iJRASET

International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 9 Issue: V Month of publication: May 2021

DOI: https://doi.org/10.22214/ijraset.2021.34241

Comparison of Logistic Regression and Decision Tree method for Credit Card Fraud Detection

Tulika Kumar[1], Anusha Lathi[2], Yukti Mathur[3], Prateeksha Shanoj[4]
[1]Department of Electronics and Telecommunication, MPSTME NMIMS, Pune, India
[2]Department of Electronics and Telecommunication, MPSTME NMIMS, Bhilwara, India
[3]Department of Electronics and Telecommunication, MPSTME NMIMS, Gotan, India
[4]Assistant Professor EXTC Dept MPSTME NMIMS, Mumbai India

Abstract: The use of Credit cards has drastically increased as they become one of the vital and most used modes of payment. Along with this the Credit Card. Fraud has also grown to a greater extent. Thus, it becomes vital for Credit card companies or banks to identify fraudulent transactions.

Machine learning algorithms are used to analyze and identify suspicious transactions. Credit Card Fraud detection is a typical case of classification and data analysis, in this, we have focused on analyzing and comparing the data on parameters like accuracy, precision on the basis of two supervised learning classification algorithm Logistics Regression and Decision tree.

Keywords: Logistic Regression, Decision Tree, Credit card Fraud

I. INTRODUCTION

'Fraud' in credit card transactions is unauthorized and unwanted use of an account by someone other than the owner of that account. Required preventive steps can be taken to avoid this abuse and the behaviours of such fraudulent activities can be observed to mitigate it and protect against similar incidents in the future. Frauds come in a variety of forms. If someone forges a check or pays for something with a check knowing there isn't enough money, they are committing fraud. Sales from internet are fraud in which fraudsters sell false goods or counterfeit products or take payment without delivering the product. There are a few more, including charity fraud, identity theft, credit card fraud, debt elimination. Credit card fraud is one of the most prevalent frauds, due to the growing prevalence of cashless transactions. If the credit card issuer is unaware of the fraud, it is referred to as credit card fraud. The person involved in fraud uses the credit card for their needs. In 2018, fraudulent purchases carried out using credit cards obtained worldwide grew by 18.4 percent and are still growing. There are two forms of fraud involving credit cards. One is physical card theft, and another is stealing confidential card details, such as the number of the card, card verification value, type of the card, and others. A large sum of money may be taken by the person doing the fraud.or make a large amount of payment by stealing credit card data before the cardholder finds out. Businesses use different machine learning techniques because these companies use various methods to identify fraudulent cases. In this paper firstly, we analysed various credit card fraud detection techniques like Logistic Regression (LR), Random Forest (RF), Naïve Bayes (NB), Multilayer Perceptron (MLP), Decision Tree by reading different research paper and performing a rigorous literature survey. As a result of this, it was found that logistic regression and decision tree are most widely used because of various parameters which will be focused on further in this paper.

II. EXPERIMENTAL SETUP AND WORKS

A. Logistic Regression

Logistic Regression is a classification supervised technique which is used when the response variable is categorical. It is just like Linear Regression. It uses a logistic curve for fraud detection. It is used to predict the probability outcome either true or false, yes or no and one or zero i.e., it has two outcomes, here in case of credit card fraud detection it is fraudulent transaction/ legitimate transaction.

$$y = e^{(bo+b1*x)}/ (1 + e^{(b0+b1*x)})$$

1) Dataset is imported with the additional libraries to the notebook for the in-depth analysis.

2) Data Cleaning is a process where inaccurate records or missing data is identified and then corrected, replaced with some relevant data or modified or deleted if irrelevant.

3) Preprocessing Data: If the data is accurate preprocessing data then the model is likely to give good relevant results and when compared to with a model for which data isn't well preprocessed.

It is further having four steps for preprocessing of data:

a) *Train test Split:* Here the original dataset is divided into 2 parts i.e., train set and test set where both consist of data and labels.

b) *Taking Care of Missing Values:* If we have NaN values or garbage values, then the model will surely be garbage too. Therefore, considering missing values is important.

c) Takin Care of Categorical Features:

d) *Normalization Data:* It is better on a normalized data rather than a dataset not normalized.

The main of normalization is changing values to a common scale without distorting the difference between the range of values.

4) *Data Modelling of Imbalanced data directly:* Imbalanced data is a problem of classification observation per class is not uniformly distributed. Often, we have a majority class which has a larger number of data and other is few observations as a minority class.

5) *Changing Threshold:* Default value for normalized predicted score is 0.5.

6) *Creating Over-Sampling Data And Fitting The Model:* Oversampling incorporates choosing random models from the minority class with substitution training data with different duplicates and hence it is possible that a single instance may be selected multiple times. So, after oversampling , the model is fitted into the dataset.

7) *Assigning Weights To The Model Class:* it is done to lower the rate class. If the class weight is set for the positive as the ratio of non-Fraud / Fraud, the result is approximately close to the over-sampling.

8) *Final Result:* We get different parameter’s results.

B. Decision Tree

Decision trees and their classes are popular algorithms for machine learning tasks of classification and regression. Decision trees are commonly used because they are simple to use, interpret, treat categorical characteristics, apply to the setting for multiclass classification, feature scaling is not necessary, and non-linearities and features can be captured easily.

Tree ensemble algorithms, such as random forests and boosting, for classification and regression tasks, are among the top performers. Because of these factors, decision trees often perform well on top of rules-based models in terms of facets and are often as a strong point of entry for fraud detection.

1) First of all, the data obtained is pre-processed before the modeling begins. The models are then able to learn the characteristics of both the ordinary and fraud profile of the transaction by using stratified sampling to under sample normal data.

2) To implement this, the most critical variables for distinguishing between fraudulent and legitimate transactions are observed.

3) The parameters are hence used to shape stratified samples of the legitimate transactions.

4) Subsequently, these stratified samples of genuine transactions are combined with the fraudulent ones to form three samples with different fraudulent to normal record ratios. In fraud detection systems, the factors that form the card user profile, as well as the modeling techniques used, make the difference.

5) Our purpose in defining the variables used to form the data-mart is to differentiate the fraudulent card usage profile of the fraudsters from the card holders.

C. Comparative Study

A comparison of the usage pattern and transactions leads to classifying a transaction as fraud or legitimate in this paper [1]. The techniques used are KNN, Naïve Bayes, Logistic Regression, Chebyshev functional Link Artificial Neural Network (CFLANN), Multi Layer Perceptron and Decision Trees which are evaluated on basis of their result evaluated in terms of various accuracy metrics.It can be concluded in this paper that KNN is having highest accuracy, highest sensitivity and highest specificity. is shown by KNN and Naive Bayes.

In this paper [2], some machine learning algorithms were used to compare in which gave the best results, according to state of art but applied to the same set of data. The objective of this study is to choose the best credit card fraud detection techniques to implement in our future work.

In this work [3], a high imbalanced dataset was considered and then some known supervised and unsupervised machine learning techniques were applied to detect credit card fraudulent transactions. The best results were given by unsupervised machine learning algorithms as it could handle the skewness.

Different metrics were used to evaluate a number of algorithms in supervised a number of popular algorithms in supervised, ensemble and even unsupervised categories. So, to conclude that rather than data metric,supervised techniques have got a higher edge to handle dataset skewness in excellent ways and perform extremely well over other algorithms. Few NaN values were there in the result table and the classifier couldn't detect even a single true positive or true negative value. Therefore, while dealing with highly imbalanced datasets, unsupervised, especially IF and LOF are the overall the best.

A quantitative understanding is done in this paper [4] based on its suitability for use in to credit card fraud detection technique. The researchers prepared the appropriate dataset which will be used to feed into the machine learning algorithms with the help of under sampling method. The datasets were with "Time" and without "Time" attributes. It can be concluded in this paper that Logistic Regression has a better precision and accuracy even with time and without time

In this paper [5] KNN method can be used to determine the anomaly of the target instance by performing over sampling and extracting the principal direction of the data. As a result, the KNN method can be used to detect fraud when memory is limited, in the meantime, the outlier identification mechanism aids in the detection of credit card fraud while requiring less memory and computation. Outlier detection, in particular, works quickly and effectively.

This paper [8] aims to perform comparative analyses of recognizing of fraudulent activity on credit card using support vector machine, k-nearest neighbour technique, naïve bayes and logistic regression techniques with biased information, depending on the reliability to determine the most reliable method of classifying a credit card transaction as fraudulent or non-fraudulent by which some algorithm and combination of factors are considered. Logistic Regression is the most accurate technique in detecting fraudulent activity when tested under realistic conditions. With an accuracy of 99.074% and a precision of 93.61% thus makes it a suitable technique out of the four used from this paper.

Advantages and disadvantages of the techniques are stated as conclusion in this paper [9] ,from this we can understand that decision tree, rule based methods and CNN were easy to implement and require less training whereas artificial neural networks and HMM were very slow to train. Require a lot of power. Hard to interpret. Decision tree though being easy to implement and widely used has a disadvantage of not being able to handle complex data. The objective of this paper was to give the advantage and disadvantages of all techniques mentioned but does not prove a clear conclusion about using which technique.

III. RESULTS

In this paper, two machine learning techniques were used to predict the fraudulent transactions. So, to evaluate both the methods, 70% of the dataset is used as a training set and the remaining 30% is used as a test dataset. Four parameters are calculated as follows: accuracy, precision, F1-sore and recall are used to test their evaluation of the performance. We can observe that the Decision Tree has a better accuracy than Logistic Regression. Precision of Decision Tree is much better than Logistic Regression. Similarly, F1 score and Recall of Decision tree is good when compared to Logistic Regression.

A. Comparison Table

TABLE I. Comparison Of Techniques Used

Technique used	Parameters			
	Accuracy	*Precision*	*F1-Score*	*Recall*
Logistic Regression	0.99926	0.81720	0.78351	0.75248
Decision Tree	0.99947	0.88172	0.84536	0.81188

Sample of a Table footnote. (Table footnote)

IV. CONCLUSION AND FUTURE SCOPE

Credit card fraud is a very serious concern around the globe. From the above comparative analysis of the two techniques, it is clear that Decision Tree performs the best. But the drawbacks of this paper by using the above techniques is that we cannot identify before the fraud happens, so this can be prevented. It tells us after fraud has happened. For further development of this project, we can work to solve this problem by designing an Alert-based notification system that will ask the user to accept or decline the payment process.

REFERENCES

[1] D. Dighe, S. Patil and S. Kokate, "Detection of Credit Card Fraud Transactions Using Machine Learning Algorithms and Neural Networks: A Comparative Study," 2018 Fourth International Conference on Computing Communication Control and Automation (ICCUBEA)

[2] I. SADGALI, N. SAEL and F. BENABBOU, "Fraud detection in credit card transaction using machine learning techniques," 2019 1st International Conference on Smart Systems and Data Science (ICSSD)

[3] Sangeeta Mittal and Shivani Tyagi, " Performance Evaluation of Machine Learning Algorithms for Credit Card Fraud Detection" 2019 IEEE 9th International Conference on Cloud Computing, Data Science & Engineering (Confluence)

[4] Samidha Khatri, Aishwarya Arora and Arun Prakash Agrawalr, "Supervised Machine Learning Algorithms for Credit Card Fraud Detection: A Comparison" 2020 IEEE I10th International Conference on Cloud Computing, Data Science & Engineering (Confluence)

[5] Krishna Modi and Reshma Dayma "Review On Fraud Detection Methods in CreditCard Transactions " 2017 International Conference on Intelligent Computing and Control (I2C2'17)

[6] S.Rajora et al., "A Comparative Study of Machine Learning Techniques for Credit Card Fraud Detection Based on Time Variance," 2018 IEEE Symposium Series on Computational Intelligence (SSCI)

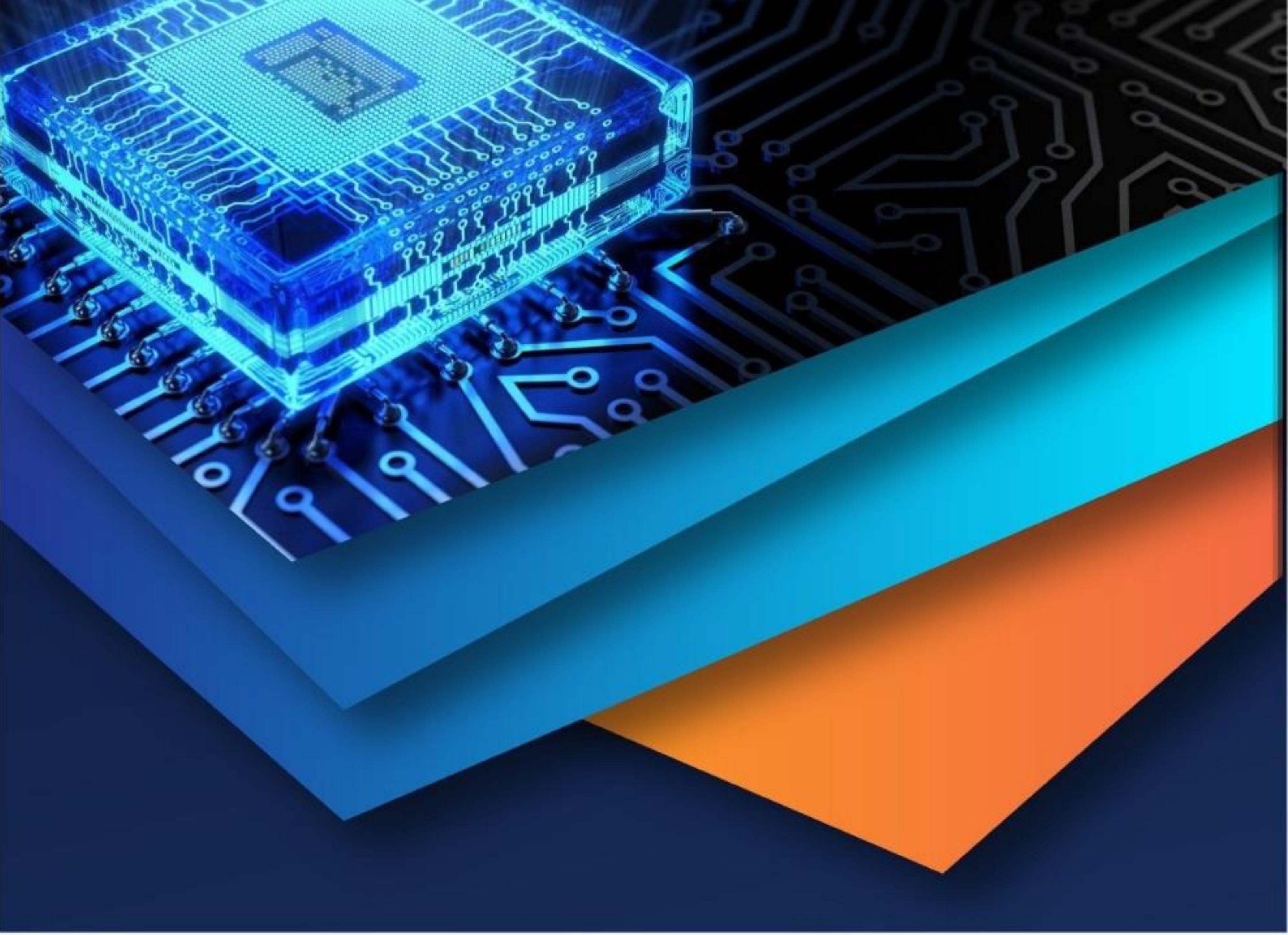
[7] Olawale Adepoju, Julius Wosowei, Shiwani lawte, Hemaint Jaiman. "Comparative Evaluation of Credit Card Fraud Detection Using Machine Learning Techniques", 2019 Global Conference for Advancement in Technology, (GCAT), 2019S

[8] N.Malini and Dr.M.Pushpa"Analysis on Credit Card Fraud Identification Techniques based on KNN and Outlier Detection"

[9] IbtissamBenchaji and Samira Douzi "Using Genetic Algorithm to Improve Classification of Imbalanced Datasets for credit card fraud detection" 2018 Cyber Security in Networking Conference (CSNet-2018)

[10] 3rd International Conference on Advances in Electrical, Electronics, Information, Communication and Bio-Informatics (AEEEICB17

[11] Dhankhad, E. Mohammed and B. Far, "Supervised Machine Learning Algorithms for Credit Card Fraudulent Transaction Detection: A Comparative Study," 2018 IEEE International Conference on Information Reuse and Integration (IRI)



10.22214/IJRASET

INDEX COPERNICUS
45.98

IMPACT FACTOR:
7.129

TOGETHER WE REACH THE GOAL
IMPACT FACTOR:
7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY