**Lending Club Loan Default Prediction**

Somayeh Youssefi

Fall 2021

**Executive Summary**

[LendingClub](https://www.lendingclub.com/) used to be the biggest peer-to-peer lending company and the world's largest peer-to-peer lending platform. In this project, I build a machine learning model to predict if a loan on Lending Club will default. The models could help Lending Club investors make better-informed investment decisions. As the data source, I used a 1.8 GB dataset with more than 2 million loans and 151 features

For training the models, I only used features that are known to investors before they choose to invest in the loan. These features include the applicant's income, FICO score, and debt-to-income ratio, and the loan amount, interest rate and grade. The modeling process includes cleaning and addressing missing data, transforming, and visualizing the data; creating dummy variables for categorical features; and fitting four models: logistic regression, decision tree, random forest, gradient boosting. I use pipelines to combine standardization and model and I optimized hyperparameters by grid search through cross validation. Here is a summary of modeling:

1. During 2007 to 2018, about 22% of borrowers have failed to return their loans. As a result, Lending Club has lost about $6 Billion.
2. Lending Club would have saved about $4 Billion (63% of the total loss), if they had used our model.
3. ***Loan term*** and ***interest rate*** are very influential in classification.
4. Other important factors arenumber of ***mortgage accounts*** and ***Fico score*** followed by less effective features: ***funded amount*, *home******ownership*** and ***debt to income ratio***
5. *Employment length*, number of *tax liens*, number of *collections* and *charged off* incidents within 12 months prior to application are insignificant in fate of a loan.
6. ***Loan term*** and ***interest rate*** are decided on by the Lending Club based on applicant's information.
7. Number of ***mortgage accounts*** and ***Fico score*** are both indicative of the applicant's credit history.
8. Lending Club collects a great deal of information about an applicant's credit history, as well as employment history, income, and home ownership. None of these items are effective in fate of a loan. Therefore, Lending Club can save money by avoiding collecting ineffective information.

**Background**

Lending Club used to be the biggest peer to peer lending platform until 2020, when they changed their business focus. During its operation, Lending Club would establish a platform for borrowers and investors where borrowers were allowed to create loan requests on its website. They were also required to provide their information like credit score, credit history, desired loan amount and the debt-to-income ratio.

Based on the data, Lending Club would decide if the loan request would be accepted and what the interest rate would be. Allowable loan range was between $1,000–40,000 and the return period was 3 or 5 years. Investors would make money from interest rates which were varied from 6% to 35%.

Lending Club would make money from charging an origination fee to its borrowers and a service fee to its investors. The interest rates that Lending Club was offering were better for borrowers and lenders than most of banks and financial services, as a result, Lending Club was highly received.

When a request was made, based on the requester’s information such as credit history, Lending Club would decide to either accept or reject the request. For the accepted loans, a credit grade ranging from A to G would be assigned. The credit grade along with other factors would determine the interest rate.

Credit grade was also important in determining the amount of the origination fee that the requester would be charged, and it would be between 1.1–5.0% of the loan amount. The investor would be charged a service fee which would be 1% of the total amount the borrower would pay. The total amount was the loan amount plus the paid interest. Therefore, higher interest rates would generate more revenue. However, higher interest rates also indicate a riskier loan which may end up on default and may cause investment loss for Lending Club. Therefore, having a method to predict the risk factor accurately is very desirable.

## **Data Source**

The dataset used for this project is from Kaggle website and can be found here. The dataset has 2,260,701 observations of loans granted between 2007 to 2018 with 151 features which include information about applicants such as credit score, income, and employment, as well as data about loans, such as their amounts, and terms.

## **Objective**

The primary goal of this project is to predict if a loan applicant would fail to pay back its loan fully at the time of application. Therefore, only features available during application are considered in modeling. The approach consists of following steps: data wrangling, exploratory data analysis, model training, model selection and applying.

## **Data Wrangling**

Chart, bar chart

Description automatically generatedThe focus of the project is to find a better model to decide about loan applications, **any feature that is not defined prior to the decision on the loan application is dropped**. Many of the available features in the source dataset refer to an applicant’s performance during the loan and are irrelevant for initial decision making. General features such as member identification numbers or url that do not provide meaningful information are also dropped. Furthermore, less than 5 percent of applications are joint. For this project, the focus is on single applications. Therefore, any columns and rows with data of the second applicants are dropped. More than 100 columns and some rows are dropped, when we check for the missing data, any columns with more than 70% missing points are also dropped. Furthermore, the loans in the dataset are in 5 categories: Current, Fully paid, Charged-off, Default and Late. The scope of this project is a binary classification; therefore, Current loans are dropped. A closer look at the data revealed that default is when the borrower is late and charged off happens when the creditor gives up hope on getting the loan back. So, default is the beginning of the process which may or may not lead to charge off. Therefore, Charged-off and Default categories are combined in one category and is called **Default**. As shown in the graph, defaulted loans make 21.6% of the data, which indicates our dataset is imbalanced and proper technique should be adopted to reach the optimum model.

The name and description of the features available in our data frame are as follows:

* annual\_inc: The self-reported annual income provided by the borrower during registration.
* chargeoff\_within\_12\_mths: Number of charge-offs within 12 months
* collections\_12\_mths\_ex\_med: Number of collections in 12 months excluding medical collections
* dti: A ratio calculated using the borrower’s total monthly debt payments on the total debt obligations, excluding mortgage and the requested LC loan, divided by the borrower’s self-reported monthly income
* earliest\_cr\_line: The month the borrower's earliest reported credit line was opened
* emp\_length: Employment length in years. Possible values are between 0 and 10 where 0 means less than one year and 10 means ten or more years.
* fico\_range\_high: The upper boundary of the borrower’s FICO at loan origination belongs to.
* fico\_range\_low: The lower boundary of the borrower’s FICO at loan origination belongs to.
* funded\_amnt: The total amount committed to that loan at that point in time.
* home\_ownership: The home ownership status provided by the borrower during registration or obtained from the credit report. The values are RENT, OWN, MORTGAGE, OTHER
* inq\_last\_6mths: The number of inquiries in past 6 months (excluding auto and mortgage inquiries)
* Installment: The monthly payment owed by the borrower if the loan originates.
* Int\_rate: Interest Rate on the loan
* mort\_acc: Number of mortgage accounts.
* open\_acc: The number of open credit lines in the borrower's credit file.
* pub\_rec\_bankruptcies: The number of open credit lines in the borrower's credit file.
* revol\_util: Revolving line utilization rate, or the amount of credit the borrower is using relative to all available revolving credit
* sub\_grade: LC assigned loan subgrade
* Tax\_lien: Number of tax liens
* term: The number of payments on the loan. Values are in months and can be either 36 or 60.

## **Exploring effects of different features on loan status distribution**

Below is a summary of our analysis followed by a more detailed explanation.

1. Length of employment, as well as home ownership is not a great predictor of fate of a loan
2. Applicants with higher credit scores are also better borrowers and it is shown that defaulted loans have applicants with lower credit scores.
3. Previous charged off accounts or bankruptcies in an applicant's credit history or the length of the applicant's credit history do not affect how he/she performs on his/her loan return. Furthermore, the applicant's balance and how much he/she uses their credit on the time of application (ie utilization rate) do not affect how he/she will return the loan. While the number of credits an applicant owns is not a determinant in the fate of a loan, applicants with more mortgages tend to perform better.
4. As the installment increases, the probability of a loan defaulting also increases. Furthermore, higher interest rates also result in higher chance of default. Also, while 30% of loans with 30 months return period default, this value doubles for loans with 60 months return period.
5. DTI, debt-to-income ratio which compares an applicant's total monthly debt obligations to their annual gross income is a good predictor of default, and as DTI increases chances of defaulting also increases.
6. There are 35 grades for loans starting from A1 and ending in G5. Briefly, as a loan moves from A1 to G5, the riskier the loan will be. It is clear from data analysis that as loan grade decreases from A1 to G5, the probability of defaulting increases; while less than 10% of grade-A loans default, there is about 50% probability that a G-grade loan defaults. So, grade can be a strong predictor of if a loan defaults. Furthermore, as shown moving from A1 grade to G5, interest rate also increases. It is seen that in G category, the interest rate reaches a plateau of about 30%.
7. We also checked to see if the state where an applicant lives affect how they perform on their loans. while in some states higher number of loans are issued per capita, applicants in different states are similar in their loan performance.
8. Principle component analysis on the numerical features also showed that, loan amount, dti, Fico scores and revolving balance are important in the first 2 principle components. Only about 32% of the variance is explained by the first two components.
9. We should note that, installment and interest rate, as well as subgrade are features determined by the lending club based on an application.

### **Loan status distribution vs grade**

There are 35 grades for loans starting from A1 and ending in G5. More details about grades can be found [here](https://www.lendingclub.com/foliofn/rateDetail.action). But briefly, as we move from grade A to G, the riskier the loan will be. Loan distribution vs grade is shown below. It is clear from the chart that as the grade level decreases from A to G, the probability of default increases; while less than 10% of grade-A loans default, there is about 50% probability that a G-grade loan defaults. So, grade can be a strong predictor if a loan defaults. Chart, bar chart, histogram

Description automatically generatedThere are 5 grades (A to G) and in each grade there are 7 sub-levels. It is clear from the chart that as the grade level decreases from A to G, the probability of default increases; while less than 10% of grade-A loans default, there is about 50% probability that a G-grade loan defaults. So, grade can be a strong predictor if a loan Chart

Description automatically generateddefaults. Below, a boxplot of interest rate versus sub grade is shown. Based on an applicant's qualifications, A1 is the best subgrade an applicant can have and G5 is the worst. Interest rates increases from grade A to G and reaches a plateau in grade G.

Chart, bar chart

Description automatically generated**Loan status vs Fico score**

[FICO credit scores](https://www.consumerfinance.gov/ask-cfpb/what-is-a-fico-score-en-1883/) are a method of quantifying and evaluating an individual’s creditworthiness. Scores range from 300 to 850, with scores in the 670 to 739 range considered to be “good”. Fico score is also one important factor to decide about a loan application. As shown below, Fully-paid loans have higher scores compared to defaulted loans. As an applicant's credit score increases, the chances that he/she defaults on his/her loans decreases.

**Loan status distribution vs interest rate and loan term**

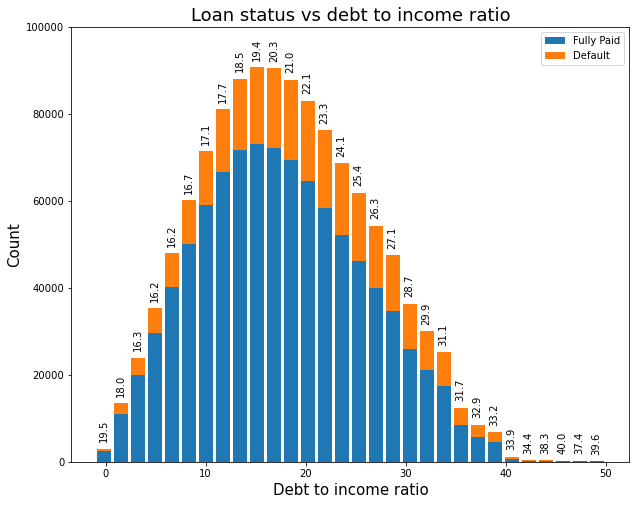
The effect of the loan interest rate and loans' term are shown below. Interest rates and term are factors which are determined by Lending Club. It seems that loans with higher amounts that result in higher installment and longer terms are riskier. When interest rate increases, default ratio also increases; for loans under lower than 10% interest rates, less than 10% of loans default, however, loans with 30% or higher interest rates has more than 50% chance of default.

Chart, bar chart

Description automatically generated

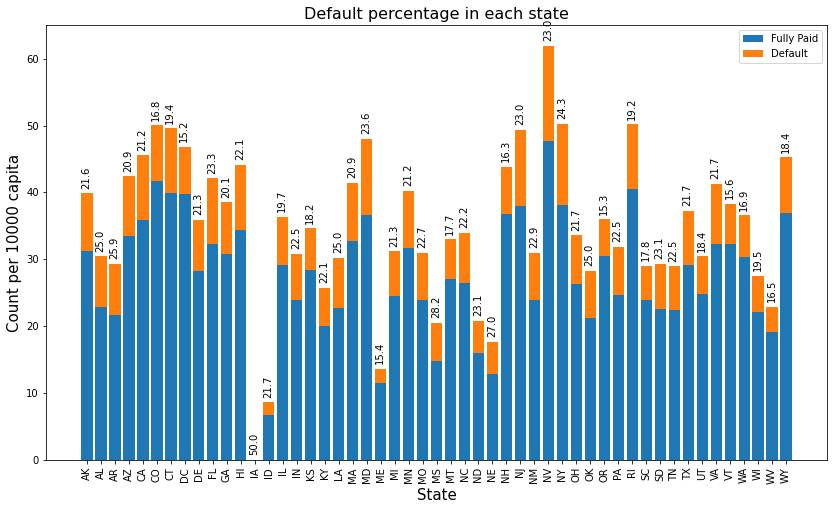
**Debt to income ratio**

Debt-to-income ratio compares one's total monthly debt obligations to their monthly gross income (before taxes). The DTI ratio gives lenders a clearer picture of an applicant's current debt and income, and it is used to determine how much money the applicant can afford to responsibly borrow. [DTI](https://www.lendingclub.com/loans/resource-center/calculating-debt-to-income) histogram is shown below and when DTI is high, the chances of an applicant defaults on their loan increases.



## **Does the applicant's residential state affect their performance?**

To find out if states are different, first the population of each state was found, and the number of loans issued per 10,000 was estimated in each state. Then the percentage of defaulted and fully paid loans were determined and shown in the figure. NV has the highest number of loans issued per 10,000 capita and IA has the lowest number of loans issued per 10,000 capita. However, 50% of loans issued in IA are finally defaulted. Does this mean that IA residents perform significantly worse than other states' residents? To answer this question, we had a closer look at IA. There are only two loans issued in this state, which is a very small number to make any hypothesis.



## **Model training**

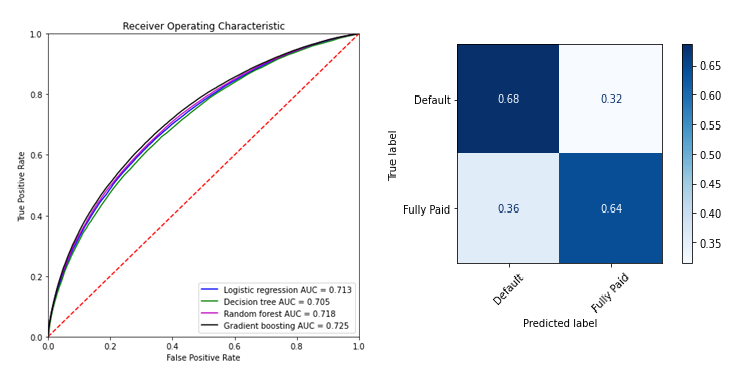
As mentioned earlier our dataset is imbalanced, and only 22% of the loans are in default category. Therefore, to overcome imbalanced data, **undersampling** was applied. To reach an optimum model, the data was separated into training and test sets and four classifiers were applied to the training data. For each model training, cross validation with 5 folds was applied to prevent overfitting. In each model training, a pipeline including scaling and modeling steps was created and by using GridSearchCV, hyperparameters were tuned, followed by feature importance analysis. The classifiers used are: Logistic regression, Decision tree Classifier, Random Forest, and Gradient boosting.

Given that the data is mildly imbalanced, accuracy score is not a good metrics to evaluate performance of a model. It is [recommended](https://www.kdnuggets.com/2017/06/7-techniques-handle-imbalanced-data.html#:~:text=1%20Use%20the%20right%20evaluation%20metrics.%20Applying%20inappropriate,keep%20the%20models%20as%20a%20fixed%20component.%20) to use evaluation metrics such as precision, recall, f1 score or balanced accuracy. We use **balanced accuracy** as our evaluation metrics to compare model performance.

The results of model training on the training data are summarized in the table. The standard deviation in cross validation is very small, which indicates models are not overfitting.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Logistic Regression | Decision Tree | Random Forest | XGBoost |
| Mean | 0.6533 | 0.6481 | 0.6532 | 0.6627 |
| Standard Deviation | .0012 | 0.0019 | 0.0018 | 0.0007 |

To choose the optimum the model, all models are applied on the test data and Receiver Operating Characteristic curve is plotted. The area under the curve is shown in the plot. As shown, Gradient Boosting performs slightly better than other models and we chose it as our final model. The classification report also is shown. The model predicts Default loans correctly in 68% of times.



## **Feature importance analysis**

All models reached to almost same balanced accuracy score of about %65. XGBoost reached 67% accuracy. As it is shown in the Feature importance plot, ***interest rate***, ***loan term***, the number of ***mortgage accounts***, followed by***Fico score*** are the most influential parameters in the classification. interest rate and term have more significant effect compared to other parameters. E*mployment length*, number of *tax liens*, number of *collections* and *charged off* incidents within 12 months prior to application insignificant.

A picture containing graphical user interface

Description automatically generated

## **Business implication**

Our model shows that a loan *interest rate* and its *term*, number of *mortgages* an applicant has, and his/her *fico score* are detrimental factors in fate of a loan. The first two terms are decided on by the Lending Club based on an applicant's information. The last two terms -the applicant's fico score and the number of mortgages he/she has- are both indicative of the applicant's credit history.

We should note that Lending Club asks for more than 20 pieces of information about an applicant's credit history, while only two of these items are important. Furthermore, other factors such as employment length, annual income and home ownership are not important in fate of a loan. Therefore, Lending Club can save money by avoiding collecting ineffective information.

The more important question to answer is that how much money would have been saved, if our selected model had been used prior to making decisions on loan applications in the data set. To find out the answer, we look through the original data file and estimate *loss of investment* for each defaulted loan. *Loss of investment* for a loan is defined as the difference between the expected amount of money by the end of the loan term and total payment received from an applicant. The expected amount is combination of the original funded amount and the interest paid and it can also be estimated by the loan term multiplied by monthly installment. After estimation how much money is lost through defaulted loans. We then used our selected model and estimate how much would have been saved if the defaulted loans had not been approved in the first place.

Our analysis showed that during 2012 and 2018, Lending Club has lost $5,998,214,435 in defaulted loans. If they had used our model to avoid approving high risk loans, they would have saved $3,764,828,296 (63% of their total loss). So, it can be concluded that although, building a predictive model is challenging, it is absolutely a wise financial decision, which can save any credit company great deal of fortune.