.43%	1.04	1.05
Re	Na	46.48%	46.28%	53.30%	52.94%	6.82%	6.66%	1.15	1.14
	Tr	63.47%	60.06%	66.02%	62.16%	2.55%	2.10%	1.04	1.03
	Re	61.49%	56.88%	64.36%	59.12%	2.87%	2.25%	1.05	1.04
	D1	77.56%	69.24%	68.41%	67.31%	-9.15%	-1.93%	0.88	0.97
D1	Tr	38.15%	34.75%	11.67%	13.12%	-26.48%	-21.63%	0.31	0.38
	Re	56.68%	43.24%	58.31%	44.17%	1.63%	0.93%	1.03	1.02
	D1	32.48%	27.65%	51.47%	38.51%	19.00%	10.86%	1.58	1.39
	D2	60.04%	44.29%	63.92%	49.01%	3.88%	4.72%	1.06	1.11
D2	Re	65.55%	50.27%	75.11%	59.67%	9.55%	9.41%	1.15	1.19
	D1	37.86%	29.05%	62.93%	49.21%	25.07%	20.16%	1.66	1.69
	D2	26.34%	22.57%	68.96%	55.33%	42.62%	32.76%	2.62	2.45
	Df	60.28%	48.83%	68.96%	55.33%	8.67%	6.50%	1.14	1.13

Generally, multinomial logistic regression has shown better results than the ordinal one. Only for models Delinquency 1 to Transactor and Revolver to Delinquency 1 the Ordered logistic regression has higher coefficients of Gini and KS: 38.15% and 34.75% vs. 11.67% and 13.12%, and 77.56% and 69.24% versus 68.41% and 67.31% respectively. In the majority of cases, the multinomial regression has a slight advantage over ordinal regression. However, for prediction of Transactor to Transactor and

Delinquent 2 to Delinquent 2 the Multinomial regression gives better results Gini and KS 45.82% and 35.04% vs .20.00% and 15.36% respectively.

According to the results of the empirical investigation, the multinomial logistic regression has shown better predictive power given by Kolmogorov-Smirnov and Gini coefficients than ordered logistic regression. However, the results of the two regressions are not significantly different except in a few segments with a lower number of observations (for example, Delinquent 1 to Transactor) than in the segments with stable validation results (for example, Revolver to Revolver). Ordinal regression has also given higher KS and Gini values for a couple of segments, and this may be caused by algorithms features. This may be a topic of the further investigations. Thus we have chosen multinomial logistic regression from the set of generalised regression models and will use it for further investigation. As the next step, we test conditional binary logistic regression and compare it with the multinomial logistic regression.

6.3 The multistage binary logistic regression model

6.3.1 Regression coefficients estimation

Prediction models priority and order is an important factor in the estimation of multistage binary logistic regression coefficients and the predictive power of the model. For example, it could be more important to have a stronger model for a transition to the delinquency states than to the transactor state. Another factor which impacts on the sub-models order is the number of observations in a stage. For example, for 90% of the population in state A, 5% in B, and 5% in state C we build the model based on the sample with 90% outcomes in A versus 10% in B and C merged in a single group at the first stage, and then at the second stage we build the model with a sample based on 50% outcomes in B versus 50% outcomes in C excluding all observations with outcome A.

We apply the following order of the predicted outcomes (targets):

Model 1: NA -> NA -> Re

Model 2: Tr -> NA -> Re

Model 3: Re -> D1 -> Re -> NA

Model 4: D1 -> D2 -> Re

Model 5: D2 -> Df -> Re -> D1

This means that, for example, for the Model 1 we select all non-active accounts. At the first step, we split accounts into non-active and active and build a model for the estimation of the probability to stay in the non-active state. Then for all accounts, which became active, we have segments of revolvers and transactors, and we build the model for the estimation of the probability to move to revolver state (otherwise to stay in transactor state).

For Model 2 we select all transactors and at the first step build the model for the probability to move to the inactive state, at the second step for the segment of account, which stay active, we built the model for the probability to move in revolver state (otherwise to stay in transactor state).

For Model 3 for revolvers at the first step we build the model for the probability to be delinquent, then for a non-delinquent segment we build the model for the probability to stay revolver, then for non-revolvers the probability to move to a non-active state (otherwise to move to transactor state).

Moreover, we use the same logic for delinquent Model 4 and Model 5 segments. For the SAS code, which reflects a general logic of the conditional models building sees *Appendix 2*.

The binary logistic regression can be used for detailed analysis of the variables' significance and their impact on the probability of transition. Table 6.6, Table 6.8, Table 6.10, Table 6.12, and Table 6.14 present the estimated coefficients of binary logistic regression, standard errors, and results of Chi-Squared test for a coefficients significance. The first row of the table heading shows origin state, the second row of the table heading show the destination state. The period of prediction is selected the same as for multinomial and ordinal regression -1 month.

Because of the transactor state is selected as a basic state, the probability of transition to this state is defined as a one minus sum of transitions from state S_t to all possible states S_{t+1} except the transactor state $1 - \sum Pr(S_{t+1}|S_t)$.

We estimate the binary logistic regression with event marked as '1' and non-event marked as '0' for the event probability modelled as Target=0. So the higher values of the coefficients of regression correspond to the higher probability of an event.

The transition from the *Inactive state* is estimated with binary logistic regression as a transition to inactive and revolver state (Table 6.6).

Month on Book is a significant predictor for the transitions from Inactive state to Inactive state with a negative sign -0.0298 can explain that a more extended period on book corresponds to lower probability to stay inactive in the next month, but insignificant for the transition to Revolver state.

The change of credit limit in the previous month (*l_ch1_flag*) with coefficient -0.477 decreases the probability to stay in an inactive state; however, Chi-Squared equal to 0.12 makes this conclusion questionable.

All behavioural variables for the current month (observation point *t* marked with prefix _1) have estimated coefficients equal to zero because all variables related to outstanding balance and transactions must have the same value, zero. Generally, behavioural variables for previous 2nd and 3rd months are insignificant. Application characteristics are generally insignificant for transitions from an inactive state.

Among macroeconomic variables, we can mention that higher monthly foreign currency rate (*LCY_EURRate_lnmom*) and quarterly change in CPI (*CPI_Inqoq*) decrease the probability of an account to stay inactive.

Table 6.5 Number of observations in Non-active state

Transition Stage	Stage 1 - NA vs TR & RE	Stage 2 - TR vs RE
Inactive	4 141	856
Transactor		3 285
Revolver	21 588	

Table 6.6 Conditional binary logistic regression Estimations for Non-Active state

From To	NA NA	Re						
		Parameter	Estimate	Standard Error	Pr > ChiSq	Estimate	Standard Error	Pr > ChiSq
	Intercept	-9.834	236.6	0.9669	-33.0561	218.4	0.8797	
	UTO_1	0	.	.	0	.	.	
	UTO_2	0.0197	0.4524	0.9652	-1.9592	1.0485	0.0617	
	UTO_3	0.2624	0.2834	0.3544	0.4052	0.6772	0.5496	
	avg_balance_1	0	.	.	0	.	.	
	avg_balance_2	0.00006	0.000059	0.3044	0.000369	0.000128	0.0039	
	avg_balance_3	-8.91E-06	0.000036	0.8043	-0.00017	0.000087	0.0533	
b_AvgOB1_to_MaxOB1_In	0	.	.	0	.	.	.	
b_AvgOB2_to_MaxOB2_In	-0.1401	0.1223	0.2518	-0.1391	0.1571	0.376		
b_AvgOB3_to_MaxOB3_In	0.0056	0.0868	0.9485	0.2848	0.2167	0.1887		
b_TRmax_deb1_To_Limit	0	.	.	0	.	.	.	
b_TRmax_deb2_To_Limit	-0.1064	0.1644	0.5173	0.1465	0.4666	0.7535		
b_TRmax_deb3_To_Limit	0.0531	0.1533	0.7291	0.4408	0.2869	0.1245		
b_TRavg_deb1_to_avgO	0	.	.	0	.	.	.	
b_TRavg_deb2_to_avgO	-0.2526	0.1173	0.0312	-0.1697	0.2496	0.4966		

From To	NA NA	Re					
		Parameter	Estimate	Standard Error	Pr > ChiSq	Estimate	Standard Error
	b_TRavg_deb3_to_avgO	-0.3132	0.0929	0.0008	-0.1605	0.1883	0.3938
	b_TRsum_deb1_to_avgO	0	.	.	0	.	.
	b_TRsum_deb2_to_avgO	0.3	0.1157	0.0095	0.2722	0.2449	0.2663
	b_TRsum_deb3_to_avgO	0.334	0.0912	0.0002	0.2287	0.1835	0.2128
	b_TRsum_deb1_to_TRsumcrd	0	.	.	0	.	.
	b_TRsum_deb2_to_TRsumcrd	0.0199	0.0122	0.1038	-0.0161	0.0262	0.5383
	b_TRsum_deb3_to_TRsumcrd	0.00566	0.00985	0.5657	0.0138	0.0194	0.4773
	sum_deb_num_1	0	.	.	0	.	.
	sum_deb_num_2	-0.0393	0.0307	0.2012	-0.078	0.0669	0.244
	sum_deb_num_3	-0.0123	0.0217	0.57	-0.0641	0.0426	0.1325
	b_OB_avg_to_eop1ln	-0.0151	0.026	0.5609	0.0519	0.0478	0.2779
	b_pos_flag_1	0	.	.	0	.	.
	b_atm_flag_1	0	.	.	0	.	.
	l_ch1_ln	1.4251	0.4706	0.0025	0.8542	0.8497	0.3147
	l_ch1_flag	-0.477	0.3068	0.12	-0.4391	0.6592	0.5053
	Mob	-0.0298	0.00497	<.0001	-0.00156	0.011	0.8871
	b_TRsum_crd1_to_OB1	0	.	.	0	.	.
	b_payment_lt_5p_1	-0.113	0.4293	0.7924	-1.136	0.7608	0.1354
	b_maxminOB_limit_1_ln	0	.	.	0	.	.
	b_maxminOB_limit_2_ln	-0.0217	0.0147	0.1394	-0.0344	0.0302	0.2553
	b_maxminOB_limit_3_ln	-0.0364	0.0146	0.0128	-0.0152	0.0346	0.6607
	b_OBbias_1_ln	0	.	.	0	.	.
	b_OBbias_2_ln	-0.0742	0.0582	0.2023	-0.0514	0.0964	0.5937
	b_OBbias_3_ln	-0.032	0.04	0.4231	0.0706	0.0976	0.4698
	b_maxminOB_avgOB_1_ln	0	.	.	0	.	.
	b_maxminOB_avgOB_2_ln	-0.0109	0.07	0.8767	-0.1395	0.1211	0.2495
	b_maxminOB_avgOB_3_ln	0.0919	0.0421	0.029	0.1618	0.1067	0.1293
	AgeGRP1	0.1498	0.0637	0.0187	0.043	0.1406	0.7595
	AgeGRP3	0.0452	0.0556	0.4163	0.0302	0.1246	0.8084
	customer_income_ln	-0.021	0.0409	0.6075	0.0534	0.0884	0.5459
	Edu_High	0.049	0.0638	0.4425	0.2181	0.1485	0.142
	Edu_Special	-0.00922	0.0643	0.8859	0.0782	0.1516	0.6057
	Edu_TwoDegree	0.1	0.1209	0.408	0.3585	0.2583	0.1651
	Marital_Civ	0.00709	0.0982	0.9424	-0.393	0.2387	0.0997
	Marital_Div	-0.0368	0.0638	0.5644	-0.0625	0.1502	0.6772
	Marital_Sin	-0.0944	0.0659	0.1518	0.1292	0.1453	0.374
	Marital_Wid	-0.1814	0.1253	0.1478	0.1295	0.2916	0.657
	position_Man	0.0206	0.0557	0.711	-0.0437	0.1243	0.7251
	position_Oth	0.1333	0.0616	0.0304	-0.1024	0.1422	0.4716
	position_Tech	0.0983	0.059	0.0956	0.0976	0.1321	0.46
	position_Top	0.0303	0.0924	0.7433	0.1828	0.192	0.3412
	sec_Constr	0.0513	0.1415	0.7171	-0.00451	0.3237	0.9889
	sec_Energy	-0.164	0.0884	0.0635	0.0687	0.1936	0.7228
	sec_Fin	0.0265	0.0535	0.6203	0.1303	0.1141	0.2536
	sec_Service	-0.04	0.0466	0.3899	-0.1109	0.1076	0.3027
	sec_Trans	0.0909	0.136	0.504	0.734	0.2622	0.0051
	car_Own	0.0307	0.0443	0.489	0.2884	0.0961	0.0027
	car_coOwn	0.0297	0.0706	0.6736	-0.0514	0.1643	0.7542
	real_Own	-0.027	0.0478	0.5722	-0.0595	0.1069	0.5775
	real_coOwn	-0.0501	0.0521	0.3357	0.0964	0.1151	0.402
	reg_ctr_Y	-0.0967	0.0725	0.182	0.00297	0.1554	0.9847
	reg_ctr_N	-0.1355	0.0719	0.0597	-0.0421	0.1553	0.7866
	child_1	0.0836	0.0615	0.1739	-0.1325	0.1377	0.3357
	child_2	0.0159	0.0343	0.6423	0.0233	0.0763	0.7599
	child_3	0.1368	0.1241	0.2702	0.0486	0.2744	0.8595
	Unempl_Inyoy	-1.0077	0.7669	0.1889	2.9161	1.738	0.0934
	UAH_EURRate_Immom	-3.6896	0.997	0.0002	-1.5121	2.1791	0.4878
	UAH_EURRate_Inyoy	0.658	0.5907	0.2653	3.0189	1.3176	0.0219
	CPI_Inqoq	-6.9226	1.5933	<.0001	-2.5203	3.5705	0.4803
	SalaryYear_Inyoy	-2.3284	1.3453	0.0835	-2.7389	3.076	0.3732
	d_State_2_NA	-2.4997	1.4687	0.0888	8.6428	212.6	0.9676
	d_State_2_Tr	-1.4601	1.464	0.3186	9.8773	212.6	0.9629
	d_State_2_Re	-1.3217	1.4555	0.3638	9.1979	212.6	0.9655
	d_State_2_D1	0	.	.	0	.	.
	d_State_2_D2	0	.	.	0	.	.

The estimations for transition probabilities from *Transactor state* to Inactive and Revolver states are presented in Table 6.8. The significant behavioural characteristics for transition of a transactor to the inactive state in the next month are the following covariates: the ratios of the sum of debit transactions (purchases) to the average outstanding balance at the same month for the current month, 1 month before observation point, and 2 months before observation point (*b_TRsum_deb1_to_avgO*, *b_TRsum_deb2_to_avgO*, and *b_TRsum_deb3_to_avgO* respectively) with positive signs and the sum of debit transactions (purchases) to the sum of credit transactions (payments) at the observation point month (*b_TRsum_deb1_to_TRsumcrd*) with a negative sign. The sign of the first characteristic can be explained as the higher values of the ratio can be obtained not only because of a significant amount of spending, but low outstanding balance at the current and previous months, and this is typical for transactors. If purchases exceed the outstanding balance for a long period, it may cause a transition to the revolver state. However, if the outstanding balance is stable, this may mean that a customer pays back the same amounts as they spend.

Month on Book is not a significant variable for both transitions to inactive and revolver states. Credit limit change is insignificant for the transition to revolver states from transactor state but relatively significant for the transition to an inactive state and with a positive sign.

From macroeconomic variables monthly change in the local currency exchange rate to Euro (*LCY_EURRate_lnmom*) affect the transition from the Transactor to the Inactive state and it has a negative sign. Yearly change in the local currency exchange rate to Euro (*UAH_EURRate_lnyoy*) is also significant with a positive sign. Yearly change in average salary level (*SalaryYear_lnyoy*) is significant with a negative sign for the Revolver state transition – the growth in salary means the decrease of transitions to the Revolver state.

The state at the previous month is not a significant parameter for movements from the transactor state.

Table 6.7 Number of observations in Transactor state

Transition Stage	Stage 1 - NA vs TR&RE	Stage 2 - TR vs RE
Inactive	3 199	
Transactor		808
Revolver	3 578	2 391

Table 6.8 Conditional binary logistic regression Estimations for Transactor state

From To	Tr			Re		
Parameter	Estimate	Standard Error	Pr > ChiSq	Estimate	Standard Error	Pr > ChiSq
Intercept	-0.2101	1.9357	0.9136	-8.0356	206.1	0.9689
UT0_1	0.0211	0.4282	0.9606	-1.0013	0.7308	0.1707
UT0_2	-0.414	0.3712	0.2647	0.6578	0.6103	0.2811
UT0_3	0.6571	0.3061	0.0318	-0.8031	0.5039	0.111
avg_balance_1	0.000011	0.000053	0.8296	0.000145	0.000084	0.0833
avg_balance_2	0.000064	0.000046	0.1641	-0.00008	0.000074	0.2681
avg_balance_3	-0.00007	0.000036	0.0665	0.000059	0.000057	0.3021
b_AvgOB1_to_MaxOB1_In	-0.0941	1.4233	0.9473	2.2854	1.4092	0.1048
b_AvgOB2_to_MaxOB2_In	0.0877	0.0868	0.3122	0.1373	0.164	0.4027
b_AvgOB3_to_MaxOB3_In	0.3074	0.1209	0.011	-0.1684	0.2291	0.4623
b_TRmax_deb1_To_Limit	0.0609	0.1495	0.6836	-0.2002	0.1744	0.2509
b_TRmax_deb2_To_Limit	0.0309	0.1381	0.8229	-0.0741	0.1637	0.651
b_TRmax_deb3_To_Limit	-0.1147	0.13	0.3777	0.1909	0.1579	0.2267
b_TRavg_deb1_to_avgO	-0.1283	0.1091	0.2399	0.0191	0.1366	0.8886
b_TRavg_deb2_to_avgO	-0.2476	0.0894	0.0056	0.104	0.1201	0.3866
b_TRavg_deb3_to_avgO	-0.2914	0.0887	0.001	0.1524	0.1257	0.2254
b_TRsum_deb1_to_avgO	0.6642	0.1336	<.0001	0.1054	0.1777	0.5531
b_TRsum_deb2_to_avgO	0.3333	0.0846	<.0001	-0.0286	0.113	0.8005
b_TRsum_deb3_to_avgO	0.3472	0.0845	<.0001	-0.0588	0.1173	0.6161
b_TRsum_deb1_to_TRsumcrd	-0.3818	0.0776	<.0001	-0.0956	0.1269	0.4511
b_TRsum_deb2_to_TRsumcrd	-0.00776	0.0125	0.5339	-0.0488	0.0213	0.022
b_TRsum_deb3_to_TRsumcrd	-0.00144	0.0123	0.9071	-0.026	0.0201	0.1962
sum_deb_num_1	0.0365	0.0236	0.1208	-0.017	0.0212	0.4219
sum_deb_num_2	-0.0174	0.0179	0.3301	0.0143	0.019	0.453
sum_deb_num_3	-0.0222	0.0164	0.1738	0.0265	0.0188	0.1591
b_OB_avg_to_eop1In	-0.0271	0.0454	0.5507	-0.0487	0.0814	0.5495
b_pos_flag_1	0.2861	0.0855	0.0008	0.2528	0.127	0.0465
b_atm_flag_1	0.1462	0.0896	0.1026	-0.00134	0.1258	0.9915
I_ch1_In	-0.4988	0.5639	0.3764	-0.7475	1.0102	0.4593
I_ch1_flag	0.6908	0.3389	0.0415	0.1797	0.5565	0.7467
Mob	0.00478	0.00765	0.5321	-0.00436	0.012	0.7168
b_TRsum_crd1_to_OB1	-0.3878	0.0768	<.0001	0.0423	0.1181	0.7206
b_payment_lt_Sp_1	0.3251	0.4365	0.4564	0.6726	0.9421	0.4752
b_maxminOB_limit_1_In	0.0898	0.0253	0.0004	0.0506	0.0459	0.2705
b_maxminOB_limit_2_In	-0.0594	0.0245	0.0152	0.0242	0.0382	0.5272
b_maxminOB_limit_3_In	-0.0576	0.0257	0.0248	-0.0458	0.0398	0.2497
b_OBBias_1_In	0.0563	0.1005	0.5752	0.1313	0.1884	0.486
b_OBBias_2_In	-0.0684	0.0372	0.066	0.0585	0.0765	0.4446
b_OBBias_3_In	0.0192	0.0436	0.66	-0.1454	0.0906	0.1087
b_maxminOB_avgOB_1_In	-0.3789	1.3748	0.7828	2.1146	1.2989	0.1035
b_maxminOB_avgOB_2_In	0.1073	0.0458	0.0192	0.1416	0.0882	0.1082
b_maxminOB_avgOB_3_In	0.1532	0.0535	0.0042	-0.0408	0.1019	0.6887
AgeGRP1	0.2135	0.0959	0.0261	-0.1888	0.1586	0.2339
AgeGRP3	0.0714	0.0856	0.4044	0.2224	0.1396	0.1112
customer_income_In	-0.1689	0.0631	0.0074	0.0146	0.0965	0.8795
Edu_High	-0.2703	0.0921	0.0033	0.145	0.1516	0.3386
Edu_Special	-0.1432	0.0923	0.1208	0.0846	0.1555	0.5864
Edu_TwoDegree	-0.3065	0.1787	0.0864	-0.1416	0.278	0.6104
Marital_Civ	-0.0547	0.1405	0.6971	0.0426	0.2239	0.849
Marital_Div	0.0321	0.0946	0.7346	-0.2438	0.1618	0.1319
Marital_Sin	-0.0129	0.0993	0.8965	0.2109	0.1619	0.1929
Marital_Wid	-0.0256	0.2035	0.9	-0.3246	0.3587	0.3655
position_Man	0.1658	0.0837	0.0474	-0.0364	0.13	0.7792
position_Oth	0.024	0.0912	0.792	0.1651	0.1455	0.2565

From To	Tr Na	Re				
Parameter	Estimate	Standard Error	Pr > ChiSq	Estimate	Standard Error	Pr > ChiSq
position_Tech	-0.101	0.086	0.24	-0.1043	0.1513	0.4905
position_Top	0.2449	0.1348	0.0693	-0.2901	0.2072	0.1616
sec_Constr	-0.3647	0.2186	0.0953	0.2521	0.383	0.5103
sec_Energy	-0.081	0.1288	0.5293	0.00121	0.2198	0.9956
sec_Fin	0.066	0.0851	0.4379	-0.1897	0.1301	0.1449
sec_Service	0.0362	0.0691	0.5999	0.00468	0.1112	0.9664
sec_Trans	0.1537	0.1986	0.439	-0.4081	0.3449	0.2368
car_Own	-0.0708	0.0674	0.2936	-0.0971	0.1094	0.3746
car_coOwn	-0.1185	0.1091	0.2773	0.0136	0.1767	0.9384
real_Own	0.0179	0.0724	0.8051	0.0109	0.1198	0.9274
real_coOwn	0.0681	0.0777	0.381	0.2333	0.1253	0.0627
reg_ctr_Y	-0.3307	0.1041	0.0015	0.0244	0.1521	0.8725
reg_ctr_N	-0.3348	0.1039	0.0013	0.1288	0.1541	0.4034
child_1	0.1092	0.0934	0.2424	-0.00352	0.1565	0.982
child_2	0.00958	0.0516	0.8528	-0.00792	0.0863	0.9269
child_3	0.1692	0.1703	0.3205	-0.0731	0.2746	0.79
Unempl_Inoy	-1.0652	1.1467	0.353	1.4636	1.8377	0.4258
LCY_EURRate_Inmom	-2.9506	1.4789	0.046	-1.9416	2.3883	0.4162
LCY_EURRate_Inoy	-0.9868	0.9093	0.2779	3.0281	1.4607	0.0382
CPI_Inqoq	-3.3757	2.4601	0.17	-1.6733	3.9484	0.6717
SalaryYear_Inoy	1.4554	2.0787	0.4838	-6.1727	3.3833	0.0681
d_State_2_NA	-0.9407	1.9575	0.6308	0.5934	168.3	0.9972
d_State_2_Tr	-0.1788	1.9445	0.9267	0.839	168.3	0.996
d_State_2_Re	0.1208	1.9396	0.9503	0.6002	168.3	0.9972
d_State_2_D1	-0.0515	1.9625	0.979	0.2816	168.3	0.9987
d_State_2_D2	-0.9604	2.3539	0.6833	-4.9531	206.1	0.9808

The coefficients estimates for the transition from the *Revolver state* are presented in Table 6.10. The most of the significant coefficients among all tested characteristics have been obtained for the transition from the Revolver state to the Delinquency 1 state. This transition corresponds with the estimation of the probability of delinquency and default, which is widely used in credit scoring.

One of the important remarks is that the covariates, which use the previous state, are significant for the transition from revolver state to the delinquency 1 state only but for transitions to other states the covariates are not significant.

High current utilisation rate (UT0_1) decreases the probability of the revolver account to stay at the revolver state (estimate coefficient value is equal to -1.2343 with Pr>Chi-Squared <0.0001). For the change of utilisation rate by 0.1 from 0.5 to 0.6, the probability of the transition from the revolver state to the delinquent state will decrease by 0.0279. The elasticity of the probability of transition to revolver state conditional on the delta of the utilisation rate can be computed, for example, for the change of the utilisation rate by 10% from 0.5 to 0.6 as

$$PD(S_{t+1} = RE | S_t = RE, UT_1 = 0.5) - PD(S_{t+1} = RE | S_t = RE, UT_1 = 0.6) = \\ 1/(1 + e^{(-1.2343*0.5)}) - 1/(1 + e^{(-1.2343*0.6)}) \sim 0.027$$

for the given case, where the value of the sum of all coefficients of the logistic regression equation, excluding coefficient related to the utilisation rate at time t (UT_1), is equal to zero:

$$e^{b_0+b_1w_1+\dots+b_nw_n} = e^{0+b_{UT_1}\cdot UT_1},$$

where b_1, \dots, b_n are estimated coefficients of regression, w_1, \dots, w_n are values of covariates, n is the number of covariates in the equation.

The transition to the delinquent 1 state is less sensitive to the current utilisation rate value with an estimated coefficient -0.4263.

Some behavioural characteristics are significant for all transitions. The logarithm of the ratio of the monthly sum of purchases to the monthly sum of payments (b_TRsum_deb1_to_TRsumcrd) is significant for all three states, has a positive sign for the probability of transition being inactive and negative signs for the chances to stay a revolver or to go to delinquency – the excess of spending overpayments causes an account to go to revolver state rather than to inactive one. A positive POS transaction flag says that a customer's chances to stay a revolver become slight lower (coefficient estimate is 0.0782) and a positive ATM cash withdrawals transaction gives more chances to stay a revolver (-0.28).

Application characteristics, as for inactive and transactor states, are mostly insignificant. Only some segments such as High education, Energy sector, one child in the family have Chi-Squared values close to zero for the transition to the revolver state. On the other hand, for the transition to delinquent state all application characteristics are significant. For example, the logarithm of the ratio of the customer income and to the ratio of average market salary has a positive sign with the probability to be inactive – higher salary means higher usage of credit cards, and negative sign to the chance of going to delinquency – credit cardholder may overestimate their own ability to pay the debt.

Some macroeconomic characteristics are significant and show an impact on the probabilities of transition. For example, the logarithm of the yearly change of unemployment rate has a positive correlation with the chance of a transition to inactive state (coefficient estimation -1.4603) and a negative correlation with the chance of a transition to delinquent state (coefficient estimation 2.5014).

Table 6.9 Number of observations in Revolver state

Transition Stage	Stage 1 - D1 vs NA&TR&RE	Stage 2 - RE vs NA&TR	Stage 3 - NA vs TR
Inactive		5 832	5 079
Transactor	159 855		753
Revolver		154 023	
Delinquent 1	3 201		

Table 6.10 Conditional binary logistic regression Estimations for Revolver state

From To	Re NA	Re		D1		
Parameter	Estimate	Pr > ChiSq	Estimate	Pr > ChiSq	Estimate	Pr > ChiSq
Intercept	11.6269	0.9415	-0.4331	0.6252	1.9826	<.0001
UTO_1	-0.6045	0.2367	-1.2343	<.0001	-0.4263	<.0001
UTO_2	1.8955	0.0024	0.1307	0.5509	-0.2742	<.0001
UTO_3	-1.3101	0.0101	-0.1059	0.5424	0.013	<.0001
avg_balance_1	-0.00002	0.7758	-0.00004	0.0871	0.000033	<.0001
avg_balance_2	-0.00003	0.6673	-0.00002	0.4027	7.64E-06	<.0001
avg_balance_3	0.000019	0.7684	6.52E-06	0.7635	-0.00003	<.0001
b_AvgOB1_to_MaxOB1_In	0.347	0.0576	0.2153	0.0019	0.0397	<.0001
b_AvgOB2_to_MaxOB2_In	0.1367	0.4422	-0.1015	0.0893	0.3687	<.0001
b_AvgOB3_to_MaxOB3_In	0.1839	0.308	-0.1072	0.0611	0.0595	<.0001
b_TRmax_deb1_To_Limit	0.4715	0.1074	0.0519	0.1862	-0.3037	<.0001
b_TRmax_deb2_To_Limit	0.0238	0.9243	0.0816	0.0238	-0.0732	<.0001
b_TRmax_deb3_To_Limit	-0.1112	0.6723	0.0759	0.042	0.0431	<.0001
b_TRavg_deb1_to_avgO	-0.0394	0.8264	0.2156	<.0001	-0.0037	<.0001
b_TRavg_deb2_to_avgO	0.2049	0.1908	0.0614	0.1199	0.1942	<.0001
b_TRavg_deb3_to_avgO	-0.0348	0.8235	0.1242	0.0077	0.0551	<.0001
b_TRsum_deb1_to_avgO	-0.2937	0.1775	-0.0504	0.4033	0.257	<.0001
b_TRsum_deb2_to_avgO	-0.2223	0.1299	-0.0888	0.0151	-0.0116	<.0001
b_TRsum_deb3_to_avgO	-0.00868	0.9536	-0.1005	0.0226	-0.00909	<.0001
b_TRsum_deb1_to_TRsumcrd	0.6123	<.0001	-0.2428	<.0001	-0.2977	<.0001
b_TRsum_deb2_to_TRsumcrd	0.00302	0.8797	-0.0222	0.0048	-0.188	<.0001
b_TRsum_deb3_to_TRsumcrd	-0.0305	0.1354	0.00353	0.6537	-0.0136	<.0001
sum_deb_num_1	0.0496	0.1847	-0.00397	0.5405	0.0118	<.0001
sum_deb_num_2	0.0245	0.4336	0.00504	0.3175	0.0185	<.0001
sum_deb_num_3	0.0156	0.6573	-0.0055	0.4957	0.00875	<.0001
b_OB_avg_to_eop1ln	0.00328	0.9216	0.3025	<.0001	0.0168	<.0001
b_pos_flag_1	0.0235	0.8607	0.0782	0.0765	-0.4063	<.0001
b_atm_flag_1	-0.4455	0.0013	-0.2831	<.0001	-0.0179	<.0001
l_ch1_In	1.5192	0.1465	-0.5201	0.0666	-0.1135	<.0001
l_ch1_flag	-0.4399	0.4097	0.1364	0.4187	0.323	<.0001
Mob	-0.00271	0.8227	-0.0028	0.5086	-0.015	<.0001
b_TRsum_crd1_to_OB1_	0.1716	0.0936	-0.109	0.0001	0.0985	<.0001
b_payment_lt_5p_1	-0.1587	0.4749	0.0989	0.0988	-0.0106	<.0001
b_maxminOB_limit_1_In	0.1821	0.0002	-0.2575	<.0001	-0.099	<.0001
b_maxminOB_limit_2_In	-0.0368	0.421	0.0309	0.0612	-0.1751	<.0001
b_maxminOB_limit_3_In	-0.00487	0.9171	0.0711	<.0001	-0.1112	<.0001
b_OBbias_1_In	0.1764	0.0034	-0.1031	<.0001	0.0987	<.0001
b_OBbias_2_In	-0.0681	0.2533	0.00175	0.9286	0.0782	<.0001
b_OBbias_3_In	-0.0159	0.7905	-0.0102	0.6008	0.0505	<.0001
b_maxminOB_avgOB_1_In	-0.1372	0.1163	0.5983	<.0001	-0.1203	<.0001
b_maxminOB_avgOB_2_In	0.1939	0.0188	0.0906	0.0016	0.1382	<.0001
b_maxminOB_avgOB_3_In	0.1094	0.1939	0.0789	0.0053	0.0295	<.0001
AgeGRP1	0.3151	0.0328	-0.0975	0.0617	-0.0569	<.0001
AgeGRP3	-0.0119	0.9263	-0.041	0.3727	0.2138	<.0001
customer_income_In	0.2956	0.0071	0.0293	0.4451	-0.089	<.0001
Edu_High	-0.2704	0.0806	0.1025	0.0404	0.0465	<.0001
Edu_Special	-0.222	0.1583	0.0654	0.1939	0.0117	<.0001
Edu_TwoDegree	-0.3484	0.2176	0.1225	0.2182	0.1866	<.0001
Marital_Civ	-0.2417	0.2397	0.0489	0.5115	-0.031	<.0001
Marital_Div	-0.2159	0.1196	0.000846	0.9867	-0.0554	<.0001
Marital_Sin	0.1976	0.1906	-0.0631	0.2542	-0.0527	<.0001
Marital_Wid	-0.6092	0.0235	-0.1121	0.3032	0.0686	<.0001

From To	Parameter	Re		Re		D1	
		NA	Estimate	Pr > ChiSq	Estimate	Pr > ChiSq	Estimate
	position_Man	-0.04	0.755	0.00267	0.9545	-0.00444	0.0013
	position_Oth	0.0973	0.5084	-0.0854	0.088	0.0218	<.0001
	position_Tech	0.4326	0.0032	-0.0647	0.1654	0.0458	<.0001
	position_Top	-0.1603	0.4487	0.2677	0.0006	-0.0455	<.0001
	sec_Constr	0.4915	0.3067	-0.0551	0.6589	-0.1167	<.0001
	sec_Energy	0.2209	0.3379	0.1449	0.0425	0.0602	<.0001
	sec_Fin	-0.7702	<.0001	-0.00815	0.8647	0.1464	<.0001
	sec_Service	0.1747	0.1323	-0.0249	0.5172	-0.0263	<.0001
	sec_Trans	0.1209	0.7205	0.0562	0.6173	-0.012	0.0002
	car_Own	-0.0984	0.3549	0.0882	0.0202	0.0406	<.0001
	car_coOwn	0.00809	0.9624	-0.0461	0.4469	-0.0326	<.0001
	real_Own	0.00104	0.9927	0.0571	0.1485	-0.0491	<.0001
	real_coOwn	-0.0683	0.5744	-0.00102	0.981	-0.0459	<.0001
	reg_ctr_Y	0.0609	0.7187	0.00106	0.9856	0.0557	<.0001
	reg_ctr_N	0.1626	0.3388	0.0208	0.7207	0.1004	<.0001
	child_1	0.2719	0.0472	-0.1978	<.0001	0.0843	<.0001
	child_2	0.2041	0.0076	-0.0594	0.0355	0.046	<.0001
	child_3	0.4025	0.1483	-0.138	0.1475	-0.0437	<.0001
	Unempl_Inoy	-1.4603	0.4243	1.9512	0.0029	2.5014	<.0001
	LCY_EURRate_Inmom	3.1119	0.1677	-1.7105	0.0316	-0.7381	<.0001
	LCY_EURRate_Infoy	-1.6841	0.2241	0.8147	0.0988	0.865	<.0001
	CPI_Inqoq	-0.00936	0.998	-2.9802	0.0257	-0.6156	<.0001
	SalaryYear_Infoy	3.8199	0.2359	-0.208	0.8553	-2.7582	<.0001
	d_State_2_NA	-23.8422	0.971	-0.8671	0.5082	-5.7195	<.0001
	d_State_2_Tr	-23.7592	0.9711	-1.3751	0.2919	-5.3755	<.0001
	d_State_2_Re	-23.2228	0.9717	-1.7518	0.1784	-5.47	<.0001
	d_State_2_D1	-23.0015	0.972	-1.7236	0.1893	-6.1373	<.0001
	d_State_2_D2	-22.6246	0.9715	-1.1535	0.4222	31.7992	0.9997

The transition from Delinquent DPD 1-30 state has two models: go forward to deeper delinquency 30-61 days in the next month or return to the revolver state (Table 6.12). The probability of staying in the same Delinquency 1 state is computed according to the full probability formula as the difference between one and the sum of probabilities to go to the Delinquency 2 and the Revolver states. In a customer pays back the full amount, he/she goes to the inactive or the transactor state. However, because we have few such cases in the data sample, these transitions have been merged with the revolver state.

Table 6.12 shows that the utilisation rate at the current month (UT0_1) is a significant characteristic for the transition from the delinquency 1 state to the delinquency 2 state with a positive sign – higher utilisation rate increases the probability of the transition to deeper delinquency state, but the utilisation rate in the previous month has the opposite sign. This contradicts with the view that customers try to use full credit limit before going to default.

A large number of insignificant covariates suggests that it can be difficult to make accurate predictions from some states because of the low number of observations.

Early delinquency stage accounts might have good payment history and be similar rather to a revolver state segment. If the account had a poor payment history, it may have a higher probability of transition to higher buckets of delinquency and default. However, in the scope of this Chapter, we investigate short-term observation period limited with 3 months of behaviour.

Table 6.11 Number of observations in Delinquent 1 state

Transition Stage	Stage 1 - D2 vs RE&D1	Stage 2 - RE & D1
Inactive		
Transactor	3 287	
Revolver		2 331
Delinquent 1		956
Delinquent 2	743	

Table 6.12 Conditional binary logistic regression Estimations for Delinquency 1 month state

From To	D1			Re			
	Parameter	Estimate	Standard Error	Pr > ChiSq	Estimate	Standard Error	Pr > ChiSq
Intercept	4.0591	2.0508	0.0478	<.0001	-9.5174	377.3	0.9799
UT0_1	11.5532	2.735	0.0015	1.7123	2.6301	0.515	
UT0_2	-9.1867	2.8948	0.0015	-4.1241	2.7708	0.1366	
UT0_3	-0.1147	1.1483	0.9205	0.4895	0.99	0.621	
avg_balance_1	-0.00015	0.000187	0.428	-0.00006	0.000188	0.7357	
avg_balance_2	0.00018	0.000209	0.3895	0.000296	0.000248	0.2323	
avg_balance_3	-0.00008	0.000103	0.457	-0.0002	0.000121	0.0963	
b_AvgOB1_to_MaxOB1_I	3.587	4.9911	0.4723	0.3782	4.8674	0.9381	
b_AvgOB2_to_MaxOB2_I	0.8031	1.7702	0.6501	-0.0803	1.5653	0.9591	
b_AvgOB3_to_MaxOB3_I	1.088	0.5396	0.0438	0.1623	0.5049	0.7478	
b_TRmax_deb1_To_Limi	-0.8859	1.078	0.4112	-3.096	2.8232	0.2728	
b_TRmax_deb2_To_Limi	-0.1799	0.3709	0.6276	-1.4437	0.8076	0.0739	
b_TRmax_deb3_To_Limi	1.6974	0.8108	0.0363	0.1262	0.519	0.808	
b_TRavg_deb1_to_avgO	0.0568	0.4816	0.9062	-0.2474	0.3892	0.525	
b_TRavg_deb2_to_avgO	-0.2779	0.2249	0.2167	-0.00475	0.2084	0.9818	
b_TRavg_deb3_to_avgO	0.4343	0.2467	0.0783	-0.0742	0.1741	0.67	
b_TRsum_deb1_to_avgO	-0.5446	0.5391	0.3124	0.4324	0.4337	0.3188	
b_TRsum_deb2_to_avgO	0.2037	0.2141	0.3413	0.1847	0.1857	0.3199	
b_TRsum_deb3_to_avgO	-0.5131	0.2213	0.0204	-0.0242	0.1521	0.8739	
b_TRsum_deb1_to_TRsu	0.00234	0.2371	0.9921	0.2679	0.1733	0.1221	
b_TRsum_deb2_to_TRsu	-0.1577	0.0143	<.0001	-0.0268	0.0135	0.0472	
b_TRsum_deb3_to_TRsu	-0.0245	0.0145	0.0903	-0.0202	0.0133	0.1286	
sum_deb_num_1	0.0568	0.105	0.5882	-0.0463	0.0705	0.5116	
sum_deb_num_2	-0.0547	0.0423	0.1958	-0.0421	0.0358	0.2395	
sum_deb_num_3	0.1206	0.0494	0.0147	0.0131	0.024	0.5864	
b_OB_avg_to_eop1n	-9.3464	3.8895	0.0163	-2.3099	3.8422	0.5477	
b_pos_flag_1	0.3535	0.6278	0.5734	12.9892	377.3	0.9725	
b_atm_flag_1	0.1705	0.206	0.4077	-0.0336	0.1835	0.8545	
l_ch1_ln	5.2903	3.7529	0.1586	4.3006	3.5094	0.2204	
l_ch1_flag	-0.319	0.9088	0.7256	-0.4023	0.8651	0.6419	
mob	0.0155	0.0151	0.3034	-0.00839	0.0135	0.5354	
b_TRsum_crd1_to_OB1_	0.3391	0.2675	0.2049	0.1415	0.1921	0.4612	
b_payment_lt_5p_1	-0.6609	0.213	0.0019	1.1321	0.1611	<.0001	
b_maxminOB_limit_1_ln	-6.4558	2.3061	0.0051	1.153	2.0079	0.5658	
b_maxminOB_limit_2_ln	3.4422	2.2912	0.133	2.3032	1.9857	0.2461	
b_maxminOB_limit_3_ln	-0.0955	0.3003	0.7505	-0.037	0.1953	0.8497	
b_OBbias_1_ln	0.2216	0.1042	0.0335	-0.1505	0.0839	0.073	
b_OBbias_2_ln	0.0465	0.0663	0.4832	0.0821	0.0602	0.1727	
b_OBbias_3_ln	-0.0122	0.057	0.8301	0.0021	0.053	0.9683	
b_maxminOB_avgOB_1_ln	6.9516	2.3124	0.0026	-1.3109	2.0134	0.515	

From To	D1		Re				
	Parameter	Estimate	Standard Error	Pr > ChiSq	Estimate	Standard Error	Pr > ChiSq
b_maxminOB_avgOB_2_In	-3.2276	2.3582	0.1711	-2.486	2.0248	0.2195	
b_maxminOB_avgOB_3_In	0.1378	0.3267	0.6732	0.17	0.2161	0.4316	
AgeGRP1	-0.2357	0.1448	0.1035	-0.0571	0.1353	0.673	
AgeGRP3	0.1581	0.159	0.3199	0.0459	0.1453	0.7518	
customer_income_In	-0.0735	0.1737	0.6722	0.1076	0.1631	0.5094	
Edu_High	0.0949	0.1331	0.4758	-0.33	0.1231	0.0074	
Edu_Special	-0.1356	0.1228	0.2694	-0.1676	0.1181	0.1559	
Edu_TwoDegree	0.5403	0.4582	0.2383	-1.0129	0.4473	0.0235	
Marital_Civ	-0.4594	0.1899	0.0156	-0.0948	0.1954	0.6278	
Marital_Div	-0.1482	0.1676	0.3767	-0.1308	0.1582	0.4084	
Marital_Sin	-0.1798	0.1465	0.2197	-0.1243	0.1375	0.3661	
Marital_Wid	0.2722	0.4356	0.532	-0.0868	0.3475	0.8027	
position_Man	-0.2235	0.1697	0.1877	0.0585	0.1544	0.7047	
position_Oth	-0.3622	0.1559	0.0202	-0.1414	0.155	0.3616	
position_Tech	-0.3698	0.1243	0.0029	0.045	0.1188	0.7045	
position_Top	-0.0517	0.4082	0.8993	0.0568	0.3159	0.8574	
sec_Constr	-0.3851	0.3096	0.2135	-0.2676	0.331	0.4187	
sec_Energy	0.2035	0.2591	0.4321	-0.1128	0.2313	0.6257	
sec_Fin	0.3665	0.2839	0.1967	0.0877	0.2176	0.6869	
sec_Service	-0.0704	0.1123	0.5307	-0.1175	0.108	0.2765	
sec_Trans	0.4136	0.3272	0.2062	-0.1578	0.289	0.5852	
car_Own	0.5932	0.1676	0.0004	-0.4136	0.1469	0.0049	
car_coOwn	0.2413	0.2247	0.2828	-0.034	0.1861	0.8552	
real_Own	0.0151	0.1233	0.9026	-0.00113	0.1164	0.9922	
real_coOwn	-0.0255	0.1166	0.827	-0.1607	0.1113	0.149	
reg_ctr_Y	-0.2174	0.1921	0.2578	0.1316	0.1676	0.4325	
reg_ctr_N	-0.3406	0.1866	0.0679	0.1191	0.1639	0.4676	
child_1	-0.3265	0.1473	0.0266	-0.0667	0.1382	0.6292	
child_2	-0.0397	0.0996	0.6901	-0.1216	0.0916	0.1845	
child_3	-0.2968	0.378	0.4323	-0.774	0.3642	0.0336	
Unempl_Inyoy	4.3508	1.9704	0.0272	-4.3536	1.8464	0.0184	
LCY_EURRate_Inmom	-5.6224	2.4532	0.0219	0.1325	2.3027	0.9541	
LCY_EURRate_Inyoy	3.4791	1.7284	0.0441	0.2704	1.6106	0.8667	
CPI_Inqoq	-5.2657	4.7308	0.2657	9.1525	4.3197	0.0341	
SalaryYear_Inyoy	-11.7977	3.9415	0.0028	1.235	3.6883	0.7377	
d_State_2_NA	0	.	.	0	.	.	
d_State_2_Tr	0	.	.	0	.	.	
d_State_2_Re	1.5854	0.8189	0.0529	-2.1282	0.8653	0.0139	
d_State_2_D1	-0.3309	0.7447	0.6568	-0.1065	0.8313	0.8981	
d_State_2_D2	-0.7159	0.7042	0.3093	0.2456	0.823	0.7654	

The transition from the Delinquent 2 (DPD 31-60) state reflects transitions to two main states from the credit risk point of view: to go to the current state or to go to the default state. The conditional binary logistic regression estimates for Delinquency 2 months state (Table 6.14) predict the probability to be a Revolver, go to Delinquent 1 state, or to go to the Default state at the next month. The probability of staying in the same Delinquency 2 state is the basic event for this model. The estimated coefficients are mainly insignificant. However, this does not necessarily indicate a poor predictive power of the model (see the next section 6.3.2 for the model validation results). Low significance of the coefficients can be caused by low number of observations for this state.

One of the significant behavioural variables is the ratio of the sum of debit transactions to the sum of credit transactions in the last month (b_TRsum_deb1_to_TRsum_crd). This has a negative sign for the probability of being a revolver and a positive sign for the probability of being in default. So, if a customer spends, but does not make payments, this increases the probability of default. The indicator of payment of less than 5% of the outstanding balance in the last month (b_payment_lt_5p_1) is significant for the probability of being delinquent. This means that a customer who pays an amount, which is not enough to cover the full amount due to return to revolver or inactive state will have an increased probability of default. Month on Book and credit limit changes are insignificant variables for the transition from Delinquent 31-60 days state.

Table 6.13 Number of observations in Delinquent 2 state

Transition Stage	Stage 1 - Default vs RE&D1&D2	Stage 2 - RE vs D1&D2	Stage 3 - D1 vs D2
Inactive			
Transactor			
Revolver	361	191	
Delinquent 1		170	105
Delinquent 2			65
Default	426		

Table 6.14 Conditional binary logistic regression Estimations for Delinquency 2 month state

From To	D2		D1		Df	
	Parameter	Estimate	Pr > ChiSq	Estimate	Pr > ChiSq	Estimate
Intercept	-74.0865	0.7052	189	0.1966	49.0665	0.0003
UTO_1	-13.9109	0.3917	47.0635	0.3005	7.5467	0.3087
UTO_2	-3.1757	0.855	-33.5843	0.4563	4.5272	0.5543
UTO_3	7.6665	0.3435	5.6426	0.7528	-3.8545	0.2691
avg_balance_1	-0.00318	0.1609	0.0244	0.0005	-7.82E-06	0.9949
avg_balance_2	0.00322	0.1879	-0.0242	0.0008	-0.00019	0.89
avg_balance_3	0.000116	0.8868	-0.00039	0.8337	0.000267	0.4647
b_AvgOB1_to_MaxOB1_In	-5.2311	0.9087	-40.2245	0.7837	35.2376	0.1079
b_AvgOB2_to_MaxOB2_In	7.9236	0.6826	97.866	0.2336	-1.0127	0.9268
b_AvgOB3_to_MaxOB3_In	-10.799	0.3491	26.8607	0.2491	-4.2441	0.2486
b_TRmax_deb1_To_Limit	41.0323	0.8589	-960.9	0.0435	-219.1	<.0001
b_TRmax_deb2_To_Limit	36.816	0.2088	-86.5632	0.0678	-11.6003	0.2892
b_TRmax_deb3_To_Limit	-3.5792	0.3044	-5.6933	0.6725	-1.7291	0.1379
b_TRavg_deb1_to_avgO	34.3413	0.9076	-5.9922	0.5772	2.3698	0.374
b_TRavg_deb2_to_avgO	-0.8985	0.7338	-0.4905	0.9167	1.1079	0.3784
b_TRavg_deb3_to_avgO	-1.1585	0.2045	-1.0529	0.6105	0.0174	0.9689
b_TRsum_deb1_to_avgO	-33.2334	0.9106	19.237	0.2186	2.8348	0.4039
b_TRsum_deb2_to_avgO	0.2555	0.9175	0.9682	0.8284	-0.5735	0.6464
b_TRsum_deb3_to_avgO	1.1778	0.1933	0.2594	0.8897	0.844	0.0579
b_TRsum_deb1_to_TRsum	-11.8969	0.0009	20.2828	0.0524	4.8685	0.0106
b_TRsum_deb2_to_TRsum	0.0397	0.6549	-0.3613	0.0479	-0.143	0.0076
b_TRsum_deb3_to_TRsum	-0.0223	0.6251	-0.1337	0.1463	-0.0821	0.001
sum_deb_num_1	13.8165	0.9083	-4.6895	0.2514	0.6564	0.4991
sum_deb_num_2	0.1635	0.797	0.0537	0.958	0.0666	0.8197
sum_deb_num_3	-0.079	0.5711	-0.00078	0.9981	0.00305	0.9676
b_OB_avg_to_eop1ln	-4.7008	0.8504	-138.1	0.0646	-24.0447	0.0356

From To		D2		D1		Df	
	Parameter	Estimate	Pr > ChiSq	Estimate	Pr > ChiSq	Estimate	Pr > ChiSq
	b_pos_flag_1	30.2501	0.861	0	.	1.404	0.3878
	b_atm_flag_1	9.469	0.9321	-16.4948	0.8518	-2.007	0.2507
	l_ch1_ln	54.3859	0.3835	-191.7	0.941	2.8193	0.8149
	l_ch1_flag	-10.2068	0.411	39.475	0.9413	-0.8952	0.7355
	Mob	0.029	0.6478	0.0302	0.816	0.0347	0.2556
	b_TRsum_crd1_to_OB1	-13.6491	0.0008	23.6023	0.0496	5.8659	0.0071
	b_payment_lt_5p_1	0.7613	0.464	4.3169	0.0794	-0.5138	0.4403
	b_maxminOB_limit_1_ln	2.6488	0.8889	-85.3489	0.1782	10.9885	0.2205
	b_maxminOB_limit_2_ln	11.1022	0.5665	169.6	0.0225	-15.6847	0.1002
	b_maxminOB_limit_3_ln	-8.0752	0.3242	-57.1553	0.0213	4.1069	0.251
	b_OBBias_1_ln	-0.8178	0.1272	-0.565	0.6875	0.5557	0.0593
	b_OBBias_2_ln	0.1133	0.7335	1.8575	0.0466	0.1713	0.4023
	b_OBBias_3_ln	-0.5139	0.0808	1.5694	0.0104	0.2327	0.089
	b_maxminOB_avgOB_1_ln	-1.2565	0.9478	91.0386	0.1483	-8.8678	0.3275
	b_maxminOB_avgOB_2_ln	-10.8707	0.5758	-161.9	0.0313	15.6403	0.104
	b_maxminOB_avgOB_3_ln	7.322	0.397	57.3742	0.0215	-4.4975	0.2297
	AgeGRP1	0.3079	0.5927	0.1851	0.8666	-0.167	0.544
	AgeGRP3	1.0068	0.1001	-1.2481	0.3518	-0.5403	0.0862
customer_income_ln		-0.3625	0.6483	-2.6914	0.078	-0.6535	0.0782
Edu_High		0.6192	0.2312	-1.5932	0.1646	0.2515	0.3476
Edu_Special		0.921	0.0372	-0.2468	0.7838	0.1664	0.4801
Edu_TwoDegree		-10.5034	0.9347	0	.	0.1292	0.8999
Marital_Civ		-0.2248	0.7152	2.9577	0.0315	-0.0449	0.8939
Marital_Div		0.0251	0.9655	-0.2851	0.794	0.2991	0.344
Marital_Sin		1.0036	0.0984	-0.00915	0.9933	-0.4206	0.1512
Marital_Wid		5.9888	0.0243	12.4816	0.9427	-1.0556	0.3608
position_Man		-0.779	0.22	-0.4608	0.7395	-0.1254	0.7035
position_Oth		-0.6948	0.2652	-1.7335	0.2013	-0.3225	0.3048
position_Tech		-0.0846	0.8566	-1.2132	0.2369	-0.1821	0.4488
position_Top		1.1215	0.5787	-9.123	0.8995	0.6766	0.4849
sec_Constr		3.194	0.0207	3.8485	0.0129	0.2656	0.6226
sec_Energy		0.7641	0.4558	-1.7131	0.3907	-0.4438	0.4248
sec_Fin		-4.7471	0.1675	-51.4286	0.7658	-0.4034	0.5394
sec_Service		0.2303	0.5722	-0.601	0.4923	0.1316	0.5465
sec_Trans		-0.5559	0.7309	9.9887	0.0425	-0.1715	0.8083
car_Own		0.4102	0.6109	0.9631	0.5225	-0.2808	0.4323
car_coOwn		1.0108	0.181	2.0375	0.1354	-0.0433	0.9174
real_Own		0.6973	0.1687	1.0266	0.3394	-0.0118	0.9613
real_coOwn		0.5831	0.2121	0.6715	0.4907	0.0706	0.755
reg_ctr_Y		-2.0297	0.0144	0.3396	0.8219	-0.3325	0.3833
reg_ctr_N		-1.3508	0.0898	-0.0957	0.9461	-0.198	0.5952
child_1		0.9028	0.0938	-1.7941	0.1399	-0.5411	0.0651
child_2		0.5773	0.1546	0.6789	0.3473	-0.2076	0.2802
child_3		-2.7276	0.0781	4.0427	0.2478	0.2947	0.6719
Unempl_Inoy		14.2819	0.0977	-2.2198	0.8882	-6.4254	0.1054
LCY_EURRate_Inmom		-4.19	0.6228	6.1444	0.724	-0.0268	0.9952
LCY_EURRate_Infoy		5.807	0.3997	38.8168	0.0178	1.2429	0.7254
CPI_Inqoq		16.3816	0.341	-4.255	0.9111	-14.3136	0.1579
SalaryYear_Infoy		6.7358	0.6676	-108.7	0.0122	-17.253	0.0416
d_State_2_NA		0	.	0	.	0	.
d_State_2_Tr		0	.	0	.	0	.
d_State_2_Re		0	.	0	.	0	.
d_State_2_D1		-0.2367	0.8889	-0.8748	0.8371	1.6898	0.1123
d_State_2_D2		-0.2747	0.8382	-0.0121	0.9975	-0.242	0.7769

As for multinomial and ordinal regression, the variables that described the history of states are insignificant for the conditional binary logistic regressions. Also, many estimated coefficients for behavioural characteristics are insignificant as in the case of multinomial regression. This may happen because of high correlations between: i)

different behavioural characteristics, ii) the same behavioural characteristic in a different time. Despite insignificant coefficients, the validation results of model predictive power are high, especially, for Delinquent states.

6.3.2 Validation Results for binary logistic regression models

For each transition model, we calculate the Gini and Kolmogorov-Smirnov coefficients (Table 6.15). Models for transition probabilities from the Non-active state show that Gini and KS values are not high – 0.3-0.35 and 0.2-0.26 for staying an inactive and going to the revolver state respectively. This is expected for application scoring models. A lot of non-active accounts do not have a behavioural history before activation. In the non-active segment is split into those activated before and those completely non-active before the scoring point, the predictive power of these two models might be lower for a model, which does not use behavioural characteristics, and higher for a model, which uses application plus behavioural characteristics.

Table 6.15 Gini and KS values for binary logistic regression (test sample)

Current State	Predicted State	KS	GINI
Non-active	Non-active	0.26	0.35
	Revolver	0.22	0.3
Transactor	Non-active	0.35	0.47
	Revolver	0.29	0.69
Revolver	Non-active	0.68	0.78
	Revolver	0.64	0.75
	Delinquent 1	0.4	0.52
Delinquent 1	Delinquent 2	0.5	0.64
	Revolver	0.44	0.57
Delinquent 2	Default	0.5	0.63
	Revolver	0.66	0.8
	Delinquent 1	0.7	0.82

The best prediction results were shown by models predicting transitions from Revolver to Non-active (Gini = 0.78), Revolver to Revolver (Gini = 0.75), Delinquency 1 to Delinquency 2 (Gini = 0.64, and Delinquency 2 to Default (Gini = 0.63). These high prediction accuracy results have been obtained despite a lot insignificant estimated coefficient, for example, for delinquent state models. The prediction accuracy for transactors and non-active states is also low as given by the multinomial and ordinal regression models. This can be explained by less behavioural history and that the

models for inactive and transactor customers are similar to application scoring models. Generally, binary logistic regression has shown better predictive power for the majority of transitions.

6.4 New Definition of States and an Introduction of Additional States

We include additional state definition in Chapter 6, related to the modelling results, because till this section the results do not include new ‘Revolver Paid’ state. The necessity of the inclusion of new state has been found in the process of the estimation of transition probabilities because we found cases which can be defined both as transactor and revolver state according to the initial state definition.

The model described in Chapter 5 consists of five states. However, there are several issues, which can impact on prediction accuracy. An account, which has a positive balance at the beginning of the month and zero balance at the end of the month is a repaid revolver, or a customer who repaid the full amount in the current month. However, according to the definition from the Chapter 5, if an account has debit (or purchase) transactions during the month and the outstanding balance at the end of the month is equal to zero, this account has a transactor state. So to be a transactor credit card holder must have zero outstanding balance at the beginning of the month, some debit (or purchase) transactions during the month, and repay the full amount at the end of the month. For accounts, which have positive, outstanding balance at the beginning of the month and repaid full debt amount at the end of the month (and it does not matter whether these accounts have debit transactions inside of the month or not) we add new state called ‘Revolvers Paid in Full’ (RP). A revolver can go from revolver state to inactive or transactor only via ‘Paid in Full Revolver’ state or go to delinquency. Thus we have a ‘pure’ transactor state and a ‘pure’ revolver state. Moreover, accounts, which generate interest income during the last active period, but become inactive (zero balance) at the end of the month, are related to the separate state – Revolver Paid (RP).

Credit card states are traditionally defined by the level of delinquency and score band, or the level of risk (for example, So and Thomas, 2011; Leow and Crook, 2014). Kallberg and Saunders (1983) split the current state into sub-states by the opening balance and used ‘Paid-up’ state when the account has no outstanding balance. In addition to the credit card segments, traditionally used in the papers, such as

transactors and revolver (see Bertaut et al., 2008; So and Thomas, 2010; Tan and Yen, 2011; So et al, 2014), we proposed to use transactor and revolver segments as income related states and include new state - Revolver Repaid as an intermediary state from revolver and delinquent states. Revolver repaid state is used for the identification of the transition of revolver account to an inactive state because the customer who fully repays the debt amount at the end of month formally cannot be allocated either to transactor, or inactive, or revolver state. Such revolver paid account generates income in the repayment months, but according to the definitions should be related to inactive or transactor accounts because of the zero outstanding balance at the end of the month.

The main characteristic of a revolver state is a non-zero outstanding balance at the end of the period. The ‘full’ state definition is considering in the current research a partitioning of the revolver state into four categories: revolver activated, revolver existing, cash-user (or convenience user), and revolver deactivated. ‘Activated revolver’ state means a transition to the positive, outstanding balance from non-active or transactor states, where the outstanding balance at the end of the previous period (or beginning of the current period) is equal to zero. ‘Revolver existing’ (RE) state means positive balance at the beginning and the end of the current period without delinquent payments. The separation of credit card holders who do not have purchases (or debit transactions) monthly and who factually use a credit card as a standard cash loan from other behavioural types of credit card holders can increase the predictive power of a model because of different usage drivers. They generate the interest rate income only, but not interchange and ATM fees, and can be interesting for marketing segmentation. Moreover, the final revolvers subcategory is revolvers who paid back the full amount in the observed period and have a zero balance at the end of month (RP), so they are becoming non-active at the end of the period. At the next month, they can stay inactive, go to the transactor state by making purchases, but paying back the full amount again, or return to the revolver state in case of new purchase transaction. In this research, for an empirical investigation, we use only two states RE and RP states, because two other states – revolver activated and cash-user – have few cases.

The definitions of full set of states with use of four dimensions: Average outstanding balance for month at the end of month, Balance at the end of month, Sum of Debit

transactions at the end of month, and Days Past Due at the end of month is presented in Table 6.16.

Table 6.16 Definitions for new full set of states (as of the end of month)

State	Average outstanding balance for the month as of the end of the month	Outstanding Balance at the end of period	Sum of Debit transactions for the month as of the end of the month	Days Past Due at the end of the month
NA	= 0	= 0	0	= 0
TR	> 0	= 0	> 0	= 0
RE	> 0	> 0	> 0	= 0
RP	> 0	= 0		= 0
D1	> 0	> 0		<=30
D2	> 0	> 0		<=60
Df	> 0	> 0		>60

Each of definition means that all five dimensions are computed and available at the end of period. So we may say that we use the state definition (such as Kallberg and Saunders, 1983) instead of segment definition (such as Frydman et al., 1985; So et al., 2014).

A problem for the one of the mentioned states system can be caused by the number of observations for each category and subcategory. We need to decide whether: 1) to exclude those segments from the logistic regression and use pool level estimation for the segment with a low number of observations, and 2) to merge segments with a few observations with other segment, which is close by logic, for example, merge delinquency 1 month with delinquency 2 months for prediction of transition from non-active state for a 3 months outcome period.

We have contributed to the literature by defining states more precisely. We adopt the following new definitions:

- i) Transactor is not only a borrower who pays back the full amount as is mentioned in some sources such as So and Thomas (2010), So et al. (2014); Cheu and Loke (2010), Tan and Yen (2010) – for so-called ‘Convenience User’; Hsieh (2004), but also has a zero balance at the beginning of the period (month);
- ii) If revolver (user with positive outstanding balance at the beginning of the period) has paid back full debt amount at the current month, (s)he becomes a transactor by

many definitions, but actually (s)he generated interest rate income and had a positive outstanding balance at the beginning of the period. Thus (s)he should be allocated to a separate state such as Revolver Paid (RP).

Because the conditional binary logistic regression has shown better results in comparison with multinomial and ordinal logistic regression, we will use this method for the full model.

The outcome prediction horizon is a question under discussion. The number of state for the $t+1$ is quite limited. However, for $t+2$ and $t+3$ the number of possible transitions is higher, and for $t+4$ and more steps it is possible to move from any state to any state. In Table 6.17 we marked with ‘plus’ possible transition from the state i_t at time t to the state i_{t+N} for N steps.

According to this matrix a Revolver can go to the revolver, revolver paid, and delinquent states, but cannot go to the inactive and transactor states directly. A customer in the ‘Revolver paid’ state can become an inactive, a transactor, or a revolver, but can not return to this state directly in the next month. Delinquent customers in case of paying back the full amount go to ‘Revolver Paid’ state in the next month, which reflects the elimination from the positive outstanding balances states, but not directly to the inactive state. On the other hand, if a customer in the delinquency state pays back full amount due in arrears, but keeps non-delinquent positive outstanding balance, he/she will move to the revolver state.

Table 6.17 Possible transitions from current state to states for N steps (full set of states)

State t	T+N	NA	TR	RE	RP	D1	D2	Df
NA	1	+	+	+				
	2	+	+	+	+	+		
	3	+	+	+	+	+	+	
	4	+	+	+	+	+	+	+
	5	+	+	+	+	+	+	+
	6	+	+	+	+	+	+	+
TR	1	+	+	+	+			
	2	+	+	+	+	+		
	3	+	+	+	+	+	+	
	4	+	+	+	+	+	+	+
	5	+	+	+	+	+	+	+
	6	+	+	+	+	+	+	+
RE	1			+	+	+		
	2	+	+	+	+	+	+	
	3	+	+	+	+	+	+	+
	4	+	+	+	+	+	+	+
	5	+	+	+	+	+	+	+
	6	+	+	+	+	+	+	+
RP	1	+	+	+				
	2	+	+	+	+	+		
	3	+	+	+	+	+	+	
	4	+	+	+	+	+	+	+
	5	+	+	+	+	+	+	+
	6	+	+	+	+	+	+	+
D1	1			+	+	+	+	
	2	+	+	+	+	+	+	+
	3	+	+	+	+	+	+	+
	4	+	+	+	+	+	+	+
	5	+	+	+	+	+	+	+
	6	+	+	+	+	+	+	+
D2	1			+	+	+	+	+
	2	+	+	+	+	+	+	+
	3	+	+	+	+	+	+	+
	4	+	+	+	+	+	+	+
	5	+	+	+	+	+	+	+
	6	+	+	+	+	+	+	+

We built empirical transition matrices for the full state for one and six months performance period (see Table 6.18 and Table 6.19). Matrices include all transitions for a one year period (from June 2010 till July 2011) and one account can be presented

in the matrix 12 times, but demonstrate various transitions. These matrices reflect empirically the transition options from Table 6.17. We add the ‘Revolver Paid’ state to the transition matrices, reseated in Chapter 5.

These matrices are not a triangle, so they have not only zero values in the cells above and to the right of the diagonal cells. It is usually expected from the transition matrices, which are built for delinquency buckets such as non-delinquent, 1-30 days past due, 31-60 days-past due etc. The cells with non-zero values reflect the possible transitions, which in Table 6.18 correspond with transitions in the first months from Table 6.17. For example, from the transactor state, an account can move in the next month to inactive (40.12% of transactors) or revolver state (35.24% of transactors), or it can stay in the transactor state (24.63% of transactors). Revolvers (81.30% of the total transition cases) mainly stays in the revolver state (94.44%), 3.64% of revolvers usually pay back the full amount in the next month, and 1.92% become delinquent customers. Almost a half of the cases from the revolver paid state move in to the inactive state, but 15% become transactors and 35.77% have a positive outstanding balance again in the next month. We can see that 55.35% of customers in the delinquency 2 state become defaulters and move in to the absorbing state.

Table 6.18 Full states empirical transition matrix for t+1

t+1									
From\To	NA	TR	RE	RP	D1	D2	Df	Total	
NA	84.55%	3.96%	11.49%	0.00%	0.00%	0.00%	0.00%	12.15%	
TR	40.12%	24.63%	35.24%	0.00%	0.00%	0.00%	0.00%	1.24%	
RE	0.00%	0.00%	94.44%	3.64%	1.92%	0.00%	0.00%	81.30%	
RP	49.35%	14.89%	35.77%	0.00%	0.00%	0.00%	0.00%	2.98%	
D1	0.00%	0.00%	57.81%	0.66%	23.40%	18.13%	0.00%	1.96%	
D2	0.00%	0.00%	23.18%	0.47%	13.15%	7.85%	55.35%	0.37%	
Total	12.24%	1.23%	80.90%	2.97%	2.07%	0.38%	0.21%	100.00%	

The matrix for the transitions for 6 month prediction period (Table 6.19) are observed. Only transitions from Delinquent 2 state to Transactor state have not been observed. Generally, the percentage transition decreased for the majority of transitions because of a decrease in concentrations of transitions which are possible for $t+1$ period. For example, the share of inactive customers who stay inactive in an extended period significantly decreased: from 84.55% for $t+1$ to 62.75% for $t+6$, and the transition to

the revolver state has increased almost three times – from 11.5% to 29.05%. The chance of staying in the revolver state is still high – 87.56%. Generally, about a half of Revolvers who paid the full amount become inactive in 1 month period, but only one-third of all accounts in the RP state move to become inactive in a 6 month period. Moreover, 8% of customers in Revolver Paid state become a ‘Revolver Paid’ in 6-month period, so they use a credit limit, pay back the full amount and to have a zero outstanding balance again. The transition from the transactor to the transactor state confirms that it is non-stable state – only 10% of transactors stay in the same state in 6 months, and a big part of them – 43.76 per cent become Revolvers. However, only 1.25% of the sample are in the one year period, so this state is not very spread in the given data sample.

Table 6.19 Full states empirical transition matrix for t+6

t+6								
From\To	NA	TR	RE	RP	D1	D2	Df	Total
NA	62.75%	2.98%	29.05%	4.98%	0.18%	0.03%	0.03%	12.15%
TR	34.90%	10.61%	43.76%	10.30%	0.31%	0.03%	0.09%	1.24%
RE	5.47%	0.69%	87.56%	2.85%	2.27%	0.44%	0.73%	81.30%
RP	34.78%	4.44%	52.35%	8.01%	0.31%	0.05%	0.07%	2.98%
D1	1.43%	0.16%	56.63%	1.09%	20.12%	5.03%	15.53%	1.96%
D2	1.14%	0.00%	24.41%	0.57%	13.25%	6.81%	53.83%	0.37%
Total	13.57%	1.19%	78.02%	3.31%	2.32%	0.49%	1.10%	100.00%

Inactive and revolver states are stable and have approximately similar rates of staying. However, the new matrix is not triangle. It does not reflect the possible transition from the state to the next state, because transition from the Revolver Paid state to the Delinquent 1 state is not possible. It is possible only moving from RP state, but not possible to stay there for the next step. So, RP is only a transition state. However, we believe that the introduction of this state makes the system of states more consistent and logical.

6.5 Multinomial regression coefficients estimations results

6.5.1 Model t+1 estimations

We apply multinomial logistic regression for transition probabilities estimation. In this section we build 6 models for 6 periods: the probability to move from state S_t to state S_{t+N} , where $N = 1,2,3,4,5,6$. In Table 6.20 to Table 6.29 we present the estimated coefficients for the multinomial regression. Each table is related to a state from which an account moves, and each column is related to the state to which an account moves. At this stage, we use aggregated variables for several months (see Chapter 3 for description) instead of monthly covariates as was done for models' selection at the first stage in Section 6.2.1.

The most significant characteristics for a transition from inactive state are: Indicator of only ATM cash withdrawal (`b_atm_use_only_flag`) average number of debit transactions (purchases) for the last 3 months (`b_avgNumDeb13`), logarithm of bias of the outstanding balance from the average for the last 6 months (`b_OBbias_16_ln`), ratio of the average purchase transaction amount to the average outstanding balance for the last 6 months (`b_TRavg_deb16_to_avgOB16_ln`), and state characteristic as number of times in state Revolver, Revolver Paid, or transactor (Table 6.21).

Despite few covariates for the model of the transitions from the inactive state to the revolver state being significant (see Table 6.21) the predictive power of the model is satisfactory as was the application model, because a lot of inactive accounts do not have a behavioural history before the scoring date. The validation sample Gini is low for such models: for an inactive account to stay inactive it is - 0.29, for an inactive to move to transactor state the Gini is 0.37, and only 0.24 for the transition to revolver state (0.29 for development sample). Thus, it is possible to predict whether the customer will activate an inactive account next month, but difficult to say for the inactive client if (s)he will pay back the full amount the month after activation or stay active for a longer period. However, the current model includes all accounts: those with previous history and those without previous history.

Table 6.20 Target frequencies for t+1, non-active state

Ordered value	Target_SI_dev	Total freq
1	NA	20424
2	RE	2766
3	TR	974

Table 6.21 Multinomial regression estimations for t+1, from non-active state

Char	NA		RF	
	Estimate	Pr > ChiSq	Estimate	Pr > ChiSq
<i>Intercept</i>	2.0673	0.0092	0.5508	0.5424
<i>b_atm_flag_use13vs46</i>	0.0619	0.7272	-0.1263	0.5305
<i>b_atm_use_only_flag_</i>	-0.2404	0.0779	0.0149	0.9214
<i>b_avgNumDeb13</i>	-0.1498	0.0018	-0.00687	0.8925
<i>b_AvgOB1_to_MaxOB1_I</i>	-0.1544	0.9349	-1.2836	0.5368
<i>b_AvgOB16_to_MaxOB16</i>	-0.3631	0.8038	0.69	0.6649
<i>b_max_dpd16</i>	0.3021	0.739	0.2908	0.7485
<i>b_maxminOB_avgOB_1_I</i>	0.0479	0.9403	-0.4526	0.5176
<i>b_maxminOB_avgOB_16_</i>	-0.1872	0.8889	0.1182	0.9355
<i>b_maxminOB_limit_1_I</i>	-0.0157	0.8491	0.0202	0.8272
<i>b_maxminOB_limit_16_</i>	0.0738	0.02	-0.0231	0.5262
<i>b_OB_avg_to_eop1ln</i>	0.1007	0.2227	0.0598	0.5161
<i>b_OB1_to_OB2_ln</i>	0.0144	0.5125	-0.00638	0.8019
<i>b_OBBias_1_ln</i>	-0.0342	0.9686	-0.5672	0.5503
<i>b_OBBias_16_ln</i>	-0.2729	0.2759	0.4914	0.107
<i>b_payment_lt_5p_1</i>	0.4195	0.7627	1.3789	0.3591
<i>b_payment_lt_5p_2</i>	0.0163	0.9795	-0.5873	0.4336
<i>b_payment_lt_5p_3</i>	-0.1426	0.3064	-0.0353	0.8216
<i>b_pos_flag_use13vs46</i>	-0.2174	0.2112	-0.1235	0.5332
<i>b_pos_use_only_flag_</i>	-0.1708	0.234	-0.0571	0.7235
<i>b_TRavg_deb1_to_26_I</i>	-0.0675	0.0028	-0.0143	0.5817
<i>b_TRavg_deb16_to_avgOB16_ln</i>	-0.0856	0.0006	-0.0785	0.0067
<i>b_TRsum_deb1_to_OB1</i>	0.00226	0.4466	0.0055	0.1078
<i>b_TRmax_deb16_To_Lim</i>	-0.0395	0.1711	0.0092	0.7796
<i>b_TRsum_crd1_to_2_ln</i>	-0.00284	0.9591	0.0788	0.2272
<i>b_TRsum_crd13_to_46_</i>	-0.00801	0.6334	0.0118	0.54
<i>b_TRsum_crd13_to_OB1</i>	-0.0345	0.1574	-0.0316	0.2564
<i>b_TRsum_deb1_to_2_ln</i>	0.00679	0.636	0.0131	0.4143
<i>b_TRsum_deb1_to_TRsu</i>	0.0563	0.676	-0.0231	0.8714
<i>b_TRsum_deb16_to_TRs</i>	0.00223	0.9404	0.00493	0.8884
<i>b_UT13to46ln</i>	-0.00531	0.685	-0.00727	0.6296
<i>b_UT1to2ln</i>	-0.00245	0.8654	-0.021	0.1958
<i>avg_balance_6</i>	0.00222	0.2196	0.000256	0.9014
<i>age</i>	0.0119	0.0127	0.00703	0.1894
<i>car_coOwn</i>	-0.0465	0.7357	-0.0361	0.8144
<i>car_Own</i>	-0.2913	0.0003	-0.2977	0.0011
<i>child_1</i>	-0.0537	0.6384	0.0652	0.6113
<i>child_2</i>	-0.0913	0.1572	-0.0717	0.323
<i>child_3</i>	-0.3267	0.1559	-0.1812	0.4834
<i>customer_income_ln</i>	-0.066	0.5393	-0.067	0.5819
<i>Edu_High</i>	-0.06	0.6268	-0.06	0.6632
<i>Edu_Special</i>	-0.0437	0.7247	-0.0641	0.6428
<i>Edu_TwoDegree</i>	-0.3188	0.1352	-0.3111	0.2032
<i>I_ch1_flag</i>	0.9421	0.1499	0.3153	0.6559
<i>I_ch1_ln</i>	-1.7803	0.0484	-0.3354	0.7286
<i>I_ch6_flag</i>	-0.0901	0.4663	0.009	0.948
<i>limit_6</i>	0.000008185	0.5118	0.000001912	0.8928
<i>Marital_Civ</i>	-0.0639	0.7362	-0.00102	0.9961
<i>Marital_Div</i>	-0.0264	0.831	-0.0202	0.8834
<i>Marital_Sin</i>	-0.0565	0.6425	-0.1582	0.2513
<i>Marital_Wid</i>	-0.1596	0.4789	-0.2683	0.2973
<i>mob</i>	0.0173	0.0768	-0.0143	0.1961
<i>position_Man</i>	0.0752	0.4803	0.1146	0.3365
<i>position_Oth</i>	0.042	0.7371	0.1891	0.171
<i>position_Tech</i>	-0.1647	0.1487	-0.0605	0.636
<i>position_Top</i>	-0.2127	0.1908	-0.188	0.3138
<i>real_coOwn</i>	0.0113	0.9063	-0.0942	0.3866
<i>real_Own</i>	0.0444	0.6245	0.0109	0.9146
<i>reg_ctr_N</i>	0.1426	0.278	0.0428	0.7742
<i>reg_ctr_Y</i>	0.1454	0.261	0.0684	0.6414
<i>s_been_D1_full</i>	-3.4637	0.714	-3.023	0.749
<i>s_been_Tr_full</i>	0.1246	0.4393	0.1185	0.5238
<i>s_cons_full</i>	-0.00445	0.9605	0.0064	0.9507
<i>s_month_since_NA_ful</i>	0	0	0	0

Char	NA		RE	
	Estimate	Pr > ChiSq	Estimate	Pr > ChiSq
s_month_since_RP_full	-0.0642	0.2366	0.0056	0.928
s_times_RE_full	-0.1506	0.073	0.1429	0.1325
s_times_RP_full	-0.7571	0	-0.2243	0.1225
s_times_TR_full	-0.5584	0	-0.2664	0.0361
sec_Agricult	0.2462	0.2664	0.2521	0.2996
sec_Constr	-0.152	0.5577	-0.0447	0.878
sec_Energy	0.2861	0.1116	0.111	0.581
sec_Fin	-0.1628	0.1169	-0.1862	0.1154
sec_Industry	-0.3622	0.3697	0.1284	0.7698
sec_Manufact	0.2759	0.4108	0.2877	0.4329
sec_Mining	0.1275	0.5465	0.0485	0.8374
sec_Service	0.0304	0.7549	0.0362	0.7395
sec_Trade	-0.2928	0.0272	-0.2435	0.1067
sec_Trans	-0.3602	0.1092	-0.4123	0.1194
UT0_6	-26.5583	0.1396	-7.7723	0.6886
UT0_7	-1.2778	0.0438	-1.258	0.0741
UT0_8	0.2219	0.5375	0.6641	0.0925
m_Unempl_Inyoy_6	1.3048	0.3542	0.2669	0.8666
m_UAH_EURRate_Inmom_6	0.2847	0.8629	-3.4661	0.0664
m_CPI_Inqoq_6	4.0051	0.1603	-3.4292	0.282
m_SalaryYear_Inyoy_6	-0.7736	0.5648	-1.9483	0.1941

As Table 6.23 shows the most significant characteristics for a transition from the transactor state to the revolver state are: sum of debit transactions in the last month to the sum of debit transactions in months 2-6, the sum of payments in months 1-3 to payments in months 4-6. However, for the transition from transactor to non-active, we have other significant characteristics such as the average number of debit transactions (purchases) for the last 3 months (b_avgNumDeb13) and changes in the utilisation rate in the last 3 months to previous 4-6 months. All other predictors show low significance. However, the model predictive power is satisfactory: Gini index is 0.4 and 0.31 for the transition to inactive and revolver states.

Table 6.22 Target frequencies for t+1, from transactor state

Ordered value	Target_SI_dev	Total freq
1	NA	992
2	RE	851
3	TR	575

Table 6.23 Multinomial regression estimations for t+1, from transactor state

Char	NA		RE	
	Estimate	Pr > ChiSq	Estimate	Pr > ChiSq
Intercept	1.1991	0.2304	1.632	0.1103
b_atm_flag_use13vs46	-0.00442	0.9802	0.1865	0.3044
b_atm_use_only_flag_	0.225	0.1829	0.1724	0.3045
b_avgNumDeb13	-0.1825	0	0.00759	0.6061
b_AvgOB1_to_MaxOB1_I	-0.0718	0.5802	-0.021	0.8614
b_AvgOB16_to_MaxOB16	-2.8782	0.1247	-1.1941	0.4533
b_max_dpd16	2.1496	0.1502	1.5619	0.2142
b_maxminOB_avgOB_1_I	-0.1336	0.6725	0.1245	0.664
b_maxminOB_avgOB_16_	-2.6879	0.0823	-1.1939	0.3423
b_maxminOB_limit_1_I	-0.04	0.3664	-0.036	0.4271
b_maxminOB_limit_16_	0.0907	0.2243	-0.0576	0.4461
b_OB_avg_to_eop1n	-0.0666	0.3263	-0.0825	0.2368
b_OB1_to_OB2_ln	0.028	0.3618	-0.0143	0.651

Char	NA		RE	
	Estimate	Pr > ChiSq	Estimate	Pr > ChiSq
b_OBbias_1_ln	0.163	0.4286	-0.1235	0.5125
b_OBbias_16_ln	-0.3145	0.4091	-0.1736	0.6447
b_payment_lt_5p_1	-0.8886	0.2431	-0.4659	0.582
b_payment_lt_5p_2	-0.0316	0.9447	-0.1593	0.7327
b_payment_lt_5p_3	0.1173	0.5121	-0.0715	0.6974
b_pos_flag_use13vs46	0.4098	0.0197	0.1009	0.5862
b_pos_use_only_flag_	0.1035	0.5312	-0.0673	0.6763
b_TRavg_deb1_to_26_l	-0.0653	0.0398	0.0516	0.1273
b_TRavg_deb16_to_avg	0.000778	0.9891	-0.0392	0.5072
b_TRmax_d*b_TRsum_de	-0.0034	0.6141	-0.0169	0.0219
b_TRmax_deb16_To_Lim	-0.1738	0.0197	0.0529	0.4753
b_TRsum_crd1_to_2_ln	0.0358	0.4794	0.0314	0.5456
b_TRsum_crd13_to_46_	-0.00638	0.8168	-0.0574	0.043
b_TRsum_crd13_to_OB1	0.0353	0.4354	0.0546	0.2153
b_TRsum_deb1_to_2_ln	-0.0379	0.1678	-0.0213	0.4385
b_TRsum_deb1_to_TRsu	0.0984	0.2143	0.1812	0.044
b_TRsum_deb16_to_TRs	-0.1138	0.1175	0.1208	0.0819
b_UT13to46ln	0.0591	0.0031	0.027	0.2024
b_UT1to2ln	-0.0243	0.1873	-0.00297	0.8733
age	0.0206	0.0107	-0.00904	0.2722
avg_balance_6	-0.00023	0.2147	-0.00031	0.045
car_coOwn	0.2667	0.2755	0.2517	0.3029
car_Own	0.0253	0.8586	-0.0542	0.7093
child_1	-0.3312	0.111	-0.2192	0.294
child_2	-0.1308	0.2639	-0.0379	0.7488
child_3	-0.4931	0.1919	-0.1425	0.7074
customer_income_ln	0.1161	0.5034	-0.0133	0.9389
Edu_High	-0.1254	0.5665	-0.3658	0.0887
Edu_Special	0.2251	0.3149	-0.1681	0.4502
Edu_TwoDegree	-0.4105	0.2703	-0.4753	0.1933
I_ch1_flag	-0.5537	0.56	-0.0908	0.9146
I_ch1_ln	0.7134	0.6613	1.3822	0.3465
I_ch6_flag	-0.3219	0.149	-0.1738	0.4377
limit_6	0.000015	0.4832	0.000006981	0.7459
m_CPI_Inqoq_6	0.9247	0.8464	5.4587	0.2641
m_SalaryYear_Inyoy_6	4.262	0.0631	4.1572	0.0701
m_UAH_EURRate_Inmom_6	0.000528	0.9999	1.7002	0.5521
m_Unempl_Inyoy_6	3.1601	0.1942	1.0169	0.6781
Marital_Civ	0.3862	0.2691	0.5333	0.108
Marital_Div	0.3052	0.1847	0.3166	0.1808
Marital_Sin	-0.1283	0.5502	-0.3953	0.0689
Marital_Wid	-0.3074	0.5267	0.3361	0.4789
mob	-0.0148	0.3766	-0.0242	0.1485
position_Man	-0.0596	0.7373	0.1515	0.3898
position_Oth	-0.2552	0.1926	-0.0517	0.7913
position_Tech	0.0826	0.6838	-0.0704	0.7367
position_Top	0.0822	0.7666	-0.0156	0.9561
real_coOwn	-0.1956	0.2349	-0.1934	0.247
real_Own	-0.3318	0.0347	-0.1417	0.3732
reg_ctr_N	0.1169	0.5967	-0.2326	0.2837
reg_ctr_Y	0.2258	0.2921	-0.2045	0.3278
s_been_D1_full	-22.2976	0.1516	-15.4286	0.2313
s_been_Tr_full	0	0	0	0
s_cons_full	0.5152	0.0002	0.3199	0.0135
s_month_since_NA_ful	-0.1148	0.0465	-0.1853	0.0012
s_month_since_RP_ful	-0.0249	0.7389	0.0265	0.7288
s_times_RE_full	0.0992	0.478	0.3484	0.0135
s_times_RP_full	-0.11	0.573	0.1791	0.3475
s_times_TR_full	-0.4655	0.0012	-0.085	0.543
sec_Agricult	-0.48	0.1754	-0.3444	0.3386
sec_Constr	0.8304	0.1209	0.3984	0.4859
sec_Energy	-0.1194	0.6822	0.097	0.7451
sec_Fin	-0.2459	0.1544	0.0419	0.8084
sec_Industry	0.2763	0.6423	0.3631	0.5087
sec_Manufact	0.9645	0.1476	0.0318	0.9658
sec_Mining	-0.3101	0.3778	-0.2688	0.4787
sec_Service	-0.0976	0.5569	0.000845	0.996
sec_Trade	-0.1818	0.4364	0.2341	0.3141
sec_Trans	-0.1571	0.7093	-0.1123	0.8097
UT0_1	0.5316	0.7282	1.7442	0.2063
UT0_2	0.4324	0.4681	-0.3336	0.5721
UT0_3	-0.103	0.834	0.5848	0.2174

The multinomial logistic regression for the revolver state for t+1 (Table 6.25) gives coefficients for the transition to the Revolver paid state and to the delinquent one state. Staying in the revolver state is predicted as the full probability minus the probability of transitions to the revolver paid and the delinquent states. The behavioural variables are mainly significant and may have different signs. For example, the growth of the logarithm of the ratio of average outstanding balance in the current month to the average outstanding balance in the previous months (b_OB1_to_OB2_ln) causes a decrease of the chance to be delinquent (-0.3477) and causes an increase in the chance of paying back the full amount (0.094). It is interesting that if an account has been in Delinquent 1 state in the last 6 months, it has more chances to be paid than go to delinquency DPD 1-30 state again (coefficient s Been_D1_full 0.88 versus -0.83).

Table 6.24 Target frequencies for t+1, revolver state

Ordered value	Target_SI_dev	Total freq
1	D1	3201
2	RE	154064
3	RP	5838

Table 6.25 Multinomial regression estimations for t+1, from revolver state

Char	D1		RP	
	Estimate	Pr > ChiSq	Estimate	Pr > ChiSq
Intercept	-5.671	0	0.3745	0.265
b_atm_flag_use13vs46	0.2006	0.0233	-0.0627	0.2136
b_atm_use_only_flag_	-0.4986	0	-0.145	0.0002
b_avgNumDeb13	-0.0264	0.0176	-0.00957	0.0841
b_AvgOB1_to_MaxOB1_I	0.6517	0.0358	0.0777	0.2684
b_AvgOB16_to_MaxOB16	0.0767	0.8002	-0.2691	0.0282
b_max_dpd16	-0.00745	0.0027	-0.0106	0.0152
b_maxminOB_avgOB_1_I	-1.6619	0	0.6023	0
b_maxminOB_avgOB_16_	0.1805	0.3215	-0.1684	0.012
b_maxminOB_limit_1_I	1.9874	0	-0.3053	0
b_maxminOB_limit_16_	-0.1564	0.3703	0.4473	0
b_OB_avg_to_eop1n	-0.6019	0.0117	0.262	0
b_OB1_to_OB2_In	-0.3477	0.0044	0.094	0
b_OBBias_1_In	-0.1552	0	-0.1042	0
b_OBBias_16_In	-0.1324	0.0076	-0.2242	0
b_payment_lt_5p_1	-0.1448	0.0279	0.2197	0.0004
b_payment_lt_5p_2	-0.1474	0.0123	0.1441	0.0058
b_payment_lt_5p_3	-0.0262	0.6018	-0.1452	0.0002
b_pos_flag_use13vs46	0.034	0.6528	-0.0486	0.3062
b_pos_use_only_flag_	0.3338	0	0.2544	0
b_TRavg_deb1_to_26_I	-0.0788	0.0425	-0.0389	0.0021
b_TRavg_deb16_to_avg	-0.52	0	0.3504	0
b_TRmax_d*b_TRsum_de	0.0778	0	-0.00897	0.0135
b_TRmax_deb16_To_Lim	0.0359	0.5152	-0.2065	0
b_TRsum_crd1_to_2_In	0.197	0	-0.057	0
b_TRsum_crd13_to_46_	-0.0952	0.0007	-0.0225	0.0522
b_TRsum_crd13_to_OB1	0.0476	0.3179	0.0175	0.3065
b_TRsum_deb1_to_2_In	-0.0574	0.0455	0.00699	0.4314
b_TRsum_deb1_to_TRSu	0.7085	0	-0.1701	0
b_TRsum_deb16_to_TRs	0.1008	0.1016	-0.1429	0
b_UT13to46ln	0.0438	0.408	0.0185	0.0886
b_UT1to2ln	0.2395	0.0159	-0.0235	0.0194
age	-0.0161	0	0.00356	0.0838
avg_balance_6	-0.00014	0	-0.00003	0.0603
car_coOwn	0.0228	0.8067	-0.0566	0.3538

Char	D1		RP	
	Estimate	Pr > ChiSq	Estimate	Pr > ChiSq
car_Own	-0.0393	0.5486	0.067	0.0796
child_1	-0.1369	0.0524	-0.1933	0.0001
child_2	-0.0962	0.0293	-0.0652	0.0206
child_3	0.1113	0.4525	-0.1612	0.0933
customer_income_In	0.149	0.0481	0.1001	0.0332
Edu_High	-0.1128	0.0711	0.0904	0.0739
Edu_Special	-0.042	0.4785	0.0681	0.1775
Edu_TwoDegree	-0.3415	0.0861	0.1443	0.1496
I_ch1_flag	-0.3158	0.2891	0.0991	0.5553
I_ch1_In	-0.3325	0.6717	-0.2085	0.4585
I_ch6_flag	-0.1322	0.135	-0.1912	0.0005
limit_6	0.000094	0	-0.00002	0.0017
m_CPI_Inqoq_6	4.576	0.0184	-2.3655	0.0598
m_SalaryYear_Inyoy_6	1.5242	0.0923	1.0987	0.0654
m_UAH_EURRate_Inmom_6	0.393	0.7009	-1.1772	0.1098
m_Unempl_Inyoy_6	-6.7609	0	1.7149	0.0079
Marital_Civ	0.0488	0.6126	0.023	0.7584
Marital_Div	0.0247	0.7426	-0.00493	0.923
Marital_Sin	0.0127	0.8607	-0.0515	0.3527
Marital_Wid	-0.00194	0.9903	-0.1493	0.1765
mob	0.021	0.0014	0.00783	0.076
position_Man	-0.0175	0.8225	-0.0136	0.7729
position_Oth	-0.0462	0.5063	-0.0811	0.1089
position_Tech	-0.0764	0.223	-0.0578	0.2375
position_Top	0.1444	0.3318	0.2551	0.0012
real_coOwn	0.1042	0.0634	-0.00012	0.9978
real_Own	0.1062	0.0606	0.0567	0.154
reg_ctr_N	-0.1036	0.2181	-0.00949	0.8721
reg_ctr_Y	-0.1283	0.132	-0.0179	0.7598
s_been_D1_full	0.8836	0	-0.8334	0
s_been_Tr_full	1.6466	0.2165	-0.0199	0.8432
s_cons_full	-0.0864	0.0131	-0.0145	0.7193
s_month_since_NA_ful	0.4847	0	0.0674	0.0012
s_month_since_RP_ful	0.1159	0.2444	-0.00881	0.7879
s_times_RE_full	-0.6164	0	-0.4231	0
s_times_RP_full	-0.4215	0.2593	-0.1153	0.1798
s_times_TR_full	-2.1759	0.0906	-0.2032	0.0082
sec_Agricult	0.0161	0.8944	0.0488	0.5807
sec_Constr	0.1883	0.2069	-0.0899	0.4797
sec_Energy	-0.0593	0.6186	0.1491	0.0454
sec_Fin	-0.1842	0.0806	-0.0972	0.0554
sec_Industry	0.1052	0.6381	-0.1447	0.3789
sec_Manufact	-0.0195	0.901	-0.0136	0.9114
sec_Mining	-0.00247	0.983	0.0644	0.4605
sec_Service	0.1251	0.0392	-0.0338	0.424
sec_Trade	0.1134	0.133	-0.1116	0.0635
sec_Trans	0.2598	0.0969	0.0399	0.7263
UT0_1	-0.3952	0.3345	-1.0624	0
UT0_2	1.3293	0	-0.1684	0.1962
UT0_3	-0.4704	0.048	-0.1295	0.2244

We first investigate the coefficient estimates for new ‘Revolver Paid’ state (Table 6.27). The estimates for the transition to the Revolver state are mainly insignificant, and it is necessary to look at the total predictive power of the model to decide whether to include it into the general system of transition prediction models or not (see section 6.5.3). However, for the transition to the inactive state, we can find several significant covariates. For example, b_avgNumDeb13 with a negative sign which means that higher number of debit transactions in the month corresponds with lower probability to move to inactive in the next month.

Table 6.26 Target frequencies for t+1, from revolver paid state

Ordered value	Target_SI_dev	Total freq
1	NA	2936
2	RE	2067
3	TR	874

Table 6.27 Multinomial regression estimations for t+1, from revolver paid state

Char	NA		RE	
	Estimate	Pr > ChiSq	Estimate	Pr > ChiSq
Intercept	1.5292	0.0746	1.7381	0.0443
b_atm_flag_use13vs46	0.0968	0.5006	0.1326	0.3537
b_atm_use_only_flag_	0.0824	0.457	0.09	0.4169
b_avgNumDeb13	-0.1464	0	-0.0127	0.3251
b_AvgOB1_to_MaxOB1_I	-0.6261	0.0713	1.4463	0.3336
b_AvgOB16_to_MaxOB16	-0.6314	0.7243	-2.7778	0.1103
b_max_dpd16	0.099	0.3729	0.0977	0.3788
b_maxminOB_avgOB_1	-1.7443	0	0.8052	0.5688
b_maxminOB_avgOB_16	-0.1582	0.9248	-2.7287	0.0978
b_maxminOB_limit_1_In	-0.0697	0.0745	-0.078	0.0484
b_maxminOB_limit_16	0.0368	0.769	-0.0687	0.5798
b_OB_avg_to_eop1ln	0.0306	0.5052	0.0106	0.8256
b_OB1_to_OB2_In	0.0331	0.4914	0.0233	0.6379
b_OBbias_1_In	0.889	0	0.523	0.0003
b_OBbias_16_In	-0.3417	0.094	-0.1612	0.4092
b_payment_lt_5p_1	0.4541	0.3299	-1.0403	0.0321
b_payment_lt_5p_2	0.3349	0.0348	-0.0751	0.638
b_payment_lt_5p_3	0.0633	0.569	0.0778	0.4854
b_pos_flag_use13vs46	0.026	0.8478	-0.017	0.9
b_pos_use_only_flag_	-0.2916	0.0203	0.034	0.7773
b_TRavg_deb1_to_26_I	-0.1925	0	-0.0562	0.0734
b_TRavg_deb16_to_avg	-0.2025	0.0399	-0.1901	0.0489
b_TRmax_d*b_TRsum_de	0.00048	0.9328	0.00391	0.4878
b_TRmax_deb16_To_Lim	0.0403	0.7215	0.1104	0.3149
b_TRsum_crd1_to_2_In	-0.0202	0.3886	0.0282	0.2355
b_TRsum_crd13_to_46_	0.0326	0.3122	-0.0421	0.2043
b_TRsum_crd13_to_OB1	-0.3265	0.0013	-0.1509	0.1476
b_TRsum_deb1_to_2_In	0.0475	0.0602	-0.0095	0.716
b_TRsum_deb1_to_TRsu	-0.0629	0.0392	0.0915	0.0034
b_TRsum_deb16_to_TRs	-0.0945	0.5924	0.2626	0.1337
b_UT13to46ln	-0.00511	0.8402	0.0406	0.1171
b_UT1to2ln	-0.0112	0.5144	-0.00253	0.8873
age	0.00509	0.3807	-0.00754	0.1925
avg_balance_6	-0.00008	0.1481	-0.00005	0.2477
car_coOwn	-0.0449	0.7862	-0.1385	0.4071
car_Own	0.1531	0.1522	0.0773	0.4678
child_1	-0.1325	0.3505	0.0462	0.741
child_2	-0.1254	0.114	-0.0577	0.4661
child_3	-0.2627	0.3308	0.0514	0.8473
customer_income_In	0.0823	0.5213	-0.1383	0.269
Edu_High	0.1037	0.4668	-0.1163	0.408
Edu_Special	0.1609	0.2636	0.0641	0.654
Edu_TwoDegree	0.4071	0.161	0.3117	0.2504
I_ch1_flag	-0.2931	0.5642	0.3855	0.4323
I_ch1_In	-0.0285	0.9722	-0.7118	0.3815
I_ch6_flag	0.0765	0.6174	-0.0539	0.726
limit_6	-0.00001	0.5444	-0.00000148	0.9276
m_CPI_Inqoq_6	7.7992	0.0316	-1.4838	0.6811
m_SalaryYear_Inyoy_6	2.5546	0.1137	0.3673	0.8197
m_UAH_EURRate_Inmom_6	1.6558	0.4408	-1.3464	0.5295
m_Unempl_Inyoy_6	2.6908	0.1204	0.0913	0.9574
Marital_Civ	-0.1632	0.4271	-0.2304	0.2549
Marital_Div	0.1463	0.3254	0.1122	0.4456
Marital_Sin	-0.1682	0.2719	-0.1671	0.2652
Marital_Wid	0.4452	0.1901	0.1646	0.6425
mob	0.0215	0.0797	0.0149	0.2191
position_Man	-0.0879	0.4924	-0.0298	0.813
position_Oth	-0.3575	0.0096	-0.2411	0.081
position_Tech	-0.0599	0.6753	-0.1176	0.4159
position_Top	-0.0802	0.7159	0.3526	0.0931
real_coOwn	-0.1887	0.1198	-0.1036	0.3843
real_Own	-0.1609	0.1574	-0.0687	0.5413

Char	NA		RE	
	Estimate	Pr > ChiSq	Estimate	Pr > ChiSq
reg_ctr_N	0.1482	0.3716	-0.1221	0.4425
reg_ctr_Y	0.0605	0.7092	-0.1531	0.3208
s_been_D1_full	-1.6134	0.2005	-1.0751	0.3889
s_been_Tr_full	-0.1076	0.6797	0.214	0.3808
s_cons_full	0	0	0	0
s_month_since_NA_ful	0.0295	0.6369	0.0752	0.2292
s_month_since_RP_ful	0	0	0	0
s_times_RE_full	-0.1607	0.2369	-0.0693	0.6131
s_times_RP_full	-0.2939	0.0358	-0.21	0.1371
s_times_TR_full	-0.2884	0.1662	-0.3823	0.054
sec_Agricult	0.123	0.6419	0.5038	0.0536
sec_Constr	-0.1074	0.7447	-0.2578	0.4475
sec_Energy	0.0253	0.9042	-0.0339	0.8746
sec_Fin	-0.106	0.4438	0.171	0.2032
sec_Industry	-0.5641	0.2246	0.00884	0.9838
sec_Manufact	0.1401	0.7333	0.5526	0.1783
sec_Mining	0.2057	0.4356	0.3696	0.1652
sec_Service	0.00895	0.9403	0.1142	0.3376
sec_Trade	-0.2085	0.2041	-0.0463	0.773
sec_Trans	0.6653	0.1016	0.9483	0.0179
UT0_1	-1.4194	0.0085	-0.0925	0.8502
UT0_2	-0.1429	0.6704	-0.1771	0.5914
UT0_3	-0.2065	0.466	-0.057	0.8377

The multinomial regression estimations for the Delinquency 1 state are computed for transitions to Delinquent 1, Delinquent 2, and Revolver states (Table 6.29). Revolver Paid state is selected as a basic state. The estimates are mainly insignificant, the common problem of the coefficient estimates for the multinomial regressions. However, we will pay attention to the model prediction accuracy estimation, and in case of appropriate KS and Gini coefficients (section 6.5.3), the model can be used for the further prediction.

Table 6.28 Target frequencies for t+1, from delinquent 1 state

Ordered value	Target_SI_dev	Total freq
1	D1	956
2	D2	743
3	RE	2302
4	RP	29

Table 6.29 Multinomial regression estimations for t+1, from delinquents 1 state

Char	D1		D2		RE	
	Estimate	Pr > ChiSq	Estimate	Pr > ChiSq	Estimate	Pr > ChiSq
Intercept	-1.3421	0.8991	-6.338	0.5548	1.8252	0.8581
b_atm_flag_use13vs46	-0.4986	0.5354	-0.4587	0.5723	-0.2759	0.7277
b_atm_use_only_flag_	-0.761	1	13.0536	0.9571	-6.5417	0.9997
b_avgNumDeb13	0.00264	0.9801	-0.0679	0.5439	-0.0179	0.863
b_AvgOB1_to_MaxOB1_I	8.4943	0.5481	2.0638	0.8854	11.6657	0.393
b_AvgOB16_to_MaxOB16	7.0383	0.0314	2.7895	0.3911	5.3258	0.076
b_max_dpd16	-0.0226	0.0693	-0.00383	0.7431	-0.0163	0.1759
b_maxminOB_avgOB_1_I	3.0178	0.3217	5.1875	0.084	5.9062	0.0396
b_maxminOB_avgOB_16_	-0.3248	0.8916	-3.1205	0.181	-0.0644	0.9764
b_maxminOB_limit_1_I	-2.3762	0.4138	-4.7275	0.0989	-5.303	0.0524
b_maxminOB_limit_16_	1.4723	0.5243	3.6282	0.1108	1.1344	0.593
b_OB_avg_to_eap1n	-0.2154	0.9847	11.2853	0.3177	-6.1745	0.5724
b_OB1_to_OB2_In	2.6226	0.8699	19.9849	0.2133	8.9288	0.5733
b_OBbias_1_In	0.2223	0.5154	0.0989	0.7761	0.3624	0.2817
b_OBbias_16_In	0.262	0.6465	0.1753	0.761	0.1624	0.7727
b_payment_lt_5p_1	0.7148	0.377	0.6493	0.429	-0.3665	0.6478

Char	D1		D2		RE	
	Estimate	Pr > ChiSq	Estimate	Pr > ChiSq	Estimate	Pr > ChiSq
b_payment_lt_5p_2	1.1722	0.203	0.7249	0.4353	1.0627	0.2465
b_payment_lt_5p_3	0.0173	0.9751	0.2097	0.7072	0.0834	0.8795
b_pos_flag_use13vs46	-0.203	0.7576	-0.2088	0.7532	-0.5011	0.4377
b_pos_use_only_flag_	0.8359	0.2291	0.9107	0.1919	0.9083	0.186
b_TRavg_deb1_to_26_l	-0.2349	0.7714	-0.2063	0.8003	-0.1991	0.8034
b_TRavg_deb16_to_avg	0.4259	0.6809	0.3895	0.7098	0.6176	0.5459
b_TRmax_d*b_TRsum_de	-0.1526	0.437	-0.1515	0.4476	-0.1183	0.5384
b_TRmax_deb16_To_Lim	-0.9289	0.0963	-0.4939	0.3795	-0.8993	0.1025
b_TRsum_crd1_to_2_ln	-0.0493	0.601	0.0867	0.3601	-0.0585	0.533
b_TRsum_crd13_to_46	0.2057	0.4141	0.2244	0.3724	0.1442	0.5633
b_TRsum_crd13_to_OB1	0.3834	0.3041	0.1942	0.5983	0.4775	0.1902
b_TRsum_deb1_to_2_ln	0.2971	0.3778	0.1926	0.5715	0.1659	0.6135
b_TRsum_deb1_to_TRsu	0.1319	0.3128	0.4101	0.0019	0.0364	0.7786
b_TRsum_deb16_to_TRs	0.826	0.0274	0.9793	0.0119	0.5314	0.1335
b_UT13to46ln	-0.2433	0.5772	-0.9847	0.0356	-0.2869	0.4877
b_UT1to2ln	8.3982	0.6029	-6.0523	0.7078	2.4585	0.8768
age	0.0777	0.0862	0.0469	0.3022	0.072	0.1098
avg_balance_6	0.000163	0.4681	0.000188	0.4106	0.000354	0.0975
car_coOwn	2.9558	0.2503	2.7535	0.2848	2.9991	0.2426
car_Own	-0.376	0.5553	-0.6304	0.3284	-0.0681	0.9136
child_1	-1.0154	0.1639	-0.6349	0.3869	-0.9707	0.1793
child_2	0.5261	0.4548	0.6859	0.3312	0.6458	0.3565
child_3	-3.8231	0.0065	-2.9829	0.0344	-3.0667	0.0255
customer_income_ln	-1.353	0.0724	-1.2315	0.1047	-1.3569	0.0673
Edu_High	0.2643	0.6844	0.2714	0.6784	0.5446	0.3981
Edu_Special	0.8604	0.194	1.0521	0.1137	1.0267	0.1182
Edu_TwoDegree	2.9884	0.6835	2.952	0.6872	3.8177	0.602
I_ch1_flag	48.4936	0	49.3013	0	48.8607	0
I_ch1_ln	190.9	0	172.7	0	184	0
I_ch6_flag	-1.2462	0.0744	-0.9694	0.1698	-1.0851	0.1104
limit_6	-0.00002	0.9042	-0.00003	0.8615	-0.00023	0.1786
m_CPI_Inqoq_6	35.4269	0.1134	31.0217	0.1682	24.558	0.2684
m_SalaryYear_Inyoy_6	-16.5295	0.1767	-12.3532	0.3149	-18.4469	0.1295
m_UAH_EURRate_Inmom_6	26.3634	0.0598	30.6295	0.0293	26.4345	0.0577
m_Unempl_Inyoy_6	-10.0062	0.3591	-11.394	0.298	-6.0146	0.5789
Marital_Civ	1.1026	0.4833	1.6383	0.2978	1.3042	0.405
Marital_Div	-0.8895	0.2073	-0.6423	0.3661	-0.777	0.2643
Marital_Sin	-0.3045	0.6784	-0.0159	0.9829	-0.1392	0.8483
Marital_Wid	0.1341	0.9758	0.0262	0.9953	0.2024	0.9635
mob	0.0634	0.3825	0.053	0.4675	0.0762	0.2895
position_Man	0.2882	0.7234	0.5036	0.5396	0.2051	0.7992
position_Oth	-0.3166	0.6735	0.1255	0.8678	-0.1559	0.8336
position_Tech	0.3394	0.6428	0.6049	0.41	0.2915	0.6883
position_Top	0.5379	0.7224	0.6515	0.6698	0.5702	0.7018
real_coOwn	0.1387	0.8139	0.3473	0.5574	0.3885	0.5054
real_Own	0.2648	0.6574	0.3294	0.5829	0.3345	0.5715
req_ctr_N	-1.0018	0.3083	-0.7606	0.4422	-1.1075	0.2561
req_ctr_Y	-0.9505	0.3355	-0.8415	0.3973	-1.0926	0.2641
s_been_D1_full	0	0	0	0	0	0
s_been_Tr_full	22.5829	0.9496	20.0615	0.9553	16.3833	0.9632
s_cons_full	0.969	0.1346	1.1306	0.0817	0.697	0.281
s_month_since_NA_ful	0.2493	0.598	-0.0782	0.8665	-0.2658	0.5435
s_month_since_RP_ful	-0.148	0.9028	0.2741	0.8239	-0.0156	0.9894
s_times_RE_full	-0.3045	0.3078	-0.3025	0.3117	0.0496	0.8673
s_times_RP_full	0.2391	0.9521	1.8465	0.6503	0.5293	0.8899
s_times_TR_full	-10.2718	0.9644	-7.9971	0.9724	-3.7208	0.9869
sec_Agricult	-1.4749	0.1082	-1.7343	0.0637	-1.9968	0.0276
sec_Constr	4.6666	0.6571	5.0736	0.6293	4.7794	0.6492
sec_Energy	1.9244	0.5614	1.5075	0.6495	1.8873	0.5686
sec_Fin	-0.2014	0.8279	-0.6604	0.4849	-0.4903	0.5902
sec_Industry	-1.8601	0.2038	-3.0442	0.0457	-2.0348	0.1571
sec_Manufact	-0.8791	0.594	-0.3653	0.8243	-0.6445	0.6915
sec_Mining	-0.0574	0.9647	0.0735	0.9549	-0.0402	0.975
sec_Service	-0.114	0.8677	-0.0417	0.9517	-0.1685	0.8039
sec_Trade	1.0709	0.2333	0.8092	0.3699	0.6865	0.442
sec_Trans	2.1253	0.5938	1.6862	0.6726	2.0163	0.6123
UT0_1	-8.6805	0.2157	-14.2172	0.0405	-8.1359	0.2025
UT0_2	23.0446	0.0035	28.2151	0.0003	21.5992	0.0034
UT0_3	-5.8804	0.0153	-4.6498	0.058	-4.35	0.0669

We have put the same set of predictors for each model. As it can be seen from the estimation p-level values, some predictors have high significance for one model, but low significance for another one, or are insignificant. These models predict the transition from the state at time t to state at time t+1 for the 1-month horizon. Because our final aim is income prediction for the 6-month period, we need to build the models for longer periods.

6.5.2 Model t+3

We will not present all estimates for the multinomial logistic regression model for t+3 and t+6 months period of prediction. For the comparative analysis of the covariates signs for different transitions, we give an example of the coefficient estimations for 3 months outcome period for transitions from the revolver state (see Table 6.30).

The same covariates have different signs for various state transitions. For example, the usage of a credit card for ATM cash withdrawals only (*b_atm_use_only_flag*) decreases the probability to go to delinquent and default states in the 3-months period (negative coefficients values -0.3396, -0.2528, and -0.0988) and increases the chance to stay a Revolver (positive coefficient 0.1216). On the other hand, the usage of a credit card for Point-of-Sales (POS) transactions only (*b_pos_use_only_flag*) decreases the probability of staying a Revolver and increases the probability of default in a 3-month period (positive coefficient 0.0841). If a revolver customer did not use a card for ATM cash withdrawals for the period of 4-6 months ago, but has started to withdraw cash for the last 1-3 months, he/she has higher chances to pay back the full amount, or become Revolver Paid, in the next 3 months, and the opposite tendency – lower chances to paid back the full amount due – can be concluded from the negative coefficient for whose clients, who did not use, but started to make POS transactions (*b_pos_flag_use13vs46* = -0.1262). Higher current utilisation rate *UT0_1* decreases the chance to become inactive or pay back the full amount in 3 months (negative coefficients -0.6549 and -0.25) but increases the chances to stay a Revolver or go to Delinquent and Default states. The increase of the credit limit in the last month (*l_ch1_flag*) increase the chance to become an Inactive in a 3-month period, but the increase of the credit limit in the six-months period before the observation point decreases the chance to become an inactive and increase of chance of transition to

other states. For the more extended period on the book, a customer will become an inactive, a repaid revolver or a delinquent 1 than a revolver or a default.

Table 6.30 Coefficients estimations for Revolver state transition for t+3 states prediction

Parameter	t+3					
	NA	RE	RP	D1	D2	Df
Intercept	2.3672	2.0085	1.2332	-2.0829	-5.2086	-8.8038
b_atm_flag_use13vs46	0.0134	-0.0409	0.0329	0.1538	0.03	-0.1772
b_atm_use_only_flag	-0.00505	0.1216	-0.018	-0.3396	-0.2528	-0.0988
b_avgNumDeb13	-0.1064	-0.0117	-0.0329	-0.0204	-0.0653	-0.074
b_AvgOB1_to_MaxOB1_I	0.2511	0.0507	0.191	-0.6123	-0.6573	5.2972
b_AvgOB16_to_MaxOB16	-0.2023	0.00974	0.1136	-1.1549	-2.586	-0.7165
b_max_dpd16	0.907	0.9156	0.8978	0.9135	0.916	0.9223
b_maxminOB_avgOB_1_I	0.1942	-0.3645	-0.1125	-0.6375	-1.0681	-1.6514
b_maxminOB_avgOB_16_	-0.2543	-0.0777	-0.0243	-1.1009	-2.9329	-2.2452
b_maxminOB_limit_1_I	-0.1371	0.1298	0.0768	0.5646	0.7448	1.8163
b_maxminOB_limit_16_	0.4015	-0.1933	0.111	0.6969	2.3638	1.6146
b_OB_avg_to_eop1ln	0.0833	-0.1198	-0.0867	-0.0966	-0.4	-5.9041
b_OB1_to_OB2_ln	-0.00185	-0.0838	0.0534	-0.182	0.4179	-1.8056
b_OBbias_1_ln	0.1098	0.1189	0.133	0.038	0.105	0.1245
b_OBbias_16_ln	-0.108	0.0824	0.0311	-0.1033	-0.1127	-0.1171
b_payment_lt_5p_1	0.1233	-0.0567	-0.1953	-0.2749	-0.2301	-0.1795
b_payment_lt_5p_2	0.0151	-0.0253	0.0532	-0.2361	-0.563	-0.4389
b_payment_lt_5p_3	0.0475	0.0735	-0.0106	0.1507	0.1536	-0.3388
b_pos_flag_use13vs46	-0.0124	0.0471	-0.1262	0.0363	-0.0792	0.0385
b_pos_use_only_flag_	-0.1615	-0.24	-0.0956	-0.0935	-0.1469	0.0841
b_TRavg_deb1_to_26_I	0.00443	0.0365	0.0444	-0.0667	0.00243	-0.1358
b_TRavg_deb16_to_avg	0.0578	-0.3004	0.0304	-0.673	-0.8864	-1.0777
b_TRmax_deb16_To_Lim	-0.1968	0.0899	-0.0653	0.1136	0.224	0.3127
b_TRsum_crd1_to_2_ln	0.0128	0.0683	0.0235	0.1934	0.2501	0.2955
b_TRsum_crd13_to_46_	0.000694	0.0259	0.0191	0.00295	0.0137	-0.0916
b_TRsum_crd13_to_OB1	-0.043	-0.0446	-0.0222	-0.0824	-0.0574	-0.1753
b_TRsum_deb1_to_2_ln	0.0363	-0.00265	-0.00264	-0.00921	-0.0309	0.0165
b_TRsum_deb1_to_avgO	-0.2109	-0.0685	-0.1759	-0.2824	-0.1146	-0.5002
b_TRsum_deb1_to_TRsu	-0.00096	0.1426	0.0894	0.4896	0.5365	0.8928
b_TRsum_deb16_to_TRs	-0.0121	0.1327	-0.0274	0.289	0.1944	0.2898
b_UT13to46ln	0.0355	0.000112	0.000021	-0.00416	0.0955	0.4376
b_UT1to2ln	-0.0136	0.00219	0.00745	-0.0335	-0.4414	0.9033
age	0.00314	-0.00439	-0.00029	-0.0255	-0.0446	-0.0407
avg_balance_6	-0.00003	-0.0000895	-0.00003	-0.00005	-0.00005	-0.00004
car_coOwn	0.0242	-0.0186	-0.00673	0.0154	-0.2831	-0.2699
car_Own	0.0398	-0.0737	0.00355	-0.1269	-0.5943	-0.3239
child_1	-0.2654	0.02	-0.2071	-0.1306	-0.0867	0.1979
child_2	-0.1598	-0.0418	-0.1492	-0.1219	-0.2008	0.00375
child_3	-0.4846	-0.0964	-0.2523	-0.1395	-0.1617	0.5553
CPI_Inqoq_6	0.861	3.0329	-0.7811	10.6709	5.2326	-1.1934
customer_income_ln	0.0145	-0.1682	-0.0715	-0.2648	-0.3422	-0.2559
Edu_High	0.0249	-0.1324	0.00887	-0.2625	-0.4456	-0.6018
Edu_Special	0.1296	0.0394	0.0996	-0.0613	-0.0128	0.00587
Edu_TwoDegree	-0.0151	-0.00611	0.2496	-0.2518	-0.2917	-0.7037
I_ch1_flag	0.0987	-0.0678	-0.0374	-0.0589	-0.6869	-0.6694
I_ch1_ln	-0.5867	0.2579	-0.4053	0.5093	2.032	3.5491
I_ch6_flag	-0.0517	0.2039	0.0891	0.1428	0.0976	0.2457
limit	0.0004834	0.00021	0.00016	0.00039	0.00064	0.000085
Marital_Civ	-0.2299	-0.2054	-0.1502	-0.1718	0.0596	0.1101
Marital_Div	0.0904	0.0449	0.0792	0.0817	0.2908	0.3835
Marital_Sin	-0.2001	-0.1597	-0.2491	-0.1026	-0.1995	0.0995
Marital_Wid	-0.00074	0.1081	-0.1745	0.0982	0.2792	0.5055
mob	0.00189	-0.00735	0.00897	0.00863	-0.0137	-0.0202
position_Man	-0.1216	-0.07	-0.0597	0.0126	0.1655	0.136
position_Oth	-0.3465	-0.1612	-0.2276	-0.1739	0.0463	-0.1197
position_Tech	-0.1438	-0.0997	-0.134	-0.1856	0.00146	0.0943
position_Top	-0.1056	-0.1999	0.049	0.0131	0.4946	0.5225
real_coOwn	0.00992	0.00396	-0.0259	-0.0114	-0.1169	0.0197
real_Own	-0.1038	-0.1677	-0.1261	-0.1137	-0.3072	-0.2643
reg_ctr_N	0.0674	-0.0995	-0.0195	-0.2239	0.1919	-0.1972
reg_ctr_Y	-0.1051	-0.205	-0.1546	-0.3163	0.071	-0.1956
s_been_D1_full	-10.1256	-9.2002	-9.5944	-8.1691	-9.2788	-9.9159
s_been_Tr_full	-0.0976	-0.00473	0.0689	-0.6165	7.5269	4.925
s_cons_full	0.00563	-0.0588	-0.0798	-0.0552	-0.2579	-0.4154
s_month_since_NA_ful	-0.00309	-0.0318	-0.0234	0.2493	0.5778	0.9577

Parameter	t+3					
	NA	RE	RP	D1	D2	Df
s_month_since_RP_full	0.0241	0.0827	0.1016	0.0645	0.1461	0.2489
s_times_RE_full	-0.1181	0.3193	0.1396	-0.1505	-0.2438	-0.4153
s_times_RP_full	-0.2411	0.0414	0.1576	-0.299	-0.5251	0.147
s_times_TR_full	-0.4381	-0.1961	-0.2385	0.0435	-8.233	-5.8404
SalaryYear_Inoy	1.75	-1.5081	-1.5898	-3.2829	5.6549	5.4588
sec_Agricult	0.4753	0.4883	0.5824	0.6981	0.5736	0.0433
sec_Constr	-0.0113	-0.0213	0.0286	0.2918	0.7954	-0.1509
sec_Energy	0.0281	-0.111	0.0704	-0.1986	-0.2146	-0.3584
sec_Fin	0.0908	0.1957	0.1994	-0.0325	-0.0952	0.0774
sec_Industry	-0.0921	0.2508	0.1325	0.2091	-0.3263	0.2633
sec_Manufact	0.6016	0.559	0.5483	0.6597	0.9179	0.3023
sec_Mining	0.5694	0.4679	0.5462	0.6113	0.6187	0.7475
sec_Service	0.0673	0.0561	0.0495	0.1556	0.3725	0.2825
sec_Trade	-0.066	0.0337	0.00403	0.3138	0.3829	0.2514
sec_Trans	0.496	0.3895	0.4656	0.4952	0.2609	0.2576
UAH_EURRate_Inmom	2.8445	1.3307	2.6285	3.3438	-0.6151	-1.8219
Unempl_Inoy	0.5962	-3.2983	-2.4683	-12.2072	-9.8972	-12.9246
UTO_1	-0.6549	0.7108	-0.25	0.9232	0.5098	3.5912
UTO_2	0.1179	0.1826	0.1324	0.0102	0.2594	-2.7467
UTO_3	0.00916	0.1973	-0.00784	0.4347	0.113	0.0219

We do not provide with analysis of coefficient estimates for a 6-months period because it mainly replicates the described results for shorter periods. The more critical issue in the scope of our research is the predictive power of partial logistic regression models for the prediction of transitions from any state to all possible states.

We define the probability of transition to the longer period which is multiple to the estimated models by multiplication of the estimations for the shorter periods. For example, the probability of the transition for two months period is computed as a product of two probabilities of transition for one-month period.

6.5.3 Multinomial regression models validation results

We compare the Kolmogorov-Smirnov and Gini indices for the development and validation samples for a full set transition models built with multinomial logistic regression for t+1, t+2, t+3, and t+6 prediction period. In Table 6.31 the column Model means the current state a transition *from* which we predict. The column Target means the state a transition *to* which we predict.

The predictive power of the t+1 models is strong, especially, for the transitions from the revolver state for which we obtained Gini index values are around 0.7 - 0.77 for the transition to revolver and revolver paid states, and 0.87 for the transition to delinquent 1 for the validation sample. On the other hand, the predictive power for transitions from inactive and transactor states is weak – around 0.2-0.5 for the

validation sample (Table 6.31). For delinquent states, it is easier predict the next month transition to revolver or deeper delinquency, default states (Gini around 0.5 – 0.6) than the staying in the current delinquent state (Gini around 0.05-0.4). The new state ‘Revolver Paid’ has a moderate Gini value for the transition to inactive state – 0.53, but significantly lower Gini for their transitions to the transactor and revolver states (0.33 and 0.38 respectively), but these indicators are higher than for prediction of the transitions from the transactor state.

Table 6.31 KS and Gini coefficients for development and validation sample target t+1

Model	Target	Development		Validation	
		KS	GINI	KS	GINI
NA	NA	22.64	0.3096	20.68	0.2899
	TR	27.85	0.3664	29.85	0.3784
	RE	21.04	0.2871	17.54	0.2385
TR	NA	37.76	0.4813	31.89	0.4012
	TR	28.39	0.3669	22.27	0.2702
	RE	31.54	0.4211	25.23	0.3140
RE	RE	60.77	0.7346	61.54	0.7428
	RP	65.17	0.7662	65.74	0.7777
	D1	74.11	0.8709	75.39	0.8707
RP	NA	39.84	0.5172	41.42	0.5306
	TR	32.68	0.4298	27.20	0.3340
	RE	29.65	0.3900	29.91	0.3826
D1	RE	45.36	0.5959	46.78	0.5694
	RP	67.93	0.7879	27.72	0.1184
	D1	37.91	0.4946	36.80	0.4612
D2	D2	51.16	0.6678	47.54	0.6048
	RE	57.50	0.713	55.27	0.6501
	D1	53.68	0.6569	38.01	0.3968
	D2	54.20	0.7055	21.42	0.0595
	Df	53.25	0.6589	50.16	0.5680

The predictive power of models decreases with the increase in the period of prediction. For example, the Gini index for the transition from Revolver state to Delinquent 1 state is 0.87 for 1-month prediction, 0.66 for 3-months prediction (Table 6.32), and only 0.57 for a 6-month period (Table 6.33). The Gini index for staying inactive is 0.29 for the next month period, but 0.24 for 3 and 6 months predictions. It is quite difficult to predict what will happen with transactors in 3 months because of low Gini indexes - 0.27, 0.31, and 0.22 for the transition to the inactive, transactor, and revolvers state respectively. We also have an estimation for the transition of transactors to the Revolver Paid state, but because of few cases and quite a complicated way to this state – a transactor must go to revolver state and at the 3rd-month payback full amount – the

Gini index is very low as 0.06.

Table 6.32 KS and Gini coefficients for development and validation sample target t+3

Model	Target	Development		Validation	
		KS	Gini	KS	Gini
NA	NA	19.64	0.28	17.40	0.24
	TR	25.72	0.35	22.69	0.25
	RE	18.92	0.26	14.91	0.20
	RP	21.58	0.30	20.43	0.26
TR	NA	29.95	0.40	21.75	0.27
	TR	36.77	0.49	26.72	0.31
	RE	26.28	0.36	18.33	0.22
	RP	27.99	0.37	10.51	0.06
RE	NA	58.63	0.72	58.29	0.72
	TR	58.60	0.71	58.01	0.70
	RE	44.24	0.56	44.16	0.55
	RP	44.98	0.56	44.25	0.55
D1	D1	49.57	0.66	50.04	0.66
	D2	66.36	0.81	65.74	0.82
	Df	90.96	0.97	92.41	0.97
	NA	32.60	0.44	31.69	0.42
RP	TR	34.57	0.47	30.74	0.39
	RE	26.26	0.36	22.03	0.29
	RP	31.59	0.42	22.23	0.24
	RE	34.83	0.47	37.89	0.48
D2	RP	59.71	0.75	31.30	0.35
	D1	28.53	0.37	23.48	0.30
	D2	39.64	0.51	26.94	0.36
	Df	50.37	0.68	45.25	0.61
D2	RE	50.34	0.65	28.10	0.28
	D1	49.21	0.63	34.17	0.29
	D2	61.21	0.75	38.04	0.37
	Df	51.23	0.68	29.87	0.38

The predictive power decreases for a 6-month period of the prediction (Table 6.33). For example, Gini index for the transition from Delinquent 2 to Default state for 1 month is 0.57 versus 0.27 for a 6-month period, from Delinquent 1 state to Default state for 3 months is 0.61 versus 0.55 for a 6-month period.

We do not give the predictive power of the transition from the transactor state t the default state in 6-month period because of the lack of observations – only 3 cases of transactors who became defaulters have been observed.

Table 6.33 KS and Gini coefficients for development and validation sample target t+6

Model	Target	Development		Validation	
		KS	Gini	KS	Gini
NA	NA	19.83	0.2811	18.02	0.2440
	TR	27.02	0.3661	21.58	0.2493
	RE	19.95	0.2669	17.94	0.2420
	RP	20.04	0.2691	14.47	0.1706
	D1	66.57	0.8197	21.69	0.0970
TR	NA	30.39	0.4075	22.27	0.2583
	TR	35.94	0.4778	19.27	0.1649
	RE	25.86	0.342	20.24	0.2401
	RP	29.57	0.3832	12.69	0.0459
RE	NA	47.94	0.6184	48.03	0.6189
	TR	47.51	0.6075	43.62	0.5612
	RE	35.28	0.4633	35.13	0.4506
	RP	37.21	0.4849	34.96	0.4455
	D1	47.35	0.6152	43.74	0.5764
	D2	60.00	0.7642	50.35	0.6057
	Df	67.83	0.8147	60.48	0.7726
RP	NA	25.16	0.3497	26.83	0.3459
	TR	27.62	0.3753	30.02	0.3744
	RE	20.42	0.2856	20.33	0.2703
	RP	25.55	0.3386	19.97	0.2303
	D1	62.23	0.7406	47.69	0.3854
D1	RE	31.82	0.4309	25.64	0.3416
	RP	51.90	0.6607	42.93	0.4463
	D1	24.66	0.3276	19.15	0.2205
	D2	38.32	0.4928	8.27	0.0500
D2	Df	43.66	0.5737	43.93	0.5562
	RE	46.46	0.5471	15.08	0.1733
	D1	50.21	0.6291	21.19	0.1361
	D2	51.00	0.6841	15.41	0.0093
	Df	44.69	0.5767	22.69	0.2662

The predictive power of some model with the additional state ‘Revolver Paid’ is weaker than for the initial set of states, discussed in section 6.2.2, especially, for moving to transactor state and from inactive and transactor states.

The number of observations in the updated transactor states has been decreased around two times. However, despite a decrease in model predictive power, it can be interesting from the business point of view to separate clients who pay back the full amount to a separate category because they generate an interest income, unlike transactors who generate only transactional income.

6.6 Conclusion

We compared multinomial and ordinal logistic regression for the prediction of the transition between credit card states. The multinomial regression gives better predictive power results (KS and Gini coefficients) than ordinal regression.

We tested *multitarget logistic regression* models for the probability of transition between states. The comparative empirical analysis of multinomial logistic regression and conditional multistage binary logistic regression has shown that both methods do not have strict preferences or advantages and both of them give satisfactory validation results of transition prediction for different types of account statuses. Conditional binary logistic regression model efficiency depends on the order of stages. Multinomial regression gives a more convenient model in use and helps to avoid the problem of stage ordering choice. However, the order of stages in multistage binary logistic regression can be useful if it is known which segment is a more critical in the sense of quality prediction.

Multinomial logistic regression is the relatively innovative approach in risk modelling. On the other hand, the decision tree of conditional binary logistic regressions has given similar results. Both models have moderate, but not strong predictive power. Prediction accuracy for the decision tree depends on the order of stages for conditional binary logistic regression. An examination of possible options is a complicated and lengthy process.

Transactors and revolver are usually used as segments (see Bertaut et al., 2008; So and Thomas, 2010; Tan and Yen, 2011; So and Thomas, 2014). Kallberg and Saunders (1983) proposed the following states: i) inactive account, ii) ‘true revolver’ (total outstanding balance repaid within the period), iii) payment less than the total outstanding, but greater than the minimum required payment, iv) payment within \$.50 of the minimum required payment, v) payment less than the minimum required payment, but still positive, and vi) no payment. We include revolver repaid state, which is according to the definition proposed by Kallberg and Saunders (1983) the same state as ‘true revolver’. However, in our definition a revolver is a customer who does not pay total outstanding balance within the period. Revolver repaid state is used for the identification of the transition of revolver account to an inactive state because

the customer who fully repays the debt amount at the month end of the last repayment formally cannot be allocated either to transactor, or inactive, or revolver state. The implementation of this new state: i) increases the predictive accuracy of the transition probability to the transactor and revolver states; and ii) allocates the individual state for the prediction of full debt amount repayment, and as a result, gives a background for the attrition and churn scoring for revolver clients

We show Gini index values for the probability of transition to a certain state for the test sample. The probability to stay in the state and to move to the next state for 1 month / 6 months: inactive – $(0.28 - 0.37) / (0.24 - 0.81)$, transactor – $(0.22 - 0.31) / (0.47 - 0.34)$, revolver – $(0.74 - 0.87) / (0.45 - 0.77)$, repaid – $(NA - 0.53) / (0.23 - 0.38)$, delinquent 1-30 – $(0.46 - 0.60) / (0.22 - 0.55)$, delinquent 31-60 – $(0.70 - 0.65) / (0.68 - 0.57)$.

These Gini values can be used as benchmarks for further empirical investigation of the predictions of account level transition probabilities prediction for different behavioural types of credit card holders.

One of the tasks for further investigation is the achievement of higher predictive power than obtained in existing models for the transition probabilities in multistage conditional models. It is recommended to try all possible combination of the order of states for conditional binary logistic regression. Then to start from the segment with the best validation results or more important from the business point of view. Further research areas can be concentrated on relatively new models for credit cards' such as the discrete nested logit from discrete choice models to use for multistate transition probabilities modelling.

So et al. (2014) predict the profitability of credit card and use the score of being revolver and transactor and Good/Bad score. We use revolver and transactor behavioural types as an account state and use a separate model for income amount prediction for each behavioural type of credit card holder. We consider the probability of transition to the default state for the total income prediction and profitability estimation but predict the probabilities for all possible transition from any state.

We tried to fill a gap in empirical evidence of the prediction of the transition probabilities between credit card account states at account level, especially, for a full

set of income source-based states. Predictive models for risk and profit parameters can be built for a credit card portfolio at the pooled level with, for example, a Markov Chain (So and Thomas, 2011). However, significant differences between credit card usage types can decrease the predictive accuracy of such models, because the different forms of credit card usage have individual behavioural drivers for risk, utilisation, purchases, and profit (So et al., 2014; Tan and Yen, 2010). We have found only Volker (1982) used *multinomial logistic regression* for modelling of bankcards utilisation at the account level, and So and Thomas (2014) use *multinomial logistic regression* for the prediction of the transition between inactive, closed, and active account segments, and *cumulative (or ordinal) logistic regression* for the prediction the probability of transition between credit score bands. Kim, Y. and Sohn, S.Y. (2008) estimated of transition probabilities of credit ratings with using a *random effect multinomial regression* model. So, we generally filled the gap in usage of the multinomial and ordinal logistic regression, and multistate models with binary logistic regression for the prediction of credit cards states transition probabilities. Our contribution is that we proposed and extended an approach of individual transition probabilities between account states for each credit card holder depending on the individual behaviour instead of pool level probabilities of transition computed with the transition matrix.

7 Chapter Seven. Transactional Income Prediction

7.1 Introduction

A credit card generates income of different types from several sources for the issuing bank. The first type of income is interest income, generated from payments of a customer for the use of credit as an accrued interest depending on i) the interest rate and ii) the outstanding balance for a period. The second type of income is non-interest, or transactional, income. It can be generated as fees, commissions, penalties, and other payments from a variety of operations. In this chapter we investigate the fees and commissions from Point-of-Sales transactions and ATM cash withdrawals.

In this chapter, we try to predict the non-interest income with regressions based on panel data. Each account has observations at different time points. So, both cross-section and time components are considered for each account in the models. We apply several methods for panel data regression to find a technique which gives the highest goodness of fit. Also, we apply one-stage direct estimation and two-stage estimation conditional on the probability of the transaction during the performance period. This Chapter aims to find the most accurate estimation of the transactional income from credit card activities for the aggregation of the total income model.

We consider two sources of non-interest income: i) interchange fees and foreign exchange fees from transactions via point-of-sale (POS) and ii) ATM fees from cash withdrawals.

The interchange fee is a payment between banks for the acceptance of a card transactions, which is received from a merchant's bank to a customer's bank. Credit card interchange fees vary for different countries. Typically, merchants pay the interchange fee to banks.

Credit card interchange fee rates in European countries are around 1-1.5%, in the US – around 2%. (Hayashi, 2010). Also, a credit card can have fees charged annually and per transaction. Annual fees of 24 Euros are charged in 24 EU members, and usually not charged in US, Canada, and Australia. However, the reward value for per transaction fees can be 0.5%-1%.

In the current research, we consider interchange fees paid by merchants and from the foreign currency conversion rate. The fees volume are usually around 1%. Because merchant's fees are mainly generated in shops and restaurants, we call them Point-of-Sales, or POS, fees.

The ATM fees may be divided into various types depending on the category, frequency, source, and generating subject (Hayashi, Sullivan, and Weiner, 2003). In the current research, we consider total ATM fees for a period as fees from cash machine cash withdrawals. Cash withdrawal fees usually can be around 2.5%, or higher, 3% as, for example, in Lloyds Bank and RBS¹.

Cash withdrawal fees for the given data sample from the East European bank are 2.5% for the internal bank's ATM network and up to 5% for other local banks networks and foreign banks. The investigated data sample lacks information about internal and external, domestic and abroad transactions. So we do not consider ATM fees for the income amount calculation, but use given ATM income from all cash withdrawals aggregated on a monthly basis.

The problem of modelling a consumer's choice between POS and ATM depends on a set of parameters such as the cash using cost, the POS coverage etc. Humphrey, Kim and Vale (2001) estimate the consumer's demand for payment choices: checks, ATMs and POS. They use an indirect utility function to model separability between demand on POS and ATM payment instruments. In the current research, we analyse the difference between POS and ATM usage by the customer in the context of the income, which is produced by this instrument for the bank.

We perform an empirical investigation of different methods for the income amount prediction for the following segments:

- i) Models for point-of-sale (POS) and cash withdrawals (ATM) prediction

We build separate models for income from POS transactions and cash withdrawal (ATM) because of different business logic and occurrence of these two types of operations.

¹ From open sources, - <https://www.lloydsbank.com/travel/using-debit-credit-cards-abroad.asp> , <http://www.financechoices.co.uk/credit-cards/rbs-classic-credit-card/>

ii) A one-stage and a two-stage model

The one-stage model means a prediction of the non-interest income amount with a single linear regression model both for the positive income from transactions and for zero income in case of no transactions during the month. We estimate transactional income with the direct approach, which means the use of income amount as a dependent variable. A two-stage model consists of the prediction of the probability of transactions, and then the prediction of the non-interest income amount with a linear regression model for the positive income only.

iii) Different time horizons for prediction: 6 months and one month.

We build models for different periods of prediction. The six-months period is selected to be consistent with utilisation rate prediction and transition probabilities prediction as the 6-month period has been selected as the maximum possible period from accurate prediction considering data sample period and business logic.

We test five approaches for random effect estimation and pooled data and perform a comparative analysis of the following methods: pooled OLS, and four random-effect methods: Wansbeek and Kapteyn (1989), Fuller and Battese (1974), Wallace and Hussain (1969), Nerlove's Method (1971). The aim of the chapter is to determine which of these methods gives the most accurate prediction.

This chapter consists of four sections. The first section gives an overview of predictive methods for panel data and approaches to the random-effect calculation. The second section explains the model setup and estimation methods for panel regression. The third section describes the modelling results such as a comparative analysis of regression coefficients and goodness-of-fit for pooled and random-effect regression methods for POS income. The fourth section gives the modelling results in the same form for ATM income. The fifth section discusses the comparative analysis of pooled and random effect regression coefficient estimations for POS and ATM transactions income. Moreover, the conclusion highlights the main findings and contribution.

This chapter has two main findings. Firstly, we performed the empirical tests of the random-effect panel data methods with credit cards transactional income and found that Wansbeek and Kapteyn (1989) gives the best predictive accuracy from the four

variance component methods, but the pooled model predictive accuracy is even higher. Secondly, we estimated the individual models for transactional income from Point-of-Sale Interchange fees (POS) and cash withdrawals (ATM) with one-stage and two-stage models and found that two-stage model, which includes the estimation of the probability of transaction, has higher predictive accuracy than the direct estimation of the income.

We tried to fill the gap of the application of specific methods to panel data in the credit card income modelling. Baltagi et al. (2002) tested with a Monte-Carlo simulated data sample different random-effect variance component methods for panel data designed by Fuller and Battese (1974), Wansbeek and Kapteyn (1989), Wallace and Hussain (1969). However, we *first* applied these random-effect methods for credit cards panel data for the transactional income amount prediction.

7.2 Panel models

7.2.1 Panel models description

Econometrics data can be divided into two types: i) Cross-sectional which has dimensions by economic items at the same point of time without relation to time, ii) time series or observation of the economic values ranked in time. Panel data is a two-dimension array both cross-sectional and time series where cross-sectional characteristics are ranked as time series. Panel data is often called ‘longitudinal’ or ‘cross-sectional time-series’ data.

Cross-sectional and time series data are joined in different ways. In industry practice, the time component is excluded for the simplification of the regression analysis and model building and implementation. The simplest way of the transformation of panel data to cross sections is an assumption that all observations at time t_1, t_2, \dots, t_n of the same case (account) i are independent (not ranked in time). For example, 12 periods transform to 12 independent rows in the data sample. So, we transform one case which has n observations in different time to n cases, and excluded information that these observations belong to a single case (account). This type of data is called pooled cross-sectional data. For example, data slices dated monthly as the balance as of the end of

January, balance as of the end of February and so on are added to the data sample as independent observations.

For pooled data we investigate the impact of predictors of $n \times m$ observations, where n is a number of time series observations, m is a number of cases or cross-sections. We assume that observations are independent. So, each of the m cases has n identical values of the outcome.

For the panel time depending data, we investigate how predictors of m cases, which have observations at n periods, impact on $n \times m$ outcomes. We assume that observations, which belong to the same case (account), are not independent and the volatility of the values of predictors has an impact on the outcome. The outcome can be explained, for example, by an average in the time value of each predictor. The number of time series components can be equal to n for all cases (balanced panel) and equal to a different number of observations for cases (unbalanced panel).

Generally, researchers mark out the following advantages of panel data (Baltagi, B. H. (2001), Ahn, S. C. et al. (2007), Baltagi, B. H. et al. (2007)):

- i) a higher number of observations, which causes an increase in the degrees of freedom, and gives more efficient estimations,
- ii) heterogeneity of the objects is under control, because we investigate the explanatory variables within groups and between groups, and consider time variation of covariates.
- iii) testing of the effects which is impossible to identify separately in cross-sections and time series,
- iv) the decrease in multicollinearity.

It was decided to use cross-sectional data only with behavioural characteristics calculated at a point in time for the initial investigation. The main assumption was that the customer behavioural characteristics are homogeneous in time, and the number of observations is big enough to level all possible time and structure fluctuations. However, because of some changes in customer behaviour and accounts dynamics in the period 2011-2012 years it makes more sense to apply panel data model approaches to consider both cross-sectional and time-series changes.

Let's introduce the panel data model. Generally, the prediction for some outcome y_{it} at time t for the case i can be presented with the following equation:

$$y_{it} = \alpha + X'_{it}\beta + Z'_{it}\gamma + u_{it}, \quad i=1,\dots,N, \quad t=1,\dots,T \quad (7.1)$$

where

X is a vector of observed factors (or covariates);

Z is unobserved factors vector, $Z_{it} = Z_t$, so unobserved factor is considered the same for all cases and vary within time only.

So the panel model equation can rewritten as follows:

$$y_{it} = \alpha_i + X'_{it}\beta + \gamma_t + u_{it} \quad (7.2)$$

where

N is the number of observations;

T is the number of time periods;

β and γ - regression slope coefficients;

$$u_{it} = \mu_i + \lambda_t + v_{it} \quad (7.3)$$

μ_i , λ_t are non-observed individual and time effects, v_{it} – residual idiosyncratic components, or remainder disturbance.

The within-effect is the mean of the change for the average individual case and can be referred to as time-series (ANOVA) estimations. The between-effect determines whether outcomes vary on the dependent variables and can be referred to regression analysis.

A panel model can be pooled, fixed effect, and random effect depending on the consideration of variance and a time component.

A pooled model does not consider the time component and has the same form as a general linear regression model (OLS):

$$y_{it} = X'_{it}\beta + \alpha + \varepsilon_{it} \quad (7.4)$$

α и β – intercept and slope is independent of observation and time;

X_{it} - a vector of regressors (predictors).

The approach with an assumption, that each time slice for N observations is considered as an independent observation and T elements time series gives $N \times T$ observations, is widely applied as an industry standard in banks for credit risk modelling. For instance,

it is used to create development and validation samples from the data set with not enough observations at the certain point in time or to consider different seasons.

The use of a pooled panel data approach requires the next assumption:

- ✓ dependence between factors is stable in time;
- ✓ correlation between observations is not considering (it is assumed that all observations are independent).

However, in practice, these conditions often are not satisfied. Thus, generally the time component can be used as part of intercept or variance component in the fixed and random effects.

A fixed effect model can be written as follow:

$$y_{it} = (\alpha + u_i) + X'_{it}\beta + v_{it} \quad (7.5)$$

where

u_i – an individual intercept (group or specific effect) is time-invariant and is a part of intercept. Intercept is varying across groups and times. An individual part of the intercept u_i can correlate to other regressors;

X_{it} is a vector of characteristics for i observation at time t ;

β is a vector of regression coefficients;

v_{it} is an error variance and is a constant.

The problem of fixed effect method is the exclusion of time-constant effects. A fixed effect model considers only within variation, but does not use between-unit changes. So, fixed effects methods can be inefficient, because they throw out information.

A random effect model can be written as follows:

$$y_{it} = \alpha + X'_{it}\beta + (u_i + v_{it}) \quad (7.6)$$

where $(u_i + v_{it})$ is a random effect. u_i is a part or error and should not correlate to regressors. The intercept α in the random effect model is constant, and error variance v_{it} is varying across groups and/or times. The error is independent and identically-distributed random variable: $v_{it} \sim IID(0, \sigma_v^2)$.

The coefficients of a random-effect regression can be estimated with the generalised least squares method, as, for example, the groupwise heteroscedastic regression model (Greene, 2003). Baltagi and Cheng (1994) applied various estimation methods for feasible generalised least squares to exploit within correlation, which is used in the case when the variance structure is not known.

The random effect model can be used in case the variance across entities has some impact on the dependent variable.

The estimator of the coefficients for a random effect model is a product of vectors of covariates, correlation matrix and outcome as follows:

$$\hat{\beta}_{RE} = [\sum_{i=1}^N (X_i' \Omega^{-1} X_i)]^{-1} \sum_{i=1}^N (X_i' \Omega^{-1} y_i) \quad (7.7)$$

where

Ω is the correlation matrix between individual variances of individual for each period,

X is the vector of covariates for case i ,

N is a number of cases,

y_i is a vector of outcomes for case i .

We select random effect because we need to estimate the population mean and variance instead of means of the individual factor levels. We investigate how behavioural characteristics impacts on the transactional income amount. But we do not investigate income amount for each group which can be used for fixed effect models.

In case either random effect or fixed effect cannot be used pooled regression can be applied. This means that each observation in a panel data set is considered as independent and any explanation of the impact of variance between cases or within cases on the dependent variable is avoided. However, pooled OLS does not consider the autocorrelation in the composite error term, and so it can be inefficient.

Hausman test

The application of a fixed or random effect method to panel data can be tested statistically with a Hausman test.

The error for i case at time t for efficient estimators can be explained as the sum of variances for the u and v components of error for diagonal elements of matrix $T \times T$ for generalised least squares estimator as follows:

$$E(\varepsilon_{it}^2) = \sigma_u^2 + \sigma_v^2, t = s,$$

and for elements out of diagonal of the variance structure matrix as follows:

$$E(\varepsilon_{it}\varepsilon_{is}) = \sigma_u^2, t \neq s.$$

The assumption, which the expectation of the product of the covariates for all cases at each period and the random effect component u_i as the part of the composite error ε_{it} should be equal to zero, is required for consistency of the random-effect model, and can be written as follows:

$$E(x_{kit}u_i) = 0 \text{ for all } k, i, t$$

However, it is not required for the fixed-effect models. Hausman (1978) test is used to test this assumption. The Hausman test is applied to test independence of the error component u_i and covariates x_{kit} . The null hypothesis is that there is no correlation and, consequently, no difference between fixed effect and random effect estimators $\hat{\beta}_{FE} - \hat{\beta}_{RE}$.

The Hausman test formula can be written as follows:

$$H = (\beta_{RE} - \beta_{FE})' (Var(\beta_{RE}) - Var(\beta_{FE}))^{-1} (\beta_{RE} - \beta_{FE}) \sim \chi^2(k), \quad (7.8)$$

where $(Var(\beta_{RE}) - Var(\beta_{FE}))^{-1}$ is a covariance matrix of estimators with inverse.

In the case Hausman test is significant, the random effect estimator should not be used. The null hypothesis H_0 , is that the individual effects u_i are uncorrelated with the other regressors in the model. If this is rejected (or the Hausman test cannot be computed for the set of covariates), that we should use the fixed-effect model rather than a random-effect model.

7.2.2 R-Squared for Panel data

The coefficient of determination R-square shows the proportion of variance of the dependent variable which is explained by predictors and used as a measure of Goodness-of-Fit of the linear model.

The standard R-square is defined as

$$R^2 = \frac{SS_{reg}}{SS_{tot}} = 1 - \frac{SS_{res}}{SS_{tot}} \quad (7.9)$$

where

$SS_{tot} = \sum_i (y_i - \bar{y})^2$ is the total sum of squares,

$SS_{reg} = \sum_i (f_i - \bar{y})^2$ is the explained sum of squares.

$SS_{res} = \sum_i (y_i - f_i)^2$ is the residual sum of squares.

The standard coefficient of determination R-square measure is inapplicable for the assessment of the Goodness-of-Fit of linear models based on panel data because a number out of the [0,1] range might be obtained.

‘The total variation in y_{it} can be written as the sum of the within variation and the between variation’ (Verbeek, 2004):

$$\frac{1}{NT} \sum_{i,t} (y_{it} - \bar{y}_i)^2 = \frac{1}{NT} \sum_{i,t} (y_{it} - \bar{y}_i)^2 + \frac{1}{N} \sum_i (\bar{y}_i - \bar{y})^2 \quad (7.10)$$

The within R-square for a fixed-effect model, which is used for the within variation explanation, is a squared correlation coefficient between $(x_{it} - \bar{x}_i)'\hat{\beta}_{FE}$ and difference of observed and average outcome:

$$R_{within}^2(\hat{\beta}_{FE}) = \text{corr}^2\{\hat{y}_{it}^{FE} - \hat{y}_i^{FE}, y_{it} - \bar{y}_i\}, \quad (7.11)$$

where FE means fixed effect and \hat{y}_{it}^{FE} is an estimation for a fixed-effect model.

The between R-square for the estimator of OLS of individual means is a squared correlation coefficient between $\hat{y}_i^B = \bar{x}_i\beta_B$ and average outcome:

$$R_{between}^2(\hat{\beta}_B) = \text{corr}^2\{\hat{y}_i^B, \bar{y}_i\} \quad (7.12)$$

So the overall R-square can be defined as the squared correlation coefficient

$$R_{overall}^2(\hat{\beta}) = \text{corr}^2\{\hat{y}_{it}, y_{it}\} \quad (7.13)$$

The current research contains a comparative analysis of pooled data OLS and four variance component methods of regression analysis for random-effect models:

- i) Wallace and Hussain
- ii) Wansbeek and Kapteyn
- iii) Fuller and Battese
- iv) Nerlove’s.

The first three methods were selected by Baltagi and Chang (1994) and Nerlove's method is a simple method selected for the comparison with others.

7.2.3 Methods for Random-Effects Models Estimation

The random-effect methods are used for the panel data regression analysis in cases of unobserved heterogeneity and when the individual specific effects are unobserved.

Initially, we introduce the one-way random-effects as

$$u_{it} = \nu_i + \varepsilon_{it} \quad (7.14)$$

where ν_i is unobserved over periods, but constant over time, ε_{it} is a time-varying idiosyncratic error.

A one-way random-effect model is estimated in two stages. Firstly, the variance σ_ε^2 and σ_ν^2 is calculated with one of the methods, described below. Then the variance estimations are used for the weighting factor θ_i calculation. The model estimation is obtained with the Ordinary Least Squares regression method, built on the partial deviations from the means, which are calculated for account i .

The following notations for vectors are used in the random-effect methods description.

\mathbf{j}_{T_i} is a vector of ones with dimension T_i . \mathbf{J}_{T_i} is a square matrix of ones with dimension T_i .

$\mathbf{Z}_0 = \text{diag}(\mathbf{j}_{T_i})$, $\mathbf{P}_0 = \text{diag}(\bar{\mathbf{J}}_{T_i})$, and $\mathbf{Q}_0 = \text{diag}(\mathbf{E}_{T_i})$,

with $\bar{\mathbf{J}}_{T_i} = \mathbf{J}_{T_i}/T_i$ and $\mathbf{E}_{T_i} = \mathbf{I}_{T_i} - \bar{\mathbf{J}}_{T_i}$.

The following transformation is used for the estimations: $\tilde{\mathbf{X}}_s = \mathbf{Q}_0 \mathbf{X}_s$ and $\tilde{\mathbf{y}} = \mathbf{Q}_0 \mathbf{y}$.

The method, used for the random-effect estimation, is an ANOVA method. The estimates are calculated as quadratic sums of squares to their expectations. For the unbalanced panel, the variance components are estimated as ‘a function of the variance components themselves’ (Townsend and Searle, 1971).

Baltagi and Chang (1994) have described several methods for the estimation of random-effects models. Such methods as Fuller and Battese (1974) method, Wansbeek and Kapteyn (1989) method, and Wallace and Hussain (1969) method demonstrated good performance. Nerlove (1971) method is selected as a relatively simple approach

for random-effect estimation. The difference between the methods is in the approaches to variance components estimation.

7.2.3.1 Fuller and Battese Method

The error and cross-section variance component in the method, proposed by Fuller and Battese (1974), is also called the ‘fitting of constants’ method. The variance is estimated as follows:

$$\hat{\sigma}_\epsilon^2 = \left\{ \mathbf{y}' \mathbf{y} - R(\beta|\nu) - R(\nu) \right\} / \{M - N - (K - 1)\} \text{ for the error variance,}$$

$$\hat{\sigma}_v^2 = \left\{ R(\nu|\beta) - (N - 1) \hat{\sigma}_\epsilon^2 \right\} / \left\{ M - \text{tr}(\mathbf{Z}_0' \mathbf{X} (\mathbf{X}' \mathbf{X})^{-1} \mathbf{X}' \mathbf{Z}_0) \right\} \text{ for cross-section variance,}$$

where

\mathbf{X} is a vector of covariates

$$R(\nu) = \mathbf{y}' \mathbf{Z}_0 (\mathbf{Z}_0' \mathbf{Z}_0)^{-1} \mathbf{Z}_0' \mathbf{y}$$

$$R(\beta|\nu) = \tilde{\mathbf{y}}' \tilde{\mathbf{X}}_s' (\tilde{\mathbf{X}}_s' \tilde{\mathbf{X}}_s)^{-1} \tilde{\mathbf{X}}_s' \tilde{\mathbf{y}}$$

$$R(\beta) = \mathbf{y}' \mathbf{X}' (\mathbf{X}' \mathbf{X})^{-1} \mathbf{X}' \mathbf{y}$$

$$R(\nu|\beta) = R(\beta|\nu) + R(\nu) - R(\beta)$$

The error variance estimation is equal to the residual variance, obtained from the within estimation.

First, the variance residual sums of squares is calculated, and then - the residual sums of squares of the regression. The Fuller and Battese method can be used for both balanced and unbalanced data.

7.2.3.2 Wansbeek and Kapteyn Method

We use a specialisation (Baltagi and Chang, 1994) of the method, which is developed by Wansbeek and Kapteyn (1989) for unbalanced two-way models. This approach uses a quadratic unbiased estimation (QUE) method for the variance estimation.

Let us introduce the residuals $\tilde{\mathbf{u}} = (\mathbf{I}_M - \mathbf{J}_M) \{ \mathbf{y} - \mathbf{X}_s (\tilde{\mathbf{X}}_s' \tilde{\mathbf{X}}_s)^{-1} \tilde{\mathbf{X}}_s' \tilde{\mathbf{y}} \}$ and then two variables:

$$q_1 = \tilde{\mathbf{u}}' \mathbf{Q}_0 \tilde{\mathbf{u}}$$

$$q_2 = \tilde{\mathbf{u}}' \mathbf{P}_0 \tilde{\mathbf{u}}$$

The expected values of the $q1$ and $q2$ are defined as follows:

$$E(q_1) = (M - N - (K - 1))\sigma_\epsilon^2$$

$$\begin{aligned} E(q_2) &= (N - 1 + \text{tr}[(\mathbf{X}'_s \mathbf{Q}_0 \mathbf{X}_s)^{-1} \mathbf{X}'_s \mathbf{P}_0 \mathbf{X}_s] - \text{tr}[(\mathbf{X}'_s \mathbf{Q}_0 \mathbf{X}_s)^{-1} \mathbf{X}'_s \bar{\mathbf{J}}_M \mathbf{X}_s])\sigma_\epsilon^2 + \\ &+ [M - (\sum_i T_i^2 / M)]\sigma_v^2 \end{aligned}$$

The error variance $\hat{\sigma}_\epsilon^2$ and cross-sectional variance $\hat{\sigma}_v^2$ are computed as a result of the equation, where quadratic forms are equating to the expected values $E(qi)$. The error variance estimation is equal to the residual variance, obtained from the within estimation.

7.2.3.3 Wallace and Hussain Method

Wallace and Hussain (1969) estimates use OLS residuals for the variance components estimation. Baltagi and Chang (1991) modified this method for unbalanced one-way models with the assumption of groupwise heteroscedasticity.

Wallace and Hussain method started with the same definition of q_i , as Wansbeek and Kapteyn method:

$$q_1 = \tilde{\mathbf{u}}'_{OLS} \mathbf{Q}_0 \tilde{\mathbf{u}}_{OLS}$$

$$q_2 = \tilde{\mathbf{u}}'_{OLS} \mathbf{P}_0 \tilde{\mathbf{u}}_{OLS}$$

However, the observed errors from the Wansbeek and Kapteyn method are replaced with residuals, obtained from OLS estimation. The error variance $\hat{\sigma}_\epsilon^2$ and cross-sectional $\hat{\sigma}_v^2$ variance values are computed as a solution of the following system of equations:

$$E(\hat{q}_1) = E(\hat{\mathbf{u}}'_{OLS} \mathbf{Q}_0 \hat{\mathbf{u}}_{OLS}) = \delta_{11} \hat{\sigma}_v^2 + \delta_{12} \hat{\sigma}_\epsilon^2$$

$$E(\hat{q}_2) = E(\hat{\mathbf{u}}'_{OLS} \mathbf{P}_0 \hat{\mathbf{u}}_{OLS}) = \delta_{21} \hat{\sigma}_v^2 + \delta_{22} \hat{\sigma}_\epsilon^2$$

Where $\delta_{11}, \delta_{12}, \delta_{21}, \delta_{22}$ are constants, defined as follow:

$$\delta_{11} = \text{tr} \left((\mathbf{X}' \mathbf{X})^{-1} \mathbf{X}' \mathbf{Z}_0 \mathbf{Z}'_0 \mathbf{X} \right) - \text{tr} \left((\mathbf{X}' \mathbf{X})^{-1} \mathbf{X}' \mathbf{P}_0 \mathbf{X} (\mathbf{X}' \mathbf{X})^{-1} \mathbf{X}' \mathbf{Z}_0 \mathbf{Z}'_0 \mathbf{X} \right)$$

$$\delta_{12} = M - N - K + \text{tr} \left((\mathbf{X}' \mathbf{X})^{-1} \mathbf{X}' \mathbf{P}_0 \mathbf{X} \right)$$

$$\delta_{21} = M - 2 \text{tr} \left((\mathbf{X}' \mathbf{X})^{-1} \mathbf{X}' \mathbf{Z}_0 \mathbf{Z}'_0 \mathbf{X} \right) + \text{tr} \left((\mathbf{X}' \mathbf{X})^{-1} \mathbf{X}' \mathbf{P}_0 \mathbf{X} \right)$$

$$\delta_{22} = N - \text{tr} \left((\mathbf{X}' \mathbf{X})^{-1} \mathbf{X}' \mathbf{P}_0 \mathbf{X} \right)$$

where $\text{tr}()$ is the trace operator or the sum of the elements on the main diagonal of a square matrix.

The Wallace and Hussain method may give negative values, which are replaced with zero.

7.2.3.4 Nerlove's Method

Nerlove's method (Nerlove, 1971; Baltagi, 1995,) presents a relatively simple approach to the variance components estimation in comparison with previous methods. Nerlove's method gives positive variance components.

Nerlove's method estimates cross-sectional variance $\hat{\sigma}_v^2$ as the variance of the fixed effects, where γ_i is the i th fixed effect. The cross-sectional variance is calculated as follows:

$$\hat{\sigma}_v^2 = \sum_{i=1}^N \frac{(\gamma_i - \bar{\gamma})^2}{N-1},$$

where $\bar{\gamma}$ is the mean fixed effect.

The error variance $\hat{\sigma}_\epsilon^2$ formula uses the residual of the one-way fixed-effects regression. The number of observations divides into the sum of squares of residuals.

The next step is the estimation of the regression model. For each account i a weight (θ_i) is defined as follows:

$$\theta_i = 1 - \sigma_\epsilon / w_i$$

$$w_i^2 = T_i \sigma_v^2 + \sigma_\epsilon^2$$

where T_i is the cross section's time observations for i .

The partial deviations are calculated with a weight θ_i as follows:

$$\tilde{y}_{it} = y_{it} - \theta_i \bar{y}_i$$

$$\tilde{x}_{it} = x_{it} - \theta_i \bar{x}_i$$

where \bar{y}_i and \bar{x}_i are cross section means of the outcome and explanatory variables.

The result of the previous steps is the random effects β , which is calculated as a simple OLS on the transformed data.

We use in the current research SAS® Software for the model parameters estimation. For panel data regression running both for pooled and random-effects we use the SAS/ETS® package. Four random-effect methods are implemented in this software as procedures and we use and compare the results generated by the SAS® after running PROC PANEL.

There are not preferences for any method. So we need to test all of them with the given data sample to select which method can show higher predictive accuracy.

7.3 Transactional income modelling conception

Credit card income is generated from two sources: interest rate payments and payment to the bank from transactions or non-interest sources of income.

The interest income can be calculated by using the utilisation rate as follows:

- i) $Utilization = Balance / Credit\ Limit$
- ii) $IR_Income = Utilization\ Rate \times Limit \times IR$

The total Interest Income is equal to the product of the average utilisation rate for period T ($Avg\ Ut(T)$), interest rate (IR), and the average credit limit for period T ($Limit(T)$):

$$Avg\ UT(T) \times IR \times Limit(T).$$

For non-interest rate (or transactional) income, we apply a different concept. Generally, transactional income from POS and ATM activities are calculated (in a simple form) as

- i) $POS\ Income = TR\ Debit_POS \times POS_fees_rate$
- ii) $Cash\ Withdrawal\ Income = TR_Debit_ATM \times ATM_fees_rate,$

where

TR_Debit_POS is the sum of spending transactions (or purchases) via Point-of-Sales for the period,

TR_Debit_ATM is the sum of ATM cash withdrawals for the period,

POS_fees_rate is applied interchange fees rate for POS transactions,

ATM_fees_rate is applied rate for ATM cash withdrawals.

However, we use a direct estimation of the income amount for both transactional income from POS and ATM. We do not predict the sum of transactions and multiply it by fees rate because it complicates the model as different rates are applied for different products and at a different point of time.

7.3.1 Direct estimation

The first approach is the direct estimation of the amount of income generated by a transaction of the account i at time t . The predictors used for the prediction of transaction income of an account are behavioural, customer application, and macroeconomic characteristics. We use a linear regression (OLS) model for pooled data and four linear regressions for random-effect models.

$$I_{1POS\,it} = \sum_{k=1}^K \beta_{POS\,k} \cdot \mathbf{B}_{bi,t-1} + \sum_{l=1}^L \alpha_{POS\,a} \cdot \mathbf{A}_{ai} + \sum_m^M \gamma_{POS\,m} \mathbf{M}_{m,t-1} \quad (7.15)$$

$$I_{1ATM\,it} = \sum_{k=1}^K \beta_{ATM\,k} \cdot \mathbf{B}_{bi,t-1} + \sum_{l=1}^L \alpha_{ATM\,a} \cdot \mathbf{A}_{ai} + \sum_m^M \gamma_{ATM\,m} \mathbf{M}_{m,t-1} \quad (7.16)$$

α, β, γ – regression coefficients (slopes)

\mathbf{B} – vector of behavioural variables (for example, average balance to maximum balance, maximum debit turnover to average outstanding balance or limit)

\mathbf{A} – vector of application variables - client's demographic, financial and product characteristics

\mathbf{M} – vector of macroeconomic factors (GDP, FX, Unemployment rate changes, etc.)

We predict directly the amount of income after the transaction. Generally, it is possible to predict a transaction amount and then calculate an income. For an ATM withdrawal, an income amount depends on the cash withdrawal fees. However, depending on a

number of cash machine factors such as domestic or foreign withdrawal, issuer bank ATM network, other banks, the fees and commissions may vary.

For POS transactions, an income amount depends on interchange fees. However, depending on the type of transaction, merchant and acquiring terms, a credit card rates, which are specific for certain product, a number of different fees may be applied. Thus, it becomes a complicated task to calculate the amount of income after the transaction, and we do not have the required information in our database to split transactions and apply appropriate fees and commissions. Thus, we use for prediction the generated income amount and do not consider how exactly this income has been accrued.

7.3.2 Two-stage model – indirect estimation

Two-stage model means income estimation conditional on the probability of a transaction. At the model development stage for the income amount prediction, we use data sample with transactions only and exclude zero income. For the income prediction, the probability of transition function can be used as a weight (the probability of transition is multiplied by the estimated income amount) or as a switch function with the threshold for the probability.

1st stage – the estimation of the probability that the client will use credit cards for POS/ATM transaction during the performance period

$$\begin{aligned} \ln\left(\frac{\Pr(TR_{POS,it} > 0)}{1 - \Pr(TR_{POS,it} > 0)}\right) = \\ = \sum_{k=1}^K \beta_{POS_LOG_k} \cdot \mathbf{B}_{ki,t-1} + \sum_{l=1}^L \alpha_{POS_LOG_a} \cdot \mathbf{A}_{ai} + \sum_{m=1}^M \gamma_{POS_LOG_m} \mathbf{M}_{m,t-lag} \end{aligned} \quad (7.17)$$

$$\begin{aligned} \ln\left(\frac{\Pr(TR_{ATM,it} > 0)}{1 - \Pr(TR_{ATM,it} > 0)}\right) = \\ = \sum_{k=1}^K \beta_{ATM_LOG_k} \cdot \mathbf{B}_{ki,t-1} + \sum_{l=1}^L \alpha_{ATM_LOG_a} \cdot \mathbf{A}_{ai} + \sum_{m=1}^M \gamma_{ATM_LOG_m} \mathbf{M}_{m,t-lag} \end{aligned} \quad (7.18)$$

where $\Pr(TR_{POS,it})$ and $\Pr(TR_{ATM,it})$ is the probability of a POS and ATM transaction for account i at time t ;

β_{POS_LOG} , α_{POS_LOG} , and γ_{POS_LOG} are coefficient estimates for logistic regression (in contrast to linear regression coefficients).

We estimate the above model using logistic regression. We denote the absence of a transaction during a performance period as 0, otherwise 1.

2nd stage – income amount for the period

$$\begin{aligned} I_{POS_{it}}(\mathbf{B}, \mathbf{A}, \mathbf{M} | \Pr(TR_{POS_{it}}) > 0) = \\ = \sum_{k=1}^K \beta_{POS_k} \cdot B_{bi,t-1} + \sum_{l=1}^L \alpha_{POS_a} \cdot A_{ai} + \sum_m \gamma_{POS_m} M_{m,t-lag} \end{aligned} \quad (7.19)$$

$$\begin{aligned} I_{ATM_{it}}(\mathbf{B}, \mathbf{A}, \mathbf{M} | \Pr(TR_{ATM_{it}}) > 0) = \\ = \sum_{k=1}^K \beta_{ATM_k} \cdot B_{bi,t-1} + \sum_{l=1}^L \alpha_{ATM_a} \cdot A_{ai} + \sum_m \gamma_{ATM_m} M_{m,t-1} \end{aligned} \quad (7.20)$$

Expected income is equal to the product of two functions: i) the probability that customer will use cards for a transaction such as point-of-sales transaction, ATM cash withdrawal) and ii) the estimation of the income amount from this transaction.

$$\begin{aligned} \Pr(TR_{POS_{it}} > 0) = \\ = \frac{1}{1 + \exp\left(-\left(\sum_{k=1}^K \beta_{POS_LOG_k} \cdot B_{ki,t-1} + \sum_{l=1}^L \alpha_{POS_LOG_a} \cdot A_{ai} + \sum_m \gamma_{POS_LOG_m} M_{m,t-lag}\right)\right)} \end{aligned} \quad (7.21)$$

$$\begin{aligned} I_{2POS_{it}} = \Pr(TR_{POS_{it}} > 0 | \mathbf{B}, \mathbf{A}, \mathbf{M}) \cdot I_{POS_{it}}(\mathbf{B}, \mathbf{A}, \mathbf{M} | \Pr(TR_{POS_{it}}) > 0) + \\ + (1 - \Pr(TR_{POS_{it}} = 0 | \mathbf{B}, \mathbf{A}, \mathbf{M}) \cdot 0) \end{aligned} \quad (7.22)$$

7.3.3 Income as a proportion of the credit limit – indirect estimation

The sum of income for the period can be estimated indirectly as a part of the credit limit and as part of the outstanding balance. In the first case, the credit limit is equal to its initial value and is a constant for all period. If it has been changed during the observation or performance period, the new credit limit value is used from the limit change point. The Income from POS for an account i at time t ($IncPOS_{it}$) depends on the credit limit and the result of the estimation of the ratio of the POS income at the performance period to the credit limit at the observation point ($POS_LimitRate_{it}$).

$$\frac{I_{POS_{it}}}{Limit} = \sum_{k=1}^K \beta_k \cdot B_{bi,t-1} + \sum_{l=1}^L \alpha_a \cdot A_{ai} + \sum_m^M \gamma_m M_{m,t-1} \quad (7.23)$$

$$POS_{it} = Limit \cdot POSLimitRate_{it} = \\ = Limit \cdot \left(\sum_{k=1}^K \beta_k \cdot B_{bi,t-1} + \sum_{l=1}^L \alpha_a \cdot A_{ai} + \sum_m^M \gamma_m M_{m,t-1} \right) \quad (7.24)$$

It is expected that the outstanding balance denominator is changed during the observation period and at the observation point for the same account, it will be different at the different time. The Income from POS for an account I at time t (IncPOS_{it}) depends on the outstanding balance at t-1 OB_{t-1} and the result of the estimation of the ratio of the POS income at the performance period to the outstanding at the observation point (POS_OBRate_{it}).

$$\frac{IncPOS_{it}}{OB_{t-1}} = \sum_{k=1}^K \beta_k \cdot B_{bi,t-1} + \sum_{l=1}^L \alpha_a \cdot A_{ai} + \sum_m^M \gamma_m M_{m,t-1} \quad (7.25)$$

$$IncPOS_{it} = OB_{t-1} \cdot POS_OBRate_{it} = \\ = OB_{t-1} \cdot \left(\sum_{k=1}^K \beta_k \cdot B_{bi,t-1} + \sum_{l=1}^L \alpha_a \cdot A_{ai} + \sum_m^M \gamma_m M_{m,t-1} \right) \quad (7.26)$$

7.3.4 Data description and covariates selection

In our current research, we use panel data. All characteristics such as the outstanding balance, delinquency state, account state, spending transactions amount, number of transactions, payment amount, and all aggregated computed characteristics like average values for a period and ratios of indicators, are presented as a time-series. Each variable has a history of values for several periods. Panel data can be balanced, where each character has an observation for each period, or unbalanced, where some observations can be missed for some periods or periods have different number of observations. We use unbalanced data in the sense that i) observations are not at the same time points, and ii) a different number of periods may exist for each account. However, i) all observations are consecutive, so there are no gaps in time series, and ii) there are at least 12 observations for each account, so we have one year or a longer period for investigation. A detailed description of the data sample is given in Chapter Three.

We have selected the following covariates because of our expectation on how these covariates impact on POS and ATM income (Table 7.1).

Table 7.1 Selected covariates and expectations for the impact on transactional income

Variable	POS income expectations	ATM income expectations
Mob	The higher MOB might decrease credit card usage and POS transactions amount	The higher MOB might decrease credit card usage, but customers might use a credit card for cash withdrawals
Limit	Higher credit limit might slightly increase the credit card usage and POS transactions amount	Higher credit limit might slightly increase the credit card usage and cash withdrawals amount
UT0_	High utilisation is correlated to POS transactions amount	High utilisation is correlated to ATM transactions amount
b_UT1to2In	Increase in the utilisation rate is correlated to POS transaction amount	Increase in the utilisation rate is correlated to ATM transaction amount
b_UT1to6In	Increase in the utilisation rate is correlated to POS transaction amount	Increase in the utilisation rate is correlated to ATM transaction amount
avg_balance_1	Higher outstanding balance might slightly increase the credit card usage and POS transactions amount	Higher outstanding balance might slightly increase the credit card usage and POS transactions amount
avg_deb_amt_1	Higher debit transactions amount (spending) might increase POS income	Higher debit transactions amount (spending) might increase ATM income
sum_crd_amt_1	Higher credit transactions (payments) might decrease POS income because if a customer pays back more he/she probably spends less amount	Higher credit transactions (payments) might decrease ATM income because if a customer pays back more he/she probably spends less amount
sum_deb_amt_1	Higher debit transactions amount (spending) might increase POS income	Higher debit transactions amount (spending) might increase ATM income
max_deb_amt_1	Higher maximum debit transaction (spending) might decrease POS income because this might mean a low number of transactions	Higher maximum debit transactions (spending) might decrease ATM income because this might mean a low number of transactions
b_AvgOB1_to_MaxOB1_In	In case the average balance is close to maximum balance, a customer probably does not make significant spending transactions. So for a higher ratio of average balance to maximum balance POS income might decrease	In case the average balance is close to maximum balance, a customer probably does not make significant spending transactions. So for a higher ratio of average balance to maximum balance ATM income might decrease
b_TRmax_deb1_To_Limit	Higher debit transactions (spending) might increase POS income	Higher debit transactions (spending) might increase ATM income
b_TRavg_deb1_to_avgO	A higher ratio of average debit transactions (spending) to average balance might increase POS income	A higher ratio of average debit transactions (spending) to average balance might increase ATM income
b_TRsum_deb1_to_TRsumcrd	In case a customer spends more than pay back POS income might increase	In case a customer spends more than pay back ATM income might increase

Table 7.1 Selected covariates and expectations for the impact on transactional income

Variable	POS income expectations	ATM income expectations
b_avgNumDeb13	Higher number of debit transactions (spending) might increase POS income	Higher number of debit transactions (spending) might increase ATM income
b_OB13_to_OB46ln	In case the outstanding balance for the last three months is higher than balance for 4-6 months, the purchase amount increases. So if balance Increase continues, POS income will also increase	In case the outstanding balance for the last three months is higher than balance for 4-6 months, the purchase amount increases. So if balance Increase continues, ATM income will also increase
b_OB1_to_OB2_ln	In case the outstanding balance for the last month is higher than balance for the previous month, the purchase amount increases. So if balance Increase continues, POS income will also increase	In case the outstanding balance for the last month is higher than balance for the previous month, the purchase amount increases. So if balance Increase continues, ATM income will also increase
b_pos_flag_use13vs46	If customer start to use a credit card for POS transactions, it is expected that POS income will increase	the effect is not explicit
b_atm_flag_use13vs46	The effect needs to be tested	The effect needs to be tested
b_pos_use_only_flag	If customer use a credit card for POS transactions only, it is expected that POS income will increase	If customer use a credit card for POS transactions only, it is expected that ATM income will decrease
b_atm_use_only_flag	If customer use a credit card for ATM transactions only, it is expected that POS income will decrease	If customer start to use a credit card for ATM transactions only, it is expected that ATM income will increase
b_TRsum_crd1_to_OB1_ln	It is expected that if customer pays back more than he/she will spend a higher amount in the next period. So higher sum of a credit transaction to the outstanding balance might increase POS income amount	It is expected that if customer pays back more than he/she will spend a higher amount in the next period. So higher sum of a credit transaction to the outstanding balance might increase ATM income amount
b_payment_lt_5p_1	If customer pay back less than the minimum required payment, an account would go to delinquency, blocked and a customer will not be able to make transactions. So flag of payment less than 5% decrease expected POS income amount	If customer pay back less than the minimum required payment, an account would go to delinquency, blocked and a customer will not be able to make transactions. So flag of payment less than 5% decrease expected ATM income amount
b_maxminOB_avgOB_1_ln	A higher ratio of the difference between the minimum and maximum outstanding balance to average outstanding balance mean higher volatility of balance, which can be caused by spending transactions, and the higher POS income might be expected in the next period	A higher ratio of the difference between the minimum and maximum outstanding balance to average outstanding balance mean higher volatility of balance, which can be caused by spending transactions, and the higher ATM income might be expected in the next period
b_TRsum_deb1_to_2_ln	If an increase of debit transaction amount in the last month in comparison to the previous month	If an increase of debit transaction amount in the last month in comparison to the previous month continues, the

Table 7.1 Selected covariates and expectations for the impact on transactional income

Variable	POS income expectations	ATM income expectations
b_TRsum_crd1_to_2_ln	continues, the POS income in the next months might increase If an increase of credit transaction amount in the last month in comparison to the previous month continues, the POS income in the next months might decrease because a customer pays back more and probably spends less amount	ATM income in the next months might increase If an increase of credit transaction amount in the last month in comparison to the previous month continues, the ATM income in the next months might decrease because a customer pays back more and probably spends less amount
I_ch1_ln	An increase in the credit limit in the last month might have a slight impact on POS income in the next months	An increase of the credit limit in the last month might have a positive impact on ATM income in the next months, because customers with low financial literacy might have a signal to withdraw additional money after credit limit increase
I_ch1_flag	An increase in the credit limit in the last month might have a slight impact on POS income in the next months	An increase of the credit limit in the last month might have a positive impact on ATM income in the next months, because customers with low financial literacy might have a signal to withdraw additional money after credit limit increase
I_ch6_flag	An increase in the credit limit in the last 6 months might have positive impact on POS income in the next months, because customers can spend more money	An increase in the credit limit in the last 6 months might have positive impact on ATM income in the next months, because customers can spend more money
age	Older customers probably make fewer transactions than younger cardholders, so POS income for older cardholders might be less than for young customers	Older customers probably make fewer transactions than younger cardholders, so ATM income for older cardholders might be less than for young customers. However, older people might have a tendency to withdraw cash from a credit card
customer_income_ln	A customer with high income might have high credit limits and high spending. So we expect the positive correlation between customer's income and POS income	A customer with high income might have high credit limits and high spending. So we expect the positive correlation between customer's income and ATM income
Edu_	Customers with higher education have higher credit limits and, consequently, might spend more and generate higher POS income than customers with secondary or special education	Customers with higher education have higher credit limits and, consequently, might spend more and generate higher ATM income than customers with secondary or special education. However, customers with secondary or special education might make more cash withdrawals and generate higher ATM income
Marital_	We expect that married customers might have higher usage of credit card and generate higher POS	We expect that married customers might have higher usage of credit card

Table 7.1 Selected covariates and expectations for the impact on transactional income

Variable	POS income expectations	ATM income expectations
position_	income than other categories of customers Customers with Top manager and manager positions have higher credit limits and, consequently, might spend more and generate higher POS income than customers with other positions	and generate higher ATM income than other categories of customers Customers with Top manager and manager positions have higher credit limits and, consequently, might spend more and generate higher ATM income than customers with other positions. However, customers with Technical staff position might make more cash withdrawals and generate higher ATM income
sec_	Customers from financial and energy sectors have higher credit limits and, consequently, might spend more and generate higher POS income than customers from other sectors	Customers from financial and energy sectors have higher credit limits and, consequently, might spend more and generate higher ATM income than customers from other sectors. However, customers from agriculture and construction sector might make more cash withdrawals and generate higher ATM income
car_Own	Customers with a car might have more spending and generate more POS income than customers without a car	Customers with a car might have more spending and generate more ATM income than customers without a car
real_Own	Customers who rent a flat might have more spending and generate more POS income than customers with real estate	Customers who rent a flat might have more spending and generate more ATM income than customers with real estate
reg_ctr_Y	Customers who live in capital and region centres might generate more POS income	Customers who live in capital and region centres might generate less ATM income from cash withdrawals than customers from the province
child_	Customers with children might have more spending and generate more POS income than customers without children	Customers with children might have more spending and generate more ATM income than customers without children
Unempl_Inoy	A higher level of the unemployment rate may cause the increase in demand for money, but decrease the appetite for spending. So we expect the negative correlation between the unemployment rate and POS income	A higher level of the unemployment rate may cause the increase in demand for money, but decrease the appetite for spending. So we expect the negative correlation between the unemployment rate and ATM income
UAH_EURRate_Inmom	The increase in local currency to Euro exchange rate (LCY/EUR exchange rate) might cause the increase in demand for money, but decrease the appetite for spending. So we expect the negative correlation between the unemployment rate and POS income	The increase in local currency to Euro exchange rate (LCY/EUR exchange rate) might cause the increase in demand for money, but decrease the appetite for spending. So we expect the negative correlation between the unemployment rate and ATM income

Table 7.1 Selected covariates and expectations for the impact on transactional income

Variable	POS income expectations	ATM income expectations
UAH_EURRate_Inoy	The increase in local currency to Euro exchange rate (LCY/EUR exchange rate) might cause the increase in demand for money, but decrease the appetite for spending. So we expect the negative correlation between the unemployment rate and POS income	The increase in local currency to Euro exchange rate (LCY/EUR exchange rate) might cause the increase in demand for money, but decrease the appetite for spending. So we expect the negative correlation between the unemployment rate and ATM income
CPI_Inqoq	An increase in the Consumer Price Index might increase the usage of credit cards and, as a result, an increase in POS income	An increase in the Consumer Price Index might increase the usage of credit cards and, as a result, increase in ATM income
SalaryYear_Inoy	An increase in Salary might decrease the usage of credit cards and, as a result, decrease in POS income	An increase in Salary might decrease the usage of credit cards and, as a result, decrease in ATM income
s_month_since_NA	The long period after the inactive state might decrease the chance of debit transaction and, consequently, POS income in the next months	The long period after the inactive state might decrease the chance of debit transaction and, consequently, ATM income in the next months
s_month_since_Tr	The short period after the transactor state might increase the chance of debit transaction and, consequently, POS income in the next months	The short period after the transactor state might increase the chance of debit transaction and, consequently, ATM income in the next months
s_month_since_Re	The long period after the revolver state might decrease the chance of debit transaction and, consequently, POS income in the next months	The long period after the revolver state might decrease the chance of debit transaction and, consequently, ATM income in the next months
s_month_since_RP	The long period after the revolver repaid state might decrease the chance of debit transaction and, consequently, POS income in the next months	The long period after the revolver repaid state might decrease the chance of debit transaction and, consequently, ATM income in the next months
s_month_since_D1	The long period after the delinquent state might increase the chance of debit transaction and, consequently, POS income in the next months	The long period after the revolver repaid state might increase the chance of debit transaction and, consequently, ATM income in the next months
s_month_since_D2	The long period after the delinquent state might increase the chance of debit transaction and, consequently, POS income in the next months	The long period after the revolver repaid state might increase the chance of debit transaction and, consequently, ATM income in the next months
s_times_NA	The high number in the inactive state might decrease the chance of debit transaction and, consequently, POS income in the next months	The high number in the inactive state might decrease the chance of debit transaction and, consequently, ATM income in the next months

Table 7.1 Selected covariates and expectations for the impact on transactional income

Variable	POS income expectations	ATM income expectations
s_times_TR	The high number in the transactor state might increase the chance of debit transaction and, consequently, POS income in the next months	The high number in the transactor state might increase the chance of debit transaction and, consequently, ATM income in the next months
s_times_RE	The high number in the revolver state might increase the chance of debit transaction and, consequently, POS income in the next months	The high number in the revolver state might increase the chance of debit transaction and, consequently, ATM income in the next months
s_times_RP	The high number in the revolver repaid state might decrease the chance of debit transaction and, consequently, POS income in the next months	The high number in the revolver repaid state might decrease the chance of debit transaction and, consequently, ATM income in the next months
s_times_D1	The high number in the delinquent state might increase the chance of debit transaction and, consequently, POS income in the next months	The high number in the delinquent state might increase the chance of debit transaction and, consequently, ATM income in the next months
s_times_D2	The high number in the delinquent state might increase the chance of debit transaction and, consequently, POS income in the next months	The high number in the delinquent state might increase the chance of debit transaction and, consequently, ATM income in the next months
d_StateFull_1_NA	The current inactive state might decrease the chance of debit transaction and, consequently, POS income in the next months	The current inactive state might decrease the chance of debit transaction and, consequently, ATM income in the next months
d_StateFull_1_Tr	The current transactor state might increase the chance of debit transaction and, consequently, POS income in the next months	The current transactor state might increase the chance of debit transaction and, consequently, ATM income in the next months
d_StateFull_1_Re	The current revolver state might increase the chance of debit transaction and, consequently, POS income in the next months	The current revolver state might increase the chance of debit transaction and, consequently, ATM income in the next months
d_StateFull_1_RP	The current revolver repaid state might decrease the chance of debit transaction and, consequently, POS income in the next months	The current revolver repaid state might decrease the chance of debit transaction and, consequently, ATM income in the next months
d_StateFull_1_D1	The current delinquent state might increase the chance of debit transaction and, consequently, POS income in the next months	The current delinquent state might increase the chance of debit transaction and, consequently, ATM income in the next months

7.4 Non-interest income modelling results

7.4.1 The probability of transaction

Initially, we estimate the probability of transaction using the two-stage model. The distribution of the target variable values for active customers shows that 57.2% of observations have POS transaction and 31.3% have an ATM money withdrawal as shown in Table 7.2.

Table 7.2The distribution of binary target for the probability of transaction

Value	Description	POS		ATM	
		frequency	%	frequency	%
0	No transactions	84 945	42.8%	136 319	68.7%
1	Transactions exist	113 511	57.2%	62 137	31.3%
	Total	198 456	100.0%	198 456	100.0%

The parameter estimates for the probability of POS and ATM transactions for six months with binary logistic regression are given in Table 7.3. The coefficient estimates are given for equations 7.17 and 7.18 with covariates described in Chapter 3, which include behavioural, application, macroeconomic, and state characteristics. We use target equal to 1 as the event definition. So positive sign of the estimate means positive correlation with the probability of event.

We test the significance of estimated coefficients by use of Wald tests and p-level tests. The most significant characteristics are the credit limit, the utilisation rate at the observation month, the number and the sum of debit (spending) transactions, the payment amount less than 5 percent of the debt amount for the probability of both POS and ATM transactions, age for the probability of POS transactions which is less significant for the probability of ATM transactions, number of times in the inactive state for the probability of ATM transaction, and also previous usage of the credit card for POS and ATM transactions. The changes in credit limit are not significant for the prediction of the probability of making a POS or an ATM transaction.

Table 7.3 The logistic regression coefficient estimation for POS and ATM transaction probability during 6 months

Parameter	POS 6 months				ATM 6 months			
	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept	1.2266	0.6258	3.8415	0.05	-1.1823	0.4268	7.6737	0.0056
Mob	-0.00967	0.00175	30.4888	<.0001	0.00388	0.00192	4.0735	0.0436
Limit	0.000036	2.93E-06	154.1914	<.0001	0.000056	3.36E-06	281.5332	<.0001
UT0_1	1.0116	0.0948	113.9461	<.0001	1.3739	0.0987	193.6241	<.0001
UT0_2	0.4752	0.1237	14.7588	0.0001	0.5102	0.1257	16.4752	<.0001
UT0_3	0.1898	0.1083	3.074	0.0796	0.0193	0.1109	0.0304	0.8615
UT0_4	0.2012	0.1042	3.7267	0.0535	-0.044	0.1074	0.1679	0.682
UT0_5	0.2273	0.0959	5.6205	0.0178	0.0535	0.0998	0.2875	0.5918
UT0_6	0.059	0.0671	0.7727	0.3794	0.094	0.0704	1.7818	0.1819
b_UT1to2ln	-0.00888	0.00574	2.3871	0.1223	0.00354	0.00717	0.2433	0.6218
b_UT1to6ln	0.0293	0.0035	70.1563	<.0001	0.0235	0.00416	31.797	<.0001
avg_balance_1	-0.00009	0.000018	27.8559	<.0001	-0.00007	0.000017	17.3271	<.0001
avg_balance_3	-0.00007	0.00002	12.5565	0.0004	-0.00007	0.000019	13.577	0.0002
avg_balance_4	-9.9E-06	0.000019	0.2615	0.6091	-0.00001	0.000019	0.2956	0.5867
avg_balance_5	0.000021	0.000019	1.2253	0.2683	0.000027	0.000019	2.1655	0.1411
avg_balance_6	0.000011	0.000018	0.3816	0.5368	0.000023	0.000018	1.6409	0.2002
avg_balance_7	0.000038	0.000014	7.1887	0.0073	0.000023	0.000014	2.5544	0.11
avg_deb_amt_1	0.000027	0.000032	0.7479	0.3871	0.000112	0.000033	11.6087	0.0007
avg_deb_amt_2	0.000047	0.000031	2.2655	0.1323	0.000123	0.000032	15.3038	<.0001
avg_deb_amt_3	0.000073	0.000031	5.7879	0.0161	0.000086	0.000031	8.0016	0.0047
avg_deb_amt_4	-0.00006	0.00003	3.8682	0.0492	-0.00009	0.000029	9.5019	0.0021
avg_deb_amt_5	-0.00011	0.000027	15.9574	<.0001	-0.00008	0.000027	9.4436	0.0021
avg_deb_amt_6	-0.00011	0.000021	24.3128	<.0001	-0.00015	0.000022	46.0382	<.0001
sum_crd_amt_1	-0.00005	0.00001	23.4939	<.0001	-0.00001	0.00001	1.3502	0.2452
sum_crd_amt_2	-0.00012	0.000016	54.8161	<.0001	-0.00007	0.000015	22.2117	<.0001
sum_crd_amt_3	-0.00013	0.000017	53.3064	<.0001	-0.00008	0.000016	26.2879	<.0001
sum_crd_amt_4	-0.00009	0.000016	31.3582	<.0001	-0.00008	0.000016	23.4846	<.0001
sum_crd_amt_5	-0.00004	0.000014	9.7614	0.0018	-0.00004	0.000013	11.3052	0.0008
sum_crd_amt_6	0.000013	8.58E-06	2.4735	0.1158	-0.00001	8.25E-06	3.2425	0.0717
sum_deb_amt_1	0.000524	0.000046	128.9539	<.0001	0.000157	0.00003	27.1069	<.0001
sum_deb_amt_2	0.000399	0.000044	83.5303	<.0001	0.00015	0.000032	21.5956	<.0001
sum_deb_amt_3	0.000505	0.000047	116.768	<.0001	0.000235	0.000035	46.5173	<.0001
sum_deb_amt_4	0.000726	0.000048	226.2419	<.0001	0.000334	0.000034	98.8746	<.0001
sum_deb_amt_5	0.000675	0.000044	240.1676	<.0001	0.000304	0.00003	101.2782	<.0001
sum_deb_amt_6	0.000408	0.000033	151.6679	<.0001	0.000232	0.000025	86.0364	<.0001
max_deb_amt_1	-0.00059	0.000048	147.27	<.0001	-0.00031	0.000033	86.0823	<.0001
max_deb_amt_2	-0.0004	0.000044	80.8443	<.0001	-0.00021	0.000033	41.4772	<.0001
max_deb_amt_3	-0.00047	0.000047	103.3656	<.0001	-0.00023	0.000034	43.278	<.0001
max_deb_amt_4	-0.00063	0.000048	170.7748	<.0001	-0.00027	0.000033	64.5904	<.0001
max_deb_amt_5	-0.00061	0.000044	195.3235	<.0001	-0.00026	0.00003	75.6362	<.0001
max_deb_amt_6	-0.00038	0.000034	124.5189	<.0001	-0.00019	0.000026	54.5104	<.0001
b_AvgOB1_to_MaxOB1_ln	-0.0511	0.0314	2.6411	0.1041	-0.2374	0.0357	44.1465	<.0001
b_AvgOB2_to_MaxOB2_ln	-0.0248	0.0317	0.6154	0.4328	-0.1482	0.0353	17.5914	<.0001
b_AvgOB3_to_MaxOB3_ln	0.0345	0.0316	1.1928	0.2748	-0.0392	0.0353	1.2317	0.2671
b_TRmax_deb1_To_Limit	0.1957	0.0802	5.9591	0.0146	0.2687	0.0781	11.8323	0.0006
b_TRmax_deb2_To_Limit	-0.1583	0.0777	4.1491	0.0417	-0.0973	0.0782	1.55	0.2131
b_TRmax_deb3_To_Limit	-0.1212	0.0785	2.3852	0.1225	-0.2073	0.0777	7.116	0.0076
b_TRev_avg_deb1_to_avgO	-0.2322	0.0139	278.0881	<.0001	-0.4618	0.0148	972.2819	<.0001
b_TRev_avg_deb2_to_avgO	-0.0733	0.0138	28.3322	<.0001	-0.168	0.0144	135.5269	<.0001
b_TRev_avg_deb3_to_avgO	-0.0273	0.0132	4.275	0.0387	-0.0769	0.0139	30.8334	<.0001
b_TRsum_deb1_to_TRsumcrd	0.2639	0.0122	468.2899	<.0001	0.4824	0.0129	1402.108	<.0001
b_TRsum_deb2_to_TRsumcrd	0.1692	0.0118	206.8988	<.0001	0.2843	0.0125	520.2947	<.0001
b_TRsum_deb3_to_TRsumcrd	0.0801	0.0111	52.3843	<.0001	0.1266	0.0115	121.5069	<.0001

Table 7.3 The logistic regression coefficient estimation for POS and ATM transaction probability during 6 months

Parameter	POS 6 months				ATM 6 months			
	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
b_avgNumDeb13	0.1583	0.00779	413.2637	<.0001	0.0872	0.00571	233.5596	<.0001
b_OB13_to_OB46ln	0.0609	0.00568	114.7272	<.0001	0.0427	0.00751	32.3233	<.0001
b_OB1_to_OB2_ln	0.0397	0.0114	12.0561	0.0005	0.0396	0.0134	8.6918	0.0032
b_OB2_to_OB3_ln	-0.00448	0.00785	0.3253	0.5684	-0.00746	0.00909	0.673	0.412
b_OB3_to_OB4_ln	0.00653	0.00409	2.554	0.11	0.00138	0.005	0.0755	0.7835
b_pos_flag_use13vs46	-0.2262	0.0196	133.173	<.0001	-0.3909	0.0192	415.5543	<.0001
b_atm_flag_use13vs46	0.3861	0.0214	325.4562	<.0001	0.4557	0.0242	354.718	<.0001
b_pos_use_only_flag	1.0176	0.0235	1879.161	<.0001	0.8991	0.0212	1790.977	<.0001
b_atm_use_only_flag	-0.9266	0.0145	4087.033	<.0001	-1.348	0.0181	5556.232	<.0001
b_TRsum_crd1_to_OB1_ln	0.2344	0.0128	334.3418	<.0001	0.418	0.0144	842.6531	<.0001
b_TRsum_crd2_to_OB2_ln	0.0286	0.0118	5.8311	0.0157	0.1185	0.0131	82.0747	<.0001
b_TRsum_crd3_to_OB3_ln	0.0603	0.0108	31.205	<.0001	0.1078	0.0117	85.486	<.0001
b_payment_lt_5p_1	-0.4092	0.0177	534.5078	<.0001	-0.7724	0.0182	1797.02	<.0001
b_payment_lt_5p_2	-0.0826	0.0181	20.8111	<.0001	-0.0457	0.0187	5.9947	0.0143
b_payment_lt_5p_3	-0.1066	0.0178	35.9951	<.0001	-0.1177	0.0184	41.1117	<.0001
b_maxminOB_avgOB_1_ln	0.1448	0.0144	100.5461	<.0001	0.1507	0.0173	75.5469	<.0001
b_maxminOB_avgOB_2_ln	0.1054	0.0144	53.8517	<.0001	0.0544	0.017	10.2628	0.0014
b_maxminOB_avgOB_3_ln	0.1131	0.0138	67.3349	<.0001	0.0349	0.0162	4.6651	0.0308
b_TRsum_deb1_to_2_ln	0.0243	0.00524	21.5848	<.0001	0.0377	0.00605	38.7494	<.0001
b_TRsum_crd1_to_2_ln	-0.1179	0.00954	152.8142	<.0001	-0.1469	0.0102	208.4463	<.0001
l_ch1_ln	-0.0898	0.112	0.6432	0.4226	0.2042	0.1169	3.0536	0.0806
l_ch1_flag	0.1439	0.0582	6.1056	0.0135	-0.1219	0.06	4.1315	0.0421
l_ch6_flag	0.1274	0.0215	35.0101	<.0001	0.2211	0.0234	89.3605	<.0001
age	-0.0317	0.000765	1720.809	<.0001	-0.0181	0.00087	431.4165	<.0001
customer_income_ln	0.1871	0.0199	88.035	<.0001	0.0347	0.0213	2.6536	0.1033
Edu_High	0.0774	0.0178	18.9156	<.0001	0.0316	0.0188	2.8405	0.0919
Edu_Special	-0.0717	0.0167	18.4143	<.0001	-0.0567	0.0182	9.7628	0.0018
Edu_TwoDegree	0.2896	0.0453	40.9066	<.0001	0.2016	0.0453	19.8303	<.0001
Marital_Civ	0.1485	0.0279	28.2708	<.0001	0.0214	0.0293	0.5327	0.4655
Marital_Div	0.1429	0.0187	58.2154	<.0001	0.0899	0.0206	19.1029	<.0001
Marital_Sin	0.058	0.0211	7.5287	0.0061	-0.00992	0.0219	0.2056	0.6502
Marital_Wid	0.1376	0.0336	16.7291	<.0001	0.0658	0.0404	2.6521	0.1034
position_Man	0.0337	0.0191	3.1189	0.0774	0.048	0.021	5.2177	0.0224
position_Oth	-0.0154	0.0179	0.746	0.3877	-0.0236	0.0199	1.4062	0.2357
position_Tech	-0.00825	0.0169	0.2369	0.6264	-0.0437	0.0184	5.647	0.0175
position_Top	0.0216	0.0357	0.3669	0.5447	-0.0246	0.0421	0.343	0.5581
sec_Agricult	-0.034	0.0318	1.1427	0.2851	-0.06	0.0377	2.5383	0.1111
sec_Constr	0.1671	0.0432	14.9453	0.0001	0.0433	0.0476	0.8259	0.3635
sec_Energy	0.0719	0.0274	6.8843	0.0087	0.1181	0.0308	14.6658	0.0001
sec_Fin	0.4746	0.0225	445.4059	<.0001	0.3498	0.0231	228.6328	<.0001
sec_Industry	0.4194	0.0586	51.1472	<.0001	0.1748	0.0645	7.3371	0.0068
sec_Manufact	0.07	0.0405	2.9909	0.0837	-0.0343	0.0455	0.5685	0.4508
sec_Mining	0.0125	0.0297	0.1777	0.6734	-0.00587	0.0338	0.0301	0.8622
sec_Service	0.1172	0.0156	56.7706	<.0001	0.1046	0.0172	36.8336	<.0001
sec_Trade	0.4036	0.0237	290.1792	<.0001	0.2467	0.0231	114.275	<.0001
sec_Trans	0.0296	0.0429	0.4767	0.4899	0.098	0.0468	4.3754	0.0365
car_Own	0.0505	0.0153	10.9364	0.0009	0.00002	0.0173	0	0.9991
car_coOwn	-0.0524	0.0225	5.4117	0.02	-0.0778	0.0259	9.033	0.0027
real_Own	-0.0179	0.0148	1.4506	0.2284	0.00985	0.016	0.3795	0.5378
real_coOwn	-0.0201	0.0157	1.6351	0.201	0.0164	0.0165	0.9933	0.3189
reg_ctr_Y	-0.2676	0.0262	104.4564	<.0001	-0.1753	0.0245	51.1296	<.0001
reg_ctr_N	-0.6165	0.0257	577.2042	<.0001	-0.395	0.0244	261.5991	<.0001
child_1	0.0554	0.0192	8.346	0.0039	0.0214	0.0205	1.0973	0.2949

Table 7.3 The logistic regression coefficient estimation for POS and ATM transaction probability during 6 months

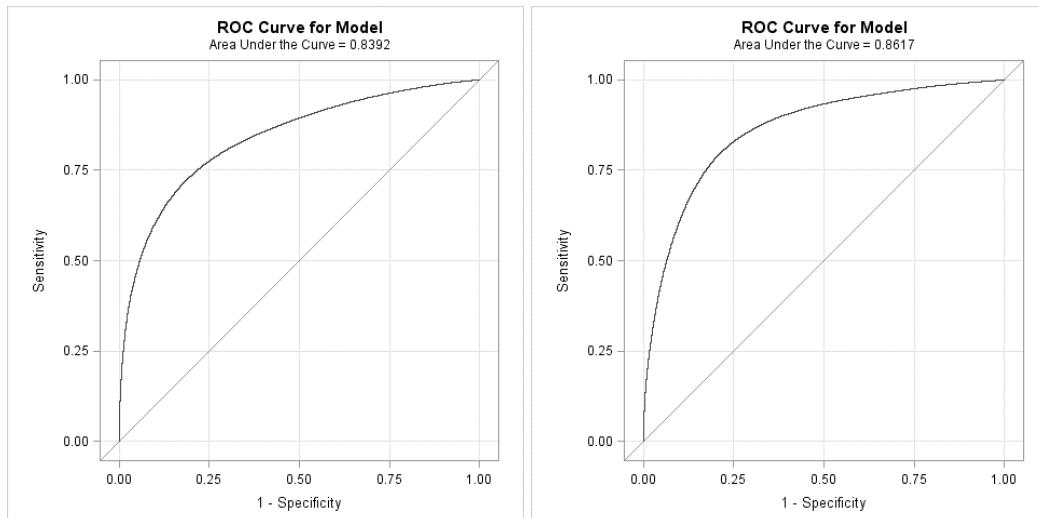
Parameter	POS 6 months				ATM 6 months			
	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
child_2	0.0381	0.0109	12.2922	0.0005	0.00413	0.0121	0.1169	0.7324
child_3	-0.047	0.0355	1.7466	0.1863	-0.0568	0.0422	1.8137	0.1781
Unempl_Inyoy	-3.4052	0.2464	191.0362	<.0001	-0.3296	0.2735	1.4528	0.2281
UAH_EURRate_Inmom	-1.0525	0.304	11.9898	0.0005	-3.7064	0.3409	118.1775	<.0001
UAH_EURRate_Inyoy	-2.6996	0.1907	200.4078	<.0001	-2.3032	0.2101	120.2076	<.0001
CPI_Inqoq	7.0965	0.5127	191.5759	<.0001	1.6548	0.557	8.8245	0.003
SalaryYear_Inyoy	0.5045	0.43	1.3765	0.2407	-0.177	0.4709	0.1412	0.7071
s_month_since_NA	0.017	0.0118	2.0725	0.15	-0.0265	0.0128	4.2714	0.0388
s_month_since_Tr	-0.00091	0.013	0.0049	0.9442	-0.0063	0.0143	0.1927	0.6607
s_month_since_Re	-0.0115	0.0167	0.4792	0.4888	0.0253	0.019	1.7653	0.184
s_month_since_RP	0.0514	0.0151	11.6568	0.0006	0.0489	0.0164	8.8389	0.0029
s_month_since_D1	0.0387	0.0214	3.265	0.0708	0.0751	0.0175	18.4936	<.0001
s_month_since_D2	0.1277	0.0696	3.3653	0.0666	0.1983	0.036	30.2599	<.0001
s_times_NA	-0.1994	0.0716	7.7534	0.0054	-0.4625	0.051	82.3257	<.0001
s_times_TR	0.1581	0.078	4.1076	0.0427	-0.0475	0.0588	0.6512	0.4197
s_times_RE	0.1041	0.0686	2.3001	0.1294	-0.073	0.0451	2.62	0.1055
s_times_RP	0.3163	0.0799	15.6903	<.0001	0.0444	0.0626	0.5047	0.4775
s_times_D1	0.7835	0.0925	71.7036	<.0001	0.3469	0.0546	40.312	<.0001
s_times_D2	0.6532	0.3016	4.69	0.0303	0.4631	0.1554	8.8852	0.0029
d_StateFull_1_NA	-2.9133	0.3929	54.9723	<.0001	-1.8044	0.2225	65.7501	<.0001
d_StateFull_1_Tr	-2.3666	0.3933	36.2044	<.0001	-0.9343	0.2166	18.5966	<.0001
d_StateFull_1_Re	-2.6155	0.3818	46.9277	<.0001	-1.5534	0.1765	77.4851	<.0001
d_StateFull_1_RP	-2.4446	0.3909	39.1047	<.0001	-0.8805	0.2069	18.1104	<.0001
d_StateFull_1_D1	-2.1913	0.388	31.8884	<.0001	-1.4692	0.1732	71.993	<.0001

The covariate ‘Month on Book’ has a negative sign for POS transactions and a positive sign for ATM transactions. This means that accounts with long time usage generate more ATM income and less POS income. Generally, this tendency is expected because longtime credit card usage usually means the motivation of the customer to use POS for payments. However, this can be explained by the specifics of the local market to use the credit card for the first time for cash withdrawals.

Higher utilisation rate increases the probability of transactions. Covariates related to purchases and outstanding balance such as the sum of debit transactions (sum_deb_1), the ratio of debit transactions to credit transactions in the observation month (b_TRsum_deb1_to_TRsumcrd), the logarithm of the sum of payments to the outstanding balance (b_TRsum_crd1_to_OB1_ln) have similar slopes for POS and ATM and this generally means that high spending in the past explains high spending in the future. Current state characteristics are significant. Transactors and repaid revolvers have a lower probability of usage of credit cards for both POS and ATM

transactions. Higher number of times in inactive state (`s_times_NA`) increase the probability of ATM transaction and is significant for the probability of POS transaction. The average debit transaction amount to average outstanding balance (`b_TAvg_deb1_to_avgO`) has a negative sign logically explain the decrease of the probability of transaction for an active credit card user. The usage of a credit card for the last 6 months for POS transactions only (`b_pos_use_only_flag`) increase the probability of both POS and ATM transactions, and vice versa the usage of the credit card for ATM transactions only ((`b_atm_use_only_flag`)) decreases the chance of transaction. However, if a customer did not use a credit card for ATM transactions, but has started to use it for cash withdrawals, this increases the chance of any transaction (positive sign for `b_atm_flag_use13vs46`), and if a customer did not use a credit card for POS transactions, but have started to use it for non-cash payments, this decreases the chance of any transaction (positive sign for `b_pos_flag_use13vs46`). From the set of macroeconomic variables, only CPI quarterly changes are a significant character and CPI growth decreases the probability of credit card usage.

Figure 7.1 ROC Curves for POS and ATM probability of transaction for 6 months



The predictive power of the model is measured by the area under the ROC curve and the Gini coefficient. The results of the development and validation sample performance tests are high: ROC around 0.82 and 0.85 for out-of-sample validation for POS and ATM respectively (Table 7.4). There are relatively high values. The ROC curves are shown in Figure 7.1

Table 7.4 AUC and Gini for POS and ATM development and validation samples

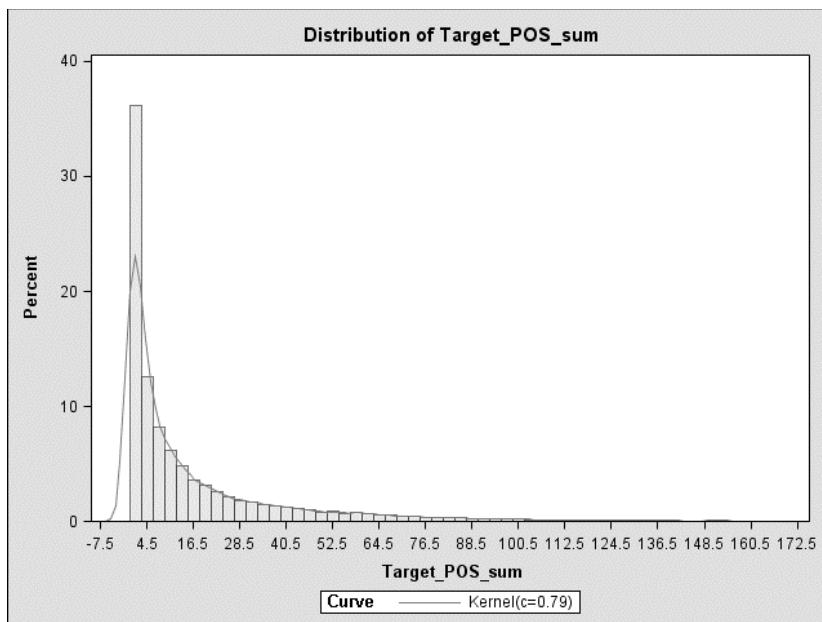
Model	AUC		Gini	
	Development	Out-of-sample	Development	Out-of-sample
POS	0.8392	0.8205	0.6784	0.6410
ATM	0.8617	0.8492	0.7234	0.6984

7.4.2 Distributions for POS and ATM income

We consider the following distributions and statistics only for cases with positive POS transactions for 6 months, that is, active cases only.

The shape of the POS income distribution in Figure 7.2 is positively skewed with a long thin tail at high POS income values and has a high concentration of income around zero (or minimum income equal to 0.1 money units).

Figure 7.2 Distribution of the target – POS amount for 6 months



The total number of observations is 44,056. However, each account has an observation for several time points, and all related characteristics are presented as time series. For the current distributions, we use a pooled method of aggregation: each account at a certain time point is presented as a separate independent observation.

Table 7.5 Distribution characteristics of the target – POS amount for 6 months

Moments			
N	44056	Sum Weights	44056
Mean	17.23	Sum Observations	759192.82
Std Deviation	25.93	Variance	672.72
Skewness	2.51	Kurtosis	7.14

The POS income for 6 months is skewed and has high asymmetry (Skewness is 2.5112). Mean and median are quite far from each other: 17.23 and 6.39 respectively are shown in Table 7.5. and Table 7.6.

Table 7.6 Distribution quantiles of the target – POS amount for 6 months

Quantiles	
Quantile	Estimate
100% Max	164.7723
99%	126.9305
95%	72.9532
90%	50.299
75% Q3	21.2832
50% Median	6.3903
25% Q1	1.3375
10%	0.3234
5%	0.1885
1%	0.1189
0% Min	0.1001

We consider the following distributions and statistics only for cases with positive ATM transactions for 6 months (or active cases only). The shape of the ATM income distribution is similar to a lognormal with a long thin tail for high ATM income values with a high concentration of income around zero (or minimum income equal to 6 money units) as shown in Figure 7.3.

The total number of observations is 94107 (Table 7.7). However, each account has an observation for several time points, and all related characteristics are presented as time series. For current distributions, we use a pooled method of aggregation: each account at a certain time point is presented as a separate independent observation.

Figure 7.3 Distribution of the target – ATM amount for 6 months

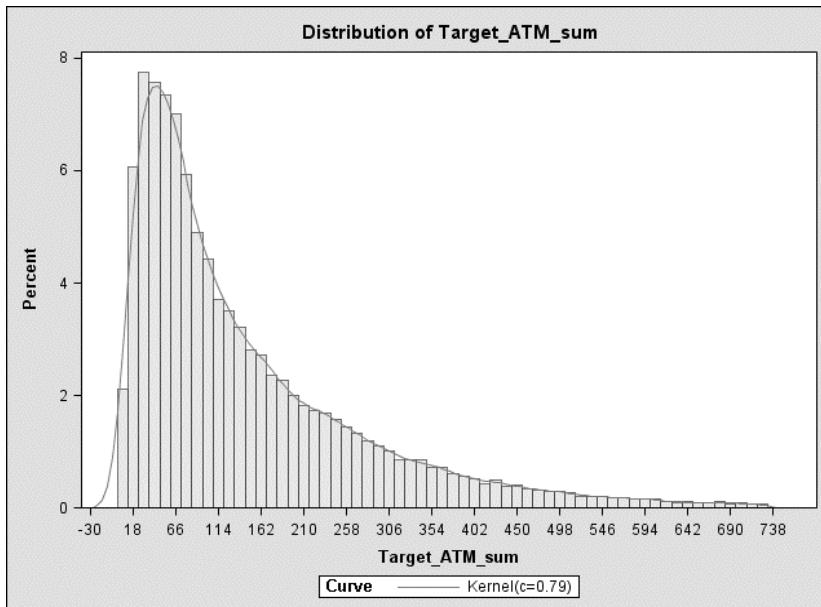


Table 7.7 Distribution characteristics of the target – ATM amount for 6 months

Moments			
N	94107	Sum Weights	94107
Mean	144.77	Sum Observations	13623924
Std Deviation	131.68	Variance	17340.36
Skewness	1.60	Kurtosis	2.59

The distribution of summed ATM income over 6 months' values is skewed and has high asymmetry (Skewness is 1.6085) Mean and median are far from each other: 144.77 and 99.4 respectively. However, the skewness of ATM income is less than for POS income as shown in Table 7.7 and Table 7.8.

Table 7.8 Distribution quantiles of the target – ATM amount for 6 months

Quantiles (Definition 5)	
Quantile	Estimate
100% Max	729.95
99%	607
95%	424.75
90%	330.1
75% Q3	199.5
50% Median	99.4
25% Q1	50
10%	26.5
5%	18
1%	10
0% Min	6.13

7.4.3 Explanatory Variables Distribution

We have selected four characteristics for both POS and ATM income to show visually the correlation between explanatory variables and the dependent variable. These are Sum of Purchases to Sum of Payments for one month, Average Number of Debit Transactions for 1-3 months, the Credit Limit Utilization rate for 6 months, and Age. These variables represent different trends and categories of variables. The axis X reflects the values of characteristics, the axis Y – the POS/ATM income amount respectively, and the additional axis Y (right side) – number of pooled observations. We also put a linear trend line on the graph, and R-squared value for observed income and trend line to understand how far the real dependence of income on independent variable is from the ideal linear equation.

Figure 7.4 Dependence of Average POS income for 6 months on the Logarithm of Sum of Purchases to Sum of Payments for one month

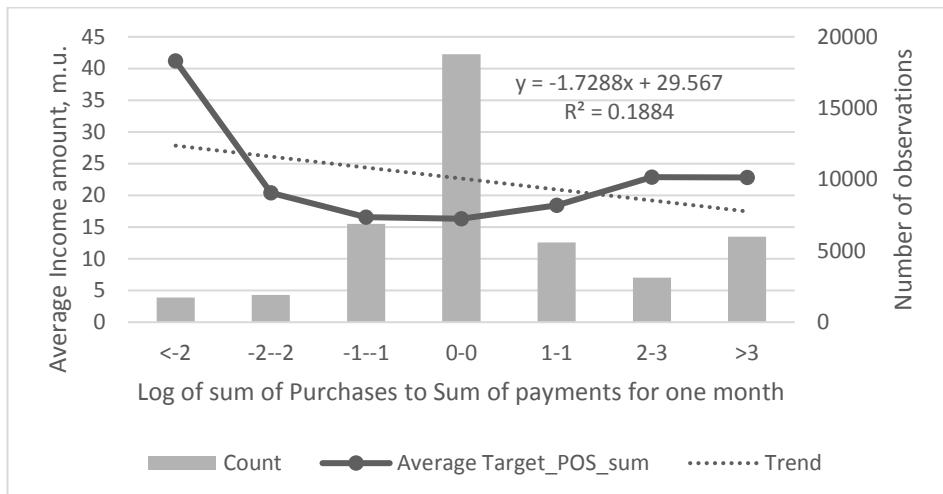


Figure 7.5 Dependence of Average POS income for 6 months on Average Number of Debit Transactions for 1-3 months

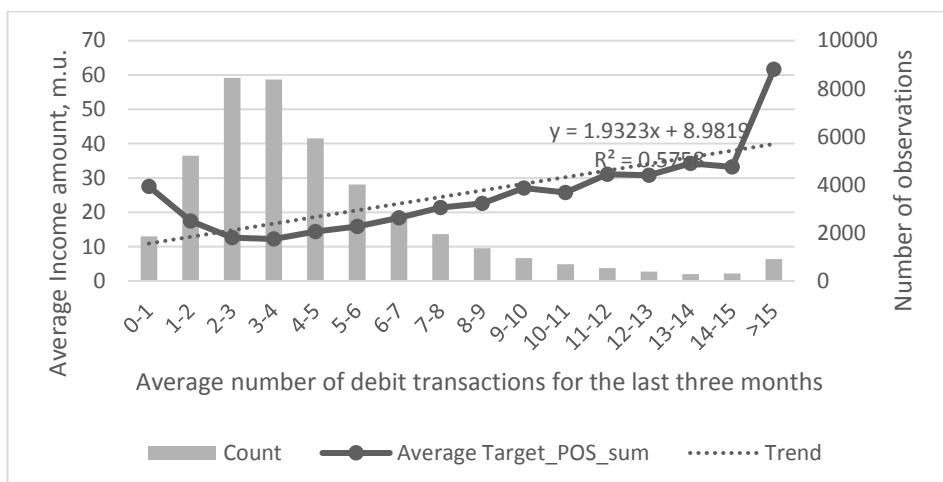


Figure 7.6 Dependence of Average POS income for 6 months on the Credit Limit Utilization rate for 6 months

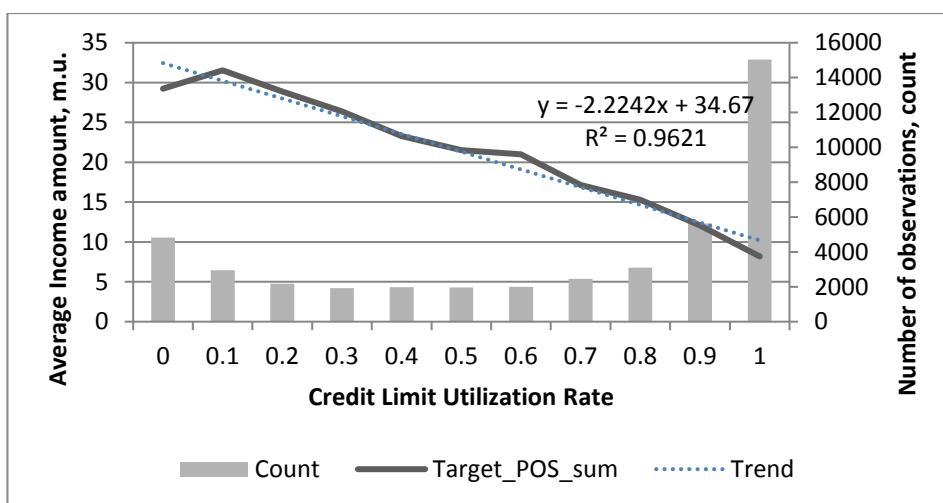
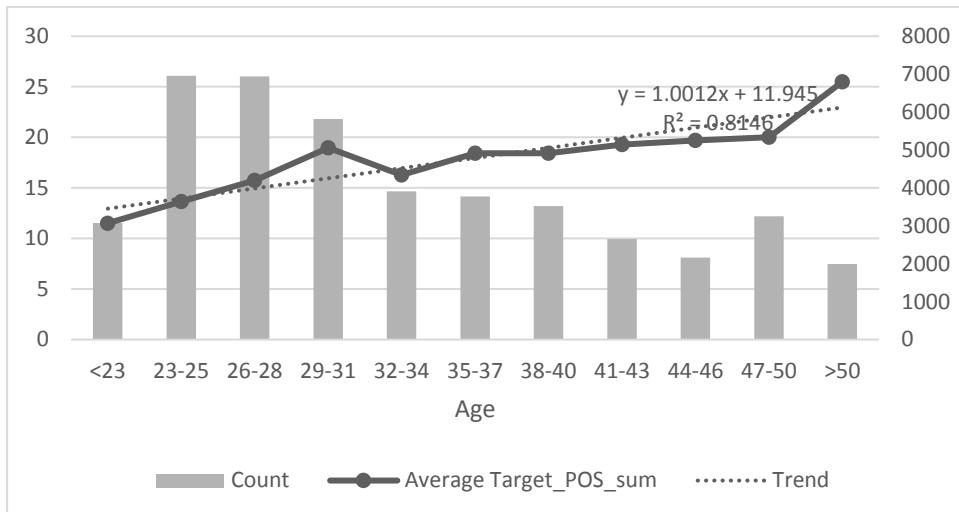


Figure 7.7 Dependence of Average POS income for 6 months on Age



The average POS and average ATM income scales are different. Scale ranges are 0-30 for POS income amount and scale ranges are 0-200 for ATM income amount. This difference means that income per transaction, which is generated by fees from ATM cash withdrawals, significantly exceed the income per POS transaction. However, the total sum of POS income can exceed the ATM income amount due to the number of POS transactions.

Figure 7.8 Dependence of Average ATM income for 6 months on the Logarithm of Sum of Purchases to Sum of Payments for one month

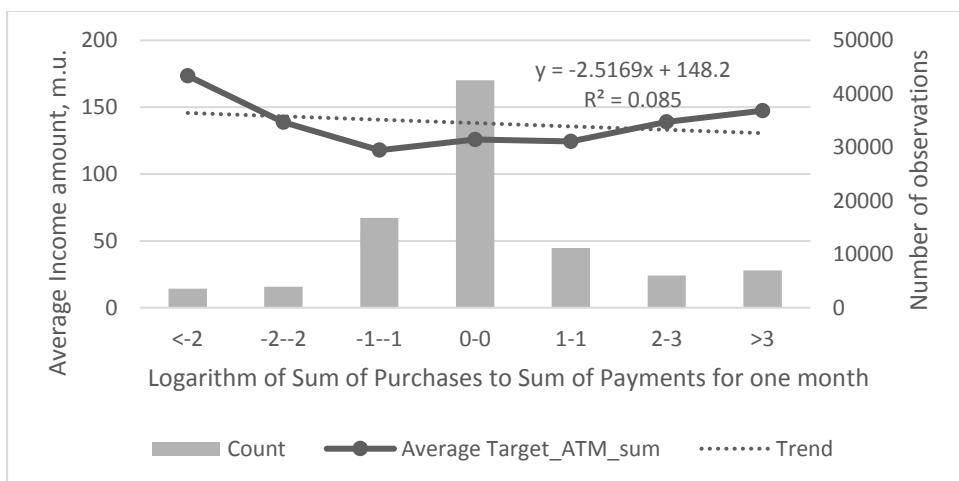


Figure 7.9 Dependence of Average ATM income for 6 months on Average Number of Debit Transactions for 1-3 months

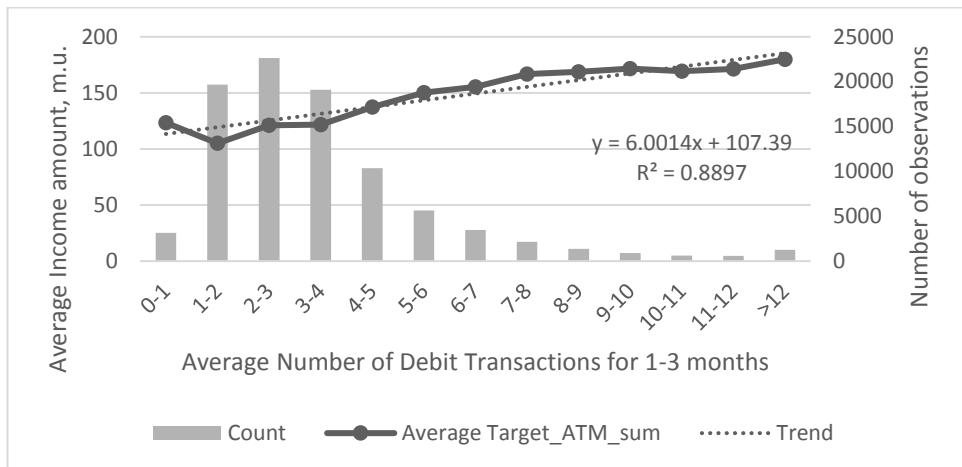


Figure 7.10 Dependence of Average ATM income for 6 months on the Credit Limit Utilization rate for 6 months

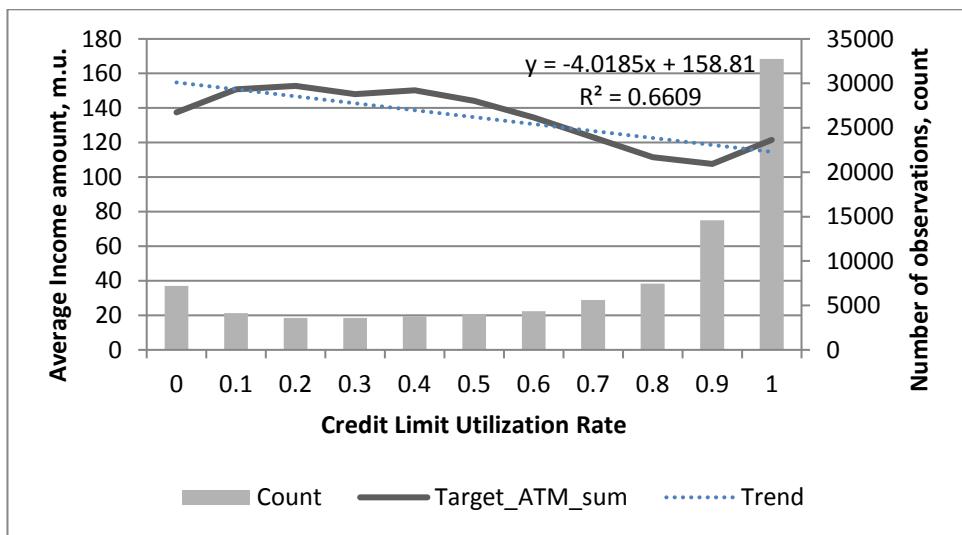
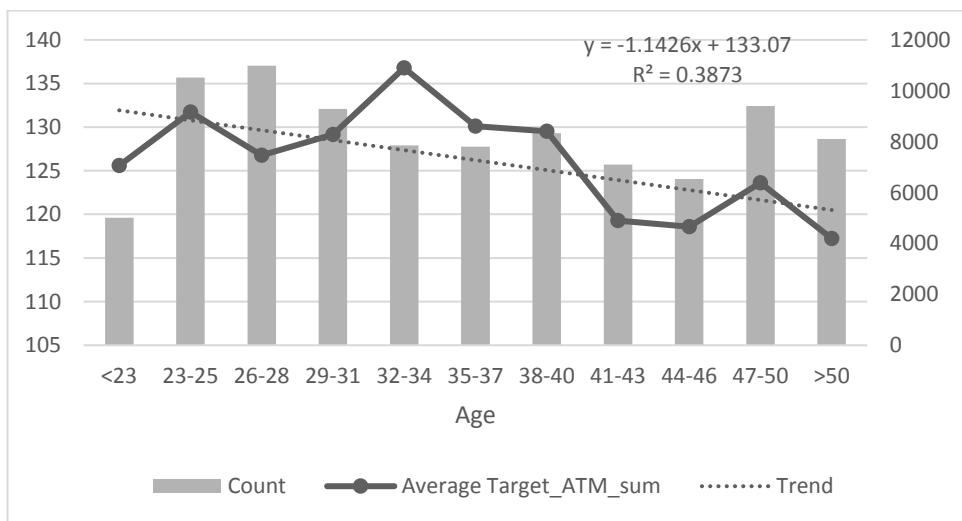


Figure 7.11 Dependence of Average ATM income for 6 months on Age



The univariate analysis shows that some characteristics as, for example, Age (Figure 7.7 and Figure 7.11) and Utilization Rate (Figure 7.6 and Figure 7.10Figure 7.5) have a relatively strong relations with transactional income. Higher number of debit transactions increase the expected average POS and ATM income in the next 6 months (Figure 7.5 and Figure 7.9). However, the trends for POS and ATM income can have opposite directions. For example, higher Age generates more POS income (Figure 7.7), but older people give less ATM income for the bank (Figure 7.11). Some behavioural covariates might be insignificant. For example, for the Logarithm of Sum of Purchases to Sum of Payments is a behavioural characteristic (Figure 7.4), the trend line is close to flat, in particular, at the area of high concentration of the distribution for the logarithm values around zero, and it is mostly because of a non-linear relationship between the values of the characteristic and income amount (Figure 7.4).

We have selected the same set of covariates as described in Chapter 3 and used for the prediction of the transition probabilities between credit card states in Chapter 6. We want to investigate how this set of predictors will explain the transactional income amount from POS purchases and ATM cash withdrawals with panel data, which consider both cross-sectional and time series components.

7.5 POS income estimation

7.5.1 Comparative analysis of the regression coefficients for pooled and random effect estimation methods

We use four methods for the random-effect estimation and apply the following notation: WK - Wansbeek and Kapteyn Method, FB - Fuller and Battese Method, WH - Wallace and Hussain Method, and NL - Nerlove's Method. The first three methods were selected by Baltagi and Chang (1994) as methods, which have demonstrated the highest predictive accuracy and efficiency for a sample of panel data. We also selected Nerlove's method is a simple random-effect method, which we use for the comparison with others. These four methods are the most popular and implemented in statistical software (for example, SAS).

In this section, we show two parameters for each method: OLS regression coefficient estimates and p-values as a test of a coefficient's level of significance. We do not apply any benchmarks for the p-value but use them for the comparative analysis of the regression methods.

The comparative analysis of the results of the regression estimation is contained in Table 7.9 with different methods to highlight i) the significance of the coefficient and ii) the trend of dependence between the predictor and the dependent variable. For example, Month on Book is a significant predictor for all random-effect methods (p-value is <0.0001), but not for the pooled method estimation. All cases show that longer usage of the credit card gives as a higher expected income from the point-of-sales transactions.

The credit limit (Limit_1) is a significant characteristic with a positive sign for the pooled regression method, but not significant for the random-effect methods and has a negative relation between the credit limit and the income amount. However, even for the pooled regression estimation, the value of the coefficient means that for each money unit of the credit limit the average generated income is 0.000159. For example, if the size of a credit limit equal to 1000 money units this gives only 0.16 money units of additional income. This means that the credit limit size can be excluded from the model only minimal effect on the total predictive accuracy.

However, the characteristics related to the credit limit changes can be predictive. The logarithm of the credit limit changes (l_ch_1_ln) is relatively significant (with p-level value 0.0776). The growth of the logarithm of the limit changes by 1, which is equal to the growth of the credit limit value by 2.7182 money units, this adds 2.3172 money units to the total POS non-interest income estimation. The coefficients given by the random effect methods are insignificant with p-levels values around 0.38. The binary flag of the credit limit changes for the last 6 months (l_ch6_flag) has got a high significance for all methods and has a negative sign. This means that, if the limit changes, an account gets -1.55 money units in case of the pooled effect model and only -0.7715 money unit in case of the random effect model with the Wallace and Hussain Method.

Table 7.9 POS Income 6 months – Linear Regression Estimation Results

Variable	Pooled		WK		FB		WH		NL	
	Estimate	Pr > t	Estimate	Pr > t	Estimate	Pr > t	Estimate	Pr > t	Estimate	Pr > t
Intercept	-13.258	(0.0003)	2.065767	(0.5533)	1.068244	(0.7466)	0.098645	(0.9757)	2.239272	(0.5276)
mob	0.02774	(0.3116)	0.237426	(<.0001)	0.220577	(<.0001)	0.205106	(<.0001)	0.240476	(<.0001)
limit_1	0.000159	(0.0002)	-0.00004	(0.363)	-0.00002	(0.5977)	-7.8E-06	(0.8624)	-0.00004	(0.328)
UTO_1	-6.69484	(<.0001)	-5.97111	(<.0001)	-6.09816	(<.0001)	-6.19607	(<.0001)	-5.94577	(<.0001)
UTO_2	3.669884	(0.0179)	0.673567	(0.5397)	0.828207	(0.4586)	0.989989	(0.3839)	0.648058	(0.554)
UTO_3	2.446091	(0.0673)	0.581465	(0.5358)	0.655739	(0.4928)	0.739464	(0.4474)	0.569957	(0.5427)
UTO_4	-0.30908	(0.8108)	-0.58488	(0.5179)	-0.61882	(0.5016)	-0.63693	(0.4969)	-0.57711	(0.5222)
UTO_5	0.103936	(0.933)	-0.30802	(0.7219)	-0.36009	(0.6827)	-0.39287	(0.6614)	-0.29679	(0.7308)
UTO_6	-1.5758	(0.0779)	-0.04502	(0.9452)	-0.30177	(0.6498)	-0.5128	(0.447)	0.004441	(0.9946)
b_UT1to2In	-0.02676	(0.7478)	-0.0252	(0.6667)	-0.02622	(0.6599)	-0.02701	(0.6561)	-0.02501	(0.6682)
b_UT1to6In	0.08954	(0.0718)	0.090115	(0.0134)	0.091715	(0.0133)	0.092876	(0.0136)	0.08979	(0.0135)
avg_balance_1	0.000069	(0.7328)	0.000571	(<.0001)	0.000568	(0.0001)	0.00056	(0.0002)	0.000571	(<.0001)
avg_balance_2	-0.0008	(0.0006)	-0.00024	(0.1467)	-0.00026	(0.1193)	-0.00028	(0.0951)	-0.00023	(0.1517)
avg_balance_3	-0.00049	(0.0257)	-0.00027	(0.0844)	-0.00027	(0.0872)	-0.00027	(0.088)	-0.00027	(0.0836)
avg_balance_4	0.000113	(0.5999)	-0.00008	(0.5751)	-0.00007	(0.6458)	-0.00006	(0.71)	-0.00009	(0.5623)
avg_balance_5	0.000052	(0.8061)	-0.00022	(0.1283)	-0.00021	(0.1654)	-0.00019	(0.2051)	-0.00023	(0.1222)
avg_balance_6	0.0000576	(0.0009)	-0.00004	(0.7375)	4.95E-06	(0.9689)	0.000047	(0.7174)	-0.00005	(0.6852)
avg_deb_amt_1	0.000127	(0.7854)	-0.00189	(<.0001)	-0.00179	(<.0001)	-0.00169	(<.0001)	-0.0019	(<.0001)
avg_deb_amt_2	0.002277	(<.0001)	-0.00066	(0.0447)	-0.00052	(0.1174)	-0.00038	(0.258)	-0.00068	(0.0372)
avg_deb_amt_3	0.002203	(<.0001)	-0.00035	(0.2581)	-0.00022	(0.4808)	-0.00009	(0.7705)	-0.00038	(0.2281)
avg_deb_amt_4	0.001359	(0.0002)	-0.00032	(0.2373)	-0.00025	(0.3567)	-0.00018	(0.5067)	-0.00033	(0.22)
avg_deb_amt_5	0.000553	(0.1011)	-0.0005	(0.0418)	-0.00047	(0.0634)	-0.00043	(0.0934)	-0.00051	(0.0388)
avg_deb_amt_6	0.000602	(0.0364)	-0.00018	(0.3972)	-0.00015	(0.4827)	-0.00012	(0.5747)	-0.00018	(0.3832)
sum_crd_amt_1	0.001207	(<.0001)	0.001036	(<.0001)	0.001061	(<.0001)	0.001084	(<.0001)	0.001031	(<.0001)
sum_crd_amt_2	0.000576	(0.0013)	0.000862	(<.0001)	0.000869	(<.0001)	0.000873	(<.0001)	0.000861	(<.0001)
sum_crd_amt_3	0.000044	(0.8237)	0.000639	(<.0001)	0.000633	(<.0001)	0.000625	(<.0001)	0.00064	(<.0001)
sum_crd_amt_4	-0.0001	(0.6053)	0.000329	(0.0146)	0.00033	(0.0159)	0.00033	(0.0177)	0.000328	(0.0144)
sum_crd_amt_5	-0.00005	(0.7737)	0.000114	(0.3175)	0.000125	(0.28)	0.000135	(0.2507)	0.000112	(0.3247)
sum_crd_amt_6	0.000187	(0.0575)	-0.00015	(0.0288)	-0.00012	(0.0836)	-0.0001	(0.1944)	-0.00016	(0.0235)
sum_deb_amt_1	0.002086	(<.0001)	-0.00084	(0.0005)	-0.00074	(0.0023)	-0.00064	(0.0097)	-0.00085	(0.0004)
sum_deb_amt_2	0.001519	(<.0001)	-0.00119	(<.0001)	-0.0011	(<.0001)	-0.001	(0.0002)	-0.0012	(<.0001)
sum_deb_amt_3	0.001589	(<.0001)	-0.00166	(<.0001)	-0.00155	(<.0001)	-0.00143	(<.0001)	-0.00168	(<.0001)
sum_deb_amt_4	0.002183	(<.0001)	-0.00153	(<.0001)	-0.00137	(<.0001)	-0.00121	(<.0001)	-0.00156	(<.0001)
sum_deb_amt_5	0.002049	(<.0001)	-0.0012	(<.0001)	-0.00105	(<.0001)	-0.00089	(0.0002)	-0.00123	(<.0001)
sum_deb_amt_6	0.002314	(<.0001)	-0.00075	(0.0001)	-0.00061	(0.0023)	-0.00046	(0.0239)	-0.00078	(<.0001)
max_deb_amt_1	-0.00295	(<.0001)	0.000199	(0.4797)	0.000099	(0.7285)	-0.00001	(0.9686)	0.000215	(0.4442)
max_deb_amt_2	-0.00207	(<.0001)	0.000374	(0.1717)	0.000297	(0.2851)	0.000211	(0.4544)	0.000386	(0.157)
max_deb_amt_3	-0.00151	(<.0001)	0.001137	(<.0001)	0.001047	(0.0002)	0.000948	(0.0011)	0.001151	(<.0001)
max_deb_amt_4	-0.00193	(<.0001)	0.001082	(<.0001)	0.000947	(0.0003)	0.000808	(0.0026)	0.001105	(<.0001)
max_deb_amt_5	-0.00181	(<.0001)	0.000978	(<.0001)	0.000844	(0.0005)	0.000708	(0.004)	0.001	(<.0001)
max_deb_amt_6	-0.00239	(<.0001)	0.00072	(0.0008)	0.000575	(0.0083)	0.000428	(0.0532)	0.000744	(0.0005)
min_deb_amt_1	0.003612	(<.0001)	0.000512	(0.0367)	0.000645	(0.0095)	0.000782	(0.002)	0.000489	(0.0453)
min_deb_amt_2	0.001715	(<.0001)	-0.00003	(0.9005)	0.000065	(0.7668)	0.000156	(0.4824)	-0.00004	(0.8423)
min_deb_amt_3	0.001637	(<.0001)	-0.00002	(0.9287)	0.00007	(0.7353)	0.000158	(0.4533)	-0.00003	(0.8694)
min_deb_amt_4	0.000795	(0.0019)	-0.00041	(0.0267)	-0.00035	(0.0644)	-0.00029	(0.1354)	-0.00043	(0.0227)
min_deb_amt_5	0.00148	(<.0001)	0.000197	(0.2432)	0.000262	(0.127)	0.000328	(0.0605)	0.000186	(0.2694)
min_deb_amt_6	0.000923	(<.0001)	0.000125	(0.3645)	0.000164	(0.2404)	0.000204	(0.1507)	0.000118	(0.3894)
b_AvgOB1toMaxOB1_In	1.453313	(0.0003)	0.47561	(0.1103)	0.556922	(0.0657)	0.630547	(0.0403)	0.460767	(0.1208)
b_AvgOB2toMaxOB2_In	1.558011	(0.0001)	0.994931	(0.0008)	1.049044	(0.0005)	1.096775	(0.0003)	0.984899	(0.0008)
b_AvgOB3toMaxOB3_In	1.535759	(0.0002)	0.928429	(0.0023)	0.983894	(0.0015)	1.032815	(0.001)	0.91814	(0.0025)
b_TRmaxdeb1ToLimit_In	-2.19418	(0.0578)	-1.57719	(0.0664)	-1.74674	(0.0453)	-1.88591	(0.0332)	-1.5446	(0.0715)
b_TRmaxdeb2ToLimit_In	-1.90483	(0.0851)	0.250774	(0.7581)	0.0425	(0.959)	-0.14805	(0.86)	0.288486	(0.7224)
b_TRmaxdeb3ToLimit_In	-4.51509	(<.0001)	-1.09766	(0.1643)	-1.37557	(0.086)	-1.63463	(0.0444)	-1.04788	(0.1832)
b_TRevavgdeb1toavgOB1In	-0.1038	(0.5331)	-0.16542	(0.1803)	-0.15494	(0.2168)	-0.14569	(0.253)	-0.16735	(0.174)

Table 7.9 POS Income 6 months – Linear Regression Estimation Results

Variable	Pooled		WK		FB		WH		NL	
	Estimate	Pr > t								
b_TRavgdeb2toavgOB2ln	0.108285	(0.5165)	0.139438	(0.2545)	0.147658	(0.2352)	0.154482	(0.2218)	0.137879	(0.2585)
b_TRavgdeb3toavgOB3ln	-0.029	(0.8546)	-0.12853	(0.2703)	-0.11676	(0.3246)	-0.10576	(0.3799)	-0.13063	(0.2613)
b_TRsumdeb1tocrd1_ln	0.149861	(0.2974)	0.121187	(0.2557)	0.119032	(0.2718)	0.117633	(0.2851)	0.121647	(0.2526)
b_TRsumdeb2tocrd2_ln	-0.1374	(0.3355)	-0.08481	(0.421)	-0.09127	(0.3941)	-0.09736	(0.371)	-0.08366	(0.426)
b_TRsumdeb3tocrd3_ln	-0.09649	(0.4581)	0.045972	(0.6339)	0.038547	(0.6943)	0.030956	(0.756)	0.047225	(0.6237)
b_avgNumDeb13	0.731195	(<.0001)	-0.01189	(0.7803)	0.044872	(0.2954)	0.099816	(0.0206)	-0.02179	(0.609)
b_OB13_to_OB46ln	0.2522	(0.0062)	-0.00635	(0.9235)	0.016817	(0.8028)	0.038719	(0.5719)	-0.01046	(0.8741)
b_OB1_to_OB2_ln	0.195806	(0.2192)	0.159666	(0.1637)	0.165026	(0.1571)	0.169582	(0.153)	0.158656	(0.1651)
b_OB2_to_OB3_ln	0.192529	(0.0763)	0.179691	(0.0209)	0.182584	(0.0211)	0.184777	(0.0218)	0.179115	(0.021)
b_OB3_to_OB4_ln	-0.00162	(0.9787)	0.007994	(0.8547)	0.008734	(0.8441)	0.009146	(0.8396)	0.007829	(0.8572)
b_OB_avg_to_eop1ln	-0.33916	(0.021)	-0.41969	(0.0002)	-0.42382	(0.0002)	-0.42685	(0.0002)	-0.41886	(0.0001)
b_pos_flag_use13vs46	-3.22164	(<.0001)	-0.06108	(0.772)	-0.21164	(0.3228)	-0.36521	(0.0929)	-0.03562	(0.8654)
b_atm_flag_use13vs46	4.206632	(<.0001)	-0.51435	(0.0335)	-0.31488	(0.1997)	-0.1103	(0.6579)	-0.54803	(0.0231)
b_pos_use_only_flag_13	7.612892	(<.0001)	-0.42913	(0.0503)	-0.06776	(0.7597)	0.30062	(0.1793)	-0.49035	(0.0251)
b_atm_use_only_flag_13	-2.88585	(<.0001)	0.553366	(0.0183)	0.395786	(0.0963)	0.235011	(0.33)	0.58004	(0.0132)
b_TRsum_crd1toOB1_ln	0.229749	(0.1523)	0.146164	(0.2173)	0.152815	(0.2044)	0.158957	(0.1939)	0.144969	(0.2198)
b_TRsum_crd2toOB2_ln	0.050826	(0.7338)	-0.1271	(0.2469)	-0.1198	(0.2829)	-0.11267	(0.3204)	-0.12837	(0.2409)
b_TRsum_crd3toOB3_ln	0.019486	(0.8817)	-0.018	(0.8531)	-0.01546	(0.8756)	-0.01346	(0.8933)	-0.0185	(0.8486)
b_payment_lt_5p_1	-0.1205	(0.6163)	-0.49178	(0.0062)	-0.4767	(0.009)	-0.46199	(0.0127)	-0.49443	(0.0058)
b_payment_lt_5p_2	-0.18662	(0.4402)	-0.55114	(0.0021)	-0.53688	(0.0032)	-0.52267	(0.0047)	-0.55362	(0.0019)
b_payment_lt_5p_3	-0.01249	(0.9583)	-0.58614	(0.001)	-0.56847	(0.0016)	-0.55012	(0.0027)	-0.58912	(0.0009)
b_maxminOBavgOB1_ln	1.115153	(<.0001)	0.550279	(0.0004)	0.605786	(0.0001)	0.655634	(<.0001)	0.54012	(0.0005)
b_maxminOBavgOB2_ln	1.013658	(<.0001)	0.444675	(0.0031)	0.492999	(0.0013)	0.536954	(0.0005)	0.435891	(0.0037)
b_maxminOBavgOB3_ln	1.108759	(<.0001)	0.455535	(0.0016)	0.509524	(0.0005)	0.558734	(0.0002)	0.44573	(0.002)
b_TRsum_deb1_to_2_ln	0.040732	(0.5334)	0.01449	(0.7526)	0.015365	(0.7427)	0.016094	(0.7356)	0.01432	(0.7547)
b_TRsum_crd1_to_2_ln	0.081813	(0.4876)	-0.00857	(0.9186)	-0.00512	(0.9522)	-0.00148	(0.9864)	-0.00914	(0.9129)
I_ch1_ln	2.317187	(0.0776)	0.818812	(0.3884)	0.833954	(0.3877)	0.853429	(0.3846)	0.816598	(0.3883)
I_ch1_flag	-0.62339	(0.3848)	0.466294	(0.3631)	0.44362	(0.3952)	0.417368	(0.4319)	0.469802	(0.358)
I_ch6_flag	-1.55193	(<.0001)	-0.56799	(0.011)	-0.67665	(0.0028)	-0.77157	(0.0008)	-0.54774	(0.0139)
Age	0.077702	(<.0001)	0.124086	(0.0047)	0.119754	(0.0005)	0.115884	(<.0001)	0.124883	(0.0073)
customer_income_ln	3.570669	(<.0001)	8.851626	(<.0001)	8.291828	(<.0001)	7.799788	(<.0001)	8.955508	(<.0001)
Edu_High	0.710991	(0.0022)	1.138817	(0.1833)	1.072622	(0.1075)	1.017094	(0.0719)	1.151424	(0.2048)
Edu_Special	-0.52136	(0.0314)	-0.92587	(0.3016)	-0.90794	(0.1929)	-0.88936	(0.1323)	-0.92886	(0.3288)
Edu_TwoDegree	0.924364	(0.0642)	2.812073	(0.1278)	2.676063	(0.0626)	2.549067	(0.0364)	2.836444	(0.1478)
Marital_Civ	-0.19314	(0.5905)	-0.09605	(0.9427)	-0.07182	(0.9449)	-0.05741	(0.948)	-0.10141	(0.9429)
Marital_Div	-0.09685	(0.7135)	-0.1643	(0.8666)	-0.13187	(0.8624)	-0.10581	(0.8696)	-0.17061	(0.8694)
Marital_Sin	0.555087	(0.0333)	0.741467	(0.444)	0.763778	(0.3109)	0.777218	(0.2236)	0.736572	(0.4736)
Marital_Wid	-0.91638	(0.1631)	-1.26513	(0.605)	-1.18018	(0.5351)	-1.10832	(0.4918)	-1.28119	(0.6217)
position_Man	0.596089	(0.0204)	0.755453	(0.4283)	0.763017	(0.3038)	0.766665	(0.2227)	0.753687	(0.4565)
position_Oth	-0.32252	(0.2188)	-0.12061	(0.9006)	-0.13587	(0.8565)	-0.15103	(0.8126)	-0.118	(0.9083)
position_Tech	-1.5069	(<.0001)	-2.13091	(0.0184)	-2.06899	(0.0033)	-2.01431	(0.0007)	-2.14238	(0.0256)
position_Top	-2.14614	(<.0001)	-3.05558	(0.1245)	-3.01372	(0.0516)	-2.97409	(0.0234)	-3.06305	(0.1467)
sec_Agricult	-3.19065	(<.0001)	-4.81419	(0.0424)	-4.73331	(0.0104)	-4.65058	(0.003)	-4.82781	(0.0551)
sec_Constr	0.997211	(0.1357)	1.017228	(0.6767)	1.052863	(0.5792)	1.078618	(0.5027)	1.009941	(0.6964)
sec_Energy	-1.97811	(<.0001)	-2.15029	(0.143)	-2.14983	(0.0598)	-2.14707	(0.0265)	-2.15012	(0.1675)
sec_Fin	0.744777	(0.0045)	2.095365	(0.0285)	1.943657	(0.0092)	1.810648	(0.0042)	2.123561	(0.0364)
sec_Industry	-1.16178	(0.1865)	-0.70728	(0.8274)	-0.75343	(0.7654)	-0.7942	(0.7104)	-0.69875	(0.8392)
sec_Manufact	0.101362	(0.8773)	-0.99453	(0.6809)	-0.95686	(0.6111)	-0.91389	(0.5667)	-1.00033	(0.6967)
sec_Mining	-1.90572	(<.0001)	-2.69578	(0.1368)	-2.67037	(0.0582)	-2.64168	(0.027)	-2.69974	(0.1603)
sec_Service	-0.73702	(0.0008)	-0.54696	(0.5011)	-0.57078	(0.3669)	-0.59107	(0.2703)	-0.54246	(0.5296)
sec_Trade	-0.41385	(0.1086)	-0.0893	(0.9254)	-0.12272	(0.8686)	-0.15184	(0.8091)	-0.08306	(0.9345)
sec_Trans	0.222185	(0.7314)	0.147435	(0.9511)	0.156785	(0.9332)	0.167933	(0.9157)	0.146079	(0.9544)
car_Own	-0.01382	(0.9506)	0.916784	(0.2654)	0.852859	(0.1831)	0.793317	(0.144)	0.92826	(0.288)
car_coOwn	1.673402	(<.0001)	2.414495	(0.0811)	2.377673	(0.0273)	2.341412	(0.0103)	2.420877	(0.0993)

Table 7.9 POS Income 6 months – Linear Regression Estimation Results

Variable	Pooled		WK		FB		WH		NL	
	Estimate	Pr > t								
real_Own	0.652033	(0.0012)	0.918841	(0.2179)	0.909218	(0.1171)	0.898144	(0.0678)	0.920314	(0.2449)
real_coOwn	0.414044	(0.037)	0.259337	(0.7246)	0.273345	(0.6332)	0.285729	(0.556)	0.256737	(0.7424)
reg_ctr_Y	-2.36859	(<.0001)	-3.1327	(0.0008)	-3.09025	(<.0001)	-3.04892	(<.0001)	-3.14011	(0.0015)
reg_ctr_N	-3.78955	(<.0001)	-5.3446	(<.0001)	-5.25387	(<.0001)	-5.16544	(<.0001)	-5.36042	(<.0001)
child_1	0.177353	(0.4776)	-0.04986	(0.9573)	-0.02157	(0.9763)	0.002723	(0.9965)	-0.05519	(0.9555)
child_2	-0.06476	(0.6753)	-0.31379	(0.5847)	-0.2839	(0.5251)	-0.25759	(0.4961)	-0.31933	(0.6002)
child_3	-2.69971	(<.0001)	-2.55734	(0.2661)	-2.54744	(0.1548)	-2.53997	(0.0943)	-2.55932	(0.2943)
Unempl_InyoY_1	0.035535	(0.9921)	-0.29373	(0.9095)	-0.12135	(0.9632)	0.012229	(0.9963)	-0.32786	(0.8988)
UAH_EURRate_Inmom_1	-11.4444	(0.0079)	-7.99451	(0.008)	-8.17763	(0.0077)	-8.35796	(0.0075)	-7.96283	(0.008)
UAH_EURRate_InyoY_1	0.501516	(0.8487)	1.563732	(0.3993)	1.536134	(0.416)	1.505267	(0.4337)	1.568146	(0.3965)
CPI_Inqoq_1	-49.7415	(<.0001)	-37.8649	(<.0001)	-39.0081	(<.0001)	-40.0428	(<.0001)	-37.6563	(<.0001)
SalaryYear_InyoY_1	2.229274	(0.7044)	-3.80628	(0.3622)	-3.41472	(0.4218)	-3.03847	(0.4825)	-3.87518	(0.3521)
s_cons_full	-0.16889	(0.3863)	-0.20888	(0.161)	-0.21295	(0.1595)	-0.21566	(0.1606)	-0.20803	(0.1616)
s_month_since_NA_full	0.806958	(<.0001)	-0.06459	(0.5919)	-0.03163	(0.7958)	0.003005	(0.9807)	-0.07006	(0.56)
s_month_since_Tr_full	1.199013	(<.0001)	0.912417	(<.0001)	0.936257	(<.0001)	0.958671	(<.0001)	0.90818	(<.0001)
s_month_since_Re_full	0.537111	(0.0099)	0.54074	(0.0007)	0.545559	(0.0007)	0.549691	(0.0008)	0.539842	(0.0007)
s_month_since_RP_full	0.262324	(0.1713)	0.152338	(0.2932)	0.165753	(0.2603)	0.177582	(0.2351)	0.149864	(0.2998)
s_month_since_D1_full	0.471895	(0.0213)	0.312537	(0.0488)	0.331104	(0.0398)	0.347192	(0.0336)	0.309077	(0.0508)
s_month_since_D2_full	0.150248	(0.5632)	-0.04941	(0.8006)	-0.0396	(0.8421)	-0.03024	(0.881)	-0.05115	(0.7932)
s_times_NA_full	3.242604	(<.0001)	1.492361	(<.0001)	1.610968	(<.0001)	1.726316	(<.0001)	1.471706	(<.0001)
s_times_TR_full	3.153459	(<.0001)	3.48941	(<.0001)	3.550579	(<.0001)	3.598056	(<.0001)	3.477448	(<.0001)
s_times_RE_full	1.836206	(<.0001)	1.383159	(<.0001)	1.433565	(<.0001)	1.4777	(<.0001)	1.373801	(<.0001)
s_times_RP_full	2.845491	(<.0001)	1.669929	(0.0001)	1.78279	(<.0001)	1.886521	(<.0001)	1.649575	(0.0001)
s_times_D1_full	0.370538	(0.349)	0.215944	(0.4994)	0.222637	(0.4917)	0.226959	(0.4888)	0.214492	(0.5013)
s_times_D2_full	-0.70809	(0.4072)	-0.00204	(0.9975)	-0.02403	(0.9711)	-0.05119	(0.9393)	0.001127	(0.9986)
d_StateFull_1_NA	0.142227	(0.9493)	1.307431	(0.431)	1.203796	(0.4757)	1.098436	(0.5219)	1.324844	(0.4236)
d_StateFull_1_Tr	-1.57948	(0.4641)	3.023116	(0.0602)	2.839309	(0.0824)	2.649077	(0.1108)	3.053946	(0.0569)
d_StateFull_1_Re	-4.18555	(0.0051)	-0.41069	(0.7112)	-0.64278	(0.5687)	-0.86787	(0.4489)	-0.37015	(0.7379)
d_StateFull_1_RP	-2.86488	(0.1549)	1.743598	(0.2445)	1.533918	(0.3137)	1.321861	(0.393)	1.779302	(0.2336)
d_StateFull_1_D1	-2.65288	(0.0576)	0.14816	(0.8865)	-0.00171	(0.9987)	-0.1508	(0.8882)	0.173917	(0.8666)

Behavioural characteristics

As we use behavioural characteristics for the monthly panel data the use of longer periods for the behavioural characteristics, which are based on several months, will give high multicollinearity and overlapping observations. For example, the characteristic ‘an average outstanding balance for 6 months’ period is likely not to have significant changes for the next month (time series slice), because even if the balance changes exactly in the next month, this has only a 1/6 weight for the characteristic value. To avoid the overlapping of characteristics based on the several months we use only monthly behavioural characteristics for a set of basic factors. The first category of characteristics is primary or original factors such as the average monthly utilization rate, the average monthly outstanding balance, the average monthly debit turnover (purchase transactions), the minimum and maximum monthly

debit turnover (purchase transactions), the sum of monthly debit turnover (purchase transactions), the sum of monthly credit turnover (payment transactions). We use the values of these characteristics for the previous 1 to 6 months. The second category of characteristics is factors derived from primary as, for example, b_AvgOB1toMaxOB1_ln, b_AvgOB2toMaxOB2_ln, b_AvgOB3toMaxOB3_ln – the logarithm of the average outstanding balance to the maximum outstanding balance in the previous month 1, 2 and 3 respectively. We use the following variables derived from original characteristics on a monthly basis only: logarithm of maximum debit transaction to the credit limit (b_TRmaxdeb1ToLimit_ln), logarithm of average debit transaction to average outstanding balance (b_TRavgdeb1toavgOB1ln), logarithm of sum of debit transactions to sum of credit transactions (b_TRsumdeb1tocrd1_ln), logarithm of sum of credit transactions to outstanding balance (b_TRsum_crd1toOB1_ln), logarithm of the difference between maximum and minimum balance to credit limit (b_maxminOB_limit_1_ln), and logarithm of the difference between maximum and minimum balance to average balance (b_maxminOBavgOB1_ln). We investigate how these predictors from the different periods impact on the dependent variable, and total amounts of the characteristics like the outstanding balance or total sum of monthly ratios are reflected in the outcome as the additive regression function is used. However, the changes between periods are not investigated in this approach. So we have also included characteristics, which reflect changes in some factors such as logarithm of utilisation rate in the last month to utilisation rate in the previous month (b_UT1to2ln), logarithm of number of debit transactions in the last three months to the number of debit transactions in months 4-6 (b_NumDeb13to46ln), the flag of POS transactions for the last 3 months and no POS transactions in months 4-6 (b_pos_flag_use13vs46), the flag of ATM transactions for the last 3 months and no ATM transactions in months 4-6 (b_atm_flag_use13vs46), the logarithm of average outstanding balance for the lasrt 3 months to the average outstanding balance in months 4-6 (b_OB13_to_OB46ln). We have selected one month as a basic observation and performance window and six months' period as a maximum observation and performance window.

In our terminology under the ‘deeper month(s)’ we mean month(s), which has (have) longer period to the current or mentioned months. For example, if we talk about a

characteristic at the current month X1, say July, the characteristics at deeper months mean X2, X3 etc., i.e. June, May etc. According to this concept the first month is the current month, or the observation point, the second month is the previous to the current month, and so on.

Current outstanding balance (avg_balance_1) is not a significant characteristic for the pooled method, but it is significant for other random effect methods. However, for previous months 2 and 3 the average outstanding balance is significant for the pooled method and loses significance for random effect estimations. The current month coefficient has a positive sign, which means that higher average outstanding balance in the current month causes higher income amount in the next 6 months, but for months 2-6 of the observation period the trend is the opposite for the random effect estimations and varied for the pooled method. The average outstanding balance at the month, which is 6 months before the observation point, is still a significant characteristic for the pooled regression, but the random effect shows that months with deeper history are not significant for the prediction – p-level values from 0.447 for WH to 0.99 for WK methods.

The amount of debit transaction (sum_deb_1, sum_deb_2 etc.) are significant for all months, except the first one, under the pooled method and it is less significant for lagged months for the random effect methods. For the pooled estimation, the correlation between income amount and purchases amount is positive, but for all random effect estimations, the correlation is negative. The negative trend for the random effect looks illogical, because it is expected that more purchases generate more income, and can be as an issue for the following investigation.

The sum of credit transaction – credit payments (sum_crd_1, sum_crd_2 etc.) – has mainly positive signs and is significant for the majority of periods and methods, except the pooled method after 3 months. It looks like higher payments in previous months cause higher amount income. However, it can be explained by the behaviour of a customer, who paid more in the past because of high spending, and such a tendency will be kept in the future. So the high payments are evidence of high spending in the past.

We have also included minimal and maximum spending monthly amount into the model. The coefficients are significant and positive for the pooled effect model for all months, but are insignificant for a random effect model.

It is interesting that maximum debit transaction to credit limit for a 2nd month ($b_{TRmaxdeb2ToLimit_ln}$) is insignificant for all random effect methods but is significant at 10% for the pooled effect method. The characteristics for the first and the third months ($b_{TRmaxdeb1ToLimit_ln}$, $b_{TRmaxdeb3ToLimit_ln}$) are significant both for pooled and random effect estimations.

We use not only interval and categorical variables as income predictors. For the average number of debit transactions during the observed period of 1-3 months ($b_{avgNumDeb13}$) the coefficient is significant for pooled, and WH random effect models and has a positive correlation with predicted income. But the relationship is insignificant for the WK, FB, and NL random effect models. For the pooled method each transaction in the observed period gives an additional 0.73 money units for predicted income.

We have included several characteristics based on a 6 months' period to compare the first 3 months' value with the next 3 months. For instance, the logarithm of the sum of the outstanding balance for 1-3 months to the sum of the outstanding balance for 4-6 months ($b_{OB13_to_OB46ln}$). The pooled estimation has positive and significant coefficients 0.2522 (p-level 0.0062), but all random effect methods give insignificant estimates and negative signs (FB -0.00635 (p-level 0.9235) and NL -0.01046 (p-level 0.8741) indicate a negative relationship between quarterly changes of the outstanding balances and POS income amount.

The same correlation is related to the ratio of the number of spending debit transactions in months 1-3 to the number of spending debit transactions in months 4-6 ($b_{NumDeb13to46ln}$). The pooled method gives a significant negative coefficient - 0.32936 (0.0002). However, all random effect estimations are insignificant, and WK, FB, and NL method give positive coefficients.

If a credit card holder made only POS transactions during the last 3 months ($b_{pos_use_only_flag_13}$) this resulted in a high additional 7.61 money units for the total POS income according to the pooled method. However, WK, FB, and NL random

effect have shown negative coefficients close to zero, and FB the method gave an insignificant coefficient.

If a credit card holder made only ATM transactions during the last 3 months (b_atm_use_only_flag_13) this resulted in a decrease of 2.86 money units for the total POS income estimation according to the pooled method. However, WK, FB, and NL random effect have shown positive coefficients close to zero.

Application characteristics

We have included application variables that are fixed over time both into the pooled and random effect methods. When static variables are included, the Hausman test cannot be calculated for the random effect. Thus the fixed effect model should be used. However, the fixed effect considers the estimation of the separate Cross-Sectional Effect either for each account (for the account-level model) or each segment (for the pool-level model). This means that the estimated regression model can be used for the same set of accounts or the same segments, but at the pool level. However, we would like to build the income estimation model applicable at account level for any account with the appropriate set of characteristics. We have tested the model with only a set of time-varying characteristics to keep Hausman test successful, but the predictive accuracy results (R-squared, RMSE, MAE) were at the same level or worse than with time-stationary application characteristics. We kept the time-stationary characteristics for the comparative analysis purposes to test how the random-effect methods change the significance of the application characteristics in comparison with the polled regression method.

From all application characteristics, only two: age and the logarithm of the ratio of customer monthly income to average income of all customers (customer_income_ln) are significant for both pooled and random-effect methods. For other factors, we use dummy, or binary, variables. For example, customer education is presented as a set of binary characteristics equal to the number of education categories minus one. For current binning of education characteristic these are high education (Edu_High), special professional education (Edu_Special), two high educations or doctoral degree (Edu_TwoDegree), and simple secondary education, which is not at the list of

characteristics, but the coefficient for this category of education is equal to zero, and the income estimation for secondary education is included into intercept.

As it can be seen from Table 7.9 the majority of application dummy characteristics are significant for the pooled method only, but are not significant for random-effect estimation. From the random-effect methods, WH gave higher number of coefficients with low p-values than the WK, NL, and FB methods. Despite the difference in the significance of the predictors the signs of coefficients are the same for all methods, and hence the marginal effects of the characteristics are consistent in sign. For example, the highest income is generated by the two-degrees category – these customers in average give an additional 0.92 money units according to pooled estimation and 2.55 money units according to WH random-effect estimation to the total POS income in comparison with secondary education. On the other hand, the Special education category gives -0.52 money unit according to the pooled estimation and -0.89 money units according to WH random-effect model in comparison with secondary education. The difference between the highest and the lowest value is not large – for pooled effect $(2.55 - (-0.52)) = 3.07$ for a customer with two-degrees in comparison with a customer with special education.

Marital status is an insignificant characteristic for all methods of estimation. The highest income is generated by single customers – plus 0.55 money units in comparison with married customers according to the pooled effect estimation.

It is unexpected that from the position categories the lowest income is generated by top managers - 2.15/-2.97 money units (pooled/WH) and the highest income – by managers +0.60/+0.76 money units (pooled/WH) in comparison with customers with an employee position. Generally, top managers have higher credit limits than other categories of credit card holders, and it is expected that they should generate more income than others. However, as it can be seen from the utilisation rate analysis (see Chapter 3) top managers do not use credit cards so actively as, for example, employees do. These can be explained that top managers have enough money for consumer spending and try to use debit cards.

The sector of the economy is an insignificant factor, especially for the random effect. The pooled effect model suggests customers from the construction sector are the most

profitable 0.99 and customers from the agriculture sector are the least profitable -3.19. This was expected, because in the sample agriculture customers have very poor access to the usage of credit cards in any other way than cash withdrawal at ATMs or via bank branches.

It is also expected and is confirmed by results that real estate owners and car owners generate higher income than other categories, but these predictors are not significant for current estimations.

State characteristics

We have included the current state (at the observation point) and characteristics derived from the state's history as predictors into the models.

The current state of the account is not a significant characteristic for the income amount prediction. Only the revolver state for pooled effect with a coefficient -4.1855 and transactor state for the random effect model with coefficients from 2.649077 to 3.053946 have p-level values close to zero.

On the other hand, the characteristics based on states dynamics are predictive. Number of months since a specific state is significant both for pooled and random-effect methods. It gives quite small income values per one month, but the income amount increases depending on the number of months since the certain state. The number of the month since being in the transactor state is the weightiest factor from among these characteristics with 1.2 money units per month and number of months since being in the repaid revolver state with 0.26 has a smaller estimated parameter. The coefficient given by random-effect methods are also significant.

The number of times in the certain state for whole period of observation (s_times_NA, s_times_TR, etc) is an ordinal characteristic. The pooled method gives significant coefficient estimates for all states, except the number of delinquencies for 1 month and delinquency for two months (D2). The highest marginal income – 3.24 and 3.15 money units have a number of inactive and transactor states. However, if we look at random effect estimation, only the number of times in the transactor state is a significant predictor and has values close to the pooled method (around 3.5) for all random effect methods.

Macroeconomic characteristics

We included macroeconomic characteristics into the regression along with other features. Macroeconomic characteristics have the same values for all observations at each period, so for the panel data where the time component is a calendar month, each macroeconomic characteristic is constant for each account at time t. However, we use 20 periods for panel data and respectively 20 cross-sections are used for the regression model. So, we investigate how these portfolio-level time-series characteristics impact on the general model estimates for pooled and random-effect methods.

The change of Consumer price index to the previous quarter (CPI_Lnqoq_1) is the only characteristic which is significant for all methods of estimation. The logarithm of the month-to-month change of the current local currency to EURO exchange rate (UAH_EURRate_Lnmom_1) does not have a noticeable difference between the pooled and random-effect method but is less significant. Both characteristics have negative signs for all coefficients. This can be interpreted as any increase of CPI and of the exchange rate of the local currency to EURO will decrease the expected amount of the non-interest income from POS transactions. However, it is not concern with the theory of rational expectations when people start spending more money than in stable times. The unemployment rate, salary changes and year-to-year currency exchange rate are not significant characteristics for all methods.

7.5.2 Comparative analysis of the goodness-of-fit of the pooled and random effect models for POS income

We compare the R-squared, RMSE, and MAE values to select the regression method with the best fitting results for 6 months income prediction. As it can be seen from the Table 7.10, the pooled method gives the highest R-square - 0.2692/0.2608 (training/validation data) and the lowest error coefficients: RMSE - 11.14/11.43 and MAE - 5.96/6.07 for both the development/validation samples. Among the random effect methods, the WH (Wallace and Hussain Variance Component) with two-way (cross-sectional and time series) shows the best fitting results: R-square - 0.1435/0.1587 and the lowest error coefficients: RMSE - 12.06/12.17 and MAE -

6.57/6.62 for both development/validation samples (see Table 7.10), and the validation results for WH one-way approach are close to two-way one.

Table 7.10 Assessing the fit of One-stage 6 months income model for full (positive POS transaction and zero income) data sample

Model description Conditional equation	POS Income amount - direct estimation POS Sum 6 ALL						
	Development			Validation			
	Coefficient	R^2	RMSE	MAE	R^2	RMSE	MAE
Regression Type							
Pooled	POOLED	0.2692	11.1399	5.9627	0.2608	11.4229	6.0669
Random effect - One-way - Overall							
Wansbeek and Kapteyn Variance Components	RAN1_WK	0.1017	12.3700	6.8111	0.1345	12.4928	6.8593
Fuller and Battese Variance Components	RAN1_FB	0.1208	12.2239	6.6977	0.1076	12.3376	6.7464
Wallace and Hussain Variance Component	RAN1_WH	0.1435	12.0637	6.5760	0.1587	12.1702	6.6264
Nerlove	RAN1_NL	0.0962	12.4146	6.8462	0.1136	12.5405	6.8941
Random effect - Two-ways - Overall							
Wansbeek and Kapteyn Variance Components	RAN2_WK	0.1020	12.3676	6.8093	0.1346	12.4902	6.8574
Fuller and Battese Variance Components	RAN2_FB	0.1209	12.2234	6.6954	0.1063	12.3372	6.7440
Wallace and Hussain Variance Component	RAN2_WH	0.1435	12.0632	6.5711	0.1587	12.1698	6.6211
Nerlove	RAN2_NL	0.0943	12.4291	6.7969	0.1140	12.5505	6.8426

The comparative analysis of regression models has shown that there is almost no difference between the one-way and the two-way random effect models. So, the random effect from time component does not impact on the linear regression results and only the random effect of the cross-sectional variance is important.

The development (training) and validation (out-of-sample) data sets show similar results, so i) the regression model is not overtrained, ii) the regression model is stable and can be used for estimation from the current data sample.

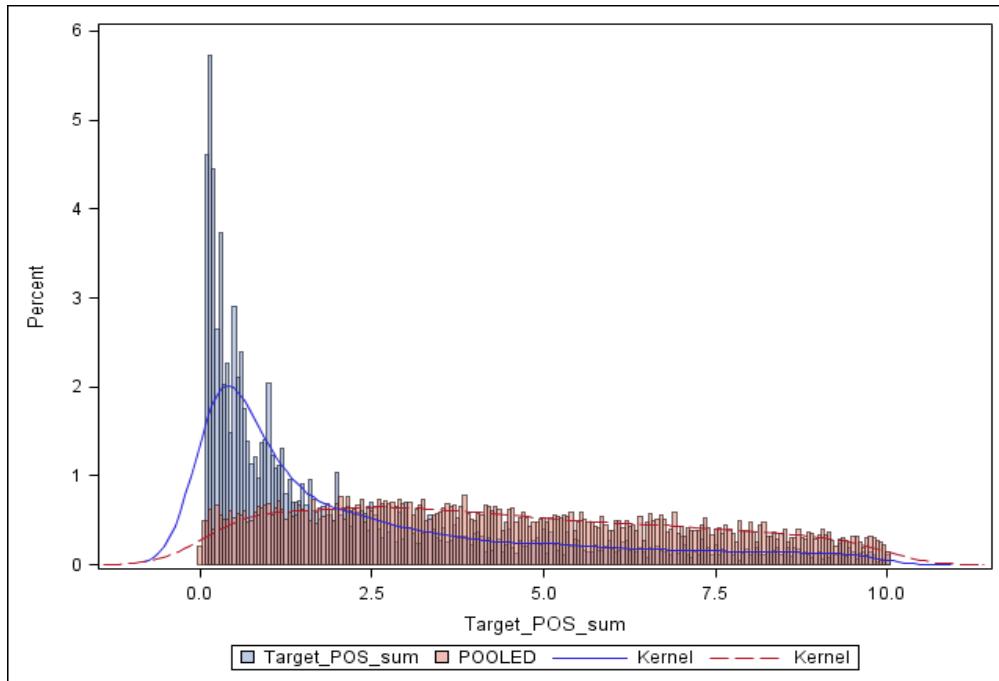
Table 7.11 Assessing the fit of POS income model - second stage: conditional on positive POS transaction (POS Sum for 6 month > 0)

Model description Conditional equation	POS Income conditional on positive POS transactions OLS: POS Sum 6 POS Sum 6 > 0						
	Development			Validation			
	Coefficient	R^2	RMSE	MAE	R^2	RMSE	MAE
Regression Type							
Pooled	POOLED	0.2909	15.6793	10.6822	0.2962	15.8619	10.8615
Random effect - One-way							
Wansbeek and Kapteyn	RAN1_WK	0.1633	17.0360	11.7869	0.1518	17.4598	12.3974
Fuller and Battese	RAN1_FB	0.1784	16.8774	11.6539	0.1671	17.2891	12.2507
Wallace and Hussain	RAN1_WH	0.1925	16.7335	11.5346	0.1816	17.1320	12.1148
Nerlove	RAN1_NL	0.1606	17.0648	11.8113	0.1491	17.4905	12.4237
Random effect - Two-ways							
Wansbeek and Kapteyn	RAN2_WK	0.1645	17.0245	11.7472	0.1533	17.4372	12.3462
Fuller and Battese	RAN2_FB	0.1789	16.8726	11.6452	0.1676	17.2828	12.2408
Wallace and Hussain	RAN2_WH	0.1931	16.7279	11.5150	0.1823	17.1216	12.0906
Nerlove	RAN2_NL	0.1616	17.0559	11.7568	0.1507	17.4640	12.3492

We compare two methods of aggregation of the models into a two-stage model. The first method is to use the two-stage method. The first stage is a logistic regression for the probability of the POS transaction. The second stage is a linear regression for the POS income amount from positive transactions only. The second method is to try to predict with a one-stage linear regression where we use full data sample with both POS transactions and zero income during the performance window for POS income estimation. As can be seen from Table 7.11, the pooled method gives the highest R-square - 0.2909/0.2962 and the lowest error coefficients: RMSE - 15.6793/15.8619 and MAE - 10.6822/10.8615 for both the development/validation samples. Among the random effect methods, the WH (Wallace and Hussain Variance Component) with two-way (cross-sectional and time series) shows the best fitting results: R-square - 0.1931/0.1823 and the lowest error coefficients: RMSE - 16.7279/17.1216 and MAE - 11.5150/12.0906 for both development/validation samples (see Table 7.11). For the two-stage model, the linear regression for amount shows better results than for the one-stage model. The test sample R-squared for the pooled method is equal to 0.2963 instead of 0.2608 for the one-stage model (Table 7.10). This may be because we concentrate on active accounts only. The WH method also showed the best result among random-effect models – 0.1816 and 0.1823 for one and two-way methods.

The distributions of observed vs predicted POS income values for the range (0,10] shows that we have an overestimation of the income amount (Figure 7.12). The predicted distribution is flat in comparison with the distribution of observed values. But with low R-squared values, the accuracy of the models for business implementation is questionable.

Figure 7.12 The distributions of observed vs. predicted POS 6 months income values for the range (0,10]



For the second-stage model (Table 7.12 and Table 7.13) we test two further options. The first option is considered the use of a threshold for the probability of a transaction. For the probabilities under the threshold the income amount is considered as zero. For the probabilities above the threshold the income amount is calculated according to the POS income estimation linear model. The second option means that we use the probability of transaction as a weight for the amount estimated with linear regression (POS Sum X Pr(POS > 0)).

Table 7.12 Assessing the Fit of Two-stage 6 months income model result – Option 1: Non-zero income condition is Pr(POS) > 0.5

Model description Conditional equation	POS Income Amount: Option 1 (POS Sum Pr(POS>0), Pr(POS>0) >= 0.5; 0, Pr(POS>0) < 0.5)						
	Development			Validation			
	Coefficient	R^2	RMSE	MAE	R^2	RMSE	MAE
Regression Type							
Pooled		0.26576	15.5863	7.17152	0.27374	16.2713	7.42932
Random effect - One-way							
Wansbeek and Kapteyn		0.12630	17.0021	8.17317	0.14944	17.6088	8.38949
Fuller and Battese		0.13964	16.8719	8.08361	0.16219	17.4763	8.30305
Wallace and Hussain		0.15201	16.7501	8.00099	0.17396	17.3531	8.22348
Nerlove		0.12392	17.0253	8.18929	0.14714	17.6326	8.40504
Random effect - Two-ways							
Wansbeek and Kapteyn		0.1271	16.9948	8.1771	0.1501	17.6021	8.3922
Fuller and Battese		0.1401	16.8678	8.0815	0.1626	17.4723	8.3010
Wallace and Hussain		0.1524	16.7460	8.0018	0.1743	17.3494	8.2240
Nerlove		0.1244	17.0210	8.1993	0.1475	17.6291	8.4128

Table 7.13 Assessing the Fit of Two-stage 6 months income model result – Option 2: POS Sum X Pr(POS > 0)

Model description Conditional equation	POS Income Amount: Option 2 POS Sum X Pr(POS > 0)						
	Development			Validation			
	Coefficient	R^2	RMSE	MAE	R^2	RMSE	MAE
Regression Type							
Pooled		0.3039	15.1762	7.3214	0.30779	15.8853	7.61005
Random effect - One-way							
Wansbeek and Kapteyn		0.1793	16.4787	8.5385	0.19608	17.1192	8.79218
Fuller and Battese		0.1909	16.3611	8.4428	0.20735	16.9988	8.69906
Wallace and Hussain		0.2018	16.2509	8.3526	0.21776	16.8868	8.61174
Nerlove		0.1772	16.4996	8.5555	0.19405	17.1408	8.80873
Random effect - Two-ways							
Wansbeek and Kapteyn		0.1803	16.4683	8.5396	0.19702	17.1092	8.79121
Fuller and Battese		0.1914	16.3567	8.4406	0.20775	16.9945	8.69668
Wallace and Hussain		0.2023	16.2454	8.3525	0.21825	16.8815	8.61068
Nerlove		0.1782	16.4898	8.5605	0.19491	17.1317	8.81041

Income as a proportion of the outstanding balance – indirect estimation – shows a high fit for the logarithm of the rate prediction model – R-squared around 0.55/0.52 for development/validation sample accordingly. However, after rates transformation to the income amount the model accuracy drop down because of use of exponent for the transformation – R-squared close to the same values as for direct model – 0.28/0.26. High values give extremely high residual values. However, for low values – close to zero – this model can give high predictive accuracy.

For all estimated models for POS income prediction: direct and two-stage for pooled and random-effect methods, all the R-squared, MAE, and RMSE, computed for the

training, and test samples are close and differ insignificantly (for example, R-squared equal to 0.26 and 0.27 for train and test samples for the two-stage model in Table 7.12). Thus, overfitting is not observed for estimated models for POS income.

7.6 Estimation of income from ATM cash withdrawals

For the ATM income model, the results are similar to the POS estimation model. As can be seen from Table 7.14, the pooled method gives the highest R-square - 0.28/0.27 and the lowest error coefficients: RMSE - 103/106 and MAE - 73/74 for both the development/validation samples. Among the random effect methods, the WH (Wallace and Hussain Variance Component) with two-way (cross-sectional and time series) shows the best fitting results: R-square - 0.152/0.149 and the lowest error coefficients: RMSE - 122.8/114.9 and MAE - 80.6/82.7 for both development/validation samples, and the validation results for WH one-way approach are close to the two-way approach.

Table 7.14 Assessing the fit of One-stage ATM model for full (positive ATM transaction and zero income) data sample

Model description Conditional equation	ATM Income amount - direct estimation					
	Development			Validation		
Coefficient	R^2	RMSE	MAE	R^2	RMSE	MAE
Regression Type						
Pooled	POOLED	0.2832	103.821	72.948	0.27357	106.147
Random effect - One-way						
Wansbeek and Kapteyn	RAN1_WK	0.02778	120.912	87.477	0.02176	123.177
Fuller and Battese	RAN1_FB	0.09987	116.342	83.843	0.09488	118.485
Wallace and Hussain	RAN1_WH	0.15417	112.779	80.985	0.14966	114.843
Nerlove	RAN1_NL	0.01824	121.503	87.946	0.01207	123.786
Random effect - Two-ways						
Wansbeek and Kapteyn	RAN2_WK	0.01	221.123	180.879	0.01	222.279
Fuller and Battese	RAN2_FB	0.0999	116.335	83.413	0.0961	118.405
Wallace and Hussain	RAN2_WH	0.1526	112.881	80.583	0.1492	114.871
Nerlove	RAN2_NL	0.01	224.89	184.313	0.01	226.023

We compare two methods of aggregation of the models into a two-stage model. The first method is to use a two-stage method. The first stage is a logistic regression for the probability of the ATM transaction. The second stage is a linear regression for the ATM income amount from positive transactions only. The second method is to try to predict with a one-stage linear regression where we use the full data sample with both ATM transactions and zero income during the performance window for ATM income

estimation. As can be seen from Table 7.15, the pooled method gives the highest R-square - 0.249/0.234 for both the development/validation samples. From the random effect methods, the WH (Wallace and Hussain Variance Component) with one-way shows the best fitting results: R-square - 0.15/0.12 for both development/validation samples.

Table 7.15 Assessing the fit of ATM income model - second stage: conditional on positive POS transaction (POS Sum 6 month > 0)

Model description Conditional equation	ATM Income conditional on ATMitive ATM transactions OLS: ATM Sum 6 ATM Sum 6 > 0					
	Development			Validation		
Coefficient	R^2	RMSE	MAE	R^2	RMSE	MAE
Regression Type						
Pooled	POOLED	0.2496701	91.828	69.116	0.2347403	94.649
Random effect - One-way						
Wansbeek and Kapteyn	RAN1_WK	0.1031373	102.36	77.494	0.0742999	107.761
Fuller and Battese	RAN1_FB	0.1328894	99.458	75.317	0.1015314	104.29
Wallace and Hussain	RAN1_WH	0.1555514	97.696	73.98	0.1238618	102.137
Nerlove	RAN1_NL	0.0993132	102.806	77.827	0.0709636	108.287
Random effect - Two-ways						
Wansbeek and Kapteyn	RAN2_WK	0.0384748	216.207	179.651	0.0214271	216.842
Fuller and Battese	RAN2_FB	0.1269283	99.754	75.093	0.100109	104.324
Wallace and Hussain	RAN2_WH	0.145527	98.306	73.967	0.1186458	102.506
Nerlove	RAN2_NL	0.0394181	219.477	182.609	0.0329858	220.028

The estimates of pooled and random effect regression are given in Appendix 1.

7.7 The comparative analysis of pooled and random effect regression coefficient estimates for POS and ATM transactions income

We compare the estimated coefficients for both the pooled and random effect methods for POS and ATM income prediction (see Table 7.16). In this analysis, we focus on the sign of the coefficient to find the similar and opposite trends. Similar signs indicate that the same behaviour of the cardholder leads to an increase or decrease in the POS and ATM transactional income when the covariate increases. Thus the factors, which decrease both incomes from POS and ATM simultaneously, can also be managed together in the same manner.

For example, the logarithm of the ratio of maximum purchase (debit) transaction to the credit limit in months 1 (2,3 respectively) - b_TRmax_deb1(2,3)_To_Limit_In – has a negative sign for both POS and ATM transactions. This means that if the

maximum transaction amount is close to the credit limit, the logarithm of the ratio is close to zero and the income amount is also close to zero. On the other hand, for low ratio values, the logarithm has negative values, and its multiplication by negative coefficients give a high positive increase in the predicted income. This relation can be explained as that for credit cards with a high level of the credit limit utilisation the cardholder has few chances to repay sufficient amount and make significant purchases or cash withdrawals to generate POS or ATM income. On the other hand, if credit cardholder has a large unused credit limit (s)he has higher chances to make money spending in the future during the performance (predicted) period. So, if we would like to motivate customers to generate high transactional income for the bank in future, we need to prompt them to keep unused credit limit now. However, it can decrease the income from the interest rate. Moreover, it is the analysts choice as what is bank's strategy at the current and predicted periods, and what is more profitable: interest rate income or transactional income.

Another example of both negative signs is the current revolver state (d_StateFull_1_Re), and negative signs can be explained by the same factor as for the ratio of maximum purchase (debit) transaction to the credit limit. A revolver has a low unused credit limit. On the other hand, transactors and non-active customers can generate high transactional income in the future, but not the interest rate income, which is currently made by revolvers.

The opposite trends of POS and ATM transactional income mean that credit cardholders have behavioural patterns, which work asymmetrically for total transactional income. For example, High Education segment has a positive relationship with POS and an negative are with ATM. We can assume that cardholders with high education i) are more literary than cardholders with secondary education, and ii) have easy access to a POS terminal, for example, because they mainly live in cities.

We can observe cases with opposite signs for pooled and random effect estimates. For example, the maximum sum of debit transaction in the current month (max_deb_amt_1) has a negative sign for POS and a positive sign for ATM income for pooled model. However, the estimates for this covariate have opposite signs for

random-effect model: a positive sign for POS and a negative sign for ATM income. In both cases the estimated coefficients are significant (p-level close to zero). The difference in signs for various methods of coefficient estimates can be caused by the correlation between covariates which is considered in different ways: random effect considers the variance with between effect estimation method, and pooled regression considers each observation – several time observations for the same individual – as an independent observation.

Table 7.16 Comparative analysis of pooled and random effect regression coefficient estimations for POS and ATM income

Variable	Pooled						Random effect - WK					
	POS		ATM		Signs	POS		ATM		Signs		
	Estimate	Pr > t	Estimate	Pr > t		Estimate	Pr > t	Estimate	Pr > t		Estimate	Pr > t
Intercept	-13.2580	(0.0003)	-110.2630	(0.0906)	POS - ATM-	2.0658	(0.5533)	104.8890	(0.0848)	POS + ATM+		
mob	0.0277	(0.3116)	2.4124	(<.0001)	POS + ATM+	0.2374	(<.0001)	4.6455	(<.0001)	POS + ATM+		
limit	0.0002	(0.0002)	0.0007	(0.0006)	POS + ATM+	0.0000	(0.363)	0.0001	(0.7236)	POS - ATM+		
UT0_1	-6.6948	(<.0001)	-66.2825	(<.0001)	POS - ATM-	-5.9711	(<.0001)	-40.8486	(<.0001)	POS - ATM-		
UT0_2	3.6699	(0.0179)	45.3828	(<.0001)	POS + ATM+	0.6736	(0.5397)	22.1332	(<.0001)	POS + ATM+		
UT0_3	2.4461	(0.0673)	41.3497	(<.0001)	POS + ATM+	0.5815	(0.5358)	16.8217	(<.0001)	POS + ATM+		
UT0_4	-0.3091	(0.8108)	22.8843	(<.0001)	POS - ATM+	-0.5849	(0.5179)	7.5689	(0.059)	POS - ATM+		
UT0_5	0.1039	(0.933)	-1.4798	(0.7635)	POS + ATM-	-0.3080	(0.7219)	-3.8061	(0.3107)	POS - ATM-		
UT0_6	-1.5758	(0.0779)	17.3266	(<.0001)	POS - ATM+	-0.0450	(0.9452)	8.4145	(0.0029)	POS - ATM+		
b_UT1to2ln	-0.0268	(0.7478)	-0.7358	(0.0539)	POS - ATM-	-0.0252	(0.6667)	-1.0760	(0.0002)	POS - ATM-		
b_UT1to6ln	0.0895	(0.0718)	1.8825	(<.0001)	POS + ATM+	0.0901	(0.0134)	2.2610	(<.0001)	POS + ATM+		
avg_balance_1	0.0001	(0.7328)	0.0079	(<.0001)	POS + ATM+	0.0006	(<.0001)	0.0003	(0.7122)	POS + ATM+		
avg_balance_2	-0.0008	(0.0006)	-0.0042	(0.0001)	POS - ATM-	-0.0002	(0.1467)	-0.0055	(<.0001)	POS - ATM-		
avg_balance_3	-0.0005	(0.0257)	-0.0021	(0.0353)	POS - ATM-	-0.0003	(0.0844)	-0.0026	(0.0008)	POS - ATM-		
avg_balance_4	0.0001	(0.5999)	0.0009	(0.3462)	POS + ATM+	-0.0001	(0.5751)	0.0001	(0.9433)	POS - ATM+		
avg_balance_5	0.0001	(0.8061)	0.0017	(0.0821)	POS + ATM+	-0.0002	(0.1283)	0.0007	(0.3584)	POS - ATM+		
avg_balance_6	0.0006	(0.0009)	-0.0010	(0.1949)	POS + ATM-	0.0000	(0.7375)	0.0008	(0.2032)	POS - ATM+		
avg_deb_amt_1	0.0001	(0.7854)	0.0078	(0.0004)	POS + ATM+	-0.0019	(<.0001)	-0.0006	(0.7233)	POS - ATM-		
avg_deb_amt_2	0.0023	(<.0001)	0.0085	(<.0001)	POS + ATM+	-0.0007	(0.0447)	-0.0027	(0.1251)	POS - ATM-		
avg_deb_amt_3	0.0022	(<.0001)	0.0071	(0.0007)	POS + ATM+	-0.0004	(0.2581)	-0.0028	(0.0932)	POS - ATM-		
avg_deb_amt_4	0.0014	(0.0002)	0.0029	(0.1071)	POS + ATM+	-0.0003	(0.2373)	-0.0016	(0.2712)	POS - ATM-		
avg_deb_amt_5	0.0006	(0.1011)	0.0001	(0.9504)	POS + ATM+	-0.0005	(0.0418)	-0.0036	(0.0066)	POS - ATM-		
avg_deb_amt_6	0.0006	(0.0364)	0.0006	(0.6256)	POS + ATM+	-0.0002	(0.3972)	-0.0014	(0.1576)	POS - ATM-		
sum_crd_amt_1	0.0012	(<.0001)	0.0139	(<.0001)	POS + ATM+	0.0010	(<.0001)	0.0089	(<.0001)	POS + ATM+		
sum_crd_amt_2	0.0006	(0.0013)	0.0103	(<.0001)	POS + ATM+	0.0009	(<.0001)	0.0040	(<.0001)	POS + ATM+		
sum_crd_amt_3	0.0000	(0.8237)	0.0076	(<.0001)	POS + ATM+	0.0006	(<.0001)	0.0005	(0.5323)	POS + ATM+		
sum_crd_amt_4	-0.0001	(0.6053)	0.0059	(<.0001)	POS - ATM+	0.0003	(0.0146)	-0.0022	(0.0013)	POS + ATM-		
sum_crd_amt_5	-0.0001	(0.7737)	0.0069	(<.0001)	POS - ATM+	0.0001	(0.3175)	-0.0016	(0.0051)	POS + ATM-		
sum_crd_amt_6	0.0002	(0.0575)	0.0059	(<.0001)	POS + ATM+	-0.0002	(0.0288)	-0.0014	(0.0001)	POS - ATM-		
sum_deb_amt_1	0.0021	(<.0001)	-0.0069	(<.0001)	POS + ATM-	-0.0008	(0.0005)	-0.0026	(0.0532)	POS - ATM-		
sum_deb_amt_2	0.0015	(<.0001)	-0.0052	(0.0051)	POS + ATM-	-0.0012	(<.0001)	0.0016	(0.283)	POS - ATM+		
sum_deb_amt_3	0.0016	(<.0001)	-0.0095	(<.0001)	POS + ATM-	-0.0017	(<.0001)	0.0017	(0.2563)	POS - ATM+		
sum_deb_amt_4	0.0022	(<.0001)	-0.0091	(<.0001)	POS + ATM-	-0.0015	(<.0001)	0.0008	(0.5743)	POS - ATM+		
sum_deb_amt_5	0.0020	(<.0001)	-0.0088	(<.0001)	POS + ATM-	-0.0012	(<.0001)	0.0022	(0.0884)	POS - ATM+		
sum_deb_amt_6	0.0023	(<.0001)	-0.0075	(<.0001)	POS + ATM-	-0.0008	(0.0001)	0.0001	(0.9058)	POS - ATM+		
max_deb_amt_1	-0.0030	(<.0001)	0.0013	(0.5153)	POS - ATM+	0.0002	(0.4797)	-0.0059	(0.0001)	POS + ATM-		
max_deb_amt_2	-0.0021	(<.0001)	-0.0012	(0.5262)	POS - ATM-	0.0004	(0.1717)	-0.0066	(<.0001)	POS + ATM-		
max_deb_amt_3	-0.0015	(<.0001)	0.0076	(0.0001)	POS - ATM+	0.0011	(<.0001)	-0.0019	(0.2237)	POS + ATM-		
max_deb_amt_4	-0.0019	(<.0001)	0.0084	(<.0001)	POS - ATM+	0.0011	(<.0001)	-0.0018	(0.2141)	POS + ATM-		
max_deb_amt_5	-0.0018	(<.0001)	0.0074	(<.0001)	POS - ATM+	0.0010	(<.0001)	-0.0031	(0.0196)	POS + ATM-		
max_deb_amt_6	-0.0024	(<.0001)	0.0068	(<.0001)	POS - ATM+	0.0007	(0.0008)	-0.0010	(0.4011)	POS + ATM-		
min_deb_amt_1	0.0036	(<.0001)	0.0005	(0.7725)	POS + ATM+	0.0005	(0.0367)	0.0046	(0.0004)	POS + ATM+		
min_deb_amt_2	0.0017	(<.0001)	-0.0007	(0.682)	POS + ATM-	0.0000	(0.9005)	0.0029	(0.0274)	POS - ATM+		
min_deb_amt_3	0.0016	(<.0001)	-0.0028	(0.0678)	POS + ATM-	0.0000	(0.9287)	0.0020	(0.102)	POS - ATM+		
min_deb_amt_4	0.0008	(0.0019)	-0.0064	(<.0001)	POS + ATM-	-0.0004	(0.0267)	0.0011	(0.3638)	POS - ATM+		
min_deb_amt_5	0.0015	(<.0001)	-0.0037	(0.0033)	POS + ATM-	0.0002	(0.2432)	0.0014	(0.1759)	POS + ATM+		
min_deb_amt_6	0.0009	(<.0001)	-0.0038	(<.0001)	POS + ATM-	0.0001	(0.3645)	0.0012	(0.0772)	POS + ATM+		
b_AvgOB1_to_MaxOB1_ln	1.4533	(0.0003)	16.8303	(<.0001)	POS + ATM+	0.4756	(0.1103)	12.8305	(<.0001)	POS + ATM+		
b_AvgOB2_to_MaxOB2_ln	1.5580	(0.0001)	14.1589	(<.0001)	POS + ATM+	0.9949	(0.0008)	10.2451	(<.0001)	POS + ATM+		
b_AvgOB3_to_MaxOB3_ln	1.5358	(0.0002)	16.4131	(<.0001)	POS + ATM+	0.9284	(0.0023)	9.8930	(<.0001)	POS + ATM+		
b_TRmax_deb1_To_Limit_ln	-2.1942	(0.0578)	-19.8859	(<.0001)	POS - ATM-	-1.5772	(0.0664)	-29.7165	(<.0001)	POS - ATM-		
b_TRmax_deb2_To_Limit_ln	-1.9048	(0.0851)	-1.5464	(0.7578)	POS - ATM-	0.2508	(0.7581)	-15.2945	(0.0002)	POS + ATM-		
b_TRmax_deb3_To_Limit_ln	-4.5151	(<.0001)	-12.8431	(0.0086)	POS - ATM-	-1.0977	(0.1643)	-17.3832	(<.0001)	POS - ATM-		
b_Travg_deb1_to_avgOB1_ln	-0.1038	(0.5331)	-8.1992	(<.0001)	POS - ATM-	-0.1654	(0.1803)	-0.8908	(0.1407)	POS - ATM-		
b_Travg_deb2_to_avgOB2_ln	0.1083	(0.5165)	-4.1130	(<.0001)	POS + ATM-	0.1394	(0.2545)	1.9330	(0.0013)	POS + ATM+		
b_Travg_deb3_to_avgOB3_ln	-0.0290	(0.8546)	-3.6251	(<.0001)	POS - ATM-	-0.1285	(0.2703)	2.0146	(0.0005)	POS - ATM+		
b_Trsum_deb1_to_TRsum_cr	0.1499	(0.2974)	8.2682	(<.0001)	POS + ATM+	0.1212	(0.2557)	1.4622	(0.0061)	POS + ATM+		
b_Trsum_deb2_to_TRsum_cr	-0.1374	(0.3355)	7.5888	(<.0001)	POS - ATM+	-0.0848	(0.421)	1.0798	(0.0367)	POS - ATM+		
b_Trsum_deb3_to_TRsum_cr	-0.0965	(0.4581)	5.2396	(<.0001)	POS - ATM+	0.0460	(0.6339)	-0.1559	(0.7402)	POS + ATM-		
b_NumDeb13to46ln	-0.3294	(0.0002)	-4.0216	(<.0001)	POS - ATM-	0.0368	(0.5622)	-1.3934	(<.0001)	POS + ATM-		
b_avgNumDeb13	0.7312	(<.0001)	1.6447	(<.0001)	POS + ATM+	-0.0119	(0.7803)	0.0442	(0.8596)	POS - ATM+		

Table 7.17 Comparative analysis of pooled and random effect regression coefficient estimations for POS and ATM income (continue)

Variable	Pooled						Random effect - WK					
	POS		ATM		Signs	POS		ATM		Signs		
	Estimate	Pr > t	Estimate	Pr > t		Estimate	Pr > t	Estimate	Pr > t		Estimate	Pr > t
b_OB13_to_OB46ln	0.2522	(0.0062)	4.1127	(<.0001)	POS + ATM+	-0.0064	(0.9235)	0.9758	(0.0055)	POS - ATM+		
b_OB1_to_OB2_ln	0.1958	(0.2192)	3.8985	(<.0001)	POS + ATM+	0.1597	(0.1637)	2.1611	(0.0001)	POS + ATM+		
b_OB2_to_OB3_ln	0.1925	(0.0763)	3.4058	(<.0001)	POS + ATM+	0.1797	(0.0209)	1.8459	(<.0001)	POS + ATM+		
b_OB3_to_OB4_ln	-0.0016	(0.9787)	1.6037	(<.0001)	POS - ATM+	0.0080	(0.8547)	1.4416	(<.0001)	POS + ATM+		
b_OB_avg_to_eop1ln	-0.3392	(0.021)	-6.1233	(<.0001)	POS - ATM-	-0.4197	(0.0002)	-3.0319	(<.0001)	POS - ATM-		
b_pos_flag_use13vs46	-3.2216	(<.0001)	-12.0898	(<.0001)	POS - ATM-	-0.0611	(0.772)	2.2882	(0.0198)	POS - ATM+		
b_atm_flag_use13vs46	4.2066	(<.0001)	0.0000	(.)	POS + ATM+	-0.5144	(0.0335)	0.0000	(.)	POS + ATM+		
b_pos_use_only_flag_13	7.6129	(<.0001)	14.1852	(<.0001)	POS + ATM+	-0.4291	(0.0503)	6.2720	(<.0001)	POS - ATM+		
b_atm_use_only_flag_13	-2.8859	(<.0001)	0.0000	(.)	POS + ATM+	0.5534	(0.0183)	0.0000	(.)	POS + ATM+		
b_Tsum_crd1_to_OB1_ln	0.2297	(0.1523)	10.1403	(<.0001)	POS + ATM+	0.1462	(0.2173)	3.0973	(<.0001)	POS + ATM+		
b_Tsum_crd2_to_OB2_ln	0.0508	(0.7338)	6.6696	(<.0001)	POS + ATM+	-0.1271	(0.2469)	0.8699	(0.1022)	POS - ATM+		
b_Tsum_crd3_to_OB3_ln	0.0195	(0.8817)	5.2643	(<.0001)	POS + ATM+	-0.0180	(0.8531)	0.3075	(0.514)	POS - ATM+		
b_payment_lt_5p_1	-0.1205	(0.6163)	-0.9715	(0.2799)	POS - ATM-	-0.4918	(0.0062)	0.2000	(0.7856)	POS - ATM+		
b_payment_lt_5p_2	-0.1866	(0.4402)	-2.1368	(0.0191)	POS - ATM-	-0.5511	(0.0021)	-0.9933	(0.1781)	POS - ATM-		
b_payment_lt_5p_3	-0.0125	(0.9583)	-3.2227	(0.0003)	POS - ATM-	-0.5861	(0.001)	-1.3453	(0.0667)	POS - ATM-		
b_maxminOB_limit_1_ln	-0.3903	(0.0101)	-9.6919	(<.0001)	POS - ATM-	-0.5789	(<.0001)	-8.0530	(<.0001)	POS - ATM-		
b_maxminOB_limit_2_ln	-0.4768	(0.001)	-9.0928	(<.0001)	POS - ATM-	-0.5785	(<.0001)	-7.6113	(<.0001)	POS - ATM-		
b_maxminOB_limit_3_ln	-0.3344	(0.011)	-11.8167	(<.0001)	POS - ATM-	-0.4821	(<.0001)	-9.0365	(<.0001)	POS - ATM-		
b_OBbias_1_ln	0.0669	(0.4934)	-0.5725	(0.1132)	POS + ATM-	0.0276	(0.7042)	-0.9186	(0.0017)	POS + ATM-		
b_OBbias_2_ln	0.0538	(0.5781)	-0.7235	(0.0445)	POS + ATM-	0.0954	(0.1832)	-0.6886	(0.0178)	POS + ATM-		
b_OBbias_3_ln	-0.0355	(0.7133)	-1.0160	(0.0046)	POS - ATM-	0.0851	(0.2346)	-0.9293	(0.0013)	POS + ATM-		
b_maxminOB_avgOB_1_ln	1.1152	(<.0001)	14.1657	(<.0001)	POS + ATM+	0.5503	(0.0004)	8.6063	(<.0001)	POS + ATM+		
b_maxminOB_avgOB_2_ln	1.0137	(<.0001)	11.0891	(<.0001)	POS + ATM+	0.4447	(0.0031)	6.4525	(<.0001)	POS + ATM+		
b_maxminOB_avgOB_3_ln	1.1088	(<.0001)	14.4965	(<.0001)	POS + ATM+	0.4555	(0.0016)	7.7300	(<.0001)	POS + ATM+		
b_Tsum_deb1_to_2_ln	0.0407	(0.5334)	0.9067	(0.0188)	POS + ATM+	0.0145	(0.7526)	0.5856	(0.0481)	POS + ATM+		
b_Tsum_crd1_to_2_ln	0.0818	(0.4876)	-1.1776	(0.0313)	POS + ATM-	-0.0086	(0.9186)	-0.7784	(0.0662)	POS - ATM-		
I_ch1_ln	2.3172	(0.0776)	21.6054	(0.0001)	POS + ATM+	0.8188	(0.3884)	15.2791	(0.0006)	POS + ATM+		
I_ch1_flag	-0.6234	(0.3848)	-7.6029	(0.007)	POS - ATM-	0.4663	(0.3631)	-6.7599	(0.0021)	POS + ATM-		
I_ch6_flag	-1.5519	(<.0001)	-7.4687	(<.0001)	POS - ATM-	-0.5680	(0.011)	-7.0360	(<.0001)	POS - ATM-		
age	0.0777	(<.0001)	-0.3486	(<.0001)	POS + ATM+	0.1241	(0.0047)	-0.8484	(<.0001)	POS + ATM-		
customer_income_ln	3.5707	(<.0001)	10.8002	(<.0001)	POS + ATM+	8.8516	(<.0001)	62.3436	(<.0001)	POS + ATM+		
Edu_High	0.7110	(0.0022)	-2.7310	(0.0034)	POS + ATM-	1.1388	(0.1833)	-7.1902	(0.0414)	POS + ATM-		
Edu_Special	-0.5214	(0.0314)	-0.1814	(0.8368)	POS - ATM-	-0.9259	(0.3016)	-1.4521	(0.6639)	POS - ATM-		
Edu_TwoDegree	0.9244	(0.0642)	-6.6170	(0.0105)	POS + ATM-	2.8121	(0.1278)	-5.5463	(0.5704)	POS + ATM-		
Marital_Civ	-0.1931	(0.5905)	5.0122	(0.0006)	POS - ATM+	-0.0961	(0.9427)	12.1761	(0.0284)	POS - ATM+		
Marital_Div	-0.0969	(0.7135)	-1.1562	(0.257)	POS - ATM-	-0.1643	(0.8666)	-0.5561	(0.8851)	POS - ATM-		
Marital_Sin	0.5551	(0.0333)	5.2133	(<.0001)	POS + ATM+	0.7415	(0.444)	6.3568	(0.1318)	POS + ATM+		
Marital_Wid	-0.9164	(0.1631)	2.6193	(0.1694)	POS - ATM+	-1.2651	(0.605)	2.5629	(0.7237)	POS - ATM+		
position_Man	0.5961	(0.0204)	2.9730	(0.0059)	POS + ATM+	0.7555	(0.4283)	3.9920	(0.3284)	POS + ATM+		
position_Oth	-0.3225	(0.2188)	2.2267	(0.0214)	POS - ATM+	-0.1206	(0.9006)	4.5398	(0.2152)	POS - ATM+		
position_Tech	-1.5069	(<.0001)	3.9602	(<.0001)	POS - ATM+	-2.1309	(0.0184)	5.1092	(0.1335)	POS - ATM+		
position_Top	-2.1461	(<.0001)	4.2572	(0.0592)	POS - ATM+	-3.0556	(0.1245)	10.7105	(0.21)	POS - ATM+		
sec_Agricult	-3.1907	(<.0001)	1.3889	(0.4238)	POS - ATM+	-4.8142	(0.0424)	3.1139	(0.6353)	POS - ATM+		
sec_Constr	0.9972	(0.1357)	-9.4891	(<.0001)	POS + ATM+	0.1012	(0.6767)	-13.7435	(0.1283)	POS + ATM-		
sec_Energy	-1.9781	(<.0001)	-3.8266	(0.0114)	POS - ATM-	-2.1503	(0.143)	-4.0412	(0.481)	POS - ATM-		
sec_Fin	0.7448	(0.0045)	-14.6702	(<.0001)	POS + ATM-	2.0954	(0.0285)	-15.4275	(0.001)	POS + ATM-		
sec_Industry	-1.1618	(0.1865)	11.9233	(0.0002)	POS - ATM+	-0.7073	(0.8274)	5.5027	(0.6516)	POS - ATM+		
sec_Manufact	0.1014	(0.8773)	-8.3872	(0.0003)	POS + ATM-	-0.9945	(0.6809)	-12.2259	(0.1623)	POS - ATM-		
'	-1.9057	(<.0001)	-9.2743	(<.0001)	POS - ATM-	-2.6958	(0.1368)	-6.9684	(0.26)	POS - ATM-		
sec_Service	-0.7370	(0.0008)	-2.8736	(0.0007)	POS - ATM-	-0.5470	(0.5011)	-3.6364	(0.2564)	POS - ATM-		
sec_Trade	-0.4139	(0.1086)	-1.0752	(0.3763)	POS - ATM-	-0.0893	(0.9254)	5.9340	(0.1952)	POS - ATM+		
sec_Trans	0.2222	(0.7314)	-4.5933	(0.0546)	POS + ATM-	0.1474	(0.9511)	-4.9803	(0.5831)	POS + ATM-		
car_Own	-0.0138	(0.9506)	-1.0202	(0.2414)	POS - ATM-	0.9168	(0.2654)	-0.8760	(0.7899)	POS + ATM-		
car_coOwn	1.6734	(<.0001)	3.6328	(0.0042)	POS + ATM+	2.4145	(0.0811)	6.1777	(0.2005)	POS + ATM+		
real_Own	0.6520	(0.0012)	-0.1320	(0.8701)	POS + ATM-	0.9188	(0.2179)	-0.8931	(0.7701)	POS + ATM-		
real_coOwn	0.4140	(0.037)	-1.4975	(0.0721)	POS + ATM-	0.2593	(0.7246)	-2.7961	(0.3759)	POS + ATM-		
reg_ctr_Y	-2.3686	(<.0001)	4.4979	(0.0006)	POS - ATM+	-3.1327	(0.0008)	-2.2667	(0.6457)	POS - ATM-		
reg_ctr_N	-3.7896	(<.0001)	10.5285	(<.0001)	POS - ATM+	-5.3446	(<.0001)	3.9792	(0.4115)	POS - ATM+		
child_1	0.1774	(0.4776)	2.0355	(0.0452)	POS + ATM+	-0.0499	(0.9573)	1.1913	(0.7581)	POS - ATM+		
child_2	-0.0648	(0.6753)	0.5999	(0.3061)	POS - ATM+	-0.3138	(0.5847)	-0.4660	(0.8343)	POS - ATM-		
child_3	-2.6997	(<.0001)	5.781627	(0.0037)	POS - ATM+	-2.5573	(0.2661)	11.3768	(0.1311)	POS - ATM+		

Table 7.18 Comparative analysis of pooled and random effect regression coefficient estimations for POS and ATM income (continue)

Variable	Pooled						Random effect - WK					
	POS		ATM		Signs	POS		ATM		Signs		
	Estimate	Pr > t	Estimate	Pr > t		Estimate	Pr > t	Estimate	Pr > t		Estimate	Pr > t
Unempl_Inyoy	0.0355 (0.9921)	-417.8600 (<.0001)	POS + ATM-	-0.2937 (0.9095)	-360.6400 (<.0001)	POS - ATM-						
UAH_EURRate_Inmom	-11.4444 (0.0079)	-26.3807 (0.1303)	POS - ATM-	-7.9945 (0.008)	-35.7482 (0.0073)	POS - ATM-						
UAH_EURRate_Inyoy	0.5015 (0.8487)	33.8882 (0.001)	POS + ATM+	1.5637 (0.3993)	55.6125 (<.0001)	POS + ATM+						
CPI_Inqoq	-49.7415 (<.0001)	-127.8610 (<.0001)	POS - ATM-	-37.8649 (<.0001)	9.9571 (0.6519)	POS - ATM+						
SalaryYear_Inyoy	2.2293 (0.7044)	151.9027 (<.0001)	POS + ATM+	-3.8063 (0.3622)	38.7931 (0.0306)	POS - ATM+						
s_cons_full	-0.1689 (0.3863)	-1.9085 (0.047)	POS - ATM-	-0.2089 (0.161)	-2.5004 (0.0018)	POS - ATM-						
s_month_since_NA_full	0.8070 (<.0001)	-0.0699 (0.9191)	POS + ATM-	-0.0646 (0.5919)	0.0576 (0.9218)	POS - ATM+						
s_month_since_Tr_full	1.1990 (<.0001)	1.3650 (0.0889)	POS + ATM+	0.9124 (<.0001)	2.0221 (0.0031)	POS + ATM+						
s_month_since_Re_full	0.5371 (0.0099)	1.2011 (0.2755)	POS + ATM+	0.5407 (0.0007)	-0.1684 (0.8541)	POS + ATM-						
s_month_since_RP_full	0.2623 (0.1713)	1.7354 (0.0719)	POS + ATM+	0.1523 (0.2932)	2.6061 (0.001)	POS + ATM+						
s_month_since_D1_full	0.4719 (0.0213)	1.3639 (0.1723)	POS + ATM+	0.3125 (0.0488)	1.6614 (0.0503)	POS + ATM+						
s_month_since_D2_full	0.1502 (0.5632)	3.3740 (0.1552)	POS + ATM+	-0.0494 (0.8006)	1.8563 (0.3473)	POS - ATM+						
s_times_NA_full	3.2426 (<.0001)	6.1954 (0.5485)	POS + ATM+	1.4924 (<.0001)	-8.7698 (0.3609)	POS + ATM-						
s_times_TR_full	3.1535 (<.0001)	6.7916 (0.5167)	POS + ATM+	3.4894 (<.0001)	-3.3999 (0.7268)	POS + ATM-						
s_times_RE_full	1.8362 (<.0001)	10.8331 (0.2894)	POS + ATM+	1.3832 (<.0001)	-1.1146 (0.9069)	POS + ATM-						
s_times_RP_full	2.8455 (<.0001)	10.7154 (0.3075)	POS + ATM+	1.6699 (0.0001)	-1.4987 (0.8776)	POS + ATM-						
s_times_D1_full	0.3705 (0.349)	1.8801 (0.8556)	POS + ATM+	0.2159 (0.4994)	-6.2820 (0.5133)	POS + ATM-						
s_times_D2_full	-0.7081 (0.4072)	17.3887 (0.2334)	POS - ATM+	-0.0020 (0.9975)	2.3044 (0.8613)	POS - ATM+						
d_StateFull_1_NA	0.1422 (0.9493)	42.6642 (0.0023)	POS + ATM+	1.3074 (0.431)	17.5843 (0.1118)	POS + ATM+						
d_StateFull_1_Tr	-1.5795 (0.4641)	34.1794 (0.0121)	POS - ATM+	3.0231 (0.0602)	8.6746 (0.4203)	POS + ATM+						
d_StateFull_1_Re	-4.1856 (0.0051)	-15.1935 (0.1981)	POS - ATM-	-0.4107 (0.7112)	-19.4144 (0.0362)	POS - ATM-						
d_StateFull_1_RP	-2.8649 (0.1549)	25.4616 (0.0523)	POS - ATM+	1.7436 (0.2445)	2.2449 (0.8282)	POS + ATM+						
d_StateFull_1_D1	-2.6529 (0.0576)	-13.1737 (0.2576)	POS - ATM-	0.1482 (0.8865)	-12.1347 (0.1839)	POS + ATM-						

Selection of the final model for prediction of transactional income

We apply Hausman test for the selection of the approach for the final model, which will be used for the prediction of transactional income. The null hypothesis H_0 , that the individual effects u_i are uncorrelated with the other regressors in the model, is rejected (or the Hausman test cannot be computed for the set of covariates). This means that we should use the fixed-effect model rather than a random-effect model. However, this does not allow us to predict the income amount for other periods and other cases than used in the data sample for the model training. The fixed effect model can explain group difference in intercepts, and throw away time-invariant variables. However, we need to examine both between and within effects to make predictions.

Thus, we will use the Pooled method as a final selection for the aggregated model, when the model fails the Hausman test. However, we need to consider that the Pooled OLS is inefficient because it does not consider the autocorrelation in the composite error term.

7.8 Summary of the non-interest income functions performance

For the prediction of the income from POS and ATM transactions we estimated 40 linear models: two types of transactions (POS and ATM) x two types of models (one-stage and two-stage) x five types of methods for panel data (pooled and four random-

effect) and also two binary logistic regression models for the prediction of the probability of transaction.

For further implementation in the general model for the total credit card income prediction for each type of transaction we selected the two-stage model, which consists of one logistic regression and one linear regression with the highest fitting accuracy – pooled method. Table 7.19 gives the results of the performance quality of the selected models for the transactional income prediction for six months period.

Pooled linear regression gave similar fitting accuracy for both POS and ATM income: R-squared 0.29 and 0.28 respectively. The random-effect method (Wallace and Hussain Variance Component) demonstrated an R-squared between 0.18 and 0.15 only for test data set. These results cannot be considered as a good performance model. On the other hand, the models for the prediction of the probability of transaction have demonstrated very good performance – Area Under Curve indexes are equal to 0.83 and 0.86 for POS and ATM transactions respectively.

Table 7.19 Performance quality of the income prediction models

Model	Regression equation	Target	Results
Probability of POS transaction	Logistic regression $\ln\left(\frac{P_i}{1-P_i}\right) = \sum_{k=1}^K \beta_k \cdot B_{ki,t-1} + \sum_{l=1}^L \alpha_a \cdot A_{ai} + \sum_{m=1}^M \gamma_m M_{m,t-1}$	POS transaction next 6 months	AUC=0.83
Probability of ATM withdrawal	Logistic regression $\ln\left(\frac{P_i}{1-P_i}\right) = \sum_{k=1}^K \beta_k \cdot B_{ki,t-1} + \sum_{l=1}^L \alpha_a \cdot A_{ai} + \sum_{m=1}^M \gamma_m M_{m,t-1}$	ATM withdrawal next 6 months	AUC=0.86
POS income (interchange)	Panel regression: polled $POS_i = \sum_{k=1}^K \beta_k \cdot B_{ik} + \sum_{l=1}^L \alpha_a \cdot A_{ai} + \sum_{m=1}^M \gamma_m M_m$	POS Income next 6 months	R ² ~0.29
POS income (interchange)	Panel regression: random-effect $POS_{it} = \sum_{k=1}^K \beta_k \cdot B_{bi,t-1} + \sum_{l=1}^L \alpha_a \cdot A_{ai} + \sum_{m=1}^M \gamma_m M_{m,t-1}$	POS Income next 6 months	R ² ~ 0.18
ATM withdrawal income	Panel regression: polled $ATM_i = \sum_{k=1}^K \beta_k \cdot B_{ik} + \sum_{l=1}^L \alpha_a \cdot A_{ai} + \sum_{m=1}^M \gamma_m M_m$	ATM withdrawal income next 6 months	R ² ~ 0.28
ATM withdrawal income	Panel regression: random-effect $ATM_{it} = \sum_{k=1}^K \beta_k \cdot B_{bi,t-1} + \sum_{l=1}^L \alpha_a \cdot A_{ai} + \sum_{m=1}^M \gamma_m M_{m,t-1}$	ATM withdrawal income next 6 months	R ² ~ 0.15

7.9 Conclusion

As a result of this Chapter we built models for the prediction of income from Point-of-Sales transactions and ATM cash withdrawals. We tested one-stage models, which use the income amount directly as a dependent variable, and two-stage models, which predict the probability of POS or ATM transaction and estimate the transactional income conditional of the probability of a transaction. For the direct model we consider observations without transactions and generated income, for two-stage models we use income amount prediction only for observations with related POS and ATM transactions. We used linear regression for the income amount prediction. Because of the usage of panel data, we tested different methods, which consider a random-effect variance component.

Baltagi et al. (2002) tested different random-effect variance component methods for panel data. These are Fuller and Battese (1974), Wansbeek and Kapteyn (1989), Wallace and Hussain (1969), and Nerlove (1971) with Monte-Carlo simulated data sample. However, these random-effect methods for panel data have been *firstly* applied for the transactional income amount prediction in this research.

Random-effect models show lower prediction accuracy than pooled effect model. However, the estimators can be more efficient (or unbiased) because of the use of methods, which consider between and within variance component. The Wallace and Hussain (WH) variance component method has shown the highest prediction accuracy among tested random-effect methods.

Two-stage models, which estimate the POS/ATM income amount conditional on the probability of the corresponding transaction, have given more accurate fit than one-stage direct income amount estimation models.

The prediction of the logarithms of income amount to the outstanding balance ratios give significantly more accurate regression models than direct amount prediction. But after the transformation of indirect results as exponent from ratio to amount, the income prediction via ratios gives even higher error than direct amount estimation because of a high variance due to relatively high values.

The predictive models for the probability of transaction in the current research have shown high goodness of fit and confirm that it is relatively easy to predict whether a

cardholder will use a credit card for certain type of transaction or not. The predictive models for the income amount have demonstrated low fitting accuracy, and an improvement of non-interest income prediction is a field for further researches.

We provide a list of the most significant explanatory characteristics for the transactional income with related positive or negative correlation between characteristics and outcome. A positive sign means a marginal increase of the income for an increase of the characteristic or value equal to one for dummy variables.

For the transactional income from Point-of-sales (POS) transactions the following characteristics are significant:

Behavioural: the credit limit utilisation rate at the observation point month and ATM use flag with negative signs, loan payment amounts for last month, maximum and sum of purchase amount for the last 6 months, average number of purchase transactions for the last 3 months, and POS use only flag with positive signs;

Application: logarithm of the customer income to the average income, no region living, sector – Finance – with positive signs, sector – agriculture – with negative sign;

States: not significant;

Macroeconomic: CPI change for the last quarter with negative sign.

For transactional income from ATM cash withdrawal transactions the following covariates are significant:

Behavioural: month on book (+), logarithm of the sum of purchase transactions to the sum of payments in month 1, loan payment amounts for last 3 months, average number of purchase transactions for the last 6 months, ATM use only flag – with positive signs, the difference between monthly minimum and maximum outstanding balance for 3 months, the credit limit utilisation rate at the observation point month – with negative signs;

Application: logarithm of the customer income to the average income - with positive sign, sector – Finance – with negative sign;

States: a customer has the transactor state at the observation point - with positive sign;

Macroeconomic: the unemployment rate change for the last year (negative sign for 6 months, positive sign for 1 month), the exchange rate of the local currency to EUR for the last month – positive sign, average salary changes for the last year – negative sign.

This research revealed the weakness of linear regression methods, for instance, for the prediction of income from credit card transactions. Many relationships between predictors and dependent characteristics are non-linear. The problem can be partially solved with the use of characteristics binning and the transformation of predictors to dummy variables. Another way is the usage of high degree polynomial for a description of dependencies between factors and outcome. However, in case of information complexity and big volumes of data, non-linear regression analysis and machine learning methods (see Crook et al., 2007; Finlay, 2010) can give more efficient substitution of the traditional linear methods.

8 Chapter 8. Total income prediction with an aggregated model

8.1 Overview

Profitability predictions for credit cards, as well as their dual nature as both payment tools and convenient loans, are well-represented in the literature (Crook et al., 1992; Ma et al., 2010; So & Thomas, 2008; Cheu & Loke, 2010; Tan et al., 2011). Traditionally, the decision whether to grant a loan is based on estimated credit risk. However, for credit cards, low credit risk often means low usage and, consequently, low profit. Thus, information about a borrower's credit card usage can be more attractive to grantors than a risk score, because usage is related to profit (Banasik et al., 2001). Loan decisions based on credit card usage create a conflict of interest for the lender, forcing them to choose between risk and profitability.

For credit cards, reducing risk means reducing profit, and vice versa. Therefore, it is necessary for lenders to establish a way to optimise the balance between risk and usage (that is, risk and profit). However, we propose another approach to risk-profit estimation and decision making. Instead of the risk-profit decision, total profit can be predicted from a set of credit card activity, and the decision to lend can be made based on total profit maximisation criterion—that is, based on the total profit generated, rather than on risk or profit alone.

In this paper, we compare three approaches to the prediction of total credit card income for several months, as well as during specific months. The predictive horizon is six months. We have tested seven income models for months $t+1, t+2, \dots, t+6$, where t is the current month and total income for six months as the sum of monthly incomes.

The first approach is an aggregation of interest income and transactional income models as an algebraic sum of partial incomes predicted utilising these models. Interest income is predicted as a product of utilisation rate (see Results, Chapter 4), credit limit at the time of prediction, and monthly interest rate. Transactional income is predicted as the sum of income from POS transactions and ATM cash withdrawals (see Results, Chapter 7). Thus, the total income is calculated as interest income + POS transactional income + ATM transactional income.

The second approach is an aggregation of interest income and transactional income models as the sum of partial incomes, weighted by the probability of transition from the current state of the account to the set of possible states for a certain number of months. The results of transition probability models and states (defined in Chapters 5 and 6) are used as weights for the models from Chapter 4 and Chapter 7. We consider this approach to be the primary model in this work, based on the key hypothesis that a model with partial income predictions weighted by the states' transition probabilities provides the most accurate fit, compared to non-weighted and direct estimation.

The third approach to the prediction of credit card income is a direct estimation of total income without consideration of the results given in previous chapters. Monthly incomes and total income over a six-month period are predicted using linear regression using the same set of predictors as with other models. This approach is used as a fixed point to determine whether complication of the models improves predictive accuracy, or whether the simple models are more efficient than complex ones.

Total income prediction is a behavioural model built on, and implemented for, accounts open for at least six months; behavioural characteristics are therefore calculated for six months or longer. For an application model without behavioural history, and for the first five months on balance, the same approach can be applied, but with appropriate predictors.

This chapter describes first, the aggregated total income model; second, the approach for total income calculation; third, the total income prediction results; fourth, the comparative analysis of the validation results of aggregated and direct estimation models for total income; firth, the expected loss modelling; and sixth, total profit prediction; scenario analysis and business contribution examples.

The main difference between the existing papers in the modelling of revenue and profitability prediction (Andreeva et al., 2007; Ma et al., 2010; So et al., 2014) and our research is the following. First, we use a lot of behavioural variables, which have not been used before. Second, we aggregate income from different sources: interest and transactional income, into the total income. Third, we use income conditional on the current state and predicted state of the account.

8.2 An approach to the total income calculation

8.2.1 The relationship between an account state and income

This section describes the main aggregated formulas for the calculation of total income for different types of aggregations and relationships between account states and corresponding income sources.

Generally, risk management efforts approach distributed accounts by segments with the use of delinquency bins, such as current, day past due (DPD) 1-30 (Bucket 1), DPD 31-60 (Bucket 2), and so on (see So and Thomas, 2011). We have defined credit card states based on balances outstanding and delinquency – see Tables 6.16 and 5.1.

Accounts in any state other than inactive and defaulting can generate income. However, the sources of income are different for each state. For instance, delinquent accounts can produce non-interest income due to interchange fees from merchants and penalties but do not produce interest income because of unpaid debt.

Revenue is generated from two sources: i) interest revenue, and ii) transaction revenue. Interest revenue is generated as a percentage of an account's active outstanding balance, usually for monthly accruals, and is averaged each month. Transaction revenue is generated from interchange fees from debit transaction purchases, and as ATM usage fees from cash withdrawal.

Revolver accounts generate both interest rate and transactional revenues and can generate losses in a period of t+4 (at the shortest) if the default state is defined as 4 or more missed payments. In this case, a revolver progresses from its current state through a series of delinquent states, from 0 delinquency bucket to 1, from 1 to 2, from 2 to 3, and finally from 3 to 4.

Inactive accounts do not generate revenue, because they have no active outstanding balance and no debit transactions, and require an additional one more month than revolver accounts—a period of t+5—to generate losses, as they must move to the revolver state at the first stage. Transactor accounts generate transaction revenue only, and also generate risk for a period of t+4, because transactors are inactive at the end of the accrual period; thus, by the definition, they do not have required payments.

Delinquent accounts can generate transaction revenue and penalties. However, penalties are accrued in the current period, but earned after the resumption of payments. A 1-month delinquency generates risk for t+3, because an account transitioning from up to 30 days past due (1 missed payment) to 60 or more days past due (2 or more missed payments) must miss at least 3 payments and move forward from the first delinquency bucket to the second, and from the second to the third, or default state. Respectively, a 2-month delinquent state requires a 1-month period of transition to reach the default state.

For detailed definitions of states, see Chapter 5. Table 8.1 is a version of Table 5.1 updated to show the full set of states, including revenue (income) and risk (losses) for each state. The inactive state is a state with zero average outstanding balance and spending amount. The transactor state is a state with zero outstanding balance at the end of the period but positive spending amount for the month. Both inactive and transactor states are not able to generate expected losses for the next month, so the risk level is defined as ‘No risk’. Transactors generate transactional income from spending transactions. The revolver state is defined as a state with a positive outstanding balance at the end of a period and with zero days past due. Income from revolvers is defined as a sum of the interest income, which computed as the product of the credit limit, the utilisation rate, and the interest rate, and the transactional income. The risk is estimated as the probability of transition to Delinquent 1 State. The new state ‘Revolver Paid’ is introduced as a state with a positive outstanding balance at the end of the previous month and zero outstanding balance at the end of the observation month. The income sources are the same as for the revolver state, but the revolver paid state does not generate risk because the only possible transitions for the next month are to one of the inactive, transactor, and revolver states only.

Table 8.1 Account state definition and related assessments

Account status	Symbol	Definition	Risk assessment	Revenue assessment
Inactive	NA	Average OB = 0 and Spending Amount = 0	No	No
Transactor	TR	OB end of period (eop) = 0 and Spending Amount > 0	No	Transactional Income Amount
Revolver (current)	RE	OB eop > 0 and DPD = 0	Transition to Delinquent 1 State	Limit x Utilisation Rate x Interest Rate + Transactional Income Amount
Revolver Paid (Repaid)	RP	OB eop (t-1) > 0 and OB eop (t) = 0	No	Limit x Utilization Rate x Interest Rate + Transactional Income Amount
Delinquent 1 (Days Past Due 1-30)	D1	OB eop > 0 and (DPD > 0 and DPD <=30)	Transition to Delinquent 2 State	Transactional Income Amount
Delinquent 2 (Days Past Due 31-60)	D2	OB eop > 0 and (DPD > 31 and DPD <=60)	Transition to Default State	No
Defaulted (DPD 61 and more)	Df	OB eop > 0 and DPD >= 61	LGD	-

Delinquent 1 state is a state with the positive outstanding balance at the end of a period and the number of days past due between 1 and 30. The level of risk is assessed as the probability of transition to higher delinquency bucket. An account in this state can still generate the income from spending transactions, but not from interest income.

Income in the early delinquent state can occur in cases where a cardholder has made a partial payment of the amount due, but the account has nonetheless progressed to a delinquent state because of the delinquent amount incurred. Therefore, it is possible for delinquent accounts to generate interest income. However, we assume (for the purpose of simplification) that delinquent accounts do not make payments towards the amount due.

Also, the bank can get penalty payments from the delinquent accounts, but we do not consider this source of income in the scope of this research.

Delinquent 2 state is a state with a positive outstanding balance at the end of a period and the number of days past due between 30 and 60. The level of risk is assessed as

the probability of transition to a higher delinquency state – default bucket. Usually accounts are blocked at the second month of delinquency, so we consider that an account does not generate any income at this state. The last state is an absorbing defaulted state, which is defined as a positive outstanding balance at the end of a month and days past due counter higher than or equal to 61. Because the default event has already occurred at this stage, the risk component of Expected Loss is estimated as Loss Given Default. So potentially we can estimate the share of the outstanding balance, which can be recovered after collection actions. However, this is out of scope of this research, and accrued and collected interest income is not considered in the total income.

An account in any of the states except inactive, delinquent 2, and default can generate income. The income sources are different for each state (see Table 8.2). For instance, delinquent accounts can generate non-interest income due to penalties and interchange fees from merchants, but it does not generate interest income because of unpaid debt. Types of income generated within particular account states are marked with crosses in the appropriate cell of Table 8.2.

Table 8.2 Sources of income for each state

<i>Status</i>	<i>Interest</i>	<i>Fees/Interchange</i>	<i>Penalty</i>
<i>Inactive</i>	-	-	-
<i>Transactor</i>	-	+	-
<i>Revolver</i>	+	+	-
<i>Revolver paid</i>	+	+	-
<i>Delinquent</i>	-	+	+
<i>Defaulted</i>	-	-	-

Depending on an account's state, each account has an individual set of models: the probability of transition to possible states, the probability of action, and the income estimation for each possible action. Thus, the total income prediction model is presented as the sum of results of three-level conditional models:

- i) the probability of being in state s ;
- ii) the probability of action;
- iii) the income estimation after action for the specific state.

For each state we have defined indicators, which are used in the income model: outstanding balance, interest income, debit turnover (purchases transactions amount), transactional income amount, and the risk, or the expected loss estimation. The total scope of the model, used for the final aggregated model for the full set of states, is described in Table 8.3.

Table 8.3 Matrix of models for the set of account states

Status (s)	Inactive (NA)	Transactor (TR)	Revolver (RE)/ Repaid(RP)	Delinquent (D1)	Defaulted (Df)
S at t+1	Pr(NA s≠D)	Pr(Tr s=NA,TR,RP)	Pr(RE s≠D) Pr(RP s≠D)	Pr(D1 s=R) Pr(D2 s=D1)	Pr(D s=D2)
Balance	N/A	N/A	UR x Limit	UR x Limit	UR x CF x Limit
Interest Income	N/A	N/A	UR x Limit x IR	0	0
Debit Turnover	N/A	Pr(POS s=TR) Pr(ATM s=TR)	Pr(POS s=RE/RP) Pr(ATM s=RE/RP)	Pr(POS s=D1) Pr(ATM s=D1)	N/A
Transact. Income	N/A	Inc(TR Pr(POS)=1) Inc(TR Pr(ATM)=1)	Inc(RE/RP Pr(POS)=1) Inc(RE/RP Pr(ATM)=1)	Inc(D1 Pr(POS)=1) Inc(D1 Pr(ATM)=1)	N/A
Risk	N/A	0	Pr(D s=RE)	Pr(D s=D2)	LGD x UT x CCF x Limit

‘S at t+1’ is the set of the probability of transition models, conditional on the current state. So, for example, the probability of being a transactor at t+1 can be defined in a common format with a model $\Pr(Tr|s=NA,TR,RE)$, which means the probability of being in the transactor state at time t+1 conditional on being inactive, transactor, or revolver paid in period t. A revolver in the proposed model cannot be a transactor in the next month because a revolver account, which has a positive outstanding balance at the end of observed month, will have a positive outstanding balance at least for one day in the next month, and so the interest income will be accrued, but a transactor must not have the interest income. So, we use the ‘Revolver Paid’ state as an obligatory transition state between the revolver state and other non-delinquent states such as inactive and transactor due to income source purposes (see Chapter 5 and 6).

Balance at the end of the period and interest income are not applicable parameters for the transactor state. The probability of Debit turnover, or spending transaction, is computed as $\Pr(POS/s=TR)$ and $\Pr(ATM/s=TR)$ respectively for the type of POS or ATM transaction conditional on being in the current transactor state. The expected income amount after the transaction is computed with separate models for each income source as $\text{Inc}(TR | \Pr(POS)=1)$ and $\text{Inc}(TR | \Pr(ATM)=1)$. Expected income is then computed as a product of the probability and the income amount (see Chapter 7).

The final model is defined as the sum of the products of three estimations, such as the probability of being in state s , the probability of the action (POS/ATM), and the estimated income for each state.

For the states of revolver and revolver paid the interest income is computed as a product of the average outstanding balance and the interest rate. The outstanding balance is computed as the product of the average utilisation rate for the period and the credit limit.

The expected outstanding balance for default state is computed as the product of the expected utilisation rate, the credit limit, and the credit conversion factor (CCF), which reflects the increase of the outstanding balance at the time of default in comparison with a non-default period (see section 8.6 about expected loss). A risk component for default state is defined as expected losses and computed as the product of the outstanding balance, credit conversion factor, and Loss Given Default.

Income is generated by accounts in the four following states: transactor, revolver, revolver paid, and delinquent. In states Delinquent 2 and Defaulted, interest income can be accrued, but the cardholder does not make payments and the interest income is not actually generated. The credit card is blocked and transactions are not available, so the transactional income is not generated either.

8.2.2 Income calculation for states

We investigated total credit card income originating from two sources: interest rate income, and transactional income from POS and ATM cash withdrawals. Each account can be in one of the possible states at any time, and the income sources are active or inactive, depending on the account state at the end of the period of income prediction. Thus, each state has a corresponding individual formula for income calculation. In

total, there are seven states in the model: inactive, transactor, revolver, revolver paid (repaid), delinquent 1, delinquent 2, and default. Income is generated in 4 states: transactor, revolver, revolver paid (repaid), and delinquent 1. The income formula for each state is based on data for one period; in this case, one month. Because the state of an account can change over time, the expected income sources from an account can also vary. Throughout several months, the monthly incomes are calculated individually, then aggregated as a sum for the given period.

Four formulas are used to compute expected income for an account in a given state. The formulae related to utilisation rate, transactional revenue, the probability of certain types of transactions, and the probability that an account is in a given state.

Income functions for account i at time $t+n$ are represented as

$$I(i, t + n | s_{i,t+n}),$$

where $s_i \in \{TR, RE, RP, D1\}$

and $S = \{TR, RE, RP, D1\}$ is the set of states for which transaction income generation is possible.

Transactor income consists of transactional revenue only and is written as follows:

$$\begin{aligned} I(i, t + n | s_{i,t+n} = TR) &= \Pr(s_{i,t+n} = TR | s_{i,t} \neq Df) \times \\ &\times \left(\Pr(a_{i,t+n} = POS | s_{i,t+n} = TR) \cdot R(\mathbf{x}_{it} | a_{i,t+n} = POS) + \right. \\ &\quad \left. + \Pr(a_{i,t+n} = ATM | s_{i,t+n} = TR) \cdot R(\mathbf{x}_{it} | a_{i,t+n} = ATM) \right) \end{aligned} \quad (8.1)$$

where

$R(\mathbf{x}_i | a = POS/ATM)$ is the revenue (transactional income) function for the account i subject to vector of predictors \mathbf{x}_i depending on the action type,

$a_{i,t+n}$ is the type of transaction: POS or ATM of the account i at time $t+n$,

$\Pr(s_{i,t+n} = TR | s_{i,t} \neq Df)$ is the probability of transition of the account i at time $t+n$ to the state s , so long as the account does not default at current time t ,

t is the current time point,

n is the number of months for prediction (performance window), and

$s_{i,t}$ is the state of account i at time t .

$\Pr(a_{i,t+n} = POS \mid s_{i,t+n} = TR)$ is the probability of POS transaction of the account i at time $t+n$, so long as the account is in state ‘transactor’ at time $t+n$.

The ‘probability of transaction’ function is used in the income equation in two ways: first, as a weight of the income from the corresponding source; and second, as a binary indicator function of transaction-making with a particular threshold, i.e. if the probability of transaction is higher than the threshold, the transaction is carried out; otherwise the transaction is predicted as non-existent.

The indicator function 1_{P^*} for the second approach for the probability of any POS or ATM transaction can be written as

$$1_{P^*}(\Pr(a_{i,t+n} = POS / ATM \mid s_{i,t+n} = S)) = \begin{cases} 1, & \Pr(a_{i,t+n} = POS / ATM \mid s_{i,t+n} = S) \geq P^* \\ 0, & \Pr(a_{i,t+n} = POS / ATM \mid s_{i,t+n} = S) < P^* \end{cases}$$

where P^* is the threshold for the probability of transaction, $P^* \in [0,1]$

Hereafter, we use $\Pr(a_{i,t+n} = POS/ATM \mid s_{i,t+n} \in S)$ in formulas for the probability of POS or ATM transaction identification; this corresponds with the first approach of the use of the probability of transaction as a weight in the income amount equation. However, in further calculations, we ceased use of the second approach with the indicator function and transaction threshold for the probability of a transaction.

Revolver income consists of the interest and transactional revenue and is written as follows:

$$\begin{aligned} I(i, t+n \mid s_{i,t+n} = RE) &= \Pr(s_{i,t+n} = RE \mid s_{i,t} \neq DF) \times \\ &\times \left(\text{Ut}(\mathbf{x}_{it} \mid s_{i,t+n} = RE) \times IR \times \text{Limit}_{it} + \right. \\ &\quad \left. + \Pr(a_{i,t+n} = POS \mid s_{i,t+n} = RE) \cdot R(\mathbf{x}_{it} \mid a_{i,t+n} = POS, s_{i,t+n} = RE) + \right. \\ &\quad \left. + \Pr(a_{i,t+n} = ATM \mid s_{i,t+n} = RE) \cdot R(\mathbf{x}_{it} \mid a_{i,t+n} = ATM, s_{i,t+n} = RE) \right) \end{aligned} \quad (8.2)$$

where

$\text{Ut}(\mathbf{x}_i \mid s_{i,t+1} = RE)$ is the utilisation rate function of the vector of predictors \mathbf{x}_i ,

depending on the revolver state at time $t+1$;

IR is the interest rate for the account. We assume the interest rate is a constant for all accounts for any period.

$Limit_{it}$ is the credit limit for the account i at time t . We assume that the credit limits can vary.

We use the same model for accounts with changes in credit limit and with constant credit limits throughout the research, but credit limit changes are reflected in predictors used in regression models.

The *revolver paid (repaid)*: the income formula is the same as for the revolver state, and consists of the interest and transactional revenue, and is written as follows:

$$\begin{aligned} I(i, t+n | s_{i,t+n} = RP) &= \Pr(s_{i,t+n} = RP | s_{i,t} \neq DF) \times \\ &\times \left(\text{Ut}(\mathbf{x}_{it} | s_{i,t+n} = RP) \times IR \times Limit_{it} + \right. \\ &\times \left. + \Pr(a_{i,t+n} = POS | s_{i,t+n} = RP) \cdot R(\mathbf{x}_{it} | a_{i,t+n} = POS, s_{i,t+n} = RP) + \right. \\ &\left. + \Pr(a_{i,t+n} = ATM | s_{i,t+n} = RP) \cdot R(\mathbf{x}_{it} | a_{i,t+n} = ATM, s_{i,t+n} = RP) \right) \end{aligned} \quad (8.3)$$

The ‘utilisation rate at the month of repayment’ and ‘transactional revenue’ functions may differ from the corresponding functions for accounts in the revolver state. The revolver paid state is included in the model to ensure the accuracy of transition probability calculations, because accounts in both transactor and revolver states—in which the full amount of debt is paid — have, by definition, outstanding balances of zero at the end of the month.

Delinquent 1 (Days past due 1-30): income from an account in this state consists of transactional revenue only, and is computed as follows:

$$\begin{aligned} I(i, t+n | s_{i,t+n} = D1) &= \Pr(s_{i,t+n} = D1 | s_{i,t} \neq Df) \times \\ &\times \left(\Pr(a_{i,t+n} = POS | s_{i,t+n} = D1) \cdot R(\mathbf{x}_{it} | a_{i,t+n} = POS, s_{i,t+n} = D1) + \right. \\ &\left. + \Pr(a_{i,t+n} = ATM | s_{i,t+n} = D1) \cdot R(\mathbf{x}_{it} | a_{i,t+n} = ATM, s_{i,t+n} = D1) \right) \\ &+ \text{Penalty} \end{aligned} \quad (8.4)$$

where *Penalty* is a fee for delinquency. Penalties can be a significant income source, especially for short-term consumer loans such as cash loans, sale finances, and payday loans. In the current model, penalty is assumed to be equal to zero.

Income in the early delinquent state can be generated from POS-transaction interchange fees and cash withdrawal from ATM fees, but not from the interest, because the cardholder does not make payments.

The revenue (transactional income) functions (Chapter 6) and the utilisation rate function (Chapter 3) are the results of regressions build using the same set of predictors from vector \mathbf{x} . These equations are examples of account predictions for $t+n$, where for the current model $n = 1,2,3,4,5,6$.

8.2.3 Total income calculation

After calculating the income for each state, it is necessary to aggregate them to produce the total income prediction for individual months and a period.

Total income can be calculated in three ways: i) direct total income estimation; ii) as a sum of income from multiple sources (interest income and transactional income); and iii) sum of income from multiple sources, assuming the appropriate state is weighted by the probability of transition to that state. The final method is split into two options according to the source of the probability of transition: from the transition matrix, and from the individual transition probabilities.

8.2.3.1 Direct total income estimation

Direct total income estimation is a result of income amount prediction with a regression equation. We have selected linear regression (Ordinary Least Squares - OLS), for the simplification of the calculations and as a benchmark for the main complex model. The total income by direct estimation (TID) is predicted as follows:

$$TID(i, t + n) = \boldsymbol{\beta}^T \mathbf{x}_i + \varepsilon,$$

where \mathbf{x}_i is a vector of predictors for the account i ,

$\boldsymbol{\beta}$ is a vector of coefficients of linear regression, estimated using OLS method.

For n months, the total income amount for the period from $t+1$ to $t+n$ is calculated as the sum of n monthly income functions:

$$TID(i; t+1, \dots, t+n) = \sum_{k=1}^n TID(i; t+k),$$

where k is month's counter.

Therefore, the procedure for direct estimation of total income for n months is the prediction of monthly income for each of n months, using linear (or generally any appropriate) regression, and the summation of the results of the monthly models. For

a brief description of the model, see section ‘Direct Total income prediction with linear model’ in this Chapter.

8.2.3.2 Total income as a simple sum of interest and transactional incomes

The direct total income prediction does not consider the part of the income consisting of the interest income and transactional income separately, and these parts can i) be significantly unequal and therefore differently impact the total income, and ii) make up different proportions of the total in different periods. In contrast to the direct model for total income estimation, the proposed approach is to predict income for each source, and then to sum the parts for the estimation of total income.

Total income (TI) as a simple sum of interest and transactional income is calculated as follows:

$$TI(i, t+n) = Interest_Income(i, t+n) + POS_Transaction_Income(i, t+n) + ATM_Transaction_Income(i, t+n)$$

The total income as a simple sum of interest and transactional income excludes inappropriate states for income prediction. The utilisation rate for the interest rate prediction should not include the inactive, transactor, delinquent 1 (1-30 days past due), delinquent 2 (31-60 days past due), and defaulted states, as the cardholder does not pay the interest payments at these states. The transactional income for POS and ATM transactions should not include inactive, delinquent 2 (30-59 days past due), and defaulted states, as transactions are not possible in these states because the credit card is inactive or is blocked. The full formula for the total income (TI) as a simple sum of interest and transactional income can be written as:

$$TI(i, t+n) = Ut(\mathbf{x}_{it} | s_{i,t+n} \notin (NA, TR, D1, D2, Df)) \times IR \times Limit_{it} + \\ + Pr(a_{i,t+n} = POS | s_{i,t+n} \notin (NA, D2, Df), \mathbf{x}_i) \cdot R(\mathbf{x}_{it} | a_{i,t+n} = POS, s_{i,t+n} \notin (NA, D2, Df)) + , (8.5) \\ + Pr(a_{i,t+n} = ATM | s_{i,t+n} \notin (NA, D2, Df), \mathbf{x}_i) \cdot R(\mathbf{x}_{it} | a_{i,t+n} = ATM, s_{i,t+n} \notin (NA, D2, Df))$$

where

$Ut(\mathbf{x}_{it} | s_{i,t+n} \notin (NA, TR, D2, Df))$ means the utilization rate function, which depends on the vector of predictors \mathbf{x} of the account i at time t ;

$\Pr(a_{i,t+n} = POS \mid s_{i,t+n} \notin (NA, D2, Df), \mathbf{x}_i)$ is the probability of POS transaction of the account i at time $t+n$, assuming that the account will not be in a state of inactive or delinquency or default at time $t+n$;

$\Pr(a_{i,t+n} = ATM \mid s_{i,t+n} \notin (NA, D2, Df), \mathbf{x}_i)$ is the probability of ATM transaction of the account i at time $t+n$, assuming on that the account will not be in the inactive or delinquent or default state at time $t+n$.

However, the simple sum approach does not consider the possible future states of each account, which can impact the structure of the received income. The interest income and the transactional income models can make estimations for any account, but do not consider the probability an account will be in the relevant income earning state. For example, interest income for the account can be predicted at time $t+n$ using the general model, but if the account becomes inactive or defaults, the interest income prediction will no longer make sense. As we are not able to predict the exact state of the account—only the probability—the weight of each state at time $t+n$ can be used.

8.2.3.3 Total income as a sum of state incomes weighted by the states transition probabilities.

Total income is calculated as a weighted sum of individual models of expected income prediction for each account state. In the context of total income, a model for income prediction for a particular state is described as ‘partial’ because total income is computed as a sum of expected income from all possible states. Therefore, $I(i, t+n \mid s_{i,t+n} = RE)$ is a partial income model for the revolver state of the account i at time $t+n$, and so on.

Total monthly income weighted by the transition probabilities (TIW) is the sum of the products of the interest and transactional income functions, multiplied by the probabilities of transition to the corresponding states. For an n-month prediction period, the total income amount at month $t+n$ is as follows:

$$\begin{aligned}
TIW(i, t+n) = & I(i, t+n | s_{i,t+n} = TR) \cdot \Pr(s_{i,t+n} = TR) + \\
& + I(i, t+n | s_{i,t+n} = RE) \cdot \Pr(s_{i,t+n} = RE) + \\
& + I(i, t+n | s_{i,t+n} = RP) \cdot \Pr(s_{i,t+n} = RP) + \\
& + I(i, t+n | s_{i,t+n} = D1) \cdot \Pr(s_{i,t+n} = D1)
\end{aligned} \tag{8.6}$$

The full formula for the total monthly income *weighted* by the transition probabilities can be written as follows:

$$\begin{aligned}
TI(i, t+n) = & Ut(\mathbf{x}_{it} | s_{i,t+n} = RE) \times IR \times Limit \times \Pr(s_{i,t+n} = RE | \mathbf{x}_{it}) + \\
& + Ut(\mathbf{x}_{it} | s_{i,t+n} = RP) \times IR \times Limit \times \Pr(s_{i,t+n} = RP | \mathbf{x}_{it}) + \\
& + Ut(\mathbf{x}_{it} | s_{i,t+n} = D1) \times IR \times Limit \times \Pr(s_{i,t+n} = D1 | \mathbf{x}_{it}) + \\
& + Pr(a_{i,t+n} = POS | s_{i,t+n} = TR, \mathbf{x}_{it}) \cdot R(\mathbf{x}_{it} | a_{i,t+n} = POS, s_{i,t+n} = TR) \cdot \Pr(s_{i,t+n} = TR | \mathbf{x}_{it}) + \\
& + Pr(a_{i,t+n} = ATM | s_{i,t+n} = TR, \mathbf{x}_{it}) \cdot R(\mathbf{x}_{it} | a_{i,t+n} = ATM, s_{i,t+n} = TR) \cdot \Pr(s_{i,t+n} = TR | \mathbf{x}_{it}) + \\
& + Pr(a_{i,t+n} = POS | s_{i,t+n} = RE, \mathbf{x}_{it}) \cdot R(\mathbf{x}_{it} | a_{i,t+n} = POS, s_{i,t+n} = RE) \cdot \Pr(s_{i,t+n} = RE | \mathbf{x}_{it}) + \\
& + Pr(a_{i,t+n} = ATM | s_{i,t+n} = RE, \mathbf{x}_{it}) \cdot R(\mathbf{x}_{it} | a_{i,t+n} = ATM, s_{i,t+n} = RE) \cdot \Pr(s_{i,t+n} = RE | \mathbf{x}_{it}) + \\
& + Pr(a_{i,t+n} = POS | s_{i,t+n} = RP, \mathbf{x}_{it}) \cdot R(\mathbf{x}_{it} | a_{i,t+n} = POS, s_{i,t+n} = RP) \cdot \Pr(s_{i,t+n} = RP | \mathbf{x}_{it}) + \\
& + Pr(a_{i,t+n} = ATM | s_{i,t+n} = RP, \mathbf{x}_{it}) \cdot R(\mathbf{x}_{it} | a_{i,t+n} = ATM, s_{i,t+n} = RP) \cdot \Pr(s_{i,t+n} = RP | \mathbf{x}_{it}) + \\
& + Pr(a_{i,t+n} = POS | s_{i,t+n} = D1, \mathbf{x}_{it}) \cdot R(\mathbf{x}_{it} | a_{i,t+n} = POS, s_{i,t+n} = D1) \cdot \Pr(s_{i,t+n} = D1 | \mathbf{x}_{it}) + \\
& + Pr(a_{i,t+n} = ATM | s_{i,t+n} = D1, \mathbf{x}_{it}) \cdot R(\mathbf{x}_{it} | a_{i,t+n} = ATM, s_{i,t+n} = D1) \cdot \Pr(s_{i,t+n} = D1 | \mathbf{x}_{it})
\end{aligned} \tag{8.7}$$

This formula uses individual probabilities of transitions to particular states at the account level. On the other hand, the transition probabilities from Transition matrix model (TMM) can be used, calculated at the portfolio level. For the TMM for the transition probabilities, see Chapter 6, section 6.4. The total monthly income weighted is computed as follows:

$$\begin{aligned}
TI(i, t+n) = & Ut(\mathbf{x}_{it} | s_{i,t+n} = RE) \times IR \times Limit \times \Pr(s_{i,t+n} = RE) + \\
& + Ut(\mathbf{x}_{it} | s_{i,t+n} = RP) \times IR \times Limit \times \Pr(s_{i,t+n} = RP) + \\
& + Ut(\mathbf{x}_{it} | s_{i,t+n} = D1) \times IR \times Limit \times \Pr(s_{i,t+n} = D1) + \\
& + Pr(a_{i,t+n} = POS | s_{i,t+n} = TR, \mathbf{x}_{it}) \cdot R(\mathbf{x}_{it} | a_{i,t+n} = POS, s_{i,t+n} = TR) \cdot \Pr(s_{i,t+n} = TR) + \\
& + Pr(a_{i,t+n} = ATM | s_{i,t+n} = TR, \mathbf{x}_{it}) \cdot R(\mathbf{x}_{it} | a_{i,t+n} = ATM, s_{i,t+n} = TR) \cdot \Pr(s_{i,t+n} = TR) + \\
& + Pr(a_{i,t+n} = POS | s_{i,t+n} = RE, \mathbf{x}_{it}) \cdot R(\mathbf{x}_{it} | a_{i,t+n} = POS, s_{i,t+n} = RE) \cdot \Pr(s_{i,t+n} = RE) + , \\
& + Pr(a_{i,t+n} = ATM | s_{i,t+n} = RE, \mathbf{x}_{it}) \cdot R(\mathbf{x}_{it} | a_{i,t+n} = ATM, s_{i,t+n} = RE) \cdot \Pr(s_{i,t+n} = RE) + \\
& + Pr(a_{i,t+n} = POS | s_{i,t+n} = RP, \mathbf{x}_{it}) \cdot R(\mathbf{x}_{it} | a_{i,t+n} = POS, s_{i,t+n} = RP) \cdot \Pr(s_{i,t+n} = RP) + \\
& + Pr(a_{i,t+n} = ATM | s_{i,t+n} = RP, \mathbf{x}_{it}) \cdot R(\mathbf{x}_{it} | a_{i,t+n} = ATM, s_{i,t+n} = RP) \cdot \Pr(s_{i,t+n} = RP) + \\
& + Pr(a_{i,t+n} = POS | s_{i,t+n} = D1, \mathbf{x}_{it}) \cdot R(\mathbf{x}_{it} | a_{i,t+n} = POS, s_{i,t+n} = D1) \cdot \Pr(s_{i,t+n} = D1) + \\
& + Pr(a_{i,t+n} = ATM | s_{i,t+n} = D1, \mathbf{x}_{it}) \cdot R(\mathbf{x}_{it} | a_{i,t+n} = ATM, s_{i,t+n} = D1) \cdot \Pr(s_{i,t+n} = D1)
\end{aligned} \tag{8.8}$$

where $\Pr(s_{i,t+n} = S)$ is the probability of transition to state $S \in \{TR, RE, RP, D1\}$ in accordance with the transition matrix.

The difference between formulae (8.7) and (8.8) is that in formula (8.7) the probability of transition to state is conditional on the vector of covariates \mathbf{x}_{it} . So each account has an individual probability of transition to the state. On the other hand, in formulae (8.8) the probability of transition to the state is not conditional on the characteristics of the account except the current state of the account.

For a one-month prediction period, the total income amount at month $t+1$ can be computed as follows:

$$TIW(i, t+1) = I(i, t+1 | s_{i,t+1} = TR) \cdot \Pr(s_{i,t+1} = TR) + I(i, t+1 | s_{i,t+1} = RE) \cdot \Pr(s_{i,t+1} = RE) + \\ + I(i, t+1 | s_{i,t+1} = RP) \cdot \Pr(s_{i,t+1} = RP) + I(i, t+1 | s_{i,t+1} = D1) \cdot \Pr(s_{i,t+1} = D1)$$

$t+1$ replaced with $t+2$, $t+3$, and so on, for the second, third, and ongoing months.

For n months, the total income amount for the period from $t+1$ to $t+n$ is a sum of monthly income functions for both the simple sum and the weighted sum approaches:

$$TIW(i; t+1, \dots, t+n) = \sum_{k=1}^n TIW(i; t+k), \text{ where } k \text{ is the month's counter.}$$

This facilitates estimation of total income over seven periods: six month predictions, and one prediction for the entire period from the first to the sixth month as a sum of monthly results. The description of the regressions and the model selection process for the final estimations inputs for each block of the aggregated model were discussed in the following chapters: Chapter 4 for the utilisation rate, used for the interest income, in the step two of the aggregated model, Chapter 7 for the transactional income for step three, and Chapter 5 and 6 for the transition probabilities for step one of the aggregated model.

The section below describes i) distributions of the observed total income and total income, predicted with the simple sum and weighted sum models; ii) validation results for the development, testing, and out-of-time samples for all seven-period models; and iii) back-testing results of observed versus predicted values for $t+1$ and six-month periods.

Additionally, the section contains a brief description of the Expected Loss model for deduction from the income part of the model for the final profit calculation. The overall contribution and business applications of the model are discussed in the final section.

8.3 Total Income distributions and calculation results

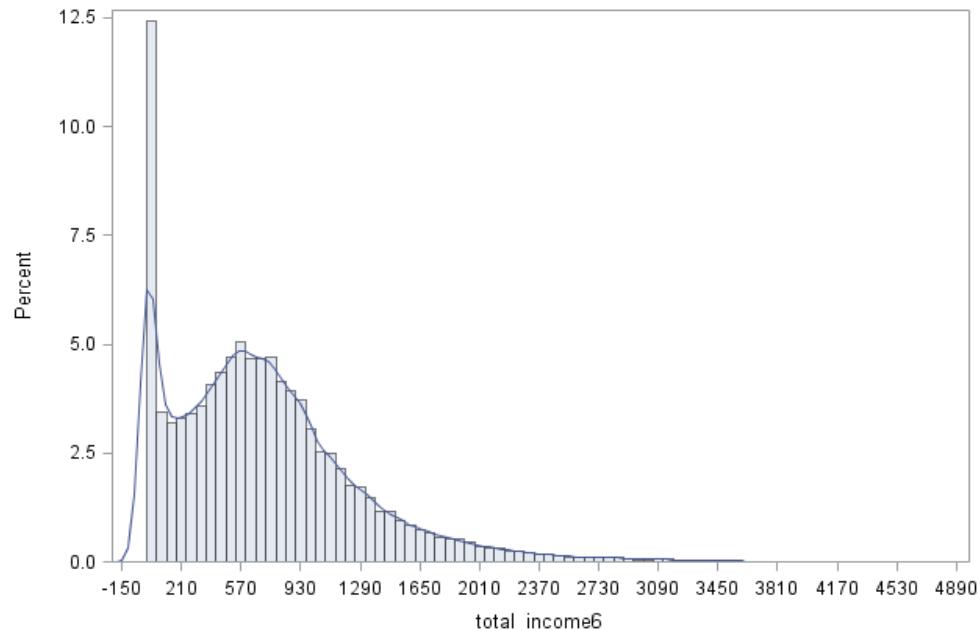
8.3.1 The distribution of target variables

First, we investigated the distribution of the target variables: interest and transactional income. This is necessary to test the fitting accuracy of the predictions, and to facilitate understanding of possible gaps and weaknesses in the predictive models.

In the figures from Figure 8.1 to Figure 8.3 below, we illustrate the density distribution of observed total income over six months, transactional income over six months, interest income over six months. The x-axis represents income values, while the y-axis represents the percentage of the income range under the bar in the total population.

Total income over six months has the highest proportion of cases, around 12.5% of the population, at zero, then a peak value is observed around 600 money units with a long thin tail to 5000 money units with an insignificant number of observations.

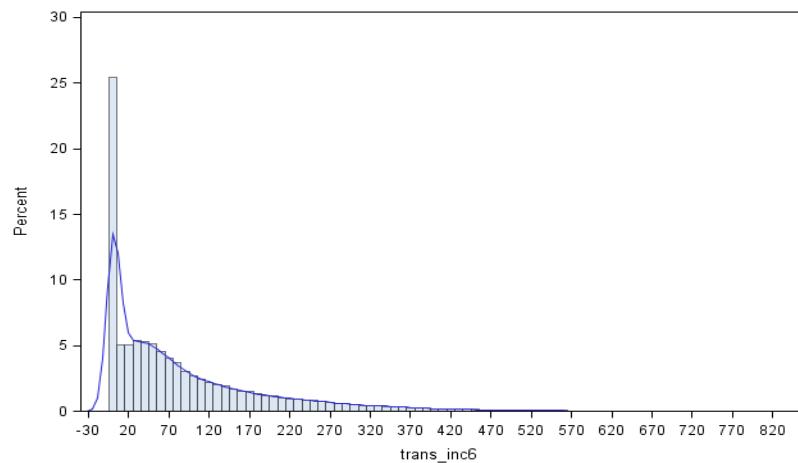
Figure 8.1 Density distribution of total income over six months (total_income6)



Total income over six months has the highest number of cases around 12.5% of the population for zero value. A peak value is observed around 600 money units with a long thin tail to 5000 money units with insignificant number of observations.

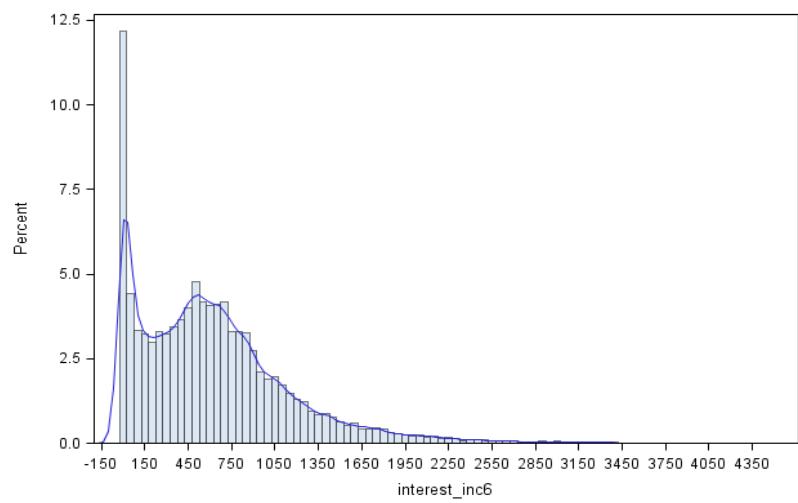
Transactional income for six months (trans_inc6) has a flat distribution in the area between 0 and 50 money units of income, and then a slight decrease to 850 units with an insignificant number of observations.

Figure 8.2 Density distribution of the transactional income over 6 months (trans_inc6)



The distribution density for the interest income (interest_inc6) has a shape similar to total income in Figure 8.1, but differs from the transactional income in Figure 8.2. This because the interest income amounts exceed the transactional income amounts and so dominate the total income.

Figure 8.3 Density distribution of interest income over 6 months (interest_inc6)



This distribution of total income for six months, conditional on the current state at time t , differs between the states as shown in Table 8.4. Because an account for 6-months can be in any state, it is expected that an inactive account at the observation point, will generate on an average income of 112.5. The current transactor and revolver repaid will give similar incomes, 215 and 242. Significantly higher income is expected from accounts, which are in the revolver state at the observation point, 824 money units. However, the highest total income for 6 months is expected from accounts, which are in early delinquency state 1, 961.4 money units. However, expected total 6-months income distributions are skewed. Income from accounts in the Revolver and Delinquent 1 states have the lowest skewness, and the median values of the distributions are 722 and 838 respectively (see Table 8.5). Also, these states have the lowest coefficients of variation. So, the expected total income for 6 months for accounts in the revolver and delinquent states are more highly concentrated around mean and less variative than inactive, transactor, and revolver repaid states. For accounts in the inactive state the median is equal to zero, so less than half of the inactive population give any income for the next 6 months. Total income for 6 months for accounts in the transactor and revolver repaid states are similar by the shape of distribution and values.

Table 8.4 Statistics: Distribution of Total income over 6 months, by account states at time t

Characteristic	Inactive	Transactor	Revolver	Repaid	Delinquent 1
	NA	TR	RE	RP	D1
N	27 265	2 339	176 016	5 906	4 517
Mean	112.53	215.14	824.39	241.62	961.42
Std Deviation	219.37	291.25	569.57	319.60	544.25
Skewness	3.12	2.64	1.51	2.17	1.77
Coeff Variation	194.95	135.38	69.09	132.28	56.61

Table 8.5 Quartiles of the total income for 6 months' distribution by account states at time t

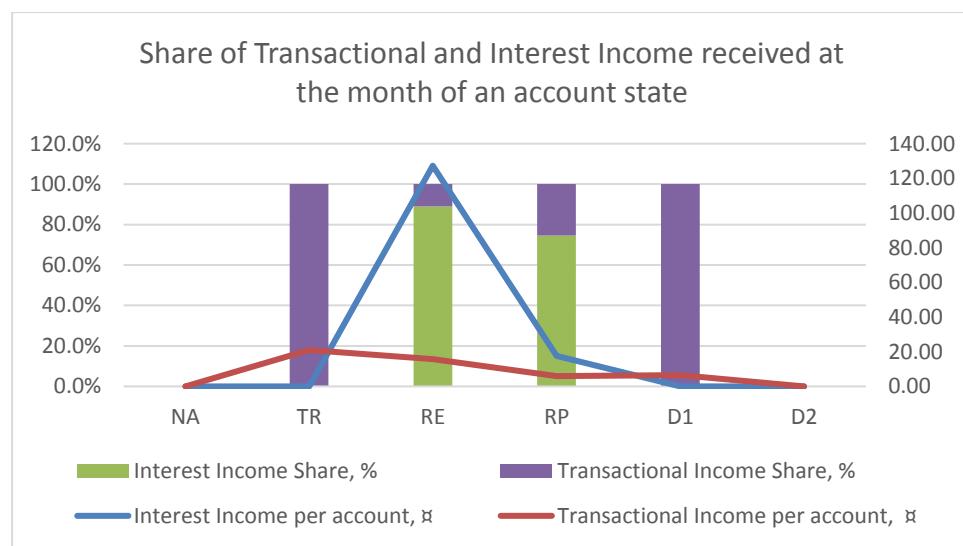
Quartile	Inactive	Transactor	Revolver	Repaid	Delinquent 1
	NA	TR	RE	RP	D1
100% Max	2 902.92	2 835.00	4 728.16	2 801.45	4 479.65
0.99	1 015.18	1 347.71	2 823.07	1 401.11	2 885.09
0.95	565.18	817.18	1 907.70	879.90	2 024.00
0.9	373.40	569.16	1 533.05	668.37	1 621.27
75% Q3	134.80	311.10	1 072.22	358.76	1 155.07
50% Median	0.00	106.42	722.89	120.95	838.60
25% Q1	0.00	15.85	445.50	0.86	606.53
0.1	0.00	0.00	211.67	0.00	442.53
0.05	0.00	0.00	102.46	0.00	351.00
0.01	0.00	0.00	5.40	0.00	129.60
0% Min	0.00	0.00	0.00	0.00	0.00

The interest and transactional incomes represent different shares of the total income, depending on the state. Table 8.6Table 8.4, Table 8.7, Figure 8.4 and Figure 8.5 below depict these differences in income periods of one month and six months. The current state is the initial, or observation point, state. In Table 8.6 we can see interest income per account, transactional income per account, total income per account, the share of interest income, and share of transactional income. As is expected, the highest interest income is produced from accounts in the revolver state, and a low amount of income can be obtained from accounts in the revolver paid state in the observed month. For the current month, the transactors do not generate interest income, but over six months, they can move to other states and the interest income can be generated. However, transactors give the highest transactional income per account – around 21 money units. Interest income gives 88.9% of the total income for revolvers and 74.5% of the total income for the revolver paid state. Delinquent and transactor state accounts can generate only transactional income in the month of observation.

Table 8.6 Share of Transactional and Interest Income received during the month of the current account state

Current State	Interest Income per account, ₽	Transactional Income per account, ₽	Total Income per account, ₽	Share of Interest Income, %	Share of Transactional Income, %
NA	0.00	0.00	0.00	-	-
TR	0.00	20.98	20.98	0.0	100.0
RE	127.26	15.81	143.08	88.9	11.1
RP	17.56	6.01	23.57	74.5	25.5
D1	0.00	6.46	6.46	0.0	100.0
D2	0.00	0.00	0.00	-	-

Figure 8.4 Share of Transactional and Interest Income received during the month of an account state



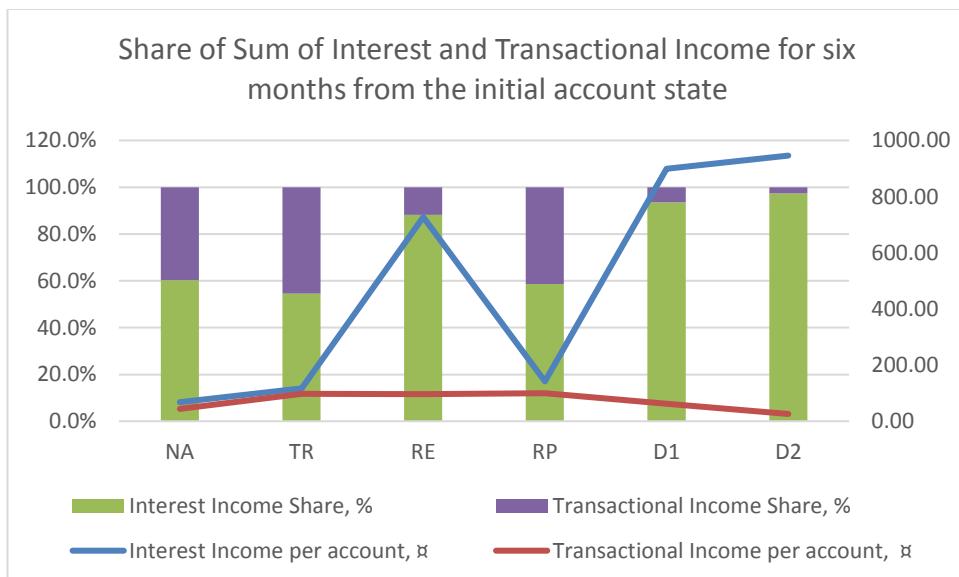
A completely different situation can be observed for a six-months prediction period. Each state can generate any type of income for the performance period. The highest interest income is obtained from delinquent 1 and delinquent 2 states, 898.78 and 946.18 money units respectively, in comparison with the revolver state, which gives in average only 727.43 money units per account for a six-month period.

Transactors give an average 117.27 money units of interest income per account, but it is higher than the sum of transactional income for six months. Transactors at the observation point have the biggest share of transactional income – 45% - in comparison with other states. The biggest share of interest income in the 6-month total income belongs to Delinquent states: 93.5% for Delinquent 1 and 97.3% for Delinquent 2. This means that clients with arrears amount mostly prefer not to spend money, but pay back the principal and accrued interest and return to the revolver state.

Table 8.7 Share of Sum of Transactional and Sum of Interest Income received for six months from the initial account state

State in the first month	Interest Income per account, ₽	Transactional Income per account, ₽	Total Income per account, ₽	The share of Interest Income, %	The share of Transactional Income, %
NA	67.87	44.66	112.53	60.3	39.7
TR	117.27	97.87	215.14	54.5	45.5
RE	727.43	96.95	824.39	88.2	11.8
RP	141.72	99.89	241.62	58.7	41.3
D1	898.78	62.64	961.42	93.5	6.5
D2	946.18	26.00	972.18	97.3	2.7

Figure 8.5 Share of Sum of Interest and Transactional Income over six months from the initial account state



Repaid accounts at the observation point generate mostly interest income during the one-month period, so clients try not to make purchase transactions in the months, and after they decide to pay back the full debt amount. However, the repaid state represents a significant share of transactional income during the six-month period, similar to transactors. This means that if we observe a payback it has a high chance of being repeated, so such a customer tends to spend money and repay the full amount. Revolver repaid can be an attractive category of clients in the sense of risk-return ratio: these clients understand and use the main benefits of the credit card as a payment tool and as a source of funds.

8.3.2 Data used in the Direct Estimation Method

In this thesis, we calculate the predicted total income in three different ways. We calculate it, firstly, as a simple sum of interest and transactional income, secondly, as the sum of state incomes weighted by the states transition probabilities, and, thirdly, as a sum of state incomes weighted by the Transition Matrix Model (TMM) state by formulas 8.5, 8.7, and 8.8.

The sample data and characteristics that are used for total income predictions were described in Chapter 3. The Transition matrix for method a sum of state incomes weighted by the Transition Matrix Model (TMM) computed with formulae 8.8 is described in Chapter 6.

We now describe an example of account data with the relevant calculations over a six-month period from the database used for prediction.

Each account (ID) is presented as a panel for a particular number of months, which differ between accounts; therefore, the panel is unbalanced. Table 8.8 contains the following observed data in highlighted columns: POS and ATM transaction income amounts over 6 months (Target_POS sum_6 and Target_ATM sum_6), average utilisation rate over 6 months (Target_UT_6), credit limit, transactional income (trans_inc), interest income, and total income.

Predicted values appear in columns with prefix p_: utilisation rate (p_UT_6); probability of POS and ATM transaction (p_POS Prob and p_ATM Prob); amount of POS and ATM transactional income over six months (p_POS amt 6 and p_ATM amt 6); estimated income from POS and ATM weighted by probability of transaction (p_inc_pos6 and p_inc_atm6); interest income amount over six months (interest_inc 6_pred); transactional income amount (trans_inc 6_pred); and total income weighted (total_inc_weight6 pred).

Table 8.8 The example of accounts data sample with calculations

ID	Month	State	Target_POS sum_6	Target_ATM sum_6	Target_UT_6	limit	p_UT_6	p_POS Prob	p_ATM Prob	p_POS amt 6	p_ATM amt 6	p_inc_pos6	p_inc_atm6	interest_inc6	interest_inc6_pred	trans_inc6	trans_inc6_pred	total_income6	total_income6_pred	total_in_weight6 pred
Observed (OBS)										Predicted (PRE)										
5058	201108	NA	0	0	15000	0.04	0.45	0.42	32.95	190.59	0.00	0.00	0.00	99.41	0.00	0.00	0.00	99.41	38.41	
5058	201109	NA	52.3018	0	0.01	15000	0.03	0.26	0.35	31.42	155.96	0.00	0.00	27.00	83.10	52.30	0.00	79.30	83.10	27.86
5058	201110	NA	52.3018	0	0.07	15000	0.04	0.15	0.29	27.33	74.66	0.00	0.00	189.00	118.76	52.30	0.00	241.30	118.76	35.66
5058	201111	NA	52.3018	0	0.09	15000	0.04	0.11	0.28	27.70	60.65	0.00	0.00	243.00	109.33	52.30	0.00	295.30	109.33	27.82
5058	201112	NA	52.3018	0	0.09	15000	0.04	0.19	0.32	30.68	112.17	0.00	0.00	243.00	102.25	52.30	0.00	295.30	102.25	25.88
5058	201201	NA	52.3018	0	0.09	15000	0.03	0.19	0.27	31.81	123.03	0.00	0.00	243.00	90.68	52.30	0.00	295.30	90.68	21.69
5058	201202	NA	52.3018	0	0.09	15000	0.03	0.20	0.25	32.53	131.91	0.00	0.00	243.00	85.53	52.30	0.00	295.30	85.53	19.47
5058	201203	RE	28.9645	0	0.08	15000	0.17	0.59	0.13	43.88	122.48	43.88	0.00	216.00	461.52	28.96	43.88	244.97	505.41	407.09
5058	201204	RE	43.5411	0	0.05	15000	0.11	0.42	0.10	34.52	58.15	0.00	0.00	135.00	287.11	43.54	0.00	178.54	287.11	214.17
5058	201205	RP	43.5411	0	0.05	15000	0.03	0.48	0.10	41.84	108.31	0.00	0.00	135.00	87.69	43.54	0.00	178.54	87.69	21.56
5058	201206	NA	44.8184	0	0.05	15000	0.02	0.54	0.11	39.72	158.23	39.72	0.00	135.00	43.86	44.82	39.72	179.82	83.58	9.89
5058	201207	NA	69.6481	0	0.06	15000	0.02	0.53	0.17	36.99	194.02	36.99	0.00	162.00	56.90	69.65	36.99	231.65	93.89	11.24
5058	201210	RE	92.1945	0	0.1	15000	0.09	0.80	0.47	42.17	246.19	42.17	0.00	270.00	233.02	92.19	42.17	362.20	275.19	199.22
5058	201211	RP	99.2506	0	0.09	15000	0.03	0.76	0.26	40.23	245.13	40.23	0.00	243.00	84.25	99.25	40.23	342.25	124.49	34.66
1613	201201	RE	0.2243	37.25	1	7000	0.91	0.62	0.94	10.38	99.99	10.38	99.99	1260.00	1149.56	37.47	110.37	1297.47	1259.93	1222.28
1613	201202	RE	0.2171	72	1	7000	0.91	0.60	0.94	10.76	105.90	10.76	105.90	1260.00	1143.87	72.22	116.65	1332.22	1260.52	1230.95
1613	201203	RE	0.2097	73.65	1	7000	0.93	0.71	0.88	8.54	100.39	8.54	100.39	1260.00	1169.04	73.86	108.93	1333.86	1277.97	907.26
1613	201204	D1	0.1675	73.65	1	7000	0.87	0.65	0.79	7.00	114.13	7.00	114.13	1260.00	1100.14	73.82	121.13	1333.82	1221.28	943.67
1613	201205	RE	0.2271	57.4	1	7000	0.93	0.70	0.85	3.31	93.31	3.31	93.31	1260.00	1174.13	57.63	96.62	1317.63	1270.75	609.65
1613	201206	D1	0.2303	57.4	1	7000	0.94	0.73	0.65	1.29	87.01	1.29	87.01	1260.00	1183.83	57.63	88.31	1317.63	1272.13	495.27
1613	201207	D2	0.2286	57.4	1	7000	0.89	0.78	0.30	3.17	130.35	3.17	130.35	1260.00	1118.19	57.63	3.17	1317.63	1121.36	419.93
1613	201208	RE	0.2809	13.9	1	7000	0.89	0.48	0.74	0.91	134.54	0.00	134.54	1260.00	1124.04	14.18	134.54	1274.18	1258.58	623.67
1613	201209	RE	0.3333	0	1	7000	0.92	0.66	0.66	0.24	143.22	0.24	143.22	1260.00	1156.57	0.33	143.46	1260.33	1300.03	99.18
1613	201210	D1	0.3503	0	1	7000	0.92	0.72	0.50	-0.02	142.25	0.00	0.00	1260.00	1164.67	0.35	0.00	1260.35	1164.67	259.59
1613	201211	D2	0.7646	0	1	7000	0.96	0.88	0.23	-2.97	151.61	0.00	0.00	1260.00	1205.05	0.76	0.00	1260.76	1205.05	59.89
7099	201108	TR	60.7617	21.5	0.04	5000	0.12	0.97	0.68	27.78	91.42	27.78	91.42	36.00	107.59	82.26	119.20	118.26	226.79	90.82
7099	201109	RE	56.5367	5	0.03	5000	0.28	0.99	0.99	24.91	161.90	24.91	161.90	27.00	247.57	61.54	186.81	88.54	434.38	436.23
7099	201110	RP	60.4915	5	0.03	5000	0.21	0.98	0.96	24.61	154.03	24.61	154.03	27.00	191.46	65.49	178.64	92.49	370.10	361.22
7099	201111	TR	63.5507	5	0.03	5000	0.14	0.95	0.83	25.94	81.03	25.94	81.03	27.00	125.99	68.55	106.97	95.55	232.95	131.49
7099	201112	TR	36.7764	29.6	0.02	5000	0.13	0.95	0.60	45.26	123.26	45.26	123.26	18.00	118.19	66.38	168.52	84.38	286.72	104.36
7099	201201	TR	33.7564	29.6	0.01	5000	0.10	0.97	0.57	42.85	101.27	42.85	101.27	9.00	88.36	63.36	144.12	72.36	232.48	82.69
7099	201202	TR	29.1015	24.6	0.01	5000	0.12	0.96	0.80	33.21	114.08	33.21	114.08	9.00	105.42	53.70	147.29	62.70	252.70	101.88
7099	201203	NA	34.4677	118.3	0.03	5000	0.07	0.93	0.39	33.48	79.23	33.48	0.00	27.00	67.23	152.77	33.48	179.77	100.71	60.44
7099	201204	TR	37.984	118.3	0.03	5000	0.10	0.97	0.51	32.49	103.02	32.49	103.02	27.00	85.78	156.28	135.51	183.28	221.29	89.17
7099	201205	TR	56.5375	118.3	0.08	5000	0.07	0.96	0.44	34.62	91.73	34.62	0.00	72.00	64.19	174.84	34.62	246.84	98.81	69.16
7099	201206	TR	47.2222	172.7	0.09	5000	0.10	0.98	0.97	29.78	204.44	29.78	204.44	81.00	94.05	219.92	234.23	300.92	328.27	430.06
7099	201207	RE	50.3519	309.8	0.11	5000	0.09	0.98	0.95	30.52	249.00	30.52	249.00	99.00	78.39	360.15	279.52	459.15	357.90	529.88

8.4 Direct Total income prediction with a linear model

In this section we present the estimated coefficients of the OLS model for the direct prediction of total income over a six-month period (Table 8.9).

The total profit prediction for the non-active state is the most difficult. Many predictors are insignificant—the Chi-square values are not close to zero, and as a result, the R-square values are close to 0.05. The delinquent state also contains many insignificant coefficients, but the R-square is very high 0.92. R-squared for the revolver state test sample is around 0.8, the transactor state – 0.7.

Higher age and customer income decrease the total income amount, but these covariates are significant for the revolver state only. The increase of credit limit has a positive impact on the total income for revolver state by 146 money units and for the transactor states by 130 money units for the change of the logarithm of the last month credit limit change (*l_ch1_ln*) by 1 unit.

In case a customer started to use a card for POS transactions, but did not use it for such transactions 4 months ago and earlier (*b_pos_flag_use13vs46*), the total income will in average increase by 34.59 for transactors and by 18.48 for revolvers. However, if a customer used a credit card for ATM cash withdrawals only for the last 3 months (*b_atm_use_only_flag_13*), the total income will in average increase by 62.22 money units for transactors and by 30.07 money units for revolvers. These relationships between total income and predictors have been expected.

The same predictors have the different significance and impact on the predicted income for different states. For accounts in the revolver state the utilisation rate in the previous months has the opposite relationship - higher current utilisation rates mean lower total income over the next six months. However, for accounts in the transactor state, the outstanding balance has a negative impact on the total income, but the utilisation rate for the previous month has a positive impact. It appears that the transactor represents mostly unstable behaviour, and the customer tends i) to take the money and become a revolver in the next months, ii) make large transactions, bringing in income over the next period. The significant utilisation rate for transactors means that (s)he uses a limit actively and because of high amounts has a significant chance

to become a revolver. On the other hand, a revolver with high utilisation will be similar to a cash-user, paying the interest rate, but not making spending transactions.

Table 8.9 OLS Estimation coefficients for Direct Total income

State Variable	Non-active		Transactor		Revolver		Revolver Paid	
	Parameter Estimate	Pr > t						
Intercept	-235.0408	0.8048	1911.9302	0.0783	-333.8257	0.0059	3736.6843	0.0139
mob	-4.4962	<.0001	11.9882	0.0032	-0.4178	0.2975	3.1578	0.2383
limit_1	0.0030	0.0022	0.0199	0.0001	0.0102	<.0001	0.0035	0.2650
UT0_1			389.0922	0.1674	-517.4580	<.0001	197.4470	0.0303
UT0_2	-5.0209	0.9584	-43.3012	0.8189	87.5115	<.0001	-68.4617	0.3835
UT0_3	22.1977	0.6980	250.1062	0.1857	130.5186	<.0001	80.0282	0.3834
UT0_4	7.7120	0.8748	163.7291	0.3672	95.5045	<.0001	57.9348	0.5587
UT0_5	-13.1688	0.7728	4.1587	0.9784	10.6612	0.4652	-56.3959	0.5359
UT0_6	89.6131	0.0058	-62.4943	0.6069	57.4571	<.0001	84.5126	0.2053
b_UT1to2ln	-2.5624	0.0712	2.7755	0.4076	-7.8656	<.0001	-3.5495	0.1574
b_UT1to6ln	1.4662	0.1146	0.4270	0.8808	1.5418	0.0273	1.4633	0.4953
avg_balance_1	1.2493	0.0284	-0.0127	0.8015	0.1046	<.0001	-0.0331	0.0240
avg_balance_2	0.0129	0.4230	0.0204	0.5198	0.0216	<.0001	0.0547	0.0002
avg_balance_3	0.0088	0.3976	-0.0687	0.0588	0.0139	<.0001	0.0207	0.2469
avg_balance_4	0.0064	0.4763	0.0117	0.7442	0.0084	0.0066	0.0028	0.8774
avg_balance_5	-0.0001	0.9898	0.0435	0.1559	0.0130	<.0001	0.0257	0.1604
avg_balance_6	-0.0112	0.0966	0.0284	0.2740	0.0021	0.3089	0.0006	0.9666
avg_deb_amt_1	0.0000	.	0.0126	0.5565	-0.0832	<.0001	-0.0153	0.7788
avg_deb_amt_2	-0.0154	0.5619	0.0389	0.5374	-0.0157	0.0025	0.0550	0.0111
avg_deb_amt_3	0.0057	0.7622	0.1227	0.0114	-0.0063	0.2102	-0.0422	0.0661
avg_deb_amt_4	0.0094	0.5071	0.0095	0.8158	-0.0083	0.0713	0.0274	0.0924
avg_deb_amt_5	-0.0021	0.8550	-0.0362	0.1818	-0.0177	<.0001	-0.0073	0.7334
avg_deb_amt_6	-0.0083	0.3736	-0.0192	0.5390	-0.0219	<.0001	-0.0511	0.0032
sum_crd_amt_1	0.0000	.	-0.0927	<.0001	-0.0752	<.0001	-0.0605	<.0001
sum_crd_amt_2	0.0055	0.6450	-0.0261	0.3584	-0.0471	<.0001	-0.0579	<.0001
sum_crd_amt_3	0.0023	0.8073	-0.0505	0.0537	-0.0302	<.0001	-0.0497	0.0005
sum_crd_amt_4	0.0030	0.6869	-0.0568	0.0223	-0.0144	<.0001	-0.0276	0.0494
sum_crd_amt_5	0.0054	0.3489	-0.0398	0.0710	-0.0031	0.1173	-0.0015	0.8944
sum_crd_amt_6	0.0055	0.0818	-0.0013	0.9284	0.0054	<.0001	0.0129	0.0699
sum_deb_amt_1	0.0000	.	0.2493	<.0001	0.1034	<.0001	0.2958	<.0001
sum_deb_amt_2	0.0557	0.6407	0.1274	0.2075	0.0745	<.0001	0.0724	0.0303
sum_deb_amt_3	-0.0038	0.9036	0.0751	0.2586	0.0500	<.0001	0.0727	0.0171
sum_deb_amt_4	-0.0026	0.9109	0.1759	0.0027	0.0464	<.0001	0.0282	0.2930
sum_deb_amt_5	-0.0216	0.1725	0.0367	0.4216	0.0308	<.0001	-0.0261	0.3439
sum_deb_amt_6	-0.0142	0.2811	-0.0372	0.2872	0.0128	0.0002	0.0467	0.1064
max_deb_amt_1	0.0000	.	-0.1747	0.0042	0.0094	0.1070	-0.2000	0.0079
max_deb_amt_2	-0.0760	0.5349	-0.0960	0.3599	-0.0296	<.0001	-0.0208	0.5476
max_deb_amt_3	0.0025	0.9387	-0.0050	0.9401	-0.0234	0.0001	-0.0206	0.5026
max_deb_amt_4	0.0091	0.6980	-0.1164	0.0423	-0.0233	<.0001	-0.0043	0.8665
max_deb_amt_5	0.0225	0.1626	0.0056	0.8964	-0.0207	<.0001	0.0650	0.0206
max_deb_amt_6	0.0153	0.2748	0.0435	0.2326	-0.0098	0.0070	-0.0394	0.1900
min_deb_amt_1	0.0000	.	0.0156	0.3418	-0.0026	0.4973	-0.0189	0.5796
min_deb_amt_2	0.0117	0.5043	0.0180	0.6605	-0.0167	<.0001	-0.0317	0.0193
min_deb_amt_3	-0.0108	0.3864	-0.0608	0.0562	-0.0172	<.0001	0.0174	0.2259
min_deb_amt_4	-0.0169	0.0941	-0.0098	0.7345	-0.0191	<.0001	-0.0120	0.2195
min_deb_amt_5	-0.0147	0.0478	-0.0132	0.3681	-0.0054	0.0635	-0.0293	0.0310
min_deb_amt_6	-0.0063	0.3327	-0.0099	0.4573	0.0004	0.8387	0.0168	0.1805
b_AvgOB1_to_MaxOB1_ln	0.0000	.	-51.7490	0.7411	-40.5055	<.0001	-65.3040	0.7221
b_AvgOB2_to_MaxOB2_ln	-81.1139	0.2415	45.4964	0.7223	-18.7149	0.0014	13.2845	0.5955
b_AvgOB3_to_MaxOB3_ln	8.0449	0.6367	-42.6858	0.4254	-27.6186	<.0001	-44.0584	0.0903
b_TRmax_deb1_To_Limit_ln	0.0000	.	-58.3982	0.4860	-163.0808	<.0001	-83.3526	0.3465
b_TRmax_deb2_To_Limit_ln	54.1539	0.5106	-145.5845	0.2777	91.4932	<.0001	-76.6131	0.2325
b_TRmax_deb3_To_Limit_ln	-15.0229	0.7643	-166.0484	0.1722	59.2583	<.0001	-9.6660	0.8412
b_TRavg_deb1_to_avgOB1_ln	0.0000	.	-2.5947	0.8145	-8.2637	0.0003	-5.6732	0.5454
b_TRavg_deb2_to_avgOB2_ln	15.0480	0.0246	-2.1298	0.8375	-9.5314	<.0001	-21.0222	0.0497
b_TRavg_deb3_to_avgOB3_ln	0.1795	0.9681	-15.7711	0.2499	-13.5178	<.0001	-16.9930	0.0701
b_TRsum_deb1_to_TRsum_crdln	0.0000	.	-3.5583	0.8076	41.0523	<.0001	9.5616	0.2214
b_TRsum_deb2_to_TRsum_crdln	-9.2941	0.1119	16.9119	0.1131	17.5999	<.0001	19.9808	0.0262
b_TRsum_deb3_to_TRsum_crdln	0.9809	0.7988	22.1813	0.0557	17.2636	<.0001	15.6117	0.0483
b_avgNumDeb13	8.5522	0.2202	6.0188	0.3159	2.6036	0.0002	3.9605	0.3330
b_OB13_to_OB46ln	2.1339	0.0617	0.2408	0.9462	3.2806	0.0502	0.9214	0.8681
b_OB1_to_OB2_ln	0.0000	.	1.8232	0.8011	-6.9187	0.0290	10.9591	0.2533
b_OB2_to_OB3_ln	-1.4075	0.6214	13.0207	0.1170	2.1184	0.1893	4.9143	0.4324
b_OB3_to_OB4_ln	-0.0805	0.9318	6.7997	0.0330	1.3348	0.1408	-1.9457	0.4949
b_pos_flag_use13vs46	63.6665	0.0002	34.5961	0.2497	18.4850	<.0001	5.8515	0.7584
b_atm_flag_use13vs46	-11.8406	0.4746	-2.2626	0.9390	-14.0816	<.0001	-18.7954	0.3569
b_pos_use_only_flag_13	-16.6429	0.2510	-23.3244	0.4652	-36.0445	<.0001	-24.9574	0.2066
b_atm_use_only_flag_13	25.7608	0.0368	62.2215	0.0324	30.0773	<.0001	51.2260	0.0007
b_TRsum_crd1_to_OB1_ln	0.0000	.	7.0594	0.3524	23.5742	<.0001	12.4377	0.1462

State	Non-active		Transactor		Revolver		Revolver Paid	
	Variable	Parameter Estimate	Pr > t	Parameter Estimate	Pr > t	Parameter Estimate	Pr > t	Parameter Estimate
b_TRsum_crd2_to_OB2_ln	-7.3674	0.2126	10.6409	0.2745	23.0466	<.0001	17.1322	0.0299
b_TRsum_crd3_to_OB3_ln	3.7845	0.3017	9.4479	0.3752	12.5294	<.0001	9.4760	0.1664
b_payment_lt_5p_1	-157.8319	0.3478	125.9144	0.3405	12.6907	<.0001	95.0427	0.1332
b_payment_lt_5p_2	122.4910	0.0458	-59.5578	0.6441	-6.1153	0.0126	-31.2332	0.2122
b_payment_lt_5p_3	36.5350	0.0876	-68.1271	0.2659	-9.1963	0.0001	-12.5324	0.5963
b_maxminOB_limit_1_ln	0.0000	.	3.7668	0.6230	48.0957	<.0001	-3.9713	0.4468
b_maxminOB_limit_2_ln	2.1147	0.4936	4.1243	0.6393	-3.4750	0.1444	-5.7140	0.5702
b_maxminOB_limit_3_ln	-3.3808	0.1187	-1.0789	0.8937	0.2888	0.8830	13.8319	0.0407
b_maxminOB_avgOB_1_ln	0.0000	.	-29.8891	0.8154	-51.9694	<.0001	-43.8204	0.8026
b_maxminOB_avgOB_2_ln	-59.6838	0.2650	26.4432	0.7907	2.6231	0.3690	19.4277	0.1790
b_maxminOB_avgOB_3_ln	18.1268	0.0208	-8.5400	0.7462	-1.2263	0.6303	1.5827	0.9002
b_TRsum_deb1_to_2_ln	-0.1100	0.9338	-1.0042	0.8299	-2.4434	0.0699	-2.6522	0.3348
b_TRsum_crd1_to_2_ln	-6.9567	0.2201	1.8017	0.8280	13.5950	<.0001	8.0611	0.2525
l_ch1_ln	179.8925	0.0005	146.3823	0.3688	130.0067	<.0001	1.0564	0.9902
l_ch6_flag	24.8852	0.0008	19.2078	0.5662	28.2985	<.0001	31.2779	0.1212
age	-0.6091	0.0310	-0.9253	0.4384	-0.8403	<.0001	-1.1482	0.1372
customer_income_ln	-1.6400	0.8248	-91.7212	0.0054	-7.9746	0.0079	-1.4031	0.9431
Edu_High	-26.7341	0.0002	-41.6721	0.1616	-16.5252	<.0001	-38.3720	0.0345
Edu_Special	-19.2416	0.0075	-44.2406	0.1460	-4.1288	0.0734	-22.2793	0.2125
Edu_TwoDegree	-40.5420	0.0044	-143.4059	0.0217	-17.0145	0.0088	1.8073	0.9667
Marital_Civ	14.3118	0.2035	1.7144	0.9698	3.5382	0.3492	39.3280	0.1579
Marital_Div	9.4890	0.1971	-21.4041	0.5208	3.4837	0.1883	4.3224	0.8241
Marital_Sin	-3.3218	0.6542	39.3665	0.2485	0.6171	0.8322	-9.3967	0.6458
Marital_Wid	6.5367	0.6155	50.4158	0.4259	19.4113	<.0001	-79.2919	0.0467
position_Man	-0.3482	0.9566	12.0796	0.6747	-3.6818	0.1808	25.8197	0.1510
position_Oth	0.8326	0.9063	10.2911	0.7226	2.9719	0.2350	-4.0766	0.8241
position_Tech	-14.2129	0.0406	-32.1548	0.2555	3.1157	0.1877	0.1943	0.9913
position_Top	9.1111	0.4199	75.7831	0.1489	21.0803	0.0001	29.1430	0.3625
sec_Agricult	12.8569	0.2797	-48.4994	0.3847	1.3601	0.7703	20.3047	0.5174
sec_Constr	8.0387	0.6437	-102.3876	0.1839	2.2730	0.7175	-11.2458	0.8114
sec_Energy	-40.8001	0.0001	-90.0116	0.0634	-16.2030	<.0001	-10.7020	0.7097
sec_Fin	-29.7894	<.0001	-33.9982	0.2361	-42.9095	<.0001	-31.6848	0.0923
sec_Industry	19.3131	0.4712	-70.9464	0.5692	2.1598	0.7838	145.0555	0.0226
sec_Manufact	-38.8963	0.0262	-78.1956	0.2920	10.0629	0.0870	17.5372	0.7195
sec_Mining	31.9662	0.0072	104.1105	0.0417	6.8048	0.1075	12.9988	0.6842
sec_Service	-8.9685	0.1103	-44.2758	0.0728	1.3449	0.5387	12.1376	0.4305
sec_Trade	-22.0818	0.0124	-1.2108	0.9742	-1.2615	0.6898	28.1998	0.2151
sec_Trans	-20.7656	0.2365	-33.8928	0.6060	-15.1999	0.0162	-19.0469	0.6686
car_Own	13.4583	0.0091	34.3382	0.1201	-18.3762	<.0001	-13.9929	0.3265
car_coOwn	4.8163	0.5556	4.2632	0.9054	1.9333	0.5480	27.4163	0.2150
real_Own	4.3655	0.4310	27.0000	0.2491	-3.0400	0.1478	-12.5779	0.4035
real_coOwn	-3.9487	0.5088	-0.2588	0.9918	-3.2531	0.1347	-3.8080	0.8111
reg_ctr_Y	-11.6195	0.1764	-2.0251	0.9538	-8.6583	0.0116	-38.0577	0.0998
reg_ctr_N	2.3505	0.7835	-6.0394	0.8660	-3.0620	0.3643	-38.5265	0.0969
child_1	-1.5569	0.8163	-0.2503	0.9936	5.1440	0.0517	12.1960	0.5117
child_2	5.0335	0.1798	32.9969	0.0607	3.3972	0.0249	1.4564	0.8872
child_3	57.7875	0.0001	166.1472	0.0029	19.7793	<.0001	75.0831	0.0337
Unempl_Infoy_1	204.1431	0.5024	-2104.5697	0.1111	-363.2844	0.0029	-1082.115	0.1975
UAH_EURRate_Inmom_1	-110.2308	0.4986	68.8632	0.9189	-0.4507	0.9946	-362.9538	0.4237
UAH_EURRate_Infoy_1	-218.0091	0.0078	-353.2218	0.3229	-139.8792	<.0001	-426.0816	0.0653
CPI_Inqoq_1	-21.8747	0.9005	411.2813	0.5851	320.0180	<.0001	-134.0550	0.7898
SalaryYear_Infoy_1	194.6950	0.2944	59.9236	0.9421	45.1840	0.5510	146.8802	0.7753
s_month_since_NA_full	0.0000	.	6.5576	0.4541	-2.8154	0.2499	-6.0906	0.5460
s_month_since_Tr_full	4.0120	0.2713	0.0000	.	-0.2357	0.9371	-16.3031	0.1215
s_month_since_Re_full	-11.1590	0.1714	-2.0799	0.9521	0.0000	.	-272.9239	<.0001
s_month_since_RP_full	20.9229	0.0131	15.9932	0.5906	-1.0244	0.8085	0.0000	.
s_month_since_D1_full	-18.7271	0.5282	106.0457	0.3097	13.2627	0.0025	-53.7497	0.0397
s_month_since_D2_full	-4.9917	0.9639	0.0000	.	21.6215	0.0051	-90.5510	0.3090
s_times_NA_full	50.9270	0.3502	-506.2063	0.0833	44.6066	0.0032	-393.6392	0.0259
s_times_TR_full	92.7364	0.0947	-465.6810	0.1116	49.8734	0.0025	-411.0650	0.0204
s_times_RE_full	85.1878	0.1169	-482.6232	0.0983	64.0895	<.0001	-362.8708	0.0379
s_times_RP_full	102.4887	0.0612	-473.2225	0.1064	27.1926	0.1170	-371.4692	0.0344
s_times_D1_full	19.4158	0.8127	19.4158	0.8127	77.7868	<.0001	-345.5933	0.0541
s_times_D2_full	297.0374	0.5024	297.0374	0.5024	122.1262	0.0005	-524.7011	0.2976

R-squared values for direct income OLS models for test sample are: Non-active – 0.35, Transactor – 0.38, Revolver – 0.75, and Revolver Paid – 0.55. The obtained values of the coefficients of determination show the satisfactory predictive accuracy of OLS for the total income direct model.

8.5 Comparative analysis of the validation results of aggregated and direct estimation models for the total income

In this section, we compare the results of the validation of three models: Simple Sum of Interest and Transactional Income (TI), Sum of State Incomes Weighted by State Transition Probabilities (TIW), and Direct Total Income Estimation (TID) Models.

The development samples have the following number of observations: 152,925; out-of-sample: 64,002; total: 216,92; out-of-time: 76,798. However, the number of accounts in the sample are significantly lower: development, 9,919; out-of-sample, 4,160; and 13,601 accounts in out-of-time. The out-of-time validation sample incorporates the last six months of the general data sample—that is, the period from July 2012 to December 2012. Respectively, development and out-of-sample validation samples are randomly selected accounts from the period from July 2010 to June 2012.

For this research, we use the panel data, whereby each account has a time series observation for each characteristic. Thus, each account contains as many records in the data sample as there are periods observed for the account. In Chapter 7, we used the panel data regression at the account level and investigated the random-effect model using various methods.

Although the random-effect models have more efficient estimations of the regression coefficients, the pooled-effect models demonstrate higher fitting accuracy (see Chapter 7 – Comparative analysis of the results of different approaches). We have selected the pooled data approach for the final model because of higher predictive accuracy. Pooled data are presented as a set of independent observations, where an account provides the number of observations by the time series length and each row is assumed as a new independent case. Thus, the same account can provide several observations in the data sample, and depending on the history of account states, the

same account can take part in multiple models; for example, inactive, revolver, and even default states.

This is how 13,601 accounts for Out-of-time samples represent 76,798 observations in total. The same account can appear in the development sample and in the out-of-time validation sample. However, we have assumed that the observations at different periods for the same account are independent and uncorrelated. We are testing our models on the same accounts, but they produce independent cases as we used pooled data in the model building.

The total income predictive models cover a six-month prediction horizon, with six monthly income models and one for the income amount for the entire six-month period. Each of the seven total incomes are estimated using three approaches: Simple Sum of Interest and Transactional Income, Sum of State Incomes Weighted by the State Transition Probability, and Direct Total Income estimation, used for a comparison of simple and complex models. Altogether, we have $7 \times 3 = 21$ models for the prediction of total income. These models should be validated for use with the out-of-sample and out-of-time data, to draw conclusions regarding which model-building approach has provided the most accurate fitting for the sample data.

For the comparison of predicted and observed total income, we used a set of descriptive statistics as measures of central tendency and variability: i) Mean, standard deviation, skewness, the coefficient of variation, and ii) Quantile distribution statistics such as the maximum, minimum, and median. We present the statistics for the validation sample data for two periods: six-month and first-month predictions.

The average predicted total income over six months is less than the observed one by 66.49 or $(644.41 - 710.90) / 710.90 = -9.35\%$. The observed income distribution had a higher standard deviation (581.82 versus 546.11), but a lower coefficient of variation than the predicted income distribution (see Table 8.10).

Table 8.10 Descriptive statistics for Total income for 6 months

CHARACTERISTIC	OBSERVED	PREDICTED
N	64 002	64 002
MEAN	710.90	644.41
STD DEVIATION	581.82	546.11
SKEWNESS	1.30	1.41
COEFF VARIATION	81.84	84.75

Table 8.11 Descriptive statistics for Total income for 6 months: quantiles

<i>Quantile</i>	<i>Observed</i>	<i>Predicted</i>
100% Max	4 536.05	4 561.55
99%	2 642.68	2 470.08
95%	1 813.80	1 680.14
90%	1 448.71	1 337.18
75% Q3	990.84	900.13
50% Median	627.79	559.81
25% Q1	276.50	223.32
10%	18.00	25.58
5%	0.00	13.81
1%	0.00	7.44
0% Min	0.00	0.00

The data was positively skewed for both observed and predicted values. This means that the right tail of the density distribution is longer, and the mass is concentrated in the left part of the distribution. The predicted distribution is more skewed than the observed distribution (skew coefficient 1.41 versus 1.30), and consequently has a longer or heavier right tail and a mass concentration closer to zero than for the observed distribution. The positive skewness and distribution asymmetry correspond to a gap between the mean and median values. The median 627.79 is less than the mean, 710.90, for the observed income distribution, while the median 559.81 is less than the mean, 644.41, for the predicted income distribution (see Table 8.11).

We can attribute the relatively slight difference in the descriptive statistics of the observed and predicted total income distributions to the positive points of the model predictions. The distributions have similar shapes and dispersion.

To the positive points of the model predictions, we can attribute the prediction of some income for the observed zero income values and the difference between means and medians of the observed and predicted total income distributions around 10%. The

excess of the observed overpredicted one may lead to underestimation of the total income. For the first quantile, the difference between the average observed and predicted total income is less in relative measurement than for the fourth quantile; 10% versus 19-38%.

For the shortest prediction horizon of one month, the model shows very similar descriptive statistics values. The mean (115.93 for observed and 114.25 for predicted) and median (102.00 for observed and 100.94 for predicted) values have a difference of 2% and are nearly identical (see Table 8.12). The standard deviation for predicted income is less than for observed one— 98.99 versus 102.49. The shape of the distributions is skewed, with the relatively long right tail and the mass concentration on the left part of the distribution.

Table 8.12 Descriptive statistics for total income for the first month

<i>Characteristic</i>	<i>Observed</i>	<i>Predicted</i>
<i>N</i>	64 002	64 002
<i>Mean</i>	115.93	114.25
<i>Std Deviation</i>	102.49	98.99
<i>Skewness</i>	1.26	1.34
<i>Coeff Variation</i>	88.40	86.64

Table 8.13 Descriptive statistics for total income for first month: quartiles

<i>Quartile</i>	<i>Observed</i>	<i>Predicted</i>
100% Max	799.82	771.68
99%	450.00	442.25
95%	309.52	300.47
90%	247.00	239.82
75% Q3	167.40	161.39
50% Median	102.00	100.94
25% Q1	32.40	36.21
10%	0.00	0.49
5%	0.00	0.25
1%	0.00	0.12
0% Min	0.00	0.00

Figure 8.6 shows that the shape of the density distribution of the observed and predicted total income values are close to each other. The highest concentration of values is observed at the zero and low-income values. However, in this area, the

difference between observed and predicted is the most significant. Several predicted cases have zero total income values.

The density of observed cases with values higher than zero is close to uniform distribution, but the distribution of predicted cases has the high concentration for values a bit greater than zero, with reduction of the distribution to the observed level of the income around 100 (Figure 8.6). This means that the prediction value of the total income at the low-income area may be overestimated.

Two curves in Figure 8.6, Figure 8.7 and Figure 8.8 show the smoothed distribution of observed and predicted values, presented with the histogram. Thin lines show the spline approximation of the distribution function for both observed and predicted values.

Figure 8.6 Total income observed and total income predicted with TID for six-month period (density histogram cut for high frequent values).

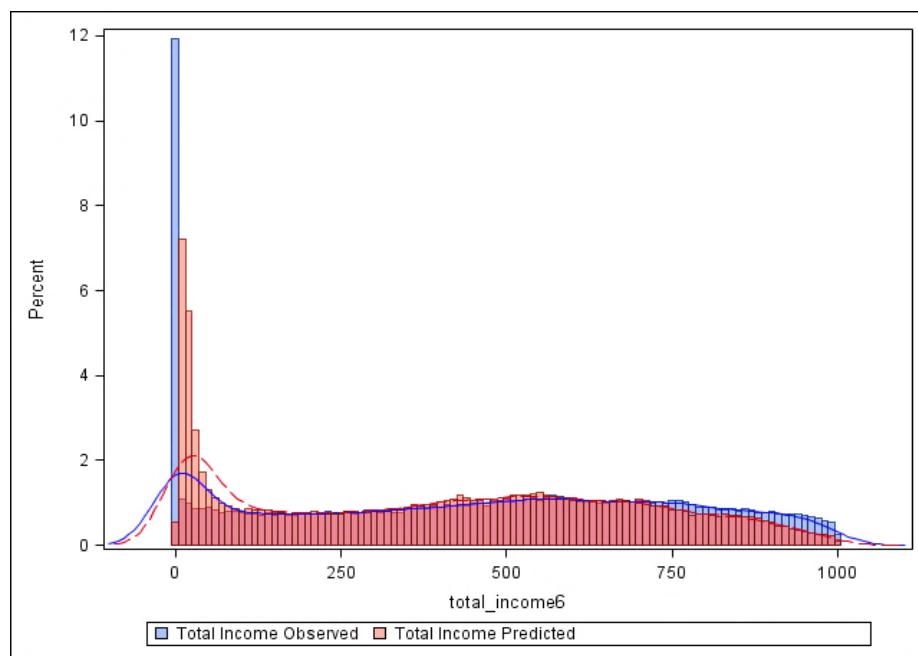


Figure 8.7 Total income observed and total income predicted with TID for six month period density histogram for all values.

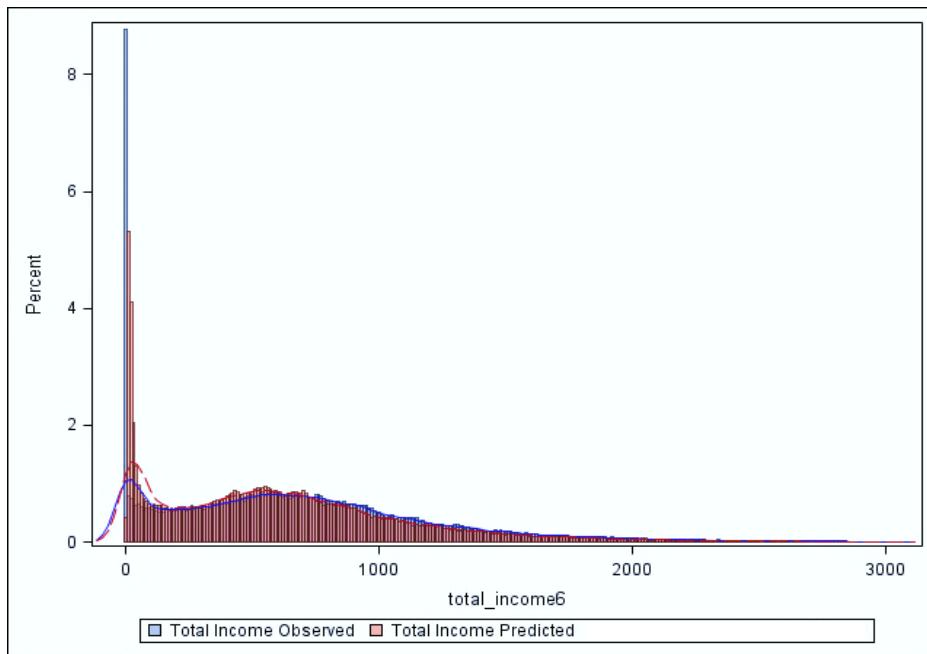
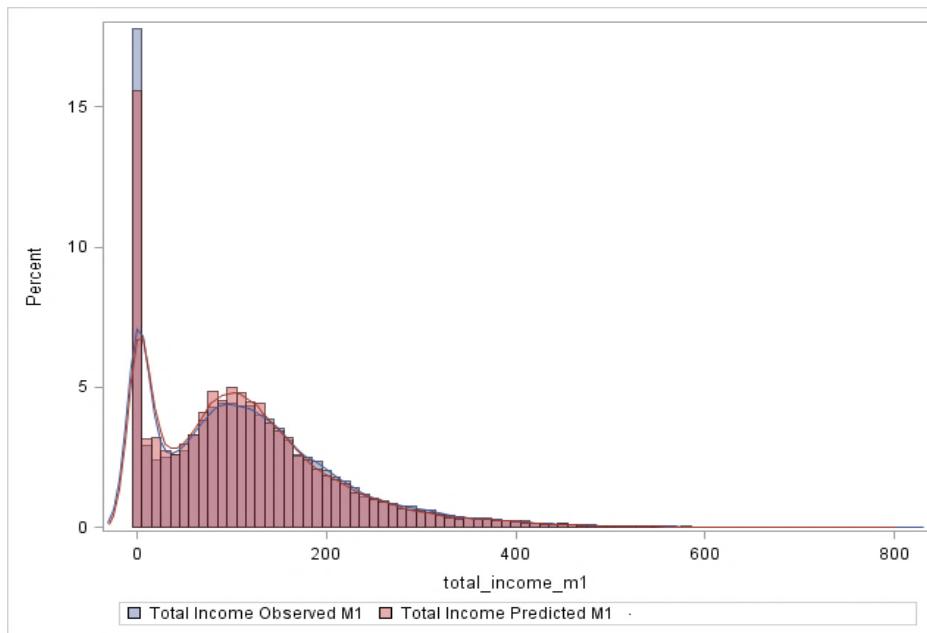


Figure 8.8 Total income observed and total income predicted with TID for +1 month period density histogram.



For validation of the models, we have selected three statistical coefficients: the coefficient of determination (R-squared), Mean Absolute Error (MAE), and Root-mean-square error (RMSE). The use of R-squared, which is applied to linear models, is possible because we assume that the final estimation from the aggregated model is

a linear combination of some estimations (interest and transactional income), obtained from other models, but is a non-linear function. The higher R-squared and lower MAE and RMSE values indicate the better fitting accuracy of the model than in the opposite case. A detailed description of the statistical indicators, selected for the models' validation, appears in Chapter 4.

Table 8.14 contains R-squared, MAE, and RMSE values in three columns for the development, validation out-of-sample, and validation out-of-time data sets for four aggregated models: the simple sum of incomes and the sum of state incomes weighted by the state transition probabilities, sum of state incomes weighted by the states individual transitions probabilities, and one direct total income estimation.

The main difference between Sum of State Incomes Weighted by TMM (TIW1) and Sum of State Incomes Weighted by the States Individual Transition Probabilities (TIW2) models is that TIW1 uses the transition frequencies from the transition matrix from section 6.4 of Chapter 6 as weights for income from different account states and TIW2 uses the individual transition probabilities obtained from the final multinomial regression models from section 6.5 of Chapter 6.

Table 8.14 Validation results of aggregated and direct for the total income prediction

Model name	Development Sample			Validation - Out-of-sample			Validation - Out-of-time		
	R^2	MAE	RMSE	R^2	MAE	RMSE	R^2	MAE	RMSE
Simple Sum of Interest and Transactional Income (TI)									
month 1	0.7673	24.50	47.40	0.7663	24.71	46.97	0.7420	30.18	57.38
month 2	0.7077	29.99	53.26	0.7040	30.27	53.00	0.6538	38.65	67.19
month 3	0.6636	33.88	57.39	0.6553	34.16	57.47	0.5890	45.05	73.69
month 4	0.6239	37.02	60.59	0.6194	37.20	60.36	0.5339	50.18	79.00
month 5	0.6006	39.11	62.23	0.5919	39.38	62.29	0.5030	53.46	81.91
month 6	0.5517	42.50	66.37	0.5422	42.92	66.54	0.4566	56.17	84.72
6 months	0.8336	145.01	227.76	0.8227	148.44	232.19	0.8181	184.41	273.37
Sum of State Incomes Weighted by TMM (TIW1)									
month 1	0.7866	23.75	45.39	0.7817	24.08	45.40	0.7693	28.85	54.26
month 2	0.7237	29.72	51.78	0.7182	30.00	51.71	0.6790	37.39	64.70
month 3	0.6746	34.27	56.44	0.6671	34.40	56.47	0.6143	43.56	71.39
month 4	0.6339	37.81	59.78	0.6302	37.85	59.50	0.5613	48.51	76.64
month 5	0.6052	39.98	61.87	0.5980	40.12	61.83	0.5210	52.18	80.41
month 6	0.5572	43.86	65.96	0.5509	44.16	65.91	0.4773	54.80	83.09
6 months	0.7924	164.12	254.40	0.7843	166.31	256.11	0.7588	204.43	314.77
Sum of State Incomes Weighted by the States Individual Transition Probabilities (TIW2)									
month 1	0.8053	23.18	43.36	0.8011	23.52	43.33	0.7943	27.94	51.24
month 2	0.7323	29.12	50.97	0.7244	29.48	51.14	0.6947	36.72	63.10
month 3	0.6850	33.25	55.53	0.6752	33.52	55.78	0.6300	42.78	69.92
month 4	0.6443	36.61	58.92	0.6374	36.79	58.92	0.5735	47.84	75.57
month 5	0.6136	39.16	61.21	0.6054	39.36	61.26	0.5296	51.75	79.69
month 6	0.5700	42.39	65.00	0.5614	42.79	65.14	0.4877	54.14	82.26
6 months	0.7928	161.45	254.17	0.7821	163.96	257.41	0.7464	207.94	322.80
Direct Total Income Estimation (TID)									
month 1	0.80746	23.708	43.115	0.79411	25.817	46.207	0.78591	29.683	51.721
month 2	0.73837	29.934	50.382	0.71303	32.944	54.908	0.69375	38.386	62.533
month 3	0.68888	34.591	55.185	0.65646	37.695	60.404	0.63445	43.35	68.77
month 4	0.64937	37.804	58.501	0.61191	41.784	64.4	0.58534	48.87	73.726
month 5	0.62435	39.891	60.354	0.58383	44.25	66.764	0.55086	52.187	77.051
month 6	0.58083	43.299	64.173	0.5366	47.908	70.448	0.50104	55.827	80.325
6 months	0.8357	146.28	226.33	0.8201	162.43	245.22	0.8187	185.80	270.02

First, we compare the fitting accuracy of two aggregated models for the entire period prediction and for the monthly predictions. The Simple Sum of Interest and Transactional Income Model (TI) shows higher fitting accuracy than the Sum of State Incomes Weighted by the States Transition Probabilities Model (TWI1) over the six-month period. Simple Sum of Interest and Transactional Income (TI) has a coefficient of determination equal to 0.8336 for the development sample, 0.8227 for the validation out-of-sample, and 0.8181 for validation out-of-time data set. On the other hand, the Sum of State Incomes Weighted by the States Transition Probabilities (TIW1) model has shown lower values of the coefficient of determination, equal to 0.7924 for the development sample, 0.7843 for the validation out-of-sample, and 0.7588 for the validation out-of-time data set.

validation out-of-time data set. Also, the values of Mean Absolute Error and Root-Mean-Square Error of the first model is less than the errors values of the second model. The Simple sum of incomes model has MAE and RMSE values equal to 141 and 227, 148 and 232, and 184 and 273 for the development, out-of-sample, and out-of-time data sets versus 164 and 254, 166 and 256, and 204 and 314 for the respective data sets for the Sum of State Incomes Weighted model. The best fitting accuracy results are highlighted with bold in the corresponding row for Out-of-sample and Out-of-time in Table 8.14. Final results for the development sample are not considered in the model selection procedure.

Thus, the simple summation of monthly income (TI) provides more accurate predictions than the model, which considers the transition between states (TIW1 and TIW2). This can be explained by the incorporation of the inaccuracy of estimations, because the transition probability models have moderate predictive power (GINI Indexes around 0.5 – 0.6) and Type I and Type II estimation errors. The incorporation of errors in six submodels results in the decrease of the total income fitting accuracy for the six-month period prediction.

Analysis of the fitting accuracy of monthly prediction reveals the opposite picture. The coefficient of determination values for one-month total income prediction (equal to 0.8011) is higher for the second model Sum of State Incomes Weighted by the States Transition Probabilities (TIW1) than for the first model Simple Sum of Interest and Transactional Income (TI), which has shown 0.7663 for the validation sample. The MAE and RMSE values for the second model are also lower than the error values for the first model for the development, out-of-sample, and out-of-time datasets respectively.

The predictive accuracy for higher order months decreases with each month. For example, the coefficient of determination for the model Sum of State Incomes Weighted by the States Individual Transition Probabilities (TIW2) for the Validation Out-of-time sample is equal to 0.7244 for month 2, 0.6752 for month 3, 0.6374 for month 4, 0.6054 for month 5, and only 0.5614 for month 6. The same trend is observed for the development and validation out-of-sample data sets. MAE and RMSE values increase with the higher month order what also indicates the decrease in the model's

fitting accuracy. The Simple Sum of Interest and Transactional Income (TI) model has the same tendency – an increase in the prediction horizon results in decreased fitting accuracy of the model. However, the TIW2 model shows higher coefficients of determination and lower MAE and RMSE values for all monthly predictions than TI model.

The advantages of the Sum of State Incomes Weighted by the States Individual Transition Probabilities TIW2 model for monthly income prediction can be explained by the fact that the future state is a significant parameter for predicted income and the fitting accuracy of such models is higher than the simple sum of incomes TI.

Because of only one period and one estimated set of probabilities for the transition probability, the cumulation of estimation errors does not occur. In Markov Decision Processes the overlapping for high-order prediction horizons can be observed. It happens when the prediction for n months is given as the multiplication of the transition matrix by itself $n-1$ times. In the current model the probability of transition of the account is estimated for a n month period at the account level, but not as an $n-1$ product of the 1-period estimation with MDP at the portfolio or pool level. So, the account level approach with a n -month prediction and the probability of transition weighting (TIW2) can improve the accuracy of prediction in comparison with Markov Decision Processes (TIW1). For the Transition Matrices see the section of Chapter 6. We have tested Sum of State Incomes Weighted by the States TMM (TIW2) with formula (8.8) and the results with the weighting by the individual probability of transition show better fitting accuracy.

Direct Total Income Estimation (TID) also shows good results for the prediction of the total income for a 6 months period and for month 4, month 5, and month 6 prediction. However, the prediction accuracy for month 1, month 2, and month 3 income is also less than for weighted method TIW2. Direct estimation of total income demonstrates good predictive accuracy for the total income but does not give a view of the income components.

Backtesting with confusion matrix

An alternative measurement of a model's fitting accuracy is the confusion matrix or backtesting of the number of observed values versus the number of predicted values

in each selected bin (or segment). For total income, the bin ranges have been designed to have ranges of 50 money units. In the case of high accuracy of prediction, many observations would concentrate around the diagonal, and few observations would be at the top right and a bottom left corners far from the matrix diagonal. This means that for many cases we predicted income values close to observed income, and for a small number of cases we have predictions with a high deviation. The observed versus predicted backtesting table could be used as an analogue of mean and standard deviation measurement of residuals.

For total income with a $t+1$ month prediction period the results are shown in Table 8.15, the significant number of cases are on or close to the diagonal. For the range of observed income values from 0 to 49 the model has predicted the same range for 80.75% of cases, and for 12.90% the predicted values are higher – from 50 to 99, and for only 4.21% of cases the model has given overestimated income values from 100 to 149. For the range of observed income values from 50 to 99 the model returned 63.72% of cases in the same range, 15.18% of cases are underestimated with predicted values between 0 and 49, and for the other 21.1% of cases the predicted income is overestimated, but 18.25% of these cases are in the next bin 100-149. For higher income values the concentration of cases at the same income bins is lower than for low income, but dispersion is not high. For example, in a bin of income 750-799 the model has predicted the same values for 43.24% cases, and for the remaining cases the income is underestimated, but mainly in the range from 650.

Table 8.15 Confusion matrix of number of observed vs. predicted values for total income t+1 month

Predicted Observed	0-49	50-99	100-149	150-199	200-249	250-299	300-349	350-399	400-449	450-499	500-549	550-599	600-649	650-699	700-749	750-800	800-849
0-49	80.75%	12.90%	4.21%	1.18%	0.50%	0.23%	0.11%	0.06%	0.03%	0.01%	0.01%	0.01%	0.00%	0.00%	0.00%	0.00%	0.00%
50-99	15.18%	63.72%	18.35%	2.24%	0.34%	0.10%	0.06%	0.01%	0.01%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%
100-149	4.21%	15.66%	66.69%	11.89%	1.24%	0.23%	0.04%	0.03%	0.01%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%
150-199	3.11%	4.29%	21.58%	59.07%	10.40%	1.30%	0.19%	0.04%	0.01%	0.01%	0.01%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%
200-249	2.61%	2.65%	5.87%	22.08%	53.18%	11.36%	1.87%	0.25%	0.07%	0.04%	0.02%	0.01%	0.01%	0.00%	0.00%	0.00%	0.00%
250-299	2.79%	2.01%	3.98%	6.95%	22.26%	47.72%	11.77%	2.00%	0.40%	0.10%	0.00%	0.01%	0.00%	0.00%	0.00%	0.00%	0.00%
300-349	2.15%	1.27%	2.54%	4.64%	7.53%	21.99%	46.44%	10.45%	2.34%	0.47%	0.13%	0.02%	0.04%	0.00%	0.00%	0.00%	0.00%
350-399	1.75%	1.29%	2.11%	2.90%	4.81%	7.28%	19.93%	44.58%	12.29%	2.24%	0.56%	0.20%	0.07%	0.00%	0.00%	0.00%	0.00%
400-449	1.14%	0.69%	1.26%	1.77%	3.03%	4.75%	7.67%	23.87%	39.27%	12.76%	2.46%	1.14%	0.06%	0.06%	0.00%	0.00%	0.00%
450-499	0.89%	0.53%	0.98%	1.87%	1.87%	4.01%	5.08%	7.13%	20.23%	42.60%	11.59%	2.32%	0.62%	0.27%	0.00%	0.00%	0.00%
500-549	0.15%	0.76%	1.07%	0.92%	1.22%	1.68%	3.36%	4.27%	6.26%	19.24%	42.14%	15.88%	2.75%	0.15%	0.15%	0.00%	0.00%
550-599	0.22%	0.22%	0.66%	0.22%	0.88%	0.88%	1.76%	3.96%	7.05%	7.27%	21.81%	35.02%	17.84%	1.54%	0.66%	0.00%	0.00%
600-649	0.00%	0.00%	0.00%	0.00%	0.85%	0.85%	2.56%	2.99%	5.56%	10.26%	25.21%	33.76%	14.10%	2.56%	0.43%	0.00%	0.00%
650-699	0.88%	0.00%	1.77%	0.00%	0.88%	1.77%	0.88%	1.77%	3.54%	6.19%	7.96%	16.81%	38.94%	17.70%	0.88%	0.00%	0.00%
700-749	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	6.10%	8.54%	8.54%	26.83%	30.49%	19.51%	0.00%	0.00%
750-799	0.00%	0.00%	0.00%	0.00%	2.70%	2.70%	0.00%	0.00%	2.70%	0.00%	2.70%	5.41%	21.62%	18.92%	43.24%	0.00%	0.00%
800-849	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	14.29%	0.00%	0.00%	0.00%	14.29%	28.57%	42.86%	0.00%	0.00%

Table 8.16 Confusion matrix of number of observed vs. predicted values for total income for 6-month period

Predicted Observed	<100	100-299	300-499	500-699	700-899	900-1099	1100-1299	1300-1499	1500-1699	1700-1899	1900-2099	2100-2299	2300-2499	2500-2699	2700-2899	2900-3099	3100-3299	3300-3500	>3500
<100	77.79%	16.09%	3.87%	1.42%	0.50%	0.21%	0.06%	0.03%	0.02%	0.01%	0.01%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%
100-299	27.90%	40.38%	20.78%	7.11%	2.28%	0.86%	0.33%	0.14%	0.10%	0.05%	0.02%	0.02%	0.02%	0.01%	0.00%	0.00%	0.00%	0.00%	0.00%
300-499	10.23%	19.59%	45.22%	17.85%	4.58%	1.54%	0.46%	0.28%	0.13%	0.04%	0.04%	0.01%	0.01%	0.02%	0.00%	0.00%	0.00%	0.00%	0.00%
500-699	3.98%	7.81%	23.77%	47.63%	12.23%	2.93%	0.97%	0.36%	0.15%	0.08%	0.03%	0.03%	0.01%	0.01%	0.00%	0.00%	0.00%	0.00%	0.00%
700-899	2.22%	4.41%	8.84%	31.00%	40.87%	9.33%	2.08%	0.74%	0.25%	0.12%	0.06%	0.03%	0.03%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%
900-1099	1.61%	2.52%	5.29%	11.07%	35.72%	32.85%	7.68%	2.12%	0.59%	0.24%	0.12%	0.08%	0.05%	0.04%	0.01%	0.00%	0.00%	0.00%	0.00%
1100-1299	1.11%	1.93%	3.06%	5.72%	12.32%	35.39%	29.37%	7.77%	2.13%	0.69%	0.23%	0.11%	0.07%	0.01%	0.03%	0.03%	0.01%	0.01%	0.01%
1300-1499	0.73%	1.35%	2.18%	3.46%	5.78%	13.30%	35.60%	26.89%	6.92%	2.26%	0.77%	0.31%	0.21%	0.09%	0.05%	0.02%	0.03%	0.03%	0.00%
1500-1699	0.66%	1.25%	1.46%	2.70%	4.14%	6.84%	14.90%	32.73%	25.11%	6.47%	2.14%	0.76%	0.37%	0.18%	0.19%	0.05%	0.03%	0.02%	0.02%
1700-1899	0.54%	0.92%	1.16%	1.77%	2.46%	3.95%	6.64%	15.03%	30.74%	26.54%	6.10%	2.22%	0.80%	0.52%	0.28%	0.14%	0.02%	0.12%	0.05%
1900-2099	0.31%	0.59%	0.83%	1.56%	1.66%	2.74%	4.47%	7.77%	15.95%	35.58%	18.65%	5.58%	2.84%	0.59%	0.42%	0.31%	0.07%	0.07%	0.00%
2100-2299	0.10%	0.46%	0.76%	1.22%	1.73%	1.98%	3.20%	4.77%	7.97%	16.69%	31.56%	5.63%	1.73%	1.01%	0.36%	0.05%	0.20%	0.10%	0.00%
2300-2499	0.24%	0.39%	0.63%	0.55%	0.71%	1.97%	1.57%	2.44%	4.87%	7.86%	15.09%	37.81%	16.75%	5.42%	1.65%	1.02%	0.31%	0.39%	0.31%
2500-2699	0.11%	0.77%	0.66%	0.44%	1.21%	0.99%	0.88%	1.54%	3.63%	3.74%	7.71%	18.94%	31.50%	19.49%	4.41%	2.42%	0.99%	0.44%	0.11%
2700-2899	0.65%	0.78%	0.39%	1.16%	0.39%	1.55%	1.55%	2.20%	2.58%	3.10%	4.39%	6.72%	15.50%	36.82%	15.25%	3.49%	1.55%	1.42%	0.52%
2900-3099	0.23%	0.23%	0.45%	0.23%	1.13%	0.00%	0.23%	1.80%	1.13%	2.03%	4.28%	6.08%	9.23%	22.07%	21.62%	16.89%	5.86%	4.28%	2.25%
3100-3299	0.00%	0.27%	0.27%	0.54%	0.27%	0.00%	0.27%	1.35%	1.62%	2.97%	3.51%	3.78%	2.70%	10.27%	18.92%	32.43%	14.05%	4.86%	1.89%
3300-3500	0.00%	0.30%	0.00%	0.30%	0.30%	0.61%	0.00%	0.61%	0.61%	1.52%	2.12%	2.73%	3.03%	3.64%	15.15%	17.88%	27.58%	19.09%	4.55%
>3500	0.00%	0.00%	0.00%	0.24%	0.00%	0.95%	0.00%	0.00%	0.47%	0.24%	0.71%	1.65%	0.71%	1.18%	2.60%	2.84%	11.35%	26.71%	50.35%

Total income for the six-month period (Table 8.16) has also shown a concentration along the diagonal of observed versus predicted values. The size of the range is selected as 200 because of higher values of income. For low-income values the concentration in the same range is less than for one-month model, 77.79% and 40.38% for two initial bins <100 and 100-299 respectively. This means that total income model for six months has a higher dispersion of predicted values and the model predict less accuracy than income model for one month, which is confirmed by R-squared, MAE, and RMSE indicators.

This type of model validation by use of confusion matrices can be even more informative than the use of a coefficient of determination and error indicators. In the observed versus predicted matrix we can see, firstly, the directions of biases in prediction as overestimation and underestimation, and secondly, the areas where the

model has weak and strong points. For example, a model with a low coefficient of determination can have strong fitting accuracy in some ranges, but not often. Another important issue for assessing a predictive model's validity is the objective of the model in terms of how we are going to use it: i) try to guess value close to observed at the account level, or ii) get average values for pools or the portfolio close to the average for the observed portfolio.

Thus, the total income as a part of the profit estimation by a general model schema (see Introduction) can be predicted with sufficient goodness-of-fit. High outliers can be observed for extreme values at the tails of distributions, but for mass accounts, or the area of the concentration of accounts, the predictive accuracy of the estimations is satisfactory or good and meet or exceed analogues of models in the area of credit cards performed with similar regression analysis techniques. The results of the total income modelling can be accepted both for i) academic research as some benchmarks and approaches, and for ii) business use as a model for implementation.

8.6 Expected Loss Modelling

In previous sections, we have built the estimation for the first part of a general model for credit card profit estimation. The total profit in accordance with the general model scheme (see Introduction) has two parts: income and expenditures. We have modelled income in Chapters 3 to 7. For expenditures, we need to estimate both operating expenditures and losses. We have simplified the model and consider risk-costs only, assuming that any other outflows such as funding costs or operational and administrative expenses, are beyond the scope of this research. So, the total profit, calculated in the model, is the profit from the lending activities with risk-costs only.

Within this study, we define risk-costs as expected loss. The expected loss is calculated in accordance with the Basel II Advanced IRB approach (BIS, 2005) formula as:

$$EL = PD \times LGD \times EaD,$$

where PD is the probability of default,

LGD is the Loss Given Default,

EaD is exposure at default.

At the account level for the account i at time t the Expected Loss is computed as follows:

$$EL_{i,t} = PD_{i,t} \times LGD \times EaD_{i,t}. \quad (8.9)$$

The default definition is 3 or more missing payments; that is, the equivalent of DPD 90+ (days past due = 90 or more). The default state is an absorbing state, meaning that an account is not able to return to any other state from the defaulted state.

The Probability of Default is calculated as the probability of an account having the default state at any time during the performance period. The performance period for the current investigation within this research is defined as six months. Generally, according to the Basel II Accord, the performance period for default is taken as twelve months. However, we reduced this period because of business logic—we estimate profit for six months and the default state can be achieved within three months of the current non-delinquent state of the account. Therefore, the probability of default prediction is equivalent to the transition to the default state by the sixth month from any current state. The probability of default is calculated by the account state transition formula from Chapter 5 with the use of, for example, logistic regression as:

$$PD = \Pr(s_{i,t+1} = 'Df' | s_{i,t} \neq 'Df') \Rightarrow \ln\left(\frac{p_i}{1-p_i}\right) = \beta^T \cdot \mathbf{X} \quad (8.10)$$

The default transition probability model has good predictive power: KS = 75.80, Gini = 0.898 for validation sample. For more details see Chapter 5.

The exposure at default for the account i at time $t+6$ is estimated as the outstanding balance at the time of default as:

$$EaD_{i,t+6} = Limit_{it} \times UT_{i,t+6} \times CCF, \quad (8.11)$$

where

$Limit_{it}$ is the credit limit of account i at the prediction point, and it is assumed that the credit limit is constant for all performance periods,

$UT_{i,t+6}$ is the utilisation rate of account i at time $t+6$,

CCF is a credit conversion factor. CCF is the expected gross per cent change in the total commitment or the ratio of the outstanding balance at the time of default to the outstanding balance at the time of observation.

Because we use account level prediction of the utilisation rate and the outstanding balance in six months at the account level, we do not use CCF, because it is expected that PD (or the Default state transition probability) is correlated with the outstanding balance prediction for defaulted cases. The concept with CCF, which is considered indirectly in the utilisation rate, can be written as a simplified model for Exposure at Default as follows:

$$EaD_{i,t+6} = Limit_{i,t} \times UT_{i,t+6} \quad (8.12)$$

The data sample consists of all account from total data sample, used for the total income prediction, which have moved into the default state at any time. The data sample does not contain information about recoveries. Therefore, the loss given default value is defined as a constant and equal to 1 for the most conservative estimate. As the value is constant, it has no impact on the shape of the expected loss distribution, but on the absolute value of losses only.

The final model for the estimation of Expected loss as for case i at time t for a N month prediction period is the following:

$$EL_{i,t} = Pr(s_{i,t+N} = Df | s_{i,t} \neq Df) \times Limit_{i,t} \times UT_{i,t+N} \quad (8.13)$$

where

$Pr(s_{i,t+N} = Df | s_{i,t} \neq Df)$ is the probability of transition to the default state from a non-defaulted state,

$s_{i,t+N}$ is the state of the account i at time $t+N$.

8.7 Total Profit and Profitability

As a final stage of the model calculations, we introduce two new categories:

Expected Profit and Expected Profitability. We use a simplified definition of profit, as we do not use such concepts as discounting, time value of money, or accounting standards. We consider risk cost to be expressed solely by the Expected Loss, but

assume the funding costs, administrative and operational costs to be constants, with no impact on the shape of profit. Profit in the model discussed here is therefore dependent on the expected income and expected loss only.

Expected total profit is calculated as a difference between the expected total profit and expected loss:

$$ETP_{it} = TotalIncome_{it} - EL_{it},$$

where i is an account, t is time at the observation point or scoring time.

Expected profit is calculated both at the account and at the portfolio level. We use a prediction horizon of 6 months. Expected Total Profitability is the difference between the expected total profit and expected loss divided by the credit limit:

$$ETPR = \frac{TotalIncome_{it} - EL_{it}}{Limit_{it}},$$

where $Limit_{it}$ is the credit limit of the account i at time t .

We predict the profit for a 6 months period as the difference between the sum of 6-month income and Expected Loss. Because of the different number of observations for each state, we use average profit values for the comparative analysis of the profit by states.

The observed and predicted values for average profit and profitability with distribution by the states for the whole portfolio used in training and testing samples is given in Table 8.17. The credit limit, average profit and average profitability values average are computed over all accounts and over all available observation points, the same as if we take the same account and predict the profit in January observation point for 6 months till July, in February observation point for 6 months to August and so on for all accounts for all available dates, and then compute the average value for the profit and profitability estimates. The highest observed and predicted profit values as well as the profitability value are observed for the revolver state, 809.56 and 13.1% respectively. Transactors and repaid revolver generate significantly lower average profit and profitability of 104 and 122 money unit and 1.8% and 2.3% respectively. However, all predicted values are lower than the observed values, so the profit and profitability are mainly underestimated. This can be cause by the highly pessimistic

assumption for Loss Given Default Value (equal to 1). The low LGD will result in the lower Expected Loss and cause an increase in estimated profit (average profit predicted and profitability predicted). As expected, negative profit (i.e. losses) appears in the state Delinquent 2, because its position is closest to the Default state. Other states have positive profit values. Delinquent 1 state, on the other hand, has positive observed profit but negative predicted profit.

Profitability is a relative measure of the efficiency with which lend funds generate profit. The most efficient account state in the sense of profitability is the revolver state. Transactors have generated low income in comparison with the allocated sources for the credit limit.

Table 8.17 Total profit and profitability for portfolio for all time

State	Average Credit Limit	Average Profit	Average Profit Predicted	Profitability	Profitability Predicted	% of population
NA	7101	112.53	21.30	2.0%	0.4%	12.7%
TR	7210	213.04	104.45	3.8%	1.8%	1.1%
RE	6506	809.56	735.96	13.1%	12.1%	81.3%
RP	6962	239.91	122.40	4.3%	2.3%	2.7%
D1	5706	630.22	-164.70	8.9%	-3.5%	1.9%
D2	4634	-541.35	-1640.19	-18.7%	-38.2%	0.3%

We have selected one time point in the development sample to check the profit and profitability values for one observed month at some fixed time point of the portfolio. The average profit and profitability by states for one month at the observation point using the development sample does not demonstrate a significant difference between average profit and profitability for the whole portfolio, and there used for model training and testing. So, the distribution of the estimates at the time point is not biased significantly from the distribution of the portfolio for the whole period.

However, we have got a significant bias between predicted and observed profitability at the pool level for all states except the revolver state (see Table 8.18). For the revolver state, the total observed profitability is slightly higher than the predicted one: 12.7% versus 11.6%. For a 1-month period the highest profitability after the revolver state is delinquent 1, but predicted profitability for delinquent 1 is negative (-4.2%). The model should be improved for use in business for profit amount estimation.

Table 8.18 Profit and profitability for portfolio for one month fixed observation point of the total data sample

State	Average Credit Limit	Average Profit	Average Profit Predicted	Profitability	Profitability Predict	% of population
NA	7506	101.93	18.27	1.7%	0.3%	13.6%
TR	8606	214.50	110.40	3.3%	1.6%	1.1%
RE	7242	861.36	775.36	12.7%	11.6%	81.1%
RP	7706	254.51	101.66	4.2%	1.8%	2.3%
D1	5544	516.72	-205.09	7.0%	-4.2%	1.6%
D2	5052	-124.00	-1702.32	-8.0%	-35.1%	0.4%

We assess the predictive accuracy of the total profit model as if it were a linear model to compare how well the aggregated model predicts the profit and the profitability values. The profit and profitability predictive accuracy is not high (see Table 8.19), but satisfies the benchmarks from the credit cards modelling sources, for example, R-squared for credit card usage, 0.3919 (linear bivariate correlation as alternative calculation of R-squared) (Banasik and Crook, 2001), outstanding credit card balances, 0.30 (Kim & De Vaney, 2001). These values are not related directly to income prediction, but are related to outstanding balance and usage, which are generally correlated with total income. The Validation – Out-of-Sample has lower Mean Absolute Error equal to 226.92 and Root Mean Standard Error equal to 569.27 in comparison with Validation – Out-of-time sample – 304.73 and 792.02 respectively. This means that at the last 6 months of the total data sample, which was selected as the out-of-time sample, we can investigate some changes in customers' behaviour and credit policy. For example, the credit limit increase process was run before the out-of-time sample and some profit drivers and features correlations with outcome, which could be observed after credit limit increase, might not be reflected in the development sample. R-squared is computed and the values can be considered as appropriate values – 0.38 for the profit estimation and 0.30 for the profitability. However, because the model is not a linear regression model and R-squared values computed as for conditionally linear dependence, these values are only informative.

Table 8.19 Total profit and profitability predictive accuracy

Model	Development Sample			Validation - Out-of-Sample			Validation - Out-of-time		
	R^2	MAE	RMSE	R^2	MAE	RMSE	R^2	MAE	RMSE
Profit	0.4395	230.42	589.21	0.4653	226.92	569.27	0.3815	304.73	792.02
Profitability	0.3102	0.04	0.10	0.3325	0.04	0.09	0.3081	0.05	0.11

The total profit and profitability portfolio models can be used for further research in academic and business applications. However, they require some review of assumptions and, possibly, the use of other methods as, for example, a consideration of non-linear dependencies, machine learning techniques in addition to regression models, or other approaches to model building and a client's segmentation to increase the predictive accuracy of the total model.

8.8 Scenario analysis and business contribution examples

8.8.1 Impact of the different behavioural scenarios on the total income from the credit card

The developed model concept and coefficient estimates can be used for analyse of ‘what if’ scenarios. For the same set of the initial parameters, we can simulate different customer behaviour scenarios. This can be done both manually, and by changing the input parameters at the model coefficient estimation stage. In the case of the manual scenario we can, for example, adjust the probability of default to increase the expected loss estimation and assess the decline in profit; or we can increase the probability of an account transitioning to a transactor state to see how the total income will change. The change in parameters means the marginal analysis in the estimated model as the investigation of how some changes in macroeconomic factors, which are included into the model, will affect the expected profit. For example, an increase in the unemployment rate can decrease interest income and increase expected loss.

In Table 8.20 we show a brief example of a scenario to how examine different probabilities of the transition to states can impact on the total income for given constant interest and transactional income. The scenario contains four cases. The credit limit and the interest rate (rows 1 and 2 respectively) are given as 10000 money units and 36% annually. The states transition probabilities are given from the scenario. We can see that scenarios 3 and 4 have higher transactor transition probabilities – 0.6 instead of 0.2. Scenarios two and four have a low utilisation rate of 0.1 instead of 0.5 for the first and third one. The transactional income is the same for all scenarios and equal to 105 for revolver and 184 for transactor states. This means that the same account in the same conditions will have income equal to 105 money units if the customer is in a

revolver state and will have income equal to 184 money units if the customer is in a transactor state. However, we know the probabilities of transition only. So, the estimation of the total income is calculated as a weighted sum of the predicted income from different states.

Table 8.20 An example of impact of different scenarios on the total income

		Case			
		1	2	3	4
<i>Product parameters</i>					
1	Credit Limit – GIVEN	10000	10000	10000	10000
2	Interest Rate – GIVEN	36%	36%	36%	36%
<i>Transition Probabilities</i>					
3	Inactive - PREDICTED	0.1	0.1	0.1	0.1
4	Transactor - PREDICTED	0.2	0.2	0.6	0.6
5	Revolver – PREDICTED	0.6	0.6	0.25	0.25
6	Delinquent – PREDICTED	0.1	0.1	0.05	0.05
<i>Interest Income</i>					
7	Utilization Rate – PREDICTED	0.5	0.1	0.5	0.1
8	<i>Interest Rate Income = (1)*(2)/12 * (7)</i>	150	30	150	30
<i>Transactional income</i>					
<i>For Revolver</i>					
<i>Probability of transaction</i>					
9	Probability of POS transaction	0.8	0.8	0.8	0.8
10	Probability of ATM transaction	0.5	0.5	0.5	0.5
<i>Transactional income amount</i>					
11	POS	100	100	100	100
12	ATM	50	50	50	50
13	Transactional income POS = (9)*(11)	80	80	80	80
14	Transactional income ATM = (10)*(12)	25	25	25	25
15	<i>Total transactional income = (13)+(14)</i>	105	105	105	105
<i>For Transactor</i>					
<i>Probability of transaction</i>					
16	Probability of POS transaction	0.9	0.9	0.9	0.9
17	Probability of ATM transaction	0.2	0.2	0.2	0.2
<i>Transactional income amount</i>					
18	POS	200	200	200	200
19	ATM	20	20	20	20
20	Transactional income POS = (16)*(18)	180	180	180	180
21	Transactional income ATM = (17)*(19)	4	4	4	4
22	<i>Total transactional income = (20) + (21)</i>	184	184	184	184
<i>Weighted Income</i>					
23	<i>Interest Income Weighted = (5)*(8)</i>	90	18	37.5	7.5
24	<i>Total transactional income Weighted Revolver = (5)*(15)</i>	21	21	63	63
25	<i>Total transactional income Weighted Transactor = (4)*(22)</i>	36.8	36.8	110.4	110.4
26	<i>Total income = (23) + (24) + (25)</i>	147.8	75.8	210.9	180.9

The maximum total income according to the analysis of the four scenarios is generated in case 3, where the utilisation rate is high, and the probability of being in the transactor state is also high. It can be explained that the transactional income from a transactor has the highest estimated value in comparison with interest income and transactional income for being in the revolver state. However, for other initial conditions such as high-interest income and low transactional income, the results can be opposite.

8.8.2 Examples of the business implementation of the model for credit card income prediction

The total income model can be used in the credit limit management process at the different stages of the decision-making process. The model used with application characteristics can be applied for loan granting and the initial credit limit set up. The model discussed in the current research uses behavioural characteristics after 6 months on book and more. So, it can be used for on-going credit limit management for opened and existing credit cards. The decision can be made to increase, decrease, or keep the same credit limit.

The credit limit management matrix in Table 8.21 shows various coefficients for the current credit limit. It is not based on the data, used to estimate the models, and is only for the illustration of possible business implementation of the discussed methods logic. These values in the body of the table are proposed percentage increases of credit limit. The percentage increases in credit limit depend on the profitability segment and the probability of default. They are deduced from the model we have developed. The probability of default (PD) is split into 11 segments, and the segment value means the lower boundary of the range of PD.

For example, if a customer for the existing credit card has an estimated probability of default equal to 5% and is allocated to segment 4 with the expected profitability 9%, (s)he can get a proposition from the bank to increase the credit limit by 10% of the current credit limit. If this customer has the same profitability 9%, but higher probability of risk, say equal to 10%, (s)he will not have a proposal for the credit limit increase. In case the same customer is allocated to segment 14 with high profitability equal to 25%, (s)he can get a proposal for a 45% increase of the credit limit.

Table 8.21 An example of the credit limit management matrix based on the profitability and the probability of default (is not related to thesis data sample)

Segment	Utilization	Profitability	PD										
			0.0%	1.0%	2.0%	3.0%	4.0%	5.0%	6.0%	7.0%	8.0%	9.0%	10.0%
1	10.2%	3%	5%	5%	0%	0%	0%	0%	0%	-5%	-5%	-5%	-5%
2	16.3%	5%	10%	5%	5%	5%	5%	0%	0%	0%	0%	-5%	-5%
3	18.0%	7%	10%	10%	10%	10%	5%	5%	5%	5%	0%	0%	0%
4	19.5%	9%	15%	15%	10%	10%	10%	10%	5%	5%	5%	5%	0%
5	39.5%	11%	20%	20%	15%	15%	15%	10%	10%	10%	10%	5%	5%
6	36.3%	12%	20%	20%	20%	15%	15%	15%	10%	10%	10%	10%	10%
7	85.0%	13%	25%	20%	20%	20%	15%	15%	15%	10%	10%	10%	10%
8	78.4%	14%	25%	25%	20%	20%	20%	20%	15%	15%	15%	10%	10%
9	60.3%	16%	30%	30%	30%	25%	25%	25%	20%	20%	20%	15%	15%
10	35.3%	18%	35%	35%	35%	30%	30%	25%	25%	25%	20%	20%	20%
11	78.7%	19%	40%	40%	35%	35%	30%	30%	30%	25%	25%	25%	20%
12	77.6%	21%	45%	45%	40%	40%	35%	35%	35%	30%	30%	25%	25%
13	91.9%	24%	55%	50%	50%	45%	45%	45%	40%	40%	35%	35%	30%
14	93.3%	25%	60%	55%	55%	50%	50%	45%	45%	45%	40%	40%	35%

Figure in the body of the table refers to the percentage increase in credit limit

The value for the increase of the credit limit can be computed with use of an optimisation task or with an expert approach. However, the computation of credit limit is out of scope of this thesis.

In case of the credit limit management table uses PD estimation only and does not use the expected profitability or fix the profitability and vary PD only, the rational decision will be to decrease the credit limit for higher values of risk. However, in this case we can lose the high profitability segments, which can generate high income per high credit limit and cover the potentially big expected loss.

For each profitability segment, in addition to its value, we give the utilisation rate value. The utilization rate shows that the credit line profitability does not depend on the utilization rate pro rata: for monotonic profitability, the utilization rate has no monotonic trend. It is possible to select the segment of accounts with high profitability, but low utilization rate. Consequently, the profit from this segment will be high, but the expected loss for the same probability of default is lower than for close segments because of a low utilization rate. It is recommended to build strategies in the card business with the risk – revenue principle to maximize the profitability.

Areas of application are limit management: segmentation by revolver/transactor for risk limitation and usage motivation, pricing: not only risk-based, but use motivation, and marketing: differentiate target groups.

8.9 Conclusion

This model is a further development of total income (or revenue) prediction models beyond the works of Andreeva et al. (2007), Finlay (2010), Ma et al. (2010). We use separate income drivers to predict income from different behavioural types of a credit cardholder, and then estimate profit as the difference between income and loss.

Total income is calculated using three aggregated models and one direct estimation, and a comparative analysis of fitting accuracy has revealed the best models for different prediction horizons. We have also estimated expected loss, and have calculated the implied total profit and profitability.

In section 8.5 we saw that the single model of total income prediction for several months has higher fitting accuracy than the sum of monthly results and the sum of state incomes weighted by the probability of transition. However, for monthly income prediction, the sum of state incomes weighted by the probability of transition has shown the best results from the aggregated models. Thus, estimated profit for separate months $t+1, t+2, \dots, t+n$, the weighted-by-state-transition probabilities approach is the most accurate.

In Chapters 4 and 7 we apply regression models to the prediction of the utilisation rate and transactional income for any initial state at time t , and the current state is used as one of the predictors of the models. The utilisation rate from Chapter 4 is used in Chapter 8 for the computation of the interest income. However, the high fitting accuracy of these results are valid for the entire portfolio only, which includes inactive, transactor, revolver, repaid, and delinquent states. Despite of good predictive accuracy for model, which estimate income for all states, the fitting accuracy for particular states may have poor predictive accuracy. For example, in this research, the revolver and delinquent state accounts have R-squared and RMSE values close to the values given for the entire portfolio. However, for the inactive and transactor states, R-squared is close to zero, and RMSE values are high. This means that inactive and transactor state accounts are accompanied by behavioural patterns, or signs and weights of estimates of income drivers, unlike revolvers. As the revolver estimates are significant in the portfolio and account for approximately 80% of portfolio cases, the single model for all states (using the current state as a factor) provides high fitting accuracy general

results for the entire portfolio but can produce low fitting accuracy results for separate segments. The high concentration of cases in one state dilutes the impact of other states in the total result, revealing model weaknesses when we approach a deeper level of detail or segments. Thus, the next step needed to refine the model is an application of the individual income prediction model for each state. In this model, we used behavioural characteristics for each state in regression analysis, but not in the individual models for each state.

The goodness-of-fit of the total income aggregated models are rather high, because the coefficients of determination values are close to 0.8 for period prediction, and the range is from 0.79 to 0.5 for monthly predictions. In the literature the goodness-of-fit of the models has been primarily related to the balance and debt prediction, with R-squared averaging approximately 0.3. We have not found the coefficient of determination for prediction of income amount for credit cards in the literature. The profitability prediction measures, which are used in some related papers, are error rates (Barrios et al., 2014), classification errors for observed and predicted values, and monetary expression of results (Ma et al., 2010; Andreeva et al., 2007). The total income goodness-of-fit obtained in the current research—R-squared values ranging from 0.5 to 0.8—meets or exceeds many existing related to credit card and debt prediction models.

For comparison with the main complex aggregated models, we have built a direct OLS model for prediction of total income amount. The coefficients of determination, MAE, and RMSE, have produced better fitting accuracy for almost all prediction periods and samples than have the aggregated models. Only the prediction for the Sum of State Incomes Weighted by the States Transition Probabilities model during Month 1 has produced a higher R-squared value—0.7930, for the out-of-time sample—versus 0.7859, for the direct model.

Generally, it seems that a simple linear regression model is superior to multistage models, which aggregate the results of income predictions from different sources. The direct total income model provides a more accurate fitting, because in one step estimation, it contains only one error. In the case of aggregation of multiple models, the errors from different stages are correlated.

However, estimation of total income as an aggregation of income components provides a deep analysis of total income structure and the interactions between predictors, which can have a different impact on the various sources of income, and on the outcomes. Detailed analysis of the total income components facilitates more efficient management of credit card portfolios.

Ma et al.(2010) and Andreeva et al.(2007) use only application and related to purchase characteristics for the prediction of profitability. We use different types of covariates: behavioural, application, and macroeconomic. Some of the behavioural characteristics, which we use, are transactional characteristics and reflect the behavioural type of a credit card holder. This is the *first* application of credit card transactional characteristics for the prediction of total income, which is generated by a credit card.

Ma et al.(2010) predict the expected profit conditional on the probability of acceptance and use four sources of profit from scheduled payments, early repayments, recoveries, and insurance premia. We look at the total income from a credit card as at the combination of partial incomes from different sources related to the type and amount of transaction and the outstanding balance. We also use the probability of transition between account states and the probability of transaction, but do not consider some issues, related to the decision-making process, such as the probability of acceptance. Our behavioural model is for existing clients.

Andreeva et al. (2007) investigate relationship between present value of net revenue from a credit card and times to default and to second purchase. We predict total income amount, which is similar to net revenue. We have extended the measures and show that there exists dependence between income amount and i) the probabilities of transition between credit account states, ii) usage type of the credit card, and iii) type of transactions. However, we used fixed time measure such as one month and six months.

So et al. (2014) predict the profitability of credit card and use the score of being revolver and transactor and Good/Bad score. We use revolver and transactor behavioural types as an account state and use a separate model for income amount prediction for each behavioural type of credit card holder. We consider the probability

of transition to the default state for the total income prediction and profitability estimation but predict the probabilities for all possible transition from any state.

Based on the given data sample, we have the following conclusions for the prediction of total credit card income:

- i. For estimation of total income over a period of several months, the best result is achieved using the Simple Sum of Monthly Interest and Transactional Income model (TI).
- ii. For estimation of monthly income, the best result is achieved using the Sum of State Incomes Weighted by the Individual States Transition Probabilities model (TIW2).
- iii. The use of the model with the account level transition probabilities TWI2 has shown better predictive accuracy than the use of transition probabilities from the transition matrix (MDP) TW1.
- iv. Direct estimation of total income demonstrates the best goodness-of-fit but does not facilitate understanding of the origins of income, internal distributions, the behaviour of each source of income.

The results obtained from the aggregated model for total profit estimation suggest it can be useful for academic and business applications, but requires further improvement of predictive accuracy.

9 Chapter Nine. Conclusion

This concluding chapter presents the summary of results from the substantive chapters and describes the contribution of the thesis. In this regard, the thesis mainly responds to the six questions earlier raised in the Introduction, and it corresponds with the results of Chapters 4-8.

The main aims of this project are: i) to contribute to the formulation of a methodology for a total profit prediction model for some period for credit card activities, which combines different sources of income and behavioural types of the cardholder and ii) to empirically test regression methods applied for each sub-model with panel data. The logical schema of a total profit model is given in the Introduction and shows that the research consists of four sub-projects.

9.1 Summary

The empirical investigation is based on a sample of data relating to credit cards, which contain application and behavioural characteristics from an East European bank. The panel data sample spans two and a half years of lending and payment history – a time series of 30 slices. The data sample contains 14,000 accounts and a total of 280,000 observations for all accounts. We selected a 12-month maximum observation window and a 6-month maximum performance window. However, depending on the objectives, the observation and performance windows can be reduced to six and one months respectively.

The modelling methodology consists of five stages: i) account state prediction with conditional probabilities of transition between states; ii) outstanding balance and interest income prediction; iii) non-interest, or transactional, income amount prediction; iv) expected losses estimation; and v) total income and total profit prediction (see Appendix 1). This complex model is designed to provide an efficient methodology for credit cards profit scoring. We believe that the aggregation of the individual models, which explain separate processes of credit card activities, brings higher predictive accuracy to the total profit estimation than the use of a single model. We formulated six questions for the empirical investigation in the introductory chapter. In the process of the investigation, we found responses to all of them, thus constituting the main contribution of the thesis.

The literature review identified the main papers related directly to the topics of this research such as profit scoring, credit cards usage and utilisation, Markov Decision Process use for risk and profit modelling, credit limit and pricing strategies, and the panel data application for the credit scoring tasks. Also, we review sources secondary for this research, but basic for the credit risk and profit modelling, such as credit risk scoring, optimisation problems in credit scoring, credit risk portfolio models, and correlation in credit risk models. It systematises some sources related to the methods of profit prediction in a table with task and method dimensions. As a result, the literature review discloses the gap in the profit scoring literature and areas related to the credit cards' income prediction.

In Chapter 3 we described the data set, which is used for the empirical analyse, and introduced the characteristics, which are used as covariates and outcomes in the proposed models.

The total income from a credit card consists of two components: interest income and transactional income. The first modelling chapter (Chapter 4) discusses the key element of the interest income prediction – the credit limit utilisation rate. The interest income generally depends on the credit card average outstanding balance. However, we have shown empirically that once there is a credit limit increase, the outstanding balance also increases in several months in the same proportion as the credit limit. So, the utilisation rate is a more stable indicator than the outstanding balance, because it depends on customers' behaviour, not on the controls of the bank.

We applied five a one stage methods and a two-stage methods. The one stage methods were selected according to use for LGD modelling (Bellotti and Crook, 2009a; Yao et al., 2014): OLS, fractional regression (quasi-likelihood), beta-regression (non-linear), beta-transformation + OLS/GLM, and weighted logistic regression with data binary transformation (Berkel, A. & Siddiqi, N.,2012).

We test two types of models: one-stage direct model and two-stage model with the first step as the estimation of the probability that the utilisation rate is equal to bounded values of zero and one, and the second step as the direct estimation of the utilisation rate between the bounds. Fractional regression and weighted logistic regression with a binary transformation show the best fitting accuracy. The beta transformation has the

most similar distribution shape to the observed but has the worst validation results. Two-stage models show slightly better results for all five approaches than the one-stage model. The coefficient of determination for the test sample has values from 0.8 for one-month prediction to 0.55 for six-month horizon prediction. The models for the estimation of the probabilities of the utilisation rate bound values 0 and 1 have high predictive accuracy results with Gini of 0.72 – 0.74 and KS 0.60 – 0.62.

The model for the segment of Month on Book 6 or more (MOB 6+) Changed Limit is slightly stronger than a MOB 6+ segment wit No Limit Change because of additional parameters related to the credit limit changes. Models for MOB less than 6 show weak predictive power because of short behavioural histories. Revolvers and transactors can have different utilisation rates, and more accurate estimation of the utilisation rate requires segmentation of the models by the behavioural pattern.

The credit card behavioural types and the methods for probabilities of transition between credit card states are discussed in Chapters 5 and 6. Chapter 5 gives the concept of states, the modelling methodology and regression analysis description. Chapter 6 contains the comparative analysis of the results of estimation for the account level transition models and the pool level transition matrices.

We discuss two possible ways to predict the transition probabilities: i) a pool level approach with the use of transition matrices and Markov Chains, and ii) an account level approach with the use of logistic regressions. Our data does not satisfy the criteria of Markovity and stationarity. Thus we use three account level approaches: i) conditional binary logistic regression; ii) multinomial logistic regression; and iii) ordinal logistic regression for the prediction of an account transition for 1, 2, and more months. The models show satisfactory predictive power with some variations from strong to weak models.

Multinomial regression shows higher predictive accuracy than ordered logistic regression for the majority of state transition models. The predictive power of the conditional binary logistic regression models depends on the order of the stages, and Gini coefficient values can vary from 0.2 to 0.9 for the transitions from the same state. Multinomial regression gives a balanced model, which implies that the probability of transition between different states does not have such significant variation in predictive

accuracy as the conditional binary logistic regression has - Gini from 0.3 to 0.7. However, the order of the states in the model can be useful in the case that we know which segment is more important for us and have higher priority than others.

The account state transition probabilities from the Chapter 6 is used in the aggregated total income model in Chapter 8.

The second type of income is a transactional one, and it can be generated as fees, commissions, penalties, and other payments. In Chapter 7, we try to predict the non-interest, or transactional, income with regressions based on panel data. We apply one-stage direct estimation and two-stage estimation conditional on the probability of the transaction during the performance period. We consider two sources of transactional income: i) interchange fees and foreign exchange fees from transactions via point-of-sale (POS), and ii) ATM fees from cash withdrawals.

We test and compare five approaches to panel data analysis for the transactional income amount prediction: i) pooled OLS and four random-effect methods; ii) Wansbeek and Kapteyn Method; iii) Fuller and Battese Method; iv) Wallace and Hussain Method; and v) Nerlove's Method.

The Wansbeek and Kapteyn variance component method demonstrated the highest prediction accuracy among the tested random-effect methods. However, the fitting accuracy is low – the coefficient of determination is around 0.19 and 0.15 for POS and ATM transactional income models respectively. On the other hand, the pooled data method gives R-squared values equal to 0.29 and 0.28 respectively. Thus, the pooled method has been selected for the aggregation model.

Two-stage models, which estimate the POS/ATM income amount conditional on the probability of the transaction showed more accurate fit than the one-stage model. The estimated models for the probability of transaction, demonstrate significant goodness-of-fit (Gini around 0.6 – 0.7) and confirm that it is relatively easy to predict whether a cardholder will use a credit card for some types of transactions or not.

The final stage of the project in Chapter 7 is an aggregation of the results of the sub-project to a single model for the total income and total profit estimation. The prediction of the income amount from interest rate and transactions at the different account states, and the prediction of the transition probability between states, are independent parts of

the total income estimation. The total income can be calculated in three ways: i) direct total income estimation; ii) the sum of income from different sources: interest income and transactional income; and iii) the sum of income from various sources conditional on the appropriate state weighted by the probability of transition to the state. The last method is split into two options: the transition matrices and the individual transition probabilities.

The total income goodness-of-fit obtained in the current research – R-squared values from 0.5 to 0.8 - is about the same or exceeds many existing models related to credit card and debt amount prediction models.

We have found some empirical goodness-of-fit values for our given data sample for the total income and profit predictions and their components predictions: the interest income, the transactional income, the utilisation rate, and the probability of transition for the given data sample

Previous research on credit card parameter modelling with the application of OLS has produced the following R-squared values. For income-related, but not exactly income estimation, parameters the following R-squared values are determined in the literature: estimated expenditure - 0.097 (Greene, 1992); credit card usage - 0.3919 (linear bivariate correlation as alternative calculation of R-squared) (Banasik and Crook, 2001); outstanding credit card balances - 0.30 (Kim & De Vaney, 2001); card debt - 0.10 (Tan et al., 2011); and available credit limit - 0.38 (Cohen-Cole, 2001). For loss given default prediction the following R-squared values can be found: LGD - 0.443 (Qi & Zhao, 2011), LGD – 0.8 – 0.4 (Yao et al., 2014), and for EAD modelling: LEQ - 0.0648 and 0.3697 (Qi, 2009).

We give the values of the coefficient of determination for amounts and proportion based on the test sample for the 1st month, for the 6th month, and for a 6 month period respectively. Thus, we obtained the following R-squared values: for the utilisation rate: 0.90, 0.58, and 0.79, for interest income: 0.86, 0.58, and 0.84, for transactional income from POS: 0.18, 0.10, and 0.30, for transactional income from ATM: 0.31, 0.22, and 0.25, for total income: 0.79, 0.48, and 0.81. The income prediction for a period in many cases gives more accurate results than the prediction for a certain month.

Andrade and Thomas (2007) gained a KS for a behavioural scorecard between 0.41 and 0.46. So et al. (2014) gained a Gini for the probability of default around 0.52. Barrios et al. (2014) predict the probability of repurchase with a AUROC around 0.7 (or Gini 0.4) and the probability of default with an AUROC around 0.6 (or Gini 0.2).

We show Gini index values for probability of transition to a certain state for a test sample. The probability to stay in the state and to move to the next state for 1 month / 6 months: inactive – $(0.28 - 0.37) / (0.24 - 0.81)$, transactor – $(0.22 - 0.31) / (0.47 - 0.34)$, revolver – $(0.74 - 0.87) / (0.45 - 0.77)$, repaid – $(NA - 0.53) / (0.23 - 0.38)$, delinquent 1-30 – $(0.46 - 0.60) / (0.22 - 0.55)$, delinquent 31-60 – $(0.70 - 0.65) / (0.68 - 0.57)$.

We also gained a Gini values for the probability of the credit card transaction in the next month and next 6 months' model: POS – 0.69 and 0.64, ATM – 0.59 and 0.69 respectively.

This Coefficient of determination and Gini index values can be used as possible benchmarks for further research into credit cards income and states transitions modelling.

In the scope of this work, we can make the following conclusions for the credit card total income prediction. First, for the estimation of the total income for periods of several months, the best result is given with the simple sum of monthly interest and transactional income model. Second, for the estimation of monthly income, the best result is given with the Sum of State Incomes Weighted by the Individual States from the Transition Probabilities model. The use of the account level transition probabilities shows better fitting accuracy than the use of transition probabilities from a transition matrix (Markov Chain). The direct estimation of the total income shows the best goodness-of-fit, but does not indicate the origins of income and internal distributions and behaviour of each source of income.

9.2 Contributions

9.2.1 Academic contribution

We have created an approach for the prediction of total income and partial incomes from various sources for credit card accounts, considering the existing gaps in the

literature. We have empirically tested the fitting accuracy of regression methods for the prediction of the credit limit utilisation rate, the multi-target probability of transition between account states, and the transactional income amount with a credit cards panel data sample. Moreover, we have tested empirically the goodness-of-fit of the credit card total income and profit prediction aggregated model with use of a panel data sample.

This thesis has made some relevant contributions to knowledge, about income modelling and income scoring. We identified the gaps in the academic literature on credit cards profit prediction and tried to fill them. The published papers are mainly dedicated to the probabilities of card usage and repurchase, debt amount prediction, the transition between risk states, and optimisation tasks based on the risk-revenue approach. However, very few limited studies on the prediction of the total income from the credit card activities.

Firstly, we found a lack of empirical investigations in the credit limit utilisation rate modelling and tried to fill this gap. To do this, we applied and empirically tested regression methods (which are widely used for risk parameters estimation such as LGD but never used for credit card income prediction) to the credit limit utilisation rate predictive models. The credit limit utilisation rate is used for the prediction of the interest income via the product of the credit limit, expected utilisation rate, and interest rate (see Chapter 8).

The usage of credit cards is a topic discussed in many papers (for example, Crook et al., 1992; Banasik and Crook, 2001; Hand and Till, 2003) along with credit cards outstanding balance (Kim and DeVaney, 2001; Tan et al., 2011; Leow and Crook, 2016). However, there is a lack of research on the prediction of the *credit limit utilisation rate*. We implemented some methods already used for proportions prediction such as Loss Given Default (Yao et al., 2014; Arsova et al., 2011) and applied them to *the utilisation rate* (for example, Agarwal et al., 2006). Also we used the credit card usage approaches (Crook et al., 1992; Banasik and Crook, 2001) as the probability of full use or no use of a credit card in a two-stage model. Given that the utilisation rate has outcome value bounded between 0 and 1, specific methods should be applied, such as *beta-regression*, *fractional regression*, or *weighted logistic*

regression with binary transformation (see Chapter 4). Linear regression and other unbounded outcome methods can be applied for utilisation rate prediction, but they often give results outside of the outcome range bounded between zero and one. We also tested the Arellano and Bond (1991) method for estimation of the utilisation rate with lagged endogenous variables. However, the inclusion of lagged values of endogenous variables led to a decrease in predictive accuracy of the regression model.

Secondly, we tried to fill a lack of empirical evidence of the prediction of the transition probabilities between credit card account states at the account level, especially, for a full set of income source-based states. We compared the performance and estimates of *multinomial logistic regression, ordinal regression, and multistate conditional binary logistic regression* for credit card data sample.

Predictive models for risk and profit parameters can be built for a credit card portfolio at the pooled level with, for example, a Markov Chain (So and Thomas, 2011). However, significant differences between credit card usage types can decrease the predictive accuracy of such models, because the different forms of credit card usage have individual behavioural drivers for risk, utilisation, purchases, and profit (So et al., 2014; Tan and Yen, 2010). We tested empirical transition matrices for markovity (according to Thomas et al., 2004) and found that our data sample did not satisfy the requirements of markovity. So the state transition has a high order and we use account level models with behavioural covariates to predict the transition probabilities. We tested *multitarget conditional logistic regression* models for the probability of transition between states. We have found only Volker (1982) used *multinomial logistic regression* for modelling of bankcards utilisation at the account level, and So and Thomas (2014) use *multinomial logistic regression* for the prediction of the transition between inactive, closed, and active account segments, and *cumulative (or ordinal) logistic regression* for the prediction the probability of transition between credit score bands. Kim, Y. and Sohn, S.Y. (2008) estimated of transition probabilities of credit ratings with using a *random effect multinomial regression* model. So, we generally filled the gap in usage of the multinomial and ordinal logistic regression, and multistate models with binary logistic regression for the prediction of credit cards states transition probabilities.

Thirdly, we tried to fill the gap of the application of specific methods to panel data in the credit card income modelling. Baltagi et al. (2002) tested with a Monte-Carlo simulated data sample different random-effect variance component methods for panel data designed by Fuller and Battese (1974), Wansbeek and Kapteyn (1989), Wallace and Hussain (1969). However, we *first* applied these random-effect methods for credit cards panel data for the transactional income amount prediction.

Fourthly, we have not found a comprehensive approach in the academic literature, only partial models, to the prediction of the total income and profit estimation from a set of credit card activities such as transactions and interest rate payments, at the account level. So, we tried to fill this gap in the literature and proposed a comprehensive modelling approach for the total income prediction, which accumulates several individual models for various sources of income and explores the income in the dimension of credit card behavioural types.

Some papers model parts of the general process of credit card income generation. For example, Till and Hand (2003) and Leow and Crook (2014) use delinquency states, So and Thomas (2011) investigate transition to special states such as Closed, Bad1, Bad2, 3 + cycle, Inactive, and Score1, and moving to behavioural score bands from Score 1 to Score 10. We proposed a set of behavioural states for the more accurate credit card profit prediction, which contains only profit related states: merged risk-level states and we introduced transactor and revolver repaid states.

Credit card states are traditionally defined by the level of delinquency and score band, or the level of risk (for example, So and Thomas, 2011; Leow and Crook, 2014). Kallberg and Saunders (1983) split the current states into sub-states by the opening balance and used a ‘Paid-up’ state when the account has no outstanding balance. In addition to the credit card segments, traditionally used in the papers, such as transactors and revolver (see Bertaut et al., 2008; So and Thomas, 2010; Tan and Yen, 2011; So et al, 2014), we proposed to use *transactor and revolver* segments as income related *states* and include a *new state - Revolver Repaid* as an intermediary state from revolver and delinquent states. The revolver repaid state is used for the identification of the transition of a revolver account to an inactive state because the customer who fully repays the debt amount at the end of month formally cannot be allocated either

to a transactor, or an inactive, or a revolver state. Such an account generates income in the repayment months, but according to definitions should be related to inactive or transactor accounts because of the no outstanding balance at the end of the month. The implementation of this new state: i) might increase the predictive accuracy of the transition probability to the transactor and revolver states; and ii) allocates the individual state for the prediction of full debt amount repayment, and as a result, gives a background for the attrition and churn scoring for revolver clients.

9.2.2 Practical impact

Firstly, we identified the most significant explanatory variables (covariates) for the submodels relating to credit cards income prediction, such as the credit limit utilisation rate, the transition probability between states, the POS/ATM transaction income amount, and the interest income amount.

Some papers give lists of application characteristics for risk and profit prediction with significance levels and standardised coefficients, for example, Greene (1992), Crook et al. (1992), Banasik and Crook (2001), Ma et al., (2010), Barrios et al. (2014). Thomas (2000), Belotti and Crook (2009), Malik and Thomas (2010) use macroeconomic variables for default prediction, such as CPI (specifically the rate of inflation), interest rates, and GDP growth. However, fewer papers describe customer related behavioural covariates. Crook and Leow (2014) use behavioural and loan characteristics lagged 3 months such as payment amount, the proportion of credit drawn, an indicator for improvement in the state from 3 months previous, and credit limit. Nie et al. (2010) use behavioural characteristics such as Trade times via ATM and via POS for panel data clustering. Ju et al. (2015) described a behavioural credit scoring model with time-dependent covariates for a stress test. We used similar application and macroeconomic covariates for the income and transition probabilities prediction in both types of models. However, we have significantly expanded the list of behavioural variables. Firstly, we have tested original account characteristics as covariates in predictive models such as average transaction amount, number of transactions, average outstanding balance for the period. Secondly, we created and tested different combinations of characteristics, which are derived from original variables such as average purchase amount to average outstanding balance for three

months, the sum of debit transactions to the sum of credit transactions, maximum purchase transaction to credit limit. These behavioural characteristics have shown high significance and predictive power. However, they are correlated with original characteristics and can overlap in the case of calculations for several months.

Thomas and Malik (2012) predict the transitions between ratings (behavioural score bands) and compared transition matrices for different segments (by age). We estimated the transition between credit card states and compared transition matrices for different prediction periods (see Chapter 6, section 6.4). Thomas and Malik (2012) also investigated the month on book (MOB) effect. We use a month on the book as a covariate for the probability of transition, the probability of POS/ATM transaction, the utilisation rate, and the income amount prediction. For the given data sample the MOB covariate is significant for all types of models and generally has a negative slope for income prediction. Thus, customers substantially generate income during the first year of credit card usage, and then income is slightly reduced.

These findings can be used in financial industry modelling as a benchmark and recommendations for covariate selection and features engineering.

Secondly, the general practice of credit profit modelling is modelling of the risk component as an expected loss and income component separately. The modelling of the credit card income component is usually performed as a single model at the account level and partial transactional (separately for interchange fees and ATM cash withdrawals) and interest income models at the portfolio level. We propose:

- i) to consider both risk and income parts in a single model, and
- ii) to split the prediction of the transactional income from different sources and interest income into account level models for individual income prediction.

The empirical results from Chapter 8 have shown that the aggregation of the individual models for account segments instead of single generalised models for an entire portfolio, gives a similar level of predictive accuracy. However, the parts of the aggregated model give accurate predictions for total income segments and explain the origins of the total income. The methodology of the proposed modelling approach can be applied for the development of business strategies such as credit limit management,

customer segmentation by profitability and behavioural profiles. The thesis proposes an academic contribution to empirical investigations in the profit scoring and the account level credit card behavioural modelling, which can contribute to modelling by the financial industry.

Thirdly, we perform an empirical investigation with panel data to find to following. 1) Fractional regression (quasi-likelihood) and weighted logistic regression (Arsova et al., 2011; Siddiqi, 2012; Yao et al., 2014) with binary transformation give the most reliable result for the *credit limit utilisation rate prediction*. 2) For the prediction of *multistate's transition probability* at the account level (for example, Leow and Crook, 2014) the best results with predictive accuracy among all states were given by multinomial regression (Volker, 1982), and by a decision tree with conditional binary logistic regressions for some selected segments. 3) The total income amount prediction with a single linear regression has shown the higher values of the coefficient determination (R-square around 0.82) in comparison with indirect estimation, but indirect estimation such as the sum of incomes in particular states weighted by the transition probabilities has demonstrated better fitting accuracy for monthly income prediction than the direct model. 4) Pooled regression gives the highest predictive accuracy of all methods, and the Wansbeek and Kapteyn (1989) variance component method showed the highest fitting accuracy for random-effect methods from Baltagi et al. (2002).

Finally, the proposed aggregated model for the total income prediction and its parts can be applied in the financial industry. These models are primarily used for account level strategies to propose individual terms and conditions for each customer, apply an individual decision making strategy, and collection stages. The proposed model for income prediction can be applied in the following tasks and strategies.

The final total profit can be used as a dimension in the risk-profit matrix for credit limit strategy to set up the credit limit to optimise the profit from the credit card, and not only to minimise the expected losses. Also high-risk, but high-profit, cardholders can get credit product terms, which can be declined for high-risk clients. Examples for these terms include acceptance for the credit limit increase procedure, special rates and

fees for spending transactions in retail partner networks, and soft collection procedure for early delinquency stages as 1-5 days past due (so-called loyalty for ‘lazy payers’).

The final income can be used for client segmentation for marketing purposes to propose different bank products depending on the expected profit from the customer or to set up various loyalty programs such as, for example, cash back for some retail networks for high expected profit clients.

The credit limit utilisation rate model can be used both: i) for risk management to compute the Exposure at Default, or the outstanding balance at the default point, both for Expected Loss estimation, and ii) for client segmentation by the level of credit limit usage for marketing purposes, for example, to motivate customers with low usage of the credit limit to spend more.

The transition probabilities between credit card account states can be used for client segmentation by behaviour type and for developing the strategies of moving clients to states, which are more profitable for our business strategy. For example, a bank can push inactive clients with moderate probabilities of transition to the revolver state to make a transaction. A bank can propose extended grace period for free use of credit line, propose cash back for some types of transactions and for purchases from particular retailers etc. On the other hand, a bank can do not spend a budget for promotion and communication with customers with high probability of activation, and also do not spend a budget for actions with a customer with a low probability of activation.

The transactional income from POS and ATM transactions can be used for the motivation of a customer for certain types of spending such as to propose a cash-back for the usage of a credit card for POS transactions and to decrease cash withdrawals fees for ATM. However, there can be an opposite action – the motivation not to spend money on some type of transactions.

The proposed income model can be used for budgeting purposes at the portfolio level for different scenarios of macroeconomic indicators and various profiles of the cardholders. For example, we can simulate how various behaviour groups of cardholders – revolvers, transactors, delinquent – will react to changes in the macroeconomy in the sense of the income which they bring to the bank. Also, we can

change the probabilities of moving between states in the transition matrices and simulate the dynamics of the portfolio for some period to predict total income in the future according to hypothetic scenarios.

The actual use of income models in credit card segments are generally hidden in internal bank procedures and are not made public, except in some white papers and advertisement to consultancy documents. The procedures and models used for income and profitability modelling are usually for internal use only, so the author can mention which methods and approaches are used, what are the advantages and weak points, but is not able to give a reference to real models, which were applied in industry practice.

9.3 Limitations and further research

9.3.1 Limitations

This research has some practical limitations caused by the data sample origin and modelling techniques.

In this research, we applied many simplifications and assumptions.

We used linear models only as OLS and logistic regression and did not consider non-linear relationships between covariates and outcomes. Other models like beta and fractional regressions, multinomial logistic regression are also linear because of use of the linear form of the equation of independent variables. We did not use machine learning methods for simplification of the computation and technical part because our main goal was to test a significant number of models for various segments and targets and to aggregate this variety of partial models, which explain particular types of customer behaviour, into a single integral estimator of the total income amount.

We assumed that the customer behaviour and outcome values are stable and have stable dependencies over two and a half years. We did not consider possible changes in credit policy and decision-making process, which can impact on the customer behaviour, except changes in the credit limits.

For interest income, we simplified the daily scheme of interest accrual to the monthly accrual for average outstanding balance. However, the sum of real daily accruals for

interest can differ from the interest amounts, which are used in our model. We consider only two sources of non-interest income: interchange fees and ATM cash withdrawals. However, we did not take into account the income from the penalty for delinquent customers and other possible sources of transactional income.

We investigate in detail the income sources and income drivers, and the transition probabilities consider the credit risk component as the Probability of Default and Expected Loss via the transition probability to the default state. However, to be able to say that this research is dedicated to profit scoring it must contain all of the components, in a profit computation such as operation cost, cost of finding, cost of recovery and so on. Because we consider risk costs only, the research covers profit, which is defined as the total income minus risk cost, or Expected Loss, only, but not the full definition of profit as an economic category.

We also assumed Loss Given Default as one, or no recoveries after default, and use this pessimistic assumption for the Expected Loss prediction.

We concentrated on the modelling but we have not developed the business use and strategies development that may follow from the models. We have only highlighted some possible ways of the usage of the proposed models. This can be a topic of the further research.

The research has limitations for the conclusion.

The empirical conclusions are made on a relatively short time series of 30 months. The two-and-a-half-year cycle may not be enough for an in-depth study of the relationships between macroeconomic variables and our variables of interest. The short data sample may make the out-of-time validation problematic due to a possible lack of historical data, especially, for some states of the account such as delinquent 30-59 and transactors. Because of the short time series in the panel data, methods like the random-effect variance component may not give convincing results. The six-month prediction period was also selected, in particular, due to the short time series for panel data investigation. A longer time series data sample would enable a more detailed investigation of the impact of macroeconomic factors on the target variable and get more convincing random-effect estimations.

Changes in credit policy and the credit limit management have an impact on the loan parameters and outcomes distributions. So, changes in customer behaviour can be caused not by a reason specific to the account, but have a systematic character common to all card lending processes and clients. These changes can be of two types: internal, created by a lender, and external, or systematic, caused by the macroeconomic changes. Changes in credit policy, in the collection process, and marketing actions should be logged for a more in-depth analysis of the impact of the time component on the correlations between predictors and target variables such as credit limit changes impact on the limit related parameters.

The credit card behavioural data is limited in our dataset. We only have aggregates of the monthly activities of accounts. However, valuable information may also be contained inside of monthly aggregates and disclosed in single transactions. The description of the transactions such as transaction type, place of transaction, time, merchant, goods can give additional knowledge about the client and increase the predictive accuracy of the models. This research does not use credit bureau data. So, the predictive accuracy could be increased significantly, especially, for the application (no behaviour history) models if bureau data was incorporated.

The empirical conclusions about significant explanatory characteristics and target distributions were made on high-quality data from one of the East European retail banks and might be different for other data qualities and financial institutions in the UK, Western Europe, USA, and elsewhere in the World.

9.3.2 *Further research*

This research discloses the weakness of linear and other traditional regression methods for real-life objects prediction, for instance, income from card actions. Many relationships between predictors and dependent characteristics are non-linear. The problem can be partially solved with the use of features binning and the transformation of predictors to dummy variables. Another way is the usage of high degree polynomials for a description of dependencies between factors and outcome.

One of the alternative directions for the investigation in the area of credit cards profit scoring is deep drilling in customer behaviour types, as they are significantly more varied than implemented in this research. For example, a revolver type of behaviour

can be expanded according to the frequency of purchases and payments, for example, to an active revolver, a non-active revolver, and a cash-user.

In our opinion, further investigations can be conducted in the area of application of machine learning methods (for example, Yao et al., 2015) for credit card income prediction and profit scoring tasks. They can replace the traditional regression methods as more efficient methods of prediction. For example, the segmentation and clustering methods as k-means clustering or distribution-based clustering can be used as the statistical approach to the account segments (states) search and definition instead of the used expert approach. Decision trees like Chi-square automatic interaction detection (CHAID) or Classification and Regression Tree (CART) can be applied for selection of the most predictive (powerful) covariates for regression models. Decision trees also can be used as a separate predictive model and a method for finding the segments. Non-linear relationships between covariates and dependent variables can be investigated with advanced methods of machine learning. In this research, we used a multistate tree with conditional binary logistic regression for the prediction of the probabilities of transition between more than two states. However, this problem can be solved with a Neural Network, Naive Bayes classifier, or Support-Vector Machine (SVM) algorithms. A separate class of machine learning algorithms, which demonstrate good predictive results both for classification and value prediction tasks, is ensemble methods such as boosting, random forest, reinforcement learning (Lessmann et al., 2015). Despite the high predictive accuracy of machine learning methods, they are more complicated for technical implementation and support than traditional regression methods. On the other hand, the growing volumes of data, diversity of characteristics from various sources (so-called Big Data), and short times for model development and implementation require an application of machine learning and artificial intelligence methods for income modelling and profit scoring.

Reference

- Andrade, F.W.M. de., and Thomas, L.C. (2007). Structural models in consumer credit. *European Journal of Operational Research* 183, 1569–1581.
- Andreeva, G., Ansell, J.I., and Crook, J.N. (2005). Modelling the purchase propensity: Analysis of a revolving store card. *Journal of Operational Research Society*, 56 (9), 1041–1050.
- Andreeva, G., Ansell, J.I., and Crook, J.N. (2007). Modelling profitability using survival combination scores. *European Journal of Operational Research*, 183, 1537–1549.
- Agarwal, S., Ambrose, B. W., and Liu, C. (2006). Credit Lines and Credit Utilization. *Journal of Money, Credit, and Banking*, Vol. 38, No. 1.
- Anderson, R. (2007). *The Credit Scoring Toolkit: Theory And Practice For Retail Credit Risk Management*, Decision Automation (Oxford).
- Arsova, A., Haralampieva, M., and Tsvetanova T. (2011). Comparison of regression models for LGD estimation. *Credit Scoring and Credit Control XII 2011, Edinburgh*.
- Artzner, P., and Delbaen, F. (1995). Default risk insurance and incomplete markets. *Mathematical Finance*, 1995, Vol. 5, pp. 187–195.
- Au, W.H., Chan, K.C.C., Mining Fuzzy Association Rules in a Bank-Account Database, *IEEE Transactions on Fuzzy Systems*, 2003, Vol. 11.
- Awh, R.Y. and Waters, D. (1974). Discriminant Analysis of Economic, Demographic, and Attitudinal Characteristics of Bank Charge Card Holders: A Case Study. *Journal of Finance*. Vol. 29, No. 3: 973-80.
- Bade, B., Rosch, D., and Scheule, H. (2011). Default and Recovery Risk Dependencies in a Simple Credit Risk Model. *European Financial Management*, Vol. 17, No. 1, 2011, 120–144.
- Bailey M. (2004). *Consumer credit quality: underwriting, scoring, fraud prevention and collections*, Kingswood, Bristol: White Box Publishing.
- Baltagi, B. H. and Chang, Y. (1994). Incomplete Panels: A Comparative Study of Alternative Estimators for the Unbalanced One-Way Error Component Regression Model. *Journal of Econometrics*, 62(2), 67–89.
- Baltagi, B. H., Song, Seuck H., and Jung, Byoung C. (2002). A Comparative Study of Alternative Estimators for the Unbalanced Two-Way Error Component Regression Model, *Econometrics Journal*, 5, 480–493.
- Banasik, J., Crook, J. N., and Thomas, L. C. (1999). Not if but when borrowers default. *Journal of Operational Research Society*, 50, 1185–1190.

- Banasik, J., Crook, J.N., and Thomas, L.C. (2001). Scoring by usage. *Journal of the Operational Research Society*, 52, 997-1006.
- Basel Committee on Banking Supervision (2004). Basel II: International Convergence of Capital Measurement and Capital Standards: a Revised Framework. June 2004.
- Bellotti, T., and Crook, J.N. (2009a). Loss Given Default models for UK retail credit cards. *CRC working paper*, 09/1.
- Bellotti, T., and Crook, J.N. (2009b). Credit scoring with macroeconomic variables using survival analysis. *Journal of the Operational Research Society*, 60, 1699-1707.
- Bellotti, T., & Crook, J. (2012). Loss given default models incorporating macroeconomic variables for credit cards. *International Journal of Forecasting*, 28(1), 171–182.
- Berkel, A. & Siddiqi, N. (2012). Building Loss Given Default Scorecard Using Weight of Evidence Bins in SAS®Enterprise Miner™. *SAS Institute Inc.* Paper 141-2012.
- Bertaut, C.C., Haliassos, M., and Reiter, M. (2008). Credit Card Debt Puzzles and Debt Revolvers for Self Control. *Review of Finance*, 13: 657–692.
- Bird, E., Hagstrom, P. A. & Wild, R. (1997). Credit cards and the poor. *Institute for Research on Poverty Discussion Paper*, 1148-1197.
- Black, F., and Cox, J.C.(1976). Valuing corporate securities: some effects of bond indenture provisions. *Journal of Finance*, Vol. XXXI, No.2, pp. 351–367.
- Black, F., and Scholes, M. (1973). The pricing of options and corporate liabilities. *Journal of Political Economy*. Vol. 81, pp. 81–98.
- Bluhm, C. (2003). An introduction to credit risk modeling/ Christian Bluhm, Ludger Overbeck, Christoph Wagner. P. cm. – (Chapman & Hall/ CRC financial mathematics series).
- Chakravorti, S. (2003). Theory of Credit Card Networks: A Survey of the Literature. *Review of Network Economics*, Vol.2, Issue 2 – June 2003, 50-68.
- Cheu, See-Peng. and Loke, Yiing-Jia (2010). Credit Cardholder: Convenience User or Credit Revolver? *Malaysian Journal of Economic Studies* 47(1):1-17, 2010.
- Ching, W.-K., Ng, M. K., Wong, K.-K., and Altman, E. (2004). Customer lifetime value: stochastic optimization approach. *Journal of Operational Research Society*, 55:860-868, 2004.
- Crook, J. N., Hamilton, R., and Thomas, L. C. (1992a). Credit Card Holders: Characteristics of Users and Non-Users. *The Service Industries Journal*, Vol. 12, No. 2, pp. 251-262
- Crook, J.N., Hamilton, R., and Thomas, L.C. (1992b). A Comparison of a Credit Scoring Model with a Credit Performance Model. *The Service Industries Journal*, Vol. 12, No. 4 (October 1992), 558-579.

- Crook, J.N, Edelman D.B., and Thomas L.C. (2007). Recent developments in consumer credit risk assessment. *European Journal of Operational Research*, Vol. 183(3), 1447–1465.
- Crook, J.N., and Bellotti, T. (2009). Forecasting and Stress Testing Credit Card Default using Dynamic Models, *Credit Research Centre Working Paper*, 10/01.
- Crook, J.N., and Bellotti, T. (2010). Time varying and dynamic models for default risk in consumer loans. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 173(2), 283-305.
- Crosbie P.(1999). Modeling default risk. *KMV Corporation*, <http://www.kmv.com>.
- Crouhy, M., Galai, D., and Mark, R. (2000). A comparative analysis of current credit risk models. *Journal of Banking & Finance*, 24, 59-117
- Cyert, R.M., Davidson, H.J., & Thompson, G.L. (1962). Estimation of allowance for doubtful accounts by Markov Chains. *Management Science* 8, 287-303
- Duffie, D., and Singleton, K.J. (1999). Modeling term structures of defaultable bonds. *Review of Financial Studies*, 1999, Vol. 12(4), pp. 687–720.
- Duffie, D., and Lando, D.(2001). Term structures of credit spreads with incomplete accounting information. *Econometrica*. Vol. 69, pp. 633–664.
- Desay, V. S., Convay, D. G., Crook, J. N., & Overstreet, G. A. (1997). Credit scoring models in the credit union environment using neural networks and genetic algorithms. *IMA Journal of Mathematics Applied in Business and Industry* 8, 323–346.
- Duca, J.V., and Whitesell, W.C. (1995). Credit Cards and Money Demand: A Cross-sectional Study. *Journal of Money, Credit, and Banking*, Vol. 27, No. 2.
- Durand, D. (1941). Risk elements in consumer instalment financing. *National Bureau of Economic Research*, New York.
- Dulltnann, K., and Trapp, M. (2004). Systematic Risk in Recovery Rates – an Empirical Analysis of US Corporate Credit Exposures. *Discussion Paper, Series 2; Banking and Financial Supervision*, No. 02/2004.
- Dunkelberg, W. C., and Smiley, R.H. (1975). Subsidies in the Use of Revolving Credit, *Journal of Money, Credit and Banking*, Vol. 7, November: 469-90
- Dunkelberg, W.C, and Stafford, F.P. 1971. Debt in the Consumer Portfolio: Evidence from a Panel Study. *The American Economic Review*, Vol. LXI, No. 4, September: 598-613.
- Dunn, L. and Kerr, S. (2002). Consumer Search Behavior in the Changing Credit Card Market. *Journal of Business & Economic Statistics*, 01 July 2008, Vol.26(3), p.345-353

Ethan Cohen-Cole (2011). Credit Card Redlining. *The Review of Economics and Statistics*,

May 2011, 93(2): 700–713.

Farajian, M. Ali & Mohammadi, Sh. (2010). Mining the Banking Customer Behavior Using Clustering and Association Rules Methods. *International Journal of Industrial Engineering & Production Research*. December 2010, Volume 21, Number 4, pp. 239-245

Fiorio, L., Mau, R., Steitz, J. and Welander, T. (2014). New frontiers in credit card segmentation: Tapping unmet consumer needs. *McKinsey on Payments*, Number 19, 2014.

Fulford, S.L. (2015). How Important is Variability in Consumer Credit Limits? *Journal of Monetary Economics* 72: 42–63.

Fulford, S.L. and Schuh, S. (2015). Consumer Revolving Credit and Debt over the Life Cycle and Business Cycle. *Federal Reserve Bank of Boston*, Working Paper, No.15-17.

Fulford, S.L. and Schuh, S. (2017). Credit Card Utilization and Consumption over the Life Cycle and Business Cycle. *Consumer Financial Protection Bureau Office of Research*. Working Paper No. 2017-03

Fuller, W. A. and Battese, G. E. (1974). Estimation of Linear Models with Crossed-Error Structure, *Journal of Econometrics*, 2, 67–78.

Ferrari, S. L. P. and Cribari-Neto, F. (2004). Beta Regression for Modelling Rates and Proportions. *Journal of Applied Statistics*, 31, 799-815

Finlay, S., (2008a). Towards profitability. A utility approach to the credit scoring problem. *Journal of the Operational Research Society*, 59 (7), 921–931.

Finlay, S., (2008b). The Management of Consumer Credit: Theory and Practice. *Palgrave Macmillan, Basingstoke, UK*.

Finlay, S.M. (2009). Are we modelling the right thing? The impact of incorrect problem specification in credit scoring. *Expert Systems with Applications*, 36(5), 9065–9071.

Finlay, S., (2010). Credit scoring for profitability objectives. *European Journal of Operational Research*, 202, 528–537.

Finlay, S (2010). Credit Scoring, Response Modelling and Insurance Rating: A Practical Guide to Forecasting Consumer Behaviour. *Palgrave Macmillan; 1 edition (December 7, 2010)*

Fisher, R. A. (1936). The use of multiple measurements in taxonomic problems. *Annals of Eugenics* 7, 179–188.

Frydman, H., Kallberg, J.G., and Kao, D.-L. (1985). Testing the Adequacy of Markov Chain and Mover-Stayer Models as Representations of Credit Behavior. *Operations Research*, Vol. 33, No. 6, pp.1203-1214.

- Frye, John (2000). Collateral damage. *Risk* (April): 91–94.
- Geisecke K. and Goldberg L. (2004). Forecasting default in the face of uncertainty. *Journal of Derivatives*. Vol. 12. pp. 14–15.
- Gersbach Hans and Alexander Lipponer (2003). Firm Defaults and the Correlation Effect. *European Financial Management*, Vol. 9, No. 3, 2003, 361–377
- Gordy M. (2000). A Comparative Anatomy of Credit Risk Models. *Journal of Banking and Finance*, 24:119–149, 2000.
- Greene, W. H. (2000). Econometric Analysis, Fourth Edition, New York: Macmillan Publishing Company.
- Guido Giese (2005). The impact of PD/LGD correlations on credit risk capital. [www.risk.net.](http://www.risk.net/), April 2005, Risk.
- Gupton G. M., Finger C. C., and Bhatia M.(1997). CreditMetrics, Technical Document. *Morgan Guaranty Trust Co.*, http://www.defaultrisk.com/pp_model_20.htm, 1997.
- Hayashi, F., Sullivan, R. and Weiner , S. E. (2006). A guide to the ATM and Debit Card Industry. *Federal Reserve Bank of Kansas City*.
- Hayashi , F.(2010). Payment Card Interchange Fees and Merchant Service Charges - An International Comparison". *Lydian Payments Journal*, Vol. 1, Issue 3, January 2010.
- Ho, J., Thomas, L.C., Pomroy, T.A., and Scherer, W.T., (2004). Segmentation in Markov chain consumer credit behavioural models, in Readings in Credit Scoring (ed. Thomas L.C., Edelman D.B., Crook J.N.,), Oxford University Press, Oxford, pp 295-308.
- Hsieh, Nan-Chen (2004). An integrated data mining and behavioral scoring model for analyzing bank customers. *Expert Systems with Applications* 27 (2004) 623–633.
- Hsieh, N.-C., Chu, K.-C. (2009). Enhancing Consumer Behavior Analysis by Data Mining Techniques. *International Journal of Information and Management Sciences* 20 (2009), 39-53.
- Jacobs, M.Jr. (2008). An empirical study of exposure at default. *Office of the Comptroller of the Currency Working Paper*.
- Jarrow R.A. and Fan Yu (2001). Counterparty risk and the pricing of defaultable securities, *Journal of Finance*. Vol. 56, 555–576.
- Jones, M.T. (2005). Estimating Markov Transition Matrices Using Proportions Data: An Application to Credit Risk. *IMF Working Paper*, WP/05/219.

- Ju, Y., Jeon, S.Y, and Sohn, S.Y. (2015). Behavioral technology credit scoring model with time-dependent covariates for stress test. *European Journal of Operational Research* 242 (2015) 910-919.
- Kallberg, J.G., and Saunders, A. (1983). Markov Chain Approaches to the Analysis of Payment Behavior of Retail Credit Customers. *Financial Management*, Vol. 12, No. 2 (Summer, 1983), pp. 5-14
- Keeney, R.L., and Oliver, R.M. (2005). Designing win-win financial loan products for consumers and businesses. *Journal of the Operational Research Society*, 56, 1030–1040.
- Kim, H, and DeVaney, Sh. A. (2001). The Determinants Of Outstanding Balances Among Credit Card Revolvers. *Association for Financial Counseling and Planning Education*.
- Kim, Y. & Sohn, S.Y. (2008). Random effects model for credit rating transitions. *European Journal of Operational Research*, 184, 561–573.
- Klein, M. A. (1971). Theory of Banking Firm. *The Journal of Money, Credit and Banking*, vol.3, pp.205-218, 1971.
- Lachenbruch, P. A., and Goldstein, M. (1979). Discriminant Analysis, *Biometrics*, Vol. 35, No. 1, Perspectives in Biometry (Mar., 1979), pp. 69-85.
- Lando, D. and Skødeberg,, T.M. (2002). Analyzing rating transitions and rating drift with continuous observations. *Journal of Banking and Finance*, Vol.26(2), pp.423-444
- Leow, M., Crook, J. (2014). Intensity models and transition probabilities for credit card loan delinquencies. *European Journal of Operational Research*, 236, 685–694
- Leow, M., Crook, J. (2016). A new Mixture model for the estimation of credit card Exposure at Default. *European Journal of Operational Research* 249, 487–497
- Lessmann, S., Baesens, B., Seow, H.-V., Thomas, L.C. (2015). Benchmarking state-of-the-art classification algorithms for credit scoring: An update of research. *European Journal of Operational Research*, 247, pp.124–136.
- Lewis, E., (1992). Introduction to Credit Scoring. *The AthenaPress, San Rafael*.
- Li, H. G., & Hand, D. J. (1997). Direct versus indirect credit scoring classification. In: *Proceedings of Credit Scoring and Credit Control V, Credit Research Centre, University of Edinburgh*.
- Lai, T. L., & Ying, Z. L. (1994). A Missing information principle and M-estimators in regression analysis with censored and truncated data. *Annals of Statistics* 22, 1222–1255.
- Lieli Robert P., White Halbert (2010). The construction of empirical credit scoring rules based on maximization principles. *Journal of Econometrics*, 157 (2010), 110-119.

- Lütkebohmert Eva (2009). Concentration Risk in Credit Portfolios. *Springer-Verlag Berlin Heidelberg*.
- Ma, P., Crook J., and Ansell, J. (2010). Modelling take-up and profitability. *Journal of the Operational Research Society*, 61, 430–442.
- Makowski,,P. (1985) Credit scoring branches out. *Credit World*, 75, 30-37.
- Malik, M. & Thomas, L.C. (2012). Transition matrix models of consumer credit ratings. *International Journal of Forecasting* 28, pp.261–272.
- Meier, S. and Sprenger, C. (2007). Impatience and Credit Behavior: Using Choice Experiments to Explain Borrowing and Defaulting. *Center for Behavioral Economics and Decision Making, Federal Reserve Bank of Boston*, 2007.
- Merton, R.C. (1974). On the pricing of corporate debt: the risk structure of interest rates. *Journal of Finance*, Vol. 29, pp. 449–470.
- Myers, J. H., & Forgy, E. W. (1963). The development of numerical credit evaluation systems. *Journal of American Statistics Association* 58 (September), 799–806.
- Nerlove, M. (1971). Further Evidence on the Estimation of Dynamic Relations from a Time Series of Cross Sections, *Econometrica*, 39, 359–382.
- Narain, B. (1992). Survival analysis and the credit granting decision. In: Thomas, L. C., Crook, J. N., & Edelman, D. B. (Eds.), *Credit scoring and credit control*, Oxford University Press, Oxford, pp. 109–122.
- Niea, G., Chen, Y., Zhang, L. and Guo, Y. (2010). Credit card customer analysis based on panel data clustering. *International Conference on Computational Science, ICCS 2010*. Procedia Computer Science, Vol. 1, pp.2489–2497
- Oliver, R. M. (1993). Effects of calibration and discrimination on profitability scoring. In: *Proceedings of Credit Scoring and Credit Control III*, Credit Research Centre, University of Edinburgh.
- Oliver, R.M. and Wells, E. (2001). Efficient frontier cutoff policies in credit portfolios. *Journal of the Operational Research Society*, 52, 1025-1033.
- Pykhtin, M. (2003). Unexpected Recovery Risk. *Risk*, Vol. 16, 74-78.
- Qi, M. (2009). Exposure at Default of Unsecured Credit Cards. *Office of the Comptroller of the Currency Working Paper*.
- Ramona K.Z.Heck (1987). Differences in Utilisation Behaviour Among Types of Credit Cards. *The Service Industries Journal*, Vol. 7, Issue 1, 1987

- Rosch. D, and H. Scheule (2004). Forecasting Retail Portfolio Credit Risk. *Journal of Risk Finance*, Vol. 5, No. 2, pp, 16-32
- Rosch Daniel, Scheule Harald (2005). A Multifactor Approach for Systematic Default and Recovery Risk. *The Journal of Fixed Income*, Sep.2005.
- Rosenberg, E., & Gleit, A. (1994). Quantitative methods in credit management: a survey. *Operations Research* 42, 589–613.
- Sabato Gabriele and Schmid Markus M. (2008). Estimating conservative loss given default. <http://ssrn.com/abstract=1136762>.
- Sánchez-Barrios, L.J., Andreeva, G. and Ansell, J. (2014). Monetary and relative scorecards to assess profits in consumer revolving credit. *Journal of the Operational Research Society* 65, pp.443–453.
- Sánchez-Barrios, L.J., Andreeva, G. and Ansell, J. (2016). Time-to-profit scorecards for revolving credit. *European Journal of Operational Research* 249, pp.397–406.
- Schonbucher, Rj. Credit Derivatives Pricing Models: Models. Pricing and implementation. New York: John Wiley and Sons. 2003.
- Sealey, C., and Lindley, J.T. (1977). Inputs, Outputs and a Theory of Production and Cost at Depository Financial Institutions, *Journal of Finance*, 32, 1251-1266.
- Seow H-V (2010).Question selection responding to information on customers from heterogeneous populations to select offers that maximize expected profit. *Journal of the Operational Research Society* 61, pp.443 –454.
- Serrano-Cinca, C. and Gutiérrez-Nieto, B. (2016). The use of profit scoring as an alternative to credit scoring systems in peer-to-peer (P2P) lending. *Decision Support Systems* 89, pp.113–122.
- Siddiqi, Naeem (2005). *Credit Risk Scorecards: Developing and Implementing Intelligent Credit Scoring*. Wiley and Sons.
- So, M. C., and Thomas, Lyn C. (2010). Modeling and model validation of the impact of the economy on the credit risk of credit card portfolios. *The Journal of Risk Model Validation* (93–126) Volume 4/Number 4, Winter 2010/11.
- So, M.C. and Thomas, L.C. (2011). Modelling the profitability of credit cards by Markov decision processes. *European Journal of Operational Research* 212, pp. 123–130.
- So, M. C.,Thomas, Lyn C., Seow, H-V., and Mues, C. (2014). Using a transactor/revolver scorecard to make credit and pricing decisions. *Decision Support Systems* 59, pp. 143–151.

- Stepanova, M., Thomas, L.C., 2001. PHAB scores: Proportional hazards analysis behavioural scores. *Journal of Operational Research Society* 52, 1007–1016.
- Stepanova, M., Thomas, L.C., 2002. Survival analysis methods for personal loan data. *Operations Research* 50 (2), 277–289.
- Stewart R.T. (2010). A profit-based scoring system in consumer credit: making acquisition decisions for credit cards. *Journal of the Operational Research Society* (2011) 62, 1719–1725.
- Stoyanov, S. (2009). Application LGD Model Development. A Case Study for a Leading CEE Bank. *Credit Scoring and Credit Control XI Conference, Edinburgh, August, 2009*.
- Shuai, Qing-hong, and Shi, Yu-lu. (2009). Profitable credit card business empirical analysis of factors. Oct. 2009, Volume 8, No.10 (Serial No.76) *Chinese Business Review*, ISSN 1537-1506, USA
- Tasche Dirk (2004). The single risk factor approach to capital charges in case of correlated loss given default rates.
- Till, R.J. and Hand, S.J. (2003). Behavioural models of credit card usage. *Journal of Applied Statistics*, Vol. 30, No. 10, pp.1201–1220
- Thomas, L. C., Edelman, David B., Crook, Jonathan N (2004). *Readings in credit scoring: foundations, developments, and aims*. Oxford : Oxford University Press, 2004.
- Thomas, L. C. (1994). Applications and solution algorithms for dynamic programming. *Bulletin of the IMA* 30, 116–122.
- Thomas Lyn C.(2000). A survey of credit and behavioural scoring: forecasting financial risk of lending to consumers. *International Journal of Forecasting*, 16(2000), 149-172.
- Thomas, L.C., Ho, J., and Scherer, W.T. (2001). Time will tell: Behavioural scoring and the dynamics of consumer credit assessment. *IMA Journal of Management Mathematics*, 12(1), 89-103.
- Thomas, L. C., Oliver, R. W., and Hand, D. J. (2005). A Survey of the Issues in Consumer Credit Modelling Research. *The Journal of the Operational Research Society*, Vol. 56, No. 9, 1006-1015.
- Thomas L.C., (2009). Consumer Credit Models: Pricing, Profit and Portfolios. *Oxford University Press, Oxford* (2009).
- Thomas, L.C., Crook, J.N., and Edelman, D.B. (2017). Credit Scoring and Its Applications, Second Edition. *Philadelphia: Society for Industrial and Applied Mathematics* (2017).
- Wilson, T. C. (1998). Portfolio Credit Risk. *FRBNY economic policy review/ October 1998*.
- Vasicek, O.A., 1987, Probability of Loss on Loan Portfolio, *KMV Corporation*.

- Verbraken, T., Bravo, C., Weber, R., and Baesens, B. (2014). Development and application of consumer credit scoring models using profit-based classification measures. *European Journal of Operational Research* 238, pp.505–513.
- Volker (1982). A note on factors influencing the utilization. *Australian University. Econ Record* September: 281–289.
- Wallace, T., and Hussain, A. (1969). The Use of Error Components Model in Combining Cross Section with Time Series Data, *Econometrica*, 37, 55–72.
- Wansbeek, T., and Kapteyn, Arie (1989), Estimation of the Error-Components Model with Incomplete Panels, *Journal of Econometrics*, 41, 341–361.
- Wilde, T. (1997). CreditRisk+. Wilde, T. of CSFB Credit Suisse Financial Products. *A Credit Risk Management Framework*.
- White, K. J. (1975). Consumer Choice and Use of Bank Credit Cards: A Model and Cross-Section Results. *Journal of Consumer Research*, Vol. 2, June 10-18.
- White, K.J. (1976). The Effect of Bank Credit Cards on the Household Transactions Demand for Money. *Journal of Money, Credit, and Banking*, Vol. 8, February: 51-63.
- Wiginton, J.C. (1980). A Note on the Comparison of Logit and Discriminant Models of Consumer Credit Behavior. *Journal of Financial and Quantitative Analysis*, Vol. 15, No. 3 (Sep., 1980), pp. 757-770.
- Yao, X., Crook, J., and Andreeva, G. (2014). Modeling Loss Given Default in SAS/STAT®. *SAS Forum 2014*. Paper 1593-2014.
- Yao, X., Crook, J., and Andreeva, G. (2015). Support vector regression for loss given default modelling. *European Journal of Operational Research*, Vol.240(2), pp.528-538.

Appendix 1. The polled and random-effect methods for ATM income

Table A3.1. Coefficients estimation for ATM income

Variable	Pooled		WK		WH		NL	
	Estimate	Pr > t	Estimate	Pr > t	Estimate	Pr > t	Estimate	Pr > t
Intercept	-110.263	(0.0906)	104.889	(0.0848)	44.43868	(0.4723)	109.8435	(0.071)
Mob	2.41237	(<.0001)	4.645451	(<.0001)	4.002847	(<.0001)	4.700607	(<.0001)
limit_6	0.000703	(0.0006)	0.000083	(0.7236)	0.000442	(0.0563)	0.000049	(0.8343)
UT0_6	-66.2825	(<.0001)	-40.8486	(<.0001)	-42.5291	(<.0001)	-40.811	(<.0001)
UT0_7	45.38276	(<.0001)	22.1332	(<.0001)	27.6843	(<.0001)	21.70683	(<.0001)
UT0_8	41.34974	(<.0001)	16.82168	(<.0001)	21.77318	(<.0001)	16.46136	(<.0001)
UT0_9	22.88434	(<.0001)	7.568884	(0.059)	10.56334	(0.0117)	7.356928	(0.0656)
UT0_10	-1.47977	(0.7635)	-3.80606	(0.3107)	-3.7455	(0.3398)	-3.79194	(0.3109)
UT0_11	17.3266	(<.0001)	8.414491	(0.0029)	8.848178	(0.0025)	8.41533	(0.0028)
b_UT1to2ln	-0.7358	(0.0539)	-1.07603	(0.0002)	-1.11307	(0.0003)	-1.07058	(0.0002)
b_UT1to6ln	1.88246	(<.0001)	2.261019	(<.0001)	2.380083	(<.0001)	2.24731	(<.0001)
avg_balance_6	0.007904	(<.0001)	0.000277	(0.7122)	0.00147	(0.0598)	0.000191	(0.7988)
avg_balance_7	-0.00416	(0.0001)	-0.00547	(<.0001)	-0.00517	(<.0001)	-0.0055	(<.0001)
avg_balance_8	-0.00214	(0.0353)	-0.00259	(0.0008)	-0.00215	(0.0081)	-0.00264	(0.0007)
avg_balance_9	0.000939	(0.3462)	0.000054	(0.9433)	0.000673	(0.3976)	-5.4E-06	(0.9943)
avg_balance_10	0.001659	(0.0821)	0.00067	(0.3584)	0.001461	(0.0554)	0.000594	(0.414)
avg_balance_11	-0.001	(0.1949)	0.000771	(0.2032)	0.001372	(0.0294)	0.000702	(0.2452)
avg_deb_amt_6	0.007811	(0.0004)	-0.00063	(0.7233)	0.00008	(0.9654)	-0.00065	(0.7115)
avg_deb_amt_7	0.008475	(<.0001)	-0.00267	(0.1251)	-0.00123	(0.496)	-0.00276	(0.1122)
avg_deb_amt_8	0.007099	(0.0007)	-0.00282	(0.0932)	-0.0014	(0.4207)	-0.00291	(0.0822)
avg_deb_amt_9	0.002934	(0.1071)	-0.0016	(0.2712)	-0.00139	(0.3584)	-0.0016	(0.2692)
avg_deb_amt_10	0.000104	(0.9504)	-0.00361	(0.0066)	-0.00384	(0.0055)	-0.00357	(0.007)
avg_deb_amt_11	0.000625	(0.6256)	-0.00143	(0.1576)	-0.00147	(0.1651)	-0.00142	(0.1592)
sum_crd_amt_6	0.013912	(<.0001)	0.008854	(<.0001)	0.009391	(<.0001)	0.008828	(<.0001)
sum_crd_amt_7	0.010262	(<.0001)	0.003996	(<.0001)	0.004672	(<.0001)	0.00396	(<.0001)
sum_crd_amt_8	0.007578	(<.0001)	0.000455	(0.5323)	0.001343	(0.0769)	0.0004	(0.5811)
sum_crd_amt_9	0.0059	(<.0001)	-0.00221	(0.0013)	-0.001	(0.1618)	-0.00229	(0.0008)
sum_crd_amt_10	0.006921	(<.0001)	-0.0016	(0.0051)	-0.00005	(0.9327)	-0.00171	(0.0027)
sum_crd_amt_11	0.005903	(<.0001)	-0.00138	(0.0001)	0.000295	(0.425)	-0.00151	(<.0001)
sum_deb_amt_6	-0.00686	(<.0001)	-0.00262	(0.0532)	-0.00282	(0.0456)	-0.00262	(0.0523)
sum_deb_amt_7	-0.00519	(0.0051)	0.001582	(0.283)	0.001044	(0.4958)	0.001604	(0.2748)
sum_deb_amt_8	-0.00954	(<.0001)	0.001747	(0.2563)	0.000684	(0.6691)	0.001804	(0.2397)
sum_deb_amt_9	-0.0091	(<.0001)	0.000831	(0.5743)	0.00053	(0.7303)	0.000825	(0.5759)
sum_deb_amt_10	-0.00878	(<.0001)	0.002214	(0.0884)	0.001626	(0.2287)	0.002231	(0.0851)
sum_deb_amt_11	-0.00746	(<.0001)	0.00013	(0.9058)	-0.00034	(0.7665)	0.000152	(0.8899)
max_deb_amt_6	0.001264	(0.5153)	-0.00591	(0.0001)	-0.00454	(0.0049)	-0.00602	(0.0001)
max_deb_amt_7	-0.00124	(0.5262)	-0.00658	(<.0001)	-0.0055	(0.0007)	-0.00667	(<.0001)
max_deb_amt_8	0.007604	(0.0001)	-0.00193	(0.2237)	-0.00042	(0.7995)	-0.00204	(0.1963)
max_deb_amt_9	0.008424	(<.0001)	-0.00184	(0.2141)	-0.00103	(0.5028)	-0.00188	(0.2033)
max_deb_amt_10	0.007366	(<.0001)	-0.00306	(0.0196)	-0.00225	(0.0984)	-0.0031	(0.0178)
max_deb_amt_11	0.006827	(<.0001)	-0.00098	(0.4011)	-0.00052	(0.6678)	-0.00099	(0.3922)
min_deb_amt_6	0.000473	(0.7725)	0.004596	(0.0004)	0.003925	(0.0039)	0.004645	(0.0004)
min_deb_amt_7	-0.000668	(0.682)	0.002892	(0.0274)	0.002103	(0.1233)	0.002954	(0.0239)
min_deb_amt_8	-0.00283	(0.0678)	0.002004	(0.102)	0.001198	(0.3475)	0.002063	(0.0913)
min_deb_amt_9	-0.00644	(<.0001)	0.001051	(0.3638)	-0.00017	(0.8859)	0.001137	(0.3244)
min_deb_amt_10	-0.00372	(0.0033)	0.001353	(0.1759)	0.000765	(0.4623)	0.001388	(0.1639)
min_deb_amt_11	-0.00383	(<.0001)	0.001214	(0.0772)	0.000797	(0.2655)	0.001232	(0.0722)
b_AvgOB1_to_MaxOB1_ln	16.83031	(<.0001)	12.83046	(<.0001)	14.04149	(<.0001)	12.73268	(<.0001)
b_AvgOB2_to_MaxOB2_ln	14.15887	(<.0001)	10.24508	(<.0001)	11.36293	(<.0001)	10.15279	(<.0001)
b_AvgOB3_to_MaxOB3_ln	16.41312	(<.0001)	9.892972	(<.0001)	11.27832	(<.0001)	9.785708	(<.0001)
b_TRmax_deb1_To_Limit_ln	-19.8859	(<.0001)	-29.7165	(<.0001)	-30.8465	(<.0001)	-29.5097	(<.0001)
b_TRmax_deb2_To_Limit_ln	-1.54638	(0.7578)	-15.2945	(0.0002)	-15.6548	(0.0002)	-15.1454	(0.0002)
b_TRmax_deb3_To_Limit_ln	-12.8431	(0.0086)	-17.3832	(<.0001)	-19.0808	(<.0001)	-17.1455	(<.0001)
b_TRavg_deb1_to_avgOB1_ln	-8.19918	(<.0001)	-0.89081	(0.1407)	-2.03589	(0.0012)	-0.81251	(0.1779)
b_TRavg_deb2_to_avgOB2_ln	-4.11304	(<.0001)	1.932955	(0.0013)	0.987076	(0.1146)	1.996418	(0.0009)
b_TRavg_deb3_to_avgOB3_ln	-3.62507	(<.0001)	2.014597	(0.0005)	1.054212	(0.0789)	2.081499	(0.0003)
b_TRsum_deb1_to_TRsum_crd1_ln	8.268213	(<.0001)	1.462182	(0.0061)	2.78908	(<.0001)	1.36327	(0.0104)
b_TRsum_deb2_to_TRsum_crd2_ln	7.588811	(<.0001)	1.07983	(0.0367)	2.318147	(<.0001)	0.989562	(0.0549)
b_TRsum_deb3_to_TRsum_crd3_ln	5.239561	(<.0001)	-0.15589	(0.7402)	0.914509	(0.0608)	-0.23465	(0.6167)
b_NumDeb13to46ln	-4.02164	(<.0001)	-1.39338	(<.0001)	-2.12144	(<.0001)	-1.33842	(<.0001)

b_avgNumDeb13	1.644663	(<.0001)	0.044231	(0.8596)	0.509286	(0.0443)	0.008434	(0.9731)
b_OB13_to_OB4ln	4.11269	(<.0001)	0.975792	(0.0055)	1.670346	(<.0001)	0.925682	(0.0083)
b_OB1_to_OB2_ln	3.898535	(<.0001)	2.16112	(0.0001)	2.511857	(<.0001)	2.135922	(0.0002)
b_OB2_to_OB3_ln	3.405814	(<.0001)	1.84586	(<.0001)	2.156305	(<.0001)	1.822246	(<.0001)
b_OB3_to_OB4_ln	1.603699	(<.0001)	1.441582	(<.0001)	1.522008	(<.0001)	1.433785	(<.0001)
b_OB_avg_to_eop1ln	-6.12331	(<.0001)	-3.03191	(<.0001)	-3.5366	(<.0001)	-2.99721	(<.0001)
b_pos_flag_use13vs46	-12.0898	(<.0001)	2.288242	(0.0198)	0.359586	(0.7246)	2.408116	(0.014)
b_atm_flag_use13vs46	0	(.)	0	(.)	0	(.)	0	(.)
b_pos_use_only_flag_13	14.18524	(<.0001)	6.272027	(<.0001)	6.894447	(<.0001)	6.242994	(<.0001)
b_atm_use_only_flag_13	0	(.)	0	(.)	0	(.)	0	(.)
b_Tsum_crd1_to_OB1_ln	10.14031	(<.0001)	3.097325	(<.0001)	4.440962	(<.0001)	2.999493	(<.0001)
b_Tsum_crd2_to_OB2_ln	6.669577	(<.0001)	0.869902	(0.1022)	1.944404	(0.0004)	0.791814	(0.1358)
b_Tsum_crd3_to_OB3_ln	5.264313	(<.0001)	0.307452	(0.514)	1.269449	(0.0094)	0.237797	(0.6127)
b_payment_lt_5p_1	-0.97148	(0.2799)	0.200038	(0.7856)	-0.25391	(0.7388)	0.234155	(0.7495)
b_payment_lt_5p_2	-2.13679	(0.0191)	-0.99327	(0.1781)	-1.48004	(0.053)	-0.95481	(0.1943)
b_payment_lt_5p_3	-3.22266	(0.0003)	-1.34531	(0.0667)	-2.03954	(0.0073)	-1.29014	(0.0779)
b_maxminOB_limit_1_ln	-9.69192	(<.0001)	-8.05302	(<.0001)	-8.81044	(<.0001)	-7.98609	(<.0001)
b_maxminOB_limit_2_ln	-9.0928	(<.0001)	-7.61132	(<.0001)	-8.29662	(<.0001)	-7.55049	(<.0001)
b_maxminOB_limit_3_ln	-11.8167	(<.0001)	-9.03653	(<.0001)	-9.9492	(<.0001)	-8.95903	(<.0001)
b_OBBias_1_ln	-0.57251	(0.1132)	-0.9186	(0.0017)	-0.87703	(0.0039)	-0.92132	(0.0016)
b_OBBias_2_ln	-0.72347	(0.0445)	-0.68864	(0.0178)	-0.7021	(0.02)	-0.6875	(0.0177)
b_OBBias_3_ln	-1.01596	(0.0046)	-0.92929	(0.0013)	-0.9672	(0.0013)	-0.92599	(0.0014)
b_maxminOB_avgOB_1_ln	14.16565	(<.0001)	8.606256	(<.0001)	10.08646	(<.0001)	8.490683	(<.0001)
b_maxminOB_avgOB_2_ln	11.08906	(<.0001)	6.452534	(<.0001)	7.689059	(<.0001)	6.355652	(<.0001)
b_maxminOB_avgOB_3_ln	14.49649	(<.0001)	7.730027	(<.0001)	9.312205	(<.0001)	7.60887	(<.0001)
b_Tsum_deb1_to_2_ln	0.906685	(0.0188)	0.585638	(0.0481)	0.631698	(0.0414)	0.582482	(0.0487)
b_Tsum_crd1_to_2_ln	-1.17756	(0.0313)	-0.77838	(0.0662)	-0.86055	(0.0518)	-0.77347	(0.0671)
I_ch1_ln	21.60541	(0.0001)	15.27913	(0.0006)	15.08449	(0.0011)	15.31742	(0.0006)
I_ch1_flag	-7.60285	(0.007)	-6.75986	(0.0021)	-6.32835	(0.0057)	-6.80716	(0.0019)
I_ch6_flag	-7.46872	(<.0001)	-7.036	(<.0001)	-6.90948	(<.0001)	-7.05538	(<.0001)
age	-0.34863	(<.0001)	-0.84843	(<.0001)	-0.67199	(<.0001)	-0.86398	(<.0001)
customer_income_ln	10.80017	(<.0001)	62.34363	(<.0001)	44.11195	(<.0001)	63.95935	(<.0001)
Edu_High	-2.73098	(0.0034)	-7.19017	(0.0414)	-5.54035	(0.0081)	-7.34079	(0.0511)
Edu_Special	-0.18141	(0.8368)	-1.45208	(0.6639)	-1.02036	(0.6063)	-1.49001	(0.6762)
Edu_TwoDegree	-6.61697	(0.0105)	-5.54633	(0.5704)	-5.16435	(0.3727)	-5.59868	(0.5915)
Marital_Civ	5.012159	(0.0006)	12.17613	(0.0284)	10.55409	(0.0014)	12.30085	(0.0381)
Marital_Div	-1.15616	(0.257)	-0.55611	(0.8851)	-0.55753	(0.807)	-0.56146	(0.8913)
Marital_Sin	5.213276	(<.0001)	6.356839	(0.1318)	6.332679	(0.0113)	6.348672	(0.1586)
Marital_Wid	2.619305	(0.1694)	2.562945	(0.7237)	1.968299	(0.6468)	2.626267	(0.7344)
position_Man	2.973025	(0.0059)	3.991961	(0.3284)	3.930584	(0.1044)	3.991159	(0.3601)
position_Oth	2.226672	(0.0214)	4.539805	(0.2152)	3.917341	(0.0711)	4.590149	(0.2404)
position_Tech	3.960217	(<.0001)	5.109176	(0.1335)	4.796886	(0.0174)	5.134591	(0.1578)
position_Top	4.257241	(0.0592)	10.71047	(0.21)	9.256904	(0.0675)	10.82538	(0.2352)
sec_Agricult	1.388888	(0.4238)	3.113879	(0.6353)	2.434721	(0.5315)	3.176126	(0.6504)
sec_Constr	-9.4891	(<.0001)	-13.7435	(0.1283)	-12.3429	(0.0213)	-13.8656	(0.1506)
sec_Energy	-3.82661	(0.0114)	-4.04122	(0.481)	-4.45382	(0.19)	-3.99897	(0.5136)
sec_Fin	-14.6702	(<.0001)	-15.4275	(0.001)	-14.4102	(<.0001)	-15.5466	(0.0019)
sec_Industry	11.92326	(0.0002)	5.502724	(0.6516)	7.662302	(0.288)	5.305731	(0.6834)
sec_Manufact	-8.3872	(0.0003)	-12.2259	(0.1623)	-11.4667	(0.027)	-12.2775	(0.1887)
sec_Mining	-9.27426	(<.0001)	-6.96838	(0.26)	-7.83946	(0.0324)	-6.88909	(0.2969)
sec_Service	-2.87355	(0.0007)	-3.63644	(0.2564)	-3.30893	(0.0814)	-3.67005	(0.2833)
sec_Trade	-1.07519	(0.3763)	5.934037	(0.1952)	3.450967	(0.204)	6.148047	(0.2087)
sec_Trans	-4.59325	(0.0546)	-4.98028	(0.5831)	-4.96969	(0.3552)	-4.98095	(0.6071)
car_Own	-1.0202	(0.2414)	-0.87599	(0.7899)	-0.59161	(0.7615)	-0.91029	(0.7953)
car_coOwn	3.632771	(0.0042)	6.177734	(0.2005)	5.999757	(0.0359)	6.180293	(0.2303)
real_Own	-0.13203	(0.8701)	-0.89312	(0.7701)	-0.52062	(0.7737)	-0.92759	(0.7761)
real_coOwn	-1.49752	(0.0721)	-2.79608	(0.3759)	-2.50013	(0.1814)	-2.81985	(0.4028)
reg_ctr_Y	4.497857	(0.0006)	-2.26671	(0.6457)	0.267009	(0.9272)	-2.48813	(0.6364)
reg_ctr_N	10.52846	(<.0001)	3.979193	(0.4115)	6.37523	(0.0266)	3.777141	(0.4652)
child_1	2.035539	(0.0452)	1.191263	(0.7581)	1.505958	(0.511)	1.162298	(0.7784)
child_2	0.599871	(0.3061)	-0.46601	(0.8343)	-0.15677	(0.9054)	-0.49202	(0.8361)
child_3	5.781627	(0.0037)	11.37676	(0.1311)	10.06775	(0.0242)	11.48113	(0.1535)
Unempl_Infoy_6	-417.86	(<.0001)	-360.64	(<.0001)	-370.176	(<.0001)	-359.983	(<.0001)
UAH_EURRate_Inmom_6	-26.3807	(0.1303)	-35.7482	(0.0073)	-36.4485	(0.0088)	-35.6301	(0.0073)
UAH_EURRate_Infoy_6	33.88822	(0.001)	55.61245	(<.0001)	53.09088	(<.0001)	55.77344	(<.0001)
CPI_Inqoq_6	-127.861	(<.0001)	9.957103	(0.6519)	-12.6118	(0.5828)	11.58098	(0.5987)
SalaryYear_Infoy_6	151.9027	(<.0001)	38.79312	(0.0306)	54.38204	(0.0037)	37.72661	(0.0349)
s_cons_full	-1.90849	(0.047)	-2.50043	(0.0018)	-2.83696	(0.0006)	-2.46777	(0.002)
s_month_since_NA_full	-0.06993	(0.9191)	0.057644	(0.9218)	-0.01501	(0.9803)	0.065958	(0.9103)
s_month_since_Tr_full	1.364964	(0.0889)	2.022144	(0.0031)	1.989664	(0.0049)	2.022993	(0.003)
s_month_since_Re_full	1.201071	(0.2755)	-0.1684	(0.8541)	-0.03185	(0.9732)	-0.17571	(0.8475)
s_month_since_RP_full	1.735382	(0.0719)	2.606104	(0.001)	2.931744	(0.0004)	2.574662	(0.0011)
s_month_since_D1_full	1.363895	(0.1723)	1.661353	(0.0503)	2.019412	(0.0215)	1.626861	(0.0547)
s_month_since_D2_full	3.374032	(0.1552)	1.856286	(0.3473)	2.055661	(0.3153)	1.838377	(0.3507)

s_times_NA_full	6.195435	(0.5485)	-8.76979	(0.3609)	-6.106	(0.5334)	-8.94507	(0.3506)
s_times_TR_full	6.791638	(0.5167)	-3.39992	(0.7268)	-1.12017	(0.9103)	-3.55979	(0.714)
s_times_RE_full	10.83314	(0.2894)	-1.11458	(0.9069)	1.302183	(0.8935)	-1.28298	(0.8927)
s_times_RP_full	10.71537	(0.3075)	-1.49872	(0.8776)	2.032303	(0.838)	-1.76191	(0.856)
s_times_D1_full	1.880081	(0.8556)	-6.28195	(0.5133)	-4.02945	(0.6815)	-6.45879	(0.5008)
s_times_D2_full	17.38871	(0.2334)	2.304418	(0.8613)	5.65656	(0.6757)	2.033917	(0.8772)
d_StateFull_1_NA	42.66415	(0.0023)	17.58426	(0.1118)	21.65657	(0.0602)	17.29293	(0.1168)
d_StateFull_1_Tr	34.1794	(0.0121)	8.674589	(0.4203)	13.51888	(0.2281)	8.314354	(0.4385)
d_StateFull_1_Re	-15.1935	(0.1981)	-19.4144	(0.0362)	-18.8463	(0.0512)	-19.4637	(0.0352)
d_StateFull_1_RP	25.46159	(0.0523)	2.244889	(0.8282)	6.482635	(0.5477)	1.931187	(0.8515)
d_StateFull_1_D1	-13.1737	(0.2576)	-12.1347	(0.1839)	-11.8328	(0.214)	-12.1683	(0.1814)

Table A3.2. Assessing the fit of One-stage model for full (positive ATM transaction and zero income) development data sample

Model description Conditional equation	ATM Income amount - direct estimation						
	ATM Sum 6 ALL						
Coefficient	R^2	Development			Validation		
		RMSE	MAE	R^2	RMSE	MAE	
Regression Type							
Pooled	POOLED	0.2832	103.821	72.948	0.27357	106.147	74.636
Random effect - One-way							
Wansbeek and Kapteyn	RAN1_WK	0.02778	120.912	87.477	0.02176	123.177	89.491
Fuller and Battese	RAN1_FB	0.09987	116.342	83.843	0.09488	118.485	85.682
Wallace and Hussain	RAN1_WH	0.15417	112.779	80.985	0.14966	114.843	82.719
Nerlove	RAN1_NL	0.01824	121.503	87.946	0.01207	123.786	89.986
Random effect - Two-ways							
Wansbeek and Kapteyn	RAN2_WK	-2.2516	221.123	180.879	-2.18552	222.279	181.638
Fuller and Battese	RAN2_FB	0.09999	116.335	83.413	0.0961	118.405	85.198
Wallace and Hussain	RAN2_WH	0.15264	112.881	80.583	0.14925	114.871	82.268
Nerlove	RAN2_NL	-2.36331	224.89	184.313	-2.29373	226.023	185.057

Table A3.3. Assessing the fit of ATM income model - second stage: conditional on positive POS transaction (POS Sum 6 month > 0)

Model description Conditional equation	ATM Income conditional on ATMitive ATM transactions						
	OLS: ATM Sum 6 ATM Sum 6 > 0						
Coefficient	R^2	Development			Validation		
		RMSE	MAE	R^2	RMSE	MAE	
Regression Type							
Pooled	POOLED	0.2496701	91.828	69.116	0.2347403	94.649	70.744
Random effect - One-way							
Wansbeek and Kapteyn	RAN1_WK	0.1031373	102.36	77.494	0.0742999	107.761	81.518
Fuller and Battese	RAN1_FB	0.1328894	99.458	75.317	0.1015314	104.29	78.865
Wallace and Hussain	RAN1_WH	0.1555514	97.696	73.98	0.1238618	102.137	77.228
Nerlove	RAN1_NL	0.0993132	102.806	77.827	0.0709636	108.287	81.92
Random effect - Two-ways							
Wansbeek and Kapteyn	RAN2_WK	0.0384748	216.207	179.651	0.0214271	216.842	180.366
Fuller and Battese	RAN2_FB	0.1269283	99.754	75.093	0.100109	104.324	78.732
Wallace and Hussain	RAN2_WH	0.145527	98.306	73.967	0.1186458	102.506	77.358
Nerlove	RAN2_NL	0.0394181	219.477	182.609	0.0329858	220.028	183.283

Table A3.4. Assessing the Fit of Two-stage model result – Option 1: Non-zero income condition is Pr(POS) > 0.5

Model description Conditional equation	ATM Income Amount: Option 1 (ATM Sum Pr(ATM>0), Pr(ATM>0) >= 0.5; 0, Pr(ATM>0) < 0.5)					
	Coefficient	Development			Validation	
		R^2	RMSE	MAE	R^2	RMSE
Regression Type						
Pooled	POOLED	0.23486	112.815	76.278	0.22011	115.462
Random effect - One-way						
Wansbeek and Kapteyn	RAN1_WK	0.05497	125.378	85.484	0.03383	128.514
Fuller and Battese	RAN1_FB	0.09512	122.686	83.609	0.07521	125.732
Wallace and Hussain	RAN1_WH	0.12132	120.897	82.352	0.10213	123.888
Nerlove	RAN1_NL	0.04906	125.769	85.756	0.02772	128.919
Random effect - Two-ways						
Wansbeek and Kapteyn	RAN2_WK	-1.09838	186.827	137.169	-1.08235	188.668
Fuller and Battese	RAN2_FB	0.0923	122.876	83.971	0.07386	125.823
Wallace and Hussain	RAN2_WH	0.11454	121.362	82.944	0.09693	124.246
Nerlove	RAN2_NL	-1.13139	188.291	138.433	-1.11441	190.115
						140.002

Table A3.5. Assessing the Fit of Two-stage model result – Option 2: POS Sum X Pr(POS > 0)

Model description Conditional equation	ATM Income Amount: Option 2 ATM Sum X Pr(ATM > 0)					
	Coefficient	Development			Validation	
		R^2	RMSE	MAE	R^2	RMSE
Regression Type						
Pooled	POOLED	0.28797	108.83	74.189	0.27895	111.021
Random effect - One-way						
Wansbeek and Kapteyn	RAN1_WK	0.15622	118.471	82.618	0.14159	121.135
Fuller and Battese	RAN1_FB	0.18647	116.328	80.824	0.17265	118.924
Wallace and Hussain	RAN1_WH	0.20611	114.915	79.635	0.19278	117.468
Nerlove	RAN1_NL	0.15175	118.784	82.88	0.137	121.458
Random effect - Two-ways						
Wansbeek and Kapteyn	RAN2_WK	-0.70132	168.225	126.119	-0.6894	169.937
Fuller and Battese	RAN2_FB	0.1869	116.297	80.948	0.17438	118.799
Wallace and Hussain	RAN2_WH	0.20346	115.107	79.96	0.19152	117.56
Nerlove	RAN2_NL	-0.72332	169.309	127.169	-0.71072	171.006
						128.542

Appendix 2. Classification of some profit modelling literature sources

	Profit estimation - default, usage, acceptance etc.	Linear regression model and discriminant analysis	Markov Decision Process/ Markov Chains	Survival analysis	General or complex from four classes
Classification origin	Thomas(2000) – Profit scoring models can be classified by origin into four groups				
Original/basic research	<i>Duca and Whitesell (1995)</i> – credit cards effect on money demand; cross-sectional analysis of credit card use and deposits <i>White (1975)</i> - the probability that the credit card will be used for an individual's transaction	<i>Awh and Waters (1974)</i> – A discriminant analysis of Social and economic (income, age, education, state) and attitudinal: use or non-use of other credit cards, attitude toward credit, and bank charge-cards.	<i>Thomas (2007)</i> – behavioural score dynamics based profit models; portfolio profitability models; multi-dimensional state – balance, periods, characteristics	<i>Stepanova and Thomas (2002)</i> - Cox's proportional hazards model to assess both profit and default	<i>Dunkelberg and Smiley (1975)</i> - low-income credit user subsidizes the higher-income users of retail credit is not true <i>Dunkelberg and Stafford (1974)</i> – debt, panel study
General issues	<i>Cheu, Loke (2010)</i> – credit card Convenient Users and Revolvers, separate by the type of use like instrument or loan; logistic regression (pay in full/pay partially) <i>Guangli Nie et al(2010)</i> – panel data clustering for churn prediction <i>Heck(1987)</i> – probit for the probability of using only, but not the depth of usage		<i>Malik and Thomas (2007)</i> – Markov chains transition matrix model based on behavioural scores; Macroeconomic variables and Month on Book effect included		<i>Tan, Steven, Yen (2011)</i> - separation of customers on convenience users and revolvers; tobit model . <i>Cohen-Cole(2001)</i> – utilization as demand on credit; credit availability and racial factor
EaD/OB	<i>Jacobs (2008)</i> – Credit conversion factor estimation with logistic regression techniques	<i>Kim and DeVaney(2001)</i> – use OLS for amount; characteristics related to the probability to have an outstanding balance and to the amount of outstanding balance are different; higher income and willing of higher level of life lead to the increase of amount of outstanding balance, but not the probability of card use.			<i>Qi (2009)</i> - credit score, aggregate bankcard balance, aggregate bankcard credit line utilization rate, number of recent credit inquiries, and number of open retail accounts, are significant drivers of LEQ; EaD positively correlated with default rate

	Profit estimation - default, usage, acceptance etc.	Linear regression model and discriminant analysis	Markov Decision Process/ Markov Chains	Survival analysis	General or complex from four classes
Utilization and usage	<p><i>Cheu, Loke (2010)</i> – females more likely revolvers and have higher utilization; lower income customers are likely to pay in full than high income (on Malasian data)</p> <p><i>Tan, Steven, Yen (2011)</i> – probability of positive debt, tobit model</p> <p><i>Volker (1982)</i> – Multinomial regression for usage</p>	<p><i>Crook, et al (1992)</i> - The discriminative factors of the fact if customer will use or not use credit card (predictors like age, income, residential status)</p> <p><i>Agarwal et al (2006)</i> - borrowers with lower credit scores at origination utilize a lower percentage of their credit line; use FICO score</p> <p><i>Cohen-Cole(2001)</i> – credit availability and credit limit</p>			Banasik et all (2001) – two-stage Heckman model; usage is constrained by limit
Profitability	<p><i>So et al (2014)</i> – transactors/revolvers probability, transactor a priory good; profit generated from the real cost on balance in time.</p> <p><i>Ma et al (2010)</i> (take-up) – profit consists of four sources: payment, pre-payment, recovery, insurance premium; optimal interest rate for lender</p> <p><i>Finlay (2010)</i> – profit is a difference between fixed proportions of gross payment and outstanding balance; modelling individual components of profit contribution outperform traditional (default likelihood) models</p> <p><i>Oliver and Wells, 2001</i></p> <p><i>Stewart 2011</i> – the main problem of existing models is that the revenue characteristics correlated with cost predictors; dual score-based charge-off and spends model</p> <p><i>Barrios et al. (2014)</i> - the cumulative profit relative to the outstanding debt; logistic regression models for the probabilities of default and repurchase</p>	<p><i>Finlay (2008)</i> – to replace the binary classification models with continuous models; predict payments, not probability</p>	<p><i>So, Thomas (2008)</i> – credit cards states depended on behavioural score, bad status and activity; transition depends on the limit; MDP used also for low default level accounts</p>	<p><i>Andreeva et al (2007)</i> - present value of net revenue is profit estimation as a function of survival probabilities of default and of the second purchase (card usage)</p> <p><i>Ma et al (2010)</i> (take-up) – interest rate is significant for both default and pre-payment; Loan amount – for prepayment; insurance for default.</p> <p><i>Andreeva et al (2005)</i> – remaining credit after first purchase enhance the predictive power and do not have a first-order markovian property; good and bad customers have different behaviour between first and second purchase</p>	<p><i>Tan, Steven, Yen (2011)</i> – marginal effect probability of holding credit card debt and level of debt.</p> <p><i>Keeney and Oliver (2005)</i> – model - system of iso-profit and iso-preference contours ; loan offer choice to maximise profit for a given probability of acceptance</p> <p><i>Kerr, Dunn (2002)</i> - Consumer Search Behavior in the Changing Credit Cards</p> <p><i>Crook et al. (2007)</i> – profit scoring problems</p>

	Profit estimation - default, usage, acceptance etc.	Linear regression model and discriminant analysis	Markov Decision Process/ Markov Chains	Survival analysis	General or complex from four classes
				<i>Bellotti and Crook (2009)</i> –macroeconomic variables	
Business strategy			<i>So, Thomas (2008)</i> – dynamic credit limit management and optimal decision policy defines not to change limit or to change and how		<i>Keeney and Oliver (2005)</i> win-win products: customer's 'price and quality' vs. lender's 'profit and market share'; <i>Ma et al (2010)</i> - market share and the profit are interdependent; the optimal interest rate can be chosen from the iso-profits and iso-preference contours for certain loan amount