GROUP ASSIGNMENT 2 | GROUP 1

Predicting Default Risk on Peer-to-Peer Lending Platform

Google Colab Notebook: BUSA310 - Group 6.ipynb

BUSA 310 Business Analytics III: Predictive and Prescriptive Business Analytics

Fall 2024 | November 10, 2024

Sota Fujii, Holten Heldebrand, Chan-Tran Le, Quan Nguyen, Hien Anh (Annie) Tran

Abstract

This paper explores the predictive modeling of default risk and returns in peer-to-peer (P2P) lending, a decentralized debt financing system where loans are made directly between borrowers and lenders. Unlike traditional financial institutions, P2P lending involves higher risk and requires careful assessment of borrower creditworthiness. Using a dataset of historical loans, we conduct data preprocessing and exploratory data analysis to identify key loan characteristics. Machine learning models are applied to predict default risk and evaluate potential returns, offering insights into risk management strategies and investment decision-making.

I. INTRODUCTION

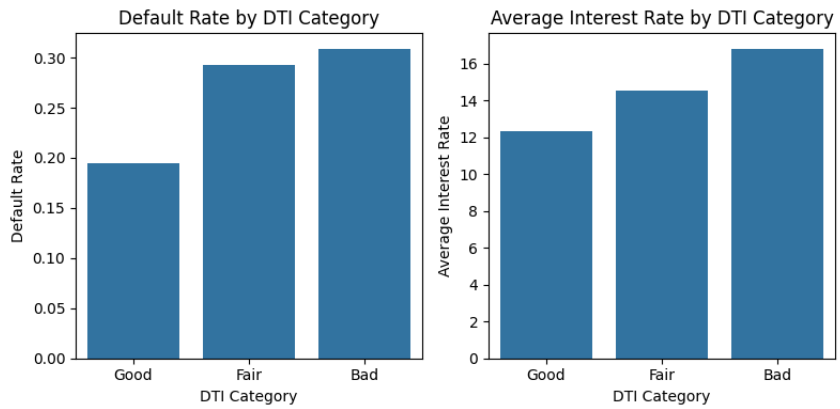
This group assignment shows the ability to analyze and compare many different ways to use code to understand and analyze data to predict default risk on peer-to-peer lending. To summarize what our group did, we applied cleaned data to multiple different sections. The goal of our project is to utilize the multiple ways of regression to best understand the different outcomes of each. Because of our large dataset as well we can understand the importance of using different regression models and then comparing them. The hardest part of this project was understanding the comparison of each model. Regression analysis is so vital to this project because of peer-to-peer lending. The whole point of doing this analysis is to find out the best private lenders and what factors contribute to the best investment. Peer-to-peer is different because it does not involve a bank or cooperation but instead a private company. We can then look into the different factors that affect the different individual loan givers. From this peer-to-peer lending, there are many different loans that consumers can apply for. The point of our summarization isn't to find these loans but to use different regression models to analyze and understand the data. Rainer Lenz of Bielefeld University states that "Web-based financial intermediation on a peer-to-peer (P2P) basis will eventually prevail as an economically superior form of organization compared to the traditional banking business model. P2P lending is the most popular type of crowdfunding, whereby an internet platform collects small amounts of funds from individuals in a crowd to finance collectively a larger loan to individuals or businesses. Unlike a commercial bank, the platform does not take risks through its own contractual positions. Whereas banks accumulate risks by taking positions on their balance sheet, platforms decentralize the risks by spreading them to their users".

II. DATA SECTION

1. EDA

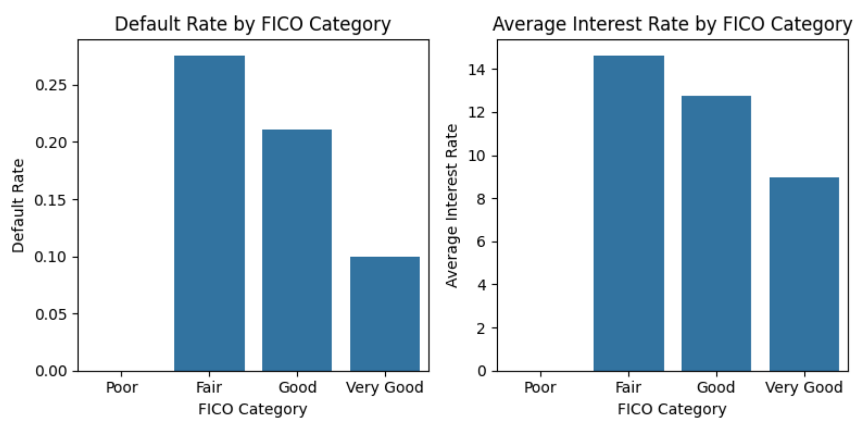
These sections started with our basic data where we loaded the provided data into our Google Colab. From this, we started our exploratory data analysis by creating different variables that helped us find what our best investment would be.

a. DTI_cat



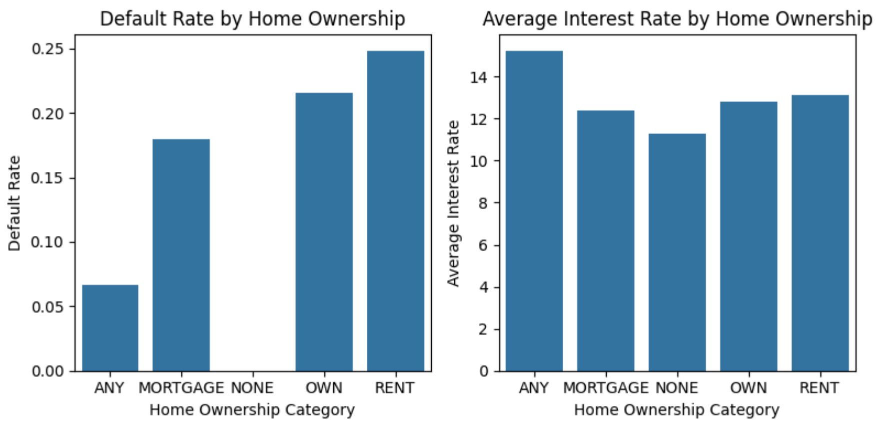
Analysis of P2P lending data shows a clear correlation between DTI categories and loan performance. The interest rate premium of 4.5% for riskier borrowers fails to adequately compensate for their 12% higher default risk, indicating that high-risk loans may not be optimal investments. Therefore, the recommended strategy is to concentrate investments on "Good" DTI borrowers (60-70% of portfolio), maintain limited exposure to "Fair" DTI (20-30%), and minimize "Bad" DTI investments (maximum 10%) to optimize returns while maintaining portfolio stability.

b. FICO_cat



Based on these findings, the optimal investment strategy would be to heavily weight the portfolio towards "Very Good" FICO borrowers (suggested 50-60% allocation) and "Good" FICO borrowers (30-40%), while maintaining minimal exposure to "Fair" FICO borrowers (10-20%), as the higher interest rates in lower FICO categories do not adequately compensate for the increased default risk. In addition, implementing strict screening criteria for "Fair" FICO borrowers, requiring additional collateral, and developing a tiered pricing model that better aligns interest rates with actual default risks across FICO categories would be options.

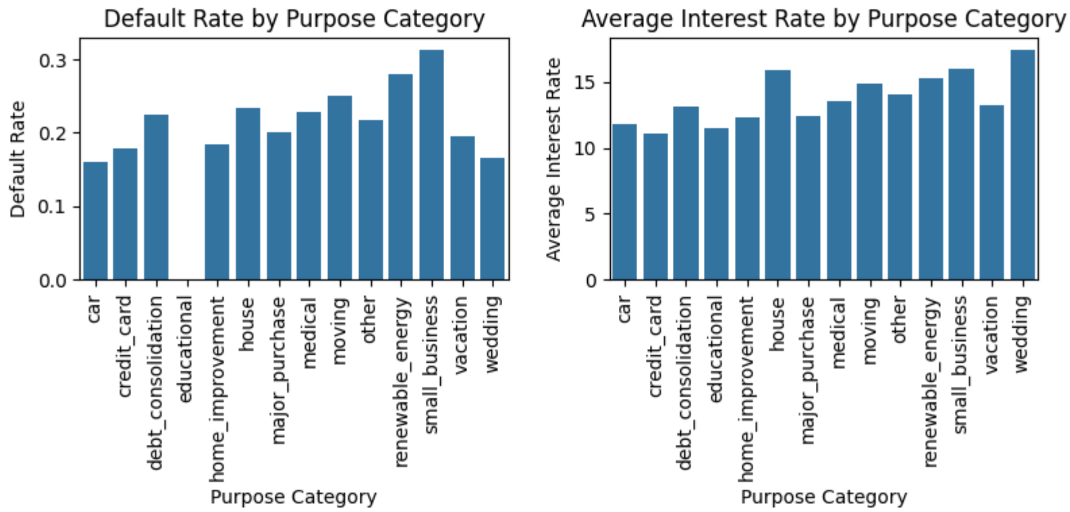
c. Home Ownership



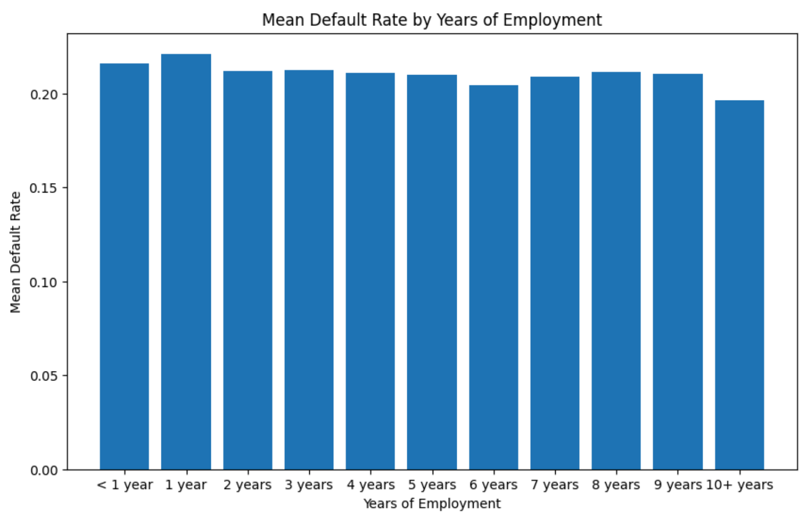
Borrowers with ANY home ownership status show the lowest default risk (7%) despite high interest rates (15%), while renters have the highest default rate (25%) with 13% interest rates, and traditional homeowners surprisingly show high defaults at 21%. This pattern challenges common assumptions about property ownership indicating creditworthiness. For optimal P2P lending strategy, prioritize borrowers with ANY home ownership status, followed by those with mortgages (18% default rate), while being cautious with renters and outright homeowners despite their attractive rates.

d. Loan Purpose

The graph reveals a risk pattern: vacation and small business loans have high default rates (around 30%) and interest rates (15-17%), contrasting with car loans and wedding expenses, which have lower default rates (15-17%) and moderate interest (11-13%). Home improvement and debt consolidation show moderate risks, with default rates of 18-22% and interest at 11-13%, making them potentially balanced options. An optimal strategy would focus on lower-risk loans like car and wedding expenses while avoiding high-risk vacation and small business loans, where high interest fails to offset default risk.

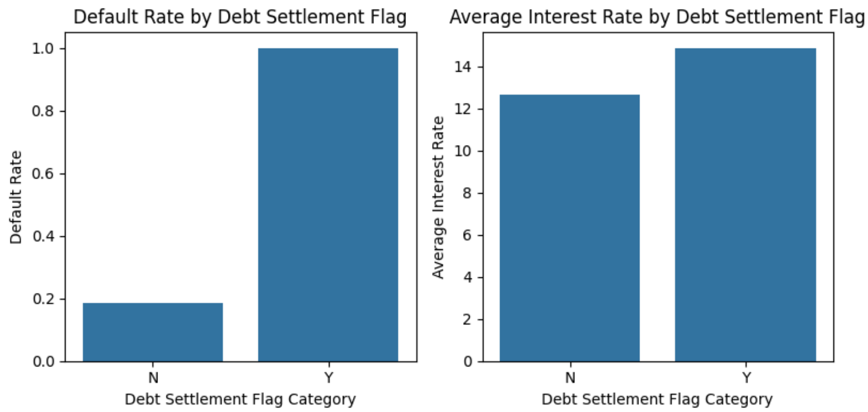


e. Employment Length



In the graph above, borrowers with less than one year and one year of employment show the highest default rates (around 21-22%), while those with 10+ years of experience demonstrate the lowest default rate at approximately 19%. The data suggests a slight downward trend in default risk as employment tenure increases, though the difference is relatively modest with only about a 2-3 percentage point spread between the highest and lowest risk categories. While employment length should be considered in lending decisions, its impact on default risk is less pronounced than other factors like FICO scores or loan purpose, suggesting it should be a secondary consideration in the investment strategy rather than a primary screening criterion.

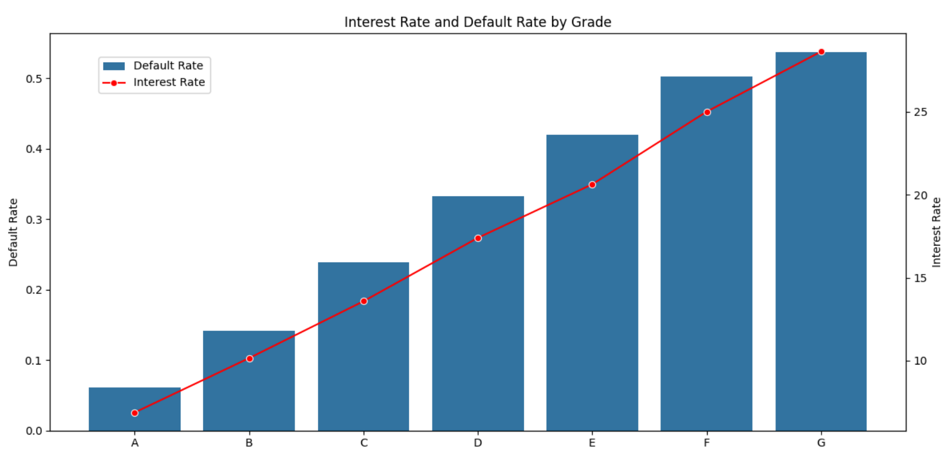
f. Debt Settlement Flag



Borrowers with a debt settlement history show an alarming default rate near 100% despite higher interest rates (15%), compared to those without such history, who default at around 20% with interest rates of 12.5%. This stark difference indicates debt settlement history as a critical risk factor. To maintain a sustainable portfolio, investors should avoid borrowers with debt settlement flags and focus on those without this history.

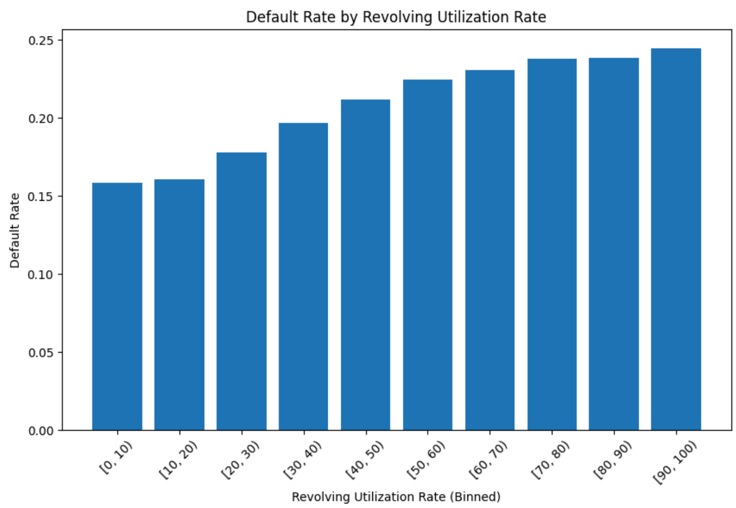
g. By Grade

Grade A borrowers show remarkable reliability with only 5% defaults despite paying the lowest interest rates of 7%, indicating strong financial stability, while Grade G borrowers face the highest interest rates of 28% yet demonstrate default rates over 50%, suggesting these are likely borrowers with limited financial alternatives.



The middle grades (C-E) represent the sweet spot in the market, where moderate default rates of 20-40% are balanced by interest rates of 15-25%, potentially offering the most efficient risk-adjusted returns for investors. This pattern reveals that beyond Grade E, higher interest rates may actually contribute to increased defaults, creating a cycle that investors should consider when building their P2P lending portfolios.

h. Revolving Utilization

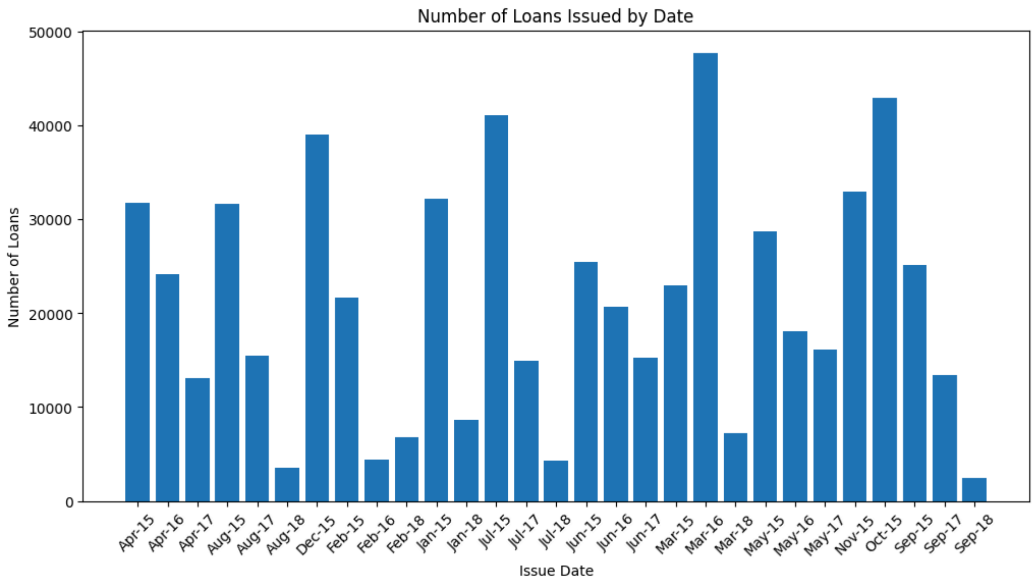


The relationship between revolving credit utilization and default rates in P2P lending shows a clear positive correlation, with default rates increasing from 15% to 25% as utilization rises from 0% to 100%. Low credit utilization (under 40%) corresponds with the lowest default rates around 15-17%, suggesting these borrowers maintain healthier financial habits. The steepest increase in default rates occurs between 20-50% utilization, indicating this range as a critical threshold for risk assessment. It suggests prioritizing borrowers with lower credit utilization rates, particularly those under 40%, to minimize default risk for investors.

i. Issue Date

The P2P lending volume shows a cyclical pattern with peaks typically occurring in March and May-June (reaching up to 47,000 loans), likely corresponding to tax season and mid-year financial planning periods. Loan volumes consistently drop to their lowest points in February and September-October, suggesting seasonal lending behavior. The data fluctuates between 3,000 to

47,000 loans per period, with the highest peak in March 2016, while showing a gradual declining trend from 2015 to 2018, indicating evolving market conditions or lending policy changes.



2. Variables

The dependent variable for this analysis, "*loan_status_dummy*", classifies each observation as either "*Fully Paid,*" "*Charged Off,*" or "*Default,*" representing the loan repayment outcomes. Observations with loan statuses outside these categories were excluded, yielding a final dataset of 611,803 observations.

A carefully selected list of predictors was used to model *"loan_status_dummy"*, guided by recommendations from Rainer Lenz of Bielefeld University. These predictors include a mix of financial, credit, and demographic factors that were converted into dummy variables, resulting in 66 final predictors.

Missing values, only in categorical variables, were imputed with the mode of each respective variable. For classification, loan status predictions were interpreted as probabilities above 50% indicating *"Fully Paid"* status; otherwise, they were classified as *"Charged Off"* or *"Default"*.

III. METHODS SECTION

1. Ordinary Least Squares (OLS) Regression

The first regression model performed was the Ordinary Least Squares (OLS) regression. The method aims to minimize the sum of squared residuals between observed and predicted values. Table 1 shows the result of OLS regression on this dataset.

Table 1: OLS Regression Results

	coef	std err	t	P>|t|	[0.025
intercept	-0.1196	0.075	-1.595	0.111	-0.267
funded_amnt	-5.748e-06	4.49e-07	-12.799	0.000	-6.63e-06
int_rate	0.0123	0.000	68.569	0.000	0.012
installment	0.0003	1.41e-05	18.606	0.000	0.000
collections_12_mths_ex_med	0.0233	0.003	7.501	0.000	0.017
annual_inc	-9.83e-09	7.9e-09	-1.243	0.214	-2.53e-08
pub_rec	0.0032	0.001	3.899	0.000	0.002
revol_bal	3.412e-08	6.65e-08	0.513	0.608	-9.61e-08
revol_util	0.0002	3.53e-05	5.490	0.000	0.000
inq_last_6mths	0.0111	0.001	16.339	0.000	0.010
total_rev_hi_lim	-2.082e-07	4.29e-08	-4.856	0.000	-2.92e-07
acc_open_past_24mths	0.0066	0.000	31.530	0.000	0.006
avg_cur_bal	-4.965e-07	4.16e-08	-11.922	0.000	-5.78e-07
bc_open_to_buy	6.154e-07	9.54e-08	6.451	0.000	4.28e-07
mo_sin_old_il_acct	-1.134e-05	1.02e-05	-1.113	0.266	-3.13e-05
mo_sin_old_rev_tl_op	-2.602e-05	6.05e-06	-4.297	0.000	-3.79e-05
mo_sin_rcnt_tl	-0.0002	6.64e-05	-2.481	0.013	-0.000
mort_acc	-0.0068	0.000	-20.771	0.000	-0.007
mths_since_recent_bc	-0.0001	1.82e-05	-6.313	0.000	-0.000
mths_since_recent_inq	-0.0003	9.53e-05	-3.155	0.002	-0.000
num_actv_rev_tl	0.0037	0.001	4.875	0.000	0.002
num_bc_tl	-0.0009	0.000	-6.103	0.000	-0.001
num_rev_tl_bal_gt_0	0.0032	0.001	4.112	0.000	0.002
num_sats	-0.0018	0.000	-11.400	0.000	-0.002
num_tl_120dpd_2m	0.0437	0.016	2.717	0.007	0.012
num_tl_30dpd	0.0190	0.008	2.499	0.012	0.004
num_tl_90g_dpd_24m	0.0037	0.001	3.783	0.000	0.002
pct_tl_nvr_dlq	-0.0003	6.24e-05	-4.287	0.000	-0.000
percent_bc_gt_75	0.0002	2.02e-05	11.438	0.000	0.000
total_bal_ex_mort	1.395e-07	3.78e-08	3.688	0.000	6.54e-08
total_bc_limit	-7.408e-07	6.33e-08	-11.711	0.000	-8.65e-07
total_il_high_credit_limit	-1.916e-07	3.76e-08	-5.090	0.000	-2.65e-07
Income_cat_Medium Income	-0.0137	0.004	-3.484	0.000	-0.021
Income_cat_High Income	-0.0357	0.004	-8.456	0.000	-0.044
DTI_cat_Fair	0.0459	0.001	30.949	0.000	0.043
DTI_cat_Bad	0.0278	0.007	4.146	0.000	0.015
FICO_cat_Fair	-0.0274	0.025	-1.095	0.274	-0.076
FICO_cat_Good	-0.0449	0.025	-1.795	0.073	-0.094
FICO_cat_Very Good	-0.0474	0.025	-1.891	0.059	-0.096
purpose_credit_card	0.0054	0.005	1.059	0.289	-0.005
purpose_debt_consolidation	0.0095	0.005	1.896	0.058	-0.000
purpose_educational	-0.0983	0.386	-0.255	0.799	-0.854
purpose_home_improvement	0.0218	0.005	4.096	0.000	0.011

2. Logistic Regression

Next, the logistic regression model was employed to predict loan status using the specified predictors. Logistic regression, a classification technique commonly applied in binary/dummy variables, estimates the probability of which class a given instance belongs to. Here, the model classifies observations as either approved or denied, with the `loan_status_dummy` variable serving as the binary response. Initially, the model was fit using the entire dataset to evaluate its performance on the full data, and predictions were generated to determine each observation's predicted class. The results for the Logistic Regression are shown in Table 2.

Table 2: Logistic Regression Results

funded_amnt: -9.406599266820648e-05	num_tl_30dpd: 8.516711086009694e-07
int_rate: 0.00010735448031347162	num_tl_90g_dpd_24m: 2.715645475117669e-06
installment: 0.0041324631350228	pct_tl_nvr_dlq: 1.5708589166970442e-05
collections_12_mths_ex_med: -0.002295927365785294	percent_bc_gt_75: -0.009021482246259293
annual_inc: 1.0633209649309819e-05	total_bal_ex_mort: 0.002900956738534818
pub_rec: -3.6420500282157234e-06	total_bc_limit: 5.786532085752422e-06
revol_bal: 5.7980806224286e-05	total_il_high_credit_limit: -6.177391792625234e-06
revol_util: 3.7749962867811547e-06	Income_cat_Medium Income: -4.16996197026451e-06
inq_last_6mths: -0.0016656637450063341	Income_cat_High Income: -7.819154891621199e-05
total_rev_hi_lim: 0.00035800249919090677	DTI_cat_Fair: -1.4934836219137869e-05
acc_open_past_24mths: -5.8576434457497304e-06	DTI_cat_Bad: 0.00012499086417279263
avg_cur_bal: 0.0014549381415792663	FICO_cat_Fair: 4.332015834771234e-06
bc_open_to_buy: -1.8932560103723965e-05	FICO_cat_Good: 8.581159404606438e-05
mo_sin_old_il_acct: -8.844996165094963e-06	FICO_cat_Very Good: -0.00010780476552902158
mo_sin_old_rev_tl_op: -0.0010524346220576154	purpose_credit_card: -7.207282118450498e-05
mo_sin_rcnt_tl: -0.000781332835954764	purpose_debt_consolidation: -0.00011554796953664225
mort_acc: -0.0032248995684202564	purpose_educational: -8.952615165107666e-06
mths_since_recent_bc: -0.00018042600757735432	purpose_home_improvement: -5.966750016673183e-09
mths_since_recent_inq: -0.006113976587936997	purpose_house: 2.1798213079575472e-06
num_actv_rev_tl: -0.002018289948759289	purpose_major_purchase: 1.0848861991059613e-06
num_bc_tl: 0.0008631972043158484	purpose_medical: -2.2003576492754167e-06
num_rev_tl_bal_gt_0: 0.00024783937661337587	purpose_moving: 6.076766289883655e-06
num_sats: 0.0008089468091179376	purpose_other: 3.736041694726648e-06
num_tl_120dpd_2m: 0.00036377952817850826	purpose renewable energy: 1.0522065016195676e-05

Additionally, a validation approach was utilized to evaluate the model on unseen data, ensuring robustness. After training the model on the training subset, its performance was evaluated on the test set using a confusion matrix and accuracy rate, which provide insights into the model's generalization capacity and predictive reliability.

3. Ridge Regression

In order to identify the optimal predictive coefficients with minimal Mean Squared Error (MSE), a 10-fold cross-validation was conducted using ridge regression across a range of regularization parameters (λ). The chosen λ values included 0.001, 0.01, 0.1, 1, 10, 100, 1,000, and 10,000, representing a comprehensive selection of possible regularization strengths. This iterative cross-validation technique was applied to enhance model generalizability by splitting the dataset into 10 subsets, training the model on nine subsets, and validating it on the remaining subset for each fold. The mean MSE across all folds was computed for each λ to determine the regularization parameter that minimized prediction error, balancing bias and variance effectively.

Upon completing the analysis, it was observed that a λ value of 10,000 yielded the lowest average MSE among all tested parameters. This suggests that stronger regularization is beneficial for this dataset, implying a higher level of multicollinearity or overfitting susceptibility in the model when λ is low. Consequently, λ = 10,000 was selected as the optimal parameter, as it achieved a balance between regularization and predictive accuracy, thus enhancing the model’s robustness in out-of-sample predictions while reducing susceptibility to overfitting. The results for Ridge Regression are shown in Table 3.

Table 3: Ridge Regression Results with Lambda equals 10,000

intercept: -0.007067382610996146	*DTI_cat_Fair: 0.0022581012611342013*
funded_amnt: 0.0643021737696877	*DTI_cat_Bad: 0.004842787489407763*
int_rate: 0.027845875643210664	*FICO_cat_Fair: -0.0024569903824991773*
installment: 0.0036555787830903973	*FICO_cat_Good: -0.0025048218334175454*
collections_12_mths_ex_med: -0.0009891346434071282	*FICO_cat_Very Good: -0.00033241322250526464*
annual_inc: 0.0020901009938716205	*purpose_credit_card: 0.0018121693707937032*
pub_rec: -2.710021606766497e-05	*purpose_debt_consolidation: -0.00013838650252053611*
revol_bal: 0.004568944235175552	*purpose_educational: 0.003916651078870033*
revol_util: 0.00970213545856756	*purpose_home_improvement: -0.001205814218772634*
inq_last_6mths: -0.006159088302704666	*purpose_house: 0.002183069652499703*
total_rev_hi_lim: 0.021276567342129595	*purpose_major_purchase: 0.003356298090505763*
acc_open_past_24mths: -0.008008652728041523	*purpose_medical: 0.0021018066024729195*
avg_cur_bal: 0.006586870060220692	*purpose_moving: 0.0021497670983307176*
bc_open_to_buy: -0.0007072147831892993	*purpose_other: 0.0013641269734796246*

mo_sin_old_il_acct: -0.002447073956687968
mo_sin_old_rev_tl_op: -0.0015443720492939361
mo_sin_rcnt_tl: -0.01315215417853771
mort_acc: -0.0036190754596852697
mths_since_recent_bc: -0.0018133573984465364
mths_since_recent_inq: 0.011137175755126589
num_actv_rev_tl: -0.004051759715925842
num_bc_tl: 0.011403493232275293
num_rev_tl_bal_gt_0: -0.009302109213555705
num_sats: 0.0013224274972039236
num_tl_120dpd_2m: 0.0012123481964689065
num_tl_30dpd: 0.0019030706590076668
num_tl_90g_dpd_24m: -0.002303360153900727
pct_tl_nvr_dlq: 0.008106256336120126
percent_bc_gt_75: 0.004692902035306173
total_bal_ex_mort: -0.014466695327945986
total_bc_limit: -0.006714558130568537
total_il_high_credit_limit: -0.002219447039341021
Income_cat_Medium Income: -0.012805990243744163
Income_cat_High Income: 0.01647177578853666
purpose_renewable_energy: 0.006616743210453061
purpose_small_business: 0.0009121621576262555
purpose_vacation: -0.00020073948701001052
purpose_wedding: 0.055015763906445726
term_ 60 months: -0.0019762863000273028
application_type_Joint App: -0.01960979709400403
emp_length_10+ years: -0.010434259589844178
emp_length_2 years: -0.00948524526985791
emp_length_3 years: -0.00899274070014643
emp_length_4 years: -0.008909240775813253
emp_length_5 years: -0.008456492549162956
emp_length_6 years: -0.007681234205174792
emp_length_7 years: -0.007248407255350627
emp_length_8 years: -0.006808798947769536
emp_length_9 years: -0.008647724667516718
emp_length_< 1 year: -0.010027501496191945
home_ownership_MORTGAGE: -0.0003684999749362719
home_ownership_NONE: -4.936359509370783e-05
home_ownership_OWN: 0.010844912934953354

4. Lasso Regression

A similar 10-fold cross-validation process and the same set of λ values were subsequently applied to a Lasso regression model to identify the optimal λ and to analyze feature selection effects. Lasso, known for its capability to enforce sparsity, has the advantage of driving certain coefficients to zero, effectively performing feature selection by excluding less relevant predictors from the model.

The cross-validation results indicated that λ = 0.001 minimized the average Mean Squared Error (MSE), suggesting this level of regularization best-balanced model fit and feature selection. At this λ, the Lasso regression assigned a coefficient of zero to a subset of predictors, implying they have little impact on predicting the target variable. These excluded features were: 'intercept,' 'pub_rec,' 'avg_cur_bal,' 'bc_open_to_buy,' 'percent_bc_gt_75,' 'purpose_credit_card,' 'purpose_moving,' 'FICO_cat_Fair,' 'purpose_debt_consolidation,' 'total_il_high_credit_limit,' 'purpose_small_business,' 'home_ownership_MORTGAGE,' and 'home_ownership_NONE.' 'purpose_vacation,' This outcome highlights the effectiveness of Lasso in enhancing model interpretability by focusing on a more concise set of variables while still optimizing prediction accuracy. Table 4 shows the results of Lasso Regression.

Table 4: Lasso Regression Result with Lambda equals 0.001IV. RESULTS

funded_amnt: 0.06989167634788425
int_rate: 0.017773429674985106
installment: 0.0027638693420359243
collections_12_mths_ex_med: -0.0003694940576866436
annual_inc: 0.0011192972943537018
pub_rec: -0.0
revol_bal: 0.002322278268249338
revol_util: 0.009035978331987623
inq_last_6mths: -0.0035368792330059944
total_rev_hi_lim: 0.019606443669168316
acc_open_past_24mths: -0.007696592663452977
avg_cur_bal: 0.0
bc_open_to_buy: -0.0
mo_sin_old_il_acct: -0.0007411733240892574
mo_sin_old_rev_tl_op: -0.0005953201489212837
mo_sin_rcnt_tl: -0.013090720176662069
mort_acc: -0.0027845861501326866
mths_since_recent_bc: -0.0008158365849056818
mths_since_recent_inq: 0.007921879304544625
num_actv_rev_tl: -0.0021563486491659087
num_bc_tl: 0.011584703348649129
num_rev_tl_bal_gt_0: -0.006731427807511788
num_sats: 0.00033023617046810537
num_tl_120dpd_2m: 4.2839205206660354e-05
num_tl_30dpd: 0.001057607912110321
num_tl_90g_dpd_24m: -0.0010576559409698843
pct_tl_nvr_dlq: 0.00673050539742072
percent_bc_gt_75: -0.0
total_bal_ex_mort: -0.010416406900316897
total_bc_limit: -0.00250589331627766
total_il_high_credit_limit: 0.0
Income_cat_Medium Income: -0.010192678473144975
Income_cat_High Income: 0.014944059515601089
DTI_cat_Fair: 0.0005709360748949231
DTI_cat_Bad: 0.006882232764587235
FICO_cat_Fair: -0.0
FICO_cat_Good: -0.00034240433014124305
FICO_cat_Very Good: -0.0009052554146035393
purpose_credit_card: -0.0
purpose_debt_consolidation: -0.0
purpose_educational: 0.0015294093003854292
purpose_home_improvement: -0.0007475940482427522
purpose_house: 0.00036539424240640984
purpose_major_purchase: 0.0016020195660343677
purpose_medical: 0.0004451184164154747
purpose_moving: 0.0
purpose_other: 0.00017767713073821518
purpose_renewable_energy: 0.004952621577589984
purpose_small_business: 0.0
purpose_vacation: -0.0
purpose_wedding: 0.05058402638864876
term_ 60 months: -8.602466480884604e-05
application_type_Joint App: -0.011343069176981293
emp_length_10+ years: -0.00474858872168142
emp_length_2 years: -0.004026778118517382
emp_length_3 years: -0.004102119848702815
emp_length_4 years: -0.004013888456770025
emp_length_5 years: -0.00423971871911838
emp_length_6 years: -0.0034613716975969763
emp_length_7 years: -0.0027974631150339217
emp_length_8 years: -0.0027008810774325166
emp_length_9 years: -0.0032950592269653514
emp_length_< 1 year: -0.009973678883676984
home_ownership_MORTGAGE: -0.0
home_ownership_NONE: 0.0
home_ownership_OWN: 0.010285503031222496

To assess model performance, a Confusion Matrix was utilized to examine the classification outcomes across all models. This matrix provides a breakdown of true positive, true negative, false positive, and false negative counts, allowing for a comprehensive evaluation of prediction accuracy. Using the Confusion Matrix results, the Accuracy Rate was calculated as the proportion of correct predictions (both true positives and true negatives) out of the total predictions made.

IV. RESULTS

The Accuracy Rate serves as a key metric to determine how well each model correctly classifies the loan status as either *"Fully Paid"* or *"Charged Off/Default."* Using this measure, the evaluation provides insights into each model's ability to generalize on unseen data, highlighting its effectiveness in correctly predicting loan repayment status. This approach ensures that models are evaluated consistently and reliably, aligning with the study's objective to achieve robust and accurate classification performance. The results for each model are as follows:

1. **Ordinary Least Squares (OLS):**
    - Confusion Matrix:

474881	7713
119068	10140

    - Prediction Accuracy Rate: **0.7928**
2. **Logistic Regression:**
    - Confusion Matrix:

481466	1128
128285	923

    - Prediction Accuracy Rate: **0.7885**
    - Confusion Matrix on Test Set:

96520	94
25657	90

    - Prediction Accuracy Rate on Test Set: **0.7895**
3. **Ridge Regression:**
    - Confusion Matrix:

475568	7026
119774	9434

- Prediction Accuracy Rate: **0.7927**

4. **Lasso Regression:**
   - Confusion Matrix:

475672	6922
120116	9092

- Prediction Accuracy Rate: **0.7924**

The models were evaluated using the Confusion Matrix and corresponding Accuracy Rate, which measures the proportion of correct predictions. Among the models, Ordinary Least Squares (OLS) achieved the highest Accuracy Rate of 0.7928, followed closely by Ridge regression at 0.7927, Lasso regression at 0.7924, and Logistic regression at 0.7885. Despite these similar rates, OLS showing the highest Accuracy Rate indicates it performs well on the training dataset.

However, OLS's strong performance on the training data does not necessarily imply superior predictive ability on new, unseen data. The Ridge and Lasso models, in contrast, incorporate regularization techniques to address potential overfitting, which intentionally introduces a trade-off between model fit on training data and generalizability. As a result, while Ridge and Lasso exhibit slightly lower Accuracy Rates on the training dataset, this reflects their efforts to improve performance on future predictions by minimizing variance. In contrast, OLS, which lacks regularization, may better fit the training data at the risk of overfitting, potentially limiting its effectiveness on new data. Thus, Ridge and Lasso offer benefits in predictive robustness, although with a marginal trade-off in training accuracy.

Business Practices

This project uses data from peer-to-peer (P2P) lending to help businesses, investors, and lending platforms make smarter decisions about loans. By analyzing which factors influence whether a loan is fully repaid or goes unpaid, the model can help predict which borrowers are likely to repay their loans and which might struggle.

For P2P lending companies, this means better loan approvals, as the model can help determine the risk of each borrower. It also allows lenders to set interest rates that match the risk, making loans fairer and potentially more profitable. Investors benefit by being able to choose safer investments, as they can see which types of borrowers tend to repay successfully.

Thus, understanding default risks helps lenders manage loans more effectively—they can plan for timely follow-ups or support borrowers to avoid defaults. It also aids in targeted marketing, allowing companies to offer personalized loan options that suit different types of borrowers.

Therefore, this project provides a clear, data-backed way to make lending and investing in P2P platforms safer, more efficient, and better aligned with the needs and behaviors of borrowers.

V. CONCLUSION

In this study, various multiple regression techniques were compared based on which provided high accuracy for predictive models. The highest accuracy was obtained by OLS at 0.7928 indicating its ability to work well with the attributes of the dataset. Ridge Regression and Lasso Regression have comparable performance; therefore, they can be used in place of the previous models if regularization is an issue. Logistic Regression showed the least accuracy of the class, which signifies that the binary classification capability of this type of Learning Model is quite constrained in this data set. Concisely, it is possible to denote that the optimal model, which was chosen for the given data, was the OLS model that guarantees balanced and further predictions.

REFERENCES

Lenz, Rainer. (PDF) Peer-to-Peer Lending: Opportunities and Risks, www.researchgate.net/publication/313442224_Peer-to-Peer_Lending_Opportunities_and_Risks. Accessed 11 Nov. 2024.

Regression Analysis, www.uoguelph.ca/lang/system/files/Regression.pdf. Accessed 11 Nov. 2024.

Simple Linear Regression, www.colorado.edu/amath/sites/default/files/attached-files/ch12_0.pdf. Accessed 11 Nov. 2024.