ITMD 524

Title: Anti-Money Laundering Detection

Team Members:

First name	Last Name	IIT Email
Sophia	Robles	Srobles2@hawk.iit.edu
Nicholas	Simpkins	nsimpkins@hawk.iit.edu
Harsh	Patel	hpatel100@hawk.iit.edu **(Using the same project for ITMD 522)**

Important Notes:

- Each group must submit ONLY one copy by a single team member!
- Do not worry! Since you will list all team members in the table above.

1. Introduction

Question:
Introduce the background of your application and give me the motivations why you want to do that.

Answer:
Money laundering has emerged as a multi-trillion-dollar international underground market amidst increasing governance of capital movement and tracking of such markets through the digitization of currency handling. The IBM Transactions for Anti-Money Laundering (AML) dataset simulates a combination of 13 million suspicious and non-suspicious banking transactions across an array of synthetic customers. This research aims to explore this dataset to develop and evaluate machine learning models capable of parsing abnormal behaviors across fraudulent transactions—assisting regulatory AML institutions in tracking criminal activity.

2. Data Sets

Question:
Briefly introduce your data sets, such as which application or domain the data belongs to, where did you collect it, how large it is, how many features there are, what is your target variable, and so forth

Answer:
We are using two datasets, both of which are filled with transaction data. Some of the transactions are involved in money laundering. The goal is to use this data to train our model to detect money laundering patterns. One dataset includes a high activity of money laundering transactions while the other one includes a lower activity of money laundering transactions. The goal is to train with the high money laundering activity dataset and then test on the lower activity dataset. Combined they have approximately 13 million transactions (rows).

Tell where the data is, such as giving Kaggle URL
https://www.kaggle.com/datasets/ealtman2019/ibm-transactions-for-anti-money-laundering-aml

Size of the data:
650.42MB, 475.66
7 million, 5 million rows

Tell the variables (X = Independent variables + Y = Dependent variables)
Timestamp *(dtype: object)*, Sender Bank Code *(dtype: int64)*, Starting Account *(dtype: object)*, Receiver Bank Code *(dtype: int64)*, Ending Account *(dtype: object)*, Amount Received *(dtype: float64)*, Receiving Currency *(dtype:*

object), Amount Paid *(dtype: float64)*, Payment Currency *(dtype: object)*, Payment Format *(dtype: object)*, isLaundering *(dtype: int64)*
Total dtypes: float64(2), int64(3), object(6)

Tell the target variables, if you are going to perform predictive tasks

- isLaundering (For main classification),
- Amount Received, Amount Paid, Payment Currency, Payment Format (Possible but depends on data exploration)

3. Research Problems

Question:

- List your research problems, that is, what kinds of the problems you want to solve.
- You cannot simply say I want to explore the data and find the patterns
- If you decide to work on a classification task, you must identify labels.
- If your project is involved with multiple data mining tasks, you should clearly mention each problem and why you want to do that.
- You should provide finer-grained research problems that can be solved by data analysis/mining techniques. If it is an implementation project, you should introduce the challenges in implementing or development and how you will evaluate them

Answer:

Data mining task: (Data Preprocessing and Sampling)

1. Clean the data
2. OneHotEncoding: Encode categorical features as a one-hot numeric array (i.e., whenever needed).
3. Possible dataset challenge:
    a. To solve target variable class imbalance: By using an undersampling and oversampling tasks.
        i. The reason is there are only 5-10% instances where isLaundering is true. (i.e., 1). Example techniques are SMOTE, Random Oversampling/Undersampling, Cost-sensitive learning.
    b. Noise in the data:
        i. We need to construct models with regularization parameters.

Possible Research Problems exploration:

- For classification, Supervised machine learning on target variable "isLaundering" – our primary label:
    - Using KNN, NB, Decision Trees, SVM, Logistic Regression, Neural Network, Ensemble classifiers (i.e., Random Forest), and its binary classification for isLaundering.
    - Possible algorithms exploration: Decision Trees, Random Forest, XGBoost, Neural Networks
    - Goal: Determine prediction if the transaction exchange between accounts is fishy/Non-fishy as target variable.
- Unsupervised Learning:
    - To detect anomalies without labeled data. (i.e., to resemble actual shady transactions)
    - What an abnormality is in the overall picture of money transactions and making sure that any abnormalities found match up with real activity that is suspicious?
    - Possible algorithms exploration: Isolation Forest, One-Class SVM, Autoencoders
- Graph-based Analysis:
    - Find groups or colluding networks (money-laundering operations).
    - Graph Theory + Graph Nural Networks, Network Analysis
- Pattern Identification:
    - Time-series Analysis (i.e., Consecutive Patterns in AI)
    - Possible algorithms exploration: Recurrent Neural Networks (RNN), LSTMs, Hidden Markov Models

We would like to identify money laundering transactions in large datasets with a high degree of accuracy and true positives.

4. Potential Solutions

Question:
For each problem you list above, figure out feasible solutions, and introduce your plan to perform experiments

Answer:
This is my overall plan to perform tasks. All work will be in python notebook.

1. Load the dataset and understand it. Check the number of unique values and how many times it appears. Check data types and overall structure of the dataset.
2. Clean the data
    a. Handling missing values if any.
    b. Normalized if needed.
    c. OneHotEncoding: Encode categorical features as a one-hot numeric array.
    d. To solve target variable class imbalance: By using an undersampling and oversampling tasks.
3. Visualize the data.
    a. Class distribution, frequency of transaction from single user, top suspicious patterns
4. Get the basic dataset patterns:
    a. Average transaction per hour, minimum, maximum and median, payment mode.
    b. Cognitive attributes: divergence of the user's average activity.
    c. Possible graph-based capabilities include the total amount of links and circulation across financial records.
5. For classification, Supervised machine learning on target variable "isLaundering" – our primary label:
    a. Determine whether an exchange is fishy.
    b. By using, Decision Trees, Random Forest, XGBoost, Neural Networks.
    c. Evaluation: Accuracy, Precision, Recall, F1, AUC-ROC, and confusion matrix (i.e., TP, FP, TN, FN)
6. Unsupervised Learning:
    a. To detect anomalies without labeled data (i.e., that set off from predictable trends.)
    b. Isolation Forest, One-Class SVM, Autoencoders, and possible clustering.
    c. Evaluation: How most identified anomalies that set off from predictable trends.
7. Graph-based Analysis:
    a. Using graph attributes (the degree and relevance) and to identify fraud and cyclic transactions, where,
        i. Nodes = Accounts and Edges = Transactions
    b. Possible solution: Graph Theory + Graph Nural Networks, Network Analysis
    c. Evaluation: Recognizing patterns (i.e., movable and clustering data)
8. Pattern Identification in a time-series analysis of the data.
    a. Possible algorithms exploration: Recurrent Neural Networks (RNN), LSTMs, Hidden Markov Models
9. See the predictions of suspicious behavior, evaluation matrices and visualization using heatmaps and graphs.

5. Evaluations

Question:
There could be multiple solutions for a same problem, You must figure out how to evaluate them and the details about your evaluations, for example, hold-out or N-folds evaluation?, which metrics you will use for evaluations.

Answer:
- We are using 80:20 (Training: Testing) dataset OR we can try out N-fold validation for more accurate results.
- The evaluations details are described in section-4 (i.e., answer-4) already.

6. Expected Outcomes

Question:
Introduce your expected outcomes for your project

Answer:
Overall goal:

- The expected outcome for this project is a model that can successfully identify transactions that are involved in money laundering. We would like to correctly identify at least 90% of all money laundering transactions in the dataset, and of all "money laundering" transactions identified, at least 90% should be a true positive.

After you finished your proposal, you should ask yourself the following questions:

1. Is my objective/goal being clear in the proposal? Am I able to decompose the high-level objectives into some practical problems which can be solved by my proposed solutions?

Ans: yes

2. Can my solutions help me solve the proposed problems? why? are there any requirements on the data given by my solutions? Did I introduce the data? Can I use my solutions on the proposed problems?

Ans: yes

3. Do I have a clear evaluation approach? Can I reasonably evaluate my solutions to tell that my solutions are good ones?

Ans: yes

4. Can the reader understand every detail in my proposal?

Ans: yes, possibly.

Some students just propose different techniques I taught in the class, such as classification, clustering, association rules, and put all the data mining tasks on the proposal. Well, what is your goal? are you sure these techniques can solve the problems you proposed? Please think deeper about it!! It is an examination about your understanding of the different knowledge and techniques. You should be able to figure out what techniques can be used to solve which problems, and which cannot.

5. You can choose to work on an easy project or a challenging one. Your project will be compared with others, and your grade will be affected by the degree of difficulty of your topic

Ans: Okay.