iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 10 **Issue:** V **Month of publication:** May 2022

DOI: https://doi.org/10.22214/ijraset.2022.42974

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 10 Issue V May 2022- Available at www.ijraset.com

Fraud Detection in Credit Card Data Using Unsupervised Machine Learning Algorithm

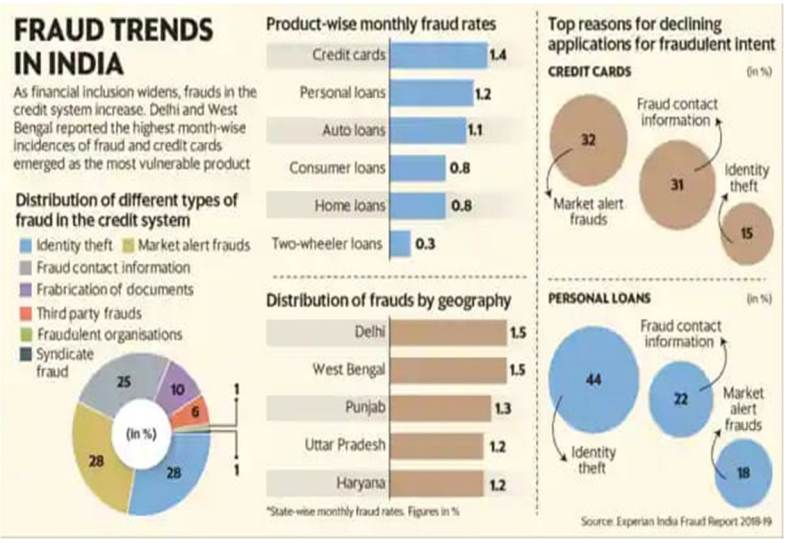
Deepali Kawade[1], Sanket Lalge[2], Dr. Manisha Bharati[3]
[1]*Student at Savitribai Phule Pune University, Department of Technology.*
[2]*Student at, Savitribai Phule Pune University, Department of Technology.*
[3]*Associate Professor at Department of Technology, Savitribai Phule Pune University, Pune, Maharashtra, India*

Abstract: ***We are living in the technology world in which Development of communication and e-commerce has made credit card as the most common technique of payment for both online and offline mode of purchases. As the e-commerce has increased, the buying and selling the product online is becoming very easy and comfortable to everyone in the daily life.***
Due to this, the online payment and online banking with credit card is increased. The fraud also happens when we lost our credit card or it get stole. Recently due to COVID-19 everything has become contactless so the use of credit card has increased. The transaction is done on online shopping and online payment is done from many places so it become difficult to recognise the real transaction and the fraud transaction. So, it becomes difficulty for the bank to stop fraud detection.
In this paper it clearly explains about how fraud can be detected by using the Unsupervised Machine Learning using the algorithm and by using the algorithm technique like Isolated Forest, Local Outlier Factor and One class SVM.
Keywords: Introduction, Machine learning, Supervised learning, Unsupervised learning.

I. INTRODUCTION

From the movement the e-commerce payment system it has found that there have always been a people who will find new ways to access someone's finances illegally. This has become major problem now a days because all the online transaction is being through the credit card. The fraud also happens by sharing the card number and the CVV of the card. In the Covid-19 situation the fraud through the online mode has become more by sharing the card detail and the phone number. So, it become difficult for bank to recognise the fraud happen.

A. Fraud trends in India



Credit Card Fraud Detection with Machine Learning is a process of data investigation by a Data Science team and the development of a model that will provide the best results in revealing and preventing fraudulent transactions.

II. MACHINE LEARNING

Machine learning (ML) is a type of artificial intelligence (AI) that allows software applications to become more accurate at predicting outcomes without being explicitly programmed to do so. Machine learning algorithms use historical data as input to predict new output values.
There are two mostly used method in Machine Learning

They are:
1) Supervised Learning
2) Unsupervised Learning

III. HOW MACHINE LEARNING HELPS WITH FRAUD DETECTION?

The key objective of any credit card fraud detection system is to identify suspicious events and report them to an analyst while letting normal transactions be automatically processed.

1) *Higher Accuracy of Fraud Detection:* Compared to rule-based solutions, machine learning tools have higher precision and return more relevant results as they consider multiple additional factors. This is because ML technologies can consider many more data points, including the tiniest details of behaviour patterns associated with a particular account.
2) *Less Manual Work Needed for Additional Verification:* Enhanced accuracy leads reduces the burden on analysts. People are unable to check all transactions manually.
3) *Fewer False Declines:* False declines or false positives happen when a system identifies a legitimate transaction as suspicious and wrongly cancels it.
4) *Ability to Identify New Patterns and Adapt to Changes:* Unlike rule-based systems, ML algorithms are aligned with a constantly changing environment and financial conditions. They enable analysts to identify new suspicious patterns and create new rules to prevent new types of scams.

A. *Rule-based vs ML-based Fraud Detection*

Rule-based	ML-based
Catching obvious fraudulent scenarios.	Finding hidden and implicit correlations in data.
Requires much manual work to enumerate all possible detection rules.	Automatic detection of possible fraud scenarios.
Multiple verification steps that harm user experience.	The reduce number of verification measure.
Long-term processing.	Real-time processing.

IV. ANOMALY DETECTION

An outlier is nothing but a data point that differs significantly from other data points in the given dataset. Anomaly detection is the process of finding the outlier in the data, i.e., points that are significantly different from the majority of the other data points.
Anomalies can be detected in many ways. In this paper we are using the unsupervised machine learning for finding the anomalies on the fraud transaction of the credit card.

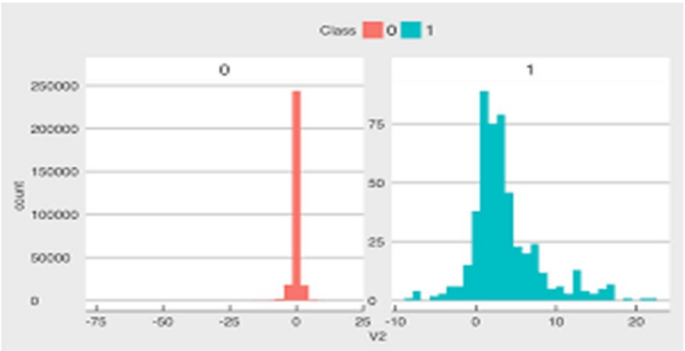


Fig 1: Anomalies detection in credit card Transaction.

In Fraud case implementation as we can see that data is imbalanced with very fee positive Fraud Cases

V. IMPLEMENTATION

In, this paper we are using the Kaggle data set of the credit card to find the fraudulent transaction by using Unsupervised algorithms. The data set we used is not trained with variable it is directly trained to the actual dataset without any labels.

The datasets contain transactions made by credit cards in September 2013 by European cardholders. This dataset presents transactions that occurred in two days, where we have 492 frauds out of 284,807 transactions. The dataset is highly unbalanced, the positive class (frauds) account for 0.132% of all transactions.

To predict the dataset, we had used the three unsupervised algorithm so to test the accuracy and compare the best algorithm. The reason to using the unsupervised in this paper because of the real world, there won't be any labelled variable and for credit card fraudulent unsupervised is the best approach.

The algorithms we are using is:

1) Local Outlier Factor
2) Isolated Forest
3) One class SVM

A. Local Outlier Factor

The LOF algorithm is an unsupervised outlier detection method which computes the local density deviation of a given data point with respect to its neighbours. It considers as outlier point that have a substantially lower density than their neighbours.

The parameter for number of neighbours is typically chosen to be greater than the minimum number of objects a cluster has to contain, so that other objects can be local outliers relative to this cluster, and smaller than the maximum number of close by objects that can potentially be local outliers.

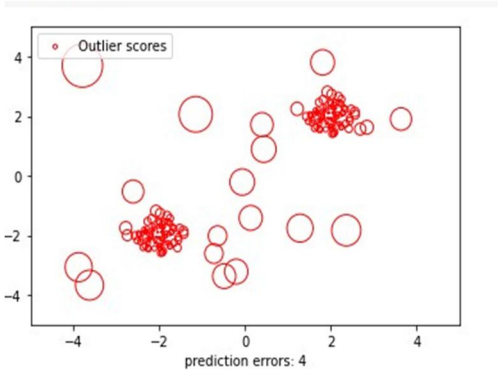


Fig 2. Local Outlier Factor

In the Fig 1 we can see the data points and the outlier scores. As the lower the density from the neighbour points are the outliers.
In the fig 2 we have used data set and showed Local Outlier Factor.

ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 10 Issue V May 2022- Available at www.ijraset.com

```
[9]  1 plt.title("Local Outlier Factor (LOF)")
     2 plt.scatter(X[:, 0], X[:, 1], color="k", s=3.0, label="Data points")
```

```
<matplotlib.collections.PathCollection at 0x7f54fb4bb650>
```

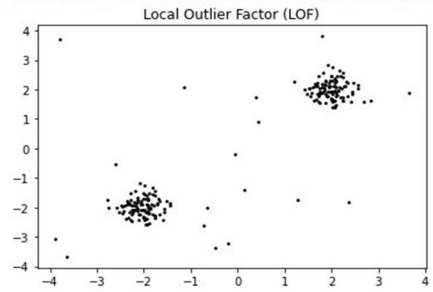


Fig. 3 Local Outlier Factor (Using Dataset)

To understand Local Outlier Factor, we should know about LRD is the Local Reachability Density it is the average reachability distance from its neighbours.

LRD is the Local Reachability Density it is the average reachability distance from its neighbours. Accordingly, to LRD formula, more the reachability distance, less density of points is present around a particular point.

This tells us how the point is far from the nearest cluster points. Low value of LRD tells us the closest cluster is far from the point.

LRD formula:

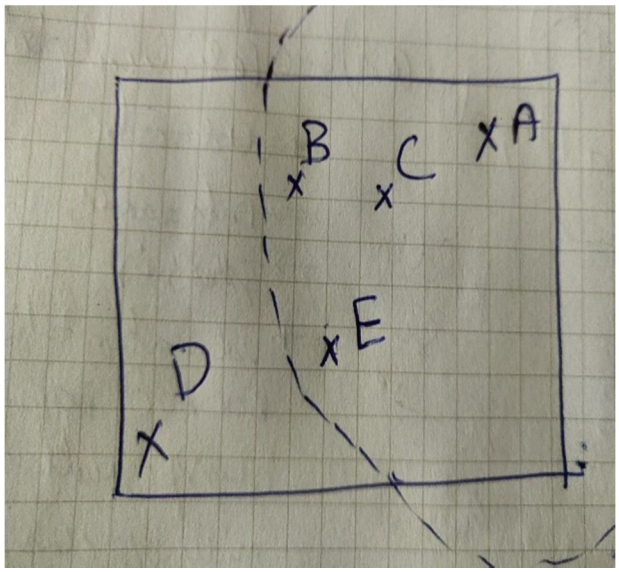
$$LRD_k(x) = \frac{1}{\sum \frac{d_k(x,0)}{|N_k(x)|}}$$

LOF Formula:

$$LOF(x) = \frac{\sum \frac{LRD_k(o)}{LRD_k(x)}}{N_k(x)}$$

LRD of each point is used to compare with the average LRD of its K neighbours. LOF is the ratio of the average LRD of the K neighbours of x to the LRD of x.

K-distance, it is the distance of a point to its k_th neighbour Let us assume k =3 and we calculate LOF of A. From below Fig k-distance means 3rd closet distance from A i.e., E (consider it has greater distance among all other).



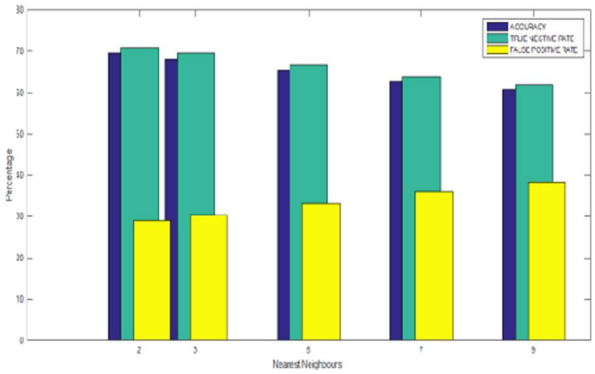


Fig 3: The graph of Accuracy, True Negative Rate, False Positive Rate

B. Isolated Forest

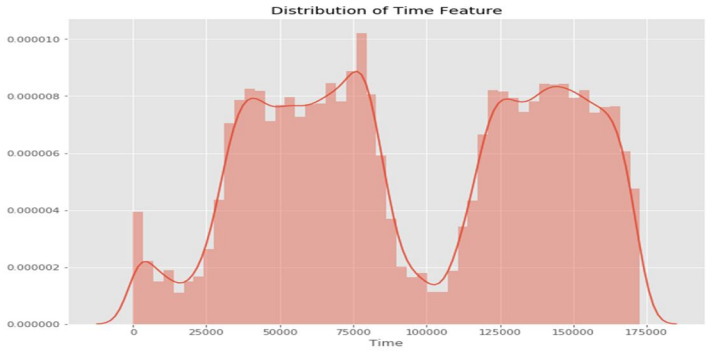
A newer technique used to detect anomalies is called Isolation Forests. The algorithm works because anomalies are data points that are far from regular ones. As a result, anomalies are susceptible to a mechanism called isolation.

The algorithm isolates points by randomly selecting a feature and then split on values between the maximum and minimum values of the feature until the points are isolated from each other. Isolating anomalous observations is easier because only a few generations are needed to separate those cases from the normal observations. On the other hand, isolating normal observations require many more generations. Therefore, an anomaly score is calculated as the path length required to separate a given observation.

This algorithm has low computation complexity and use less memory. It builds a good performing model with a small number of trees using fixed sub-sample sizes, regardless of the size of a data set

In the Fig4 here the anomalies are highlighted by the red colour points and the normal points are highlighted by the green colour points. The red colour points are the outliers.

Graph of Distribution of Time Feature



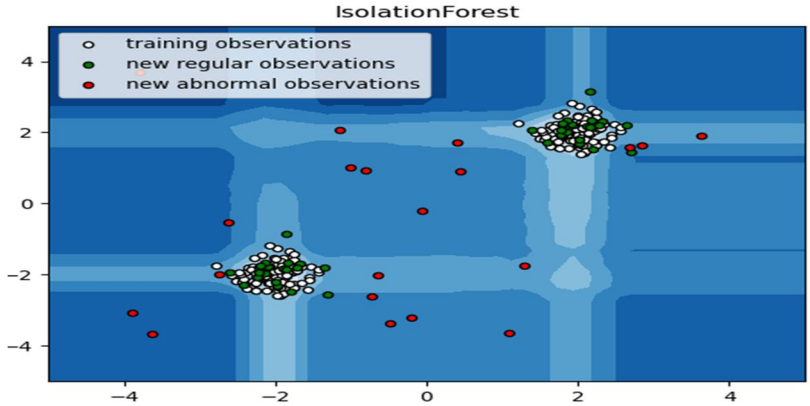


Fig 4: Isolated Forest

Bubble chart of Isolated Forest

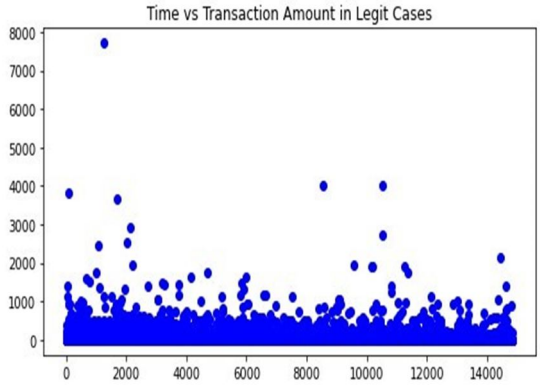
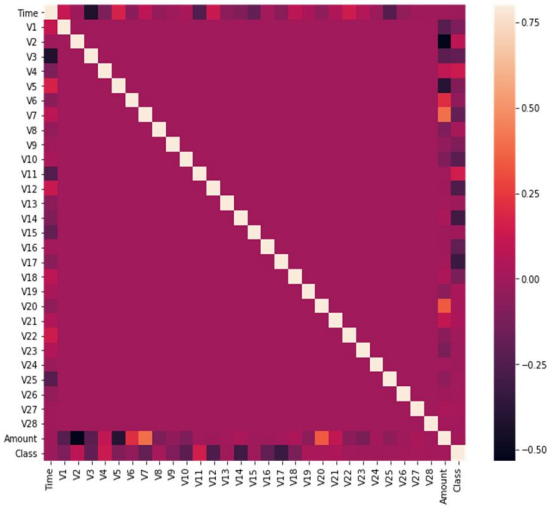


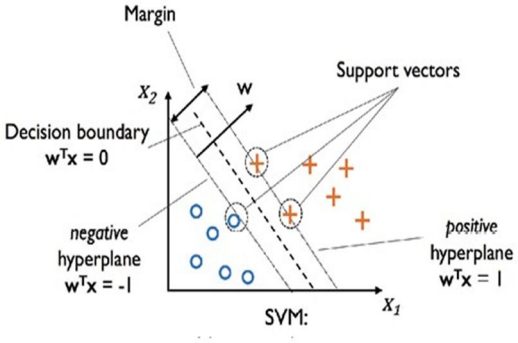
Fig 5: Correlation of Fraud Detection

C. One class SVM



One Class SVM is unsupervised learning algorithms designed for outlier detection. The model is trained on 'healthy' data. The algorithm learns on the normal transactions and creates a model that contains a representation of the data. When introduced to observations that are far away, it will be labels as outlier and return a negative number. When introduced to observations that are close, it will be labelled as inlier, a positive number.

SVM is a supervised machine learning algorithm which can be used for classification or regression problems. It uses a technique called the kernel trick to transform your data and then based on these transformations it finds an optimal boundary between the possible outputs.



D. Working of Support Vector Machine:

Support vector machines focus only on the points that are the most difficult to tell apart, whereas other classifiers pay attention to all of the points. Support vector machines focus only on the points that are the most difficult to tell apart, whereas other classifiers pay attention to all of the points.

In the Fig 4 One class SVM separate's the outlier. The yellow points in the diagram are consider as the anomalies.

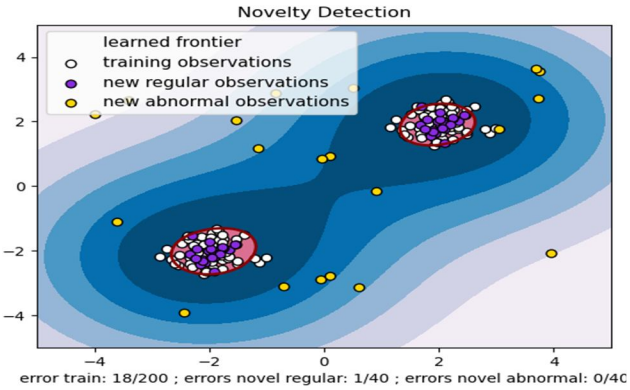


Fig 6. One class SVM

VI. RESULT

We have used the credit card data set for prediction by using the unsupervised algorithms viz, is Isolated Forest, One class SVM, Local Outlier Factor

Table of Algorithms

Algorithms	Accuracy
Local outlier Factor	87.50%
Isolated Forest	95.76%
One class SVM	70.09%

VII. CONCLUSION

From the table we can see the accuracy of the Isolated Forest has accuracy 95.76%, One class SVM has the accuracy 70.09%, Local Outlier Factor has the accuracy 87.50%.

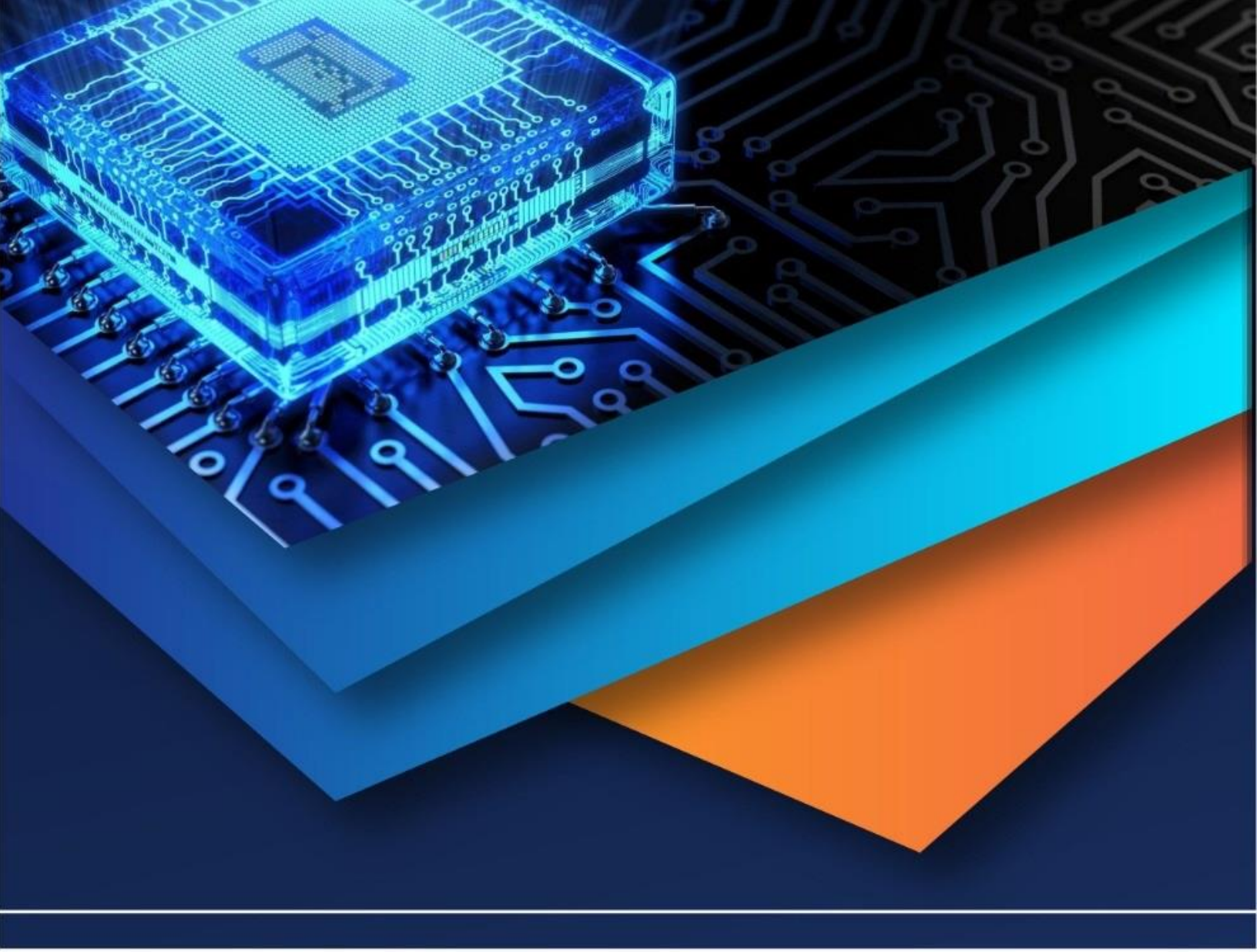
So, from the all three algorithms Isolated and Local Outlier Factor has the good accuracy than the One class SVM. The performance of Once class SVM is very poor compare to the Isolated Forest and Local outlier Factor. In this paper we have concentrated on the Unsupervised learning algorithm so that we can detect anomalies without the knowing the variables.

Unsupervised algorithm is very useful we can detect the anomalies. We can use Unsupervised algorithms to detect the anomalies to any field. We can detect the outlier or the fraudulent transaction for any field of transaction.

Comparative chart							
Algorithm		Precision	recall	F1 score	support	Accuracy Score	Fraud outlier
Local Outlier Factor	0	1	1	1	28432	87.15%	97
	1	0.02	0.02	0.02	49		
Support Vector Machine	0	1	0.46	0.63	28432	70.09%	15425
	1	0	0.46	0	49		
Isolation Forest	0	1	1	1	4987	95.76%	77
	1	0.22	0.22	0.22	49		

REFERENCES

[1] https://colab.research.google.com/drive/17KL4UVirIxpdc9NLSkr16Aq4Ivp8E5d0?usp=sharing
[2] https://colab.research.google.com/drive/1f1xAeY109m-vz9A1xx__2zBi6c9UZYb3?usp=sharing
[3] Credit Card Fraud Detection Using Unsupervised Machine Learning Algorithm, Hariteja Bodepudi,13 August 2021 Available: https://www.researchgate.net/publication/354065310
[4] Credit Card Fraud Detection Using Machine Learning. R Raja Subramanian, Reddy, Kumar, Tanouz, 2021.
[5] https://machinelearningmastery.com/imbalanced-classification-with-the-fraudulent-credit-card-transactions-dataset/
[6] https://neurospace.io/blog/2019/03/predicting-credit-card-fraud-with-unsupervised-learning/
[7] https://www.kaggle.com/code/jiedong00/anomaly-detection-with-unsupervised-learning
[8] Credit Card Fraud Detection Using Unsupervised Machine Learning Algorithms Article in International Journal of Computer Trends and Technology · August 2021
[9] Supervised Machine Learning Algorithms for Credit Card Fraudulent Transaction Detection: A Comparative Study. sahil, Emad, Behroz , 2021
[10] Relative Analysis of ML Algorithm QDA, LR and SVM for Credit Card Fraud Detection Dataset, P. Naveen, B Diwan,2021
[11] Relative Analysis of ML Algorithm QDA, LR and SVM for Credit Card Fraud Detection Dataset, Sadgali, sael, Benabbou,20 feb 2021.
[12] Supervised Machine Learning Algorithms for Credit Card Fraud Detection: A Comparison, Khatri, Arora, Agrawal, 29-31 Jan. 2020
[13] Machine learning Based-Intrusion- Prediction-System PENSEE Journal, Volume51, Issue 01, January-2021.
[14] 14. [23] Gated Recurrent Unit Deep Neural Network-Based-Intrusion-Prediction-to Classify Network-Attacks, PENSEE Journal, Volume51, Issue 01, January-2021.
[15] Network Intrusion Detection System Based on Deep and Machine Learning Frameworks with CICIDS2018 using Cloud Computing, International Conference on Smart Innovations in Design, Environment, Planning Computing (ICSIDEMPC 2020) IEEE.



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY