International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395-0056
Volume: 09 Issue: 07 | July 2022 www.irjet.net p-ISSN: 2395-0072

Analysis on Fraud Detection Mechanisms Using Machine Learning Techniques

Anusree Sanalkumar

Computer Science and Engineering Department, Toms College Of Engineering, Kottayam, Kerala, India

---

Abstract - *A blockchain is basically a decentralized digital ledger of transactions that is duplicated and distributed across the complex network systems. Each transactional data is stored as a block in the network. Every blockchain transaction is created across a peer to peer network and is authenticated by the digital signature of the owner. Hence, the information that contained in the ledger is highly secure. Even though they are secure, fraudulent activities still takes place. Many machine learning techniques are used for the reduction of fraudulent activities and its detection. In this paper various machine learning techniques are studied and combined to produce a more efficient fraud detection mechanism by Ensembling the most prominent machine learning algorithms. Algorithms like Random Forest Classifier and Adaboost is ensembled using the Stacking method and that accuracy is measured and compared with their individual accuracy results.*

Key Words: Blockchain, fraud detection, machine learning, Ensembling, online transaction

1. INTRODUCTION

Machines are a great asset at processing and validating large datasets. They are able to detect and recognize thousands of patterns on a user's browsing in networks and their transactions. We can predict fraud in a large volume of transactions by applying several computing technologies to raw data. This is one of the reasons why we use machine learning algorithms for preventing fraud for our clients.

Fraud detection process using machine learning starts with gathering and preprocessing the data. Then various machine learning models is fed with training sets to predict the probability of fraud. It is a very useful technology which allows us to find patterns of an anomaly in everyday transactions. Machine learning technologies has become a superior mechanism in finding and preventing these fraudulent activities in the network system

Different prominent machine learning algorithms like SVM, Adaboost, Random forest classifier etc. are ensemble together to provide a better fraud detection mechanism and this is analyzed with the existing system to acquire the result of finding which method is better for the detection mechanism in machine learning techniques.

1.1 OBJECTIVE

Different types of machine learning techniques have been used to solve the problem of online fraudulent activities. There are supervised and unsupervised machine learning techniques used in different methods to detect fraudulent activities in the online transaction system. The blockchain technology also uses this machine learning techniques as it is the most reliable and fast method in the present time. Different prominent machine learning algorithms are ensemble together to provide a better fraud detection mechanism and this is analyzed with the existing system to acquire the result of finding which method is better for the detection mechanism in machine learning techniques.

2. LITERATURE SURVEY

Madhuparna Bhowmik et al, [1] A method has been proposed for the detection of fraudulent transactions in a blockchain network using machine learning.they evaluate all our classification models using bootstrap sampling and processed the data using node2vec algorithm.

Yuanfeng Cai et al. [2] 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.

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.

Blaˇz 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.

Thai T. Pham et al. [6] 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.

3. METHEDOLOGY

The transactional history of user around a period of time is taken for the model creation in this approach. The obtained model is first pre-processed for the data cleaning and data reduction for omitting the missing and infinite values. The cleaned data is then normalized before it is processed into the algorithm. The obtained dataset gets split into training and validation or test set using k-means validation method. Various machine learning algorithm are modeled which showed high accuracy in the existing system such as Random Forest Classifier and AdaBoost classifier. The predicted values obtained from these models are together made into a new dataset which is used as the training model for the ensembling technique. The new dataset being the training set and previous test being the validation set the final predictions are produced and analyzed comparing with the previous models.

The work done can be divided into many phases:

1. Pre-processing phase
2. Building and training model
3. Performance evaluation of each model
4. Ensembling of models
5. Final analysis

3.1 Pre-Processing Phase

The pre-processing phase is the most time consuming and the first phase of the analysis. The dataset used for the proposed approach is the transaction data of an Ethereum blockchain network. The ski-learn library is an important library used for the pre-processing of data.

The obtained dataset is cleaned by eliminating the null and infinite values from the records. The cleaned and reduced data is then divided into two sets:

- Training Set
- Test/Validation Set

The splitting of dataset is processed by k-fold validation method. In K-fold cross-validation approach the entire input dataset into K groups of samples dataset of equal sizes and these samples are called folds. For each learning set, the prediction function uses k-1 folds as the training set to train the model and the rest of the folds are used for the test set. The k is a constant which should be inputted by the user

3.2 Building and Training Model

The Supervised Machine Learning algorithm can be broadly classified into Regression and Classification Algorithms. The method used in this process will be classification mechanism. The algorithms considered for modeling are:

1. Random Forest Classifier
2. Adaboost Classifier

Random Forest Classifier: Random Forest is a classifier that contains a number of decision trees on various subsets of the given dataset and takes the average to improve the predictive accuracy of that dataset. The greater number of trees in the forest leads to higher accuracy and prevents the problem of overfitting. Random Forest works in two-phase first is to create the random forest by combining N decision tree, and second is to make predictions for each tree created in the first phase

AdaBoost Classifier: The AdaBoost algorithm involves using very short (one-level) decision trees as weak learners that are added sequentially to the ensemble. What this algorithm does is that it builds a model and gives equal weights to all the data points. It then assigns higher weights to points that are wrongly classified. Now all the points which have higher weights are given more importance in the next model. It will keep training models until and unless a lower error is received.

3.3 Performance Evaluation

The total computational time required for the completion of training and modeling these algorithms took a total time period of 33minutes and 15seconds. The accuracy of the machine learning algorithms are determined using the accuracy score method called “cross_val_score()”.The result of the analysis determines that the accuracy obtained by each algorithm modeling is:

- Random forest classifier: 0.9999966960155324
- Adaboost classifier: 0.9999955946878819

3.4 Ensembling of Models

After the individual performances of these models are analyzed, the predicted model output is combined to form a new training set which will be used for the ensembling model. The ensembling model is developed by training the predicted values of the machine learning model using the method of stacking.

Stacking model is designed in such a way that it consists of two or more base/learner's models and a meta-model that combines the predictions of the base models. These base models are called level 0 models, and the meta-model is known as the level 1 model

Base models: These models are also referred to as level-0 models. These models use training data and provide compiled predictions (level-0) as an output. The models used in the base model are Random forest classifier and Ada boost classifier

Meta Model: The architecture of the stacking model consists of one meta-model, which helps to best combine the predictions of the base models. The Meta model used in stacking process is called linear regression.

The final prediction is developed from model and its accuracy is calculated to determine its efficiency or in this case accuracy score.

3.5 Final Analysis

After the individual performances of these models are analyzed, the predicted model output is combined to form a new training set which will be used for the ensembling model. The ensembling model is developed by training the predicted values of the machine learning model using the method of stacking. The final prediction is developed from model and its accuracy is calculated to determine its efficiency or in this case accuracy score.

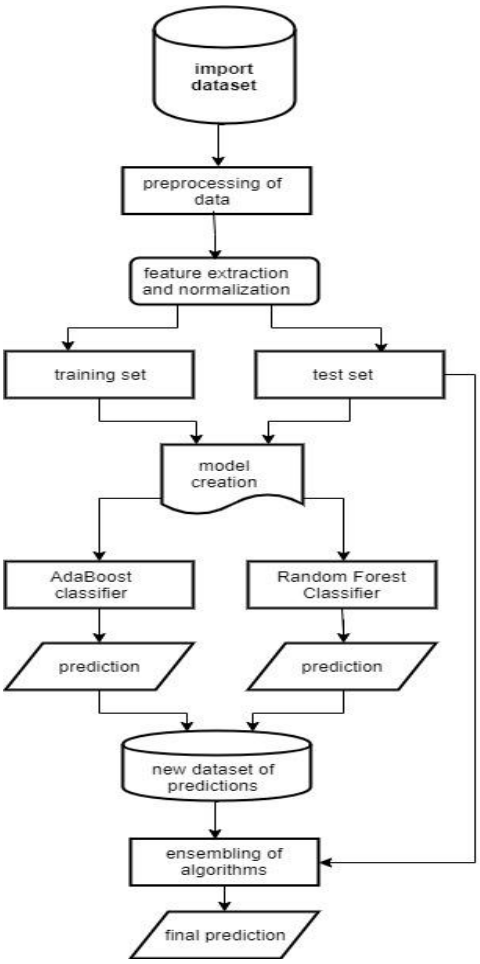

Fig 3.1 workflow of ensembling algorithms

4. FINAL RESULT

The total accuracy score of all the algorithms are analyzed and it can be seen that the accuracy obtained by individual algorithms are more in accuracy than they were while used in the existing system which were to be 97%. We can also observe that the accuracy score obtained by the ensembling is same and equal to the accuracy score of Random Forest Classifier.

The result of the analysis determines that the accuracy obtained by ensembling of these algorithms by stacking method is 0.9999966960155324

The overall result analysis of the proposed system can be shown in a tabular context:

Sl.no	Algorithm	Accuracy score
1	Random Forest Classifier	0.9999966960155324
2	Adaboost Classifier	0.9999955946878819
3	Ensemble Model (Stacking)	0.9999966960155324

Fig 4.1 Accuracy Score Table

5. CONCLUSION

Analysis of various machine learning algorithm has been observed in the proposed design method. First individual analysis of each machine learning algorithm is determined using k-fold validation method. The accuracy score obtained by the k-fold cross validation mechanism on AdaBoost classifier and Random Forest Classifier provides an accuracy of approximately 99% for Random Forest classifier having slightly higher than the AdaBoost but the overall accuracy score of these algorithm modeled by k-folk cross validation is higher than node2vec algorithm. The ensembling of these algorithms is also performed using the k-fold cross validation method by stacking. We can observe that the accuracy score obtained by Random Forest Classifier and the stacking of these algorithms is similar. The stacking process acts as an extra measure for more efficient accuracy of fraud pattern detection. The analysis can be concluded by stating that the use of k-fold cross validation mechanism for fraud pattern detection is more prominent and efficient compared to node2vec algorithm[1]

REFERENCES

[1] Madhuparna Bhowmik, Tulasi Sai Siri Chandana and Dr. Bhawana Rudra "Comparative Study of Machine Learning Algorithms for Fraud Detection in Blockchain" 2021 5th international conference on computing methodologies and communication (ICCMC) https://doi.org/10.1109/ICCMC51019.2021.9418470

[2] 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

[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., ˙ Zbikowski 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., Turkanovi´c, M. and Karakati˘c, S., 2020. A Machine Learning-Based Method for Automated Blockchain Transaction Signing Including Personalized Anomaly Detection. Sensors, 20(1), p.147.

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