**Lappeenranta-Lahti University of Technology LUT**

**School of Business and Management**

**Master's Programme in Strategic Finance and Business Analytics**

**Master's Thesis**

# Default prediction in peer-to-peer lending and country comparison

Matias Koskimäki

2021

1st examiner: Mikael Collan

2nd examiner: Jyrki Savolainen

# ABSTRACT


| | |
|---|---|
| **Author:** | Matias Koskimäki |
| **Title:** | Default prediction in peer-to-peer lending and country comparison |
| **Faculty:** | LUT School of Business and Management |
| **Major:** | Master's Programme in Strategic Finance and Business Analytics |
| **Year:** | 2021 |
| **Master's Thesis:** | 80 pages, 14 figures, 10 tables, 14 appendices |
| **Examiners:** | Professor Mikael Collan, Postdoctoral researcher Jyrki Savolainen |
| **Key words:** | P2P lending, default prediction, country comparison, |



The purpose of this thesis is to predict default in P2P lending and compare prediction performance and variable importance between countries. This research is done using feature selection (FS) and random under-sampling (RUS) in data preparation. Dataset is also split to each country. These datasets are then trained using machine learning. Selected models are Logistic regression (LR), Support vector machine (SVM), and Random Forest (RF) and parameters are optimized using hyper parameter optimization and models are trained using 10-fold cross validation. This thesis uses credit data from P2P lending site Bondora, an Estonian P2P lending platform. Classification results are evaluated using multiple metrics derived from confusion matrix and area under ROC curve (AUC)

The results show that default can be predicted very accurately with these methods. Prediction performance, according to evaluation metrics, does not get better when dividing dataset to specific countries. Overall models perform best when they are used on whole dataset. This could be due to smaller sample size when data is split to each country. Interestingly, Finnish dataset, when using RF model, managed to predict default class the best out of all other models and datasets. This gives an indication that, with enough data on each country, results could have been different. Supervised machine learning models tend to perform best with very large datasets. Also, countries have similarities in variable importance, but some variables stood out in specific countries. Also, some variables had opposite effects on default probability in different countries.

# TIIVISTELMÄ


| | |
|---|---|
| **Tekijä:** | Matias Koskimäki |
| **Otsikko:** | Luottoriskin ennustaminen vertaislainauksessa ja maakohtainen vertailu |
| **Akateeminen yksikkö:** | LUT School of Business and Management |
| **Maisteriohjelma:** | Master's Programme in Strategic Finance and Business Analytics |
| **Vuosi:** | 2021 |
| **Pro gradu:** | 80 sivua, 14 kuviota, 10 taulukkoa, 14 liitettä |
| **Ohjaajat:** | Professori Mikael Collan, Tutkijatohtori Jyrki Savolainen |
| **Hakusanat:** | Vertaislainaus, luottoriskin ennustaminen, maa vertailu |



Tämän tutkimuksen tarkoitus on ennustaa luottoriskiä vertaislainauksessa ja tarkastella tärkeitä muuttuja, sekä vertailla tuloksia maakohtaisesti. Tutkimuksessa käytettiin muuttujavalintaa sekä satunnaisotantaa, jotta ennustus mallit toimisivat mahdollisimman hyvin. Data on jaettu myös eri maihin. Data koulutettiin käyttämällä logistista regressiota, tukivektorikonetta ja satunnaista metsää. Parametrit myös optimoitiin hyper-parametrioptimoinnilla ja mallit koulutettiin 10-kertaisella ristiin validoinnilla. Tutkimuksessa käytetään dataa vertaislaina sivustolta nimeltä Bondora, joka on virolainen vertaislainapalvelu. Luokittelutulokset arvioidaan käyttämällä sekaannusmatriisista johdettuja mittareita, sekä AUC (area under ROC curve) -tunnuslukua.

Tulokset näyttävät, että luottoriskiä voidaan ennustaa hyvin tarkasti käyttämällä koneoppimisen malleja. Mallien ennustuskyky ei parane, kun data jaetaan eri maihin. Mallit ennustavat parhaiten kaiken datan avulla. Tämä voi johtua tietoaineiston koosta, sillä koko datassa on paljon enemmän tapauksia verrattuna siihen, että ne olisi jaettu maihin. Mielenkiintoinen havainto löytyy kuitenkin Suomen datasta, sillä maksukyvyttömyyttä pystyttiin ennustamaan parhaiten satunnaisella metsällä verrattuna muihin maihin ja koko dataan. Tämä osittaa, että maakohtaisia eroja löytyy, mutta niiden ennustamiseen pitäisi olla tasavertaiset tietoaineistot. Eri maiden luottoriskiin vaikuttaa pääasiassa samat muuttujat, mutta myös ainutlaatuisia muuttujia löytyy jokaisesta maasta. Jotkin muuttujat vaikuttavat myös päinvastoin luottoriskiin eri maissa.

# ACKNOWLEDGEMENTS

This journey has been amazing as a student of LUT. I have had many special moments and memorable experiences with my classmates during this 6-year stretch. The first five years I lived and studied in Lappeenranta, where I met amazing people and made a lot of friends. This last year has been very rough for all of us and I hope we can soon gaze into the future with positive thoughts soon enough. I cannot even imagine that my university journey is coming to its end, but it is time to move towards new challenges.

I want to thank my supervisors Mikael Collan and Jyrki Savolainen for their guidance in this thesis. With their help I was able to do this project in a proper manner. It has been a long road but with your help, my goals became much clearer.

I want to give big thanks to my family who have been there for me this final year. I moved back to Helsinki last year and it has been a tough year with my knee surgery, corona and thesis working at the same time, so I was extremely stressed. But they helped me to get through this tough time with endless support.

I want to give special thanks to all the friends I made in LUT. Our time in university was amazing and I will always remember fondly our excursions to Ruka and other shenanigans we had. I hope we keep in touch in the future. Thank you for all the precious memories!

In Helsinki, June 20th, 2021

Matias Koskimäki

# TABLE OF CONTENTS

# FIGURES



# TABLES

# APPENDICES

# LIST OF ABBREVIATIONS

| | |
|---|---|
| AUC | Area under the ROC curve |
| CV | Cross-validation |
| FN | False negatives |
| FP | False positives |
| FS | Feature selection |
| LR | Logistic regression |
| ML | Machine learning |
| P2P | Peer-to-peer |
| RF | Random Forest |
| ROC | Receiver operating characteristic curve |
| RUS | Random Under-Sampling |
| SVM | Support vector machine |
| TN | True negatives |
| TP | True positives |

# 1 INTRODUCTION

This master's thesis begins with an introduction chapter where the background of the topic and motivation, the focus of the study and research questions are explained. This chapter provides a starting point for this thesis.

## 1.1 Background of the topic

Peer-to-peer (P2P) financing is not a new idea. It has been used long by premarket societies but, it was embedded in social relations. People loaned money or other things to friends they knew where trustworthy and had the capabilities to pay back in time. However, modern markets act differently. Now days lending involves rationale and calculations to optimize risk and return (Granovetter 1985). Banks became the normal way to borrow, and so P2P lending declined.

However, social network sites such as Facebook and LinkedIn have changed the landscape of social embeddedness. Social relations can now be created and maintained through internet. This also makes the relations highly visible and transparent (Kane et al. 2014; Oestreicher-Singer and Sundararajan 2012). More and more online platforms emerge that are seeking to use these social relations for economic activities such as lending (Bondora) and rentals (AirBnB). As individuals connected by powerful social networking tools, it is inevitable that social relations are used for economic purposes. Such is the case with P2P lending where individual lenders can collectively bid on loan requests of other individuals in an online platform. (Liu et al. 2015)

With P2P lending made easier by platforms. Individuals can start to invest in loans like banks do. In order to do this properly, one should identify the characteristics that effect on borrowers' capabilities of taking care of liabilities. This study's purpose is to identify these variables that effect the performance of borrowers. Furthermore, in this study, a comparison of countries is being made. Also, a predictive model is being applied and tested to see whether it is accurate enough to be used for economic purposes.

These results could be used for evaluating borrowers and foresee whether they default or not based on just the characteristics. This would make the decision making for investors a lot easier. Also, this study could be helpful to determine whether P2P loans are good enough investment alternative compared to, for example, stocks.

## 1.2 Focus and contribution of the study

This is a quantitative study which uses a large amount of data. This study concentrates on P2P lending as a personal investment alternative. The focus of this study is to explore P2P lending data and find variables that effect on borrower's performance. Also, focus is on country comparison. In this study, the purpose is to see whether there is a difference in borrower performance between countries. If there is a difference, one can possibly reduce risk by investing in some particular country's loans.

Motivation of doing this study is to learn about different modelling methods and learn more about variable analysis. Data is the key in the modern world but by itself it is worth nothing. Refining data to your own needs makes all the difference and creates value for future decision making.

Secondly, P2P lending is a very interesting phenomenon, and its popularity has been increasing many folds during the years. Using P2P lending as investment alternative can possibly result in more returns than traditional means of investing. Learning how to process P2P lending data and using it can be very useful. With it there is a possibility to beat, for example, stock market in returns.

## 1.3 Research questions, objectives, and limitations

This study examines the variables that define a good borrower and whether there is difference in borrower performance between countries and can it be predicted accurately. The goal of this study is to answer the following research questions in clear manner

*“What has been previously researched in literature?*”

Hypothesis: Credit management in general is very extensively researched topic, but P2P lending is relatively new phenomena. There might be areas that have not been researched before. For example, country comparison studies in P2P lending seem to be very scarce.

*“What are the differences in country borrower populations and default predictability?”*

Hypothesis: There is a slight difference since countries have different demographics and social structures. Also, different cultures might have an impact. These differences can result in different predictability.

*“Are there identifiable characteristics that explain borrower default?”*

Hypothesis: There are factors that help to determine borrower default. Financial variables should have significant impact on performance. Still, there might be many other significant variables that are not obvious.

The main objective of this study is to learn more about P2P lending and use models necessary to evaluate possible lending options and to create a predictive model. The motivation behind this study is that any P2P investor can use these methods to evaluate and predict possible lending options.

This study does not compare P2P lending with other investment alternatives. This study only tries to predict defaulting borrowers and compare these predictions between countries.

## 1.4 Structure of the thesis

This thesis is structured in two main parts. First part is the theoretical segment, which include chapters P2P lending, machine learning, literature review, and methodology. In these chapters all the necessary information is acquired for second part of the thesis. P2P lending chapter describes this phenomenon and how it differs from traditional lending. Machine learning chapter describes what it is and how it can be used for P2P lending purposes. Literature review chapter describes all relevant research on subject of machine learning and lending, which gives the knowledge on how different methods work and how they should be applied. Methodology chapter describes all the methods chosen and how to use them in the second phase of the thesis.

Second part is empirical analysis. In this part default is predicted using P2P lending data, and different countries predictions are compared. First, data is pre-processed so it can be used in machine learning purposes. Next, feature selection and data balancing are implemented on the data so that prediction results are better. In the next part, machine learning models are trained using hyper parameter optimization and 10-fold cross validation. Then, models are tested how well they can predict correct labels on unknown data and evaluation metrics are analysed. Next, important variables of identified and researched how they affect default probability. Finally, conclusions are made based on the results. In this part the whole research process is summarized, research questions are answered, and further research possibilities are examined.

# 2 PEER-TO-PEER LENDING

In 2007 the world experienced a global economic crisis that shattered the belief in financial sector (Atz & Bholat 2016). This crisis led to actions that restricted the financial sector significantly and the traditional banks could not lend money to people who had low credit scores (Crotty 2009). This new formation of banking led to a situation where some people could not finance their investments. Thus, new form of financing, P2P lending, was born. Although this was not completely a new idea since people have always lent money to one another. But the creation of P2P lending platforms was the innovation that made lending to lower/mid income citizens possible and easier.

The roots of P2P lending go as deep as ancient Babylonian civilisation. In fact, P2P lending was the first form of financing by credit. Babylonians gave credit to individuals so that agricultural projects could develop. P2P lending continued to be the major form of financing until 1300s when banking became the central form of financing. The success and growth of modern banking was mostly due to the ability to diversify lending to a large population. This lowered the risk significantly. (Namvar 2013)

The development of internet and consumer data has grown rapidly recently which virtually eliminated previously mentioned risk-barriers of entry and re-opened the doors for P2P lending. Risks were previously much greater since investors could not assess credit risk of borrowers as well as diversifying investments was very difficult since all loans were limited geographically for both borrowers and lenders. The development of internet allows investors to reach millions of borrowers and gave the ability to diversify portfolios geographically. Furthermore, intermediary P2P operator facilitates the loan, which reduces costs for both investors and borrowers. This redirects the profits to the investor, rather than a bank. (Namvar 2013)

As mentioned previously, financial crisis led to a situation where people with lower credit score could not get a loan with acceptable terms anymore. Deutsche Bank reported in 2013 that approximately 48 million consumer borrowers with credit scores between 650-750 have less financing options than before the crisis. Thus, there is a large untapped consumer lending market. This has led to development of P2P lending platforms. (Namvar 2013)

First lending platform was created 2005 in UK. This platform was called Zopa. The Founders of Zopa recognised that the growing part of population would become contract workers not in full-time job who were creditworthy but unable to access credit from traditional banks. Also, they recognised lenders perspective in a way that savvy financiers could use this new asset class to reduce risk by diversifying their portfolios with multiple loans. (Atz & Bholat 2016) Soon other entrepreneurs recognized this uncapitalized market and started creating platforms of their own. Nowadays there are dozens of P2P lending platforms all over the world and P2P lending is growing as a financing alternative every single day.

P2P lending occurs at the intersection of e-commerce and sharing economy (Ye et al. 2018). P2P loans are usually personal loans that are unsecured and often utilized by individual borrowers. Although some loans can be issued by small companies (Namvar 2013). Lenders and borrowers are directly matched through online services, platforms (like Lending Club or Bondora) (Zhao et al. 2017). Since this direct matching happens online, platforms can operate with lower costs than traditional financial institutions. Online platforms make micro-financing possible without going through financial intermediaries (Zhao et al. 2017). For investors, P2P lending can create a predictable, high yield income from diversified portfolio of these loans. These two points are the key aspects that creates the competitive advantage of lending platforms compared to traditional banking (Namvar 2013). But there is a catch. All P2P loans, as previously mentioned, are unsecured which means that loans do not have a collateral. Also, there is an information asymmetry between lenders and borrowers. Lenders do not know much about the borrows which may lead to losses in loans. Therefore, assessing borrower's creditworthiness is very important (Pokorná & Sponer 2016).

The focus of P2P operators has been primarily personal and small business loans. But operators are expanding more and more into different loan markets such as trade credit and mortgages. P2P lending is often considered as a platform to connect borrowers and retail investors, but it has evolved such that on some platforms most investor funds comes from institutional investors. (Davis & Murphy 2016)

In this chapter, P2P lending is explained in detail. First, the lending process is examined to see how actually this kind of lending works. Next, the development of P2P financing around the world is being researched. Then, pros and cons of P2P financing for both investors and borrowers are examined. Finally, the credit risk of using P2P financing is researched.

## 2.1 P2P lending process

P2P lending mechanisms are almost the same in all the platforms. First, potential users, like borrowers and lenders, register with personal information. With this personal information, the credit rating of users is calculated, and the user verified. Next, borrowers provide information of their loan size, maximum interest rate willing to offer and some other information like loan purpose, repayment information etc. Then, lenders provide a certain amount of money and choose what lending pattern to use. Currently, there are two choices. One pattern is that the lender chooses a platform and provides the money to borrower directly. Another pattern is that lender puts the money in a pool of funds. P2P company then distributes the money to different borrowers. Downside here is that lender does not know borrower's information. When the loan is fully funded, the borrower may have to provide additional documents to verify the creditworthiness. (Wang et al. 2015)

Some platforms, like Prosper, uses auction mechanism in a way that lenders place bids on loans defining interest rate and amount. This auction lasts several days (14 days in Prosper), and the lenders can undercut each other by placing lower interest rates. This continues until the end date. Lowest interest rate wins. (Bachmann et al. 2011)

Bondora (2020b) offers its customers in-depth historical data about creditworthiness and lending trends. Using this data one can create models that help to determine good borrowers. Bondora also has algorithms that select good borrowers and loans for you automatically, recommendations. These options have already a predetermined interest rate. These interest rates are calculated from various variables, such as FICO score.

In the next page, Figure 1 shows a simplified illustration of P2P financing process. Figure from Davis & Murphy 2016 article was used as a model for this. After borrowers and lenders have registered, the process starts. Borrower pays a certain amount of service fees to access the platform and its services to apply for a loan. Then, lender decides to loan some amount in the platform. Lending in platforms also involves a service fee. Usually, platforms have recommendation systems that suggest borrowers for lenders. These recommendation systems work in a way that lender specifies what kind of borrower (interest rate, risk etc.) he/she wants. Then, the system, based on the specifications, recommends the best borrowers matching the

specifications. But there is a possibility to use own judgement without any systems. For example, Bondora provides customer data so one can make own analysis and decisions but its very time consuming. After matching borrowers and lenders, the platform then reassures borrower by asking more credentials like mentioned before. At this point the loan is set and lenders transfer the funds to borrowers directly via the platform. P2P operator performs ex post monitoring and management of borrowers for investors (Davis & Murphy 2016).

Overtime borrower pays interest and finally principal of the loan. Now, loan contract has ended, and all parties are satisfied. Borrower got loan, P2P platform got service fees from both borrowers and lenders, and lender made profit from the loan.

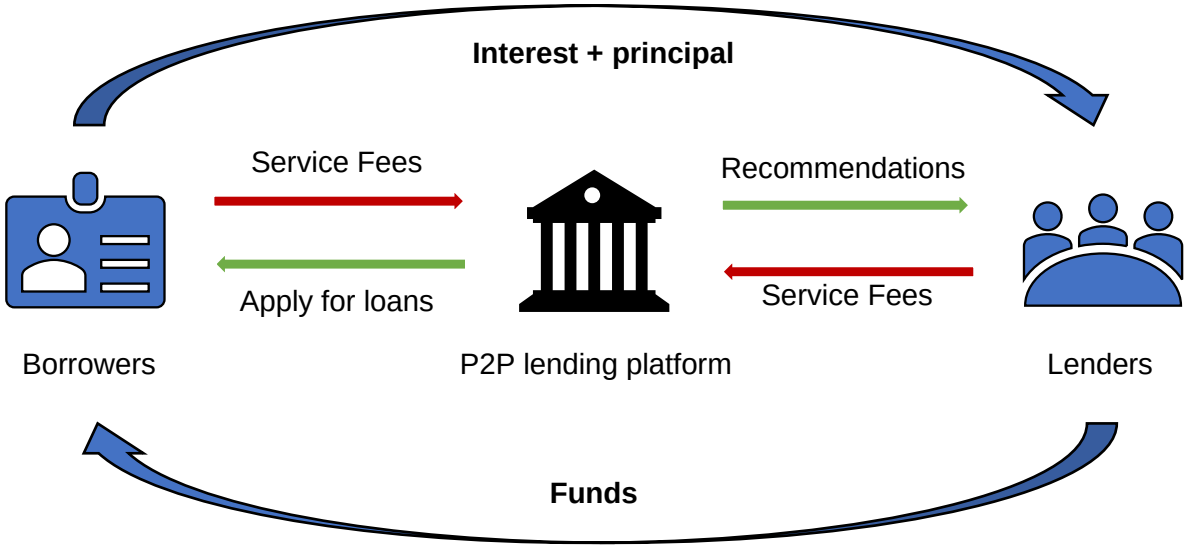



Figure 1. Simplified P2P lending process

In a nutshell P2P platforms main business compiles just of these service fees (Davis & Murphy 2016). To further increase firms profit they need economics of scale. Their goal is to get as many customers as possible. This creates a principal-agent problem for P2P operators since their short-term incentive is to maximise loan volume which could affect the assessment of creditworthiness (Davis & Murphy 2016). Bondora has set fixed rate on their service fees. Contract fee is 3.65 % of the loan amount to max value of 150 € and annual management fee is also 3.65 % to max value of 150 €. Furthermore, they have debt collection fees which are default notification letter and debt notifications. In comparison, Lending Club, which is a P2P provider from U.S, has a service fee of 2-6 % of the loan amount. So, it is possible to get lower

service fees from Lending Club, but Bondora seems to be more transparent with just one fixed percentage. (Bondora 2020c; Lending Club 2020)

## 2.2 P2P lending around the world

Ziegler et al. (2020) have constructed a very thorough report of worlds situation considering alternative financing options. These options have P2P lending in them, and they represent most of the alternative lending. According to Ziegler et al. (2020) there are three major P2P lending forms. These forms are Consumer lending, business lending, and property lending. In consumer lending individuals or institutional funders provide a loan to a consumer borrower. In business lending the process is the same, but the borrower is a business. In property lending individuals or institutional funders provide a loan, secured against a property, directly to a consumer or business borrower. In Figure 2 is shown how popular these forms are. Data for this figure is from Ziegler et al. (2020). The number in each piece is in millions of US dollars. Consumer lending has the vast majority. Business lending has one fifth of the whole P2P lending. P2P property lending only has 2 % which is understandable because why put your house on the line when you can get a loan without a collateral.

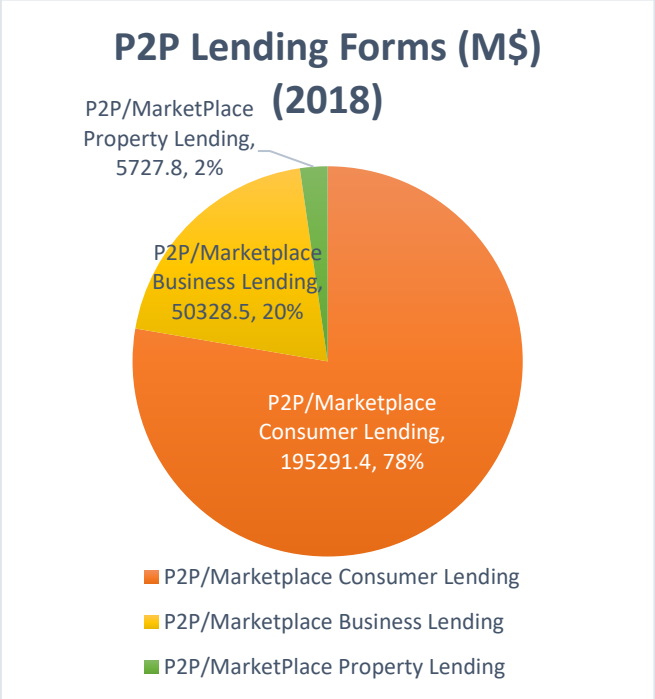
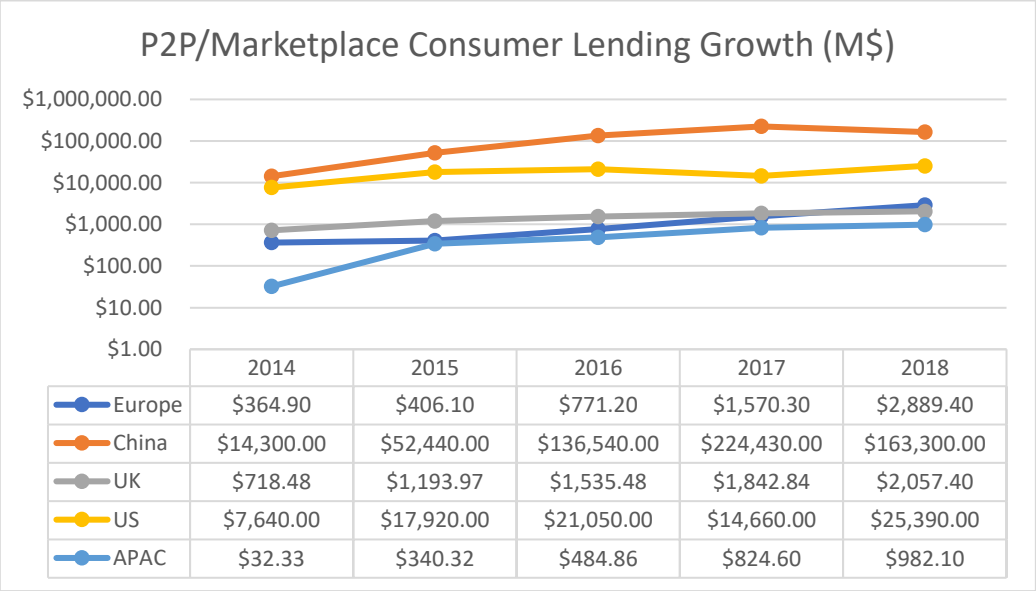



Figure 2. P2P lending form portions

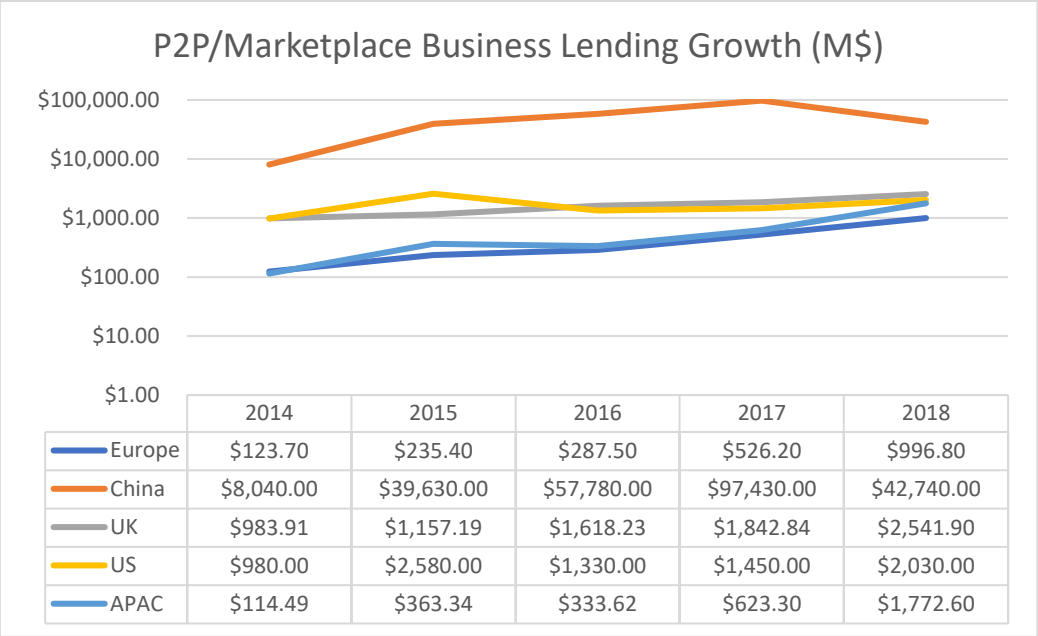
All these areas of P2P lending have seen a substantial growth in recent years. Ziegler et al. (2020) have collected data over many years from alternative financing. This data is represented in following figures. There are three figures for each form of P2P lending. All figures needed to be transformed into logarithmic scale since China has so much volume in P2P lending that it suppresses other countries or even continent's graphs. Also, only five largest markets have been chosen to represent the growth. Furthermore, pay attention to y-axis numbers. Figure 3 considers consumer lending. As we can see from the figure, the growth has been very substantial. China for example has increased its consumer P2P lending by more than 10-fold. This figure shows that P2P consumer lending is constantly increasing overtime. This means, that research is indeed needed. The more research we have, the more informed decisions every participant in P2P lending can make.




| | 2014 | 2015 | 2016 | 2017 | 2018 |
|---|---|---|---|---|---|
| Europe | $364.90 | $406.10 | $771.20 | $1,570.30 | $2,889.40 |
| China | $14,300.00 | $52,440.00 | $136,540.00 | $224,430.00 | $163,300.00 |
| UK | $718.48 | $1,193.97 | $1,535.48 | $1,842.84 | $2,057.40 |
| US | $7,640.00 | $17,920.00 | $21,050.00 | $14,660.00 | $25,390.00 |
| APAC | $32.33 | $340.32 | $484.86 | $824.60 | $982.10 |

Figure 3. P2P consumer lending growth

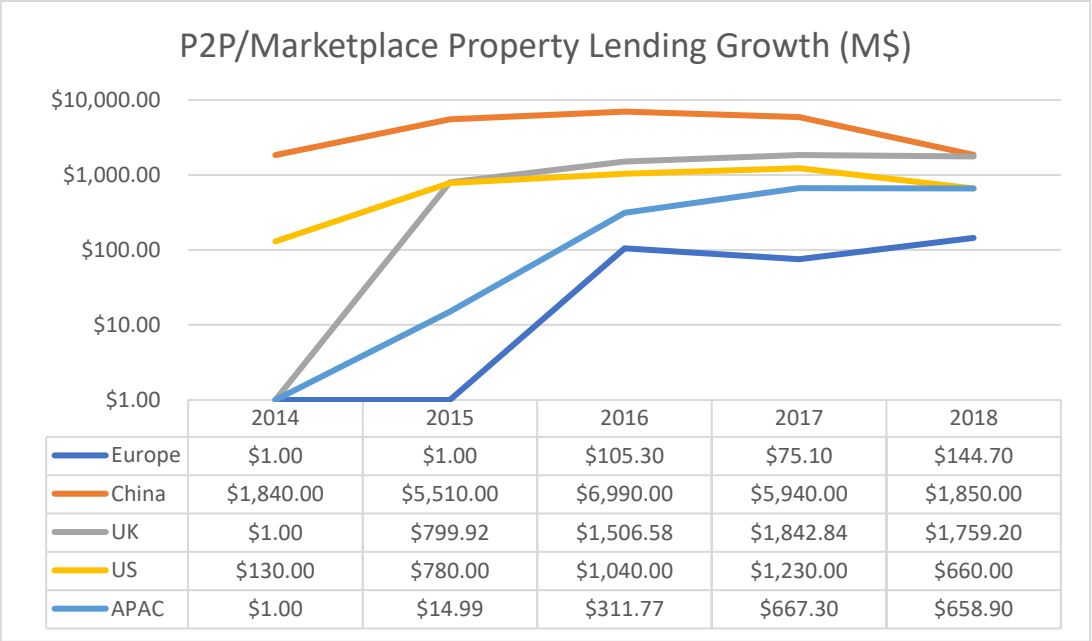
Figure 4 represents the growth of P2P business lending. The graph shows overtime growth in all representative countries. Not as much growth compared to consumer lending. What is interesting in this graph is that Chinas volume increased a lot until 2018 its P2P business lending was almost cut in half. Also, US market increased first by more that 2-fold in 2014-2015 but then it decreased almost by half in 2016. But since then, it has increased overtime. This means that even businesses use P2P lending more and more as a financing option.




| | 2014 | 2015 | 2016 | 2017 | 2018 |
|---|---|---|---|---|---|
| Europe | $123.70 | $235.40 | $287.50 | $526.20 | $996.80 |
| China | $8,040.00 | $39,630.00 | $57,780.00 | $97,430.00 | $42,740.00 |
| UK | $983.91 | $1,157.19 | $1,618.23 | $1,842.84 | $2,541.90 |
| US | $980.00 | $2,580.00 | $1,330.00 | $1,450.00 | $2,030.00 |
| APAC | $114.49 | $363.34 | $333.62 | $623.30 | $1,772.60 |

Figure 4. P2P Business Lending growth

Figure 5 represents P2P property lending. In this figure, the data needed to be altered a little bit. Values of 1 were originally zero. This alteration was done because the graph could not draw lines if values were zero. So, with value number one the increase of property lending can be illustrated better. This figure shows that P2P property lending is increasing overtime as well except for the last few years. All major markets, except Europe, decreased in 2017-2018. This slump might be explained because of the previously mentioned issue. Majority of people do not want to risk their houses for a loan. The idea of P2P lending is that you get a loan, no matter what. Yes, the interest rates are higher, but banks usually require some sort of collateral. In this form the security is property. This becomes a lot like banks loan. So, probably only people who are in very deep financial trouble would apply for this loan to lower the interest rate of the loan.




| | 2014 | 2015 | 2016 | 2017 | 2018 |
|---|---|---|---|---|---|
| Europe | $1.00 | $1.00 | $105.30 | $75.10 | $144.70 |
| China | $1,840.00 | $5,510.00 | $6,990.00 | $5,940.00 | $1,850.00 |
| UK | $1.00 | $799.92 | $1,506.58 | $1,842.84 | $1,759.20 |
| US | $130.00 | $780.00 | $1,040.00 | $1,230.00 | $660.00 |
| APAC | $1.00 | $14.99 | $311.77 | $667.30 | $658.90 |

Figure 5. P2P property lending growth

Lastly in this chapter, the total P2P lending volume is being examined in Figure 6. This figure shows the total volumes of all P2P financing forms in 2018. China is by far the biggest P2P lending market in the world with a lending volume of $ 208 billion. This amount is 83 % of all P2P lending in the world. Second biggest market is US with a volume of $ 28 billion. Third biggest market is in UK which has a volume of $ 6.3 billion. Also, Europe and Pacific Asia has noteworthy amounts $ 4 billion and $ 3,4 billion respectively. As this figure shows, there are markets that already has a lot of activity in P2P lending market. But there are many markets that do not have P2P in such large volumes. This means that P2P Lending has so much room to grow globally. Also, Europe has not grown to its fullest potential. UK has more lending volume than whole of Europe combined. This means that in future days to come, P2P lending will become a more popular financing option in Europe.

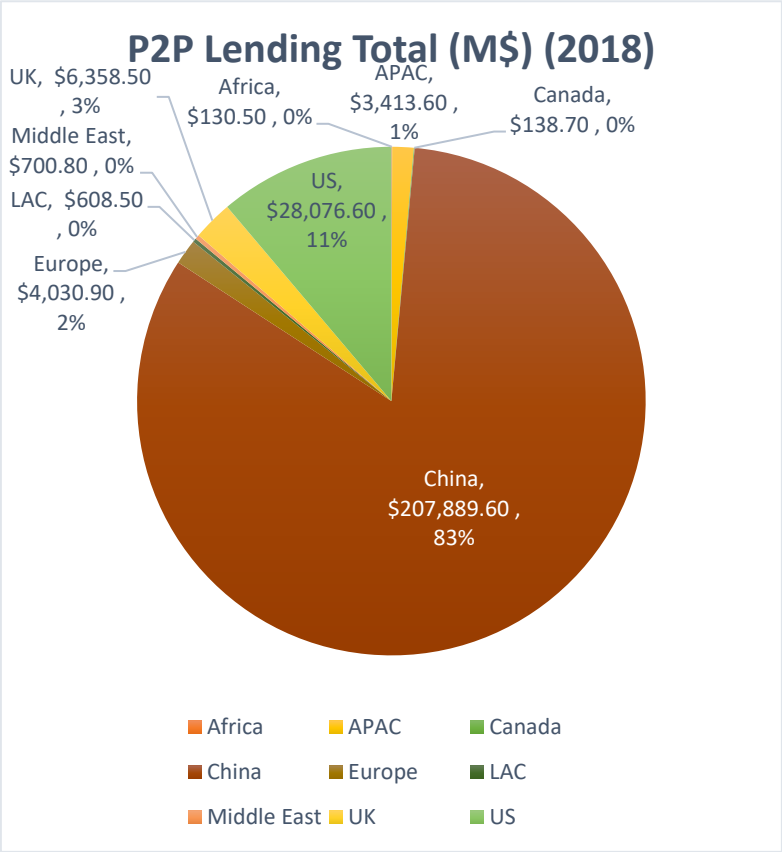


Figure 6. P2P lending in total globally

To conclude this chapter, P2P lending is a growing form of financing. It is already a large form in couple of countries, but it has so much potential to grow in other ones. Consumer loans are the most popular choice currently. To make P2P lending reliable and more transparent, a lot of research is needed for both investors and borrower's sake.

## 2.3 Pros and Cons of P2P lending

P2P lending has a lot of potential to be one of the most used way to finance investments. There is a reason why P2P lending has increased globally. But everything might not be as it seems. P2P lending also has many hazards that can cause serious financial damage to investors. Online P2P platforms often claim that they are beneficial for both borrowers and investors by eliminating expensive intermediaries and reducing transaction costs (Klafft 2008). This chapter aims to bring pros and cons of P2P lending alight.

### 2.3.1 Pros of P2P Lending

The motivation behind building P2P platforms was to go around intermediary banks. This circumvention has multiple advantages. Expensive middleman is replaced by a more cost-effective online platform, which reduces transaction costs. P2P platforms are so much more cost-effective because they are not so administrative and hierarchic overloaded like banks (Pokorná & Sponer 2016). P2P operators do not work under bank regulations since they pass on the risk to investors by passing on the credit and liquidity risk. (Davis & Murphy 2016) Furthermore, borrowers are given the chance to present their loan case in much detail. This provides investors new information that banks do not have because they have standardized decision processes that usually does not take into consideration of additional information. What is more, all bids for the loan are visible and traceable online. This means that the loan generation process is very transparent and creates a feeling of fairness. Finally, the loans are said to create higher returns than traditional bank savings and to be cheaper for borrowers. (Klafft 2008)

The main advantage of P2P lending is that a borrower can get a loan at a lower rate compared to traditional bank and without collateral, while lender can get higher return on investment. Collateral makes the lending decisions hard in traditional banking but in P2P lending's flexibility makes it an easy alternative. (Pokorná & Sponer 2016)

Since P2P lending has a lot of data available, using different decision models that relies on information of borrowers can significantly increase profits. So, with IT techniques like big data analysis, prediction models, credit audition and data mining can decrease risks in P2P lending. (Wang et al. 2015)

### 2.3.2 Cons of P2P Lending

P2P lending might not be as good for investors as claimed. From lenders perspective, it is very difficult to judge the quality of the deal beforehand because lenders have the default risk and few of them are experts in risk management. Moreover, pseudonymous environments are usually riddled with information asymmetries, which makes it easy for opportunistic borrowers to exploit lenders. (Klafft 2008)

Also, there is a possibility of misrepresentation for borrowers when considering their creditworthiness. Most of the requested loans are from people that could not get a loan from a bank. Investors may not understand this and rely, to some degree, on P2P platforms risk assessment, which is a good tool to use but not completely accurate (Davis & Murphy 2016). If the borrower cannot pay back the loan in time, there is a chance that the full amount of the loan will not be recovered. Furthermore, P2P platforms are regulated as “Providers of small sized payment services” which means that they do not have an obligation to contribute to “fund of deposit’s insurance” which means that investors do not have their investments insured. (Pokorná & Sponer 2016)

P2P operators perform a function like credit rating agencies. They create a model which calculates borrowers credit score which indicates loans performance. These credit rating models might not have as good quality as other rating agencies. This is a problem since some investors might rely on this metric when deciding whether to invest or not. (Davis & Murphy 2016)

P2P investments are also largely illiquid. The maturity of matching borrowers and lenders is long and if there are no secondary markets on these loans, the maturity of the loan increases the illiquidity. Some P2P providers do have secondary markets in place and the information flow is transparent since secondary market buyers see how the borrower has performed in the payments of the loan. (Davies & Murphy 2016)

Investors face the risk of P2P operator ceasing operations due to unprofitability or platform software failure. In this case, question arises how the assets will be managed once this agency risk manifests. One possibility is to transfer the loan book, repayments, and all, to another provider under the direction of an administrator or liquidator. This case would most likely involve significant losses for the investor. (Davis & Murphy 2016)

Even if investors understand the risks involved in P2P lending, the question of what rate of return they should expect arises. P2P investment is roughly the same as holding both equity and deposits in a depository institution specialising in same kind of loans. Considering this, investors required rate of return should be about the same as weighted average cost of funds

of a similar depository institution. This highlights the fact that P2P operators need comparative operating cost and risk assessment abilities to succeed in the long run. (Davis & Murphy 2016)

Studies have shown that in online business players exhibit herding behaviour when facing risk of uncertainty such as information asymmetry. Online platforms are destined to have herding behaviour because of two reasons. One, information overload. There is so much information on the internet, so users have difficulty to understand and use all information available. This leads to a situation where people do not have any idea where to invest money and then end up following some “experts” blindly. Second, people can easily follow others’ choices in P2P lending. They see that some loan has many bids, which can cause flawed thinking, “others seem to think that this is a good loan so it must be.”. If everyone is bidding on a loan, it does not mean that the loan will perform well. (Pokorná & Sponer 2016)

## 2.4 Credit Risk Management

As we can see, P2P loans have a lot of uncertainty bound to them. Risks in previous section seem to overwhelm the pros in lending. This means that managing the credit risk is a very important aspect of P2P lending.

The problem arises when inspecting the individuals baring the risk. In a bank, the credit risk is assumed by the bank itself. So, the bank has a great motivation to build a system that minimises credit risk to increase profits. Banks have multiple expert departments to handle credit risk assessment and the expertise is top notch. On the other hand, P2P providers are not the ones that have credit risk, it is the investors. This means that compared to banks, their credit assessment might not be as accurate. Furthermore, P2P providers’ credit scoring models do not have the same data of borrowers as banks do, such as account transaction data, financial data, and credit bureau data. For these reasons, the credit assessment might be poor. On the other hand, P2P services do provide some data of borrowers with continuous networking activities. By using this data and different mathematical models, it is possible to improve credit risk assessment accuracy significantly and make P2P lending as a viable choice for investors. (Agosto et al. 2019)

Credit scores used by P2P platforms provide important information about the borrower and is one of the most important variables when considering the creditworthiness of borrowers. Serrano-Cinca et al. (2015) Determined in their study that subgrade assigned by the P2P lending site, based on FICO score among other attributes, is the most important variable. These grades predicted defaults with accuracy of 62.0-80.6 %. This helps to reduce the information asymmetry which is very much present in P2P lending. On the other hand, the investor should not only use this score because the prediction power could be even better, and this can be achieved through individual analysis if one has expertise to do proper analysis. For example, Pan and Zhou (2019) managed to increase the prediction accuracy to 98.63 % using random forest and visual graph model. Cai and Zhang (2020) used data mining techniques and then logistic regression model to achieve accuracy of 86 %. Agosto et al. (2019) used spatial regression models to generated default prediction accuracy of 80 %. As we can see from these studies, using mathematical models to predict accuracy can increase the correct prediction of defaults. This means that through careful analysis, one can achieve higher returns in P2P lending since less loans tend to default with good models.

These studies showcase the importance of default prediction models. This is also the motivation of this study. To get a good default prediction and then prepare a country wise comparison. In the next chapter machine learning is examined briefly, what it is and how it can be utilized in P2P lending.

# 3 MACHINE LEARNING

Machine learning is essential part of this thesis. It is being used to create predictive models that can accurately predict default outcome of the P2P borrowers. This is very important from lenders perspective since earlier it was described that P2P lending tends to have risky borrower behaviour and any tool that can alleviate that risk, is more than welcome.

Machine learning can be defined as a branch of artificial intelligence. Using computing, it is possible to create systems from the data that can learn and improve with experience. These systems can predict outcomes that would be way too much to handle for a human mind. There are number of different learning algorithms that can be used for prediction purposes. The required output determines which kind of algorithm to use. Machine learning algorithms fall into two different categories, supervised and unsupervised machine learning. (Bell 2020, 3)

Supervised machine learning refers to a labelled training data. Supervised learning is used to assign correct labels for given sample. Input data of supervised learning model is very important since for the classifier to make sense of the samples, it needs a lot of input data of labels and their properties to make accurate decisions. This input data is manually inserted for the algorithm which makes it supervised learning method. This input data is used to train learning models which later can be applied on unknown data. This will result in predictions of rather good accuracy if the model is trained properly. (Bell 2020, 3)

Unsupervised machine learning is on the opposite side of the spectrum. Here, the algorithm will find, by itself, a hidden pattern in a load of data. With this method, there is no right or wrong answer. In this case the algorithm is just run on a data, and it will return some pattern or outcome which might not be expected. Unsupervised learning is more like data mining than actual learning. (Bell 2020, 4)

Machine learning algorithms cannot function without a human touch though. All models and algorithms need to be built using methods that give them the best outcome. All it needs is a human to get it started, but once all the requirements are in place, machine learning have the capabilities to predict even most complicated cases. (Bell 2020, 4)

This thesis is done with supervised machine learning. Later in this thesis, models are trained using vast amounts of data of defaulting borrowers and their properties. Machine learning algorithms will then learn to identify these defaulting characteristics and when to assign a default-label. When the model is trained, it can be applied on unknown data, and it will provide classification with certain accuracy. All these pre-processing methods and algorithms will be explained in more detail at chapter 5. In the next chapter literature review is being conducted. Literature review gives this study a better understanding of different prediction modelling techniques used and hopefully better results.

# 4 LITERATURE REVIEW

In this chapter, literature review is conducted to get a more comprehensive view of the subject. Rowley and Slack (2004) identifies six reasons why literature review is important. These reasons are:

1. It supports the identification of a research topic, question, or hypothesis
2. It helps to identify literature which the research will contribute and contextualizing the research within that literature
3. It helps to build an understanding of theoretical concepts and terminology
4. It facilitates a list of sources that are being used
5. It suggests research methods that might be useful
6. It helps analysing and interpreting results.

Since literature review plays a very important role in research it should be constructed the best way possible. Callahan (2014) has recognized from literature review research five distinctive characteristics to showcase a rigorous literature review. These aspects are called five C's. Meaning literature review should be concise, clear, critical, convincing and contributive. *Concise* means that review should be concise synthesis of a broad array of literature on the topic. *Clear* means clarity of the data from articles that creates the foundation of literature review. The methods used and research outcomes need to be reported so that correct view can be achieved. *Critical* means that rigorous literature review include critical reflection and critical analysis of each research article. *Convincing* means that after analysing data, a convincing argument must be developed. So, findings of the research need to be presented to make a convincing case of research. *Contributive* means that literature review needs to contribute to the body of research. Using these key characteristics, the literature review can be developed correctly and rigorously. It is important to develop this part well since it also helps to build knowledge of used methods. This way deciding on methods in this research is much easier and best methods can be found and used. In the next subchapter, the search process is defined and created.

## 4.1 Search Process

Conducting search process appropriately is also very important in literature review to optimise the number of key sources. Timmins and McCabe (2004) suggest some principles of search strategy in the following way. Outline the stages in the search process, this will be explained in more detail later. Keep a record of databases and keywords used in the search. Use a table format to identify databases included, number of references found, and the final number of references used in the review. Document reasons for excluding some sources. Identify the type of literature sourced, for example qualitative or quantitative studies, surveys, descriptive, reports etc. And finally, keep a record of key references included.

As mentioned before, stages of the search process need to be defined properly in order to have a rigorous literature review. To conduct the search process properly, two search processes were synthesised in Figure 7 (Timmins and McCabe 2004; Webster and Watson 2002).

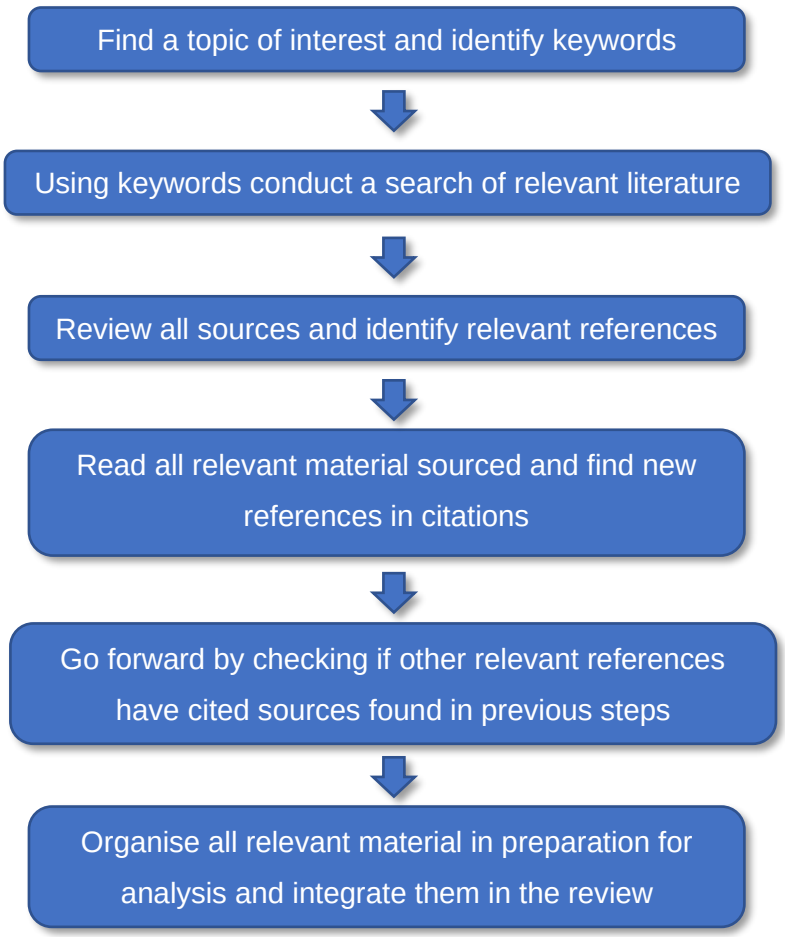



Figure 7. Illustration of proper search process

Now that the proper way to conduct literature review is defined, the process can begin. First, databases used are being addressed. This study uses LUT Primo search engine which combines many research databases into one.

Next step is to identify keywords. Obvious choices for this topic are “P2P lending” or its synonyms “peer-to-peer lending” and “social lending”. With these words one can find studies related to P2P lending. To include articles that involve general credit risk, search words “credit risk”, “credit management”, and “credit scor*” were used. But the point is also finding studies that has used machine learning based prediction methods. So, to incorporate these findings in search, words “machine learning”, “predict*” are used in every search. Keywords “predict*” and “credit scor*” were formed by using truncation. This means that search engine will use this letter sequence but also include letters that follow. For example, it will search for “prediction” and “predicting” as well. This is a very handy tool since words have many forms and this is the way to catch them all.

So, the search is conducted in two parts. First part includes P2P lending words mentioned above and default, predict, and machine learning. Second part includes credit words mentioned above and default, predict, and machine learning. This part was trickier since the amount of studies in credit prediction in general is vast. So, to tackle this problem, search was conducted only for titles. Using these methods, the results will only include studies in fields of P2P lending and credit management but also include only studies that have used machine learning based default prediction which is the goal of this study. Also, additional filters were introduced. These were peer-reviewed articles, studies after 2005 to ensure recency in research, availability online, and language is English.

Initial searches resulted in 257 articles. Next step is to start scanning these articles by reading titles and abstracts to see if the topic is relevant for this research. Topics that considered P2P in general were removed since that has already been examined. Topics that consider default prediction in other fields such as retail or bankruptcy of companies were excluded since the focus is on consumer loans. Topics that used text descriptions of loans to predict default were excluded because this study does not use text descriptions of borrowers. After scanning titles, abstracts and overall research, number of references is 27.

Next step is to conduct backward tracking of these references. Every research is scanned, and their reference list is thoroughly checked to see if other essential articles can be found. These articles were found using Google Scholar or links provided. Some of the articles found were not available online so they were excluded. Backward tracking resulted in total of 38 articles which means that method yielded 38 – 27 = 11 new articles. These articles are relevant and provided new information and insight for this thesis.

Next step is to do forward tracking. This means that articles that have referenced these current articles will be revealed. This method can be applied by using Google Scholar. There is an option to click “cited by” in Scholar. This feature helps to track all articles that have referenced current articles. There are a vast number of articles that have cited current references and number of references is already 38 in this thesis so additional articles need to have very important information to be included. Also, only articles that were cited over 50 times were included to ease the search of relevant articles. Forward tracking resulted in 42 – 38 = 4 new articles. In total, 42 articles were found.

Now that all the references are gathered in a proper manner, all relevant material are organized, and summary of all articles’ key points are gathered. This resulted in three distinguishable groups of articles, Credit scoring and credit management in general, Determinants of default and Predicting default in P2P lending. In the next three chapters all these topics are unfolded. Furthermore, few articles were removed from the list since they did not provide more information or was not related enough for this thesis. So, total number of articles being examined is now 37.

## 4.2 Credit scoring and credit management in general using machine learning

Credit analysis was born in the beginning of commerce with borrowing and lending. However, modern credit scoring system started to develop 70 years ago since Durand (1941) first realized the potential of credit data. Since then, traders have been gathering information on the applicants for credit and cataloguing purposes to decide whether to lend or not to borrowers. (Louzada et al. 2016)

Thomas et al. (2002) defines credit scoring as “a set of decision models and their underlying techniques that aid credit lenders in the granting of credit”. Nowadays, this definition has become a bit broader. Louzada et al. (2016) states that “credit scoring is a numerical expression based on a level analysis of customer credit worthiness, a helpful tool for assessment and prevention of default risk, an important method in credit risk evaluation, and an active research area in financial risk management.” This definition is more accurate since via credit scoring, it is possible to do so much more than just decide on whether to lend or not. With credit scoring it is possible to calculate, for example, expected profits beforehand which will help banks and investors to make more profit.

At the same time, data mining techniques started to develop. With increasing computational power, it was now possible to calculate predictions and expected individually for each customer based on her/his characteristics. This began a giant leap in credit scoring and lending. (Louzada et al. 2016) Table 1. Consists of a collection of different articles that have more insight in matter of credit scoring and using data mining and machine learning techniques to predict default. This table has rough descriptions of the objective and used prediction models. Also, I included in the table whether the data is balanced or not, meaning if it has as many defaulters as creditworthy borrowers. Next, articles in this table are broken down and explained to get a good image of credit scoring and machine learning in general. Also, many models are compared using AUC (Area under ROC curve) score. This metric compares correctly classified samples against falsely classified ones. So, if AUC is 70 % this means that 70 % was correctly classified and 30 % falsely classified. This metric is explained further in methodology chapter later.

Table 1. Credit scoring and machine learning articles

| | Objective | Prediction models | | | | | | | | | | Balanced data |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **Author(s) and year** | | LR | RF | NN | LDA | **SVM** | k-NN | QDA | NB | DT | Other | |
| Boughaci et al. (2020) | Examines whether clustering or segmentation is a good method in credit scoring. | | x | | | | | | | | k-means + RF | x |
| Brown & Mues (2012) | This paper works with imbalanced data and solves that problem. It also studies how different kinds of balances change the predict results. It uses many algorithms to predict default. | x | x | x | x | **x** | x | x | | x | Gboost | x |
| Dastile et al. (2020) | This is a systematic literature review that explores best statistical and machine learning models in credit scoring. This paper provides information of all the necessary steps and best methods. | | | | | | | | | | Literature review | |
| Harris (2013) | This paper uses support vector machine (SVM) based credit-scoring models and compares Broad (less than 90 days past due) and Narrow (greater than 90 days past due) default definitions. | | | | | **x** | | | | | | x |
| Keramati & Yousefi (2011) | The aim of this study is to provide a comprehensive literature review of applied data mining techniques in credit scoring context | | | | | | | | | | Literature review | |
| Kruppa et al. (2013) | This study focuses on default probability rather than classification in consumer credit scoring. So by using machine learning methods, they estimate default probabilities. | x | x | | | | x | | | | PETs, bNN, LR tuned | |
| Lessmann et al. (2015) | The aim of this research is to compare several novel classification algorithms to state-of-the-art models in credit scoring. | x | x | x | x | **x** | x | x | x | x | Total 41 models | |
| Louzada et al. (2016) | This research aims to present a systematic literature review on theory and application of binary classification techniques in credit scoring. | | | | | | | | | | Literature review | |
| Luo et al. (2009) | Credit scoring problem. Using machine learning methods to predict default | | | | | **x** | | | | | CLC | x |
| Pławiak et al. (2019) | A novel deep genetic cascade ensemble of SVM classifiers, DGCEC technique is proposed to predict the Australian credit scoring. | | | | | **x** | | | | | DGCEC | x |
| Pławiak et al. (2020) | A novel DGHNL credit score prediction model is proposed. | | | | | | x | | | | DGHNL, Fuzzy system | |
| Trivedi (2020) | This is paper studies credit scoring with different machine learning models and compares their predictive performance. | | x | | | **x** | | | x | x | Bayesian classifier | |
| van Thiel & van Raaij (2019) | This paper contains research from UK and Netherlands and examines to what extent can individual lender advance their credit decisions with risk assesment AI. | x | x | x | | | | | | | | |
| Yu et al. (2010) | A four-stage SVM based multiagent ensemble learning approach is proposed for credit risk evaluation. | | | x | x | **x** | | x | | | MV, TA, ALNN | x |
| Count | 14 | 4 | 6 | 4 | 3 | **7** | 4 | 3 | 2 | 3 | | 6 |

Objectives of these articles stay somewhat the same. All aim to find better prediction models and methods for credit scoring purposes. The most used model in credit scoring from these articles is support vector machine (SVM). Second place goes to random forests (RF) method. There are three methods in third place, logistic regression (LR), neural networks (NN) and k-NN (k-nearest neighbour). Louzada et al. (2016) literature review has similar results in terms of popularity. NN is the most used, second is SVM and third is LR from single method models. So, NN only switched from third place to first. Also, hybrid models and combined models are very popular in literature (second and third place in terms of popularity, after NN). These kinds of models tend to outperform traditional models, but traditional single model methods offer a competitive comparison. Keramati & Yousefi (2011) found in their literature review that SVM is the most popular model in recent years which is in line with my findings. Louzada et al. (2016) also found that SVM method provides best predictive performance. This means that in this thesis I should consider using SVM as one of the models. NN could also be considered but I do not have previous experience using NN model.

LR is considered as the industry standard model in credit scoring according to Lessmann et al. (2015). Many models, like ensemble, RF and ANN, perform significantly better than LR. Thus, they argue that LR should not be used as the benchmark model for new models since it does not require that much improvement in prediction performance to outdo LR. They suggest that RF should be used instead since it is easy to use and produce better prediction

performance and new models have much harder time to outperform this one which raises the bar for new publications. This thesis does not propose a new model but compares prediction performance between countries. Thus, the most common and simple models should be used to get reliable results in performance differences. Also, Keramati & Yousefi (2011) point out that LR still provides competitive performance, so it is not as bad as Lessmann et al. (2015) leads on in their results. But I agree that for new model proposals, LR is not good enough benchmark.

So according to these two literature reviews (Keramati & Yousefi 2011; Louzada et al. 2016) and one vast study of classifiers (Lessmann et al. 2015) most popular and simple models to use are SVM, RF and LR. These models provide good prediction accuracy and are simple to perform. Which is just what I need for country comparison. But more possible models are presented in other studies, and this is not tested with P2P data, but normal credit scoring data. So, we will see which models are used at the end of literature review in the summary part.

There are only 6/14 articles that use balanced data in the prediction. Dastile et al. (2020) also came to this conclusion that most of the studies do not implement balancing. Without balancing the data is biased towards the majority class. Furthermore, Brown & Mues (2012) showcased in their study that AUC scores tend to decrease if imbalance is present in the data. Also, they found that even in the presence of class imbalance, RF and GB models performed well. LR and LDA remained competitive as well. SVM does not perform well in the presence of class imbalance. To tackle imbalance problem Harris (2013) tested broad (less than 90 days due) vs. narrow (more than 90 past due) default definition using only SVM method. Harris found that using broad definition results in better AUC and accuracy in SVM models. One possible explanation is that the algorithm gets fed more with default applicants, so it has a better understanding of the characteristics and patterns of a defaulter. Best SVM model was achieved when broad definition plus random under-sampling (RUS) was used. So, to get balanced model, I should consider using broad definition of default as well as sampling techniques. More detailed research of sampling techniques will be introduced in chapter 3.4.

Feature selection is also an important part of credit scoring. With this technique, excess variables are filtered out of the data which do not provide any prediction performance improvement. Trivedi (2020) studied different feature selection techniques effects of prediction performance. Conclusion was that different feature selection techniques, Chi-square, gain-

ratio, and info-gain, did not provide that much different results in terms of accuracy in SVM, Bayesian and NB models. However, in decision tree and random forests models, chi-square technique provided better accuracy compared to other methods. According to this study, chi-square method is the best option to use.

Validation of the model needs to be done correctly so that the results will be valid. Louzada et al. (2016) found in their literature review that k-fold cross validation is the most used method in literature. Almost all studies in Table 1 use k-fold cross validation, excluding literature reviews, Kruppa et al. (2013), and van Thiel & van Raaij (2019), which totals of 9 studies. Both articles use the holdout method which means that about 70 % of the data is used to train the model and 30 % is used to validate it. This method is second most used according to Louzada et al. (2016). But since k-fold cross validation is by far the most used method according to Louzada et al. (2016) and collection of studies in Table 1, I should use this validation method.

Misclassification criteria needs to be created as well so the actual metrics and performance numbers can be evaluated properly. Louzada et al. (2016) finds in their literature review that metrics in confusion matrix is the most used method, second used is ROC curve (or AUC) and third place uses both methods. Other methods are not that much used in literature. Table 1 studies used metrics in different forms. Yu et al. (2010), Trivedi (2020), Pławiak et al. (2019) only uses metrics in confusion matrix form. Luo et al. (2009) uses only accuracy as a metric, which makes the results quite unreliable since accuracy only considers correctly classified samples. Kruppa et al. (2013) uses AUC (Area under ROC curve) and BS (Brier scores). Brown & Mues (2012) uses only AUC measure, which is better than accuracy since it also accounts for misclassified samples. Dastile et al. (2020) proposes that the following evaluation metrics should be used: accuracy, AUC, G-mean, recall and F-measure. Accuracy and AUC are the most used in literature. G-mean, recall and F-measure are great metrics to be used in imbalanced data. So, these should be the metrics I use in this thesis. Theoretical framework is now taking its form and already methods and metrics to use have been somewhat established. In chapter 3.5 the framework will be presented in figure form as the P2P part is also included in the framework.

Next, studies that use more complex methods, like combination of methods and hybrids in Table 1, are being examined. These studies provide insight that models can be made very predictive, so that default risk can be minimized as much as possible. Even though Lessmann

et al. (2015) say that progress in probability of default topic is stalling, new models still emerge and provide very good accuracies and AUC scores. Pławiak et al. (2019) states that even a fraction of increase in these scores mean a great deal of savings for banks. P2P lending might get even more use out of these complex models since individual investors are the ones that have a lot of money in line. Only problem is that usually individual investors do not have the necessary knowhow to implement these methods.

First article is from Boughaci et al. (2020). Their goal was to examine whether clustering enhances classification prediction. Six different datasets were used to get reliable results. They used k-means to group data into two clusters. They found out that clustering before the classification increases predictive performance significantly. For example, in a Japanese dataset, they predicted with RF reaching AUC score of 92,7 % but with RF and k-means clustering they managed to get 99,7 % score. Japanese dataset was also balanced so there is no bias towards the majority. Also, they used 10-fold cross-validation which ensures that the evaluations are valid. Furthermore, they used t-test to check if these increases in prediction performance are significant and they certainly were. This increase is very significant with relatively simple methods. This research may present the answer to many investors who want to lower credit risk significantly.

Second article is from, Kruppa et al. (2018). Their focus is probability of default rather than correctly classifying borrowers. They found out that RF-PET (random forest probability estimation tree) model outperformed other ones with AUC score of 95,9 %. Logistic regression, which is the most popular model of them all, only achieved AUC score of 77,9 %. The data was not balanced so the results could be biased towards the majority. They trained the data with random split of 2/3 as training and 1/3 testing which is a good way of training models. This model provides very significant improvement compared to traditional models. Although they compared RF-PET model to logistic regression and other weak models, AUC score was still high. But there might be issues with the data balancing and biases. So, these improvements in the model are somewhat questionable. If balancing was implemented, AUC score might drop a little, but I expect the difference to be small since Brown & Mues (2012) showed that RF performs quite well even in the presence of imbalance. Comparison with just RF would have been good as well since it is one of the best benchmark models to have according to Lessmann et al. (2015).

Next research is from Luo et al. (2009). They used CLC (Clustering-launched classification) and SVM. They used Australian data set which is balanced. In this data they managed to accuracy of 86,52 % using CLC and 80,43 % using SVM. So, there is a significant increase in accuracy, but this metric is not perfect since it does not account for the misclassified portion at all. AUC score would have been a better metric in this case. They used 10-fold cross-validation which is a good method. This research is okay but as said before, it lacks the proper metrics which makes the results questionable. Also, there was no confusion matrix, so it is impossible to tell what the model actually predicted. It should be included in every research regarding machine learning prediction according to Louzada et al. (2016). AUC score would have been also good. All in all, this research does not use proper metrics. Thus, the results are not reliable.

Next article is from Pławiak et al. (2019). They propose a novel deep genetic cascade ensemble of SVM classifiers model (DGCEC). This model is a deep learning model that learns the data in multiple layers. In each layer the model learns to recognize good and bad borrowing behaviour which increases its prediction performance. This research also uses balanced Australian dataset, so it is not biased. Also, feature selection was implemented using genetic algorithm. Furthermore, 10-fold cross validation was used. This research is properly done, and the results should be reliable. DGCEC model achieved prediction accuracy of 97,39 % which is very high. They also included multiple SVM-based models which had accuracy of 88 % approximately. They used confusion matrix to analyse prediction performance so on can also see falsely classified samples because accuracy does not include falsely classified samples. Only 18 samples were falsely classified so this model seems to perform very well. According to Pławiak et al. (2019), this model was the best in current literature to predict default. This shows that with good knowhow, it is possible to create models that predict default with very high accuracy. Although, this might be too complicated for individual investors in P2P, experienced individuals could use this model to their advantage and gain superior profits.

Fifth article is also from Pławiak et al. (2020). In this article they propose a new deep genetic hierarchical network of learners (DGHNL) for prediction of credit scoring. This model is a hybrid of other models Fuzzy system, kNN (k-nearest neighbour), PNN (probabilistic neural network), nu-SVC, and C-SVC (SVC models are types of SVM models). This time they use German dataset which has 70/30 ratio of good and bad borrowers, respectively. This data is not completely balanced but still it has a good ratio between good and bad borrowers. Genetic algorithm was used as feature selection method. Also, 10-fold stratified cross-validation was

used. This research seems to be performed in orderly manner, so results are reliable. The new hybrid model achieved AUC value of 91,38 % which is very good. Maybe a bit better AUC value could have been possible with more balanced data. In confusion matrix one can notice that bad borrower misclassification was 5 %. In their previous model in Pławiak et al. (2019), the model performed better and had misclassification of 1,4 % in bad borrowers which is much better compared to this study. This further demonstrates that good prediction performance can be achieved through complex and advanced methods, but balanced data can train the model to better notice patterns in bad borrowers.

Final article using advanced techniques is from Yu et al. (2010). They proposed a four-stage SVM based multiagent ensemble learning approach for credit risk evaluation. This study uses British credit card application data which is then balanced to have 50/50 ratio of good and bad borrowers. No feature selection was used in this study. ALNN multi agent model achieved the best total accuracy of all other used models which was 71,19 %. Even though they did not use AUC, this study reports type 1, and type 2 accuracies so one can determine the misclassification from there. This model did not exceed previously mentioned research and is the poorest of them all. Still, it outperforms single agent system models such as SVM (6 % better) and LR (7 % better). Maybe if feature selection methods were implemented, the accuracy could have been better.

One more article is being examined in this part. This article is from van Thiel & van Raaij (2019) and it examines whether it is possible to predict default using machine learning methods. They did not use balanced data, but they used three different datasets, two of them were from different countries. They managed to find that most predictive features that models produced, have high levels of similarities between different datasets. So, this study suggests that default is a sum of similar features which are mostly income or spending related and it stays the same even if other country or lending purpose (mortgage vs credit) is examined. Compared to this study I can use one dataset and maybe find slight differences between countries. Since different countries are in one dataset, it is possible to examine differences between countries validly because the variables are measured in the same way for everyone. I need to look for differences in variables or AUC values in each country when all models are completed.

Now that all articles from credit scoring techniques in general have been thoroughly examined, the same techniques in P2P lending can be examined more closely. Key takeaways from this

part are summarised in chapter 3.5. This chapter has helped a lot in theoretical framework building and it shows in summary chapter.

## 4.3 Determinants of default in P2P lending

The point of this chapter is to examine what variables are important when constructing models. This helps in feature selection process and helps to justify the selections for included features. After the determination of important variables or features, models can be constructed knowing that the features selected are important and in line with literature. Table 2 consists of these articles that examine important variables in P2P lending and default prediction. All these articles cover, at least to some extent, important variables that help in determination of default. Articles have four Chinese datasets, two American datasets, an online survey and one Indonesian dataset. The models used here are very simple. Logistic regression is the most popularly used in variable selection among these articles. Next, all these articles are broken down in detail and examined for important determinants of default.

Table 2. Determinants of default in P2P lending articles

| Author(s) and year | Objective | Data | Statistical model(s) |
|---|---|---|---|
| Chen et al. (2019) | This paper examines data from chinese P2P platform an assesses probability of default as well as significant impact variables. | Ppdai.com | Variable selection methods: Stepwise, LASSO, Bayesian variable selection. Default prediction: Logistic quantile regression. |
| Jagtiani & Julapa (2019) | The use of alternative data and machine learning in P2P lending. Predictive models to predict default probability | Lending Club and traditional bank loans | Logistic regression and ROC curve. |
| Li et al. (2018) | The purpose of this paper is to examine the mechanism how the platform uses undisclosed information to determine borrowers credit rating. Also, this study examines the effectiveness of credit scoring in default prediction. | Renrendai.com | OLS regression for variable selection and Logistic regression for default prediction. |
| Lin et al. (2017) | The point of this study to explore factors that determine the default risk based on the demographic characteristics of borrowers. | Yooli.com | Nonparametric tests for variable difference between good and bad borrowers. Logistic regression for default prediction. |
| Santoso et al. (2020) | The idea of this study is to investigate the determinants platform interest rate and borrowers' default status. | Indonesia Financial Services Authority (three major P2P platforms) | OLS and Logistic regression |
| Serrano-Cinca et al. (2015) | This paper studies P2P lending and the variables that explain loan default. | Lending Club | Unvariate means test and survival analysis. Logistic regression for default prediction. |
| Wang et al. (2020) | This study aims to evaluate credit risk of borrowers with psychological variables. The point is identify borrowers' default behaviour. | Online survey | General Strain Theory |
| Wu & Zhang (2020) | This study focuses on credit ratings and their reliability in P2P lending. | Renrendai.com | Logistic regression |
| Count | 8 | | |

First article is from Chen et al. (2019). They examine Chinese P2P platform and tries to predict default as well as identify significant variables. They used feature selection techniques such as stepwise method, LASSO and Bayesian variable selection. They concluded that Bayesian variable selection method provided the best quality variables. Furthermore, they identify important variables which will have a significant impact on default probability. These variables

are longer loan period, interest due, interest rate, loan type (meaning which way the loan is received) and regulation change. These significant variables make sense since almost all of them indicate to risky borrower behaviour except loan type and regulation change. Loan type is an interesting significant variable, and it shows that if loan is received via app, the bigger probability to default there is.

Second article is from Jagtiani & Julapa (2019). They use alternative data and machine learning to predict default and, they examine if FICO scores, and rating grades provided by the platforms correlate with each other. They found out that alternative data sources have allowed borrowers with fewer or inaccurate credit records to have access to credit in P2P lending. FICO score and P2P platform's credit ratings had a strong correlation of 80 % in 2007 but as years went by, the correlation has declined to as small as 35 % in 2015. This means that the use of alternative data has provided borrowers better ratings and FICO score has a smaller impact on the ratings these days. But even in these cases when correlation to FICO score and ratings were low, the ratings provided by Lending Club were accurate in predicting future loan delinquency. This study shows that even if FICO score is low, ratings provided by Lending Club are accurate and can be used in decision making. Still, I would do own analysis of variables to determine the best borrowers and not simply trust the rating.

Third article is from Li et al. (2018). The purpose of this study is to examine the mechanics of borrower credit ratings provided by the platform. They want to find how renrendai.com utilizes undisclosed information in credit ratings. Also, they want to find out the effectiveness of these ratings in default prediction. They conclude that undisclosed information embedded in the credit score has a significant role in default prediction. Predictive effectiveness is better for high-risk borrowers compared to low-risk borrowers. This means that P2P platforms may have additional data what they do not share. This data has been used to calculate credit ratings. This means that predicting solely from the data, provided by the platform, might not be enough to accurately predict default. So, considering the credit rating assigned by the platform is important in default prediction.

Next article is from Lin et al. (2017). They examined factors that determine the default risk based on the demographic characteristics of a borrower. Data was gathered from yooli.com which is a Chinese P2P lending platform. Nonparametric tests were used to find variable differences between good and bad borrowers. Then logistic regression was used to predict

default. They found out that low default risk characteristics are female gender, young adults, long working time, stable marital status, high educational level, working in large company, low monthly payment, low loan amount, low debt to income ration and no default history. All monetary related variables make sense in this case. Low monthly payment means that you do not have to pay that much interest every month. Low loan amount means that the loan itself is not that big of a liability to overall budget. Low debt to income ratio is very important since existing liabilities make you a riskier borrower. No default history is also an important variable since defaulting chances are much higher if it has happened before. This also means that borrower is more prone to default leading behaviours such as spending too much on unnecessary things.

More interesting characteristics are not monetary related in Lin et al. (2017) article since these cannot be foreseen. Being a woman seems like a very interesting variable. It might make sense since men tend to have riskier behaviour than women. Young adults also might make sense since they still have job opportunities while older people might already be retired so coming up with enough money is more difficult, but it is peculiar that young adults are more creditworthy than middle aged people. Long working time makes a lot of sense because this means that borrower has a steady job, and thus has financial security almost guaranteed. Stable marital status is an interesting variable but can be explained with the fact that married couple always help each other out so if liabilities come to overwhelming for one, the other can help. High educational level makes a lot of sense since higher education most likely means bigger salary. Finally, working in a large company is also sensible variable, because large companies have a lot of wealth which means they can keep operating for long periods of time and have the cash to pay salaries. These variables are very interesting, and I should keep eye on nonmonetary variables since there might be many variables that provide a lot of information even though, at start, it does not seem that way.

Fourth article is from Santoso et al. (2020). The point of their research is to investigate the determinants of interest rate and borrower default. Their data consists of three datasets from Indonesia. They used OLS and LR to identify these variables. Borrowers' characteristic impacts on interest rates are as follows. Young people tend to get higher interest rates, which makes sense since they do not have experience or as good jobs as middle aged people. Also, marriage is a sign of maturity in Indonesian culture, so they tend to get lower interest rates. Interestingly, young, and married couples tend to get higher interest rates since they have a lot of things on their plate at once. For example, the couple might need a bigger house, a car,

or some other big liabilities on the way, while older people have these things already covered. Characteristics that determine default are somewhat the same as in Lin et al. (2017) article, like higher interest rate, bigger loan amount, longer loan period, unmarried, and lower education. But, in this article women have bigger chance to default. In the article this is explained with cultural differences since Muslim women tend to have less experience in financing. This is completely opposite what was discovered in Lin et al. (2017) article. Not owning a house was also important variable in default determination. Income and age had mixed results. This study shows that cultural differences can influence on variables and how they effect in a different way. This means that there may be differences between countries which is what this thesis is all about.

Fifth article is from Serrano-Cinca et al. (2015). The point of this research is to identify variables which explain default. They used data from American P2P lending platform Lending Club. They used univariate means test and survival analysis in variable selection and logistic regression for prediction. They found out that variables that describe default are loan purpose, annual income, credit history, current housing situation, indebtedness and credit rating signed by the platform. These findings are mostly the same as in two previous research (Lin et al. 2017, Santoso et al. 2020). New variable here is loan purpose, which is quite interesting. For example, the least risky is loan for a wedding and the riskiest is a loan for small business funding. This makes sense since weddings are just one-time expenditures and the income of the couple stays the same, but small businesses funding is risky because these types of businesses have high probability to fail. Once that happens, the income source is gone as well, and so the default happens. Also, there might be other liabilities existing before this loan since small businesses require a lot of funding. Most important variable in this research was the grade assigned by the P2P lending platform.

Next article is from Wang et al. (2020). Their objective was to evaluate default behaviours of borrowers with psychological variables. They used online survey as a data source with 713 responses. General strain theory was used to in this study. The empirical section was done in two parts. First stage was to find significant variables and second stage measure the effects of these variables. The study finds out that economic pressure, socialization difficulty and negative effects (variables such as life dissatisfaction, perceived unfairness, inferiority feeling and loneliness) increase the probability of default. Also, income and ownership of a house have significant effects. It is interesting to see that defaulters have these antisocial tendencies. It makes sense because if one might feel these negative feelings, paying back loans is

probably not a priority in their mind. These variables are probably not present in the data I use, but it is interesting to know that people with difficulties paying back loans might have something else that is pressing down their life. These feelings might come overwhelming and totally paralyze someone. So, it is important to recognize variables that might reflect these negative effects indirectly.

Final article in this section is from Wu & Zhang (2020). Their research focuses on credit ratings and their reliability in P2P lending. They used data from Chinese P2P lending platform renrendai.com. Also, logistic regression was used in this research. This research shows that credit ratings provided by renrendai.com do not accurately predict their default risk. This is bad news for investors since they depend on this information. In previous studies by Jagtiani & Julapa (2019) and Serrano-Cinca et al. (2015) credit ratings were accurate default predictors, but they were constructed using Lending Club data. This indicates that credit rating systems are very much different between platforms. It seems that not all of them are reliable. A high probability of default in new borrowers suggests that renrendai.com platform does not screen their applicants effectively. These results suggest that constructing an effective personal credit system may be the key to healthy economy in Chinese P2P lending market. This gives the impression that credit ratings are not enough in their own and personal prediction models are essential.

To summarize these determinants of P2P default articles, investors should keep an eye on variables that give information of applicant's financial situation. Most reoccurring variables in in this research are interest rate, credit score assigned by the platforms, loan amount, debt to income ratio, credit history, and longer loan period (Chen et al. 2019; Lin et al. 2017; Santoso et al. 2020; Serrano-Cinca et al. 2015). All these variables are financial variables which give important information to lenders/investors in P2P financing. To average investor these are the variables to look for and see if they have healthy levels compared to majority. Although, Wu & Zhang (2020) points out that credit rating signed by the platform is not always accurate. But, if investor seeks superior performance, more variables should be included in the mix. Alternative data, such as demographic and psychological variables, are proven to increase predictive performance of models (Jagtiani & Julapa 2019; Lin et al. 2017; Santoso et al. 2020; Wang et al. 2020). To conclude, superior models can be created by including many variables from financial to alternative variables. To determine importance of variables, feature selection techniques should be used. Now that all important variables are known, this knowledge can be used to further develop prediction models.

## 4.4 Predicting default with machine learning in P2P lending

The point of this chapter is to explore different prediction methods, much like in chapter 3.2, but this time only P2P lending data was used to create these models. This chapter shows that credit scoring in general might differ from P2P lending credit scoring. Table 3 consists of 15 different articles that explores the use of prediction models in P2P lending. 11/15 of the articles used data from American P2P lending platform Lending Club. Rest of the articles use data from different Chinese P2P lending platforms, like renrendai.com, PPDai and we.com. 6/15 articles used some sort of balancing methods, so that the number of defaulters and creditworthy applicants is the same. These methods increase prediction capabilities as proven by Brown & Mues (2012).

Table 3 also has a list of models used in articles. Most used models got their own column in the table while once used, more advanced models, are in Other-column. If a model was used, it has a percentage in the table. The percentage is AUC score from prediction models. This helps to determine which models performed the best. Also, AUC average score with and without balancing are calculated to determine if balancing gives better prediction performance to models. Table 3 shows that 5 most used models are LR (logistic regression), RF (Random Forests), GB (Gradient Boosting), SVM (Support Vector Machine), and DT (Decision Tree), respectively. AUC scores for these most used models are 68,5% (LR), 66,6 % (RF), 68,0 % (GB), 69,0 % (SVM), 67,4 % (DT). GB models provide best prediction performance from top 5 models. Honourable mentions for good performing models are NN (neural networks) with AUC of 68,1 % and MLP (Multilayer Perceptron) with AUC of 70,6 %. Interestingly, more advanced models used in this set of articles did not perform that much better compared to these widely used, more traditional models. Furthermore, balancing of the data gives better prediction performance on average 67,7 % vs. 66,3 %. So, balancing indeed should be included in the models. But there are only a few researches that use balancing in this list so these average calculations can only be used as direction. As previously mentioned, models become more predictive because, without balancing, models become biased towards the majority class (Brown & Mues, 2012). Next this set of articles is broken down in detail so that a full picture of default prediction in P2P lending can be acquired.

Table 3. Default prediction in P2P lending articles

| | Objective | Prediction models | | | | | | | | | | Balanced data | Data |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **Author(s) and year** | | **LR** | **RF** | **NN** | **MLP** | **SVM** | k-NN | NB | **GB** | DT | Other | | |
| Ariza-Garzón et al. (2020) | Typical models provide good predictions but lack explanatory power. This study aims to use SHAP values in LR and several other models to reveal that machine learning algorithms can be predictive and transparent. | **66.6%** | **66.3%** | | | | | | **67.4%** | 64.7% | | x | Lending Club |
| Bastani et al. (2019) | This paper showcases two-stage scoring approach in P2P lending. Using Credit scoring (default prediciton) and profit scoring combined is the solution. Also, imbalance problem has been addressed. | | **70.0%** | | | **69.0%** | | | **70.0%** | | WL: 69.0 %, DL: 70.0 %, WDP: 71.0 % | x | Lending Club |
| Cho et al. (2019) | This study proposes an investment decision model which consists of fully paid loans classified with instance-based entropy fuzzy support vector machine (IEFSVM). | | **52.3%** | | | | | | **56.6%** | | AdaBoost: 59.2 %, EasyEnsemble: 55.3 %, RUSBoost: 58.4 %, EFSVM: 57.1 %, IEFSVM: 59.4 % | | Lending Club |
| Jin et al.(2015) | This study compares five different data mining methdos and determines which one is the most useful in terms of default prediction accuracy. | | | | x | **x** | | | | x | CHAID, RBF | | Lending Club |
| Li et al. (2019) | This study focuses on predicting prepayments and defaults in P2P lending. | **x** | | | | | | | | | | | Lending Club |
| Li et al. (2018) | In P2P lending, default prediction is important to reduce credit risk. This study aims to provide better models to improve prediction accuracy. | **77.5%** | | 76.7% | | | | | **78.7%** | | Ensemble: 78.9 % | | PPDai |
| Liu et al.(2018) | This research uses prediction models to predict default. | **x** | | | | | | | | | | | Renrendai.com |
| Malekipirbazari & Aksakalli (2015) | The goal of this study is to predict default by using random forests classification method. | **68.0%** | **71.0%** | | | **62.0%** | 53.0% | | | | | | Lending Club |
| Moscato et al. (2021) | This study aims to decrese lender risk by introducing most used machine learning credit scoring methods to predict P2P default. | **70.0%** | **71.0%** | | 70.0% | | | | | | | x | Lending Club |
| Serrano-Cinca & Gutiérrez-Nieto (2016) | This paper studies porfitability rather than default probability. The focus is to predict the expected profitability. | **x** | | | | | | | | | Profit scoring, CHAID | | Lending Club |
| Teply & Polena (2020) | Ranking 10 different classification methods for default prediction in P2P. | **69.8%** | **69.3%** | 69.8% | | **69.7%** | 63.6% | 66.9% | | 63.7% | SVM-Rbf: 65.2 %, B-Net: 67.9 %, LDA: 69.6 % | | Lending Club |
| Wang et al. (2018) | This study aims to use behavioral elements to default prediction. Classification models yield a static probability while borrower repayment behaviour evolves dynamically. | | | | | | | | | | EMRF, WOE | | Chinese dataset |
| Xia et al. (2017) | Most traditional loan evaluation methods assume balanced misclassification cost which is far from reality. This study aims to use cost-sensitive approach to tackle that issue. | **64.7%** | **68.4%** | | | | | | **70.0%** | | | x | Lending Club and we.com |
| Zanin (2020) | Goal of this paper is to predict default in case of imbalanced data. | | **67.3%** | | | | | 64.9% | **68.4%** | | GAM: 67.54 % | x | Lending Club |
| Zhou et al. (2019) | Credit risk is inevitable in P2P. To reduce this a default prediction model needs to be created that can effectively and accurately predict default probability of each loan. | **62.7%** | **63.5%** | 64.2% | | **65.3%** | 59.0% | | **71.6%** | 70.0% | LightGBM: 70.29 %, AdaBoost: 67.55 % | | Chinese dataset |
| Count | 15 | **10** | **9** | 3 | 2 | **5** | 3 | 2 | **7** | 4 | | 5 | 15 |
| | | | | | | | | | | | | Total average | |
| AUC average | | **68.5%** | **66.6%** | 70.2% | 70.0% | **66.5%** | 58.5% | 65.9% | **69.0%** | 66.2% | = | 66.8% | |
| AUC average with balanced data | | **67.1%** | **69.2%** | | 70.0% | **69.0%** | | 64.9% | **69.0%** | 64.7% | = | **67.7%** | Highest average |
| AUC average with imbalanced data | | **69.5%** | **64.0%** | 70.2% | | **65.7%** | 58.5% | 66.9% | **69.0%** | 66.9% | = | 66.3% | |

First, articles that cover many different prediction methods are examined since these ones provide a broad picture of what methods work and what does not. There are two articles that have examined many models at once, Teply & Polena (2020) and Zhou et al. (2019).

Teply & Polena (2020) examined 10 different prediction methods and compared which of them provide the best prediction performance. They used data provided by American platform Lending Club. Their data was not balanced. 33 780/178 500 = 18,9 % of their data was defaulters. This means that their results are a bit biased towards majority. Also, no feature selection methods were used. Important variables were hand-picked, which seems a bit odd. 5-fold cross-validation was implemented for training the data which is good. With these methods in place, the results are good enough to give broad view of the models, but some models, like SVM, might not perform to their fullest potential because of the imbalance. Five best performing models in this research, evaluated with AUC, are as follows: 1. LR (69,79 %), 2. ANN (69,75 %), 3. L-SVM (69,67 %), 4. LDA (69,55 %), and 5. RF (69,28 %). The margins seem very small but even a small increase can make a difference in investing. Logistic regression performed the best, which is interesting result since this is the simplest method to use and it provides the best result as well. K-nearest neighbour and regression tree are not recommended to be used in P2P default classification. They provided the worst results. This

research reinforces the idea of what algorithms to use. Previously in chapter 3.2, LR, SVM and RF were considered as good models, and this research shows that as well.

Zhou et al. (2019) objective in the research is to reduce credit risk by predicting default with 10 different models. They used a Chinese dataset provided by a well-known P2P platform in China. They did not balance the data. However, they used gradient boosting (GB) models to tackle this issue, which are resilient to imbalanced data. This means though that other models, which are more vulnerable to imbalance, are weaker. Feature selection was used to determine important features. With these methods in mind, results will be reliable, but some models do not get as good AUC scores as others because of the imbalance. These imbalance resilient models performed the best reaching AUC scores of 70 % approximately. More traditional models like NN (64,2 %), LR (62,7 %), SVM (65,3 %) and RF (63,5 %) did not perform as good. It is evident that gradient boosting algorithms have an edge in imbalanced datasets. This research showed that there is an option to use gradient boosting models to tackle class imbalance. Interestingly, SVM performed best from the traditional models even though class imbalance effects its prediction performance the most (Brown & Mues 2012).

Next, P2P default prediction is examined through articles that use balancing techniques to overcome misclassification. First article is from Ariza-Garzón et al. (2020), and their objective is to increase prediction models' explanatory power. They used data from Lending Club. The goal was to use SHAP values in LR and other models to see exactly what features are important. Furthermore, this article demonstrates how imbalanced data results in misclassification and can lead to models that superficially predict default in terms of accuracy. But when examined through recall measurement, it shows that imbalanced models do not predict default at all. Algorithms are not fed enough default cases to be accurate enough to examine defaulters. In the article LR first predicts default the best. But inspecting more closely with recall, the model is not good at all. But after balancing was implemented, the model performed much better, even though accuracy drops. AUC scores are as follows: LR-balanced = 66.6 %, DT = 64.7 %, RF = 66.3 %, XGB = 67.4 %. SHAP values demonstrate that same attributes, as selected from LR, are found to be important. Also, SHAP values indicate that some variables have more complex relationships on default than LR leads on. This study shows that balancing of the model is essential in default prediction.

Second balancing article is from Bastani et al. (2020). Their goal was to create a two-stage scoring approach to help lenders invest in P2P lending. In first stage they rebalance the data and predict default. In second stage profit scoring method is used. Lending Club data was used. Their results show that rebalancing of the data does not always increase precision, so the results are mixed. Except IHT (Instance Hardness Threshold) balancing method increases AUC scores significantly in every model. AUC without resampling and (with IHT): WL (Wide Learning) = 51 % (69 %), DP (Deep Learning) = 50 % (70 %), WDP (Wide and Deep learning) = 50 % (71 %), SVM = 53 % (69 %), GB = 44 % (70 %), RF = 47 % (70 %). IHT is an under resampling technique which reduces cases in majority class. IHT selects samples to remove with high hardness levels, and hardness in this case is the likelihood of misclassification for each sample. This explains a lot why AUC scores are very good. All instances that are hard to classify are removed until class sizes are the same, so the model produces very good prediction performance. This is not a very good scientific method since it seems like data manipulation. I would prefer RUS (Random Under Sampling) since it randomly removes samples from majority class. Majority class has a vast number of samples in this case, so it does not hurt evaluation that much. SMOTE is also a good method but it provides artificially new samples to minority class, so they are not “real world” samples. These two methods managed to increase AUC scores just a bit in some models and decrease in others. Results are mixed here. The second stage of the research is profit scoring method where they implemented IRR (Internal Rate of Return). They managed to get outcomes that resulted in positive rate of return which is good from investor perspective. I would still exercise caution to trust these results since IHT balancing method is somewhat questionable.

Third article that used balancing is from Moscato et al. (2021). The aim of the article is to decrease lender risk by introducing most used machine learning algorithms (RF, LR and MLP) for credit scoring and predicting default. Also, they used different balancing methods. This article used data from Lending Club. Without balancing, models had very high accuracy, but they are biased towards majority class. Balancing the data results in lower accuracy but more reliable results. Algorithms are ranked with G-mean values, which considers classification performance in both minority and majority classes. Using random under sampling method the best performing algorithm was random forests, RF-RUS had AUC of 71 %. Using Random over sampling method the best performing algorithm was logistic regression, LR-ROS had AUC score of 71 %. Using SMOTE balancing method, the best performing model was logistic regression, LR-SMOTE had AUC score of 71 %. From these three best performing models, RF-RUS had the highest G-mean score which means that it was the best of them all. This article provided important insight in regards of balancing. This solidifies that balancing is

indeed important and will give better results. Furthermore, G-mean scores compared to models without balancing are much higher, for example RF-RUS G-mean score is 0,6560 while just RF model has 0,2870. This indicates that balancing is essential for prediction with P2P data. Also, RUS balancing method provided best overall results in terms of G-mean score. This further strengthens the decision of picking this method in this thesis.

Next article that uses balancing is from Xia et al. (2017). Their objective is to create a model that predicts default and is also cost-sensitive because most traditional models assume balanced misclassification cost which is far from reality. They used data from Lending Club and we.com, which is a Chinese P2P lending platform. As benchmark models, they used LR and RF. All models were also balanced using SMOTE technique. The model proposed is cost-sensitive extreme gradient boosting model, and it exceeds both benchmark models in terms of ARR (average rate of return). This means that cost-sensitive models are better in terms of profits since it gives misclassification a cost that traditional model do not do. SMOTE balancing method managed to increase the performance of unbalanced LR model, but it also resulted in worse AUC score in RF. This Study shows that just predicting the outcome is not enough since bad outcomes have high costs. CSRF-SMOTE model even had a negative ARR which means that you would lose money if this model were used. It also shows that balancing does not always improve models. RF should be balanced using RUS as mentioned in previous research (Moscato et al 2021).

Final article that uses balancing is from Zanin (2020). The goal of this article is to predict default in presence of imbalance. Data is used is from Lending Club. This research uses four different rebalancing methods: random under-sampling (RUS), random over-sampling (ROS), random under- & over-sampling (RUOS), random over-sampling examples (ROSE). Also, it uses four supervised learning methods: generalized additive model (GAM), naive Bayes (NB), random forest (RF), and extreme gradient boosting (XGBoost). All models produce similar results regardless of the balancing methods or imbalanced data. In terms of AUC, GAM performed best with ROS, NB performed best with ROSE, RF performed best with RUS, and XGBoost performed best with imbalanced data. The differences between AUC scores are marginal between balancing techniques and imbalanced data. Predictive algorithms XGBoost and RF are resilient to imbalanced data so this might explain small margins in these algorithms (Brown & Mues 2012).

Overall, balancing methods provide better prediction performance. Some of the articles provided mixed results (Bastani et al. 2020; Xia et al. 2017; Zanin 2020), while some provided positive results (Ariza-Garzón et al. 2020; Moscato et al. 2020). There were cases when balancing methods made model prediction performance worse, but it depends on the technique used. In chapter 3.2 Dastile et al. (2020) article provides a good framework which shows how all credit scoring research should be conducted. In this article, the chosen balancing method is SMOTE, but many of the research in this chapter showed that SMOTE does not provide better results for RF algorithm (Bastani et al. 2020; Moscato et al. 2020; Xia et al. 2017). Also, RF should be one of the methods to include in research since it provides very good prediction result and is also considered as best benchmark model for new models to beat (Lessmann et al. 2015). In Moscato et al. (2020) article, RUS technique provided best overall results, so if only one balancing method was used, this should be the one. RUS technique is only weak if there are too little samples in minority side, so balancing the data might result in too small data. But the data provided by Bondora, which is used in this thesis, has plenty of samples so I doubt this will be a problem.

Next, articles that do not use balancing methods are briefly introduced. This part shows that if balancing is overlooked, the results will have very different outcomes. Even if more advanced methods are used, AUC scores tend to be lower compared to balanced datasets. Credit scoring using P2P data is very imbalanced, so models do not accurately understand defaulting behaviour.

First article is from Cho et al. (2019). They propose an investment decision model called instance-based entropy fuzzy support vector machine (IEFSVM). The data is from Lending Club. They did not use any balancing methods, except for one model. They were aware of the imbalance in the data and used models that do not get effected that much by the imbalance. They managed to get very good accuracies for their models. All of them were around 90 %. But there is in issue with AUC values. All of them are between 50-60 %. This means that models can accurately classify non defaulters but lack in predicting default. This issue arises because the data is not balanced and so models are biased towards the majority class. Balancing methods should be introduced here so that models can identify defaulters with better precision. New proposed model IEFSVM performed well, all things considered. Accuracy was 92,16 % and AUC was 59,38 %. In balancing part, models got much better AUC scores than this though.

Second article is from Jin et al. (2015). This study compares five different data mining methods and determines which one is the most useful in terms of default prediction accuracy. They used data from Lending Club site, and they used random forests for feature selection, but no balancing methods was used. Models that they used were CART (Classification and Regression trees), CHAID (chi-square automatic interaction detector), MLP (multilayer perceptron), RBF (radial basis function) and SVM (Support vector machine). SVM was the best predictor with 72.05 % accuracy. Second was MLP with 71.24 %. Third was CART with 71.23 %, Fourth was CHAID with 70.9 %. Last place was RBF with 68.11 %. They only used precision to compare models, which is not enough. This method only considers correctly classified samples overall and does not compare the ratios of how many correctly classified defaulters or non-defaulters there are. Because no balancing was used, most of the correctly classified samples are probably non-defaulters. So, the correctly classified defaulters compared to actual number of defaulters is probably low. This article is not properly conducted so the results are not comparable to other articles that use proper methods. This article shows how important it is to report all significant metrics so that models can be properly evaluated.

Third article is from Li et al. (2018). Their goal is to predict default with different models so credit risk can be decreased. They used data from a Chinese P2P lending platform called PPDai. Balancing was not used, and feature selection was performed with XGBoost model. They also used XGBoost model to pre-train the dataset, and hyperparameter optimization was used. They used models Extreme gradient boosting (XGBoost), Deep neural network (DNN), logistic regression (LR) then a linear weighted fusion of these models (Ensemble). Model managed to get an AUC scores as follows: ensemble 78,91 %, XGBoost 78,69 %, DNN 76,70%, LR 77,51 %. XGBoost model after hyperparameter optimization can beat traditional machine learning models in default prediction. The fusion of these three models with linear weighted fusion resulted in ration of 0.75:0.20:0.05 (XGB, DNN, LR respectively) and AUC of 78,91 %. This means that heterogenous models can improve the prediction performance compared to traditional single method algorithms. Even though no balancing methods were used, this article provided good AUC scores. It is possible that pre-training and feature selecting the data with XGBoost method gave these results. XGBoost model is very resilient to imbalanced data and that is why metrics had good performance (Brown & Mues 2012).

Next article is from Liu et al. (2018). Their objective is to reduce credit risk with default prediction. They used data from a Chinese platform called renrendai.com. This study did not

use balancing techniques, nor feature selection. This article used only LR as prediction method. It was well constructed and many of the requirements of the model were checked. None of the usual single number metrics were used in this study, only confusion matrix of the correctly and incorrectly classified samples. Even though the methods used are simple, they provided good prediction results as type 2 errors, which are defaults but misclassified as non-defaults, were very low (only 40 samples out of 1408 were misclassified). They managed to increase predictive performance by changing prediction threshold value. Meaning if threshold is set to 0.1, any sample that has 0,1 or higher default probability, will be categorized as defaulters. Mean default probability in all samples is 0,163 so setting threshold to 0,1 increases the number of predicted defaulters. This way type 2 errors decrease but this increases type 1 error. Type 1 error is tolerable since the goal is to predict defaulters accurately, not credit-worthy samples. This means that LR can accurately predict default if probability thresholds are implemented. By tweaking probability threshold, one can decrease credit risk significantly. The only downside is that there are less credit-worthy samples, but if the number of borrowers is high enough, this should not be a problem.

Final article without balancing is from Malekipirbazari & Aksakalli (2015). Objective of this research is to predict default by using random forests algorithm and compare it to traditional ones. The data is gathered from Lending Club. Also, feature selection was used but balancing is not. Prediction algorithms used are k-Nearest neighbours (kNN), Logistic regression (LR), Support vector machines (SVM), and random forests (RF). Random forests method provided the best AUC of 71 %. Second is LR 68 %, third SVM 62 % and fourth kNN 53 %. Differences between model performance could have been smaller if balancing were implemented since it improves models such as LR and SVM but does not affect RF that much (Brown & Mues, 2012). RF comes with a cost since some of the good borrowers are classified as bad. Random forest algorithms are superior in identifying the best of the best borrowers. Although, research shows that higher risk is not worth the higher return from a risk/expected return trade-off point of view. Thus, predicting the best of the best is not that bad of a thing. Furthermore, in P2P lending there is plenty borrowers to choose from, which makes the choice of using RF rather simple.

Table 3 showcased that balanced and imbalanced dataset using articles are close in terms AUC score, but balanced data seem to provide better prediction performance overall. After closer inspection of the articles, there seems to be other overlooked issues in some of the articles that used imbalanced data. For example, two articles did not use proper metrics (Jin

et al. 2015; Liu et al. 2018). Also, one research reported accuracy as main prediction performance indicator and declared that this novel model provides incredible performance (Cho et al. 2018). But, when looked more closely, AUC score of this model was not that different from other models. Reviewing these articles, one must be careful, since proper metrics may not be always used which results in false perceptions. All in all, balancing should be used since it provides more reliable results and makes algorithms more able to accurately identify default.

Next, three more articles are examined. These articles are not that relevant in terms of default prediction, but I think they provide good insight of different methods and point out some problems in P2P lending. First article is from Serrano-Cinca & Gutiérrez-Nieto (2016). Objective of this research is to predict expected profitability rather than default using profit scoring method. They used data from Lending Club. Data is not balanced nor feature selection was used. They found out that P2P lending is not currently efficient market. This means that data mining techniques can find ways to make arbitrage profit. Furthermore, they found that credit scoring and profit scoring has different factors. Profit scoring system outperforms traditional credit scoring (default prediction) system based on logistic regression. LR credit scoring results 5.98 % average IRR (internal rate of return), Profit scoring by means of linear multivariate regression results 11.92 % IRR, CHAID, which is a type of decision tree, results an average of 8.57 % IRR. This study differs from its peers since it examines the profitability side of the spectrum rather than just default probability which makes it very interesting. Investors should pay attention to profit scoring and considered it as an alternative for credit scoring.

Second article is from Wang et al. (2018). Objective of this research is to use behavioural scoring model for default prediction. Traditional classification models yield a static probability while borrower repayment behaviour evolves dynamically. So, more dynamic model is needed. They used data from major Chinese P2P platform. Ensemble mixture random forest (EMRF) and Weight of evidence (WOE) were used as models. The results provide default prediction in intervals as the loan matures. Results show that EMRF model predicts the defaults very well overtime and is also able to predict when a borrower is likely to default. LR only succeeds to predict default accurately at the end date. Using this model one can determine when defaults can occur and intervene if potential default is coming. This can effectively reduce loan delinquencies as well as the number of accounts that become bad debts

Third and final article of this literature review is from Li et al (2019). Their objective was to predict prepayments and default in P2P lending. They used data from Lending Club, and no balancing was used. Also, feature selection was used. This article only uses logistic regression, but it is a sufficient method for prediction, but better models of course give better results. This study finds out that prepayment and default have different patterns. They also found that prepayment is very common in P2P lending, as borrowers in urgent need of money seek instant relief for their monetary issues. This is not good news from investor perspective since this means less interest income. Without penalties, P2P lending is usually used as a short-term and low-cost loan solution for financial distress since they pay the loan back as soon as possible, even though it is designed as 36-month or longer terms. Thus, prepayment is much more likely to happen and there may be an arbitrage opportunity of abusing P2P lending loans. This is a problem since many loans are valued with interest rates and for the whole maturity of the loan, but if loans are often prepaid, the pricing of loans is wrong. There should be a penalty fee to compensate for the potential losses of loan portfolios.

## 4.5 Literature review summary

In chapter 3.2 credit scoring and credit management was covered in general form. This was done using research that included mostly traditional bank loans and their credit scoring using machine learning and statistical methods. During the review of this article, it was evident that balancing the data has significant positive impact on prediction metrics (Brown & Mues 2012; Dastile et al. 2020; Zanin 2020; Louzada et al. 2016). Also, feature selection was seen to have a positive impact. One article showed that chi-square method provided good results (Trivedi 2020). Table 1 shows that most used methods in chapter 3.2 were SVM (7 articles), RF (6 articles), and LR, NN and k-NN (4 articles). According to two literature reviews (Keramati & Yousefi 2011; Louzada et al. 2016) and one vast study of classifiers (Lessmann et al. 2015) most popular and simple models to use are SVM, RF and LR. These models also provide good prediction performance.

In chapter 3.3 factors that explain default in P2P lending were covered. This chapter helped to identify important variables that should be kept an eye on. Most reoccurring variables in in this chapter are interest rate, credit score assigned by the platforms, loan amount, debt to income ratio, credit history, and longer loan period (Chen et al. 2019; Lin et al. 2017; Santoso et al. 2020; Serrano-Cinca et al. 2015). All these variables are financial variables which give

important information to lenders/investors in P2P financing. Alternative data, such as demographic and psychological variables, are proven to increase predictive performance of models (Jagtiani & Julapa 2019; Lin et al. 2017; Santoso et al. 2020; Wang et al. 2020). Low default risk demographic characteristics are female gender, young adults, long working time, stable marital status, high educational level, working in large company, and loan purpose. This chapter helps to identify important variables in advance and a broad idea of what should be included in the algorithm of this thesis.

In chapter 3.4 credit scoring was thoroughly examined from P2P perspective. This chapter showed similar results compared to chapter 3.2. One big difference is that data in chapter 3.2 was more often better balanced. This means that research in chapter 3.4 might have differences in prediction accuracies compared to articles in chapter 3.2. Table 3 shows that balancing has somewhat mixed results in predictive performance to these most used algorithms but mostly it enhances algorithms capabilities. Only LR seems to get worse, and GB remained the same, while others improved. But on average, balancing improved algorithms (Ariza-Garzón et al. 2020; Moscato et al. 2021). In Moscato et al. (2020) article, RUS (random under-sampling) technique provided best overall results. Five most used models in this chapter were LR (10 articles), RF (9 articles), GB (7 articles), and SVM (5 articles). Popularity of models are very similar with results from chapter 3.2. In terms of AUC score in most popular models using balanced data, RF performed the best with 69.2 %, SVM and GB are tied to second spot with 69.0 %, and LR is third with 67.1 %.

Literature review constructed here has provided a lot of insight in credit scoring and default prediction. All previous research is put to good use and should provide the necessary knowhow of building prediction models in P2P lending. All decisions made, from chosen models to balancing techniques, are supported by previous research. This review made it clear how to construct theoretical framework for this thesis and should be enough information to further continue this research. In the next chapter, all chosen methods and metrics are being examined more closely to fully understand their functions.

# 5 METHODOLOGY

In this chapter, all the methods that are used are described in detail. This gives the necessary knowhow to complete this research. This chapter includes feature selection, data balancing, data modelling, and evaluation.

## 5.1 Justification of used methods

With information from literature review, the theoretical framework can be constructed properly. Furthermore, Dastile et al. (2020) provides a good theoretical framework for future research. According to this literature review, theoretical framework should be constructed as follows:

1. Exploratory data analysis and data pre-processing
2. Feature selection/Feature engineering
    - Rough set/Genetic algorithm
3. Balanced data
    - If balanced continue, if not, use SMOTE for balancing
4. Benchmark models
    - LR/DT
5. Ensemble classifier/CNN (Convolutional neural network)
6. Evaluation metrics
    - PCC, AUC, G-mean, Recall, F-measure
7. Model transparency
    - LIME

This framework gives a clear idea how credit scoring research should be constructed. However, some changes are needed to make this framework suitable for this thesis. Next, I propose how this thesis theoretical framework is constructed in Figure 8.

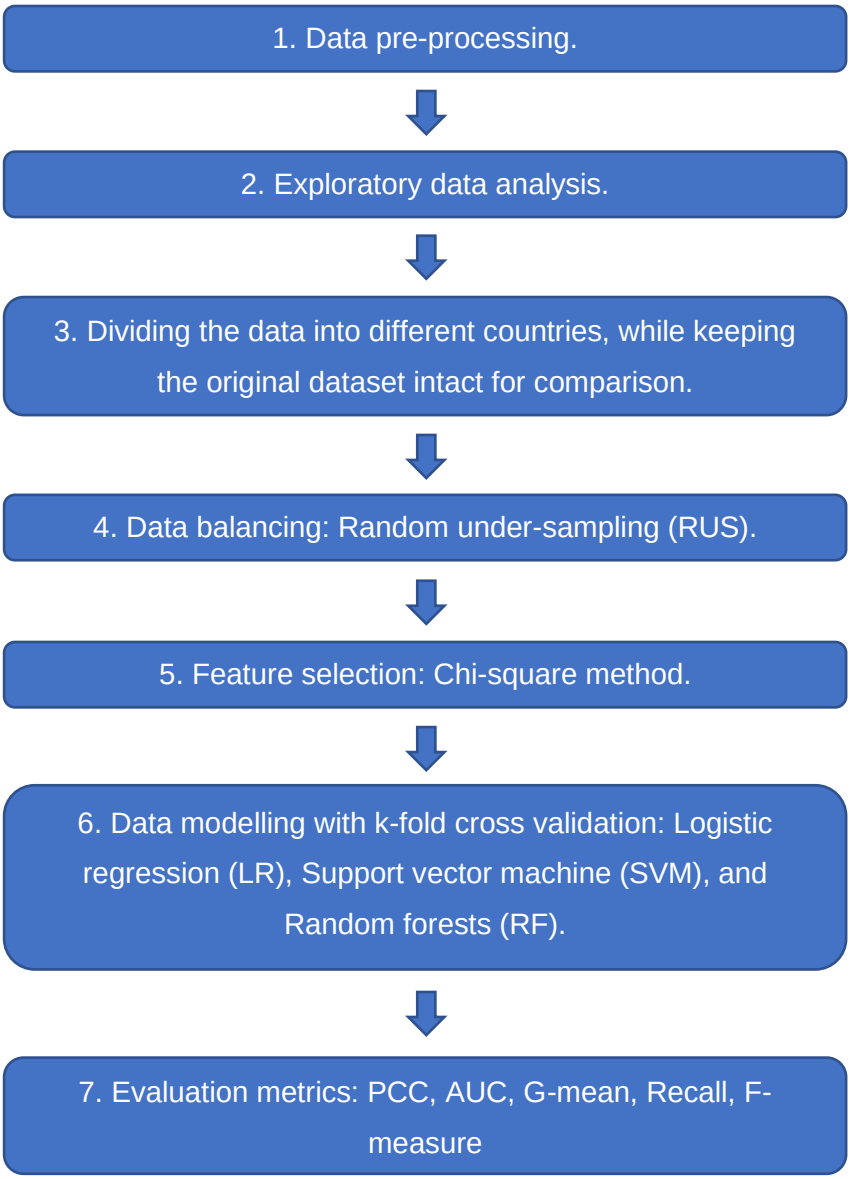



Figure 8. Illustration of theoretical framework

Figure 8 shows how empirical research is constructed in this thesis. It is similar what Dastile et al. (2020) proposed but with few differences. Step 1, pre-processing is constructed so that the data is prepared for further usage. Step 2, exploratory data analysis is constructed to see, for example, how many delinquencies there are and how many samples there are in each country. Step 3, data is split to each country for further examination while whole data set is kept intact. This provides a good comparison how each country may differ from original data. Also, defaults are examined for each country in this part to get a general idea if some countries have more default than others. Step 4, Balancing is implemented. RUS method is used instead of SMOTE. RUS provides better overall results, since SMOTE did not enhance RF model that much in previous research. Step 5, feature selection is constructed with Chi-square method.

Even though Dastile et al. (2020) proposed different methods, previously in literature review, Chi-square method was proven to be adequate method to use, and it is very simple to build. Also, in this part, features chosen are being compared between countries to see if they have some differences. Step 6, data modelling can begin. The chosen models are LR, SVM and RF since they are very popular and effective models according to literature, and they are simple to use (Keramati & Yousefi 2011; Lessmann et al. 2015; Louzada et al. 2016). The purpose of this thesis is to examine differences between countries so using advanced models would just make this thesis more complicated without any benefits. These models are trained using k-fold cross validation method. Final step 7, these methods and countries are compared using the following metrics: PCC (accuracy), AUC, G-mean, Recall, F-measure. LIME method is not used for transparency, since models that are used in this thesis are not that complex and results should be interpretable.

## 5.2 Feature selection: Chi-square method

Feature selection, as a data pre-processing step, has been proven to be an effective and efficient technique for data mining and machine learning purposes. The objective of feature selection is to narrow down high-dimensional data to make it more understandable, and only including the most important variables. Also, having many features, tend to overfit machine learning models. This may cause performance decrease on unseen data. Furthermore, high dimensional features increase the computational requirements and can be costly for data analytics. (Li et al. 2017)

Chi-square statistic is a non-parametric (distribution free) tool which is used to analyse group differences when dependent variable is nominal. Chi-square method is also very robust with respect to the distribution of the data. Specifically, it does not require homoscedasticity and equality of variance in the data among the study groups like some other methods would. Furthermore, chi-square method provides considerable information about how each variable performed in the study. This richness of information helps researchers to understand results thoroughly and thus derive more detailed information from the data. Chi-square statistic can be calculated as follows: (McHugh 2013)

$$\sum_{i-j} \chi^2 = \frac{(O-E)^2}{E} \quad (1)$$

Where:

O = Observed (the actual count of cases in each cell of the table)

E = Expected value

$\chi^2$ = The cell Chi-square value

If chi-square has a larger value than one, it means that it differs from the expected value. A positive value of chi means that observed value is higher than expected value and a negative means that observed value is smaller than expected value. Also, bigger variance from expected value means that variable has more effect. (McHugh 2013)

## 5.3 Data balancing: Random Under-Sampling (RUS)

Data balancing is important because credit score data usually has a vast majority class compared to minority. This will result in situation that algorithm cannot fully understand minority, default class, since it is fed mostly with cases that are fully paid. This will result in biased algorithms that provide good results in accuracy, but with more close inspection, usually algorithm can only predict paying borrowers well which is not the point at all. Algorithm should be capable of recognizing defaulters, not good borrowers. (Brown & Mues 2012)

To tackle this problem random under-sampling is implemented. This algorithm randomly samples the majority class (fully paid loans) by reducing the number of cases to match the minority class (loans in default). An advantage to this method is reduced size of the data, which is computationally easier to handle. Disadvantage though is the loss of information. But if there are enough cases and samples are randomly removed, the mean values of the variables should still stay the same. (Zanin 2020)

## 5.4 Validation of models: K-fold cross validation (CV)

Classification model and results need to be validated in proper way. Otherwise, results are not reliable. The most used method in machine learning is holdout validation. In this technique

data is split in two sets, training, and testing sets. Usually, the split is done in 70 % and 30 % respectively. Training set is used to train classification model and test set is used to see if the trained model can predict the test data's results properly. Holdout method is a pessimistic estimator though since only a portion of the data is given to the algorithm for training. The more instances are left for testing, the higher is the bias of the estimate. However, fewer test samples will result in a wider confidence interval for the accuracy. (Arlot & Celisse 2010; Kohavi 1995)

K-fold cross validation (CV) offers a solution to this problem since it utilizes whole training data without risking the independence of the test set. In CV process, the data is first split into k number of equally sized subsets. Then, the algorithm is trained k times using k-1 subsets for model training and the one remaining subset for validation. Once all iterations have completed and performance calculated for each iteration, all iterations are aggregated to one performance metric. This is then applied to test data to get results of the prediction. This process is illustrated in Figure 9. (Kohavi 1995)

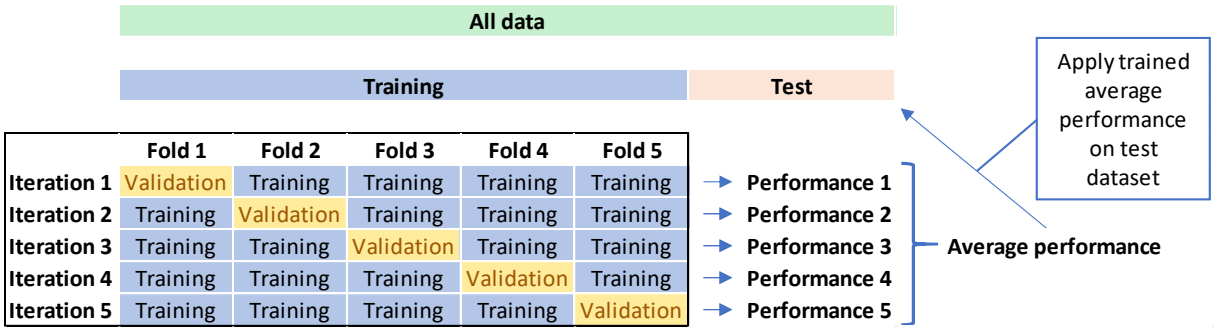



Figure 9. Illustration of 5-fold cross validation

In this research, cross validation is used for classification models. This procedure makes sure that results being interpreted at the end are valid.

## 5.5 Classification models

Classification models are used to predict certain event. In this case models are used as binary classification, which means that there are two outcomes, default, or no default. These models are then used to classify each borrower to either of these classes. Using data available, models should be able to classify borrowers correctly most of the time. In P2P lending there are always some randomness involved in default, which makes it hard to predict. In this thesis three different models are used and are defined next.

### 5.5.1 Logistic Regression (LR)

Regression analysis is a vital component of any data analysis when describing a response variable and one or more explanatory variables. It is often considered as an industry standard model and a benchmark model to beat in credit scoring (Lessmann et al. 2015). What distinguishes a logistic regression model from linear is that the outcome variable is binary or dichotomous in logistic regression. Also, in case of linear regression, parameters are calculated using ordinary least squares estimation, but in logistic regression, parameters are calculated using maximum likelihood estimation. This means that parameters chosen are the most likely values in the used data. (Hosmer et al. 2000)

Logistic regression models the chance of an outcome based on individual characteristics. Because chance is a ratio, the modelled metric is the logarithm of that ratio. This can be calculated as follows: (Sperandei 2013)

$$log(\frac{p}{p-1}) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 \ldots \beta_m x_m \quad (2)$$

where,

p = Probability of event (default)

$\beta_i$ = Regression coefficient associated with the reference group

$x_i$ = Explanatory variable

The advantages of this method are that it does not assume linear relationship between predictor and response variables. Also, it does not require normally distributed variables. Furthermore, multiple variables can be used as predictors. (Keramati et al. 2011; Sperandei 2013)

### 5.5.2 Support Vector Machine (SVM)

A support vector machine is an algorithm that learns by example to classify labels to objects. For instance, SVM is great at recognizing fraudulent credit card behaviour by examining fraudulent and nonfraudulent cases which this thesis is all about. It has been also applied in biomedical field. In essence, SVM is a mathematical algorithm that maximizes a particular mathematical function with respect to the given data. To understand SVMs function, one need to grasp four basic concepts: 1. The separating hyperplane, 2. The maximum-margin hyperplane, 3. The soft margin and 4. The kernel function. (Noble 2006)

Separating hyperplane essentially means that, in high dimensional space, data points are separated with a straight line. This line separates points in this case to defaulters and non-defaulters. Problem is that there can be multiple possible lines between classes. This is where Maximum-margin hyperplane comes in to play. Instead of choosing a line, SVM fits widest bar possible between classes. Once widest bar is found, SVM then picks the middle line of this wide bar. This gives the widest margin and maximises SVMs ability to correctly predict classification of previously unseen samples. So far, assumption was that it is possible to draw a line that separates both classes, but this hardly reflects reality. There are always outliers in the data which stretch beyond drawn separating line. Here soft margin is used to allow few outliers in the data to be beyond the separating hyperplane. The key is to define how many outliers are allowed so that the classification still performs well. Misclassification results in error proportional to the distance between the margin and misclassified datapoint. This error function is known as hinge loss. Figure 10 illustrates how SVM works in simplified way. This figure is drawn from Noble 2006 examples but with more clarity. (Noble 2006; Provost & Fawcett 2013, 92, 94)

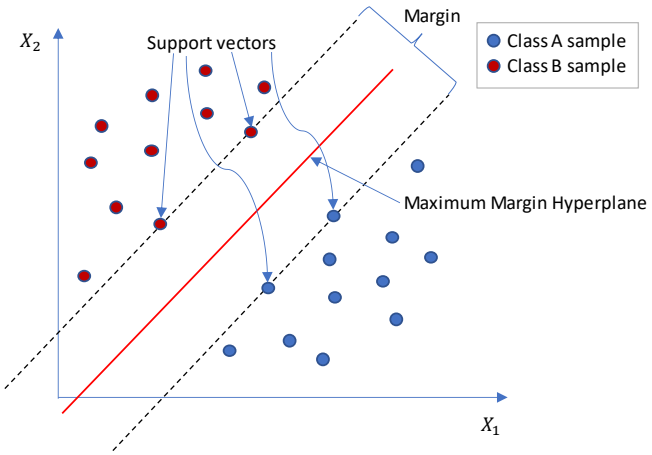



Figure 10. Illustration of simplified SVM

Sometimes there is also an issue, where measurement is done from a single data point. In this case separating hyperplane is also a single datapoint, which makes it impossible for SVM to classify variable. The kernel function provides a solution here. It modifies the data by adding an additional dimension, which is done by simply squaring the original values. This trick allows to change a one-dimensional data into two-dimensional dataset in which the separating hyperplane can be drawn. While this sounds great, too many dimensions result in over-fitted data so dimensions should be as low as possible. Unfortunately, optimal number of dimensions can only be found by trial and error. Using cross-validation can help to determine the optimal kernel function. (Noble 2006)

### 5.5.3 Random Forest (RF)

Random forest is a computationally efficient method to quickly operate with large datasets. It has been used a lot recently in research projects and real-world applications. RF is a combination of randomly produced decision tree predictors. Each tree is drawn at random from a set of possible trees, containing a specified number of attributes at each node. The term “random” means that each tree has the same possibility to be sampled. Random trees can be efficiently created, and once a large combination of trees is constructed, it leads to accurate models if they are aggregated. (Breiman 2001; Oshiro et al. 2012)

RF uses bagging method, which means that it uses different training subsets for each tree. These subsets are created using randomly chosen bootstrap replicates of the original data.

Each new training set is built with replacement from original data. Thus, every tree has new data and different results. These results are then aggregated to create accurate and lower variance results and because of the law of large numbers, this model does not overfit. Also, some of the samples are left out of the bags and these are called out-of-bag observations. These can be used to validate RF model and for comparison. Figure 11 shows simplified illustration of RF basic idea. (Breiman 2001; Oshiro et al. 2012)

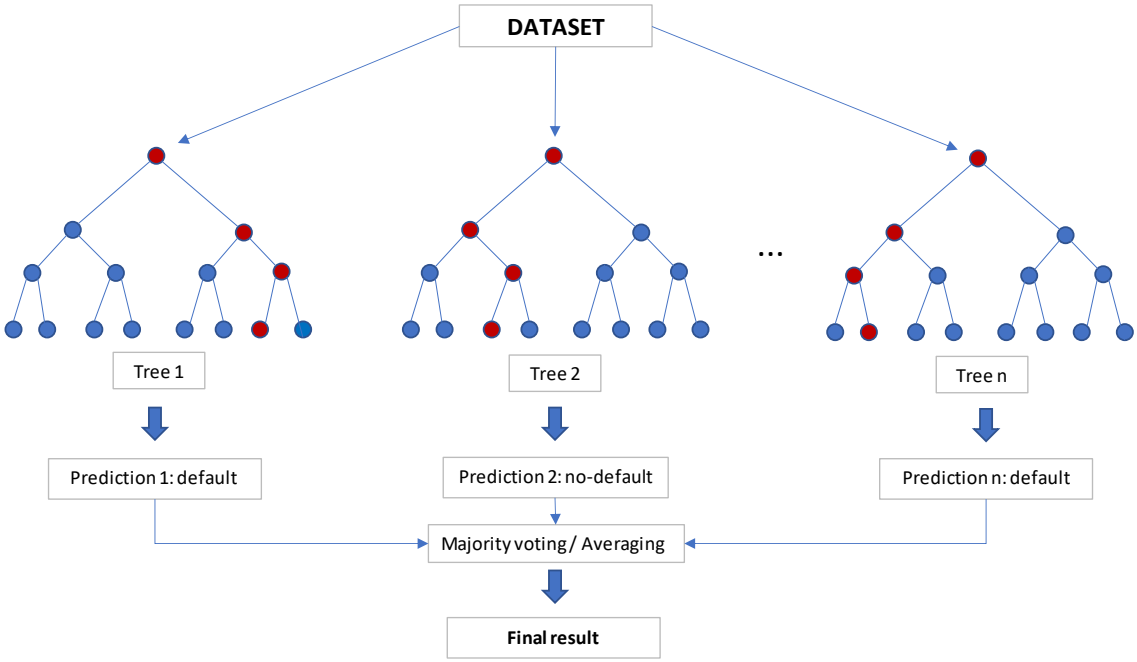



Figure 11. Simplified illustration of Random forests method

## 5.6 Evaluation metrics of classification algorithms

In the field of classifier evaluation, evaluation metrics have received the most attention by far. Accuracy is the most used overused metric to evaluate classifier performance. While it is a good single number metric, it has some imbedded problems within. Evaluation metrics plays an important role when determining the optimal classified during the classification training. Thus, selection of suitable evaluation metrics is vital. For classification problems, confusion matrix is the most used method. From this matrix, many metrics can be derived, and it provides a lot of information. Also, ROC/AUC analysis has received a lot of attention in machine learning

community. All these methods and metrics are unfolded in the next chapters (Hossin et al. 2015; Japkowich & Shah 2014, p.12-13)

### 5.6.1 Confusion Matrix

Like stated before, confusion matrix is used for classification performance evaluation. A confusion matrix is a contingency table showing the differences between actual cases and predicted cases (Bradley 1997). The correctness of classification can be evaluated by computing the number of correctly classified samples by algorithm (true positives, TP), the number of correctly classified samples that do not belong to the class (true negatives, TN), the number of samples that were incorrectly to assigned to class (false positive, FP), and the number of samples that were incorrectly excluded from the class (false negative, FN). These four counts establish the framework of confusion matrix. Figure 12 illustrates how confusion matrix is built. (Sokolova et al. 2009)

| | | Predicted | | |
|---|---|---|---|---|
| | N = 100 | Positive | Negative | Total |
| Actual | Positive | TP = 40 | FN = 10 | 50 |
| | Negative | FP = 5 | TN = 45 | 50 |
| | Total | 45 | 55 | |

Figure 12. Simplified example of confusion matrix

From this matrix, many metrics can be derived that provide more meaningful information of certain performance criteria (Bradley 1997). Most used metric of them all is accuracy. In general, accuracy metric measures the ratio of correctly predicted samples over the total number of samples. Problem with accuracy though is that it does not include a cost to misclassification. Also, there is no way of knowing from this number if the algorithm correctly predicted default. For example, if data is imbalanced, most of the predicted cases can be non-defaulters and default prediction is not successful at all. Accuracy can be calculated from the confusion matrix as follows: (Hossin et al. 2015; Japkowich & Shah 2014, p.12-13)

$$Accuracy = \frac{(TP + TN)}{(TP + FP + FN + TN)} \quad (3)$$

Because accuracy alone is not good enough metric, more information is needed. Calculating misclassification errors from confusion matrix can provide information needed to determine if algorithm was successful in default predictions (Hossin et al. 2015). Type 1 error represents ratio of non-defaulters that were classified as defaulters over actual number of non-defaulters. Type 2 error represents the ratio of defaulters that were misclassified as non-defaulters over actual number of defaulters. Type 2 error is the most important here since misclassifying defaulters as non-defaulters can become very expensive in P2P lending. Misclassification error is the combination of these two. They are calculated from confusion matrix as follows:

$$Type\ 1\ Error\ = \frac{FP}{(FP + TN)} \quad (4)$$

$$Type\ 2\ Error\ = \frac{FN}{(TP + FN)} \quad (5)$$

$$Misclassification\ = \frac{(FP + FN)}{(TP + FP + FN + TN)} \quad (6)$$

True-positive rate, also known as sensitivity or recall, measures the effectiveness of algorithm to correctly detecting default in this case. It calculates the ration of correctly predicted positives instances over all positive samples. Specificity is the complementary metric for sensitivity and measures true-negative rate. It can be calculated, in this case, as predicted non-defaulters over total number of non-defaulters. These metrics are used to see if classes are imbalanced. If they are, these estimates should be skewed. They are calculated from confusion matrix as follows: (Japkowich & Shah 2014, p.95-96)

$$Sensitivity/Recall\ = \frac{TP}{(TP + FN)} \quad (7)$$

$$Specificity\ = \frac{TN}{(TN + FP)} \quad (9)$$

G-mean is also used to check if classes are imbalanced by using both specificity and sensitivity. This metric considers the relative balance of classifier's performance on both positive and negative classes. (Japkowich & Shah 2014, p.100)

$$G - mean = \sqrt{Specificity \times Sensitivity} \quad (11)$$

Finally, focus can be directed to information retrieval from the confusion matrix. Precision is used here in conjunction with sensitivity. These are typical metrics of interest in information retrieval, not only because of relevant information identified, but also in investigating relevant information from class that is labelled as relevant. Precision metric shows how precise algorithm was to identify default in this case. It is calculated as follows: (Japkowich & Shah 2014, p.101)

$$Precision = \frac{TP}{(TP + FP)} \quad (8)$$

Precision and sensitivity are combined to calculate F-measure which is a “single-number measure” that is commonly used to evaluate algorithm performance. It is a weighted harmonic mean of precision and sensitivity. This metric takes considers false-negative predictions which is the costliest part, and the precision of the model which makes it good single number measure. However, this measure excludes true negatives meaning correctly predicting non-defaulters. Though these samples are not very important in this case since default is the phenomena every lender wants to avoid. F-measure can be calculated as follows: (Japkowich & Shah 2014, p.103-104)

$$F - measure = \frac{2 \times Sensitivity \times Precision}{Sensitivity + Precision} \quad (10)$$

### 5.6.2 Area Under the ROC Curve (AUC)

AUC is one of the most popular ranking method and is considered as one of the best single-number metrics for machine learning algorithm evaluation (Bradley 1997). Basically, the measure shows classifier’s ability to avoid false classification. It is derived from Receiving Operating Curve (ROC). It captures a single point from this curve, which is deemed the optimal point. Figure 13 illustrates an example of ROC curve and AUC is considered as the area under the curve. ROC curve is built from the confusion matrix. It plots True positive rate against False positive rate. If classifier’s prediction line goes over the random classifier line, it means it can

predict better than randomly choosing, for example default or no default. ROC curve is mainly used for visualizing different classifiers performance, but AUC value is the one number that should be evaluated. Values of AUC values range from 0 to 1. Higher the value, better the prediction performance, and less misclassification. (Bradley 1997; Sokolova et al. 2009)

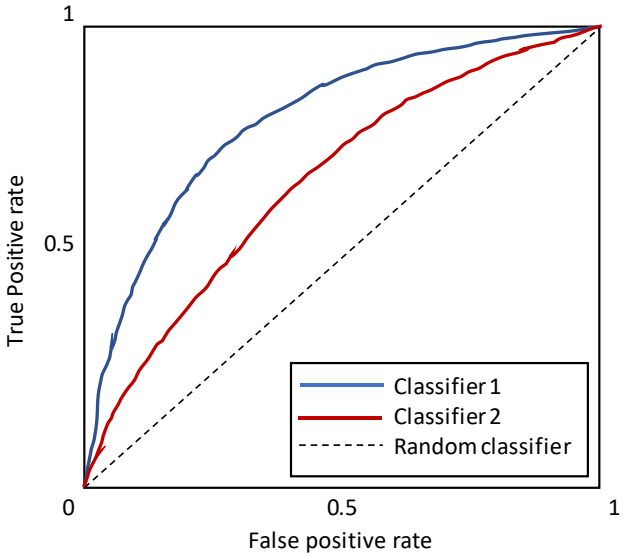



Figure 13. Example of ROC curve

AUC value helps to rank different machine learning algorithms. Also, AUC value is good with imbalanced data since other metrics, such as accuracy, can be strongly biased towards the majority class (Japkowich & Shah 2014, p.129). Though AUC is an excellent metric, it might be computationally expensive when multi class problem is evaluated (Hossin 2015). Fortunately, this thesis works on binary classification problem.

This chapter helped in defining different methods and how to use them in research. This should give sufficient knowledge for further experimenting. In the next chapter, all this knowledge is put to good use and the experimenting with Bondora P2P lending data can begin.

# 6 CASE: IDENTIFYING AND PREDICTING BORROWER DEFAULT AND COMPARING RESULTS BETWEEN COUNTRIES

This chapter consists of empirical evaluation of P2P lending data. All evaluations are done using MATLABs latest version 2021a. Also, some downloadable packages were used in empirical phase like in model building phase used package was Statistics and Machine Learning Toolbox.

## 6.1 Bondora data

The data is acquired from an Estonian peer-to-peer lending platform, Bondora. The company was founded in 2009. It operates in 4 countries, Finland, Spain, Estonia, and Slovakia. Bondora has 742 678 clients and has transacted 355 097 932 € worth of loans. Loan amounts vary from 500 € - 10 000 € and loan maturity varies from 3 - 60 months. (Bondora 2020)

Bondora serves lower- and middle-income borrowers that would otherwise have problems getting a loan from a bank. Bondora has as single digitalized platform where clients can apply and invest regardless of residency, language, and currency. This gives Bondora an extensive economics of scale advantage over traditional banks. Bondora conducts thorough background checks by verifying income and expenses on prospective borrowers. It also uses a combination of transactional, credit bureau and bid data to assess credit and fraud risks quickly and cost effectively so that loan products meet customer needs. (Bondora 2020a)

## 6.2 Data preparation and transformation

The dataset was acquired 25th of February 2020. It consists of 112 variables and 139347 observations in total, and it is cross sectional. There are observations from 4 different countries, Finland, Estonia, Spain, and Slovakia. Most of the people are from Estonia, Finland, and Spain. The data contains information about borrowers' demographics and loan information

such as loan status, default, and loan purpose. The dataset is free and is available for everyone.

The dataset consists of many variables and most of them are useless in this research case. For example, there are various dates, usernames, loanids etc. that do not provide any necessary information. These variables were hand picked out of the data in Excel to make it easier to handle. Removing useless variables resulted in total of 29 variables at this point, but the number is still going to change later in pre-processing phase.

### 6.2.1 Handling missing values and removing samples

Initially, there were a lot of samples in the data, but most of them were removed since they were not concluded yet. In the dataset there is a variable called MaturityDate_Last which tells when loan was matured. Using this variable, loans that are still active, can be removed and only concluded loans are present in evaluation phase. This will make samples more valid. If sample removal was not done, some samples could still be active and on a verge of bankruptcy. Therefore, removal must be made since we do not know the number of these cases in the data. Since the data was gathered at 25$^{th}$ of February 2020, this should be the cut-off date for maturity. All samples after this date should be removed. Sample removal results in a very significant loss of samples but there are still plenty of samples to work on. Sample removal reduced the size of data to 24189 samples from 139347 samples which is very significant loss of information. But the amount is still enough for evaluation.

Next, data should be cleansed from missing values. This was done by removing rows of data that contained missing or NaN values. This procedure resulted in 13064 number of samples from 24189. This should be the last step that results in massive loss of information and 13064 samples are still enough for making good evaluations. When number of samples drop below 100, then information loss starts to really show.

### 6.2.2 Encoding categorical variables

Initially, data contained many categorical variables. These kinds of variables cannot be interpreted in this form. Default date variable needed to be changed as well to binary form since it is the most essential variable in this thesis. This was simply done by renaming variable to “Default” and it gets values of 1, if default occurs, and 0 if not.

There are also many categorical variables with characters that must be changed to binary- or dummy-variables. One-hot-encode function in MATLAB was used to change these variables. For example, use of loan, marital status and employment status were changed to dummy-variables. Also, rating variable needed tweaking since it has character categories. These were changed to numbers instead. After all these procedures, the number of variables is now 70. Mostly several dummy variables from categorical variables increased the total number.

### 6.2.3 Handling outliers and high cardinality in categorical variables

Outliers have significant impact on the performance of machine learning algorithms. For example, when standardizing data, if outliers are present, the standardization can become extremely skewed which results in values that are not representative of the data. Outlier removal results in loss of information since entire rows are deleted. Thankfully, MATLAB has a function called filloutliers() which fills outlier cells with nearest non-outlier values. The function was set to use threshold of 0.1-99.9 %. So, all samples that do not fall in between this percentile are filled with new cell value of the nearest non-outlier value. This method made handling outliers very simple.

Next, high cardinality of categorical variables is handled. Determining, whether a variable has high cardinality is done by using histograms for each variable. If there are some categories that only have few samples, it can cause problems later in the process. Thus, categories that have only few samples, are removed from the data. Fortunately, outlier filling procedure took care of small sample sized categories and resulted in variables that contained only zero. These were easy to spot and removed from dataset. Also, samples that belonged to these small sample sized categories were removed. Furthermore, dataset had a country, Slovakia, that contained only 218 samples. This is not enough since later, the dataset is split to different

countries and further split to training and test subsets. Other ones have plenty of samples but this one is simply too small to be compared validly. These procedures resulted in slightly smaller number of samples of 12833 and 65 variables.

### 6.2.4 Data standardization

Initially, datasets contained many variables with different numerical scales. Different scales will result in distorted estimates which will be very problematic in the evaluation phase. For these reasons, standardization, or in other words, normalization must be made. MATLAB has a function called normalize, which can be specified using ‘range’ to have values between 0 and 1. This also preserves the same distribution or skewness of the data so only the values change to same range. This helps a lot when comparing different variables.

### 6.2.5 Creating sub datasets for each country

Dataset is split to each country in a way that each country is its own sub dataset now. Now there are three sub datasets for each country. Estonian with 6470 samples, Finnish with 2524 samples, and Spanish with 3839 samples. Original dataset is also kept intact for comparison purposes. Also, country variables were now removed from the datasets since they have served their purpose resulting in total of 62 variables.

## 6.3 Descriptive statistics

The final dataset consists of 62 variables, but most of them are dummy variables. Without them the number of variables would be 28, but many of the categorical variables need to be in dummy form to be analysed properly. Target variable default is being examined at Table 4. As one can see, default percentages vary a lot between countries. This is exactly what this thesis is about, recognizing the reasons why these differences are so large. Estonia has the most borrowers in the data. Since the differences in sample sized are significant, balancing should be included. This will be done by using random under sampling method (RUS). Sample sizes for each country should be large enough for evaluation. Also, noteworthy thing is that default rates seem to be very high. This can be due to the reason that many of the borrowers that

borrow through P2P lending has been declined to borrow money via traditional banks. So, naturally default rate should be higher in P2P lending since borrowers tend to have more risk overall. Furthermore, P2P lending sites probably do not have as efficient borrower screening compared to banks, which will result in riskier borrower acceptance.

Table 4. Class frequencies of target variable between countries

| | Default | Count | Percent |
|---|---|---|---|
| Estonia | No | 4102 | 63.40% |
| | Yes | 2368 | 36.60% |
| Finland | No | 825 | 32.69% |
| | Yes | 1699 | 67.31% |
| Spain | No | 904 | 23.55% |
| | Yes | 2935 | 76.45% |
| Total | No | 5831 | 45.44% |
| | Yes | 7002 | 54.56% |

Appendix 1 contains all the variables used in this research. The table gives an overall impression of the used variables and what categories they relate to most. It consists of borrower's demographics, borrower's financials, loan characteristics, and borrower's credit history variable groups. Also, it is noteworthy to mention that 13 out of 28 variables are categorical, which is a relatively high ratio.

The descriptive statistics of numerical variables are in Appendix 2. The average age of the borrower is 37,74 and usually loan appliers do not have any dependants since median value is 0. Loan amount on average is 2354,30 €, so small loans are usually favoured. Loan duration average value is 35,63 which means that most people apply for 3-year loans. Financials of the borrowers seem to be quite weak. As we can see, average total income is 1368,47 € which is quite low. Also, total liabilities are 793,74 € on average which is out of the total income. This leads to free cash average of 503.49. € Some borrower even has almost zero free cash since the min value is 0.25 €. This is also noted in interest rates which is 38.91 % on average. Also, maximum accepted interest rate is 254.84 % which is extremely high. Debt-to-income levels are on acceptable levels though with a mean of 27.73 %. All of this leads to extremely risky

borrower behaviour since there is not much free cash available for payments of the loan. This explains a lot why the default rates are so high.

Credit history of the borrowers has acceptable levels. Many of the borrowers do not even have previous loans as the mean is 0,42. Also, median values of both number, and amount of previous loans are 0. This means that many of Bondora's borrowers are borrowing for the first time. Even though it is good that many borrowers do not have previous loans, this might result in inexperience of borrowing situations. It may come as a surprise that interest rate payments are quite high, and the result is default. On average monthly payment is 127,83 € but the max value is 1377,76 €.

Most of the continuous variables used are skewed or leptokurtic. This can be examined from Appendix 2 skewness and kurtosis numbers and from visualization of Appendix 3 histograms. Most of the financial variables are strongly skewed to the right. This can affect machine learning models since they usually require normal distribution. But in this case SVM, LR and RF are used and none of them require normally distributed data, so skewness will not be an issue.

Most of the categorical variables are already coded as dummy variables and it would be too much to include all dummies in descriptive stats and visualization. So, to analyse descriptive statistics, data before dummy variables is used. This means that outliers have not been removed and some of the high cardinality categories are still present. These categories are identified in the process. Appendix 4 contains all categorical variables, their frequencies, and percentages. Some categories contained just few or no samples so these were removed. These categories were EmploymentStatus: Unemployed with 13 samples, EmploymentDurationCurrentEmployer: Other and Retiree with 0 samples, and HomeOwnershipType: Homeless with 1 sample. Also, Country: Slovakia with 218 samples was removed from final data since it does not have enough samples to be its own dataset as other countries.

Class percentage indicates that 75,80 % are new customers in Bondora. Also, very big portion of borrowers' income has not been verified at all 34,63 %, so screening of borrowers seems to be a big issue in P2P lending. Majority of the borrowers are from Estonia. Home improvement

is the most popular reason to take a loan with 25 %. Most of the samples have passed secondary education 37,81 %. A large majority of borrowers are fully employed with 81,19 % and most of them have also been more than 5 years with current employer 37,24 %. Work experience in years seems to be rather evenly distributed except 5,09 % of the samples have less than 2 years of experience. Most of the borrowers also own a home with 30,18 %. HR (high risk) rating provided by Bondora is the most common rating with 33,36 % which further demonstrates the high-risk-nature of P2P lending.

## 6.4 Balancing of the data using RUS

As we can see from Table 4, countries have different ratios of defaulters, which can become misleading when calculating the metrics. Therefore, random under-sampling was performed. This means that samples that are in majority class are removed so that default and non-default classes are the same size. This will result in loss of information, but it is a better alternative than inaccurate models. There was no straight function for random under-sampling, so MATLAB was told to find majority samples and remove them randomly until it is the same size as minority class. This resulted in data sizes of 4736 samples of Estonians, 1808 samples of Spanish and 1650 samples of Finnish borrowers. All the datasets now have 50/50 ratio of defaulters. Also, whole dataset was sampled.

## 6.5 Feature selection: Chi square

For feature selection, chi square method is used. With feature selection, most of the unnecessary variables can be filtered out of the model to increase performance. This will result in fewer variables but more efficient model. The use of chosen method is justified in chapter 5.1. Chi-square was calculated using MATLAB's function fsschi2(). It constructs feature ranking based on the give data and the target variable. The score given by the function is a logarithm of p-value provided by chi-square test. Sometimes this value is infinite.

The results of chi-square can be seen in Figure 14 for each country and the whole data. The scores are plotted in descending order. As we can see from the graph, the scores vary a lot between countries. For example, Estonia has only two very important variable while Spain and

Finland have multiple. Also, whole data set has one infinite variable which is represented in purple colour. Also, variable importance decreases relatively fast for each dataset, which means that there are only few important variables that contribute to default. 30 variables seem to be the sweet spot for each dataset since the score tends to be very low after 30 variables.

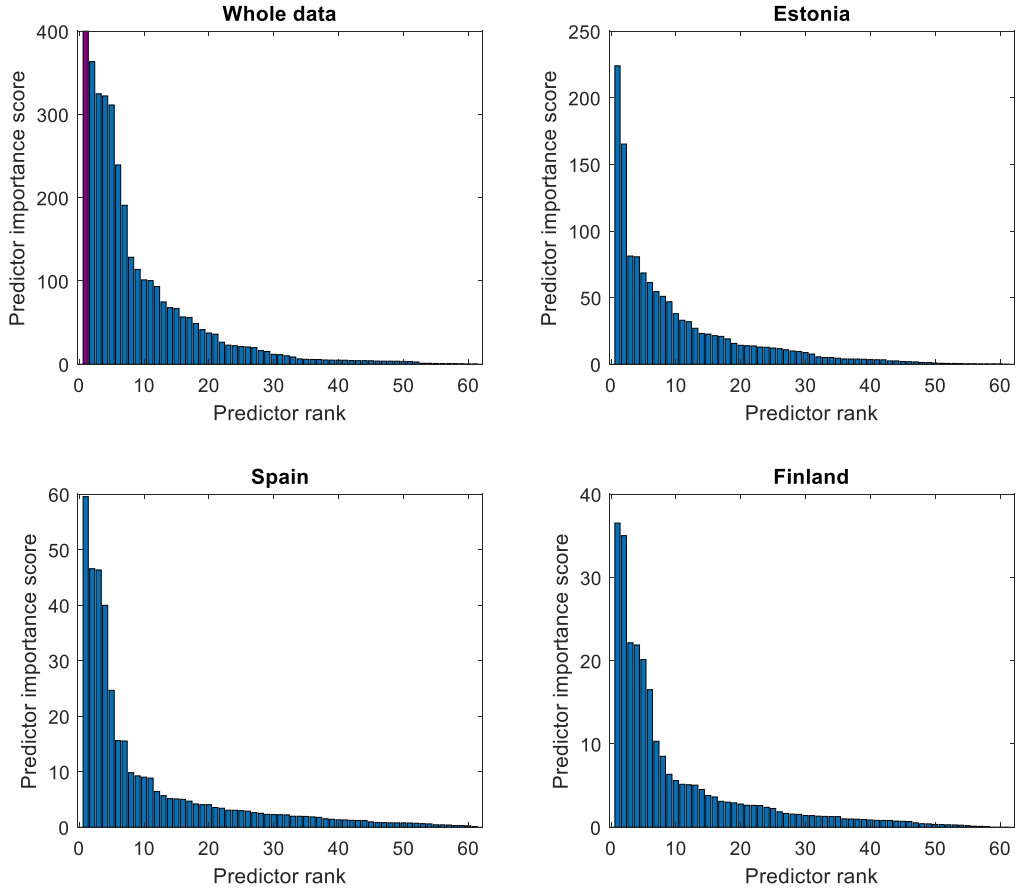



Figure 14. Visualization of Chi-square feature selection scores

Table 5 represents top ten of the most important variables for each country. Same variables pop out in the rankings, but they are in different order. Two most important variables seem to be rating provided by Bondora and monthly payment. Interestingly, Spain does not have rating ranked at top 3. It is at seventh place. Also, reoccurring variables are credit history variables like, number and amount of previous loans and new credit customer. Furthermore, loan characteristics are important too since loan duration and loan amount keep occurring in all datasets. It is curious that interest rate is important on the whole dataset but when divided to countries, it does not even reach top ten. What is more, existing, and/or total liabilities occur frequently in each dataset. From borrower characteristics, gender and education occurs frequently. Only Finland does not have either of these two in top ten. Finland has one peculiar

variable and that is home ownership type: tenant pre-furnished property. My guess would be that this variable contains risky demographic groups, like students, who already have dire monetary situation so borrowing is very risky. Also, loan consolidation is occurring only once in Estonian data, which makes sense since this means that borrower uses the loan to pay other debts which is risky behaviour. Interestingly, income was left out of top ten.

Table 5. Ten most important predictors for each country

| Rank | Whole data | Estonia | Spain | Finland |
|---|---|---|---|---|
| 1 | Rating | Rating | Monthly payment | Monthly payment |
| 2 | Monthly payment | Monthly payment | Number of previous loans | Rating |
| 3 | Number of previous loans | Loan duration | New credit customer | Amount of previous loans before loan |
| 4 | New credit Customer | Refinance liabilities | Amount of previous loans before loan | Number of previous loans |
| 5 | Amount of previous loans before loan | New credit customer | Loan duration | New credit customer |
| 6 | Interest | Number of previous loans | Liabilities total | Refinance liabilities |
| 7 | Gender | Education | Rating | Existing liabilities |
| 8 | Applied amount | Amount of previous loans before loan | Education | Amount |
| 9 | Existing liabilities | Use of loan: Loan consolidation | Amount | HOT: Tenant pre-furnished property |
| 10 | Amount | Existing liabilities | Gender | Loan duration |

Since figure 13 shows that information gain after 30 variables is very small, this could be the number of variables to run in models. There are multiple ways to determine optimal number of variables for each model. But since the point of this research is to examine differences between countries, a fixed number of predictors makes sense. It means that all countries and models have the same starting point model and country differences can be compared in more comprehensive manner.

## 6.6 Data split

The data split is done using cvpartition function in MATLAB. In this function it is possible to specify what method to use. In this case holdout method of 70-30 split was used. 70 % of the data is used for training and hyperparameter optimization and 30 % is used for testing models. Also, stratify command was used to keep the distribution of classes in 50 % in both training and test data. K-fold cross validation is used in model training phase on training data to ensure results validity.

## 6.7 Hyperparameter optimization and model training

Now that the data is pre-processed and features selected, hyperparameters of the model can be optimized and the final specifications for models can be done. In this chapter hyperparameter optimization and training process is described.

### 6.7.1 Logistic regression

Logistic regression was trained using MATLABs function fitclinear. Model was trained using function specifications ‘learner’, ‘logistic’ which specifies that model is using logistic regression. Also, model was trained using hyper parameter optimization which optimized parameters lambda and regularisation. Appendix 6 contains misclassifications of both in-sample and 10-fold cross validation models without and with hyper parameter optimization. There is relatively small difference between in-sample and 10-fold CV misclassification. In-sample misclassification is usually smaller since it has a little bias embedded since the trained model is used to predict class with the same data. Hence, 10-fold CV misclassification is better number to evaluate whether hyper parameter optimization is useful or not. Parameter optimization resulted approximately 1 % better classification in each dataset which is not that much. But since it is still slightly better, 10-fold CV hyper parameter optimized model is used for logistic regression prediction. Appendix 7 contains optimized hyper parameters.

### 6.7.2 Support vector machine

SVM was trained using fitcsvm function. In this function, following command were used to optimize the model a bit better: 'KernelFunction', 'RBF', 'KernelScale', 'auto'. SVM had a problem with hyper parameter optimization since it is much heavier to compute. Iterations of optimization took way too long to be used, so optimizing was not used for this classifier. Appendix 8 contains misclassification of in-sample and 10-fold CV prediction. Now the gap is much more noticeable. Whole data has a gap of 10 % while countries have approximately 20 % between in-sample and 10-fold CV misclassification rate. This means that SVM is a bit overfitted. Hence 10-fold CV model is used in evaluation phase.

### 6.7.3 Random Forests

Random forests model was trained using MATLAB's fitcensemble function. This uses random forests as default, so it was not needed to specify it beforehand. This method has many parameters that can be optimized. Selected ones are as follows: Number of learning cycles, Number of variables to samples, minimum leaf size, maximum number of splits, split criterion. Appendix 9 contains misclassification of in-sample and 10-fold CV for optimized and non-optimized models. Here we can see that in-sample of non-optimized model perfectly captures the training data in Spanish and Finnish dataset and almost perfectly in Estonian data. This is concerning in terms of overfitting, but the 10-fold CV model performs more realistically. Also, hyper parameter optimization provided better results compared to non-optimized 10-fold CV model. It improved whole data, Estonian, and Spanish data model predictions by 2 %. But most change was in Finnish dataset which improved by 6 % approximately. Hyper parameter optimization seems to improve RFs performance very well. Therefore, 10-fold CV parameter optimized model is used in evaluation phase. Appendix 10 contains the information of optimized hyper parameters for each country.

## 6.8 Evaluation of the models and countries predictions

Now that models are trained properly, the actual predictions of the test datasets can begin. Prediction results are calculated using confusion matrix. All numbers are derived from that. All used evaluation metrics are introduced in chapter 5.5.1 and 5.5.2. Good measures to keep an eye on are, accuracy, sensitivity, type 2 error, and AUC. Accuracy is a good overall measure that gives indication of the model performance with one glance. Sensitivity is important since it considers true positives of the confusion matrix, meaning it considers predicted defaults. Type 2 error is also important since this is the costly mistake of the prediction model, meaning samples predicted as good when they actually default. AUC is the most important performance measure since it considers true positives and false positives.

Table 6 contains performance measures of logistic regression for each country. All these measures seem to decrease while data set is further broken down to countries with least samples. Estonia has the most and Finland the least samples. It seems that country with most samples get the best prediction results. This makes sense since supervised prediction models are better if they are fed more samples. But even though the differences of default rates were

large in original data between countries (Table 4), these models show that they can be predicted to somewhat similar level. Finland and Estonia have the largest difference in prediction performance, but Spanish model seem to be relatively close to Estonian level. Whole data set predicts the best though. Estonia has smallest type 2 error and highest sensitivity which is interesting. This means that this model has predicted defaults the best and has smallest error to wrongly assign non-default class to actual defaulter.

Table 6. Evaluation metrics of logistic regression

| | Logistic regression (LR) | | | |
|---|---|---|---|---|
| **Evaluation metrics** | **Whole data** | **Estonia** | **Spain** | **Finland** |
| **Accuracy** | 0.697 | 0.667 | 0.627 | 0.582 |
| **Type 1 error** | 0.264 | 0.332 | 0.339 | 0.441 |
| **Type 2 error** | 0.342 | 0.334 | 0.406 | 0.395 |
| **Misclassification** | 0.303 | 0.333 | 0.373 | 0.418 |
| **Sensitivity** | 0.658 | 0.666 | 0.594 | 0.605 |
| **Specificity** | 0.736 | 0.668 | 0.661 | 0.559 |
| **G-mean** | 0.696 | 0.667 | 0.626 | 0.581 |
| **Precision** | 0.713 | 0.667 | 0.636 | 0.579 |
| **F-measure** | 0.684 | 0.667 | 0.615 | 0.592 |
| **AUC** | 0.761 | 0.728 | 0.663 | 0.606 |

Table 7 contains evaluation metrics of SVM for each country and whole data set. Same pattern appears here that the dataset with most samples has best overall prediction performance in terms of accuracy and AUC. SVM seems to be a lot better when it is run on whole dataset. Dividing it to different countries reduces its prediction capabilities significantly.

Table 7. Evaluation metrics of SVM

| | Support vector machine (SVM) | | | |
|---|---|---|---|---|
| **Evaluation metrics** | **Whole data** | **Estonia** | **Spain** | **Finland** |
| **Accuracy** | 0.700 | 0.631 | 0.616 | 0.558 |
| **Type 1 error** | 0.272 | 0.356 | 0.387 | 0.466 |
| **Type 2 error** | 0.328 | 0.382 | 0.380 | 0.419 |
| **Misclassification** | 0.300 | 0.369 | 0.384 | 0.442 |
| **Sensitivity** | 0.672 | 0.618 | 0.620 | 0.581 |
| **Specificity** | 0.728 | 0.644 | 0.613 | 0.534 |
| **G-mean** | 0.700 | 0.631 | 0.616 | 0.557 |
| **Precision** | 0.712 | 0.634 | 0.615 | 0.556 |
| **F-measure** | 0.691 | 0.626 | 0.618 | 0.568 |
| **AUC** | 0.758 | 0.689 | 0.634 | 0.597 |

Table 8 contains evaluation metrics from random forests model. Accuracy and AUC metrics are the same as previously. Whole dataset gets the best prediction performance, and it decreases as the dataset changes to a country that has fewer samples. Interestingly models sustain higher level of prediction performance when RF is used. Furthermore, sensitivity is the highest in Finnish dataset which is intriguing. It has the lowest sample size, and it had the highest default rate originally before data sampling. Also, type 2 error is lowest in Finnish dataset. This means that RF model seems to predict default class best on Finnish dataset. This is an interesting finding since the assumption is that bigger dataset would fare better in prediction of every class. More importantly it is precisely default we want to avoid and specifying data to regions seems to help it at least in this model's case. This gives a small indication that specifying countries in model prediction might be beneficial. Bigger sample size might even make the differences clearer.

Table 8. Evaluation metrics of random forests

| | Random forests (RF) | | | |
|---|---|---|---|---|
| **Evaluation metrics** | **Whole data** | **Estonia** | **Spain** | **Finland** |
| **Accuracy** | 0.698 | 0.656 | 0.629 | 0.626 |
| **Type 1 error** | 0.281 | 0.337 | 0.373 | 0.437 |
| **Type 2 error** | 0.322 | 0.352 | 0.369 | 0.310 |
| **Misclassification** | 0.302 | 0.344 | 0.371 | 0.374 |
| **Sensitivity** | 0.678 | 0.648 | 0.631 | 0.690 |
| **Specificity** | 0.719 | 0.663 | 0.627 | 0.563 |
| **G-mean** | 0.698 | 0.656 | 0.629 | 0.623 |
| **Precision** | 0.707 | 0.658 | 0.629 | 0.613 |
| **F-measure** | 0.692 | 0.653 | 0.630 | 0.649 |
| **AUC** | 0.781 | 0.723 | 0.700 | 0.666 |

Now that each classifier is evaluated independently, all models are compared to decide which one performs the best. Table 9 contains three previous tables in one, so it is easier to spot performance differences. Overall, when looking all values, random forests model is the best one. It has best overall values in many evaluation metrics and in each country dataset. Also, all metrics seem to be more stable between countries when using random forests. Other models seem to vary a bit more. Interestingly, logistic regression model for Estonian dataset seem to beat random forest model for Estonian dataset completely. Furthermore, Best values for each evaluation metric is bolded so it is easier to spot differences. In terms of AUC, and F-measure RF performs the best. These metrics were the best "single number" metrics to

evaluate to determine the best prediction model. So, from these three methods, RF is the king in stability between country datasets and in best metrics.

Table 9. Evaluation metrics for all models

| | Logistic regression (LR) | | | | Support vector machine (SVM) | | | | Random forests (RF) | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **Evaluation metrics** | **Whole data** | **Estonia** | **Spain** | **Finland** | **Whole data** | **Estonia** | **Spain** | **Finland** | **Whole data** | **Estonia** | **Spain** | **Finland** |
| **Accuracy** | 0.697 | 0.667 | 0.627 | 0.582 | **0.700** | 0.631 | 0.616 | 0.558 | 0.698 | 0.656 | 0.629 | 0.626 |
| **Type 1 error** | **0.264** | 0.332 | 0.339 | 0.441 | 0.272 | 0.356 | 0.387 | 0.466 | 0.281 | 0.337 | 0.373 | 0.437 |
| **Type 2 error** | 0.342 | 0.334 | 0.406 | 0.395 | 0.328 | 0.382 | 0.380 | 0.419 | 0.322 | 0.352 | 0.369 | **0.310** |
| **Misclassification** | 0.303 | 0.333 | 0.373 | 0.418 | **0.300** | 0.369 | 0.384 | 0.442 | 0.302 | 0.344 | 0.371 | 0.374 |
| **Sensitivity** | 0.658 | 0.666 | 0.594 | 0.605 | 0.672 | 0.618 | 0.620 | 0.581 | 0.678 | 0.648 | 0.631 | **0.690** |
| **Specificity** | **0.736** | 0.668 | 0.661 | 0.559 | 0.728 | 0.644 | 0.613 | 0.534 | 0.719 | 0.663 | 0.627 | 0.563 |
| **G-mean** | 0.696 | 0.667 | 0.626 | 0.581 | **0.700** | 0.631 | 0.616 | 0.557 | 0.698 | 0.656 | 0.629 | 0.623 |
| **Precision** | **0.713** | 0.667 | 0.636 | 0.579 | 0.712 | 0.634 | 0.615 | 0.556 | 0.707 | 0.658 | 0.629 | 0.613 |
| **F-measure** | 0.684 | 0.667 | 0.615 | 0.592 | 0.691 | 0.626 | 0.618 | 0.568 | **0.692** | 0.653 | 0.630 | 0.649 |
| **AUC** | 0.761 | 0.728 | 0.663 | 0.606 | 0.758 | 0.689 | 0.634 | 0.597 | **0.781** | 0.723 | 0.700 | 0.666 |

## 6.9 Determinants of default

It is important to recognize variables that has the most effect on default probability. But it is necessary to know how they effect on default probability. For this reason, logistic regression was run on all datasets using glm function on MATLAB which provides estimates, standard deviation, t-statistic, and statistical significance (p Value) for each variable. The function was run on sampled datasets, also including feature selection so these datasets contain 30 most important variables. These results can be found from appendices 11-13.

Only estimates, p values and 10 most important variables, chosen by feature selection, are shown in table 10. Bolded variables are selected to top-10 in each dataset. There are 6 reoccurring variables, and these are: Rating, Monthly payment, No. of previous loans, New credit customer, Amount of previous loans before loan, and Existing liabilities. As expected, worse rating increases the probability of default a lot since its estimates are all positive and relatively high compared to other estimates. Existing liabilities variable also makes sense, since the more liabilities you have, less money you have for interest payments, so default probability increases.

Table 10. 10 most important variables for each dataset

| Whole data | | | Estonia | | |
|---|---|---|---|---|---|
| | Estimate | pValue | | Estimate | pValue |
| **Rating** | 2.752 | 0.000 | **Rating** | 2.281 | 0.000 |
| **MonthlyPayment** | -1.068 | 0.001 | **MonthlyPayment** | -2.924 | 0.000 |
| **NoOfPreviousLoansBeforeLoan** | -3.100 | 0.000 | LoanDuration | 0.374 | 0.004 |
| **NewCreditCustomer** | 0.306 | 0.000 | RefinanceLiabilities | 0.946 | 0.051 |
| **AmountOfPreviousLoansBeforeLoan** | 1.126 | 0.001 | **NewCreditCustomer** | 0.177 | 0.097 |
| Interest | 0.136 | 0.413 | **NoOfPreviousLoansBeforeLoan** | -2.167 | 0.000 |
| Gender | 0.145 | 0.038 | Education | -0.552 | 0.000 |
| AppliedAmount | 0.177 | 0.494 | **AmountOfPreviousLoansBeforeLoan** | 1.203 | 0.001 |
| Amount | 0.333 | 0.204 | **ExistingLiabilities** | 1.520 | 0.000 |
| **ExistingLiabilities** | 0.447 | 0.100 | UOL_Loan_consolidation | -0.209 | 0.052 |
| | | | | | |
| **Spain** | | | **Finland** | | |
| **MonthlyPayment** | -0.017 | 0.979 | **MonthlyPayment** | -1.479 | 0.037 |
| **NewCreditCustomer** | 0.269 | 0.244 | **Rating** | 1.473 | 0.000 |
| **NoOfPreviousLoansBeforeLoan** | -6.423 | 0.001 | **NoOfPreviousLoansBeforeLoan** | -6.924 | 0.012 |
| **AmountOfPreviousLoansBeforeLoan** | 2.306 | 0.089 | **AmountOfPreviousLoansBeforeLoan** | 1.151 | 0.360 |
| LoanDuration | 0.302 | 0.124 | **NewCreditCustomer** | -0.107 | 0.729 |
| **ExistingLiabilities** | 1.697 | 0.019 | RefinanceLiabilities | 2.188 | 0.002 |
| Gender | 0.406 | 0.002 | Amount | 0.484 | 0.423 |
| VerificationType | -0.424 | 0.002 | AppliedAmount | 0.251 | 0.670 |
| **Rating** | 1.808 | 0.000 | **ExistingLiabilities** | 0.035 | 0.952 |
| LiabilitiesTotal | -2.210 | 0.071 | VerificationType | -0.379 | 0.014 |

Rest of the main six variables have interesting effects on default. Monthly payment variable has decreasing effect on probability of default which sounds flawed. Since more payments you make every month, should result in bigger difficulty to survive from all liabilities. But, in this case it seems to be the opposite. Interestingly, amount and interest variable, which should increase the monthly payment, have the opposite effect. They increase the probability of default. This seems very conflicted, so there might be something more to monthly payment variable than initially thought. Furthermore, chi-square feature selection failed in this variable for Spanish dataset. This should not be the most important variable since the estimate is almost zero and p value is near one which means it is far from statistical significance. Also, very interesting variable is No. of previous loans before loan since it has a negative effect on default probability. So, more loans you have had decreases chances to default. This could be explained with borrowing experience as mentioned before in feature selection chapter. More experience you have with borrowing, more likely you know how to get through of all payments. This is also backed up with other important variable, New credit customer which has positive estimate. Although, it is not statistically significant variable in country datasets. This variable means that borrower is new to Bondora, and the inexperience leads to higher probability of default. Amount of previous loans before loan is intriguing as well. Since the estimate is

positive, this variable has an increasing effect on default probability. Which is kind of odd since number of loans had opposite effect. But maybe if the amount of previous loans is very high, risky borrowing behaviour is much more probable.

Some variables, that are not reoccurring, have also interesting effects. Gender is selected as important in whole data and Spanish dataset. In both, it seems that female gender has higher probability of default since female = 1 and male = 0 in this variable. In Spanish data it is somewhat expected since the culture is more patriarchal, meaning men do financials more, so women might have less experience. When compared to Finnish and Estonian data, Gender variable is at the bottom of 30 important variables, which indicate that these countries have more gender equality. Also, in both datasets, gender variable has negative estimate, meaning that higher value decreases default risk. Which means that women default less. This is more common result in countries that have more gender equality since, in general men have riskier behaviour compared to women. So, gender variable occurring in whole datasets top 10 is because of Spanish datasets importance for variable.

Interestingly, interest, applied amount and amount variables, which all indicate riskier borrowing, are not statistically significant variables in any datasets. Education variable also takes logical path, which is more education = lower default probability. Verification type variable indicates more verified income with higher values. This makes sense as well since the estimate is negative, meaning if income is more verified, default probabilities decrease. Loan duration appears in many datasets as important variable, and it also has logical effect. Since estimate is positive, it means that longer loan duration leads to increased probability of default. Last variable is use of loan: Loan consolidation which only appears in Estonian data. It has interesting effect since this variable means that borrower finances debts with more debts. One might think that this is very risky borrower behaviour, but instead it has negative estimate, meaning it decreases default probability.

Same variables tend to appear in different countries datasets. Still, there are differences in the ordering of most important variables and there are some interesting variables that only appear in one country's dataset. This indicates that countries indeed have differences, and it might be beneficial to evaluate them separately to get better results in prediction.

# 7 CONCLUSIONS

In this thesis, P2P lending was researched. The topic was to identify defaulting borrowers and predicting them using machine learning. First, P2P lending was introduced, showing its popularity around the world, and describing its pros and cons. Then, machine learning was introduced to help lenders to make investing decisions. With machine learning one can create predictive tools to help identify borrowing behaviour in P2P lending which already is riskier compared to traditional bank lending. Then, literature review was conducted to get a better understanding of recent research. This chapter helped to identify good prediction models and practices to get most of learning algorithms. After that, used methods were chosen based on the previous research. These methods are described in detail to get a better understanding of their use. Finally, empirical research was conducted on a real-world dataset provided by Bondora P2P lending site. The dataset was split to multiple countries so that each country could be compared if prediction algorithms can perform better when models are trained with specific country's dataset rather than whole data. Prediction methods used were Logistic regression, Support vector machine and Random forests.

According to empirical research of this thesis, default prediction with machine learning can be very effective. Best AUC value was 0,781, which is considered as best prediction evaluation metric, was achieved with random forest method. This can be considered as very good AUC value since previous research had approximately lower AUC values. This model was built using feature selection, data balancing and hyper parameter optimization, so many boosting procedures were applied. Overall random forests method provided best results since countries evaluation metrics varied a lot less compared to other methods. Despite sample size being a lot smaller in Spanish and Finnish datasets, their evaluation metrics were relatively close to largest Estonian dataset when using random forests.

Country comparison was not as successful as first thought. It seems that dataset size has considerable effect on evaluation metrics, since throughout the evaluation phase, dataset with most samples had best performance across all models and tests. This was a problem since it makes it hard to determine whether the sample size was the main reason for prediction difference or was it just that a certain country can be predicted better than the rest. This could have been resolved by sampling datasets to be the same size, but then there would have been a significant information loss. Initially I thought that data balancing would solve this issue, but

it seems supervised learning models get best results when there are more samples to evaluate. Larger data could have made the difference.

But there were some interesting findings in prediction comparison between countries. For example, when dataset was trained using random forests, Finnish dataset had highest sensitivity evaluation metric of all models and datasets, which calculates the true positive rate of the prediction model. This means that this model performed on Finnish data can predict default the best from all other examined models and countries. This is very interesting finding since this model and dataset got these metrics while having the least number of samples of all datasets. Which indicates that there could be benefits in predicting each country separately, given enough data. Also, there were only 3 different countries to analyse which is quite small number. Furthermore, Estonia and Finland are very similar when considering culture. It would have been interesting to compare completely different cultures predicting capabilities.

Finally, the determinants of default were analysed for each country separately. Results indicate that there are many similarities between each country. For example, in top-10 list of each country's most important variables, six were the same. Although, ordering of these features varies a lot between countries. But there were significant differences, for example, gender variable was important in Spanish dataset but not that meaningful in Estonian or Finnish data. There were other variables as well that were ranked completely differently, and this suggests that countries should be evaluated separately when predicting default.

This thesis taught that predicting default can be very useful. From a 50/50 default rate dataset it was possible to get default prediction accuracy to 70 % which is at a good level. Also, country comparison is much more complicated than initially though. To make it valid, all countries should have same sample size and default rate. Also, some predictor variables had very different effects what one might expect. So, prejudice of variable effects should be kept to minimum.

## 7.1 Answering research questions

*"What has been previously researched in literature?"*

This Literature review was divided in three parts: Credit scoring in general, determinants of default, and credit scoring in P2P lending. All these chapters provided a lot of insight how models should be constructed and what the most important variables are. In the first part, balancing the data and feature selection had a positive impact on prediction performance. Most used methods were SVM (7 articles), RF (6 articles), and LR, NN and k-NN (4 articles) according to Table 1. Most popular and simple models to use are SVM, RF and LR. These models also provide good prediction performance.

In the second part factors that explain default in P2P lending were examined. Most reoccurring variables were interest rate, credit score assigned by the platforms, loan amount, debt to income ratio, credit history, and longer loan period. All these variables are financial variables. Alternative data, such as demographic and psychological variables, were proven to increase predictive performance. Low default risk demographic characteristics were female gender, young adults, long working time, stable marital status, high educational level, working in large company, and loan purpose.

In final part, data was usually imbalanced. Balancing had somewhat mixed results in predictive performance to these most used algorithms but mostly it enhances algorithms capabilities. Only LR seems to get worse, and GB (gradient boosting) remained the same, while others improved. But on average, balancing improved algorithms. RUS (random under-sampling) technique provided best overall results. Five most used models in this chapter were LR (10 articles), RF (9 articles), GB (7 articles), and SVM (5 articles) according to Table 3. Popularity of models are very similar with results from the first part. In terms of AUC score in most popular models using balanced data, RF performed the best with 69.2 %, SVM and GB are tied to second spot with 69.0 %, and LR is third with 67.1 %.

*"What are the differences in country borrower populations and default predictability?"*

Country borrower populations were different to some extent. When conducting feature selection, 30 most important variables were selected for each country. In the top-10 variables in Table 5, six were the same in each country so similarities can be found. Although, these features were in different order in terms of relevancy determined by chi-square. There were also some unique features in each country which also indicates that there are differences. Furthermore, each country had large differences in default rate of borrowers. This suggests that each country has own specifics and should be predicted separately.

Balancing of each country was done using RUS which resulted in same ratios of defaulters. This was not enough though since sample size seems to have significant impact on predictability. Default predictability for each country was not very comparable for this reason. But there were some interesting default predictability differences, for example, Finnish dataset with RF model had the highest default prediction rate even though it had smallest sample size. This indicates that default predictability indeed varies between countries and running prediction algorithms for specific countries could be beneficial, given enough data. Although, whole dataset got the best prediction performance metrics overall, which makes sense since supervised learning algorithms tends to perform better when there are more samples.

*"Are there identifiable characteristics that explain borrower default?"*

Yes. These characteristics are mostly loan related. For example, rating, credit history, and monthly payment variables were the most reoccurring variables between countries. Worse rating increases default probability. Interestingly, credit history variables have negative effect on default probability which could be explained with experience in borrowing liabilities. Monthly payment variable had negative effect on default probability which was strange. Also, each country had some unique variables that had completely different effect. For example, Spain had gender as top-10 variables and it had also different sign in estimate, which means it has opposite effect on default probability compared to Estonian and Finnish datasets.

## 7.2 Further research possibilities

Further research can be done on this subject since this thesis did not fully explain the country comparison aspect. It only gave an indication that it could be possible to get better prediction performance when training and testing for each country separately. It can be thoroughly examined with better data and same sample sizes for each country. Doing this would solve this issue and give a straight answer on this topic. Also, using data that has countries from very different cultures would give better results on possible prediction differences but in my opinion this data should come from the same lending platform since the variables would be the same for each country. Using very international P2P lending platforms data would be sufficient for this task.

Other prediction algorithms should be evaluated as well. P2P prediction field mostly consists of supervised machine learning techniques but using unsupervised techniques is not that common. Using unsupervised methods might lead to completely different results and it would be very interesting to see if they can perform better than supervised learning methods.

# 8 REFERENCES


Agosto, A., Giudici, P & Leach, T. (2019) Spatial Regression Models to Improve P2P Credit Risk Management. Front. Artif. Intell. 2:6. Doi: 10.3389/frai.2019.00006

Ariza-Garzón, M. J., Arroyo, J, Caparrini, A., & Segovia-Vargas, M-J (2020) Explainability of a Machine Learning Granting Scoring Model in Peer-to-Peer Lending, IEEE Access, vol. 8, pp. 64873-64890

Arlot, S., Celisse, A. (2010). A survey of cross-validation procedures for model selection. Statistics surveys, vol. 4, pp. 40-79.

Atz, U. & Bholat, D. (2016). Peer-to-peer lending and financial innovation in the United Kingdom. Staff Working Paper No. 598, Bank of England.

Bastani, K., Asgari, E., & Namavari, H. (2019). Wide and deep learning for peer-to-peer lending. Expert Systems with Applications, vol. 134, pp. 209–224.

Bell, J. (2020) Machine Learning: Hands-On for Developers and Technical Professionals. Newark: John Wiley & Sons, Incorporated.

Boughaci, D., Alkhawaldeh, A., Jaber, J., & Hamadneh, N. (2020). Classification with segmentation for credit scoring and bankruptcy prediction. Empirical Economics.

Bondora 2020. Background information about Bondora. Accessed 25.2.2020. Available https://support.bondora.com/hc/enus/articles/212499589-Background-information-about-Bondora

Bondora 2020a. Background information about Bondora. Accessed 25.2.2020. Available https://support.bondora.com/hc/en-us/articles/213073845-Competitive-advantages-of-Bondora

Bondora 2020b. In a nutshell: Bondora p2p lending platform benefits. Accessed 19.8.2020. Available https://www.bondora.com/en/peer-to-peer-lending

Bondora 2020c. Fees. Accessed 24.8.2020. Available https://www.bondora.fi/en/fees/

Bradley, A. P. (1997). The Use of the Area Uncer the ROC Curve in the Evaluation of Machine

Learning Algorithms. Pattern Recognition, vol. 30, no. 7, pp. 1145-1159.

Breiman, L. (2001) Random Forests. Machine Learning, vol. 45, no. 1, pp. 5-32.

Brown, I., & Mues, C. (2012). An experimental comparison of classification algorithms for imbalanced credit scoring data sets. Expert Systems with Applications, vol. 39, no. 3, pp. 3446–3453.

Callahan, J. L. (2014) Writing Literature Reviews: A Reprise and Update, Human Resource Development Review, vol. 13(3), pp. 271–275. Doi: 10.1177/1534484314536705

Cai, S  Zhang, J. (2020) Exploration of credit risk of P2P platform based on data mining technology. Journal of Computational and Applied Mathematics, vol. 372

Chen, C., Dong, M., Liu, N., & Sriboonchitta, S. (2019). Inferences of default risk and borrower characteristics on P2P lending. The North American Journal of Economics and Finance, vol. 50, pp. 101013

Cho, P., Chang, W., Song J.-W. (2019) Application of instance-based entropy fuzzy support vector machine in peer-to-peer lending investment decision, IEEE Access, vol. 7

Crotty, J. (2009) Structural causes of the global financial crisis: a critical assessment of the 'new financial architecture'. Cambridge Journal of Economics, vol. 33, no. 4, pp. 563-580.

Davis, K., Murphy, J. (2016) Peer-to-Peer Lending: Structures, risks and regulation. JASSA: The Finsia Journal of Applied Finance, no. 3, pp. 37-44.

Dastile, X., Celik, T., & Potsane, M. (2020). Statistical and machine learning models in credit scoring: A systematic literature survey. Applied Soft Computing, vol. 91, pp. 106263–.

Durand, D. (1941). Risk elements in consumer instalment financing, in: National Bureau of Economics, New York.

Emekter, R., Tu, Y., Jirasakuldech, B., & Lu, M. (2015) Evaluating credit risk and loan performance in online Peer-to-Peer (P2P) lending. *Applied Economics*, vol. 47, no. 1, pp. 54-70.

Granovetter, M. (2012) Economic Action and Social Structure: The Problem of Embeddedness. *American Journal of Sociology*, vol. 91, No. 3, pp. 481-510

Garvey, K., Zhang, B., Ralston, D., Ying, K., Maddock, R., Chen, H., Buckingham, E., Katiforis, Y., Deer, L & Ziegler, T. (2017) Cumulative Growth: The 2nd Asia Pacific Region Alternative Finance Industry Report. Cambridge Centre for Alternative Finance, Cambridge.

Gavurova, B., Dujcak, M., Kovac, V., Kotaskova, A. (2018) Determinants of successful loan application at peer-to-peer lending market. *Economics & Sociology*, vol. 11, no. 1, pp. 85-99.

Harris, T. (2013). Quantitative credit risk assessment using support vector machines: Broad versus Narrow default definitions. Expert Systems with Applications, vol. 40, no.11, pp. 4404–4413.

Hosmer, D. W. & Lemeshow, S. (2000). Applied Logistic Regression, 2nd ed. New York: Wiley

Hossin, M. & Sulaiman, M. (2015). A review on evaluation metrics for data classification evaluations. International Journal of Data Mining Knowledge Management Process, vol. 5, no.2

Jagtiani, J. & Lemieux, C. (2019) The roles of alternative data and machine learning in fintech lending: Evidence from the LendingClub consumer platform. Financial management. Vol. 48, no.4, pp. 1009–1029.

Japkowich, N., Shah, M. 2014. Evaluating Learning Algorithms: A Classification Perspective.

Cambridge University Press, New York.

Jin, Y., & Zhu, Y. (2015) A Data-Driven Approach to Predict Default Risk of Loan for Online Peer-to-Peer (P2P) Lending, in International Conference on Communication Systems and Network Technologies, IEEE xplore.

Kohavi, R. (1995) A Study of Cross-Validation and Bootstrap for Accuracy Estimation and Model Selection. Artificial Intelligence, vol. 14, no. 1-2, pp. 273-324.

Kane, G. C., Alavi, M., Labianca, G., and Borgatti, S. P. (2014). What's Different about Social Media Networks? A Framework and Research Agenda, MIS Quarterly, vol. 38, no. 1, pp. 275-304.

Keramati, A.,Yousefi, N. (2011). A Proposed Classification of Data Mining Techniques in Credit Scoring. Proceedings. 2nd International Conference on Industrial Engineering and Operations Management (IEOM 2011), January, Kuala Lumpur, Malaysia, pp. 416–424.

Klafft, M. (2008). Online Peer-to-Peer Lending: A Lenders' Perspective. Proceedings of the International Conference on E-Learning, E-Business, Enterprise Information Systems, and E-Government, EEE 2008. H. R. Arabnia and A. Bahrami, eds., pp. 371-375, CSREA Press, Las Vegas 2008.

Kruppa, J., Schwarz, A., Arminger, G., & Ziegler, A. (2013). Consumer credit risk: Individual probability estimates using machine learning. Expert Systems with Applications, vol. 40, no. 13, pp. 5125–5131.

Lending Club 2020. Interest Rates and Fees. Accessed 24.8.2020. Available https://www.lendingclub.com/investing/investor-education/interest-rates-and-fees

Lessmann, S., Baesens, B., Seow, H., & Thomas, L. (2015). Benchmarking state-of-the-art classification algorithms for credit scoring: An update of research. European Journal of Operational Research, vol. 247, no. 1, pp. 124–136.

Li, J., Cheng, K., Wang, S., Morstatter, F., Trevino, R., Tang, J., Liu, H. (2018) Feature Selection: A Data Perspective. ACM computing surveys, Vol.50 (6), p.1-45

Li, Q., Chen, L., & Zeng, Y. (2018). The mechanism and effectiveness of credit scoring of P2P lending platform. China Finance Review International, vol.8, no. 3, pp. 256–274.

Li, W., Ding, S., Chen, Y., & Yang, S, (2018) Heterogeneous ensemble for default prediction of peer-to-peer lending in China, IEEE Access, vol. 6, pp. 54396–54406.

Li, Z., Li, K., Yao, X., & Wen, Q. (2019). Predicting Prepayment and Default Risks of Unsecured Consumer Loans in Online Lending. Emerging Markets Finance & Trade, vol. 55, no. 1, pp.118–132.

Lin, X., Li, X., & Zheng, Z. (2017). Evaluating borrower's default risk in peer-to-peer lending: evidence from a lending platform in China. Applied Economics, vol. 49, no.35, pp. 3538–3545.

Liu, D., Brass, D., Lu, Y., Chen, D. (2015) Friendships in Online Peer-to-Peer Lending: Pipes, Prisms, and Relational Herding. *MIS Quarterly*, vol. 39 no. 3, pp. 729-742

Liu, Y., Zhou, Q., Zhao, X., & Wang, Y. (2018). Can Listing Information Indicate Borrower Credit Risk in Online Peer-to-Peer Lending? Emerging Markets Finance & Trade, vol. 54, no. 13, pp. 2982–2994.

Louzada, F., Ara, A., & Fernandes, G. (2016). Classification methods applied to credit scoring: Systematic review and overall comparison. Surveys in Operations Research and Management Science, vol. 21, no. 2, pp. 117–134.

Luo, S., Cheng, B., & Hsieh, C. (2009). Prediction model building with clustering-launched classification and support vector machines in credit scoring. Expert Systems with Applications, vol. 36, no. 4, pp. 7562–7566.

Malekipirbazari, M., & Aksakalli, V. (2015). Risk assessment in social lending via random forests. Expert Systems with Applications, vol. 42, no. 10, pp. 4621–4631.

McHugh, M.L. 2013. The Chi-Square test of independence. Biomechia Medica, vol. 23, no. 2,

pp. 143-149.

Moscato, V., Picariello, A., & Sperlí, G. (2021). A benchmark of machine learning approaches for credit score prediction. Expert Systems with Applications, vol. 165, pp-113986–.

Namvar, E. (2013). An Introduction to Peer to Peer Loans as Investments. Journal of Investment Management First Quarter, 2014. Pp. 1-20.

Noble, W. S. (2006). What is a support vector machine? Nat. Biotechnol., vol. 24, no. 12, pp. 1565-1567

Oshiro, T. M., Perez, P. S. and Baranauskas, J. A. (2012) How many trees in a random forest? In Machine Learning and Data Mining in Pattern Recognition: 8th International Conference, MLDM 2012, Berlin, Germany, July 13-20, 2012, Proceedings, vol. 7376, 154. Springer.

Oestreicher-Singer, G., and Sundararajan, A. 2012. The Visible Hand? Demand Effects of Recommendation Networks in Electronic Markets. Management Science (58:11), pp. 1963-1981.

Pans, S & Zhou, S. (2019) Evaluation Research of Credit Risk on P2P Lending based on Random Forest and Visual Graph Model. Journal of Visual Communication and Image Representation, pp. 102680

Pławiak, P., Abdar, M., & Rajendra Acharya, U. (2019). Application of new deep genetic cascade ensemble of SVM classifiers to predict the Australian credit scoring. Applied Soft Computing, vol. 84, pp. 105740–.

Provost, F., Fawcett, T. (2013) Data Science for Business: What you need to know about data mining and data-analytic thinking. O'Reilly Media, Inc., Sebastopol.

Rowley, J & Slack, F. (2004). Conducting a literature review. Management research news, vol. 27, no. 6.

Santoso, W., Trinugroho, I., & Risfandy, T. (2020). What Determine Loan Rate and Default Status in Financial Technology Online Direct Lending? Evidence from Indonesia. Emerging Markets Finance & Trade, vol. 56, no. 2, pp. 351–369.

Serrano-Cinca, C., Gutlerrez-Nieto, B., Lopez-Palacios, L. (2015). Determinants of Default in P2P Lending. *PloS one*, vol. 10, no. 10, pp. 1-22

Serrano-Cinca, C., & Gutiérrez-Nieto, B. (2016). The use of profit scoring as an alternative to credit scoring systems in peer-to-peer (P2P) lending. Decision Support Systems, vol. 89, pp. 113–122.

Singh, p., Uparna, J., Karampourniotis, P., Horvat, E., Szymanski, B., Korniss, G., Bakdash, J., Uzzi, B., & Gallos, L. (2018). Peer-to-peer lending and bias in crowd decision-making, PloS one, vol.13, no. 3, pp.e0193007

Sokolova, M. & Lapalme, G. (2009). A systematic analysis of performance measures for classification tasks. Information Processing and Management, vol 45, pp. 427-437.

Sperandi S. (2013) Understanding logistic regression analysis. Biochem Med, vol. 24(1), pp.12-18.

Teply, P., & Polena, M. (2019). Best classification algorithms in peer-to-peer lending. The North American Journal of Economics and Finance, vol. 3, no. 1

Teply, P., & Polena, M. (2020). Best classification algorithms in peer-to-peer lending. The North American Journal of Economics and Finance, vol. 51, 100904–.

Thomas, L. C., Edelman, D., & Crook, J. (2002) Credit Scoring and its Applications, Monographs on Mathematical Modeling and Computation, SIAM.

Timmins, F, McCabe, C. (2005). How to conduct an effective literature search. Nursing Standard, vol. 20, no. 11, pp. 41-47.

Trivedi, S. (2020). A study on credit scoring 86odelling with different feature selection and machine learning approaches. Technology in Society, vol. 63, 101413–.

Ye, X., Dong, L., Ma, D. (2018). Loan evaluation in P2P lending based on Random Forest optimized by genetic algorithm with profit score, Electronic Commerce Research and Applications, vol. 32, pp. 23-36.

Zanin, L. (2020). Combining multiple probability predictions in the presence of class imbalance to discriminate between potential bad and good borrowers in the peer-to-peer lending market. Journal of Behavioral and Experimental Finance, vol. 25, pp. 100272

Zhao, H., Ge, Y., Liu, Q., Wang, G., Chen, E., Zhang, H. (2017). P2P Lending Survey: Platforms, Recent Advances and Prospects. ACM Transactions on Intelligent Systems and Technology (TIST), vol. 8, no. 6, pp. 1-28.

Ziegler, T., Reedy, E, J., Zhang, B., Kroszner, R., Le, A., & Garvey, K. (2017). Hitting Stride: The 2nd Americas Alternative Finance Industry Report. Cambridge Centre for Alternative Finance. pp 12-88.

Ziegler, T., Shneor, R., Garvey, K., Wenslaff, K., Yerolemou, N., Hao, R., & Zhang, B. (2017). The 3rd European Alternative Finance Industry Report. Cambridge Centre for Alternative Finance. pp 16-105.

Ziegler, T., Shneor, R., Wenzlaff, K., Wang, W.B., Kim, J., Odorovic, A., Paes, F., Zhang, B., Johanson, D., Lopez, C., Mammadova, L., Adams, N., & Luo, D. (2020) The Global Alternative Finance Market Benchmarking Report. Cambridge Centre for Alternative Finance, pp. 1-228.

Van Thiel, D., & van Raaij, W. (2019). Artificial intelligence credit risk prediction: An empirical study of analytical artificial intelligence tools for credit risk prediction in a digital era. Journal of Risk Management in Financial Institutions, vol. 12, no. 3, pp. 268–286.

Wang, H., Chen, K., Zhu, W., & Song, Z. (2015). A process model on P2P lending. Financial In-novation, vol. 1, no. 1, pp. 1-8.

Wang, Z., Jiang, C., Ding, Y., Lyu, X., & Liu, Y. (2018). A Novel behavioral scoring model for estimating probability of default over time in peer-to-peer lending. Electronic Commerce Research and Applications, vol. 27, pp. 74–82.

Wang, X., Xu, Y., Lu, T., & Zhang, C. (2020). Why do borrowers default on online loans? An inquiry of their psychology mechanism. Internet Research, vol. 30, no. 4, pp. 1203–1228.

Webster, J., & Watson, R.T. (2002) Analyzing the Past to Prepare for the Future: Writing a Liter-ature Review. MIS Quarterly, vol. 26, no. 2, pp. 13-23.

Wu, Y., & Zhang, T. (2020). Can credit ratings predict defaults in peer-to-peer online lending? Evidence from a Chinese platform. Finance Research Letters, pp. 101724–.

Xia, Y., Liu, C., & Liu, N. (2017). Cost-sensitive boosted tree for loan evaluation in peer-to-peer lending. Electronic Commerce Research and Applications, vol. 24, pp. 30–49.

Yu, L., Yue, W., Wang, S., & Lai, K. (2010). Support vector machine based multiagent ensemble learning for credit risk evaluation. Expert Systems with Applications, vol. 37, no. 2, pp. 1351–1360.

Zhou, J., Li, W., Wang, J., Ding, S., & Xia, C. (2019). Default prediction in P2P lending from high-dimensional data based on machine learning. Physica A, vol. 534, pp. 122370–.

# 9 APPENDICES

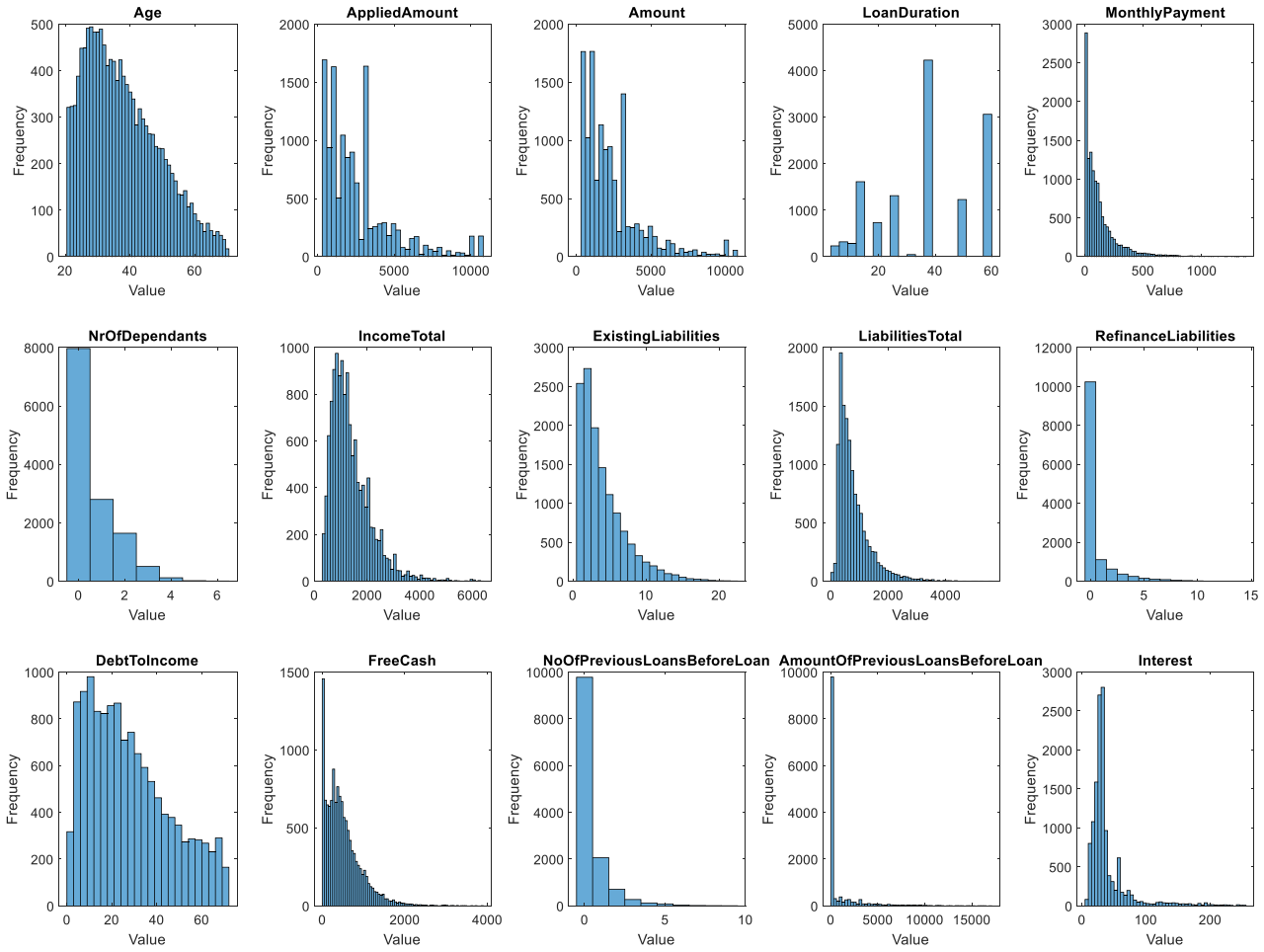
Appendix 1. Descriptions of used variables

| Variable | Definition |
|---|---|
| **Demographics of the borrower** | |
| Age | Age of the borrower |
| Gender | Gender of the borrower |
| Country | Borrowers residency |
| Education | Educational level of the borrower |
| MaritalStatus | Marital status of the borrower |
| NrofDependants | Number of children or other dependants |
| EmploymentStatus | Current status of employment |
| EmploymentDurationCurrentEmployer | Employment time with the current employer |
| WorkExperience | Borrower's overall work experience in years |
| HomeOwnershipType | Current housing situation |
| **Financials of the borrower** | |
| IncomeTotal | Borrower's total income |
| ExistingLiabilities | Borrower's number of existing liabilities |
| LiabilitiesTotal | Total monthly liabilities |
| RefinanceLiabilities | The total amount of liabilities after refinancing |
| DebtToIncome | Ratio of borrower's monthly gross income that goes toward paying loans |
| FreeCash | Discretionary income after monthly liabilities |
| Rating | Bondora Rating issued by the Rating model |
| **Characteristics of the loan** | |
| VerificationType | Method used for loan application data verification |
| AppliedAmount | The amount borrower applied for originally |
| Amount | Amount the borrower received on the Primary Market |
| LoanDuration | Current loan duration in months |
| MonthlyPayment | Estimated amount the borrower has to pay every month |
| Interest | Maximum interest rate accepted in the loan application |
| **Credit history of the borrower** | |
| NewCreditCustomer | Did the customer have prior credit history in Bondora |
| NoOfPreviousLoansBeforeLoan | Number of previous loans |
| AmountOfPreviousLoansBeforeLoan | Value of previous loans |

Appendix 2. Summary statistics of numerical variables

| Variable | Min. | Max. | Mean | Median | STD | Skewness | Kurtosis |
|---|---|---|---|---|---|---|---|
| Age | 21 | 70 | 37.74 | 36 | 11.34 | 0.62 | 2.64 |
| AppliedAmount | 500 | 10630 | 2617.95 | 2000 | 2220.70 | 1.69 | 5.87 |
| Amount | 265 | 10630 | 2354.30 | 1900 | 1957.21 | 1.82 | 6.80 |
| LoanDuration | 3 | 60 | 35.63 | 36 | 17.61 | -0.04 | 1.88 |
| MonthlyPayment | 0 | 1377.76 | 127.83 | 77.99 | 166.16 | 2.88 | 14.32 |
| NrOfDependants | 0 | 6 | 0.63 | 0 | 0.94 | 1.56 | 5.34 |
| IncomeTotal | 304 | 6400 | 1368.47 | 1200 | 761.36 | 1.66 | 7.48 |
| ExistingLiabilities | 1 | 22 | 4.09 | 3 | 3.19 | 1.61 | 6.08 |
| LiabilitiesTotal | 10 | 5594.86 | 793.74 | 618 | 589.98 | 2.28 | 10.90 |
| RefinanceLiabilities | 0 | 14 | 0.59 | 0 | 1.52 | 3.67 | 19.44 |
| DebtToIncome | 1.05 | 69.96 | 27.73 | 24.235 | 18.13 | 0.60 | 2.41 |
| FreeCash | 0.25 | 3862 | 503.49 | 408.36 | 435.28 | 1.64 | 7.50 |
| NoOfPreviousLoansBeforeLoan | 0 | 9 | 0.42 | 0 | 0.93 | 3.23 | 16.70 |
| AmountOfPreviousLoansBeforeLoan | 0 | 17095 | 799.38 | 0 | 1984.14 | 3.54 | 17.89 |
| Interest | 7.96 | 254.84 | 38.91 | 30 | 32.56 | 3.30 | 15.80 |

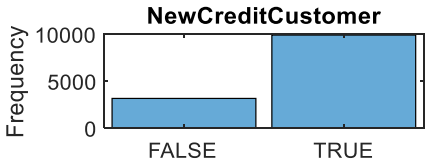
Appendix 3. Distribution of numerical variables

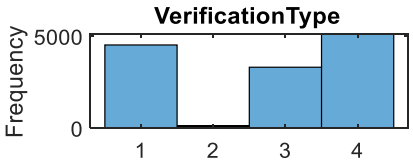


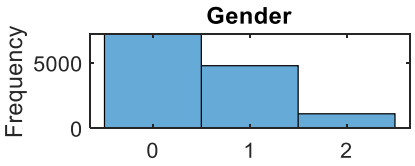
Appendix 4. Summary statistics of categorical variables

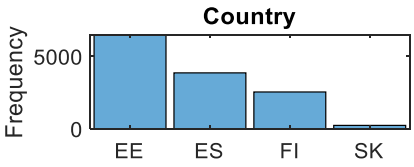
| Variable | Value | Count | Percent |
|---|---|---|---|
| NewCreditCustomer | | | |
| | 0 = FALSE | 3161 | 24.20% |
| | 1 = TRUE | 9903 | 75.80% |
| VerificationType | | | |
| | 1 = Income unverified | 4524 | 34.63% |
| | 2 = Income unverified, cross-referenced by phone | 116 | 0.89% |
| | 3 = Income verified | 3313 | 25.36% |
| | 4 = Income and expenses verified | 5111 | 39.12% |
| Gender | | | |
| | 0 = Male | 7191 | 55.04% |
| | 1 = Female | 4776 | 36.56% |
| | 2 = Other | 1097 | 8.40% |
| Country | | | |
| | EE = Estonia | 6481 | 49.61% |
| | ES = Spain | 3841 | 29.40% |
| | FI = Finland | 2524 | 19.32% |
| | SK = Slovakia | 218 | 1.67% |
| UseOfLoan | | | |
| | 0 = Loan consolidation | 2182 | 16.70% |
| | 1 = Real estate | 352 | 2.69% |
| | 2 = Home improvement | 3258 | 24.94% |
| | 3 = Business | 768 | 5.88% |
| | 4 = Education | 544 | 4.16% |
| | 5 = Travel | 725 | 5.55% |
| | 6 = Vehicle | 1112 | 8.51% |
| | 7 = Other | 3634 | 27.82% |
| | 8 = Health | 489 | 3.74% |
| Education | | | |
| | 1 = Primary education | 161 | 1.23% |
| | 2 = Basic education | 1633 | 12.50% |
| | 3 = Vocational education | 2822 | 21.60% |
| | 4 = Secondary education | 4940 | 37.81% |
| | 5 = Higher education | 3508 | 26.85% |
| MaritalStatus | | | |
| | 1 = Married | 3625 | 27.75% |
| | 2 = Cohabitant | 3006 | 23.01% |
| | 3 = Single | 5046 | 38.63% |
| | 4 = Divorced | 1195 | 9.15% |
| | 5 = Widow | 192 | 1.47% |
| EmploymentStatus | | | |
| | 0 = Unemployed | 13 | 0.10% |
| | 2 = Partially unemployed | 473 | 3.62% |
| | 3 = Fully employed | 10607 | 81.19% |
| | 4 = Self-employed | 613 | 4.69% |
| | 5 = Entrepreneur | 731 | 5.60% |
| | 6 = Retiree | 627 | 4.80% |
| EmploymentDurationCurrentEmployer | | | |
| | MoreThan5Years | 4865 | 37.24% |
| | Other | 0 | 0.00% |
| | Retiree | 0 | 0.00% |
| | TrialPeriod | 234 | 1.79% |
| | UpTo1Year | 2654 | 20.32% |
| | UpTo2Years | 1889 | 14.46% |
| | UpTo3Years | 1493 | 11.43% |
| | UpTo4Years | 1015 | 7.77% |
| | UpTo5Years | 914 | 7.00% |
| WorkExperience | | | |
| | 10To15Years | 2545 | 19.48% |
| | 15To25Years | 2793 | 21.38% |
| | 2To5Years | 1870 | 14.31% |
| | 5To10Years | 2881 | 22.05% |
| | LessThan2Years | 665 | 5.09% |
| | MoreThan25Years | 2310 | 17.68% |
| HomeOwnershipType | | | |
| | 0 = Homeless | 1 | 0.01% |
| | 1 = Owner | 3943 | 30.18% |
| | 2 = Living with parents | 2591 | 19.83% |
| | 3 = Tenant, pre-furnished property | 2240 | 17.15% |
| | 4 = Tenant, unfurnished property | 1485 | 11.37% |
| | 5 = Council house | 218 | 1.67% |
| | 6 = Joint tenant | 502 | 3.84% |
| | 7 = Joint ownership | 765 | 5.86% |
| | 8 = Mortgage | 1157 | 8.86% |
| | 9 = Owner with encumbrance | 162 | 1.24% |
| Rating | | | |
| | AA | 233 | 1.78% |
| | A | 467 | 3.57% |
| | B | 1229 | 9.41% |
| | C | 1955 | 14.96% |
| | D | 1946 | 14.90% |
| | E | 1600 | 12.25% |
| | F | 1276 | 9.77% |
| | HR | 4358 | 33.36% |
| Default | | | |
| | 0 = No | 5844 | 44.73% |
| | 1 = Yes | 7220 | 55.27% |

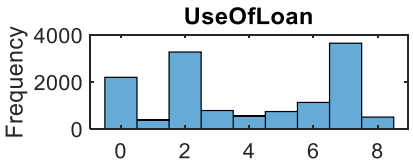
Appendix 5. Distribution of categorical variables

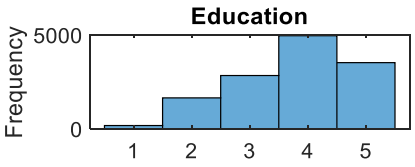


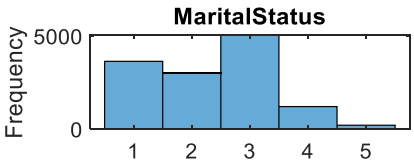



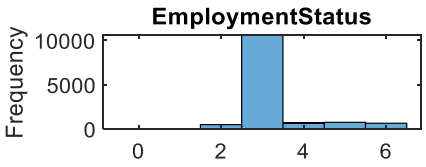



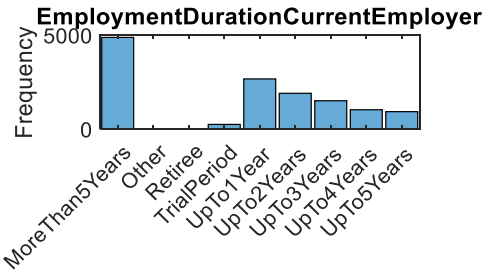



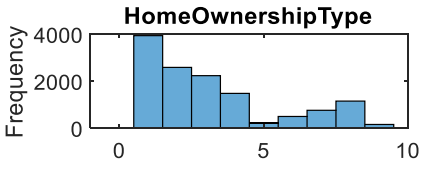



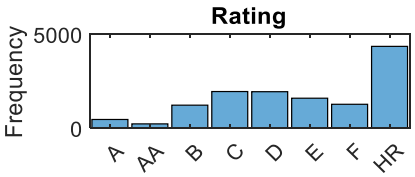



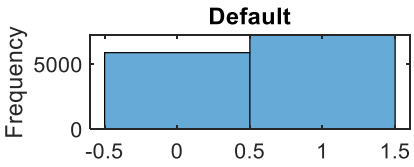







Appendix 6. In-sample and 10-fold CV for all countries using LR

| | Default hyperparameters | | Optimized hyperparameters | |
|---|---|---|---|---|
| Model | In-sample error | 10-fold CV error | In-sample error | 10-fold CV error |
| Whole data | 0.3060 | 0.3082 | 0.3056 | 0.3076 |
| Estonia | 0.3411 | 0.3528 | 0.3402 | 0.3492 |
| Spain | 0.3333 | 0.3602 | 0.3270 | 0.3581 |
| Finland | 0.3688 | 0.4009 | 0.3671 | 0.3892 |

Appendix 7. Hyper optimized parameters: Logistic regression

| Model | Lambda | Regularization |
|---|---|---|
| Whole data | 0.0003 | ridge |
| Estonia | 0.0000 | ridge |
| Spain | 0.0008 | lasso |
| Finland | 0.0031 | ridge |

Appendix 8. 10-fold CV for all countries using SVM

| | Default hyperparameters | |
|---|---|---|
| Model | In-sample error | 10-fold CV error |
| Whole data | 0.1951 | 0.3046 |
| Estonia | 0.1511 | 0.3782 |
| Spain | 0.1580 | 0.3523 |
| Finland | 0.1714 | 0.3870 |

Appendix 9. In-sample and 10-fold CV for all countries using RF

| | Default hyperparameters | | Optimized hyperparameters | |
|---|---|---|---|---|
| Model | In-sample error | 10-fold CV error | In-sample error | 10-fold CV error |
| Whole data | 0.1383 | 0.3103 | 0.2842 | 0.2949 |
| Estonia | 0.0573 | 0.3571 | 0.3341 | 0.3317 |
| Spain | 0.0000 | 0.3357 | 0.2512 | 0.3167 |
| Finland | 0.0000 | 0.4286 | 0.3333 | 0.3619 |

Appendix 10. Hyper optimized parameters: Random forests

| Model | No. of learning cycles | No. of variables to sample | Min. leaf size | Max. Number of splits | Split criterion |
|---|---|---|---|---|---|
| Whole data | 74 | 25 | 6 | 1 | gdi |
| Estonia | 10 | 9 | 88 | 15 | deviance |
| Spain | 75 | 28 | 502 | 1 | gdi |
| Finland | 10 | 26 | 12 | 46 | deviance |

Appendix 11. Determinants of default: Whole data

| | Estimate | SE | tStat | pValue |
|---|---|---|---|---|
| (Intercept) | -2.107 | 0.164 | -12.879 | 0.000 |
| Rating | 2.752 | 0.092 | 29.808 | 0.000 |
| MonthlyPayment | -1.068 | 0.310 | -3.451 | 0.001 |
| NoOfPreviousLoansBeforeLoan | -3.100 | 0.497 | -6.241 | 0.000 |
| NewCreditCustomer | 0.306 | 0.083 | 3.691 | 0.000 |
| AmountOfPreviousLoansBeforeLoan | 1.126 | 0.338 | 3.331 | 0.001 |
| Interest | 0.136 | 0.166 | 0.819 | 0.413 |
| Gender | 0.145 | 0.070 | 2.080 | 0.038 |
| AppliedAmount | 0.177 | 0.258 | 0.685 | 0.494 |
| Amount | 0.333 | 0.262 | 1.270 | 0.204 |
| ExistingLiabilities | 0.447 | 0.272 | 1.643 | 0.100 |
| Education | -0.575 | 0.084 | -6.826 | 0.000 |
| LoanDuration | 0.475 | 0.083 | 5.755 | 0.000 |
| IncomeTotal | 1.525 | 0.602 | 2.531 | 0.011 |
| FreeCash | 1.127 | 0.500 | 2.255 | 0.024 |
| RefinanceLiabilities | 1.527 | 0.346 | 4.407 | 0.000 |
| HOT_Owner | -0.236 | 0.089 | -2.663 | 0.008 |
| MS_Cohabitant | -0.217 | 0.056 | -3.865 | 0.000 |
| MS_Single | -0.016 | 0.052 | -0.301 | 0.764 |
| DebtToIncome | 0.328 | 0.147 | 2.240 | 0.025 |
| ES_Entrepreneur | -0.369 | 0.095 | -3.897 | 0.000 |
| HOT_Living_with_parents | -0.179 | 0.095 | -1.879 | 0.060 |
| LiabilitiesTotal | -1.393 | 0.480 | -2.902 | 0.004 |
| HOT_Tenant_unfurnished_property | -0.073 | 0.100 | -0.728 | 0.467 |
| HOT_Tenant_pre-furnished_property | -0.046 | 0.094 | -0.493 | 0.622 |
| UOL_Loan_consolidation | -0.178 | 0.067 | -2.665 | 0.008 |
| ES_Self-employed | 0.171 | 0.105 | 1.637 | 0.102 |
| VerificationType | -0.104 | 0.056 | -1.872 | 0.061 |
| HOT_Joint_ownership | -0.226 | 0.118 | -1.924 | 0.054 |
| HOT_Mortgage | -0.327 | 0.109 | -2.992 | 0.003 |
| ES_Retiree | 0.199 | 0.099 | 2.014 | 0.044 |

Appendix 12. Determinants of default: Estonia

| | Estimate | SE | tStat | pValue |
|---|---|---|---|---|
| (Intercept) | -1.347 | 0.240 | -5.617 | 0.000 |
| Rating | 2.281 | 0.161 | 14.180 | 0.000 |
| MonthlyPayment | -2.924 | 0.640 | -4.568 | 0.000 |
| LoanDuration | 0.374 | 0.130 | 2.874 | 0.004 |
| RefinanceLiabilities | 0.946 | 0.484 | 1.952 | 0.051 |
| NewCreditCustomer | 0.177 | 0.107 | 1.661 | 0.097 |
| NoOfPreviousLoansBeforeLoan | -2.167 | 0.526 | -4.120 | 0.000 |
| Education | -0.552 | 0.139 | -3.979 | 0.000 |
| AmountOfPreviousLoansBeforeLoan | 1.203 | 0.375 | 3.212 | 0.001 |
| ExistingLiabilities | 1.520 | 0.387 | 3.932 | 0.000 |
| UOL_Loan_consolidation | -0.209 | 0.108 | -1.942 | 0.052 |
| AppliedAmount | 0.338 | 0.609 | 0.555 | 0.579 |
| Amount | 0.812 | 0.625 | 1.299 | 0.194 |
| WE_Less_than_a_year | 0.390 | 0.135 | 2.890 | 0.004 |
| IncomeTotal | -0.241 | 0.549 | -0.439 | 0.660 |
| Age | 0.855 | 0.223 | 3.829 | 0.000 |
| Interest | 0.181 | 1.075 | 0.168 | 0.867 |
| EDCE_More_than_5_years | -0.098 | 0.086 | -1.131 | 0.258 |
| VerificationType | -0.117 | 0.085 | -1.377 | 0.169 |
| WE_More_than_25_years | -0.327 | 0.118 | -2.763 | 0.006 |
| LiabilitiesTotal | -0.822 | 0.744 | -1.105 | 0.269 |
| HOT_Tenant_pre-furnished_property | 0.258 | 0.106 | 2.423 | 0.015 |
| HOT_Mortgage | -0.211 | 0.139 | -1.511 | 0.131 |
| HOT_Living_with_parents | -0.011 | 0.106 | -0.104 | 0.917 |
| EDCE_Up_to_1_year | -0.003 | 0.092 | -0.030 | 0.976 |
| WE_2_to_5_years | 0.095 | 0.096 | 0.985 | 0.325 |
| HOT_Owner | 0.012 | 0.087 | 0.140 | 0.889 |
| Gender | -0.604 | 0.138 | -4.384 | 0.000 |
| UOL_Travel | -0.119 | 0.164 | -0.728 | 0.466 |
| EDCE_Up_to_2_years | 0.133 | 0.099 | 1.353 | 0.176 |
| UOL_Business | -0.216 | 0.127 | -1.701 | 0.089 |

Appendix 13. Determinants of default: Spain

| | Estimate | SE | tStat | pValue |
|---|---|---|---|---|
| (Intercept) | -1.172 | 0.562 | -2.086 | 0.037 |
| MonthlyPayment | -0.017 | 0.648 | -0.026 | 0.979 |
| NewCreditCustomer | 0.269 | 0.231 | 1.164 | 0.244 |
| NoOfPreviousLoansBeforeLoan | -6.423 | 1.853 | -3.466 | 0.001 |
| AmountOfPreviousLoansBeforeLoan | 2.306 | 1.356 | 1.701 | 0.089 |
| LoanDuration | 0.302 | 0.196 | 1.540 | 0.124 |
| ExistingLiabilities | 1.697 | 0.721 | 2.354 | 0.019 |
| Gender | 0.406 | 0.132 | 3.082 | 0.002 |
| VerificationType | -0.424 | 0.134 | -3.171 | 0.002 |
| Rating | 1.808 | 0.449 | 4.026 | 0.000 |
| LiabilitiesTotal | -2.210 | 1.223 | -1.807 | 0.071 |
| Education | -0.711 | 0.194 | -3.666 | 0.000 |
| Interest | -0.429 | 0.243 | -1.766 | 0.077 |
| DebtToIncome | -0.821 | 0.346 | -2.373 | 0.018 |
| HOT_Mortgage | -0.617 | 0.221 | -2.799 | 0.005 |
| FreeCash | 0.247 | 1.167 | 0.212 | 0.832 |
| UOL_Home_improvement | 0.244 | 0.186 | 1.310 | 0.190 |
| Amount | 1.550 | 0.558 | 2.779 | 0.005 |
| HOT_Joint_tenant | -0.850 | 0.359 | -2.367 | 0.018 |
| UOL_Travel | -0.374 | 0.238 | -1.576 | 0.115 |
| RefinanceLiabilities | 2.717 | 1.224 | 2.221 | 0.026 |
| WE_More_than_25_years | 0.175 | 0.186 | 0.936 | 0.349 |
| UOL_Vehicle | 0.333 | 0.262 | 1.272 | 0.203 |
| UOL_Education | -0.166 | 0.240 | -0.691 | 0.489 |
| HOT_Joint_ownership | 0.455 | 0.320 | 1.423 | 0.155 |
| AppliedAmount | -0.038 | 0.477 | -0.080 | 0.936 |
| IncomeTotal | -0.280 | 1.440 | -0.195 | 0.846 |
| UOL_Other | -0.100 | 0.169 | -0.592 | 0.554 |
| Age | 0.384 | 0.318 | 1.208 | 0.227 |
| NrOfDependants | 0.191 | 0.401 | 0.477 | 0.634 |
| UOL_Loan_consolidation | -0.148 | 0.247 | -0.598 | 0.550 |

Appendix 14. Determinants of default: Finland

| | Estimate | SE | tStat | pValue |
|---|---|---|---|---|
| (Intercept) | -0.705 | 0.480 | -1.469 | 0.142 |
| MonthlyPayment | -1.479 | 0.707 | -2.090 | 0.037 |
| Rating | 1.473 | 0.294 | 5.009 | 0.000 |
| NoOfPreviousLoansBeforeLoan | -6.924 | 2.760 | -2.509 | 0.012 |
| AmountOfPreviousLoansBeforeLoan | 1.151 | 1.256 | 0.916 | 0.360 |
| NewCreditCustomer | -0.107 | 0.308 | -0.346 | 0.729 |
| RefinanceLiabilities | 2.188 | 0.702 | 3.117 | 0.002 |
| Amount | 0.484 | 0.604 | 0.802 | 0.423 |
| AppliedAmount | 0.251 | 0.589 | 0.426 | 0.670 |
| ExistingLiabilities | 0.035 | 0.582 | 0.061 | 0.952 |
| VerificationType | -0.379 | 0.154 | -2.466 | 0.014 |
| DebtToIncome | -0.002 | 0.393 | -0.004 | 0.997 |
| FreeCash | 1.767 | 0.578 | 3.059 | 0.002 |
| ES_Entrepreneur | -0.583 | 0.255 | -2.290 | 0.022 |
| LoanDuration | -0.081 | 0.230 | -0.353 | 0.724 |
| WE_2_to_5_years | 0.164 | 0.178 | 0.917 | 0.359 |
| HOT_Council_house | 0.696 | 0.282 | 2.467 | 0.014 |
| UOL_Business | -0.413 | 0.344 | -1.201 | 0.230 |
| Age | 0.082 | 0.237 | 0.346 | 0.729 |
| UOL_Real_estate | -0.579 | 0.364 | -1.591 | 0.112 |
| Education | -0.434 | 0.220 | -1.974 | 0.048 |
| HOT_Mortgage | -0.212 | 0.162 | -1.306 | 0.191 |
| EDCE_Trial period | -0.460 | 0.345 | -1.330 | 0.183 |
| Gender | -0.471 | 0.220 | -2.135 | 0.033 |
| MS_Single | -0.002 | 0.121 | -0.020 | 0.984 |
| UOL_Home_improvement | 0.247 | 0.133 | 1.857 | 0.063 |
| UOL_Health | -0.486 | 0.379 | -1.283 | 0.200 |
| ES_Partially_employed | 0.233 | 0.235 | 0.990 | 0.322 |
| EDCE_Up_to_4_years | 0.158 | 0.195 | 0.809 | 0.419 |
| EDCE_Up_to_1_year | 0.072 | 0.148 | 0.486 | 0.627 |
| Interest | -0.180 | 0.467 | -0.385 | 0.700 |