International Journal of Recent Development in Engineering and Technology
Website: www.ijrdet.com (ISSN 2347-6435 (Online) Volume 13, Issue 3, March 2024)

Effective Fraud Detection in Blockchain Using XGBoost with Random Forest

Pooja Patel, Dr. Mamta Samal, Prof. Saurabh Sharma

Global Nature Care Sangathan Group of Institutions, Jabalpur (M.P)

Abstract: **Since credit cards were widely used in the post-demonetization era, there is a greater chance of fraudulent activity. Banks are susceptible to fraud because they have large databases that include vital business information. This problem affects a number of industries, including the financial, government, business, and consumer sectors. The issue has been made worse by our increasing dependence on cutting-edge technology like cloud and mobile computing, which has made manual detection techniques like audits inefficient and expensive. As a result, financial institutions are using numerical and computational approaches to automate procedures more and more. Detecting subtle anomalies in massive datasets is possible with data mining-based technologies, which have become a feasible solution. Using supervised algorithms, we want to improve the accuracy of fraud detection. Since fraud can take many different forms and data mining techniques are numerous, study is always being done to determine the best course of action in certain situations. Financial fraud, which includes intentional criminal activity for financial benefit, has a considerable negative impact on the economy and society. Credit card fraud alone accounts for large yearly revenue losses.**

Keywords: Fraud detection, Financial fraud, Decision tree.

I. INTRODUCTION

Fraud is the use of a profit organization's system for one's own benefit without necessarily facing immediate legal consequences. It is a widespread activity intended to deceive people or groups in order to obtain financial advantage. Identifying legitimate from fraudulent transactions—which can be generally divided into traditional card-related and online frauds—is the task of credit card fraud detection. In contrast to management fraud, which is carried out by upper management within the firm, customer fraud is committed by those who are not affiliated with the company. Fraud detection, which is essential to total fraud control, minimizes manual intervention by automating and streamlining the screening process. Unauthorized account activity, or the usage of another person's credit card without the cardholder's or issuer's knowledge, is known as credit card fraud. It is critical to detect fraud as soon as it occurs, because fraud detection techniques are always changing to counter the tactics used by criminals. Data mining is the process of gaining knowledge from large datasets through the use of two methods: supervised learning, which uses labeled fraud and real examples, and unsupervised learning, which uses unlabeled data. The problem definition, data preparation, exploration, model construction, validation, deployment, and updating phases comprise the data mining process. In today's highly competitive economy, it is essential to implement efficient fraud prevention and detection procedures in order to minimize financial losses and organizational setbacks caused by fraud. Identity theft and asset misappropriation are just two examples of the illegal behaviors that fall under the umbrella of commercial fraud, an organized crime. Developing fraud detection systems requires the extraction of actionable insights from massive databases through data mining, which makes it easier to find pertinent patterns in large datasets. Choosing features from these datasets that are pertinent to the task at hand is a major obstacle to creating reliable fraud detection algorithms.

II. LITERATURE SURVEY

[Jarrod West, Maumita Bhattacharya]"Intelligent Financial fraud detection"

This author discusses several clever methods for detecting fraud, including statistical and computational approaches. Although the effectiveness of each method varied, it was demonstrated to be fairly capable of identifying a range of financial fraud types. The capacity of computational approaches like neural networks and support vector machines to acquire new skills and adjust to them is a powerful tool in the ever-evolving arsenal of fraudsters. Early research on fraud detection primarily used neural networks and statistical models like logistic regression. Forecasting is one of the financial uses of neural networks. Neural networks

have a long history of being used to detect fraud. However, their training and operation demand a lot of processing resources, which makes them unsuitable for real-time functions. Possibility of overfitting if the training set does not accurately reflect the issue domain, necessitating frequent retraining to accommodate novel fraud techniques. The author of this paper discusses the various types of fraud, including credit card, mortgage, health insurance, and telecommunication fraud. Various methods have been developed for various types of frauds, establishing factors like as entropy and sensitivity, evaluating the effectiveness of various algorithm types, and graphically depicting the results.

[Rasa kanapickiene, Zivile Grundiene] "The model of fraud detection by means of financial ratios"

This author explains about how financial ratios are analysed in order to determine the most fraud-sensitive ratios of financial statements with regard to company managers' and employees' motivation to commit fraud. It was found out that in most cases fraud is committed to show that the company keeps growing and to fulfill obligation conditions. Literary sources offer a wide range of such ratios. Theoretical analysis showed that profitability, liquidity, activity and structure ratios are analyzed most often. Theoretical survey revealed that, in scientific literature, financial ratios are analyzed in order to designate which ratios of the financial statements are the most sensitive in relation with the motifs of executive managers and employees of companies to commit frauds. The logistic regression model of fraud detection in financial statements has been developed.

[Fletcher H. Glancy, Surya B. Yadav] "A computational model for financial reporting fraud detection"

This author explains that the computational fraud detection model is possible to detect financial exposure fraud from the text of annual filings with the Security and Exchange Commission. The model is generalizable because it specifies automatable steps that can be adapted to other domains and genres. A potential application for CFDM is to screen companies for investigation of potential fraud by the SEC (Security and exchange commission). Additional potential applications include financier analysis, e-mail spam detection, and business intelligence validation. A computational fraud detection model (CFDM) was proposed for detecting fraud in financial reporting. CFDM uses a quantitative approach on textual data. It incorporates techniques that use essentially all of information contained in the textual data for fraud detection. Extant work provides a foundation for detecting deception in high and low synchronicity computer- mediated communication (CMC). CFDM provides an analytical method that has the potential for automation. It was tested on the Management's Discussion and Analysis from 10-K filings and was able to distinguish fraudulent filings from non-fraudulent ones. CFDM can serve as a screening tool where deception is suspected.

Siddhartha Battacharya, Sanjeev jha , Kurian Thanakunnel, J Christopher Westland: Data Mining for credit card fraud:

This author says that with the growth in credit card transactions, as a share of the payment system ,there has also been increase in the credit card fraud and most of the U.S consumers are noted to be significantly concerned about identity fraud. While predictive models for credit card fraud detection are in active in use practice, reported studies on the use of web data mining approaches for credit card fraud detection are relatively few, possibly due to the lack of available data for research. In this paper the author evaluates two advanced data mining approaches , support vector machines and random forests. Together with well known logistic regression as part of an attempt to better detect credit card fraud. In this paper the Statistical fraud detection methods have been divided in to two broad categories: supervised and unsupervised. In supervised fraud detection methods , models are estimated based on the samples of fraudulent and legitimate transactions to classify new transactions as fraudulent or legitimate. In unsupervised fraud detection, outliers or unusual transactions are identified as potential cases of fraudulent transactions. Both thee fraud detection methods predict the probability of fraud in any given transaction. Predictive models for credit card fraud detection are in active use.

International Journal of Recent Development in Engineering and Technology
Website: www.ijrdet.com (ISSN 2347-6435 (Online) Volume 13, Issue 3, March 2024)

Other techniques reported for credit card fraud detection include case based reasoning and hidden Markov models. Support vector machines and random forests are sophisticated data mining techniques which have been noted in recent years to show superior performance across different applications. The choice of these two techniques together with the logistic regression is based on their accessibility for practitioners and noted performance advantages.

III. RELATED WORK

The field of financial fraud detection is always changing, with an emphasis on keeping up with the offenders. Still, there are facets of intelligent fraud detection that have not yet been thoroughly investigated. According to surveys on the subject, there are many different kinds of fraud as well as computational methods for identifying the fraudulent acts that con artists commit. These methods include graphically showing computing time and calculating various parameters for each algorithm. Scholars have employed a variety of datasets, such as the German credit card dataset and data from other nations, such as China, to create computational techniques for detecting fraud and assessing the precision of various algorithms. Existing systems analyze the percentage of fraud events and compare different factors among algorithms to detect fraud using ID3 and support vector machine methods. Fraud detection is essential to the contemporary financial sector. With an emphasis on comparing their building times, the suggested system uses supervised learning methods for fraud detection, such as decision tree learning and Naive Bayes classifier. Each technique shows a reasonable competence in identifying various types of financial fraud, despite differences in performance. In particular, computational techniques such as decision trees and Bayesian classifiers are quite useful because of their capacity to adapt and learn from changing fraud strategies. Users can be categorized as good or bad depending on their ability to repay loans by using accessible datasets; each category is represented by a positive or negative count, and sensitivity and efficiency can be computed and graphically displayed.

IV. TYPES OF FRAUDS

There are different types of frauds they are: credit card fraud, financial fraud, mortgage fraud, insurance fraud , telecommunication fraud.

Credit card fraud:

This fraud is defined as the method of purchasing and marketing goods without having money. It is a small plastic card to provide the credit service to the customer. Now a days credit card plays a important role in automated business and online money transaction area which is increasing every year. With the growth of usage of the credit card, fraudsters are finding more opportunities to commit the fraud which causes huge loss to cardholders and banks. Credit card fraud is classified in to two two types:

- Offline credit card fraud:

This kind of fraud is done physically which means the plastic card is stolen by fraudsters and using the card in stocks or supplies or stores or for different purposes as an actual owner. It is an unusual type of fraud because financial organizations will immediately block the card immediately when the card holders report about the theft.

- Online credit card fraud:

This kind of fraud is popular and it is very dangerous, the credit card's information is stolen by the fraudsters to be used in future online transactions by internet or by phone. This kind of fraud is also called as "cardholder not existing" fraud. The card holders can be obtained by the fraudsters through the skimming, phishing or credit card generators.

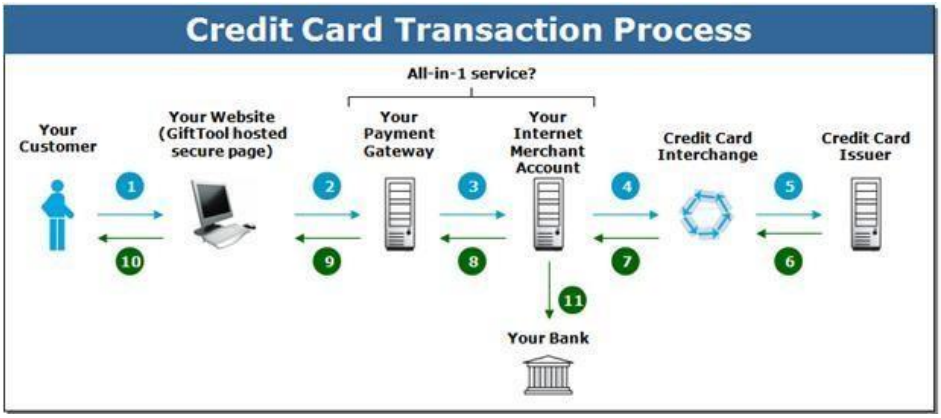
There is another classification for credit card fraud they are application fraud and behavioral fraud. This classification is based on fraudster's strategy on compelling the fraud. Application fraud occurs when the user enters any wrong evidence and wrong details in to the presentation for opening a new credit card. Fraudsters may use other persons information to obtain credit cards or get their new credit cards by using false information with the intention of the never repaying the purchases. Behavioral fraud occurs when fraudsters

International Journal of Recent Development in Engineering and Technology
Website: www.ijrdet.com (ISSN 2347-6435 (Online) Volume 13, Issue 3, March 2024)

obtain credit card holder details to use them later for sales which are made on a cardholder present basis.

V. SUPERVISED LEARNING ALGORITHMS

Supervised learning algorithms are defined as the desired output is known for the input provided in these kind of algorithms we have an input and the desired output is known and we need to map a function for these values . In these supervised learning algorithms predictions are made on the known training dataset and it will be accurate. These learning algorithms are further grouped into regression and classification problems. The Supervised learning algorithms uses a supervised training data where it contains supervised examples. The supervised learning algorithm analyzes the training dataset and produces an classifier. For this initially we need to collect the accurate training dataset and we need to find the accuracy of the function. It is the machine learning task of inferring a function from supervised training data. The training data consists of training examples. In supervised learning , each example is a pair consisting of an input object and a desired output value.. a supervised learning algorithm.



Introduction to decision tree algorithm:

To find an optimal way to classify the learning set initially we need to minimize the depth of the tree. To minimize the tree we need some function information gain. In order to define information gain precisely we need to calculate entropy first.

Entropy:

Without loss of generality, that the resulting decision tree classifies instances into two categories, we'll call them P(positive)and N(negative).

Given a set S, containing these positive and negative targets, the entropy of S related to this boolean classification is: Entropy(S)=

- P(positive)log2P(positive) - P(negative)log2P(negative) P(positive): proportion of positive examples in S P(negative): proportion of negative examples in S Information gain:

To minimize the decision tree depth, when we traverse the tree path, we need to select the optimal attribute for splitting the tree node, which we can easily imply that the attribute with the most entropy reduction is the best choice. We define information gain as the expected reduction of entropy related to specified attribute when splitting a decision tree node. The information gain, Gain(S,A) of an attribute A,

Gain(S,A)= Entropy(S) - Sum for v from 1 to n of (|Sv|/|S|) * Entropy(Sv) The intention of this ordering is:

1. To create small decision trees so that records can be identified after only a few decision tree splitting.
2. To match a hoped for minimalism of the process of decision making

Information gain:

To minimize the decision tree depth, when we traverse the tree path, we need to select the optimal attribute for splitting the tree node, which we can easily imply that the attribute with the most entropy reduction is the best

International Journal of Recent Development in Engineering and Technology
Website: www.ijrdet.com (ISSN 2347-6435 (Online) Volume 13, Issue 3, March 2024)

choice. We define information gain as the expected reduction of entropy related to specified attribute when splitting a decision tree node.

The information gain, Gain(S,A) of an attribute A,

Gain(S,A)= Entropy(S) - Sum for v from 1 to n of (|Sv|/|S|) * Entropy(Sv) The intention of this ordering is:

1. To create small decision trees so that records can be identified after only a few decision tree splitting.
2. To match a hoped for minimalism of the process of decision making Maximum(Gain(S,A)). Create child nodes of this rootNode and add to rootNode in the decision tree.

For each child of the rootNode, apply ID3(S,A,V) recursively until reach node that has entropy=0 or reach leaf node. End ID3.

Naïve bayes classifier:

Introduction to Bayesian Classification The Bayesian Classification represents a supervised learning method as well as a statistical method for classification. Assumes an underlying probabilistic model and it allows us to capture uncertainty about the model in a principled way by determining probabilities of the outcomes. It can solve diagnostic and predictive problems. This Classification is named after Thomas Bayes (1702-1761), who proposed the Bayes Theorem. Bayesian classification provides practical learning algorithms and prior knowledge and observed data can be combined. Bayesian

Classification provides a useful perspective for understanding and evaluating many learning algorithms. It calculates explicit probabilities for hypothesis and it is robust to noise in input data.

Uses of Naive Bayes classification:

1. Naive Bayes text classification The Bayesian classification is used as a probabilistic learning method (Naive Bayes text classification). Naive Bayes classifiers are among the most successful known algorithms for learning to classify text documents.
2. Spam filtering: Spam filtering is the best known use of Naive Bayesian text classification. It makes use of a naive Bayes classifier to identify spam e-mail. Bayesian spam filtering has become a popular mechanism to distinguish illegitimate spam email from legitimate email (sometimes called "ham" or "bacn").[4] Many modern mail clients implement Bayesian spam filtering. Users can also install separate email filtering programs. Server-side email filters, such as DSPAM, Spam Assassin, Spam Bayes, Bogofilter and ASSP, make use of Bayesian spam filtering techniques, and the functionality is sometimes embedded within mail server software itself. 3. Hybrid Recommender System Using Naive Bayes Classifier and Collaborative Filtering (http://eprints.ecs.soton.ac.uk/18483/) Recommender Systems apply machine learning and data mining techniques for filtering unseen information and can predict whether a user would like a given resource. It is proposed a unique switching hybrid recommendation approach by combining a Naive Bayes classification approach with the collaborative filtering. Experimental results on two different data sets, show that the proposed algorithm is scalable and provide better performance–in terms of accuracy and coverage–than other algorithms while at the same time eliminates some recorded problems with recommender systems.

Fig 1.Bar Chart for accuracy measures

VI. CONCLUSION

In recent years, credit card fraud has increased dramatically. One of the main responsibilities for merchant banks is to provide an accurate and user-friendly credit card risk monitoring system in order to raise the degree of risk management for merchants in an automated and efficient manner. Finding the user model that detects fraud instances the best is one of the study's objectives. Credit card fraud can be found in a variety of ways. The likelihood of fraudulent transactions can be anticipated shortly after credit card transactions by the banks if one of these algorithms, or a combination of them, is implemented into the system for detecting credit card fraud. Additionally, a number of anti-fraud tactics can be used to lower risks and shield institutions from significant losses in the past. This study contributes to the use of supervised learning algorithms for credit card fraud detection.

REFERENCES

[1] Linda Delamaire (UK), Hussein Abdou (UK), John Pointon (UK), "Credit card fraud and detection techniques: a review", Banks and Bank Systems, Volume 4, Issue 2, 2009 .

[2] Khyati Chaudhary, Jyoti Yadav, Bhawna Mallick, "A review of Fraud Detection Techniques: Credit Card", International Journal of Computer Applications (0975 – 8887) Volume 45– No.1, May 2012 .

[3] Vladimir Zaslavsky and Anna Strizhak," credit card fraud detection using selforganizing maps", information & security. An International Journal, Vol.18,2006.

[4] L. Mukhanov, "Using bayesian belief networks for credit card fraud detection," in Proc. of the IASTED International conference on Artificial Intelligence and Applications, Insbruck, Austria, Feb. 2008, pp. 221– 225.

[5] John T.S Quah,M Sriganesh "Real time Credit Card Fraud Detection using Computational Intelligence" ELSEVIER Science Direct,35 (2008) 1721-1732.

[6] Joseph King –Fung Pun, "Improving Credit Card Fraud Detection using a Meta Heuristic Learning Strategy" Chemical Engineering and Applied Chemistry University of Tornto 2011.

[7] Kenneth Revett,Magalhaes and Hanrique Santos "Data Mining a Keystroke dynamic Based Biometric Dtatabase Using Rough Set" IEEE