Charles University

Faculty of Social Sciences
Institute of Economic Studies

MASTER'S THESIS

Consumer Credit Risk Analysis: Evidence from the Czech Republic

Author: **Bc. Patricie Mittigová**

Supervisor: **prof. Ing. Evžen Kočenda, M.A., Ph.D., DSc.**

Academic Year: **2017/2018**

Declaration of Authorship

The author hereby declares that she compiled this thesis independently, using only the listed resources and literature, and the thesis has not been used to obtain a different or the same degree.

The author grants to Charles University permission to reproduce and to distribute copies of this thesis document in whole or in part.

Prague, May 3, 2018

Signature

Acknowledgments

First and foremost, I would like to express my gratitude to prof. Ing. Evžen Kočenda, the supervisor of this thesis, for his guidance and invaluable advice during writing of this thesis. Furthermore, I would like to thank to the bank which provided me with the data. Finally, I am proud to state that none of this thesis would be possible without the incredible support and infinite patience of my loving family. My final thanks are dedicated to Petr whose support, love and care helped me the most.

Abstract

An increase in the number of granted loans in last decades resulted in more attention paid to proper assessment of borrower's creditworthiness. For this purpose, credit scoring aims to classify good and bad applicants prior loan granting. In this thesis, I analyze a large real-world dataset of borrowers who were granted an unsecured consumer loan in the Czech Republic. The objective is to determine core default predictors while employing seven classification methods. Additionally, a performance measure is computed for each method in order to compare their suitability for examined loan types. Using logistic regression as the core model, the results suggest that borrower's age, monthly income, region of residence, and the number of children substantially influence the probability of default. Conversely, borrower's gender and education level did not prove to be significant for assessing client's creditworthiness. Comparing the performance of employed classification methods, it can be concluded that all models produced almost identical results and can be used for the purpose of credit scoring. This thesis complements rather a limited number of credit scoring studies in the Czech Republic and provides new findings about default determinants for unsecured consumer loans.

JEL Classification C10, C38, C51, D81, G21, H87
Keywords credit scoring model, default predictors, classification methods, personal characteristics, banking sector, consumer loan

Author's e-mail mit.patricie@gmail.com
Supervisor's e-mail kocenda@fsv.cuni.cz

Abstrakt

Nárůst počtu poskytnutých úvěrů v posledních desetiletích způsobil zvýšení důrazu na řádné posouzení úvěrové spolehlivosti dlužníků. Skóringové modely mají za cíl klasifikovat dobré a špatné žadatele před poskytnutím půjčky. V této práci analyzuji rozsáhlý soubor reálných dat obsahující informace o dlužnících, kterým byl v České republice poskytnut nezajištěný spotřebitelský úvěr. Cílem je stanovení hlavních indikátorů budoucího selhání pomocí použití sedmi klasifikačních metod. Pro každou metodu je vypočtena statistika hodnotící přesnost modelu, aby bylo možné porovnat jejich vhodnost pro zkoumané typy úvěrů. Výsledky logistické regrese, jakožto hlavního modelu, napovídají, že věk dlužníka, měsíční příjem, kraj, ve kterém bydlí a počet dětí značně ovlivňují pravděpodobnost selhání. Dlužníkovo pohlaví a úroveň vzdělání se naopak neprokázaly jako důležité faktory pro posouzení bonity klienta. Z porovnání použitých klasifikačních metod vyplývá, že všechny modely poskytly téměř shodné výsledky a mohou být použity jako skóringové modely. Tato práce doplňuje nízký počet studií zabývajících se ohodnocením úvěrové schopnosti v České republice a poskytuje nové poznatky o klíčových faktorech, které vedou k selhání dlužníků v případě nezajištěných spotřebitelských úvěrů.

Klasifikace JEL C10, C38, C51, D81, G21, H87

Klíčová slova ohodnocení úvěruschopnosti, indikátory selhání, klasifikační metody, osobní charakteristiky, bankovní sektor, spotřebitelský úvěr

E-mail autora mit.patricie@gmail.com

E-mail vedoucího práce kocenda@fsv.cuni.cz

Contents

List of Tables

List of Figures

Acronyms

AUC	Area under the Curve
CART	Classification and Regression Tree
FN	False Negative
FP	False Positive
IV	Information Value
KNN	K-Nearest Neighbors
ROC	Receiver Operating Characteristic
SVM	Support Vector Machines
TN	True Negative
TP	True Positive
WOE	Weight of Evidence

Master's Thesis Proposal

Author	Bc. Patricie Mittigová
Supervisor	prof. Ing. Evžen Kočenda, M.A., Ph.D., DSc.
Proposed topic	Consumer Credit Risk Analysis: Evidence from the Czech Republic

Motivation Credit risk represents the most important risk commercial banks have to manage. It accounts for approximately 70% of all risks banks face. Its appropriate qualification and management are therefore crucial. As the number of consumer loans has been growing in the last 15 years, the assessment of risk of default on credit has been paid a lot of attention. For this purpose, various credit scoring methods were developed in order to help commercial banks prevent the financial loss resulting from potential defaults. These methods classify applicants for a loan into bad and good borrowers according to probability of default. This helps to evaluate their creditworthiness.

Due to high importance in the banking sector, credit risk analysis and particular credit scoring techniques have been examined by a plethora of authors. Not only do they compare suitability and accuracy of various traditional methods, but they also investigate the application of less conventional approaches. A detailed list of credit scoring methods was assembled by Hand and Henley (1997), Vojtek and Kočenda (2006) or Abdou and Pointon (2011). The majority of the methods reviewed in these papers are widely used and evaluated. The most frequently employed approach is logistic regression which usually serves as a baseline for comparison with other methods. Other popular methods include linear discriminant analysis, k-nearest neighbours, decision trees, random forest, support vector machines and neural networks. Their application and comparison was investigated for instance by Bellotti and Crook (2007), Kruppa et al. (2013) or Abdou and Tsafack (2015).

In order to model probability of default, it is necessary to work with client's personal data which are further examined. The most important variables used in the analysis include demographic, financial, employment and behavioural indicators (Vojtek and Kočenda, 2006). Due to high data requirements, the research in this area is extremely difficult to conduct. Thus, the amount of studies which perform credit scoring analysis on real world data is very limited. To the author's knowledge, default predictors in European retail banking have been investigated only in the cases of France (Nguyen, 2015) and the Czech Republic (Kočenda and Vojtek, 2011). Nguyen (2015) performed logistic regression in order to model credit risk in the French banking sector. Furthermore, Crook et al. (1992) examined various sociodemographic and economic discriminators for default prediction among cardholders in the UK.

In the Czech Republic, Kočenda and Vojtek (2011) constructed two credit risk models in order to examine default predictors in retail credit scoring using retail-loan banking data. They compared performance of logistic regression and CART model. They discovered that both methods were comparably efficient. As far as the key determinants of default behaviour are concerned, both models detected similar financial and socio-economic indicators. As a follow-up to this research paper, the main aim of this thesis is to investigate what factors influence probability of default in the Czech Republic. Furthermore, the comparison of suitability and accuracy of various techniques is made. Both traditional credit scoring methods and less conventional approaches are applied and evaluated.

Hypotheses

Hypothesis #1: Client's gender does not affect the probability of default.

Hypothesis #2: Client's age does not affect the probability of default.

Hypothesis #3: Number of client's children does not affect the probability of default.

Hypothesis #4: Level of client's education does not affect the probability of default.

Hypothesis #5: Client's monthly income does not affect the probability of default.

Hypothesis #6: A district in which a client lives does not affect the probability of default.

Hypothesis #7: There is no difference in performance among credit scoring techniques.

Methodology The main objective of this thesis is to investigate default determinants in the Czech Republic. Additionally, the evaluation of various credit scoring methods is performed. In order to conduct this research, it is essential to work with data containing personal information. For this purpose, one of the largest Czech banks provided me with a random sample of its clients' loans. These clients have taken out either a loan for housing or they have consolidated their loans. The dataset was created in March 2017.

The anonymised data include information about 4,000 persons who were granted a loan during the period from November 2006 to March 2017. The variables included in the dataset can be divided into two groups. The first group provides information about the particular loan such as its type, amount borrowed, unpaid balance, interest rate and instalment amount. Specific dates when the loan was taken out and when it will be fully repaid are provided as well. The second group is directly related to the borrower and includes socio-demographic characteristics. The most important variable indicates a default of a client. This means that these borrowers were not able to meet their financial obligations in time. Other personal data contain client's gender, age, level of education, number of children, region of residence and monthly income based on the account transactions. The very last variable indicates a period for which the person has been a client of the bank.

Firstly, I will perform a logistic regression which is a widely used method in credit scoring analysis. The results of this approach can be used for a detection of key default determinants and predictors. Due to such useful interpretation of results, this method belongs to one of the most popular techniques in this area. Furthermore, this method will be used for testing of hypothesis related to personal characteristics. Secondly, I will apply additional classification methods which aim to classify borrowers into bad and good ones as accurately as possible. Finally, various approaches will be evaluated and compared based on their predictive power.

Expected Contribution I will perform a detailed credit risk analysis conducted on real world banking data. Due to general data unavailability, this thesis will supplement a rather limited amount of research papers both in the Czech Republic and Europe. By comparing the results with the existing work which was conducted on a different dataset, this could provide us with the dynamics of the credit risk development in the Czech Republic. Furthermore, the application of additional classification methods could result in more accurate models for default prediction. This might help to better address credit risk in the banking sector.

Outline

1. Introduction: This part will introduce a role of credit risk and credit scoring in the banking sector.

2. Literature Review: I will summarize the previous research related to credit scoring and compare the results of various authors.

3. Data Description: I will present the examined dataset including the detailed feature statistics.

4. Empirical Part: I will present the used methods and build models for default prediction by using the described techniques.

5. Results: I will discuss the results and compare the suitability of methods.

6. Conclusion: I will summarize my findings and their implications for future work.

Core bibliography

Abdou, H. A., & Pointon, J. (2011). Credit scoring, statistical techniques and evaluation criteria: A review of the literature. Intelligent Systems in Accounting, Finance and Management, 18(2-3), 59-88.

Abdou, H., & Tsafack, M. (2015). Forecasting creditworthiness in retail banking: a comparison of cascade correlation neural networks, CART and logistic regression scoring models. The 2nd International Conference on Innovation in Economics and Business ICIEB 2015, February 12-13 2015, Amsterdam, Netherlands.

Bellotti, T., & Crook, J. (2009). Support vector machines for credit scoring and discovery of significant features. Expert Systems with Applications, 36(2), 3302-3308.

Crook, J. N., Hamilton, R., & Thomas, L. C. (1992). A comparison of discriminations under alternative definitions of credit default. In L. C. Thomas, J. N. Crook, & D. B. Edelman (Eds.), Credit scoring and credit control (pp. 217-245). Oxford: Oxford University Press.

Hand, D. J., & Henley, W. E. (1997). Statistical classification methods in consumer credit scoring: a review. Journal of the Royal Statistical Society: Series A (Statistics in Society), 160(3), 523-541.

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). An introduction to statistical learning (Vol. 6). New York: springer.

Kočenda, E., Vojtek, M. (2011). Default Predictors in Retail Credit Scoring: Evidence from Czech Banking Data. Emerging Markets Finance and Trade, 47(6), 80-98.

Kruppa, J., Schwarz, A., Arminger, G., & Ziegler, A. (2013). Consumer credit risk: Individual probability estimates using machine learning. Expert Systems with Applications, 40(13), 5125-5131.

Nguyen, H. T. (2015). Default predictors in credit scoring: evidence from France's retail banking institution. The Journal of Credit Risk, 11(2), 41-66.

Vojtek, M., & Kočenda, E. (2006). Credit-scoring methods. Czech Journal of Economics and Finance (Finance a úvěr), 56(3-4), 152-167.

Author

Supervisor

Chapter 1

Introduction

Credit risk presents the most important risk which needs to be managed by commercial banks. Its appropriate assessment is crucial for banks in order to minimize potential financial losses resulting from defaults on loans. The risk can be mitigated by employing *credit scoring* procedure. Mester (1997) defines credit scoring as a tool for evaluating credit risk of loan applicants. It analyzes historical data in order to extract the effects of various borrower's characteristics which might be useful for predicting the probability of default for new applicants. By performing credit scoring procedure, the whole application process is more efficient.

As the number of loans has raised in past decades, more emphasis is put on accurate performance of credit scoring models.[1] In order to build a precise prediction model, several statistical methods have been investigated for the purpose of creditworthiness assessment. Nevertheless, the results are not identical for all datasets. Since credit scoring models need to work with borrower's personal data, the research into this area is difficult to conduct due to general data unavailability. Hence, the amount of studies which analyze credit scoring models and default predictors using a real-world banking data is limited, especially in European countries.

In this thesis, I analyze a unique dataset of borrowers who have been granted an unsecured consumer loan by one of the largest Czech commercial banks. The objectives of the analysis are twofold. The core part focuses on the investigation of certain borrower's characteristics and their effect on the probability of default. The secondary aim is to compare selected classification methods and their performances in order to find the best credit scoring model for given

[1] Detailed data on the development of consumer credit can be found in ARAD database on `https://www.cnb.cz/docs/ARADY/HTML/index.htm`.

dataset. By conducting this analysis, I complement the existing research of default predictors in the Czech Republic by considering different loan types and extending the list of applied classification methods.

In total, seven hypotheses are proposed considering both default predictors and the comparison of employed classification methods. In the first part, the effects of borrower's gender, age, the number of children, education level, region of residence, and monthly income on the probability of default are examined. In addition, the influence of other features included in the analysis, such as loan amount or the length of borrower's relationship with the bank, is discussed. Secondly, computed performance measures of all models are compared so that the best method for examined loan types can be determined.

The thesis is structured as follows. Chapter 2 presents existing research in the area of credit scoring models and default predictors. Chapter 3 describes the examined dataset. The empirical background behind selected features, descriptive statistics, data preprocessing and transformation procedure prior to analysis are introduced. As for the empirical part, Chapter 4 firstly describes employed classification methods and selected measures of model performance. This chapter is presented separately in order to introduce the conducted approach from the theoretical point of view. Secondly, the actual empirical analysis is presented in Chapter 5. Chapter 6 discusses the results of empirical analysis and provides possible explanation for discovered findings. Finally, Chapter 7 concludes the thesis and presents potential suggestions for future research.

Chapter 2

Literature Review

In this chapter, I present an overview of existing research papers which are closely related to the examined topic. Although the application of credit scoring has been paid much attention in the last decades, the amount of research conducted in this area is not as plentiful as in other fields. It is very likely caused by general data unavailability.

Nevertheless, there are many excellent studies which have presented encouraging findings in credit scoring analysis. Many authors focused on investigating the employment and the comparison of various credit scoring methods and their ability to discriminate between good and bad borrowers based on their characteristics. Since the aim of this thesis is twofold, both studies focusing on credit scoring method assessment and papers investigating potential default predictors are presented.

2.1 Credit Scoring Methods

Since credit scoring is a challenging task for bank's business, the need to develop a suitable model which would classify borrowers as precisely as possible is crucial. For that purpose, a countless number of methods have been studied and their suitability for this particular task has been evaluated. Hand & Henley (1997) reviewed statistical classification methods in credit scoring as a response to its increasing importance.

Another description of frequently used approaches was assembled by Vojtek & Kočenda (2006) who explained the intuition behind selected credit scoring methods. The described methods are linear discriminant analysis, logistic regression, k-nearest neighbor classifier, classification and regression tree, and neural networks. Similarly, Abdou & Pointon (2011) summarized various methods which are frequently used for the purpose of credit scoring. The authors carried out an overview of 214 articles, in which the application of statistical methods was studied, and compared overall results. Moreover, they also introduced evaluation criteria based on which the models could be assessed.

Louzada *et al.* (2016) composed a comprehensive summary of previously conducted credit scoring research. They primarily focused on approaches which have been examined in the existing literature. Furthermore, they compared previously discovered findings. Since the studied literature was published over a long period of time, they also mentioned the changes which had occurred over the years.

2.2 Machine Learning Utilization

Since a classification of borrowers has always been a very important topic for both banks and other lending institutions, many authors considered the suitability of different methods for the purpose of credit scoring. Even though the number of investigated methods is abundant, logistic regression still prevails in credit scoring analysis and it is usually used as a baseline for other methods. Nevertheless, the employment of other approaches has been frequently studied in the last decade, especially machine learning utilization.

As the enhancement of computers and software resulted in more computationally demanding methods to be performed, the number of techniques which might be used for credit scoring significantly increased. One of such machine learning techniques is called k-nearest neighbors. Henley & Hand (1996) investigated the application of this method for credit scoring. This technique was evaluated based on its performance on a real-world dataset. Furthermore, they proposed a modification of the rule based on which the decision line between two classes was determined. Finally, the authors compared KNN with linear and logistic regression, decision trees, and decision graphs. They claimed that KNN classifier performed very well and evinced the lowest expected bad risk rate. Additionaly, the construction of credit scoring model was described by Henley & Hand (1997). As in the previously mentioned paper, the authors emphasized the importance of distance metric selection. The results were evaluated on a sample of applications for mail-order credit and subsequently compared with other techniques.

Classification and regression trees, which are sometimes referred to as decision trees, represent a popular method mainly due to their straightforward interpretation. Feldman & Gross (2005) applied CART algorithm in order to assess its performance on mortgage data. According to the authors, CART was particularly suitable for their data structure. They analyzed a balanced dataset of 3,035 mortgage contracts which were granted during the period from 1993 to 1997 in Israel. The examined dataset included 33 variables. By comparing CART algorithm in relation to logistic regression, discriminant analysis, partial least squares classification, and neural networks, the authors discussed both advantages and disadvantages of this method. Finally, they claimed that borrower's characteristics have a much stronger impact on default probability in comparison to mortgage contract features if accepting a bad borrower is considered to be more costly than rejecting a good one. If both cases are supposed to be equal, mortgage contract features are important as well.

Similarly, Lee *et al.* (2006) analyzed the employment of CART approach on a credit card dataset which included data on 8,000 customers of a Taiwanese bank. Furthermore, they studied the effect of gender, age, marital status, educational level, occupation, job position, annual income, residential status, and credit limits. They discovered that CART demonstrated better credit scoring accuracy in comparison to traditional approaches.

The application of CART approach was investigated by Khandani *et al.* (2010) who analyzed credit bureau data, transaction data, and deposit data. Furthermore, they discovered that the employment of the proposed model would result in cost savings ranging from 6 % to 25 % of total losses.

Another machine learning technique, which has been frequently examined in relation to credit scoring, is called support vector machines proposed by Cortes & Vapnik (1995). Huang *et al.* (2007) studied performance of this classification method while analyzing two real-word datasets. The datasets are available from the UCI Repository of Machine Learning and are comprised of either Australian or German loan applicants. The Australian credit data consist of 690 observations and 15 investigated variables. The German data include information about 1,000 debtors and 24 variables which describe credit history, account balances, or other relevant personal data. Their encouraging findings suggest that the application of this method in credit risk could be beneficial.

By investigating a sample of 25,000 credit card customers, Bellotti & Crook (2009) assessed the performance of support vector machines against logistic regression, KNN, and LDA. The authors concluded that this method could compete with traditional methods which is in accord with Huang *et al.* (2007).

In the last years, neural networks have gained a strong position in the area of machine learning algorithms. They represent a very promising research field and their employment in new areas will be very likely studied in the future. Šušteršič *et al.* (2009) developed a credit scoring model using neural networks. They focused on designing such a model for financial institutions which would be suitable in the event that previously researched data are not available. They used mainly accounting data containing 581 short-term consumer loans granted in the period from 1994 to 1998 and 67 variables. The authors employed principal component analysis. It is an approach which appears very rarely in the related literature. Finally, they made a comparison of the proposed neural network model and traditional logistic regression.

West (2000) investigated the accuracy of various neural network architectures of two publicly available credit datasets from Australia and Germany. After cross-validation testing, the author claims that some models might be used in credit scoring application. Nevertheless, logistic regression still represents the most accurate method.

Contrarily, Ayouche *et al.* (2017) examined a dataset from Morocco which is a developing country. Therefore, credit scoring models might be different from models developed for advanced economies. The studied data included information about 620 applicants and 16 variables. When building a model, the authors considered gender, family situation, age, occupation, present employment, housing situation, status of existing checking account, credit history etc. Finally, the performance of proposed neural network model was compared with traditional credit scoring techniques. The authors claimed that their neural network had outperformed both discriminant analysis and linear regression.

A dataset from another developing country was examined by Blanco *et al.* (2013). They studied the employment of neural networks in order to build a model based on a Peruvian dataset. After comparison with traditional credit scoring methods, their neural network model demonstrated the best performance as opposed to West (2000). Nonetheless, the result is in accord with Ayouche *et al.* (2017).

Contrary to previously mentioned papers in which authors focused only on one particular classification technique, many researchers aimed at investigating more credit scoring methods in their papers. Kennedy (2013) provided a complex analysis of machine learning utilization in credit scoring. He studied the performance of a number of classification techniques using a set of imbalanced credit scoring data. Additionally, he addressed the issue of imbalanced datasets which might be known as a low-default portfolio problem. As the proportion of non-defaulted clients is usually much higher in comparison to defaulted clients, it exacerbates the task of building an accurate classification model. It is, therefore, necessary to take the disproportion into account and adjust the model accordingly. Finally, the author examined the employment of artificial data. This approach might overcome the issue of real-world data unavailability.

Similarly, Kruppa *et al.* (2013) considered the application of machine learning methods in credit scoring area. The investigated methods are random forest, KNN, and bagged KNN. Together with logistic regression, these methods were examined on a dataset of short-termed installment credits. Contrary to other studies, the analyzed dataset does not provide information about typical loans as granted by banks. It originates from such a company selling household appliances for which customers pay in installments. The main advantage of such data is that no credit scoring has been applied prior to granting a loan. The dataset consisted of 64,524 observations. It included both variables about the particular loan and personal borrower's characteristics. After performing a thorough analysis, the authors concluded that random forest algorithm outperformed all other investigated machine learning methods.

Bhatia *et al.* (2017) focused on a different set of machine learning methods. They examined random forest together with LDA, logistic regression, and XGBoost algorithm. The authors put more emphasis on the explanation of intuition behind these methods rather than assessing their performance on a real-world data. Nevertheless, the authors proposed a combination of models estimated by all examined methods in order to create a credit risk scorecard.

Desai *et al.* (1996) addressed the ability of neural networks, LDA, and logistic regression in implementing credit scoring models. They focused on a credit union environment and analyzed data from three such unions. They discovered that the neural network model performed very well in terms of bad loans that were correctly classified in comparison to linear methods. Nevertheless, some limitations of their analysis should be notified, especially examined data structure. The first dataset included information only about teachers and the second one contained data only about telecommunication workers. The last dataset was much more diversified. As all three datasets were used in the analysis, the results might not be applicable to the whole population.

In a very similar way, the performance of neural networks, CART, and logistic regression was examined by Abdou & Tsafack (2015) who analyzed credit data from Cameroon. They compared currently used techniques and proposed more appropriate approach for credit scoring. In order to construct the most accurate model, they compared various quality measurements. The authors emphasized the superiority of neural network model. Additionally, they performed a sensitivity analysis in order to determine which variables were the most significant in creditworthiness forecasting. Even though constructed models differentiate between the importance of particular characteristics, the results suggest that previous occupation, guarantees, borrower's account functioning, car ownership, and loan purpose play a major role in creditworthiness assessment. The same classification methods were employed by Constangioara (2011) who analyzed a sample of Hungarian consumer loans. Moreover, she studied the performance of bagging as a reaction to classification tree sensitivity to data changes. After comparison of all methods, she demonstrated the superiority of bagging estimation.

Another field of machine learning application in credit scoring is the use of ensemble classifiers. This approach is grounded in building individual models and combining their results in a predetermined way. Koh *et al.* (2006) examined this data mining technique on two available Australian and German datasets. Individual classification methods used in the analysis included logistic regression, neural networks, and classification trees. Finally, estimated models determined whether a borrower will default on their loan based on voting. The authors claim that employing such an approach in credit scoring model construction outperforms individual classification decisions. This result is in line with Dahiya *et al.* (2015) and Nanni & Lumini (2009) who analyzed identical datasets. Dahiya *et al.* (2015) constructed an ensemble of seven classification models using confidence-weighted voting. Furthermore, Nanni & Lumini (2009) complemented examined data with additional dataset from a Japanese bank. Although they investigated different individual classification methods, the ensemble of all classifiers also demonstrated the best prediction ability. The overall results show that the employment of ensemble might significantly improve the performance of credit scoring models.

As can be seen from previously mentioned studies, the ambiguity of their findings shows that there is not any classification method which significantly outperforms the other methods. The suitability of particular classification method varies with examined datasets and it cannot be determined beforehand which method will be the most appropriate for the particular credit data.

2.3 Default Predictors in European Countries

The aim of this thesis is not only a comparison of various classification methods, but also the discovery of crucial default predictors. As real-world credit data are usually not publicly available, it is difficult to conduct adequate analysis which would determine the effect of selected borrower's characteristics on the probability of default. Nonetheless, some authors focused their attention primarily on discovering potential default predictors rather than on assessing the suitability of selected classification techniques. In the following section, such research papers analyzing real-world European data are presented.

As this topic has always been very important for both banks and other lending institutions, the research into the role of personal characteristics in default prediction started a long time ago. For that purpose, Crook *et al.* (1992) examined a sample of 1,001 borrowers who were supposed to repay their credit card loan in the United Kingdom. In order to estimate the model, they applied discriminant analysis and concluded that zip code of debtor's residence was the most significant predictor.

The situation in Greece was examined by Ganopoulou *et al.* (2013). They analyzed a sample of consumer loans during the period 2007–2009 gathered from a Greek commercial bank. Overall, 11 variables were included in the analysis which were related both to clients themselves and to their loan. As a reaction to the financial crisis occurring during the observed period, the sample was divided into three parts in order to take the crisis effect into consideration. Contrary to majority of related research papers, the authors estimated a binary probit model instead of a traditional logistic model. The results suggest that debtor's gender, age, and income affect the probability of default in the most significant way. The fact that a debtor lives in a big city and owns a house decreases the probability of default. Additionally, they estimated also a model analyzing which variables influence the probability of being granted a loan. Therefore, they were able to compare the results of both models and analyze bank's decision-making process in more detail.

Similarly, Roszbach (2004) considers not only the probability of defaulting on a loan, but also the probability of being granted a loan.[1] In order to conduct such an analysis, he examined a Swedish dataset. Hence, the author was capable of comparing both events and modeling their relationship. In a

[1] Roszbach (2004) also estimated a tobit model in order to predict the period after which the borrower will default on their loan.

similar way as Ganopoulou *et al.* (2013), a probit model was constructed in order to determine the most significant default predictors. The results show that income, change in income, marital status, and existence of a guarantor significantly influence the probability of default in Sweden.

Default predictors in the Czech Republic were studied by Kočenda & Vojtek (2011) who, to the author's knowledge, have conducted the only publicly available analysis using real-world Czech banking data. As a reaction to a credit increase in European emerging countries, they focused on credit risk evaluation since it has not been paid much attention until then. The dataset included 3,403 observations about clients who were granted a mortgage during 1999–2006. Two models were estimated using logistic regression and CART algorithm. The advantage of both methods is that the effect of each variable can be easily extracted from the results. Both models yielded similar results in terms of feature importance. The results show that amount of resources, education level, marital status, loan purpose, and the length of a relationship with the bank appear to be the strongest default predictors. As for default prediction, the authors discovered one additional encouraging finding. After constructing a model in which the amount of resources was not included, they concluded that the prediction ability of such a restricted model was almost identical in comparison to the full model.

On the contrary, Nguyen (2015) investigated default predictors in the advanced European economy. He provided an evidence from French retail banking sector by examining a sample which included both application and behavior data. The author constructed various models in order to estimate the probability of default of all automobile loans in this bank. Interestingly, this model was implemented by the bank in practice. Apart from including available variables, interactions were generated and considered which appears in the credit scoring literature very rarely. The results of all models were compared using different quality measures. Finally, the author selected the most important variables based on their information value. As far as the personal characteristics are concerned, the most significant default determinants are marital status, residential status, and job type.

Chapter 3

Data Description

In this chapter, I comment on the empirical background behind the examined dataset, its origin and all necessary modifications that needed to be performed. Firstly, I introduce characteristics which are typically used in order to determine creditworthiness of a new loan applicant. Additionally, the comparison of applicant's characteristics which were used in various research papers is included. Secondly, data preprocessing procedure which was performed in order to clean the data is presented. Furthermore, I introduce the structure of the data. A detailed description of each variable which is included in the dataset is presented in order to determine general features and characteristics. This might provide us with some insight into the relations in the sample and indicate who the typical loan applicant is. Finally, the investigated data are transformed in such a way that they are suitable for building credit scoring models as suggested by Thomas *et al.* (2017).

3.1 Empirical Background

As credit scoring is the most important tool for assessing new applicants' creditworthiness, it is essential for a bank to have as much information about potential borrower as possible.

Since the main aim of this thesis is to investigate how borrower's characteristics influence the probability of default on a loan, it is crucial to examine a sample of borrowers including personal information. Unfortunately, not all necessary data are always available. Due to data unavailability, the set of variables is usually restricted, and therefore it is not always feasible to include all desired features. This might be also influenced by legislation in some countries as there exist laws which prevent particular personal characteristics from being included into credit scoring, such as Equal Opportunity Credit Act (1976). Nevertheless, some conclusions may be drawn from existing research about which characteristics should be included in the analysis.

Hand (2001) points out that the characteristics are ordinarily a combination of continuous and categorical variables. The example of *continuous variable* is e.g. age, income, or time at present employment. On the other hand, education, type of employment or marital status can be mentioned as examples of *categorical variables.* As for the information provided by various variables, Vojtek & Kočenda (2006) differentiate among demographic indicators, financial situation, employment status, and behavioral indicators.

To the author's knowledge, there is not any theory that would precisely state which variables should be included in the analysis. The reason for this might be that the significance of various variables is not identical across all countries, analyzed loan types or even methods used. Thus, no general rule can be deduced because it is not clear which variables are significant for the analysis before creating the model. This is in line with Abdou & Pointon (2011) who reviewed 214 articles addressing credit scoring application. The authors claim that none of the researches presented theoretical reasons for selecting particular variables.

Nevertheless, Abdou & Pointon (2011) summarized characteristics typically used for credit scoring and frequently appearing in the existing research. These are gender, age, marital status, number of dependents, education level, occupation, time at present address, having a telephone, and having a credit card. Additionally, information about time at present job, loan amount and duration, house owner, monthly income, purpose of loan, bank accounts, having a car, or time with a bank might be included as well. This list of variables is also in line with Hand & Henley (1997) who presented features typical for credit scoring. Besides, the authors mentioned zip code of a residence as another characteristic for credit scoring.

Table 3.1 shows the list of variables included in selected research papers.

Ngyuen (2015)	**Ganopoulou (2013)**	**Kočenda & Vojtek (2011)**	**Desai (1996)**
Personal Characteristics: - Marital Status - Age - Region - Residential Status - Time at Present Address - Number of Children - Time at Present Employment - Time at Bank **Loan Characteristics:** - Type of Credit - Vehicle Type - Vehicle Condition - Vehicle Price - Loan Amount - Loan Duration - Client's Contribution-to-Vehicle Price Ratio - Client's Initial Contribution - Existence of a Co-borrower - Financial Situation - Income - Debt-to-Income Ratio - Troubled Debt Restructuring **Other Behavioral Variables** - Number of Months in Bucket 0 - Max Bucket in the Past 12/6/3 Months - Number of Months of Non-payments in the Past 12/6 Months - Delinquency Status	**Borrower's characteristics:** - Age - Gender - Annual Income - Marital Status - Nationality - Region - Type of Residence **Loan Characteristics:** - Loan Duration - Loan Status - Warranty **Other Variables:** - Other Banking Transactions - Other Loans	**Socio-demographic variables:** - Education - Marital Status - Years of Employment - Sector of Employment - Gender - Date of Birth - Type of Employment - Number of Employments - Employment Position - Credit Ratio 1 - Credit Ratio 2 - Region **Bank-Client Relationship Variables:** - Own Resources - Amount of Loan - Purpose of Loan - Length of the Relationship - Date of Account Opening - Deposit Behavior - Loan Protection - Type of Product - Number of Co-signers - Date of Loan	- Age of Borrower - Owns home - Number of Years in Current Job - Number of Dependents - Number of Major Credit Cards - Salary plus Other Income - 'Good' depending upon Derogatory Information and Number of 01-09 - Ratings on Credit Bureau Reports - Number of Inquires in Past 7 Months - Number of Months since Trade Line Opened - Number of Trade Lines 75% Full - Delinquent Accounts in Past 12 Months - Total Debt as a Proportion of Income - Number of Open Accounts on Credit Bureau Reports - Number of Active Accounts on Credit Bureau Reports - Number of Previous Loans with Credit Union - Information based upon 01-09 Ratings on Credit Bureau Reports

Table 3.1: Examples of analyzed variables in selected research papers

3.2 Origin of Data

For the purpose of analyzing default determinants and comparing various credit scoring methods, one of the largest Czech commercial banks provided me with a unique sample of granted loans to its customers. As the bank does not want to be identified, I respect its request and I will not state its name. The provided dataset has been anonymized so that it is not possible to identify any specific person in the sample. Nevertheless, the data are confidential and cannot be published even as an attachment to this thesis.

The sample contains *unsecured* consumer loans which were granted to clients who have either taken out a loan for housing or who have consolidated their loans. The dataset was created in March 2017 and all included data are as of this date. Before any adjustments, the sample is originally comprised of 4,000 loans which were granted to clients in the period from November 2006 to March 2017.

Initially, it included in total 20 explanatory variables which can be divided into three categories: borrower's personal characteristics, loan-related variables and variables describing bank-client relationship. The first category includes gender, age, zip code of borrower's residence, monthly income, family status, the number of children, education level, household income, and household expense. The second category contains loan type, loan amount, actual balance, interest rate, date of loan granting, date of last installment, date of next installment, and installment amount. The last category provides information about the duration of borrower's relationship with the bank and whether borrower's partner is the bank's client as well. Nonetheless, not all of them are included in the empirical analysis due to specific reasons described in Section 3.3.

3.3 Data Preprocessing

Data preprocessing is a crucial part of conducting any research and might be even more difficult than the analysis itself. Before the analysis, it is necessary to prepare the data so that they are in the appropriate form. Hence, a thorough data preprocessing was performed in order to prepare the dataset. Furthermore, some variables were completely removed from the dataset as their inclusion would not make sense and might distort the final results.

Firstly, the values of age and the borrower's relationship with the bank, also referred to as *tenure profile*, were computed as of the date when the loan was granted because the original dataset included values as of the March 31, 2017 when the dataset was created. This was easily calculated since the date of loan granting is provided. Nevertheless, the value of age needed to be rounded to the whole number because the date of birth was not available. As far as tenure profile is concerned, the number of months originally provided is rounded, and therefore a few values fall into interval between -1 and 0. It is very likely that such borrowers became clients at the time of loan granting and they did not have any relationship with the bank prior to loan application. Nevertheless, 6 observations included values which are even lower than -1. These low values could not be explained by rounding. It seems that there has been an error in the data since such low values do not make sense. Therefore, these observations were removed from the sample.

Secondly, *zip code* variable providing information about the district of borrower's residence was transformed into a region so that it is possible to compare the differences across the Czech Republic. This was performed in order to reduce the number of categories for this variable. Additionally, the most of the zip code categories would include a very low number of observations. In total, 37 missing entries were removed from the data set. They were either missing or the zip code could not be traced back and transformed into the region in the Czech Republic. These untraceable zip codes denoted clients who live abroad, for instance in Slovakia, Poland, Austria or France. Similarly, 14 clients did not provide information about their education level and were removed from the sample.

Afterwards, I decided to restrict the values of monthly income and the year in which the loan was granted due to the following reasons. Firstly, the bank does not have information about client's monthly income but it estimates its value based on financial transactions on client's bank account. Unfortunately, this does not assure that this is indeed borrower's monthly income since they might have a different bank account to which the income is credited. As a consequence, some values were extremely low since they probably resulted only from credited interest. Therefore, I decided to consider this variable as a *proxy* for income. In order to do that, it was necessary to remove these low values. The threshold was set to 11,000 CZK which is the value of minimum wage as of January 1, 2017 according to Ministry of Labour and Social Affairs (2017).

Secondly, loans which were granted in 2017 were removed from the sample. As this dataset was created at the end of March 2017, no borrower could have possibly defaulted on their loan as the period of non-payment for default event needs to be at least 90 days.

Additionally, I considered which variables should be included in the further analysis since the original dataset contains some features which do not provide complete information or were provided only as a complementary piece of information. As for loan-related variables, actual balance, interest rate, and installment amount were removed because they are not usually known at the time of application for a loan. Specifically, interest rate and installment amount are derived after credit scoring is performed.

Moreover, household expenses and household income were calculated as a sum of positive and negative financial transactions on bank accounts of a client and their spouse. If a spouse is not a bank's client, the total household income was calculated as a value of borrower's positive transactions. As the number of spouses who are bank's client is very low, I do not consider this variable to be a suitable proxy for household financial situation.

The final dataset which is examined henceforward includes 2,420 observations and 11 explanatory variables. Although the sample size decreased, I assume that it was a necessary step in order to clean the data and increase their quality for the research.

3.4 Descriptive Statistics

In this part, I provide description of all variables included in the final dataset which characterize each loan and a person who has taken it out. As previously mentioned in Section 3.2, these variables are related to borrowers themselves, or they describe the particular loan and the borrower's relationship with the bank. More specifically, they provide information about borrower's gender, age, the number of children, education level, region of residence, and monthly income. The remaining variables are loan amount, loan type, loan year, loan duration, tenure profile, and whether a borrower's partner is a bank's client. Finally, the core variable for the further analysis indicates whether a client has defaulted on a loan or not, and therefore it is introduced as the first one.

3.4.1 Default

Default represents the most important piece of information since it is the explained variable in the analysis. It describes a situation when a borrower is not able to meet their financial obligations in time which might result in significant losses. For the sake of accuracy, I quote the official definition of default according to the Czech National Bank (2017, p.137) which is stated as follows:

> “Default is defined as a breach of the debtor’s payment discipline. The debtor is in default at the moment when it is probable that he will not be able to repay his obligations in a proper and timely manner, without recourse by the creditor to settlement of the claim from the security, or when at least one repayment (the amount of which deemed by the creditor to be significant) is more than 90 days past due.” - CNB Financial Stability Report 2016/2017, page 137

For the purpose of the analysis, this variable is a binary indicator which equals 1 in case of default and 0 otherwise. In the examined dataset, there are 546 clients who were not able to meet their financial obligations towards the bank in time. This accounts for approximately 22.5 % of all borrowers. Nevertheless, this does not correspond to the frequency of defaults on these loans in reality. Due to confidentiality reasons, it is not possible to publish the exact frequency but the number of bad loans is virtually negligible and it would not be possible to conduct proper analysis. As the main aim is to train a model which would predict defaults as precisely as possible, it was necessary to examine such data on which the model can be trained in the most efficient manner.

3.4.2 Gender

It might be possible that borrower’s *gender* influences the probability of default, although banks should take a cautious approach to its inclusion in credit scoring models in order to avoid potential accusation of discrimination. Nevertheless, the variable is included in the analysis in order to reveal true underlying relationship.

The statistics provided in Table 3.2 summarize values which are related to loan amount and default frequency according to gender. Hereinafter, N denotes the total number of observations, whereas D describes the number of defaults for particular group. The observed values suggest that women borrow lower amounts of money and less frequently in comparison to men. This might be caused by the fact that husband usually signs the contract for a loan in case of married couples. Nevertheless, data on marital status are not available for these loan types, and therefore this conjecture cannot be empirically tested. The ratio of defaulted loans based on gender is approximately 3:5. Overall, 20.4 % of female debtors defaulted on their loans whereas for male debtors this percentage equals 24.0 % which is 3.6 % higher.

Gender	Mean	Median	Std. Dev.	Min	Max	N	D
Female	265,744	210,000	170,762	49,001	1,000,000	987	202
Male	285,847	250,000	169,934	50,000	1,000,000	1,433	344

Table 3.2: The descriptive statistics of loan amount (CZK) and default frequencies according to gender

3.4.3 Age

The variable presents information about *borrower's age* at the time when the loan was granted. The detailed description is summarized in Table 3.3. The distribution of borrower's age in the sample can be seen in Figure 3.1.

The mean age equals 39.1 years and the range of values is relatively wide. The youngest person in the sample is 18 years old whilst the oldest person is 74 years old. The distribution is slightly skewed to the right. This tendency to borrow money at rather lower age might suggest that people do not want to incur debts when they are about to retire or have already retired. By taking out a loan, they would need to repay the debt for a certain period of time in the future which might be more difficult in later life.

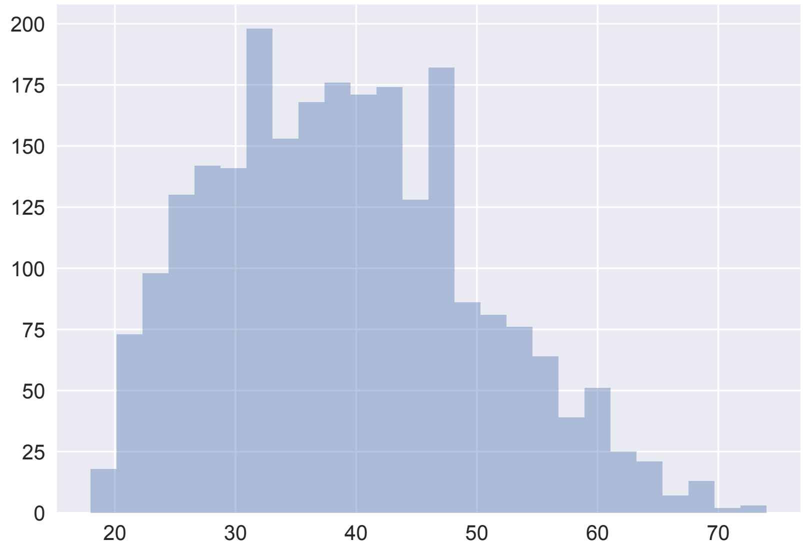


Figure 3.1: The distribution of borrower's age

Variable	Mean	Median	Std. Dev.	Min	Max
Age	39.1	38.0	10.76	18	74

Table 3.3: The descriptive statistics of borrower's age

3.4.4 Number of Children

The *number of children* might affect the probability of default as parents usually provide their children with substantial financial support. The more children a person has, the higher amount of expenses is probably required for covering needs of their children. Therefore, this might considerably reduce the amount of money which is available for monthly installments.

This variable provides information about the number of borrower's children the bank knows about. As statistics in Table 3.4 indicate, the majority of the dataset consists of childless debtors. It can be seen also in Figure 3.2 which presents the distribution of this variable. Although it is not exceptional to have rather fewer children, the number of childless persons in the sample is surprisingly high. It might be closely related to the age distribution in the sample since nowadays people tend to delay parenthood. On the other hand, having two or more children is very rare. Only two borrowers in the sample have the maximum number of four children.

Variable	Mean	Median	Std. Dev.	Min	Max
Children	0.38	0	0.62	0	4

Table 3.4: The descriptive statistics of the number of children

3.4.5 Education Level

The examined dataset includes information about borrower's educational background. This might be a very significant factor which is frequently used in credit scoring models, e.g. in Dinh & Kleimeier (2007). Borrowers in the sample attained either primary education, secondary education, vocational education, higher vocational education or they obtained a university degree. One additional category includes all other education types which could not be included into other education levels.

The number of defaults for each education level together with the frequency of each category in the sample is summarized in Table 3.5. The distribution is plotted in Figure 3.3.

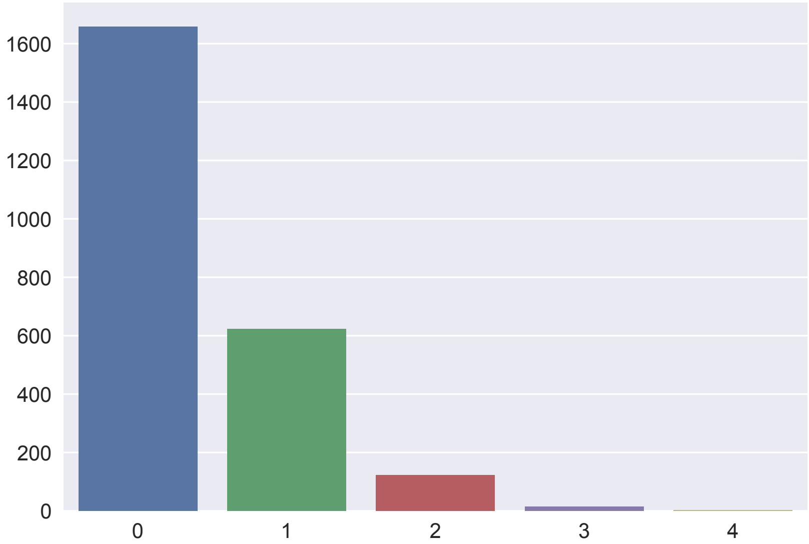


Figure 3.2: The distribution of the number of children

As can be seen, borrowers who attained secondary education comprise the most plentiful category. On the other hand, higher vocational education is the least frequently observed education category in the sample. Furthermore, clients who have obtained university degree borrow the highest loan amounts. This could be explained by the fact that the university graduates are assumed to find a job for which they get paid a high salary. Therefore, they might be confident about certainty of repaying the loan as they have a sufficient monthly income at their disposal for installment payments.

As far as default rates are concerned, debtors who attained vocational education evince the highest default frequency. It equals more than 31 %. Conversely, this frequency is almost twice as low for university graduates with value of 15 %. This is consistent with previously mentioned surmise about graduates' confidence while repaying their debts.

Education	**Mean**	**Median**	**Std. Dev.**	**Min**	**Max**	**N**	**D**
Primary	199,101	150,000	117,196	50,000	525,000	69	21
Secondary	287,803	250,00	173,350	50,000	1,000,000	909	190
Vocational	253,791	200,000	155,716	49,001	850,000	694	218
Higher Vocational	276,107	270,000	144,838	55,000	500,000	28	5
University	323,811	300,000	193,761	60,000	1,000,000	295	43
Other	275,698	230,000	174,316	50,000	800,000	425	69

Table 3.5: The descriptive statistics of loan amount (CZK) and default frequencies for various education levels

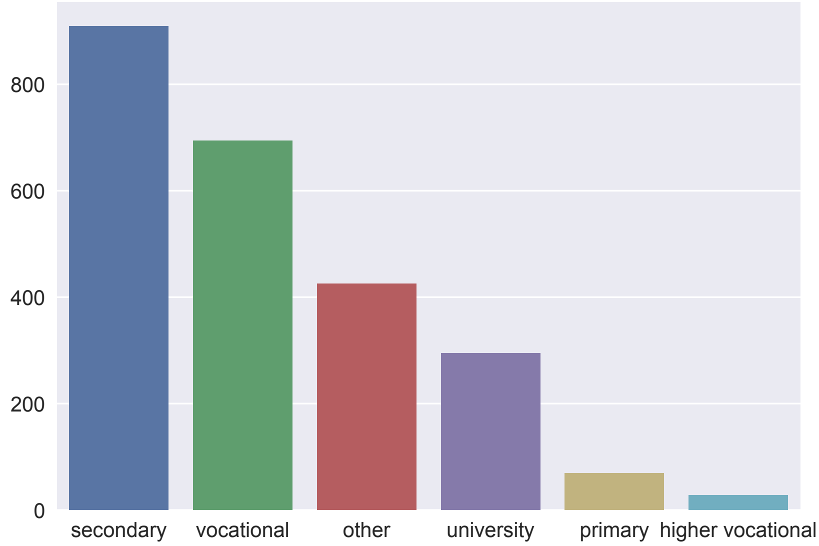


Figure 3.3: The distribution of education levels

3.4.6 Region of Residence

This generated variable might be assessed in order to compare the probability of default across the Czech Republic. It might be possible that people living in certain regions tend to default more frequently on their loans in comparison to other parts of the country. The summary statistics and the distribution of borrowers in the sample are provided in Table 3.6 and Figure 3.4, respectively.

As can be observed, Prague is the most plentiful category in the data. Furthermore, the loan amounts are the highest as living in the capital city is usually more expensive in comparison to other regions due to higher prices. Mean values for all regions ranges from 236,250 CZK for Pardubice region to 313,820 CZK for Prague.

As far as default frequencies in particular regions are concerned, the highest proportion of debtors defaulted in Liberec region. The percentage equals approximately 31 %. Conversely, only 15 % defaulted in Vysočina region.

3.4.7 Monthly Income

Financial resources are closely connected with the probability of default on a loan as discovered by Kočenda & Vojtek (2011). *Monthly income* can be used as an alternative indicator of financial situation. As previously mentioned in Section 3.3, the bank estimates the value of income based on financial transactions occurring on client's bank account. Hence, this variable was further modified so that it can be used as a proxy for income.

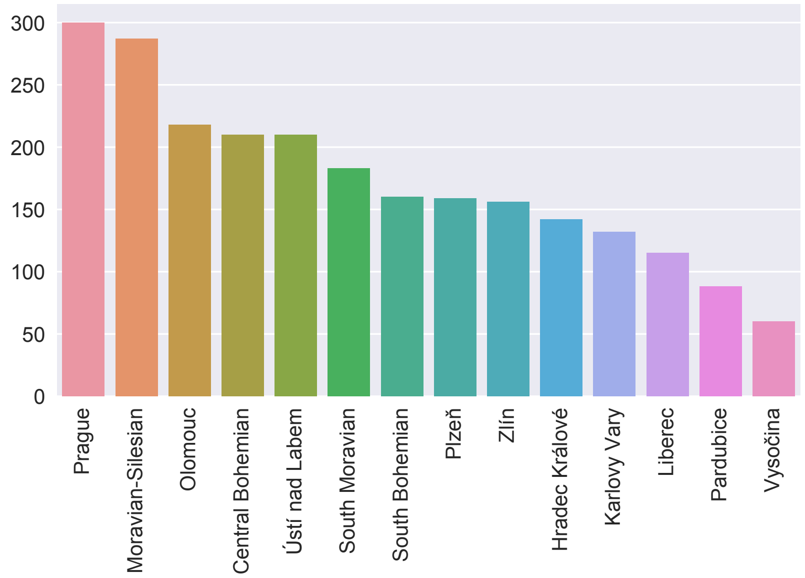


Figure 3.4: The distribution of regions of borrower's residence

A detailed description of monthly income variable is summarized in Table 3.7. Additionally, the distribution of borrower's income can be seen in Figure 3.5. The lowest income equals 11,000 CZK which is the threshold used for excluding low values. Conversely, there are few outlying observations. This can be seen in Figure 3.5a which shows a distribution with a significant heavy right tail. Therefore, for the sake of clarity, Figure 3.5b captures only a part of the complete income distribution.

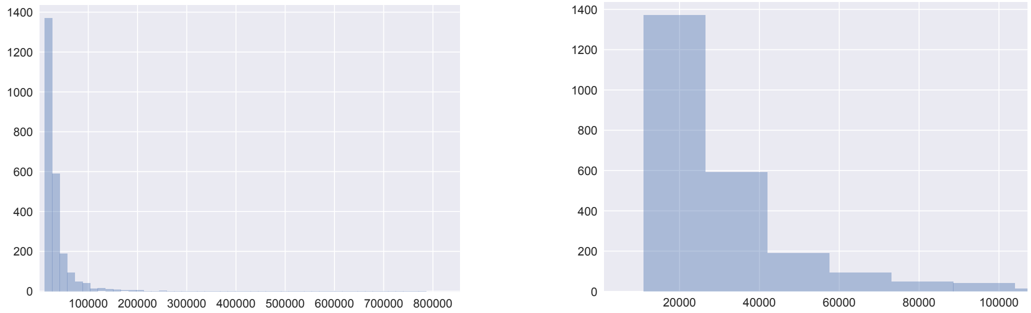
As the income is determined based on account transactions, it might be possible that these persons might have done some profitable businesses which affected their income in the examined month. Nevertheless, it still reflects borrower's financial situation. The mean value equals 34,721 CZK but might be very likely affected by unusually high monthly incomes in the sample. Thus, the median value is more suitable for assessing the average monthly income since it is not influenced by outlying observations. It equals 24,128 CZK.

3.4.8 Tenure Profile

Tenure profile denotes for how long period of time, measured in months, a borrower has been a bank's client. This might be a relevant variable for building credit scoring models because the bank has very likely much more information about potential borrowers and their credit history in comparison to new applicants. It might be possible that longer history with the bank implies lower

Region	Mean	Median	Std. Dev.	Min	Max	N	D
Prague	313,820	280,000	182.643	50,000	820,000	300	55
Moravian-Silesian	266,180	220,000	167,741	49,001	895,000	287	74
Olomouc	270,517	213500	171,233	60,000	1,000,000	218	47
Ústí nad Labem	280,283	239,500	168,182	58,000	1,000,000	210	50
Central Bohemian	284,449	257,500	175,238	53,000	1,000,000	210	54
South Moravian	262,937	225,000	151,220	50,000	980,000	183	38
South Bohemian	275,536	208,000	174,815	50,000	900,000	160	40
Plzeň	268,886	230,000	158,224	58,000	730,000	159	28
Zlín	265,993	230,000	170,987	50,000	1,000,000	156	38
Hradec Králové	273,589	220,000	165,373	50,000	800,000	142	31
Karlovy Vary	280,674	239,000	167,552	60,000	800,000	132	25
Liberec	307,856	250,000	180,614	50,000	800,000	115	36
Pardubice	236,250	200,000	159,850	50,000	800,000	88	21
Vysočina	254,324	229,000	158,731	50,000	695,000	60	9

Table 3.6: The descriptive statistics of loan amount (CZK) and default frequencies according to the region of borrower's residence



(a) The complete distribution

(b) The partial distribution

Figure 3.5: The distribution of monthly income (CZK)

probability of default. As a consequence, the bank does not need to collect such detailed data on borrower's characteristics during the application process since it has already have this information.

The results in Table 3.8 suggest that the majority of the sample consists of new clients. Nevertheless, those who have already been bank's clients have established relatively long-term relationship with the bank before taking out this type of loan. The average duration equals more than 8 years. During that time, the bank probably gathered a lot of data about client's behavior which could be used in the analysis of their creditworthiness prior to loan granting. The distribution can be seen in Figure 3.6.

Variable	Mean	Median	Std. Dev.	Min	Max
Monthly Income	34,721	24,128	40,704	11,000	786,576

Table 3.7: The descriptive statistics of borrower's monthly income (CZK)

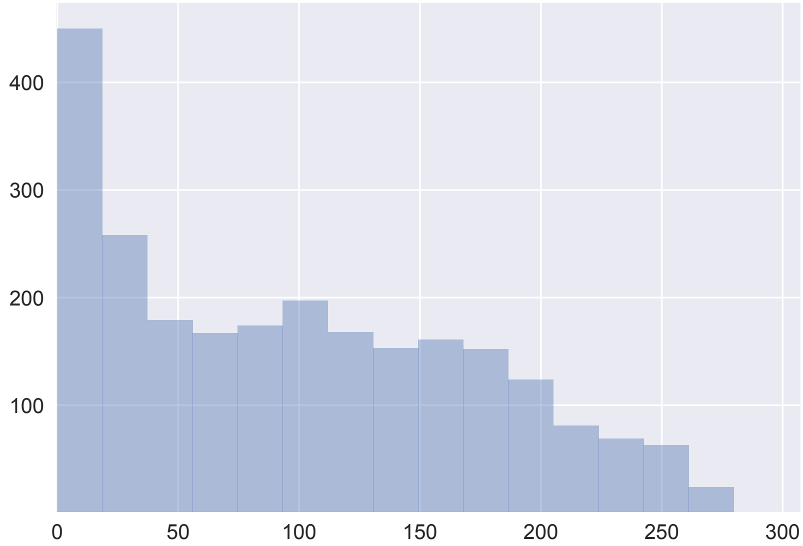


Figure 3.6: The distribution of borrower's tenure profile (months)

3.4.9 Loan Amount

The amount of money borrowed probably influences the fact whether a debtor is able to fully repay their debt as the loan might represent a major financial burden for a significant period of time. The summary of loan amounts which were granted to bank's customers is provided in Table 3.9. Furthermore, the distribution of loan amounts can be seen in Figure 3.7. The plotted graph shows that people borrow rather lower amounts of money as the distribution is skewed to the right. The values at the right tail of the distribution appear in the sample only rarely. Mean value equals 277,648 CZK and median value is 235,000 CZK. These values follow from the nature of the granted loans which is discussed in the following section.

Variable	Mean	Median	Std. Dev.	Min	Max
Tenure Profile	98	91	74	0	280

Table 3.8: The descriptive statistics of borrower's tenure profile (months)

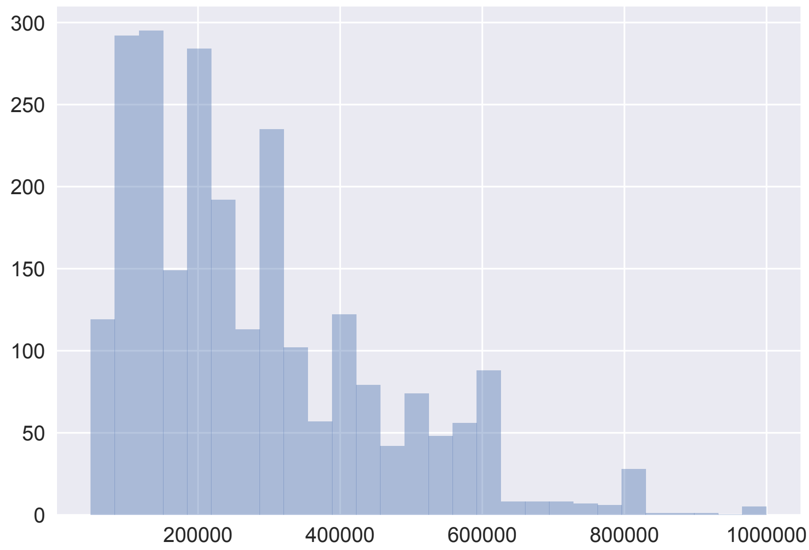


Figure 3.7: The distribution of granted loan amount (CZK)

Variable	Mean	Median	Std. Dev.	Min	Max
Loan Amount	277,648	235,000	170,559	49,001	1,000,000

Table 3.9: The descriptive statistics of granted loan amount (CZK)

3.4.10 Loan Type

The examined sample consists of two types of consumer loans. The first type is defined as a loan for housing which is unsecured and clients do not need to use any collateral. This loan type is suitable for small house improvements, reconstructions, or furniture purchases.

Although potential applicants need to deliver specific documents during the application process, it can be relatively easily obtained compared to other loan types such as mortgage. Since the amounts borrowed are not that high and the potential financial loss resulting from default is limited, the data collected by the bank during the application probably do not include such a detailed list of applicant's characteristics in comparison to mortgages.

The second loan type denotes loan consolidation which enables clients to refinance their existing debt by merging several loans into one loan and possibly negotiate better conditions.

The descriptive statistics about both loan types are summarized in Table 3.10. Approximately 45 % of the sample is comprised of loans for housing, whereas a little less than 55 % of borrowers have consolidated their loans. The similarity of observed values for both loan types suggests that it was meaningful to include these types in the sample of consumer loans in order to increase its size without causing possible distortions. According to mean and median values, the amounts of granted loans tend to be rather lower which is closely connected with loan purposes.

Loan Type	Mean	Median	Std. Dev.	Min	Max	N	D
For Housing	264,644	220,000	160,484	49,001	1,000,000	1,098	216
Consolidation	288,449	242,500	177,773	50,000	1,000,000	1,322	330

Table 3.10: The descriptive statistics of loan types (CZK)

3.4.11 Loan Year

In the sample, there are various dates which are related to the granted loan. This might provide us with information whether the probability of default varies over the examined period and loans granted in some particular year are riskier than the others. Figure 3.8 shows the number of loans according to the year in which they were taken out. The distribution follows from the loan purpose since these are short-term loans. Therefore, loans granted during last 5 years dominate in the sample.

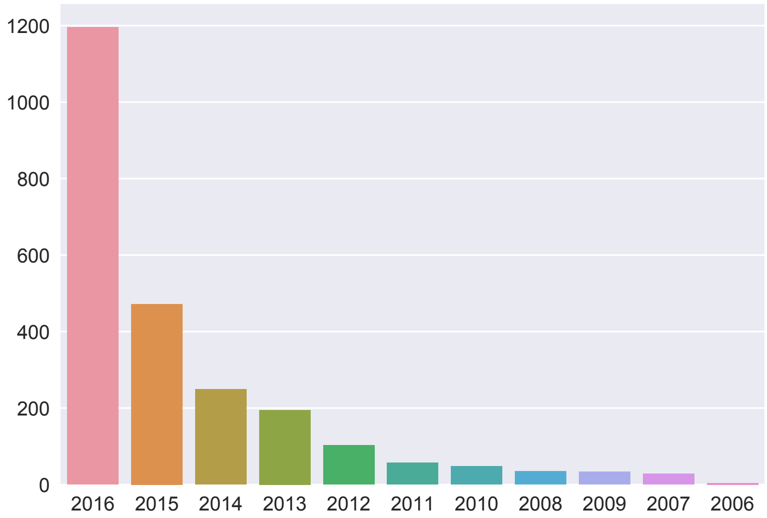


Figure 3.8: The distribution of loans according to years

3.4.12 Loan Duration

As loans pose a large financial burden for debtors, the period of loan repayment might influence the probability of default. This would mean that loans with higher duration could be riskier. On the other hand, the installment amount might be lower as the repayment is spread over a longer period of time. Therefore, monthly income is reduced by a lower amount. The distribution of loan duration is plotted in Figure 3.9. It suggests that people borrow for rather a longer period of time.

The descriptive statistics summarized in Table 3.11 show that the average loan duration equals approximately 8 years which is almost identical to the median value. The maximum duration in the sample equals almost 14 years. These are surprisingly high values considering that the sample includes short-term loans.

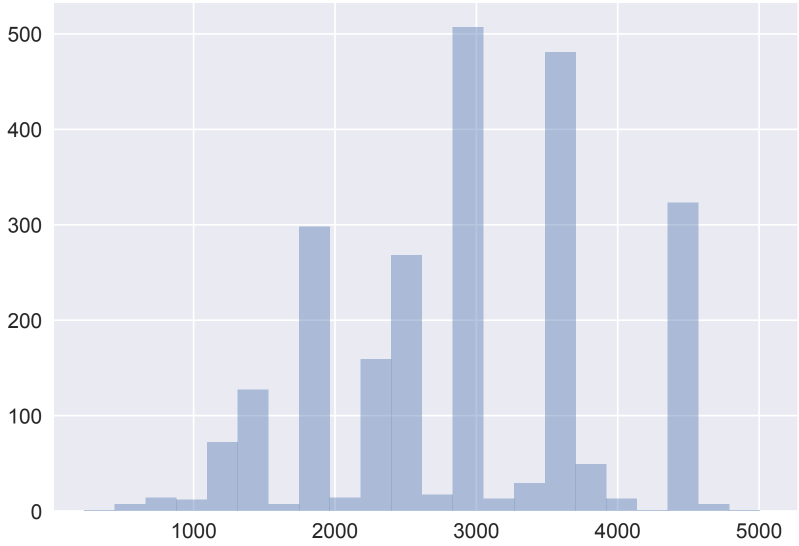


Figure 3.9: The distribution of loan duration (days)

Variable	Mean	Median	Std. Dev.	Min	Max
Loan term	2,916	2,938	949	225	5,006

Table 3.11: The descriptive statistics of duration of a granted loan (days)

3.5 Data Transformation

This section introduces data transformations performed to the majority of included variables so that they can be used in model building as suggested by Nguyen (2015), Anderson (2007), or Thomas *et al.* (2017). After preprocessing and cleaning the sample, there are some adjustments which are typically made in the area of credit scoring before the actual analysis. The procedure is called *coarse classing* of explanatory variables. This was applied to all variables in the dataset except for *gender*, *region*, and *partner client* variables as they are already in the proper form.

The aim of coarse classing is to construct classes which are very similar in terms of risk. Furthermore, coarse classing allows calculating so-called *information value* also for non-categorical variables (hereinafter referred to as IV). Based on IV, it is possible to identify features which can discriminate between good and bad applicants.

This modification technique is described by Thomas *et al.* (2017) who emphasize the necessity of splitting the characteristics into a rather small number of classes. The authors state two arguments for such modification depending on whether the variable is continuous or categorical. As for categorical variables, there might be too many categories and some of them might have insufficient number of observations if there are very rare. In case of continuous variables, this modification allows the risk to be non-linear which might provide us with better prediction. Additionally, Hand (2001) points out that the advantage of classing is its easy interpretability as it can be stated that one class is riskier than the other one.

The procedure is performed in two steps. The first step is referred to as *fine classing*. In case of continuous variables, the values are divided into at most ten intervals which contain approximately the same amount of observations. In case of categorical variables, Anderson (2007) suggests creating classes for all possible values. Hence, each class corresponds to one particular category observed in the sample.

After creating all classes for both categorized and continuous variables, IV statistic is computed. It measures the level of ability of a variable to discriminate between good and bad borrowers and is defined as

$$IV = \sum_i \left(\frac{g_i}{g} - \frac{b_i}{b}\right) \log\left(\frac{g_i b}{b_i g}\right)$$

where g denotes the number of all good borrowers and b stands for the number of all bad borrowers in the sample. Lower index i denotes the number of good, respectively bad, borrowers in the particular class. The logarithm is also known as a *weight of evidence*.

High values of IV suggest that the variable is more beneficial in differentiating between good and bad borrowers. After computing IV for all characteristics in the sample, bad rates for each category are determined.

The second step is called *coarse classing* and it describes the process of merging fine classes into a smaller number of classes for each variable. The first objective is to have a sufficient amount of observations in the classes which ensures that the model is stable. The second objective is to have as few coarse classes as possible which represent particular characteristics. This should prevent the model from overfitting as a lower number of variables is considered. Moreover, similar bad rates should be taken into account when merging classes. Naturally, the final number of classes for each variable differs depending on the structure. After creating specific coarse classes, the examined dataset is completely prepared for modeling procedure. The detailed list of all categorized variables can be found in Appendix A.

Chapter 4

Methodology

In this chapter, I present an overview of classification methods, which are used in the analysis, and introduce the principles on which they are based. Specifically, these methods include *logistic regression, linear discriminant analysis, quadratic discriminant analysis, classification tree, random forest, k-nearest neighbors*, and *support vector machines.* Since logistic regression appears to be the most frequently employed method in the existing research, it is used as a primary model for result interpretation.

Moreover, measures of model performance employed in this thesis are described. Firstly, a *confusion matrix* is presented as it provides the intuition behind employed measures. Secondly, *receiver operating characteristic curve* and corresponding *area under the curve* are introduced.

In the following section, the concept of *cross-validation* which ensures proper testing of model quality is introduced. Finally, the approach of finding the best model concludes this chapter.

4.1 Classification Methods

Generally, there are many classification methods which can be used for the purpose of credit scoring. Most of them have already been investigated and assessed, nevertheless, some are examined more frequently in comparison to other ones.

In this thesis, I study performance of several selected techniques. The aim was to choose a variety of both traditional and less frequent classification methods in the area of credit scoring. As the most suitable classification model cannot be determined beforehand, the objective was to select seemingly different methods. Classification methods described in this section can be divided into two groups: parametric and non-parametric.

Logistic regression, linear discriminant analysis, and quadratic discriminant analysis are representatives of parametric methods. As James *et al.* (2013) describe, these methods are based on predetermined assumptions about the form of the function which is used for capturing the relationship between the explained variable and explanatory variables. In addition, the number of parameters is fixed before training and does not depend on the amount of observations.

On the contrary, classification trees, random forest, k-nearest neighbors, and support vector machines belong to non-parametric approaches. In comparison, non-parametric methods are not restricted by such assumptions at all. The form of the function expressing the relationship between explanatory variables and explained variable is not known. Conversely, these methods aim to seek an estimate of this function which is as close to the true relationship as possible. Furthermore, the structure of the model depends on training data and cannot be determined beforehand.

4.1.1 Logistic Regression

Based on the existing research, logistic regression is the most popular method in credit scoring analysis. Its performance was assessed for instance by Nguyen (2015), Kočenda & Vojtek (2011), Kruppa *et al.* (2013), or Dinh & Kleimeier (2007).

As Anderson (2007) mentions, it has overcome shortcomings of linear regression which was originally used for classifying new applicants for a loan. Over the last period, it has become a traditional classification method in many fields as stated by Hosmer Jr *et al.* (2013).

This method estimates the relationship between binary dependent variable and a set of explanatory variables. Therefore, it is suitable for the purpose of credit scoring. Instead of directly classifying a person into a particular group, it estimates the *probability* of being a member of such a group.

Logistic regression uses a specific function for probability modeling. Having a set of p explanatory variables denoted $\mathbf{x}$ and the probability of default $\pi(\mathbf{x})$, logistic regression is based on the following function:

$$\pi(\mathbf{x}) = \frac{e^{\beta_0+\beta_1 x_1+\beta_2 x_2+\ldots+\beta_p x_p}}{1+e^{\beta_0+\beta_1 x_1+\beta_2 x_2+\ldots+\beta_p x_p}}$$

After adjusting the original formula, we arrive at the resulting equation for logistic model which is characterized by so-called *logit*:

$$\ln\left(\frac{\pi(\mathbf{x})}{1-\pi(\mathbf{x})}\right) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + ...\beta_p x_p$$

The main consequence of this definition is that logit is a linear function of explanatory variables. The assumption of linearity is very strict and might not be optimal for all datasets. Coefficients denoted by β are estimated by applying maximum likelihood estimation. It is an iterative computational process which searches for such coefficients which maximize the log-likelihood function. This process was initially very demanding as for computer capacity and might have taken a long period of time. Nowadays, this does not present such an issue as computational power has significantly improved during last decades.

Additionally, the results of logistic regression can be easily interpreted in comparison to other machine learning techniques which have been used in credit scoring analysis, such as deep neural networks or SVM. Therefore, this technique belongs to the most popular ones and is usually used as a baseline for other approaches. For further details see Hosmer Jr *et al.* (2013) or Friedman *et al.* (2001).

4.1.2 Linear Discriminant Analysis

Linear discriminant analysis (hereinafter referred to as LDA) used to be, and still is, a very popular in the area of classification tasks. It was presented by Fisher (1936) who aimed at differentiating between separate groups using linear combination of available variables.

The objective of this technique is to minimize the differences within particular group while maximizing the differences between groups. This is reached by finding a linear combination of explanatory variables which maximizes the difference between both groups. Furthermore, using this linear combination functions as a dimension reduction tool. Nonetheless, details on the precise mathematics behind this method are beyond the scope of this thesis and can be found in Friedman *et al.* (2001).

The core assumption of LDA is that the observations within each category come from a normal distribution. This is a very strong assumption which might not always hold in practice. Additionally, all categories are assumed to have an identical covariance matrix.

Training of a model is based on estimating a precise shape of normal distribution while taking a linear combination of features into consideration. In order to do this, covariance matrix, means of both categories, and coefficients defining the linear combination need to be estimated.

After estimating normal densities for both categories, every new observation is classified based on the probability of coming from a particular estimated normal distribution. It is possible to differentiate between groups by drawing a linear decision line using their fitted normal distributions. Unfortunately, it is very susceptible to any violation of its assumptions, and therefore it has been slowly replaced by logistic regression. According to Anderson (2007), logistic regression is preferred also because it works better in cases of imbalanced data which are very common in practice.

4.1.3 Quadratic Discriminant Analysis

The limitations of strict assumptions of LDA can be overcome by employing a slightly modified approach called *quadratic discriminant analysis* (hereinafter referred to as QDA). If we relax the assumption of equal covariances of both categories, we might arrive at a quadratic discriminant function. Covariances for each category are estimated separately which allows for different shapes of density function.

Compared to its linear predecessor, it is more flexible as it allows for more complex decision boundaries. Therefore, this technique might fit the data better. Nevertheless, its main disadvantage is that more parameters need to be estimated during training as opposed to LDA.

According to Friedman *et al.* (2001), if the examined dataset is diverse and has a large number of observations, both LDA and QDA proved to be very helpful and performed well in classification tasks.

4.1.4 Classification Tree

Classification tree, sometimes also referred to as a decision tree, is a nonparametric method which is based on completely different principles compared to previously described approaches. It was introduced by Breiman *et al.* (1984) and has become a frequently examined technique in the area of credit scoring. Due to its estimation setting, it is sometimes called recursive partitioning algorithm.

This technique recursively splits the original data into different subsets and identifies the resulting category membership depending on what the majority in the particular set is. In the first step, the dataset is divided into two parts which become more homogeneous in comparison to the original dataset. Secondly, these newly created parts are further split into another two parts resulting in even more homogeneous subsets. This step is recursively performed as long as all terminal nodes, also known as *leaves*, are pure. It means that all leaves contain only one category and no majority voting is required.

According to Anderson (2007), there are several rules when growing a classification tree for credit scoring purpose. Firstly, it is necessary to determine how to make continuous predictors discrete. Secondly, predictors which will be included in the tree have to be chosen. In order to grow the classification tree, a *stopping rule* which prevents the tree from creating new splits must be defined. This needs to be done, otherwise a very complex classification tree with pure leaves would be grown. This would lead to overfitting of the model. After reaching this rule, the algorithm terminates even though some leaves include observations of both classes.

Furthermore, so-called *pruning* may be performed when the tree becomes overly complex. This approach is applied after the tree is grown. It removes nodes which do not provide much information. As a result, pruning decreases tree complexity, and therefore forestalls overfitting. As far as other regularization techniques are concerned, one can restrict the depth of a tree prior training. This can be done by defining for instance a maximum number of leaves or minimum number of observations in each node.

As for main advantages of this technique, Anderson (2007) mentions its ability to discover underlying patterns in the data, transparency and easy implementation. James *et al.* (2013) point out that classification trees better reflect the process of human decision-making. In addition, it might be suitable for cases when the relationship between features and the predicted variable cannot be expressed by a linear model.

On the contrary, it is susceptible to substantial overfitting if the tree contains a large number of nodes which have only few observations. Nevertheless, this can be prevented by using a proper regularization technique as previously mentioned. Furthermore, it is not that robust as the final tree can be significantly influenced by a small change in the data.

4.1.5 Random Forest

Random forest utilizes classification trees in order to provide a more powerful prediction tool. Firstly introduced by Breiman (2001), the employment of random forest algorithm has resulted in improvements in classification accuracy.

When building a random forest, a higher number of trees is constructed and each of them votes for a resulting category. The prediction is then made based of averaging their individual votes. Individual classification trees are build using bootstrapped training samples. It means that each tree is grown using a different subsample of the original dataset.

While growing these trees, a random sample of m predictors is selected from the set of p predictors when creating a new node. It means that the algorithm considers only a part of all available predictors at each split. This procedure has a main advantage. If there was a feature in the sample having a very strong predictive power, the majority of individual trees would assign this feature at the top node. Therefore, all trees would resemble each other and their predictions would be highly correlated.

Random forest overcomes this problem by considering only a random subset of available predictors. Thanks to this algorithm setting, random forests can be seen as a collection of de-correlated classification trees as pointed out by Friedman *et al.* (2001). Theoretical aspects of this methods are beyond the scope of this thesis, nevertheless, an interested reader might found them in Breiman (2001) or Friedman *et al.* (2001).[1]

4.1.6 K-Nearest Neighbors

This approach is very simple and intuitive. Nevertheless, it might work very well for some datasets as shown by Henley & Hand (1996) and Henley & Hand (1997). It determines a category membership of a test observation by analyzing observations in its neighborhood.

Firstly, it identifies k observations in the training sample which are closest to this test observation. For measuring closeness, it is possible to use various metrics. Anderson (2007) mentions Euclidean distance, a square root of the sum of the squared differences, or City-block distance which is a sum of the absolute differences. After finding closest observations, the resulting prediction is based on averaging their responses.

[1] Even more complex versions of random forest algorithm have been proposed. One such example is rotation forest introduced by Rodriguez *et al.* (2006).

The main advantage of KNN is that no prior training of the model is needed. Once a new observation needs to be classified, it directly analyzes its neighbors and predicts a particular class. Hence, no parameters need to be estimated.

On the contrary, each prediction is very demanding in terms of time. The algorithm needs to compute the distance for all observations in the dataset in order to determine the closest ones which are taken into account during the classification. Furthermore, this method is not convenient for data in which classes are scattered and no clusters are presented.

4.1.7 Support Vector Machines

Support vector machines (hereinafter referred to as SVM) were proposed by Cortes & Vapnik (1995). In order to explain the rationale behind this method, it is necessary to introduce two approaches from which SVM is derived. They are called *maximal margin classifier* and *support vector classifier*. The common feature of all three methods is that they use a separating hyperplane in order to separate the training observations in the feature space. Any test observation can be then classified based on which side of the hyperplane it is located.

Let's assume we have data which can be divided into two classes using a separating hyperplane. This results in a linear decision boundary between classes derived from this hyperplane. Nevertheless, there exists an infinite number of such hyperplanes which draw a perfect boundary between classes. Each separating hyperplane might be shifted a little bit and still produce a correct boundary. The objective is to find the best separating hyperplane out of all possible ones.

Maximum margin classifier aims to find such a separating hyperplane that perfectly discriminates between classes and is the farthest from training data. This distance is called *margin* and its maximization suggests improved accuracy for testing data as pointed out by James *et al.* (2013). All training observations of which distance from the boundary is exactly the value of margin are called *support vectors*. These are the core observations based on which the hyperplane is fitted.

Nevertheless, the perfect linear decision boundary might not be always found. Furthermore, every outlying observation might substantially change the resulting decision boundary. In order to solve this problem, it is possible to intentionally classify some observations incorrectly in order to improve robustness of the classifier. This is done by support vector classifier which still aims to maximize the margin but some observations might violate the decision boundary. The algorithm takes into consideration the number of incorrectly classified observations and their distance from the boundary during the estimation. This procedure is very complex and beyond the scope of this thesis.[2]

It is possible that the observed classes are not linearly separable. SVM is an extension of support vector classifier by allowing a non-linear boundary between classes which makes it very popular for classification tasks. It is possible to extend the feature space by using a non-linear transformation function in order to accommodate the possibility of non-linear boundary using a separating hyperplane. Hence, non-linear transformation of the feature space allows for a non-linear decision line while using a linear classifier.

As the computation of a transformation function is very demanding, SVM uses *kernel functions*[3] which reduce computational demands during training and make the procedure more efficient. Nevertheless, the SVM algorithm is very complicated and it can be solve as an optimization problem using advanced quadratic programs. Since the algorithm is beyond the scope of this thesis, more details can be found in Friedman *et al.* (2001).

The rationale behind SVM somewhat resembles the idea of KNN classifier by considering few key observations. The main advantage over KNN is that SVM does not need to consider all observations during prediction. Once the model is trained, it remembers individual support vectors and every new observation is classified based on already fitted boundary. Unfortunately, it might take a long period of time to perform due to its complexity. Another disadvantage is that it does not provide much information about the patterns in the data as its primary objective is classification.

4.2 Model Performance Measures

After employing previously described classification methods, the next step is to assess their performance. Specifically, it is necessary to compare how precisely the resulting model is able to differentiate between good and bad borrowers. The aim is to identify which approach is the most appropriate for the investigated dataset and given loan types.

There are many statistics which can be used for computing discrimination

[2] An interested reader may consult for example Friedman *et al.* (2001).

[3] There are many types of kernel functions such as linear kernel, radial basis function kernel, sigmoid kernel etc.

power of the model e.g.*accuracy, F-score, precision*, or *recall.* Accuracy is one of the most frequent performance indicators which measures a total percentage of all correctly classified observations. Nonetheless, this is not appropriate for imbalanced data as even very bad model can produce relatively good results.

In this thesis, I follow the approach of Kočenda & Vojtek (2011) since they conducted a similar analysis examining Czech mortgage data. In order to measure the model quality, they computed *receiver operating characteristic curve* and *area under curve.* These are known under the acronyms ROC curve and AUC, respectively. Using identical performance measure, it is possible to identify differences between credit scoring models used for mortgage loans and credit scoring models for ordinary consumer loans. Before introducing ROC curve and AUC, *confusion matrix* is presented as it is closely connected to these core performance measures used in the analysis.

4.2.1 Confusion Matrix

Confusion matrix is one of the most convenient ways to evaluate classification model performance. It summarizes the results of specific models by inspecting their predictions. Thus, it might tell us exactly how the model performs and whether it does not favour one class over the other one which is usually a problem of imbalanced datasets. It is presented in this section since it provides crucial information for determination of a performance measure which is used in this thesis.

An example of a confusion matrix for a binary classification task is presented in Table 4.1. More specifically, it is applied to case of credit scoring. Bad borrowers are considered to be a positive class, whereas good borrowers presents a negative class.

		Predicted	
		Good	**Bad**
Actual	**Good**	True Negative	False Positive
	Bad	False Negative	True Positive

Table 4.1: Confusion matrix for bad and good borrowers

Rows correspond to the actual class, whereas columns capture what model actually predicts. In each entry, the number of observations belonging to the specific group is written. Entries on a diagonal denote correctly classified observations. Specifically, *true negative* box captures the number of correctly classified good borrowers. *True positive* box includes the number of correctly classified bad borrowers. On the contrary, *false positive* box denotes the number of good borrowers which were predicted to be bad. Similarly, *false negative* box states how many bad borrowers were classified as good.

Using the numbers in all boxes, one can calculate many performance measures in order to summarize the information provided by a confusion matrix. Most importantly, it is necessary to introduce *sensitivity* and *specificity* statistic measures as they are closely connected to the performance measure used for model comparison in this thesis.

Sensitivity, also known as *true positive rate*, measures the probability of predicting default given the actual state is default. Using confusion matrix, it can be expressed as $\frac{TP}{TP+FN}$. Similarly, specificity or *true negative rate* measures the probability of predicting non-default given the actual state is non-default. This can be expressed as $\frac{TN}{TN+FP}$.

4.2.2 ROC Curve

When classifying a new applicant, it is necessary to set a threshold for the probability of default which is used as tool for prediction. Although 0.5 is usually used as a default value for many models as pointed out by pMüller & Guido (2016), it is possible to change it accordingly and put more emphasis on particular class.

Applying concepts of sensitivity and specificity, one can assess model performance constructing *receiver operating characteristic curve*. In this thesis, it is constructed for all classification methods in order to compare their performance on the investigated dataset.

Using these two measures, ROC curve is a plot of $1-$*specificity* on the x-axis against *sensitivity* on the y-axis for various values of threshold used as a decision rule. Hence, it shows the relationship between false positive rate and true positive rate derived from a confusion matrix. In other words, it captures the trade-off between two types of error for different thresholds. An example of ROC curve can be seen in Figure 4.1.

A straight dashed diagonal line represents a random model which serves as a baseline. In order to have a high-quality model, the objective is to construct ROC curve which is above the diagonal line and as far as possible.

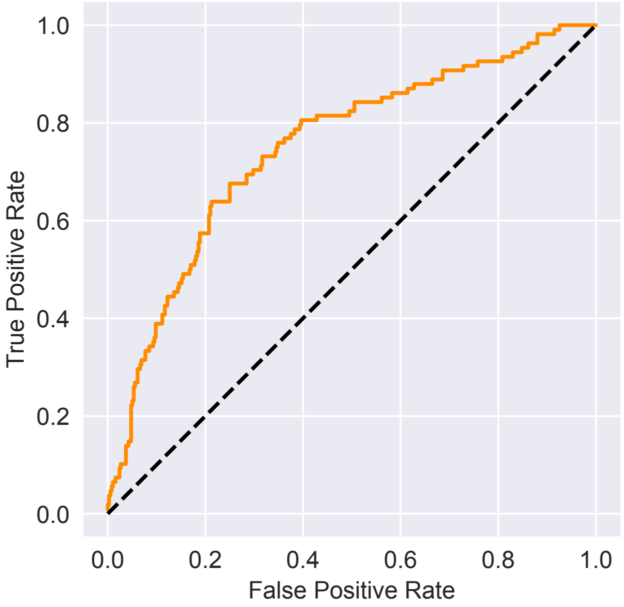


Figure 4.1: Example of ROC curve

4.2.3 AUC Statistic

Although ROC curve is a popular performance measure, it might not be always feasible to compare various ROC curves based on their graphs and conclude which curve and how much it is better compared to the other ones. Therefore, the overall performance can be expressed as the total *area under the curve* which is usually referred to as AUC or $AUROC$. It is computed as a definite integral.

The ideal value equals 1, whereas the value of 0.5 suggests that the model is equally as good as a random guess denoted by a diagonal line. Computing this statistics suggests which method is the most suitable for given dataset.

According to Abdou *et al.* (2016), model performance based on AUC values can be interpreted as follows from Table 4.2.

AUC Value	**Model Performance**
$AUC < 0.6$	fail
$0.6 \leq AUC < 0.7$	poor
$0.7 \leq AUC < 0.8$	fair
$0.8 \leq AUC < 0.9$	good
$0.9 \leq AUC$	excellent

Table 4.2: Performance assessment based on AUC value

4.3 Cross-Validation

When training a model, it is possible to reach a perfectly accurate performance on the training data, nevertheless, it might produce very bad results when applied to a different dataset. Therefore, the objective is to assess the model performance when previously unseen data are used. This approach is based on splitting the sample into training set and testing set.

Firstly, the model is trained on *training* dataset. In order to evaluate how accurately the specific classification method is capable of discriminating between two or more classes, it is given previously unseen data called *testing* dataset. By employing this approach, the results are more trustworthy and might be used for overfitting detection when the model performance on testing dataset is substantially worse in comparison to training dataset.

This might be applied to the case of credit scoring as the objective is to assess the creditworthiness of new applicants. Therefore, the objective is to train a model based on previous defaults and to identify features which might be relevant. Afterwards, the model assesses new applicants without using their data during model building.

In order to assess the overall performance of the model, it is necessary to generalize its results as the performance on the original sample split would very likely be exceptionally good. Hence, the approach employed by Kočenda & Vojtek (2011) or Kruppa *et al.* (2013) is based on splitting the sample into training and testing sets multiple times and evaluating the performance individually for each sample split.

Kočenda & Vojtek (2011) applied 1,000 repetitions, whereas Kruppa *et al.* (2013) used 2,000 repetitions but it is closely connected to the sample size. Taking the sample size of the examined dataset into account, I employ 1,000 repetitions. After computing the value of selected performance measure for each split, 95% confidence interval can be estimated assuming a normal distribution of the results.

By performing this bootstrap technique, the results are more trustworthy since they are not connected only to one particular sample split on which the model was trained. This approach is performed for all selected classification techniques. Additionally, it can be used for comparing the performance and accuracy of all methods.

4.4 Hyperparameter Tuning

After deciding on the employment of a specific classification technique and its performance measure, the aim is to specify the model properly so that it produces the best results. While training a model, each method has a specific set of so-called *hyperparameters* which determine the way models are trained. These hyperparameters might attain a number of values. Therefore, it is difficult to identify the optimal set of hyperparameters for which the model performs the best as there is a large number of combinations. Nevertheless, it can be solved by performing so-called *grid search.* During this procedure, all possible combinations of predetermined hyperparameters are applied one by one and the model is trained for each combination. At the same time, selected performance measures are computed. Hence, it is possible to identify the best set of parameters based on selected measures for the specific sample split.

Nevertheless, these hyperparameters are selected on the basis of its performance on one particular sample split. Therefore, it does not assure that the final set of hyperparameters is optimal for a different sample split. When studying a different sample split, it is expected that the performance of the final model will be worse in comparison to the sample split on which the original grid search was performed. If a grid search procedure was performed on the other sample split, it would probably yield a different set of optimal hyperparameters.

Chapter 5

Empirical Analysis

In this chapter, I describe the employed approach for conducting an empirical analysis in order to examine the effect of personal characteristics on the probability of default. As the aim of the thesis is twofold, the emphasis is not put only on the results of one particular econometric model, but also on the methods used for the analysis and their performance comparison.

Firstly, I define hypotheses to be tested including the expected effect of various debtor's personal characteristics on the probability of default and the suitability of employed methods. Secondly, I identify characteristics which have the highest predictive power, and hence they are selected for model building. Finally, econometric models are presented together with their results. Their performances are compared in order to conclude whether some methods are more suitable for the purpose of credit scoring than the others.

Before the actual analysis, it should be pointed out that modeling the probability of default is generally a difficult task. Bellotti & Crook (2009) mentions that credit data are not easily separable. Hence, the rates of misclassification ranges usually from 20 % to 30 % as suggested by Baesens *et al.* (2003). This would be considered as a poor result in other areas. Furthermore, the examined dataset includes data about borrowers who have been granted a loan. This means that the bank has already performed an advanced credit scoring procedure during application process. Nevertheless, no other data could be used for the analysis.

The complete analysis was conducted using Python 3 programming language and many applied functions were inspired by Müller & Guido (2016). More specifically, machine learning package called *scikit* introduced by Pedregosa *et al.* (2011) was used during model training.

5.1 Hypotheses

This section presents hypotheses to be tested in the thesis. As the results of logistic regression can be easily interpreted, they will be used for hypothesis testing. Furthermore, each hypothesis is complemented by an expected outcome supported by existing research in this area. All hypotheses are stated in a negative form so that they can be rejected. Therefore, they do not express a probable outcome.

As findings of Kočenda & Vojtek (2011) suggest, financial resources play a significant role while assessing credit risk as it decreases the probability of default. Therefore, the effect of personal characteristics might be reflected in the probability of default through its relationship with monthly income.

Hypothesis 1: *Client's gender does not affect the probability of default.*
According to findings of Kočenda & Vojtek (2011), gender is not a good default predictor. Although this effect applies to mortgage loans, it might be expected that the effect is similar for unsecured consumer loans. Therefore, there should not be any difference in the probability of default between male borrowers and female borrowers. This might be connected to the fact that some countries prohibited banks from including gender into credit scoring models.

Hypothesis 2: *Client's age does not affect the probability of default.*
The effect of age might be influenced by two facts. Firstly, it might be assumed that with increasing age, a person might get a better job as they gather more working experience. Therefore, their financial environment might improve over time and the probability of default would decrease. On the other hand, it might be possible that from a certain moment, person's income is considerably reduced and financial situation worsens. Moreover, it might be assumed that the relationship is not linear and might change over certain age categories. This assumption would suggest that age is a significant factor that influences borrower's creditworthiness. This would be in line with findings of Constangioara (2011) or Bellotti & Crook (2009). On the contrary, Kočenda & Vojtek (2011) cannot confirm such an effect of age on the probability of default in the Czech Republic.

Hypothesis 3: *The number of client's children does not affect the probability of default.*
The number of children a borrower has might negatively influence the probability of default. It might be expected that the more children a person has, the higher level of financial support is required for covering their needs. Therefore, lower amount of money is at borrower's disposal for repayments. This assumption is in accord with the findings of Crook *et al.* (1992) who discovered that childless people have the lowest probability of default. Nevertheless, they found out that the risk of default was not linear as for the number of children. On the contrary, people who have children might feel more responsible for their future, and therefore do not wish to incur debts. This would suggest their enhanced creditworthiness.

Hypothesis 4: *The level of client's education does not affect the probability of default.*
According to Kočenda & Vojtek (2011), education level is one the most important socio-demographic default predictors. As in the previous cases, this might be linked to its relationship with income. Thus, higher gained education might imply lower probability of default as it may be assumed that more educated people occupy better-paid employments. Kočenda & Vojtek (2011) discovered that borrowers with a university degree had the lowest probability of default. This was found also by Constangioara (2011) claiming that the effect of education is one of the strongest.

Hypothesis 5: *Client's monthly income does not affect the probability of default.*
It is very likely that a stable financial situation enhances borrower's creditworthiness. Having a sufficient amount of money at disposal implies that monthly repayments do not account for such a large proportion of income. Therefore, the financial burden is not that large as in cases of lower incomes. In the case of the Czech Republic, Kočenda & Vojtek (2011) claim that the amount of money a borrower owns is the most important default predictor. It is very likely that this effect prevails also for unsecured consumer loans.

Hypothesis 6: *The region in which a client lives does not affect the probability of default.*
This hypothesis claims that there is no region in the Czech Republic in which

borrowers are more susceptible to default on their loans in comparison to other ones. If we assume that the living standard is approximately identical across regions, this claim might be very likely true. Nevertheless, according to Czech Statistical Office (2017), there is almost 11,000 CZK gap between average wages in Prague and Karlovy Vary region. This might suggest that some regions suffer from higher probability of default as a loan presents a large financial burden in comparison to their resources. Kočenda & Vojtek (2011) did not find any evidence for supporting this claim since a region of borrower's residence evinces low predictive power as for default estimation. On the contrary, Crook *et al.* (1992) identified that a place of residence was one of the most important features while predicting the probability of default in the United Kingdom.

Hypothesis 7: *There is no difference in performance among applied credit scoring techniques.*
According to James *et al.* (2013), no statistical method is superior to other techniques over all possible datasets. This is in accord with Hand & Henley (1997) who claim that there is no general best technique. It depends on data structure, characteristics used, examined loan types etc. Therefore, all methods used in this thesis should produce comparable results. Nevertheless, James *et al.* (2013) also state that it is possible that some method dominates on a particular dataset but, on the other hand, some other methods may produce better results on a similar but different dataset. Hence, selecting the most suitable technique for particular dataset presents a challenging task.

The aim is to compare performances of applied methods in order to determine whether some model produces significantly better results in comparison to other ones or whether all techniques are identically efficient on the examined dataset.

5.2 Feature Selection

After transforming original variables into classes as described in Chapter 3, the number of variables usually considerably increases. Since they usually enter into the model as binary variables, this might be difficult for some techniques to cope with that many attributes. Hence, the objective is to reduce the number of explanatory variables while maintaining all relevant features.

In order to select the most appropriate characteristics, Thomas *et al.* (2017) suggest that variables which evince low predictive power should be eliminated.

This can be done by using the statistic which was computed and used during data transformation phase. Computing IV provides us with crude orderings of variables according to their importance as for predictive power.

After calculating values of IV for each variable, it is necessary to determine the threshold for ruling out bad predictors. Eliminating features with low predictive power ensures that all included variables are relevant and bring valuable information into the analysis.

According to Siddiqi (2006, p.81), based on the value of IV, variables can be classified as described in Table 5.1.

IV Value	Predictive Power
$IV < 0.02$	unpredictive
$0.02 \leq IV < 0.1$	weak
$0.1 \leq IV < 0.3$	medium
$0.3 \leq IV$	strong

Table 5.1: Assessment of predictive power based on IV statistic

Table 5.2 summarizes predictive power based on the value of IV statistic for all variables in the investigated dataset.

Variable	IV
Monthly Income	0.3650
Education	0.1304
Loan Year	0.0955
Age	0.0674
Loan Duration	0.0666
Tenure Profile	0.0508
Region	0.0423
Loan Amount	0.0273
Number of Children	0.0237
Loan Type	0.0230
Partner Client	0.0161
Gender	0.0100

Table 5.2: Predictive power of analyzed variables based on IV statistic

The results show that there is one strong predictor, one medium predictor and eight weak predictors. Kočenda & Vojtek (2011) included such variables whose IV equals at least 0.1. Nevertheless, only two variables would pass this value. Hence, I decided to set the threshold equal to 0.02 which is the boundary for weak predictors. Such a low threshold was selected in order to include as many personal characteristics as possible. Since the aim is to study the effect of borrower's characteristics, they would not be incorporated into the model if the threshold was higher.

It can be seen that monthly income evinces the greatest ability of discriminating between good and bad borrowers. The same finding was discovered by Kočenda & Vojtek (2011). Other socio-demographic characteristics seem to be less important in comparison to variables related directly to a loan or relationship with a bank. Nevertheless, the level of education is an important default predictor apart from client's monthly income. This is also supported by Kočenda & Vojtek (2011).

As far as unpredictive features are concerned, gender variable and the fact whether a borrower's partner is also a bank's client do not bring any valuable information into the prediction. Hence, they were eliminated from the sample. The final list of features which are included in the models contains monthly income, education level, loan year, tenure profile, loan duration, region of residence, age, loan amount, the number of children, and loan type. All categories of all final characteristics are in a form of dummy variables. It means that if a client belongs to a specific category, it equals 1 for a corresponding dummy variable. To summarize, the final list contains a combination of both socio-demographic and loan-related characteristics.

5.3 Model Building and Training

This section presents all models which were built and trained together with their performance assessment statistics. Firstly, the dataset was divided into training and testing datasets in a ratio of 4:1. As the number of observations in the dataset is not as extensive as in other studies, the training set needs to be sufficiently large so that the model can be trained. The proportion of defaults in both samples remains approximately identical.

As for the number of defaults, the dataset is imbalanced. This can be remedied by putting more emphasis on defaulted clients in order to equalize the amount of bad and good borrowers. Therefore, class weight is one of defined hyperparameters while training a model if this option is available.

As mentioned in Chapter 4, the best combination of corresponding hyperparameters for each model was found by performing an extensive grid search procedure. Models which produced the best results on the testing dataset are presented together with resulting hyperparameter setting.

5.3.1 Logistic Regression

As was already mentioned in Chapter 4, logistic regression appears to be the most frequently employed classification method in credit scoring. Hence, it is used as the primary model in this thesis. As its coefficients can be easily interpreted, final conclusions about the effects of personal characteristics on the probability of default are drawn from its results.

Firstly, a model which includes all previously selected variables is constructed. In order to prevent dummy variable trap, one reference category is selected for each characteristic. After estimating model coefficients, z-scores and p-values are computed for all variables in order to test whether a specific coefficient is equal to 0.

Secondly, as the threshold for selecting variables is relatively low, possibly insignificant features are identified and eliminated so that the model is not misspecified and produces more reliable results. Additionally, determination of the set of eliminated variables is supported by Python 3 feature importance algorithm which sorts variables according to their importance in the model.[1] Taking both calculated p-values and importance ranking into consideration, a second model which includes a reduced set of variables is built and trained. Both models are compared using ROC curves and AUC statistics.

While training a logistic model, there are many hyperparameters which can be adjusted so that the model produces the best results. More specifically, I considered three hyperparameters which appeared to be the most important ones during training. These are *shrinkage method selection, tuning parameter*, and *class weight.*

James *et al.* (2013) describe shrinkage method as a technique which constraints the value of estimated coefficients toward zero. The main advantage of this approach is that it decreases variance and prevents the model from overfitting. It can be thought as a form of regularization. The two best-known shrinkage methods for linear models are *ridge regression* and the *lasso.* Both methods are based on adding some additional argument, called *shrinkage penalty*, into the estimated equation which states the rule for estimated coefficients. Ridge regression considers the sum of squares of β coefficients, whereas lasso technique is based on the sum of absolute values of β coefficients.

In both cases, a shrinkage penalty argument is multiplied by so-called *tuning*

[1] More information regarding feature importance algorithm can be found on `http://scikit-learn.org/stable/modules/generated/sklearn.feature_selection.RFE.html#sklearn.feature_selection.RFE`

parameter which is a second hyperparameter used in the training of a logistic model. The tuning parameter adjusts the impact of this regularization. Therefore, it is possible to set the level at which the estimated coefficients are regularized.

The last hyperparameter of a logistic model adjusts the weights which are put on specific classes. As the investigated sample is imbalanced, putting more emphasis on being a bad borrower might mitigate this problem and balance both classes.

In order to find the best combination of selected model hyperparameters, an extensive grid search was performed. All previously defined combinations of a shrinkage technique, its tuning parameter, and class weights were examined and the results were compared in order to select the optimal model setting for this specific sample split. In total, 120 different logistic models were trained and the one with the highest value of AUC on testing dataset was chosen. The combination of hyperparameters of the best model is summarized in Table 5.3.

Shrinkage method	Tuning parameter	Class weights (Good:Bad)
Ridge	0.1	3:7

Table 5.3: Selected hyperparameters of logistic model

The first model, denoted as Model I, includes all previously selected variables. Its estimated coefficients are presented in Table B.1 in Appendix B. The last column of Table B.1 states the ranking of variables based on feature importance algorithm.

The value of AUC on testing sample is 0.744 which implies a fair model performance. In order to generalize model performance of different sample splits, 95% confidence interval for AUC statistic was estimated using 1,000 bootstrap samples. According to the results, 95 % of all AUC values on testing samples fall into (0.647; 0.744). As expected, the performance of the original model is very good in comparison to other possible AUC values.

The next step is to build Model II which includes a restricted set of variables based on their significance in Model I. The aim is to simplify the model in order to prevent it from overfitting while keeping all important and valuable variables in the model. Statistical significance of estimated coefficients is the primary source for assessing the effect of selected variables on the probability of default. Yet, the ranking of feature importance algorithm is also taken into account while deciding which variables might be redundant in the model. The results provided by both approaches are almost identical.

Each variable contains at least one significant category at 10% significance level, except for tenure profile and education level. Nevertheless, according to a feature importance algorithm, the category of vocational education level is the third most important variable. This is in line with its high IV which was previously computed. Therefore, education level is kept in the model. Additionally, all variables including at least one category which is significant at 1% level are kept in the model. As far as other variables are concerned, I decided to consider also loan amount, region of residence, and the number of children as potentially redundant variables as only one category was significant at 5%, resp. 10% significance level while taking also feature importance results into consideration.

After identifying potentially superfluous variables, models with all possible combinations of included variables were trained in order to identify which set of variables provided the best prediction performance. All models were trained using model hyperparameters which are identical with the ones of the original model. The final model included all previously used variables except for loan amount and tenure profile.

The results are summarized in Table B.2 in Appendix B. Furthermore, ROC curves of both logistic model can be found in Figure B.1 in Appendix B. The value of AUC on testing set equals 0.751. This is an improvement of 0.7 % in comparison to the full model. As in the previous case, the next step is to estimate 95% confidence interval for AUC values in order to generalize the results. After performing 1,000 repetitions of sample splits, the estimated confidence interval equals (0.649; 0.745).

The result shows that AUC statistic of the original model does not fall into this interval. This could have been expected as the values of hyperparameters of logistic model were optimized based on the original sample split. Hence, it might not be optimal for other splits and the overall performance is worse. On the contrary, the maximum value of AUC even reached 0.781 in one particular sample split, but such a high value is so rare that it does not fall into 95% confidence interval as in the case of originally presented model.

The restricted model contains three variables which are loan-related, whereas remaining five variables provide information about borrower. Using only eight variables, it is possible to accomplish a fair model performance for borrower's creditworthiness assessment. The results of both models provide very similar coefficients without any major change. Nevertheless, some variables became more significant after eliminating loan amount and tenure profile.

As the results suggest, monthly income is the most important variable. All categories are highly statistically significant. Similarly, loan year seems to be an important factor. Detailed discussion of the effects of various features on the probability of default and a comparison of the results with existing research findings are covered in Chapter 6.

5.3.2 Linear Discriminant Analysis

The second parametric method examined in this thesis is LDA. The first step is to decide which variables should be included in the model. Although logistic regression identified some features as insignificant, it might be possible that other classification methods will assess them as important. Therefore, they are kept in the sample for all remaining classification methods. This means that all methods are employed while having the same available information as they examine the same set of variables. Hence, it will be possible to compare their performance on this dataset.

Assuming the same identical covariance matrix for both classes, LDA fits a Gaussian density to each of the classes. Afterwards, the model assesses the probability of coming from a particular conditional density and assigns the most probable class.

Given the fact that LDA has such strict assumptions, it can be expected that the result will not be very good since the data are categorized. Hence, if possible, classed variables were transformed back to original continuous ones. Both options of coarse classed data and continuous ones were investigated.

As this method is not that complex, it does not require such a high number of hyperparameters to be defined. The only hyperparameter which proved to be important is *shrinkage*. It is a tool for estimation of covariance matrix when the size of a dataset is limited and the sample covariance is not optimal for the estimation. It can attain values from 0 to 1. Any value in this interval leads to estimation of a shrunk covariance matrix.

Afterwards, grid searches using both classed and continuous variables were performed. The results using classed variables yielded the AUC value of 0.757 on the testing dataset. It is even higher than the results of logistic models. Conversely, when using a set of continuous variables, the highest value of AUC for different shrinkage setting was only 0.700. It is quite surprising that the dataset which very likely violates model assumptions performs much better. The difference is more than 5 % on this sample split.

The optimal value of shrinkage hyperparameter as discovered by a grid search procedure equals 0.9. The estimated 95% confidence interval for AUC after 1,000 bootstrapped sample splits is (0.650; 0.748). It is very similar to the case of logistic regression as the interval is relatively wide. Additionally, the presented value does not fall into that interval. The grid search selected the best value for the specific sample split and it is exceptionally good to be in the confidence interval.

5.3.3 Quadratic Discriminant Analysis

QDA presents an extension of LDA classification method by allowing different covariance matrices for each class. As a consequence, a non-linear decision boundary can be constructed. Nevertheless, normality assumption should still hold. As in the case of LDA, both data with classed and continuous variables are examined.

The only hyperparameter to be consider during QDA model training is related to *regularization*. It regularizes the estimation of covariance matrices similarly as a shrinkage tool for LDA. A set of possible values was determined and a grid search was performed.

Once again, the results on the categorized dataset are much better in comparison to non-categorized data although the assumption of normality probably does not hold. The corresponding value of AUC for classed dataset reached 0.740, whereas the value of AUC for continuous variables was only 0.700. The optimal value for a regularization hyperparameter as discovered by grid search for both cases equals 0.3. Generalizing the results, 95% confidence interval equals (0.641; 0.745). Interestingly, some extremely low values occurred during 1,000 repetitions but they were very rare.

In comparison to LDA, this method produces slightly worse results on this particular sample split. Nevertheless, their 95% confidence intervals for AUC are almost identical.

5.3.4 Classification Tree

This is the first non-parametric method which was used in the thesis. James *et al.* (2013) point out that one of the advantages of classification tree is that it can easily handle quantitative variables, and hence it does not require a creation of dummy variables. When training a final model, the inclusion of continuous variables was also considered and both approaches were compared. Nevertheless, the model with classed variables outperformed the model with continuous variables as in the cases of LDA and QDA. Henceforward, the description of model building considers the original dataset of classed variables.

As for growing a classification tree, there are many aspects which define the final form of the model. Firstly, the objective is to have a model which is not too complicated and branched. This might be prevented by setting predetermined values of tree hyperparameters. They allow to define model training procedure and adjust it accordingly.

The most important hyperparameter adjusts the complexity of a final tree. It is possible to set the threshold for the number of maximum terminal nodes which determine its final size. If this hyperparameter was not defined, the tree could grow without any restrictions in order to provide accurate predictions. Nevertheless, the objective is to have a good and a parsimonious model which needs to be taken into account during model building. As the size of the model is controlled for, there is no need to perform pruning after the training.

The other group consists of hyperparameters which consider how nodes are created. The first hyperparameter called *criterion* selects the function which is used in order to measure the split quality. Two possible options include *gini* and *entropy*. Gini is a criterion for Gini impurity, whereas entropy is a criterion for the information gain. Second hyperparameter called *splitter* defines the strategy implemented while creating a split at each node. It can be either *best* which chooses the best split, or *random* which chooses the best random split.

Finally, it is possible to balance the classes using *class weights* as in the case of logistic models. This might improve the performance of the model and prevent it from giving priority to the larger class.

By defining a set of values for each of selected hyperparameters, a grid search procedure was performed and models were trained using all predetermined combinations. Overall, 9,072 models were trained and evaluated. The model which has the highest value of AUC on the testing set contained hyperparameter values summarized in Table 5.4. It can be seen that class weights are identical to weights used in the logistic model.

Criterion	**Splitter**	**Max terminal nodes**	**Class weights (Good:Bad)**
gini	best	8	3:7

Table 5.4: Selected hyperparameters of classification tree model

After constructing ROC curve, the value of corresponding AUC equals 0.721 on the testing set. Although James *et al.* (2013) claim that classification trees are expected to provide worse performance in comparison to parametric methods, it still provides a fair performance.

Computing AUC statistic for 1,000 sample splits, the estimated 95% confidence interval is (0.607; 0.706). As in the case of logistic model, the overall generalized performance is worse and the AUC value for the presented model does not fall into this interval. Nevertheless, the highest AUC value of 0.733 during bootstrapping was even higher.

The resulting tree diagram can be found in Appendix C. As can be seen, it is indeed very simple. It considers only three variables but still performs surprisingly well. Although only a few variables are included, the elimination of any other variable deteriorates the overall performance of the model. The top node denotes the most important feature. Furthermore, it takes education level and loan year into account. As for education level, its inclusion supports the incorporation of this variable into the logistic model. Both models hence provide very similar results as for feature importance.

Finally, other experiments during model training were performed. By experimenting with a tree size, I discovered that it was possible to grow a large complicated tree having more than 100 terminal nodes. This tree was built without any restriction on its size. Although it outperforms previously presented model, the difference was only marginal. Additionally, it performs much worse as for its overall generalized performance and could not compete with much simpler classification trees.

5.3.5 Random Forest

Random forest combines the prediction of several classification trees. Hence, it is natural to follow up on the previous section with its extension. As it is based on building multiple classification trees, its hyperparameters are almost identical as in the case of only one classification tree. Therefore, hyperparameters of previously grown classification tree were taken into account when defining possible values for grid search procedure.

More specifically, only *gini* criterion was considered as it outperformed the *entropy* option in the previous case. The most important random forest hyperparameter was the number of individual trees and their maximum depth.

Additionally, class weights are used in order to correct class imbalance in the dataset as in the previous cases.

After defining possible values, grid search was performed in order to select the optimum set of hyperparameters. The total number of 108 random forest models were trained and the best one was selected.

The optimal set of model hyperparameters is summarized in Table 5.5.

Criterion	**Max depth of a tree**	**Number of trees**	**Class weights (Good:Bad)**
gini	3	300	1:4

Table 5.5: Selected hyperparameters of random forest model

The resulting combination of model hyperparameters shows that the forest contains a high number of very small classification trees. Nevertheless, the model was also built using 600 and 1,000 trees but the results of a less complex random forest were better. As for the size of individual classification trees, rather smaller trees appeared to outperform more complex ones. Conversely to previous classification methods, even more emphasis is put on bad borrowers in the sample.

According to James *et al.* (2013), random forests usually provides a better performance in comparison to individual classification trees as they overcome drawbacks of this method mentioned in Chapter 4. This is supported by the result of constructed random forest model as can be derived from its ROC curve. The AUC statistic equals 0.757 which is the enhancement of 3.6 % over classification tree model. The estimated 95% confidence interval for AUC statistic is (0.657; 0.755). Similarly as in the previous cases, the performance of presented model does not fall into the interval as it is exceptionally good in comparison to other sample splits. The highest AUC value for one particular sample split even reached 0.785.

After computing feature importance of included variables, being in the lowest income category is the most important variable. This is in accord with the results of classification tree and it also supports the importance of income suggested by logistic models. Other most important features contain other income categories, loan year and vocational education level. Overall, the importance of all included variables is very similar to results of classification tree model and logistic models.

5.3.6 K-Nearest Neighbors

KNN classification method is a simple non-parametric approach. This implies a low number of hyperparameters to be defined prior to model training.

Firstly, one needs to define how many neighboring observations should be considered for prediction. Secondly, a metric based on which the distance is computed needs to be selected. Finally, it is possible to assign specific weights to neighboring observations. It needs to be emphasized that these are not the same class weights as in the cases of logistic model, classification tree, and random forest. For KNN, one cannot prioritize one class over the second one. More specifically, weights used in KNN might give priority to closer observations. Hence, the higher the distance between two observations, the lesser weight is put on the neighbor during model building. On the other, *uniform* weights might be more appropriate and are considered as well.

The total number of 40 models was trained and assessed using grid search procedure. The combination of selected hyperparameters is summarized in Table 5.6. The optimal number of neighbors is 19, although models using even wider neighborhood were considered as well. As for the weights, closer observations were given higher preference during prediction.

Number of neighbors	**Metric**	**Weights**
19	euclidean	distance

Table 5.6: Selected hyperparameters of k-nearest neighbors model

In comparison to other employed methods, KNN performs almost perfectly on the training sample. It was able to classify all bad borrowers correctly and the value of AUC was very close to 1.0. This shows clear overfitting of the model on the training data. Contrarily, the highest value of AUC on the testing sample equals 0.683 which implies a poor performance of KNN on this specific dataset. Despite overfitting on the training sample, the model with these specific hyperparameters outperformed all other models on the testing sample.

The estimated 95% confidence interval for AUC on this dataset is (0.598; 0.701). In comparison to other techniques, this method shows the worst results.

5.3.7 Support Vector Machines

SVM is the last employed and assessed classification method. As for its setting and computation procedure, it is the most demanding one. Nevertheless, it is very popular since it allows for complex decision boundaries. Moreover, it works very well in cases of both low-dimensional and high-dimensional datasets as suggested by Müller & Guido (2016).

There are many hyperparameters to be defined. They are mainly connected to the form of selected kernel function which is used to transform the feature space. I considered only the hyperparameter which specified kernel type. The options are *radial basis function kernel, linear kernel, polynomial kernel*, and *sigmoid kernel*. Other kernel hyperparameters were left at their default values as the type of kernel function proved to be the most important one.

Secondly, it is possible to set the regularization level as in the case of linear models in order to prevent the model from overfitting. Finally, SVM allows adjusting class weights as in the cases of logistic regression, classification tree, and random forest.

After defining a set of possible values of selected hyperparameters, a grid search yielded the optimal setting of SVM model which performed the best as for AUC statistic. The hyperparameters are summarized in Table 5.7.

Regularization	**Kernel type**	**Class weights (Good:Bad)**
1.0	sigmoid	1:4

Table 5.7: Selected hyperparameters of support vector machines

Similarly as in the previous cases, much more emphasis is put on bad borrowers in the sample in order to balance both classes. This approximately corresponds to the proportion of bad and good borrowers in the original dataset.

The value of corresponding AUC equals 0.761 which implies a fair performance of the model. It is the highest performance on this sample split in comparison to all employed classification methods. Estimated 95% confidence interval is (0.654; 0.750). As was already observed for previous methods, the value of AUC does not fall into this confidence interval. The reasons are identical as in previous cases.

5.4 Summary of Employed Classification Methods

This section summarizes the results of all employed classification methods so that their performance on the particular dataset can be compared. These models included all previously selected variables in order to compare methods having the same available information. Hence, the logistic model from which *tenure profile* and *loan amount* were removed is not considered. Nevertheless, the discussion of the effect of personal characteristics on the probability of default in Chapter 6 is based mostly on its results.

Detailed results of all methods are presented in Table 5.8. ROC curves are plotted in Figure 5.1. All applied models are compared in Chapter 6.

Classification Method	**AUC statistic**	**95% Confidence Interval**
Logistic regression	0.744	(0.647; 0.744)
LDA	0.757	(0.650; 0.748)
QDA	0.740	(0.641; 0.745)
Classification tree	0.721	(0.607; 0.706)
Random forest	0.757	(0.657; 0.755)
KNN	0.683	(0.598; 0.701)
SVM	**0.761**	(0.654; 0.750)

Table 5.8: Computed performance measures of employed classification methods

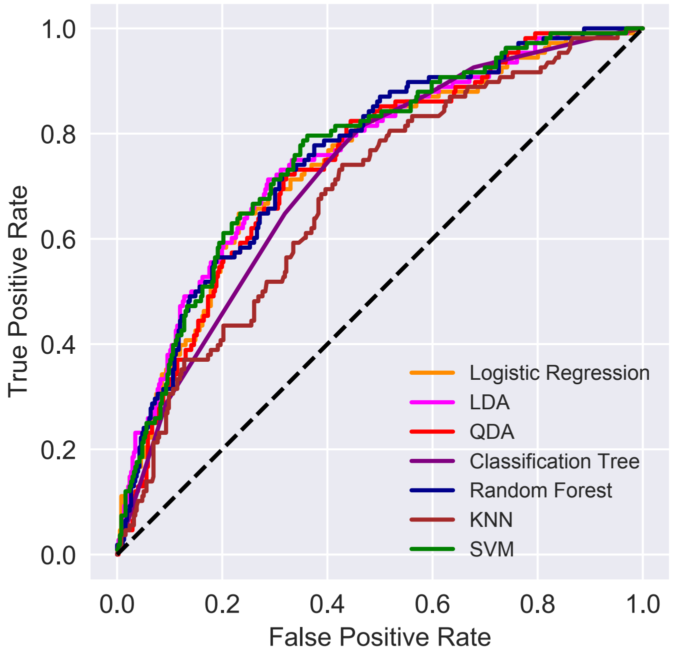


Figure 5.1: Comparison of ROC curves of employed classification methods

As mentioned in Chapter 4, AUC statistic as a performance measure was selected so that it is possible to compare the results with the existing research of Kočenda & Vojtek (2011). The authors constructed logistic models and a classification tree model using Czech mortgage data. The values of AUC statistic for these models reached 0.869 and 0.815, respectively. As can be seen, their models outperformed all classification methods employed in this thesis. This suggests that mortgage data might be better separable as for good and bad borrowers in comparison to unsecured consumer loans. Additionally, the authors might have selected more appropriate combination of analyzed variables for model building. Although the majority of investigated features is identical, there are some variables which have not been considered in this thesis, mainly information about borrower's employment status.

Chapter 6

Discussion of Results

In this chapter, I discuss the main findings which can be derived from conducted empirical analysis. As the primary objective of this study was to analyze the influence of individual borrower's characteristics on the probability of default in the Czech Republic, the first part comments on such discovered effects. The findings are mainly compared with the results of Kočenda & Vojtek (2011) as they examined a similar loan dataset from a Czech bank. Nevertheless, their research focused primarily on default predictors for mortgage loans. Hence, it is possible to compare the effect of borrower's characteristics between mortgage loans and unsecured consumer loans which are analyzed in this thesis.

The effects are presented and discussed based on the estimated coefficients of logistic model. Additionally, the predictive power of included features and feature importance provided by other models are taken into account when presenting the main findings.

Since classification models considered also variables which were not directly connected with a borrower but with their granted loan, their effect on the probability of default is presented as well.

As for the secondary aim of this study, classification methods which were employed are evaluated and potential implications are discussed. The results are compared with existing research which focuses on model performance assessment using credit data.

6.1 Discovered Effects of Examined Features

Before commenting on the estimated coefficients, it should be noted that the aim was to model the probability of default. Hence, features with negative coefficients indicate better creditworthiness and a lower risk of default. All formulated hypotheses related to the impact of personal characteristics on the probability of default are tested at significance level $\alpha = 0.05$. Furthermore, presented findings are based on estimated coefficients, z-scores and p-values in Table B.1 and Table B.2 in Appendix B, respectively.

6.1.1 Monthly Income

Monthly income is the strongest default predictor for estimation of the probability of default. Its prominent role was already discovered by Greene (1992). This finding confirms the assumption that borrower's creditworthiness is determined by their financial situation and the amount of money they have at disposal. For the analysis, borrowers have been divided into five income groups. The estimated coefficients show that higher monthly income leads to lower probability of default. This finding estimated by logistic regression model is supported by other models. Classification tree model assigned the lowest income category to the top node.

Additionally, all coefficients are statistically significant at 5% significance level. Overall, we can reject the hypothesis that income does not effect the probability of default since there are significant differences across various income groups as can be seen in Table B.2.

The dominant role of monthly income is also confirmed by the results of performed feature importance algorithm. As for the value, the coefficients belong to ones with the highest magnitude which indicates a great difference between people in the lowest income category and borrowers with the highest monthly incomes. This is in accord with the results for mortgage loans as claimed by Kočenda & Vojtek (2011). Even though mortgage loans represent a larger financial burden since the borrowed amounts are higher, monthly income is the strongest default predictor for both loan types. In conclusion, much emphasis should be put on client's financial situation as monthly income significantly affects the probability of default.

6.1.2 Gender

One of the analyzed personal characteristics and its effect on the probability of default was *gender* of a borrower. The predictive power of this variable was the lowest in the whole dataset. This suggests that it is not a good default predictor and its inclusion in credit scoring models does not provide any enhancement to its ability to discriminate between potentially bad and good borrowers. Therefore, we cannot reject the hypothesis that borrower's gender does not affect the probability of default when considering its predictive power included in Table 5.2 in Chapter 5.

This is in line with findings of Kočenda & Vojtek (2011) whose analysis provided identical conclusion. This might be closely connected to the fact that some advanced countries do not allow discriminating based on gender, and hence do not consider its inclusion into credit application process.[1] Contrarily, borrower's gender plays a major role when predicting the probability of default for African and Asian countries as discovered by Kinda & Achonu (2012) or Dinh & Kleimeier (2007).

6.1.3 Age

Another examined feature was borrower's *age*. As a reminder, the intervals for age were constructed during data transformation procedure described in Chapter 3. The values of estimated coefficients show that the risk of default is not linear across various age categories. Clients who took out a loan at the age of 30 to 36 are the most susceptible category to default on their loans in comparison to other age groups. This is an interesting finding which might indicate the situation of a borrower in this period. It is likely that all necessary expenses are very high, and hence it is difficult to cope with repayment schedule.

The second riskiest category is made up by borrowers aged 18–29 years. This could have been expected as these borrowers probably occupy junior job positions and their income is not as high as for other groups. On the contrary, borrowers who are older than 36 years have a lower probability of default. This applies to borrowers who are younger than 54 years. After reaching the age of 55 years, the probability of default increases.

People aged 49–54 years are the least risky group. It is probable that borrowers in this category have established a strong financial background and are able to repay their loans with higher probability. The estimated coefficient is highly statistically significant which indicates a substantial difference in comparison to the youngest borrowers. Furthermore, this is the fifth most important default predictor used in model building.

Borrowers aged 39–48 years are the second most creditworthy category. The

[1] Equal Opportunity Credit Act (1976) in the United States of America can be mentioned as an example.

coefficient is statistically significant at 5% significance level. This is the main difference compared to the full model. When including all features, this variable was regarded as insignificant. Since two coefficients for age categories are statistically significant at 5% level, we can reject the hypothesis that borrower's age does not affect the probability of default as can be derived from Table B.2.

Overall, the findings suggest that higher age implies lower probability of default and its consideration might improve the results of credit scoring models. This is supported by the results of Constangioara (2011). As for analyses conducted on Czech data, this finding is contrary to results discovered by Kočenda & Vojtek (2011) who claim that borrower's age is not a good default predictor.

6.1.4 Number of Children

The *number of children* a borrower has might influence financial situation in the family. The estimated coefficients show that people who have one child are less risky in comparison to childless people and people who have more than one child.

Additionally, having at least two children and being childless seem to be indifferent which is a surprising finding. The results suggest that there is not any difference between the creditworthiness of borrowers who do not have any children and borrowers who have at least two children since the estimated coefficient is insignificant at 5% significance level.

On the contrary, the difference between childless borrowers and people who have one child is statistically significant at 5% significance level (see Table B.2). Therefore, by testing the corresponding hypothesis, we can reject the null hypothesis that the number of children does not influence borrower's creditworthiness.

This result cannot be compared to other studies conducted on a Czech dataset since Kočenda & Vojtek (2011) included the number of children into their analysis. Nevertheless, the result is similar to findings of Crook *et al.* (1992) who found out that the number of children was one of the most important characteristics for default prediction in the United Kingdom. The results show that this feature should be taken into consideration when assessing credit risk but its role seems to be less important in comparison to other personal characteristics.

6.1.5 Education Level

Education level is the third strongest default predictor based on calculated information value. In addition, the category denoting borrowers with vocational education is the third most important variable for model building as predicted by feature importance algorithm. This applies to the results of logistic regression, classification tree, and random forest.

As in the previous cases, the relationship is not linear, and hence it is not possible to claim that the higher level of education generally leads to enhanced borrower's creditworthiness. The results suggest that people who have vocational education are the riskiest. This is a surprising finding because it might be expected that people who have only primary level of education probably earn lower amount of money in comparison to other categories. This would suggest that they are the riskiest category but this is not confirmed by the results. Possible explanation for this might be that banks are cautious about applicants who have primary education and do not grant a loan at all. Furthermore, people who have gained vocational education might have a tendency to work as self-employed. This might pose a threat of unstable and irregular income.

On the contrary, secondary education and university education leads to lower probability in comparison to primary education which is in accord with Kočenda & Vojtek (2011). Apart from the effect of vocational education, the values of estimated coefficients on education level are as expected since higher level of education leads to lower probability of default.

As for the hypothesis testing, we cannot reject the null hypothesis of no effect of education on the probability of default. No estimated coefficient contained in Table B.2 is significant at 5% significance level, and therefore we did not find enough evidence to claim that there is any difference across groups of people with various education levels. Although feature algorithm and the results of other models suggest the contrary, this cannot be confirmed by the estimated results of logistic regression.

6.1.6 Region of Residence

The last analyzed personal characteristic is a *region of borrower's residence.* The comparison is drawn between the capital city of Prague and other regions in the Czech Republic. As for predictive power, region does not seem to substantially affect borrower's creditworthiness. This is supported by the results of feature importance algorithm as majority of categories are not that important in comparison to other features.

Most of the coefficients are insignificant at 5% significance level which indicates that there is not any major difference between Czech regions as for the susceptibility of their residents to default on a loan. This would indicate that the living standard is very similar across the Czech Republic and no region suffers from more frequent defaults. Nevertheless, the coefficient on Karlovy Vary region suggests that there is a difference between this region and Prague as it is significant at 5% significance level. In addition, it is the tenth most important variable. The negative sign of the estimated coefficient means that people living in Prague are more likely to default on a loan in comparison to Karlovy Vary region. This is a peculiar finding as average monthly income in this region is the lowest in the Czech Republic as stated by Czech Statistical Office (2017). As income is the core default determinant, people living in Karlovy Vary should have much more difficulties while repaying their loans as opposed to the results. Nevertheless, the results do not support this assumption and it is very likely that other factors influence the probability of default in this region.

Since the coefficient on Karlovy Vary region is statistically significant at 5% significance level, we can reject the hypothesis that there is no difference across Czech regions as for the probability of default (see Table B.2). This is a new finding since Kočenda & Vojtek (2011) did not include region into their prediction models. As for other European countries, Crook *et al.* (1992) discovered that a region of borrower's residence was the most important default predictor in the United Kingdom.

6.1.7 Loan-Related Variables

Remaining variables included in the model are related directly to granted loans or the relationship with the bank. The results show that there is a significant difference between consolidation of loans and loans for housing at 5% significance level. People who have consolidated their loans into one major loan are more likely to default on their loan in comparison to loan for housing. This follows from the nature of both types. Although data on both loan types are very similar as for loan amounts and other variables, the probability of default is significantly higher for consolidation loan type. This might indicate a very difficult financial situation for borrowers who have taken out more loans at the same time, and hence their indebtedness increases.

The analyzed dataset covers a period staring in 2006 and ending in 2016. The data were categorized in four classes with loans granted in 2016 being the reference category. As can be seen from the estimated coefficients in Table B.2, loans granted in 2016 are the least risky. This follows from the structure of data as these borrowers have not had so much time to default on their loans. Nevertheless, the results for other categories are more interesting and might provide some information about the situation in the loan market over the examined period.

People who took out a loan during the period 2006–2012 are the second riskiest category. This might be affected by the financial crisis which occurred during this period. Hence, the effect of the crisis probably influenced also the loan market. Additionally, loans granted in 2013 and 2014 are the riskiest in the sample. This might be connected with macroeconomic situation during this period.

The expected effect of loan duration is not straightforward. The period for which the loan is taken out corresponds to the period for which the borrower is indebted. Nevertheless, the installment amount is lower, and hence lower proportion of monthly income is dedicated to repayment.

As the estimated coefficients suggest, loan duration does not affect the probability of default apart from one exception. Loans granted for a period of 7 to 8 years are regarded as much riskier in comparison to other loans since the estimated coefficient is significant at 5% level.

The first logistic model which included all variables provided some insight into the effect of loan amount on borrower's creditworthiness. The variable was divided into six categories. Surprisingly, the results do not confirm that higher loans are repaid with more difficulties as could be assumed. The riskiest loans are relatively low. These are loans granted for the amount between 100,001 CZK and 155,999 CZK which is the second lowest category. As for other categories, the model did not identify any significant differences.

As for the relationship with the bank, the model does not bring any evidence that there is a difference between creditworthiness of long-term clients and new clients. This is inconsistent with the results of Kočenda & Vojtek (2011), Crook *et al.* (1992), or Bellotti & Crook (2009). It might suggest that the loan application process is very thorough since the information gathered on long-term clients over the years is very likely comparable to the information gathered on new clients. The same finding applies to the fact whether a borrower's partner is also a bank's client as it has very low information value and was not included in the model.

6.2 Comparison of Classification Methods

The secondary objective of performed analysis was to assess the suitability of employed classification methods on the examined dataset of unsecured consumer loans.

Given the sample split on which the analysis was conducted, performance of all methods is very similar. The values of AUC statistic range from 0.683 to 0.761. The best performance was reached by SVM model of which AUC statistic is almost 8 % higher in comparison to the worst result. This in accord with the results of Bellotti & Crook (2009) as their SVM model outperformed LDA, KNN, and logistic regression.

On the contrary, KNN appears to be the least appropriate technique. As can be seen in Figure 5.1 in Chapter 5, ROC curve of KNN method is clearly the worst one without any need to compute AUC value for comparison. It is the only method which did not reach a fair performance level at the examined sample split. Hence, it is very likely that the dataset does not include clustered classes of good and bad borrowers for which KNN might be suitable.

Additionally, classification tree model is somewhat worse in comparison to other remaining methods. This is contrary to Lee *et al.* (2006) who constructed a classification tree model which provided the best performance among all methods. The 95% confidence interval for AUC statistic is almost identical to confidence interval of KNN method which suggests that both methods are generally identical. Nevertheless, it outperformed KNN on this particular sample split by almost 4 %. Overall, it still performs very well considering the small size of the tree having only eight terminal nodes and including information only about monthly income, loan year, and borrower's education level.

As for other classification methods, they produced almost identical performance. The values of AUC statistic for logistic regression, LDA, QDA, random forest, and SVM ranges from 0.740 to 0.761. The difference of 2.1 % between QDA and SVM is very small. Similarly, when generalizing model performance, the estimated 95% confidence intervals are only marginally different.

LDA model marginally outperformed logistic regression on this particular sample split and provided the second best performance as for predictive power.

The difference in their AUC statistics equals 1.3 %. This is in accord with the results of Bellotti & Crook (2009). Conversely, Desai *et al.* (1996), Blanco *et al.* (2013), or West (2000) analyzed such datasets on which logistic regression performed better in comparison to LDA. Furthermore, LDA outperformed QDA as in case of Blanco *et al.* (2013).

Random forest model was together with LDA the second best applied method. Its promising application for credit scoring was suggested by Kruppa *et al.* (2013) as it outperformed other investigated methods. The difference between individual classification tree and random forest is 3.6 %.

In order to test the hypothesis that there is no difference among applied credit scoring models as for their predictive power, it is necessary to consider their overall performance which is expressed by estimated 95% confidence intervals for AUC statistic.

Since all intervals are overlapping, we cannot reject the hypothesis regarding significantly different discrimination power of evaluated models. Hence, all classification methods performed only with marginal differences on the examined dataset. This result supports the statement of James *et al.* (2013) that all methods generally perform equally precisely. Additionally, the size of estimated intervals is very similar for all methods which implies that all models are equally stable. In conclusion, all employed classification methods provided a comparable solid performance and might be used for credit scoring purposes.

The results of hypothesis testing which were previously formulated are summarized in Table 6.1.

No.	**Hypothesis**	**Result**
H1	Client's gender does not affect the probability of default.	not rejected
H2	Client's age does not affect the probability of default.	rejected
H3	The number of client's children does not affect the probability of default.	rejected
H4	The level of education does not affect the probability of default.	not rejected
H5	Client's monthly income does not affect the probability of default.	rejected
H6	The region in which a client lives does not affect the probability of default.	rejected
H7	There is no difference in performance among applied credit scoring techniques.	not rejected

Table 6.1: The results of hypothesis testing

Chapter 7

Conclusion

The thesis was focused on two main objectives. The primary aim was to assess the effects of borrower's personal characteristics on the probability of default. The secondary objective was to investigate suitability of selected classification methods for a specific Czech dataset of two types of unsecured consumer loans. In total, seven hypotheses were formulated. They considered both potential default predictors and comparison of selected classification methods and their performance on the examined dataset.

To the author's knowledge, such a thorough credit risk analysis of unsecured loans has not been performed yet. The conducted research completed the work of Kočenda & Vojtek (2011) who studied default predictors on mortgage loans by employing two classification methods.

In this thesis, a real-world loan dataset provided by a major Czech commercial bank was analyzed. Investigating default predictors using such a dataset presents a challenging task since it includes clients who have been granted a loan, and hence the probability of default is supposed to be very low. These clients were approved as successful applicants for loan based on advanced credit scoring procedure conducted by the bank. Hence, the aim was to model the probability of default conditional on the fact that the client was granted a loan. This is a general problem of building credit scoring models as no other data can be used for this purpose.

Firstly, the dataset was examined in order to identify general patterns about people who have taken out an unsecured consumer loan. Subsequently, the data were transformed into more convenient form which is suggested for building credit scoring models. Secondly, seven classification methods were applied and their performances were assessed. Having a logistic regression as a core model, its estimated coefficients were used for the interpretation of effects of personal characteristics and hypothesis testing. The results suggest than monthly income, age, region of residence, and the number of children a borrower has play an important role while assessing borrower's creditworthiness. Conversely, we could not reject hypotheses that gender and education do not influence the probability of default, respectively.

After identifying the most important characteristics for prediction of default on a loan, the secondary objective was to compare constructed credit scoring models. Overall, all applied classification methods produced approximately identical performance with only marginal differences. Furthermore, all models produced solid results as for discriminating between bad and good borrowers.

Regarding future work, the following research directions might be considered. Since this analysis was partially limited by data unavailability of some desired features, the research can be extended to other personal characteristics, such as marital status, information about borrower's employment, information about household situation, or monthly expenses. Additionally, there is a potential in considering various macroeconomic features and the overall effect of macroeconomic conditions on the probability of default. As for employed classification methods, it is possible to enhance the list of applied approaches. One possibility is combining predictions of selected models via ensemble methods. Hence, the resulting class will be predicted based on the majority or other type of voting which might improve the accuracy of individual models.

Bibliography

Abdou, H. A. & J. Pointon (2011): "Credit scoring, statistical techniques and evaluation criteria: a review of the literature." *Intelligent Systems in Accounting, Finance and Management* **18(2-3)**: pp. 59–88.

Abdou, H. A. & M. D. D. Tsafack (2015): "Forecasting creditworthiness in retail banking: a comparison of cascade correlation neural networks, CART and logistic regression scoring models." In "The 2nd International Conference on Innovation in Economics and Business," Amsterdam, Netherlands.

Abdou, H. A., M. D. D. Tsafack, C. G. Ntim, & R. D. Baker (2016): "Predicting creditworthiness in retail banking with limited scoring data." *Knowledge-Based Systems* **103**: pp. 89–103.

Anderson, R. (2007): *The credit scoring toolkit: theory and practice for retail credit risk management and decision automation.* Oxford University Press.

Ayouche, S., R. Aboulaich, & R. Ellaia (2017): "Partnership credit scoring classification problem: A neural network approach." *International Journal of Applied Engineering Research* **12(5)**: pp. 693–704.

Baesens, B., T. Van Gestel, S. Viaene, M. Stepanova, J. Suykens, & J. Vanthienen (2003): "Benchmarking state-of-the-art classification algorithms for credit scoring." *Journal of the operational research society* **54(6)**: pp. 627–635.

Bellotti, T. & J. Crook (2009): "Support vector machines for credit scoring and discovery of significant features." *Expert Systems with Applications* **36(2)**: pp. 3302–3308.

Bhatia, S., P. Sharma, R. Burman, S. Hazari, & R. Hande (2017): "Credit Scoring using Machine Learning Techniques." *International Journal of Computer Applications* **161(11)**: pp. 1–4.

BLANCO, A., R. PINO-MEJÍAS, J. LARA, & S. RAYO (2013): "Credit scoring models for the microfinance industry using neural networks: Evidence from Peru." *Expert Systems with Applications* **40(1)**: pp. 356–364.

BREIMAN, L. (2001): "Random forests." *Machine learning* **45(1)**: pp. 5–32.

BREIMAN, L., J. FRIEDMAN, C. J. STONE, & R. A. OLSHEN (1984): *Classification and regression trees.* Wadsworth International Group, Belmont, CA.

CONSTANGIOARA, A. (2011): "Consumer credit scoring." *Romanian Journal of Economic Forecasting* **3**: pp. 162–177.

CORTES, C. & V. VAPNIK (1995): "Support-vector networks." *Machine learning* **20(3)**: pp. 273–297.

CROOK, J. N., R. HAMILTON, & L. C. THOMAS (1992): "A comparison of discriminators under alternative definitions of credit default." In "Proceedings of the IMA Conference on Credit Scoring and Credit Control," pp. 217–245. Claredon Press.

CZECH NATIONAL BANK (2017): "CNB Financial Stability Report 2016/2017." `http://www.cnb.cz/miranda2/export/sites/www.cnb.cz/en/financial_stability/fs_reports/fsr_2016-2017/fsr_2016-2017.pdf`. Online; accessed December 26, 2017.

CZECH STATISTICAL OFFICE (2017): "Average gross monthly wage in Q4 2017." `https://www.czso.cz/csu/xs/prumerna-hruba-mesicni-mzda-ve-4-ctvrtleti-2017`. Online; accessed April 5, 2018.

DAHIYA, S., S. HANDA, & N. P. SINGH (2015): "Credit scoring using ensemble of various classifiers on reduced feature set." *Industrija* **43(4)**: pp. 163–174.

DESAI, V. S., J. N. CROOK, & G. A. OVERSTREET (1996): "A comparison of neural networks and linear scoring models in the credit union environment." *European Journal of Operational Research* **95(1)**: pp. 24–37.

DINH, T. H. T. & S. KLEIMEIER (2007): "A credit scoring model for Vietnam's retail banking market." *International Review of Financial Analysis* **16(5)**: pp. 471–495.

Equal Opportunity Credit Act (1976): *Amendments of 1976. U.S. Code, Title 15, Section 1691 et. seq.*

Feldman, D. & S. Gross (2005): "Mortgage default: classification trees analysis." *The Journal of Real Estate Finance and Economics* **30(4)**: pp. 369–396.

Fisher, R. A. (1936): "The use of multiple measurements in taxonomic problems." *Annals of human genetics* **7(2)**: pp. 179–188.

Friedman, J., T. Hastie, & R. Tibshirani (2001): *The elements of statistical learning*, volume 1. Springer Series in Statistics, New York.

Ganopoulou, M., F. Giapoutzi, K. Kosmidou, & T. Moysiadis (2013): "Credit-scoring and bank lending policy in consumer loans." *International Journal of Financial Engineering and Risk Management* **1(1)**: pp. 90–110.

Greene, W. H. (1992): "A Statistical Model for Credit Scoring." *Working Paper, Department of Economics, Stern School of Business, New York University* **92(29)**.

Hand, D. J. (2001): "Modelling consumer credit risk." *IMA Journal of Management mathematics* **12(2)**: pp. 139–155.

Hand, D. J. & W. E. Henley (1997): "Statistical classification methods in consumer credit scoring: a review." *Journal of the Royal Statistical Society: Series A (Statistics in Society)* **160(3)**: pp. 523–541.

Henley, W. E. & D. J. Hand (1996): "A k-nearest-neighbour classifier for assessing consumer credit risk." *Journal of the Royal Statistical Society: Series D (The Statistician)* **45**: pp. 77–95.

Henley, W. E. & D. J. Hand (1997): "Construction of a k-nearest-neighbour credit-scoring system." *IMA Journal of Management Mathematics* **8(4)**: pp. 305–321.

Hosmer Jr, D. W., S. Lemeshow, & R. X. Sturdivant (2013): *Applied logistic regression*, volume 398. John Wiley & Sons.

Huang, C.-L., M.-C. Chen, & C.-J. Wang (2007): "Credit scoring with a data mining approach based on support vector machines." *Expert systems with applications* **33(4)**: pp. 847–856.

JAMES, G., D. WITTEN, T. HASTIE, & R. TIBSHIRANI (2013): *An introduction to statistical learning*, volume 112. Springer.

KENNEDY, K. (2013): *Credit Scoring Using Machine Learning.* Ph.D. thesis, Dublin Institute of Technology.

KHANDANI, A. E., A. J. KIM, & A. W. LO (2010): "Consumer credit-risk models via machine-learning algorithms." *Journal of Banking & Finance* **34(11)**: pp. 2767–2787.

KINDA, O. & A. ACHONU (2012): "Building a Credit Scoring Model for the Savings and Credit Mutual of the Potou Zone." *Consilience: The Journal of Sustainable Development* **7(1)**: pp. 17–32.

KOČENDA, E. & M. VOJTEK (2011): "Default predictors in retail credit scoring: Evidence from Czech banking data." *Emerging Markets Finance and Trade* **47(6)**: pp. 80–98.

KOH, H. C., W. C. TAN, & C. P. GOH (2006): "A two-step method to construct credit scoring models with data mining techniques." *Journal of Business and Information* **1(1)**: pp. 96–118.

KRUPPA, J., A. SCHWARZ, G. ARMINGER, & A. ZIEGLER (2013): "Consumer credit risk: Individual probability estimates using machine learning." *Expert Systems with Applications* **40(13)**: pp. 5125–5131.

LEE, T.-S., C.-C. CHIU, Y.-C. CHOU, & C.-J. LU (2006): "Mining the customer credit using classification and regression tree and multivariate adaptive regression splines." *Computational Statistics & Data Analysis* **50(4)**: pp. 1113–1130.

LOUZADA, F., A. ARA, & G. B. FERNANDES (2016): "Classification methods applied to credit scoring: Systematic review and overall comparison." *Surveys in Operations Research and Management Science* **21(2)**: pp. 117–134.

MESTER, L. J. (1997): "What's the point of credit scoring?" *Business review–Federal Reserve Bank of Philadelphia* **3(Sep/Oct)**: pp. 3–16.

MINISTRY OF LABOUR AND SOCIAL AFFAIRS (2017): "Overview of the development of minimum wage rates." `https://www.mpsv.cz/cs/871`. Online; accessed April 3, 2018.

Müller, A. C. & S. Guido (2016): *Introduction to machine learning with Python: a guide for data scientists.* O'Reilly Media, Inc.

Nanni, L. & A. Lumini (2009): "An experimental comparison of ensemble of classifiers for bankruptcy prediction and credit scoring." *Expert systems with applications* **36(2)**: pp. 3028–3033.

Nguyen, H.-T. (2015): "Default predictors in credit scoring: evidence from France's retail banking institution." *The Journal of Credit Risk* **11(2)**: pp. 41–66.

Pedregosa, F., G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, & E. Duchesnay (2011): "Scikit-learn: Machine Learning in Python." *Journal of Machine Learning Research* **12**: pp. 2825–2830.

Rodriguez, J. J., L. I. Kuncheva, & C. J. Alonso (2006): "Rotation forest: A new classifier ensemble method." *IEEE transactions on pattern analysis and machine intelligence* **28(10)**: pp. 1619–1630.

Roszbach, K. (2004): "Bank lending policy, credit scoring, and the survival of loans." *Review of Economics and Statistics* **86(4)**: pp. 946–958.

Siddiqi, N. (2006): *Credit risk scorecards: developing and implementing intelligent credit scoring.* John Wiley & Sons.

Šušteršič, M., D. Mramor, & J. Zupan (2009): "Consumer credit scoring models with limited data." *Expert Systems with Applications* **36(3)**: pp. 4736–4744.

Thomas, L., J. Crook, & D. Edelman (2017): *Credit scoring and its applications*, volume 2. Siam.

Vojtek, M. & E. Kočenda (2006): "Credit-scoring methods." *Czech Journal of Economics and Finance (Finance a úvěr)* **56(3-4)**: pp. 152–167.

West, D. (2000): "Neural network credit scoring models." *Computers & Operations Research* **27(11)**: pp. 1131–1152.

Appendix A

Categorized Variables

This part presents a detailed description of categorized variables used in the empirical analysis. Most variables were transformed into coarse classes, whereas some variables preserved original categories. Additionally, the computation of IV value for each variable is presented.

Loan Amount	**Good**	**%Good**	**Bad**	**%Bad**	**Odds**	**WOE**	**IV**
49,001–100,000	268	14.30 %	77	14.10 %	1.01	0.0139	0.00 %
100,001–155,999	278	14.83 %	106	19.41 %	0.76	-0.2690	1.23 %
156,000–200,999	267	14.25 %	89	16.30 %	0.87	-0.1346	0.28 %
201,000–290,999	301	16.06 %	73	13.37 %	1.20	0.1834	0.49 %
291,000–544,999	562	29.99 %	157	28.75 %	1.04	0.0420	0.05 %
$\geq$ 545,000	198	10.57 %	44	8.06 %	1.31	0.2708	0.68 %
Total IV							**2.73 %**

Table A.1: Coarse classes of loan amount (CZK)

Loan Type	**Good**	**%Good**	**Bad**	**%Bad**	**Odds**	**WOE**	**IV**
For housing	882	47.07 %	216	39.56 %	1.19	0.1737	1.30 %
Consolidation	992	52.93 %	330	60.44 %	0.88	-0.1325	1.00 %
Total IV							**2.30 %**

Table A.2: Classes of loan type

Loan Year	**Good**	**%Good**	**Bad**	**%Bad**	**Odds**	**WOE**	**IV**
2016	986	52.61 %	210	38.46 %	1.37	0.3133	4.43 %
2015	358	19.10 %	114	20.88 %	0.91	-0.0888	0.16 %
2014–2013	304	16.22 %	140	25.64 %	0.63	-0.4578	4.31 %
2012–2006	226	12.06 %	82	15.02 %	0.80	-0.2194	0.65 %
Total IV							**9.55 %**

Table A.3: Coarse classes of loan year

Loan Duration	**Good**	**%Good**	**Bad**	**%Bad**	**Odds**	**WOE**	**IV**
225–1,852	398	21.24 %	95	17.40 %	1.22	0.1993	0.77 %
1,853–2,549	191	10.19 %	42	7.69 %	1.32	0.2814	0.70 %
2,550–2,588	181	9.66 %	62	11.36 %	0.85	-0.1618	0.27 %
2,589–2,938	172	9.18 %	82	15.02 %	0.61	-0.4924	2.88 %
2,939–3,654	350	18.68 %	126	23.08 %	0.81	-0.2115	0.93 %
$\geq$ 3,655	582	31.06 %	139	25.46 %	1.22	0.1987	1.11 %
Total IV							**6.66 %**

Table A.4: Coarse classes of loan duration (days)

Gender	**Good**	**%Good**	**Bad**	**%Bad**	**Odds**	**WOE**	**IV**
Male	1089	58.11 %	344	63.00 %	0.92	-0.0808	0.40 %
Female	785	41.89 %	202	37.00 %	1.13	0.1242	0.60 %
Total IV							**1.00 %**

Table A.5: Classes of gender

Age	**Good**	**%Good**	**Bad**	**%Bad**	**Odds**	**WOE**	**IV**
18–29	382	20.38 %	155	28.39 %	0.72	-0.3312	2.65 %
30–36	365	19.48 %	127	23.26 %	0.84	-0.1775	0.67 %
37–38	152	8.11 %	37	6.78 %	1.19	0.1797	0.24 %
39–48	592	31.59 %	142	26.01 %	1.21	0.1944	1.09 %
49–54	206	10.99 %	37	6.78 %	1.62	0.4837	2.04 %
$\geq$ 55	177	9.45 %	488	8.79 %	1.0743	0.0717	0.05 %
Total IV							**6.74 %**

Table A.6: Coarse classes of borrower's age

Children	**Good**	**%Good**	**Bad**	**%Bad**	**Odds**	**WOE**	**IV**
0	1257	67.08 %	401	73.44 %	0.91	-0.0907	0.58 %
1	510	27.21 %	113	20.70 %	1.31	0.2738	1.79 %
$\geq$ 2	107	5.71 %	32	5.86 %	0.97	-0.0261	0.00 %
Total IV							**2.37 %**

Table A.7: Coarse classes of the number of children

Education	**Good**	**%Good**	**Bad**	**%Bad**	**Odds**	**WOE**	**IV**
Primary	48	2.56 %	21	3.85 %	0.67	-0.4065	0.52 %
Secondary	719	38.37 %	190	34.80 %	1.10	0.0976	0.35 %
Vocational	476	25.40 %	218	39.93 %	0.64	-0.4523	6.57 %
University	275	14.67 %	48	8.79 %	1.67	0.5123	3.01 %
Other	356	19.00 %	69	12.64 %	1.50	0.4076	2.59 %
Total IV							**13.04 %**

Table A.8: Coarse classes of education level

Region	Good	%Good	Bad	%Bad	Odds	WOE	IV
Prague	245	13.07 %	55	10.07 %	1.30	0.2607	0.78 %
South Bohemian	120	6.40 %	40	7.33 %	0.87	-0.1346	0.12 %
South Moravian	145	7.74 %	38	6.96 %	1.11	0.1059	0.08 %
Karlovy Vary	107	5.71 %	25	4.58 %	1.25	0.2207	0.25 %
Vysočina	51	2.72 %	9	1.65 %	1.65	0.5014	0.54 %
Hradec Králové	111	5.92 %	31	5.68 %	1.04	0.0423	0.01 %
Liberec	79	4.22 %	36	6.59 %	0.64	-0.4472	1.06 %
Moravian-Silesian	213	11.37 %	74	13.55 %	0.84	-0.1760	0.38 %
Olomouc	171	9.12 %	47	8.61 %	1.06	0.0583	0.03 %
Pardubice	67	3.58 %	21	3.85 %	0.93	-0.0734	0.02 %
Plzeň	131	6.99 %	28	5.13 %	1.36	0.3098	0.58 %
Central Bohemian	156	8.32 %	54	9.89 %	0.84	-0.1723	0.27 %
Ůstí nad Labem	160	8.54 %	50	9.16 %	0.93	-0.0700	0.04 %
Zlín	118	6.30 %	38	6.96 %	0.90	-0.1001	0.07 %
Total IV							**4.23 %**

Table A.9: Classes of region of residence

Tenure Profile	Good	%Good	Bad	%Bad	Odds	WOE	IV
0	91	4.82 %	22	4.03 %	1.21	0.1866	0.15 %
0.01–20.99	267	14.25 %	100	18.32 %	0.78	-0.2511	1.02 %
21–90.99	535	28.55 %	195	35.71 %	0.80	-0.2239	1.60 %
$\geq$ 91	981	52.35 %	229	41.94 %	1.25	0.6717	2.31 %
Total IV							**5.08 %**

Table A.10: Coarse classes of tenure profile (months)

Partner Client	Good	%Good	Bad	%Bad	Odds	WOE	IV
Yes	859	45.84 %	216	39.56 %	1.16	0.1473	0.92 %
No	1,015	54.16 %	330	60.44 %	0.90	-0.1096	0.69 %
Total IV							**1.61 %**

Table A.11: Classes of partner client

Monthly Income	Good	%Good	Bad	%Bad	Odds	WOE	IV
11,000–21,100	645	34.42 %	324	59.34 %	0.58	-0.5447	13.58 %
21,101–27,700	377	20.12 %	103	18.86 %	1.07	0.0643	0.08 %
27,771–33,000	200	10.67 %	43	7.88 %	1.35	0.3039	0.85 %
33,001–61,000	426	22.73 %	61	11.17 %	2.03	0.7103	8.21 %
$\geq$ 61,001	226	12.06 %	15	2.75 %	4.39	1.4792	13.78 %
Total IV							**36.50 %**

Table A.12: Coarse classes of client monthly income (CZK)

Appendix B

Detailed Results of Logistic Models

This appendix presents ROC curves of Model I and Model II. Furthermore, estimated coefficients of both logistic models are included.

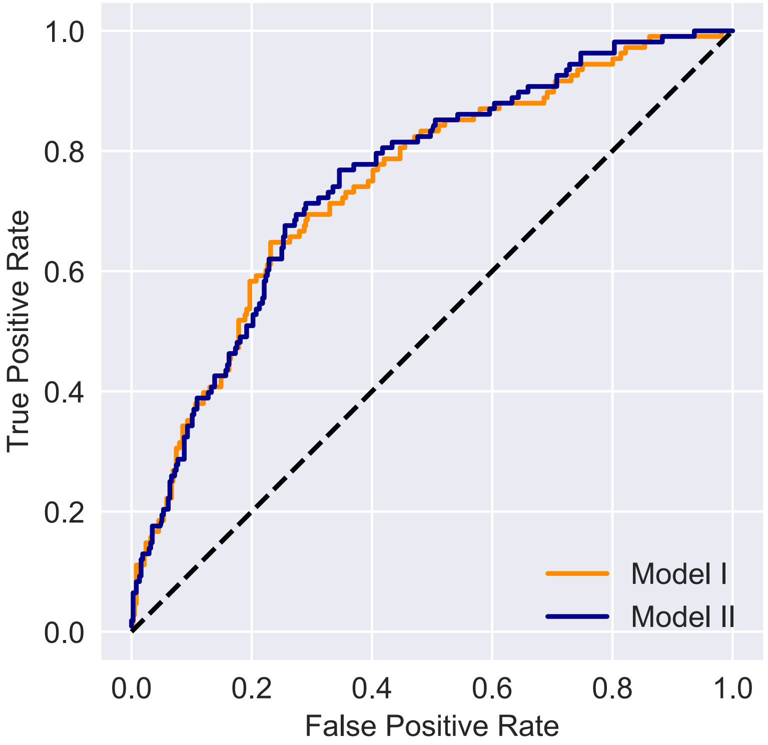


Figure B.1: Comparison of ROC curves of logistic models

Variable	Category	Coefficient	z-score	Importance
Intercept		-0.1696	-0.7749	
Loan amount	49,001–100,000	reference value		
	100,001–155,999	0.1941**	2.1874	17
	156,000–200,999	0.1136	1.2714	22
	201,000–290,999	-0.1127	-1.2498	31
	291,000–544,999	-0.0795	-0.9009	33
	≥ 545,000	0.0559	0.5510	37
Loan type	Consolidation	reference value		
	Loan for housing	-0.3518***	-5.8100	7
Loan year	2016	reference value		
	2015	0.2311***	3.3107	12
	2014–2013	0.5069***	7.0857	4
	2012–2006	0.3923***	4.5241	6
Loan term (days)	225–1,852	reference value		
	1,853–2,549	-0.1296	-1.3901	19
	2,550–2,588	0.1133	1.2907	28
	2,589–2,938	0.2960***	3.0541	8
	2,939–3,654	0.1110	1.3019	29
	≥ 3,655	-0.0257	-0.3130	42
Age	18–29	reference value		
	30–36	0.0934	1.1916	26
	37–38	-0.0395	-0.4056	40
	39–48	-0.1005	-1.3891	34
	49–54	-0.3881***	-4.0576	5
	≥ 55	-0.0615	-0.6603	36
Children	0	reference value		
	1	-0.1110*	-1.9049	25
	≥ 2	0.0277	0.2107	41
Region	Prague	reference value		
	Central Bohemian	0.1794	1.5964	18
	South Bohemian	0.0998	0.8847	32
	Plzeň	-0.1378	-1.1072	23
	Karlovy Vary	-0.2594**	-1.9738	10
	Ústí nad Labem	-0.0187	-0.1737	43
	Liberec	0.2026	1.6038	16
	Hradec Králové	-0.0407	-0.3137	39
	Pardubice	0.0137	0.0938	45
	Vysočina	-0.1029	-0.6388	30
	South Moravian	-0.0499	-0.4691	38
	Zlín	0.0196	0.1620	44
	Moravian-Silesian	0.0620	0.6301	35
	Olomouc	-0.1269	-1.1888	24
Education	Primary	reference value		
	Secondary	-0.1493	-0.9743	20
	Vocational	0.2448	1.6008	3
	University	-0.01288	-0.7802	21
	Other	-0.2509	-1.5573	14
Tenure profile (months)	0	reference value		
	0.01–20.99	0.1818	1.2960	13
	21–90.99	0.1301	0.9810	15
	≥ 91	-0.1029	-0.7844	27
Monthly income	11,000–21,100	reference value		
	21,101–27,700	-0.2365***	-3.3911	11
	27,701–33,000	-0.2890***	-3.7680	9
	33,001–61,000	-0.5668***	-8.4172	2
	≥ 61,101	-0.7752***	-8.0851	1

* $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$

Table B.1: Estimated coefficients of Model I

Variable	Category	Coefficient	z-score	Importance
Intercept		-0.1253	-0.7267	
Loan type	Consolidation	reference value		
	Loan for housing	-0.3173***	-5.3940	7
Loan year	2016	reference value		
	2015	0.2361***	3.4264	12
	2014–2013	0.5150***	7.3444	4
	2012–2006	0.4005***	4.7569	6
Loan term (days)	225-1,852	reference value		
	1,853–2,549	-0.1358	-1.4837	17
	2,550–2,588	0.1100	1.2799	23
	2,589–2,938	0.2908***	3.2072	8
	2,939–3,654	0.1055	1.3530	24
	≥ 3,655	-0.0295	-0.4074	32
Age	18–29	reference value		
	30–36	0.0796	1.0460	28
	37–38	-0.0742	-0.7845	29
	39–48	-0.1379**	-1.9953	16
	49–54	-0.4224***	-4.6675	5
	≥ 55	-0.1066	-1.1870	25
Children	0	reference value		
	1	-0.1311**	-2.2792	20
	≥ 2	0.0065	0.1291	37
Region	Prague	reference value		
	Central Bohemian	0.1787*	1.6561	15
	South Bohemian	0.1015	0.9098	26
	Plzeň	-0.1304	-1.0584	19
	Karlovy Vary	-0.2561**	-1.9657	10
	Ústí nad Labem	-0.0238	-0.2236	36
	Liberec	0.1852	1.4820	14
	Hradec Králové	-0.0244	-0.1903	35
	Pardubice	0.0275	0.1914	33
	Vysočina	-0.0872	-0.5485	27
	South Moravian	-0.0535	-0.5082	31
	Zlín	0.0258	0.2143	34
	Moravian-Silesian	0.0622	0.6366	30
	Olomouc	-0.1309	-1.2334	18
Education	Primary	reference value		
	Secondary	-0.1418	-0.9365	22
	Vocational	0.2671*	1.7632	3
	University	-0.1529	-0.9403	21
	Other	-0.2562	-1.6122	13
Monthly income	11,000–21,100	reference value		
	21,101–27,700	-0.2350***	-3.4125	11
	27,701–33,000	-0.2918***	-3.8789	9
	33,001–61,000	-0.5730***	-8.7340	2
	≥ 61,101	-0.7895***	-8.4241	1

* $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$

Table B.2: Estimated coefficients of Model II

Appendix C

Classification Tree Diagram

This appendix presents classification tree diagram. Blue color presents bad borrowers, whereas orange color stands for good borrowers. The more radiant the color in the terminal node is, the higher percentage of specific class is represented in the node. This is captured by values in square brackets which present the proportion of both classes in given node. Although only one terminal node leads to a bad category, the remaining nodes are crucial for probability computation used for ROC curve generation while using different thresholds.

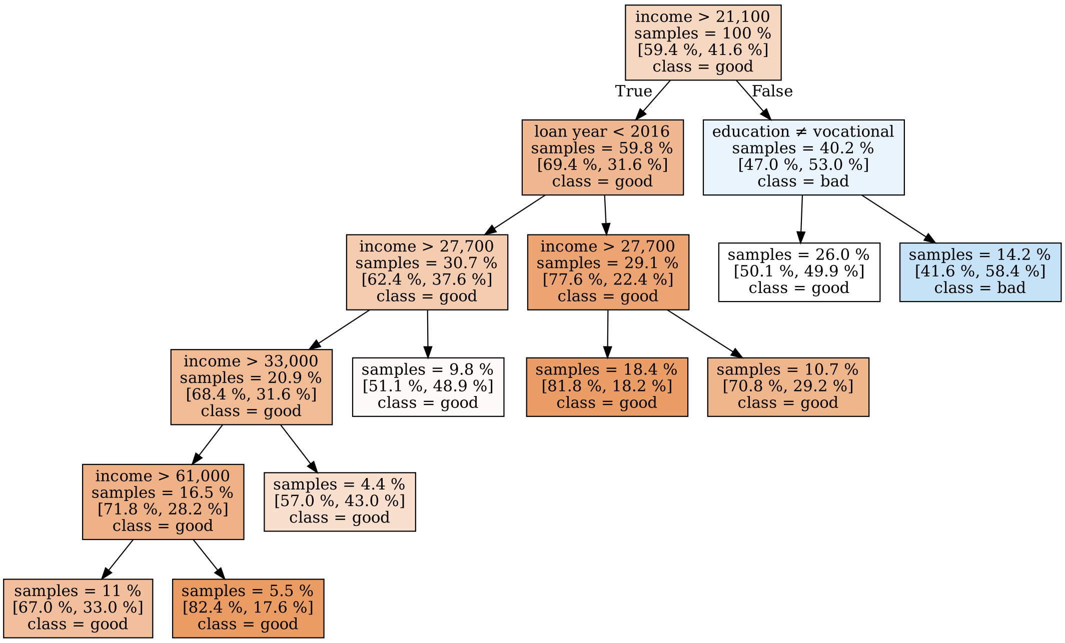


Figure C.1: Classification tree diagram

Appendix D

Confusion Matrices

This appendix presents confusion matrices for all employed classification methods on testing dataset.

Conf. Matrix		Predicted	
		Good	Bad
Actual	Good	327	49
	Bad	64	44

(a) Logistic regression - Model I

Conf. Matrix		Predicted	
		Good	Bad
Actual	Good	328	48
	Bad	66	42

(b) Logistic regression - Model II

Conf. Matrix		Predicted	
		Good	Bad
Actual	Good	328	48
	Bad	55	53

(c) LDA

Conf. Matrix		Predicted	
		Good	Bad
Actual	Good	364	12
	Bad	101	7

(d) QDA

Conf. Matrix		Predicted	
		Good	Bad
Actual	Good	341	35
	Bad	77	31

(e) Classification tree

Conf. Matrix		Predicted	
		Good	Bad
Actual	Good	183	193
	Bad	14	94

(f) Random forest

Conf. Matrix		Predicted	
		Good	Bad
Actual	Good	364	12
	Bad	101	7

(g) KNN

Conf. Matrix		Predicted	
		Good	Bad
Actual	Good	207	169
	Bad	20	88

(h) SVM

Table D.1: Confusion matrices for all models