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/IJSRSET

Credit Card Fraud Detection using Machine Learning

[*1]Revathi Simhadri, [2]Vemula Kalpana, [3] Martha Omseetha

[1]Associate Professor, Department of Information Technology, Bhoj Reddy Engineering College for Women, Hyderabad, India

[2&3,] Students, Department of Information Technology, Bhoj Reddy Engineering College for Women, Hyderabad, India

Article Info
Volume 9, Issue 5
Page Number : 321-326

Publication Issue :
September-October-2022

Article History
Accepted : 01 Oct 2022
Published: 22 Oct 2022

ABSTRACT

This Project is focused on credit card fraud detection in real world scenarios. Nowadays credit card frauds are drastically increasing in number as compared to earlier times. Criminals are using fake identity and various technologies to trap the users and get the money out of them. Therefore, it is very essential to find a solution to these types of frauds. In this proposed project we designed a model to detect the fraud activity in credit card transactions. This system can provide most of the important features required to detect illegal and illicit transactions. As technology changes constantly, it is becoming difficult to track the behavior and pattern of criminal transactions. To come up with the solution one can make use of technologies with the increase of machine learning, artificial intelligence and other relevant fields of information technology; it becomes feasible to automate this process and to save some of the intensive amounts of labor that is put into detecting credit card fraud. Initially, we will collect the credit card usage data-set by users and classify it as trained and testing dataset using a random forest algorithm and decision trees. Using this feasible algorithm, we can analyze the larger data-set and user provided current data-set. Then augment the accuracy of the result data. Proceeded with the application of processing of some of the attributes provided which can find affected fraud detection in viewing the graphical model of data visualization. The performance of the techniques is gauged based on accuracy, sensitivity, and specificity, precision. The results is indicated concerning the best accuracy for Random Forest are unit 98.6% respectively.

Keywords : Random Forest Algorithm, Criminal Transactions, Credit Card

I. INTRODUCTION

Nowadays Credit card usage has been drastically increased across the world, now people believe in going cashless and are completely dependent on online transactions. The credit card has made the digital transaction easier and more accessible. A huge number of dollars of loss are caused every year by the

criminal credit card transactions. Fraud is as old as mankind itself and can take an unlimited variety of different forms. The PwC global economic crime survey of 2017 suggests that approximately 48% of organizations experienced economic crime. Therefore, there's positively a necessity to unravel the matter of credit card fraud detection. Moreover, the growth of new technologies provides supplementary ways in which criminals may commit a scam. The use of credit cards is predominant in modern day society and credit card fraud has been kept on increasing in recent years. Huge Financial losses have been fraudulent effects on not only merchants and banks but also the individual person who are using the credits. Fraud may also affect the reputation and image of a merchant causing non-financial losses that. For example, if a cardholder is a victim of fraud with a certain company, he may no longer trust their business and choose a competitor. Fraud Detection is the process of monitoring the transaction behavior of a cardholder to detect whether an incoming transaction is authentic and authorized or not otherwise it will be detected as illicit. In a planned system, we are applying the random forest algorithm for classifying the credit card dataset. Random Forest is an associate in the nursing algorithmic program for classification and regression. Hence, it is a collection of decision tree classifiers. The random forest has an advantage over the 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 designated from a random subset of the complete feature set. Even for large data sets with many features and data instances, training is extremely fast in the 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 overfitting.

II. RELATED WORK

New methods for credit card fraud detection with a lot of research methods and several fraud detection techniques with a special interest in the neural networks, data mining, and distributed data mining. Many other techniques are used to detect such credit card fraud. When done the literature survey on various methods of credit card fraud detection, we can conclude that to detect credit card fraud there are many other approaches in Machine Learning itself. The research on credit card fraud detection uses both Machine Learning and Deep Learning algorithms. In this section, we enhance the work done in two different points:

(i) the methods that are readily available forfraud detection

(ii) The techniques that are available to handle the imbalanced data.

To handle the imbalanced data some of the techniques are available. They are

(a) classification methods

(b) sampling methods

(c) resembling techniques.

Here are some of the Machine Learning algorithms that are used for credit fraud detection are support vector machine(SVM), decision trees, logistic regression, gradient boosting, K-nearest neighbor, etc.In 2019, Yashvi Jain, NamrataTiwari, Shripriya Dubey, Sarika jain have researched various techniques[10] for credit cards fraud detection such as support vector machines(SVM), artificial neural networks(ANN), Bayesian Networks, Hidden Markov Model, K-Nearest Neighbours (KNN) Fuzzy Logic system and Decision Trees. In their paper, they have observed that the algorithms k-nearest neighbor, decision trees, and the SVM give a medium level accuracy. The Fuzzy Logic and Logistic Regression give the lowest accuracy among all the other algorithms. Neural Networks, naive bayes, fuzzy

systems, and KNN offer a high detention rate. The Logistic Regression, SVM, decision trees offer a high detection rate at the medium level. There are two algorithms namely ANN and the Naïve Bayesian Networks which perform better at all parameters. These are very much expensive to train. There is a major drawback in all the algorithms. The drawback is that these algorithms don't give the same result in all types of environments. They give better results with one type of datasets and poor results with another type of dataset. Algorithms like KNN and SVM give excellent results with small datasets and algorithms like logistic regression and fuzzy logic systems give good accuracy with raw and unsampled data.

In 2019, Heta Naik, Prashasti Kanikar, has done their research on various algorithms like Naïve Bayes, Logistic Regression, J48, and Adaboost. Naïve Bayes on among the classification algorithm. This algorithm depends upon Bayes theorem. Bayes's theorem finds the probability of an event that is occurring is given. The Logistic regression algorithm is similar to the linear regression algorithm. The linear regression is used for the prediction or forecasting the values. The logistic regression is mostly used for the classification task. The J48 algorithm is used to generate a decision tree and is used for the classification problem. The J48 is the extension of the ID3 (Iterative Dichotomieser). J48 is one of the most widely used and extensively analyzed areas in Machine Learning. This algorithm mainly works on constant and categorical variables. Adaboost is one of the most widely used machine learning algorithms and is mainly developed for binary classification. The algorithm is mainly used to boost the performance of the decision tree. This is also mainly used for the classification of the regression. The Adaboost algorithm is fraud cases to classify the transactions which are fraud and non-fraud. From their work they have concluded that the highest accuracy is obtained for both the Adaboost and Logistic Regression. As they have the same accuracy the time factor is considered to choose the best algorithm. By considering the time factor they concluded that the Adaboost algorithm works well to detect credit card fraud.

In 2019 Sahayasakila V, D.Kavya Monisha, Aishwarya, Sikhakolli Venkatavisalakshiswshai Yasaswi have explained the Twain important algorithmic techniques which are the Whale Optimization Techniques (WOA) and SMOTE (Synthetic Minority Oversampling Techniques). They mainly aimed to improve the convergence speed and to solve the data imbalance problem. The class imbalance problem is overcome using the SMOTE technique and the WOA technique. The SMOTE technique discriminates all the transactions which are synthesized are again resampled to check the data accuracy and are optimized using the WOA technique. The algorithm also improves the convergence speed, reliability, and efficiency of the system.

In 2018 Navanushu Khare and Saad Yunus Sait have explained their work on decision trees, random forest, SVM, and logistic regression. They have taken the highly skewed dataset and worked on such type of dataset. The performance evaluation is based on accuracy, sensitivity, specificity, and precision. The results indicate that the accuracy 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 concluded that the Random Forest algorithm has the highest accuracy among the other algorithms and is considered as the best algorithm to detect the fraud. They also concluded that the SVM algorithm has a data imbalance problem and does not give better results to detect credit card fraud.

III. PROPOSED SYSTEM

There are lots of issues that make this procedure tough to implement and one of the biggest problems associated with fraud detection is the lack of both the literature providing experimental results and of real-

world data for academic researchers to perform experiments on. The reason behind this is the sensitive financial data associated with the fraud that has to be kept confidential for the purpose of customer's privacy. Now, here we enumerate different properties a fraud detection system should have in order to generate proper results. The system should be able to handle skewed distributions, since only a very small percentage of all credit card transactions is fraudulent. There should be a proper means to handle the noise. Noise is the errors that is present in the data, for example, incorrect dates. This noise in actual data limits the accuracy of generalization that can be achieved, irrespective of how extensive the training set is. Another problem related to this field is overlapping data. Many transactions may resemble fraudulent transactions when actually they are genuine transactions.

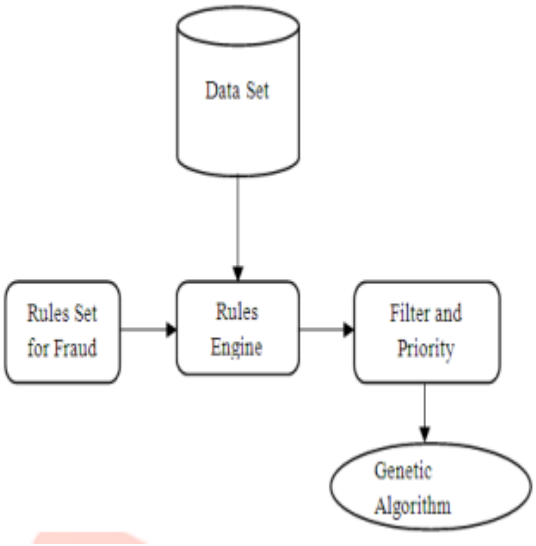

Fig 1: System Architecture

The above architectural design describes the work structure of the system. The data warehouse contains the customer data. This customer data is subjected to the rules engine and again ,the rules engine comprises of the rules set. The filter and priority module sets the priority for the data and hence. Plays a very important role in the system. then the filter data is sent to the genetic algorithm module which performance its functions and generates the output.

IV. RESULTS

In this section, we report our experimental study that we performed with selected machine learning algorithms and imbalance classification approaches. First, we provide a detailed description of the design of experiments followed by the results and a discussion. Finally, we discuss some critical shortcomings we discovered in our experiments.

Design of Experiments

This section briefly presents the workflow of our experiments, the dataset used, the selection of target variables and performance measure.

A. Workflow of Experiments

Our experimental study is organized as follows. The experiments are presented and discussed in two phases. In the first phase, eight classification methods are compared. The comparison was carried out with respect to three parameters including the following: accuracy, sensitivity, and the Area under PrecisionRecall Curve (AUPRC). This comparison results in selecting the most suitable algorithms including the SVM and ANN. In the second phase, the selected algorithms are used in comparing selected imbalance classification approaches such as Random Oversampling, One-Class Classification and Cost Sensitive. Then, the SVM is used as a binary classification tool, and compared to the One-Class Classification SVM and Cost Sensitive SVM. Also, the ANN is applied and compared to the Auto-Associative Neural Network.

B. Dataset and Variable Selection

The dataset used in our experiment contains credit card fraud labeled data. It contains ten million credit card transactions described by 8 variables listed here:

- ✓ Cust ID is an auto increasing integer value that represents the customer ID: This variable is removed later as it has no relevance for detecting fraud.
- ✓ Gender: represents the customer's gender.
- ✓ State: represents the state in which the customer lives in the United States.

- ✓ Card holder: is the number of cards that the customer holds (maximum 2).
- ✓ Balance : indicates the balance on the credit card in USD.

 Num Trans: is a discrete variable that represents the number of transactions made to date.
 NumInt Trans: is a discrete variable representing the number of international transactions made to date.
 Credit Line: denotes the customer's credit limit.
 Fraud Risk: the binary target variable, taking the values 0 denoting legitimate transaction, and 1 denoting fraudulent transaction.

V. CONCLUSION AND FUTURE SCOPE

Credit card fraud becomes a serious concern to the world. Fraud brings huge financial losses to the world. This urged Credit card companies have been invested money to create and develop techniques to reveal and reduce fraud. The prime goal of this study is to define algorithms that confer the appropriate, and can be adapted by credit card companies for identifying fraudulent transactions more accurately, in less time and cost. Different machine learning algorithms are compared, including Logistic Regression, Decision Trees, Random Forest, Artificial Neural Networks, Logistic Regression, K-Nearest Neighbors, and Kmeans clustering. Because not all scenarios are the same, a scenario-based algorithm can be used to determine which scenario is the best fit for that scenario.

VI. REFERENCES

[1] Anuruddha Thennakoon, Chee Bhagyani, Sasitha Premadasa,"Real-time Credit Card Fraud Detection Using MachineLearning",(2019) International Conference on Intelligent Computing and Control

[2] S. Xuan, G.. Liu, Z. Li, L. Zheng, S. Wang, and G. N. Surname, "Random Forest for credit card fraud," 15th Int. conf. networking, Sens. Control, 2018.

[3] Analysis Of Credit Card Fraud Detection Detection Techniques: Based On Certain Design Criteria Author:MasoumehZareapoor, Seeja.K.R, 2017.

[4] David Robertson,"Investments & Acquitions – September 2016 Top Card Issuers in Asia-Pacific Card Fraud Losses Reach,"Nilson Rep., no.1096, 1090.

[5] M.Zareapoor, S.K..Seeja.K.R, and M.Asshar Alam,"Analysis on Credit Card Fraud Detection Techinqes: Based on Certin Design Criteria,"Int. J. Comput.Appl., vol.52,no.3,pp.35-42,2012.

[6] Michael Edward Edge, Pedro R. Falcone Sampaio, "A Survey Of Signature Based Methods For Financial Fraud Detection", 2012,

[7] J. O. Awoyemi, A.O.Adetunmbi, and S.A Oluwadare, Credit Card fraud detection using machine learning techniques: A comparative Analysis,"2017 Int. Conf. Comput. Netw. Informatics, pp. 19,2017.

[8] HetaNaik,PrashastiKanikar," Credit Card Fraud Detection Based On the Machine Learning Algorithms", 2011

[9] M.rafalo,"Real time fraud detection in credit card transactions,"Data ScienceWarsaw.2017.

[10] Salvatore J. Stolfo, David W. Fan, Wenke Lee AndAndreas L.Prodromidis, " Credit Card Fraud Detection Using MetaLearning: Issues And Initial Results".

[11] Real-Time Credit Card Fraud Detection Using Streaming Analytics Author: Rajeshwari U, B SathishBabu

[12] Survey Of Credit Card Fraud Detection Techniques: Data And Technique Oriented Perspective Author: Samanehsorournejad , Zahra Zojaji, Reza

[13] A.Dal pozzolo,G.boracchi,O.Caelan, and C.Alippi,"Credit Card fraud detyection:A Realistic Modeling and A novel learning strategy,"Ieee trans.Neural Networks learn.Syst.,pp.1- 14,2018.

Cite this article as :

Revathi Simhadri, Vemula Kalpana, Om seetha, "Credit Card Fraud Detection using Machine Learning ", International Journal of Scientific Research in Science, Engineering and Technology (IJSRSET), Online ISSN : 2394-4099, Print ISSN : 2395-1990, Volume 9 Issue 5, pp. 321-326, September-October 2022.

Journal URL : https://ijsrset.com/IJSRSET229645