FRAUD DETECTION ASSIGNMENT

1. Data Cleaning:

- **Missing Values**: I identified and handled missing values by applying imputation techniques or removal where necessary, ensuring a clean dataset for model training.
- **Outliers**: I detected outliers using statistical methods (Z-scores and box plots) and addressed them appropriately to avoid skewed results.
- **Multi-Collinearity**: I evaluated the features for multi-collinearity using correlation matrices and Variance Inflation Factor (VIF). Features with high correlation were removed to improve model performance.

2. Model Development:

I implemented four machine learning models for fraud detection: Random Forest, Gradient Boosting, XGBoost, and Logistic Regression. These models were selected based on their suitability for classification tasks, particularly with imbalanced datasets.

- **Model Training**: I split the data into calibration and validation sets and fine-tuned each model to ensure optimal performance. The models were trained on various features to predict fraudulent transactions.
- **Evaluation Metrics**: I assessed the models using accuracy, precision, recall, F1-score and ROC AUC. The confusion matrices for each model were generated to assess how well they identified fraud.

3. Model Results:

Random Forest Model:

- Accuracy: 1.00
- ROC AUC Score: 0.999999982
- Precision and Recall: Both 1.00 for class 0 and class 1, indicating perfect classification.

Gradient Boosting Model:

- Accuracy: 1.00
- ROC AUC Score: 0.999966600
- Precision and Recall: Both 1.00 for class 0 and class 1, indicating excellent performance.

XGBoost Model:

- Accuracy: 1.00
- ROC AUC Score: 0.9999996317
- Precision and Recall: Both 1.00 for class 0 and class 1, showing flawless classification.

Logistic Regression Model:

- Accuracy: 1.00
- ROC AUC Score: 0.9998377
- Precision and Recall: Both 1.00 for class 0 and class 1, confirming the model’s accuracy.

4. Key Factors Predicting Fraudulent Transactions:

1. The models identified factors such as transaction amount, frequency of transactions, and transaction location as critical predictors for identifying fraud.
2. These factors align with known fraud detection patterns, where anomalous behaviors and high-value transactions are typically red flags for fraudulent activities.

5. Recommendations:

- Based on the model results, I recommend improving the fraud detection system by integrating real-time monitoring of high-risk transactions and utilizing two-factor authentication for transactions over a certain threshold.
- Preventive Measures: Enhancing security protocols and updating infrastructure to detect anomalous patterns in real-time would help further mitigate fraud risks.

6. Evaluation of Effectiveness:

The effectiveness of the fraud prevention actions can be evaluated by measuring reduced fraud incidents over time, improved detection accuracy and user feedback. I recommend conducting regular A/B tests to validate the system's performance post-implementation.

I have attached the Jupyter notebook containing the code for the data cleaning, model development, and evaluation steps, as well as the confusion matrices and detailed performance metrics for each model.