PREDICTING DEFAULT RISK

ON PEER-TO-PEER LENDING PLATFORM

April 16th, 2025

Group 2: Audrey Bui, Otto Graham, Cooper Bruce

Submitted in partial fulfillment of the requirements for
BUS 310 Business Analytics III: Predictive and Prescriptive Business Analytics

Abstract:

Peer-to-peer (P2P) lending platforms have reshaped the landscape of personal finance by directly connecting borrowers and lenders, thus providing easier access to credit for underserved borrowers while offering investors opportunities for higher returns. However, this model inherently involves elevated credit risk, making accurate prediction of loan default essential. This research applies several predictive analytics methods—including Ordinary Least Squares (OLS), Logistic Regression, Ridge and Lasso regression, and Random Forest—to evaluate default risk within a comprehensive dataset comprising over one million P2P loan records from 2015 to 2018. After rigorous data preprocessing and exploratory data analysis, logistic regression emerged as the most effective model due to its superior recall rate (0.446) and satisfactory accuracy (0.744) at an optimized classification threshold. Key determinants identified included loan grade, term length, FICO scores, debt-to-income ratios, loan purpose, and verification status. Findings underline the importance of balancing model accuracy and recall to minimize financial risks associated with defaults. This study provides valuable insights for investors and lending platforms, suggesting strategies to improve loan assessment and portfolio management.

I. Introduction:

Peer-to-peer lending has revolutionized the traditional money lending industry. It connects borrowers to individual lenders directly through a digital platform. This financing model is innovative and expands credit opportunities to underserved borrowers while offering individual investors to create potential opportunities of higher returns. The risk of not using the traditional banking methods there higher possibility of credit risk, making it more crucial for lenders and borrowers to accurately predict loan defaults. The world is evolving, and from that, we are seeing lots of new alternative finance methods. With these new finance methods, people need to have a solid understanding of default risk. Lenders need to account for a borrower's annual income and employment status to credit history, and loan purpose and access based on those things, the amount of risk in the potential loan. Another example of risk is non-individual factors like the economic conditions and market sentiment, these factors can also affect a person's ability to repay their loan. This paper will show you how we applied a group of predictive modeling techniques to help comprehend the peer-to-peer lending dataset. We are using ordinary least squares, logistic regression, and regularization methods such as ridge and lasso regression. Integrating these methods, we will be able to give an analysis of the key determinants of loan default. The goal is for us to develop a framework that displays default probabilities and help investors understand different investment strategies when it comes to the peer-to-peer lending market.

II. Literature Review:

In recent years, an increasing number of people are using peer-to-peer lending, which has led to an increase in literature aimed at understanding and predicting loan default risk. Iyer(2011) was the first article and it made the case that traditional lending falls short in capturing the dynamics that online peer-to-peer market has. This study lays an excellent framework for demonstrating that alternative sources like borrower demographics, online behavior and previous lending history can help figure out the predictability of loan performance and default rate. In the article they mention a website named Proser.com and it would be “where borrowers post loan listings and multiple individual lenders bid to fund a portion of the loan at a desired rate.” This gives the opportunity for the lenders to take advantage of seeing a borrower’s credit score and

creditworthiness, which is a crucial piece of data to have available. Emekter(2014) did a study on the effects of various borrower factors, such as employment status, income levels influence default rate. This article mentions the idea of the Lending Club. The Lending Club is a financial institution and marketplace that connects borrowers and lenders for various types of loans. “Lending Club screens out any potential high-risk borrowers based on the FICO score. The minimum FICO score to be able to participate is 640. The typical size of the loans produced in this market is small, which is under $ 35,000 at the Lending Club. Therefore, these loans are essentially microloans which pose a relatively small loss in case of default.” The final article from Aksakalli(2015) introduced machine learning techniques like Random Forest which highlights how it can handle complex patterns and interactions better than traditional models. “An empirical comparison reveals that RFs significantly outperform both FICO scores and LC grades in identification of the best borrowers in terms of low default probability.” Together these articles illustrate the evolution of peer-to-peer lending and the different methodologies of predicting default risk.

III. Data Section

1. Dataset Overview

The P2P lending dataset under our analysis includes 1,044,488 loan records with 114 variables from 2015 to 2018, offering comprehensive insights into borrower behavior, loan characteristics, and repayment outcomes. The original dataset reflects various personal loan applications with varying term lengths, interest rates, income levels, and credit backgrounds. The dataset includes essential features such as loan amount, interest rate, loan term, income, FICO score, DTI ratio, loan purpose, homeownership status, and final loan status. An initial review of the logistic regression results helped identify several variables that influenced default probability. Among them, loan grade emerged as a strong predictor, with Grades G, F, and E showing notably higher default rates as creditworthiness declined. Loan term proved to be an important predictor as loans with 60-month maturities were significantly more likely to default than those with shorter terms. In contrast, FICO scores had an inverse relationship as borrowers with “Very Good” or “Good” scores were considerably less likely to default. Additionally, the data demonstrated that borrowers with lower debt-to-income (DTI) ratios were less likely to default. In terms of loan

purpose, loans for debt consolidation, medical costs, renewable energy, and small company endeavors all continuously showed greater default rates, indicating that some use cases are more risky than others. Verification status was a significant factor, as loans that were labeled as "Verified" or "Source Verified" had higher default probability.

2. Data Pre-processing

Our approach to data pre-processing focused on creating a clean and robust dataset for default risk prediction. We filtered the loan records to retain only those with a definitive status: "Fully Paid," "Charged Off," or "Default." This allowed us to establish a clear binary outcome for classification. We created a new binary variable, coding both "Charged Off" and "Default" as 1 (indicating default), and "Fully Paid" as 0. The overall default rate in the dataset was 21.12%, meaning that about one in five loans resulted in a default. Next, we transformed categorical variables such as loan purpose, homeownership, loan grade, and verification status into dummy variables using one-hot encoding, making them suitable for inclusion in regression models. To better capture nonlinear correlations with default risk, we developed categorized versions of continuous variables. There were three categories for annual income: Low, Medium, and High. FICO scores were classified as Fair, Good, and Very Good, and the debt-to-income (DTI) ratio was split into five levels, from Very Low to Very High. Missing data was handled to preserve the quality of the dataset. Columns with more than 20 percent missing values were removed to ensure analytical reliability. For variables with minor missingness, mean or median imputation was applied depending on the distribution, while rows with missing values in key features were dropped when necessary. These steps ensured the final modeling dataset was complete and ready for analysis without introducing bias due to incomplete records.

IV. Exploratory Data Analysis (EDA)

We performed an exploratory data analysis (EDA) on borrowers and loan details to understand the default risk in P2P lending. This helped us identify key variables for additional research by revealing patterns in the relationships between default rates and interest rates and variables like income, debt-to-income ratio, credit score, property ownership, and loan purpose.

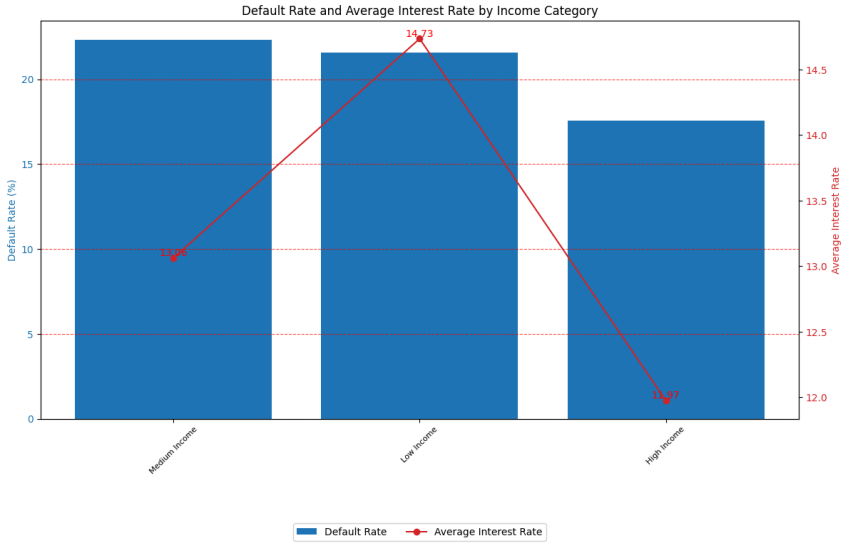

Figure 1: Default Rate and Average Interest Rate by Income Category

Figure 1 shows that borrowers with lower incomes have the highest default rates, over 25%, while those with higher incomes have lower default rates, with high-income borrowers defaulting less than 15% of the time. These borrowers also face the highest average interest rates, around 14.5%. High-income borrowers, by contrast, default less frequently and pay lower interest rates, indicating lenders view them as more creditworthy.

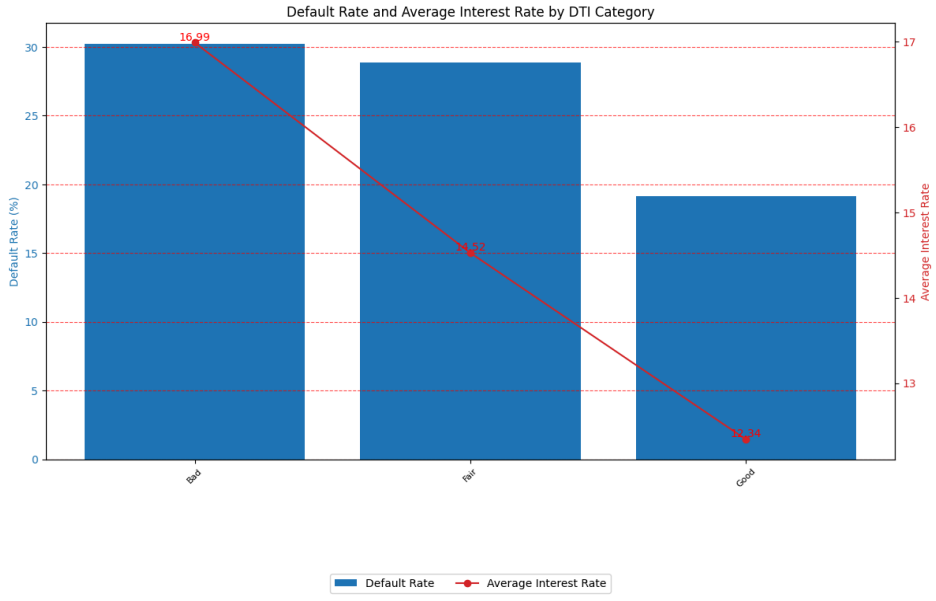

Figure 2: Default Rate and Average Interest Rate by Debt-to-Income (DTI) Category

A study found that high DTI ratios are associated with a greater incidence of mortgage default, even after controlling for other borrower and loan characteristics. (Gerardi, 2020). In Figure 2, borrowers with a "Very High" DTI have default rates above 30% and pay interest rates close to 15%. In contrast, those with lower DTI levels show default rates below 15% and benefit from lower interest rates. These results confirm that DTI is a strong indicator of credit risk.

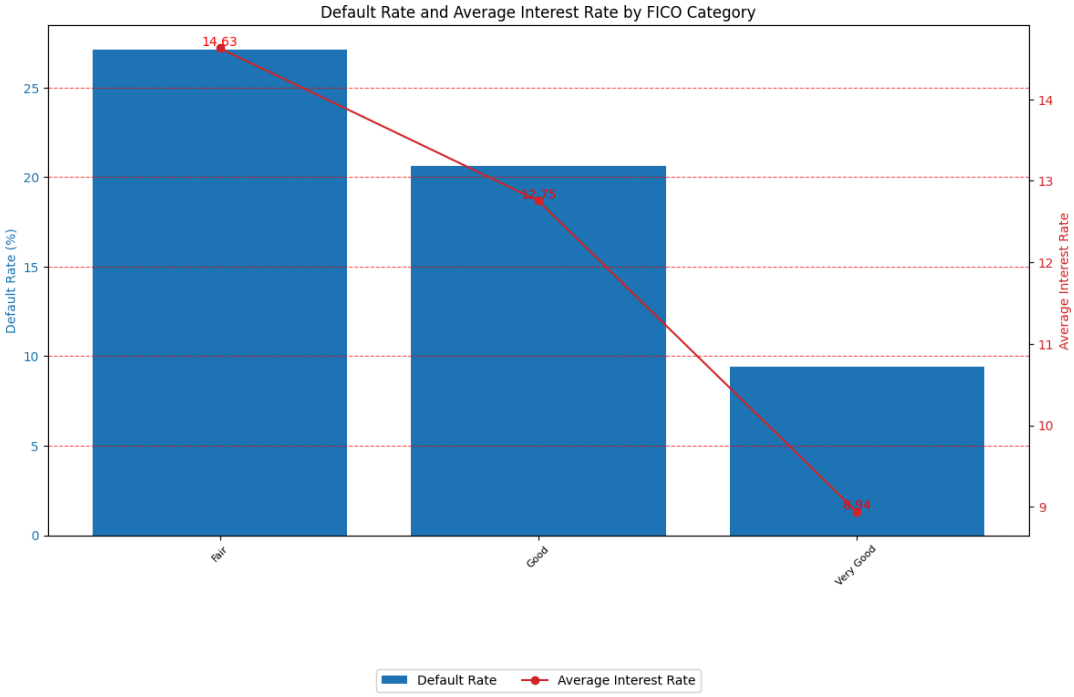

Figure 3: Default Rate and Average Interest Rate by FICO Category

Figure 3 shows that borrowers with "Fair" FICO scores default more than 25% of the time, while those with "Good" and "Very Good" have lower default rates. Average interest rates decline steadily as FICO scores improve, dropping from above 14% to about 9%. This confirms that lenders use credit scores as a key factor when assessing loan risk and setting interest rates.

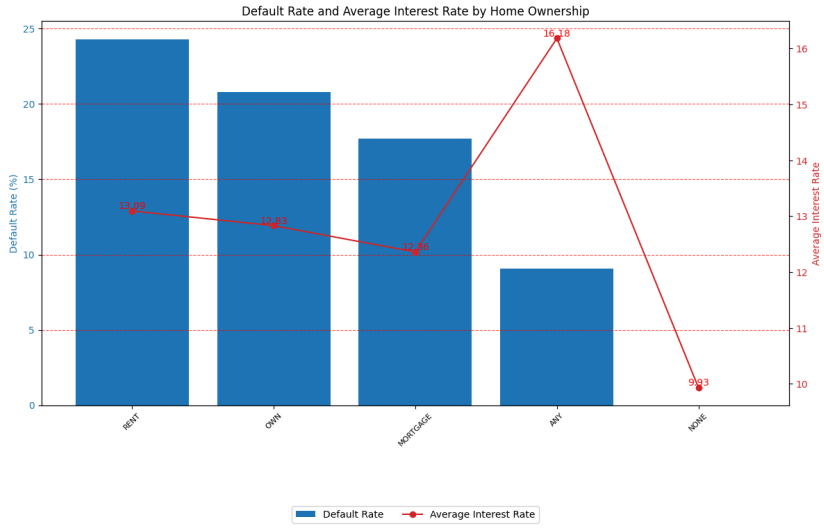

Figure 4: Default Rate and Average Interest Rate by Home Ownership

According to Figure 4, borrowers who rent have the greatest default rate of nearly 25%, while borrowers who own their homes or have mortgages have lower default rates. This implies that homeowners may be less likely to default since they are more financially stable, especially if they hold mortgages. Interest rates vary across home ownership categories, but the trend is less consistent than with other factors. Renters and those with mortgages tend to receive lower average interest rates—around 12 to 13 percent—while borrowers listed under the "ANY" category pay noticeably more, with rates above 15 percent. This suggests that lenders might associate more stable housing situations with lower risk, though the differences aren't as sharply defined as they are with income or credit score.

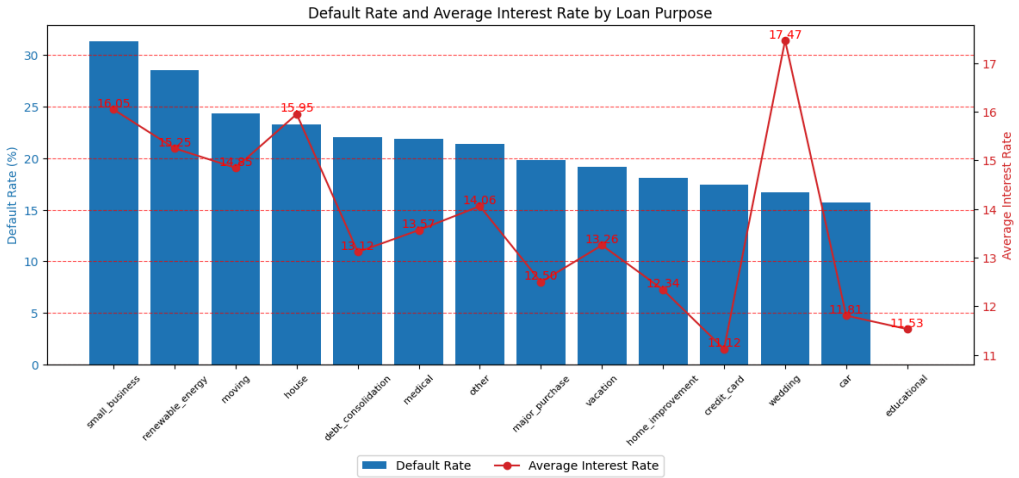

Figure 5: Default Rate and Average Interest Rate by Loan Purpose

Our findings in Figure 5 are consistent with Emekter et al. (2015), who found that loan purpose significantly influences default rates in peer-to-peer lending. They observed that loans for small businesses and medical expenses are more likely to default, while education and car loans tend to perform better. Loans taken out for small businesses have the highest default rate, topping 30%, while those used for education and car purchases tend to be much safer. Interest rates follow a similar pattern. "Wedding" loans carry the highest rates at over 17%, whereas "Educational" loans have the lowest, around 11%. This shows that lenders likely set interest rates based on how risky they think the loan's purpose is.

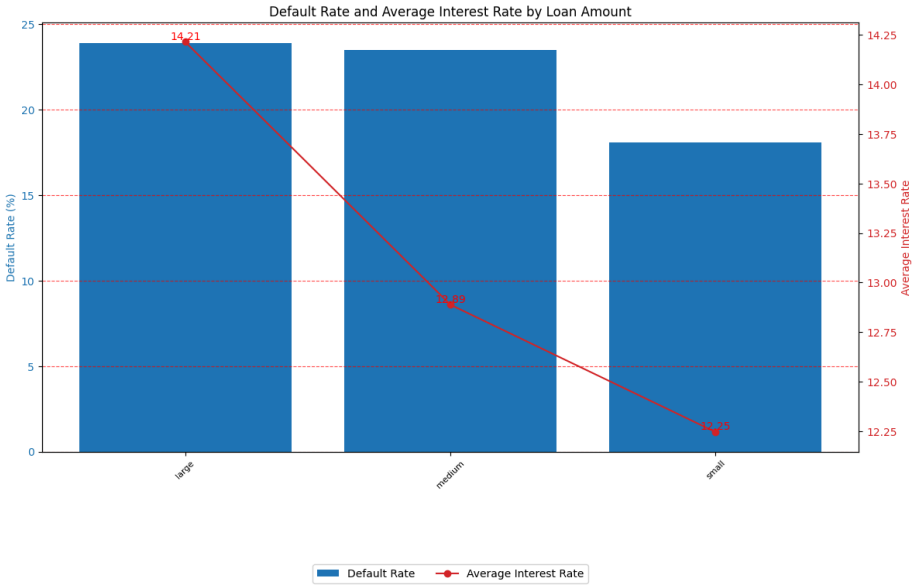

Figure 6: Default Rate and Average Interest Rate by Loan Amount

Figure 6 shows that loans between $15,000 and $25,000 had the highest rates of default. Both smaller and larger loans had lower default rates, suggesting that the extreme ends of the spectrum may attract more qualified or more cautious borrowers. Average interest rates tend to increase with loan amount at first, but then level off, possibly because lenders are pricing in additional risk for mid-sized loans. This trend reflects a relationship between what borrowers can afford to repay and how much risk lenders are willing to take on.

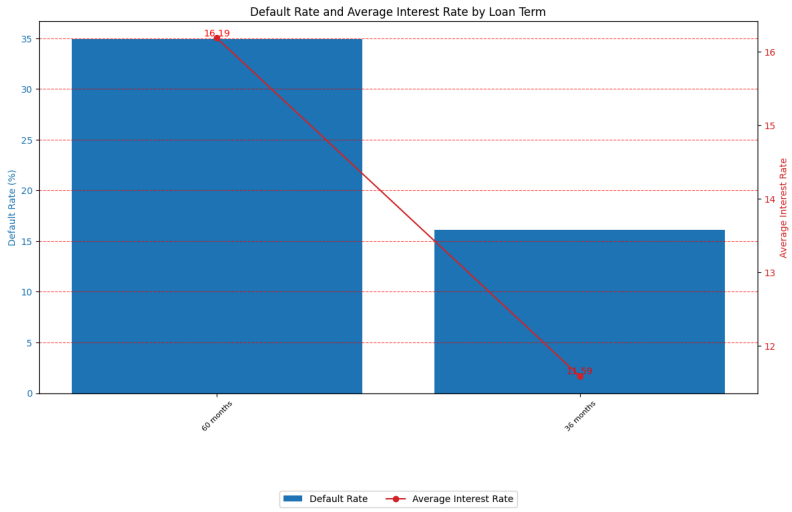

Figure 7: Default Rate and Average Interest Rate by Loan Term

Figure 7 shows that the default rates for loans with 60-month maturities are significantly higher than those for loans with 36-month periods. Due to the increased risk associated with longer payback terms, borrowers who take out longer-term loans also typically pay higher interest rates. Deferring payments for two years may seem more practical in the short term, but it also gives more time for unforeseen costs to occur. Knowing this, lenders adjust rates accordingly, especially for clients who already appear to be at a higher risk.

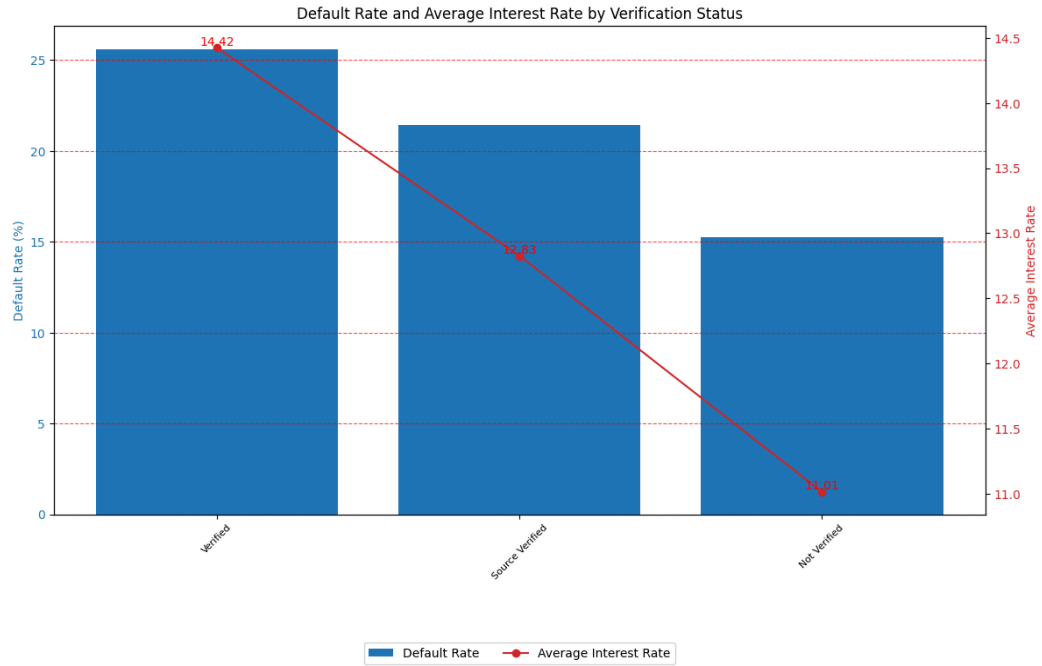

Figure 8: Default Rate and Average Interest Rate by Verification Status

Figure 8 shows that borrowers with verified income have slightly lower default rates than those with unverified income. Interest rates also tend to be lower for verified borrowers, suggesting that lenders view income verification as a positive signal of creditworthiness. The differences are not as large as other variables, and verification status still appears to play a role in how lenders assess and price risk.

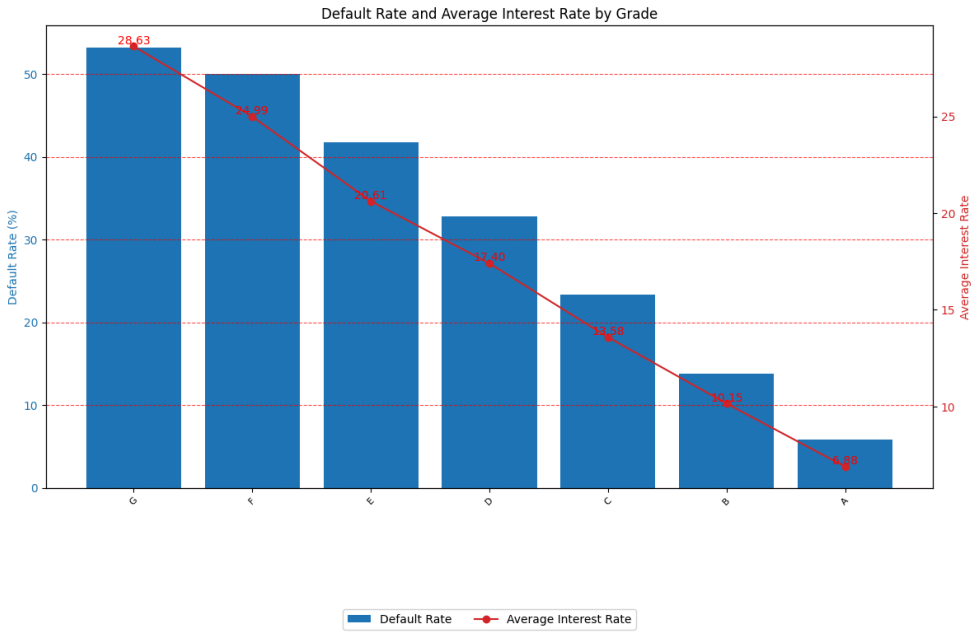

Figure 9: Default Rate and Average Interest Rate by Grade

We observed in Figure 9 that default rates increase steadily from Grade A to Grade G, with Grade G loans showing the highest likelihood of default. Interest rates follow the same pattern, rising consistently across the grades. This trend reflects how LendingClub's internal grading system aligns with credit risk, as lower-grade borrowers face both higher chances of default and higher borrowing costs. These patterns support the use of loan grade as a key predictor in default risk modeling.

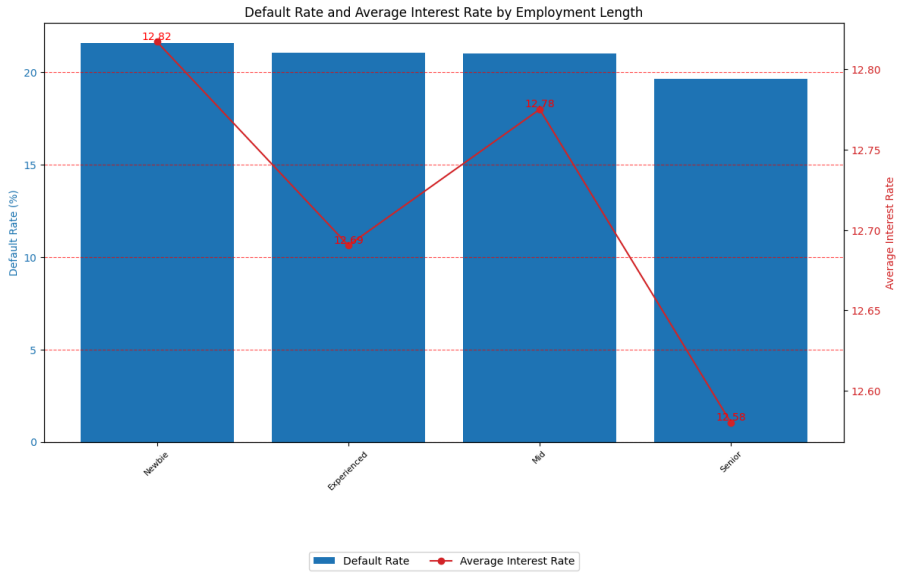

Figure 10: Default Rate and Average Interest Rate by Employment Length

Figure 10 shows that borrowers with shorter employment histories tend to exhibit slightly higher default rates compared to those with longer tenures. However, the variation in interest rates across employment lengths is minimal. This suggests that while employment length may provide some insight into borrower stability, it is not a primary factor in risk assessments. A study analyzing LendingClub data found that employment length had a negligible effect on default prediction, indicating that lenders may prioritize other borrower attributes when evaluating risk (Emekter et al.).

V. Method Section

For us to access the default rate in our dataset, we used a group of statistical and machine learning models. The models we used were Ordinary Least Squares, Logistic regression, Ridge regression, Lasso regression, and a Random Forest model. Each method was chosen to perform a specific task in our framework and provides analytical insights into the relationship between borrower trends and default probability.

Ordinary Least Squares is the baseline model for the analysis. We used OLS to help us understand the linear relationship between the predictor and the outcomes of either default or no

default. This step let us examine the significance and direction of the effects of key variables. This gives us key information to use for more analytical and classification-oriented models.

Logistic regression is a natural choice for yes or no questions, like predicting if a specific loan will default. It models probability and estimates the odds that a loan is classified as a default. Missing a defaulting borrower is costly, so we pay attention to the recall rate. A high recall rate indicates that the model is successfully identifying a large proportion of default cases.

Ridge and Lasso regression are both used to help control multicollinearity and overfitting in high-dimensional data. Ridge regression adds a penalty to the sum of squared coefficients, which helps shrink the less important variables without setting them to exactly zero, so it doesn't eliminate them. Lasso regression also applies a penalty, except it applies to all coefficients and shrinks some to zero. This effectively selects the most relevant features, which further simplifies the model. Similar to logistic regression, we again focus on the recall rate, ensuring we catch as many high-risk default loans.

We then use the Random Forest model to catch more non-linear patterns and interactions between the variables. Random Forest generates multiple decision trees, and it aggregates the prediction results, which then improves the accuracy. The importance of correctly finding defaults is so crucial; once again, the recall metric is used to measure this. We aim to maximize the true positive rate and minimize false positives because misclassifying a defaulting borrower as non-default can lead to financial consequences. In summary, our frameworks integrated simple and more modern modeling techniques to analyze the default rate in our dataset. By focusing on recall across a majority of our models, we prioritize accurately finding default rates.

VI. Results

Our selected predictors show that none of the pairwise correlations are notably high. This finding is encouraging from a modeling standpoint, as lower correlation levels help mitigate potential multicollinearity issues and ensure that each variable contributes independently to the regression analysis.

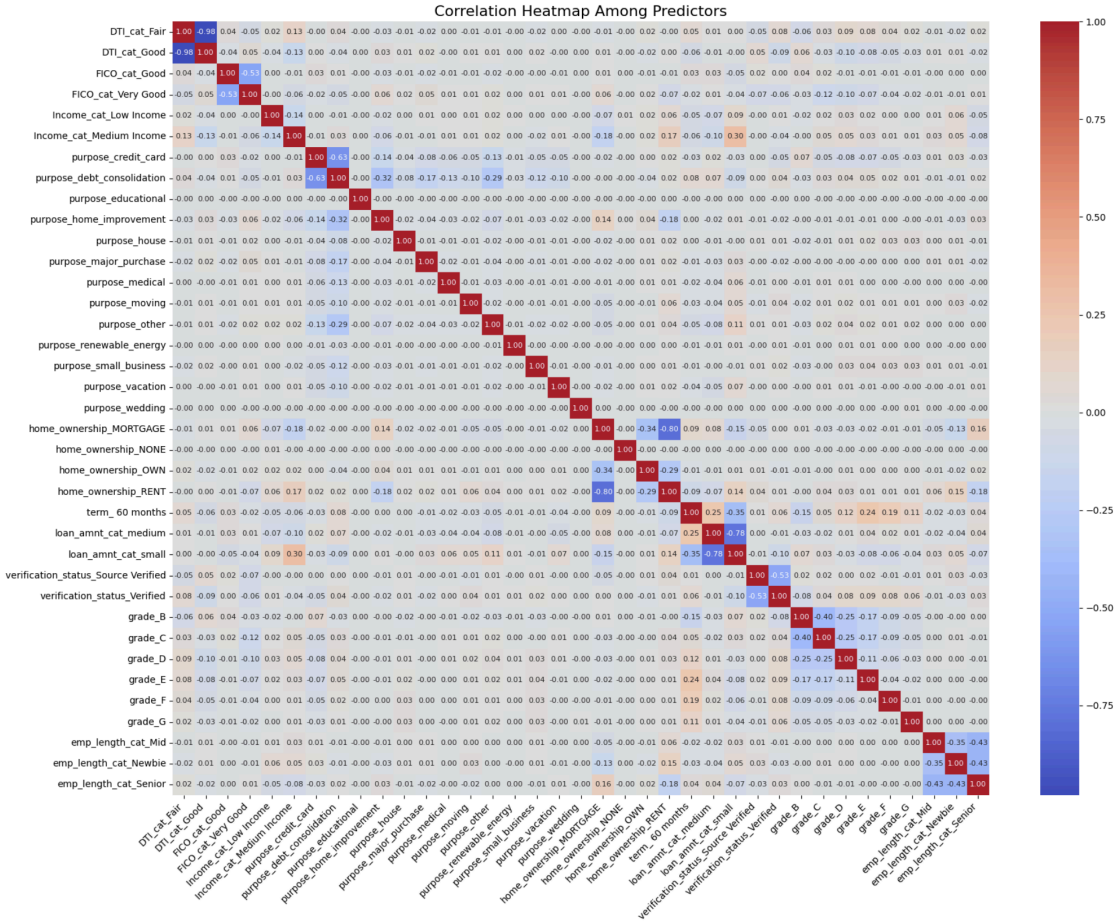

Figure 11: Heatmap of chosen predictors' correlation

1. Ordinary Least Square (OLS) Regression

OLS Regression Results

Dep. Variable:	loan_default_dummy	R-squared:	0.100
Model:	OLS	Adj. R-squared:	0.100
Method:	Least Squares	F-statistic:	1729.
Date:	Sun, 13 Apr 2025	Prob (F-statistic):	0.00
Time:	22:08:45	Log-Likelihood:	−2.6499e+05
No. Observations:	573237	AIC:	5.300e+05
Df Residuals:	573199	BIC:	5.305e+05
Df Model:	37		
Covariance Type:	nonrobust		

	Model	MSE_Train	MSE_Test
0	linear	0.14762	0.147457

Figure 12: OLS Regression Results & MSEs

As a preliminary step, an Ordinary Least Squares (OLS) regression was conducted despite the fact that loan_default_dummy is a binary outcome rather than a continuous variable. The rationale behind this choice was primarily exploratory. The model's R-squared value is 0.10, suggesting that only about 10% of the variation in default rates can be explained by the current dataset when using a linear framework. Additionally, the Mean Squared Error (MSE) on the training set is 0.14762, and on the testing set is 0.147457, indicating minimal divergence between training and testing performance. While this might suggest that the model is not overfitting, it does not mitigate the fundamental issue that OLS is not well-suited to binary classification.

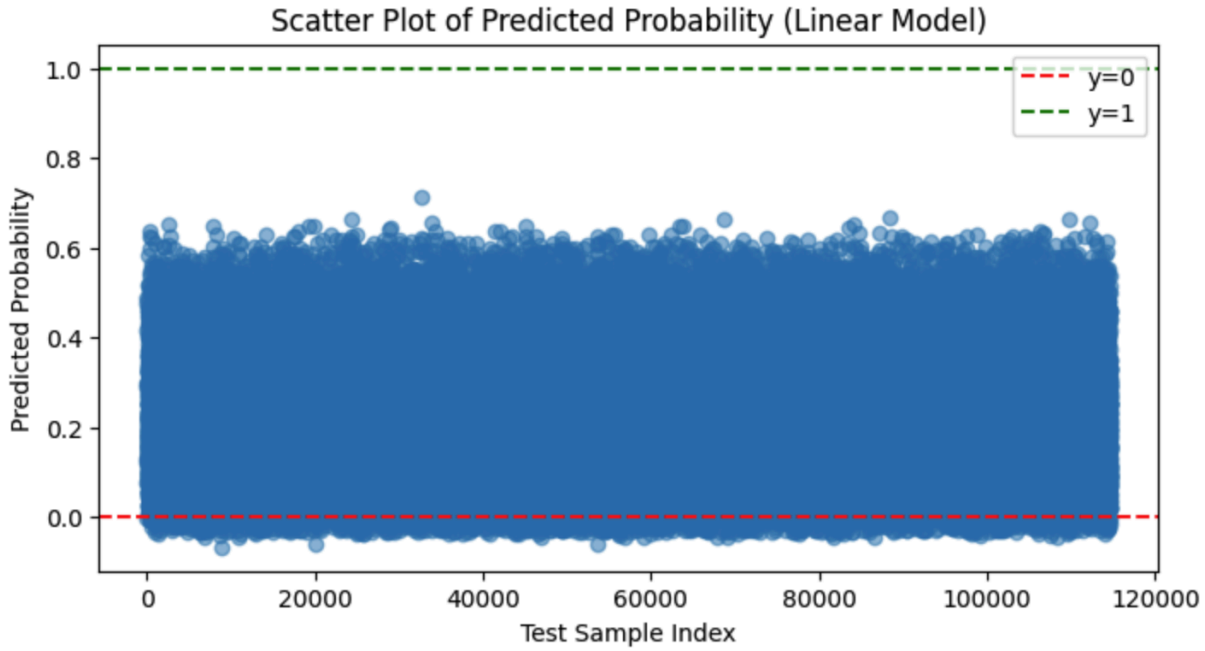

Figure 13: Scatter Plot of Predicted Probability (OLS)

By design, linear regression predicts continuous outcomes and can output values beyond the valid range of [0,1] for a probability. In contrast, classification tasks require discrete labels (default/non-default) or valid probability estimates (strictly between 0 and 1). Consequently, although it can be instructive as an initial benchmark, OLS offers limited utility for discerning and explaining default risks. Therefore, while the OLS approach highlights some variables that may correlate with default, it falls short as a classification model.

2. Logistic Regression

Our logistic regression analysis points to several key factors that significantly influence default risk on this P2P lending platform. On the borrower side, Very Good FICO scores emerge as the strongest negative predictor (coefficient = –0.5189, $p < 0.001$), indicating that these borrowers have a substantially lower likelihood of defaulting. Borrowers with Good DTI ratios also display lower default risk (coefficient = –0.1781, $p < 0.001$). Conversely, Medium-income status corresponds to a notably higher risk (coefficient = 0.1646, $p < 0.001$), while Low-income status shows no significant relationship with default ($p = 0.930$).

Regarding loan characteristics, the 60-month term is linked to a substantial increase in default probability (coefficient = 0.5586, $p < 0.001$). Both medium (coefficient = 0.0542, $p < 0.001$) and small loan amounts (coefficient = –0.1160, $p < 0.001$) are statistically significant; however, smaller loans appear less prone to default. The platform's grading system offers critical insight, with coefficients rising from Grade B (0.7145) to Grade G (2.1323)—all significant at $p < 0.001$—underscoring Grade G as the highest-risk category.

In addition, loan purposes likewise play a considerable role. Small business loans (coefficient = 0.4167, $p < 0.001$) demonstrate the greatest default risk among the listed categories, followed by renewable energy (coefficient = 0.3769, $p = 0.004$) and medical expenses (coefficient = 0.2048, $p < 0.001$). Other purposes—credit card, debt consolidation, home improvement, major purchases, moving, other, and vacation—also show significant positive coefficients (all $p < 0.05$). Meanwhile, educational, house-related, and wedding loans are not significantly related to default.

Furthermore, verification status and employment length further refine our understanding of default risk. Source Verified (coefficient = 0.1276, $p < 0.001$) and Verified (coefficient = 0.1749, $p < 0.001$) loans both show higher default probability compared to unverified loans. Among employment categories, only Senior status (coefficient = –0.0464, $p < 0.001$) is tied to reduced default risk; mid-career and new employees exhibit no significant effect. Interestingly, home ownership—whether mortgage, own, or rent—does not emerge as a significant determinant of default (all p-values > 0.05).

In the lending context, recall (or the true positive rate) becomes crucial because it focuses on identifying borrowers who are likely to default. A missed default, which is a false negative, can be extremely costly, as the lender risks losing not only the principal but also any interest and administrative costs associated with the loan. By emphasizing recall, the model flags a greater proportion of genuinely high-risk borrowers, thereby reducing the odds of accidentally approving loans to individuals who will charge off. Although a higher recall typically comes with trade-offs in other metrics, such as precision or overall accuracy, the potential financial impact of undetected defaults often justifies prioritizing recall in credit risk decisions.

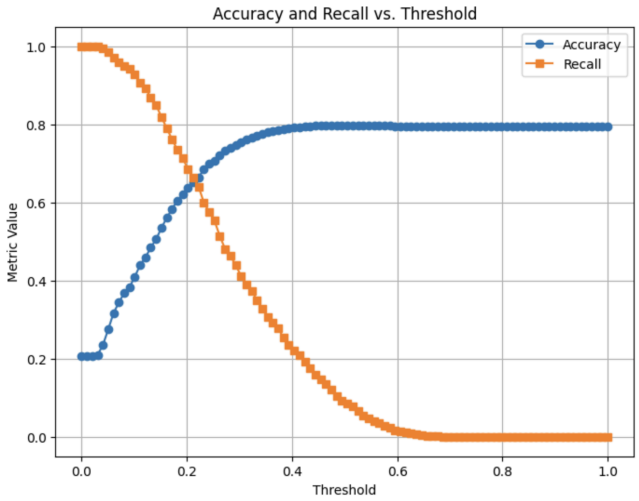

Figure 14: Accuracy and Recall with different threshold

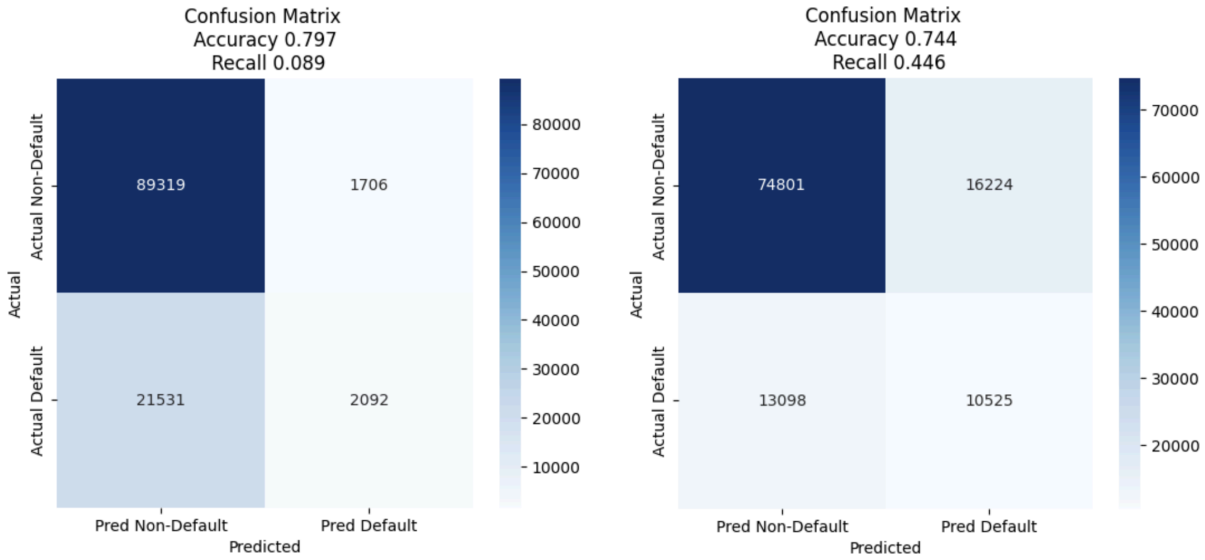

Figure 15: Confusion Matrix of Logistic Regression with 0.5 threshold (left) and 0.3 threshold (right)

There is a clear trade-off between accuracy and recall when adjusting the classification threshold. At the default threshold of 0.5, the model's accuracy stood at 0.797, yet it managed to capture only 8.9% of actual defaults (recall = 0.089). Because missing defaults can be extremely costly, we lowered the threshold to 0.3, which reduced accuracy to 0.744 (a 6.6% drop) but boosted recall to 0.446—representing a 401% improvement in identifying true defaults. By lowering the threshold, we substantially boosted the model's ability to flag potential defaults, with true positives rising from 2,092 to 10,525. Although false positives also climbed from 1,706 to 16,224, this trade-off is generally acceptable in a lending context, where the cost of missing a likely defaulter far outweighs the downside of declining a potentially creditworthy borrower.

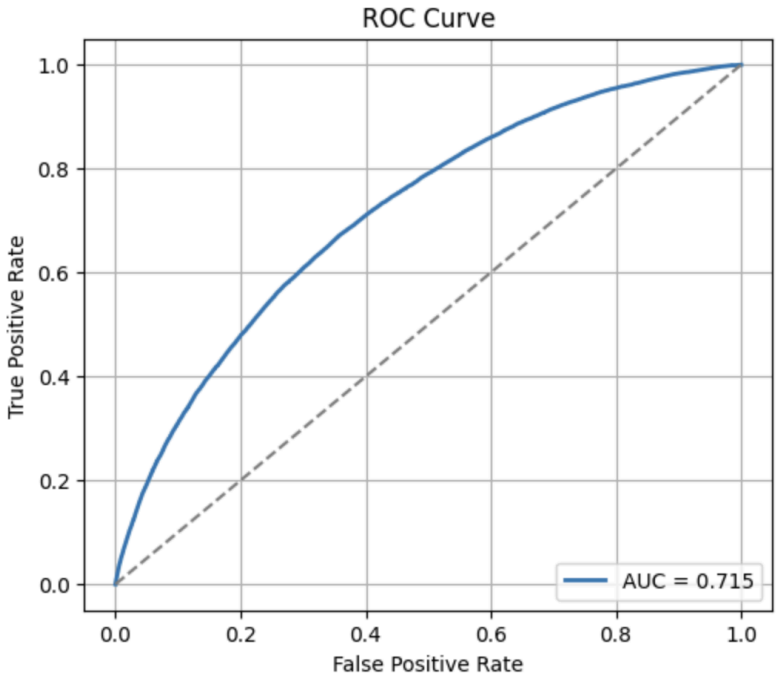

Figure 16: Logistic Regression ROC Curve

The ROC curve charts the true positive rate against the false positive rate for various threshold values. In this study, our model attained an AUC of 0.715, placing it in the "fair" category ($0.7 \leq$ AUC < 0.8) (Abdou et al., 2016). This level of performance signals that the model can meaningfully differentiate between defaulting and non-defaulting loans, outperforming random guessing (represented by the diagonal line). Nonetheless, the curve's shape suggests there is still room for advancement before reaching "good" or "excellent" classification standards.

3. Ridge Regression

Ridge regression departs from standard logistic regression by introducing an L2 penalty, which shrinks coefficient values toward zero based on their magnitude. This regularization strategy helps control overfitting by reducing the influence of less impactful features. While logistic regression aims solely to maximize likelihood, ridge balances that objective with a penalty term regulated by the hyperparameter alpha. Larger alpha enforces stronger regularization, potentially lowering model variance but increasing bias.

For the ridge regression analysis, we sampled 20,000 observations to lessen the computational overhead and cut down on the model's running time.

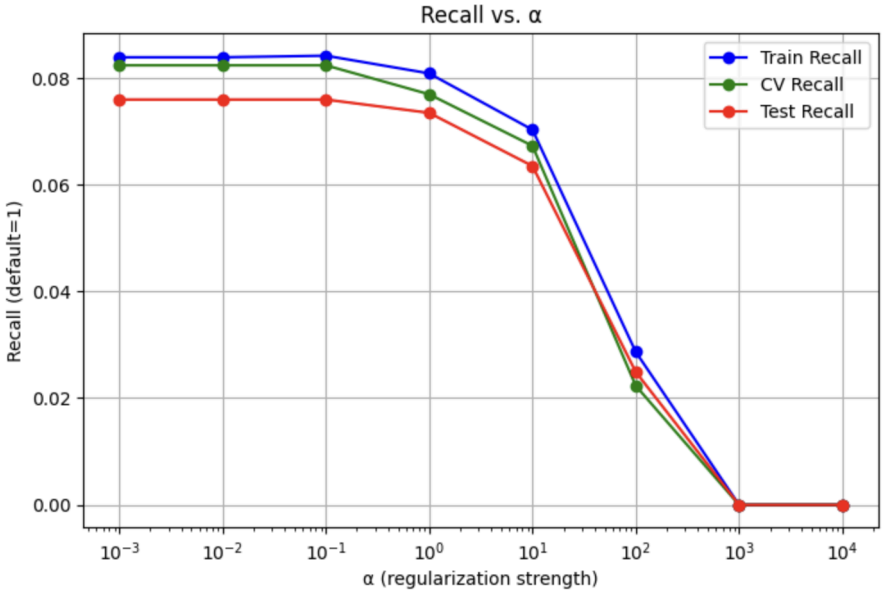

Figure 17: Recall with different alpha values (Ridge regression)

To calibrate the strength of regularization, we experimented with various alpha values, ranging from 0.001 to 10000, and chose the one that maximized the cross-validation (CV) recall without compromising accuracy. Our findings indicated that alpha = 0.01 provided the most suitable trade-off, implying that minimal regularization was sufficient given the already robust predictive value of our features.

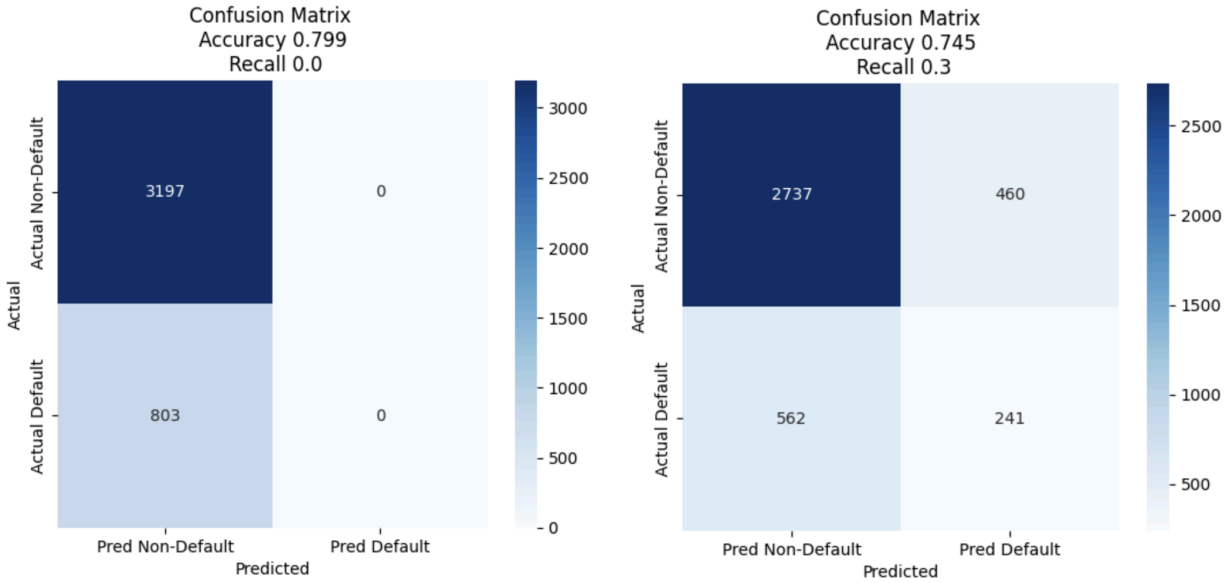

Figure 18: Confusion Matrix of Ridge Regression with 0.5 threshold (left) and 0.3 threshold (right)

With alpha set to 0.001, the ridge model initially recorded an accuracy of 0.799 but a recall of 0.0 at the default threshold of 0.5. While this accuracy might seem impressive, the model failed to detect a single default, missing all 803 true defaulters, a relatively bad outcome in lending. Lowering the threshold to 0.3 bumped recall up to 0.3, although accuracy dipped to 0.745 (a 6.8% decrease). At this new threshold, the model correctly identified 241 defaults, reducing false negatives from 803 to 562. In addition, the AUC of 0.679 places its discriminative power in the "poor" range ($0.60 \leq \text{AUC} < 0.70$), reflecting limited ability to distinguish between defaulting and non-defaulting loans.

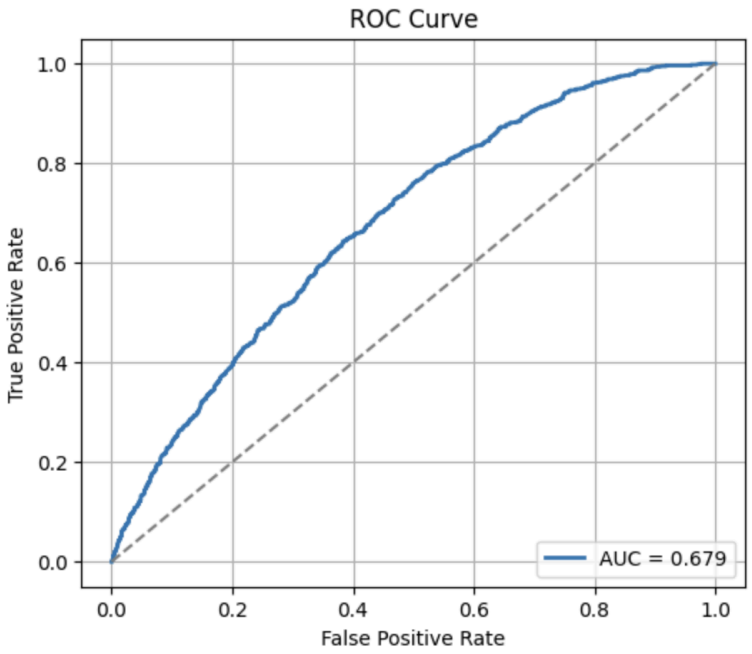

Figure 19: *Ridge Regression ROC Curve*

When matched against our logistic regression model under comparable threshold adjustments, ridge regression shows nearly equivalent performance in terms of accuracy (0.745 vs. 0.744) but has a lower recall (0.3 vs. 0.446). Its AUC also drops from 0.715 (logistic) to 0.679 (ridge). However, it is important to note that ridge regression was run on a much smaller sample for improved runtime, so the direct comparison may not fully reflect each model's capabilities on the broader dataset.

4. Lasso Regression

Lasso regression differs from standard logistic regression by incorporating L1 regularization, which not only shrinks the coefficients toward zero but can also set some coefficients exactly to zero, effectively selecting the most relevant predictors and eliminating unnecessary ones. This makes the lasso method particularly suitable for datasets with potentially redundant or irrelevant features, helping to simplify the model and reduce overfitting.

For the lasso regression analysis, we similarly sampled 20,000 observations to reduce computational overhead and improve the model's runtime efficiency.

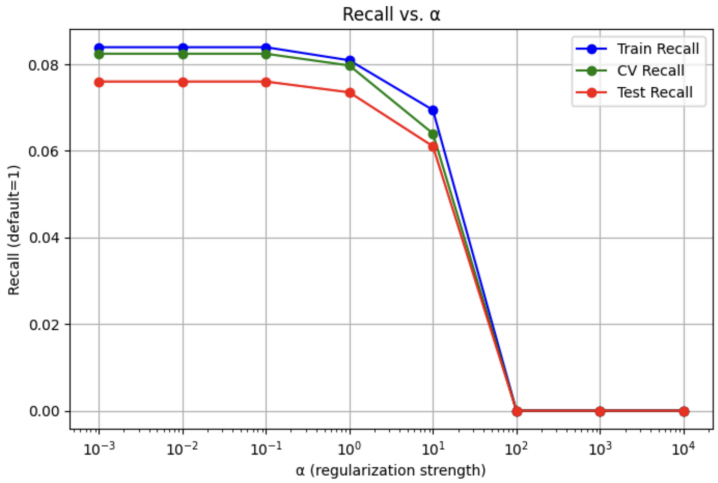

Figure 20: Recall with different alpha values (Lasso regression)

The optimal lasso model was achieved at alpha = 0.001, providing the best balance of recall and accuracy. At the default threshold of 0.5, the lasso model achieved an accuracy of 0.799 but a recall of 0.0, classifying all loans as non-default and missing all 803 actual defaults. Lowering the threshold to 0.3 resulted in a significant recall improvement to 0.3 (correctly identifying 241 defaults) while slightly decreasing accuracy to 0.745. At this adjusted threshold, false negatives decreased from 803 to 562, but false positives increased to 460.

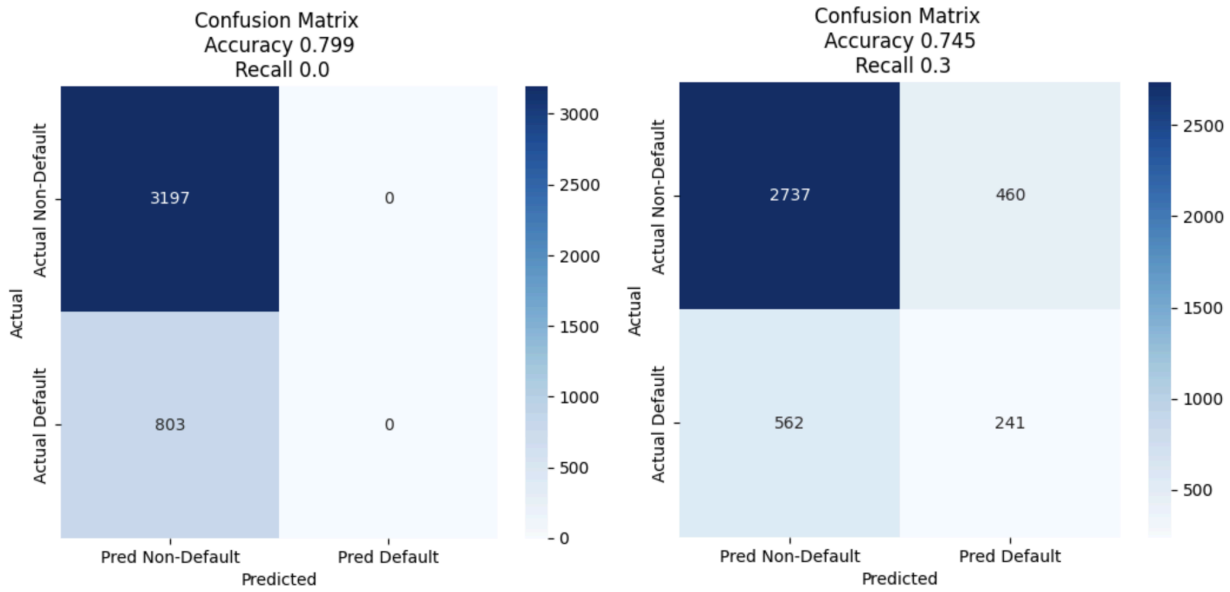

Figure 21: Confusion Matrix of Lasso Regression with 0.5 threshold (left) and 0.3 threshold (right)

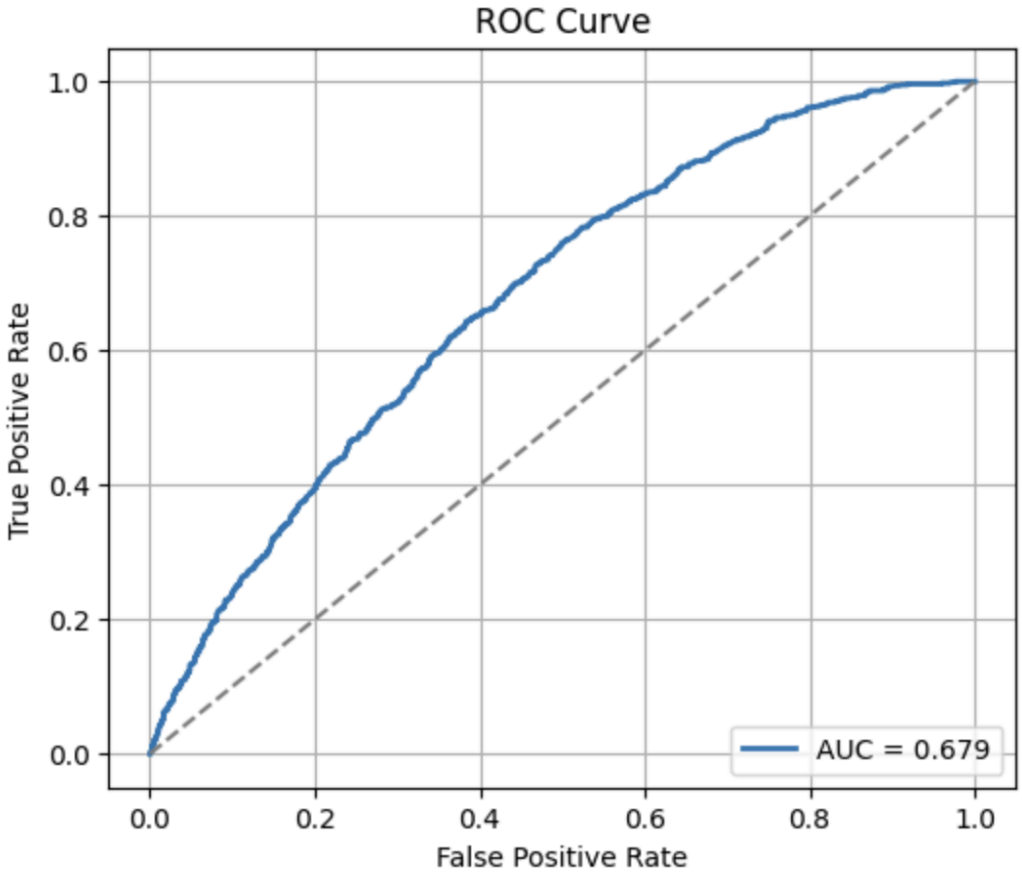

Figure 22: Lasso Regression ROC Curve

When compared with ridge and logistic regressions at the same threshold (0.3), the lasso regression achieved nearly identical accuracy (0.745, compared to 0.745 for ridge and 0.744 for logistic). However, its recall of 0.3 was lower than logistic regression (0.446) but matched the ridge model (0.3). Additionally, lasso's AUC score was 0.679, lower than logistics (0.715) and matching ridge's (0.679). It's crucial to highlight that, like ridge regression, lasso was run on a smaller sample (10,000 observations) to improve runtime efficiency, making direct comparisons to the full logistic model approximate rather than exact. Overall, while lasso regression effectively simplified the model by feature selection, it offered no clear advantage in predictive performance compared to the ridge model and performed slightly worse in terms of recall relative to the logistic model.

5. Random Forest

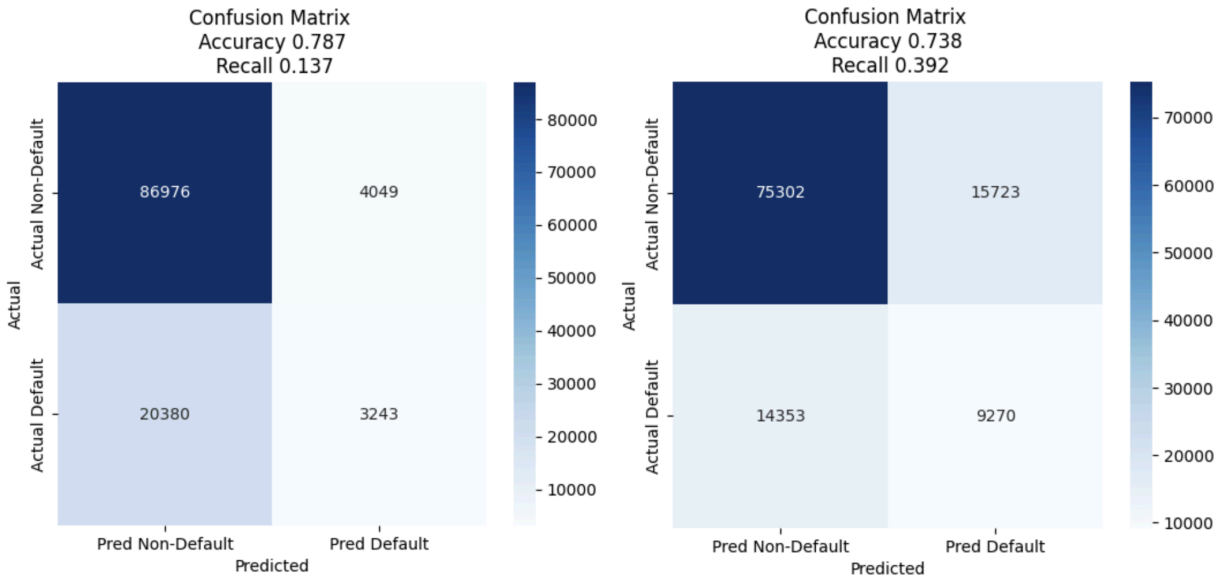

Figure 23: Confusion Matrix of Random Forest with 0.5 threshold (left) and 0.3 threshold (right)

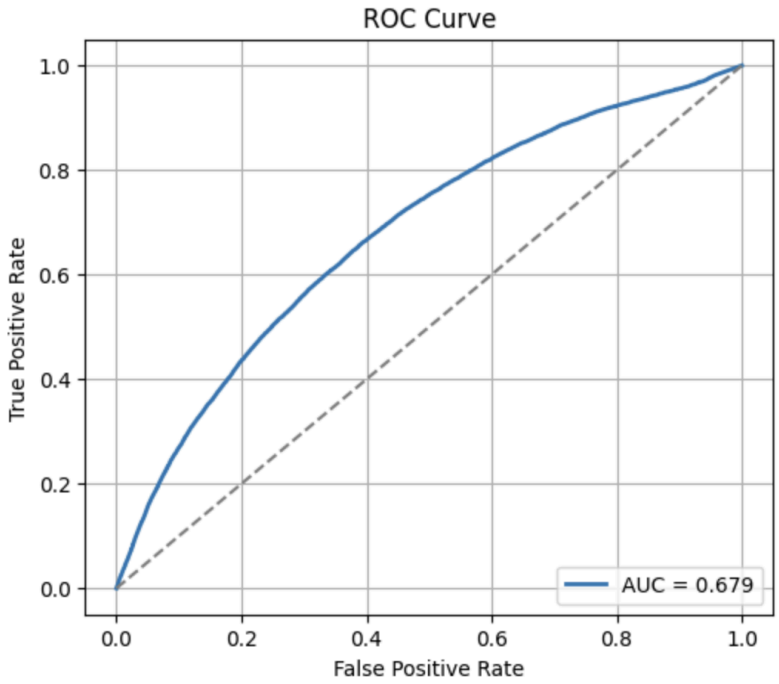

Figure 24: Lasso Regression ROC Curve

The Random Forest model initially achieved an accuracy of 0.787 but a recall of only 0.137 at the default threshold of 0.5, correctly identifying just 3,243 defaults while missing 20,380 true defaulters. Lowering the threshold to 0.3 improved recall significantly to 0.392, identifying 9,270 defaults and reducing missed defaults to 14,353, although accuracy dropped to 0.738.

In terms of overall discriminative ability, the Random Forest model achieved an AUC of 0.679, placing it in the "poor" performance category, this is identical to the ridge and lasso models (0.679) and slightly below logistic regression (0.715). When directly compared at the adjusted threshold (0.3), Random Forest's accuracy (0.738) is similar but slightly lower than Ridge (0.745) and Logistic (0.744) regression. However, Random Forest offers a notably higher recall (0.392) compared to Ridge and Lasso (0.3), though still below the Logistic model's recall (0.446). Overall, Random Forest provides a moderate balance between accuracy and recall, with performance roughly equivalent to previous models, demonstrating a reasonable but limited capability to distinguish between defaulters and non-defaulters.

6. Model Comparison

Regression Metrics (Threshold = 0.5)

	Model Type	Accuracy	Precision	Recall	F1-Score	AOC
0	ols	0.796900	0.563918	0.063116	0.113526	0.714577
0	logistic	0.797319	0.550816	0.088558	0.152584	0.715414
0	lasso	0.799250	0.000000	0.000000	0.000000	0.678836
0	random_forest	0.786922	0.444734	0.137281	0.209801	0.678909

Regression Metrics (Threshold = 0.3)

	Model Type	Accuracy	Precision	Recall	F1-Score	AOC
0	ols	0.748884	0.399298	0.433645	0.415764	0.714577
0	lasso	0.744500	0.343795	0.300125	0.320479	0.678836
0	lasso	0.744500	0.343795	0.300125	0.320479	0.678836
0	random_forest	0.737667	0.370904	0.392414	0.381356	0.678909

To assess which model offers the best prediction of default risk, we compared four different modeling approaches: OLS, Logistic regression, Lasso regression, and Random Forest. We evaluated model performance primarily on recall and accuracy, focusing especially on recall since the main goal in lending decisions is to minimize the risk of failing to identify borrowers who may default.

At the default threshold of 0.5, all models exhibited relatively high accuracy (around 0.79). However, recall was consistently low, making these models inadequate for practical decision-making, as they failed to detect a large portion of actual defaulters. Reducing the threshold to 0.3 significantly improved recall for all models, allowing for more balanced trade-offs between correctly identifying defaults and maintaining acceptable accuracy.

Comparing models at the adjusted threshold (0.3), Logistic Regression emerged as the best-performing model with a recall of 0.446 and accuracy of 0.744, as well as the highest AUC (0.715), which places its predictive performance in the "fair" category. Although OLS showed similar accuracy (0.749) and comparable recall (0.434), linear regression is inherently unsuitable for binary classification problems due to its theoretical limitations. Both Lasso and Random Forest regressions performed similarly with slightly lower accuracy (roughly 0.744 and 0.738 respectively) and notably lower recall (approximately 0.30 and 0.39 respectively). Furthermore, Lasso and Random Forest models exhibited lower AUC scores (around 0.679), indicating weaker overall discriminative power.

Given these results, we conclude that the Logistic Regression model, with its highest recall, acceptable accuracy, and strongest discriminative power, is the most suitable for predicting default risk on the P2P lending platform.

7. Important Variables

Based on the logistic regression coefficients and significance tests, the key predictors influencing default probability were identified:

- Loan Grade: Particularly Grades G, F, and E, with significantly higher default likelihood as grades worsened.
- Loan Term: 60-month loans significantly increased default risk compared to shorter terms.

- FICO Score: Higher (Very Good, Good) FICO scores strongly reduced default risk.
- Debt-to-Income (DTI) Ratio: Good DTI ratios substantially decreased default probability.
- Loan Purpose: Loans for small businesses, renewable energy, medical expenses, and debt consolidation had notably higher default risks.
- Verification Status: Verified and source-verified loans were associated with increased default probabilities.

VII. Conclusion

In this study, we used a variety of machine learning models, such as logistic regression, ridge and lasso regression, and random forest. Logistic regression offered a fundamental understanding, while other models like random forest achieved greater accuracy in classifying defaults.

Exploratory visualizations highlighted that loan grade, income level, FICO score, and debt-to-income (DTI) ratio were influential, with lower credit grades and higher DTI ratios corresponding to substantially elevated default rates. These findings were consistent with broader financial research. Gerardi and Fuster (2020) emphasize the predictive power of DTI, noting that borrowers with high DTI ratios are significantly more likely to default, especially during periods of economic stress. This external evidence reinforces the validity of our variable selection and supports the integration of these features into predictive modeling. By aligning data-driven trends with findings from the literature review, our analysis ensured that the models were informed by both statistical rigor and real-world lending dynamics.

Our study emphasizes the value of combining machine learning techniques with conventional statistical methods to improve lending prediction accuracy. The combination of these models enables more comprehensive risk assessment, assisting lenders in making data-driven decisions. Future research could concentrate on experimenting with more advanced models to enhance loan portfolio management.

Malekipirbazari and Aksakalli (2015) found that random forest models outperform traditional methods in predicting defaults within P2P lending platforms. For future research and model refinement, we suggest greater emphasis on ensemble methods like Random Forests to capture nonlinear relationships and improve predictive accuracy in identifying high-risk borrowers.

VIII. References

1. Gerardi, Kristopher, and Andreas Fuster. *How Do Mortgage Payments Affect Consumer Spending?*. Federal Reserve Bank of Dallas, 24 Mar. 2020,
2. Emekter, Riza, et al. "Evaluating credit risk and loan performance in online Peer-to-Peer (P2P) lending." *PLOS ONE*, vol. 10, no. 6, 2015, e0139427.
3. Abdou, H. A., Dongmo Tsafack, M. D., Ntim, C. G., & Baker, R. D. (2016). Predicting Creditworthiness in Retail Banking with Limited Scoring Data. Knowledge-Based Systems, 103, 89-103.
4. https://www.journals.elsevier.com/knowledge-based-systems https://doi.org/10.1016/j.knosys.2016.03.023
5. Emekter, R., Tu, Y., Jirasakuldech, B., & Lu, M. (2014). Evaluating credit risk and loan performance in online Peer-to-Peer (P2P) lending. *Applied Economics*, *47*(1), 54–70. https://doi.org/10.1080/00036846.2014.962222
6. Iyer, Rajkamal & Ijaz, Asim & Erzo, Khwaja & Luttmer, F & Shue, Kelly & Fisman, Raymond & Gentzkow, Matthew & Katz, Lawrence & Atif, Raghab & Ravina, Enrichetta & Scharfstein, David & Shapiro, Jesse & Stein, Jeremy. (2011). Inferring Asset Quality: Determining Borrower Creditworthiness in Peer-to-Peer Lending Markets.
7. Milad Malekipirbazari, Vural Aksakalli, Risk assessment in social lending via random forests, Expert Systems with Applications, Volume 42, Issue 10,
8. 2015, Pages 4621-4631, ISSN 0957-4174, https://doi.org/10.1016/j.eswa.2015.02.001.