Final Report

By Yuxuan Hu

Summary

To find a better model with higher predicting power in finding companies with higher chance of defaulting, this project employed the Linear Probability Model (LPM), K-nearest neighbor (KNN), and Logistic Regression Model, using the method of cross-validation and forming the confusion matrix and ROC curve to find the model which gives the lowest error rate in prediction.

Model Selections

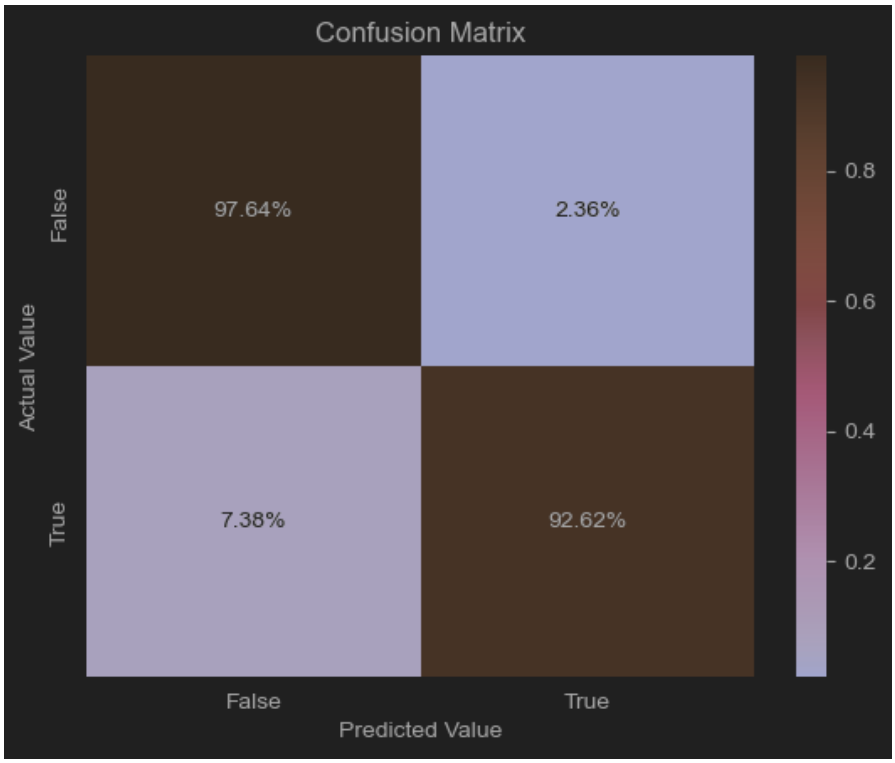
LPM

Setting the threshold at 0.5 the overall error rate of 9.5%; the false negative rate is 18.24%

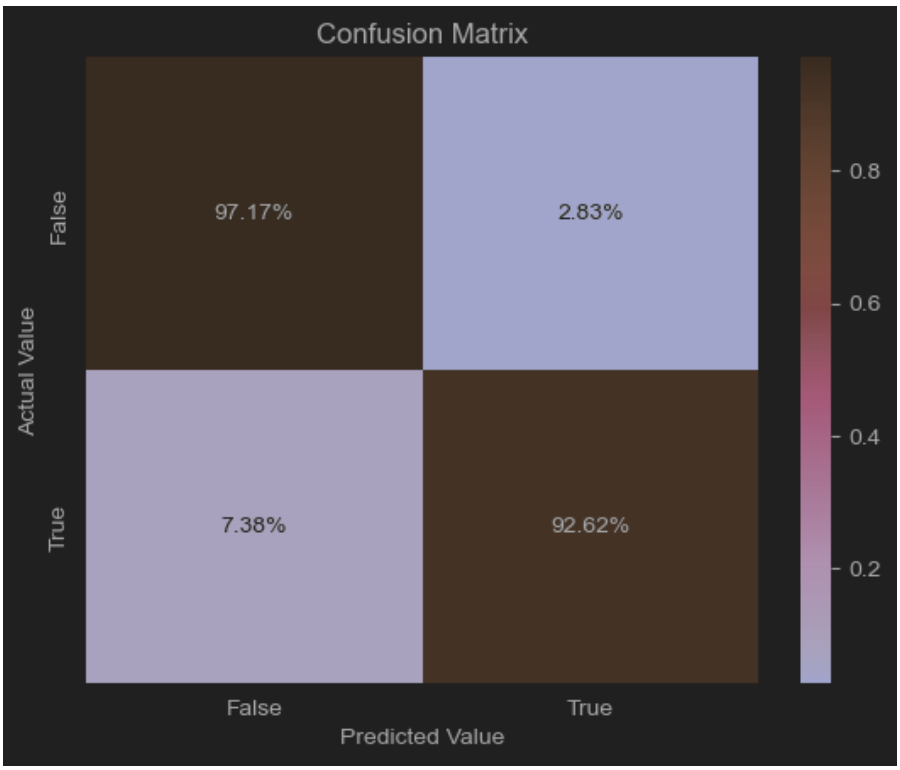
Setting the threshold at 0.6 the overall error rate is 11.08%; the false negative rate is 24.53%

A False Positive is when someone does not evade tax but is predicted as cheated. False Negative is when someone evades the tax but is predicted as not evading the tax. In this context, **False Negative is more important since we want to miss as few firms that cheated as possible.**

Setting the threshold of predicted probability of tax evasion will affect the false negative rate. When we set a lower value of the threshold, we categorize firms with a lower probability of cheating to be predicted as potential tax evaders. In this way, we have a lower false negative rate and a higher false positive rate. We would like to choose a lower threshold, a threshold at 0.5, to achieve a lower false negative rate. This model will provide a more accurate prediction for tax evasion.



(KNN model without standardization)



(KNN model with standardization)

KNN Model

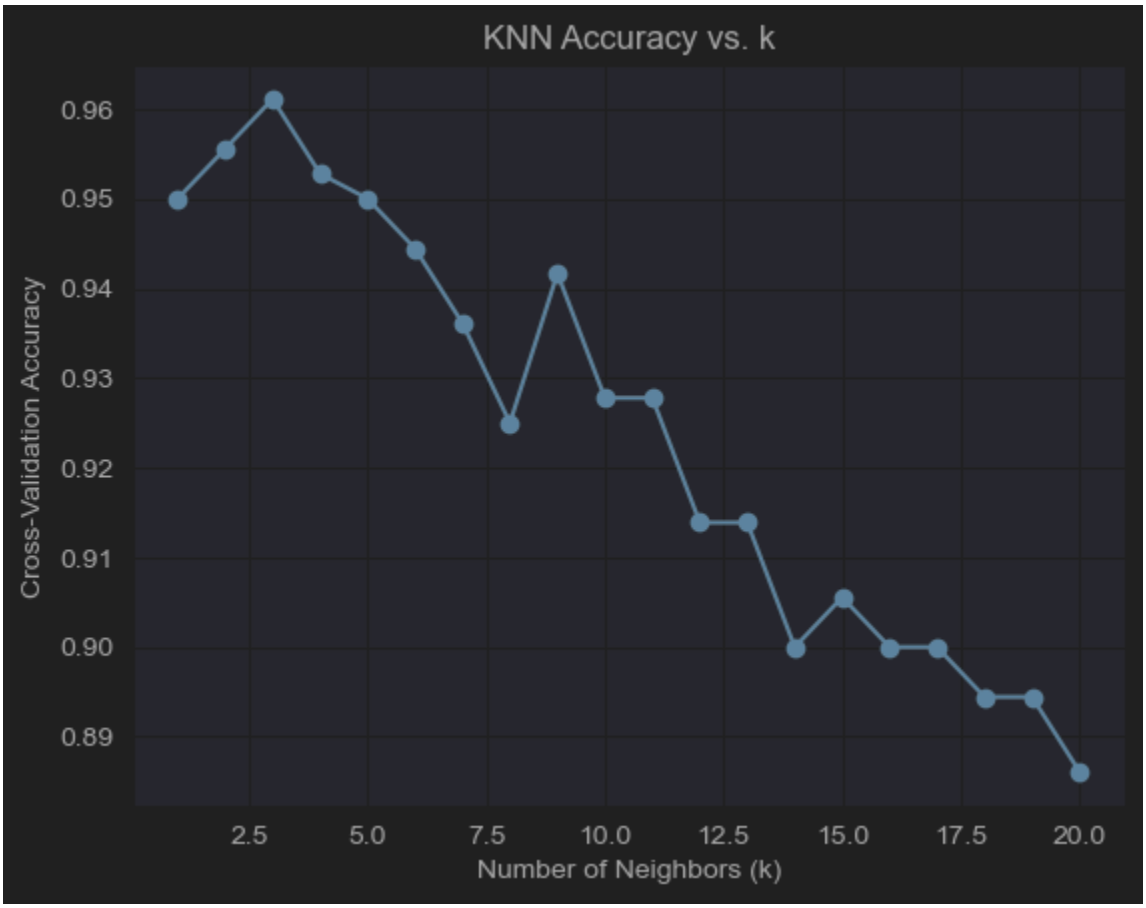
We use the KNN model with a non-standard scaler and with a standard scaler to find the model with the highest accuracy.

With Standard Scaler, the overall error rate is 4.7%; the false negative rate is 7.38%.

Without a Standard Scaler, the overall error rate is 4.4%; the false negative rate is 7.38%

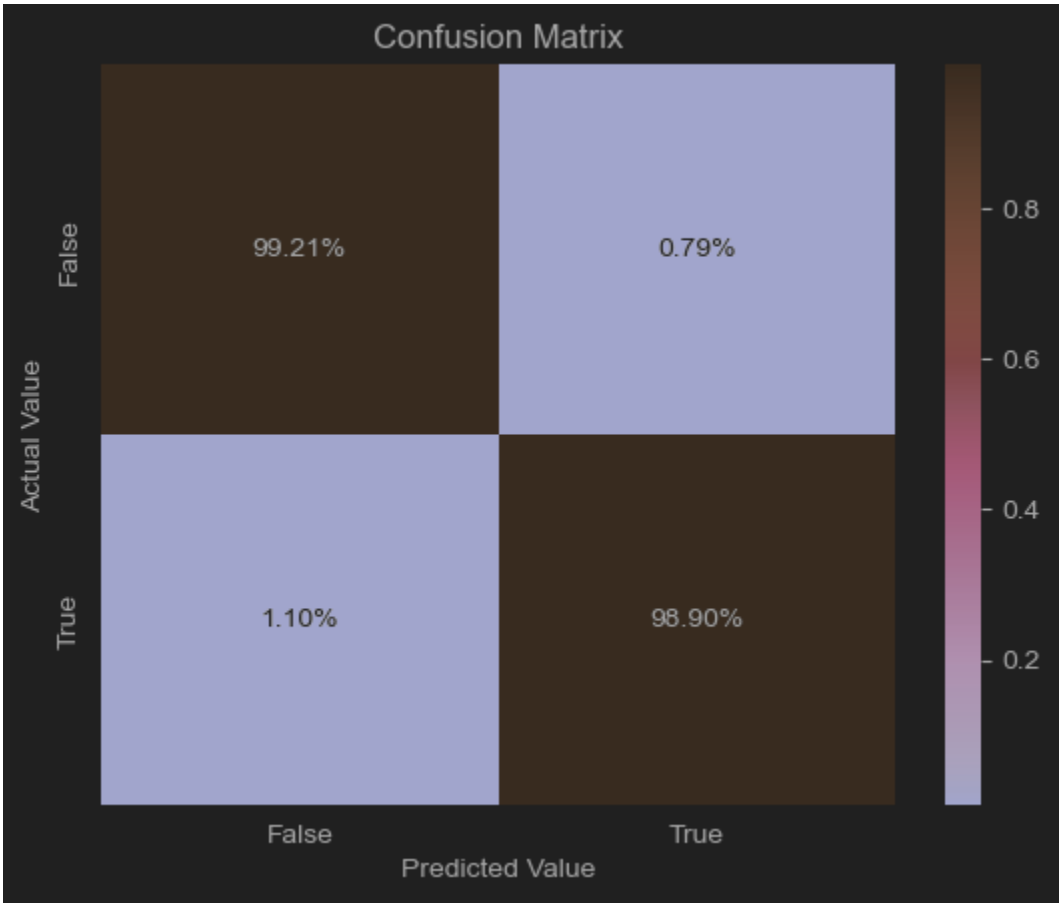
The non-scaled model gives a better prediction.

After tuning of k values, k=3 gives the most accurate model with the lowest overall error rate at 3.6%.



Logistic Regression Model

The logistic regression model gives an overall error rate of 0.0092 with a false negative rate of 1.1%.



In conclusion, the Logistic Regression model is the model with the lowest error rate and false negative rate.

However, when making further predictions, one major problem is selection bias. The data only includes firms that were audited. We don't know the performance of firms that were not audited. If the government relied on this model, it may lead to audits being concentrated on firms that resemble past audited firms, even if they are not fraudulent. Also, if there's a new fraud pattern, the model may fail to detect it.