Assignment 4 – MSDS 422 Jamia Russell

The goal of this assignment is to use two different types of regression: Linear and Logistic. We will use Logistic regression to determine the probability of a crash. Linear regression will be used to calculate the damages assuming that a crash occurs.

SEE ATTACHED PDF WITH PYTHON CODE, GRAPHS AND STATISTICAL OUTPUT

TENSOR FLOW – LOAN DEFAULTS

The findings show a classification accuracy on the training data set at 92.1% and an accuracy on the testing set at 89.6%. This indicates that the model performed well on the training data set and slightly lower on the testing data set suggesting that the model generalizes unseen data.

TENSOR FLOW – LOSS GIVEN DEFAULT

Classification:

Random Forest (RF): 91.61%

Gradient Boosting (GB): 90.44%

REG_STEPWISE: 90.44%

REG_ALL: 88.00%

REG_RF: 87.92%

REG_TREE: 87.58%

Decision Tree (TREE): 87.58%

REG_GB: 87.50%

Random Forests have the highest classification accuracy performing the best among the models. The model shows that Random Forests is the best choice for predicting the probability of a crash showing consistency and increased performance across the data.

DAMAGES

TREE = 3897.663736539515

RF = 3028.8246823918666

GB = 3077.494290676056

REG_ALL = 4027.4426338655644

REG_TREE = 4255.587745569308

REG_RF = 4243.440594173769

REG_GB = 4268.9622840948505

Random Forest has the lowest damage model accuracy. Gradient Boosting and Decision Trees have a lower damage value performing well. The Random Forest model's performance is the best solution to lower the probability of a loan default with high loss. It would be best to continue with that model. Random Forests also have the advantage of being easy to interpret, immune to outliers, and ability to handle different data types.