IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
Volume 2, Issue 3, November 2022
Impact Factor: 6.252

Use of Machine Learning in Financial Fraud Detection: A Review

Sumukh Uday Rabade
Student, Department of Computer Engineering
NBN Sinhgad School of Engineering, Pune, Maharashtra, India
sumukhrabade@gmail.com

Abstract: *Financial factors are manipulated to produce fraudulent financial activities, which are produced by overvaluing revenues, assets, sales, and profits while undervaluing costs, liabilities, or losses. Traditional approaches, such as manual auditing and inspections, are expensive, inaccurate, and time-consuming for detecting such bogus statements. Auditors can analyse several financial statements with the use of intelligent methods. In this paper, we thoroughly analyse and summarise the body of knowledge on detecting intelligent fraud in corporate financial accounts. The exploration of machine learning and data mining techniques, as well as the many datasets under investigation for financial fraud detection, is the specific emphasis of this paper. This study provides insight on selecting appropriate approach with different types of datasets while taking into consideration the trade-offs of speed, accuracy, and cost.*

Keywords: Fraud detection, financial statement, machine learning, literature survey, machine learning algorithms.

I. INTRODUCTION

Financial fraud refers to the use of fraudulent and illegal methods or deceptive tactics to gain financial benefits. Fraud can be committed in different areas of finance, including banking, insurance, taxation, and corporates, and more. Fiscal fraud and evasion, including credit card fraud, tax evasion, financial statement fraud, money laundry, and other types of financial fraud, has become a growing problem. Despite efforts to eliminate financial fraud, its occurrence adversely affects business and society as hundreds of millions of dollars are lost to fraud each year. This significant financial loss has dramatically affected individuals, merchants, and banks. Nowadays, fraud attempts have increased drastically, which makes fraud detection more important than ever.

II. WHAT DOES MACHINE LEARNING-BASED FRAUD DETECTION MEAN?

Machine learning in fraud detection refers to a group of artificial intelligence (AI) algorithms that have been trained using your past data to identify risk rules. The rules can then be put into effect to prevent or permit specific user actions, such as shady logins, identity theft, or fraudulent transactions. To prevent false positives and to increase the accuracy of your risk rules, you must trace earlier instances of fraud and non-fraud while training the machine learning engine. The rules predictions will be increasingly accurate as the algorithms run longer. The first step in ML implementation for fraud detection is the training phase, during which the machine learning algorithm is "trained" how to spot the indicative signals of a fraudulent transaction. Large data sets containing earlier fraudulent transactions must be ingested in order to accomplish this. After a thorough training process, the algorithm will begin to learn additional patterns and abnormalities that are symptomatic of malpractices those it was initially intended to recognise. Due to the complexity involved in identifying fraudulent activity in the banking, finance, and other industries, several machine learning models—as well as training techniques—are implemented in this endeavour.

III. ALGORITHMS FOR FRAUD DETECTION

Numerous algorithms are available for fraud detection. However, there isn't a single optimum machine learning method for fraud detection because it all depends on the data you currently have. The algorithms listed below are some of the more well-known ones, but this is by no means a universal list.

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
Volume 2, Issue 3, November 2022
Impact Factor: 6.252

3.1 Logistic Regression

The simplest yet most effective approach for predicting true or false (binary) values is logistic regression. By fitting the data to a logistic function, it estimates discrete values (often binary values like fraud/no fraud) from a set of independent variables.

3.2 Decision Trees

Another well-liked approach for separating or classifying data is decision trees. The model of a decision tree is a simple set of principles, which makes it extremely intriguing. You may take these principles and develop a rules-based system to improve things even further. The model, however, is in no manner a rules-based system because minute adjustments to the underlying data might lead to whole new rules.

3.3 Random Forests

An algorithm called random forest builds on several decision trees to produce more precise classifications. Because it accomplishes this by averaging the outcomes of various decision trees, its predictive power is higher. Large training datasets with many input variables and random forests perform well together. Random forests, on the other hand, are harder to understand than decision trees. You end up with numerous rules rather than just one set. This might be problematic, particularly if an explanation is required for system compliance or other regulations.

3.4 K- Nearest Neighbour

This straightforward algorithm classifies any new cases by receiving a majority vote from its top k neighbours while storing all of the existing cases. It does this by using a distance function, similar to the Euclidean distance. A model is not exactly created during the training phase. Instead, "training" and "classification" take place as needed. Because of this, the KNN technique requires slightly more computing power than other machine learning algorithms to detect fraud.

3.5 K-Means

Clustering issues are resolved by this unsupervised learning approach (Different from KNN). The technique divides a given dataset into a Number of clusters, with the goal of making each cluster's data Points as similar as feasible. It also uses a distance function, just like KNN.

IV. HOW DOES MACHINE LEARNING MODEL WORK?

4.1 Providing the Data for Input

To get started, every AI or ML system requires data. It will be transactional information in this case, such as:
Monetary exchange

- Product SKU
- Credit card type

But we'll also include information about how users access the website:

- VPN, proxy, or Tor usage, among other IP data device types.

Keep in mind that your results will be more accurate the more data you have to work with. This is especially significant if your fraud detection software does not support custom fields because you might be overlooking important data.

4.2 Setting the Rules

Two primary categories of rules can be produced:

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
Volume 2, Issue 3, November 2022
Impact Factor: 6.252

- Heuristic rules with a single parameter include the following, as an illustration: if the IP is X, block.
- Multiple parameters are included in complex rules. A score for each specified rule is displayed. To tighten or relax the circumstances for triggering, you can change the accuracy thresholds.

4.3 Analysing and Implementing the Rule

You can filter the rules using any type of data point, including the type and anticipated accuracy. The accuracy component, which is calculated using a challenging confusion matrix, is very helpful.
Machine-learning suggestions are disabled by default. The ON/OFF switch will enable them instantly. Additionally, it is possible to manually establish and modify the rule's triggering thresholds.

4.4 Algorithm Training

The secret to improving accuracy and modifying rules is to provide feedback data.

4.5 Using Historical Data to Test Rules

You should be able to go back and review earlier cases in a good fraud prevention programme to see if the rules would have been helpful. The rules can be turned on and off in a sandbox setting where you can see for yourself how accurate they are.
When a test is run, a confusion matrix based on prior transactions over the chosen time period is generated, highlighting the rule's projected accuracy rate.
A confusion matrix, also known as an error matrix, is a table design used in machine learning that enables the display of an algorithm's performance. This gives you the option to determine accuracy over a range of dates, which you can choose from the most recent hour to the most recent year.

V. THE ADVANTAGES OF MACHINE LEARNING FOR FRAUD MANAGEMENT

- **Faster and efficient detection:** Because machines can handle vast datasets considerably faster than humans, they can slice and dice massive volumes of data. That implies: Faster and more effective detection: The technology can spot suspicious patterns and actions that could have taken human agents months to discover.
- **Reduced manual review time**: In a similar manner, letting computers analyse all the data points for you can significantly cut down on the time spent manually examining information.
- **Larger datasets improve predictions**: A machine learning engine gets better trained the more data you feed it. Consequently, while enormous datasets can occasionally make it difficult for people to identify patterns, the situation is exactly the opposite with an AI-driven system.
- **Cost-effective remedy**: You only need one machine-learning system to process all the data you put at it, regardless of volume, as opposed to adding more Risks agents. This is perfect for companies who see seasonal fluctuations in traffic, checkouts, or signups. A machine learning system can help your business grow without significantly raising risk management expenses at the same time.
- Not to mention, algorithms don't require rest, holidays, or pauses. Even the greatest fraud managers could show up to work on Monday morning with a backlog of manual reviews. Fraud attacks can occur around the clock. By separating the situations that are obviously false or acceptable, machines can speed up the process.

VI. DISADVANTAGES OF ML

- **Cost:** Companies will need to hire a team of data scientists to create and maintain the system if they want to build a machine learning system in-house. Businesses must spend in data management and storage since there are vast volumes of data involved, and this takes up additional network space. Racketeers will be able to get around the machine learning model if it is constructed improperly, which will result in a flood of fraudulent transactions and subsequent chargebacks.
- **High degrees of necessary technical proficiency**: Building a proper, error-free machine learning model is essential. To do that, businesses require educated professionals with experience creating such systems and in-depth knowledge of the peculiar subject of payment fraud. It may be wiser to delegate the tough lifting to

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
Volume 2, Issue 3, November 2022
Impact Factor: 6.252

qualified service providers since a poor data scientist is likely to make fatal errors and neglect crucial information.

- **Difficult to gather reliable data:** It takes a lot of high-quality data, which is hard to come by, to develop a machine learning model that is functional over the long term. There must be a sufficient number of transactions and sufficient data in each for datasets to be statistically meaningful. This also offers details on how those transactions turned out.

VII. USE CASES

1. **Online Retailers and Fraud in Transactions:** It can be challenging to analyse data for thousands of transactions. For this reason, a lot of fraud managers for a lot of big E-Commerce websites employ machine learning to comprehend why some transactions weren't at first identified by the system as fraudulent. As a result, after letting your machine learning system run for a while, you can discover which products fraudsters target the most, what shipping information poses the greatest danger, which card payments should be banned to reduce the likelihood of chargebacks, among other things
2. **Compliance with Financial Institutions:** To avoid regulatory penalties, fintech businesses, well-established financial institutions, and even insurance providers must adhere to stringent compliance rules. In other words, companies must make sure they are working with legitimate users and not scammers. To be competitive, they must also work quickly. This is how fake profiles get past the filtering system. Many of these businesses can obtain crucial information about what distinguishes a real user profile from a phoney one by implementing a machine learning system.
3. **False Insurance Claims:** The insurance industry, particularly in the healthcare industry, can benefit from the implementation of machine learning systems to improve fraud detection. Algorithms can quickly spot erroneous and redundant claims; for instance, we might be dealing with a consumer who claimed an inaccurate diagnosis or overstated the expense of their medical insurance. Natural language processing, which guarantees a thorough examination of unstructured data such as medical reports, is one of the most useful machine learning approaches in the healthcare industry. These systems use machine learning to scan patient, insurance, or client-written papers for apparent discrepancies.
4. **Fraudulent Chargebacks and Payment Gateways:** Another illustration of how difficult it is to manually review every transaction, particularly when speed is crucial. Employing human agents to review every transaction would be practically difficult given that payment gateways must handle thousands of transactions as rapidly as possible. You can teach a machine learning engine to recognise fraudulent transactions that might otherwise result in chargeback costs, acting as a sort of fraud monitoring analytics system.

VIII. LITERATURE SURVEY

The numerous works that have been done in computer fraud and security are presented in this section.

- In this study, the effectiveness of a QML system for fraud detection is compared to systems based on 12 machine learning methods. The results demonstrate the potency of our suggested approach for detecting fraud and the exceptional performance of QML, particularly when used to a time-series-based, severely unbalanced, highly dimensional dataset. By providing a direction for future QML research, our study adds to the detection literature. Despite the increased interest in using quantum to tackle practical issues, the findings warn that, in order to offer the best business solution, practitioners must take into account the trade-offs between accuracy, speed, and computational cost as well as the type of data the system uses. Applications of quantum computing can process time-series, high-dimensional, and highly unbalanced datasets, which are difficult for regular machine learning methods to process. Regarding non-time-series data, traditional machine learning continues to play a large role since it is the more practical and affordable solution before quantum computing makes substantial advances. Given the incredibly quick detection speed, the QML system used in this study provides data that are almost immediately available, which is a significant step in the direction of real-time fraud detection [1].

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
Volume 2, Issue 3, November 2022
Impact Factor: 6.252

- In this paper, Following the results analysis and review, the following conclusions will be withdrawn: For fraudulent credit card transactions, the Logistic Regression classification accuracy is 80%, with precision of 78%, recall of 100%, and F1-Score of 88%.For fraudulent credit card transactions, the K-Nearest Neighbour classification accuracy is 78.96%, with precision of 83%, recall of 91%, and F1-Score of 77%.Naive Bayes' classification accuracy for fraudulent credit card transactions is 73.90%, with precision of 82.25%, and recall of 80.52%.Comparative analysis depicts that the basis on parameters, i.e. Precision/Recall and F1-Score the K-Nearest Neighbour, is the better approach for detecting fraudulent transactions than the Logistic Regression and Naïve Bayes. However, the accuracy is marginal high of Logistic Regression, but the False Positive parameters cannot identify the imbalanced data; therefore, they disguise the results and accuracy of Logistic Regression.[2]
- The ability to make payments without having currency on hand is made incredibly convenient for the user in this article by cashless transactions. As more people switch to digital payments, the likelihood of identity theft increases. The best acceptable algorithm for identifying fraud using an unbalanced data set and several supervised machine learning algorithms is decision tree. [3]
- According to the study in this paper, the Random Forest algorithm can prove to be a good choice because of its advantages like higher dimensionality and accuracy. It is capable to solve both classification and regression issues.
- Each decision in Random Forest has a high variance and several decision trees are used. Since each decision tree is trained using a separate sample set of data, the majority output from all the decision trees is taken into consideration and the low variance is attained. Therefore, when a small portion of the dataset is altered, the decision tree's accuracy and output are unaffected. The most crucial aspect of random forest is this. The most crucial aspect of random forest is this. When using a classifier, the majority of all outputs are used; when using a regression model, the total number of outputs is used. [6]

IX. Results Based From Survey

Identified ML/DM techniques, the data sources, and the research trends. However, there are still some limitations and validity threats. We found out that the best acceptable algorithm for identifying fraud using an unbalanced data set and several supervised machine learning algorithms is decision tree.After Comparative analysis we found that on the basis on parameters, i.e. Precision/Recall and F1-Score, the K-Nearest Neighbour, is the better approach for detecting fraudulent transactions than the Logistic Regression and Naïve Bayes. The Random Forest algorithm can prove to be a good choice because of its advantages like higher dimensionality and accuracy. It is capable to solve both classification and regression issues. Further Research should focus on how the accuracy of the models can be improved by including complex parameters, use unstructured information and train them accordingly to give us the best results.

Models	Recall value	F1-score	MCC value
SVM	0.92	1.00	0.558
Naïve Bayes	0.91	0.98	0.761
Logistic Regression	0.83	1.00	0.761
KNN	0.84	1.00	0.793
Random Forest	0.89	1.00	0.848

Table 1:Comparative Study of MCC Values of ML Models.

The Matthews correlation coefficient (MCC) is a more dependable statistical measure that only yields a high score if the prediction performed well in each of the four categories of the confusion matrix (true positives, false negatives, true negatives, and false positives), proportionally to the size of the dataset's positive and negative elements.

By first describing the mathematical principles of MCC, followed by the advantage of MCC in six hypothetical use cases and different scenarios, MCC generates a more informative and true result when evaluating categorical

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
Volume 2, Issue 3, November 2022
Impact Factor: 6.252

classifications than accuracy and F1 score. Researchers believe that all scientific communities should choose the Matthews correlation coefficient above accuracy and F1 score when assessing binary classification tasks.

Now if we look at the comparative table we see that the Random forest algorithm has a mcc value of 0.848 which is the highest amongst all and the SVM has the mcc value of 0.558.However the Recall value of SVM is 0.92 which is the highest amongst all.

After comparing the MCC values of different ML algorithms, the Random forest algorithm also proves to be a good method for detecting financial frauds

Method Used	Frauds	Genuines	MCC
Naïve Bayes	83.130	97.730	0.219
Decision Tree	81.098	99.951	0.775
Random Forest	42.683	99.988	0.604
Gradient Boosted Tree	81.098	99.936	0.746
Decision Stump	66.870	99.963	0.711
Random Tree	32.520	99.982	0.497
Deep Learning	81.504	99.956	0.787
Neural Network	82.317	99.966	0.812
Multi Layer Perceptron	80.894	99.966	0.806
Linear Regression	54.065	99.985	0.683
Logistic Regression	79.065	99.962	0.786
Support Vector Machine	79.878	99.972	0.813

Table 2: List of different ML Algorithms and their accuracy.

X. FUTURE SCOPE

Future studies should concentrate on how to train the models to produce the best outcomes by adding additional factors to improve the models' accuracy. Future studies on fraud detection should concentrate on unsupervised, semi-supervised, as well as bio-inspired and evolutionary heuristic techniques. Future study is anticipated to use textual and auditory data in terms of datasets. This unstructured data necessitates additional investigation because, despite creating new difficulties, it can produce intriguing findings for intelligent fraud detection. Organizations must collaborate and share their prior fraud impacts in order to effectively prevent fraud. A walled approach is not a viable paradigm.

In order to support a global fraud prevention strategy, a platform is urgently needed. The idea is to create an operating framework and model to encourage businesses all over the world to use it by abiding by its standards. By doing this, they can share and use fraud patterns to proactively alert and be alerted on fraudulent transactions, thereby strengthening the layer of security for their applications. Businesses collaborate and share their fraud experiences. With the proper technology in place and contemporary approaches being applied to the data already available, fraud will be prevented internationally, benefiting all the organizations.

REFERENCES

[1]. Haibo Wang; Wendy Wang; Yi Liu; Bahram Alidaee "Integrating Machine Learning Algorithms With Quantum Annealing Solvers for Online Fraud Detection" July 2022.

[2]. IG. Jaculine Priya, Dr.S. Saradha "Fraud Detection and Prevention Using Machine Learning Algorithms", 2021.

[3]. Samidha Khatri.,Aishwarya Arora., and Arun Prakash Agarwal., "Supervised Machine Learning Algorithms for Credit Card Fraud Detection: A Comparion".,IEEE.,2020.

[4]. R. Sailusha, V. Gnaneswar, R. Ramesh, and G. R. Rao, ''Credit card fraud detection using machine learning,'' in Proc. 4th Int. Conf. Intell. Comput. Control Syst. (ICICCS), May 2020, pp. 1264–1270.

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
Volume 2, Issue 3, November 2022
Impact Factor: 6.252

[5]. NayanUchhana, Ravi Ranjan, Shashank Sharma, Deepak Agrawal, Anurag Punde."Literature Review of Different Machine Learning Algorithms for Credit Card Fraud Detection" April 2021.

[6]. SudeepDogga,"The Role of Random Forest in Credit Card Fraud Analysis "December 2020.