TRANSFORMING THE FINANCE WORLD

Unveiling the Hidden Depths of Ethereum Transactions: A Machine Learning Approach to Fraud Detection

How technology has impacted our financial dealings

By Arina Nurjanah

Outline

Points to discuss

- Background
- About Ethereum and ERC20
- Objectives
- About Dataset
- Data Preprocessing
- Data Features
- Business Questions
- Machine Learning Model
- Conclusion
- Recommendation Action
- Profound Impact

Background

Cryptocurrency has become a preferred tool for criminals to launder money and engage in illegal activities. Criminals are becoming more adept at using cryptocurrencies for terrorism financing, hiding their financial transactions, and using them for payment and investment fraud.

About Ethereum and ERC20

Ethereum enables decentralized applications, organizations, and transactions without requiring a central authority. Users have control over their data and sharing preferences. Ethereum's cryptocurrency, Ether, is used for specific activities within the Ethereum network. While ERC20 is a token standard developed on top of the Ethereum platform, making it one of the most commonly traded token types on the Ethereum blockchain.

Objectives

Valuable Insights

Understand more about transaction patterns that occur on the platform.

Prediction

To predict whether an Ethereum transaction is categorized as fraud or not (not fraud).

Recommendation Action

Transaction targets and prevention implementation methods based on modelling and Exploratory Data Analysis results.

About Dataset

This study utilized one dataset licensed by Open Database, which contains detailed Ether (ETH) transaction information.

The classification dataset has a target with two values (fraud and not fraud) and 50 features with a total of 9,841 rows.

Data Preprocessing

After the process, only 9,288 rows will be utilized for analysis and modeling. This data preprocessing makes features more informative and useful to improve model performance.

Redundant Features

The columns are unique, and the values of these variables are all 0s, as zero variance indicates constant or near-constant behavior in the variables.

Missing Value

At 8.42% the empty rows were medianly imputed. Most of the columns with missing values are right-skew distributions.

Duplicates

The same 553 rows exist after deleting unique columns. The rows were dropped.

Multicollinearity

Only nine features have correlated with the threshold value of 0.8. Dropping them is a good way except for one feature that has the most highly correlated with the target.

Data Features

Out of the 38 features under analysis, only 29 features will be utilized in the modeling process.

The features in this study are divided into several category.

Time

The columns show the average difference in transaction time in minutes.

Ether (ETH)

The columns provide important Ether transaction details for a wallet address, including sending/receiving minimum, maximum, and average values, total Ether transactions, and smart-contract interactions.

ERC20 Token

The ERC20 token transaction columns show sending and receiving transactions to specific addresses and smart contracts.

Smart Contract

Ether and ERC20 token transactions are related to smart contracts.

Wallet Address

Columns containing transactions to/from unique account addresses.

Business Questions

01 How can the timing pattern between user transactions affect the risk of fraud in Ethereum transactions?

02 What is the comparison between the average value of Ether sent by an address and the average value of Ether received by the particular address in a specific period of time?

03 What is the relationship between total ERC20 transactions and unique ERC20 transactions from unique accounts against fraud?

04 What is the influence of receive and send transaction to unique addresses against fraud?

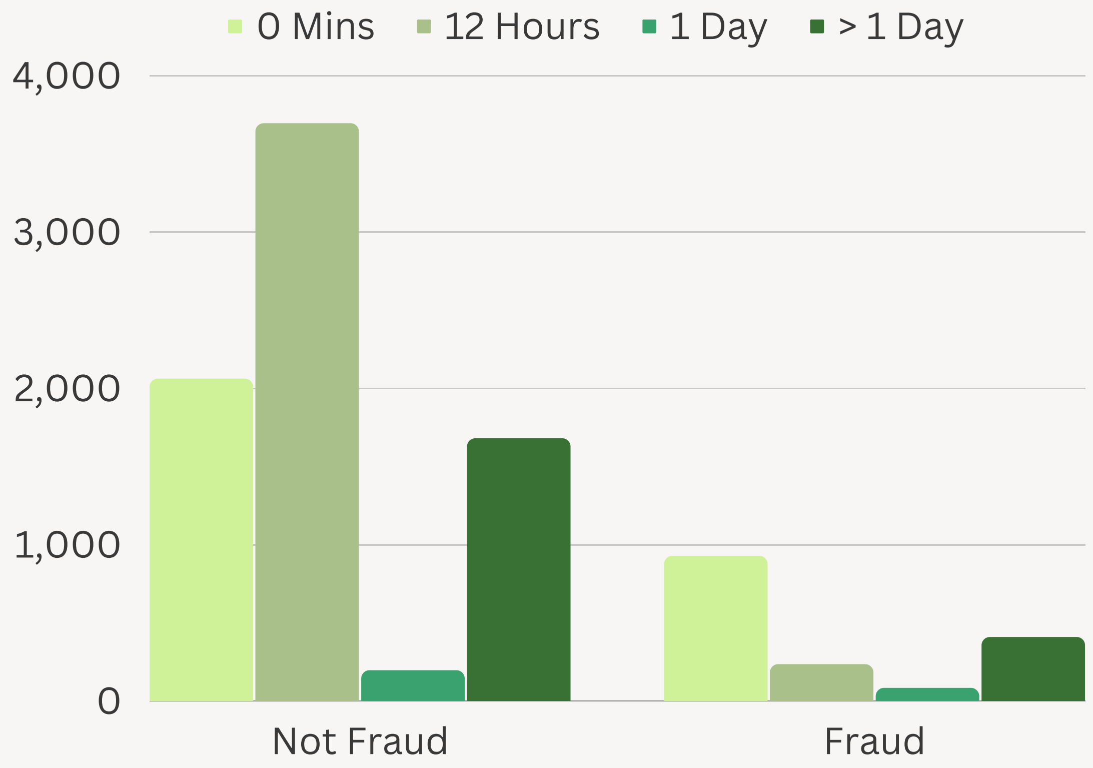
05 What is the difference between the wallet addresses that have made ETH transactions sent and received on total transactions based on fraud classification?

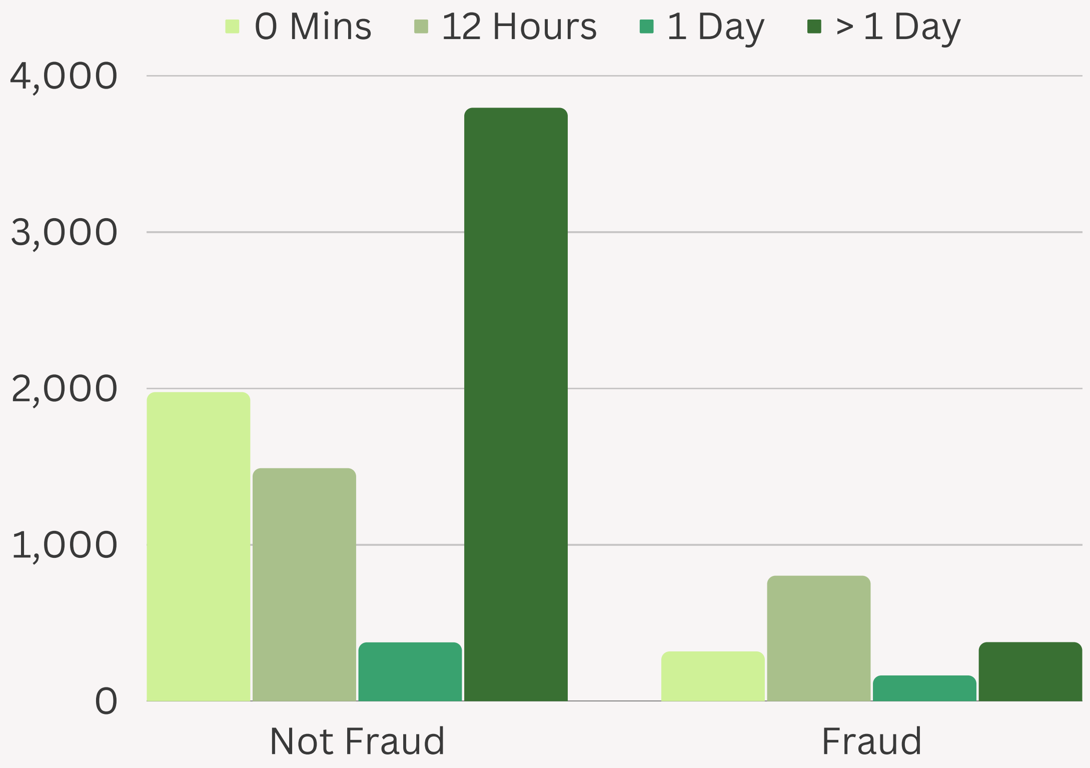
Let's deep dive!

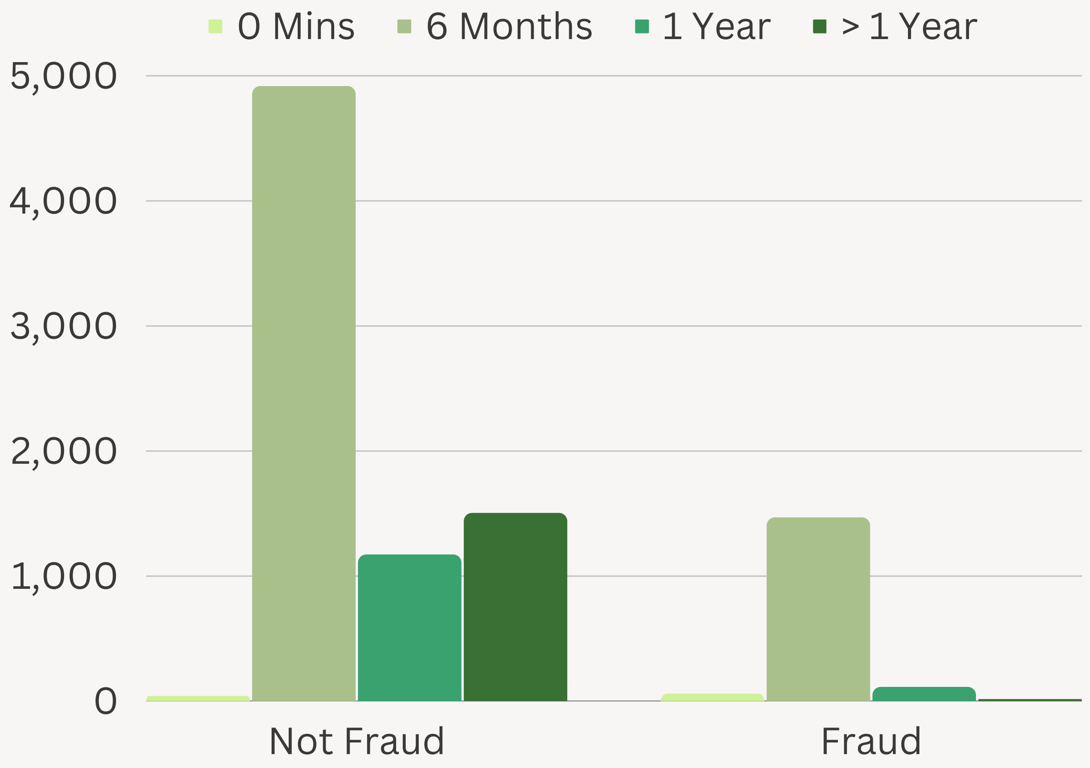
The timing pattern between user transactions

Why is it important? They have a substantial impact on the risk of fraud in ETH transactions.

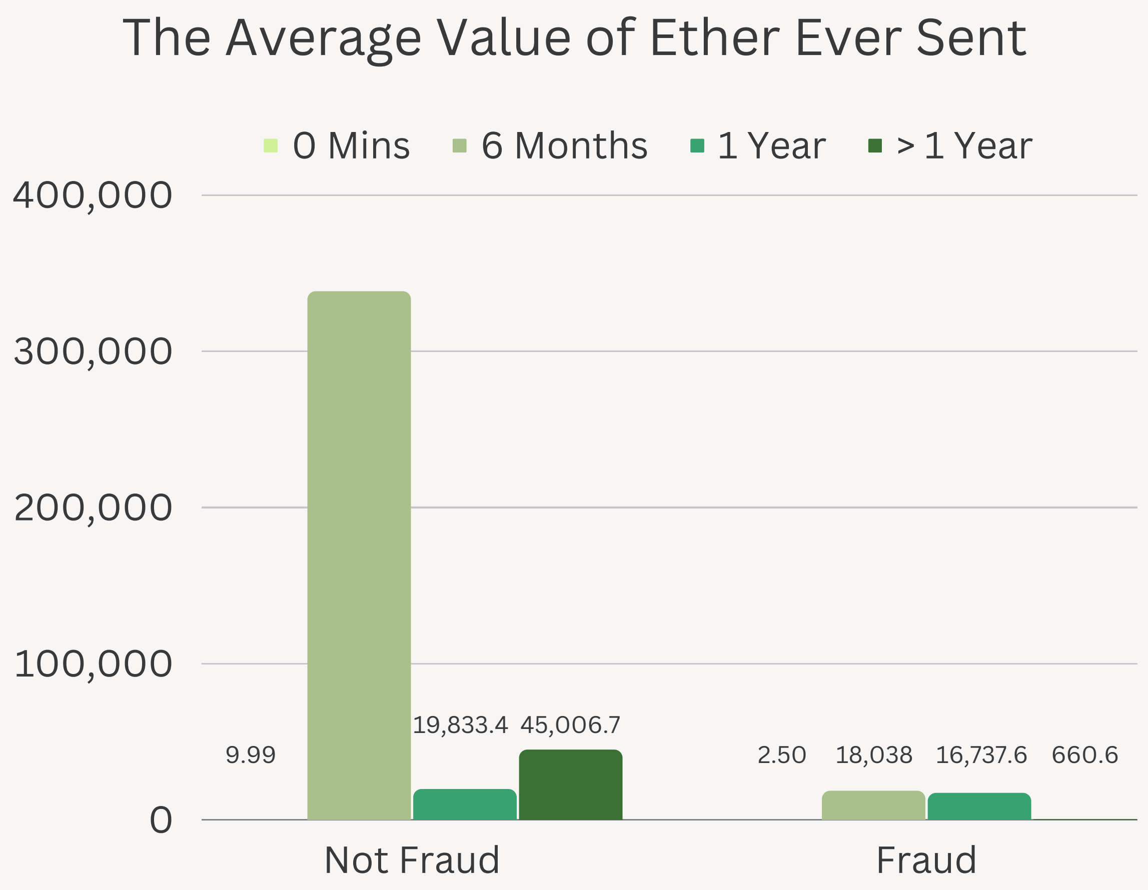
A high frequency of transactions within a short time frame, especially when combined with small transaction amounts, can indicate an attempt to obfuscate the true purpose of the transactions or mask fraudulent activity.


Average Time Between Sent Transaction

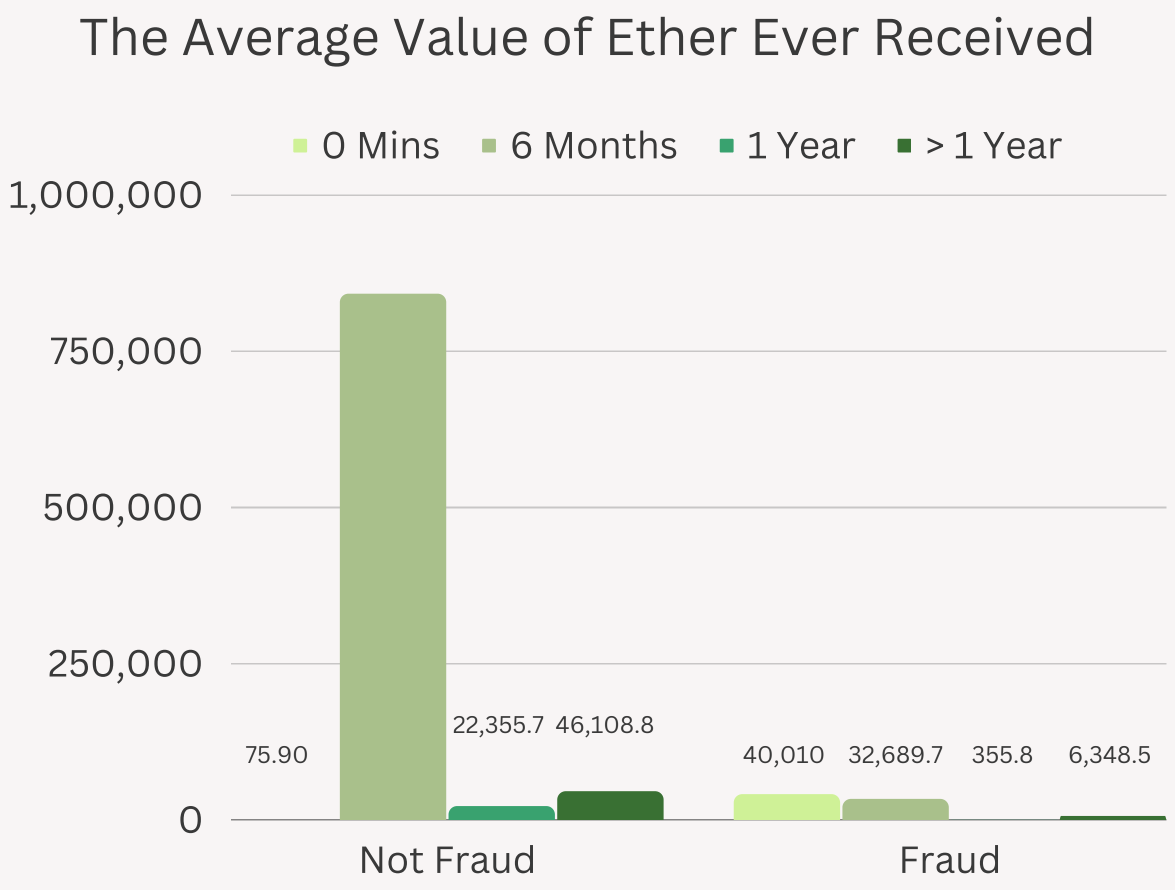

Average Time Between Received Transaction


Time Different Between First and Last Transaction

Understanding the different types of sending and receiving Ether transactions



The Average Value of Ether Ever Sent



The Average Value of Ether Ever Received

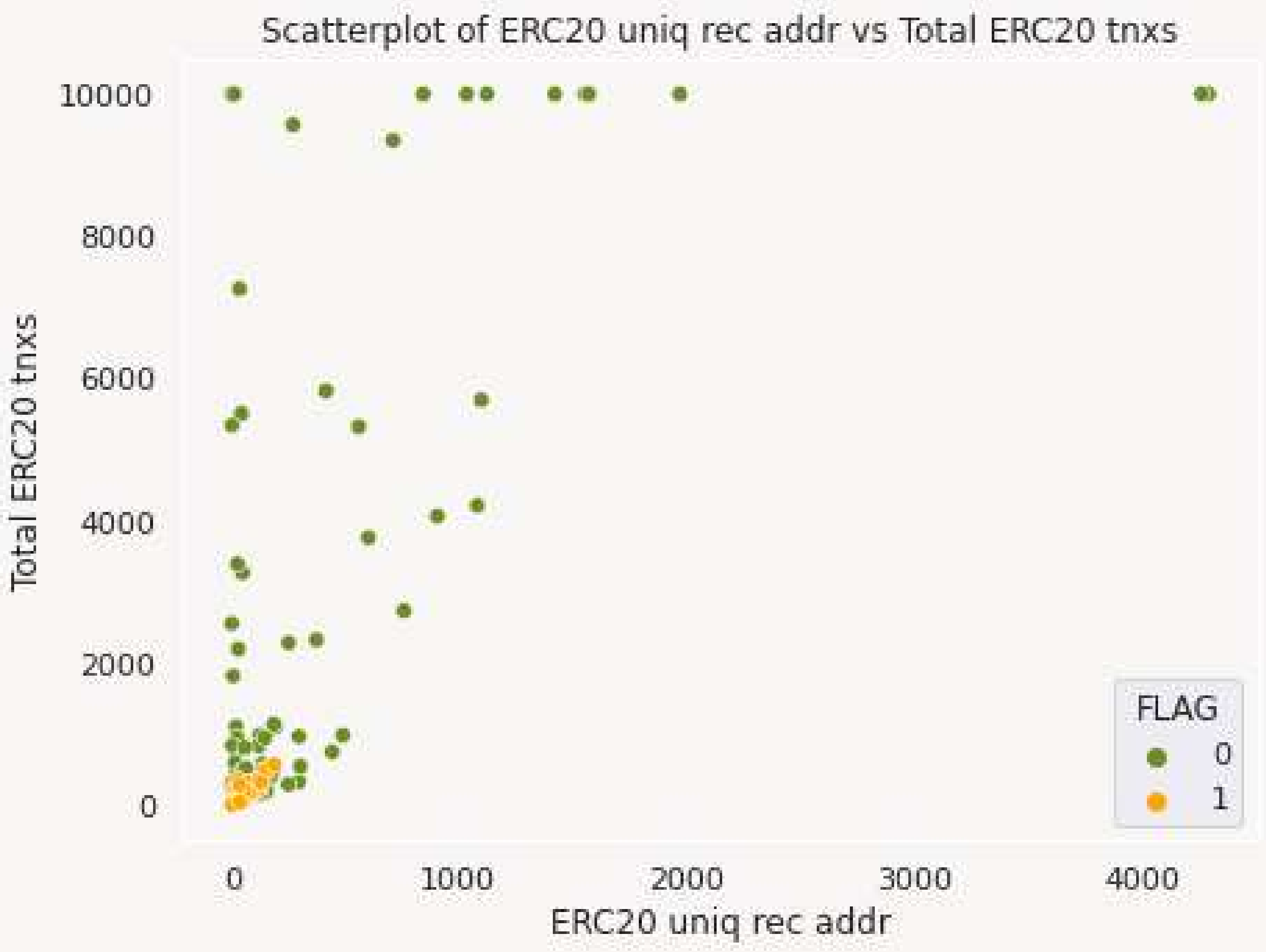
Monitoring the scheme

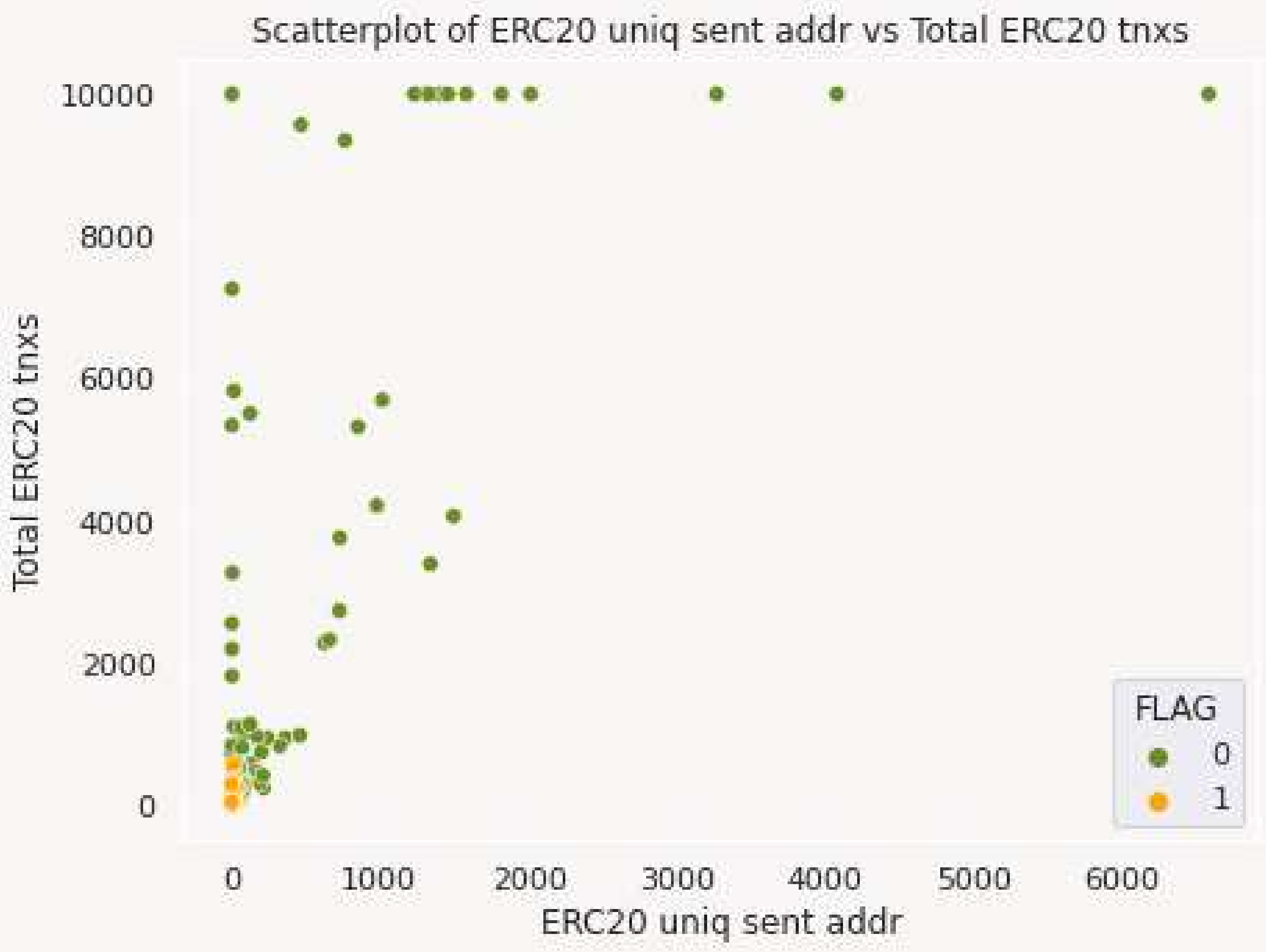
The average amount of Ether sent or received by that address is lower than usual, it may indicate suspicious behavior.

Significant differences in the average amount of Ether sent or received by addresses over a certain period of time could indicate fraudulent activity.

ERC20 token transactions with the unique ERC20 tokens on the Ethereum network

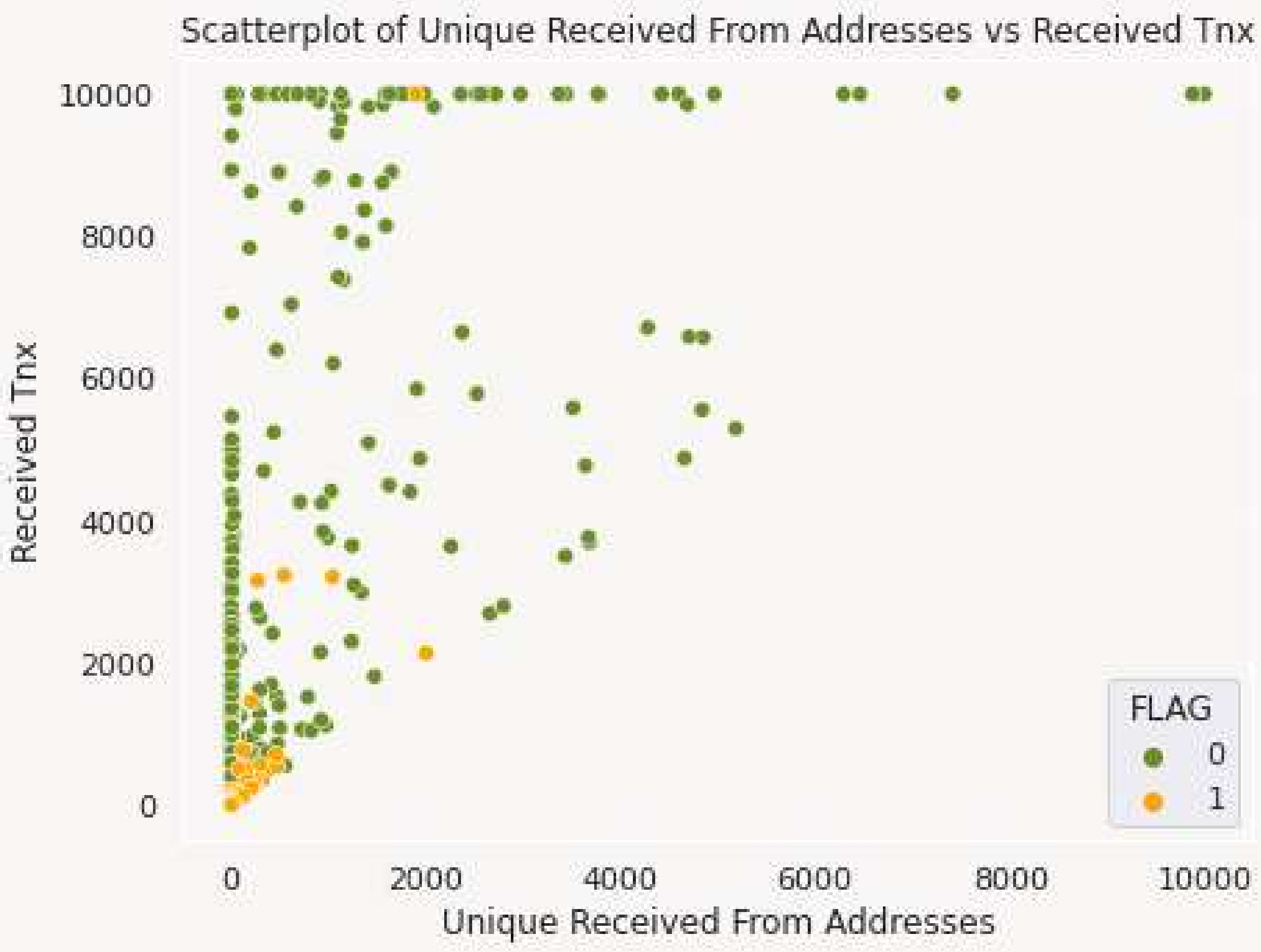
The fewer ERC20 transactions on assets sent and received by different address accounts tend to reveal fraud.

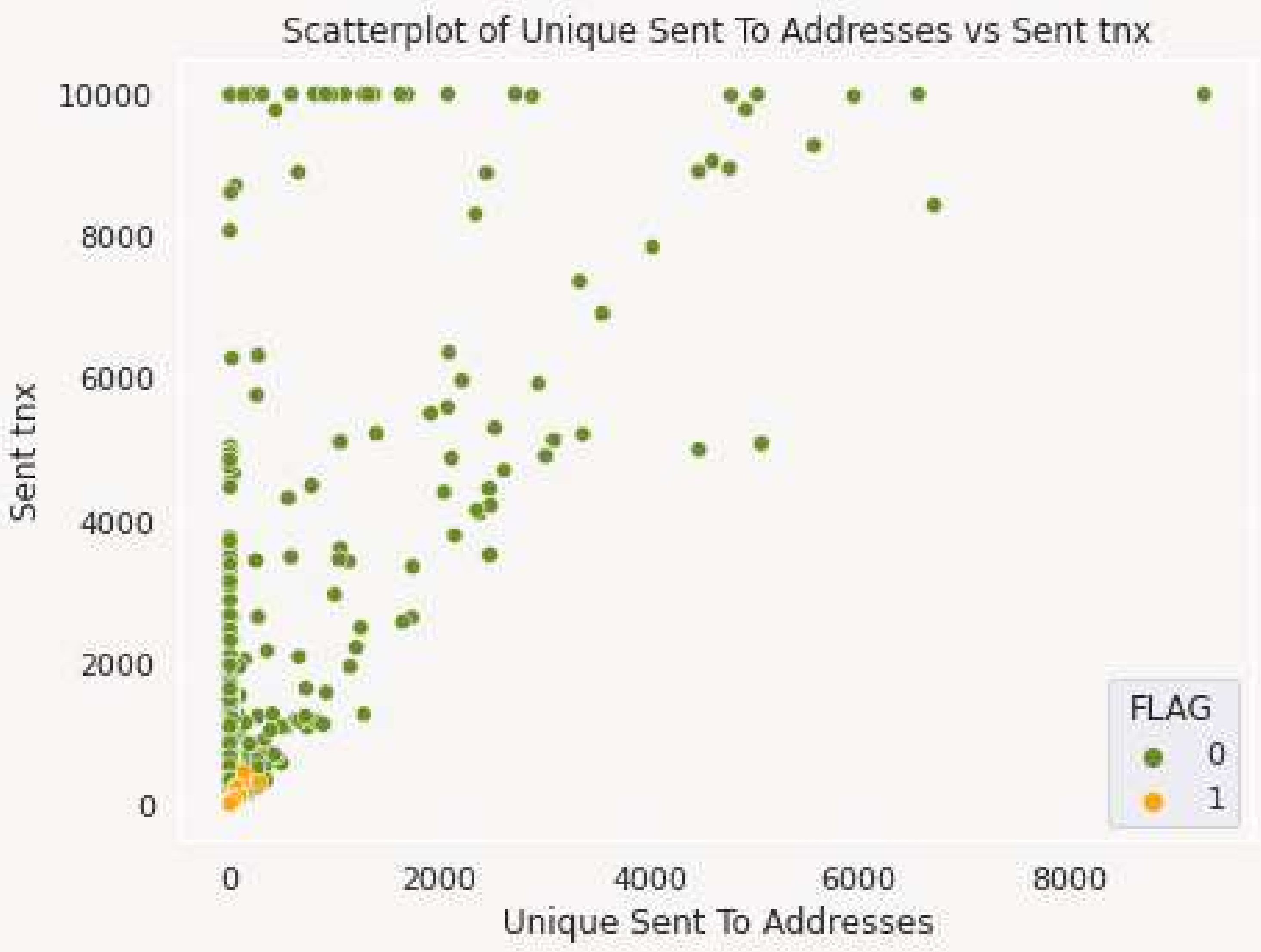




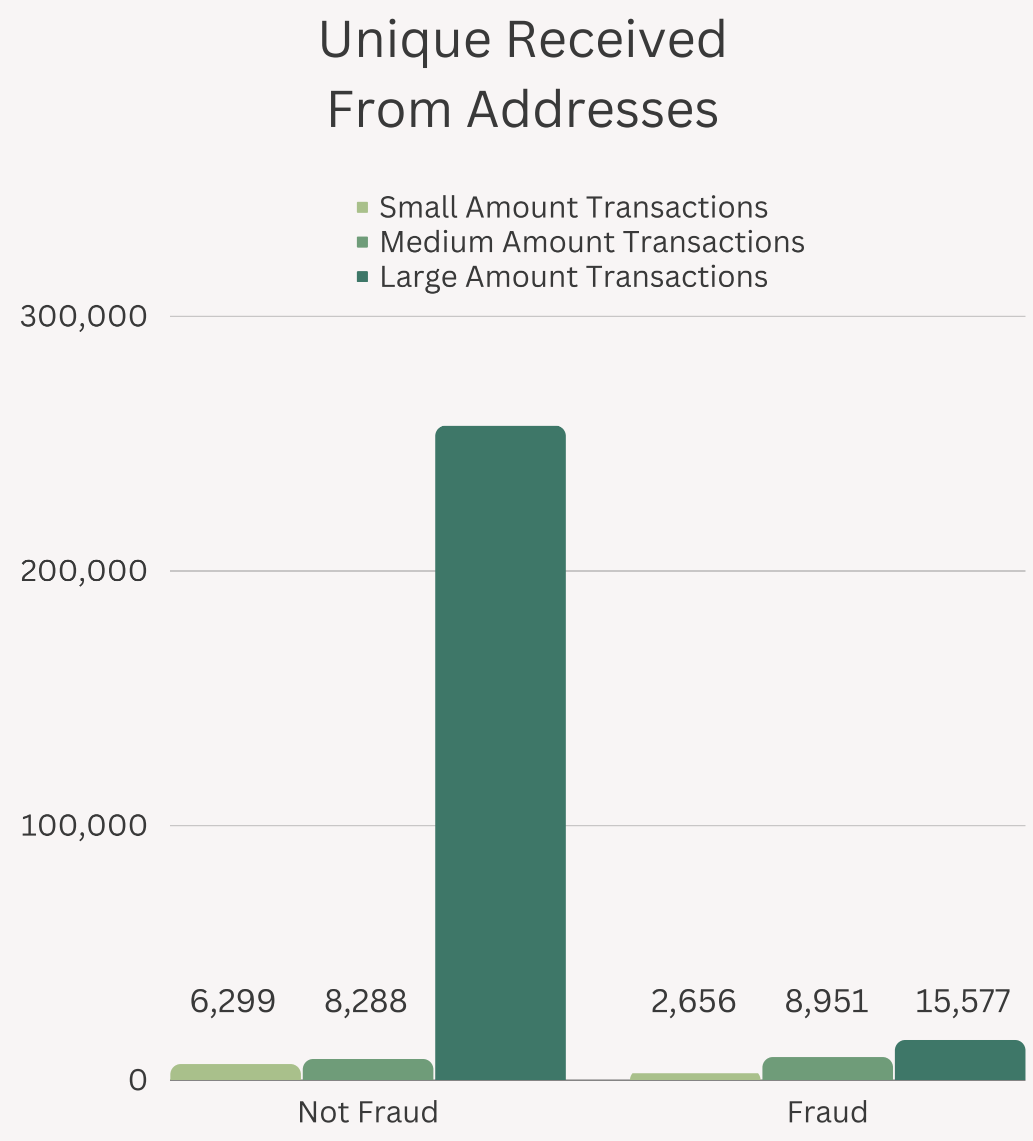
Transactions by a wallet address with the unique account addresses

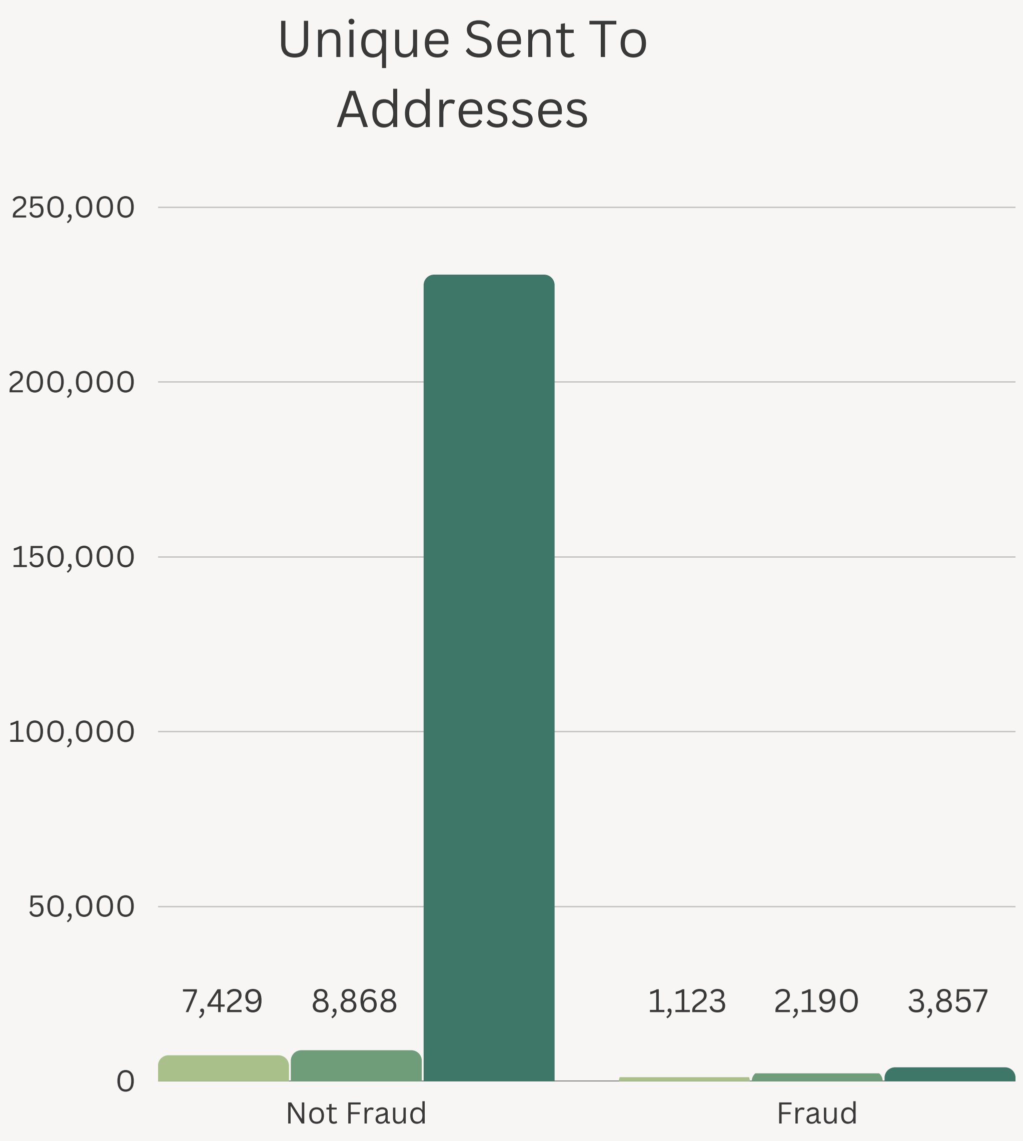
The fewer transactions and the smallest number of different wallet addresses further strengthen the suspicion of a fraudulent transaction.





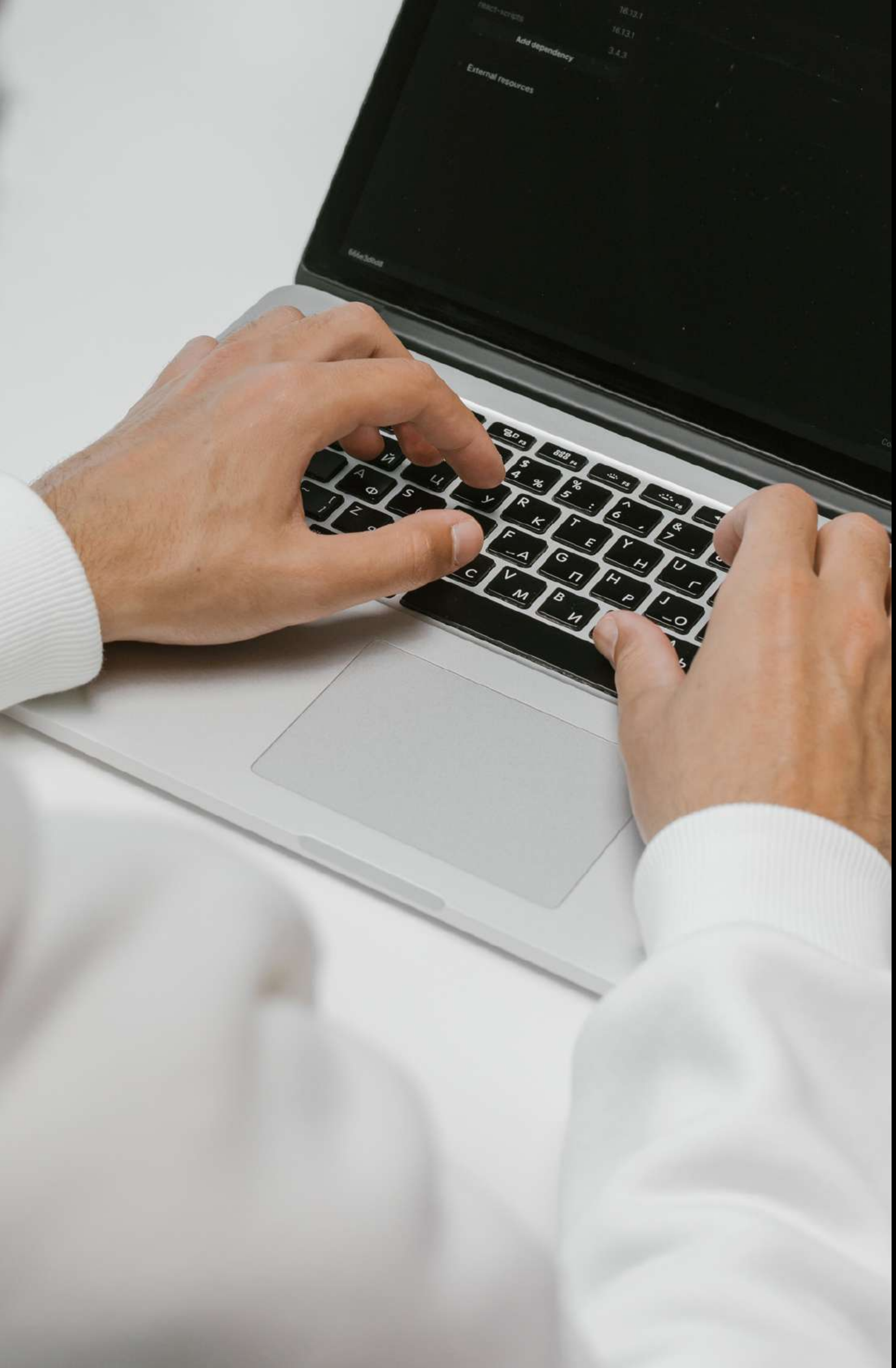
The transactions of unique addresses on total transactions





Observing the habit

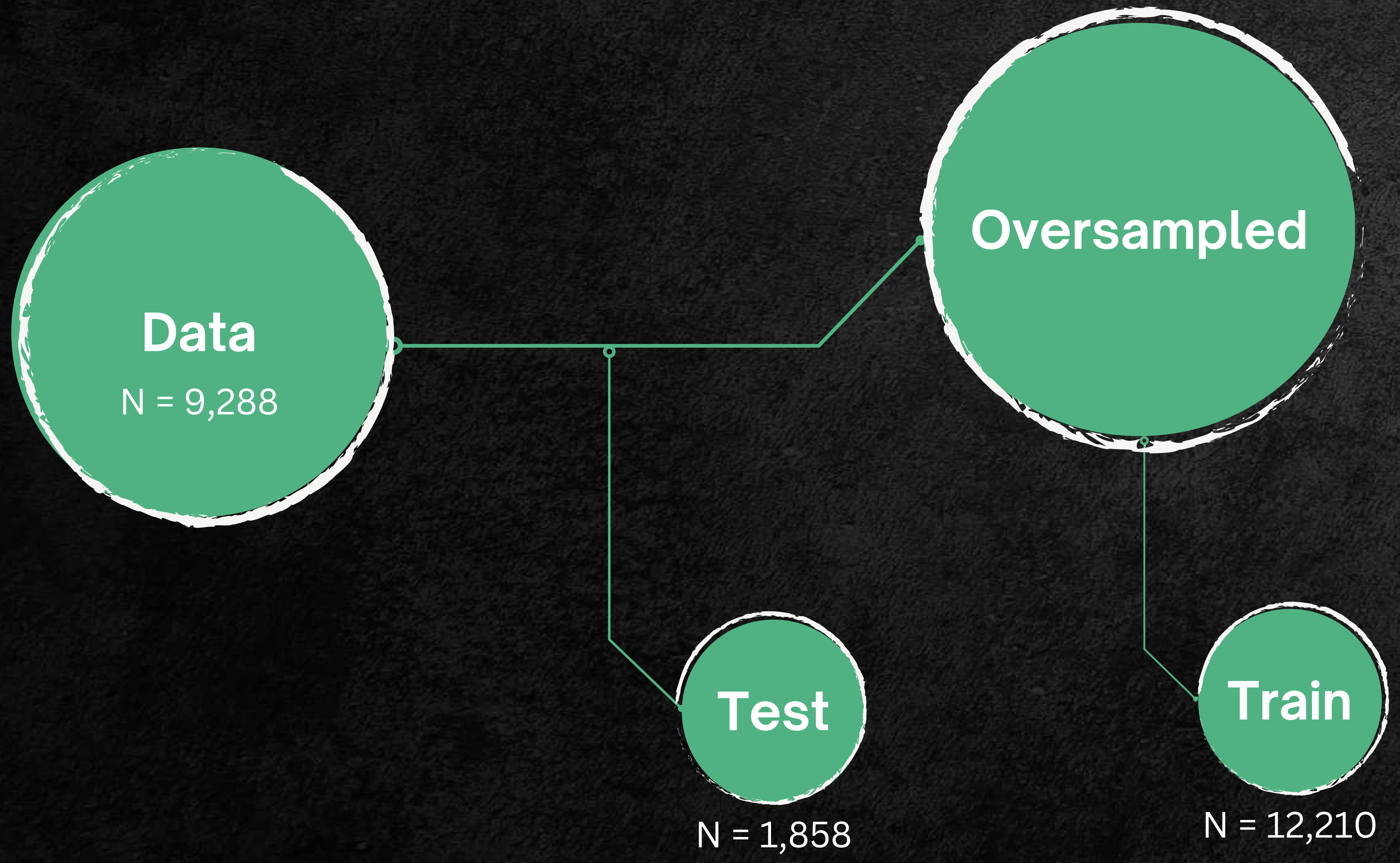
The fraud transaction actor has a total number of different wallet addresses that have **received** Ether from a particular wallet address **more than** transactions of unique account addresses when they **sent** Ether.



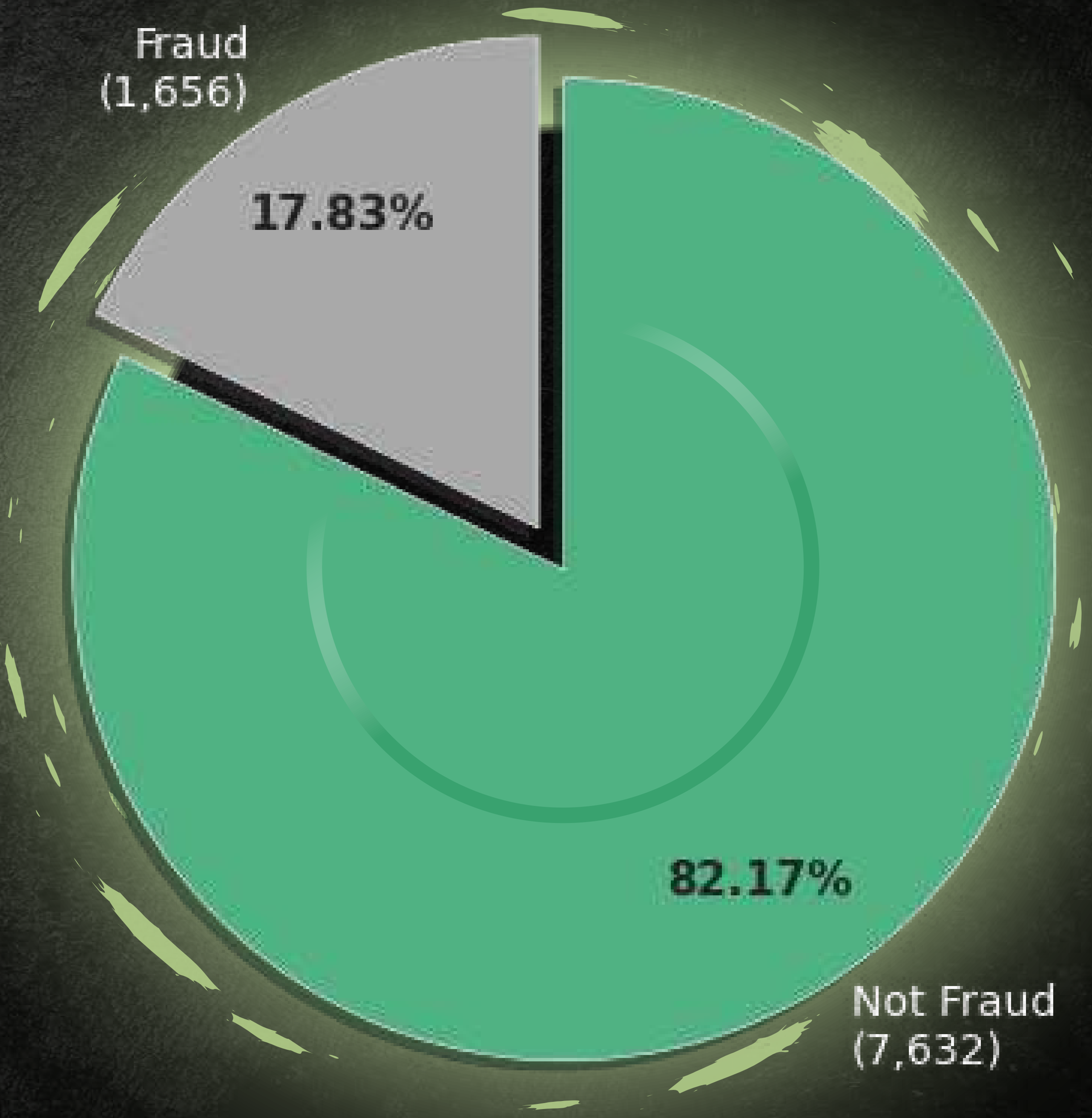
Machine Learning Model

The model was trained on 80% of the data, where the imbalanced data was addressed by applying the SMOTE oversampling method.

This method is utilized to increase the minority class synthetically.



Percentage of Fraud and Not Fraud Transaction Before Balancing



The study was considered using accuracy, AUC, precision, recall, and F1 score, but more strict about using recall.

MODEL	ACCURACY	AUC	PRECISION	RECALL	F1 SCORE
LOGISTIC REGRESSION	82.6695371	55.6188012	55.5555556	13.5951662	21.8446602
K-NEAREST NEIGHBOR	84.4994618	83.470937	54.3086172	81.8731118	65.3012048
DECISION TREE	90.9580194	91.1862408	68.3972912	91.5407855	78.2945736
RANDOM FOREST	91.2809473	91.5010179	69.2482916	91.8429003	78.961039
GRADIENT BOOSTING	97.3089343	96.4697875	90.2578797	95.1661631	92.6470588
XGBOOST	98.277718	97.177492	94.8948949	**95.4682779**	95.1807229
LIGHT GBM	98.4930032	97.1901543	96.3302752	95.1661631	95.7446809
CAT BOOST	98.2238967	96.7898076	95.4268293	94.5619335	94.9924127

In percent (%)

Best Possible XGBoost Classifier Model with Hyperparameter Tuning using GridSearchCV

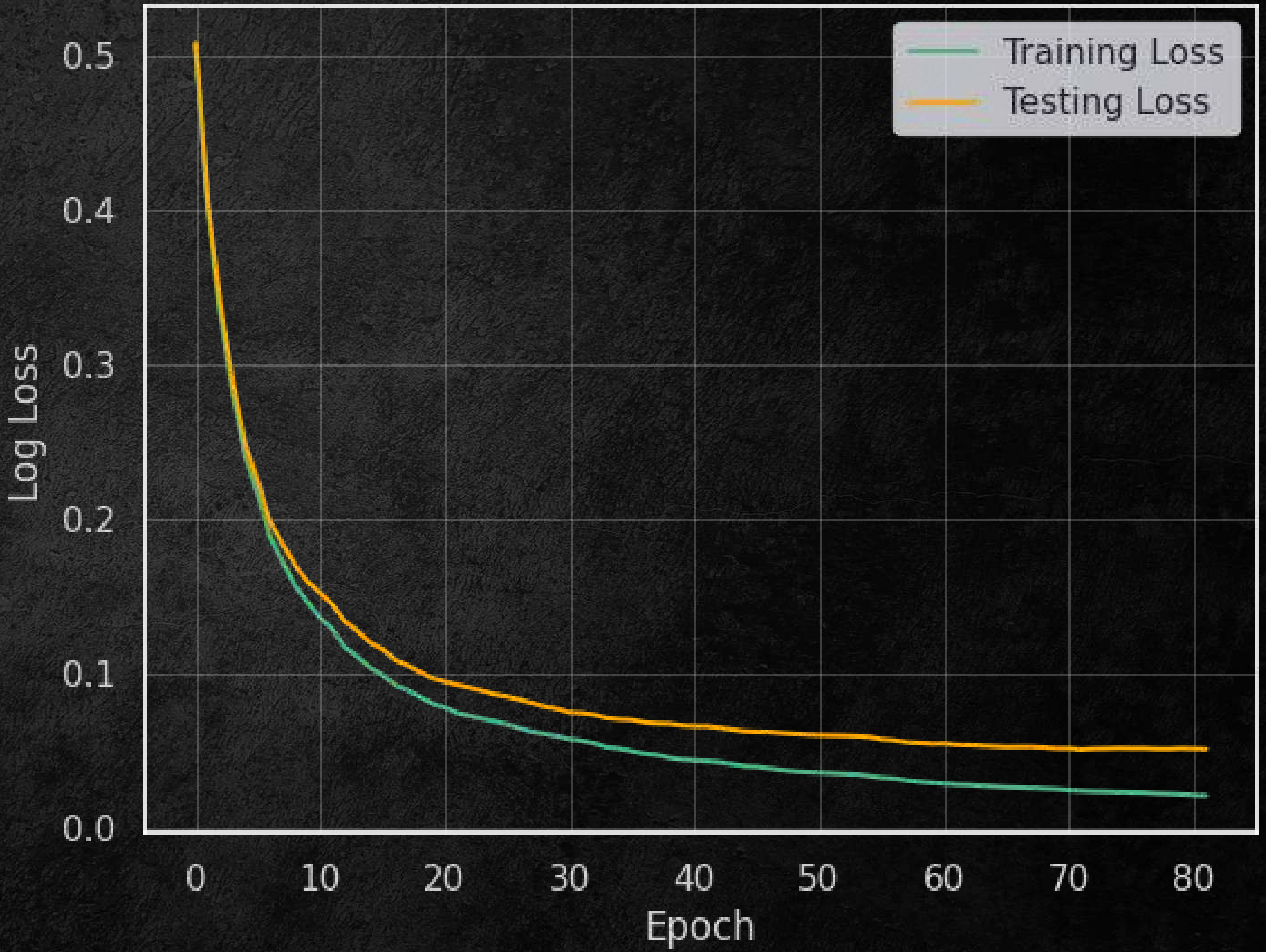
The model is able to catch **95.47%** of transaction actors who are actually fraudulent transactions among all fraudulent transactions predicted fraud.

Recall Score = 95.4682779

		Predicted Label	
		Not Fraud	Fraud
True Label	Not Fraud	1510	17
	Fraud	15	316

How many times should training be done on the dataset, with the aim of achieving optimal model performance?

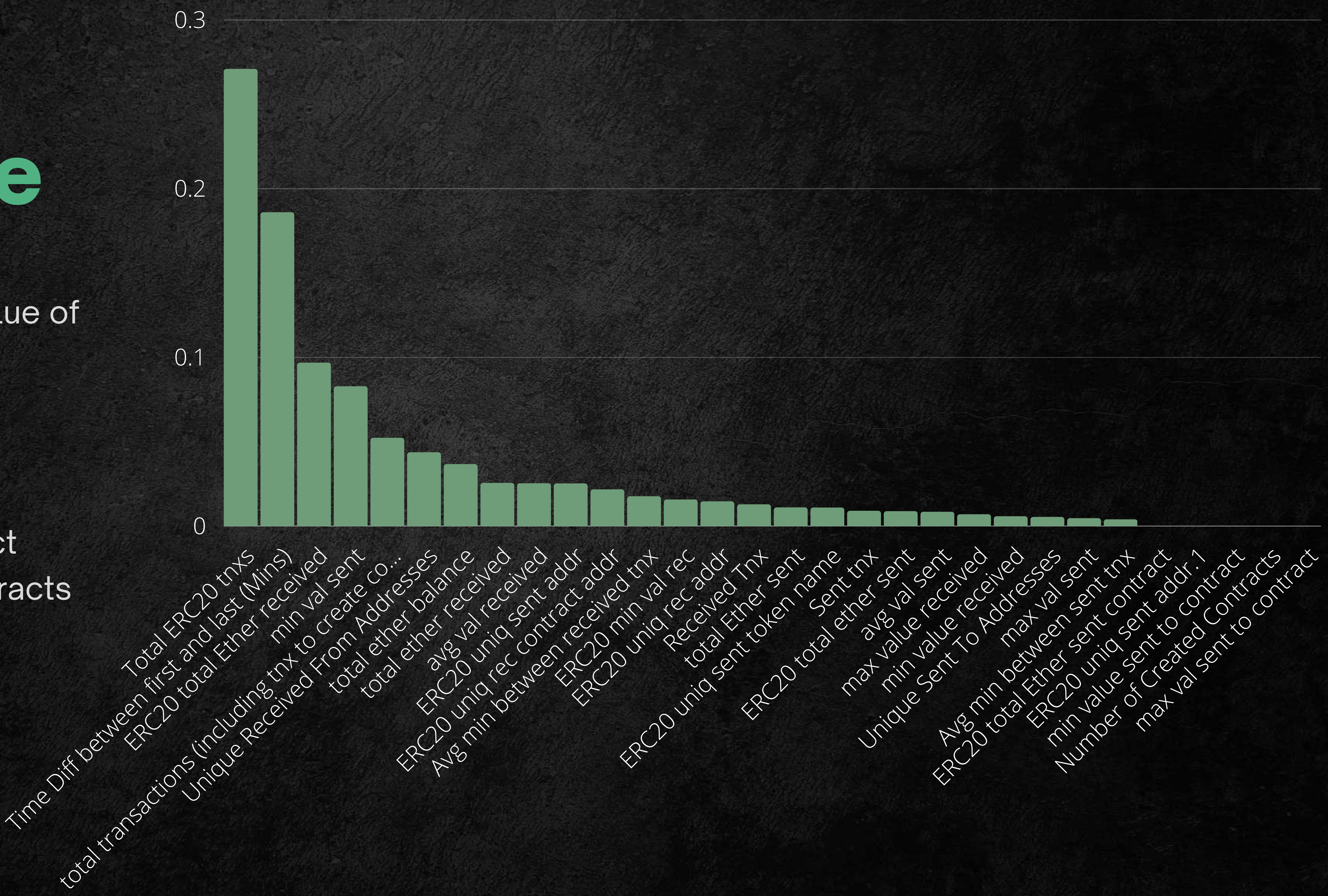
The **XGBoost Classifier** model learns and reduces errors in predicting the correct outcome. The improvement in model learning has stabilized around 20 to 80 epochs, as log loss no longer significantly decreases.



Feature Importance

Out of **30 features,** only the following features have a value of **0.00.**

- ERC20 total Ether sent contract
- ERC20 uniq sent addr.1
- min value sent to contract
- Number of Created Contracts
- max val sent to contract



Conclusion

By implementing SMOTE on our imbalanced dataset, we were able to address the issue of label imbalance, which resulted in a higher number of non-fraudulent transactions compared to fraudulent ones.

The use of hyperparameter tuning in the Extreme Gradient Boosting Classifier modeling has resulted in the model achieving optimal performance.

Key **Points**

- In cases where fraud occurs in the average minutes of sending and receiving transactions, the number of transactions is **below the range** of 1,000 transactions. The number of transactions on the assets sent is **less than** the accepted. The number of assets received conducted within 12 hours is **greater** than the other timeframe, whereas more transactions on the assets received that were not fraudulent usually took **more than one day to complete**.
- The gap between the first and last transactions was much during the **six months**, but the number was less than in non-fraud transactions, which is illogical.
- **The fewer** ERC20 token transactions for unique ERC20 tokens that are sent or received tend to indicate fraud.
- **The fewer** transactions with a unique account address, tend to indicate fraud.
- Based on total transactions (including transactions to create contracts), most fraud perpetrators make transactions at a **lower value** than usual in the **large transaction category.**

The fight against fraud demands **unwavering commitment** to leverage the latest tools and techniques **to safeguard the integrity of financial systems.**

Recommendation **Action**

For transaction participant

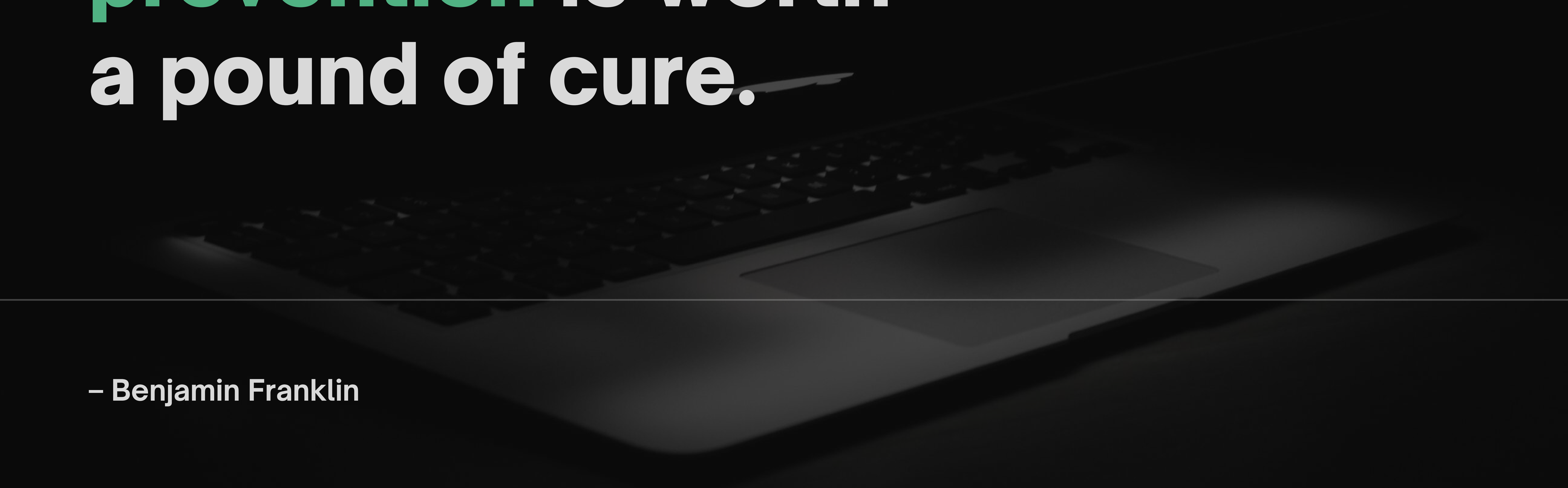
- Always **be vigilant and careful** in conducting transactions, such as ensuring the identity of the recipient of the transaction and not sharing sensitive information such as passwords and OTP codes with other people.
- Consistently **follow the security guidelines** of the platform used and update the security system on the device used to make transactions.

For centralized exchange platform

- Establish **strict rules** on suspected transactions, such as doing ECDD and blocking the account or transaction directly.
- Use **technology and data analysis** to monitor financial transactions and identify suspicious patterns or behavior using AML.
- Strength **cooperation** between platforms and related financial institutions in building a better security system and sharing information about detected fraud.
- Implement **appropriate sanctions** against accounts or actors found to be involved in fraud by reporting suspicious transactions to the authorities, such as supervisory and law enforcement authorities.
- Improve **systems and processes** that are vulnerable to fraud to prevent fraud from recurring in the future.

Profound Impact

- The risk of financial loss due to fraud can be minimized by **95.47%.**
- Increase customer trust in the platform.
- Companies can maintain or even increase their customer base and transaction volume, as well as provide a good reputation.
- With the adoption of more sophisticated technology to combat fraud, the company can position itself as an industry leader who is cutting-edge and innovative.



An ounce of prevention is worth a pound of cure.

– Benjamin Franklin

GitHub

https://github.com/moonarina