AUTHOR: UTTAM BISWAS

# Fraud Detection Model :

## 1. Data cleaning including missing values, outliers and multi-collinearity.

Cleaning and pre-processing the data is a crucial step in preparing it for machine learning. Here's how we can handle missing values, outliers, and multi-collinearity in your dataset:

- Missing Values: Identify where data is missing. Fill missing values with appropriate strategies, like mean for numbers or 'Unknown' for categories.

- Outliers: Find outliers in numerical features. Decide whether to remove, transform, or keep them based on their validity and impact.

- Multi-collinearity: Calculate correlations between numerical features. Consider removing one of highly correlated variables or use dimensionality reduction techniques like PCA.

These steps help prepare your data for machine learning by addressing missing values, outliers, and multi-collinearity.

## 2.Describe your fraud detection model in elaboration.

The fraud detection model used in this case is an XGBoost classifier. XGBoost is an ensemble learning technique based on decision trees, known for its high predictive power and ability to handle imbalanced datasets. Here's an elaboration on the model:

- Algorithm Choice: XGBoost was chosen because it is a robust algorithm capable of handling imbalanced datasets and capturing complex relationships in the data.

- Ensemble Approach: XGBoost is an ensemble method that combines multiple decision trees to make predictions. This ensemble approach helps improve model accuracy and generalization.

- Model Interpretability: XGBoost provides feature importance, allowing us to understand which features are most important for making predictions.

### 3. How did you select variables to be included in the model?

Variable selection is crucial for model performance. In this case, the following steps were likely taken:

- Domain Knowledge: Variables relevant to the financial domain and fraud detection were likely included. This might include transaction amount, type, and balance information.

- Feature Engineering: Additional features might have been engineered to capture relevant information, such as the difference between old and new balances.

- Correlation Analysis: Correlation analysis may have been performed to identify highly correlated features, and only one of each correlated pair would be included to reduce multi-collinearity.

- Feature Importance: The feature importance from the XGBoost model would have guided the selection of variables. Features with higher importance scores are more likely to be included in the model.

### 4. Demonstrate the performance of the model by using the best set of tools.

The code provided in the python file demonstrates the performance of the XGBoost model using standard tools. Model can be evaluated using various tools like visualisation and by using various kinds of metrics. In our code, we have evaluated it using confusion matrix, classification report and AUC_ROC score. Let's look into how the code proceeds and how we evaluated it:

First of all, we loaded the data, pre-processed it, and split it into training and validation sets. Then, we trained an XGBoost model on the training data. We made predictions on the validation data.

We evaluated key performance metrics such as:

- Confusion Matrix: This shows the number of true positives, true negatives, false positives, and false negatives.
- Classification Report: This provides precision, recall, F1-score, and support for each class.
- AUC-ROC Score: This measures the area under the ROC curve, which is useful for binary classification tasks.

### 5. What are the key factors that predict fraudulent customer?

The key factors that predict fraudulent customers are likely derived from the feature importance of the XGBoost model. These factors might include:

- Transaction Amount: Large or unusual transaction amounts could be indicative of fraud.
- Transaction Type: Certain transaction types may be more susceptible to fraud such as transfer or cash out.
- Balance Changes: Sudden and significant changes in account balances could raise suspicion.
- Historical Behaviour: Patterns of past transactions and behaviour may also be important.

### 6. Do these factors make sense? If yes, How? If not, How not?

The factors mentioned do make sense in the context of fraud detection:

- Transaction Amount: Fraudulent activities often involve transferring large sums of money.
- Transaction Type: Certain types of transactions, such as transfers and cash-outs, are more likely to be associated with fraud.
- Balance Changes: Sudden balance changes may indicate unauthorized access to an account.
- Historical Behaviour: Understanding a customer's historical behaviour helps identify deviations from the norm.

These factors align with common patterns of fraudulent behaviour, making them sensible for inclusion in the model**.**

### 7. What kind of prevention should be adopted while the company updates its infrastructure?

Prevention measures can include:

- Two-Factor Authentication: Implementing two-factor authentication for certain transactions.
- Transaction Limits: Setting transaction limits, especially for high-risk transactions.
- Real-time Monitoring: Using real-time monitoring systems to detect anomalies.
- Machine Learning Models: Continuously updating and improving machine learning models for fraud detection.
- User Education: Educating users about safe online practices and how to recognize phishing attempts.

### 8. Assuming these actions have been implemented, how would you determine if they work?

To determine the effectiveness of prevention measures we can do the following tasks:

- Monitoring: Continuously monitor key metrics, such as the number of fraudulent transactions, false positives, and financial impact.
- A/B Testing: Implement measures in a controlled manner, perhaps applying them to a subset of users or transactions, and compare the results with a control group.
- Feedback Loop: Maintain a feedback loop with fraud analysts and experts to gather insights and make adjustments.
- Model Performance: Regularly assess the performance of your fraud detection model to ensure it adapts to new fraud patterns.

The effectiveness of prevention measures should be assessed over time, and adjustments should be made as necessary to stay ahead of evolving fraud tactics.