Credit Report Analysis

The purpose of this analysis is to train and evaluate a model based on loan risk and identify the credit worthiness of its borrowers. A dataset was used from a peer to peer lending services company. The dataset included the loan size amount, interest rate, borrower income, total debt etc. From this dataset we were analyzing the “loan_status” column, where if the loan_status had a value of 0 it meant the loan was in a healthy standing, and a value of 1 meant the loan was at risk of defaulting or risky. We needed to make a model that would try to predict whether the loan was healthy or at risk of defaulting.

Before processing the data with machine learning we first analyzed the dataset. We found that the data was not scaled and severely imbalanced. To improve training and model performance we scaled the data to put into the linear regression model. Next the columns were separated to exclude loan status in the new scaled dataframe. Where X=df_scaled and Y= loan_status. Then the data set was split to be incorporated into a train-test split. Where the training set is used to teach the model, looking for patterns and relationships, and the testing set puts the model against new or unseen data. We then fit the logistic regression model to the trained data.

Results:

Class 0: Healthy Loan Class

- Train:
  - Precision: 1.00
  - Recall: 0.99
  - F1-score: 1.0
- Test:
  - Precision: 1.00
  - Recall: 0.99
  - F1-score: 1.00

Class 1: Unhealthy Loan Class (Risky)

- Train:
  - Precision: 0.85

  - Recall: 0.98
  - F1-score: 0.91
- Test:
  - Precision: 0.86
  - Recall: 0.98
  - F1-score: 0.92

Summary:

After analysis we found that the linear model performs the best in predicting loan risk. When compared to the random forest predictive model the linear regression performed had a lower false negative report of 15/1001 or a recall of 99%. Also, a high F1-score of 91%-92% indicates a good performance of the model for both precision and recall. It is more important to predict risky loans as these are the loans that are at risk of defaulting. Overall, the model does not show over fitting and has 99% accuracy. I would recommend the linear regression model as it balances both classes well even with the data being imbalanced and having multicollinearity.