IJCRT.ORG ISSN : 2320-2882

INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS (IJCRT)

An International Open Access, Peer-reviewed, Refereed Journal

Detecting Frauds In Credit Card Using KNN And Random Forest Machine Learning Approach

Rucha Narkhede [1], Nilesh Chaudhari [2]

PG Student [1], Assistant Professor [2]

Department of Computer Engineering,

Godavari College of Engineering, Jalgaon, India

Abstract: Now-a-days online transactions and e-trade platforms are grown to be necessary factor of our lives. Additionally economic fraud is the most frequently occurring problem in financial enterprise, authorities and company businesses. Fraud may be described as wrongdoer deception with movie of acquiring financial advantage. Credit card frauds are without problems targets by fraudsters. As e-commerce and some different online sites have improved the online payment transactions, increases the risk for online frauds. Credit card fraud normally occurs when the card was stolen for any of the unauthorized objectives, or even when the fraudster uses the credit card information for his use. In the project, machine learning algorithms used such as Random Forest, K-Nearest Neighbor, Local Outlier Function, K-Means Clustering. Similar set of Algorithmic program are implemented and tested utilizing an online dataset. The performance based on the accuracy and low false positive rate. Results show that each algorithm can be used for credit card detection with higher accuracy. But as compare to other algorithms, K-Nearest Neighbor is considered as the best algorithm that is used to detect fraud.

Index Terms – Random Forest, K-Nearest Neighbor, Local Outlier Function, K-Means Clustering

Introduction

Due to increasing use of E-trade, over there has been large use of credit cards for online shopping which brought about a large wide variety of frauds narrated to credit cards. The most important goal is to make a fraud detection set of rules, which unearths the fraud transactions with less time and lofty accuracy with the aid of making use of system acquiring based class algorithms. As technology is transferring send swiftly, the price by way of cash is reduced and online payment receives multiplied, this enables way for the fraudsters to make unidentified transactions[2].

There're a lot of two types of credit card frauds. One is theft of physical card, and different one is stealing sensitive facts from the card, for instance card number, cvv code, type of card and other. Through stealing credit card information, a fraudster can negate a massive wide variety of money or make a huge amount of buy prior to cardholder reveals out. As a consequence of that, corporations use several machine acquiring strategies to look which transactions are fake and which are not[3].

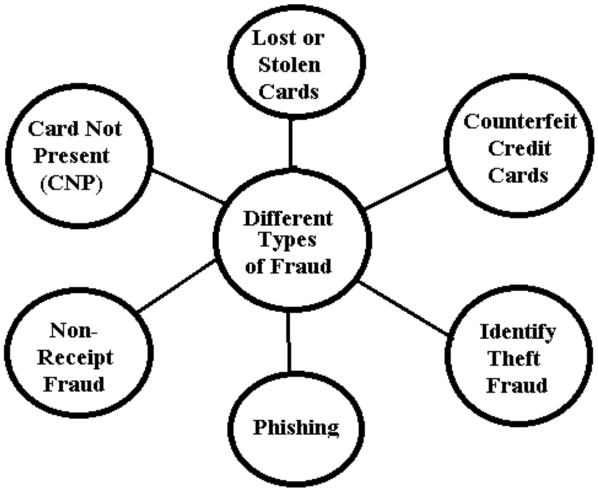

Figure 1: Different Types of Fraud

The surge in charges reaches as consumer spending plummets, leaving card issuers and consumers at a swiftly blooming risk of account fraud. The rise in such attempts reaches as millions of people failed to make card payments.

Fraud detection entails monitoring and analyzing the behavior of several users in order to estimate detect or avoid undesirable behavior. In order to identify credit card fraud detection productively, we need to understand the various machine learning technologies, algorithms and types involved in detection of credit card frauds.

The objective of this paper is to analyze several machine learning algorithms, for instance Random Forest (RF), K-Nearest Neighbor (KNN), Local Outlier Fuction (LOF) to work out which algorithm is the best for credit card fraud detection.

I. Literature survey

The research on credit card fraud detection utilizes both Machine Learning and Deep Learning algorithms. In this section, we improve the task done in two distinct points : (i) The methods that are readily accessible for fraud detection, and (ii) The strategies that are accessible to the imbalanced information in the dataset. To handle the imbalanced information several of the strategies are available. They're as (A) classification methods (B) sampling methods (C) resembling techniques. In this place are several of the Machine Learning algorithms that are utilized for credit fraud detection are support vector machine (SVM), decision trees, logistic regression, gradient boosting, k-nearest neighbor, etc[1].

Deep learning is part of the main and considerable strategies being used for the detection of fraud within the credit card. These kinds of networks have a posh distribution of knowledge that is incredulous troublesome to acknowledge. Deep auto encoders are used in several stages to extract the vast majority of efficient features of the info and for classification functions. Also, higher accuracy and low variance are accomplished amongst these networks[4].

Supervised learning categorizes the dataset utilizing training information whereas unsupervised learning categorizes it utilizing clustering technique. A current near was to solve this plight utilizing deep learning as said by AutoEncoder and Confined Boltzmann Machine. The writers concluded that supervised learning is greater proper for historical database in credit card fraud detection. The set aside of Undersampling on the posterior probability of machine learning version is analyzed by utilizing Bayesian minimum risk theory. Support Vector Machines besides with Random Forests are discussed. The entity utilized in paper utilizes decision tree which is price tag sensitive. The paper uses data mining technology to create credit card acquiring fraud analysis version as said by mass credit card transaction information and merchant materials, and additionally developed merchant fraud risk management system[5].

In this paper, discuss about the imbalance of class and how to handle it and moreover discuss how to work on huge dataset. Detection of credit card fraud for new frauds might be problematic if new information has severe changes in fraud styles. Replacing the version is hazardous as machine learning algorithm take big time for training in spite of predicting[6].

Due to the increasing use of credit cards as the primary means of payment (both online and for everyday purchases), the rate of fraud tends to increase. The use of big data has made manual methods of detecting fraudulent transactions. The use of intelligent techniques has been embraced by financial institutions impractical because they are time consuming and inaccurate. An algorithmic intelligence (CI)-based fraud strategy is included in these intelligent fraud strategies. Statistical fraud detection methods can be divided into two broad categories: supervised and unsupervised. A supervised fraud detection method estimates versions based on a sample of fraudulent or legitimate transactions to classify new transactions as fraudulent or legitimate, whereas an unsupervised fraud detection method detects outliers' transactions as potential fraudulent instances [7].

Classification of credit card transactions is normally a binary classification problem. The credit card transaction can be viewed as a legitimate transaction (negative class) or as a fake transaction (positive class). Fraud detection can be viewed as a data mining classification problem, in which credit card transactions are appropriately classified as legitimate or fraudulent [8].

Countless versions are implemented for fraud detection. In any version distinct algorithm are used. New information that has severe changes in fraud patterns will make it difficult to detect credit card fraud for new frauds. It is risky to replace the version since machine learning algorithms require much training time while predicting [8].

II. System Architecture

An overview of the complete process of credit card fraud detection is shown in the diagram in Figure 2, each of whose steps are explained in this section. The preliminary step for the detection of fraud and non-fraud cases is data collection and necessary preprocessing to convert it into a form, which can be used by machine learning algorithms. The project is implemented in python platform.

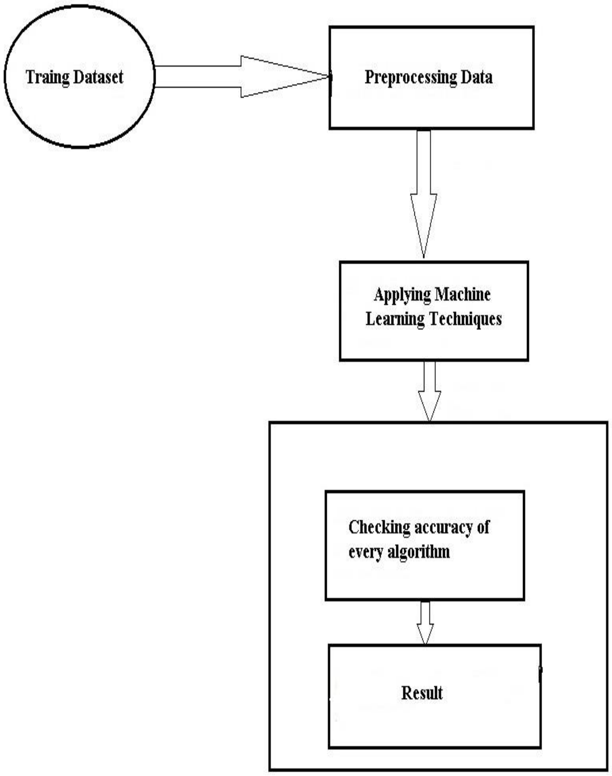
Traing Dataset
Preprocessing Data
Applying Machine Learning Techniques
Checking accuracy of every algorithm
Result

Figure 2: System Architecture

III. Proposed Methodology

A. Dataset

In this research, the Credit Card Fraud Detection is used, which is downloaded from Kaggle. This dataset contains transactions, occurred in two days, made in September 2013 by European cardholders.

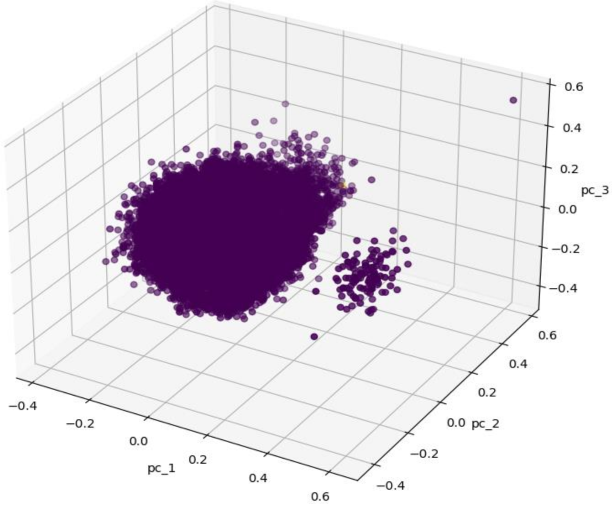

Figure 3: Principal Component Analysis (PCA)

The dataset contains 31 numerical features. Since several of the input variables contains financial information, the Principal Component Analysis (PCA) transformation of these input variables were performed in order to keep these input variables were performed in order to keep these information anonymous. Three of the delivered features weren't transformed. Option "Time" shows the time between initially transaction and the any other transaction in the dataset Option "Amount" is the number of the transactions made by the credit card. Option "Class" serves the label, and takes only 2 values: either 1 if fraud case or 0 if not the fraud case.

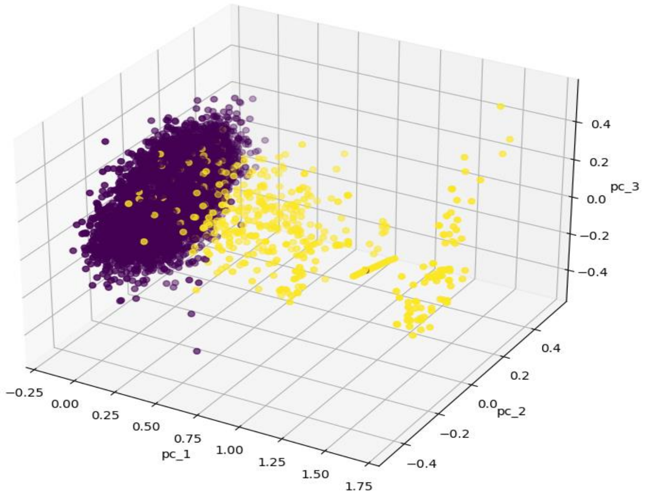

Figure 4: Principal Component Analysis Based on Attributes

Hence distribution ratio of classes plays an important role in algorithm accuracy and precision, pre-processing of the data.

B. Preprocessing

Data preprocessing is a data mining technique to spin the raw information collected from different sources into cleaner info that's more proper for work. In other words, it's a preliminary step that takes all of the accessible info to organize it, sort it, and merge it.

Option choice is a essential technique, which chooses the variables that are the vast majority of pertinent in the delivered dataset. Carefully selecting necessary option and removing the less important one can decrease overfitting inhace accuracy and reduce training time.

C. Results

1. K-Means Clustering

Using this algorithm, input data is categorized into specified numbers of groups. The unsupervised learning algorithm is used when there is no prior knowledge about a particular class of observations in a dataset. An algorithm that classifies n data points into k clusters based on predetermined criteria. Clustering will be done using k-clusters based on similarities among the clusters.
The K-Means Clustering algorithm obtained following results:

Precision: 46.51%
Recall: 54.0%
Accuracy: 79.63%

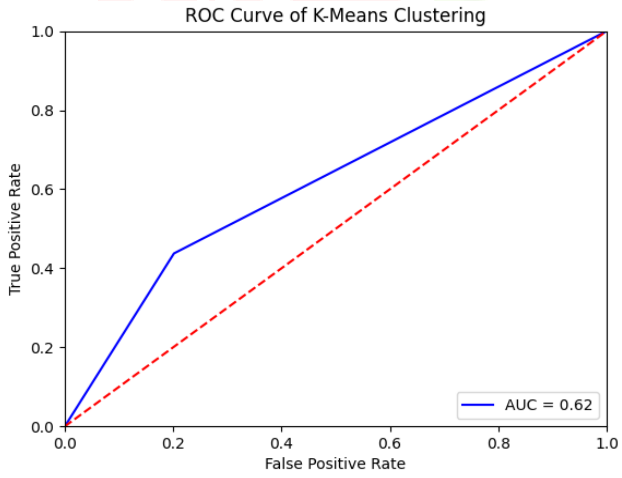

Figure 5: ROC Curve of K-Means Clustering

In the above graph, at point 0.2 false positive rate - the true positive rate is in between point 0.4 to 0.6. It shows that the false positive rate is minimum and true positive rate is maximum. Because this algorithm divide input data into different predefined groups. Each group would hold the data points most similar to itself, and points in different group would be dissimilar to one another. By applying this algorithm before implementation of other algorithms on the dataset the accuracy of algorithm is increases. The accuracy of this algorithm after preprocessing is 79.63%.

2. Random Forest

The Random Forest algorithm is a supervised learning algorithm. This algorithm is utilized for both regression and classification purposes. But, this algorithm is mostly used for classification problems. Generally, a forest is made up of trees and similarly, the Random Forest algorithm made the decision trees on the sample information and receive the prediction from each of the sample data. Later Random Forest algorithm is a co-ordinates method. This algorithm is greater than the single decision trees due to it reduces the over-fitting by equating the result.[1]

	precision	recall	F1-score	support
0	1.00	1.00	1.00	28432
1	0.28	0.29	0.28	49
Accuracy			1.00	28432
Macro avg.	0.64	0.64	0.64	28481
Weighted avg.	1.00	1.00	1.00	28481

Table 1: Output of Random Forest

In the above figure, precision, recall and f1-score is shown for random forest algorithm. The accuracy of this algorithm is 99.75%.

3. Local Outlier Function

Local Outlier Function is a unsupervised machine learning algorithm that identifies outliers regarding to the local neighborhoods as opposed to using the all data distribution. LOF is a density-based technique that uses the nearest neighbor quest to identify the anomalous points. The benefit of using an LOF is identifying points that are outliers relative to a local cluster of points. For instance, when using the local outlier factor technique, neighbors of a certain points are identified and compared against the density of the neighboring points.

	precision	recall	F1-score	support
0	1.00	1.00	1.00	28432
1	0.02	0.02	0.02	49
Accuracy			1.00	28481
Macro avg.	0.51	0.51	0.51	28481
Weighted avg.	1.00	1.00	1.00	28481

Table 2: Output of Local Outlier Function

In the above figure, precision, recall and f1-score is shown for local outlier function algorithm. The accuracy of this algorithm is 99.65%.

4. K-Nearest Neighbor

K-nearest neighbors (KNN) algorithm is a type of supervised machine learning algorithm which can be used for both classification as well as regression predictive problems. However, it is mainly used for classification predictive problems in industry.

The KNN algorithm obtained following results:

Precision: 92.95%
Recall: 80.43%
Accuracy: 99.95%

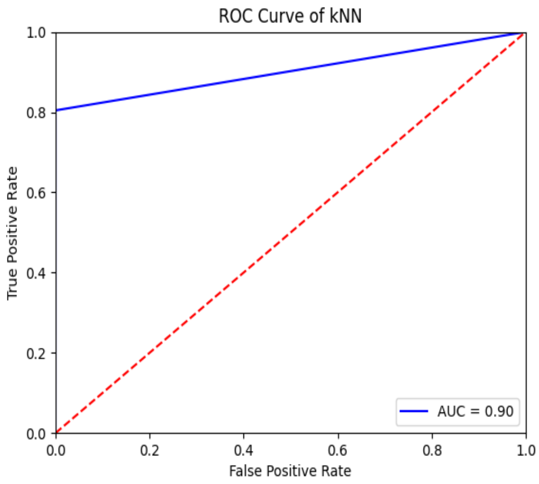

Figure 6: ROC curve for K-Nearest Neighbor

In the above graph, false positive rate at point 0.0 - the true positive rate is at point 0.8. This algorithm shows the highest accuracy i.e. 99.95% and give best results than other algorithms.

IV. CONCLUSION

In credit card fraud detection, we frequently deals with highly imbalanced datasets. For the chosen dataset from Kaggle, we shows that our proposed algorithms are able to detect fraud transactions with very high accuracy and low false positive rate.

Hence for better performance, our result shows that classification of algorithms done by preprocessing data rather than raw data. Because of applying preprocessing technique and K-Means algorithm on the dataset, output of algorithms is with high accuracy and give best results. Hence comparison was done and it was concluded that K-Nearest Neighbor gives the best results. This was established using accuracy, precision and recall. Balancing of dataset and feature selection is important in achieving significant results.

In future, to enhance the system, other machine learning algorithms or artificial neural networks approaches can be used to detect frauds in credit card.

REFERENCES

[1] R. Sailusha, V. Gnaneswar, R. Ramesh and G. R. Rao, "Credit Card Fraud Detection Using Machine Learning," *2020 4th International Conference on Intelligent Computing and Control Systems (ICICCS)*, 2020, pp. 1264-1270, 2020

[2] D. Tanouz, R. R. Subramanian, D. Eswar, G. V. P. Reddy, A. R. Kumar and C. V. N. M. Praneeth, "Credit Card Fraud Detection Using Machine Learning," *2021 5th International Conference on Intelligent Computing and Control Systems (ICICCS)*, pp. 967-972, 2021.

[3] D. Varmedja, M. Karanovic, S. Sladojevic, M. Arsenovic and A. Anderla, "Credit Card Fraud Detection - Machine Learning methods," *2019 18th International Symposium INFOTEH-JAHORINA (INFOTEH)*, pp. 1-5, 2019.

[4] https://www.jetir.org/papers/JETIR2204420.

[5] 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)*, 2018, pp. 1-6.

[6] Varun Kumar K S , Vijaya Kumar V G , Vijayshankar A , Pratibha K, 2020, Credit Card Fraud Detection using Machine Learning Algorithms, INTERNATIONAL JOURNAL OF ENGINEERING RESEARCH & TECHNOLOGY (IJERT) Volume 09, Issue 07 (July 2020).

[7] J. O. Awoyemi, A. O. Adetunmbi and S. A. Oluwadare, "Credit card fraud detection using machine learning techniques: A comparative analysis," *2017 International Conference on Computing Networking and Informatics (ICCNI)*, 2017, pp. 1-9,

[8] https://www.ijert.org/credit-card-fraud-detection-using-machine-learning-algorithms.