Credit Card Fraud Pattern Discovery

Submitted by:

Md. Shakib Ahamed
(ID: 2215151130)
Rawnak Tasnim Ruku
(ID: 2215151152)
Sec: 8C(2)

Submitted to:

Tahmina Yeasmin Lima
&
Paban Shaha
Lecturer,
Dept of CSE,UITS

Introduction

- Credit card fraud is a major challenge in financial security.
- Highly imbalanced dataset: fraud transactions are extremely rare.
- Traditional supervised models are not always effective.
- This project applies multiple Data Mining algorithms to detect fraud patterns.
- **Goal**: Understand why fraud happens, not just detect it.

Objectives

- Apply different Data Mining techniques to identify abnormalities.
- Compare Outlier Detection models.
- Extract frequent fraud patterns using Association Rule Mining.
- Analyze structural behaviour of transactions using graph algorithms.
- Integrate results to understand fraud behaviour comprehensively.

Dataset Description

Source:

- Kaggle – Credit Card Fraud Detection
- Size: 284,807 transactions
- Key Challenge: Class imbalance (fraud ≈ 0.17%)

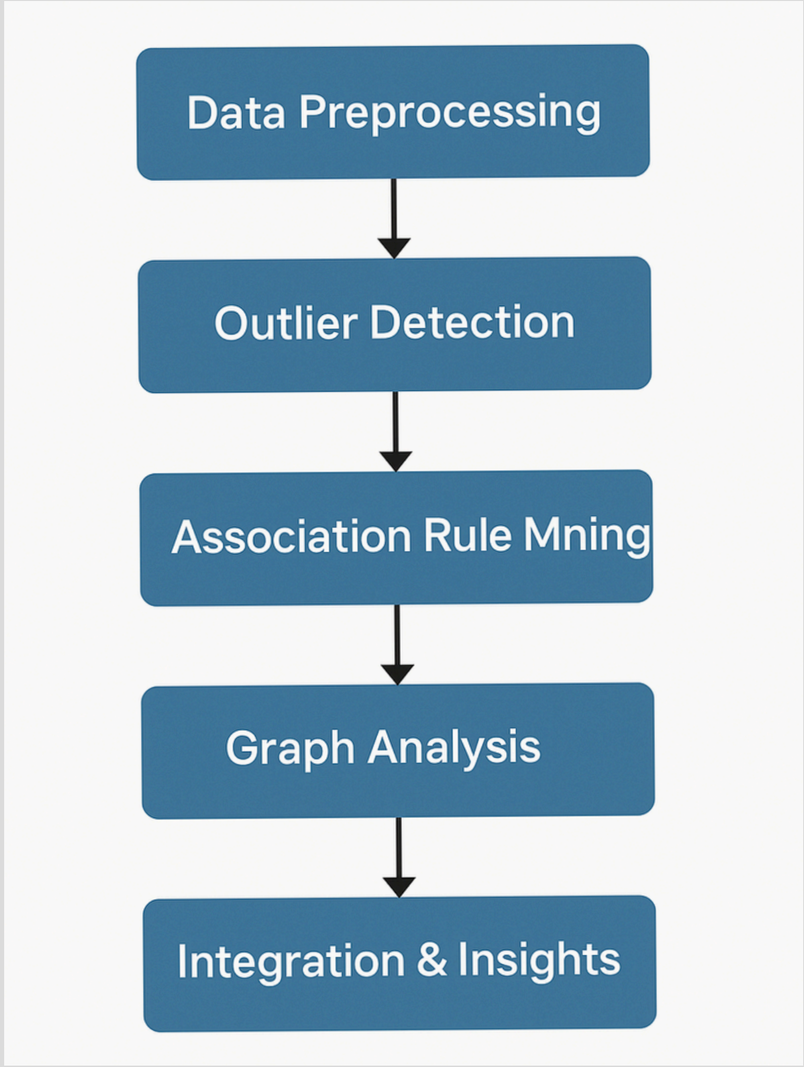
Features:

- 28 PCA-transformed features (V1–V28)
- Time, Amount
- Class (0 = Normal, 1 = Fraud)

Data Preprocessing

- Load dataset from Google Drive
- Check missing values
- Remove duplicates
- Analyze fraud vs non-fraud distribution
- Standard scaling on all features
- 80/20 train-test split with stratification

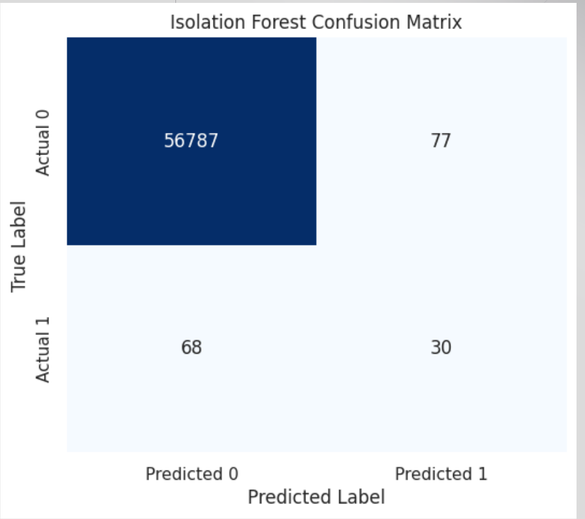
Methodology



Outlier Detection Algorithms

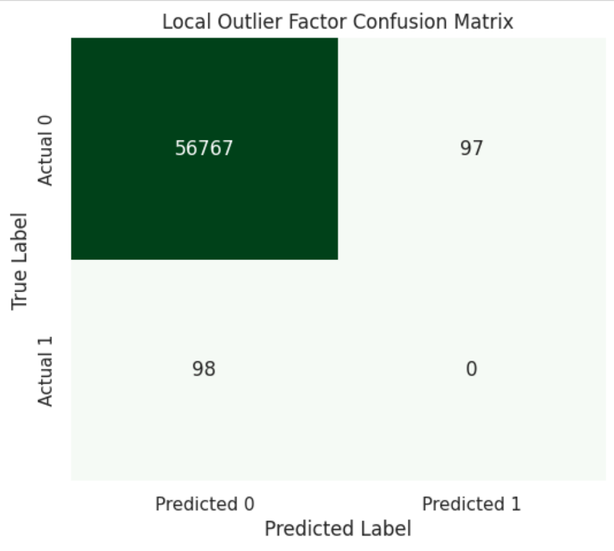
Isolation Forest

- Random tree splitting is the basis for the isolation forest.
- Outliers are "isolated" more quickly.
- Effectively handles high-dimensional data



Local Outlier Factor (LOF)

- Detects anomalies by local density
- Works poorly on highly imbalanced data



Outlier Detection Results

Isolation Forest

Isolation Forest

- Normal detection ≈ 99.8%
- Fraud recall ≈ 31%
- Better model among the two

Conclusion: Isolation Forest performs significantly better.

Local Outlier Factor (LOF)

LOF

- Predicted almost everything as normal
- Fraud recall = 0%
- Not suitable without heavy tuning

Association Rule Mining (Apriori)

- Fraud-only subset used
- Continuous features binned (Amount, Time)
- PCA features converted to positive/negative sign
- One-hot encoded for Apriori

Findings:

- Fraud tends to cluster in specific PCA patterns
- Rules with high Lift show consistent patterns

Graph-Based Analysis

kNN Graph Creation

- 2,000 sampled transactions
- 10 nearest neighbours per node
- Constructed directed graph

Graph Algorithms Used:

- PageRank → identifies globally important nodes
- HITS → identifies hub & authority scores

PageRank & HITS Results

Client : Fradel and Spies

PageRank

- Fraud nodes have lower PageRank
- Non-fraud dominate network structure

HITS

- Fraud has lower hub score
- Fraud has almost zero authority score

Conclusion:
Fraud transactions lie at the network periphery, not in central positions.

Conclusion

- This project demonstrates a complete multi-algorithm fraud pattern analysis
- Unsupervised, association, and graph-based approaches combined
- Fraud behaviour is explained through:
  - Outliers
  - Frequent itemsets
  - Graph centralities
- Strong foundation for future supervised models

Future Works

- Integrating graph features into ML classifiers
- Using SMOTE for class imbalance
- Applying XGBoost/LightGBM
- Time-series fraud sequence modelling
- Real-time fraud detection pipeline

References

[1] F. T. Liu, K. M. Ting and Z.-H. Zhou,“Isolation Forest,” 2008 IEEE International Conference on Data Mining, Pisa, Italy, 2008, pp. 413–422.
[2] R. Agrawal and R. Srikant, “Fast Algorithms for Mining Association Rules,” Proc. 20th Int’l Conf. Very Large Data Bases (VLDB), pp. 487–499, 1994.
[3] S. Brin, R. Motwani and T. Winograd, “The PageRank citation ranking: Bringing order to the web,”Technical Report, Stanford InfoLab, 1998.

Thank you

by : Shakib & Rawnak