Polena, Michal; Regner, Tobias

Working Paper

Determinants of borrowers' default in P2P lending under consideration of the loan risk class

Jena Economic Research Papers, No. 2016-023

Provided in Cooperation with:
Friedrich Schiller University Jena, Faculty of Economics and Business Administration

JENA ECONOMIC RESEARCH PAPERS

2016 – 023

Determinants of borrowers' default in P2P lending under consideration of the loan risk class

by

Michal Polena
Tobias Regner

www.jenecon.de

ISSN 1864-7057

The JENA ECONOMIC RESEARCH PAPERS is a joint publication of the Friedrich Schiller University Jena, Germany. For editorial correspondence please contact markus.pasche@uni-jena.de.

Impressum:

Friedrich Schiller University Jena
Carl-Zeiss-Str. 3
D-07743 Jena
www.uni-jena.de

Determinants of borrowers' default in P2P lending under consideration of the loan risk class

Michal Polena ♠ Tobias Regner ♠ *
♠ *University of Jena, Germany*

Abstract

We study the determinants of borrowers' default in P2P lending with a new data set consisting of 70,673 loan observations from Lending Club. Previous research identified a number of default determining variables but did not distinguish between different loan risk levels. We define four loan risk classes and test the significance of the default determining variables within each loan risk class. Our findings suggest that the significance of most variables depends on the loan risk class. Only few variables are consistently significant across all risk classes. The debt-to-income ratio, inquiries in the past 6 months and a loan intended for a small business are positively correlated with the default rate. Annual income and credit card as loan purpose are negatively correlated.

JEL classifications: D14, E41, G23

Keywords: crowdfunding, peer-to-peer lending, P2P, credit grade, FICO score, default risk

*This paper is based on the master thesis of Michal Polena (michal.polena@outlook.com) submitted at the University of Jena. Tobias Regner (tobias.regner@uni-jena.de) gratefully acknowledges support by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) - project number 628902.

1 Introduction

Peer-to-peer (P2P) lending platforms are new financial intermediaries connecting borrowers and lenders. Both might benefit from using P2P lending platforms. Borrowers get on average lower interest rates on their loans than at banks. Lenders with a well diversified loan portfolio earn more money than with bank's saving accounts. Technologically innovative P2P lending platforms facilitate loans with low intermediation costs and thus pose a threat for traditional banks [Deloitte, 2014]. Unsurprisingly, the popularity of P2P lending is rising rapidly. For example, Lending Club, the biggest P2P lending platform in the world, almost doubled the amount of issued loans from USD 4.4 billion in 2014 to USD 8.4 billion in 2015. The remarkable growth of P2P lending is present in Europe [Wardrop et al., 2015] as well as in China [Wang et al., 2015].

The fundamental problem of lending is information asymmetry between borrowers and lenders: borrowers have more information about their creditworthiness than lenders have. P2P lending platforms try to decrease this information asymmetry. They apply credit scoring techniques and assign a risk grade to each loan that may serve as a signal for lenders. Indeed, existing research [Emekter et al., 2015, Carmichael, 2014, Serrano-Cinca et al., 2015] finds a positive correlation between a loan's default and the assigned risk grade. They also find further determinants of the default rate, for instance, the debt-to-income ratio or revolving credit utilization.

We conjecture that the significance of these default determinants depends on the loan's risk grade. Thus, the goal of our study is to evaluate known determinants of borrowers' default for each risk grade separately. We test this with a new data set consisting of 70,673 loan observations from Lending Club. Loans in our data set have a 36-month duration and were issued between January 2009 and December 2012, thus avoiding a structural break in the data due to the financial crisis in 2007/2008.

We identify *Annual Income*, the *Debt-to-Income* ratio, *Inquiries in Past 6 Months* and the loan purposes *Credit Card* and *Small Business* as significant determinants of default in the full data set and also across all loan risk classes. The significance of other variables depends on the loan risk class. For example, *Revolving Credit Utilization* which is significant in our full data set and in less risky loan classes is not significant in loan classes with riskier loans.

We conclude that whether loan/borrower characteristics can be used to predict a loan's default chances actually depends on the loan's risk class. We connect our findings to the literature on funding success of P2P loans in an effort to understand to what extent insights about default determinants are anticipated by lenders' choices when funding a loan. Generally, our results

contribute to a better understanding of the mechanisms of P2P lending. Potential lenders, especially those investing in high risk loans, can use our findings for their advantage and allocate their money more effectively.

The remainder of the paper is organized as follows. Related literature is reviewed in Section 2 and in Section 3 we develop our hypotheses. We describe our data set in Section 4 and report our findings in Section 5. Section 6 concludes.

2 Related literature

P2P lending platforms are currently experiencing exponential growth[1] with the USA being the biggest P2P lending market. According to Wardrop et al. [2015], there was an average yearly growth of 113% in P2P consumer lending between 2012 and 2014 in Europe (excluding the UK). The amount of funded P2P consumer loans increased from EUR 62.5 million in 2012 to EUR 274.6 million in 2014. Furthermore, Wang et al. [2015] add that P2P lending has been rapidly growing in China since its inception in 2007. According to Deer et al. [2015], there were 1,575 P2P lending platforms in 2014 with an estimated volume of funded loans between USD 20 and 40 billion by the end of 2015. These numbers would make China the second largest P2P lending market in the world.

2.1 Funding success of P2P loans

A number of studies explore what factors contribute to the funding success of P2P loans. Most are based on data from the platform Prosper which used to be the biggest P2P lending platform in the USA. Prosper had many social features, such as a discussion forum and detailed borrowers' characteristics including their photos. Studies, among others, by Lin et al. [2013] and Freedman and Jin [2014], stress the importance of social relationships for funding success. They find that borrowers with better social ties are more likely to get their loans funded and to receive a lower interest rate. However, social features were completely removed by Prosper in 2008.[2]

Several studies focus on herding behavior in P2P lending. Herzenstein et al. [2011] conclude that a 1% increment in the number of bids represents a 15% increase of the probability of an additional bid (until the loan is fully funded). They also control for borrower/loan characteristics and find that the debt-to-income ratio is negatively correlated with funding, while the

[1] Two reasons for the rapid emergence of P2P lending platforms are put forward in the literature: low intermediation costs [Wu, 2014, Namvar, 2013] and credit rationing after the financial crisis in 2007/2008 [Mills, 2014].

[2] The Securities and Exchange Commission (SEC) ordered P2P lending companies to register their loans as securities and provide them through a bank.

credit grade is positively correlated with funding. They find no relationship between funding and home ownership or the requested loan amount. Zhang and Liu [2012] find that lenders observe their peers' lending decisions and use this information to infer creditworthiness of borrowers. Among their control variables, the debt-to-income ratio is negatively correlated with funding, while the credit grade, home owner status and the amount requested are positively correlated with funding.

2.2 Determinants of borrowers' default

Investing at P2P lending platforms is a risky activity, because the offered loans are not secured. In order to decrease the information asymmetry between lender and borrower, borrowers are obliged to provide some personal information, such as annual income or the loan's purpose. For example, borrowers at Lending Club are required to provide detailed information about themselves and their credit history. P2P lending platforms use this information to assess the likelihood of borrowers' default and assign him or her an appropriate interest rate with a given grade class.[3] It is generally assumed that the better the grade the more likely is the borrower to repay his or her debt.

There are several studies, such as Iyer et al. [2009] and Freedman and Jin [2014], examining borrowers' characteristics and their influence on borrowers' default based on data from Prosper. We do not review these studies in detail because of the differences between the platforms Prosper and Lending Club.[4] Instead, we focus on studies examining borrowers' default determinants based on Lending Club data: Emekter et al. [2015], Carmichael [2014] and Serrano-Cinca et al. [2015].

All three are in consensus that the *Credit Grade*[5] assigned by Lending Club is the best predictor for borrowers' default. Moreover, *Revolving Credit Line Utilization* is another variable influencing the default rate mentioned in all three papers. Findings of other default determinants vary. The discrepancy between the findings of Emekter et al. [2015], Carmichael [2014] and Serrano-Cinca et al. [2015] might be caused by three different factors. The first factor is the selection of variables potentially having an impact on borrowers' default. For example, Carmichael [2014] and Emekter et al. [2015] found out that the *FICO score* has an influence on

[3] An accurate credit scoring predictive model is crucial for P2P lending platforms. Abdou and Pointon [2011] conduct an extensive literature review of more than 200 articles about credit scoring models. They conclude that there does not exist a single best statistical technique used for the creation of credit scoring models.

[4] Before the SEC regulation, as discussed above, Prosper used the Dutch auction to determine the appropriate interest rate for borrowers. Moreover, Prosper used social features enabling social network effects between borrowers and lenders. Even after the SEC regulation, there are still significant differences between the platforms. These differences might make the comparison of determinants influencing borrowers' default inaccurate.

[5] In order to better differentiate and highlight variables, we write them with capital letters and in italics.

Table 1: Summary of Borrowers' Default Determinants

Name of study	Data set	Method used	Findings
Emekter et al. (2015)	May 2007 - June 2012 (36- & 60-month loans)	Binary logistic regression	Credit Grade, FICO score, Debt-to-Income and Revolving Credit Utilization
Serrano-Cinca et al. (2015)	January 2008 - December 2011 (36-month loans)	Univariate means test and Cox regression	Credit Grade, Annual Income, Loan Purpose, Debt-to-Income, Current Housing Situation, Credit History Length, Revolving Credit Utilization, Recent Inquiries, Deliquency in Past 2 Years, Open Credit Lines
Carmichael (2014)	June 2007 - October 2013 (36-month loans)	Dynamic logistic regression	Credit Grade, Annual Income, Loan Purpose, FICO score, Revolving Credit Utilization, Recent Inquiries, Credit History Length, Time since Last Delinquency, Loan Amount, Loan Description

default. Serrano-Cinca et al. [2015] did not choose the *FICO score* as an independent variable in their study. The second factor potentially creating discrepancy between the findings is the data set used. Specifically, differences in time frames, classification of loan status or type of loan length might be the cause. For example, Emekter et al. [2015] and Serrano-Cinca et al. [2015] used only 36-month loans. Instead, Carmichael [2014], used both, 36- and 60-month loans. The last factor which might cause the discrepancy is the research method used. Carmichael [2014] used dynamic logistic regression to assess determinants influencing default rate in P2P lending. Serrano-Cinca et al. [2015] conducted their study with a combination of univariate means test and Cox regression. Emekter et al. [2015] chose binary logistic regression for their analysis. For better clarity, we summarize this information in Table 1.

3 Hypotheses

Seven different loan credit grades, from A to G, can be assigned to a loan at Lending Club. Some investors at Lending Club intentionally invest into loans with high-risk credit grades, such as E, F or G. Riskier loans have higher net annualized returns after accounting for defaults than less risky loans, such as A or B. For example, loans with credit grades F or G have an average net annualized return of 8.64% compared to the 5.25% from A-graded loans and 7.29% from B-graded loans.[6]

Weiss et al. [2010] argue that the loan grade assigned by the P2P lending platform is the most important factor considered by investors when allocating their money. However, as discussed in the literature review, investors should also take into account characteristics of borrowers and loans. It could help them to increase their profit by allocating their funds more effectively. According to Emekter et al. [2015], Serrano-Cinca et al. [2015] and Carmichael [2014] the *Annual Income*, *Revolving Credit Line Utilization* and the *Debt-to-Income ratio*, among others, have an influence on the borrowers' default rate.

[6]Lending Club Statistics: https://www.lendingclub.com/info/demand-and-credit-profile.action

It is questionable, however, whether a lender investing, for example, in riskier loans with credit grades D or E can rely on these findings. Results from these studies are based on data including all types of loan grades (from A to G). Generally, the majority of loans is issued as A- or B-graded, that is, less risky loans. Therefore, the findings might be biased because of the high number of issued loans with less risky credit grades. Thus, the significance of borrowers' default determinants might not be valid for loans with riskier grades. To the best of our knowledge, this study is the first one to test determinants of borrowers' default within given loan risk classes.

Besides the *Loan Grade* assigned by Lending Club, Emekter et al. [2015], Serrano-Cinca et al. [2015] and Carmichael [2014] find *Revolving Credit Utilization* to be a positive determinant of borrowers' default. However, *Revolving Credit Utilization* might not be a determinant across all risk classes. On the one hand, people who are assigned to low risk classes tend to have few experience with handling negative external shocks to their available income. Hence, an increase of *Revolving Credit Utilization* might be a precursor of an upcoming default. On the other hand, people from high risk classes tend to be close to their maximum debt limit. Any need for an additional credit might then not translate in *Revolving Credit Utilization* but immediately in borrowers' default because of insufficient credit repayment reserves.

H1: *Revolving Credit Utilization is a significant determinant of borrowers' default in all loan risk classes.*

Emekter et al. [2015] and Serrano-Cinca et al. [2015] conclude that the *Debt-to-Income* ratio is a default determinant in P2P lending. We test whether the *Debt-to-Income* ratio is a reliable predictor across all risk levels.

H2: *The Debt-to-Income ratio is a significant determinant of borrowers' default in all loan risk classes.*

Serrano-Cinca et al. [2015] find that the current *Housing Situation* influences the borrowers' probability of default. Home ownership (whether it is mortgaged or not) is associated with lower chances of default in comparison to renting. However, home ownership may not be a reliable indicator of loan default for different risk levels.

H3: *The Current Housing Situation is a significant determinant of borrowers' default in all loan*

risk classes.

Carmichael [2014] finds that borrowers' self-claimed creditworthiness and the loan description lacking complete sentences are significant determinants of default. In addition to that, Nowak et al. [2015] studies non-missing loan descriptions of small businesses' loans at Lending Club. They find that loan descriptions with more words and characters as well as descriptions with misspelled words are less likely to be funded by investors. Thus, we expect that creditworthy borrowers invest higher effort in the explanation of their loan purpose resulting in, on average, a higher number of characters in the loan description. We test whether this is the case in all loan classes.

H4: *Creditworthy borrowers do not use more characters in their loan descriptions than borrowers who defaulted.*

Based on the studies of Serrano-Cinca et al. [2015] and Carmichael [2014] some loan purposes exhibit a lower frequency of default than others, for instance, if the loan is used for a *Wedding*, *Car* or *Credit Card* consolidation. People with a high *FICO score* can be regarded as being able to meet their liabilities or not having any liabilities at all. These people usually get a lower *Loan Grade*. We hypothesize that people in low risk classes borrow money solely on well-considered purposes. Therefore, their potential default is unlikely to be related to a specific *Loan Purpose* but rather by unexpected circumstances. As a consequence we expect that there are no default rate differences in low loan risk classes across the various *Loan Purposes* but only in high risk classes.

H5: *The Loan Purpose is a significant determinant of borrowers' default in low loan risk classes.*

Serrano-Cinca et al. [2015] and Carmichael [2014] find that the *Length of Credit History* is negatively correlated to borrowers' default. The longer the credit history is, the less likely is the borrower to default. We do, however, hypothesize that people in low loan risk classes are not more likely to default, if they have a shorter credit history. Instead, we believe the effect of the credit history's length is only relevant in high risk classes.

H6: *The Length of Credit History is a significant determinant of borrowers' default in low loan risk classes.*

4 Data description

The aim of our study is to evaluate determinants of borrowers' default within given loan grade classes in P2P lending. The data we use come from Lending Club, the biggest P2P lending platform in the world with total loan issuance of almost $16 billion by the end of 2015.[7] First of all, we explain the Lending Club process and the way how a prospective borrower can apply for a loan. Secondly, we describe our data set. Thirdly, we explain the main variables of interest. At the end of this section, we provide descriptive statistics of our variables and correlational matrixes.

4.1 The Lending Club process

Lending Club connects people who want to borrow money with people who are willing to lend their money. Before applying for a loan at Lending Club, a prospective borrower should find out the value of his or her *FICO score*. The *FICO score* is a credit score which is widely used by banks and credit providers in the USA.[8] The *FICO score* represents the creditworthiness of a person, that is, it shows the likelihood that a borrower will meet his or her liabilities. The *FICO score* is computed based on a borrower's personal credit report provided by national credit bureaus in the USA. The exact formula for the *FICO score* computation is held secret. Only approximate weights of given categories are made public. The total *FICO score* is made up from five categories from a person's financial history. The highest weight, about 35%, gets the *payment history*[9] with information, such as bankruptcy, charge offs or late payments. The second category with approximately 30% weight is *debt burden*. The *debt burden* category is associated with debt metrics, such as the amount owed on all accounts, the credit utilization ratio on revolving accounts or the number of accounts with balances. The *length of credit history* is the third category with 15% weight. The metrics of this category are linked to the age of a borrower's credit accounts. The last two categories, *types of credit used* and *recent searches for credit*, have both 10% weight in the *FICO score*. As the name suggests, the *types of credit* category is computed based on the types of credit the borrower has, such as consumer loan or mortgage. The *recent searches for credit* category consists of information about recent credit inquiries. About 90% of borrowers' applications at Lending Club is rejected because of an insufficient *FICO score*. Only potential borrowers with a *FICO score* of at least 600 are allowed to apply for a loan at Lending Club.

[7] Source: https://www.lendingclub.com/info/statistics.action (accessed April 13, 2016)

[8] According to http://www.myfico.com/, up to 90% of top lenders use the FICO score.

[9] The FICO score categories are written in *italics* in order to improve readers' comprehension. Moreover, they are written in small letters to differentiate them from our variables.

The potential borrower is further asked to provide some personal and loan information. The self-reported information is his or her *Annual Income*, the current *Home Situation* (potential options are own, mortgage or rent), the *Length of Employment*, the *Loan Purpose* and a *Loan Description*. All of this information, except *Loan Description*, are mandatory.

After checking a borrower's *FICO score* and his or her self-reported information, Lending Club assigns him or her a risk *Loan Grade*, from A to G, followed by a more accurate risk *Loan Subgrade*, from A1 to G5, and a corresponding interest rate. The interest rate charged for A1 was 5.32% and 28.99% for G5 in the first quarter of 2016. Lending Club's credit scoring model is kept secret. The P2P lending platform, however, affirms that the risk *Loan Grade* and *Subgrade* are computed based on the borrower's *FICO score* and his or her personal and loan information.

If the offered loan conditions and the interest rate are accepted by the borrower, Lending Club announces the loan on its website. Potential lenders can then view the loan online and start to fund it. During the loan's funding period, Lending Club might ask the borrower to verify the self-reported information. The loan might be removed from Lending Club's website, if the borrower's self-reported information cannot be verified. However, if the loan gets funded before the verification is done, the verification is not needed anymore and the loan is issued.

4.2 Our data set

In an effort to be fully transparent about company and loan performance, Lending Club makes public the data of every loan they have ever issued. The information about these loans used to be updated daily, then monthly and currently is updated quarterly. Our Lending Club data set was downloaded in February 2016. It contains information about 884,633 loans issued between June 2007 and December 2015.

From the whole data set we have, we chose only loans issued between January 2009 and December 2012 with 36-months duration. We focus on this period because of the following reasons. First, the default rate (16.10%) of loans issued before January 2009 is higher than the default rate (12.49%) of loans issued between January 2009 and December 2012. We found that this difference is highly statistically significant (two-sided t-test, $p < 0.001$). This difference might be caused by the financial crisis in 2007/2008 which hit hard many US households. In order to avoid a structural break in our data set, we decided to use only observations after 2008. Moreover, some of the loans issued before the SEC regulation in 2008 came with different loan conditions. Furthermore, Lending Club published less information about these loans. We have

not included loans issued after December 2012 as their maturity has not yet been reached. For a similar reason we have neither included loans with 60-month duration. Loans with 60-month duration were firstly introduced in 2010. Therefore, their maturity has not yet been reached in the data set we have.

In order to test our hypotheses, we need to classify loans in our data set as 'Fully Paid' or as 'Charged Off'. This classification will help us to differentiate between good (Fully Paid) and bad (Charged Off) loans. The loans in our data set, however, have six different statuses: Fully Paid, Charged Off, Current, Default, Late (31-120 days), Late (16-30 days) and In Grace Period. For the distribution of loan statuses in our data set see Part A in Table 2. A loan is marked as Fully Paid when the full principal with interest rates is paid back. A loan with status Charged Off is a loan where a borrower defaulted on the loan and the loan will never be paid back in full amount. Even though we have chosen our dataset's time span so that all loans are supposed to have already reached their maturity, there are still some loans which have not been completely paid back or charged off. This is usually caused by some delayed installments in the credit life span. Delayed installments extend the whole maturity of a loan. These loans have statuses Current, In Grace Period, Late (16-30 days), Late (31-120 days) or Defaulted. There are 33 loans with status Current. These loans are currently being paid back. We do not include loans with status Current into our analyses, because we do not know whether they will or will not be paid back. Furthermore, there are similarly six loans with status In Grace Period and 6 loans with status Late (16-30 days). In Grace Period means that a loan installment is delayed by at most 15 days. A loan with status Late (16-30 days) has a delayed installment between 16-30 days. We do not consider loans with statuses In Grace Period and Late (16-30 days) as Charged off, because these loans are not delayed by more than 30 days. We believe that they might be paid back. On the other hand, we do not say that these loans will be paid back. Therefore, we do not include loans with status In Grace Period or Late (16-30 days) at all. According to the Lending Club statistics, 75% of loans with status Late (31-120 days) are charged off.[10] We believe that this percentage is even higher for loans with installment delayed by 90 or more days. There are 81 loans with status Late (31-120 days) and 46 of them are delayed by more than 90 days. We consider all these 46 loans as Charged Off. Loans with status Default have delayed installment by more than 120 days. We consider all 12 loans with status Default as Charged Off. The proportion of Fully Paid and Charged Off loans are shown in Part B of Table 2. The number of Charged Off loans in Part B includes all Charged Off loans from Part A, 46 loans with status Late (31-120 days) and 12 Default loans.

[10]Source: https://www.lendingclub.com/info/demand-and-credit-profile.action (accessed May 1, 2016)

Table 2: Distribution of Loan Statuses

Part A
Initial data set distribution

Loan Status	**# of loans**
- Fully Paid	61,836
- Charged Off	8,779
- Current	33
- In Grace Period	6
- Late (16-30 days)	6
- Late (31-120 days)	81 (46)
- Default	12 (12)
Total number	**70,753**

Part B
Final data set distribution

Loan Status	**# of loans**
- Fully Paid	61,836
- Charged Off	8,837
Total number	**70,673**

Note: The number in parentheses of loans with status Late (31-120 days) or Default denote how many of these loans are considered as Charged Off in Part B.

In order to test our hypotheses, we distinguish between the following loan risk classes. A-graded loans belong to the *Low-Risk Class*. The *Medium-Risk Class* consists of B-graded loans. C-graded loans are in the *High-Risk Class*. Loans graded with letters D, E, F and G are aggregated to the *Very High-Risk Class* in order to make the classes somewhat comparable in terms of the number of observations. Table 3, part A, provides an overview of the four loan classes and their corresponding loan grades, average default rates and average *FICO scores*. For the composition of the *Very High-Risk Class* see part B in Table 3. Loans in the *Very High-Risk Class* are pretty similar in terms of *FICO score*. The difference between the average *FICO score* of the best loan grade D and the average *FICO score* of the worst loan grade G is only 5 points. Moreover, D-, E- and F-graded loans are also fairly similar in terms of the default rate. The default rate of G-graded loans is above the default rate of the remaining loan grades in the *Very High-Risk Class*. However, as there are only 76 G-graded loans, it would not be useful to create a separate group for these loans. Therefore, we added G-graded loans to the same class as D-, E- and F-graded loans.

4.3 Variables of interest

There are 78 variables in the data set provided by Lending Club.[11] Not all are of interest for us as some do not include any values (such as *Personal Finance Inquiries* and *Finance Trades*) or do not contain useful information for our purposes (like *Loan URL* and *Loan ID*).

[11] We have downloaded the data from the download data section at the Lending Club website. Moreover, the download data section's Data Dictionary provides variable descriptions. Source: https://www.lendingclub.com/info/download-data.action (accessed April 30, 2016)

Table 3: Loan Risk Classes

Part A: Overview of Loan Risk Classes

Type of class	Loan grade	# of loans	Default rate	FICO score
Low-Risk Class	A	20,041	6.6 %	750
Medium-Risk Class	B	25,539	11.8 %	707
High-Risk Class	C	15,117	16.5 %	687
Very High-Risk Class	D, E, F, G	9,976	20.1 %	677
All Loan Classes		**70,673**	**12.5 %**	**710**

Part B: Composition of High-Risk Class

Loan grade	# of loans	Default rate	FICO score
D	8,045	19.7 %	677
E	1,569	21.2 %	675
F	286	22.0 %	673
G	76	30.2 %	672
High-Risk Class	**9,976**	**20.0 %**	**677**

Our variables of interest can be divided into two sources of information origin. The first source is the borrower's self-reported information. Borrower's self-reported information are *Annual Income*, *Housing Situation*, *Length of Employment*, *Loan Amount*, *Loan Purpose*, and *Loan Description*. The second source of information origin is the borrower's credit file provided by one of three national credit bureaus in the USA. We choose the following variables from a borrower's credit file: *Debt−to−Income*, *Delinquency in Past 2 Years*, *Date of First Credit Line*, *Inquiries in Past 6 Months*, *Months since Last Delinquency*, *Months since Last Record*, *Open Credit Lines*, and *Revolving Credit Utilization*. The description of our variables is included in Table 6 in Appendix B.

We modified two variables from the original data set. The first variable is *Loan Description*. It is provided by a borrower when applying for a loan. There are many ways to use *Loan Description* as an independent variable that might be the predictor for borrowers' default. For example, Carmichael [2014] extracted two dummy variables from *Loan Description*: *Borrower's Self-Claimed Creditworthiness* and *Description Lacking Full Stop*. He found that both variables are significant for default prediction. Our approach is different from Carmichael [2014]. We count the number of characters in *Loan Description* and call this new variable *Number of Characters*. The second variable of interest from the original data set that was modified is *Date of First Credit Line*. It is a variable in the form of 'month-year' and represents the reported date of the first open credit line. We transformed this variable into the number of years since the first

reported credit line was opened. The name of this new variable is *Length of Credit History*.

Thus, our variables are fairly similar to variables used by Emekter et al. [2015], Carmichael [2014] and Serrano-Cinca et al. [2015]. Unlike these papers though, we do not include information about the *Loan Subgrade* and the *FICO Score*. Furthermore, we neither include the *Loan Grade* nor the *Interest Rate*. All these variables are highly correlated, because *Loan Subgrade*, which is more specific than *Loan Grade*, is largely based on the *FICO score*. The interest rate is then assigned based on the *Loan Subgrade*. More importantly, we do not need to include these variables, because we analyze our data within given loan risk classes.

4.4 Descriptive statistics

Table 7 in Appendix B contains the correlation matrix table of all non-categorical variables. The correlation matrix is based on our full data set of 70,673 observations. The highest correlation (0.33) is between *Debt-to-Income* and *Open Credit Lines*. The second largest correlation, which is 0.29, is between *Annual Income* and *Loan Amount*. All correlations between *Default* and other variables are less than 0.1. The most correlated variable with *Default* is *Revolving Credit Utilization* with a correlation of 0.08.

Table 8 in Appendix B contains descriptive statistics of our full data set. There are 82 missing values of *Revolving Credit Utilization* and 2,538 missing values of *Length of Employment*. We have excluded all 82 observations with missing values of *Revolving Credit Utilization* from our data set. Excluding 2,538 observations with missing values of *Length of Employment* from our dataset would mean a significant loss of information for our hypotheses testing. However, we have found that the *Length of Employment* is not a significant determinant of borrowers' default. This finding allows us to exclude the *Length of Employment* from our further analysis.

The maximum value of *Annual Income* is USD 7,141,778. It appears suspicious that a borrower with a self-reported annual income of USD 7,141,778 would ask for a loan of USD 14,825. Overall, there are 15 observations in our data set with a self-reported income exceeding USD 1,000,000 and we have decided to exclude these outliers. Thus our final data set for the remaining analyses includes 70,579 observations.

Table 9 in Appendix B presents mean values of the non-categorical variables, in particular loan classes. Interestingly, the highest mean of *Annual Income* is in the *Very High-Risk Class*. Furthermore, borrowers from the *Very High-Risk Class* wrote, on average, the longest loan descriptions. They might be afraid that their loan will not be funded because of their inferior credit grade. Therefore, they might try to provide a sound explanation of the loan need to their potential funders. Concerning *Loan Amount*, *Delinquency in past 2 Years*, *Inquiries in Past*

6 Months, *Months since Last Delinquency* and *Months since Last Record*, *Open Credit Lines* and *Revolving Credit Utilization* variables, we can observe a rising trend of variable mean values from *Low-Risk Class* to *Very High-Risk Class*. The only variable with a declining trend of its mean value is the *Length of Credit History*.

Finally, we look at the default statistics of our categorical variables. Table 10 in Appendix B contains the *Loan Purpose* default statistics. The trend of the default rate is clearly rising with the riskiness of a given loan class - starting with a default rate of 6.59% in the *Low-Risk Class* and ending with 20.06% in the *Very High-Risk Class*. The two most frequent loan purposes are *Debt Consolidation* (51.94% of all loans) and *Credit Card* (18.28%). Furthermore, the purposes *Car* and *Major Purchase* have the smallest default rates across all classes. On the other hand, loans with purpose *Small Business* or *Renewable Energy* have the highest default rate in the *All Classes* category.[12] It is interesting to observe how default rates of given *Loan Purposes* change in particular loan classes. For example, loans with the purpose *Moving* have a higher default rate in the *Medium-Risk Class* (17.97%) than in the *High-Risk Class* (14.74%). Similar examples are loans with *Home Improvement*, *Vacation* or *Car* purpose. Table 11 in Appendix B contains the *Home Situation* default statistics. As expected and similarly to the *Loan Purpose* the default rate is rising with the riskiness of a given loan class. The most frequent *Home Situations* are *Rent* (48.80% of all loans) and *Mortgage* (42.86%). The frequency of the *Home Situation Other* (0.16% of all loans) and *No Information* (0.05%) are negligible.

5 Results

We generally use binary logistic regression specifications to analyze the determinants of borrowers' default.[13] We use backward stepwise elimination to find the most suitable model specification, that is, we start with a full model including all 13 variables of interest. We then drop every variable with a p-value higher than 0.1 starting with the variable with the highest p-value. Backward stepwise elimination is sometimes criticized for producing models which do not fit the data well. Critics of this approach argue that other models might dominate the model achieved by backward stepwise elimination in terms of the Akaike information criterion (AIC), a measurement of relative model quality for a given data set. As a robustness check, we have run additional regressions which employ an automated selection of the best model with AIC as criterion. All of our specifications reached by backward stepwise elimination are the

[12] We do not further comment loans with *Renewable Energy* purpose, because they make up only a small percentage (0.20%) of all loans. The same applies to the loans with purpose *Education* (0.34%).

[13] All statistical analyses are performed using the software *R* (version 3.2.3) with its integrated development environment called *RStudio*. We use the *glm* function of the family *binomial*.

same as the specifications chosen by using AIC as selection criterion.

We first run a logistic regression on the full data set (*All Classes*) because of two reasons. The first reason is that we want to compare our *All Classes* findings with results of Carmichael [2014], Emekter et al. [2015] and Serrano-Cinca et al. [2015]. The second reason is that it allows us to highlight the differences between our regression results from given loan classes and the regression findings based on the whole data set.

Results from the *All Classes* regression are in Table 4. The coefficients of *Loan Amount*, *Debt-to-Income*, *Delinquency in Past 2 Years*, *Inquiries in Past 6 Months*, *Revolving Credit Utilization*, *Months since Last Record* are all positive and highly significant. The coefficients of *Annual Income*, *Number of Characters* and *Lenght of Credit History* are all negative and highly significant. The variable *Open Credit Lines* is not significant. The loan purposes *Car*, *Credit Card*, *Debt Consolidation*, *Home Improvement*, *Major Purchase*, and *Wedding* are negatively correlated with loan default. The loan purposes *Renewable Energy*, and *Small Business* are positively correlated with loan default. Home ownership (statuses *Own* and *Mortgage*) is negatively correlated with loan default.

We proceed with regressions for the four loan risk classes, see also Table 4. Results for the *Low Risk* and *Medium Risk* classes only differ slightly from the *All Classes* results. In both the *Lenght of Credit History* and the loan purpose *Home Improvement* are not significant anymore. *Months since Last Record* is not significant in the *Low Risk* class, while it is significant in the *Medium Risk* class. *Delinquency in Past 2 Years* is not significant in the *Medium Risk* class, while it is highly significant in the *Low Risk* class. In the *High Risk* and *Very High Risk* classes, the *Number of Characters* are not significant anymore as well as the loan purposes *Car*, *Debt Consolidation*, *Home Improvement* and *Renewable Energy*. The *Lenght of Credit History* is not significant in the *High Risk* class but it is significant in the *Very High Risk* class. The loan purpose *Major Purchase* is not significant anymore in the *Very High Risk* class.

Table 4: Regressions results for all classes and each class separately

	All Classes		Low-Risk Class		Medium-Risk Class		High-Risk Class		Very High-Risk Class	
Variable	**Coefficients**	**Std. Errors**	**Coefficients**	**Std. Errors**	**Coefficients**	**Std. Errors**	**Coefficients**	**Std. Errors**	**Coefficients**	**Std. Errors**
Intercept	-2.144e+00***	6.231e-02	-2.276e+00***	1.298e-01	-1.782e+00***	1.049e-01	-1.389e+00***	1.057e-01	-1.099e+00***	1.184e-01
Annual Income	-8.136e-06***	4.445e-07	-1.191e-05***	1.170e-06	-7.856e-06***	7.588e-07	-6.892e-06***	8.413e-07	-4.664e-06***	8.006e-07
HS: None	3.625e-01	4.531e-01	-9.928e+00	1.439e+02	1.281e+00*	6.192e-01	-2.796e-01	1.066e+00	not significant	
HS: Other	4.792e-01*	2.434e-01	7.905e-01	7.726e-01	8.527e-01*	3.857e-01	-6.056e-01	6.128e-01	not significant	
HS: Own	-2.450e-02***	4.299e-02	-1.052e-01	1.053e-01	-8.105e-02	7.344e-02	1.544e-01	8.493e-02	not significant	
HS: Mortgage	-1.740e-01***	2.655e-02	-2.235e-01***	6.524e-02	-1.185e-01**	4.458e-02	-1.581e-01**	4.967e-02	not significant	
Loan Amount	1.867e-05***	2.000e-06	1.093e-05*	5.481e-06	1.231e-05***	3.637e-06	-1.147e-05**	3.694e-06	8.051e-06*	3.744e-06
Number of Characters	-1.954e-04***	4.424e-05	-4.222e-04**	1.287e-04	-4.736e-04***	8.823e-05	not significant		not significant	
LP: Car	-3.766e-01***	9.473e-02	-2.999e-01	1.693e-01	-2.521e-01	1.637e-01	-3.147e-01	2.075e-01	-5.691e-01*	2.767e-01
LP: Credit Card	-5.386e-01***	5.034e-02	-6.680e-01***	1.203e-01	-4.739e-01***	8.619e-02	-3.614e-01***	9.738e-02	-4.256e-01***	1.119e-01
LP: Debt Consolidation	-2.385e-01***	4.373e-02	-4.126e-01***	1.009e-01	-1.823e-01*	7.646e-02	-1.223e-01	8.576e-02	-1.387e-01	9.310e-02
LP: Education	1.015e-01	1.885e-01	-6.597e-01	6.049e-01	4.358e-01	3.085e-01	-4.047e-01	3.672e-01	4.045e-01	3.914e-01
LP: Home Improvement	-1.931e-01***	6.617e-02	-3.305e-01*	1.429e-01	-1.265e-01	1.143e-01	-1.247e-02	1.283e-01	-2.252e-01	1.574e-01
LP: House	4.084e-03	1.314e-01	-1.700e-01	2.865e-01	8.798e-03	2.326e-01	1.649e-01	2.715e-01	2.558e-01	2.868e-01
LP: Major Purchase	-4.570e-01***	7.973e-02	-5.694e-01***	1.627e-01	-4.732e-01***	1.434e-01	-4.343e-01**	1.420e-01	-1.543e-01	1.788e-01
LP: Medical	1.070e-01	9.553e-02	-8.276e-02	2.115e-01	3.001e-01	1.628e-01	-3.725e-02	1.891e-01	1.153e-01	2.176e-01
LP: Moving	8.584e-02	1.051e-01	1.551e-01	2.249e-01	3.415e-01*	1.686e-01	-3.755e-01	2.399e-01	5.225e-02	2.369e-01
LP: Renewable Energy	5.345e-01*	2.176e-01	8.869e-01*	4.278e-01	7.853e-01*	3.440e-01	1.228e-01	5.109e-01	2.310e-01	5.351e-01
LP: Small Business	5.470e-01***	6.788e-02	5.964e-01***	1.475e-01	5.621e-01***	1.235e-01	4.343e-01**	1.420e-01	3.945e-01**	1.336e-01
LP: Vacation	3.449e-02	1.217e-01	4.863e-01*	2.182e-01	-1.836e-02	2.129e-01	-1.800e-01	2.482e-01	-4.309e-01	3.408e-01
LP: Wedding	-4.035e-01***	1.007e-01	-5.484e-01*	2.380e-01	-2.278e-01	1.721e-01	-4.207e-01*	1.917e-01	-5.569e-01*	2.251e-01
Debt-to-Income	9.643e-01***	1.667e-01	1.441e+00***	4.127e-01	7.793e-01**	2.821e-01	2.028e+00***	3.067e-01	1.113e+00**	3.609e-01
Delinquency in Past 2 Years	1.075e-01***	1.825e-01	2.691e-01***	7.154e-02	not significant		-8.452e-02*	3.477e-02	not significant	
Length of Credit History	-8.094e-03***	1.880e-03	not significant		not significant		not significant		-1.393e-02***	4.223e-03
Inquiries in Past 6 Months	2.114e-01***	1.091e-02	2.092e-01***	2.774e-02	1.487e-01***	1.939e-02	1.098e-01***	2.161e-02	1.146e-01***	2.436e-02
Months since Last Delinquency	1.523e-03*	5.176e-04	not significant		-1.649e-03*	8.908e-04	-2.391e-03*	9.689e-04	not significant	
Months since Last Record	3.724e-03***	5.851e-04	not significant		2.328e-03*	9.872e-04	1.968e-03*	9.898e-04	2.656e-03*	1.242e-03
Open Credit Lines	not significant		not significant		not significant		not significant		not significant	
Revolving Line Utilization	1.045e+00***	4.883e-02	1.067e+00***	1.330e-01	3.620e-01***	9.486e-02	not significant		not significant	
Akaike Information Criterion	**51 294**		**9 354**		**18 188**		**13 433**		**9 882**	

*Note: Significance code is '***' for 0, '**' for 0.001 and '*' for 0.01.*

Revolving Credit Utilization has been found to be a significant predictor for borrowers' default in all related studies [Carmichael, 2014, Emekter et al., 2015, Serrano-Cinca et al., 2015] as well as in our *All Classes* data. However, it is only significant in our *Low-Risk* and *Medium-Risk Classes*. It is not a significant determinant in the *High-Risk* and *Very High-Risk Class*.

Result 1: *Revolving Credit Utilization is a significant determinant of borrowers' default only in low loan risk classes.*

The *Debt-to-Income* ratio is significant in all loan classes. Thus, we cannot reject hypothesis 2. In fact, the *Debt-to-Income* ratios for defaulted/non-defaulted loans have almost identical values across risk classes.

Result 2: The *Debt-to-Income ratio is a significant determinant of borrowers' default in all loan risk classes.*

The current *Housing Situation* is a significant determinant of default in *All Classes*, as well as in the *Low-Risk*, *Medium-Risk* and *High Risk* classes. It is, however, not significant in the *Very High-Risk Class*. Defaulting on a loan when having a mortgage on a house would mean the loss of the house. Therefore, there might be a higher motivation for borrowers to avoid default when having the mortgage than living in a rented home. One of the possibilities to avoid default is to take a further loan. Borrowers from the *Very High-Risk Class* may not have such an opportunity which might explain that there is no effect of the *Current Housing Situation*.

Result 3: *Home ownership is not a significant determinant of borrowers' default in the highest loan risk class.*

Overall, creditworthy borrowers write, on average, 169 characters in their loan descriptions compared to 157 characters in loan descriptions of defaulted loans. This difference is highly significant ($p < 0.001$). Moreover, it is interesting to observe that borrowers in the *Very High-Risk Class* write, on average, the most characters in their *Loan Description* compared to borrowers from other classes. Borrowers from the *Very High-Risk Class* might feel that their *Loan Description* must be comprehensive in order to get funding with a risky loan grade. However, the *Number of Characters* are neither significant in the *Very High-Risk Class* nor in the *High-Risk Class*, while they are in low risk classes. We can, therefore, reject hypothesis 4.

Result 4: *In low loan risk classes, creditworthy borrowers write, on average, a longer Loan Description than borrowers who defaulted.*

We can only partially reject hypothesis 5, because some loan purposes are significant in all loan classes. It seems that a loan used for *Credit Card* consolidation has a significantly higher chance to be paid back even in the *Very High-Risk Class*, while loans used for a *Small Business* generally bear a higher risk of default independently of the associated risk class. For example, the default rates of loans with purpose *Small Business* are twice as high as default rates of loans with *Car* or *Wedding* as the purpose.

Results 5: *The loan purposes Credit Card and Small Business are significant determinants of borrowers' default in all loan risk classes.*

The *Length of Credit History* is negatively correlated with loan default in our *All Classes* regression results. This finding is in line with Carmichael [2014] and Serrano-Cinca et al. [2015]'s results. However, it is only supported in the *Very High-Risk Class*. The *Length of Credit History* is not a significant determinant of default in the *Low-Risk*, *Medium-Risk* and *High-Risk* classes. This finding is in line with our hypothesis 6. It seems that experience with loans in the *Very High-Risk Class* is of advantage as people get used to live close to their credit limits. For example, a young man without any previous credit experiences classified to be in the *Very High-Risk Class*, also without any financial buffer, can easily overdraw his credit. This might cause a default because of insufficient credit experience and a lack of possibilities of obtaining an additional loan.

Result 6: *The Length of Credit History is a significant determinant of borrowers' default only in the High-Risk Class.*

5.1 Discussion

In our full data set, all variables of interest turn out to be significant determinants of default except the variable *Open Credit Lines*. Table 5 provides a comparison of our *All Classes* findings and the previously mentioned studies. Generally, discrepancies of results could be due to the fact that our data avoids the structural break of loan defaults possibly caused by the 2007/08

financial crisis.[14] See Table 1 for differences of the data and methodology. The only difference to Carmichael [2014]'s results is that *Debt-to-Income* is not a significant predictor of borrowers' default in his study. This difference might be caused by the fact that Carmichael [2014] used loans with status 'current' in his analyses. Comparing our results to Serrano-Cinca et al. [2015], two discrepancies are worth to note. *Loan Amount* is not significant in their study but in ours and *Open Credit Lines* is significant in theirs but not in ours. Finally, our *All Classes* results are quite different from Emekter et al. [2015]'s results. Besides differences in the time frame of the data set, Emekter et al. [2015] include the *Loan Credit Grade* and *FICO score* as explanatory variables in their regression. A high correlation between *FICO score* and other variables of interest is to be expected, because the *FICO score* is computed based on these values. The same may apply to the *Loan Grade*.

[14] We show that loans issued before 2009 have significantly higher default rates than loans issued between 2009 and 2012.

Table 5: Comparison of Findings

				Our Classes				
Variable / Paper	**Serrano-Cinca et al. (2015)**	**Carmichael (2014)**	**Emekter et al. (2015)**	**All**	**Low-Risk**	**Medium-Risk**	**High-Risk**	**Very High-Risk**
Annual Income	x	x		x	x	x	x	x
Loan Amount		x		x	x	x	x	x
Number of Characters	not measured	not measured	not measured	x	x	x		
Debt-to-Income	x		x	x	x	x	x	x
Delinquency in Past 2 Years	x	not measured		x	x		x	
Length of Credit History	x	x		x				x
Inquiries in Past 6 Months	x	x		x	x	x	x	x
Months since Last Delinquency		x		x		x	x	
Months since Last Record	not measured	not measured		x		x	x	x
Open Credit Lines	x	not measured						
Revolving Credit Utilization	x	x	x	x	x	x		
Loan Purpose	x	x		x	x	x	x	x
Home Situation	x	not measured	not measured	x	x	x	x	

Note: The mark ***x*** *denotes that a given variable was found to be a significant determinant of borrowers' default.*

Overall, we find the following determinants of borrowers' default which are significant in *All Classes* as well as in the loan classes separately: *Annual Income*, *Debt-to-Income*, *Inquiries in Past 2 Years* and the loan purposes *Credit Card* and *Small Business*. The significance of other variables varies class by class. *Revolving Credit Utilization*, *Number of Characters* or the *Lenght of Credit History* are significant for *All Classes* but not in given loan classes. Only *Open Credit Lines* is neither significant in any class nor on the full data set.

Finally, we address to what extent these insights are taken into account by lenders, when they decide whether to fund a loan or not. For this purpose, we draw on evidence from existing studies who analyze the funding success of P2P loans. According to both Herzenstein et al. [2011] and Zhang and Liu [2012] loans have a higher chance to attract funding, the lower the debt-to-income ratio is and the better the credit grade is. Moreover, Zhang and Liu [2012] find a positive correlation between funding success and the amount requested as well as whether the borrower's home is owned. Our analysis reassures the positive attitute of lenders towards a borrower's debt-to-income ratio. Other characteristics warrant more caution. We find that the home ownership status is only a good indicator of a loan getting paid back if the loan is not from the highest risk class.

Our results provide insights for potential P2P lenders, especially those who seek to strictly maximize their profit. As mentioned in section 3 investing in the *Very High-Risk Class* at Lending Club historically yields the highest net profit (after accounting for defaulted loans). Thus, investors whose primary goal is to achieve a high return on their investment will target the high risk segment and will try to optimize their loan portfolio choices. Our results contribute to a better understanding to what extent our existing knowledge about loan default determinants applies in this high risk segment. It seems that for high risk loans *Revolving Credit Utilization* or the *Home Situation* status are treacherous predictors of default. Instead, the mindful investor should target the *Length of Credit History*, *Inquiries in Past 2 Years*, *Annual Income*, the *Debt-to-Income* ratio and the loan purposes *Credit Card* or *Small Business* as reliable predictors of default.

6 Conclusion

P2P lending connects people in need for a loan with people willing to lend their money. The intermediation of credit is handled through more or less automated online platforms with very low transaction costs. The benefits of automation transform into lower interest rates for borrowers and higher interest earnings for lenders in comparison to traditional banks.

However, information asymmetries between borrowers and lenders remain a central issue

faced by P2P lending platforms. Credit scoring techniques are employed to address this. They assign a credit grade to each loan based on the perceived risk of default. Riskier loans are associated with higher interest rates as higher interest rates serve as compensation for a potential loan default. Besides the credit grade and interest rate, P2P lending platforms usually provide a prospective lender with a large amount of information about a loan's and borrower's characteristics.

Previous research [Emekter et al., 2015, Carmichael, 2014, Serrano-Cinca et al., 2015] identifed some of the borrower's and loan's information as useful determinants for borrowers' default. We hypothesize that the significance of default determining variables might not be the same in different loan risk classes. In other words, some variables are only significant default determinants in specific loan classes.

While results on our full data set are largely in line with findings of previous studies, our set of separate regressions for each loan risk class identifies only *Annual Income*, *Debt-to-Income*, *Inquiries in Past 2 Years* and the loan purposes *Credit Card* and *Small Business* as significant determinants of loan default in all loan risk classes. *Revolving Credit Utilization*, *Delinquency in Past 2 Years* and *Number of Characters* are only significant for low loan risk classes. *Lenght of Credit History* is only significant for high loan risk classes.

Our analysis confirms that loan/borrower characteristics can indeed be used to predict a loan's default chances. However, since default determinants depend on the loan's risk class, caution is warranted. What seems to be a good predictor of loan default based on overall data may not be reliable in the highest loan risk class. This is relevant since the high risk segment is most attractive to some lenders due to the highest returns that can be reached.

References

Hussein Abdou and John Pointon. Credit scoring, statistical technique and evaluation criteria: a review of the literature. *Intelligent systems in accounting, finance and management*, 18(2-3): 59–88, 2011. ISSN 1055615X. doi: 10.1002/isaf. URL http://onlinelibrary.wiley.com/doi/10.1002/isaf.329/full.

Don Carmichael. Modeling default for peer-to-peer loans. *Available at SSRN: http://ssrn.com/abstract=2529240*, 2014. ISSN 1556-5068. doi: 10.2139/ssrn.2529240.

Luke Deer, Jackson Mi, and Yu Yuxin. The rise of peer-to-peer lending in China: An overview and survey case study. *The Association of Chartered Certified Accountants Report*, 2015. URL http://www.accaglobal.com/uk/en/technical-activities/technical-resources-search/2015/december/p2p-lending.html.

Deloitte. Banking disrupted: How technology is threatening the traditional European retail banking model. Technical report, 2014.

Riza Emekter, Yanbin Tu, Benjamas Jirasakuldech, and Min Lu. Evaluating credit risk and loan performance in online Peer-to-Peer (P2P) lending. *Applied Economics*, 47(1):54–70, 2015. ISSN 0003-6846. doi: 10.1080/00036846.2014.962222. URL http://www.tandfonline.com/doi/abs/10.1080/00036846.2014.962222.

Seth Freedman and Ginger Zhe Jin. The Information value of online social networks: lessions from peer-to-peer lending. *NBER Working paper. Available at: http://www.nber.org/papers/w19820*, 2014. doi: 10.3386/w19820.

Michal Herzenstein, Utpal M. Dholakia, and Rick L. Andrews. Strategic Herding Behavior in Peer-to-Peer Loan Auctions. *Journal of Interactive Marketing*, 25(1):27–36, 2011. ISSN 10949968. doi: 10.1016/j.intmar.2010.07.001. URL http://dx.doi.org/10.1016/j.intmar.2010.07.001.

R. Iyer, A.I. Khwaja, E.F.P. Luttmer, and K. Shue. Screening in New Credit Markets Can Individual Lenders Infer Borrower Creditworthiness in Peer-to-Peer Lending? *Harvard Kennedy School Faculty Research Working Papers Series*, 2009.

Mingfeng Lin, Nagpurnanand Prabhala, and Siva Viswanathan. Judging borrowers by the company they keep: friendship networks and information asymmetry in online peer-to-peer lending. *Management Science*, 59(1):17–35, 2013. ISSN 0025-1909. doi: 10.1287/mnsc.1120.1560.

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

E Namvar. An Introduction to Peer to Peer Loans as Investments. *Journal of Investment Management*, 12(1):1–18, 2013. URL http://dx.doi.org/10.2139/ssrn.2227181.

Adam Nowak, Amanda Ross, and Christopher Yencha. Small Business Borrowing and Peer-to-Peer Lending: Evidence from Lending Club. *West Virginia University Working Paper*, (No. 15-28), 2015.

Carlos Serrano-Cinca, Begoña Gutiérrez-Nieto, and Luz López-Palacios. Determinants of Default in P2P Lending. *PLOS ONE*, 10(10):1–22, oct 2015. ISSN 1932-6203. doi: 10.1371/journal.pone.0139427. URL http://dx.plos.org/10.1371/journal.pone.0139427.

Jiazhuo Wang, Hongwei Xu, and Jun Ma. *Financing the Underfinanced*. Springer Berlin Heidelberg, ISBN: 978-3-662-46524-0, 2015. ISBN 978-3-662-46524-0. doi: 10.1007/978-3-662-46525-7. URL http://link.springer.com/10.1007/978-3-662-46525-7.

Robert Wardrop, Bryan Zhang, Raghavendra Rau, and Mia Gray. The European Alternative Finance Benchmarking Report. *Universtiy of Cambridge Report*, 2015. URL http://www.jbs.cam.ac.uk/index.php?id=6481{#}.VTOtICGqpBd.

G Weiss, K Pelger, and A Horsch. Mitigating Adverse Selection in P2P Lending - Empirical Evidence from Prosper.com. *Available at SSRN: http://ssrn.com/abstract=1650774*, 2010. ISSN 1556-5068. doi: http://dx.doi.org/10.2139/ssrn.1650774. URL http://papers.ssrn.com/sol3/papers.cfm?abstract{_}id=1650774.

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

J. Zhang and P. Liu. Rational Herding in Microloan Markets. *Management Science*, 58(5):892–912, 2012. ISSN 0025-1909. doi: 10.1287/mnsc.1110.1459.

Appendix

A: Variables of Interest

Table 6: Variables of Interest and their Descriptions

Borrower's self-reported information

Name of variable	Description of variable
Annual Income	The self-reported annual income provided by the borrower during registration.
Housing Situation	The home ownership status provided by the borrower during registration. Our values are: RENT, OWN, MORTGAGE, OTHER.
Length of Employment	Employment length in years. Possible values are between 0 and 10 where 0 means less than one year and 10 means ten or more years.
Loan Amount	The listed amount of the loan applied for by the borrower.
Loan Purpose	A category provided by the borrower for the loan request.
Number of Characters	The number of characters used by borrower for loan description.

Information from borrower's credit file

Name of variable	Description of variable
Debt-to-Income	A ratio calculated using the borrower's total monthly debt payments on the total debt obligations, excluding mortgage and the requested LC loan, divided by the borrower's self-reported monthly income.
Delinquency in Past 2 Years	The number of 30+ days past-due incidences of delinquency in the borrower's credit file for the past 2 years.
Length of Credit History	The number of years since the first reported credit line was opened.
Inquiries in Past 6 Months	The number of inquiries in past 6 months (excluding auto and mortgage inquiries).
Months since Last Delinquency	The number of months since the borrower's last delinquency.
Months since Last Record	The number of months since the last public record.
Open Credit Lines	The number of open credit lines in the borrower's credit file.
Revolving Credit Utilization	Revolving credit line utilization rate or the amount of credit the borrower is using relative to all available revolving credit.

B: Descriptive Statistics

Table 7: Correlation Matrix

	Default	Loan Amount	Length of Employment	Annual Income	Number of Characters	Debt-to-Income	Delinquency in Past 2 Years	Lenght of Credit History	Inquiries in Past 6 Months	Months since Last Deliquency	Months since Last Record	Open Credit Lines	Revolving Credit Utilization
Default	1	-0.01	0.00	-0.05	-0.01	0.06	0.01	-0.04	0.06	0.01	0.02	0.00	0.08
Loan Amount	-0.01	1	0.12	0.29	0.05	0.04	-0.01	0.16	-0.01	-0.01	-0.06	0.19	0.07
Length of Employment	0.00	0.12	1	0.09	-0.09	0.07	0.04	0.25	0.00	0.06	0.04	0.08	0.05
Annual Income	-0.05	0.29	0.09	1	0.00	-0.15	0.04	0.17	0.05	0.03	-0.01	0.14	0.00
Number of Characters	-0.01	0.05	-0.09	0.00	1	-0.05	-0.02	-0.01	0.00	-0.03	0.01	-0.02	-0.04
Debt-to-Income	0.06	0.04	0.07	-0.15	-0.05	1	0.01	0.03	0.00	0.02	-0.02	0.33	0.28
Delinquency in Past 2 Years	0.01	-0.01	0.04	0.04	-0.02	0.01	1	0.09	0.01	-0.01	-0.01	0.06	0.00
Length of Credit History	-0.04	0.16	0.25	0.17	-0.01	0.03	0.09	1	0.01	0.11	0.05	0.17	-0.04
Inquiries in Past 6 Months	0.06	-0.01	0.00	0.05	0.00	0.00	0.01	0.01	1	0.02	0.03	0.10	-0.09
Months since Last Deliquency	0.01	-0.01	0.06	0.03	-0.03	0.02	-0.01	0.11	0.02	1	0.02	0.07	0.05
Months since Last Record	0.02	-0.06	0.04	-0.01	0.01	-0.02	-0.01	0.05	0.03	0.02	1	-0.01	0.01
Open Credit Lines	0.00	0.19	0.08	0.14	-0.02	0.33	0.06	0.17	0.10	0.07	-0.01	1	-0.06
Revolving Credit Utilization	0.08	0.07	0.05	0.00	-0.04	0.28	0.00	-0.04	-0.09	0.05	0.01	-0.06	1

Table 8: Overall Descriptive Statistics

Variable / Statistic	**N**	**Mean**	**St. Dev.**	**Min**	**Median**	**Max**
Default	70 673	0.125	0.330	0.000	0.000	1.000
Loan Amount	70 673	10 888	6 878	1 000	10 000	35 000
Length of Employment	68 135	5.0	3.5	0.0	5.0	10.0
Annual Income	70 673	67 154	61 531	4 000	57 000	7 141 778
Number of Characters	70 673	167	281	0	71	3 853
Debt-to-Income	70 673	0.151	0.074	0.000	0.149	0.349
Delinquency in Past 2 Years	70 673	0.176	0.581	0.000	0.000	18.000
Length of Credit History	70 673	17.63	6.85	6.00	16.00	69.00
Inquiries in last 6 Months	70 673	0.812	1.013	0.000	0.000	8.000
Months since Last Deliquency	70 673	14.07	22.24	0.00	0.00	152.00
Months since Last Record	70 673	3.31	17.63	0.00	0.00	119.00
Open Credit Lines	70 673	9.96	4.44	1.00	9.00	49.00
Revolving Credit Utilization	70 591	0.537	0.263	0.000	0.563	1.044

Table 9: Mean Comparison Table

	All Classes		Low-Risk Class		Medium-Risk Class		Risk Class		High-Risk Class	
Mean of	**Default**	**Non-Default**	**Default**	**Non-Default**	**Default**	**Non-Default**	**Default**	**Non-Default**	**Default**	**Non-Default**
Annual Income	58 507	67 870	55 842	69 301	56 946	65 961	56 843	65 698	64 720	73 363
Employment Length	5.27	5.29	5.29	5.32	5.36	5.34	5.19	5.20	5.23	5.23
Loan Amount	10 798	10 900	9 295	10 017	10 289	10 720	10 315	10 820	13 172	13 624
Number of Characters	157	169	148	170	138	163	159	162	189	190
Debt-to-Income	0.163	0.150	0.151	0.134	0.164	0.157	0.168	0.158	0.161	0.155
Deliquency in Past 2 Years	0.198	0.173	0.080	0.060	0.155	0.167	0.224	0.262	0.309	0.313
Lenght of Credit History	16.99	17.72	18.69	19.04	17.15	17.44	16.59	16.82	16.10	16.83
Inquiries in Last 6 Months	0.977	0.788	0.783	0.635	0.816	0.694	1.207	1.088	1.063	0.941
Months since Last Deliquency	14.66	13.99	7.68	7.78	13.56	14.97	16.90	18.07	18.14	19.37
Months since Last Record	4.48	3.15	1.80	1.31	4.34	3.59	5.73	4.64	4.90	3.87
Open Credit Lines	9.97	9.96	9.68	9.83	9.80	9.89	10.13	10.08	10.22	10.27
Revolving Credit Utilization	0.591	0.529	0.387	0.342	0.566	0.561	0.637	0.636	0.707	0.708

Table 10: Loan Purposes

	All Classes		Low-Risk Class		Medium-Risk Class		Risk Class		High-Risk Class	
Loan Purpose	**Default Rate**	**% (#) of Loans**	**Default Rate**	**% (#) of Loans**	**Default Rate**	**% (#) of Loans**	**Default Rate**	**% (#) of Loans**	**Default Rate**	**% (#) of Loans**
Car	9.16 %	2.37 % (1 671)	6.05 %	4.29 % (860)	10.92 %	1.86 % (476)	14.55 %	1.46 % (220)	14.78 %	1.16 % (115)
Credit Card	10.10 %	18.28 % (12 900)	5.33 %	16.58 % (3 321)	9.39 %	20.45 % (5 220)	13.88 %	18.76 % (2 832)	15.91 %	15.38 % (1 527)
Debt Consolidation	13.14 %	51.94 % (36 659)	6.56 %	43.98 % (8 809)	12.05%	53.69 % (13 706)	16.93 %	55.85 % (8 430)	20.34 %	57.56 % (5 714)
Education	14.88 %	0.34 % (242)	4.62 %	0.32 % (65)	18.18 %	0.30 % (77)	13.04 %	0.46 % (69)	32.26 %	0.31 % (31)
Home Improvement	10.16 %	6.04 % (4 263)	5.08 %	8.84 % (1 771)	10.83 %	5.35 % (1 367)	17.45 %	4.78 % (722)	17.12 %	4.06 % (403)
House	13.90 %	0.78 % (554)	6.98 %	1.07 % (215)	13.26 %	0.71 % (181)	20.88 %	0.60 % (91)	28.36 %	0.67 % (67)
Major Purchase	8.43 %	3.95 % (2 789)	4.46 %	6.38 % (1 277)	8.71 %	3.19 % (815)	13.00 %	2.96 % (446)	19.52 %	2.53 % (251)
Medical	14.91 %	1.53% (1 080)	7.77 %	1.93 % (386)	17.22 %	1.30 % (331)	18.58 %	1.50 % (226)	23.36 %	1.38 % (137)
Moving	15.63 %	1.18% (832)	10.11 %	1.33 % (267)	17.97 %	1.16 % (295)	14.74 %	1.03 % (156)	23.68 %	1.15 % (114)
Other	14.48 %	7.46 % (5 263)	8.61 %	8.35 % (1 672)	13.78 %	6.79% (1 734)	18.88 %	6.98 % (1 054)	22.42 %	8.09 % (803)
Renewable Energy	20.86 %	0.20 % (139)	14.89 %	0.23 % (47)	24.49 %	0.19 % (49)	20.83 %	0.16 % (24)	26.32 %	0.19 % (19)
Small Business	20.74 %	3.13 % (2 208)	12.94 %	3.43 % (688)	20.44 %	2.47 % (631)	24.63 %	2.66 % (402)	28.95 %	4.91 % (487)
Vacation	14.26 %	0.90 % (638)	13.27 %	1.13 % (226)	13.36 %	0.85 % (217)	17.19 %	0.85 % (128)	14.93 %	0.67 % (67)
Wedding	9.92 %	1.90% (1 341)	5.15 %	2.13 % (427)	10.72 %	1.68 % (429)	12.97 %	1.94 % (293)	14.06 %	1.93 % (192)
Total	**12.50 %**	**100 % (70 579)**	**6.59 %**	**100 % (20 031)**	**11.81 %**	**100 % (25 528)**	**16.51 %**	**100 % (15 093)**	**20.06%**	**100 % (9 927)**

Table 11: Home Situations

	All Classses		Low-Risk Class		Medium-Risk Class		Risk Class		High-Risk Class	
Home Situation	**Default Rate**	**% (#) of Loans**	**Default Rate**	**% (#) of Loans**	**Default Rate**	**% (#) of Loans**	**Default Rate**	**% (#) of Loans**	**Default Rate**	**% (#) of Loans**
Mortgage	10.79 %	42.86 % (30 248)	5.50 %	51.42 % (10 299)	10.64 %	41.59 % (10 617)	15.29 %	37.55 % (5 667)	19.08 %	36.92 % (3 665)
No information	17.14 %	0.05 % (35)	0.00 %	0.02 % (5)	33.33 %	0.05 % (12)	11.11 %	0.06 % (9)	11.11 %	0.09 % (9)
Other	20.00 %	0.16 % (110)	12.50 %	0.08 % (16)	23.08 %	0.15 % (39)	10.00 %	0.20 % (30)	32.00 %	0.25 % (25)
Own	13.28 %	8.14 % (5 746)	7.40 %	8.23 % (1 648)	12.03 %	8.13 % (2 076)	17.42 %	8.08 % (1 220)	22.43 %	8.08% (802)
Rent	13.84 %	48.80 % (34 440)	7.81 %	40.25 % (8 063)	12.69 %	50.08 % (12 784)	17.25 %	54.11 % (8 167)	20.38 %	54.66 % (5 426)
Total	**12.50 %**	**100 % (70 579)**	**6.59 %**	**100% (20 031)**	**11.81 %**	**100 % (25 528)**	**16.51 %**	**100 % (15 093)**	**20.09 %**	**100% (9 927)**