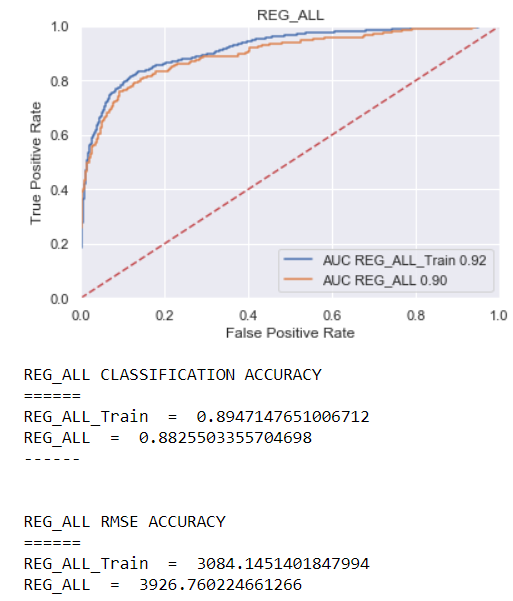
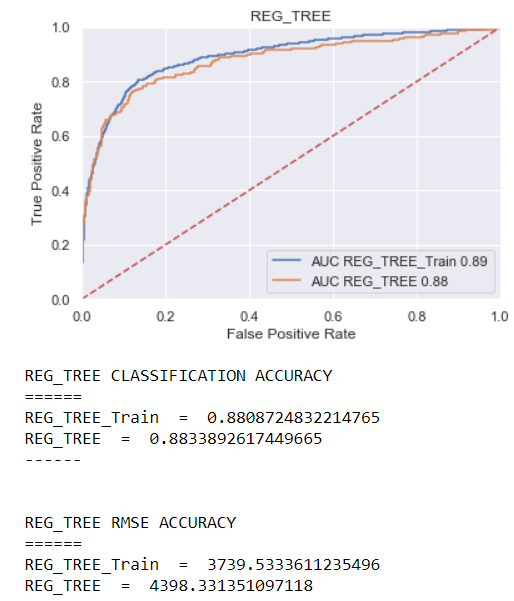
MSDS-422-SEC56\_Assignment 3 Daniel Zaremba

**Regression Observations**

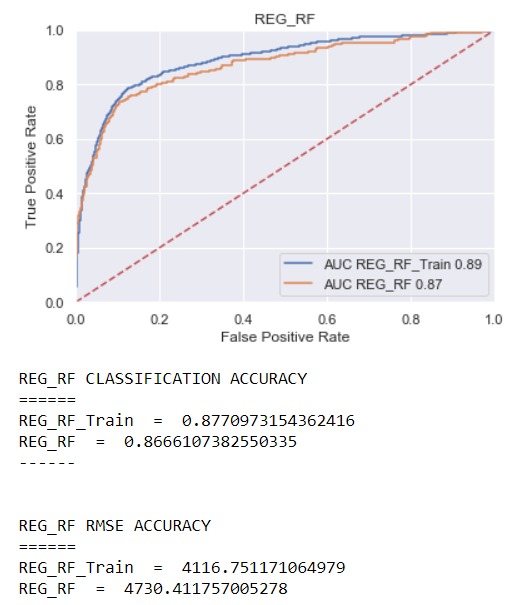
**Logistic Regression and Linear Regression models using all variables:**



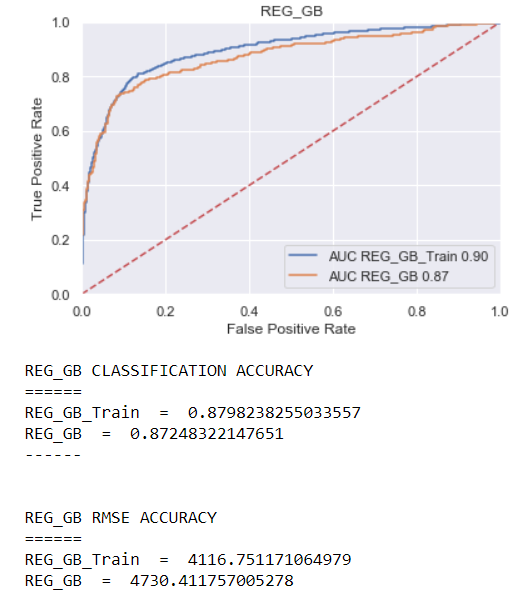
**Logistic Regression and Linear Regression models using Decision Tree variables:**

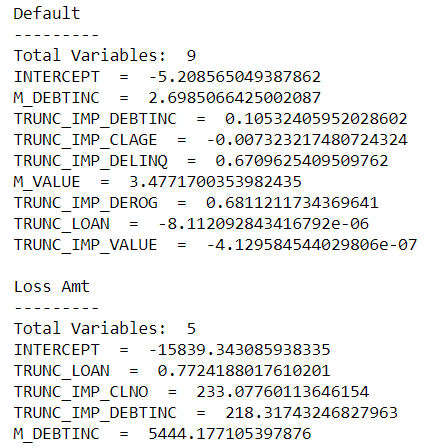


**Logistic Regression and Linear Regression models using Random Forest variables:**

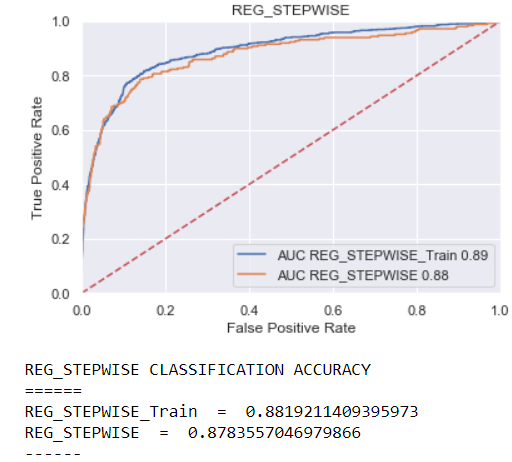


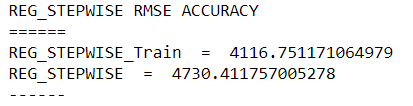
**Logistic Regression and Linear Regression models using Gradient Boosting variables:**

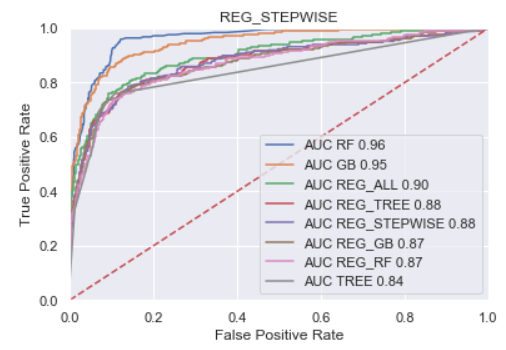


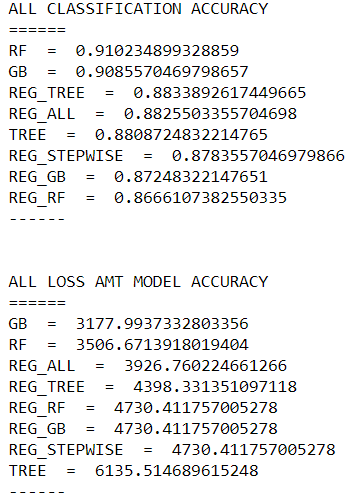


**Logistic Regression and Linear Regression models using Stepwise Selection variables:**









**Observations**

Having played around with the LogisticRegression Solver parameters, I settled on ‘newton-cg’, as the other alternatives I tried, i.e. ‘sag’, ‘saga’ and ‘lbfgs’ degraded the model. I would recommend using the Random Forest model as it has the highest classification accuracy. I would also recommend using the Gradient Boosting model for predicting Loss Amounts. The Default predictive variables for the most part make sense, M\_DEBTINC makes sense as individuals that don’t provide their debt to income ratio may be hiding high debt and be more risky, on the other hand I’m not sure about the M\_VALUE field, I would reach out to a subject matter expert to see if a missing home value is predictive of default. With respect to the Loss Amount variables again we see M\_DEBTINC and also TRUNC\_IMP\_DEBTINC which makes sense, also showing as predictive is TRUN\_IMP\_CLNO which is a double edge sword, the fewer the tradelines the riskier but any individual may have a lot of tradelines and run up a lot of debt which is also risky.