LAPPEENRANTA-LAHTI UNIVERSITY OF TECHNOLOGY LUT
School of Business and Management
Business Administration

Toni Erola

ADVANTAGES OF PROFIT CONSIDERING SCORING APPROACH IN CREDIT CARD BUSINESS: TOWARDS PREFERABLE RATIO OF PROFIT AND RISK

Examiners: Professor Mikael Collan
Post-Doctoral Researcher Jan Stoklasa

ABSTRACT

Author :	Toni Erola
Title :	Advantages of profit considering scoring approach in credit card business: towards preferable ratio of profit and risk.
Facility :	School of Business and Management
Degree :	Master of Science in Economics and Business Administration
Master's Program :	Strategic Finance and Business Analytics
Year :	2020
Master's Thesis :	LUT University, 89 pages, 19 figures, 9 tables, 10 equations, 13 appendices
Examiners :	Professor Mikael Collan Post-Doctoral researcher Jan Stoklasa
Key words :	Credit card, Credit risk, Credit application scoring, Risk-adjusted return

The purpose of the thesis is to build a framework for developing credit card application scoring-model, which ranks applications based on the ratio of expected profits and credit losses, and to find evidence whether such a model could outperform a scoring-model, which ranks applications based on the credit risk only. The literature review of this paper focuses on challenges and advantages around different approaches used in profit-based scoring models, and the findings of review are used for the framework development. The results attained with a data of a case company indicate that the suggested scoring approach can result higher net profits and ratio of profits and defaults compared to a risk-based scoring approach. The findings are consistent with the existing literature, which provides evidence that ignoring expected revenue streams in the credit card application scoring is likely to result suboptimal profitability outcome.

TIIVISTELMÄ

Tekijä :	Toni Erola
Otsikko :	Tuoton huomioivan pisteytysmenetelmän edut luottokorttiliiketoiminnassa: kohti haluttua tuotto-riskisuhdetta
Tiedekunta :	Kauppatieteellinen tiedekunta
Tutkinto :	Kauppatieteiden maisteri (KTM)
Maisteriohjelma :	Strateginen rahoitus ja liiketoiminta-analytiikka
Vuosi :	2020
Pro Gradu -tutkielma :	LUT-yliopisto, 89 sivua, 19 kuvaajaa, 9 taulukkoa, 10 laskentakaavaa, 13 liitettä
Tarkastajat :	Professori Mikael Collan Tutukijatohtori Jan Stoklasa
Avainsanat :	Luottokortti, Luottoriski, Luottohakemusten pisteytys, riskikorjattu tuotto

Tämän pro gradu -tutkielman tarkoitus on luoda viitekehys luottokorttihakemusten pisteytysmallille, joka luokittelee hakemukset odotettujen tuottojen ja luottotappioiden suhteen perusteella, sekä löytää näyttöä voiko tällainen malli suoriutua paremmin kuin malli, joka luokittelee hakemuksia pelkän luottoriskin perusteella. Tutkimuksen sisältämä kirjallisuuskatsaus keskittyy haasteisiin ja hyötyihin, joita liittyy erilaisiin tuottoperusteisiin pisteytysmenetelmiin. Katsauksen löydöksiä hyödynnetään viitekehyksen kehittämisessä. Case-yrityksen datalla saadut tulokset indikoivat, että ehdotettu pisteytysmenetelmä voi johtaa korkeampiin nettotuottoihin sekä korkeampaan tuottojen ja maksukyvyttömien vastuiden suhdelukuun verrattuna riskiperusteiseen lähestymistapaan. Löydökset ovat linjassa aiemman kirjallisuuden kanssa, joka tarjoaa näyttöjä siitä, että odotettujen tulovirtojen sivuuttaminen luottokorttihakemusten pisteytyksessä ei todennäköisesti johda optimaaliseen kannattavuuteen.

ACKNOWLEDGEMENTS

The focus of this Master's thesis is something I'm really interested as a professional, and despite the fact that the process was quite at some points quite heavy to complete along long workdays, I didn't consider compromising my goals for a moment. I think the knowledge I acquired during this research will be invaluable for my future career around finance and analytics, and I would be dishonest if I would claim that I'm not excited about the findings of this paper, and their practical value.

I want to thank the case company for providing such an opportunity to carry out this research and all the support they have provided for me along the way. I'm also thankful for Christoph Lohrmann and all the other professors and post-doctoral researchers of LUT who have enabled me to acquire necessary analytical skills to complete this research. I'm also very grateful to all my fellow students who have been sparring with me during my studies. I see great future ahead of all of us.

Helsinki 26th of September 2020

Toni Erola

TABLE OF CONTENTS

1 INTRODUCTION 5

1.1 BACKGROUND 5

1.2 FOCUS OF THE RESEARCH 6

1.3 OBJECTIVES AND RESEARCH QUESTIONS 7

1.4 RESEARCH METHOD AND MOTIVATION 7

1.5 STRUCTURE OF THE RESEARCH 8

2 THEORETICAL BACKGROUND 10

2.1 STATISTICAL TECHNIQUES USED IN CREDIT SCORING 10

2.1.1 CLASSIFICATION 11

2.1.2 TIME SCALE AND SURVIVAL ANALYSIS 13

2.1.3 CONTINUOUS MODELS 14

2.2 MODEL EVALUATION 16

2.2.1 CONFUSION MATRIX 16

2.2.2 AUC AND GINI INDEX 17

2.2.3 R-SQUARED AND ADJUSTED R-SQUARED 17

2.3 REGULATORY FRAMEWORK 18

2.4 REVENUE STREAMS OF CREDIT CARD BUSINESS 20

3 LITERATURE REVIEW 20

3.1 LITERATURE SELECTION PROCESS 20

3.2 PROFIT-BASED SCORING APPROACHES 21

3.2.1 ADVANTAGES OF PROFIT-BASED MODELS AND ALTERNATIVE APPROACHES FOR MODEL DEVELOPMENT 22

3.2.2 CHALLENGES OF PROFIT-BASED SCORING-MODELS 25

3.3 RESULT AND DISCUSSION OF REVIEW 26

4 THEORETICAL FRAMEWORK FOR RISK-ADJUSTED RETURN MODEL AND HYPOTHESIS FOR MODELLING METHODS 30

5 EMPIRICAL STUDY OF RISK-ADJUSTED RETURN CONSIDERING PROFIT-BASED CREDIT APPLICATION SCORING APPROACH 33

5.1 Presentantion of the Case 34
5.2 Data and Methodology 34
5.3 Descriptive Analytics of the Data 37
5.4 Data Pre-Processing 41
5.5 Predictive Analytics 41
5.5.1 Hypothesis 1: Classification of Applications as Active and Passive Credit Accounts 45
5.5.2 Hypothesis 2: Create Separate Profit Models for Applications with Different Level of Credit Risk 52
5.6 Results 53

6 SUMMARY AND CONCLUSION 58

6.1 Discussion of Results 61
6.2 Limitations and validity 62
6.3 Further reasearch 62

REFERENCES 64

APPENDICES 70

Appendix 1 - Data Descriptives Before Pre-Processing 70
Appendix 2 – Data Descriptives After Pre-Processing 72
Appendix 3 – Summary of Base Model Statistics 74
Appendix 4 – Summary of Model 1 Statistics 75
Appendix 5 – Summary of Model 1B Statistics 76
Appendix 6 – Summary of Base Model 2 Statistics 77
Appendix 7 – Summary of Model 2 Statistics 78
Appendix 8 – Profit Outcomes of Model 2 by Cutoff Level 79
Appendix 9 – Summary of Model 1C Statistics 81
Appendix 10 – Summary of Model 3A Statistics 82
Appendix 11 – Summary of Model 3B Statistics 83
Appendix 12 – Summary of Model 3C Statistics 84
Appendix 13 – Summary of Model 3D Statistics 85

LIST OF FIGURES

Figure 1 Focus of the Thesis ... 6
Figure 2 Structure of the Research ... 9
Figure 3 ROC curve of Model X ... 13
Figure 4 Example of Predictor Transformation Effect ... 15
Figure 5 Advantages and Challenges of Profit-Based Credit Scoring-Models in Credit Card Business ... 28
Figure 6 Theoretical Framework for Developing Risk-Adjusted Return Model ... 32
Figure 7 Overview of Model Development Process ... 34
Figure 8 Summary of Methodology ... 37
Figure 9 Profit Density: All Credit Card Accounts vs. Activated Accounts ... 38
Figure 10 Outliers of Profit ... 39
Figure 11 Outliers Within Numeric Application Information ... 39
Figure 12 Relationship of Realized Defaults and Profits ... 40
Figure 13 Relationships of variable AA and Profit (plot A) and log(AA+1) and Profit (plot B) (Does not include the highest outlier observation of AA for sake of interpretability) ... 42
Figure 14 Residual Distribution of Model 1B with Test-Data (Plot A) and Residuals by Realized Profits (Plot B) ... 44
Figure 15 WOE and IV of Binned Variables Z and DD ... 46
Figure 16 ROC curve of Model 2 with Test-Data ... 47
Figure 17 Outcome Densities by Probability of Ending up Passive ... 49
Figure 18 Residual Distribution of Combination Model (3A, 3B, 3C and 3D) with Test-Data (Plot A) and Residuals by Realized Profits (Plot B) ... 52
Figure 19 Updated Framework for Developing Risk-Adjusted Return Model ... 54

LIST OF TABLES

Table 1 Example of Confusion Matrix (Model X) ... 16
Table 2 Modelling Methods and Key Findings ... 29
Table 3 Profit Outcomes of Model 2 by Cutoff Level ... 48
Table 4 Confusion matrix of Model 2 ... 49
Table 5 Comparison of Stand-Alone Profit-Model and a Model Using the Probability of Passiveness as a Predictive Variable ... 51
Table 6 Results of Risk-Adjusted Return-Based Scoring-Model on Validation Dataset ... 56
Table 7 Results of Risk-Based Scoring-Model on Validation Dataset ... 56
Table 8 Results of Risk-Adjusted Return-Based Scoring-Model on Whole Dataset ... 57
Table 9 Answers for the Reasearch Questions ... 60

LIST OF EQUATIONS

Equation 1 Formula for Logistic Regression (Crook et al. 2007) ... 11
Equation 2 Formula for Weight of Evidence (Fair Isaac Corporation 2014) ... 12
Equation 3 Formula for Information Value (Fair Isaac Corporation 2014) ... 12
Equation 4 Formula for Linear Regression (Brooks 2014) ... 14

Equation 5 Example of Quadratic Function Form in Linear Regression (Olive 2017) 15
Equation 6 Formula for Gini Index (Doumpos, Lemonakis et al. 2019) 17
Equation 7 Formula for Expected Credit Loss (Basel II 2005)........ 18
Equation 8 Suggested Formula for Risk Adjusted Return........ 30
Equation 9 Formula for Estimated Risk-Adjusted Return........ 55
Equation 10 Formula for Realized Risk-Adjusted Return........ 55

1 INTRODUCTION

This master's thesis aims to study alternative approaches for scoring credit card applications. Application scoring has often been approached from the credit risk perspective or from the profitability perspective. As both of these aspects are crucial for long term success in credit business, it's meaningful to study alternative methods considering the both in order to achieve higher profitability without losing the touch with good credit risk management.

1.1 Background

Credit application scoring-systems are essential tools for credit risk management in financial companies. Not just to gain more insight about the level of credit risk, but also to process the huge volumes of certain credit products, such as credit cards. According to payments statistics of European central bank (ECB) from 2019, by the end of 2018, almost 28 million payment cards with credit function had been issued within the Euro area, and there was an increase of four percentage from 2017. The scoring-algorithms allow financial companies issuing the credit cards (issuers) to evaluate these huge volumes of incoming applications fast, efficiently and reliably.

Credit scoring-systems typically refer to models with creditworthiness rating or probability of default (PD) as an output (Doumpos, Lemonakis et al. 2019). These models which rank customers based on their creditworthiness are discussed in this paper as risk-based scoring-models. However, in case of credit cards and other revolving credit products, profits arise from the activity of the customer and their usage behavior. And, as the ultimate goal of issuers is not to minimize the risk, but to maximize the profits in long-term, it is not necessarily enough to evaluate and compare the levels of credit risk, but also the levels of expected profits as well.

Previous studies have already found evidence that profit-based scoring-models, which consider the expected usage and revenue, can lead to a higher overall profitability in credit card business (Finlay 2008; Stewart 2011) — at least with considerably higher level of credit risk (Andreeva et al. 2007). However, investors are not only looking to maximize profits either. Therefore, measures such as Sharpe ratio are used for evaluating the trade-off between risk and return when comparing alternative investment possibilities. The objective of this study is to create such a method, which could be used to develop well performing credit application scoring-model that doesn't either rank credit applications based on credit risk nor the expected profit but based on the ratio of these two (risk-adjusted return). Considering the information available from public sources, this paper is likely to be the first one to address the approach described.

1.2 Focus of the Research

This research focuses on credit card application scoring methods. To address the topic in comprehensive manner, background information is provided about the credit card business, credit risk management, and the regulation related to these two. Risk-based scoring methods are introduced for the sake of big picture, but the profit-based models are more centric for the objectives and methods used. Focus of empiricism is in identifying good practices for profit-modelling, and in developing and testing application scoring method, which aims to maximize the risk-adjusted return. Figure 1 illustrates the focus of the thesis on high-level.

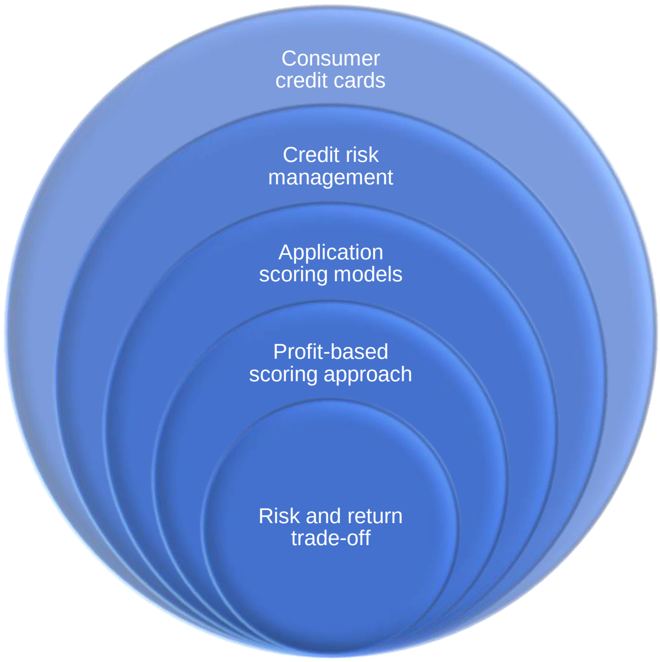

Figure 1 Focus of the Thesis

Even though credit card customers and accounts are evaluated not just when the card is applied, but almost all the time in order to identify increased risk (Witzany 2017), this study discusses the scoring methods only from credit decisions point of view. Scope doesn’t either include scoring-models used for defining appropriate credit limit for the customer. Methods for defining threshold value of score assigning the approval or decline for application (cutoff) are discussed for the sake of the entity, but these are not the core of focus.

1.3 Objectives and Research questions

This research has two primary objectives. Firstly, it aims to create a framework for developing credit card application scoring-model, which ranks applications based on the ratio of expected profits and credit losses. In the suggested approach, profits refer to difference of revenue and other costs but write-offs. The other objective is to find evidence whether such a model could perform better than traditional risk-based application scoring-model. Based on the objectives, following research questions and sub-questions were formed:

1) How to estimate profits of credit card accounts in order to conduct a scoring-model which ranks applications based on risk-adjusted returns?

2) How does an application scoring-model which ranks applications based on the risk-adjusted returns perform compared to risk-based model?
   a. Which model ranks applications better in terms of realized profits?
   b. Which model ranks applications better in terms of realized defaults?
   c. Which model ranks applications better in terms of realized net profits?
   d. Which model ranks applications better in terms of ratio between profits and amount of defaults?

1.4 Research Method and Motivation

This study is executed as a case study together with European financial company XZY. Hypothetico-deductive research method is used to complete the primary objectives and to provide answers for the research questions. In case of this research, this means that theory about the challenges and solutions related to different score modelling approaches is used to create a theoretical model, and form hypothesis around the suggested method, which ranks applications based on the ratio of expected profits and credit losses. This theoretical model and the hypothesis are then tested with an empirical study to provide answers for the research question number one. In order to do this, quantitative data provided by the case company is used. The theory introduced focuses on profit-based models over risk-based models as the existing credit risk model of the case company is used in the empirical study to measure the credit risk. In other words, there is no need to develop a credit-risk model in order to complete the objectives of this research as the PD can be estimated with the model already used in the case company. The answers for the sub-questions of second research question are provided with comparative research method. The performance of the risk-based scoring-model used in the case company is compared to an alternative approach suggested in this paper.

The study was motivated by the fact that even small improvements in the application scoring process can bring great improvements in profitability of the business. The idea of suggested alternative scoring-model has arisen from the practical issues that the author has encountered while working in consumer credit business. Based on the public sources, it also seems that the suggested approach for application scoring modelling has not been introduced before.

1.5 Structure of the Research

This paper is divided to six Chapters, the content of which are all introduced within the summary of Figure 2. The structure of the study is intended to ensure that a reader with a moderate understanding of finance and statistics can comprehend the research topic, methods, results, and possible problems around validity of the results. A reader with more comprehensive knowledge from finance and statistics is likely to be able to skip the chapter two of theoretical background.

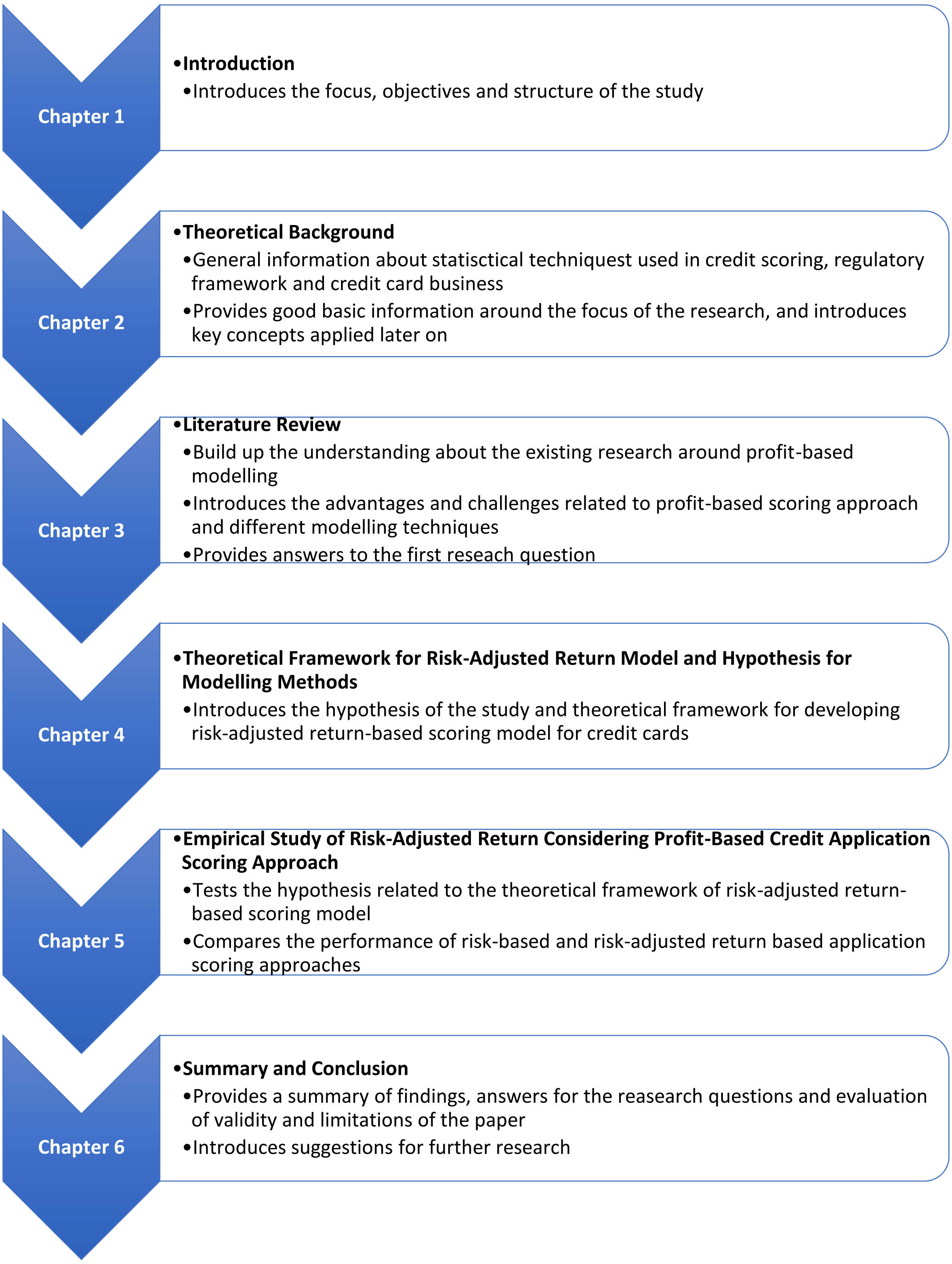

Figure 2 Structure of the Research

2 THEORETICAL BACKGROUND

Consumer lending business has existed over 750 years since the pawn brokers and the usurers of the Middle Ages, but the mass market started to develop in 1920s when Henry Ford and AP Sloan realized that they had to develop a way to finance the cars they sold (Thomas et al. 2005). Back in time, credit decisions were traditionally made by individual experts based on their subjective judgement (Marqués et al. 2013). The rise of credit card business in late 1960s made financial companies realize the benefits of scoring-models in the credit decision process; although there was really no other option but to increase the automation in order to handle the masses of applications (Thomas et al. 2002). In 1980s scoring-models became common in other lending products as well due to the success and lower default rates (bad rates) achieved with credit cards. During last few decades, credit scoring-models have been harnessed in use to answer other questions as well in addition to whether customer will default or not:

- Whether customer is likely to grab a marketing offer?
- Is the customer going to use the product?
- Is she going to keep using it in the future as well?

(Thomas et al. 2005)

Credit risk assessment methods in general have not had too much attention — at least until the financial crisis of 2008. The crisis proved how severe the consequences can be, if credit risk assessment and management fail. Since the crisis, regulatory framework related to the credit risk management and the tools of credit risk models has seen fast development, and thus forced financial operators to develop their models and methods used for risk evaluation.

2.1 Statistical Techniques Used in Credit Scoring

Credit scoring-models employ statistical models such as linear- and logistic regression, classification trees and survival analysis; operations research models like linear-, quadratic- and dynamic programming; but also more sophisticated techniques such as neural networks, support vector machines and fuzzy systems (Marqués et al. 2013). This chapter introduces few of the statistical models based on their relevancy for the study in hand.

2.1.1 Classification

First research related to credit risk assessment by scoring-models was done by Durand already in 1941. He used quadratic discriminant analysis to classify applicants as goods and bads (creditworthy and unworthy), but also to identify factors affecting the credit risk (Durand 1941). In earlier days, the specific probability for bad debt was not concerned as much as ordering the applicants correctly based on their creditworthiness because lenders tended to choose the score cutoff subjectively based on measures like approval rate (Thomas et al. 2005). Despite the fact that there is evidence of superiority of other classification methods like decision trees, random forest and more advanced techniques such as neural networks and support vectors machines over logistic regression (de Paula, Daniel Abreu Vasconcellos et al. 2019; Srinivasan 1987; Dong et al. 2010), logistic regression is nowadays the most popular application scoring method (Crook et al. 2007; Witzany 2017). Dong and his colleagues (2010) for example, have explained the popularity of logistic regression by its robustness and transparency. They also note that it might match better for certain regulations.

Logistic regression provides a probability of applicant belonging to class of Goods or Bads by using maximum likelihood method to fit following logistic function to estimate coefficients (β) with independent variables (x):

$$\mathrm{Log}\left(\frac{p_{gi}}{1-p_{gi}}\right) = \alpha + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \cdots + \beta_k x_k + u_i;$$

$$p_{gi} = \frac{e^{\beta^T x_i}}{1 + e^{\beta^T x_i}}$$

Equation 1 Formula for Logistic Regression (Crook et al. 2007)

where p_{gi} is probability of being good and u is unobservable error term. As applications can be ranked based on p_{gi}, logistic regression can also be used for other purposes than drawing a credit decision; such as for individual credit pricing (Crook et al. 2007).

In order to come up with robust model, some industry practitioners like FICO prefer binning of numeric variables in order to capture the non-linear relationships between the predictors and the predicted, and to address outliers. In their white paper for building scorecards, FICO suggest using statistics of weight of evidence (WOE) and information value (IV) in the process of identifying and binning the predictive variables. WOE can be used as a measure of standalone prediction power of single a single bin, while IV informs about the prediction power of binned variable as a whole. Formulas of WOE and IV are presented in Equations 2 and 3.

$$WOE = \log\left(\frac{G(i)}{B(i)}\right)$$

Equation 2 Formula for Weight of Evidence (Fair Isaac Corporation 2014)

where the $G(i)$ is the percentage of Goods in bin i, and $B(i)$ is the percentage of Bads in bin i.

$$IV = \sum_{i=1}^{q} \frac{Gn(i) - Bn(i)}{100} \times WOE$$

Equation 3 Formula for Information Value (Fair Isaac Corporation 2014)

where $Gn(i)$ is number of Goods in bin i, and $Bn(i)$, is number of Bads in bin i. (Fair Isaac Corporation 2014)

Unlike other classification techniques like decision trees and K-nearest neighbor, logistic regression include somewhat complicated task of defining appropriate cutoff. For logistic regression, the receiver operating characteristic curve (ROC curve) can be used to visualize the relationship of “sensitivity” and “specificity” with every different cutoff value. Sensitivity refers to the proportion of Goods classified as correctly and specificity to the share of Bads classified as Bads (Doumpos, Lemonakis et al. 2019). Figure 3 provides an example of ROC curve for Model X.

In case of Model X, achieving specificity of 75% for instance, would result misclassification for approximately 40% of creditworthy applications. Losing share this large of potential well performing customers is of course undesirable, but also underlines the problematics around cutoff determination.

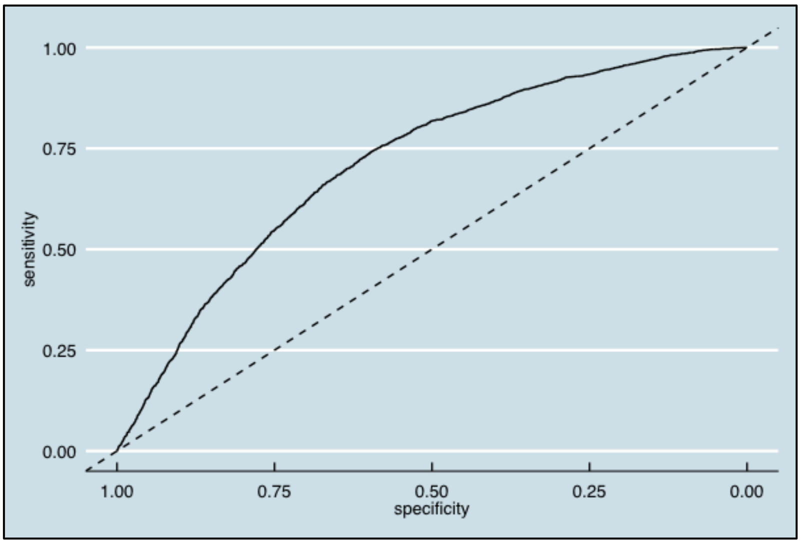

Figure 3 ROC curve of Model X

As a general approach, Verbraken and his colleagues suggest Expected Maximum Profit (EMP) measure as a proxy for the additional profits that classification model generates. The idea of the EMP is based on comparing the expected profits of alternative models to a base scenario where all loans would be accepted. In EMP, profits generated by the scoring-model rise from correctly predicted defaults as these save money by the amount of expected loss. On the other hand, costs are faced when customer is incorrectly predicted to default as the financial company loses potential income. The costs in EMP is equal to ROI which is assumed to be constant across the loan accounts. Optimal cutoff point with EMP measure, lies in a point where the EMP is maximized. (Verbraken et al. 2014)

2.1.2 Time Scale and Survival Analysis

Classification models are unable to catch the time scale of default, but sometimes it’s necessary to make prediction for periods of different length. Survival analysis is an application that can be used to estimate time to some event (Witzany 2017), such as default. Researchers like Andreeva have found evidence that survival analysis is also competitive in terms of classification accuracy compared to logistic regression for instance (Andreeva 2006). Thus, it can be used to provide more information about the nature of the risk without failing in the primary task of classification. There is also empirical evidence that for most of credit products, PD depends on the aging of the credit (time on the books) (Witzany 2017).

2.1.3 Continuous Models

It is not always the case that prediction task in scoring is related to classification at all. One might need to estimate continuous monetary value like exposure at default (EAD) or loss given default (LGD), which are other components of expected credit losses (ECL). Linear regression is optional application for modelling both EAD and LGD, but several other approaches have been suggested for consumers credits as well (Joubert et al. 2019, Tong et al. 2016, Leow, Crook 2016).

In linear regression or multiple linear regression, predicted variable (dependent or response variable) is assumed to be somewhat random (stochastic) in a way that it has probability distribution. Independent/predictor variables used to predict dependent variable are assumed to have fixed (non-stochastic) values in repeated samples. Regression finds coefficients (β) values that minimize the sum of squared unobservable error terms (residuals) in Equation 4.

$$y_i = \alpha + \beta_1 x_{1_i} + \beta_2 x_{2_i} + \beta_3 x_{3_i} + \ldots + \beta_k x_{k_i} + u_i$$

Equation 4 Formula for Linear Regression (Brooks 2014)

In Equation 4 above, y_i is the dependent variable on time i, x_{1_i} is one of the independent variables on time i and u_i is the unobservable error term on time i. (Brooks 2014)

Assumption of linear relationship can be a problem if relationship between parameter and response variable is actually nonlinear. In such case, transformation of predictors can be used to address the problem. Scatterplots can be first used to analyze the relationship between variables before fitting a transformation of independent variable with square, square root or logarithm for instance (Olive 2017). The process of predictor transformation can be useful with logistic regression as well.

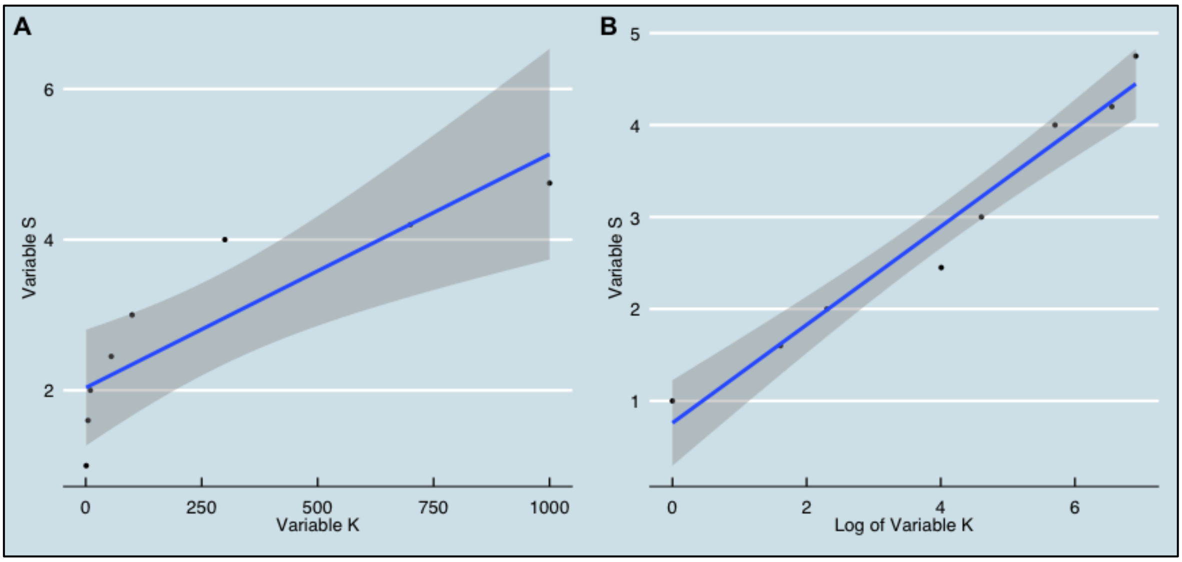

Figure 4 Example of Predictor Transformation Effect

In plot A of Figure 4, two variables K and S seem to have relationship, but it seems more like polynomial one than linear. Variable K is really skewed and is not able to explain changes of S very well through linear function. However, when S is explained by the logarithm of K (plot B), the skewness is reduced, and the linear function seem to fit a lot better. Quadratic function form can also be applied if relationship seem to quadratic (Olive 2017). In this type of case: both, predictor and its transformation are included in the function (see Equation 5).

$$y_i \; = \; \alpha + \beta_1 x_{1_i} + \beta_2 \log{(x_{1_i})} + \ldots + \beta_k x_{k_i} + u_i$$

Equation 5 Example of Quadratic Function Form in Linear Regression (Olive 2017)

Also binning can be used to catch non-linearities and in order to reduce effect of outliers (Fair Isaac Corporation 2014).

Neural networks can also be used to provide continuous estimates in addition to probabilities. Neural network is often described to mimic the functioning of human brains. It emulates networks of interconnected neurons, which are trained with historical data. Neural network models have been criticized of being hard to interpret and they are also often characterized as black boxes. (Witzany 2017)

2.2 Model Evaluation

There are number of methods for evaluation of statistical models. This sub-chapter introduces some of the most common techniques used for classification- and continuous regression models. In order to evaluate the how the model will perform, it’s important that separate test-data is used also for evaluation besides the data used in model development or training (Forsyth 2019; Olive 2017).

2.2.1 Confusion Matrix

Performance of any classification model can be easily summarized with a measure of accuracy, which is the percentage rate of correctly classified cases (Forsyth 2019). However, Accuracy doesn’t draw very comprehensive picture of the model performance. To get a better idea how the model performs with a chosen cutoff, one can examine Confusion matrix, which also indicates the level of sensitivity and specificity. Accuracy of example-model X introduced in Chapter 2.1.1 with probability cutoff of 7% is 65,15%, and it results following Confusion matrix.

Table 1 Example of Confusion Matrix (Model X)

	Actual class	
Predicted class	Good	Bad
Good	4 941	954
Bad	2 890	2245

Confusion matrix of Table 1 can be used to calculate sensitivity (63,10%) and specificity (70,18%) of the model. Model can be evaluated also based on its' precision or negative precision value. Model precision is the ratio of correctly predicted Goods and all cases classified as Goods, while negative precision value is the ratio of correctly classified Bads and all cases predicted as Bads (Witzany 2017). So, based on Table 1, precision of Model X would be 83,82 %, and negative prediction value would be 43,72 %.

2.2.2 AUC and Gini Index

Area under the ROC curve (AUC) is a goodness-of-fit measure, which indicates the probability measure of good classification (Witzany 2017). As AUC is a probability measure, it is always a value between zero (0) and one (1) (Witzany 2017). AUC of one would indicate that there are no classification errors, and zero would mean that prediction of the model is always incorrect.

Cumulative accuracy profile (CAP) curve is another popular measure for classification accuracy. In the CAP curve, false positive rate is replaced with the number of approved applications (probability default < cutoff). CAP curve plots the approval and true positive rate for all possible cutoff points. The area under the CAP curve is known as the Gini index. Gini index can be calculated also by using AUC:

$$Gini\ index\ =\ 2 \times AUC - 1$$

Equation 6 Formula for Gini Index (Doumpos, Lemonakis et al. 2019)

(Doumpos, Lemonakis et al. 2019)

2.2.3 R-squared and Adjusted R-squared

R-squared is a fraction of the variance in the dependent variable that is explained by the continuous model (Forsyth 2019). Like AUC, R-squared is a value between zero and one, where one would indicate that model can explain 100 % of the variance in the dependent variable and the sum of residuals would be zero. I.e. the model is perfect. One of major flaws in R-squared is that it never decreases when independent variables are added to the regression function, and therefore adjusted R-squared is often used (Brooks 2014).

In adjusted R-squared the, R-squared is adjusted with the number of independent variables, which helps to decide whether variable should be included in a regression model or not (Brooks 2014). I.e. adjusted R-squared can be used to eliminate predictors from a regression model if they do not improve the model more than what is predicted by chance.

2.3 Regulatory Framework

In addition to providing great benefits for financial companies, scoring-models are nowadays also necessary due to the regulatory framework of financial business. One could even argue that regulators would deserve part of the credits about the sophisticated scoring-models in use today. In 2005 Basel banking supervisory committee introduced Basel II framework, which allowed financial companies to use their own risk estimation methods for calculating capital requirements for credit risk. This is also known as advanced internal ratings-based approach (A-IRB). As an alternative, so called standardized approach could be used. Main difference between these two is that A-IRB enables financial houses with low and moderate level of risk in their portfolios to achieve lower capital requirements by modelling the account level expected credit loss (EL), which is discussed in this paper by abbreviation "ECL". While the standardized approach sets the capital requirements on one size fits all basis mostly through the type and collateral of credit, A-IRB sets it based on the sum of account level ECLs, which are estimated through parameters of probability of default (PD), exposure at default (EAD) and loss given default (LGD):

$$ECL = PD \times EAD \times LGD$$

Equation 7 Formula for Expected Credit Loss (Basel II 2005)

(Basel II 2005)

In Equation 7 above, PD indicates the likelihood of default in 12 months. In Basel framework, default in brief refers to a situation in which bank has a reason to believe that debtor is unlikely to pay his credit in full, or the credit is more than 90 days past due. EAD is the estimated amount outstanding on the time of default and LGD the percentage of EAD that is expected to be lost. (Basel II 2005)

In order to gain the capital requirement benefits, Basel II requires the implementation of A-IRB models across the business (also to the credit decision process). However, the competitive advantage of using A-IRB is shrinking as reforms introduced in Basel III (2017) set higher capital requirement floors for low risk portfolios of A-IRB compliant banks. The higher capital requirements of Basel III aim to improve the ability of banks to absorb shocks after events of the financial crises (Basel III 2017).

In 2018, IFRS 9 accounting standard came into force. The standard is required to use in almost all European public companies (IFRS Foundation 2009). The accounting standard of IFRS 9 aim to increase the transparency of financial statements from point of view of financial assets and liabilities (IFRS 9 2014). Until IFRS 9, only actual credit losses were recognized, but the new standard urge financial houses to recognize expected losses as well. Prediction of credit losses should even consider the expected changes in macroeconomic environment (IFRS 9 2014). Consequently, the requirements considering ECL modelling are more forward-looking than ones set by A-IRB. Although the concept of ECL and modelling requirements are somewhat similar as in Basel framework, there are also some other differences in addition to the requirements concerning the forward-looking approach. However, these differences are not discussed further in this paper.

In addition to the requirements related to ECL modelling, European banking authority (EBA) guides the credit risk management practices used by credit institutions. Based on EBAs guidelines, the management of financial company should limit the credit risk that the institutions exposures to by “approving and regularly reviewing a credit institution's credit risk management strategy and the main policies and processes for identifying, measuring, evaluating, monitoring, reporting and mitigating credit risk consistent with the approved risk appetite" (EBA/GL/2017/06).

2.4 Revenue Streams of Credit Card Business

Credit card issuers have multiple revenue streams to compensate the credit risk taken. Firstly, they receive interchange fees from merchant's bank for handling the payments. These fees were capped to the maximum of 0,3 % of the transaction value by the European commission in 2015 (2015/751). Other major revenue source is typically the interest, which is paid by customers for the revolving balance — the part of the balance that is not paid after the billing cycle. Customers who pay their monthly balance just partly and generate interest profits for issuers are also called "revolvers" (So et al. 2014). Pricing structures vary among the issuers but annual or monthly fees, invoicing fees, cash withdrawal fees and foreign currency conversion premiums are examples of revenue streams which are not unusual.

3 LITERATURE REVIEW

Purpose of this literature review is to build up the understanding about the existing research in the area of different consumer credit application scoring methods and modelling approaches, which consider expected profits related to the credit card or the applicant (profit-based approach). The focus of the review is in the advantages and challenges of different modelling approaches used in profit-based scoring-models of credit card applications. The information and knowledge retained is used for conducting the theoretical framework for the new application scoring approach suggested, which ranks credit applications based on the ratio of expected profits and expected credit losses (ECL).

3.1 Literature Selection Process

This chapter describes the literature selection process used in this study. Article search was started by using Finna-service which provides articles from multiple data databases, and arranges the search results based on their relevancy. Search term primarily used in the first phase was “Credit” AND “Scoring” AND “Profit” OR “Profitability”. This term was chosen to get comprehensive idea of profit-based modelling methods used. The search was limited to only peer-reviewed full articles. This resulted 1 421 objects. About 80 most relevant results were explored mostly based on their titles and abstracts in order to identify appropriate articles. For example, articles which were focused on defining cutoff or addressed corporate loans were disregarded. Also, articles addressing fixed-term loans were mostly ignored except if they had content that could be easily generalized to revolving credits. In the second phase, the most fundamental studies from point of view of this paper were recognized by reviewing citations of appropriate articles found in the first phase. In third phase, more articles were searched by authors names cited, and about specific subjects with more precise search terms such as “Finlay” AND “credit” AND “scoring”. These further searches were used to find more articles and further findings of leading authors from the areas that they have been specialized in. Most of the articles used were found from ProQuest and Elsevier, but also Google Scholar was used in step three. As a result of the process presented also in summary below, a good understanding of the literature around the focus of the study was obtained and review of this entity is presented in following chapters. RefWorks-service by ProQuest was used to store and manage articles cited during the process.

1. Search relevant articles by search term “Credit scoring” AND “Credit card”.
2. Search backwards based on citations of articles found in first phase to recognize the most fundamental ones.
3. Get further information about specific subjects based on authors cited in fundamental studies and with more precise search terms that have come up during step 1 and 2.

3.2 Profit-Based Scoring Approaches

As the primary goal of a bank is to make profit, it is somewhat surprising how scarce the amount of literature available on profit-based models is compared to the traditional risk-based models, even though several of the studies done introduce valid arguments and similar results — If the objective of the model is to achieve as a high profitability as possible, profit-based models seem to outperform risk-based models in terms of an outcome (Andreeva et al. 2007; Finlay 2008; Stewart 2011).

3.2.1 Advantages of Profit-Based Models and Alternative Approaches for Model Development

Andreeva and her colleagues (2007) address the necessity of profit-based models in credit card business and with other revolving credits by the argument that risk-based approaches such as time to default can't be used to estimate the expected profit of a credit card, as the revenue streams generated by the card usage include much more uncertainty than streams of fixed-term loans. Risk-based binary models don't include discrimination between goods and bads of different worth either (Finlay 2008). Account expect to default for 10 euros is considered as bad as account defaulting 10 000 euros regardless of the expected profits. Risk-based binary model might also perform poorly with "indeterminate" cases where the class might be ambiguous for some reason. Example of such case is a credit card which is never used as it doesn't provide neither upside in terms of revenue nor downside in terms of default (Finlay 2008).

Most popular method of conducting a risk-based model with logistic regression has been extended to consider the aspect of profits by determining revolver segment as Bads and segment of transactors (full payers) as Goods. So and his colleagues (2014) compared how model which uses credit card usage segments performs compared to one which discriminate applicant based on credit risk. They justified their approach by the argument that as full payers always pay their balance in full, they can't default and are therefore to be classified as Goods. On the other hand, by classifying customers based on the usage segment, the model includes more information about the expected usage behavior and therefore about the profitability as well. (So et al. 2014)

To consider the credit card usage besides the credit risk when applying survival analysis technique, Andreeva and her colleagues proposed a method where the survival analysis of time to default was expanded to a survival combination model which also consider the usage of a credit card. In their study, time to second purchase was used to measure the probability of card usage in the same way that time to default is used to evaluate credit risk. The results suggested that the method described could be used to achieve higher profitability — at least with higher level of risk (Andreeva et al. 2007). Survival analysis method has been fitted to different type of profit-based models also formerly, but most of these have addressed fixed-term loans (Banasik et al. 1999; Hand, Kelly 2001; Stepanova, Thomas 2002). Andreeva has later on developed the survival analysis approach further with Sanchez-Barrios and Ansell (2016). Their quite recent study compared results achieved with model, which measures time to profit by survival analysis and profits in monetary terms by linear regression (Barrios et al. 2016). Results suggested that survival analysis can outperform linear regression at least when measured with relative measure such as return on asset (ROA) (Barrios et al. 2016). Barrios and his colleagues (2014) were actually first ones to emphasis the importance of using relative measure over monetary measure in profit-based scoring. They have rationalized the usage of relative measure with an argument that if two customers provide the same expected profit and score, one with lower amount of capital bound is likely to be more attractive for the issuer due to productivity of funds (Barrios et al. 2014). However, it seems that models using monetary units are likely to achieve higher overall profitability in monetary terms (Barrios et al. 2014; Barrios et al. 2016).

Finlay (2008) have examined whether modeling account worth or profitability should be done with classification model or some alternative approach which provides continuous outcome. He ended up comparing results of three different models.

A. A risk-based binary model by logistic regression (default = bad)
B. A profit-based binary model by logistic regression (unprofitable = bad)
C. A profit-based continuous model of account worth by linear regression

In his study, accounts were sorted to ten score bins with the three different models and the results suggested that continuous profit-based model is likely to outperform a binary model when these two are compared from point of view of ranking accounts based on their worth. To address the possible problematics of linearity assumption, continuous independent variables were coded as dummy variables, which each included a decile of observations (Finlay 2008). Although continuous model seemed superior in terms of overall profitability, it also led to significantly higher bad rate compared to classification models A and B, which both led to very similar outcomes (Finlay 2008). Later on, Finlay (2010) compared the results achieved with continuous models conducted with linear regression, neural network (NN) and genetic algorithm (GA). He found out evidence that both NN and GA could outperform linear regression, and GA seemed superior (Finlay 2010). However, the GA based model didn't provide an estimate of profit, but a continuous and relative score (Finlay 2010).

Serrano-Cinca and Gutiérrez-Nieto (2016) have also compared the performance of profit-based model by multivariate regression and risk-based model by logistic regression, although their study used peer-to-peer loans. However, they had similar results as Finlay. Best 100 credits in terms of score provided internal rate of return (IRR) 5,98 % with risk-based model while continuous profit model achieved IRR of 11,92. One of the key contributions of Serrano-Cinca and Gutiérrez-Nieto (2016) were related to the relationships of independent variables and profitability. They found that relationships between profits and independent variables are often non-linear and thus, non-linear regression models might be needed. In case of peer-to-peer loans, it also seemed that credit risk and profit could be explained with different factors. (Serrano-Cinca, Gutiérrez-Nieto 2016).

As a solution for the possible problem of reckless increase in risk caused by strong correlation between risk and profits when using continuous profit-based model, Stewart (2011) suggested that risk- and profit-based models would be used side by side and multiple profit-models would be developed for different risk-bands to identify such independent variables for the models which correlate with the profits but not with the risk (Stewart 2011). Based on the results, this type of approach can be used achieve higher profits without discarding a good credit risk management. Idea of building a score by modelling individual components of profitability separately have been proposed already before Stewarts' study (Finlay 2010).

Fitting multiple models to the credit decision process might although set some additional challenge for defining cutoff. A method proposed by Oliver and Wells (2001) is an alternative one to find a cutoff point when two different models are used. They fitted the concept of efficient frontier to credit scoring. Based on efficient frontier method, a cutoff point should be chosen so that any other cutoff point doesn't provide the same expected profit with lower level of expected losses for instance. The idea of efficient frontier can also be applied to other objective trade-offs as well; such as profit-volume or risk-volume trade-off (Oliver, Wells 2001) to ensure that higher market share could not be achieved with different cutoff resulting the same profit- or risk-level.

3.2.2 Challenges of Profit-Based Scoring-Models

There have been several arguments for the unpopularity of profit-based models besides the problematic of risk-profit correlation and many of those concern the problems around measuring the target variable itself: profitability. Firstly, assigning fixed costs for account level can be tricky (Thomas et al. 2002; Finlay 2008). For instance, allocating costs such as personnel cost over different products, customers and accounts is a complex task and the procedures are also likely to change over time. In previous researches, the problem of defining and measuring profitability of credit cards have been approached by some type of simplification. To simplify the profitability analysis, variables such as amount of credit card spend and charge-offs or write-offs have been used as proxies for revenue and costs for instance (Andreeva et al. 2007; Stewart 2011).

From the modelling point of view, the characteristics of profit distribution can also cause some problems. Due to defaults and high number of totally passive accounts, profits are likely to be quite far from normally distributed (Stewart 2011). Accuracy of profit-based model is also likely to suffer from the changes that take place during the prediction period. Finlay (2008) refers to this as a tracing problem and gives an example where a single credit card can be highly profitable because of a credit limit increase made during the modelling period. And, due to this kind of change, the customer profile at the start of the modelling period can differ a lot from the profile at the end the period. Also changes in macroeconomic conditions are likely cause changes in realized profits (Stewart 2011) through interest spread movements for instance.

3.3 Result and Discussion of Review

Profit-based models include the problem of measuring the target variable (profits) reliably enough. It's of course complex to predict something, which one can't even measure. Previous studies have used variables such as credit card spend and charge-offs or write-offs as proxies for revenue and costs. However, these types of solutions are suitable only in single occasions since the different customer segments generate different amount of profits with their spend. So called transactors or full payers do not usually produce any interest profits, whereas the segment of revolvers doesn't necessarily generate as much transaction related fees like interchange fee (So et al. 2014). Using write-off as a proxy for cost can also be misleading as the LGD and thus the net exposure differs by account. Levels of LGD affect costs also through the capital requirements set in the Basel regulatory framework and the IFRS accounting standards. For the reasons mentioned, the reliable measure of profit is vital for successful profit-based model.

Whereas risk-based models suffer from overlooking the expected profits, profit-based models do not necessarily consider the practical fact that most of the financial institutions must follow some pre-determined credit risk management strategy and risk appetite which set the limits for the risk taken. Therefore, it's likely to be necessary to include some risk-limiting factor to the profit-based model or use risk- and profit-models side by side in the application decision process as Stewart have suggested. This also means increased amount of required resources in developing, monitoring and upkeeping the models used as risk-models are anyway necessary to be maintained due to IFRS 9 for instance. And, even though the credit risk would be modelled separately, profitability distribution is likely to be skewed due to the numbers of passive and unprofitable cards. This sets some additional challenge for the modelling and is likely to affect the model accuracy.

Despite the limitations and problematics around the profit-considering scoring approaches, risk-based approach is able to inform only about the possible downsides related to the application processed. Thus, it is not fitting to the credit card business where uncertainty related to revenue streams is on whole another level than in fixed-term loan business. Compared to good/bad binary risk model, a continuous profit-based models can provide more information about the “goodness” or “badness” of the application. As Finlay stated, it is relevant whether expected default is 10 or 10 000 euros as the former one is covered with the revenue streams rather quickly. Credit card business can also benefit the profit-based models at least in theory to reduce the number of non-active credit cards as these would be approved in most risk-based systems despite the negative overall profit due to fixed costs. As discussed earlier in this paper, defining a score cutoff can be a challenging task. In case of continuous profit-models, this can be seen relative easier as in many cases rational choice would be to approve all applications which are expected to turn into net profit contributing credit cards. Figure 5 provides a summary of possible advantages and shortcomings related to profit-based credit scoring-models in credit card business.

+ Is likely to lead higher profitability

+ Can be used to identify and delcine credit accounts which are likely to end up passive and thus unprofitable

+ Continuous profit-models provide more information than risk-based binary models

+ Defining score cutoff for continuous profit-model can be seen relative easier than for classifiaction models

- Reliable measure for profitability is needed

- Skewed distribution of profits is likely to deteriorate the accuracy of the model

- Due to high correlation of expected risk and revenue, it might be hard to identify applications with sub-optimal profitability expectations

- Model must include credit risk component as credit institutions must follow their credit risk management strategies

Figure 5 Advantages and Challenges of Profit-Based Credit Scoring-Models in Credit Card Business

This review also revealed which kind of measures have been used to address the challenges around profit-based models, and what type of results and findings have been done related to these proceedings. Table 2 below provides the review of different modelling approaches and techniques used in previous researches.

Table 2 Modelling Methods and Key Findings

Authors	Suggested method	Key findings
Andreeva et al. (2007)	Modelling both, expected risk and usage by survival combination model	A survival combination model with separate survival analyses for default and card usage can be used to achieve higher profits than simple risk-based survival model.
Finlay (2008)	Modelling profits in monetary terms by linear regression	Continuous monetary measure and linear regression for modelling profits can be used to achieve higher profitability in monetary terms, but it is also likely to increase credit risk significantly compared to risk- and profit based binary models.
Finlay (2010)	Genetic algorithm with linear function	Genetic algorithm can outperform linear regression and neural network as modelling techniques in profit-based model, and models using separate sub-models for components of overall profits, are likely to outperform single aggregated models.
Stewart (2011)	Combination model (classification and linear regression)	By modelling returns and write-offs separately, higher profitability can be achieved without losing control over credit risk.
So et al. (2014)	Using logistic regression to which distribute applicant as full payers and revolvers	In credit card business, logistic regression which consider segment of full payers as Goods and revolver as Bads, is likely to achiveve higher profits than traditional risk-based logistic regression model, which assign credit decision based on the risk of default.
Sanchez-Barrios et al. (2016)	Modelling time to profit by Survival analysis	Survival analysis, which provides time to profitability can be used to achieve higher ROA compared to linear regression model which provides monetary outcome.
Serrano-Cinca & Gutiérrez-Nieto (2016)	Modelling profits in monetary terms by non-linear regression	Credit risk and profits can be explained with different factors, the relationships of independent variables are often non-linear and continuous non-linear regression for modelling profits can be used to achieve higher profitability in monetary terms compared to logistic regression.

As table 2 illustrates, results of previous researches around profit-based application scoring-models are encouraging. It also seems that there are multiple different options for statistical techniques, which are all likely to outperform risk-based models in terms of realized profits. So, silver bullet of profit-based models is yet to be found, and therefore the best modelling approach around profit-based models is likely to depend on unique needs of credit card issuers.

4 THEORETICAL FRAMEWORK FOR RISK-ADJUSTED RETURN MODEL AND HYPOTHESIS FOR MODELLING METHODS

Both risk- and profit-based scoring approaches have their own shortcomings which are mostly related to the fact that neither minimizing risk at cost of reduced profits nor maximizing profits at any cost doesn't necessarily work very well in credit business. To achieve higher profits with limited risk, this paper concentrates on evaluating a new method, which aims to rank applications based on the expected risk-adjusted return. The theoretical framework for modelling risk-adjusted return, which tries to consider the challenges of profit-based models will be introduced in this chapter.

As it is necessary for most of European issuers to model ECL for the credit cards, it is somewhat reasonable to use the existing models in credit decision process to evaluate the credit risk component in the equation of risk-adjusted return. The risk-adjusted return is calculated in the suggested method as follows:

$$Risk\ adjusted\ return\ = \frac{Profits}{Credit\ risk}$$

Equation 8 Suggested Formula for Risk Adjusted Return

where Profits is the difference of revenue streams and costs; other than write-offs. For the credit risk component, ECL is preferred as a proxy for risk over PD to reckon the post-write-off revenue streams. The measure of profits should consider the internal policies for assigning fixed costs and time-period should match the credit risk component. These two components of risk-adjusted return are suggested to be modelled separately.

As the distribution of profit component is likely to be skewed especially because of passive credit card users, the effect of two-stage profit-model on the prediction accuracy is tested and compared to single-stage profit-model. In the two-stage model, credit accounts with high probability of being passive are identified with classification model and assigned with profit equal to profit/cost generated by passive credit card before providing an expected profit with continuous model. In the single-stage model, only continuous model is used. So, first hypothesis (H1) proposes that two-stage profit-model which includes separate classification model for identifying passive credit accounts can be used to achieve better prediction accuracy.

$H1_0$ = Two-stage profit-model which includes separate classification model for identifying passive credit accounts achieves better prediction accuracy than single-stage model which ignores the classification of applications.

$H1_1$ = Two-stage profit-model which includes separate classification model for identifying passive credit accounts doesn't achieve better prediction accuracy than single-stage model which ignores the classification of applications.

Due to high correlation of credit risk and expected revenue, the profit-model might not recognize factors that differ from the risk model. Thus, the profit-model would not necessarily be able to create any additional information for the credit decision process. In order to tackle this possible problem, approach originally proposed by Steward is suggested to be used for predicting profits within different risk-bands. The second hypothesis (H2) proposes that using separate continuous profit-models for different risk-bands is likely to improve the prediction accuracy.

$H2_0$ = Using separate continuous profit-models for applications with different level of credit risk improves the prediction accuracy.

$H2_1$ = Using separate continuous profit-models for applications with different level of credit risk doesn't improve the prediction accuracy.

Based on the result of literature review, theoretical framework for developing risk-adjusted return considering application scoring-model was conducted and is presented in Figure 6. Ranking applications based on the risk-adjusted return has several advantages over other approaches discussed in this paper. These advantages and practical features of the system are addressed in following paragraphs.

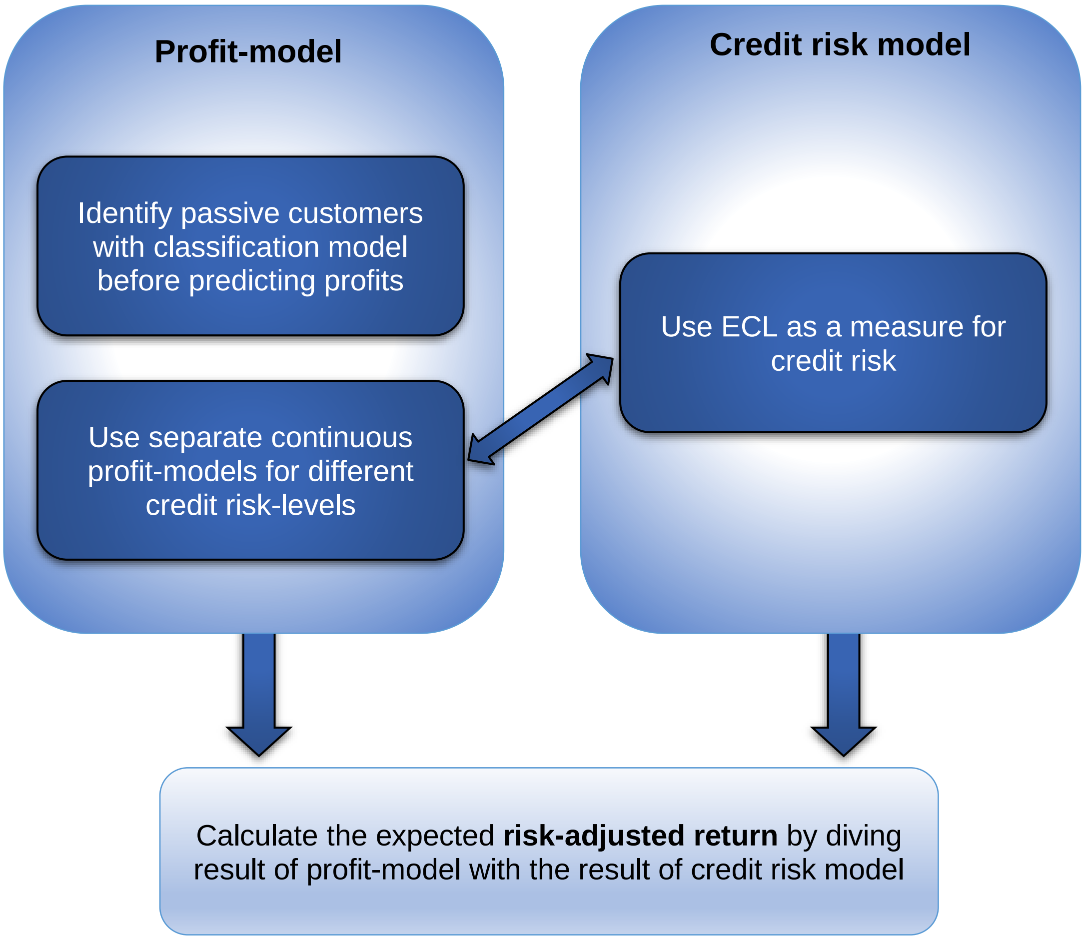

Figure 6 Theoretical Framework for Developing Risk-Adjusted Return Model

When ranking the applications based on the ratio of expected profits and credit losses, preferable cutoff point is relatively easy to derive from risk appetite noted in company's credit risk management strategy. In aggressive strategy, even ratio of one could be used. This would basically be equal to simplistic profit-based approach as all applications with positive expected outcome would be approved. By choosing desirable ratio as a cutoff, financial company can also determine an adjustable buffer between risk and profit. This is valuable feature in a scoring-system as the level of risk is easy adjusted to match its credit risk appetite during constantly changing market environment.

Whereas simplistic profit-based scoring-model would accept an application with expected profits of 100 euros and expected credit losses of 90 euros due to expected positive contribution to profits, risk-adjusted return considering system with cutoff ratio of 1,5, would decline such application because it doesn't meet the required risk-buffer. In theory, risk-adjusted return considering profit-based approach could also lead to higher profits compared to risk-based approach as high risk applications could also be approved if the expected excess returns are high enough to compensate the high risk taken. Thus, there would be less potential profits lost. Whereas the traditional risk-based system would decline application with PD over the cutoff threshold, risk-adjusted return considering system would do the same only if expected excess profit for the high risk taken would not meet the requirements set in the credit risk strategy.

The suggested method has been inspired by earlier studies related to profit-based credit scoring. It uses separate models for the risk and profit components as it's been suggested by Finlay (2010) and Stewart (2011) for instance. It provides a continuous outcome, which is likely to deliver more information than a good/bad binary model like Finlay has proposed (2008). The score is also relative figure as Barrios and his colleagues suggested (2016).

5 EMPIRICAL STUDY OF RISK-ADJUSTED RETURN CONSIDERING PROFIT-BASED CREDIT APPLICATION SCORING APPROACH

Aim of this empirical study is to test the hypothesis related to the conducted framework of modelling risk-adjusted return for credit cards and to compare the results of finalized modelling approach to those achieved with a risk-based scoring approach used in the case company. As it's been discussed, credit institutions have guidelines for credit risk modelling which arise from the regulation. Thus, the empirical study concentrates on the subject that can be approached more freely: modelling the profits. The objectives of this research are designed to be achieved by the findings of this empirical study: (1.) *create a framework for developing application scoring-model, which ranks applications based on the ratio of expected profits and credit losses* and (2.) *find evidence whether such a model could perform better than traditional risk-based application scoring-model.*

5.1 Presentantion of the Case

The case company is using risk-based scoring approach in its' credit decision process for credit cards. However, the company and the author have recognized the shortcomings of using risk-based approach with revolving products and alternative options are therefore explored. Simplistic profit-based approach, which aims to maximize profits is not seen as an option as good credit risk management is considered vital by the company. After exploring the literature and considering possible alternative approaches, idea of risk-adjusted return model was born. Now the idea of this new approach is developed further by using application data of the case company. An overview of this process is provided below.

Figure 7 Overview of Model Development Process

Model development process is started with data gathering in co-operation with the case company. Before testing hypothesis by conducting profit-models with alternative approaches, data is processed to a form which suits the methodology. After hypothesis are tested, final profit-model is built, and its' results are compared with the risk-based model.

5.2 Data and Methodology

The dataset used in this study includes application information of approved and declined applications from time period less than a year. However, only approved applications are used in the modelling dataset as profits or losses cannot be retained for the declined observations. The dataset includes mostly information that applicant has given for the issuer regarding her financial position, but also internal segmentation information and demographics retained from the public databases. Within the data gathering and selection process, the previous knowledge acquired by the company from different credit risk and customer behavior analyses were used to identify possible variables, which might be related to credit card usage and profits. Criteria whether the variable is available for all applicants; new and existing ones was also used as criteria. This is because the resulted model is expected to be suitable for all credit applications.

The data was validated by using logical checks and by verifying especially outlier values through the credit card ledger user interface of the case company. Logical checks were used to make sure that logical errors do not exist. As an example, profits cannot exist without card usage. Variables chosen for the modelling are either numeric, binary or categorical. All numeric values including the profits were multiplied with a constant value in order to protect the sensitive data content.

In the process of testing hypothesis 2 and the framework conducted based on the theory, multivariate linear regression with backward elimination process is used to predict the 12-month profit of credit card account. One year was chosen as observation period so that the length of the profit-model observation period would match the period length of the most common PD-model used in the company. As discussed earlier in this paper, calculating fixed costs precisely is very complicated task, and therefore this study relies on the estimate provided by the case company.

Multivariate linear regression was chosen as a modelling method because it can be used to predict continuous value such as profit and because of good results achieved in previous studies of Finlay (2008), Serrano-Cinca and Gutiérrez-Nieto (2016). Although there is evidence that for example NN and GA models could outperform regression when applied for profit-based scoring-model, these techniques were not considered transparent, and therefore practical enough. In order to avoid some possible problems related to the assumption of linearity, also non-linear function forms are used in model development.

For testing the hypothesis 1, logistic regression is used for classifying the credit accounts to two categories; ones, which are expected to be passive and ones expected to be active. This approach was chosen due to its' transparency and the assumption that high misclassification costs might rise from declining profitable applications. As logistic regression enables to set the tradeoff between the sensitivity and specificity on desired level, it's considered suitable for the purpose unlike decision trees and K-nearest neighbor for example. The fact that logistic regression is widely used within the case company, also favors the choice as model developed with familiar technique is likely to be easier to communicate, monitor and maintain.

During the development of logistic regression model, Akaike information criteria (AIC) is tried to minimize using backward elimination process of predictors. I.e. model development is started by using all independent variables, and the predictors are then removed from the model one by one by starting from highest p-value, so that only predictors, which don't contribute for the AIC are removed permanently.

After the classification model is conducted, a continuous profit-model for accounts, which are expected to be active based on the classification is developed and the results achieved with the two-staged profit-model (logistic regression + linear regression) is compared to results of single-stage profit-model with linear regression. Mean absolute error (MAE) will be used for the comparison. MAE was chosen as a measure instead of root mean squared error (RMSE) as the goal is to achieve good prediction accuracy for the masses, and MAE does not give as much weight for large singular errors as RMSE.

One-sample T-test is used for testing of both hypotheses. After overall prediction errors of different models have been compared by MAE, the absolute observation-level prediction errors are compared by T-test to test whether the mean of absolute errors is significantly different. 0.05 will be used as significance level (Alpha = 0.05).

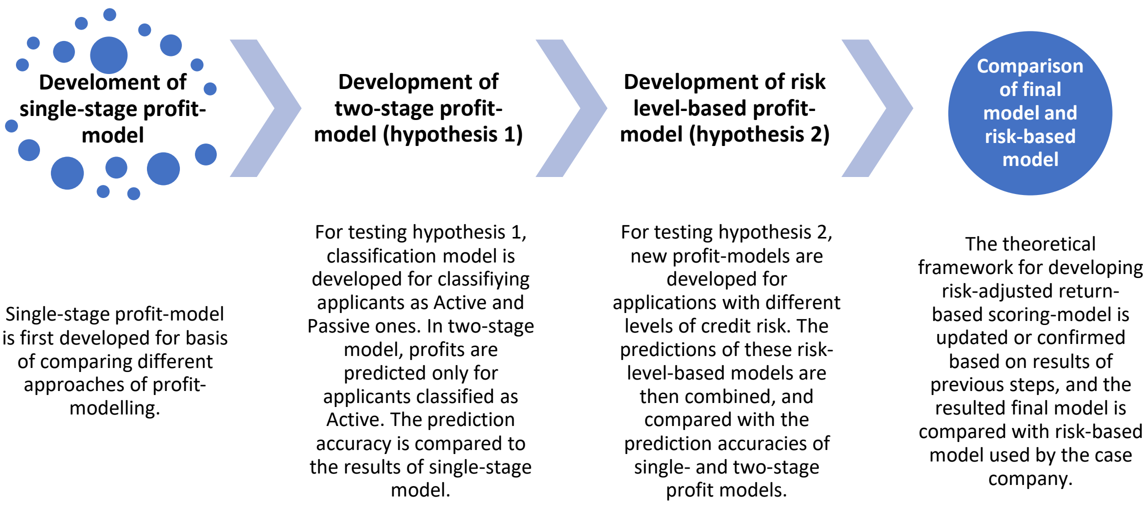

Figure 8 Summary of Methodology

The methodology of the study is summarized in Figure 8. Confidence level of 5 % is used for the testing of hypotheses. Model development process is completed by using R programming software and following libraries: mgcv, psych, dplyr, plyr, ggplot2, vtreat, tidyr, nnet, OptimalCutpoints, WVPlots, broom, sigr, devtools, rlang, gridExtra, reshape2, scales, rpart, rpart.plot, caTools, class, caret, purrr, pROC, cowplot, stringr, tibble, ROCR, grid, ggthemr, ggthemes, data.table, ggpubr, ggfortify, ggResidpanel, pscl, logiBin, woeBinning, glmdisc, e1071, regclass, margins and lmtest. Microsoft Excel is also used for Data pre-processing and editing different tables presented later in this paper.

5.3 Descriptive Analytics of the Data

The credit card application dataset retained from the case company includes 52 776 observations. Appendix 1 provides descriptive analytics of the original dataset before pre-processing. Most of the numeric variables are very skewed and long tailed, including the target variable: Profit (kurtosis = 23.246). This was also expected based on the literature review though. As discussed earlier, skewness of the dependent variable might deteriorate the accuracy of the profit-model. In Figure 9, density plot A includes all credit accounts while plot B includes only active credit accounts: accounts, which had been used at least once during the 12-month period.

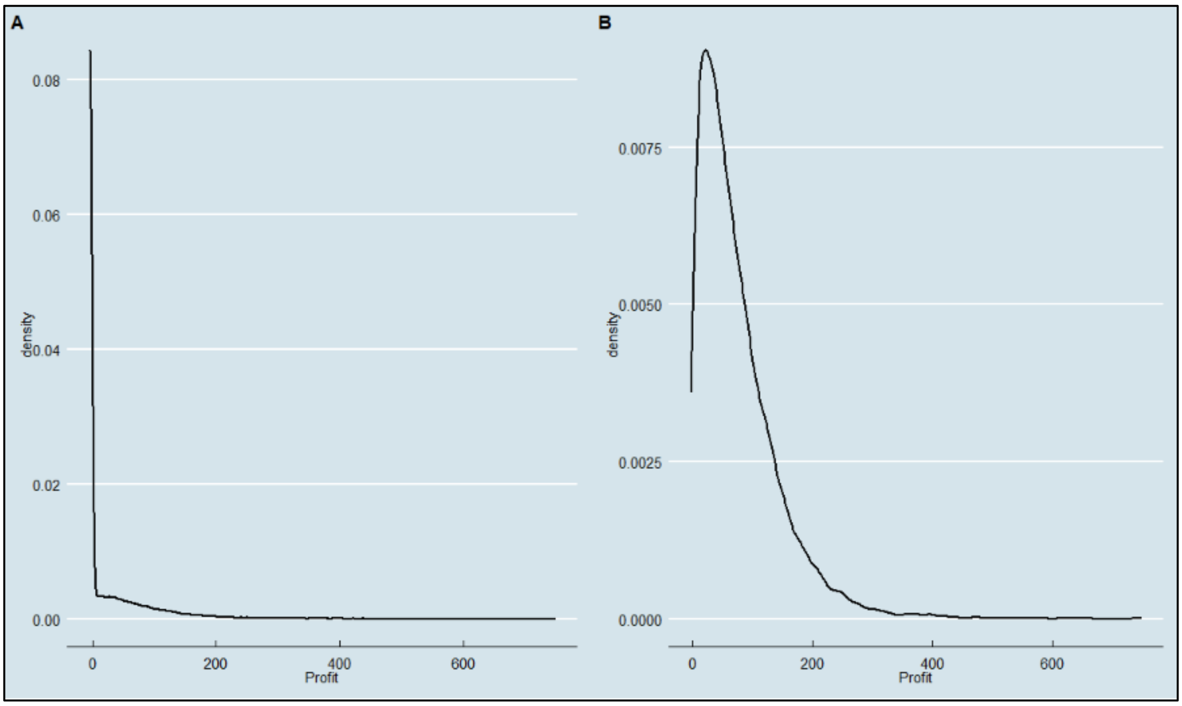

Figure 9 Profit Density: All Credit Card Accounts vs. Activated Accounts

Although, the profits from activated credit accounts are neither normally distributed, there is a lot less skewness. Even though the outliers of Profit are likely to deteriorate the prediction accuracy of the profit-model, there is no justification to exclude these observations from the modelling dataset, as discarding these exceptional and interesting observations could also lead to worse model performance with unseen data. Boxplot of Figure 10 illustrates the extremeness of outliers in Profit variable.

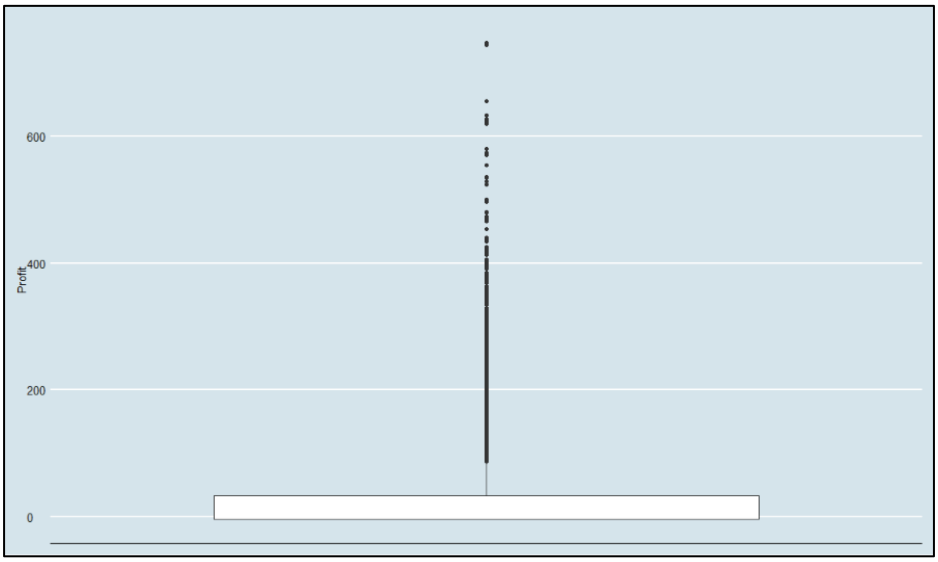

Figure 10 Outliers of Profit

Many of the numeric independent variables, such as variables T, U, X, Y, Z and AA presented within boxplots of figure 11 contain information entered by the customer herself. Therefore, in case of these variables, the high-end outlier values can be caused by typing errors for example. However, even these types of observations were not excluded from the modelling as similar observations are expected occur in the future as well.

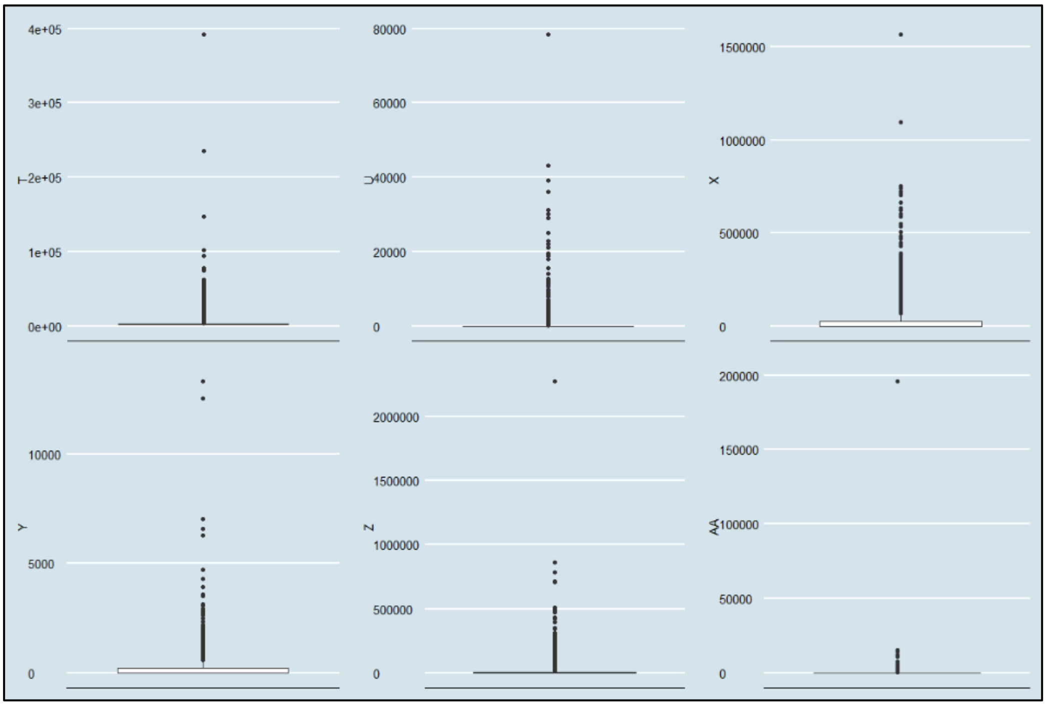

Figure 11 Outliers Within Numeric Application Information

From point of view of hypothesis 2, the relationship between credit risk and Profits is also very interesting. To provide insight of this important matter, Figure 12 illustrates the relationship between PD and Profit during first 12-months of credit card lifecycle. 409 PD outlier observation were excluded from the figure for sake of interpretation.

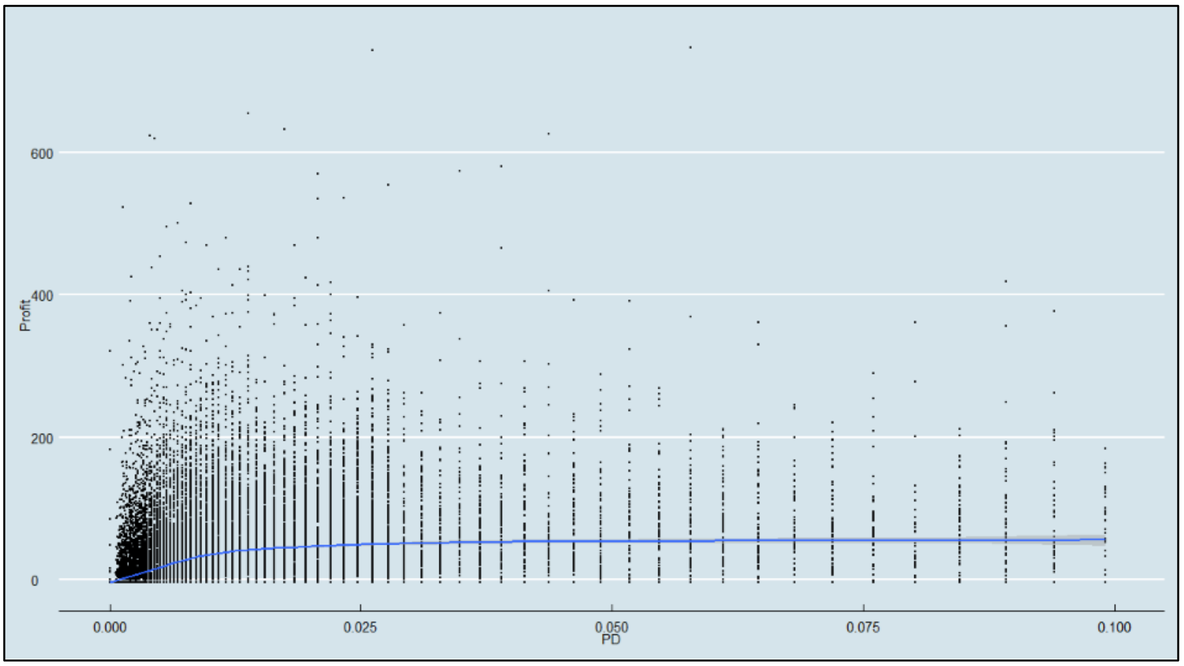

Figure 12 Relationship of Realized Defaults and Profits

Based on the plot, there seems to be somewhat linear relationship between PD and Profit roughly up to PD of 1 %, after which the increase of PD does not seem to have a similar effect in Profit variable. The blue smoothed line is locally estimated scatterplot smoothing (LOESS), and its' purpose is to help with the detection of trend within the relationship of the variables. Although the relationship between the credit risk and profits seems to be somewhat linear only to certain extent based on Figure 12, majority (~65.4 %) of credit applications have a PD lower than 1 %. Thus, examining hypothesis 2 and method suggested by Stewart (2011) (modelling profits separately for different risk-levels) seems justifiable based on the data as well.

5.4 Data Pre-Processing

In data pre-processing, binary and categorical variables were transformed as dummy variables, where value of one represent true or belonging to the category. The transformation is necessary due to the chosen modelling method (regression). Most of the numeric variables are very skewed, but as discussed; no justification for excluding the outliers was found. The original dataset included information from 52 776 applications, but two observations were removed due to missing data. Appendix 2 provides data descriptive analytics after the pre-processing phase.

For the model development, only approved applications were used (33 395 observations in total) as profits cannot be modelled with declined applications, which do not generate profits. Approved applications of pre-processed dataset were split into three subsets, one for model development purpose (training data), second for model testing (test-data) and last one for validation purposes (validation data). The split was done randomly so that 60 % of observations were assigned to the training dataset and test was split evenly to test-data (10 %) and validation data (30 %).

5.5 Predictive Analytics

Profit-model development was started by setting up a Base model: a multivariate linear regression model, which includes all the predictive variables. Appendix 3 provides a summary of Base model including its' coefficients and p-values of predictors.

The Base model achieves adjusted R-squared of 0.2143 and MAE of 32.52178 with the test-data. Before starting to eliminate variables, which do not contribute to the model, all numeric variables were plotted against the profit variable in order to identify potential transformation methods, which could improve the prediction power of the model. As descriptive analysis of Appendix 2 shows, most of numeric variables are very skewed. Reducing the skewness and extreme outliers by square roots or logarithms for instance is likely to lead better fit of model for the bulk of the data as the heavy tails of data might have heavy effects on the coefficients. And, if relationships between independent and dependent variables seem for example polynomial instead of linear, transformation of variable is done to find better fitting function form for the regression.

In Figure 14, plot A illustrates the relationship of variable AA and Profit whereas plot B shows the relationship between logarithm of AA plus one and Profit. In the latter, the LOESS curve seems to follow the black regression line more closely, which indicates for better function form. Adding one to the value of variable is necessary in case of variable AA and all the other variables with possible zero-values if log-transformation is used as logarithm can't be calculated for zero.

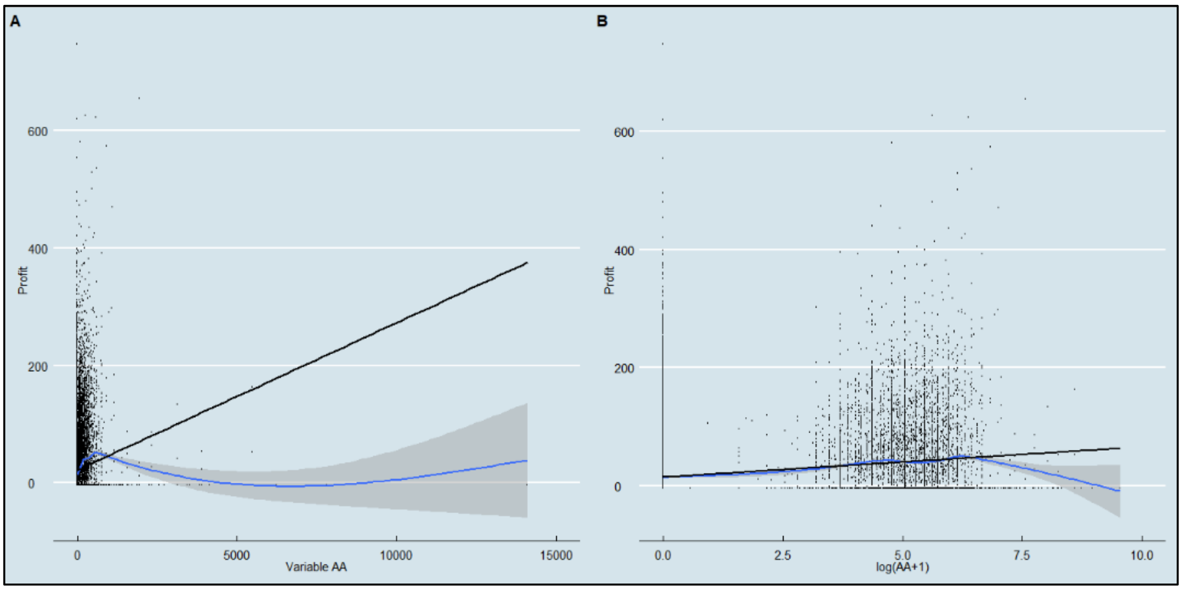

Figure 13 Relationships of variable AA and Profit (plot A) and log(AA+1) and Profit (plot B) (Does not include the highest outlier observation of AA for sake of interpretability)

In case of variable AA, transformation also increases the Pearson's correlation between the predictor and the response variable from 0.015 to 0,230. Using log(AA+1) within the Base model instead of actual value of variable AA, improves the Adjusted R-squared from 0.2143 to 0.2319, and the transformation is therefore applied to the developed model. Alternative transformations methods were used by trial and error for all numeric variables in the model development. In addition, quadratic form was also applied for some predictors.

After finding appropriate transformations, unnecessary variables were removed from the model by using backward elimination process, where the predictors with highest p-values were removed one by one in case the removal didn't have negative impact on the Adjusted R-squared. As an outcome of this process, Model 1 was developed. The model achieves Adjusted R-squared of 0.2548 and MAE of 31.57579 with the test-data resulting a decrease of 0.94599 in MAE compared to the Base model. However, result of Ramsey RESET test (p-value < 2.2e-16) provided evidence of non-linearity and problems in the function form of the model.

Due to the results of Ramsey RESET test and in order to find a better function form, new independent variables were formed by binning all the numeric variables, and by trying to replace the numeric variables with binned ones. The replacement was applied only if it seemed to contribute the overall fitness of the Model 1 measured by Adjusted R-squared. The binning itself was done with a similar approach as the variable transformation: by plotting the independent variables against Profit, and by identifying patterns within the relationships. Binned versions of variables were named by adding "2" to the original variable name. For instance, variable U2 is a binned version of variable U. Model 1B was formed after replacements of numeric variables were completed. The Model 1B includes both: transformed numeric variables and binned ones, and it lifts the model Adjusted R-squared up to 0.2607 and pushes the MAE down to 31.44568. Also, exponents of independent variables were added to the regression function, but better fitting function form couldn't be found.

The Model 1B was also examined from multicollinearity, heteroskedasticity and residual distribution point of view. High correlation (variance inflation factor (VIF) > 5), was recognized with 10 variables in total. However, this is not considered as a problem as the covariance between variables should be similar within all datasets, and the primary goal is to achieve high prediction accuracy instead of trying to understand the variable relationships profoundly. Breusch-Pagan test was run to find out whether the Model 1B has a constant variance. As the test resulted a p-value close to zero (< 2.2e-16), the model is clearly heteroskedastic. The problem of heteroskedasticity was considered difficult to remove due to the very skewed nature of the data. Plot B of Figure 15 illustrates the heteroskedasticity.

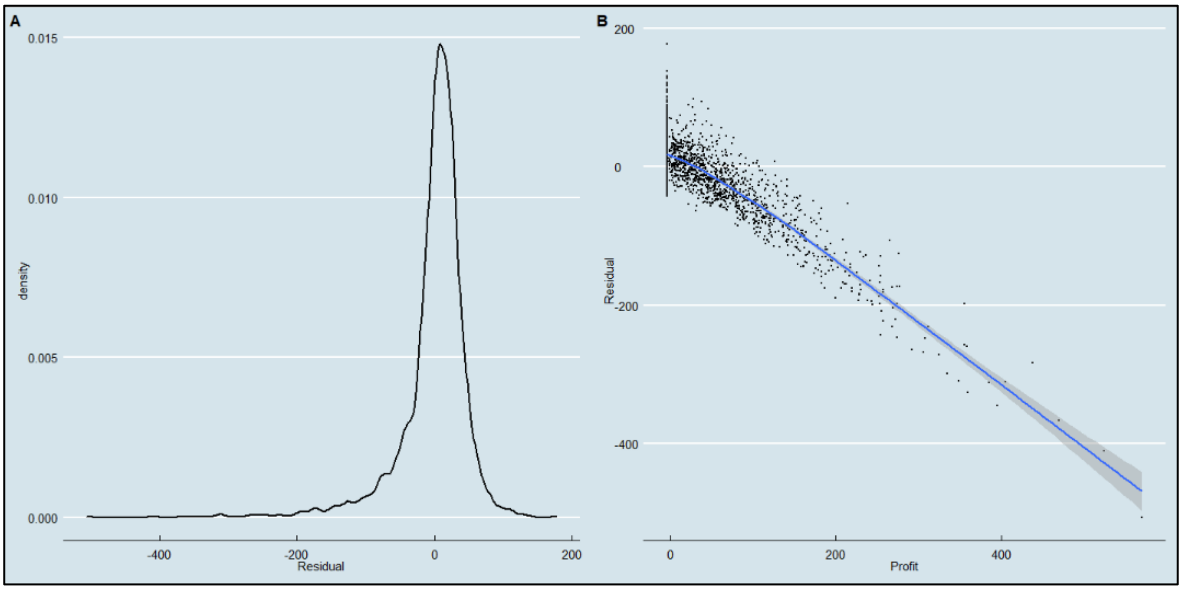

Figure 14 Residual Distribution of Model 1B with Test-Data (Plot A) and Residuals by Realized Profits (Plot B)

Plot A provides the distribution of residuals. Plot B also indicates that in overall the model tends to predict too high profits for low profiting credit cards while high-end outliers are not caught very well either. The prediction error within low profiting credit cards emphasize the importance of Hypothesis 1 and identifying applicants, which are likely to be passive. The residuals are not either normally distributed due to the fact that Profit outliers are not caught very well by the model. The mean of residuals with the test-data is -0.87 meaning that on average the predictions are a bit too low. Kolmogorov-Smirnov test was used to confirm that this is not caused by significant Profit distribution differences between the datasets used for the modelling and testing.

Lowest value predicted by Model 1B on test-data is around -45,73. Although yearly fixed costs cannot be this high, there is no reason to limit the predicted value as the model is also supposed to rank applicants, and information would be lost if predictions would have a minimum value. However, if the predictions less than possibilistic minimum would be forced to the minimum value (-3.83768), the MAE of Model 1B would decrease from 31.44568 to 29.99078. For comparison, intercept only model has a MAE of 38.31985. The statistics of Model 1B is presented in Appendix 5, and the model is next compared to the outcome of alternative model development approach suggested in Hypothesis 1.

5.5.1 Hypothesis 1: Classification of Applications as Active and Passive Credit Accounts

As with the profit-model, development of logistic regression model for identifying to be passive credit cards was started with a model, which included all the predictors. This model (Base model 2) and its' key statistics are presented in Appendix 6. 2. Base model 2 has AUC of 0.7661 and Gini of 0.5322 with test-data.

To construct a robust model, which could catch non-linear relationships, all the numeric variables were binned in the next phase. Binnig was done by using weight of evidence (WOE), which is a measure of the separation in the predicted class. The process was completed by merging initial classes (20) based on WOE. The merging was repeated until information value (IV) decreased more than three percentage compared to the previous binning step. Three percentage was chosen as a threshold value for IV decrease as it was considered to be a level where the information loss would be acceptable, but the overfitting would likely to be avoided in most parts. Binning was discarded if the resulted bins were not useful for prediction based on low IV (< 0.02). This was the case for variable T (IV = 0.019). Highest IVs were retained for bins of Z (0.271) and DD (0.268). Bins of these variables are presented in Figure 16. Binned versions of variables were named by adding "3" to the original variable name. For instance, variable Z3 is the binned version of variable Z.

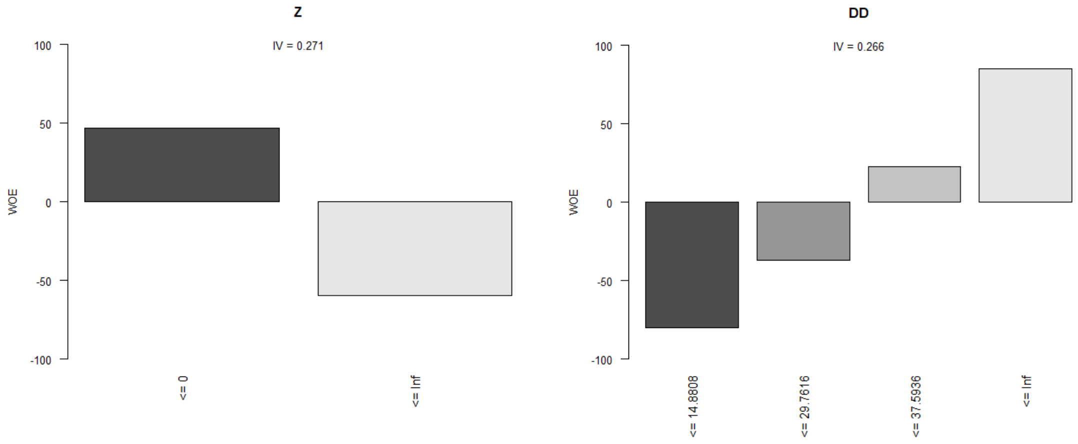

Figure 15 WOE and IV of Binned Variables Z and DD

During the next phase of model development, all unnecessary predictors, which didn't contribute for AIC were removed from the model one by one starting from highest p-value. The Base model 2 had AIC of 21 962, whereas the model formed as a result of variable removal process (Model 2) has AIC of 21 048. The Model 2 resulted AUC of 0.7838 and Gini of 0.5676. As with the Model 1B, VIF outcomes provided evidence of multicollinearity. Eight variables of Model 2 resulted VIF greater than five. The summary of model statistics is provided in Appendix 7, and the ROC curve is presented in Figure 17.

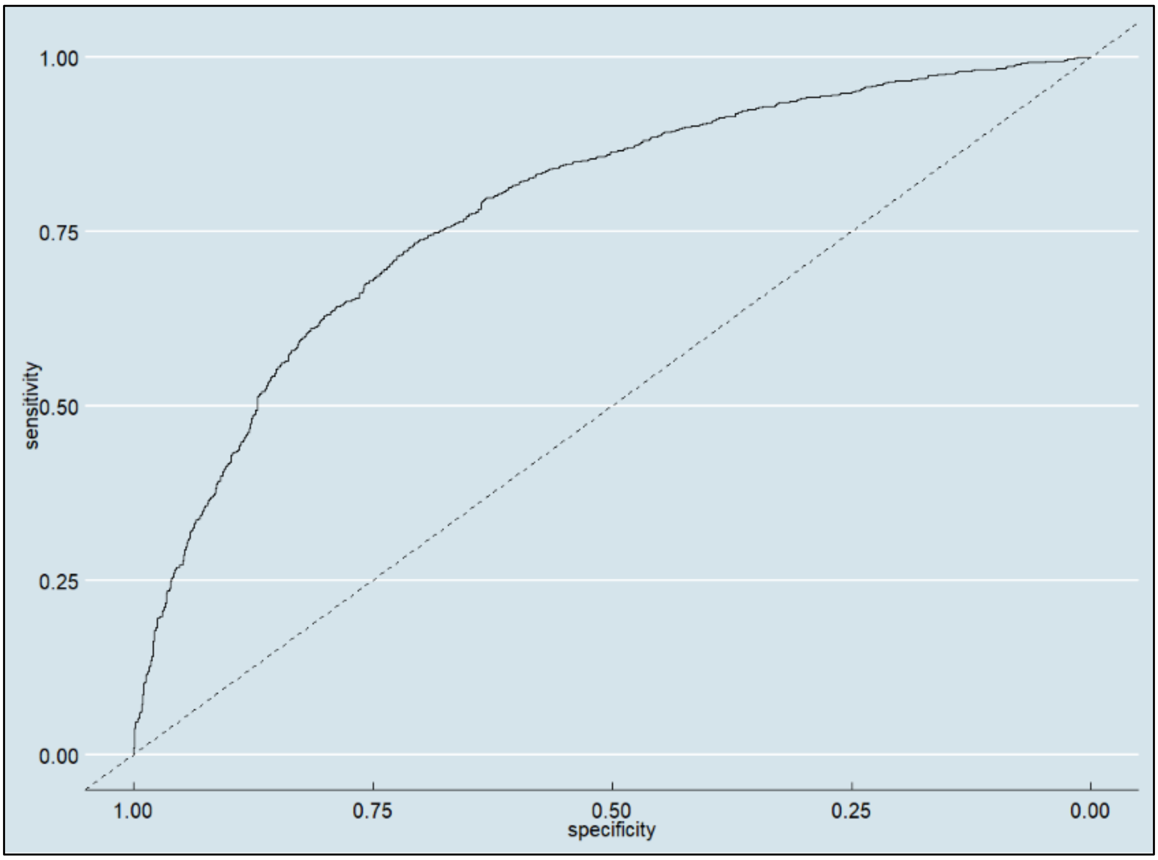

Figure 16 ROC curve of Model 2 with Test-Data

As discussed earlier, misclassification of passive credit card users might result significant misclassification cost due to lost profit opportunities, and therefore high sensitivity and negative precision value is required for the Model 2. To find appropriate cutoff point, EMP maximizing approach suggested by Verbraken and his colleagues was first applied. However, using average ROI or average profit of active credit card as a proxy for misclassification cost was realized as inapplicable because the profits of those active credit card accounts, which had a high probability of being passive based on the Model 2 had significantly lower average profits than those active accounts, which had a low probability score. I.e., the EMP would be significantly too low for high cutoff points. And on the other hand, too high for low cutoff points. Therefore, a similar method was applied, but by using realized profits. Table 3 introduces the profitability outcomes on test-data by different cutoff levels.

Table 3 Profit Outcomes of Model 2 by Cutoff Level

Probability cutoff	Approval rate	Profits from approved	Profit improvement (€)	Profit improvement (%)
100 %	100.00 %	82 693.00 €	- €	0.00 %
95 %	95.57 %	83 003.77 €	310.77 €	0.38 %
90 %	85.83 %	83 521.08 €	828.08 €	1.00 %
85 %	75.32 %	83 127.49 €	434.49 €	0.53 %
80 %	66.58 %	81 495.90 €	-1 197.10 €	-1.45 %
75 %	58.19 %	80 000.15 €	-2 692.85 €	-3.26 %
70 %	50.73 %	76 360.13 €	-6 332.87 €	-7.66 %
65 %	43.91 %	72 055.37 €	-10 637.63 €	-12.86 %
60 %	37.98 %	66 978.45 €	-15 714.55 €	-19.00 %
55 %	32.35 %	62 062.24 €	-20 630.76 €	-24.95 %
50 %	27.88 %	55 310.62 €	-27 382.38 €	-33.11 %

In the Table 3, Approval rate is the share of applications below the cutoff (classified as Active), and the Profits from approved is the sum of realized profits from those credit cards. The baseline for profit improvement is a model which would approve all applications (cutoff = 100 %). Based on the Table 3, the profit maximizing cutoff level for test-data seems to be somewhere around 90 %. Appendix 8 however provides more detailed information on the matter and reveals that optimal cutoff point would be probability of 88 % if alternative cutoff levels are compared by percentage point. I.e. based on the results of Appendix 8, all applications with probability of being passive greater than 88 % should be declined in order to maximize the profits. The 88 % cutoff level would result improvement of 954.44 Euros in total profits (1.15 %). The contribution of Model 2 can therefore be considered as quite minor.

Figure 18 illustrates the discrimination power of Model 2 through densities of different outcomes on both sides of the optimal cutoff point (black line).

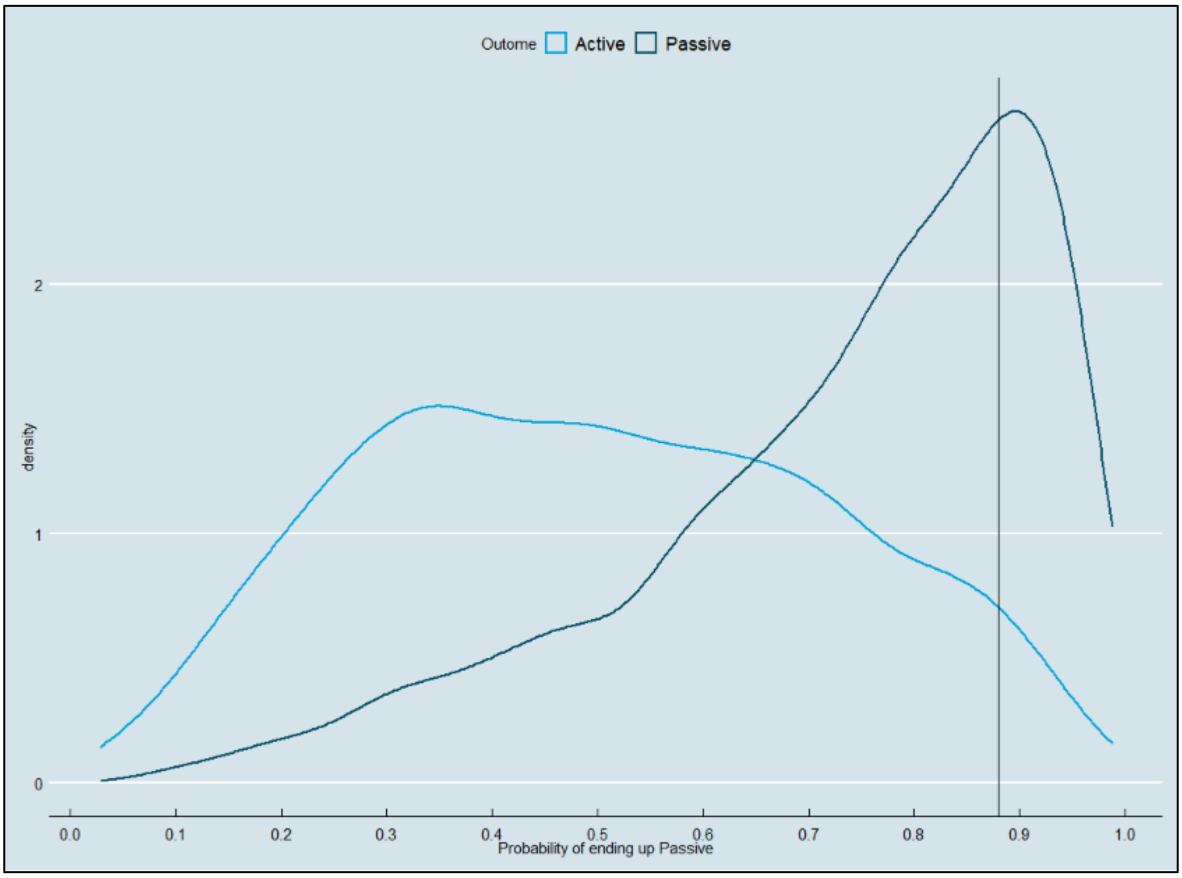

Figure 17 Outcome Densities by Probability of Ending up Passive

Figure 18 also shows that the selected cutoff point (0.88) is likely to result really low model specificity as majority of actual passive observations have probability of ending up passive under the cutoff point. The confusion matrix of Table 4 illustrates this a bit further.

Table 4 Confusion matrix of Model 2

	Actual class	
Predicted class	Active	Passive
Active	1 161	1 550
Passive	56	572

Confusion matrix shows that out of 2 122 actual passive customers only 572 are classified correctly by Model 2, and 56 of 1217 active customers would have been declined due to negative profit expectation. Misclassification cost rising from these 56 applicants would have been around 1 242 euros, which explains the weak contribution of Model 2 for the overall profits despite the high sensitivity of (95.54 %) and negative prediction value (91.08 %).

During the next phase of testing Hypothesis 1, prediction accuracy of Model 1B is compared to two-stage profit-model in which the prediction of Model 1B is applied only for applicants classified as active by Model 2. The fixed cost of credit card (-3.83768) is assigned as a predicted profit for applicants classified as passive. So that prediction accuracy of approaches would be comparable, the minimum value of prediction generated by Model 1B is also limited the fixed costs of credit card as the actual profit cannot be lower than this. As mentioned earlier Model 1B has a MAE of 29.99078 with test-data when predictions are limited to minimum possible value. The comparison is however done by using validation data as test-data cannot be used because the cutoff level of Model 2 was defined based on the results on test-data, and it's therefore biased. The MAE of Model 1B with validation data and limited minimum prediction is 29.93223. Two stage-model in which the applicants are classified as active and passive before applying the continuous prediction results a MAE of 29.47802, which means decrease of 0.45421 compared to the one-stage model. Although there is lower prediction error when two-stage model is applied, one sample T-test on the absolute prediction error returns a p-value of 0.3776 indicating that the difference in prediction error is not statistically significant. As the two-stage model doesn't seem to result significantly lower prediction error, the null hypothesis is rejected: *Two-stage profit-model which includes separate classification model for identifying passive credit accounts doesn't achieve better prediction accuracy than single-stage model which ignores the classification of applications.*

Regardless of the outcome of hypothesis testing, the probability of passiveness provided by Model 2 was next applied to Model 1B as a new independent variable. By adding the outcome of Model 2 to Model 1B, the Adjusted R-squared of Model 1B increases from 0.2607 to 0.2700 and the MAE on test-data decreases from 31.44568 to 30.64434. Based on this discovery, the Model 1C was developed with similar method as Model 1B but using the probability of passiveness as additional predictor. The MAE of Model 1C on validation data is 1.7306 lower than one of Model 1B and the T-test on the absolute prediction error returns a p-value of 0.0006496, which indicates that profit-model using the probability of applicant to end up as a passive credit card holder, is likely to achieve lower prediction error than stand-alone profit-model like Model 1B.

Table 5 Comparison of Stand-Alone Profit-Model and a Model Using the Probability of Passiveness as a Predictive Variable

Score decile	Profit of Model 1B	Cumulative profit	Profit of Model 1C	Cumulative profit
10	90 288.66 €	90 288.66 €	92 838.05 €	92 838.05 €
9	52 970.55 €	143 259.21 €	53 565.27 €	146 403.32 €
8	37 845.47 €	181 104.68 €	37 868.76 €	184 272.08 €
7	27 440.55 €	208 545.23 €	27 554.48 €	211 826.56 €
6	17 997.44 €	226 542.67 €	16 134.00 €	227 960.56 €
5	12 164.89 €	238 707.56 €	11 524.39 €	239 484.95 €
4	7 734.66 €	246 442.22 €	8 444.25 €	247 929.20 €
3	3 666.02 €	250 108.24 €	2 223.52 €	250 152.72 €
2	578.40 €	250 686.64 €	367.43 €	250 520.15 €
1	-2 188.84 €	248 497.80 €	-2 022.35 €	248 497.80 €

Table 5 illustrates the gain achieved with Model 1C compared to the stand-alone model (Model 1B). In the Table 5, the credit card accounts are split to score deciles based on the predicted profit so that the highest decile (tenth) includes those 10 % of credit card accounts within the validation data, which are expected to generate the highest profits. Table 5 shows that the Model 1C can identify the best profiting credit card applications better than Model 1B as the highest score deciles of Model 1C generate higher profits than those of Model 1B. For example, by approving only highest score decile, Model 1C would have generated 2 549.39 Euros (2.82 %) more than Model 1B with the application stream of validation data. Model 1C is also superior over Model 1B in terms of cumulative profit until the two lowest score deciles. Based on these discoveries, model using the probability of passiveness as a predictive variable is preferred for profit-modelling when risk-adjusted return-based scoring-model is applied.

5.5.2 Hypothesis 2: Create Separate Profit Models for Applications with Different Level of Credit Risk

Hypothesis two suggest that using separate continuous profit-models for applications with different level of credit risk improves the prediction accuracy. Consequently, the training data was next split to quartiles based on PD. Four additional multivariate regression models were then developed to predict the profits of each PD-quartile. For the model development, similar process as with Model 1C was applied. Summaries of models 3A, 3B, 3C and 3D are presented in Appendix 10, 11, 12 and 13. The Model 3A is applied for the applicants who belong to the lowest PD quartile based on training data and Model 3B for the second quartile and so forth.

After the models were developed, Profit was predicted with test-data and validation data by combining predictions of the models developed for each risk class. Residuals were similar with the ones of Model 1B as can be seen from Figure 19.

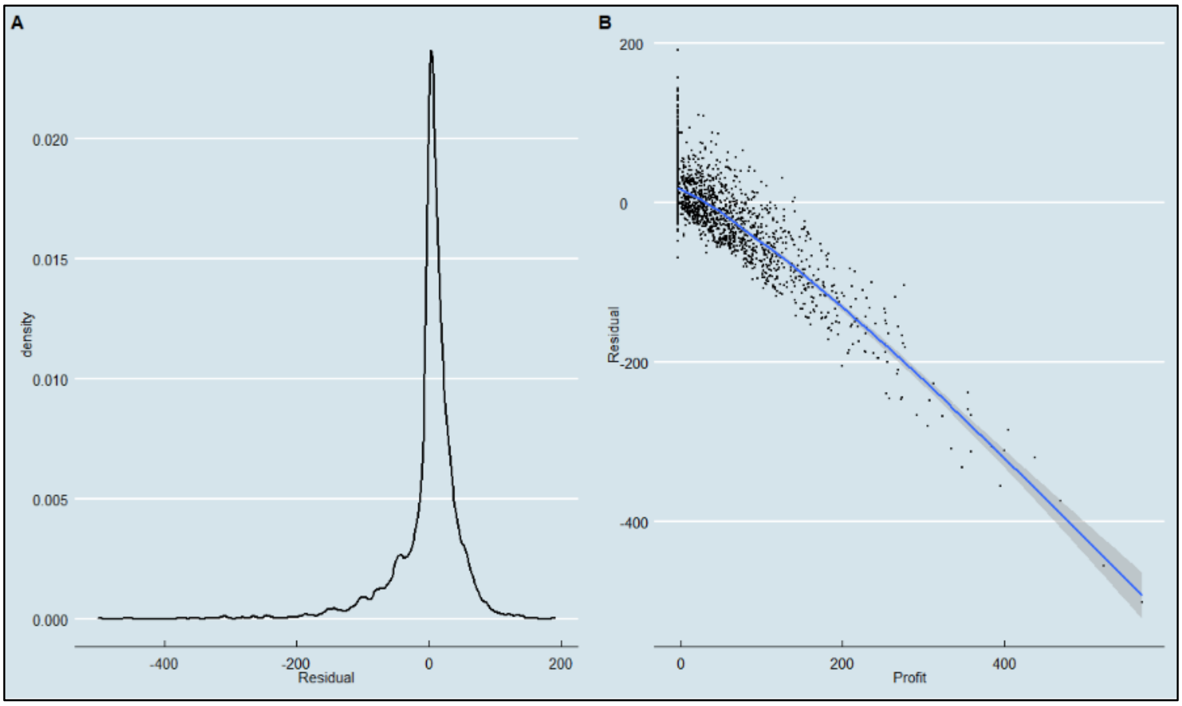

Figure 18 Residual Distribution of Combination Model (3A, 3B, 3C and 3D) with Test-Data (Plot A) and Residuals by Realized Profits (Plot B)

The overall prediction error with validation data measured by MAE was 29.27334, which is 0,5217 and around 1.75 % lower than with Model 1C. The difference in MAE is a bit higher with the test-data (29.78236 vs. 30.59135). One sample T-test on absolute prediction errors of Model 1C and the Combination Model returns a p-value below 2.2e-16. Thus, the superiority of Combination Model in terms of prediction accuracy seems to be statistically significant, and the null hypothesis is confirmed: *Using separate continuous profit-models for applications with different level of credit risk improves the prediction accuracy.*

5.6 Results

Regarding the research question one (*How to estimate profits of credit card accounts in order to conduct a scoring-model which ranks applications based on risk-adjusted returns?*) and the suggested theoretical framework for modelling profits, the predictive analytics proved that prediction accuracy of profit-model can be enhanced by including output of separate classification model for identifying passive credit accounts as a predictive variable for the profit-model and by developing separate models for applications with different level of credit risk. However, the method in which higher overall profits would be retained by declining applications based on high probability of ending up passive before providing a profit-score was found as fairly ineffective. Therefore, an adjustment is required for the theoretical framework developed based on the literature review. The updated framework is presented in Figure 19.

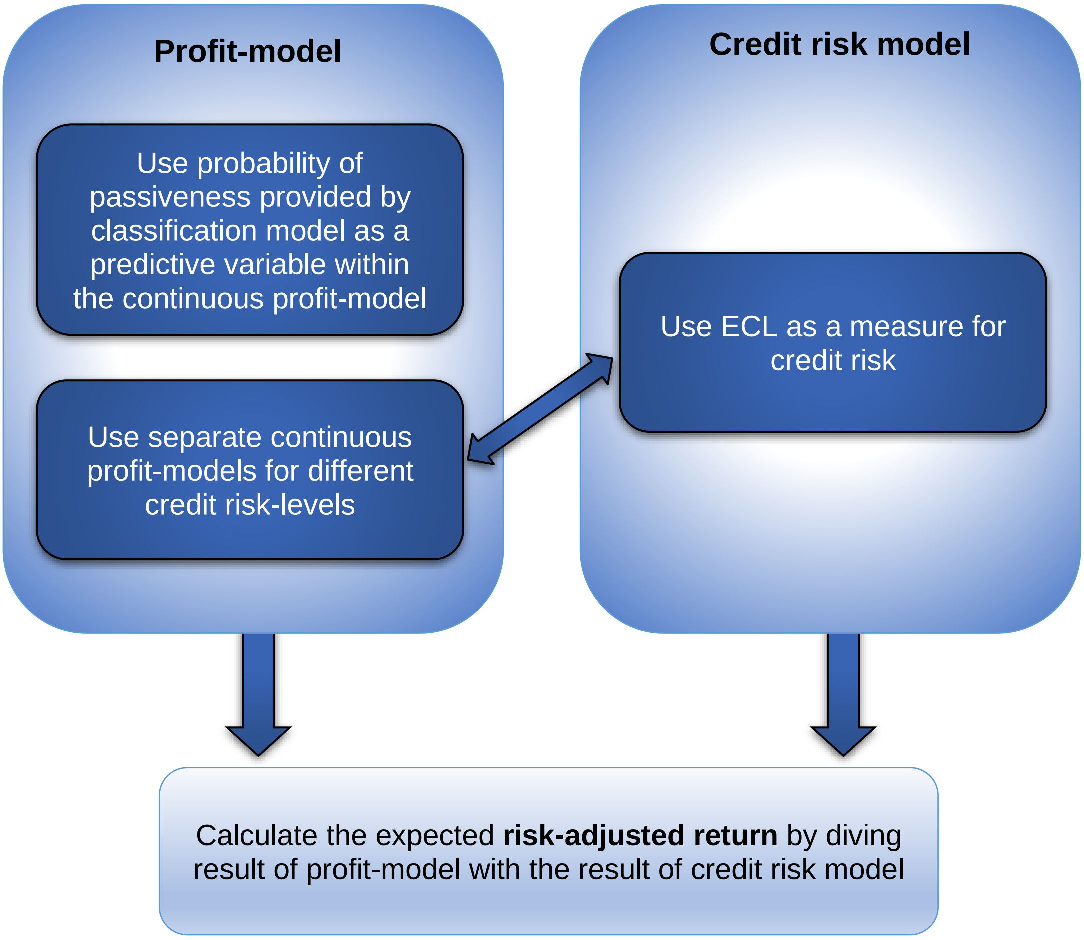

Figure 19 Updated Framework for Developing Risk-Adjusted Return Model

In the framework, it is suggested that ECL would be used as a proxy for the credit risk. However, the data available does not include actual ECL, and the study had to rely on a proxy of it, which is PD times the credit limit applied. I.e. it is expected that in case default takes place, the defaulted amount would be equal to the credit limit, and possible recoveries after the default are ignored (LGD = 1). The estimated risk-adjusted return used in the final scoring-model of the study is calculated as presented in Equation 9, and the realized risk-adjusted return with the formula of Equation 10.

$$Estimated\ risk\ adjusted\ return\ = \frac{Predicted\ profits}{PD \times Credit\ limit}$$

Equation 9 Formula for Estimated Risk-Adjusted Return

$$Realized\ risk\ adjusted\ return\ = \frac{Realized\ profits}{Amount\ of\ realized\ defaults}$$

Equation 10 Formula for Realized Risk-Adjusted Return

Tables 6 and 7 presents the comparison of risk-adjusted return (Table 6) and risk-based scoring-model (Table 7). In order to compare the alternative approaches, validation dataset was split into score-deciles in a way that the highest (tenth) score decile include applicants with highest estimated risk-adjusted return (risk-adjusted return-based model) or lowest PD (risk-based model).

Based on Table 6, the ranking power of the risk-adjusted return-based model seems to be somewhat limited. Although the highest score decile results by far the highest realized risk-adjusted return and net profit proxy (realized profits minus realized defaults), seventh decile results higher risk-adjusted return than eighth and ninth decile. Inconsistencies between the scores and realized risk-adjusted returns exist also within the lower score deciles. Lowest score decile seems to have the worst contribution for the overall net profit though. The negative profits of the lowest decile can be tracked to extremely high share (93.41 %) of completely passive credit card accounts. For comparison, the share of passive accounts in the highest score decile is 42.56 %. Yet it seems that it would be possible to improve the overall profitability of the product by identifying the applicants, which are likely to end up passive.

Table 6 Results of Risk-Adjusted Return-Based Scoring-Model on Validation Dataset

Score decile	Realized profits	Share of total profits	Realized Defaults	Share of total defaults	Realized Risk-adjusted return	Cumulative Risk-adjusted return	Net profit proxy (Profits-Defaults)	Share of net profit
10	42 092 €	16.9 %	4 306 €	4.6 %	9.78	9.78	37 786 €	24.5 %
9	37 172 €	15.0 %	11 890 €	12.6 %	3.13	4.89	25 283 €	16.4 %
8	37 165 €	15.0 %	17 468 €	18.5 %	2.13	3.46	19 697 €	12.8 %
7	39 284 €	15.8 %	10 118 €	10.7 %	3.88	3.56	29 165 €	18.9 %
6	36 398 €	14.6 %	20 590 €	21.8 %	1.77	2.98	15 808 €	10.3 %
5	31 299 €	12.6 %	17 849 €	18.9 %	1.75	2.72	13 450 €	8.7 %
4	18 771 €	7.6 %	8 085 €	8.6 %	2.32	2.68	10 686 €	6.9 %
3	6 100 €	2.5 %	2 890 €	3.1 %	2.11	2.66	3 210 €	2.1 %
2	2 143 €	0.9 %	1 301 €	1.4 %	1.65	2.65	842 €	0.5 %
1	-1 927 €	-0.8 %	- €	0.0 %	N/A	2.63	-1 927 €	-1.3 %

Table 7 Results of Risk-Based Scoring-Model on Validation Dataset

Score decile	Realized profits	Share of total profits	Realized Defaults	Share of total defaults	Realized Risk-adjusted return	Cumulative Risk-adjusted return	Net profit proxy (Profits-Defaults)	Share of net profit
10	91 €	0.0 %	- €	0.0 %	N/A	N/A	91 €	0.1 %
9	4 256 €	1.7 %	1 004 €	1.1 %	4.24	4.33	3 252 €	2.1 %
8	6 952 €	2.8 %	- €	0.0 %	N/A	11.25	6 952 €	4.5 %
7	12 200 €	4.9 %	1 682 €	1.8 %	7.25	8.75	10 518 €	6.8 %
6	17 524 €	7.1 %	8 171 €	8.6 %	2.14	3.78	9 353 €	6.1 %
5	28 543 €	11.5 %	4 899 €	5.2 %	5.83	4.42	23 645 €	15.4 %
4	33 257 €	13.4 %	8 993 €	9.5 %	3.70	4.15	24 264 €	15.8 %
3	40 282 €	16.2 %	16 083 €	17.0 %	2.50	3.50	24 199 €	15.7 %
2	48 719 €	19.6 %	21 488 €	22.7 %	2.27	3.08	27 231 €	17.7 %
1	56 673 €	22.8 %	32 180 €	34.1 %	1.76	2.63	24 493 €	15.9 %

Based on the realized defaults of Table 7, the risk-based scoring-model seems to be a bit more consistent in terms of ranking power. However, the highest score decile is barely profitable, and the highest net profit arise from second worst decile. Although, the highest score deciles of risk-based model result high risk-adjusted returns, when cumulative risk-adjusted returns are observed, the ranking power of risk-adjusted return-based model seems superior. In overall and regarding the second research question, it seems that the risk-adjusted return-based scoring-model performs better than the risk-based model used in the case company by all other metrics considered besides the realized defaults.

When the findings are analyzed from the case company point of view, it can be stated that usage of risk-based scoring approach leads to suboptimal profitability within the credit card portfolio compared to risk-adjusted return-based model since relatively higher risk applicants are often declined even though they would provide higher profit-risk ratio than many of approved. Table 8 includes similar information as Tables 6 and 7, but it has been extended by including all the application data used in the study (including declined applications and data used for model training and testing) and approval rate for each score decile. Content of the table illustrates somewhat well the problematics in usage of risk-based application scoring model in credit card business.

Table 8 Results of Risk-Adjusted Return-Based Scoring-Model on Whole Dataset

Score decile	Realized profits	Share of total profits	Realized Defaults	Share of total defaults	Realized Risk-adjusted return	Cumulative Risk-adjusted return	Net profit proxy (Profits-Defaults)	Share of net profit	Approval rate
10	186 706 €	22.9 %	47 902 €	14.4 %	3.90	3.90	138 804 €	28.7 %	83.2 %
9	153 525 €	18.8 %	46 020 €	13.8 %	3.34	3.62	107 505 €	22.2 %	74.6 %
8	134 545 €	16.5 %	47 717 €	14.3 %	2.82	3.35	86 828 €	18.0 %	65.4 %
7	110 080 €	13.5 %	47 269 €	14.2 %	2.33	3.10	62 812 €	13.0 %	54.8 %
6	79 559 €	9.7 %	44 687 €	13.4 %	1.78	2.84	34 872 €	7.2 %	46.9 %
5	64 153 €	7.9 %	35 169 €	10.6 %	1.82	2.71	28 984 €	6.0 %	39.2 %
4	46 368 €	5.7 %	28 277 €	8.5 %	1.64	2.61	18 091 €	3.7 %	33.8 %
3	32 267 €	4.0 %	22 954 €	6.9 %	1.41	2.52	9 313 €	1.9 %	59.4 %
2	14 502 €	1.8 %	10 394 €	3.1 %	1.40	2.49	4 108 €	0.8 %	86.7 %
1	-5 587 €	-0.7 %	2 317 €	0.7 %	-2.41	2.45	-7 904 €	-1.6 %	88.7 %

When all the applications are rescored with the risk-adjusted return-based model, lowest two score deciles have the highest approval rate but the worst net profits. I.e. minimization of credit risk seems to lead quite far from minimization of business risk. With the assumption that all the declined applications would have resulted similar profits and defaults as approved ones, simply by increasing the approval rate in the highest score decile by 10 %-units, the net profit proxy would increase by 16 677.29 Euros, which equals around 3.45 % in overall net profit. Although, there are likely to be justified reasons behind many credit decisions, which have led to decline of an application; based on the results examined, it is likely that the case company could improve its' profitability considerably without compromising good credit risk management by using the suggest method of scoring credit card applications based on the expected risk-adjusted return instead of probability of default.

6 SUMMARY AND CONCLUSION

The purpose of this study was to come up with a framework for developing credit card application scoring-model, which ranks applications based on the ratio of expected profits and credit losses, and to find evidence whether such a model could outperform a scoring-model, which ranks applications based on credit risk only. The results of literature review around advantages and problems of profit-based models were used for framework development. In the process, following hypotheses were set regarding profit-models of credit cards:

$H1_0$ = *Two-stage profit-model which includes separate classification model for identifying passive credit accounts achieves better prediction accuracy than single-stage model which ignores the classification of applications.*

$H2_0$ = *Using separate continuous profit-models for applications with different level of credit risk improves the prediction accuracy.*

Hypothesis one was first tested by developing two different profit-models: Model 1B and two-stage model (Model 2 and Model 1B) in which the applicants were first classified as passive and active, and the profits were then predicted with the continuous model only for those expected to be active. The ones expected to be passive, were assigned with profit prediction equal to a proxy of fixed costs of the credit card. **Although the two-stage model resulted slightly lower prediction error, the difference in predictions errors was statistically insignificant, and therefore the null hypothesis $\mathrm{H1}_0$ was rejected.** However, in a later phase of predictive analytics it was found that the output of developed classification model for identifying passive credit accounts improved the prediction accuracy of Model 1B significantly. Thus, it was included in the finalized framework as a step for profit-model development.

For testing hypothesis two, model training data was split to quartiles based on the PD related to the applications. Then, four different continuous profit-models were developed for each risk class (Model 3A, 3B, 3C and 3D). The predictions of these models were then combined and compared to prediction of Model 1C, which was developed during testing of hypothesis one. **The prediction error of combination model on validation data was 1.75 % percent lower, and the difference was found to be statistically significant. Consequently, the null hypothesis $\mathrm{H2}_0$ was confirmed.**

The framework for developing risk-adjusted return-based scoring-model was finalized based on the results of hypotheses testing. The finalized framework, which is presented in Figure 19 also answers the research question one:

How to estimate profits of credit card accounts in order to conduct a scoring-model which ranks applications based on risk-adjusted returns?

Answer for the research question one in briefly is that the probability of ending up as a passive credit card holder provided by classification model could be used as an independent variable in the continuous profit model to identify passive and low profiting credit accounts a bit better. And, instead of developing just one continuous model, separate models could be developed for applications with different level of credit risk in order to identify such predictors for the models which correlate with the profits but not with the credit risk.

When the framework was finalized, it was tested and compared to a PD model of the case company in order to provide answers for the research question two and its' sub-questions:

How does an application scoring-model which ranks applications based on the risk-adjusted returns perform compared to risk-based model?

a. *Which model ranks applications better in terms of realized profits?*
b. *Which model ranks applications better in terms of realized defaults?*
c. *Which model ranks applications better in terms of realized net profits?*
d. *Which model ranks applications better in terms of ratio between profits and amount of defaults?*

Outcomes of Tables 6, 7 and 8 provided evidence that the risk-adjusted return-based scoring-model is likely to rank applications better in terms of profits, net profits and ratio between profits and amount of defaults. These results underline the problematics around risk-based application scoring in credit card business: the minimization of credit risk might actually increase the business risks considerably as low level of credit risk does not signify high net profits nor risk-reward ratio.

Table 9 Answers for the Reasearch Questions

Research question	Sub-question	Answer
1) How to estimate profits of credit card accounts in order to conduct a scoring-model which ranks applications based on risk-adjusted returns?	-	Prefer using parallel models for applicants with different level of credit risk in order to identify predictors, which do not correlate too much with the risk and ultimately to identify applicants with high risk-reward ratio. And, use output of classification model predicting the probability of ending up as passive credit card holder as an independent variable in the profit model to improve predictive accuracy.
2) How does an application scoring-model which ranks applications based on the risk-adjusted returns perform compared to risk-based model?	a) Which model ranks applications better in terms of realized profits?	Risk-adjusted return-based scoring-model
	b) Which model ranks applications better in terms of realized defaults?	Risk-based scoring-model
	c) Which model ranks applications better in terms of realized net profits?	Risk-adjusted return-based scoring-model
	d) Which model ranks applications better in terms of ratio between profits and amount of defaults?	Risk-adjusted return-based scoring-model

The answers for the research questions are summarized in Table 9 above.

6.1 Discussion of Results

The existing literature around credit application scoring-models is rather divided between two approaches: risk- and profit-based ones. This study supplements the area of literature between these main approaches. The results of predictive analytics with the data of a case company provided promising evidence of possible performance of suggested risk-adjusted return-based modelling. This alternative method could be found suitable and useful especially by those issuers, which wish to consider expected profits in their credit decision processes, but who find traditional profit-models as inapplicable due to their credit risk management strategies for instance. By choosing risk-adjusted return-based approach, financial companies can manage their credit portfolios towards preferable ratio of profit and risk. This can be considered as valuable feature as the score cutoffs level can be naturally derived from credit risk management strategies. And, as the model set includes the traditional PD-parameter, the method enables to hold on with the PD-cutoff as simultaneously. Issuer using risk-adjusted return-based approach could set 1.75 as a minimum level for profit-risk ratio and 5 % as a maximum level for probability of default for instance. This is how the profits could be considered without ignoring firm credit risk management.

Although the results of developed framework seem promising, the suggested modelling method can be seen quite heavy to monitor and manage as it contains several parallel models: classification model for to be active card holders, multiple continuous profit-models for applicants with different risk-levels and required risk model(s), such as PD-model or ECL-models. For example, changes in pricing are likely to deteriorate the predictive power of profit model, and therefore the profit-model is likely to require continual retraining. The matter does not concern only profit-levels in general but also the differences between profits of certain customer groups. For example, increase of cash withdrawal fee is likely to effect more on certain type of customers than others. Hence, quite high application volumes would be required in order to cover the cost related to more complex modelling method with the marginal benefit of better prediction accuracy. Thus, it might be reasonable to simplify the suggested framework in case of smaller companies.

6.2 Limitations and validity

As the dataset used for the modelling included data of approved applications of single company, the data included some pre-selections, and therefore the applicants and applications used for the model training might not represent average credit card applicant information. The study relied also on proxies of fixed costs and ECL, and the tested framework might have resulted different type of results with more accurate values for these two parameters. What comes to the methodology used, the study relied on multiple linear regression and logistic regression. It is also probable that different type of results would have been acquired with different type of statistical techniques.

There are also some limitations related to time frames around the study. Firstly, the dataset included data from period less than a year. Longer time period might have resulted more inaccurate predictions and different type of results as profits might vary among similar applicants due to macroeconomic changes for instance. Secondly, the length of the observation period might affect the results as well. 12-month observation period is somewhat short if the model is supposed to provide information about the overall profit expectations. It’s also worth noting that the paper discussed the net profits only on level of a single product. Although, it was pointed that higher net profits could be achieved by identifying and declining passive and low profiting credit card applications, it could be the case that the due to the declined decision, the applicant does not apply or use some other product provided by the issuer. E.g. the product-level profit maximization does not guarantee customer-level nor overall profit maximization.

6.3 Further reasearch

As discussed, there are some limitations especially related to the way the profits and net profits are addressed within this study. Consequently, it would beneficial to study whether the suggested method could be extended to cover whole credit lifetime instead of the 12-month period. Also, the further studies could be extended to examine the scoring model performance from customer-level profitability point of view. After all, the issuers are likely to chase the high overall profitability instead of product- or account-level profits.

The results of this study provided further insight of problems around risk-based scoring approach, and the strong correlation of risk and profit. Although Table 7 for instance shows how applications with worst scores generate highest defaults, they also generate significantly higher profits and net profits compared to low risk applicants, which are ranked as best by the risk-based model. Scope of this study didn’t include models used for defining appropriate credit limit for the applicant, but it might be that risk-based models have some advantages when limit strategies are applied. Therefore, further research could be extended to examine the pros and cons of different modelling approaches from point of view of limit strategy management. For example, the defaults rising from high-risk customers could be limited with a well-managed limit strategy without losing the relative amount of profits; at least when fixed revenue streams such as monthly or yearly fees exist.

In overall, this study was no exception of many before as it proved how scoring-models can have significant effects on the profitability of creditor, and therefore further research on the suggested or similar credit application scoring approaches can have a massive contribution for the practices used by financial companies, and enhance their credit risk and profitability management.

REFERENCES

Abid L, Masmoudi A and Ghorbel SZ (2016). The Consumer Loans Payment Default Predictive Model: An Application in a Tunisian Commercial Bank. Asian Economic and Financial Review, vol. 6, no. 1, pp. 27-42.

Akkoç S (2012). An empirical comparison of conventional techniques, neural networks and the three stage hybrid Adaptive Neuro Fuzzy Inference System (ANFIS) model for credit scoring analysis: The case of Turkish credit card data. European Journal of Operational Research, Oct 1, vol. 222, no. 1, pp. 168-178.

Andreeva G (2006). European generic scoring models using survival analysis. Journal of the Operational Research Society, Oct 1, vol. 57, no. 10, pp. 1180-1187.

Andreeva G, Ansell J and Crook J. (2007). Modelling profitability using survival combination scores. European Journal of Operational Research, vol. 183, no. 3, pp. 1537-1549.

Banasik J, Crook J.N and Thomas L.C (1999). Not if but when will borrowers default. Journal of the Operational Research Society, Dec 1, vol. 50, no. 12, pp. 1185-1190.

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, Mar 1, vol. 65, no. 3, pp. 443-453.

Basel II (2005) International Convergence of Capital Measurement and Capital Standards. Basel committee on banking supervision.

Basel III (2017) Finalising post-crisis reforms. Basel committee on banking supervision.

Beling P, Covaliu Z and Oliver R.M (2005). Optimal scoring cutoff policies and efficient frontiers. Journal of the Operational Research Society, Sep 1, vol. 56, no. 9, pp. 1016-1029.

Bellotti T (2010). A simulation study of Basel II expected loss distributions for a portfolio of credit cards. Journal of Financial Services Marketing, Mar, vol. 14, no. 4, pp. 268-277.

Brooks C (2014) Introductory Econometrics for Finance. Third edition. New York. Cambridge University Press.

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, no. 3, pp. 1447-1465.

De Paula D.A.V, Artes R, Ayres F and Minardi A.M.A.F (2019). Estimating credit and profit scoring of a Brazilian credit union with logistic regression and machine-learning techniques. RAUSP Management Journal, Jul 8, vol. 54, no. 3, pp. 321-336.

Dong G, Lai K.K and Yen J (2010). Credit scorecard based on logistic regression with random coefficients. Procedia Computer Science, May, vol. 1, no. 1, pp. 2463-2468.

Doumpos M, Lemonakis C, Niklis D and Zopounidis C (2019). Analytical Techniques in the Assessment of Credit Risk. Cham: Springer International Publishing.

Durand D (1941) Risk Elements in Consumer Instatement Financing. New York. Bureau of Economic Research.

EBA/GL/2017/06 (2017) Guidelines on credit institutions' credit risk management practices and accounting for expected credit losses. European banking authority (EBA).

European central bank (2019) Payments statistics: methodological notes [www document]. [Accessed 31 October 2019]. Available http://sdw.ecb.europa.eu/servlet/desis?node=1000004051

2015/751 (2015) Regulation (EU) 2015/751 of the European parliament and of the council of 29 April 2015 on interchange fees for card-based payment transactions. European commission.

Fair Isaac Corporation (2014) Building Powerful, Predictive Scorecards - An overview of Scorecard module for FICO® Model Builder [www document]. [Accessed 21 March 2020]. Available https://www.fico.com/en/resource-download-file/3477

Finlay S.M (2008). Towards profitability: a utility approach to the credit scoring problem. Journal of the Operational Research Society, Jul 1, vol. 59, no. 7, pp. 921-931.

Finlay S.M (2010). Credit scoring for profitability objectives. European Journal of Operational Research, vol. 202, no. 2, pp. 528-537.

Forsyth D (2019). Applied Machine Learning. Cham: Springer International Publishing ISBN 3030181138.

Hand D.J and Kelly M.G (2001). Lookahead scorecards for new fixed term credit products. Journal of the Operational Research Society, Sep 1, vol. 52, no. 9, pp. 989-996.

Hand D.J (2001). Modelling consumer credit risk. IMA Journal of Management Mathematics, Oct 1, vol. 12, no. 2, pp. 139-155.

IFRS 9 (2014) IFRS Foundation.

IFRS Foundation (2019) Who uses IFRS Standards? [www document]. [Accessed 4 November 2019]. Available https://www.ifrs.org/use-around-the-world/use-of-ifrs-standards-by-jurisdiction/

Joubert M, Verster T and Raubenheimer H (2019). Making use of survival analysis to indirectly model loss given default. ORiON, Jan 14, vol. 34, no. 2, pp. 107-132.

Leow M and Crook J (2016). A new Mixture model for the estimation of credit card Exposure at Default. European Journal of Operational Research, Mar 1, vol. 249, no. 2, pp. 487-497.

Marqués A.I, García V and Sánchez J.S (2013). A literature review on the application of evolutionary computing to credit scoring. Journal of the Operational Research Society, Sep 1, vol. 64, no. 9, pp. 1384-1399.

Nie G, Rowe W, Zhang L, Tian Y and Shi Y (2011). Credit card churn forecasting by logistic regression and decision tree. Expert Systems with Applications, vol. 38, no. 12, pp. 15273-15285.

Olive D.J (2017). Linear regression. Cham: Springer.

Oliver R.M and Wells E (2001). Efficient frontier cutoff policies in credit portfolios. Journal of the Operational Research Society, Sep 1, vol. 52, no. 9, pp. 1025-1033.

Qi M and Zhao X (2011). Comparison of modeling methods for Loss Given Default. Journal of Banking and Finance, vol. 35, no. 11, pp. 2842-2855.

Sanchez-Barrios L.J, Andreeva G, Ansell J (2016) Time-to-profit scorecards for revolving credit. European Journal of Operational Research, vol. 249, p. 397–406.

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, Sep, vol. 89, pp. 113-122.

So M.C, Thomas L.C, Seow H and Mues C (2014). Using a transactor/revolver scorecard to make credit and pricing decisions. Decision Support Systems, Mar, vol. 59, pp. 143-151.

So M.C. and Thomas L.C (2011). Modelling the profitability of credit cards by Markov decision processes. European Journal of Operational Research, vol. 212, no. 1, pp. 123-130.

Srinivasan V (1987). Credit granting. The Journal of Finance, vol. 42, no. 3, pp. 665-681.

Stepanova M and Thomas L.C (2002). Survival Analysis Methods for Personal Loan Data. Operations Research, Mar 1, vol. 50, no. 2, pp. 277-289.

Stewart R.T (2011). A profit-based scoring system in consumer credit: making acquisition decisions for credit cards. Journal of the Operational Research Society, Sep 1, vol. 62, no. 9, pp. 1719-1725.

Thomas L.C, Edelman D.B and Crook I.N (2002). Credit Scoring and Its Applications. SIAM: Philadelphia, USA.

Thomas L.C, Oliver R.W and Hand D.J (2005). A survey of the issues in consumer credit modelling research. Journal of the Operational Research Society, Sep 1, vol. 56, no. 9, pp. 1006-1015.

Tong E.N.C, Mues C, Brown I and Thomas L.C (2016). Exposure at default models with and without the credit conversion factor. European Journal of Operational Research, Aug 1, vol. 252, no. 3, pp. 910-920.

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, Oct 16, vol. 238, no. 2, pp. 505-513.

Witzany J (2017). Credit Risk Management : Pricing, Measurement, and Modeling. Cham: Springer International Publishing AG.

APPENDICES

Appendix 1 - Data Descriptives Before Pre-Processing

variable	n	mean	sd	median	min	max	skew	kurtosis	se
A	52776	25411.949	14495.864	25646.500	1.000	50274.000	-0.033	-1.198	63.099
B	52776	26388.500	15235.263	26388.500	1.000	52776.000	0.000	-1.200	66.318
A1	52776	0.633	0.482	1.000	0.000	1.000	-0.551	-1.697	0.002
D*	52776	104.529	60.957	101.000	1.000	211.000	0.083	-1.189	0.265
E	52776	820.325	945.560	626.560	0.000	7832.000	1.927	6.155	4.116
H	52776	5.399	22.372	0.000	0.000	1022.970	9.518	188.331	0.097
F	52776	5.846	17.545	0.000	0.000	319.440	4.888	36.336	0.076
G	52776	14.378	34.836	0.000	0.000	1042.000	4.993	51.779	0.152
CF	52776	415.661	1171.847	0.000	0.000	35279.190	6.342	71.911	5.101
CFA	52776	1.496	4.219	0.000	0.000	127.010	6.342	71.914	0.018
Profit	52775	15.464	45.803	0.000	-3.84	747.560	3.976	23.246	0.199
Passive	52776	0.775	0.418	1.000	0.000	1.000	-1.316	-0.268	0.002
I	52776	752.990	28.418	754.222	0.000	798.864	-14.831	389.675	0.124
II	52776	0.031	0.047	0.012	0.000	0.783	4.046	33.934	0.000
J	52776	14.902	82.441	0.000	0.000	764.400	5.613	30.793	0.359
K	52776	0.025	0.157	0.000	0.000	1.000	6.041	34.498	0.001
L	52776	0.473	0.499	0.000	0.000	1.000	0.106	-1.989	0.002
M	52776	0.046	0.209	0.000	0.000	1.000	4.338	16.820	0.001
N	52776	2.529	1.805	1.000	1.000	7.000	0.682	-1.010	0.008
O	52776	0.944	0.230	1.000	0.000	1.000	-3.855	12.859	0.001
Q	52776	1386.362	1135.414	783.200	626.560	7832.000	2.809	10.457	4.942
R	52776	0.637	0.916	0.000	0.000	2.000	0.778	-1.355	0.004
S	52776	0.637	0.916	0.000	0.000	2.000	0.778	-1.355	0.004
T	52776	2231.142	3455.637	1879.680	0.000	391600.000	43.691	3766.230	15.042
U	52776	156.450	811.623	0.000	0.000	78320.000	37.039	2400.598	3.533
TU	52776	2387.592	3654.302	1958.000	0.000	391600.000	39.899	3161.885	15.907
V	52776	2.101	1.155	2.000	1.000	5.000	0.758	-0.426	0.005
W	52776	1.982	1.155	1.000	1.000	5.000	0.654	-0.674	0.005
X	52776	19587.810	39759.420	0.000	0.000	1566400.000	5.196	85.801	173.070
Y	52776	128.532	247.619	0.000	0.000	13314.400	8.833	316.513	1.078
Z	52776	4604.394	19706.590	0.000	0.000	2271280.000	40.381	3613.347	85.781
AA	52776	92.584	884.538	0.000	0.000	195800.000	205.620	45408.999	3.850
AAA	52776	221.116	923.283	78.320	0.000	195800.000	180.688	38155.150	4.019
AAAA	52776	2166.476	3699.706	1762.200	-186401.600	390914.700	35.731	3125.421	16.105
ZAAA	52776	2.219	2.704	1.833	0.000	250.000	26.849	1666.965	0.012
BB	52776	0.638	0.566	1.000	0.000	2.000	0.182	-0.754	0.002
DD	52776	35.168	12.734	34.000	14.000	73.000	0.319	-0.957	0.055

EE	52776	1007.662	6424.801	33.932	-970.541	465352.652	23.939	998.288	27.967
FF	52775	2.404	1.526	2.000	1.000	5.000	0.647	-1.149	0.007
GG	52775	0.118	0.322	0.000	0.000	1.000	2.373	3.633	0.001
HH	52776	0.203	0.402	0.000	0.000	1.000	1.474	0.173	0.002
III	52776	8.311	6.351	7.000	0.000	20.000	0.306	-1.358	0.028
JJ	52776	0.355	0.479	0.000	0.000	1.000	0.605	-1.634	0.002
LL	52776	0.019	0.136	0.000	0.000	1.000	7.049	47.688	0.001
MM	52776	0.020	0.141	0.000	0.000	1.000	6.824	44.574	0.001
NN	52776	0.883	0.322	1.000	0.000	1.000	-2.379	3.659	0.001
OO	52776	5.121	6.816	2.350	0.000	83.019	2.711	10.084	0.030
Default	52776	0.005	0.072	0.000	0.000	1.000	13.669	184.840	0.000
Default amount	52776	20.369	202.569	0.000	0.000	8082.726	14.855	311.559	0.882

Appendix 2 – Data Descriptives After Pre-Processing

variable	n	mean	sd	median	min	max	skew	kurtosis	se
A	52774	25412.500	14495.750	25647.500	1.000	50274.000	-0.033	-1.198	63.100
B	52774	26389.283	15235.007	26389.500	1.000	52776.000	0.000	-1.200	66.318
A1	52774	0.633	0.482	1.000	0.000	1.000	-0.551	-1.696	0.002
D*	52774	104.526	60.955	101.000	1.000	211.000	0.083	-1.189	0.265
E	52774	820.342	945.571	626.560	0.000	7832.000	1.927	6.155	4.116
H	52774	5.399	22.372	0.000	0.000	1022.970	9.518	188.324	0.097
F	52774	5.846	17.545	0.000	0.000	319.440	4.888	36.334	0.076
G	52774	14.378	34.836	0.000	0.000	1042.000	4.993	51.777	0.152
CF	52774	415.677	1171.867	0.000	0.000	35279.190	6.342	71.908	5.101
CFA	52774	1.496	4.219	0.000	0.000	127.010	6.342	71.912	0.018
Profit	52774	15.464	45.803	0.000	-3.84	747.560	3.976	23.245	0.199
Passive	52774	0.775	0.418	1.000	0.000	1.000	-1.316	-0.268	0.002
I	52774	752.990	28.419	754.222	0.000	798.864	-14.831	389.686	0.124
II	52774	0.031	0.047	0.012	0.000	0.783	4.046	33.939	0.000
J	52774	14.903	82.443	0.000	0.000	764.400	5.613	30.792	0.359
K	52774	0.025	0.157	0.000	0.000	1.000	6.041	34.497	0.001
L	52774	0.473	0.499	0.000	0.000	1.000	0.106	-1.989	0.002
M	52774	0.046	0.209	0.000	0.000	1.000	4.338	16.819	0.001
N1	52774	0.509	0.500	1.000	0.000	1.000	-0.036	-1.999	0.002
N2	52774	0.072	0.258	0.000	0.000	1.000	3.313	8.976	0.001
N3	52774	0.102	0.302	0.000	0.000	1.000	2.633	4.933	0.001
N4	52774	0.073	0.260	0.000	0.000	1.000	3.286	8.800	0.001
N5	52774	0.206	0.404	0.000	0.000	1.000	1.454	0.114	0.002
N6	52774	0.020	0.139	0.000	0.000	1.000	6.897	45.568	0.001
N7	52774	0.019	0.135	0.000	0.000	1.000	7.113	48.595	0.001
O	52774	0.944	0.230	1.000	0.000	1.000	-3.855	12.858	0.001
P	52774	408.963	236.821	408.000	0.000	1174.800	1.503	2.933	1.031
Q	52774	1386.377	1135.432	783.200	626.560	7832.000	2.808	10.456	4.943
QO	52774	0.535	0.499	1.000	0.000	1.000	-0.141	-1.980	0.002
R3	52774	0.532	0.499	1.000	0.000	1.000	-0.130	-1.983	0.002
R5	52774	0.170	0.376	0.000	0.000	1.000	1.758	1.089	0.002
R10	52774	0.196	0.397	0.000	0.000	1.000	1.535	0.355	0.002
R100	52774	0.102	0.303	0.000	0.000	1.000	2.631	4.921	0.001
S0	52774	0.667	0.471	1.000	0.000	1.000	-0.708	-1.499	0.002
S1	52774	0.029	0.168	0.000	0.000	1.000	5.608	29.454	0.001
S2	52774	0.304	0.460	0.000	0.000	1.000	0.851	-1.275	0.002
T	52774	2231.152	3455.702	1879.680	0.000	391600.000	43.691	3766.089	15.043
U	52774	156.448	811.638	0.000	0.000	78320.000	37.038	2400.519	3.533
TU	52774	2387.601	3654.371	1958.000	0.000	391600.000	39.899	3161.768	15.908
V1	52774	0.411	0.492	0.000	0.000	1.000	0.361	-1.870	0.002
V2	52774	0.251	0.434	0.000	0.000	1.000	1.147	-0.685	0.002
V3	52774	0.198	0.398	0.000	0.000	1.000	1.518	0.304	0.002
V4	52774	0.105	0.306	0.000	0.000	1.000	2.585	4.682	0.001
V5	52774	0.035	0.184	0.000	0.000	1.000	5.042	23.424	0.001
W	52774	1.982	1.155	1.000	1.000	5.000	0.654	-0.674	0.005
WA	52774	0.551	0.497	1.000	0.000	1.000	-0.207	-1.957	0.002
WB	52774	0.005	0.072	0.000	0.000	1.000	13.719	186.208	0.000
WC	52774	0.388	0.487	0.000	0.000	1.000	0.461	-1.787	0.002
WD	52774	0.021	0.144	0.000	0.000	1.000	6.634	42.014	0.001
WO	52774	0.034	0.182	0.000	0.000	1.000	5.108	24.095	0.001

X	52774	19588.553	39759.991	0.000	0.000	1566400.000	5.196	85.799	173.076
Y	52774	128.537	247.623	0.000	0.000	13314.400	8.833	316.507	1.078
Z	52774	4604.553	19706.947	0.000	0.000	2271280.000	40.380	3613.221	85.785
AA	52774	92.585	884.555	0.000	0.000	195800.000	205.617	45407.297	3.850
AAA	52774	221.122	923.300	78.320	0.000	195800.000	180.685	38153.800	4.019
AAAA	52774	2166.478	3699.775	1762.200	- 186401.600	390914.700	35.730	3125.306	16.105
ZAAA	52774	2.219	2.704	1.833	0.000	250.000	26.848	1666.914	0.012
BB0	52774	0.407	0.491	0.000	0.000	1.000	0.381	-1.855	0.002
BB1	52774	0.549	0.498	1.000	0.000	1.000	-0.197	-1.961	0.002
BB2	52774	0.045	0.206	0.000	0.000	1.000	4.417	17.512	0.001
DD	52774	27.544	9.974	26.629	10.965	57.174	0.319	-0.957	0.043
EE	52774	1007.700	6424.920	33.936	-970.541	465352.652	23.939	998.251	27.968
EE1	52774	0.010	0.100	0.000	0.000	1.000	9.847	94.957	0.000
EE2	52774	1007.944	6424.876	33.936	0.000	465352.652	23.939	998.275	27.968
EEO	52774	0.240	0.427	0.000	0.000	1.000	1.219	-0.513	0.002
FF1	52774	0.413	0.492	0.000	0.000	1.000	0.353	-1.875	0.002
FF2	52774	0.234	0.423	0.000	0.000	1.000	1.256	-0.423	0.002
FF3	52774	0.055	0.228	0.000	0.000	1.000	3.893	13.156	0.001
FF4	52774	0.131	0.337	0.000	0.000	1.000	2.186	2.779	0.001
FF5	52774	0.167	0.373	0.000	0.000	1.000	1.790	1.205	0.002
GG	52774	0.118	0.322	0.000	0.000	1.000	2.373	3.633	0.001
HH	52774	0.203	0.402	0.000	0.000	1.000	1.474	0.173	0.002
II0	52774	0.016	0.126	0.000	0.000	1.000	7.673	56.883	0.001
II1	52774	0.140	0.347	0.000	0.000	1.000	2.074	2.301	0.002
II2	52774	0.163	0.369	0.000	0.000	1.000	1.825	1.330	0.002
II3	52774	0.054	0.226	0.000	0.000	1.000	3.945	13.567	0.001
II4	52774	0.047	0.211	0.000	0.000	1.000	4.286	16.371	0.001
II5	52774	0.029	0.169	0.000	0.000	1.000	5.573	29.055	0.001
II6	52774	0.030	0.169	0.000	0.000	1.000	5.561	28.924	0.001
II7	52774	0.021	0.144	0.000	0.000	1.000	6.666	42.436	0.001
II8	52774	0.042	0.200	0.000	0.000	1.000	4.586	19.031	0.001
II9	52774	0.008	0.088	0.000	0.000	1.000	11.129	121.864	0.000
II10	52774	0.028	0.165	0.000	0.000	1.000	5.732	30.855	0.001
II11	52774	0.003	0.058	0.000	0.000	1.000	16.987	286.562	0.000
II12	52774	0.107	0.309	0.000	0.000	1.000	2.545	4.476	0.001
II13	52774	0.092	0.289	0.000	0.000	1.000	2.826	5.989	0.001
II14	52774	0.011	0.106	0.000	0.000	1.000	9.226	83.112	0.000
II15	52774	0.029	0.168	0.000	0.000	1.000	5.600	29.364	0.001
II16	52774	0.022	0.148	0.000	0.000	1.000	6.476	39.935	0.001
II17	52774	0.058	0.233	0.000	0.000	1.000	3.794	12.398	0.001
II18	52774	0.034	0.180	0.000	0.000	1.000	5.169	24.715	0.001
II19	52774	0.036	0.186	0.000	0.000	1.000	4.997	22.974	0.001
II20	52774	0.030	0.172	0.000	0.000	1.000	5.477	27.993	0.001
JJ	52774	0.355	0.479	0.000	0.000	1.000	0.605	-1.634	0.002
KK	52774	7.276	3.278	8.615	1.566	13.314	-0.574	-1.307	0.014
LL	52774	0.019	0.136	0.000	0.000	1.000	7.049	47.686	0.001
MM	52774	0.020	0.141	0.000	0.000	1.000	6.824	44.572	0.001
NN	52774	0.883	0.322	1.000	0.000	1.000	-2.379	3.660	0.001
OO	52774	5.121	6.816	2.350	0.000	83.019	2.711	10.084	0.030
OOO	52774	0.080	0.271	0.000	0.000	1.000	3.101	7.616	0.001
Default	52774	0.005	0.072	0.000	0.000	1.000	13.669	184.833	0.000
Default amount	52774	20.369	202.573	0.000	0.000	8082.726	14.854	311.547	0.882

Appendix 3 – Summary of Base Model Statistics

```
Residuals:
    Min      1Q  Median      3Q     Max 
-162.49  -25.69  -10.09   10.33  597.47 

Coefficients: (11 not defined because of singularities)
              Estimate Std. Error t value Pr(>|t|)    
(Intercept)  8.796e+00  1.009e+01   0.872 0.383338    
J            1.065e-03  4.236e-03   0.251 0.801465    
K           -4.781e+00  1.979e+00  -2.416 0.015722 *  
L            3.774e+00  7.519e-01   5.019 5.25e-07 ***
M            5.474e+00  1.547e+00   3.539 0.000402 ***
N1           1.318e+01  8.440e+00   1.562 0.118340    
N2           8.126e+00  8.503e+00   0.956 0.339253    
N3           7.849e+00  8.490e+00   0.925 0.355193    
N4           1.072e+01  8.554e+00   1.253 0.210110    
N5           9.154e+00  8.458e+00   1.082 0.279160    
N6          -4.200e+00  9.516e+00  -0.441 0.658937    
N7                  NA         NA      NA       NA    
O           -7.592e+00  2.512e+00  -3.022 0.002510 ** 
P            1.187e-02  1.267e-03   9.372  < 2e-16 ***
Q            6.455e-03  4.584e-04  14.082  < 2e-16 ***
R3           1.582e+01  1.146e+00  13.804  < 2e-16 ***
R5           1.822e+01  1.353e+00  13.467  < 2e-16 ***
R10          8.861e+00  1.237e+00   7.162 8.24e-13 ***
R100                NA         NA      NA       NA    
S0          -1.224e+01  8.915e-01 -13.732  < 2e-16 ***
S1           1.732e+00  2.474e+00   0.700 0.483799    
S2                  NA         NA      NA       NA    
T            4.986e-04  2.316e-04   2.153 0.031333 *  
U           -1.256e-03  6.076e-04  -2.068 0.038663 *  
TU                  NA         NA      NA       NA    
V1           9.233e-01  1.910e+00   0.483 0.628892    
V2           1.011e+00  2.073e+00   0.488 0.625752    
V3           1.244e+00  2.088e+00   0.596 0.551444    
V4           2.012e+00  2.124e+00   0.947 0.343503    
V5                  NA         NA      NA       NA    
WA          -9.085e+00  2.374e+00  -3.826 0.000131 ***
WB           1.018e+01  5.773e+00   1.763 0.077967 .  
WC           2.228e+00  2.362e+00   0.944 0.345388    
WD           1.391e+00  3.222e+00   0.432 0.665943    
WO                  NA         NA      NA       NA    
X            5.114e-05  1.215e-05   4.208 2.58e-05 ***
Y            1.704e-03  1.895e-03   0.899 0.368685    
Z            1.274e-04  2.203e-05   5.785 7.37e-09 ***
AA           8.922e-05  2.579e-04   0.346 0.729335    
AAA                 NA         NA      NA       NA    
AAAA                NA         NA      NA       NA    
ZAAA        -7.115e-01  2.517e-01  -2.827 0.004698 ** 
BB0          1.139e+00  1.724e+00   0.661 0.508782    
BB1         -6.379e+00  1.578e+00  -4.043 5.30e-05 ***
BB2                 NA         NA      NA       NA    
DD          -1.035e-01  6.596e-02  -1.568 0.116785    
EE           1.953e-01  5.600e-02   3.487 0.000490 ***
EE1          3.665e+01  5.318e+00   6.892 5.68e-12 ***
EE2         -1.959e-01  5.600e-02  -3.498 0.000470 ***
EE0          1.210e+00  1.309e+00   0.925 0.355198    
FF1          7.761e+00  1.757e+00   4.418 1.00e-05 ***
FF2         -3.186e+00  1.703e+00  -1.871 0.061418 .  
FF3          5.634e+01  2.403e+00  23.448  < 2e-16 ***
FF4          1.409e+01  2.295e+00   6.141 8.38e-10 ***
FF5                 NA         NA      NA       NA    
GG           1.575e+00  1.683e+00   0.936 0.349462    
HH           1.729e+01  1.232e+00  14.031  < 2e-16 ***
II0          3.130e+00  4.525e+00   0.692 0.489079    
II1          2.623e-01  2.063e+00   0.127 0.898838    
II2          5.538e-01  2.155e+00   0.257 0.797222    
II3          5.623e+00  2.416e+00   2.328 0.019941 *  
II4         -2.835e+00  2.480e+00  -1.143 0.252955    
II5          3.457e+00  2.770e+00   1.248 0.212037    
II6         -2.549e+00  2.754e+00  -0.926 0.354663    
II7         -3.744e+00  3.026e+00  -1.237 0.216109    
II8         -2.775e+00  2.593e+00  -1.070 0.284706    
II9         -3.756e+00  4.417e+00  -0.850 0.395121    
II10        -3.943e+00  2.870e+00  -1.374 0.169547    
II11        -1.165e+01  6.459e+00  -1.804 0.071247 .  
II12        -3.931e+00  2.146e+00  -1.832 0.067017 .  
II13         7.252e-01  2.185e+00   0.332 0.739927    
II14         1.770e+00  3.846e+00   0.460 0.645315    
II15         2.505e+00  2.713e+00   0.923 0.355909    
II16        -1.979e+00  3.023e+00  -0.654 0.512807    
II17        -1.575e+00  2.371e+00  -0.664 0.506462    
II18        -5.773e+00  2.594e+00  -2.225 0.026095 *  
II19         1.808e+00  2.587e+00   0.699 0.484570    
II20                NA         NA      NA       NA    
JJ           2.551e+00  7.493e-01   3.404 0.000665 ***
KK          -5.055e-01  1.854e-01  -2.726 0.006410 ** 
LL          -8.106e+00  2.423e+00  -3.345 0.000824 ***
MM           6.297e+00  2.744e+00   2.295 0.021768 *  
NN           3.030e+00  2.116e+00   1.432 0.152119    
OO          -8.145e-01  8.056e-02 -10.111  < 2e-16 ***
OOO         -9.261e+00  1.578e+00  -5.869 4.45e-09 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 49.4 on 19963 degrees of freedom
Multiple R-squared:  0.2171,     Adjusted R-squared:  0.2143 
F-statistic: 75.84 on 73 and 19963 DF,  p-value: < 2.2e-16
```

Appendix 4 – Summary of Model 1 Statistics

```
Residuals:
    Min      1Q  Median      3Q     Max 
-183.98  -25.65   -8.32   12.44  593.99 

Coefficients:
                 Estimate Std. Error t value Pr(>|t|)    
(Intercept)     6.950e+01  3.736e+00  18.603  < 2e-16 ***
K              -3.858e+00  1.913e+00  -2.016 0.043777 *  
L               3.380e+00  7.264e-01   4.652 3.31e-06 ***
M               3.408e+00  1.493e+00   2.282 0.022508 *  
N2             -4.248e+00  1.202e+00  -3.534 0.000410 ***
O              -1.165e+01  2.397e+00  -4.861 1.18e-06 ***
N3             -4.356e+00  1.074e+00  -4.054 5.06e-05 ***
P               1.273e-02  1.233e-03  10.323  < 2e-16 ***
R3              1.090e+01  1.118e+00   9.752  < 2e-16 ***
R5              1.343e+01  1.322e+00  10.164  < 2e-16 ***
R10             5.530e+00  1.206e+00   4.584 4.59e-06 ***
S0             -9.664e+00  8.403e-01 -11.501  < 2e-16 ***
I(T^0.5)        2.574e-01  5.016e-02   5.132 2.90e-07 ***
I(U^0.1)        1.936e+00  4.875e-01   3.970 7.21e-05 ***
Q               4.631e-03  6.380e-04   7.259 4.05e-13 ***
WA             -1.179e+01  9.079e-01 -12.990  < 2e-16 ***
I(log(X + 1))   5.979e-01  7.841e-02   7.625 2.54e-14 ***
I(log(AA + 1))  2.810e+00  1.466e-01  19.174  < 2e-16 ***
EE1            -2.468e+01  5.306e+00  -4.652 3.31e-06 ***
EE2            -8.420e-05  5.369e-05  -1.568 0.116849    
I(EE2^0.1)     -3.353e+01  1.285e+00 -26.089  < 2e-16 ***
EE0            -5.062e+01  2.207e+00 -22.938  < 2e-16 ***
FF2            -6.520e+00  8.337e-01  -7.821 5.51e-15 ***
FF3             4.438e+01  1.758e+00  25.240  < 2e-16 ***
FF4             4.265e+00  1.533e+00   2.782 0.005405 ** 
HH              1.278e+01  1.189e+00  10.751  < 2e-16 ***
JJ              2.114e+00  7.213e-01   2.931 0.003379 ** 
LL             -6.438e+00  2.339e+00  -2.752 0.005923 ** 
MM              4.554e+00  2.500e+00   1.822 0.068518 .  
I(OO^0.5)       1.936e+00  5.202e-01   3.722 0.000198 ***
OOO            -6.804e+00  1.553e+00  -4.380 1.19e-05 ***
I(ZAAA^0.5)    -9.006e+00  1.572e+00  -5.729 1.02e-08 ***
BB0             8.984e+00  8.527e-01  10.536  < 2e-16 ***
II12           -3.592e+00  1.071e+00  -3.354 0.000797 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 48.11 on 20003 degrees of freedom
Multiple R-squared:  0.256,	Adjusted R-squared:  0.2548 
F-statistic: 208.6 on 33 and 20003 DF,  p-value: < 2.2e-16
```

Appendix 5 – Summary of Model 1B Statistics

```
Residuals:
    Min      1Q  Median      3Q     Max
-172.07  -25.22   -8.03   12.45  599.03

Coefficients:
                 Estimate Std. Error t value Pr(>|t|)
(Intercept)     56.387647   5.903415   9.552  < 2e-16 ***
K               -4.095514   1.910032  -2.144 0.032028 *
L                3.158354   0.732225   4.313 1.62e-05 ***
M                3.052865   1.497585   2.039 0.041511 *
N2              -4.487372   1.202367  -3.732 0.000190 ***
O               -9.778590   2.411045  -4.056 5.02e-05 ***
N3              -4.881335   1.083462  -4.505 6.67e-06 ***
P                0.023134   0.007642   3.027 0.002470 **
P2 0-250         3.999108   4.203350   0.951 0.341408
P2 250-500       1.333797   4.986362   0.267 0.789095
P2 500-1000     -4.740270   7.134499  -0.664 0.506432
P2 >1000        -6.004915   9.881563  -0.608 0.543401
R3              10.712639   1.118748   9.576  < 2e-16 ***
R5              13.331401   1.320990  10.092  < 2e-16 ***
R10              5.582283   1.204325   4.635 3.59e-06 ***
S0              -9.500663   0.838240 -11.334  < 2e-16 ***
I(T^0.5)         0.138475   0.058701   2.359 0.018333 *
Y2 0-250        -0.220021   3.553370  -0.062 0.950628
Y2 250-500      -3.415800   3.763931  -0.908 0.364149
Y2 500-1000     -0.530181   4.053315  -0.131 0.895933
Y2 >1000        -1.445834   5.466174  -0.265 0.791393
U2 0-100         1.281566   2.053057   0.624 0.532488
U2 100-400       4.449539   1.132577   3.929 8.57e-05 ***
U2 400-1000      1.949661   1.573295   1.239 0.215278
U2 >1000        -1.176813   2.247886  -0.524 0.600618
Q2 800-1600      5.379480   1.314640   4.092 4.29e-05 ***
Q2 1600-4000    15.143333   2.108112   7.183 7.04e-13 ***
Q2 4000-7000    49.372968   4.829226  10.224  < 2e-16 ***
Q2 >7000        35.926993   5.388140   6.668 2.66e-11 ***
WA             -11.116808   0.989548 -11.234  < 2e-16 ***
I(log(X + 1))    0.680476   0.351591   1.935 0.052953 .
I(log(AA + 1))   3.436611   1.118352   3.073 0.002123 **
AA2 0-200       -3.880527   5.177718  -0.749 0.453585
AA2 200-500     -3.326543   6.451209  -0.516 0.606107
AA2 500-1000     2.829623   7.603776   0.372 0.709797
AA2 >1000      -16.758741   9.496485  -1.765 0.077624 .
EE1            -25.896613   5.234926  -4.947 7.60e-07 ***
I(EE2^0.1)     -33.198048   1.159655 -28.628  < 2e-16 ***
EE0            -50.788245   2.175673 -23.344  < 2e-16 ***
FF2             -5.956356   0.859268  -6.932 4.28e-12 ***
FF3             43.760647   1.705832  25.654  < 2e-16 ***
FF4              4.570548   1.459418   3.132 0.001740 **
HH              12.415762   1.193609  10.402  < 2e-16 ***
JJ               2.078968   0.720576   2.885 0.003916 **
KK2 2-4         -0.096203   1.698651  -0.057 0.954836
KK2 4-7          2.187926   1.972953   1.109 0.267461
KK2 7-9          0.581175   1.872201   0.310 0.756242
KK2 >9          -1.616967   1.695194  -0.954 0.340169
LL              -5.938656   2.328102  -2.551 0.010753 *
MM               3.915382   2.481565   1.578 0.114630
OO2 0-5          8.439711   1.494165   5.648 1.64e-08 ***
OO2 5-20        15.894986   1.854587   8.571  < 2e-16 ***
OO2 20-40        9.968548   2.635513   3.782 0.000156 ***
OO2 >40          4.855519   5.506694   0.882 0.377923
I(ZAAA^0.5)     -4.674794   1.976427  -2.365 0.018026 *
BB0              9.145582   0.857530  10.665  < 2e-16 ***
II12            -3.391121   1.068003  -3.175 0.001500 **
DD2 20-30        3.394561   1.086500   3.124 0.001785 **
DD2 30-45        0.205018   1.148284   0.179 0.858299
DD2 45-60        0.162241   1.951115   0.083 0.933731
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 47.92 on 19977 degrees of freedom
Multiple R-squared:  0.2629,    Adjusted R-squared:  0.2607
F-statistic: 120.7 on 59 and 19977 DF,  p-value: < 2.2e-16
```

Appendix 6 – Summary of Base Model 2 Statistics

```
Deviance Residuals:
    Min       1Q   Median       3Q      Max
-3.2878  -0.9949   0.5364   0.8465   2.6054

Coefficients: (11 not defined because of singularities)
              Estimate Std. Error z value Pr(>|z|)
(Intercept)  1.505e+00  6.110e-01   2.464 0.013758 *
J           -5.941e-04  1.931e-04  -3.077 0.002091 **
K           -5.867e-02  9.474e-02  -0.619 0.535752
L           -1.085e-01  3.572e-02  -3.037 0.002390 **
M           -4.974e-01  6.963e-02  -7.143 9.12e-13 ***
N1          -1.324e+00  5.543e-01  -2.388 0.016940 *
N2          -1.227e+00  5.564e-01  -2.206 0.027383 *
N3          -1.148e+00  5.563e-01  -2.063 0.039108 *
N4          -1.374e+00  5.583e-01  -2.461 0.013854 *
N5          -1.060e+00  5.556e-01  -1.908 0.056440 .
N6          -6.651e-01  5.859e-01  -1.135 0.256248
N7                  NA         NA      NA       NA
O            6.976e-02  1.130e-01   0.617 0.537161
P           -4.996e-04  5.864e-05  -8.520  < 2e-16 ***
Q            4.292e-05  2.286e-05   1.878 0.060419 .
R3          -1.175e+00  6.735e-02 -17.452  < 2e-16 ***
R5          -1.262e+00  7.432e-02 -16.986  < 2e-16 ***
R10         -1.005e+00  7.084e-02 -14.183  < 2e-16 ***
R100                NA         NA      NA       NA
S0           5.330e-01  3.974e-02  13.412  < 2e-16 ***
S1          -1.156e-01  1.114e-01  -1.038 0.299423
S2                  NA         NA      NA       NA
T           -5.272e-05  1.317e-05  -4.002 6.28e-05 ***
U           -4.617e-05  3.349e-05  -1.378 0.168083
TU                  NA         NA      NA       NA
V1          -2.265e-02  9.881e-02  -0.229 0.818704
V2           2.361e-02  1.052e-01   0.224 0.822451
V3           2.819e-02  1.059e-01   0.266 0.790136
V4          -1.469e-01  1.071e-01  -1.372 0.170197
V5                  NA         NA      NA       NA
WA           4.651e-01  1.059e-01   4.392 1.12e-05 ***
WB          -2.678e-01  2.555e-01  -1.048 0.294525
WC          -7.792e-02  1.044e-01  -0.746 0.455401
WD           3.607e-02  1.437e-01   0.251 0.801761
WO                  NA         NA      NA       NA
X           -2.258e-06  5.955e-07  -3.792 0.000150 ***
Y           -1.594e-04  9.625e-05  -1.657 0.097613 .
Z           -3.906e-06  1.043e-06  -3.745 0.000180 ***
AA          -3.728e-06  1.053e-05  -0.354 0.723396
AAA                 NA         NA      NA       NA
AAAA                NA         NA      NA       NA
ZAAA         7.249e-02  1.489e-02   4.869 1.12e-06 ***
BB0          7.504e-02  7.914e-02   0.948 0.343013
BB1          3.234e-01  7.251e-02   4.460 8.20e-06 ***
BB2                 NA         NA      NA       NA
DD           2.007e-02  3.085e-03   6.505 7.75e-11 ***
EE          -5.445e-03  2.601e-03  -2.093 0.036332 *
EE1         -1.269e+00  2.527e-01  -5.020 5.17e-07 ***
EE2          5.472e-03  2.602e-03   2.103 0.035440 *
EE0         -1.714e-01  6.063e-02  -2.827 0.004700 **
FF1         -5.214e-01  8.114e-02  -6.426 1.31e-10 ***
FF2          2.746e-01  7.951e-02   3.454 0.000553 ***
FF3         -1.111e+00  1.116e-01  -9.955  < 2e-16 ***
FF4         -7.036e-01  1.055e-01  -6.667 2.61e-11 ***
FF5                 NA         NA      NA       NA
GG          -1.427e-01  7.736e-02  -1.844 0.065138 .
HH          -5.764e-01  5.538e-02 -10.408  < 2e-16 ***
II0         -1.785e-02  2.159e-01  -0.083 0.934119
II1          4.792e-02  9.908e-02   0.484 0.628627
II2          4.372e-02  1.019e-01   0.429 0.667867
II3          5.933e-02  1.147e-01   0.517 0.604941
II4          3.173e-01  1.180e-01   2.688 0.007194 **
II5          1.522e-01  1.324e-01   1.150 0.250343
II6          3.742e-01  1.334e-01   2.806 0.005013 **
II7          3.798e-01  1.466e-01   2.591 0.009572 **
II8          3.817e-01  1.243e-01   3.070 0.002141 **
II9          3.280e-01  2.077e-01   1.580 0.114217
II10         3.360e-01  1.362e-01   2.468 0.013601 *
II11         3.949e-01  3.071e-01   1.286 0.198446
II12         3.660e-01  1.027e-01   3.562 0.000367 ***
II13         1.511e-01  1.042e-01   1.450 0.147192
II14         2.800e-01  1.823e-01   1.536 0.124524
II15         2.525e-01  1.300e-01   1.942 0.052172 .
II16         2.722e-01  1.410e-01   1.931 0.053543 .
II17         2.164e-01  1.131e-01   1.913 0.055719 .
II18         5.847e-01  1.288e-01   4.540 5.64e-06 ***
II19         2.236e-01  1.235e-01   1.811 0.070155 .
II20                NA         NA      NA       NA
JJ          -1.311e-01  3.513e-02  -3.733 0.000190 ***
KK           1.861e-02  8.669e-03   2.146 0.031858 *
LL          -3.396e-01  1.121e-01  -3.030 0.002445 **
MM          -4.990e-01  1.294e-01  -3.856 0.000115 ***
NN           1.395e-01  9.613e-02   1.451 0.146746
OO           1.700e-02  3.919e-03   4.337 1.44e-05 ***
OOO          5.092e-01  7.681e-02   6.630 3.36e-11 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 25993  on 20036  degrees of freedom
Residual deviance: 21814  on 19963  degrees of freedom
AIC: 21962

Number of Fisher Scoring iterations: 5
```

Appendix 7 – Summary of Model 2 Statistics

```
Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.7174  -0.9258   0.4762   0.8147   2.4844

Coefficients:
                      Estimate Std. Error z value Pr(>|z|)
(Intercept)          2.2258192  0.2612408   8.520  < 2e-16 ***
J                   -0.0005990  0.0001964  -3.050 0.002285 **
L                   -0.1111592  0.0370924  -2.997 0.002728 **
M                   -0.4256604  0.0715269  -5.951 2.66e-09 ***
N1                  -0.7198379  0.1935824  -3.719 0.000200 ***
N2                  -0.5951854  0.2002016  -2.973 0.002950 **
N3                  -0.5681078  0.2016588  -2.817 0.004845 **
N4                  -0.8588084  0.2076389  -4.136 3.53e-05 ***
N5                  -0.7037640  0.2085565  -3.374 0.000740 ***
O                    0.1932951  0.1131798   1.708 0.087663 .
P3 <=391.6          -0.3482140  0.0425534  -8.183 2.77e-16 ***
P3 <=783.2          -0.3640711  0.0549113  -6.630 3.35e-11 ***
P3 <=Inf            -0.5699699  0.0727849  -7.831 4.84e-15 ***
Q3 <=783.2          -0.0366089  0.0498245  -0.735 0.462487
Q3 <=1174.8         -0.3407769  0.0733206  -4.648 3.36e-06 ***
Q3 <=1566.4         -0.1601427  0.0648290  -2.470 0.013503 *
Q3 <=Inf            -0.2226230  0.0760177  -2.929 0.003405 **
R3                  -0.9937449  0.0684836 -14.511  < 2e-16 ***
R5                  -1.0929594  0.0757953 -14.420  < 2e-16 ***
R10                 -0.9018225  0.0720578 -12.515  < 2e-16 ***
S0                   0.4508576  0.0394217  11.437  < 2e-16 ***
U3 <=383.768        -0.2517453  0.0517694  -4.863 1.16e-06 ***
U3 <=Inf            -0.0924184  0.0594671  -1.554 0.120159
V4                  -0.1670144  0.0582744  -2.866 0.004157 **
WA                   0.6448307  0.0515091  12.519  < 2e-16 ***
WD                   0.2468116  0.1122507   2.199 0.027895 *
X3 <=Inf            -0.2122265  0.0691325  -3.070 0.002142 **
Y3 <=274.12         -0.3019420  0.0688993  -4.382 1.17e-05 ***
Y3 <=Inf            -0.1776305  0.0770550  -2.305 0.021153 *
Z3 <=Inf            -0.3495289  0.0922332  -3.790 0.000151 ***
AA3 <=Inf           -0.4225471  0.0937352  -4.508 6.55e-06 ***
ZAAA3 <=4            0.0460656  0.0503611   0.915 0.360346
ZAAA3 <=Inf          0.2245126  0.0849160   2.644 0.008195 **
BB0                 -0.3789517  0.0428717  -8.839  < 2e-16 ***
BB2                 -0.2650050  0.0741854  -3.572 0.000354 ***
DD3 <=29.7616        0.0803691  0.0821789   0.978 0.328085
DD3 <=37.5936        0.2512806  0.0916984   2.740 0.006138 **
DD3 <=Inf            0.5372814  0.1188531   4.521 6.17e-06 ***
EE3 <=15.664        -0.5683743  0.0638991  -8.895  < 2e-16 ***
EE3 <=49.44341      -0.1905828  0.0681248  -2.798 0.005149 **
EE3 <=163.4742032    0.0507787  0.0619282   0.820 0.412238
EE3 <=Inf            0.4650190  0.0584470   7.956 1.77e-15 ***
FF1                 -0.6139144  0.0475792 -12.903  < 2e-16 ***
FF3                 -0.8653653  0.0920942  -9.397  < 2e-16 ***
FF4                 -0.6932186  0.0878462  -7.891 2.99e-15 ***
GG                  -0.1501409  0.0761943  -1.971 0.048781 *
HH                  -0.4154386  0.0562500  -7.386 1.52e-13 ***
II1                 -0.2473608  0.0510873  -4.842 1.29e-06 ***
II2                 -0.3051841  0.0551255  -5.536 3.09e-08 ***
II3                 -0.1758377  0.0781725  -2.249 0.024490 *
II13                -0.1349562  0.0596799  -2.261 0.023738 *
II18                 0.2432302  0.0970596   2.506 0.012211 *
II20                -0.3640831  0.0951439  -3.827 0.000130 ***
JJ                  -0.1313480  0.0358543  -3.663 0.000249 ***
KK3 <=3.916          0.4367423  0.0870314   5.018 5.22e-07 ***
KK3 <=7.0488         0.2758018  0.0716684   3.848 0.000119 ***
KK3 <=9.3984         0.3746380  0.0720355   5.201 1.98e-07 ***
KK3 <=Inf            0.4554883  0.0656862   6.934 4.08e-12 ***
LL                  -0.2839237  0.1119606  -2.536 0.011215 *
MM                  -0.3831608  0.1303391  -2.940 0.003285 **
NN                   0.1933072  0.0967208   1.999 0.045651 *
OO3 <=7.832         -0.2212313  0.0572650  -3.863 0.000112 ***
OO3 <=Inf           -0.2457493  0.0695144  -3.535 0.000407 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 23993  on 20036  degrees of freedom
Residual deviance: 20922  on 19974  degrees of freedom
AIC: 21048

Number of Fisher Scoring iterations: 5
```

Appendix 8 – Profit Outcomes of Model 2 by Cutoff Level

Probability cutoff	Profits from approved	Profit improvement (€)	Profit improvement (%)
1 %	- €	- 82 693.00 €	-100.00 %
2 %	- €	- 82 693.00 €	-100.00 %
3 %	134.06 €	- 82 558.94 €	-99.84 %
4 %	427.32 €	- 82 265.68 €	-99.48 %
5 %	860.96 €	- 81 832.04 €	-98.96 %
6 %	1 282.06 €	- 81 410.94 €	-98.45 %
7 %	1 676.30 €	- 81 016.70 €	-97.97 %
8 %	2 448.43 €	- 80 244.57 €	-97.04 %
9 %	2 598.57 €	- 80 094.43 €	-96.86 %
10 %	2 937.47 €	- 79 755.53 €	-96.45 %
11 %	3 516.65 €	- 79 176.35 €	-95.75 %
12 %	3 976.42 €	- 78 716.58 €	-95.19 %
13 %	4 304.13 €	- 78 388.87 €	-94.80 %
14 %	5 507.85 €	- 77 185.15 €	-93.34 %
15 %	6 696.37 €	- 75 996.63 €	-91.90 %
16 %	7 443.64 €	- 75 249.36 €	-91.00 %
17 %	9 172.88 €	- 73 520.12 €	-88.91 %
18 %	10 639.45 €	- 72 053.55 €	-87.13 %
19 %	11 135.68 €	- 71 557.32 €	-86.53 %
20 %	12 022.21 €	- 70 670.79 €	-85.46 %
21 %	13 350.83 €	- 69 342.17 €	-83.85 %
22 %	14 574.01 €	- 68 118.99 €	-82.38 %
23 %	16 225.97 €	- 66 467.03 €	-80.38 %
24 %	18 183.39 €	- 64 509.61 €	-78.01 %
25 %	18 901.71 €	- 63 791.29 €	-77.14 %
26 %	20 200.16 €	- 62 492.84 €	-75.57 %
27 %	21 368.20 €	- 61 324.80 €	-74.16 %
28 %	22 613.61 €	- 60 079.39 €	-72.65 %
29 %	24 262.05 €	- 58 430.95 €	-70.66 %
30 %	25 881.09 €	- 56 811.91 €	-68.70 %
31 %	27 362.59 €	- 55 330.41 €	-66.91 %
32 %	28 341.69 €	- 54 351.31 €	-65.73 %
33 %	30 495.80 €	- 52 197.20 €	-63.12 %
34 %	33 055.71 €	- 49 637.29 €	-60.03 %
35 %	35 708.27 €	- 46 984.73 €	-56.82 %
36 %	36 891.28 €	- 45 801.72 €	-55.39 %
37 %	38 859.62 €	- 43 833.38 €	-53.01 %
38 %	40 406.81 €	- 42 286.19 €	-51.14 %
39 %	41 809.34 €	- 40 883.66 €	-49.44 %
40 %	42 908.80 €	- 39 784.20 €	-48.11 %
41 %	44 584.48 €	- 38 108.52 €	-46.08 %
42 %	45 302.71 €	- 37 390.29 €	-45.22 %
43 %	46 425.67 €	- 36 267.33 €	-43.86 %
44 %	47 724.15 €	- 34 968.85 €	-42.29 %
45 %	49 043.00 €	- 33 650.00 €	-40.69 %
46 %	50 642.41 €	- 32 050.59 €	-38.76 %
47 %	51 628.39 €	- 31 064.61 €	-37.57 %
48 %	52 616.30 €	- 30 076.70 €	-36.37 %
49 %	54 299.04 €	- 28 393.96 €	-34.34 %
50 %	55 310.62 €	- 27 382.38 €	-33.11 %
51 %	56 337.06 €	- 26 355.94 €	-31.87 %

52 %	58 203.32 €	-	24 489.68 €	-29.62 %
53 %	59 516.87 €	-	23 176.13 €	-28.03 %
54 %	60 841.34 €	-	21 851.66 €	-26.43 %
55 %	62 062.24 €	-	20 630.76 €	-24.95 %
56 %	63 278.79 €	-	19 414.21 €	-23.48 %
57 %	64 314.61 €	-	18 378.39 €	-22.22 %
58 %	65 402.15 €	-	17 290.85 €	-20.91 %
59 %	66 160.93 €	-	16 532.07 €	-19.99 %
60 %	66 978.45 €	-	15 714.55 €	-19.00 %
61 %	67 601.86 €	-	15 091.14 €	-18.25 %
62 %	68 593.80 €	-	14 099.20 €	-17.05 %
63 %	69 926.86 €	-	12 766.14 €	-15.44 %
64 %	71 299.20 €	-	11 393.80 €	-13.78 %
65 %	72 055.37 €	-	10 637.63 €	-12.86 %
66 %	72 696.16 €	-	9 996.84 €	-12.09 %
67 %	74 014.33 €	-	8 678.67 €	-10.50 %
68 %	74 788.20 €	-	7 904.80 €	-9.56 %
69 %	75 129.92 €	-	7 563.08 €	-9.15 %
70 %	76 360.13 €	-	6 332.87 €	-7.66 %
71 %	77 624.45 €	-	5 068.55 €	-6.13 %
72 %	78 326.19 €	-	4 366.81 €	-5.28 %
73 %	79 138.75 €	-	3 554.25 €	-4.30 %
74 %	79 568.57 €	-	3 124.43 €	-3.78 %
75 %	80 000.15 €	-	2 692.85 €	-3.26 %
76 %	80 796.61 €	-	1 896.39 €	-2.29 %
77 %	81 016.31 €	-	1 676.69 €	-2.03 %
78 %	81 392.52 €	-	1 300.48 €	-1.57 %
79 %	81 419.90 €	-	1 273.10 €	-1.54 %
80 %	81 495.90 €	-	1 197.10 €	-1.45 %
81 %	81 626.31 €	-	1 066.69 €	-1.29 %
82 %	82 166.93 €	-	526.07 €	-0.64 %
83 %	82 532.45 €	-	160.55 €	-0.19 %
84 %	82 622.18 €	-	70.82 €	-0.09 %
85 %	83 127.49 €		434.49 €	0.53 %
86 %	83 200.92 €		507.92 €	0.61 %
87 %	83 371.63 €		678.63 €	0.82 %
88 %	**83 647.44 €**		**954.44 €**	**1.15 %**
89 %	83 597.98 €		904.98 €	1.09 %
90 %	83 521.08 €		828.08 €	1.00 %
91 %	83 614.46 €		921.46 €	1.11 %
92 %	83 442.12 €		749.12 €	0.91 %
93 %	83 285.28 €		592.28 €	0.72 %
94 %	83 271.07 €		578.07 €	0.70 %
95 %	83 003.77 €		310.77 €	0.38 %
96 %	82 987.20 €		294.20 €	0.36 %
97 %	82 844.63 €		151.63 €	0.18 %
98 %	82 731.40 €		38.40 €	0.05 %
99 %	82 693.00 €		- €	0.00 %
100 %	82 693.00 €		- €	0.00 %

Appendix 9 – Summary of Model 1C Statistics

```
Residuals:
    Min      1Q  Median      3Q     Max 
-162.15  -23.61   -6.25    9.84  605.86 

Coefficients:
                Estimate Std. Error t value Pr(>|t|)    
(Intercept)     93.79399    5.76826  16.260  < 2e-16 ***
BBBProb        -83.52316    3.42246 -24.404  < 2e-16 ***
K               -3.94187    1.89629  -2.079 0.037655 *  
L                2.22141    0.71956   3.087 0.002023 ** 
M               -5.25336    1.52970  -3.434 0.000595 ***
N2              -2.81483    1.19334  -2.359 0.018344 *  
O               -4.77231    2.40803  -1.982 0.047512 *  
N3              -2.60140    1.07565  -2.418 0.015596 *  
P2 0-250         6.35404    4.00647   1.586 0.112768    
P2 250-500       5.27098    4.00696   1.315 0.188371    
P2 500-1000      7.51565    4.06509   1.849 0.064498 .  
P2 >1000        12.14292    4.19193   2.897 0.003775 ** 
R10             -3.58151    0.82525  -4.340 1.43e-05 ***
S0              -2.06510    0.91027  -2.269 0.023299 *  
I(T^0.5)         0.12181    0.05324   2.288 0.022144 *  
Y2 0-250         0.91838    1.13326   0.810 0.417725    
Y2 250-500      -1.52962    0.92339  -1.657 0.097632 .  
Y2 500-1000      1.79149    1.35408   1.323 0.185840    
Y2 >1000         0.85114    3.68884   0.231 0.817524    
Q2 800-1600      1.72669    1.22722   1.407 0.159445    
Q2 1600-4000    10.68694    1.92307   5.557 2.78e-08 ***
Q2 4000-7000    43.95459    4.60930   9.536  < 2e-16 ***
Q2 >7000        31.19541    5.11463   6.099 1.09e-09 ***
WB               8.31599    5.14076   1.618 0.105752    
I(log(AA + 1))   3.41802    1.11035   3.078 0.002085 ** 
AA2 0-200      -16.18606    5.18043  -3.124 0.001784 ** 
AA2 200-500    -15.28198    6.43670  -2.374 0.017597 *  
AA2 500-1000    -8.50417    7.57300  -1.123 0.261468    
AA2 >1000      -27.40294    9.43553  -2.904 0.003686 ** 
I(EE2^0.1)     -18.82258    1.21557 -15.485  < 2e-16 ***
EE0            -27.19895    2.24759 -12.101  < 2e-16 ***
FF2              2.11146    0.92073   2.293 0.021843 *  
FF3             34.83670    1.61664  21.549  < 2e-16 ***
HH               4.62364    1.22076   3.787 0.000153 ***
KK2 2-4         -1.84573    1.68751  -1.094 0.274072    
KK2 4-7         -2.46769    1.96528  -1.256 0.209260    
KK2 7-9         -2.66376    1.86104  -1.431 0.152351    
KK2 >9          -3.22881    1.67970  -1.922 0.054588 .  
LL             -11.61748    2.32151  -5.004 5.65e-07 ***
MM              -4.33493    2.48662  -1.743 0.081297 .  
OO2 0-5          8.01130    1.47928   5.416 6.18e-08 ***
OO2 5-20        12.01920    1.80508   6.659 2.84e-11 ***
OO2 20-40        3.92426    2.50207   1.568 0.116802    
OO2 >40         -3.09755    5.40579  -0.573 0.566647    
I(ZAAA^0.5)     -4.35191    1.75596  -2.478 0.013207 *  
BB0              3.52734    0.89026   3.962 7.45e-05 ***
DD2 20-30        4.39020    1.07382   4.088 4.36e-05 ***
DD2 30-45        5.09330    1.15255   4.419 9.96e-06 ***
DD2 45-60        5.87398    1.94718   3.017 0.002559 ** 
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 47.61 on 19988 degrees of freedom
Multiple R-squared:  0.2719,	Adjusted R-squared:  0.2702 
F-statistic: 155.5 on 48 and 19988 DF,  p-value: < 2.2e-16
```

Appendix 10 – Summary of Model 3A Statistics

```
Residuals:
   Min     1Q Median     3Q    Max
-62.29  -8.38  -2.89   1.26 338.84

Coefficients:
                    Estimate Std. Error t value Pr(>|t|)
(Intercept)        68.504503   8.458472   8.099 6.86e-16 ***
BBBProb           -62.718512   4.375476 -14.334  < 2e-16 ***
N1                 -1.770616   0.898995  -1.970  0.04894 *
O                 -16.055981   5.165075  -3.109  0.00189 **
N3                 -0.788676   0.992784  -0.794  0.42699
S1                  3.695361   3.462733   1.067  0.28594
P2 0-250            3.724937   3.744150   0.995  0.31985
P2 250-500          3.373326   3.750187   0.900  0.36842
P2 500-1000         3.951492   3.839547   1.029  0.30345
P2 >1000            6.019837   4.094928   1.470  0.14160
V1                 -2.233273   1.328694  -1.681  0.09286 .
V2                 -3.197005   1.617619  -1.976  0.04817 *
V3                 -1.997821   1.613069  -1.239  0.21558
V4                 -4.391697   1.842221  -2.384  0.01717 *
X2 0-20000          0.654575   3.981711   0.164  0.86943
X2 20000-80000      0.556450   3.934607   0.141  0.88754
X2 80000-160000     2.956250   4.155777   0.711  0.47689
X2 >160000         -3.760988   5.058729  -0.743  0.45723
Y2 0-250           -0.276949   4.023041  -0.069  0.94512
Y2 250-500          1.207532   3.941324   0.306  0.75933
Y2 500-1000         3.765290   4.045722   0.931  0.35206
Y2 >1000            3.969512   4.861985   0.816  0.41429
Q                  -0.003767   0.001194  -3.156  0.00161 **
Q2 800-1600         2.907267   1.231768   2.360  0.01830 *
Q2 1600-4000       12.520357   2.755549   4.544 5.65e-06 ***
Q2 4000-7000       49.443668   8.219766   6.015 1.92e-09 ***
Q2 >7000           55.228732  10.052864   5.494 4.12e-08 ***
WA                  3.219517   2.834189   1.136  0.25603
WC                 15.517440   7.100824   2.185  0.02891 *
I(log(AA + 1))      2.427201   1.905925   1.274  0.20290
AA2 0-200         -11.749181   9.095977  -1.292  0.19652
AA2 200-500       -13.093821  11.077171  -1.182  0.23724
AA2 500-1000      -13.239893  12.695990  -1.043  0.29707
AA2 >1000         -26.547845  15.446249  -1.719  0.08572 .
EE2B 0-50           1.091576   1.968353   0.555  0.57922
EE2B 50-2500       -0.441467   1.778733  -0.248  0.80400
EE2B 2500-10000    -6.645495   2.254640  -2.947  0.00322 **
EE2B >10000       -12.086701   2.813880  -4.295 1.78e-05 ***
FF1                -1.554175   0.717258  -2.167  0.03029 *
FF3                12.586391   2.217099   5.677 1.45e-08 ***
LL                -17.302398   3.625538  -4.772 1.87e-06 ***
OO2 0-5             3.101043   2.103410   1.474  0.14046
OO2 5-20            4.484318   2.298424   1.951  0.05111 .
OO2 20-40           1.314932   3.092736   0.425  0.67073
OO2 >40            -5.282386   8.893130  -0.594  0.55255
DD2 20-30           0.492976   2.117176   0.233  0.81589
DD2 30-45           2.864940   2.172912   1.318  0.18740
DD2 45-60           3.169045   2.489025   1.273  0.20300
BB2                -1.955838   1.398094  -1.399  0.16189
HH                 -2.026838   1.822535  -1.112  0.26615
ZAAA2 1-2.5        -3.707306   1.500959  -2.470  0.01355 *
ZAAA2 2.5-5        -3.628083   1.691333  -2.145  0.03199 *
ZAAA2 >5           -4.590954   2.137309  -2.148  0.03176 *
II3                 1.775300   1.521457   1.167  0.24333
II10                2.925890   2.200727   1.330  0.18374
II19                2.401132   1.681463   1.428  0.15335
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 23.37 on 5157 degrees of freedom
Multiple R-squared:  0.1571,    Adjusted R-squared:  0.1482
F-statistic: 17.48 on 55 and 5157 DF,  p-value: < 2.2e-16
```

Appendix 11 – Summary of Model 3B Statistics

```
Residuals:
    Min      1Q  Median      3Q     Max
-117.06  -18.14   -6.64    4.60  540.79

Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)      5.448e+01  4.149e+01   1.313 0.189196
BBBProb         -8.314e+01  5.878e+00 -14.143  < 2e-16 ***
K               -5.133e+00  3.301e+00  -1.555 0.119987
S0               3.232e+00  1.617e+00   1.999 0.045688 *
P2 0-250         7.228e+00  6.778e+00   1.066 0.286304
P2 250-500       5.116e+00  6.787e+00   0.754 0.451001
P2 500-1000      8.906e+00  6.901e+00   1.291 0.196911
P2 >1000         1.074e+01  7.122e+00   1.508 0.131611
Y2 0-250        -4.102e+00  7.650e+00  -0.536 0.591876
Y2 250-500      -6.450e-01  7.571e+00  -0.085 0.932107
Y2 500-1000      8.699e-01  7.749e+00   0.112 0.910622
Y2 >1000         5.298e+00  9.417e+00   0.563 0.573703
X2 0-20000       5.974e+00  7.665e+00   0.779 0.435823
X2 20000-80000   4.739e+00  7.583e+00   0.625 0.532048
X2 80000-160000  7.605e+00  7.864e+00   0.967 0.333548
X2 >160000       1.338e+01  9.187e+00   1.456 0.145498
Q                5.258e-03  7.572e-04   6.944 4.31e-12 ***
WB              -1.140e+01  8.468e+00  -1.346 0.178229
AA2 0-200       -4.661e+00  1.708e+00  -2.729 0.006370 **
AA2 200-500     -3.362e-01  1.897e+00  -0.177 0.859322
AA2 500-1000     1.111e+01  3.722e+00   2.985 0.002849 **
AA2 >1000       -7.241e+00  5.811e+00  -1.246 0.212813
EE1              2.202e+01  1.094e+01   2.013 0.044153 *
EE2             -1.488e-04  9.767e-05  -1.524 0.127632
EE2B 0-50        5.351e+00  2.396e+00   2.233 0.025594 *
EE2B 50-2500     4.225e+00  2.150e+00   1.965 0.049419 *
EE2B 2500-10000 -5.270e+00  3.390e+00  -1.555 0.120085
EE2B >10000     -1.600e+01  5.443e+00  -2.940 0.003292 **
FF1             -2.412e+00  1.365e+00  -1.767 0.077223 .
FF3              3.730e+01  3.429e+00  10.876  < 2e-16 ***
LL              -1.351e+01  4.028e+00  -3.355 0.000800 ***
MM              -1.159e+01  4.533e+00  -2.558 0.010559 *
OO2 0-5          2.880e+00  2.376e+00   1.212 0.225607
OO2 5-20         3.152e+00  2.905e+00   1.085 0.277935
OO2 20-40       -4.176e+00  4.579e+00  -0.912 0.361851
OO2 >40         -1.159e+01  1.134e+01  -1.021 0.307135
DD               6.295e-01  1.769e-01   3.558 0.000377 ***
DD2 20-30       -4.546e+00  2.537e+00  -1.792 0.073180 .
DD2 30-45       -5.194e+00  3.890e+00  -1.335 0.181837
DD2 45-60       -1.326e+01  5.829e+00  -2.275 0.022939 *
ZAAA2 0-1       -1.984e+01  4.050e+01  -0.490 0.624242
ZAAA2 1-2.5     -1.352e+01  4.046e+01  -0.334 0.738312
ZAAA2 2.5-5     -1.263e+01  4.047e+01  -0.312 0.754963
ZAAA2 >5        -1.337e+01  4.053e+01  -0.330 0.741518
KK              -4.584e-01  2.266e-01  -2.023 0.043162 *
II1             -4.060e+00  1.726e+00  -2.353 0.018679 *
II2             -4.280e+00  2.326e+00  -1.840 0.065835 .
II3              6.732e+00  2.693e+00   2.500 0.012464 *
II10            -6.288e+00  3.568e+00  -1.762 0.078128 .
II12            -2.267e+00  1.873e+00  -1.211 0.226081
II13            -2.535e+00  2.042e+00  -1.241 0.214657
II16            -4.446e+00  3.779e+00  -1.176 0.239484
II18             4.441e+00  2.927e+00   1.517 0.129253
II19             4.330e+00  2.916e+00   1.485 0.137683
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 40.05 on 5001 degrees of freedom
Multiple R-squared:  0.2016,    Adjusted R-squared:  0.1932
F-statistic: 23.83 on 53 and 5001 DF,  p-value: < 2.2e-16
```

Appendix 12 – Summary of Model 3C Statistics

```
Residuals:
    Min      1Q  Median      3Q     Max 
-137.51  -30.37  -11.82   15.77  576.84 

Coefficients:
                  Estimate Std. Error t value Pr(>|t|)    
(Intercept)      4.413e+01  1.513e+01   2.916 0.003558 ** 
EEEProb         -1.115e+02  9.367e+00 -11.907  < 2e-16 ***
K               -1.048e+01  5.859e+00  -1.789 0.073702 .  
M               -4.004e+00  3.662e+00  -1.093 0.274257    
O               -9.363e+00  5.288e+00  -1.771 0.076672 .  
SO               3.625e+00  2.010e+00   1.804 0.071353 .  
P               -2.076e-02  1.844e-02  -1.126 0.260110    
P2 0-250         1.801e+01  9.926e+00   1.815 0.069614 .  
P2 250-500       1.846e+01  1.181e+01   1.563 0.118139    
P2 500-1000      2.596e+01  1.702e+01   1.525 0.127258    
P2 >1000         4.118e+01  2.359e+01   1.745 0.080993 .  
Y2 0-250        -6.580e+00  9.644e+00  -0.682 0.495046    
Y2 250-500      -1.255e+01  9.486e+00  -1.323 0.185914    
Y2 500-1000     -5.460e+00  9.877e+00  -0.553 0.580423    
Y2 >1000        -1.937e+01  1.470e+01  -1.318 0.187565    
X2 0-20000       9.739e+00  9.840e+00   0.990 0.322309    
X2 20000-80000   5.442e+00  9.524e+00   0.571 0.567705    
X2 80000-160000  9.637e+00  9.870e+00   0.976 0.328909    
X2 >160000      -3.757e+00  1.331e+01  -0.282 0.777819    
Q                3.661e-03  2.306e-03   1.587 0.112470    
Q2 800-1600      6.763e+00  2.732e+00   2.476 0.013336 *  
Q2 1600-4000     1.571e+01  5.413e+00   2.902 0.003719 ** 
Q2 4000-7000     4.337e+01  1.289e+01   3.364 0.000773 ***
Q2 >7000         5.766e+00  1.786e+01   0.323 0.746824    
WA               4.929e+00  2.482e+00   1.986 0.047117 *  
AA2 0-200       -4.194e+00  2.399e+00  -1.748 0.080479 .  
AA2 200-500     -1.217e+00  2.734e+00  -0.445 0.656178    
AA2 500-1000     3.944e+00  5.710e+00   0.691 0.489820    
AA2 >1000       -5.639e-01  9.657e+00  -0.058 0.953434    
EE1              3.383e+01  1.083e+01   3.124 0.001795 ** 
EE2             -5.664e-04  3.048e-04  -1.858 0.063175 .  
EE2a 0-50        5.747e-01  3.014e+00   0.191 0.848786    
EE2a 50-2500    -1.739e+00  2.847e+00  -0.611 0.541322    
EE2a 2500-10000 -1.378e+01  5.090e+00  -2.707 0.006816 ** 
EE2a >10000     -2.020e+01  9.807e+00  -2.059 0.039502 *  
FF2              3.865e+00  2.436e+00   1.587 0.112632    
FF3              3.892e+01  3.429e+00  11.349  < 2e-16 ***
LL              -1.572e+01  4.667e+00  -3.368 0.000763 ***
MM              -1.441e+01  4.767e+00  -3.023 0.002513 ** 
NN               1.092e+01  4.695e+00   2.326 0.020061 *  
HH               6.358e+00  2.625e+00   2.422 0.015472 *  
OO              -5.407e-01  3.527e-01  -1.533 0.125366    
OO2 0-5          6.720e+00  3.174e+00   2.117 0.034306 *  
OO2 5-20         1.252e+01  4.895e+00   2.558 0.010551 *  
OO2 20-40        1.344e+01  1.033e+01   1.302 0.193072    
OO2 >40          3.353e+01  2.150e+01   1.559 0.118971    
DD               3.041e-01  2.588e-01   1.175 0.240055    
DD2 20-30        7.203e+00  2.988e+00   2.410 0.015969 *  
DD2 30-45        5.903e+00  5.379e+00   1.097 0.272579    
DD2 45-60        2.924e+00  9.004e+00   0.325 0.745406    
ZAAA2 1-2.5      7.447e+00  3.000e+00   2.483 0.013071 *  
ZAAA2 2.5-5      1.113e+01  3.862e+00   2.883 0.003960 ** 
ZAAA2 >5         1.092e+01  5.982e+00   1.826 0.067904 .  
KK2 2-4         -5.005e+00  3.726e+00  -1.343 0.179192    
KK2 4-7         -4.582e+00  4.411e+00  -1.039 0.299002    
KK2 7-9         -2.145e+00  4.092e+00  -0.524 0.600127    
KK2 >9          -6.630e+00  3.721e+00  -1.782 0.074813 .  
II0              1.852e+01  9.852e+00   1.880 0.060210 .  
II3              4.342e+00  3.542e+00   1.226 0.220275    
II4              5.630e+00  3.801e+00   1.481 0.138600    
II5              1.066e+01  4.861e+00   2.194 0.028301 *  
II6              6.001e+00  4.638e+00   1.294 0.195795    
II12             3.197e+00  2.658e+00   1.203 0.229190    
II14             1.401e+01  6.947e+00   2.017 0.043761 *  
II15             1.117e+01  4.539e+00   2.461 0.013870 *  
II17             4.024e+00  3.426e+00   1.174 0.240269    
II19             8.988e+00  4.384e+00   2.050 0.040430 *  
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 56.53 on 5005 degrees of freedom
Multiple R-squared:  0.2302,    Adjusted R-squared:   0.22 
F-statistic: 22.67 on 66 and 5005 DF,  p-value: < 2.2e-16
```

Appendix 13 – Summary of Model 3D Statistics

```
Residuals:
    Min      1Q  Median      3Q     Max
-212.43  -37.09  -11.09   28.45  596.82

Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)      6.484e+01  8.266e+00   7.845 5.34e-15 ***
BBBProb         -9.947e+01  7.111e+00 -13.988  < 2e-16 ***
L                4.676e+00  1.884e+00   2.482 0.013082 *
M               -8.355e+00  3.114e+00  -2.683 0.007324 **
P                8.236e-03  3.176e-03   2.593 0.009534 **
R5               3.499e+00  2.311e+00   1.514 0.130055
V1               6.336e+00  3.322e+00   1.907 0.056537 .
V2               6.024e+00  3.344e+00   1.802 0.071654 .
V3               7.636e+00  3.468e+00   2.202 0.027712 *
V5              -9.615e+00  8.317e+00  -1.156 0.247678
X               -6.398e-05  4.266e-05  -1.500 0.133782
Q2 800-1600      3.423e+00  2.362e+00   1.449 0.147295
Q2 1600-4000     1.891e+01  3.172e+00   5.963 2.66e-09 ***
Q2 4000-7000     6.792e+01  9.360e+00   7.257 4.61e-13 ***
Q2 >7000         5.291e+01  1.117e+01   4.738 2.23e-06 ***
WB               2.021e+01  1.152e+01   1.755 0.079346 .
WC              -6.302e+00  2.334e+00  -2.700 0.006950 **
AA2 0-200       -4.918e+00  2.343e+00  -2.099 0.035843 *
AA2 200-500      3.248e+00  3.405e+00   0.954 0.340207
AA2 500-1000     2.807e+01  7.757e+00   3.619 0.000299 ***
AA2 >1000        1.340e+01  2.757e+01   0.486 0.626876
EE1              1.627e+01  9.224e+00   1.763 0.077886 .
EE2B 0-50       -1.434e+00  2.845e+00  -0.504 0.614173
EE2B 50-2500    -8.751e+00  2.854e+00  -3.066 0.002181 **
EE2B 2500-10000 -3.044e+01  4.860e+00  -6.264 4.10e-10 ***
EE2B >10000     -4.622e+01  7.753e+00  -5.962 2.68e-09 ***
FF3              3.730e+01  3.487e+00  10.695  < 2e-16 ***
LL              -8.069e+00  5.717e+00  -1.411 0.158204
I(ZAAA^0.5)     -3.019e+00  2.423e+00  -1.246 0.212875
OO2 0-5          1.667e+01  4.042e+00   4.124 3.79e-05 ***
OO2 5-20         2.308e+01  4.814e+00   4.794 1.68e-06 ***
OO2 20-40        1.727e+01  5.926e+00   2.914 0.003585 **
OO2 >40          1.660e+01  1.032e+01   1.609 0.107762
DD2 20-30        6.812e+00  2.280e+00   2.988 0.002824 **
DD2 30-45        8.996e+00  2.962e+00   3.037 0.002402 **
DD2 45-60        4.249e+01  9.864e+00   4.308 1.68e-05 ***
II5              9.086e+00  5.563e+00   1.633 0.102452
II9             -1.345e+01  1.100e+01  -1.222 0.221726
II11            -2.347e+01  1.934e+01  -1.213 0.225034
II13             8.615e+00  2.957e+00   2.914 0.003589 **
II16             6.145e+00  5.904e+00   1.041 0.298018
AAA             -1.750e-03  4.637e-03  -0.377 0.705820
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 60.81 on 4655 degrees of freedom
Multiple R-squared:  0.2262,    Adjusted R-squared:  0.2194
F-statistic: 33.19 on 41 and 4655 DF,  p-value: < 2.2e-16
```