Charles University

Faculty of Social Sciences
Institute of Economic Studies

MASTER'S THESIS

Portfolio diversification on P2P loan markets

Author: **Bc. Petr Polák**

Supervisor: **PhDr. Ing. Jiří Skuhrovec**

Academic Year: **2016/2017**

Declaration of Authorship

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

Prague, January 6, 2017

Signature

Acknowledgments

I am grateful especially to my family members for the support during my studies. Further I want to thank my colleagues, namely Adam Kučera, Hana Hejlová, Oldřich Koza, and Marek Rusnák for their helpful comments. My specially thanks are to my supervisor PhDr. Ing. Jiří Skuhrovec, for his comments and willingness to help.

Abstract

This thesis presents ways how investors can construct optimal portfolios on online peer-to-peer lending platforms. Thesis uses standard portfolio theory and unique dataset from Lending Club platform of over 886 thousand loans issued since 2008 till the end of 2015. Firstly, this thesis shows that there is a nonzero covariance between loans from different credit grades and it is necessary to include it in portfolio management optimization. Secondly, the thesis with the help of a logistic regression identifies loan default determinants. Using the default predictions, the portfolio performance can be improved significantly. Thirdly, the thesis simulates diversification benefits stemming from investing into multiple loans.

JEL Classification	D12, G11, E41, E44, G21
Keywords	Peer-to-peer lending, diversification, portfolio, default risk

Author's e-mail	64212974@fsv.cuni.cz
Supervisor's e-mail	jskuhrovec@gmail.com

Abstrakt

Tato práce prezentuje způsoby, jakými mohou investoři sestavit optimální portfolio na on-line peer-to-peer úvěrové platformě. Práce vychází ze standardní teorie portfolia, kterou aplikuje na unikátní data obsahující 886 tisíc půjček poskytnutých mezi lety 2008 a 2015 prostřednictvím platformy Lending Club. Nejprve ukazuje, že mezi různými kreditními skupinami existuje nenulová kovariance, kterou je nutno zohlednit při optimalizaci portfolia. Poté za využití logistické regrese zkoumá, jaké faktory předpovídají selhání úvěrů. S využitím předpovědí se výkonost portfolia významně zlepší. Následně práce provádí simulační cvičení, které dokumentuje přínosy diverzifikace plynoucí z investování do vyššího počtu úvěrů.

Klasifikace JEL	D12, G11, E41, E44, G21
Klíčová slova	Peer-to-peer půjčky, diversifikace, portfolio, riziko selhání

E-mail autora	64212974@fsv.cuni.cz
E-mail vedoucího práce	jskuhrovec@gmail.com

Contents

List of Tables

List of Figures

Acronyms

CAPM Capital Asset Pricing Model

CML Capital market line

DTI Debt to income

FICO Fair, Isaac and Company

MLE Maximum Likelihood Estimation

P2P Peer-to-peer

VCV variance-covariance

Master's Thesis Proposal

Author	Bc. Petr Polák
Supervisor	PhDr. Ing. Jiří Skuhrovec
Proposed topic	Portfolio diversification on P2P loan markets

Motivation Loans to consumers or entrepreneurs in developed countries have typically been provided by financial institutions. Recently, however, online social lending platforms (also called P2P loan markets) have emerged – first lending platform was launched in UK in 2005. Today, one of these is Prosper.com in the US, on which lenders can give their money directly to borrowers without the intermediation of a financial institution. Several studies used prosper data for analysis of this website as it is publicly available. Nevertheless, there are more platforms in different countries which I want to use it for analysis.

So far the research focused mainly on determinants of interest rate and success rates using only information from Prosper. I work with an idea that further examinations are necessary to identify similarities and differences between the traditional banking and the P2P lending market and we will focus on some of them, that we are able to quantify and influence the loan provision. Both markets differ very much in the average size of the funded loan, the screening process, as well as the knowledge and resources to evaluate and manage risks. That kind of research might clarify whether results from former research in the traditional banking market are applicable to the P2P lending market and vice versa.

This these will focus on investor perspective with the aim to compare investing into P2P loans with investment into diversified portfolio We will examine empirically what is the optimal portfolio constitution for an investor and if investor can increase expected return by pre-selecting loans and credit grades compared to random investing.

Hypotheses

Hypothesis #1: Diversification can be achieve via investing into multiple loans.

Hypothesis #2: Investing into more credit grades will result in more efficient portfolio.

Hypothesis #3: Expected returns of different credit grades are correlated.

Methodology Portfolio analysis is an optimization problem with given conditions, we will solve it numerically, because we are interested in the portfolio composition. Furthermore, probability of the defaults will be examined using binary logistic regression as that is the proper framework for it. Lastly, diversification effects will be simulated using own programme in R/Python, similiar to the previous studies.

Expected Contribution This thesis will bring three key contributions to the academic literature. Firstly, only one previous study tried to use portfolio theory on P2P loan markets, but they claimed that covariance between credit grades can be ignored. This thesis will calculate the variance-covariance matrix and provide empirical evidence about this statement. Secondly, this thesis will analyse, if diversification effects which can be applied on stock market (by investing into multiple securities, the risk will decrease) can be also observed on the P2P loan markets and how many loans should optimal portfolio have. Thirdly, this thesis will evaluate, how can investors increase their expected returns if they pre-select the loans by calculation of probability of default.

Outline

1. What evidence about P2P lending platforms is found in the economic literature
2. Description of investigated P2P lending platform: How the P2P lending works
3. Data gathering and description: I will describe the used data and provide summary characteristics
4. Theoretical models: Portfolio theory, diversification effects
5. Results: I will discuss my baseline results together with some robustness check
6. Concluding remarks: I will summarize my findings and their implications for future research and investor behaviour

Core bibliography

Jefferson Duarte, Stephan Siegel, and Lance Young, Trust and Credit: The Role of Appearance in Peer-to-peer Lending, Rev. Financ. Stud. (2012) 25(8): 2455-2484 doi:10.1093/rfs/hhs071

Ashta, Arvind and Assadi, Djamchid, Do Social Cause and Social Technology Meet? Impact of Web 2.0 Technologies on Peer-to-Peer Lending Transactions (October 9, 2008). Cahiers du CEREN, Vol. 29, pp. 177-192, 2009

Berger, Sven C. and Gleisner, Fabian, Emergence of Financial Intermediaries in Electronic Markets: The Case of Online P2P Lending (March, 11 2010). BuR Business Research Journal, Vol. 2, No. 1, May 2009

Ashta, Arvind and Assadi, Djamchid, An Analysis of European Online Micro-Lending Websites (June 4, 2009). Cahiers du CEREN, Vol. 29, pp. 147-160, 2009

Iyer, R., A. Khwaja, E. Luttmer, and K. Shue (2010): Inferring Asset Quality: Determining Borrower Creditworthiness in Peer-to-Peer Lending Markets, NBER Working Paper No. 15242.

Bachmann, Alexander; Becker, Alexander; Buerckner, Daniel; Hilker, Michel; Kock, Frank; Lehmann, Mark; Tiburtius, Phillip; Funk, Burkhardt, Online Peer-to-Peer Lending – A Literature Review, Journal of Internet Banking & Commerce;Aug2011, Vol. 16 Issue 2, Special section p1

Ian Galloway, 2009. Peer-to-peer lending and community development finance, Community Development Investment Center Working Paper, 2009-06, Federal Reserve Bank of San Francisco.

Markowitz, H. (1952). Portfolio selection. The journal of finance, 7(1), 77-91.

Sharpe, W. F. (1963). A simplified model for portfolio analysis. Management science, 9(2), 277-293.

Evans, J. L., & Archer, S. H. (1968). Diversification and the reduction of dispersion: an empirical analysis. The Journal of Finance, 23(5), 761-767.

Author Supervisor

Chapter 1

Introduction

Peer-to-peer (P2P) lending is a form of alternative finance which is based on direct transactions between private individuals. These transactions are done mostly on-line using specialized platforms, where financial institutions serve only as intermediaries. First commercial on-line P2P lending platforms were launched in 2005, which makes this type of business relatively new to economists and therefore it is also quite a young and only little investigated research area. Nevertheless, within last years this field of research attracted scientists' attention not only from department of economics, but also from other social sciences and information technology.

Existing research has been mainly focused on relationships between lenders and borrowers and determinants which play the key role in the loan provision process. The aim of this thesis follows recommendation of some previous studies and shifts the focus from behavioural economics more towards the financial perspective and puts loan provision into perspective of an investor and connects it to the microeconomic portfolio theory.

On-line P2P lending platforms create market environment for borrowers and investors (lenders). Loan provision process is thus governed by private agents, not institutions. Borrowers in general describe the reason and purpose of their loan request and accompany it with information about their financial status (e.g. income, credit history, home ownership, other debts etc.). Using these information, lenders make offers on loans by specifying amount they are willing to invest and under what interest rate. Using mostly auction mechanisms, the loan is granted or not. This process is completely different from traditional banking system, where borrower approaches some banking institution who makes the decision. With no such financial institution involved, borrower may get loan that traditional system would not provide him with or get it under better conditions. For lenders on-line P2P lending platforms provide investment possibilities with again specific terms. Since the platform provides the environment for deal making, it charges some fees when the loan is successfully provided (Galloway 2009; Bachmann *et al.* 2011).

Success of P2P lending platforms has several reasons. The Internet and information technology progress provide broader information access, better credit scoring and also accessibility to almost everyone. In addition to that, consuming way of life leads to significant increase of individual debts of people. On the opposite site, there are low returns on savings and overall consumers' lending and borrowing behaviour changes. This creates very favourable environment for development of alternatives to traditional ways of investing and banking and one of the area is P2P on-line lending. People also want to be more independent in their finance decisions and depend less on financial institutions. Consumers do not trust them very much according to a Forrester[1] study as they believe that banks put their own interests ahead of their customers. P2P lending platforms provide a suitable marketplace where borrowers and lenders make decisions on their own without a bank dictating the conditions (Slavin 2007). In addition, low interest rate environment created after global financial crisis forces investors to seek new forms of investments that provide more interesting results. In comparison with saving account returns of roughly 1%, it is more attractive for households and small investors to join the P2P platforms and claim returns of 5% or higher.

The aim of this thesis is to provide analytical solution to portfolio decision in context of P2P markets. We will use portfolio theory to create optimal portfolio on P2P platform, which is something investors should be aiming at, because that approach gets them to the best available investment opportunities. This thesis uses standard technique of default prediction and investigates whether the portfolio created with the default prediction performs better than a random portfolio. Key part of the portfolio management is diversification. This thesis further analyses the benefits of diversification into multiple loans. Finally, we test whether randomly selected portfolio or equal weights portfolio perform

[1] forrester.com

worse, than portfolio created using statistical techniques.

In this thesis we focus on the common aspects of all P2P lending platforms. P2P platforms differ very widely in the type and amount of information they provide about borrowers. In 2015 a new platform Zonky[2] was launched in the Czech Republic. As an investor, all you know about the borrower is the credit rating, income type and country region. Then along with the loan term and amount and borrower written story, investors make their decisions. These key characteristics of loans are reported by every P2P platform and investigation of these aspects makes this study replicable on datasets from other platforms and results can be compared. We also capture the current trends on the P2P platforms from the investors' perspective and bring a light into unclear investment making process on these platforms. P2P lending platforms can only exist if there are investors willing to invest and ability to diversify credit risk might attract investors. Academic literature does not provide sufficient evidence on P2P platforms and this thesis tries to fill the gap by providing unique evidence on diversification.

The thesis is organized as follows. The next Chapter 2 presents P2P lending market with detailed description of lending process and differences between platforms. Chapter 3 summarizes all available and related literature findings about P2P platforms. Chapter 4 presents data and model used for quantitative analysis followed by Chapter 5 where we present results of our analyses and finally Chapter 6 concludes.

[2] www.zonky.cz

Chapter 2

P2P lending market

2.1 Idea of P2P lending

As Everett (2010) points us, the concept of private loans is rather the traditional model and definitely not completely new business model. Originally private subjects or individuals provided loans without intermediaries or any other institution, later on banking system emerged and put money lending process without mediation aside. Today, when the internet creates a possibility to transform the process into the virtual world, on-line P2P lending platforms emerged and new phenomenon arose. Garman *et al.* (2008) identified 24 platforms worldwide, half of it alone in the United States of America, where the first platform Prosper.com was launched in 2006. The first platform – Zopa – was launched one year earlier in United Kingdom and in 2005 won "Internet Innovation of the Year" CNet Technology award. Since then the number of platforms emerged.

All platforms are using the same basic idea – they connect investors (or lenders) and borrowers and act as intermediary in the process. Borrowers seek for best credit conditions given their various level of credit risks and history, while lenders search for best investment opportunities in the form of highest possible profit for certain risk levels or highest utility in the form of social satisfaction (Greiner & Wang 2009). For the mediation of the process and provision of necessary institutional environment the platform charges fees which are used to cover its costs and earn profit.

The success of these platforms or online lending communities can be judged by the evolution of number of their members and volume of loans provided there. By the end of 2016 there are over 360 000 consumer loans currently funded for UK borrowers totalling in £1.3 billion of outstanding principal.[1] These are just platforms in the UK. In the USA the P2P platforms overgrew it. Morse (2015) extensively discusses both borrowers and investors perspectives and claims that P2P platforms should be able to offer pricing benefits for both sides and Käfer (2016) considers P2P lending as a part of the shadow banking sector.

2.2 P2P lending platforms – new form of alternative financing

Lending platforms adopted several different business models, some of them changed it over time and this thesis will provide a closer look at different approaches in the following lines. Legal requirement which also cause that today most of online lending platforms operate only on a national level (Berger & Gleisner 2009), influence business models too. Ashta & Assadi (2009) identifies 2 types of P2P lending platforms: non-commercial and commercial. Main difference makes the lender's incentive and his expectations about the loan and return of his investment. Non-commercial platforms often operate globally and lenders get little or no reward for risks they undertake. This concept is close to "donating" small loans to projects in regions that are very poor or underdeveloped and other forms of loans would never be provided. Commercial platforms are usually focused on one national market and lenders seek reward in the form of interest for the risk they are taking.

Commercial platforms can be further separated into two groups by focus – those who aim at individuals with loans usually up to about 20 000 USD and those who focus on businesses or invoice trading. Mostly the loan provision process is using crowd-funding methods, which means that several investors provide funding for one loan or project, but there are also several platforms, where each loan is funded by one individual investor, who sets the conditions.

Banking theory separates two types of credit for corporate customers based on

[1] Author's own calculation based on number from these platforms and current exchange rate.

information provided: relationship loans and transactional loans. Former type is characterized by inside information and also well documented by empirical literature since the data lending relationship is more available. Information content determines the availability and pricing of credit and is therefore highly valuable. Recent development of technology make new data available for credit-market transaction that fit description of transactional loans. While lending standards are identical for these two models, loan officers are using more subjective credit-assessment when loans are provided in-person compared to on-line application (Agarwal & Hauswald 2008). Since all loans request in P2P lending industry are made on-line, lenders make their decisions based on mostly on public information and such process is closer to transactional form of loans, than relationship loans.

Modern P2P platforms are created with the aim to make the consumer lending more pleasant and rewarding to every stakeholder. These on-line marketplaces works in a very similar way as well-known auction platforms like eBay or any e-shop (Slavin 2007). Instead of items, lenders are bidding on investments, borrowers can use their stories and pictures to attract the investors to bid.

Crowdfunding is a term used for financing businesses and small enterprises. This alternative way of financing is quite popular among new start-up companies. They can get necessary starting capital in exchange for small reward or specific product. Investors do not necessary look for monetary reward, but are keen to support new or innovative ideas by small amount just because they like it. For the businesses it is also a way how to present itself. Situation in which investors borrow money and expect to get them back together with the interest is called *peer-to-peer lending*.

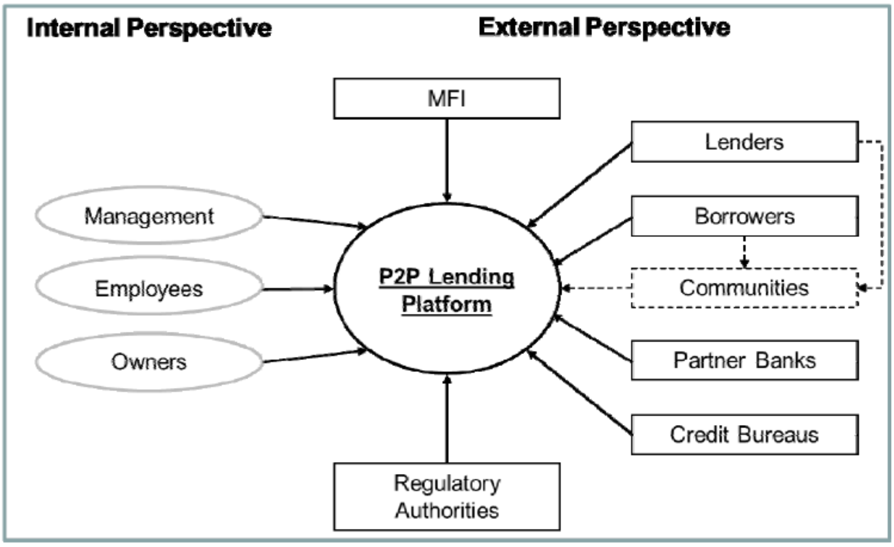
P2P platforms use different business models in placing bids as well. There are platforms who let investors to bid for specific loans, but there are also others who let lenders to bid on a specific packages of loans – this approach works in a similar way as investment funds. In the Czech Republic platform Ferratump2p.cz uses this type of financing and offers investments into packages of consumer loans. Some platforms still use Dutch auction mechanism to determine the target interest rate of a loan. Such mechanism was originally used by the pioneering platform Prosper.com, however in 2009 their business model changed and interest rate is given by the platform itself. How this rate is set is part of know-how of this platform and only general principles are known. Since the platform does not provide publicly all data used for the credit scoring, perfect reverse engineering of the scoring process is not possible.

2.3 Lending process

In 1983 Grameen Bank was established and micro loans to the poor were offered by it. Extensive development of micro-credit companies followed and was further bound with spread of the Internet. Principles of microcredit have been applied to peer-to-peer lending that can have different target group of borrowers and also investors. New models offer an opportunity for non-bankable borrowers to have access to a line of credit. Retail banks with risk management require collateral or credit history to provide a loan, this requirement is removed and system use also mutual trust. In addition to that P2P platforms offer small personal loans, filling a gap on the credit market. Moreover, peer-to-peer lenders evaluate the whole loan request containing also soft information that bank models are not accounting for, especially when motivation of these lenders is driven not only by rate of return, but additional factors like doing social good or willingness to invest in some specific area (Krumme & Herrero-Lopez 2009).

Figure 2.1 presents functional chart of a P2P lending platform. Internal perspective including management, owners and employees does not lie in the centre of our focus. We assume that these entities want the platform to prosper and thus follow recommendation of regulatory authorities and laws. The internal part sets the rules of the platform, determine business model and oversees the marketplace. External perspective comprises of other players and institutions that influence the P2P platform and are part of the lending process. Lenders and borrowers can form communities, in some cases these communities are form within the platform in other cases the communities are formed in the first step and then enter the platform as one entity. That is not always true, some platforms do not work with communities and therefore dashed line is used. Important players are partner banks – P2P platforms serves as intermediary in the lending process, but loans are provided though banks, since the money are sent to bank accounts and also banks take the repayments often directly from the account. Credit Bureaus are used by all platforms to access the credibility of the borrowers, these institutions are often closely connected to the local le-

Figure 2.1: Functional chart of a P2P lending platform



Source: Bachmann *et al.* (2011, Fig.1).

gal environment and collect credit data. For P2P platform a close cooperation with such institutions is very valuable and often almost necessary for effective business.

Firstly, the platform needs to verify identity of each party. It is therefore necessary for each stakeholder (borrower and lender) to create an account on the platform and have a bank account to transfer funds. Then the process in the mainstream is as follows (there might be tiny differences for specific platforms):

- Borrower applies for the loan and is required by the platform to provide all necessary documents to prove his/her income and credit history.
- The platform evaluates the loan request, verifies information with other available data sources (Internet, social networks, insolvency registry, credit bureaus, etc.) and if approved, assigns credit score and interest rate.
- Approved loan request is put on the platform together with some information (interest rate, income type, credit score, loan purpose, story, Debt to income (DTI), etc.).
- Investors browse through loan requests and make lending decisions – amount of their participation.

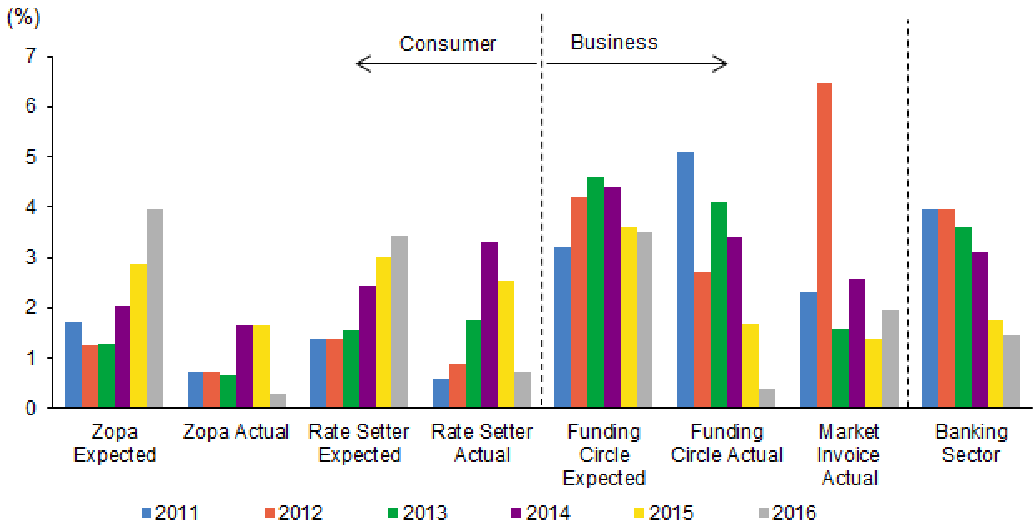
- If the whole amount is collected, loan is provided to the borrower.
- Borrower repays monthly instalments and the platform deducts its fees and distributes the principal and interest payments between investors.
- Platform communicates with the borrower in the case of default and takes legal action on behalf of the investors to recover as much as possible from the claim.
- From legal perspective, the platform is only an intermediary, borrowers enter into legally binding contracts with the lenders.

Originally, the lending process was a Dutch auction limited by time, usually 2 weeks. Borrower indicated the highest acceptable interest rate and lenders made the bid (money they are willing to lend) on listings together with the indication of a minimum rate they are willing to accept. Even if there are enough funds from the lenders to provide the loan, other investors can still place their bids and undercut others with lower interest rates. When the designated time has elapsed, bids are sorted from the lowest to the highest interest rate. For the loan to be provided, sum of amounts in the auction offered by lenders has to at least cover the amount requested by the borrower. The level of loan's interest rate is given by the highest interest rate that is still needed to fund the loan. All lenders will receive the determined interest rate no matter how low their required interest rates were. If in total insufficient amount in offered, the loan is not granted. As stated, this is no longer true, since the majority of platforms determine the interest rate ex ante, so the borrowers can choose, if their loan application should be offered on the marketplace.

2.4 Default rates

Peer-to-peer lending is consumer lending without collateral. Platforms take legal actions if there is a problem with the borrower, but there is a high risk for investor of losing everything. Information about default rates and expected returns is important, yet some platforms do not provide this information publicly. Figure 2.2 provides a look at the peer-to-peer lending market in the United Kingdom and separates it into two parts – consumer lending and businesses.

Figure 2.2: Expected and actual default rates in UK



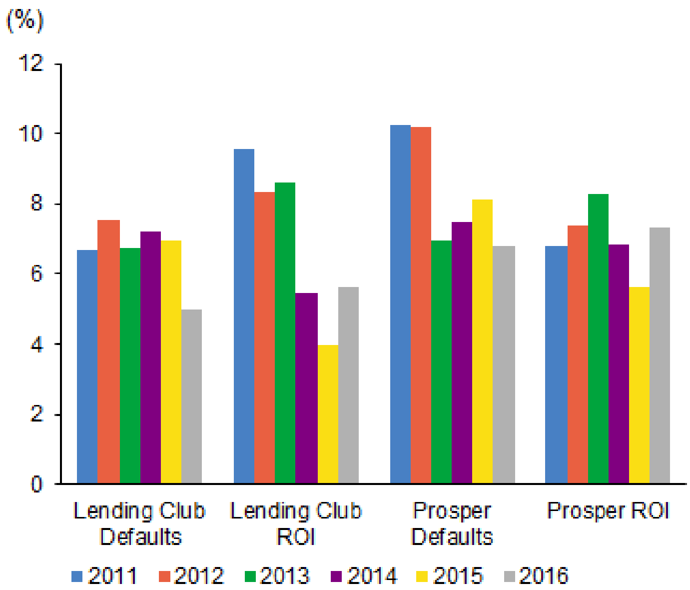
Source: Data from Zopa, RateSetter, Funding Circle, Market Invoice, World Bank.

Figure 2.2 illustrates the P2P consumer lending market. Consumer lending is less risky and expected rates are higher than actual. Default rates for 2011 and 2012 are final, since those loans are already repaid or defaulted. Some of the loans provided in later years are still active, hence actual default rates can change significantly especially for the last two years. Rate Setter platform runs provision fund to cover loses from defaulted loans and was able to ensure, that no investor has lost anything so far. Interesting is the comparison with the banking sector as whole (majority of loans in the banking sector are provided to households and large portion is for mortgages). This shows that risk management of P2P platforms is good and similar to the banking sector.

P2P platforms in the USA are riskier, but still offering interesting total returns (Figure 2.3). Default rates on the two largest platforms in the USA (Lending Club and Prosper) are more than double than in the UK, but return on investment (ROI) is around 5% on Lending Club and 7% on Prosper.com. These numbers can change significantly with the economic situation as default rates on Lending Club reached 14% in 2014[2] and ROI was negative on Prosper during the Global Financial Crisis[3]. Prosper changed its business model after the crisis and interest rates are no longer determined by the auction, but by the platform based on the credit scoring process.

[2] https://www.nsrplatform.com
[3] http://www.lendacademy.com/lending-club-prosper-default-rates/

Figure 2.3: Return on investment and default rates – US platforms



Source: Data from Nsrplatform.com.

2.5 Credit rating

One of the highly discussed and investigated problem of P2P platforms is the information asymmetry between borrowers and lenders. Lenders are provided with only limited information about the borrower. Borrower knows his credibility well and in order to get the loan he might be tempted not to reveal all information about himself. Next to it, people might turn to P2P platforms if they are rejected by banks which could result in adverse selection problem. To reduce risks connected with infromation asymmetry and to help lenders understand credit risk of borrowers, each borrower receives credit rating from the platform. Usually A* for the lowest risk and D or E for high risky borrower. Lending Club uses Fair, Isaac and Company (FICO) scoring to determine the credit rating. Minimum FICO score for considering the borrower is 640 (the generic or classic FICO score is between 300 and 850).

FICO score is calculated from different pieces of credit data gathered from US national credit bureaus. Score is designed to measure the risk of default by evaluation and combination of several different factors in one's personal financial history. Exact formulas are not public, but 35% of the score is payment history, 30% debt burden and balance on accounts, 15% length of credit history, 10% credit mix (types of credit used), 10% recent searches for new credit.

2.6 Interest rate determination

The interest rate is calculated using borrower's characteristics. Loan is then funded as soon as the sum of lenders bids covers the required amount of the loan (Collier & Hampshire 2010). A perfect example is Lending Club[4] in USA whose data we also use in our empirical analysis. They set the final interest rate as *Lending Club Base Rate + Adjustment for Risk & Volatility*. The starting point for base interest rate is the middle of the spread between the interest rate for unsecured consumer credit as published in Federal Reserve Board Consumer Credit G.19 Release and the average interest rate for 6-month certificates of deposit as published in Federal Reserve Board Selected Interest Rate H.15 Release. Furthermore Lending Club modifies this initial base rate for economic slowdowns or expansions, whether borrowing requests exceed investor commitments or vice-versa, and rates set by other lending platforms and financial institutions. The reason for taking these two rates published by Federal Reserve Board is to make it profitable for borrowers as well as lenders. The former rate reflects the average interest rate at which borrower members could generally obtain other financing, the latter rate reflects a widely available risk-free alternative investment. After the base rate is set, assumed default rates are estimated. Lending Club uses 35 sub-grades for loan credit score. Afterwards further adjustment based on volatility factors and other not specifically mentioned factors is done.

2.7 Diversification and investor preferences

All platforms publish on their websites, that investors should invest in several loans to diversify their portfolio and based on their risk attitude consider investing into different credit ratings. P2P platforms differ widely in the amount of information they provide about each loan. Current trend among platforms is to create autoinvesting schemes, where investors define terms (duration of the investment) and credit rating composition of their portfolio and platforms use these mechanisms to bid on loans. In the end, what matters the most is the credit grade, which scale is almost the same on all platforms (credit grade represents the labelling from A to G with A being the low risk loan and G

[4]http://www.lendingclub.com/public/how-we-set-interest-rates.action

with the highest risk). For each credit group platforms also announce expected return, expected loss and expected net return after losses. In call cases, the credit category with highest risk is always the one with the highest expected return.

Recently, platforms started to offer auto investment options based on the investors' stated preferences – key factors are credit grade, term and amount. Aim is to attract wider audience, because not everybody fully understands and knows how to evaluate all the financial information provided about the borrower. That is the current trend in this area, like when investing into stocks you do not look at the company itself, but relate to the information provided by your broker or credit rating agency that does it for you. Same situation is present on P2P loan markets.

Chapter 3

Literature overview

Most of the research has focused on these stakeholders (borrowers and platforms), since they are the most important part of the lending process, and also what determinants do play key role in the success of loan provision (Freedman & Jin 2008). Some elaborate more on the regulatory restrictions present in different countries and other legal requirements under which P2P platforms operate. Original banks, credit bureaus or other external agencies are in some cases required to facilitate process of lending or for verification of borrower's data (Galloway 2009). For fully funded loans, verification processes of borrower's ability to pay, including verification of income have been implemented even directly by some platforms like prosper.com. (Garman *et al.* 2008)

Two key indicators of loans – whether the loan is provided and what is the interest rate – have been of the major focus of investigation in scientific research. Lending process can be also seen as information asymmetry issue. The challenge is to overcome the principal-agent problem (Jensen & Meckling 1976). Lenders want to get as much information about borrowers as possible, so they can make an informed decision about their investments. P2P lending platforms are there to help lenders with their decision by providing them with valid data about borrowers, who might be tempted to hide some information or provide false data about themselves to get better conditions, mostly lower interest rates. Borrowers are required to provide valid information about themselves as well as their financial and credit history. To avoid fraud and falsified data, external agencies' verification are required. P2P platforms also use demographic and social information provided by borrowers, but those are mostly not standardized.

Determinants that have major influence on the success of the loan provision and its interest rate are called characteristic determinants (Bachmann *et al.* 2011).

Puro *et al.* (2010) creates borrower decision aid, to help them fund their loan request and is followed by Wu & Xu (2011). Yum *et al.* (2012) uses data from largest South Korean platform and focuses on group wisdom to overcome information asymmetry problem. Lee & Lee (2012) investigates herding behaviour using data from South Korean and compares his result with Herzenstein *et al.* (2011)

One of the most important characteristics determinants are obviously financial ones, since they are the key indicator of creditworthiness. Examples of financial characteristics of borrower are information about income and expenses, debt ratios, credit ratings, or property ownerships. Credit ratings, that aggregate financial and personal information info a number of scale, are often estimated by external agencies. Some of P2P platforms including prosper.com operate also with current open credit lines of borrowers (Klafft 2008; Bachmann *et al.* 2011).

One of the most discussed financial determinant is the borrower's credit rating and its effect on funding success and interest rate of the loan. Klafft (2008) seeks determinants used in traditional banking for interest rate determination also in P2P lending systems, namely in Prosper.com. His analysis shows that credit rating has the highest impact on the interest rate, followed by the debt-to-income ratio. Other information like property ownership were found to have almost no effect on interest rate. Interesting fact is possession of verified bank account, since it does not influence the interest rate, but is the most significant variable determining the funding success itself. Klafft (2008) points out, that credit rating despite its sophistication and in fact incorporation information about the bank account, is the second most significant variable. Another important remark made by the author is, that borrowers with weak credit-ratings that are unable to get a loan in the traditional banking system, are also very unlikely to get a loan via P2P platform.

Iyer *et al.* (2009) evaluate how lenders are able to use available information other than credit score to determine the creditworthiness of a borrower. Authors depicted in a chart (see Figure **??**) the relationship between credit scores and interest rates. There is a clear connection between credit score and the interest rate of the loan. AA being the best rating the borrower can get and HR (high risk) the lowest possible. Prosper.com uses more precise measure up to 1000 and also letters for easier division into performance groups. The research asked, if lenders are able to distinguish between borrowers belonging into different credit score groups using other information provided in the loan request. They conclude, that lenders are able to derive creditworthiness also from other data about the borrowers than credit score. The key indicators were debt ratios, credit inquiries and number delinquencies, which are all standardized measures. That implies an important finding that P2P markets have screening ability and thus these emerging markets have the ability to complement traditional lending markets.

P2P lending platforms operate also with demographic data about borrowers. Several studies investigated correlations between those characteristics and loan performance, but some part of the research also focused on the question of discrimination based on the race of the borrower. African Americans have been found to have less change of getting funded compared to other races (Pope & Sydnor 2011; Herzenstein *et al.* 2008). Ravina (2012) finds that African Americans have to pay higher interest rates compared to other races. This raise a question if P2P markets are efficient and how can this be explained by the economics theory. Nevertheless, Pope & Sydnor (2011) find that estimated net return of African American loan listings is lower due to higher relative default rates.

Pope & Sydnor (2011) finds that age is also an important determinant of loan funding success. Borrowers reliability can change with age and experience and also young people usually do not have any credit history. As a base group were selected 35-60 year olds, and compared to them, there is a 0.4 to 0.9 percentage points higher chance of getting funded for those who appear younger than 35. For the other end, those who appear to be 60 years and older have between 1.1 and 2.3 percentage points less likelihood to succeed in a loan acquiring process.

Barasinska (2009) and Pope & Sydnor (2011) also find gender differences in ex-ante loan risk and return. Former author finds surprising, that women lenders are less risk-averse compared to their male peers – female lenders are funding lower credit ratings and low interest rate loans with higher probability than male lenders. Author provide as an explanation for such behavior altruistic motives and that males are seeking profit. Pope & Sydnor (2011) concludes that the estimated return on loans to single women is about 2 percentage points less than for single men.

3.1 Default risk

Successful investments require either low default rates or high recovery ratios (low losses). Given the provided data, we are able to calculate both, but for better investment decision, ex-ante identification of loan that is going to default is crucial.

There is an extensive literature that indetifies the determining factors of default risk. One of the first studies is by Iyer *et al.* (2009) on early pre-crisis Prosper data with focus on screening abilities of P2P markets. Their methodology incorporating aggregate credit category shows, that credit category rather than individual credit score allows lenders to make better decision about target rate when bidding for a loan. Duarte *et al.* (2012) uses data from Prosper and finds that borrowes who appear more trustworthy default less often. Same data uses in his study Everett (2010) and next to information asymmetry focus, studies also default rates. He finds that not only creditscore, but also borrower age, home ownership and amount size are determinants of loan default. Herzenstein *et al.* (2011) again on Prosper data investigates herding behaviour and finds out, that such strategic behavior benefits both the borrower and lender as loans are funded and have less defaults.

Emekter *et al.* (2015) evaluates credit risk and loan performance on Lending Club platform. After estimating the default risk he compares it to the general US level and concludes, that risky loans are not worth it. Emekter *et al.* (2015) is followed by Malekipirbazari & Aksakalli (2015) who exploits random forest method for prediction of borrower status. Using data from Lending Club from year 2012 and 2014, their method outperforms FICO credit scores as well as credit grades given by Lending Club platform, but only by a small proportion. Serrano-Cinca *et al.* (2015) also explains default rate on Lending Club by logistic regression and find that also loan purpose and annual income are important determinants of default along with the purpose of the loan. Guo *et al.* (2016) uses kernel regression to improve estimation of default rates and touches the portfolio selection problem. This thesis follows mainly Emekter *et al.* (2015); Guo *et al.* (2016) who provide comparable results.

This thesis looks at P2P loan markets as investment possibility. In order to make optimal investment decision, we use portfolio management perspective. This approach allows for formalized investment guide that can be further used by investors and embedded into automated mechanisms on P2P platforms that invest in loans given specific criteria. Academic literature is really limited in this so far. The first attempt that focuses on risk and return of investment into loans by Singh *et al.* (2008) who used early Prosper data. They calculated ROI for each loan and then applied decision tree analysis to separate loans into groups. Based on risk and return of each group, they calculated optimal portfolio, but did not account for covariance between portfolio assets. They find that most of the groups have negative expected return, but it might be thanks to high discount rate they use for cash flows (size of discount rate is not given).

Second study by Guo *et al.* (2016) uses datasets from Prosper and Lending Club, yet very limited compared to the available data (2016 loans from Lending Club and 4128 loans from Prosper which is less than Singh *et al.* (2008)) and states again, that correlation between loans is negligible without any attempt to calculate it. Decision process in this study does not uses credit scoring by the platform, and makes decisions between available loans per se. Modern approach to default is taken by Byanjankar *et al.* (2015). Authors use neural networks and data from Bondora platform. They conclude, that financial attributes are more influencing than demographic attributes in determining credit risk, but dataset they use is quite limited and is not clear, if default rates can further change or not.

This thesis argues, that loan performances across credit grades are correlated and correlation cannot be simply ignored as by Guo *et al.* (2016); Malekipirbazari & Aksakalli (2015). Reasons for correlation can be argued not only by adverse selection (Freedman & Jin 2008; Iyer *et al.* 2010), but also as by dependency of consumer ability to repay on current economic situation (Iyer *et al.* 2010).

Chapter 4

Theoretical concepts

4.1 Portfolio theory

Portfolio theory is microeconomic area investigating ideal composition of assets in a portfolio to achieve optimal risk and return properties. One of the first economists to deal with portfolio was Hicks (1934) who said that investors use statistics and probability distributions to improve their returns. As a founding father of modern portfolio theory is known Markowitz (1952), who deals with the relation between risk and return. He constructs effective frontier, which represents maximal return for given level of risk. Markowitz emphasizes, that risk and return are two contradictory targets. Investors want to achieve the highest possible returns, but they are volatile. Investors can also aim at specific return with the lowest risk. Markowitz in his analysis assumes, that investors have certain amount of capital that they want to invest for specific time and then reinvest it or use it for their own consumption. Hence investing is a periodical activity when investor chooses between different forms of investment with various returns and risk level (uncertainty of the return).

Markowitz uses risky assets and effects of diversification to construct the efficient portfolio. The aim of diversification is to lower the risk by investing in multiple assets. Necessary condition for the diversification to work is at least partial independence of asset returns. As the returns are not perfectly dependent on each other, the portfolio created from such assets purchased separately has a better ratio between risk and return than each of the assets individually.

Markowitz's portfolio of risky assets was enhanced by a risk-free asset by Sharpe (1964) and new Capital Asset Pricing Model (CAPM) model was created. This model is facing criticism from the day it has been published (e.g. Roll (1977)). The subject of criticism are (a) the existence of perfect capital markets, (b) determination of risk-free rate, (c) not accounting for taxes and transaction costs, (d) different expectations about risk and return by the investors, (e) that all investors are risk averse etc.

According to the portfolio theory investors choose their optimal portfolio in following three steps. This thesis will follow these steps as well.

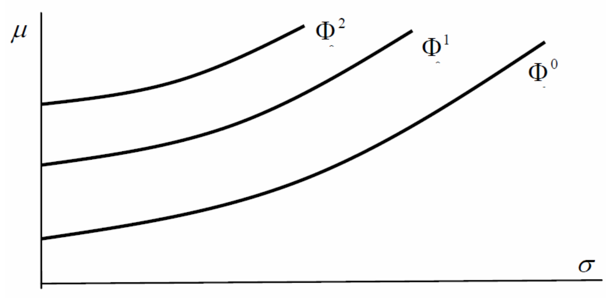
1. Asset analysis – estimate risk and return of assets and calculation of variance-covariance (VCV) matrix.
2. Portfolio analysis – decision about the weights of different assets in the portfolio No preference or expectation is taken into account at this point, it is merely technical operation.
3. Choosing the optimal portfolio – based on the investors preferences (level of risk, return, area, etc.) optimal portfolio is chosen.

4.1.1 Investor utility

Every investor has different attitude towards risk. In portfolio theory they differ by how much of additional risk they are willing to take in exchange for additional return. These preferences can be depicted using indifference curves that capture same utility level for various combinations of risk and return. Indifference curves cannot intersect because all portfolios on the curve are equally good for the investor. Sample indifference curves for an investor are depicted in Figure 4.1. Where μ represents return, σ states for risk and $\phi(\mu, \sigma)$ represents expected utility from expected wealth.

It is assumed investors will always prefer higher return for the same level of risk and at the same time investor will prefer portfolio with lower risk for the same level of return. Investors will therefore try to get to the highest indifference curve. For the rational investors indifference curves are in the context of Figure 4.1 upward-sloping from these reasons. Given the risk-averse or risk-seeking feature of an investor, the steepness of the curves is determined.

Figure 4.1: Indifference map in the risk-return space



Source: Dědek (2016).

The more risk-averse an investor is the steeper are the curves (investor needs higher increase in expected return to compensate for increased risk).

4.1.2 Expected return and risk of portfolio

Expected return of a portfolio is a weighted average of expected returns of all assets in the portfolio. For n assets we can compute expected return of the portfolio using Equation 4.1:

$$r_p = \sum_{i=1}^{n} w_i \cdot r_i, \quad \sum_{i=1}^{n} w_i = 1 \tag{4.1}$$

where, r_p is the expected return of the portfolio, w_i and r_i is the weight and expected return of asset i, respectively. Expected return of the portfolio can never be higher than expected return of the asset with the highest expected return. If investors want to maximize the expected return regardless of the risk, they should invest only in the asset with the highest expected return (if we do not consider short positions).

Risk of the portfolio is measured as the standard deviation or variance of returns. For n assets in the portfolio, we can compute the risk using Equation 4.2

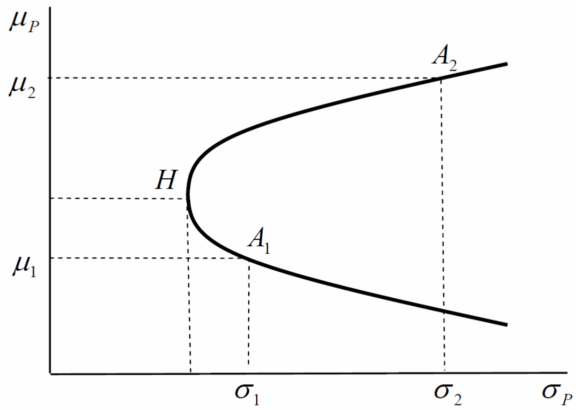
$$\begin{aligned} \sigma_p &= \sqrt{\sum_{i=1}^{n}\sum_{j=1}^{n} w_i w_j \sigma_i \sigma_j \rho_{ij}}, \quad \rho_{ij} = \frac{\sigma_{ij}}{\sigma_i \sigma_j} \\ &= \sqrt{\sum_{i=1}^{n}\sum_{j=1}^{n} w_i w_j \sigma_{ij}}, \quad \sum_{i=1}^{n} w_i = 1 \end{aligned} \tag{4.2}$$

where σ_p is the expected standard deviation of portfolio returns, w_i and w_j are weights of assets i and j in the portfolio, σ_i and σ_j are standard deviations of assets i and j, ρ_{ij} means correlation coefficient between asset i and j, and σ_{ij} is the covariance between returns of assets i and j. Since for the most of the assets $\rho_{ij} < 1$, proper selection of multiple assets into the portfolio reduces the whole risk. That is the mathematical reason behind the diversification.

4.1.3 Efficient portfolio

Having two or more available assets, investors can create infinite amount of portfolios (by using different weights of the assets in the portfolio). As we assume rationality of the investor, we do not need to calculate the whole set of available portfolios, but only the effective frontier of the set. By effective frontier we understand combinations of assets that offer maximal expected return for given level of risk.

Figure 4.2: Investment opportunity set formed by two uncorrelated assets

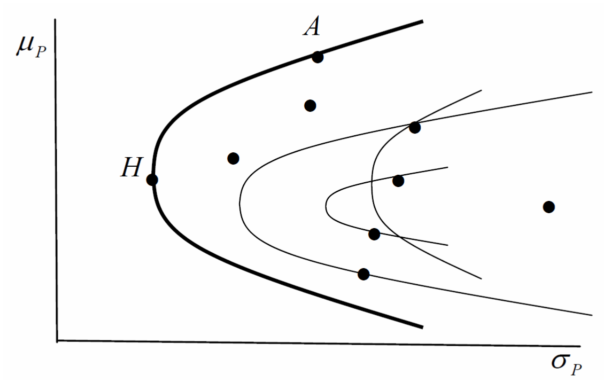


Source: Dědek (2016).

The opportunity set depicted on Figure 4.2 has the efficient part (the upper part of parabola) and inefficient part (the lower part of the parabola. The opportunity set is created using two assets A_1 and A_2 which are uncorrelated. Portfolio H is the minimum variance portfolio – the least risky portfolio.

We can create opportunity set for many assets – depicted on Figure 4.3. All portfolios inside the set are inefficient because it is possible to either get the same return with lower risk or for the same risk get higher return. Therefore

Figure 4.3: Investment opportunity set formed by many risky assets



Source: Dědek (2016).

the individual points of the investment opportunity set can be determined by solving problems formalized by Equation 4.3:

$$\begin{aligned} \sigma_p &\rightarrow \text{min subject to a given } \mu_p \\ \mu_p &\rightarrow \text{min subject to a given } \sigma_p \end{aligned} \tag{4.3}$$

4.1.4 Risk-free asset

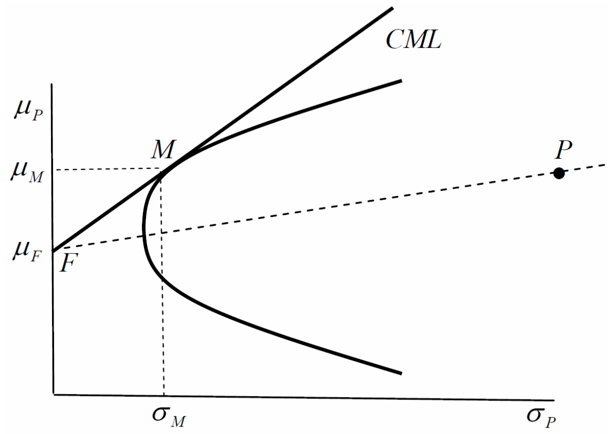
By introduction of risk-free asset F the investment opportunity set changes to a straight line going from portfolio F and touching efficient frontier of all risky assets – situation is depicted on Figure 4.4. This tangent line is called the Capital market line (CML). All points on CML provide higher return for the same risk compared to the efficient frontier except the tangent portfolio M. The tangent portfolio generates the steepest possible line when combined with the riskfree investment and thereofe provides the highest reward to volatility and maximizes Sharpe ratio formalized by Equation 4.4

$$s_r = \frac{\mu_P - \mu_F}{\sigma_P} \tag{4.4}$$

To compute the tangent portfolio, we need to solve the optimization problem (formalized by Equation 4.5) for which Sharpe ratio is maximized:

$$\max\, s_r = \frac{\mu_P - \mu_F}{\sigma_P} = \frac{\sum_{i=1}^n w_i \mu_i - \mu_F}{\sqrt{\sum_{i=1}^n \sum_{j=1}^n w_i w_j \sigma_{ij}}}, \quad \text{s.t.} \sum_{i=1}^n w_i = 1 \tag{4.5}$$

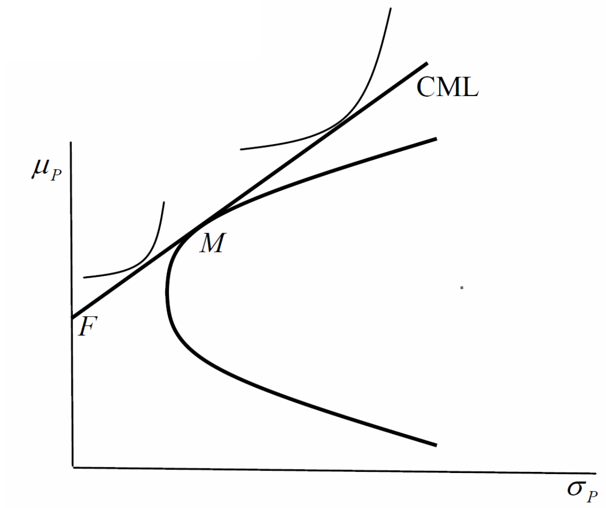
Figure 4.4: Risk-free asset, CML and efficient frontier



Source: Dědek (2016).

By adding investors (Figure 4.5) with their indifference curvers described in Subsection 4.1.1 we can find the their optimal choices, which will be the point where CML touches their indifference curves. Selected portfolio of each investor

Figure 4.5: Optimal choice of risk-averse investors



Source: Dědek (2016).

consists only of risk-free asset F and market portfolio M, because investors get to their highest indifference curve by portfolios that belong to CML. As market portfolio M is identical for all investors, they all have the same combination of risky assets regardless of their risk attitude. The differ only in the amount of risk-free asset held.

CAPM model uses following assumptions, summarized by Sharpe (2000):

- There are no transaction costs or taxes
- Investors are rational and risk averse
- All investors have the same expectations about asset characteristic (volatility, expected returns, correlations)
- Inventors can borrow or lend at the risk-free interest rate
- Markets are efficient (all information is freely available)
- Number of assets is fixed and assets are perfectly divisible
- Investors are price takers (cannot influence prices)
- Assets are perfectly divisible

CAPM model therefore allows investors to borrow at the risk free rate and invest more than their initial capital into the market portfolio This creates leverage, that increases expected return. We can impose also restriction that $w_i \geq 0$ for each asset i, including the risk free asset. This condition does not allow to enter short position and is more realistic for a smaller investor that we aim to investigate in this thesis.

Compared to stock market, P2P platforms allow for cheaper diversification, because there are fees calculated from total amount invested and so they are not restricted to each investment. There are no transaction fees when investing into the loans, the only transaction fees are connected with reselling the position. But rebalancing of the position can be also achieved by reinvesting collected interest rate payments.

4.2 Credit risk (default estimation)

Very large portion of research using P2P data is oriented on estimation of defaults of single loans as described in Chapter 3 (Emekter *et al.* 2015; Everett 2010; Malekipirbazari & Aksakalli 2015; Guo *et al.* 2016; Iyer *et al.* 2009; Herzenstein *et al.* 2008; Serrano-Cinca *et al.* 2015). Using historical data and logistic regression framework for ex-ante estimation of default of a loan is created. By investing only into the loans what are expected not to default investors can achieve higher results. Problem might arise if all investors would use the same strategy, because properties of funded loans would be the same and later default would depend on something completely different and unpredictable. Past performance, however, cannot guarantee good performance in the future. We will use default estimation and selection of a new portfolio as robustness check of our results.

To determine the precise effect of each loan variable on the default probability, we follow logic described by Bastos (2010); Emekter *et al.* (2015); Malekipirbazari & Aksakalli (2015); Serrano-Cinca *et al.* (2015). Our ultimate target is to estimate the probability of default, which should be between 0 and 1, for which we need binary response model (Equation 4.6):

$$P(y=1 \mid \mathbf{x}) = P(y=1 \mid x_1, x_2, \ldots, x_n) \tag{4.6}$$

where $\mathbf{x}$ denotes the full set of explanatory variables. The most important disadvantage of linear probability model is that the fitted probabilities can be lower than 0 or higher than 1. We need transformation function T formalized by Equation 4.7, that will ensure the output values are strictly between 0 and 1.

$$P(y=1|\mathbf{x}) = T(\beta_0 + \beta_1 x_1 + \ldots + \beta_k x_k) = T(\boldsymbol{x\beta}) \tag{4.7}$$

In the logit model transformation function is the logistic function, with real number input d, expressed by Equation 4.8:

$$p_i = T(d) = \frac{e^d}{1+e^d} = \frac{1}{1+e^{-d}}, \quad d = \boldsymbol{x\beta} \tag{4.8}$$

Probability of default is therefore determined in two steps. In the first step real number representing likelihood of default (d_i) is estimated for each loan by linear regression 4.9, where we assume, that d_i in linearly related to independent variables in the binary logistic regression model:

$$d_i = \beta_0 + \beta_1 x_1 + \ldots + \beta_k x_k + \epsilon_i \tag{4.9}$$

In the second step the likelihood of default d_i is transformed into probability of default p_i between zero and one using Equation 4.8. Due to the non-linear nature of logit models, we have to estimate them using the Maximum Likelihood Estimation (MLE). As the transformation function $T(\mathbf{x}\beta)$ is non-linear, the partial effects of the explanatory variables on the default rates are not constant. The explicit partial effect of variable x_i on the default rate is given by Equation 4.10:

$$\frac{\partial T(\boldsymbol{x\beta})}{\partial x_i} = \frac{dT(\boldsymbol{x\beta})}{d(\boldsymbol{x\beta})}\beta_i \tag{4.10}$$

Because the transformation function $T(\boldsymbol{x\beta})$ is strictly monotonic the sign of the coefficients gives us the direction of the partial effects.

Emekter *et al.* (2015) included in total 13 variables[1] to estimate the binary regression model, but only four variables turned out to be significant. Namely credit grade, FICO score, debt-to-income ratio and revolving credit line utilization. Guo *et al.* (2016) use FICO score, number of inquiries of the borrower in the last 6 months, monetary amount of the loan, home-ownership status and debt-to-income ratio. Serrano-Cinca *et al.* (2015) finds that factors explaining the default rate are loan purpose, annual income, current housing situation, credit history and indebtedness. Byanjankar *et al.* (2015) finds on different platform as significant determinants loan amount, income and if underwriters restructured the initial application (they offer change of term, loan amount etc.).

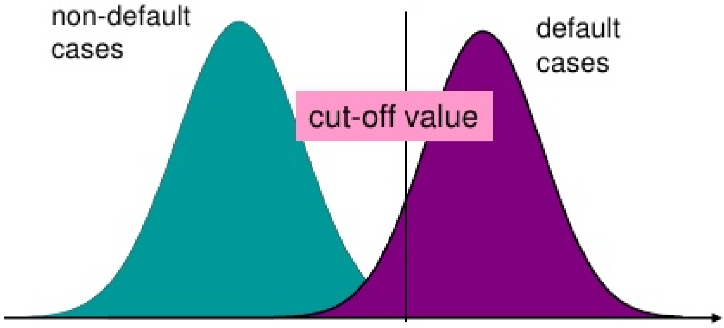
4.2.1 Cut-off rate

After the defaults are predicted on the testing dataset, one should optimize cut-off ratio to obtain optimal results. One way to set up the cut-off rate is to get the same number of predicted defaults as there are real defaults (default ration will be the same). Other possibility is to look as gains and losses from cut-off rate and optimize total gains or minimize total losses. Our model predicts the probability of default to each loan. In the historical dataset we know which loans really defaulted and which survived. By plotting the distribution of default rates we should get a plot similar to Figure 4.6.

By changing the cut-off rate, we can change the proportion between successfully predicted defaults and survivals. By lowering it, we can increase the prediction accuracy of defaults, on the other hand, we reject larger proportion of good

[1] All variables were: credit grade, debts to income ratio, monthly income, FICO score, open credit lines, total credit lines, revolving credit balance, revolving credit line utilization, inquires six months, delinquent amount, delinquencies two years, months since last delinquency and month since last received.

Figure 4.6: Cut off rate



Source: www.slideshare.net.

loans. The idea can be explained using Figure 4.6 where x-axis represents the probability od default and y-axis the number of loans. The green plot captures the distribution of probability of default for non-defaulted loans and the purple for defaulted ones. The cut-off rate is a minimum creditworthiness threshold (value of probability of default) selected by investor. All loans with probability higher than cut-off rate are considered as predicted default and all loans with probability smaller then cut-off rate are considered as not to default (creditworthy).

To calibrate the cut-off rate (how many defaulted and non-defaulted loans are predicted correctly) we need to balance between losses and gains as captured by Table 4.1. For investors is important to maximize the sum of profit from loans and loss from defaults. If defaults are really costly, it is important for the investor, that the model is able to detect as much defaults as possible, so that the investor can avoid granting such loans. For credit institutions, it is also important to minimize losses from opportunity costs (rejected loans that would perform). For our investor these losses are negligible, because he does not have any costs with them.

Table 4.1: Cut-off rate calibration conditions

	Predicted		
Actual	Non Default	Default	Correct
Non Default	Profit from loans	Opportunity costs	Ratio of correctly estimated good loans
Default	Loss from default	-	Ratio of correctly estimated defaults

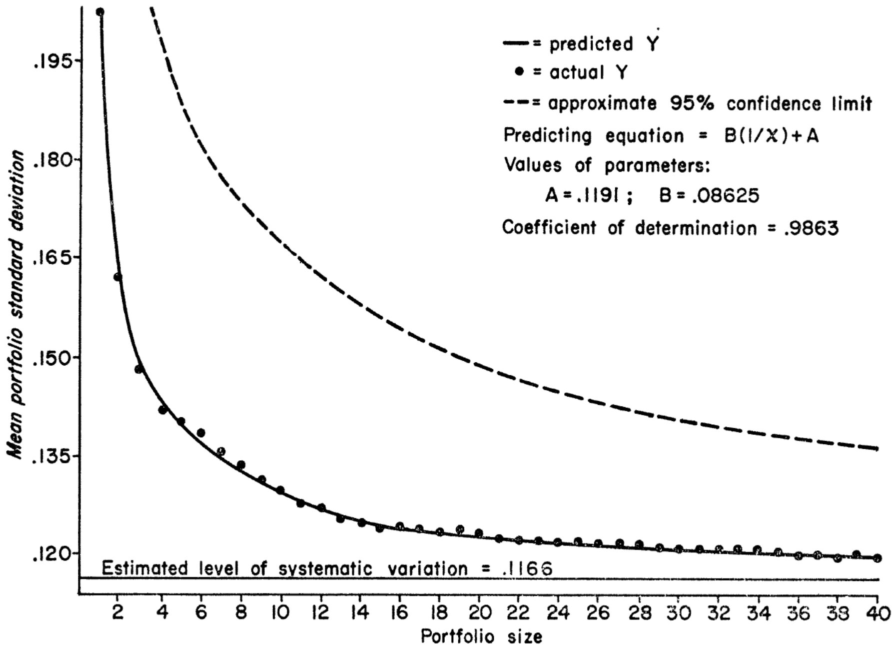
4.3 Diversification

Risk, which is one of the key determinants of a portfolio depends on the weights of individual assets in the portfolio, their variances, and their covariances. If any of these characteristics change, overall risk of the portfolio will be affected. For an equally weighted portfolio (each stock or security has the same weight) it is true that the increase of the number of stocks will reduce portfolio risk (Statman 1987). That is the benefit of diversification.

Markowitz (1952) states that diversification is the only free lunch in finance. Markowitz is further followed by Sharpe (1964) who suggests, that overall variation of the returns of portfolios can be split into two forms. Firstly, the systematic part, which results of covariance of returns between different securities and the market return. Secondly, the non-systematic part, which account for security idiosyncratic attributes. Evans & Archer (1968) discusses necessity of marginal analysis of portfolio diversification. Portfolio analysis per se does not discriminate between portfolios that differ only in number of securities. However, adding an extra security to a portfolio always has some costs. Then the question is, what is the marginal added value of additional security to the portfolio and what are the costs associated with such action.

Evans & Archer (1968) claim that reduction of portfolio return variation that is a result of diversification must be related ot the number of securities in the portfolio. After empirical analysis of stock data from year 1958 to 1967, they conclude, that the largest proportion of unsystematic variation is eliminated by the time portfolio consists of 8 securities, as depicted on Figure 4.7. They raise doubts, if it is profitable to increase number of securities in portfolio whose size is more than 10 securities. The answer lies in the marginal analysis of each

Figure 4.7: Elimination of unsystematic variation by number of securities



Source: Evans & Archer (1968, Figure 1).

specific portfolio.

Statman (1987) raises doubts that 10 securities is enough to seize the benefits of diversification and argues, that well-diversified portfolio of randomly chosen stocks must consists of at least 30 stocks for a borrowing investor and 40 stocks for a lending investor. He also agrees with Evans & Archer (1968), that diversification should be increased as long as the marginal benefits are larger than marginal costs. Moreover, Statman (1987) finds that individual investors do not hold well diversified portfolios and concludes, that investors should not be rushed into holding fully diversified portfolios, because their preferences are unknown. Later on Statman (2004) continues his research and solves this diversification puzzle by using behavioural portfolio theory (Statman 2004).

Forty years later is Evans & Archer (1968) followed by Benjelloun (2010), who firstly argues that question about size of a diversified portfolio is still very discussed and important in academic finance. Benjelloun (2010) uses two weighting schemes and measures of risk and concludes, that 40-50 stocks are needed to achieve sufficient diversification in the US stock market. Aim of that study is to contradict previous research like Domian *et al.* (2007) claiming needs for hundreds of stocks in a portfolio to be diversified.

This thesis follows the above summarized previous research. P2P market places emphasize that investors should invest into multiple loans to diversify their portfolios. However, no empirical evidence has been provided by academics that what is valid for stock markets is also valid for P2P market places. We do so in Section 5.4

Chapter 5

Empirical part

The empirical part is structured as follows. Firstly we describe the dataset thoroughly, to provide an idea about its extensiveness and basic characteristics of the dataset. Then we estimate credit risk for each loan by default prediction. Then we will use this information in the portfolio creation and find effective portfolio – what is the ideal composition of credit grades. Lastly we will show how many of the loans from each credit grade investors should invest into to fully enjoy the only free lunch in portfolio management – the effects of diversification.

5.1 Description of data

In this section, the sample of the loan data is firstly described, followed by the distribution tables by credit grades and loan status. This thesis uses in total 886837 loan applications funded in the Lending Club from January 2008 to December 2015, which we downloaded from www.lendingclub.com. Lending club is currently the largest P2P lending platform in the USA and traded on the New York Stock Exchange since December 2014. Data between May 2007 to January 2008 are not used in the analysis, as the beginning of the platform might add additional noise in the data. We believe, that within the first months of the functioning the platform solved possible problems and also the number of funded loans increased and allows for more robust results. Data from 2016 are available, but we prefer to run our analysis on loans for which we have at least

12 months history. Over this whole period of the Lending Club's existence it lent about 13.1 billion to borrowers. Borrowers can borrow for 36 or 60 months.

Table 5.1: Loan distributions by loan purpose

Purpose	Number of loans	Per cent	Amount	Per cent
Debt consolidation	524 046	59.09	8 086 510 300	61.78
Credit card repay.	206 107	23.24	3 162 006 125	24.16
Home improvement	51 798	5.84	742 059 825	5.67
Other	42 783	4.82	423 660 725	3.24
Major purchase	17 267	1.95	200 323 075	1.53
Small business	10 322	1.16	159 137 250	1.22
Car	8 851	1	78 644 200	0.6
Medical	8 532	0.96	76 955 775	0.59
Moving	5 400	0.61	42 660 775	0.33
Vacation	4 732	0.53	29 770 150	0.23
House	3 699	0.42	54 802 825	0.42
Wedding	2 338	0.26	24 519 650	0.19
Renewable energy	575	0.06	5 720 275	0.04
Educational	387	0.04	2 577 575	0.02
Grand Total	886 837	100	13 089 348 525	100

Table 5.1 provides overview of the borrower's claimed reasons for the loan. Majority of the requests (59%) is related to debt consolidation followed by credit card repayments (23%), which means that 82% of loan requests are debt repayments. On the other side of the table are loans used for educational purposes or investments into renewable energy in total only 0.1% of all loan requests. Borrowers report that Lending Club offers them better borrowing conditions (repayment options, interest rate) then their current debt providers (Emekter *et al.* 2015). We can compare these numbers to the largest P2P platform in the Czech Republic, Zonky. On Zonky 42% of the loan requests are for debt refinance, followed by car loans which make 19% of the loan requests[1]

Table 5.2 presents Lending Club distribution of credit grades and year. The Lending Club uses FICO score as the main determinant of the credit grade, but makes adjustment based on other information (such as credit history length, revolving line utilization, requested amount, total open credit accounts) pro-

[1] Author's own calculation based on automatically collected data from Zonky market place.

vided by the borrower. Credit grades are further divided into subgroups that determine the final interest rate.[2]

Table 5.2: Loan distributions by year and credit grade

Year	Credit grade	Number of loans	Per cent	Amount	Per cent
2008	A	318	0.04	1 982 575	0.02
	B	594	0.07	5 449 775	0.04
	C	580	0.07	5 127 250	0.04
	D	419	0.05	3 759 325	0.03
	E	285	0.03	2 561 825	0.02
	F	111	0.01	1 236 275	0.01
	G	86	0.01	1 002 225	0.01
2009	A	1 203	0.14	8 700 675	0.07
	B	1 445	0.16	15 695 475	0.12
	C	1 348	0.15	13 182 425	0.1
	D	817	0.09	8 700 125	0.07
	E	308	0.03	3 665 175	0.03
	F	105	0.01	1 328 225	0.01
	G	55	0.01	656 150	0.01
2010	A	2 830	0.32	24 824 350	0.19
	B	3 687	0.42	40 253 325	0.31
	C	2 729	0.31	27 447 875	0.21
	D	1 885	0.21	20 708 000	0.16
	E	962	0.11	11 791 925	0.09
	F	311	0.04	4 570 050	0.03
	G	133	0.01	2 397 025	0.02
2011	A	5 754	0.65	51 657 300	0.39
	B	6 565	0.74	74 656 100	0.57
	C	3 942	0.44	48 099 400	0.37
	D	2 796	0.32	37 321 675	0.29
	E	1 739	0.2	31 071 225	0.24
	F	722	0.08	14 562 175	0.11
	G	203	0.02	4 315 950	0.03
2012	A	10 901	1.23	121 476 925	0.93
	B	18 507	2.09	219 668 150	1.68
	C	11 875	1.34	153 419 000	1.17
	D	7 323	0.83	113 862 325	0.87
	E	3 185	0.36	70 276 075	0.54
	F	1 315	0.15	32 853 525	0.25

Continued on next page

[2] Exact weights of each parameter and procedure of credit grade determination is not a publicly available know how of Lending Club.

Table 5.2: Loan distributions by year and credit grade (continued)

Year	Credit grade	Number of loans	Per cent	Amount	Per cent
	G	261	0.03	6 855 025	0.05
2013	A	17 679	1.99	270 784 350	2.07
	B	44 116	4.97	595 898 525	4.55
	C	38 134	4.3	567 091 400	4.33
	D	20 569	2.32	283 641 225	2.17
	E	9 059	1.02	161 329 875	1.23
	F	4 393	0.5	84 437 275	0.64
	G	864	0.1	19 582 625	0.15
2014	A	36 108	4.07	523 412 775	4
	B	61 935	6.98	844 292 375	6.45
	C	66 565	7.5	963 542 750	7.36
	D	42 992	4.84	689 507 475	5.27
	E	20 121	2.27	345 506 050	2.64
	F	6 223	0.7	104 298 375	0.8
	G	1 685	0.19	33 280 375	0.25
2015	A	73 336	8.26	1 077 445 350	8.23
	B	117 606	13.25	1 676 097 950	12.8
	C	120 567	13.59	1 777 831 825	13.58
	D	62 654	7.06	999 154 850	7.63
	E	34 948	3.94	645 584 850	4.93
	F	9 817	1.11	197 226 225	1.51
	G	2 167	0.24	44 267 125	0.34
Grand Total		887 440	100	13 094 326 000	100

The Lending Club uses borrower's FICO scores along with other information to classify a loan from A being the lowest risk and G being the highest risk. Total amount number of loans as a percentage of total loans is 16.1%, 28.7%, 27.7%, 15.7%, 8%, 2.6% and 0.6% for credit grade group A, B, C, D, E, F, G respectively.

Majority (almost 90%) of the loan requests belong to credit grades between A and D, grades E, F and G account for 11% of total loan requests. Largest number of request belong to grades B and C, both around 28% representing over 250 000 requests and each worth \$3.5 billion. Worst credit group G with only 5454 loan requests with total amount of \$112 million accounts for not even one per cent of all loans.

Lastly, Table 5.3 shows the loan status for all loan requests as per December 2016. Actual default rate is 9.63% with written off losses about \$1.3 billion,

Table 5.3: Loan distributions by the loan status

Loan status	Loan count	Per cent	Amount (USD)	Per cent
Current	434 568	49	6 723 905 775	51.37
Fully Paid	342 794	38.65	4 699 231 750	35.9
Charged Off (Default)	85 402	9.63	1 280 256 650	9.78
Late (31-120 days)	14 743	1.66	234 991 800	1.8
In Grace Period	6 379	0.72	103 965 475	0.79
Late (16-30 days)	2 951	0.33	46 997 075	0.36
Grand Total	886 837	100	13 089 348 525	100

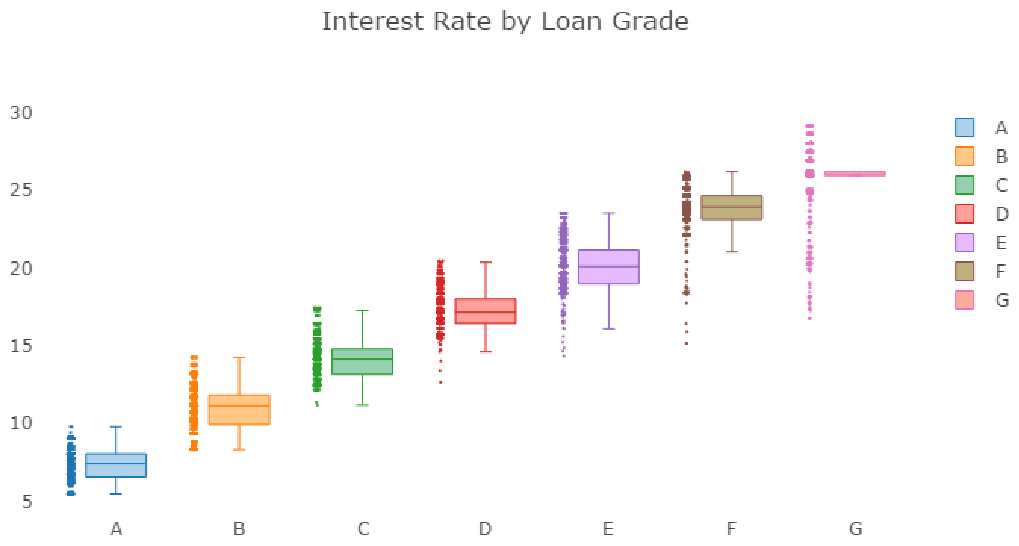
which is double of what reported Emekter *et al.* (2015). Another 2% of loans are late and not following the expected payment schedule. Positive side is, that 38.65% of loans with principal value of $4.7 billion is already fully repaid and 49% of loans with original principal worth $6.7 billion are in current status (following expected repayment schedule). Lowest credit grades naturally reported the highest default rates (see Table 5.5 for more details). For investors is important to balance default rate and interest rates to achieve optimal returns. If default rates are too high, even high interest rates do not provide sufficient revenues to maintain positive returns.

Table 5.4: Descriptive statistics of explanatory variables

Variable	Mean	SD	Min	Max
Interest rate	13.25	4.38	5.32	28.99
DTI	18.14	8.42	0	1092.52
Loan amount	14759.59	8435.02	500	35000
Installment	436.82	244.17	15.67	1445.46
Term	43.21	11	36	60
Home ownership	0.6	0.49	0	1
Revolving line utilization	55.04	23.86	0	892.3
Annual income	75033.39	64696.32	1200	9500000

Table 5.4 reports the general characteristics of the dataset from Lending Club we use in empirical analysis. Based on our sample of almost 900 thousand loans, the average interest rate is 13.25% and average size of a loan is $14760. Borrowers earn on average $6253 monthly and their debt-to-income ratio is 18%. Average monthly payment on a loan is $437 and average remaining principal is $9166 on non-repaid loans.

Figure 5.1: Interest rate disribution by grades



5.2 Estimation of default rates

From a lender's perspective a very important concern is if a borrower will default and if one can predict that from available information. Lender can benefit if some loan or borrower characteristics can help to determine that the loan request is more likely to default than others (Emekter *et al.* 2015). We will try to find these characteristic using learning sample of 84550 loans (matured loans issued before 2013) from Lending club platform described in Section 5.1. From 174486 matured loans, 23114 loans are not paid back, which can be translated as default rate of 13%, but as one would expect and Table 5.5 presents in more detail, default rate varies across credit grades. Actual default rates are 3.66%, 7.74%, 12.41%, 17.69%, 22.38%, 28.83% and 33.54% for loans with credit grade A, B, C, D, E, F and G, accordingly.

Firstly, we will compare variables that will be used in our default estimation model using non-parametric test for good and defaulted loans. Then we model the default risk by employing binary logit regression (see Section 4.2 more model derivation). As defaulted loans for our analysis we consider charged-off loans and loans which are late more than 30 days. Table 5.6 summarizes the differences between defaulted and good loans (table reports means of those for

Table 5.5: Loan defaults by credit grades

Credit grade	Defaulted loans	Total loans	Default ratio
A (lowest risk)	5418	148129	0.0366
B	19698	254455	0.0774
C	30505	245740	0.1241
D	24663	139455	0.1769
E	15805	70607	0.2238
F	6631	22997	0.2883
G (highest risk)	1829	5454	0.3354
Total	104549	886837	0.1179

Notes: Each credit grade is determined by Lending Club using FICO score and other information. Data for calculation are from January 2008 to December 2015.

a better idea). To determine if these groups are different in these variables, we use non-parametric Kruskal–Wallis test. This test can either test for different shape distributions, but in our case we assume similar shapes so we test for the difference in means. The chi-square statistic values of Kruskal Wallis test show that all listed variables between the two groups are significantly different at the 1% level. We find that interest rate on a defaulted loan is higher, amount is lower, borrowers have lower FICO score, lower credit grade, higher DTI and credit line utilization compared to non-defaulted loans which borrowers are more likely to own a home.

Table 5.6: Comparison of variables differences between defaulted and good loans

Variable	Defaulted loans	Good loans
Interest rate	15.77	12.91
FICO score	689	698
Credit grade	3.4	2.7
Home ownership	0.5	0.6
Monthly income	5599	6333
DTI	19.5	17.9
Revolving line utilization	58.17	54.67
Loan amount	14994	14691

Notes: Credit grade is Lending club grade (A-F) translated into numerical values with 1 being the low risk category 'A' and 6 being the high risk category 'G'.

To determine the precise effect of each variable on the likelihood of a loan default, we use binary logistic regression presented in Section 4.2. Let us recall Equation 4.9:

$$d_i = \beta_0 + \beta_1 x_1 + \ldots + \beta_k x_k + \epsilon$$

which for the given variables has following form:

$$\begin{aligned} d_i = \beta_0 &+ \beta_1(creditgrade) + \beta_2(FICOscore) + \beta_3(Homeownership) + \\ &+ \beta_4(creditratio) + \beta_5(Revolvingline) + \beta_6(Monthlyincome) + \\ &+ \beta_7(Purpose) + \beta_8(Loanamnout) + \beta_9(Term) + \beta_{10}(DTI) + \epsilon_i \end{aligned} \tag{5.1}$$

We aim to use the prediction for portfolio selection. From all available loans, we use only part of the dataset and second part will be used for measurement of model performance. Following common practice, 70% of observations is used for training and 30% for model testing and dataset is divided randomly into two parts (Byanjankar *et al.* 2015). From investor perspective, it is logical to use the historical data for creating the model that is used in the decision process for new investments. Similar to financial markets, past performance does not necessary predict future returns yet it is the best estimate of it. For these reasons, we selected years 2008–2012 as learning (or training) and 2013 as testing year. Available dataset consisting of years 2008-2015 consists of matured loans granted between 2008 and 2013. Table 5.7 shows the distribution of loan origination and maturity in the dataset.

Table 5.7: Loan counts by year

Year	All loans	Per cent	Matured loans	Per cent	Not matured (%)
2008	2393	0.27	2393	1.38	0
2009	5281	0.6	5281	3.04	0
2010	12537	1.41	12537	7.21	0
2011	21721	2.45	20869	12	3.92
2012	53367	6.02	43470	25	18.55
2013	134814	15.2	89333	51.38	66.26
2014	235629	26.57			100
2015	421095	47.48			100
Total	886837	100	173883	100	

5.2.1 Logistic regression results

Results of model 5.1 for whole dataset are provided in Table 5.8. The binary logistic regression is estimated using iterative maximum likelihood method. In total we used 6 single and 3 category variables in the binary regression model to estimate the probability of default, and from that only 3 purpose types are not significantly (estimated coefficients are significant at 1% level) affecting the odds of loan default. When we use loans from early years 2008–2012, also home ownership is not significant. We test for the goodness of fit using Hosmer and Lemeshow test, and we report also final R^2 for each model.

Table 5.8: Binary logistic regression results

Variable	Learning sample	2008–2012	All matured	All loans
Credit grade				
B grade	0.467***	0.465***	0.468***	0.671***
	(11.67)	(12.16)	(15.61)	(40.83)
C grade	0.768***	0.757***	0.825***	1.140***
	(16.71)	(17.33)	(25.32)	(68.51)
D grade	0.952***	0.949***	1.070***	1.514***
	(18.12)	(19.06)	(30.34)	(85.96)
E grade	1.006***	1.076***	1.154***	1.797***
	(15.26)	(18.12)	(25.55)	(93.13)
F grade	1.186***	1.231***	1.300***	2.133***
	(13.77)	(17.15)	(19.67)	(92.94)
G grade	1.268***	1.380***	1.390***	2.341***
	(10.74)	(14.02)	(12.87)	(68.49)
FICO score				
750-779	0.295***	0.307***	0.319***	0.204***
	(3.51)	(3.93)	(4.38)	(4.90)
715-749	0.502***	0.517***	0.498***	0.291***
	(6.32)	(7.03)	(7.37)	(7.76)
680-714	0.544***	0.554***	0.586***	0.383***
	(6.53)	(7.19)	(8.49)	(10.29)
660-679	0.659***	0.677***	0.670***	0.430***
	(7.43)	(8.25)	(9.44)	(11.45)
< 660	0.974***	0.935***	1.001***	0.655***
	(6.17)	(6.10)	(6.86)	(5.01)
Purpose				
Car	-0.145**	-0.193***	-0.0837	-0.0377
	(-2.01)	(-2.80)	(-1.35)	(-1.01)
Credit card	-0.253***	-0.243***	-0.175***	-0.108***
	(-8.30)	(-8.71)	(-8.94)	(-12.00)

Continued on next page

Table 5.8: Binary logistic regression results (continued)

Variable	Learning sample	2008–2012	All matured	All loans
Education	0.431***	0.426***	0.512***	0.750***
	(3.30)	(3.26)	(3.93)	(5.75)
Home improvement	0.126***	0.114***	0.147***	0.0313**
	(3.14)	(3.06)	(4.91)	(2.21)
Major purchase	0.138***	0.135***	0.142***	0.00575
	(4.06)	(4.27)	(5.39)	(0.42)
Medical	0.325***	0.301***	0.298***	0.0951***
	(4.16)	(3.95)	(4.71)	(2.91)
Small business	0.766***	0.775***	0.749***	0.472***
	(16.15)	(17.61)	(18.24)	(18.52)
Vacation, wedding	-0.0427	-0.0442	-0.0324	0.0256
	(-0.65)	(-0.70)	(-0.60)	(0.69)
Loan amount	0.137***	0.155***	0.187***	0.184***
	(7.35)	(8.75)	(13.94)	(26.42)
Term	0.0181***	0.0196***	0.0190***	-0.00292***
	(13.26)	(17.67)	(15.83)	(-7.96)
DTI	0.00902***	0.00899***	0.0125***	0.00643***
	(6.08)	(6.67)	(12.31)	(15.18)
Home ownership	-0.0167	-0.0185	-0.0843***	-0.192***
	(-0.78)	(-0.95)	(-5.54)	(-27.13)
Revolving util.	0.00177***	0.00163***	0.000482	-0.000575***
	(3.76)	(3.72)	(1.36)	(-3.58)
Monthly income	-0.536***	-0.548***	-0.538***	-0.339***
	(-23.57)	(-25.61)	(-31.83)	(-41.26)
Constant	-0.614***	-0.728***	-1.181***	-2.160***
	(-2.87)	(-3.71)	(-7.48)	(-28.50)
Observations	84550	95299	173883	886835
r^2	0.0540	0.0635	0.0516	0.0598
χ^2	3722.1	5234.7	7025.1	38453.7

Notes: Credit grade is Lending club grade (A-F) translated into numerical values with 1 being the low risk category 'A' and 6 being the high risk category 'G'. Base lending purpose is debt consolidation. Base FICO score is above 780.
Dependent variable: default; t statistics in parentheses

Results presented in Table 5.8 are similar to Emekter *et al.* (2015) who also found credit grade, DTI, FICO score and revolving credit line utilization significant). On the other hand, we also found monthly income, loan amount, term, income and purpose to be significant. The purpose with the lowest probability of default is credit card repayment followed by car loans. As expected, the lower credit grade, the more risky the loan is. For the FICO score the result is also in-line with the expectation – lower FICO higher probability of default. Investments into small businesses seem to be the riskiest purpose when other variables are kept equal. Results for learning sample are used further in the thesis for default prediction.

Based on the binary regression result, we can determine the probability of default for a loan from the estimated coefficient from Equation 5.1 that are presented in Table 5.8 and plug them into formula Equation 4.8 presented in Section 4.2:

$$p_i = \frac{1}{1 + e^{-d_i}}$$

As an example we can take a safe borrowers with credit grade 'A' (translated into 1), FICO score 700 (average FICO score that is translated into FICO category 3), DTI ratio 18% (average DTI of all loans), asking for $14800 for 36 months, monthly income of $6250, and revolving line utilization of 55%. $d_i = -2.53$ and $p_i = 0.074$. If we take a borrower with same characteristics but credit grade 'E', the probability of default would be 0.179.

Table 5.9: Logistic regression results by credit grades

Credit grade	Estimate of β	t-stat	exp(β)	PD	Actual DR
Constant (A)	-2.799	0.0228	0.0609	0.0037	0.06
B	0.7083	0.026	2.0305	0.11	0.11
C	1.1575	0.0265	3.182	0.1623	0.16
D	1.4666	0.0278	4.3345	0.2088	0.21
E	1.6612	0.0365	5.2656	0.2427	0.24
F	1.8993	0.0578	6.6812	0.2891	0.29
G	2.0352	0.0993	7.6538	0.3178	0.32

Notes: PD = probability of default. DR = default rate. Matured loans only. The base value of the model for credit grade is grade 1, being the lowest risk category. The highest grade is represented by the constant term of the model which is -2.799

To make a general expectation about the default probability in each credit grade, logistic regression employing matured loans and only credit grades is presented in Table 5.9. Resulted coefficients for credit grades are in line with the expectation that with the increasing risk of a borrower, also probability of default increases. The probability of default for the highest credit grade A is 0.4%, for B 11%, for C 16.23%, for D 20.88%, for E 24.27%, for F 28.29%, and for the lowest credit grade G 31.78%. These result are also consistent with overall loan default ratios presented in Table 5.5.

5.2.2 Predicting accuracy and cut-off rate optimization

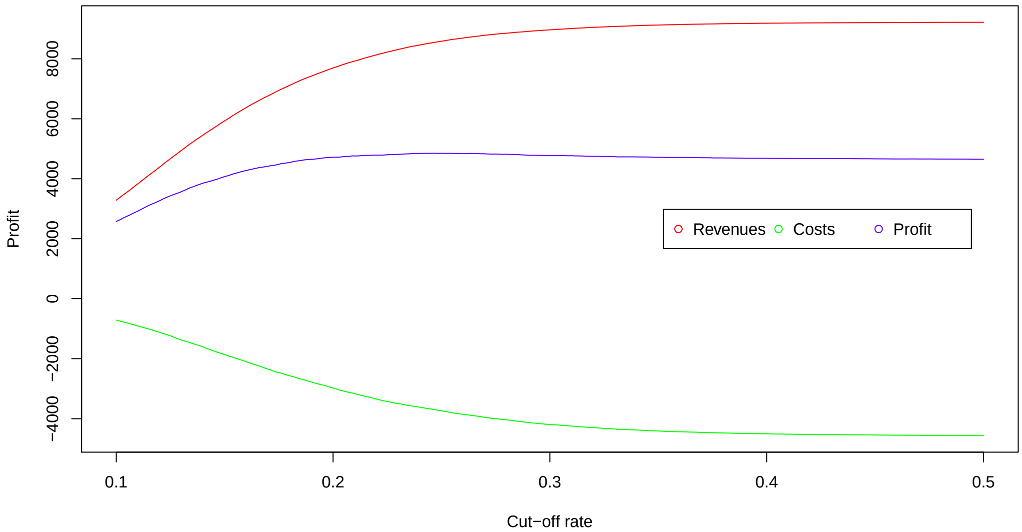
Table 5.10: Logit model predicting accuracy

	Actual	Predicted		
		Non Default	Default	Correct (%)
Training	Non Default	63915	8689	88.03
	Default	8689	3257	27.26
Testing	Non Default	72026	6201	92.07
	Default	9129	1977	17.8
Full dataset	Non Default	596343	185944	76.23
	Default	59323	45225	43.26

To verify how good our prediction model is in classifying default and non-default loans, we predict the default for each loan using results presented in Table 5.8. The threshold values for the classification of the loans is chosen as 0.219483 so that we get the same proportion of defaulted loans in the training part of data set which is 14.7%. Table 5.10 summarizes the predicting accuracy of our model. The model correctly classified 27.26% of default loans and 88.03% non-defaulted loans in the training dataset. For the testing dataset the ratios of corrected default and non-default loans are 17.8%, 92.07% respectively. Our model is much better in predicting non-default than default loans. We will use loans, that are predicted not to default in the portfolio management decision.

Possible optimization of cut-off rate is described in Subsection 4.2.1. We will try to maximize gains, by knowing, that average profit on defaulted loans is -0.383 and average return on good loan is 0.127. This ratio is quite good, because losses are not too costly – to break even we are need 3 performing loans for each defaulted. As profits we thus have number of properly predicted non-defaulting loans times 0.127 and for losses we multiply the number of loans that were predicted not to default but defaulted by -0.383. As the prediction of default is a function of cut-off rate, we will maximize the overall profit with respect to cut-off rate. Situation is illustrated on Figure 5.2.

Figure 5.2: Gains and losses for given level of cut-off rate



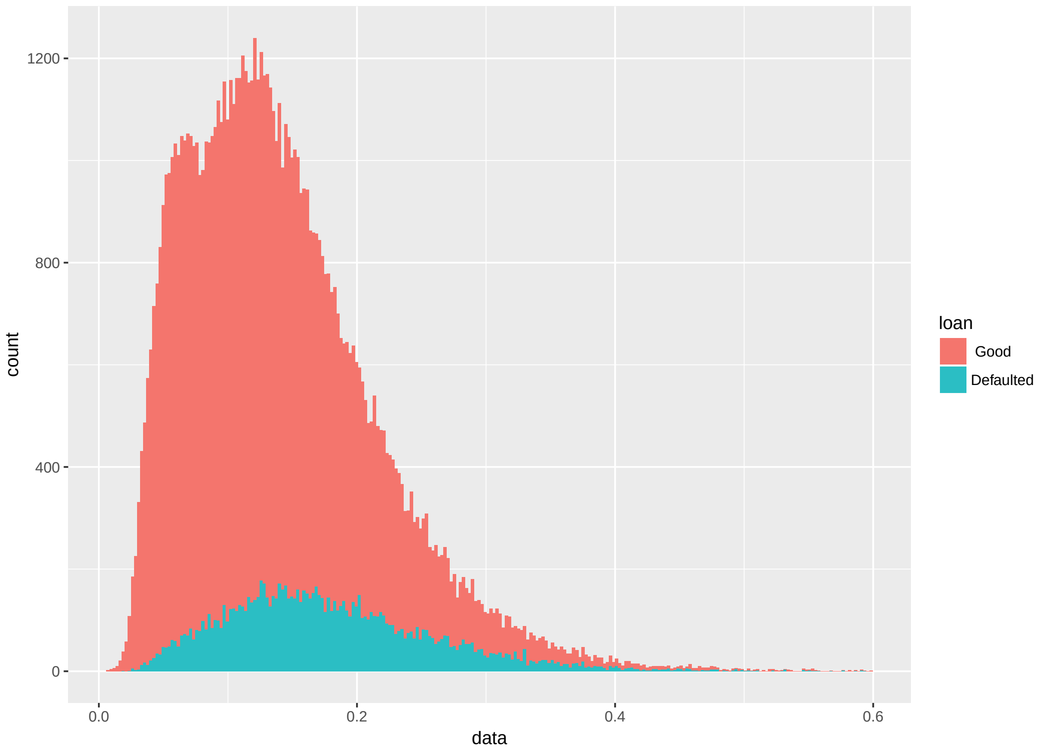
Investor achieves maximal profit for cut-off rate 0.25 for which the predicting accuracy is summarized in Table 5.11. If we compare efficiency of default

Table 5.11: Logit model predicting accuracy optimized

	Actual	Predicted Non Default	Default	Correct (%)
Training	Non Default	67615	4989	93.13
	Default	9748	2194	18.37
Testing	Non Default	75418	2809	96.41
	Default	10031	983	8.93
Full dataset	Non Default	665650	116637	85.1
	Default	62661	27470	30.48

predicting before and after the profit optimization, we can clearly see that on the testing dataset the predicting power for defaulted loans decreased by half and for non-defaulted loans increased only by 4 percentage points. The expected rate of return on the testing sample is 0.0727 for cut-off rate 0.25 and 0.0735 for cut-off rate 0.219483. If we look at expected return of the whole dataset, the non-optimized cut-off rate is again better (0.0675 vs. 0.0669). In the next section, where we will compare portfolio of all loans with those predicted no to default, we will use the naive cut-off rate of 0.219483.

Figure 5.3: Distribution of probability of default for defaulted and non-defaulted loans



Actual distribution of probability of defaults for each category of loans depicted by Figure 5.3 is not ideal. Distributions have different means (see Figure A.2), but are overlapping a lot. This is caused by the available data and resulting inability of performed estimation to separate defaulted and non-defaulted loans. Or even with better data both groups are very similar and it is statistically impossible to obtain better distributions.

5.3 Optimal portfolio finding

Theory of portfolio by Markowitz (1952); Sharpe (1964) is presented in Section 4.1. Let us recall the key aspects. Investors want to optimize their portfolio and have N securities available. In our case, investors can invest into loans that are differentiated by their credit grade, which gives us 7 "securities" available to investors. Each credit grade can be described by its expected value of return r_i and variance σ_i. Covariance between loan grades i and j is represented by σ_{ij}. Let w_i represent weight of credit grade i in the portfolio. Expected return

r_p and variance σ_p of the portfolio is given by Equation 4.1 and Equation 4.2:

$$r_p = \sum_{i=1}^{n} w_i \cdot r_i$$
$$\sigma_p = \sqrt{\sum_{i=1}^{n} \sum_{j=1}^{n} w_i w_j \sigma_{ij}}$$

Investors want to determine the efficient portfolio – such that (a) one cannot achieve higher expected return at the same level of variance (b) one cannot find portfolio with lower variance for the same level of expected return (Sharpe 1964). Investors cannot short sell loans and are therefore facing following constraints:

$$\sum_{i=1}^{n} w_i = 1, \quad w_i \geq 0 \quad \forall i = 1, \ldots, n \tag{5.2}$$

Investors can either maximize return for given level of risk or calculate minimum variance portfolio or maximize Sharpe ratio (Equation 4.4). Firstly we will obtain variance-covariance matrix and then compute all possible choices of investors.

5.3.1 Covariance between loan grades

Previous studies Singh *et al.* (2008); Guo *et al.* (2016) claim that covariance is zero and can be ignored without providing further reasoning or empirical evidence. Covariance is a measure of joint variability of two random variables X_i and X_j and is calculated as

$$cov\left(X_i, X_j\right) = E\left[\left(X - E\left[X_i\right]\right)\left(X_j - E\left[X_j\right]\right)\right] \tag{5.3}$$

From covariances we can easily create covariance matrix $\Sigma_{ij} = cov\left(X_i, X_j\right)$. To create covariance matrix for loan grades, we firstly need the time series of returns. For that purpose, we firstly calculate ROI of each loan. For simplicity and possibility to compare it with other forms of investments, we will not discount payments in time and we further assume, that investors reinvest their money. For fully paid and current loans the ROI is equal to their interest rates. For defaulted loans the $ROI = \frac{\text{total payment}-\text{loan amount}}{\text{loan amount}}$. Having ROI for each loan, we calculate averages for credit grades and month in which the loan was issued. This gives us 96 observations for each of 7 credit grades. Using Equation 5.3 we obtain unique VCV matrix presented in Table 5.13 and the corresponding correlation matrix captured by Table A.1 in the Appendix.

Table 5.12: Expected return and standard deviation of each credit grade

	A	B	C	D	E	F	G
Section A: Full sample							
Expected return	0.0502	0.0495	0.0442	0.0389	0.031	0.0164	0.0128
σ	0.0143	0.0276	0.0316	0.0377	0.0428	0.0861	0.1032
Section B: Years 2010–2015							
Expected return	0.0493	0.0578	0.0544	0.0486	0.0376	0.0247	0.0066
σ	0.01	0.0124	0.0167	0.0202	0.0305	0.0435	0.0781
Section C: Default predicted full sample							
Expected return	0.0511	0.0526	0.0547	0.0614	0.0584	0.0527	0.0866
σ	0.0151	0.0259	0.0324	0.0463	0.0701	0.1749	0.2033
Section D: Default predicted years 2010–2015							
Expected return	0.0498	0.0609	0.0665	0.0749	0.0737	0.0861	0.0884
σ	0.0105	0.0122	0.0147	0.0196	0.0392	0.1013	0.2155

Table 5.12 summarizes expected return for each credit grade along with its standard deviation (measure of risk). Figure 5.1 presented differences in interest rates between credit grades. Quite surprisingly, higher interest rates are not sufficient to provide higher returns for more risky credit grades. Credit grade 'A' offers the highest expected return along with the lowest risk. The lower the credit grade, the lower the expected return and higher risk. If we would deduct the 1% fee charged by Lending club and taxes, the riskiest credit grades 'F' and 'G' would not provide positive returns. For years 2010–2015 the situation a slightly different, but still the high risk credit grades provide lowest returns and highest risk at the same time.

Section C and section D of Table 5.12 capture loans that are predicted not to default. The expected return is higher for every credit grade, and the highest difference is for high risk grades. At the same time, standard deviation for low risky grades is almost the same, but for the high risky grades is significantly higher. After predicting the default, expected returns are increasing with the increase of risk (lower credit grade), which is the opposite of what we can see in section A and B of the table. We can easily conclude, that default prediction should improve performance of optimal portfolios which are constructed mainly from low risky loans.

Table 5.13: Variance-covariance matrix of credit grades – full sample

σ_{ij}	A	B	C	D	E	F	G
A	0.000204	0.000068	-0.000033	0.000056	0.000065	-0.000121	0.000345
B	0.000068	0.000763	0.00042	0.000602	0.000515	0.000134	-0.000243
C	-0.000033	0.00042	0.000996	0.000739	0.000488	0.000223	-0.000544
D	0.000056	0.000602	0.000739	0.001424	0.000975	0.000753	0.000322
E	0.000065	0.000515	0.000488	0.000975	0.001834	0.00054	0.000991
F	-0.000121	0.000134	0.000223	0.000753	0.00054	0.007409	0.000848
G	0.000345	-0.000243	-0.000544	0.000322	0.000991	0.000848	0.010652

Note: Negative covariance is marked with red colour.

Table 5.13 summarizes covariances between credit grades and their variance. These numbers are different from zero, which means, that there is a relation between credit grades expected return, which can be expected from the nature of the underlying securities – consumer loans. Interestingly, covariance between credit grades A and C, A and F, B and G, C and G is negative (marked red in the table), which means, that if one goes up, the other goes down. This can be effect of global financial crisis, because if we use data from 2010 onwards we get negative relation only between G and B, C Table 5.14.

Table 5.14: Variance-covariance matrix of credit grades – 2010-2015

σ_{ij}	A	B	C	D	E	F	G
A	0.0001	0.000077	0.00008	0.000071	0.000125	0.000129	0.000097
B	0.000077	0.000154	0.00013	0.000136	0.000123	0.000044	-0.000065
C	0.00008	0.00013	0.000277	0.000148	0.000187	0.00001	-0.000077
D	0.000071	0.000136	0.000148	0.000407	0.00035	0.000275	0.000086
E	0.000125	0.000123	0.000187	0.00035	0.000928	0.000406	0.000848
F	0.000129	0.000044	0.00001	0.000275	0.000406	0.00189	0.000889
G	0.000097	-0.000065	-0.000077	0.000086	0.000848	0.000889	0.006094

Note: Negative covariance is marked with red colour.

If we compare correlation matrices Table A.1 and Table A.2, we can observe greater correlation between low credit grades in the 2010–2015 sub-sample than in the whole period.

5.3.2 Optimal portfolio evaluation

Section A in Table 5.15 shows result of different portfolio optimizations when we take into account covariance between credit grades. We find credit grade shares for minimum variance portfolio, portfolio maximizing Sharpe ratio and equally weighted portfolio. For optimization we use full dataset and then subset of loans between 2010–2015. Expected return is around 5% and majority of the portfolio is invested into low risky loans. Both types of optimizations give us better portfolios than equally weighted – higher expected return and lower risk measured by the standard deviation.

Section B in Table 5.15 shows results for portfolio optimization with zero covariance between credit grades, as previous studies Singh *et al.* (2008); Guo *et al.* (2016) assumed. According to the new optimization it is best to invest in all credit grades, expected returns are smaller (with one exception) and risk is higher or same, which means that the portfolio optimization underestimates the returns and overestimates the risks. Both ways of optimization give us higher expected return than equally weighted portfolio, but for newer data, the equally weighted portfolio has lower risk.

In the next step, we perform the optimization of portfolio only using loans that we predict not to default. This results in the new variance-covariance matrix and also in the portfolio shares. Optimal portfolio is presented in section C in Table 5.15. Compared to the Section A, where portfolio random of loans in each credit grades is selected, default prediction increased expected returns for all portfolio selection by at least half percentage point. That is quite interesting and important improvement given the probability of default precision. The shares of credit grades is very similar for section A and section C, majority of the portfolio should be invested in the safest credit grade A and investor should not invest in credit grade D and lower.

There is a very little difference between portfolios constructed by different optimization targets – Sharpe ratio and minimum variance. When we employ the default prediction into our portfolio selection, expected returns increase with higher risk as shown in Table 5.12. This implicates, that equally weighted portfolio performs better in terms of expected return, but also standard deviation of such portfolio increases significantly.

We showed, that default prediction can significantly improve the portfolio performance and for risk seeking investors it makes now sense to invest also into the high risk loans. For investor seeking for the highest expected return regardless of the risk, credit grade A is the best option if we consider the full sample and credit grade B if only data from 2010. However, if investors make the decision also using the predicted defaults, credit grade G is the best choice.

Table 5.15: Portfolio optimization results

	Shares in portfolio in %							Standard deviation	Expected return	Sharpe ratio
	A	B	C	D	E	F	G			
Section A: With estimated covariances										
Minimum Variance full	77.59	4.02	15.61	0	0	2.78	0	0.0124	0.0483	3.9088
Minimum Variance 2010–2015	74.6	20.37	1.87	2.6	0	0	0.56	0.0097	0.0508	5.253
Sharpe ratio	79.01	5.5	14.02	0	0	1.47	0	0.0124	0.0488	3.9301
Sharpe ratio 2010–2015	60.77	38.19	1.03	0	0	0	0	0.0099	0.0526	5.3294
Equally weighted full	14.29	14.29	14.29	14.29	14.29	14.29	14.29	0.0277	0.0347	1.2535
Equally weighted 2010–2015	14.29	14.29	14.29	14.29	14.29	14.29	14.29	0.0192	0.0399	2.0804
Section B: With zero covariances										
Minimum Variance full	56.37	15.08	11.55	8.08	6.28	1.55	1.08	0.0107	0.0463	4.3193
Minimum Variance 2010–2015	41.04	26.7	14.85	10.12	4.44	2.18	0.68	0.0064	0.0509	7.9322
Sharpe ratio	61.05	16.11	11.01	6.78	4.2	0.55	0.3	0.0109	0.0475	4.3751
Sharpe ratio 2010–2015	39.71	30.32	15.88	9.67	3.27	1.06	0.09	0.0065	0.0519	8.0119
Equally weighted full	14.29	14.29	14.29	14.29	14.29	14.29	14.29	0.0218	0.0347	1.5923
Equally weighted 2010–2015	14.29	14.29	14.29	14.29	14.29	14.29	14.29	0.0002	0.0399	2.8113
Section C: Predicted defaults and covariances										
Minimum Variance full	76.2	5.21	16.9	0	0	1.05	0.64	0.0125	0.052	4.1708
Minimum Variance 2010-2015	65.17	12.44	2.51	16.22	2.82	0.84	0	0.0091	0.0567	6.248
Sharpe ratio	75.47	5.06	17.47	0.13	0	0.99	0.88	0.0125	0.0521	4.1751
Sharpe ratio 2010–2015	44.64	20.21	11.72	19.87	2.59	0.97	0	0.0093	0.06	6.4274
Equally weighted full	14.29	14.29	14.29	14.29	14.29	14.29	14.29	0.0485	0.0596	1.2296
Equally weighted 2010–2015	14.29	14.29	14.29	14.29	14.29	14.29	14.29	0.0385	0.0715	1.8581

5.4 Diversification

Theory of diversification is presented in Section 4.3 where from the academic discussion can be derived that optimal portfolio consists of about 50 stocks. P2P platforms emphasize to the investors that they should invest in multiple loans in order to diversify their portfolios and reduce risks. There are hundreds of thousands of loan requests published every year on P2P platforms, so there is a huge room for diversification possibilities. We will use simulation exercise using matured loans only to investigate the optimal size of portfolio in terms of number of loans in each credit grade.

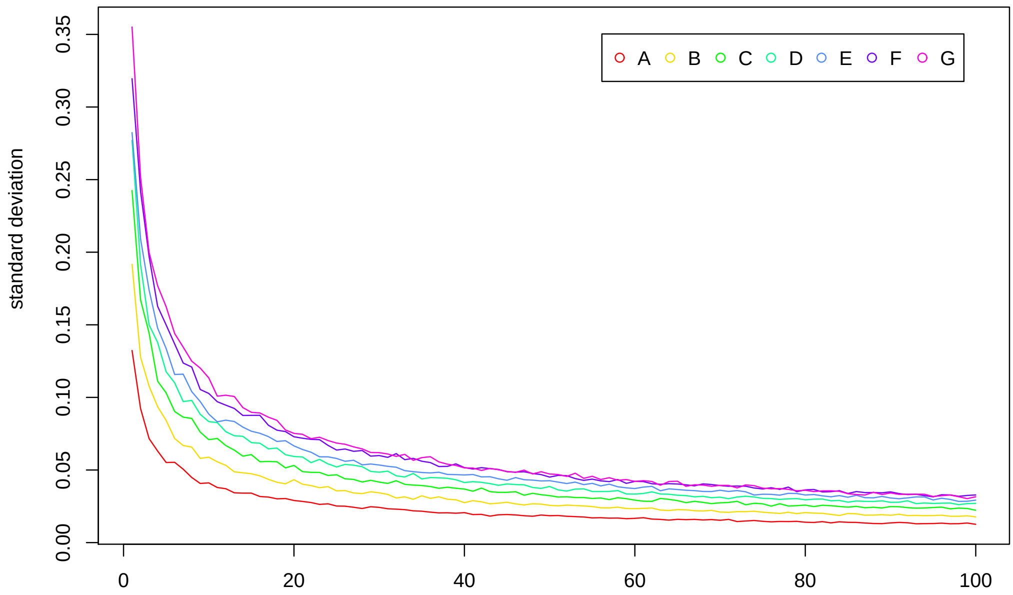
We should not forget another claim, that diversification should be performed as long as marginal costs are lower than marginal returns (Statman 1987). As a sample situation can be taken minimal investment of $25 into a "A" graded loan on Lending Club platform. Average interest rate for this credit grade is 7%. If the loan is for the term of 36 months, the total interest gain is about $2.4, because Lending Club takes 1% fee from the remaining principal of the loan. Investor should pay the income tax on that which reduces the gain to about $2, not accounting for possible losses. Average wage in the US is about $22[3]. For an investor, the search for such loan should take less than 6 minutes in order to cover the search costs.

In order to examine the diversification benefits on P2P platforms, we will run simulations similar to original study by Evans & Archer (1968). Diversification simulation can be described by the following procedure:

1. We use dataset of matured loans only.
2. Then we select all loans of specific score grade.
3. From these loans we randomly select n loans which determines our portfolio size.
4. We calculate mean return of the portfolio.
5. We random select portfolio of the same size 1000 times.
6. We calculate mean and standard deviation based on these 1000 observations.

[3] www.tradingeconomics.com

Figure 5.4: Effect of diversification for different credit grades – standard deviation



7. Then portfolio size is increased by 1 and process repeated up to the size of portfolio of 100.

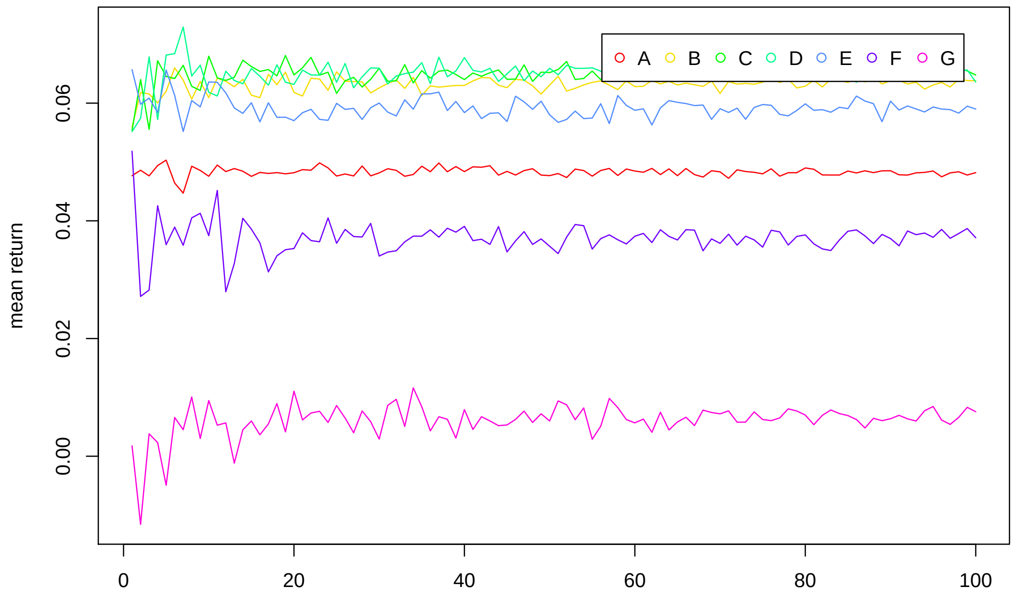
8. Mean and standard deviation is plotted.

9. We iterate for all credit grades.

In estimation performed by Evans & Archer (1968) which is depicted on Figure 4.7, not only the variance was decreasing with the number of securities in the portfolio, but also the mean return. In the simulation performed by this thesis using Lending Club data, the means are stable for each credit grade group as shows Figure 5.5. For each credit grade the mean is different – for more risky assets, the expected returns should be higher to cover also higher risks. Level of risk measured as standard deviation of portfolio returns is depicted on Figure 5.4. We can observe the reduction of variance with increasing number of loans in the portfolio. For each credit grade, the number of loans that should be included in the portfolio to reduce risk is different.

For the lowest risk credit grade 'A', adding 20 additional loans to portfolio of

Figure 5.5: Effect of diversification for different credit grades – mean return



20 loans has smaller effect than for highest risk credit grade 'G'. For each credit grade, the risk is decreasing with the number of loans added to the portfolio, but even for the largest portfolios of 100 loans, the risk is different between credit grades.

This last section provides insights into diversification effects on the P2P platforms. Investors should invest at least in about 10-20 loans to significantly reduce the variance of returns and hence reduce risk. We also proved, that there is a significance difference between credit grades not only in terms of variance, but also in therms of expected return. Larger variance of returns or in other terminology higher risk is correctly found by lower credit grades, which should be also more risky. This shows good credit rating ability of P2P platforms. However, the expected return is larger for the lowest risk loans graded 'A' then for highest risk loans graded 'G', which means that interest rates charged on highly risky loans are insufficient which is in-line with findings provided by Emekter *et al.* (2015). On the other hand, we find positive expected return for all credit grade groups and refute results by Singh *et al.* (2008) who found mostly negative returns on data from Prosper.com

Chapter 6

Conclusion

The aim of this thesis is to discuss P2P lending, a new alternative way of financing and type of crowdfunding, from investors perspective and portfolio theory. This perspective is quite unique, because P2P oriented research is mainly focused on the determinants of loans funding success and borrowers characteristics. The portfolio optimization oriented analysis follows previous research on optimal portfolios but uses unique dataset and type of market. We also enhance the first attempts to incorporate portfolio management thinking into P2P loan research by Singh *et al.* (2008); Guo *et al.* (2016), yet this thesis is the first one to incorporate non-zero covariance between loans into the decision marking process.

Portfolio performance depends on the performance of the underlying assets, loans in case of this thesis. For any loan provider the key is to filter out requests from applicants which are more likely to fail with their payments as defaulted loans most often result in a loss of part of the investment. This thesis uses standard methodology to predict loan defaults and compares performance of portfolio constituting of randomly selected loans with portfolio using loans which are predicted not to default. Following existing research, we estimate factors which influence the default rate. Along with the previous studies using different data, significant determinants of probability of default are credit grade and FICO score (Malekipirbazari & Aksakalli 2015; Emekter *et al.* 2015),and then loan purpose (Serrano-Cinca *et al.* 2015), income and loan amount (Byanjankar *et al.* 2015).

Second key finding of this thesis is that investors can reduce risk by investing into multiple loans of the same credit grade and to paraphrase Markowitz (1952) enjoy the only free lunch in finance – the benefits of diversification. Statman (1987) emphasizes that investors should diversify as long as the marginal costs are smaller than the marginal gains from diversification. Investing into P2P loans is not limited by fixed transaction costs related to each loan, but investors only pay fee based on the total investment. Investors should therefore invest into multiple loans, and the number of loans should increase with the increase of the risk for the given credit grade. Up to my best knowledge, such analysis has never been done on this type of data.

Third and unique finding of this thesis is: there is a non-zero covariance between credit grades. Portfolio optimization provides very different result when accounting for the covariance between credit grades. Covariance is significantly different from zero and if investors want to build a well diversified portfolio, it is a mistake to ignore it as Singh *et al.* (2008); Guo *et al.* (2016). We also find that covariance changes over time, but changes in the optimal portfolio are only small. We also agree with Emekter *et al.* (2015) that higher risk loans do not provide sufficient returns to cover the risks, because highest expected return show low risk loans.

Optimal investor choice should consist of three steps. In the first step, investors should predict defaults in order to determine criteria for loan selection. In the second step, investors should use their risk preferences to determine optimal distribution of investments between credit grades. In the last step, investor should invest smaller amount into multiple loans to really seize all the benefits of portfolio diversification.

Future research should focus on the robustness of our results by using different datasets. Investigating effects of diversification between platforms and even countries might provide a guidance on how much investors should invest on different platforms to get well optimized portfolio of P2P loans. The evolution of P2P lending will continue together with the increasing importance on the financial markets. Recent publications EBA (2015) or the FCA (2015) seem to be promising in the area of regulatory approaches. It is critical, that regulatory bodies follow current trends and support development and innovation of alternative finance.

Bibliography

Agarwal, S. & R. Hauswald (2008): "The choice between arm's-length and relationship debt: Evidence from eloans." *Technical report*, Working Paper, Federal Reserve Bank of Chicago.

Ashta, A. & D. Assadi (2009): "An analysis of european online micro-lending websites." *Cahiers du CEREN* **29**: pp. 147–160.

Bachmann, A., A. Becker, D. Buerckner, M. Hilker, F. Kock, M. Lehmann, P. Tiburtius, & B. Funk (2011): "Online peer-to-peer lending–a literature." *Journal of Internet Banking and Commerce* **16(2)**.

Barasinska, N. (2009): "The role of gender in lending business: Evidence from an online market for peer-to-peer lending." *The New York Times* **217266**: pp. 1–25.

Bastos, J. A. (2010): "Forecasting bank loans loss-given-default." *Journal of Banking & Finance* **34(10)**: pp. 2510 – 2517.

Benjelloun, H. (2010): "Evans and archer–forty years later." *Investment Management and Financial Innovations* **7(1)**: pp. 98–104.

Berger, S. C. & F. Gleisner (2009): "Emergence of financial intermediaries in electronic markets: The case of online p2p lending." *BuR-Business Research* **2(1)**: pp. 39–65.

Byanjankar, A., M. Heikkilä, & J. Mezei (2015): "Predicting credit risk in peer-to-peer lending: A neural network approach." In "Computational Intelligence, 2015 IEEE Symposium Series on," pp. 719–725. IEEE.

Collier, B. C. & R. Hampshire (2010): "Sending mixed signals: multilevel reputation effects in peer-to-peer lending markets." In "Proceedings of the

2010 ACM conference on Computer supported cooperative work," pp. 197–206. ACM.

Dědek, O. (2016): "Portfolio theory lecturer's notes no. 1." `http://dedeklegacy.cz/documents-pa.html` [Accessed: 2015-12-31].

Domian, D. L., D. A. Louton, & M. D. Racine (2007): "Diversification in portfolios of individual stocks: 100 stocks are not enough." *Financial Review* **42(4)**: pp. 557–570.

Duarte, J., S. Siegel, & L. Young (2012): "Trust and credit: The role of appearance in peer-to-peer lending." *Review of Financial Studies* **25(8)**: pp. 2455–2484.

EBA (2015): "Opinion of the european banking authority on lending-based crowdfunding." *Technical report*, European Banking Authority.

Emekter, R., Y. Tu, B. Jirasakuldech, & M. Lu (2015): "Evaluating credit risk and loan performance in online peer-to-peer (p2p) lending." *Applied Economics* **47(1)**: pp. 54–70.

Evans, J. L. & S. H. Archer (1968): "Diversification and the reduction of dispersion: an empirical analysis." *The Journal of Finance* **23(5)**: pp. 761–767.

Everett, C. R. (2010): "Group membership, relationship banking and loan default risk: the case of online social lending." *Relationship Banking and Loan Default Risk: The Case of Online Social Lending (March 15, 2010)* .

FCA (2015): "A review of the regulatory regime for crowdfunding and the promotion of non-readily realisable securities by other media." *Technical report*, UK Financial Conduct Authority.

Freedman, S. & G. Z. Jin (2008): "Do social networks solve information problems for peer-to-peer lending? evidence from prosper. com." *Technical report*.

Galloway, I. (2009): "Peer-to-peer lending and community development finance."

Garman, S., R. Hampshire, & R. Krishnan (2008): "A search theoretic model of person-to-person lending." *Technical report*, Heinz College Carnegie Mellon University.

Greiner, M. E. & H. Wang (2009): "The role of social capital in people-to-people lending marketplaces." *ICIS 2009 Proceedings* **(29)**.

Guo, Y., W. Zhou, C. Luo, C. Liu, & H. Xiong (2016): "Instance-based credit risk assessment for investment decisions in p2p lending." *European Journal of Operational Research* **249(2)**: pp. 417–426.

Herzenstein, M., R. L. Andrews, U. Dholakia, & E. Lyandres (2008): "The democratization of personal consumer loans? determinants of success in online peer-to-peer lending communities."

Herzenstein, M., U. M. Dholakia, & R. L. Andrews (2011): "Strategic herding behavior in peer-to-peer loan auctions." *Journal of Interactive Marketing* **25(1)**: pp. 27–36.

Hicks (1934): "Application of mathematical methods to the theory of risk." .

Iyer, R., A. Ijaz, K. Erzo, F. Luttmer, & K. Shue (2010): "Inferring asset quality: determining borrower creditworthiness in peer-to-peer lending markets." .

Iyer, R., A. I. Khwaja, E. F. Luttmer, & K. Shue (2009): "Screening in new credit markets: Can individual lenders infer borrower creditworthiness in peer-to-peer lending?" *Technical report*, National Bureau of Economic Research.

Jensen, M. C. & W. H. Meckling (1976): "Theory of the firm: Managerial behavior, agency costs and ownership structure." *Journal of Financial Economics* **3(4)**: pp. 305–360.

Käfer, B. (2016): "Peer-to-peer lending–a (financial stability) risk perspective." *Joint Discussion Paper Series in Economics* **(22-2016)**.

Klafft, M. (2008): "Peer to peer lending: auctioning microcredits over the internet." In "Proceedings of the International Conference on Information Systems, Technology and Management, A. Agarwal, R. Khurana, eds., IMT, Dubai," .

Krumme, K. & S. Herrero-Lopez (2009): "Do lenders make optimal decisions in a peer-to-peer network?" In "Web Intelligence and Intelligent Agent Technologies, 2009. WI-IAT '09. IEEE/WIC/ACM International Joint Conferences on," volume 1, pp. 124–127.

Lee, E. & B. Lee (2012): "Herding behavior in online {P2P} lending: An empirical investigation." *Electronic Commerce Research and Applications* **11(5)**: pp. 495 – 503.

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

Markowitz, H. (1952): "Portfolio selection." *The journal of finance* **7(1)**: pp. 77–91.

Morse, A. (2015): "Peer-to-peer crowdfunding: Information and the potential for disruption in consumer lending." *Annual Review of Financial Economics* **7**: pp. 463–482.

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

Puro, L., J. E. Teich, H. Wallenius, & J. Wallenius (2010): "Borrower decision aid for people-to-people lending." *Decision Support Systems* **49(1)**: pp. 52 – 60.

Ravina, E. (2012): "Love & loans: the effect of beauty and personal characteristics in credit markets." *Available at SSRN 1101647* .

Roll, R. (1977): "A critique of the asset pricing theory's tests part i: On past and potential testability of the theory." *Journal of financial economics* **4(2)**: pp. 129–176.

Serrano-Cinca, C., B. Gutiérrez-Nieto, & L. López-Palacios (2015): "Determinants of default in p2p lending." *PloS one* **10(10)**: p. e0139427.

Sharpe, W. F. (1964): "Capital asset prices: A theory of market equilibrium under conditions of risk." *The journal of finance* **19(3)**: pp. 425–442.

Sharpe, W. F. (2000): *Portfolio Theory and Capital Markets*. McGraw-Hill series in finance. McGraw-Hill.

Singh, H., R. Gopal, & X. Li (2008): "Risk and return of investments in online peer-to-peer lending." *Technical report*, University of Texas. 2008: l-5.

Slavin, B. (2007): "Peer-to-peer lending – an industry insight." *Retrieved from bradslavin. com* pp. –.

Statman, M. (1987): "How many stocks make a diversified portfolio?" *Journal of Financial and Quantitative Analysis* **22(03)**: pp. 353–363.

Statman, M. (2004): "The diversification puzzle." *Financial Analysts Journal* **60(4)**: pp. 44–53.

Wu, J. & Y. Xu (2011): "A decision support system for borrower's loan in p2p lending." *Journal of Computers* **6(6)**: pp. 1183–1190.

Yum, H., B. Lee, & M. Chae (2012): "From the wisdom of crowds to my own judgment in microfinance through online peer-to-peer lending platforms." *Electronic Commerce Research and Applications* **11(5)**: pp. 469 – 483. Poverty, Technology and Microfinance.

Appendix A

Additional tables and plots

Table A.1: Correlation matrix of credit grades – full sample

ρ_{ij}	A	B	C	D	E	F	G
A	1	0.172	-0.074	0.103	0.107	-0.098	0.234
B	0.172	1	0.482	0.577	0.435	0.056	-0.085
C	-0.074	0.482	1	0.62	0.361	0.082	-0.167
D	0.103	0.577	0.62	1	0.604	0.232	0.083
E	0.107	0.435	0.361	0.604	1	0.147	0.224
F	-0.098	0.056	0.082	0.232	0.147	1	0.095
G	0.234	-0.085	-0.167	0.083	0.224	0.095	1

Note: Negative correlation is marked with red colour.

Table A.2: Correlation matrix of credit grades – 2010–2015

ρ_{ij}	A	B	C	D	E	F	G
A	1	0.618	0.477	0.351	0.411	0.296	0.124
B	0.618	1	0.627	0.543	0.326	0.081	-0.067
C	0.477	0.627	1	0.439	0.368	0.014	-0.06
D	0.351	0.543	0.439	1	0.57	0.314	0.055
E	0.411	0.326	0.368	0.57	1	0.307	0.356
F	0.296	0.081	0.014	0.314	0.307	1	0.262
G	0.124	-0.067	-0.06	0.055	0.356	0.262	1

Note: Negative correlation is marked with red colour.

Figure A.1: Effect of diversification for credit grade A

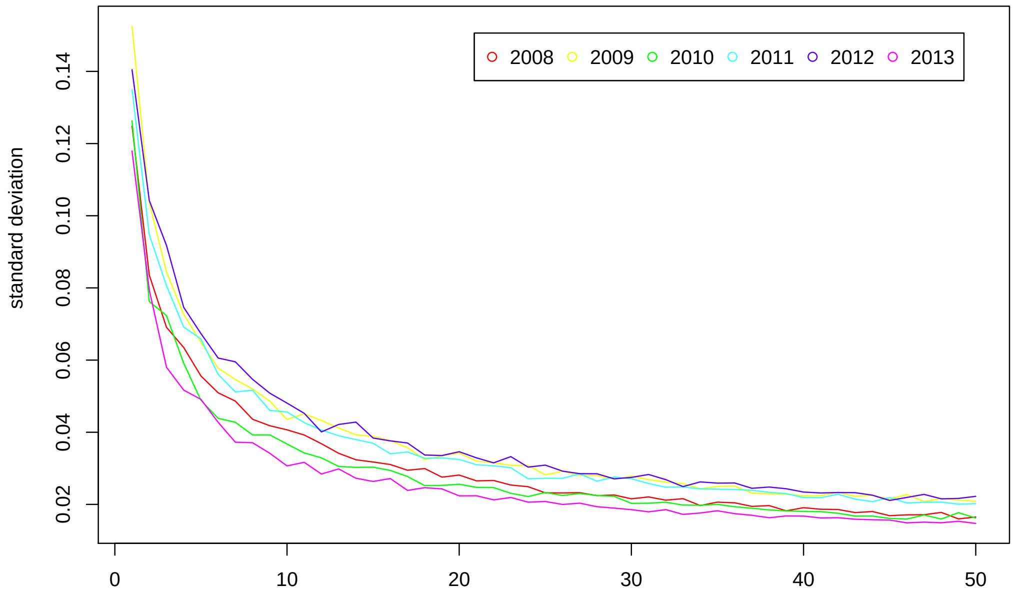


Figure A.2: Distribution of probability of defaults