Assignment 3 – MSDS 422 Jamia Russell

The goal of this assignment is to complete regression based predictive models to analyze the Home Equity Line of Credit data set. The defining variables in this data set include "FLAG" data which indicates whether or not a loan has been defaulted, and "LOSS" which displays the amount of money lost by the bank for loans that went poorly.

SEE ATTACHED PDF WITH PYTHON CODE, GRAPHS AND STATISTICAL OUTPUT

ALL VARIALE FINDINGS:

The findings show a classification accuracy on the training data set at 89.3% and an accuracy on the testing set at 88.1%. This indicates that the model performed well on the training data set and slightly lower on the testing data set suggesting the model does well without overfitting data. In addition, RMSE accuracy for the training data set is 3763.23 compared to the testing data set at 3754.74 the testing dataset RMSE being lower shows that the model performs better on the data trained than the testing data. Although relatively small, the difference shows how well the model responds to new data.

Coefficients:

Flag: The coefficients allow is the to see the direction and extent of each coefficients predicting power on the baseline by the model. The FLAG model (logistic regression) has shown us positive impacts on coefficients such as debt income ratio, number of inquiries, derogatory marks, value of property, manager job position, debt consolidation, and home improvement all increase the likelihood of loan default. While credit line age, years on job, and office jobs decrease the likelihood of loan default. This is reflective of the real world.

Loss: The damage/loss model coefficients represent the impact of coefficient predictors on the total loss amount taken on by the bank. The model found debt consolidation, job classified as "other", sales jobs, value of the property, derogatory credit marks, delinquencies, credit lines, inquiries, and debt to income ratios having positive impact indicating a higher predicted loss. Age of credit line, years on job, mortgage amount, and manager positions had negative impacts suggesting lower loss amount.

DECISION TREE FINDINGS:

Findings show a classification accuracy on the training data set at 88.1% and an accuracy on the testing set at 86.6%. This indicates that the model performed well on the training data set and slightly lower on the testing data set suggesting the model does well without overfitting data. In addition, RMSE accuracy for the training data set is 3763.24 compared to the testing data set at 3850.54 the testing dataset RMSE being lower shows that the model performs better on the data trained than the testing data. Although relatively small, the difference shows how well the model responds to new data.

Coefficients:

Flag: The model has shown us positive impacts on coefficients such as derogatory marks, delinquencies, and debt to income ratio indicate a higher baseline probability of loan default. While credit line age, indicates a lower baseline probability of loa default. This is reflective of the real world.

Loss: The model found derogatory credit marks, delinquencies, credit lines, and debt to income ratios having positive impact indicating a higher predicted loss. Age of credit line had negative impacts suggesting lower loss amount.

RANDOM FOREST FINDINGS:

Findings show a classification accuracy on the training data set at 87.4% and an accuracy on the testing set at 86.5%. This indicates that the model performed well on the training data set and slightly lower on the testing data set suggesting the model does well without overfitting data. In addition, RMSE accuracy for the training data set is 3859.48 compared to the testing data set at 3845.85 the testing dataset RMSE being lower shows that the model performs better on the data trained than the testing data. Although relatively small, the difference shows how well the model responds to new data.

Coefficients:

Flag: Coefficients debt to income ratio, and delinquencies show to be significant predictors of loan default. While credit line age has a minor negative effect.

Loss: The model found delinquencies, and debt to income ratios having positive impact indicating a higher predicted loss. Age of credit line has a substantial negative impact suggesting lower loss amount.

GRADIENT BOOSTING:

Findings show a classification accuracy on the training data set at 87.7% and an accuracy on the testing set at 87.5%. This indicates that the model performed well on the training data set and slightly lower on the testing data set suggesting the model does well without overfitting data. In addition, RMSE accuracy for the training data set is 3788.53 compared to the testing data set at 3850.89 the testing dataset RMSE being lower shows that the model performs better on the data trained than the testing data. Although relatively small, the difference shows how well the model responds to new data.

Coefficients:

Flag: Coefficients debt to income ratio, and delinquencies show to be significant predictors of loan default. Credit line age has a positive effect but lesser compared to the others.

Loss: The model found delinquencies, and debt to income ratios having high positive coefficients indicating a higher predicted loss. Age of credit line has a substantial negative impact suggesting lower loss amount.

STEP-WISE SELECTION:

Findings show a classification accuracy on the training data set at 88.0% and an accuracy on the testing set at 86.7%. In addition, RMSE accuracy for the training data set is 6926.15 compared to the testing data set at .31. I am unclear on why these values seems to be irregularly high and low. There could be an issue in scaling, however I am unsure how to address. I followed the lectures and videos in great detail.

Coefficients:

Flag: Coefficients debt to income ratio, and delinquencies show to be significant predictors of loan default.

Loss: The model's findings are unclear on the effect on predicted loss value. The coefficients are small and show minor changes in prediction. Thus, suggesting there is a limited impact from these variables.

CONCLUSION

In sum, the models performed well. The Gradient Boosting models proved to be most accurate with a higher classification accuracy. It proves to handle complex relationships with accuracy. This would be the model I chose to put into production compared to the others. There were some coefficients that did not make much sense during model performance testing, such as the questionable RMSE values for Stepwise testing.