MONEY LAUNDERING DETECTION USING MACHINE LEARNING METHODS

Y B. Shashank Raj
KV SubbaReddy Engineering College, Kurnool, A.P, India
D. Eswara Reddy, V. Veeresh, S. Moinuddin Nausheer, Samson Paul
KV SubbaReddy Engineering College, Kurnool, A.P, India

Abstract: Money Laundering is the process of creating the appearance that large amounts of money obtained from serious crimes, such as drug trafficking or terrorist activity, originated from a legitimate source. Through money laundering, the launderer transforms the monetary proceeds derived from criminal activity into funds with an apparently legal source. The system that works against Money laundering is Anti-Money Laundering (AML) system. The existing system for Anti-Money Laundering accepts the bulk of data and converts it to large volumes reports that are tedious and topsy-turvy for a person to read without any help of software. To develop a structure to research in datamining, we create a taxonomy that combines research on patterns of observed fraud schemes with an appreciation of areas that benefit from a productive application of data mining. The aim of this study was to review research conducted in the field of fraud detection with an emphasis on detecting honey laundering and examine deficiencies based on data mining techniques. Which include a set of predefined rules and threshold values. In addition to this approach, data mining techniques are very convenient to detest money laundering patterns and detect unusual behavior. Therefore, unsupervised data mining technique will be more effective to detect new patterns of money laundering and can be crucial to enhance learning models based on classification methods. Of course, the development of new methods will be very useful to increase the accuracy of performance.

Keywords: Money Laundering**,** Anti-Money Laundering (AML)**,** Data Mining Techniques**,** Fraud Detection**,** Unsupervised Learning

I. INTRODUCTION

Detecting management fraud is a difficult task when using normal audit procedures. First, there is a shortage of knowledge concerning the characteristics of management fraud. Secondly, given its infrequency, most auditors lack the experience necessary to detect it. Finally, managers deliberately try to deceive auditors. For such managers, who comprehend the limitations of any audit, standard auditing procedures may prove insufficient. These limitations suggest that there is a need for additional analytical procedures for the effective detection of management fraud. It has also been noted that the increased emphasis on system assessment is at odds with the profession's position regarding fraud detection since most material frauds originate at the top levels of the organization, where controls and systems are least prevalent and efficient. Applying data mining to fraud detection as part of a routine financial audit can be challenging and, as we will explain, data mining should be used when the potential payoff is high. In general, when it comes to fraud detection for a given audit client, the audit team would make three major decisions: (1)What specific types of fraud (e.g., revenue recognition, understated liabilities, etc.) should be included in the audit plan for a particular client? (2) What sources of data (e.g., journal entries, emails, etc.) would be provided evidence of each type of fraud? (3) Which data mining technique(s) (e.g., directed or undirected techniques) would be the most effective for finding potential evidence of fraud in the selected data? Developing answers for each of these questions are significant individually, but, in combination, answering these questions is challenging.

Money Laundering

Money laundering is the process of taking cash earned from illicit activities such as drug trafficking, and making the cash appears to be earnings from a legal business activity. The money from the illicit activity is considered dirty and the process “launders” the money to make it look clean. Money laundering is the generic term used to describe the process by which criminals disguise the original ownership and control of the proceeds of criminal conduct by making such proceeds appear to have derived from a legitimate source. Illegally earned money needs laundering for the criminal organization to use it effectively. Dealing with large amounts of illegal cash is inefficient and dangerous. The criminals need a way to deposit the money in financial institutions, yet they can only do so if the money appears to come from legitimate sources. There are many ways to launder money. These methods span from the very simple to the very complex. One of the most common ways is to launder the money

through a legitimate cash-based business owned by the criminal organization.

For instance, if the organization owns a restaurant, it might inflate the daily cash receipts to funnel its illegal cash through the restaurant and into the bank. Then they can distribute the funds to the owners out of the restaurant's bank account.

Steps of Money Laundering

Money-laundering is a dynamic three-stage process that requires

a. **Placement:** This is the movement of cash from its source. On occasion, the source can be easily disguised or misrepresented. This is followed by placing it into circulation through financial institutions, casinos, shops, bureau de change and other businesses, both local and abroad. The process of placement can be carried out through many processes.

b. **Layering:** The purpose of this stage is to make it more difficult to detect and uncover a laundering activity. It is meant to make the trailing of illegal proceeds difficult for the law enforcement agencies.

c. **Integration:** This is the movement of previously laundered money into the economy mainly through the banking system, and thus such monies appear to be normal business earnings. This is dissimilar to layering, for in the integration process detection and identification of laundered funds is provided through informants.

II. LITERATURE SURVEY

Money laundering poses a significant threat to global financial systems, and its detection is critical for ensuring transparency and compliance with regulatory frameworks. Traditional rule-based systems often fall short in detecting complex laundering schemes, leading researchers to explore machine learning (ML) as a powerful alternative due to its ability to analyze large-scale data and uncover hidden patterns.

1. Traditional Methods and Their Limitations

Historically, Anti-Money Laundering (AML) systems have relied on rule-based engines and manual investigation**,** which are prone to high false positives and inefficiencies. According to Delamaire et al. (2009), these systems struggle to adapt to evolving laundering tactics. This has motivated a shift towards intelligent and adaptive solutions like ML.

2. Supervised Learning in AML

Supervised learning algorithms have been widely used to classify suspicious and non-suspicious transactions. Logistic Regression, Random Forest, Support Vector Machines (SVM), and Neural Networks have shown promise in identifying patterns indicative of laundering (Weber et al., 2018). These methods require labeled datasets, which can be challenging due to confidentiality and limited availability of confirmed laundering cases.

3. Unsupervised and Semi-Supervised Techniques

Due to the scarcity of labeled data, unsupervised methods such as clustering (e.g., K-Means, DBSCAN) and anomaly detection are often employed to detect outliers and unusual transaction behavior. Autoencoders and isolation forests are also popular in this context. Semi-supervised approaches, which leverage a small set of labeled data alongside a larger unlabeled dataset, have also gained attention for improving detection accuracy without extensive annotation requirements.

4. Deep Learning Approaches

With the rise of deep learning, models like Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs) have been explored to capture temporal and spatial patterns in financial transaction sequences. For example, RNNs can identify laundering behaviors that occur over time, such as structuring or layering (Roy et al., 2020).

5. Graph-Based Machine Learning

Since laundering often involves networks of transactions across multiple entities, graph-based ML has emerged as a robust solution. Graph Neural Networks (GNNs) and community detection algorithms can identify collusive groups and transactional cycles indicative of laundering. Research by Weber et al. (2019) using GNNs on the Elliptic Bitcoin dataset demonstrated significant improvements in suspicious activity detection.

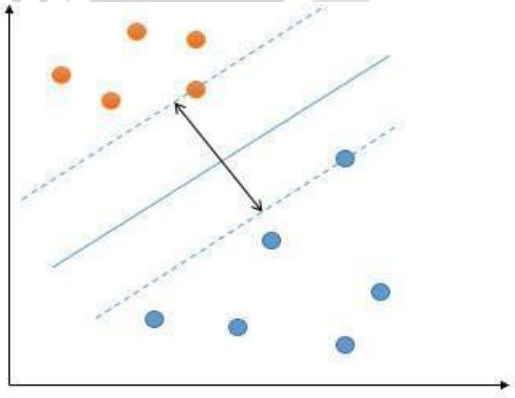
6. Challenges and Future Directions

Key challenges include data privacy**,** imbalanced datasets**,** interpretability of models**,** and real-time detection requirements. Researchers are focusing on explainable AI (XAI) to improve model transparency and federated learning to build models across institutions without compromising data privacy.

III.EXISTING SYSTEM

1. **Clustering**:

Clustering is a process that classifies the data into different groups, and the members of each group have the most similarities to each other and the members of each group have the least similarities to another group. The best performance of a clustering algorithm will be apparent when the clusters are away from each other as far as possible. In anti-money laundering, clustering is typically used for grouping transactions with bank accounts in different clusters that have the most similarities with each other. These techniques help us to detect patterns for suspicious transaction sequence or present models to identify the accounts or the riskier customers. One of the most important challenges facing the clustering of financial transactions is the size and the amount of data, for example, we are facing thousands or millions of transactions per unit time in this method. Table 1 shows different clustering methods used.



2. Rule-based methods:

We can observe two approaches in data mining, classification - prediction and clustering approach. Rule-based methods are considered classification and prediction methods. In rule-based methods, we are facing a set of rules that are expressed in the language of logic, and actually, we use a series of logical rules to classify the factors.

2. Support vector machines:

SVM is a supervised learning method, which is used for classification. Support vector machines similar to neural

networks can obtain approximations with the desired degree of accuracy for each multivariable function. So, it is very useful to model the nonlinear and complex systems and processes, including detecting activities related to financial fraud. SVM goal is to find a separator super-vector of data points belonging to two classes, with a maximal margin. From a the geometric perspective, it is the gap between the super-vector and the nearest training samples. From another perspective, the margin is defined as the amount of space or separation between the two classes, which is defined by the super -vector.

Fig : Sample of Support Vector Machine

3. Social networks:

In recent years, the Social Network theory has attracted increasing attention. Social network analysis regarding data mining is called link analysis or link mining. For the modelling of social networks, the relationship between entities will be displayed in the form of links in a graph. A social network represents relationships between social entities such as friends, professionals or writers. In the last decade, due to the increasing growth of communication technologies and Web-
based services, the growth and penetration of these networks have been widespread, and many people had the experience and interacted with social networks. Through online social networks, a huge amount of facilities will be provided for interaction and cooperation between independent individuals, regardless of the geographical distance between them. Each network is a massive database of millions of users and their activities. Unfortunately, due to the liberalization of the use of most social networks, the information contained on these networks is a good place to delinquent users. Therefore, network analysis will be instrumental to explore relationships or suspicious transactions for money laundering detection. Social networks are dynamic. New links show a new inter action between objects.In the prediction of links, a snapshot of the social network at the time "t" will be placed at our disposal, and we are asked to predict the edges that will be added to this network, in the period t to t + 1. In this case, we are looking forward to use the real attributes of the model and to expose the development that can model the evolution of a social network.

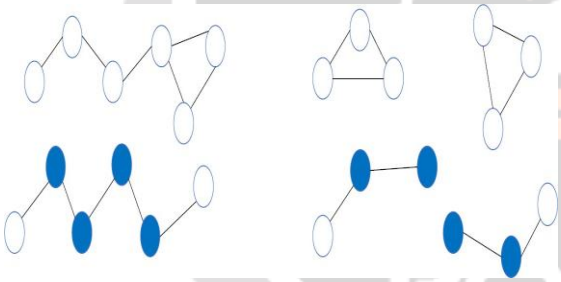


Fig : Networks with Different Relationships

IV. PROPOSED SYSTEM

Logistic Regression Model:

Logistic regression is a supervised learning classification algorithm used to predict the probability of a target variable. The nature of target or dependent variable is dichotomous, which means there would be only two possible classes.

The dependent variable is binary in nature having data coded as either 1 (stands for success/yes) or 0 (stands for failure/no). Mathematically, a logistic regression model predicts P(Y=1) as a function of X. It is one of the simplest ML algorithms that can be used for various classification problems such as spam detection, Diabetes prediction, cancer detection etc.

Types of Logistic Regression

a. **Binary or Binomial:** In such a kind of classification, a dependent variable will have only two possible types either 1 and 0

b. **Multinomial:** In such a kind of classification, dependent variable can have 3 or more possible unordered types or the types having no quantitative significance.

c. **Ordinal:** In such a kind of classification, dependent variable can have 3 or more possible ordered types or the types having a quantitative significance.

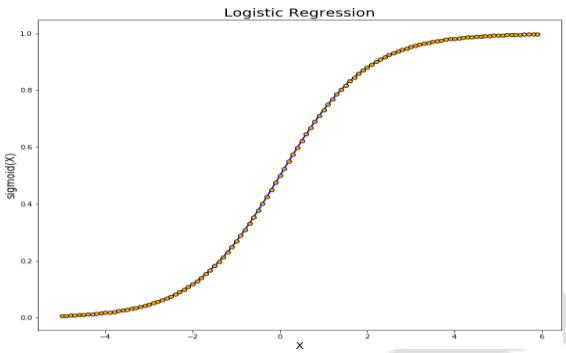


Fig : Sample of Logistic Regression Model

Decision Tree Model:

The structure of a decision tree is a tree topology similar to a flowchart. The highest node in the tree is the root node, and the leaf nodes represent categories and distribution of categories. The decision tree is a classification or prediction technique that each non- leaf test node test specifies a feature and every branch out from this node shows the result of this test. Unlike neural networks, decision trees deal with the production rules. The prediction obtained from a tree is explained in the form of a series rule in a decision tree, a while the result of prediction is only expressed in neural networks and how to achieve them is hidden in the network itself. Also, unlike neural networks, there is no requirement for the data to be necessarily numerical in the decision tree. Decision forest or random forest is a collection of several decision trees that avoids instability and excessive risk education (bias) that may occur in a single tree.

Split Creation

1. **Part one: Calculating Gini Score:** We have just discussed this part in the previous section.

2. **Part two: Splitting a dataset:** It may be defined as separating a dataset into two lists of rows having index of an attribute and a split value of that attribute. After getting the two groups - right and left, from the dataset, we can calculate the value of split by using Gini score calculated in first part. Split value will decide in which group the attribute will reside.

3. **Part three: Evaluating all Splits:** Next part after finding Gini score and splitting dataset is the evaluation of all splits.

For this purpose, first, we must check every value associated with each attribute as a candidate split. Then we need to find the best possible split by evaluating the cost of the split. The best split will be used as a node in the decision tree.

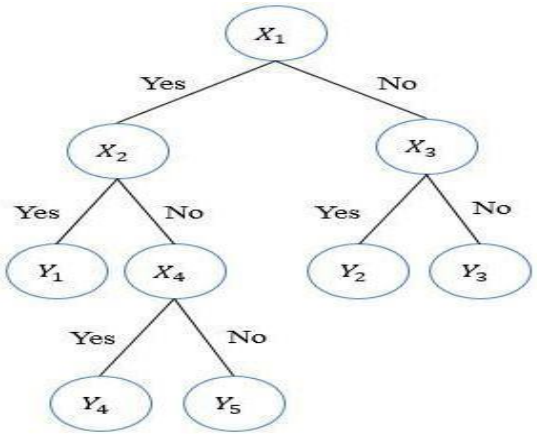


Fig 16: Sample of Decision Tree Model

Random Forest Model:

Random forest is a supervised learning algorithm which is used for both classification as well as regression. But however, it is mainly used for classification problems. As we know that a forest is made up of trees and more trees means more robust forest. Similarly, random forest algorithm creates decision trees on data samples and then gets the prediction from each of them and finally selects the best solution by means of voting. It is an ensemble method which is better than a single decision tree because it reduces the over-fitting by averaging the result.

Working of Random Forest Model:

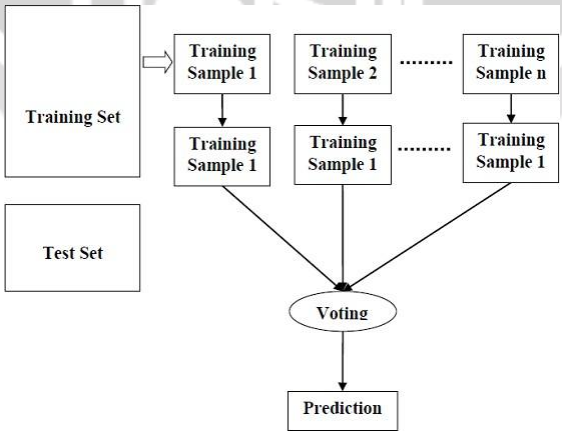
We can understand the working of Random Forest algorithm with the help of following steps –

Step-1: First, start with the selection of random samples from a given dataset.

Step-2: Next, this algorithm will construct a decision tree for every sample. Then it will get the prediction result from every decision tree.

Step-3: In this step, voting will be performed for every predicted result.

Step-4: At last, select the most voted prediction result as the final prediction result.



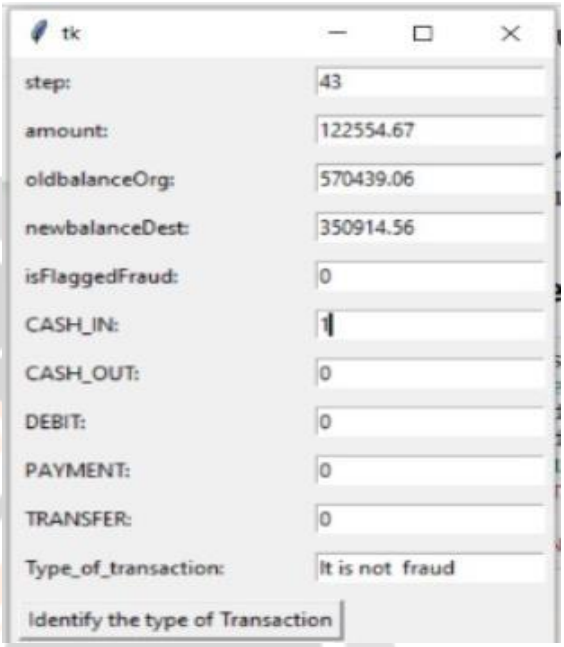
V. RESULTS

Input 1:

Step	43
Type	CASH_IN

Amount	122554.67
Old balance Org	570439.06
New balance Org	350914.56
Is Flagged Fraud	0

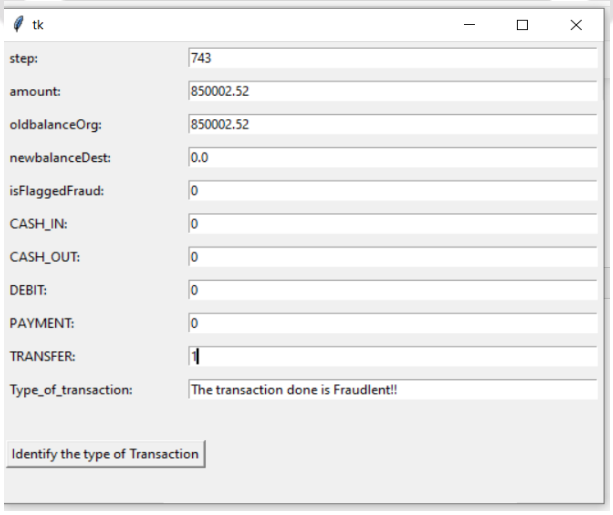
Output:



Input 2:

Step	743
Type	TRANSFER
Amount	850002.52
Old balance Org	850002.52
New balance Org	0.0
Is Flagged Fraud	0

Output:



VI. CONCLUSION

Data mining is a process to extract knowledge from existing data. It is used as a tool in banking and finance, in general, to discover useful information from the operational and historical data to enable better decision-making. It is an interdisciplinary field, the confluence of Statistics, Database technology, Information science, Machine learning, and Visualization. It involves steps that include data selection, data integration, data transformation, data mining, pattern evaluation, knowledge presentation. Banks use data mining in various application areas like marketing, fraud detection, risk management, money laundering detection and investment banking. According to what was mentioned in the previous parts, detecting activities related to money laundering is necessary and inevitable for the economy, industries, banks and financial institutions.

The future scope for this project is – This activity of identifying suspicious activity can be extended to be implemented in electronic payment applications to detect fraudulent transactions immediately.

VII. REFERENCES

1. Batista, G. E. A. P. A., & Monard, M. C. (2002). A study of K-nearest neighbour as an imputation method. *His*, **87**, 251-260.
2. Delamaire, L., Abdou, H., & Pointon, J. (2009). Credit card fraud and detection techniques: A review. *Banks and Bank Systems*, **4**(2), 57–68.
3. Roy, S., Ghosh, D., & Ganguly, N. (2020). Anti-money laundering: Experiments with deep learning on real transaction data. *Proceedings of the 2020 IEEE International Conference on Big Data (Big Data)*, 4410–4419.
4. Weber, M., Domeniconi, G., Chen, J., Weidele, D., Bellei, C., Robinson, T., & Leiserson, C. (2019). Anti-money laundering in Bitcoin: Experiments with graph convolutional networks. *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, 456–466.
5. Weber, M., Gombar, S., & Bellei, C. (2018). Scalable graph-based anomaly detection for cryptocurrency transaction networks. *arXiv preprint arXiv:1809.07476*.
6. Zhang, Y., Yin, Y., & Zhang, Z. (2021). Missing data imputation method based on K-nearest neighbors for the health examination data. *Mathematical Biosciences and Engineering*, **18**(4), 3967–3983.
7. Duman, E., & Ozcelik, M. H. (2011). Detecting credit card fraud by genetic algorithm and scatter search. *Expert Systems with Applications*, **38**(10), 13057–13063. https://doi.org/10.1016/j.eswa.2011.04.181
8. Ngai, E. W. T., Hu, Y., Wong, Y. H., Chen, Y., & Sun, X. (2011). The application of data mining techniques in financial fraud detection: A classification framework and an academic review of literature. *Decision Support Systems*, **50**(3), 559–569. https://doi.org/10.1016/j.dss.2010.08.006
9. Zanin, M., & Papo, D. (2016). Functional networks in economics: A dynamic approach for the study of fraud detection. *Scientific Reports*, **6**, 34190. https://doi.org/10.1038/srep34190
10. Akoglu, L., Tong, H., & Koutra, D. (2015). Graph-based anomaly detection and description: A survey. *Data Mining and Knowledge Discovery*, **29**(3), 626–688. https://doi.org/10.1007/s10618-014-0365-y