Finding Fake Transactions Using Semi Supervised Models

Ms.S. Revathi AP/CSE, M. Akash, K. Barani, S. Kishore Kumar, G. Manoj
Dept of CSE, Erode Sengunthar Engineering College,
Erode, Tamil Nadu.

Abstract

Due to the growth of e-commerce and online transactions, credit card fraud has become the most prevalent problem in the modern world. It involves the theft of a card and its information for unauthorized use or personal gain. As a result, we are seeing a lot of credit card issues. Fake transaction is discovered after the crime has been committed, and it is discovered following the cardholder's complaint. Three algorithms are compared: decision trees,logistic regression and XG Boost. The model we utilize is a semi-supervised model. The primary topic of the paper is machine learning algorithms. Decision Trees, XG Boost, and Logistic Regression. By comparing accuracy, precision, recall, and F1-score, the outputs of the three algorithms are contrasted. The ROC curve is generated from the confusion matrix. The algorithm that is best at detecting fraudulent transactions should have high accuracy, precision, recall, and F1-score.

Keywords—credit card, fraudulent activities, decision trees, Logistic Regression boost, ROC curve.

I. INTRODUCTION

The prevalence of fake transactions has increased substantially over the years, as it has become an ever-growing threat in businesses, government agencies, the financial sector, and many other institutions. The rise of the internet has been a major contributing factor to this surge in credit card fraud, with both online and offline transactions affected. Despite efforts to use data mining methods for fake identification, the results are not always accurate. To avoid these losses, effective algorithms must be employed to detect fake transactions, which is a highly promising strategy for mitigating credit card crime. When internet usage increases, financial businesses provide people with credit cards, which might be used fraudulently if someone uses another person's card without their permission. It is possible to detect credit card fraud and distinguish between new transactions that are fraudulent or legitimate, even if someone has stolen the PIN or account details to conduct unauthorized transactions without the physical card.

If someone else uses your card in your absence without your permission, it is called fake transaction. Even without taking the actual card, fraudsters can make any illegal transactions by getting the PIN or card details from card. We can identify whether the new transactions are fraudulent or real with the help of fake transaction detection.

Fraudsters often target credit cards due to their potential to make money quickly with relatively low risk; a fraudulent transaction may take weeks to be detected. The card itself may be used as a source for the fraud, such as a

credit card or debit card, and the perpetrator's motivation may be to gain products without paying money or to obtain an unlawful fund.

Given the high frequency of internet use, there is a heightened risk of fake transactions as more people are opting to shop online rather than going to a store. Although important to find the effective strategy for preventing such frauds. Various alternative methods are being utilized to combat fake transactions. After conducting a literature review, it shows that machine learning can offer a magnificent number of additional approaches for detecting fake transactions.

II. RELATED WORK

Research on fake transactions has utilized both ML and DL algorithms [1][2][7], as well as methods and techniques for handling imbalanced data [11]. These include classification methods, sampling methods, and

resembling techniques. Frequently used Machine Learning techniques in the detection of fake transactions are SVM, Decision Trees, Logistic Regression, Gradient Boosting, and K-Nearest Neighbour.

In 2019, Yashvi Jain, NamrataTiwari, Shripriya Dubey, and Sarika Jain studied different methods of detecting credit card fraud, comprising Support Vector Machines (SVM), Artificial Neural Networks (ANN), Bayesian Networks, K-Nearest Neighbours (KNN), Fuzzy Logic systems, and Decision Trees. The authors of the paper noted that K-Nearest Neighbor, Decision Trees, and the SVM deliver a medium level of accuracy, while Fuzzy Logic and Logistic Regression algorithms provide the lowest accuracy. Neural Networks, Naive Bayes, Fuzzy Systems, and KNN has high detection rate. Logistic Regression, SVM, and Decision Trees all show a high detection rate at a medium level. Ann and the Naive Bayesian Networks demonstrate better overall results, though they are costlier to train. For example, KNN and SVM offer excellent results with small datasets, and Logistic Regression and Fuzzy Logic systems provide good accuracy.

XG Boost, a commonly employed Machine Learning technique, is principally designed for binary classification. It is regularly used to boost the efficiency of the trees and can be applied to both regression and classification tasks.It has find to be especially useful in fake cases, as it might be used to classify transactions as either fraud or non-fraud with a higher accuracy. When compared to Logistic Regression, which is also used for classification, XG Boost and Logistic Regression have the same accuracy; however, XG Boost is faster, making it the preferred algorithm for detecting credit card fraud. Other popular classification algorithms include Naive Bayes, which based on Baye's theorem, and J48, which is an extension of the ID3 algorithm and it is used to generate decision trees.

The SMOTE method was employed to address the class imbalance issue, while the WOA approach was used to optimize the accuracy of the re-sampled synthesized data. The system's convergence speed, reliability, and efficiency were also enhanced by the algorithm.

Navanushu Khare and Saad Yunus Sait evaluated the performance of random forests, SVM and logistic regression on a high valid dataset in 2018 [5]. It was concluded that Random Forest had the highest accuracy among the algorithms and was the best choice for fake transaction detection. However, the SVM algorithm was found to have a data imbalance problem and therefore was not suitable for detecting the fake transactions.

III. PROPOSED WORK

The objective of this paper is to use algorithms like the Random Forest and XG Boost to classify the dataset's transactions, which include both fake and non fake ones.The best algorithm for detecting fake transactions will be determined by measuring the performance of both algorithms. The stages of data division, model training, model deployment, and evaluation criteria form the credit fraud detection problem process flow as shown in [Figure.1].

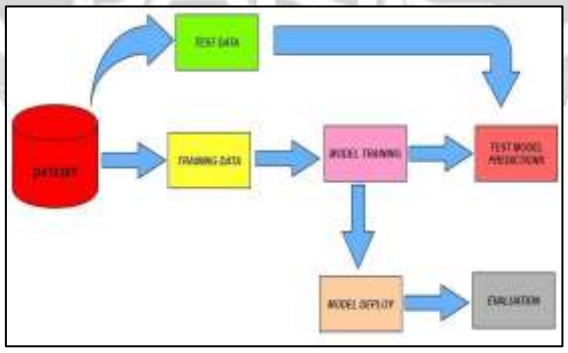

Figure.1 Process Flow

The comparison of the accuracy, recall and precision and F1-score of the Random Forest and XG Boost models created from the Kaggle credit card dataset is made after the data undergoes pre-processing and is separated into training and testing sets. A comprehensive architectural diagram [Figure 2] for the fake transaction detection system outlines the use of this comparison to create a more realistic depiction of fake transactions.

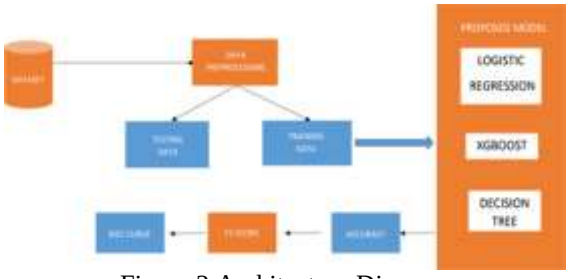

Figure.2 Architecture Diagram

A. Decision Tree Algorithm

Decision trees are a popular supervised learning algorithm used for both classification and regression. They are usually applied to categorization problems. Random Forest is a technique that applies multiple decision trees on sample data and then averages the results to create predictions, similar to how a forest is composed of its trees. As an ensemble method, Random Forest reduces over-fitting more than single decision trees.

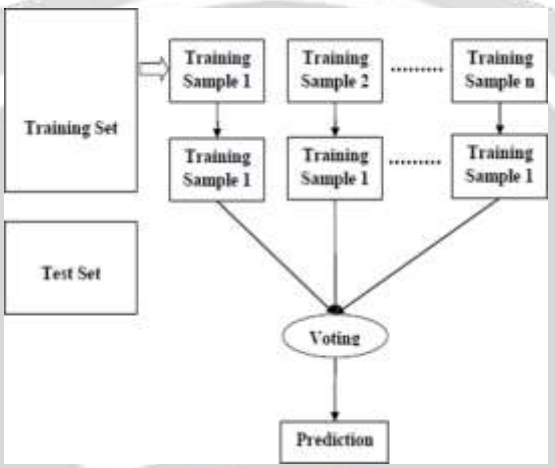

Figure.3 Decision Trees

Steps for Decision Trees

1. Randomly select some sample data from the Kaggle credit card fraud dataset that has been trained..
2. Construct Decision Trees from the randomly generated sample data to classify cases into fake and non-fake.
3. Splitting the nodes to form Decision Trees, with the node having the most Information gain chosen as root node, allows the classification of fake and non-fake cases.
4. After the voting has been completed, the Decision Tree may yield 0 as the result, indicating that these are non-fraudulent cases.
5. We have finally determined the accuracy, precision, recall, and F1-score for both fake and non-fake cases.

B. XGboost Algorithm

Boosting is one of the ensemble strategies which may be used to construct stronger classifiers starting from weaker ones. This is done by creating a model with the training data and subsequently utilizing it to produce the second model, thus correcting the mistakes of the first. XG boost, one of the most successful boosting algorithms developed for binary classification, is a process that is repeated until either the maximum number of models are added or the whole training dataset is predicted correctly.

XG boost, also referred to as adaptive boosting, is most effective when teaching weak learners. This boosting algorithm takes multiple weak classifiers and combines them into one strong classifier, as seen in [Figure.4].

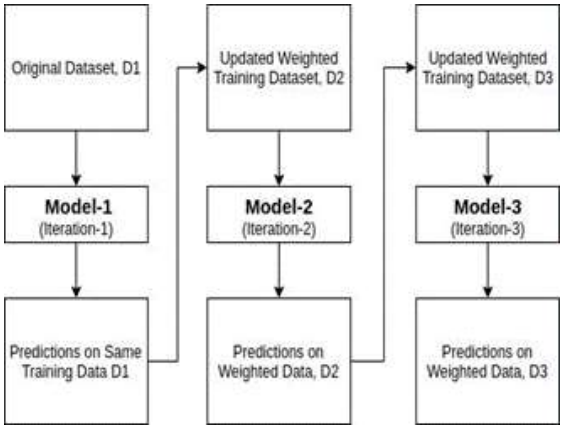

Figure.4 XG Boost workflow

Short decision trees might be used with the XG boost method, which looks at the performance of each tree's node and assigns a weight to it, giving more importance to data that is difficult to predict. Although XG boost can make strong classifiers that work well in simple and complex scenarios, it is sensitive to noise and outliers, which might be a problem.

Steps for XGboost Algorithm

1. Randomly selecting the sample data from Kaggle credit card dataset and using it for training.
2. Construct decision trees sequentially with the randomly generated sample data to classify fake and non-fake cases.
3. Initially, decision trees can be created by splitting the node with the highest information gain, making it the root node, and then classifying fraud and non-fraud cases.
4. Determine the error rate, assess performance, and adjust the weights of the misclassified fake and non-fake transactions.
5. A majority vote is taken, and decision trees algorithm might be used to identify non-fraudulent cases.
6. The output of decision tree may be 1, indicating that it as a fraudulent case.
7. We have calculated the accuracy, precision, recall, and F1-score for both fraud and non-fraud cases.

IV. EVALUATION AND RESULT ANALYSIS

A. Dataset

A European credit card provider provided the dataset for fake transactions, which was taken from Kaggle website. The dataset consists of 284,807 total transactions that was made by cardholders in September of 2013, including two-day transactions, and 492 of these transactions were identified as fraudulent. It means that only 0.172% of all transactions are fraudulent. In order to maintain secrecy, the PCA transformation is used to convert the dataset, which contains input variables, into numerical values. However, the 'Time' and 'Amount' features are not PCA transformable. The feature labeled 'Time' displays the duration in seconds between the first transaction and the current transaction, whilst the 'Amount' feature specifies the monetary sum that has been exchanged. Lastly, the 'Class' feature is used to define whether a transaction is fake or legitimate, with 1 indicating a fake transaction and 0 indicating a non-fake transaction.

B. Evaluation Criteria

We must analyze accuracy, precision, recall, and F1-score parameters to compare different methods. The confusion matrix, composed of a 2*2 matrix, provides four outputs: true positive rate (TPR), true negative rate (TNR), false positive rate (FPR), and false negative rate (FNR). Sensitivity, specificity, accuracy, and error rate can then be calculated from the confusion matrix. Ultimately, the best method to identify credit card fraud can be determined.

The output of the confusion matrix is

1. The TPR can be defined as the fake transactions correctly identified by the system.
2. The system correctly classified the valid transactions as True Negative Rate.
3. A FPR can be defined as the number of legal transactions are classified as fake.
4. The FNR is the proportion of fake transactions that have been classified as legitimate.

C. Results Analysis

Using a different algorithm, the outputs from the dataset are applied in the random forest model may differ from the confusion matrix and ROC curve we obtained.

	precision	recall	f1-score	support
0	1.00	1.00	1.00	93825
1	0.95	0.77	0.85	162
accuracy			1.00	93987
macro avg	0.97	0.89	0.93	93987
weighted avg	1.00	1.00	1.00	93987

Figure.5 Output of Decision Tree

The precision, recall, and F1-score for the non-fraud cases are the same, but they differ for the fake cases, as explained in [Figure.5] of the evaluation criteria.

```
Confusion Matrix on train data
[[190490      0]
 [     0    330]]

Confusion Matrix on test data
[[93818    37]
 [    7   125]]
```

Figure.6 Confusion Matrix of Decision Tree

Confusion matrix declares that there were 190490 true positives and 0 false positives for the train data, and 330 false negatives. The test results revealed 93818 true positives, 37 false positives, 7 true negatives, and 125 false negatives.

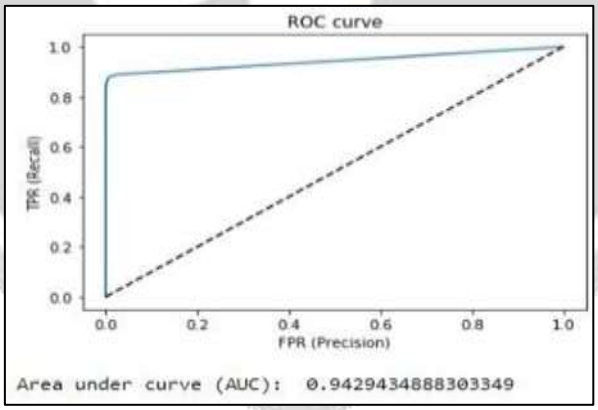

Figure.7 ROC curve of Decision Tree

The results obtained from the XG boost algorithm are comparable to the Random Forest Algorithm.

```
Accuracy = 0.9990743400683073
              precision    recall  f1-score   support

           0  0.99938202 0.99969091 0.99953644     93825
           1  0.78195489 0.64197531 0.70508475       162
```

Figure.8 Output for XG boost

Although the F1-score, recall, and precision varied very less in non-fraud instances and significantly in fraud situations, as demonstrated by the assessment criteria in [Figure.9].

```
Confusion Matrix on train data
[[190464    120]
 [    26    210]]
Confusion Matrix on test data
[[93811    65]
 [   14    97]]
```

Figure.9 Confusion Matrix of XG boost

The train data had 190464 true positives and 120 false positives, 26 true negatives and 201 false negatives, regarding confusion matrix [Figure.9]. The test results had 93811 true positives, 65 false positives, 14 true negatives, and 97 false negatives.

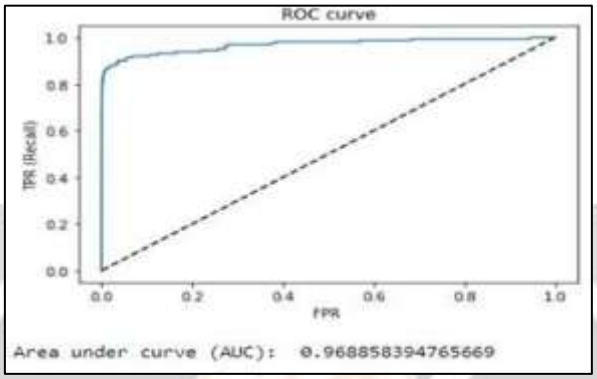

Figure.10 ROC curve for XG boost

On comparing XG boost and Random forest methods, [Figure.10] shows that, although the accuracy of those two techniques are same, they have different recall, precision and F1 scores. The XG Boost algorithms have the highest accuracy, recall, and F1-score.

Summary

```
results.sort_values(by="ROC", ascending=False)
```

	Model Name	Accuracy	F1-score	ROC
1	XG Boost	0.999456	0.822857	0.978537
0	Logistic Regression	0.998982	0.641975	0.977696
2	Decision Tree	0.998771	0.615385	0.921750

Figure.11 Results of Algorithms

V. CONCLUSION

We find that XGboost is better at detecting fake transaction than the Decision Tree algorithm, as it has better precision, recall, and F1-score in comparison. Despite the various fake transaction techniques available, its not possible to guarantee that this particular algorithm will identify all cases of fraud. Our analysis has concluded that accuracy is same for XG boost, Logistic regression and the Decision Trees algorithms.

VI. FUTURE SCOPE

From the above analysis, it is clear that many machine learning techniques are used to find the fake transactions, but we can observe that the results would get better. So, we would like to implement deep learning algorithms to detect fake transactions accurately.

REFERENCES

[1] Adi Saputra, and Suharjito, (2019) “Fraud Detection Using Machine Learning” in E-Commerce,

International Journal of Advanced Computer Science and Applications, Vol.10, Iss.9.

[2] Changjun Jiang, Jiahui Song, Guanjun Liu, (2018) "Credit Card Fraud Detection: A Novel Approach Using Aggregation Strategy and Feedback Mechanism", IEEE Internet of Things Journal, Vol. 5, pp. 3637 – 3647, DOI. 10.1109/JIOT.2018.2816007.

[3] Tanouz, D., Raja Subramanian, D., Eswar, G. and Parameswara Reddy, V. (2021) "Credit Card Fraud Detection Using Deep Learning", IEEE International Conference On Intelligent Computing And Control System,DOI:10.1109/ICICCS51141.2021.9432308.

[4] Ruttala Sailusha, Gnaneswar, V. Ramesh, R. and Ramakoteswara Rao, G. (2020) "Credit Card Fraud Detection Using Machine Learning", IEEE International Conference on Intelligent Computing and ControlSystems,DOI.10.1109/ICICCS48265.2020.9121114.

[5] Filippov, V, Mukhanov, L. and Shchukin, B. (2018) "Credit Card Fraud Detection System", IEEE International Conference on Cybernetic Intelligent Systems, DOI: 10.1109/UKRICIS.2008.4798919.

[6] Sahil Negi, Sudipta Kumar Das and Rigzen Bodh, (2022) "Credit Card Fraud Detection Using Deep and Machine Learning", IEEE International Conference on Applied Artificial Intelligence and Computing, DOI: 10.1109/ICAAIC53929.2022.9792941.

[7] Olawale Adepoju, Julius Wosowei, Shiwani lawte, and Hemaint Jaiman, (2019)" Comparative Evaluation of Credit Card Fraud Detection Using Machine Learning Techniques", IEEE Global Conference for Advancement in Technology,DOI:10.1109/GCAT47503.2019.8978372.

[8] John O. Awoyemi, Adebayo O. Adetunmbi , and Samuel A. Oluwadare, (2022) "Credit Card Fraud Detection Using Machine Learning Techniques: A Comparative Analysis", IEEE International Conference On Computing Networking And Informatics,DOI:10.1109/ICCNI.2017.8123782.

[9] Vaishnavi Nath Dornadulaa and Geetha, S. (2019) "Credit Card Fraud Detection Using Machine Learning Algorithms", International Conference on Recent Trends In Advanced Computing (ICRTAC).

[10] Francisca Nonyelum Ogwueleka, (2020) "Data Mining Application in Credit Card Fraud Detection System", Journal of Engineering Science and Technology, Vol. 6, No. 3 (2011) 311 – 322.