FINANCIAL TRANSACTION ANOMALY DETECTION USING K-MEANS CLUSTERING

Author: Oreoluwa Anjorin
Date: August 2025
Project: Anomaly Detection in Financial Transactions (Synthetic Dataset)

EXECUTIVE SUMMARY

This report presents a comprehensive analysis of anomaly detection in financial transactions using K-Means clustering algorithm on a **synthetic dataset** of 100 financial transactions. The dataset was artificially generated to simulate real-world transaction patterns for educational and portfolio demonstration purposes. The analysis successfully identified 5 anomalous transactions (5.00% anomaly rate) using distance-based detection with a 95th percentile threshold. The findings reveal significant patterns in fraudulent transaction behavior, with anomalous transactions averaging 10.87 times the amount of normal transactions.

---

1. METHODOLOGY AND IMPLEMENTATION

1.1 Dataset Overview

Note: This project uses **synthetic/artificial data** generated specifically for this analysis. The dataset simulates realistic financial transaction patterns but contains no real customer information or actual transaction data.

The synthetic dataset contains 100 simulated financial transactions with the following characteristics:

- **Features analyzed**: Transaction Amount (₦) and Time
- **Amount range**: ₦23.80 - ₦1,200.00
- **Time range**: 20.41 - 500.00 time units
- **Average normal transaction amount**: ₦55.95
- **Standard deviation of amounts**: ₦168.41

1.2 Algorithm Implementation

The K-Means clustering approach was implemented with the following specifications:

- **Number of clusters (k)**: 4
- **Features**: Standardized Amount and Time using StandardScaler
- **Random state**: 42 for reproducibility
- **Initialization**: 10 random initializations (n_init=10)

2. DISTANCE-BASED ANOMALY DETECTION FRAMEWORK

2.1 Distance Calculation Mechanism

The anomaly detection system employs minimum distance to cluster centers as the primary anomaly indicator. For each transaction, the algorithm:

1. Standardizes features using Z-score normalization to ensure equal weight between Amount and Time dimensions
2. Computes distances to all 4 cluster centers using Euclidean distance
3. Selects minimum distance as the anomaly score for each transaction
4. Applies threshold to classify transactions as normal or anomalous

2.2 Threshold Selection and Justification

Selected Threshold: 0.1668 (95th percentile of distances)

The 95th percentile threshold was chosen based on several analytical considerations:

- **Statistical Foundation**: Captures the top 5% of distances, aligning with standard anomaly detection practices where 5-10% anomaly rates are typical in financial datasets
- **Business Context**: In financial fraud detection, a 5% false positive rate is acceptable given the high cost of missing actual fraud
- **Data Distribution**: Analysis of the distance distribution showed a natural break at the 95th percentile, indicating a clear separation between normal and anomalous patterns

Alternative Threshold Analysis:

- 90th percentile (0.1521): Would identify 10 anomalies (too sensitive)
- 99th percentile (0.6803): Would identify only 1 anomaly (too conservative)
- 95th percentile provides optimal balance between detection sensitivity and false positive rate

3. BOUNDARY TRANSACTION HANDLING

3.1 Boundary Detection Methodology

Transactions falling close to cluster boundaries pose unique challenges in anomaly detection. The system addresses this through:

Boundary Transaction Identification:

- **Method**: Calculate standard deviation of distances to all cluster centers
- **Low standard deviation**: Indicates equidistant positioning from multiple clusters
- **Identified boundary transaction**: Synthetic Customer ID 97
    - Amount: ₦500.00
    - Time: 250.00
    - Cluster: 2
    - Distance to center: 0.6803

3.2 Boundary Handling Strategy

For boundary transactions, the system implements a conservative classification approach:

- Primary classification based on minimum distance to any cluster center
- Secondary validation using cluster membership stability
- Risk assessment considering the transaction's position relative to multiple clusters

Business Implication: Boundary transactions require additional manual review due to their ambiguous classification, representing potential false positives or sophisticated fraud attempts.

---

4. IDENTIFIED ANOMALIES - DETAILED ANALYSIS

4.1 Anomaly Summary

- **Total Anomalies Detected**: 5 transactions
- **Anomaly Rate**: 5.00%
- **Most Anomalous Transaction**: Synthetic Customer ID 96 with 407.9% of threshold

4.2 Individual Anomaly Analysis

Transaction ID	Amount (₦)	Time	Distance Score	Anomaly Level
96	300.00	200.00	0.6803	407.9% of threshold
97	500.00	250.00	0.6803	407.9% of threshold
98	700.00	400.00	2.9234	1,752.5% of threshold
99	1,000.00	450.00	3.5481	2,127.4% of threshold
100	1,200.00	500.00	4.1728	2,502.4% of threshold

4.3 Anomaly Pattern Analysis

Amount-Based Patterns:

- Average anomalous transaction: ₦608.38
- Average normal transaction: ₦55.95
- Anomaly multiplier: 10.87x normal amount
- 80% of anomalies (4/5) fall in the top 10% of transaction amounts

Time-Based Patterns:

- Average anomalous transaction time: 288.72
- Average normal transaction time: 33.91
- Anomalies occur at 8.52x the average normal time
- Suggests after-hours or unusual timing patterns

Cluster Distribution:

- Cluster 1: 2 anomalies (high-value, high-time transactions)
- Cluster 2: 2 anomalies (moderate-value, high-time transactions)
- Cluster 0: 1 anomaly (low-value, moderate-time outlier)

5. CLUSTER ANALYSIS AND BUSINESS INSIGHTS

5.1 Cluster Characteristics

Cluster	Count	Avg Amount (₦)	Avg Time	Pattern Description
0	95	49.03	30.16	Normal Operations - Standard business hours, typical amounts
1	2	1,100.00	475.00	High-Value Late - Large amounts, after-hours
2	2	400.00	225.00	Moderate Unusual - Medium amounts, unusual timing
3	1	700.00	400.00	Single Outlier - Unique pattern

5.2 Risk Assessment by Cluster

Cluster 0 (Normal Operations): Risk Level: LOW

- Business Context: Regular customer transactions during business hours
- Monitoring: Standard automated monitoring sufficient

Clusters 1 & 2 (Anomaly Clusters): Risk Level: HIGH

- Business Context: Unusual timing and amounts suggest potential fraud or VIP transactions
- Monitoring: Enhanced manual review required

Cluster 3 (Single Transaction): Risk Level: MEDIUM

- Business Context: Requires investigation to determine if pattern emerges

6. TECHNICAL IMPLEMENTATION

6.1 Code Structure

The project consists of several Python modules:

- `anomaly_detection.py`: Main analysis script
- `threshold.py`: Threshold calculation utilities
- `boundary_handling.py`: Boundary transaction analysis
- `transactions_nigeria.csv`: Synthetic dataset

6.2 Key Libraries Used

- **pandas**: Data manipulation and analysis
- **numpy**: Numerical computations
- **scikit-learn**: Machine learning algorithms (K-Means, StandardScaler)
- **matplotlib/seaborn**: Data visualization

---

7. LIMITATIONS AND FUTURE IMPROVEMENTS

7.1 Current Limitations

1. **Feature Limitation**: Only uses Amount and Time features
2. **Algorithm Constraints**: K-Means assumes spherical clusters
3. **Static Threshold**: Fixed 95th percentile may not adapt to evolving patterns
4. **Synthetic Data**: Results based on artificially generated data

7.2 Future Enhancements

1. **Feature Engineering**: Include transaction frequency, merchant category, location
2. **Hybrid Algorithms**: Combine with Isolation Forest or One-Class SVM
3. **Dynamic Thresholding**: Implement adaptive threshold based on recent patterns
4. **Real-time Processing**: Stream processing capabilities for live detection

---

8. CONCLUSION

This personal project demonstrates the effectiveness of K-Means clustering for anomaly detection in financial transactions using synthetic data. The analysis successfully identified 5 suspicious transactions representing 5.00% of the dataset. The key findings include:

- **Anomalous transactions are characterized by**:
  - 10.87x higher amounts than normal transactions
  - 8.52x longer processing times suggesting after-hours activity
  - Clear clustering patterns separating anomalies from normal operations

- **System Performance**:
  - 5.00% anomaly detection rate (appropriate for financial data)
  - Clear separation of anomalous patterns (407.9% of threshold for top anomalies)
  - Actionable insights for investigation prioritization

While K-Means provides a solid foundation for anomaly detection in this synthetic financial dataset, the identified limitations suggest that a more sophisticated, multi-algorithm approach would enhance detection accuracy and reduce false positives.

The system demonstrates particular strength in detecting high-value, unusual-timing transactions but may require enhancement for detecting more subtle fraud patterns. The 95th percentile threshold provides appropriate balance between detection sensitivity and operational feasibility.

---

9. PROJECT REPOSITORY

This project is part of my data science portfolio demonstrating:

- Unsupervised learning techniques
- Anomaly detection methodologies
- Financial data analysis
- Python programming skills
- Statistical analysis and reporting

Note: All data used in this project is synthetic and generated for educational purposes. No real financial or personal data was used in this analysis.

This project was created as a learning exercise and portfolio demonstration. The synthetic dataset and analysis provide insights into anomaly detection techniques while maintaining data privacy and security.