**Predicting Loan Default with Tree-Based Models**

This report will apply tree-based supervised learning methods to the default data. The R codes and graphs are attached in the RMarkdown files.

**1 Data Preprocessing**

Due to high dimensionality of data, we focus on only 10 out of 626 predictors from the original data set for the section. Other dimensions of data such as the response `bad\_good` and the number of observations 285,285 remains the same. The features selected are `GENDER`, `LOAN\_FLAG`, `OS\_PRCP\_SUM\_THREE`, `OS\_PRCP\_SUM\_SIX`, `G\_OS\_PRCP\_SUM`, `L6\_CUST\_DEBT\_AVG\_AMT`, `CUST\_DEBT\_AMT`, `L3\_CUST\_DEBT\_AVG\_AMT`, `DEP\_SA\_OPEN\_TENURE\_DAYS`, and `DEP\_SA\_AVG\_TENURE\_DAYS`.

Before training models, we preprocess the data of the selected predictors and the response. The data type of the categorical variable `bad\_good` is converted from integer to factor. We conduct a brief exploratory data analysis with a summary of the variables and pairwise plots between variables. We split the 60% of the data into the train set and 40% into the test set by stratified sampling according to the stratification of the response. The seed is set at 123. We normalize and standardize all 8 numeric features. We create two dummy variables for nominal features `GENDER` and a dummy variable for `LOAN\_FLAG` by one-hot encoding. On top of the feature engineering, we prepare the ordinal features of the data for XGBoost by label encoding.

**2 Models**

We fit three tree-based methods to the training data. The methods consist of random forest (RF), basic gradient boosting machines (GBM), and extreme gradient boosting (XGBoost).

**2.1 Random Forest**

In terms of RF, we train a model with default hyperparameters. `mtry` accounts for approximately one third of 13. We tune the model hyperparameters by full Cartesian grid search. The best hyperparameters are using a random subset of 6 variables for split, at least 10 observations in a leaf node, sampling scheme of sampling without replacement, 80% of observations to sample. After tuning, the cross-validated (CV) error rate reduces by nearly a half to .

From the variance importance graph, the most important variable deciding the target variable is the gross outstanding balance, followed by the loan flag. The outstanding balance within 3 months and within 6 months both have a lower importance score but much higher than the remaining features. The partial dependence plots picture the predicted probability of default versus the four most important features proposed by the model. High probability of default is associated with near zero gross outstanding sum, no loan flag, and near zero outstanding sum within 3 months. A special effect is that outstanding sum within 3 months exceeding 120 is linked with moderately high probability of default.

**2.2 Basic GBM**

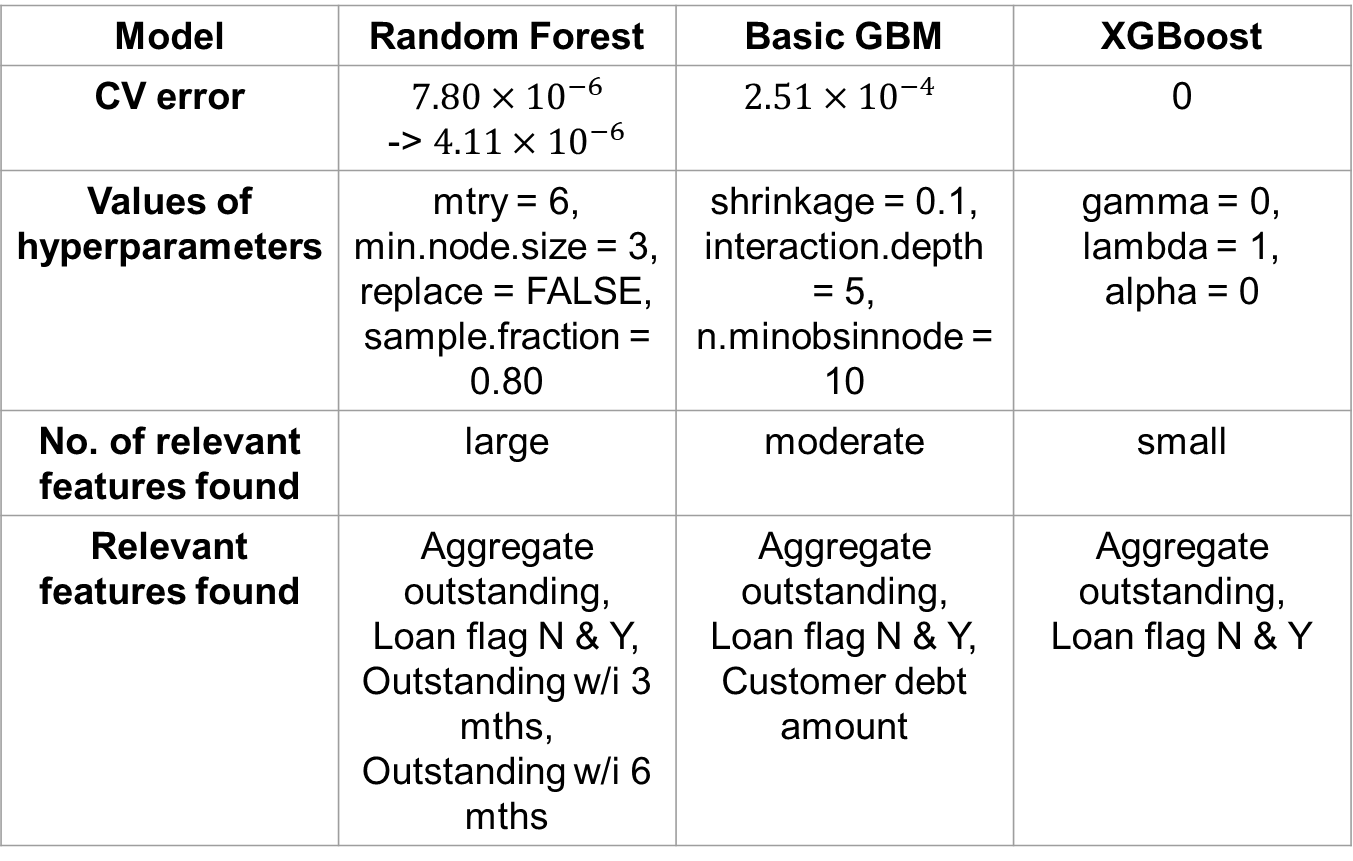
Concerning basic GBM, we train a default model untuned because the large data size leads to R session abortion. Like in RF, the important variables in GBM contain the gross outstanding sum and the loan flag. Observations with default labels tend to have extremely low gross outstanding sum and have loan label. In addition, as customer debt amount raises, the probability of default increases. However, the marginal effect of customer debt amount is not so significant as the three major variables.

**2.3 XGBoost**

Regarding XGBoost, we train a model with default hyperparameters. The CV error is 0, so tuning is not needed. The importance graph displays that only gross outstanding balance and loan flag is valued in the model. The gross outstanding balance is twice as important in determining default as the variable of no loan flag.

**3 Comparison**

The following chart summaries the comparison of the results of the three models.

****

**4 Final Model**

Having compared CV errors, we conclude that XGBoost is the most suitable method for the data among the three tree-based methods. We use the same hyperparameters to refit the model to the whole training data and make predictions on the test data. From the confusion matrix, only one client who is not default on the loan is wrongly predicted as default. There is only one false positive and no false negative compared with over a thousand true positives and a hundred thousand true negatives. The test error is . The F1 score is 99.97%. Overall, the model correctly classifies the vast majority of new data.