FRAUDULENT E-COMMERCE TRANSACTIONS

Members:

Adwait Dani
Ashish Singh Songara
Kiran Mayi Vijaya Kumar
Paras Sharma
Rishabh Mahajan
Sidhant Bajaj
Silja Johny

36103 Statistical Thinking for Data Science - Autumn 2024

Assessment Task 2B - Data analysis project: presentation

Prepared by: Group 6

AGENDA

1 INTRODUCTION

- RELEVANCE OF FRAUD DETECTION IN E-COMMERCE TRANSACTIONS
- DATASET OVERVIEW

2 BUSINESS CONTEXT & SCOPE PROJECT

- AIM & OBJECTIVES
- RESEARCH QUESTIONS

3 METHODOLOGY

- DATA PRE-PROCESSING
- EDA
- MODEL SELECTION & TECHNIQUES

4 RESULT

- PERFORMANCE COMPARISION

5 CONCLUSION

INTRODUCTION

RELEVANCE OF FRAUD DETECTION IN E-COMMERCE TRANSACTIONS

01

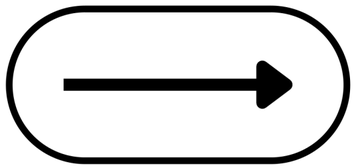
- Fraudulent activities include money laundering, cyberattacks, fraudulent banking claims, forged bank checks, identity theft, and many such illegal practices.
- As a result, organizations implement modern fraud detection and prevention technologies and risk management strategies to combat growing fraudulent transactions across diverse platforms.
- These techniques apply adaptive and predictive analytics (i.e., machine learning) to create a fraud risk score along with real-time monitoring of fraudulent events.
- This allows continuous monitoring of transactions and crimes in real-time.
- It also helps decipher new and sophisticated preventive measures via automation.

INTRODUCTION

DATASET OVERVIEW

01

- Number of Transactions: (23,634)
- Features: I6



Feature Details

1. **Transaction ID:** A unique identifier for each transaction.
2. **Customer ID:** A unique identifier for each customer.
3. **Transaction Amount:** The total amount of money exchanged in the transaction.
4. **Transaction Date:** The date and time when the transaction took place.
5. **Payment Method:** The method used to complete the transaction (e.g., credit card, PayPal, etc.).
6. **Product Category:** The category of the product involved in the transaction.
7. **Quantity:** The number of products involved in the transaction.
8. **Customer Age:** The age of the customer making the transaction.
9. **Customer Location:** The geographical location of the customer.
10. **Device Used:** The type of device used to make the transaction (e.g., mobile, desktop).
11. **IP Address:** The IP address of the device used for the transaction.
12. **Shipping Address:** The address where the product was shipped.
13. **Billing Address:** The address associated with the payment method.
14. **Is Fraudulent:** A binary indicator of whether the transaction is fraudulent (1 for fraudulent, 0 for legitimate).
15. **Account Age Days:** The age of the customer's account in days at the time of the transaction.
16. **Transaction Hour:** The hour of the day when the transaction occurred.

BUSINESS CONTEXT & SCOPE PROJECT

AIM & OBJECTIVES

02

AIM

To develop and evaluate effective fraud detection model using machine learning algorithms to accurately identify fraudulent transactions in e-commerce datasets, thereby enhancing security and trust in online transactions.

OBJECTIVE

- Initial objectives involve understanding the dataset's structure, identifying anomalies, and cleansing the data to ensure its integrity.
- Subsequently, exploratory data analysis aims to uncover patterns, distributions, and correlations within the dataset.
- Utilising machine learning models, such as Random Forest Classifier, the project aims to develop predictive model for fraud detection, evaluating their performance using appropriate metrics.

02 BUSINESS CONTEXT & SCOPE PROJECT

RESEARCH QUESTIONS

- What are the most important features or characteristics that contribute to the likelihood of a transaction being fraudulent?
- Are there specific product categories or payment methods more prone to fraudulent activity?
- Can we identify distinct profiles or characteristics of customers who are more likely to be involved in fraudulent transactions?
- Can we build an accurate machine learning model to predict whether a transaction is fraudulent or not?

03

METHODOLOGY

DATA PRE-PROCESSING

Handling Missing Values

The dataset had no missing values, indicating no further action was required in this regard.

Handling Duplicate Rows

Duplicate rows were checked, found to be absent, ensuring data integrity.

Dropping columns

Certain columns deemed irrelevant for analysis, including "Transaction ID", "Customer ID", "Customer Location", "Transaction Date", "IP Address", "Shipping Address", and "Billing Address", were dropped.

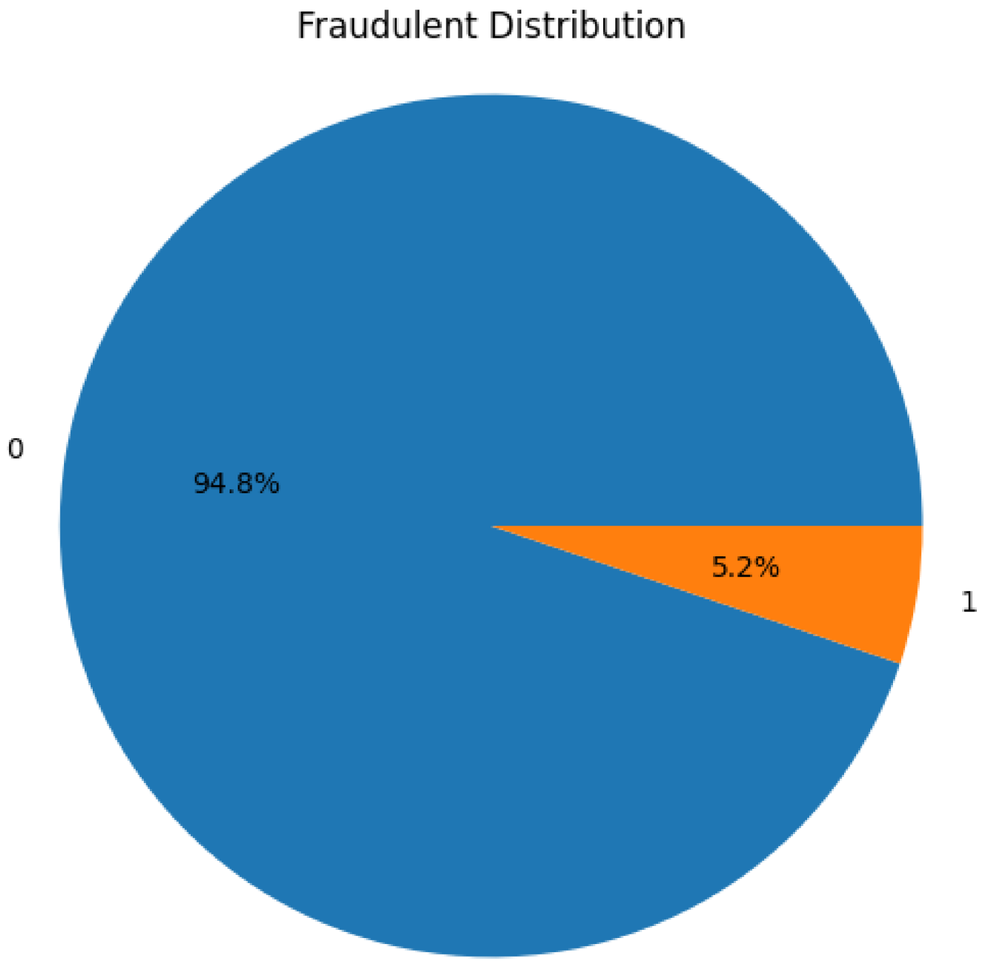
Handling Categorical Data

Categorical columns such as "Payment Method", "Product Category", and "Device Used" were mapped to numerical values to facilitate machine learning model training

03

EXPLORATORY DATA ANALYSIS (EDA)

Fraudulent Distribution



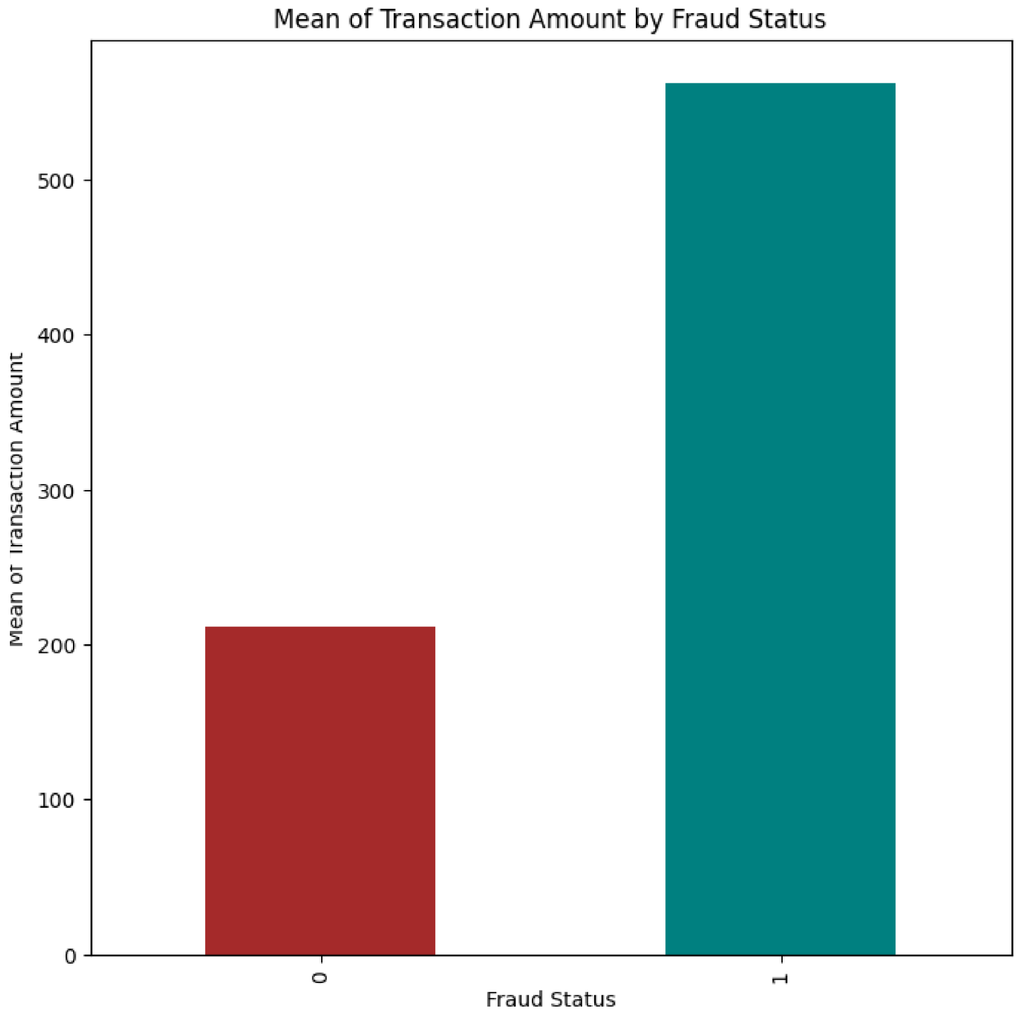
Fraudulent Distribution

- The pie chart demonstrates that **5.2%** of transactions in our dataset are labeled as fraudulent.
- This **relatively low** prevalence of fraud is typical for e-commerce transactions but still demands attention for effective detection and prevention strategies.

03

EXPLORATORY DATA ANALYSIS (EDA)

Transaction Amount Distribution



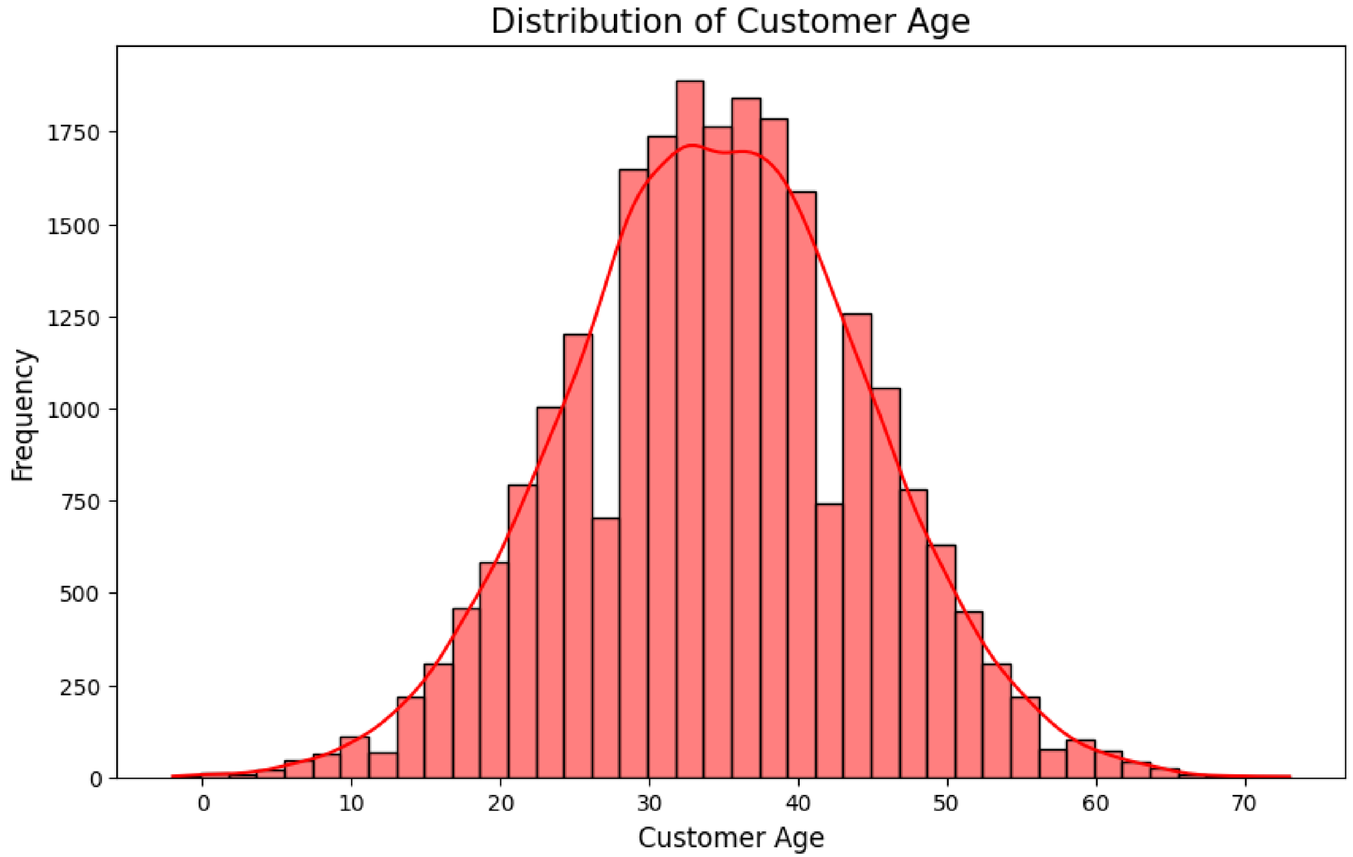
Transaction Amount Distribution

- The distribution of transaction amounts is **positively skewed**, with a mean transaction amount of approximately **$229.37**.
- This suggests that fraudulent activity tends to involve transactions of **larger monetary value**.

03

EXPLORATORY DATA ANALYSIS (EDA)

Customer Age Distribution



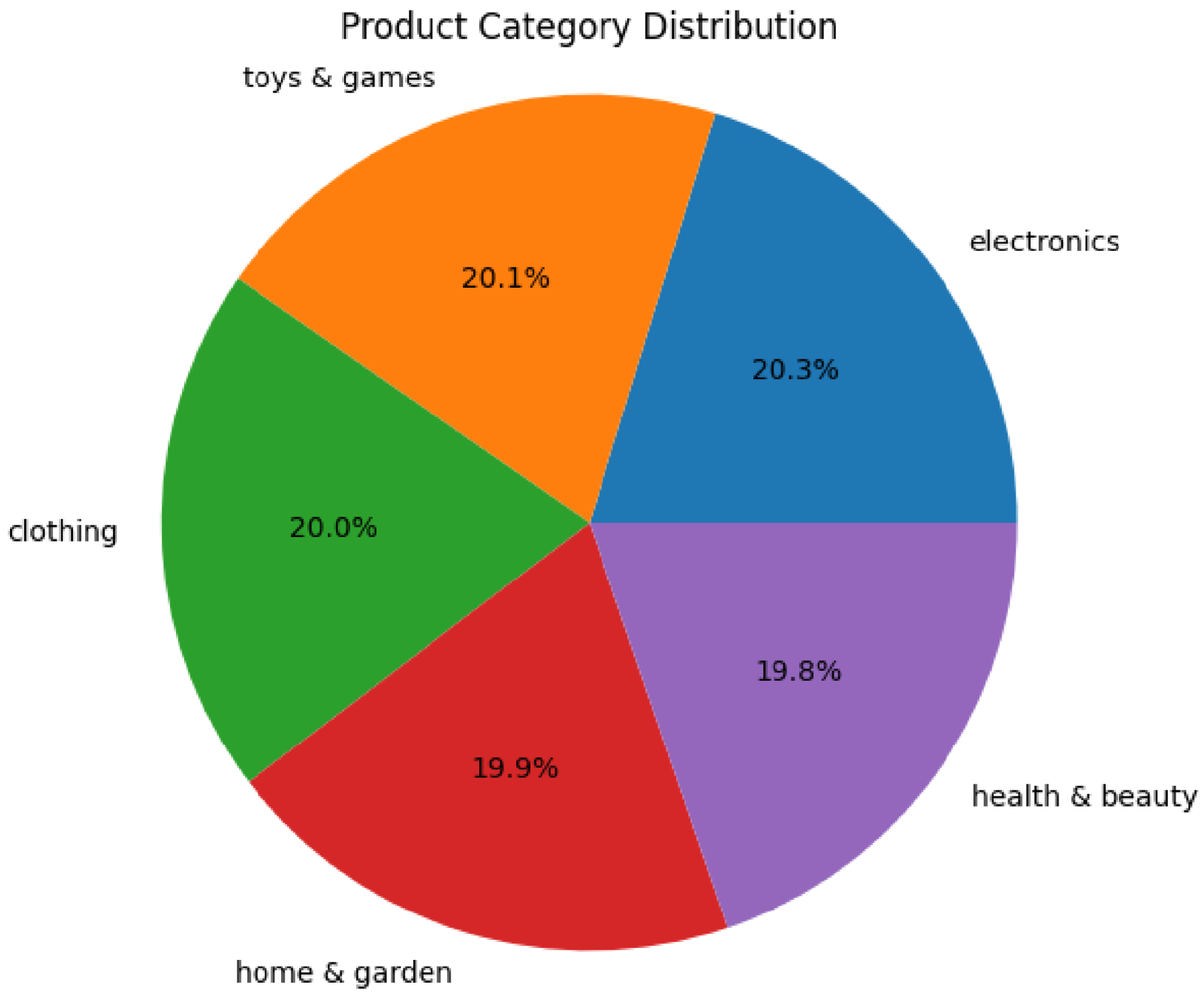
Customer Age Distribution

- The distribution of customer ages appears to be **relatively normally distributed** around a mean age of approximately **34.56 years.**
- This suggests a **diverse customer base across different age groups**.

03

EXPLORATORY DATA ANALYSIS (EDA)

Product Category Distribution



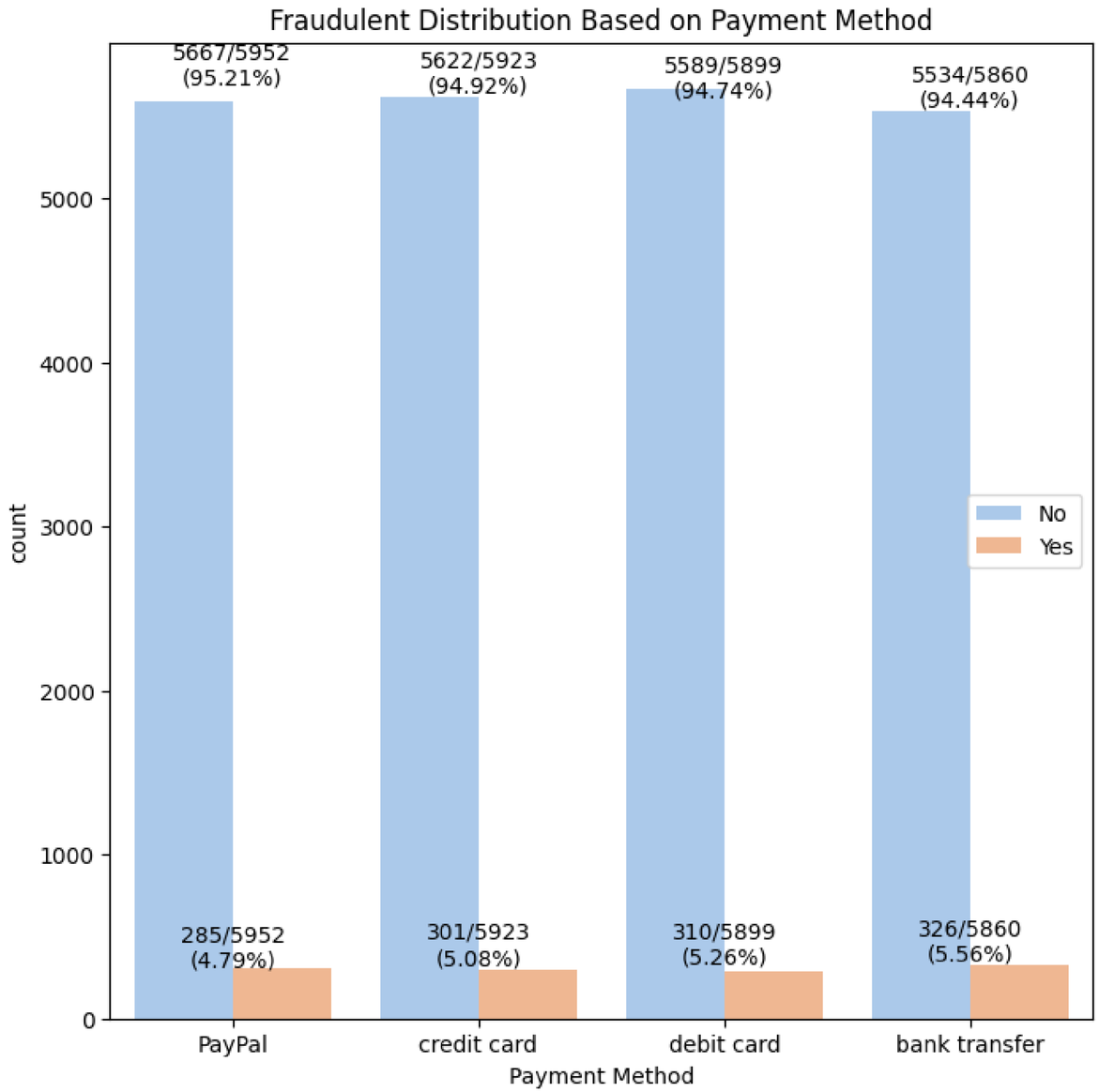
Product Category Distribution

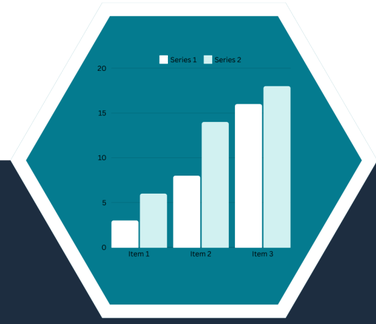
- The dataset contains transactions across various product categories, including home & garden, electronics, toys & games, clothing, and health & beauty.
- The distribution of transactions among these categories is **relatively balanced, with no single category dominating the dataset.**

03

EXPLORATORY DATA ANALYSIS (EDA)

Payment Method Distribution





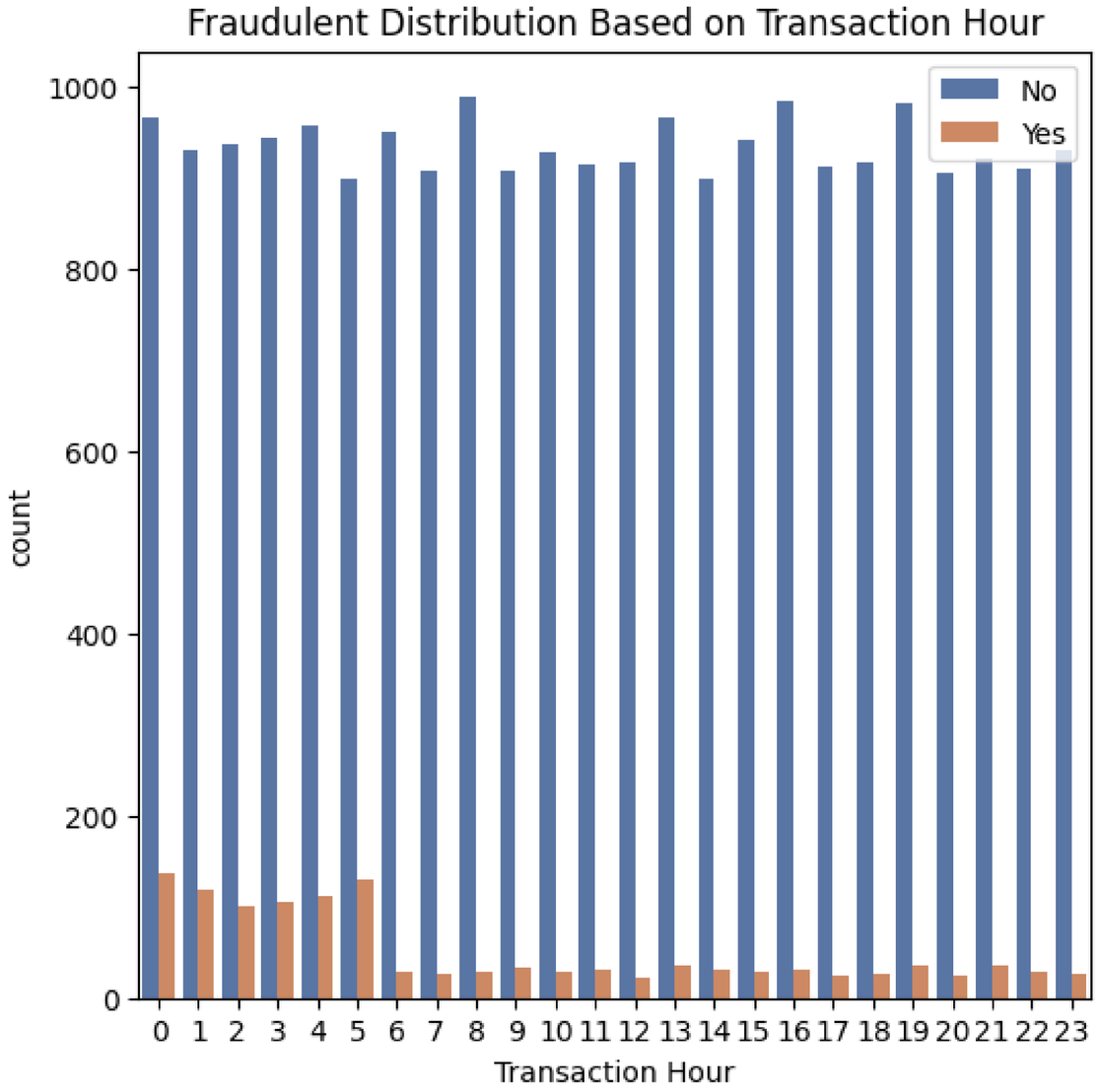
Payment Method Distribution

- **Bank Transfer** has the highest percentage of fraudulent transactions at **5.56%,** followed by Debit Card, Credit Card, and PayPal.

03

EXPLORATORY DATA ANALYSIS (EDA)

Transaction Hour Distribution



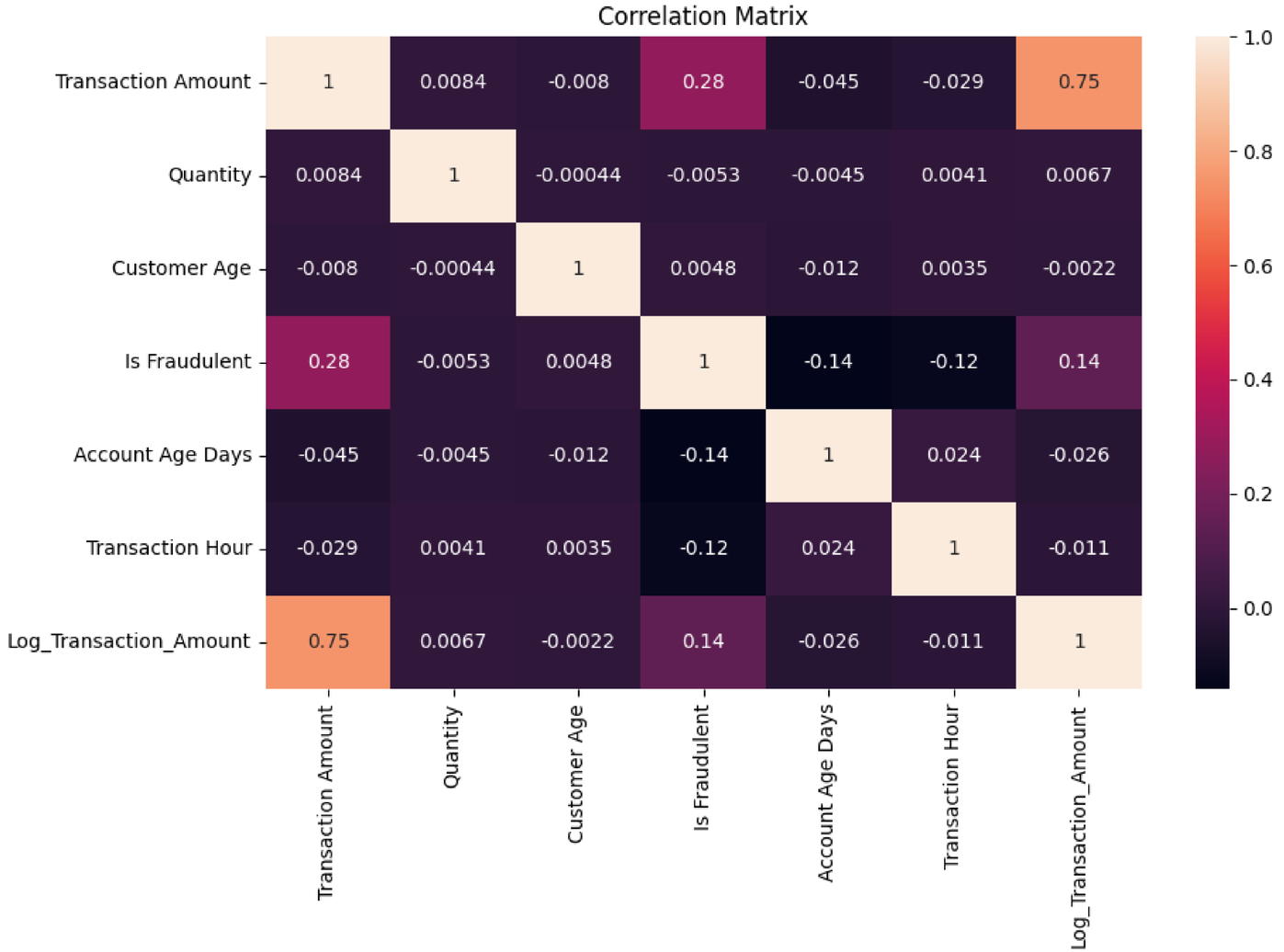
Transaction Hour Distribution

- The **majority** of fraudulent transactions tend to occur **during the early hours of the day**, specifically **between midnight (hour 0) and 5 AM.**

03

EXPLORATORY DATA ANALYSIS (EDA)

Correlation Analysis



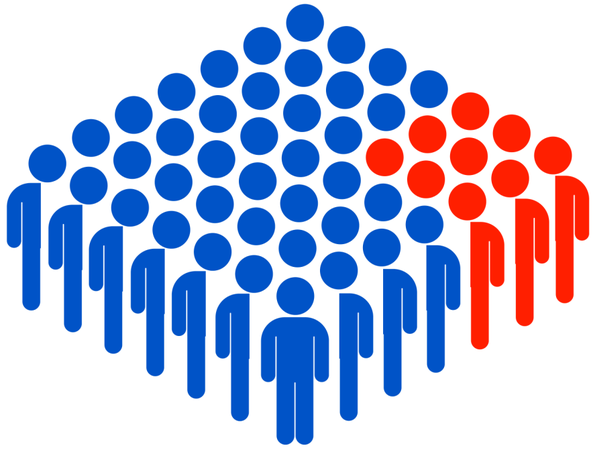
Correlation Analysis

- The heat-map shows that **transaction amount is the most important characteristic that contribute to** likelihood transaction being fraudulent or not, but the correlation between the features doesn't show much.

03 MODEL SELECTION & TECHNIQUES

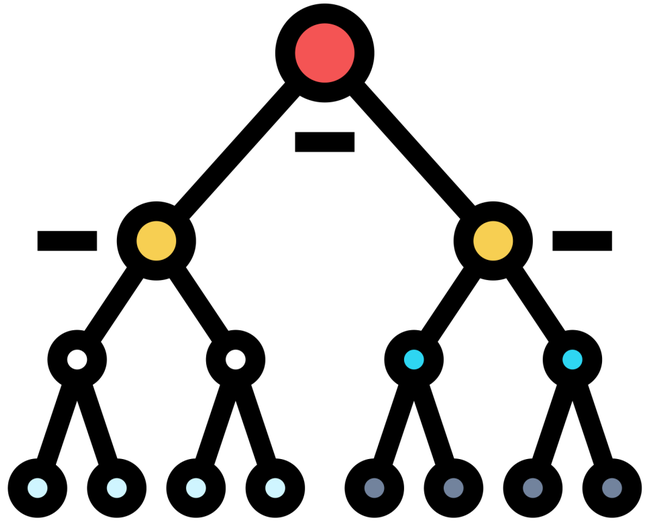
01

Techniques - RUS & SMOTE



02

MODEL - Random Forest Classifier

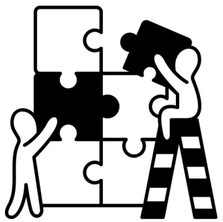


03

TECHNIQUES

Random Under Sampler

- Addresses class imbalance by randomly removing instances from the majority class.
- Reduces the number of instances in the majority class to match the number of instances in the minority class.
- Helps mitigate the impact of class imbalance on model performance by creating a more balanced dataset for training.



SMOTE

- Addresses class imbalance by generating synthetic samples of the minority class.
- Oversamples the minority class to achieve a balanced dataset.
- Helps improve the model's ability to detect fraudulent transactions by providing more representative training data.

03

MODEL

Random Forest Classifier

- Ensemble learning method that constructs multiple decision trees during training.
- Combines the predictions of individual trees to make more accurate predictions.
- Suitable for classification tasks, including fraud detection in e-commerce transactions.
- Offers robustness against overfitting and noise in the data.

Advantages:

- Handles high-dimensional data well.
- Provides feature importance analysis.
- Less prone to overfitting compared to individual decision trees.

04

RESULTS

PERFORMANCE COMPARISION

Random Under Sampler (RUS):

- Utilizes random under-sampling to balance the class distribution by reducing the majority class instances.
- Achieves an **overall accuracy of 77%** on the test set.
- Precision for **detecting fraudulent transactions (class 1) is 78%**, indicating the percentage of correctly classified fraud cases among all predicted fraud cases.
- **Recall for class 1 is 74%**, representing the percentage of actual fraud cases that were **correctly identified** by the model.
- F1-score, which combines precision and recall into a **single metric, is 0.76 for class 1.**

	precision	recall	f1-score	support
0	0.76	0.80	0.78	248
1	0.78	0.74	0.76	241
accuracy			0.77	489
macro avg	0.77	0.77	0.77	489
weighted avg	0.77	0.77	0.77	489

04

RESULTS

PERFORMANCE COMPARISION

SMOTE (Synthetic Minority Over-Sampling Technique):

- Addresses class imbalance by generating synthetic instances of the minority class (fraudulent transactions).
- **Achieves a higher overall accuracy of 93%** on the test set compared to RUS.
- Exhibits precision, recall, and **F1-score of 91% for class 1**, indicating a balanced performance in detecting fraudulent transactions.

	precision	recall	f1-score	support
0	0.94	0.91	0.93	4555
1	0.91	0.94	0.93	4410
accuracy			0.93	8965
macro avg	0.93	0.93	0.93	8965
weighted avg	0.93	0.93	0.93	8965

05

CONCLUSION

PERFORMANCE COMPARISION

The analysis provided valuable insights into the distribution of fraudulent transactions across different product categories, payment methods, and transaction hours.

Additionally, it shed light on the effectiveness of various features in predicting fraudulent transactions

05

CONCLUSION

The modelling phase concludes that employing Synthetic Minority Over-sampling Technique (SMOTE) for handling class imbalance in fraudulent transactions yields superior performance compared to Random Under Sampler (RUS). SMOTE demonstrates higher accuracy and balanced precision-recall scores, showcasing its effectiveness in detecting fraudulent transactions in e-commerce datasets.

This highlights the importance of data preprocessing and model selection in enhancing fraud detection systems. Continued optimization and monitoring of machine learning models are crucial for adapting to evolving fraud patterns and improving detection capabilities over time.

THANK YOU

DO YOU HAVE ANY QUESTION?