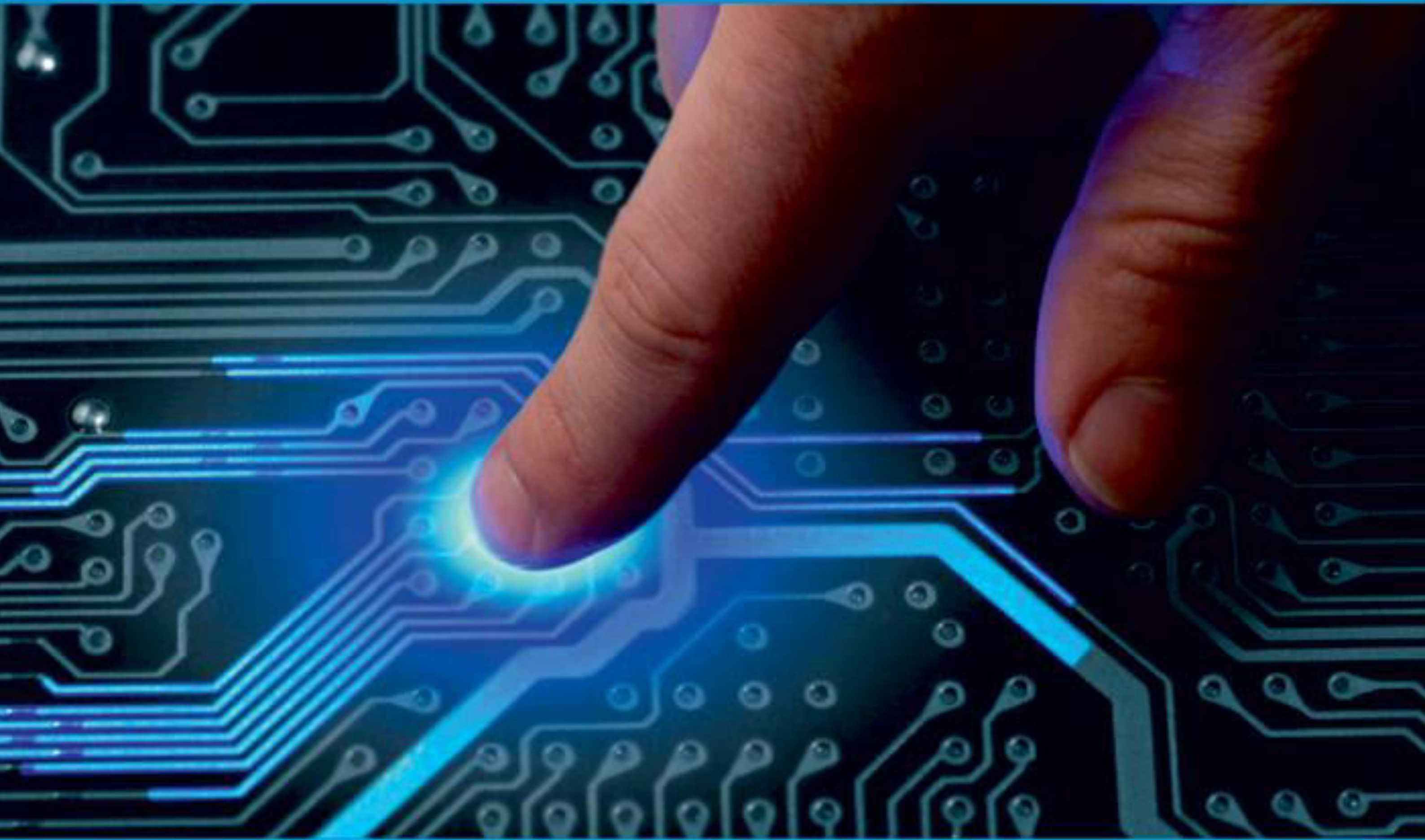
IJIRCCE

e-ISSN: 2320-9801 | p-ISSN: 2320-9798



INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN COMPUTER & COMMUNICATION ENGINEERING

Volume 10, Issue 5, May 2022

INTERNATIONAL
STANDARD
SERIAL
NUMBER
INDIA

Impact Factor: 8.165

International Journal of Innovative Research in Computer and Communication Engineering

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | |Impact Factor: 8.165 |

|| Volume 10, Issue 5, May 2022 ||

| DOI: 10.15680/IJIRCCE.2022.1005243|

Performance Evaluation of Credit Card Fraud Detection Models Using Imbalanced Classification

M. Parvathi[1], M. Madhavi[2], P. Naga Sivani[3]

Department of ECE, Vasireddy Venkatadri Institute of Technology, Andhra Pradesh, India[1,2,3]

ABSTRACT: Credit card fraud is an extremely popular fraud technique worldwide, which results in loss of many billion dollars. Credit card fraud occurs when skimmers steal card information through various methods to create fake cards for use in person or for online purchases. This fraud is extremely easy because in this day and age with all information and data that we have floating out, there are so many points that we can get our information stolen. All banks and credit companies are looking for all sorts of ways to mitigate this scam so that customers are not charged for items that they did not purchase. To tackle this problem, Data Science along with Machine Learning is an obvious solution. One of the inevitable challenges of Credit Card Fraud detection model is dealing with imbalanced dataset. Another challenge includes incorrect flagging of legitimate transaction, due to which real customers might feel harassed. This paper tends to illustrate a model which decides whether a transaction is fraud or not. Our objective is to detect all fraudulent transactions in the meantime minimizing incorrect flagging of legitimate transaction.In this model we are determined on exploratory data analyzing, preprocessing dataset and deploying multiple anomaly detection algorithms such as Isolation Forest, Local Outlier Factor and one-class SVM to handle imbalanced classification problem. The dataset we considered undergoes a type of dimensional data reduction technique called Principal Component Analysis (PCA).

KEYWORDS:Credit card fraud, Imbalanced classification, Principal Component Analysis, Anomalies, Isolation Forest, Local Outlier Factor, Support Vector Machine.

I.INTRODUCTION

Credit card fraud is a form of identity theft, in which a person steals our credit card and obtains cash advances. It is an unauthorized use of the card. It arises when both the cardholder and the card issuer are not conscious of the fact that the card is being used by other people. In addition to credit card being stolen, thieves can use personal credentials of card holder like name, date of birth, address etc. and take loans on our personal bank account. Credit card trickster are exuberant in the usage of new technologies in their schemes for extracting out credit card numbers and PINs for their usage.

Based on the nature of fraudulent activities credit card is classified into different types. They are described as the following:

- **Simple theft (offline fraud)**–It is the most straightforward type of credit card fraud i.e, a stolen card and also the fastest fraud that can be detected.
- **Application fraud**– In this a new card is obtained by the person using false personal credentials.
- **Bankruptcy fraud**– The card holder uses the card though he knows that he is not able to pay and purchases the goods.
- **Internal fraud**– This fraud is due to the bank employees; they steal the card and use it remotely.
- **Counterfeit fraud –** Here only the details of a legitimate credit card are needed, card holder presence is not needed.

To prevent the credit card fraud, there are different methods for the banks and credit card companies to detect the fraudulent transactions. A few of them are Neural networks, Decision trees, Fuzzy clustering approach. Concepts of Machine Learning can be used in fraud detection. They are Supervised Anomaly detection and Unsupervised anomaly detection algorithms.

II.LITERATURE REVIEW

Several studies are done on the topic credit card fraud detection in order to understand it in detail. Literature survey of some of them are given as:

There is an existing model [2] for credit card fraud detection in which fraud detection was done by using machine learning algorithms.First standard models are applied and then hybrid methods are used which are based on AdaBoost and majority voting methods are used publicly available credit card data set is used in order to evaluate the efficiency. Additionally, noise is added to the data samples in order to estimate the rigidity of the algorithm. The results says that majority voting method got good accuracy rate in detecting the fraud.The MCC score of 0.823 says that majority voting is the best method for detecting the fraudulent transactions.

There is another model [3] in fraud detection which was stated as an intelligent model uses an optimized light gradient boosting machine [OLightGBM] for detecting fraud transactions. In this approach to tune the parameters of a light gradient boasting machine (light GBM) a Bayesian -based hyperparameters optimization algorithm is intelligently integrated. Several experiments were done using two real -world public credit card transaction dataset containing fraudulent transactions and legitimate ones in order to find the efficiency and effect of OLight BGM that was proposed. After performing several Operations on data set that provided and comparing other approaches recorded accuracy as 98.40%, Area under receiver characteristic curve (AVC) as 92.40%, precision as 97.34% and FI-score as 56.95.

In other paper [4] we examine the application of linear and nonlinear statistical modelling and nonlinear statistical modelling and machine learning model on real credit card transaction data. These all are supervised used to find which transactions are most likelyto be fraudulent transactions. And five different types of supervised models are explored and compared containing logistic regression neutral networks, random forest, boosted tree and support vector machine model. Among all those models boosted tree model gives the best fraud detection results of (FDR=49.83%) for given particular data.

In another paper [5] several machine learning algorithms like logistic regression (LR), Random Forest (RF), NaiveBayes (NB),multiplayer perception (MLP) are studied in order to find which one give the best results in detecting fraud. Here the aim of this model is comparing all those algorithms that which one will give best accuracy of fraudulent transactions. Hence studies that on all these concluded that Random Forest algorithm give the finest and the best accuracy result. It was build using several metrics such as recall, accuracy and precision. In order to achieve best results, it is important to have feature selection and balancing of data.

In the paper [6] several machine learning algorithms were proposed in order to get good Accuracy and they are Logistic regression, multilayer perception, naive Bayes and Random Forest algorithm by adopting Waikato environment for knowledge analysis (WEKA) tool. Dataset that was used in this model are from European cardholder containing 284,807 transactions. This paper mainly concentrated on which modern method will provide the best results of accurate fraud transactions. And anti-fraud methods were used in order to save banks from major disorders and minimize damages. Here the four machine learning algorithms were compared for accuracy using given data set. It states that among all those algorithms Random Forest algorithm gave the best result of accuracy 99.9%.

III.METHODOLOGY

This paper proposes an approach in which outlier detection algorithms are implemented on credit fraud detection model to detect fraudulent transactions and their performances are evaluated to determine the best algorithm for fraud detection problem based on selected performance measures.

Building credit card fraud detection model includes the following stages:

a) DATASET:

The dataset used in this fraud detection model is obtained from Kaggle[x], the largest platform for data scientists and machine learning experts. This dataset includes credit card transactions made by the European cardholders in September,2013. It contains about 284807 transactions out of which only 492 transactions are proclaimed to be fraud. This means only 0.172% of all transactions account for the positive class(fraud) which clearly demonstrate that dataset is highly imbalanced.

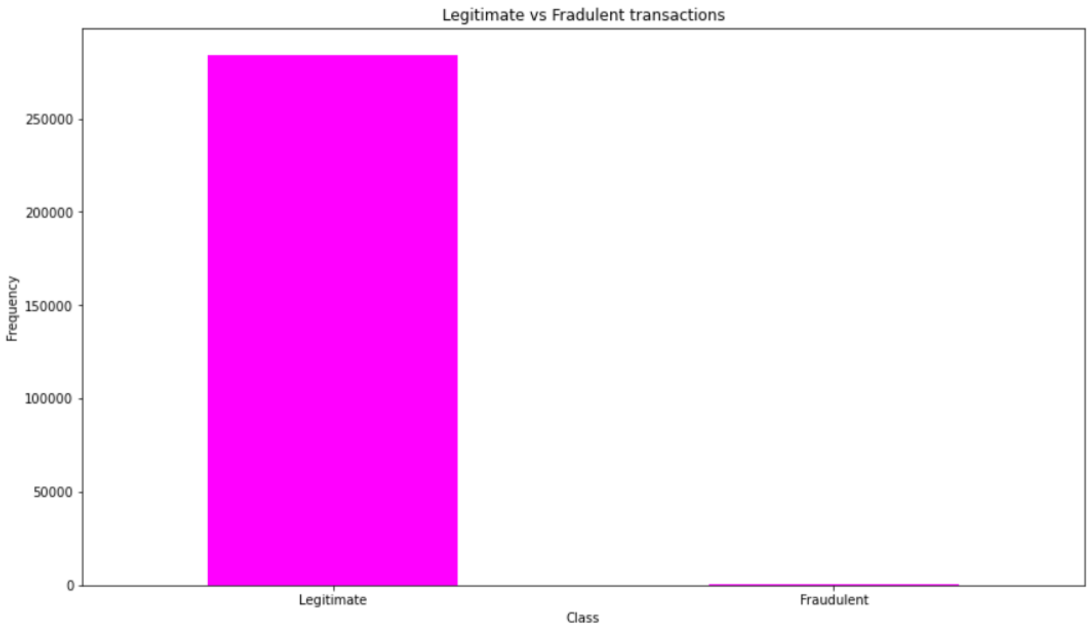
A dimensional data reduction technique called Principal Component Analysis (PCA) is applied to dataset to handle high dimensionality of data. Using this PCA transformation we can reduce the number of features while preserving the important hidden patterns of data. The features in this data set are named as v1, v2…v28 to because of the confidentiality issues whereas 'Time', 'Amount' are the features which does not undergo PCA transformation and 'Class' is the other feature in which value 1 indicates fraud transaction, value 0 indicates legitimate transaction.

b) IMPORTING LIBRARIES:

Firstly, we started building this model by importing the various libraries and necessary packages. Using pandas, we loaded the dataset which is a csv file into pandas data frame and checked for the null values if any.

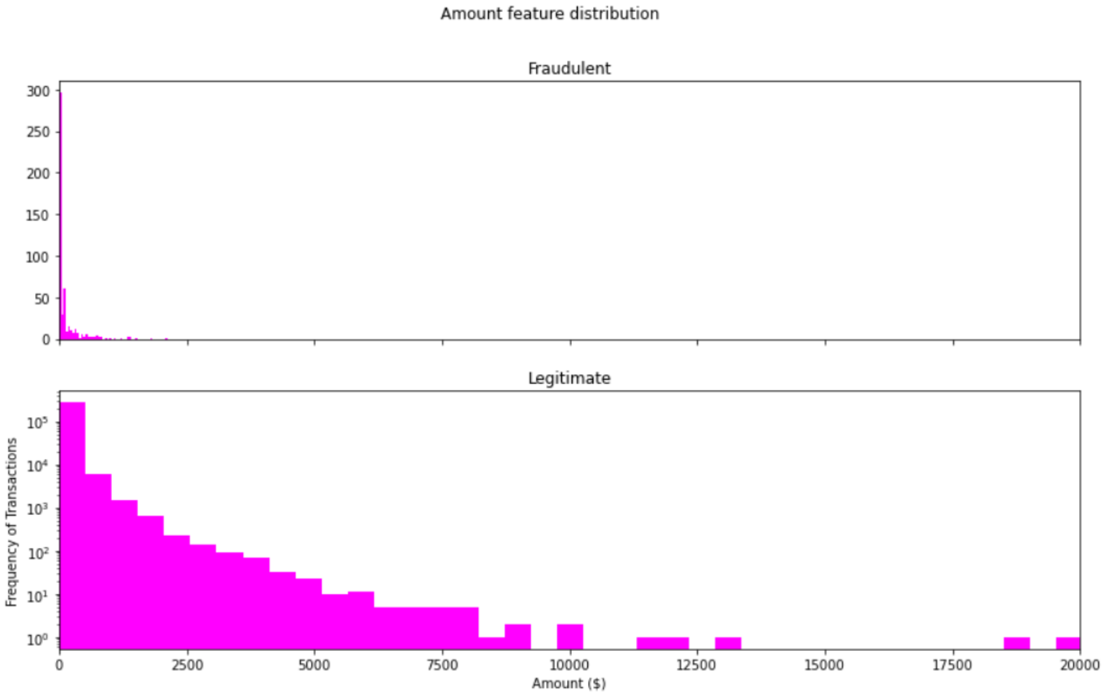
c) EXPLORATORY DATA ANALYSIS /DATA EXPLORATION:

In this section, we perform some basic data exploration to gain better perception. We plot transaction class distribution graph on dataset considered



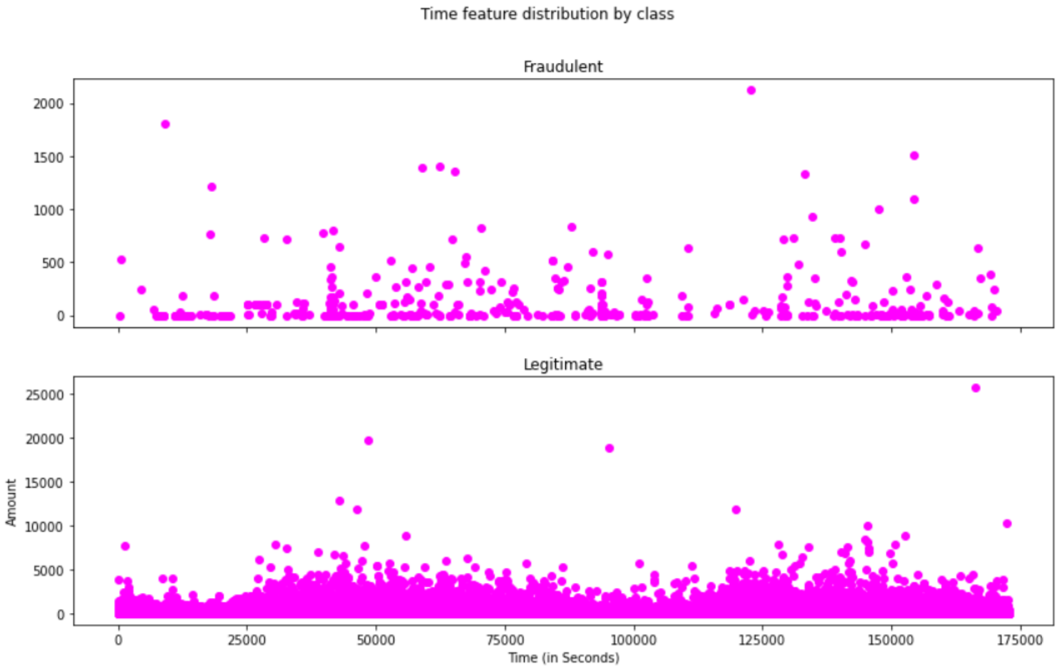
This graph depicts frequency of fraudulent transaction vs non fraudulent transaction. It demonstrates that fraudulent transactions are much inferior than legitimate transaction which clearly portrays that it is an imbalanced classification.

We plot another graph using the dataset which detects amount per transaction by class as shown in below figure.



This graph demonstrates the amounts that are transacted in legitimate case and in fraud case. From this plot we can draw the conclusion that transaction amount is small in fraudulent cases when compared to the amount that is transacted in legitimate case.

In the process of analyzing the dataset we plotted another graph which is transaction time versus class and is as shown below



This plot illustrates the times at which transactions takes place for both legitimate and fraudulent cases. By just looking at the plot we cannot draw an exact conclusion but on careful analysis it is observed that fraudulent transactions strike high during late hours and early hours of a day. And in the mid-day count is significantly low.

d) FEATURE CORRELATION:

Correlation among the features provides the relationship between the features or attributes in the data set. Significantly, it measures the degree of dependance between the features available in the dataset. Correlation among the features in our data is depicted as shown in below depicted figure

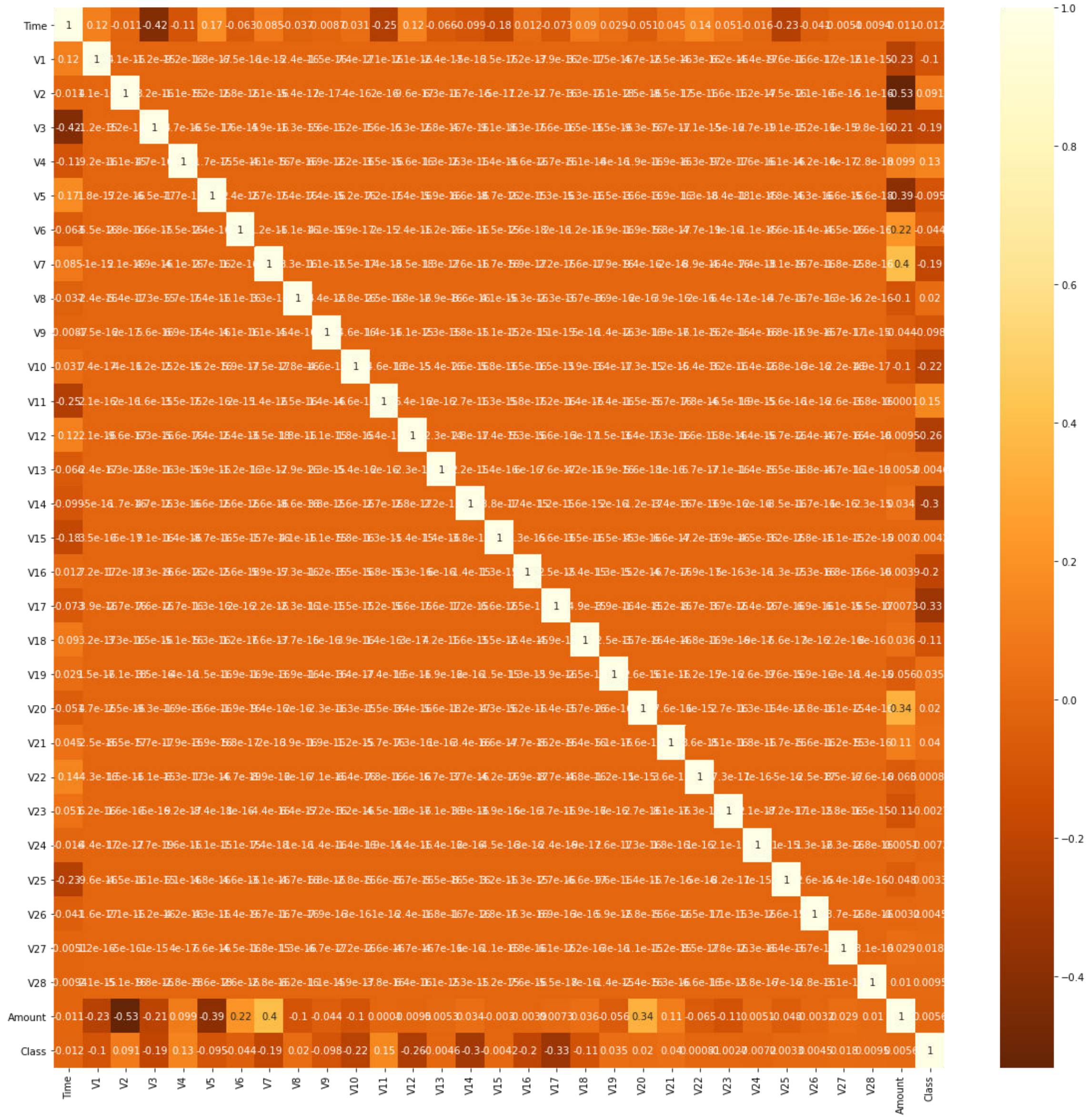


Fig. Heatmap of correlation

From this plot we can draw out the conclusion that there is no strong relationship among the features so based on this correlation analysis it is not necessary to remove any features.

e) IMPLEMENTING OUTLIER DETECTION ALGORITHMS

In this model we are comparing the performances of outlier detection algorithms such as Isolation Forest, Local Outlier Factor and Support Vector Machine. And these algorithms are incorporated within the python Scikit-Learn. The python program that is used to illustrate the approach that this paper suggests is done on jupyter notebook platform, an open-source web application.In this process instead of taking the entire data, only sample of 0.1 fraction of whole data is considered because it requires much preprocessing time as it is a huge data.

ISOLATION FOREST ALGORITHM:

Isolation Forest is significantly used to detect anomalies and accounts for the fact that the anomalies are the data points that are ''few and different". This method requires small memory and has low linear time complexity.This algorithm works similar to the random forest, it consists of multiple decision trees which are constructed by randomly selecting a split value between max and min value of the feature that is selected. This tree separates the anomalies from the other observations as they end up in shorter branches.

LOCAL OUTLIER FACTOR ALGORITHM:

This algorithm evaluates the local density deviation of a given data point based on local neighborhood. The samples that have lower density than their neighbors are considered as outliers. For this algorithm to work well neighbors=20 is considered.

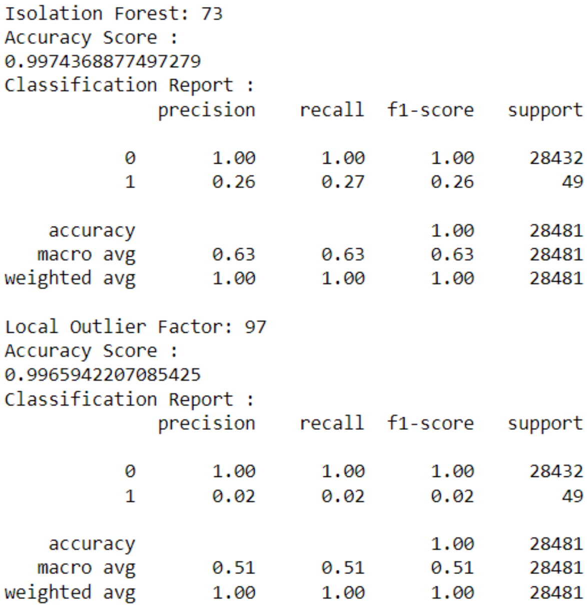
SUPPORT VECTOR MACHINE:

Usually, SupportVector Machine algorithm is a machine learning model that is used to analyze the data and recognize the pattern. Specifically in this model we have used One-class SVM.One-class SVM is used in the situations where the outliers are not depicted well in the training dataset. This algorithm differentiates the target class from the other classes using only the target class training data.

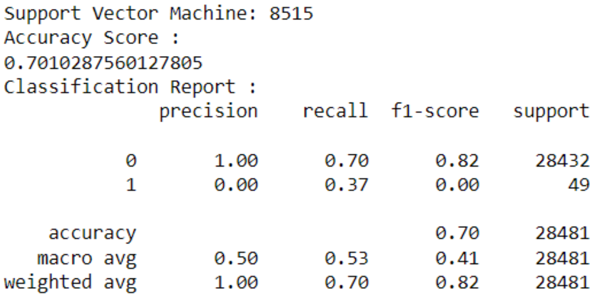
All the three algorithms are applied on this fraud detection model by enclosing them in a dictionary which followed iteration and executing the corresponding code.

IV.RESULTS & DISCUSSION

Only 10% of the entire dataset is used for the evaluation of this fraud detection model. After execution of the machine learning models, the code imprints number of outliers detected i.e, number of False Positives and is compared with actual values .It also prints out the classification report which contains values of performance metrics such as Accuracy Score, recall ,precision, f1-score and support for all the three Machine learning models used. The result obtained is as shown below.



```
Isolation Forest: 73
Accuracy Score :
0.9974368877497279
Classification Report :
              precision    recall  f1-score   support

           0       1.00      1.00      1.00     28432
           1       0.26      0.27      0.26        49

    accuracy                           1.00     28481
   macro avg       0.63      0.63      0.63     28481
weighted avg       1.00      1.00      1.00     28481

Local Outlier Factor: 97
Accuracy Score :
0.9965942207085425
Classification Report :
              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
```



```
Support Vector Machine: 8515
Accuracy Score :
0.7010287560127805
Classification Report :
              precision    recall  f1-score   support

           0       1.00      0.70      0.82     28432
           1       0.00      0.37      0.00        49

    accuracy                           0.70     28481
   macro avg       0.50      0.53      0.41     28481
weighted avg       1.00      0.70      0.82     28481
```

From the above obtained result we can observe that:

- Isolation forest detected 73 outliers or errors, Local Outlier Factor detected 97 outliers and Support Vector Machine detected 8516 errors
- Isolation Forest results in 99.74 % of accuracy and Local outlier Factor has an accuracy of 99.65% whereas SVM givens an accuracy score of 70.09%.
- Fraud detection rate of Isolation Forest is about 27%, for Local Outlier Factor it is 2% and SVM it is 0%.
- Based on comparison between the performance metrics, Isolation Forest out performed both Local Outlier Factor and SVM.
- We can have much better accuracy by employing deep learning algorithms but at higher computational expenses.

V.CONCLUSION

Credit card fraud undoubtedly has to be prevented as effectively as possible. This paper has listed out the different types of frauds and the existing work on the credit card fraud detection model. This paper has also demonstrated in detail the anomaly detection algorithms, their implementation and the result analysis.

Isolation forest gives at most accuracy of 99.75% which certainly out performs the both Local Outlier Factor and SVM whose accuracies are 99.65% and 70.1% and proves to be the best for fraud detection model. Since it is machine learning based model, its efficiency is increased when more amount of data is given.

REFERENCES

[1] Makki, S., Assaghir, Z., Taher, Y., Haque, R., Hacid, M. S., and Zeineddine, H. (2019) An experimental study with imbalanced classification approaches for credit card fraud detection. IEEE Access, 7, pp. 93010-93022.

[2] Randhawa, K., Loo, C. K., Seera, M., Lim, C. P., and Nandi, A. K. (2018) Credit card fraud detection using AdaBoost and majority voting. IEEE Access, 6, pp. 14277-14284.

[3] Taha, A. A., & Malebary, S. J. (2020) An intelligent approach to credit card fraud detection using an optimized light gradient boosting machine. IEEE Access, 8, pp. 25579-25587.

[4] Vardhani, P. R., Priyadarshini, Y. I., & Narasimhulu, Y. (2019) CNN data mining algorithm for detecting credit card fraud. In Soft Computing and Medical Bioinformatics, pp. 85-9. Springer, Singapore.

[5] Varmedja, D., Karanovic, M., Sladojevic, S., Arsenovic, M., & Anderla, A. (2019, March) Credit card fraud detection-machine learning methods. (INFOTECH), pp. 1-5. IEEE.

[6] Singh, A., & Jain, A. (2019) Adaptive credit card fraud detection techniques based on feature selection method. In Advances in computer communication and computational sciences, pp. 167-178). Springer, Singapore.

INNO SPACE
SJIF Scientific Journal Impact Factor
Impact Factor: 8.165

doi cross ref

INTERNATIONAL STANDARD SERIAL NUMBER INDIA

निस्केयर
NISCAIR

INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN COMPUTER & COMMUNICATION ENGINEERING

9940 572 462 6381 907 438 ijircce@gmail.com

www.ijircce.com

Scan to save the contact details