**Feature Engineering**

* Feature Selection:
* Drop 58 features which has missing rate over 30%, since it's difficult to impute data accurately with more than 30% missing values.
* Drop 60 features which may not available to investors before make investment decisions.
* Drop 'installment', 'total\_acc', 'pub\_rec\_bankruptcies', 'grade', 'title', since they are highly correlated with other features or their information has been covered by other features.
* Grab the year information from 'issue\_d' and 'earliest\_cr\_line', and drop these two features. Name the new features as 'issue\_y' and 'earliest\_cr\_line\_y'. For 'earliest\_cr\_line\_y', convert the year into 'age' of the earliest reported credit line, i.e., the number of years until 2023.
* Drop 'emp\_title', since it contains too many unique values in 'emp\_title' value, and it’s difficult to be converted to a dummy variable feature.
* Drop 'emp\_length', since there's no significant change in loan status as the year of employment changed.

*List of selected features:* 'loan\_amnt', 'term', 'int\_rate', 'sub\_grade', 'home\_ownership', 'annual\_inc', 'verification\_status', 'purpose', 'addr\_state', 'dti', 'open\_acc', 'pub\_rec', 'revol\_bal', 'revol\_util', 'initial\_list\_status', 'application\_type', 'mort\_acc', 'charged\_off', 'issue\_y', 'earliest\_cr\_line\_y'

* Dealing with Missing Values:

There are two features have missing value: 'dti' and 'revol\_util'. Since the missing rate is quite low (0.08%), I simply dropped the rows with missing value.

* Remove Outliers: According to EDA, I choose to remove the outliers for four of the continuous features: 'annual\_inc','dti','revol\_bal', and 'revol\_util', which has relatively large range of data. In other works, I only kept the data that are within +3 to -3 standard deviations of these four columns.
* Create Dummies: Create dummy variables for 7 categorical features ('sub\_grade', 'home\_ownership', 'verification\_status', 'purpose', 'addr\_state', 'initial\_list\_status', 'application\_type').

Final train set:

*List of selected features before get\_dummies:*

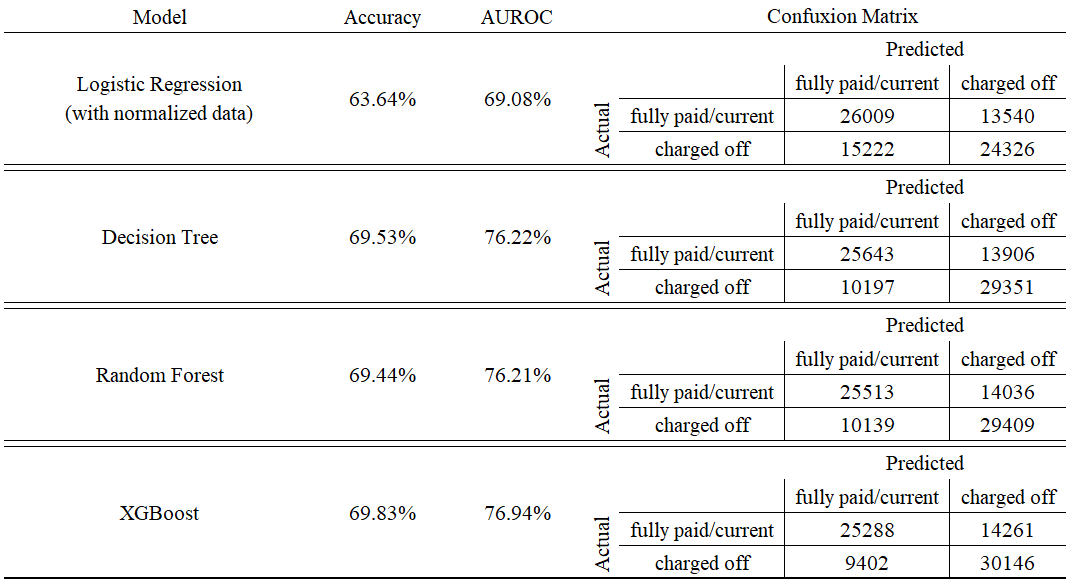
'loan\_amnt', 'term', 'int\_rate', 'sub\_grade', 'home\_ownership', 'annual\_inc', 'verification\_status', 'purpose', 'addr\_state', 'dti', 'open\_acc', 'pub\_rec', 'revol\_bal', 'revol\_util', 'initial\_list\_status', 'application\_type', 'mort\_acc', 'charged\_off', 'issue\_y', 'earliest\_cr\_line\_y'

*Shape:*

* Before get\_dummies: (1787930, 20)
* After get\_dummies: (1787930, 118)

**Modeling**

* Train/Validation Set Split: I stratified split the dataset, and used 70% of the data as the training set, and the other 30% as the validation set.
* Downsampling: I downsampled the data to make the number of fully paid/current loans equal to the number of charged-off loans. After downsampling, both of two classes have 131,828 observations.
* Model Selection:

I tried to apply 4 predictive models, and the result is shown as below:

The XGBoost model achieved 69.83% Accuracy Score and 76.94% AUROC Score, which is considered to be the best model. Then I tuned the XGBoost model using randomized search cross-validation. The best parameters: {'learning\_rate': 0.1000695984086799, 'max\_depth': 6, 'n\_estimators': 188, 'subsample': 0.6355582882297452}.

* Result: The tuned XGBoost Model slightly improved the original XGBoost model, and got 70.08% Accuracy Score and 77.2% AUROC Score. The most important features including: 'issue\_y' and 'int\_rate'.

**Future Work**

* The XGBoost model seems to have the problem of overfitting, we can complete further analysis to check.
* Since issue year is a crucial attribute in the final model, we can try to find the macroeconomics index of different years to analyze if the variation of economic situation year by year is the reason behind it.
* It's not quite reasonable to regard current loans as fully paid loans since the current loan would also be defaulted in the future. We can complete the analysis only based on fully paid loans and charged off loans to see if there’s any improvement.
* The relatively weak performance of the model is due to the fact that the dataset only included accepted loans and the model only used origination time information. In practice, we can choose a high threshold close to 1 to elevate the precision by sacrificing the recall and use this model as a preventive tool for the investors to avoid risky loans.