Toronto Metropolitan University

DS 8012 - Research Skills - W2025 - Major Research Project Proposal

Detecting Money Laundering Using Advanced Machine Learning Techniques

Submitted to:
Dr. Ceni Babaoglu

Submitted by:
Avi Kaushik (Student ID - 501338331)

1. Introduction

One serious financial crime that seriously jeopardizes international financial stability, economic security, and regulatory frameworks is money laundering. Criminal groups use complex techniques to hide the source of illegal cash by mixing them with legal transactions, which makes detection very difficult. In order to detect and stop such operations, financial institutions around the world are under tremendous pressure to adhere to anti-money laundering (AML) legislation.

Conventional AML systems mostly rely on frameworks based on rules, which make it difficult for them to adjust to new money laundering methods. These systems frequently produce an excessive amount of false positives, which puts an enormous burden on compliance teams and raises operating expenses significantly. Furthermore, complicated laundering patterns involving several tiers of transactions and accounts might go undetected by these rule-based methods.

A possible substitute is provided by advanced machine learning techniques, which use transaction data patterns to more accurately and efficiently identify questionable activity. ML models can improve the detection of suspicious transactions, lower false positive rates, and automatically adjust to evolving money laundering strategies. In order to overcome the drawbacks of conventional rule-based frameworks, this project intends to create a sophisticated machine learning (ML)-based AML detection system using the IBM Transactions for Anti-Money Laundering dataset.

2. Problem Statement: Despite ongoing efforts to combat financial crime, money laundering detection remains a persistent challenge for financial institutions. The key issues include:

- **High False Positive Rates:** Traditional AML systems generate excessive false alerts, reducing the efficiency of compliance teams and increasing costs.
- **Complex Laundering Techniques:** Criminal organizations utilize multi-step processes such as layering and integration to obscure transaction trails, making it difficult to track suspicious activity.
- **Data Imbalance:** Money laundering data is highly imbalanced, with genuine suspicious transactions being a small minority among normal transactions, making supervised learning models prone to bias.
- **Limited Real-World Model Adoption:** While machine learning approaches have shown promise in research, their real-world adoption is hindered by scalability issues, data quality constraints, and regulatory complexities.

To address these challenges, this project aims to build a robust AML detection system utilizing advanced ML techniques that leverage anomaly detection, behavioral analysis, and transaction scoring to improve detection accuracy and reduce false positives.

3. Dataset Description: The IBM Transactions for Anti-Money Laundering dataset is a comprehensive dataset designed to simulate real-world financial transaction patterns. It provides transaction records that are rich in attributes relevant to identifying suspicious behavior. Key features of the dataset include:

- **Transaction Details:** Information about transaction amounts, types (e.g., cash-in, cash-out, transfer), and timestamps to analyze patterns over time.
- **Customer Profile Data:** Includes metadata such as customer account details, transaction frequency, and average transaction size to support behavioral analysis.
- **Labeled Transactions:** The dataset includes labels identifying suspicious and non-suspicious transactions, enabling the development and evaluation of supervised learning models.
- **Complex Transaction Chains:** The dataset reflects common money laundering patterns such as structuring, layering, and integration, providing a robust foundation for anomaly detection and ML model training.

By leveraging this dataset, the project aims to build a comprehensive AML detection framework capable of identifying suspicious activities efficiently and effectively.

4. Proposed Machine Learning Models: To achieve robust money laundering detection, the following machine learning models will be explored:

- **Random Forest (RF):** Effective for identifying feature importance, suitable for high-dimensional data, and robust in handling imbalanced datasets.
- **Isolation Forest:** An unsupervised anomaly detection model designed to isolate anomalies efficiently, ideal for detecting rare laundering transactions.
- **Autoencoders:** Deep learning models effective for capturing complex patterns in transaction data, particularly useful for unsupervised anomaly detection.
- **XGBoost:** A powerful gradient boosting algorithm that is widely known for its high performance in classification tasks.

- **Deep Neural Networks (DNN):** Capable of learning intricate patterns in transactional data, especially effective when combined with engineered features.

Each model will be evaluated using key metrics such as Precision, Recall, F1-Score, and ROC-AUC to ensure robust and accurate detection of suspicious activities. Additionally, ensemble learning techniques will be explored to combine the strengths of individual models and enhance overall performance.