iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 7 **Issue:** IV **Month of publication:** April 2019

DOI: https://doi.org/10.22214/ijraset.2019.4209

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 6.887
Volume 7 Issue IV, Apr 2019- Available at www.ijraset.com

Credit Card Fraud Detection Using Random Forest and Local Outlier Factor

Abhilasha Kulkarni[1], Priyanka Ghare[2], Apoorva Dharadhar[3], Anushka Dhekne[4], Aditi Helaskar[5]
[1, 2, 3, 4 ,5]*Department of CSE, Marathwada Mitra Mandals College Of Engineering, Pune, India*

Abstract: ***With the evolution of new technology, the use of credit cards has augmented. As credit card becomes the trendiest style of payment for online payment as well as manual payment, the numbers of credit card frauds are increasing day by day. It is necessary to curb the credit card frauds as it causes huge amount of financial loss. Many modern techniques based on Artificial Intelligence, Data mining, Machine Learning, Sequence Alignment, Genetic programming, etc are available that can be used in detecting the fraudulent transactions. We have used machine learning based algorithms. Machine Learning is a subset of Artificial Intelligence. It is a scientific study of algorithms and statistical models that computer systems use to effectively perform a specific task without using any explicit instructions, relying on patterns and inference instead. Machine learning is the technology in which we train the machine by using various algorithms and make the machine capable enough to take its own decisions. Machine Learning consists of many algorithms that can be used in fraud detection such as Random Forest, Local Outlier Fraction, Isolation Forest, Naïve Bayes, K-nearest Neighbour, Hidden Markov Model, Neural Networks, etc that can be used in fraud detection. In this paper we have done comparative study of Random Forest algorithm and Local Outlier Factor.***
Keywords: Credit Card, Fraud Detection, Random forest, Local Outlier Factor, Financial Loss

I. INTRODUCTION

One way by which credit card fraud takes place is by getting access to the stolen credit cards and second is by exploiting the details of the card via online transaction without the knowledge of the genuine card holder. While detecting credit card fraud we face many challenges. Millions of transactions take place per minute all over the world. Detecting which among all the transactions is fraudulent is a challenging task.

The amount, time, place at which the transactions take place is different. The amount of research carried out in this field is low due to the lack of availability of data sets. As the details of the users are confidential, working on the real data sets becomes impossible. Machine learning algorithms can be classified into supervised and unsupervised algorithms. Supervised algorithms consist of a predetermined set of data that is provided for training the system. The system tries to predict the results based on the previous examples or training data.

On the other hand, in case of unsupervised algorithms the system tries to find the patterns directly from the example provided. Therefore, if the dataset is labelled then it comes under supervised algorithm and if the dataset is unlabelled it comes under unsupervised algorithm.

We have used random forest algorithm and local outlier factor. Random forest algorithm is a supervised classification algorithm. It is used for both regression as well as classification kinds of problems. Local Outlier fraction is an anomaly detection algorithm. Outlier is a synonym for anomalies. Anomaly refers to the abnormal behaviour or a deviation from normal behaviour or expected behaviour of a certain data points with the respect to certain attributes. Outliers have a different statistical property. Applications of these algorithms are speech recognition, banking sector, healthcare, pattern recognition, etc.

II. LITRATURE SURVEY

A. In this paper, the developments and improvements of Random Forest in the last 15 years are presented. This paper also presents the description of usage of Random Forest in various fields like Medicine, Agriculture, Astronomy, etc.

B. Local outlier factor is an anomaly detection technique and has various applications in numerous fields. This paper deals with abnormal usage of data and unauthorized access in large-scale critical networks. It mainly focuses on healthcare infrastructures.

C. The two basic types of outliers are global outlier and local outlier. A large amount of data is generated continuously by various applications. Outlier detection is a technique of detecting the data whose behaviour is deviated from normal behaviour or expected behaviour. It is necessary to detect outliers from streaming data as well as static data. This paper focuses on detection techniques for static and streaming data. The work also focuses on various local and global detection techniques.

D. We have studied various anomaly detection algorithms and isolation forest is one of them. Isolation forest works on the principle of randomly selecting a split value between the maximum and minimum values based on the selected feature. The feature is randomly selected. This paper presents an extension to the model-free anomaly detection algorithm.

E. In this paper 4 different algorithms are compared namely KNN, AdaBoost, Random Tree and Logistic Regression gor imbalanced data.

III. DATA ANALYSIS

Selected dataset contains records of card holders who made transactions using credit card in September 2013. In the dataset of 2,84,807 transactions, 492_are fraudulent. Selected dataset is in the comma- separated values format i.e. CSV format. CSV file format is used to store the data in tabular form. Dataset values are in numerical form as PCA (Principal Component Analysis) transformation is done on input values. This conversion is done so that the user's personal details remain hidden and the user's security is maintained. Columns having heads as V1 to V28 show PCA transformed numeric values but time, amount and class features show their genuine values. Sometimes while dealing with huge databases it is not possible to do a detailed observation on each value, hence graphical representation of data makes observation easier. In this dataset time, amount, class and columns V1 to V28, total 31 features are represented in the form of Histogram. Histogram is an accurate representation of the distribution of numerical data. Time feature shows the elapsed time between transactions while amount shows actual transaction amount. Class is the result variable which gives values in the form of 0 and 1, 1 for fraudulent transactions and 0 for valid transactions.

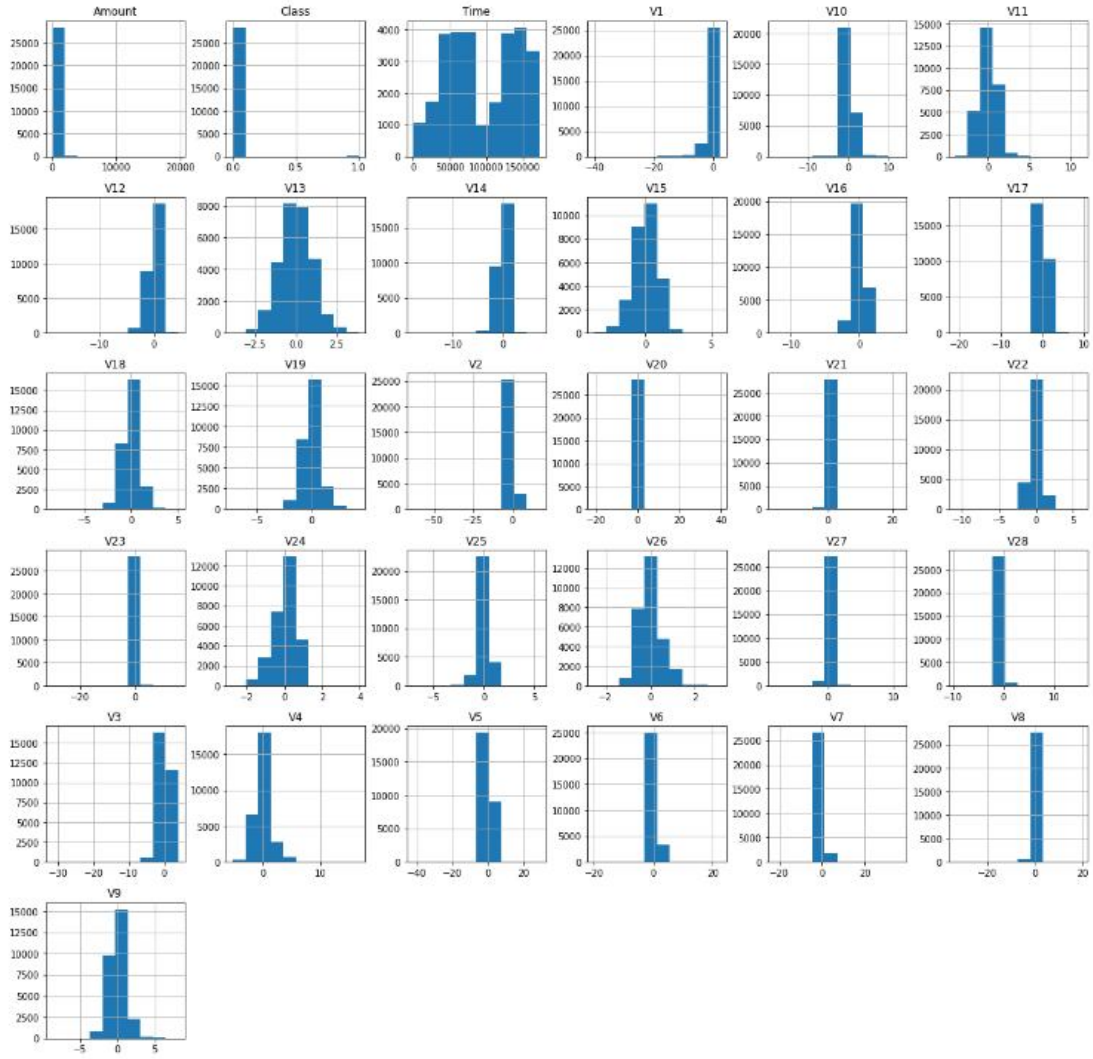


Figure 1: Histogram Showing Dataset Features

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 6.887
Volume 7 Issue IV, Apr 2019- Available at www.ijraset.com

IV. DETECTION METHODS

A. Random Forest Algorithm

In Random forest algorithm, Decision trees are the main components. Decision tree is used for both classification and Regression. Decision tree is one of the powerful and popular method for classification and prediction. It is tree like structure where internal nodes denotes test on attribute, each branch represents an outcome of a particular test in terms of binary classification(answer is in the form of true or false, 1 or 0, yes or no)and leaf node (terminal node) holds decision or classification. For Construction of Decision tree source test is split into subsets based on an attribute value test. Now for each derived subset this process is repeated called as recursive partitioning. When splitting no longer add value to the predictions, recursion is completed.

Example of Decision Tree: Mark will play cricket today or not

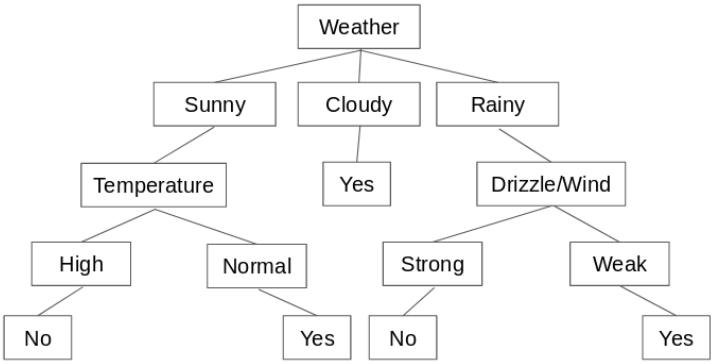


Figure 2: Sample Decision Tree

1) Advantages Of Decision tree
a) It clearly indicates important fields for classification
b) It Does Classification without much complex computations
c) It handles both continuous and categorical variables
d) It generates simple and understandable rules

Random Forest Algorithm is Supervised Learning Algorithm. It is capable of doing both classification and regression. Random forest is method that operates by constructing multiple decision trees during training of the model . The decision voted by maximum trees is considered by the random forest algorithm. Number of trees in forest and results are directly related to each other as higher number of trees in forest leads to higher efficiency .For Implementation of random forest algorithm Decision tree is the support tool. We have already discussed decision tree. We input a training dataset with labels and pass to decision tree module and it formulates some rules. These rules can be used to perform predictions.

Working:

B. Random Forest Creation

1) Randomly select ‘r’ features from all total features, r << total features
2) 2. Among r features calculate node using best split point
3) Split the node into child nodes using splitting method
4) Repeat the process for further nodes
5) Follow above steps ‘n’ times to create ‘n’ number of trees into the forest

C. Random Forest Predictions

1) Take features and use rules of each randomly created decision tree to predict outcome and store it for further use
2) Calculate the votes for each predicted feature
3) Consider highest voted answer as the final prediction from random forest algorithm

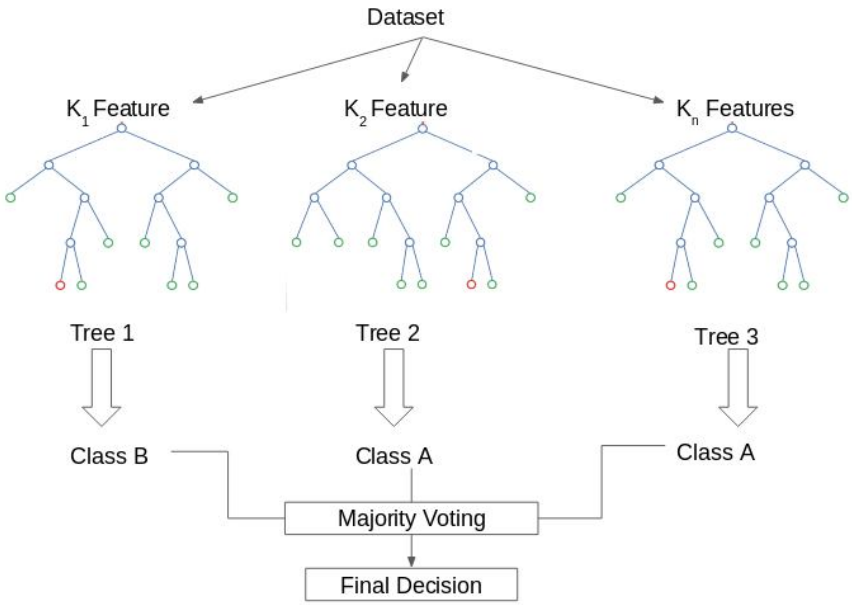


Figure 3: Random Forest

1) Advantages of Random Forest Algorithm:
a) It is capable of doing both classification and regression
b) It can handle missing values
c) Random forest classifier can be modelled for categorical values
d) No over fitting of trees

D. Local Outlier Factor

The Local Outlier Factor or LOF algorithm is an unsupervised anomaly detection method. It computes the local deviation of a given a data point with respect to its neighbours. Local Outlier Factor considers as outliers the samples that have a substantially lower density than their neighbours. Below example shows how to use Local Outlier Factor for outlier detection. Note that when LOF is used for outlier detection it has no predict, decision function and scoe_samples methods. The number of neighbors is typically set 1) greater than the minimum number of samples a cluster has to contain, so that other samples can be local outliers relative to this cluster, and 2) smaller than the maximum number of close by samples that can potentially be local outliers.

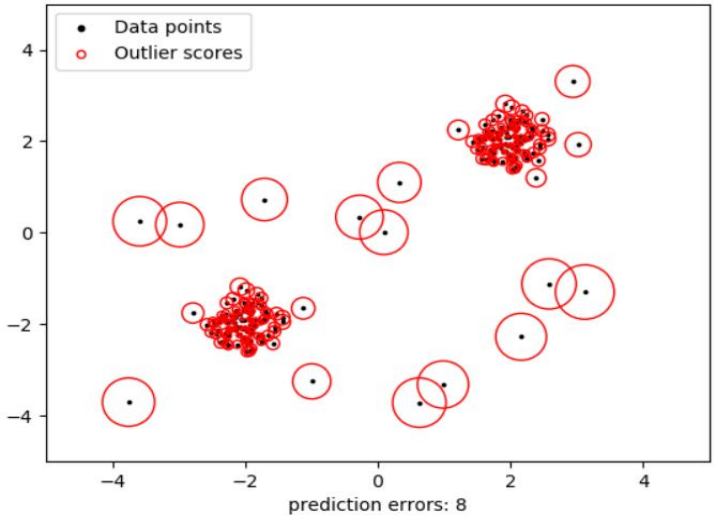


Figure 4: Local Outlier Factor

1) Types of Outliers: Outliers can be classified into three types, namely

a) Point Outlier: If an individual data instance is considered as anomalous with respect to the rest of the data, then the instance is termed as point outlier. As a real-life example, if we consider the amount spent by the user using a credit card, a normal range of amount being spent is observed and if suddenly a transaction of higher amount (i.e. the amount higher than the normal amount range in which the person spends his money) is observed than that transaction will be a point outlier.

b) Contextual Outlier: Object deviates significantly from the rest of the data set based on a selected context. For example, 28°C temperature in Moscow in winter season is an outlier whereas, 28°C in summer in Moscow is not an outlier. Here the context is temperature, that is, the outliers are detected based on the context of temperature. The temperature that deviates the most from a group of common temperatures is a contextual outlier. Another similar example is credit card fraud detection. For example, if an individual's weekly expenditure is $200 and suddenly during the week of Christmas his expenditure rises and reaches to $1200 then it will be considered as a contextual outlier. The reason for it being considered as a contextual outlier is that it doesn't confirm to the normal spending behaviour of that individual. However, if the same individual spends the same amount during the same week then it will be considered normal.

c) Collective Outlier: When a subset of data objects collectively deviate significantly from the whole data set, even if the individual data objects may not be outliers is a collective outlier. Example of collective outlier is as follows

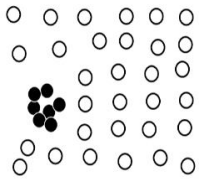


Figure 5: Collective Outlier

Usually, a data set may contain different types of outliers and at the same time may belong to more than one type of outliers.

There are many outlier detection techniques- Supervised methods Unsupervised methods Semi-supervised methods Statistical methods Parametric Methods, Non-parametric methods Proximity-based algorithms Density based methods In our case, we have used supervised methods for outlier detection. Supervised methods – In this method, the samples that are tested by experts are used training and testing purposes. Challenges faced in outlier detection approach are as follows- Classes are unbalanced.

That is, the population of outliers is typically much smaller than the normal objects. Methods used for handling unbalanced classes such as sampling techniques can be used. Catch as many outliers as possible, i.e. recall is more important than accuracy (i.e., not mislabelling normal objects as outliers).

2) Advantages of Local Outlier Factor

a) Due to the local approach, LOF is able to identify outliers in a data set that would not be outliers in another area of the data set. For example, a point at a "small" distance to a very dense cluster is an outlier, while a point within a sparse cluster might exhibit similar distances to its neighbours. It outperforms network intrusion detection and on processed classification benchmark data.

b) The LOF family of methods can be easily generalized and then applied to various other problems such as detecting outliers in geographic regions.

V. METHODOLOGY USED

We are using Random Forest Algorithm and Local Outlier Factor for detecting fraudulent credit card transactions from the dataset. Here given dataset is in labelled format. For analysing efficiency of the algorithms, we use split function on database. Split function divides the dataset in training data and testing data. Amount of data that is to be divided into training and testing data is upon user. User can decide how much data to be used for training and testing purposes as per the need. Training data is the data that is to be passed to the module for building its logic. After model is trained with the training data, testing data is passed to the model to check efficiency of algorithms. Here we have used 80% of the total credit card transactions for training purpose and remaining 20% of the transactions for testing purpose.

Selected 80% of training data is used to train fraud detection module, module defines its logic for dealing with further transactions, algorithms used can be Random Forest Algorithm or Local Outlier Factor, Testing Data is passed to the module as training of module is complete.

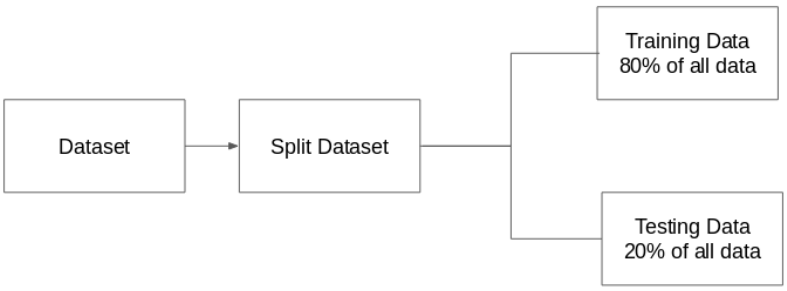


Figure 6: Splitting Of Dataset

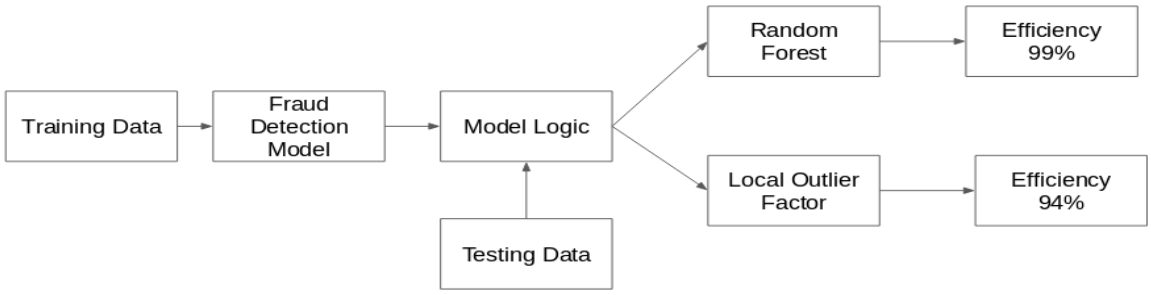


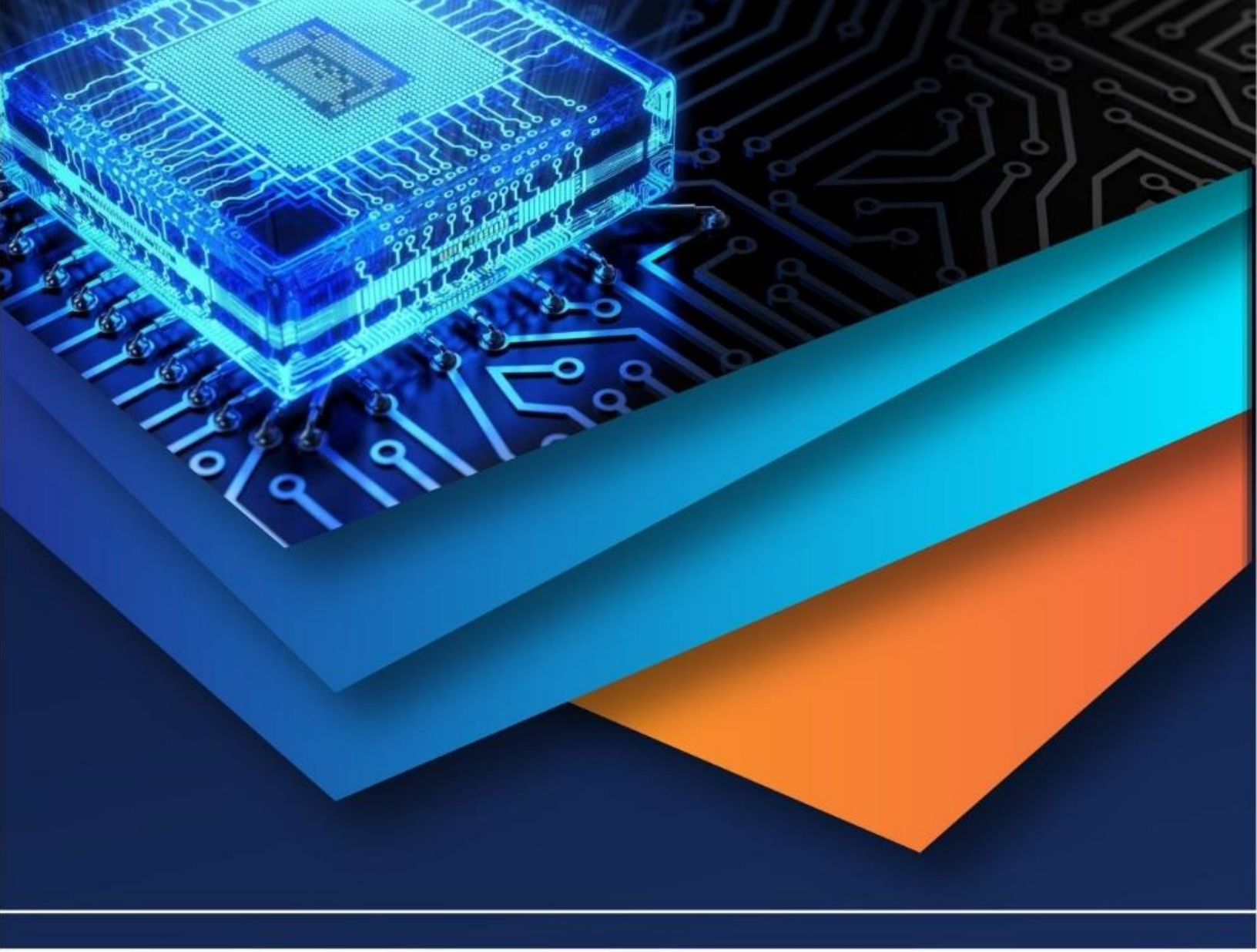
Figure 7: Method Used For Calculating Efficiency

VI. CONCLUSION

In this paper, two algorithms Random Forest Algorithm and Local Outlier Factor are compared for detecting fraudulent transactions from the given dataset. Random Forest Algorithm is better in detecting frauds than Local Outlier factor. Efficiency is 99% for Random Forest Algorithm and 94% for Local Outlier factor. Credit card fraud detection is efficient by both of these algorithms but every algorithm has its own specific advantages and disadvantages. Combining more than one algorithm will give higher efficiency.

REFERENCES

[1] Eesha Goel, Er. Abhilasha Computer Science & Engineering &GZSCCET Bhatinda,Punjab, India, "Random Forest: A Review", Volume 7, Issue 1, January 2017, International Journal of Advanced Research in Computer Science and Software Engineering
[2] AARON J. BODDY , WILLIAM HURST , MICHAEL MACKAY , AND ABDENNOUR EL RHALIBI" Density-Based Outlier Detection for Safeguarding Electronic Patient Record Systems", March 20, 2019
[3] Ms. D. R.Gupta, Prof. Dr. S. M. Kamalapur," A Review on Outlier Detection Techniques", International Journal for Research in Applied Science & Engineering Technology (IJRASET) *volume 5 Issue XII December 2017*
[4] Sahand Hariri, Matias Carrasco Kind, Robert J. Brunner,"Extended Isolation Forest" 6 Nov 2018
[5] Heta Naik,Credit Card Fraud Detection for Online Banking Transactions,International Journal for Research in Applied Science & Engineering Technology (IJRASET) Volume 6 Issue IV, April 2018
[6] Sunil Bhatia,Rashmi Bajaj,Santosh Hazari, Analysis of Credit Card Fraud Detection Techniques, International Journal of Science and Research (IJSR), March 2016
[7] Samaneh Sorournejad, Zahra Zojaji, Reza Ebrahimi Atani, Amir Hassan Monadjemi A Survey of Credit Card FraudDetection Techniques: Data and Technique Oriented Perspective, November 2016
[8] K. Veeramachaneni, I. Arnaldo, V. Korrapati, C. Bassias, and K. Li, ''AI2 : Training a big data machine to defend,'' in Proc. IEEE 2nd Int. Conf. Big Data Secur. Cloud (BigDataSecurity), IEEE Int. Conf. High Perform. Smart Comput. (HPSC), IEEE Int. Conf. Intell. Data Secur. (IDS), Apr. 2016, pp. 49–54.
[9] Zhang H, Wang M, (2009): Search for the smallest Random Forest, Statistics and Its Interface Volume.2, pp 381-388.
[10] E Tripoli, D Fotiadis, G Manis, 2010: "Dynamic Construction of Random Forests: Evaluation using Biomedical Engineering Problems", IEEE.
[11] KULDEEP RANDHAWA,CHU KIONG LOO, MANJEEVAN SEERA, CHEE PENG,ASOKE K. NANDI Credit Card Fraud Detection Using AdaBoost and Majority Voting, February 10, 2018



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY