The Effect of Borrower Identities on Loan Application in the P2P Market

Russell Liu, Xiaokai Xu, Yezi Liu

2. Summary of Research Questions and Results

1. Question: How do borrower identities(FICO score, annual income, etc) impact the probability of qualifying for loans in the P2P market, either positively or negatively? We will run logistic regression and analyze the coefficients.

Result: We obtained the coefficients for all borrower identities from the logistic regression and listed several interesting findings below.

- When a borrower's annual income percentage increases, the probability of meeting borrowing criteria at LendingClub increases under the 5% significance level.
- When a borrower has a higher FICO score, the probability of meeting borrowing criteria at LendingClub increases under the 5% significance level.
- When the number of days a borrower has a credit line increases, the probability of meeting borrowing criteria at LendingClub increases under the 5% significance level.
- When the times a borrower requests credit reports in the last 6 months increase, the probability of meeting borrowing criteria at LendingClub decreases under the 5% significance level.
- When a borrower has a record of loan default at LendingClub, the probability of meeting borrowing criteria at Lending Club decreases under the 5% significance level.

2. Question: Which borrower identities are the most significant for predicting loan qualification in the P2P market? We will run LASSO regression and analyze the independent variables with non-zero coefficients.

Result: The LASSO regression analysis indicates that these features have a more significant impact on the dependent variable.

- The borrower's number of inquiries by creditors in the last 6 months
- The FICO credit score of the borrower
- The borrower's revolving line utilization rate
- The monthly installments owed by the borrower if the loan is funded
- The number of days the borrower has had a credit line
- The borrower's revolving balance

Also, requesting credit reports in the last 6 months many times and owning a high revolving balance have a significantly negative impact on meeting borrowing criteria for a loan. Other features have a significantly positive impact on meeting borrowing criteria for a loan, which aligns with our logistic regression analysis.

3. Question: How could we use these significant features and apply machine learning to accurately predict whether a borrower qualifies for a loan or not in the P2P market? We will use the Decision Tree and Random Forest to build models and choose the one with the highest accuracy.

Result:

The accuracy scores for both the cross-validation with 5 folds and the test set are really high for the Decision Tree model, with values of 99.1% and 99.3% respectively. It suggests that the model would perform well on future data and is not subject to overfitting. On the other hand, to solve the imbalanced data issue, we built the Random Forest model with accuracy scores of 98.9% for cross-validation and 98.8% for the test set. Similarly, it suggests that the model would perform well on future data and is not subject to overfitting.

4. Question: Are there possible biases present in our dataset and potential limitations to our machine learning models, and why might they have occurred? We will carefully research our dataset and reflect on the results of our models.

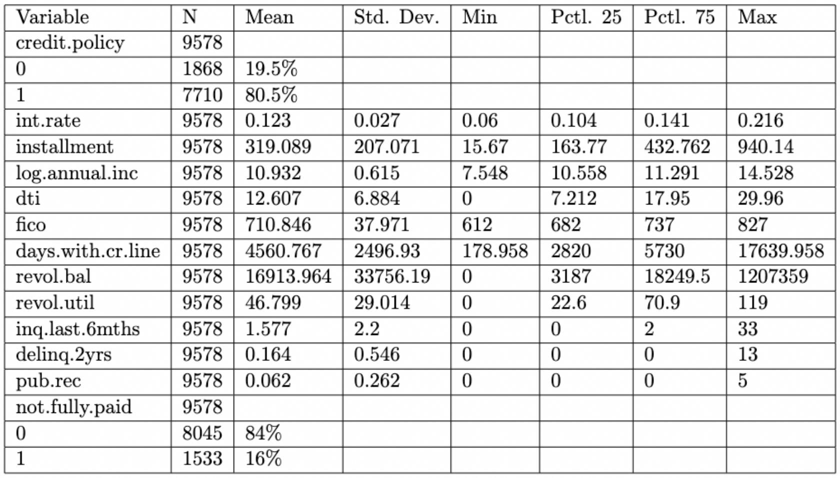
Result: There are several main biases in our dataset, like sampling bias, measurement bias, and historical bias. There are also limitations to our Random Forest model. Despite having the highest accuracy, its predictions tend to favor borrowers more than lenders at LendingClub from the types of errors it makes, which doesn't meet our expectations for fairness. Our assumptions for individual fairness could also cause limitations to our model in reality.

3. Motivation

The project mainly focuses on the loan market and how certain borrowers' features on record can affect the investor's decision on lending. A P2P lender typically has a lower credit requirement than a traditional lender, allowing individuals with poor credit to qualify for the loan. Having a holistic understanding and avoiding imbalanced information between "peers" is important for both financial company lenders and personal borrowers. We want to determine whether loan applicants meet certain borrowing criteria and are likely to repay the loan by examining crucial parameters from LendingClub. There are enormous underprivileged groups all over the world that are desperate due to the lack of access to loans or financing. Simultaneously, many companies and personal investors encounter bankruptcy due to borrowers' default loans. Therefore, we care about this issue since it helps borrowers to adjust their financial behaviors and increases the probability of obtaining loans, and supports lenders to avoid risky borrowers that tend to default.

4. Dataset

LendingClub is a financial services company headquartered in San Francisco, California. It's the first and largest lending platform that registered its offerings as securities with the Securities and Exchange Commission (SEC) and offered loan trading on a secondary market to various types of urban customers. The dataset is on Kaggle.com called "Loan Data" which was publicly available cross-sectional data from LendingClub.com (URL: https://www.kaggle.com/datasets/itssuru/loan-data). The dataset contained 9578 observations and 14 attributes, profiling the key parameters of risk analysis to each borrower during the years from 2007 to 2010. All of the data points were collected within the border of the USA at the individual level. There was no missing value (NA) inside.



Variable	N	Mean	Std. Dev.	Min	Pctl. 25	Pctl. 75	Max
credit.policy	9578						
0	1868	19.5%					
1	7710	80.5%					
int.rate	9578	0.123	0.027	0.06	0.104	0.141	0.216
installment	9578	319.089	207.071	15.67	163.77	432.762	940.14
log.annual.inc	9578	10.932	0.615	7.548	10.558	11.291	14.528
dti	9578	12.607	6.884	0	7.212	17.95	29.96
fico	9578	710.846	37.971	612	682	737	827
days.with.cr.line	9578	4560.767	2496.93	178.958	2820	5730	17639.958
revol.bal	9578	16913.964	33756.19	0	3187	18249.5	1207359
revol.util	9578	46.799	29.014	0	22.6	70.9	119
inq.last.6mths	9578	1.577	2.2	0	0	2	33
delinq.2yrs	9578	0.164	0.546	0	0	0	13
pub.rec	9578	0.062	0.262	0	0	0	5
not.fully.paid	9578						
0	8045	84%					
1	1533	16%					

Table 1. Summary Statistics for Loan Dataset

Summary Statistics from the Table above exhibited the diverse features in variables. Companies have to make decisions for loan approvals based on the loan applicants' profiles. The dummy variable *'credit.policy'* used 1 to represent that the customer meets the credit underwriting criteria of LendingClub, and 0 otherwise. Investors want to invest in people whose profiles have a high probability of paying back. Twelve crucial numerical variables can help LendingClub to judge the applicants' qualifications. The categorical variable *'purpose'* records each borrower with seven different borrowing intentions. The Figure below shows that most people borrow money from LendingClub for debt consolidation purposes, taking up 41.3% of the total sample.

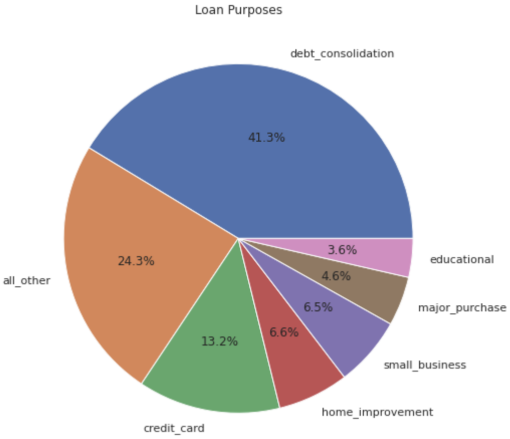
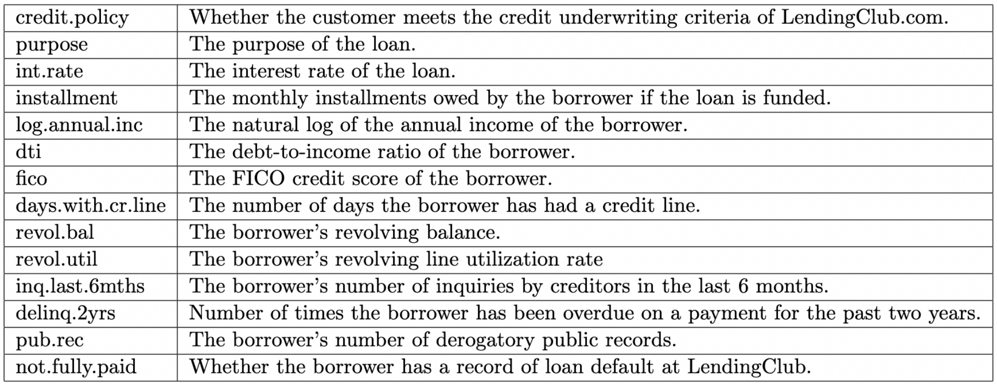


Figure 1. Pie Chart Showing the Variable *'purpose'*

A detailed explanation for all of the variables in this dataset is listed below:



credit.policy	Whether the customer meets the credit underwriting criteria of LendingClub.com.
purpose	The purpose of the loan.
int.rate	The interest rate of the loan.
installment	The monthly installments owed by the borrower if the loan is funded.
log.annual.inc	The natural log of the annual income of the borrower.
dti	The debt-to-income ratio of the borrower.
fico	The FICO credit score of the borrower.
days.with.cr.line	The number of days the borrower has had a credit line.
revol.bal	The borrower's revolving balance.
revol.util	The borrower's revolving line utilization rate
inq.last.6mths	The borrower's number of inquiries by creditors in the last 6 months.
delinq.2yrs	Number of times the borrower has been overdue on a payment for the past two years.
pub.rec	The borrower's number of derogatory public records.
not.fully.paid	Whether the borrower has a record of loan default at LendingClub.

Table 2. A Comprehensive Explanation of Fourteen Attributes

5. Method

1. Explore the dataset with data visualization methods to get a general and comprehensive picture of the borrowers' situations in the P2P market. How many people meet the borrowing criteria? What is the distribution of the FICO score? These questions are important for our regression results.
2. Perform logistic regression to see how independent variables (borrowers' identities) have an effect on the dependent variable (meeting the borrowing criteria or not). (research question 1)
3. Employ the statistical method of hypothesis testing to see whether the coefficients are statistically significant.
4. Employ LASSO regression on the dataset to find the independent variables that have the strongest impact on the dependent variable. (research question 2)
5. Split the dataset into the training set (80%) and test set (20%).
6. To build our Decision Tree model that predicts whether a borrower meets the borrowing criteria or not, use the "GridSearchCV" method with 5-fold cross-validation on the training set to select the optimal value for the hyperparameter of the maximum depth.
7. Use this best hyperparameter value to train a Decision Tree on the entire training data set and make predictions on the test set. (research question 3)
8. Visualize this Decision Tree model with the best depth.
9. To build our Random Forest model that predicts whether a borrower meets the borrowing criteria or not, similarly use the "GridSearchCV" method with 5-fold cross-validation on the training set to select the optimal values for 2 hyperparameters: 'maximum depths of each tree' and 'the number of trees in a forest'.
10. Use the best hyperparameter values to train a Random Forest on the entire training data set and make predictions on the test set. (research question 3)
11. Analyze and compare the accuracy scores on the test set of these two models.
12. Graphically display the prediction results of the Random Forest model on the test set in matrixes that consist of 4 areas: false positives, false negatives, true positives, and true negatives (Confusion Matrix). Positives mean the borrower meets the criteria for getting a loan and negatives mean the borrower doesn't.
13. Discuss whether our model with the highest accuracy has certain limitations and biases for different groups. (research question 4)
14. Discuss the potential bias in the dataset and possible reasons. (research question 4)

6. Main Results

The United States is a modern commercial society built upon various types of lending. People who seek loans would first instinctively go to traditional institutions, such as banks or credit unions [1]. But today, borrowers are offered more alternative options for funding. For example, a peer-to-peer (P2P) loan means that individuals lend money to other individuals using intermediary websites, such as LendingClub, Peerform, and Prosper. These intermediaries allow people who are 18 years old or older to enter their personal information into their databases [2]. The noteworthy benefit of borrowing money from a P2P platform is the streamlined application processes. Different from banks, P2P platforms require far less paperwork and inform their applications of the loan-approval decision almost immediately [3]. Furthermore, P2P lending offers loan opportunities to borrowers with imperfect credit because P2P lenders typically have a lower credit requirement than traditional lenders.

Proper use of loans improves the efficient allocation of social resources while irresponsible use of loans could harm the economies and individual trust between people, leading to economic crises. Although P2P lending lowers the borrowing standards, there are still a large number of borrowers who fail to meet the borrowing criteria. Financial services companies, as the intermediaries between borrowers and investors, need to guarantee the benefits of both sides. Based on our team's research, before sending loans to borrowers, financial institutions require them to fulfill a comprehensive application form including detailed information like annual incomes, past default records, etc [4]. Such information is critical for institutions to make decisions on loan lending. Therefore, we believe that these intermediaries, specifically LendingClub, have standardized parameters to help differentiate whether loan applicants meet the borrowing criteria or not. Derived from the investigation of these crucial parameters, the empirical results can help future borrowers to adjust their financial behaviors and successfully meet borrowing criteria in the P2P market. It also supports lenders to avoid risky borrowers that tend to default.

6.1 Logistic Regression

Theoretically, when the dependent variable is dichotomous or binary in nature, linear regression is not the best way to perform analysis. Figure 2 makes a comparison between two statistical techniques. As x moves along the horizontal axis, y is likely to exceed the range between 0 and 1 in the case of linear regression [5]. It is counterintuitive since the dependent variable *'credit.policy'* is a dummy variable, and 1 means meeting borrowing criteria and 0 otherwise.

Logistic regression is used to predict the relationship between independent variables and a dependent variable where the dependent variable is binary (a dummy variable that has two values such as "male" and "female", "true" and "false", "success" and "failure") rather than continuous. As a result, logistic regression is a better model to fit the data [6].

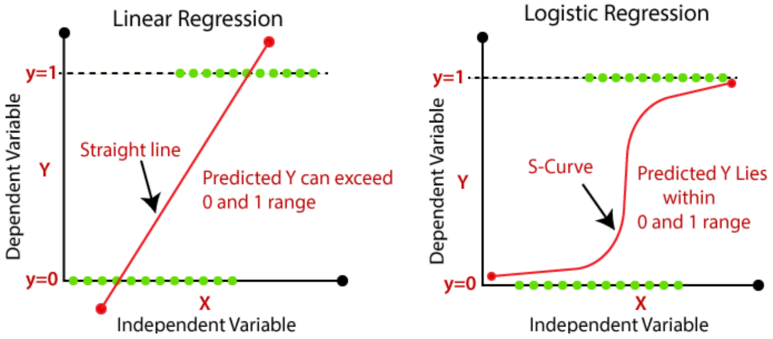


Figure 2. Linear Regression vs Logistic Regression [5]

One assumption made by our team is to employ a Hypothesis Test with a significant level of 5 % for all of the variables. Multiple logistic regression uses the following null and alternative hypotheses [7].

H0 (Null Hypothesis): $\beta1 = \beta2 = \beta3 = \beta4 = 0$

H1 (Alternative Hypothesis): coefficient $\neq 0$

The null hypothesis states that all coefficients in the model are equal to zero, or none of the predictor variables have a statistically significant relationship with the response variable. The alternative hypothesis states that at least one of the coefficients is not equal to zero.

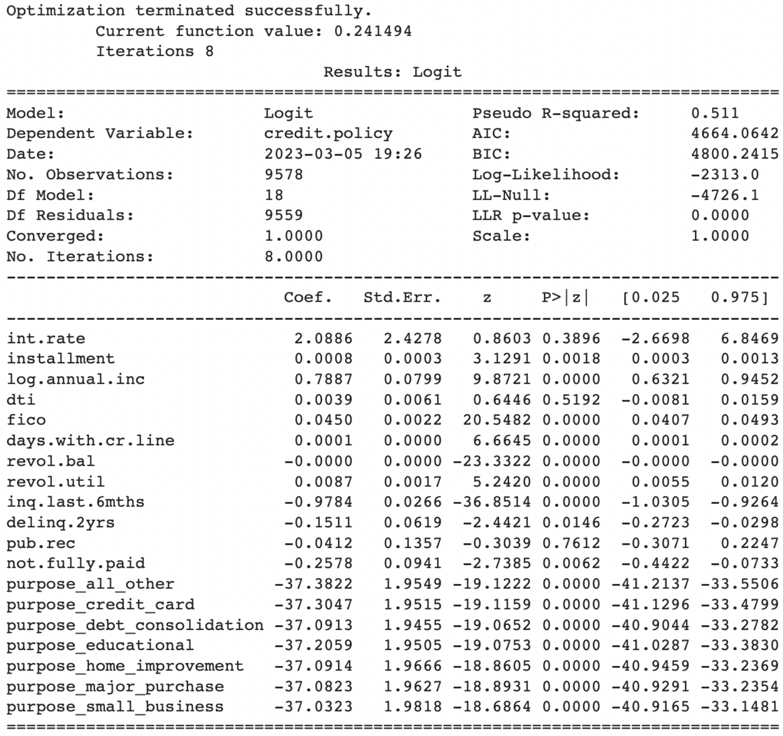
For example:

- Annual income is a significant predictor of borrowing criteria under a 5% significance level.
- Fico score is a significant predictor of borrowing criteria under a 5% significance level.
- The borrower's number of inquiries by creditors in the last 6 months is a significant predictor of borrowing criteria under a 5% significance level.

The coefficients in Table 3 represent the change in the log odds of the outcome variable associated with a one-unit change in the corresponding predictor variable while holding all other variables constant. To interpret a coefficient, we can exponentiate it to obtain the odds ratio, which represents the multiplicative effect of a one-unit increase in the predictor variable on the odds of the outcome variable. For example, if the coefficient for a predictor variable is 0.5, then exponentiating it gives an odds ratio of e^0.5 = 1.65. This means that for a one-unit increase in the predictor variable, the odds of the outcome variable increase by a factor of 1.65, assuming all other variables are held constant.

Again, the significance of the coefficients can be evaluated by looking at their associated p-values. A p-value less than 0.05 indicates that the coefficient is statistically significant, meaning that there is evidence of a non-zero effect of the predictor variable on the outcome variable. In addition to the coefficients and their associated p-values, the result of Table 3 also provides standard errors, confidence intervals, and various goodness-of-fit statistics that can be used to assess the overall fit of the model.

According to Table 3 below, we reject the null hypothesis partially, except for variables like *'int.rate', 'dti', and 'pub.rec'.* Other coefficients are statistically significant at a high level. Thus, most of the independent variables have an effect on the odds of meeting the borrowing criteria at LendingClub. Specifically, each borrower's annual income, FICO scores, the number of inquiries in the last 6 months, etc all contribute to the likelihood of borrowing a loan successfully. In the next four sub-sections, we further elaborate on 3 parameters and their corresponding coefficients to hopefully provide some insights for future borrowers in the P2P market.



```
Optimization terminated successfully.
         Current function value: 0.241494
         Iterations 8
                                Results: Logit
==============================================================================
Model:                    Logit                Pseudo R-squared:     0.511
Dependent Variable:       credit.policy        AIC:                  4664.0642
Date:                     2023-03-05 19:26     BIC:                  4800.2415
No. Observations:         9578                 Log-Likelihood:       -2313.0
Df Model:                 18                   LL-Null:              -4726.1
Df Residuals:             9559                 LLR p-value:          0.0000
Converged:                1.0000               Scale:                1.0000
No. Iterations:           8.0000
------------------------------------------------------------------------------
                            Coef.  Std.Err.    z     P>|z|   [0.025   0.975]
------------------------------------------------------------------------------
int.rate                     2.0886   2.4278   0.8603 0.3896  -2.6698   6.8469
installment                  0.0008   0.0003   3.1291 0.0018   0.0003   0.0013
log.annual.inc               0.7887   0.0799   9.8721 0.0000   0.6321   0.9452
dti                          0.0039   0.0061   0.6446 0.5192  -0.0081   0.0159
fico                         0.0450   0.0022  20.5482 0.0000   0.0407   0.0493
days.with.cr.line            0.0001   0.0000   6.6645 0.0000   0.0001   0.0002
revol.bal                   -0.0000   0.0000 -23.3322 0.0000  -0.0000  -0.0000
revol.util                   0.0087   0.0017   5.2420 0.0000   0.0055   0.0120
inq.last.6mths              -0.9784   0.0266 -36.8514 0.0000  -1.0305  -0.9264
delinq.2yrs                 -0.1511   0.0619  -2.4421 0.0146  -0.2723  -0.0298
pub.rec                     -0.0412   0.1357  -0.3039 0.7612  -0.3071   0.2247
not.fully.paid              -0.2578   0.0941  -2.7385 0.0062  -0.4422  -0.0733
purpose_all_other          -37.3822   1.9549 -19.1222 0.0000 -41.2137 -33.5506
purpose_credit_card        -37.3047   1.9515 -19.1159 0.0000 -41.1296 -33.4799
purpose_debt_consolidation -37.0913   1.9455 -19.0652 0.0000 -40.9044 -33.2782
purpose_educational        -37.2059   1.9505 -19.0753 0.0000 -41.0287 -33.3830
purpose_home_improvement   -37.0914   1.9666 -18.8605 0.0000 -40.9459 -33.2369
purpose_major_purchase     -37.0823   1.9627 -18.8931 0.0000 -40.9291 -33.2354
purpose_small_business     -37.0323   1.9818 -18.6864 0.0000 -40.9165 -33.1481
==============================================================================
```

Table 3. The Multiple Logistic Regression Results

(1) Ceoffceint Interpretation for *'log.annual.inc'*

We believe that wage is a crucial parameter for LendingClub to make lending decisions. The initial assumption is that if one borrower has a higher annual income, he or she might be employed in a promising industry with a steady source of revenue. Therefore, we expect annual income to have a positive impact on meeting underwriting borrowing criteria at LendingClub. The lenders must be more confident that a relatively wealthier borrower will be able to repay the loan on time.

Before plugging *'log.annual.inc'* into our logistic regression equation, we noticed that the wage is right-skewed and would cause bias to the estimator by having outliers, resulting in an inaccurate relationship between the right-hand side and left-hand side variables. Logarithmic transformation is a convenient means of transforming a highly skewed variable to become more compatible with the data. In Figure 3, after we logged the annual income, the independent variable *'log.annual.inc'* is quite close to the Gaussian normal distribution. Also, the interpretation of the variable becomes the percentage of change in this variable, not the absolute change.

To interpret the coefficient 0.7887 of *'log.annual.inc'* in Table 3, we reason that as the borrower increases one percentage unit of annual income, the odds of meeting the borrowing criteria will increase by a factor of e^0.7887, holding all other independent variables constant. The empirical result is compatible with the assumption that under a 5% significant level, when a borrower's annual income percentage increases, the probability of meeting borrowing criteria at LendingClub increases.

The empirical result aligns with the real-life context. Income, as a crucial factor in eligibility at the LendingClub platform, convinces lenders that borrowers have the means to pay back the loan. LendingClub, therefore, empowers borrowers who have a sound annual income with a high borrowing capacity [8]. Most lenders in the P2P market set a minimum income limit for loan approvals, which depends on the reported amount from the borrowers. Gig workers or freelancers who don't meet minimum income requirements or have steady paychecks are less likely to obtain loans unless they supply additional proof of their incomes or average out their payments on a month-by-month basis [9].

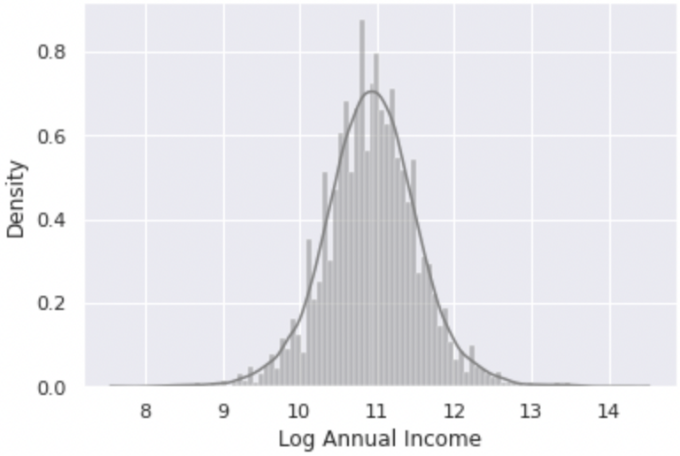


Figure 3. Distribution Visualization of *'log.annual.inc'*

(2) Ceoffceint Interpretation for *'fico'*

FICO score was introduced in 1989 by the Fair Isaac Corp [10]. The terms "credit score" and "FICO score" are often used interchangeably. A higher FICO score indicates better credit to one borrower, so our team assumes *'fico'* to have a positive impact on *'credit.policy'*.

FICO generally defines good credit as scores ranging from 670 to 739 [11]. Combined with Table 1 and Figure 4, we observe that the score range varies from 612 to 827, and the variable profiled different score intervals for different groups of borrowers. Some borrowers have FICO scores in the low range, while others have good FICO scores over 800. Hence, the numerical variable *'fico'* is a really reliable parameter to perform regression analysis since it contains information on all different interval ranges. To be noticed, we do realize that one individual's FICO score can be adjusted or changed through his or her financial behaviors, like paying bills on time. However, the cross-sectional dataset records data entries that occurred at the same specific time point, so we could ignore this effect in this study.

Recalled from Table 3, the regression result indicates that the coefficient for *'fico'* is 0.045. The empirical interpretation is that as the borrower increases one unit of FICO score, the odds of meeting the borrowing criteria will increase by a factor of e^0.045, holding all other independent variables constant. Therefore, a higher FICO score results in a higher chance to obtain a loan in the P2P market generally.

Again, the empirical result is consistent with our assumption that when a borrower's FICO score increases, the probability of meeting borrowing criteria at a Lending Club increases. Realistically, lenders use FICO scores to determine whether borrowers are responsible and hence assess their likelihood of loan default. Achieving a high FICO score requires borrowers to have a mix of credit accounts and maintain an excellent payment history; maxing out credit cards, paying late, and applying for new credit haphazardly could lower FICO scores, increasing the probability of not meeting borrowing criteria of LendingClub [12]. As shown in Figure 5, a borrower with a FICO score of 650 has a low, 50% probability of meeting borrowing criteria at LendingClub while a borrower with a FICO score of 800 has a really high probability of successful borrowing.

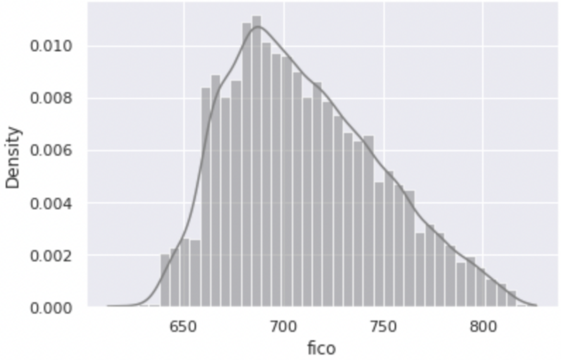


Figure 4. Distribution Visualization of *'fico'*

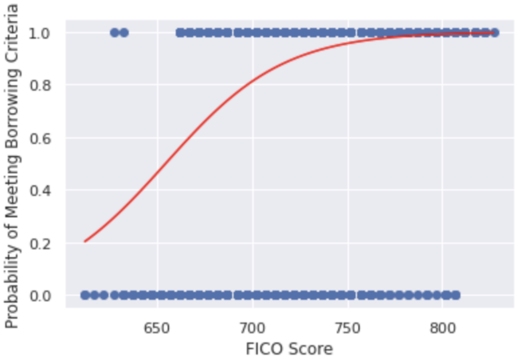


Figure 5. Logistic Regression of *'fico'*

(3) Ceoffceint Interpretation for *'inq.last.6mths'*

The numerical variable *'inq.last.6mths'* represents the number of credit reports a borrower applied in the last 6 months. A credit report is a statement that has information about one's credit activity and current credit situation, such as loan paying history and the balance of credit accounts [13]. To determine how much risk borrowers pose, they must submit a report each time to request loans from any loan institution. Combined with Table 1 and Figure 6, we observe that some borrowers did not submit any credit report before applying to LedningClub for a loan, while others submitted multiple credit reports. We then assumed that a high number of credit reports might affect the probability of meeting borrowing criteria at LendingClub.

The regression result indicates that the coefficient for *'inq.last.6mths'* is -0.9785 in Table 3. The interpretation is that when the borrower increases one number of the credit report in the last 6 months, the odds of meeting the borrowing criteria will decrease by a factor of e^-0.9785, holding all other independent variables constant. Therefore, when the number of times a borrower requests credit reports in the last 6 months increases, the probability of meeting borrowing criteria at LendingClub will decrease. It's consistent with the data visualization in Figure 7: borrowers with more than 10 inquiries have an extremely low chance to meet borrowing criteria while borrowers with zero inquiries have a significantly higher chance to meet borrowing criteria.

This coefficient is intuitively meaningful in real-life situations. Individuals with multiple inquiries on hand could indicate that they have tried to borrow money from different places. A higher number of inquiries indicates a higher financial demand, and it's possible that the borrower is at financial risk or trying to pay back loans by borrowing other loans, which increases the chance of loan default. Therefore, LendingClub is less likely to approve loans to borrowers with a high number of inquiries in the last six months.

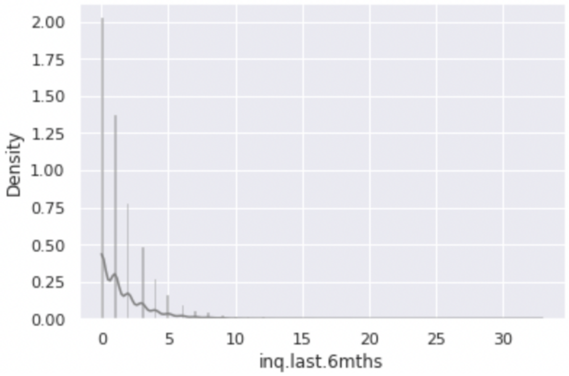


Figure 6. Distribution Visualization of *'inq.last.6mths'*

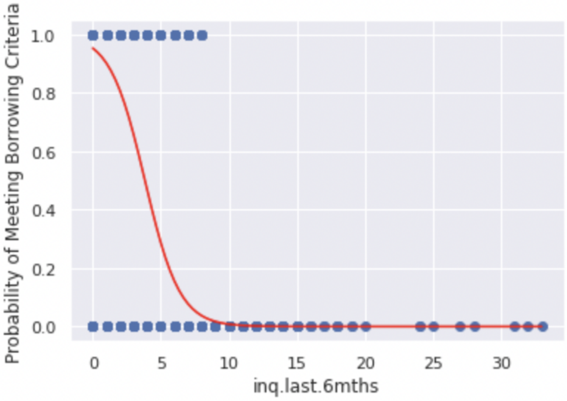


Figure 7. Logistic Regression of *'inq.last.6mths'*

(4) Ceoffceint Interpretation for all other independent variables

- When a borrower's monthly installments increase, the probability of meeting borrowing criteria at LendingClub increases under a 5% significant level.
- When the number of days a borrower has had a credit line increases, the probability of meeting borrowing criteria at LendingClub increases under a 5% significant level.
- When a borrower's revolving balance increases, the probability of meeting borrowing criteria at LendingClub decreases under a 5% significant level.
- When a borrower's revolving line utilization rate increases, the probability of meeting borrowing criteria at LendingClub increases under a 5% significant level.
- When the number of times the borrower has been overdue on a payment for the past two years increases, the probability of meeting borrowing criteria at LendingClub decreases under a 5% significant level.

- When a borrower has a record of loan default at LendingClub, the probability of meeting borrowing criteria at Lending Club decreases under the 5% significance level.

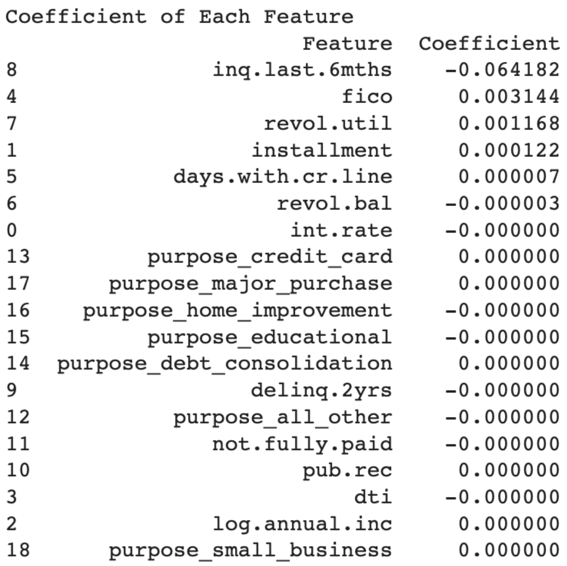
6.2 LASSO Regression

After we selected all features from the dataset, we noticed that dependent variables have different magnitudes or levels of impact on the independent variable. Lasso regression is a powerful tool for feature selection and model regularization, and it helps improve the performance and interpretability of our logistic model.

Lasso regression shows which independent variables have the strongest correlation or impact on the dependent variable since coefficient magnitude is directly related to the strength of the association between each independent variable and the dependent variable [14]. During the fitting process, our Lasso regression model is created with an alpha (hyperparameter) value of 0.1.

The larger magnitude of the coefficient, the stronger impact the corresponding independent variable has on the dependent variable. Based on Table 4, we found that the independent variables *'inq.last.6mths', 'fico', ' revol.util', 'installment' 'days.with.cr.line', and 'revol.bal'* are more important for predicting the target variable. The sign of the coefficients indicates the direction of the relationship between the predictor variable and the target variable. A positive coefficient means that as the predictor variable increases, the target variable also increases. In our result, *'fico', 'revol.util', 'installment',* and *'days.with.cr.line'* all have positive impacts on meeting borrowing criteria at LendingClub. A negative coefficient means that as the predictor variable increases, the target variable decreases. For example, as the variables *'inq.last.6mths'* and *'revol.bal'* increase, the probability of meeting borrowing criteria at LendingClub decreases.

Some coefficients are close to zero, suggesting that the corresponding features don't have a significant impact on the target variable [15]. For example, the features of *"purpose_credit_card", "purpose_major_purchase",* and *"purpose_home_improvement"* with coefficients close to zero are unimportant predictors of the target variable.



```
Coefficient of Each Feature
                       Feature  Coefficient
8               inq.last.6mths    -0.064182
4                         fico     0.003144
7                   revol.util     0.001168
1                  installment     0.000122
5            days.with.cr.line     0.000007
6                    revol.bal    -0.000003
0                     int.rate    -0.000000
13         purpose_credit_card     0.000000
17      purpose_major_purchase     0.000000
16    purpose_home_improvement    -0.000000
15         purpose_educational    -0.000000
14  purpose_debt_consolidation     0.000000
9                  delinq.2yrs    -0.000000
12           purpose_all_other    -0.000000
11              not.fully.paid    -0.000000
10                     pub.rec     0.000000
3                          dti    -0.000000
2               log.annual.inc     0.000000
18      purpose_small_business     0.000000
```

Table 4. The LASSO Regression Results

By examining only the non-zero coefficients of the variables, we can narrow down only the independent variables with the strongest correlation or impact on the dependent variable based on their coefficient magnitudes and signs.

6.3 Decision Tree

A Decision Tree is a type of predictive model in machine learning used to classify or predict outcomes based on a set of input data. It is a tree-structured model where each internal node represents a test on a specific feature of the data, and each leaf node represents a decision or a prediction. To build the machine learning model, we split 80% data into the training set and 20% data into the test set.

We employed cross-validation to choose the optimal settings for our machine-learning models [16]. It uses the available data as effectively as possible: it overcomes overfitting issues and selects the best hyperparameters for our Decision Tree classifier. Specifically, we used 5-fold cross-validation: it splits the training set into five subsets and uses one subset for validation and choosing the best hyperparameters and the other four for training each time, and this process repeats five times.

The GridSearchCV function is used to search for the best hyperparameters for both the Decision Tree and Random Forest [17]. Specifically, it's used to tune the hyperparameter of the maximum depth of the Decision Tree in our study. Choosing an appropriate value for max_depth is important because it helps to prevent overfitting or underfitting of the data. In general, a larger max_depth value allows the model to capture more complex relationships in the data but increases the risk of overfitting, while a smaller max_depth value reduces the risk of overfitting but may result in underfitting. By controlling the max_depth hyperparameter, we can balance the trade-off between model complexity and generalization performance.

We allow GridSearchCV to test each value from the list of 6, 8, 10, 12, 14, and 16 for the maximum depth using cross-validation. Then, the model is trained on the entire training set using the selected best hyperparameter value. Finally, we estimate the future performance of our model based on the accuracy score for its predictions on the test set.

The resulting Decision Tree has the best hyperparameter which is a depth of 14. The accuracy score of 99.1% for the cross-validation and accuracy score of 99.3% for the test set are extremely high. This suggests that our model is not subject to overfitting and will perform really well on future data. However, it is important to note that we should be careful about generalizing these results to other datasets since this dataset is relatively small.

Finally, we visualized how this Decision Tree with the best depth of 14 makes specific decisions in Figure 8.

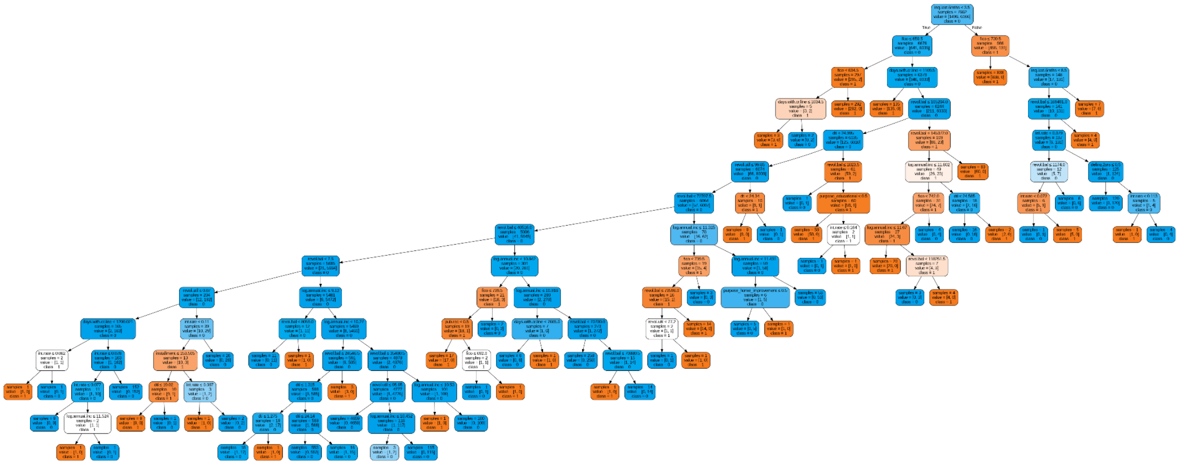


Figure 8. Visualizing Decision Tree Model with a Depth of 14

6.4 Random Forest

After performing the Decision Tree model, we found out that our dataset is suffering from an imbalanced nature, which is a typical problem for real-world datasets. Data imbalance can be best described by looking at a binary classification task [18]. In binary classification, the data set is imbalanced if the number of samples between classes zero and one is uneven. One major feature is the skewed distribution. A group having more data points is known as the majority class whereas the group having fewer data points is known as the minority class. In Figure 9, it's obvious that there are more people meeting the borrowing criteria in LendingClub, than people who don't. The imbalanced data could harm our regression and machine learning results [19]. Therefore, we employed a more advanced machine learning model called Random Forest, which is a type of ensemble learning method that uses multiple Decision Trees to make predictions. The Random Forest creates a diverse set of Decision Trees that work together to improve the overall accuracy and model stability as well as avoid overfitting issues [20].

The Random Forest solves the issue of imbalanced data by using bootstrapped samples and random subset features for each tree so that individual trees are trained on different data subsets, including the minority class data. The model then takes the majority vote of each tree to make the final prediction and the minority class data is weighted more in the voting process, which leads to better predictions [21].

We chose these two important hyperparameters of max_depth and n_estimators for the Random Forest Classifier. Similar to the Decision Tree, the max_depth hyperparameter controls the maximum depth of each Decision Tree in the forest. Selecting an appropriate value for max_depth is important to balance the bias-variance trade-off and improve the model performance. Secondly, the n_estimators hyperparameter controls the number of Decision Trees in the forest. Increasing the number of trees in the forest typically improves the accuracy of the model, but also increases the computational cost and may lead to overfitting. Therefore, selecting an appropriate value for n_estimators is important to both improve the model performance and maintain computational efficiency. We tuned max_depth from values of 15, 20, 25, and 30 and n_estimators from values of 150, 200, 250, and 300. The best hyperparameter values are 150 for n_estimators and 25 for max_depth. Our best Random Forest model has accuracy scores of 98.9% for cross-validation and 98.8% for the test set, indicating an estimation of good future performance.

When comparing the accuracy scores on the test set of these two machine learning models, we discovered that the Decision Tree's accuracy score of 99.3% is higher than the Random Forest's accuracy score of 98.8% even if the Random Forest is considered to be more advanced than Decision Trees. Through analysis, we found that two potential problems could cause this result.

Firstly, the machine learning technique focuses on tuning better hyperparameters. For Decision Trees, it costs little time to run since there is only one hyperparameter - max_depth in the model. We found that our Decision Tree model has easily reached the state of convergence, which refers to the point at which a machine learning algorithm has reached its optimal solution [22]. However, the Random Forest required more computing power than the Decision Tree, and every run required 6-10 minutes to tune different two combinations of hyperparameters. The time-consuming nature might prevent us from finding a Random Forest model with a better accuracy score [23]. Secondly, our dataset contains 9578 observations but Random Forests tend to work better on large datasets to effectively reduce overfitting [24]. If more objects are added to our dataset, the Random Forest model can perform better than the Decision Tree.

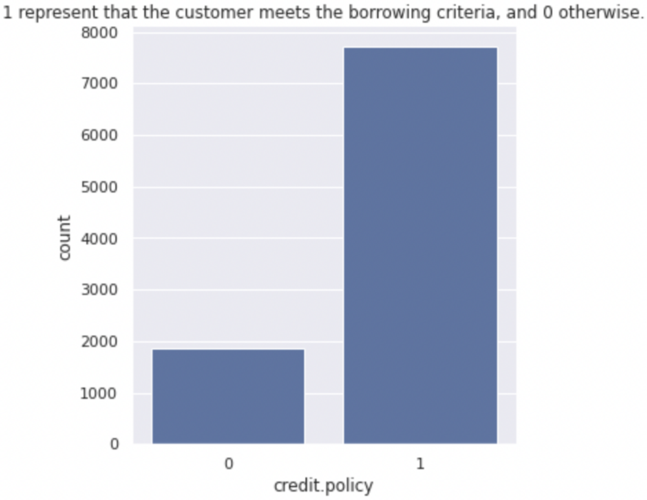


Figure 9. Visualization of *'credit.policy'*

6. 5 Confusion Matrix

We graphically displayed the prediction results of our Random Forest model on the test set in a Confusion Matrix that consists of 4 areas: false positives, false negatives, true positives, and true negatives [25]. Positives mean that the borrower meets the criteria for getting a loan and negatives mean otherwise. The matrix is shown below:

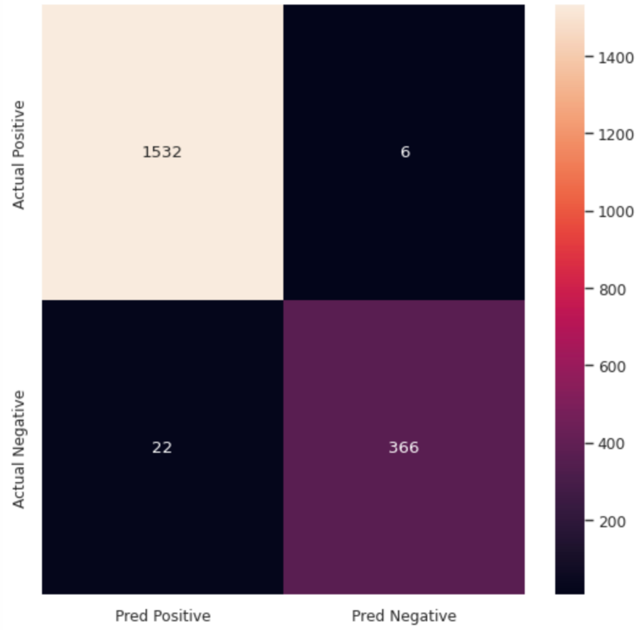


Figure 10. Confusion Matrix

Based on the number of total observations in the test set, the prediction results include 1532 true positives, 366 true negatives, 22 false positives, and 6 false negatives. A false positive means that our model predicts that the borrower meets the borrowing criteria but actually he/she doesn't; a false negative means that our model predicts that the borrower doesn't meet the borrowing criteria but actually he/she does. We could also calculate the False Positive Rate and the False Negative Rate:

$$FPR = FP / (FP + TN) = 22 / (22 + 366) \approx 0.0567 = 5.67\%$$
$$FNR = FN / (FN + TP) = 6 / (6 + 1532) \approx 0.0039 = 0.39\%$$

Among all the borrowers who don't meet the borrowing criteria, our model predicts 5.67% of them as meeting the criteria; among all the borrowers who meet the borrowing criteria, our model predicts only 0. 39% of them as not meeting the criteria. Comparing these two percentage terms, we classify a much higher percentage of unqualified borrowers as qualified, which harms the lenders from the LendingClub since there is a larger portion of borrowers they lend money to are actually unqualified and could default; on the other hand, we classify a much lower percentage of qualified borrowers as unqualified, which benefits the borrowers because our model's probability of depriving qualified borrowers of their opportunity to obtain financing is extremely low. As a result, our Random Forest model does have certain limitations due to its different impacts on different groups, which is elaborated further in the *Impact and Limitations* section.

7. Impact and Limitations

First, there are biases in the LendingClub dataset that might impact our results:

1. Sampling Bias: The dataset might not be representative of the entire population, as it was obtained from LendingClub.com. The data might only reflect the characteristics of the customers who applied for loans on

LendingClub and not the characteristics of the entire population of borrowers. For example, the variable called *'not.fully.paid'* limits the objects in the dataset to only borrowers who have had a borrowing history in LendingClub [26].

2. Measurement Bias: There might be bias in the way the data was collected, recorded, and measured due to errors in data entry, incomplete data, or inconsistencies in different variables' measurements [27]. For example, the codebook indicates that the variable *'log.annual.income'* is self-reported (reporting bias) which contains bias and harms the Regression and ML model results.
3. Historical Bias: Historical bias in the context of the LendingClub dataset refers to the potential bias that may exist in the time period of data collection when certain groups were systematically disadvantaged in the lending process. For example, historically, there has been a pattern of discrimination against minority groups and women in the lending industry, which might lead to biased data and hence biased results [28]. Also, since our data were collected decades ago in the time period from 2007 to 2010, it might not be representative of the current situation. As a result, our models would fail to provide accurate predictions on the most recent borrowers.

Second, our Random Forest model has limitations on fairness that favor the borrower group over the lender group in the LendingClub.

As analyzed in the Confusion Matrix section under *Results*, we discovered that the False Positive Rate of our model on the test set is 5.67% and the False Negative Rate is 0.39%. It's estimated that in the future, our model will mistakenly predict 5.67% of unqualified borrowers as qualified but only predict 0.39% of qualified borrowers as unqualified. The huge difference in these two rates indicates that, despite having the highest accuracy, our model benefits borrowers for almost always giving credits to qualified borrowers but harms lenders for convincing them to lend money to a relatively larger percentage of unqualified borrowers. These biased impacts would only be enlarged when the model is utilized in real life on more and more borrowers beyond our 9578 observations in this dataset. There is always a tradeoff between accuracy and fairness in ML models and people who tend to deploy our model should be aware of its limitations and ensure that it aligns with their expectations for fairness and accuracy [29].

Also, when allowing our model to make predictions based on these borrowers' identities in our dataset, we assume that the WYSIWYG(What You See Is What You Get) worldview is true [30]. We believe that these observed borrower features accurately reflect how they are responsible and capable enough to be qualified to pay back the loan. However, it's possible that the observed space of these limited borrower identities doesn't accurately represent the constructed space of our interest and there are structural biases or other excluded but relevant factors due to the lack of data. This source of limitation could also negatively impact our model predictions for the decision space.

8. Challenge Goals

Challenge 1 (**Result Validity**): We employ **p-value** and **hypothesis testing** with a significant level of 5 % for all the coefficient results of the multiple logistic regression derived by the library called statsmodels.api. We use the following null and alternative hypotheses to **test the statistical significance of our results**. The null hypothesis states that all coefficients in the model are equal to zero. The alternative hypothesis states that not all coefficients are equal to zero.

$$H_0 \text{ (Null Hypothesis): } \beta_1 = \beta_2 = \beta_3 = \beta_4 = \beta_5 = 0 \qquad H_1 \text{ (Alternative Hypothesis): } coefficient \neq 0$$

We further validate our results by interpreting the coefficients. For the logistic regression, by exponentiating coefficients to obtain the odds ratio, we obtain the multiplicative effect of a one-unit increase in the predictor variable on the odds of the outcome variable. We then analyze these effects' magnitude and signs in the real-world context to examine their validity. For the LASSO regression, we select features with non-zero coefficients that have the strongest impact on the dependent variable and again analyze their magnitudes and signs in real-world situations.

Challenge 2 (**Machine Learning**): We split our data into the training set and the test set. We train and use cross-validation to select **our Decision tree model** and **Random Forest model** with the optimal hyperparameter values. We then estimate and compare these two models' future performance for predicting whether a borrower meets the borrowing criteria or not based on their accuracy scores on the test set. Lastly, we analyze the errors of our Random Forest model on the test set.

9. Work Plan Evaluation

Our proposed work plan turned out to be quite accurate for estimating the total amount of working hours, but not accurate for estimating individual tasks' time. Some tasks were easier than expected and took less time and effort, such as initial data visualizations and LASSO regression. Since Huisheng has learned the pandas, seaborn, and matplotlib libraries really well in this class, he quickly created efficient tables and graphs to display a solid dataset introduction and some interesting facts. Also, since all three members are taking the machine learning course together, we are familiar and comfortable with the LASSO concept. The computational workload for LASSO turned out to be not intensive. However, certain tasks were more time-consuming than planned because we encountered new, unexpected smaller problems to solve along the way, such as implementing the Decision Tree and Random Forest models. We took time to select and decide which hyperparameters to include in our models and learned how to find the best hyperparameter values using the GridSearchCV function with cross-validation to help choose the optimal model. We also managed to create a function that plots the Decision Tree model and visually displays how it makes decisions. Additionally, we had a lot more discussions than expected about potential biases in our datasets and limitations with our machine learning models by incorporating the TA mentor's suggestions and the relevant course content. Thinking critically and comprehensively about any machine learning model is a necessary and serious step. In conclusion, the implementation stage was different from the planning stage and we encountered some unexpected issues to solve and decisions to make. But overall, easier tasks were balanced with more challenging tasks and we successfully finished this project within a manageable time.

10. Testing

11. Collaboration

In this section, we state online resources that we consulted during the project aside from the TAs and team members.

- The following resources are for introducing the P2P market as a whole, and how LendingClub works with both sides of lenders and borrowers.

[1] Epstein, L. (2022, September 23). Credit Unions vs. banks: Which one is the best for you? Investopedia. Retrieved December 16, 2022, from https://www.investopedia.com/credit-unions-vs-banks-4590218

[2] *What you need to know about prosper loans - consumer reports*. What You Need to Know About Prosper Loans - Consumer Reports. (n.d.). Retrieved December 16, 2022, from https://www.consumerreports.org/cro/news/2015/01/what-to-know-about-lending-club-and-prosper-peer-t o-peer/index.htm

[3] *What is peer-to-peer lending and how does it work?* Business News Daily. (n.d.). Retrieved December 16, 2022, from https://www.businessnewsdaily.com/16480-peer-to-peer-lending.html

[4] Martin, A. (n.d.). *What documents are required for a personal loan?* Bankrate. Retrieved December 16, 2022, from https://www.bankrate.com/loans/personal-loans/documents-required-for-personal-loan/

- The following resources explain why we choose to use Logistic regression instead of Linear regression, and why employing Hypothesis Testing in statistical methods can be crucial.

[5] *Linear regression vs logistic regression - javatpoint*. www.javatpoint.com. (n.d.). Retrieved December 16, 2022, from https://www.javatpoint.com/linear-regression-vs-logistic-regression-in-machine-learning

[6] *Logistic Regression Analysis - sagepub.com*. (n.d.). Retrieved March 10, 2023, from https://www.sagepub.com/sites/default/files/upm-binaries/36660_8.pdf

[7] Zach. (2021, September 29). *Understanding the null hypothesis for logistic regression*. Statology. Retrieved December 16, 2022, from https://www.statology.org/null-hypothesis-of-logistic-regression/

- The following resources explain that under the context of LendingClub (the representative of the secondary market), how our empirical results align with real-life situations. This information can further approve the validity of regression results. However, due to content limitation, we only explain the coefficients from *'log.annual.inc', 'fico', and 'inq.last.6mths'.*

[8] Cupler, J. (2022, August 7). *How to increase your borrowing power and get more credit*. Tally. Retrieved December 16, 2022, from https://www.meettally.com/blog/how-to-increase-your-borrowing-power

[9] *Personal loan eligibility criteria lenders look for*. LendingClub. (n.d.). Retrieved December 16, 2022, from https://www.lendingclub.com/loans/resource-center/personal-loan-eligibility

[10] About the author: Bev O'Shea is a former credit writer at NerdWallet. Her work has appeared in the New York Times. (n.d.). *What is a FICO score? FICO score vs credit score*. NerdWallet. Retrieved December 16, 2022, from https://www.nerdwallet.com/article/finance/fico-score

[11] DeNicola, L. (2023, February 22). *What is a good credit score?* Experian. Retrieved March 10, 2023, from https://www.experian.com/blogs/ask-experian/credit-education/score-basics/what-is-a-good-credit-score/

[12] *What is a FICO score and why does it matter?* LendingTree. (n.d.). Retrieved December 16, 2022, from https://www.lendingtree.com/credit-repair/what-is-a-fico-credit-score/

[13] *What is a credit report?* Consumer Financial Protection Bureau. (n.d.). Retrieved December 16, 2022, from https://www.consumerfinance.gov/ask-cfpb/what-is-a-credit-report-en-309/#:~:text=A%20credit%20repo rt%20is%20a,status%20of%20your%20credit%20accounts

- The following resources focus on an explanation of LASSO regression. Specifically, how to interpret coefficients with numbers and zero coefficients in this regression method.

[14] *Regression analysis*. Regression Analysis - an overview | ScienceDirect Topics. (n.d.). Retrieved March 10, 2023, from https://www.sciencedirect.com/topics/medicine-and-dentistry/regression-analysis

[15] *5.4 - the lasso: Stat 508*. PennState: Statistics Online Courses. (n.d.). Retrieved March 10, 2023, from https://online.stat.psu.edu/stat508/lesson/5/5.4

- Since in CSE163, we have already learned about the Decision Tree model. The following resources are more focused on explaining the mechanism of cross-validation, and how the GridSearchCV function can help us to find the best hyperparameters for ML models.

[16] Joby, A. (n.d.). *What is cross-validation? comparing machine learning models*. Learn Hub. Retrieved March 10, 2023, from https://learn.g2.com/cross-validation

[17] *Sklearn.model_selection.GRIDSEARCHCV*. scikit. (n.d.). Retrieved March 10, 2023, from https://scikit-learn.org/stable/modules/generated/sklearn.model_selection.GridSearchCV.html

- The following resources help us to understand what is Data Imbalance, and why the Random Forest can help us to minimize the negative impact of Data Imbalance on prediction results. After we built the ML model, we surprisingly found that the Random Forest model did not return a better accuracy than the Decision Tree model. Although the differences between the two accuracy scores are not big (1% difference), we still want to explore this unusual situation under the evidence from the internet.

[18] Mazumder, S. (2022, December 1). *What is imbalanced data: Techniques to handle imbalanced data*. Analytics Vidhya. Retrieved March 10, 2023, from https://www.analyticsvidhya.com/blog/2021/06/5-techniques-to-handle-imbalanced-data-for-a-classification-problem/

[19] Cheruku, S. K. (n.d.). *What is imbalanced dataset and its impacts on machine learning models?* LinkedIn. Retrieved March 10, 2023, from https://www.linkedin.com/pulse/what-imbalanced-dataset-its-impacts-machine-learning-models-cheruku/

[20] *What is Random Forest?* IBM. (n.d.). Retrieved March 10, 2023, from https://www.ibm.com/topics/random-forest#:~:text=Random%20forest%20is%20a%20commonly,both%20classification%20and%20regression%20problems.

[21] *Using random forest to learn imbalanced data - University of California* ... (n.d.). Retrieved March 10, 2023, from https://statistics.berkeley.edu/sites/default/files/tech-reports/666.pdf

[22] *On the adaptive properties of Decision Trees - NeurIPS*. (n.d.). Retrieved March 10, 2023, from https://proceedings.neurips.cc/paper/2004/file/6412fef87392ae8c987b0ecc79da1902-Paper.pdf

[23] *What is Random Forest?* IBM. (n.d.). Retrieved March 10, 2023, from https://www.ibm.com/topics/random-forest#:~:text=Time%2Dconsuming%20process%3A%20Since%20random,for%20each%20individual%20decision%20tree.

[24] Singh, A. (2023, January 7). *Random forests: The go-to algorithm for complex data sets*. Medium. Retrieved March 10, 2023, from https://medium.com/@Ambarish_224/random-forests-the-go-to-algorithm-for-complex-data-sets-aacbfaef57a2

- The following resources talked about the concept of the Confusion Matrix and the definition of its labels.

[25] Narkhede, S. (2021, June 15). *Understanding confusion matrix*. Medium. Retrieved March 10, 2023, from https://towardsdatascience.com/understanding-confusion-matrix-a9ad42dcfd62

- The following resources help us to understand study limitations through the ideas of bias (Sampling Bias & Measurement Bias & Historical Bias) and ideas of fairness (the Confusion Matrix & WYSIWYG worldview)

[26] Wikimedia Foundation. (2023, February 8). *Sampling bias*. Wikipedia. Retrieved March 10, 2023, from https://en.wikipedia.org/wiki/Sampling_bias

[27] Metwalli, S. A. (2021, February 24). *5 types of machine learning bias every data scientist should know*. Medium. Retrieved March 10, 2023, from https://towardsdatascience.com/5-types-of-machine-learning-bias-every-data-science-should-know-efab28041d3f

[28] PhD, M. R. (2021, April 2). *Understanding bias and fairness in AI Systems*. Medium. Retrieved March 10, 2023, from https://towardsdatascience.com/understanding-bias-and-fairness-in-ai-systems-6f7fbfe267f3

[29] Bhandari, A. (2023, March 9). *Confusion matrix: Interpret & implement confusion matrices in ML*. Analytics Vidhya. Retrieved March 10, 2023, from https://www.analyticsvidhya.com/blog/2020/04/confusion-matrix-machine-learning/

[30] Schafer, H. (n.d.). *Introduction to machine learning*. CSE/STAT 416. Retrieved March 10, 2023, from https://courses.cs.washington.edu/courses/cse416/21sp/