Credit Card Fraud Detection – Project Report

---

1. Introduction

Credit card fraud is a significant concern in the financial sector, as unauthorized or illegitimate transactions result in substantial monetary losses. Identifying these transactions in real-time has become critical.

This project implements a machine learning pipeline to detect fraudulent transactions using both unsupervised and supervised learning approaches. We aim to build a deployable system that not only provides accurate predictions but also features an intuitive interface for real-time and batch inference.

---

2. Dataset

The dataset used in this project is the **Credit Card Fraud Detection Dataset** from Kaggle.

Key Details:

- Total transactions: **284,807**
- Fraudulent transactions: **492** (~0.172%)
- Features: **V1 to V28** (PCA-transformed), **Time**, and **Amount**
- Label: **Class** (1 = Fraud, 0 = Normal)

The data is highly imbalanced, which poses a challenge for most classification models. Special care was taken in preprocessing and model evaluation to handle this.

---

3. Techniques Used

We used a hybrid approach combining anomaly detection and supervised classification:

3.1 Anomaly Detection

- **Isolation Forest**: Works by isolating anomalies instead of profiling normal data.
- **Local Outlier Factor (LOF)**: Detects anomalies by comparing the local density of a point with its neighbors.

3.2 Supervised Classification

- **XGBoost Classifier**: A scalable and highly accurate boosting model.

The model was trained on balanced data using under-sampling techniques and evaluated using proper classification metrics.

3.3 Evaluation Metrics

To assess performance:

- **Confusion Matrix**: For true/false positives/negatives.
- **Precision, Recall, F1-Score**: Especially important due to imbalance.
- **ROC Curve & AUC**: For visualizing performance across thresholds.

---

4. Streamlit Dashboard

A user-friendly, interactive dashboard was created using **Streamlit**. It allows:

4.1 Single Transaction Prediction

- Sidebar inputs for Time, Amount, V1 to V28 features.
- Prediction displayed as **Fraud** or **Normal**, along with confidence scores.

4.2 Batch CSV Upload

- Upload a file containing multiple transactions.
- Results displayed with prediction labels.
- If true labels are provided, shows a confusion matrix and ROC curve.

4.3 Visualizations

- **Bar graph** of prediction probabilities.
- **Confusion Matrix** (interactive heatmap using Plotly).
- **ROC Curve** with AUC score.

This makes the tool accessible to analysts and operations teams without deep technical knowledge.

---

5. Project Structure

bash

CopyEdit

credit-card-fraud-detection/

├── app.py # Streamlit dashboard logic

├── models/

| └── project1.pkl # Trained XGBoost model

├── data/ # Placeholder for dataset

```
├── notebooks/
|   └── fraud_detection_pipeline.ipynb  # Full development notebook
├── utils.py               # Helper functions (e.g., preprocessing)
├── requirements.txt       # Dependency file
└── README.md              # Project overview and instructions
```

---

6. Conclusion

This project demonstrates a successful application of machine learning for credit card fraud detection. Key takeaways:

- **Hybrid model approach** improves robustness.
- **Streamlit dashboard** adds usability and practical value.
- Model performs well despite class imbalance, achieving high precision for the minority class.

Future Work

- Automate model retraining with incoming data.
- Deploy as a REST API or integrate with transaction systems.
- Expand dataset with real-world features if possible.