iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 11 **Issue:** VI **Month of publication:** June 2023

DOI: https://doi.org/10.22214/ijraset.2023.54509

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 11 Issue VI Jun 2023- Available at www.ijraset.com

Credit Card Fraud Detection using Ensemble Machine Learning Method - Gradient Boosting Framework

SK.Asiff[1], SK.Muzafar[2]
[1]*Faculty,* [2]*Student, CSE Department, GIST, Nellore*

Abstract: ***The use of credit cards has become commonplace in our daily lives, transforming the way we make cashless payments and making all types of transactions more convenient for buyers. However, credit card fraud has also become increasingly prevalent as the number of users has risen. Criminals may illegally obtain an individual's credit card information and use it for fraudulent activities. This project involved collecting a sample of a publicly available dataset from Kaggle, consisting of 284,807 credit card transactions. Extreme Gradient Boost was used to detect fraudulent transactions, and it was found to be the most effective method, achieving the highest AUC value of 97.67% after thorough analysis.***
Keywords: Fruad Detection, Machine Learning, Ensemble Learning Technique.

I. INTRODUCTION

A Fraud refers to a deliberate act of deception carried out with in the intention of gaining something, particularly money. It is an unfair practice that is being increasingly common. The surge in the use of electronic payment methods such as credit and debit card have a lead to a corresponding increase in credit card fraud. These cards can be used for both online and offline transactions, with the former being more vulnerable to attacks by hackers and cybercriminals. Such fraudulent activities result in substantial financial losses each year. To combat this issue, several algorithms and detection approaches are being developed.

Credit card transactions are ubiquitous, but they present their own set of problems, particularly when it comes to detecting fraud. The acceptance or rejection ofa transaction occurs within a fraction of a second, making it necessary for frauddetection systems to be swift and efficient. Additionally, the sheer volume of similar transactions happening simultaneously makes it challenging to monitor each one individually to determine if it is fraudulent. Therefore, an effective Fraud Detection System must learn the user-specific usage behavior of the card to differentiate between genuine and fraudulent transactions. To accomplish this, existing supervised and unsupervised machine learning techniques can be applied to the data.

A. Types Of Credit Card Frauds

Lost/Stolen credit card fraud: This type of frauds occurs when one lost his/her credit card or dropped off somewhere and gets stolen, and your credit card is used by a thief as their own.as the fraudsters have a physical card, they also have CCV numbers they can make the transaction without any issue. The owner is not able to know about the transaction unless they receive the monthly statement expenses.

Skimming: Skimming is very risky and difficult, but today there are many intelligent fraudsters, who can perform this type of fraud. Somehow the fraudster obtains the card details like number and the PIN of the card .so, whenever the owner performs any transactions or uses the credit card, the fraudster gets a certain percent of balance every time and for each transaction. The amount is not so high that it can come into notice of the owner, it can be a few pennies.

False application fraud: This type of fraud is done by identity theft. The fraudster selects the person who does not have any credit card or have a good credit score, and then try to obtain the details like Date of Birth, Social Security Number via calls or fake emails from Social Security Department, Police Department. After obtaining the details they apply for the credit card using their own identity.

Fraud by data breaches: This fraud is done by breaching the database owners of credit cards. By breaching the database, the fraudster can have access to all the details of the owner and the credit/debit card. So, owner details can be used for fraud application or to do any large number of transactions.

Mail Intercept fraud: when your credit/debit card is lost or expired, you order the new one. The new credit/debit card is sent via mails, while you are waiting for the mail even before the owner receives the mails, the thieves take away the mail and use it, till the cardholder gets to know about the mail, the credit card is already stolen. These are the not only ways to fraudsters uses for credit card frauds. Such frauds result I n a high number of losses to the banks, its reputation and customers.

The objective of this paper is to evaluate an dataset with the help of machine learning model and to determine which one of those is the best suited model for detecting credit card frauds.

II. LITERATURE REVIEW

These papers focus on credit card fraud detection using various techniques and methodologies. Here's a summary of each paper

"Automatic credit card fraud detection based on non-linear signal processing" (2020): This paper proposes a method for automatic fraud detection in credit card transactions using non-linear signal processing. It involves feature extraction, training and classification, decision fusion, and result presentation. The method utilizes discriminant-based classifiers and a non-Gaussian mixture classification method to distinguish between legitimate and fraudulent transactions.

"Generation and Interpretation of Temporal Decision Rules" (2019): The paper addresses the problem of understanding a system that produces temporally ordered observations. It presents a solution based on generating and interpreting a set of temporal decision rules. These rules can predict or retrodict the value of a decision attribute using condition attributes observed at different times. The authors demonstrate the effectiveness of their method through experiments with synthetic and real temporal data.

"Surrogate techniques for testing fraud detection algorithms in credit card operations" (2018): This paper tackles the issue of limited access to real credit card data for testing fraud detection algorithms. The authors propose using surrogate techniques to generate synthetic credit card data that closely resemble the original data in terms of statistical properties. They test the performance of fraud detection algorithms using a mix of real and surrogate data, considering receiver operating characteristic (ROC) curves.

"Data mining application in credit card fraud detection system" (2018): With the increasing use of credit cards for online transactions, credit card fraud has become a significant concern. This paper explores the application of data mining techniques, particularly unsupervised anomaly detection algorithms, for real-time fraud detection in internet transactions. The system classifies transactions as legitimate, suspicious fraud, or illegitimate based on anomaly detection. The paper also discusses different types of fraudsters and techniques used in online credit card fraud.

"A Comprehensive survey of Datamining-based fraud detection research" (2017): This survey paper provides an overview of fraud detection research using data mining techniques. It compares various models based on artificial intelligence, such as Naïve Bayesian classifier, Bayesian Networks, and clustering models. The paper evaluates the accuracy of these models and offers recommendations for improving them.

"Credit card fraud detection using Machine Learning Models and collating Machine Learning Models" (2017): This paper focuses on credit card fraud detection using machine learning models. It highlights the challenges posed by the changing profiles of normal and fraudulent behavior and highly skewed datasets. The authors evaluate the performance of decision tree, Random Forest, and logistic regression models on raw and preprocessed data, considering accuracy, sensitivity, and precision as evaluation metrics.

Overall, these papers contribute to the ongoing research and development of fraud detection techniques in the context of credit card transactions, utilizing approaches such as non-linear signal processing, temporal decision rules, surrogate data generation, data mining, and machine learning models.

III. METHODOLOGY

If there are enough trees in the forest. Gradient Boosting is a popular machine learning algorithm used for both regression and classification problems. It is an ensemble method that combines multiple weak learners (typically decision trees) to create a strong learner. The basic idea behind gradient boosting is to iteratively add weak learners to the model, with each new learner learning from the errors of its predecessors.

The pseudocode of how Gradient Boosting works is given below:

Split the data into training and validation sets. The training set will be used to train the model, while the validation set will be used to evaluate the performance of the model.

Choose a weak learner. In most cases, decision trees are used as the weak learners.

Train the weak learner on the training data.

ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 11 Issue VI Jun 2023- Available at www.ijraset.com

Evaluate the performance of the weak learner on the validation set.
Calculate the residuals (i.e., the differences between the predicted values and the actual values) of the weak learner on the validation set.
Use the residuals as the target variable for the next weak learner.
Train a new weak learner on the residuals.
Combine the new weak learner with the previous learners to create a stronger learner.
Repeat steps 4-8 until the desired level of accuracy is achieved.
Use the final model to make predictions on new data.
Some of the important hyperparameters that can be tuned in Gradient Boosting are the number of weak learners, the learning rate, the maximum depth of the decision trees, and the minimum number of samples required to split an internal node.
Gradient Boosting is a powerful algorithm that can achieve high accuracy on a wide range of machine learning tasks. However, it can be prone to overfitting, so it's important to carefully tune the hyperparameters and use regularization techniques to prevent overfitting.In this project we are using python XGBoost algorithm to detect fraud transaction from credit card dataset, we downloaded this dataset from 'Kaggle's' web site from below URL
Dataset URL: https://www.kaggle.com/mlg-ulb/creditcardfraud
To provide privacy to user's transaction data Kaggle's peoples have converted transaction data to numerical format using PCA Algorithm. Below are some example from dataset
All variables in the dataset are numerical. The data has been transformed using PCA transformation(s) due to privacy reasons. The two features that haven't been changed are Time and Amount.
Time contains the seconds elapsed between each transaction and the first transaction in the dataset."Time","V1","V2",V28","Amount","Class"
Above bold names are the column names of this dataset and others decimal values are the content of dataset and in above 3 rows last column contains class label where 0 means transaction values are normal and 1 means contains fraud values.
Using above 'CreditCardFraud.csv' file we will train Random Forest algorithm and XGBoost algorithm, then we will upload test data file and this test data will be applied on Random Forest and XGBoost to train model to predict whether test data contains normal or fraud transaction signatures. When we upload test data then it will contain only transaction data no class label will be there application will predict and give the result. See below test data file

A. Evaluation Criteria

Area under the curve(AUC) :The area under the curve (AUC) is the measure of the ability of a classifier to distinguish between classes and is used as a summary of the ROC curve. The higher the AUC, the better the performance of the model at distinguishing between the positive and negative classes.so, we consider this AUC curve to find the efficient algorithm among all supervised Machine Learning algorithm The area under the curve (AUC) is a metric used to evaluate the performance of a binary classification model. It measures the ability of the model to distinguish between positive and negative classes. The AUC is calculated by plotting the true positive rate (TPR) against the false positive rate (FPR) at different classification thresholds. TPR is the ratio of correctly classified positive instances to the total number of positive instances, while FPR is the ratio of incorrectly classified negative instances to the total number of negative instances.
The AUC ranges from 0 to 1, where 0 indicates that the model is making random predictions, and 1 indicates that the model is making perfect predictions.
XGBoost (eXtreme Gradient Boosting) is a popular open-source machine learning algorithm that belongs to the family of boosting algorithms.
It is used for supervised learning problems, including regression, classification, and ranking tasks. XGBoost is designed to handle large-scale, structured and unstructured data in a distributed environment.
The key idea behind XGBoost is to iteratively add new models to an ensemble, with each new model correcting the errors made by the previous models.
This process is called boosting. XGBoost uses gradient boosting, which involves minimizing a loss function by adding models that make the largest contribution to reducing the gradient of the loss function.

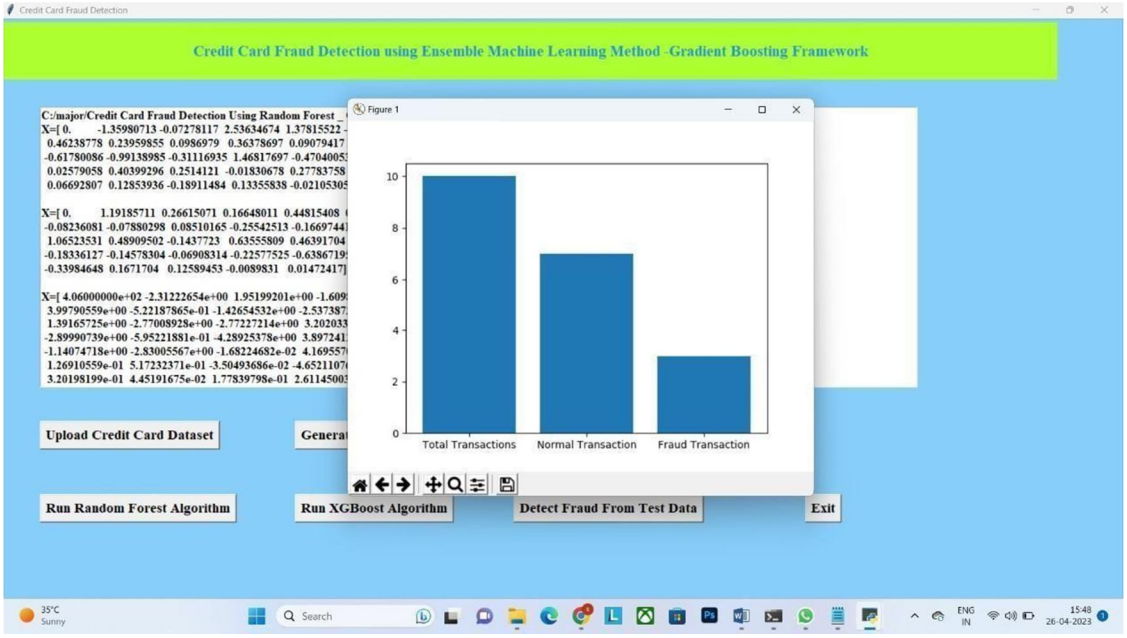


Fig 1. Output Screen

In above screen we can see Random Forest generate 88.43% percent accuracy and XGBoost generate 97.66% while building model on train and test data. In above screen I am uploading test dataset and after uploading test data will get below prediction details Comparison table for finding the best approach among algorithms used: The below table consists of evaluation metric values for the algorithms used, based on which we are finding the best approach.

Table :1 Comparison table for finding the best approach among algorithms used

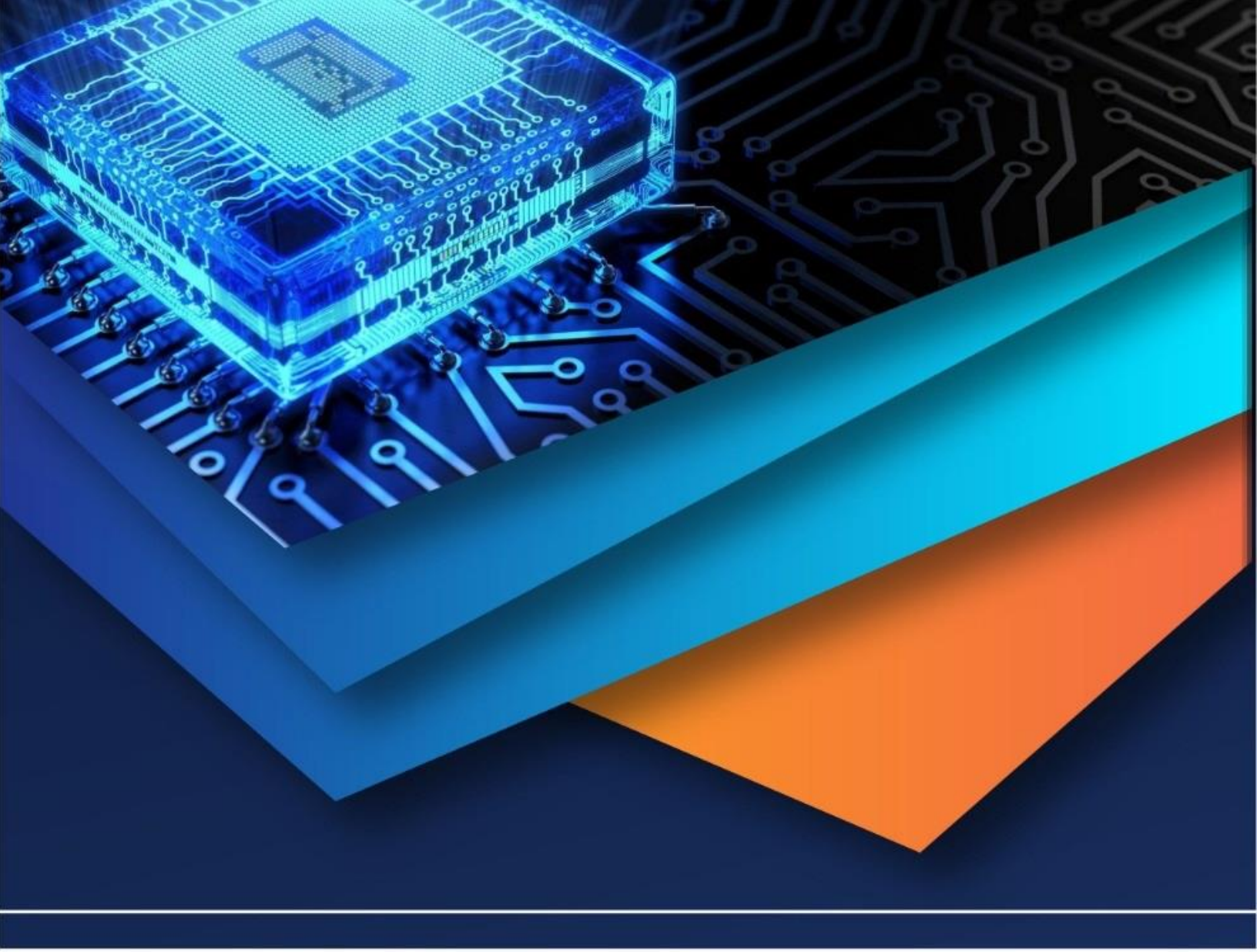
	Random forest	XGBoost
AUC	88.43	97.67

IV. CONCLUSIONS

In this project, we used a dataset to check the suitability of machine learning model to predict the chances of occurrence of a fraudulent transaction. We used AUC as the deciding parameters to come to a particular conclusion. Accuracy as a parameter was not used as it is not sensitive to imbalanced data and does not give a conclusive answer. We analyzed the XGBoost algorithm to detect these fraudulent transactions. XGBoost is found to be the best approach to detect the credit card frauds as it got the highest AUC value i.e., about 97.67% on thorough examination.

REFERENCES

[1] https://inoxoft.com/blog/why-use-python-for-machine-learning/
[2] https://www.analyticsvidhya.com/blog/2021/12/evaluation-of-classification- model/
[3] https://www.shiksha.com/online-courses/articles/roc-auc-vs-accuracy/
[4] Deepak Pawar, SwapnilRabse, Sameer Paradkar, NainaKaushi, “Detection of Fraud in Online Credit Card Transactions”, International Journal of Technical Research and Applications e-ISSN: 2320-8163.
[5] A. Dal Pozzolo, G. Boracchi, O. Caelen, C. Alippi and G. Bontempi,” Credit Card Fraud Detection: A Realistic Modeling and a Novel Learning Strategy,” in IEEE Transactions on Neural Networks and Learning Systems, vol. 29, no. 8, pp. 3784- 3797, Aug. 2018.
[6] S. Xuan, G. Liu, Z. Li, L. Zheng, S. Wang and C. Jiang, ’’ Random forest for credit card fraud detection,” 2018 IEEE 15th International Conference on Networking, Sensing and Control(ICNSC), Zhuhai, 2018, pp. 1-6.



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY