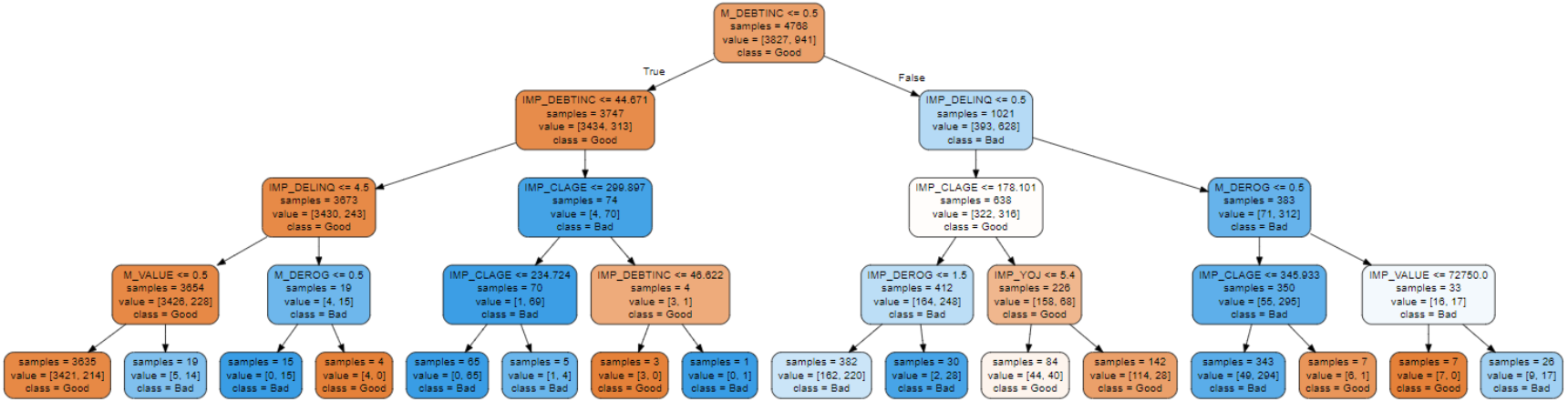
MSDS-422-SEC56_Assignment 2 Daniel Zaremba

Decision Trees

Probability of Default



Running the code multiple times using different random state numbers revealed that the following variables the predict loss showed up consistently:

M_VALUE

IMP_VALUE

IMP_YOJ

M_DEROG

IMP_DEROG

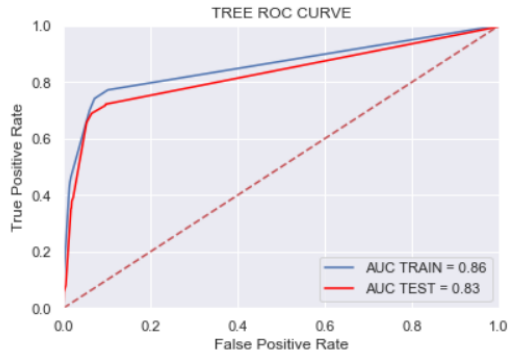
IMP_DELINQ

IMP_CLAGE

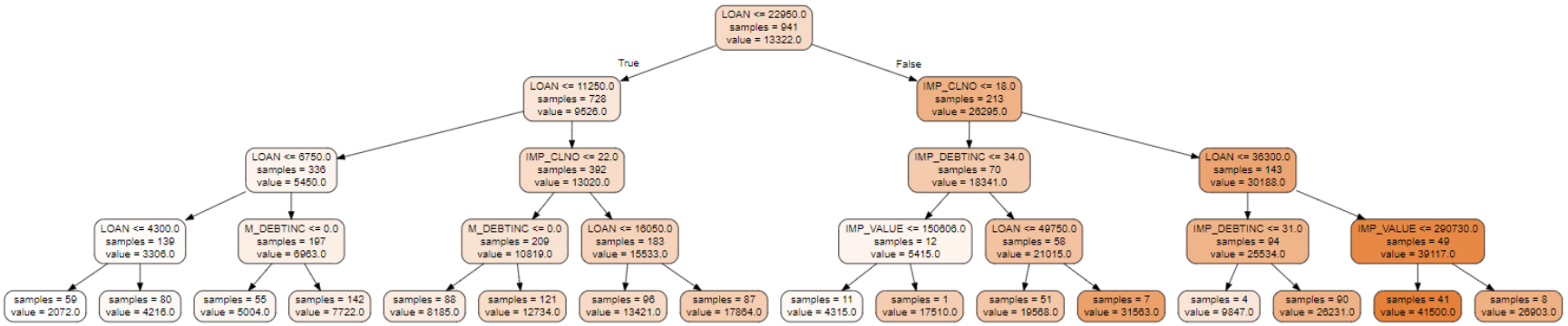
M_DEBTINC

IMP_DEBTINC

```
=============

DECISION TREE

Probability of Default
Accuracy Train: 0.8928271812080537
Accuracy Test: 0.886744966442953
```



Decision Tree - Predict Loss Amount



MEAN Train 13321.757704569607
MEAN Test 13154.826612903225

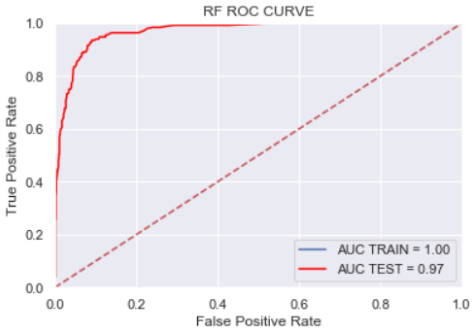
TREE RMSE Train: 4373.688039245928
TREE RMSE Test: 5179.4450845715855

LOAN
IMP_VALUE
IMP_CLNO
M_DEBTINC
IMP_DEBTINC

Random Forest

Probability of Default

```
=============

RANDOM FOREST

Probability of crash
Accuracy Train: 1.0
Accuracy Test: 0.9203020134228188


('M_DEBTINC', 100)
('IMP_DEBTINC', 71)
('IMP_CLAGE', 43)
('IMP_DELINQ', 40)
('LOAN', 37)
('IMP_VALUE', 36)
('IMP_MORTDUE', 35)
('IMP_CLNO', 33)
('IMP_YOJ', 28)
('IMP_DEROG', 22)
('IMP_NINQ', 19)
```



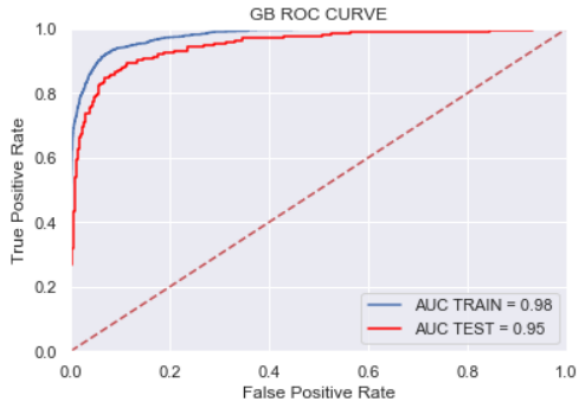
Random Forest – Predict Loss Amount

```
RF RMSE Train: 1139.1895440401704
RF RMSE Test: 2734.3656311917653
('LOAN', 100)
('IMP_CLNO', 13)
('IMP_DEBTINC', 6)
```

Gradient Boosting

Probability of Default

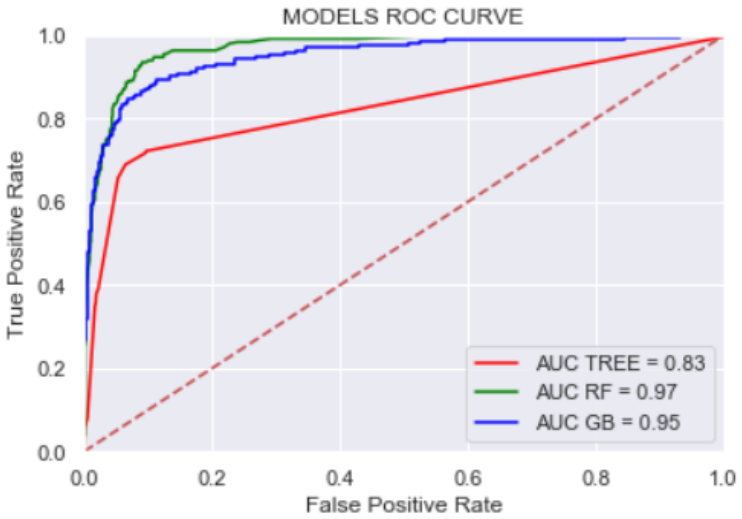
```
GRADIENT BOOSTING

Probability of Loss
Accuracy Train: 0.9425335570469798
Accuracy Test: 0.9161073825503355


('M_DEBTINC', 100)
('IMP_DEBTINC', 30)
('IMP_DELINQ', 20)
('IMP_CLAGE', 16)
('IMP_VALUE', 9)
('IMP_DEROG', 8)
```



Random Forest – Predict Loss Amount

```
GB RMSE Train: 850.1949815796479
GB RMSE Test: 1882.7168256622888
('LOAN', 100)
('IMP_CLNO', 15)
('IMP_DEBTINC', 6)
('M_DEBTINC', 5)
```

ROC Curves



Observations

- Include a discussion of the Decision Tree diagrams. Do they appear to make sense? After reviewing the variables that the decision tree diagrams reflected, they all appear to make sense and the if else relationship between the variables and their values also makes sense for ex. In the decision tree for probability of default, looking at IMP_DELINQ, if the value is less than 4.5 than that population is less risky than those that have a higher amount of delinquencies reflected on the credit report.
- Which variables appear to be most predictive of loan default? Do they make sense? The most predictive variable is M_DEBTINC which makes sense because if individuals have their debt to income ratio missing those individuals might have done so on purpose, as they are probably hiding their debt and their inability to pay back the loan, hence they are riskier.
- Which variables appear to be most predictive of loss amount? Do they make sense? The most predictive variable was LOAN, which makes sense because the bigger the loan the more risky a person is.
- If you were to select one of these models to put into production, which would it be? Why would you select this model? The Random Forest Model as it has the highest Test Accuracy.
- I played around with the parameters, for each of the models for the decision tree I primarily focused on tuning the max_depth and then also looked at min_samples_leaf and min_samples_split, neither of these two provided significant lift so I did not include. For the Random Forest and Gradient Boosting models I adjusted the parameters by setting n_estimators to 200 this provided a slight lift to the test accuracy in both models.