Building An Intelligent Fraud Detection System

Group 4 Capstone Project

Project Overview

Fraud trends vary by region, with West and Southern Africa facing increased scams, while Asia deals with telecommunication fraud. Commercial banks and health insurers are the most affected, with identity fraud making up 45% of cases in 2023 and expected to reach 50% by year-end. Scam-related frauds have surged by 56% in 2024, surpassing digital payment fraud, with scams now comprising 23% of fraudulent transactions. In Kenya, financial fraud is rising, exemplified by Kiwipay Kenya Limited's Ksh2.3 billion freeze due to suspected debit card fraud. The Central Bank of Kenya links the surge to ICT adoption, low security awareness, and cyber threats, stressing the need for stronger security and public education.

PROJECT OUTLINE

- Business Problem
- Objectives
- Data Source
- Visualizations
- Modelling
- Conclusion

Business Problem

- The Central Bank of Kenya (CBK) has identified key risk factors, including the rapid adoption of digital financial services, low consumer awareness of financial security, and evolving cyber threats.

- Additionally, the rise in scam-related frauds, identity theft, and card skimming has led to significant financial losses, eroding public trust in the financial sector. To address these challenges, we aim to develop a robust fraud detection and prevention model that leverages machine learning, artificial intelligence, and real-time transaction monitoring.

- This model will enhance the ability of financial institutions in Kenya to proactively detect fraudulent activities, mitigate risks, and strengthen cybersecurity measures, ensuring real-time fraud prevention and minimizing financial losses.

Objectives

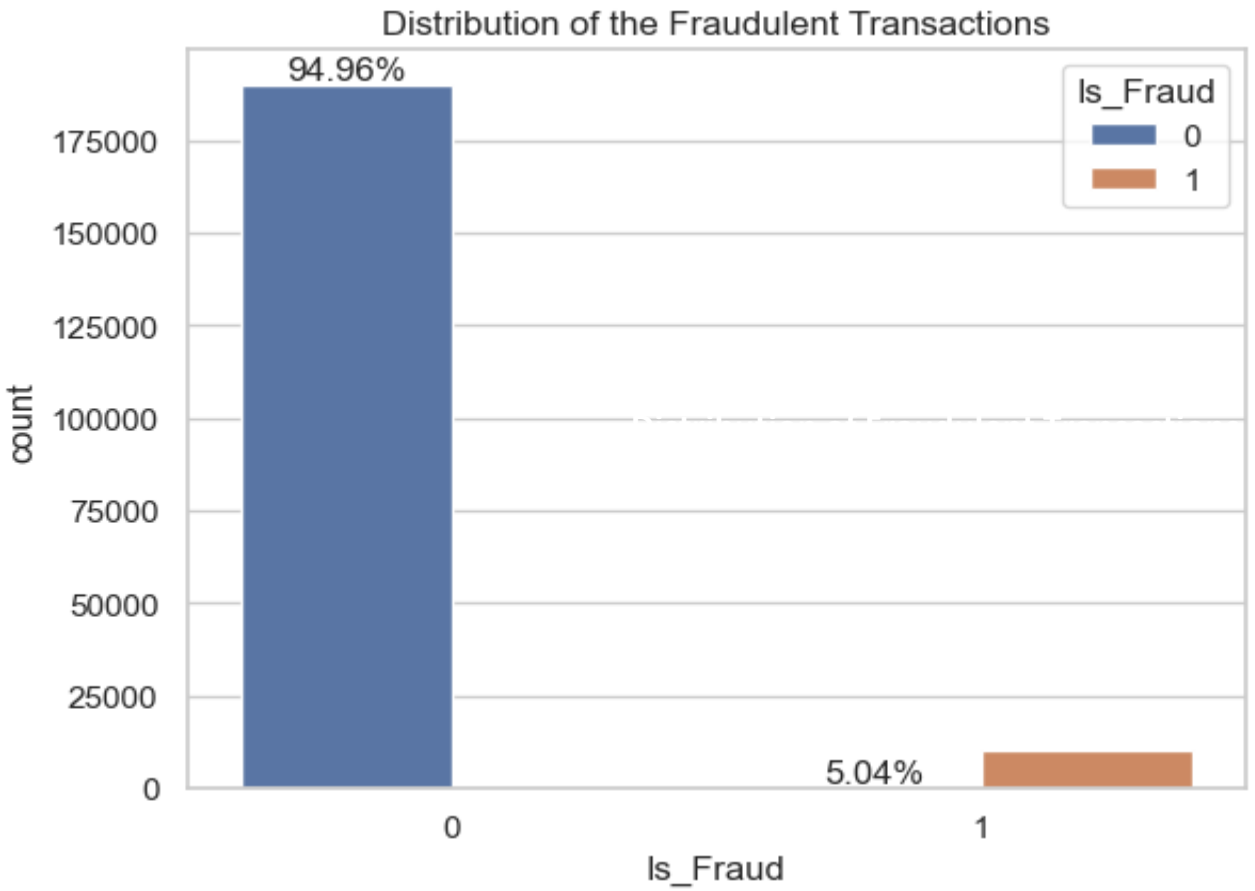
1. Analyze card transaction patterns to detect fraudulent activity.
2. Develop predictive models to accurately classify transactions as fraudulent or legitimate.
3. Examine the impact of demographics, such as age and gender, on fraud risks.
4. Identify peak fraud periods based on transaction dates and times.
5. Design a real-time fraud detection model for identifying suspicious card transactions.

Data Source and Understanding

- The dataset was obtained from [https://www.kaggle.com/datasets/marusagar/bank-transaction-fraud-detection]
- The dataset has 200000 rows and 24 columns.
- The dataset has 2 columns with Float data type, 2 column with integer data type and 20 columns with categorical data types.
- The Transaction_Date and Transaction_Time columns are indicated as object data type. For analysis and feature engineering processes, the data types for the two columns will be converted to Datetime format.

Visualization

Distribution of Fraudulent Transactions



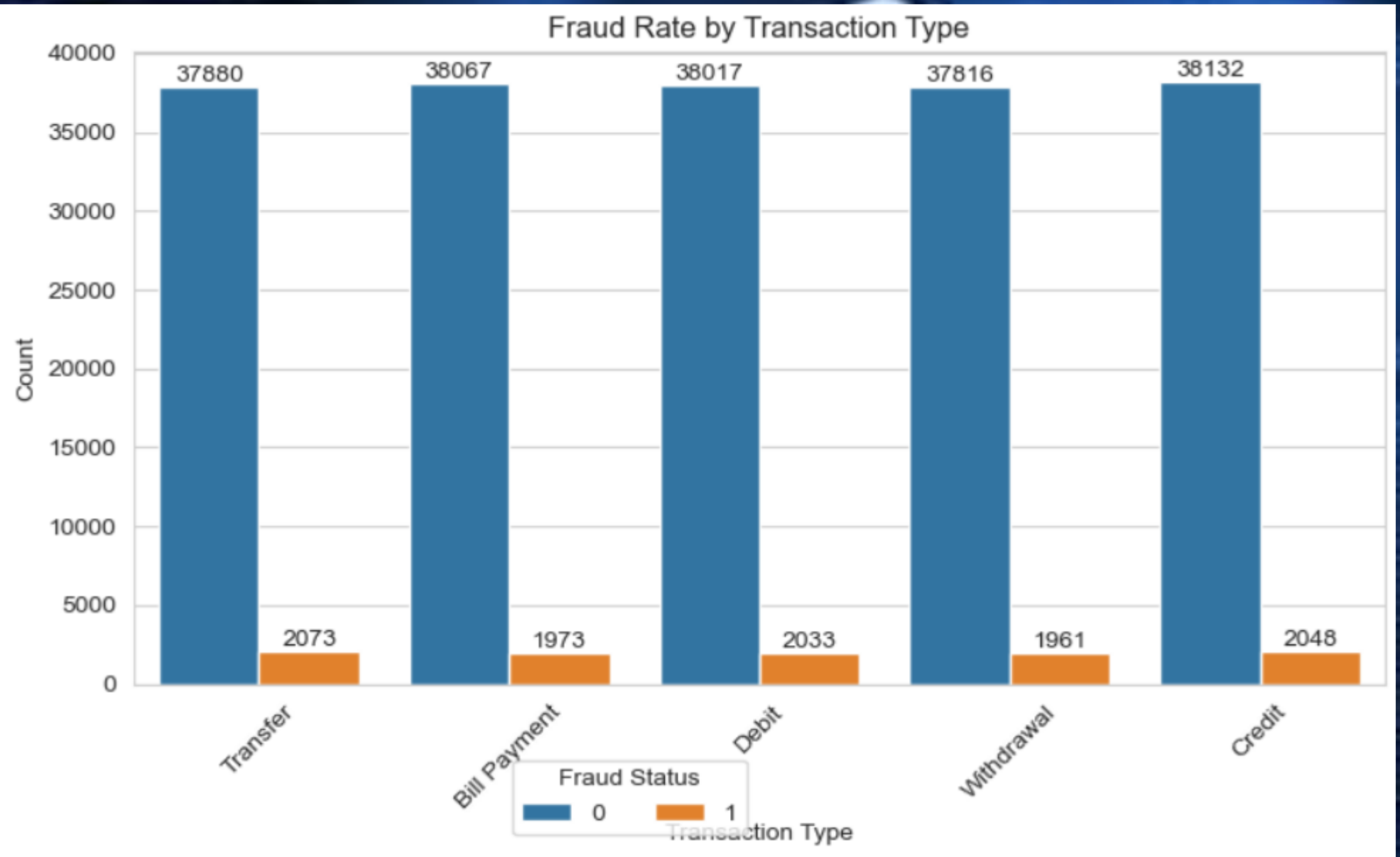
Analysis of Fraud Cases by Gender

For the class 0 indicating (Non-fraud cases) which is 94.956% of the data while for class 1 (fraud cases) 5.044% of the data.

The distribution points to a slightly higher number of reported fraud cases affecting males as compared to females.

Visualization

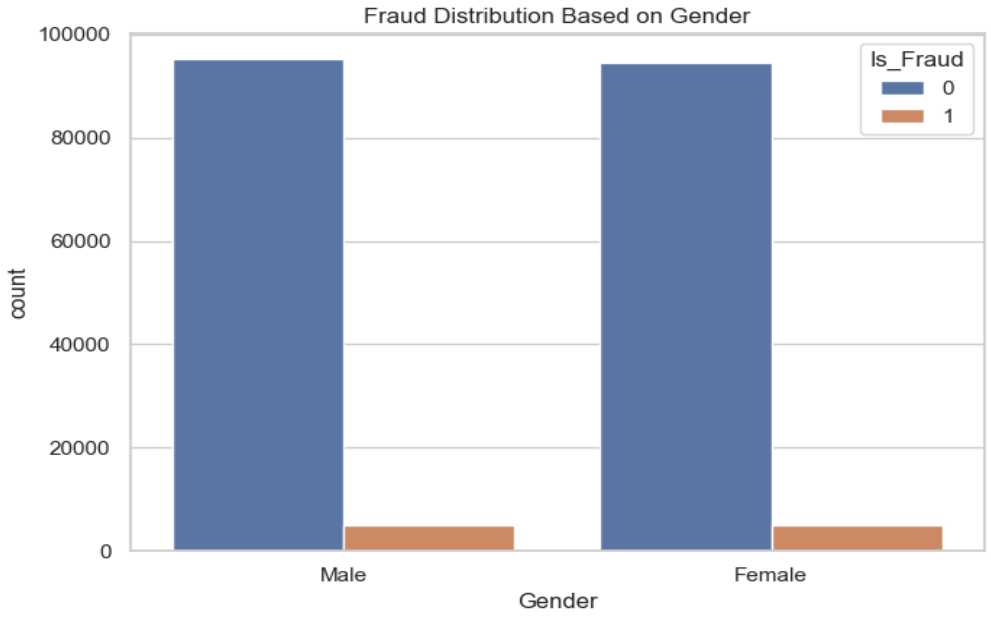
Distribution of Fraud Cases By Transaction Type

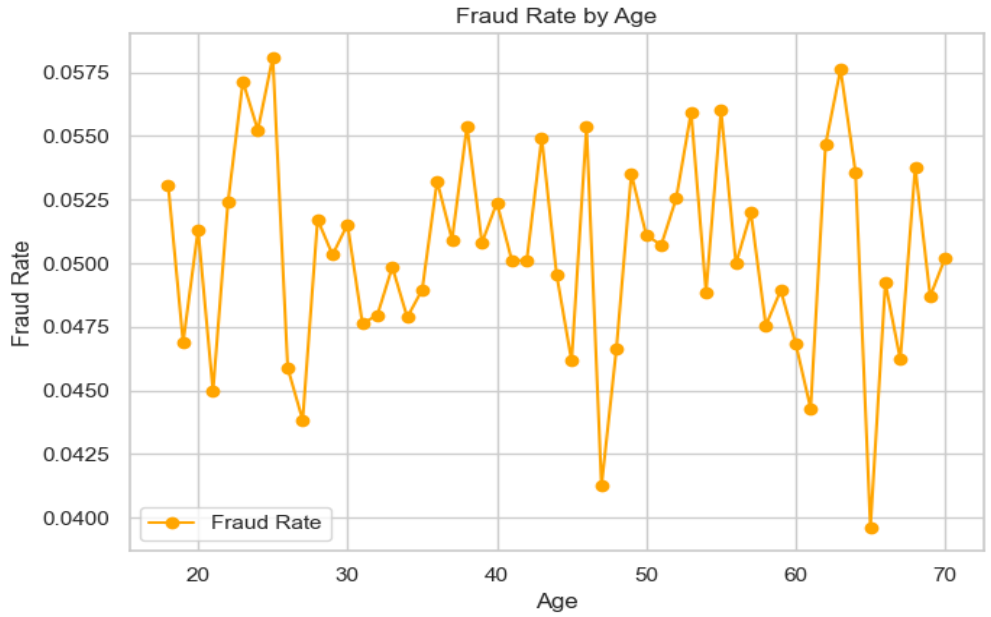


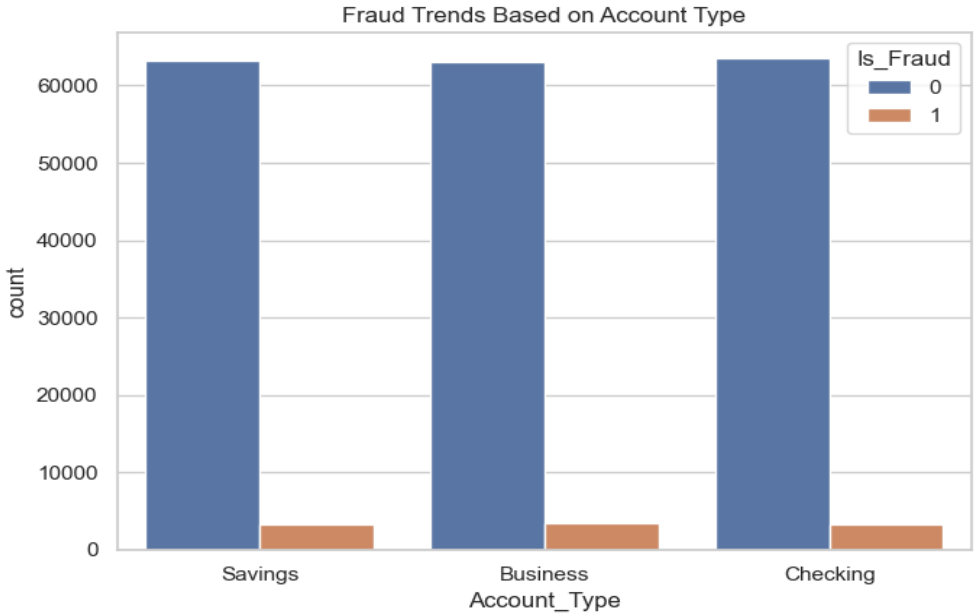
The transaction type with the highest cases of fraud is: Transfer

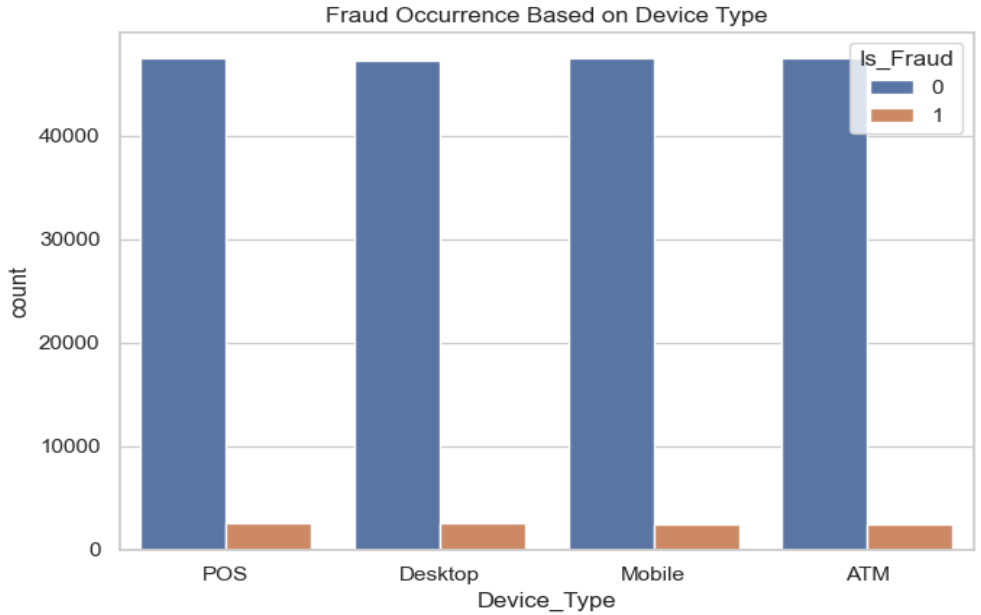
Visualization

Distribution of Fraud Cases By Gender, Age, Account Type and Device Type









The count plot shows that there are more non-fraudulent transactions for both genders, but the proportion of fraudulent transactions is slightly higher among males compared to females.

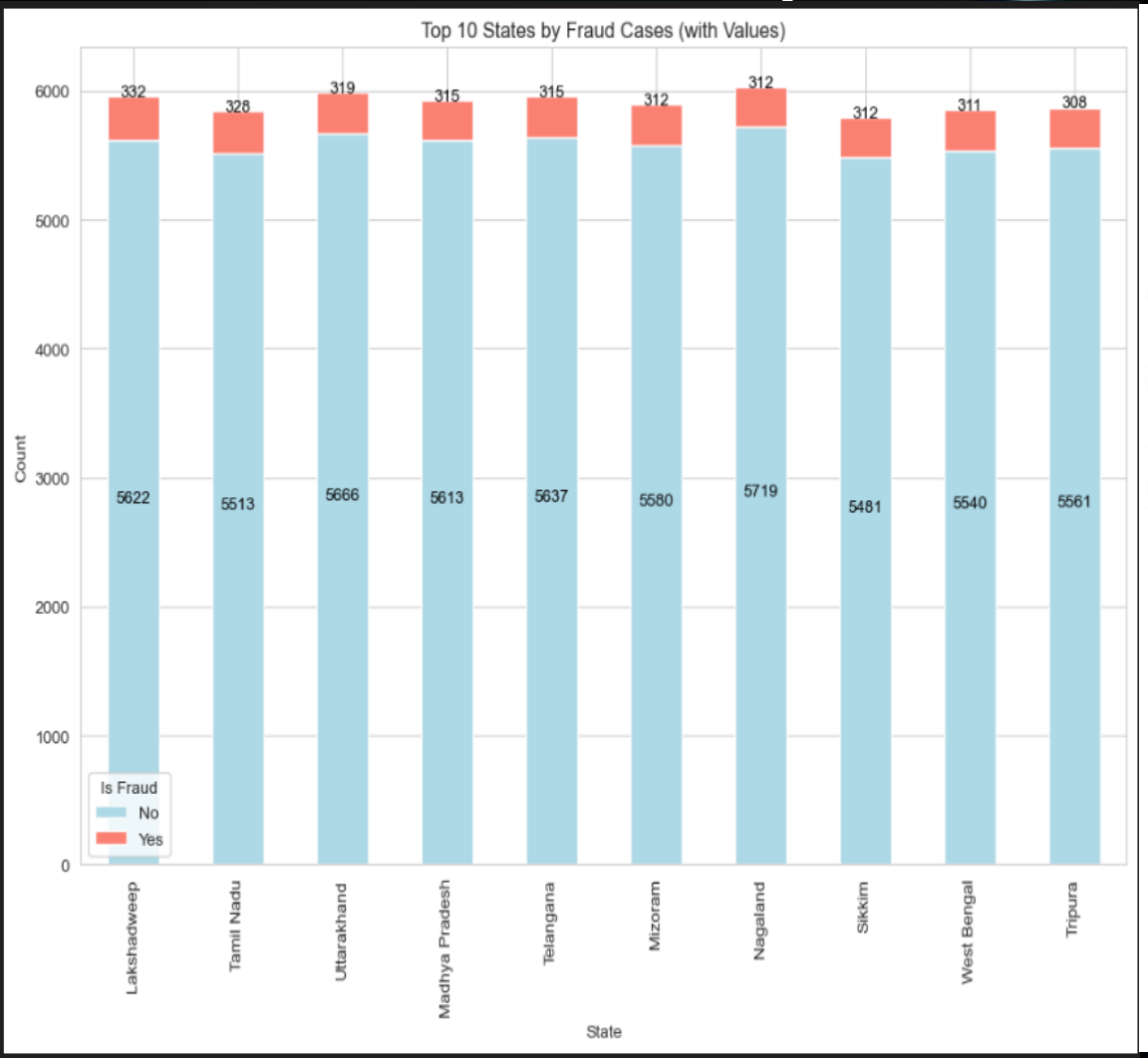
The age distribution shows a more uniform spread, with a slight concentration in the middle age range (around 30-50 years).

The distribution on fraud based on account type Fraud does not seem to be strongly biased toward any particular account type. Fraud cases are present in all account types but remain relatively low

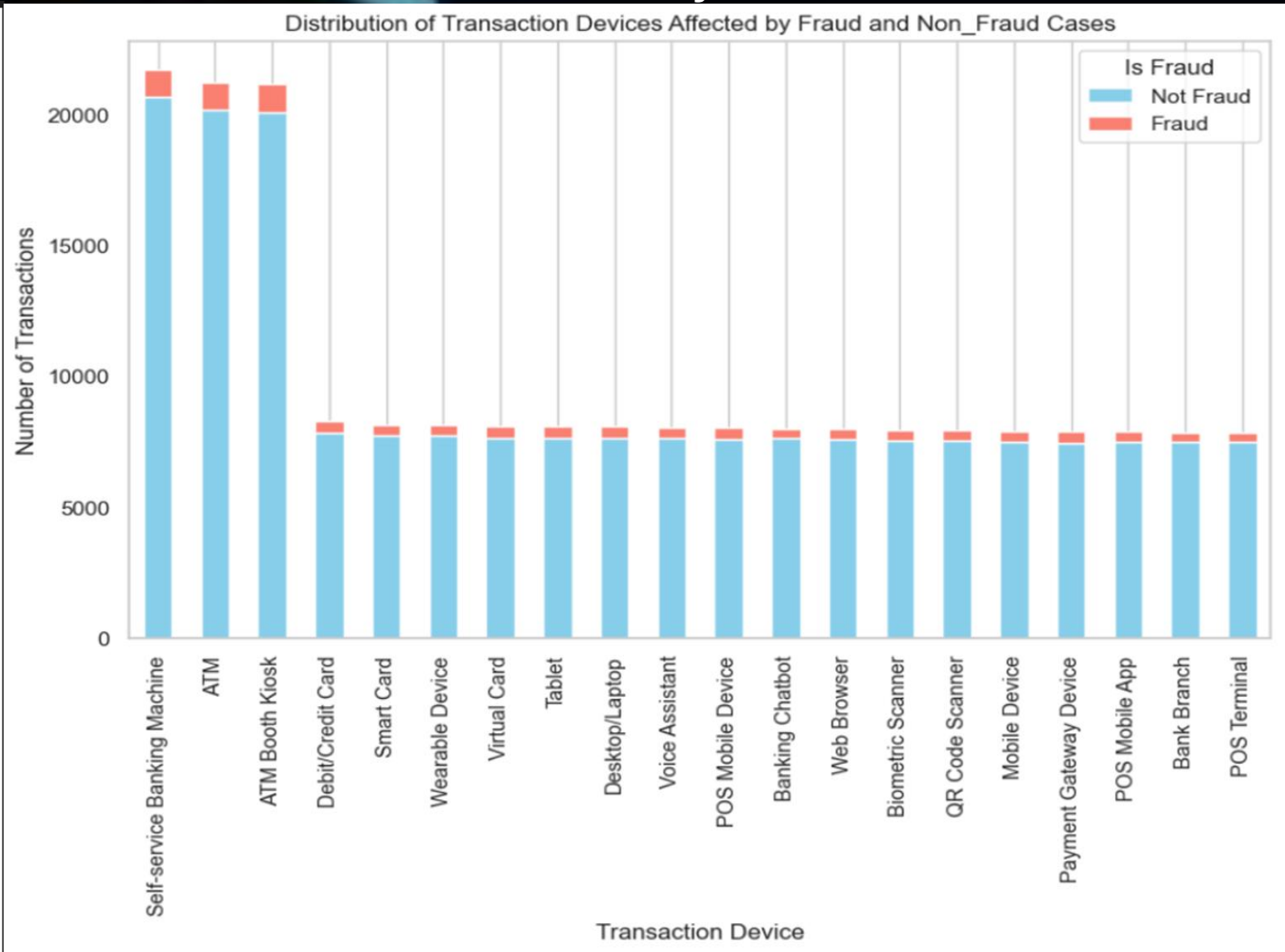
The distribution of fraudulent transactions across different devices is relatively equal, indicating that there may not be a significant difference in fraud occurrence based on the device used.

Visualization

Distribution of Fraud Cases By State



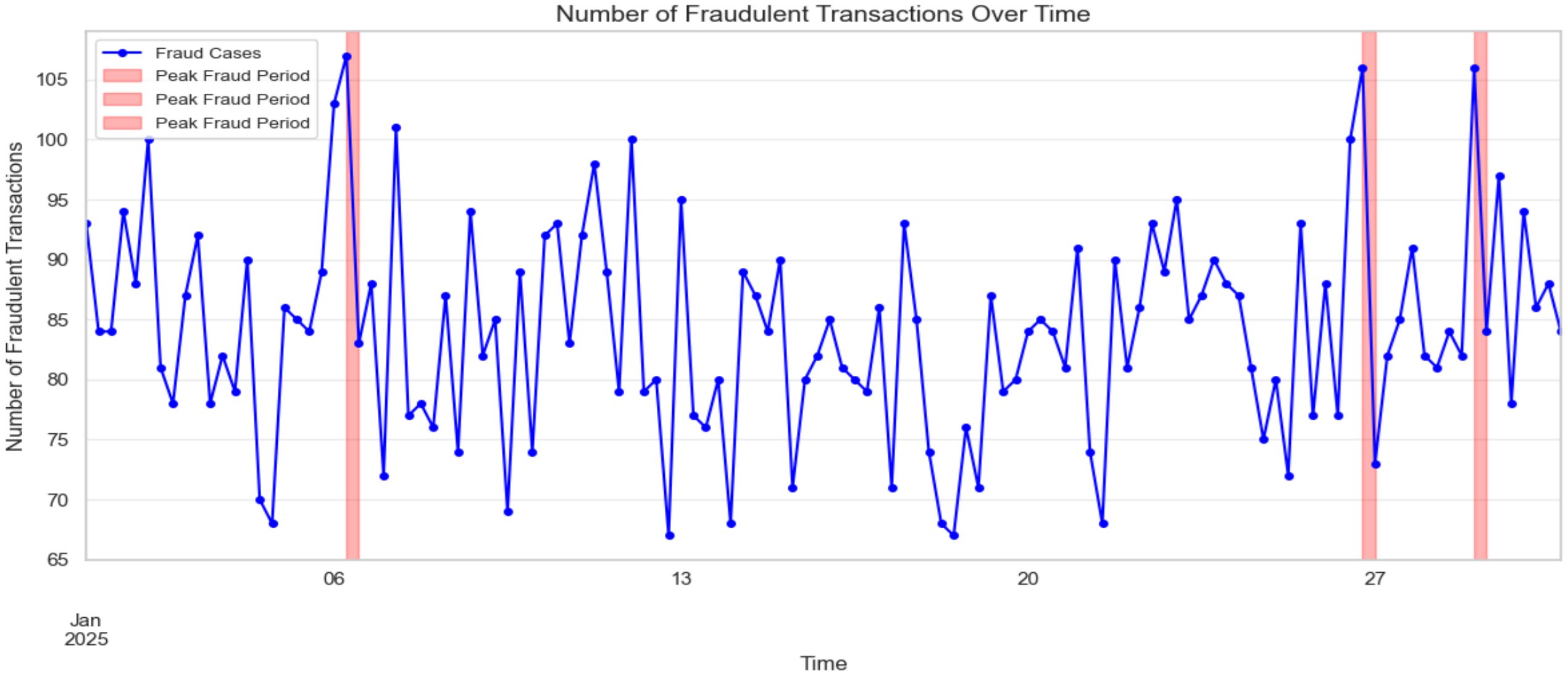
Distribution of Fraud Cases By Transaction Devices



The state with the highest fraud cases is: Lakshadwee

The transaction device with the highest fraud cases is: ATM Booth Kiosk
The transaction device with the lowest fraud cases is: POS Terminal

Distribution of Fraudulent Transactions Over Time



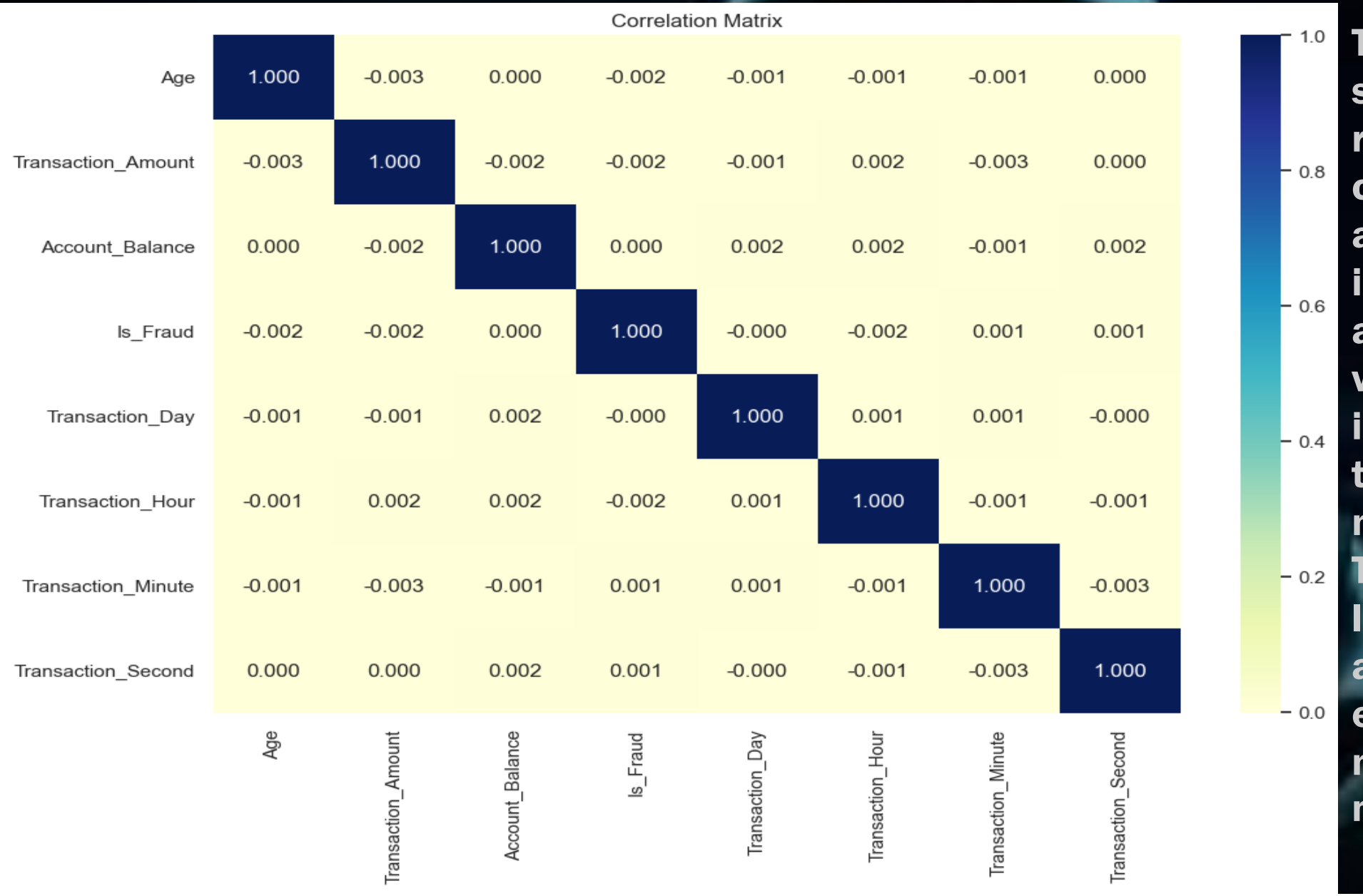
Top 3 Peak Fraud Periods:

Peak 1: Start = 2025-01-06 06:00:00, End = 2025-01-06 12:00:00, Count = 107

Peak 2: Start = 2025-01-29 00:00:00, End = 2025-01-29 06:00:00, Count = 106

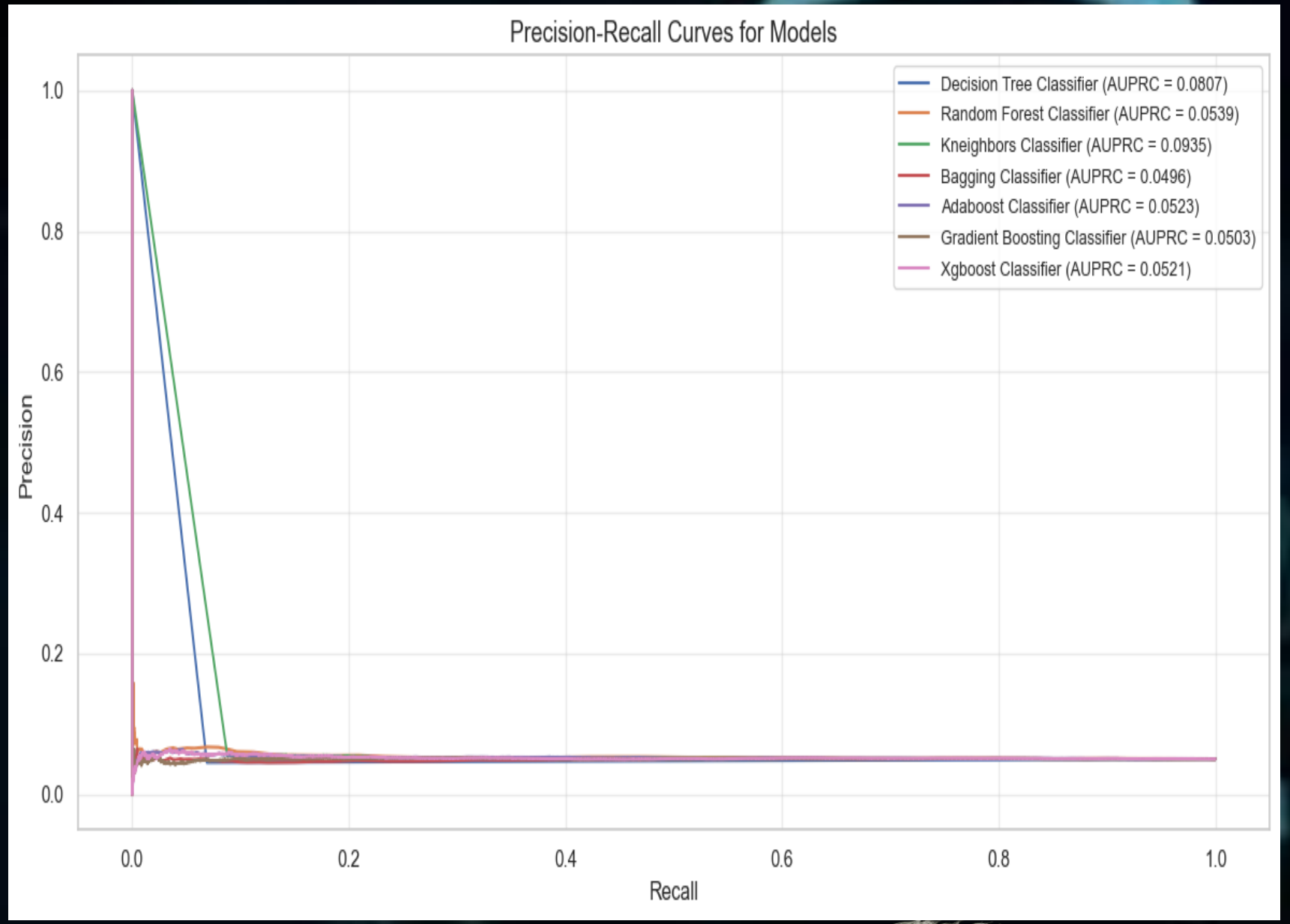
Peak 3: Start = 2025-01-26 18:00:00, End = 2025-01-27 00:00:00, Count = 106

Distribution of Fraudulent Transactions Over Time



The correlation matrix shows very weak linear relationships (correlation coefficients near zero) among variables, indicating minimal direct association. Most variables are linearly independent, suggesting that changes in one do not strongly affect others. This implies that non-linear patterns or additional feature engineering may be necessary for predictive modeling.

Modelling


Precision-Recall Curves for Models
Decision Tree Classifier (AUPRC = 0.0807)
Random Forest Classifier (AUPRC = 0.0539)
Kneighbors Classifier (AUPRC = 0.0935)
Bagging Classifier (AUPRC = 0.0496)
Adaboost Classifier (AUPRC = 0.0523)
Gradient Boosting Classifier (AUPRC = 0.0503)
Xgboost Classifier (AUPRC = 0.0521)
Precision
1.0
0.8
0.6
0.4
0.2
0.0
0.0
0.2
0.4
0.6
0.8
1.0
Recall

Conclusions

1. There seem to be a minor difference in the number of fraud cases between genders suggesting that the fraud occurrences is relatively balanced across the genders.

2. For customers at the age 19-30 and 51 - 60 years show significantly higher numbers of fraud cases which further indicates that individuals in the age groups are more vulnerable to fraud.

3. The transaction type with the highest number of frauds is Transfer with 2,073 cases reported, followed closely by credit transactions with 2048 cases.

4. The ATM Booth Kiosk, the ATM and the Self-service Machine channels posed the highest risk of fraud among the transaction devices.

5. The peak periods of fraud incidents are during holidays.

Recommendation

1. Financial institutions to use fraud by age-group analysis to perform risk assessment to inform and come up with awareness campaigns towards the targeted group to reduce chances of fraud.

2. Financial institution to do a deeper analysis on the how to establish controls to mitigate risks of fraud in areas with highest frequency including transfers and credit transactions.

3. More controls need to be established on the ATM Booth Kiosk, the ATM and the Self-service Machine as they reported the most frauds.

4. Financial institution should heighten monitoring of fraudulent activities during holidays and special days marked in the country.

5. Understanding the distribution of fraud cases by age group can aid in risk assessment and the development of targeted fraud prevention strategies. Financial institutions and security agencies can use this information to implement age-specific awareness campaigns and security measures.

6. Understanding the distribution of fraud cases by gender can help in designing targeted fraud prevention strategies. For example, if females have a higher number of fraud cases, awareness campaigns and security measures can be tailored specifically for female users.

Thank you