# Module 12 Report

For this module, we were provided a dataset of lending activity to predict creditworthiness of borrowers. The dataset consisted of loan size, interest rate, borrower income, debt to income, number of accounts, derogatory marks, total debt, and loan status. Once the data was read in the mean, median, std, min, max points were spread-out over a broad range, so I decided to scale the data. I also performed a corrs function and the correlation between all the datapoints were pretty similar. This meant that there is a lot of multicollinearities between the data points and hence all the features we kept. We could have reduced some due to multicollinearities, but I wanted to see if there were any slight discrepancies. I also executed a .info() function to determine if the data needed to be cleaned in any way. Below is a visual map of multicollinearities between the features.

A graph of different colored squares

Description automatically generated

To analyze the data, we set the loan\_status datapoints as the variable y and the rest as the X variable. We then checked the balance between our target feature of loan\_status by doing a value count of those who defaulted and those who did not. It showed a great imbalance between the two statuses with 75,036 showing as not defaulting and 2500 as defaulting.

At this point the data was split into two groups, one for training and one for test. We used logistic regression model on both groups and compared results. Plotting the training metrics showed a confusion matrix with 277 false positives but most of the training data was identified as a true negative (55,994) and 1700 as true positive so the model was a bit reliable to use if you want to predict creditworthiness of borrowers. Below is the output from the linear regression analysis performed on the training data.

A screenshot of a computer screen

Description automatically generated

For the test data, it was about the same as the training data. There was some overfitting with none of the scores falling below 85% but overall, this model could be used to predict credit worthiness of borrowers. The output for the test data is below:

A screenshot of a computer screen

Description automatically generated

We were asked to re-sample the data and perform regression analysis again to see if there was a difference, but I did not see one. The output was exactly the same which leads me to believe that I coded the re-sampled data wrong OR the data is so closely correlated that the results are the same.

Overall, I would use the linear regression model to determine credit worthiness of borrowers. The f1 scores were high (close to one) meaning that there was excellent precision and recall. This suggests that the model performance is pretty good. The model was not perfect but the true negative and positives seem to be a high enough percentage to rely on it for predictions on who to provide loans to.