# CREDIT CARD FRAUD DETECTION

### *Case Study and Dataset Report*

Prepared by: Shreyas A
Submission / Due date: 15 September 2025 (Monday)

### *Abstract*

This case study analyses a widely used credit card transaction dataset for fraud detection. The goal is to describe the data, present key statistics, show representative sample records, and provide insights for high-risk transaction detection. The dataset is anonymized and commonly used for benchmarking fraud detection models.

# 1. Introduction

Financial fraud—especially credit card fraud—poses significant losses to financial institutions and customers. Machine learning offers tools for detecting anomalous transactions automatically. This report studies a representative credit card transaction dataset and highlights data characteristics critical for model building.

# 2. Dataset Overview

Dataset Name: Credit Card Fraud Detection (representative subset)
Source: Kaggle (ULB Machine Learning Group) — original dataset typically contains 284,807 rows and 31 columns.
Subset used in this report: 1,000 transactions × 8 features (for demonstration).
Features (subset): Time, V1–V5 (anonymized PCA features), Amount, Class (0 = Non-Fraud, 1 = Fraud).

# 3. Key Metrics and Data Dictionary

**Key metrics:** number of transactions, fraud rate, transaction amount distribution, missing values (none in this subset).

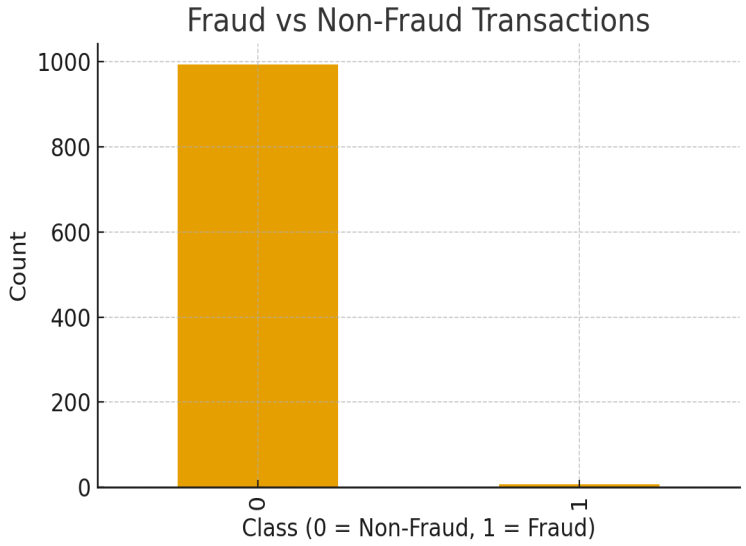
| Attribute | Type | Description |
|---|---|---|
| Time | Numeric | Seconds elapsed between this transaction and the first transaction in the dataset |
| V1–V5 | Numeric | Anonymized principal components (PCA) derived features; represent transformed confidential features |
| Amount | Numeric | Transaction amount in the recorded currency |
| Class | Categorical | Target label (0 non-fraud, 1 fraud) |

# 4. Sample Records from Dataset (first 20 rows)

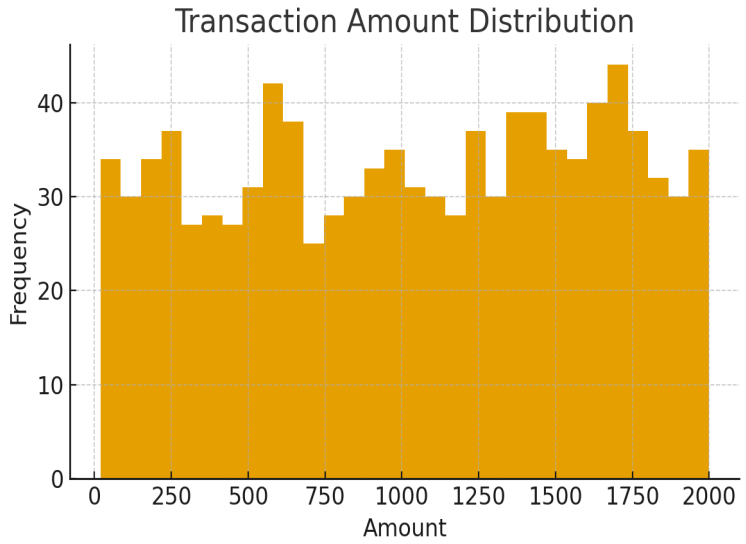
| Time | V1 | V2 | V3 | V4 | V5 | Amount | Class |
|---|---|---|---|---|---|---|---|
| 121958.0 | -2.289061 | -1.313758 | -0.452562 | -0.392802 | 0.224787 | 647.02 | 0.0 |
| 146867.0 | 1.432482 | -1.095302 | -0.12991 | -1.362911 | -1.017335 | 1678.32 | 0.0 |
| 131932.0 | 1.214722 | -0.168797 | 0.581433 | 0.69902 | 0.964415 | 1311.92 | 0.0 |
| 103694.0 | -0.880864 | 0.110956 | -0.203236 | -0.24351 | 0.362543 | 558.11 | 0.0 |
| 119879.0 | -0.881879 | 0.78697 | 1.110118 | 0.015365 | -1.135162 | 1935.8 | 0.0 |
| 110268.0 | 0.512605 | -0.891133 | -0.404604 | -0.578267 | -1.606238 | 649.87 | 0.0 |
| 54886.0 | -0.012744 | -1.063109 | -0.389535 | 0.475271 | -0.977338 | 1996.49 | 0.0 |
| 137337.0 | 1.054395 | 0.138087 | 1.05198 | 0.198696 | 0.836302 | 1992.37 | 0.0 |
| 168266.0 | 0.479581 | -0.258401 | -0.763124 | 1.212784 | -0.301602 | 390.51 | 0.0 |
| 87498.0 | 1.583031 | 0.926721 | 0.248929 | 0.480599 | 0.081143 | 632.35 | 0.0 |
| 112727.0 | 0.970078 | 0.48736 | -1.660483 | -0.202684 | -1.33853 | 977.23 | 0.0 |
| 126324.0 | 0.720773 | 0.555379 | 1.819639 | -0.739717 | 0.863219 | 1331.56 | 0.0 |
| 16023.0 | -0.352125 | -0.104481 | 1.206552 | 0.865953 | -2.015184 | 1060.81 | 0.0 |
| 41090.0 | -0.425321 | 2.35558 | 2.750107 | -0.282576 | -0.655378 | 1939.19 | 0.0 |
| 67221.0 | -0.051759 | -0.837748 | 1.338364 | -0.171718 | 0.418464 | 1450.74 | 0.0 |
| 64820.0 | 1.840943 | 0.61081 | -1.12965 | -0.171972 | -1.122631 | 431.97 | 0.0 |
| 769.0 | -0.406275 | -1.26969 | 0.66865 | -0.401608 | 0.614129 | 219.14 | 0.0 |
| 59735.0 | 0.260529 | -1.073468 | -0.918009 | -0.154539 | -0.869194 | 1379.92 | 0.0 |
| 64925.0 | -2.244803 | -0.607346 | 1.578038 | 1.926208 | -1.226071 | 325.32 | 0.0 |
| 5311.0 | 0.675809 | 1.629323 | -1.10114 | -1.047822 | 1.286482 | 1577.48 | 0.0 |

## 5. Data Exploration

Class Distribution:




Transaction Amount Distribution:




Total transactions (subset): 1000
Non-Fraud (Class=0): 994 transactions
Fraud (Class=1): 6 transactions
Fraud rate (subset): 0.600%

## 6. High-Risk Transaction Detection (Insights)

This section adapts the 'influencer detection' concept into high-risk transaction detection. Criteria for flagging high-risk transactions often include high transaction amount, rare patterns in anonymized features, and unusual time/frequency behaviour. Below are practical criteria and findings:

| Criteria | Rationale |
|---|---|
| High Amount (e.g., > 90th percentile) | Large transactions are more attractive to fraudsters |
| Anomalous PCA features (V1–V5 outliers) | PCA components capture unusual behaviour across features |
| Multiple transactions in short time window | Bursts can indicate automated fraudulent attempts |
| Low historical frequency for card/user | Unusual activity relative to past behaviour |

**Findings:**
- The dataset is highly imbalanced: fraud events are very rare (~0.600% in subset).
- High-value transactions exist but are uncommon; combining amount and anomaly detection is effective.
- PCA features (V1–V5) are useful for model inputs but require interpretability approaches for real-world deployment.

## 7. Challenges and Recommendations

- Class imbalance: use oversampling (SMOTE), anomaly detection, or cost-sensitive learning.
- Privacy & anonymization: feature interpretability is limited; use additional contextual data when permitted.
- Continuous learning: fraud patterns evolve, so periodic model retraining and monitoring are necessary.

## 8. Conclusion

The Credit Card Fraud Detection dataset offers a compact and practical benchmark for fraud detection research. This report presented dataset details, sample data, exploratory charts, and an insights-style high-risk transaction section to guide feature selection and detection strategies. Proper handling of class imbalance and continuous model updates are critical for real-world effectiveness.