Credit Card Fraud Detection

Major project report submitted in partial fulfilment of the requirement for the degree of Bachelor of Technology

in

Computer Science and Engineering

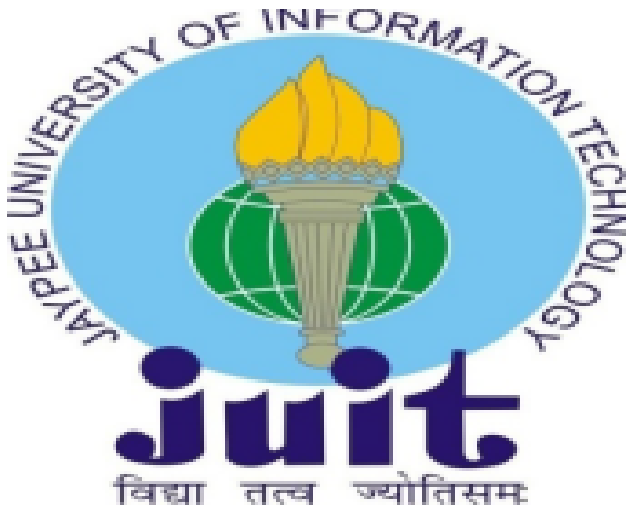
By

SAHIL DHIMAN (181226)

UNDER THE SUPERVISION

OF

Dr. Ravindara Bhatt



Department of Computer Science & Engineering and Information Technology,Jaypee University of Information Technology, Waknaghat, 173234, Himachal Pradesh, INDIA

TABLE OF CONTENT

I
DECLARATION

I hereby declare that this project has been done by me under the supervision of Dr. Ravindara Bhatt, Associate Professor Jaypee University of Information Technology. I also declare that neither this project nor any part of this project has been submitted elsewhere for award of any degree or diploma.

Supervised by:
Dr. Ravindara Bhatt
Associate Professor,
Department of Computer Science & Engineering and Information Technology
Jaypee University of Information Technology

Submitted by:
Sahil Dhiman (181226)
Computer Science & Engineering Department
Jaypee University of Information Technology

II
CERTIFICATE

This is to certify that the work which is being presented in the project report titled "Credit Card Fraud Detection" in partial fulfilment of the requirements for the award of the degree of B.Tech in Computer Science And Engineering and submitted to the Department of Computer Science And Engineering, Jaypee University of Information Technology, Waknaghat is an authentic record of work carried out by "Sahil Dhiman (181226)" under the supervision of Dr. Ravindara Bhatt, Department of Computer Science and Engineering, Jaypee University of Information Technology, Waknaghat.

Sahil Dhiman (181226)

The above statement made is correct to the best of my knowledge.

Dr. Ravindara Bhatt
Associate Professor,
Computer Science & Engineering and Information Technology
Jaypee University of Information Technology, Waknaghat,

III
ABSTRACT

This put forward a novel mechanism in which the payment is used to find credit card fraud detection.This methodology is known as the 'No Cash' which is a very famous application.There is not required for NFC i.e that is mobile based technique,which helps us to exceeding the cargo of taking the card outside.It is NFC based algorithm which is ver sufficient.It gives us the best information of every transaction.It extracts the normal and abnormal transaction in two different sections.it is a very good algorithm and it is highly efficient.It extracts the normal and abnormal transaction in two different sections.it is a very good algorithm and it is highly efficient.To hold the vast amount of data collected from online credit card card fraud detection and to predict he fraud cases in credit card detection model.

This give out the appraisal of the presentation of sampling techniques which they can find credit card fraud this is also known as the principal component analysis to find out result that is credit card fraud.It can be differanted by classifiers and performance which they can detect the normal transactions and abnormal transactions therefore with the help of random forest we can detect the credit card fraud.There is 29 principal components are used which they can find credit card fraud this is also known as the principal component analysis to find out result that is credit card fraud.There is very good adaptability is used which we can detect the credit card fraud.

CHAPTER 1:INTRODUCTION

1.1 INTRODUCTION

The online shopping trend is growing big and fast and with this various frauds are also growing and has become a major threat to government ,finance industries and the general population .The biggest of fraud in these online transactions is credit card fraud. So as the credit card transactions are becoming a major mode of payment , it is required to handle the problems through softwares which can be used to prevent frauds in these businesses.As data provided by google that over 30 billion US dollars worth is lost due the various kinds of frauds , in which 17.8 billion USD worth is lost due to the credit card frauds.credit card frauds comes under two categories inner card fraud which are done using false identity and external card fraud which is done using the stolen credit cards.The methods that the scammers use to get information regarding your credit card details are Phishing in which you receive a link that would seem like legit website, Skimming in this the scammers use a device called “skimmer” which captures the credit card information, Counterfeit cards are stolen cards which they hit and try until they found a card which is not already blocked , Dumpster dive is also same and in vishing scammer makes personal calls pretending to officials from banks and ask for otp and cvv and other critical information.

There are many techniques that can be used to solve credit card fraud detection.But Isolation Forest Algorithm is most effective because it has very high accuracy rate when compared to other techniques.Fraud detection is basically the process of separating the transactions in two different types of classes , a fraudulent transaction class and a legit transaction class.Although these two classes tends to be similar ,But they are different in many ways for example they vary in basic elements like time , amount , frequency and location.And these are the variable that are used in this detection technique because the credit card fraud detection technique can greatly affected by type of dataset used and the type of variable used . So these variables help our model get the results.From the experiment the result has been concluded that Isolation forest which has accuracy of 99.75 and 30 % more effective than other Machine learning techniques is best for our credit card fraud detection system.

1.2 PROBLEM STATEMENT

The one the problem in credit card system is that if anyone got their hands on information regarding your credit card can make fraud transaction without even your confirmation.And mostly when a fraudulent transaction takes place user is notified after the transaction is done which also major downside.Adding to that there is no mechanism to check that the upcoming transaction is fraudulent or not. In this project we will try to solve all of these problems.

1.3 OBJECTIVES

There are some proposed methods to develop a mechanism to determine that the upcoming transaction is fraud or not.The fraud transaction will recognized with the help of location where the transaction took place,Frequency the interval of the time between two transactions, Amount what was the amount that was withdrawn from the transaction.And the comparison of different Machine Learning algorithms will be shown.The figure below shows the overall system framework.

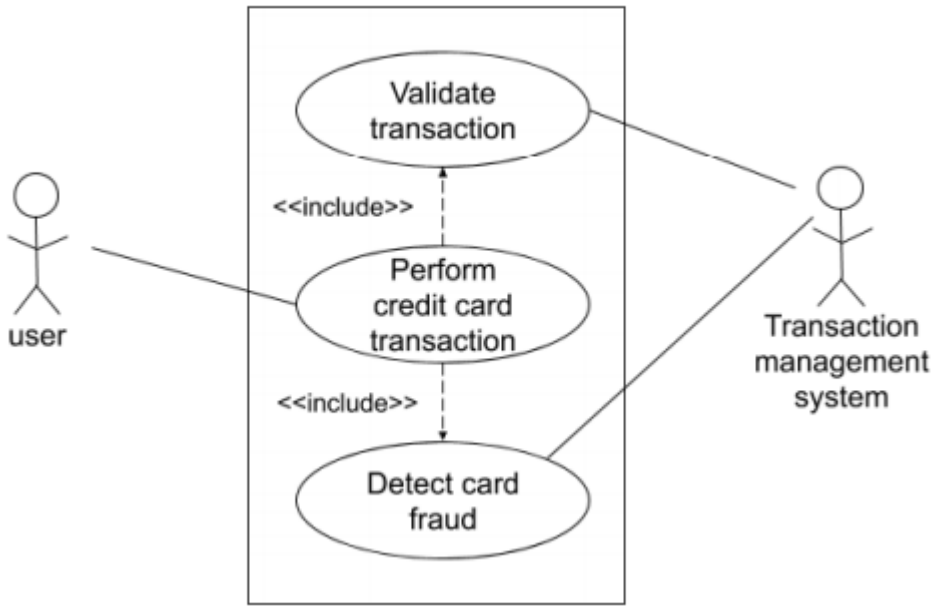


Fig 1.1

CHAPTER 2: LITERATURE SURVEY

K.Randhawa put forward a technique using machine learning to detect the fraud of the credit card. This project is known as the credit card fraud detection.To begin with,standard models were used after that hybrid model.Then,used the data set which is analyzed by various financial experts which can easily detect the fraud.With the this type of method we can easily achieve our goal with very good accuracy in order to detect the fraud of credit card with very good accurateness.But, the exactness the quality of being accurate is low as compared to other models.

A.Roy is used deep learning methodology to predict the for the detection of credit card fraud.large classification over fraud cases.Credit card fraud detection is unbalanced data which gives less accuracy.It cannot predict the hidden patterns.To hold the vast amount of data collected from online credit card card fraud detection and to predict he fraud cases in future.So,there is need to improve the very good accurateness for credit card fraud detection model.

S.Xuan is used machine learning and deep learning concept together and they used two types of random forest that can detect the actions of abnormal transactions and normal transactions.It can be differanted by classifiers and performance which they can detect the normal transactions and abnormal transactions therefore with the help of random forest we can detect the credit card fraud.

Z.Kazemi uses the methodology of deep learning to differentiate the normal transactions and abnormal transactions.It gives us the best information of every transaction.It extracts the normal and abnormal transaction in two different sections.it is a very good algorithm and it is highly efficient.

S.Dhankar is used supervised machine learning algorithm is used to extract the best characteristics of normal and abnormal transactions.This project is known as the credit card fraud detection.To begin with,standard models were used after that hybrid mode.To hold the vast amount of data collected from online credit card card fraud detection and to predict he fraud cases in future.So,there is need to improve the very good accurateness for credit card fraud detection model.

J.O.Awomeyi used the KNN better to detect the credit card fraud and after the use of this KNN algorithm then they use the Bayes logistic regression result.Raw and unprocessed data they use the different techniques of python.It can be differanted by classifiers and performance which they can detect the normal transactions and abnormal transactions therefore with the help of random forest we can detect the credit card fraud.It gives us the best information of every transaction.It extracts the normal and abnormal transaction in

two different sections.it is a very good algorithm and it is highly efficient.

S.Dutta in there research tell us that this is very rare then it is very hard to detect then they face very uncertain issue.they used Communal Detection and Spike Detection which initiate novel layers.It cannot predict the hidden patterns.To hold the vast amount of data collected from online credit card card fraud detection and to predict he fraud cases in future.So,there is need to improve the very good accurateness for credit card fraud detection model.The cd and sd algorithm used to determine the credit card fraud which give us very good accuracy.there is very good adaptability is used which we can detect the credit card fraud.

K.Modi is uses various techniques that help to find credit card fraud.This model have to increment some more feature for best result i.e best accuracy.they use comparative method to find the credit card fraud detection.they can compare normal and abnormal transaction then find credit card fraud.They use comparative analysis to find credit card fraud.So,there is need to improve the very good accurateness for credit card fraud detection model.It is very hard method to find that fraud with comparative analysis.

D.Pojee then put forward a novel mechanism in which the payment is used to find credit card fraud detection.This methodology is known as the 'No Cash' which is a very famous application.There is not required for NFC i.e that is mobile based technique,which helps us to exceeding the cargo of taking the card outside.It is NFC based algorithm which is ver sufficient.It gives us the best information of every transaction.It extracts the normal and abnormal transaction in two different sections.it is a very good algorithm and it is highly efficient.It extracts the normal and abnormal transaction in two different sections.it is a very good algorithm and it is highly efficient.To hold the vast amount of data collected from online credit card card fraud detection and to predict he fraud cases in credit card detection model.

D.Singh Sisodia give out the appraisal of the presentation of sampling techniques which they can find credit card fraud this is also known as the principal component analysis to find out result that is credit card fraud.It can be differanted by classifiers and performance which they can detect the normal transactions and abnormal transactions therefore with the help of random forest we can detect the credit card fraud.There is 29 principal components are used which they can find credit card fraud this is also known as the principal component analysis to find out result that is credit card fraud.There is very good adaptability is used which we can detect the credit card fraud.

L.Vergara put forward the graph theory in which with the help curves you can easily differentiate normal transaction and abnormal transaction.They can compare normal and abnormal transaction then find credit card fraud.They use comparative analysis to find credit card fraud.It gives us the best information of every transaction.It extracts the normal and abnormal transaction in two different sections.it is a very good algorithm and it is highly efficient.These techniques is used in various financial businesses it is very accurate with the help of comparing the curves of graphs we can find the credit card fraud very easily.

2.1 TABLE OF COMPARISON DIFFERENT METHODOLOGY

Writer's name	Methodology	AdvantagesDisadvantages /Upgradation
K.Randhaawa, C.Lo,M.Sarena, C.P.Linn,Ashok K	They are using high level ma learning to detect the credit fraud detection.	It gives us very good Credibility,factualnes s,correctness at this rate we find very good results in credit card fraud detection. more noise than v more improvement accuracy. Only achieve 80 percent of accuracy rate with the noise level of only 20 percent but in the real world th think so we need to find the best
A.Rooy,J.Sunn, R.Mohneey, P.Belleing,S.Addar L.Alfonzi	They are using deep learning methodology to detect the card fraud detection.	It can very easily to perform or we can detect the credit card fraud detection during online transactions of credit card. We need to improve accuracy of the these project .

G. Liu, Z. Li, Lutao S.Wang	They are using B2C datas detect the credit card fraud.	They are using Data they are using random forest which is very unequal in holds good during structure dataset small datasets. needs to improve noise level.B2C are very stable but they need more stable data than for solution.
Z.Kazemi	Deep learning is used in this r which they can very differentiate on the basis characteristics and features.	Classification is Number of done very easily on variations seems to basis of very less. characteristics and features which we can very differentiate.
J.O.Awomeyi	The performance of various methodology they are using ve stable dataset.	They are using We can not get logistics regression result accurately we and KNN classifiers only get some it gives them the information. best result.

Writer's name	Methodology	Advantages Disadvantages /Upgradation
S.dutta	They are using(CD) communal detection to detect the credit ca fraud.they are also using (SD)s detection for better results.	CD and SD gives us The results are not very good results shown in this project during the properly it can not implementation of give us accuracy this algorithm.they that we want in real

		are using a high number of parameters.
K. Modi,D.Pojee D.Singh Sisodia,L.Vergara	They are using various methods to protect the processing of algorithm.they are comparative analysis. There is 29 principal componer are used which they can find credi fraud this is also known as principal component analysis.	This methodology is The execution time known as the 'No of this model is Cash' which is a excellent but very famous accuracy of this application.There is model is not gives not required for NFC us expected re i.e that is mobile based is NFC based algorithm which is very sufficient There is very good These techniques is used in various adaptability is used financial businesses which we can detect it is very accurate the credit card fraud. with the help of comparing the curves of graphs we can find the credit card fraud very easily.

2.2 LITERATURE REVIEW OF DIFFERENT METHODOLOGY

We can use many methods to detect credit card fraud. We have to differentiate between normal transactions and abnormal transactions. We find in our data set that credit card fraud is very rare.So it is very hard to detect credit card fraud. we use in this credit card fraud detection model machine learning (ML),Artificial intelligence techniques we can differentiate normal transaction and abnormal transaction.Credit card fraud detection is classified with the help of supervised and unsupervised learning.we can use various techniques as well as supervised learning and unsupervised learning to find the credit card fraud detection but we can only find few of them give us very good accuracy.With the help of KNN,Logistic regression,vector machines help us find credit card fraud detection.with the help of supervised learning and unsupervised learning we can easily find abnormal transaction which help us to detect the credit card fraud detection.Credit card fraud detection is classified with the help of supervised and unsupervised learning.we can use various techniques as well as supervised learning and unsupervised learning to find the credit card fraud detection but we can only find few of them give us very good accuracy.In supervised learning we use different techniques like SVM(support vector machine),ANN(Artificial Neural Network) etc.to detect the credit card fraud detection.In unsupervised learning ASI,Fuzzy system etc. to detect the credit card fraud detection.With the help of SVM(support vector machine),ANN(Artificial Neural Network) help us find credit card fraud detection.with the help of supervised learning and unsupervised learning we can easily find abnormal transaction which help us to detect the credit card fraud detection.With the help of ASI,Fuzzy system,HMM help us find credit card fraud detection.with the help of supervised learning and unsupervised learning we can easily find abnormal transaction which help us to detect the credit card fraud detection.We use SVM in credit card fraud detection because it gives us the best result with unequal structure. It gives us the very best result to find the credit card fraud detection.Credit card fraud detection is classified with the help of supervised and unsupervised learning.we can use various techniques as well as supervised learning and unsupervised learning to find the credit card fraud detection but we can only find few of them give us very good accuracy..Using KNN classifier it is very easily handle the noise level of the structure which we can very easily differentiate between normal and abnormal transaction.we can use various techniques as well as supervised learning and unsupervised learning to find the credit card fraud detection but we can only find few of them give us very good accuracy.Only achieve 80 percent of accuracy rate with the noise level of only 20 percent but in the real world there is more noise than we think so we need more improvement to find the best accuracy.We can use many methods to detect credit card fraud.We have to differentiate between normal transactions and abnormal transactions.

The performance of various methodology they are using very stable dataset.They are using logistics regression and KNN classifiers it gives them the best result.We can not get result

accurately we only get some information.Raw and unprocessed data they use the different techniques of python.It can be differanted by classifiers and performance which they can detect the normal transactions and abnormal transactions therefore with the help of random forest we can detect the credit card fraud.It gives us the best information of every transaction.It extracts the normal and abnormal transaction in two different sections.it is a very good algorithm and it is highly efficient.S.dutta describe in this project that they are using they are using(CD) communal detection to detect the credit card fraud.they are also using (SD)spike detection for better results.S.Dutta in there research tell us that this is very rare then it is very hard to detect then they face very uncertain issue.they used Communal Detection and Spike Detection which initiate novel layers.It cannot predict the hidden patterns.To hold the vast amount of data collected from online credit card card fraud detection and to predict he fraud cases in future.So,there is need to improve the very good accurateness for credit card fraud detection model.The cd and sd algorithm used to determine the credit card fraud which give us very good accuracy.there is very good adaptability is used which we can detect the credit card fraud.The results are not shown in this project properly it can not give us accuracy that we want in real world.

KNN classifier also gives us very good noise handling in the data structure. After the noise handling we can use classification and regression in the dataset which we can very easily detect the credit card fraud detection.Only achieve 85 percent of accuracy rate with the noise level of only 40 percent but in the real world there is more noise than we think so we need more improvement to find the best accuracy.Using KNN classifier it is very easily handle the noise level of the structure which we can very easily differentiate between normal and abnormal transaction.Classification of the unequal structure is very hard to detect then we have to apply regression also so we can easily detect the credit card fraud detection.The selection on parameter in the dataset is very hard and we can not find easily animal point in the dataset.these are difficulties that's why we use KNN classifiers which can very easily detect the credit card fraud detection.

What other Data Scientists got

Method Used	Frauds	Genuines	MCC
Naïve Bayes	83.130	97.730	0.219
Decision Tree	81.098	99.951	0.775
Random Forest	42.683	99.988	0.604
Gradient Boosted Tree	81.098	99.936	0.746
Decision Stump	66.870	99.963	0.711
Random Tree	32.520	99.982	0.497
Deep Learning	81.504	99.956	0.787
Neural Network	82.317	99.966	0.812
Multi Layer Perceptron	80.894	99.966	0.806
Linear Regression	54.065	99.985	0.683
Logistic Regression	79.065	99.962	0.786
Support Vector Machine	79.878	99.972	0.813

Fig 2.1

CHAPTER 3:METHODOLOGY AND SYSTEM DEVELOPMENT

3.1 METHODS AND RESULTS

For the purpose of this project, distinctive types of algorithms have been selected. Following are the different kinds of Anomaly Detection algorithms and techniques that were used for this project:

1. Classification Based – One Class SVM, Random Forest Classifier, Isolation Forest

2. Nearest Neighbor Based – KNN and Local Outlier Factor (LOF)

3. Statistical Based – the use of Mahalanobis distance and Log Likelihood approach

4. Transduction Based – StrOUD Algorithm using KNN and LOF as strangeness function

5. Graph Based – Convex Hull method

6. Resampling the Data

Since the original facts set is sorted with the aid of the 'Time' attribute, for the reason of this project, the records set was shuffled (randomly) and saved as a separate CSV file. Once the dataset is shuffled, 80% of data was used for training and move validation (4-fold cross validation) and final 20% of the statistics was used for testing. For the purpose of this task (6), resampling of facts followed by way of trying out on famous algorithms (like KNN, Naïve Bayes, Decision Trees, Neural Networks and SVM) has been finished in Weka 3.8. And the rest of the mission (1 to 5), all the distinct type of algorithms and techniques, have been applied using Python 2.7. Popular libraries like numpy, sklearn, scipy, matplotlib, time, csv have been used. Initially all the algorithms and strategies have been tested on 10% of records (randomly chosen facts factors but also keeping in thought of ratio the anomalous to non-anomalous information factors i.e, 0.001728 or 0.1728%). Once that used to be done, algorithms have been run on entire (original) data set. Except Convex Hull and Chi Square Test method, all the implementations gave nearly the equal results. In some case, even better results than before. The source code archives comprise code for training and move validation followed by means of testing.

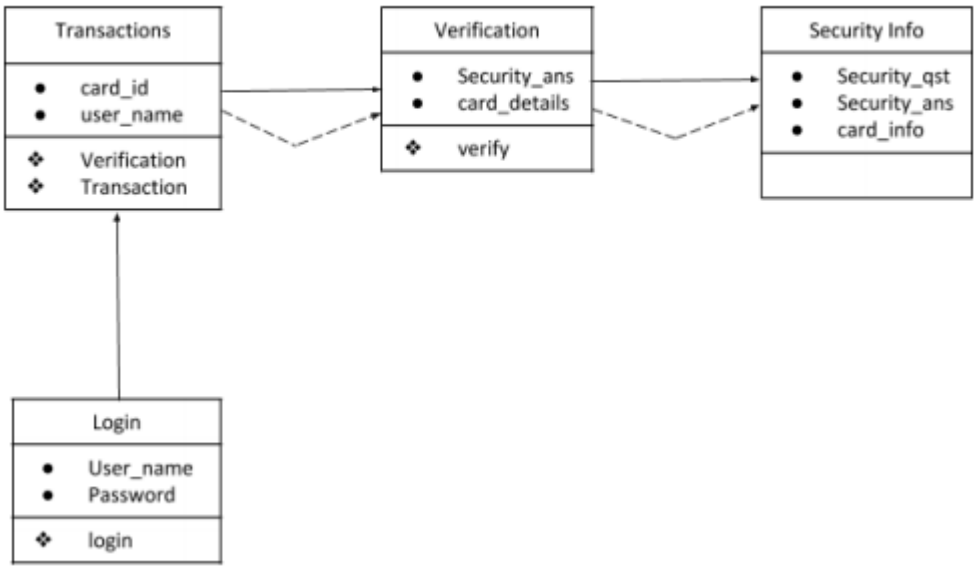


Fig3.1

1. CLASSIFICATION BASED

These type of Anomaly Detection algorithms can be supervised as well as unsupervised. But since number of anomaly category in check set is unbounded and education may lead to overfitting. General classification methods:

• One Class SVM

• Neural Networks Based – Auto Encoders

• Forest Based Classifiers – Random Forest Classifier and Isolation Forest Classifier

• Bayesian Network Approaches

For this project, One Class SVM, Random Forest Classifier and Isolation Forest were chosen as the Classification Based Anomaly Detection algorithm. ONE-CLASS SVM Main concept is to separate the entire set of education information from the origin, i.e. to discover a small region where most of the facts lies and label records factors in this region as one class. Following consequences are obtained by applying the algorithm on test information set. Parameter ‘nu’ used to be set to 0.05 and RBF kernel used to be used.

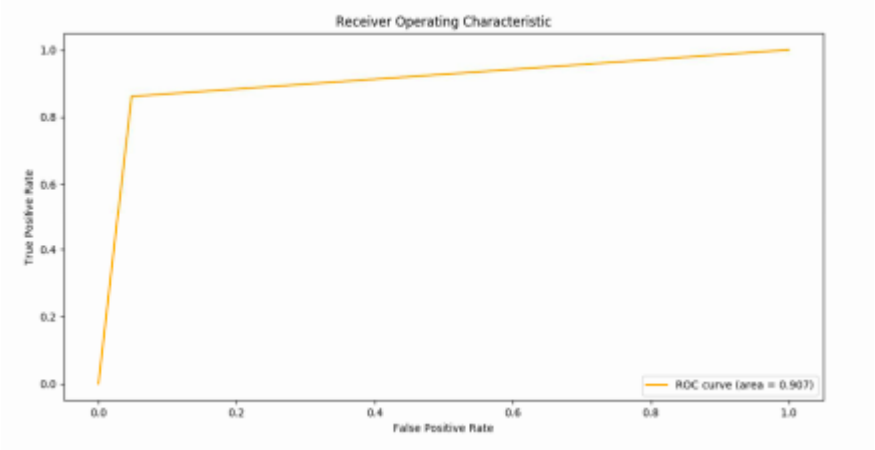


Fig3.2

RANDOM FOREST CLASSIFIER

A random wooded area classifier fits a wide variety of choice tree classifiers on various sub-samples of the dataset and use averaging to enhance the predictive accuracy and manage over-fitting. This classifier is very popular to handle unbalanced classes. Following consequences are acquired by means of making use of the algorithm on test data set. All the points (maximum elements = 30) have been used for this method.

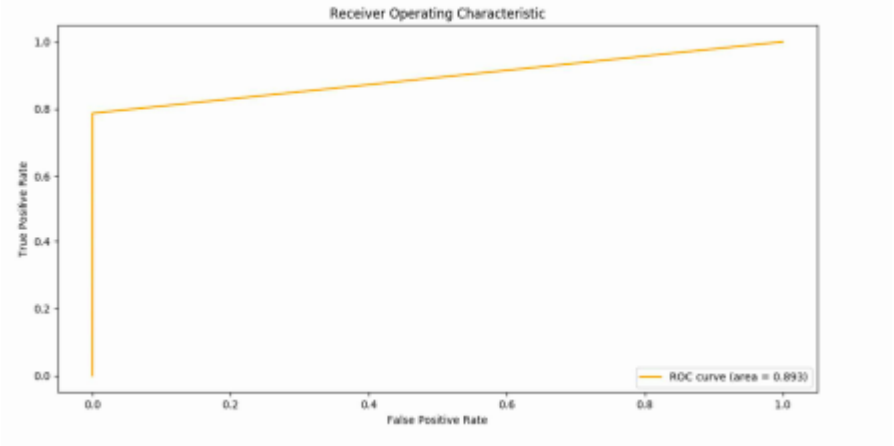


Fig3.3

ISOLATION FOREST

This the algo that is prone to a mechanism that find different Anomaly called isolation. Following results are obtained by way of making use of the algorithm on test statistics set. Parameter ‘contamination’ was once set to 0.1 for this method.

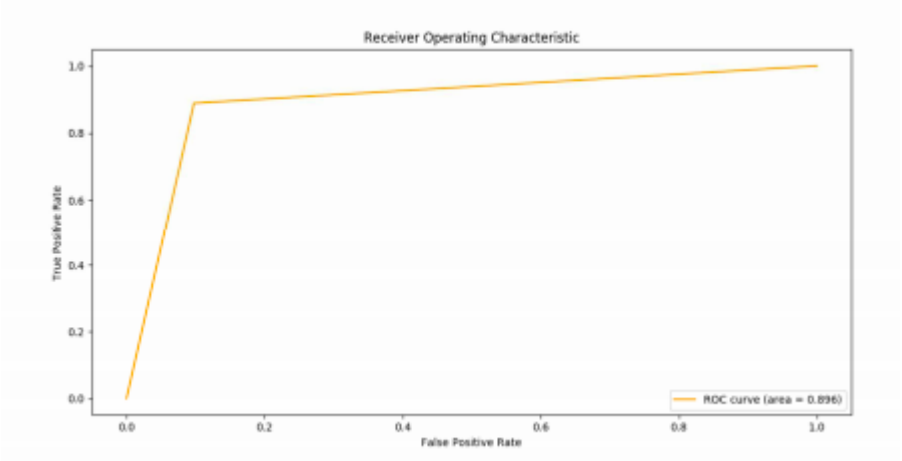


Fig3.4

2. NEAREST NEIGHBOR BASED

These sorts of Anomaly Detection Algorithm methods have a simple two-step approach:

• Compute regional for every facts record

• Analyze the nearby to decide whether records report is anomaly or not.

They can be further labeled as:

• Distance Based – KNN algorithm

• Density Based – LOF, COF, LOCI, etc.

For this project, KNN and LOF were chosen as Nearest Neighborhood Based Anomaly Detection

Algorithms.

K-NEAREST NEIGHBOUR BASED APPROACH

It's slightly different from the KNN classification approach. It can be explained in the following points:

• For every records point, compute the distance to the Kth nearest neighbor.

• Sort all information points according to the calculated distance.

• Outliers are factors that have the largest calculated distance and consequently are placed in the more sparse neighborhoods.

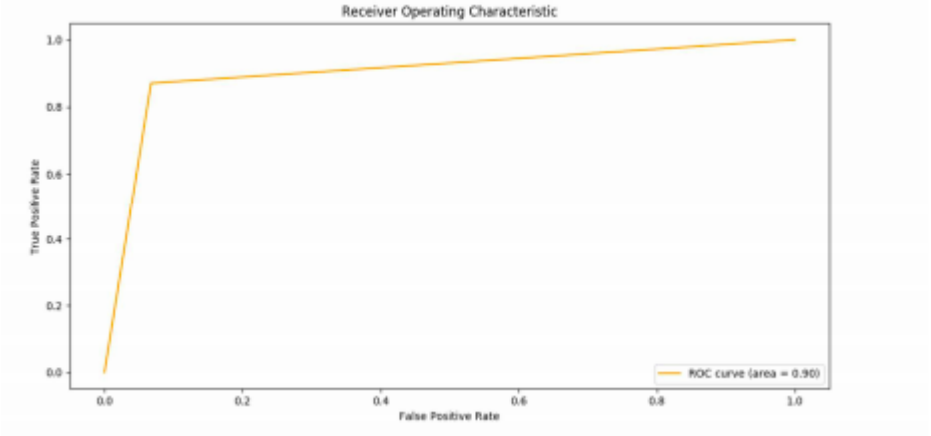


Fig3.5

LOCAL OUTLIER FACTOR BASED APPROACH

LOF scores can be used to discover outliers. Following points provide a quick of how to do it:

• Find the k-nearest-neighbors

• For each instance, compute the local density

• For every instance compute the ratio of neighborhood densities

• Normal examples have ratings shut to 1.0

• Anomalies have rankings > (1.2 ... 2.0)

3. STATISTICAL BASED APPROACH

In these sorts of Anomaly Detection approaches, records points are modeled using stochastic distribution and factors are decided to be outliers depending on their relationship with this model. For the purpose of this challenge two methods have been used:

• Mahalonobis distance from the centroid can be used to locate outliers

• Log Likelihood measure can also be used to detect outliers through checking if the log possibility of the dataset changes drastically by getting rid of a records point.

MAHALONOBIS DISTANCE

The reality that outliers have a greater mahalanobis distance than inliers, can be used to observe anomalies. Since the mahanalobis distances can be huge, a logarithm of base 10 was once used. And for every facts point whenever that handed the fee of 1.838 (determined the use of pass validation), that point was once declared as an outlier. Following consequences are received via making use of the algorithm on take a look at data set.

LOG LIKELIHOOD APPROACH

We can use the fact that if an anomaly is removed from the statistics set there is a considerable exchange in log likelihood of the complete data set. Following algorithms

explains how we can take advantage of that fact:

• Let M be the set of everyday information factors and A be the set of anomalous points.

• Initially anticipate all points to be in M (A is empty) and find Loglikelihood of M.

• Remove a point from M and discover the new Loglikelihood of M.

• If the Loglikelihood modifications drastically (change in likelihood is increased than some threshold) then cross it to A

• Otherwise maintain it in M

4. TRANSDUCTION BASED

Transduction is the procedure that reasons from unique cases (training) to specific instances (test). A strangeness function measures how atypical an item is. Given a distribution of strangeness values for the general population, compute the possibility of being an outlier for a given point. It avoids growing a model by way of making solely choices about character points at a time. For the purpose of this project, the StrOUD algorithm has been carried out the usage of KNN, LOF and ChiSquare as the strangeness functions.

STROUD ALGORITHM

The algorithm can be temporarily described in the following points:

• Sample the information set. Make sure solely ordinary records factors are selected.

• Calculate strangeness value of every point. This is known as the baseline.

• Sort the baseline (strangeness).

• For each test point, compute strangeness value with admire to preceding statistics points.

• Find how many of these preceding points had strangeness greater than or equal to the calculated

strangeness of the check point. Let's name it b.

• Compute p-value = (b+1)/(N+1) where N is the wide variety of preceding records points.

• If p-value < (1-confidence) then the test point is an outlier in any other case it is not.

CHAPTER 04:WORKING OF PROJECT

Dataset

The dataset that we used in our model is available in kaggle which was originally provided by European Bank in 2013 . And this dataset contains the translation that took place in that region for two days, where the number of total transaction were 284,807 and only 492 fraudulent were recorded among them.Thus the dataset is highly imbalance because the fraudulent transaction accounts for only 0.172% of all the transactions in dataset.

	V3	V4	V5	V6	V7	V8	V9	...	V21	V22	V23	V24	V25	V26	V27	V28	Amount	Class
	2.536347	1.378155	-0.338321	0.462388	0.239599	0.098698	0.363787	...	-0.018307	0.277838	-0.110474	0.066928	0.128539	-0.189115	0.133558	-0.021053	149.62	0
	0.166480	0.448154	0.060018	-0.082361	-0.078803	0.085102	-0.255425	...	-0.225775	-0.638672	0.101288	-0.339846	0.167170	0.125895	-0.008983	0.014724	2.69	0
	1.773209	0.379780	-0.503198	1.800499	0.791461	0.247676	-1.514654	...	0.247998	0.771679	0.909412	-0.689281	-0.327642	-0.139097	-0.055353	-0.059752	378.66	0
	1.792993	-0.863291	-0.010309	1.247203	0.237609	0.377436	-1.387024	...	-0.108300	0.005274	-0.190321	-1.175575	0.647376	-0.221929	0.062723	0.061458	123.50	0
	1.548718	0.403034	-0.407193	0.095921	0.592941	-0.270533	0.817739	...	-0.009431	0.798278	-0.137458	0.141267	-0.206010	0.502292	0.219422	0.215153	69.99	0

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 284807 entries, 0 to 284806
Data columns (total 31 columns):
Time        284807 non-null float64
V1          284807 non-null float64
V2          284807 non-null float64
V3          284807 non-null float64
V4          284807 non-null float64
V5          284807 non-null float64
V6          284807 non-null float64
V7          284807 non-null float64
V8          284807 non-null float64
V9          284807 non-null float64
V10         284807 non-null float64
V11         284807 non-null float64
V12         284807 non-null float64
V13         284807 non-null float64
V14         284807 non-null float64
V15         284807 non-null float64
V16         284807 non-null float64
V17         284807 non-null float64
V18         284807 non-null float64
V19         284807 non-null float64
V20         284807 non-null float64
V21         284807 non-null float64
V22         284807 non-null float64
V23         284807 non-null float64
V24         284807 non-null float64
V25         284807 non-null float64
V26         284807 non-null float64
V27         284807 non-null float64
V28         284807 non-null float64
Amount      284807 non-null float64
Class       284807 non-null int64
dtypes: float64(30), int64(1)
memory usage: 67.4 MB
```

Fig 4.1

Exploratory Data Analysis

The transactions that accounts for fraudulent is 0.17% of the dataset and legit transaction is 99.83% of the whole dataset this proves how imbalance our dataset is because most of the transactions are not fraud . So normal machine learning techniques will assume that transactions are not fraud , we don't want that to happen because we want that our model should learn the pattern and assume .

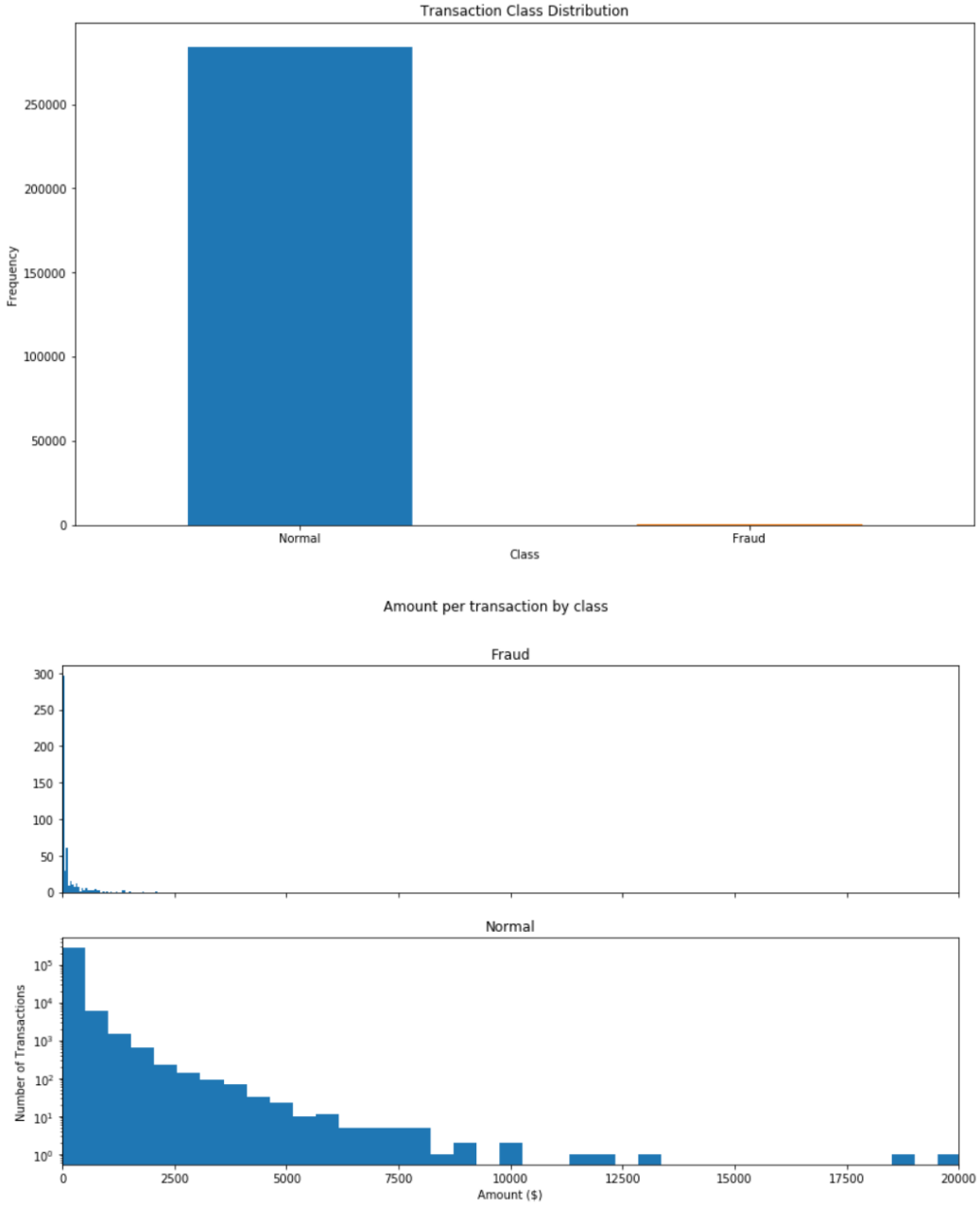


Fig 4.2

We will check that fraudulent transactions occur more often during certain time frame with the help of visual representation.

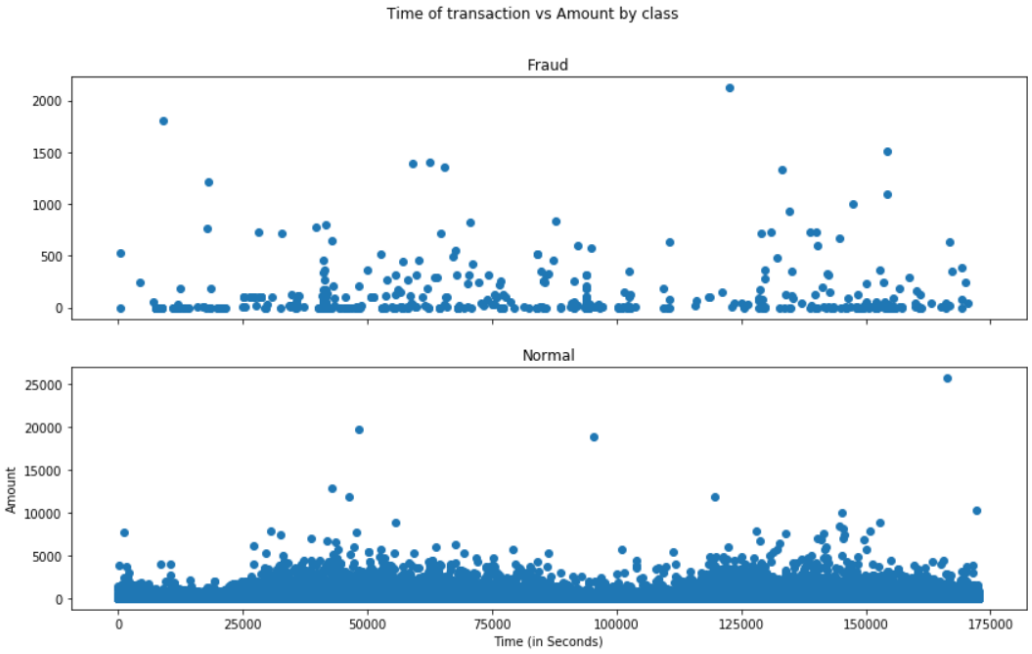


Fig 4.3

Here we have taken some samples of the data which 0.1 % of whole dataset because running the model using whole dataset will be time consuming and waste of the computational power .

And further below we have determine the fraud class as 1 and Valid class 0 ,which means in the dataset point which are recorded as 1 will be considered as fraudulent transactions

```
data1= data.sample(frac = 0.1,random_state=1)
data1.shape
```

```
(28481, 31)
```

```
data.shape
```

```
(284807, 31)
```

```
#Determine the number of fraud and valid transaction
Fraud = data1[data1['Class']==1]
Valid = data1[data1['Class']==0]
outlier_fraction = len(Fraud)/float(len(Valid))
```

Fig 4.4

Model Prediction

First we will create independent and Dependent Features and than Filter the columns to remove data we do not want and Store the variable we are predicting then Define a random state 42 and finally Print the shapes of X & Y

```
columns = data1.columns.tolist()
# Filter the columns to remove data we do not want
columns = [c for c in columns if c not in ["Class"]]
# Store the variable we are predicting
target = "Class"
# Define a random state
state = np.random.RandomState(42)
X = data1[columns]
Y = data1[target]
X_outliers = state.uniform(low=0, high=1, size=(X.shape[0], X.shape[1]))
# Print the shapes of X & Y
print(X.shape)
print(Y.shape)
```

Fig 4.5

Isolation Forest Algorithm :

The isolation forest algorithm here isolates the data points while selecting a feature and a split value between minimum and maximum values of that feature randomly .the algorithm here constructs the decision tree on basic of the decision tree the score is calculated as the path length of the observations.

Local Outlier Factor Algorithm:

The local outlier factor algorithm is also an unsupervised algorithm which works on the density of deviation of a given observations with respect to its neighbours.in this model information regarding the neighbours (fraud transactions) is not available so , we are taking neighbours = 20 which works well for our model.

```
##Define the outlier detection methods

classifiers = {
    "Isolation Forest":IsolationForest(n_estimators=100, max_samples=len(X),
                                       contamination=outlier_fraction,random_state=state, verbose=0),
    "Local Outlier Factor":LocalOutlierFactor(n_neighbors=20, algorithm='auto',
                                              leaf_size=30, metric='minkowski',
                                              p=2, metric_params=None, contamination=outlier_fraction),
```

Fig 4.6

These are some the Observations that were concluded after running the model .So the Isolation Forest detected 73 frauds and Local Outlier Factor detecting 97 frauds and also its shown that Isolation Forest has a 99.74% more accurate than LOF of 99.65% When comparing error precision & recall for 2 models , the Isolation Forest performed much better than the LOF as we can see that the detection of fraud cases is around 27 % versus LOF detection rate of just 2 % . So overall Isolation Forest Method performed much better in determining the fraud cases which is around 30%.

```
Isolation Forest: 73
Accuracy Score :
0.9974368877497279
Classification Report :
              precision    recall  f1-score   support

           0       1.00      1.00      1.00     28432
           1       0.26      0.27      0.26        49

   micro avg       1.00      1.00      1.00     28481
   macro avg       0.63      0.63      0.63     28481
weighted avg       1.00      1.00      1.00     28481

Local Outlier Factor: 97
Accuracy Score :
0.9965942207085425
Classification Report :
              precision    recall  f1-score   support

           0       1.00      1.00      1.00     28432
           1       0.02      0.02      0.02        49

   micro avg       1.00      1.00      1.00     28481
   macro avg       0.51      0.51      0.51     28481
weighted avg       1.00      1.00      1.00     28481
```

Fig 4.7

APPLICATION OF CREDIT CARD FRAUD DETECTION

We find that there are 91 cents per 100 dollar in loss frauds every year through paypal credit card and there are 4.64 billion dollar transactions then the loss per 100 dollar is 91 cent.so we can save many 3.5 million dollar money every year of paypal customers so that's how we use credit card fraud detection.Raw and unprocessed data they use the different techniques of python.It can be differanted by classifiers and performance which they can detect the normal transactions and abnormal transactions therefore with the help of random forest we can detect the credit card fraud.It gives us the best information of every transaction.It extracts the normal and abnormal transaction in two different sections.it is a very good algorithm and it is highly efficient.in our research we find that there are 91 cents per dollar in loss frauds every year through paypal credit card and 95 cents per dollar in loss frauds every year through visa cards.Only achieve 80 percent of accuracy rate with the noise level of only 20 percent but in the real world there is more noise than we think so we need more improvement to find the best accuracy.As we seen in this research the performance evaluation of classifiers which we are used in this models like naive bayes ,knn classifiers, logistic regression.we find that there are 91 cents per 100 dollar in loss frauds every year through paypal credit card and 95 cents per 100 dollar in loss frauds every year through visa cards.

Only achieve 80 percent of accuracy rate with the noise level of only 20 percent but in the real world there is more noise than we think so we need more improvement to find the best accuracy.There is 29 principal components are used which they can find credit card fraud this is also known as the principal component analysis to find out result that is credit card fraud.There is very good adaptability is used which we can predict the credit card fraud regression,vector machines help us find credit card fraud detection.with the help of supervised learning and unsupervised learning we can easily find abnormal transaction which help us to detect the credit card fraud detection.our fraud detection is work on model where we begin with incoming transaction then user input goes through the OCSVM/ T2 control chart which help us to differentiate that It can be different by classifiers and performance which they can detect the normal transactions and abnormal transactions therefore,T2 Control chart sends the user's information to in our algorithm help us to find the that it is legitimate transaction then it allows the transaction or Fraudulent transaction then it is alarm to bank.If output of our algorithm is 1 then this means it is fraudulent transaction then alarm to the bank quickly.if the output of our algorithm is zero then it means it is normal transaction,after the transaction the history of transaction is recorded in to database of bank.

Classification of the unequal structure is very hard to detect then we have to apply regression also so we can easily detect the credit card fraud detection.The selection on parameter in the dataset is very hard and we can not find easily animal point in the dataset.these are difficulties that's why we use isolation forest .

CONCLUSION

As we know there are many other technique available for fraud detection but not even single one is able to detect all frauds when they are happening ,they are detected when the fraudulent transaction is already done.The main reason for this because there are very small number of fraud transactions as compared to the legitimate ones.So we require the mechanism that can detect the fraud transactions when they are taking place , so it can be stop before it takes place . To do that major task in hand is to build a detection technique which is fast , precise and accurate for credit card fraud ,it also should be able to detect fraud that are done using various techniques for example phishing , vishing, skimming and many other type of frauds . But all of that is possible when we have good dataset other wise there is no guarantee that it will give the results that we are looking for.

Using our dataset , our detection system is able to detect huge amount of of cases with very accurate and precise results. The algorithm that we use also plays an important role in this because the isolation forest algorithm which is an anomaly detection algorithm is perfect for the credit card fraud detection ,it detect the anomaly using the variable that are provided to it by the dataset which contains features like v1 ,v2,v3 to v28 these are transaction details which are coded by the bank from which the our dataset is taken and other time , amount and frequency.Using all of these variable and data machine learning technique like decision tree, Logistic regression ,anomaly detection were used to detect the fraud transaction in the credit card fraud detection system.Isolation forest algorithm gives us accuracy of 99.75 % which more than any other technique , logistic regression gives the accuracy of 94% ,SVC and KNN gives 93% and only technique closest to Isolation Forest is LOF which gives 99% accuracy.So we conclude that Isolation Forest is best for our system.

REFERENCES

Raj, S. B. E., & Portia, A. A. (2011, March). Analysis on credit card fraud detection methods. In *2011 International Conference on Computer, Communication and Electrical Technology (ICCCET)* (pp. 152-156). IEEE.

Shen, A., Tong, R., & Deng, Y. (2007, June). Application of classification models on credit card fraud detection. In *2007 International conference on service systems and service management* (pp. 1-4). IEEE.

Awoyemi, J. O., Adetunmbi, A. O., & Oluwadare, S. A. (2017, October). Credit card fraud detection using machine learning techniques: A comparative analysis. In *2017 International Conference on Computing Networking and Informatics (ICCNI)* (pp. 1-9). IEEE.

Xuan, S., Liu, G., Li, Z., Zheng, L., Wang, S., & Jiang, C. (2018, March). Random forest for credit card fraud detection. In *2018 IEEE 15th International Conference on Networking, Sensing and Control (ICNSC)* (pp. 1-6). IEEE.

John, H., & Naaz, S. (2019). Credit card fraud detection using local outlier factor and isolation forest. *Int. J. Comput. Sci. Eng*, *7*(4), 1060-1064.