2023 7th International Conference on Intelligent Computing and Control Systems (ICICCS) | 979-8-3503-9725-3/23/$31.00 ©2023 IEEE | DOI: 10.1109/ICICCS56967.2023.10142404

Fraud Detection in Online Payments using Machine Learning Techniques

U.Siddaiah
Department of InformationTechnology
Velagapudi Ramakrishna Siddhartha
Engineering College
Vijayawada, India
upputurisiddaiah096@gmail.com

P.Anjaneyulu
Department of InformationTechnology
Velagapudi Ramakrishna Siddhartha
Engineering College
Vijayawada, India
posanianjaneyulu12345@gmail.com

Y.Haritha
Department of Information Technology
Velagapudi Ramakrishna Siddhartha
Engineering College
Vijayawada, India
harithayamarthi@gmail.com

M Ramesh
Department of Information Technology
Velagapudi Ramakrishna Siddhartha
Engineering College
Vijayawada, India
ramesh.welcome@gmail.com

***Abstract* — People rely on online transactions for nearly everything in today's environment. Online transactions offer several benefits, such as ease of use, viability, speedier payments, etc., but they also have some drawbacks, such as fraud, phishing, data loss, etc. As online transactions grow, there is a continuing risk of frauds and deceptive transactions that could violate a person's privacy. In order to prevent high risk transactions, numerous commercial banks and insurance firms invested millions of rupees in the development of transaction detection systems. This research study has introduced a feature-engineered machine learning-based model for detecting transaction fraud. By processing as much data as it can, the algorithm can gain experience, strengthen its stability, and increase its performance. The effort to detect online fraud transactions can use these algorithms. In this, a dataset of specific online transactions is obtained. Then, with the aid of machine learning algorithms, unusual or distinctive data patterns that will be helpful in identifying any transactions that are fraudulent are discovered. The XGBoost algorithm is a cluster of decision trees, which will be utilized in order to achieve the best outcomes. This algorithm has recently taken control of the ML industry. Comparing this approach to other ML algorithms reveals that it is faster and more accurate.**

Keywords— Fraud detection, Machine learning, Random Forest, Gradient Boosting algorithms, classification, Data pre-processing, Prediction.

I. INTRODUCTION

The world is rapidly moving toward becoming cashless. Many surveys and studies have shown a growth in the number of people making purchases online, and it is predicted that this trend will continue in the coming years. Although this may seem like exciting news, on the Fraudulent transactions on the other side are also on the rise [1]. Even with the implementation of numerous security measures, a sizable sum of money continues to be lost as a result of fraudulent transactions. Online fraud transactions happen when someone makes unauthorized personal online purchases using another person's credit card without the cardholder's or the cards issuer's knowledge.

Fraud detection is the process of keeping track on user populations' activities to gauge, identify, or prevent undesirable behaviour including fraud, intrusion, and defaults. The majority of the time, a person who has fallen victim to such scam doesn't realize it until it's too late. Although it is not without challenges, implementing a fraud detection system in the real world is not easy. Samples drawn from the real world demonstrate how quickly automated tools analyse the massive volume of payment requests to choose which transactions to approve. Machine learning techniques are used to examine every permitted transaction and to identify any suspect activity. Before speaking with the specialists who investigate these accusations, cardholders must decide if a transaction was valid or fraudulent. [2]

To steadily improve the accuracy of fraud detection, the automated system that develops and improves the algorithm integrates ideas from the detectives. In this project, what are the with the aid of various tools, attempted to construct a Web App for the detection of these kinds of scams using machine learning.

II. LITERATURE REVIEW

To understand and analyse, this study has examined several existing research works and concluded that several drawbacks must be taken care of for achieving better performance and accuracy.

In the paper [1] Virjanand, Rajkishan Bharti, Shubham Chauhan, Suraj Pratap This review paper lists the various techniques for detecting fraud in online transactions. It offers an understanding of various research papers in the area of detecting fraud in online transactions that can be used to efficiently address the problems posed by fraud detection and prevention. Future research may use different input and output consideration configurations with machine learning algorithms to identify fraudulent online transactions. [2] Fraudulent transactions should be able to be recognized by a highly accurate and effective fraud detection system. The majority of transactions are genuine, and there are only a small number of fraudulent ones, according to many readily available databases. Researchers face a significant problem in creating an effective fraud detection system that can accurately identify fraudulent behaviour while generating minimal false positives. In our study, We use a labelled dataset made consisting of payment transaction data to perform several binary classification approaches, including Logistic Regression, Linear SVM, and SVM with RBF kernel. Our objective is to create binary classifiers that can distinguish between honest and dishonest transactions. We compare the effectiveness of various techniques for spotting fraudulent transactions. In this paper [3] The proposed system uses a combination of SMOTE technique followed by the XGBoost classification algorithm to classify fraud activities. This is a technique to increase the number of cases in your dataset in a balanced way. In our study, in order to use a number of binary classification algorithms, including Logistic Regression, Linear SVM, and SVM with RBF kernel, we use a labelled dataset made up of payment transaction records. The development of binary classifiers that can distinguish among fraudulent and legitimate transactions is our goal. We compare how well different methods detect fraudulent transactions. This study [4] Balanced delivery strategies are the most common. The most widely used proposed solutions can be grouped into three categories: data layer, algorithm, and synthesis solutions. Reducing the preprocessing phase as a resample to apply data-level solutions can have a negative impact on class-imbalance. The goal of the algorithmic level solution is to create new algorithms or change the existing ones' learning biases in the minority class. This paper [5] They started by cleansing the data in order to identify any irregularities. Also, the minority class was oversampled using the SMOTE (Synthetic Minority Oversampling Method) to address the label uneven distribution problem. The multimillion-dollar fraud industry is continually expanding. Many financial institutions and insurance companies invested millions of dollars in the creation of transaction detection systems in order to stop high risk transactions. Using the IEEE-CIS dataset, we develop an XGBoost-based model to detect consumer transaction fraud. in this study [6]. In section II, we discuss certain data mining approaches such feature engineering, feature selection, and data cleansing. In part III, we present a variety of algorithms for machine learning, including XGBoost, random forests, logistic regression, and support vector machines. The four models' effectiveness is shown through experiments in the fourth portion of the paper. Based on the findings, We can get the conclusion that, in terms of accuracy score and Accuracy Boost score, Xgboost performs better than other machine learning models.

III. FRAUD DETECTION

The system as it currently stands is basically expanded by the online fraud transaction detection system. The algorithms will be developed using this framework to process the dataset and deliver the desired output. This approach will be helpful in the long run since it offers a productive way to develop a safe transaction system for analyzing and identifying fraudulent transactions. The XGBoost technique, which is extensively used in open-source software, successfully implements the gradient boosted trees method. In order to more accurately forecast a target variable, the supervised learning process known as gradient boosting combines the predictions of a series of weaker, simpler models. [3] By making a sizable dataset available for model training, this accuracy can be improved even more. This application's reach is extremely broad. This strategy is utilized to spot fraud-related in a dataset that is highly useful to a variety of industries, including banking, insurance, e-commerce, money transfers, and bill payment. Indeed, doing so will help to improve security.

A. Machine Learning:

The goal of the portion machine learning (ML), an application of artificial intelligence (AI) is to create frameworks that Study the information they consume and become more effective. Better decision-making is the main application of machine learning. Systems can perform better and learn without explicit design thanks to machine learning. Machine learning approaches can be broadly divided into three categories:

a. Supervised learning

A form of learning where the knowledge and the desired result are both provided is referred to as "supervised learning algorithms." Marked data is used in controlled AI to teach robots to comprehend and recognize connections between various data sources and outcomes. The majority of machine learning used by frameworks globally is directed. The relationship between the information variable (x) and the result variable (y) is established through the use of a calculation.

b. Unsupervised Learning

The methodology needs useful indexes and is unable to forecast how most issues that arise during solo learning will turn out. In simple terms, the ML point is buried when the AI framework is put to use. Even though the framework has numerous, large-scale intelligent activities at its disposal, the lack of appropriate data and result calculations makes engagement more challenging. The analysis and solution of problems containing an infinite quantity of data are both possible with unsupervised learning.

c. Re-inforcement Learning

Reinforcement learning instead of supervised learning, which incorporates the correct answer key so that the model is prepared with it, relies on the reinforcement agent to decide how to complete the task. Rewarding the computer when it is successful in learning from prior mistakes is the goal of reinforcement learning. The expert travels alone and observes the world. An expert in learning support works to boost performance by securing the finest incentives.

IV. PROPOSED WORK

A. Dataset:

Acquired a data set from Kaggle that shows whether fraud has happened in a transaction. The dataset offers information on Transactions, which has a total of 10 attributes and over 636263 instances. The dataset will include the following features:

Table. 1. FRAUD DETECTION DATASET

Feature Name	Information
Step_1	Represent unit of time
Type_1	Type
Amount_1	Amount
Name Orig	Starting transaction
Old Balance Org	Before transaction
New Balance Orig	After transaction
name Destination	Recipient of the transaction
Old balance Destination	Initial balance
New Balance Destination	After the transaction
Isfraud_1	Fraud transaction in data

Dataset does not contain records having values that are mismatched, missing values, or noisy data. Comma-Separated Value (CSV) format is used for the dataset.

B. Design Methodology

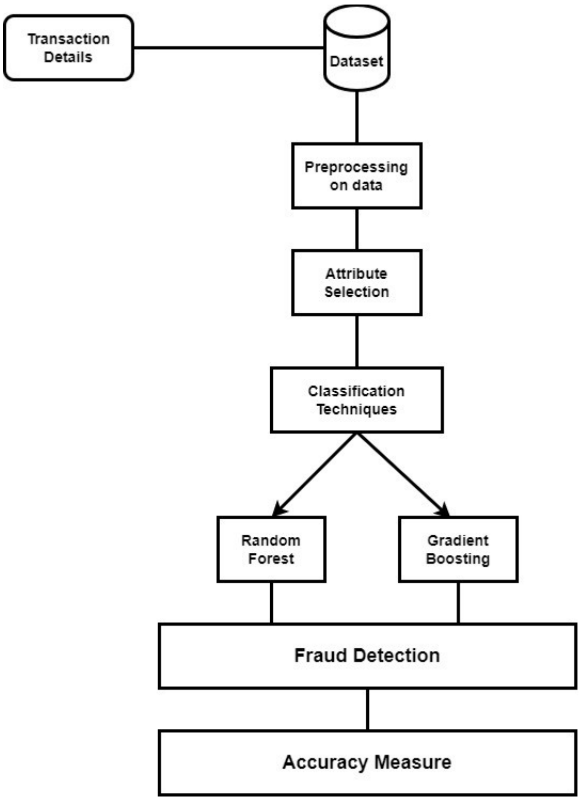

Fig. 1. Architecture of the fraud detection

Fig. 1 shows the proposed work's architecture. Fraud Detection in Online Payments. We collect the dataset from Kaggle, the dataset contains transaction details. Important attributes are extracted from the collected dataset then apply preprocessing on data. Now, apply classification techniques such as random forest and gradient boosting. After implementation of the both algorithms find the fraud detection and its accuracy.

C. Data Preprocessing

In the pre-processing stage, various techniques are applied. Before implementing the model, firstly converted all the string value data type features to float values using python Panda's library built-in functions. Out of a total of 10 features, 3 features are of string values. These features have been converted to numeric values. After that removed all the null values present over the dataset by using python Pandas dropna() function. Feature scaling is also done to standardize the values to a fixed range. Finally iloc[] method of Pandas library is used to extract independent and dependent variable.

D. Description of Algorithm:

The model makes an assumption or guess based on the preliminary data provided, and it will choose a category or classification for the data. Three different types of ML models are available from machine learning: regression, binary classification, and multiclass classification. The kind of model you should employ will depend on the type of target you wish to predict.

The following are the algorithms used to build models for the fraud detection**:**

1) Random Forest

In directed AI (ML), the irregular woodland agreement strategy is used to address relapse and grouping concerns. Every irregular forest is made up of a variety of desirable trees that work together to form a single conjecture. Even if certain trees in the forest lack information, erratic computations done in the woods can nevertheless produce accurate estimates. According to measurements, the number of trees in the group will directly increase the accuracy of the result. Several decision trees are created simultaneously with bootstrapping, totaling, and stowing using a group process called as RF classifier. Using various selections of the available attributes as well as multiple subsets of the preparation dataset, bootstrapping represents the equally prepared preparation of numerous distinct choice trees.

Algorithm :

```
fun BESTSPLITRANDOM( $\mathcal{L}_t$, $K$ )
  $\Delta = -\infty$
  Where Δ is the variable
  ∞ is negative infinity
  Draw K random $j_k$ from 1 ..., p
  Where K is the attribute
  For K = 1 ,…., K
    Find the best split $S^*_{jk}$ defined on $X_{jk}$
    if $\Delta i\ (S^*_{jk}, t) > \Delta$
      $\Delta = \Delta i\ (S^*_{jk}, t)$
      $S^* = S^*_{jk}$
    End if
  End for
  Return $S^*$
End fun
```

2) Gradient Boosting Classifier

Gradient boosting is the most often used forward learning ensemble methods in machine learning. It is a useful technique for building predictive regression models. Tasks that require categorization Through the use of decision trees and other weak prediction models, gradient boosting helps us create a predictive model. When a decision tree performs poorly as a learner, gradient-boosted trees is the resulting algorithm's moniker. A class of machine learning methods known as gradient boosting classifiers combines a number of smaller, weake models into a powerful, large model that generates outputs that are highly predicative. The ability of certain types of models to precisely identify datasets makes them popular. Gradient boosting is one popular boosting method. Gradient boosting uses predictions to identify and repair previous errors. Adaboost trains each predictor by changing the weights of the training examples, whereas each predictor in this method is trained using the residual of the predecessor.

Algorithm

```
1.  Initialize $\hat{f}_0$ with constant
2.  For t = 1 to M
        Where M is the variable
3.      Compute the negative gradient $g_t(x)$
4.      Fit a base-learner function $h(x, \theta_t)$
5.      Find the gradient descent step-size $\rho_t$:
            $\rho_t = \arg\min \rho \sum_{i=1}^{N} \psi [y_i, \hat{f}_{t-1}(x_i) + \rho h(x_i, \theta_t)]$
6.      Update the function estimate:
            $\hat{f}_t - \hat{f}_{t-1} + \rho_t h(x, \theta_t)$
7.  End for
```

V. PERFORMANCE EVALUATION

Assessment metrics are employed to gauge how effectively an AI or factual model is doing. Any AI models or calculations used in a project must be evaluated. There are many evaluation tools available to test a model.

1) Confusion Matrix

In any usage scenario, including an order difficulty, a disarray grid provides an examination overview of the intended outcomes and the actual outcomes. For inspecting the model's presentation after preparation on specific preparation facts, the examination rundown is crucial. When attempting to address characterization troubles, disarray frameworks are a commonly used estimation. It can be used to address concerns with multiclass characterization and duplicate organisations. Counts between expected and noticed values are shown in disarray grids. The True Negative outcome denotes the number of events that were accurately identified as being classified negatively. The word "TP" means to True Positive and refers to the quantity of correctly identified positive cases in a manner similar to this. The percentage of cases that were wrongly labelled as negative even though they were positively validated is known as the False Negative Value. The number of genuine negative models that have been incorrectly labelled with certain is termed as false negative value (FP). One of the measurements that is most frequently used in arrangements is probably exactness.

$$\begin{pmatrix} TP & FN \\ FP & TN \end{pmatrix}$$

To assess a model's accuracy, apply the formula below (through a confusion matrix).

$$\text{Accuracy} = \frac{TN + TP}{TN + FP + FN + TP}$$

The number of predictions in a dataset that were actually relevant can be determined via recall. Using the formula below, the recall is calculated based on the number of TP and FP

$$\text{Recall} = \frac{TP}{TP + FN}$$

The percentage of pertinent facts (expectations) that were accurately inferred as evident and were in fact true/valid is known as accuracy. The equation shown below is used to calculate accuracy.

$$Precision = \frac{TP}{TP + FP}$$

Since accuracy and review are mutually exclusive processes, the F1 score disqualifies excessive amounts of either one. Because it is not symmetric between the classes, this estimation depends on whether each class is presented as either positive or negative. These days, a high F1 score indicates a high review and a high accuracy. When resolving problems like uneven grouping, it demonstrates a respectable

Blend of accuracy and memory.

$$\text{F1} = \frac{2}{\frac{1}{\text{precision}} + \frac{1}{\text{recall}}}$$

VI. RESULTS AND OBSERVATIONS

The confusion matrix was used to calculate the accuracy of each classifier. The experiment's findings demonstrate that Random Forest classifiers help achieve an accuracy of 86.68%. An accuracy of 97.91 was supplied by the Gradient Boosting classifier.

Table 2. Using RF Classifier

	Precision	Recall	F1-score	Support
0	1. 00	1. 00	1. 00	101632
1	0. 97	0. 53	0. 68	170
Accuracy			1. 00	101802
Macro Average	0. 98	0. 76	0. 84	101802
Weighted	1. 00	1. 00	1. 00	101802

Average				

Table 3. Using GB Classifier

	Precision	Recall	F1-score	Support
0	1. 00	1. 00	1. 00	101632
1	0. 99	0. 47	0. 64	170
Accuracy			0. 97	101802
Macro Average	0. 99	0. 74	0. 82	101802
Micro	1. 00	1. 00	1. 00	101802
Average				

A. Comparison of Results:

For the supplied dataset, we noticed that the two methods yielded results that were rather comparable.

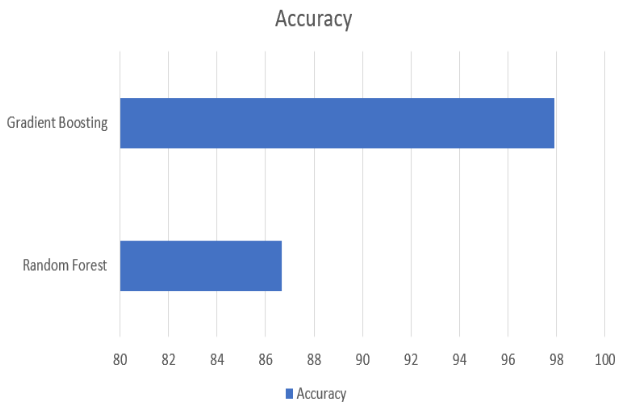

Fig. 2. Comparison of Result

Fig.2 shows that comparison between GB and RF, the Gradient Boosting gives the best accuracy of 97.91% and, Random Forest gave less accuracy of 86.68%.

VII CONCLUSION AND FUTURE WORK

Over the past 20 years, research on fraud detection has used variety of techniques, including manual inspection and consumer end authentication. In this field, machine learning model have also achieved great success. This paper includes empirical research comparing multiple machine learning models on various data sets in order to identify fraudulent transactions. The major goal of this study is to identify the methodologies that are most suited for different dataset types. This study may help professionals and businesses comprehend how various methodologies operate on various sorts of datasets since many businesses nowadays invest in new technologies to enhance their operations According to the proposed research, RF are the best methods for spotting fraud on larger datasets, and they can be combined with GB to deliver a more dependable result. Random Forest and Gradient Boosting ensemble techniques can offer good improvements for smaller datasets. Although gradient boosting is appealing and produces excellent results, they are not effective in dynamic contexts Because they usually evolve over time, fraudulent actions can be challenging to notice. Both new datasets and retraining of the ML algorithms would be required.

Table 4. Accuracy vs Algorithms

Algorithm	Accuracy (%)
Random Forest (RF)	86.68%
Gradient Boosting (GB)	97.91%

The passing of time and are challenging to detect. It would be necessary to gather fresh data sets and retrain ML algorithms. Autoencoders are a perfect option in this situation because they are trained exclusively on ordinary communications. Transactions that are fraudulent are recognised as being unusual. Despite being relatively expensive at first, The labelling of data sets can benefit from autoencoder training. Once there is enough labelled data, it may be used to either build brand-new supervised learning models or retrain ones that already exist. The results show that the Gradient Boosting gives the best accuracy of 97.91% and, Random Forest gave less accuracy of 86.68%.

References

[1] Online Transaction Fraud Detection System1Based on Machine Learning by Virjanand, Rajkishan Bharti, Shubham Chauhan, and Suraj Pratap Singh-2022

[2] Aditya Oza - aditya19 - Fraud Detection using Machine Learning-2021

[3] Kanika and J. Singla, "A Survey of Deep Learning based Online Transactions Fraud Detection Systems," 2020 International Conference on Intelligent Engineering and Management (ICIEM), London, UK, 2020, pp. 130-136, doi: 10.1109/ICIEM48762.2020.9160200.

[4] Yixuan Zhang, Jialiang Tong, Ziyi Wang, Fengqiang Gao-Customer Transaction Fraud Detection Using Machine Learning-2020

[5] San Jose State University Researchers T. Singh, F. Di Troia, C. Vissagio, and Mark Stamp published a study in October 2015 on support vector machines and malware detection.

[6] N. Sharma and V. Ranjan, "Credit Card Fraud Detection: A Hybrid of PSO and K-Means Clustering Unsupervised Approach," 2023 13th International Conference on Cloud Computing, Data Science & Engineering (Confluence), Noida, India, 2023, pp. 445-450

[7] "Fraud Detection Based on Transaction Behavior" by John Richard D. Kho and Larry A. Vea was released in the proceedings of the 2017 IEEE Region 10 Conference (TENCON), which took place in Malaysia from November 5-8, 2017.

[8] S. Nami and M. Shajari, "Cost-sensitive payment card fraud detection based on dynamic random forest and knearest neighbors," Expert Syst. Appl., vol. 110, pp. 381– 392, Nov. 2018.

[9] "Credit Card Fraud Detection Based on Transaction Be haviour -by John Richard D. Kho, Larry A. Vea" published by Proc. of the 2017 IEEE Region 10 Conference (TENCON), Malaysia, November 5-8, 2017.

[10] Kamesh, V., Karthick, M., Kavin, K., Velusamy, M., & Vidhya, R. (2019). Real-Time Fraud Anamaly Detection in E-banking Using Data Mining Algorithm

[11] A. D. Pozzolo, G. Boracchi, O. Caelen, C. Alippi, and G. Bontempi, "Credit Card Fraud Detection: A Realistic Modeling and a Novel Learning Strategy," IEEE Trans. Neural Netw. Learn. Syst., vol. 29, no. 8, pp. 3784–3797, Aug. 2018.

[12] P. Kaur, A. Sharma, J. K. Chahal, T. Sharma and V. K. Sharma, "Analysis on Credit Card Fraud Detection and Prevention using Data Mining and Machine Learning Techniques," 2021 International Conference on Computational Intelligence and Computing Applications (ICCICA), Nagpur, India, 2021, pp. 1-4, doi: 10.1109/ICCICA52458.2021.9697172.

[13] N. Boutaher, A. Elomri, N. Abghour, K. Moussaid and M. Rida, "A Review of Credit Card Fraud Detection Using Machine Learning Techniques," 2020 5th International Conference on Cloud Computing and

Artificial Intelligence: Technologies and Applications (CloudTech), Marrakesh, Morocco, 2020, pp. 1-5.

[14] S. Negi, S. K. Das and R. Bodh, "Credit Card Fraud Detection using Deep and Machine Learning," 2022 International Conference on Applied Artificial Intelligence and Computing (ICAAIC), Salem, India, 2022, pp. 455-461.

[15] A. M. Mubalaike and E. Adali, "Deep Learning Approach for Intelligent Financial Fraud Detection System," 2018 3rd International Conference on Computer Science and Engineering (UBMK), Sarajevo, Bosnia and Herzegovina, 2018, pp. 598-603

[16] R. San Miguel Carrasco and M. -Á. Sicilia-Urbán, "Evaluation of Deep Neural Networks for Reduction of Credit Card Fraud Alerts," in IEEE Access, vol. 8, pp. 186421-186432, 2020,doi: 10.1109/ACCESS.2020.3026222.

[17] S. Sharma, A. Kataria, J. K. Sandhu and K. R. Ramkumar, "Credit Card Fraud Detection using Machine and Deep Learning Techniques," 2022 3rd International Conference for Emerging Technology (INCET), Belgaum, India, 2022, pp. 1-7.