Building An Intelligent Fraud Detection System

Group 4 Capstone Project

Members:
Gilbert Cheruiyot, James Muthee, Vivian Muiruri, Jackline Nyaguthii, Joseph Mulwa and Mika Benson

Project Overview

- **Fraud schemes are rising,** with scam-related fraud surging by 56% in 2024 (PYMNTS, 2024).
- Scams now account for **23%** of fraudulent transactions, driven by relationship and product scams.
- Kenya faces escalating financial fraud, with Kiwipay Kenya's Ksh 2.3 billion frozen over suspected debit card fraud (Kenyan Wall Street, 2024).
- CBK (2025) links fraud growth to increased ICT adoption, low financial security awareness, and cyber threats.
- Stronger security measures and public education are essential to combat digital financial fraud.

PROJECT OUTLINE

1. Business Problem
2. Business Objectives
3. Data Source
4. Data Understanding
5. Exploratory Data Analysis
6. Data Preparation
7 Modelling
8. Evaluation
9. Conclusion
10. Recomendations

Business Problem

Rising banking fraud in Kenya has led to significant financial losses and reduced customer trust, highlighting the inadequacy of traditional detection systems, whereas AI-powered machine learning solutions can analyze past transactions, detect anomalies in real time, adapt to evolving fraud techniques, and prevent fraud before it occurs

Business Objectives

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

Data Source

The data used for this project was sourced from Kaggle

Data Understanding

1. The dataset has 200000 rows and 24 columns.

2. The dataset has 2 columns with Float data type, 2 column with integer data type and 20 columns with categorical data types.

3. The dataset has no missing values

4. The Transaction_Date and Transaction_Time columns are indicated as object data type. The data types for the two columns will be converted to Datetime format.

5. The dataset has no duplicate rows

6. The dataset has no outliers

7. The descriptive analysis of the categorical data was:

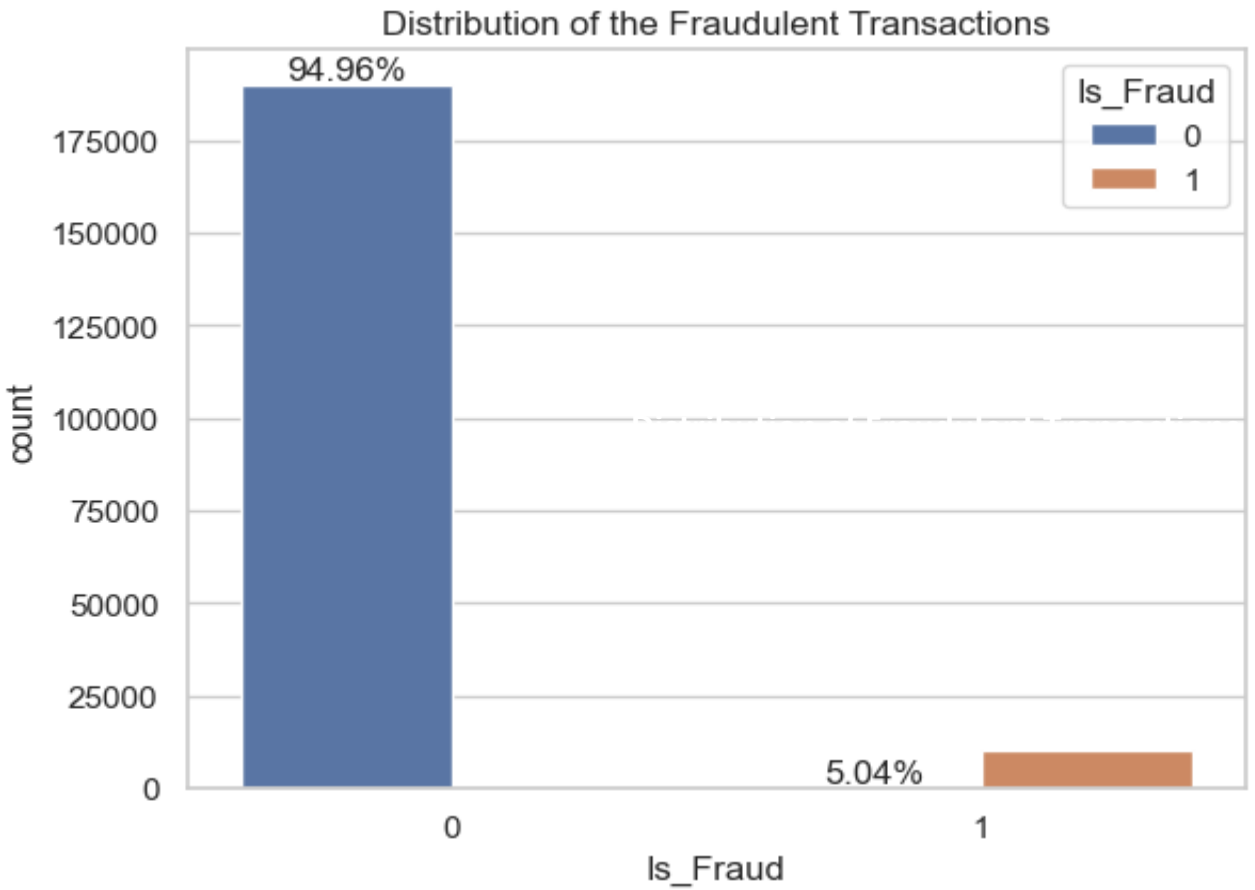
i) The mean age, transaction amount and account Balance is 44 years, 49538 INR and 53437 INR respectively.

ii) The standard deviation of the age transaction amount is 15 years, 28551 INR and 27399 INR respectively.

iii) The minimum age and maximum age is 18 and 70 years

Exploratory Data Analysis

Distribution of Fraudulent Transactions



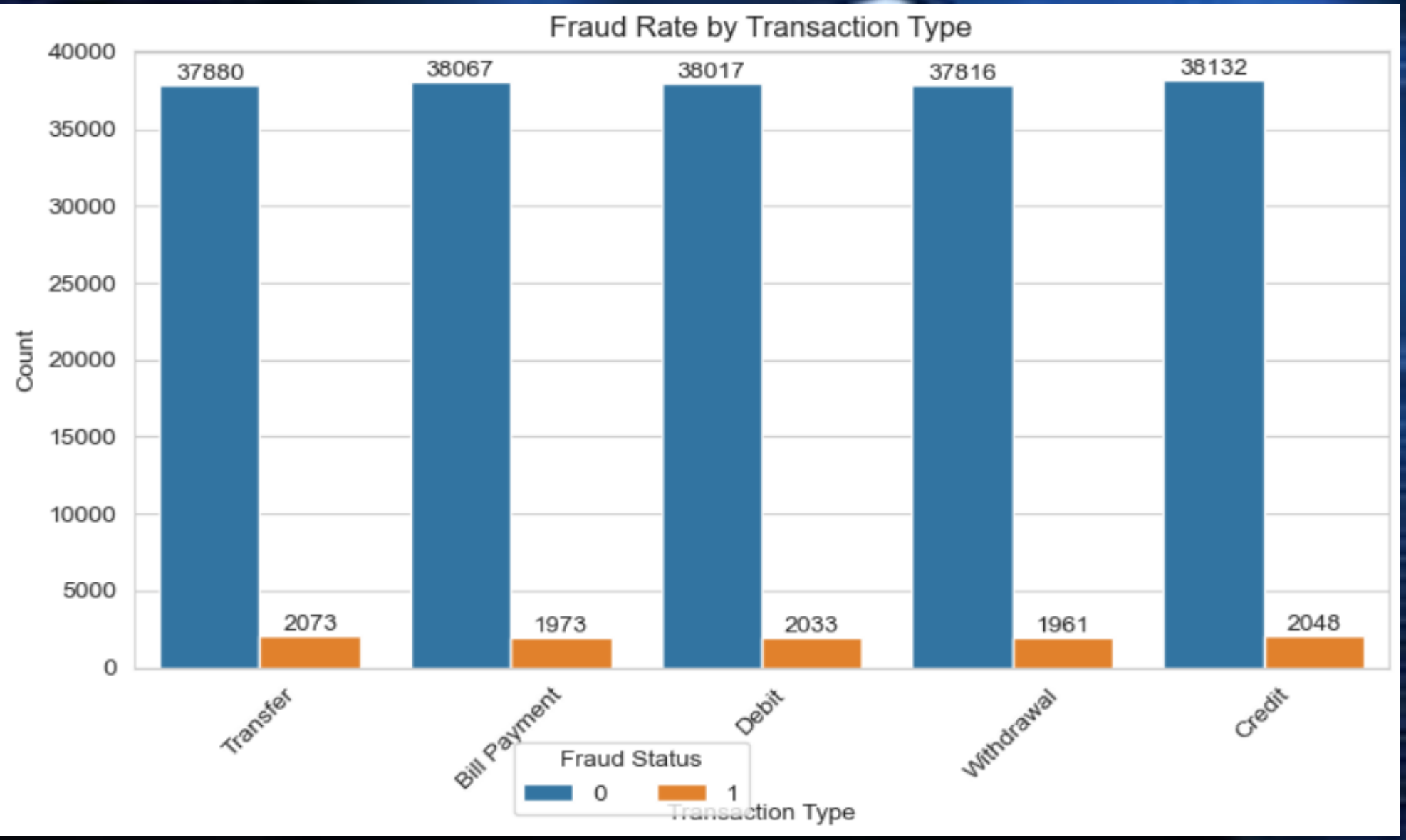
Analysis of Fraud Cases by Gender

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

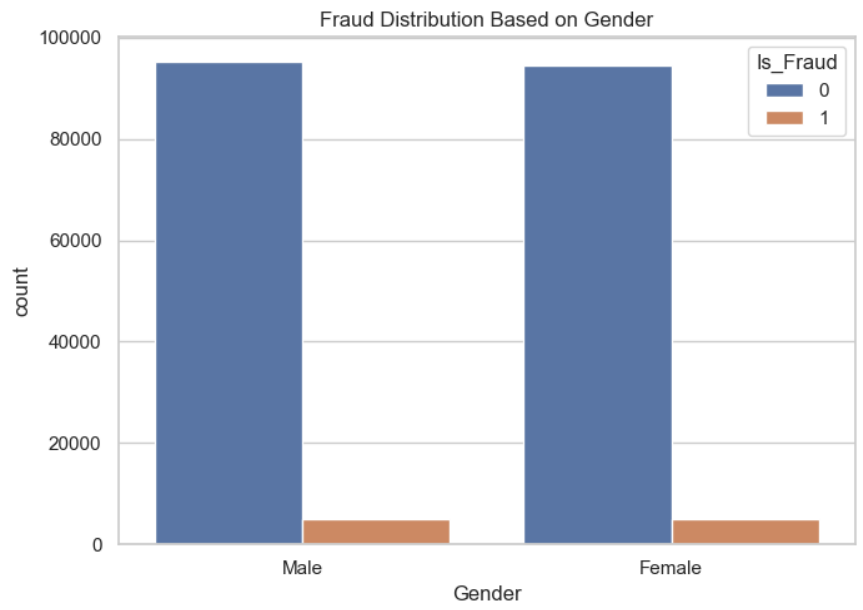
Distribution of Fraud Cases By Transaction Type

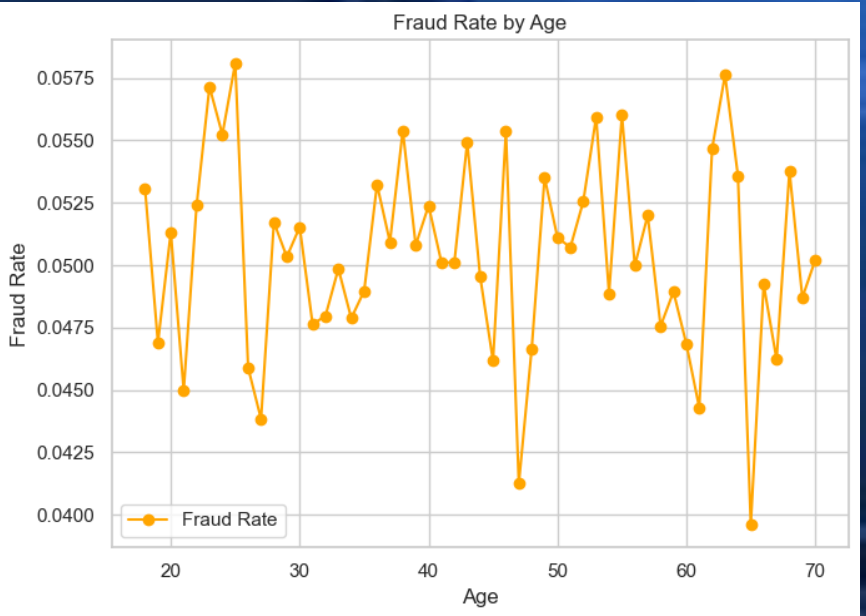


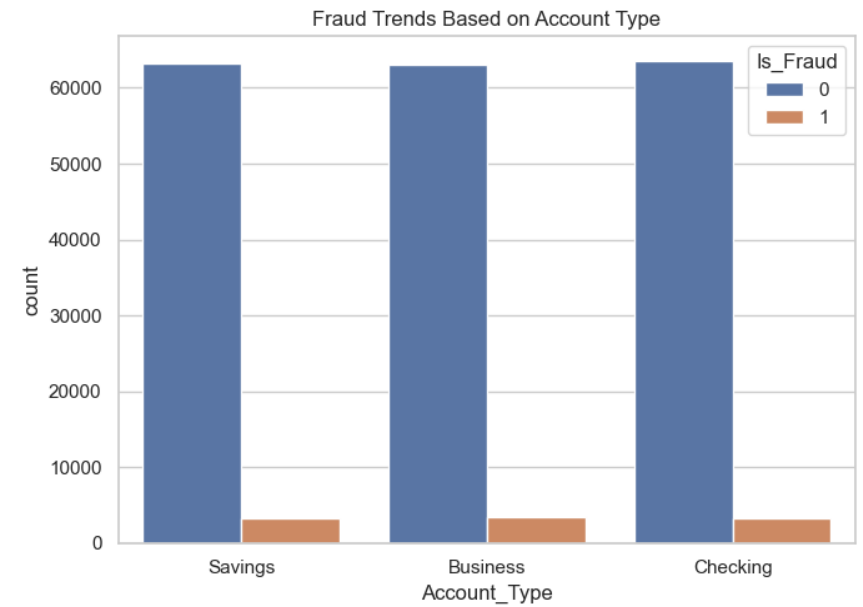
The transaction type with the highest cases of fraud is: Transfer

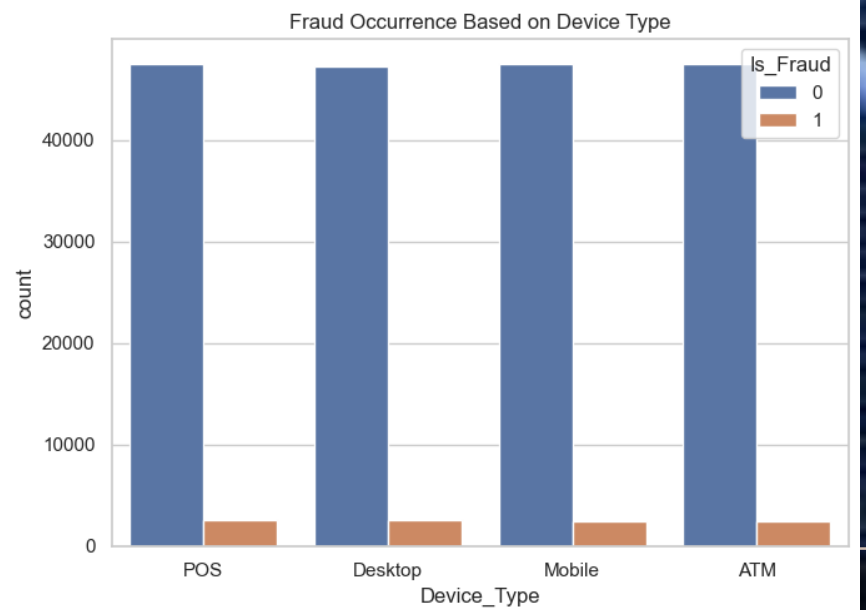
Visualization

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









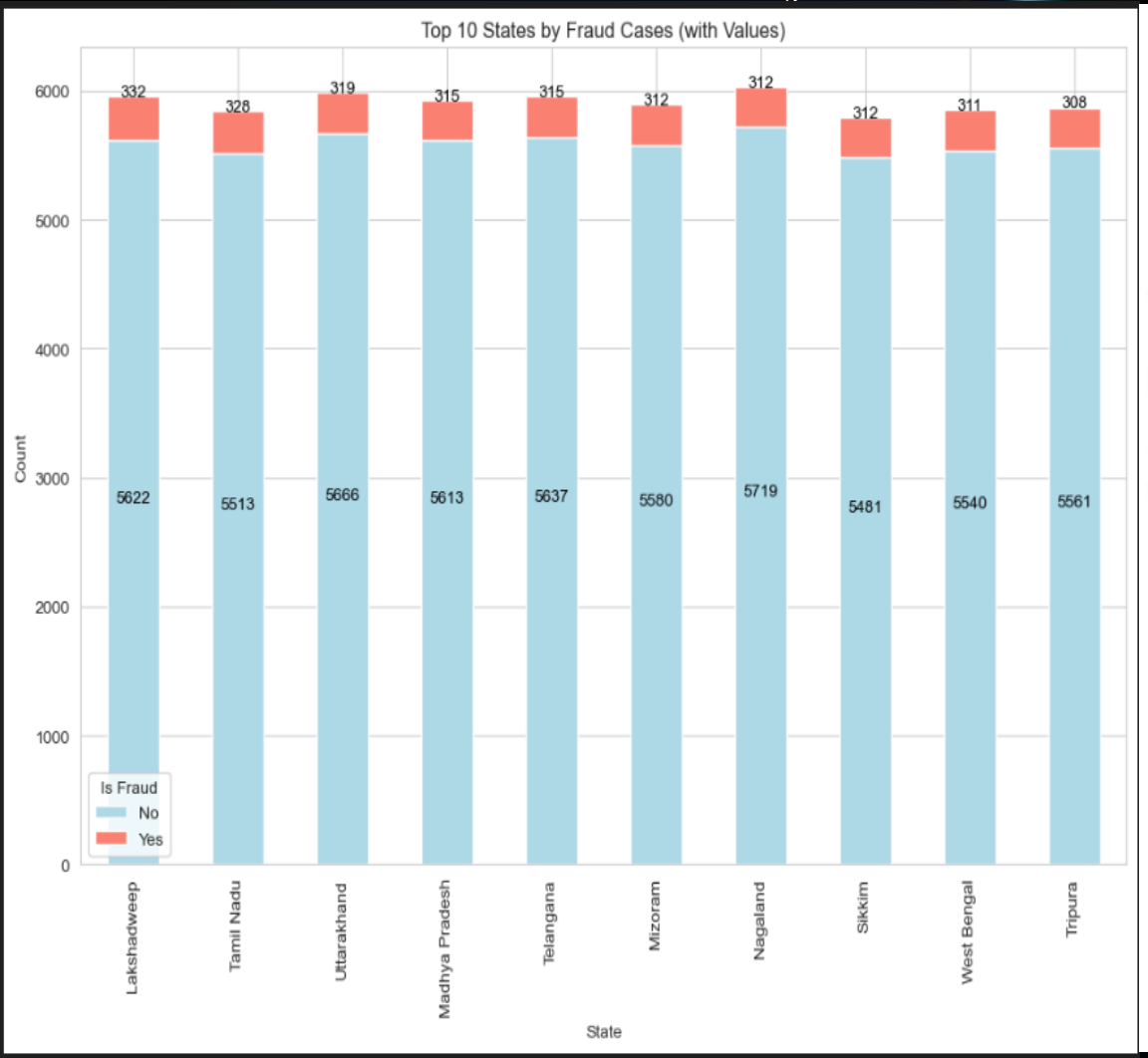
1.The proportion of fraudulent transactions is slightly higher among males compared to females.

2. The distribution on fraud based on account type Fraud does not seem to be strongly biased toward any particular account type.

3. The distribution of fraudulent transactions across different devices is relatively equal.

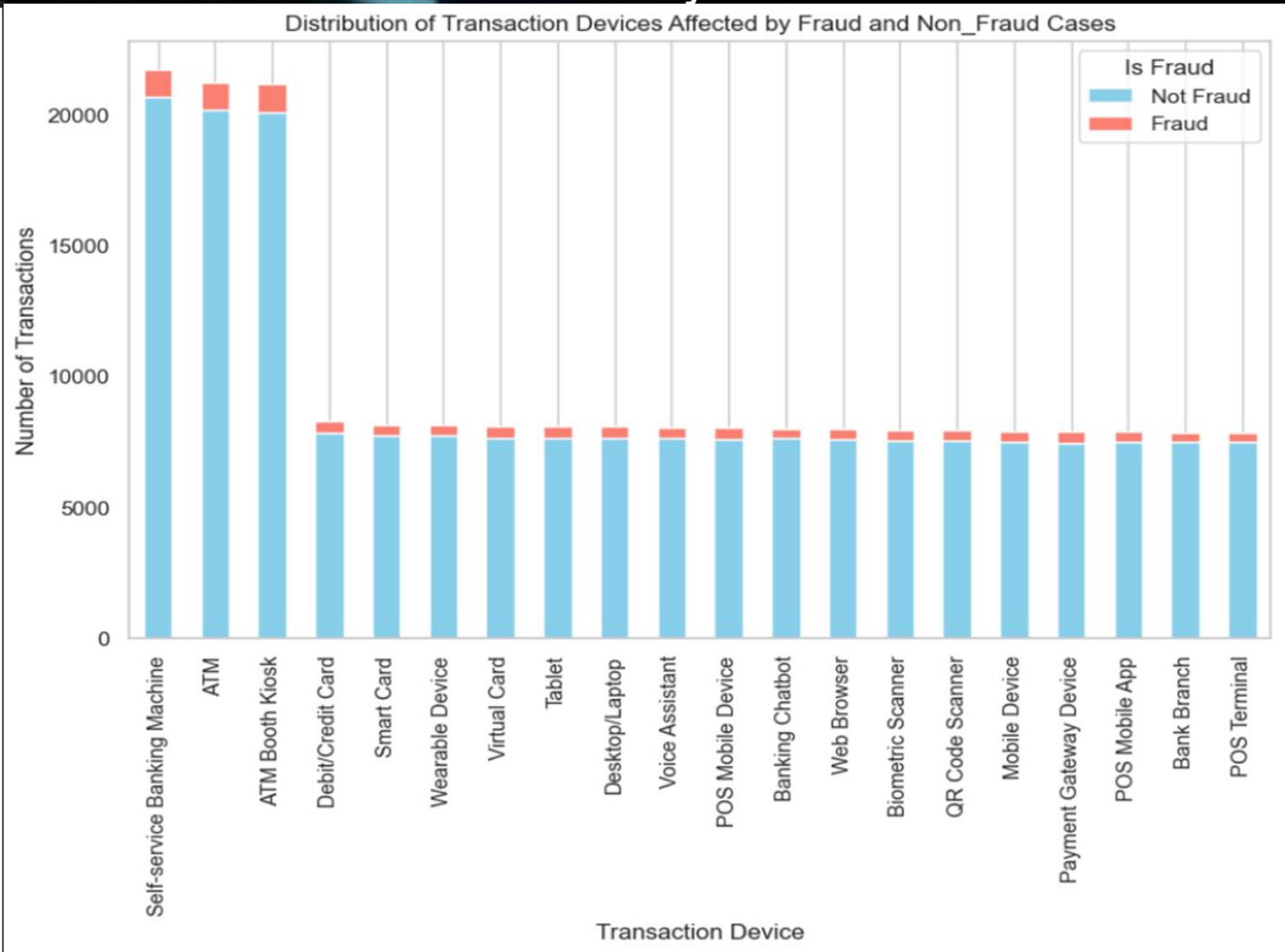
Visualization

Distribution of Fraud Cases By State



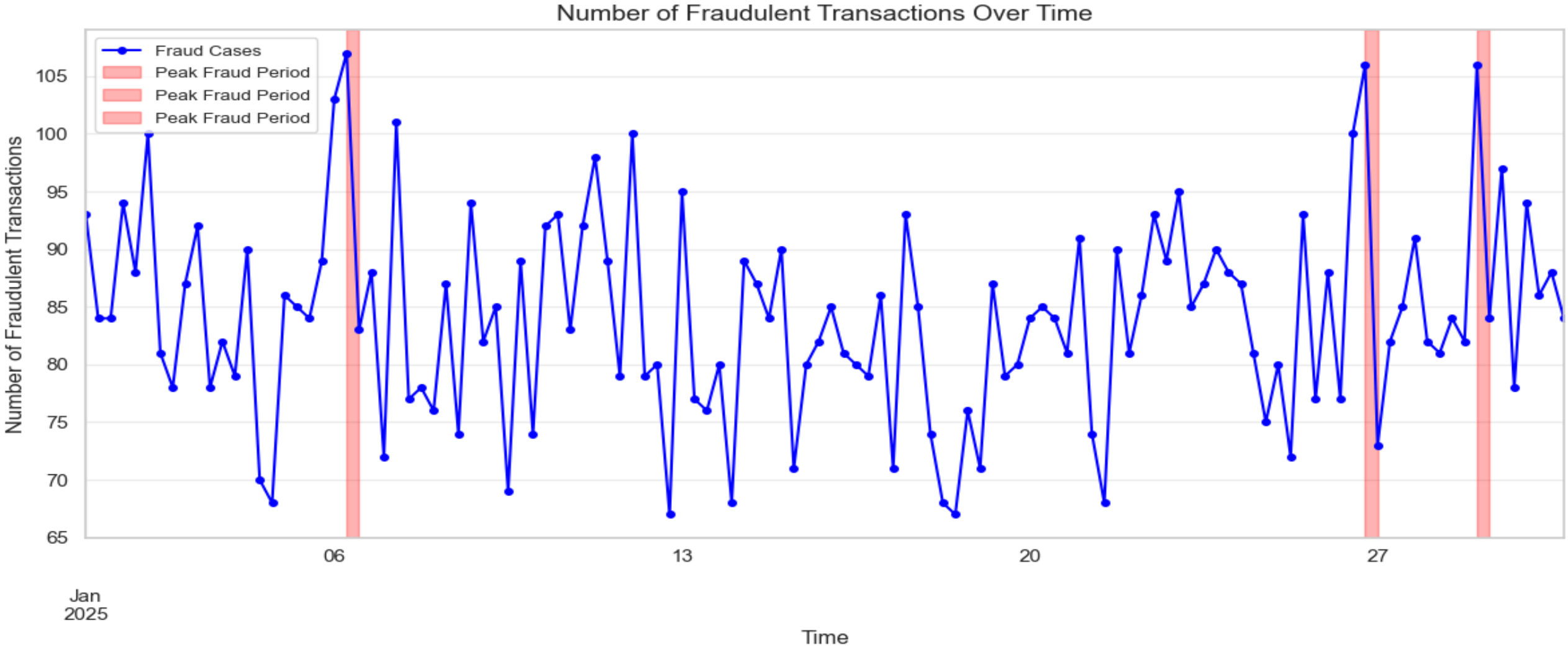
The state with the highest fraud cases is: **Lakshadwee**

Distribution of Fraud Cases By Transaction Devices



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



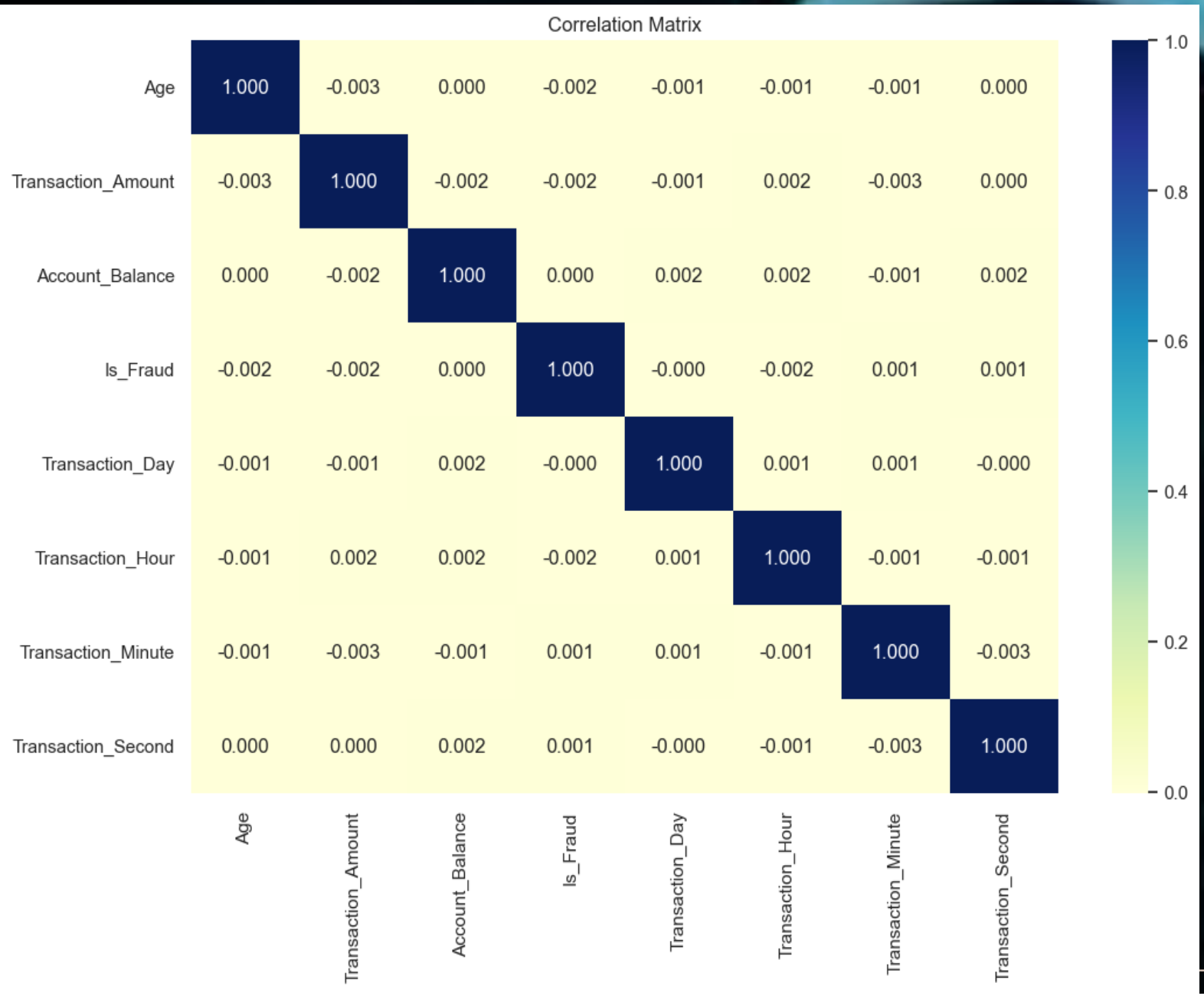
Top 3 Peak Fraud Periods:

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

Correlation Matrix



The correlation matrix shows very weak linear relationships among variables, indicating minimal direct association.

Most variables are linearly independent.

Data Preparation

Data Preparation: The dataset required preprocessing before model building.

Preparation Steps:

i) Converted categorical data into numerical format.

ii) Adjusted data types for certain columns.

iii) Removed unnecessary columns for better model performance.

iv) Standardized numerical columns to maintain a consistent scale.

v)Addressing the Class imbalance since the data was highly imbalanced

Modelling

Tested Models:

1. Logistic Regression (Baseline)
2. Decision Tree Classifier
3. Random Forest Classifier
4. Adaboost Classifier
5. XGBoost Classifier
6. Gradient Boosting Classifier
7. Stacking Classifier
8. K-Nearest Neighbors (KNN) Classifier

Models Tested with Tuned Parameters:

1. K-Nearest Neighbors (KNN) Classifier
2. Random Forest Classifier

Evaluation

The following Metrics were used to evaluate the performance of the models:

1. **Recall**: Measures how well a model identifies actual fraudulent transactions.

2. **Precision**: Measures how many flagged transactions are actually fraudulent.

3. **AUCPRC(Area Under the Precision Recall Curve)** : Measures the trade-off between precision and recall across different thresholds, with a higher value indicating better model performance

Evaluation Results

#	Model	Accuracy	Precision	Recall	AUPRC
2	KNeighbors Classifier	0.705900	0.050793	0.273043	0.093486
0	Decision Tree Classifier	0.880125	0.045499	0.068880	0.080677
1	Random Forest Classifier	0.949550	0.000000	0.000000	0.053888
4	Adaboost Classifier	0.949550	0.000000	0.000000	0.052331
6	XGBoost Classifier	0.949400	0.000000	0.000000	0.052140
5	Gradient Boosting Classifier	0.949550	0.000000	0.000000	0.050345
3	Bagging Classifier	0.949175	0.000000	0.000000	0.049623

From the results:

1. Best Model: K-Nearest Neighbors (KNN) Classifier had the highest recall (0.273) and AUPRC (0.094) among all models.

2. Accuracy vs. Recall: Decision Tree, Random Forest, Adaboost, and Gradient Boosting had high accuracy (>0.88) but lower recall and AUPRC.

3. Poor Performers: Random Forest, Adaboost, XGBoost, Gradient Boosting, and Bagging Classifiers had AUPRC < 0.054 and 0 precision & recall, failing to detect fraud.

4. Overall Best: KNN performed best in fraud detection, excelling in recall and handling imbalanced fraud data.

Evaluation Results

#	Model	Accuracy	Precision	Recall	AUPRC
0	Decision Tree Classifier	0.880125	0.045499	0.068880	0.080677
1	Random Forest Classifier	0.949550	0.000000	0.000000	0.053888
2	KNeighbors Classifier	0.705900	0.050793	0.273043	0.093486
3	Bagging Classifier	0.949175	0.000000	0.000000	0.049623
4	Adaboost Classifier	0.949550	0.000000	0.000000	0.052331
5	Gradient Boosting Classifier	0.949550	0.000000	0.000000	0.050345
6	XGBoost Classifier	0.949400	0.000000	0.000000	0.052140
7	Stacking Classifier	0.942700	0.069182	0.010902	0.052049
8	Tuned Random Forest	0.910825	0.046280	0.039148	0.050789
9	Tuned KNeighbors	0.761975	0.054189	0.225966	0.084625

i) The Stacking Classifier performs well in accuracy. However, its recall for fraud cases remains low(0.01), meaning it fails to identify most fraud case

ii) K-Nearest Neighbors (kNN) model still stands out as the best performer for fraud detection due to its high recall and superior AUPRC, both critical for identifying fraud in imbalanced datasets.

Conclusions

1. Gender Balance: Fraud cases are relatively balanced between genders.
2. Age Vulnerability: Ages 19-30 and 51-60 experience the highest fraud cases, indicating greater vulnerability.
3. High-Risk Transactions: Transfers (2,073 cases) and credit transactions (2,048 cases) have the most fraud incidents.
4. Risky Transaction Channels: ATM booths, ATMs, and self-service machines pose the highest fraud risks.
5. Peak Fraud Periods: Fraud incidents spike during holidays.
6. Best Fraud Detection Model: K-Nearest Neighbors (KNN) Classifier performed the best in detecting fraud.

Recommendations

1. 1. Age-Group Analysis: Financial institutions should analyze fraud by age group to design targeted awareness campaigns and reduce fraud risks.

2. 2. Risk Mitigation: Conduct deeper analysis on fraud-prone areas like transfers and credit transactions to establish better controls.

3. 3. ATM & Self-Service Security: Strengthen security measures on ATMs, kiosks, and self-service machines due to high fraud incidents.

4. 4. Holiday Monitoring: Heighten fraud monitoring during holidays and special national events.

5. 5. Age-Based Strategies: Use fraud distribution by age group to implement targeted prevention strategies and awareness efforts.

6. 6. Gender-Based Strategies: Analyze fraud by gender to design specific awareness campaigns and security measures, especially if one group is more affected.

7. 7. The model that we recommend for this Fraud detection is Knearest Neigbours because of it’s high recall and AUPRC.

Thank you