Problem Overview

The goal of this project is to build a smart model that can accurately detect fraudulent credit card transactions.
One of the biggest challenges here is that fraud cases are very rare compared to genuine transactions. This kind of class imbalance often causes normal machine learning models to favor the majority class (legitimate transactions), making it hard to catch the rare fraud cases.

Why is this important?
Because effective fraud detection helps protect both financial institutions and cardholders from heavy financial losses.

Preprocessing Steps

Before training any models, a few important data preparation steps were done:

1. Handling Missing Values
   First, the data was checked for any missing values. Any incomplete rows were removed to ensure clean data.
2. Addressing Class Imbalance
   To fix the problem of too few fraud cases, the Synthetic Minority Over-sampling Technique (SMOTE) was considered.
   It creates artificial examples of fraud transactions to balance the classes.
   Note: Although SMOTE was initially applied, it was later commented out because the project shifted to a time-based data split instead.
3. Time-Based Data Splitting
   Since credit card transactions happen over time, the dataset was sorted by the 'Time' feature and split into training (first 80%) and testing (last 20%) sets.
   This method mimics real-world conditions — the model trains on past data and predicts on future unseen data.
4. Feature Scaling
   The features were standardized using StandardScaler so they all have a mean of 0 and standard deviation of 1.
   This helps many machine learning algorithms perform better.

Models Trained and Compared

Three different models were trained:

1. Random Forest Classifier
   - Builds multiple decision trees and combines their predictions.
2. Gradient Boosting Classifier
   - Builds trees sequentially, with each one trying to correct the mistakes of the previous ones.
3. XGBoost Classifier
   - An advanced and optimized version of Gradient Boosting, known for its speed and accuracy.

Evaluation focused mainly on the Random Forest model, and here’s how it performed:

- F1 Score: 0.8125
- ROC AUC Score: 0.9444

The Precision-Recall curve shows that the model starts with high precision at low recall, which then slowly declines — which is typical but desirable in fraud detection.

In simple terms:

- A high ROC AUC score means the model is good at telling fraud apart from genuine transactions.

- A strong F1 score indicates a good balance between catching fraud (recall) and avoiding false alarms (precision).

So far, the Random Forest model looks very promising!

Conclusion and Recommendations

What we know:

- Random Forest performed well, with high scores for both ROC AUC and F1.
- The model handles the imbalanced data fairly well and has potential for real-world use.

What still needs to be done:

1. Evaluate All Models
   We should also evaluate the Gradient Boosting and XGBoost models using the same metrics — F1, ROC AUC, Precision-Recall curves — to ensure Random Forest is truly the best.
2. Hyperparameter Tuning
   Use techniques like GridSearchCV to find the best settings for each model and possibly boost their performance further.
3. Threshold Optimization
   Find the best probability cutoff value that gives the right balance between missing fraud and flagging genuine users incorrectly.
4. Explore More Metrics
   Dive deeper into confusion matrices and classification reports to better understand how often the model gets things right or wrong.
5. Feature Importance Analysis
   Look at the SHAP summary plots to see which features are most important for fraud detection.
   This can guide future improvements in data preprocessing and feature selection.
6. Feature Engineering
   Think about creating new features, like "number of transactions in the last hour" or "average transaction amount," to help the model catch subtle fraud patterns.
7. Deployment and Monitoring
   Once the best model is ready, deploy it into a real system — and keep monitoring and retraining it regularly, because fraud patterns change over time.