JURNAL EMACS
[Engineering, MAthematics and Computer Science] Vol.6 No.1 January 2024: 13-17
e-ISSN: 2686-2573
DOI: 10.21512/emacsjournal.v6i1.10744

Optimization of Fraud Detection Model with Hybrid Machine Learning and Graph Database

Aan Albone

Data Science Program, Computer Science Department, School of Computer Science,
Bina Nusantara University,
Jakarta, Indonesia 11480
aan.albone@binus.ac.id

*Correspondence: aan.albone@binus.ac.id

Abstract – *Machine learning and the graph database work well together. By concentrating on the relationships between fraudsters or fraud cases, graph databases can provide an additional layer of security, while machine learning uses statistics and data analytical tools to categorize information and identify patterns within data. In doing so, it can transcend rigid rules and scale human insights into algorithms. When combined with a graph, machine learning alone can increase the accuracy of fraud signals to 90% or higher. On its own, it can reach 70–80%. Graphs also improve machine learning's explainability.*

Keywords: *Fraud; Graph Database; Machine Learning*

I. INTRODUCTION

According to Guven, Ozlem *et al.*, (2022), fraud is one of the biggest problems with the payment system. Fraud lowers snow margins, lowers customer satisfaction, and incurs significant costs for the business. Consequently, it's critical to identify and stop fraudsters. (Guven, Ozlem & Serkan Aras, 2022).

According to Amna Sajid (2018), the time-consuming and ineffective nature of traditional manual detection methods for fraudulent activities is compounded by the impracticality of big data and machine learning. (Amna Sajid, 2018).

According to Shamil Magomedov, *et al.,* (2018), We think that machine learning (ML) models can be applied to fraud detection in order to reliably solve the anomaly detection problem. Data collection is turning into one of the main bottlenecks in machine learning, among its many other challenges. It is well known that data preparation, which includes gathering, cleaning, analyzing, visualizing, and feature engineering, takes up most of the time when running machine learning end-to-end. Despite the time-consuming nature of each step, data collection has recently become more difficult for the reasons listed below. (Roh, *et al.*, 2018).

According to Shamil Magomedov, *et al.,* (2018), additionally, machine learning technologies are not without limitations. When the initial data set is small, one of these limitations is their inability to see connections in the data. Models of machine learning operate on activities, behavior, and actions. For instance, the model may fail to recognize a relationship that would seem obvious, like a shared card between two accounts. Graph databases can be used to improve machine learning models in order to combat this. Graph databases tackle the fifth fraud prevention layer identified by Gartner: analysis of entity links [Shamil Magomedov, *et al.,* 2018).

Graph databases make it possible to see connections between discrete analysis data points rather than just the individual data points themselves. Therefore, for each fake actor that is stopped through scoring, the graph technique can identify multiple of them. Blocking phony and suspicious accounts before they've committed any fraud is possible with graph databases. Graph databases are an invaluable asset to any fraud prevention solution because of their innate ability to calculate relationships quickly.

According to Shamil Magomedov1, *et al.,* (2018), millions of connections can be made per second by the engine that manages the connections between nodes because the relationships in a graph database are given the

same consideration as the actual database records. With the use of graph databases, new information can be quickly extracted from huge, intricate databases to help identify previously unidentified relationships and interactions. (Shamil Magomedov1, *et al.,* 2018).

II. METHODS

According to Abhirami, *et al.,* (2021), the extremely uneven nature of datasets—there are fewer fraudulent transactions in the dataset than real transactions—presents one of the major obstacles to using ML in fraud detection. Machine learning techniques such as Logistic Regression, Support Vector Machine, Decision Tree, and Random Forest are employed to analyze datasets comprising thousands of transactions and distinguish the legitimate from fraudulent transactions. Prior to a fraudulent transaction occurring, fraud detection systems must identify the fraudulent and normal transaction. (Abhirami, *et al.,* 2021).

Together, graph solution and machine learning (ML) can produce even greater outcomes, shown in Figure 1.

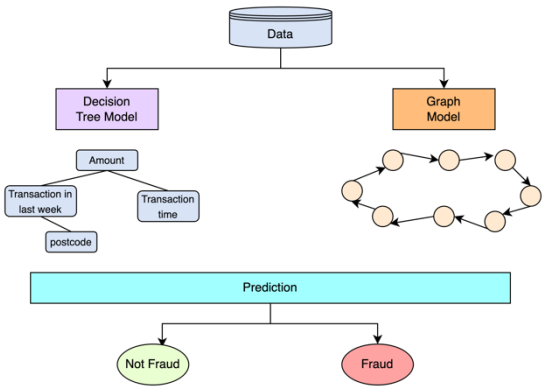


Figure 1. Hybrid ML & Graph DB Design

III. RESULTS AND DISCUSSION

According to Smt.S.Rajani, *et al.,* (2012), Fraud patterns in a rule-based fraud detection system are characterized as rules. One or more conditions can make up a rule. An alert is raised once all requirements are satisfied. (Smt.S.Rajani, *et al.,* 2012).

How many cases were identified as truly fraudulent and how many as false alarms determines the success of a fraud rule model.

Example for rules may be:

Credit-rating=C AND daily international calls duration> 2hrs => alert

Deposit= X AND normalized-daily-duration standard deviations >4 = > alert

Rules to Identify Known Fraud-based (1=Fraud; 0=No Fraud):

```
df = df.withColumn("label",
      F.when(
        (
          (df.oldbalanceOrg  56900) &
          (df.newbalanceOrig  56900) &
          (df.newbalanceOrig > 12) &
          (df.amount > 1160000)
              ), 1
        ).otherwise(0))
```

Fraud score is another rule-based fraud system. A fraud score is a number that represents the degree of risk associated with a specific transaction.

According to Michaela Baumann, *et al.,* (2021), an expert in fraud examines the automatically generated rules more closely and evaluates their quality, explanability, and meaningfulness by looking at the assigned weights. The interaction of the original rules, the combinations, and how the combinations raise or lower the suspicion of the original rules should receive special attention. (Michaela Baumann, *et al.,* 2021).

According to Andrea Dal Pozzolo, *et al.,* (2018), Another type of expert-driven model that takes the form of if-then (-else) statements is the scoring rule. On the other hand, these work with feature vectors and give each approved transaction a score; the higher the score, the higher the probability that the transaction is fraudulent. Investigators manually create the scoring rules and arbitrary assign scores to them. An example of scoring rule can be "IF previous transaction in a different continent AND less than 1 h from the previous transaction THEN fraud score = 0.95."

Sadly, scoring rules are only able to identify fraudulent strategies that investigators have already identified and that display patterns involving a small number of the feature vectors' components. Furthermore, because different experts create different rules, scoring guidelines are somewhat arbitrary. (Andrea Dal Pozzolo, *et al.,* 2018).

In this example, it is evident that the company believes that a disposable phone number is the most telling sign that someone is a scammer. They have assigned it a score of +10.

Table I. Fraud Scoring

No	Rule Items	Score
1.	Phone is disposable	10
2.	Browser version age is greater or equal to 5 years	5
3.	Customer is using harmful IP address	2
4.	Customer is using Private Email Relay Service	2

This sample demonstrates how this business has determined that a disposable phone number is the most convincing indicator that someone is a fraudster. They have assigned it a score of +10.

By obtaining a decision boundary in the feature space defined by input transactions, the machine learning model for fraud detection seeks to distinguish between fraudulent

and non-fraudulent transactions. Several fields of study use machine learning techniques to derive computational intelligence. It makes it possible to generalize particular examples that are useful for dataset modeling, prediction, and classification.

One such popular machine learning method that works well for regression or classification is the decision tree. Algorithms that divide a dataset into numerous branching segments based on decision rules produce decision trees. The relationship between the input attributes and the outputs is used to determine these decision rules, shown in Figure 2 & Figure 3.

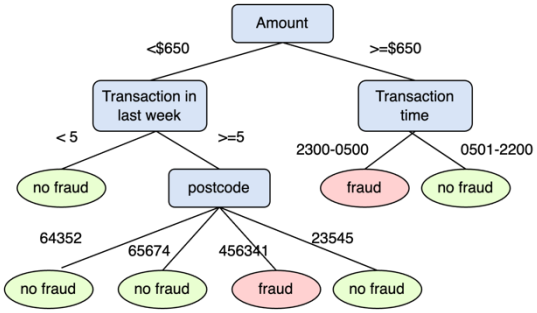
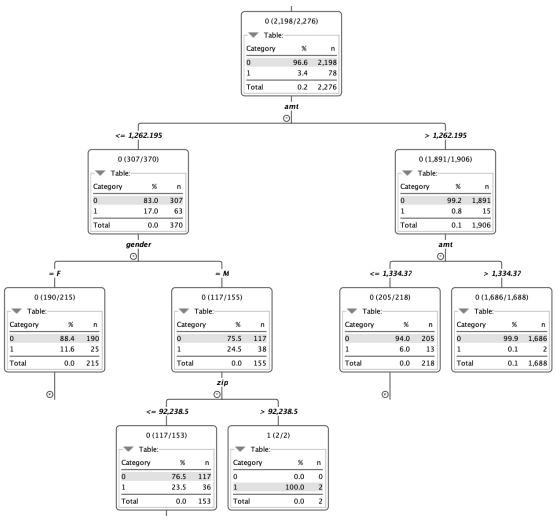


Figure 2. Fraud Decision tree



0: no fraud
1: fraud

Figure 3. Fraud Decision tree (Case Study)

A black box approach to fraud detection is frequently unworkable. Prior to anything else, the domain experts must be able to explain the reason behind a transaction's fraudulent identification. The evidence must then be presented in court if action is to be taken. For this use case, the decision tree model is a great place to start because it is simple to understand, shown in Figure 4.

#	Row...	amt *Number (double)*	gender *String*	zip *Number (integer)*	is_fraud *String*
1	0	4.97	F	28654	0
2	1	107.23	F	99160	0
3	2	220.11	M	83252	0
4	3	45	M	59632	0
5	4	41.96	M	24433	0
6	5	94.63	F	18917	0
7	6	44.54	F	67851	0
8	7	71.65	M	22824	0
9	8	4.27	F	15665	0
10	9	198.39	F	37040	0

Figure 4. Fraud Data Training and Testing

According to Shivaram Kalyanakrishnan, *et al.,* (2014), a decision tree divides the universe of keys into partitions, with each partition's cell representing a predicted response. As a result, the prediction linked to the cell that the key belongs to is the expected response for a (test) key. The participation is taught and shown in a hierarchical manner as a tree, with each cell denoting a leaf. (Shivaram Kalyanakrishnan, *et al.,* 2014).

According to Richard Henderson, *et al.,* (2020), Making connections is the key to detecting fraud. By examining the connections between individuals, phones, and bank accounts, among other things, graph techniques can be used to combat financial fraud. This helps banks identify suspicious activity in a sea of data and provides them with the means to explain what's happening. (Richard Henderson, *et al.,* 2020)

Combating financial fraud is difficult. It specifically entails being able to identify possible fraud cases in sizable datasets and be able to discriminate between legitimate cases and false positives, or cases that appear suspicious but aren't.

According to Amna Sajid, *et al.,* (2018), Conventional fraud detection systems concentrate on customer activity-related thresholds. For instance, making numerous purchases of the same item or making a large number of transactions with one credit card or person are suspicious activities. Furthermore, machine learning, a branch of artificial intelligence, combines statistical modeling with a variety of computer algorithms to enable tasks to be completed by computers without the need for hard coding. (Amna Sajid, *et al.,* 2018).

Graph analysis, which emphasizes the connections between fraud cases or fraudsters, can provide an additional degree of security.

For example, in online retail operation. It includes:

- Order details: product, amount, order-id, date.
- Personal details: first name, last name.
- Contact info: phone, email.
- Payment: credit card.
- Shipping: address, zip, city, country.
- Tracking: IP address.

In order to examine the relationships within our data, we kept it in a graph database. The graph approach uses nodes and edges to represent data, shown in Figure 5.

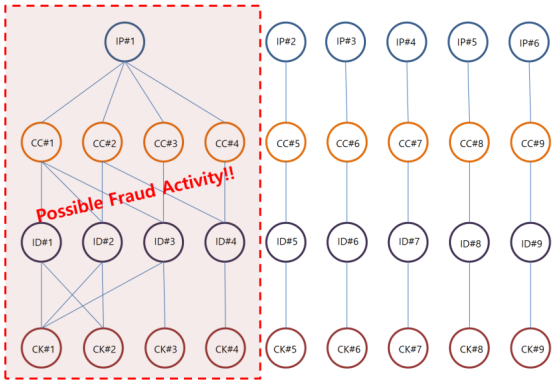


Figure 5. Graph Fraud Activities

Think about how much information banks have about their clients. By looking at related elements, we can analyze customer behavior and determine what constitutes good and bad customer behavior. After that, we can begin to categorize our clientele into two groups: good and bad.

Case 1 Multi Transaction

When given a transaction, it locates the user networks associated with the sender and the recipient. After that, all of the transactions between the two networks are found.

This query aids in the identification and visualization of transaction patterns in cases where the transaction is a component of a money laundering scheme by data analysts, shown in Figure 6.

1. Starting from a given transaction, find its sender and receiver.
2. Starting from the sender, traverse 4 steps via Device_ Token and
3. Payment_Instrument edges to find connected Users.
4. Starting from the receiver, traverse 4 steps via Device_ Token and
5. Payment_Instrument edges to find connected Users.
6. Detect transactions between the sender and receiver networks.

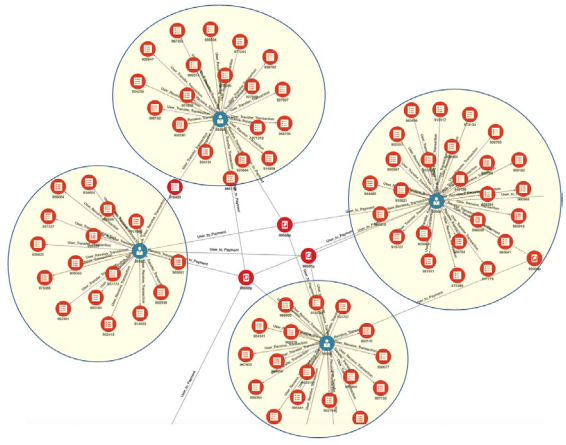


Figure 6. Case Multi Transaction

Case 2 Circle Detection

Find every transaction path that started with the input user and ended with the user, given a user ID. A circular flow like this could be a sign of money laundering.

We start at a specific transaction from account A to account B and watch transactions that ultimately link back to account A in order to identify such a circular flow. We check with a similar amount of money at a later time for transactions from account B. In order to determine whether we return to account A, we keep tracking the transactions from account to account. A graph would be created from a series of transactions demonstrating this pattern of money laundering, shown in Figure 7 & Figure 8.

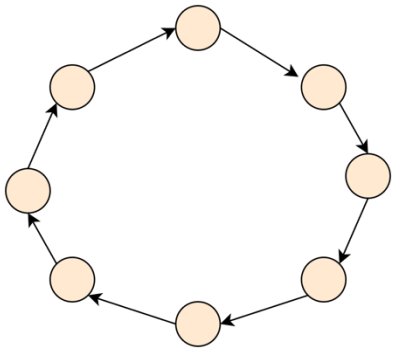


Figure 7. Example of Circle Detection

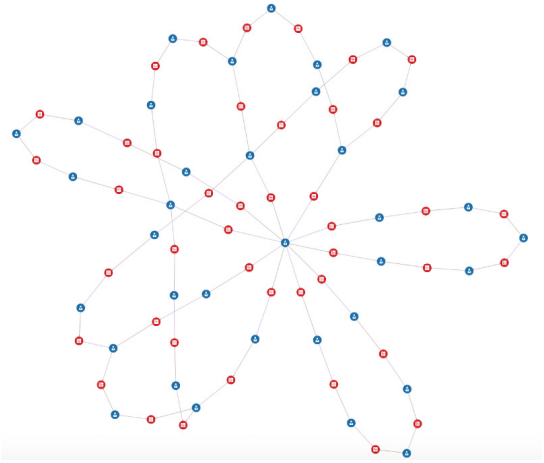


Figure 8. Case 2 Circle Detection

Case 3 Same Receiver Sender

Find all occurrences in which the sender is linked to the recipient through Device_Token or Payment_Instrument in 4 steps, given an input transaction. Self-transactions of this kind are more prone to be fraudulent, shown in Figure 9.

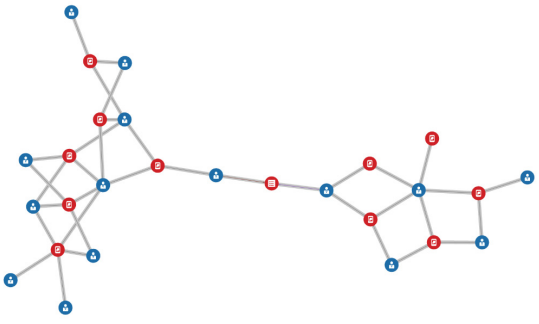


Figure 9. Case Same Receiver Sender

Ascertain the amount of money that has been transferred out of a user's connected users during a specified time frame.

1. From a user "sender", find all other users connected by Device_Token or Pay-ment_Instrument within k steps.
2. From these connected users, find transfer (send) transactions between "start_date" and "end_date".
3. Calculate total transferred money of the transactions in step 2)

Most of the time, the relationships between these identifiers should be one-to-one. Naturally, some variations take into account things like families using a single credit card number, people using multiple computers, and shared machines.

However, fraud should be taken seriously as soon as there is a significant correlation between these variables and a reasonable number. There is more reason to be concerned the more connections there are between identifiers. Big, closely-knit graphs are excellent markers of fraud activity. We can therefore detect, watch for potential fraud activities, and stop fraudsters' nefarious attempts to defraud people and organizations in real time by incorporating checks and event triggers into the pathways of transactions.

IV. CONCLUSION

Machine learning techniques can be used to accomplish classification, grouping, regression, and other tasks by ingesting graph data. Graphs also improve machine learning's explainability. Graphs are used to identify intricate patterns and provide analysis in a visual context.

Combined, graph and machine learning yield faster insights and higher analytical accuracy:

1. When combined with a graph, machine learning alone can increase the accuracy of fraud signals to 90% or higher. On its own, it can reach 70–80%.
2. Machine learning and graph databases work well together. By concentrating on the relationships between fraudsters or fraud cases, graph databases can provide an additional layer of security, while machine learning uses statistics and data analytical tools to categorize information and identify patterns within data. In doing so, it can transcend rigid rules and scale human insights into algorithms.

REFERENCES

Amna Sajid. (2023). Fraud Detection of Credit Cards Using Supervised Machine Learning Techniques, *Journal of Emerging Science and Technologies (PJEST).*

Andrea Dal Pozzolo, Giacomo Boracchi, Olivier Caelen, Cesare Alippi. (2018). Credit Card Fraud Detection: A Realistic Modeling and a Novel Learning Stra-tegy. *IEEE Transactio on Neural Network and Learning System.*

Guven, Ozlem & Serkan Aras (2022). Fraud Detection by Machine Learning Algorithms: A Case From A Mobile Payment. *International Journal of Mana-gement Economics and Business*, Vol. 18, No. 3.

K. Abhirami, A. K. Pani, M. Manohar, and P. Kumar, (2021). An Approach for De-tecting Frauds in E-Commerce Transactions using Machine Learning Tech-niques, in 2021 2nd. *International Conference on Smart Electronics and Com-muniation* (ICOSEC), IEEE, pp. 826-831.

Michaela Baumann. (2021). Improving a Rule-based Fraud Detection System with Classification Based on Association Rule Mining *INFORMATIK*.

Roh, Y., Heo, G., Whang, S.E. (2019). A Survey on Data Collection for Machine Learning: A Big Data - AI Integration perspective. *IEEE Transactions on Knowledge and Data Engineering* (Volume: 33)

Richard Henderson. (2020). Using graph databases to detect financial fraud. *Computer Fraud & Security,* Volume 2020, Issue 7, July 2020, Pages 6-10.

Shamil Magomedov1, Sergei Pavelyev, Irina Ivanova. (2018). Anomaly Detection with Machine Learning and Graph *Databases in Fraud Management. International Journal of Advanced Computer Science and Application* (IJACSA), Vol. 9, No. 11.

Shivaram Kalyanakrishnan, Olivier Caelen, Cesare Alippi. (2015). On Building Decis-ion Trees from Large-scale Datain Applications of On-line Adverti-sing, Proceedings of the 23rd *.ACM International Conference on Information and Knowledge Management*, pp. 669--678, ACM.

Smt.S.Rajani & Padmavathamma. (2012). A Model for Rule Based Fraud Detection in Telecommunications. *International Journal of Engineering Research & Technology.*