**MFIT5001**

**AI for FinTech**

**Predicting Default Risk of Lending Club Loans**

**Group number: 12**

|  |  |
| --- | --- |
| **Name** | **Student I.D.** |
| **SUN Yiying** | **20653607** |
| **YE Weikeng** | **20638695** |
| **LI Xinyi** | **20636398** |

**Introduction**

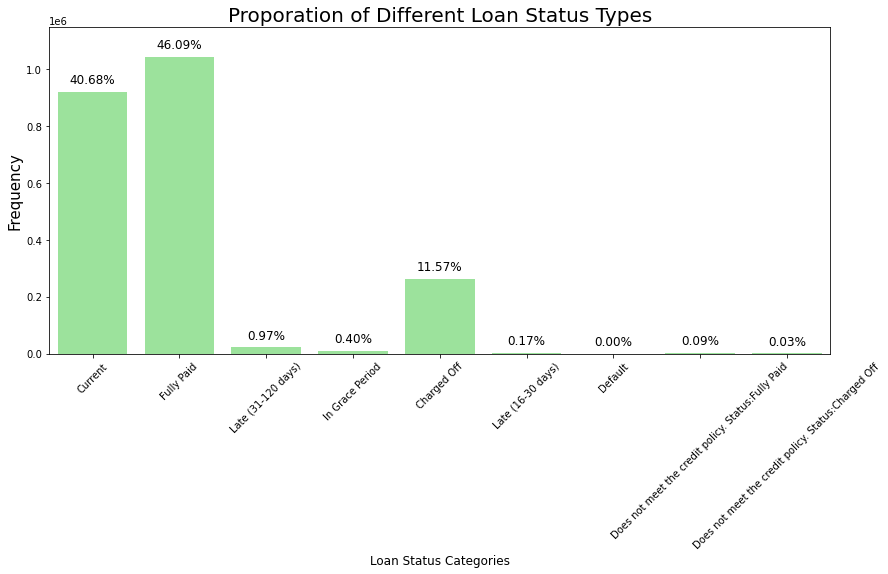
Lending Club is one of the largest peer-to-peer lending company in the world with $15.98 billion originated loans. It allows supply and demand for loans to be exchanged directly between investors and borrowers to avoid intermediaries, thus bringing down the cost of personal loans.  However, the default rate of P2P lending platform is much higher than that of traditional platform, and investors run the risk of investing in a default loan. In this project, we will try to predict the probability of default of the borrowers, in order to help avoid investment in high-risk loans.

**Exploratory Data Analysis**

This dataset contains complete loan data for loan issued during 2007 – 2015, including latest payment information and current loan status. The size of dataset is around 22 million 145. Features include loan amount, loan type, credit scores, borrower personal information, number of finance inquiries, address and collections among others.

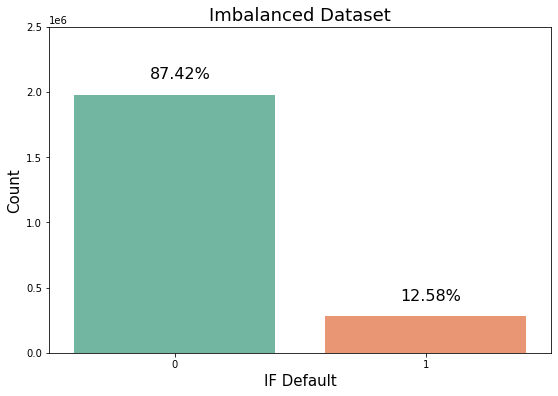
*“Target Variable: loan\_status”*

The chart below indicates the distribution of our target variable “*loan\_status”*:



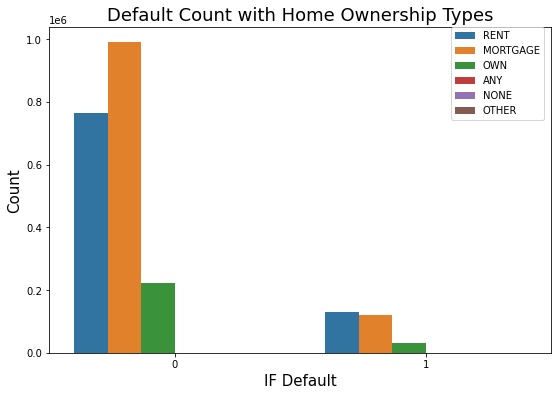
As shown in the chart above, the target variable “loan\_status” has 9 different categories: “Current”, “Fully Paid”, “Late (31-120 days)”, “In Grace Period”, “Charged Off”, “Late (16-30 days)”, “Default”, “Does not meet the credit policy. Status: Fully Paid”and “Does not meet the credit policy. Status: Charged Off”. The statuses “Current” and “Fully Paid” account for most of status types, at 40.68% and 46.09% respectively, followed by “Charged Off”.

In order to perform binary classification, these types are classified into two types. Statuses "Default", "Does not meet the credit policy. Status: Charged Off", "Late (31-120 days)" and "Charged Off" are classified as “Default” and others are classified as “Not Default”. The distribution for the binary target variable “Default” is shown in the following chart:



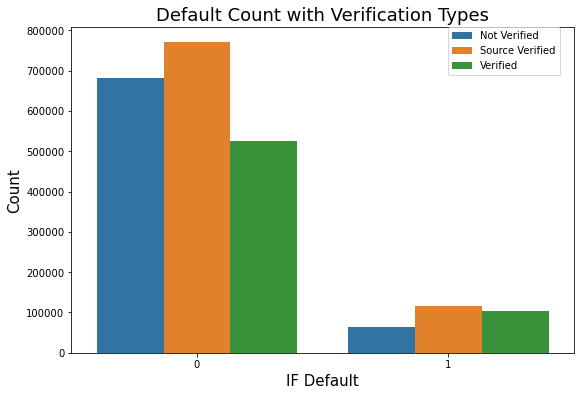
As shown in the chart, this dataset is largely imbalanced with only 12.58% of loans classified as default. Next, let’s explore the distributions for various features and their relationships with binary target “Default”.

*“Home Ownership Types”*



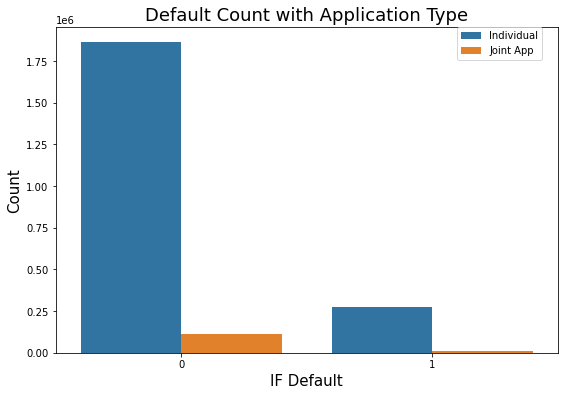
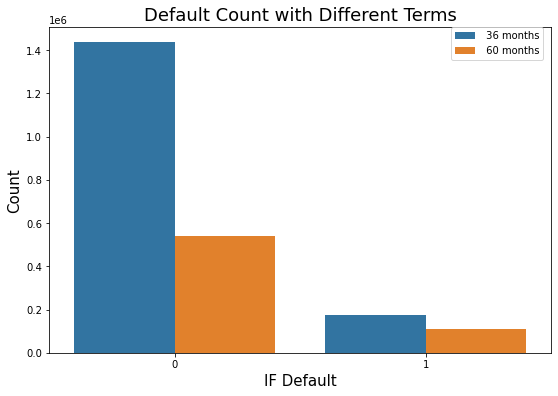
Among loans not default, most borrowers mortgage on their houses while among loans default, most borrowers rent to live. Only a small number of borrowers have their own house in both cases.

*“Verification Types”*



Borrowers with source verified account for the largest proportion in both cases while the proportion of borrowers not verified in the case of default is much samller than that in the case of not default.

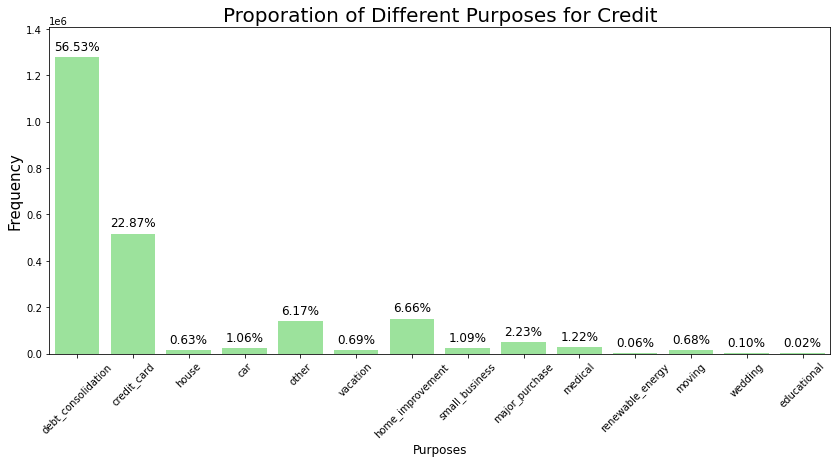
*“Application Type” & “Terms”*

Distributions for Application Type and Terms are nearly the same in both cases.

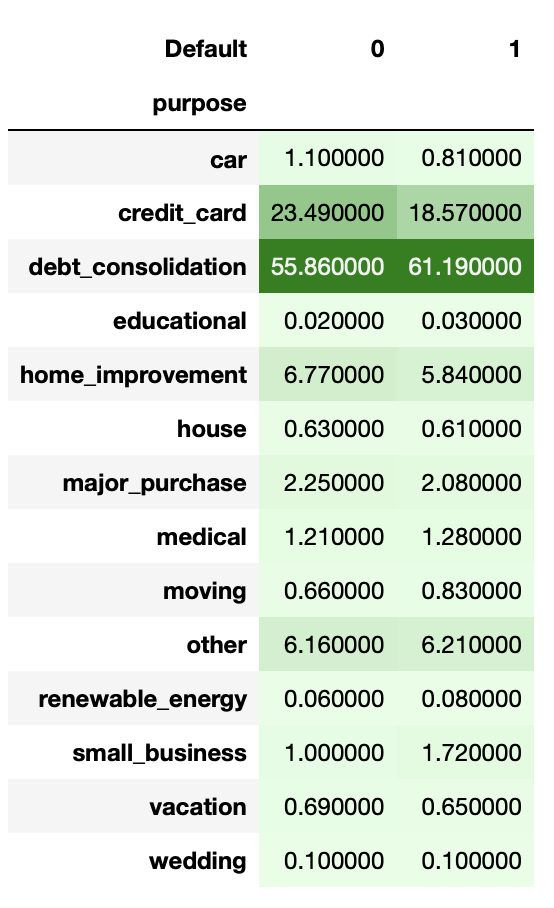
*“Purpose”*

Distribution for the categorical variable “purpose for credit” is indicated in the following chart:



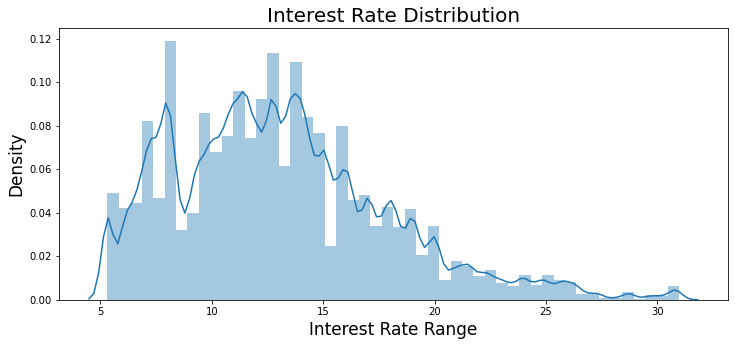
As shown in the chart, more than half borrowers borrow for debt consolidation, accounting for 56.53% of all borrowers. People borrow for credit card are the second largest group of borrowers, accounting for 22.87% of all. Only few people borrow for educational purpose, accounting for only 0.02% of all borrowers.

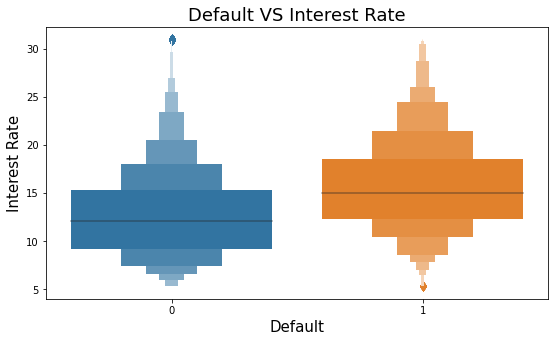
Next, let’s explore the relationship between purpose and default.



Distributions for two cases are nearly the same. Debt consolidation is the main purpose for default, accounting for 61.19% of default cases, followed by credit card. One thing noteworthy is that small business accounts for 1.72% of all default cases while it only accounts for 1% of all not default cases. Hence, compared with other purposes, lending to people who borrow for small business implies higher default risk.

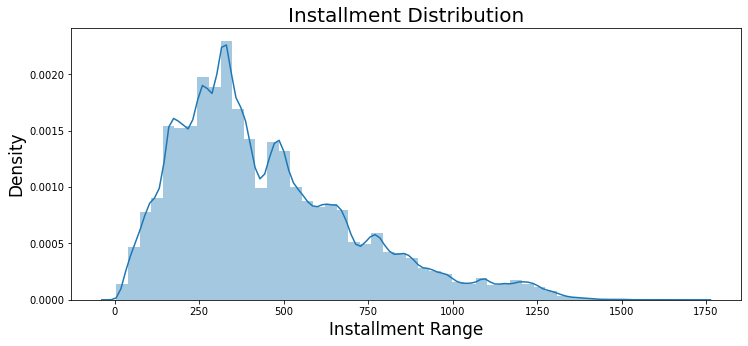
*“Interest Rate”*

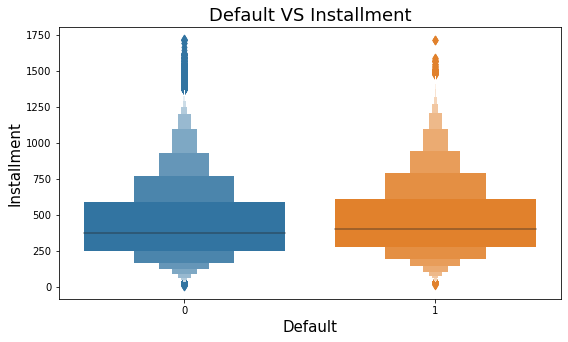




As shown in the PDF plot, variable interest rate is heavily right skewed with fat right tail implying high but rare interest rates. Interest rate is closely linked to risk. High interest rate is often required as a compensation for high default risk. As indicated in the box plot, default loans are generally issued with higher interest rate, implying higher risks.

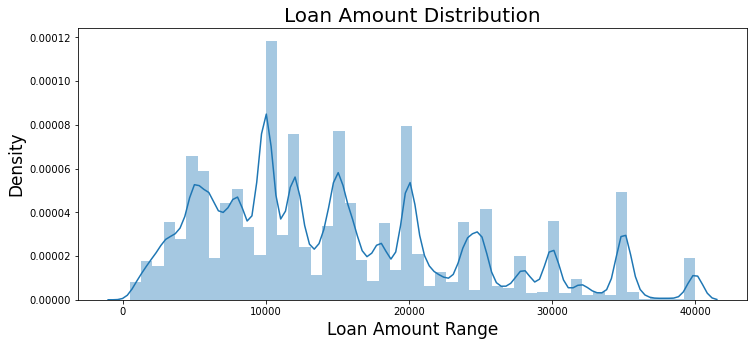
*“Installment”*

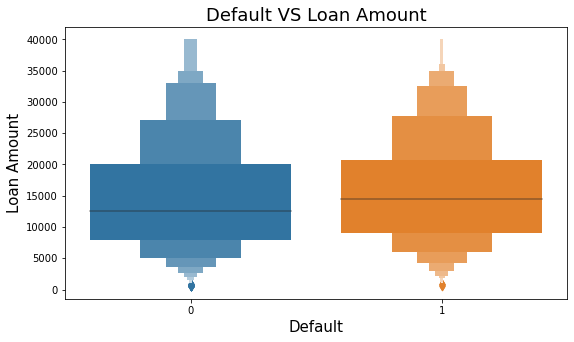




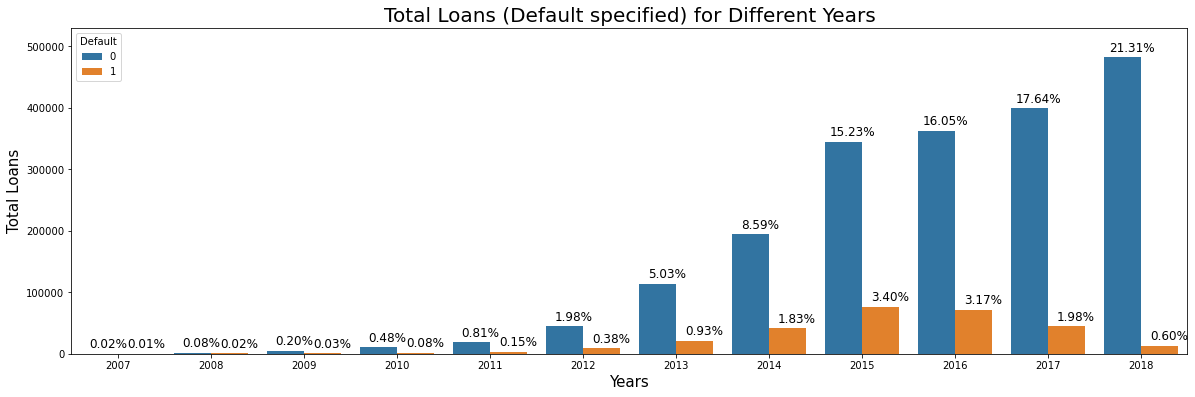
The distribution for installment is also heavily right skewed. As indicated by the box plot, there’s no big difference between these two cases although average installment for default loans is a little bit higher.

*“Loan Amount”*



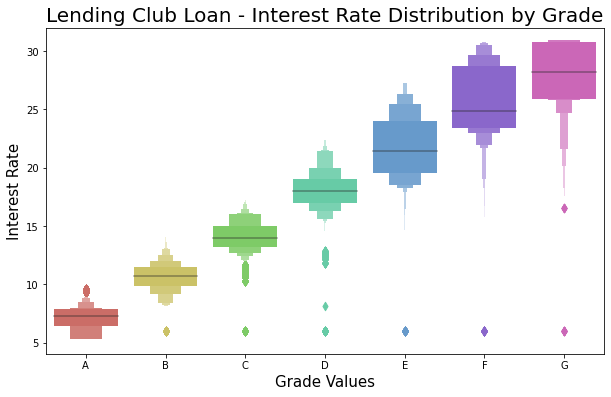
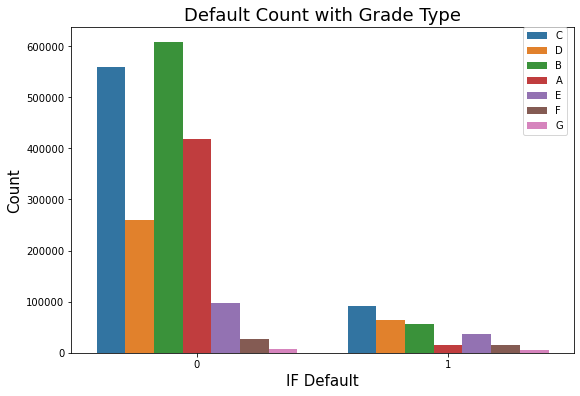


Variable loan amount is also right-skewed with a fat right tail. As indicated in the boxplot, default loans are generally associated with higher loan amount. Let’s look deeper into the relationship between loan amount and year.

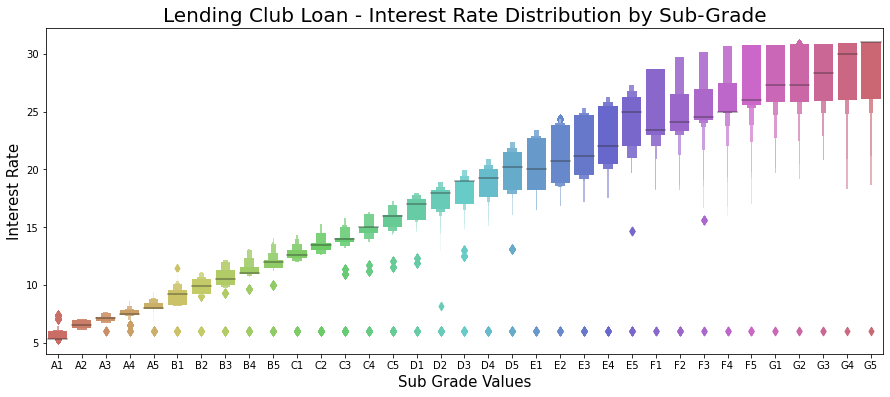


As shown in the plot, total loan amount continuously increases year by year from 2007 to 2018. However, the proportion of loan default amount peaks at 2015 and then starts decreasing. In 2018, the proportion of default loan amount is extremely low compared with that the not default amount.

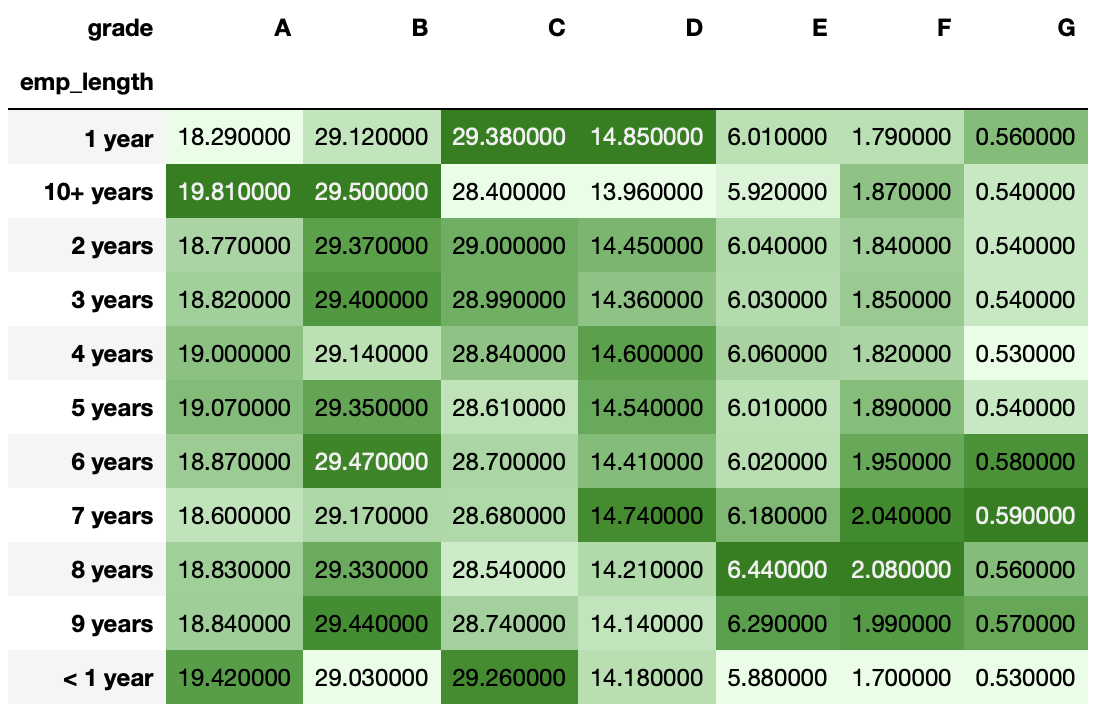
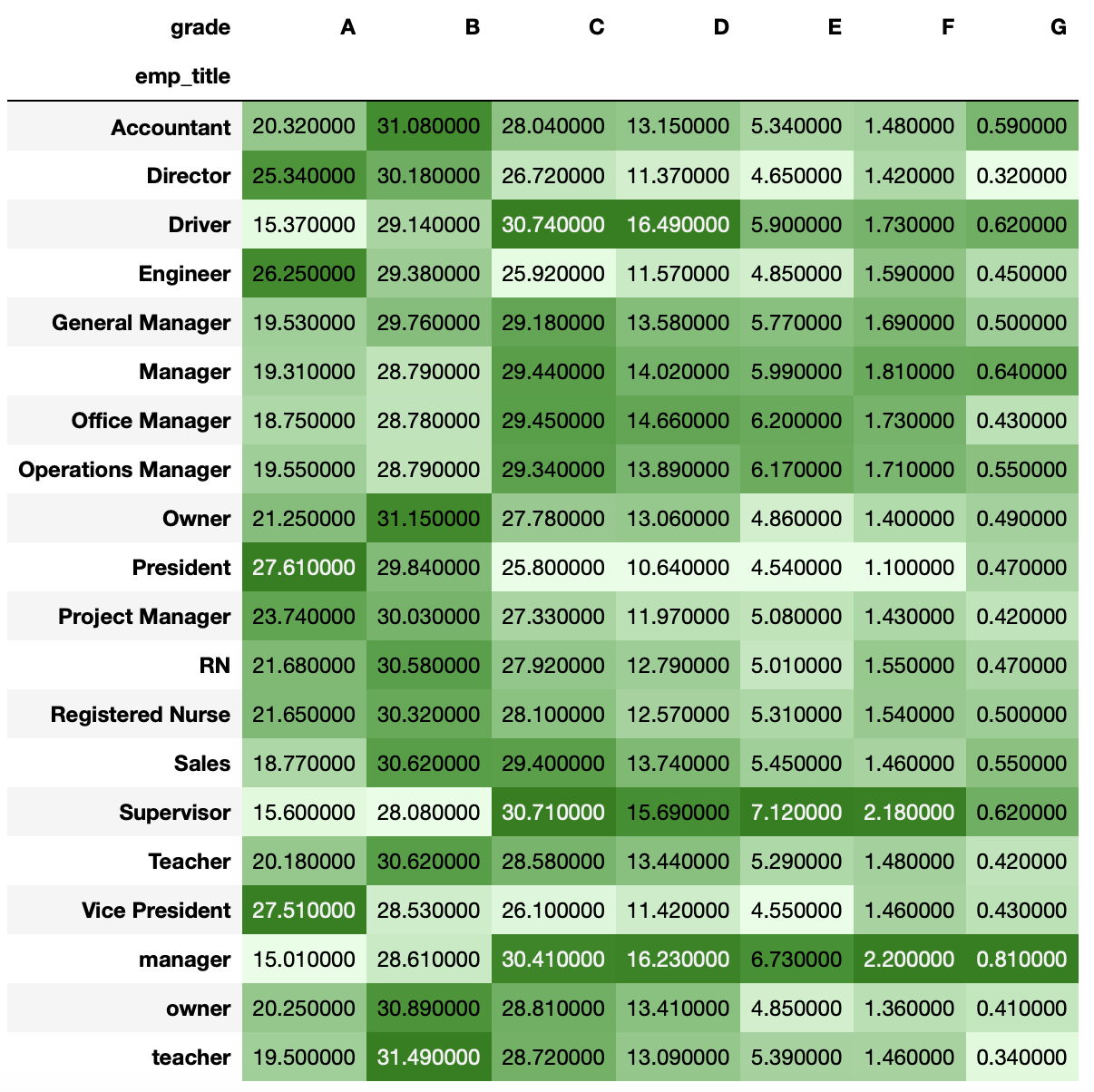
*“Grade” & “Sub\_Grade”*



As shown in the left upper plot, borrowers graded as “C” account for the most default cases while borrowers graded as “A” account for few default cases which corresponds to the fact that the lower the grade, the higher risk of default. As indicated by the right upper plot, the lower the grade, the higher the interest rate which corresponds to the fact that higher interest rate generally implies higher risk of default.



The boxplot distribution of interest rate for different sub grades is similar to the previous one for grades. Finally, let’s explore the relationship between employment title / employment length and grades.



As shown in the plots, Director, Engineer, President and Vice President are Top 4 title that are most likely to be graded as “A” while Drivers, Managers and Supervisors are more likely to be graded as “G”. Also, people working for more than 10 years are more likely to receive high grades while those working for 7 or 8 years are more likely to receive low grades.

**Data Pre-processing**

The data obtained in this project contains a large number of missing data, noise, and irrelevant features. In literature research, when there is a large number of missing data, it is difficult to get an accurate and reliable prediction. On P2P platform, if the borrower cannot provide some sensitive information, there will be a lot of missing data. At the same time, some customers fill in the registration casually and do not provide the real information, which may also cause problems such as outliers, noise data, etc. Therefore, in order to build an efficient risk identification model and obtain more accurate prediction results, it is necessary to first obtain a clean dataset.

The prerequisite of modelling is data cleaning. The first step is usually to deal with a large number of missing values. We take two measures to deal with missing values: first, directly delete variables with too much missing data; second, when the data of a variable is only a small part missing, the median or mean of a variable can be used to impute the missing value.

**Narrowing Down Columns in the Dataset**

To deal with missing data, we first calculate the percentage of missing data for each feature. From the below histogram, we can see there's a large gap between features missing "some" data (<20%) and those missing "lots" of data (>30%). Because it's generally very difficult to accurately impute data with more than 30% missing values, we drop such 44 columns with more than 30% missing value.



Figure 1. Fraction of data missing

Among the remaining variables, the missing value of numerical variables will be replaced by median, mean and 0 according to different situations. For object variables, "999" will be given.

Next, we will check the features mainly according to the following criteria：

(1) features that are recorded only after the loan has already been funded, which will leak information from the future

(2) features that contain redundant information.

(3) features that have a large number of unique values or only one unique value

The dataset also contains features that were not available to the investor when lending decision is to be made. We drop these features which are added later for the purpose of default prediction model is to determine whether a borrower would default before approving the loan. We examine the Lending Club website and Data Dictionary to determine which features would have been available to potential investors. The dataset is reduced to 30 features. Table 1 shows the names and description of the selected features.

|  |  |
| --- | --- |
| Feature | Description |
| addr\_state | The state where the application address belongs |
| annual\_inc | A Numerical variable about the Annual income of the customer. |
| application\_type | Indicates whether the loan is an individual application or a joint application with two co-borrowers |
| dti | The debt-to-income ratio of the borrower (amount of debt divided by annual income). |
| delinq\_2yrs | The number of times the borrower had been 30+ days past due on a payment in the past 2 years |
| earliest\_cr\_line | the date when the first credit line was established |
| emp\_length | A Numerical variable which tells us about for how long the customer has been working at his/her current workplace. |
| emp\_title | The job title supplied by the Borrower when applying for the loan |
| grade | A Categorical variable which tells us how important the customer is to the bank with A being highly important and G being least important |
| home\_ownership | A Categorical variable which tells us about the type of house owned by the customer |
| Id | A unique LC assigned ID for the loan listing |
| initial\_list\_status | The initial listing status of the loan |
| installment | The monthly installments owed by the borrower if the loan is funded. |
| int\_rate | The interest rate of the loan, as a proportion |
| issue\_d | The month which the loan was funded |
| loan\_amnt | A Numerical variable which tells us about the last loan amount taken by the said customer |
| loan\_status | A Numerical variable which tells us about the last loan amount taken by the said customer |
| mort\_acc | Number of mortgage accounts. |
| open\_acc | Number of open credit lines on the borrower’s credit file. |
| pub\_rec | The borrower’s number of derogatory public records |
| pub\_rec\_bankruptcies | Number of public record bankruptcies |
| purpose | The purpose of the loan (takes values “credit\_card”, “debt\_consolidation”, “educational”, “major\_purchase”, “small\_business”, and “all\_other”) |
| revol\_bal | The borrower’s revolving balance (amount unpaid at the end of the credit card billing cycle) |
| revol\_util | The borrower’s revolving line utilization rate (the amount of the credit line used relative to total credit available) |
| sub\_grade | A further partitioning of loan quality - A1, A2, etc |
| term | The term length of the loan. Either 36-months (standard) or 60-months |
| title | The loan title provided by the borrower |
| total\_acc | Total number of historical credit lines on the borrower’s file |
| verification\_status | Status of Lending Club’s verification of the borrower |
| zip\_code | The first 3 numbers of the zip code provided by the borrower in the loan application. |

Table 1. Retained features and description

For the remaining features, we’ll inspect individually. First, we will remove redundant feathers. When considering features with overlapping information, we tend to choose features that record information about clusters of borrowers instead of individuals. Subgrade is implied by the grade, so we will only keep grade. Then, we delete features with too many unique values. There are 61466 different loan titles, which makes it hard to make sense, so we remove the title column. Also, there are 483753 job titles, too many for the feature to be useful, so we drop emp\_titile. The information of zip code is also shown by the state where the application address belongs. Because there are a lot of different values of zip codes, we will just keep the addr\_state column and remove zip codes column.

**Outlier Detection**

When observing 5-number summary statistic or making a boxplot or scatter plot of some features, we can see that there are many outliers. The scatter plot shows annual income versus loan amount. From the plot, we can see a few income outliers with questionable value. From the 5-number summary we know the maximum value of annual income is $ 110,000,000 USD, while median is $ 65,000 USD, and 75% of all data is below $ 93,000 USD. Detail inspection shows that a borrower with highest annual income of $ 110,000,000 USD requested a loan of only $ 30,000 USD for debt consolidation. It is clear that we have outliers due to mistakes.

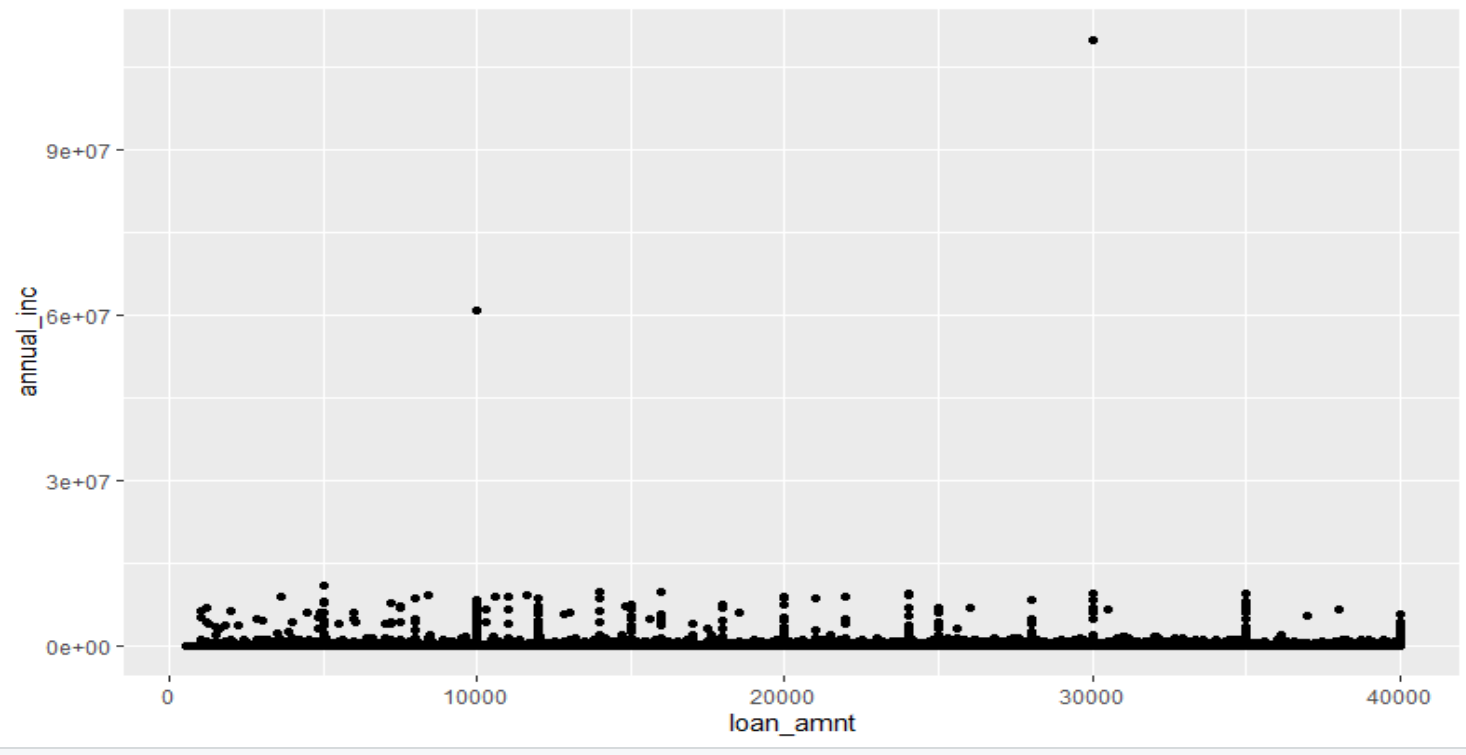


Figure 2. Outliers of annual account

We mainly use two methods to handle these outliers. The first method is to compute upper limit 1.5 \* IRQ, with IRQ = 3rd Quartile – 1st Quartile. Therefore, upper limit for outliers would be $ 70,500 USD, noted it is not suitable to use as the threshold to remove outliers in this case. The second method is using expert judgment. In this case all annual incomes which are larger than $ 15,000,000 USD will be considered as outliers and will be removed from dataset.

After removing outliers, we plot a histogram of annual income. The plot shows annual income has a large range and skewed to the right , so we take a log transform of the variable.

The same producers was conducted for all other variables. The outliers and missing values were treated in accordance to methods described before. The variables that are taken log transformation also include loan amount and revolving balance.

In the final step of data preprocessing, we apply feature standardization to scale all features to the range [0,1].

**Defining the target variable**

In this project, we would like to predict of loan defaults from a given set of observations by selecting explanatory variables that produce a good model performance.

For loans data we would discover more about a delay or a default on required payments. There are 3 potential variables that may indicate a default / delay in payments:

(1) loan\_status: Current status of the loan

(2) delinq\_2yrs: The number of 30+ days past-due incidences of delinquency in the borrower’s credit file for the past 2 years

(3) mths\_since\_last\_delinq: The number of months since the borrower’s last delinquency.

The variable delinq\_2yrs shows only a few unique values and the variable mths\_since\_last\_delinq has surprisingly large values. Both variables only indicate a delinquency in the past, which cannot help with the default definition. Only the variable loan status can be an indicator of the current state a loan is in.

We create a new column in the database called “y” that is “1” for loan status“Charged Off”, “Does not meet the credit policy. Status: Charged Off”, “Default”, “Late (31-120 days)” and 0 otherwise. This leaves us with a dataset with 12.6% negative (labeled 0) and 87.4% positive (labeled 1) examples.

It’s clear that our dataset is highly imbalanced. The target label is described in the table below.

|  |  |  |  |
| --- | --- | --- | --- |
| Loan status | Count | Percentage | Label |
| Charged Off | 261655 | 11.574% | 1 |
| Does not meet the credit policy.  Status: Charged Off | 761 | 0.0337% | 1 |
| Default | 31 | 0.001% | 1 |
| Late (31-120 days) | 21897 | 0.969% | 1 |
| Current | 919695 | 40.682% | 0 |
| Does not meet the credit policy.  Status: Fully Paid | 1988 | 0.0879% | 0 |
| Fully Paid | 1041952 | 46.090% | 0 |
| In Grace Period | 8952 | 0.396% | 0 |
| Late (16-30 days) | 3737 | 0.165% | 0 |

Table 2. Loan status and its statistics

It’s clear that our dataset is highly imbalanced. The target label is described in the table above. To deal with imbalanced dataset, we use oversampling method by tripling the number of default labels in the dataset.

**Model Fitting**

* 1. Data split

The merged dataset is split into two subsets, training set and testing set. Training set contains 75% of data, the rest of them are testing data.

* 1. Feature selection

A tree model was used for prediction therefore we use a Light GMB model for feature selection. Selected feature with feature importance greater than 0.001 and 41 features are selected.

* 1. Create Dummy variables

An encoding method was applied to a string variable, then create dummy variable on them.

* 1. Logistic Regression
     1. Logistic model concept

We use logistic regression as our base model. Even this method named regression, but it is actually a model used for classification. Logistic regression uses a sigmoid function to calculate a probability for an input. If the probability is greater than 0.5 hence, we classify the input as class 1 otherwise class 0.

* + 1. Fitted Logistic model

First, we use all 41 variables as input to fit the logistic model but there are some redundant variables which require to remove.

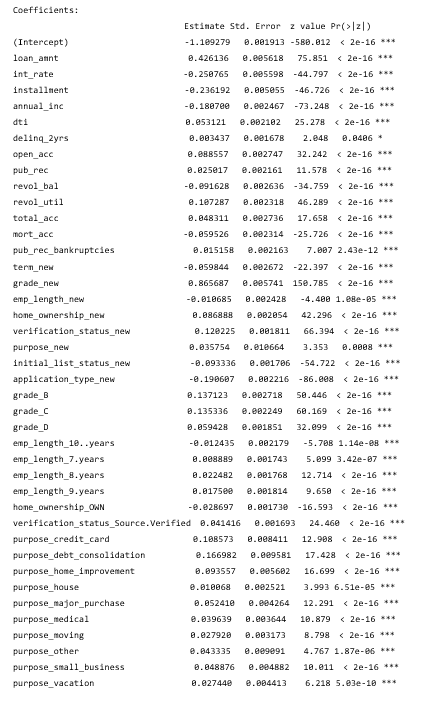
Secondly, we use stepwise selection method to choose those significantly important features. We used a stepwise from both directions to select the best variables which 40 variables get selected. The selected variables of logistic regression is displays below:

y ~ loan\_amnt + int\_rate + installment + annual\_inc +

dti + delinq\_2yrs + open\_acc + pub\_rec + revol\_bal + revol\_util + total\_acc + mort\_acc + pub\_rec\_bankruptcies + term\_new + grade\_new + emp\_length\_new + home\_ownership\_new + verification\_status\_new + purpose\_new + initial\_list\_status\_new + application\_type\_new + grade\_B + grade\_C + grade\_D + emp\_length\_10..years+emp\_length\_7.years+ emp\_length\_8.years+emp\_length\_9.years +home\_ownership\_OWN + verification\_status\_Source.Verified + purpose\_credit\_card + purpose\_debt\_consolidation + purpose\_home\_improvement + purpose\_house + purpose\_major\_purchase + purpose\_medical + purpose\_moving +

purpose\_other + purpose\_small\_business + purpose\_vacation)

The estimated value and its corresponding p-value are showed below:



* + 1. Logistic model result

A confusion matrix is displayed below.

|  |  |  |
| --- | --- | --- |
|  | Predicted 1 | Predicted 0 |
| True 1 | 46712 | 16443 |
| True 0 | 199001 | 250884 |

The precision of logistic model is 19% and recall rate is 74%.

The odd ratio is displays below:

|  |  |
| --- | --- |
| Feature | Odd Ratio |
| grade\_new | 1.862971 |
| loan\_amnt | 1.271664 |
| purpose\_debt\_consolidation | 1.083684 |
| verification\_status\_new | 1.080428 |
| revol\_util | 1.069587 |
| home\_ownership\_new | 1.062733 |
| open\_acc | 1.057468 |
| purpose\_home\_improvement | 1.050603 |
| purpose\_credit\_card | 1.04587 |
| total\_acc | 1.035035 |
| purpose\_small\_business | 1.034853 |
| dti | 1.034362 |
| purpose\_major\_purchase | 1.029072 |
| purpose\_other | 1.024083 |
| purpose\_medical | 1.02251 |
| pub\_rec | 1.020829 |
| verification\_status\_Source\_Verified | 1.019361 |
| purpose\_vacation | 1.017199 |
| purpose\_moving | 1.015083 |
| emp\_length\_8year | 1.014454 |
| emp\_length\_9year | 1.010847 |
| pub\_rec\_bankruptcies | 1.008445 |
| purpose\_new | 1.008038 |
| emp\_length\_7year | 1.00548 |
| delinq\_2yrs | 1.004249 |
| purpose\_house | 1.003539 |
| emp\_length\_new | 0.995883 |
| emp\_length\_10+year | 0.992835 |
| grade\_B | 0.990744 |
| grade\_C | 0.983111 |
| home\_ownership\_OWN | 0.981303 |
| grade\_D | 0.978051 |
| term\_new | 0.973798 |
| mort\_acc | 0.967029 |
| revol\_bal | 0.941889 |
| initial\_list\_status\_new | 0.937137 |
| application\_type\_new | 0.900253 |
| annual\_inc | 0.896425 |
| installment | 0.876596 |
| int\_rate | 0.797725 |

The top two largest odd ratio are grade and loan amount which mean having large grade and loan amount the probability of default will increase. On the other hand, the smallest odd ration is interest rate which mean the lower interest rate corresponding to lower default probability.

* 1. Decision Tree

We will start with a simple tree model before we fit a boosting tree model. Fit a decision tree model with maximum depth 25. The loss of one default lending is much greater than one successful lending therefore we want to control the recall rate around 70%. The summary and confusion matrix display as below:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Model Score | Precision | Recall | Accuracy |
| Dtree | 0.877 | 0.171 | 0.68 | 0.554 |

|  |  |  |
| --- | --- | --- |
|  | Predicted 1 | Predicted 0 |
| True 1 | 43151 | 20004 |
| True 0 | 208848 | 241037 |

The precision is 17% and recall rate is 68%. In business perspective, the lending club would not reject the clients based on the model result directly. But these people require an addition audit to determine the final decision.

* 1. LightGBM
     1. LightGBM concept

The main reason for using LightGBM is that has a reasonable performance and relatively low computational cost. Since the dataset contains huge amount of data, the advantage of LightGBM is suitable for this case.

LightGBM is developed by Ke et al. in 2017 which is improved over Gradient boosting decision tree (GBDT) in terms of efficiency. The most time-consuming part for GBDT is that it estimates the information gain of all possible splits for each feature.

To address this issue, Ke et al. (2017) propose two novel techniques: Gradient-based One-Side Sampling (GOSS) and Exclusive Feature Bundling (EFB). GOSS retains instances with large gradients and randomly drops instances with small gradients in order to retain accuracy while enhancing efficiency. As proved by Ke et al., this treatment dominants uniformly random sampling since it leads to more accurate estimates for sampling split gain. EFB is used to deal with sparse feature space, for example one-hot features, by bundling these features safely.

* + 1. Parameter test

Our model construction begins with grid search for optimal parameters. Our searching procedures and the corresponding results as shown as follows:

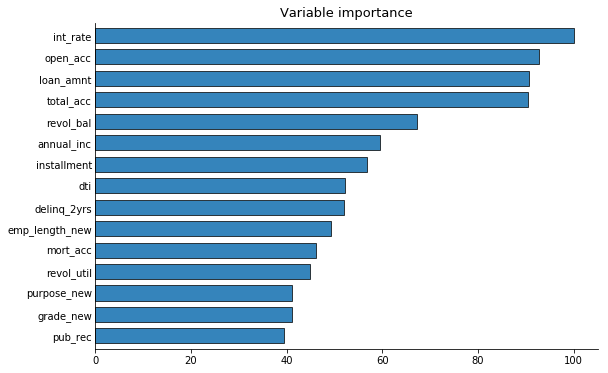
|  |  |
| --- | --- |
| Searching Steps | Optimal Parameters (Default value) |
| STEP 1 | learning\_rate = 0.2 (0.1) |
| STEP 2 | max\_depth = 15(3); |
| STEP 3 | colsample\_bytree = 0.1(1), subsample = 0.1 (1) |
| STEP4 | Gamma = 0(0) |

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Model Score | Precision | Recall | Accuracy |
| LGMB | 0.877 | 0.212 | 0.748 | 0.627 |

A confusion matrix is displayed below.

|  |  |  |
| --- | --- | --- |
|  | Predicted 1 | Predicted 0 |
| True 1 | 44978 | 18177 |
| True 0 | 160532 | 289353 |

The TOP 15 important features for LGB are plotted below:



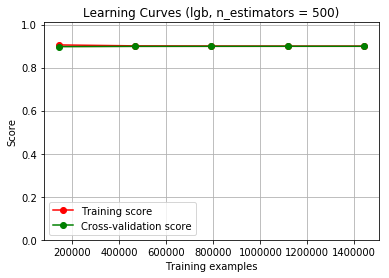
* 1. Interpreting of the result

Because of the particularity of the banking industry, recall rate is the point we pay most attention to. The profit of lending club is mostly come from the interest, if one lending was defaulted there might need multiple successful lending to make up the loss.

For classification model, confusion matrix is a good indicator for the result. The model predicted 44978 true response (true response indicates as 1) there is 63155 true positive response in test set therefore the recall rate is 0.712. The model also correctly predict majority of the true positive. The overall accuracy of the model is 0.652.

From feature importance, the top 15 most important feature were plot. The first three features are interest rate, loan amount and open account which mean there three features were closely related to the probability of default.

The cross-validation and train score quickly converge due to the relatively large learning rate.



Compare to the other two models, LightGBM has the best performance in recall rate and overall accuracy.

Reference

Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., ... & Liu, T. Y. (2017). LightGBM: A highly efficient gradient boosting decision tree. *In Advances in Neural Information Processing Systems* (pp. 3149-3157). Retrieved from: http://papers.nips.cc/paper/6907-lightgbm-a-highly-efficient-gradient-boosting-decision-tree.pdf.