Anti-Money Laundering (AML) System Development

Rohit Eerabattini, y72@umbc.edu
Shashank raj gupta Gunta, shashag1@umbc.edu
Yeshwanth Thalapaneni, rohite1@umbc.edu

DATA 603: Platforms for Bigdata Processing,
Data Science. Prof. Melih Gunay

INTRODUCTION

In the complex landscape of global finance, money laundering emerges as a critical challenge, threatening the integrity and stability of financial systems worldwide. This illicit activity involves disguising the origins of illegally obtained money, enabling criminals to infuse their gains into the economy as legitimate funds. The repercussions of unchecked money laundering are profound, influencing not just economic stability but also facilitating further criminal activities.

Addressing this critical issue, our project focuses on the development of an advanced machine learning-based Anti-Money Laundering (AML) detection system. This system is designed to sift through vast amounts of transactional data to identify anomalies and patterns indicative of laundering activities. The primary goal is to flag suspicious transactions for further investigation, thus aiding financial institutions in their compliance with regulatory requirements and their fight against financial crimes.

Utilizing a dataset of 2.77 GB, specifically the LI-Medium_Trans.csv, our approach integrates comprehensive data cleaning, exploratory data analysis (EDA), feature engineering, and the application of sophisticated machine learning models, including Logistic Regression and Random Forest Classifier. The choice of technologies—Pyspark, pandas, seaborn, matplotlib, and scikit-learn—supports an efficient analysis pipeline capable of handling large-scale data with high performance.

By bridging advanced data analytics and machine learning, our system not only enhances the detection of suspicious transactions but also contributes to a broader understanding of money laundering dynamics. This report outlines the objectives, methodologies, challenges encountered, and the significant results achieved through our system, providing a blueprint for leveraging technology in the fight against money laundering.

OBJECTIVE

The primary objective of this project is to develop a highly effective machine learning-based Anti-Money Laundering (AML) detection system capable of identifying suspicious financial transactions in real-time. Given the sophistication with which illicit activities are masked within voluminous financial data, our system aims to enhance the detection capabilities of financial institutions, ensuring the integrity and stability of the financial ecosystem. By pinpointing potentially fraudulent transactions, the system assists in compliance with regulatory standards and plays a crucial role in curtailing the financial underpinnings of global criminal enterprises.

STATE OF ART

The conventional methodologies in AML detection largely revolve around rule-based systems and traditional statistical methods. While these have served as foundational tools, they are often plagued with limitations such as high rates of false positives and a lack of adaptability to new, sophisticated laundering techniques. In response to these challenges, recent advancements in machine learning and big data analytics have paved the way for more dynamic and precise detection systems.

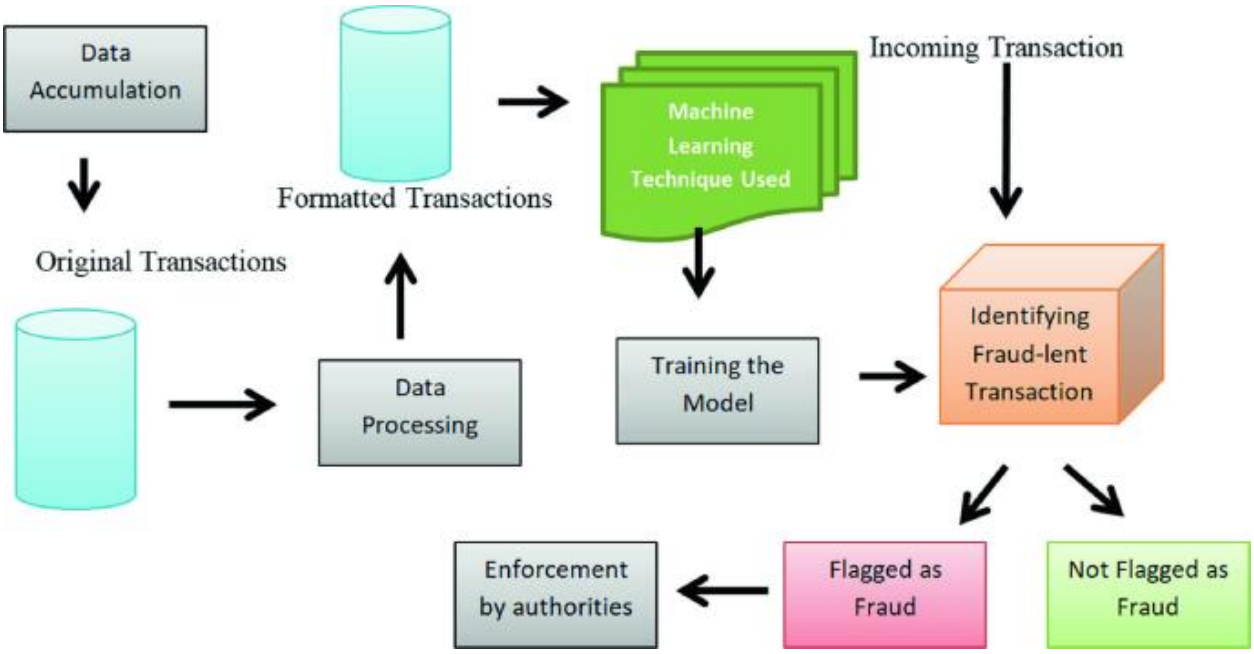
Machine learning models such as decision trees, random forests, and logistic regression have significantly improved the identification of complex patterns indicative of money laundering. These models are not only more flexible but also more powerful in detecting subtle anomalies in data compared to traditional methods. Furthermore, the advent of deep learning techniques, especially neural networks, has revolutionized the field. These models excel in modeling intricate relationships between diverse features and uncovering hidden patterns that may elude conventional approaches.

To further refine detection accuracy, the integration of time-series analysis and anomaly detection methods has been instrumental. These techniques are adept at recognizing sudden, irregular shifts in transaction patterns that may signify illicit activities. Additionally, the adoption of ensemble methods, which combine multiple machine learning models, has emerged as a superior strategy in AML detection. By leveraging the strengths of various predictive models, ensemble approaches reduce false positives and enhance the overall robustness of the detection system.

This project seeks to harness these state-of-the-art methodologies to construct a robust AML detection system that not only addresses the current gaps in detection capabilities but also sets a new standard for future developments in the field.

Proposed Architecture and Technologies

Our AML detection system is built on a stack comprising Pyspark, pandas, seaborn, matplotlib, and scikit-learn. The architecture is designed to process large datasets efficiently, leveraging Pyspark for distributed data processing, pandas for data manipulation, seaborn and matplotlib for visualization, and scikit-learn for implementing machine learning models.



Problems and Challenges

Data Cleaning:

Encountering missing values and inconsistent data formats posed significant challenges. We implemented stringent data preprocessing protocols using pandas to ensure data quality and consistency.

Performance Issues:

The large dataset size (2.77 GB) led to prolonged processing times. By employing Pyspark for its distributed computing capabilities, we managed to optimize performance and reduce time delays in data processing and model training.

Methodology

Data Cleaning:

We began by cleaning the dataset, addressing missing values and standardizing formats for uniform processing.

Exploratory Data Analysis (EDA):

EDA was conducted to understand the underlying patterns and anomalies in the data, using seaborn and matplotlib for visualization.

Feature Engineering:

Critical features were engineered to enhance model performance, focusing on aspects that are indicative of laundering activities.

Model Building:

Two models were built:

i) Logistic Regression
ii) Random Forest Classifier

The logistic regression model outperformed the random forest, providing better accuracy in detecting laundering activities.

Results and Discussion

In our efforts to develop a robust Anti-Money Laundering (AML) detection system, we employed two primary machine learning models: Logistic Regression and Random Forest Classifier. The comparative analysis of these models yielded insightful results, which are critical in understanding their effectiveness in detecting potentially fraudulent transactions.

Model Performance Comparison:

The Logistic Regression model outperformed the Random Forest Classifier in terms of overall accuracy. Specifically, the accuracy metrics indicated that Logistic Regression was more adept at identifying true laundering cases as well as legitimate transactions. This superior performance can largely be attributed to the model's ability to capture linear relationships between key features and the laundering outcome, which appeared to be more pronounced in our dataset.

On the other hand, while the Random Forest model generally provides excellent performance across various scenarios by capturing non-linear interactions, it did not perform as well in this

particular setting. This observation suggests that the key indicators of money laundering in our dataset were more linearly separable, a scenario where logistic regression typically excels.

Detailed Metrics Evaluation:

To provide a comprehensive evaluation, we assessed the models using several metrics:

Accuracy: As noted, logistic regression showed higher accuracy than random forest.

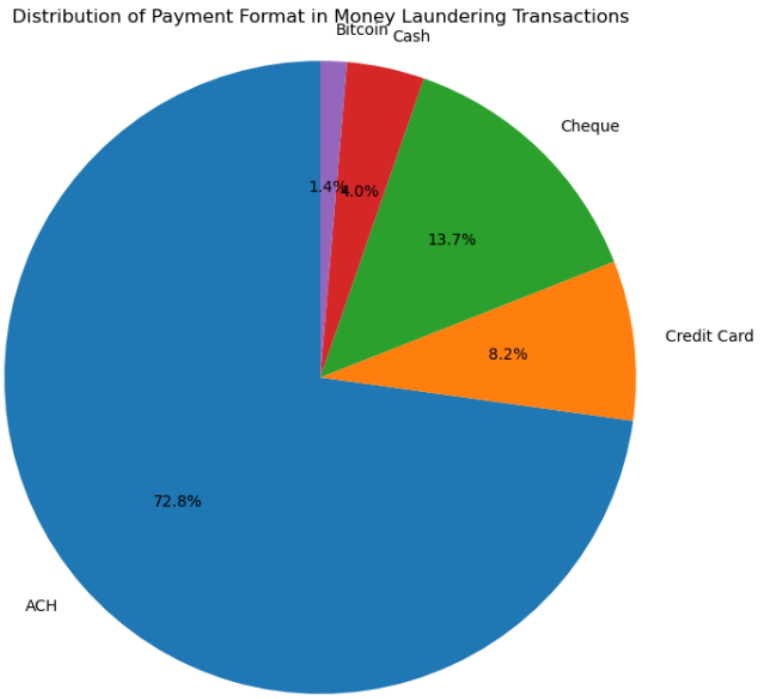
Precision: Indicates the proportion of identified positive identification results that were actually correct. Logistic Regression also led in this metric, suggesting it was less likely to falsely label legitimate transactions as suspicious.

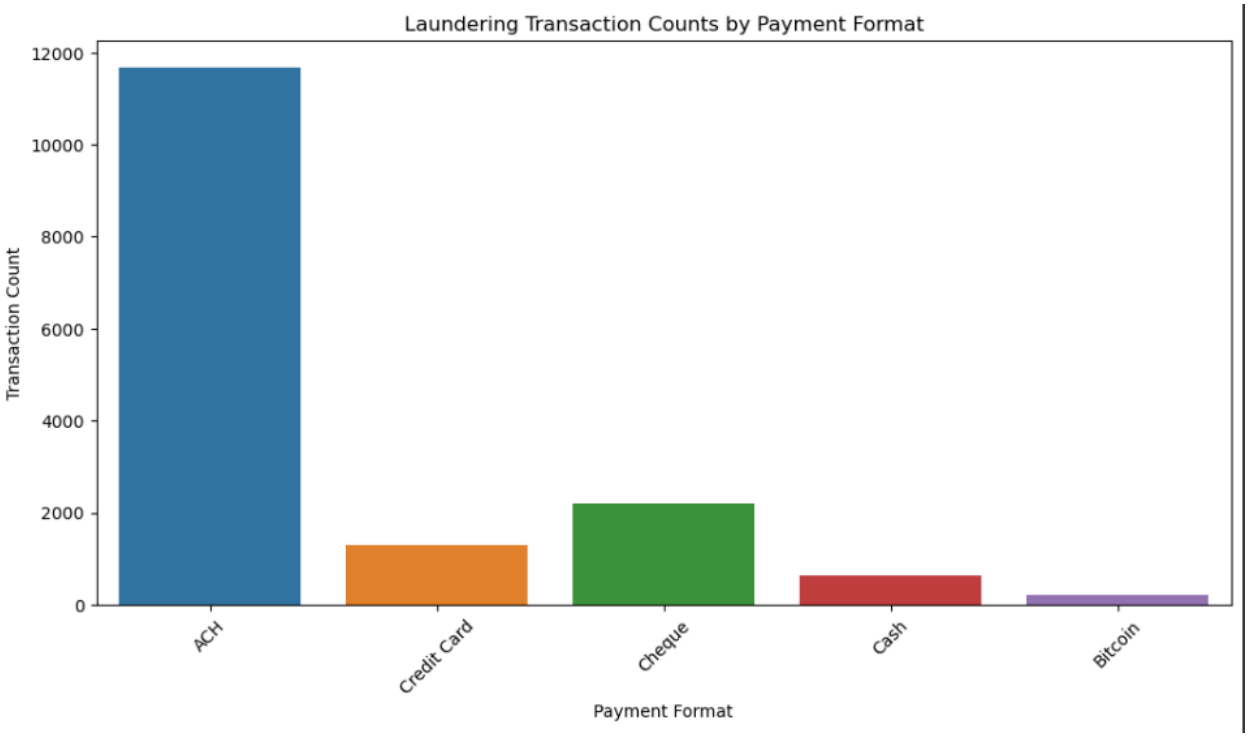
Recall: Measures the ability of the model to detect all relevant instances. Here, Logistic Regression had a slight edge, indicating it missed fewer actual laundering cases.

Analysis by Currency and Transaction Type:

Further analysis was conducted based on different currencies and transaction types. This segmentation provided deeper insights into the model's performance across various transaction scenarios. The Logistic Regression model demonstrated consistent superiority across most currency types, particularly in handling major currencies where laundering activities are often more nuanced and complex.

Moreover, when evaluating transaction types—ranging from simple transfers to more complex financial instruments—the Logistic Regression model maintained higher accuracy and reliability, reinforcing its suitability for diverse financial environments.


Distribution of Payment Format in Money Laundering Transactions
Bitcoin
Cash
Cheque
1.4%
4.0%
13.7%
Credit Card
8.2%
72.8%
ACH


Laundering Transaction Counts by Payment Format
Transaction Count
12000
10000
8000
6000
4000
2000
0
ACH
Credit Card
Cheque
Cash
Bitcoin
Payment Format

Conclusion

The findings underscore the importance of choosing the right model based on the characteristics of the data at hand. For datasets where linear relationships predominate, as in our case, Logistic Regression can provide more accurate and reliable results. These insights not only validate the effectiveness of our AML detection system but also guide future enhancements and optimizations to ensure even greater efficacy in combating money laundering activities.