Charles University

Faculty of Social Sciences
Institute of Economic Studies

MASTER'S THESIS

Performance Analysis of Credit Scoring Models on Lending Club Data

Author: **Bc. Michal Polena**

Supervisor: **doc. PhDr. Petr Teplý, Ph.D.**

Academic Year: **2016/2017**

Declaration of Authorship

The author hereby declares that he 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 18, 2017

Signature

Acknowledgments

I would like to thank doc. PhDr. Petr Teplý, Ph.D. for his helpful advices and specifically for his ability to keep me well-motivated. Further, I want to thank my family and close friends for their support and love during my entire life.

Abstract

In our master thesis, we compare ten classification algorithms for credit scoring. Their prediction performances are measured by six different classification performance measurements. We use a unique P2P lending data set with more than 200,000 records and 23 variables for our classifiers comparison. This data set comes from Lending Club, the biggest P2P lending platform in the United States. Logistic regression, Artificial neural network, and Linear discriminant analysis are the best three classifiers according to our results. Random forest ranks as the fifth best classifier. On the other hand, Classification and regression tree and k-Nearest neighbors are ranked as the worse classifiers in our ranking.

JEL Classification	C10, G10, C80, C58
Keywords	Credit scoring, P2P Lending, Classification, Classifiers' ranking

Author's e-mail	michal.polena@outlook.com
Supervisor's e-mail	petr.teply@fsv.cuni.cz

Abstrakt

V naší magisterské práci jsme porovnávali deset klasifikačních algoritmů pro kreditní skórování. Jejich prediktivní schopnosti byly měřeny šesti rozdílnými technikami pro měření klasifikační přesnosti. Pro porovnání klasifikátorů jsme použili unikátní datový set z P2P půjčování s více jak 200.000 záznamy a 23 proměnnými. Tento datový set pochází z Lending Clubu. Lending Club je největší platforma pro P2P půjčování ve Spojených státech amerických. Logistická regrese, umělá neuronová síť a lineární diskriminační analýza jsou tři nejlepší klasifikátory podle našich výsledků. Náhodný les se umístil jako pátý nejlepší klasifikátor. Na druhou stranu, klasifikační a regresní strom a k-nejbližší okolí se umístily jako nejhorší klasifikátory v našem žebříčku.

Klasifikace JEL	C10, G10, C80, C58
Klíčová slova	Kreditní skóre, P2P půjčování, Klasifikace, Žebříček klasifikátorů

E-mail autora	michal.polena@outlook.com
E-mail vedoucího práce	petr.teply@fsv.cuni.cz

Contents

List of Tables

List of Figures

Acronyms

ANN Artificial neural network

B-Net Bayesian network

CAGR Compound annual growth rate

CART Classification and regression tree

P2P Peer-to-peer

k-NN k-Nearest neighbors

NB Naïve Bayes

LDA Linear discriminant analysis

LR Logistic regression

L-SVM Linear support vector machine

SVM-Rbf Support vector machine with radial basis kernel function

RF Random forrest

Master's Thesis Proposal

Author	Bc. Michal Polena
Supervisor	doc. PhDr. Petr Teplý, Ph.D.
Proposed topic	Performance Analysis of Credit Scoring Models on Lending Club Data

Motivation P2P lending platforms, a new financial intermediary between borrowers and lenders, experience an astonishing growth since their inception. For example, the biggest P2P lending platform in USA, Lending Club, almost double the amount of issued loans each year[1]. P2P lending is rapidly growing in Europe as well. Wardrop *et al.* (2015) has showed that P2P consumer lending more than doubled the amount of lend money each year since 2012.

Our master thesis will be based on the data provided by Lending Club. Lending Club publishes information about all issued loans on their websites. For purpose of our thesis, we have taken a data set of loans issued between January 2009 and December 2013. Our data set contains more than 200,000 loans and we know the final status of all loans. We can extract training and testing data sample from our data set. Moreover, our data set is large enough to ensure inter-temporal validation.

It is essential for P2P lending platforms to decrease the information asymmetry between lenders and borrowers. Therefore, the borrowers are required to provide some information about themselves and the loan characteristics. Based on this information, P2P lending platforms use their credit scoring models to properly assess borrowers's credit risks. Well performing credit scoring model is pivotal for P2P lending platforms' success. Nevertheless, as researched by Abdou & Pointon (2011)'s meta-analyses including more than 200 articles about credit scoring models, there is no single credit scoring method outperforming others.

[1]Statistics from Lending Club webpage: https://www.lendingclub.com/info/statistics.action

There are, nevertheless, many papers comparing different classification methods. We do consider Baesens *et al.* (2003) and Lessmann *et al.* (2015)'s to be the most comprehensive papers comparing classifiers. The performance ranking of chosen classification methods from Lessmann *et al.* (2015) is taken as our baseline.

To the best of our knowledge, there are only four papers (Wu, 2014; Tsai et al., 2014; Chang et al., 2015; Malekipirbazari & Aksakalli, 2015) comparing classification methods based on the Lending Club data. The performance ranking of given classifiers differs in the above-mentioned papers. We do see several shortcomings, such as data preparation and performance measurements, in these studies. Moreover, only limited number of classifiers were used in these studies.

The purpose of our master thesis is comprehensive performance comparison of various classification methods based on the Lending Club data set. Furthermore, we want to overcome methodological shortcomings of afore-mentioned papers. Our main contribution might be divided into three parts: data set used, number of classifiers scored and performance measurement used.

Hypotheses

Hypothesis #1: Random forest is the best classification method based on the Lending Club data.

Hypothesis #2: Artificial neural network outperforms Logistic regression based on the Lending Club data.

Hypothesis #3: Linearly based classification methods under-perform on Partial Gini index compared to other performance measurements.

Hypothesis #4: Lending Club data is only weekly non-linear.

Hypothesis #5: Logistic regression outperforms Support Vector Machine based on the Lending Club data.

Methodology Our data set with loans issued between January 2009 and December 2013 contains more than 200,000 loans. We know the final status of issued 36-months loans from this given period as they had time to mature. Furthermore, we know the final status of 60-months loans issued in 2010. Furthermore, we do not use loans from 2007 and 2008 because they have higher default rate and might be influenced by the financial crisis in 2007/9.

We compare ten different classification methods which makes our comparison comprehensive. The classification methods used are Random forest, Artificial neural network, Logistic regression, Linear discriminant analysis, Support vector machine with radial basis kernel function, Linear support vector machine, Bayesian network, Naive Bayes, k-nearest neighbor, Classification and regression tree.

The models are evaluated based on several performance measurements because each measuremnt may favour some classification method. The mesaurement method used are: Area under curve (AUC), Brier Score (BS) and Partial Gini index (PG).

Expected Contribution Our master thesis is the first study comprehensively comparing several classification methods based on the Lending Club (LC) data. Our contribution has three main areas. The first area of contribution is the magnitude of data set used. We have the largest data set among the studies using LC data. Moreover, our data preparation approach is more exhaustive and accurate than in papers mentioned above. The second area of contribution is the number of used classification methods. The previous studies based on LC data have used at most 4 methods. We use ten methods which makes our comparison comprehensive. Finally, we use three different types of performance measurements. Using different measurement techniques make our findings robust. We do believe that the above-mentioned reasons make our master thesis unique.

The logistic regression is an industry standard for classification. Our results might show that some classification methods significantly outperform logistic regression. Therefore, these classification methods could be interesting for companies willing to improve their current credit scoring.

Outline

1. Introduction
2. Literature review
3. Hypotheses development
4. Methodological background
5. Data preparation
6. Results
7. Discussion
8. Concluclusion

Core bibliography

Abdou, H. & J. Pointon (2011): "Credit scoring, statistical technique and evaluation criteria: a review of the literature." Intelligent systems in accounting, finance and management 18(2-3): pp. 59-88.

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.

Chang, S., S. D.-o. Kim, & G. Kondo (2015): "Predicting Default Risk of Lending Club Loans." Machine Learning pp. 1-5.

Mills, K. G. (2014): "The State of Small Business Lending: Credit Access during the Recovery and How Technology May Change the Game." Harvard Business School Working Paper (No. 15-004).

Malekipirbazari, M. & V. Aksakalli (2015): "Risk assessment in social lending via random forests." Expert Systems with Applications 42(10): pp. 4621-4631.

Lessmann, S., B. Baesens, H. V. Seow, & L. C. Thomas (2015): "Benchmarking state-of-the-art classification algorithms for credit scoring: An update of research." European Journal of Operational Research 247(1): pp. 124-136.

Serrano-Cinca, C., B. Gutierrez-Nieto, & L. Lopez-Palacios (2015): "Determinants of Default in P2P Lending." Plos One 10(10): p. 1-22.

Tsai, K., S. Ramiah, & S. Singh (2014): "Peer Lending Risk Predictor." Stanford University CS229 Project Report.

Wardrop, R., B. Zhang, R. Rau, & M. Gray (2015): "The European Alternative Finance Benchmarking Report." Universtiy of Cambridge Report

Wu, J. (2014): "Loan default prediction using lending club data." Available at http://www.wujiayu.me/assets/projects/loan-default-prediction-Jiayu-Wu.pdf pp. 1-12.

Author Supervisor

Chapter 1

Introduction

Warren Buffet famously said that he follows two rules for investing his money. The first rule says: "Never lose money." and the second rule adds: "Never forget the first rule." This rule in the context of retail banking or P2P lending means that the money should not be lent to someone who will not pay them back. There exist numerous classification algorithms, such as Logistic regression or Random forest, for assessment of borrower's creditworthiness. These classification techniques support the decision-making process of whether to lend money to a borrower or not.

The remaining question is what the best credit scoring algorithm is?[1] There already are some comparison studies of classification algorithms, such as Baesens *et al.* (2003) and Lessmann *et al.* (2015), that provide their ranking of classifiers. In line with Salzberg (1997) and Wu (2014), we are, however, concerned about the relevance of these findings for applications in the real world because these studies are usually based on small data sets with unknown source of origin. This might be a problem because classifier's predictions are only as good as the data sets used for its training.

We propose five hypotheses about classifiers' performance based mainly on the Baesens *et al.* (2003) and Lessmann *et al.* (2015)'s findings. The first and second hypotheses are that Random forest and Artificial neural network are better classifiers than Logistic regression. The third hypothesis says that linearly based classifiers rank in the better half of classifiers ranking. The next hypothesis proposes that Logistic regression outperforms Support vector ma-

[1] Throughout our master thesis, we use terms credit scoring algorithm, classification algorithm, classification technique, and classifier interchangebly.

chines with different kernel functions. The last hypothesis declares that linearly based classifiers underperform when their performance is measured by Partial Gini index in comparison to other performance measurements.

The main contribution of our master thesis is the robust ranking of ten different classification techniques based on a real-world P2P lending data set. We have at hand unique P2P lending data set with more than 200,000 records and 23 variables from Lending Club where the final loan resolution is known.[2] Besides that, our ranking is robust because we use 5-fold cross-validation method and six different classifiers' performance measurements. The hypotheses from the previous paragraph are then tested with help of our classifiers' performance (Table 7.1) and ranking results (Table 7.2).

The thesis is organized as follows. In Chapter 2, we theoretically describe P2P lending and credit scoring. Next chapter is devoted to the overview of current literature comparing classification techniques. Based on the literature review, we state our hypotheses in Chapter 4. Chapter 5 is dedicated to the introduction and description of our unique P2P lending data set. The classification algorithms and performance measurements are presented in the next chapter. Our performance results and classifiers ranking is included in Chapter 7. Furthermore, Chapter 7 includes a summary of our key findings and recommendations for further research. The last chapter concludes our master thesis.

[2]Lending Club is the biggest P2P lending platform in the United States with more than 2 million of issued loans in the total value of $24.6 billion by the end of 2016. Source of information: https://www.lendingclub.com/info/statistics.action

Chapter 2

Theoretical Background

Before starting the literature review of classification techniques in the financial industry including P2P lending in Chapter 3, we provide the in-depth explanation of terms P2P lending and credit scoring. We do believe that a reader will benefit and better understand the rest of our master thesis after having clear elucidation of P2P lending and credit scoring terms. Furthermore, we briefly describe the history and current literature research of P2P lending and credit scoring. Moreover, we include a part devoted to the loan application and lending process at Lending Club. We expect that description of these procedures helps to better illuminate the way the data in our data set has been generated.

2.1 P2P Lending

Peer-to-peer (P2P) lending is a new on-line based financial intermediary connecting people willing to borrow (borrowers) with people willing to lend their money (lenders/investors).[1]. Borrowers and lenders are connected through on-line available P2P lending platforms. P2P lending platforms can provide loans with lower intermediation costs than traditional banks because of their on-line functioning. That is to say, P2P lending platforms do not have to pay for costly branches. This fact allows offering more favorable conditions for borrowers and lenders. Borrowers get on average lower interest rates on P2P

[1]There are two different kinds of people investing in P2P loans. The first kind is common individuals who are called *lenders* These *lenders* usually buy only a fraction of a loan as their personal investment. On the other hand, *investor* is term usually used for institutional investors, such as asset managers, in P2P lending terminology. The *investors* usually buy the whole loans in large quantities. These two terms are commonly used interchangeably in many research papers. We use, however, only the term *lender* throughout our master thesis for everyone on the money supply side of P2P lending.

lending platforms than in banks. Similarly, lenders with well-diversified loan portfolio achieve higher returns than on traditional saving accounts (Namvar, 2013; Serrano-Cinca et al., 2015). This makes P2P lending very popular for borrowers as well as lenders.

Besides borrowers and lenders, P2P lending is popular for researchers too. There are several different streams of P2P lending research topics. It is beyond the scope of our master thesis to cover them all. Therefore, we choose only the most relevant research topics. First, we introduce the research concerned with the reason for P2P lending platforms emergence. Afterwards, the current tremendous growth of P2P lending platforms is described. There are P2P lending platforms, such as Prosper or Lending Club, which make their data public. In the part 2.1.3 Research based on P2P lending data, we introduce what research has been done based on the publicly available P2P lending data. Our master thesis can be categorized in this group as it is based on P2P lending data from Lending Club. In the last part, we present the possible future of P2P lending.

2.1.1 Emergence of P2P Lending

There are different competing hypotheses for the explanation of the rapid emergence of P2P lending platforms. Havrylchyk *et al.* (2016) present three main hypotheses as possible explanations. Their first hypothesis is competition-related. The on-line based P2P lending platforms can operate efficiently with low financial intermediation cost which enables to offer lower interest rate to borrowers and higher return to lenders than traditional banks do. Namvar (2013), Wu (2014) and Tsai *et al.* (2014) are all advocates of this hypothesis. The second hypothesis, named crisis-related, is connected to the financial crisis in 2007/2008. Banks limited their supply of credit which caused credit rationing.[2] Mills (2014) mainly supports this hypothesis. Moreover, Atz & Bholat (2016) state that mistrust in the banking industry after financial crises could favor P2P lending for lenders as well as for borrowers. The third hypothesis, called internet-related, explores the readiness of society to use on-line based financial services without the need to physically visit a bank branch. In

[2]Credit rationing is a situation when lenders (the supply side of credit) limit their offers of credit even though borrowers (the demand side of credit) are willing to pay high-interest rates (Mills, 2014).

conclusion, Havrylchyk *et al.* (2016)'s findings support the competition-related hypothesis. They, furthermore, add that P2P lending spread more in areas with lower density of banks and their branches.

2.1.2 P2P Lending Growth

The first P2P lending platform was ZOPA. ZOPA was establisted in the United Kingdom in 2005. Since then, it has been possible to observe the substantial growth of P2P lending platforms all around the world. In the last couple of years, we can even speak about the exponential growth of P2P lending platforms. The compound annual growth rate (CAGR) of P2P lending in the United States and the United Kingdom is 151% since 2010 according to Leech (2015)'s study. Wardrop *et al.* (2015) state that the CAGR of P2P consumer lending in continental Europe was 113% between 2012 and 2014. The compound annual growth rate in China is expected to be even higher than in geographical areas mentioned above. Deer *et al.* (2015) and Leech (2015) estimate the CAGR of China P2P lending to be more than 200% since its inception in 2007. It is, however, tough to estimate the growth of China P2P lending platforms as there are hundreds of P2P lending platforms. Deer *et al.* (2015) states that there were 1,575 P2P lending platform in China in 2014. Comparing the countries based on the total value of issued loans, Leech (2015) claims that loans in the value of $40 billion were issued in China in 2014. This number makes China the biggest P2P lending marketplace in the world. For comparison, the loans in the value of about $11 billion were issued in the United States and in the value of about $4.6 billion in the United Kingdom in 2014. The continental Europe is significantly behind with only $0.5 billion in issued P2P loans in 2014.

2.1.3 Research Based on P2P Lending Data

The majority of scientific papers based on publicly available P2P lending data use Prosper data. Prosper was the first P2P lending platform which made its P2P lending data public in 2007.[3] Bachmann *et al.* (2011) say that availability of Prosper P2P lending data has triggered a wave of scientific contributions and interest in P2P lending. The popularity of Prosper data has been mainly

[3]Official Prosper website: www.prosper.com

caused by social network features and Dutch auction for interest rate determination that used to be part of Prosper P2P lending platform. Prosper removed the social network features from its platform after the Securities and Exchange Commission (SEC) regulation in 2008. Similarly, the Dutch auction system for interest rate determination has been removed from Prosper as well. The optimal interest rates for borrowers are currently determined by Prosper, which is considered to be a standard at P2P lending platforms nowadays. Despite the early popularity of Prosper data, P2P lending data from Lending Club are currently more popular because of their better quality and higher number of independent variables.

The research based on P2P lending data can be divided into four areas. The first area of research is mainly focused on circumstances before the loans were funded. This research is primarily based on Prosper data issued before 2008 because the data included social features and Dutch auction system as discussed in the previous paragraph. Freedman & Jin (2014) and Lin *et al.* (2013) point out to the importance of social network connections for loan funding success and associated interest rate. People with better social network connections are more likely to get their loans funded and have lower interest rates. Furthermore, Duarte *et al.* (2012) found out that borrowers who included their photo and were perceived to be trustworthy were more likely to get funded. The second area of research examines the determinants of borrower's default. Serrano-Cinca *et al.* (2015) and Carmichael (2014) found out several determinants, such as annual income or loan purpose, which are significant variables for borrower's default prediction. These two studies are then further discussed in Chapter 5, called Data Preparation. The next research area is portfolio management based on P2P lending data. For instance, Singh *et al.* (2008) divided with the help of decision tree P2P lending loans into different groups based on the loan's risk and return. They calculated then optimal portfolio based on these groups. The last area of research focuses on the comparison of classification methods based on P2P lending data. Studies based on Lending Club data are discussed in Section 3.2, named Comparison of classifiers based on Lending Club data. Our master thesis can be categorized into this research area.

2.1.4 Future of P2P Lending

P2P lending is often called the UBER of the financial industry. McMillan (2014) is a supporter of this idea and considers P2P lending as one of the biggest threads for the banking industry. Deloitte (2014)'s study: Banking disrupted adds that P2P lending platforms are connecting borrowers with lenders in a highly efficient manner with very low intermediation costs, which causes a thread for the current banking business model. These findings are in line with PwC (2015)'s study: Peer pressure. On the other hand, Foottit *et al.* (2016) in their Deloitte (2016) study argue that at least in the United Kingdom the competitive advantage of P2P lending platforms is not sufficient to threaten bank's mainstream lending. Moreover, they expect that P2P lending platforms will more collaborate and integrate with traditional banks as it has already been witnessed in the United States.

2.2 Credit Scoring

There are plenty of credit scoring definitions. We like the most the definition provided by Anderson (2007). Anderson (2007) says that the term *credit scoring* should be firstly divided into two parts - *credit* and *scoring*. The first word *credit* comes from the Latin word *credo*. *Credo* means I trust in or I believe in Latin. The word *credit* as we use it today means buy now and pay later. The second word *scoring* refers to the use of numerical methods that helps us to rank order cases to be able to differentiate between their qualities. In other words, *scoring* is a method which assigns a score or a grade describing case quality. Combining the meaning of *credit* and *scoring*, Anderson (2007) states that *credit scoring* is the use of statistical models to transform relevant data into a numerical score describing the likelihood of a prospective borrower's default.

Abdou & Pointon (2011) say that even though the history of credit can be traced back to around 2000 BC, the history of credit scoring is very short. They estimate the length of credit scoring history to be only about six decades. Moreover, Abdou & Pointon (2011) add that the credit scoring literature is very limited. According to them, the use of credit scoring started to be broadly popular at the beginning of the 21st century. The increased popularity of credit

scoring has been mainly caused by huge technological advancements and by the introduction of advanced credit scoring techniques. Credit scoring falls into the risk management category of banks and other financial institutions. Apart from that credit scoring is regarded as an indispensable part of risk management helping to maximize the expected profit from clients.

The expected profit can be maximized when borrower's default is minimized. To minimize the borrower's default, it is necessary to decrease the information asymmetry between borrowers and lenders. Borrowers have more information than lenders about their ability to pay back their liabilities. Therefore, borrowers are asked to provide some information about themselves and a loan itself as a part of their loan application. The loan application process at Lending Club is described in the next section. Based on the loan application information, a credit scoring model can predict the borrower's creditworthiness. Nevertheless, to be able to predict the borrower's creditworthiness, the credit scoring model must be firstly trained on similar past loan applications data with known repayment results.

2.3 Lending Process at Lending Club

The knowledge of a lending process at Lending Club can help a reader better understand what the criteria for loan application approval are. Moreover, the reader gets to know the way the Lending Club data are generated. The borrower's credit characteristics, such as FICO score, needed for loan application approval has changed several times so far. For example, the minimum FICO score for loan application has been reduced to 600 points.[4] Nevertheless, the minimal borrower's FICO score used to be about 650 points according to our information. This number is in line with our Descriptive Statistics in Table 5.3 where the minimal FICO score in our data set is 662. As far as we know, the fundamental process of loan application has not changed, even though the loan application requirements and options alter in time.

[4]We do not have a personal experience with the lending process at Lending Club. Therefore, we do use two websites as the source of our information. The first website is the official Lending Club page with the How to Apply information: https://help.lendingclub.com/hc/en-us/articles/214496857-How-do-I-get-a-loan-. The second website is a detailed review for potential Lending Club borrowers: http://www.magnifymoney.com/blog/personal-loans/lendingclub-review-borrowers-insiders-reveal578301843.

The primary criterion that is expected to be a minimum requirement for borrower's loan application at Lending Club is the value of FICO score. As being already discussed above, the minimum FICO application score should be at least 600. The FICO score ranges from 300 to 850. The higher the FICO score, the more creditworthy is the borrower. Borrower's credit file information of national credit bureaus in the United States are the main source for FICO score computation. The exact formula for FICO score calculation is, however, secret. Nevertheless, it has been disclosed that FICO score is computed based on following five components with associated weights in percentage: 35% payment history, 30% debt burden, 15% length of credit history, 10% type of credit used and 10% of recent credit inquiries.[5]

After passing the minimum requirement for loan approval described above, the borrower needs to provide some more information about himself or herself and a required loan. At first, the borrower is asked about his or her self-reported annual income. Afterwards, the borrower should choose his or her current home situation with possible options: mortgage, rent, own or other. The employment status is asked next. The length of employment is known from a borrower file based on his or her Social Security Number (SSN). Concerning the loan information, the borrower is asked for a loan amount, a loan purpose and a loan description. The loan amount ranges from $1,000 to $35,000 in our data set. The loan purpose has fourteen different categories described in Table 5.4. The information about loan purpose is mandatory. The loan description is optional and is, therefore, often left blank. Our descriptive statistics show that the median length of loan description is 0 and the mean value is 103 characters.

Based on the afore-mentioned borrower's credit file information and his or her inputs, Lending Club's credit scoring algorithm determines a borrower's creditworthiness. The borrower's creditworthiness is represented by assigned credit grade with a related interest rate. Immediately after being scored, loan listing offer with obtained interest rate is offered to the borrower. If the borrower accepts given loan offer, the loan is listed on the Lending Club platform. A potential lender can right away find and fund the loan among Lending Club loan listings. During the loan funding period, the borrower might be asked by Lending Club to verify his or her self-reported annual income. If the loan is,

[5]The official FICO score website with information about FICO score composition: http://www.myfico.com/credit-education/whats-in-your-credit-score/

however, funded in the meantime, then the loan is issued and verification is not anymore needed. 65.1% of loans in our final data set are verified. The default rate of verified loans is surprisingly higher (17.8%) than the default rate of not verified loans (12.3%).[6] The Lending Club might know based on the borrower's credit file if the verification is needed or not. If the borrower, however, fails to verify his or her self-reported information, Lending Club removes the listed loan from its platform.

[6]For further information regarding the descriptive statistics of our data sets, please, refer to Chapter 5.

Chapter 3

Literature Review

We divide our literature review into two parts. The first part is devoted to the current research papers comparing the classification techniques for credit scoring. This part is connected to the credit scoring part discussed in Chapter 2. The second part of our literature review explores the current literature comparing classifiers based on the Lending Club data.

3.1 Comparison of Classification Techniques

A proper credit scoring technique is a vital part of long-term success for financial institutions including P2P lending platforms. Abdou & Pointon (2011) conducted an in-depth review of 214 articles and books concerned with applications of credit scoring in various areas of business. They found out that there does not exist single overall best classification technique for creation of credit scoring models. Abdou & Pointon (2011) in line with Hand & Henley (1997) argue that performance of classification techniques depends on many characteristics. These characteristics might be available variables in data set, data structure or just the objective of classification.

Even though one single best credit scoring technique might not exist according to the Abdou & Pointon (2011), the amount of literature comparing different classification algorithms is very rich. The majority of those studies, such as Yeh & Lien (2009), Tsai *et al.* (2009) or Akkoc (2012), introduce some new classification method. These new classifiers are then usually compared with a limited number of classifiers including Logistic regression. Logistic regression is regarded as an industry standard for credit scoring models (Ala'raj

& Abbod (2015)). This approach is criticized by Lessmann *et al.* (2015). Lessmann *et al.* (2015) argue that comparing some new classification method, often specifically fine-tuned and without any prior hypotheses, to the limited number of classifiers and showing its performance superiority to Logistic regression is not a signal of methodological advancement.

Another issue, we observe in studies comparing different classification techniques, is the choice of dataset. Some studies, such as Zhang *et al.* (2007) and Chuang & Lin (2009), use the Lichman (2013)'s data sets called Australian and German credit data. Both these data sets are freely downloadable from UCI Machine Learning Repository. The Australian credit data set has 690 observations with 14 independent variables and default rate of 44.5%. The German credit data set has 1,000 observations with 20 independent variables and default rate of 30%. We regard both these data sets as inappropriate for classifiers' comparison because of a low number of observations. Wu (2014) has the same opinion regarding the data set size. Furthermore, although high default rates of 44.5% and 30% ensure balanced data sets, they do not correspond to the reality as we see it.

Based on our literature review comparing classification techniques, we consider two studies methodologically outstanding. These studies are Baesens *et al.* (2003) and Lessmann et al. (2015). The latter study is an update of the former study incorporating new findings, such as new classifiers, performance criteria and statistical testing procedures. Furthermore, Lessmann *et al.* (2015) include more data sets than Baesens *et al.* (2003). Altogether, Lessmann *et al.* (2015) compare 41 different classification algorithms based on eight data sets measured by six various measurement methods.

For the purpose of our master thesis, we take Lessmann *et al.* (2015)'s results of nine chosen individual classifiers and one homogenous ensemble classifier. These results were relatively recalculated for classifiers of our interest and are depicted in Table 3.1. The nine individual classifiers are: Artificial neural network (ANN), Logistic regression (LR), Linear discriminant analysis (LDA), Support vector machine with radial basis kernel function (SVM-Rbf), Linear support vector machine (L-SVM), Bayesian network (B-Net), Naïve Bayes (NB), k-Nearest neighbors (k-NN), Classification and regression tree (CART). The only homogenous ensemble classifier is Random forest (RF). We include

Random forest (RF) in our comparison because of its popularity and wide range of usage. Each classifier was tested on eight different data sets, and its results were measured by six various performance measurement techniques in Lessmann *et al.* (2015)'s study. The six performance measures are Area under the curve (AUC), Percentage correctly classified (PCC), Brier score (BS), H-measure (H), Partial Gini index (PG) and Kolmogorov-Smirnov statistic (KS). For example, looking at classifier Logistic regression, abbreviated as LR in Table 3.1, a reader can see that Logistic regression was third best algorithm measured by area under the curve, abbreviated as AUC, methodology based on the eight data sets. Overall, the Logistic regression (LR) ranks as third best classifier out of our ten chosen classifiers.

We do consider Lessmann *et al.* (2015)'s results in Table 3.1 as a baseline ranking for classifiers of our interest. The main goal of our master thesis is, however, to find the best classification technique based on the Lending Club data set. Lessmann *et al.* (2015) do not use Lending Club data in their study. Therefore, as Abdou & Pointon (2011) and Hand & Henley (1997) argue, we might arrive at completely different results than Lessmann *et al.* (2015) based on the Lending Club data set.

3.2 Comparison of Classifiers Based on Lending Club Data

As far as we know, there are only three P2P lending platforms that make their data about issued loans and borrowers characteristics public. These platforms are Bondora, Prosper and Lending Club. We have not found any studies comparing classification techniques based on Bondora or Prosper data sets. Bondora is the youngest P2P lending platform among these three platforms. Even though Bondora was founded in 2009, it experienced the first rapid growth in 2013. The number of issued loans in January 2013 was almost fourteen thousand and more than three times more in January 2014.[1] The average loan duration at Bondora is 47 months. It means that majority of loans has not yet reached their maturity to be properly analyzed. We assume that current immaturity of loans issued at Bondora might be the reason why we have not

[1]Statistics from Bondora webpage: https://www.bondora.com/en/public-statistics

Table 3.1: Classifiers ranking according to the Lessmann et al. (2015)

Classifier	Performance Measurement						Avg. Score	Total Ranking
	AUC	PCC	BS	H	PG	KS		
RF	1	1	1	1	2	1	**1.2**	**1**
ANN	2	2	7	2	1	2	**2.7**	**2**
LR	3	3	2	3	6	3	**3.3**	**3**
LDA	6	4	3	4	8	6	**5.2***	**4**
SVM-Rbf	4	5	8-9	5	4	4	**5.2***	**5**
SVM-L	5	6	8-9	6	3	5	**5.7**	**6**
B-Net	7	7	4	7	7	7	**6.5**	**7**
NB	9	8	5	8	5	8	**7.2**	**8**
k-NN	8	9	6	9	10	9	**8.5**	**9**
CART	10	10	10	10	9	10	**9.8**	**10**

Source: *Authors' own recalculation of Lessmann et al. (2015)'s ranking results.*
Classifiers: *RF - Random forest, ANN - Artificial neural network, LR - Logistic regression, LDA - Linear discriminant analysis, SVM-Rbf - Support vector machine with radial basis kernel function, SVM-L - Linear support vector machine, B-Net - Bayesian network, NB - Naïve Bayes, k-NN - k-Nearest neighbors, CART - Classification and regression tree.*
Performance measurements: *AUC - Area under curve, PCC - Percentage correctly classified, BS - Brier score, H - H-measure, PG - Partial Gini index, KS - Kolmogorov-Smirnov statistic*
Avg. Score: *Average score computes the average ranking of classifier based on rankings achieved under different performance measurements.*
** LDA and SVM-Rbf have in our simplified version of Lessmann et al. (2015)'s findings the same average score. This is because of the simplification. In Lessmann et al. (2015), LDA outperforms SVM-Rbf.*

Total Ranking: *Total ranking ranks classifiers based on their average score.*

yet seen any study comparing classifiers based on these data. There has already been written many papers, such as Herzenstein *et al.* (2011) and Zhang & Liu (2012), based on the Prosper data set. Most of these papers, such as Pope & Sydnor (2011) and Duarte *et al.* (2012), are mainly concerned with the social features of Prosper. For example, Lin *et al.* (2013) state that borrowers with stronger network relationships are less likely to default. We suppose that it is impossible to isolate the effect of the social features in the Prosper data. Therefore, these data might not be suitable for the comparison of classification techniques. Lending Club does not support any social features, and all loans in

our data set are matured. Therefore, we firmly believe that the data, we have at hand, are convenient for the comparison of classification methods.

To the best of our knowledge, there are only four studies (Wu, 2014; Tsai et al., 2014; Chang et al., 2015; Malekipirbazari & Aksakalli, 2015) comparing classification methods based on the Lending Club data.[2] We focus on three aspects of these studies. The first aspect is the data set used. We are interested to know from which year the data are, how many observations is used for classification and how many variables are considered. The second aspect is the use of classifiers. Only one study uses five classifiers, two studies use four classifiers and once even only two classifiers are used. The last aspect, we are interested in, is the use of performance measurement techniques. Most of the studies used three performance measurements. The most popular measurement technique is Percentage correctly classified (PCC) that is used in three studies.

Each of the above studies was written with a different purpose. For example, Wu (2014) argues that the performance of classifiers in the Kaggle.com competition called 'Give me some credit' cannot be taken as credible. Wu (2014) criticizes the Kaggle data set for being artificially created, and that's why dubious. The primary goal of her study is to compare Logistic regression and Random forest on the real data set. Next, the purpose of Tsai *et al.* (2014) research is to avoid as many false positive predictions as possible. They, therefore, use precision as performance measurement. Moreover, Tsai *et al.* (2014) use modified version of Logistic regression with penalty factor to avoid false positive predictions. Chang *et al.* (2015) compares the performance of different Naïve Bayes distributions and kernel methods for Support vector machine. Chang *et al.* (2015) found that Naïve Bayes with Gaussian distribution and Support vector machine with linear kernel have the best performance based on the LC data. The main aim of Malekipirbazari & Aksakalli (2015) was to compare different machine learning algorithms. Even though Malekipirbazari & Aksakalli (2015) identified the Random forest as best scoring classifier, the ranking of remaining classifiers is not in line with Lessmann *et al.* (2015)'s findings.

[2] Jin & Zhu (2015) compare classification methods based on the Lending Club data set as well. They classify, however, predictions into three or four categories instead of two which makes the comparison with above-mentioned papers challenging. We do not, therefore, include this paper into our comparison.

Table 3.2: Classifiers comparison based on the Lending Club data

Credit scoring studies based on Lending Club data	**Data**			**Classifiers**							**Performance measurement technique**
	Year	**# of observations**	**# of variables**	**LR**	**RF**	**k-NN**	**NB**	**L-SVM**	**SVM-P**	**SVM-Rbf**	
Wu (2014)	2007-2011	33 571	22	1	2						PCC, AUC
Tsai et al. (2014)	2007-2013	91 520	n/a	1	3		4			2	PVV
Chang et al. (2015)	2007-2015	n/a	n/a	3			1	2	4	5	PCC, G-mean
Malekipirbazari & Aksakalli (2015)	2012-2014	68 000	16	4	1	3			2		PCC, AUC, RMSE

Source: *Authors' information extraction and ranking computation based on Wu(2014), Tsai et al. (2014), Chang et al. (2015) and Malekipirbazari & Aksakalli (2015)'s research.*
Classifiers: *LR - Logistic regression, RF - Random forest, k-NN - k-Nearest neighbors, NB - Naïve Bayes, L-SVM - Linear support vector machine, SVM-P - Support vector machine with polynomial kernel function, SVM-Rbf - Support vector machine with radial basis kernel function.*

Performance measurements: *PCC - Percentage correctly classified, AUC - Area under curve, PVV - Postive predictive value / Precision, RMSE - Root-mean-square error*

We do consider Lessmann *et al.* (2015) classifier ranking as our baseline. Nevertheless, none of the above-discussed papers using Lending Club data includes classifier rankings which resembles Lessmann *et al.* (2015)'s findings. One of the possible explanations, as stated by Hand & Henley (1997) and Abdou & Pointon (2011), might be the particular structure of data set, available variables or purpose of classification. In other words, studies based on the Lending Club data may produce different rankings than the one from Lessmann *et al.* (2015). To be able to confirm this, we need to overcome shortcomings of studies from Table 3.2.

In our opinion, the studies based on Lending Club data set have three different types of shortcomings. The first shortcoming is the type of used data set. We explain this deficit more deeply in Chapter 5 which is devoted to the data preparation. The second shortcoming is the comparison of a limited number of classifiers. At most five classifiers are compared. Moreover, none of the studies includes nowadays very popular Artificial neural network (ANN) or other classification techniques, such as k-Nearest neighbors (k-NN). The last issue, we observe, is the choice of classifiers performance measurements. For example, Tsai *et al.* (2014) use only one performance measurement for classifier comparison. We firmly believe that performance results based on one performance measurement cannot be robust. Nevertheless, more performance measurements from the same measurement group might not improve the situation as well.[3] Chang *et al.* (2015) uses, for instance, Percentage correctly classified, Precision and G-mean as performance measures. All of them are, however, from the same performance measurement group based on the confusion matrix. Thus, the Chang *et al.* (2015)'s results might not be robust because measurement techniques from this performance measurement group could favor some classification techniques. Furthermore, none of the studies from Table 3.2 has a general framework for classifiers performance ranking. In other words, there is no overall classifier performance ranking based on the chosen performance measurements. In most cases, classifiers performance is compared based on individual performance measurements. We create classifiers performance ranking in Table 3.2 based on average classifier ranking across chosen performance measures. For example, in Malekipirbazari & Aksakalli (2015)'s paper in Table 3.2, Random forest (RF), labeled with 1, is the best classifier based on average

[3]More information about different performance measurement groups and their advantages and disadvantages might be found in Section 6.2 Performance Measurements.

ranking of all three performance measurements. The last shortcoming, we observe, is the absence of modern performance measurement techniques, such as Partial Gini index or Kolmogorov-Smirnov statistics. Neither of studies from Table 3.2 uses these measures.

The main contribution of this master thesis is to overcome shortcomings of the papers mentioned above, and to carry out the proper comparison ranking of classifiers based on the Lending Club data set. Our contribution has three parts. The first part is the data set used. Compared to the studies in Table 3.2, we use the largest data set. Our data set has more than 200,000 records. All of our records are matured. In other words, we know their final loan resolution status. Moreover, our data preparation approach is more exhaustive and accurate than in papers from Table 3.2. The second part of our contribution is the number of classification methods used. The previous studies based on Lending Club data have used at most five methods. We use ten classification methods, which makes our comparison comprehensive. Finally, we use six performance measurements from three different performance measurement groups. Using different measurement techniques makes our findings robust. We do believe that the aforementioned reasons make our master thesis unique.

Chapter 4

Hypotheses

As stated in the literature review part, we consider Lessmann *et al.* (2015)'s findings as the baseline for our classifiers performance ranking. Their findings are depicted in Table 3.1. Lessmann *et al.* (2015)'s findings state that Random forest is the best classifier among the classification algorithms of our interest. This finding is in line with Malekipirbazari & Aksakalli (2015) who conducted their research based on the Lending Club data. Malekipirbazari & Aksakalli (2015) used, however, only four classification methods altogether. Moreover, Wu (2014) and Tsai *et al.* (2014) argue that Logistic regression is better classifier than Random forest based on the Lending Club data. Neither of these studies based on the Lending Club data, nevertheless, compares Random forest to other individual classification techniques, such as Artificial neural network or Bayesian network. Based on our baseline ranking from Table 3.1, we state following hypothesis:

H1: *Random forest is the best classifier among our classifiers of interest based on the Lending Club data.*

Artificial neural network is an attractive alternative to conventional classification techniques, such as Logistic regression or Linear discriminant analysis. Abdou *et al.* (2008) in line with Lessmann *et al.* (2015)'s findings found out that Artificial neural network outperforms Logistic regression that is considered to be an industry standard. Nevertheless, Artificial neural network was not used in either of studies based on the Lending Club data from Table 3.2. Based on the above stated findings, we suggest:

H2: *Artificial neural network outperforms Logistic regression based on the Lending Club data.*

Baesens *et al.* (2003) conducted their comparison of classification methods based on eight real-life credit scoring data sets. They measured the performance of given classifiers by the Percentage of correctly classified cases (PCC) and by the Area under the curve (AUC). Baesens *et al.* (2003) found out that Logistic regression and Linear discriminant analysis performed very well. According to Baesens *et al.* (2003), this finding indicates that the credit scoring data sets used in their study are only weakly non-linear. Moreover, looking at our benchmark ranking in Table 3.1, we can observe that Logistic regression is the third and Linear discriminant analysis is the fourth best classifier. As our Lending Club data is a credit scoring data, we anticipate a good ranking of linearly based classifiers. That's why we introduce following hypothesis:

H3: *Linearly based classifiers rank in the first half of classifiers' ranking based on Lending Club data.*

There are several different distributions of Support vector machine. The distributions differ from each other in used kernel functions. In the papers from Table 3.2., we have Support vector machine (SVM) with linear, polynomial and radial basis kernel function. The comparison of these distributions with Logistic regression yields ambiguous results. Chang *et al.* (2015) argue that Support vector machine with linear kernel function outperforms Logistic regression, but Logistic regression yields better performance than SVM with radial basis kernel function. On the other hand, Tsai *et al.* (2014) state that their modified Logistic regression has higher performance than SVM with linear kernel function. Furthermore, Malekipirbazari & Aksakalli (2015) found out that Logistic regression dominates in performance SVM with polynomial kernel function. Only one performance measure was applied in studies of Chang *et al.* (2015) and Tsai *et al.* (2014). Therefore, their results might be biased because the performance measurement could favor either Logistic regression or SVM. Our benchmark ranking says that Logistic regression outperforms SVM with linear as well as radial basis kernel function. This inference gives rise to our next hypothesis:

H4a: *Logistic regression outperforms Support vector machine with linear kernel function based on the Lending Club data.*

H4b: *Logistic regression outperforms Support vector machine with radial basis kernel function based on the Lending Club data.*

Looking at classifiers ranking in Table 3.1, linearly based classifiers, namely Logistic regression and Linear discriminant analysis, have significantly lower rankings when measured by Partial Gini index compared to the rankings achieved by remaining performance measurements. Logistic regression is ranked as the third best classifier with the average ranking score of 3.3. Based on the Partial Gini index measurement, Logistic regression is, however, the sixth best classifier. This finding is similar to ranking differences for Linear discriminant analysis too. Linear discriminant analysis is the fourth best classifier with average ranking score of 5.2. Nevertheless, Linear discriminant analysis is the eighth best classifier based on the Partial Gini index. We find this outcome as a clear indication that Partial Gini index might handicap linearly based classifiers. To test this surmise, we pose following hypothesis:

H5: *Linearly based classification methods underperform when measured by Partial Gini index in comparison with other performance measurements.*

Chapter 5

Data Description

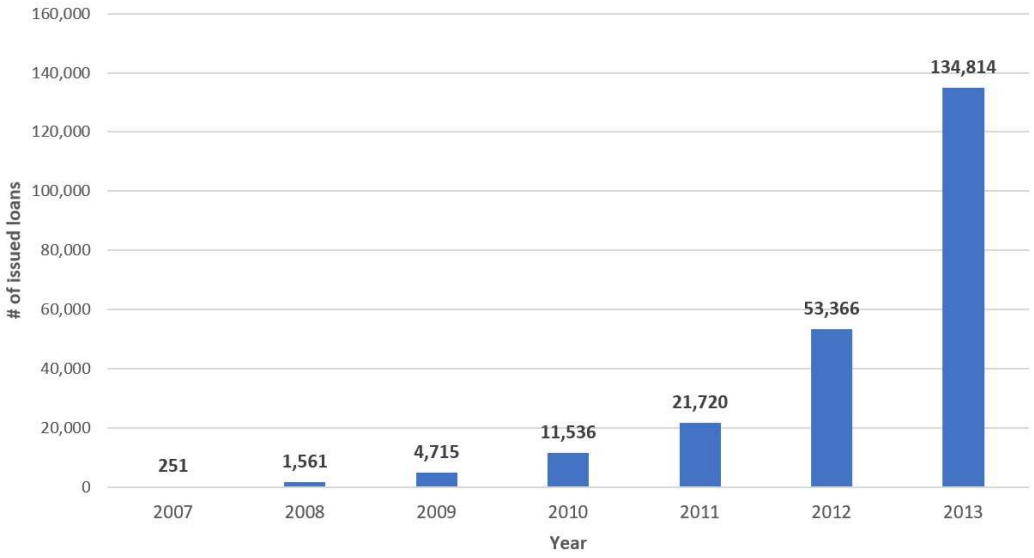
We divide our Data Description chapter into four parts. In the first part, called Data Preparation, we introduce all available data we have at hand. Furthermore, we describe why we choose the time frame from 2009 to 2013 for our data set and which variables are not appropriate for our final data set. Data Transformation is the second part of our Data Description. We describe here which variables have been transformed and how was the transformation done. The descriptive statistics of our data set are described in the third part, called Descriptive Statistics. This part includes correlation coefficient matrix of individual Lending Club variables too. The last part, called Training, Validating and Testing Data, describes our 5-fold cross-validation approach we choose for training, fine-tuning and testing of our classifiers. Besides all the above mentioned, we show the advantages of our data preprocessing methodology in comparison with papers from Table 3.2 that use Lending Club data too.

5.1 Data Preparation

Our data was downloaded from a registered Lending Club account on the 3rd March 2017. The data can be downloaded from the Lending Club statistics web page without registration as well.[1] Nevertheless, as stated by Lending Club, the full version of data files might be downloaded only after registration. For example, we observe that the size of uncompressed CSV data file from years 2012-2013 downloaded with the registered account is about 4 MB bigger than

[1]Lending Club web page for data download: https://www.lendingclub.com/info/download-data.action

Figure 5.1: Number of issued loans by years



Source: *Authors' production based on downloaded Lending Club data.*

the data file from same years downloaded without registration.[2] What are the exact differences between the data sets with and without registration is beyond the scope of our master thesis.

Our downloaded data sets from years 2007 - 2013 has 227,963 observations and 115 variables. Figure 5.1 shows the number of issued loans by years. Looking at Figure 5.1, we can observe a rapid growth of issued loans. The number of issued loans at Lending Club doubles each year. We have decided not to include loans from years 2007 and 2008 into our final data set. We believe that these loans might be influenced by the Great Financial crisis in 2007/2008. Moreover, only 1,812 loans was issued in 2007 and 2008. We do not, therefore, lose many observations in our final data set. Furthermore, there are 115 variables in the downloaded data set. All 115 variables with their description can be found in Appendix A. Majority of these variables, however, have a large number of missing values or is not relevant for our default prediction. We do discuss the detailed selection process of suitable variables later.

[2]The size of 2012-2013 CSV data file is 157,421 KB with registration and 153,016 KB without registration

We choose the time frame from 2009 to 2013 for our data set because the majority of loans has already had its maturity. The loans at Lending Club can be either issued at 36 or 60 months. Also, all our loans with 36-month duration should have already had their maturity because they were issued no later than in 2013. Besides, the majority of issued loans has 36-month duration. There are about four times more issued loans with 36-month than with 60-month duration. Further, loans with 60-month duration started to be first issued at Lending Club in 2010. Not all loans with the duration of 60 months issued from 2011 to 2013 have, however, reached their maturity yet. Such loans are filtered out. This process is described in the following part of this chapter. Even though not all our loans with 60-month duration have reached their maturity yet, we do believe that our chosen time frame is better than in papers from Table 3.2 using likewise Lending Club data. For example, Chang *et al.* (2015) uses data from 2007 to 2015. Considering only loans with 36-month duration, none of these loans issued between 2012 and 2015 could have reached their pre-arranged maturity yet. In other words, all 36-month loans issued between 2012 and 2015 having a final status are either prematurely paid off or defaulted. We do not know whether most loans are prematurely paid off or defaulted. As the number of loans, however, doubles each year, the majority of loans in the final data set can be prematurely paid off or defaulted, which might skew the default rate. Such default rate does not then correspond to the reality and makes the data set biased. Similarly to Chang *et al.* (2015), Tsai *et al.* (2014) used data set from years 2007-2013 and Malekipirbazari & Aksakalli (2015) used even data from 2012 to 2014.

Out of the 115 variables, we choose only 23 variables including the variable with final loan status result. These variables might be found in Table 5.1.[3] Otherwise, there are four main reasons for leaving out the majority of remaining variables. The first reason is that many variables, such as *open_acc_6m* and *total_bal_il*, do not include any values.[4] Another reason for leaving out variable is a high number of missing values. We leave out variables, such as *total_bc_limit* and *pct_tl_nvr_dlq* with more than five percent of missing values. Variables with

[3]The description of variables with label *No* in column Transformed are taken from the Lending Club official websites. The description of transformed variables is authors' production.

[4]We use abbreviated names of variables provided by Lending Club. These abbreviated names of variables are written in *italics* for better clarity. The full description of abbreviated names can be found in Appendix A.

Table 5.1: Included Lending Club variables

Transformed	Abbreviated Name	Description
No	acc_now_delinq	The number of accounts on which the borrower is now delinquent.
No	annual_inc	The self-reported annual income provided by the borrower during registration.
No	chargeoff_within_12_ mths	Number of charge-offs within 12 months.
No	delinq_2yrs	The number of 30+ days past-due incidences of delinquency in the borrower's credit file for the past 2 years.
No	delinq_amnt	The past-due amount owed for the accounts on which the borrower is now delinquent.
No	dti	A ratio calculated using the borrower's total monthly debt payments on the total debt obligations, excluding mortgage and the requested LC loan, divided by the borrower's self-reported monthly income.
No	home_ownership	The home ownership status provided by the borrower during registration or obtained from the credit report.
No	inq_last_6mths	The number of inquiries in past 6 months (excluding auto and mortgage inquiries).
No	loan_amnt	The listed amount of the loan applied for by the borrower.
No	open_acc	The number of open credit lines in the borrower's credit file.
No	pub_rec	Number of derogatory public records.
No	pub_rec_bankruptcies	Number of public record bankruptcies.
No	purpose	A category provided by the borrower for the loan request.
No	revol_util	Revolving line utilization rate, or the amount of credit the borrower is using relative to all available revolving credit.
No	tax_liens	Number of tax liens.
No	term	The number of payments on the loan. Values are in months and can be either 36 or 60.
No	total_acc	The total number of credit lines currently in the borrower's credit file.
No	verication_status	Indicates if income was verified by LC or not verified.
Yes	loan_status	Final status of loan has binary outcome. 0 for Fully paid loans and 1 for Charged off loans.
Yes	emp_length	Number of years in employment represented by continues variable going from 0 to 10.
Yes	desc	Number of characters included in loan description.
Yes	earliest_cr_line	Number of years since the first credit line has been opened.
Yes	fico_range_avg	The average value of fico_range_low and fico_range_high.

Legend: *The column Transformed signifies if a variable has its original form or if it has been transformed. For overview of variables' original descriptions see Appendix A. For overview of variables' transformation see section 5.2 Data Transformation.*

less than five percent of missing values, such as *tax_lieans* and *revol_util*, have their missing values replaced.[5] We replace these missing values with the help of MICE R package. The abbreviation MICE stands for Multivariate Imputation via Chained Equations. This package is well described in van Buuren & Groothuis-Oudshoorn (2011). Under the assumption that missing data are Missing at Random (MAR), the missing values are predicted by regression on

[5]Variables *tax_lieans* and *revol_util* have even less than one percent of missing records.

observed values. Continues missing values are by default predicted by linear regression. Another reason for leaving out variables was their constant values. For example, variables *pymnt_plan* and *application_type* have only one constant value for all observations. These variables do not have any significant impact on borrower's default and have been, therefore, removed. Last reason for leaving out a variable is a lack of its information value. For example, variable *url*, which represents web link to the loan listing, or variable *member_id*, assigning a unique number to a borrower, have in our opinion no information value for default prediction.

Serrano-Cinca *et al.* (2015) and Carmichael (2014) research what are the determinants of borrowers' default based on the Lending Club data set. Serrano-Cinca *et al.* (2015) identify ten variables which are significant for default prediction. Carmichael (2014) identify ten variables too. These two papers have six variables in common, for example *annual_inc* and *earliest_cr_line*. We included nine out of ten from Serrano-Cinca *et al.* (2015) and similarly nine out of ten variables from Carmichael (2014) in our final data set. According to us, this fact denotes the high quality of variables we have chosen. The only variable similar for both studies which we have not included in our data set is *grade*. The variable *grade* represents the grade assigned to the loan by Lending Club. This variable has been created by Lending Club's credit scoring system. Therefore, the default identification power of this variable is very strong. That is the reason why we leave out this variable. All variables included in our data set are depicted in Table 5.1.

5.2 Data Transformation

There are several variables we have transformed. These variables are *loan_status*, *emp_length*, *desc*, *earliest_cr_line*, *fico_range_low*, and *finco_range_high*.

Official Lending Club description of the untransformed version of these variables is included in Table A.1 in Appendix A. They are labeled with *INC**. It means that these variables are included in our data set, but they have been transformed. All the transformed variables with authors' description can be found in Table 5.1. The *loan_status* is evidently the most important variable for our purpose. It used to describe what is the current status of a loan. There were seven possible loan statuses. Overview of possible loan statuses includ-

Table 5.2: Number of loans by loan status

Loan Status	**# of loans**
Fully Paid	178,500
Charged Off	33,333
Current	13,307
Late (31-120 days)	543
In Grace Period	346
Late (16-30 days)	105
Default	17
Total	**226,151**

ing their issued numbers in our data set is displayed in Table 5.2. We are, however, only interested in having *loan_status* with the binary outcome - 0 for paid back and 1 for defaulted loans. We have, therefore, labeled all loans with status *Fully Paid* as 0 because they have been paid back.[6] Otherwise, we have filtered out all loans with status *Current* as we do not know their final status. Loans with status *Charged Off* are defaulted loans. We have labeled them with 1. There are four more loan statuses used for loans with delayed payments. A loan status *In Grace Period* means that a borrower is at most 15 days late with loan repayments. Loan statuses with names *Late (16-30 days)* and *Late (31-120 days)* are self-explaining in our opinion. Loans with status *Default* are more than 120 days past due. According to the Lending Club statistics, loans that are more than 90 days past due have 85% chance of not being paid back at all.[7] Based on this statistic, we have marked all loans with status *Default* as defaulted, thus with 1. Moreover, all loans with more than 90 days past due from *Late (31-120 days)* has been marked as defaulted too. All other loans that are past due but not more than 90 days have been filtered out. Altogether 13,871 loans have been filtered out. Our data set has 212,280 records. Comparing the size of our data set to the papers from Table 3.2 or to Lessmann *et al.* (2015), we have by far the biggest data set.

[6]For better clarity, we write the type of loan status, such as *Fully Paid*, in *italics*.

[7]Official Lending Club statistics: https://www.lendingclub.com/info/demand-and-credit-profile.action

The variable *emp_length* describes how long has been a borrower employed before asking for a loan. The values of *emp_length*, such as 1 year, 2 years and 10+ years, make this variable categorical. For better usage, we have decided to make this variable continues going from 0 to 10. The 0 value of our *emp_length* variable means that a borrower worked less than 1 year before applying for a loan. The maximal value of *emp_length* which is 10 includes all the borrowers who have worked 10 or more years by the same employer.

Every borrower can describe why he or she needs to borrow money. This loan description is included in Lending Club data under the variable *desc*. Instead of text description meaning provided by a borrower, we are interested in a number of characters a borrower used for his or her description. That's why, our variable *desc* contains number of characters used in loan description.

As being already described in Chapter 2, the length of credit history is an important part of FICO score. Furthermore, Serrano-Cinca *et al.* (2015) and Carmichael (2014) argue that the length of credit history is a significant determinant of borrower's default. In our data set, we have variable *earliest_cr_line*, in the form month-years, which represents the month and the year when the first borrower's credit line was opened. We have transformed this variable to show how many years have passed since the first credit line was opened. As we have the data from the end of 2016, we do consider the year 2017 as our reference year. For example, a borrower with *earliest_cr_line* in value of 5 has opened his or her first credit line 5 years ago.

The last variables, we have modified are *fico_range_low* and *fico_range_high*. The Lending Club data does not contain the exact value of FICO score. It contains FICO score in a range of four points with lower and upper bounds. In other words, the difference between *fico_range_high* and *fico_range_low* is four points. For our purpose, we have taken an average of these two variables. The newly created variable is called *fico_range_avg*.

5.3 Descriptive Statistics

We have chosen 23 variables including the *loan_status* for our final data set. 18 out of 23 variables are continues. During our descriptive statistics of continues variables, we found out that there were some borrowers with suspiciously high annual income. For example, there was a borrower with reported annual income of $ 7,141,778 who applied for a loan in the value of $ 14,825. We considered this record to be erroneous. Furthermore, there were altogether 28 records with reported annual income higher than $ 1 million. We deleted all these records from our data set because they might have been erroneous and skewed so our results. After this change, there are 212,252 records in our final data set. Otherwise, we have not found any suspicious values by exploration of remaining variables. The descriptive statistics of continuous variables including *loan_status* is shown in Table 5.3.

Looking at last column of Table 5.3, called t-test, we can see that for majority of our continues variables there are significant differences between their average values for loans with status *Fully Paid* and *Charged Off*. *Fully Paid* loans have significantly lower loan amount (*loan_amnt*) and longer loan description (*desc_count*) than *Charged Off* loans.[8] Borrowers who paid off their loans have higher annual income (*annual_inc*), higher FICO score (*fico_range_avg*) and longer credit history (*earliest_cr_line*) than defaulted borrowers. Besides, borrowers with *Fully Paid* loans have lower debt-to-income ratio (*dti*), were less deliquent in past two years (*delinq_2yrs*) and asked for less loans in past six months than (*inq_last_6mths*) borrowers with *Charged Off* loans. For further information about significant differences of continues variables please refer to Table 5.3.

The remaining five variables are categorical. These variables are *loan_status*, *home_ownership*, *purpose*, *term*, and *verification_status*. The descriptive statistics of categorical variables, except for *loan_status*, is depicted in Table 5.4. Table 5.4 contains a column with default rate of given categorical variables. Default rate is calculated based on *loan_status* as ratio of *Charged Off* loans to total number of loans. The overall default rate is 15.91% in our final data set.

[8]For better clarity, we add abbreviated names of variables in parentheses. The descriptive statistics of these abbreviated variable names cas be found in Table 5.3.

Table 5.3: Descriptive statistics of continuous variables

Statistics / Abbreviated Name	Mean	St. Dev.	Min	Median	Max	Average value for loan_status:		
						Fully Paid	Charge Off	t-test
loan_amnt	13,406	7,958	1,000	12,000	35,000	13,161	14,704	-31.23***
emp_length	5.849	3.585	0.000	6	10	5.836	5.919	-3.93***
annual_inc	70,986	45,017	4,000	60,000	1,000,000	72,103	65,086	28.45***
desc_count	103.1	214.2	0	0	3,959	104.1	98.11	4.67***
dti	16.29	7.56	0.00	16.01	34.99	16.02	17.72	-37.97***
delinq_2yrs	0.22	0.67	0	0	29	0.212	0.232	-3.43***
earliest_cr_line	18.88	7.00	6	18	71	18.96	18.42	13.23***
fico_range_avg	702	32	662	697	848	704	694	62.51***
inq_last_6mths	0.82	1.04	0	0	8	0.79	0.97	-27.17***
open_acc	10.65	4.61	0	10	62	10.62	10.84	-7.89***
pub_rec	0.098	0.385	0	0	54	0.097	0.104	-3.07***
revol_util	0.565	0.243	0.000	0.587	1.404	0.556	0.608	-37.02***
total_acc	24.04	11.19	2	23	105	24.1	23.71	5.77***
acc_now_delinq	0.002	0.052	0	0	5	0.002	0.003	-2.26**
chargeoff_within_12_mths	0.004	0.074	0	0	5	0.004	0.004	1.71*
delinq_amnt	6.81	476.48	0	0	65,000	6.62	7.82	-0.40
pub_rec_bancruptcies	0.078	0.279	0	0	8	0.077	0.083	-3.33***
tax_liens	0.011	0.222	0	0	53	0.012	0.011	0.87

Source: *Authors' descriptive statistics of continues variables based on Lending Club data.*

Legend: *Stars in the column t-test signify whether the difference in average values for Fully Paid and Charge Off loans is significant. *** denote significance at 1% level, ** at 5% level and * at 10% level.*

Examining Table 5.4, it is clear that the main purpose for loan application at Lending Club is *debt_consolidation* (56.7% of all loans) or repayment of *credit_cards* (21.3% of all loans). Considering the default rate for different loan purposes, the loans with purpose *small_business* are by far riskier ones (default rate of 26.2%). On the other hand, loans with purpose *car* (default rate of 11.1%) and *major_purchase* (11.9%) belong to the safest loans. Looking at the variable *home_ownership*, *mortgage* (49.6% of all loans) and *rent* (42.2%) are by borrowers the most commonly chosen options for description of their home situation. Surprisingly, people who own their home have significantly higher default rate (16.2%) than people who has mortgage (14.6%).[9] Furthermore, the majority of loans in our data set (80.6% of all loans) has 36-months duration.

[9]The p-value of this difference is $1.009 \cdot 10^{-7}$.

Table 5.4: Descriptive statistics of categorical variables

Variable name - Levels of variable	**Percentage (number) of loans**	**Default rate (%)**
purpose		
- debt_consolidation	56.7 % (120 285)	16.9 %
- credit_card	21.3 % (45 119)	13.1 %
- home_improvement	5.8 % (12 369)	13.8 %
- other	5.7 % (12 160)	18.5 %
- major_purchase	2.6 % (5 574)	11.9 %
- small_business	2.0 % (4 337)	26.2 %
- car	1.6 % (3 364)	11.1 %
- wedding	1.0 % (2 186)	12.4 %
- medical	1.0 % (2 150)	16.9 %
- moving	0.7 % (1 573)	16.7 %
- house	0.7 % (1 394)	16.3 %
- vacation	0.6 % (1 263)	15.8 %
- educational	0.1 % (260)	16.5 %
- renewable_energy	0.1 % (218)	19.3 %
home_ownership		
- mortgage	49.6 % (105 229)	14.6 %
- rent	42.2 % (89 523)	17.3 %
- own	8.2 % (17 346)	16.2 %
- other	0.1 % (114)	21.9 %
- none	0.02 % (40)	17.5 %
term		
- 36-months	80.6 % (171 137)	12.4 %
- 60-months	19.4 % (41 115)	30.5 %
verification_status		
- verified	65.1 % (138 148)	17.8 %
- not verified	34.9 % (74 104)	12.3 %

Source: *Authors' descriptive statistics of categorical variables based on Lending Club data.*

Moreover, 36-months loans have a significantly lower default rate (12.4%) than 60-months loans (30.5%). As being already discussed in section 5.1, we think that the high default rate of 60-months loans is caused by the earlier default of some loans and mainly because of the 60-months loans with the current status which were filtered out. The last categorical variable is *verification_status*. Most of the loans (65.1%) has been *verified*. They have, however, significantly higher default rate (17.8%) than loans with *not verified* information (12.3%). Based on borrower's credit file, Lending Club might not require the verification of borrower's self-reported information because the borrower is considered to be creditworthy. This approach is justifiable because of the lower default rate of loan with *not verified* information. For further information about the descriptive statistics of categorical variables see Table 5.4.

The last table of our descriptive statistics part is correlation matrix of continues variables and is included in Appendix C. We were interested to know which variables are correlated and how strong is the potential correlation. The correlation which is higher than 0.5 in absolute terms is bold highlighted. There are four cases when the correlation is higher than 0.5 in absolute terms. The strongest correlation is between *pub_rec* and *pub_rec_bankrp*. This correlation coefficient is 0.76. Moreover, variable *pub_rec* is considerably correlated (r = 0.63) with variable *tax_liens*. We have, however, decided not to exclude variable *pub_rec*. By excluding variable *pub_rec*, we could potentially lose some important information about borrower's public records which are not included in variables *pub_rec_bankrp* and *tax_liens*.

Looking at correlation matrix at Table B.1, we can observe that there are almost no correlations between *loan_status* and other variables. The variable *fico_range_avg* has in absolute terms the strongest correlation with *loan_status* and this correlation coefficient is only -0.12. Interpreted in words, the higher the FICO score, the lower the chance that the borrower will default. A reader interested in correlation between variables is advised to see Table B.1 in Appendix C.

5.4 Training, Validating and Testing Data

To be able to measure the classifiers' performance, it is needed to divide the original data set into at least two parts. One of the parts is then the testing data set on which the classifiers can be scored. There exist many different methods and their variations to split the original data set, and to create the testing data set. We describe three most common methods. The first method is a simple separation into training and testing data set. In the first step, the classifier is trained based on the training set. In the second step, a trained classifier is scored on the testing data set. The separation ratio of original data set is usually between 70% and 80% for training data and remaining 30% or 20% for testing data. The second method separates the original data set into three parts - training, validation and testing data set. After the training of classifier, the validation data set is used for model fine-tuning. The validation data set helps to improve the classifiers out of training data set accuracy. Some classifiers might have near perfect accuracy based on the training data. However, these classifiers might perform then very poorly on the testing data. Therefore, the validation data set is used to fine-tune the classifier before being scored on the testing data set. The last method is called k-fold cross-validation. We have chosen this approach because it provides valid and robust results (Salzberg, 1997; Huang et al., 2007) compared to the first and second method. The detailed explanation of k-fold cross-validation is provided in the following paragraph.

In the k-fold cross-validation method is the original data set randomly divided into k subsets. Each of the k subsets is used as testing data set in one of the k iterations. The remaining *k-1* subsets are used for model training and fine-tuning. Salzberg (1997) argues that this approach minimizes the impact of data dependency. In other words, the risk that the performance of a classifier depends on the choice of testing set is minimized because the classifier is scored sequentially on the whole data set. Moreover, Huang *et al.* (2007) add that use of k-fold cross-validation serves as guarantee of results validity.

We do use 5-fold cross-validation in our master thesis. So, our original data set has been randomly divided into five subsets. The partition was done with the help of IBM SPSS Modeler 18.0. These partitioned subsets are fairly even in terms of default rates and numbers of observations. The subsets are depicted

in the Table 5.5. We use three subsets for model training, one subset for model fine-tuning and the last subsets for model testing. As suggested by Salzberg (1997), after the model is trained and fine-tuned, we retrain the prepared model on the training and fine-tuning subsets. This process should increase the classifier's predictive power because it is retrained on a larger training set. The sequence of subsets used in different iterations is displayed in Table 5.6 below. For example in the first iteration, the subsets SS 1, SS 2 and SS 3 are used as a training data sets. Subset SS 4 is used for model fine-tuning. The fine-tuned model is then retrained on subsets SS 1, SS 2, SS 3 and SS 4. The model performance is tested based on subset SS 5. We record classifier's performance for each iteration.

The last thing, we want to cover in this section, is the imbalance of defaulted and non-defaulted loans in our data set. It has been already shown in the section 5.3 Descriptive Statistics that the default rate in our data set is 15.91%, which causes our target variable *loan_status* to be imbalanced. Brown & Mues (2012) show in their study that the higher the imbalance of target variable in data set, the worse the classifier performance. The prediction performance decreases because classifiers tend to overpredict the majority class the more the data are imbalanced. HE & Garcia (2010) proposes under-sampling of majority class, over-sampling of minority, synthetic minority over-sampling

Table 5.5: Partitioned subsets

Subset name	**# of observations**	**Default rate**
SS 1	42 533	15.92 %
SS 2	42 312	16.14 %
SS 3	42 583	15.79 %
SS 4	42 351	15.95 %
SS 5	42 473	15.77 %
Total	**212 252**	**15.91 %**

Note: *The original data set was divided into five almost equal subsets. We use abbreviations for our subsets. For example, SS 1 stands for the first subset.*

technique (SMOTE) and other techniques as a remedy for imbalanced data sets. In agreement with Lessmann *et al.* (2015)), we do, however, refrain from balancing the target variable in our data set for following reasons. Lessmann *et al.* (2015) argue that if the imbalanced data affects all classifiers in the same way, then only the absolute classifiers performance is affected. Thus, this is not an issue because the relative performance comparison of classifiers, we are interested in, is not influenced. Nevertheless, if imbalanced data affects classifiers differently, then any means of balancing data would hide the robustness of given classifier. The second reason for not balancing our data is that the severity of imbalance might be rather low considering the size of our data set. We have 212,252 observations in our data sets. Therefore, even with the default rate of 15.91%, there are enough defaulted loans which can be used for classifiers training. This might, however, be an issue in Brown & Mues (2012) because their two biggest data sets have only 7,190 and 2,974 records. If Brown & Mues (2012) use then highly imbalanced data sets with 2.5% or even 1% of the minority class, it is debatable whether the classifiers performance deteriorates because of highly imbalanced data or because of a low number of defaulted loans available for training. Due to the above-stated reason, we believe that there is no need for balancing our data set.

Table 5.6: Iteration sequence of testing

Iteration	**Subsets**		
	Training	**Fine-tuning**	**Testing**
1.	SS 1, SS 2, SS 3	SS 4	SS 5
2.	SS 5, SS 1, SS 2	SS 3	SS 4
3.	SS 4, SS 5, SS 1	SS 2	SS 3
4.	SS 3, SS 4, SS 5	SS 1	SS 2
5.	SS 2, SS 3, SS 4	SS 5	SS 1

Legend: *Five iterations of our 5-fold cross-validation method, and application of our subsets in different steps of cross-validation.*

Chapter 6

Methodological Background

Our methodological background is divided into two parts. The first part, called Classification Techniques, theoretically describes all ten individual classifiers which we use in our master thesis. The second part of our methodological background is named Performance Measurement. We introduce and describe here the performance measurement techniques we use to measure the performance of our classifiers.

6.1 Classification Techniques

The classification techniques we use in our master thesis might be divided into the three groups based on the type of algorithm they use. The classifiers use a linear, non-linear or rule-based algorithm. Logistic regression (LR) and Linear discriminant analysis (LDA) are classification techniques based on linear algorithms. The two linearly based classifiers are presented first. The classifier using non-linear algorithm are described next. These are Support Vector Machine (SVM), Artificial neural network (ANN), k-nearest neighbor (k-NN), Naïve Bayes (NB) and Bayesian network (B-Net). The last group of rule-based classifiers contains Classification and regression tree (CART) and Random forest (RF). A comprehensive description of our classifiers is beyond the scope of our master thesis. Our classifiers description is briefly highlighting a key classifier's algorithm idea with accompanied equations. An interested reader is provided with a link to literature explaining the classification techniques more in depth.

Before describing given classifiers, we introduce a formal notion which is used throughout this chapter. $x = (x_1, x_2, \ldots, x_j) \in R^j$ be a j-dimensional vector containing borrower's and loan's characteristics. Let $p(y = 1|x_i)$ be a probability that a borrower i will default on his or her loan given the input vector x_i with information about the borrower and the loan. Let $y \in \{0; 1\}$ be a binary outcome of a predicted loan status. The value $y = 0$ means that borrower is predicted to pay off his or her loan and $y = 1$ signifies that borrower will default on paying back the loan.

6.1.1 Logistic Regression

Logistic regression is the most widely used classification technique for credit scoring. This algorithm is even considered to be an industry standard for classification (Ala'raj & Abbod, 2015). Among the main advantages of Logistic regression belong an easy implementation, relatively high predictive power and clear interpretation of input variables value for prediction. The following paragraphs and formulas describing Logistic regression are mainly based on Wendler & Gröttrup (2016) and Kuhn & Johnson (2013).

Logistic regression does not directly predict if borrower will or will not pay back his or her loan. Instead, it estimates the probability $p(y = 1|x_i)$ that borrower i will default given the information x_i. The probability $p(y = 1|x_i)$ of Logistic regression is estimated in two steps. The first 'regression' step estimates a linear regression function $g(x_i)$ based on the input variables from vector x_i:

$$g(x_i) = \beta_0 + \beta_1 \cdot x_{i1} + \ldots + \beta_j \cdot x_{ij}. \tag{6.1}$$

In the second 'logistic' step, results of regression function $g(x_i)$ must be transformed to be bounded between 0 and 1 which represents then the probability $p(y = 1|x_i)$. The transformation of regression results is done by logistic function. The logistic function can be mathematically expressed as follows:

$$f(t) = \frac{e^t}{1 + e^t}. \tag{6.2}$$

Combining regression function (6.1) with logistic function (6.2), we get formula for Logistic regression probability $p(y = 1|x_i)$ prediction:

$$p(y = 1|x_i) = \frac{e^{g(x_i)}}{1 + e^{g(x_i)}}. \tag{6.3}$$

The formula 6.3 converges to 1 for high positive values of regression function $g(x_i)$. On the other hand, probability $p(y = 1|x_i)$ goes to 0 for negative values of regression function $g(x_i)$.

6.1.2 Linear Discriminant Analysis

There are two competing approach for explanation of Linear discriminant analysis (LDA). These approaches are from Welch (1939) and Fisher (1936). We have decided for Welch (1939)'s approach described in Kuhn & Johnson (2013) because it is more straightforward. According to this approach, Linear discriminant analyses (LDA) minimizes the total probability of misclassification based on class probabilities and distribution of input variables. We explain Linear discriminant analysis (LDA) in two steps. In the first step, we explain the general idea behind LDA on an example with single input variable. Linear discriminant function is introduced in the second step as a solution for more generalized cases with multiple inputs and multiple classes.

Considering binary classification with single input variable, the total probability of misclassification would be minimized when x is classified into class 1 if $p(y = 1|x) > p(y = 0|x)$.[1] Using Bayes' theorem and previous inequality, x is classified into class 1 if:

$$p(y = 1) \cdot p(x|y = 1) > p(y = 0) \cdot p(x|y = 0). \tag{6.4}$$

To solve more general cases with multiple variable inputs or classes, we introduce a linear discriminant function. Before defining linear discriminant function, we must accept an assumption that multivariate distribution of input

[1]The x can represent in our case single variable vector including some information about borrower. The $p(y = 1|x)$ is posterior probability calculated with the help of Bayes' theorem. For Bayes' theorem, we need prior probability $p(y = 1)$, known in our case as default rate, and conditional probability $p(x|y = 1)$.

variables is normal. Such multivariate distribution has then two parameters. The first parameter is a multidimensional mean vector μ_C. It is assumed that mean vector μ_C is different for each class. The second parameter of multivariate distribution is covariance matrix Σ_C. Here, we assume that covariance matrices are identical for different classes. If the above stated conditions are satisfied, we can mathematically express linear discriminant function for class C as follows:

$$X'\Sigma^{-1}\mu_C - 0.5\mu'_C\Sigma^{-1}\mu_C + log(p(Y = C)). \tag{6.5}$$

The equation 6.5 is linear in its input variables and defines separating boundaries for given classes.

For more information about Linear discriminant analysis or Fisher (1936)' approach please refer to Kuhn & Johnson (2013).

6.1.3 Support Vector Machine

Support vector machine (SVM) is very versatile and effective algorithm. It can be used for classification, regression as well as novelty detection. Although being categorized as a non linear classification technique, Support vector machine (SVM) can be considered as a connection between linear and non-linear classifiers (Wendler, 2016). As described in Karatzoglou *et al.* (2006), SVM uses a simple linear method to classify data in a high-dimensional feature space which is derived by non linear methods from the original input space. In other words, input data are transformed into the high-dimensional feature space in which are the data linearly separable.

The transformation of the input data into the high-dimensional feature space is done by a kernel function k. There are many different kernel functions as discussed below. Karatzoglou *et al.* (2006) generally defines kernel function k as:

$$k(x, x') = \langle \Phi(x), \Phi(x') \rangle, \tag{6.6}$$

where $\Phi : X \to H$ is a projection from feature space into high-dimensional

feature space.

The data are then separated in the high-dimensional feature space by a hyper-plane. The optimal hyper-plane has the maximal separation margins between the two classified classes. The maximal separation is achieved by solving quadratic optimization problem with constrains. The classification decision function with hyper-plane as parameter can be mathematically expressed as follows:

$$f(x) = sign(\langle w, \Phi(x) \rangle + b), \tag{6.7}$$

where w is a solution of quadratic optimization and b is a constant.

There are eight different kernel function described in Karatzoglou *et al.* (2006). We use linear (SVM-L) and Gaussian radial basis (SVM-Rbf) kernel function in our master thesis. We have chosen these kernel functions because they have been both compared in Chang *et al.* (2015)'s study based on the Lending Club data. Furthermore, Tsai *et al.* (2014) used SVM-L in their study too.

The Support Vector Machine with linear kernel function (SVM-L) is the simplest kernel functions with following expression:

$$k(x, x') = \langle x, x' \rangle. \tag{6.8}$$

Gaussian radial basis kernel function includes parameter gamma which is achieved by fine-tuning the SVM-Rbf model. The gamma parameter determines the shape of hyperplane. An increase in parameter gamma usually means an increased number of support vectors. The support vectors are the closest data points which uniquely define the decision boundary. The Gaussian radial basis kernel function can be expressed as:

$$k(x, x') = exp(-\gamma \|x - x'\|^2). \tag{6.9}$$

For more information regarding Support vector machines and additional

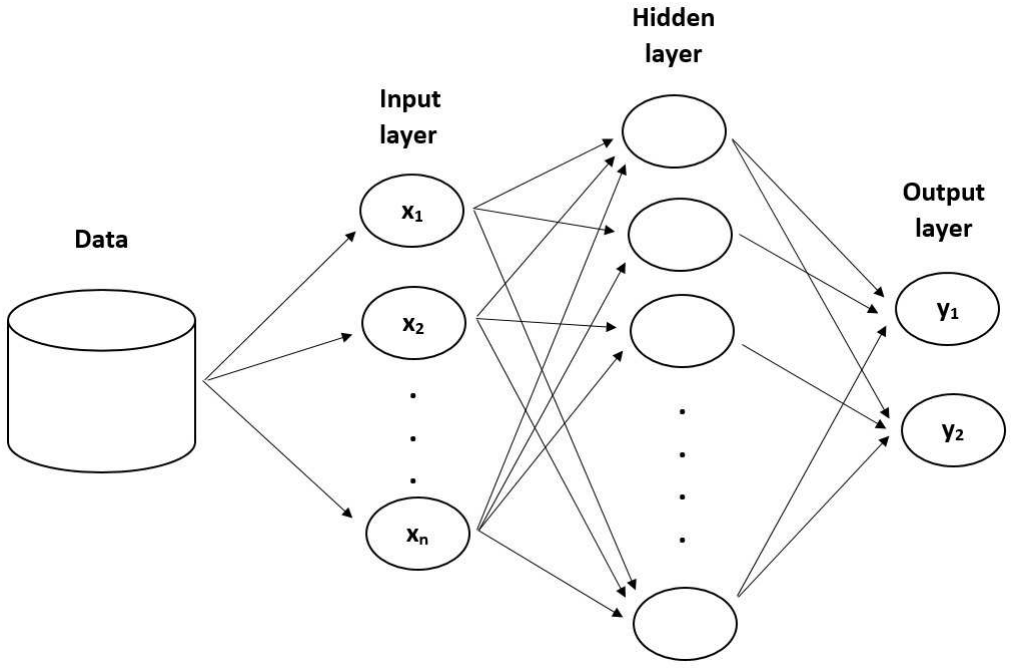
kernel functions, refer please to Karatzoglou *et al.* (2006) and Scholkopf *et al.* (2000).

6.1.4 Artificial Neural Network

Similarly to the Support vector machines (SVM), an Artificial neural network (ANN) is a black box algorithm. The ANN's algorithm is hardly comprehensible and interpretable because of its neuron mechanism with hidden layers. Despite its black box nature, Artificial neural network is very popular and powerful algorithm which might be applied to a variety of complex problems (Wendler, 2016).

Wendler & Gröttrup (2016) says that the human brain served as an inspiration for origin of ANN's algorithm. Artificial neural network contains multiple different neurons that functioning corresponds to basic brain processes. To better explain ANN's mechanism, we have included figure 6.1. There are three different types of layers in ANN. The first layer is called an *Input layer*. Each input variable from data set is represented by one neuron in the *Input layer*.

Figure 6.1: Artificial neural network's mechanism



Source: Authors' own production inspired by Wendler (2016).

These neurons are then transformed with an activation function and passed to neurons in a new layer. This new layer is called *Hidden layer*. There might be one or multiple *Hidden layer(s)*. The processed neurons from *Hidden layer(s)* are then passed to the last layer. Neurons in the last layer, called *Output layer*, belong to one of the final classes or contain prediction information, such as probability prediction.

There exist different models of Artificial neural network which differ in use of the activation function. We use *Multilayer Perceptron (MLP)* model of ANN with hyperbolic tangent activation function φ in our master thesis. The whole mechanism of ANN can be simply described with following formula:

$$y_c = \varphi(\sum_{i=0}^{n} \omega_i \cdot x_i), \tag{6.10}$$

where $\omega_0, \omega_1, \ldots, \omega_n$ are weights for input neurons $(x_0, x_1, \ldots, x_n)$, φ is the activation function and y_c represents the class of output neurons.

6.1.5 k-Nearest Neighbors

The k-Nearest neighbors (k-NN) is one of the simplest classification methods according to the Wendler & Gröttrup (2016). To classify a new observation from a testing set, the k-NN classifier simply identifies k nearest observations from training sample, hence the name k-Nearest neighbors, and the prediction for new observation is made based on the mean class of k nearest neighbors from training set. Kuhn & Johnson (2013) add that the class prediction can be based on median class of k nearest neighbors instead of mean class. In our master thesis, we use the mean class prediction which is more common.

The accuracy of k-NN's classification highly depends on the size of neighborhood, i.e. the value of k. Small k value will include only several observations which have then high impact on classification. It means that the final classification might be affected by outliers or some noise in data. On the other hand, the classification with large value of k is more robust and less prone to outliers or noise. Nevertheless, in case of imbalanced data set, the majority class might easily suppress the minority class in favour of majority class.

There are several methods for measuring the distance between two observations. Kuhn & Johnson (2013) describe following two methods: Euclidian and Minkowski distance. The Euclidian distance can be expressed with following formula:

$$\left(\sum_{i=1}^{N}(x_{ai} - x_{bi})^2\right)^{1/2}, \tag{6.11}$$

where x_a and x_b are individual observations. The Minkowski distance has generalized form and is defined as follows:

$$\left(\sum_{i=1}^{N}|x_{ai} - x_{bi}|^p\right)^{1/p}, \tag{6.12}$$

where $p > 0$. For $p = 1$ is the Minkowski distance same as Manhattan distance which is frequently used in k-NN algorithm. Furthermore, we can see that for $p = 2$, the Minkowski distance is equivalent to Euclidian distance. We use Euclidian distance in our master thesis. Lastly, it should be added that the data must be normalized before distance measurement. The normalization is done to put all variables on the same measurement scale.

For more information about k-Nearest neighbors refer please to Wendler & Gröttrup (2016) and Kuhn & Johnson (2013).

6.1.6 Naïve Bayes and Bayesian Network

The last two non-linear classifiers in our classifiers' description are Naïve Bayes and Bayesian Network. These two classifiers are very similar because they both use a Bayes rule and differ only in the strength of made assumptions. Therefore, we have decided to describe both classifiers in one subsection.

The building block of Naïve Bayes and Bayesian Network is the Bayes rule. The Bayes rule, also known as Bayes theorem, has following definition according to the Kuhn & Johnson (2013):

$$p(y = c_l|x) = \frac{p(y = c_l) \cdot p(x|y = c_l)}{p(x)}, \tag{6.13}$$

where $p(y = c_l \mid x)$ is probability that y belongs to the class c_l given variable x. We have already used Bayes rule for explanation of Logistic regression and Linear discriminant analysis in the previous subsections. Specifically, we have used there following condition probability: $p(y = 1|x)$. It estimates the probability that borrowers will default given the input vector x containing borrower' and loan's characteristics. The probability $p(y = c_l|x)$ is called posterior probability. On the other hand, $p(y = c_l$ is the prior probability and signifies unconditional probability of class c_l being outcome y. The $p(x)$ estimates the frequency of input vector x in the sample. The last expression $p(x|y = c_l)$ is conditional probability showing probability of observing input vector x given the outcome class c_l.

As being described above, the Naïve Bayes and Bayesian Network differ only in the strength of their assumptions. The Naïve Bayes makes stronger assumptions than Bayesian Network classifier. For instance, the Naïve Bayes assumes that all input variables in vectors x are independent of each other. Kuhn & Johnson (2013) says that this assumption is extremely strong and difficult to claim. However, the independence assumption simplifies the complexity of computation. For example, under independence assumption the conditional probability can be computed with the following formula:

$$p(x|y) = \prod_{j=1}^{q} p(x_j|y), \tag{6.14}$$

where $p(x_j|y)$ is the conditional probability of individual input variables.

The Bayesian Network relaxes the independence assumption made in Naïve Bayes. On the contrary, the Bayesian Network assumes correlation between independent variables in input vector x. For instance, we use in our master thesis Bayesian Network with tree augmented network (TAN) which allows dependencies between individual input variables. In other words, the impact of independent variable x_i on output y depends as well on an independent variable x_j.

6.1.7 Classification and Regression Tree

Classification and regression tree (CART) belongs to a rule-based classifiers class. The rule-based classifiers, including Random forest (RF), have a different approach to classification than classifiers based on linear or non-linear algorithms. Wendler & Gröttrup (2016) says that rule-based classifiers try to find rules, hence the name, or structures in raw data for determination of final class. The classification technique is then based on the found rules. These rules can usually be represented by decision trees which are easily interpretable and understandable.

There are various versions of decision trees. These versions mainly differ in the method of node splitting. The Classification and regression tree (CART) uses the binary splitting method which means that each non-leaf node splits into two new branches, as described in Wendler & Gröttrup (2016). The node splitting is determined with the split dispersion measure, called *Gini coefficient*.[2] Wendler & Gröttrup (2016) defines the *Gini coefficient* as follows:

$$Gini(\sigma) = 1 - \sum_{l} \left(\frac{N(\sigma, l)}{N(\sigma)} \right)^2, \tag{6.15}$$

where σ signifies the node, l is a class category, $N(\sigma, l)$ is the count of observation with category l in node σ and $N(\sigma)$ is the total count of observations in node σ. There is no need for further splitting if *Gini index* is close to 0. It happens when the ratio of $N(\sigma, l)$ and $N(\sigma)$ is close to 1 which means that the vast majority of observation in the node σ belongs to the same category.

Gini Gain is used to determine whether and how the next split should be made. The *Gini Gain* is defined by Wendler & Gröttrup (2016) with following formula:

$$GiniGain(\sigma, s) = Gini(\sigma) - \frac{N(\sigma_q)}{N(\sigma)} Gini(\sigma_q) - \frac{N(\sigma_r)}{N(\sigma)} Gini(\sigma_r), \tag{6.16}$$

where σ_q and σ_r represent two split nodes from σ and s describes the criteria for

[2] Please do not confuse the *Gini coefficient* used for splitting of decision trees with *Gini index* used for classifiers performance measurement. *Partial Gini index* is used in our master and its description might be found in subsection 6.2.5.

splitting. From numerous binomial splits is chosen that one which maximizes the *GiniGain*. The child nodes should be purer after such split. In other words, one of the classes will be more prevalent in the child nodes.

6.1.8 Random Forest

Random Forest (RF) is the only homogenous ensemble classifier in our master thesis. All other classifiers here are individual classifiers. The homogenous ensemble classifiers combine prediction results of multiple base models. This approach is supposed to increase the predictive performance of such classifier. Lessmann *et al.* (2015) describe homogenous ensemble classification as two stages process in their on-line appendix. A set of base models is created in the first stage. In the second stage, final prediction is made by combination of base model predictions. Lessmann *et al.* (2015) generally describes the ensemble prediction, $E(x, M)$, based on input vector x with following formula:

$$E(x, M) = \frac{1}{S} \sum_{s=1}^{S} \beta_s M_s(x), \tag{6.17}$$

where S is a set of base models $M = (M_1, M_2, \ldots, M_s)$, $M_s(x)$ represents prediction of individual base model M_s and β_s is the weight of this prediction on final prediction.

Cichosz (2015) explains that Random forest uses decision tree as base classifier for its predictions. To create different base models M_s, the independent variables are randomly sampled at each node split. The first meta-parameter of Random forest determines the number of randomly sampled variables. For example, if the value of this meta-parameter is 5, then there will be 5 randomly sampled variables used as candidates for node split. The size of set S is the second meta-parameter of Random forest. This number determines how many base models will be used for final prediction.

6.2 Performance Measurements

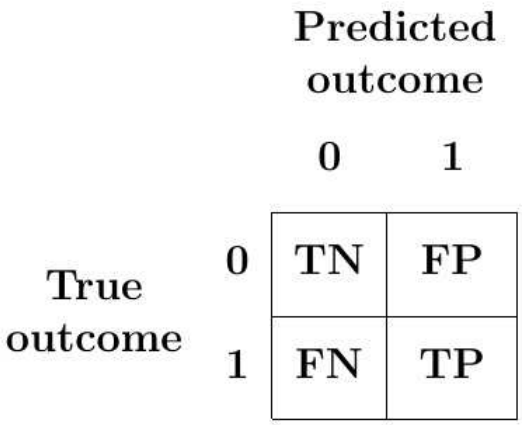
We use six different performance measurements to evaluate classifiers performance. These performance measurements might be divided into three groups. The first group of performance measurements evaluates the correctness of classifiers' categorical predictions, such as Percentage correctly classified (PCC) or Kolmogorov-Smirnov statistics (KS). The second group contains performance measurement which evaluate the accuracy of classifiers' probability predictions, such as Brier score (BS). The performance measurements using discriminatory ability of classifier, such as Area under the curve (AUC), Partial Gini index (PG) and H-measure (H), belongs to the last group. The classifiers' performance results based on several performance measurements from different measurement groups are more robust than results based on one performance measurement or performance measurement from the same group.

Before explaining individual performance measurement techniques used in our master thesis, we firstly introduce a confusion matrix, also known as an error matrix. The introduction of confusion matrix is a prerequisite for explanation of Percentage correctly classified (PCC) and Area under the curve (AUC). Moreover, understanding confusion matrix is beneficial for explanation of other performance measurements, such as Precision and Recall, used in studies from Table 3.2 because they are derived from confusion matrix.

The aim of classification is to predict the binary outcome of target variable. However, as being already explained in the previous section 6.1 Classification Techniques, classifiers do not directly predict a binary outcome of target variable. The classifiers rather predict the probability of borrower's default. Mathematically expressed, let $y \in \{0; 1\}$ be a binary outcome where $y = 0$ signifies that borrower paid back his or her loan and $y = 1$ means that borrower defaulted on paying back the loan. Let $x = (x_1, x_2, \ldots, x_j) \in R^j$ be a j-dimensional vector containing borrower's and loan's characteristics. The classifier output is then $p(y = 1|x_i)$, which is the probability that a borrower i will default on his or her loan. To be able to decide whether a loan will be granted to a borrower i, the borrowers i default probability $p(y = 1|x_i)$ must be compared to a chosen threshold τ. If $p(y = 1|x_i) > \tau$, the borrower i is predicted not to pay back his or her loan. Otherwise, classifier predicts that borrower i repays his or her loan. In other words, having a threshold τ, we can classify

classifier's probability predictions as binary outcome. In our case, *loan_status* is the target binary variable with true outcome values 0 and 1. Similarly to above described binary outcome, 0 is for good and 1 is for bad borrowers. The binary classification prediction might or might not be the same as the true outcome of *loan_status*. This originates 2x2 matrix, called confusion matrix, with four possible states of classification prediction and true outcome. The four possible states are described below. This information is based on Wendler & Gröttrup (2016).

Figure 6.2: Confusion matrix



Source: Authors' own production inspired by Wendler (2016).

- **True positive (TP)** - The true outcome is 1 and classification prediction is 1. The borrower defaulted on his or her loan and the loan default was predicted.
- **True negative (TN)** - The true outcome is 0 and classification prediction is 0. The borrower paid back his or her loan and the loan repayment was predicted.
- **False positive (FP)** - The true outcome is 0 and classification prediction is 1. The borrower paid back his or her loan but the loan default was predicted.
- **False negative (FN)** - The true outcome is 1 and classification prediction is 0. The borrower defaulted on his or her loans but the loan repayment was predicted.

The confusion matrix with four possible classification states is displayed in Figure 6.2.

There are many different performance measures, such as Precision and Recall, derived from confusion matrix. For example, Precision was used as performance measure in studies of Tsai *et al.* (2014) and Chang *et al.* (2015). Precision, also known as Positive predictive value (PVV), measures the percentage of correctly classified cases in positive prediction.[3] It is defined as follows:

$$PVV = \frac{TP}{TP + FP}. \tag{6.18}$$

The next performance measurement, used by Wu (2014), is called Recall. There are several other names, such as Sensitivity or True positive rate (TPR), usually used for Recall. Recall measures the percentage of correctly classified cases in true positive outcome. This can be mathematically expressed as follows:

$$TPR = \frac{TP}{TP + FN}. \tag{6.19}$$

6.2.1 Percentage Correctly Classified

The most important and widely used performance measurement derived from confusion matrix is, however, called Accuracy or Percentage correctly classified (PCC). As the name suggests, the Percentage correctly classified (PCC) measures the percentage of correctly classified cases in confusion matrix. The definition of PCC is following:

$$PCC = \frac{TP + TN}{TP + FP + TN + FN}. \tag{6.20}$$

This performance measurement was used by Wu (2014), Chang *et al.* (2015) and Malekipirbazari & Aksakalli (2015). In defiance of issues regarding the ap-

[3]We mark prediction 1 as positive, which means that a borrower is predicted to default. On the hand, prediction 0, named as negative prediction, means that a borrower is predicted to pay back his or her loan. We use the same notion of positive and negative cases for true outcome too.

propriate threshold value τ, we have decided to include Percentage correctly classified (PCC) in our master thesis as well. Specifically speaking, we have two reasons to include PCC measurement. The first reason is that this measurement has been used in the three above mentioned studies. Using the same performance measurement as they use will help us to compare our results on the same basis. Moreover, Percentage correctly classified was used as performance measurement in our benchmark classifier ranking in Table 3.2.

In the previous paragraph, we mentioned the issue of finding the appropriate threshold value τ. This issue is typical for all performance measurements based on confusion matrix. As being explained at the beginning of this chapter, the threshold τ is chosen to classify the classifiers' probability prediction into binary predicted outcome - either 0 or 1. To compute appropriate τ, the misclassification costs should be known. The misclassification costs represent the trade-off between False negative (FN) and True negative (TN) errors. Obviously, the False negative (FN) error is much worse than the True negative (TN) error because the costs of loan granted to a borrower who defaults are higher than revenue loss incurred by rejection of a good borrower. Hand (2006) argues that misclassification costs significantly differ case by case and might even change in time. Hand (2006) even expressed concern about the possibility of knowing the right misclassification costs. We do, unfortunately, lack the information about our misclassification cost for τ computation. Therefore, we have decided to use the same approach for τ computation as Lessmann *et al.* (2015). So the threshold τ is computed such that the percentage of predicted positive cases is equal to the default rate in our data set. In other words, the fraction of predicted positive cases, which are True positive (TP) and False positive (FP), to all cases in confusion matrix is equal to roughly 15.9% which is the default rate.

6.2.2 Kolmogorov-Smirnov Statistic

The Kolmogorov-Smirnov statistics (KS) is from the same performance measurement group as Percentage correctly classified. Furthermore, the Kolmogorov-Smirnov statistics uses classifiers predicted probability $p(y = 1|x_i)$ too but with a fixed threshold value. Mays (2001) describes Kolmogorov-Smirnov statistics as the maximum difference between the cumulative distribution function of

negative ($F_{n,negative}(q)$) and positive cases ($F_{n,positive}(q)$). This can be mathematically express as follows:

$$KS = \max_{q \in [L,H]} |F_{n,positive}(q) - F_{n,negative}(q)|, \tag{6.21}$$

where L represents the minimum and H the maximum value of probability score prediction.

6.2.3 Brier Score

The Brier score (BS) assess the accuracy of classifiers' probability prediction. The Brier score (BS) can be described as mean squared error of probability prediction and true outcome. Let $p(y = 1|x_i)$ be the probability prediction that borrower i defaults, N be the number of observations and θ_i be the true binary outcome, then the Brier score (BS) can be mathematically described as follows:

$$BS = \frac{1}{N} \sum_{i}^{N} (p(1|x_i) - \theta_i)^2. \tag{6.22}$$

For more information about Brier score please refer to Hernandez-Orallo *et al.* (2011) or Rufibach (2010).

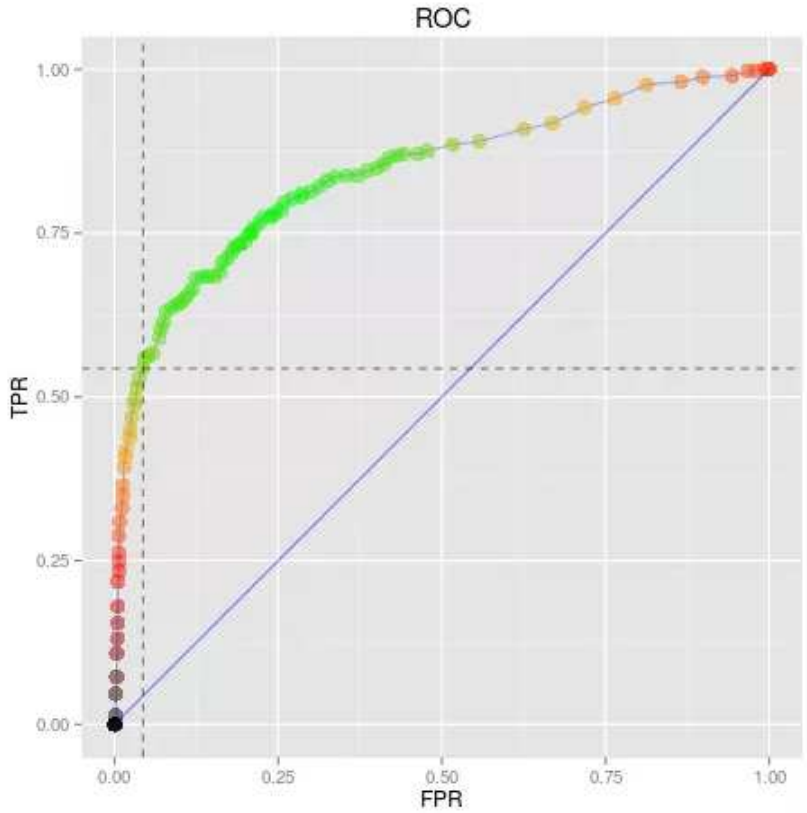
6.2.4 Area Under Curve

The Area Under the Curve (AUC) is well-known and widely used performance measure. The Area Under the Curve (AUC) measures the area under the Receiver Operating Characteristic curve (ROC). The ROC curve is based on two performance measurements derived from the confusion matrix. These measurements are True Positive Rate (TPR) and False Positive Rate (FPR). The True Positive Rate (TPR) has already been described above and defined in equation 6.19. The False Positive Rate (FPR) is defined by following equation:

$$FPR = \frac{FP}{FP + TN}. \tag{6.23}$$

As being previously discussed, the values of four possible states in confusion matrix depends on the choice of threshold value τ. The same holds for TPR

Figure 6.3: Receiver operating characteristics curve



Source: Authors' recreation of code from *joyofdata* Github repository.

and FPR metrics. It means that values of TPR and FPR are functions of threshold value τ - formally written TPR(τ) and FPR(τ). Let p be a point in a 2-dimensional space with y-coordinate being TPR(τ) value and x-coordinate being FPR(τ) value. Then by varying τ, we get collection of points p_τ. The Receiver Operating Characteristics curve is created by connecting all points p_τ for threshold value τ varying from 0 to 1. The ROC curve is depicted in Figure6.3.[4]. As an example, for threshold value of 0.70 we have point $p_{0.70}$ with TPR of 0.55 and FPR of 0.05 lying on the intersection of dotted lines in Figure 6.3. Looking at Figure 6.3, we can see a clear trade-off between TPR and FPR. If we want to increase the TPR value, the value of FPR will increase as well. The closer the ROC curve is to the point [0.00;1.00] in upper left corner, the better the trade-off between TPR and FPR and the higher the area under the ROC curve. For AUC performance measurement holds the higher the area

[4]The Figure 6.3 has been created in *R* by master thesis authors based on the code from *joyofdata* GitHub repository. Link to the *joyofdata* GitHub repository: https://github.com/joyofdata/joyofdata-articles/tree/master/roc-auc

under the ROC curve, the better the classifier. The Area Under the Curve can be mathematically defined as follows:

$$AUC = \int_{-\infty}^{\infty} TPR(\tau)FPR'(\tau)d\tau. \tag{6.24}$$

6.2.5 Partial Gini Index

It should be highlighted at the beginning that there is a difference between classical Gini index and Partial Gini (PG) index. As described in Wendler & Gröttrup (2016), the classical Gini index can be computed by following formula:

$$Gini = 2AUC - 1. \tag{6.25}$$

It is obvious from the equation 6.25 that classical Gini index is only linear transformation of Area Under Curve (AUC). It means that we would get completely same classifiers ranking by using classical Gini index as by using AUC measure. Therefore, we have decided to use a novel approach introduced by Pundir & Seshadri (2012). This novel approach is called Partial Gini index (PG) and is usually used for evaluation of income inequalities. The motivation for using PG is to divide the whole data sample into portions and individually analyze the portions. For example, in evaluation of income inequalities the whole society can be divided into several classes, such as poor class or middle class, which are then individually measured. We believe that in the case of credit scoring the classifier probability prediction $p(y = 1|x_i)$ might be divided into several portions too. For instance, borrower i may be very likely to default (represented by high value of $p(1|x_i)$), somewhat likely or unlikely (average value of $p(1|x_i)$) or very unlikely to default (low value of $p(1|x_i)$). We are mainly interested to measure the Partial Gini index (PG) in the "somewhat likely or unlikely" class with the mean values of $p(1|x_i)$. That's why, we use two cut off values l and u for $p(1|x_i)$ values. The lower cut off value l is defined as mean value of $p(1|x_i)$ minus one standard deviation of $p(1|x_i)$. Similarly, the upper cut off value u is defined as mean value of $p(1|x_i)$ plus one standard deviation of $p(1|x_i)$. These values are calculated for each classifier separately. The Partial Gini index (PG) can be then mathematically described with following equation:

$$PG_{[l,u]} = 1 - \frac{2\int_l^u L(p)dp}{(l+u)(u-l)}, \tag{6.26}$$

where $L(p)$ is Lorenz curve and p is the cumulative distribution function of $p(1|x_i)$. Further information regarding Partial Gini index is described in Pundir & Seshadri (2012).

6.2.6 H-Measure

Even though the AUC, described in 6.2.4, is very popular performance measurement, there is one serious deficiency in this measurement according to Hand (2009). Specifically, he argues that the fundamental incoherence of misclassification costs usage is the main deficiency of AUC. It means that different misclassification cost distributions are applied for different classifiers. Hand (2009) states that this fact causes the fundamental incoherence because the relative severity of misclassification costs depends on the choice of classifier. Hand (2009) proposes performance measurement, called H-measure, as a remedy for AUC's imperfection.

The main advantage of H-measure is that it uses a weight function which is independent of classifier probability score distribution. This weight function used in H-measure is a Beta distribution. Hand (2009) says that using the Beta distribution in H-measure makes the classifier comparison fair. For computation of H-measure, we use the *R*'s *hmeasure* package described in depth in Anagnostopoulos *et al.* (2012).

Chapter 7

Empirical Results

This chapter is divided into four parts. In the first part, called Overview and Benchmark Comparison, we describe the results of our classifiers' performance comparison. Next, we compare these results to Lessmann *et al.* (2015)'s classifiers benchmark ranking from Table 3.1. The second part of this chapter compares our results with other studies based on Lending Club data from Table 3.2. Hypotheses about classifiers' performance from Chapter 4 are tested in the third part of this chapter. The last part is devoted to the discussion of our findings and recommendation of further research areas.

7.1 Overview and Benchmark Comparison

We compare ten classifiers on six different performance measurements with 5-fold cross-validation method in our master thesis. The 5-fold cross-validation method consists of five iterations. The performance results from these iterations can be found in Appendix D. The aggregated results from single iterations are displayed in Table 7.1. The Table 7.1 shows the average overall performance of our classifiers across the six different performance measurements. The best classifier according to the given performance technique is determined based on the total performance. The best classifiers in Table 7.1 are underscored and in bold face. We further calculate the standard deviation of classifier's overall performance as based on the results from iterations. The last metric included in Table 7.1 in column M-W is a Mann-Whitney U statistic. The values of Mann-Whitney U statistic is accompanied by stars showing if the classifier's performance is significantly different from the best classifier.

Table 7.1: Average performance results

Classifier	Performance measurement																	
	PCC			KS			BS			AUC			PG			H		
	Perf.	St. Dev.	M-W	Perf.	St. Dev.	M-W	Perf.	St. Dev.	M-W	Perf.	St. Dev.	M-W	Perf.	St. Dev.	M-W	Perf.	St. Dev.	M-W
LR	**0.7913**	0.0016	/	**0.2885**	0.0056	/	**0.1239**	0.0008	/	**0.6979**	0.0028	/	0.2502	0.0046	24**	**0.1319**	0.0039	/
ANN	0.7905	0.0011	17	0.2848	0.0039	17	0.1240	0.0009	11	0.6975	0.0019	14	0.2453	0.0081	25***	0.1305	0.0031	15
LDA	0.7904	0.0021	16	0.2833	0.0048	20	0.1245	0.0008	8	0.6955	0.0028	19	0.2586	0.0075	17	0.1285	0.0039	19
L-SVM	0.7887	0.0034	18	0.2854	0.0053	17	0.1585	0.0003	0***	0.6967	0.0029	18	0.2309	0.0068	25***	0.1300	0.0040	17
RF	0.7883	0.0014	23**	0.2789	0.0074	20	0.1248	0.0010	6	0.6928	0.0032	23**	0.2236	0.0062	25***	0.1243	0.0040	23**
B-Net	0.7878	0.0018	23**	0.2555	0.0032	25***	0.1257	0.0007	1**	0.6787	0.0027	25***	0.2108	0.0054	25***	0.1122	0.0039	25***
SVM-Rbf	0.7818	0.0019	25***	0.2104	0.0044	25***	0.1306	0.0008	0***	0.6519	0.0034	25***	**0.2641**	0.0070	/	0.1089	0.0044	25***
NB	0.7836	0.0072	24**	0.2425	0.0066	25***	0.1502	0.0076	0***	0.6689	0.0031	25***	0.2043	0.0226	25***	0.1029	0.0028	25***
CART	0.7659	0.0144	25***	0.2557	0.0090	25***	0.2019	0.0073	0***	0.6373	0.0048	25***	0.1495	0.0249	25***	0.0801	0.0053	25***
k-NN	0.7655	0.0170	25***	0.2022	0.0026	25***	0.1322	0.0006	0***	0.6360	0.0015	25***	0.1502	0.0082	25***	0.0678	0.0015	25***

Source: *Authors' computation based on the five partial iteration results from Appendix D.*

Legend: *The abbreviation M-W stands for Mann-Whitney U test. The Mann-Whitney U statistics in column M-W are accompanied with stars signifying whether the classifier's performance is significantly different from the performance of the best classifier. *** denote significance at 1% level, ** at 5% level and * at 10% level.*

Three stars (***) shows that the classifier's performance is significantly different from the best classifier at 1% significance level. Two stars (**) denote significance at 5% level and one star (*) at 10% level. For example, Logistic regression (LR) is the best classifier based on the Percentage correctly classifier (PCC) measure according to our results. The performance of Logistic regression (LR) measured by PCC is, however, not significantly different from the performance of Linear discriminant analysis (LDA) and two other classifiers at 5% significance level. On the hand, PCC performance of LR is significantly different from Random forest (RF) at 5% significance level, and from k-Nearest neighbors (k-NN) at even 1% significance level. If not stated otherwise, we always refer to the 5% significance level when speaking about significant differences between classifiers' performance.

Table 7.2: Classifiers' ranking

Classifier	**Performance Measurement**						**Avg. Score**	**Total Ranking**
	PCC	**KS**	**BS**	**AUC**	**PG**	**H**		
LR	1	1	1	1	3	1	**1.3**	**1**
ANN	2	3	2	2	4	2	**2.5**	**2**
LDA	3	4	3	4	2	4	**3.3**	**3**
L-SVM	4	2	9	3	5	3	**4.3**	**4**
RF	5	5	4	5	6	5	**5.0**	**5**
B-Net	6	7	5	6	7	6	**6.2**	**6**
SVM-Rbf	8	9	6	8	1	7	**6.5**	**7**
NB	7	8	8	7	8	8	**7.7**	**8**
CART	9	6	10	9	10	9	**8.8**	**9**
k-NN	10	10	7	10	9	10	**9.3**	**10**

Source: *Authors' ranking based on the average performance results from Table 7.1.*

Avg. Score: *Average score computes the average ranking of classifier based on rankings achieved under different performance measurements.*

Total Ranking: *Total ranking ranks classifiers based on their average score.*

To be able to compare and rank our classifiers across all performance measurements, we have created a Table 7.2. The Table 7.2 displays ranking of classifiers based on given measurement techniques as well as total ranking. The classifiers are ranked based on their performance. The best performing classifier gets ranking 1, the second best performing classifier gets ranking 2 and so on. Looking at Table 7.2, we see that Logistic regression (LR) has ranking 1 based on Percentage correctly classified (PCC) because it has the highest PCC performance in Table 7.1. The rankings across different performance measurement are averaged in column *Avg. Score* in Table 7.2. The total ranking of classifiers, displayed in column *Total Ranking*, is derived from the values of column *Avg. Score*.[1]

Our Table 7.2 is similar in structure to the Lessmann *et al.* (2015)'s Table 3.2 from Chapter 3. We regard the Lessmann *et al.* (2015)'s ranking as our baseline ranking. Comparing our baseline ranking from Table 3.2 and our ranking from Table 7.2, we identify several differences between these rankings. The biggest difference is in the ranking of the best classifier. In Lessmann *et al.* (2015)'s ranking is the best classifier Random forest (RF), and Logistic regression (LR) is the third best classifier. According to our results is Logistic regression (LR) the best classifier, and Random Forest (RF) is only fifth best classifier based on the Lending Club data. We see two possible reasons as the explanation of these ranking differences.

The first possible explanation is fine-tuning of model's meta-parameters. The model's meta-parameters, we have used, are described in Appendix C. We demonstrate this potential explanation based on the lower performance of Random forest (RF) in comparison with Logistic regression (LR). Similar to Lessmann *et al.* (2015), we have not fine-tuned Logistic regression (LR) and we have used two meta-parameters for fine-tuning of Random forest. These meta-parameters are a number of grown trees and a number of randomly sampled variables for node splitting. As shown in Appendix C, we have used 5 randomly sampled variables for node splitting and 800 grown trees because it give us the best performance of Random forest (RF) across individual iterations.

[1]There are several approaches to make classifiers' ranking. Following a similar approach used in Lessmann *et al.* (2015), we could rank our classifiers based on the performance in each iteration rather than ranking them based on the average performance across five iterations as described above. We tried both approaches and both do lead to the same ranking shown in the Table 7.2.

Malekipirbazari & Aksakalli (2015) fine-tuned Random forest (RF) with five randomly sampled variables in their study based on Lending Club data too. We were interested to know if we can replicate the Malekipirbazari & Aksakalli (2015)'s results based on our Lending Club data set. Moreover, we wanted to check if Random forest (RF)'s performance achieved by our meta-parameters in R is comparable to the performance of Malekipirbazari & Aksakalli (2015)'s approach. Therefore, we have used the same analytical software, called WEKA, as Malekipirbazari & Aksakalli (2015) used. Furthermore, we have used the same meta-parameters, also 5 randomly sampled variables and 80 grown trees.[2] With this set-up, we have achieved AUC of 0.684 with Random forest (RF) based on our data.[3] However, Malekipirbazari & Aksakalli (2015) achieved AUC of 0.71 based on their Lending Club data. When searching for the cause of this difference, we found out that the default rate in Malekipirbazari & Aksakalli (2015)'s data set is 20.5% which is significantly higher than default rate in our data set (15.9%). Brown & Mues (2012) argues that classifiers' performance deteriorates with higher data imbalance. That's why we think that performance of Random forest (RF) might deteriorate more than the performance of Logistic regression (LR) based on Lending Club data. After all, we do not believe that the worse performance of Random forest (RF) in our master thesis is caused by inappropriate fine-tuning of meta-parameters, but we cannot deny this possibility.

The second possible reason for classifier's ranking difference between our master thesis and Lessmann *et al.* (2015) is the use of data sets. In line with Salzberg (1997) and Baesens *et al.* (2003), we believe that classifiers' prediction performance highly depends on the data sets used for model training. Lessmann *et al.* (2015) used eight different data sets for their analysis. These data sets together with our Lending Club data set are displayed in Table 7.3. Looking at Table 7.3, we can clearly see that none of the Lessmann *et al.* (2015)'s data sets resembles our data sets. Therefore, we suppose that the classifiers ranking differences are caused by use of different data sets.

[2]Malekipirbazari & Aksakalli (2015) argue that growing more than 80 does not yield considerable performance increase but significantly increases the computation run time.

[3]We have achieved average Random forest (RF)'s AUC of 0.6928 by using our meta-parameters and randomForest's R package. It means that our set-up is better than Malekipirbazari & Aksakalli (2015)'s set-up applied to our data set.

Table 7.3: Comparison of data sets

Lessmann et al. (2015)'s data sets

Name	# of observations	# of variables	Default rate	Source
AC	690	15	0.445	Lichman (2013)
GC	1,000	21	0.300	Lichman (2013)
Th02	1,225	18	0.264	Thomas et al. (2002)
Bene 1	3,123	28	0.667	Baesens et al. (2003)
Bene 2	7,190	29	0.300	Baesens et al. (2003)
UK	30,000	15	0.040	Baesens et al. (2003)
PAK	50,000	38	0.261	http://sede.neurotech.com.br/PAKDD2010/
GMC	150,000	13	0.067	http://www.kaggle.com/c/GiveMeSomeCredit

Our Lending Club data set

Name	# of observations	# of variables	Default rate	Source
LC	212,252	23	0.159	https://www.lendingclub.com/ info/download-data.action

Source: *Authors' depiction of Lessmann et al. (2015)'s and own data set.*

Despite the significant ranking differences of Random forest (RF) and Support vector machine with radial basis kernel function (SVM-Rbf) in our master thesis and Lessmann *et al.* (2015)'s study, we think that our common results bear more resemblance than difference. For example, Logistic regression (LR), Linear discriminant analysis (LDA) and Artificial neural network (ANN) are in both studies placed within the fourth first places. Moreover, three other classifiers - Naïve Bayes (NB), Classification and regression tree (CART) and k-Nearest neighbors (k-NN) are in both studies placed in the last three places. We believe that this is a clear sign of resemblance of our and Lessmann *et al.* (2015)'s findings.

7.2 Comparison with Other LC-based Studies

We have already discussed in Section 3.2 and shown in Table 3.2 four studies comparing classifiers based on Lending Club data. Having our final classifiers ranking, we would like to extend the Table 3.2 with our results. Therefore, we have created Table 7.4 and added there our results. It is evident that we have used by far the largest data sets with 212,252 records in comparison with remaining studies. Furthermore, our data set has the most variables too.

It contains 23 variables including the dependent variable of loan status. We have included ten classifiers to make our comparison comprehensive. As being already discussed, the primary goal of remaining studies is not classifiers comparison, but rather the introduction of new classifiers for default prediction. That's why these studies are not a comprehensive comparison as ours and compare at most five classifiers. The last thing that differentiates our master thesis from the remaining studies is the number of performance measurement used. Altogether, we have used six different measurement techniques from three different performance measurement's groups.[4] Using a broad range of evaluation techniques makes our results robust.

Comparing our final classifiers ranking to four studies from Table 7.4, we observe that Wu (2014) and Tsai *et al.* (2014) rank Logistic regression (LR) as the best classifier in their studies.[5] On the other hand, Chang *et al.* (2015) rank Logistic regression (LR) as the third and Malekipirbazari & Aksakalli (2015) as the fourth classifier in their studies. Generally speaking, we believe that comparison of these studies with our findings is unequal because authors pursue different goals. For example, the main goal of Chang *et al.* (2015) is to compare different distributions of Naïve Bayes (NB) and Support vector machines (SVM). That's why, the data preprocessing and other steps are done to suit these classifiers. For instance, Chang *et al.* (2015) say that the data set has been rebalanced because Support vector machines (SVM) underperform with imbalanced data sets. With this set-up, it might not be surprising that Naïve Bayes (NB) is ranked as the best classifies, and Linear support vector machine (L-SVM) as the second best classifier in Chang *et al.* (2015)'s study. Moreover, these classifiers have been specifically fine-tuned to fit the data. Studies of Tsai *et al.* (2014) and Malekipirbazari & Aksakalli (2015) are conducted in similar manner to Chang *et al.* (2015). We, therefore, refrain from comparing our results with these studies because of unequal conditions for classifier comparison.

[4]For more information about performance measurement's groups, please, refer to Section 6.2 Performance Measurements.

[5]It should be mentioned that Tsai *et al.* (2014) used a modified version of Logistic regression (LR). For more information about this modification, please, refer to Tsai *et al.* (2014).

Table 7.4: Final classifiers' comparison based on the LC data

Classification studies based on Lending Club data	**Data**			**Classifiers**											**Performance measurement technique**
	Year	**# of observations**	**# of variables**	**LR**	**ANN**	**LDA**	**L-SVM**	**RF**	**B-Net**	**SVM-Rbf**	**NB**	**CART**	**k-NN**	**SVM-P**	
Wu (2014)	2007-2011	33 571	22	1				2							PCC, AUC
Tsai et al. (2014)	2007-2013	91 520	n/a	1				3		2	4				PVV
Chang et al. (2015)	2007-2015	n/a	n/a	3			2			5	1			4	PCC, G-mean
Malekipirbazari & Aksakalli (2015)	2012-2014	68 000	16	4				1					3	2	PCC, AUC, RMSE
This Study	2009-2013	212 252	23	1	2	3	4	5	6	7	8	9	10		PCC, KS, BS AUC, PG, H

Source: *Authors' information extraction and ranking computation based on Wu(2014), Tsai et al. (2014), Chang et al. (2015), Malekipirbazari & Aksakalli (2015)'s, and own research.*

7.3 Hypotheses Testing

In this section, we refer to the stated hypotheses from Chapter 4. For each hypothesis, we provide a result below. For example, result 1, denoted as **R1**, is our outcome for hypothesis 1, labeled as **H1** in Chapter 4. The results of our hypotheses are based on summary Tables 7.1 and 7.2.

R1: *Random forest (RF) is not the best classifier among our classifiers of interest based on the Lending Club data.*

We reject our first hypothesis (H1) because Logistic regression (LR) is better classifier than Random forest (RF) based on our Lending Club data. Looking at Table 7.1, we can see that Logistic regression (LR) significantly dominates Random forest (RF) based on four performance measurements (PCC, AUC, PG and H). In the remaining two measurement techniques (KS and BS) has Logistic regression (LR) higher overall performance but this performance is not significantly better than the performance of Random forest (RF). Therefore, Random forest (RF) cannot be the best classifier.

R2: *Artificial neural network (ANN) does not outperform Logistic regression (LR) based on the Lending club data.*

Our second hypothesis (H2) is rejected because we have not found any evidence that Artificial neural network (ANN) outperforms Logistic regression (LR). On the contrary, Logistic regression (LR) outperforms in absolute numbers Artificial neural network (ANN) based on all performance measurements. As can be seen in Table 7.1, the performance differences between LR and ANN are, however, not significant.

R3: *Linearly based classifiers rank in the first half of classifiers' ranking based on Lending Club data.*

We cannot reject our third hypotheses (H3) because both of our linear classifiers ranked within the first half of ranking. More specifically, Logistic regression (LR) ranked as the best classifier and Linear discriminant analysis (LDA) as the third best classifier. This can be seen in Table 7.2. In line with Baesens *et al.* (2003), we believe that this finding is a clear sign that linearly

based classifiers are well suited for predictions using credit scoring data.

R4a: *Logistic regression (LR) outperforms Support vector machine with linear kernel function (L-SVM) based on the Lending Club data.*

R4b: *Logistic regression (LR) outperforms Support vector machine with radial basis kernel function (SVM-Rbf) based on the Lending Club data.*

We reject neither hypothesis 4a (H4a) nor 4b (H4b). Considering H4a and looking at Table 7.1, we can see that Logistic regression (LR) dominates in absolute performance numbers Linear support vector machine (L-SVM). In five out of six performance measurements are these superior performances, however, not significant. Nevertheless, LR's performance is significantly better than L-SVM's performance based on Partial Gini index (PG) measure. The p-value of this difference is 0.008 according to the Mann-Whitney U test. Therefore, we cannot reject H4a.

Support vector machine with radial basis kernel function (SVM-Rbf) is the best classifier measured by Partial Gini index (PG). Moreover, SVM-Rbf is significantly better than LR based on PG measurement. Logistic regression (LR), however, significantly outperforms SVM-Rbf based on all five remaining measurements. This fact does not allow us to reject H4b.

R5: *Linearly based classification methods do not under-perform when measured by Partial Gini index (PG) in comparison with other performance measurements.*

To test our last hypotheses (H5), we have added up the rankings of our linearly based classifiers for different performance measurements from Table 7.2. We have got following numbers by adding up rankings of Logistic regression (LR) and Linear discriminant analysis (LDA): 4 for Percentage correctly classifier (PCC), 5 for Kolmogorov-Smirnov statistics (KS), 4 for Brier score (BS), 5 for Area under the curve (AUC), 5 for Partial Gini index (PG) and 5 for H-measure (H). The average added ranking based on all performance measurements, but PG is 4.6. We do not believe that added ranking 5 of PG is significantly different from the average value of 4.6. Moreover, there are three other performance measurements (KS, AUC, and H) that have the same value

of added ranking as our Partial Gini index (PG). This is for us a sufficient evidence to reject the hypothesis (H5) that linearly based classifiers underperform when measured by Partial Gini index (PG).

7.4 Key Findings

Several interesting findings are introduced in the previous sections. These findings are based on Tables 7.1 and 7.2. We would like to, however, highlight only three of them. According to us, following three findings, labeled with **F**, are the key findings of our master thesis. Moreover, they should serve as take-way messages for our readers.

F1: ***Logistic regression and Linear discriminant analysis are proper classification algorithms for credit scoring.***

Baesens *et al.* (2003), Lessmann *et al.* (2015) and our results provide sufficient evidence that the above-mentioned classifiers perform very well on various credit scoring data sets. Artificial neural network (ANN) performs very well too. Nevertheless, as stated in Table 7.5, ANN is a black box algorithm. It means that the model's decision-making process is unknown. That's why it is almost impossible to determine the relevance and importance of given independent variables. This is particularly problematic for credit scoring as we, for example, cannot explain which variables are the best determinants for borrower's default. On the other hand, it is easy to interpret Logistic regression and Linear discriminant analysis models. Furthermore, as stated in Table 7.5, there is no need for fine-tuning of linearly based classifiers, which makes them easy to use. Based on the arguments mentioned above, it is not surprising that Logistic regression is considered industry standard for credit scoring and is nowadays used by many banks.

F2: ***Support vector machines with linear and radial basis kernel functions are unsuitable classification algorithms for credit scoring.***

We do not recommend Support vector machines with linear and radial basis kernel distributions because their medium performance is not justifiable

in the context of their shortcomings. We identify two main shortcomings of Support vector machines in Table 7.5. The first shortcoming is their black box nature that is described in the previous paragraph. The second shortcoming of Support vector machines is their very long-lasting model training. For example, the training process of Support vector machines took up to 27 hours. We believe that mediocre performance and these two shortcomings entitle us not to recommend Support vector machines for credit scoring.

Table 7.5: Pros and cons of chosen classifiers

Total Ranking	**Classifier's**		
	Name	**Performance**	**Pros & Cons**
1.	Logistic Regression (LR)	High	+ No need for fine-tuning + Model's interpreability
2.	Artificial Neural Network (ANN)	High	- Black box
3.	Linear Discriminant Analysis (LDA)	High	+ No need for fine-tuning + Model interpretability
4.	Support Vector Machine with Linear Kernel Function (L-SVM)	Medium	- Black box - Demanding model training
7.	Support Vector Machine with Radial Basis Kernel Function (SVM-Rbf)	Medium	- Black box - Demanding model training
9.	Classification and Regression Tree (CART)	Low	+ Model's interpretability
10.	k - Nearest Neighboor (k-NN)	Low	+ Easy model training - Prolonged prediction process

Source: *Authors' processing of chosen classifier's results and feature.*

Total Ranking: *The total ranking is the same as our classifier's ranking in Table 7.2.*

Pros and Cons: *Classifier's pros are marked with plus sign (+) and cons with minus sign (-).*

F3: ***Classification and regression tree and k-Nearest neighbors are <u>not well-suited</u> algorithms for credit scoring.***

These two algorithms do not simply perform well based on the credit scoring data. Our as well as Lessmann *et al.* (2015)'s results ranked both classifiers at the very bottom of the ranking. Even though there are advantages of using

Classification and regression tree and k-Nearest neighbors, such as clear model interpretability or easy model training, these advantages do not justify the low performance of classifiers for their usage. That's why we recommend avoiding the use of these algorithms for credit scoring.

7.5 Further Research Opportunities

During the writing of our master thesis, we came across several interesting research topics that deserve more investigation according to us. The first suggested topic for further research is a general framework of rules used for classifiers comparison. We think that such framework should be derived from studies of Salzberg (1997), Hand (2006) and similar to them. We recognize a need for a definition of such framework because the classifier comparison studies are otherwise hardly comparable. For example, classifier's ranking from studies which do not use k-fold cross validation or several performance measurements cannot be considered as robust. The second proposal for further research might be classifiers performance in the context of imbalanced data sets. We have already mentioned in our master thesis the main finding of Brown & Mues (2012) who made such study. We are, however, concerned with the data sets used for Brown & Mues (2012)'s study. Brown & Mues (2012) used five different data sets with the largest data set having 7,190 records and second largest having 2,974 records. We would be interested to know if their results are for example reproducible based on our Lending Club data. As for last suggestion for further research, we recommend searching for other P2P lending platforms making their data public. Most of nowadays studies is based on data sets from UCI Machine Learning Repository. These data sets have usually unclear origin and rather small size (Salzberg,1997). It is then questionable whether findings of studies based on the UCI repository are meaningful for real world data sets, such as P2P credit scoring data. We, therefore, believe that having a new data set at hand would trigger many interesting research questions.

Chapter 8

Conclusion

The main contribution of our master thesis is ranking of ten classification techniques, displayed in Tables 7.1 and 7.2, based on our Lending Club's data set. To ensure robustness of our results, we use 5-fold cross-validation method and six different performance measurements for classifiers' predictions. We propose five hypotheses based on our literature review. These hypotheses are tested with the help of our classifiers' ranking. Besides that, we highlight three key findings that should serve as the take-away message for our readers.

According to our classifiers' ranking, Logistic regression, Artificial neural network, and Linear discriminant analysis are the three best algorithms based on the Lending Club data. Random forest, the best algorithm in Lessmann *et al.* (2015)'s study, ranks as the fifth best classifier in our master thesis. The performance results of top classifiers are, however, often not significantly different from each other. On the other hand, Classification and regression tree and k-Nearest neighbors rank at very bottom of our ranking.
Despite some differences, such as ranking of Random forest, our ranking results do bear a resemblance to ranking of Lessmann *et al.* (2015).

Comparing our classifiers' ranking with results of Wu (2014), Tsai *et al.* (2014), Chang *et al.* (2015) and Malekipirbazari & Aksakalli (2015) who use Lending Club data too, yield some ambiguous outcomes. This comparison bears some resemblance to Wu (2014) and Tsai *et al.* (2014), but no resembles to studies of Chang *et al.* (2015) and Malekipirbazari & Aksakalli (2015). Nevertheless, it must be pointed out that the primary goal of these studies is not a comprehensive and robust comparison of classifiers but rather an introduction

of some new approaches.

Considering our hypotheses, we summarize them with their rejection status in Table 8.1 below.

Table 8.1: Hypotheses' results

#	Hypothesis	Rejected
H1:	Random forest is the best classifier among our classifiers of interest based on the Lending Club data.	NO
H2:	Artificial neural network outperforms Logistic regression based on the Lending Club data.	NO
H3:	Linearly based classifiers rank in the first half of classifiers' ranking based on the Lending Club data.	YES
H4a:	Logistic regression outperforms Support vector machine with linear kernel function based on the Lending Club data.	YES
H4b:	Logistic regression outperforms Support vector machine with radial basis kernel function based on the Lending Club data.	YES
H5:	Linearly based classication methods underperform when measured by Partial Gini index in comparison with other performance measurements.	NO

Source: *Authors' processing of hypothese' testing result from Section 7.3.*

Finally, we present three key findings that we wish our readers would remember.

F1: *Logistic regression and Linear discriminant analysis are proper classification algorithms for credit scoring.*

F2: *Support vector machines with linear and radial basis kernel functions are unsuitable classification algorithms for credit scoring.*

F3: *Classification and regression tree and k-Nearest neighbors are not well-suited algorithms for credit scoring.*

Bibliography

Abdou, H. & J. Pointon (2011): "Credit scoring, statistical technique and evaluation criteria: a review of the literature." *Intelligent systems in accounting, finance and management* **18(2-3)**: pp. 59–88.

Abdou, H., J. Pointon, & A. El-Masry (2008): "Neural nets versus conventional techniques in credit scoring in Egyptian banking." *Expert Systems with Applications* **35(3)**: pp. 1275–1292.

Akkoc, S. (2012): "An empirical comparison of conventional techniques, neural networks and the three stage hybrid Adaptive Neuro Fuzzy Inference System (ANFIS) model for credit scoring analysis: The case of Turkish credit card data." *European Journal of Operational Research* **222(1)**: pp. 168–178.

Ala'raj, M. & M. F. Abbod (2015): "Classifiers consensus system approach for credit scoring." *Knowledge-Based Systems* **104**: pp. 89–105.

Anagnostopoulos, C., D. J. Hand, & N. M. Adams (2012): "Measuring classification performance : the hmeasure package ." pp. 1–15.

Anderson, R. (2007): *The Credit Scoring Toolkit - Theory and Practice for Retail Credit Risk Management and Decision Automation*. Oxford University Press: New York, 1st edition.

Atz, U. & D. Bholat (2016): "Peer-to-peer lending and financial innovation in the United Kingdom Peer-to-peer lending and financial innovation in." **(598)**.

Bachmann, A., A. Becker, D. Buerckner, M. Hilker, F. Kock, M. Lehmann, & P. Tiburtius (2011): "Online P2P Lending - A Literature Review." *Journal of Internet Banking and Commerce* **16(2)**: pp. 1–18.

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.

Brown, I. & C. Mues (2012): "An experimental comparison of classification algorithms for imbalanced credit scoring data sets." *Expert Systems with Applications* **39(3)**: pp. 3446–3453.

van Buuren, S. & K. Groothuis-Oudshoorn (2011): "MICE: Multivariate Imputation by Chained Equations in R." *Journal of Statistical Software* **45(3)**.

Carmichael, D. (2014): "Modeling default for peer-to-peer loans." *Available at SSRN: http://ssrn.com/abstract=2529240* .

Chang, S., S. D.-o. Kim, & G. Kondo (2015): "Predicting Default Risk of Lending Club Loans." *Machine Learning* pp. 1–5.

Chuang, C. L. & R. H. Lin (2009): "Constructing a reassigning credit scoring model." *Expert Systems with Applications* **36(2 PART 1)**: pp. 1685–1694.

Cichosz, P. (2015): *Data Mining Algorithms: Explained Using R.*

Deer, L., J. Mi, & Y. Yuxin (2015): "The rise of peer-to-peer lending in China: An overview and survey case study." *The Association of Chartered Certified Accountants Report* .

Deloitte (2014): "Banking disrupted: How technology is threatening the traditional European retail banking model." *Technical report.*

Duarte, J., S. Siegel, & L. Young (2012): "Trust and credit: The role of appearance in peer-to-peer lending." *Review of Financial Studies* **25(8)**: pp. 2455–2483.

Fisher, R. A. (1936): "the Use of Multiple Measurements in Taxonomic Problems." *Annals of Eugenics* **7(2)**: pp. 179–188.

Foottit, I., N. Tomlinson, & M. Doyle (2016): "Marketplace lending: A temporary phenomenon?" p. 40.

Freedman, S. & G. Z. Jin (2014): "The Information value of online social networks: lessions from peer-to-peer lending." *NBER Working paper. Available at: http://www.nber.org/papers/w19820* .

Hand, D. J. (2006): "Classifier Technology and the Illusion of Progress." *Statistical Science* **21(1)**: pp. 1–14.

Hand, D. J. (2009): "Measuring classifier performance: A coherent alternative to the area under the ROC curve." *Machine Learning* **77(1)**: pp. 103–123.

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.

Havrylchyk, O., C. Mariotto, T. Rahim, & M. Verdier (2016): "What drives the expansion of the peer-to-peer lending?" p. 29.

HE, H. & E. a. Garcia (2010): "Learning from Imbalanced Data Sets." *IEEE Transactions on knowledge and data engineering* **21(9)**: pp. 1263—-1264.

Hernandez-Orallo, J., P. Flach, & C. Ferri (2011): "Brier Curves: a New Cost-Based Visualisation of Classifier Performance." *Proceedings of the 28th International Conference on Machine Learning (ICML-11)* pp. 585–592.

Herzenstein, M., U. M. Dholakia, & R. L. Andrews (2011): "Strategic Herding Behavior in Peer-to-Peer Loan Auctions." *Journal of Interactive Marketing* **25(1)**: pp. 27–36.

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.

Jin, Y. & Y. Zhu (2015): "A data-driven approach to predict default risk of loan for online peer-to-peer (P2P) lending." *Proceedings - 2015 5th International Conference on Communication Systems and Network Technologies, CSNT 2015* pp. 609–613.

Karatzoglou, A., D. Meyer, & K. Hornik (2006): "Support Vector Algorithm in R." *Journal of Statistical Software* **15(9)**: pp. 1–28.

Kuhn, M. & K. Johnson (2013): *Applied Predictive Modeling [Hardcover].*

Leech, C. (2015): "Direct Lending : Finding value / minimising risk." **44(5912554)**.

Lessmann, S., B. Baesens, H. V. Seow, & L. C. Thomas (2015): "Benchmarking state-of-the-art classification algorithms for credit scoring: An update of research." *European Journal of Operational Research* **247(1)**: pp. 124–136.

Lichman, M. (2013): "UCI Machine Learning Repository." *University of California, Irvine, School of Information and Computer Sciences"* p. 2013.

Lin, M., N. Prabhala, & S. Viswanathan (2013): "Judging borrowers by the company they keep: friendship networks and information asymmetry in online peer-to-peer lending." *Management Science* **59(1)**: pp. 17–35.

Malekipirbazari, M. & V. Aksakalli (2015): "Risk assessment in social lending via random forests." *Expert Systems with Applications* **42(10)**: pp. 4621–4631.

Mays, E. (2001): *Handbook of Credit Scoring*. June. Global Professional Publishing.

McMillan, J. (2014): *The End of Banking: Money, Credit, and the Digital Revolution*.

Mills, K. G. (2014): "The State of Small Business Lending: Credit Access during the Recovery and How Technology May Change the Game." *Harvard Business School Working Paper* **(No. 15-004)**.

Namvar, E. (2013): "An Introduction to Peer to Peer Loans as Investments." *Journal of Investment Management* **12(1)**: pp. 1–18.

Pope, D. G. & J. R. Sydnor (2011): "What's in a Picture? Evidence of Discrimination from Prosper.com." *Journal of Human Resources* **46(1)**: pp. 53–92.

Pundir, S. & R. Seshadri (2012): "A novel concept of partial lorenz curve and partial gini index." *International Journal of Engineering, Science and Innovative Technology* **1(2)**: pp. 296–301.

PwC (2015): "Peer pressure: How peer-to-peer lending are transforming the cosumer lending industry." *Technical report*.

Rufibach, K. (2010): "Use of Brier score to assess binary predictions." *Journal of Clinical Epidemiology* **63(8)**: pp. 938–939.

Salzberg, S. (1997): "On Comparing Classifiers : Pitfalls to Avoid and a Recommended Approach." *Data Mining and Knowledge Discovery* **328**: pp. 317–328.

Scholkopf, B., A. J. Smola, R. C. Williamson, & P. L. Bartlett (2000): "New Support Vector Algorithms." *Neural computation* **12(5)**: pp. 1207–1245.

Serrano-Cinca, C., B. Gutiérrez-Nieto, & L. López-Palacios (2015): "Determinants of Default in P2P Lending." *PLOS ONE* **10(10)**: pp. 1–22.

Singh, H., R. Gopal, & X. Li (2008): "Risk and Return of Investments in Online Peer-to-Peer Lending." *University of Texas* pp. 1–6.

Tsai, K., S. Ramiah, & S. Singh (2014): "Peer Lending Risk Predictor." *Stanford University CS229 Project Report* .

Tsai, M. C., S. P. Lin, C. C. Cheng, & Y. P. Lin (2009): "The consumer loan default predicting model - An application of DEA-DA and neural network." *Expert Systems with Applications* **36(9)**: pp. 11682–11690.

Wardrop, R., B. Zhang, R. Rau, & M. Gray (2015): "The European Alternative Finance Benchmarking Report." *Universtiy of Cambridge Report* .

Welch, B. (1939): "Note on discriminant functions." *Biometrika* **31(1)**: pp. 218–220.

Wendler, T. & S. Gröttrup (2016): *Data Mining with SPSS Modeler*. Springer International Publishing.

Wu, J. (2014): "Loan default prediction using lending club data." *Available at http://www.wujiayu.me/assets/projects/loan-default-prediction-Jiayu-Wu.pdf* .

Yeh, I. C. & C. h. Lien (2009): "The comparisons of data mining techniques for the predictive accuracy of probability of default of credit card clients." *Expert Systems with Applications* **36(2 PART 1)**: pp. 2473–2480.

Zhang, D., H. Huang, Q. Chen, & Y. Jiang (2007): "A comparison study of credit scoring models." *Proceedings - Third International Conference on Natural Computation, ICNC 2007* **1(Icnc)**: pp. 15–18.

Zhang, J. & P. Liu (2012): "Rational Herding in Microloan Markets." *Management Science* **58(5)**: pp. 892–912.

Appendix A

All Lending Club Variables

There are all together 115 variables in downloaded Lending Club data set. All these variables with their descriptions are depicted in Table A.1. The abbreviated names of variables as well as their descriptions have been taken from LC_Data_Dictionary data set downloaded in download section at Lending Club official web site[1].

Each variable in Table A.1 has a status. Three possible values of variable status are OUT, INC and INC*. The most common status is OUT. Variables with status OUT are not included in our final data set. There are several different reasons why majority of our variables is not part of our final data set. For explanation see part 5.1 Data Preparation. Variables with status INC has been unchanged included in our final data set. Furthermore, variables with status INC* has been transformed. For detailed explanation of our transformation steps see part 5.2 Data Transformation.

[1]Link to the data download section at Lending Club official web site: https://www.lendingclub.com/info/download-data.action

Table A.1: All Lending Club variables with description

Status	Abbreviated Name	Description
INC	acc_now_delinq	The number of accounts on which the borrower is now delinquent.
OUT	acc_open_past_24mths	Number of trades opened in past 24 months.
OUT	addr_state	The state provided by the borrower in the loan application.
OUT	all_util	Balance to credit limit on all trades.
INC	annual_inc	The self-reported annual income provided by the borrower during registration.
OUT	annual_inc_joint	The combined self-reported annual income provided by the co-borrowers during registration.
OUT	application_type	Indicates whether the loan is an individual application or a joint application with two co-borrowers.
OUT	avg_cur_bal	Average current balance of all accounts.
OUT	bc_open_to_buy	Total open to buy on revolving bankcards.
OUT	bc_util	Ratio of total current balance to high credit/credit limit for all bankcard accounts.
INC	chargeoff_within_12_mths	Number of charge-offs within 12 months.
OUT	collection_recovery_fee	Post charge off collection fee.
OUT	collections_12_mths_ex_med	Number of collections in 12 months excluding medical collections.
INC	delinq_2yrs	The number of 30+ days past-due incidences of delinquency in the borrower's credit file for the past 2 years.
INC	delinq_amnt	The past-due amount owed for the accounts on which the borrower is now delinquent.
INC*	desc	Loan description provided by the borrower.
INC	dti	A ratio calculated using the borrower's total monthly debt payments on the total debt obligations, excluding mortgage and the requested LC loan, divided by the borrower's self-reported monthly income.
OUT	dti_joint	A ratio calculated using the co-borrowers' total monthly payments on the total debt obligations, excluding mortgages and the requested LC loan, divided by the co-borrowers' combined self-reported monthly income.
INC*	earliest_cr_line	The month the borrower's earliest reported credit line was opened.
INC*	emp_length	Employment length in years. Possible values are between 0 and 10 where 0 means less than one year and 10 means ten or more years.
OUT	emp_title	The job title supplied by the borrower when applying for the loan.
INC*	fico_range_high	The upper boundary range the borrower's FICO at loan origination belongs to.
INC*	fico_range_low	The lower boundary range the borrowerâ€™s FICO at loan origination belongs to.
OUT	funded_amnt	The total amount committed to that loan at that point in time.
OUT	funded_amnt_inv	The total amount committed by investors for that loan at that point in time.
OUT	grade	LC assigned loan grade.
INC	home_ownership	The home ownership status provided by the borrower during registration or obtained from the credit report.
OUT	id	A unique LC assigned ID for the loan listing.
OUT	il_util	Ratio of total current balance to high credit / credit limit on all install acct.
OUT	initial_list_status	The initial listing status of the loan.
OUT	inq_fi	Number of personal finance inquiries.

Status	Abbreviated Name	Description
OUT	inq_last_12m	Number of credit inquiries in past 12 months.
INC	inq_last_6mths	The number of inquiries in past 6 months (excluding auto and mortgage inquiries).
OUT	installment	The monthly payment owed by the borrower if the loan originates.
OUT	int_rate	Interest rate on the loan.
OUT	issue_d	The month which the loan was funded.
OUT	last_credit_pull_d	The most recent month LC pulled credit for this loan.
OUT	last_fico_range_high	The upper boundary range the borrower's last FICO pulled belongs to.
OUT	last_fico_range_low	The lower boundary range the borrower's last FICO pulled belongs to.
OUT	last_pymnt_amnt	Last total payment amount received.
OUT	last_pymnt_d	Last month payment was received.
INC	loan_amnt	The listed amount of the loan applied for by the borrower.
INC*	loan_status	Current status of the loan.
OUT	max_bal_bc	Maximum current balance owed on all revolving accounts.
OUT	member_id	A unique LC assigned Id for the borrower member.
OUT	mo_sin_old_il_acct	Months since oldest bank installment account opened.
OUT	mo_sin_old_rev_tl_op	Months since oldest revolving account opened.
OUT	mo_sin_rcnt_rev_tl_op	Months since most recent revolving account opened.
OUT	mo_sin_rcnt_tl	Months since most recent account opened.
OUT	mort_acc	Number of mortgage accounts.
OUT	mths_since_last_delinq	The number of months since the borrower's last delinquency.
OUT	mths_since_last_major_derog	Months since most recent 90-day or worse rating.
OUT	mths_since_last_record	The number of months since the last public record.
OUT	mths_since_rcnt_il	Months since most recent installment accounts opened.
OUT	mths_since_recent_bc	Months since most recent bankcard account opened.
OUT	mths_since_recent_bc_dlq	Months since most recent bankcard delinquency.
OUT	mths_since_recent_inq	Months since most recent inquiry.
OUT	mths_since_recent_revol_delinq	Months since most recent revolving delinquency.
OUT	next_pymnt_d	Next scheduled payment date.
OUT	num_accts_ever_120_pd	Number of accounts ever 120 or more days past due.
OUT	num_actv_bc_tl	Number of currently active bankcard accounts.
OUT	num_actv_rev_tl	Number of currently active revolving trades.
OUT	num_bc_sats	Number of satisfactory bankcard accounts.
OUT	num_bc_tl	Number of bankcard accounts.
OUT	num_il_tl	Number of installment accounts.
OUT	num_op_rev_tl	Number of open revolving accounts.
OUT	num_rev_accts	Number of revolving accounts.
OUT	num_rev_tl_bal_gt_0	Number of revolving trades with balance >0.
OUT	num_sats	Number of satisfactory accounts.
OUT	num_tl_120dpd_2m	Number of accounts currently 120 days past due (updated in past 2 months).

Status	Abbreviated Name	Description
OUT	num_tl_30dpd	Number of accounts currently 30 days past due (updated in past 2 months).
OUT	num_tl_90g_dpd_24m	Number of accounts 90 or more days past due in last 24 months.
OUT	num_tl_op_past_12m	Number of accounts opened in past 12 months.
INC	open_acc	The number of open credit lines in the borrower's credit file.
OUT	open_acc_6m	Number of open trades in last 6 months.
OUT	open_il_12m	Number of installment accounts opened in past 12 months.
OUT	open_il_24m	Number of installment accounts opened in past 24 months.
OUT	open_il_6m	Number of currently active installment trades.
OUT	open_rv_12m	Number of revolving trades opened in past 12 months.
OUT	open_rv_24m	Number of revolving trades opened in past 24 months.
OUT	out_prncp	Remaining outstanding principal for total amount funded.
OUT	out_prncp_inv	Remaining outstanding principal for portion of total amount funded by investors.
OUT	pct_tl_nvr_dlq	Percent of trades never delinquent.
OUT	percent_bc_gt_75	Percentage of all bankcard accounts >75% of limit.
OUT	policy_code	Publicly available policy_code = 1. New products not publicly available policy_code = 2.
INC	pub_rec	Number of derogatory public records.
INC	pub_rec_bankruptcies	Number of public record bankruptcies.
INC	purpose	A category provided by the borrower for the loan request.
OUT	pymnt_plan	Indicates if a payment plan has been put in place for the loan.
OUT	recoveries	Post charge off gross recovery.
OUT	revol_bal	Total credit revolving balance.
INC	revol_util	Revolving line utilization rate, or the amount of credit the borrower is using relative to all available revolving credit.
OUT	sub_grade	LC assigned loan subgrade.
INC	tax_liens	Number of tax liens.
INC	term	The number of payments on the loan. Values are in months and can be either 36 or 60.
OUT	title	The loan title provided by the borrower.
OUT	tot_coll_amt	Total collection amounts ever owed.
OUT	tot_cur_bal	Total current balance of all accounts.
OUT	tot_hi_cred_lim	Total high credit / credit limit.
INC	total_acc	The total number of credit lines currently in the borrower's credit file.
OUT	total_bal_ex_mort	Total credit balance excluding mortgage.
OUT	total_bal_il	Total current balance of all installment accounts.
OUT	total_bc_limit	Total bankcard high credit / credit limit.
OUT	total_cu_tl	Number of finance trades.
OUT	total_il_high_credit_limit	Total installment high credit / credit limit.
OUT	total_pymnt	Payments received to date for total amount funded.
OUT	total_pymnt_inv	Payments received to date for portion of total amount funded by investors.

Status	Abbreviated Name	Description
OUT	total_rec_int	Interest received to date.
OUT	total_rec_late_fee	Late fees received to date.
OUT	total_rec_prncp	Principal received to date.
OUT	total_rev_hi_lim	Total revolving high credit / credit limit.
OUT	url	URL for the LC page with listing data.
INC	verification_status	Indicates if income was verified by LC, not verified, or if the income source was verified.
OUT	verified_status_joint	Indicates if the co-borrowers' joint income was verified by LC, not verified, or if the income source was verified.
OUT	zip_code	The first 3 numbers of the zip code provided by the borrower in the loan application.

Appendix B

Descriptive Statistics

There is only one descriptive statistics table included in Appendix C. It is Table B.1 with correlation coefficients between variables in our data set. The result description of this correlation matrix is provided in section 5.3 Descriptive Statistics. Further descriptive statistics of our Lending Club data set might be found in the above mentioned section 5.3.

	loan_ status	loan_ amnt	emp_ length	annual_ inc	desc_ count	dti	delinq_ 2yrs	earliest_ cr_line	fico_ range_ avg	inq_ last_ 6mths	open_ acc	pub_ rec	revol_ util	total_ acc	acc_ now_ delinq	chargeoff_ within_ 12_mths	delinq_ amnt	pub_ rec_ bankr	tax_ liens
loan_status	1	0.07	0.01	-0.06	-0.01	0.08	0.01	-0.03	-0.12	0.06	0.02	0.01	0.08	-0.01	0.01	0.00	0.00	0.01	0.00
loan_amnt	0.07	1	0.12	0.42	0.01	0.06	0.01	0.17	0.08	0.02	0.20	-0.06	0.09	0.25	0.01	0.01	0.00	-0.08	0.00
emp_length	0.01	0.12	1	0.09	-0.08	0.06	0.03	0.22	0.00	-0.01	0.06	0.04	0.05	0.13	0.01	0.01	0.00	0.05	0.01
annual_inc	-0.06	0.42	0.09	1	0.00	-0.20	0.07	0.21	0.10	0.09	0.19	-0.02	0.03	0.29	0.02	0.02	0.01	-0.05	0.03
desc_count	-0.01	0.01	-0.08	0.00	1	-0.06	-0.03	0.00	0.10	0.01	-0.04	-0.04	-0.04	-0.03	-0.01	-0.01	0.00	-0.04	-0.01
dti	0.08	0.06	0.06	-0.20	-0.06	1	0.00	0.03	-0.16	0.01	0.32	-0.04	0.24	0.24	0.01	-0.01	0.00	-0.03	-0.01
delinq_2yrs	0.01	0.01	0.03	0.07	-0.03	0.00	1	0.09	-0.18	0.02	0.06	-0.02	-0.01	0.13	0.11	0.12	0.03	-0.03	0.00
earliest_cr_line	-0.03	0.17	0.22	0.21	0.00	0.03	0.09	1	0.16	0.01	0.15	0.05	-0.02	0.32	0.02	0.02	0.00	0.05	0.02
fico_range_avg	-0.12	0.08	0.00	0.10	0.10	-0.16	-0.18	0.16	1	-0.03	-0.03	-0.15	**-0.54**	0.04	-0.03	-0.04	-0.01	-0.16	-0.03
inq_last_6mths	0.06	0.02	-0.01	0.09	0.01	0.01	0.02	0.01	-0.03	1	0.12	0.01	-0.09	0.15	0.00	0.01	0.00	0.01	0.01
open_acc	0.02	0.20	0.06	0.19	-0.04	0.32	0.06	0.15	-0.03	0.12	1	-0.02	-0.10	**0.67**	0.01	0.01	0.00	-0.03	0.00
pub_rec	0.01	-0.06	0.04	-0.02	-0.04	-0.04	-0.02	0.05	-0.15	0.01	-0.02	1	-0.02	-0.01	0.01	-0.01	0.00	**0.76**	**0.63**
revol_util	0.08	0.09	0.05	0.03	-0.04	0.24	-0.01	-0.02	**-0.54**	-0.09	-0.10	-0.02	1	-0.08	-0.01	-0.01	0.00	-0.02	-0.01
total_acc	-0.01	0.25	0.13	0.29	-0.03	0.24	0.13	0.32	0.04	0.15	**0.67**	-0.01	-0.08	1	0.03	0.04	0.01	-0.01	0.00
acc_now_delinq	0.01	0.01	0.01	0.02	-0.01	0.01	0.11	0.02	-0.03	0.00	0.01	0.01	-0.01	0.03	1	0.06	0.23	0.00	0.01
chargeoff_12_m	0.00	0.01	0.01	0.02	-0.01	-0.01	0.12	0.02	-0,04	0.01	0.01	-0.01	-0.01	0.04	0.06	1	0.02	-0.01	0.00
delinq_amnt	0.00	0.00	0.00	0.01	0.00	0.00	0.03	0.00	-0.01	0.00	0.00	0.00	0.00	0.01	0.23	0.02	1	0.00	0.00
pub_rec_bankrp	0.01	-0.08	0.05	-0.05	-0.04	-0.03	-0.03	0.05	-0.16	0.01	-0.03	**0.76**	-0.02	-0.01	0.00	-0.01	0.00	1	0.06
tax_liens	0.00	0.00	0.01	0.03	-0.01	-0.01	0.00	0.02	-0.03	0.01	0.00	**0.63**	-0.01	0.00	0.01	0.00	0.00	0.06	1

Table B.1: Correlation matrix of LC variables

Appendix C

Meta-parameters of classifiers

Table C.1 below displays following information: meta-parameters of classifiers that we have fine-tuned, fine-tuned values of meta-parameters and analytical software we have used. The symbol n/a in meta-parameters description denotes that it was not needed to fine-tuned given classifier. We would like to point out that it might happen that different analytical software require different meta-parameters. Our main software for classifier's scoring was IBM SPSS Modeler 18.0. We have chosen SPSS Modeler because of its reliability, ease of use and authors' proficiency in this software. Two classification methods, Naïve Bayes and Random forest, are not covered by SPSS Modeler. Therefore, our second software is R 3.4.0. The R's packages we have used are included in parentheses.

Table C.1: Meta-parameters of Classifiers

Classifier	**Meta-parameter**	**Value**	**Software**
Artificial neural network (ANN)	# of hidden nodes # of units in hidden nodes	2 22;10	SPSS Modeler
Bayesian net (B-Net)	Structure of network	TAN	SPSS Modeler
Classification and regressio tree (CART)	Tree depth Min. leaf size	6 2%	SPSS Modeler
k-Nearest Neighbor (k-NN)	# of nearest neighbors	10	SPSS Modeler
Linear discriminant analysis (LDA)	n/a		SPSS Modeler
Logistic regression (LR)	n/a		SPSS Modeler
Linear support vector machine (L-SVM)	Epsilon Lambda	0.1 5	SPSS Modeler
Naïve Bayes (NB)	n/a		R (e1071)
Support vector machine - radial (SVM-Rbf)	Epsilon Gamma	0.1 0.1	SPSS Modeler
Random forest (RF)	# of grown trees # of randomly sampled variables	800 5	R (randomForest)

Appendix D

Results of Iterations

The 5-fold cross validation approach has been used to measure classifiers' performance. Below are, therefore, five partial results from five cross validation iterations. The results of best classifiers based on given performance measurements are underlined. These partial results serve as basis for average classifier performance results displayed in Table 7.1.

Table D.1: Results from first iteration of 5-fold cross validation

Classifier	**Performance measurement**					
	PCC	**KS**	**BS**	**AUC**	**PG**	**H**
RF	0.7892	0.2793	0.1240	0.6935	0.2213	0.1245
ANN	0.7906	0.2796	0.1232	0.6957	0.2347	0.1285
LR	0.7906	0.2874	0.1235	0.6952	0.2517	0.1271
LDA	0.7879	0.2846	0.1243	0.6934	0.2658	0.1243
SVM-Rbf	0.7819	0.2093	0.1295	0.6501	0.2572	0.1073
L-SVM	0.7876	0.2850	0.1583	0.6942	0.2365	0.1254
B-Net	0.7859	0.2544	0.1254	0.6748	0.2028	0.1066
NB	0.7864	0.2340	0.1423	0.6657	0.2146	0.0997
k-NN	0.7792	0.1972	0.1313	0.6334	0.1418	0.0663
CART	0.7484	0.2536	0.2114	0.6347	0.1356	0.0763

Table D.2: Results from second iteration of 5-fold cross validation

Classifier	Performance measurement					
	PCC	**KS**	**BS**	**AUC**	**PG**	**H**
RF	0.7865	0.2693	0.1253	0.6892	0.2158	0.1206
ANN	0.7917	0.2864	0.1239	0.6994	0.2370	0.1339
LR	0.7925	0.2890	0.1237	0.6998	0.2457	0.1356
LDA	0.7929	0.2852	0.1242	0.6979	0.2521	0.1328
SVM-Rbf	0.7802	0.2146	0.1315	0.6501	0.2716	0.1065
L-SVM	0.7920	0.2876	0.1585	0.6988	0.2247	0.1342
B-Net	0.7892	0.2590	0.1254	0.6820	0.2126	0.1167
NB	0.7696	0.2455	0.1517	0.6698	0.2023	0.1045
k-NN	0.7442	0.2039	0.1328	0.6370	0.1527	0.0696
CART	0.7786	0.2587	0.2065	0.6440	0.1937	0.0867

Table D.3: Results from third iteration of 5-fold cross validation

Classifier	Performance measurement					
	PCC	**KS**	**BS**	**AUC**	**PG**	**H**
RF	0.7905	0.2912	0.1235	0.6981	0.2334	0.1316
ANN	0.7915	0.2914	0.1230	0.6999	0.2479	0.1345
LR	0.7935	0.2968	0.1228	0.7002	0.2524	0.1359
LDA	0.7930	0.2886	0.1234	0.6980	0.2599	0.1325
SVM-Rbf	0.7852	0.2162	0.1298	0.6577	0.2723	0.1131
L-SVM	0.7919	0.2926	0.1582	0.6995	0.2352	0.1343
B-Net	0.7902	0.2585	0.1247	0.6800	0.2065	0.1154
NB	0.7900	0.2485	0.1624	0.6708	0.1734	0.1070
k-NN	0.7451	0.2041	0.1321	0.6360	0.1648	0.0666
CART	0.7527	0.2665	0.1998	0.6392	0.1421	0.0863

Table D.4: Results from fourth iteration of 5-fold cross validation

Classifier	Performance measurement					
	PCC	**KS**	**BS**	**AUC**	**PG**	**H**
RF	0.7874	0.2808	0.1262	0.6937	0.2277	0.1240
ANN	0.7885	0.2834	0.1256	0.6975	0.2551	0.1285
LR	0.7890	0.2901	0.1252	0.7004	0.2569	0.1336
LDA	0.7888	0.2838	0.1258	0.6973	0.2670	0.1291
SVM-Rbf	0.7822	0.2080	0.1314	0.6537	0.2640	0.1145
L-SVM	0.7828	0.2854	0.1590	0.6986	0.2374	0.1308
B-Net	0.7878	0.2552	0.1269	0.6803	0.2167	0.1137
NB	0.7858	0.2496	0.1420	0.6731	0.2403	0.1035
k-NN	0.7792	0.2041	0.1329	0.6379	0.1441	0.0697
CART	0.7643	0.2602	0.2018	0.6296	0.1202	0.0763

Table D.5: Results from fifth iteration of 5-fold cross validation

Classifier	Performance measurement					
	PCC	**KS**	**BS**	**AUC**	**PG**	**H**
RF	0.7881	0.2737	0.1251	0.6896	0.2199	0.1206
ANN	0.7903	0.2831	0.1243	0.6953	0.2519	0.1270
LR	0.7910	0.2794	0.1243	0.6939	0.2443	0.1274
LDA	0.7894	0.2743	0.1250	0.6911	0.2480	0.1236
SVM-Rbf	0.7796	0.2041	0.1307	0.6479	0.2554	0.1027
L-SVM	0.7891	0.2763	0.1587	0.6924	0.2207	0.1252
B-Net	0.7857	0.2502	0.1260	0.6763	0.2155	0.1086
NB	0.7862	0.2352	0.1526	0.6649	0.1908	0.0997
k-NN	0.7796	0.2018	0.1319	0.6357	0.1478	0.0667
CART	0.7856	0.2397	0.1898	0.6392	0.1558	0.0748