Machine Learning I

Instructor: Dr. Richard Xie

Ye In Jeon

**Final Project Report**

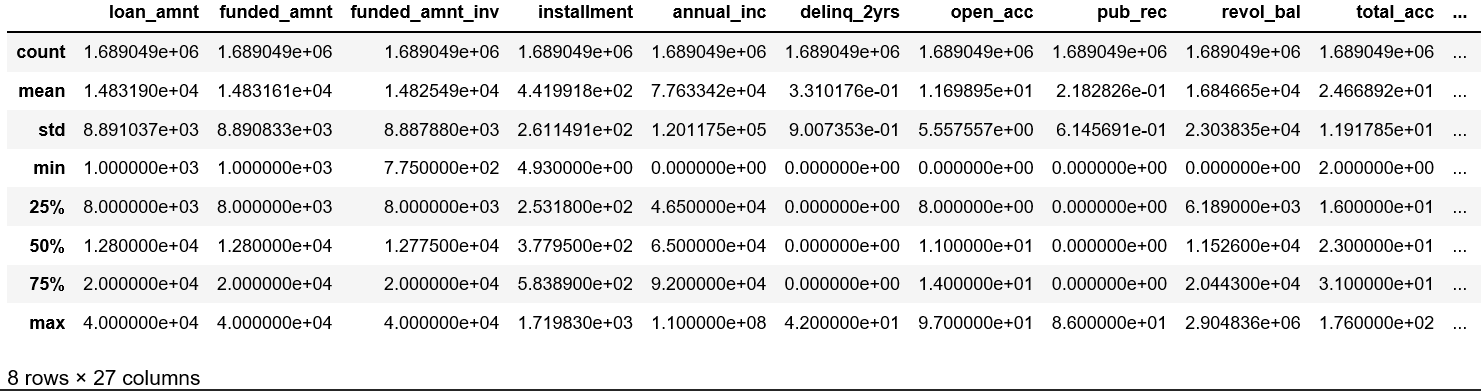
Loan Prediction

Introduction

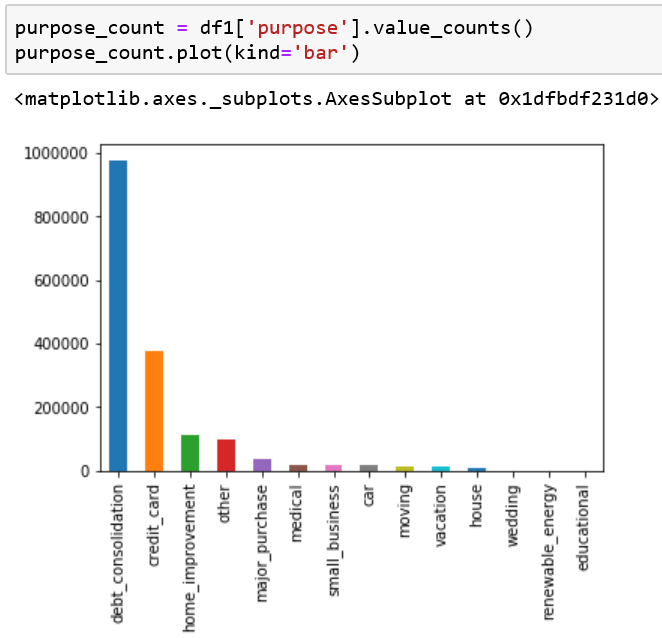
The loan status, which states if the loan is defaulted or not, is the most considerable factor to the companies related to loan. This report discusses loan status prediction. First, it states the importance of the problem statement and the description of datasets. In addition, it provides the exploratory data analysis. Furthermore, it states the different methods of model building. Additionally, it addresses how the models are compared and evaluated. Moreover, it shows the results of the analysis with the selected datasets. Base on the result, the comparisons are made and find the best model for the analysis.

Importance of the Problem statement & Dataset

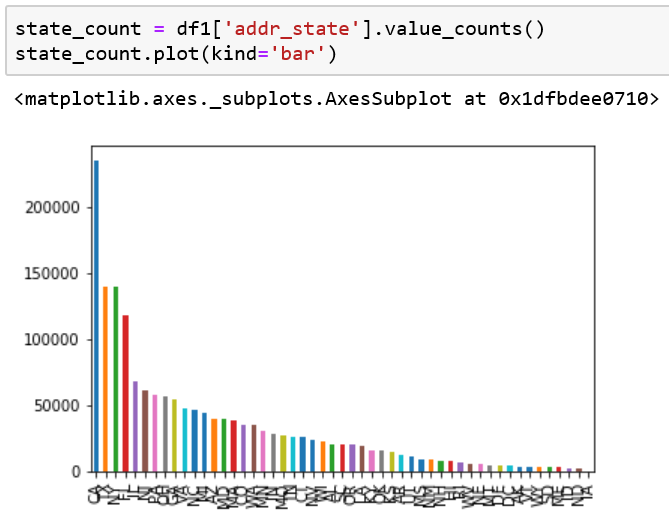
In this paper, it will mainly talk about ‘How to predict a current loan will default in the end?’ For this analysis, the datasets are downloaded from ‘Lendingclub’ homepage (<https://www.lendingclub.com/info/download-data.action>). There are 11 datasets which contain loan data from 2012 to 2017. For this analysis, all datasets has to be merged. After merging process, all the columns with NaN values are dropped. Moreover, the rows with NaN values are eliminated. Following is the part of summary statistics of the dataset as below:



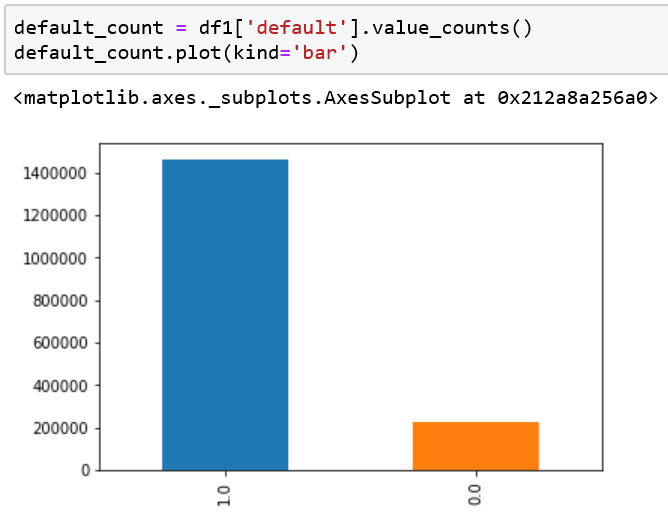
Exploratory Data Analysis

To understand the dataset better and find the trend in the dataset, the data visualization is proceeded. First, as it can be found in the bar plot below, the loan purpose of majority people are debt consolidation and credit card.

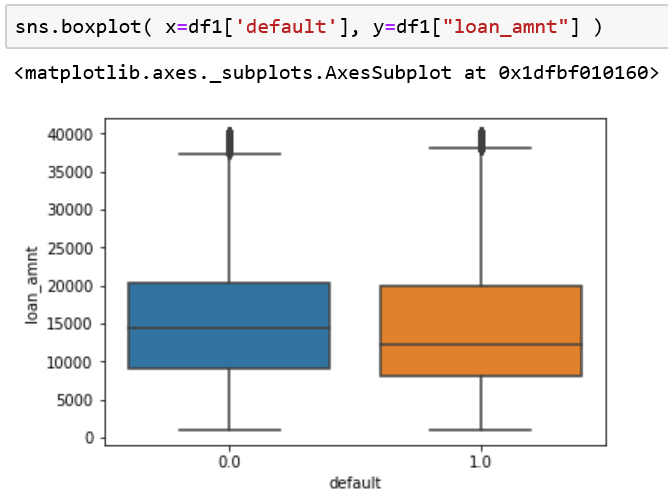
In addition, it can also be found that top three states with highest number of loan cases are California, Texas and New York.



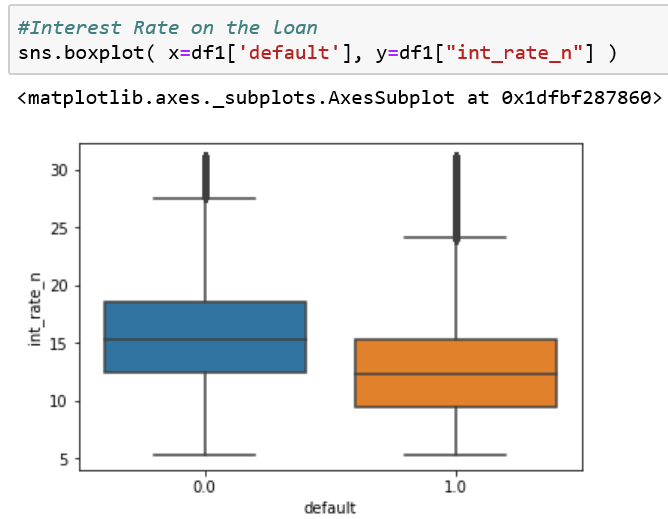
Moreover, it can be seen that the data is biased. There are more loan cases which are not defaulted compared to the number of that which are defaulted.



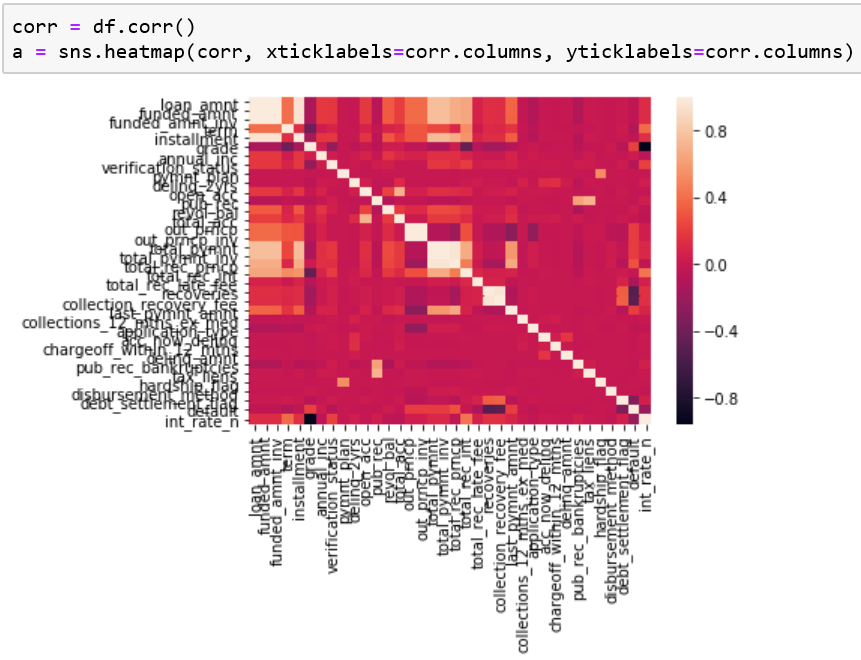
With the boxplot below, the loan amount’ median value of the loan which are defaulted are smaller than that of the loan which are not defaulted.



Moreover, it can be found that the median of the interest rate tend to be higher with the loan cases which are not defaulted.



When finding the feature relevance, the correlation plot is useful. It can be seen that some variables are highly correlated in negatively and positively while most of the variables are not relevant to each other.



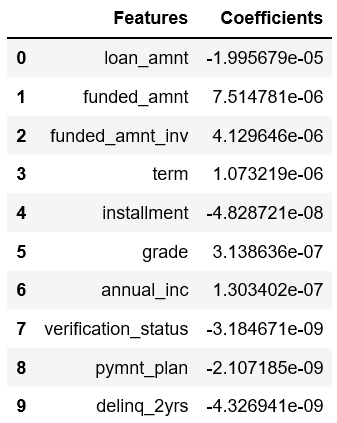
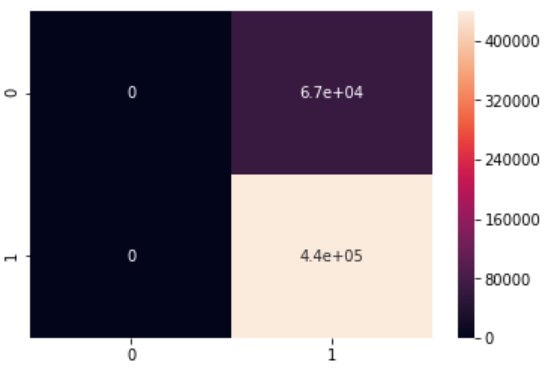
Model building method

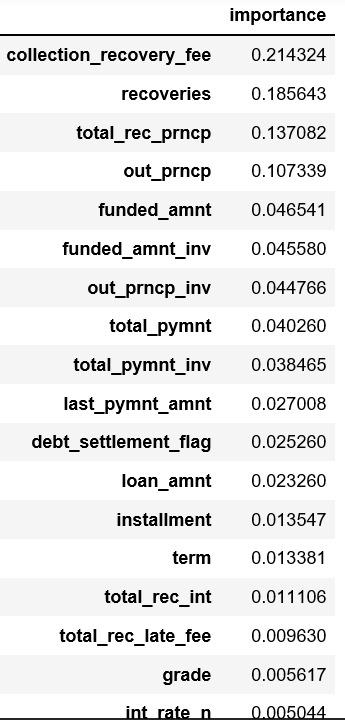
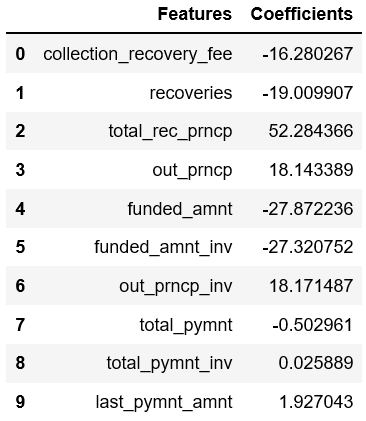
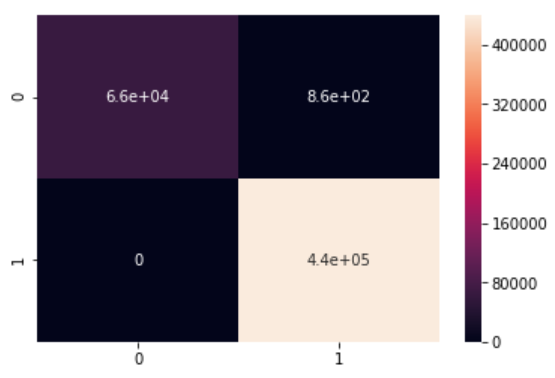
Before selecting the model building methods, the dependent variable should be chosen. With this dataset, the target variable is ‘loan\_status’. The logistic regression models are built in two ways by different feature selection methods. First, the 10 features are selected by principal component analysis, which is called as PCA in this paper. Moreover, the 10 features are selected by the highest number of random forest feature importance. With the selected 10 variables, the logistic regression is proceeded. Next, Naïve Bayesian model is built. Since ‘loan\_status’ is binary variable, Naïve Bayesian model can be appropriate model for this analysis. Last, Lasso method is used for this analysis. Before starting the model building, the data is split into test and train and standardized for better analysis as below. However, standardized data is not used for LASSO model.

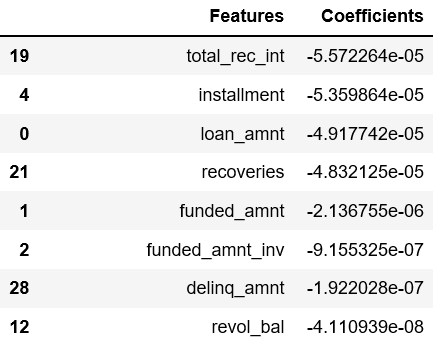
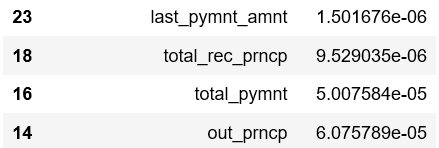
Model Evaluation

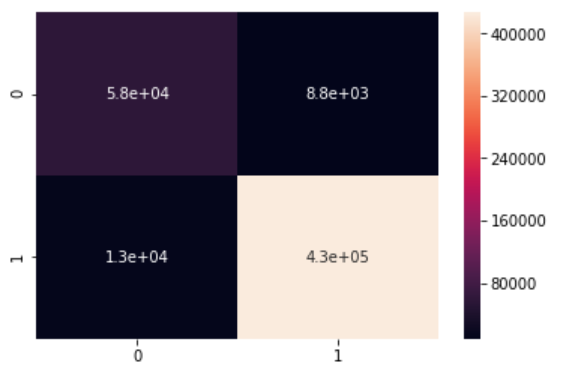
In this report, three methods are used for model evaluation: Confusion matrix, Accuracy, and R squared. First confusion matrix is a plot of predictions. It shows the number of loan cases which the prediction values are 0 or 1 when the actual values are 0 or 1. Therefore, it can be interpreted as following: The more cases, which the prediction value and the actual value are equal, the better the model is. The accuracy value and R squared value are between 0 and 1. The accuracy is the ratio of the number of cases, which are predicted correctly, and the number of all cases. Furthermore, R squared value shows the goodness of fit. With three different method, the best model can be selected among conducted models for this project.

Models Result

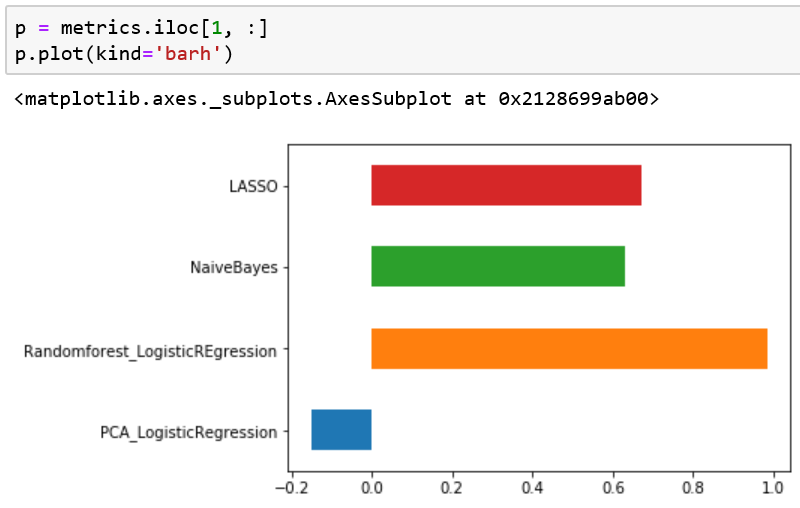
 The models are created with the methods mentioned above. To deliver the result of the models in a better way, the features’ coefficients and confusion matrixes are shown. First of all, the 10 features chosen with the PCA method are ‘loan\_amnt’, ‘funded\_amnt’, ‘funded\_amnt\_inv’, ‘term’, ‘installment’, ‘grade’, ‘annual\_inc’, ‘verification\_status’, ‘pymnt\_plan’ and ‘delinq\_2yers.’ For more information, ‘term’ is the number of payments on the loan. Furthermore, ‘installment’ indicates the monthly payment owed by the borrower if the loan originates. In addition, ‘annual\_inc’ is the self-reported annual income provided by the borrower during registration. Moreover, ‘grade’ is LC assigned loan grade. Additionally, ‘delinq\_2yers’ shows the number of 30+ days past-due incidences of delinquency in the borrower's credit file for the past 2 years. With these 10 variables, logistic regression is used to predict ‘loan\_status’ which is t arget variable. As it can be seen in the matrix on the right, it can be seen there are 0 variable which predict value is 0 when actual value is also 0. In addition, there are high number of loan cases when predicted value is 1 and actual value is 0. Therefore, it is hard to say this is a good model.

 With the Random forest method, features are sorted by the importance of the variables. Likewise in PCA method, top 10 features are selected for another logistic regression. As a result, ‘collection\_recovery\_fee’, ‘recoveries’, ‘out\_prncp’, ‘funded\_amnt’, ‘out\_prncp\_inv’, ‘total\_pymnt’, ‘total\_pymnt\_inv’, and ‘last\_pymnt\_amnt.’ For better understanding, ‘collection\_recovery\_fee’ indicates the post charge off collection fee. Moreover, ‘recoveries’ shows the post charge off gross recovery. Furthermore, ‘out\_prncp\_inv’ is the remaining outstanding principal for total amount funded. As it can be found in the confusion matrix plot, there are majority number of loan cases which predicted value and actual value are equal.

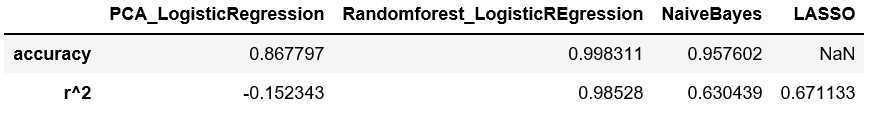
With Lasso method, the number of features is not decided before running the regression. The features of this method are ‘total\_rec\_int’, ‘installment’, ‘loan\_amnt’, ‘recoveries’, ‘funded\_amnt’, ‘funded\_amnt\_inv’, ‘deliq\_amnt’, ‘revol\_bal’, ‘last\_pymnt\_amnt’, ‘total\_rec\_prncp’, ‘total\_amount’, and ‘out\_prncp.’ For the explanation of the features, ‘total\_rec\_int’ is interest received to date. Moreover, ‘delinq\_amnt’ is the past-due amount owed for the accounts on which the borrower is now delinquent. ‘revol\_bal’ is the total credit revolving balance. In addition, ‘last\_pymnt\_amnt’ is the last month payment was received. ‘total\_rec\_prncp’ is the principal received to date. Furthermore, ‘total\_pymnt’ is the payments received to date for total amount funded. Unlike other methods, instead of the confusion matrix plot, Lasso are compared with other models by R squared value.

With Naïve Bayesian regression, it is not available to show the feature names and coefficients of them. Instead, confusion matrix plot is drawn. Compared to the plot of the logistic  regression based on the random forest feature importance, it can be found that the number of loan is higher which the predictions are failed.

Model Comparison

 To find the most appropriate model, 4 models have to be compared by confusion matrix plots, accuracy values, and R squared values. The R squared value of the logistic regression with random forest feature importance is highest. It can be also shown in the bar plot below.

In addition, in case of accuracy value, it point out that the best model is the logistic regression with random forest feature importance.



Moreover, with the confusion matrix plot, the logistic regression with random forest feature importance is the best model because its plot shows the highest number of loan cases which are correctly predicted.

Conclusion

This report is to find out if the loan case will default in the end. With the analysis, it is found that the logistic regression base on Random Forest feature importance is the best model among other 4 models. According to the analysis, ‘collection\_recover\_fee’, ‘recoveries’, ‘total\_rec\_prncp’, ‘out\_prncp’, ‘funded\_amnt’, ‘funded\_amnt\_inv’, ‘out\_prncp\_inv’, ‘total\_pymnt’, ‘total\_pymnt\_inv’, and ‘last\_pymnt\_amnt’ are found to affect the loan default status.

For the future studies, several suggestions can be made for better analysis. First, more loan cases, which are defaulted, can be collected. In addition, other features can be added because there may be factors which severely affect the loan default status but not included in this analysis. Moreover, analysis base on the state can be made. This is because the laws and the circumstance can different by each state. Therefore, the importance of the features may different by state.