# Bootstrapped Ensemble Machine Learning Approach to Money Laundering Monitoring.

Presented By: Ombok Patrick Ouma

African Institute for Mathematical Sciences CAMEROON

February 15, 2025

Supervised By: Prof. Olawale Awe

# Table of Contents

# Introduction

As of October 2024, the updated grey list by Financial Action Task Force (FATF) had 13 out of 24 countries in this list from the African continent (54%). This is an indicator that their is need to improve the monitoring strategies in the identified jurisdictions.

While some financial institutions have invested in different money laundering monitoring systems,others still use manual ineffective techniques due to the high costs of such systems.

This presentation unveils how Bootstrapped Ensemble, Machine learning approach, specifically bagging and random forest algorithms can be used to identify suspicious transactions which is the starting point to curbing money laundering activities.

# Overview of Money Laundering

**Definition**

Money Laundering is defined as the process where illegally obtained money is made legitimate by passing it through financial systems.

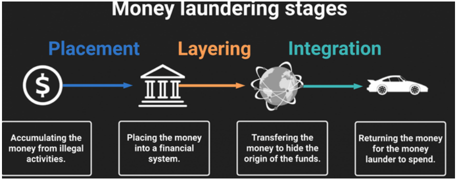



Figure: Stages of Money Laundering, Source:brittontime.com

# Overview of Money Laundering

**Effects of Money Laundering**

The effects have been categorized into two; at industry level and the overall impact in the economy of a country. Below are some of the effects;

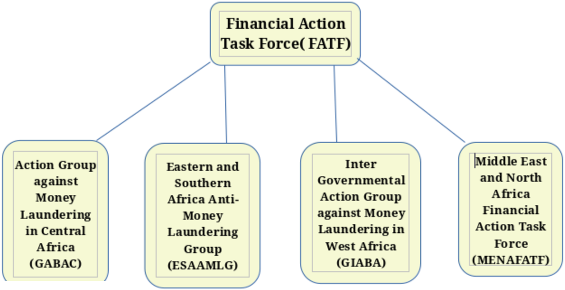
| Effects to the country | Effects to the entity |
|---|---|
| Reputational damage | Reputational damage |
| Increased corruption | Fines and penalties |
| Discourages foreign investments | Imprisonment. |
| International sanctions | Industry-based sanctions |

# Overview of Money Laundering

**The role of financial institutions in Combating Money Laundering**

- Conducting a proper KYC (Know Your Customer) and KYB (Know Your Business) on customers.
- Monitoring and reporting Suspicious Transactions to regulators.

**Financial Action Task Force (FATF) and It's African Network**



# Overview Of Money Laundering

**How Bootstrapped Ensemble Machine learning algorithms can be used to combat money laundering:**

- To improved detection accuracy of suspicious transactions.
- To reduce incidences of false positive cases thereby limiting resource wastage.

In this study the following machine learning algorithms are used;

- Bagging with Decision Trees.
- Random Forest

# Methodology

**Bagging with Decision Trees**

Bagging ,a short for bootstrap aggregating, was introduced by Leo Breiman in 1996. The name refers to how bagging achieves ensemble diversity through model aggregation.

Below, we walk through a step by step explanation of how bagging works and the mathematical principle behind it.

**Mathematical Principle of Bagging**

We have a labeled dataset:

$$D = \{(x_1^{(i)}, x_2^{(i)}, \ldots, x_x^{(i)}, y^{(i)}) \mid i = 1, 2, \ldots, N\}$$

The goal is to train $B$ decision trees and combine their predictions by majority voting.

# Methodology

**Step 1: Bootstrap Sampling**
For each $b$-th tree $(b = 1, 2, \ldots, B)$: Create a bootstrapped dataset $D^{(b)}$ by sampling $N$ data points with replacement from $D$.
**Step 2: Building a Decision Tree**
To build a decision tree, the algorithm selects the best feature to split the dataset. This is evaluated using metrics such as **Gini Impurity**.

**Objective Function**
The **Gini Impurity** for a node $T$ is defined as:

$$Gini(T) = 1 - \sum_{k=1}^{K} p_k^2,$$

# Methodology

When evaluating the split, the weighted Gini impurity for the child nodes is computed as:

$$Gini_{split} = \frac{N_L}{N} Gini(L) + \frac{N_R}{N} Gini(R),$$

where:

- $N_L$ and $N_R$ are the number of instances in the left and right child nodes, respectively.
- $N$ is the total number of instances at the current node.

The objective is to minimize $Gini_{split}$ across all possible splits.

**Step 3: Prediction Aggregation**

For a new instance $x$, the final prediction is determined by majority voting:

$$\hat{y} = \text{mode}(T_1(x), T_2(x), \ldots, T_B(x))$$

# Methodology

**Random Forest**

Random Forest is an ensemble method combining multiple decision trees for improved classification.

It was introduced by Leo Breiman in the 2000s and enhances generalization by adding randomness in feature selection.

The core principle is similar to Decision Trees, with the key distinction being randomized feature selection.

- Create a bootstrapped dataset $D^{(b)}$ by sampling $N$ data points with replacement from $D$.
- At each node, select $m$ features (where $m < d$) randomly from the total $d$ features.

# Methodology

- Choose the best split point based on criteria such as Gini Impurity.
  The Objective function is;

$$Gini_{split} = \frac{N_L}{N} Gini(L) + \frac{N_R}{N} Gini(R)$$

- Aggregate the prediction using majority voting technique;

$$\hat{y} = \text{mode}(T_1(x), T_2(x), \ldots, T_B(x))$$

# Application and Results

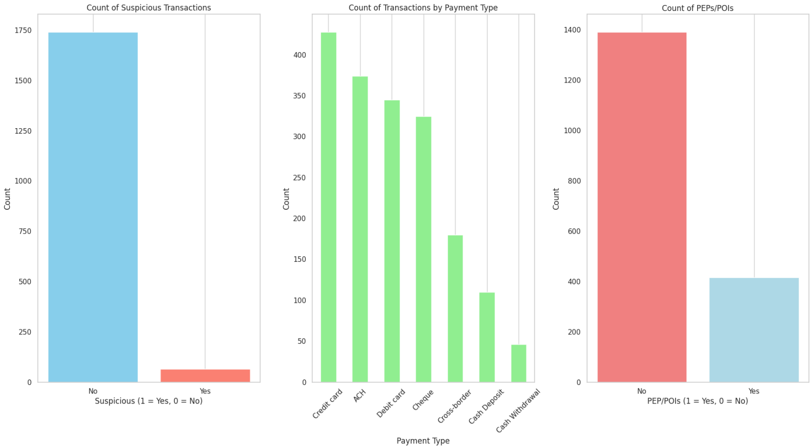
**Data Description**

The dataset used had 1808 observations and 10 features.

| Features | Description |
|---|---|
| Sender_account | The account number sending the money. |
| Receiver_account | The account number receiving the money. |
| Amount | The total monetary value. |
| Payment_currency | The currency sent. |
| Received_currency | The currency received. |
| Sender_bank_location | Country of sender's account. |
| Receiver_bank_location | Country of receiver's account. |
| Payment_type | The method of the transaction |
| PEP/POIs | Politically Exposed or Person of Interest. |
| Suspicious | suspicious transaction or not |

# Application and Results

## Exploratory Data Analysis




AIMS African Institute for Mathematical Sciences CAMEROON

# Application and Results

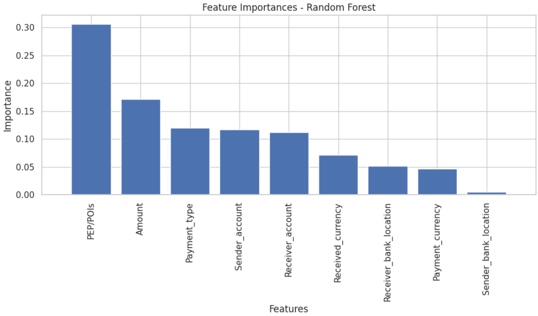
Below we present the results of the two models;

Table: Model Performance Metrics

| Metric | Bagging | Random Forest |
|---|---|---|
| Balanced Accuracy | 0.9670 | 0.9742 |
| F1 Score | 0.9650 | 0.9729 |
| Recall | 0.9759 | 0.9738 |
| AUC | 0.9949 | 0.9961 |
| Matthews Correlation Coeff | 0.9332 | 0.9483 |

The results clearly indicate that the Random Forest model outperformed Bagging Decision Tree classifier model.

# Application and Results




The most influencing feature in the prediction of the Random Forest Model is PEP/POIs, followed by the amount. The least influential feature is the bank location of the sender.

# Limitations

- The findings are limited to a financial institution that considers these type of features when assessing a suspicious transaction,commonly banking sector, and thus may not extend to other domains or use cases without further validation.
- Class imbalance between the suspicious and the non suspicious cases was addressed using SMOTE oversampling technique,however, generating synthetic instances in SMOTE can increase the computational load, especially if the dataset is too large.

# Conclusion

- Random Forest with it's high accuracy as compared to bagging is a cost-effective technique that can be used to identify suspicious transactions enabling financial institutions file Suspicious Transaction Reports (STR) promptly.
- With the use of this technique, reporting entities are allowed enough time to scrutinize a transaction thoroughly before reporting it to the regulators thus limiting cases of false positives.
- The feature importance can be used as a guide to a quick check on what features are of great weight to why a transaction is considered suspicious. However, care should be taken to ensure all the supporting evidences are tabled without purely depending on the model.

# References


- The Three Stages of Money Laundering and How Money Laundering Works by Financial Crime Academy: https://financialcrimeacademy.org/the-three-stages-of-money-laundering/.
- Financial Action Task Force website (FATF):https://www.fatf-gafi.org/en/home.html
- Awe, O. O., Ojumu, J. B., Ayanwoye, G. A., Ojumoola, J. S., and Dias,R. (2024). Machine Learning Approaches for Handling Imbalances in Health Data Classification. In Sustainable Statistical and Data Science Methods and Practices: Reports from LISA 2020 Global Network, Ghana, 2022 (pp. 375-391). Cham: Springer Nature Switzerland.

# Thank you for listening!