Credit Card Fraud Detection using Machine Learning

Abstract

This project aims to detect fraudulent credit card transactions using supervised machine learning. The dataset is highly imbalanced, so SMOTE (Synthetic Minority Oversampling Technique) is applied to balance classes. Models like Logistic Regression and Random Forest are trained, and their performance is evaluated using metrics such as precision, recall, F1-score, and ROC-AUC.

1. Introduction

Credit card fraud poses a significant threat to financial institutions and consumers. Automated fraud detection systems using machine learning can greatly reduce fraud losses by identifying suspicious patterns in real-time.

2. Dataset Overview

The dataset used for this project is from Kaggle and contains 284,807 transactions with 492 labeled as fraud. It includes anonymized numerical features (V1 to V28), along with 'Time', 'Amount', and 'Class' indicating fraud (1) or non-fraud (0).

3. Data Preprocessing

- Dropped 'Time' feature due to irrelevance
- Scaled 'Amount' using StandardScaler
- Split dataset into train/test (80/20)
- Applied SMOTE on training data to address class imbalance

4. Algorithms Used

Two classification algorithms were implemented:

1. Logistic Regression: A linear model suitable for binary classification.
2. Random Forest: An ensemble method using decision trees for improved accuracy.

Credit Card Fraud Detection using Machine Learning

5. Evaluation Metrics

The performance of each model was evaluated using:

- Confusion Matrix
- Precision, Recall, F1-Score
- ROC-AUC Score

6. Results

Random Forest achieved better results due to its ensemble nature, handling non-linear patterns and overfitting better than Logistic Regression.

7. Conclusion

Machine learning algorithms can effectively detect fraudulent transactions. With SMOTE balancing, Random Forest showed high accuracy and recall. Future enhancements could involve deep learning models and real-time analysis.

8. Future Work

- Explore deep learning (LSTM, Autoencoders)
- Real-time detection with streaming data
- Unsupervised anomaly detection methods

9. References

- Kaggle Dataset: https://www.kaggle.com/mlg-ulb/creditcardfraud
- Scikit-learn documentation
- imbalanced-learn documentation

Results Comparison Table

Credit Card Fraud Detection using Machine Learning

Model	Precision	Recall	F1-Score	ROC-AUC
Logistic Regression	0.93	0.87	0.90	0.94
Random Forest	0.97	0.92	0.94	0.97