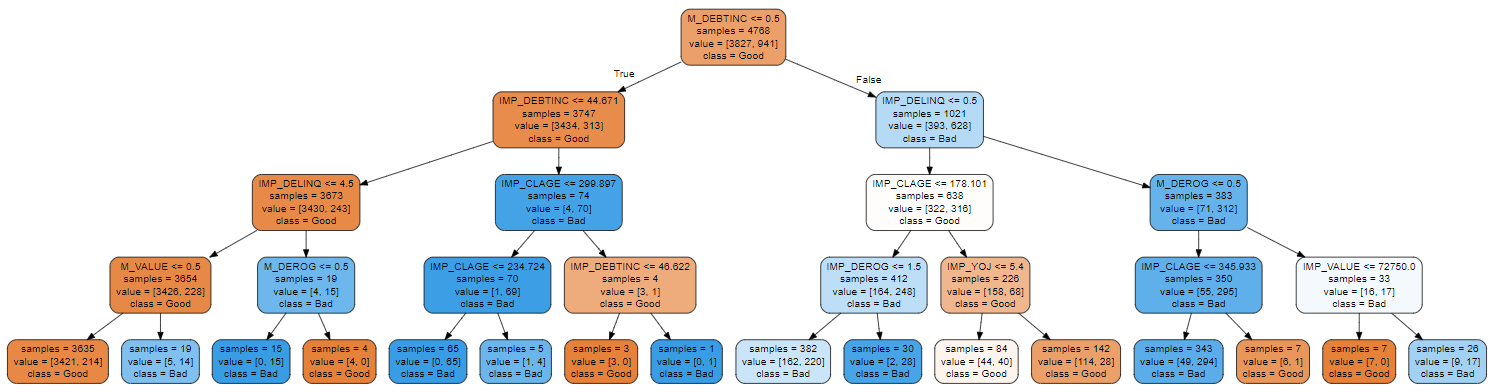
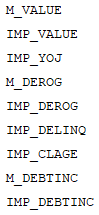
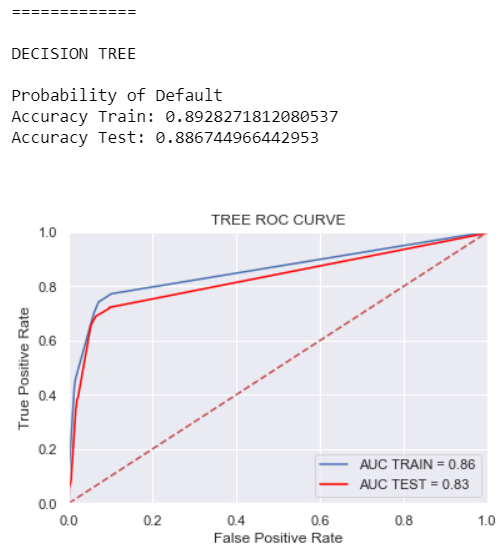
MSDS-422-SEC56\_Assignment 2 Daniel Zaremba

**Decision Trees**

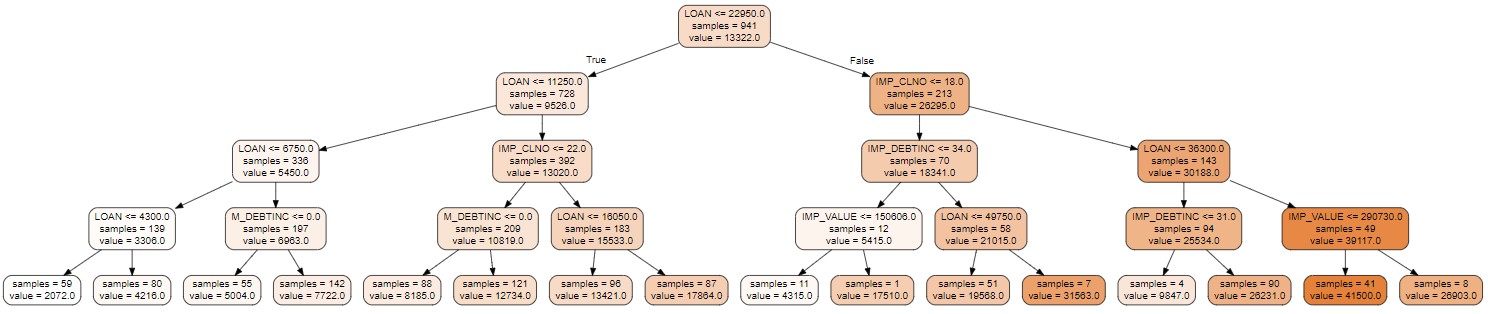
**Probability of Default**

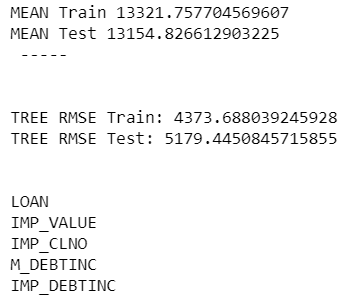
Running the code multiple times using different random state numbers revealed that the following variables the predict loss showed up consistently:





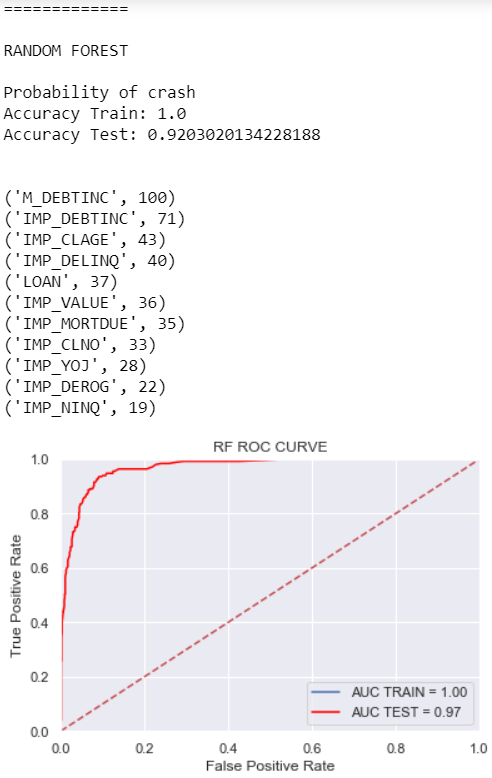
**Decision Tree - Predict Loss Amount**



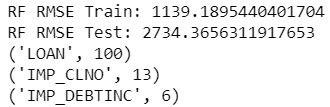


**Random Forest**

**Probability of Default**

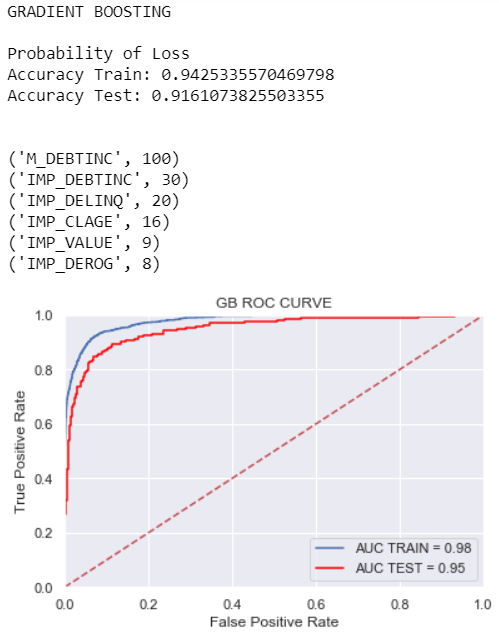


**Random Forest – Predict Loss Amount**

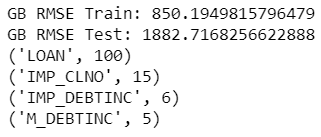


**Gradient Boosting**

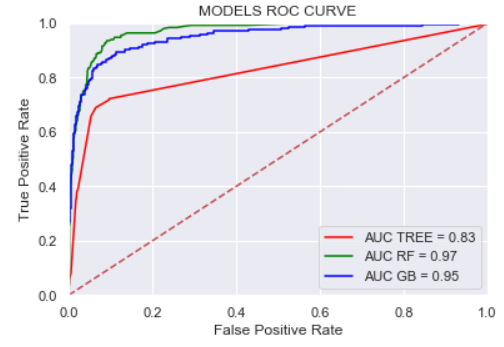
**Probability of Default**



**Random Forest – Predict Loss Amount**



**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.