ijar
ISSN NO. 2320-5407

Journal Homepage: -www.journalijar.com

INTERNATIONAL JOURNAL OF ADVANCED RESEARCH (IJAR)

Article DOI:10.21474/IJAR01/16824
DOI URL: http://dx.doi.org/10.21474/IJAR01/16824

RESEARCH ARTICLE

CREDIT CARD FRAUD DETECTION USING MACHINE-LEARNING

Prateeksha M.S[1], B. Naga Swetha[1] and Manjula Patil[2]

1. Dept of ISE, RYM Engineering College, Ballari.
2. Professor, Dept of ISE, RYM Engineering College, Ballari.

Manuscript Info

Manuscript History
Received: 28 February 2023
Final Accepted: 31 March 2023
Published: April 2023

Key words:-
Machine Learning, Genetic Algorithm, Fraud Detection, Cybersecurity

Abstract

The recent advances of e-commerce and e-payment systems have sparked an increase in financial fraud cases such as credit card fraud. It is therefore crucial to implement mechanisms that can detect the credit card fraud. Features of credit card frauds play important role when machine learning is used for credit card fraud detection, and they must be chosen properly. This paper proposes a machine learning (ML) based credit card fraud detection engine using ML classifiers: Decision Tree (DT), Logistic Regression (LR), Artificial Neural Network (ANN). To validate the performance, the proposed credit card fraud detection engine is evaluated using a dataset generated from European cardholders. The result demonstrated that our proposed approach outperforms existing systems.

Copy Right, IJAR, 2023,. All rights reserved.

Introduction:-

In the last decade, there has been an exponential growth of the Internet. This has sparked the proliferation and increase in the use of services such as e-commerce, tap and pay systems, online bills payment systems etc. As a consequence, fraudsters have also increased activities to attack transactions that are made using credit cards. There exists a number of mechanisms used to protect credit cards transactions including credit card data encryption and tokenization. Although such methods are effective in most of the cases, they do not fully protect credit card transactions against fraud. Machine Learning (ML) is a sub-field of Artificial Intelligence (AI) that allows computers to learn from previous experience (data) and to improve on their predictive abilities without explicitly being programmed to do so. In this work we implement Machine Learning (ML) methods for credit card fraud detection. Credit card fraud is defined as a fraudulent transaction (payment) that is made using a credit or debit card by an unauthorized user. According to the Federal Trade Commission (FTC), there were about 1579 data breaches amounting to 179 million data points whereby credit card fraud activities were the most prevalent. Therefore, it is crucial to implement an effective credit card fraud detection method that is able to protect users from financial loss. One of the key issues with applying ML approaches to the credit card fraud detection problem is that most of the published work are impossible to reproduce. This is because creditcard transactions are highly confidential. Therefore, the datasets that are used to develop ML models for credit card fraud detection contain anonymized attributes. Furthermore, credit card fraud detection is a challenging task because of the constantly changing nature and patterns of the fraudulent transactions. Additionally, existing ML models for credit card fraud detection suffer from a low detection accuracy and are not able to solve the highly skewed nature of credit card fraud datasets. Therefore, it is essential to develop ML models that can perform optimally and that can detect credit card fraud with a high accuracy score. This research focuses on the application of the following supervised ML algorithms for credit card fraud detection: Decision Tree (DT), Artificial Neural Network (ANN), and Logistic Regression (LR).

Corresponding Author:- Prateeksha M.S, B.Naga Swetha
Address**:-** Dept of ISE,RYM Engineering College, Ballari.

ML systems are trained and tested using large datasets. In this work, a credit card fraud dataset generated from European credit cardholders is utilized. Often times, these datasets may have many attributes that could have a negative impact on the performance of the classifiers during the training process.

Literature Survey:-

The authors implemented a credit card fraud detection system using several ML algorithms including logistic regression (LR), decision tree (DT), support vector machine (SVM) and random forest (RF). These classifiers were evaluated using a credit card fraud detection dataset generated from European cardholders in 2013. In this dataset, the ratio between non-fraudulent and fraudulent transactions is highly skewed; therefore, this is a highly imbalanced dataset. The researcher used the classification accuracy to assess the performance of each ML approach. The experimental outcomes showed that the LR, DT, SVM and RF obtained the following accuracy scores: 97.70%, 95.50%, 97.50% and 98.60%, respectively. Although these outcomes are good, the authors suggested that the implementation of advanced pre-processing techniques could have a positive impact on the performance of the classifiers

Varmedja et al. proposed a credit card fraud detection method using ML. The authors used a credit card fraud dataset sourced from Kaggle. This dataset contains transactions made within 2 days by European credit card holders. To deal with the class imbalance problem present in the dataset, the researcher implemented the Synthetic Minority Oversampling Technique (SMOTE) oversampling technique. The following ML methods were implemented to assess the efficiency of the proposed method: RF, NB, and multilayer perceptron (MLP). The experimental results demonstrated that the RF algorithm performed optimally with a fraud detection accuracy of 99.96%. The NB and the MLP methods obtained accuracy scores of 99.23% and 99.93%, respectively. The authors concede that more research should be conducted to implement a feature selection method that could improve on the accuracy of other ML methods.

Khatri et al. conducted a performance analysis of ML techniques for credit card fraud detection. In this research, the authors considered the following ML approaches: DT, k-Nearest Neighbor (KNN), LR, RF and NB. To assess the performance of each ML method, the authors used a highly imbalanced dataset that was generated from European cardholders. One of the main performance metric that was used in the experiments is the precision which was obtained by each classifier. The experimental outcomes showed that the DT, KNN, LR, and RF obtained precisions of 85.11%, 91.11%, 87.5%, 89.77%, 6.52%, respectively

Awoyemi et al. presented a comparison analysis of different ML methods on the European cardholder’s credit card fraud dataset. In this research, the authors used an hybrid sampling technique to deal with the imbalanced nature of the dataset. The following ML were considered: NB, KNN, and LR. The experiments were carried out using a Python based ML framework. The accuracy was the main performance metric that was utilized to assess the effectiveness of each ML approach. The experimental results demonstrated that the NB, LR, and KNN achieved the following accuracies, respectively: 97.92%, 54.86%, and 97.69%. Although the NB and KNN performed relatively well, the authors did not explore the possibility to implement a feature selection method.

The authors utilized several ML learning-based methods to solve the issue of credit card fraud. In this work, the researchers used the European credit cardholder fraud dataset. To deal with the highly imbalanced nature of this dataset, the authors employed the SMOTE sampling technique. The following ML methods were considered: DT, LR, and Isolation Forest (IF). The accuracy was one of the main performance metrics that was considered. The results showed that the DT, LR, and IF obtained the accuracy scores of 97.08%, 97.18%, and 58.83%, respectively

Existing System

The previous detecting technique (SVM) takes a long time to catch fraud which is basically depend on the database, not that much accurate and not give the result in-time. After that algorithm which is used for the detection of credit card fraudulent is generally on basis of analysis, fraudulent detection based on credit card transaction made by cardholder and the credit rate for cardholders. There are certain limits of meta-learning. There are two features which is introduced here in our report is True Positive and False alarm. Both these features play an important role in catching fraudulent because the rate of determining fraudulent behavior is quick. For the better performance of model, we need a better classifier. Different classifier can be combined together with help of meta-learning. Previously attempts have been made to work out Credit Card Fraud Detection system using SVM (Select Vector Machine). SVM makes use of hyperplane to classify the data points in a collection. A good hyperplane associates greater number of data points within its margin. This is not efficient for a large amount of data sets. As, in large amount of data sets there is a probability of redundant data which will take more time to process. Therefore, it usually delayed in calculating the fraud or there might be probability to not calculate in time

Proposed System

In the proposed system, Dataset is collected from the Kaggle website. The dataset is trained and tested using the following techniques: logistic regression, decision trees, artificial neural network. If our algorithm is applied into bank credit card fraud detection systems, the probability of fraud transactions can be predicted soon after credit card transaction occurs. Thereafter a series of anti-fraud strategies can be adopted to prevent banks from great losses and reduce risks.

Objectives:-

1. To organize the calculation for recognition of the slump shame beneficially.
2. To recommend the systems for dysfunctional behaviour arrangement.
3. To survey how to sort the mental issue of an individual.

Methodology:-

Logistic Regression:

The Logistic Regression (LR) classifier, sometimes referred to as the Logit classifier, is a supervised ML method that is generally used for binary classification tasks. LR is a special type of linear regression whereby a linear function is fed to the logit function

$$y = \alpha_0 + \alpha_1 X_1 + \alpha_2 X_2 + \cdots + \alpha_n X_n$$

$$q = \frac{1}{1 + e^{-y}}$$

where the value of q will be between 0 and 1. q is the probability that determines the prediction of a given class. The closer q is to 1, the more accurately it predicts a particular class

Decision Tree:

This represents the given data into a tree like structure in a hierarchal structure which consists of N number of nodes which is also called as attributes, there are some directed links or edges so that it establishes a relation between all the entities available. A data set divided into large number of rows and columns where last column is referred as a class and the rows or tuples are referred as instance and other columns are called as attributes or features. Decision tree has a primary node or a root node which has no incoming edges and it has two or more number of out-going edges. At the end of the tree there are leaf nodes or terminal nodes which has incoming edges but no out-going edges. And there are some nodes in between root and leaf nodes which are called as internal nodes it has exactly one incoming node and several out-going nodes. Edges consists of conditions known as attribute test conditions. This algorithm is represented in the form of a linked list in memory map. This tree is traversed in the left to right format.

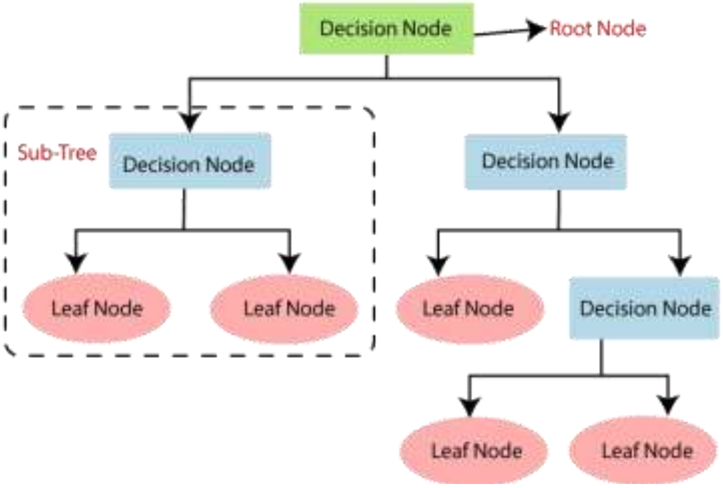


Fig:- Decision Tree.

Artificial Neural Networks:

Artificial Neural Network (ANN) is a supervised ML method that is inspired from the inner workings of the human brain. The simplest ANN have the following basic structure: an input layer, one hidden layer and an output layer. The input layer size is based on the number of features in a given dataset. The hidden layer size can be varied based on the complexity of a task and the output layer size depends on the type of problems to be solved. The most basic component of an ANN is a node or neuron. In this research, we consider feed forward ANNs. Therefore, the information flows in one direction (from its input to its output) through a neuron. Figure 1 depicts a graphical representation of a simple ANN with 3 nodes in the input layer, a hidden layer with 4 nodes and an output layer with 1 node

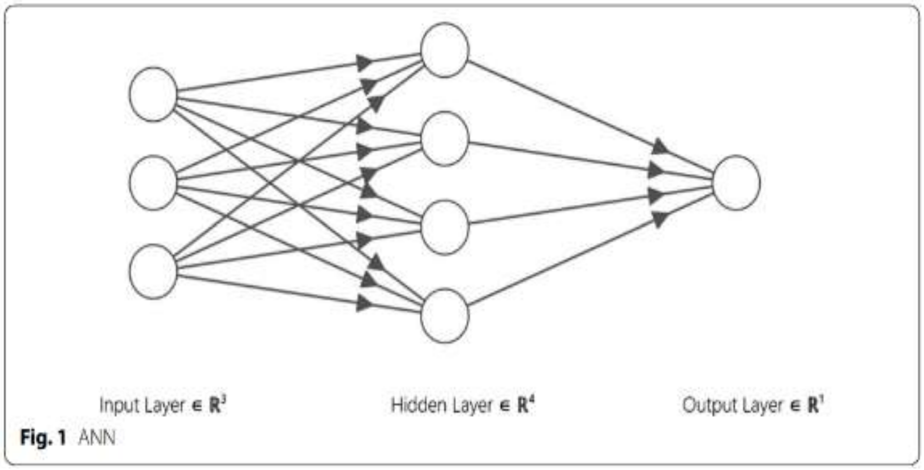


Fig:- Artificial Neural Network.

Problem Statement

Not al1 the doubtful transactions consider as fraudulent. It is commonly called as false positive (FP) which means that the case was not fraud although it was flagged as being potentially scam. This process of affirming each transaction those outliers from the cardholder's normal routine brings doubt about possible client disappointment. Additionally, the expenses related with exploring an enormous no. of false positives are high.

Experiment Results:-

Evaluation Criteria

To evaluate the results of the classification algorithms there are various parameter such as Accuracy score, classification report, F1-score, confusion matrix etc. Some important definitions are:

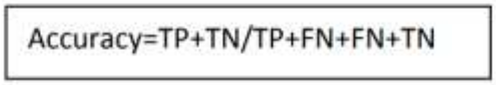
o True positive (TP)- It is an outcome in which the model accurately predicts the positive class.

o False positive (FP)- It occurs when the positive class is predicted wrongly by the model.

o True negative (TN)- It is an outcome in which the model accurately predicts the negative class.
o False negative (FN)- It is an outcome in which the model predicts the negative class inaccurately.

Accuracy-

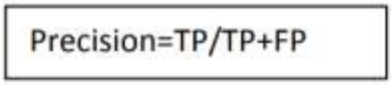
The number of correct predictions divided by the total number of input samples is known as Accuracy



Accuracy=TP+TN/TP+FN+FN+TN

Accuracy score for Logistic Regression, Decision Tree and Neural networks are 0.97,0.98,0.99.

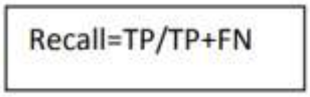
Precision (Specificity)–
It's the number of correct positive outcomes divided by the classifier's projected number of positive findings.



Precision=TP/TP+FP

Precision score for Logistic Regression, Decision Tree and Neural networks are 0.76,0.66,0.74

Recall (Sensitivity)-
It's calculated by dividing the number of correct positive results by the total number of relevant samples (all samples that should have been identified as positive).



Recall=TP/TP+FN

Recall score for Logistic Regression, Decision Tree and Neural networks are 0.62,0.70,0.70

Results:-

Three machine learning methods were employed to detect fraud in the credit card system in this article. Data from 80% of the training dataset and 20% of the testing dataset were utilised to evaluate the algorithms. Accuracy, precision, and recall score are used to analyze the performance these four approaches. As shown in the observations of accuracy outcomes. The accuracy score for Decision tree, Logistic Regression and Neural networks. As we compare 3 algorithms, we can clearly see that the Neural networks predicts the fraudulent transaction with accuracy score, precision and recall score.

Conclusion:-

With the development of electronic financial transaction technology and the emergence of simple payment, the risk of fraudulent payment and fraudulent payment increases as the authentication process is simplified. The types of fraudulent use of credit cards include theft and loss, identity theft, new card not received, card forgery, and card information theft. In particular, as phishing, pharming as well as card information leakage due to card information leakage, card information theft accidents are occurring. In response, the government tried to deal with electronic financial fraud by implementing the 'e-financial fraud prevention service'. It is difficult to cope with financial fraud by simply setting the existing keyboard security, public certificate, and additional password. The abnormal transaction detection system is used to analyze the user's data and payment data in real time to inform the financial institution and the user of the detection if it is different from the usual pattern, and further to arbitrarily stop the transaction. Therefore, an abnormal transaction detection system is important for fast and accurate detection, and research is needed to improve the algorithm. In this study, the method of detecting anomalous transactions using the electronic payment log analysis and machine learning technique was investigated. Results show the significance of algorithms used over the dataset and efficient classification is performed.

In future machine learning concepts can be applied using convolution networks for improved accuracy. Also, some other datasets can be used for further testing of proposed mechanisms.

References:-

1. Machine Learning for Credit Card Fraud Detection System, Lakshmi S V S, Selvani Deepthi Kavila, November 2018
2. Credit Card Fraud Detection using Data science and Machine learning, S P Maniraj, Aditya Saini, Shadab Ahmed, Swarna Deep Sarkar, September 2019.
3. A. Mishra, C. Ghorpade, "Credit Card Fraud Detection on the Skewed Data Using Various Classification and Ensemble Techniques" 2018 IEEE International Students' Conference on Electronics, Electrical and Computer Science (SCEECS) pp. 1-5. IEEE
4. P. Kumar, F. Iqbal, Credit card fraud identification using machine learning approaches, in: Proceedings of the 1st International Conference on Innovations in Information and Communication Technology (ICIICT), CHENNAI, India, 2019, pp. 1–4, doi:10.1109/ICIICT1.2019.8741490
5. A.A Taha, S.J. Malebary, an intelligent approach to credit card fraud detection using an optimized light gradient boosting machine, IEEE Access 8 (2020)25579-25587, doi:10.1109/ACCESS.2020.2971354.