|| Volume 5 || Issue 12 || April 2021 || ISSN (Online) 2456-0774

INTERNATIONAL JOURNAL OF ADVANCE SCIENTIFIC RESEARCH & ENGINEERING TRENDS

Multidisciplinary Journal

Double-Blind Peer Reviewed Refereed Open Access International Journal

CREDIT CARD FRAUD DETECTION USING MACHINE LEARNING

Aditya Raj Singh[1], Hrishikesh Morade[2], Sanket Sable[3], Sunil Hule[4] Guide:- B.R. Patle[5]

Department of Computer Science, SKN Sinhgad Institute of Technology and Science, Lonavala.[1,2,3,4,5]

---***---

Abstract: It's vital that mastercard companies are ready to identify fraudulent credit card transactions so that customers are not charged for items that they didn't purchase. Such problems are often tackled with Data Science and its importance, along with Machine Learning, cannot be overstated. This project intends for instance the modelling of a knowledge set using machine learning with mastercard Fraud Detection. The Credit Card Fraud Detection Problem includes modelling past credit card transactions with the info of those that clothed to be fraud. This model is then wont to recognize whether a replacement transaction is fraudulent or not. Our objective here is to detect 100% of the fraudulent transactions while minimizing the incorrect fraud classifications. Credit Card Fraud Detection is a typical sample of classification. In this process, we have focused on analysing and pre-processing data sets also because the deployment of multiple anomaly detection algorithms such as Local Outlier Factor and Isolation Forest algorithm on the Principle Component Analysis transformed Credit Card Transaction data.

Keywords— *Credit card fraud, applications of machine learning, data science, isolation forest algorithm, local outlier factor, automated fraud detection.*

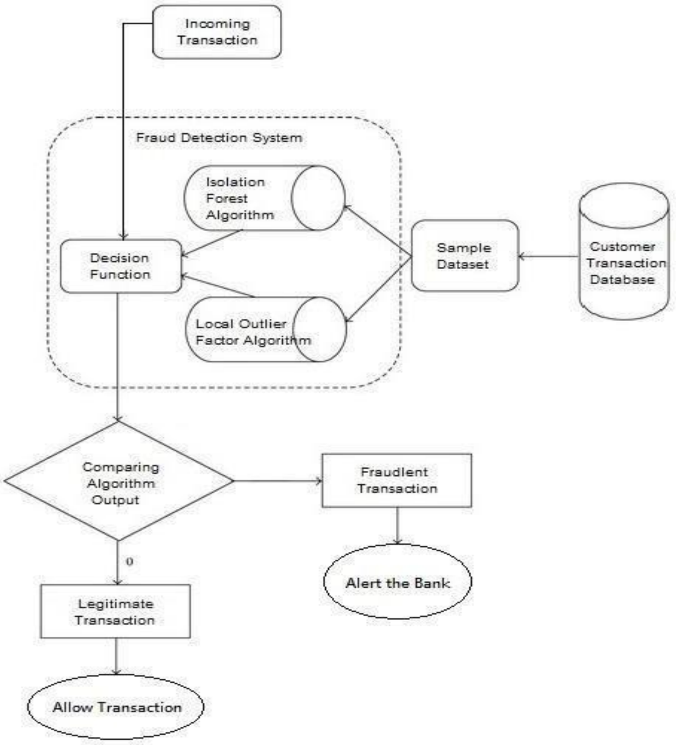
--***--

I INTRODUCTION

'Fraud' in credit card transactions is unauthorized and unwanted usage of an account by someone aside from the owner of that account. Necessary prevention measures can be taken to prevent this abuse and therefore the behaviour of such fraudulent practices are often studied to attenuate it and protect against similar occurrences within the future.In other words Credit Card Fraud are often defined as a case where an individual uses someone else’s mastercard for private reasons while the owner and therefore the card issuing authorities are unaware of the very fact that the cardboard is being used. Fraud detection involves monitoring the activities of populations of users so as to estimate, perceive or avoid objectionable behaviour, which consist of fraud, intrusion, and defaulting.

This is a really relevant problem that demands the eye of communities like machine learning and data science where the solution to this problem can be automated. This problem is particularly challenging from the perspective of learning, because it is characterized by various factors like class imbalance. The number of valid transactions far outnumber fraudulent ones. Also, the transaction patterns usually change their statistical properties over the course of time.

These aren't the sole challenges within the implementation of a real-world fraud detection system, however. In real world examples, the huge stream of payment requests is quickly scanned by automatic tools that determine which transactions to authorize. Machine learning algorithms are employed to review all the authorized transactions and report the suspicious ones. These reports are investigated by professionals who contact the cardholders to verify if the transaction was genuine or fraudulent. The investigators provide a response to the automated system which is employed to coach and update the algorithm to eventually improve the fraud-detection performance over time.



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

●Credit Card Frauds: Online and Offline

●Card Theft

●Account Bankruptcy

●Device Intrusion

●Application Fraud

Some of the currently used approaches to detection of such fraud are:

●Artificial Neural Network

●Fuzzy Logic

●Genetic Algorithm

●Logistic Regression

●Decision tree

●Support Vector Machines

●Bayesian Networks

●Hidden Markov Model

●K-Nearest NeighbourCounterfeit Car

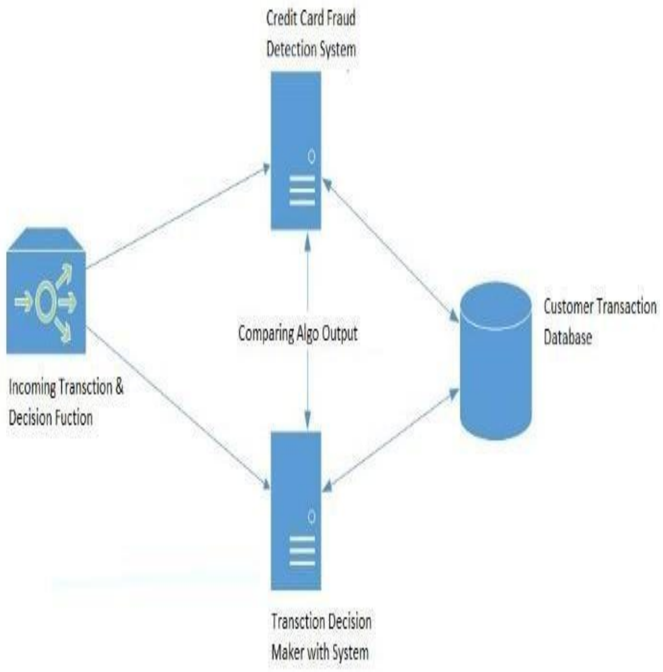
II.LITERATURE REVIEW

Fraud act because the unlawful or criminal deception intended toresult in financial or personal benefit. It is a deliberate act thatis against the rule or policy with an aim to achieveunauthorized financial benefit. Numerous literatures pertaining to anomaly or fraud detection in this domain are published already and are available for public usage. A comprehensive survey conducted by Clifton Phua and his associates have disclosed that techniques employed in this domain include data mining applications, automated fraud detection, adversarial detection. In another paper, Suman, Research Scholar, GJUS&T at Hisar HCE presented techniques like Supervised and Unsupervised Learning for credit card fraud detection. Even though these methods and algorithms fetched an unexpected success in some areas, they didn't provide a permanent and consistent solution to fraud detection. A alike research domain was presented by Wen-Fang YU and Na Wang where they used Outlier detection, Outlier detection mining and Distance sum algorithms to accurately predict fraudulent transaction in an emulation experiment of credit card transaction data set of 1 certain commercial bank. Outlier mining may be a field of knowledge mining which is basically used in monetary and internet fields. It deals with detecting objects that are detached from the most system i.e. the transactions that aren't genuine. They have taken attributes of customer's behaviour and supported the worth of these attributes they've calculated that distance between the observed value of that attribute and its predetermined value. Unconventional techniques such as hybrid data mining/complex network classification algorithm is able to recognize illegal instances in an actual card transaction data set, based on network reconstruction algorithm that allows creating representations of the deviation of 1 instance from a reference group have showed efficient typically on medium sized online transaction. There have also been efforts to progress from a totally new aspect. Attempts have been made to improve the alert feedback interaction in case of fraudulent transaction. In case of fraud transaction, the authorised system would be alerted and a feedback would be sent to deny the continued transaction. Artificial Genetic Algorithm, one among the approaches that shed new light during this domain, countered fraud from a special direction. It proved accurate find out the fraudulent transactions and minimizing the number of false alerts. Even though, it was accompanied by classification problem with variable misclassification costs.

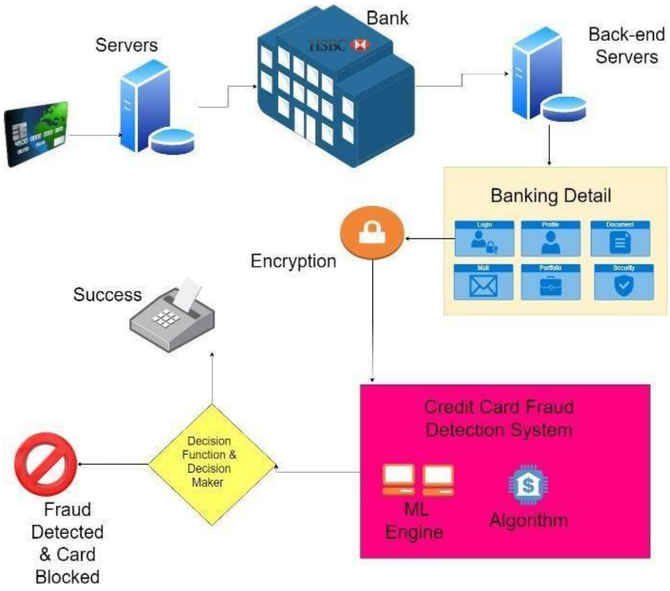
III.METHODOLOGY

The approach that this paper proposes, uses the newest machine learning algorithms to detect anomalous activities, called outliers.

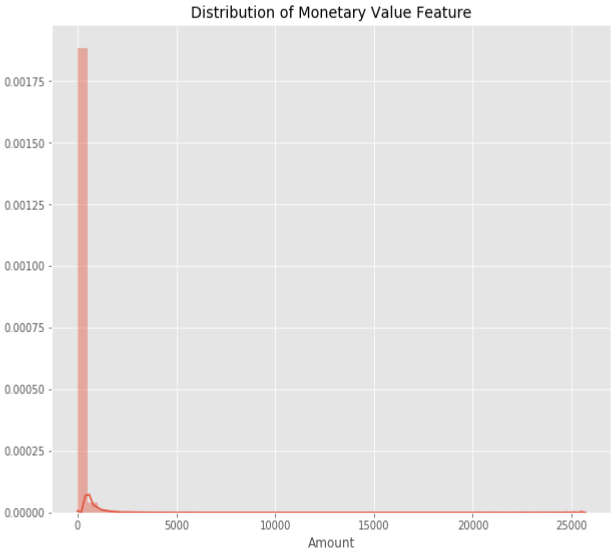
The basic rough architecture diagram are often represented with the following figure:

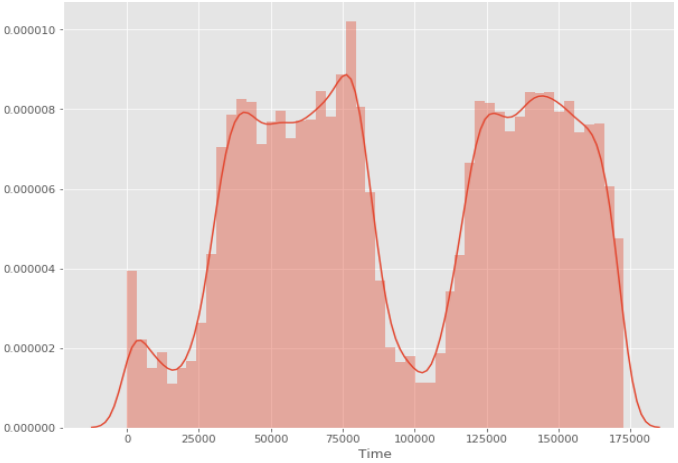


When checked out intimately on a bigger scale alongside real world elements, the complete architecture diagram are often represented as follows:



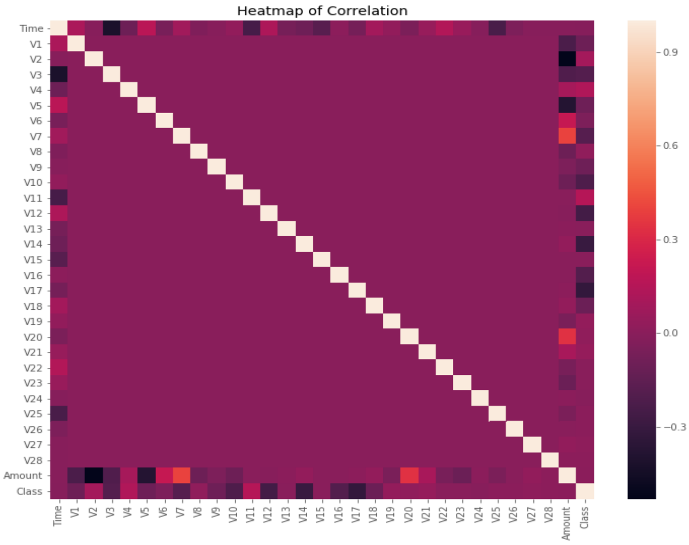
Firstly, we obtained our dataset from Kaggle, a data analysis website which provides datasets. Inside this dataset, there are 31 columns in which 28 are named as v1-v28 to protect sensitive data. The other columns represent Time, Amount and sophistication . Time shows the time gap between the primary transaction and therefore the following one. Amount is the amount of money transacted. Class 0 represents a legitimate transaction and 1 represents a fraudulent one. We plot different graphs to see for inconsistencies within the dataset and to visually comprehend it.





This graph shows the days at which transactions were done within two days. It can be seen that the least number of transactions were made during night and highest during the days.

After checking this dataset, we plot a histogram for each column. this is often done to urge a graphical representation of the dataset which may be wont to verify that there are not any missing any values in the dataset. This is done to ensure that we don’t require any missing value imputation and the machine learning algorithms can process the dataset smoothly. After this analysis, we plot a heatmap to urge a coloured representation of the info and to review the correlation between out predicting variables and the class variable. This heatmap is shown below:



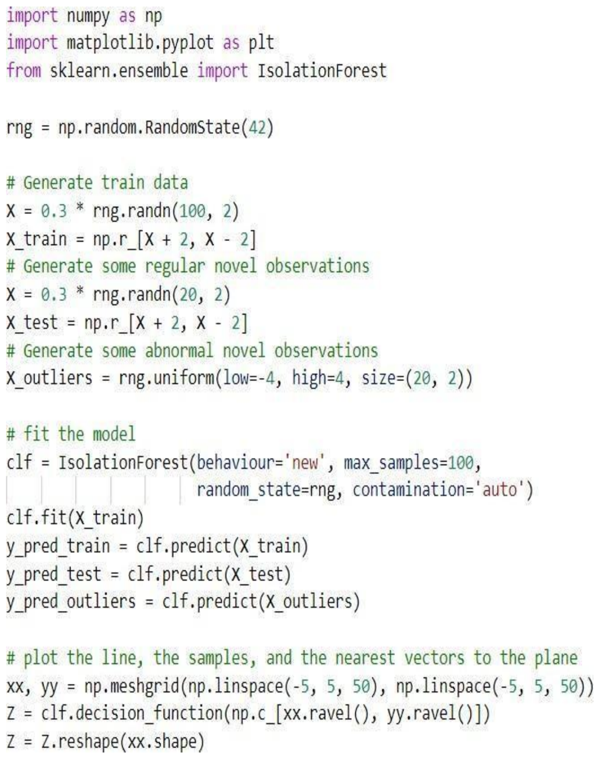
The dataset is now formatted and processed. The time andamount column are standardized and therefore the Class column isremoved to ensure fairness of evaluation. The data is processed by a set of algorithms from modules. The following module diagram explains how these algorithms work together: This data is fit into a model and then the following outlier detection modules are applied on it:

- Local Outlier Factor
- Isolation Forest Algorithm

These algorithms are a part of sklearn. The ensemble module in the sklearn package includes ensemble-based methods and functions for the classification, regression and outlier detection. This free and open-source Python library is made using NumPy, SciPy and matplotlib modules which provides tons of simple and efficient tools which may be used for data analysis and machine learning. It features various classification, clustering and regression algorithms and is designed to interoperate with the numerical and scientific libraries. We've used Jupyter Notebook platform to form a program in Python demonstrate the approach that this paper suggests. This program can also executed on the cloud using Google Collab platform which supports all python notebook files. Detailed explanations about the modules with pseudocodes for their algorithms and output graphs are given as belows:

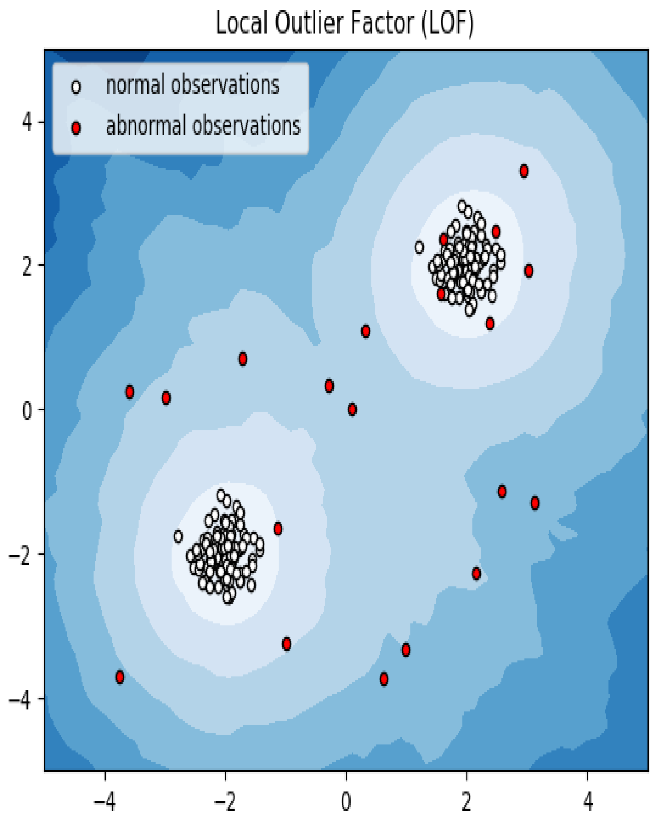
A. Local Outlier Factor

It is an Unsupervised Outlier Detection algorithm. 'Local Outlier Factor' refers to the anomaly score of every sample. It measures local deviation of the sample data with regard to its neighbours. More precisely, locality is given by k-nearest neighbours, whose distance is employed to estimate the local data. The pseudocode for this algorithm is written as:



```
import numpy as np
import matplotlib.pyplot as plt
from sklearn.ensemble import IsolationForest

rng = np.random.RandomState(42)

# Generate train data
X = 0.3 * rng.randn(100, 2)
X_train = np.r_[X + 2, X - 2]
# Generate some regular novel observations
X = 0.3 * rng.randn(20, 2)
X_test = np.r_[X + 2, X - 2]
# Generate some abnormal novel observations
X_outliers = rng.uniform(low=-4, high=4, size=(20, 2))

# fit the model
clf = IsolationForest(behaviour='new', max_samples=100,
                      random_state=rng, contamination='auto')
clf.fit(X_train)
y_pred_train = clf.predict(X_train)
y_pred_test = clf.predict(X_test)
y_pred_outliers = clf.predict(X_outliers)

# plot the line, the samples, and the nearest vectors to the plane
xx, yy = np.meshgrid(np.linspace(-5, 5, 50), np.linspace(-5, 5, 50))
Z = clf.decision_function(np.c_[xx.ravel(), yy.ravel()])
Z = Z.reshape(xx.shape)
```

On plotting the results of Local Outlier Factor algorithm, weget the following figure:



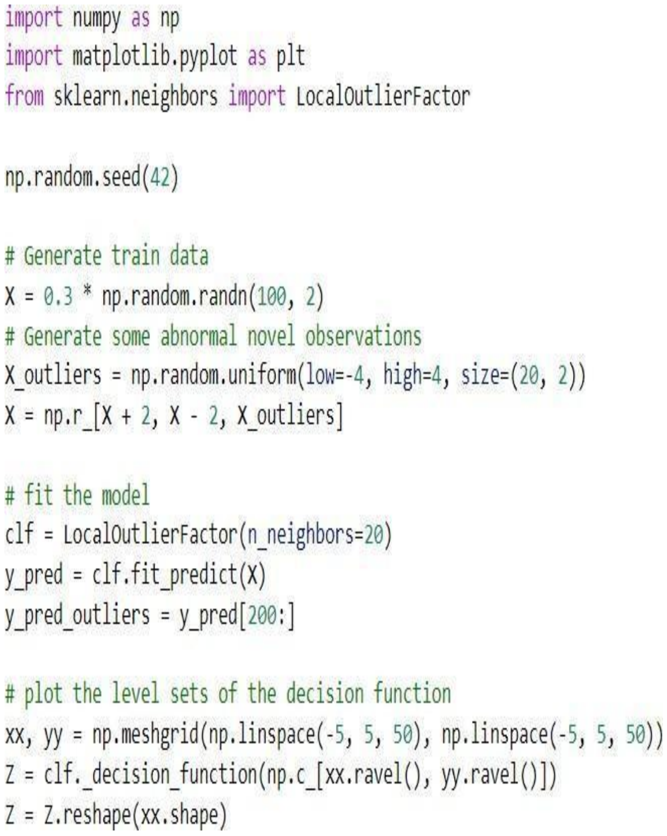
By comparing the local values of a sample to that of its neighbours, one can identify samples that are substantially lower than their neighbours. These values are quite amanous and they are considered as outliers. As the dataset is very large, we used only a fraction of it in out tests to reduce processing times. The final result with the entire dataset processed is additionally determined and is given within the results section of this paper.

B. Isolation Forest Algorithm

The Isolation Forest 'isolates' observations by arbitrarily selecting a feature then randomly selecting a split value between the utmost and minimum values of the designated feature. Recursive partitioning are often represented by a tree, the number of splits required to isolate a sample is like the path length root node to terminating node. The average of this path length gives a measure of standardness and the decision function which we use.

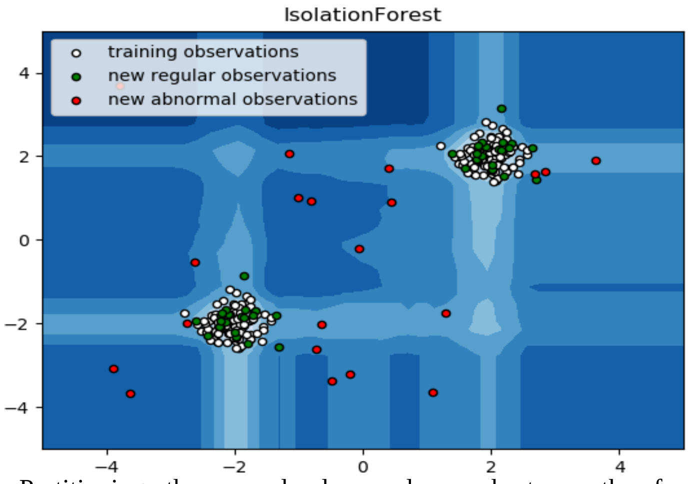
The pseudocode for this algorithm are often written as:

On plotting the results of an Isolation Forest algorithm, we get the following figure:



```
import numpy as np
import matplotlib.pyplot as plt
from sklearn.neighbors import LocalOutlierFactor

np.random.seed(42)

# Generate train data
X = 0.3 * np.random.randn(100, 2)
# Generate some abnormal novel observations
X_outliers = np.random.uniform(low=-4, high=4, size=(20, 2))
X = np.r_[X + 2, X - 2, X_outliers]

# fit the model
clf = LocalOutlierFactor(n_neighbors=20)
y_pred = clf.fit_predict(X)
y_pred_outliers = y_pred[200:]

# plot the level sets of the decision function
xx, yy = np.meshgrid(np.linspace(-5, 5, 50), np.linspace(-5, 5, 50))
Z = clf._decision_function(np.c_[xx.ravel(), yy.ravel()])
Z = Z.reshape(xx.shape)
```

Plotting the results of Isolation Forest algorithm, we get the following figure:



Partitioning them randomly produces shorter paths for anomalies. When the forest of the random trees mutually produces shorter path lengths for specific samples, they're extremely likely to be anomalies. Once the anomalies are detected, the system are often wont to report them to the concerned authorities. For testing purposes, we are comparing the outputs of those algorithms to work out their accuracy and precision.

IV.IMPLEMENTATION

This idea is difficult to implement in real world because it requires the cooperation from banks, which are not willing to share information thanks to their market competition, and also due to legal reasons and protection of knowledge of their users. Therefore, we looked up some reference papers which followed similar approaches and gathered results. As stated in one of these reference papers:

"This technique was applied to the full application data set supplied by a German bank in 2006. For banking confidentiality reasons, only a summary of the results gained is presented below. After applying this technique, the level 1 list encompasses a couple of cases but with a high probability of being fraudsters.

All individuals mentioned during this list had their cards closed to avoid any risk due to their high-risk profile. The condition is more complex for the other list. The level 2 list is still restricted sufficiently to be checked on a case by case basis. Credit and collection officers considered that half the cases in this list might be considered as suspicious fraudulent behaviour. For the last list and therefore the huge, the work is equitably heavy. Less than a third of them are suspicious. In order to maximise the time efficiency and therefore the overhead charges, an opportunity is to incorporate a replacement element within the query; this element are often the 5 first digits of the phone numbers, the email address, and therefore the password, as an example , those new queries are often applied to the extent 2 list and level 3 list.".

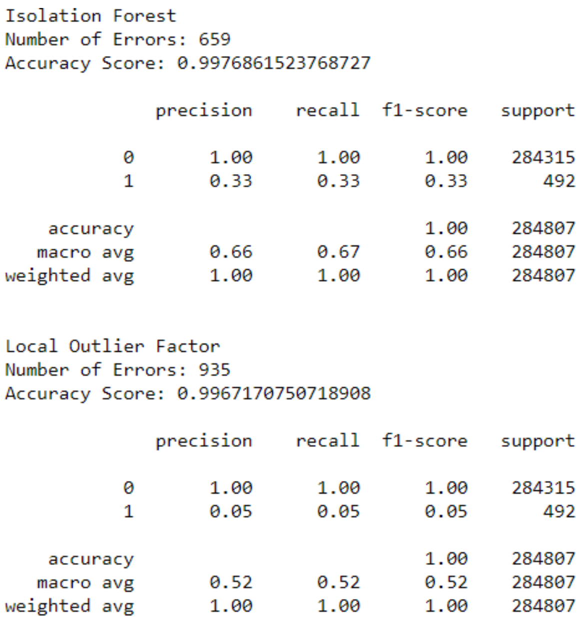
V.RESULTS

The code prints out the amount of false positives it detected and compares it with the actual values. This is used to calculate the accuracy score and precision of the algorithms. The fraction of knowledge we used for faster testing is 10% of the entire dataset. The complete dataset is additionally used at the top and both the results are printed.

These results along side the classification report for every algorithm is given within the output as follows, where class 0 means the transaction decided to be valid and 1 means it was determined as a fraud transaction.

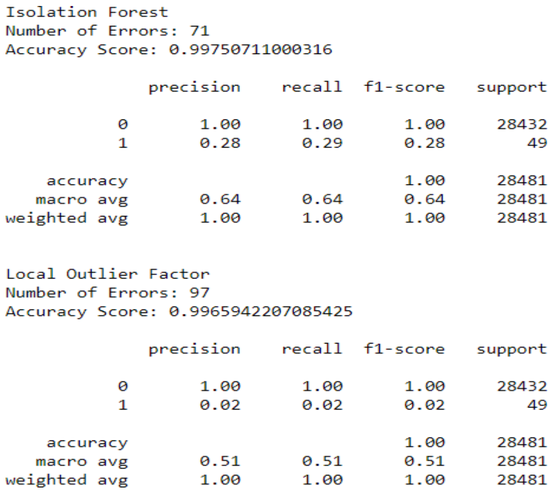
This result matched against the category values to see for false positives.

Results when 10% of the dataset is used:



```
Isolation Forest
Number of Errors: 659
Accuracy Score: 0.9976861523768727

              precision    recall  f1-score   support

           0       1.00      1.00      1.00    284315
           1       0.33      0.33      0.33       492

    accuracy                           1.00    284807
   macro avg       0.66      0.67      0.66    284807
weighted avg       1.00      1.00      1.00    284807

Local Outlier Factor
Number of Errors: 935
Accuracy Score: 0.9967170750718908

              precision    recall  f1-score   support

           0       1.00      1.00      1.00    284315
           1       0.05      0.05      0.05       492

    accuracy                           1.00    284807
   macro avg       0.52      0.52      0.52    284807
weighted avg       1.00      1.00      1.00    284807
```

Results with the complete dataset is used:



```
Isolation Forest
Number of Errors: 71
Accuracy Score: 0.99750711000316

              precision    recall  f1-score   support

           0       1.00      1.00      1.00     28432
           1       0.28      0.29      0.28        49

    accuracy                           1.00     28481
   macro avg       0.64      0.64      0.64     28481
weighted avg       1.00      1.00      1.00     28481

Local Outlier Factor
Number of Errors: 97
Accuracy Score: 0.9965942207085425

              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
```

VI.CONCLUSION

Credit card fraud is without a doubt a move of criminal dishonesty. this text has listed out the foremost common methods of fraud along side their detection methods and reviewed recent findings in this field. This paper has also explained intimately , how machine learning are often applied to get better leads to fraud detection along side the algorithm, pseudocode, explanation its implementation and experimentation results. While the algorithm does reach over 99.6 percentage accuracy, its precision remains only at 28% when a tenth of the info set is taken into consideration.

However, when the entire dataset is fed into the algorithm, the precision rises to 33 percentage. This high percentage of accuracy is to be expected thanks to the large imbalance between the amount of valid and number of genuine transactions.Since the whole dataset consists of only two days' transaction records, its only a fraction of knowledge which will be made available if this project were to be used on the billboard scale. Being based on machine learning algorithms, program will only increase its efficiency over time the more data is put into it.

VII.FUTURE ENHANCEMENTS

While we couldn't reach out goal of 100 percentage accuracy in fraud detection, we did find yourself creating a system which will , with enough time and data, get very on the brink of that goal. As with any such project, there's some room for improvement here. The very nature of the project allows for multiple algorithms to be integrated together as a modules and their results are often combined to extend the accuracy of the ultimate result. This model can be further improved with the addition of more algorithms into it. However, the output of these algorithms needs to be within the same format because the others. Once that condition is satisfied, the modules are easy to feature as wiped out the code. This provides a great degree of modularity and versatility to the project. More room for improvement are often found within the dataset. As demonstrated before, the precision of the algorithms increases when the size of dataset is increased. Hence, more data will surely make the model more accurate in detecting the frauds and reduce the number of false positives. However, this requires official support from the banks themselves.

VIII.REFERENCES

[1]"Credit Card Fraud Detection supported Transaction Behaviour - published by Proc. of the 2017 IEEE

[2]CLIFTON PHUA1, VINCENT LEE1, KATE SMITH1 and ROSS GAYLER2 " A Survey of knowledge Mining-based Fraud Detection Research" - by the School of Business, Faculty of Information Technology, Monash University, Wellington Road, Clayton, Victoria 3800,

Australia

[3]"Survey Paper on the Credit Card Fraud Detection by Suman" , Scholar, GJUS&T Hisar HCE, Sonepat published by International Journal of the Advanced Research in Computer Engineering & Technology (IJARCET) Volume 3 Issue 3, March 2014

[4]"Research on Credit card Fraud Detection Model supported Distance Sum – by Wen-Fang YU and Na Wang" published on 2009 International Joint Conference on Artificial Intelligence

[5]"Credit Card Fraud Detection through the Parenclitic Network Analysis - Massimiliano Zanin, Miguel Romance, Regino Criado, and SantiagoMoral" published by Hindawi

Complexity Volume 2018, Article ID 5764370, 9 pages

[6]“Credit Card Fraud Detection:- a sensible Modeling and a completely unique Learning Strategy” published by IEEE TRANSACTIONS ON NEURAL NETWORKS AND LEARNING SYSTEMS, VOL. 29, NO. 8, AUGUST 2018

[7]“Credit Card Fraud Detection- Ishu Trivedi, Monika, Mrigya, Mridushi” published by the International Journal of Research in Computer and Communication Engineering Vol. 5, Issue 1, January 2016

[8]David J.Wetson,David J.Hand,M Adams,Whitrow and Piotr Jusczak “Plastic Card Fraud Detection using coevals Analysis” Springer, Issue 2008.