STAT4510 Progress Update

In my research, my aim is to build an optimal model to help a company predict customer reliability based on credit score data. I initially tested three classifier models—KNN, LDA, and QDA—and found that KNN outperformed the others. Although smaller K-values can lead to overfitting, my dataset showed that a K-value of 1 yielded the best results, suggesting that relying on the closest data point provides the most accurate predictions. My research also indicated that using the Random Forest model might further enhance prediction accuracy.

To refine my predictive models and gain deeper insights into the data, my research plan includes exploring various classifier models, analyzing feature importance, and conducting unsupervised learning using K-means clustering. The Random Forest importance plot revealed that key variables for predicting customer reliability are Outstanding Debt, Credit Mix, Interest Rate, Delayed from Due Date, and Credit History Age. Outstanding Debt had the highest MeanDecreaseGini, indicating that it plays a pivotal role in reducing impurity in the dataset. Meanwhile, Changed Credit Limit ranked highest for MeanDecreaseAccuracy, suggesting it has the most significant impact on model accuracy when its values are shuffled.

To explore various models, I experimented with the Support Vector Machine, but it did not perform as well as expected. It had a misclassification rate of about 0.33, which was worse than KNN but slightly better than LDA and QDA. Additionally, SVM is computationally expensive when tuning parameters like cost and gamma, so I turned to tree-based models due to their robustness with complex data. I first tried a pruned classification tree, but it didn't meet expectations. Subsequently, I used Random Forest, which had the lowest misclassification rate among all models, validating its reputation as a robust classifier.

I also used K-Means Clustering with K=3 to understand the data's structure. The results indicated that most data points fell into the second cluster, suggesting a relatively balanced distribution among the clusters. When clustering by age and annual income, the three clusters showed distinct characteristics, indicating that these variables might play a significant role in credit score predictions, aligning with common expectations.

These findings underscore that Outstanding Debt is critical for reducing impurity in my model, while Changed Credit Limit is essential for maintaining prediction accuracy. By focusing on these influential variables, I can build a more reliable model for predicting customer reliability. This approach not only strengthens my current research but also provides valuable insights for guiding the company's future creditworthiness assessment strategies.