International Journal of Innovative Research & Growth
A Peer Reviewed Journal
ISSN: 2455-1848
Journal homepage: http://www.ijirg.com/ijirg/
Volume: 12 Issue: 2, April 2023

A comprehensive review on credit card fraud detection using machine learning techniques

Arvind Rawat[1], Sandeep Kumar Tiwari[2*]
[1,2,3]Vikrant University, M. P, India
E-mail: sandeep72128@gmail.com, arvind@vitm.edu.in
* Corresponding Author

Article Info
Received 13 March 2023
Received in revised form 19 April 2023
Accepted for publication 30 April 2023
DOI: 10.26671/IJIRG.2023.2.12.103
Citation:
Rawat, A., Tiwari, S. K. (2023). A comprehensive review on credit card fraud detection using machine learning techniques. *Int J Innovat Res Growth*, 12, 25-29.

Abstract
The rapid advancement in E-Commerce industry has led to an exponential increase in the use of credit cards for online purchasing and consequently they have been surging in the fraud related to it .In recent years, For banks has become very difficult for detecting the fraud in credit card system and online banking and a large financial loss has greatly affected persons and also merchants and banks. Basically, Fraud is an unlawful way to obtain commodities and resources. The goal of such illegal transaction might be to get yield without paying or gaining an unconstitutional access to an account. That's why there is a need for an efficient technique that can be used for detecting the credit card fraud. Machine learning algorithms play an important role for detecting the credit card fraud in transactions.
Keywords: - ANN, Random forest, SVM, Regression, Classification.

1. Introduction

Now a day's digitization is increasing rapidly in every field like electronic commerce, government agencies, corporate industries, banking sector and in various other organizations due to this digitization the credit card frauds are also increasing not only in online transactions as well as offline transactions. Credit card fraud is the criminal use of someone else's personal credentials, as well as their credit standing, to borrow money or use credit cards to purchase goods or services with no intention of repaying the debt. Credit card fraud is the common type of identity robbery. According to the Federal Trade Commission, more than 270,000 Americans reported new or existing account fraud in 2019.
There are two types of card fraud: first one in which card is present, this type of fraud is not common and the other one in which card is not present which is very common now a days.

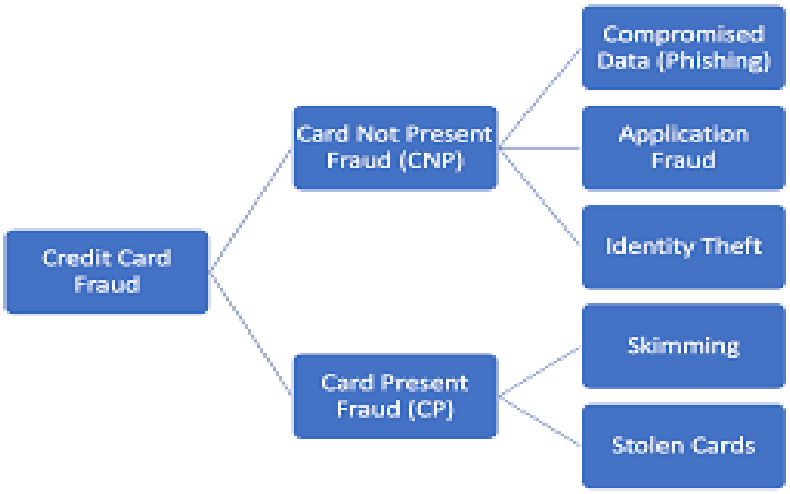

Figure-1Types of Credit card fraud

There are various types of payment credit card frauds like application fraud, account takeover, social engineering fraud, skimming.

A. PoS Fraud

In this type of fraud, small skimming devices are attached to Point-of-Sale devices to hack data. These devices can scan and store the card information while the customer completes a swipe transaction. Generally, a store owner who shares details with attackers or hackers. Similarly attachments may be fastened on to ATM card slots to copy card information, while a hidden camera is placed over the keypad to capture PIN [16].

B. Phishing

This type of attack involves impersonating official communication mail from the bank which in turns acts as you to click on that links. This will usually take to websites that look to be authentic. Once you enter your card details on these fake links, fraudsters can access the details and use them for their benefit. Another version is when fraudsters impersonate bank officials on phone calls, asking you to share an OTP to 'verify your card' or 'avail the reward points' or 'extend the validity of your reward points [16].'

C. Keystroke Logging

Today, since most financial transactions are online, hackers have started relying on keystroke logging through malicious software to grab credit card details. This usually begins after you have clicked on a suspicious link and unknowingly installed malware on your system. The software records every key pressed on the system, eventually stealing card details, PIN and more [16].

D. Application Fraud

This is a type of identity theft where fraudulent actors impersonate a genuine customer by using their stolen or counterfeited documents to obtain a credit card. While this might be detected after thorough background checks, if carried out, this will allow criminals to use a valid credit card with a false paper trail. A similar type of fraud involves taking over a valid credit card account by posing as the customer using a similar fake paper trail [16].

E. Theft or Loss of Card

If your physical credit card gets stolen or misplaced, there are chances it could be misused. Fraud avoidance and fraud recognition both are the way to handle the fraud. In fraud prevention, the main intend is to prevent the fraudulent action; it observes the transaction and prevents the illegal transactions. Whereas in fraud detection, the aim is to distinguish the fraudulent transaction and valid transactions or authorized transactions. Machine learning algorithms can be used for credit card fraud detection. Machine learning is a subset of artificial intelligence. It focuses mainly on the designing of system thereby allowing them to learn and make predictions based on some experience which is data in case of machines. Machine learning enables computer to act and make data driven decisions rather than being explicitly programmed to carry out certain task these programs are designed to learn and improve over time when exposed to new data. The process starts with loading good quality data and then training machines by building machine learning models using the data and different algorithms [2].

The aim of this research paper is to represent the state of the art in the field of credit card fraud detection. In this paper section II describes the types of machine learning techniques and section III describes the related work in this field and section IV elaborates the problem identification and future scope and finally section V describes the conclusion of this research paper.

2 Types of Machine Learning Techniques and Algorithms

A. Machine Learning Techniques

Machine learning techniques can be broadly classified into three categories [3]:

1) Supervised Learning: Supervised learning is that learning in which the model is trained on a labeled data set. Labelled dataset is one which has both input and output values. Supervised learning as the name indicates the presence of a supervisor as a teacher. Basically supervised learning is learning in which train or test the machine using data.

Supervised learning algorithm consist of two types of technique- Classification and Regression.

a) Classification: It is a Supervised Learning task where output is having defined labels (discrete value). It can be either binary or multi class classification. In binary classification, model predicts either 0 or 1 ; yes or no but in case of multi class classification, model predicts more than one class.

b) Regression: It is a Supervised Learning task where output is having continuous value.

2) Unsupervised Learning- It is a type of machine learning algorithm used to draw consequences from data set consisting of input data without output values. It is the training of machine using data that is uncategorized and unlabelled and algorithm to act on that data without any supervision. It means that teacher is not present in this type of learning.

Unsupervised learning consists of two types of techniques Clustering and association rule mining.

3) Reinforcement Learning: It is a type of machine learning algorithm where an agent learns to behave in an environment by performing actions. In this learning agent decides what action is performed for a task, according to the action's agent got rewards by the environment. These rewards may be positive or negative.

B. Machine Learning Algorithms

There are various machine learning algorithms in which some important machine learning algorithms are as follows:

1) Logistic Regression: Logistic Regression is one of the classification algorithms, used to predict a binary values in a given set of independent variables (1 / 0, Yes / No, True / False). To represent binary / categorical values, dummy variables are used. For the purpose of special case in the logistic regression is a linear regression, when the resulting variable is categorical then the log of odds are used for dependent variable and also it predicts the probability of occurrence of an event by fitting data to a logistic function [1].
2) Linear regression: Linear regression is a supervised learning algorithm; it performs a regression task. It describes a linear relationship between input variable and output variable. It is used for predictive analysis. Linear regression is a linear approach for modeling the relationship between the criterion or the scalar response and the multiple predictors or explanatory variables. Linear regression focuses on the conditional probability distribution of the response given the values of the predictors [3].
3) Multiple Linear regression: Multiple linear regression is a supervised learning algorithm which is used for regression. It describes the relationship between two or more independent variable and single dependent variable by fitting an algorithm to data.
4) Polynomial Regression: Polynomial regression is a type of linear regression in which the relationship between the independent variable and dependent variable is describes as an mth degree of polynomial. It fits the nonlinear relationship between the dependent and independent variable. It is used for curvilinear data. Polynomial regression is fit with the method of least squares. The goal of regression analysis to model the expected value of a dependent variable y in regards to the independent variable x.
5) Decision tree: Decision tree is the most powerful and popular tool for classification and prediction. A Decision tree is a flowchart like tree structure, where each internal node denotes a test on an attribute, each branch represents an outcome of the test, and every leaf node (terminal node) holds a class label. Decision Tree is a tree structured framework. Decision Tree (DT) is a white box type of ML algorithm. Primarily, it is used for classification. However DT can also be used for regression. It works on the principle called "Decision-making logic". The decision tree is a distribution-free or non-parametric method, which does not depend upon probability distribution assumptions. Decision trees can handle high dimensional data with good accuracy [1].
6) Random Forest: It belongs to the family of supervised learning approaches, suitable for the classification and regression problems as well. Basic working ideas behind this approach are multiple collections of tree-structured classifiers. Random forest is an ensemble learning method. It is used when size of dataset is large and the very large number of input variables approximately hundreds or thousands [1].
7) Support Vector Machine: In machine learning, support vector machines (SVMs, also support vector networks) are supervised learning models with associated learning algorithms that analyze data used for classification and regression analysis. A Support Vector Machine (SVM) is a discriminative classifier formally defined by a separating hyper plane. In other words, given labeled training data (supervised learning), the algorithm outputs an optimal hyper plane which categorizes new examples. An SVM model is a representation of the examples as points in space, mapped so that the examples of the separate categories are divided by a clear gap that is as wide as possible.
In addition to performing linear classification, SVMs can efficiently perform a non-linear classification, implicitly mapping their inputs into high-dimensional feature spaces [1].
8) Artificial Neural Network: An ANN often just called a neural network. ANN is an efficient and most used approach in the field of the computing system. The idea behind the ANN is based on Biological Neurons in the human brain.
It comprised of an interrelated group of neurons and process instruction using connectionist approach to computation. ANN has an adaptive property because neural network changed their structure according to internal information or external information passes through this network.
Artificial neural network has various components are as input layer, hidden layer, output layer, weights, activation function, threshold. ANN represents a basic attempt to simulate the type of nonlinear learning that occurs in the networks of neurons present in nature [5].

3. Literature Review

Yashvi Jain et.al. proposed a model for credit card fraud detection. In the proposed methodology researchers have used various machine learning algorithms such as support vector machine (SVM), artificial neural network (ANN), Bayesian Networks, K-Nearest Neighbors (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 byes, 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 improved results with one type of datasets and poor results with another kind 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 [1].
Navanushu Khare et.al. have used the decision tree, random forest, SVM, and logistic regression algorithms for credit card fraud detection. Researchers have taken the highly skewed dataset. And they concluded that random forest provides the best results and good accuracy as comparison to others algorithm and also concluded that SVM algorithm has a data imbalance problem [2].

Ruttala Sailusha et.al. proposed a model for credit card fraud detection. In the proposed methodology researcher have used random forest algorithm and ada-boost algorithm for credit card fraud detection. Researcher also concluded that both algorithm provides the same results [3].
Sikdar Md. S Askari et.al. proposed a model for e-transactional credit card fraud detection. In this proposed methodology the researcher used Intuitionistic fuzzy logic based decision tree algorithm for credit card fraud detection and also used both normal and derived attributes. The mathematical logic, intuitionistic fuzzy logic based decision tree presented here classifies the transactions into different classes fraud, normal and doubtful based on the information gain ratio calculated from membership degree and non-membership degree to the fuzzy sets defined for different attributes. The proposed algorithm IFDTC4.5 is applied on the training data set which is splitted into multiple windows, covering the Indian scenarios of e-transactions. Conditional attributes are selected based on information gain ratio. The algorithm is executed on the test data set, then test on some other transactions with different values of the attributes in different situations are also conducted to determine the sensitivity and specificity. Then the performance measures are computed [4].
Ghosh and Reilly have used a feed-forward radial basis function (RBF) a three layer neural network having two training passes that produce fraud score in two hours interval for any new transaction [5].
Sayeda et.al. have used granular neural network (GNN), rule based approach and fuzzy neural network, for the detection of customer specific credit card frauds, the fuzzy neural network is trained with training data set that produces fuzzy rules as result for checking the transactions authenticity [6].
Vaishnavi Nath Dornadula et al. proposed a novel approach for fraud detection, in which customers are grouped based on their transactions and extract behavioral patterns to develop a profile for every cardholder. Then different classifiers are applied on three dissimilar groups later rating scores are generated for every sort of classifier. These vibrant changes in parameters lead the method to adapt to fresh cardholder's transaction behaviors timely. Followed by a feedback method to solve the problem of concept drift. We observed that the Matthews Correlation Coefficient was the better parameter to deal with imbalance dataset. MCC was not the only solution. By applying the SMOTE, researcher tried balancing the dataset, where we found that the classifiers were performing better than before. The other way of handling imbalance dataset is to use one-class classifiers like one-class SVM and researcher concluded that Logistic regression, decision tree and random forest are the algorithms that gave better results [7].
Akila et al. proposed an ensemble model namely called as Risk Induced Bayesian Inference Bagging model, RIBIB. They propose a three-step approach a bagging architecture with a constrained bag creation method, Risk Induced Bayesian Inference technique as base learner, and a weighted voting combiner. Bagging is a procedure of combining multiple training datasets and utilizing them independently to train multiple classifier models. They evaluated their solution on Brazilian Bank Data and excelled at cost minimizing compared to the other state-of-the-art models [8].
Andrea Dal Pozzolo et. al. reviewed that credit card fraud problem comprises a number of relevant issues, namely: concept drift , class imbalance , and verification latency. In this work paper has author firstly proposed by taking the help of industrial partner, a formalization of the fraud-detection problem that logically explains the working conditions of FDSs that everyday evaluate huge streams of credit card transactions. Then they designed and assessed a novel learning approach that efficiently addresses class imbalance, concept drift, and verification latency and detailed the impact of class unbalance and concept drift in a real-world data stream consists of more than 75 million transactions, authorized over an instant of time window of three years. They used two vast data sets of real-world transactions to get precise alerts, it is mandatory to assign larger importance to feedbacks during the learning problem. They have used the supervised learning for aggregation [9].
Altyeb Altaher Taha et. al. proposed an intelligent approach for finding fraud in credit card transactions by means of an optimized light gradient boosting machine(O Light GBM).In the proposed method, a Bayesian based hyper parameter optimization algorithm is smartly integrated to tune the parameters of alight gradient boosting machine(Light GBM). To demonstrate the effectiveness of our proposed O Light GBM for detecting fraud in credit card transactions, experiments were performed by two real-world public credit card transaction data sets consisting of fraudulent transactions and legitimate ones [10].
Rafael San Miguel Carrasco et.al. proposed a methodology in which a set of deep neural networks tested to measure their ability to detect false positives, by processing alerts triggered by a fraud detection system. The performance achieved by each neural network setting is presented and discussed. In this research work, deep neural networks assessed from a perspective of credit card fraud alert reduction. The goal was to reproduce the ability to capture and automate decision criteria used by humans reported by previous literature. A set of alerts triggered by an FDS (associated with suspicious transactions) classified as either valid alerts, representing real fraud cases, or wrong alerts, representing false positives, by ten neural network architectures [11].

4. Problem Identification and Future Scope

A lot of previous work has been done by researchers for credit card fraud detection but still there is a need for an efficient model which can easily detect the fraud transactions on the basis of given data. There are various problems that researches might faced during the classification of transactions are as follows: like imbalanced dataset problem, and the real dataset is not available for fraud detection, the pure transaction information fetched from the organization database is quite limited. The balance of cardholder, credit limit, transaction time, transaction amount is some of them. When only these ready-to-use features are used to train ordinary machine learning algorithms, the performance is not likely to differ among them. In the past decades the researchers used the a variety of machine learning algorithms like decision tree, logistic regression,

artificial neural network, support vector machine and random forest, granular neural network etc. but there is a need for a methodology for classification and detection of fraud transactions which can overcome all the issues occurred while the credit card fraud detection and also provides better performance and accuracy.
Future work concerns the study of adaptive and possibly nonlinear aggregation methods for the classifiers trained on feedbacks and delayed supervised samples. We also expect to further increase the alert precision by implementing a learning to rank approach that would be specifically designed to replace the linear aggregation of the posterior probabilities. Finally, a very promising research direction concerns semi supervised learning methods for exploiting in the learning process also few recent unlabeled transactions.

5. Conclusion

Credit card fraud detection is an approach of detecting or classifying fraudulent transactions because credit card fraud detection is an act of criminal dishonesty. There are two types of fraud. And now days credit card frauds are increasing exponentially day by day. That's why there is need for an efficient model which is used for credit card fraud transactions using the machine learning algorithms. As we have seen that various machine learning algorithms can be used for credit card fraud detection which can provides better results. Researchers have used various machine learning algorithms like linear regression, decision tree, logistic regression, support vector machine, artificial neural network for detecting the frauds. This research paper will be helpful for researchers for finding out the state of the art in the field of credit card fraud detection.

Conflict of Interest

In this manuscript the authors declare that there is no conflict of interest.

References

i. Jain, Y., Tiwari, N., Dubey, S., and Jain, S. (2019). A Comparative Analysis of Various Credit Card Fraud Detection Techniques. *International Journal of Recent Technology and Engineering (IJRTE)*, 7.
ii. Khare, N., and Sait, S. Y. (2018). Credit Card Fraud Detection Using Machine Learning Models and Collating Machine Learning Models. *International Journal of Pure and Applied Mathematics*, 118, 825-838.
iii. Sailusha, R., Gnaneswar, V., Ramesh, R., and Rao, G. R. (2020). Credit Card Fraud Detection Using Machine Learning. *Proceedings of the International Conference on Intelligent Computing and Control Systems (ICICCS 2020) IEEE Xplore* Part Number: CFP20K74-ART.
iv. Askari, S. M. S., and Hussain, M. A. (2020). IFDTC4.5: Intuitionistic fuzzy logic based decision tree for E-transactional fraud detection. *Journal of Information Security and Applications*, 52, 102469.
v. Ghosh, Reilly. Credit card fraud detection with a neural-network. *In: 1994 proceedings of the twenty-seventh Hawaii international conference on system sciences*, 3, 621–30. doi:10.1109/HICSS.1994.323314.
vi. Syeda, M, Zhang, Y-Q, and Pan, Y. (2002). Parallel granular neural networks for fast credit card fraud detection. *Fuzzy systems, 2002. FUZZ-IEEE'02. Proceedings of the 2002 IEEE international conference*, 1, 572-577.
vii. Dornadula, V. N., and Geetha, S. (2019). Credit Card Fraud Detection using Machine Learning Algorithms. *Procedia Computer Science*, 165, 631-641.
viii. Akila, S., and Reddy, U. S. (2018). Cost-sensitive Risk Induced Bayesian Inference Bagging (RIBIB) for credit card fraud detection. *Journal of Computational Science*, 27, 247–254. doi: 10.1016/j.jocs.2018.06.009.
ix. Pozzolo, A. D., Boracchi, G., Caelen, O., and Alippi, C. (2018). Credit Card Fraud Detection: A Realistic Modeling and a Novel Learning Strategy. *IEEE transactions on neural networks and learning systems*, 29.
x. Taha, A. A., and Malebary, S. J.(2020). An Intelligent Approach to Credit Card Fraud Detection Using an Optimized Light Gradient Boosting Machine. *IEEE Access*, 2020.
xi. Carrasco, R. S. M., and Urbán, M. A. S. (2017). Evaluation of Deep Neural Networks for Reduction of Credit Card Fraud Alerts. *IEEE ACCESS*, 2017.
xii. Zhang, X., Han, Y., Xu, W., and Wang, Q. (2019). HOBA: A novel feature engineering methodology for credit card fraud detection with a deep learning architecture. *Inf. Sci., May 2019.* Accessed: Jan. 8, 2019.
xiii. Carneiro, N., Figueira, G., and Costa, M. (2017). A data mining based system for credit-card fraud detection in e-tail. *Decis. Support Syst.*, 95, 91–101.
xiv. Lebichot, B., Le, Y.-A., Borgne, L., He-Guelton, Oblé, F. and Bontempi, G.(2019). Deep-learningdomainadaptationtechniquesforcreditcardsfrauddetection.*Proc.INNSBigDataDeepLearn.Conference,Genoa,Italy*,78–88.
xv.Naaz, H.J.S. (2019). Credit card fraud detection using local outlier factor andisolationforest. *Int.J.Comput.Sci.Eng.*,7, 1060–1064.
xvi. Delamaire, L., Abdou, H.,and Pointon, J. (2009). Credit card fraud and detection techniques: a review. *Banks and Bank Systems*, 4.