CREDIT CARD FRAUD DETECTION

A PROJECT REPORT

Submitted by

G.NAGA SAI PRANEETH -19MIM10044
MADA SAI SURYA -19MIM10095
KADIYALA MEGHANATH -19MIM10097
ABHISHEK SWAMI -19MIM10117

in partial fulfillment for the award of the degree of

BACHELOR OF TECHNOLOGY

in

COMPUTER SCIENCE AND ENGINEERING

Specialization in

Artificial intelligence and machine learning

SCHOOL OF COMPUTING SCIENCE AND ENGINEERING

VIT BHOPAL UNIVERSITY

KOTHRI KALAN, SEHORE
MADHYA PRADESH - 466114

NOV 2020

BONAFIDE CERTIFICATE

Certified that this project report titled **“CREDIT CARD FRAUD DETECTION”** is the bonafide work of “ **K.MEGHANATH(19MIM10097) ;M.SAI SURYA(19MIM10095);ABHISHEK SWAMI(19MIM10117);G.NAGA SAI PRANEETH(19MIM10044)”** who carried out the project work under my supervision. Certified further that to the best of my knowledge the work reported here does not form part of any other project / research work on the basis of which a degree or award was conferred on an earlier occasion on this or any other candidate.

PROGRAM CHAIR	**PROJECT GUIDE**
Dr.PANDIMURAGAN	PROF.ASHISH KUMAR SAHU
School of AI & ML division	School of AI & ML division
VIT BHOPAL UNIVERSITY	VIT BHOPAL UNIVERSITY

The Project Exhibition I Examination is held on 3-9-2020.

ACKNOWLEDGEMENT

First and foremost I would like to thank the Lord Almighty for His presence and immense blessings throughout the project work.

I wish to express my heartfelt gratitude to **Dr.V.Pandimurugan**, Head of the Department, School of Aeronautical Science for much of his valuable support and encouragement in carrying out this work.

I would like to thank my internal guide **Mr.Ashish kumar sahu**,for continually guiding and actively participating in my project, giving valuable suggestions to complete the project work.

I would like to thank all the technical and teaching staff of the School of Aeronautical Science, who extended directly or indirectly all support.

Last, but not the least, I am deeply indebted to my parents who have been the greatest support while I worked day and night for the project to make it a success.

LIST OF FIGURES

FIGURE NO.	TITLE	PAGE NO.
1.	**LOCAL OUTLIER**	12
2.	**ISOLATION FOREST**	13
3.	**SYSTEM ARCHITECTURE**	15

ABSTRACT

It is vital that credit card companies are able to identify fraudulent credit card transactions so that customers are not charged for items that they did not purchase. Such problems can be tackled with Data Science and its importance, along with Machine Learning, cannot be overstated. This project intends to illustrate the modelling of a data set using machine learning with Credit Card Fraud Detection. The Credit Card Fraud Detection Problem includes modelling past credit card transactions with the data of the ones that turned out to be fraud. This model is then used to recognize whether a new 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 preprocessing 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.

TABLE OF CONTENTS

CHAPTER NO.	TITLE	PAGE NO.
	List of Abbreviations	iii
	List of Figures	iv
	List of Tables	v
	Abstract	vi
1	**INTRODUCTION**	
	1.1 Introduction	8
	1.2 Motivation for the work	8
	1.3 Problem Statement	9
2	**LITERATURE SURVEY**	
	2.1 Introduction	10
	2.2 Existing Algorithms	10
	2.3.1 Local Outlier	10
	2.3.2 Isolation Forest	11
	2.3 Research issues/observations from literature Survey	12
3	**SYSTEM ANALYSIS**	
	3.1 Introduction	13
	3.2 Existing system	13
	3.3 Proposed System	13
	3.4 Summary	13
4	**SYSTEM DESIGN AND IMPLEMENTATION**	
	4.1 Architectural Design	14
	4.2 System Design & Implementation	15
	4.3 Hardware/Software interface	16
5	**CODING AND TESTING**	
	5.1 Testing the DataSet(Graphs)	17
	5.2 Implementation of Algorithms using Python	18
6	**FUTURE ENHANCEMENT AND CONCLUSION**	
	6.1 Introduction	19
	6.2 Limitation/Constraints of the System	19
	6.3 Future Enhancements	19
	6.4 Conclusion And Result	20

9	References	21

INTRODUCTION

1.1 INTRODUCTION

WHAT ARE FRAUDULENT TRANSACTIONS ?

- Fraudulent transactions are orders and purchases made using a credit card or bank account that does not belong to the buyer.
- One of the largest factors in identity fraud, these types of transactions can end up doing damage to both merchants and the identity fraud victim.
- Avoiding fraudulent transactions is in the interest of both merchants and buyers, so it is important to take proper precautions when managing money accounts.

WHAT IS FRAUD DETECTION ?

- Fraud detection involves monitoring the behaviour of users in order to estimate, detect, or avoid undesirable behaviour. To counter the credit card fraud effectively, it is necessary to understand the technologies involved in detecting credit card frauds and to identify various types of credit card frauds.
- Credit card fraud is a wide-ranging term for theft and fraud committed using a credit card as a fraudulent source of funds in a given transaction. credit card fraudsters employ a large number of techniques to commit fraud. To combat credit card fraud effectively, it is important to first understand the mechanisms of identifying a credit card fraud. over the years credit card fraud has stabilized much due to various credit card fraud detection and prevention mechanisms
- This project is to propose a credit card fraud detection system using genetic algorithms. Genetic algorithms are evolutionary algorithms which aim at obtaining better solutions as time progresses. When a card is copied or stolen or lost and captured by fraudsters it is usually used until its available limit is depleted. Thus, rather than the number of correctly classified transactions, a solution which minimizes the total available limit on cards subject to fraud is more prominent. It aims in minimizing the false alerts using genetic algorithms where a set of interval valued parameters are optimized.

Motivation of work

We found that many innocent people are losing money by fraud transactions done by Fraud People.
So we are taught we can help innocent people by using Machine Learning algorithms.

Problem Statement

To develop a credit card fraud detection system using genetic algorithms. During the credit card transaction, the fraud is detected and the number of false alerts is being minimized by using genetic algorithms. Instead of maximizing the numbers of correctly classified transactions we defined an objective function where the misclassification costs are variable and thus, correct classification of some transactions are more important than correctly classifying the others.

The algorithm begins with multi-population of randomly generated chromosomes. These chromosomes undergo the operations of selection, crossover and mutation. Crossover combines the information from two parent chromosomes to produce new individuals, exploiting the best of the current generation, while mutation or randomly changing some of the parameters allows exploration into other regions of the solution space. Natural selection via a problem specific cost function ensures that only the best fit chromosomes remain in the population to mate and produce the next generation. Upon iteration, the genetic algorithm converges to a global solution.

LITERATURE SURVEY

Introduction

Fraud detection has been usually seen as a data mining problem where the objective is to correctly classify the transactions as legitimate or fraudulent. For classification problems many performance measures are defined, most of which are related to the correct number of cases classified correctly.

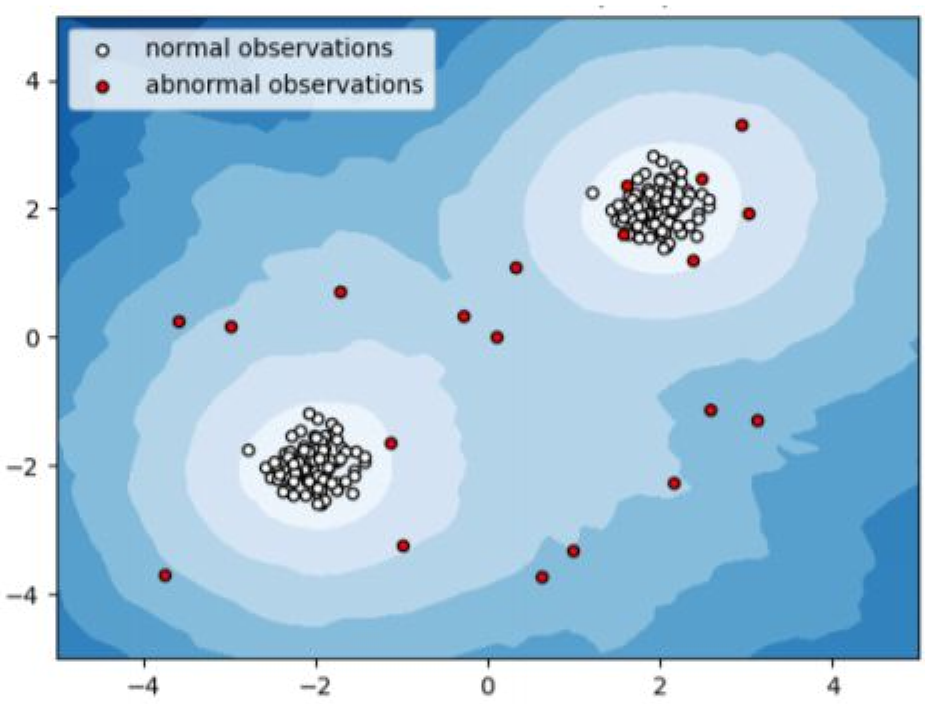
A more appropriate measure is needed due to the inherent structure of credit card transactions. When a card is copied or stolen or lost and captured by fraudsters it is usually used until its available limit is depleted. Thus, rather than the number of correctly classified transactions, a solution which minimizes the total available limit on cards subject to fraud is more prominent.

Since the fraud detection problem has mostly been defined as a classification problem, in addition to some statistical approaches many data mining algorithms have been proposed to solve it. Among these, decision trees and artificial neural networks are the most popular ones. The study of Bolton and Hand provides a good summary of literature on fraud detection problems.

Existing Algorithms

A. Local Outlier Factor

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 to its neighbours. More precisely, locality is given by k-nearest neighbours, whose distance is used to estimate the local data.



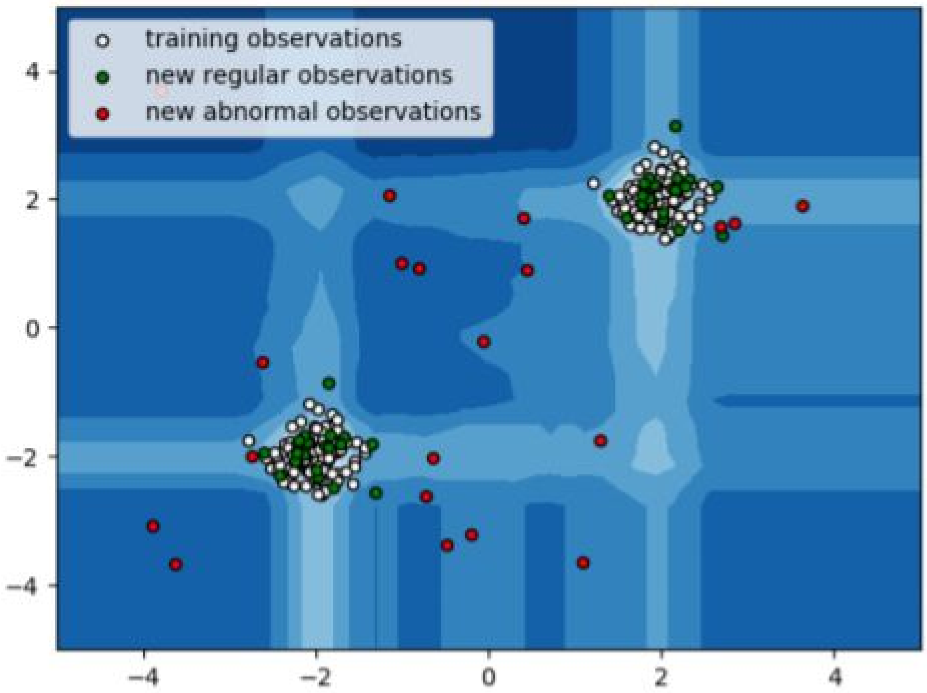
B. Isolation Forest

The IsolationForest ‘isolates’ observations by randomly selecting a feature and then randomly selecting a split value between the maximum and minimum values of the selected feature.

Since recursive partitioning can be represented by a tree structure, the number of splittings required to isolate a sample is equivalent to the path length from the root node to the terminating node.

This path length, averaged over a forest of such random trees, is a measure of normality and our decision function.

Random partitioning produces noticeably shorter paths for anomalies. Hence, when a forest of random trees collectively produce shorter path lengths for particular samples, they are highly likely to be anomalies.



Research issues/observations from literature Survey

Of the many major issues facing the graduate student, a primary one is the identification of a research problem. Problems may arise from real-world settings or be generated from theoretical frameworks. The source of research problems will vary according to the experience of the person contemplating an investigation, but it is generally agreed that the process begins with a question or need.

SYSTEM ANALYSIS

INTRODUCTION

This chapter gives the information regarding analysis done for the proposed system. System Analysis is done to capture the requirement of the user of the proposed system. It also provides the information regarding the existing system and also the need for the proposed system. The key features of the proposed system and the requirement specifications of the proposed system are discussed below.

EXISTING SYSTEM

The Traditional detection method mainly depends on the database system and the education of customers, which usually are delayed, inaccurate and not in-time. After that methods based on discriminant analysis and regression analysis are widely used which can detect fraud by credit rate for cardholders and credit card transactions. For a large amount of data it is not efficient.

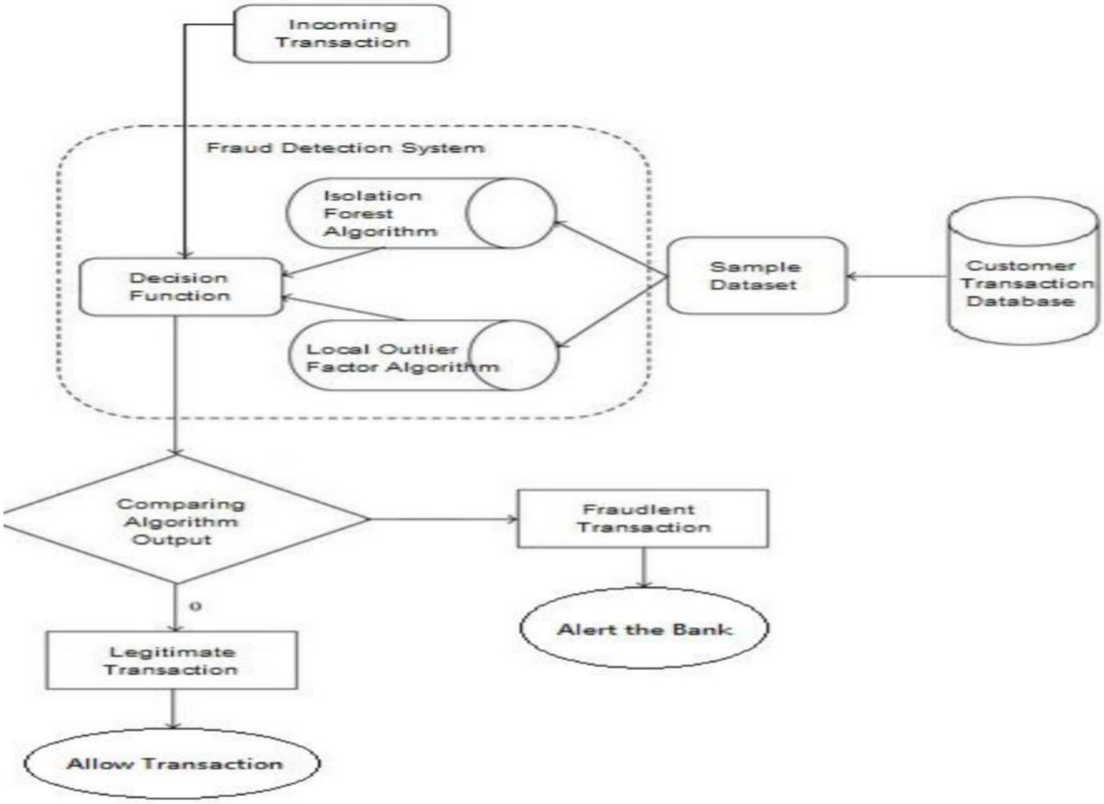
PROPOSED SYSTEM

The proposed system overcomes the above mentioned issue in an efficient way. Using outlier algorithms the fraud is detected and the false alert is minimized and it produces an optimized result. The fraud is detected based on the customers behavior. A new classification problem which has a variable misclassification cost is introduced. Here the genetic algorithm is made where a set of interval valued parameters are optimized

SYSTEM DESIGN AND IMPLEMENTATION

ARCHITECTURAL DESIGN:

Describing the overall features of the software is concerned with defining the requirements and establishing the high level of the system. During architectural design, the various web pages and their interconnections are identified and designed. The major software components are identified and decomposed into processing modules and conceptual data structures and the interconnections among the modules are identified. The following modules are identified in the proposed system.



The above architecture describes the work structure of the system.

The customer data in the data warehouse is subjected to the rules engine which consists of the fraud rule set. The filter and priority module sets the priority for the data and then sends it to the genetic algorithm which performs its functions and generates the output

DETAILED SYSTEM DESIGN :

Detailed design deals with the various modules in detail explaining them with appropriate Diagrams and notations. The Use case diagram is designed to see the working logic of the proposed system. The sequence diagram is designed to describe how the client and the server interact with each other when processing a content. The flow of the proposed system is described with the activity diagram. We know where the application starts and when it ends after processing the keywords and the current URL link. This will help the programmers to implement the internal logic for the module in the given specification.

In this part of the design phase, the design is carried out using the top-down strategy. First the major modules are identified. Then they are divided into sub modules so that each module at the lowest level would address a single function of the whole system. Each module design is explained in detail. This chapter tells us how the input module is designed in getting the users requirements. The detailed input design provides information regarding what tools are used in getting inputs and sent to the server.

Output design gives the user a good interacting option on the screen. The information delivered to the users through the information system. Useful output is essential to ensure the use and acceptance of the information system. Users often judge the merit of a system based upon its output. Productive output can only be achieved via close interaction with users. The output is designed in an attractive and effective way that users can access them with a problem.

IMPLEMENTATION

This idea is difficult to implement in real life because it requires the cooperation from banks, which aren’t willing to share information due to their market competition, and also due to legal reasons and protection of data 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 a full application data set supplied by a German bank in 2006. For banking confidentiality reasons, only a summary of the results obtained is presented below. After applying this technique, the level 1 list encompasses a few cases but with a high probability of being fraudsters. All individuals mentioned in 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 adequately to be checked on a case by case basis. Credit and collection officers considered that half of the cases in this list could be considered as suspicious fraudulent behaviour. For the last list and the largest, the work is equitably heavy. Less than a third of them are suspicious. In order to maximize the time efficiency and the overhead charges, a possibility is to include a new element in the query; this element can be the five first digits of the phone numbers, the email address, and the password, for instance, those new queries can be applied to the level 2 list and level 3 list.".

HARDWARE AND SOFTWARE REQUIREMENTS:

Hardware:

- Processor - intel
- Ram - 4GB
- Hard Disk - 260GB

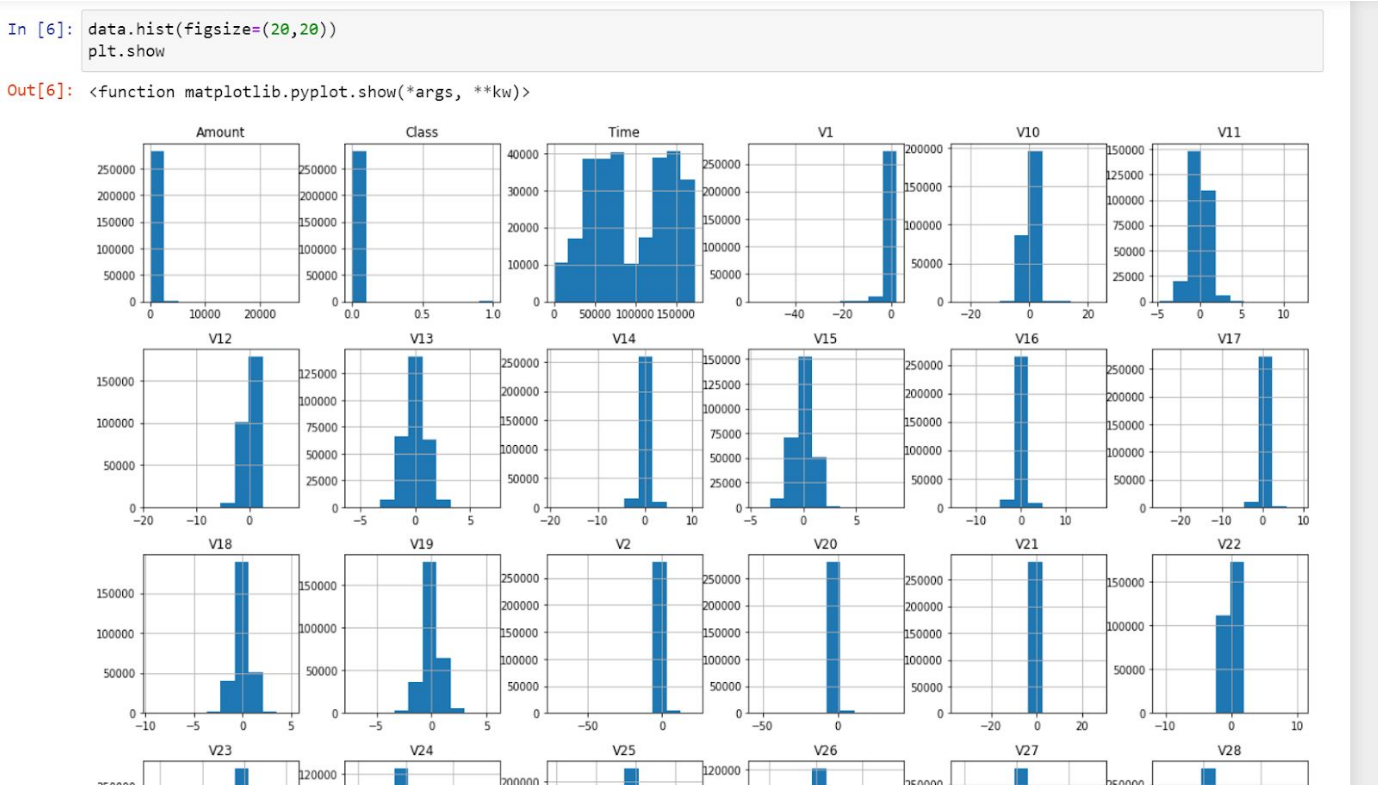
Software:

Python Anaconda

OS- Windows 7,8 and 10(32 and 64 bit)

Coding And Testing

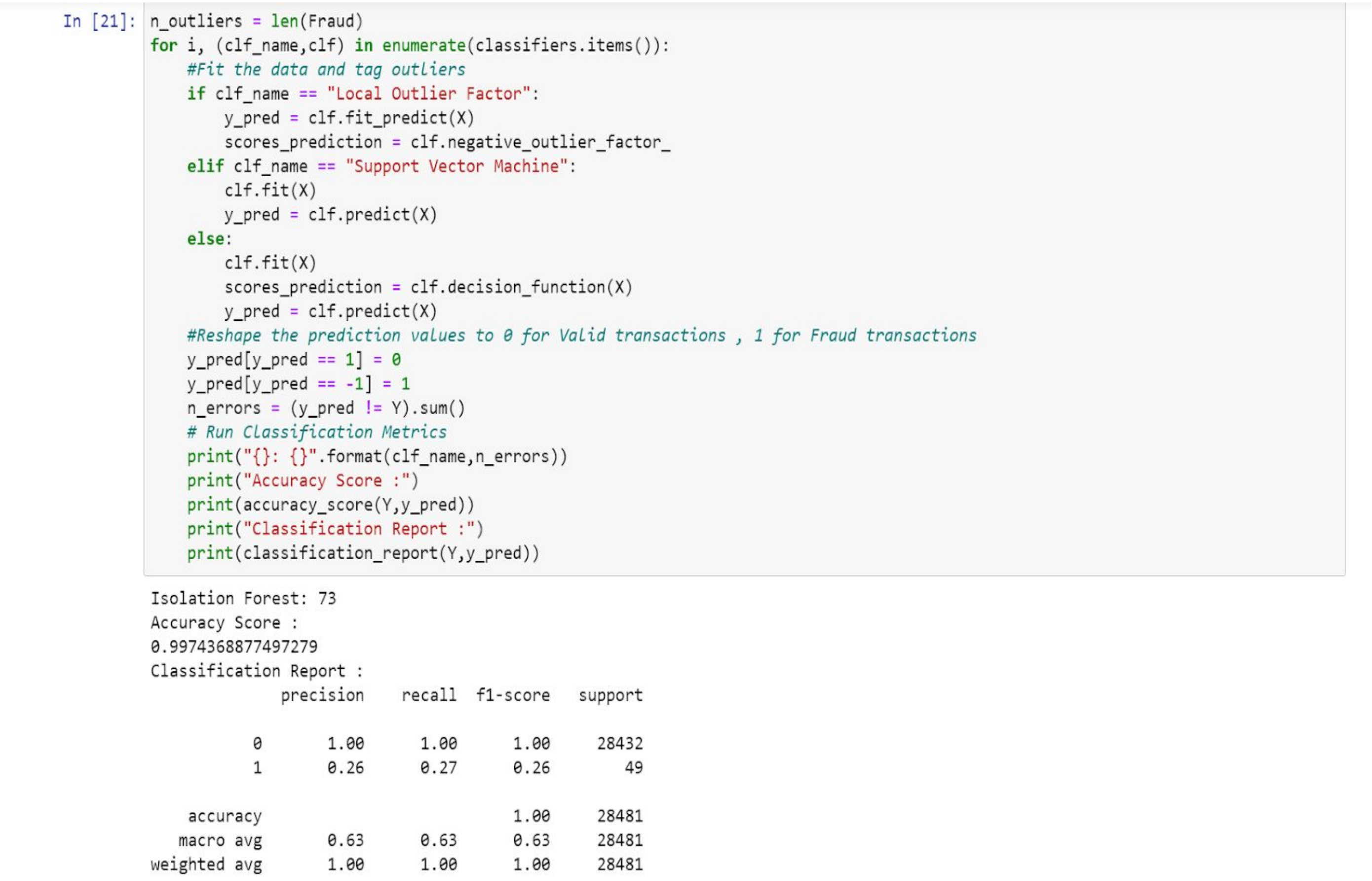
Testing the DataSet(Graphs)



```
In [6]: data.hist(figsize=(20,20))
        plt.show

Out[6]: <function matplotlib.pyplot.show(*args, **kw)>
```

Implementation of Algorithms using Python



In [21]:

```
n_outliers = len(Fraud)
for i, (clf_name,clf) in enumerate(classifiers.items()):
    #Fit the data and tag outliers
    if clf_name == "Local Outlier Factor":
        y_pred = clf.fit_predict(X)
        scores_prediction = clf.negative_outlier_factor_
    elif clf_name == "Support Vector Machine":
        clf.fit(X)
        y_pred = clf.predict(X)
    else:
        clf.fit(X)
        scores_prediction = clf.decision_function(X)
        y_pred = clf.predict(X)
    #Reshape the prediction values to 0 for Valid transactions , 1 for Fraud transactions
    y_pred[y_pred == 1] = 0
    y_pred[y_pred == -1] = 1
    n_errors = (y_pred != Y).sum()
    # Run Classification Metrics
    print("{}: {}".format(clf_name,n_errors))
    print("Accuracy Score :")
    print(accuracy_score(Y,y_pred))
    print("Classification Report :")
    print(classification_report(Y,y_pred))
```

```
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
```

Future Enhancements And Conclusion

Introduction

The findings obtained here may not be generalized to the global fraud detection problem. As future work, some effective algorithms which can perform well for the classification problem with variable misclassification costs could be developed.

Limitation/Constraints of the System

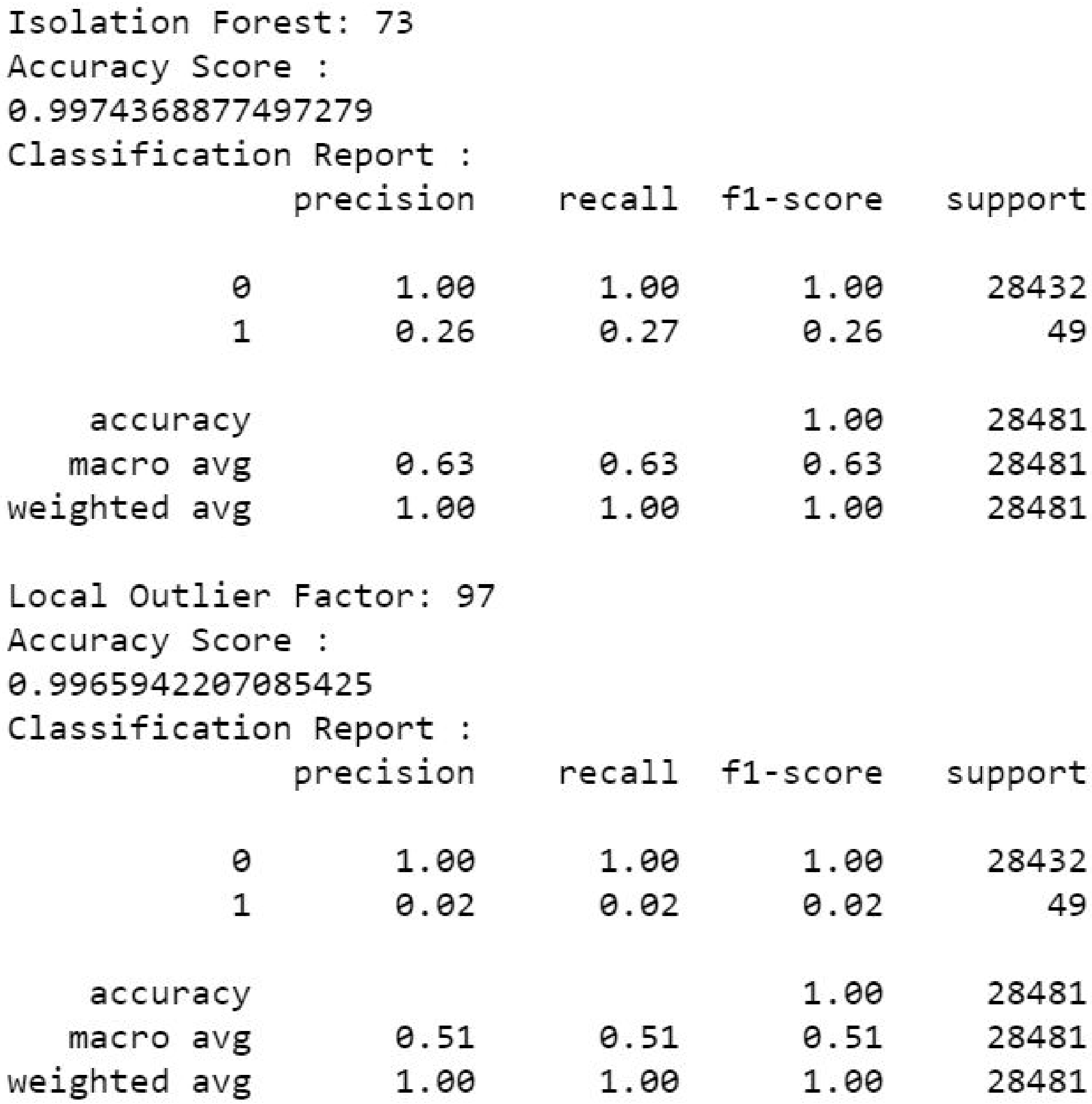
- Hardware Limitations: There are no hardware limitations.
- Interfaces to other Applications: There shall be no interfaces.
- Parallel Operations: There are parallel operations.
- Audit Functions: There shall be no audit functions.
- Control Functions: There shall be no control functions.

Conclusion

This method proves accurate in detecting fraudulent transactions and minimizing the number of false alerts. Genetic algorithm is a novel one in this literature in terms of application domain. If this algorithm is applied into bank credit card fraud detection systems, the probability of fraud transactions can be predicted soon after credit card transactions. And a series of anti fraud strategies can be adopted to prevent banks from great losses and reduce risks.

The objective of the study was taken differently than the typical classification problems in that we had a variable misclassification cost. As the standard data mining algorithms do not fit well with this situation we decided to use a multi population genetic algorithm to obtain an optimized parameter.

Result



```
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
```

REFERENCES

1.http://www.doc.ic.ac.uk/~nd/surpri se_96/journal/vol4/tcw2/report.html.

2.http://python.sun.com/developer/o nlineTraining/Programming//front.pyth on.html

3.http://www.faqs.org/patents/app/2 0100094765

4.Pressman, Roger S. Software engineering

5.Pattern Recognition and Machine Learning

6.The Elements of Statistical Learning: Data Mining, Inference, and Prediction

7.Python with Machine Learning