A Review Paper on Credit Card Fraud Detection Using Machine Learning

Asst.Prof Anush Sharma

Sahil, Anish Kumar, Simran

Himachal institute of engineering and technology,
Shahpur , Himachal Pradesh
INDIA

Abstract - The increasing spread of cyberattacks and crimes makes cyber security a top priority in the banking industry. Credit card cyber fraud is a major security risk worldwide. Conventional anomaly detection and rule-based techniques are two of the most common utilized approaches for detecting cyber fraud, however, they are the most time-consuming, resource-intensive, and inaccurate. Machine learning is one of the techniques gaining popularity and playing a significant role in this field. This study examines and synthesizes previous studies on the credit card cyber fraud detection. This review focuses specifically on exploring machine learning/deep learning approaches. In our review, we identified 181 research articles, published from 2019 to 2021. For the benefit of researchers, review of machine learning/deep learning techniques and their relevance in credit card cyber fraud detection is presented.Features of credit card frauds play important role when machine learning is used for credit card fraud detection, and they must be chosen properly.Credit Card Fraud Detection is a typical sample of classification. In this process, we have focused on analysing and pre-processing data sets as well as the deployment of multiple anomaly detection algorithms such as Local Outlier Factor and Isolation Forest algorithm on the PCA transformed Credit Card Transaction data.

Keywords –   Introduction, Machine Learning ,Methodology, Linear Regression, Analytics and Literature Review, Conclusion

INTRODUCTION

Credit card fraud detection is a critical component of financial security and risk management in the modern digital economy. It refers to the process of identifying and preventing unauthorized or fraudulent credit card transactions. This type of fraud occurs when individuals or criminal organizations use stolen or counterfeit credit card information to make unauthorized purchases or withdrawals.

'Fraud' in credit card transactions is unauthorized and unwanted usage of an account by someone other than the owner of that account. Necessary prevention measures can be taken to stop this abuse and the behaviour of such fraudulent practices can be studied to minimize it and protect against similar occurrences in the future.

There are now machine learning algorithms that are used to conquer the fraud occur in the credit card transaction. Nowadays, the machine learning and artificial intelligence are the growing technologies. In today's world and these technologies can help us In the different traits of our society.

Let us take a slight understanding of machine learning in the credit card fraud detection project.There are algorithms like forest algorithms , genetic algorithms and also trained the data-set in such a way that it can easily recognize the fraudulent transaction and non-fraudulent transactions.

Here, is a overview of the steps involved in the credit card fraud detection using machine learning.
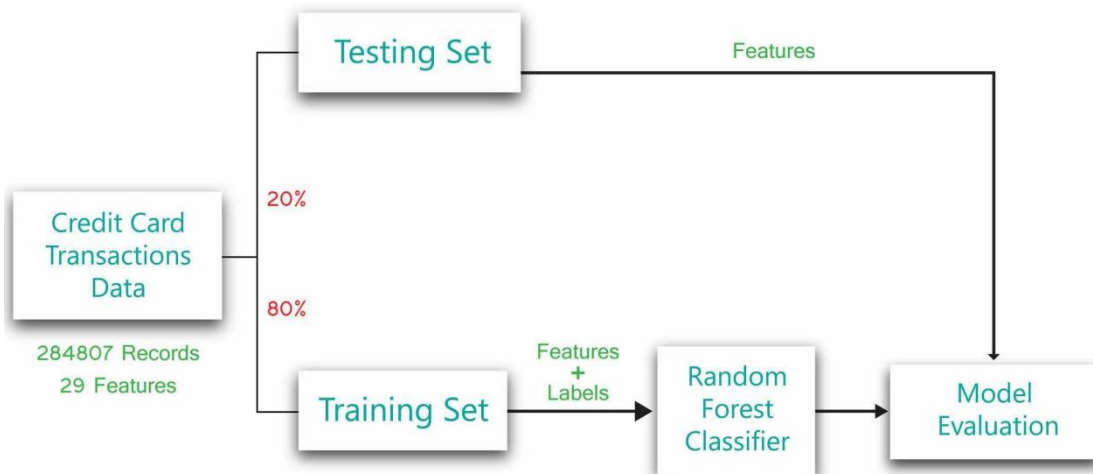
Fig(1). credit card fraud detection flowchart

Here, in the flowchart there is a method called random forest classifier but in the place of this method we can also use the linear regression method.

Fraud detection methods are continuously developed to defend criminals in adapting to their fraudulent strategies. These frauds are classified as:

- ➢ Credit Card Frauds: Online and Offline
- ➢ Card Theft
- ➢ Account Bankruptcy
- ➢ Device Intrusion
- ➢ Application Fraud
- ➢ Counterfeit Card
- ➢ Telecommunication Fraud

Machine learning

Machine learning (ML) have grown rapidly in recent years in the context of data analysis and computing that typically allows the applications to function in an intelligent manner [95]. ML usually provides systems with the ability to learn and enhance from experience automatically without being specifically programmed and is generally referred to as the most popular latest technologies in the fourth industrial revolution (4IR or Industry 4.0) [103, 105]. "Industry 4.0" [114] is typically the ongoing automation of conventional manufacturing and industrial practices, including exploratory data processing, using new smart technologies such as machine learning automation. Thus, to intelligently analyze these data and to develop the corresponding real-world applications, machine learning algorithms is the key. The learning algorithms can be categorized into four major types, such as supervised, unsupervised, semi-supervised, and reinforcement learning in the area.

1. *Supervised:* Supervised learning is typically the task of machine learning to learn a function that maps an input to an output based on sample input-output pairs [41]. It uses labeled training data and a collection of training examples to infer a function. Supervised learning is carried out when certain goals are identified to be accomplished from a certain set of inputs [105], i.e., a *task-driven approach.*

2. *Unsupervised:* Unsupervised learning analyzes unlabeled datasets without the need for human interference, i.e., a *data-driven process* [41]. This is widely used for extracting generative features, identifying meaningful trends and structures, groupings in results, and exploratory purposes.

3. *Semi-supervised:* Semi-supervised learning can be defined as a *hybridization* of the above-mentioned supervised and unsupervised methods, as it operates on both labeled and unlabeled data [41, 105]. Thus, it falls between learning "without supervision" and learning "with supervision".

4. *Reinforcement:* Reinforcement learning is a type of machine learning algorithm that enables software agents and machines to automatically evaluate the optimal behavior in a particular context or environment to improve its efficiency [52], i.e., an *environment-driven approach.*

METHODOLOGY

This review paper is about the credit card fraud detection. So let us review the methodologies that are used in the credit card fraud detection. So first we have to take the data sets from the kaggle which is known for its datasets . so firstly we have to take the first 5 rows to operate on that.

There are columns frrom v1-v28 which has the records of the transactions and with this there are columns named as TIME,AMOUNT,CLASS which show us the time of the transaction and amount of that transaction and class identifies whether that transaction is fraudulent or non-fraudulent.And after that we have to train that datasets for the datasets to identify the fraud and non fraudulent transactions. There are classes of 0 means that transaction is non-fraudulent and class 1 means the transactions are fraudulent.
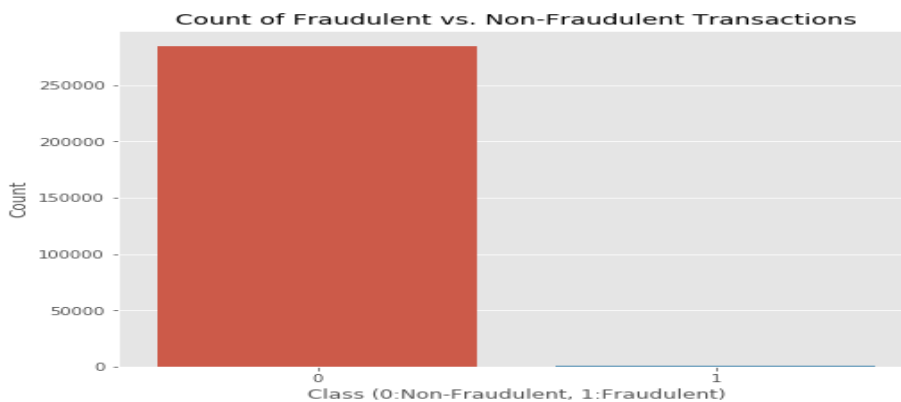
Fig(2) count of fraudulent and non-fraudulent transactions

This graph shows that the no. Of legit transactions are more as compared to fraudulent transactions. So this graph tells us that in our datasets there are some transactions which are fraudulent to some extent.

After checking our datasets for the values now we are going to check the datasets for missing values so that we can work along with whole data values and originates accurate results. Now we will check the data sets for missing values. The next step is to train the dataset to get the result of the credit card datsets. The next step is to identify the credit card transactions to fraud or non fraudulent. There are various machine learning methodologies which can be used .Here are those methods:

1. Linear regression

2. Isolated forest algorithm

Linear 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 [6]. LR is a special type of linear regression whereby a linear function is fed to the logit function.

Regression:
1. Regression analysis is a predicting modeling technique.
2. It estimate the relationship between a dependent(target) and an independent variable
Predictor.

Logistic Regression:
•Logistic Regression Is One Of The Most Popular Machine Learning Algorithms

•Logistic Regression Is A Machine Learning Model Which Is Used To The Binary Classification Model.Binary Classification Is Just Like 0\1 And True And False.

•Linear Regression Is Used For Solving Regression Problems, Whereas Logistic Regression Is Used For Solving The Classification Problems.

•Logistic Regression Can Be Used To Classify The Observations Using Different Types Of Data And Can Easily Determine The Most Effective Variables Used For The Classification.

•   Logistic Regression Is Much Similar To The Linear Regression Except That How They Are Used.

•   Equation of logistics regression are

$Y = 1/1 + e^{-(a_0 + a_1 x)}$

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.
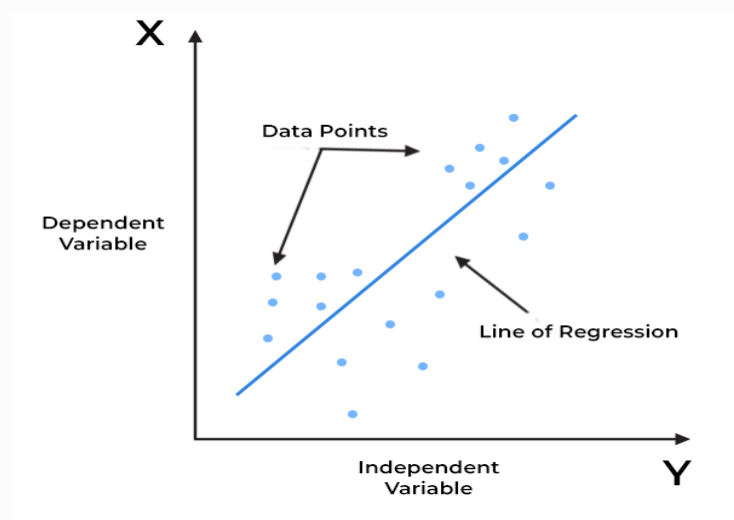

Fig -4 linear regression

Isolated Forest algorithm: It isolates observations by arbitrarily selecting and then randomly selecting a split value between maximum and minimum values of designated feature. The average of this path lengh gives a measure of normality and the decision ucntion which we use.

```
        from sklearn.ensemble import IsolationForest

# Create an Isolation Forest model
model = IsolationForest(contamination=0.05)  # You can adjust the contamination parameter

# Fit the model to your data
model.fit(X)  # X is your feature matrix

# Predict anomalies
predictions = model.predict(X)  # 1 for inliers (normal data points), -1 for outliers (anomalies)
```
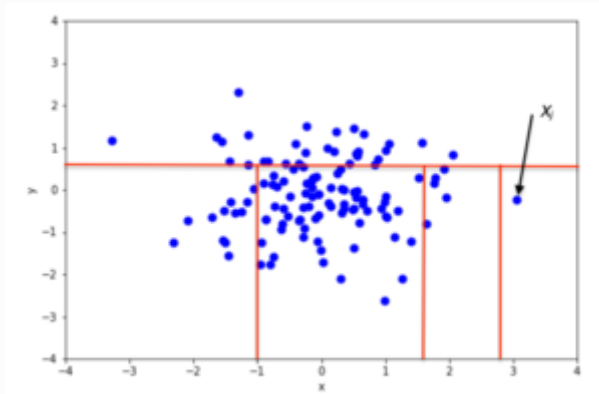

Fig.4 isolated forest model

Result

The result of credit card fraud detection is to find the accuracy of finding the fraudulent transaction among the datasets provided by the banks or any other websites. Now with accurate time and class we can get the accuracy rate of 100%. So for the better results we have to take the better inputs.

The result is being the classification report which said that when the class 0 is being hit then that transaction is non fraudulent transaction and when class 1 is being hit then that transaction is fraudulent.
The result can be get better with better inputs.

CONCLUSION

As we all know the credit card fraud is no doubt an act of crime and to avoid and to stop this crime the credit card fraud detection system plays a vital role. But as we have seen in the review paper that we are using the linear regression and isolated forest algorithm to find the fraudulent transactions.
      But the main problem in this system is that these algorithm can handle a small scale data very easily.

On the technical level we seen that when the 40-50% of datasets are involved in the project then the accuracy level is upto 90% but when the records get increasing the accuracy level is decreasing. So to get a higher accuracy we have to train the system in such a way that it can handle a large amount of data very easily.

And after all these, we can also increase the accuracy of this system by giving the system a proper set of data to get the goal. As the system has full information about the transaction then it will easily find the result accurately.

REFRENCES

1). https://fraud.net/d/credit-card-fraud-detection/#:~:text=It%20involves%20using%20various%20techniques,stopping%20unauthorized%20or%20suspicious%20transactions.

2). Credit Card Fraud Detection using Machine Learning and Data Science by S P Maniraj Assistant Professor (O.G.) Department of Computer Science and Engineering SRM Institute of Science and Technology Research paper.

3) https://favtutor.com/blog-details/credit-card-fraud-detection-using-machine-learning

4) Andreas C Müller and Sarah Guido. Introduction to machine learning with Python: a guide for data scientists. O'Reilly Media, Inc., 2016 [MullerG16]

5) Wes McKinney. Python for data analysis: Data wrangling with Pandas, NumPy, and IPython - 2nd Edition. O'Reilly Media, Inc., 2017 [McK17]

6) Sarker, I.H. Machine Learning: Algorithms, Real-World Applications and Research Directions. *SN COMPUT. SCI.* **2**, 160 (2021). https://doi.org/10.1007/s42979-021-00592-x

7) Cic-ddos2019 [online]. available: https://www.unb.ca/cic/datasets/ddos-2019.html/

8) Ileberi, E., Sun, Y. & Wang, Z. A machine learning based credit card fraud detection using the GA algorithm for feature selection. *J Big Data* **9**, 24 (2022). https://doi.org/10.1186/s40537-022-00573-8

9) https://www.spiceworks.com/tech/artificial-intelligence/articles/what-is-linear-regression/

10) https://machinelearninginterview.com/topics/machine-learning/explain-isolation-forests-for-anomaly-detection/