# Credit Card Fraud Detection using ML

Project report submitted in partial fulfillment of the requirement
for the degree of Bachelor of Technology

In

**Computer Science and Engineering/Information Technology**

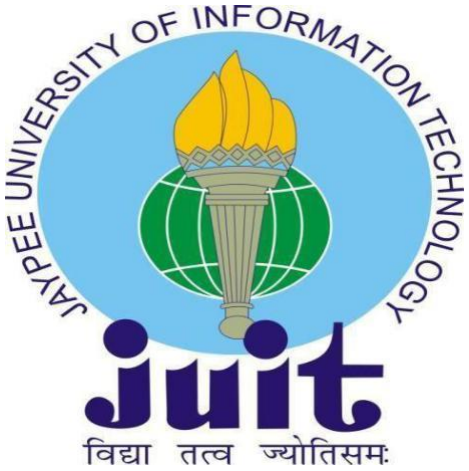
By

**Adhiraj Singh Jasrotia(171316)**
**Gaurav Dhiman(171323)**

Under the

Supervision of

**Mr. Surjeet Singh**




Department of Computer Science & Engineering

**Jaypee University of Information Technology**
**Waknaghat, Solan-173234, Himachal Pradesh**

# Candidate's Declaration

I hereby declare that the work presented in this report entitled **"Credit Card Fraud Detection"** in partial fulfillment of the requirements for the award of the degree of **Bachelor of Technology** in **Computer Science and Engineering/Information Technology** submitted in the department of Computer Science & Engineering and Information Technology**,** Jaypee University of Information Technology Waknaghat is an authentic record of my own work carried out over a period from Jan 2021 to May 2021 under the supervision of **Mr. Surjeet Singh**(Assistant Professor, Computer Science & Engineering and Information Technology).

The matter embodied in the report has not been submitted for the award of any other degree or diploma.

Gaurav Dhiman,171323 Adhiraj Singh Jasrotia,171316

This is to certify that the above statement made by the candidate is true to the best of my knowledge.

Mr. Surjeet Singh

# ACKNOWLEDGMENT

We have put a lot of effort into this project. But this was not possible without the kind support and help of many people. We thank all of them sincerely. **Mr. Surjeet Singh** is very grateful to us for providing leadership and continuous supervision as well as providing us with the necessary information about our project and for supporting our project completion. We would like to give special thanks and appreciation to our friends and colleagues who have given us time and attention.

Gaurav Dhiman(171323) Adhiraj Singh Jasrotia(171316)

Dated:

# TABLE OF CONTENTS

# LIST OF FIGURES

# ABSTRACT

With the evolution of recent technology, employment of credit cards has accumulated. As credit card becomes the trendiest quite payment for on-line payment what is more as manual payment, the numbers of master card frauds square measure increasing day by day. It's necessary to curb the master card frauds as a result of it causes amount of economic loss. So, we tend to search out the dishonorable and real transactions by victimization the conception of sophistication and additionally the accuracy of Isolation Forest and native Outlier issue is calculated to suggests the foremost effective rule for fraud detection. Machine Learning consists of the many algorithms which is able to use in fraud detection like Random Forest, native Outlier Factor, Isolation Forest, Naive man of science, K nearest Neighbors, Neural Networks, etc which is able to be employed in fraud detection which we tend to square measure near to use Isolation Forest and native Outlier in our project.

# CHAPTER I

# INTRODUCTION

Misrepresentation in credit card exchanges is unapproved and unwanted usage of companion diploma account via way of means of a person beside the owner of that account. Vital obstruction measures is probably taken to stop this maltreatment and for this reason the behavior of such stunning practices is probably focused to decrease it and protect towards comparative activities interior whats to come. In optional words, credit card Fraud is probably laid out as a case any vicinity an person makes use of every other humans credit card for personal motives aleven though the owner and in the end the cardboard flexibly experts rectangular degree uninformed of the very reality that the card is being utilized.

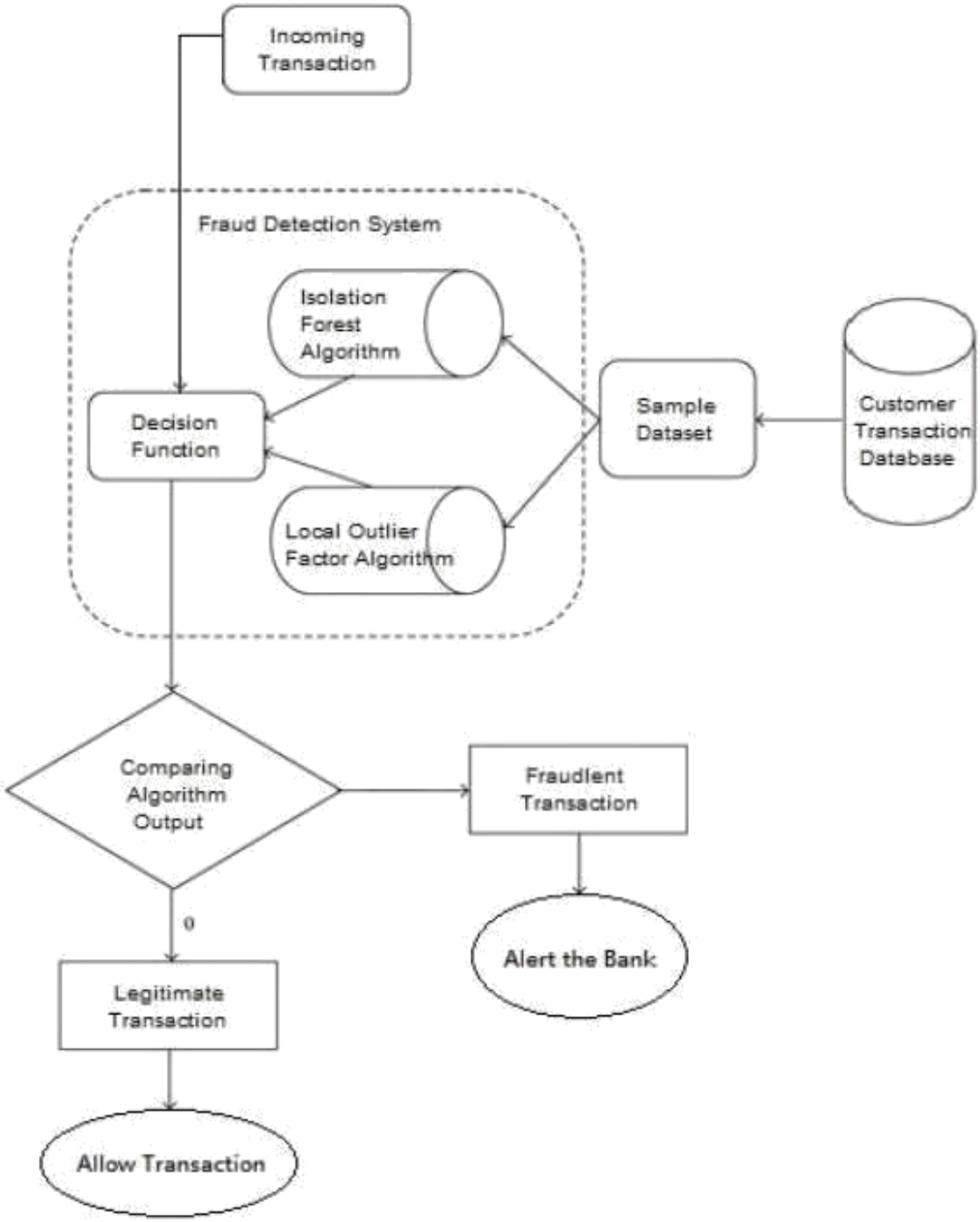
Extortion recognition includes recognition the exercises of populaces of clients to appraise, comprehend or keep away from questionable conduct, that obliges misrepresentation, interruption, and defaulting. This is a dreadfully important disadvantage that requests the eye of networks like AI and data science any place the answer for the current downside might be programmed.

This downside is particularly troublesome from the point of learning, since it is portrayed by changed variables like class irregularity. the amount of legitimate exchanges such a great amount out number disreputable ones. Likewise, the gathering activity designs frequently alteration their applied science properties throughout the course of time.

These are not using any and all means the main challenges in the execution of a genuine deception acknowledgment system, regardless. In authentic world models, the massive stream of portion requests is quickly sifted by means of modified gadgets that figure through which trades to favor. Computer based intelligence figurings are used to separate all the endorsed trades and report the questionable ones. These reports are investigated by specialists who contact the cardholders to insist if the trade was authentic or counterfeit.

The agents give an input to the mechanized framework which is utilized to prepare and refresh the calculation to ultimately improve the misrepresentation location execution over the long haul.

Credit Card Fraud Detection


Incoming Transaction
Fraud Detection System
Isolation Forest Algorithm
Decision Function
Local Outlier Factor Algorithm
Sample Dataset
Customer Transaction Database
Comparing Algorithm Output
Fraudlent Transaction
Alert the Bank
0
Legitimate Transaction
Allow Transaction

Flowchart

The Visa blackmail acknowledgment features uses customer lead and territory inspecting to check for amazing models. These models consolidate customer credits, for instance, customer spending plans similarly as regular customer geographic zones to affirm his character. If any weird model is recognized, the system requires revivification. The system assessments customer charge card data for various traits.One of the strategies by that Visa extortion is achievable is by acquiring admittance to the purloined credit cards and other is by abusing or to take the fundamental purposes of the card by means of online exchange. While identifying this kind of misrepresentation we will in general face a few challenges. Millions and billions of exchanges happen every moment wherever the planet. Identifying that among all the exchanges are happened furthermore, which one is genuine could be an undertaking. The quantity of investigation dispensed during this field to discover the deceitful exchanges are low a result of the lack of accessibility of datasets. Since the subtleties of the clients are classified, following up on the significant informational collections gets unrealistic. AI calculations are inexactly isolated into directed and unaided calculations. In directed calculations, a preset arrangement of information is accommodated preparing the framework.

The framework attempts to anticipate the outcomes upheld the all around advertised results i.e., preparing dataset. Though if there should arise an occurrence of solo calculations the framework attempts to look out the examples straightforwardly from the occasion gave. The irregular timberland calculation is a regulated arrangement recipe. It's utilized for characterization since it gives right outcomes than the other arrangement algorithmic principle. Nearby Outlier factor is an inconsistency identification algorithmic guideline. The exception is just a word for peculiarities. Inconsistency speaks to the irregular conduct or a deviation from the customary conduct of a data focuses concerning certain credits. Exceptions have an interesting factual property.

The Isolation Forest detaches perceptions via way of means of unpredictably selecting an detail and in a while each what path selecting a break up an incentive among the maximum severe and least estimations of the picked highlight. Since recursive parceling is depicted via way of means of a tree structure, the degree of parting anticipated to disconnect an instance is corresponding to the manner duration from the foundation hub to the finishing hub. Utilizations of these calculations are discourse acknowledgment, banking area, scientific services, layout acknowledgment, and so on At the top of the undertaking, calculations Local Outlier Factor and Isolation Forest is contrasted with discover which one offers the least complicated final results for misrepresentation location.

# PROBLEM STATEMENT

Charge card misrepresentation discovery gets troublesome as a result of two significant reasons first the profile of conventional and offensive practices change interminably and second, credit card data sets are very slanted.

Due to digitalization of the multitude of administrations like banking, selling and so forth The use of any very installment cards for exchange could be a horribly customary technique for each person. Are it turns out to be horribly hard to work out the genuine and erroneous exchanges. person. Are it turns out to be horrendously hard to work out the genuine and fraudulent transactions classification algorithmic guideline. Nearby Outlier factor is a peculiarity location algorithmic principle. The anomaly is simply a word for abnormalities. Peculiarity speaks to the strange conduct or a deviation from the customary conduct of a data focuses concerning certain credits. Exceptions have a special factual property.

The Isolation Forest confines perceptions with the aid of using unpredictably selecting an detail and later on each what route selecting a cut up an incentive among the best and least estimations of the picked highlight. Since recursive apportioning is depicted with the aid of using a tree structure, the degree of parting predicted to detach an instance is same to the manner duration from the foundation hub to the finishing hub.

Uses of those calculations are discourse acknowledgment, banking area, medical care, design acknowledgment, and so on At the tip of the task, two calculations Local Outlier Factor and Isolation Forest is contrasted with learn which one gives the easiest outcome for extortion location.

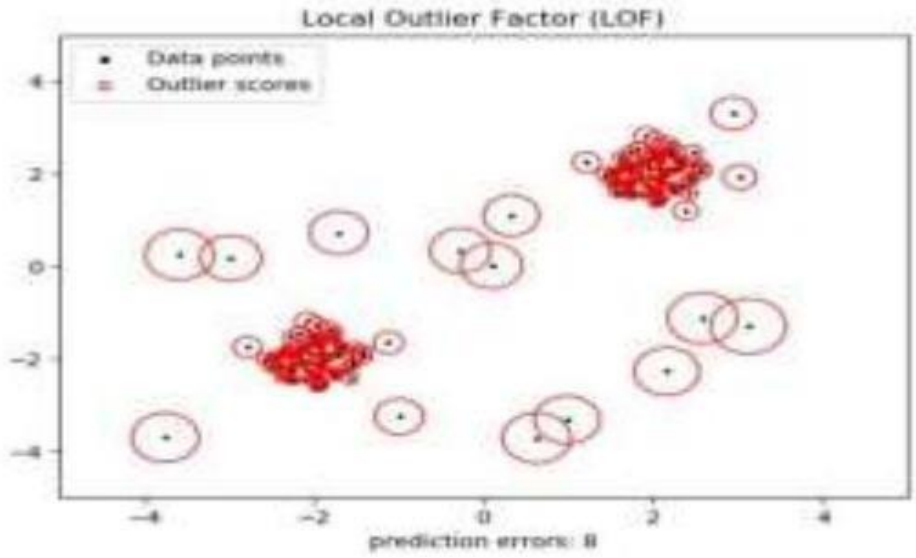
# OBJECTIVE

To preprocess dataset and split it into preparing and testing stage. Perform execution of Isolation forest calculation. Perform correlation with nearby exception calculation for precision examination.To plan a classifier to order misrepresentation and genuine exchanges to expand precision level.

# METHODOLOGY

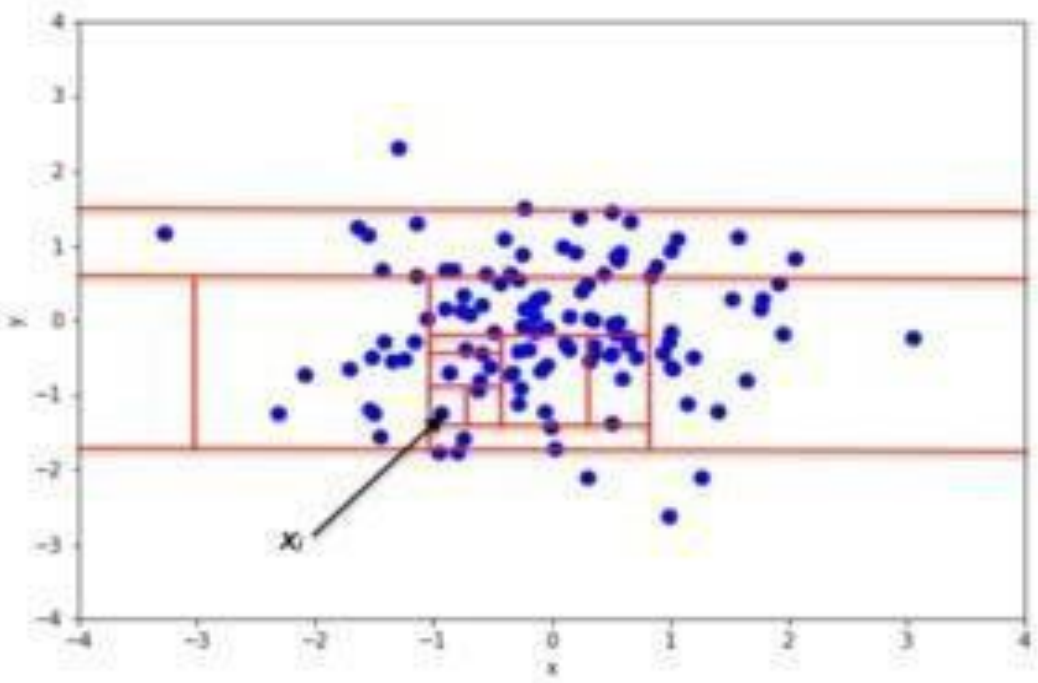
The most essential job has been battled by the dataset. Since it is absurd to expect to figure on a real time information base, we tend to are utilizing the creditcard.csv dataset from Kaggle which fuses 2,83,807 passages. The different boundaries utilized in the information base are time, class, sum, area and, and so on Such sort of complete of thirty one boundaries is utilized. V1 - V28 are the fields of a PCA dimensionality decrease to shield the personality and delicate alternatives of a client. In this undertaking, we are pointing towards assurance of exchanges with a high probability to shoulder credit card misrepresentation.

Local Outlier Factor: The peculiarity rating of every instance is called the close by Outlier factor. It ascertains the nearby deviation of the thickness of a given example in connection to its neighbors. It's known as nearby since the irregularity score relies upon the object secluded the thing is with connection to the enclosing area.



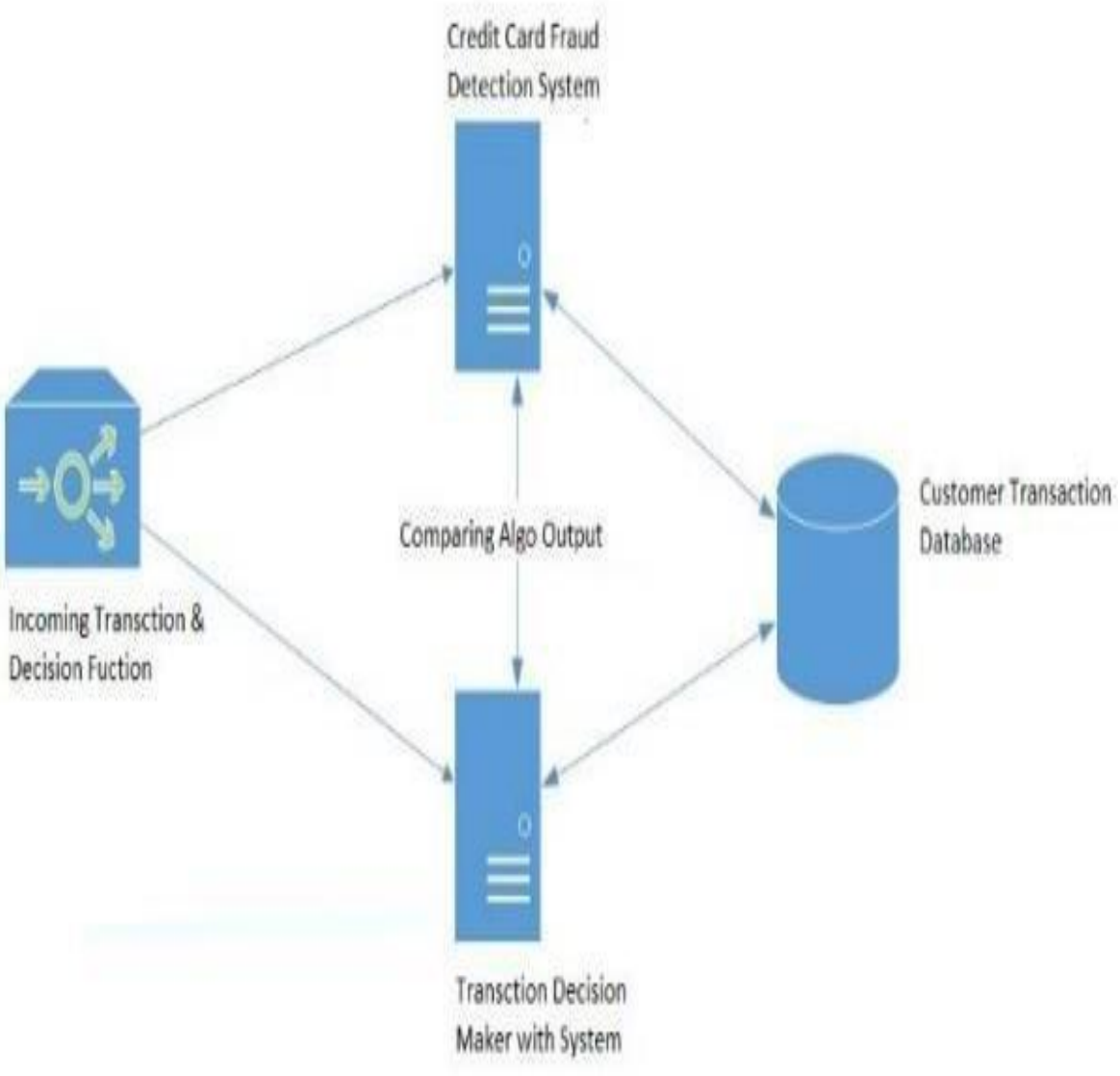

Local Outlier Factor

Isolation Forest Algorithm: The Isolation Forest confines the perceptions by unpredictably picking an element by haphazardly choosing a worth and parting it in between the atmost most extreme and least estimations of the picked include. Additionally the recursive dividing strategy is signified by a tree-like structure, the amount of parting needed to confine an example is similar to the path length from the root hub to the ending hub.

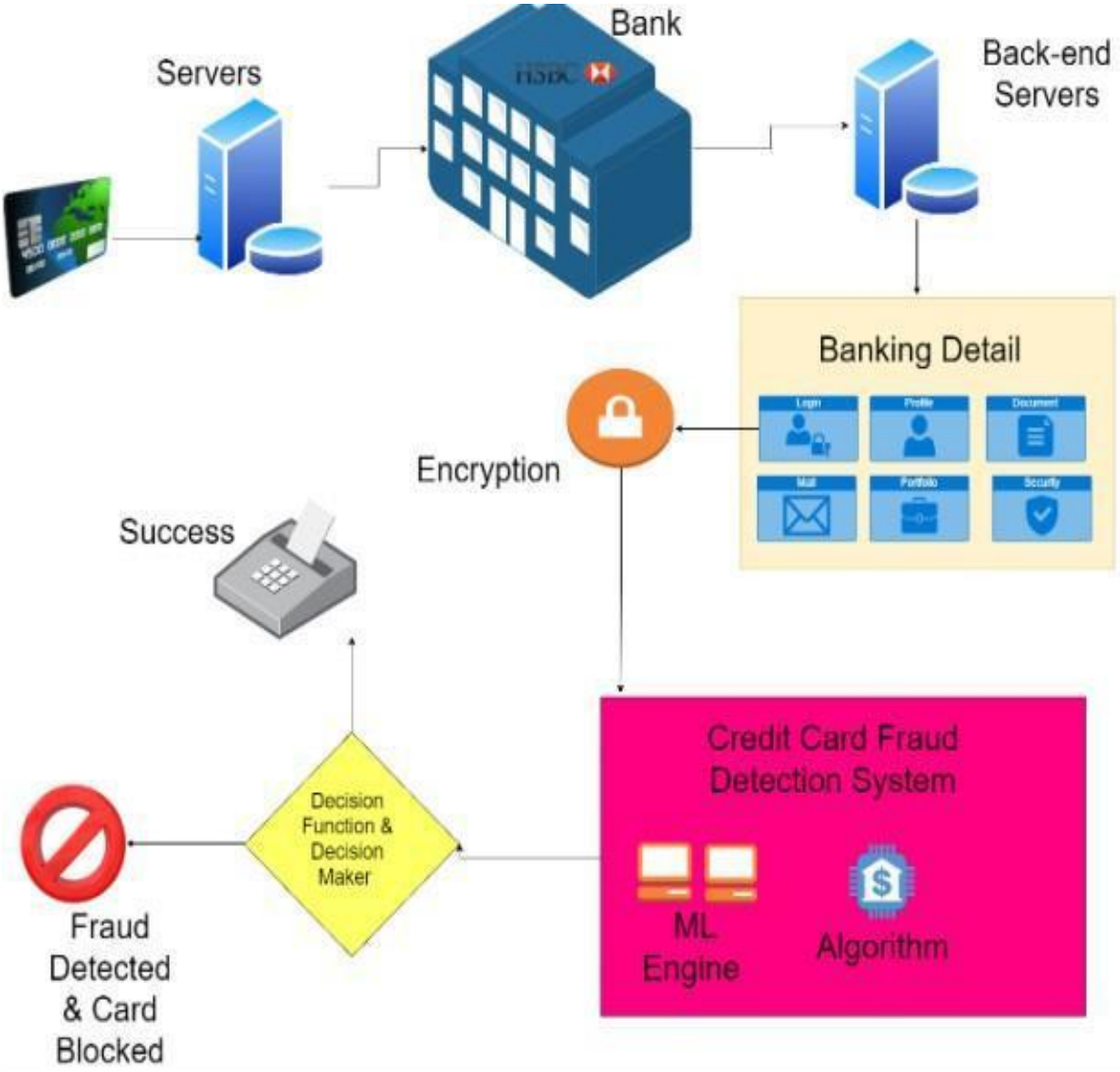



Isolation Forest Algorithm

Basic Architecture Diagram



Full Architecture Diagram



# ORGANIZATION

The principal stage includes stacking the dataset too known as the Data investigation stage. Information investigation is the cycle simply like information examination, we have utilized visual investigation to comprehend what is there in the dataset alongside its trademark. We have utilized the informational index from the Kaggle site, it contains different boundaries, for example, sum, class, time and others are decreased utilizing the PCA dimensionality decrease measure. The dataset is investigated and spoken to produces engaging insights that sum up the focal inclination, scattering, and type of a dataset's dissemination for the given arrangement object. All the counts are performed by barring Null qualities. By exploitation this spoke to information, the histogram is produced to show it.

The subsequent stage includes Data Preprocessing. It again stacks the dataset and eliminates all the invalid qualities and trash esteems from the dataset to improve its proficiency. In this stage itself, we need to split the dataset into preparing and testing stages. Here we generally take a shot at the preparation stage by depicting class 0 as veritable exchange and class 1 as a deceitful one. In preparing the dataset, deceitful and certifiable passages are given haphazardly to intensify the quality and subsequently more practical information is produced. A connection network is given to sum up information, as a contribution to a more progressed examination, and as a symptomatic for cutting edge examination.

Third and the last stage is Data grouping. It is just an assignment of contributing preparing informational index for which the classes are pre-marked for the calculation to learn from. The model is then utilized by contributing an alternate dataset for which the classes are not characterized and at that point the model predicts the class to which it takes after utilizing the gaining from the preparation set. Both the calculation must be applied to get a profitable outcome deciding the result utilizing terms as exactness, review, f1-score, and backing.

# LITERATURE SURVEY

Coercion cross possibly because the illegal or crook deceptive anticipated to gain monetary or singular favored position. It is an intentional display this is illicit, rule or manner with an association to gain unapproved cash associated little bit of leeway. Different composed works figuring out with variant from the norm or blackmail revelation on this area were circulated beginning at now and are open for public use. An big audit drove via way of means of Clifton Phua and his accomplices have found out that techniques used on this area fuse records mining applications, computerized distortion disclosure, badly organized acknowledgment. In some other paper, Suman, Research Scholar, GJUS&T at Hisar HCE offered technique like Supervised and Unsupervised Learning for Mastercard blackmail acknowledgment. In spite of the manner that those techniques and figurings were given an remarkable fulfillment in multiple regions, they fail to provide a long lasting and reliable reaction for coercion place. A comparable evaluation area turned into offered via way of means of Wen-Fang YU moreover, Na Wang in which they used Outlier mining, Outlier place mining and Distance sum estimations to appropriately are expecting bogus change in an impersonating evaluation of Mastercard change academic report of 1 positive commercial enterprise bank. Irregularity mining is a area of records mining which is largely utilized in financial and net fields. It oversees spotting items which are disengaged from the important gadget as an instance the trades that arent authentic. They have taken characteristics of clients direct and reliant at the evaluation of these credit theyve established that distance among the noticed evaluation of that characteristic and its destined worth. Uncommon techniques, as an instance, cream records mining/complicated affiliation request computation can see illegal occasions in a real card change academic assortment, considering affiliation multiplication figuring that offers making depictions of the deviation of 1 version from a reference percent have proven succesful constantly on medium predicted on line change.

There have likewise been endeavors to increase from a very new viewpoint. Endeavors had been made to enhance the alert comments communique if there must get up an prevalence of faux exchange. If there must be an prevalence of faux exchange, the permitted framework might be alarmed and an enter might be

shipped off deny the progressing exchange.

Fake Genetic Algorithm, one of the methodologies that shed new mild on this space, countered extortion from an change bearing. It confirmed precise in coming across the deceitful exchanges whats more, proscribing the amount of bogus cautions. Despite the truth that, it became joined with the aid of using characterization problem with variable misclassification costs.

# SYSTEM DEVELOPMENT

The most imperative job has been fought by the dataset. Since it is preposterous to expect to figure on a constant information base, we tend to are utilizing the creditcard.csv dataset from Kaggle which consolidates 2,83,807 sections. The different boundaries utilized in the information base are time, class, sum, area and, and so on Such sort of all out of 31 boundaries is utilized. V1 - V28 are the fields of a PCA dimensionality decrease to protect the character and delicate choices of a client.

In this venture, we are pointing towards assurance of exchanges with a high probability to tolerate Visa misrepresentation. We'll assemble and utilize the resulting two AI calculations:

•Local Outlier Factor (LOF)

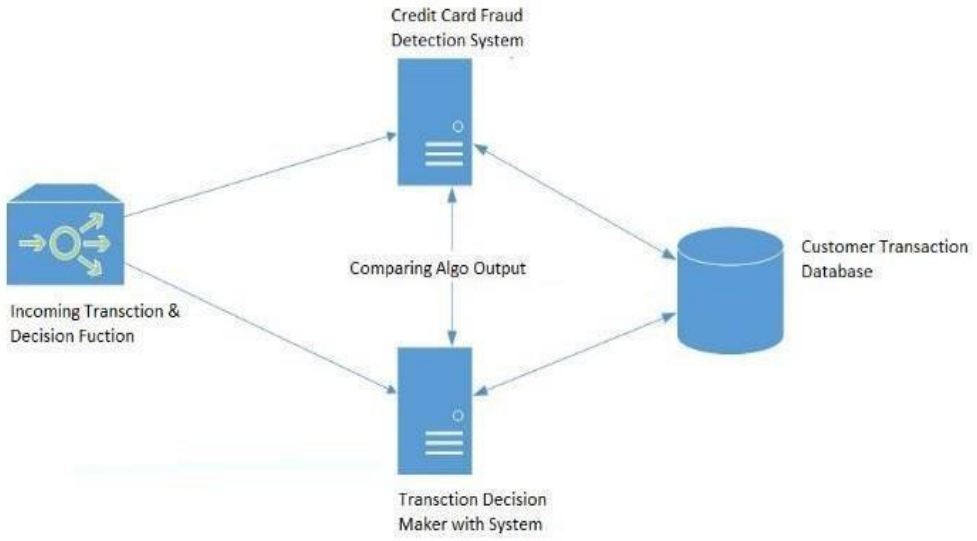
The irregularity rating of every instance is known as the close by Outlier factor. It computes the nearby deviation of the thickness of a given example comparable to its neighbors. It's known as nearby on the grounds that the peculiarity score relies upon the article disengaged the thing is with connection to the circling neighborhood.

•Isolation Forest calculation

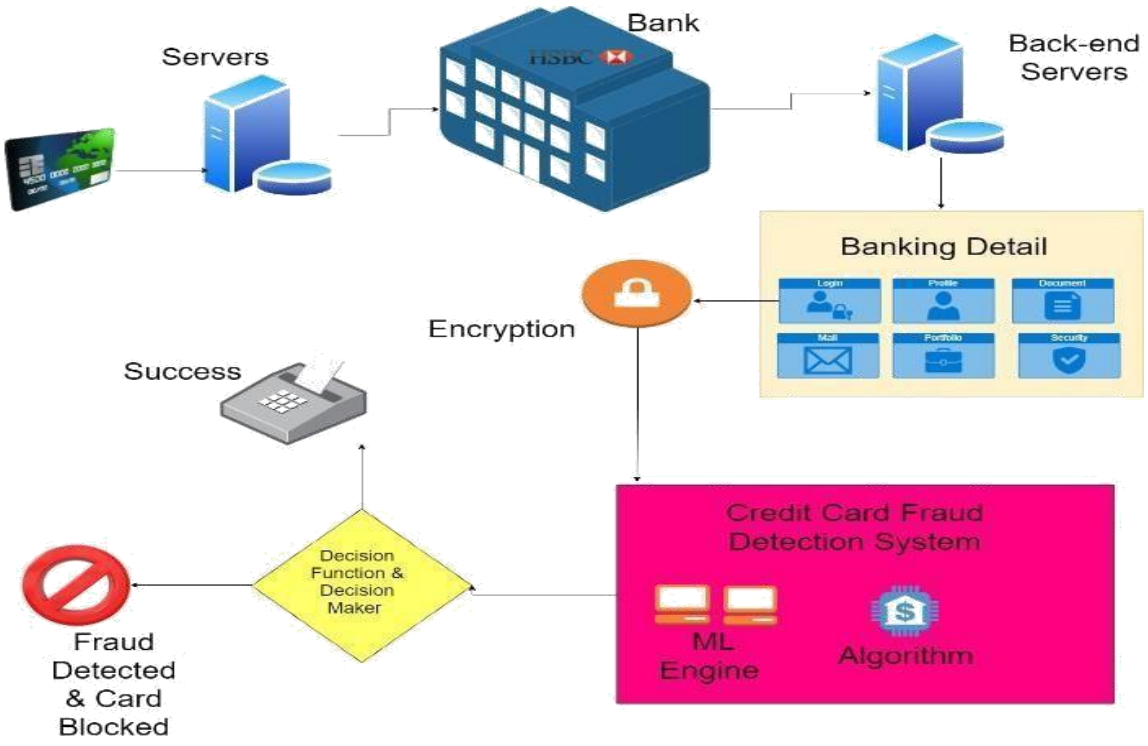
The Isolation Forest disconnects the perceptions by unpredictably picking a component by haphazardly choosing a worth and parting it in the middle of the atmost most extreme and least estimations of the picked include. Additionally the recursive parceling strategy is meant by a tree-like structure, the amount of parting needed to separate an example is similar to the path length from the root hub to the ending hub.

# Modeling & Analysis

The fundamental unpleasant design outline can be spoken to with the accompanying figure:



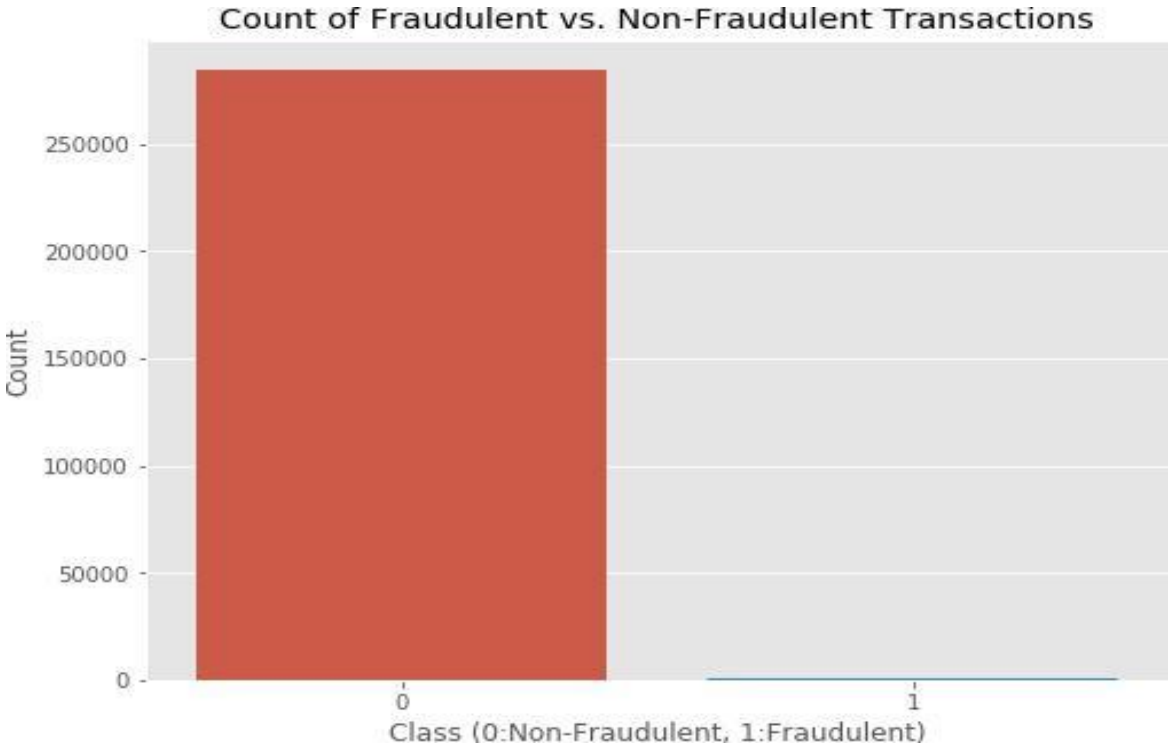

When taken a gander at in element for a larger scope along authentic components, the entire layout define may be spoken to as follows:First of all, we were given our dataset from Kaggle, an statistics research webweb page which offers datasets.



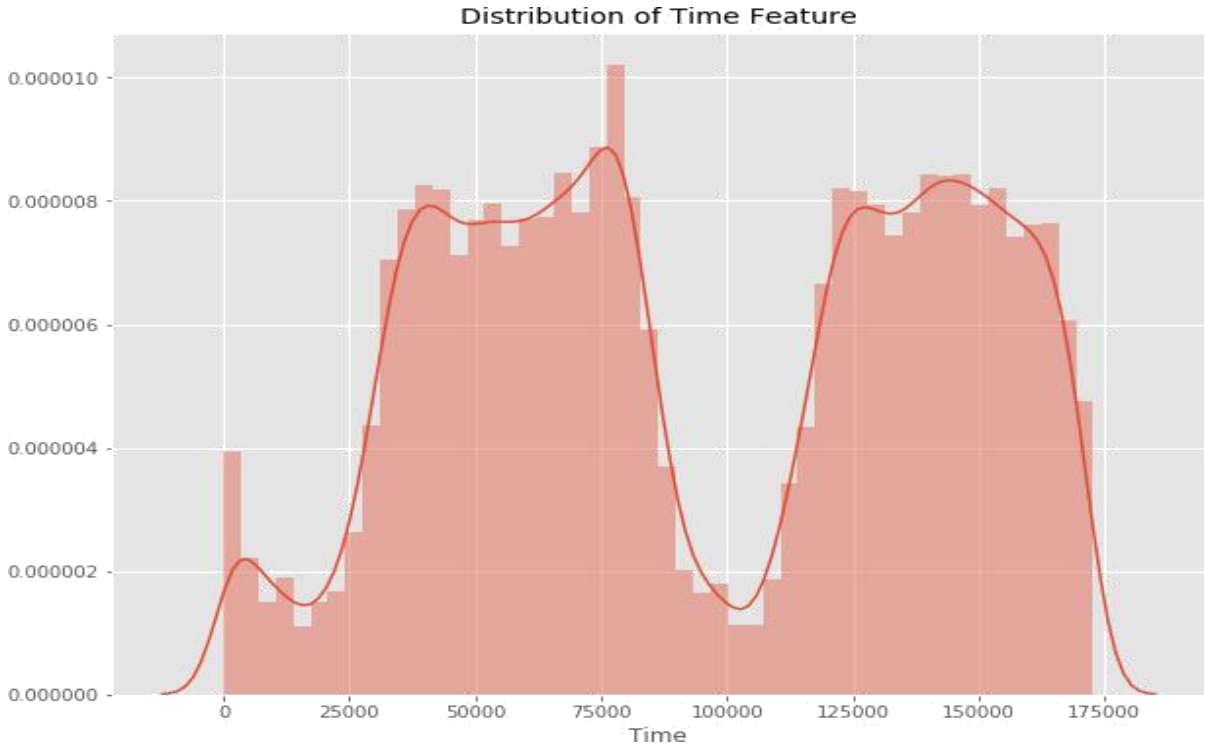

Inside this dataset, there are 31 segments out of which 28 are named as v1-v28 to make certain sensitive information.

Different sections communicate to Time, Amount and Class. Time suggests the postpone among the number one alternate and the accompanying one. Sum is the degree of coins executed. Class zero speaks to a valid alternate and 1 speaks to a fake one.

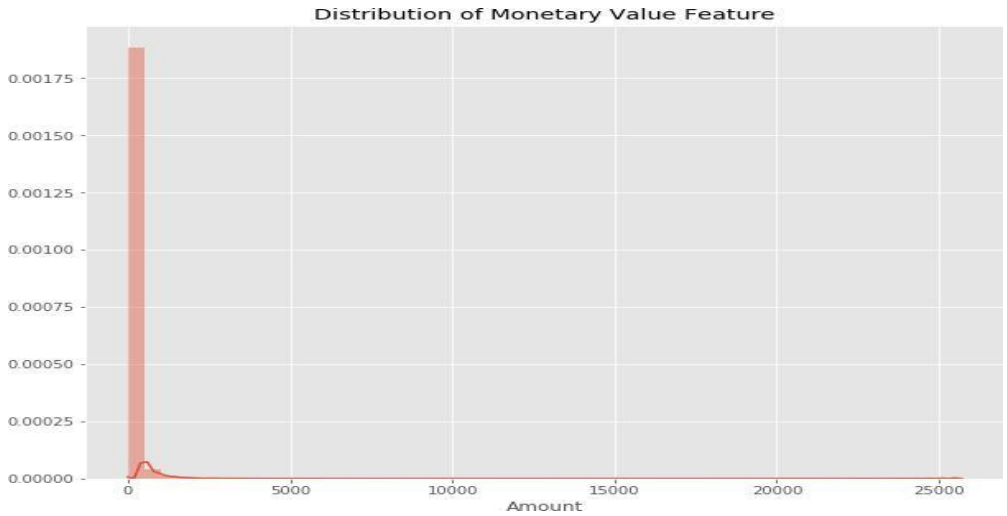
We plot various charts to check for irregularities in the dataset and to outwardly understand it:




This chart shows that the quantity of false exchanges is a lot of lower than the authentic ones.

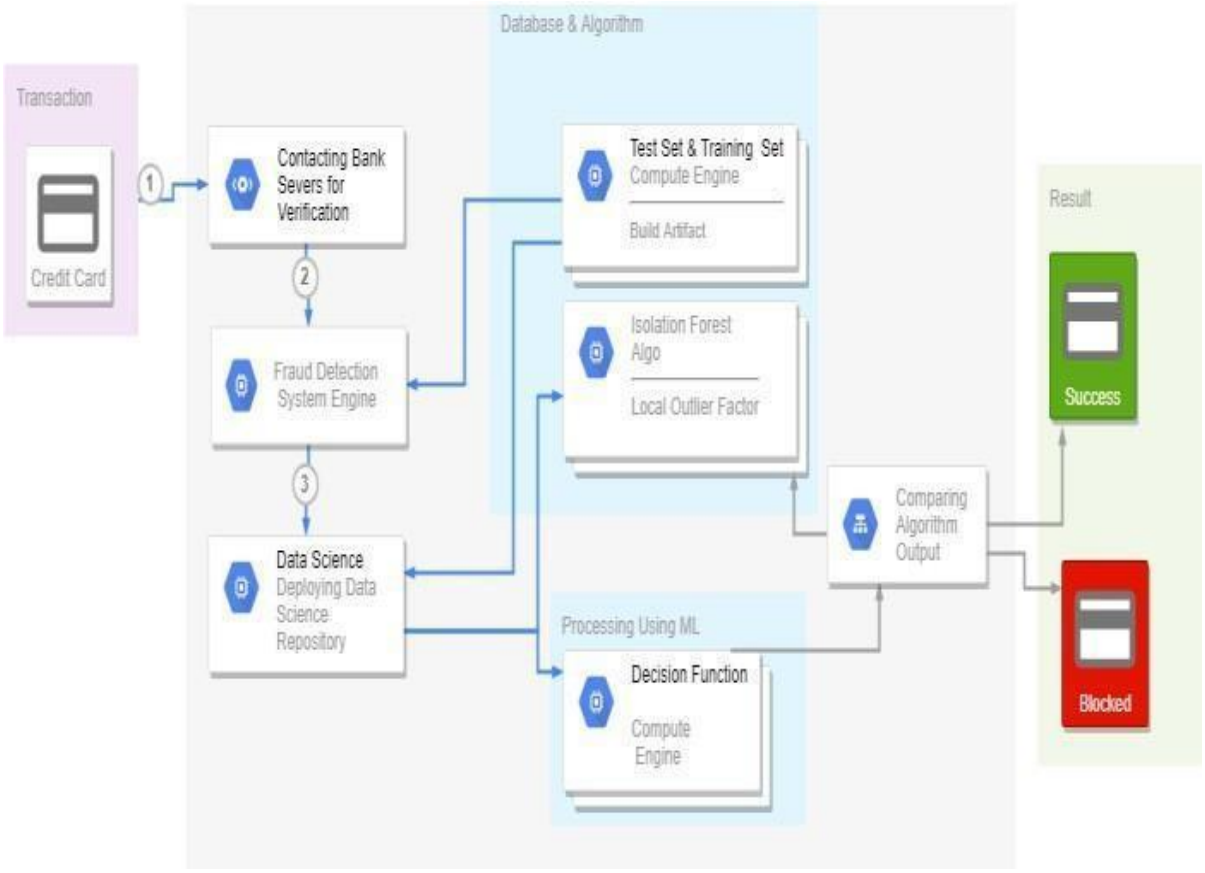



This diagram suggests the activities at which exchanges have been carried out interior days. It has a tendency to be visible that the maximum un-wide variety of exchanges have been made at some point of night and maximum noteworthy at some point of the days.

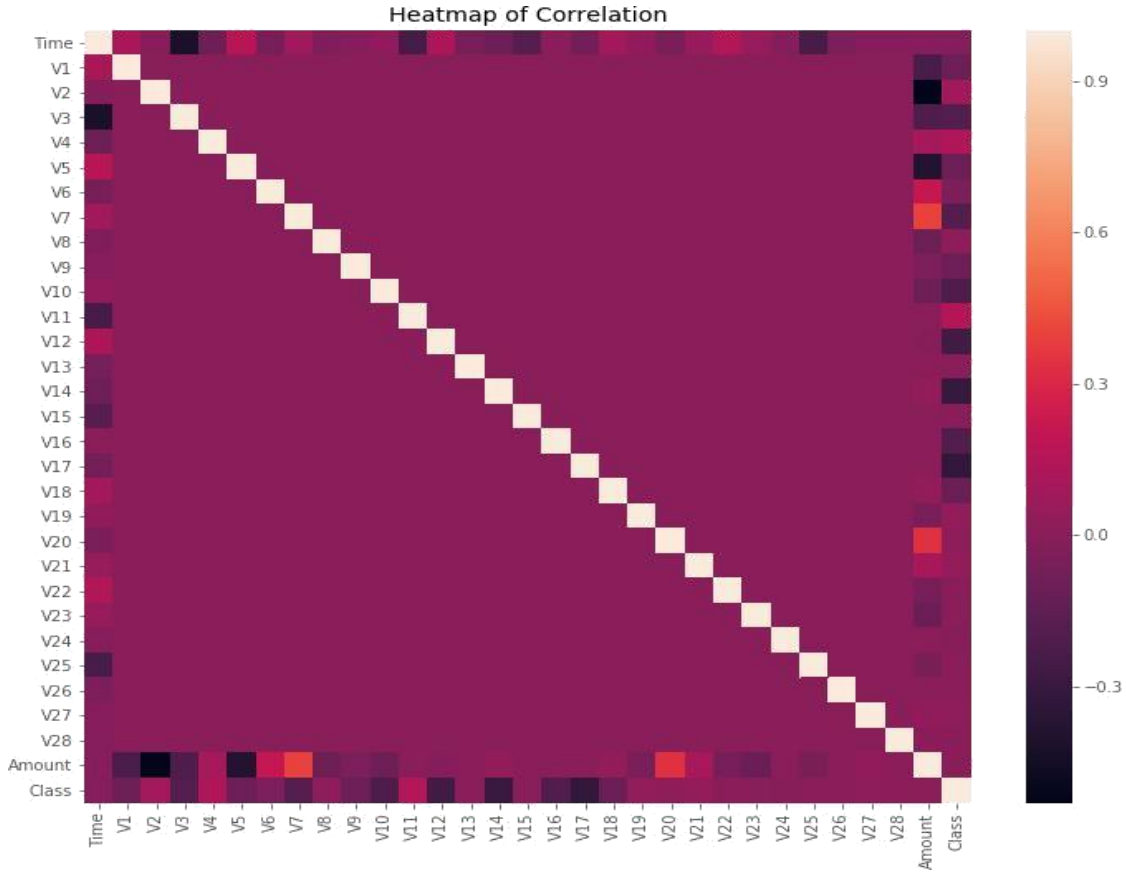



This chart speaks to the sum that became accomplished. A lions proportion of exchanges are reasonably little and only a small bunch of them method the best accomplished sum.

In the wake of checking this dataset, we plot a histogram for every segment. This is finished to get a graphical portrayal of the dataset which may be applied to verify that there aren't anyt any lacking any characteristics withinside the dataset. This is finished to assure that we dont want any lacking really well worth ascription and the AI calculations can degree the dataset easily.




After this investigation, we plot a heatmap to get a shaded portrayal of the facts and to ponder the connection betweens out looking forward to elements and the magnificence variable. This heatmap is confirmed as follows:




The dataset is currently prepared and prepared. The time and sum section are normalized and the Class segment is taken out to assure decency of assessment. The statistics is dealt with through a group of calculations from modules. The accompanying module chart clarifies how those calculations cooperate: This statistics is determined a manner right into a version and the accompanying anomaly identity modules are carried out on it:

• Local Outlier Factor

• Isolation Forest Algorithm

These calculations are a bit of sklearn. The outfit module withinside the sklearn package carries troupe primarily based totally techniques and capacities for the arrangement, relapse and exception location.

This unfastened and open-supply Python library is fabricated making use of NumPy, SciPy and matplotlib modules which offers a high-quality deal of fundamental and gifted gadgets which may be applied for statistics examinationfurthermore, AI. It highlights specific order, grouping and relapse calculations and is supposed to interoperate with the mathematical and logical libraries.

Weve applied Jupyter Notebook degree to make a application in Python to reveal the technique that this paper recommends. This application can likewise be performed at the cloud making use of Google Collab

degree which bolsters all python notice pad documents.

# PHASES OF PROJECT

We are completing this fraudulent transactions detection activity in following three phases,

1)Data Exploration

Steps: a) Load dataset

b) Preprocess dataset

c) Perform graphing

d) Display dataset

2)Data Preprocessing

Steps: a) Load dataset

b)Remove Null values

c)Split dataset

d)Move to training phase

3)Data Classification

Steps: a) Train the dataset

b)Develop classifier

c)Isolation Forest

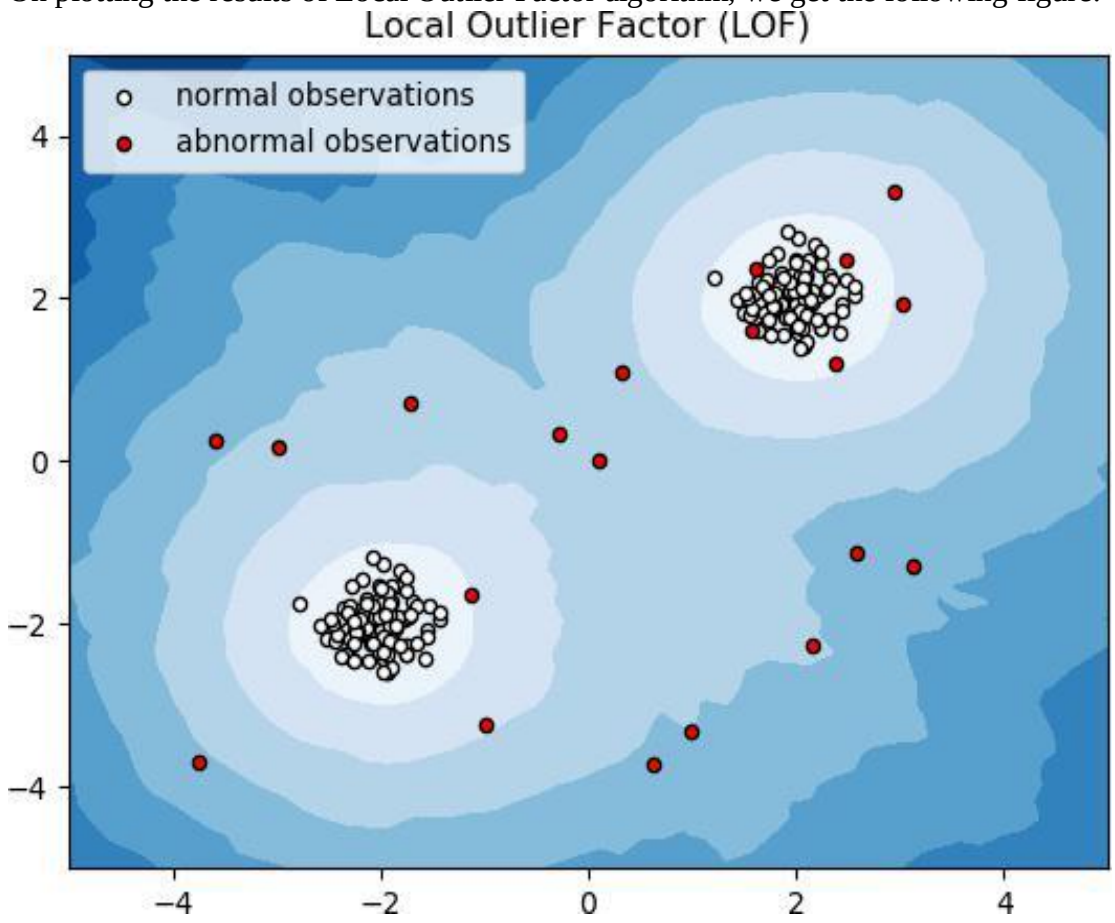
d)Perform Classification

1. Local Outlier Factor

It is an Unsupervised Outlier Detection calculation. Neighborhood Outlier Factor alludes to the oddity rating of every example. It quantifies the close by deviation of the instance records as for its neighbors. All the greater exactly, area is given through k-closest neighbors, whose distance is applied to gauge the close by records. The pseudocode for this calculation consists 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 searching on the close by estimations of an instance to that of its neighbors, you'll distinguish exams which can be drastically decrease than their neighbors. These traits are very amanous and they're taken into consideration as exceptions.

As the dataset is extraordinarily huge, we applied only a small quantity of it in out exams to reduce getting ready times. The eventual final results with the overall dataset dealt with is also determined and is given withinside the results section of this paper.

2. Isolation Forest

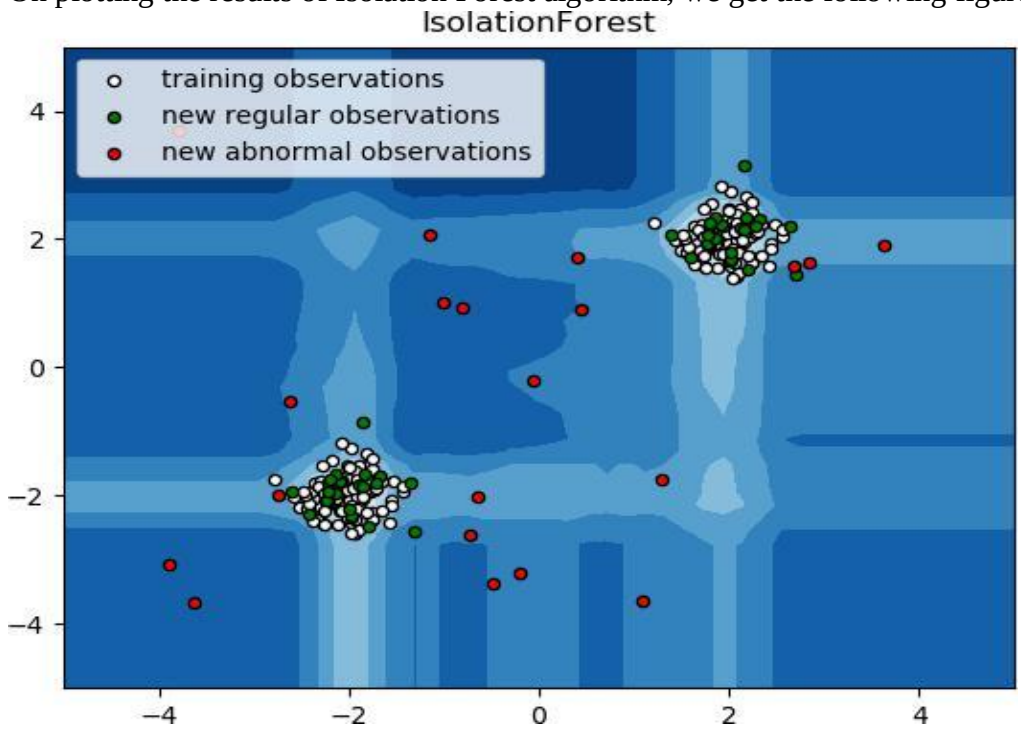
The Isolation Forest secludes perceptions through discretionarily selecting an detail and later on arbitrarily selecting a break up an incentive among the best and least estimations of the assigned detail.Recursive apportioning may be spoken to through a tree, the amount of components had to segregate an instance is

equal to the manner duration root hub to finishing hub.

The everyday of this manner duration offers a percentage of ordinariness and the selection ability which we use.The pseudocode for this calculation can be composed 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:




Parceling them arbitrarily creates greater restrained methods for inconsistencies. At the factor while a wooded area of abnormal timber normally creates greater restrained manner lengths for specific examples, they may be enormously at risk of be peculiarities.

When the irregularities are distinguished, the framework may be applied to record them to the worried specialists. For checking out purposes, we're contrasting the yields of those calculations with determine their exactness and accuracy.

# PERFORMANCE ANALYSIS

This concept is difficult to actualize, all matters taken into consideration, for the reason that it calls for the collaboration from banks, which arent glad to proportion facts due to their marketplace rivalry, and moreover due to lawful motives and safety of statistics in their clients. Subsequently, we regarded into a few reference papers which observed comparative methodologies and assembled results. As expressed in this type of reference papers:

"This technique changed into implemented to a complete software informational series supplied through a German financial institution in 2006. For banking secrecy motives, simply an define of the effects were given is added beneath. Subsequent to making use of this technique, the extent 1 rundown envelops more than one instances but with a excessive chance of being fraudsters.

All human beings referenced on this rundown had their playing cards close to avoid any risk due to their excessive-threat profile. The circumstance is extra thoughts boggling for the alternative rundown. The degree 2 rundown is as but constrained sufficiently to be minded a made to reserve premise. Credit and collection officers concept approximately that a huge a part of the instances on this rundown will be taken into consideration as doubtful faux conduct. For the closing rundown and the biggest, the paintings is impartially weighty. Not precisely 33% of them are doubtful.

To enhance the time talent and the overhead charges, a risk is to recall some other issue for the inquiry; this issue may be the 5 first digits of the phone numbers, the e-mail address, and the name of the game key, for example, the ones new inquiries may be implemented to the extent 2 rundown and degree three rundown".

# Dataset

Utilizing a dataset of virtually 284,807 Visa exchanges and one-of-a-kind unaided inconsistency region calculations, exchanges with a excessive probability of being Visa extortion are identified.The datasets includes trades made through Visas in September 2013 through ecu cardholders. This dataset offers trades that happened in days, in which we've got 492 cheats out of 284,807 trades.

The dataset is uncommonly inconsistent, the nice class (cheats) communicate to zero.172% of all trades.

It includes definitely numerical records elements which can be the final results of a PCA change. Tragically, in mild of thriller issues, we cant provide the most important capabilities and greater established order records approximately the data. Features V1, V2, … V28 are the focal sections were given with PCA, the number one capabilities that have now no longer been modified with PCA are Time and Entirety. Feature Time includes the seconds sneaked beyond among every change and the essential change withinside the dataset. The thing Whole is the change Amount, this thing may be used as an example dependant cost-senstive learning. Feature Class is the reaction variable and it takes regard 1 if there need to be an occasion of deception and zero regardless.

# IMPLEMENTATION

1.Importing Necessary Libraries

In [1]:
```
import sys
import numpy
import pandas
import matplotlib
import seaborn
import scipy

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__))
```

```
Python: 2.7.13 |Continuum Analytics, Inc.| (default, May 11 2017, 13:17:26) [MSC v.1500 64 bit (AMD64)]
Numpy: 1.14.0
Pandas: 0.21.0
Matplotlib: 2.1.0
Seaborn: 0.8.1
Scipy: 1.0.0
```

2. Data Set

In [3]:
```
# Load the dataset from the csv file using pandas
data = pd.read_csv('creditcard.csv')
```

In [4]:
```
# Start exploring the dataset
print(data.columns)
```

```
Index([u'Time', u'V1', u'V2', u'V3', u'V4', u'V5', u'V6', u'V7', u'V8', u'V9',
       u'V10', u'V11', u'V12', u'V13', u'V14', u'V15', u'V16', u'V17', u'V18',
       u'V19', u'V20', u'V21', u'V22', u'V23', u'V24', u'V25', u'V26', u'V27',
       u'V28', u'Amount', u'Class'],
      dtype='object')
```

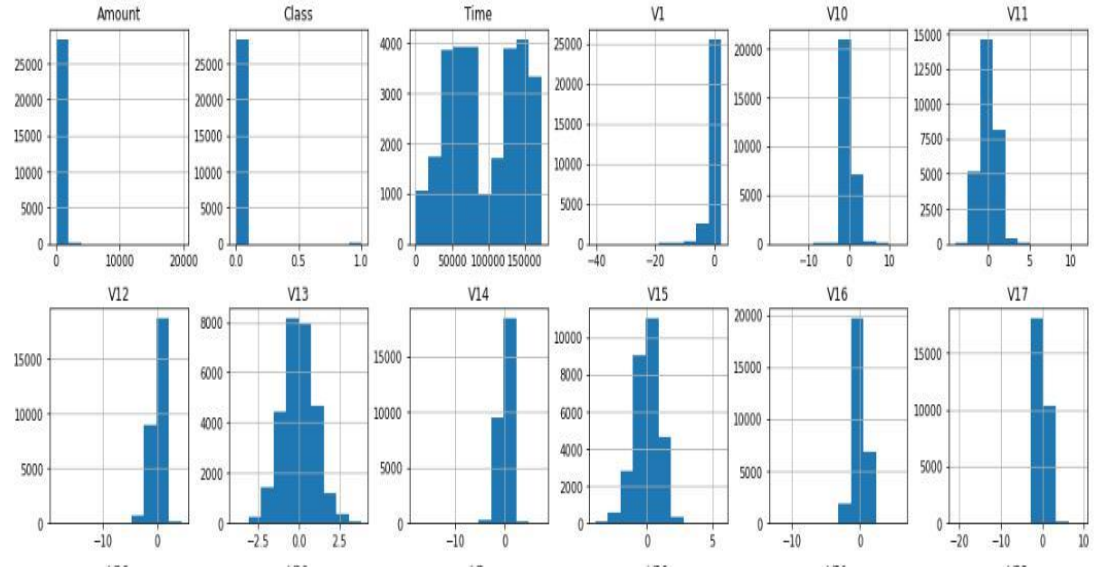
3. Dataset Exploration

```
In [5]: # Print the shape of the data
        data = data.sample(frac=0.1, random_state = 1)
        print(data.shape)
        print(data.describe())

        # V1 - V28 are the results of a PCA Dimensionality reduction to protect user identities and sensitive features
```

```
(28481, 31)
                Time            V1            V2            V3            V4  \
count   28481.000000  28481.000000  28481.000000  28481.000000  28481.000000
mean    94705.035216     -0.001143     -0.018290      0.000795      0.000350
std     47584.727034      1.994661      1.709050      1.522313      1.420003
min         0.000000    -40.470142    -63.344698    -31.813586     -5.266509
25%     53924.000000     -0.908809     -0.610322     -0.892884     -0.847370
50%     84551.000000      0.031139      0.051775      0.178943     -0.017692
75%    139392.000000      1.320048      0.792685      1.035197      0.737312
max    172784.000000      2.411499     17.418649      4.069865     16.715537
```

4. Plotting histograms of each parameter

```
In [6]: # Plot histograms of each parameter
        data.hist(figsize = (20, 20))
        plt.show()
```



5. Determining number of fraud cases in dataset

```
In [7]: # Determine number of fraud cases in dataset

        Fraud = data[data['Class'] == 1]
        Valid = data[data['Class'] == 0]

        outlier_fraction = len(Fraud)/float(len(Valid))
        print(outlier_fraction)

        print('Fraud Cases: {}'.format(len(data[data['Class'] == 1])))
        print('Valid Transactions: {}'.format(len(data[data['Class'] == 0])))
```

```
0.00172341024198
Fraud Cases: 49
Valid Transactions: 28432
```

6. Unsupervised outlier detection using Local Outlier Factor(LOF) and Isolation Forest Algorithm.

```
In [11]: from sklearn.metrics import classification_report, accuracy_score
         from sklearn.ensemble import IsolationForest
         from sklearn.neighbors import LocalOutlierFactor

         # define random states
         state = 1

         # define outlier detection tools to be compared
         classifiers = {
             "Isolation Forest": IsolationForest(max_samples=len(X),
                                                 contamination=outlier_fraction,
                                                 random_state=state),
             "Local Outlier Factor": LocalOutlierFactor(
                 n_neighbors=20,
                 contamination=outlier_fraction)}
```

In [15]:
```
# Fit the model
plt.figure(figsize=(9, 7))
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 prediction values to 0 for valid, 1 for fraud.
    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(Y, y_pred))
    print(classification_report(Y, y_pred))
```

7. Precision,recall and f1-score

```
Local Outlier Factor: 97
0.9965942207085425
              precision    recall  f1-score   support

           0       1.00      1.00      1.00     28432
           1       0.02      0.02      0.02        49

avg / total       1.00      1.00      1.00     28481

Isolation Forest: 71
0.99750711000316
              precision    recall  f1-score   support

           0       1.00      1.00      1.00     28432
           1       0.28      0.29      0.28        49

avg / total       1.00      1.00      1.00     28481
```

Results with the complete 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
```

# CONCLUSIONS

Credit card coercion is glaringly an exhibition of crook trickiness. This article has penetrated down the maximum high-quality methodologies for blackmail nearby their distinguishing evidence techniques and assessed progressing revelations on this field. This paper has further defined in detail, how AI may be implemented to enhance achieves distortion revelation nearby the figuring, pseudocode, rationalization its execution and experimentation results.

While the be counted number involves over 99.6% precision, its exactness stays precisely at 28% while a 10th of the instructive file is examined. Nevertheless, while the complete dataset is handled into the estimation, the precision rises to 33%. This expanded degree of precision is to be countless deliver of the massive lopsidedness among the quantity of genuine and variety of authentic exchanges.Since the complete dataset consists of most effective days alternate records, its solitary a modest amount of records that may be made to be had if this undertaking had been for use on a enterprise scale. Being set up on AI estimations, this system will without a doubt gather its functionality as time is going on as greater records is about into it.

# REFERENCES


- Jain, Yashvi, ShripriyaDubey NamrataTiwari, and Sarika Jain. "A comparative analysis of various credit card fraud detection techniques." *Int J Recent Technol Eng* 7.5S2 (2019): 402-407.

- John, Hyder, and Sameena Naaz. "Credit card fraud detection using local outlier factor and isolation forest." *Int. J. Comput. Sci. Eng.* 7 (2019): 1060-1064.

- Waspada, Indra, et al. "Performance Analysis of Isolation Forest Algorithm in Fraud Detection of Credit Card Transactions." *Khazanah Informatika: Jurnal Ilmu Komputer dan Informatika* 6.2 (2020).
- "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 2014

- “Research on Credit Card Fraud Detection Model Based on Distance Sum – by Wen-Fang YU and Na Wang”

# PR_Frauddetection(new).docx

ORIGINALITY REPORT

| 23% | 22% | 3% | 13% |
|---|---|---|---|
| SIMILARITY INDEX | INTERNET SOURCES | PUBLICATIONS | STUDENT PAPERS |

PRIMARY SOURCES

| | | |
|---|---|---|
| 1 | realpython.com<br>Internet Source | 5% |
| 2 | codewithharry.com<br>Internet Source | 3% |
| 3 | marswebsolutions.files.wordpress.com<br>Internet Source | 3% |
| 4 | files.gitter.im<br>Internet Source | 2% |
| 5 | stackoverflow.com<br>Internet Source | 2% |
| 6 | link.springer.com<br>Internet Source | 1% |
| 7 | www.cs.fsu.edu<br>Internet Source | 1% |
| 8 | www.coursehero.com<br>Internet Source | 1% |
| 9 | Submitted to University of Wisconsin System<br>Student Paper | 1% |

| | | |
|---|---|---|
| 10 | Prince Bose, Apurva Malpthak, Utkarsh Bansal, Ashish Harsola. "Digital assistant for the blind", 2017 2nd International Conference for Convergence in Technology (I2CT), 2017<br>Publication | 1% |
| 11 | www.geeksforgeeks.org<br>Internet Source | 1% |
| 12 | www.ir.juit.ac.in:8080<br>Internet Source | 1% |
| 13 | DalSpace.library.dal.ca<br>Internet Source | <1% |
| 14 | Submitted to Charotar University of Science And Technology<br>Student Paper | <1% |
| 15 | affiliation.oaasisbamu.org<br>Internet Source | <1% |
| 16 | Submitted to American University of the Middle East<br>Student Paper | <1% |
| 17 | Submitted to Vaasan yliopisto<br>Student Paper | <1% |
| 18 | www.rigginginnovations.com<br>Internet Source | <1% |
| 19 | sycet.org<br>Internet Source | <1% |