

A Project/Dissertation Review-ETE Report

on

FRAUD CREDIT CARD DETECTION

*Submitted in partial fulfillment of the
requirement for the award of the degree of*

BTECH IN COMPUTER SCIENCE AND ENGINEERING



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

Under The Supervision of

Name of Supervisor: Dr SPS Chauhan

Designation: Program Chair

Submitted By

Karishma Dhameja 18SCSE1010742 / 18021190006

Rishi Kumar 18SCSE1010224 / 18021011464

**SCHOOL OF COMPUTING SCIENCE AND ENGINEERING
DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING
GALGOTIAS UNIVERSITY, GREATER NOIDA
INDIA**

October, 2021

Table of Contents

Title	
Abstract	1
Chapter 1- Introduction	1
1.1 Introduction	2
1.2 Formulation of Problem	3
1.2.1 Tool and Technology Used	
Chapter 2- Literature Survey / Project Design	5

ABSTRACT

It is important for credit card companies to be able to detect fraudulent credit card transactions so that customers are not charged for items they did not purchase. Such problems can be addressed with Scientific Data and its value, as well as machine learning, cannot be overstated. This project aims to demonstrate data processing using credit card learning with Credit Card Fraud Detection. The Credit Card Determination Issue involves modeling a previous credit card transaction with details of those that have been identified as fraud. This model is used to determine whether a new transaction is fraudulent or not. Credit Card Fraud Detection is a common sample of segmentation. The process is focusing on analyzing and analyzing data sets and deploying multiple detection algorithms.

INTRODUCTION

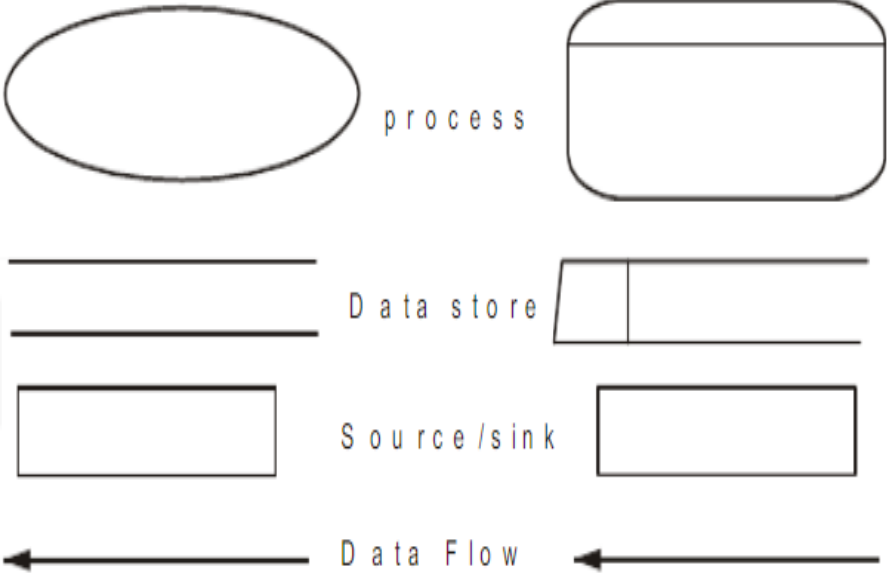
In recent years, the prevailing data mining concerns people with credit card Fraud detection model based on data mining. Since our problem is approached as a classification problem, classical data mining algorithms are not directly applicable. So an alternative approach is made by using general purpose meta heuristic approaches like genetic algorithms. This project is to propose a credit card fraud detection system using genetic algorithm. 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 algorithm where a set of interval valued parameters are optimized.

Data Flow Diagram

LITERATURE SURVEY

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 with 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. 4 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. However, when the problem is approached as a classification problem with variable misclassification costs as discussed above, the classical data mining algorithms are not directly applicable; either some modifications should be made on them or new algorithms developed specifically for this purpose are needed. An alternative approach could be trying to make use of general purpose meta heuristic approaches like genetic algorithms.

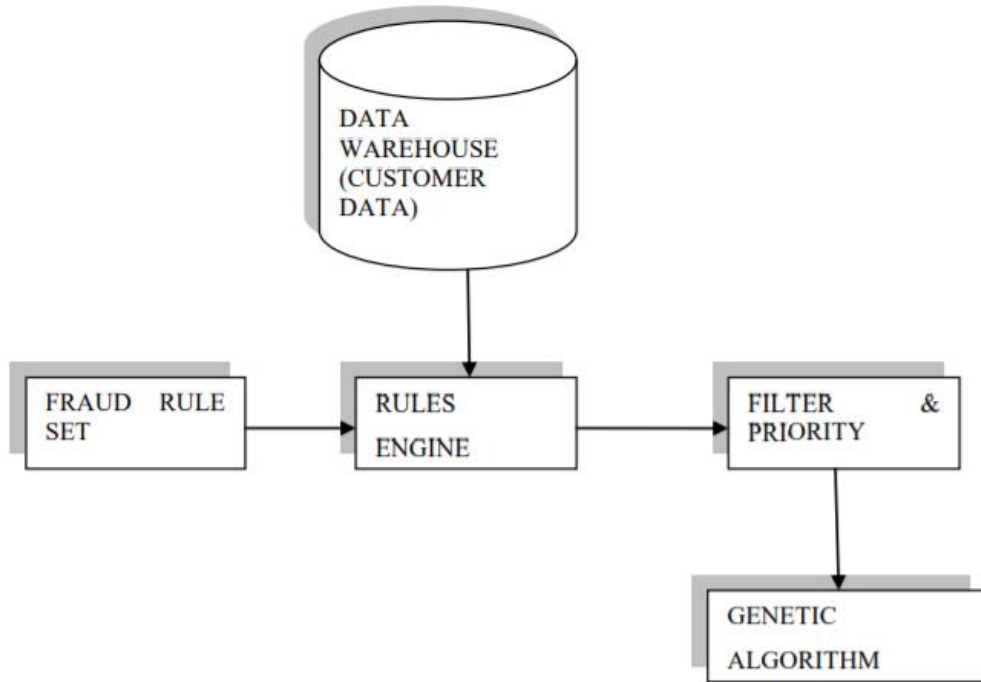
Data Flow Diagram



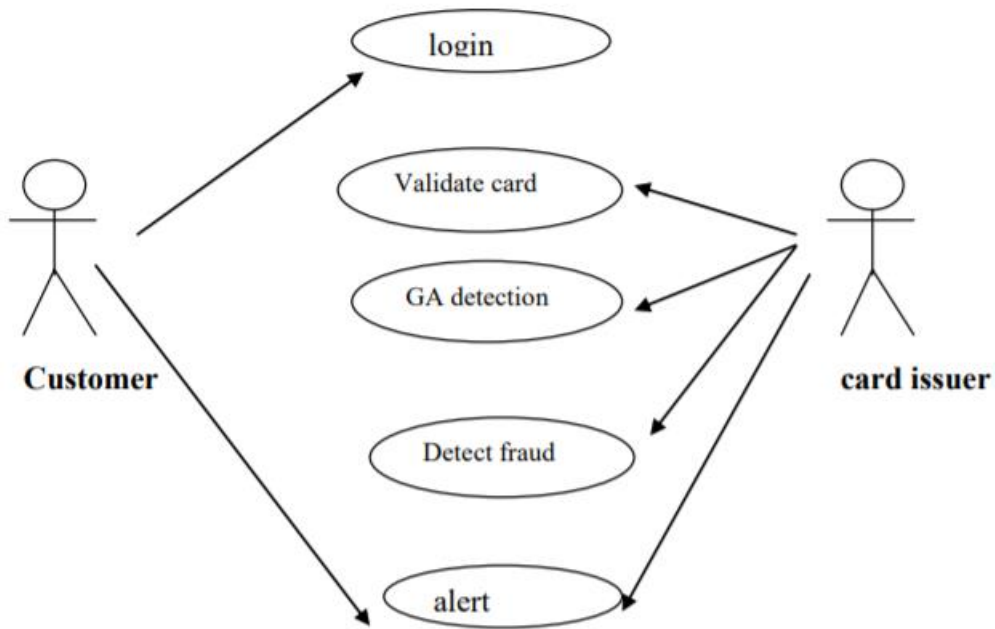
DeMarco &
Yourdon
symbols

Gane & Sarson
symbols

ARCHITECTURE SURVEY

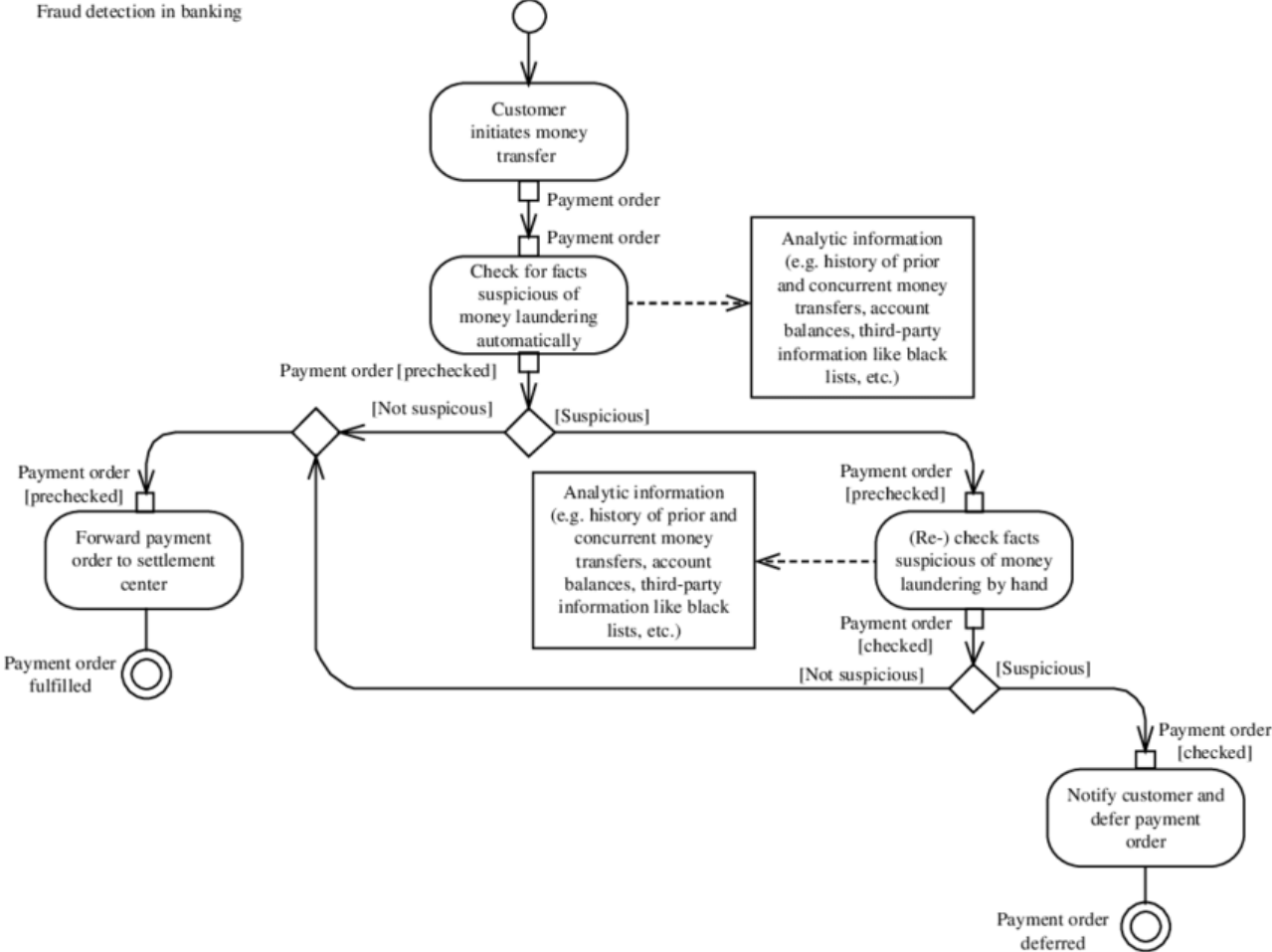


USE CASE DIAGRAM

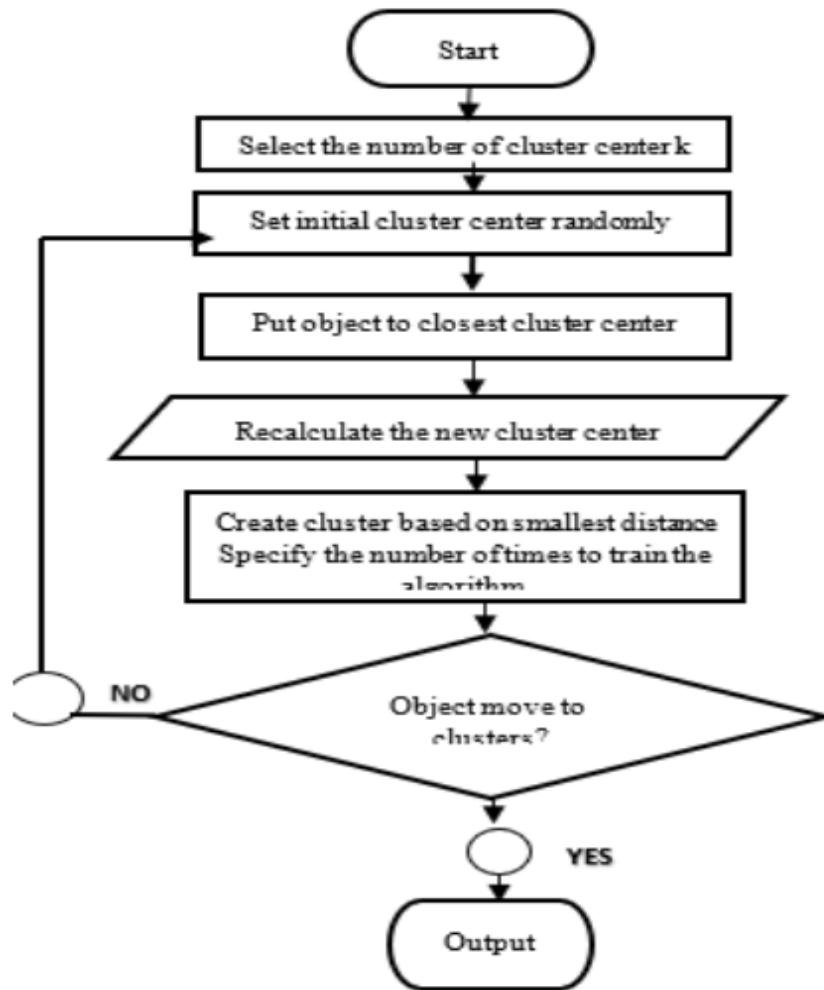


UML DIAGRAM

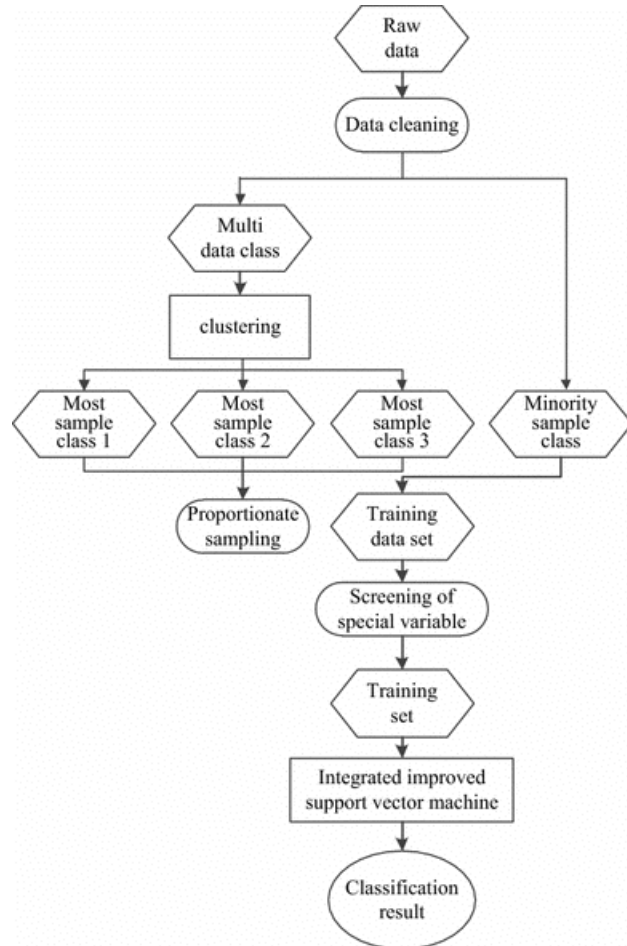
Fraud detection in banking



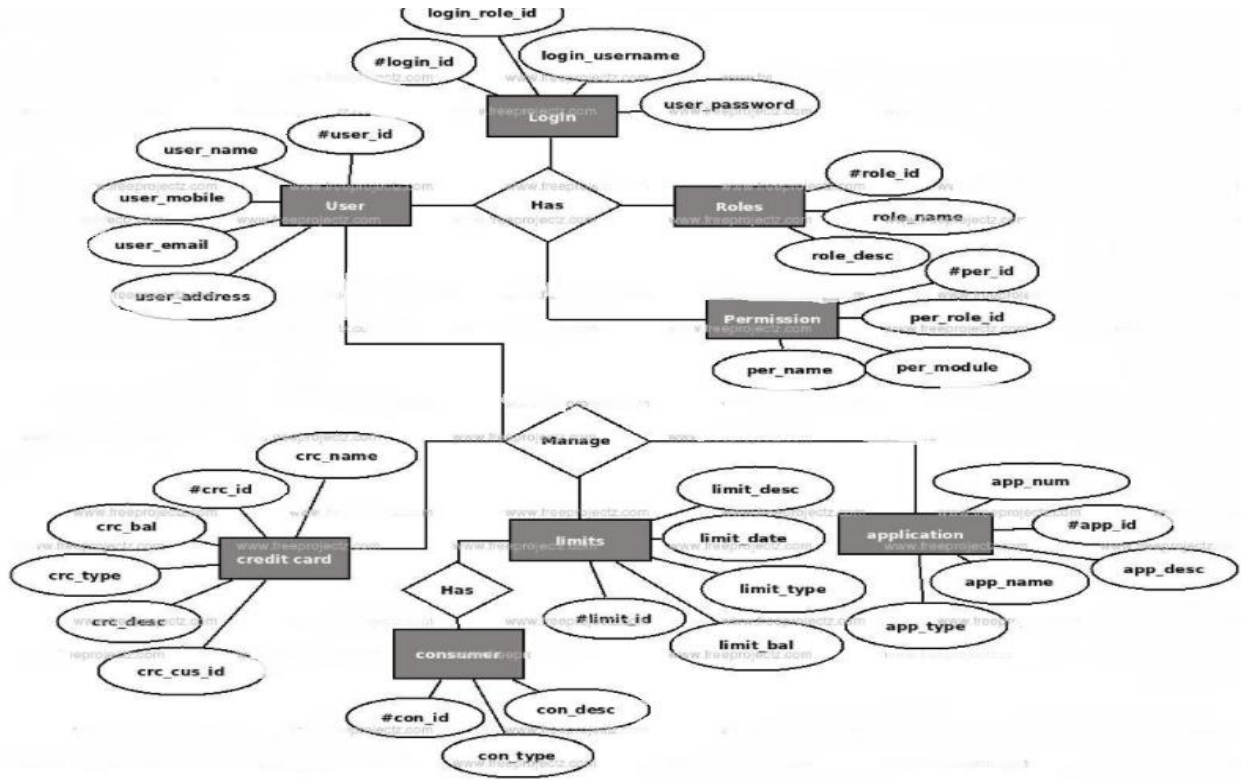
Flowchart Diagram



Schematic Diagram



ER Diagram



Module Description

System Requirements

The Software Requirements is produced at the culmination of the analysis task. The function and performance allocated to software as part of system engineering are refined by establishing a complete information description, a detailed functional and behavioral description, an indication of performance requirements and design constraints, appropriate validation criteria and other data pertinent to requirements

The Proposed system has the following requirements

1. System needs store information about new entry of credit card
2. System needs to help the internal staff to keep information of Transactions and find them as per various queries
3. System needs to maintain quantity record
4. System needs to keep record of Datasets

Software Requirements:

- Windows Xp, Windows 7(ultimate, enterprise)
- Sql 2008
- Visual studio 2010

Hardware Components:

- Processor – i3
- Min Hard Disk – 4 GB
- Min Memory – 1GB RAM

Functional Requirements

- Product and component based
- Creating & changing Issues at ease
- Accuracy in work
- Robust Database Back-end
- It contain better storage capacity
- Well designed reports
- Easy and Fast retrieval of information
- Simple Status & resolution
- Multi-Level priorities & severities
- Query issue list to any depth

Source Code

Importing all the necessary Libraries

```
# import the necessary packages
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns
from matplotlib import gridspec
```

Code : Loading the Data

```
# Load the dataset from the csv file using pandas
# best way is to mount the drive on colab and
# copy the path for the csv file
data = pd.read_csv("credit.csv")
```

Code : Understanding the Data

```
# Grab a peek at the data
data.head()
```

	Time	V1	V2	V3	V4	V5	V6	V7	V8	V9	V10	V11	V12	V13	V14
0	0.0	-1.359807	-0.072781	2.536347	1.378155	-0.338321	0.462388	0.239599	0.098698	0.363787	0.090794	-0.551600	-0.617801	-0.991390	-0.311169
1	0.0	1.191857	0.266151	0.166480	0.448154	0.060018	-0.082361	-0.078803	0.085102	-0.255425	-0.166974	1.612727	1.065235	0.489095	-0.143772
2	1.0	-1.358354	-1.340163	1.773209	0.379780	-0.503198	1.800499	0.791461	0.247676	-1.514654	0.207643	0.624501	0.066084	0.717293	-0.165946
3	1.0	-0.966272	-0.185226	1.792993	-0.863291	-0.010309	1.247203	0.237609	0.377436	-1.387024	-0.054952	-0.226487	0.178228	0.507757	-0.287924
4	2.0	-1.158233	0.877737	1.548718	0.403034	-0.407193	0.095921	0.592941	-0.270533	0.817739	0.753074	-0.822843	0.538196	1.345852	-1.119670

Code : Describing the Data

```
# Print the shape of the data

# data = data.sample(frac = 0.1, random_state = 48)

print(data.shape)

print(data.describe())
```

Output :

```
(284807, 31)
   Time          V1  ... Amount      Class
count 284807.000000  2.848070e+05  ... 284807.000000 284807.000000
mean   94813.859575  3.919560e-15  ...    88.349619    0.001727
std   47488.145955  1.958696e+00  ...   250.120109    0.041527
min     0.000000 -5.640751e+01  ...    0.000000    0.000000
25%   54201.500000 -9.203734e-01  ...    5.600000    0.000000
50%   84692.000000  1.810880e-02  ...   22.000000    0.000000
75%  139320.500000  1.315642e+00  ...   77.165000    0.000000
max  172792.000000  2.454930e+00  ... 25691.160000    1.000000
```

Code : Imbalance in the data

Time to explain the data we are dealing with.

```
# Determine number of fraud cases in dataset
fraud = data[data['Class'] == 1]
valid = data[data['Class'] == 0]
outlierFraction = len(fraud)/float(len(valid))
print(outlierFraction)
print('Fraud Cases: {}'.format(len(data[data['Class'] == 1])))
print('Valid Transactions: {}'.format(len(data[data['Class'] == 0])))
```

```
0.0017304750013189597
Fraud Cases: 492
Valid Transactions: 284315
```


Only 0.17% fraudulent transaction out all the transactions. The data is highly Unbalanced. Lets first apply our models without balancing it and if we don't get a good accuracy then we can find a way to balance this dataset. But first, let's implement the model without it and will balance the data only if needed.

Code : Print the amount details for Fraudulent Transaction

```
print("Amount details of the fraudulent transaction")
```

```
fraud.Amount.describe()
```

Output:

```
Amount details of the fraudulent transaction
count    492.000000
mean     122.211321
std      256.683288
min       0.000000
25%       1.000000
50%       9.250000
75%      105.890000
max     2125.870000
Name: Amount, dtype: float64
```

Code : Print the amount details for Normal Transaction

```
print("details of valid transaction")
```

```
valid.Amount.describe()
```

Output:

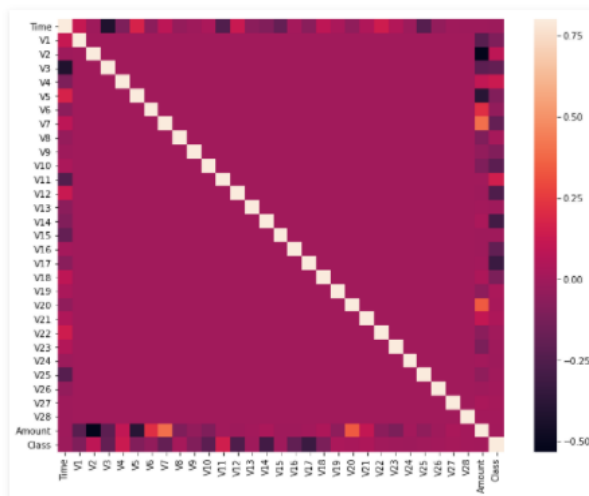
```
Amount details of valid transaction
count  284315.000000
mean    88.291022
std    250.105092
min     0.000000
25%     5.650000
50%    22.000000
75%    77.050000
max   25691.160000
Name: Amount, dtype: float64
```

As we can clearly notice from this, the average Money transaction for the fraudulent ones is more. This makes this problem crucial to deal with.

Code : Plotting the Correlation Matrix

The correlation matrix graphically gives us an idea of how features correlate with each other and can help us predict what are the features that are most relevant for the prediction.

```
# Correlation matrix
corrmat = data.corr()
fig = plt.figure(figsize = (12, 9))
sns.heatmap(corrmat, vmax = .8, square = True)
plt.show()
```



In the HeatMap we can clearly see that most of the features do not correlate to other features but there are some features that either has a positive or a negative correlation with each other. For example, V2 and V5 are highly negatively correlated with the feature called *Amount*. We also see some correlation with V20 and *Amount*. This gives us a deeper understanding of the Data available to us.

Code : Separating the X and the Y values

Dividing the data into inputs parameters and outputs value format

```
# dividing the X and the Y from the dataset

X = data.drop(['Class'], axis = 1)

Y = data["Class"]

print(X.shape)

print(Y.shape)

# getting just the values for the sake of processing

# (its a numpy array with no columns)

xData = X.values

yData = Y.values
```

Output :

```
(284807, 30)
```

```
(284807, )
```

Training and Testing Data Bifurcation

We will be dividing the dataset into two main groups. One for training the model and the other for Testing our trained model's performance.

```
# Using Skicit-learn to split data into training and testing sets

from sklearn.model_selection import train_test_split
```

```
# Split the data into training and testing sets

xTrain, xTest, yTrain, yTest = train_test_split(

    xData, yData, test_size = 0.2, random_state = 42)
```

Code : Building a Random Forest Model using skicit learn

```
# Building the Random Forest Classifier (RANDOM FOREST)

from sklearn.ensemble import RandomForestClassifier

# random forest model creation

rfc = RandomForestClassifier()

rfc.fit(xTrain, yTrain)

# predictions

yPred = rfc.predict(xTest)
```

Code : Building all kinds of evaluating parameters

```
# Evaluating the classifier

# printing every score of the classifier

# scoring in anything

from sklearn.metrics import classification_report, accuracy_score

from sklearn.metrics import precision_score, recall_score

from sklearn.metrics import f1_score, matthews_corrcoef

from sklearn.metrics import confusion_matrix
```

```

n_outliers = len(fraud)
n_errors = (yPred != yTest).sum()
print("The model used is Random Forest classifier")

acc = accuracy_score(yTest, yPred)
print("The accuracy is {}".format(acc))

prec = precision_score(yTest, yPred)
print("The precision is {}".format(prec))

rec = recall_score(yTest, yPred)
print("The recall is {}".format(rec))

f1 = f1_score(yTest, yPred)
print("The F1-Score is {}".format(f1))

MCC = matthews_corrcoef(yTest, yPred)
print("The Matthews correlation coefficient is{}".format(MCC))

```

Output :

```

The model used is Random Forest classifier
The accuracy is 0.9995611109160493
The precision is 0.9866666666666667
The recall is 0.7551020408163265
The F1-Score is 0.8554913294797689
The Matthews correlation coefficient is0.8629589216367891

```

Code : Visualizing the Confusion Matrix

```

# printing the confusion matrix

LABELS = ['Normal', 'Fraud']

conf_matrix = confusion_matrix(yTest, yPred)

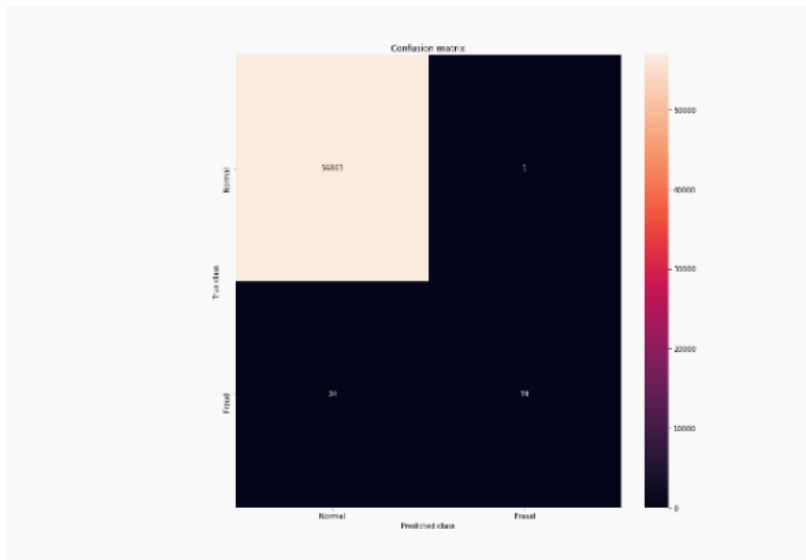
plt.figure(figsize =(12, 12))

sns.heatmap(conf_matrix, xticklabels = LABELS,

```

```
yticklabels = LABELS, annot = True, fmt = "d");  
  
plt.title("Confusion matrix")  
  
plt.ylabel('True class')  
  
plt.xlabel('Predicted class')  
  
plt.show()
```

Output :



Results



CREDIT CARD FRAUD DETECTION SYSTEM USING MACHINE LEARNING



File Management System
An online trading community provides participants with a one-to-one interface for trading, learning, or selling goods and services.



User Management System
A trading circle is a form of online trading designed to facilitate trading of financial or non-financial physical assets.



Credit Dataset System
Physical assets such as automobiles, DVDs and CDs are recharged via mail. Factors like age, gender, job, or website or to the next number.



ABOUT CREDIT CARD FRAUD DETECTION SYSTEM

The importance of Machine Learning and Data Science cannot be overstated. If you are interested in building good models and trading machines to learn with data how to define scenarios, identify unlabeled events, or predict a value in the present or future, data science is of the essence. It is essential to study the underlying data and model it by selecting an appropriate algorithm to apply with any new data. The success or failure of the algorithm used for the model is the data set. As a result, the developed application algorithms and business processes are subject to the problem of data bias, which will have a significant impact on the model of a value set using supervised learning paradigm classification, with Credit Card Fraud Detection being the best. Classification is machine learning paradigm that involves defining a function that will separate data into categories, or classes, where the best way to define a function that will separate data into categories is category identification is known. This function is often used to identify multiple of the categories a new observation belongs.

CONTACT INFO

Address: No.20000000000000000000
Hobbies: Country
Mobile: (111) 999-9999
Phone: (111) 999-9999
Email:

MODULES

- Credit Dataset Module
- User Module
- Fraud Detect (ML) Module
- Login Module
- Order Module

PROJECT LINKS

- Home
- About Us
- Contact
- Login
- Credit Us

ABOUT PROJECT

The importance of Machine Learning and Data Science cannot be overstated. If you are interested in building good models and trading machines to learn with data how to define scenarios, identify unlabeled events, or predict a value in the present or future, data science is of the essence.

ABOUT



About Credit Card Fraud Detection System

The importance of Machine Learning and Data Science cannot be overstated. If you are interested in analyzing trends and identifying patterns to learn with intention to define economic identity and label events, or predict a value in the present or future, data science is of the essence. It is essential to study the underlying dataset models for selecting an optimal data algorithm to approach any such case. The various control parameters of the algorithm need to be tweaked for the dataset. As a result, the developed application improves and becomes more efficient in solving the problem. In this, we have attempted to illustrate the modeling of values set using a machine learning paradigm. In this, we have Credit Card Fraud Detection being the best. Classification is a machine learning paradigm that involves deriving a function that will separate data into categories or classes, characterized by a training set of data containing observations. Data science predicts category membership between. This function is then used in identifying in which of the categories a new observation belongs.

CONTACT INFO

- Address: No. 00000 Street
- Map City, Country
- Mobile: (123) 456-7890
- Phone: (123) 456-7890
- Email: info@creditcardfraud.com

MODULES

- Credit Dataset Module
- User Module
- Fraud Detection Module
- Login Module
- Order Module

PROJECT LINKS

- Home
- About Us
- Contact
- Login
- Sign Up

ABOUT PROJECT

The importance of Machine Learning and Data Science cannot be overstated. If you are interested in analyzing trends and identifying patterns to learn with intention to define economic identity and label events, or predict a value in the present or future, data science is of the essence.

© Copyright Credit Card Fraud Detection System

About Project Screen

CONTACTS US

ADDRESS

Block 000000, 000000 Street,
New York, USA
100000

BUSINESS HOURS

Monday - Friday 9am to 6pm
Saturday - 10am to 5pm
Sunday - Closed

TELEPHONE

+000-0000000000
+000-0000000000

DROP US A LINE SO THAT WE CAN HEAR FROM YOU!

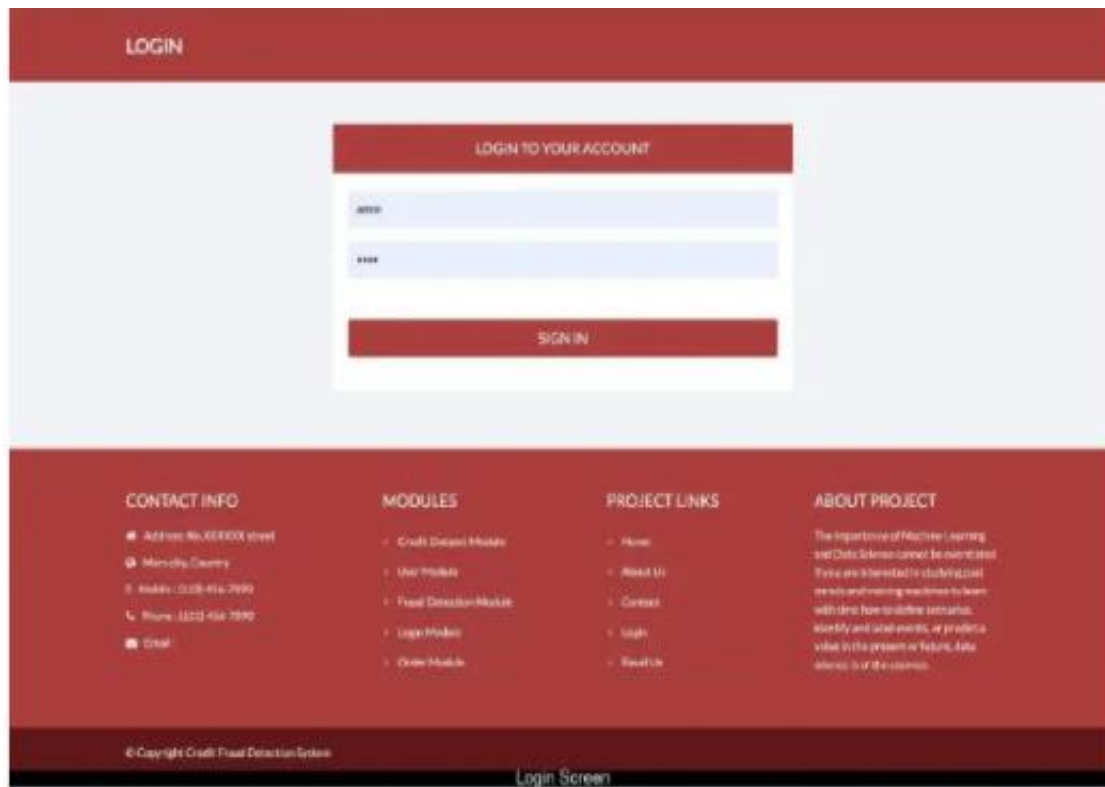
Name:

Email:

Phone:

Message:





ALL CREDIT CARD DATASET

Sr.No	File Name	Original File Name	Action
1	Heart Disease Data Set	heart.csv	View Data Delete Download Refresh

CONTACT INFO

- Address: No.0000000000000000
- Map City: Country
- Mobile: (123) 456-7890
- Phone: (123) 456-7890
- Email:

MODULES

- Credit Dataset Module
- User Module
- Fraud Detection Module
- Login Module
- Order Module

PROJECT LINKS

- Home
- About Us
- Contact
- Login
- Signup Us

ABOUT PROJECT

The importance of Machine Learning and Data Science cannot be overstated. If you are interested in studying past trends and training models to learn with data how to define anomalies, identify and label events, or predict a value in the present or future, data science is the essence.

© Copyright Credit Fraud Detection System

Admin All Dataset Screen

ADD FILES

UPLOAD CREDIT CARD DATASET

File Name

File Name

Select Credit Card Dataset File
 No File chosen

Description of the File

SUBMIT

CONTACT INFO

- Address: No.0000000000000000
- Map City: Country
- Mobile: (123) 456-7890
- Phone: (123) 456-7890
- Email:

MODULES

- Credit Dataset Module
- User Module
- Fraud Detection Module
- Login Module
- Order Module

PROJECT LINKS

- Home
- About Us
- Contact
- Login
- Signup Us

ABOUT PROJECT

The importance of Machine Learning and Data Science cannot be overstated. If you are interested in studying past trends and training models to learn with data how to define anomalies, identify and label events, or predict a value in the present or future, data science is the essence.

© Copyright Credit Fraud Detection System

Admin Dataset Upload Screen

CHANGE PASSWORD

CHANGE PASSWORD

New Password

Confirm Password

[CHANGE PASSWORD](#)

CONTACT INFO

- Address: No. XXXXXX Street
- Ministry, Country
- Mobility: (122) 454-7890
- Phone: (122) 434-7890
- Email

MODULES

- Credit Default Module
- User Module
- Fraud Detection Module
- Login Module
- Order Module

PROJECT LINKS

- Home
- About Us
- Contact
- Login
- Forgot Us

ABOUT PROJECT

The importance of Machine Learning and Data Science cannot be overstated. If you are interested in studying past trends and making machines to learn with time how to define scenarios, identify and label events, or predict a value in the present or future, data science is the resource.

© Copyright Credit Fraud Detection System

Change Password Screen

FRAUD DETECTION

CHOOSE TYPE OF DATA YOU WANT TO CHECK FOR DETECTING FRAUD

[ENTER FORM DATA MANUALLY](#)

[UPLOAD SINGLE CSV RECORD](#)

[UPLOAD MULTIPLE CSV RECORD](#)

CONTACT INFO

- Address: No. XXXXXX Street
- Ministry, Country
- Mobility: (122) 454-7890
- Phone: (122) 434-7890
- Email

MODULES

- Credit Default Module
- User Module
- Fraud Detection Module
- Login Module
- Order Module

PROJECT LINKS

- Home
- About Us
- Contact
- Login
- Forgot Us

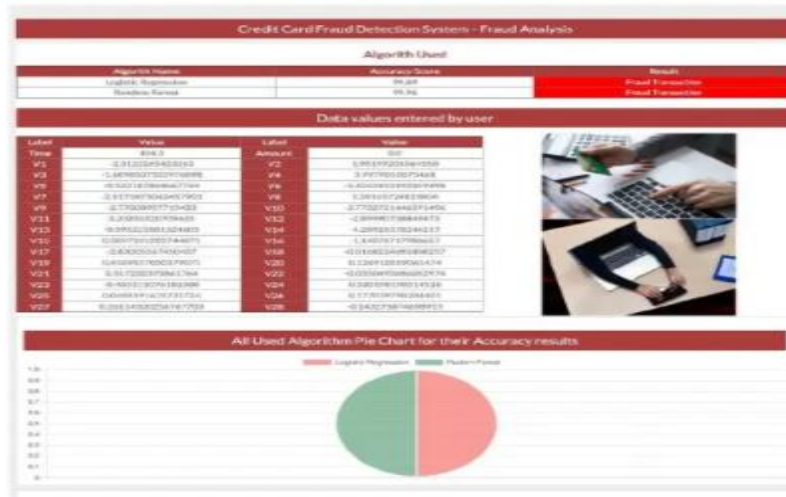
ABOUT PROJECT

The importance of Machine Learning and Data Science cannot be overstated. If you are interested in studying past trends and making machines learn with time how to define scenarios, identify and label events, or predict a value in the present or future, data science is the resource.

© Copyright Credit Fraud Detection System

Fraud Detection Forms Screen

FRAUD DETECTION



Div	CV	Text	Pin	Search				
3No.	Profile	Time	V1	V2	V3	V4	V5	V6
1	Fraud	1515	1.234230041246799	3.019746420703396	4.364346664796493	4.7327613843287	3.63433003055369	-1.36774
2	Valid	1525	4.14394309184470	-0.0704917939536	3.13645186904754	3.822734640749734	-0.96658025143017	0.49923
3	Fraud	1545	0.2299943301862037	0.712897924103	0.7194203801669	0.07026193560028	1.6673190321146	-0.42019
4	Valid	1546	0.0283997843428	0.69479547133174	1.107034446702797	2.0233200908236	0.1454802299923	-0.7184
5	Fraud	1551	0.02840398044492	3.8990754667829	5.61813903319437	0.007465300347793	1.03402516160172	-0.6132
6	Valid	0	-0.5248071334758	0.070781473896477	0.0634670769954	1.0761022427443	0.3365076994101803	0.6435
7	Valid	0	1.191607111149402	0.0641807289543	0.1664801300321	0.46134878460911	0.069615649283243	-0.9623
8	Valid	1	0.33823408101823	1.340343074780902	1.7732094240310	0.0777797302943278	0.300196123330131	1.80049

Showing 1 to 8 of 8 entries. Previous Next

CONTACT INFO

- Address: 76, XXXXXX street
- Mercuria Country
- Mobile: (020) 416-7660
- Phone: (020) 416-7670
- Fraud

MODULES

- Credit Default Module
- User Module
- Fraud Detection Module
- Login Module
- Order Module

PROJECT LINKS

- Home
- About Us
- Contact
- Login
- Event Us

ABOUT PROJECT

The integration of Machine Learning and Data Science cannot be overestimated. If you are interested in studying just trends and training machine to learn with time how to detect anomalies, identify and label events, or predict a value in the presence of feature data, refer to it of the content.

© Copyright Credit Fraud Detection System

Multiple Transaction Check Screen

ADD FILES

UPLOAD CSV FILE WITH SINGLE RECORD

Choose Single Record CSV File

Choose file:

SUBMIT

CONTACT INFO

- Address: No. XXXXXX Street
- Member: Country
- Mobile: (020) 456 7890
- Phone: (020) 456 7890
- Email

MODULES

- Credit Dataset Module
- User Worklog
- Fraud Detection Module
- Login Module
- Order Module

PROJECT LINKS

- Home
- About Us
- Contact
- Login
- Forgot Us

ABOUT PROJECT

The importance of Machine Learning and Data Science cannot be overstated. If you are interested in staying on top trends and creating machine learning with tools such as TensorFlow, identify and label events, or predict a value in the present or future data science is at the core.

Single Record File Upload Screen

FRAUD DETECTION


Credit Card Fraud Detection System - Fraud Analysis

Algorithm Used

Algorithm Name	Accuracy Score	Result
Logistic Regression	95.87	Valid Transaction
Random Forest	96.95	Valid Transaction

Data values entered by user

Label	Value	Label	Value
ID	0	Amount	149.62
V1	3.2998573384738	V2	4.0737611718986407
V3	3.33634623791434	V4	3.3782332427942
V5	-0.088202999433883	V6	0.84280777762282
V7	0.22998914041281	V8	0.098437902403907
V9	0.332769494031263	V10	0.0907441719389314
V11	-0.50199653368813	V12	-0.433608651613448
V13	0.991389641239408	V14	0.3212492049997897
V15	1.4841340739487	V16	0.4034651031047796
V17	0.28797124142924232	V18	0.0257803882982991
V19	0.483992946208733	V20	0.251412096234705
V21	-0.08326777944033	V22	0.277927070048899
V23	-0.10472993386767	V24	0.0849380718346731
V26	1.13493045417060833	V26	-0.58915484988864
V27	0.1320293719488948	V28	-0.023030024138833



All Used Algorithm Pie Chart for their Accuracy results

Algorithm	Accuracy Score
Logistic Regression	95.87
Random Forest	96.95

Conclusion

Our project is only a humble venture to satisfy the needs to manage their project work. Several user friendly coding have also adopted. This package sDatasets prove to be a powerful package in satisfying all the requirements of the school. The objective of software planning is to provide a framework that enables the manager to make reasonable estimates made within a limited time frame at the beginning of the software project and should be updated regularly as the project progresses.

At the end it is concluded that we hae made effort on following points

- A description of the backdatasets and context of the project and its relation to work already done in the area.
- Made statement of the aims and objective of the project
- The description of Purpose, Scope, and applicability
- We define the problem on which we are working in the project
- We can conclude that as the technology is developing day by day there are also fraudsters developing.
- Hence it is everyone's responsible to update about the technology and use it in a correct way.
- We should know about the Do's and Don'ts about the credit card before we start to use it and act accordingly to avoid any serious issues.

REFERENCES

- **Google for problem solving**
- **Head First PYTHON 2nd Edition**
- <http://www.jdbc-tutorial.com/>
- <http://www.PYTHON.net/>
- **PYTHON and SOFTWARE Design concepts by Apress**
- <https://www.tutorialspoint.com/sql/index.html>
- <https://www.docs.oracle.com/en>
- <https://www.wampserver.com/en/>
- **Database Programming with JDBC and PYTHON by O'Reilly**
- <https://www.tutorialspoint.com/python>
- <https://www.tutorialspoint.com/html>