GALGOTIAS UNIVERSITY

(Established under Galgotias University Uttar Pradesh Act No. 14 of 2011)

CYBER FRAUD DETECTION

A Report for the Evaluation 3 of Project 2

Submitted by

SHIKHAR JAISWAL

(1613101675)

In partial fulfillment for the award of the degree

of

BACHELOR OF TECHNOLOGY

IN

COMPUTER SCIENCE AND ENGINEERING

SCHOOL OF COMPUTING SCIENCE AND ENGINEERING

Under the Supervision of

Dr. L. Godlin Atlas

Assistant Professor

APRIL / MAY- 2020

GALGOTIAS UNIVERSITY

(Established under Galgotias University Uttar Pradesh Act No. 14 of 2011)

SCHOOL OF COMPUTING AND SCIENCE AND ENGINEERING

BONAFIDE CERTIFICATE

Certified that this project report **"CBER FRAUD DETECTION"** is the bonafide work of **"SHIKHAR JAISWAL (1613101675)"** who carried out the project work under my supervision.

SIGNATURE OF HEAD

School of Computing Science & Engineering

SIGNATURE OF SUPERVISOR

Dr. L. Godlin Atlas

Assistant professor

School of Computing Science & engineering

TABLE OF CONTENTS

ABSTRACT

Fraud is one of the major ethical issues in the credit card industry. The main aims are, firstly, to identify the different, types of credit card fraud, and, secondly, to review alternative techniques that have been used in fraud detection. The sub-aim is to present, compare and analyze recently published findings in credit card fraud detection. This article defines common terms in credit card fraud and highlights key statistics and figures in this field. Depending on the type of fraud faced by banks or credit card companies, various measures can be adopted and implemented. The proposals made in this are likely to have beneficial attributes in terms of cost savings and time efficiency. The significance of the application of the techniques reviewed here is in the minimization of credit card fraud. Yet there are still ethical issues when genuine credit card customers are misclassified as fraudulent. Credit card fraud events take place frequently and then result in huge financial losses . The number of online transactions has grown in large quantities and online credit card transactions holds a huge share of these transactions. The use of credit cards is prevalent in modern day society. But it is obvious that the number of credit card fraud cases is constantly increasing in spite of the chip cards worldwide integration and existing protection systems.

1. INTRODUCTION

Overall introduction

Fraud has been increasing drastically with the progression of state-of-art technology and worldwide communication. Fraud can be avoided in two main ways: prevention and detection. Prevention avoids any attacks from fraudsters by acting as a layer of protection. Detection happens once the prevention has already failed.

Therefore, detection helps in identifying and alerting as soon as a fraudulent transaction is being triggered. The two types of frauds that can be mainly identified in a set of transactions are Card-not-present (CNP) frauds and Card-present (CP) frauds. Those two types can be described further by bankruptcy fraud, theft/counterfeit fraud, application fraud, and behavioural fraud.

The data which is being used in this study is analysed in two main ways: as categorical data and as numerical data. The dataset originally comes with categorical data. The

raw data can be prepared by data cleaning and other basic pre-processing techniques. First, categorical data can be transformed into numerical data and then appropriate techniques are applied to do the evaluation. Secondly, categorical data is used in the machine learning techniques to find the optimal algorithm.

Machine learning is this generation's solution which replaces such methodologies and can work on large datasets which is not easily possible for human beings.

This paper consists of selecting optimal algorithms for the four fraud patterns through an extensive comparison of machine learning techniques via an effective performance measure for the detection of fraudulent credit card transactions.

Fraud means obtaining services/goods and/or money by unethical means, and is a growing problem all over the world nowadays. Fraud deals with cases involving criminal purposes that, mostly, are difficult to identify. Credit cards are one of the most famous targets of fraud but not the only one; fraud can occur with any type of credit products, such as personal loans, home loans, and retail. Furthermore, the face of fraud has changed dramatically during the last few decades as technologies have changed and developed. A critical task to help businesses, and financial institutions including banks is to take steps to prevent fraud and to deal with it efficiently and effectively.

In two countries, credit cards have no competitors in terms of transaction product. Those two countries are the United Kingdom and Ireland. On the other hand, another group of country uses mostly debit cards; it is especially the case for Sweden. However, for this group, the standard deviation between the two types of transaction product is less visible than for the other group. As to Germany, for example, the German market appears to be underserved by credit cards. Indeed, payment by cards

has been increasing in the German market over the past few years. The market for credit and charge cards is forecast to grow by 23.3% from 2004 to 2009, to reach a value of €56,477 million.

The use of credit cards is prevalent in modern day society. But as in other related fields, financial fraud is also occurring in spite of the chip cards worldwide integration and existing protection systems. This is why most software developers are trying to improve existing methods of fraud detection in processing systems. The majority of such methods are rules based models. Such models allow bank employees to create the rules describing transactions that are suspicious. But the number of transactions per day is large and new types of the fraud appear quickly. Therefore, it is very difficult to track new types of fraud and to create corresponding rules in time. It would require a significant increase in the number of employees. Such problems can be avoided using of artificial intelligence.

The use of Bayesian Networks is suitable for this type of detection, but results from previous research showed that some input data (attributes of transaction) representation method should be used for effective classification [3]. For transaction monitoring by bank employees the clustering model was developed. This model allows provision of fast analysis of transactions by attributes.

A general description of the developed credit card fraud detection system, the clustering model, the Naïve Bayesian Classifier and the model based on Bayesian Networks with the data representation method are considered. Finally, conclusions about results of models' evaluative testing are made.

List of Figures

Figure Name **Page No.**

List of Abbreviation

ACRONYM	EXPANSION
IF	Isolation Forest
IT	Information Technology
LOF	Local Outlier Factor
SP	Service Provider
IDP	Identity Provider
kNN	K Nearest Neighbor

RFA	Random Forest Algorithm
DF	Data Frame

2.Hardware Configuration

- Servers: We will need a local host to implement an algorithm to tackle the situation. The local server will get implemented on any browser present on the system.
- Terminals: Jupyter .
- Processor Pentium –IV and above
- RAM 128 MD SD EAM
- Monitor 800* 600 resolution
- Hard disk 10 GB
- Processor 64 bit Intel core I3 and above
- Recent dataset uploaded by any bank
- Key board Standard 102 keys
- Mouse 3 buttons, scroll able
- Software – Windows 10,7,8.
- Prompt- Anaconda prompt.

3. Software Requirements

- Operating system Windows 10,7,8
- Coding Language Python
- Text editor- Jupyter
- Platform – Anaconda Prompt
- Machine Learning Algorithms

4. PURPOSE

The main objective of the project is to have a fraud free credit card transactions.

This program will be able to identify frauds actually various types of frauds which happens oftenly and identifies the types of patterns followed by fraudsters to attempt to loot user.

This system will definitely reduce the time, energy and money wasted in manually information spreading the details of how to save from being looted.

With the help of this program a real time program can be implemented to the program of ATM and by unsupervised learning it will detect by itself when there will be attempt to fraud without being programmed explicitly.

This will hopefully stop the fraud because this type of fraud is staying with us for decades and no successful solution has come through to stop it.

5. **Motivation and scope**

- The main motivation behind protecting people from being a victim of credit card fraud is money that lot of people earn is by working hard for throughout their life and when that money goes away from your hand will be the worst thing that can happen to a person.
- The fraudsters are very intelligent people but evil as they earn money by stealing from others and the worst thing is they get this for free.
- To stop the fraudster from making those fraud attempts we have to start thinking like them because only then we will get to know the pattern they are following and it will be easier for us to come up to a solution easily.
- Scope is to terminate this problem from the source and end it permanently so that these kind of things never show up again in future.
- Frauds can be stopped from both by bank and people if they stay more careful.
- The best scope to stop fraud is to get an effective solution that can help in terminating this permanently a decade long problem.
- There should be a lot of research to do so that this could stop very soon else the pattern of stealing and getting robbed will be continued for another decade.

6. PROBLEM WITH CURRENT SYSTEM

- The present scenario offers a preventive approach rather than tackling it**.**

- If a new type of fraud is detected the software will not be able to terminate the fraud process it is the biggest drawback for detection system.

- The old approach is only based on supervised learning and we have to implement a unsupervised method for a permanent solution**.**

- Foreign countries have been very progressive regarding credit card detection and they are approaching towards the permanent effective solution.

- Well different countries have different approach and different fraudsters so the solution might variate geographically.

- Current system is more about manually while this is the time we should have switched to the digital payments.

- To replace the current method there should be competitions that should be held regularly so that talent from various parts of the world could come forward and we can get something.

7. PROPOSED SYSTEM

- The proposed system is to have everything completely automated and computerized.
- The software is very easy to use and manage even for a non -technical person.
- The method is not supervised but it tells the pattern that is being followed by the fraudsters for over a decade.
- The project can implemented by professionals because the algorithms used here hybrid and probably never used before because it was implemented by me.
- We will need a new dataset as the number of transactions update daily and millions of transactions have been done daily.
- The best method to implement an effective is actually unsupervised method of machine learning because by that way the software can tackle the situation by itself no need to explicitly program.

- Hybrid algorithms have been used in our approach and multiple algorithms comes up to some effective solution because only then the multiple patterns could be detected.

Figure 1

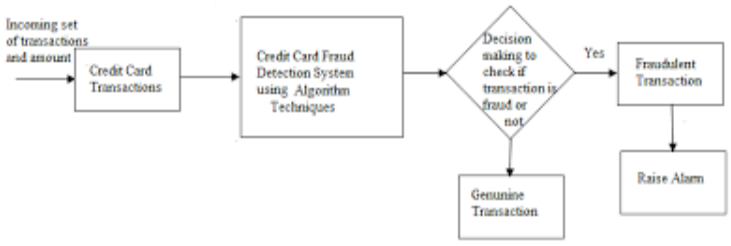
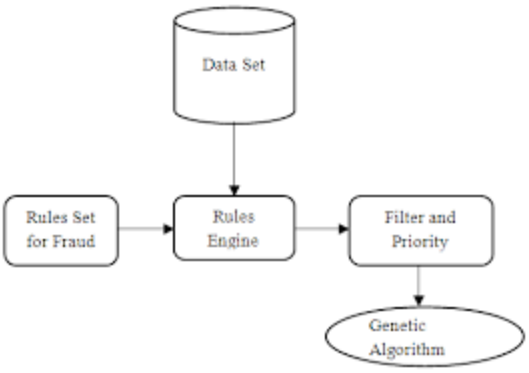


Figure 2



ALGORITHMS USED

1- **Isolation Forest**- It is an unsupervised learning algorithm for anomaly detection that works on the principle of isolating anomalies, instead of the most common techniques of profiling normal points. At the basis of the Isolation Forest algorithm there is the tendency of anomalous instances in a dataset to be easier to separate from the rest of the sample

(isolate), compared to normal points. In order to isolate a data point the algorithm recursively generates partitions on the sample by randomly selecting an attribute and then randomly selecting a split value for the attribute, between the minimum and maximum values allowed for that attribute.

2- **Local Outlier Factor**- The local outlier factor is based on a concept of a local density, where locality is given by nearest neighbors, whose distance is used to estimate the density. By comparing the local density of an object to the local densities of its neighbors, one can

identify regions of similar density, and points that have a substantially lower density than their neighbors. These are considered to be outliers.

3- **K nearest neighbor**- The k-nearest neighbors (KNN) algorithm is a simple, easy-to-implement supervised machine learning algorithm that can be used to solve both classification and regression problems. A **supervised machine learning** algorithm (as opposed to an unsupervised machine learning algorithm) is one that relies on labeled input data to learn a function that produces an appropriate output when given new unlabeled data.

1. Load the data
2. Initialize K to your chosen number of neighbors
3. For each example in the data

3.1 Calculate the distance between the query .

3.2 Add the distance and the index of the example to an ordered collection

4. Sort the ordered collection of distances and indices from smallest to largest

5. Pick the first K entries from the sorted collection

6. Get the labels of the selected K entries

7. If regression, return the mean of the K labels

8. If classification, return the mode of the K labels

4- **Random Forest Algorithm-** Random Forest algorithm is a supervised classification algorithm. We can see it from its name, which is to create a forest by some way and make it random. There is a direct relationship between the number of trees in the forest and the results it can get: the larger the number of trees, the more accurate the result. But one thing to note is that creating the forest is not the same as constructing the decision with information gain or gain index approach. The difference between Random Forest algorithm

and the decision tree algorithm is that in Random Forest, the process es of finding the root node and splitting the feature nodes will run randomly.

8. IMPLEMENTATION

Project is implemented in python. Below is the whole implementation:-

```
import sys

import numpy
```

```
import pandas

import matplotlib

import seaborn

import scipy

import sklearn

print('Python : ()',format(sys.version))

print('Numpy: ()',format(numpy.__version__))

print('Pandas: ()',format(pandas.__version__))

print('Matplotlib: ()',format(matplotlib.__version__))

print('Seaborn: ()',format(seaborn.__version__))

print('Scipy: ()',format(scipy.__version__))




print('sklearn: ()',format(sklearn.__version__))

#import important pacakges

import numpy as np
```

```
import pandas as pd

import  matplotlib.pyplot as plt

import seaborn as sns

#import the dataset from the file using pandas

data= pd.read_csv('F:\Project\creditcard.csv')

#exploring the dataset

print(data.columns)

print(data.shape)

print(data.describe)

data = data.sample(frac = 0.1 , random_state = 1)

print(data.shape)

#plot a histogram of each paramaeter

data.hist(figsize = (20,20))
```

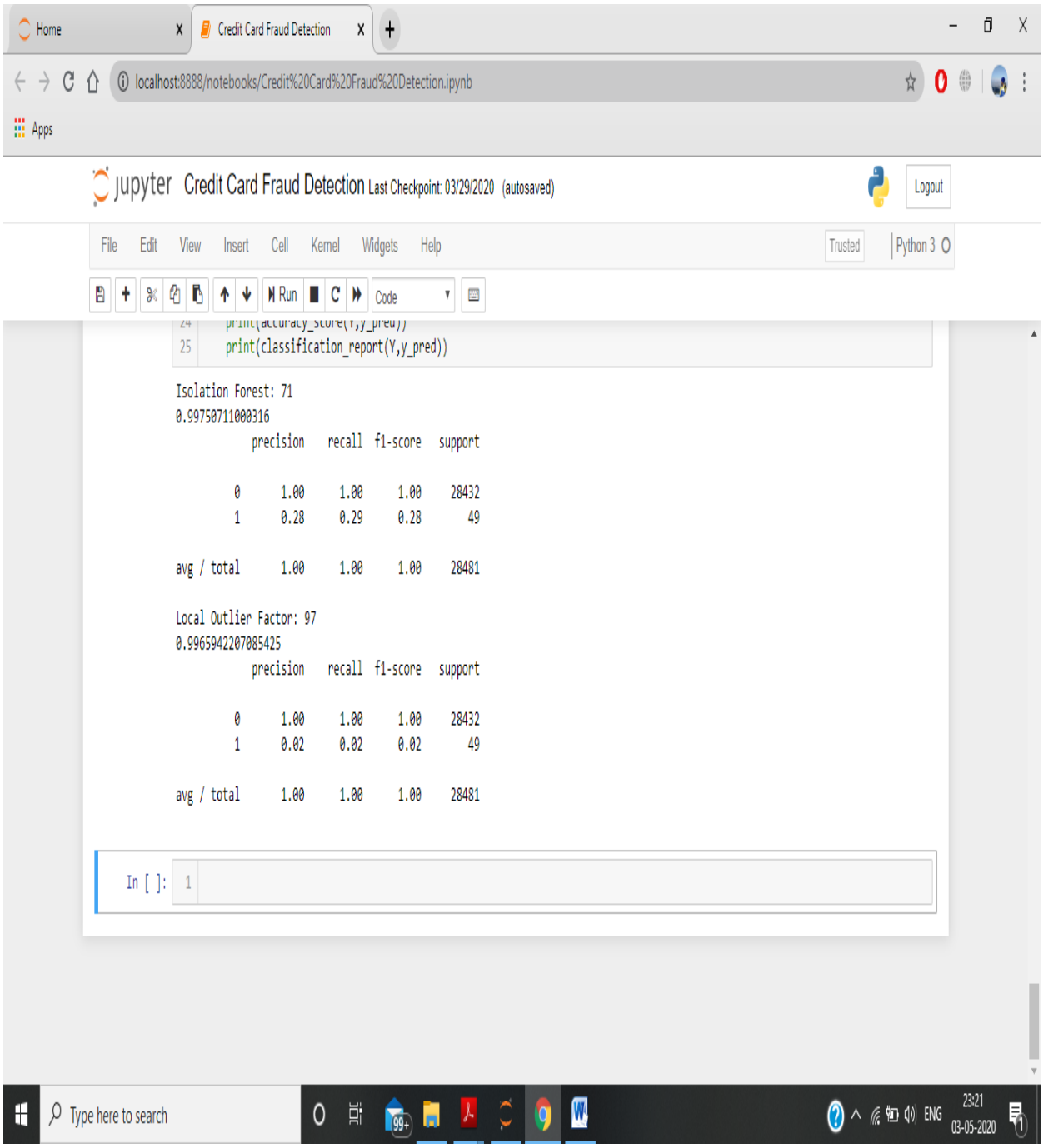
```
plt.show()

fraud= data[data['Class'] == 1]

valid= data[data['Class'] == 0]

outlier_fraction= len(fraud) / float(len(valid))

print(outlier_fraction)

print('Fraud Case ()', format(len(fraud)))

print('Valid Case ()',format(len(valid)))

#Corelation Matrix- Which measure is important to take action against it

corrmat= data.corr()

fig= plt.figure(figsize=(12,9))

sns.heatmap(corrmat, vmax= .8, square= True)

plt.show()

#This is unsupervised learning

#Get all the dataframe



columns= data.columns.tolist()
```

```
#Filter the column to remove the data we do not want

columns= [c for c in columns if c not in ['Class']]

target = 'Class'

x=data[columns]

y=data[target]

from sklearn.metrics import classification_report, accuracy_score

from sklearn.ensemble import IsolationForest

from sklearn.neighbors import LocalOutlierFactor

#Define a random state

state = 1

#Define the outlier methods

classifiers={

    "Isolation Forest": IsolationForest(max_samples=len(x),

                                        contamination = outlier_fraction,
```

```
 random_state =state),

    "Local Outlier Factor": LocalOutlierFactor(

    n_neighbors= 20,

    contamination = outlier_fraction)

}

#Fit the model

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_pred= clf.negative_outlier_factor_

    else:

        clf.fit(x)

        scores_pred= clf.decision_function(x)

        y_pred= clf.predict(x)
```

```
        #Reshape the production values to 0 for a valid, 1 for fraud




y_pred[y_pred ==1]=0

        y_pred[y_pred == -1] = 1

        n_errors= (y_pred != y).sum()

        #Run classification matrics

        print('() : ()',format(clf_name, n_errors))

       print(accuracy_score(Y,y_pred))

       print(classification_report(Y,y_pred))
```

9.OUTPUT



10. RESULT

- The Local Outlier Factor algorithm does not have that good impact as Isolation Forest algorithm on the credit card dataset.

- The above evaluation was done by setting up some parameters on which the precision was calculated.

- The Errors were more in Local Outlier Factor rather than Isolation Forest.

- That does not mean Local Outlier Factor is a bad algorithm its just that it does not fit well with credit card dataset.

- We were pretty successful in finding out the fraudulent cases because the precision rate was quite as it was expected.

Figure 3

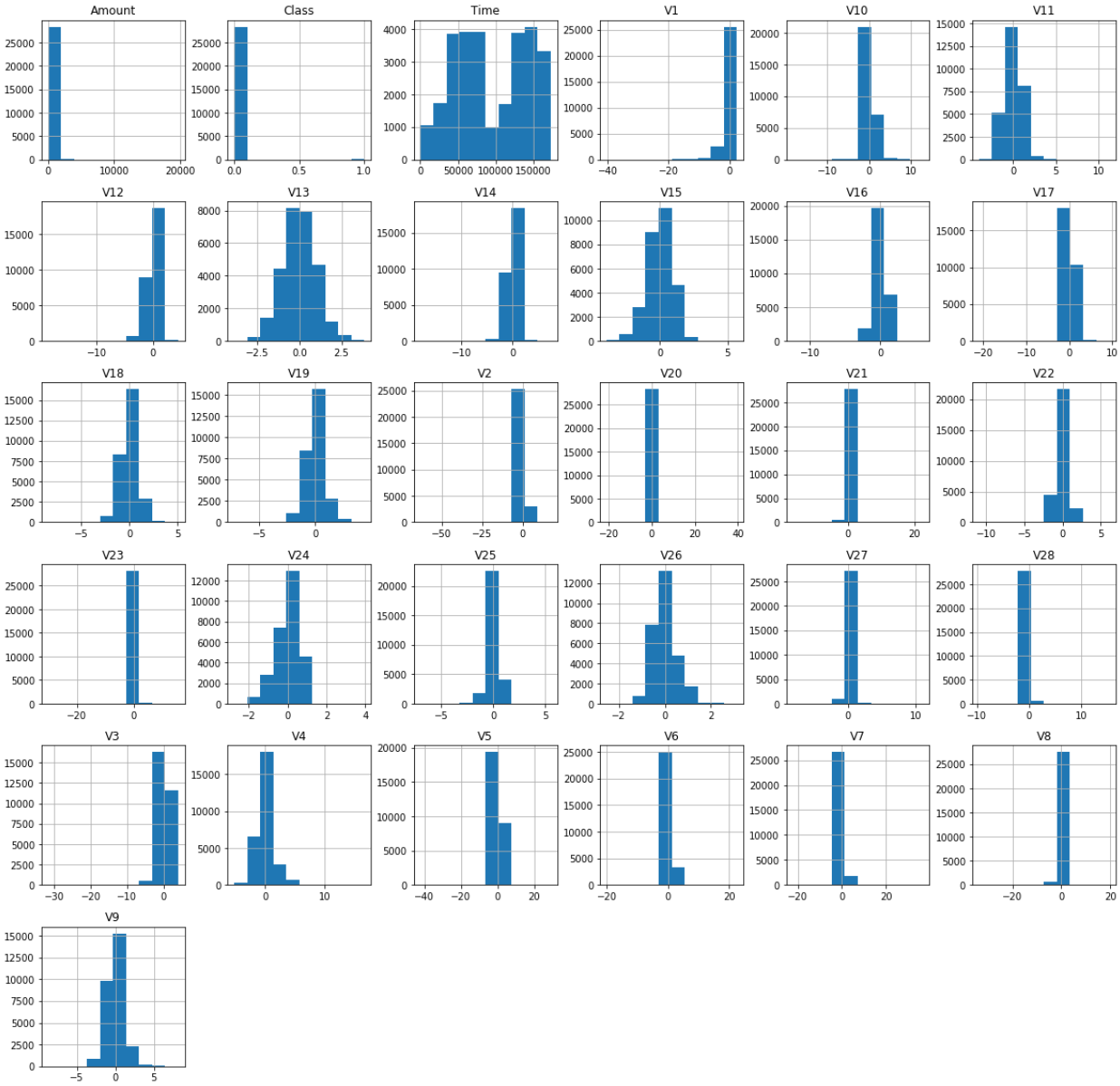
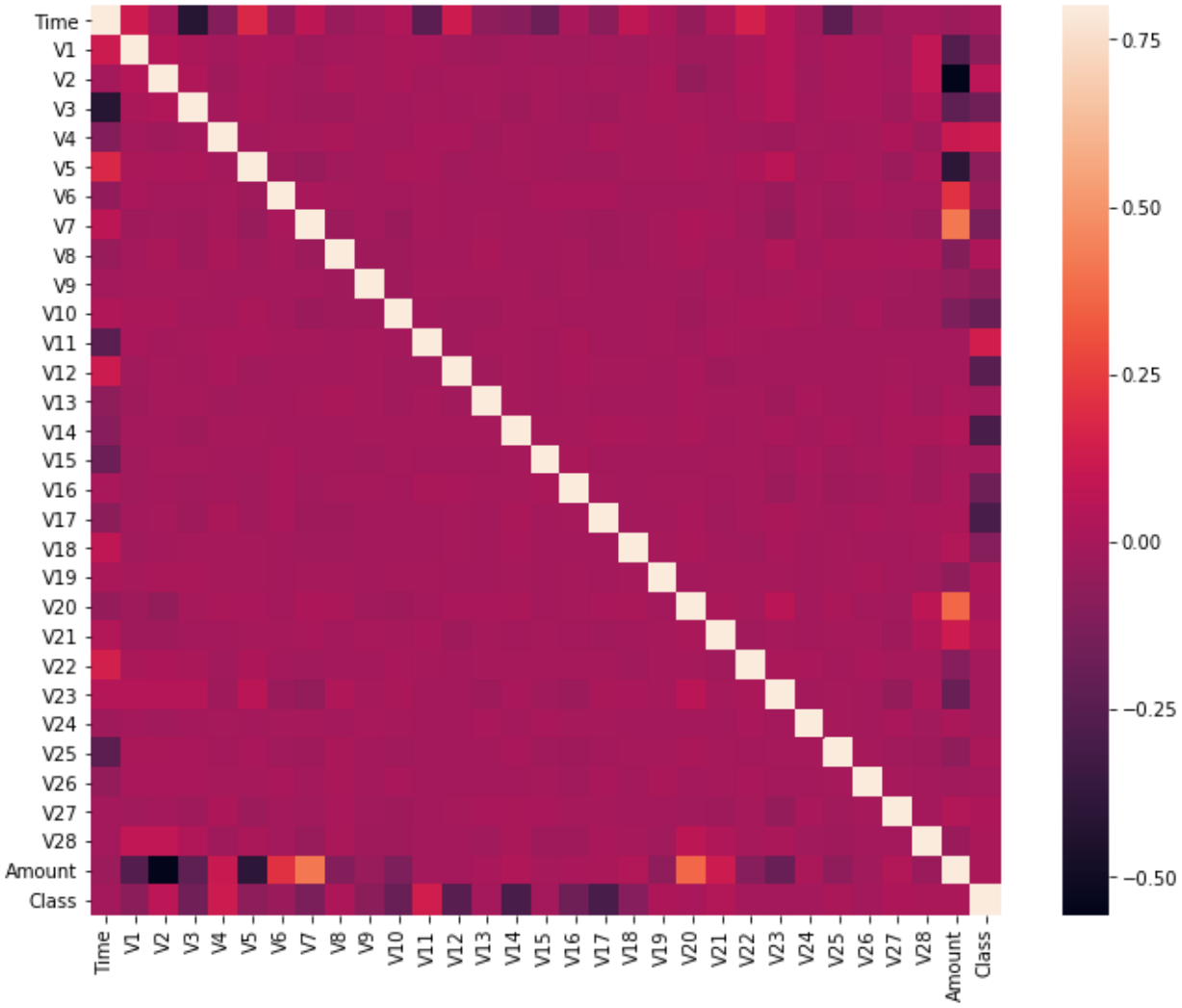

Amount
Class
Time
V1
V10
V11
V12
V13
V14
V15
V16
V17
V18
V19
V2
V20
V21
V22
V23
V24
V25
V26
V27
V28
V3
V4
V5
V6
V7
V8
V9

Figure 4



The GRAPH-2 is showing the correlation matrix. The correlation is between the amount if that is the fraudulent case or not.

The matrix lies over 0 so it means that we have a strong correlation between the amount if the case is fraudulent or not. The output is as expected.

11. CONCLUSION

From all the above we can conclude that if an effective solution is not detected then we will have to face this decade long problem to another decade. There are many algorithms that have been used into this projects of machine learning those were: Random Forest Algorithms , K nearest neighbors, Isolation Forest and Local Outlier Factor.

These algorithms have been used as hybrid algorithms because single algorithms were not effective that much. This method can be both used as supervised and unsupervised because it is implemented that way that it can be used as both. It is finely developed code program which can be used by various banks for increasing the security and protecting its customers from getting robbed by fraudsters.

Credit card fraud detection has been a keen area of research for the researchers for years and will be an intriguing area of research in the coming future. This happens majorly due to continuous change of patterns in frauds. In this paper, we propose a novel credit-card fraud detection system by detecting four different patterns of fraudulent transactions using best suiting algorithms and by addressing the related problems identified by past researchers in credit card fraud detection. A general description of the developed fraud detection system and comparison of base models have been presented. When comparing these models, the special evaluative testing was conducted. This evaluative testing was intended to simulate a typical use of credit cards.

This model allows banking employees to provide fast monitoring of incoming transactions. But the accuracy of classification for this model is not enough, because of the fact that the correlation between attributes is not taken into account in this model.

Clearly, credit card fraud is an act of criminal dishonesty. This article has reviewed recent findings in the credit card field. This paper has identified the different types of fraud, such as bankruptcy fraud, counterfeit fraud, theft fraud, application fraud and behavioral fraud, and discussed measures to detect them. Such measures have included pair-wise matching, decision trees, clustering techniques, neural networks, and genetic algorithms. From an ethical perspective, it can be argued that banks and credit card companies should attempt to detect all fraudulent cases. Yet, the unprofessional fraudster is unlikely to operate on the scale of the professional fraudster and so the costs to the bank of their detection may be uneconomic. The bank would then be faced with an ethical dilemma. Should they try to detect such fraudulent cases or should they act in shareholder interests and avoid uneconomic costs?

12. REFERENCE

http://ieeexplore.ieee.org/xpl/articleDetails.jsp?tp=&arnumber=5159014&queryText%3DCredit+Card+Fraud+Detection

http://ieeexplore.ieee.org/xpl/articleDetails.jsp?tp=&arnumber=323314&queryText%3DCredit+Card+Fraud+Detection