Proceedings of the Fifth International Conference on Computing Methodologies and Communication (ICCMC 2021)
IEEE Xplore Part Number: CFP21K25-ART

2021 5th International Conference on Computing Methodologies and Communication (ICCMC) | 978-1-6654-0360-3/20/$31.00 ©2021 IEEE | DOI: 10.1109/ICCMC51019.2021.9418470

Comparative Study of Machine Learning Algorithms for Fraud Detection in Blockchain

Madhuparna Bhowmik
Information Technology Department
NITK, Surathkal
Mangalore, India
madhuparnabhowmik04@gmail.com

Tulasi Sai Siri Chandana
Information Technology Department
NITK, Surathkal
Mangalore, India
tsschand@gmail.com

Dr. Bhawana Rudra
Information Technology Department)
NITK, Surathkal
Mangalore, India
bhawanarudra7@gmail.com

***Abstract*—Fraudulent transactions have a huge impact on the economy and trust of a blockchain network. Consensus algorithms like proof of work or proof of stake can verify the validity of the transaction but not the nature of the users involved in the transactions or those who verify the transactions. This makes a blockchain network still vulnerable to fraudulent activities. One of the ways to eliminate fraud is by using machine learning techniques. Machine learning can be of supervised or unsupervised nature. In this paper, we use various supervised machine learning techniques to check for fraudulent and legitimate transactions. We also provide an extensive comparative study of various supervised machine learning techniques like decision trees, Naive Bayes, logistic regression, multilayer perceptron, and so on for the above task.**

***Index Terms*—fraudulent transactions, blockchain, machine learning**

I. Introduction

The problem of detecting fraudulent transactions is being studied for a long time. Fraudulent transactions are harmful to the economy and discourage people from investing in bitcoins or even trusting other blockchain-based solutions. Fraudulent transactions are usually suspicious either in terms of participants involved in the transaction or the nature of the transaction. Members of a blockchain network want to detect Fraudulent transactions as soon as possible to prevent them from harming the blockchain network's community and integrity. Many Machine Learning techniques have been proposed to deal with this problem, some results appear to be quite promising [4], but there is no obvious superior method. This paper compares the performance of various supervised machine learning models like SVM, Decision Tree, Naive Bayes, Logistic Regression, and few deep learning models in detecting fraudulent transactions in a blockchain network. Such comparative study will help decide the best algorithm based on accuracy and computational speed trade-off. Our goal is to see which users and transactions have the highest probability of being involved in fraudulent transactions.

II. Literature Survey

Yuanfeng Cai et al. [1] discussed the objective and subjective frauds. They conclude that blockchain effectively detects objective fraud but not subjective fraud and thus uses Machine Learning to mitigate the weakness.

Hissu Hyvärinen et al. [2] worked on illegal applications for tax returns and caused huge loss to the government. They mainly focus on international transactions, including double tax at the country where the transaction is generated and the country where the transaction is received, insight into a different type of fraudulent activity related to tax returns.

Jennifer J. Xu [3] discussed the types of fraudulent activities that blockchain can detect and the ones that blockchain is still vulnerable to. This paved a path towards ideas about what problems a Machine learning part needs to consider. She specifies that attacks like Identity theft and system hacking are still possible and challenging to detect using blockchain as it just uses some predetermined rules.

Michał Ostapowicz et al. [4] used Supervised Machine Learning methods to detect fraudulent activities. They focused on the fact that malicious actors can steal money by applying well-known malware software or fake emails. Therefore they used the capabilities of Random Forests, Support Vector Machines, and XGBoost classifiers to identify such accounts based on a dataset of more than 300 thousand accounts.

Blaž Podgorelec et al. [5] devised a method using Machine Learning for the automated signing of transactions in the blockchain. Hence, it also uses a personalized identification of anomalous transactions.

Steven Farrugia et al. [6] detected illicit accounts in the Ethereum Blockchain based on their transaction history. They found out that 'Time difference between first and last (Mins)', 'Total Ether balance' and 'Min value received' are the three major contributing factors for detecting illicit accounts.

Thai T. Pham et al. [7] focused on detecting an anomaly, particularly in bitcoin transaction networks. They used k-means clustering, Mahalanobis distance, and unsupervised support vector machines to detect suspicious users and transactions. They used the dataset consisting of two graphs, one for users as nodes and another one as transactions as nodes.

Further, Patrick Monamo et al. [8] also used unsupervised learning algorithms for detecting fraud in bitcoin networks. They specifically focused on the use of trimmed k-means for fraud-detection in a multivariate setup.

Fa-Bin Shi et al. [9] used a different method and focused on using financial index or normalized logarithmic price return

to detect anomalies. They suggested that abnormal ask and bid prize potentially means prize manipulation or money laundering.

Li Ji et al. [10] present an exhaustive survey of data-mining techniques, including the study of deep learning techniques used for anomaly detection. They also summarized the different universal and specific detection methods. They also talk about the disadvantages and advantages of the different methods used and provide information about how this field's future may look.

Bartoletti et al. [14] also used data-mining techniques for detecting Ponzi schemes in Bitcoins. Ponzi schemes are fraudulent activities where funds from a recent investor are paid to earlier investors. They made use of features of real-life Ponzi schemes for training machine learning classifiers.

Recently, Patel et al. [11] used a sentiment analysis framework for fraud schemes detection in cryptocurrency. The decentralized framework proposed by them, KaRuNa, includes three phases of trust modeling. They employed the use of Machine Learning for measuring social trends, cryptocurrency prizes, etc. They used LSTM (Long-Short Term Memory) classifier for this purpose. Hence, many of the recent techniques used for detecting fraud are making use of machine learning for its robust nature and accuracy.

Christian Brenig et al. [12] presents an economic analysis of money laundering using cryptocurrencies. They discuss the structure of money laundering and also propose defensive techniques. This paper further motivates our work on finding an effective way to find such fraudulent activities in cryptocurrencies and blockchain in general.

Though in the presence of label scarcity, such tasks may be difficult to solve using traditional machine learning approaches, and thus Lorenz et al. [13] proposed a method to detect money laundering when not enough labeled data is present. Their solution employs a real-life situation as, in many cases, labels are not present in abundance. However, their active learning solutions worked pretty well in detecting money laundering, even with very less labeled data.

We studied the different types of fraudulent activities in Banking Systems, including people external to the system and employees within the Bank involved in fraudulent activities.

III. Methodology

The workflow for detecting fraudulent activity is summarised in Figure 1. Essentially, after the Blockchain network has approved a transaction after all basic checks, our proposed system kicks in and does additional checks to detect if the transaction can be fraudulent. This approach makes sure that there is no extra overhead of even checking the transactions that the Blockchain network itself can easily invalidate.

The work done can be divided mainly into three phases:
1. Preprocessing phase
2. Building and training various models
3. Performance evaluation of all the models.

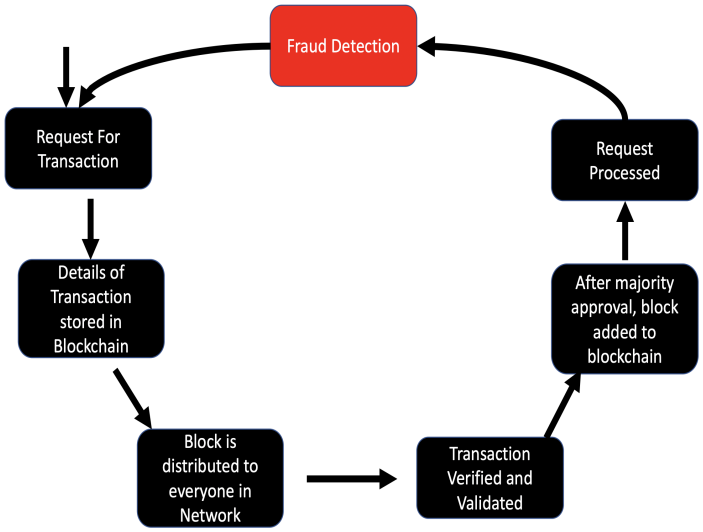

Fig. 1. Workflow of applying check for Fraud Detection

Phase I- Preprocessing

We preprocess using node-embedding in the network using the node2vec algorithm. Then, we read and convert the shorter version of concatenated rating dataset into a dataframe. Then, we create a function for the perception store features. This function extracts the features of a node using the "source" and "target" columns of the dataset. These features are then stored in a CSV file. We then run the node2vec algorithm in python and create a dictionary of nodes and corresponding embeddings. We also create a network edge list file and then reduce embeddings dimensionality for 2D projections. This dimensionality reduction can be obtained using algorithms like t-SNE.
We then normalize the features extracted from the node2vec algorithm and create a file that contains the normalized values. We assign a score of 1 if the transaction is rated badly (fraud) and 0 otherwise. We then calculate the mean and standard deviation of the node features and save it to a CSV file. We then divide all our obtained data into train and test sets.

Phase II- Building various models, training and testing them.

We divide our data into train(0.8) and test(0.2) data. We then check the ratio of fraudulent and honest transactions in our train and test sets. We use machine following machine learning and deep learning models to predict if a transaction is fraudulent:
1. Logistic Regression: This is a simple linear classifier. Logistic regression works well for binary classification problems.

2. Multilayer Perceptron: Multilayer perceptron helps in separation data that cannot be classified using a linear classifier by introducing non linearity.
3. Naive Bayes: This model uses the Bayes theorem to calculate the probability of a transaction being fraudulent.
4. Adaboost: This is an ensemble learning method to boost the performance of binary classifiers.
5. Decision Tree: This classifier has a sequence of conditions and questions on data based on various features.
6. SVM: It uses a kernel method to transform the data in the dataset, and based on these transitions, it finds a boundary between all possible outputs.
7. Random Forest Classifier: This classifier fits a number of decision trees on small batches of the dataset.
8. Neural Network: This model consists of six dense layers and four hidden layers. Relu and sigmoid were used as activation functions.

Phase III-Evaluation of models on test set

We evaluate all our classification models using bootstrap sampling. In machine learning, bootstrap sampling involves drawing sample data with replacement from the dataset to estimate a parameter. So we first choose the number of bootstrap samples. Then, we choose the sample size. Then, for each bootstrap sample, we draw a sample with chosen bootstrap size (with replacement) and test the sample's data. For this purpose, we use the accuracy metric, which is a standard metric used in machine learning problems. We then take the mean of all accuracies obtained in this fashion to evaluate the skill of our model.

IV. Results

We applied eight different supervised learning algorithms to the dataset. The dataset contains information about trust on different nodes or ratings given to them. This information is useful in detecting if a certain node's transaction can be fraudulent or not. The following table summarizes the accuracy obtained in each case.

Sl. No.	Algorithm	Accuracy
1.	Logistic regression	0.96
2.	Multi-Layer Perceptron (MLP)	0.91
3.	Naive Bayes	0.89
4.	Ada Boost	0.97
5.	Decision Tree	0.96
6.	Support Vector Machine (SVM)	0.97
7.	Random Forest Classifier	0.97
8.	Deep Neural Network	0.94

We observed that using Ada Boost, SVM, and Random Forest classifier gave the best results among the seven different algorithms. Also, since these algorithms already provide an accuracy of 97% we would like to build a fraud detector that will use the scores and decisions from the three algorithms together to decide if a transaction is fraudulent or not finally.

V. Conclusion

A method has been proposed for the detection of fraudulent transactions in a blockchain network using machine learning. In this method, various supervised learning approaches like support vector machines, decision trees, logistic regression, and dense neural networks were analyzed. A thorough comparative analysis of all the approaches is performed through accuracy. This work can be extended for the comparative study of unsupervised algorithms like clustering. In the future, we also plan to do an exhaustive study on fraudulent activities in a private blockchain.

References

[1] Cai, Y., Zhu, D. Fraud detections for online businesses: a perspective from blockchain technology. Financ Innov 2, 20 (2016). https://doi.org/10.1186/s40854-016-0039-4

[2] Hyvärinen, H., Risius, M. & Friis, G. A Blockchain-Based Approach Towards Overcoming Financial Fraud in Public Sector Services. Bus Inf Syst Eng 59, 441–456 (2017). https://doi.org/10.1007/s12599-017-0502-4

[3] Xu, J.J. Are blockchains immune to all malicious attacks?. Finance Innov 2, 25 (2016). https://doi.org/10.1186/s40854-016-0046-5

[4] Ostapowicz M., Żbikowski K. (2019) Detecting Fraudulent Accounts on Blockchain: A Supervised Approach. In: Cheng R., Mamoulis N., Sun Y., Huang X. (eds) Web Information Systems Engineering – WISE 2019. WISE 2020. Lecture Notes in Computer Science, vol 11881. Springer, Cham. https://doi.org/10.1007/978-3-030-34223-4_2

[5] Podgorelec, B., Turkanović, M. and Karakatič, S., 2020. A Machine Learning-Based Method for Automated Blockchain Transaction Signing Including Personalized Anomaly Detection. Sensors, 20(1), p.147.

[6] Farrugia S, Ellul J, Azzopardi G. Detection of illicit accounts over the Ethereum blockchain. Expert Systems with Applications. 2020 Jul 15;150:113318.

[7] Pham, Thai, and Steven Lee. "Anomaly detection in bitcoin network using unsupervised learning methods." arXiv preprint arXiv:1611.03941 (2016).

[8] Monamo, Patrick, Vukosi Marivate, and Bheki Twala. "Unsupervised learning for robust Bitcoin fraud detection." 2016 Information Security for South Africa (ISSA). IEEE, 2016.

[9] Shi, Fa-Bin, et al. "Anomaly detection in Bitcoin market via price return analysis." PloS one 14.6 (2019): e0218341.

[10] Li, Ji, et al. "A Survey on Blockchain Anomaly Detection Using Data Mining Techniques." International Conference on Blockchain and Trustworthy Systems. Springer, Singapore, 2019.

[11] P. N. Sureshbhai, P. Bhattacharya and S. Tanwar, "KaRuNa: A Blockchain-Based Sentiment Analysis Framework for Fraud Cryptocurrency Schemes," 2020 IEEE International Conference on Communications Workshops (ICC Workshops), Dublin, Ireland, 2020, pp. 1-6, doi: 10.1109/ICCWorkshops49005.2020.9145151.

[12] Brenig, Christian, and Günter Müller. "Economic analysis of cryptocurrency backed money laundering." (2015).

[13] Lorenz, Joana, et al. "Machine learning methods to detect money laundering in the Bitcoin blockchain in the presence of label scarcity." arXiv preprint arXiv:2005.14635 (2020).

[14] Bartoletti, Massimo, Barbara Pes, and Sergio Serusi. "Data mining for detecting Bitcoin Ponzi schemes." 2018 Crypto Valley Conference on Blockchain Technology (CVCBT). IEEE, 2018.