Impact of interest rate on loan default in P2P platforms

NOCERA Alessio, LATIL Quentin, WANG Qi, GUO Ruiqi

March 2024

Abstract

The motivation of this study is to analyze the impact of interest rates on loan default rates using a large-scale dataset obtained from a leading P2P lending company in the United States of America during the first quarter of 2016. We employ a logistic model to regress the interest rate on loan default, controlling for various factors including debt-to-income ratio, number of current credit lines, delinquency in the last two years, number of mortgage accounts, and loan term. The inclusion of these control variables allows for a comprehensive analysis of the relationship between interest rates and loan default rates. Our findings indicate that all variables have a positive impact on the probability of default , except for the number of mortgage accounts that have a negative impact. The result obtained for our main interest variable is consistent with existing literature.

Contents

1 Introduction

In this study we aim to analyze how the interest rate influences loan default on P2P (Peer-to-peer) platforms.

To address this question effectively, it is imperative to establish a clear understanding of key terms central to our investigation. Specifically, defining concepts such as ”interest rate”, ”loan default” and ”P2P platform”, lays the groundwork for an analysis of their interplay within the P2P lending ecosystem. The interest rate is “the interest percent that a bank or other financial company charges you when you borrow money, or the interest percent it pays you when you keep money in an account”[1] but it can also be defined as the cost of borrowing money. Loan default is defined as the failure to repay a loan according to agreed terms and finally a peer-to-peer lending platform is an online marketplace that connects lenders and borrowers, allowing them to directly engage in borrowing and lending activities without traditional financial institutions serving as intermediaries.

It is an important problem that needs to be tackled because borrowers who default cause the largest amount of loss to the lenders. On P2P platforms, where borrowers are often considered more risky than traditional borrowers, we believe it is crucial to study whether increasing the interest rate serves as a mean to protect lenders against such risky borrowers. We aim to understand whether this rate increase, intended to secure the lender, might instead exacerbate his situation.

Furthermore, we want to add that loan defaults have negative repercussions not only for lenders but also for borrowers. Defaulting on a loan can lead to damaging consequences such as damage to credit scores[2], increased difficulty in obtaining future loans or credit, potential legal action, and even asset seizure. These consequences can significantly impact a borrower’s financial stability and future opportunities, underscoring the importance of understanding the factors contributing to loan default

[1] From https://dictionary.cambridge.org/dictionary/english/interest-rate

[2] numerical ranking of an individual’s financial creditworthiness based on spending and credit history. (from https://www.dictionary.com/browse/credit-score)

rates.

To find an answer to the problem, we begin by studying previous empirical analysis on the subject to establish a foundation and identify gaps in existing research. Using data from a P2P lending platform, we perform descriptive statistics and an econometric regression to explore the relationship between interest rates and loan default, and to determine the factors influencing default rates on the platform.

For the econometric regression, we employ a logistic model due to the binary nature of our independent variable (Charged Off or Fully Paid). Our analysis reveals that all variables exhibit a positive impact on the probability of default, with the exception of the number of mortgage accounts, which had a negative effect. To interpret these results, we use the Average Marginal Effects that allows an interpretation of the coefficient in terms of change of probability of default. Subsequently, we assess the predictive capability of the model by conducting a prediction test. While the results are satisfactory, they do not demonstrate exceptional predictive performance.

2 Review of Literature

2.1 Interest rate

A study conducted by Santoso et al.(2020) investigated the relationship between certain variables, namely the interest rate and the default status on P2P platforms. The data for the study were collected from three P2P platforms kept anonymous (Gamma, Alpha, and Beta) in Indonesia spanning the years 2014-2018. According to the findings from the study, there is a positive association between loan interest rates and loan default for Gamma and Beta (the result is not significant for Alpha). Echoing this sentiment, Jote (2018) found a negative association between interest rates and loan repayment likelihood. Supporting these findings, Edet et al. (2014) observed a significant positive effect of

interest rates on loan default rates. Collectively, these studies suggest that higher interest rates lead to an increase in the rate of loan defaults.

2.2 Term of loan

Nadeesha and Madhushani (2023) conducted a study on credit worthiness of Personal Loan Borrowers of a Bank in Sri Lanka, finding a negative impact of the loan term on credit worthiness which can be translated to a positive impact on loan default. In the view of Awunyo-Vitor (2012), repayment period significantly affects loan repayment default. This result is consistent with the research of Zhichao et al.(2020) who found that compared to the short term, the medium term loans and longer term loans are both more likely to default.

2.3 Number of mortgage accounts

Uddin (2019) asserted that collateral can serve as a loan's security. It can reduce the risk of loan as the lenders can sell it to recover the loan. If the borrowers value their collateral, they will be more motivated to repay the loans. Furthermore, Zhichao et al.(2020) who studied the impact of collateral on the agriculture related loan default thought that collateral is a practical way to guarantee borrower's behavior. In summary, the existence of such collateral might decrease the risk of loan default.

2.4 Delinquency

Serrano-Cinca et al.(2015) analyzed the key factors of default by using the cox regression. The result shows that there exists a positive correlation between the delinquency and loan default, where the positive sign implies a k-fold increase in risk.

3 Data Collection

3.1 Data Source

This last decade, the United States saw a high demand for loans driven by favorable interest rates. Homebuyers sought mortgages, small businesses wanted funds for expansion, and students needed loans for education. Traditional banks and emerging fintech[3] (financial technology) companies played crucial roles in meeting this demand. Fintech introduced innovative borrowing methods, like peer-to-peer lending and online loan platforms, reshaping the lending landscape.

To study the default on loans, we use data from LendingClub. LendingClub is a peer-to-peer online lender that has 4.7 million members, making it one of the United States' biggest online banks.

LendingClub has made available the information on all transactions that have been conducted on the platform from 2007 to 2019. Our database is built upon cross-sectional data capturing the characteristics and attributes of the lenders and borrowers. The original database comprises over 2 million rows and 151 variables, making it exceptionally large. To manage its size, we decide to focus on a specific time frame. After careful consideration, we select data from January 2016 to March 2016, resulting in just over 135,000 records.

For this study, our independent variable is loan status. Out of the five modalities available (Fully Paid, Charged-Off, Current, In Grace Period, Late), we restrict our analysis to only include "Fully Paid" and "Charged-Off" statuses. This modification enables us to concentrate on these two specific outcomes, knowing that the other modalities can not be considered as default, ensuring relevance in our investigation (See Table 1). After the exclusion of those loans, we are left with a data frame of 115,000 records. However, even this subset is substantial, prompting us to further streamline our analysis. Finally, we opt for a random sample of 15,000 records.

[3]Digital technological innovations utilized by customers or institutions in the financial services industry (from https://www.dictionary.com/browse/fintech)

To isolate the specific effect of the explanatory variable on the dependent variable, we add control variables to capture variations independent of the main explanatory variable (interest rate). We end up with five control variables including four continuous (debt to income ratio[4], the number of current credit lines[5], delinquency in the last two years, the number of mortgage accounts[6] and one binary (term of the loan). An additional modification is on "delinquency in the last two years", indeed we decided to code it as 1 if one or more default, and 0 otherwise, turning this continuous variable into a binary variable (See Table 2).

3.2 Dealing with outliers

After conducting a thorough examination of our variables, we opt to exclude outlier observations from our analysis. The identification of outliers is performed using the definition of box plots. Visually they are the points lying beyond the "whiskers" of the plot, and mathematically they are values falling outside 1.5 times the interquartile range beyond the upper or lower quartiles. We do not remove all observations located above the "whiskers". After removing the top extreme observation, we can visualize and remove values that we consider can distort our model (Figure 1 and 2). We end up having a data frame with 14,754 records after removing outliers.

We do some descriptive statistics for all our variables: we compute the mean, standard deviation and the quartiles for our continuous variables and the frequency and proportion for our binary variables. (Table 3 and 4).

To have a first insight over the relation between the interest rate and loan default we use Box Plots to see if there is a difference between the interest rate in the group that defaulted on their loans and

[4]Ratio calculated using the borrower's total monthly debt payments, excluding mortgage and the requested LC loan, divided by the borrower's self-reported monthly income.

[5]Also called credit limit, line of credit. It is the maximum amount of credit that a customer of a store, bank, etc., is authorized to use. (from https://www.dictionary.com/browse/credit-line)

[6]A mortgage loan or simply mortgage, is a loan used either by purchasers of real property to raise funds to buy real estate, or by existing property owners to raise funds for any purpose while putting a lien on the property being mortgaged. The loan is "secured" on the borrower's property through a process known as mortgage origination. (from https://en.wikipedia.org/wiki/Mortgage)

the group that reimbursed it entirely. We indeed notice a distinct difference in the distribution of the interest rates across the two groups (Figure 3).

4 Econometric Model

4.1 Logistic regression

We use a logistic regression for the econometric study given that loan default is a binary variable (Table 1):

$$\Pr(\text{loan\_status}_i = 1|\mathbf{x}_i) = \Lambda(\beta_0 + \beta_1 \text{interest\_rate}_i + \beta_2 \text{delinq}_i + \beta_3 \text{open\_acc}_i + \beta_4 \text{term}_i + \beta_5 \text{mort\_acc}_i + \beta_6 \text{dti}_i)$$

for i = 1, . . . , n

$$\Lambda(\mathbf{X}) = \frac{1}{1+e^{-\mathbf{X}}}$$

In a logistic regression we cannot interpret the betas as a change of probability of loan default. We can only interpret the sign of the beta. If we want to interpret the beta, the magnitude of the beta represents the change in the log-odds[7] of loan default. That's why we use the Average Marginal Effect (AME) because the results are expressed in terms of probabilities, which is easier to interpret and communicate. We choose the AME over the "Marginal Average Effect" which is the marginal effect at the mean of the covariates and over "AME at Representative Values" as it is the typical default choice. To control for heteroscedasticity we use robust standard errors and not standard errors as they might be biased. (See Table 5)

[7]Log-odds is the logarithms of the odds of loan default. And the "odd" is the probability that there is a loan default on the probability that there isn't.

4.2 Prediction

As a reminder, the objective of our model is to assist lenders on P2P platforms in estimating the probability of default for borrowers based on the interest rates they impose and other variables seen previously. We want to assess whether our econometric model provides accurate predictions of default probability. Using a confusion matrix where each value in the confusion matrix is divided by the sum of the corresponding column we can easily interpret the results we obtain in the prediction model.

We initially considered a threshold of 0.5, which is a common default choice. This would classify any loan with a predicted probability greater than 0.5 as Defaulted and otherwise as Fully Paid. However, only 14% of defaults are actually predicted as default (Table 6). Therefore, we opt for a lower threshold of 0.2. While this may result in classifying some Fully Paid loans as Defaulted, it prioritizes accurately predicting loans that actually defaulted. This approach aims to minimize the risk of overlooking potential defaults. Now we predict as defaulted 66% of actual defaults (Table 7). Furthermore we do a mosaic plot which allows us to easily understand visually the proportions of the confusion matrix (See Figure 5).

4.3 Endogeneity

After analyzing the correlation of the interest rate and the error term through graphical examination, we spotted particular patterns (see Figure 4 and explanation). Indeed the pattern of residuals may be indicative of potential endogeneity issues due to the correlation of the interest rate with the error term in the regression model. We think that there might be a bidirectional relationship between the interest rate and loan default (our dependent variable). This simultaneity could lead to biased model estimates and residuals. In light of these findings, for future research, we aim to explore these outlier groups to understand the characteristics of individuals within them, with the goal of refining

our model and enhancing predictive accuracy. We are aware that the use of an IV regression with appropriate instruments could disentangle the causal relationship between the variables and we acknowledge this potential endogeneity issue but unfortunately we can't address it because of the lack of empirical paper treating this.

5 Result

Upon examining the outcomes of our logistic regression model, we can draw some economic intuitions.

For a one percentage point increase in the interest rate, the average change in the probability of loan default increases by 1.7 percentage points, on average across the distribution of other explanatory variables. As anticipated, the interest rate exhibits a positive Average Marginal Effect on default, aligning with the understanding that the interest rate represents the cost of the loan. Furthermore it is consistent with findings by Wimboh et al. (2020) and other literature. When high interest rates are imposed on borrowers, managing loan repayments becomes more challenging as it is harder for borrowers to keep up with payments thus elevating the risk of default.

Meanwhile for a loan with a term duration of 60 months (compared to 36), the average change in the probability of loan default increases by 10.6 percentage points (avg.)[8], it exhibits a more noticeable impact, consistent with findings by Nadeesha and Madhushani (2023) and Wimboh et al.(2020).

Both the debt-to-income (dti) ratio and the number of credit lines contribute minimally, each showcasing a marginal positive impact on default likelihood. Indeed, for a one percentage point increase in the debt-to-income ratio, the average change in the probability of loan default increases by 0.2 percentage points (avg.) and for a one-unit increase in the number of open accounts, the average change

[8] Abbreviation 'avg.' denotes 'on average across the distribution of other explanatory variables'.

in the probability of loan default increases by 0.2 percentage point (avg.). The results obtained for the dti ratio is very intuitive, let's say we fix the income, if the debt goes up, the part of income that you can allocate to the debt goes down thus making it harder to repay the loan.

For individuals that committed one or more default in the last two years, the average change in the probability of loan default increases by 3.2 percentage point (avg.). Delinquency emerges as a more discernible factor, revealing a relatively higher influence on the probability of default and is consistent with findings by Serrano-Cinca et al. (2015).

Finally, for a one-unit increase in the number of mortgage accounts, the average change in the probability of loan default decreases by 2 percentage points (avg.). The negative Average Marginal Effect suggests a strategic use of mortgages for debt consolidation. It implies that individuals may be using their homes as collateral to consolidate debts, leading to a more stable financial position and reducing the risk of default.

Using the same line of logic, we would have expected a similar result for the number of credit lines, however, we observed an unexpected positive Average Marginal Effect. One possible interpretation is that individuals might be strategically using loans, such as P2P loans, to reimburse other debts such as the debts created by the credit lines thus influencing the observed increase in default probability. This counter-intuitive finding prompts further investigation into credit line and mortgage account management. It highlights the complex nature of financial behaviors within the lending context.

6 Conclusion

Knowing that an increase in the interest rate increase the probability of default in our study, increasing interest rates might not be the best strategy to reduce the risk of loan default. This insight emphasizes the need for exploring alternative strategies, especially for non-experts, to achieve more

effective risk management in the P2P lending landscape.

Our research study stands out from others in the field due to our unique approach in examining the impact of interest rates on loan default. Indeed we distinguish ourselves by adding specific control variables that have often been overlooked. We bring in factors like the number of credit lines and mortgage accounts. This lets us dig deeper into how lenders handle their finances and it helps us see how interest rates fit into the larger financial decisions made by both borrowers and lenders.

It is important to acknowledge the limitations of our study. We think adding key demographic variables such as gender, ethnicity, and age will help strengthen our analysis. We propose that future research endeavors consider incorporating those personal information's from borrowers as additional control variables alongside the selected ones. The inclusion of these additional variables has the potential to enrich our understanding of the factors influencing the probabilities of loan defaults and make our prediction model more accurate.

We've previously highlighted the potential endogeneity issue, which unfortunately we're unable to address due to the absence of empirical evidence. Selecting instruments in our data pose a challenge as we lack studies justifying its suitability for this role.

Appendix

To study the default on loans we use data from LendingClub, who began offering loans in 2007 as a peer-to-peer online lender. It expanded its banking services to checking and savings accounts. Lending Club permits loan amounts that range from $1,000 to $40,000 for terms of 36 or 60 months. Each loan has an interest rate, from 9.57%-36%, which is based on credit score and debt-to-income ratio.

We add control variables to capture variations independent of the main explanatory variable which will allow us to have a more accurate coefficient for the interest rate. We choose our control variables by ensuring they are correlated with the outcome variable, while also avoiding strong correlation with the explanatory variable to prevent multicollinearity, which could complicate the differentiation of their individual effects.

Table 1: Specification of the data extract and sample selection procedures

Variables	**Label**	**Restriction**	**Value label**
loan_status	Status of the loan	Keep	1=Charged off 0=Fully paid
int_rate	Interest rate	NA	
term	Term of the loan	NA	
dti	Debt to income	NA	
open_acc	Number of current credit lines	NA	
mort_acc	Number of current mortgage accounts	NA	
delinq	Delinquency in the last two years	NA	

This table provides the labels for each variable in the final dataset sourced from Lending Club on loan transactions between 2007 and 2019. The following two columns, ”Restriction” and ”Value Label” highlight the categories of each variable retained for our study if not all categories are used. Now, considering our dependent variable ”loan status”, it have 5 possible modalities:

1.Fully Paid: The applicant has successfully paid back both the principal and the interest rate.

2.Charged-Off: The applicant has failed to make timely installments for an extended period, lead-

ing to a default on the loan.

3.Current: The applicant is actively in the process of repaying the installments, indicating that the loan tenure has not yet concluded.

4.In Grace Period: During this defined period after the due date, the borrower can make a payment without incurring a late fee.

5.Late: The applicant is behind schedule but has not yet reached the threshold to be labeled as defaulted.

We found it necessary to exclude all loans where the borrower was in any of those three situations—'Current', 'In Grace Period' and 'Late'—as they are not classified as defaulted.

Table 2: Description of variables

Variables	**Type of variable**	**Modalities**
Interest rate	continuous	NA
Debt to income	continuous	NA
Number of current mortgage accounts	continuous	NA
Number of current credit lines	continuous	NA
Delinquency in the last two years	binary	1 = one or more defaults
		0 = otherwise
Term of the loan	binary	1 = 60 months
		0 = 36 months
Loan status	binary	1= Charged off
		0 = Fully paid

This table illustrates the type of variable (continuous or binary) for each variable selected for our study. It also shows how we code each modality of our binary variables. Term was already binary but delinquency was continuous. To be more precise about the chosen variable we have as continuous variables the Interest Rate expressed in percentages, Debt-to-Income Ratio (dti) , also in percentages. The Number of Open Credit Lines, and Number of Open Mortgage Accounts. Additionally, as binary variables we have the Delinquency in the Last Two Years that we code as 1 if there was one or more defaults, 0 otherwise and Term of the Loan that we code as 1 for 60 months and 0 for 36 months was considered.

Now concerning the treatment of outliers: After removing the top observation, we can indeed observe more precisely the distribution of the other outliers. Visually it is shown in the before (Figure 1) and after (Figure 2) box plots.

Figure 1: Box plot of Interest rate before removing the top observation

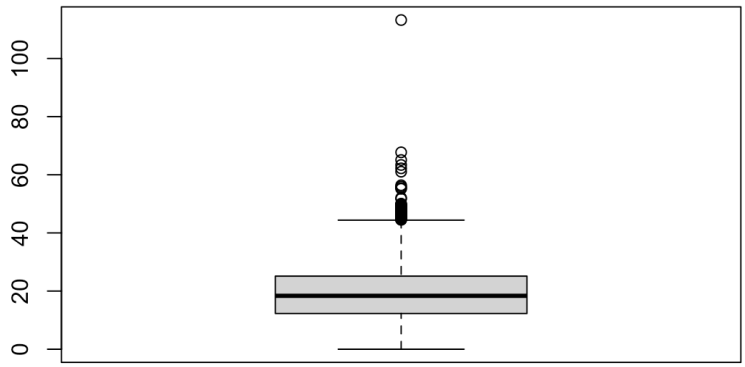
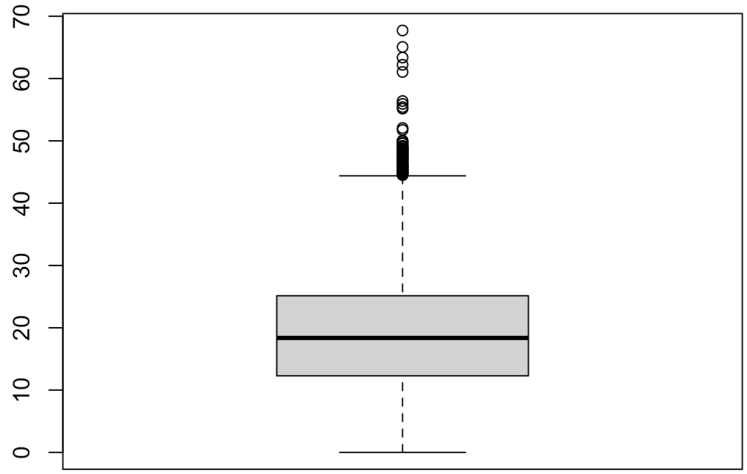


Figure 2: Box plot of Debt to income after removing the top observation



After controlling for outliers we end up having a database with 14,754 records and it is the final sub-sample we used to do all our descriptive statistics and econometrics regression.

Table 3: Transformation of variables and descriptive statistics

Variables	Value	Modality	Frequency	Proportion (%)
Loan status	Charge off	1	2954	20
	Fully paid	0	11800	80
Term	60 months	1	2992	20.3
	36 months	0	11762	79.7
Delinquency	More than 1	1	3038	20.6
	no	0	11716	79.4

This table displays the frequencies and proportions for each modalities of our binary variables.

Table 4: Descriptive statistics of continuous variables

Variables	Mean	SD	Min	Q1	Median	Q3	Max	Range
Interest rate	12.25	4.83	5.32	8.39	11.47	14.85	28.99	23.67
Debt to income	19.07	9.09	0.00	12.27	18.34	25.12	52.06	52.06
Open accounts	11.90	5.26	1.00	8.00	11.00	15.00	30.00	29.00
Mortgage accounts	1.59	1.78	0.00	0.00	1.00	3.00	8.00	8.00

This table presents some statistics for our continuous variables. ”SD” corresponds to the standard deviation, ”Min” to the minimum, ”Max” to the maximum, ”Q1” to the first quartile, and ”Q3” to the third quartile

Table 5: Result table

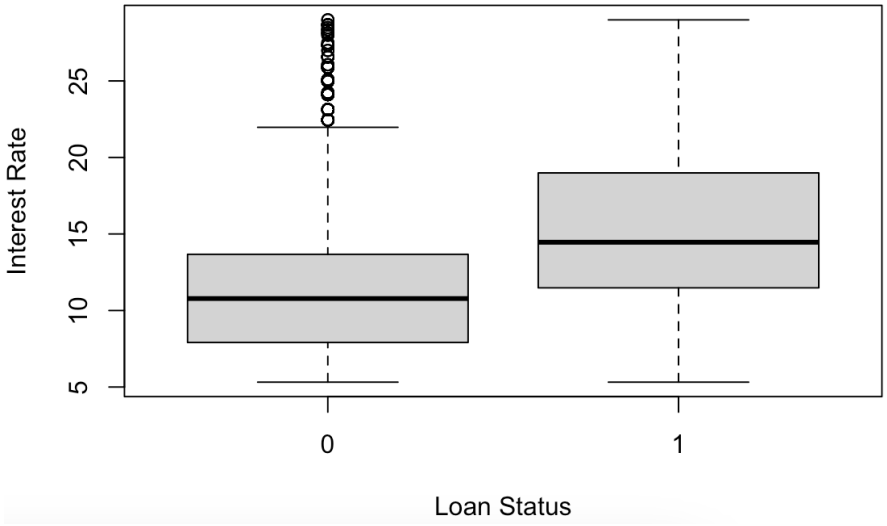
	logit	AME logit
(Intercept)	-3.412***	
	(0.084)	
Interest rate	0.120***	0.017
	(0.005)	
Term	0.672***	0.106
	(0.054)	
Debt to income	0.011***	0.002
	(0.003)	
Open accounts	0.017***	0.002
	(0.004)	
Delinquency	0.223***	0.032
	(0.054)	
Mortgage accounts	-0.143***	-0.020
	(0.014)	
N	14754	
AIC	13166.47	
BIC	13219.66	
Pseudo R2	0.16	

*** p <0.001; ** p <0.01; * p <0.05.

This table initially showcases the coefficients obtained from the logistic regression. The ”AME

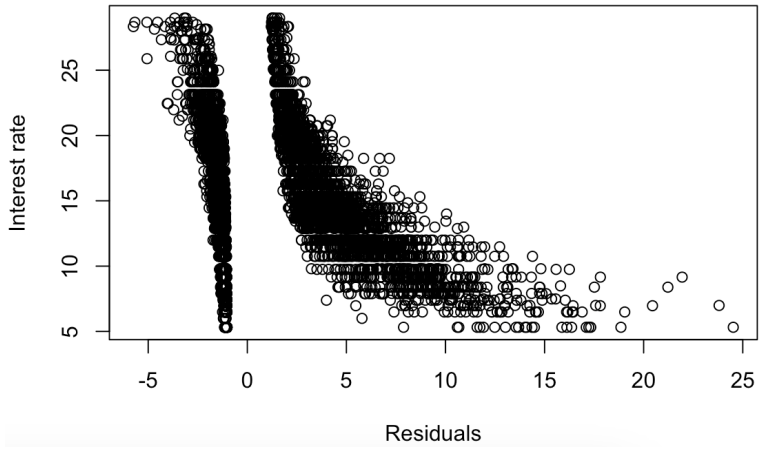
logit" column quantifies the average marginal effect of each variable. Remarkably, within the logistic model, all coefficients are deemed significant at the 1% level, marked by the presence of three stars.

Figure 3: Difference between the interest rate in the group of defaulted and the group of FullyPaid



Upon examining the box plot of the interest rate based on the two modalities of the binary variable "loan status", where 1 indicates a default and 0 indicates no default, we notice a distinct difference in the distribution of interest rates. For loans with no defaults (0), the median interest rate appears to be lower, with both the box and upper whisker extending to lower values compared to loans with defaults (1). This suggests that, on average, loans that have not resulted in defaults have been associated with lower interest rates than those that have defaulted. This observation could indicate that loans granted at lower interest rates are less likely to result in defaults.

Figure 4: Scatter Plot to explore the relationship between the Residuals and the Interest Rate



In our analysis, we do a scatter plot in R, where the y-axis represents the interest rate, and the x-axis represents the residuals obtained from our logistic regression model. Upon examination, we identify two discernible groups based on the values of the residuals.

In the first group, the residuals cluster around 0, however, for cases with higher interest rates, we notice that some residuals fell bellow that. This indicates that the model overpredicts. Conversely, in the second group, we found residuals equal to 1 and for lower interest rates it can go as far as 25. This implies that our model underpredicts for these cases. Those results suggests potential areas of improvement for our model.

Table 6: Confusion Matrix with Threshold at 0.5

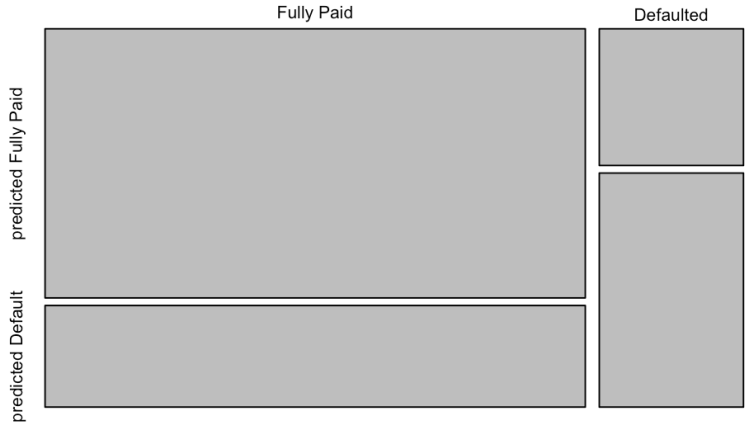
	Fully Paid	**Charged Off**
Predicted Fully Paid	0.97	0.86
Predicted Charged Off	0.03	0.14

Table 7: Confusion Matrix with Threshold at 0.2

	Fully Paid	**Charged Off**
Predicted Fully Paid	0.71	0.34
Predicted Charged Off	0.29	0.66

To do this we separate our dataframe into two subset using one for the econometric regression and the other one to predict.

Figure 5: Mozaique plots



We do a mosaic plot which allows us to easily understand visually the proportions of the confusion matrix.

References

[1] Santoso, W., Trinugroho, I. and Risfandy, T. (2020), "What Determine Loan Rate and Default Status in Financial Technology Online Direct Lending? Evidence from Indonesia", *Emerging Markets Finance & Trade*, 56:351–369.

[2] Main Uddin, Md. (2019), "Determinants of Loan Default of Low-Income Borrowers in Urban Informal Credit Markets: Evidence from Dhaka City", *European Journal of Business and Management*, Vol.11, No.26, 2019.

[3] Gudde Jote, G. (2018) "Determinants of Loan Repayment: The Case of Microfinance Institutions in Gedeo Zone, SNNPRS, Ethiopia", *Universal Journal of Accounting and Finance*, 6(3): 108-122.

[4] Edet, B.N., Atairet, E.A., Nkeme, K.K. and Sunday, U.E. (2014),"Determinants of Loan Repayment: A Study of Rural Women Fish Traders in Akwa Ibom State, Nigeria", *British Journal of Economics, Management & Trade*, 4(4): 541-550.

[5] Nadeesha, R.P.S. and Madhushani, P.W.G. (2023), "Predictors of Consumer Creditworthiness: Evidence from Personal Loan Borrowers of a Leading Public Bank in Sri Lanka", *South Asian Journal of Finance*, 3(1), 67–83.

[6] Zhichao Yin, Lei Meng, and Yezhou Sha (2020) "Determinants of agriculture-related loan default: evidence from China", *Bulletin of Monetary Economics and Banking*, pp. 129 - 150.

[7] Awunyo-Vitor, D. (2012), "Determinants of loan repayment default among farmers in Ghana", *Journal of Development and Agricultural Economics*, Vol. 4(13), pp. 339-345.

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