Credit Card Fraud Detection Report

1 Introduction

Credit card fraud detection is a critical aspect of financial security, involving the identification of unauthorized and suspicious transactions to prevent financial losses for both cardholders and financial institutions. With the increasing volume and sophistication of fraudulent activities, traditional detection methods have become inadequate, necessitating advanced approaches leveraging machine learning and data analysis.

This study explores the development and implementation of a machine learning model to detect fraudulent credit card transactions. The primary objectives include analyzing transaction data to identify patterns indicative of fraud, evaluating various machine learning algorithms, and optimizing the chosen model for high accuracy and low false positive rates.

2 Data Collection and Preprocessing

2.1 Data Collection

The dataset used for this study contains transactions made using credit cards. It consists of a large number of legitimate transactions and a small number of fraudulent transactions, making it highly imbalanced. The dataset includes 30 features resulting from a PCA transformation, with a binary target variable indicating whether a transaction is fraudulent (1) or not (0).

2.2 Data Preprocessing

1. **Adequacy of Data:** Checking the number of records and features in each dataset to ensure they are adequate for analysis. The datasets used have the following shapes:
    - Dataset 1: 56,962 records, 30 features
    - Dataset 2: 284,807 records, 31 features
    - Dataset 3: 568,630 records, 31 features
2. **Data Quality:** Examining the structure of each dataset to understand the number of columns, data types, and any missing values. There were no missing values found in any of the datasets.

3. **Data Imbalance:** Visualizing the class distribution in each dataset reveals a significant imbalance, with the non-fraudulent class having many more instances than the fraudulent class.

4. **Descriptive Analysis:** Performing descriptive analysis to summarize and interpret raw data. This includes measures of central tendency (mean, median, mode), measures of dispersion (range, variance, standard deviation, percentiles/quartiles), data distribution, and the relationship between variables (correlation).

5. **Checking for Outliers:** Using boxplots to visually identify outliers in each dataset.

6. **Handling Missing or Irrelevant Values:** Replacing null values with the mean of the column, if necessary.

3 Model Selection and Training

3.1 Model Selection

Three machine learning algorithms were selected for training: Logistic Regression, Support Vector Machine (SVM), and Decision Tree.

3.2 Oversampling

Due to the imbalanced nature of the dataset, the Synthetic Minority Oversampling Technique (SMOTE) was applied to oversample the minority class (fraudulent transactions).

3.3 Training

Each model was trained on the datasets after applying SMOTE. The models were evaluated using the following metrics:

- Accuracy
- Precision
- Recall
- F1-Score
- ROC AUC (Receiver Operating Characteristic - Area Under Curve)

3.4 Training Process

1. Splitting the datasets into training and test sets.
2. Training each model on the training set.
3. Evaluating the model on the test set.
4. Plotting ROC curves for each model to visualize the trade-off between true positive rate and false positive rate.

4 Results and Analysis

4.1 Dataset 1

- **Logistic Regression:**
  - Accuracy: 0.9498
  - Precision: 0.9039
  - Recall: 0.9039
  - F1-Score: 0.9039
  - ROC AUC: 0.9848
- **SVM:**
  - Accuracy: 0.9494
  - Precision: 0.9376
  - Recall: 0.8477
  - F1-Score: 0.8899
  - ROC AUC: 0.9732
- **Decision Tree:**
  - Accuracy: 0.9995
  - Precision: 1.0000
  - Recall: 0.9990
  - F1-Score: 0.9995
  - ROC AUC: 0.9999

4.2 Dataset 2

- **Logistic Regression:**
  - Accuracy: 0.9505
  - Precision: 0.8995
  - Recall: 0.9130
  - F1-Score: 0.9062
  - ROC AUC: 0.9834
- **SVM:**
  - Accuracy: 0.9486
  - Precision: 0.9311
  - Recall: 0.8436
  - F1-Score: 0.8849
  - ROC AUC: 0.9714
- **Decision Tree:**
  - Accuracy: 0.9994
  - Precision: 0.9990
  - Recall: 0.9990
  - F1-Score: 0.9990
  - ROC AUC: 0.9999

4.3 Dataset 3

- **Logistic Regression:**
  - Accuracy: 0.9438
  - Precision: 0.8429
  - Recall: 0.8934
  - F1-Score: 0.8675
  - ROC AUC: 0.9736
- **SVM:**
  - Accuracy: 0.9432
  - Precision: 0.8871
  - Recall: 0.8114
  - F1-Score: 0.8474
  - ROC AUC: 0.9654

- **Decision Tree:**
  - Accuracy: 0.9994
  - Precision: 0.9990
  - Recall: 0.9980
  - F1-Score: 0.9985
  - ROC AUC: 1.0000

5 Observations

- Decision Tree consistently demonstrates high accuracy, precision, recall, F1-score, and ROC AUC across all datasets, making it a strong candidate for fraud detection in this scenario.
- Logistic Regression and SVM also show good performance, especially Logistic Regression in terms of overall balance between precision and recall.
- ROC Curves illustrate the trade-off between true positive rate and false positive rate, helping us visualize the model's performance in distinguishing between fraudulent and legitimate transactions.

6 Conclusion

This study demonstrates the effectiveness of machine learning algorithms, particularly Decision Tree, in detecting fraudulent credit card transactions. The advanced data preprocessing techniques, including handling data imbalance with SMOTE, significantly improved the model performance. This approach can be extended and refined for real-world applications, ensuring robust financial security and minimizing financial losses due to fraud.