Fraud Detection in Financial Transactions: Feature Analysis and Predictive Modelling.

Introduction

Fraud in financial transactions is a growing problem, costing businesses billions and damaging customer trust. With the increase in digital transactions in banking, retail, and e-commerce, detecting fraud has become more challenging because of the large data volumes and ever-changing fraud tactics. Traditional rule-based systems struggle to keep up, making data-driven analysis essential for fraud detection.

This project focuses on preparing transaction data, analyzing patterns, and then training and exploring a few machine learning models to uncover insights linked to fraud. We focus on key features like transaction amounts, locations, times, and channels to identify patterns. The aim is to provide actionable insights to help financial institutions manage fraud more effectively. [1]

We also explore challenges in fraud detection, such as the rarity of fraud cases in datasets and the need for models that are easy to interpret. Tools like SHAP (SHapley Additive exPlanations) help us explain why certain transactions are flagged as fraud, making the results both accurate and understandable.

By combining exploratory data analysis with machine learning, this project aims to provide practical insights that can help reduce financial losses and improve fraud prevention strategies.

1. Analytical Questions and Data

1.1 Research Questions

The project is driven by the following research questions, which aim to address key aspects of fraud detection:

1. **What transaction features (e.g., amount, location, channel, time) have the highest correlation with fraudulent activity?**
   (Understanding which features are most indicative of fraud helps prioritize risk factors and improve fraud detection models.)

2. **How do high-value transactions and user location (distance from home) relate to fraud across different countries?**
   (This question investigates whether certain transaction behaviours (e.g., high-value, long distances) are stronger predictors of fraud in specific regions.)

3. **How does the time of transaction (e.g., time of day) influence fraud likelihood, and are there identifiable peak periods for fraudulent activity?**
   (Exploring temporal patterns helps identify when fraudulent activities are most likely to occur, enabling targeted prevention strategies.)

4. **Which machine learning model works best for detecting fraud, and how can tools like SHAP help us understand their decisions?**
   Comparing model performance and identifying influential features ensures actionable and interpretable fraud predictions.

1.2 Dataset Overview

The dataset is a synthetic version of real-world transaction data, created to mimic actual transactions while ensuring customer privacy. It includes 26 columns covering: [2]

- **Transaction Details**: Amount, currency, merchant category, merchant type.
- **Geographic Data**: Country, city, city size.
- **Customer Behavior**: Card present, device, channel.
- **Fraud Indicator**: The target variable, is_fraud, shows whether a transaction is fraudulent.

Limitations
As the data is synthetically generated, it may not fully reflect real-world patterns. Some variables may be skewed, which could affect the applicability of insights to real scenarios. [3]

2. Analysis

This section outlines the stages of our analysis to identify patterns and insights from transactional data. [3]
Stages:

1. Load, Clean, and Preprocess Data
2. Data Visualization
3. Model Training and Evaluation
4. SHAP Analysis and Interpretability

2.1 Load, Clean, and Preprocess Data:

2.1.1 Data Pre-processing and Sampling

A sample of 300,000 rows was extracted from the 1000,000-row dataset to manage the large size efficiently.

2.1.2 Selective Data Loading for Optimized Processing:

Irrelevant columns were removed, focusing on key features like timestamp, merchant category, amount, country, city size, card type, and others related to transaction data.

2.1.3 Handling Missing Values:

Missing data in critical columns (e.g., "amount" and "card_type") was addressed using statistical imputation techniques. The median was selected for the "amount" column to mitigate the influence of outliers, while the mode was used for "card_type" due to its categorical nature.

2.1.4 Date-Time Handling & Creating New Features:

The timestamp column (e.g., 2024-09-30 13:11:59.525302+00:00) is split into separate features: year, month, day, day of the week, hour, minute, second, and microsecond to facilitate temporal pattern analysis.

2.1.5 Feature Engineering

The "velocity_last_hour" feature, which contained nested data, was expanded into individual columns like "num_transactions" "total_amount" and “unique_countries”.

2.1.6 Feature Transformation

For categorical variables, both one-hot encoding and label encoding were tested. Label encoding was chosen as it handled the data better, avoided creating numerous columns like one-hot encoding, and worked well for analysis and visualization. Boolean features were also converted for consistency.

2.2 Data Visualization

2.2.1 Correlation Matrix:

We start by plotting a Correlation Matrix of all features to identify features strongly correlated with fraud. This gives us a strong foundation for determining which variables to focus on going further in our analysis.

2.2.2 Distribution Plots and Scatter Plots:

After we find key features we explore further by plotting various distribution plots, scatter plot, and histograms. These visualizations revealed patterns, such as the relationship between transaction amount and fraud.

2.2.3 Outlier Detection and Boxplot Visualization for Numerical Features: (Fig. 1)

Boxplots showed many outliers, particularly in "amount" and "unique_merchants," which could skew results. We used the MinMaxScaler and median to mitigate their effect.

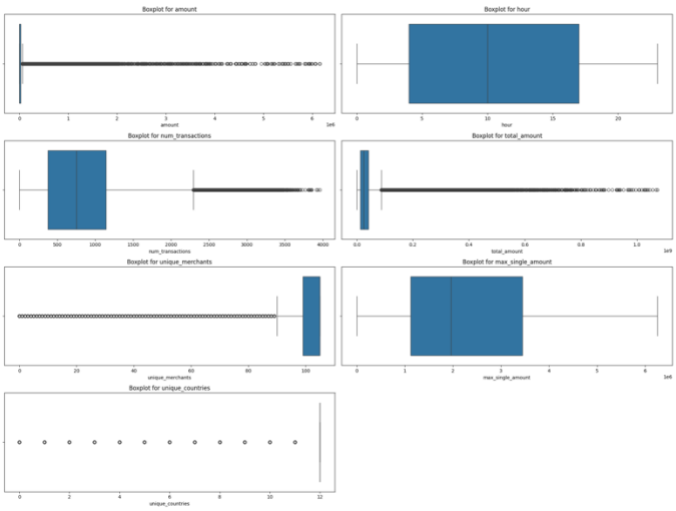

Figure 1: Boxplots for Numerical Columns

To better understand the outliers, we calculated their counts for each feature using the interquartile range (IQR) method.

Output:

```
Outlier Count for amount: 46224
Outlier Count for hour: 0
Outlier Count for num_transactions: 3684
Outlier Count for total_amount: 24995
Outlier Count for unique_merchants: 46673
Outlier Count for max_single_amount: 0
Outlier Count for unique_countries: 32890
```

2.2.4 Time-Based Analysis

Line plots were used to detect fraud patterns by hour, day, or week.

2.3 Model Training and Evaluation

2.3.1 Model Selection

Random Forest, Gradient Boosting, and Decision Tree were chosen for their ability to handle complex, non-linear data.

- **Random Forest**: Handles a wide range of features.
- **Gradient Boosting**: Detects complex patterns.
- **Decision Tree**: Provides a baseline for comparison.

2.3.2 Feature Selection:

The 20 most important features are selected based on their correlation with the target variable (is_fraud) using the correlation matrix. While feature selection was initially part of data preprocessing, this step is revisited after analysis to refine the feature set further. This ensures the model focuses on the most relevant features, improving performance and reducing complexity.

2.3.3 Data Splitting:

The dataset is divided into training and testing sets, with an 80:20 split. The target variable (is_fraud) is separated from the feature set, ensuring a clean structure for model training and evaluation.

2.3.4 Data Scaling:

Using MinMaxScaler, features are normalized to a common scale, preventing disproportionate influence from variables with larger numerical ranges (e.g., amount vs. distance_from_home). This step enhances the model's ability to learn effectively.

2.3.5 Evaluation Metrics

We focused on metrics like accuracy, recall, F1-score, and AUC to assess model performance. Recall was given special attention because misclassifying fraudulent transactions has a higher cost.

2.3.6 Custom Function :

We created a custom function to evaluate classification models. It calculates metrics like accuracy, recall, AUC, and F1-score, and generates a ROC curve. The function also displays a confusion matrix, offering a comprehensive performance evaluation.

2.3.7 Model Interpretation and Validation

SHAP Explanation: (fig 16)
SHAP (SHapley Additive exPlanations) interprets model predictions by detailing feature contributions, offering insights into why the model predicts fraud for specific transactions.

Why SHAP?

- Provides both global and local interpretability.
- Ensures explainability, critical in sensitive domains like fraud detection.
- Offers actionable insights for domain experts.

3. Findings, reflections and further work

3.1. Data Visualization

To address the research questions, several visualizations were employed to explore the relationship between transaction features and fraudulent activity.

Addressing Research Question 1: What transaction features (e.g., amount, location, channel, time) have the highest correlation with fraudulent activity?

3.1.1 Correlation Matrix:

The correlation matrix in figure 2 provides an overall view of which features are most strongly correlated with fraud and serves as a solid starting point for the analysis.

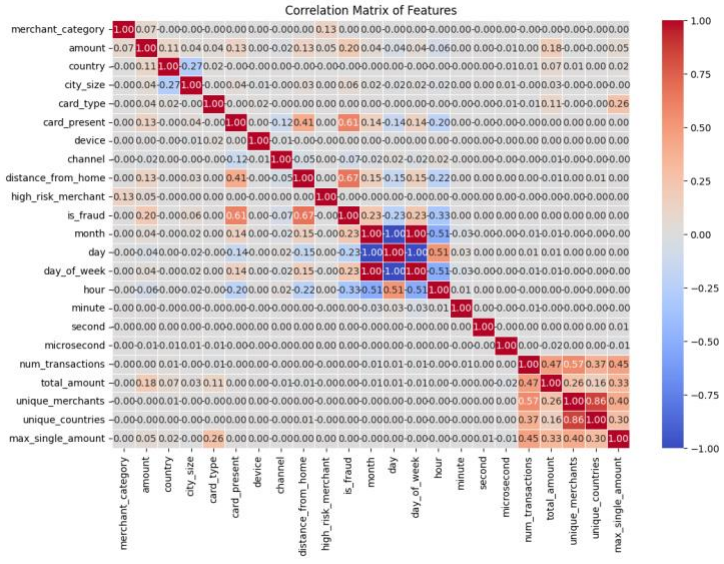

Figure 2: Confusion Matrix of Features

The heatmap highlights features strongly correlated with fraud (is_fraud):

- **distance_from_home (0.67):** Fraud is more likely for transactions far from the user's usual location.
- **card_present (0.61):** In-store transactions are less prone to fraud than card-not-present ones.
- **amount (0.20):** Larger amounts have a slight association with fraud.

Key features like "distance_from_home," "card_present," "hour," "day_of_week," and "amount" are most impactful for analysis and modelling.

3.1.2 Distance from Home Distribution:

The distribution of transaction distances from the user's home is examined to assess how it influences the likelihood of fraud. It was found that fraud tends to increase as the distance from home increases, whereas transactions closer to home have a lower chance of being fraudulent. As seen in figure 3.

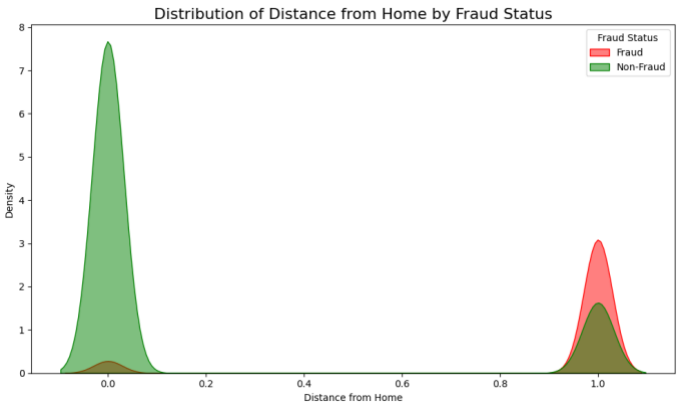

Figure 3

3.1.3 Amount vs. Max Single Amount (Scatter Plot):

To explore the relationship between transaction amount and the maximum amount transacted in the last hour, a scatter plot is used (Fig 4).

From the below graph, there's no clear linear relationship between the max amount (in the last hour) and the transaction amount. While a slight linear trend exists, many points deviate, indicating a non-linear relationship. This suggests that other factors may influence fraud more strongly, and the connection between these features is more complex than initially thought. This warrants further exploration to better understand the dynamics of fraud.

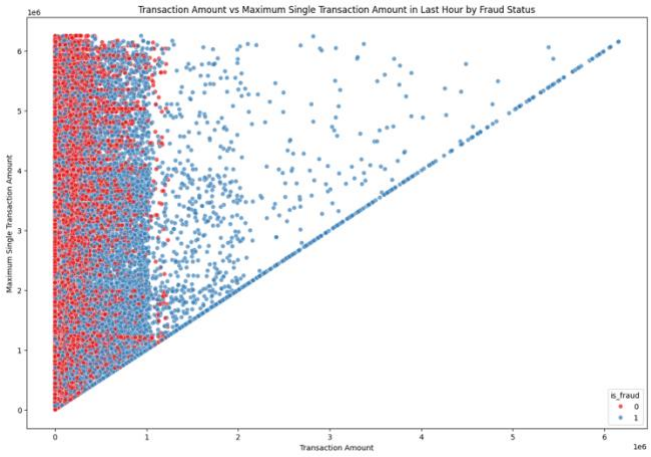

Figure 4

Addressing Research Question 2: How do high-value transactions and user location (distance from home) relate to fraud across different countries?

3.1.4 Country-Specific Fraud Insights

To examine how different countries contribute to fraud, a graph is plotted comparing the "median" transaction amounts for fraud and non-fraud transactions across countries.
The median is used to focus on the central tendency of transaction amounts, minimizing the impact of outliers and offering a clearer comparison of fraud patterns across countries.

Nigeria, Russia, and Japan exhibit the highest fraud rates as shown in figure 5. The further analysis will focus on these countries.

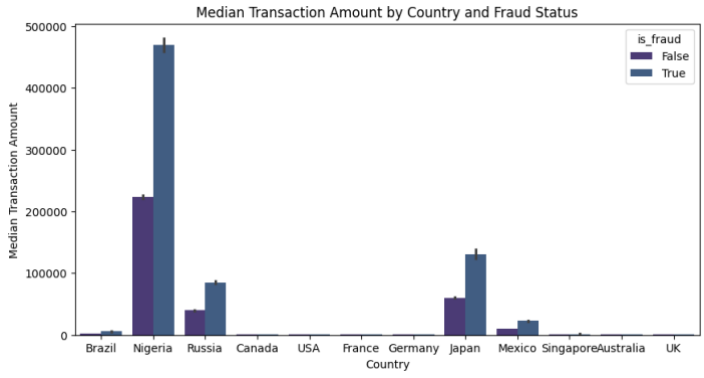

Figure 5

3.1.5 High-Value Transactions in Specific Countries:

High-value transactions are filtered and measured to determine if they are more likely to be fraudulent. Factors such as **distance**

from home and **card presence** are analyzed to explore their relationship with fraud based on earlier insights.

The first country analyzed is Nigeria. In Figure 6, the output can be seen, showing the distribution of high-value transactions.

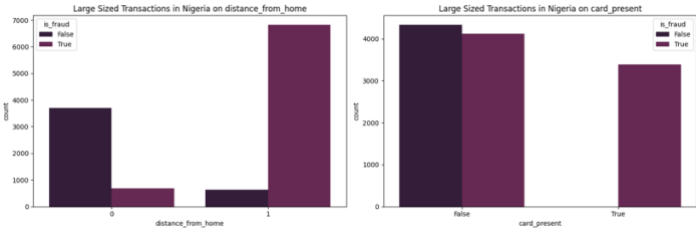

Figure 6

Insights from the graph for Nigeria:

- **Distance from Home:** High-value transactions in Brazil are predominantly from clients far from home, showing higher fraud rates.
- **Card Present**: Online transactions exhibit higher fraud, while offline transactions are mostly fraudulent, with very few non-fraudulent cases.
- **Data Skew:** Offline transactions are mainly fraudulent, indicating skewed data.

These observations remain consistent for other countries as well, including Russia (Figure 7) and Japan (Figure 8).

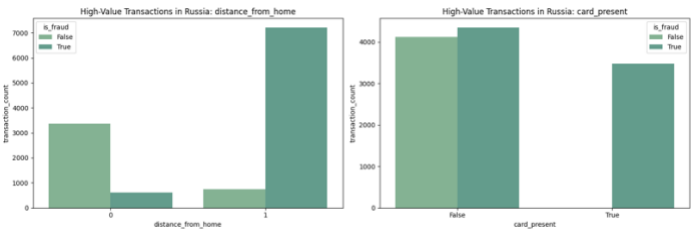

Figure 7

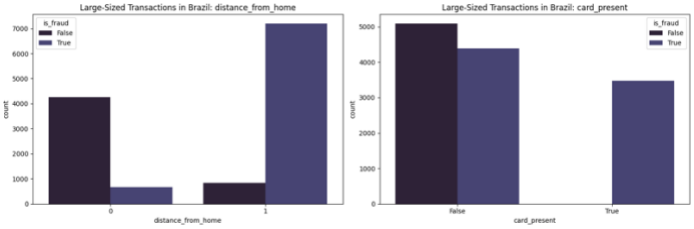

Figure 8

Addressing Research Question 3: How does the time of transaction (e.g., time of day) influence fraud likelihood, and are there identifiable peak periods for fraudulent activity?

3.1.6 Transaction Hour and Fraud Detection (Amount-Based Analysis)

The median transaction amount is calculated for each transaction hour and fraud status. A line plot is created to visualize the trends in the median transaction amount across different transaction hours, with fraudulent and non-fraudulent transactions represented by distinct colours. (Fig. 9)

Vertical reference lines are added for clarity, marking the bank's opening time at 9 AM (red) and closing time at 5 PM (blue).

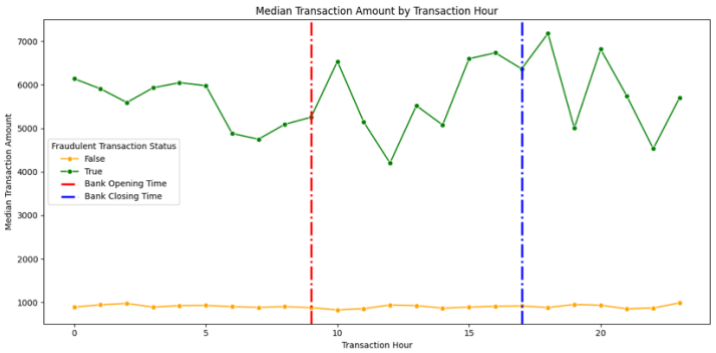

Figure 9

Results:
Fraudulent transactions show peaks around hour 10 and hour 16, indicating higher fraud activity at specific times.

- Non-fraudulent transactions remain stable throughout the day.
- Fraud seems to increase during bank opening hours (hours 9 and 10), suggesting potential vulnerabilities during those times.

3.1.7 Transaction Hour and Fraud Detection (Frequency-Based Analysis)

The median number of transactions is calculated by transaction hour and fraud status. A line plot highlights trends in transaction frequency, using the same reference lines for bank operating hours.

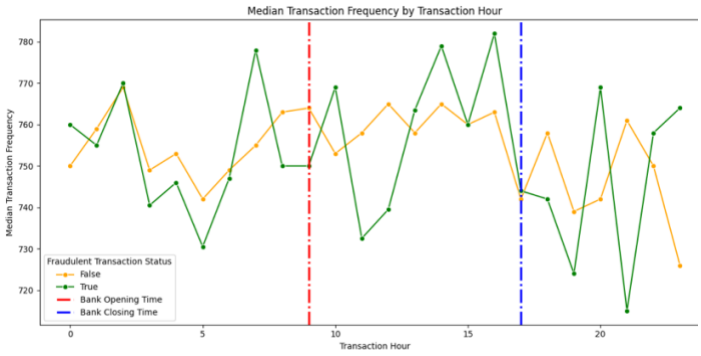

Figure 10

Results: Frequency-Based Analysis (Fig 10)

- This analysis shows more fluctuations throughout the day.
- Both fraudulent and non-fraudulent transactions display multiple peaks and dips, reflecting higher variability in frequency compared to the steadier patterns in transaction amounts.

We can see that fraud peaks during specific hours, helping us understand real-world vulnerabilities and focus on those times for better prevention.

Addressing Research Question 4: Which machine learning model works best for detecting fraud, and how can tools like SHAP help us understand their decisions?

3.2. Model Training and Evaluation

3.2.1. Models Performance

Random Forest Classifier: (Fig 11, 12)

- Accuracy: 96%
- Recall: 89.29%
- AUC: 93.72%
- F1-Score: 91.62%
- Strengths: High accuracy and balanced performance for fraudulent and non-fraudulent transactions.

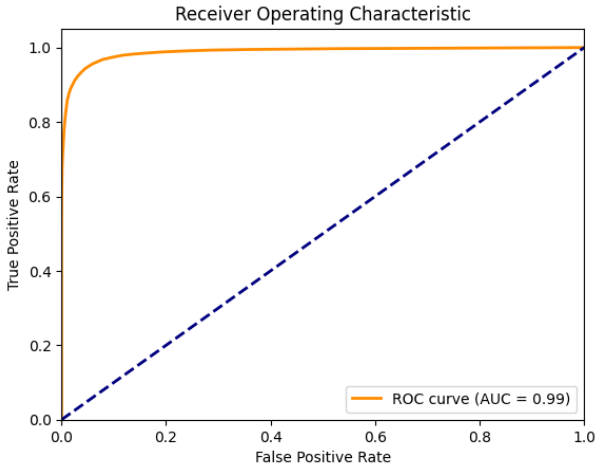

Figure 11

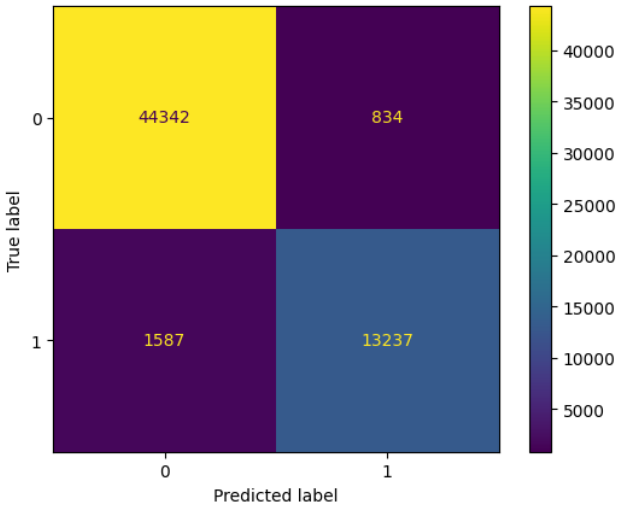

Figure 12

Gradient Boosting Classifier:(13, 14)

- Accuracy: 97%
- Recall: 91.80%
- AUC: 95.33%
- F1-Score: 94.01%
- Strengths: Slightly better recall and AUC compared to Random Forest, making it more robust for imbalanced datasets.

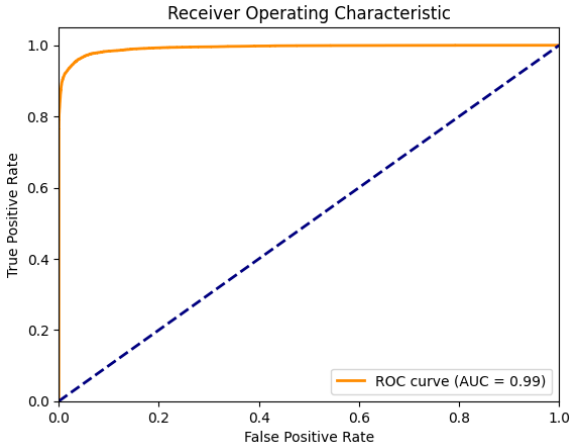

Figure 13

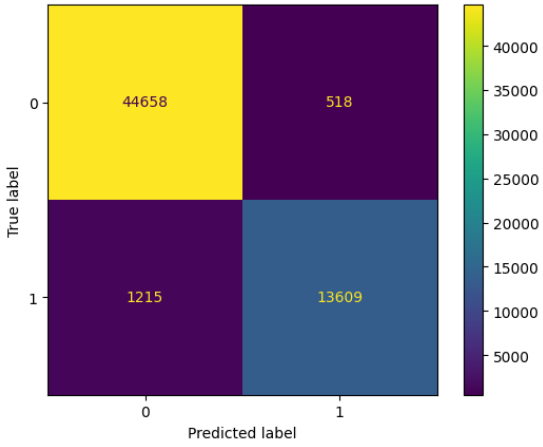

Figure 14

Decision Tree Classifier(15, 16)

- Accuracy: 96%
- Recall: 92.32%
- AUC: 94.75%
- F1-Score: 91.91%
- Strengths: Simpler and faster but slightly less robust than Gradient Boosting for highly complex patterns.

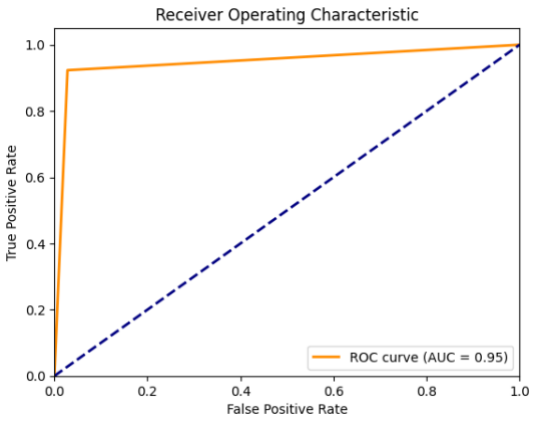

Figure 15

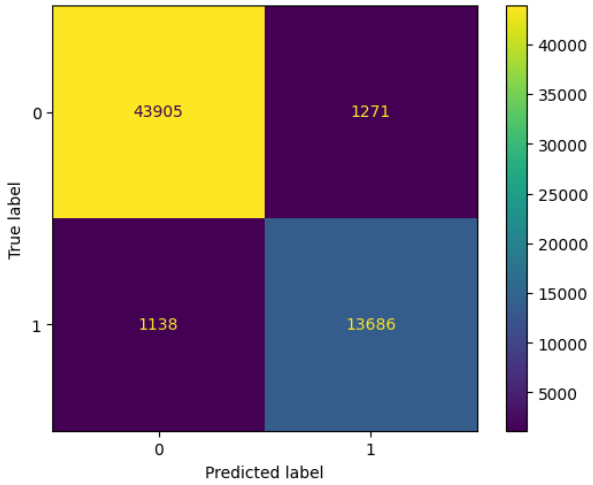

Figure 16

Model Comparison Table

Model	Accuracy	Recall	AUC	F1-Score
Random Forest	96%	89.29%	93.72%	91.62%
Gradient Boosting	97%	91.80%	95.33%	94.01%
Decision Tree	96%	92.32%	94.75%	91.91%

3.2.2. SHAP Analysis

Key Features Identified by SHAP:

- Positive SHAP values increase the likelihood of fraud, while negative values decrease it.
- Colour coding: Blue dots represent lower feature values, while red dots indicate higher feature values.
- Key features like distance_from_home, card_present, amount, and hour significantly impact predictions.
- The feature importance derived from SHAP aligns with earlier analyses from data visualization, confirming that the model is focused on the most influential predictors.

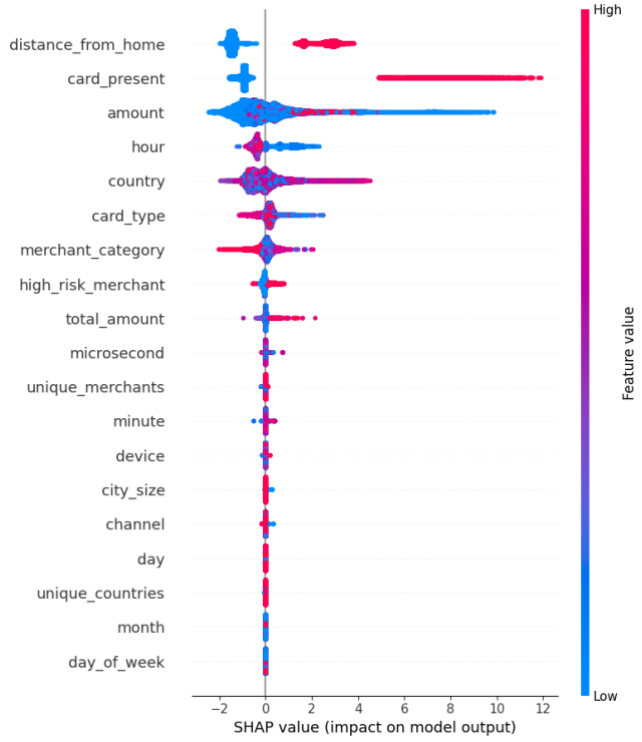

Figure 17

3.3. Reflections

This project demonstrated that transaction data holds valuable insights for detecting fraud, with key features like distance from home, transaction amount, and time of day proving significant. We explored three models—Random Forest, Gradient Boosting, and Decision Tree—and found that Random Forest (F1-score) performed the best. However, using synthetic data limited our findings, and real-world data could have provided more accurate results. Additional features, like geolocation, could have enhanced fraud

detection. While SHAP helped with interpretability, it could have been used more extensively to uncover deeper insights and better understand feature interactions, potentially improving model performance.

3.4. Future Work

Future work will explore more advanced models and incorporate additional features such as geolocation or transaction metadata. Expanding the analysis to include other cryptocurrencies and integrating real-time fraud detection could enhance the practical application of the system. Additionally, the focus will shift towards blockchain and cryptocurrencies, specifically analyzing fraudulent transactions associated with far-right groups and terrorist organizations, to identify patterns and improve tracking mechanisms.

Introduction : 187
Analytical questions and data: 254
Analysis: 630
Findings, reflections and further work: 650

References

1 Bibliography

[1] A. Oza, “Fraud Detection using Machine Learning,” [Online]. Available: https://cs229.stanford.edu/proj2018/report/261.pdf.

[2] S. N. Jallepalli, “Kaggle,” [Online]. Available: https://www.kaggle.com/code/sainikhiljallepalli/fraud-detection.

[3] S. A. Razak, “MDPI,” [Online]. Available: https://www.mdpi.com/2076-3417/12/19/9637.

[4] M. S. R. Barrero, “nature,” [Online]. Available: https://www.nature.com/articles/s41599-024-03606-0.

[5] A. Oza, Fraud Detection using Machine Learning.

[6] S. A. Razak, Financial Fraud Detection Based on Machine Learning: A Systematic Literature Review.

Word Count: