CREDIT CARD FRAUD DETECTION USING RANDOM FOREST ALGORITHM

1st Ankita Vijay Sontakke
sontakkeankita3@gmail.com
Department of Computer Science
And Engineering
Shri Sai College of Engineering and
Technology,Chandrapur ,India

2nd Prof Ashish Deharkar
ashish.deharkar@gmail.com
Department of Computer Science
And Engineering
Shri Sai College of Engineering and
Technology,Chandrapur ,India

3rd Prof Pushpa Tandekar
p.tandekar@yahoo.in
Department of Computer Science
And Engineering
Shri Sai College of Engineering and
Technology,Chandrapur ,India

ABSTRACT:This paperfocuses on credit card fraud detection in real world. Here the credit card fraud detection is based on fraudulent transactions. 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 in order to find the online fraud transactions various methods have been used in existing system. In this paper we use Random Forest Algorithm(RFA) for finding the fraudulent transactions and the accuracy of those transactions. The performance of Random Forest Algorithm is evaluated based on the confusion matrix. The results obtained from processing the dataset gives accuracy of about 90%.

KEYWORDS: Credit Card, Fraud Detection, Random Forest.

INTRODUCTIONThere are various fraudulent activities detection techniques has implemented in credit card transactions to develop models based on Artificial Intelligence , data mining, fuzzy logic and machine learning. Credit card fraud detection is significantly difficult, but also popular problem to solve. In this paper we built the credit card fraud detection using Machine learning. Machine learning has been identified as a successful measure for fraud detection. A large amount of data is transferred during online transaction processes, resulting in a binary result: genuine or fraudulent. we are applying random forest algorithm for classification of the credit card dataset. Random Forest is an algorithm for classification and regression. Summarily, it is a collection of decision tree classifiers. Random forest has advantage over decision tree as it corrects the habit of over fitting to their training set. A subset of the training set is sampled randomly so that to train each individual tree and then a decision tree is built, each node then splits on a feature selected from a random subset of the full feature set. Even for large data sets with many features and data instances training is extremely fast in random forest and because each tree is trained independently of the others. The Random Forest algorithm has been found to provide a good estimate of the generalization error and to be resistant to over fitting.

METHODOLOGYRandom Forest Algorithm is used to detect the accuracy of the fraud in the transaction. These are a categorization, regression, and other tasks that use an ensemble learning strategy that involves teaching a greater number of decision trees and then determining the norm of the classifications (categorization) or the overall prediction (regression) of each tree. Random Forest is a supervised classification technique that uses ensemble learning. Ensemble model is a type of machine learning in which multiple versions of a same algorithm are combined to create a far more effective predictive model. The

important steps in executing the algorithm are as described in the following: Should pick N records chosen at random from the dataset Based on these N records, need to create a decision tree.By repeating above steps with the number of trees you wish to provide in your approach The final value can be obtained by taking the mean of all the readings projected by all the nodes in the algorithm. As an outcome, the new record is delivered to the classification that obtains the most votes.

There are various methods to investigate Credit Card Fraud detection as Randomforest, SVM, naïve bayes models and test its performance from accuracy, recall, precision and f1 score.

A. NAÏVE BAYES

This algorithm learns the probability of an object with certain feature belonging to a particular class. It can be described as probability of an event will occur based on another which has already occurred. It can be written as

$$P(A/B) = \frac{P\left(\frac{B}{A}\right).P(A)}{P(B)}$$

1. P (A/B): Conditional probability- Probability of occurrence of event A given the event B is true.
2. P (B/A): Likelihood probability- Probability of the occurrence of event B given the event A is true.
3. P (A), P (B) are occurrence of event A and B respectively.

B. Support Vector Machine (SVM)

It is supervised learning algorithms, which is used for both classifications as well as regression problems. It is mainly used for classification problems in Machine learning. SVM classifies the two classes using hyper plane. This hyper plane should have the largest margin in a high dimensional space to separate given data into classes.
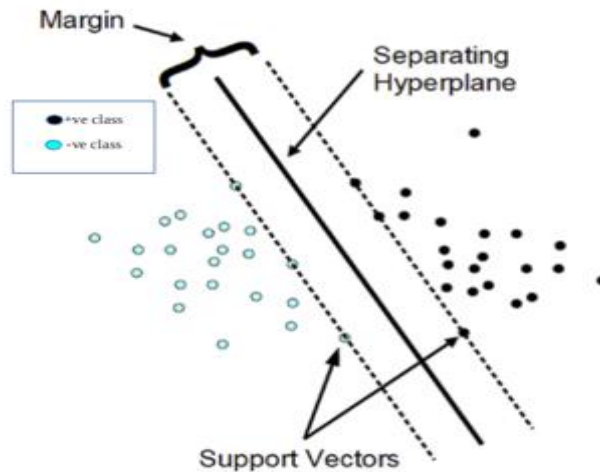
Fig No.1 Support Vector Machine

C. Random Forest

It is a machine learning technique that constructs multiple decision trees. The final decision is made based on the outcome of the majority of the decision trees. It can be used for both regression and classification purposes. But this algorithm is mainly used for classification purpose. This algorithm creates decision trees on the sample data and gets the prediction from each sample data.
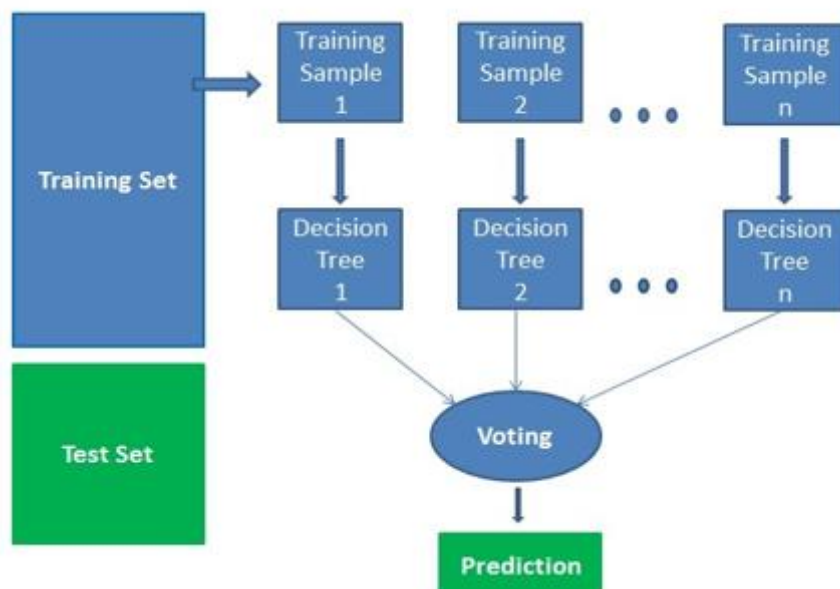

Fig No.2 Random Forest

D. Confusion Matrix

It is a specific table that is used to measure the performance of the algorithm. It is used to summarize the performance of the classification algorithm. It shows the error in performance of algorithm in the form of matrix hence it's called as error matrix. The matrix is based on actual and predicted parameters.

Fig No.3Confusion Matrix

1. *True Positive (TP):* The model has predicted YES, and the actual value was also true (YES).
2. *False Positive (FP):* The model has predicted YES, but the actual value was NO. It is also called as Type-I error.
3. *False Negative (FN):* The model has given prediction NO, and the actual value was YES, it's also called as Type-II error.
4. *True Negative (TN):* The model has given prediction NO, and the actual value was also NO.

E. Accuracy

It is the proportion of correct classifications (true positives and negatives) from overall number of cases. It defines hoe often the model predicts the correct output. The formula for accuracy is given below

$$Accuracy = \frac{TP+TN}{TP+FP+TN+FN}$$

SOFTWARE DESCRIPTION

PYTHON:- Python is an open source programming language. Python was made to be easy-to-read and powerful. Python is a good programming language for beginners. It is a high-level language, which means a programmer can focus on what to do instead of how to do it. Writing programs in Python takes less time than in some other languages. Python has a very easy-to-read syntax.

LIBRARY MODULES:-

1. NUMPY:- NumPy is a Python package. It stands for 'Numerical Python'. It is a library consisting of multidimensional array objects and a collection of routines for processing of array. Numeric, the ancestor of NumPy, was developed by Jim Hugunin. Another package Numarray was also developed, having some additional functionalities.

2. PANDAS:- Pandas is an open-sourcePython library providing high-performance, easy-to-use data structures and data analysis tools for the Python programming languageIn this tutorial, we will learn the various features of Python Pandas and how to use them in practice. Pandas deals with the following three data structures : , Series, DataFrame , Panel

 3.MATPLOITLIB:- It is a collection of command style functions that make matplotlib work like MATLAB. Each pyplot function makes some change to a figure: e.g., creates a figure, creates a plotting area in a figure, plots some lines in a plotting area, decorates the plot with labels, etc.

4.SCIKIT-LEARN Scikit-learn is a machine learning library for Python. It features several regression, classification and clustering algorithms including SVMs, gradient boosting, k-means, random forests and DBSCAN. It is designed to work with Python Numpy and Scipy.

CONCLUSIONIn earlier stage fraudulent cases has very less because everyone have direct cash withdrawal from bank or if any urgency they go directly but today fraudulent cases reported day by day. Although random forest obtains good results on small set data, there are still some problems such as imbalanced data. But data should be balanced by using the accuracy level compared to other algorithm it gives moreand also choose best feature

extraction and pre-processing technique to enhance the algorithm performance. From this project we detect the fraudulent cases taken in society.

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] Devi Meenakshi. B1, Janani. B2, Gayathri. S3, Mrs. Indira. N ,"CREDIT CARD FRAUD DETECTION USING RANDOM FOREST" International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395-0056 Volume: 06 Issue: 03 | Mar 2019 p-ISSN: 2395-0072

[2] Kaduru Thulasi Teja*, Parlapalli Anuhya*, Ponnapoola Ruchitha*, Tamizhmalar*," CREDIT CARD FRAUD DETECTION USING RANDOM FOREST ALGORITHM IN MACHINE LEARNING AND DATA SCIENCE "International Research Journal of Modernization in Engineering Technology and Science Volume:04/Issue:05/May-2022 Impact Factor- 6.752

[3] K. Deepika, M. Pavan Sai Nagendra, M Vamshi Ganesh, N. Naresh, "IMPLEMENTATION OF CREDIT CARD FRAUD DETECTION USING RANDOM FOREST ALGORITHM" IJRASET Journal for Research in Applied Science and Engineering Technology

[4] 1Prof. Teena Varma, 2Mahesh Poojari, 2 Jobin Joseph, 2Ainsley Cardozo,"Credit Card Fraud Detection Using Random Forest Algorithm" International Journal of Trendy Research in Engineering and Technology Volume 5 Issue 3 June 2021 ISSN NO 2582-0958

[5] Mrs. B. Rama Devi, 1 Sk. Neha,2M. V Jyothsna,3M. Srilekha,4N. Mounika,5N. Maruthi Priya6," Credit Card Fraud Detection Using Random Forest Algorithm" JOURNAL OF CRITICAL REVIEWS ISSN- 2394-5125 VOL 9, ISSUE 04, 2022

[6] Technique B. Mohankumar1 , Dr. K. Karuppasamy2," Credit Card Fraud Detection Using Random Forest "International Journal of Innovative Research in Science, Engineering and Technology (A High Impact Factor, Monthly, Peer Reviewed Journal) Vol. 8, Issue 4, April 2019

[7] Hardik Jethava1 , Feon Jaison2,"Credit Card Fraud Detection Using Random Forest Algorithm" INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH IN TECHNOLOGYApril 2022 | IJIRT | Volume 8 Issue 11 | ISSN: 2349-6002

[8] S. Monika1 , K. Venkataramanamma , P. Pritto Paul , M. Usha ," Credit Card Fraud Detection using Random Forest Algorithm "International Journal of Research in Engineering, Science and Management Volume-2, Issue-3, March-2019 ISSN (Online): 2581-5792

[9] B. Mohankumar1 , Dr. K. Karuppasamy2,"Credit Card Fraud Detection Using Random Forest Technique" International Journal of Innovative Research in Science, Engineering and Technology (A High Impact Factor, Monthly, Peer Reviewed Journal) Vol. 8, Issue 4, April 2019

[10] M. Mohanapriya1 , M. Kalaimani2,"Credit Card Fraud Detection"International Advanced Research Journal in Science, Engineering and Technology Vol. 7, Issue 9, September 2020 DOI 10.17148/IARJSET.2020.7903

[11] Lowlesh Nandkishor Yadav, "Predictive Acknowledgement using TRE System to reduce cost and Bandwidth" IJRECE VOL. 7 ISSUE 1 (JANUARY- MARCH 2019) pg no 275-278