A Quantitative Analysis to Estimate Transaction Fraud using Machine Learning

[1]Amruta Dhole, IT dept., Dhole Patil College of Engineering, Pune
[2]Rajat Kumar, IT dept., Dhole Patil College of Engineering, Pune
[3]Prajakta Jadhav, IT dept., Dhole Patil College of Engineering, Pune
[4]Sakshi Pawar IT dept., Dhole Patil College of Engineering, Pune
[5]Rishabh Yadav, IT dept., Dhole Patil College of Engineering, Pune
[6]Prajyot Yawalkar, IT dept., Dhole Patil College of Engineering, Pune

---

Abstract — Fraud is a highly nefarious and self-centered crime that is happening quite frequently on the various platforms. As the increase in the users has also led to an increase in fraud being committed on the financial portals. The fraud on financial portals is quite varied and is governed by a plethora of parameters that are highly difficult to ascertain. There is a wide variety of researches that facilitates the detection of financial fraud. But most of these approaches have been directed towards credit cards, money laundering, etc. these researches fail to consider the overall attributes specifically. Therefore, to combat this problem, this publication deals with the identification of fraud on a variety of transactions. The proposed system implements innovative concepts such as linear clustering, entropy estimation, and Frequent Itemset Mining along with the Hypergraph, Artificial Neural Networks, and Decision making for identification of transaction fraud.

Keywords: *Feature Extraction, Linear Clustering, Entropy Estimation. Hypergraph, Artificial Neural Networks..*

I. INTRODUCTION

Transactions are any type of give and take or a swapping of goods or services that occurs between two individuals or entities. This interactions have been utilized since millennia for the purpose of trade and commerce by human beings. These transaction have been effective in realization of the capitalist commerce that have nowadays. These transaction have been essential in providing the much needed progress and advancements in the society. Legitimate trade and transaction catalyzes growth which has been the driving factors for the realization of an improved living standards for every single human being.

The financial institutions have been the main components in the realization of the various transactions and their details. This is due to the fact that the financial authorities have certain rules and regulations to ensure that every transaction is fair and just for a number of individuals. This is crucial to determine as any kind of impartial transaction would be unjust and the institution would be termed as biased which could lead to the loss of trust in such organization. Therefore, it is essential for the financial institutions to judge each and every transaction that occurs without any bias or prejudice and ensure that each and every transaction abides by the regulation stipulated by the institution or governmental organization.

There are a number of different scenarios as well as individuals with nefarious intents that are performing fraudulent transactions that can be problematic to identify and process. There has always been the possibility of a person with malicious intents being present among a gathering of otherwise regular individuals since the start of time. This is because there are bad apples in every basket, and there is little that can be done about it apart from being equipped for it and inventing procedures to detect any harmful action undertaken by the user. This sort of behavior has resulted in several wars and other confrontations, which have already been elevated to the global level. As a result, it is inappropriate conduct, and there is an urgent need to minimize such behavior. There are laws and other rules in place to keep individuals in check and preserve a state of peace.

Fraud is a serious wrongdoing that is equivalent to robbing someone is hard-earned earnings or possessions through unethical ways. Fraud has a detrimental effect on many individuals and lowers the integrity of the service provided to users and customers on financial portals. Strategies should be developed to minimize the prevalence of fraud. Much study has been conducted on this paradigm, however the problem of fraud is a very complicated phenomenon with a plethora of methods to perform the deception. This makes fraud more difficult to identify and substantiate, giving the perpetrators a benefit over law enforcement officers.

For this purpose an effective methodology for the detection of fraudulent transactions is required which is achieved through the use of efficient analysis of related works which have been outlined in this survey article. The analysis of these works have helped in devising an effective strategy for fraud detection which will be outlined in the upcoming editions of this research.

This literature survey paper segregates the section 2 for the evaluation of the past work in the configuration of a literature survey, and finally, section 3 provides the conclusion and the future work.

II. RELATED WORKS

P. Raghavan et al. [1] present an actual study comparing several machine learning and deep learning models for the identification of fraudulent transactions on diverse data sets. The primary goal of this research is to determine which strategies are best suited for certain types of datasets. Because many businesses are investing in innovative strategies to enhance their bottom lines these days, this research might assist practitioners and businesses in better understanding how different methods operate on different datasets. SVMs, maybe paired with CNNs for more dependable performance, are the best approaches for detecting fraud with larger datasets, according to research. SVM, Random Forest, and KNN ensemble techniques can yield good improvements for smaller datasets. Convolutional Neural Networks (CNN) outperform other deep learning approaches such as Autoencoders, RBM, and DBN in most cases.

M. Erfani et al. provided an effective methodology for detecting fraud. The proposed architecture included a unique clustering-based subsampling phase, followed by a deep support vector data description step for fraud detection. Unbalanced datasets are one of the most difficult problems in fraud datasets [2]. Their solution addresses this issue in two ways: they propose the subsampling method to effectively pick a subset of non-fraud data, and train one-class DeepSVDD as an unsupervised one classification method. Based on ROC-AUC and AP, a recommended measure in unbalanced scenarios, their model produces encouraging results. For both versions of DeepSVDD and state-of-the-art machine learning classifiers, namely SVM and RF, the authors offered a trend analysis based on the size of the test dataset and the size of the training dataset. Deep one-class classifiers beat state-of-the-art machine learning classifiers in both assessment metrics for fraud detection.

D. Huang et. al. presented CoDetect a novel framework that can detect fraud using a graph-based similarity matrix and a feature matrix at the same time. It presents a novel technique of revealing the nature of financial operations, such as fraud patterns and dubious property. Furthermore, the framework offers a more understandable method of detecting fraud on sparse matrices [3]. The suggested system can efficiently detect fraud patterns as well as suspicious traits, according to the experimental outcome on synthetic and real-world data sets. Executives in financial supervision may use this co-detection framework to not only detect fraud trends but also to track down the source of fraud with suspicious characteristics.

N. Ruan et al. present a Cooperative Fraud Detection methodology to detect sophisticated fraudsters that use many operators to hide their bad conduct by broadcasting phone calls. The authors present a complete and effective matching approach to detect fraudsters, as well as an efficient and accurate profiling method to profile the behavior of mobile phone users. Meanwhile, they effectively limit privacy leaks in the cooperative model. To evaluate the viability of their methodology, they built a real-world scenario utilizing genuine CDR data given by a large telecom company in China. The outcome demonstrates that the presented approach still performs admirably in a real-world context [4].

The relevance of examining the language data in financial reports in detecting financial statement fraud was explored by A. Bhardwaj et al.. They also presented a text mining method for finding financial statement fraud by uncovering hidden indicators in financial statements. The procedure begins with the gathering of financial statements from both fraudulent and non-fraudulent businesses, followed by text extraction. Pre-processing and lexical analysis of the text is required [5]. The next step is to employ a fraud detection algorithm to identify a pattern or connection that may be utilized to differentiate between fraudulent and non-fraudulent reports. The algorithm should accurately distinguish fraudulent and non-fraudulent businesses, and the classification accuracy should be tested in conventional methods.

R. Wang et al. [6] provide a set of collaborative anomaly detection algorithms that can aid in the identification of data manipulations in current data pipelines and data centers. Unlike other methods for detecting collective abnormalities, their methodology uses statistical distance to determine similarity. The authors looked into numerous technical aspects of the algorithm's design and ran a comprehensive experiment to see how effective it was. The benefits of their method were also demonstrated in the comparative experiment. It may be inferred that their method effectively detects abnormalities in data sets and that the classifier is sensitive enough to detect real-world data manipulations.

G. Castaneda and colleagues examine and contrast four max-out functions with typical activation functions as tanh, ReLU, LReLU, and SeLU. They also compare how long it takes to train four different activation functions. The authors investigate whether marginal performance gains from max out are due to the activation function or just a 2x increase in the number of convolutional filters as compared to ReLU networks. They also decide whether max-out techniques converge faster than regular activation functions and if they outperform them inaccuracy [7].

X. Wang et al. concentrate on the demands of banking transaction anti-fraud modeling and transaction detection and researches bank anti-fraud modeling and application using K-means clustering and the Hidden Markov model [8]. Simulated and real-world bank data verification studies have shown that their approach can identify bank transaction data to a certain extent and that it can perform well for low, medium, and large user groups, resulting in an effective solution to bank fraud concerns.

A. Eshghi et al. described the research and practice in the actual world, as well as a framework consisting of three basic components, namely RBC, TAC, and SBC. Transaction aggregation and derived characteristics have been demonstrated to be beneficial in fraud detection in prior research. Rather than using classifier algorithms to extract rules, trend analysis was performed as a semisupervised approach, and it was revealed that, while semisupervised methods had lower detection rates than classifiers or rule-based methods, combining them leads to better results. Furthermore, unsupervised and semi-supervised approaches must be used for fraud detection systems, especially when there is no labeled data to use for supervised methods, which is a prevalent issue in most research in this field [9].

E. Kurshan et al. provide a hands-on look at the use of graph computing in financial fraud detection applications. They discuss the challenges that development organizations confront while developing and deploying graph-based solutions in financial transaction processing systems. Financial crime detection and graph computing are both vast and quickly expanding topics. Because of infrastructure and tool restrictions, many particular use cases necessitate customized efforts [10]. If robustness is not addressed as a fundamental design goal in solution development, adversarial methods are likely to grow more difficult in the future. Finally, by concentrating on application needs and implementation challenges, present and new graph-based systems have the potential to greatly increase their performance.

B. N. Pambudi et al. present a method for improving the performance of a data mining strategy for fraud detection in a dataset with a large number of unbalanced financial transactions. The Random Under Sampling (RUS) approach is utilized to improve the machine learning methodology based on Support Vector Machine (SVM) [11]. The classifier can detect fraud more successfully with this combo strategy. Metrics like accuracy, recall, f1-score, and Area Under Precision-Recall Curve are used to evaluate model performance (AUPRC).

By combining data completeness with sampling, R. Jing et al. suggested a unique strategy for improving data quality for credit card fraud detection. To limit the impact of missing data, they first apply the spectral regularization approach to complete the sparse matrix of the dataset [12]. In addition, they use an over-sampling approach to address the flaws of the imbalance of positive and negative samples, ensuring that the proportion of samples in the dataset is constant. On the dataset, they examine the performance of discarding missing values, standard matrix completion procedures, and spectral regularization approaches.

III. 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. There has been an inordinate increase in the number of different transactions that are being performed on a daily basis. These transactions have been increased due to the increase in the spending power of the individuals across the globe. This is useful as it allows the business to flourish and commerce to sustain a country and the economy. But this increase in the number of transactions also increases the number of fraudulent transactions that leads a lot of problematic scenarios. Fraud nowadays is not as straightforward or easy to detect as the individuals deploy ingenious ways to evade the already existing approaches to detect fraud. Therefore, this survey article analyzes a number of innovative and accurate fraud detection techniques that have been useful for deriving our approach that leverages Machine Learning techniques such as Artificial Neural Networks and Decision making and will be discussed in the upcoming researches on this topic.

REFERENCES

[1] 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.

[2] 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.

[3] D. Huang, D. Mu, L. Yang and X. Cai, "CoDetect: Financial Fraud Detection With Anomaly Feature Detection," in IEEE Access, vol. 6, pp. 19161-19174, 2018, DOI: 10.1109/ACCESS.2018.2816564.

[4] 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.

[5] 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.

[6] 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.

[7] 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.

[8] 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.

[9] 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.

[10] 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.

[11] 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.

[12] 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.