

CREDIT CARD FRAUD DETECTION USING RANDOM FOREST ALGORITHM

¹Prof. Teena Varma, ²Mahesh Poojari, ²Jobin Joseph, ²Ainsley Cardozo
¹Professor, ²Final year Student, Department of Computer Engineering
Xavier Institute of Engineering, Mumbai-400016
teena.v@xavier.ac.in

Received 10 June 2021 Received in revised form 18 June 2021 Accepted 19 June 2021
Available Online 25 June 2021

ABSTRACT

Credit card fraud avoidance has been the most popular problem in the developed world. In this case, credit card fraud is identified by fraudulent transactions. Since e-commerce sites are becoming more popular, credit card fraud is becoming more common. When a credit card is stolen it is used for dishonest reasons, a fraudster uses the credit card information for his own purposes, and it is called credit card theft. In order to track online fraud transactions, the new technology employs a variety of methods. To increase the consistency of the proposed scheme, we used a random forest algorithm to find suspicious transactions. It is built on supervise learning algorithm, which classifies the dataset using decision trees. After the dataset has been categorized, a confusion matrix is established. The confusion matrix is used to test the Random Forest Algorithm's accuracy.

Keywords— Credit Card, Fraud Detection, Random Forest, Classification technique, Transactions.

INTRODUCTION

Fraud involving credit cards is on the rise. Both online and offline transactions can result in credit card fraud. Physical cards are used for offline transactions, while virtual cards are required for online transactions when engaging in illegal or fraudulent activity. As a result of these credit card fraud activities, a large number of fraudulent transactions can occur without the knowledge of the perpetrators. To carry out transactions, fraudsters seek confidential information such as credit card numbers, bank account numbers, and other user data. In the case of offline transactions, fraudsters must steal the user's credit card to complete the transaction, while in the case of online transactions, fraudsters must steal the user's identification as well as online data to complete the transaction. As a result, credit card fraud has emerged as a major concern in today's technological environment, with a significant impact on bank transactions. Many fraud transactions are difficult for both the customer and the banking Authority to detect, resulting in the loss of sensitive data.

There are several models for detecting fraud transactions based on transaction behavior, and these approaches can be divided into two different categories: supervised learning and unsupervised learning algorithms. They used methods such as Cluster Analysis, Support Vector Machine, Nave

Bayer's Classification, and others in the current system to find the accuracy of the fraudulent activities. Using the Random Forest Algorithm, the aim of this paper is to detect the accuracy of fraudulent transactions.

RELATED WORK

In 2019 Sahayasakila V, D.Kavya Monisha, Aishwarya, Sikhakolli Venkatavisalakshiswshai Ysaswi have clarified the Twain significant algorithmic strategies which are the Whale Optimization Techniques (WOA) and SMOTE (Manufactured Minority Oversampling Techniques). They were meant to improve the union speed and to address the information lop-sidedness issue. The class irregularity issue is survived utilizing the SMOTE strategy and the WOA procedure. The Destroyed procedure separates every one of the exchanges which are orchestrated is again re-tested to check the information exactness and is enhanced utilizing the WOA procedure. The calculation likewise improves the intermingling speed, unwavering quality, what's more, proficiency of the framework.

In 2018 Navanushu Khare and Saad Yunus Sait have clarified their work on choice trees, Random Forest, SVM, and strategic relapse. They have taken the profoundly slanted dataset and dealt with such kind of dataset. The execution assessment depends on exactness, affectability, explicitness, and exactness. The outcomes show that the exactness for the Logistic

Regression is 97.7%, for Decision Trees is 95.5%, for Random Forest is 98.6%, for SVM classifier is 97.5%. They have reasoned that the Random Forest calculation has the most noteworthy precision among the other calculations and is considered as the best calculation to recognize the misrepresentation. They likewise reasoned that the SVM calculation has an information lop-sidedness issue and doesn't give better outcomes to identify Visa misrepresentation.

PROPOSED WORK

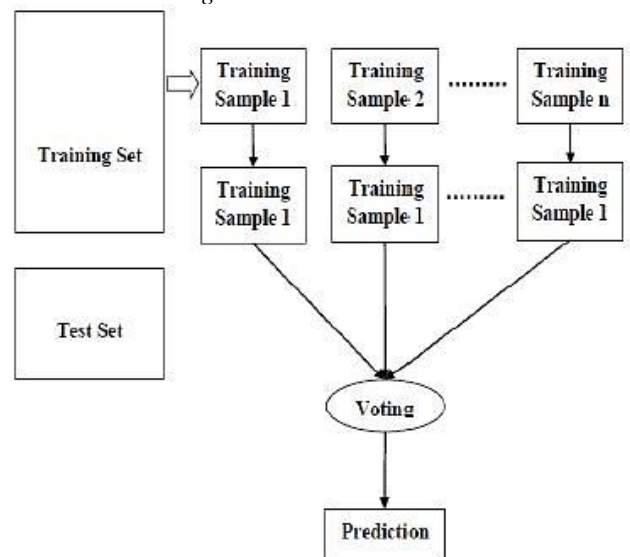
The main aim of this paper is to classify the transactions that have both the fraud and non-fraud transactions in the dataset using algorithm Random Forest algorithm. The Random Forest algorithm [Figure. 5] is one of the widely used supervised learning algorithms. This can be used for both regression and classification purposes. But, this algorithm is mainly used for classification problems. Generally, a forest is made up of trees and similarly, the Random Forest algorithm creates the decision trees on the sample data and gets the prediction from each of the sample data. Then Random Forest algorithm is an ensemble method. This algorithm is better than the single decision trees because it reduces the over-fitting by averaging the result. Random Forest is a directed learning calculation. The "Forest" it constructs, is a group of choice trees, normally prepared with the "stowing" technique. The overall thought of the packing strategy is that a mix of learning models expands the general outcome. Set forth plainly: Random Forest assembles various choice trees and consolidates them to get a more exact and stable expectation.

One major benefit of Random Forest is that it very well may be utilized for both order and relapse issues, which structure most of current AI frameworks. How about we see Random Forest in the grouping since the order is now and again thought to be the structure square of AI.

Random woods has almost similar hyper parameters as a choice tree or a stowing classifier. Luckily, there's no compelling reason to consolidate a choice tree with a sacking classifier since you can without much of a stretch utilize the classifier-class of arbitrary woods. With Random forest, you can likewise manage relapse errands by utilizing the calculation's regressor.

Random Forest adds extra arbitrariness to the model while developing the trees. Rather than looking for the main component while parting a hub, it looks for the best element among an irregular subset of highlights.

Random Forest algorithm



Algorithm Random Forest:

To generate c classifiers:

For i=1 to c do

Randomly select the training data D with replacement to produce Di

 Create a root node N containing Di and cell

 Build Tree (N)

End for Majority Vote

Build Tree (N)

 Randomly select x% of all the possible splitting features in N

 Select the features F that has the highest Information a gain for further splitting

Gain (T, X)=Entropy (T)-Entropy (T,X)

Now to calculate the entropy we use,

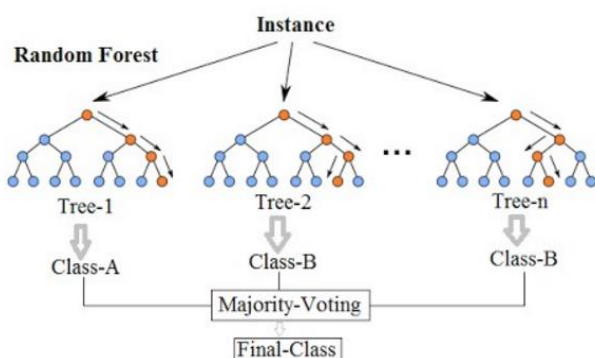
 Create f child nodes for i=1 to f do

 Set contents f N to Di Call Build Tree (Ni)

 End for

End

Random Forest Simplified



IMPLEMENTATION PART:

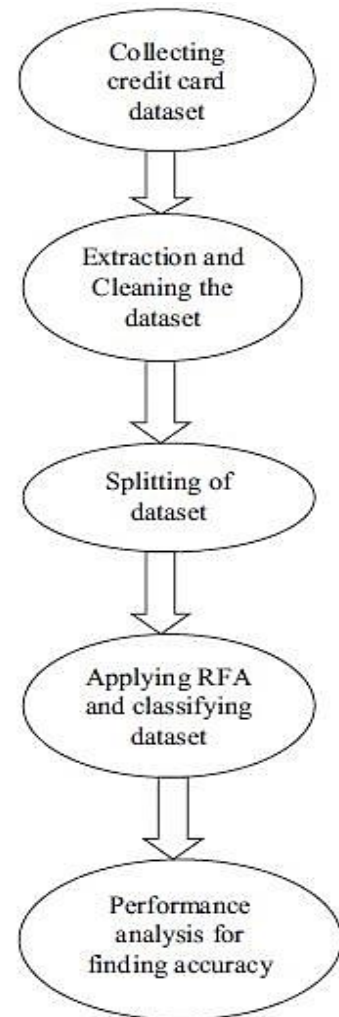
- In this module, we'll gather all of the credit card data and save it to a folder. The dataset would then be subjected to descriptive analysis.
- After reviewing the dataset, we must clean the data in the next phase. Both redundant values and null values in the dataset will be deleted during this cleaning step, and a new dataset will be created.
- The cleaned dataset will be reprocessed in this module, with the dataset being grouped by volume and transaction period.
- The dataset will be split into two parts in this module: qualified dataset and testing dataset. The Random Forest Algorithm is used after the data has been partitioned. Finally, a confusion matrix is obtained after using the Random Forest Algorithm.

Evaluation Now that the resulting data in the form of an uncertainty matrix has been obtained, it can be analyzed using a graphical representation, which provides greater precision.

FURTHER IMPROVEMENT:

- We have observed that the dataset which we are using from Kaggle which known as "CreditCard.csv" is imbalanced i.e. it consist of 284315 non-fraud transactions and 492 fraud transaction.
- So here, we can improve the recall and precision of our model by applying sampling techniques.
- There are 2 types of sampling
 1. Under Sampling
 2. Over Sampling
- In the oversampling technique, samples are repeated, and the dataset size is larger than the original dataset.
- In the under sampling technique, samples are not repeated, and the dataset size is less than the original dataset.
- We are implementing under sampling technique in our model, by choosing the no. of random samples from non-fraud transaction equals to number of fraud transaction in our dataset.
- Concatenate both indices of fraud and non-fraud.
- Extract all features from whole data for under sample indices only.
- Now we have to divide under sampling data to all features & target.
- Now split dataset to train and test datasets as before.

- Hence by performing under sampling the ROC value is improved.



OUTPUT (SCREENSHOTS):

```

0    284315
1     492
Name: Class, dtype: int64
    
```

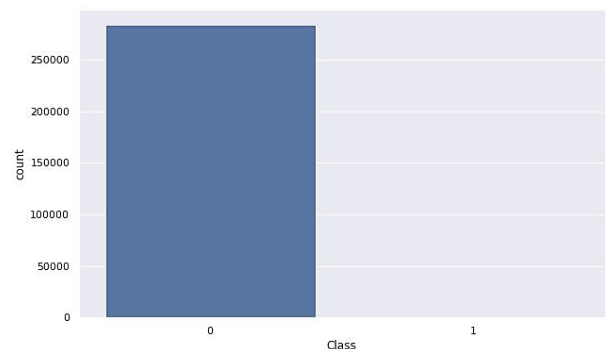


Figure 1. Before Under Sampling Technique

```
Prediction Accuracy of model on test dataset :- 0.9995201479348805
Confusion Matrix on train Dataset :-
[[85290  6]
 [ 35 112]]
Classification Report :-
precision    recall  f1-score   support

0           1.00     1.00     1.00     85296
1           0.95     0.76     0.85      147

accuracy          0.97
macro avg         0.97     0.88     0.92     85443
weighted avg      1.00     1.00     1.00     85443

AROC score :-
0.8809172093147896
```

Figure2. Output for Under Sampling Technique

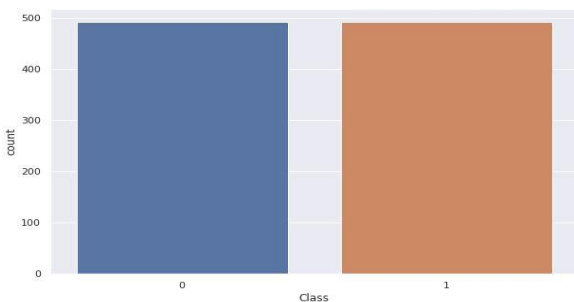


Figure3. After Under Sampling Technique

```
Model training start.....
Accuracy of model on test dataset :- 0.9593908629441624
Confusion Matrix :-
[[102  4]
 [ 4 87]]
Classification Report :-
precision    recall  f1-score   support

0           0.96     0.96     0.96     106
1           0.96     0.96     0.96      91

accuracy          0.96
macro avg         0.96     0.96     0.96     197
weighted avg      0.96     0.96     0.96     197

AROC score :-
0.9591540534936762
```

Figure4. Output After Under Sampling Technique

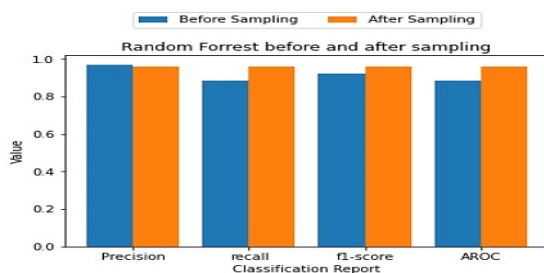


Figure5. Random Forest Before and After Under Sampling Technique

CONCLUSION

Even though there are many fraud detection techniques we can't say that this particular algorithm detects the fraud completely. From our analysis, we

can conclude that the accuracy of Random Forest is really good and after applying under sampling technique f1-score and AROC score is improved.

FUTURE SCOPE:

It is evident from the above review that several machine learning algorithms are used to detect fraud, but the findings are not satisfactory. As a result, we'd like to use deep learning algorithms to reliably detect credit card fraud.

REFERENCES

- [1] Adi Saputra1, Suharjito2L: Fraud Detection using Machine Learning in e-Commerce, International Journal of Advanced Computer Science and Applications, Vol. 10, No. 9, 2019.
- [2] Dart Consulting, Growth Of Internet Users In India And Impact On Country's Economy: [https:// www.dartconsulting.co.in/market - news/ growth- of-internet-users-in-india-and-impact -on-countrys-economy/](https://www.dartconsulting.co.in/market-news/growth-of-internet-users-in-india-and-impact-on-countrys-economy/)
- [3] Ganga Rama Koteswara Rao and R.Satya Prasad, "Shielding The Networks Depending On Linux Servers Against Arp Spoofing, Int. Journal of Engi-neering and Technology(UAE), Vol.7, PP.75-79, 2018, ,
- [4] Heta Naik , Prashasti Kanikar: Credit card Fraud Detection based on Machine Learning Algorithms, International Journal of Computer Applications (0975 – 8887) Volume 182 , March 2019.
- [5] Randula Koralage, , Faculty of Information Technology, University of Moratuwa, Data Mining Techniques for Credit Card Fraud Detection.
- [6] Roy, Abhimanyu, et al: Deep learning detecting fraud in credit card transactions, 2018 Systems and Information Engineering Design Symposium (SIEDS), IEEE, 2018.
- [7] Credit Card Fraud Detection Anonymized credit card transactions labeled as fraudulent or genuine www.kaggle.com/mlg-ulb/creditcardfraud
- [8] Working on scikit-learn library in Python to classify - Anonymized credit card transactions labeled as fraudulent or genuine <https://github.com/mohitgupta-omg/Kaggle-Credit-Card-Fraud-Detection>
- [9] Credit Card Fraud Detection using Neural Networks [https:// github.com/SimarjotKaur/Credit-Card-Fraud-Detection](https://github.com/SimarjotKaur/Credit-Card-Fraud-Detection)