MSDS 422
Assignment 2
Juhwi Kim

Introduction

This week's exercise builds on the Exploratory Data Analysis we conducted on the Home Equity Line of Credit (HMEQ) data set and using tree based models to predict home equity loan being defaulted and subsequent lost loan amount.

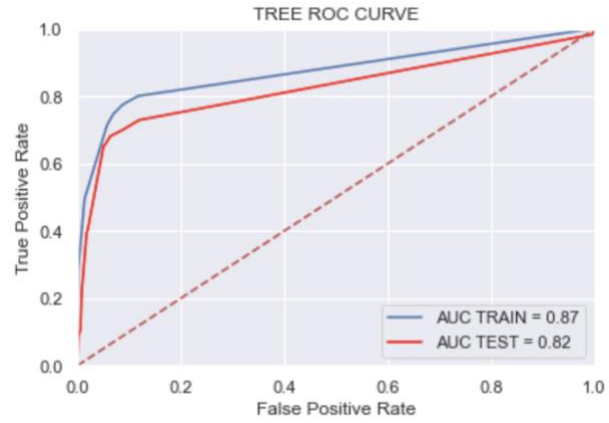
Create a Training and Test Data Set:

Using the SciKit Learn 'train_test_split' with an 80%/20% split, I then produced a train and test data sets. In the training data set, I have a total 4768 rows of records and 1192 rows of records in the test data set. Following the splitting the dataset specific to defaulted loans and reviewing the Decision Tree, it is noted through the chart on the right hand side that the maximum lost loan amount is $73946 in the training data set and $78987 in the test data set. These values far over exceeds the mean value of $13421 in training data set and $13387 in the test data set. Therefore, I decided to cap the maximum loss loan amount to be $25000 for Decision Tree models.

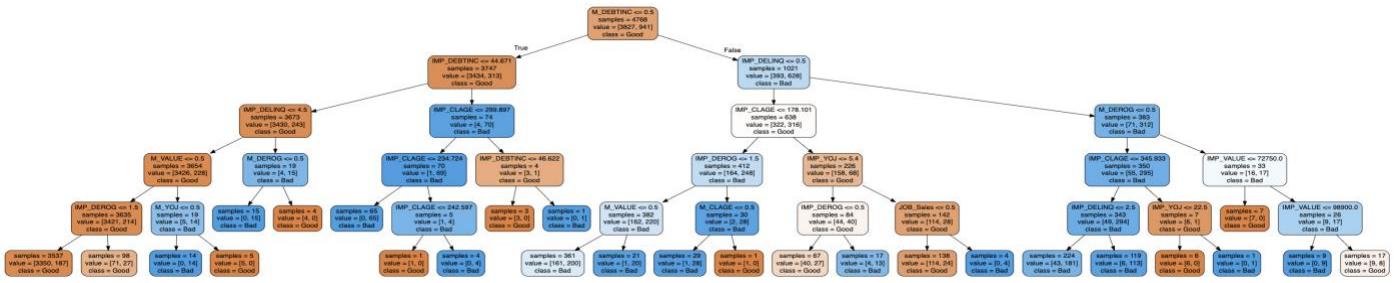
	TARGET_BAD_FLAG	TARGET_LOSS_AMT
count	941.0	941.000000
mean	1.0	12163.315622
std	0.0	7584.525527
min	1.0	224.000000
25%	1.0	5817.000000
50%	1.0	10959.000000
75%	1.0	17635.000000
max	1.0	25000.000000

	TARGET_BAD_FLAG	TARGET_LOSS_AMT
count	248.0	248.000000
mean	1.0	11946.959677
std	0.0	7549.372558
min	1.0	320.000000
25%	1.0	5214.500000
50%	1.0	11336.500000
75%	1.0	16734.000000
max	1.0	25000.000000

Decision Trees to predict the probability of defaults



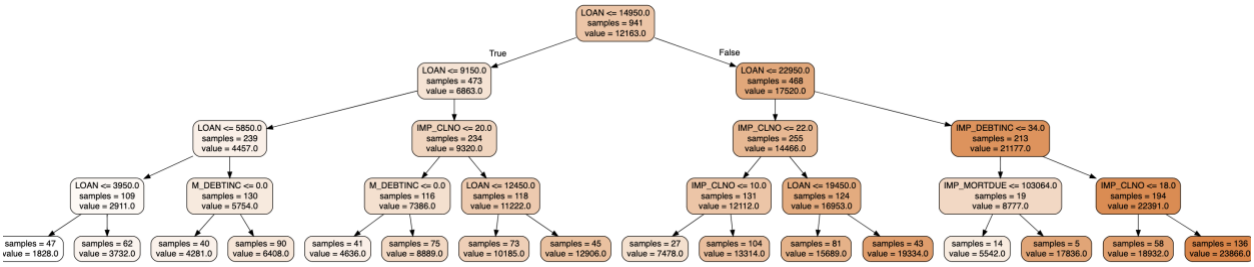
The Decision Tree model was able to successfully predict whether loan will default in 89.7% of train data set and 88.7% of the test data set. While there is a slight discrepancy between train and test model prediction where it could perform slightly better against training data over test data, the difference is small enough to rule out the overfitting of the model.



MSDS 422
Assignment 2
Juhwi Kim

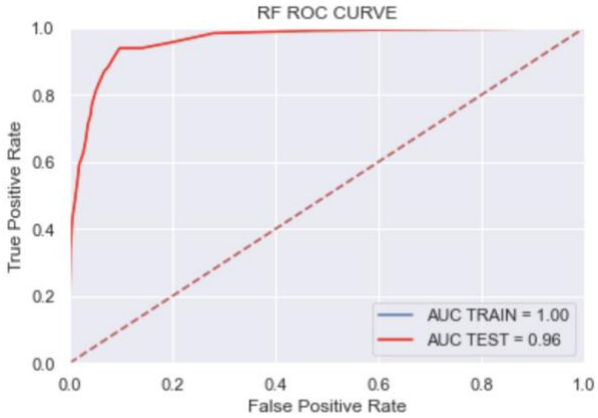
The variables included in the decision tree that predict loan default is following: HMEQ Credit Loan Amount, having Office and Sales related jobs, Current Outstanding Mortgage Balance, Value of the house, Year on the Job, Derogatory Marks on Credit Record, Delinquencies on your current credit report, Credit Line Age, and Debt to Income Ratio.

Decision Trees to predict the loss amount assuming that the loan defaults



The Root Mean Square Error in predicting amount of the loss amount using Decision Tree in train data set is 3216.22 while RMSE for the test data set is 3942.997. The variables that were included in the decision tree that predict loss amount is the total HMEG Credit Loan Amount, Current Outstanding Mortgage Balance, Number of credit lines you have, and Debt to Income Ratio.

Random Forests to predict the probability of default



In comparison to the simple Decision Tree, Random Forests model produced noticeably higher accuracy. The model was able to predict the probability of flagging the bad loan at 99.9% of training data and 91.2% of the test data.

The ROC curve for Random Forests indicate that the model was successfully able to fit the entirely of the train data with perfect accuracy while the area under the curve for the test data to be 96% of the test data, which is quite higher than the previous decision tree.

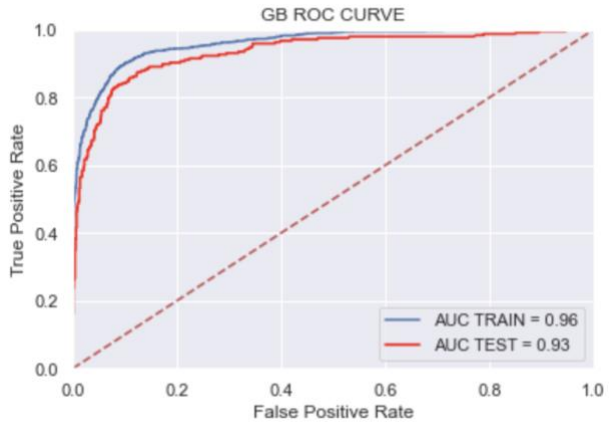
The variables included in the Random Forest that predict loan default is following from the most significant to lesser: Debt-to-Income Ratio, Credit Line of Age, Delinquencies on current credit report, HMEQ Loan Amount, Home Value, Number of Credit Lines one has, Total Mortgage Due, Year On Jobs Derogatory Marks on Credit Record, and Number of Credit Inquires.

Random Forests to predict the loss amount assuming that the loan defaults

The Root Mean Square Error in predicting amount of the loss amount using Random Forests in training data is 823.488, while RMSE in test data is 2178.544. The variables that were used in predicting the loss amount in Random Forests models are following: The total HMEQ Loan Amount, total line of credits borrower has, and the debt-to-income ratio.

Gradient Boosting model to predict the probability of default

MSDS 422
Assignment 2
Juhwi Kim



The Gradient Boosting model successfully predicted the probability of HMEQ loan being defaulted at the accuracy of 92.3% of the train data set and 90.1% of the test data set.

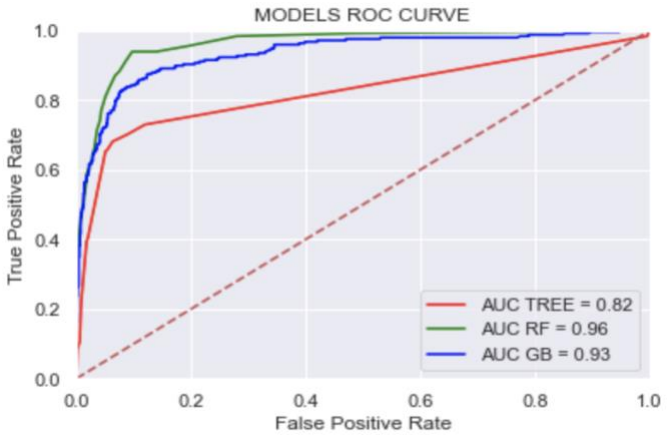
The ROC Curve for Gradient Boosting suggests that the accuracy of the Gradient Boosting model to be highly accurate at almost 96% of the train data set and 93% of the test data set.

The variables that were used in the model are Debt-to-Income Ratio, Delinquencies on current credit report, Credit Line Age, and Derogatory Marks on Credit Record.

Gradient Boosting model to predict the loss amount assuming that the loan defaults

The Root Mean Square Error in predicting amount of the loss amount using Gradient Boosting in traiing data is 1046.818 and RMSE for the test data set is 1833.696. The variables that were used in predicting the loss amount in Gradient Boosting models are following: The total HMEQ Loan Amount, total line of credits borrower has, and the debt-to-income ratio.

ROC Curves Comparison Between All Three Models & Conclusion



When we compare all three models against each other using the ROC curve, we noticed that the Random Forest was most successful in predicting the status of the HMEQ loan being defaulted at the accuracy of the model being 0.96 while Gradient Boosting following right behind the Random Forest and the simple decision tree having the lowest prediction accuracy.

The Root Meat Square Error in predicting the amount of loss HMEQ loan is following for three models:

- TREE 5732.6842719501765
- RF 3203.798055210686
- GB 2641.9375614891906

This indicates that although Random Forests was best in predicting loan being defaulted, Gradient Boosting model performed better at predicting the loss loan amount with lower RMSE value. As a conclusion, I recommend using Random Forest method in predicting whether a loan could default and Gradient Boosting method to predict the severity of the defaulted loan amount.

BINGO BONUS: For this week, I tried using R to set up a simple Decision Tree model, and I noticed the SciKit's precise code for splitting Training and Test dataset to be much easier and Python seem to have a better visualization than R.

MSDS 422
Assignment 2
Juhwi Kim