Vivian Xia

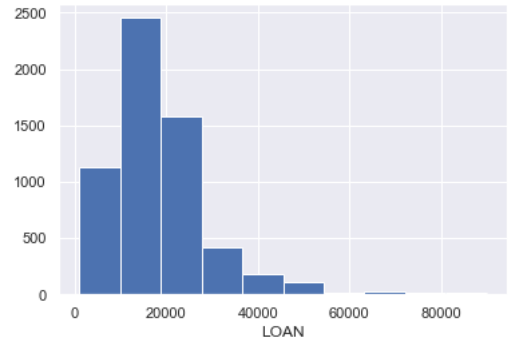
Assignment 3: Statistical Output

Remove Outliers

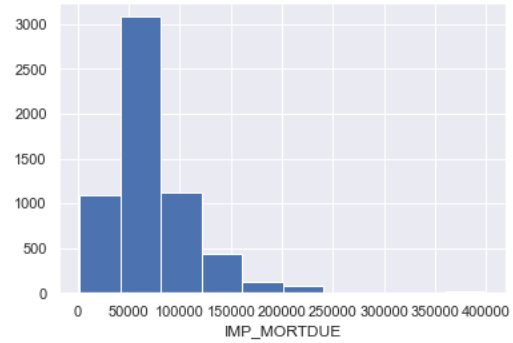
Original dataframe:

	count	mean	std	min	25%	50%	75%	max
TARGET_BAD_FLAG	5960.0	0.199497	0.399656	0.000000	0.000000	0.000000	0.000000	1.000000
TARGET_LOSS_AMT	1189.0	13414.576955	10839.455965	224.000000	5639.000000	11003.000000	17634.000000	78987.000000
LOAN	5960.0	18607.969799	11207.480417	1100.000000	11100.000000	16300.000000	23300.000000	89900.000000
z_IMP_REASON_HomeImp	5960.0	0.298658	0.457708	0.000000	0.000000	0.000000	1.000000	1.000000
z_IMP_REASON_MISSING	5960.0	0.042282	0.201248	0.000000	0.000000	0.000000	0.000000	1.000000
z_IMP_JOB_Mgr	5960.0	0.128691	0.334886	0.000000	0.000000	0.000000	0.000000	1.000000
z_IMP_JOB_Office	5960.0	0.159060	0.365763	0.000000	0.000000	0.000000	0.000000	1.000000
z_IMP_JOB_Other	5960.0	0.400671	0.490076	0.000000	0.000000	0.000000	1.000000	1.000000
z_IMP_JOB_ProfExe	5960.0	0.214094	0.410227	0.000000	0.000000	0.000000	0.000000	1.000000
z_IMP_JOB_Sales	5960.0	0.018289	0.134004	0.000000	0.000000	0.000000	0.000000	1.000000
z_IMP_JOB_Self	5960.0	0.032383	0.177029	0.000000	0.000000	0.000000	0.000000	1.000000
M_MORTDUE	5960.0	0.086913	0.281731	0.000000	0.000000	0.000000	0.000000	1.000000
IMP_MORTDUE	5960.0	73001.041812	42552.726779	2063.000000	48139.000000	65019.000000	88200.250000	399550.000000
M_VALUE	5960.0	0.018792	0.135801	0.000000	0.000000	0.000000	0.000000	1.000000
IMP_VALUE	5960.0	101540.387423	56869.436682	8000.000000	66489.500000	89235.500000	119004.750000	855909.000000
M_YOJ	5960.0	0.086409	0.280991	0.000000	0.000000	0.000000	0.000000	1.000000
IMP_YOJ	5960.0	8.756166	7.259424	0.000000	3.000000	7.000000	12.000000	41.000000
M_DEROG	5960.0	0.118792	0.323571	0.000000	0.000000	0.000000	0.000000	1.000000
IMP_DEROG	5960.0	0.224329	0.798458	0.000000	0.000000	0.000000	0.000000	10.000000
M_DELINQ	5960.0	0.097315	0.296412	0.000000	0.000000	0.000000	0.000000	1.000000
IMP_DELINQ	5960.0	0.405705	1.079256	0.000000	0.000000	0.000000	0.000000	15.000000
M_CLAGE	5960.0	0.051678	0.221394	0.000000	0.000000	0.000000	0.000000	1.000000
IMP_CLAGE	5960.0	179.440725	83.574697	0.000000	117.371430	173.466667	227.143058	1168.233561
M_NINQ	5960.0	0.085570	0.279752	0.000000	0.000000	0.000000	0.000000	1.000000
IMP_NINQ	5960.0	1.170134	1.653866	0.000000	0.000000	1.000000	2.000000	17.000000
M_CLNO	5960.0	0.037248	0.189386	0.000000	0.000000	0.000000	0.000000	1.000000
IMP_CLNO	5960.0	21.247819	9.951308	0.000000	15.000000	20.000000	26.000000	71.000000
M_DEBTINC	5960.0	0.212584	0.409170	0.000000	0.000000	0.000000	0.000000	1.000000
IMP_DEBTINC	5960.0	34.000651	7.644528	0.524499	30.763159	34.818262	37.949892	203.312149

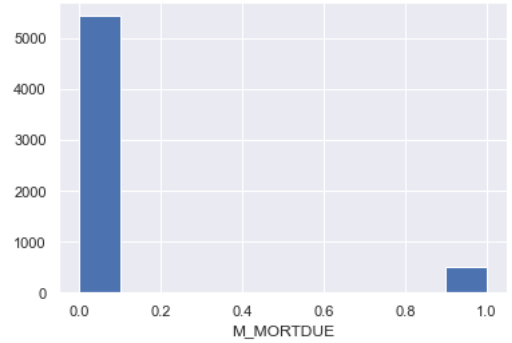
LOAN



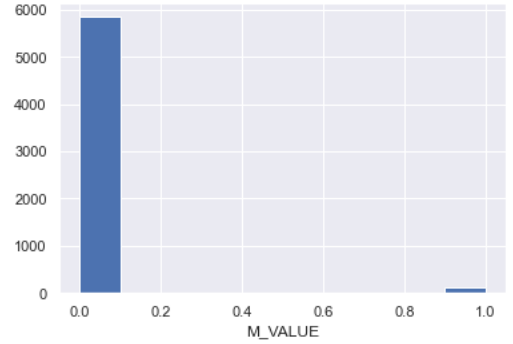
IMP_MORTDUE



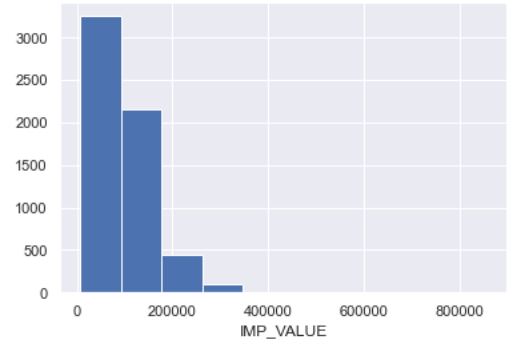
M_MORTDUE



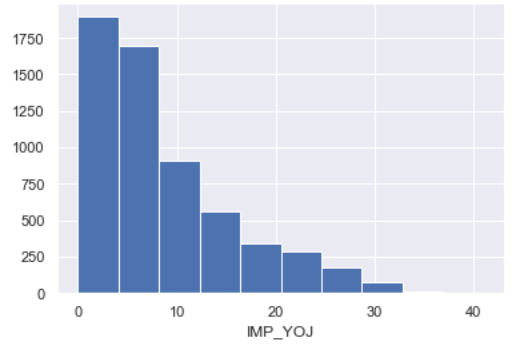
M_VALUE



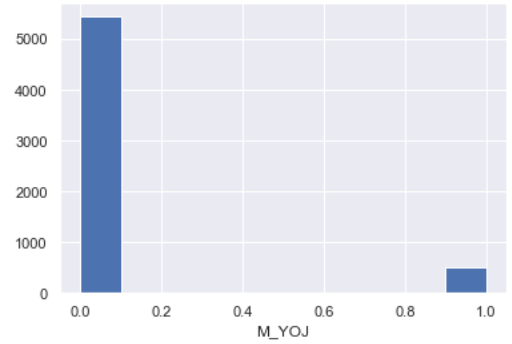
IMP_VALUE



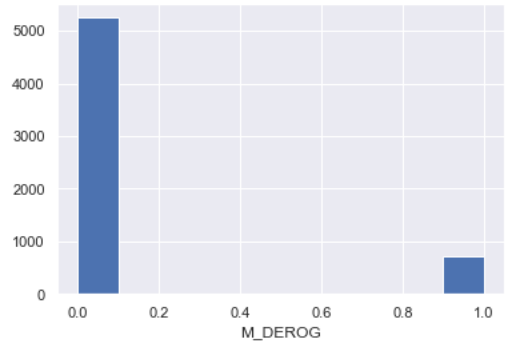
IMP_YOJ



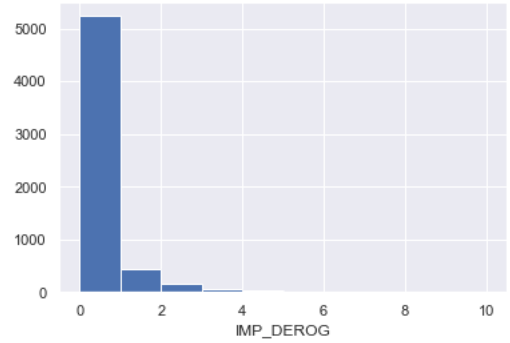
M_YOJ



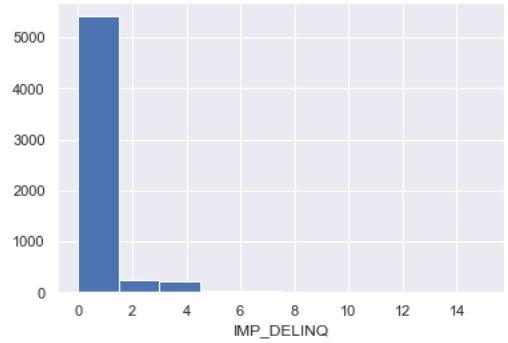
M_DEROG



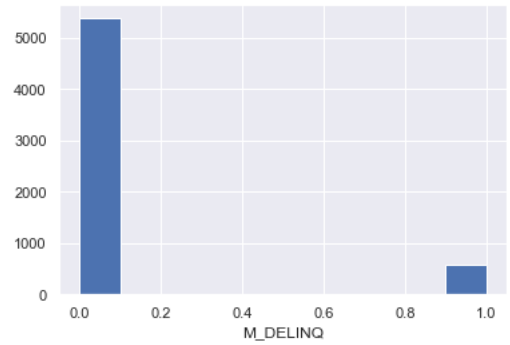
IMP_DEROG



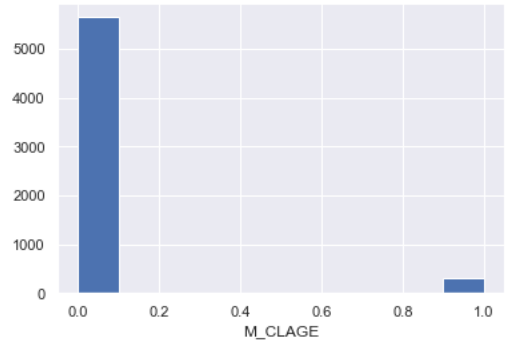
IMP_DELINQ



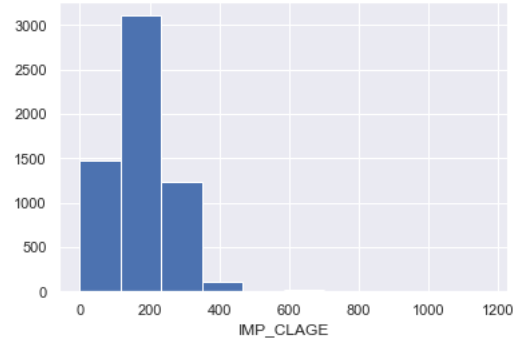
M_DELINQ



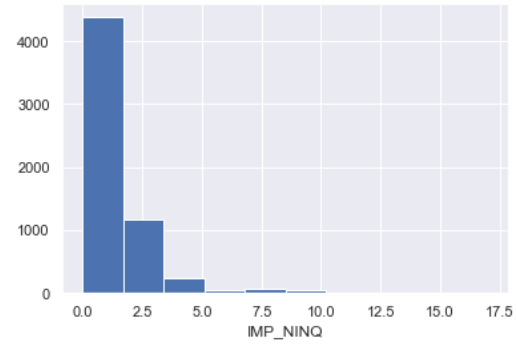
M_CLAGE



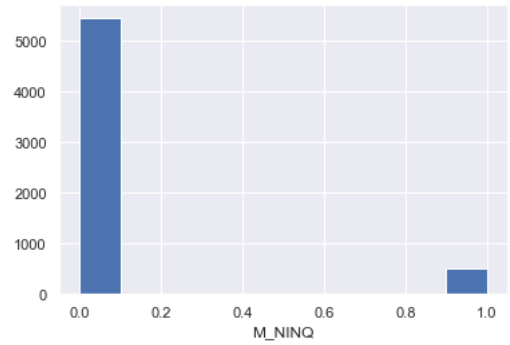
IMP_CLAGE



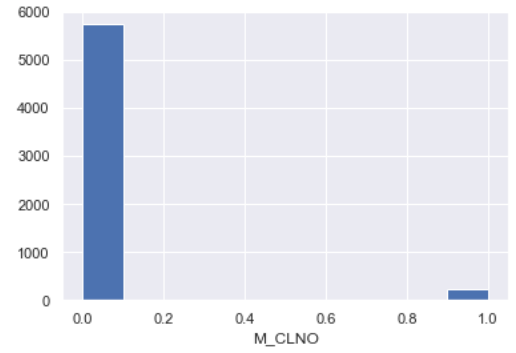
IMP_NINQ



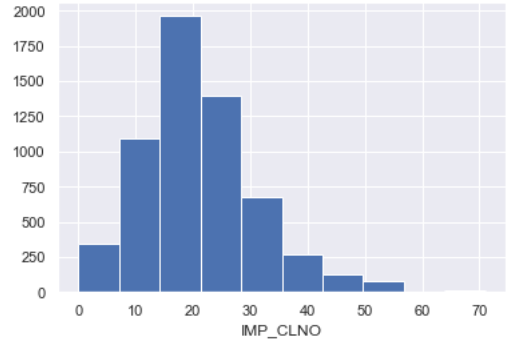
M_NINQ



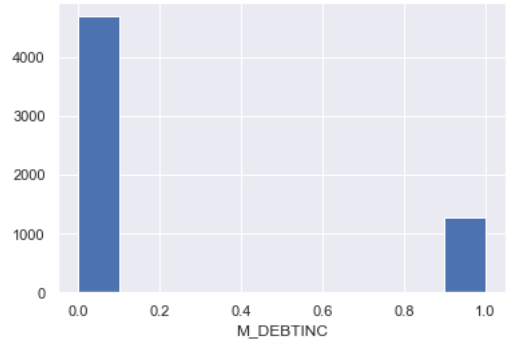
M_CLNO



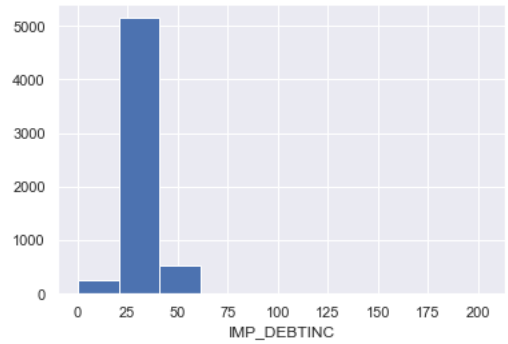
IMP_CLNO



M_DEBTINC



IMP_DEBTINC



The histograms for the columns that are not representing missing values but the actual values (LOAN and all the IMP_ columns) are generally skewed to the right because of outliers creating that right tail. In this step, the outliers are cutoff at the value 3 standard deviations away from their corresponding column means.

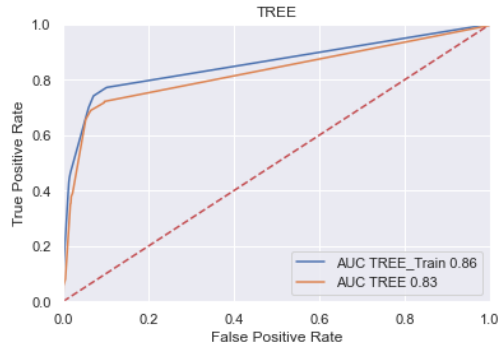
Dataframe with Cutoff:

	count	mean	std	min	25%	50%	75%	max
TARGET_BAD_FLAG	5960.0	0.199497	0.399656	0.000000	0.000000	0.000000	0.000000	1.0
TARGET_LOSS_AMT	1189.0	13414.576955	10839.455965	224.000000	5639.000000	11003.000000	17634.000000	78987.0
z_IMP_REASON_HomeImp	5960.0	0.298658	0.457708	0.000000	0.000000	0.000000	1.000000	1.0
z_IMP_REASON_MISSING	5960.0	0.042282	0.201248	0.000000	0.000000	0.000000	0.000000	1.0
z_IMP_JOB_Mgr	5960.0	0.128691	0.334886	0.000000	0.000000	0.000000	0.000000	1.0
z_IMP_JOB_Office	5960.0	0.159060	0.365763	0.000000	0.000000	0.000000	0.000000	1.0
z_IMP_JOB_Other	5960.0	0.400671	0.490076	0.000000	0.000000	0.000000	1.000000	1.0
z_IMP_JOB_ProfExe	5960.0	0.214094	0.410227	0.000000	0.000000	0.000000	0.000000	1.0
z_IMP_JOB_Sales	5960.0	0.018289	0.134004	0.000000	0.000000	0.000000	0.000000	1.0
z_IMP_JOB_Self	5960.0	0.032383	0.177029	0.000000	0.000000	0.000000	0.000000	1.0
O_LOAN	5960.0	0.015940	0.125252	0.000000	0.000000	0.000000	0.000000	1.0
TRUNC_LOAN	5960.0	18362.256711	10148.976515	1100.000000	11100.000000	16300.000000	23300.000000	52230.0
O_M_MORTDUE	5960.0	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.0
TRUNC_M_MORTDUE	5960.0	0.086913	0.281731	0.000000	0.000000	0.000000	0.000000	1.0
O_IMP_MORTDUE	5960.0	0.018121	0.133400	0.000000	0.000000	0.000000	0.000000	1.0
TRUNC_IMP_MORTDUE	5960.0	72201.825034	39017.412773	2063.000000	48139.000000	65019.000000	88200.250000	200659.0
O_M_VALUE	5960.0	0.018792	0.135801	0.000000	0.000000	0.000000	0.000000	1.0
TRUNC_M_VALUE	5960.0	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.0
O_IMP_VALUE	5960.0	0.015101	0.121964	0.000000	0.000000	0.000000	0.000000	1.0
TRUNC_IMP_VALUE	5960.0	100462.133732	50412.742922	8000.000000	66489.500000	89235.500000	119004.750000	272149.0
O_M_YOJ	5960.0	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.0
TRUNC_M_YOJ	5960.0	0.086409	0.280991	0.000000	0.000000	0.000000	0.000000	1.0
O_IMP_YOJ	5960.0	0.003020	0.054877	0.000000	0.000000	0.000000	0.000000	1.0
TRUNC_IMP_YOJ	5960.0	8.741737	7.208863	0.000000	3.000000	7.000000	12.000000	31.0
O_M_DEROG	5960.0	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.0
TRUNC_M_DEROG	5960.0	0.118792	0.323571	0.000000	0.000000	0.000000	0.000000	1.0
O_IMP_DEROG	5960.0	0.012081	0.109255	0.000000	0.000000	0.000000	0.000000	1.0
TRUNC_IMP_DEROG	5960.0	0.192114	0.582947	0.000000	0.000000	0.000000	0.000000	3.0
O_M_DELINQ	5960.0	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.0
TRUNC_M_DELINQ	5960.0	0.097315	0.296412	0.000000	0.000000	0.000000	0.000000	1.0
O_IMP_DELINQ	5960.0	0.015101	0.121964	0.000000	0.000000	0.000000	0.000000	1.0
TRUNC_IMP_DELINQ	5960.0	0.371309	0.886332	0.000000	0.000000	0.000000	0.000000	4.0
O_M_CLAGE	5960.0	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.0
TRUNC_M_CLAGE	5960.0	0.051678	0.221394	0.000000	0.000000	0.000000	0.000000	1.0
O_IMP_CLAGE	5960.0	0.005201	0.071939	0.000000	0.000000	0.000000	0.000000	1.0
TRUNC_IMP_CLAGE	5960.0	178.519248	78.853277	0.000000	117.371430	173.466667	227.143058	430.0
O_M_NINQ	5960.0	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.0
TRUNC_M_NINQ	5960.0	0.085570	0.279752	0.000000	0.000000	0.000000	0.000000	1.0
O_IMP_NINQ	5960.0	0.020302	0.141043	0.000000	0.000000	0.000000	0.000000	1.0
TRUNC_IMP_NINQ	5960.0	1.115101	1.408926	0.000000	0.000000	1.000000	2.000000	6.0
O_M_CLNO	5960.0	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.0
TRUNC_M_CLNO	5960.0	0.037248	0.189386	0.000000	0.000000	0.000000	0.000000	1.0
O_IMP_CLNO	5960.0	0.007047	0.083657	0.000000	0.000000	0.000000	0.000000	1.0
TRUNC_IMP_CLNO	5960.0	21.200671	9.783956	0.000000	15.000000	20.000000	26.000000	51.0
O_M_DEBTINC	5960.0	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.0
TRUNC_M_DEBTINC	5960.0	0.212584	0.409170	0.000000	0.000000	0.000000	0.000000	1.0
O_IMP_DEBTINC	5960.0	0.005201	0.071939	0.000000	0.000000	0.000000	0.000000	1.0
TRUNC_IMP_DEBTINC	5960.0	33.877119	6.649744	0.524499	30.763159	34.818262	37.949892	57.0

The truncated columns have a maximum value that is much smaller and closer to the 75% value in its respective columns.

Decision Tree Classifier

Loan Default Probability



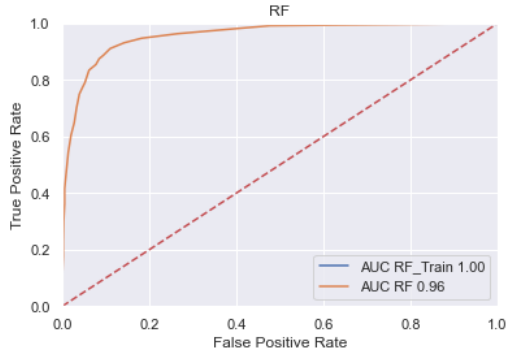

```
TREE CLASSIFICATION ACCURACY
======
TREE_Train  =  0.8928271812080537
TREE  =  0.886744966442953
------
```

Loss Amount

```
TREE RMSE ACCURACY
======
TREE_Train  =  4587.556685671267
TREE  =  5763.9837632219205
------
```

Random Forest Classifier

Loan Default Probability



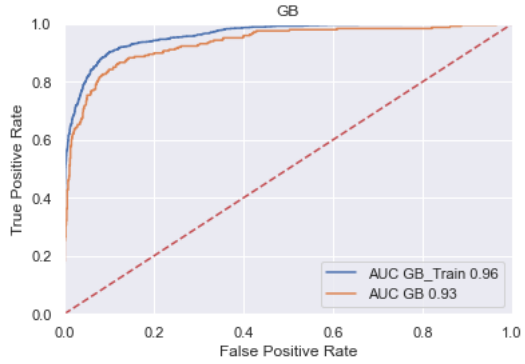

```
RF CLASSIFICATION ACCURACY
======
RF_Train  =  0.9993708053691275
RF  =  0.9119127516778524
------
```

Loss Amount

```
RF RMSE ACCURACY
======
RF_Train  =  1317.8895879806714
RF  =  3428.650015145327
------
```

Gradient Boosting Classifier

Loan Default Probability




```
GB CLASSIFICATION ACCURACY
======
GB_Train  =  0.9238674496644296
GB  =  0.9060402684563759
------
```

Loss Amount

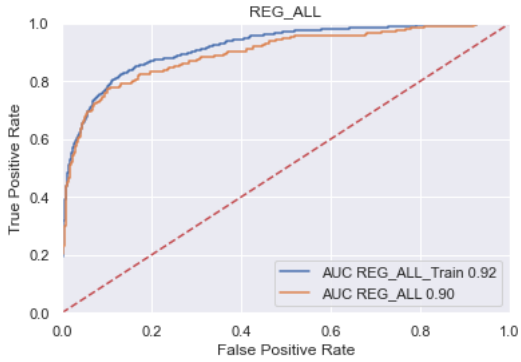
```
GB RMSE ACCURACY
======
GB_Train  =  1245.4603241383124
GB  =  2794.947030828998
------
```

For the loan default probability, the Random Tree Classifier model was the most accurate from its classification accuracy score and area under the curve.

The Gradient Boosting model was the most accurate for predicting loss amount because it had the smallest RMSE value compared to the other models.

Regression All Variables

Loan Default Probability



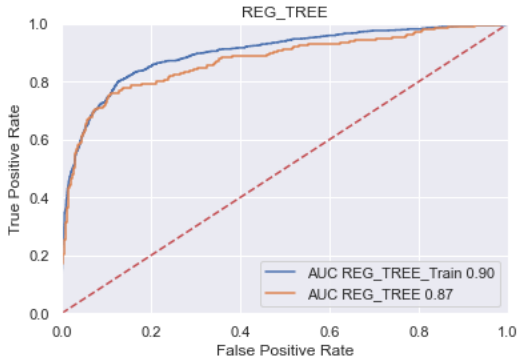

```
REG_ALL CLASSIFICATION ACCURACY
======
REG_ALL_Train  =  0.893246644295302
REG_ALL  =  0.886744966442953
------
```

Loss Amount

```
REG_ALL RMSE ACCURACY
======
REG_ALL_Train  =  3555.90846939125
REG_ALL  =  3615.1006384119664
------
```

Regression Decision Tree

Loan Default Probability



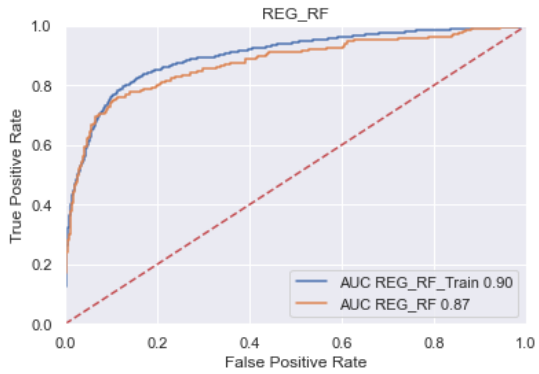

```
REG_TREE CLASSIFICATION ACCURACY
======
REG_TREE_Train  =  0.8863255033557047
REG_TREE  =  0.8800335570469798
------
```

Loss Amount

```
REG_TREE RMSE ACCURACY
======
REG_TREE_Train  =  4570.988960768495
REG_TREE  =  5114.69180925806
------
```

Regression Random Forest

Loan Default Probability




```
REG_RF CLASSIFICATION ACCURACY
======
REG_RF_Train  =  0.8800335570469798
REG_RF  =  0.87751677852349
------
```

Loss Amount

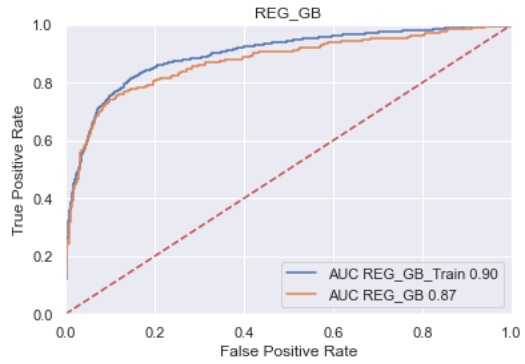
```
REG_RF RMSE ACCURACY
======
REG_RF_Train  =  4238.056494916955
REG_RF  =  4834.631645572038
------
```

Coefficients for Model

```
LOAN DEFAULT
---------
Total Variables:  13
INTERCEPT  =  -5.277350491799424
TRUNC_M_DEBTINC  =  2.6977038943919815
TRUNC_IMP_DEBTINC  =  0.10337974267018832
TRUNC_IMP_CLAGE  =  -0.005881064244565169
TRUNC_LOAN  =  -5.044107866009968e-06
TRUNC_IMP_VALUE  =  1.5802338406434643e-06
TRUNC_IMP_MORTDUE  =  -1.1578982148828966e-06
TRUNC_IMP_CLNO  =  -0.018749336039457866
TRUNC_IMP_DELINQ  =  0.7101946567812808
TRUNC_IMP_YOJ  =  -0.011837837969276883
TRUNC_IMP_DEROG  =  0.7205428184021219
TRUNC_IMP_NINQ  =  0.1325985504151847
O_M_VALUE  =  3.584420795677956

LOSS AMOUNT
---------
Total Variables:  6
INTERCEPT  =  -13269.322689831992
TRUNC_LOAN  =  0.7957096877509179
TRUNC_IMP_CLNO  =  295.8661953290628
TRUNC_IMP_DEBTINC  =  198.11601519151213
TRUNC_M_DEBTINC  =  5813.596604342766
TRUNC_IMP_CLAGE  =  -25.63038798406322
```

Regression Gradient Boosting

Loan Default Probability




```
REG_GB CLASSIFICATION ACCURACY
======
REG_GB_Train  =  0.881501677852349
REG_GB  =  0.8800335570469798
------
```

Loss Amount

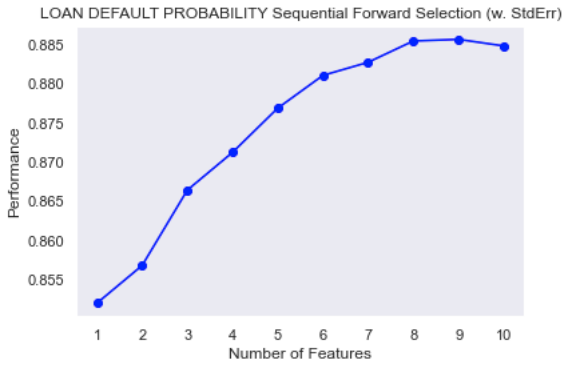
```
REG_GB RMSE ACCURACY
======
REG_GB_Train  =  4238.056494916955
REG_GB  =  4834.631645572038
------
```

Coefficients for Model

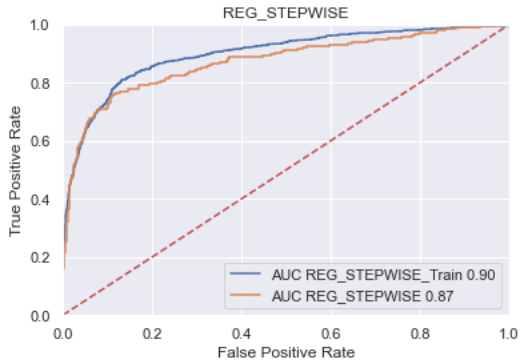
```
LOAN DEFAULT
---------
Total Variables:  11
INTERCEPT  =  -5.24430143027247
TRUNC_M_DEBTINC  =  2.738760511600896
TRUNC_IMP_DEBTINC  =  0.10578807558039724
TRUNC_IMP_DELINQ  =  0.7065760942356596
TRUNC_IMP_CLAGE  =  -0.00610884912245422
TRUNC_IMP_DEROG  =  0.7576642576894351
O_M_VALUE  =  3.5432677105315036
TRUNC_IMP_VALUE  =  5.686942832087663e-07
TRUNC_IMP_YOJ  =  -0.013075526298365439
TRUNC_IMP_CLNO  =  -0.016243098349593033
TRUNC_LOAN  =  -2.546639561948394e-06

LOSS AMOUNT
---------
Total Variables:  6
INTERCEPT  =  -13269.322689831992
TRUNC_LOAN  =  0.7957096877509179
TRUNC_IMP_CLNO  =  295.8661953290628
TRUNC_IMP_DEBTINC  =  198.11601519151213
TRUNC_M_DEBTINC  =  5813.596604342766
TRUNC_IMP_CLAGE  =  -25.63038798406322
```

Regression Stepwise

Loan Default Probability




```
 ..................
argmax
feature_names    (0, 1, 2, 3, 4, 6, 7, 8, 9)
avg_score                                0.885697
Name: 9, dtype: object
 ..................
('0', '1', '2', '3', '4', '6', '7', '8', '9')
```




```
REG_STEPWISE CLASSIFICATION ACCURACY
======
REG_STEPWISE_Train  =  0.885486577181208
REG_STEPWISE  =  0.8842281879194631
------
```

The variation of variables that were used in this stepwise model used the variables chosen from the list of variables used in the decision tree. The stepwise model found that the model with 9 variables out of the 11 total from those used in the decision tree were most accurate in predicting loan default probability. The 9 variables include:

```
TRUNC_IMP_CLAGE
TRUNC_M_DEBTINC
TRUNC_IMP_DEBTINC
O_M_VALUE
TRUNC_IMP_VALUE
TRUNC_M_DEROG
TRUNC_IMP_DEROG
O_IMP_DELINQ
TRUNC_IMP_DELINQ
```

Loss Amount


```
 ..................
argmax
feature_names    (0, 1, 2, 3, 4)
avg_score               0.824332
Name: 5, dtype: object
 ..................
('0', '1', '2', '3', '4')
```

```
REG_GB RMSE ACCURACY
======
REG_GB_Train  =  4238.056494916955
REG_GB  =  4834.631645572038
------
```

The stepwise model resulted in five variables being used to result in the best accuracy. The variables used included the following:

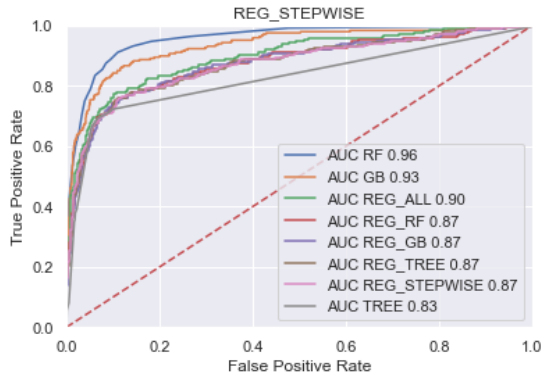
```
TRUNC_LOAN
TRUNC_IMP_CLNO
TRUNC_IMP_DEBTINC
TRUNC_M_DEBTINC
TRUNC_IMP_CLAGE
```

Coefficients for Model

```
LOAN DEFAULT
---------
Total Variables:  10
INTERCEPT  =  -5.353530813656673
TRUNC_IMP_CLAGE  =  -0.006837020088604123
TRUNC_M_DEBTINC  =  2.744201933282898
TRUNC_IMP_DEBTINC  =  0.10528069849571041
O_M_VALUE  =  3.709820349727899
TRUNC_IMP_VALUE  =  -8.007909005536421e-07
TRUNC_M_DEROG  =  -0.7852268386953548
TRUNC_IMP_DEROG  =  0.6922553233448552
O_IMP_DELINQ  =  2.0537201808929995
TRUNC_IMP_DELINQ  =  0.6176089000725961

LOSS AMOUNT
---------
Total Variables:  6
INTERCEPT  =  -13269.322689831992
TRUNC_LOAN  =  0.7957096877509179
TRUNC_IMP_CLNO  =  295.8661953290628
TRUNC_IMP_DEBTINC  =  198.11601519151213
TRUNC_M_DEBTINC  =  5813.596604342766
TRUNC_IMP_CLAGE  =  -25.63038798406322
```

Regression Comparison

Loan Default Probability




```
ALL CLASSIFICATION ACCURACY
======
RF  =  0.9119127516778524
GB  =  0.9060402684563759
REG_ALL  =  0.886744966442953
TREE  =  0.886744966442953
REG_STEPWISE  =  0.8842281879194631
REG_GB  =  0.8800335570469798
REG_TREE  =  0.8800335570469798
REG_RF  =  0.87751677852349
------
```

As mentioned above, the Random Forest Classifier model was the most accurate among the other Classifier models. For the loan default probability regression, the model with all the variables was the most accurate, but the use of all variables is not realistic and not all variables would make sense in the model. I would recommend using the regression stepwise model that included 9 variables, especially as the accuracy between the regression with all variables and the regression stepwise model differed by 0.002. The following are the 9 variables and intercept as well as their corresponding coefficient values:

```
LOAN DEFAULT
---------
Total Variables:  10
INTERCEPT  =  -5.353530813656673
TRUNC_IMP_CLAGE  =  -0.006837020088604123
TRUNC_M_DEBTINC  =  2.744201933282898
TRUNC_IMP_DEBTINC  =  0.10528069849571041
O_M_VALUE  =  3.709820349727899
TRUNC_IMP_VALUE  =  -8.007909005536421e-07
TRUNC_M_DEROG  =  -0.7852268386953548
TRUNC_IMP_DEROG  =  0.6922553233448552
O_IMP_DELINQ  =  2.0537201808929995
TRUNC_IMP_DELINQ  =  0.6176089000725961
```

These variables seem to make sense in predicting loan default probability. TRUNC_IMP_CLAGE decreases the loan default probability by 0.0068 for every additional month of that user's credit line age, which makes sense because those who have had their credit for longer are typically less risky. A missing debt-to-income ratio increases the chances of loan default which makes sense as those who leave it off usually may possess a bad or high ratio, so it also makes sense that the TRUNC_IMP_DEBTINC would increase the chances when the person's ratio is higher. It is interesting to see that the value of the house

and if it was missing would affect the chances of the loan default since this is usually compared to outstanding mortgage balance to evaluate how risky the person is. The coefficient value of TRUNC_IMP_VALUE is very small, so some consideration on removing this variable should be considered. It is also interesting to see that a missing derogatory marks on credit record value would result would decrease the chance loan default for that user. TRUNC_IMP_DEROG, O_IMP_DELINQ, TRUNC_IMP_DELINQ variables make sense since the more marks or delinquencies on the user's report, the greater the chances of the loan defaulting.

Loss Amount

```
ALL DAMAGE MODEL ACCURACY
======
GB  =  2794.947030828998
RF  =  3428.650015145327
REG_ALL  =  3615.1006384119664
REG_RF  =  4834.631645572038
REG_GB  =  4834.631645572038
REG_STEPWISE  =  4834.631645572038
REG_TREE  =  5055.795128862214
TREE  =  5763.9837632219205
------
```

The most accurate model in predicting loss amount is the Gradient Boosting Classifier model due to its low RMSE value. For the regression, the model with all the variables had the lowest RMSE score, so it was the most accurate. It is not realistic to use all the variables, so I would recommend using the regression with the Gradient Boosting model. The regression using Random Forest and Gradient Boosting have very similar RMSE values.

The RF regression uses the following variables:

```
LOSS AMOUNT
---------
Total Variables:  6
INTERCEPT  =  -13269.322689831992
TRUNC_LOAN  =  0.7957096877509179
TRUNC_IMP_CLNO  =  295.8661953290628
TRUNC_IMP_DEBTINC  =  198.11601519151213
TRUNC_M_DEBTINC  =  5813.596604342766
TRUNC_IMP_CLAGE  =  -25.63038798406322
```

The GB regression uses the following variables:

```
LOSS AMOUNT
---------
Total Variables:  6
INTERCEPT  =  -13269.322689831992
TRUNC_LOAN  =  0.7957096877509179
TRUNC_IMP_CLNO  =  295.8661953290628
TRUNC_IMP_DEBTINC  =  198.11601519151213
TRUNC_M_DEBTINC  =  5813.596604342766
TRUNC_IMP_CLAGE  =  -25.63038798406322
```

They both use the same number of variables and have the same values for coefficients and intercept of the model. There does not seem to be any difference in these two regression models, so I recommend

using either the Random Forest or Gradient Boosting regression model. The variables used all seem to make sense in predicting loss amount. It is interesting to see that a missing debt-to-income ratio would increase the loss amount by more than five thousand dollars, but it makes sense because those who leave it off may generally know it is a bad ratio. TRUNC_IMP_CLAGE makes sense since for every additional month of that person's credit line age, the less risky that person is and for the bank to probably would pay in losses. TRUNC_IMP_CLNO also makes sense as for every additional credit line the person has, the person may be riskier because of the potential to run up more debt, so the loss amount also increases. A higher debt-to-income ratio usually implies that the person is riskier, so TRUNC_IMP_DEBTINC makes sense to increase in loss amount value if the debt-to-income ratio also increases.