ISSN: 2278-4632
Vol-10 Issue-5 No. 4 May 2020

A Supervised Learning Algorithm for Credit Card Fraud Detection

Sandhya Surannagari[1], Vishal Kumar Narra[2], Charan Kumar J. S.[3], K. Suparna[4]

[1,2,3]UG Student, [4]Assistant Professor, Department of CSE
[1,2,3,4]Kommuri Pratap Reddy Institute of Technology, Ghatkesar, Hyderabad, India.

ABSTRACT

In this paper we mainly focus on credit card fraud detection in real world. Here the credit card fraud detection is based on fraudulent transactions. Generally, credit card fraud activities can happen in both online and offline. But in today's world online fraud transaction activities are increasing day by day. So, to find the online fraud transactions various methods have been used in existing system. In proposed system we use random forest algorithm (RFA) for finding the fraudulent transactions and the accuracy of those transactions. This algorithm is based on supervised learning algorithm where it uses decision trees for classification of the dataset. After classification of dataset a confusion matrix is obtained. The performance of RFA is evaluated based on the confusion matrix.

Keywords: Credit card fraud detection, transactions, classification technique, random forest algorithm.

I. INTRODUCTION

Credit card fraud is increasing day by day. Credit card fraud can be done in both online and offline transactions. In offline transactions Physical cards are required while in online transactions the virtual cards are required for doing illegal or fraud activities. Thus, these fraud activities in credit card may lead to many fraud transactions without the knowledge of the actual users. The fraudsters are looking for sensitive information such as credit card number, bank account and other user details to perform transactions. In case of offline transactions, the fraudsters has to steal the credit card of the user to do the transactions and for the online transactions the fraudsters has to steal the user's identity and online details to perform the online transactions. Thus, the credit card fraud has become the major issue in today's technological world which has a massive problem in bank transactions. There are many fraud transactions which cannot be easily identified by the user and by the banking authority which leads to loss of sensitive data. There are various models which are used for detecting the fraud transactions based on the behavior of the transactions and these methods can be classified as two broad categories such as supervised learning and unsupervised learning algorithm. In existing system for finding the accuracy of the fraudulent activates they have used methods such as cluster analysis, support vector machine, naïve Bayer's classification etc. The aim of this paper is to detect the accuracy of the fraudulent transactions by using RFA.

II. EXISTING SYSTEM

In existing System, a research about a case study involving credit card fraud detection, where data normalization is applied before Naïve Bayer's and cluster Analysis and with results obtained from the use of these methods on fraud detection has shown that by clustering attributes neuronal inputs can be minimized and promising results can be obtained by using normalized data. This research was based on unsupervised learning. Significance of this paper was to find new methods for fraud detection and to

 Copyright © 2020 Authors

ISSN: 2278-4632
Vol-10 Issue-5 No. 4 May 2020

increase the accuracy of results. The data set for this paper is based on real life transactional data by a large European company and personal details in data is kept confidential. Accuracy of an algorithm is around 50%. Thus, the accuracy of the results obtained from these methods are less when compared with the proposed system.

A comprehensive understanding of fraud detection technologies can be helpful for us to solve the problem of credit card fraud. The work in [1] provides a comprehensive discussion on the challenges and problems of fraud detection research. Mohammad et.al., [2] review the most popular types of credit card fraud and the existing nature-inspired detection methods that are used in detection methods. Basically, there are two types of credit card fraud: application fraud and behavior fraud [3]. Application fraud is that criminals get new credit cards from issuing companies by forging false information or using other legitimate cardholders' information. Behavior fraud is that criminals steal the account and password of a card from the genuine cardholder and use them to spend. Recently, a kind of fraud detection method is popular in some commercial banks which is to check behaviors of the associated cardholder [7]. Almost all the existing work about detection of credit card fraud is to capture the behavior patterns of the cardholder and to detect the fraud transactions based on these patterns. Srivastava et.al. [5] model the sequence of transaction features in credit card transaction processing using a hidden markov model (HMM) and demonstrate its effectiveness on the detection of frauds. An HMM is initially trained with the normal behavior of the cardholder. If the current transaction is not accepted by the trained HMM with a high probability, it is fraudulent. However, they only consider the transaction amount as the feature in the transaction process. Amlan et.al [8] propose a method using two-stage sequence alignment which combines both misuse detection and anomaly detection [15]. In their method, a profile analyzer is used to determine the similarity of an incoming sequence of transaction on a given credit card with the legitimate cardholder's past spending sequence. Then, the unusual transactions traced by the profile analyzer are passed to a deviation analyzer for possible alignment with the past fraudulent behavior. The final decision about the nature of a transaction is taken based on the observations by the two analyzers. However, this method cannot detect frauds in real time. Elaine et.al. [9] propose a user behavior model which treats the transaction features independently. Gabriel et.al [13] propose an alternative method to prevent fraud in E-commerce applications, using a signature-based method to establish a user's behavior deviations and consequently detect the potential fraud situations in time. However, they only consider the click stream as the element of the signature. We believe that instead of using only one transaction feature for a fraud detection, it is better to consider multiple transaction features.

III. PROPOSED SYSTEM

In proposed system we use RFA for classification and regression of dataset. First, we will collect the credit card dataset and analysis will be done on the collected dataset. After the analysis of dataset then cleaning of dataset is required. Generally, in any dataset there will be many duplicate and null values will be present, so to remove all those duplicate and null values cleaning process is required. Then we must split the dataset into two categories as trained dataset and testing dataset for comparing and analyzing the dataset. After dividing the dataset, we must apply the RFA where this algorithm will give us the better accuracy about the credit card fraud transactions. By applying the RFA, the dataset will be classified into four categories which will be obtained in the form of confusion matrix. Based on the above classification of data performance analysis will be done. In this analysis the accuracy of credit card fraud transactions can be obtained which will be finally represented in the form of graphical representation.

A. RFA

Random forest is also called as random decision forest which is used for classification, regression and other tasks that are performed by constructing multiple decision trees. This RFA is based on supervised

 Copyright © 2020 Authors

ISSN: 2278-4632
Vol-10 Issue-5 No. 4 May 2020

learning and the major advantage of this algorithm is that it can be used for both classification and regression. RFA gives you better accuracy when compared with all other existing systems and this is most used algorithm. In this paper the use of RFA in credit card fraud detection can give you accuracy of about 90 to 95%.

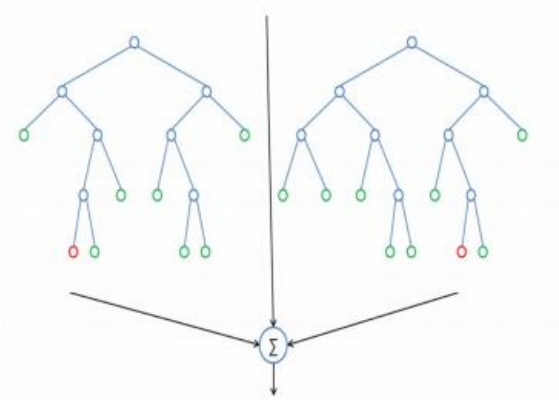

Fig. 1: Decision tree

B. RFA IMPLEMENTATION IN CREDIT CARD FRAUD DETECTION

In credit card fraud detection, the RFA gives better accuracy in results. First all the dataset will be collected and analyzed. During analysis process all the duplicate values and also the null values will be removed from the dataset. Now the dataset will be preprocessed based on the amount and transaction time for finding the accuracy of the resultant dataset. After the preprocessing of dataset into amount and transaction time now the dataset will be divided into two categories. The dataset is classified in two categories as trained data and test dataset. Here for dataset classification we use a software called 'Scikit-learn'. Scikit-learn is a free software for machine learning library in python where it contains features like classification, regression, clustering algorithms and various algorithms to interoperate with Python. After the preprocessing of the dataset now we apply the RFA. By applying RFA the preprocessed dataset will be analyzed again and then a confusion matrix will be obtained. In confusion matrix the dataset will be partitioned into four blocks as True Positive (TP), True Negative (TN), False Positive (FP) and False Negative (FN). Now the dataset will be partitioned continuously until all the data is validated. Now all these partitioned data will be evaluated and finally it will be represented as separate graphs. These separate graphs will give only less accuracy about the resultant dataset. So, in order to obtain better accuracy, we use RFA where it takes all the graph values and give us only necessary values with better accuracy when compared with all other algorithms.

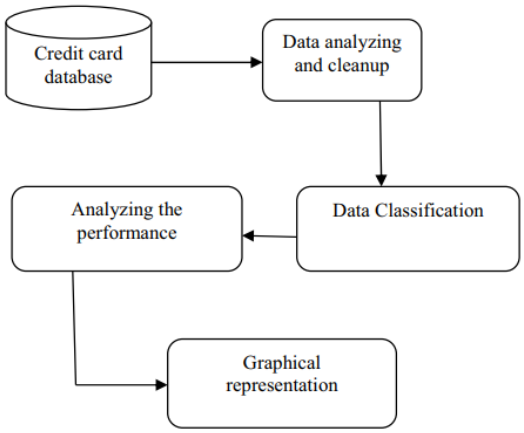

Fig. 2: System architecture.

 Copyright © 2020 Authors

ISSN: 2278-4632
Vol-10 Issue-5 No. 4 May 2020

In our architecture first we have a credit card dataset where this contains all the details about credit card. But here we take only Amount and Transaction time for analysis and preprocessing of dataset. The next step is the process of data cleaning where the dataset will be analyzed, and all the duplicate and null values will be eliminated from the dataset taken. The next step is the data partition where the credit card dataset will be partitioned into two partitions as trained dataset and testing dataset. After that RFA will be applied and a confusion matrix will be obtained. Now the performance analysis will be done on the obtained confusion matrix. This Performance analysis will give the accuracy of about 90% in this credit card fraud detection system.

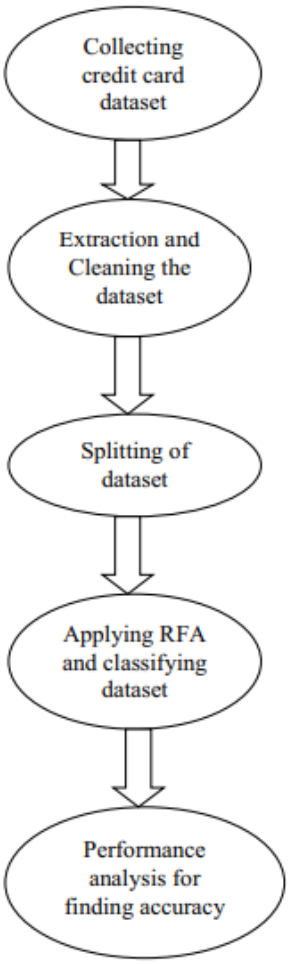

Fig. 3: RFA Implementation

MODULES

Module 1: Exploratory Data Analysis In this module we will first collect all the credit card dataset and store it in a database. Then we will perform some descriptive analysis about the dataset.

Module 2: Data Cleaning In the next step, after analyzing the dataset then we have to clean the data. In this cleaning process all the duplicate values and null values that are present in the dataset will be removed and a new dataset will be obtained.

Module 3: Preprocessing of dataset In this module the cleaned dataset will be preprocessed where the dataset will be divided based on amount and transaction time.

 Copyright © 2020 Authors

ISSN: 2278-4632
Vol-10 Issue-5 No. 4 May 2020

Module 4: Dataset Partition In this module first the dataset will be divided into two partitions as trained dataset and testing dataset. After the data partitions the RFA is applied. After applying RFA finally a confusion matrix is obtained.

Module 5: Evaluation Now the resultant data obtained in the form of confusion matrix can be evaluated by using graphical representation which gives better accuracy

4. EXPERMENTAL RESULTS

This section shows the details and results of experiments. Firstly, a performance comparison is made on the same subset. Then we explore the relation between a model's performance and the ratio of legal and fraud transactions in a subset. Finally, it shows the performances of models on a much bigger dataset, which is more closed to the actual result.

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O
1	Customer_	Customer_Name	Customer_City	Customer_State	Time	Tansaction_Mo	Transaction_ty	Amount	Gender	No_of_Car	Amount_Balanc	No_of_Transactions	Credit_Limit	Age	Self_I
2	AD125786	VENKAT	HYDERABAD	ANDHRA	00:30:00	ONLINE	NET_BANKING	25	MALE	2	75	3	150	27	NO
3	AS987878	JAYA	WARANGAL	TELANGANA	00:50	OFFLINE	CHEQUE	100	MALE	2	300	3	200	38	YES
4	BC456324	LAKSHMI	MUMBAI	MAHARASHTRA	00:55:00	ONLINE	POS	100	FEMALE	1	55	2	90	24	NO
5	AB528932	RIYA	BANGLORE	KARNATAKA	00:05	ONLINE	POS	18	FEMALE	3	100	12	85	54	YES
6	CF894532	KIYA	CHENNAI	TAMIL NADU	01:34:02	OFFLINE	CASH_WITHDR	1	FEMALE	1	18	21	200	28	YES
7	PL889966	SOMYA	KOLKATA	WEST BENGAL	13:34:00	ONLINE	NET_BANKING	60	FEMALE	2	200	10	90	32	YES
8	AW749623	AKSHITH	JAIPUR	RAJASTHAN	23:12:03	ONLINE	POS	12	FEMALE	1	50	5	185	34	NO
9	ED563217	AZZAR	KOCHI	KERALA	19:11:00	ONLINE	NET_BANKING	20	MALE	2	90	3	200	32	NO
10	ET789456	SHASHI	KOLKATA	WEST BENGAL	16:10:03	OFFLINE	CASH_WITHDR	28	MALE	2	70	5	100	29	YES
11	AZ123654	LIKESH	LUCKNOW	UTTAR PRADESH	02:54:00	ONLINE	NET_BANKING	34	MALE	1	78	4	100	27	YES
12	AF486248	SUMAN	PUNE	MAHARASHTRA	05:09:00	ONLINE	POS	36	MALE	1	28	6	150	26	NO
13	AR753951	ARVIND	VISHAKAPATANA	ANDHRA	09:22:00	OFFLINE	CASH_WITHDR	40	MALE	1	60	21	200	24	YES
14	UE159753	TARUN	BHOPAL	MADHYA PRADESH	11:33:35	ONLINE	NET_BANKING	26	MALE	3	36	8	250	31	NO
15	ER478532	BHASKER	INDORE	MADHYA PRADESH	08:06:54	ONLINE	NET_BANKING	36	MALE	2	120	19	300	30	YES
16	WQ964183	ASHISH	KANPUR	UTTAR PRADESH	00:30:00	OFFLINE	CHEQUE	16	MALE	1	85	17	150	20	YES
17	OO852147	BINDU	AGRA	UTTAR PRADESH	00:50	OFFLINE	CASH_WITHDR	2	FEMALE	3	12	20	100	24	NO
18	QS753698	KIRAN	VARANASI	UTTAR PRADESH	00:55:00	OFFLINE	CASH_WITHDR	9	MALE	3	60	20	300	41	YES

Fig. 4: Credit card fraud detection.

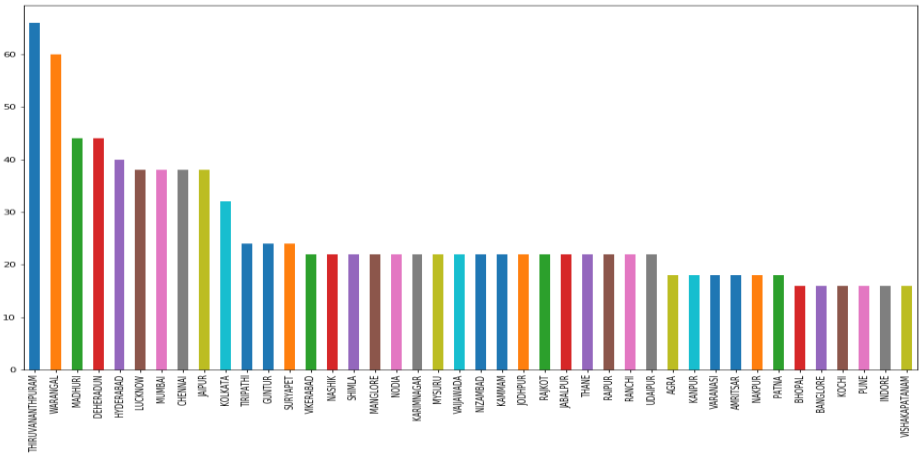

Fig. 5: Credit card fraud analysis in different locations

 Copyright © 2020 Authors

ISSN: 2278-4632
Vol-10 Issue-5 No. 4 May 2020

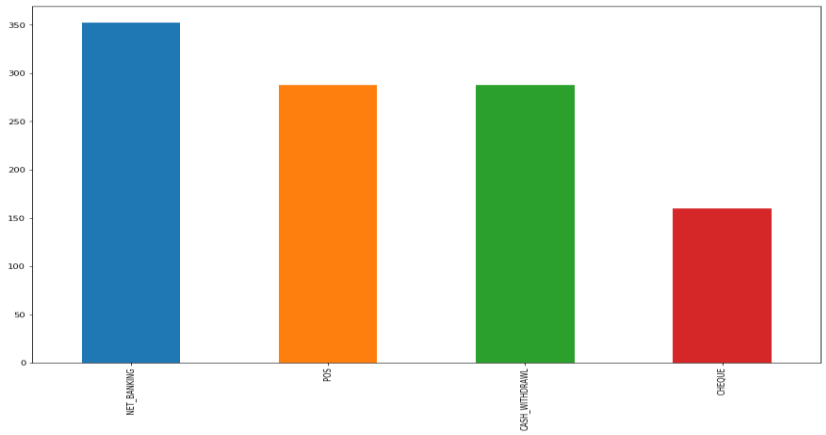

Fig. 6: Credit card fraud analysis in different banking environments.

```
In [476]: from sklearn.metrics import accuracy_score
          from sklearn.metrics import precision_score, recall_score, confusion_matrix
          from sklearn.metrics import f1_score, roc_auc_score, roc_curve



          generate_model_report(y_test, Y_Test_Pred)

          Accuracy =  0.9632352941176471
          Precision =  0.8947368421052632
          Recall =  0.85
          F1 Score =  0.8717948717948718
```

Fig . 7: Results of RFA.

5. CONCLUSION

This paper has examined the performance of two kinds of random forest models. A real-life B2C dataset on credit card transactions is used in our experiment. Although random forest obtains good results on small set data, there are still some problems such as imbalanced data. Our future work will focus on solving these problems. The algorithm of random forest itself should be improved. For example, the voting mechanism assumes that each of base classifiers has equal weight, but some of them may be more important than others. Therefore, we also try to make some improvement for this algorithm.

REFERENCES

[1] Gupta, Shalini, and R. Johari. ”A New Framework for Credit Card Transactions Involving Mutual Authentication between Cardholder and Merchant.” International Conference on Communication Systems and Network Technologies IEEE, 2011:22-26.

[2] Y. Gmbh and K. G. Co, “Global online payment methods: Full year 2016,” Tech. Rep., 3 2016. [3] Bolton, Richard J., and J. H. David. ”Unsupervised Profiling Methods for Fraud Detection.” Proc Credit Scoring and Credit Control VII (2001):5– 7.

[4] Seyedhossein, Leila, and M. R. Hashemi. ”Mining information from credit card time series for timelier fraud detection.” International Symposium on Telecommunications IEEE, 2011:619-624.

[5] Srivastava, A., Kundu, A., Sural, S., and Majumdar, A. (2008). Credit card fraud detection using hidden markov model. IEEE Transactions on Dependable and Secure Computing, 5(1), 37-48.

 Copyright © 2020 Authors

ISSN: 2278-4632
Vol-10 Issue-5 No. 4 May 2020

[6] Drummond, C., and Holte, R. C. (2003). C4.5, class imbalance, and cost sensitivity: why under-sampling beats oversampling. Proc of the Icml Workshop on Learning from Imbalanced Datasets II, 1–8.

[7] Quah, J. T. S., and Sriganesh, M. (2008). Real-time credit card fraud detection using computational intelligence. Expert Systems with Applications, 35(4), 1721-1732.

[8] Kundu, A., Panigrahi, S., Sural, S., and Majumdar, A. K. (2009). Blastssaha hybridization for credit card fraud detection. IEEE Transactions on Dependable and Secure Computing, 6(4), 309-315.

[9] Shi, E., Niu, Y., Jakobsson, M., and Chow, R. (2010). Implicit Authentication through Learning User Behavior. International Conference on Information Security (Vol.6531, pp.99-113). Springer-Verlag.

[10] Duman, E., and Ozcelik, M. H. (2011). Detecting credit card fraud by genetic algorithm and scatter search. Expert Systems with Applications, 38(10), 13057-13063. [

11] Bhattacharyya, S., Jha, S., Tharakunnel, K., and Westland, J. C. (2011). Data mining for credit card fraud: a comparative study. Decision Support Systems, 50(3), 602-613.

[12] Sahin, Y., and Duman, E. (2011). Detecting credit card fraud by decision trees and support vector machines. Lecture Notes in Engineering and Computer Science, 2188(1).

[13] Mota, G., Fernandes, J., and Belo, O. (2014). Usage signatures analysis an alternative method for preventing fraud in E-Commerce applications. International Conference on Data Science and Advanced Analytics (pp.203-208). IEEE. [14] Behdad, M., Barone, L., Bennamoun, M., and French, T. (2012). Natureinspired techniques in the context of fraud detection. IEEE Transactions on Systems Man and Cybernetics Part C, 42(6), 1273-1290.

[15] Ju, W. H., and Vardi, Y. (2001). A hybrid high-order markov chain model for computer intrusion detection. Journal of Computational and Graphical Statistics, 10(2), 277-295.

 Copyright © 2020 Authors