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 intricate financial environment of the world, money laundering stands out as a major problem that threatens the integrity and stability of financial systems all over the globe. This criminal activity is about illegal money origin, which is then disguised as legal money. Thus, the criminals can put their gains into the economy as a legal amount of money. The consequences of unchecked money laundering are hopeless, which means that it affects not only the economy but also enables other crimes.

The main purpose of our project is to solve the important problem of the AML detection system, which is based on machine learning. This system focuses on the fact that it will filter through the huge amount of transaction data to find the anomalies and patterns that are characteristic of money laundering. The main purpose of blocking such transactions is to alert the suspicion of the authorities for further investigation that is beneficial to the financial institutions in their compliance with the laws or combating of financial crimes.

Through the application of a dataset of 2, the professional speaker has utilized a useful tool to measure the quality of their speech. The limit is 77 GB, which is the LI-Medium_Trans specific one. CVS, our strategy involves thorough data cleaning, exploratory data analysis (EDA), feature engineering, and the use of complex machine learning models, which are Logistic Regression and Random Forest Classifier. The selection of technologies such as pyspark, pandas, seaborn, matplotlib, and scikit-learn provides an efficient analysis pipeline that can process large-scale data with great performance.

By combining the latest data analytics and machine learning, our system not only improves the detection of suspicious transactions but also helps the development of overall knowledge about money laundering schemes. This paper is about the objectives, methods, difficulties, and the main results that our system has achieved in the struggle against money laundering as a result of the technology use.

OBJECTIVE

The main idea of this project is to create a highly efficient machine learning-based Anti-Money Laundering (AML) detection system which will be able to find suspicious financial transactions in real time. The system aims to strengthen the ability of financial institutions to detect the illicit activities hidden in a large amount of financial data so that the financial system of the world will be more stable and integrity-based. The system does the job of finding potentially fraudulent transactions; hence, it helps to comply with the regulations and plays a vital role in the reduction of the financial backing of global criminal groups.

STATE OF ART

The staple of AML detection techniques are usually rule-based systems and statistical methods. On the one hand, these have been the basic tools, but on the other hand, they are usually hampered by problems such as a high rate of false positives and the incapability of new methods to deal with the new, sophisticated laundering techniques. Thus, in the present day, machine learning and big data analytics technology have solved the problems of the past and improved detection systems.

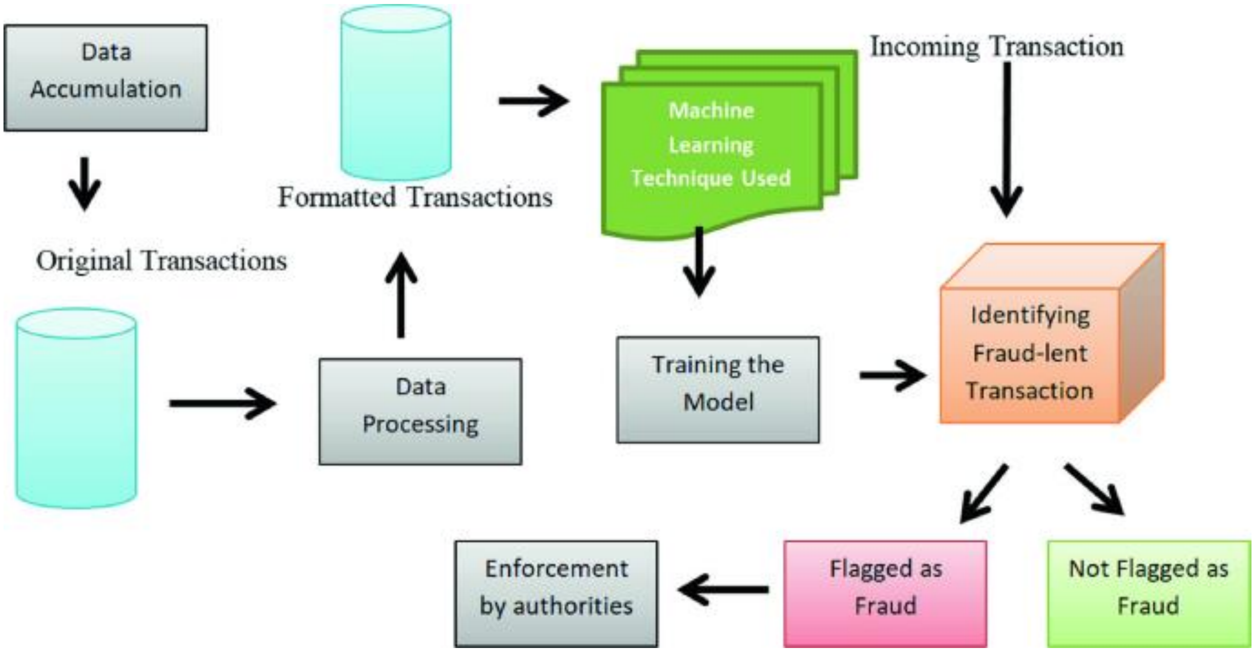
Machine learning models like decision trees, random forests, and logistic regression have been used in the identification of complex patterns that are the indicators of money laundering, and this has been done in a remarkable way. These models are not only more convenient but also stronger in the detection of slight anomalies in data than the conventional ones. Besides, the emergence of deep learning techniques, mainly neural networks, has totally changed the scene. These models are the ones that are good at making the complex relations between the different features and the hidden patterns those conventional methods may not discover.

The combination of time-series analysis and anomaly detection techniques has been a great help in the improvement of detection accuracy and, thus, the development of an advanced detection system. These methods are good at the identification of unexpected, irregular changes in the transaction patterns that one can use to prove a case of illegal activities. Besides, the usage of ensemble methods, which are the combination of several machine learning models, has appeared as the best method in AML detection. Through the use of the strengths of various predictive models, ensemble methods decrease the false positives and, at the same time, increase the accuracy of the whole detection system.

This project aims to use these cutting-edge technologies to build a strong AML detection system that will not only cope with the existing problems in detection capabilities but also set a new benchmark for future developments in the field.

Proposed Architecture and Technologies

Our anomaly detection system is comprised of Pyspark, pandas, seaborn, matplotlib, and scikit-learn. The architecture is created to process large datasets in a fast and efficient manner, concentrating Pyspark for distributed data processing, pandas for data manipulation, seaborn and matplotlib for visualization, and scikit-learn for the implementation of 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 are two linear and non-linear models for classification. The report that was done on the comparison of these models was very useful, as it gave an interesting result, which was the key to understanding their ability to detect possible fraudulent transactions.

Model Performance Comparison:

The Logistic Regression model was the best in the comparison with the Random Forest Classifier in terms of the overall accuracy. For instance, the accuracy metrics showed that Logistic Regression was better at detecting the real ones as well as the real transactions. The above, we owe the best results to the model's capability of picking linear relationships between the important features and the laundering outcome which was more visible in our data.

To sum up, the Random Forest model usually delivers high-quality results in different situations by taking into account the non-linear interactions, but it could not do well in this case. This result stated that the main features of money laundering in the dataset were more linearly separable, a situation where logistic regression usually does better.

Detailed Metrics Evaluation:

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

Accuracy: As it is well known, logistic regression was the one that demonstrated the highest accuracy compared to the random forest.

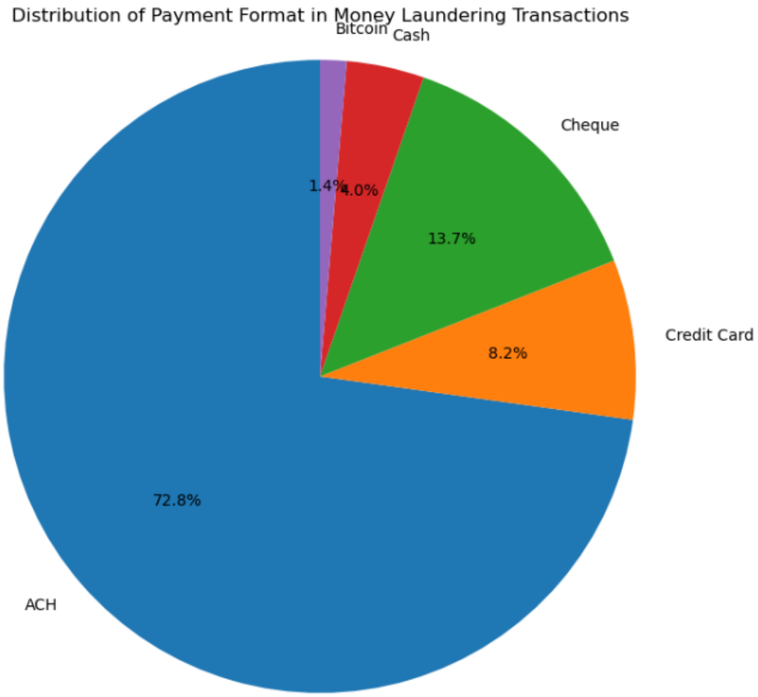
Precision: It is the percentage of the tested positive results that were indeed true. Logistic Regression was also the winner of this metric, which meant it reduced the chances of false identification of real transactions as suspicious.

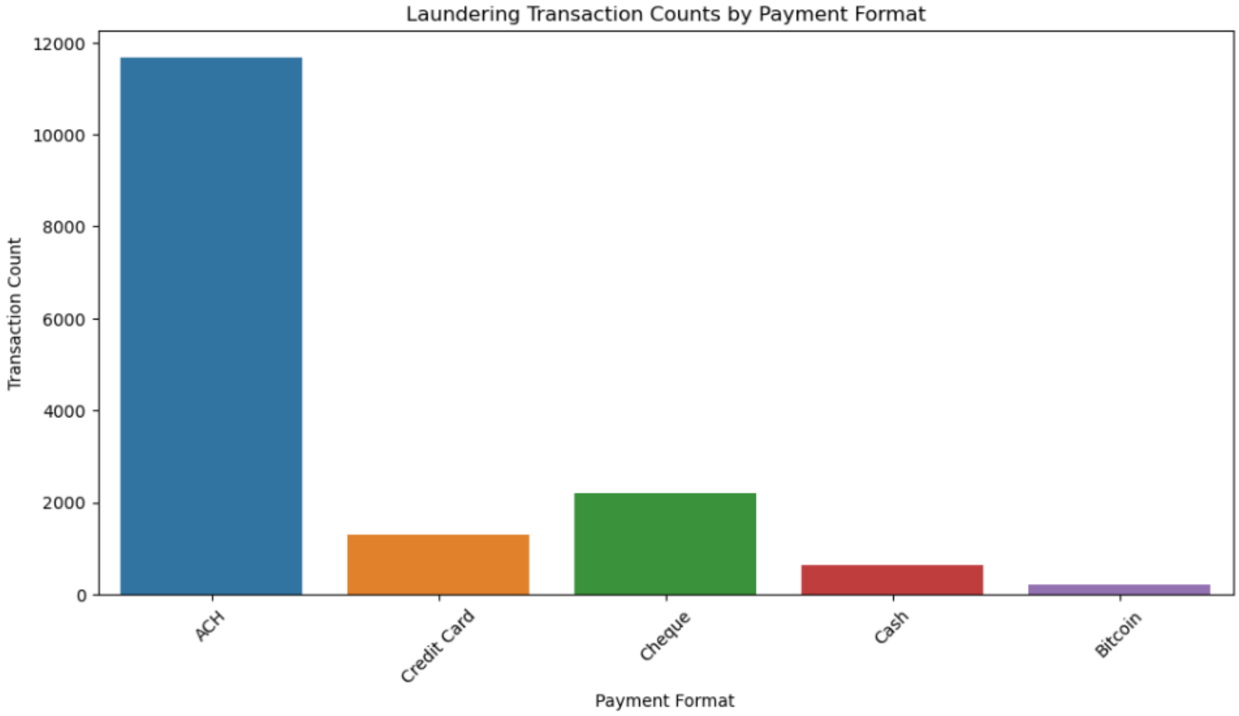
Recall: The approach determines the skill of the model to spot all the relevant ones. Here, Logistic Regression had a barely noticeable superiority because of the fact that it overlooked fewer actual laundering cases.

Analysis by Currency and Transaction Type:

The analysis was done by using different currencies and the types of transactions. This breaking down of the data brought to the surface the more profound ideas about the model's performance in different transaction situations. The logistic Regression model showed its steady performance in most of the currencies, especially in the case of the major currencies where laundering was more complicated and sophisticated.

Besides, even though the Logistic Regression model was applied to various transaction types, from simple transfers to more complicated financial instruments, it showed higher accuracy and reliability and, thus, was proven to be a good model for different financial environments.


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


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 discoveries reiterate the necessity of the correct choice of the model according to the data features. In the cases where the linear relationships rule the data, Logistic Regression can give more accurate and dependable results. These insights not only prove the effectiveness of our AML detection system but also control the course for future improvements and optimizations and thus guarantee even greater efficacy in the battle against money laundering.

Github Repo: https://github.com/shashank080/Anti-MoneyLaunnderingDetectionSystem/invitations