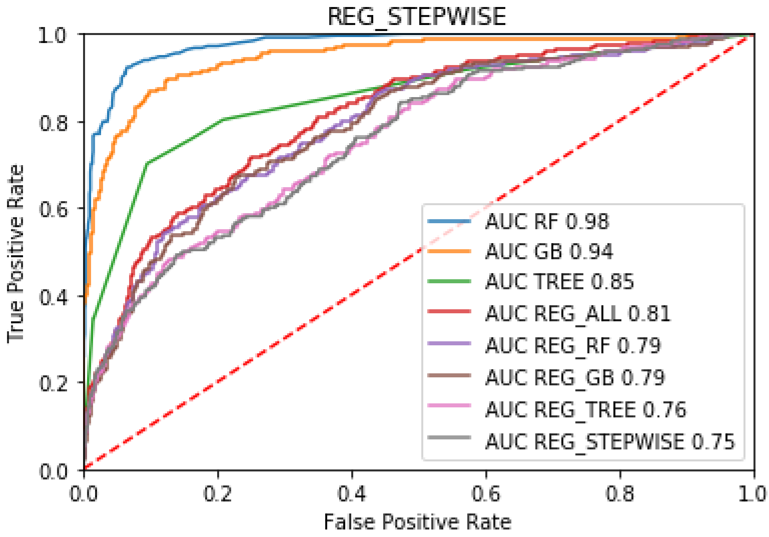
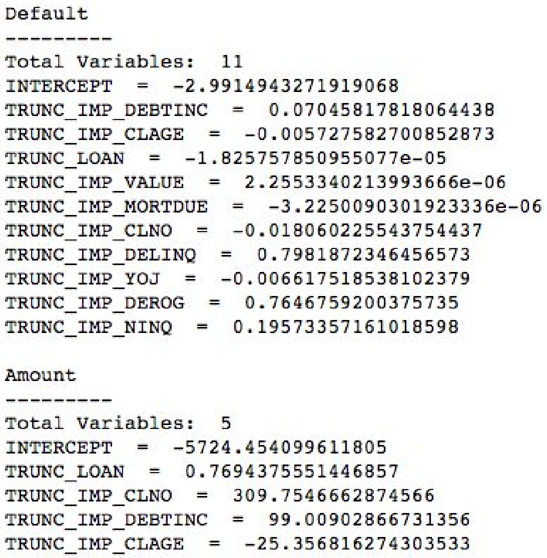
In this assignment, we leveraged Linear and Logistic Regression to help predict whether an individual would default on their loan and the amount that would be associated to the default. We took several approaches around variable selection; first, we ran a regression model leveraging all variables in our data set. Then we leveraged decision trees, Random Forest, Gradient Boosting, and Stepwise selection in order to determine which variables we would feed our regression models with.

Below is a graph that helps visualize the ROC curve for each model we built in order to allow us to easily compare which was the most accurate.



It was seen that the Random Forest-based Regression AUC score is the highest, showing the most accuracy when comparing the predicted results to the true data set. Based on the above results, I would recommend leveraging this model.



```
Default
---------
Total Variables:  11
INTERCEPT  =  -2.9914943271919068
TRUNC_IMP_DEBTINC  =  0.07045817818064438
TRUNC_IMP_CLAGE  =  -0.005727582700852873
TRUNC_LOAN  =  -1.825757850955077e-05
TRUNC_IMP_VALUE  =  2.2553340213993666e-06
TRUNC_IMP_MORTDUE  =  -3.2250090301923336e-06
TRUNC_IMP_CLNO  =  -0.018060225543754437
TRUNC_IMP_DELINQ  =  0.7981872346456573
TRUNC_IMP_YOJ  =  -0.006617518538102379
TRUNC_IMP_DEROG  =  0.7646759200375735
TRUNC_IMP_NINQ  =  0.19573357161018598

Amount
---------
Total Variables:  5
INTERCEPT  =  -5724.454099611805
TRUNC_LOAN  =  0.7694375551446857
TRUNC_IMP_CLNO  =  309.7546662874566
TRUNC_IMP_DEBTINC  =  99.00902866731356
TRUNC_IMP_CLAGE  =  -25.356816274303533
```

The above shows the coefficients for the variables selected for the regression model via Random Forest. These coefficients make the most sense because it is the model that produces the lowest amount of variables but also produces coefficients that trust just ‘make sense’, for the most part. What leads me to raise questions are items like TRUNC_IMP_DELINQ being positive in regards to classification; I would assume an individual would multiple delinquencies would be defaulting on their loan. While I am not an expert in home equity loans, I would partner and consult with someone who is to help interpret these variables as inputs for a potential production-grade model.

In regards to the linear regression and predicting the default amount. The Gradient Boosting-based regression model showed to be the most accurate (below). Again, while I am not an expert in the home equity space, I would consult one before implementing a specific model into production.

```
ALL DEFAULT MODEL ACCURACY
======
GB  =  3006.3484671784527
RF  =  3770.8264515729347
REG_ALL  =  4561.9550473980025
REG_RF  =  5687.399121822825
REG_GB  =  5687.399121822825
REG_STEPWISE  =  5760.055039945557
TREE  =  5770.601714157058
REG_TREE  =  5865.435239762463
------
```