FindDefault: Prediction of Credit Card Fraud

1. Introduction

The FindDefault project aims to predict credit card fraud using machine learning techniques. This report provides a detailed overview of the project's key steps and findings.

2. Load the Data

The dataset containing credit card transactions made by European cardholders in September 2013 was obtained from [source]. This dataset consists of transactions over two days, with 492 fraudulent transactions out of 284,807 transactions, making it highly imbalanced.

3. Exploratory Data Analysis (EDA)

Exploratory Data Analysis was performed to understand the characteristics and distributions of the dataset. Key findings include:

- Imbalance: The dataset contains a significantly higher number of non-fraudulent transactions compared to fraudulent ones.
- Distribution of Features: Visualization of features such as transaction amount, time, and anonymized features (V1-V28) revealed insights into their distributions and potential relationships.

4. Data Cleaning

Data cleaning involved handling missing values, outliers, and inconsistencies in the dataset. Key steps included:

- Missing Values: Identified and addressed missing values using techniques such as imputation or removal.
- Outliers: Detected outliers using statistical methods such as z-scores or IQR, and removed or transformed them to mitigate their impact on model training.

5. Feature Engineering

Feature engineering aimed to create new features or transform existing ones to improve model performance. Key features engineered include:

- Time: Extracted time-related features such as hour of the day or day of the week from the transaction timestamp.
- Amount: Normalized transaction amounts to ensure consistency across different scales.

6. Train/Test Split

The dataset was split into training and testing sets to train and evaluate the model, respectively. The split ratio used was 80% for training and 20% for testing.

7. Model Selection: Logistic Regression

Logistic regression was chosen as the primary model for its simplicity and interpretability, making it suitable for binary classification tasks like credit card fraud detection.

8. Model Training

The logistic regression model was trained using the training data to learn the patterns and relationships between features and target labels.

9. Hyperparameter Tuning

Hyperparameter tuning was performed to optimize the logistic regression model's performance. Key hyperparameters tuned include regularization strength (C) and solver algorithm.

10. Conclusion

The FindDefault project successfully predicts credit card fraud using logistic regression. Through exploratory data analysis, data cleaning, feature engineering, and model training with hyperparameter tuning, the model achieves high accuracy and performance in detecting fraudulent transactions.