wjert, 2021, Vol. 7, Issue 5, 170-179. Research Article ISSN 2454-695X

World Journal of Engineering Research and Technology

WJERT

www.wjert.org

SJIF Impact Factor: 5.924

CREDIT CARD FRAUD DETECTION USING MACHINE LEARNING ALGORITHMS

Dr. S. Anitha*[1] and Dr. N. Sridevi[2]

[1]Assistant Professor, Department of Computer Science, Avinashilingam Institute for Home Science and Higher Education for Women, Coimbatore, India.

[2]Assistant Professor, College of Computer Studies, University of Technology Bahrain, Kingdom of Bahrain.

Article Received on 23/06/2021 **Article Revised on 13/07/2021** **Article Accepted on 03/08/2021**

***Corresponding Author**
Dr. S. Anitha
Assistant Professor, Department of Computer Science, Avinashilingam Institute for Home Science and Higher Education for Women, Coimbatore, India.

ABSTRACT

Use of credit card urges people to purchase items online through the Internet. People will in general do a lot of buying on the web or in person by using the credit card. Charge cards have ended up being the most conspicuous office accessible to the individuals around the world to energize paperless exchanges at a colossal speed. At whatever point any such exchange occurs in trades or net advertising by utilizing a paperless structure, it is oppressed under high danger of fake exchanges because of numerous traps in the security arrangement of the utilization of credit cards on the systems. In this paper, supervised learning algorithms namely Logistic Regression, Decision Tree and Random Forest algorithms are involved to discover if the credit card transactions are fake or not. Performance evaluation metrics like precision, recall, F1-score and Matthews Correlation Coefficient are used to measure the performance of the algorithms. After the experimental outcomes, it is witnessed, the Random Forest algorithm outperforms better when compared to the other two algorithms.

KEYWORDS: Credit card fraud detection, Logistic regression, Decision tree, Random Forest, Machine Learning algorithms.

1. INTRODUCTION

A credit card is a payment card given to clients (cardholders) to empower the cardholder to

pay a vendor for merchandise and enterprises dependent on the cardholder's guarantee to the card backer to pay them for the sums in addition to the next concurred charges. The card guarantor (typically a bank) makes a rotating record and awards a credit extension to the cardholder, from which the cardholder can get cash for installment to a dealer or as a loan. With the board use of internet business and web based shopping, the user's credit card passwords, cvv numbers and other imperative data are consistently powerless. The fraud users effectively split into the essential data and subsequently misrepresentation cases are on the ascent. The financial frameworks likewise are defenseless against online fake conduct of fraud users. The fraud in online networking is on ascend as the quantity of misrepresentation cases are getting expanded. As the fraudsters are finding new roads or ways for misrepresentation, the fraud avoidance is a consistent developing procedure. Because of this, the conventional method of dealing with deceitful conduct is gradually getting supplanted with online fraud identification programming utilizing various data mining algorithms.

Machine Learning is a part of Artificial Intelligence. As original information is given to these calculations, they study and upgrade their errands to increase execution, making 'information' after some time. Presently, four categories of machine learning algorithms: supervised, semi-supervised, unsupervised and reinforcement are available. In supervised learning, the system is trained by model. The overseer outfits the Machine Learning calculation with a well-known dataset that joins needed sources of inputs and outputs, and the calculation must find a technique to conclude how to appear at those information sources and outputs. While the director knows the correct reactions to the issue, the calculation perceives designs in information, gains from discernments and makes desires. The calculation makes conjectures and is balanced by the chairman – and this strategy continues until the calculation achieves a critical degree of exactness or execution. Classification, Regression and Forecasting are the three subtypes of supervised learning calculations. In Classification assignments, the Machine Learning techniques must make a determination from watched esteems and decide to what classification new perceptions have a place. In this paper three classification algorithms namely Logistic Regression, Decision tree, Random Forests are used to detect credit card fraud detection.

2. RELATED WORK

In the paper they have clarified the idea of fraud identified with cards. The creators have executed diverse AI calculations like logistic regression, naive Bayes, random forest with

ensemble classifiers using boosting techniques on an imbalanced dataset. Distinctive characterization models are applied to the information and the exhibition of the calculations are assessed based on exactness, accuracy, review, f1 score and confusion matrix. From the test results the creators presumed that supervised techniques give better results.[1]

The specialist's fundamental point is to plan and build up a novel misrepresentation discovery strategy for data, with a target to dissect the previous exchange subtleties of the clients and concentrate the personal conduct standards. Sliding window methodology is utilized to bunch the cardholders into various gatherings dependent on their exchange sum. At that point various classifiers are applied over the prepared information and the classifier with better appraising score is picked to be probably the best technique to foresee cheats. European charge card misrepresentation dataset is utilized in their paper.[2] The paper portrays the likelihood of fake exchanges in commonness and setting of Visa use.[3]

The analyst's made another model which is utilized to perceive whether another exchange is deceitful or not. The primary target of their work is to identify 100% of the false exchanges while limiting the wrong extortion groupings. Here they have conveyed various abnormality recognition calculations, for example, Local Outlier Factor and Isolation Forest calculation on the PCA changed charge card exchange information.[4]

A half and half strategy called "AdaBoost and majority vote technique is applied to the charge card dataset. The exhibition of this half breed technique is contrasted and the standard model. The mixture technique gives great exactness appraised in charge card misrepresentation recognition.[5]

A review about the direct and non-straight AI calculations used to anticipate the false exchanges by contemplating the examples of the charge card exchanges. The creators utilized Random Forest, Support Vector Machine and Artificial Neural Network classifiers to order a concealed Visa exchange is false or not.[6]

Relapse calculations are based on AI and are utilized to distinguish Mastercard misrepresentation. The test results show calculated relapse based methodologies gives most elevated exactness and it tends to be successfully utilized for extortion specialists.[7]

3. METHODOLOGY

The Classification algorithm is a Supervised Learning technique that is used to identify the category of a new data on the basis of training data. One of the key structures of supervised learning algorithms is that they model dependencies and associations between the objective output and input types to forecast the value for new data. The system learns from the given dataset and updates its knowledge about the classes. It performs the task of classification for new data by categorizing it based on the features during the training phase. The Figure:1 depicts the important steps involved in the classification process. For the research work three supervised algorithms namely Logistic Regression, Decision Tree and Random Forest Algorithm were considered.

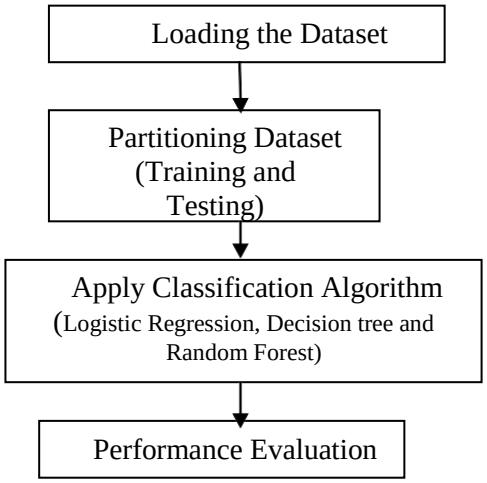

Figure 1: Steps involved in Classification.

3.1 Logistic Regression Algorithm

Logistic regression is one of the generally utilized Supervised Learning strategies. It is utilized for estimating the objective variable utilizing a given arrangement of predictor factors. It is utilized to foresee the likelihood of a dependent variable where the needy variable is a binary factor that contains information coded as 1 for progress or 0 for failure and will have just two potential results. The objective of logistic regression is to locate a best-fitting connection between the dependent variable and independent variable. Logistic regression is explicitly appropriate for binary grouping issues and can likewise be reached out to multi-class classification issues.

Pseudo code of Logistic Regression

1. The dependent variable is categorical: $y \epsilon \{0, 1\}$

2. A binary dependent variable can have only two values, like 0 or 1
3. In this case, the probability distribution of output y as 1 or 0. This is called the sigmoid probability (σ).
4. If $\sigma(\theta Tx) > 0.5$, set y = 1, else set y = 0
5. Maximum likelihood estimation is used to find the optimal weights of Logistic Regression.
6. It can be used to compute the probability of a given outcome in a binary model, like the probability of being classified as honest or fraud credit card user.

The curve from the logistic function indicates the likelihood of whether the person is fraud or not.

3.2 Decision Tree Algorithm

The decision tree algorithm builds the grouping model as a tree structure where leaf hub relates to a class mark and characteristics are speaks to an inner hub. It applies if-then rule that decides that are similarly broad and totally unrelated in grouping. The procedure goes on with separating the information into littler structures and in the long run connecting it with a steady choice tree. The last structure resembles a tree with hubs and leaves. Each time a standard is found out consecutively utilizing the preparation information each in turn and the tuples covering the principles are expelled. The procedure proceeds on the preparation set until the end point is fulfilled.

Pseudo code of Decision Tree Algorithm

1. Begin the tree with the root hub R, which contains the total dataset.
2. Find the best property in the dataset utilizing Attribute Selection Measure.
3. Divide the R into subsets that contain potential qualities for the best characteristics.
4. Generate the decision tree hub, which contains the best property.
5. Recursively settle on new decision trees utilizing the subsets of the dataset made in step - 3.
6. Proceed with this procedure until a phase is arrived at where you can't further classify the hubs and called the last hub as a leaf hub.

Decision tree speaks to every single imaginable answer for an issue dependent on given conditions. It makes a training model which can use to foresee class of target factors by taking in choice guidelines surmised from preparing information. They are reasonable for

classification issues where characteristics or highlights are efficiently checked to decide a last classification.

3.3 Random Forest Algorithm

Random forest is a directed learning algorithm which is utilized for both classification and regression. It is broadly utilized for classification issues. For the most part when a forest is viewed as it is comprised of trees and more trees implies more vigorous forest. So also, random forest algorithm makes decision trees on information tests and afterward gets the expectation from every one of them lastly chooses the best arrangement by methods for casting a voting. It is a group technique which is better than a solitary decision tree since it decreases the over-fitting by averaging the outcome.

Pseudo code of Random Forest Algorithm

1. Start with the determination of random examples from a given dataset.
2. Construct a decision tree for each example. At that point it will get the forecast outcome from each decision tree.
3. Voting will be performed for each anticipated outcome.
4. Finally select the most casted a ballot forecast result as the last expectation result.

Random forest adds extra randomness to the model, while developing the trees. Rather than looking for the most significant component while parting a hub, it scans for the best element among an irregular subset of highlights and in this manner it makes more adaptable.

4. RESULTS AND DISCUSSION

In this paper Python is used to implement the classification algorithms where it is capable of extracting the necessary knowledge from given data automatically. Python has become most popular language for machine learning due to some of its features like

- Python is Easy To Use
- Python has multiple Libraries and Frameworks
- Python has Community and Corporate Support
- Python is portable and Extensible

Python enables machine learning algorithms concrete by implementing them with a user-friendly, well-documented, and robust library. It includes two packages namely pandas and sklearn for data manipulation and algorithm implementation.

4.1 Dataset Description

The dataset contains transactions made by a cardholder for two days in the month of September 2013. The dataset contains only numerical input variables which are the result of a PCA transformation. To maintain the confidentiality of the transactions features V1, V2… V28 are transformed using the principal components, the only features which have not been transformed with PCA are 'Time' and 'Amount'. Feature 'Class' is the response variable and it takes value 1 in case of fraud and 0 otherwise.

4.2 Performance Evaluation Metrics

In the proposed methodology, the performances of the machine learning algorithms are evaluated using Confusion Matrix, Precision, Recall, F1-Score and Support. These metrics are always considered as the base parameter to evaluate any machine learning algorithms Precision quantifies the number of positive class predictions that actually belong to the positive class.

$$Precision = \frac{TP}{TP + FP}$$

Recall quantifies the number of positive class predictions that actually belong to the positive class.

$$Recall = \frac{TP}{TP + FN}$$

F1- Score also known as F-Score or F-Measure is the weighted average of Precision and Recall. The formula used to calculate the F1-Score is given below

$$F1 - Score = 2.\frac{Precision \times Recall}{Precision + Recall}$$

Support represents the number of samples of the true response that lies in each class of target values. Finally confusion matrix provides the performance of a predictive model with the information such as correct, incorrect prediction and the type of errors made in prediction. Since the dataset used in this paper has highly imbalanced data, precision, recall and F1-score will not be enough to measure the performance of the machine learning algorithms. Hence Matthews Correlation Coefficient (MCC) is used in this paper to evaluate the performance. MCC is a machine learning measure which is used to check the balance of the binary (two-class) classifiers. It takes into account all the true and false values that is why it is generally regarded as a balanced measure which can be used even if there are different classes. The

formula used for MCC is

$$MCC = \frac{TP \times TN - FP \times FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}}$$

4.3 Experimental Results

Confusion Matrix for the Logistic Regression, Decision tree and Random forest algorithms are given in the Table 1, as the True Positive and True Negative rate is high, the accuracy level of prediction is also high. While comparing the values of TP, TN, FP and FN in the confusion matrix Random forest algorithm is found better in identifying fraudulent and genuine customers.

Table 1: Confusion Matrix of Classification Algorithm.

Classification Algorithm	Actual Class	Predicted Class	
		Class 1	Class 0
Logistic Regression	Class 1	85278	30
	Class 0	55	80
Decision Tree	Class 1	85280	36
	Class 0	30	97
Random Forest	Class 1	85283	16
	Class 0	58	86

The experimental results of other evaluation metrics are given in Table 2. An F1 score is considered perfect when it's 1, while the model is a total failure when it's 0. From the table 2, we can see that the F1-Score of the Random Forest algorithm for class 1(i.e. fraud customers). is closer to 1 when compared to other algorithms. Hence it is proven that Random forest algorithm performs better in identifying the Class 1 customers.

Table 2: Performance Evaluation Metrics.

Classification Algorithm	Prediction	Precision	Recall	F1-Score	Support
Logistic Regression	Class 0	1.00	1.00	1.00	85308
	Class 1	0.73	0.59	0.65	135
Decision Tree	Class 0	1.00	1.00	1.00	85299
	Class 1	0.84	0.60	0.70	144
Random Forest	Class 0	1.00	1.00	1.00	85316
	Class 1	0.73	0.76	0.75	127

Table 3, shows the values of Matthews Correlation Coefficient obtained for various classification algorithms.

Table 3: MCC values for Classification Algorithms (MCC).

Classification Algorithms	MCC
Logistic Regression	0.6560022514861822
Decision Tree	0.7504193023003802
Random Forest	0.7643173040414987

The higher the correlation between true and predicted values, the better the prediction. Hence from the experimental results it is shown that the Random Forest algorithm performs better in classifying the fraud customers when compared to other two algorithms.

5. CONCLUSION

In recent years, credit card utilization has broadened significantly. Deceitful tasks and abuse of the credit card is very regular in numerous nations. Numerous methods have been accounted for in the writing to distinguish false exchanges that are performed after credit card robbery. Finding the most productive strategy to recognize the misrepresentation from a credit card is under testing. In any case, it is conceivable to check the quantity of fakes through exact analysis of the information gathered from a few credit card exchanges. In this paper, three supervised learning algorithms namely Logistic Regression, Decision Tree and Random Forest are considered for credit card fraud detection. The performances of the algorithms are evaluated using the evaluation metrics such as Confusion matrix, Precision, Recall, F1-Score and Mathews Correlation Coefficient. From the experiment results of the performance evaluation metrics it is clear that the Random Forest classification algorithm outperforms the other two algorithms in credit card fraud detection.

REFERENCES

1. Bhanusri, A., Valli, K. R. S., Jyothi, P., Sai, G. V., & Rohith, R (2020),” Credit card fraud detection using Machine learning algorithms”. *Journal of Research in Humanities and Social Science,* Volume 8, Issue 2 pp.: 04-11, ISSN (Online): 2321-9467.
2. Dornadula, V. N., & Geetha, S. (2019),”Credit Card Fraud Detection using Machine Learning Algorithms”, *Procedia Computer Science, 165:* 631-641.
3. Kaithekuzhical Leena Kurien & Ajeet Chikkamannur(2019). Detection and Prediction of Credit Card Fraud Transactions Using Machine Learning. *International Journal of Engineering Sciences & Research Technology*, Volume 8, Issue 3, pp.: 199-208, ISSN: 2277- 9655.
4. Maniraj, S. P., Saini, A., Sarkar, S. D., & Ahmed, S. Credit Card Fraud Detection using

Machine Learning and Data Science. *International Journal of Engineering Research & Technology (IJERT),* Vol. 8 Issue 09, pp.:110-115, ISSN: 2278-0181.

5. Ramyashree. K, Janaki K, Keerthana. S, B.V. Harshitha, Harshitha. Y.V (2019), "A Hybrid Method for Credit CardFraud Detection Using Machine Learning Algorithm", *International Journal of Recent Technology and Engineering (IJRTE),* Volume-7, Issue-6S4, ISSN: 2277- 3878.
6. Vidyashree V, Akram Pasha, Udayarani V, Vinay Kumar M (2019). Machine learning Classifiers for Credit Card Fraud Detection: A Brief Survey. *International Journal of Computer Sciences and Engineering*, Vol.-7, Special Issue- 14, E-ISSN: 2347-2693.
7. Suryanarayana, S. V., Balaji, G. N., & Rao, G. V. (2018). Machine Learning Approaches for Credit Card Fraud Detection. *Int. J. Eng. Technol*, *7*(2): 917-920.