STAT 571 HW4 Lending Club

Chenxi Leng, Jihan Zhang, Ziyi You

Mar 2022

1 Executive Summary

Before starting diving into the data and technical parts, we will start some background information on the data set we are working with. Lending Club is a peer to peer lending company based in the United States. It is the biggest P2P lending platform in the U.S. Lending Club investors provide funds for potential borrowers and investors earn a profit depending on the risk they take, which is the borrowers credit score. Lending Club provides the "bridge" between investors and borrowers. The platform cuts the middleman, a traditional financial institution, and connects multiple investors and potential borrowers to invest capital and to borrow credit. The borrower can put up a loan request which consists of a description of the loan purpose with their personal financial information. The investor then has the privilege to choose the amount of capital they would like to invest and also have the ability to choose the borrower. In our report, we will look at the data obtained from this platform.

We will restrict our analysis to the period between 2007-2011. There are around 39,000 observations and 38 attributes for each of these loans. Financial decision making regarding to the credit risks is one of the crucial operations for the lending businesses. In this analysis, I present here exploratory data analysis, visualizations, and lots of other interesting insights.

In this report, we will start by exploring through the data set through visualization, we will look at the distribution of demographics, loan purpose, employment length of the borrowers. Then, we will do feature engineering to build a logistic regression model to characterize the relationship between risk factors to the chance of a loan being defaulted. We try to identify important risk factors that a loan will be defaulted and build a classifier that maximizes the return. The factors include number of payments on the loan, interest rate, annual income, public records etc. Overall, our model is quite robust. For 7795 transactions, Lending Club's profit may increase amount ranging from 13,487 dollars to 32,111 dollars applying our model under the estimate loss ratio of 2:1. However, there are still limits. We can improve the model if we have factors relating to the macro economic situations because each year differs. The results, details, and limits will be discussed fully in the "Conclusions" part.

2 Exploratory Data Analysis

2.1 Nature of the data

There are two versions of the data, full and clean. For this report, we will mainly focus on the clean version of the data. The data set contains 38971 rows and 38 columns and no missing values.

Among those, we have 33503 observations with fully paid loan status and 5468 observations with charged off loan status.

2.2 Quantitative and Graphical Summaries

We first find out the mean of loan amount by the loan status. The "Charged Off" loan status has an average amount of 12202.94 dollars and "Fully Paid" status, on the other hand, has an average of 11114.58 dollars. We can see that although there is no significant difference in dollar amount, people who "charged off" still has a slightly higher mean than those who paid in full. We then plot the distribution of the LC grade, which is a category for credit scores. The lower the grade the higher the interest the customer had to pay back to investors. Because the lower the grade of the credit score, the higher the risk for investors. Below is a side-by-side histogram that shows the distribution of loan status by LC grade, we can see that grade B is the most common one. It is worth noticing that a borrower not paying their loan can occur in all grade assignments. Thus, only investing in borrowers that have an assignment of 'A' would still in some cases observe the borrower in not being able to pay.

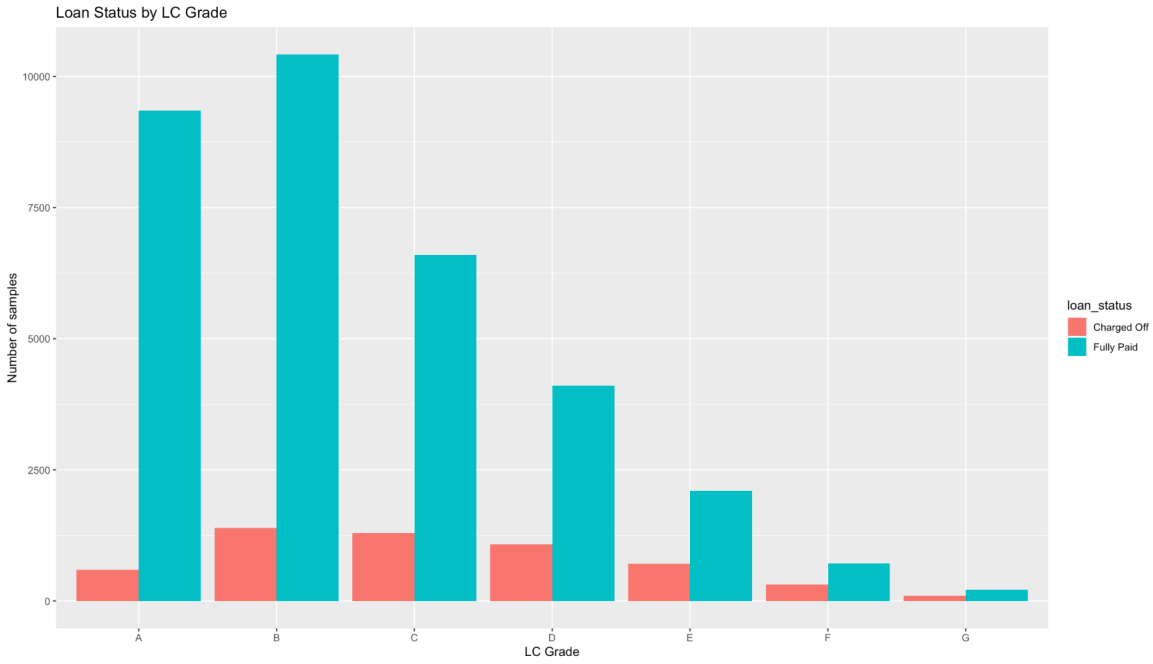

Figure 1: Loan Status by LC Grade

Next, we move on to explore the home ownership type of borrowers. "Loan Status by Home Ownership" graph in Appendix below shows that most borrowers do not own houses, they do rent or mortgage instead. People who rent houses have the greater percentage of "charging off". This

makes sense because people who rent houses probably are less stable or have a lower income than those who own houses, they might also be younger than those who have houses. So we want to figure out the financial background of the borrowers and looked at the average annual income, people who charged off have an average of 62637.80 and those who paid in full have an average of 70082.01.

After a general understanding of the financial backgrounds of the borrowers, we now look at the purpose of loans. The top three reasons for loan are: debt consolidation, credit card, and home improvement. The distribution of other categories can be shown below.

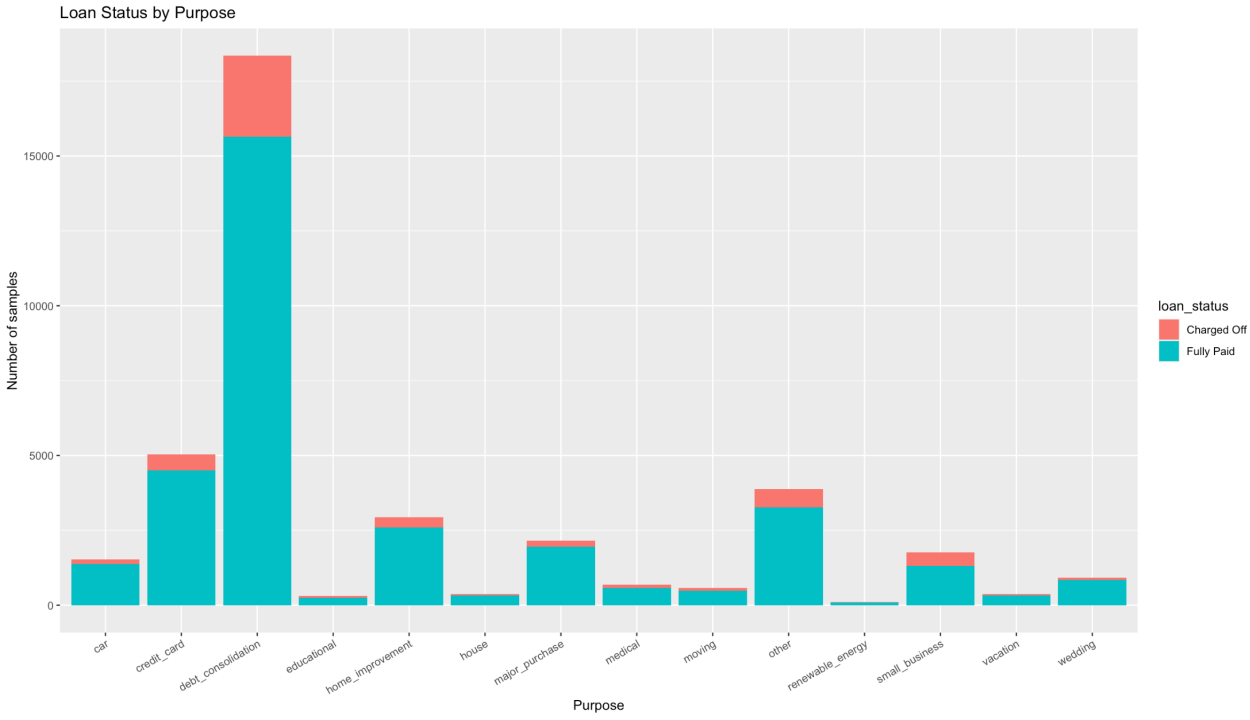

Figure 2: Loan Status by Purpose

Then we observe the demographics of the borrowers. The top five largest number of observations is in CA, NY, FL, TX, and NJ. There are 810 categories of different zip codes, so we might consider dropping this variable when modeling. The top five largest number of observations is with zip code 100xx, 945xx, 112xx, 606xx, 070xx.

We also take a look at the employment length. "Loan Status by Employment Length" graph in Appendix shows that the majority of borrowers have worked for more than 10 years. There are a significant number of people who have only worked for a year or even less.

3 Analysis

3.1 Feature Engineering

Because 'addr_state' includes the geographical information of the borrower and there are too many categories of 'zip_code', we drop 'zip_code'. There 526 categories of 'earliest_cr_line' and we only keep the year of this variable. After changing the variable type to numeric, we calculate the mean for 2 groups of different 'loan_status'. The two mean values are 1996.905 and 1996.515. We drop 'earliest_cr_line'. The month of 'issue_d' is not very informative so we only keep the year of this variable. After changing the variable type to numeric, we calculate the mean for 2 groups of different 'loan_status'. The two mean values of 'issue_d' are 2010.427 and 2010.371. We drop 'issue_d'. 'last_pymnt_d' and 'last_credit_pull_d' is not very informative for the 'loan_status', so we drop them as well. Most 'funded_amnt_inv' is equal to 'funded_amnt'. We keep 'funded_amnt' and drop 'funded_amnt_inv'. Similarly, we drop 'total_pymnt_inv'. Each category of 'emp_length' has a similar ratio of 2 groups of 'loan_status'. There are 1070 observations with "n/a" 'emp_length'. So we drop 'emp_length'.

After exploratory data analysis and feature engineering, we split the data into 'data.train', 'data.test', and 'data.val' with portions of 0.6, 0.2, and 0.2 respectively of data size *N*.

3.2 Model 1: Multiple Logistic Regression

We first fit all variables, excluding 'sub_grade' for simplicity. But the fitted probabilities are extremely close to zero or one. Some backward eliminations are conducted for model selection.

After several trial and errors, we used loan_status as the y variable and limited x variables to only 6, including term (the number of payments on the loan), int_rate (interest rate on the loan), annual_inc (the self reported annual income provided by the borrower during registration), inq_last_6mths (the number of inquiries in past 6 months), pub_rec (number of derogatory public records), and revol_util (revolving line utilization rate, or the amount of credit the borrower is using relative to all available revolving credit). All variables are significant. This model has an AIC score of 17737.

3.3 Model 2: LASSO in Logistic Regression

We then proceed to use LASSO to reduce the x variables. The final significant variables are addr_state (the state provided by the borrower in the loan application), term, int_rate, annual_inc (the self-reported annual income provided by the borrower during registration), inq_last_6mths, pub_rec, and revol_util. Compared with the previous model, the factors do not vary significantly.

3.4 Model Comparisons and Business Implications

We calculate the AUC score, the first model has an AUC score of 0.688 and the second model has an AUC score of 0.700. We estimate that the loss ratio of picking up a bad loan (false positive) to that of missing a good loan (false negative) is about 2 to 1. According to the Bayes' Rule, we pick the threshold of 1/3. Below are the confusion matrix of the two models.

fit1.pred	Charged Off	Fully Paid
0	177	297
1	1015	6306

Figure 3: Confusion Matrix of Model 1

fit2.pred	Charged Off	Fully Paid
0	204	313
1	988	6290

Figure 4: Confusion Matrix of Model 2

Overall, the two models do not vary much on their performances. We pick model 1 as our final model. Model 1 will increase Lending Club's profit under the estimated loss ratio of 2:1. Based on information we gathered of Lending Club (https://en.wikipedia.org/wiki/LendingClub#Business_model), it earns money by charging borrowers origination fee and investors service fee. Varying from the credit grade, the origination fee ranges from 1.1 percent to 5 percent of the loan amount and the service fee is 1 percent of borrower's payment. The average loan amount is 11267.29 dollars. For each transaction, Lending Club's average profit ranges from 236.62 dollars to 563.36 dollars. For 7795 transactions, Lending Club's profit may increase amount ranging from 13,487 dollars to 32,111 dollars applying our model under the estimate loss ratio of 2:1.

4 Conclusions

In this report, we did an end-to-end analysis on the Lending Club data. From exploratory data analysis, we can see that borrowers who charges off accounts for a higher percentage in people who rent houses, people who just started working, and people usually loan money to pay off debt or buy houses. We then worked on feature engineering and come up with a multiple logistic regression model and use LASSO to reduce the factors. We can see that the factors that determines the loan status include number of payments on the loan, interest rate, annual income, public records etc. For 7795 transactions, Lending Club's profit may increase 13,487 dollars to 32,111 dollars, applying our model under the estimate loss ratio of 2:1.

Despite the analysis and modelling, there are several limits to this report that can be improved upon. First, there are some factors the data set did not take into consideration, for example the last payment amount, which is which is the amount paid by the borrower on their last trade balance. Second, credit risk model could be further improved by incorporating the financial health of the economy per fiscal year. Since each year has different inflation rate and economic situations, the issuance of number of loans will be greatly impacted by the macro economics. Thus, if we have more time, we will incorporate other factors into the analysis through feature engineering, such as the last payment amount. We will also include economic indicator variables to improve our current model.

Appendix A Visualization

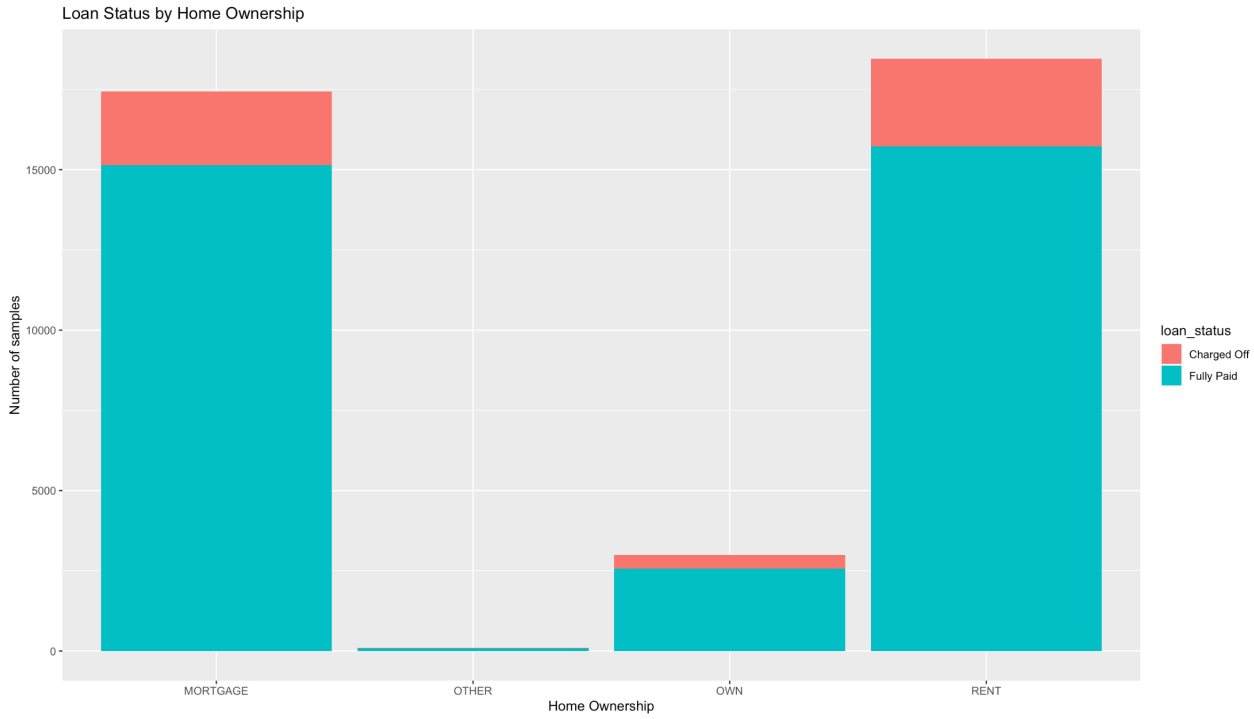

Figure 5: Loan Status by Home Ownership

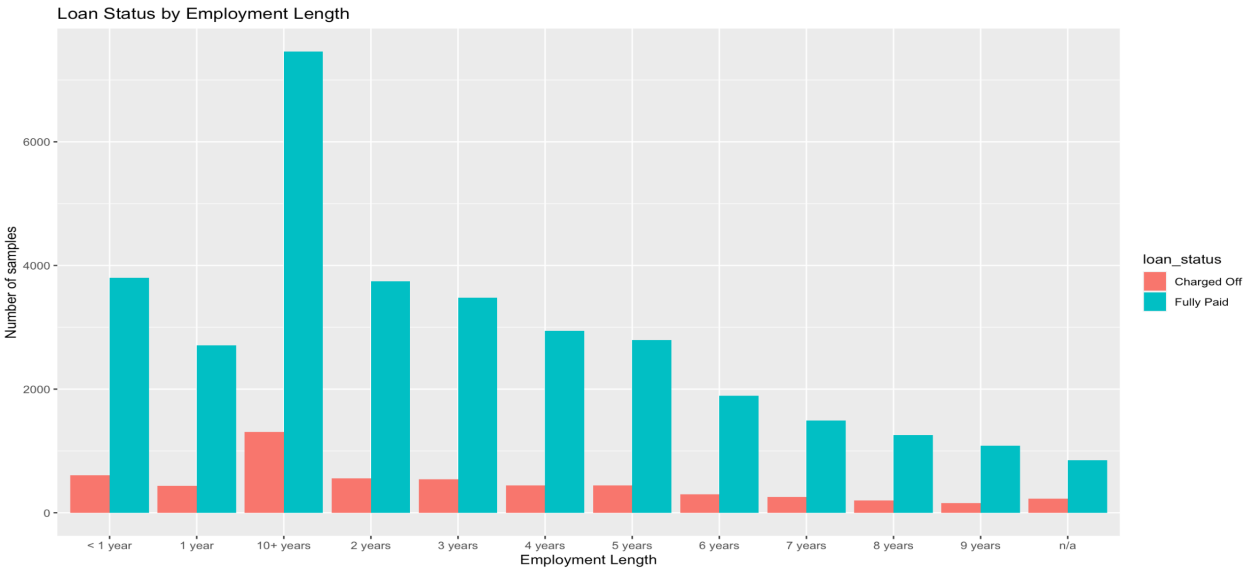

Figure 6: Loan Status by Employment Length