Fraud Detection Project

# Summary

We were tasked with identifying fraud, waste, and abuse within the payment streams for a large financial institution. We were presented with a dataset of 125,000 records of transactions and a holdout set to test the predictive power of our best model. We trained and evaluated three different models: Logistic Regression, Random Forest, and Gradient Boosting Machine.

# Data Exploration and Preprocessing

## Exploratory Data Analysis

Upon receiving the dataset, we began investigating if there were any null values in the data, and there was [[Appendix A](#_Appendix_A)]. Then, we attempted to understand the components of the dataset such as which features are numerical and categorical data types. We found that the dataset contains various numerical columns such the **age of the account** in days, **transaction amount** in dollars, the **adjusted dollar value** of the transaction, measure of the **historic USD amount used to purchase goods and services**, and the **amount of the first transaction** in dollars.

We investigated these numerical features by seeing the correlation they had with one another [[Appendix B](#_Appendix_B)]. Also, we sought to see these numerical features against the target variable of is it legit or fraud using histograms [[Appendix C](#_Appendix_C)]. The output showed that much of the data was normally distributed with very few outliers. To assure ourselves that the data did not hold extreme outliers we created box plots for each numerical data against the target which confirmed what our histograms told us [[Appendix D](#_Appendix_D)].

You may have noticed that zip code and ‘card\_bin’, which is the first 6 digits of a credit card are included. However, we will not be utilizing any of these features because they have high cardinality. This means that there are a lot of unique values associated with zip codes and encoding those unique values would affect our model and take up a lot of computational time. That said, many of the categorical columns also contain identifiers like IP addresses, billing addresses, etc. which we will not use to avoid further problems.

## Data Processing

Since we know that our dataset has various missing values, we decided to address this issue by creating pipelines which automate processes like imputing and scaling. For numerical features we decided to use a simple imputer to insert the average of a column to any row that has a null value and a scaler to standardize the features in the data to begin the modeling process.

For categorical variables, we decided to use a simple imputer to insert the most frequent occurrence on the column for any row in that column that has a null value and a one hot encoder to change the categorical variables into binary vectors of 0 and 1. Lastly, we converted the target variable to 0 and 1 where 0 is legit and 1 is fraud.

# Model Development

## Model Training

We are training three models: logistic regression which will be used as the baseline for performance model comparison, a random forest, and a gradient boosting machine model.

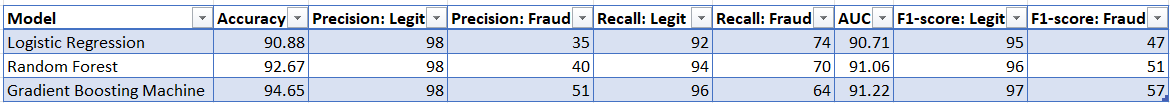
By its feature randomness, a random forest model can produce many diverse trees (models) and then combines the trees to produce higher performance over any single tree. Now with gradient boosting, it is still developing trees, but not at random. Instead, gradient boosting is learning sequentially to keep improving the trees it produces.

For all models except for logistic regression, we conducted a grid search with a list of possible parameters to find the best model based on the highest area under the curve score. We are using an AUC scoring because it visually represents the true positive rate against the false positive rate. This means that it is measuring the instances where the model incorrectly predicts fraud versus the instances where the model correctly predicts fraud.

## Parameter Tuning

After we conducted a grid search, we found the best parameters for a random tree and a gradient boosting machine and refitted the models.

However, we found an imbalance in the data between the two classes, fraud and legit [[Appendix E](#_Appendix_E)]. Notice that majority of the data, about 94% is labeled as “legit” where only 5.4% is labeled as “fraud.” Since we are dealing with fraud detection, having this kind of imbalance can pose big challenges so we opted to see the class distribution using various techniques like over-sampling, under sampling, and the synthetic minority oversampling technique [[Appendix F](#_Appendix_F)]. We found that the synthetic minority oversampling technique did much better and refitted our models on the new sample data [[Appendix G](#_Appendix_G)].



Above is the table which includes the three models we trained. We are measuring these models based on accuracy, the ratio of correctly predicted instances to the total instances. Precision which is the ratio of correctly predicted fraud cases to the total predicted fraud cases. Recall which is a measure of how many of the actual positive instances are correctly predicted. The AUC (Area Under the Curve) provides a single scalar value that represents the model's ability to discriminate between positive and negative instances. A higher AUC indicates better discrimination. Finally, F1-score is the average of precision and recall, providing a balance. So, for example, “precision: legit” means that out of all the instances predicted as “fraud,” only that percentage was actual fraud. For recall, it shows the ratio of true positive predictions to the total actual positives meaning that the model only captured X percentage of all actual between the two.

At first sight, the GBM model is outperforming all models in accuracy and AUC [[Appendix H](#_Appendix_H)]. However, it is important to note that all three models do extremely well at identifying legit cases but do poorly in detecting fraud cases except for the GBM model which has a higher F1 score for fraud cases than any other model. This means that there is a better balance between precision and recall when detecting fraud cases in the GBM model than logistic and random forest. We ran a confusion matrix to assess how well the model can make accurate predictions and when it missed instances of fraud [[Appendix I](#_Appendix_I)]. A confusion matrix compares the predicted classes of a model with the actual classes in a dataset. The matrix typically consists of four values: true positive (correctly predicted positive instances), true negative (correctly predicted negative instances), false positive (incorrectly predicted as positive), and false negative (incorrectly predicted as negative). These values help assess the model's accuracy, precision, recall, and other performance metrics. Here, our model was able to detect 34,171 fraudulent transactions, but missed 1,269 transactions that were fraudulent.

We further inspected the model by checking which features it thought were important. We ran a partial dependence plot to provide insights into the relationship between a feature and the target prediction. A flat line indicates that the target prediction is not sensitive to changes in the feature value. A sloped line, either negative or positive, indicates that the target prediction is affected by the feature. Finally, non-linear relationships like curves indicate a more complex relationship between the feature and the target prediction. As one can see, account age, transaction amount, adjusted transaction amount and historic velocity all affect the target prediction and are the top features for this model [[Appendix J](#_Appendix_J)].

## Insights and Recommendations

Let us say we want to operate at a 5% false positive rate, meaning that on average, 5% of the instances predicted will be false positives. Operating at a 5% FPR helps strike a balance between identifying fraud and minimizing false positives like incorrectly flagging a legitimate transaction as fraud which can lead to customer dissatisfaction and unnecessary costs to the company.

Using our GBM model, to operate at 5% FPR means that the model will only correctly detect approximately 49% of cases [[Appendix K](#_Appendix_K)]. By this model only catching 49% of legit transactions, the model is not effectively identifying a larger portion of legitimate transactions. This could still be due to the imbalances in the majority class so further sampling techniques can help with that. Operating at a 5% FRP means that you would have to have a threshold of 96%. By having a threshold of 96%, it signifies that any predicted probability above this threshold will be classified as “fraud” and anything below is “legit.” Operating at a 5% FPR involves a trade-off with Recall because by limiting false positives, there might be cases of actual fraud that the model misses.

Adjusting the threshold to achieve a lower FPR might lead to a decrease in recall and vice versa. For future consideration it is best to monitor and evaluate the model's performance to ensure it continues to meet the desired goals of the company.

# Appendix

## Appendix A

A screenshot of a computer screen

Description automatically generated

## Appendix B

A screenshot of a graph

Description automatically generated

## Appendix C

A graph of a normal distribution

Description automatically generated with medium confidence

A graph of a function with Ryugyong Hotel in the background

Description automatically generated with medium confidence

A graph of a post and histogram of a post

Description automatically generated

A graph of two people

Description automatically generated with medium confidence

## Appendix D

A graph of blue squares

Description automatically generated with medium confidence

A graph of a diagram

Description automatically generated with medium confidence

A diagram of a box and a box

Description automatically generated

A graph of blue squares

Description automatically generated with medium confidence

## Appendix E



A graph of a class distribution

Description automatically generated

## Appendix F

A graph of a number of blue squares

Description automatically generated with medium confidence

## Appendix G



## Appendix H

A graph of a logistic curve

Description automatically generated

## Appendix I

A graph of a diagram

Description automatically generated with medium confidence

A black and white text

Description automatically generated

## Appendix J

A graph of data set up

Description automatically generated with medium confidence

## Appendix K

A screenshot of a graph

Description automatically generated