**CREDIT CARD FRAUD DETECTION: DATASET, PREPROCESSING, MODEL PERFORMANCE, AND APPLICATION FEATURES**

**Submitted by Vishnu Chandrasekharan**

**C0924468**

**1. Dataset and Preprocessing Steps**

Dataset Overview:

The dataset used in this project consists of credit card transactions made by European cardholders over a two-day period in September 2013. It contains 284,807 transactions, of which 492 are fraudulent, accounting for approximately 0.172% of the data. Each transaction includes 30 features: 28 anonymized variables resulting from a Principal Component Analysis (PCA) transformation (labeled V1 to V28), along with 'Time' and 'Amount'. The target variable 'Class' indicates the transaction status, where '0' denotes legitimate transactions and '1' denotes fraudulent ones.

Preprocessing Steps:

1. **Feature Scaling:** The 'Time' and 'Amount' features are normalized using the StandardScaler to ensure they contribute proportionately to the model.
2. **Handling Class Imbalance:** Given the significant imbalance between legitimate and fraudulent transactions, the Synthetic Minority Over-sampling Technique (SMOTE) is applied to generate synthetic samples of the minority class, thereby balancing the dataset.

**2. Model Performance Metrics**

The logistic regression model's performance is evaluated using several key metrics:

* **Confusion Matrix:** This matrix provides a summary of prediction results, detailing true positives, true negatives, false positives, and false negatives.
* **Classification Report:** This report includes precision, recall, and F1-score for both classes, offering a comprehensive view of the model's performance.
* **ROC-AUC Score:** The Receiver Operating Characteristic - Area Under Curve (ROC-AUC) score measures the model's ability to distinguish between classes, with a higher score indicating better performance.

The classification report for the model is as follows:

| **Class** | **Precision** | **Recall** | **F1-Score** | **Support** |
| --- | --- | --- | --- | --- |
| 0 | 0.95 | 0.99 | 0.97 | 56,777 |
| 1 | 0.98 | 0.90 | 0.94 | 28,518 |

Overall, the model achieves an accuracy of 96%, with a macro average F1-score of 0.95 and a weighted average F1-score of 0.96. The ROC-AUC Score is 0.9884, indicating a high ability to distinguish between fraudulent and legitimate transactions.

**3. Key Features of the Front-End Application**

The user interface, developed using Streamlit, offers the following features:

* **User-Friendly Input:** Users can input transaction details, including 'Time', 'Amount', and PCA-transformed features (V1 to V28), through intuitive input fields.
* **Real-Time Prediction:** Upon entering the transaction details, users can initiate the fraud detection process by clicking the "Predict Fraudulent Transaction" button.
* **Immediate Feedback:** The application provides instant feedback, indicating whether the transaction is likely fraudulent or legitimate, along with the associated probability.

This streamlined interface ensures that users can efficiently assess the fraud risk of transactions with minimal effort.

**Conclusion**

This project demonstrates the application of logistic regression in detecting credit card fraud, addressing challenges such as class imbalance through SMOTE and providing a user-friendly interface for real-time predictions. Future enhancements could involve exploring more advanced machine learning algorithms and incorporating additional features to further improve detection accuracy.