## Project Description

This project focuses on developing and comparing multiple machine learning models to detect fraudulent transactions. The primary objectives are to accurately identify fraudulent activities while minimizing false positives, ensuring a balance between precision and recall. The project encompasses data preprocessing, model training, evaluation, and preparation for deployment.

**Key Components:**

1. **Data Understanding and Preparation:**
    - Loading and exploring the dataset.
    - Handling missing values and outliers.
    - Feature scaling and encoding.
    - Addressing class imbalance using SMOTE.
2. **Model Training and Evaluation:**
    - Training various classifiers: Random Forest, LightGBM, XGBoost, MLPClassifier, and Logistic Regression.
    - Evaluating models based on ROC-AUC, GINI coefficient, Precision, Recall, F1-Score, and computational speed.
3. **Model Comparison and Selection:**
    - Comparing models using consistent metrics.
    - Selecting the best-performing model for deployment.

## Dataset

The dataset (Credit Card Fraud Detection) contains transactions made by credit cards in **September 2013** by European cardholders. This dataset presents transactions that occurred in two days, where we have **492 frauds** out of **284,807 transactions**. The dataset is **highly unbalanced**, the **positive class (frauds)** account for **0.172%** of all transactions.
It contains only numerical input variables which are the result of a **PCA transformation**.
Due to confidentiality issues, there are not provided the original features and more background information about the data.

- Features **V1**, **V2**, ... **V28** are the **principal components** obtained with **PCA**;
- The only features which have not been transformed with PCA are **Time** and **Amount**. Feature **Time** contains the seconds elapsed between each transaction and the first transaction in the dataset. The feature **Amount** is the transaction Amount, this feature can be used for example-dependant cost-sensitive learning.
- Feature **Class** is the response variable and it takes value **1** in case of fraud and **0** otherwise.

## Key Observations

**Count:**

- **Non-Fraudulent (Class 0)**: 284,315 transactions.
- **Fraudulent (Class 1)**: 492 transactions.
    - Fraudulent transactions make up a very small fraction of the dataset, highlighting a significant class imbalance.

**Mean (Average Transaction Amount):**

- **Non-Fraudulent (Class 0)**: $88.29
- **Fraudulent (Class 1)**: $122.21
    - Fraudulent transactions have a slightly higher average value compared to non-fraudulent ones.
    - This could indicate that fraudsters often target transactions of larger amounts, but the difference is not drastic.

**Standard Deviation (std):**

- **Non-Fraudulent (Class 0)**: $250.11
- **Fraudulent (Class 1)**: $256.68
    - Both classes have high variability in transaction amounts.
    - Fraudulent transactions show slightly higher variability, suggesting that fraud involves a mix of both small and large transaction amounts.

**Minimum (min):**

- Both classes have a minimum transaction amount of $0.00.
    - These could represent:
        - Zero-value test transactions.
        - Refunds or system errors.

**Percentiles:**

- **25% (1st Quartile)**:
    - **Non-Fraudulent**: $5.65
    - **Fraudulent**: $1.00
        - A significant number of fraudulent transactions have very small amounts, possibly indicating "test transactions" by fraudsters to check the validity of cards.
- **50% (Median)**:
    - **Non-Fraudulent**: $22.00
    - **Fraudulent**: $9.25
        - The median for fraudulent transactions is significantly lower than for non-fraudulent ones, suggesting a tendency for smaller-value fraud.
- **75% (3rd Quartile)**:
    - **Non-Fraudulent**: $77.05
    - **Fraudulent**: $105.89
        - The upper quartile for fraudulent transactions is noticeably higher, showing that fraudsters often conduct transactions in the higher-value range.

**Maximum (max):**

- **Non-Fraudulent (Class 0)**: $25,691.16
- **Fraudulent (Class 1)**: $2,125.87
    - Non-fraudulent transactions have significantly higher maximum values, likely due to large legitimate purchases.
    - Fraudulent transactions are capped at $2,125.87, which may reflect fraud detection systems or fraudsters avoiding excessively high-value transactions to minimize suspicion.

## Model Performance Comparison for Fraud Detection

| Model | ROC-AUC | GINI | Precision_Fraud | Recall_Fraud | F1_Fraud | Train_Time (s) | Pred_Time (s) |
|---|---|---|---|---|---|---|---|
| Random Forest | 0.9561 | 0.9122 | 0.8539 | 0.7677 | 0.8085 | 38.83 | 0.0927 |
| LightGBM | 0.9649 | 0.9297 | 0.6061 | 0.8081 | 0.6926 | 2.02 | 0.0650 |
| XGBoost | 0.9668 | 0.9337 | 0.7879 | 0.7879 | 0.7879 | 1.81 | 0.0316 |
| MLPClassifier | 0.9440 | 0.8881 | 0.7701 | 0.6768 | 0.7204 | 29.45 | 0.0362 |
| Logistic Regression | 0.9718 | 0.9436 | 0.0558 | 0.8788 | 0.1049 | 3.30 | 0.0055 |

## Analysis of Results

**1. Random Forest**

- **Strengths:**
  - **High Precision (0.8539):** When predicting fraud, it is correct approximately 85.4% of the time, minimizing false positives.
  - **Good ROC-AUC (0.9561) and GINI (0.9122):** Demonstrates strong overall discriminative ability.
- **Weaknesses:**
  - **Long Training Time (≈38.83s):** Slower compared to other models, which might be a concern for scalability.
  - **Moderate Recall (0.7677):** Misses about 23.2% of actual fraud cases.
- **Implications:** Suitable for scenarios where reducing false positives is critical, even if it means missing some fraudulent transactions.

**2. LightGBM**

- **Strengths:**
  - **Strong ROC-AUC (0.9649) and GINI (0.9297):** Excellent discriminative performance.
  - **Balanced Recall (0.8081):** Captures a significant portion of fraud cases.
  - **Fast Training (≈2.02s) and Prediction Time (≈0.0650s):** Efficient for both development and deployment.
- **Weaknesses:**
  - **Moderate Precision (0.6061):** Higher false positives, leading to potential customer inconvenience.
  - **Lower F1-Score (0.6926):** Indicates room for improvement in balancing precision and recall.
- **Implications:** Ideal for environments where catching more fraud cases is prioritized, even at the expense of increased false alarms.

**3. XGBoost**

- **Strengths:**
  - **Highest ROC-AUC (0.9668) and GINI (0.9337):** Top-tier discriminative power among the evaluated models.
  - **Balanced Precision (0.7879) and Recall (0.7879):** Offers a good trade-off between minimizing false positives and false negatives.
  - **Fast Training (≈1.81s) and Prediction Time (≈0.0316s):** Highly efficient, making it suitable for real-time applications.
- **Weaknesses:**
  - **No Significant Weaknesses Identified:** Exhibits balanced performance across all metrics.
- **Implications: XGBoost** emerges as the best-performing model, providing robust performance metrics suitable for effective fraud detection.

**4. MLPClassifier**

- **Strengths:**
  - **Good Precision (0.7701):** Reduces false positives compared to some models.
  - **Decent ROC-AUC (0.9440) and GINI (0.8881):** Solid discriminative ability.
- **Weaknesses:**
  - **Lower Recall (0.6768):** Misses approximately 32.3% of actual fraud cases.
  - **Longer Training Time (≈29.45s):** Slower than gradient boosting models.
- **Implications:** While offering a reasonable balance, **MLPClassifier** underperforms compared to tree-based models like XGBoost and LightGBM in this context.

**5. Logistic Regression**

- **Strengths:**
  - **Highest ROC-AUC (0.9718) and GINI (0.9436):** Exceptional overall discriminative capability.
  - **Fast Training (≈3.30s) and Prediction Time (≈0.0055s):** Extremely efficient for deployment.
- **Weaknesses:**
  - **Extremely Low Precision (0.0558):** Predicts fraud with only about 5.6% accuracy, leading to a high number of false positives.
  - **Low F1-Score (0.1049):** Indicates poor balance between precision and recall.
- **Implications:** Despite excellent ROC-AUC and GINI scores, **Logistic Regression** is impractical for fraud detection in its current state due to its inability to accurately predict fraud cases without significant adjustments (e.g., threshold tuning).

## Model Selection Recommendation

**XGBoost** and **LightGBM** are the top performers for a fraud detection task, offering a strong balance between accuracy, efficiency, and balanced metrics. **Random Forest** provides high precision but at the cost of longer training times and moderate recall. **Logistic Regression** and **MLPClassifier** require further refinement to meet the practical demands of fraud detection.

## License