International Journal of Scientific Research in Science, Engineering and Technology
Print ISSN: 2395-1990 | Online ISSN : 2394-4099 (www.ijsrset.com)
doi : https://doi.org/10.32628/IJSRSET122937

A Survey on Fraudulent Transaction Detection using Random Forest

Prof. Girija. V[1], Gowthami. V[2], Nayana. H[2], Divya. V[2], J. Prathibha[2]

[1]Assistant Professor, CiTech, Bangalore, Karnataka, India

[2]CiTech, Bangalore, Karnataka, India

Article Info
Volume 9, Issue 3

Page Number : 106-111

Publication Issue :
May-June-2022

Article History
Accepted : 01 May 2022
Published: 13 May 2022

ABSTRACT

In the evolution of the electronic money system, frequent transaction fraud has been a shadow behind the prosperity. It not only endangers the property security of users, but also hinders the development of digital finance in the world. With the development of data mining and machine learning, some mature technologies are gradually applied to the detection of transaction fraud. This paper proposes a transaction fraud detection model based on random forest. The experimental results of IEEE CIS fraud dataset show that the method of this model is better than the benchmark model, such as logistic regression, support vector machine. Finally, the accuracy of our model reached 97.4%, and the AUC ROC score was 92.7%.

The random forest classifier is composed of a group of decision trees. Each tree is generated by independent sampling random vectors, and each tree votes to find the most popular category to classify the input. Random forest has both sample randomness and characteristic randomness, and its generalization performance is superior. At the same time, random forest has good processing ability for high-dimensional data sets, which is very suitable for IEEE CIS data sets. It can process a large number of inputs and determine the most important characteristics. Therefore, further feature mining is carried out on the data extracted by RFECV.

Keywords : RFECV, AUC ROC, Digital Finance, Random Forest Classifier

I. INTRODUCTION

In the evolution of the electronic money system, frequent transaction fraud has been a shadowbehindtheprosperity.Itnotonlyendangersthepropertysecurityofusers, butalso hinders the development of digital finance in the world.

Fraudulent transactions are usually small probability events hidden in a large amount of data, and the transaction for mis extremely flexible. Machine Learning and Data Mining are useful to build detection system of fraudulent transaction.

The core detection algorithms of the system are mainly based on classification. In order to face a large amount of data, data mining related processing methods are introduced into transaction fraudtask

Data mining is a process of exploring information hidden in a large amount of data through algorithms. Through the cleaning, correction, extraction, selection and summary of a large number of data features, the hidden knowledge behind the data is obtained. For our problem, it is to extract the difference information between the behavior patterns of real users and fraud

behavior patterns in the data, so as to help the subsequent classification model better achieve the purpose of detecting fraudulent transactions.

Our data mining methods mainly include data cleaning, missing value filling, data transformation and feature extraction. In the data cleansing section, we deleted columns with a large percentage of Nan values (missing values). For example, if more than90% of the value in a feature column is Nan, the column is deleted. For the time code, we will also convert it into more accurate time information for betteruse.

In addition, we also generate many descriptive statistical features, such as the mean and extreme value of numerical characteristics such as transaction amount, billing address and mailing address. Finally, a large part of the digital features hascorrelation.

It can greatly improve the efficiency of data fitting and improve the performance of data classification. Therefore, in our work, recursive feature elimination with cross validation (RFECV) is used to eliminate each feature iteratively.

Of course, the processing process is more complex, so we need to clean the datacarefully and select the most valuable features carefully. In our work, we first carry out the truth of the data to eliminate some outliers and excessive missingdata.

In the further feature engineering cycle, the data will be transformed and the statistical data such as maximum, mean and standard deviation will be extracted. Then, Recursive feature elimination (RFECV) is used to eliminate some unimportant features. Finally, we implement a binary classifier based on random forest according to the data and features.

Random forest is a classifier with multiple decision trees. It has flexible model and fast training. It can solve the classification error caused by the extremely unbalanced data of fraud transaction detection. In order to show its superiority, we compare it with KNN. The experimental results show that the random forest model has achieved good results in accuracy and ROC AUC score. The second section introduces the characteristic engineering of financial data and non-financial data.The third section introduces the model based on random forest. In the fourth part, the performance of this algorithm is compared with other classical machine learning models by accuracy and AUC ROC score. Finally, the fifth part summarizes thispaper.

Random forest has both sample randomness and characteristic randomness, and its generalization performance is superior. At the same time, random forest has good processing ability for high-dimensional datasets, which is very suitable for IEEE CIS data sets. It can process a large number of inputs and determine the most important characteristics. Therefore, further feature mining is carried out on the data extracted by RFECV.

II. LITERATURE SURVEY

Fraud Analysis and Prevention in e-Commerce Transactions [1]

This work aims to apply and evaluate computational intelligence techniques (e.g., datamining and machine learning) to identify fraud in electronic transactions, more specifically in credit card operations performed by Web payment gateways. In order to evaluate the techniques, we apply and evaluate them in an actual dataset of the most popular Brazilian electronic payment service

In supervised strategy with labelled data, algorithms examine every transaction, previously labelled, to mathematically determine the profile of a fraudulent transaction and estimate your risk. Neural Networks, Support Vector Machines (SVM), Decision Trees and

Bayesian Networks are some of the techniques used by this strategy. SVM and random forests are sophisticated data mining techniques, which have been noted in recent years to show superior performance across different applications.

Mae's used the STAGE algorithm for Bayesian networks and "back propagation" algorithm for neural networks to detect fraud in credit card transactions. The results show that Bayesian networks are more accurate and faster training, but are slower when applied to new instances.

In order to compare the techniques, we adopt an Economic Efficiency (EE) function, which describes the financial improvement relative to the actual scenario from the corporation. In the best case, we have achieved a gain of 43.66%.

Fraud Detection by Monitoring Customer Behaviour and Activities [2]

The purpose is to prevent the customer from online transaction by using specific technique i.e., based on Data Mining and Artificial Intelligence technique. The risk score is calculated by Bayesian Learning Approachtoanalysewhetherthetransactionisgenuineor fraudulentbased on the two parameters: Customer Spending Behaviour and Geographical Locations. The customer than spending behaviour that can be identified by KMEAN clustering algorithm and in geographical location the current geographical location is compared with the previous location.

The aim is to propose a security system that can prevent from fraudulent transactions. The security mechanism must be able to identify whether the transaction is genuine or fraudulent. In case of fraudulent transaction, the security system must be able to prevent from online transactions. The main objective is to identify the customer behaviour in case of spending behaviour and their location. The customer spending behaviour can be identified with the help of KMEAN clustering algorithm. In case of spending behaviour, the customer usually performs similar type of transactions in terms of amount which can be visualized as a part of cluster and suddenly the customer performs transaction of huge amount which can be seen as outlier

The advantages of proposed system are: 1. The detection of the fraud is found much faster than the existing system. 2. We can find the most type of fraud by usingTRSGM.

Baye's theorem is used in the model, so the model adapts to changing behaviour of genuine customer as well as fraudster. Fraudster behaviour changes constantly, so security process needs to be updated regularly.

Improved Fraud Detection in e-Commerce Transactions [3]

The main objective is to detect the fraudulent transactions by using Adaptive Neuro-Fuzzy Inference System, which is a hybrid of neural networks along with fuzzy inference, wherein the system can adapt to newer instances of fraud. A fraud detection system based on Adaptive Neuro Fuzzy Approach(ANFIS) isproposed. This technique allows to utilize the advantage of both neural networks and fuzzy inference system. The advantage of self-learning from neural networks and the advantage ofspecifying or generating fuzzy rules and inferences based on the newer instances of fraud are combined together.

The only constraint is that, the data has to be provided to the network in the way it requires it, i.e., in a matrix form. Although, the proposed system gives good results with large number of inputs, future work will concentrate on reducing the number of inputs required to predict fraud, i.e., inputreduction.

Customer Transaction Fraud Detection Using Random Forest [4]

With the development of data mining and machine learning, some mature technologies are gradually applied to the detection of transaction fraud. The accuracy of our model using Random Forest has reached 97.4%, and the AUC ROC score was 92.7%.

We implement a binary classifier based on random forest according to the data and features. Random forest is a classifier with multiple decision trees. It has flexible model and fast training. It can solve the classification error caused by the extremely unbalanced data of fraud transaction detection. In order to show its superiority, we compare it with support vector machine and logistic regression. The experimental results show that the random forest model has achieved good results in accuracy and ROC AUC score.

III. Methodology

A. Transaction table:

The transaction table has 394 characteristic variables, including 22 classification features and 372 numerical features. Most digital features are anonymous with fixed prefixes. To give a specificandcleardescription,wesummarizethesevariablesinTable1.TransactionDTrefers to the transaction date and time, which can be parsed into precise time information, such as year, month, day, and week. Transaction AMT refers to the amount of transaction payment in U.S. dollars. A small part of the amount with irregular decimal, may represent the transaction for remittance calculation.

B. Identificationtable

The identification table contains 41 features, including identity information, network connection information associated with transactions (IP, ISP, agent, etc.) and behaviour information. The field names are masked and no paired dictionaries will be provided to protect privacy and sign contracts. When the two tables are then processed separately, they are joined on the transaction id key to generate a new table. It also records behavioral fingerprints, such as account login time & login failure time, account duration staying on the page, and so on.

IV. RANDOM FOREST FRAUD DETECTION MODEL

The random forest classifier is composed of a group of decision trees. Each tree is generated by independent sampling random vectors, and each tree votes to find the most popular category to classify the input. Random forest has both sample randomness and characteristic randomness, and its generalization performance is superior.

At the same time, random forest has good processing ability for high-dimensional data sets, which is very suitable for IEEE CIS data sets. It can process a large number of inputs and determine the most important characteristics. Therefore, further feature mining is carried out on the data extracted by RFECV.

We propose a new approach utilizing random forest to detect fraud based on IEEE-CIS Fraud dataset. Just like all machine learning pipeline, data pre-processing especially feature extraction is critical to the performance of our model.

By using RFECV we can reduce many features that are redundant or highly correlated which can easily bias our model. It is shown that the random forest performs better on this dataset when comparing to such of KNN. We infer that this is probably caused by the extreme class unbalance inherent to the IEEE-CIS Fraud data. Finally, the accuracy of our model reached 99.949%.

V. SYSTEM ANALYSIS

Analysis is the process of breaking a complex topic or substance into smaller parts to gain a better understanding of it. Gathering requirements is the

main attraction of the Analysis Phase. The process of gathering requirements is usually more than simply asking the users what they need and writing their answers down. Depending on the complexity of the application, the process for gathering requirements has a clearly defined process of its own.

Proposed System

The aim of the project is to design a model which identifies the fraudulent transactions. The model is mainly divided into three phases. The first phase was collecting datasets. Datasetswas collected from various online transactions and process them to remove noisy data to get useful information. The second phase was training the model and testing repeatedly till the desired accuracy is reached. The third phase was to analyze ROC and AUC score to give the efficiency of Random Forest Algorithm in finding the fraudulent transactions among thedatasets.

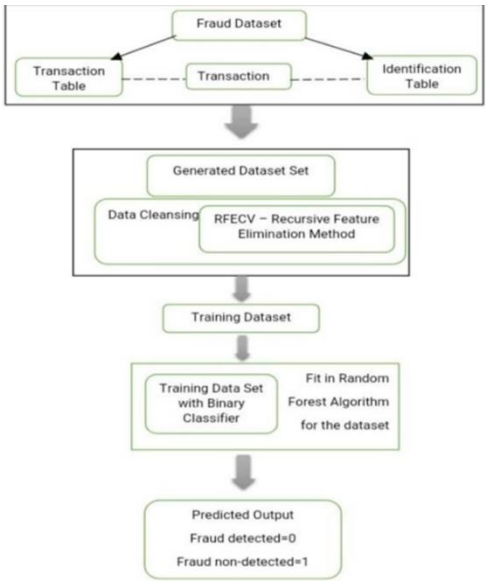


Fig : System Architecture

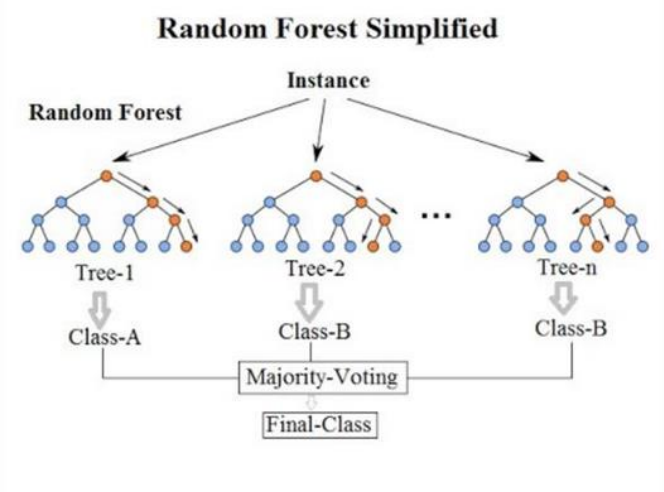
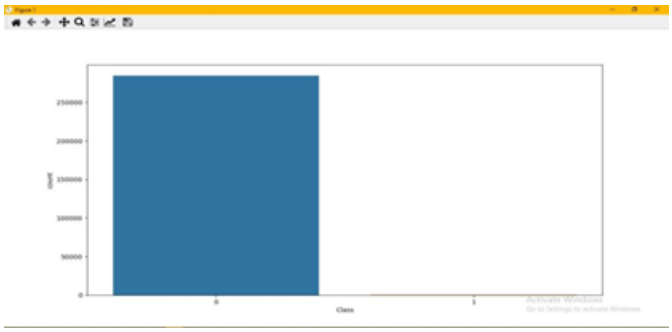


Fig: Random Forest Classification

Models	Accuracy
KNN	0.959
Random Forest	0.974

VI. RESULTS

```
##------------RANDOM FOREST ALGORITHM-----------------##
The number of Fraudulent transactions is: 492
The number of Valid transactions is: 284315
The ratio of fraudulent transactions is: 0.0017274856306200 34
Precision Score: 0.9245283018867925
Recall Score: 0.7777777777777778
Accuracy Score: 0.9994943962248252
##------------KNN ALGORITHM-----------------##
The number of Fraudulent cases is: 492
The number of valid transactions is: 284315
The ratio of fraudulent transactions is: 0.0017274856306200 34
Precision Score: 1.0
Recall Score: 0.0703125
Accuracy Score: 0.9983286986320609
```



VII. REFERENCES

[1]. https://www.javatpoint.com/machine-learning-random-forest-algorithm
[2]. https://aws.amazon.com/solutions/implementations/fraud-detection-using-machine- learning/
[3]. https://ieeexplore.ieee.org/Xplore/home.jsp
[4]. https://www.ieee.org/
[5]. https://en.wikipedia.org/wiki/Random_forest
[6]. https://towardsdatascience.com/random-forests-algorithm-explained-with-a-real-life-example-and-some-python-code-affbfa5a942c
[7]. https://youtu.be/v6VJ2RO66Ag

Cite this article as :

Prof. Girija. V, Gowthami. V, Nayana. H, Divya. V, J. Prathibha, "A Survey on Fraudulent Transaction Detection using Random Forest", International Journal of Scientific Research in Science, Engineering and Technology (IJSRSET), Online ISSN : 2394-4099, Print ISSN : 2395-1990, Volume 9 Issue 3, pp. 106-111, May-June 2022. Available at doi : https://doi.org/10.32628/IJSRSET122937
Journal URL : https://ijsrset.com/IJSRSET122937