Credit Card Fraud Detection

1 Introduction

Credit card fraud has become a major concern in the financial sector due to the rapid growth of online and digital transactions. Fraudulent activities lead to significant financial losses for banks and customers and require automated systems for early and accurate detection. However, fraud detection is a challenging machine learning problem because fraudulent transactions are extremely rare compared to genuine ones, leading to highly imbalanced datasets.

This project aims to build and evaluate machine learning models for detecting fraudulent credit card transactions. The focus is on preprocessing the data, handling class imbalance, training classification models, and evaluating performance using appropriate metrics such as precision, recall.

2 Methodology

2.1 Dataset Description

The dataset contains credit card transactions labeled as:

- **0**: Genuine transaction
- **1**: Fraudulent transaction

The dataset is highly imbalanced, with fraudulent transactions representing a very small fraction of the total data.

2.2 Data Preprocessing

The following preprocessing steps were applied:

- Missing values for each column was checked.
- Features were separated from the target variable.
- Data was split into training and testing sets using in a 80:20 ratio.
- Feature scaling was performed using `StandardScaler` to normalize the data.

2.3 Handling Class Imbalance

Since our data was highly imbalanced, **SMOTE** was applied to the training data. SMOTE generates synthetic samples for the minority class to create a more balanced training set, improving the model's ability to detect fraud cases.

2.4 Model Training

Two classification algorithms were trained and evaluated:

- **Logistic Regression** with class weights set to `balanced`
- **Random Forest Classifier**

Both models were trained on the resampled training data and evaluated on the original test set.

2.5 Evaluation Metrics

Due to the imbalanced nature of the dataset, accuracy alone is not sufficient. The following metrics were used:

- Precision
- Recall
- F1-score
- ROC-AUC
- Precision-Recall AUC (PR-AUC)

3 Results

3.1 Model Performance

The performance comparison of Logistic Regression and Random Forest models is shown in Table 1.

Table 1: Model Performance Comparison

Metric	**Logistic Regression**	**Random Forest**
Precision	0.0580	0.8632
Recall	0.9184	0.8367
F1-score	0.1092	0.8497
ROC-AUC	0.9699	0.9754
PR-AUC	0.7249	0.8749

3.2 Analysis

Logistic Regression achieved very high recall, indicating that it successfully identified most fraudulent transactions. However, its precision was extremely low, resulting in many false positives.

The Random Forest model provided a much better balance between precision and recall, leading to a significantly higher F1-score. This makes Random Forest more suitable for practical fraud detection systems, where both detecting fraud and minimizing false alarms are important.

4 Conclusion

In this project, machine learning models were developed to detect fraudulent credit card transactions using an imbalanced dataset. Data preprocessing, feature scaling, and SMOTE oversampling were applied to improve model performance.

Experimental results show that while Logistic Regression is effective in maximizing recall, Random Forest significantly outperforms it in terms of precision, F1-score, and overall robustness. Therefore, Random Forest is the preferred model for this fraud detection task.

Future work may include experimenting with advanced ensemble methods, cost-sensitive learning, and real-time fraud detection systems.