E-Commerce Transaction Fraud Detection through Machine Learning

Rajat Kumar[1], Prajakta Jadhav[2], Rishabh Yadav[3], Sakshi Pawar[4], Prajyot Yawalkar[5], Nehali Shinde[6]

[1,2,3,4,5,6] *IT dept, Dhole Patil College of Engineering, Pune*

---***---

Abstract **— The number of transactions have been steadily increasing consistently over the past few years. The development of online financial services in the form of credit cards, online funds transfer and United Payments Interface or UPI have catalyzed the growth further which has led to the astronomical number of transactions. The number of fraudsters or scammers has also been increasing consistently, which are performing fraudulent transactions. There are numerous fraudulent transaction detection techniques that are put in place by the financial institutions but are unable to detect the ingenious frauds committed by the criminals. Therefore, this paper defines an effective approach for the purpose of fraudulent transaction detection through the use of Linear Clustering, Entropy Estimation and frequent itemset extraction along with Hypergraph formation, Artificial Neural Networks and Decision Making. The extensive evaluation has been performed for quantifying the approach which has resulted in the expected outcomes.**

Keywords: *Artificial Neural Network, Information gain, Hyper graph, neo4j, transaction Fraud.*

I. INTRODUCTION

In the last decade, digital payments have grown at an unprecedented rate. Overall transaction amounts are up over last year, and transfer volumes have increased significantly. Correspondingly, in India, the amount of electronic transactions has increased dramatically in recent years. Internet banking as well as mobile payment have given billions of individual's access to banking services throughout the world. They have also brought tangible opportunities to consumers, organizations, and financial intermediaries, including as the potential to expand, lower operational costs, simplicity of use, convenience, and improved efficiency.

Illegal strategies, on the other side, have quickly adapted to take advantage of the new fast-paced electronic payments scenario. Historically, embezzlement and economic fraud identification depended on a vast number of regulations and fixed criteria, such as maximum transaction restrictions, to identify questionable activities. Such procedural and rule-based procedures have been unsuccessful in past few years as scammers have figured out how to circumvent the inflexible regulations. Contrary to the most recent industry statistics, the amount of worldwide financial corruption is expected to skyrocket. Embezzlement is a big proportion of the unauthorized charges in this category, accounting for a considerable share of the frauds.

The expense of fighting and recovering from fraud has also risen significantly. In conjunction to the rise in fraudulent activity, fraud tactics have evolved significantly. Card embezzlement has decreased dramatically in recent years as digital transactions have been more widely used. However, there has been a significant increase in potential digital transaction fraud. Due to the worldwide epidemic and the resulting substantial surge in digital trade volumes, electronic fraudulent credit card incidents increased dramatically.

Credit card transaction is becoming one of the most popular means of payment in recent years, thanks to the fast advancement of digital transaction. Nevertheless, the advent of contactless electronic monetary operations has resulted in the introduction of new forms of financial fraud. Scam artists frequently collect information and data from investors in order to conduct illicit transactions in a brief span of time. As a result, banking firms should employ a variety of approaches in the physical world and in internet to strengthen credit card fraud monitoring and defend consumers' security.

These cases of frauds are getting increasingly sophisticated and are highly problematic as the current approaches have been insufficient in the detection and identification of the frauds. This has become an increasingly difficult to detect as the static conditions are easily identified and circumvented by these scammers. The development of counter-measures is extremely speedy and these techniques can remain undetected for a long period of time allowing the criminals to perpetuate their crime. Therefore, there is a need for an effective and useful approach for the purpose of achieving the fraudulent transaction detection. The paradigm of machine learning techniques has been one of the most significant for the purpose of transaction fraud detection.

The approach stipulated in this publication utilizes the paradigm of linear clustering. The linearly clustered transactional data along with the features that are extracted are utilized for the purpose of fraud detection. The entropy estimation and frequent itemset mining is performed for the creation of the hypergraph. The hypergraph formed through the use of the Neo4j has been one of the most useful to determine the connection between the various features and attributes. This analysis provides a valuable insight but can be problematic to assess due to the massive and complex structure of the graphs. Therefore, machine learning approaches are a useful tool for insightful information extraction and the detection of transactional fraud which is effectively classified using decision making to achieve the desired results.

In this research article related works are mentioned in the section 2. The proposed technique is deeply narrated in the section 3. The experimental evaluation is performed in section 4 and whereas section 5 concludes this research article with the scope for future enhancement.

II. LITERATURE SURVEY

Masoud Erfani [1] explains that there has been an increase in the number of individuals that utilize the online platform for the purpose of achieving their financial and other banking transactions. This is due to the fact that the process of banking online is very streamlined and can be easily performed with a few clicks. This convenience has led to the majority of the transactions to be performed on the internet which has become a new target for the criminals to achieve their nefarious intents. This is highly problematic as the fraud can lead to a loss of large sums of money and cost a large number of individuals and institutions dearly. Therefore, the authors have proposed the use of support vector data description in a deep manner for the purpose of detection of financial fraud.

Bayu Nur Pambudi [2] elaborates that there has been an increased incidences of utilization of online payment options for the purpose of achieving the transactions. There has been an increased number of transactions have been performed on the online platform in comparison to the offline or cash transactions. This improvement has not been without the repercussions as there has been an increase in the number of frauds and scammers on the online platform which has led to a considerable increase in the number of frauds and other illegal activities such as money laundering. Therefore, the authors in this publication have proposed the use of an optimized SVM or Support Vector Machine for the purpose of achieving the detection of money laundering effectively.

Eren Kurshan [3] narrates that there has been an increase in the number of online transactions or the digital payments that have been increasing in popularity across the world. This increase in the usage can be attributed to the ease of use and the convenience that is offered by these platforms. The improved characteristics have been useful in achieving a large number transactions on this platform every single day with new users being added continuously. There is also a visible increase in the number of different frauds, scams and other criminal activities online which needs to be curbed to increase the reliability. Therefore, the authors have proposed the use of graph computing along with artificial intelligence to detect the presence of fraud.

Abdollah Eshghi [4] expresses the fact that the number of fraudulent transactions are increasing considerably over the past few years. This can be attributed to the enormous increase in the number of transactions in the recent years due to the use of internet platform for the purpose of performing various banking and other financial transactions. This has led to a drastic increase in the number of transactions which has shifted the focus of the criminal individuals. These criminal activities have been highly dangerous to the reliability of the platform which has led to a number of different approaches that have been designed to detect the fraudulent transactions which have been insufficient or imprecise in its applications. To improve this paradigm of fraud detection the authors in this publication have proposed the use of a novel combination of semi-supervised and supervised techniques.

Xiaoguo Wang [5] introduces the exponential increase in the number of users utilizing the online platform for the purpose of achieving an effective improvement in the paradigm of financial transactions. The increase in the number of users and the considerable increase in the number of transactions on the online realm has attracted a considerable number of fraudsters and individuals with nefarious intents. This leads to a large amount of fraud being committed on the online platforms that leads to a problematic scenario which can lead to largescale losses. The authors in this approach have provided an innovative scheme to prevent fraud through the use of K means clustering along with Hidden Markov Model.

Gabriel Castaneda [6] narrates that there has been an increased focus towards maintenance of health due to the problematic scenarios that have been encountered by the general populous. With the pandemic looming large a lot of individuals have shifted towards a healthier lifestyle. This also means that there has been an increase in the number of fraudulent activities that are performed in this sector that can be quite debilitating. The medical fraud is one of the most problematic occurrences that are difficult to identify and detect, and to provide a solution to this problem, the authors have presented an innovative approach that performs fraud detection on big data using the maxout neural network for fraud identification.

Ruoyu Wang [7] states that there has been a large amount of fraud that is being committed online which has been increasing in volume every day. This is due to the fact that the inline transaction shave paved a way for the criminals and the individuals with nefarious intents a lot of leverage in conducting these fraudulent activities. There are several different approaches for the purpose of identification and prevention of fraudulent transactions that have been employed by the financial institutions have a difficult time in assessment. The criminals have become increasingly sophisticated in subverting the conventional detection approach, therefore, the researchers in this approach have proposed the use of statistical identification for the purpose of detection of fraud performed collectively.

Aastha Bhardwaj [8] explains the occurrence of fraudulent transactions that have increased in its frequency in the recent years. The detection of fraud in the financial transactions have been one of the most used strategies for the purpose of attaining a reduction in these practices and make the financial sector more secure and reliable. There have been a large number of different approaches that have been designed to facilitate the detection of a fraudulent transaction but most of the approaches have been insufficient in the perceived accuracy and analysis of the transactions. The authors in this paper have defined an effective strategy for the purpose of achieving analysis in a qualitative manner for the financial statements and identify the fraud.

Na Ruan [9] discusses the prevalence of fraud in various sectors of life. These are the individuals with malicious intents that are highly motivated to commit criminal activities to get their motive. There have been effective techniques that have been developed for the purpose of combating the increasing fraudulent transactions that have been plaguing the financial sector. But the criminals have been getting increasingly industrious and have developed techniques to circumvent the detection mechanisms that are put in place by these institutions. Therefore, the authors in this publication have proposed the use of data mining on Call Detail Records to identify fraud performed in a cooperative manner while preserving the privacy.

Pradheepan Raghavan [10] narrates that since the onset of online transactions and credit cards becoming the norm, there has been an increase in the average transactions that are being done using these platforms. This increase in the volume of transactions has prompted the scammers and other criminals to get attracted to this platforms to fulfil their nefarious intents. There have been increased number of fraudulent activities that are being performed on these platforms due to the lack of awareness and the ingenuity of the scammers. The fraud has no patterns as it is dynamic in nature which makes it highly difficult to detect and identify. Therefore, there is a need for an effective mechanism using deep learning and machine learning to identify fraud.

Rongrong Jing [11] describes the fact that the internet platform has increased the convenience and ease of living for a large sub-section of society. This has been highly useful in the paradigm of financial services that are offered through the use of the internet platform. These allow for effective transactions using the online paradigm that improves the quality of life for their users. This is the reason most of the individuals have been utilizing the internet platform for the purpose of achieving improvement in their lives and enabling online transactions. The online platform introduces the criminals to achieve the goals of fraud on credit cards which can be a problematic occurrence. To improve the missing data and the imbalance of labels the authors have proposed a data quality improvement methodology to achieve better fraud detection on credit cards.

III PROPOSED METHODOLOGY

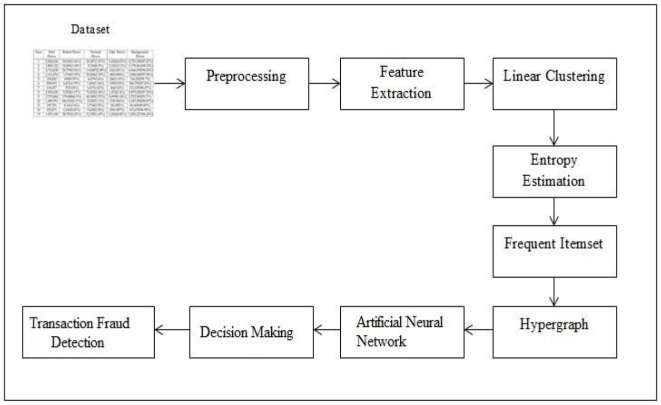


Figure 1: Overview of Transaction Fraud System

The proposed model for E-commerce Fraud detection system is depicted in the figure 1. The steps that are involved in the process are properly narrated below.

Step 1: Dataset Preprocessing and Feature Extraction – This is the initial step of the proposed model, where an E-Commerce dataset has been obtained from the URL - https://www.kaggle.com/vbinh002/fraudecommerce/data.

The obtained dataset is having many attributes like user_id, signup_time, purchase_time, purchase_value, device_id, source, browser, sex, age, ip_address and class. This dataset is stored in a workbook and then it is fed to the proposed model of fraud detection. As the dataset is fed to the system it is being read in the double dimension list, which is then subjected to the preprocessing task.

In preprocessing major attributes are selected and the rest are just vomited or will be used later. Here in this step attributes like source, browser and class attributes are dropped and the rest are aligned in a proper list. From this preprocessed list three attributes are considered as the important features which are eventually plays a vital role in the process of Fraud detection in E-Commerce transaction. For this purpose system selects three attributes like signup_time, purchase_time and purchase_value in three different lists.

Step 2: Linear Clustering and Entropy Estimation – Once the pre-processed list is ready, and then list data is segregated into the different clusters. By preparing clusters it becomes easy to select a cluster data on which neural network can be applied.

In the process of Linear clustering the preprocessed list is divided into 5 equal ranges of indices. Then for each of the range indices the respective single cluster is created and then it is added into the final linear cluster list. This process is indicated in the below mentioned algorithm 1.

ALGORITHM 1: Cluster Formation

```
//Input : Preprocessed List $P_L$,
//Output:Linear Cluster List $C_L$
Start
Index=∅ [Index List]
DIV=$P_L$ size / N [N= Number of Cluster]
begin=0, end=0
for i=0 to N
    Range=∅ [Range List]
    Range[0]=begin
    end=begin+DIV
    Range[1]=end
    Index= Index+Range
    begin=end
end for
for i=0 to Size of Index
    $T_L$=∅ [Temp List]
    R = $Index_{[i]}$
    MIN= R[0]
    MAX=R[1]
    for j=MIN to MAX
        $T_L = T_L + P_{L[j]}$
    end for
    $C_L = C_L + T_L$
end for
return $C_L$
Stop
```

The proposed system uses a protocol to estimate the E-commerce fraud in all the obtained clusters. According to this protocol each and every cluster are measured for the number of rows with the same signup and purchase time and count them as P. Then the cluster size is measured as S. Then by applying Shannon information gain on each cluster a entropy value is estimated that indicates the distribution of number of rows that are eventually matching this protocol as mentioned in the Equation 1.

$$\mathrm{IG} = -\frac{P}{S}\log\frac{P}{S} - \frac{(S-P)}{S}\log\frac{(S-P)}{S} \quad \_\_\_\_(1)$$

Where

IG = Information Gain of the cluster

The obtained Information gain value of the cluster from the Shannon information gain equation is generally lies in between 0 to 1. Any value nearer to 1 indicates the importance of the cluster with the view of the fraud detection process. Each of the cluster index and its respective gain value are stored in a double dimension list. This list is sorted in the descending order to select the best top clusters which eventually contain some quality data for the process of fraud detection.

Step 3: Frequent itemset and Hypergraph estimation – The obtained top cluster data is then merged into a single list. Then from this single list all the transactions where signup time and purchase time is same are identified to collect the respective purchase value in a list. Purchase value list is then subjected to hash set function to get the unique purchase value in a separate list to call it as frequent item list.

The obtained frequent item list is subject to perform insight evaluation of the parameter in single input list. Here signup time and purchase time are evaluated for their equal-ness to extract the purchase value. And the this purchase value is matched with the frequent item list's attribute to form a hyper graph object which contain user ID and purchase value as the nodes of the graph and edge String as "purchase value". This obtained hyper graph is stored in an advanced graph database like neo4j which can be viewed through the browser.

Step 4: Artificial Neural network – The Obtained frequent item list is fed as the input list to the ANN model. Where each of the transaction rows in the frequent item list is subject to evaluate the hidden and output layers for the attributes User ID and purchase value. This is done by considering the same as the target values for the random weights W1,W2,W3,W4,W5,W6,W7,W8,B1,B2. Here B1 and B2 are the bias values which are used to stabilize the neurons. Then by using the Equation 2 and 3 for Hidden layer and Activation function Output layers are estimated. The obtained output layers are aggregated with the target values to achieve the new prediction list as a fraud detection probability list.

$$X = (AT1 * W1) + (AT2 * W2) + B1 \_\_\_\_ (2)$$

$$H_{LV} = \frac{1}{(1+\exp(-X))} \_\_\_\_\_\_ (3)$$

Where AT1 is the USERID and AT2 is the Purchase Value. Then the sigmoid function is given by Equation 3 of the neural network. H_{LV} – Indicates the hidden layer value.

Step 5: Decision Making – The obtained fraud detection probability list from the prior step is used as the input for the Decision Classification process to determine the E-Commerce fraud IDs. In this process initially the minimum and maximum purchase values are estimated from the fraud detection probability list of ANN. And then by using these minimum and maximum values a distance is evaluated in between them as dividend, then this dividend is divided by divisor 5 to get the quotient Q. This Quotient is used to form five decision crisp values like VERY LOW, LOW, MEDIUM, HIGH and VERY HIGH.

The obtained decision crisp values are measured with the fraud detection probability list for their respective ranges to extract the USER ID. And then these USER IDs are classified into different cluster with respect to the decision crisp ranges. These classified clusters eventually indicates the different level of fraud right from the VERY LOW to VERY HIGH range, which is displayed on an interactive user interface.

III. RESULTS AND DISCUSSIONS

The proposed methodology for implementing an accurate e-commerce fraud detection mechanism has been developed utilizing the NetBeans IDE and the Java programming language. The proposed methodology has been developed on a machine running on an Intel Core i5 processor assisted with the 500GB of hard drive as storage and 4GB of physical memory as RAM. For maintaining the database, the MySQL database server is utilized.

Extensive experiments were performed to ascertain the performance metrics of the proposed methodology. For the evaluation of the accuracy of the proposed methodology, the RMSE and MSE performance metrics were used which can meticulously convey the performance of the proposed technique. The performance metrics were evaluated in detail to explain that the e-commerce fraud detection mechanism based on the Artificial Neural Networks and Decision classification elaborated in this paper has been executed appropriately.

Performance Evaluation based on Root Mean Square Error

For the evaluation of the predictive accuracy of the E-commerce Fraud detection system, the Root Mean Square Error (RMSE) is implemented. The RMSE approach is utilized for the calculation of the error rate between the two correlated and continuous entities. In the proposed system, the two correlated and continuous entities are obtained fraudulent transactions and the actual number of fraudulent transactions. This can be measured using equation 4 below.

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n}\left(x_{1,i} - x_{2,i}\right)^2}{n}} \quad \_ (4)$$

Where,

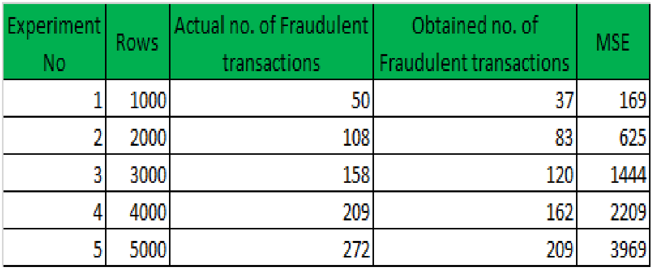
$\sum$ - Summation

$(x_1 - x_2)^2$ - Differences Squared for the summation in between the actual number of fraudulent transactions and the obtained number of fraudulent transactions

n - Number of samples or Trails

An extensive assessment is performed using the RMSE technique, and the values are listed in table 1 below.

Table 1: Mean Square Error measurement



Experiment No	Rows	Actual no. of Fraudulent transactions	Obtained no. of Fraudulent transactions	MSE
1	1000	50	37	169
2	2000	108	83	625
3	3000	158	120	1444
4	4000	209	162	2209
5	5000	272	209	3969

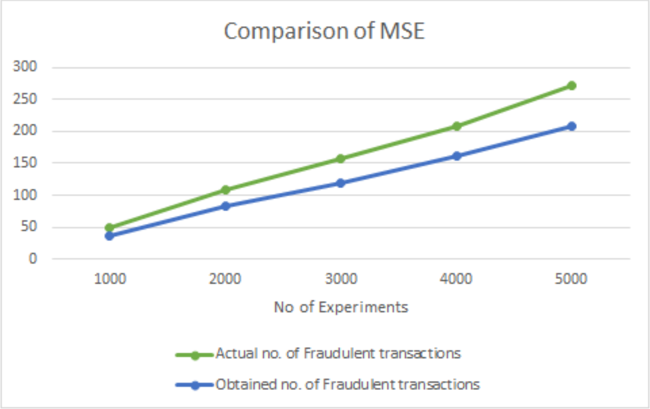


Figure 2: Comparison of MSE in between No of Actual construction waste management labels V/s No of obtained construction waste management labels

Table 1 above and the graph plotted in figure 2 depicts the mean square error rate between the No of obtained fraudulent transactions and No of fraudulent transactions for a set of 5 experiments performed. The first experiment consisted of 1000 entries with subsequent experiments facing an increment of 1000 iteratively. The experimental results achieve an average RMSE of 41.02. The realized RMSE values are calculated for the prediction of fraudulent transactions. Any obtained value for RMSE which is less than 50 is considered a successful system. The achieved RMSE value indicates that the performance derived for the initial implementation of such a system is a highly successful and exceptional achievement.

V. CONCLUSION AND FUTURE SCOPE

There have been large advancements and technological breakthroughs in recent years, the internet paradigm is one of the most significant contributors to the present data scenario. The e-commerce websites have allowed the customers to place orders for products from the comfort of their homes which is highly useful for certain disabled/ movement restricted individuals. This is a novel concept that has also been utilized by individuals with malicious intent to commit fraud which is not as straightforward or easy to detect as the other frauds committed on different platforms. Therefore, this publication details an innovative and accurate fraud detection technique that leverages Machine Learning techniques such as Artificial Neural Networks and Decision Classification. The proposed system has been evaluated through the use of the RMSE technique for its accuracy and the results indicate that it outperforms conventional fraud detection techniques by a large margin. The RMSE value obtained from the system is 41.02 which is acceptable for a first-time implementation.

For future research, the presented technique can be extended further and its accuracy can be increased significantly by the implementation of much more elaborate algorithms.

REFERENCES

[1] M. Erfani, F. Shoeleh and A. A. Ghorbani, "Financial Fraud Detection using Deep Support Vector Data Description," 2020 IEEE International Conference on Big Data (Big Data), 2020, pp. 2274-2282, doi: 10.1109/BigData50022.2020.9378256.

[2] B. N. Pambudi, I. Hidayah and S. Fauziati, "Improving Money Laundering Detection Using Optimized Support Vector Machine," 2019 International Seminar on Research of Information Technology and Intelligent Systems (ISRITI), 2019, pp. 273-278, doi: 10.1109/ISRITI48646.2019.9034655.

[3] E. Kurshan, H. Shen and H. Yu, "Financial Crime & Fraud Detection Using Graph Computing: Application Considerations & Outlook," 2020 Second International Conference on Transdisciplinary AI (TransAI), 2020, pp. 125-130, doi: 10.1109/TransAI49837.2020.00029.

[4] A. Eshghi and M. Kargari, "Introducing a Method for Combining Supervised and Semi-Supervised Methods in Fraud Detection," 2019 15th Iran International Industrial Engineering Conference (IIIEC), 2019, pp. 23-30, doi: 10.1109/IIIEC.2019.8720642.

[5] X. Wang, H. Wu and Z. Yi, "Research on Bank Anti-Fraud Model Based on K-Means and Hidden Markov Model," 2018 IEEE 3rd International Conference on Image, Vision and Computing (ICIVC), 2018, pp. 780-784, doi: 10.1109/ICIVC.2018.8492795.

[6] G. Castaneda, P. Morris and T. M. Khoshgoftaar, "Maxout Neural Network for Big Data Medical Fraud Detection," 2019 IEEE Fifth International Conference on

Big Data Computing Service and Applications (BigDataService), 2019, pp. 357-362, doi: 10.1109/BigDataService.2019.00064.
[7] R. Wang et al., "Statistical Detection Of Collective Data Fraud," 2020 IEEE International Conference on Multimedia and Expo (ICME), 2020, pp. 1-6, doi: 10.1109/ICME46284.2020.9102889.

[8] A. Bhardwaj and R. Gupta, "Qualitative analysis of financial statements for fraud detection," 2018 International Conference on Advances in Computing, Communication Control and Networking (ICACCCN), 2018, pp. 318-320, doi: 10.1109/ICACCCN.2018.8748478.

[9] N. Ruan, Z. Wei and J. Liu, "Cooperative Fraud Detection Model With Privacy-Preserving in Real CDR Datasets," in IEEE Access, vol. 7, pp. 115261-115272, 2019, doi: 10.1109/ACCESS.2019.2935759.

[10] P. Raghavan and N. E. Gayar, "Fraud Detection using Machine Learning and Deep Learning," 2019 International Conference on Computational Intelligence and Knowledge Economy (ICCIKE), 2019, pp. 334-339, doi: 10.1109/ICCIKE47802.2019.9004231.

[11] R. Jing et al., "Improving the Data Quality for Credit Card Fraud Detection," 2020 IEEE International Conference on Intelligence and Security Informatics (ISI), 2020, pp. 1-6, doi: 10.1109/ISI49825.2020.9280510.