CHARLES UNIVERSITY

FACULTY OF SOCIAL SCIENCES

Institute of Economic Studies

Bachelor thesis

2018 **Jolana Čermáková**

CHARLES UNIVERSITY

FACULTY OF SOCIAL SCIENCES

Institute of Economic Studies

Jolana Čermáková

Credit Risk of P2P lending on the Czech Market

Bachelor thesis

Prague 2018

Author: Jolana Čermáková

Supervisor: prof. Ing. Oldřich Dědek CSc.

Academic Year: 2017/2018

Bibliographic note

ČERMÁKOVÁ, Jolana. *Credit Risk of P2P lending on the Czech Market* Prague 2018. 56 pp. Bachelor thesis (Bc.) Charles University, Faculty of Social Sciences, Institute of Economic Studies. Thesis supervisor prof. Ing. Oldřich Dědek CSc.

Abstract

This thesis analyzes an emerging peer-to-peer lending industry, while introducing its main features and risks, where the risk of default and its moderation gets the most attention. Uniquely provided data from the front Czech platform Zonky containing nearly 6 000 observations serve as a baseline for credit risk modeling. It has been investigated which variables have the largest effect on default on the Czech P2P market. The final model is used to predict the associated probability of default and to compute the credit score for potential borrowers using these online platforms. Results support the fact that education, age, way of living, expenses, marital and employment status, income and the number of children are significant variables when determining the risk of default. Many of these findings are in accordance with previous international papers published on this topic.

Abstrakt

Tato práce se zabývá analýzou rozšiřujícího se odvětví peer-to-peer půjček, přibližuje jeho hlavní charakteristiky a zkoumá rizika s ním spojená. Nejpodrobněji se věnuje riziku úvěrového selhání a možným technikám sloužícím k zmírnění jeho dopadu. Modelování úvěrového rizika vychází z téměř 6 000 pozorování unikátně poskytnutých přímo českou přední platformou Zonky. Cílem bylo zjistit, jaké proměnné mají na českém trhu na úvěrové selhání nejvýznamnější dopad. Finální model tedy slouží k odhadnutí pravděpodobnosti nesplacení dluhu (defaultu) a k výpočtu kreditního skóre potenciálních vypůjčovatelů využívajících online platformy. Jako nejdůležitější proměnné se ukázaly dosažené vzdělání, věk, způsob bydlení, výdaje, rodinný stav, zaměstnání, příjem a počet dětí. Dosažené výsledky se do značné míry shodují se závěry podobných studií provedených v zahraničí.

Keywords

P2P Lending; Credit Default; Czech Market; Zonky; Credit Risk; Information Asymmetry; FinTech

Klíčová slova

P2P půjčky; Úvěrové selhání; Český trh; Zonky; Úvěrové riziko; Informační asymetrie; FinTech

Declaration of Authorship

I hereby proclaim that I wrote my bachelor thesis on my own under the leadership of my supervisor and that the references include all resources and literature I have used.

I grant a permission to reproduce and to distribute copies of this thesis document in whole or in part.

Prague, 10 May 2018

Signature

Acknowledgment

I wish to express gratitude to the thesis supervisor prof. Ing. Oldřich Dědek CSc. for his useful advices and guidance. Further I am grateful to my family and friends for the support and encouragement during my bachelor studies. Last but not least, I thank to Zonky team, especially to Vít Ficl and Aleš Černý, for providing me with all necessary facilities essential for credit risk modeling in a friendly environment.

Bachelor Thesis Proposal

Author	Jolana Čermáková
Supervisor	prof. Ing. Oldřich Dědek CSc.
Proposed topic	Credit Risk of P2P Lending on the Czech Market

Research question and motivation:

The default risk of P2P lending is on average higher than lending from conventional financial institutions. This is because it is riskier to lend to individuals as opposed to established financial institutions. The leading P2P platform on the Czech market, Zonky, does not provide credit default insurance. However, they score the loans based on the risk and incentivize borrowers to pay promptly.

Because of higher risk, P2P lending also has higher yields, which attracts more risk seeking investors because they have potential to earn higher interest than from financial institutions. Moreover, there is a very small barrier to enter the P2P market, in terms of the minimum amount being lent or borrowed. Because of the low minimum investment, it is easy for investors to diversify the risk. Thus, P2P lending has become increasingly popular throughout the world, especially when the central bank rates are lower. Because P2P technically functions as a free market economy, if there is high demand for relatively safe high yield loans, the market becomes very competitive. This has resulted in decreasing banks' revenue from loans as more and more P2P lending platforms emerge on the market. P2P lending platforms are increasingly owned or bought out by large financial institutions to hedge for future possible losses in revenue. For example, Zonky is owned by Home Credit.

Contribution:

Previously, many academic studies on this topic have focused on analyzing data from the Lending Club (the largest P2P platform on the U.S. market)

- mainly because of the availability of public data and also high number of observations. The leading P2P platform on the Czech market is Zonky, which makes it the most appropriate resource to analyze. However, because P2P market is relatively immature on the Czech market, the data is not publicly available and there will be less observations than on the U.S. market.

Therefore, my contribution to the topic will be acquiring and analyzing the data on the Czech market as opposed to more common publicly available data.

Methodology:

On most of the P2P lending platforms, the lenders bear the risk of default, which creates a problem of information asymmetry. Even though the platforms score the risk of each borrower, it is not a fully transparent process and the lender often has to trust the platform as a reliable source of evaluating risk.

I will analyze the probability of credit default on Zonky by using either a logit or a probit model. As a dependent variable, I will be using a binary variable for credit default and as explanatory contributory variables, I will be using variables such as purpose of a loan, age, gender, marital status, educational level, monthly income, Zonky risk score, current indebtedness, a loan amount, an interest rate and loan term. The model will attempt to predict defaults on Zonky and then compare the results with other studies, which were focused on other platforms. Based on the results, I will evaluate possible causes.

Outline:

1. Key characteristics
2. P2P lending in the Czech Republic
3. Zonky and the main competitors
4. Future of P2P lending

5. Data gathering and its description

6. Regression model

7. Analysis and Results

Core bibliography:

(a) Yao, F., Sui, X. (2016): *The Research to the Influential Factors of Credit Risk in the P2P Network Loan Under the Background of Internet Financial.* Paris.

(b) Serrano-Cinca, C., Gutiérrez-Nieto, B. and López-Palacios, L. (2015): *Determinants of default in P2P lending.* PLoS ONE 10(10): e0139427.

(c) Lin, X., Li, X. and Zheng, Z. (2017): *Evaluating borrower's default risk in peer-to-peer lending: evidence from a lending platform in China.* Applied Economics, 49(35), pp.3538-3545.

(d) Chang, S., Kim, S.D.O. and Kondo, G. (Autumn 2015-2016): *Predicting Default Risk of Lending Club Loans.* Stanford Engineering Everywhere.

(e) Berger, S.C. and Gleisner, F. (2009): *Emergence of financial intermediaries in electronic markets: The case of online P2P lending.* Official Open Access Journal of VHB.

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

(g) Klafft, M. (2008): *Online peer-to-peer lending: a lenders' perspective.* Fraunhofer ISST, Berlin, Germany.

Acronyms

AIC	Akaike Information Criterion
BBB	The British Business Bank
CNB	The Czech National Bank
CZK	Czech Crown
FinTech	Financial Technology
IV	Information Value
KS	The Kolmogorov-Smirnov Statistics
MLE	Maximum Likelihood Estimation
P2P	Peer-to-Peer
PwC	PricewaterhouseCoopers
SEC	The Securities and Exchange Commission
SVM	Support Vector Machines
UK	The United Kingdom
USA	The United States of America
WOE	The Weight of Evidence

Contents

1 Introduction

Financial sector offers several innovations each year. This thesis focuses on P2P lending, standing for English peer-to-peer, a recent trend in the loan market. As the name suggests, peers are at the heart of things. The British platform Zopa was in 2005 the first one to introduce this concept and it has not stop growing ever since [1]. By excluding an intermediary in a get-the-loan process, this new branch has risen an awareness among both the borrowers and the lenders. For borrowers, platforms represent a quick source of finance with lower interest rates and ease of use. For investors, on the other hand, mediated loans serve as considerable fixed returns unfolding from the freely chosen level of risk. Thanks to the straightforward process, platforms can still make profit while maintaining forenamed advantages.

Previous research considering the P2P lending industry has largely focused on associated risks and regulations mitigating them. That being achieved primarily by proposing credit risk models using particular techniques and by their elaborate evaluation. The level of development still diametrically differs across countries which also corresponds to the number of empirical studies issued. Among front countries setting the pace is the USA, the UK and China [2]. The Czech P2P market is still considered as young, even though its lending volumes are growing. Front domestic platform Zonky is in a planned loss (according to its director Pavel Novak [3]) since its birth in 2015, however it should change in the next three years thanks to the sophisticated marketing attracting both new borrowers and investors.

Lending money through financial institutions is still the most common way, however in the case of risk clients, clients with financial difficulties, or during the times of financial crises, people are more inclined to seek alternatives. It is also generally thought that younger generations are gradually switching to Fintech (financial technology) providers excluding the typical financial institutions as they often do not trust them and rather trust the online platform that puts more emphasis on sympathetic environment and supposedly a more personal approach. Later in this paper, comparison especially with banks will help to demonstrate advantages and disadvantages of P2P lending.

Throughout this thesis, the author will shed a light on several challenges that P2P industry faces. By properly approaching the topic, a reader will be able to identify them and apply the knowledge further. Focusing mainly on the Czech market, there will be things discussed that were presented only at the foreign level until now. These include regulations specific for the Czech industry and examination of the front P2P platform. A basic foreign context will help to compare the state of regulations in different countries.

The main contribution of the thesis is therefore a proper analysis of the Czech P2P industry where purpose of this research is to find *which variables have the largest effect on default on the Czech P2P market*. Introducing the industry information and building a framework of the legislative will serve as a baseline for this analytical part of the thesis. In practical part, author will analyse and model sociodemographic data of the platform Zonky and their impact on the probability of default.

This thesis will be structured as follows. In Chapter 2, the concept of P2P lending as a whole will be presented, including the matching process and an overview of the world situation with great emphasis on the Czech market. Chapter 3 is devoted to the literature review, mapping relevant publications on the topic. Chapter 4 describing the examined data opens the practical part, followed by Chapter 5 focusing on the methodology and Chapter 6 implementing several testing procedures. Finally, Chapter 7 concludes.

2 Theoretical Background

2.1 P2P lending

P2P lending is a form of lending money without going through traditional financial institutions, where banks, trust companies, or insurance companies are the main representatives [4]. In the case of P2P, these institutions are no longer needed as it is based on peers communicating directly with each other, more accurately through online platforms. Platforms serve as an online environment with markets where the loans are being intermediated. The core processes take place at the virtual marketplace. Since lending is the core source of profit in the commercial banking, it is reasonable that it gave rise to this online alternative [5].

Similar users that utilize crowdfunding sites tend to use P2P platforms too, whereas in this case, the borrowers are obliged to repay the debt. Besides start-ups and trustworthy small individual borrowers, high-risk people are attracted by this form of lending money as it could often be a last resort for them after their request is rejected at a bank.

Thanks to the online environment, the overall process is generally smooth, even though platforms serving as a middleman do not have much power in the decision-making process of the final matching [6]. They are however very important in putting the individuals on the same marketplace and trying to avoid defaults by assigning appropriate credit ratings to borrowers. Credit rating evolves from the information acquired. If the platform has enough information with noticeable value and does its job correctly, then the lower the rating, the higher the risk of default [7]. Therefore, the risk should be captured in a higher interest rate associated to the borrower.

2.1.1 Differences between classic bank and P2P lending

As opposed to the traditional model where financial institutions stand on the supply side, at these marketplaces, individuals stand on both demand and supply side of the market. Borrowers are looking for loans with low interest rates and lenders are, on the other hand, seeking the investments with the highest possible return. There are many lenders with different levels of risk aversion, therefore demand tends to be satisfied more often. For borrowers, marketplaces offer a new source of funds, similar to banks and other financial institutions. However, for investors, this form of financing represents a new asset class [8]. Investors have an option to choose which loans to fund. That is one of the advantages of these platforms, where among others the process also tends to be quick and transparent.

By connecting individuals online on marketplaces, there are much lower consequent costs. A cheaper business model results in nearly no expenses connected to the maintenance. Banks, in contrast, use sophisticated systems where a continuous attention is required. Operating costs associated to the banking sectors but not so much to the P2P platforms are the reason, why these platforms offer higher returns to investors, while borrowers pay lower interest rates at the same time [7]. Advantages as lower costs are balanced with several disadvantages connected with this industry. For instance, many borrowers still rather rely on conventional banking and physical actions by themselves. They fear to put their personal data online and of possible identity theft [9].

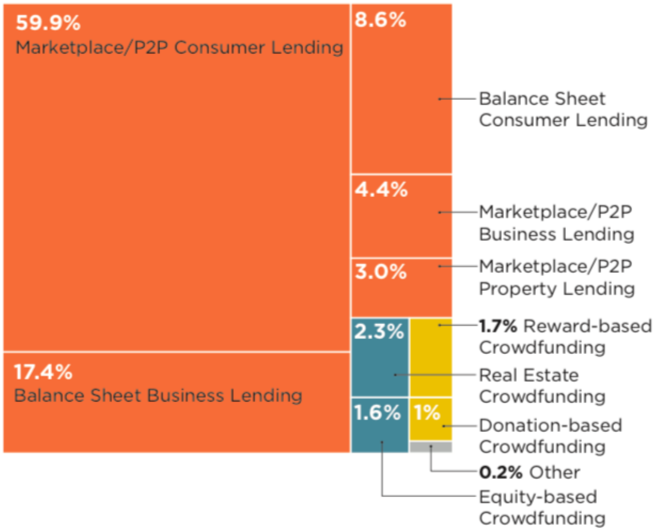


Figure 2.1: Americas Alternative Finance 2016 (market share by model)

Source: The Americas Alternative Finance Industry Report [10]

It is questionable whether the electronic marketplaces can someday replace or at least be a sufficient alternative to conventional financial institutions. In some countries (USA, China, UK) it does not form a negligible part of lending anymore and it is a leader in the alternative finance market share (to see in Figure 2.1).

2.1.2 Risk insurance

P2P investors are often not very familiar with associated risks, where in the most cases, there is no collateral for loans [7]. In fact, investors are the ones exposed to risk associated. Even so, it can be an attractive alternative of a fixed income.

A lot of platforms conduct some kind of a loan insurance in the case of default to protect investors' money. For instance, a British P2P company RateSetter has built a Provision Fund for investors to have a protection in a case of missed payments. This fund maintains the expected amount of default loans and up to February 2018 had a 100% track record [11]. Another example is a platform Lending Works, which uses so-called "Lending Works Shield" insuring investors against defaults, fraud and cybercrime [12]. Next to these particular safety precautions, nearly all platforms recommend carefully diversified portfolios to investors. By diversifying investments, final returns can be quite steady. For instance, compared with the stock market, P2P investing is associated with less volatility. To be seen in the Figure 2.2, returns of British platform Funding Circle are in general much smoother than these of other investment options, including the Financial Times Stock Exchange 100 Index.

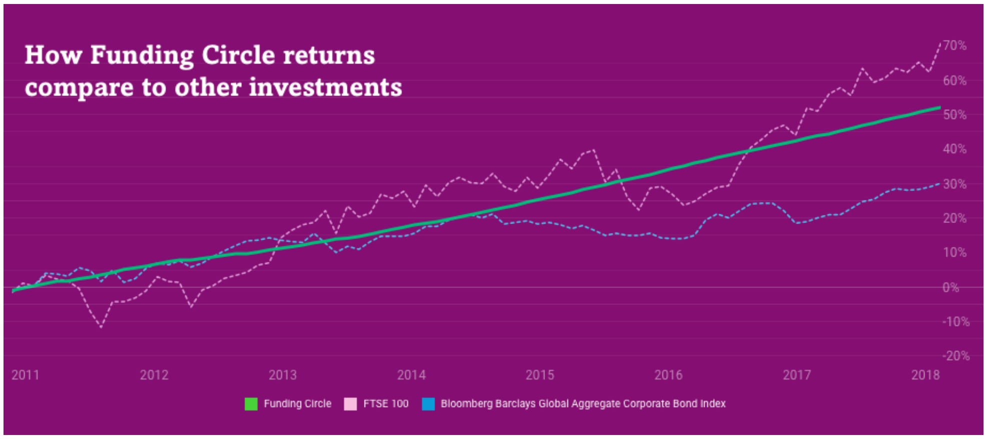


Figure 2.2: Funding Circle compared to other investments

Source: Funding Circle [13]

Moreover, compared with fundamentally risk-free government bonds, promised yields are in most cases significantly higher. On Zonky, average annual return is 6.03%, whereas a long term Czech government bonds range about 2% (Figure 2.3)

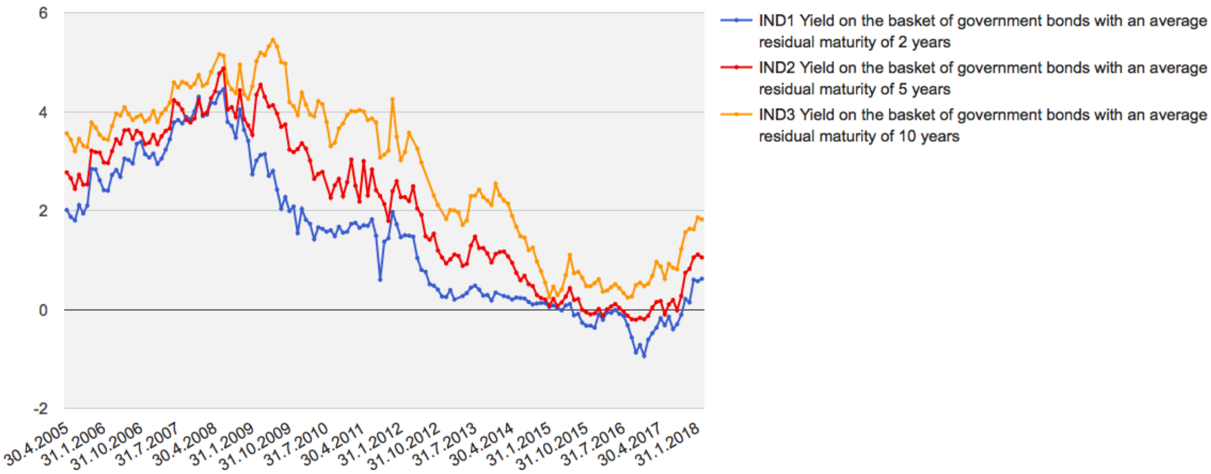
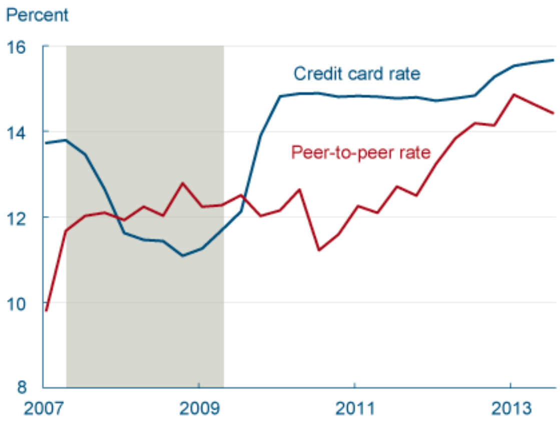


Figure 2.3: Government bond basket yields (end of month) (%)

Source: Czech National Bank [14]

2.1.3 Commercial vs. Non-commercial platforms

Receiving approved interest rate is also out of the authorities of platforms. Individual borrowers ought to pay these directly to investors. However, as platforms are usually for-profit, they receive a fee for this kind of service. Platforms earning profit are called commercial [16]. Even when the fee is included, platforms still, in the most cases, offer more favourable conditions than the banks do. Figure 2.4 compares interest rates both on credit card and on P2P platforms.



Note: Shaded bar indicates a recession.
Sources: Bankrate.com; Lending Club.

Figure 2.4: Average peer-to-peer interest rates compared to credit card rates

Source: www.crowdfundinsider.com [17]

Beside the for-profit platforms, there are also websites that focus on funding people facing financial difficulties. Similar to the for-profit ones called commercial, these are called non-commercial. They diminish the problem of credit rationing, which means that even borrowers willing to pay high interest rates do not receive the loan in the end. These non-profit websites offer loans with no interest rates, therefore investors stand here as social helpers [7].

2.2 Growth and magnitude of the P2P industry

2.2.1 Initial impulses

A lot of newly created instruments tend to arise during breakthrough events. For instance, when the famous financial crisis took place, banks were forced to hold more funds because of the capital requirements regulatios and stricter credit checks. In the US, “lending volume in the fourth quarter of 2008 was 47% lower than it was in the prior quarter and 79% lower than at the peak of the credit boom (2007:Q2)“ [18]. That was the time when a lot of new marketplaces mediating loans started to emerge.

Electronic marketplaces have the advantage of only connecting individuals, without any need of collecting capital [19]. Therefore, it seemed as a great substitute in the time of crisis. Moreover, people found it more and more comfortable to do everyday life duties from home or on their cell phones.

2.2.2 Reshaping financial industry

Since the upswing of the internet in the 1990s, entirely new industries started to develop. Internet is not only changing processes in the financial system but the whole market structure too [6]. The Fintech industry began to form a noticeable part of a market recently. According to the PwC study, ”most bankers see personal loans (64%) and personal finance (50%) most at risk in moving to a Fintech company" [20].

Fintech tends to reshape the financial industry by improving the quality of financial services, reducing costs and creating a more diverse and stable financial landscape [21]. One of its drivers, sharing economy inclusive of the P2P lending, aims to achieve all these characteristics, therefore undoubtedly belongs to the flourishing Fintech sector. Many other companies using that model, like Uber or Airbnb, have already become a perceived substitute in their own industry. In addition, a launch of Web 2.0 in early 2000s made the creation of an online markets easier and more accessible [22]. All together, following the Fintech pattern, online markets help to ease the economic activity, reduce both transaction and information costs and can replace the traditional intermediaries, at least to a significant extent [23].

Since the launch of the first platform in 2005, the P2P lending industry has grown enormously, on the global scale [6]. This rapid growth includes both the loan amount and number of emerging platforms [24]. In the Figure 2.5, exponential growth of total loan issuance on Lending Club can be seen. According to the Transparency Market Research, the opportunity in the global P2P lending market was USD 26.16 billion in 2015 and is predicted to reach USD 897.85 billion in 2024 [25].

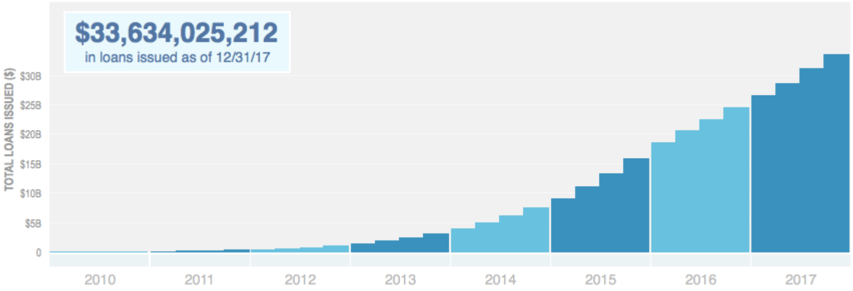


Figure 2.5: Total loan issuance (Lending Club data)

Source: Lending Club [26]

In the Czech Republic, P2P consumer lending is the second most used form of crowdfunding, right after the reward-based crowdfunding (e.g.

the Hithit platform). According to Martin Strecha (CEO Crowder.cz and Managing partner at Investree), these two are also expected to grow the fastest in the future [16]. However, the fraction of P2P lending of the total consumer lending in the Czech Republic is yet neglecting. According to the Czech National Bank, consumer loans reached CZK 230 billion at the end of January 2018, where Zonky platform arranged something about CZK 2 billion [27].

2.3 Information needed and the matching process

Matching of individuals differs across the various platforms and countries, but the principle stays the same. Both borrowers and lenders first need to register on the platform and deliver a specific information about themselves. Especially in the case of borrowers, many things are considered before assigning a credit rating to them. This information then helps the platform to determine the final credit rating and the associated interest rate based on the acquired data. Some say that the credit evaluation is the most important task of P2P platforms [6]. If they pass through this scanning process, loan is put at the marketplace where investors come into play. All investors are also required to fill in the information before they can decide to whom and how much they are going to invest.

As opposed to conventional banking, lending through P2P platforms occurs anonymously, users are usually registered under nick names. As the process is anonymous, it protects the data of both sides that can be sensitive, but on the other hand it creates an asymmetric division of knowledge about each other.

2.3.1 Information Asymmetry

The problem of information asymmetry arises on P2P platforms and has been highlighted by plenty of papers already. It simply means that one party corresponds with more information than the other one. Davis and Murphy (2016) [24] discuss possible information imbalance between investors and the platform, as investors do not always understand the potential risks and rely on the information provided by the platform. To diminish this problem, platforms publish statistics of defaults in particular risk classes for investors to be in the picture.

However, the information asymmetry between the platform and an investor is not the one widely discussed. Matching borrowers to lenders and associated information is of a bigger concern instead. Borrowers know their financial situation better which puts them in an advantageous position towards investors. Due to the fact that they put the data about themselves on the platform on their own, there could be some tendencies to hide things that would otherwise not be favourable. When the final credit rating is computed, investors usually see only the rating and associated interest rate, therefore there is no possibility for them to track the data. This can lead to moral hazard or adverse selection [22]. To avoid it, platforms are trying to improve the trust of investors through several vehicles. These are well known in the conventional banking but more difficult to acquire online.

It is in the own interest of particular platform to deliver as much reliable information as possible because if the platform would serve only as a provider of an online marketplace, not many investors would be interested in funding loans that they have no information about. For that reason, nearly all platforms cooperate with independent and trustworthy risk analysts, both individual and institutional professionals.

2.3.2 Hard vs. Soft data

The majority of the qualitative information is fixed and cannot be influenced by the borrower, as it involves default history, monthly income, family status, spending and other stable numbers. So-called "hard" data are checked in social indexes and through banks and other institutions keeping track of credit history of the borrower [29]. Moreover, this necessary information is demanded by the majority of platforms, therefore these cannot be even hidden or somehow swapped. For instance, the examined Czech platform Zonky checks loan applicants in both Bank and Non-bank Client Information Register, Insolvency Register, Register of Debtors, Central Evidence of Executions, and many others [30].

However, there are also possibilities to include other things, that can help borrowers to get a loan. Specifically, the so-called ”soft” data provides qualitative information. When we talk about differences from the common financial institutions, the soft information definitely stands out. Because of the more personal approach, many customers have a better confidence in P2P platforms and in people they are lending money to. Data with this feature are very important for borrowers as they are controllable and can have a substantial impact on the results [29].

It is not completely clear whether this kind of information should be included, as it can also result in investors lending money to risk clients that would otherwise never get the loan without the additional information. An example of this voluntary supplement are borrowers telling stories. This example is included because of the paper’s main focus on Zonky, which is primarily known for that. Stories and other kinds of soft data are not usually checked, as it is challenging to do it systematically. It therefore puts investors in difficult position where they have to decide, whether they do or do not believe it and if they want to take it as an important factor during the decision-making process. Next to stories, people can include for example a photograph, the reason of the loan or other certificates proving their clean credit shield [29].

2.3.3 Auction vs. fixed rate assigned

In general, there are two main options how borrowers are connected to lenders. In both of them, lenders are free to choose, where they are going to invest on the first come first served basis. Preferably a simpler option corresponds to a computation of a fixed interest rate by the platform. Platform assigns a credit rating to the borrower, who is then put at the marketplace and investors can decide according to the appropriate interest rates. This approach is used by the majority of platforms today, including Zonky.

The other matching option is derived through the auction process, where the borrower states the highest possible interest that he or she is able to eventually pay (this rate has to be higher than some simply calculated minimum rate accounting for the risk) [24] and then lenders bid on that interest for the fixed period of time. After the auction is finished, the lowest interest bid is chosen and both sides are acknowledged and connected afterwards. Then the process of regular instalments is usually same as in the first case. The second largest US platform Prosper (according to the total loan issuance) used to match individuals through the auction, however right after the stricter regulation, including mandatory registration at the Securities and Exchange Commission (SEC), it rather switched to the first method.

2.4 Regulations

As the industry of P2P lending is still quite new to the financial sector, legislators face several challenges. There are no precedents for the P2P loans yet, therefore it is completely up to the government how it is going to deal with it. Since numerous approaches towards regulation are present across the world, mainly in the countries with a more mature P2P market, countries as the Czech Republic can inbreathe them. As a matter of fact, observed practices beyond the borders can be helpful but also completely inapplicable, as the regulatory framework can be made-to-measure for the corresponding country. Coupled with that, it is challenging to spread these businesses abroad.

Due to the fact that in the case of P2P platforms investors are the ones facing the credit risk, regulations tend to protect both them and borrowers. Lending money through financial institutions is much easier regarding the risk, which these institutions carry and are experts in it [7]. The best possible regulation would be the one where all marketplace participants are protected from fraud and data leakage while maintaining the features thanks to which these platforms gained the popularity. However, acquiring this state of regulation is very challenging when the legislation needs to protect other industries and the whole nation at the same time.

As Davis and Murphy (2016) argues [24], P2P lending platforms "combine the functions of a market operator and a provider of financial services". Nowadays, both of these industries are often regulated separately but their combination offers many more loopholes that has to be taken care of. Moreover, financial regulations undoubtedly belong to the most fragile one. Even if they are implemented with right intentions, they often have a significant impact on the participants.

In many countries, P2P industry is still regulated only through already existing laws which is not sufficient in a lot of cases. By applying these general laws on the specific P2P sector, it happens that several platforms operate on a thin border with legislation [28].

2.4.1 Current state in the Czech Republic

The Czech National Bank's (CNB) supervision of P2P platforms is chiefly based on the two acts in the Czech legislation. Firstly the Payment System Act (Zakon o platebnim styku) and secondly the Consumer Credit Act (Zakon o spotrebitelskem uveru). However, in order to determine the relevant regulation, it is necessary to rely on the recognition of the business model and the activities carried out by the respective P2P platform. Platforms can differ from each other mainly in the relation to both financial providers and borrowers.

Generally speaking, P2P platform activities typically fulfil the features of provider of payment services. In that case, it is essential that the platform has authorisation to activities of the payment institution or a registration of a small-scale payment service provider. Specifically, in December 2016, an amendment of the Consumer credit act was enacted in order to make the market more transparent. Since then, every institution providing credit has to fulfil several legislative requirements. For instance, receive a licence from the CNB containing the condition of having at least CZK 20 million as the initial capital. This legal provision aims to protect consumer's rights and increase the market supervision.

2.4.2 Foreign context (US, UK, China)

Emerging boom of P2P platforms hit the world several years ago and the legislation was not prepared for that. For instance in China, from 2013 to mid-2016, 26% of all platforms were either completely closed, in bankruptcy or running on deposit [28]. As opposed to that, in some countries P2P industry has a good reputation in terms of the relation with the government. For instance in 2016, the British Business Bank (BBB) owned by state even invested Ł85 million in the P2P lending sector, which basically means that it invested the taxpayer money. A BBB spokesman told Business Insider: "Peer-to-peer lending platforms have the potential to be a successful delivery model for small business finance. Investing in these, and other kinds of platforms is a vital part of our remit to foster a more diverse small lending market for smaller businesses; indeed more than 10,000 smaller businesses across the country have already benefited from our partnership with Funding Circle" [31].

On the US market, the industry was for a few years operating under nearly no restrictions since the emergence of the front platforms in 2005. It has been discussed for a long time which authority should have a main word in P2P industry. This "market freedom" last only until 2008 when the Securities and Exchange Commission (SEC) stepped in. The SEC claimed that platform Prosper is selling securities, therefore has to be registered with the SEC to stand with securities laws [32]. This ordinance, relating all P2P platforms, caused a huge market purge. As a result, Lending Club and Prosper ended up as the only ones successfully registered and this current state makes it challenging for the new operators to fulfil ample requirements and enter the market.

2.5 Zonky

Zonky was established in 2015 in Prague. The main investor of the project is a Dutch Innovation Fund Home Credit Lab N.V., subsidiary company of Home Credit, which belongs to the PPF group. Until April 2017, Zonky has been led by startup incubator CreativeDock, headed by Lucie Tvaruzkova. Since then, mentioned Home Credit is the owner and Pavel Novak is the director.

There are few things that distinguish this platform from the foreign ones. For instance, provided that the loan is approved, a particular borrower can be usually assured that it will be financed. Even if there were not enough investors to cover the full amount, Zonky will in the most cases fund the loan. Since Zonky presents itself rather as a competitor in the banking sector than rescue for people whose loan requirement was declined in a bank, it offers portfolios entailing a much lower risk. As the head of the company, Pavel Novak, said: ”If someone has a problem to get a loan at a classic bank, he or she will have a problem with us as well” [3].

2.5.1 CNB categorization

According to the Czech National Bank (CNB) P2P lending is a hybrid form of lending. From the viewpoint of the Consumer Credit Act a key aspect of P2P platform categorization is whether the platform operator acts as a provider or consumer credit intermediary.

The Zonky model is based on the principle of providing consumer credit directly from the platform operator. To operate this type of P2P platforms, the authorisation to operate a non-bank consumer loan provider granted by CNB, or authorisation to provide payment services (depending on the nature of the accompanying services) is necessary.

2.5.2 Investors

Investors have to follow the registration process. They are required to be at least 18 years old, need to deliver two copies of a proof of identification and have funds in an investor account. Individual investors indicate the selected loan, choose the amount of participation on the loan and confirm it. The amount of investment is limited by the platform, in order to diversify the potential risk. The minimum amount to invest is CZK 200 and the maximum depends on the number of active investments. According to the platform, 122 investments ensure the highest return with the lowest number of participations to investors.

The investors' right to participate in the selected loan arises on the basis of the "Framework Agreement on Payment Services and Participation in Consumer Credits" closed with Zonky. By the time the order is executed, the amount is blocked in the investors' account. If a sufficient volume of confirmed contributions is collected from investors, Zonky will provide a loan and settle payment orders by debiting the blocked amounts from the investors' account. If the credit is not granted within three business days of confirmation, it will be dissolved and the blocked amount in the investors' account will be released.

Investors are obliged to pay the platform a fee from the currently invested money, which is calculated on a daily basis and charged monthly. A fee amount unfolds from the agreed interest rate associated to the loan provided loan according to the pricelist, ranging from 0.2% to 5% p.a (Table 2.1).

	A**	**A***	**A++**	**A+**	**A**	**B**	**C**	**D**
interest	*3.99% p.a.*	*4.99% p.a.*	*5.99% p.a.*	*8.49% p.a.*	*10.99% p.a.*	*13.49% p.a.*	*15.49% p.a.*	*19.99% p.a.*
fee	*0.2%*	*0.5%*	*1.0%*	*2.5%*	*3.0%*	*3.5%*	*4.0%*	*5.0%*

Table 2.1: Interests and fees

Source: www.zonky.cz [30]

Investors are not entitled to dispose with the participation shares elsewhere than on the so-called secondary market which also operates under Zonky's rules. At the secondary market, in operation since August 2018, investors have an option to sell or buy investments from other investors. Moreover, even investors with no investments yet can begin to invest right on the secondary market. It is primarily aimed to get liquidity.

After the participation is filed for the relevant loan, a lien on the receivable is set up for the client to the benefit of the investor. It does not restrain Zonky from recovering credit claims. Moreover, Zonky has no responsibility for the repayment and does not provide any guarantees to investors. Its only duty is to make appropriate effort accordant with its professional experience to recover the highest amount possible.

2.5.3 Borrowers

All published loans are anonymized and contain the minimum loan parameters of the interested party. These include required amount, amount invested so far, repayment period, assigned interest rate, rating of the borrower, several verifications in the registers and of the applicant's income, purpose of the loan and finally the period when demand is open to investors (usually two days). All information is published not before signing the contract between Zonky and the potential borrower. Resulting interest rate is directly computed by Zonky and subsequently approved by the borrower.

Borrowers' side is supported by the appropriate legislation too, in particular by the already mentioned Consumer Credit Act. Registration conditions are more thorough than these of investors. In addition to the age and identity validation, consent to the processing of personal data, consent to inspect the Non-Banking Client Information Register, affirmed earnings and expenses are required.

Zonky provides loans from CZK 20 000 to CZK 500 000 for the period from 6 months to 7 years and an option to repay the entire loan anytime.

On the basis of provided data, credit score is calculated and appropriate interest rate assigned, starting at 3.99% p.a. The lowest interest rate is connected to the best rating, specifically A**. On the other hand, the highest one, 19.99% p.a., stands for rating D (Table 2.1). However, Zonky also tries to understand the individuality of each client, thus they have the ability to include some personal achievements and curiosities.

Borrowers are obliged to repay the loan according to the schedule of instalments in their individual profiles plus a 2% one-time fee to the platform of the loan amount. If either two consecutive monthly instalments or one instalment for the period longer than 3 months is not repaid, Zonky will require the borrower to reimburse the entire outstanding principal, interest owed included.

3 Literature Review

Despite the fact, that the P2P industry is still considered as young and quite new to the existing environment, many academic papers are published every year on this topic. Due to the unexplored parts of this online complex system, there is still a lot of questions to ask and even more answers to give.

In the majority of cases, the main focus of these papers is on considering the data from markets where the P2P industry grows at the fastest rate and where they are publicly available. In particular the US, the UK and China [5]. Data-based papers' main concern is to estimate the loan profitability and associated probability of default by all kinds of techniques. Other, rather theoretical, papers investigate the state of regulation in individual countries, problem of information asymmetry arising between the opposite market sides or the general role of this kind of intermediaries.

Beside these specifications, each of these works, at least in part, concerns the risk of P2P platforms. Risk can take many different forms, involving fraud, identity theft and naturally the most resonant default risk. As loans and their repayments are not in no matter guaranteed, Meyer et al. (2007) [33] considered the main instruments how to limit the associated credit risk. According to them, this ought to be done by providing information about borrowers, by diversifying investors' funds across many loans and by the direct peer pressure to delinquent debtors. A decade later, Liang (2017) [28] rather aims to solve the financial risks by suggesting the right way towards the effective regulation. As he proposes, it should comprise general conditions about capital and organizational structure and especially consider risk control requirements and giving right to governments for them to accordingly enforce. Similar to him, Milne and Parboteeah (2016) [34] are having the same approach in directing the P2P industry regulations, focusing on minimizing the risk of fraud and also operational and security costs.

Davis and Murphy (2016) [24] consider different approaches to regulate P2P platforms. They claim that the traditional business models are not suitable for the completely new environment and suggest a more efficient regulatory structure which accounts for the distinct risk structure. Even though some level of regulation is already present in the most countries, P2P industry still benefits from regulatory savings in comparison to costly banking industry. That is mainly due to the fact, that banks need to rely on deposits which these platforms do not need at all. Acting as an intermediary, reserve or capital requirements do not apply to them. By that, costs are reduced and both borrowers and investors are usually offered more favourable and profitable lending conditions [35].

3.1 Probability of default

As already mentioned, great emphasis of published papers is on examining the probability of default. That is usually acquired by collecting and modelling data from the US Lending Club, Prosper or front Chinese platforms.

To have a precise model predicting the loan performance is very valuable in the P2P lending industry, as having less defaults makes the platform trustworthy and more attractive for investors with resources.

In 2014, Tsai et al. [36] presented 4 machine learning algorithms to determine which is the best in predicting the default in order to avoid these loans and to invest only in the good ones. They found out that the modified Logistic Regression suits prediction the best, in comparison to Naive Bayes, Random Forest or Support Vector Machines (SVM). As opposed to this finding, one year later (2015) Malekipirbazari and Aksakalli [37] rather proposed the Random Forest method and after comparing it with other methods, they claim that it is the best method to predict borrower's status. However, after the examination of these

4 main methods got used up, Wang et al. (2017) [38] proposed novel behavioural scoring model to predict the dynamic probability of default. By dynamic it is meant that it will predict not only whether the borrower will default, but also when it is going to be.

There are papers that either proceed further after knowing the probability of default, or they can propose a way how to mitigate that probability. By using Lending Club data, Chang et al. (2016) [19] tried to predict the expected returns and maximize them by using the probability of default. This maximization of returns was aimed to be done mainly by avoiding high-risk loans, therefore knowing the probability that the borrower will not repay a debt seemed to be the right instrument. Their finding was that the credit score assigned to the borrower is the best predictor of default, which was achieved by employing the Naive Bayes model.

3.2 Text analysis

Some papers focus not only on the probability of default using sociodemographic data, but on text analysis too. Loan descriptions are an option for the borrower to include important or valuable facts in order to make the relationship with investors less anonymous and more trustworthy. Supporting Chang et al. [19] findings, Serrano-Cinca et al. (2015) [7] agree that the rate assigned by the platform is the best predictor of default, however the model serves best if other variables are added to it, in particular borrower's debt level. By examining the relation between interest rate, grade assigned and default, they show that a higher interest rate results in a higher probability of default.

In 2014, Carmichael [39] went through the Lending Club loan descriptions and found several key words and phrases that were commonly used by borrowers. Among the key words with a positive feature were for example ”responsible”, ”trustworthy”, ”never late” or ”reliable”. By employing the discrete-time hazard model he found out that next to the usual significant variables as borrower income, recent credit inquiries or purpose of loan, facts like whether the borrower claims to be creditworthy and whether he or she writes in complete sentences also stands out in prediction of default. Thus, if a borrower's loan description lacks complete sentences, he or she is more likely to default. Similarly, claims about creditworthiness are believable.

Han et al. (2017) [29] included the loan description analysis too, this time using Renrendai data which is a Chinese P2P platform. Following up on Carmichael, they found out that applicants for loans who use longer sentences were less likely to be successfully funded. Moreover, they included the summary of main determinants for funding success and divided them into four categories. The first one is a loan characteristic (loan rate, amount and duration), subsequent category are borrower's personal information (gender, age, . . .), the third are voluntary information (photograph, loan description) and the last one is soft information (friendships, groups).

A paper based primarily on the loan description analysis by Herzenstein et al. (2011) [40] uses data from the first US P2P platform, Prosper. Authors found out that this unverifiable information in loan descriptions affect investors' decisions above the verifiable ones. On the whole, this paper supports the fact that borrowers use loan descriptions strategically to attract investors and that it usually helps them to get the demanded interest rate. Because of this strategic approach on one side and often unskilled investors on the other side, the US SEC aims to improve verification processes through further regulations. SEC claims that lenders have too much reliance on this incorrigible information which leads them to make rash decisions [32].

3.3 Mitigating information asymmetry

A significant part of research concerning P2P lending is considering the arise of information asymmetry between individual borrowers and investors. As Freedman and Jin (2017) [41] mention in their paper, platform Prosper tries to reduce the asymmetry by institution of the social networking features, namely option of several group memberships or connecting with each other and identifying them as friends. They find out that registered users with social ties have a higher probability of their loans to be funded and of getting a lower interest rate, meaning that investors see social ties as a positive sign in terms of repayments. However, connected with that finding they also emphasize that it can lead to misinterpretation of particular groups and wrong conclusions about trustworthiness. Lin et al. (2013) [42] came to the similar findings and claim that these online friendships on Prosper act as a signal of credit quality, therefore results in a funding success and lower interest rates associated.

Zhang et al. (2017) [43] use data from the largest Chinese platform, Paipaidai, to highlight the increasingly prominent problems associated with online lending industry. Among others, they see an ineffective credit ranking and loan approval system as one of them. Therefore, the purpose of the paper is to help the platform to improve its loan approval system and to reduce operational costs. They find that for loan to be successfully funded, significant factors are annual interest rate, repayment period, description, credit grade, successful loan number, failed loan number, gender, and borrowed credit score.

Apart from the general information about borrower and a loan, demographic information is also usually acquired. Lin et al. (2017) [42] focus on borrowers' demographic characteristics and analyze them by using a credit risk evaluation model. Results show that gender, age, marital status, educational level, working years, company size, monthly payment, loan amount, debt to income ratio and delinquency history are significant variables explaining the possible default. By including these personal characteristics, they claim that it improves the overall model quality and predicts the probability of default better.

As opposed to all the models and calculations, Meng (2016) [44] adopts an online questionnaire to discover substantive determinants influencing lending decisions. Considering the results, he suggests that these factors are "verified documents", "safety protection from platforms", "endorsement from borrower's friend" and "number of borrower's friend bid".

Even though it is difficult for often unskilled investors to correctly predict the borrower's creditworthiness, Klafft (2008) [45] claims that careful lenders who use easy selection criteria can be profitable in the end. Apart from that, he suggests that the P2P industry can be successful in a long run, if platforms highlight an issue of bad investments and substandard loan performance. However, even if the market can be sustainable, it will not solve the problem of people in financial distress, as only good debtors with high credit rating will be applicable.

To conclude, many academic papers have been published, touching several topics concerning the P2P industry. Although slightly different, the basis of all of them is always similar, therefore it leads to the similar conclusions too. Further works exploring better scoring models are expected, as it is going to be core of problem during the existence of the P2P phenomenon.

4 Data

The aim of the practical part is to analyze the Czech P2P market. For the purpose of this thesis, the sociodemographic data set acquired from Zonky has been used. The data set includes information about each mediated loan since approximately April 2016 until August 2017. This time frame contains 5 692 observations with 26 explanatory variables in the original data set. However, due to some missing or erroneous values, final number gets thinner, which is specified in Chapter 5. After further examination, author has decided to exclude variables that were not relevant for the following model. These include for instance variables with all observations following the same pattern and these not having any noticeable value.

Data includes both numerical (income, age, etc.) and categorical (marital status, housing type, etc.) variables.

4.1 Data description

In this subsection, Zonky data set will be further described, primarily by bringing the particular variables closer to the reader. Description will serve as a baseline for follow-up modelling.

Among the observed variables, there are three binary variables - default (yes/no), sex (male/female) and purpose (refinancing/others).

In general, a male requesting a loan is a more common scenario than a female applicant, specifically men create 69% of the examined sample. A loan purpose falls either to "refinancing" (39%) or to "others" (61%) category, which is not further specified. According to Zonky website, this usually includes for instance a loan for a new car, household equipment, major unexpected repairs or other everyday-life things.

The majority of people applying for a loan have no children (58%), followed by a categories with 1 child (22%) and 2 children (17%). Only few observations (3%) involve an individual with 3 or more children. Detailed summary can be seen in the Table 4.1.

CHILDREN/TARGET	*0*	*1*	*row total*
0	3194	87	3281
	0.973	0.027	0.578
1	1220	21	1241
	0.983	0.017	0.219
2	926	10	936
	0.989	0.011	0.165
3	183	4	187
	0.979	0.021	0.033
4	28	1	29
	0.966	0.034	0.005
5	4	0	4
	1.000	0.000	0.001
column total	5555	123	5678

Table 4.1: Children vs. target cross table

Achieved education level of borrowers is depicted in the Figure 4.1, where exactly 50% of the sample contains people with high school education and only 20% have received either bachelor's or master's degree.

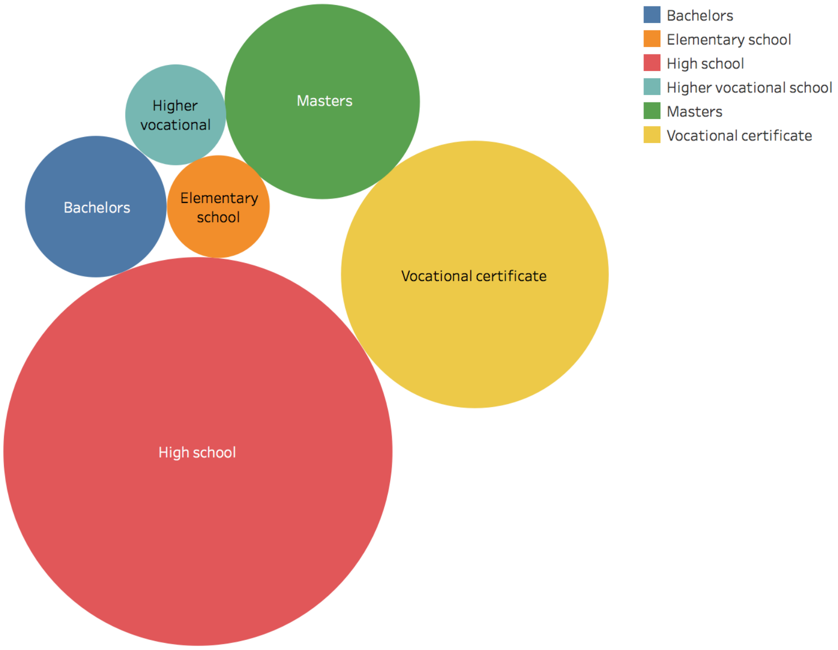


Figure 4.1: Distribution of education levels

A variable considering a permanent residence indicates that most loan applicants live in Prague (16%), Central Bohemian region (13%) or

Moravian-Silesian region (12%). On the other hand, Karlovy Vary and Highlands region belong to the least busy (both 3%) (Figure 4.2).

With respect to housing type, 30% live in his or her own flat or in a house with mortgage, 22% pay rent and 19% live in his or her own flat or in a house without mortgage.

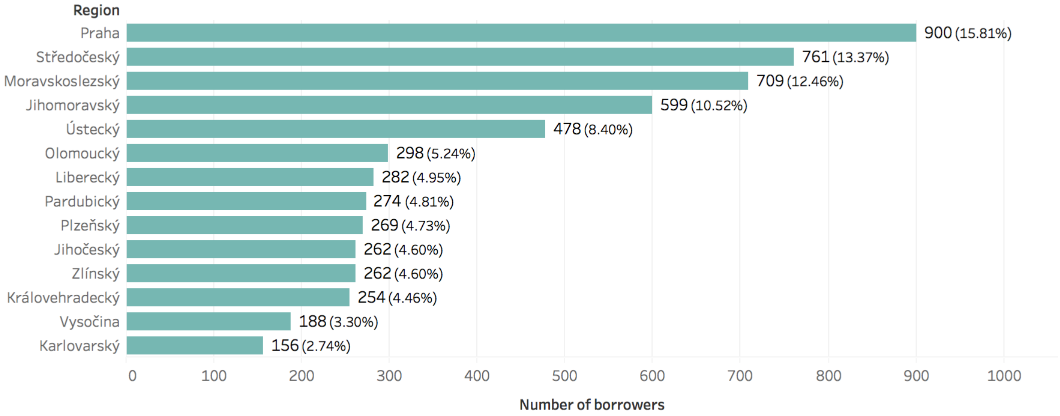


Figure 4.2: Region distribution

Regarding the marital status, the majority of borrowers is married (36%), followed by single people (27%) and by these living with his or her life partner (26%).

As a type of primary income, 75% indicate themselves as employed, 15% self-employed. A less considerable part belongs to entrepreneurs (3%) and pensioners (4%).

Numerical variables are summarized in the Table 4.2, where common statistical measures are used (median, mean, standard deviation, minimum, maximum).

Variable	**Median**	**Mean**	**SD**	**Min**	**Max**
Age	35	36.55	10.8443	18	88
Primary income (CZK)	23 000	26 479	19 062.58	0	400 000
Secondary income (CZK)	0	4 113	9 974.974	0	200 000
Total income (CZK)	25 100	30 593	21 025.38	0	415 000
Primary income (sex)	0.8627	1	0.723	0	19.0142
Secondary income (sex)	0	1	2.465	0	47.332
Total income (sex)	0.8432	1	0.689	0	15.9534
Total expense (CZK)	14 066	15 911	11 248.64	0	175 002
Expense to income	0.5471	0.5408	0.2716	0	6.5
Requested annuity (CZK)	3 000	3 642	2 663.796	500	20 000
Requested amount (CZK)	140 000	160 280	114 884.3	20 000	500 000
Term	51.094	57.197	56.277	2.015	1284.718

Table 4.2: Numerical variables summary statistics

An average age of borrower is 36.55, which supports the idea from the theoretical part that mainly younger generation is concerned with the P2P lending concept.

The average wage in 2017 in the Czech Republic is summarized in the Table 4.3 based on the data acquired from the Czech National Bank. Considering the values, a mean total income in our data set was slightly higher during the whole examined period (ending in August 2017). But in general, a typical Zonky borrower corresponds to an average Czech resident.

period	***CZK total***
Q4 2017	31 646
Q3 2017	29 063
Q2 2017	29 352
Q1 2017	27 907

Table 4.3: The average wage (the Czech Republic)

A mean Expense to income ratio is 0.54, thus an average client spends 54% of his or her income and saves the rest. The most common requested amount is CZK 140 000 which is a little lower than its mean.

5 Methodology

Resulting analysis will be primarily based on detecting the most significant variables, computing associated credit scoring and testing the applied model.

Several methods examining credit risk and their quick evaluation were mentioned in Chapter 3. In general, credit scoring models are present in order to evaluate a potential borrower based on his or her characteristics. Financial institutions then decide whether to grant a loan or not [48].

5.1 Empirical model

In this thesis, the basic logistic regression (LR) technique will be used, supporting the claim of Tsai et al. (2014) [36] that this model is the most appropriate for the prediction. Besides that, according to Ala'raj and Abbod (2016) [48], the LR is still considered the industry-standard model and Deloitte research denotes it as a prominent and one of the most successful methods to do credit scoring, among others due to its transparency and simplicity [46].

Simply put, the LR is a function that inputs the information x_i about borrower i and outputs the probability of default, called a binary response model $p(y = 1|x_i)$ [46]. A center of interest is the linear function f (5.1), where k denotes the number of explanatory variables, x_i is an explanatory factor i and β_i is a regression coefficitent of that explanatory factor i:

$$f(x_i) = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_k x_{ik} \tag{5.1}$$

To separate good from bad loans, a model of binary choice is needed with possible outcomes either zero or one. This restriction belongs to the main advantages of this model. Compared to the linear probability model where fitted probabilities can also be negative or greater than one. In order to predict the probability of default, results of the regression function need to occur between these two numbers. It can be achieved by implementing an increasing logistic function h (5.2), which fulfills the restrictive condition for all real numbers z [47]:

$$h(z) = \frac{exp(z)}{1 + exp(z)} \tag{5.2}$$

Putting together the regression function f (5.1) with the logistic function h (5.2), we get the resulting probability function (5.3) bounded by zero and one:

$$p(y = 1|x_i) = \frac{exp(f(x_i))}{1 + exp(f(x_i))} = \frac{exp(\beta_0 + \beta_1 x_{i1} + \cdots + \beta_k x_{ik})}{1 + exp(\beta_0 + \beta_1 x_{i1} + \cdots + \beta_k x_{ik})} \tag{5.3}$$

To estimate the logistic regression, we can use maximum likelihood estimation (MLE).

5.2 Building the model

The resulting model (logistic regression) has been built in the R Software. When preprocessing the data, some variables were excluded from the final data set and some were newly created - either by creating an interaction between included variables, or by doing simple numeric operations between them. Specifically, variable ”secondary income”, which was in the original data set divided in more detailed categories, was simplified by creating a new summarized variable by subtracting variable ”primary income” from ”total income”.

Besides that, all three income variables were further adjusted for gender, as their magnitude usually differs between males and females. In this sample, an average male salary is CZK 33 205, whereas an average female salary yields CZK 25 073. It was done by firstly computing a mean of the men’s income and the women’s income separately and then dividing the original income by the accordant computed mean. Thus, these three variables are transformed to decimal numbers, where values larger than 1 shows that the income is above average, whereas values lower than 1 shows an income below standard.

Before the classing of variables, several visualisations were implemented to see eventual outliers, erroneous values and the overall distribution of observed variables. In particular, these visualisations included mainly histograms and scatter plots of continuous variables and cross tables for all present classes.

From the total of 5 962 observations in the beginning, after detecting erroneous values, 14 observations were exluded, therefore the final data set includes 5 678 observations. A small summary in the form of a cross table can be seen in the Table 5.1, which shows the total number of males and females in the sample with respect to their loan status and a corresponding percentage of defaults.

SEX/TARGET	*0*	*1*	*row total*
FEMALE	1728	33	1761
	0.981	0.019	0.310
MALE	3827	90	3917
	0.977	0.023	0.690
column total	5555	123	5678

Table 5.1: Sex vs. target cross table

The default rate does not differ much between men and women, as it is only slightly higher (2.3%) for men than for women (1.9%). That is also a reason why the gender variable did not evidence high enough significance to participate in the final model mentioned later in the thesis.

Final list of used variables and their specifications to be seen in 5.2. All these were grouped and included in the first regression, however the final model includes only significant ones. It will be presented in the Chapter 7.

Variable	**Class**	**Levels**	**Specification**
target	num	-	-
domain_1	factor	13	seznam, gmail, email, centrum, post, volny, atlas, tiscali, hotmail, icloud, outlook, yahoo, others
domain_2	factor	3	cz, com, others
purpose_name	factor	2	refinancing, other
vek	int	-	-
sex	factor	2	male, female
children	factor	6	0, 1, 2, 3, 4, 5
education_name	factor	6	masters, bachelors, higher vocational school, high school, vocational certificate, elementary school
housing_type_name	factor	7	shared housing, partner, parents or family members, cooperative apartment, rental housing, own flat or a house without a mortgage, own flat or a house with a mortgage
marital_status_name	factor	6	registered partnership, divorced, single, married, widowed, partner
region_name	factor	14	praha, jihocesky, jihomoravsky, karlovarsky, vysocina, kralovehradecky, liberecky, moravskoslezsky, olomoucky, pardubicky, plzensky, stredocesky, ustecky, zlinsky
pop_type_name	factor	5	village(<1000), small town(<5000), town(<50000), large town(50000+), Prague
total_income_avg	int	-	-
primary_income_avg	num	-	-
secondary_income_avg	num	-	-
type_primary	factor	8	employment, entrepreneur, liberal profession, maternity leave, pension, self employment, student, others
total_expense	int	-	-
expense_to_income	num	-	-
term	num	-	-
requested_annuity	int	-	-
requested_amount	int	-	-

Table 5.2: Final list of used variables

Already listed in the Table 5.2, all non-continuous variables were converted to factors in order to be able to divide them later into specific groups. Besides that, the dependent variable ”target” has been defined more precisely by variable ”good” containing the good loans (non-defaults, 0) and ”bad” containing the bad loans (defaults, 1). Moreover, a special category for missing values was created for all explanatory variables.

The next step was to split the data set into a training set with 80% of data (4 542 observations) for building a model and a test set with 20% of data (1 136 observations) to test it. This has been done in order to test the model properly and not to rely on eventual biased results in the case of its low predictive power. If it would be tested on the same data set at which the model was trained, results could be too optimistic. By testing the model on the yet untouched data, it should give more accurate results.

Even when the further procedure consists of specific applications on two separate sets, the default ratio has been intentionally preserved. Before implementing a final model on training data, the ratio in both sets has been checked and it was found out that defaulted loans compose 2.2% of the sample.

5.3 Classing

Further procedure includes so called fine classing and following coarse classing. This technique is used in order to tackle a nonlinear relationship, potential outliers and for results to be easily interpretable. That being done by breaking variables down into categories which will be represented in the regression. It has been processed separately for numerical and categorical variables, as for the latter, fine classing is not much needed - categorical variables are already divided into categories by their nature. Furthermore, the fine and coarse classing has to be considered separately for each variable.

Numerical fine classing consists of numerical variables cut to equally sized groups. In this case, by 0.05 quantiles, therefore divided into 20 categories. This fine division ensures a detail analysis. After plugging specific variables into a general function, table and plot is generated. A plot visualizes the Weight of Evidence (WOE) of each category and serves as a base for coarse classing. A table, giving more tangible outcomes, summarizes each class with a total number of "Goods" and "Bads" in that class (inclusive of accordant good rate and bad rate), lower and upper bound and both Information Value (IV) and WOE computed as follows:

$$goodrate = \frac{good}{totalgood} \tag{5.4}$$

$$badrate = \frac{bad}{totalbad} \tag{5.5}$$

$$WOE = log(\frac{goodrate}{badrate}) \tag{5.6}$$

$$IV = WOE * (goodrate - badrate) \tag{5.7}$$

The WOE is used to capture a relative risk of each class. When the WOE is known, the IV can be computed. It gives a more interpretable number, where values over 0.3 are likely to feature in the final model and values under 0.1 are rather viewed as weak [2]. The IV can be computed for each data class.

Categorical "fine classing" consists only in grouping and computation of WOE. Supposedly the best practice is to create fine classes for each potential value. There is no lower nor upper bound, otherwise the summary table and a plot stay the same.

In general, coarse classing consists in grouping acquired classes by merging those with similar levels of risk or, less often, those having some logical relation. These should be chosen so that the information value stays preserved. Grouping categorical variables is generally straightforward as they are not monotonous, thus even unneighbourly classes can be merged. Though in some cases, even categorical variables can follow a certain sequence (education levels). Usually, when grouping continuous variables, we have to be more cautious and follow the order. After a group division is done, individual IV's can be checked in order to preserve its desired size. If the IV's appear to be too low, a different grouping is usually recommended.

The mentioned plot can be seen in the Figure 5.1 - fine classing of a categorical variable "education_name". From the left, depicted bars denote: $bachelors$, $masters$, $vocationalcertificate$, $highschool$, $highervocationalschool$

and *elementaryschool*. Coarse classing was determined easily by merging first two bars, third to fifth bar and lastly a subgroup *elementaryschool* was represented by its own category.

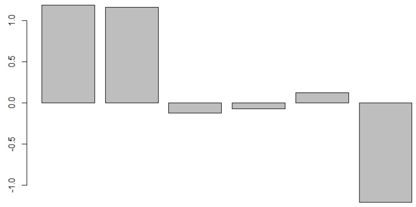


Figure 5.1: Coarse classing (education_name)

For instance, when groups acquired by fine classing are further divided into three larger groups by coarse classing, a data frame with three columns and a dummy for each of them will arise. This data frame can be once more plotted to see the correctness of group distribution. If the plot still contains similar bars, these can be always grouped.

When groups for both numerical and categorical variables are specified, a stepwise logistic regression gradually drops variables. Among that, the best model, in terms of goodness of fit, is determined according to the AIC (Akaike information criterion). The AIC can be obtained when the log-likelihood is maximized, which in this case holds as we use process called maximum likelihood estimation (MLE). MLE transforms target variable into a log function. The smaller the AIC, the better the fit. The AIC assess whether the variable improves the predictive power of the model. There is a command executing the step function and determining the best model on its own, therefore it does not include any manual selection.

As a result, only significant variables remain in the final model summary. However, to re-check the correctness of a group selection, an extra tool has been implemented. To avoid a too strong mutual relationship, variables in the final model were further examined with each other. A correlation matrix of these variables was created in order to avoid multicollinearity (correlations over 0.5 could result in inaccurate predictions). Several groups of the same variable with too high correlation were detected and later grouped in the other manner. This procedure was repeated until the model did not show any suspicious correlation.

When the model is completed, its resulting coefficients are scaled and rounded in order to interpret them as a more tangible final credit score. In this thesis, I have chosen to multiply the coefficients by 100 and round them. Results are presented in the Chapter 7 in the Table 7.1.

5.4 Scoring procedure

After the estimation and transformation of final coefficients, a corresponding credit score can be counted. Once again, it is being approached separately to each variable used in the final model and to its included categories.

In the case of categorical variables, each occurring category is assigned the appropriate treated coefficient. For the numerical variables, the process stays almost identical, but rather valid intervals are included and attached to the value. After these are taken care of, zero is assigned to every other category not listed and to the possible missing values too.

The final credit score is counted as the sum of all amended coefficients assigned to matching categories, including the intercept. Score distributions of a training set and a test set can be seen in Figures 5.2 and 5.3. Both distributions look quite similar and display the highest frequency around the score 500.

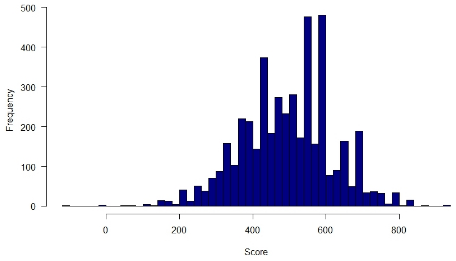


Figure 5.2: Score distribution (training set)

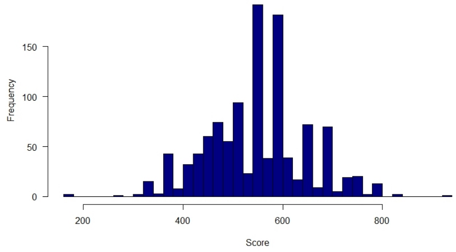


Figure 5.3: Score distribution (test set)

Achieving this state, all valuable information is known. Each observation has its unique credit score and is marked as "default" or "non-default". For instance, we can say that a person with the loan ID x, age y, income z and so forth, has a credit score of w, which is low/high. Being able to claim that is the purpose of this credit risk model.

The probability of repayment (as the dependent variable is "Good", not "Bad") can be counted by the coefficients (credit score divided by 100) entering the logistic function.

Each category and the appropriate score is stated in the Chapter 7 covering the Results in the Table **??**. When setting the final credit score of a potential borrower, this table can serve as a labour-saving tool.

6 Model Evaluation

After the score is computed and each loan properly defined either as "Good" or "Bad", cumulative metrics can be carried out. A creation of distributions of defaulted and non-defaulted loans is the first step in evaluation of the model. The total number of "Goods" and "Bads" is computed for each individual credit score value. To obtain the cumulative distribution, cumulative sums have to be computed first. These cumulative sums are then divided by the total number of "Goods" and "Bads" in order to have the resulting number in percentage rather than absolute numbers. The final cumulative distribution is then plotted (Figure 6.1 and 6.2). Because of the limited number of observations, graph is not very smooth (especially in the case of a test set), however it still has noticeable value.

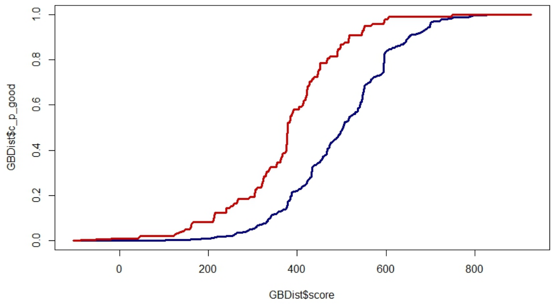


Figure 6.1: Cumulative distribution (training set)

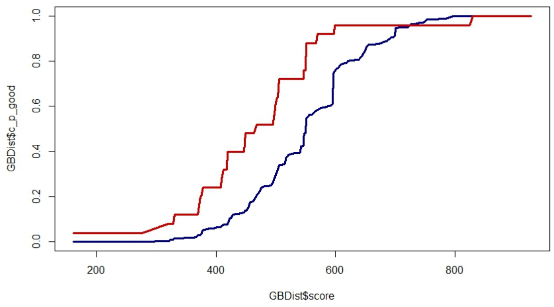


Figure 6.2: Cumulative distribution (test set)

The red line represents bad clients, whereas the blue line represents good clients. A score spread is depicted on the x axis and for its each point, the y axis shows the corresponding cumulative distribution. Therefore, it can be stated that for instance a score 500 or less corresponds to a particular percentage of both good and bad clients population.

It is reasonable, that evaluating the model using training data is straightforward and should not give any distorted outputs. However, more important part is when the remaining 20% of data, saved for the testing, intervenes. For this purpose, the whole process is repeated and the same metrics are computed. It is important that the model works properly with data that were unknown before. Ideally, the train and the test should not look much different. Plotted distributions confirm this claim. Except for the absent smoothness, both graphs look rather similar.

6.1 Basic statistical metrics (KS, GINI)

The cumulative distribution and its plot serves as the basis for the computation of more seizable statistical metrics. Specifically, the Kolmogorov-Smirnov statistics (KS) and the GINI coefficient.

The KS measures the distance between the two cumulative distribution functions, namely its maximum value. The higher the KS, the better the model - the distance between cumulative distributions of ”Goods” and ”Bads” should be as high as possible. Final value occurs between 0 and 100, where 100 means a perfect model.

As opposed to the KS measuring the distance, GINI coefficient captures a share of the area between the diagonal and the curve to the total area above the diagonal [50]. Similar interpretation counts for the GINI, where values close to 100 show that the model has a high predictive power and separates good from bad loans with the best precision. A model with GINI of 100 perfectly predicts which client is going to default and assign to him or her an adequately high score.

Training set				
KS	$GINI$	TB	TG	$TB/(TB+TG)$
43.29	52.49	98	4444	0.02
TB = total bad, TG = total good				

Table 6.1: Statistical metrics (training set)

Test set				
KS	$GINI$	TB	TG	$TB/(TB+TG)$
38.07	46.35	25	1111	0.02
TB = total bad, TG = total good				

Table 6.2: Statistical metrics (test set)

Not surprisingly, both the KS and the GINI values came out larger for the training set than for the test set (Tables 6.1 and 6.2). However, they do not differ radically. According to one of the risk experts in Zonky, Vit Ficl, the GINI of their final model yielded 57, which is not very different from results in this thesis, where both risk experience and the general data knowledge is limited.

Stated in the book The Credit Scoring Toolkit [49] "a Gini coefficient of 50 per cent is more than satisfactory, while less than 35 per cent is suspect, and 30 per cent possibly unacceptable", nearly 50% acquired on the test set seems reasonable.

6.2 Relating WOE and Credit Score

Another interesting metric exploring the model power is rather graphical. It puts the WOE and the credit score against each other by plotting it and examining if both the sign and magnitude resemble themselves. The WOE shows how it manages the division. In the best scenario, the accompanying score should follow its direction and more or less even the magnitude. When it goes to the other side, it could infer some subtle population moves.

After the WOE and the score is computed for each category, a bar plot with resulting outputs is plotted. Every larger bar depicting the WOE contains a smaller bar depicting the score. The WOE is displayed in a light blue colour, whereas the score in a navy blue tint.

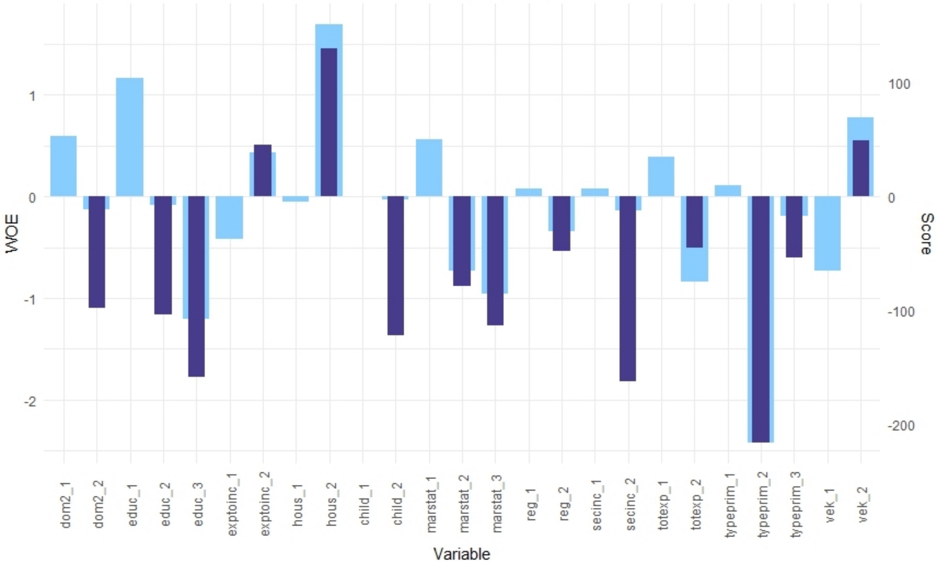


Figure 6.3: Variables performance (training set)

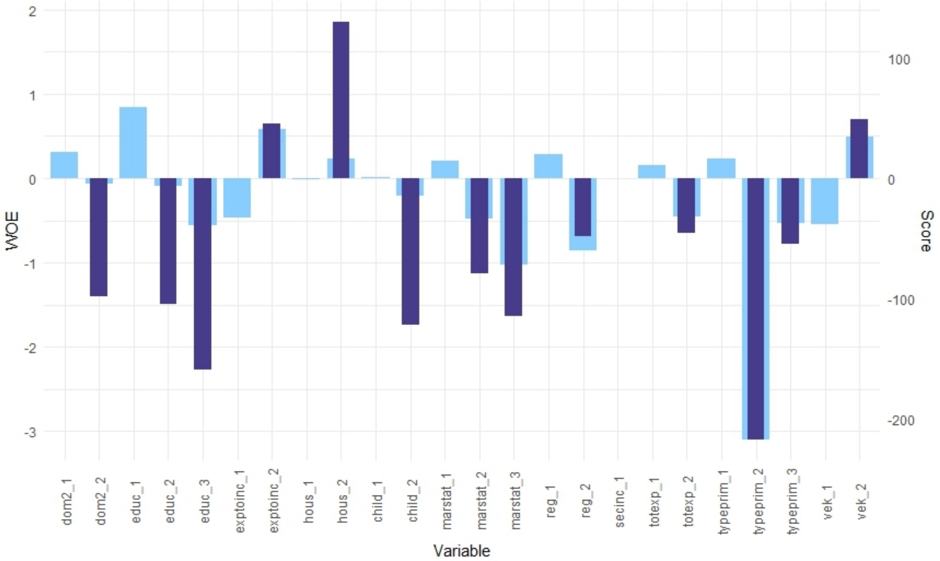


Figure 6.4: Variables performance (test set)

When looking at Figures 6.3 and 6.4, several things can be noticed. First, results look similar for both sets, which again supports the idea of a powerful model. Second, all assigned scores follow the direction of resulting WOEs. It means that the positive WOE is connected to the positive score, therefore according to these graphs, meaning and significance of variables remain largely preserved. Thirdly, some variables are overrated or underrated with respect to the assigned score. For instance, people that fall within a category educ_2 (representing a group of people from high school, with a vocational certificate or with a finished higher vocational school) are assigned a very low score in comparison to its WOE. In contrast with that, a category typeprim_2 (representing students) perfectly matches score with WOE.

However, the main objective of the score is to sort borrowers, which in this case is done correctly.

By implementing a wide range of evaluating metrics, it has been showed that the final model, which is the most important output of this thesis, can be well interpreted and serve as an advisable resource for further studies and conclusions. As the model has been evaluated both by graphic outputs and factual statistical metrics, a reader can develop its own view on the model quality.

7 Results

The main output of this thesis is the final model. It contains only significant categories of chosen variables. In total, 20 variables were chosen to be further grouped and examined. There was not any predetermined or particular number of categories that should be extracted from each variable. Individual treatment of variables was chosen instead, however three to four categories occurred most often.

As a result, 61 categories originated by completing fine and coarse classing and were included in the model, which was later narrowed by the step function earlier described. Among 61 inserted categories, 14 appear to be significant on at least 10% level and are included in the final model presented in the Table 7.1 and by the equation 7.1.

$$\begin{aligned} f(x_i) = \beta_0 + \beta_1(educ\_2) + \beta_2(educ\_3) + \beta_3(dom2\_3) \\ +\beta_4(child\_4) + \beta_5(housing\_4) + \beta_6(maritalstat\_2) \\ +\beta_7(maritalstat\_5) + \beta_8(typepriminc\_2) + \beta_9(typepriminc\_4) \\ +\beta_{10}(region\_2) + \beta_{11}(inc\_avg\_gs\_cc) + \beta_{12}(vek\_cc) \\ +\beta_{13}(totexp\_2) + \beta_{14}(expense\_to\_income\_cc) + \varepsilon_t \end{aligned} \tag{7.1}$$

As was thoroughly described in the Methodology Chapter 6, logistic function was chosen for the regression. It means, that estimated coefficients using MLE are not easily interpretable because of the non-linear input function. For the purpose of this thesis, mainly the sign and a relative magnitude is therefore investigated.

If the sign is positive, people included in that category are more likely to repay the loan. On the other hand, a negative sign represents groups of people that are more likely to default. The same interpretation would be in the case where the coefficients were put to the exponential. Then values larger than one would have the same interpretation as positive values and values lower than one as negative values in the previous case.

Following this method, namely categories *educ_2*, *educ_3*, *dom2_3*, *child_4*,

	Estimate	**Std. Error**	**z value**	**Pr(>** $|z|$**)**
(*Intercept*)	7.0385	1.1445	6.150	7.74e-10***
educ_2	-1.0378	0.4368	-2.376	0.017510*
educ_3	-1.5883	0.5488	-2.894	0.003800**
dom2_3	-0.9781	0.3108	-3.147	0.001648**
child_4	-1.2184	0.5513	-2.210	0.027091*
housing_4	1.2955	0.2838	4.564	5.01e-06***
maritalstat_2	-0.7916	0.2315	-3.419	0.000628***
maritalstat_5	-1.1360	0.6603	-1.720	0.085348 .
typepriminc_2	-2.1724	0.6421	-3.383	0.000717***
typepriminc_4	-0.5393	0.2459	-2.193	0.028286*
region_2	-0.4757	0.2587	-1.839	0.065921 .
inc_avg_gs_cc	-1.6346	1.0228	-1.598	0.109994
vek_cc	0.4872	0.2244	2.171	0.029926*
totexp_2	-0.4526	0.2319	-1.952	0.050979 .
expense_to_income_cc	0.4455	0.2310	1.929	0.053777 .

$^{***}p < 0.001$, $^{**}p < 0.01$, $^{*}p < 0.05$, $^{\cdot}p < 0.1$

Table 7.1: Estimated model

maritalstat_2, *maritalstat_5*, *typepriminc_2*, *typepriminc_4*, *region_2*, *inc_avg_gs_cc* and *totexp_2* have a negative effect on the debt repayment. According to the model, only categories *housing_4*, *vek_cc* and *expense_to_income_cc* have a positive effect, therefore people in this category are more likely to repay the loan.

It is important to emphasise that only categories significant at least on the 10% level are mentioned here. Other categories not included are likely to affect a final loan status too, however in a less considerable manner and they do not contribute in rising the goodness of fit of the model, measured by the AIC.

As already mentioned earlier, a particular credit score is assigned to each significant category and can be computed for each borrower (Table 7.2). According to that score, P2P platforms then distribute potential clients to groups with similar levels of risk.

Variable	Group	Specification	Coefficient	Score
intercept			7.0385	704
education_name	educ0	masters, bachelors	0	0
	educ1	higher vocational school, high school, vocational certificate	-1.0378	-104
	educ2	elementary school	-1.5883	-159
domain_2	dom2_0	com, others	0	0
	dom2_1	cz	-0.9781	-98
children	child0	0, 1, 2, 5	0	0
	child1	3, 4	-1.2184	-122
housing_type_name	housing0	shared housing, partner, parents or family members, rental housing	0	0
	housing1	own flat or a house without a mortgage, own flat or a house with a mortgage, cooperative apartment	1.2955	130
marital_status_name	maritstat0	divorced, married, partner, registered partnership	0	0
	maritstat1	single	-0.7916	-79
	maritstat2	widowed	-1.1360	-114
type_primary	typeprim0	employment, liberal profession, maternity leave, others	0	0
	typeprim1	student	-2.1724	-217
	typeprim2	entrepreneur, pension, self employment	-0.5393	-54
region_name	region0	praha, jihocesky, karlovarsky, vysocina, kralovehradecky, moravskoslezsky, olomoucky, pardubicky, plzensky, stredocesky, ustecky, zlinsky	0	0
	region1	jihomoravsky, liberecky	-0.4757	-48
secondary_income_avg	secinc0	$0 - 0.4733156$	0	0
	secinc1	$0.4733156 - 47.3315589$	-1.6346	-163
vek	vek0	$18 - 31$	0	0
	vek1	$31 - 88$	0.4872	49
total_expenses	totexp0	$0 - 2and9250 - 175002$	0	0
	totexp1	$2 - 9250$	-0.4526	-45
expense_to_income	exptoinc0	$0 - 4.946429e - 1$	0	0
	exptoinc1	$4.946429e - 1 - 6.5$	0.4455	45

Table 7.2: Scoring groups

The probability of repayment and a particular credit score was counted for two randomly chosen observations, from which first is marked as a default and second as a non-default. Characteristics of these two clients are depicted in the Table 7.3.

	1st Observation	Coefficient	Score	2nd Observation	Coefficient	Score
target	0	7.0385	704	1	7.0385	704
dom1	gmail	-	-	centrum	-	-
dom2	com	0	0	cz	-0.9781	-98
purpose	other	-	-	other	-	-
age	37	0.4872	49	28	0	0
sex	MALE	-	-	MALE	-	-
children	0	0	0	1	0	0
education	masters	0	0	elementary school	-1.5883	-159
housing	own flat or a house with a mortgage	1.2955	130	rental housing	0	0
marital status	married	0	0	single	-0.7916	-79
region	praha	0	0	liberecky	-0.4757	-48
population type	praha	-	-	large city	-	-
priminc (sex)	30000	-	-	9500	-	-
secinc (sex)	0	0	0	6500	-1.6346	-163
totinc (sex)	30000	-	-	16000	-	-
type primary	employment	0	0	self employment	-0.5393	-54
total expenses	11455	0	0	8850	-0.4526	-45
expense to income	0,381833333	0	0	0,553125	0.4455	45
term	10,283794	-	-	17,44437	-	-
requested annuity	5000	-	-	1500	-	-
requested amount	50000	-	-	25000	-	-
	Final Score (1st Observation)	8.8212	883	Final Score (2nd Observation)	1.0238	103

Table 7.3: Random observations (probability of repayment, credit score)

The first observation, which is represented by a 37-year-old man is associated with an above-average score of 883. Appropriate probability is computed as follows (7.2):

$$p(y=0|x_{obs1}) = \frac{exp(8.8212)}{1+exp(8.8212)} = 0.99985 \tag{7.2}$$

It means that there is a probability of 99.99% that this borrower will repay the loan. According to his characteristics and their clasification, it seems that this very high probability is reasonable.

The second observation, which is represented by a 28-year-old man is associated with a score below the average of 103. Appropriate probability is computed in the same manner (7.3):

$$p(y = 0|x_{obs2}) = \frac{exp(1.0238)}{1 + exp(1.0238)} = 0.73571 \tag{7.3}$$

For this observation, the probability of repayment is 73.57%. There is a 26.41%-high difference between these two observations.

By including these two examples in this section, the author wanted to demonstrate that due to the built model, both computations are straightforward and easily interpretable.

When banks or in our case platforms implement more sophisticated models with a very high predictive power, these are the outcomes that help them to classify all their clients and to set a cut-off value below which they will not provide a loan anymore.

Conclusion

The P2P lending concept is not well known yet, especially on the Czech market. The aim of this thesis was therefore to introduce the topic closer to the reader, to describe a position of Czech law towards the P2P lending and to examine an uniquely obtained data from the front Czech platform Zonky. To the best of the author's knowledge, it is the first academic work using their directly provided data.

The main contribution is the implementation of a well tested model including only precisely classed categories of chosen variables from the original data set. These models are usually kept from public and their potentially high predictive power is taken as the key ability of the firm to recognise good from bad loans. This counts not only for P2P platforms, but for the front world banks and other institutions facing nonnegligible level of risk. Therefore, an insight into a normally secret market is another thing that this thesis offers.

Moreover, related to model results, an extensive table depicting each significant category with appropriate credit score, extracted by transforming model coefficients, was created. Having "Good" loans as dependent variable, negative coefficients in the final model represent categories of people with a higher likelihood of default. Compared to that, positive coefficients include people characteristics with a positive impact on the loan repayment.

Consistently with other studies (e.g. Lin et al.), borrowers with higher education levels are more likely to repay the loan than these with lower education levels achieved. This could be a case of a financial literacy and presumably a better financial security. Next to that, people living in their own house or flat are a good sign for the lending platform too, as they are associated with early repayments. An age over 31 is adding up to the final credit score which is in accordance with the finding that being a student increases the likelihood of default. One more to mention, a borrower in default having three to four children is a more common scenario than not repaying the loan while having less than 2 children. It probably shows that expenses of these individuals taking care of more children are usually higher.

These findings along with a more detailed list presented in the Results section answers the research question of the thesis. Moreover, the final determinants could serve as a guideline in a marketing or scoring approach for Zonky or other P2P platforms having a similar business plan.

As the Zonky platform is still quite young and its mediated loans are expected to grow, a follow-up modelling should be implemented in the future to compare it with actual results on a more representative data sample. Besides that, text analysis is recommended for further research, as stories written by potential borrowers in order to emphasize their personal achievements are one of the main nuance of Zonky from regular P2P platforms.

To conclude, the P2P lending market has faced several obstacles until now and due to the growing regulatory network will probably face many more in the future. However, increasing lending volumes and reconciliation with still emerging new technologies will help the market to overcome associated risks and to grow further. Precisely elaborated risk models will admittedly belong to useful tools while achieving that.

References

[1] Zopa.com. (2018): *About Zopa*. Zopa.com. [online] Available at: https://www.zopa.com/about.

[2] Bank Underground. (2018): *Peer to Peer - Scale and Scalability*. [online] Available at: https://bankunderground.co.uk/2018/03/01/peer-to-peer-scale-and-scalability/.

[3] Lidovky.cz. (2018): *Sef Zonky: Na reklamu jsme dali jeste vic nez loni. Nase ztrata je planovana*. Lidovky.cz. [online].

[4] Staff, I. (2018): *Financial Institution - FI*. [online] Investopedia. Available at: https://www.investopedia.com/terms/f/financialinstitution.asp.

[5] Deloitte LLP. (2016): *A temporary phenomenon? Marketplace lending - Deloitte*. An analysis of the UK market.

[6] Qi, E., Shen, J. and Dou, R. eds. (2016): *Proceedings of the 22nd International Conference on Industrial Engineering and Engineering Management 2015*. Core Theory and Applications of Industrial Engineering (Vol. 1). Springer.

[7] Serrano-Cinca, C., Gutiérrez-Nieto, B. and López-Palacios, L. (2015): *Determinants of default in P2P lending*. PLoS ONE 10(10): e0139427.

[8] Oxera. (2016): *The economics of peer-to-peer lending*. Oxera Consulting LLP.

[9] Financial Times. (2018): *A crushing blow to peer-to-peer lending dreams (updated)*. [online] Available at: https://ftalphaville.ft.com/2017/11/13/2195766/a-crushing-blow-to-peer-to-peer-lending-dreams/.

[10] Ziegler, T., Reedy, E.J., Le, A., Zhang, B., Kroszner, R.S. and Garvey, K. (2017): *The 2017 Americas Alternative Finance Industry Report*. Cambridge Centre for Alternative Finance.

[11] Ratesetter.com. (2018): *RateSetter Provision Fund - 100% Record - RateSetter*. [online] Available at: https://www.ratesetter.com/aboutus/protection.

[12] Lending Works. (2018): *Quick guide to the Lending Works Shield — Lending Works*. [online] Available at: https://www.lendingworks.co.uk/blog-post/quick-guide-lending-works-shield.

[13] Fundingcircle.com. (2018): *Investment Statistics — Funding Circle*. [online] Available at: https://www.fundingcircle.com/uk/statistics/.

[14] Cnb.cz. (2018): *ARAD - Time Serie System - Czech National bank*. [online] Available at: https://www.cnb.cz/.

[15] Ashta, A. and Assadi, D. (2009): *An analysis of European online micro-lending websites*. Cahiers du CEREN 29. pp.147-160

[16] CrowdfundingHub. (2016): *Current State of Crowdfunding in Europe. An Overview of the Crowdfunding Industry in more than 25 Countries: Trends, Volumes & Regulations*. European Expertise Centre for Alternative and Community Finance.

[17] Alois, J. (2018): *Federal Reserve: Peer to Peer Lending Poised to Grow — Crowdfund Insider.* [online] Crowdfund Insider. Available at: https://www.crowdfundinsider.com/2014/08/47125-federal-reserve-peer-peer-lending-poised-grow/.

[18] Ivashina, V. and Scharfstein, D. (2010): *Bank lending during the financial crisis of 2008.* Journal of Financial economics, 97(3), pp.319-338.

[19] Chang, S., Kim, S.D.O. and Kondo, G. (Autumn 2015-2016): *Predicting Default Risk of Lending Club Loans.* Stanford Engineering Everywhere.

[20] PricewaterhouseCoopers LLP. (2015): *Peer pressure: How peer-to-peer lending platforms are transforming the consumer lending industry.*

[21] The Economist Group Limited. (2015): *The FinTech revolution: A wave of startups is changing finance for the better.* The Economist, 415(8937), 13.

[22] Lin, X., Li, X. and Zheng, Z. (2017): *Evaluating borrower's default risk in peer-to-peer lending: evidence from a lending platform in China.* Applied Economics, 49(35), pp.3538-3545.

[23] Berger, S.C. and Gleisner, F. (2009): *Emergence of financial intermediaries in electronic markets: The case of online P2P lending.* Official Open Access Journal of VHB.

[24] Davis, K. and Murphy, J. (2016): *Peer to Peer lending: structures, risks and regulation.* JASSA The Finsia Journal of Applied Finance.

[25] Transparencymarketresearch.com. (2018): *Peer-to-Peer Lending Market -Transparency Market Research.* [online] Available at: https://www.transparencymarketresearch.com/pressrelease/global-peer-to-peer-lending-market-size.htm.

[26] Anon. (2018): *Lending Club - Statistics.* [online] Available at: https://www.lendingclub.com/info/statistics.action.

[27] Cnb.cz. (2018): *Bankovni statistika - Ceska narodni banka.* [online].

[28] Liang, J. (2017): *A Study on the Financial Risk and Legal Supervision of P2P Lending Model in the Context of Internet Finance.* Revista de la Facultad de Ingenieria, 32(14).

[29] Han, J.T., Chen, Q., Liu, J.G., Luo, X.L. and Fan, W. (2018): *The persuasion of borrowers' voluntary information in peer to peer lending: An empirical study based on elaboration likelihood model.* Computers in Human Behavior, 78, pp.200-214.

[30] Zonky s.r.o. (2018): *Zonky.* [online] Zonky.cz. Available at: https://zonky.cz.

[31] Williams-Grut, O. (2018): *The UK government invests Ł85 million in peer-to-peer lending sector where the watchdog has 'concerns'.* [online] Business Insider. Available at: http://uk.businessinsider.com/british-business-banks-investment-in-peer-to-peer-platforms-after-fca-review-2016-12.

[32] Lo, B. (2016): *It Ain't Broke: The Case For Continued SEC Regulation of P2P Lending*. Harvard Business Law Review (HBLR).

[33] Meyer, T., Heng, S., Kaiser, S. and Walter, N. (2007): *Online P2P lending nibbles at banks' loan business*. Deutsche Bank Research, 2(1), pp.39-65.

[34] Milne, A. and Parboteeah, P. (2016): *The business models and economics of peer-to-peer lending*. The European Credit Research Institute (ECRI).

[35] Verstein, A. (2011): *The misregulation of person-to-person lending*. UCDL Rev., 45, p.445.

[36] Tsai, K., Ramiah, S. and Singh, S. (2014): *Peer Lending Risk Predictor*. Stanford University CS229.

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

[38] Wang, Z., Jiang, C., Ding, Y., Lv, X. and Liu, Y. (2017): *A novel behavioral scoring model for estimating probability of default over time in Peer-to-Peer lending*. Electronic Commerce Research and Applications, pp.74-82.

[39] Carmichael, D. (2014): *Modeling default for peer-to-peer loans*. University of Houston, C.T. Bauer College of Business.

[40] Herzenstein, M., Sonenshein, S. and Dholakia, U.M. (2011): *Tell me a good story and I may lend you money: The role of narratives in peer-to-peer lending decisions*. Journal of Marketing Research, 48(SPL), pp.S138-S149.

[41] Freedman, S. and Jin, G.Z. (2017): *The information value of online social networks: lessons from peer-to-peer lending*. International Journal of Industrial Organization, 51, pp.185-222.

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

[43] Zhang, Y., Li, H., Hai, M., Li, J. and Li, A. (2017): *Determinants of loan funded successful in online P2P Lending*. Procedia Computer Science, 122, pp.896-901.

[44] Meng, F. (2016): *What are the determinants of lending decisions for Chinese Peer-to-Peer lenders?*. University of Twente, Master Thesis, Profile of Financial Management.

[45] Klafft, M. (2008): *Online peer-to-peer lending: a lenders' perspective*. Fraunhofer ISST, Berlin, Germany.

[46] Deloitte LLP. (2016): *Credit scoring - Case study in data analytics.*

[47] Wooldridge, J.M. (2013): *Introductory econometrics: A modern approach*. Cengage learning.

[48] Ala'raj, M. and Abbod, M.F. (2016): *Classifiers consensus system approach for credit scoring*. Knowledge-Based Systems, 104, pp.89-105.

[49] Anderson, R. (2007): *The Credit Scoring Toolkit.* Oxford: Oxford University Press, UK.

[50] Rezac, M., Rezac, F. (2011): *How to Measure the Quality of Credit Scoring Models.* Czech Journal of Economics and Finance. Masaryk University, Brno.

List of tables and figures

List of Figures

List of Tables