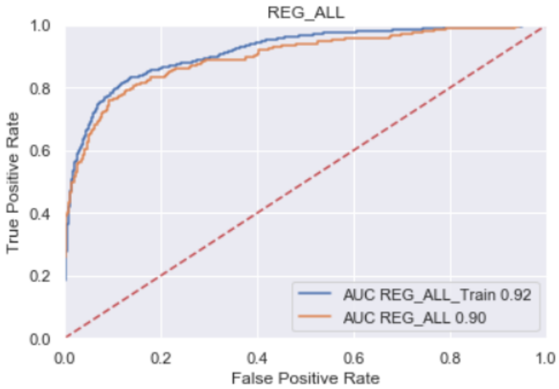
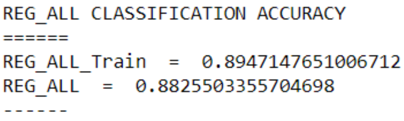
MSDS-422-SEC56_Assignment 3 Daniel Zaremba

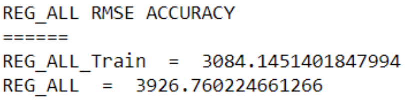
Regression Observations

Logistic Regression and Linear Regression models using all variables:



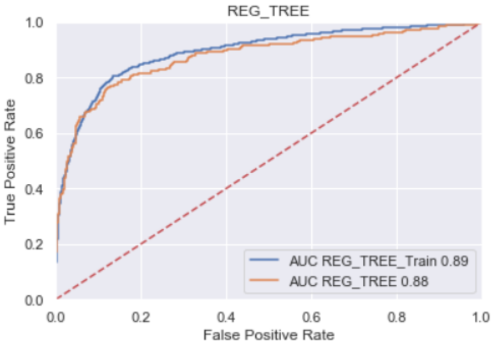


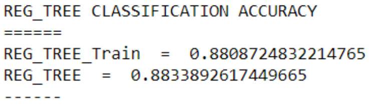
```
REG_ALL CLASSIFICATION ACCURACY
======
REG_ALL_Train  =  0.8947147651006712
REG_ALL  =  0.8825503355704698
------
```



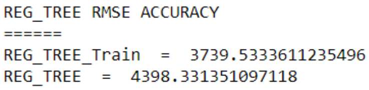
```
REG_ALL RMSE ACCURACY
======
REG_ALL_Train  =  3084.1451401847994
REG_ALL  =  3926.760224661266
```

Logistic Regression and Linear Regression models using Decision Tree variables:



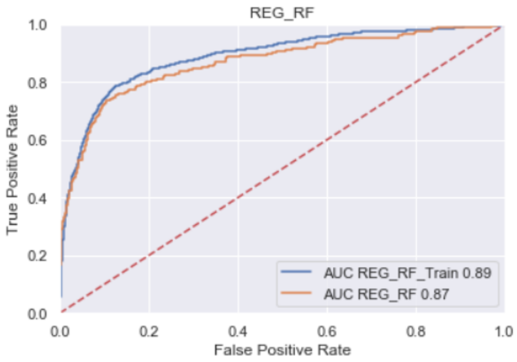


```
REG_TREE CLASSIFICATION ACCURACY
======
REG_TREE_Train  =  0.8808724832214765
REG_TREE  =  0.8833892617449665
------
```



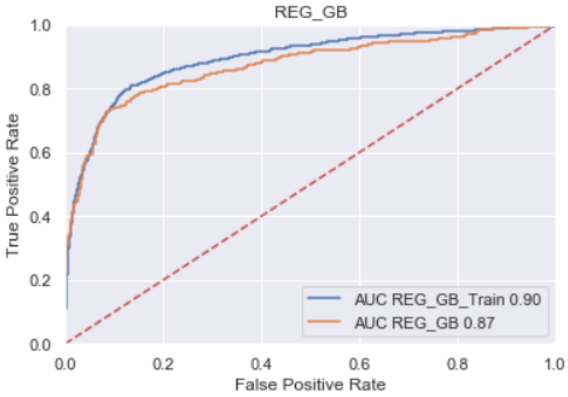
```
REG_TREE RMSE ACCURACY
======
REG_TREE_Train  =  3739.5333611235496
REG_TREE  =  4398.331351097118
```

Logistic Regression and Linear Regression models using Random Forest variables:

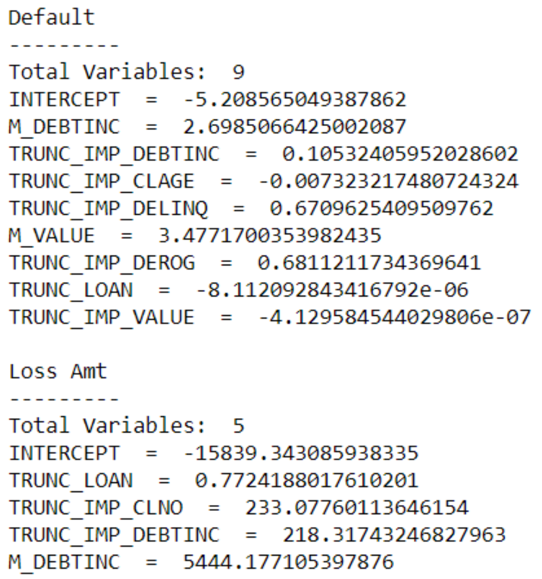


```
REG_RF CLASSIFICATION ACCURACY
======
REG_RF_Train  =  0.8770973154362416
REG_RF  =  0.8666107382550335
------


REG_RF RMSE ACCURACY
======
REG_RF_Train  =  4116.751171064979
REG_RF  =  4730.411757005278
```

Logistic Regression and Linear Regression models using Gradient Boosting variables:

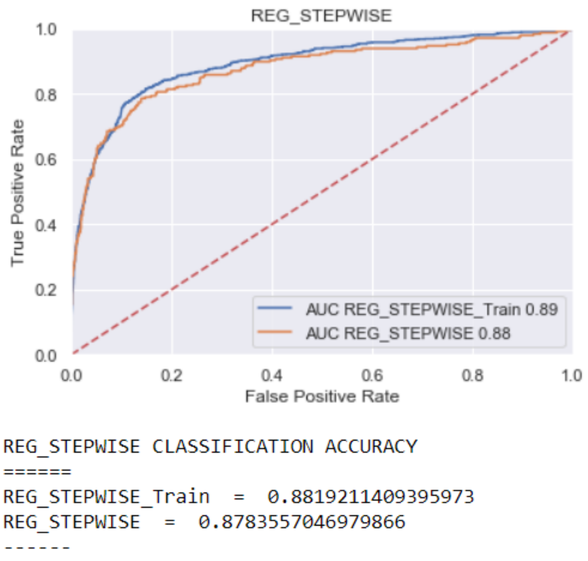


```
REG_GB CLASSIFICATION ACCURACY
======
REG_GB_Train  =  0.8798238255033557
REG_GB  =  0.87248322147651
------


REG_GB RMSE ACCURACY
======
REG_GB_Train  =  4116.751171064979
REG_GB  =  4730.411757005278
```

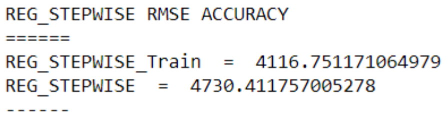


```
Default
---------
Total Variables:  9
INTERCEPT  =  -5.208565049387862
M_DEBTINC  =  2.6985066425002087
TRUNC_IMP_DEBTINC  =  0.10532405952028602
TRUNC_IMP_CLAGE  =  -0.007323217480724324
TRUNC_IMP_DELINQ  =  0.6709625409509762
M_VALUE  =  3.4771700353982435
TRUNC_IMP_DEROG  =  0.6811211734369641
TRUNC_LOAN  =  -8.112092843416792e-06
TRUNC_IMP_VALUE  =  -4.129584544029806e-07

Loss Amt
---------
Total Variables:  5
INTERCEPT  =  -15839.343085938335
TRUNC_LOAN  =  0.7724188017610201
TRUNC_IMP_CLNO  =  233.07760113646154
TRUNC_IMP_DEBTINC  =  218.31743246827963
M_DEBTINC  =  5444.177105397876
```

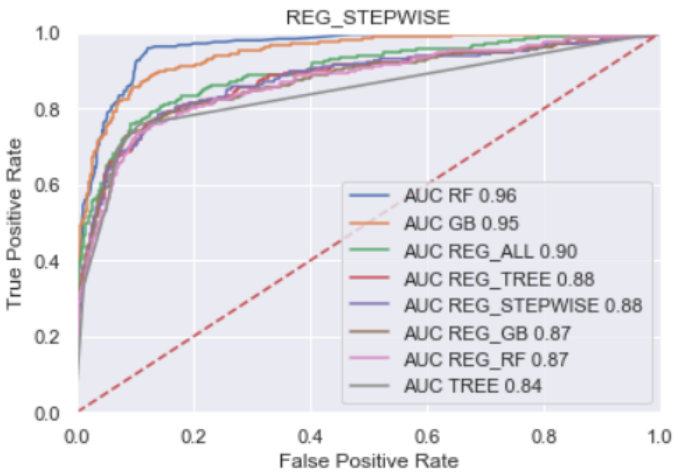
Logistic Regression and Linear Regression models using Stepwise Selection variables:

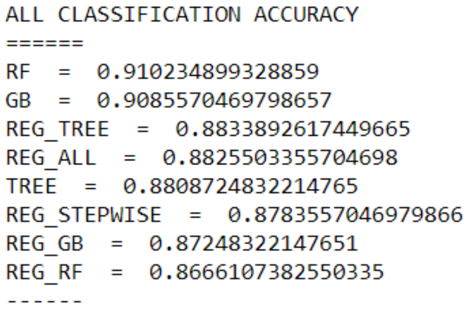


```
REG_STEPWISE CLASSIFICATION ACCURACY
======
REG_STEPWISE_Train  =  0.8819211409395973
REG_STEPWISE  =  0.8783557046979866
------
```

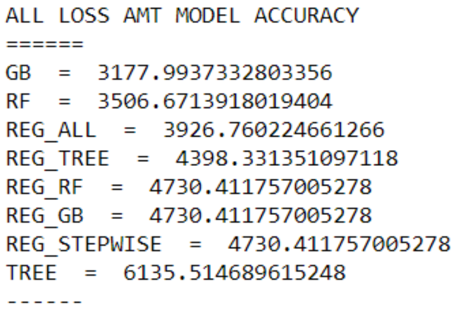


```
REG_STEPWISE RMSE ACCURACY
======
REG_STEPWISE_Train  =  4116.751171064979
REG_STEPWISE  =  4730.411757005278
------
```





```
ALL CLASSIFICATION ACCURACY
======
RF  =  0.910234899328859
GB  =  0.9085570469798657
REG_TREE  =  0.8833892617449665
REG_ALL  =  0.8825503355704698
TREE  =  0.8808724832214765
REG_STEPWISE  =  0.8783557046979866
REG_GB  =  0.87248322147651
REG_RF  =  0.8666107382550335
------
```



```
ALL LOSS AMT MODEL ACCURACY
======
GB  =  3177.9937332803356
RF  =  3506.6713918019404
REG_ALL  =  3926.760224661266
REG_TREE  =  4398.331351097118
REG_RF  =  4730.411757005278
REG_GB  =  4730.411757005278
REG_STEPWISE  =  4730.411757005278
TREE  =  6135.514689615248
------
```

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.