# Project Overview

This project focuses on detecting fraudulent transactions in credit card datasets. The dataset includes transactions made by European cardholders during two consecutive days in September 2013, containing 492 fraudulent transactions out of a total of 284,807. Due to the dataset's highly imbalanced nature, specialized techniques and models are used to enhance prediction accuracy.

# Dataset Details

- **Source**: Kaggle
- **Features**:
  - **Time**: Seconds elapsed between each transaction and the first transaction in the dataset.
  - **Amount**: Transaction amount, useful for cost-sensitive learning.
  - **Class**: Response variable; 1 indicates fraud, 0 indicates non-fraud.
  - **V1-V28**: Principal components obtained via PCA.
- The dataset is preprocessed to address confidentiality issues, with features anonymized.

# Key Explorations

1. **Class Distribution**:
   - Fraudulent transactions constitute only 0.172% of the total, highlighting the imbalance.
2. **Feature Analysis**:
   - `V3`, `V4`, `V10`, `V11`, `V17-V19` exhibit distinct distributions for fraudulent and non-fraudulent transactions.
   - Fraudulent transactions are evenly distributed across time and primarily involve amounts below $2500.
3. **Correlation Analysis**:
   - Minimal correlation between features, except `Amount` showing weak correlation with `V7` and `V20`.

# Data Processing

1. **Normalization**:
   - `Time` and `Amount` are standardized using `StandardScaler`.
2. **Handling Imbalance**:
   - **Under-sampling**: RandomUnderSampler
   - **Over-sampling**: SMOTE (Synthetic Minority Oversampling Technique)
   - **Combination**: SMOTETomek

# Feature Engineering

Top 10 selected features based on ANOVA F-tests: `V17`, `V14`, `V12`, `V10`, `V16`, `V3`, `V7`, `V11`, `V4`, `V18`

# Model Training

### Algorithms Evaluated:

1. Logistic Regression
2. Support Vector Classifier (SVC)
3. Decision Tree
4. Random Forest
5. Bagging Classifier
6. Stochastic Gradient Descent (SGD)
7. Gradient Boosting
8. XGBoost

### Model Performance:

- **Evaluation Metrics**:
  - Precision
  - Recall
  - F1 Score
  - ROC-AUC

- **Best Model**:
  - XGBoost achieved the best results with significant improvement in performance.

### Hyperparameter Tuning:

GridSearchCV was employed to fine-tune the parameters of Random Forest and XGBoost.

# Results and Insights

- Fraudulent transactions are successfully detected with high precision and recall.
- Feature selection and data balancing significantly enhanced model performance.
- XGBoost outperformed other algorithms even without extensive hyperparameter tuning.

# Deployment

- The final XGBoost model was saved as a `fraud_detection_model.pkl` file using `joblib` for deployment.

# Future Work

- Explore deep learning models like neural networks.

- Extend analysis to real-time detection using streaming data.
- Incorporate external features like geolocation and user behavior data.

# Requirements

- **Python Libraries**:
  - pandas, numpy, matplotlib, seaborn, plotly
  - scikit-learn, imbalanced-learn
  - XGBoost, joblib