iJRASET
International Journal For Research in Applied Science and Engineering Technology

# INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

**Volume:** 9 **Issue:** VII **Month of publication:** July 2021

**DOI:** https://doi.org/10.22214/ijraset.2021.37200

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: 7.429*
*Volume 9 Issue VII July 2021- Available at www.ijraset.com*

# Credit Card Fraud Detection using Machine Learning and Data Science

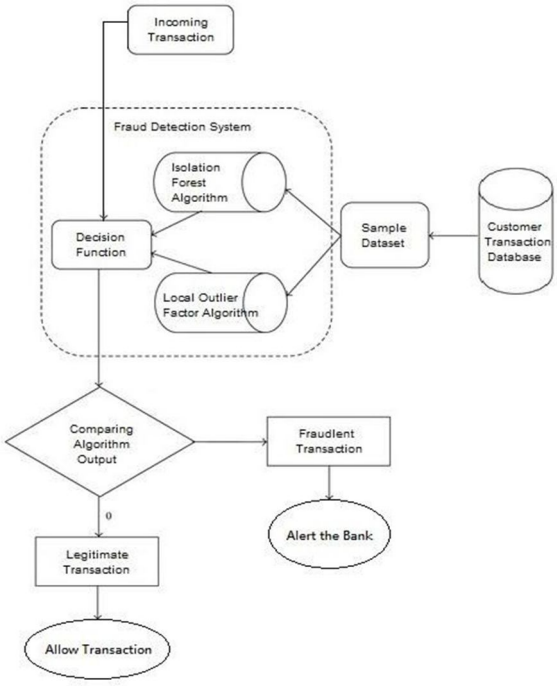
Aman[1], Arpit Mishra[2], Ashish Kumar[3], Naveen Pandey [4]
[1, 2, 3, 4]*Inderprastha Engineering College*

***Abstract:*** ***It is important that companies are able to identify fraudulent credit card transactions so that customers are not charged for items that they did not purchase. These problems can be handled with Data Science and its importance, along with Machine Learning. This project aim is to illustrate the modelling of a data set using machine learning with Credit Card. Our objective is to detect 100% of the fraudulent transactions while minimizing the incorrect fraud classifications. Credit Card Fraud Detection is a sample of classification. In this process, we have focused on analysing and pre-processing data sets as well as the deployment of multiple anomaly detection algorithms such as Local Outlier Factor and Isolation Forest algorithm on the PCA 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 illegal and unwanted usage of an account by someone other than the owner of that account. Various measures can be taken to stop this and protect against similar occurrences in the future.So we can say that Credit Card Fraud can be defined as a case where a person uses someone else's credit card for personal reasons while the owner and the card issuing authorities are unaware of the fact that the card isbeing used. Fraud detection involves monitoring the activities of populations of users in order to estimate, perceive or avoid objectionable behaviour, which consist of fraud, intrusion, anddefaulting. This is a very relevant problem that demands the attention of communities such as machine learning and data science where the solution to this problem can be automated. The number of valid transactions far outnumber fraudulent ones. Also, the transaction patterns often change their statistical properties over the course of time. Machine learning algorithms are used to analyse all the transactions and report the suspicious ones. These reports are investigated by professionals who contact the cardholders to confirm if the transaction was genuine or fraudulent. The investigators provide a feedback to the automated systemwhich is used to train 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 frauds are classified as:

*1)* Credit Card Frauds: Online and Offline
*a)* Card Theft
*b)* Device Intrusion
*c)* Application Fraud
*d)* Telecommunication Fraud

*2)* Some of the currently used approaches to detection ofsuchfraud are
*a)* Logistic Regression
*b)* Decision tree
*c)* Support Vector Machines
*d)* Bayesian Networks K-Nearest Neighbour

## II. LITERATURE REVIEW

Fraud act as the illegal or criminal work intended to result in financial or personal benefit. It is a knowingly act that isagainst the law, rule or policy with an aim to attain unauthorized financial benefit.

Numerous literatures pertaining to anomaly or fraud detection in this domain have been published already and are available for public usage. A comprehensive survey conducted by Clifton Phua and his associates have revealed 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 failed to provide a permanent and consistent solution to fraud detection.

A similar research domain was presented by Wen-Fang YU and Na Wang where they used Outlier mining, Outlier detection mining and Distance sum algorithms to accurately predict fraudulent transaction in an emulation experiment of credit card transaction data set of one certain commercial bank. Outlier mining is a field of data mining which is basically used in monetary and internet fields. It deals with detecting objects that are detached from the main system i.e. the transactions that aren't genuine. They have taken attributesof customer's behaviour and based on the value of thoseattributes they've calculated that distance between the observed value of that attribute and its predetermined value.

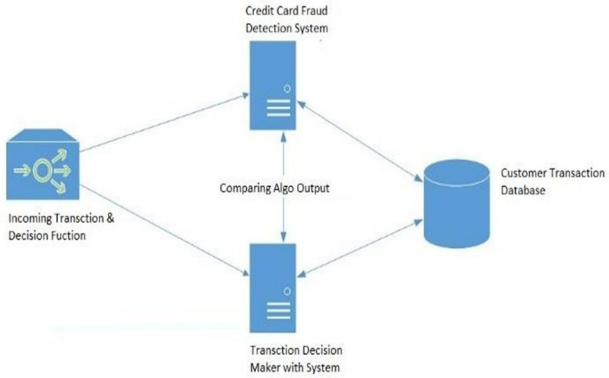
There have also been efforts to progress from a completely new aspect. Attempts have been made to improve the alert- feedback interaction in case of fraudulent transaction.

In case of fraudulent transaction, the authorised system would be alerted and a feedback would be sent to deny the ongoing transaction.

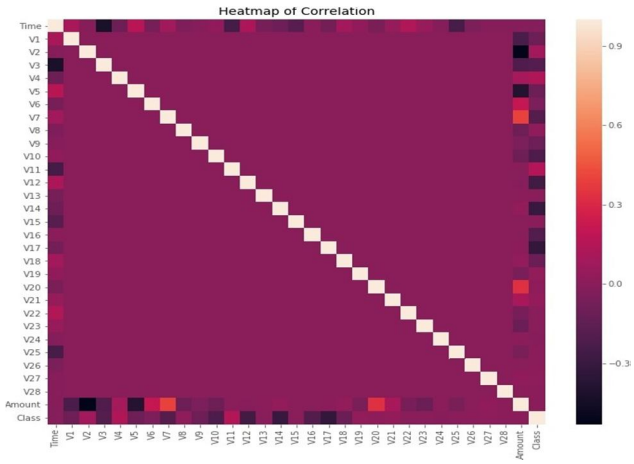
## III. METHODOLOGY

In this paper we uses the latest machinelearning algorithmsto detect anomalous activities, called outliers.

The basic rough architecture diagram can be representedwiththe following figure:



First, we obtained the dataset from Kaggle, which is a data analysis website which provide data sets.
Inside this dataset, there are 31 columns out of which 28 are named as v1-v28 to protect sensitive data.
The other columns represent Time, Amount and Class. Time shows the time gap between the first transaction and the following one. Amount is the amount of money transacted. Class 0 represents a valid transaction and 1 represents a fraudulent one.
After this, we plot a heatmap to get a coloured representation of the data and to study the correlation betweenout predicting variables and the class variable.
This heatmap isshown below:




Now the dataset is formatted and processed. The time and amount column are standardized and the Class column isremoved to ensure fairness of evaluation. The following module diagram explains how these algorithms work together. This data is fit into a model and the following outlier detectionmodules are applied on it:

*1)* Local Outlier Factor
*2)* Isolation Forest Algorithm

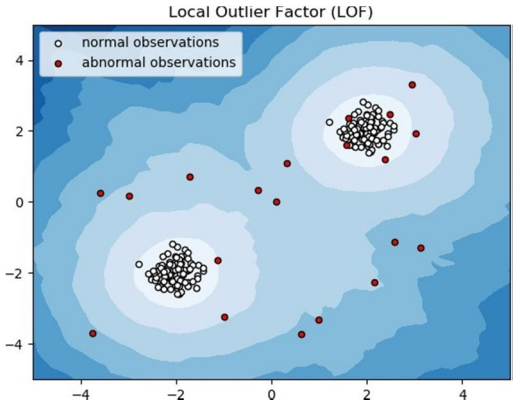
These algorithms are a part of sklearn. The ensemble module in the sklearn package includes the combination of two or more algorithm in model. The free and open-source Python library is built using NumPy, SciPy and matplotlib modules that provides a simple and important tools for the analysis of datasets.
We've used Jupyter Notebook platform to make a program in Python. This program can also be executed on the cloud using Google Collab platform which supports all python notebook files.

### *A. Local Outlier Factor*

It is defined as it is an Unsupervised Outlier Detection algorithm. 'Local Outlier Factor' refers to the anomaly score of each sample. It measures the local deviation of the sample data with respect toits neighbours.
More accurately, locality is given by k-nearest neighbours, whose distance is used 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, we get the following figure:



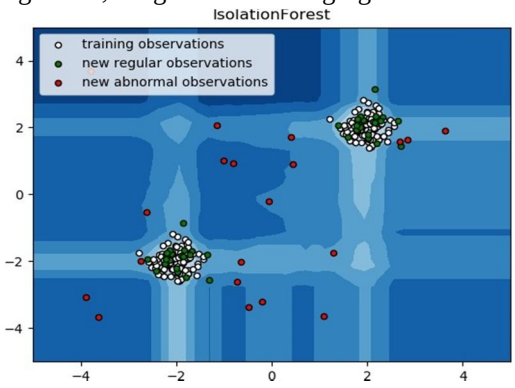

By comparing the local values of a sample to that of its neighbours, we can identify samples that are precisely lower than their neighbours. These values are quite different and they are considered as outliers.
As the dataset is very large, we used only a fraction of it in outtests to reduce processing times.

*B. Isolation Forest Algorithm*

The Isolation Forest 'isolates' observations by randomly selecting a feature and then randomly selecting a split value between the maximum and minimum values of the designatedfeature.
The average of this path length gives a measure of normality and the decision function which we use.
The pseudocode for this algorithm can be written as:

```
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)
```

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




Once the anomalies are detected, we can easily report to concerned authorities. For testing purposes, we are comparing the outputs of these algorithms to determinetheir accuracy and precision.

## IV. RESULTS

The code prints out the number 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 data we used for faster testing is 10% of the entire dataset. The complete dataset is also used at the end andboth the results are printed.

These results along with the classification report for each algorithm is given in the output as follows, where class 0 means the transaction was determined to be valid and 1 meansit was determined as a fraud transaction.

This result matched against the class values to check for falsepositives.

Results when 10% of the 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
```

## V. CONCLUSION

Credit card fraud is undoubtly an act of criminal dishonesty. This article has listed out the most common methods of fraud along with their detection methods and reviewed recent findings in this field. In this paper we have also explained in detail, how machine learning can be applied to get better results in fraud detection along with the algorithm, pseudocode, explanation its implementation and experimentation results.

While the algorithm does reach over 99.6% accuracy, its precision remains only at 28% when a tenth of the data set is taken into consideration. However, when the entire dataset is fed into the algorithm, the precision rises to 33%. This high percentage of accuracy is to be expected due to the huge imbalance between the number of valid and number of genuine transactions.

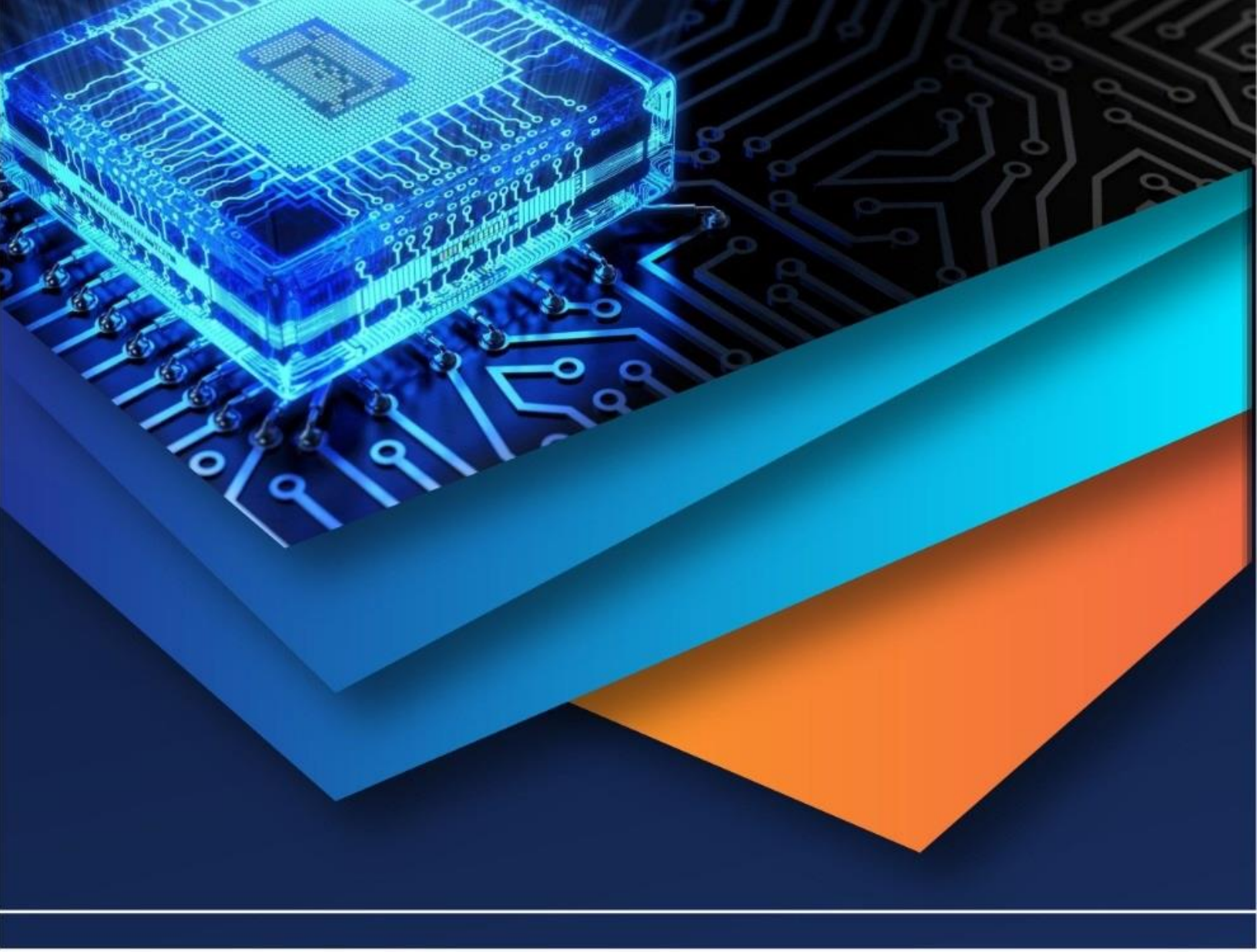
## REFERENCES

[1] “Credit Card Fraud Detection Based on Transaction Behaviour -by John Richard D. Kho, Larry A. Vea” published by Proc. of the 2017 IEEE Region 10 Conference (TENCON), Malaysia, November 5-8, 2017

[2] CLIFTON PHUA1, VINCENT LEE1, KATE SMITH1 & ROSS GAYLER2 “ A Comprehensive Survey of Data Mining-based Fraud Detection Research” published by School of Business Systems, Faculty of Information Technology, Monash University, Wellington Road, Clayton, Victoria 3800, Australia

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

[4] Research on Credit Card Fraud Detection Model Based on Distance Sum – by Wen-Fang YU and Na Wang” published by 2009International Joint Conference on Artificial Intelligence2014.



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

# INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY