Fraud Detection in E-commerce Transactions:

Leveraging Machine Learning for Enhanced Security

By Shafira Tasya WIjanarko

PROJECT BACKGROUND

With the internet's growing accessibility and the increasing ease of online shopping, many people are turning to digital platforms for their buying needs. This shift in consumer behavior, driven by the convenience of purchasing from home, comes with challenges, including rising fraud activity such as payment scams in e-commerce transactions

OBJECTIVES

VALUABLE INSIGHT

Develop a deep understanding about the data to comprehend the intricate distribution and patterns present within it

PREDICTION

To predict whether an online transaction is classified as fraudulent or not

RECOMMENDATION

Strategies for transaction targeting and prevention, informed by modeling and results from Exploratory Data Analysis

ABOUT DATASET

Customer ID	Transaction Amount	Transaction Date	Payment Method	Product Category	Quantity	Customer Age	Customer Location	Device Used	
d1b87f62-51b2-493b-ad6a-77e0fe13e785	58.09	2024-02-20 05:58:41	bank transfer	electronics	1	17	Amandaborough	tablet	2
37de64d5-e901-4a56-9ea0-af0c24c069cf	389.96	2024-02-25 08:09:45	debit card	electronics	2	40	East Timothy	desktop	20
1bac88d6-4b22-409a-a06b-425119c57225	134.19	2024-03-18 03:42:55	PayPal	home & garden	2	22	Davismouth	tablet	
2357c76e-9253-4ceb-b44e-ef4b71cb7d4d	226.17	2024-03-16 20:41:31	bank transfer	clothing	5	31	Lynnberg	desktop	2
45071bc5-9588-43ea-8093-023caec8ea1c	121.53	2024-01-15 05:08:17	bank transfer	clothing	2	51	South Nicole	tablet	1

The dataset comprises 1,472,952 transactions with 16 collected features from January 2024 to April 2024, with two target values (fraud and not fraud). The impact of this fraudulent activity resulted in a total loss of $40,449,950

DATA PREPROCESSING

0

MISSING VALUE

There’s 0 missing values

33

OUTLIER

There is 1 outlier in the Customer Age feature and 32 outliers in the Transaction amount feature. However, these outliers are still reasonable, so there is no need to drop them.

0

DUPLICATES

There’s 0 duplicates

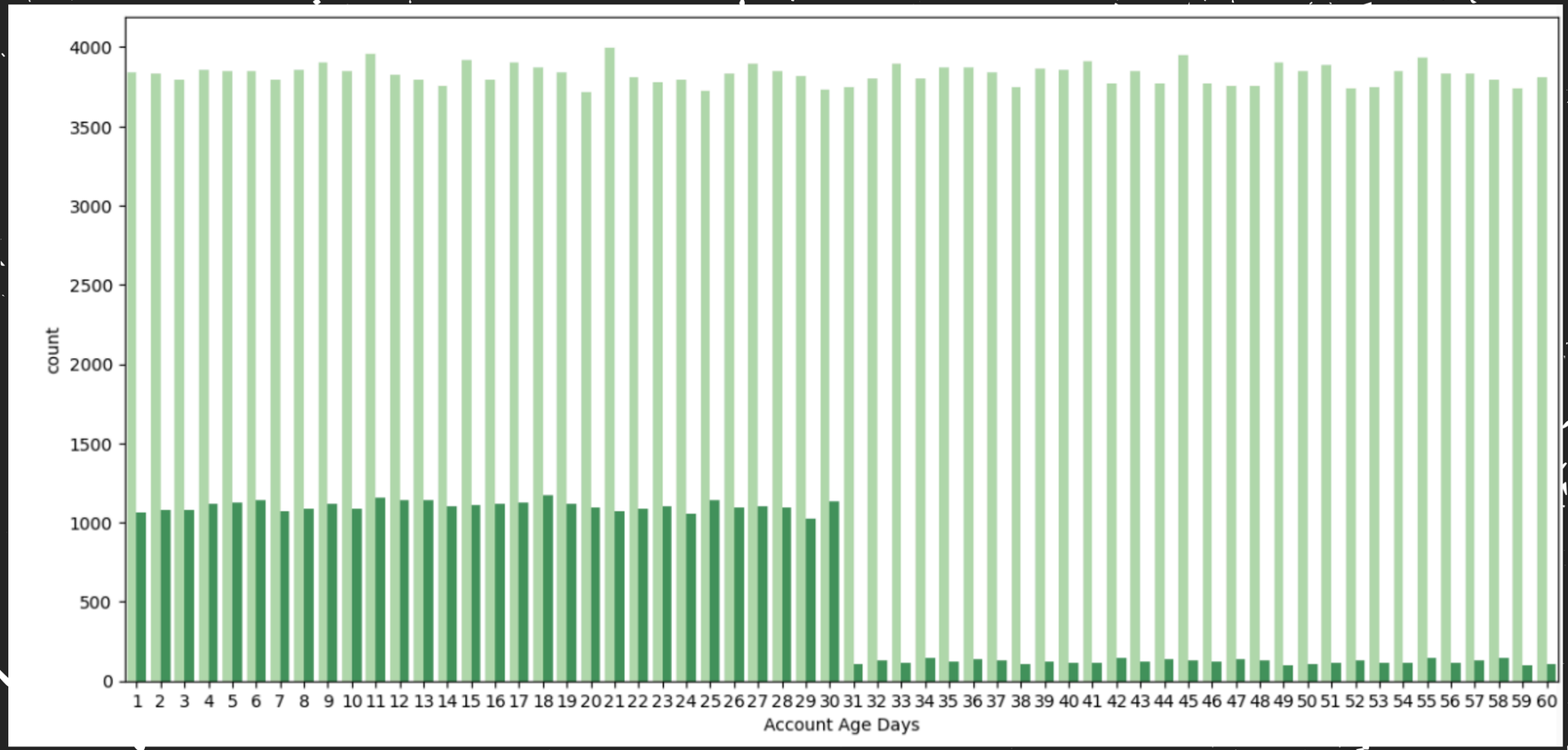
DATA FEATURES

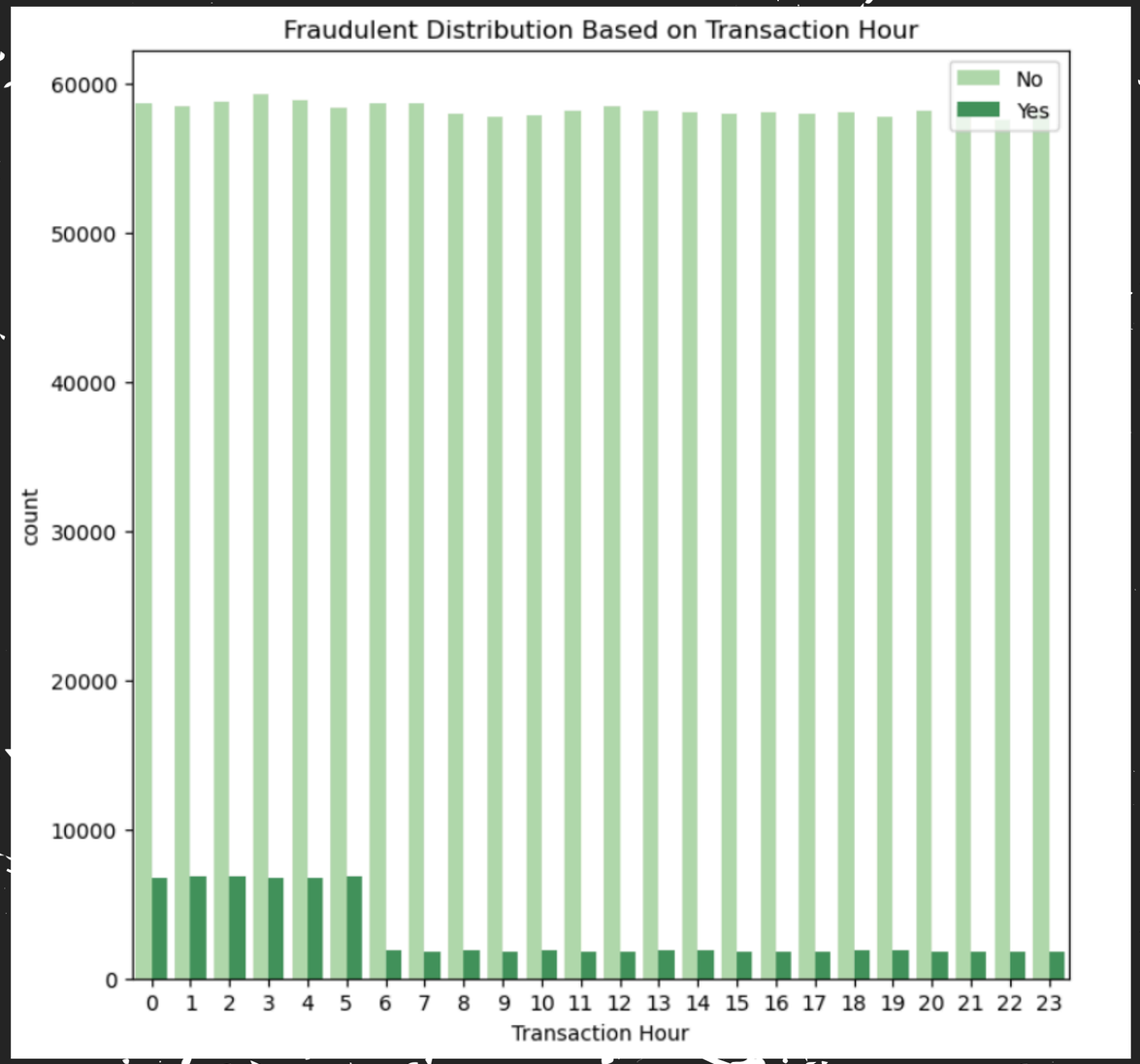
Out of the 16 features under analysis, only 13 will be utilized in the modeling process

NO.	VARIABLE	DEFINITION
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	Account Age Days	The age of the customer's account in days at the time of the transaction
15	Transaction Hour	The hour of the day when the transaction occurred.
16	Is Fraudulent	A binary indicator of whether the transaction is fraudulent (1 for fraudulent, 0 for legitimate).

THE PATTERN OF THE USER'S ACCOUNT AGE (IN DAYS) BETWEEN TRANSACTIONS

The age of customer accounts with the highest number of fraud incidents is approximately one month. This indicates that the majority of fraud cases occur on accounts that have been newly created within one month of their creation.



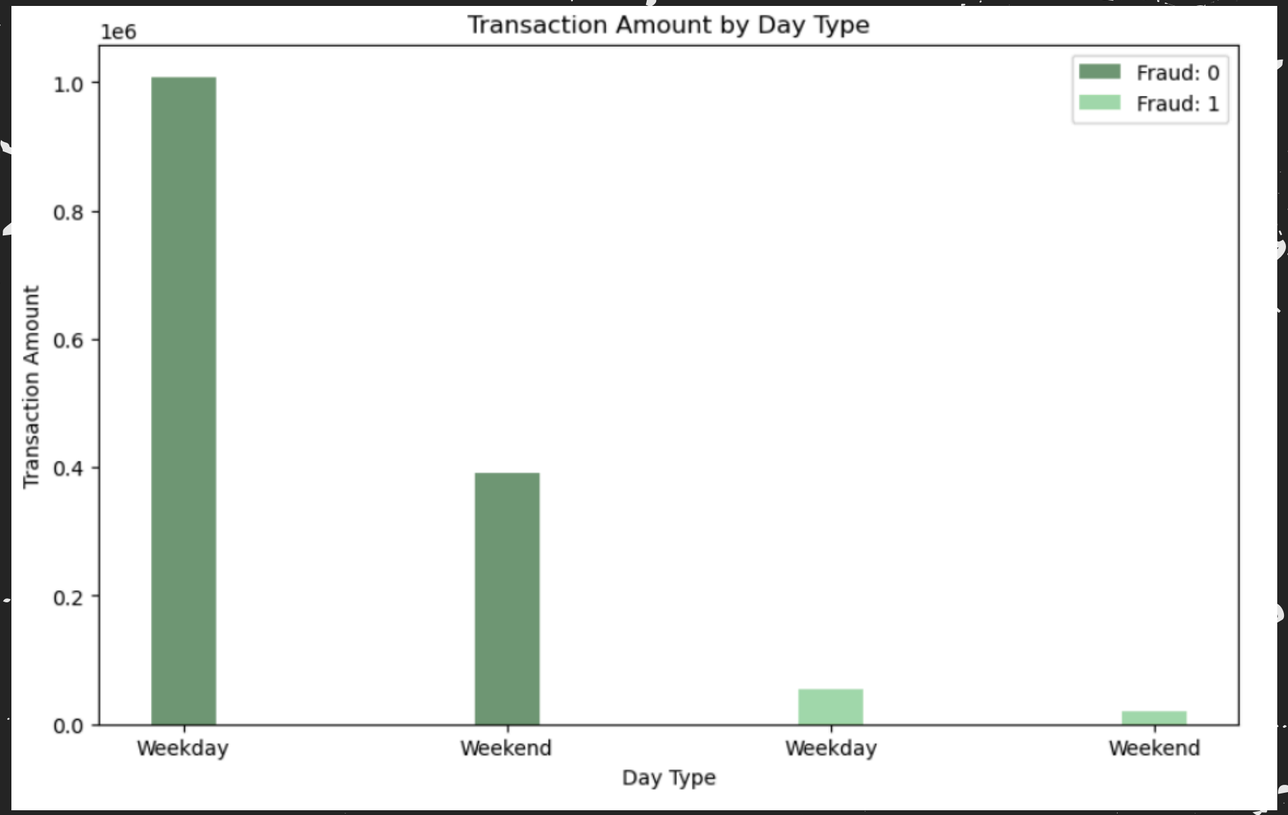


SPECIFIC HOURS WHEN FRAUDULENT TRANSACTIONS ARE MORE PREVALENT

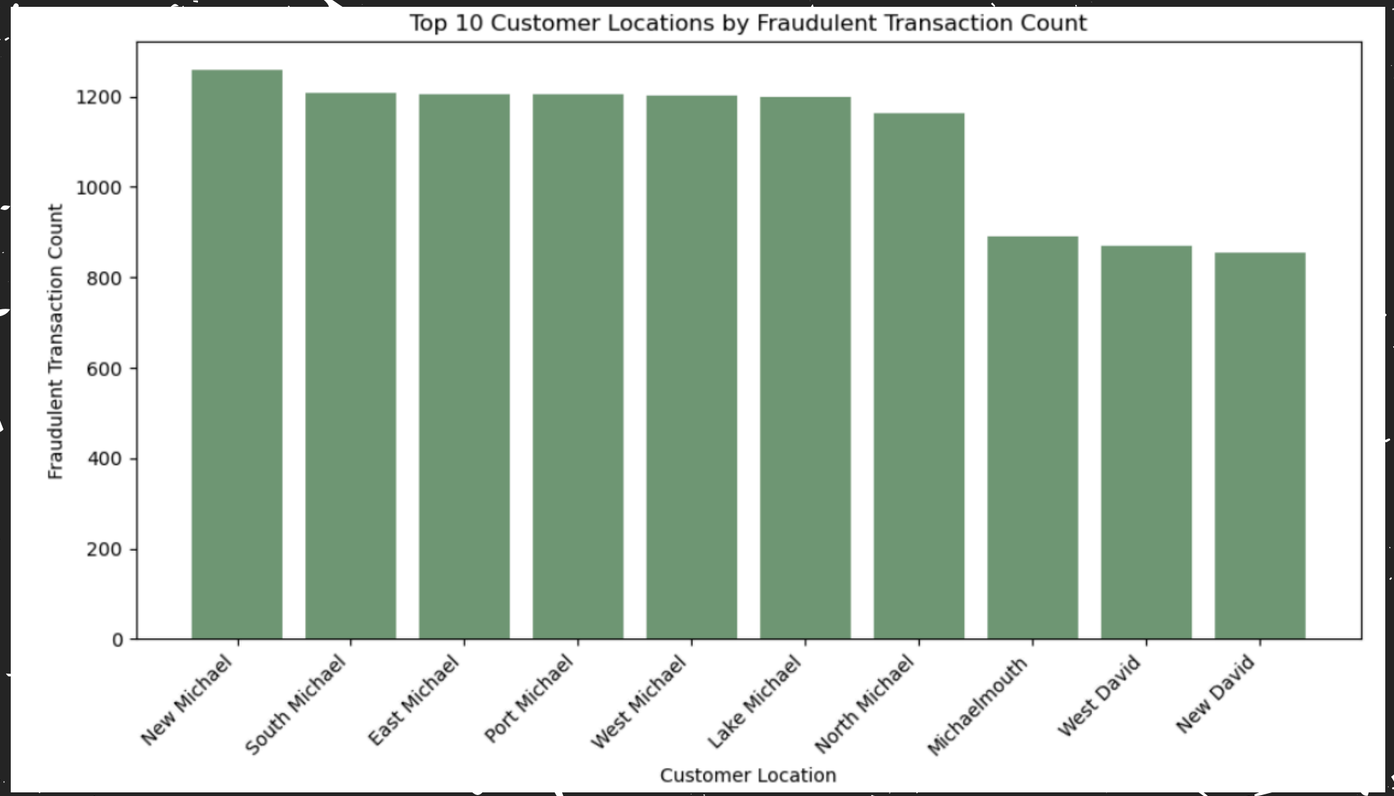
Fraud cases often occur within one to five hours after a transaction is made. This indicates that this time period is the most vulnerable for users to fraud attempts. Fraudsters may exploit security gaps or users' lack of awareness during the initial hours following a transaction.

DIFFERENCES IN TRANSACTION AMOUNTS BETWEEN WEEKDAYS AND WEEKENDS FOR FRAUDULENT

Fraud cases are more prevalent on weekdays compared to weekends. This phenomenon may be attributed to several factors, including the high volume of transactions conducted by users during weekdays



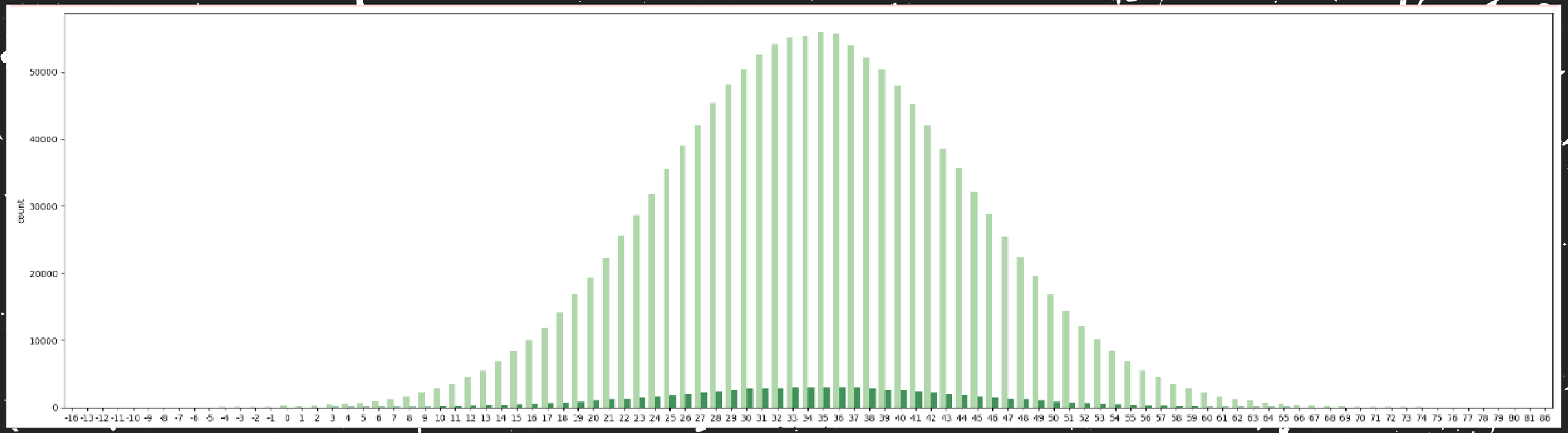
THE TOP 10 LOCATIONS WITH THE HIGHEST NUMBER OF FRAUDULENT TRANSACTIONS

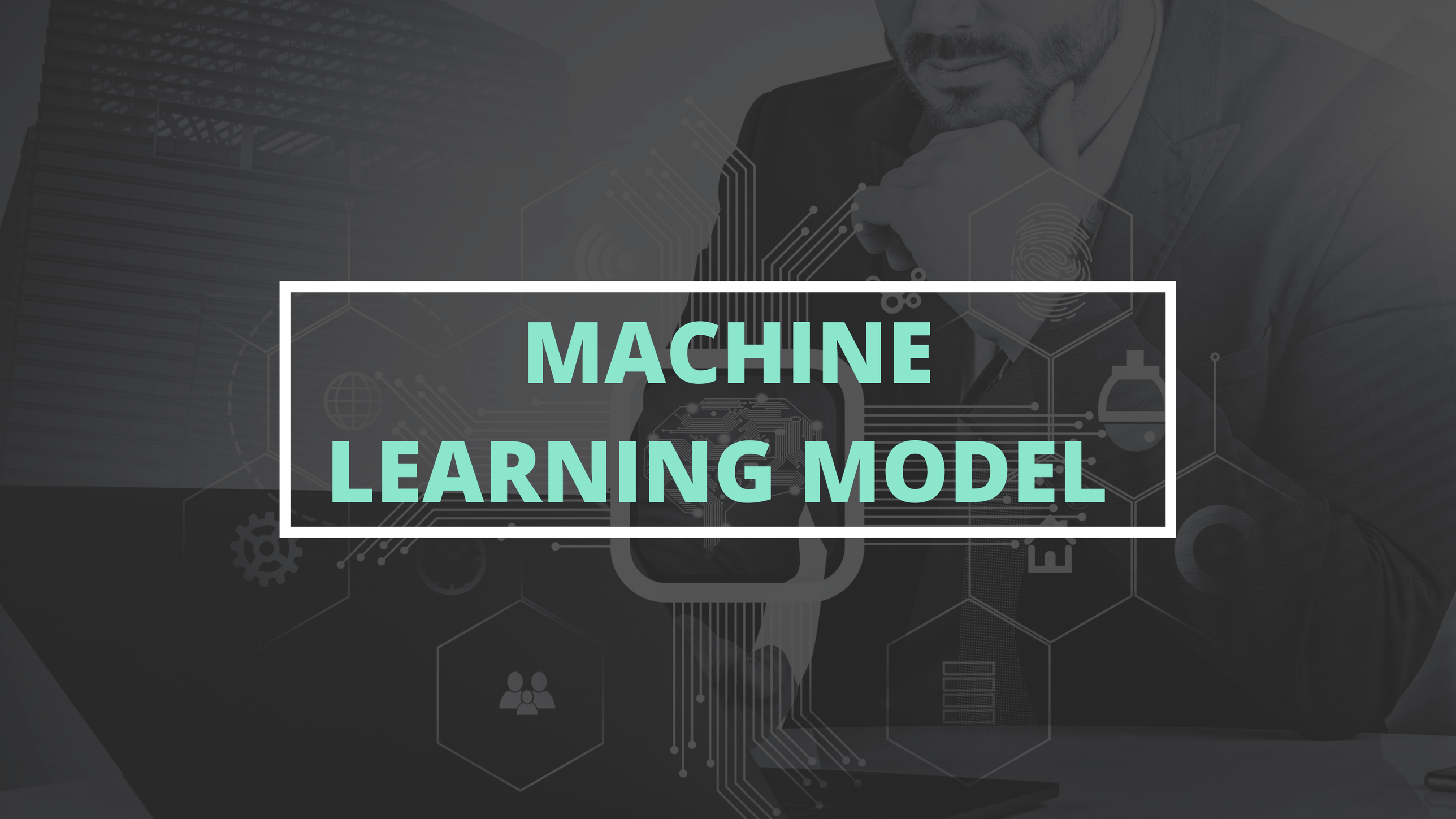


Fraud cases are more frequent in the following locations, with the highest number of cases recorded in New Michael, totaling 1258. This indicates that there is a particular pattern where certain locations, such as New Michael, become prime targets for fraudsters to carry out their fraudulent activities.

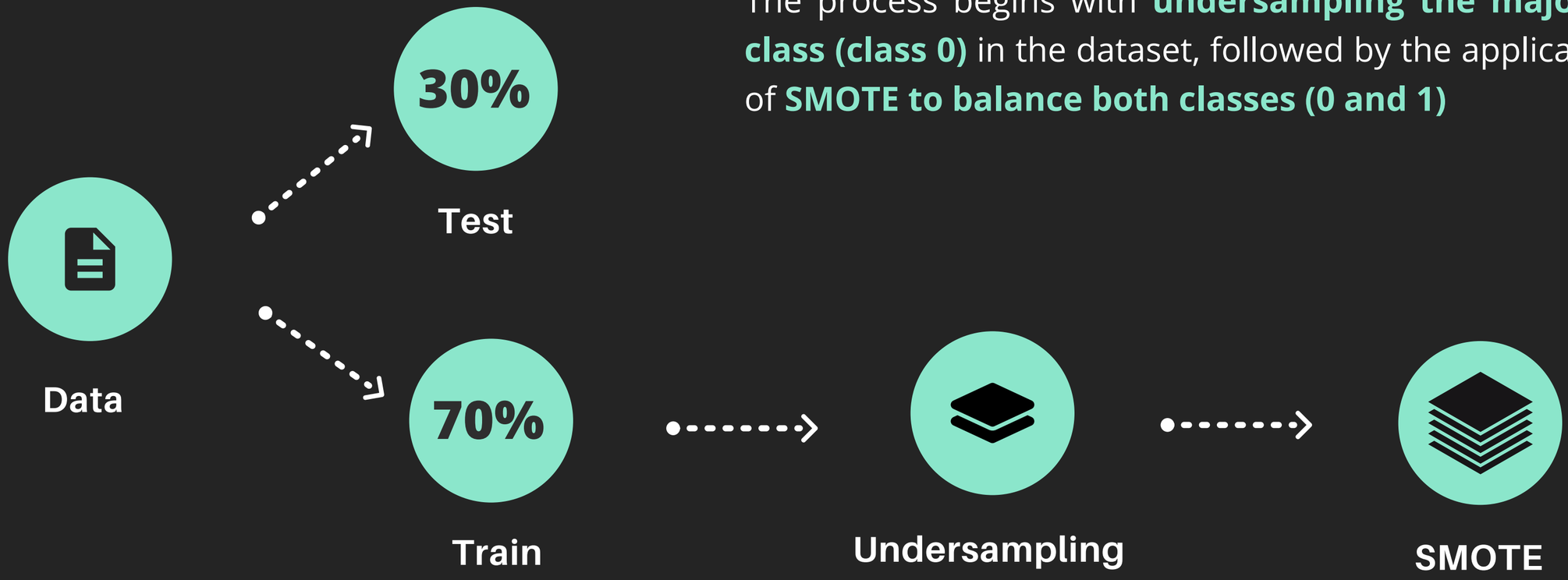
THE DISTRIBUTION OF CUSTOMER AGES IN THE DATASET

Individuals aged 30-39 are found to be the most vulnerable group to e-commerce fraud. This suggests that this age segment may have a high level of trust in e-commerce platforms and frequently engage in online transactions.




MACHINE
LEARNING MODEL

The model was trained on 70% of the data, where the data is imbalanced, necessitating the need for undersampling and SMOTE (Synthetic Minority Over-sampling Technique).



The process begins with **undersampling the majority class (class 0)** in the dataset, followed by the application of **SMOTE to balance both classes (0 and 1)**

This project used various performance metrics including accuracy, precision, recall, and the F1 score.

BEFORE BALANCING

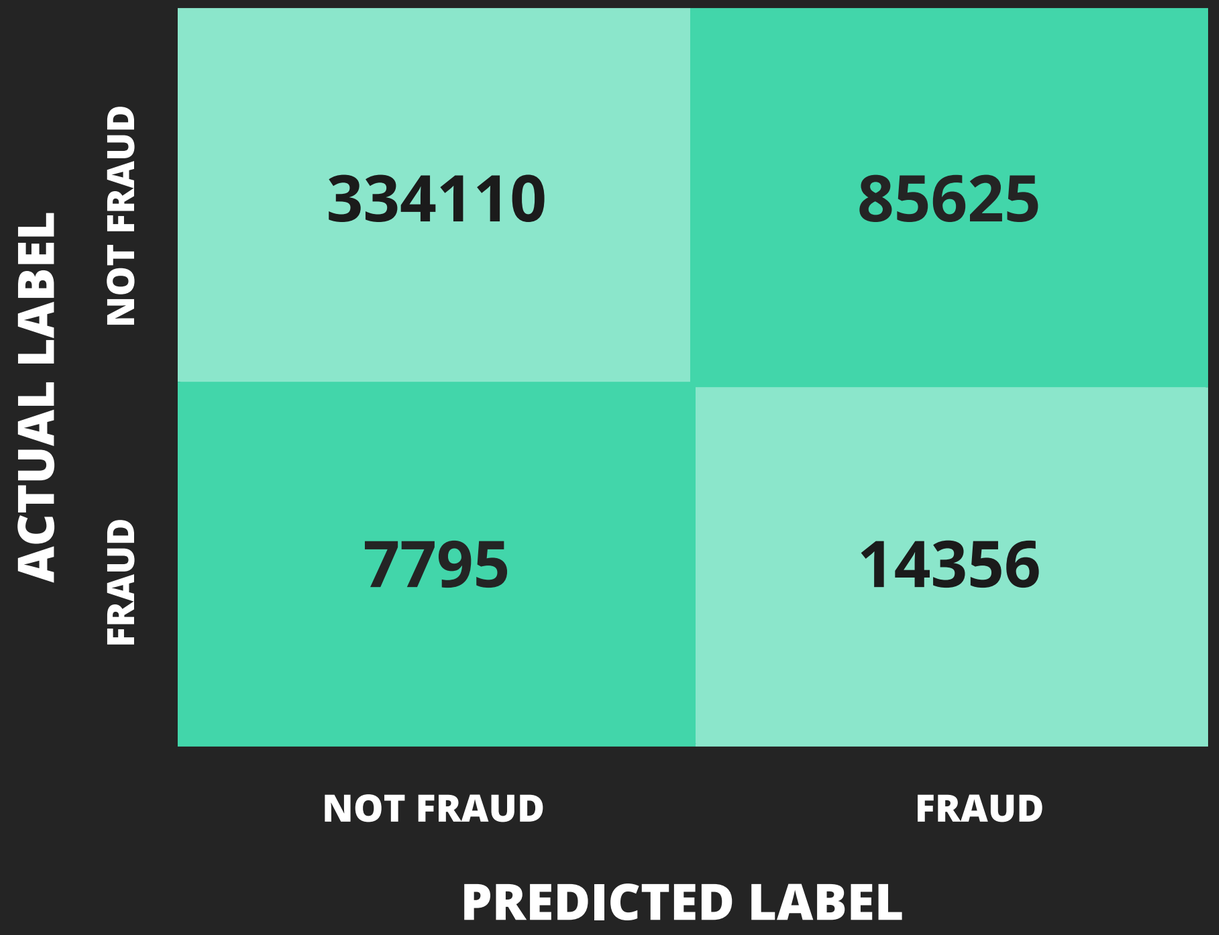
MODEL	ACCURANCY	PRECISION	RECALL	F1-SCORE
DECISION TREE	0.92	0.21	0.25	0.23
RANDOM FOREST	0.96	0.80	0.15	0.25
CAT BOOST	0.96	0.80	0.15	0.25
XGBOOST	0.96	0.80	0.15	0.26

AFTER BALANCING

MODEL	ACCURANCY	PRECISION	RECALL	F1-SCORE
DECISION TREE	0.67	0.09	0.61	0.16
RANDOM FOREST	0.80	0.15	0.63	0.24
CAT BOOST	**0.79**	**0.14**	**0.65**	**0.23**
XGBOOST	0.79	0.14	0.64	0.23

The Cat Boost Classifier model stands out as the optimal choice based on metrics Recall

The model has successfully identified fraudulent transactions with an impressive level of accuracy, reaching a value of 79%. However, it's not just that. The metric recall, often deemed more critical in the context of identifying fraud cases, has reached 65%. This indicates the model's ability to uncover the majority of all actual fraudulent transactions.



CONCLUSION

In conclusion, the fraud detection model shows strong performance with an accuracy of 79% and a recall of 65%. While accuracy reflects the model's overall capability to classify transactions correctly, recall emphasizes its effectiveness in detecting a significant portion of actual fraudulent activities. This balance of high accuracy and decent recall underscores the model's reliability in flagging fraudulent transactions while reducing the risk of missing genuine fraud cases. However, there's room for further optimizations to potentially improve the model's performance, especially in enhancing its F1 score—a composite metric that balances precision and recall—thus strengthening its overall efficacy in combating fraudulent activities.

RECOMMENDATION

01 Enhancing education and awareness regarding e-commerce fraud among users, particularly the 30-39 age group, to mitigate the risk of fraud occurrences.

02 Implementing stricter monitoring of transactions conducted by new accounts and limiting the activities that new accounts can perform during the first month. For example, restricting the number of transactions or the total transaction value that can be conducted.

03 Utilizing real-time monitoring technology to detect and promptly respond to suspicious activities during vulnerable hours. This system can provide alerts to the security team.