Fraudulent Transaction Prediction

Purpose of the Project:

The purpose of this project is to predict if the financial transactions are fraudulent or genuine based on given transaction features.

DataSet:

The dataset available for this project contained credit card transaction data with 650K rows and 10 columns. The columns contain the information that included the type of transaction, amount, balance before transfer and balance after transfer on both ends. There was also a column with the target variable that classified the transaction either as fraud or not Fraud.

Data Preparation and Feature Engineering:

We looked at the data and realized that the data was pretty clean but there were some columns that were not particularly important for the project. WE dropped the name code for origination and destination. We check for the correlation between different columns to gauge the importance of different features.

We further checked for multicollinearity using the VIF score and ignored columns that had a VIF score of more than 12. This resulted in 11 total relevant features for our model building. Furthermore the data had target features heavily skewed, at a ratio of almost 1000:1. We balanced the data using a resampling function.

Modeling:

We considered the following 5 different machine learning models and calculated their AUC score to determine which would fare out best.

Model	**AUC Score**
Logistic Regression	0.6162
Random Forest	0.9599
XG Boost	0.9962
Decision Tree	0.8856

Naive Bayes	0.7291

Out of the 5, we picked two models, Random Forest and XGBoost which had the two highest AUC scores. We then determined the most important features that would predict if the transactions were fraudulent or not.

Conclusions and Recommendation:

The feature importances tool showed that Random Forest used the amount in the transaction to be the most important indicator if the transaction was fraudulent with the transaction type ‘transfer’ and ‘cashout’ following closely. The XGBoost had ‘transfer ’as the most important indicator with ‘cash out’ and amount following closely.

Upon running SHAP analysis on these two models, both agreed transaction type transfer and cash out had the highest impact on the classification as fraudulent. This indicates and would agree with general consensus that if someone were to initiate a fraudulent transaction they would either want to transfer the money to the account where they could access later or cash out the money.

So the companies should be aware and implement a second factor authentication for transactions like transfer and cash out.