Olga Fomicheva

DATA 698

Loan Predictions for Lending Club Platform

INTRODUCTION

Lending Club is s the world's largest peer-to-peer lending platform that enables borrowers to obtain loans and investors to purchase notes backed by payments made on loans. Lending Club receives applications from individuals or small institutions looking to borrow money, and evaluates the loan decision exclusively based on the information provided by the applicant. The company then assigns a rating of the loan, similar to how a rating agency such as Standard and Poor's or Moody's assigns a rating to a publicly traded security. Assigned rating determines the interest rate on the loan. Lending Club then makes the loan available on the marketplace, where investors are able to evaluate the loan before deciding to invest or not to invest.

Lending Club made the data related to loans that were issued since 2007 publicly available. It gives an investor the opportunity to see what loans were paid off and what loans were charged off or defaulted. An investor earns money when loan is fully paid off and loses money when loan is charged off. If an investor is able to predict loan creditability he can make a better investment decision. The goal of the project is to help an investor to make the right invest decision and determine the following:

1. Predict whether an application is funded or not.
2. Determine which factors impact loan funding success.
3. Calculate the probability of weather the load will be paid off or not.
4. Classify a loan as paid off or charged off.
5. Figure out how various loan features such as loaner income or credit rating affect loan creditability.

LITERATURE REVIEW

Peer-to-peer (P2P) lending platforms act as financial intermediaries that connect borrowers with lenders. Deloitte (2014) suggests that due to technological innovations such lending platforms issue loans with low intermediation costs and as a result pose a threat for traditional banks. The two major reasons for the rapid emergence of P2P lending platforms are low intermediation costs and credit rationing after the financial crisis in 2007-2008 (Mills, 2013). SEC (The Securities and Exchange Commission) required P2P lending companies to register their loans as securities and provide them through a bank.

Unsurprisingly, the popularity of P2P lending has grown rapidly. For instance, Lending Club, one of the biggest P2P lending platforms in the world, almost doubled the amount of issued loans from USD 4.4 billion in 2016 to USD 8.4 billion in 2017. The significant growth of P2P lending occurred in Europe as well as in China. The critical problem of general lending is information inequality between borrowers and lenders. It’s known that borrowers usually get more information about their creditworthiness than lenders. P2P lending platforms try to resolve the problem of information asymmetry. The platform uses credit scoring techniques to evaluate each loan and assign a risk grade. By analyzing risk grades potential lenders can try to predict whether a certain loan will be fully paid or not. Indeed, existing research that were performed by Emekter and Tu in 2015 and Carmichael in 2014 found a positive correlation between assigned risk grade and likelihood of a loan’s default. They also concluded that default rate depend on revolving credit utilization and the debt-to-income ratio.

This project aims to verify that significance of these default determinants depends on the loan’s risk grade. Thus, one of the goals of the project is to evaluate known determinants of borrowers’ default for each risk grade separately. Several studies determined what factors leads to the funding success of P2P loans. Majority of those studies used the data that was collected from Prosper which used to be the biggest P2P lending platform in the USA a decade ago. Prosper lending platform used many social features, such as a discussion forum and detailed borrowers' characteristics including their photos. Other studies (Lin & Prabhala, 2013 and Freedman & Jin, 2014) emphasize the importance of social relationships for funding success. The studies concluded that borrowers with better social ties are more likely to get their loans funded and to get a lower interest rate. Unfortunately, in 2018 Prosper had decided to remove social from its website. In 2014 Barasinska and Schäfer analyzed data from the German platform Smava and concluded that males and females are more likely to get funded. Moreover, another two researchers Herzenstein and Dholakia (2011) stated 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. Moreover, the studies found that the funding is negatively correlated with debt-to-income ratio while the funding is positively correlated with credit grade. Furthermore, they determined no relationship between home ownership and funding or the requested loan amount and funding.

The researches Zhang and Liu (2011) stated that lenders observe their peers' lending decisions and use this information to evaluate creditworthiness of borrowers. Also, they found that the funding is negatively correlated with debt-to-income ratio, while the credit grade, home owner status and the amount requested are positively correlated with funding.

Investing at P2P lending platforms is considered to be a risky activity, because the offered loans are not secured. To decrease the information inequality between lenders and borrowers, borrowers are required to provide some personal information, such as the loan's purpose or annual income. For instance, borrowers at Lending Club are obliged to provide detailed information about their credit history and themselves Lending Club use this information to evaluate the likelihood of borrowers' default and assign him a grade and an appropriate interest. It's believed that the better the grade the more likely is the borrower to repay his or her debt.

There are several researches (Freedman & Jin, 2014 and Iyer & Khwaja, 2015) studying how borrowers' characteristics impact borrowers' default based on data from Prosper. Three similar studies (Serrano-Cinca & Gutiérrez-Nieto, 2015 and Carmichael, 2014) analyzed the data collected from Lending Club agreed that credit grade assigned by Lending Club is the best predictor for borrowers' default. Furthermore, the studies concluded that revolving credit line utilization is another variable impacting the borrower's default rate. However, the studies didn't agree on features that can affect borrower's default. The discrepancy between the findings in the studies might be caused by three different factors. Firsts, the selection of variables that might have an impact on borrowers' default. For example, the researchers Emekter and Tu (2015) and Carmichael (2014) found out that the FICO score has an influence on default. While scientists Serrano-Cinca & Gutiérrez-Nieto (2015) did not select the FICO score as an independent variable in their study. Second, discrepancy might be caused by differences in classification of loan status or type of loan length or differences in time frames. For example, Emekter and Tu (2015) and Serrano-Cinca and Gutiérrez-Nieto (2015) used only 36-month loans. While Carmichael (2015) used both, 36 and 60-month loans. Third, discrepancy might be cause by research technique used. Carmichael (2015) used dynamic logistic regression to assess factors that determine what influence default rate in P2P lending whereas Serrano-Cinca and Gutiérrez-Nieto (2015) conducted their study with a combination of Cox regression and univariate means. Emekter and Tu (2015) chose binary logistic regression for their analysis.

DATA EXPLORATION

DATASET

Lending Club made available loan data at https://www.lendingclub.com/info/download-data.action. The information about these loans is updated daily, then monthly and then quarterly. Lending Club data set for the project was downloaded in February 2019. It contains information about 986,634 loans that were issued between June 2007 and January 2019. For the analysis I chose only loans issued between January 2009 and December 2013 with 36-months duration. I focus on this period because the default rate of loans issued before January 2009 is higher than the default rate of loans issued between January 2009 and December 2013. This difference in rates might be caused by the financial crisis in 2007-2008 which negatively affected a lot of US residents. I believe that selecting only loans that were issued after 2008 helps to avoid a structural break in the data set. In addition, I haven't included loans issued after December 2013 as their maturity hasn't yet been reached. For a similar reason I haven't included loans with 60-month duration. Since loans with 60-month duration were firstly introduced in 2010 their maturity hasn't yet been reached.

RESPOSE VARIABLE

For the analysis, I classify loans in the data set as 'Fully Paid' or as 'Charged Off'. Such classification helps to differentiate between good and loans. Indeed, the loans in the data set have six different statuses such as 'Fully Paid', 'Charged Off', 'Current', 'Late (31–120 days)', 'Late (16–30 days)', 'In Grace Period' and 'Default'. A loan is labeled as 'Fully Paid' when the loan principal and the loan interest are fully paid back. A loan is labeled as 'Charged Off 'when a loan borrower defaulted on the loan and the loan will never be paid back in full amount. Although I've chosen the dataset's time frame so that all loans in the dataset are supposed to have already reached their maturity, there are still some loans which have not been completely paid back or charged off. Such situation is usually caused when a borrower makes a payment after a payment due date. Delayed payments increase the maturity of a loan. Such loans are usually labeled as 'Current', 'In Grace Period',' Late (16–30 days)', 'Late (31–120 days) 'or 'Defaulted'. The loans labeled as 'Current' are currently being paid back. I didn't include them in the analysis because it's uncertain whether they will or will not be paid back. Similarly, the dataset contains a few loans loans with status 'In Grace Period' and several loans with status 'Late (16–30 days)'. 'In Grace Period' status means that a loan instalment is delayed by at most 15 days. A loan with status 'Late (16–30 days)' has a delayed instalment between 16 to 30 days. I don't consider loans with statuses 'In Grace Period and Late (16–30 days)' as Charged off as these loans are not delayed by more than 30 days and in theoretically might be paid off. Lending Club statistics shows that 75% of loans with status 'Late (31–120 days)' are never fully paid. The dataset contains 91 loans with status 'Late (31–120 days)' and 50 of them are delayed by more than 90 days. I labeled them as 'Charged Off' since I assumed that those loans would never be repaid. Loans with marked as 'Default' have delayed instalment by more than 120 days. They are labeled as 'Charged Off' in the project as well. The proportion of 'Fully Paid' and 'Charged Off' loans are shown in Table 1.

Table 1. ***Distribution of Loan Statuses***

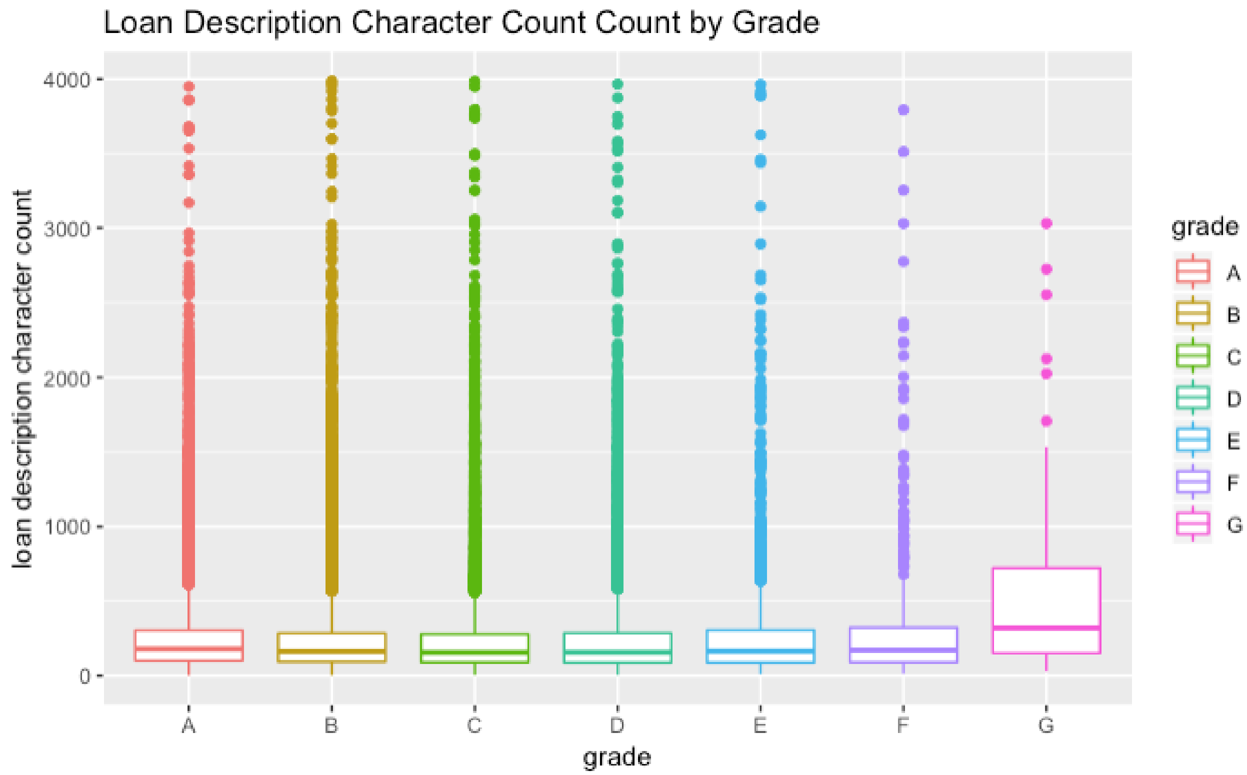
Initial Data Set		**Modified Data Set**	
Loan Status	**# of Loans**	**Loan Status**	**# of Loans**
Charged Off	60,836	Charged Off	61836
Does not meet the credit policy. Status: Charged Off	8779		
Does not meet the credit policy. Status: Fully Paid	1001		
Fully Paid	6	Fully Paid	8837
Current	6		
In Grace Period	81		
Late (31-120 days)	81		
Default	12		
Late (16-30 days)	6		
Total Number	***70753***	***Total Number***	***70673***

EXPLANATORY VARIABLES

The ORIGINAL dataset contains 145 variables. Several variables such as Loan URL, Loan ID, Personal Finance Inquiries and Finance Trades were excluded for as they don't include any values and don't contain any useful information for the purposes of the project. The variables of interest can be divided into two groups of information origin. The first group is the information that was reported by borrower. Borrower's self-reported information are Annual Income, Length of Employment, Loan Amount, Loan Purpose, Housing Situation, Loan Description, etc. The second group of information origin is the borrower's credit file provided by one of three national credit bureaus in the USA. Credit bureaus reported information are Debt-to-Income Ratio, 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, Revolving Credit Utilization, etc. The description of all variables is included in Table A1 in Appendix A.

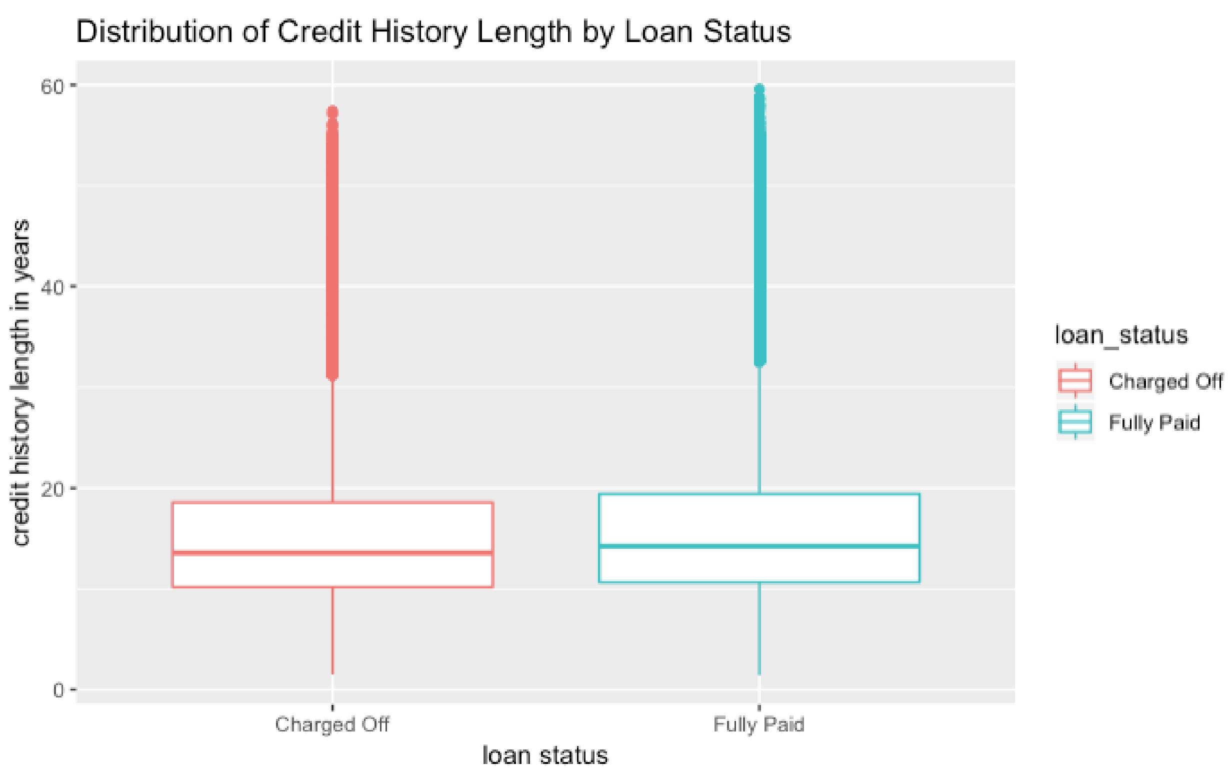
I modified several variables from the original dataset. The first variable that was modified is Loan Description. It is provided by a borrower when applying for a loan. There are many ways to use Loan Description as an independent variable that might be the predictor for borrowers' default. I counted the number of characters in Loan Description and replace the description with description's character count. Graph 1 shown below proves that loans with the lowest loan grade "G" on average have longest loan descriptions.

Graph 1. ***Distribution of loan description character count by grade***



The second variable that was modified is Date of First Credit Line. The variable represents the reported date (in form of month and year) of the first open credit line. I transformed the variable into the number of years since the first reported credit line was opened and replaced the reported date with the number of years since. Graph 2 shows that distributions of credit history length is pretty similar for loans that were fully paid and the loans that were charged off. Loans with status “Fully Paid” on average have a longer credit history than loans with status “Charged Off”. Thus, borrowers with longer credit history are more likely to paid off loans. However, the difference in credit history length for two categories of loans is so small that it can be concluded that credit history length doesn’t have a significant impact on loan default status and most likely credit history length determine whether the loan will be issued or not.

Graph 2. ***Distribution of credit history length in years by loan status***

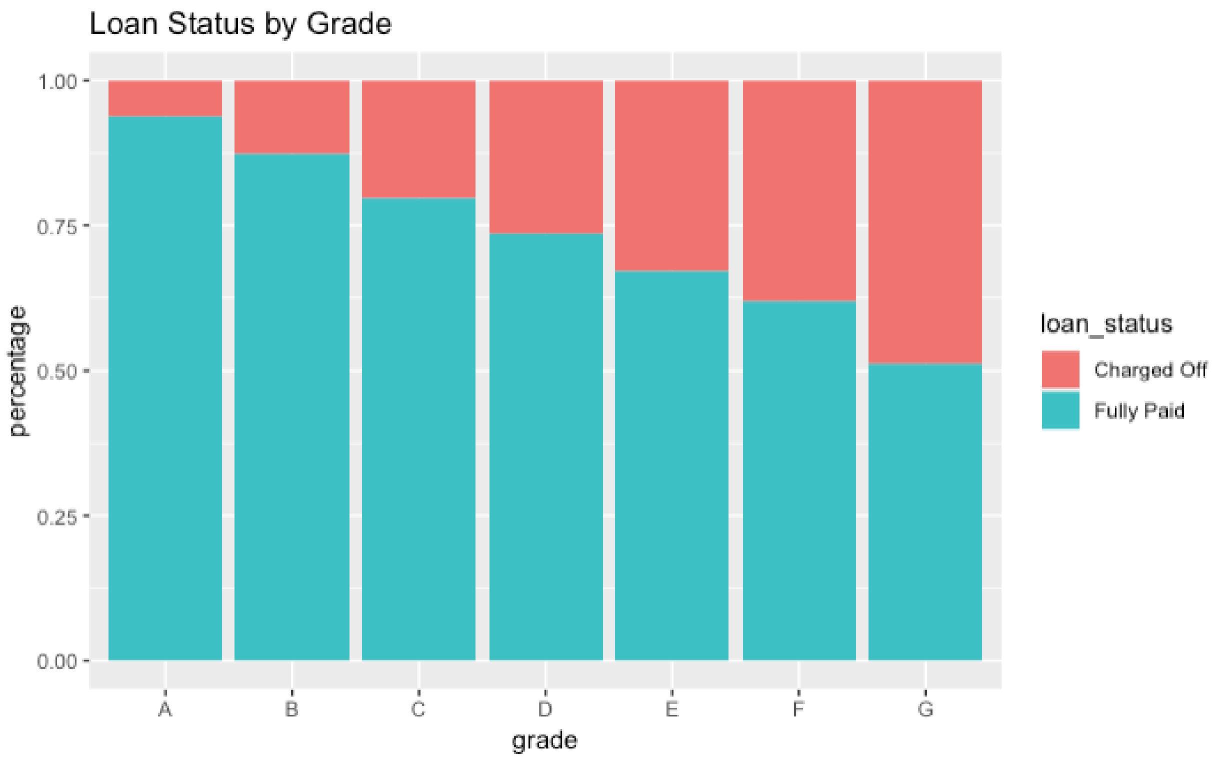


I distinguish between loan risk classes in the analysis. Loans with ‘A’ grade belong to the low-risk class, loans with ‘B’ grade belong to the medium-risk and loans with ‘C’ grade belong to the high-risk class. Loans graded with ‘D’, ‘E’, ‘F’ and ‘G’ are aggregated to very high-risk class in order to make the classes somewhat comparable in terms of the number of observations. Loans in the very high-risk class are quite similar in terms of default rate and FICO score. Only the default rate of G-graded loans stands out. However, as there are only about 80 loans with grade ‘D’, it would not be useful to create a separate group for these loans. Therefore, I added G-graded loans to the same class as ‘D’, ‘E’ and ‘F’-graded loans. Table 2 provides an overview of the four loan classes and their corresponding loan grades and average default rates. Graph 3a shows that grades determine loan default risk. Loans with higher grades have higher chances to be paid off. While Graph 3a depicts that grades determine interest rate. Loans with higher grades have lower interest rates.

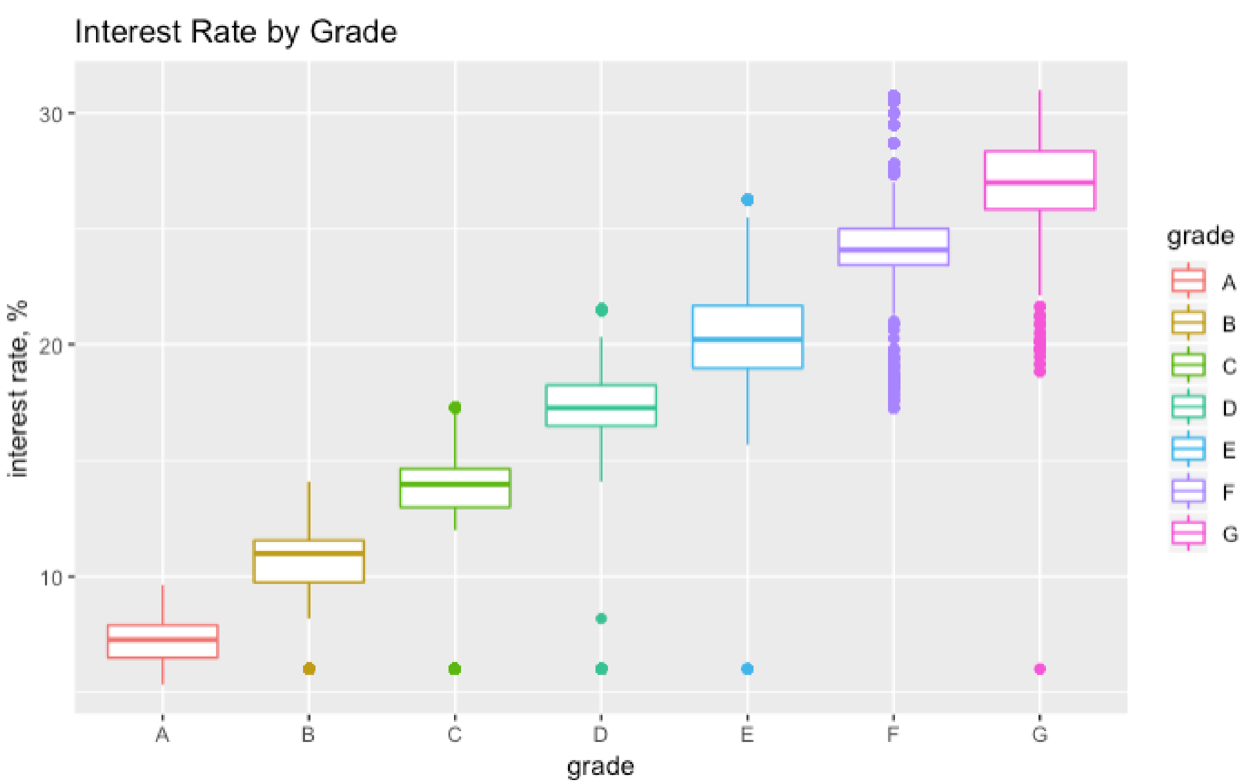
Table 2. ***Loan classes and their corresponding loan grades and average default rates***

Type of Class	**Loan Grade**	**Default Rate, %**
Low- Risk Class	A	6,6
Medium-Risk Class	B	11.8
High-Risk Class	C	16.5
Very High-Risk Class	D, E, F, G	20.1
All Loan Classes		***12.5***

Graph 3a. ***Loan statuses by grade***



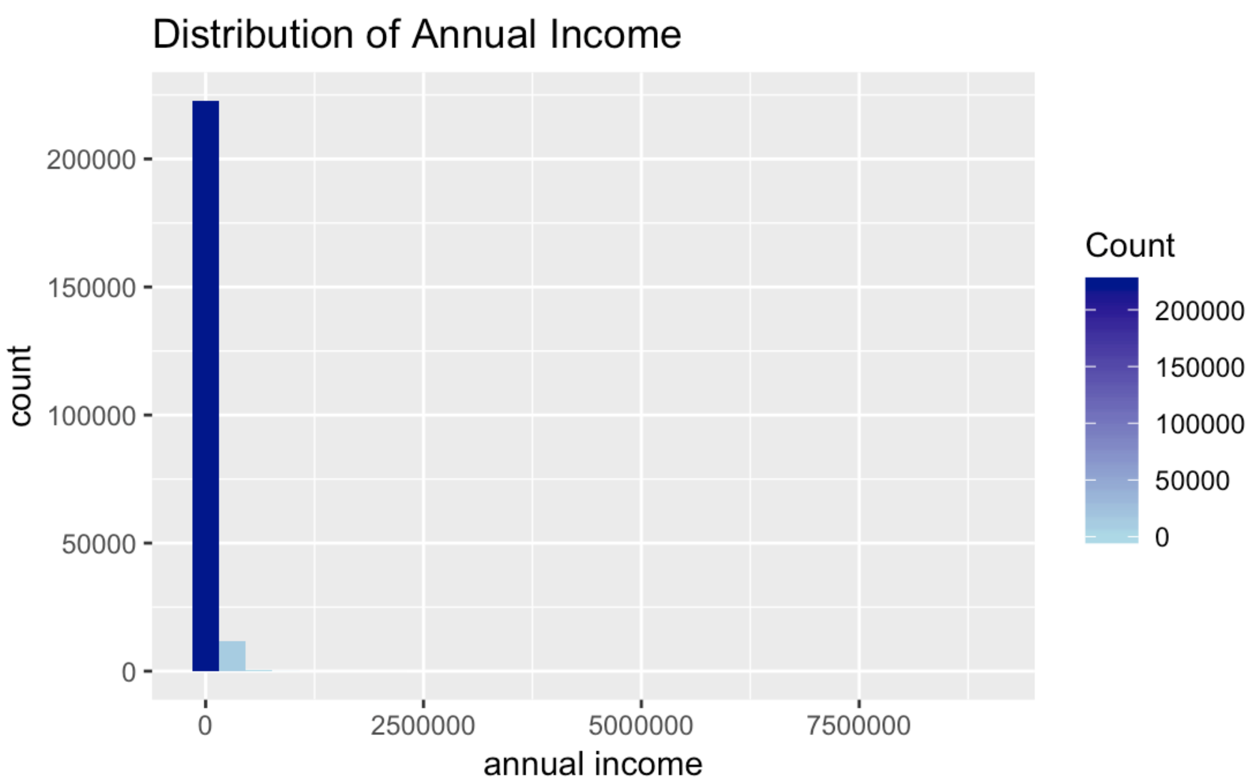
Graph 3b. ***Distribution of interest rate by grade***



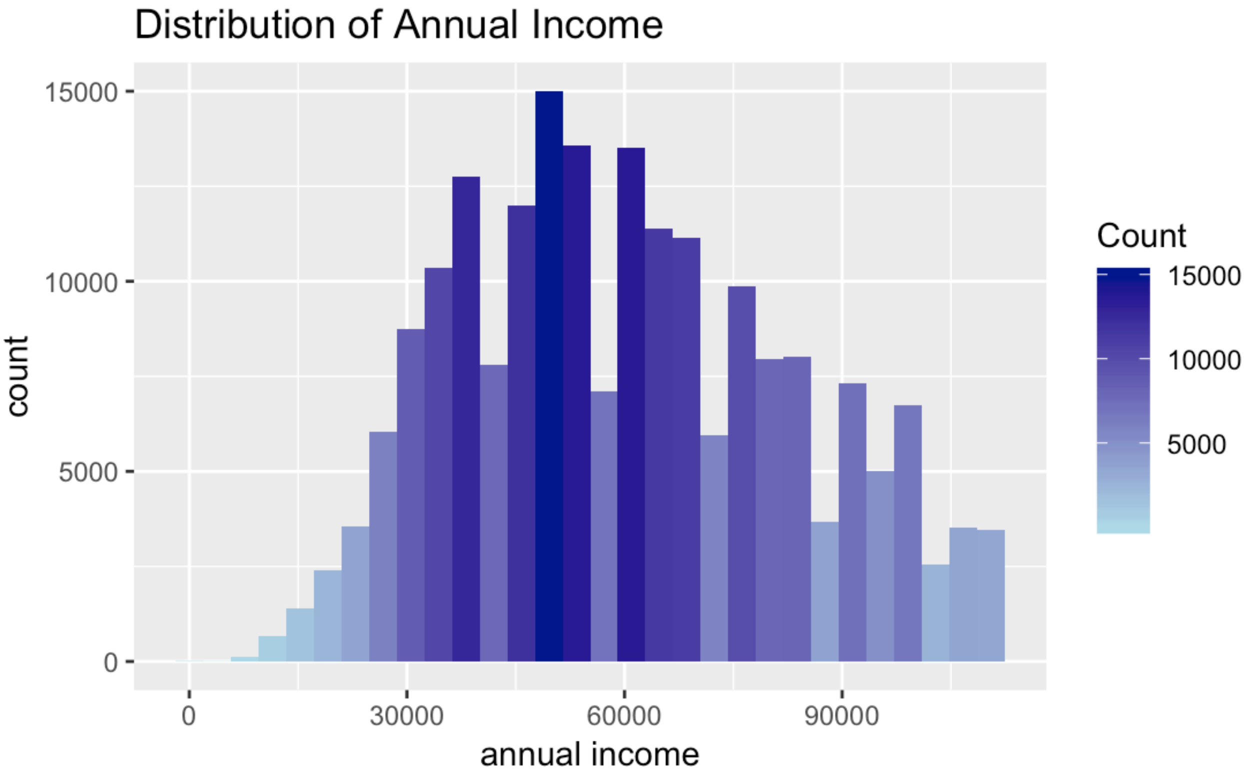
OUTLIERS

The dataset contains about 15 records with a self-reported income exceeding 1M USD. Such observations were considered to be outliers and were removed from the dataset. See graphs below. Graph 4a displays the distribution of annual income before outliers were removed whereas Graph 4b depicts the distribution of annual income after outliers were removed.

Graph 4a. ***Distribution of annual income in original dataset***



Graph 4b. ***Distribution of annual income after removal of outliers***

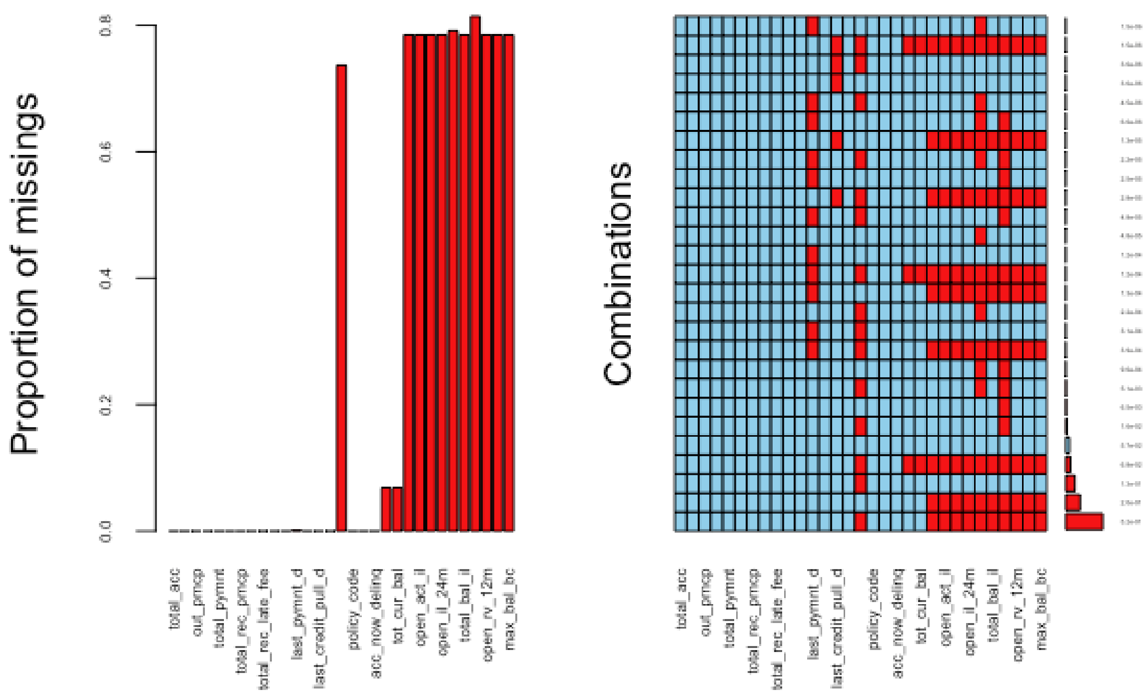


MISSING VALUES

The dataset contains a lot of missing values. Graph 5 depicts the distribution missing values for several variable in the dataset. For example, the variable Joint Revolving Balance and Hardship Status are missing more than 99% of their values. The variables that miss more than 96% of values were removed since they don't have any statistical significance and have a very little or no effect on the response variable.

By looking at the dataset it can be concluded that missing values of several variables might be predictive of the response variable Loan Status. For instance, missing values of Home Ownership were replaced by "Not Provided". Instead of throwing out missing values of the remaining variables containing missing data were restored. So that, I got a lot more data to feed to a model. In order to restore the missing data, I applied multiple imputation technique. The method ‘PMM’ (replacement with mean) that is also known as Predictive Mean Matching was chosen for imputation procedure.

Graph 5. ***The distribution of the missing values for several variables of the dataset.***



TESTING AND TRAINING DATASETS

The dataset was split into 66% training data and 34% testing data. Training dataset contains 442291 observations and testing dataset contains 227847 observations.

METHODS

BINIRY LOGISTIC REGRESSION

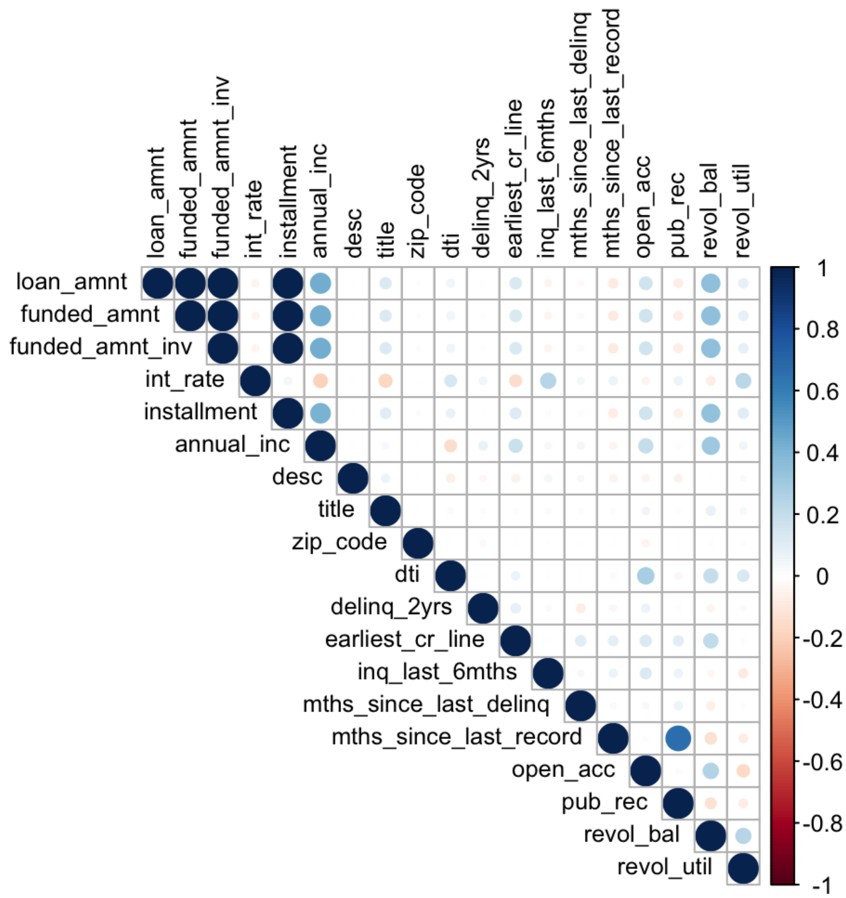
As the response variable Loan Status is the binary categorical variable (that can take values of "Paid Off" or "Charged Off") I decided to use binary logistic to analyze the determinants of borrowers' default. I selected backward stepwise elimination approach to find the most critical features. I started with a full model including all variables of interest. Then I dropped every variable with a p-value higher than 0.1 starting with the variable with the highest p-value. Backward stepwise elimination generated a few models. In order to select the model that fits data the best I used AIC (Akaike's criterion). Lower values of AIC indicate the preferred model, that is, the one with the fewest parameters that still provides an adequate fit to the data.

ASSUMPTIONS

The logistic regression must satisfy the assumptions described below.

1. Cases are independent. This requirement is met as the majority of borrowers don't relate to each other.
2. Multicollinearity. For example, the variable Loan Amount is highly correlated with Funded Amount, Funded Amount Invested and Installment. See the Graph 6 shown below. Highly correlated predictor variables were removed from the dataset.

Graph 6. *Correlation Matrix.*



3. Linear relationship between the logit of the explanatory variables and the response. The variables that didn't meet this requirement were replaced by their logs.
4. Errors need to be independent but not normally distributed. The variables that didn't meet this requirement were replaced by their logs.
5. Large sample. The training dataset containing more that 400000 observations can be considered to be large enough.

REGRESSION RESULTS

The binary logistic regression identifies that features like Annual Income, 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 Length 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, while Renewable Energy, and Small Business are positively correlated with loan default. Home ownership (status Mortgage) is negatively correlated with loan default.

I proceed with regressions for the four loan risk classes. Table 3 describes loan default determents based on logistic regression results for different loan risk classes. Results for the Low Risk (grade A) and Medium Risk classes (grade B) only differ slightly from the All Classes results. In both the Length 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 (grades D-G), the Number of Characters is not significant anymore as well as the loan purposes Car, Debt Consolidation, Home Improvement and Renewable Energy. The Length 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. Revolving Credit Utilization has been found to be a significant predictor for borrowers’ default in all related studies as well as in the All Classes data. However, it is only significant in the Low-

Risk and Medium-Risk Classes. It is not a significant determinant in the High-Risk and Very High-Risk Class.

LOGISTIC REGRESSION RESULTS

Revolving Credit Utilization is a positive determinant of borrowers' default only in low loan risk classes. The Debt-to-Income ratio is significant in all loan classes. In fact, the Debt-to-Income ratios for defaulted/non-defaulted loans have almost identical values across risk classes.

The Debt-to-Income ratio is a positive 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. 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 is neither significant in the Very High-Risk Class nor in the High-Risk Class, while they are in low risk classes.

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

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 All Classes regression results. This finding is in line with 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. 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.

The Length of Credit History is a negative determinant of borrowers' default only in the High-Risk Class. In the full data set, all variables of interest turn out to be significant determinants of default except the variable Open Credit Lines. Generally, discrepancies of results could be due to the fact that the data avoids the structural break of loan defaults possibly caused by the 2007-2008 financial crisis. The only difference to previous studies 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 [used loans with status 'current' in his analyses. Comparing the results to studies, 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, All Classes results are quite different from studies results. Besides differences in the time frame of the data set, the study includes 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.

Table 3. ***Loan default determents based on logistic regression for different loan risk classes.***

Features	**All Classes**	**Low Risk Class (A)**	**Medium-Risk Class (B)**	**High-Risk Class(C)**	**Very High-Risk Class(D-G)**
Loan Amount	x	x	x	x	
Annual Income	x				x
Loan Description Character Count	x			x	
Debt-to-Income	x	x		x	x
Credit History Length	x			x	
# Inquiries Past 6 Months	x				x
# Months since Last Record	x				x
# Open Credit Lines	x				
Late fees received to Date	x	x	x	x	x
Recoveries	x				x
# Months since Last Payment	x			x	
Last Payment Amount	x				
# Months since Last Credit Pull Date	x		x	x	x
Average Current Balance	x				

# of Installment Accounts Opened in Past 12 Months	x	x	x	x	x
Balance to Credit Limit on All Trades	x		x		
# of Finance Trades	x			x	
# Inquiries Last 12 months	x				
# Accounts Opened Past 24 Months	x	x	x	x	
Total Open to Buy on Revolving Bankcards	x				x
Ratio of total current balance to high credit/credit limit for all bankcard accounts	x	x	x		x
# of Currently Active Bankcard Accounts	x				x
Grade	B-G				
Employment length	n/a				
Issue date	2014-2016				
Debt Settlement Flag	Y				

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 can 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 identified some of the borrower's and loan's information as useful determinants for borrowers' default. I conjecture 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 the full data set are largely in line with findings of previous studies, the 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. Length of Credit History is only significant for high loan risk classes.

The 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.

APPENDIX A

Table A. ***Description of the variables.***

Variable	Description	Data Type
acceptD	The date which the borrower accepted the offer	Date
accNowDelinq	The number of accounts on which the borrower is now delinquent.	Discrete
accOpenPast24Mths	Number of trades opened in past 24 months.	Discrete
addrState	The state provided by the borrower in the loan application	Nominal Values: AK, MA, NY etc.
all_util	Balance to credit limit on all trades	Continues
annual_inc_joint	The combined self-reported annual income provided by the co-borrowers during registration	Continues
annualInc	The self-reported annual income provided by the borrower during registration.	Continues
application_type	Indicates whether the loan is an individual application or a joint application with two co-borrowers	Nominal Value: Individual, Joint, Co-borrower
avg_cur_bal	Average current balance of all accounts	Continues
bcOpenToBuy	Total open to buy on revolving bankcards.	Discrete
bcUtil	Ratio of total current balance to high credit/credit limit for all bankcard accounts.	Continues
chargeoff_within_12_mths	Number of charge-offs within 12 months	Discrete
collections_12_mths_ex_med	Number of collections in 12 months excluding medical collections	Discrete
creditPullD	The date LC pulled credit for this loan	Date
delinq2Yrs	The number of 30+ days past-due incidences of delinquency in the borrower's credit file for the past 2 years	Discrete
delinqAmnt	The past-due amount owed for the accounts on which the borrower is now delinquent.	Continues
desc	Loan description provided by the borrower	Can be converted to categorical. Values: home loan, student loan, car loan etc.

dti	A ratio calculated using the borrower's total monthly debt payments on the total debt obligations, excluding mortgage and the requested LC loan, divided by the borrower's self-reported monthly income.	Continues
dti_joint	A ratio calculated using the co-borrower's total monthly payments on the total debt obligations, excluding mortgages and the requested LC loan, divided by the co-borrowers' combined self-reported monthly income	Continues
earliestCrLine	The date the borrower's earliest reported credit line was opened	Date
effective_int_rate	The effective interest rate is equal to the interest rate on a Note reduced by Lending Club's estimate of the impact of uncollected interest prior to charge off.	Continues
emp_title	The job title supplied by the Borrower when applying for the loan.*	Can be converted to categorical. Values: IT, Finance etc.
empLength	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.	Discrete
expD	The date the listing will expire	Date
expDefaultRate	The expected default rate of the loan.	Continues
ficoRangeHigh	The upper boundary range the borrower's FICO at loan origination belongs to.	Discrete
ficoRangeLow	The lower boundary range the borrower's FICO at loan origination belongs to.	Discrete
fundedAmnt	The total amount committed to that loan at that point in time.	Continues
grade	LC assigned loan grade	Ordinal. Values: A, B, C etc.
homeOwnership	The home ownership status provided by the borrower during registration.	Nominal. Values: Rent, Own, Mortgage, Other.
id	A unique LC assigned ID for the loan listing.	Discrete

il_util	Ratio of total current balance to high credit/credit limit on all install acct	Continues
ils_exp_d	wholeloan platform expiration date	Date
initialListStatus	The initial listing status of the loan.	Nominal. Values: W, F
inq_fi	Number of personal finance inquiries	Discrete
inq_last_12m	Number of credit inquiries in past 12 months	Discrete
inqLast6Mths	The number of inquiries in past 6 months (excluding auto and mortgage inquiries)	Discrete
installment	The monthly payment owed by the borrower if the loan originates.	Continues
intRate	Interest Rate on the loan	Continues
isIncV	Indicates if income was verified by LC, not verified, or if the income source was verified	Nominal. Values: Verified and not verified
listD	The date which the borrower's application was listed on the platform.	Date
loanAmnt	The listed amount of the loan applied for by the borrower. If at some point in time, the credit department reduces the loan amount, then it will be reflected in this value.	Continues
max_bal_bc	Maximum current balance owed on all revolving accounts	Continues
memberId	A unique LC assigned Id for the borrower member.	Discrete
mo_sin_old_rev_tl_op	Months since oldest revolving account opened	Discrete
mo_sin_rcnt_rev_tl_op	Months since most recent revolving account opened	Discrete
mo_sin_rcnt_tl	Months since most recent account opened	Discrete
mortAcc	Number of mortgage accounts.	Discrete
mths_since_last_major_derog	Months since most recent 90-day or worse rating	Discrete
mths_since_oldest_il_open	Months since oldest bank installment account opened	Discrete
mths_since_rcnt_il	Months since most recent installment accounts opened	Discrete
mthsSinceLastDelinq	The number of months since the borrower's last delinquency.	Discrete
mthsSinceLastRecord	The number of months since the last public record.	Discrete

mthsSinceMostRecentInq	Months since most recent inquiry.	Discrete
mthsSinceRecentLoanDelinq	Months since most recent personal finance delinquency.	Discrete
reviewStatus	The status of the loan during the listing period.	Ordinal Values: Approved, Not Approved
disbursement_method	The method by which the borrower receives their loan	Nominal Values: Cash, Direct Pay

REFERENCES

Barasinska, N.; Schäfer, D. Is crowdfunding different? Evidence on the relation between gender and funding success from a German peer-to-peer lending platform. Available online: https://onlinelibrary.wiley.com/doi/abs/10.1111/geer.12052 (accessed on 21 March 2019).

Carmichael, D. Modeling Default for Peer-to-Peer Loans. 2014. Available online: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=%202529240 (accessed on 20 March 2019).

Deloitte. Banking Disrupted: How Technology Is Threatening the Traditional European Retail Banking Model. 2014. Available online: https://www2.deloitte.com/content/dam/Deloitte/global/Documents/Financial-Services/dttl-fsi-uk-Banking-Disrupted-2014-06.pdf (accessed on 15 March 2019).

Emekter, R.; Tu, Y.; Jirasakuldech, B.; Lu, M. Evaluating credit risk and loan performance in online Peer-to-Peer (P2P) lending. Appl. Econ. 201 Available online: https://www.tandfonline.com/doi/abs/10.1080/00036846.2014.962222 (accessed on 20 March 2019).

Freedman, S.; Jin, G.Z. The Information Value of Online Social Networks: Lessons From Peer-to-Peer Lending. 2014. Available online: https://www.nber.org/papers/w19820 (accessed on 20 March 2019).

Herzenstein, M.; Dholakia, U.M.; Andrews, R.L. Strategic Herding Behavior in Peer-to-Peer Loan Auctions. J. Interact. Mark. Available Online: https://www.sciencedirect.com/science/article/pii/S1094996810000435?via%3Dihub (accessed on 20 March 2019).

Iyer, R.; Khwaja, A.; Luttmer, E.; Shue, K. Screening in New Credit Markets Can Individual Lenders Infer Borrower Creditworthiness in Peer-to-Peer Lending? 2015. Available online: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=1570115 (accessed on 20 March 2019).

Lin, M.; Prabhala, N.; Viswanathan, S. Judging borrowers by the company they keep: Friendship networks and information asymmetry in online peer-to-peer lending. Available online: https://pubsonline.informs.org/doi/abs/10.1287/mnsc.1120.1560 (accessed on 20 March 2019).

Mills, K.G. The State of Small Business Lending: Credit Access during the Recovery and How Technology May Change the Game. 2014. Harvard Business School Working Paper. Available online: https://www.hbs.edu/faculty/Publication%20Files/15-004_09b1bf8b-eb2a-4e63-9c4e-0374f770856f.pdf (accessed on 20 March 2019).

Serrano-Cinca, C.; Gutiérrez-Nieto, B.; López-Palacios, L. Determinants of Default in P2P Lending. Available Online: https://www.ncbi.nlm.nih.gov/pubmed/26425854 (accessed on 21 March 2019).