INTERNATIONAL RESEARCH JOURNAL
OF
ENGINEERING & APPLIED SCIENCES

ISSN: 2322-0821(O) VOLUME 10 ISSUE 3
ISSN: 2394-9910(P) Jul 2022 - Sep 2022

www.irjeas.org

Original Article

Analysis and Detection of Fraud in Credit Card Using SILOF

*Franklin Ore-Areche[1], Kelyn Nataly Muñoz-Alejo[2], Cledi Puma-Condori[3]

[1]*Teacher, Department of Agroindustrial Engineering, National University of Huancavelica, Huancavelica - 09381 – Peru*
franklin.ore@unh.edu.pe https://orcid.org/0000-0002-7168-1742
[2]*Agroindustrial Engineer, Department of Agroindustrial Engineering, National University of Huancavelica, Huancavelica - 09381– Peru 2015122021@unh.edu.pe https://orcid.org/0000-0002-2938-4648*
[3]*Studentn, Professional School of Engineering in Food Industries, National University of Juliaca, Juliaca - Peru*
c.pumac@unaj.edu.pe https://orcid.org/0000-0002-5486-773X

Corresponding author: franklin.ore@unh.edu.pe

https://doi.org/10.55083/irjeas.2022.v10i03006

Received: 24 June 2022; Accepted: 05 August 2022

Abstract: Nowadays e-commerce and online transaction is growing rapidly. For online and offline transaction most of the customer uses credit card. Credit card used globally for online transaction, buy goods, product, and payment. The rising use of credit card can increase the chances of fraud in credit card. Credit card system is at risk now. The effect of this fraudulent transaction is on the bank and institute causing a financial loss to them. For the detection of distinguish frauds, several machine learning models are utilized for better prediction. The major objective of this article is to identify the fraudulent transaction and outlier in credit card transaction. The dataset of credit card is unbalanced. There are various techniques by which fraudulent transaction can be detected and we have used these techniques such as isolation forest method, local outlier factor and support vector machine to determine fraud in credit card. We have used different matrices for enhancing the performance and accuracy. At last comparison analysis is done by using isolation forest, support vector machine, and local outlier which give the better result.

Keywords: Credit Card Frauds, Detection of Frauds, Local outlier factor, Machine learning model, Random Forest, SVM.

1. INTRODUCTION

In our daily life the problem of credit card fraud is taking place at a very high rate. The rate of fraudulent activities is rapidly increasing. We can use credit card by both offline and online modes. As now; a lot of transactions are online but still there are full chances of credit card frauds; due to which people worldwide face a lot of problems. The fraudulent activities cause financial loss to many organization, companies, and government agencies as well as trust breach. The credit card frauds can occur by stealing other person credit card or to if someone shares the number of credit card to other. The numbers are increasing day-by-day and can be increased in near future more; so to prevent this, many researchers have focused in this field. There are two major reasons of frauds in the banking transactions. First is that the behaviours of fraudulent attempting and second is the highly imbalanced data. To tackle this problem many techniques and approaches are given.

The dataset which is used here are taken from Kaggle which consist transaction done by credit card by the customer in September 2013 in Europe. Dataset is highly imbalanced Credit card transaction is categorizing in two categories e.g., fraudulent and non-fraudulent. These two classes create anomalies which can be detected by using machine learning algorithm. Therefore, the current article is used to build a model for the detection of frauds in credit card. The different techniques that we are applying are isolation forest, support vector machine, and local outlier factor. The article is organized in five distinguish parts. Sections II describe the existing literature. Section III explains the proposed architecture. Section IV explains the detail of dataset. Section V describes the detail explanation of methodology. Section VI shows the experimental setup and Results. Section VII shows the conclusion of this paper.

2. LITERATURE REVIEW

Changjun Jiang [1] to see the transaction behaviour of cardholder; proposed a novel fraud detection method in which it uses the historical transactions data of cardholder and divide them into various groups. It finds that the behaviour of same groups of members is similar on the basis of transactions amount it distribute cardholder into three levels, low, medium and high with the help of clustering method in each of group the transaction is combined with the help of proposing a theory of window sliding. The cardholder's behaviour form is characterized by deriving surplus feature taken from window. On the basis of combined transactions for every single cardholder it removes assemble behavioural patterns and then classifier is trained for every group respectively. Detection of fraud online takes place by classifier and if it found a forge in a newly transactions then feedback mechanism plays role to resolve the drift problem.

Eunji Kima, Jehyuk Lee [2] proposed a hybrid ensemble method. To handle the card and association issue it concludes the deep learning method it introduced two framework, champion and challenger. Respectively both champion and challenger are the attitude of hybrid ensemble and deep learning model. To estimate the large amount of real-world transactions dataset it manipulated several realistic metrics. For the post-lunch execution, it set up these models into FDS after completing the offline testing. The winning model divided into two parts one is off-line and other one is post-launch testing. The main task of champion and challenger framework task is to compare the two models it was also used deep learning model to work with FDS.

Vaishnavi Nath Dornadula [3] proposes a method called a fraud detection method used for streaming transactions data. The objective of this is to examine the details of previous transactions and the behavioural pattern is to remove of the customer. For detecting fraud multiple supervised and semi supervised techniques are use. Based on transactions amount cardholders are clustered in various groups. Sliding window technique is used to remove the behavioural pattern of the groups. Next is to train the classifier. And then that classifier will be chosen which has better rating to detect fraud. Generally, in real world researcher resolve the concept of drift problem, to deal with the imbalance data it used Matthews's correlation coefficient. And to balance the dataset it used SMOTE, where the classifier was working better than in case of previously one. To deal with imbalance data it can also use one class classifier. For better performance and outcome, it uses random forest, decision tree, and logistic regression.

Ankit Mishra[4] focus to misclassifying transactions. Nowadays all the payment and transfer of money is done by credit card due to this the possibility of fraud is increases. Cybercriminal and hackers done the fraud in transaction so the author try to judge the several classifier and metric by examine the different classification problem. The model deal with the problem of genuine transaction which is fraud and it also find the fraud.

Yvan Lucas, P.E. Portier, L. Laporte [5] For the problem of sequence classification they proposed the sate-of-the art method. It uses HMM technique to increase the effectiveness of classification task it used a HMM feature model to check whether transactions is fraudulent or not in credit card it acquire the state of the art approach. Presenting the experimental result and ecommerce transaction with various classifier. It demonstrates the heftiness of the approach by a hyper parameter by relating various solution the issue of structural missing values can be resolved. HMM based feature is used to detect the anomaly in credit card transactions. This work is helpful for feature engineering in sequential data.

3. PROPOSED TECHNIQUE

The proposed technique can be utilized for the identification and detection of transaction frauds in credit card. In this paper, we detect the fraudulent transaction and genuine transaction. There are distinguished machine learning models used to build an intelligent machine such as Isolation Forest, support vector machine, and local outlier factor. Comparison analysis is doing by using these machines learning techniques and finding the best technique which give the highest accuracy and good performance. Figure 1 show the flow work of system architecture.

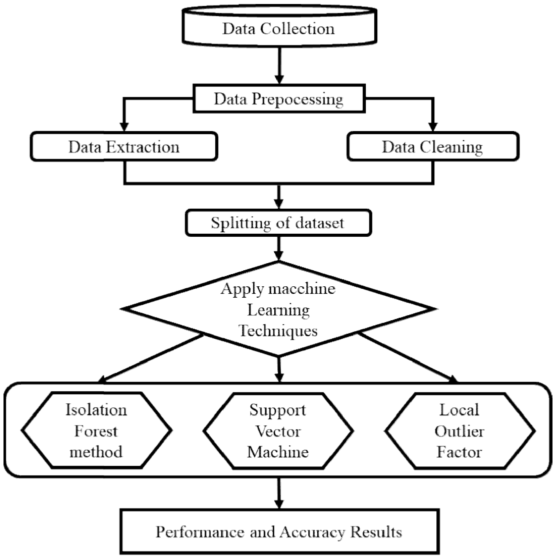

Figure 1 - System architecture

4. DATASET

The dataset which we used in our research is founded in Kaggle. The dataset was built for transactions done during the September 2013 in Europe. There is total 284,807 number of transactions which have approx. 500 fraud transactions within two working days which indicates the imbalance of dataset with only 0.172% fraud transactions. The result of PCA transformation is input numeric variables in dataset. Because of confidentially matters it does not give original feature and information of background of the data. The principal components such as feature V1, V2, V28 are achieved with PCA. Time and Amount are the only features which cannot be transformed with PCA. The feature time can be defined as consisting of the second which elapsed in between every transaction respectively in the dataset whereas the feature amount can be defined as the transaction amount which helps in dependent cost learning. Class is defined to give response to variable now if it gives value 1 it means the case is fraud and otherwise it will give value.

5. METHODOLOGY

5.1. Isolation Forest Method

Isolation method focuses on something unusual and different, so which is why it is called anomaly and we use it here for recognition of unusual patterns. So eventually we call it Anomaly Detection algorithm. These days, Anomaly Detection is having wide application across all industries (Banking, Finance, Healthcare, Manufacturing and Networking). This works same as Decision Tree algorithm, which starts with node root and keeps going on other space. Say suppose, we have to identify a mole in something similar data set by eyes, we will not be able to identify. Therefore, we have used such methods to find and improve the fraud detection in credit card use.

If in a large dataset many data are same and one of them is different from others, this method helps in isolating anomalous data in whole dataset. Main benefit of this method is chances of exploiting sampling methods to a dimension which are not allowed to profile based methods. Isolation Forest method aims on anomalies data as it has kind of shorter path as compared to profiling normal data. It mainly helps in creation of fast algorithm to detect the anomalies with very less in memory consumption.

Algorithm is given to figure out the anomalous part.

- Create a profile those data which is normal
- Observe the whole csv data clearly import it
- Report anything which can’t be retained as normal
- Use the formula to get the anomalies score

$$s\text{ (a, b)} = 2^{-E(h(\text{a}))/c(\text{b})}\text{And } c(\text{b}) = 2H\text{ (b} - 1) - 2(\text{b} - 1)/\text{b})$$

In the above equation, b indicates the number of data points, c(b) represents the average path length in binary search tree for unsuccessful search. It normalizes the score in between 0 to 1.

$$E\big(h(a)\big) = \begin{cases} c(b) thens = 0.5 \\ 0\ thens = 1 \\ n - 1\ thens = 0 \end{cases}$$

Identify the score value it indicates the anomalous point in case 1 otherwise considered as the normal point. So, once we find the anomalous part, it isolates the anomalies explicitly in a large data set. Isolation Forest method creates multiple partitions on whole dataset based on random selected patterns and randomly split those by those patterns and features. It normalizes the dataset in different small partitions. And it has become less lousy to find out the final output.

5.2. Local Outlier Factor Method

This method is used to detect outliers which are having different observations from the rest of the observations. Local outlier factor method comes in mind where data is multidimensional. It basically helps in detecting the outliers in data. This method is based on density which indicates the closest region of datapoints. This method is used to calculate the deviation from the local density by comparing the datapoints with neighbours. Outliers are some patterns which do not come as expected outcome. So, figuring out such patterns are the main important objective here in credit card fraud detection.

Outlier detection mainly works in data analysis, anomalies finding, and also helps out to discover upcoming activities in the crucial safety systems. It helps in pre-prediction of many fraudulent activities like credit/debit card theft, claiming fake insurance, stealing taxes, monitoring real time systems, medical areas and various online transactions. So here we used local outlier factor method especially for credit cards fraud detection. So, we have used K-nearest neighbour's detector and followed some algorithm steps to find the local outlier.

In general, we are taking n-neighbour = 20 by default to work correctly. And we have used 'auto' algorithm for calculating n-neighbour and the metric and auto algorithm is consisting of few steps which are given below.

- Calculate distance between points
- Evaluate the distance of kth neighbour.
- Measure the kth neighbour of each datapoints.
- Find out the local outlier.
- Evaluate the local reachability density.
- Compare the results.

So, if n-neighbor is having value of 20 or greater than 20 that is normal observation and if it comes less than 20, then that will be counted as outliers.

5.3. Support Vector Machines

SVM can be applicable in both techniques such as classification and regression. But the main use of SVM algorithm is based on classification problems in other aspects of fraudulent field. It has a very different implementation approach as compared to other algorithms which are present in Machine learning and it has capacity to manage various categories' variables simultaneously. The purpose of Support Vector Machine algorithm is to find the maximum marginal specifier for n-dimensional data points which can be segregated in distinguish classes. In other words, we can say that, it basically helps in choosing the best vectors or points in space for creation of hyperplane and those best vectors are known as Support Vectors. So, that is why this algorithm is known as Support Vector Machine algorithm.

Margin: It shows the gap between those two classes A and B which is separated by line.

Hyperplane: The line which separates the two sets of objects in different categories (Class A and Class B).

Support vectors: Support vectors are indicating the nearest data points to the hyperplane. So, if we talk about types of Support Vector Machine, we will get two types which are linear and Non-Linear. We have already given an example of linear one. If it is possible to separate two classes with a straight line that means it's a Linear and it's not possible to divide with straight line that is called non-linear. So, in term of solution, we have used kernel trick. Kernel uses the fraud transactional data set to create the classification model and it use the Support vector machine to find out the fraud data.

6. EXPERIMENTAL SETUP AND RESULTS

6.1. Evaluation Metric

There are many classifications task use evaluation metric. Evaluation metric are used for the purpose of accuracy and performance of system. For better result and improving the performance used metric. To finding the fraud and genuine transaction we need some standard measurement tools are Precision, Recall, F1-score, and Support and confusion matrix.

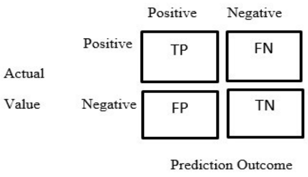

Figure 2 - Confusion matrix

- **True Positive (TP):** It indicates the true predictions those are actually true.
- **False Negative (FN):** It indicates the false predictions those are actually true.
- **True Negative (TN):** It indicates the false predictions those are actually false.
- **False Positive (FP):**It indicates the true predictions those are actually false.

Recall: Recall is the ratio of true positive to the actual positive situation which is show in equation i.e. recall is a value of true positive which find in out of all positive situation. It is also called sensitivity

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

Precision: It is the ratio of true positive over the true positive and false positive which is shown in equation i.e. how many found cases are true positive.

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

F_1 Score: F1 score is also known as F-measure is the harmonic mean of recall and precision. Its value in between 0 and 1 where 1 shows best and 0 shows worst. Shown in equation

$$\text{F1} = \frac{2 * (\text{Precision} * \text{Recall})}{Precision + Recall}$$

Support: it represents the object values which are correctly predicted.

Table 1- Classification report of isolation forest

	Precision	F1-score	Recall	Support
0	1.00	1.00	1.00	28432
1	0.22	0.22	0.22	49

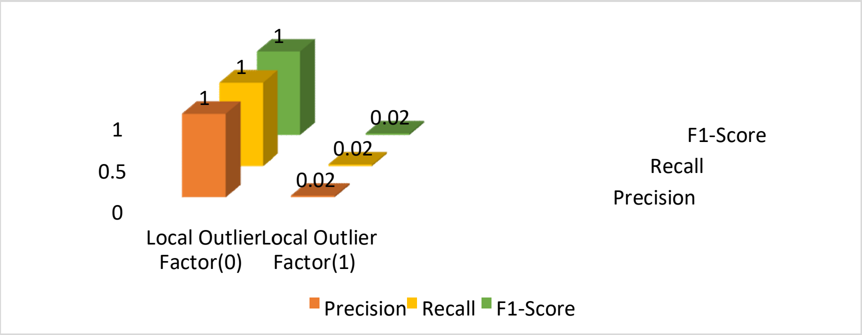

Figure 3- Representation of parameters for isolation forest

Table 2 - Classification report of Local Outlier Factor

	Precision	Recall	F1-score	Support
0	1.00	1.00	1.00	28432
1	0.02	0.02	0.02	49

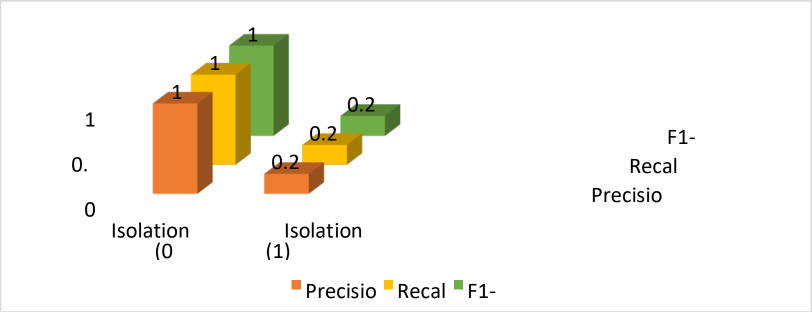

Figure 4 - Representation of parameters for local outlier factor

Table 3- Classification report of Support Vector Machine

	Precision	Recall	F1-score	Support
0	1.00	0.70	0.82	28432
1	0.00	0.37	0.00	49

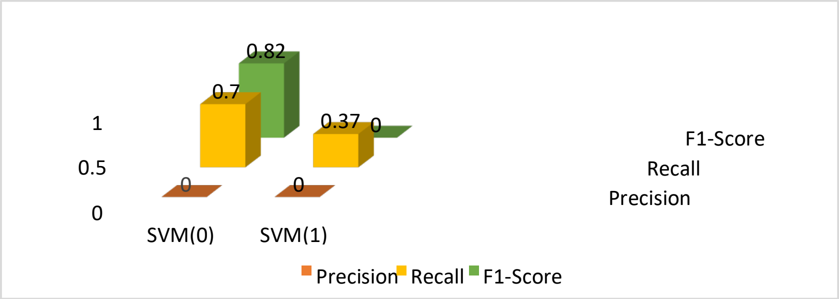

Figure 5 - Representation of parameters for support vector machine

6.2. Experimental Results

By comparing all the three models in this graph on the bases of accuracy which showing that Isolation Forest method is the best method to detect fraud in credit card transaction. Isolation forest techniques giving the highest accuracy i.e. 99.74% as compared to local outlier factor and support vector machine.

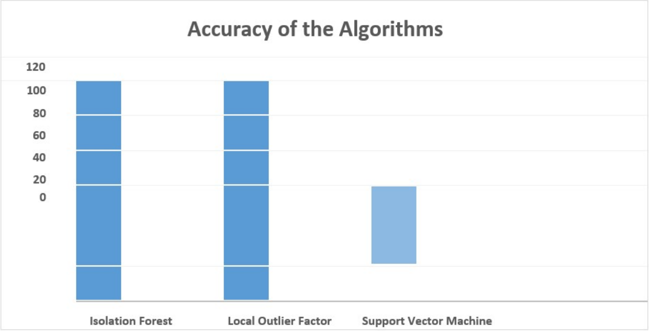

Figure 6 - Graph representing accuracy of various fraud detecting algorithms

In table 4 it shows the accuracy of all three techniques such as isolation forest, support vector machine and local outlier factor.

Table 4- Comparison of various techniques

Techniques	Accuracy
Isolation Forest	98.65
Support Vector Machine	69.87
Local Outlier Factor	98.58

This pie chart shows the percentage of fraud detection method by using different machine learning techniques. Local outlier factor shows 37 %, fraud detection rate, support vector machine shows 26% and isolation forest method also show the fraud detection rate.

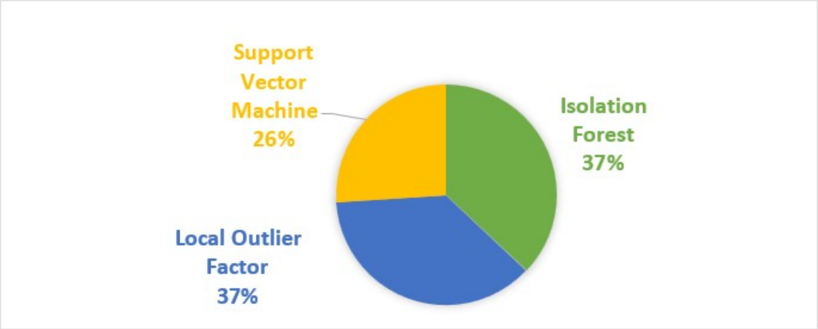

Figure 7 - Fraud detection rate

7. CONCLUSION AND FUTURE SCOPE

The paper proposes three different machine learning techniques that focus on outlier detection. The technique which used in this paper are isolation forest, local outlier factor and support vector machine. The paper examined the performance of credit card transaction on the bases of these techniques. Isolation forest gives the highest accuracy as compared to local outlier factor and support vector machine. The accuracy of isolation is 98.65%, local outlier 98.58% and SVM 69.87%. On the bases of the performance and accuracy isolation forest is the best method to detect the fraud in credit card transaction. Future work can be done in deep learning in term of increasing accuracy by enhancing the sample size, at the cost of computational expense.

REFERENCES

[1] Tanupriya Choudhary, “An Efficient Way to Detect Credit Card Fraud Using Machine Learning Methodologies” IEEE Second International Conference on Green Computing and Internet of Things, pp. 591-597, 2018.

[2] Fabrizio Carcillo, “Combining unsupervised and supervised learning in credit card fraud detection” Elsevier Inc. pp.1-15, May 2019.

[3] Sahil Dhankhad, “Supervised Machine Learning Algorithms for Credit Card Fraudulent Transaction Detection: A Comparative Study” International Conference on Information Reuse and Integration for Data Science, IEEE, pp.122-125, 2018.

[4] N. Malini, “Analysis on Credit Card Fraud Identification Techniques Based On KNN and Outlier Detection” 3rd International Conference on Advances in Electrical, Electronics, Information, Communication and Bio-informatics, IEEE, 2017.

[5] Maja Puh, “Detecting Credit Card Fraud using Selected Machine Learning Algorithms” MIPRO, pp. 1250-1255, May 2019.

[6] Sangeeta Mittal, “Performance Evaluation of Machine Learning Algorithms for Credit Card Fraud Detection” 9th International Conference on Cloud Computing, Data Science & Engineering (Confluence), pp.320-324, 2019.

[7] M. Suresh Kumar, “Credit Card Fraud Detection Using Random Forest Algorithm” 3rd International Conference on Computing and Communication Technologies (ICCCT), IEEE, pp.149-153, 2019.

[8] Changjun Jiang, Jiahui Song, "Credit Card Fraud Detection: A Novel Approach Using Aggregation Strategy and Feedback Mechanism" IEEE Internet of Things Journal, March 2018.

[9] Eunji Kima, Jehyuk Lee, "Champion-challenger analysis for credit card fraud detection: Hybrid ensemble and deep learning" Elsevier Ltd, pp.214-224, 2019.

[10] Vaishnavi Nath Dornadula, "Credit Card Fraud Detection using Machine Learning Algorithms" International conference on recent trends in advanced computing (ICRTAC),pp. 631–641,2019.

[11] Anuruddha Thennakoon, "Real-time Credit Card Fraud Detection Using Machine Learning" 9th International Conference on Cloud Computing, Data Science & Engineering (Confluence), IEEE, pp.488-493, 2019.

[12] Aman Srivastava, "Credit Card Fraud Detection at Merchant Side using Neural Networks" 3rd 2016 International Conference on Computing for Sustainable Global Development (INDIACom)" IEEE, pp.667-676, 2016.

[13] Ong Shu Yee, "Credit Card Fraud Detection Using Machine Learning as DataMining Technique" e-ISSN: 2289-8131, Vol. 10, pp.1-4, 2016.

[14] Dejan Varmedja, "Credit Card Fraud Detection - Machine Learning methods" IEEE, May 2019.

[15] Swati Patel, "Ensemble predictive model for detecting credit card fraud in e-commerce", International Research Journal of Engineering & Applied Sciences, ISSN(O): 2322-0821, ISSN(P): 2394-9910, Volume 6 Issue 3, pp.30-32, 2018

Conflict of Interest Statement: *The authors declare that there is no conflict of interest regarding the publication of this paper.*