Credit Card Fraud Detection Project Report

Abstract

This project focuses on detecting fraudulent credit card transactions using machine learning models. Financial institutions face significant losses due to fraudulent activities, making real-time detection crucial. We used Logistic Regression, XGBoost, and Autoencoder models to classify transactions as fraudulent or legitimate.

Problem Statement

Credit card fraud detection is a challenging problem due to highly imbalanced datasets, where fraudulent cases are extremely rare compared to legitimate transactions. The goal is to build a robust model that can accurately identify fraudulent transactions while minimizing false positives.

Dataset Details

Dataset: Credit Card Fraud Detection Dataset (Kaggle) - Total records: 284,807 - Features: 30 (Time, Amount, anonymized V1-V28) - Target variable: Class (0 = Legitimate, 1 = Fraud) - Fraud cases: 492 (~0.17% of dataset)

Methodology

1. **Data Preprocessing**: - Scaled Time & Amount using StandardScaler - Applied SMOTE to handle class imbalance 2. **Modeling**: - Logistic Regression: Baseline linear model - XGBoost: Gradient boosting-based classifier - Autoencoder: Anomaly detection using neural network 3. **Evaluation Metrics**: - ROC-AUC Score - Precision, Recall, F1-Score - Confusion Matrix - Precision-Recall Curves

Model Performance Results

Model	ROC-AUC	Precision (Fraud)	Recall (Fraud)	F1-Score (Fraud)
Logistic Regression	0.9808	0.06	0.92	0.11
XGBoost	1.0000	0.81	1.00	0.89
Autoencoder	0.9508	0.12	0.82	0.21

Insights & Observations

- XGBoost achieved the highest ROC-AUC of 1.0000 and excellent recall (100%) for fraud detection. - Logistic Regression performed well but had low fraud precision. - Autoencoder successfully detected anomalies but had moderate performance compared to XGBoost. - Precision-Recall curves show that XGBoost balances precision and recall best.

Conclusion

XGBoost emerged as the most effective model for credit card fraud detection, providing high recall and precision. The project demonstrates that ensemble models outperform linear and deep anomaly detection methods for highly imbalanced fraud detection tasks.

Future Scope

- Deploying the model in real-time production environments using FastAPI or Streamlit. - Implementing advanced deep learning models (e.g., LSTM, Transformers) for time-series fraud detection. - Enhancing interpretability using SHAP and LIME to explain fraud predictions.