DATA SCIENCE AND ARTIFICIAL INTELLIGENCE

Credit Card Fraud Detection Project Report

Name: UZAIR AHMAD AWAN

Instructor: Mam Maria

INTRODUCTION

Credit card fraud is a serious problem in the financial sector, causing huge monetary losses every year. Detecting fraudulent transactions in real time is crucial to prevent financial damage. This project aims to develop a machine learning model that can classify transactions as either normal or fraudulent, using a real-world dataset. The motivation behind this work is to understand the complete machine learning workflow: data preprocessing, model selection, training, evaluation, and interpretation.

Data Description and Preprocessing

The dataset used contains 284,807 transactions with 31 columns, including anonymized features V1 to V28, Time, Amount, and the target Class (0 = normal, 1 = fraud).

Preprocessing steps performed:

- Scaled Time and Amount using Standard Scaler to normalize the data.
- Dropped original Time and Amount after scaling.
- Separated features (X) and target (y).
- Split the data into **training (80%)** and **testing (20%)** sets using stratified sampling to maintain the same fraud ratio in both sets.

Class imbalance was observed: only **0.17%** of transactions are fraudulent. This is typical in real-world fraud datasets.

Methodology

- Model chosen: Logistic Regression with class_weight="balanced" to handle class imbalance.

- The choice was made because Logistic Regression is simple, fast, and interpretable, suitable for a baseline model.
- Features were fed into the model, trained on the training set, and predictions were made on the test set.

Experiments and Results

Evaluation metrics used:

- Confusion Matrix
- Precision, Recall, F1-Score

Key results:

- **Accuracy:** 97.55%
- **Recall for frauds:** 91.84% → the model successfully detected most fraud cases.
- **Precision for frauds:** 6.09% → some false positives exist (normal transactions predicted as fraud).

Confusion Matrix Summary:

- True Positives (fraud detected correctly): 90
- False Negatives (fraud missed): 8
- True Negatives (normal correct): 55,475
- False Positives (normal flagged as fraud): 1,389

The model performs well in detecting frauds (high recall), which is critical for real-world applications, even if some false alarms occur.

Discussion

- **Interpretation:** Logistic Regression provides a baseline model that effectively identifies fraud. High recall ensures minimal missed fraud cases.
- **Limitations:** Precision is low, resulting in false positives. Advanced models like Random Forest or XGBoost could improve precision.
- **Future Work:** Implementing ensemble models, hyperparameter tuning, and exploring additional features could enhance performance.

Conclusion & References

The project demonstrates a complete machine learning workflow on a real-world credit card dataset. Logistic Regression, along with proper preprocessing and stratified splitting, provides a strong baseline. The results highlight the importance of balancing recall vs precision in fraud detection.

References:

- Kaggle Credit Card Fraud Detection Dataset: https://www.kaggle.com/mlg-ulb/creditcardfraud