International Journal for Innovative Engineering and Management Research
A Peer Revieved Open Access International Journal
www.ijiemr.org

CREDIT CARD FRAUD DETECTION

Aishwarya Vuppala[1], Navya Reddy Boddu[2]
Department of Computer Science and Engineering, Stanley College of Engineering and Technology for Women, Telangana, India

Abstract: Now a day's online transactions have become an important and necessary part of our lives. It is vital that credit card companies are able to identify fraudulent credit card transactions so that customers are not charged for items that they did not purchase. As frequency of transactions is increasing, number of fraudulent transactions are also increasing rapidly. Such problems can be tackled with Machine Learning with its algorithms. This project intends to illustrate the modelling of a data set using machine learning with Credit Card Fraud Detection. Fraud detection has become an important tool and probably the best way to stop such frauds. The Credit Card Fraud Detection Problem includes modelling past credit card transactions with the data of the ones that turned out to be fraud. This model is then used to recognize whether a new transaction is fraudulent or not. Our objective here is to detect 100% of the fraudulent transactions while minimizing the incorrect fraud classifications.

Keywords: Machine Learning, Logistic Regression, Random Forest, Classification, Pandas, Numpy, Seaborn, Credit Card Fraud Detection, Pre-Processing, Sci-Kit Learn, Matplotlib.

1.INTRODUCTION

Credit Card: A credit card is a payment card issued to users (cardholders) to enable the cardholder to pay a merchant for goods and services based on the cardholder's promise to the card issuer to pay them for the amounts plus the other agreed charges.
The card issuer (usually a bank) creates a revolving account and grants a line of credit to the cardholder, from which the cardholder can borrow money for payment to a merchant or as a cash advance.

Fraud: Fraud is intentional deception to secure unfair or unlawful gain, or to deprive a victim of a legal right. Fraud can violate civil law (i.e., a fraud victim may sue the fraud perpetrator to avoid the fraud or recover monetary compensation), a criminal law (i.e., a fraud perpetrator may be prosecuted and imprisoned by governmental authorities), or it may cause no loss of money, property or legal right but still be an element of another civil or criminal wrong. The purpose of fraud may be monetary gain or other benefits, for example by obtaining a passport, travel document, or driver's license, or mortgage fraud, where the perpetrator may attempt to qualify for a mortgage by way of false statements.

Credit card fraud Detection: Most of us are using electronic payments as they are more restful, seamless, adequate, and simple to use; however, we must not overlook the losses associated with electronic commerce. Organizations and banks must have to use good security procedures in order to minimize the losses. So, it is critical to improve detection and prevention techniques. It is vital to understand the mechanisms for carrying a fraud in order to combat the fraud effectively. The gadget for identifying credit score card fraud relies upon on the fraud manner itself. To accomplish this, provide the transaction details to the verification module, which will classify them as either fraud or non-fraud. If it classified as fraudulent, it will be rejected. Otherwise, the transaction is accepted. Fraud detection techniques such as statistical data analysis and artificial intelligence can be used to distinguish between the two. AI technique includes data mining that used to detect fraud, which can classify, group, and segment data to search through millions of transactions to find patterns and detect fraud. Machine learning is a technique for automatically detecting fraud characteristics. One method of dealing with fraud is through both prevention and detection. Fraud detection and prevention's primary goal is to tell the difference between legitimate and fraudulent transactions and to prevent fraudulent activity. Using historical data, the user's pattern and behaviour are analysed to determine if a transaction is fraudulent or not. When the system fails to detect and prevent fraudulent activities, fraud detection takes over. In supervised fraud detection systems, new transactions are classified as fraudulent or genuine based on characteristics of deceptive and legitimate activities, whereas outliers' transactions are identified as prospective fraudulent transactions in unsupervised fraud detection systems. Diversity of studies have been conducted on several methods to solve the issue of card fraud detection. These approaches include, ANN, K-means Clustering, Decision Trees, etc.

- Types Of Frauds :-

1. Application fraud: Application fraud takes place when a person uses stolen or fake documents to open an account in another person's name.
2. Account takeover: An account takeover refers to the act by which fraudsters will attempt to assume control of a customer's account (i.e credit cards, email, banks, SIM card and more).Control at the account level offers high returns for fraudsters.
3. Social engineering fraud: Social engineering fraud can occur when a criminal poses as someone else which results in a voluntary transfer of money or information to the fraudster.
4. Skimming: Skimming is the theft of personal information having used in an otherwise a normal transaction. The thief can procure a victim's card number using basic methods such as photocopying receipts or more advanced methods such as using a small electronic device (skimmer) to swipe and store hundreds of victims' card numbers.
5. Unexpected repeat billing: Online bill paying or internet purchases utilizing a bank account are a source for repeat billing known as "recurring bank charges". These are standing orders or banker's orders from a customer to honor and pay a certain amount every month to the payee.
6. Stealing the card: In this type of credit card fraud, fraudsters steal the credit card from your pocket or purse and use it to conduct unauthorized transactions. In case of stolen credit cards, it is difficult for fraudsters to swipe your card at POS machines, as it requires a PIN number. So, they use it to make online purchases through various tricks. It is advisable to block your credit card as soon as you lose it. Make sure to keep your credit card safely, so that it’s not stolen.
7. Phishing emails: Another trick used to steal credit card information is a phishing email. Phishing scams are rising each day as people using emails increase. Under this fraud, fraudsters send emails which look exactly like those of a genuine company. This is mainly done to collect personal information and credit card details. Many people think these emails are sent by genuine companies and provide credit card details. When credit card details are given by the cardholders, it is easy for fraudsters to steal the money. So, make sure not to entertain such emails and do not give your credit card information to anyone.
8. Fraud calls and SMS: Fraud calls or messages are another technique used by fraudsters to steal your credit card details. Fraudsters call customers claiming to be bank officials and try to collect personal information. They say that your credit card will expire and it has to be renewed to prevent it getting blocked. They also promise help in renewing the credit card and you will be asked to provide credit card details and the OTP. One thing we should always keep in mind is that bank officials will never ask you to give personal information over the phone. In case you receive such calls, report them to the police.

1.1About Project

Now a day’s online transactions have become an important and necessary part of our lives. It is vital that credit card companies are able to identify fraudulent credit card transactions so that customers are not charged for items that they did not purchase. As frequency of transactions is increasing, number of fraudulent transactions are also increasing rapidly. Such problems can be tackled with Machine Learning with its algorithms. This project intends to illustrate the modelling of a data set using machine learning with Credit Card Fraud Detection. Fraud detection has become an important tool and probably the best way to stop such frauds. The Credit Card Fraud Detection Problem includes modelling past credit card transactions with the data of the ones that turned out to be fraud. This model is then used to recognize whether a new transaction is fraudulent or not. Our objective here is to detect 100% of the fraudulent transactions while minimizing the incorrect fraud classifications.

1.2 Objectives of the Project

The main objective of this is to perform predictive analysis on credit card transaction dataset using machine learning techniques and detect the fraudulent transactions from the given dataset. The focus is to identify if a transaction comes under normal class or fraudulent class using predictive models. Different sampling techniques will be implemented to tackle the class imbalance problem and series of machine learning algorithms like logistic regression, random forest will be implemented on the dataset, and the results will be reported by analysing the performance of proposed algorithm with test dataset.

1.3 Scope of the Project

Technology changes can be added to know the pattern of fraudulent transactions and alert the respective card holders and bankers when fraud activity is identified. To this extent, the incremental approach is necessary in making the system to learn from past transactions and making it capable of handling the fraud transactions.

1.4 Advantages

- One of the biggest advantages of random forest is its versatility. It can be used for both regression and classification tasks.
- Random forest algorithm builds multiple decision trees and merges them together to get a more accurate and stable prediction.
- The random forest algorithm also works well when data possess missing values.
- Reduction in number of fraud transactions.
- User can safely use his/her credit and debit cards for online transactions.
- Higher Security: Credit card Fraud Detection recognition software helps businesses and banks to achieve higher security by facilitating quicker detection of fraud.
- Cost Reduction: Opting for credit card Fraud Detection will help businesses on cutting down on hiring professionals to carry out data extraction, which is one of the most important benefits of credit card fraud detection data entry methods. Therefore, this machine learning detection model eliminates the cost of misplaced or lost data.

1.5 Disadvantages

- Although random forest can be used for both classification and regression tasks, it is not more suitable for Regression tasks.
- Not 100% accurate, there are likely to be some mistakes made during the method.
- Not worth doing for little amounts of data.

1.6 Applications

Credit card fraud detection are used in financial intuitions such as banks, in online transactions, businesses which will accept contactless payments.

Nowadays most of us are purchasing shopping from online platforms such as amazon, flipkart, myntra to buy their daily requirements such as groceries, clothes, shoes etc and from the time Covid has come these times of transactions have increased tremendously. So, it has become very much important to detect frauds related to credit card. We also should be careful by not sharing our OTP or CVV numbers so as to prevent these frauds. This proposed model will detect the frauds of the credit card transactions.

1.7 Hardware and Software Requirements:

Operating system: Windows 10
Programming Language: Machine Learning
IDE: Jupyter/Spyder, Anaconda Python 3.2

Packages:

1) NUMPY: NumPy is a general-purpose array-processing package. It provides a high-performance multidimensional array object, and tools for working with these arrays.
   It is the fundamental package for scientific computing with Python.

It contains various features including these important ones:

- A powerful N-dimensional array object
- Sophisticated (broadcasting) functions
- Tools for integrating C/C++ and Fortran code
- Useful linear algebra, Fourier transform, and random number capabilities.

NumPy can also be used as an efficient multi-dimensional container of generic data. Arbitrary data-types can be defined using Numpy which allows NumPy to seamlessly and speedily integrate with a wide variety of databases.

2) PANDAS: Pandas is the most popular python library that is used for data analysis. It provides highly optimized performance with back-end source code is purely written in C or PYTHON.

3) SCIKIT LEARN: Scikit-learn is the most useful library for machine learning in Python. The Sklearn library contains a lot of efficient tools for machine learning and statistical modeling including classification, regression, clustering and dimensionality reduction. sklearn is used to build machine learning models.It should not be used for reading the data, manipulating and summarizing it. There are better libraries for that (e.g. NumPy, Pandas etc.)

4) MATPLOTLIB: Matplotlib is an amazing visualization library in Python for 2D plots of arrays. Matplotlib is a multiplatform data visualization library built on NumPy arrays and designed to work with the broader SciPy stack. It was introduced by John Hunter in the year 2002.
One of the greatest benefits of visualization is that it allows us visual access to huge amounts of data in easily digestible visuals. Matplotlib consists of several plots like line, bar, scatter, histogram etc.

Fig 1.7.1: Software and Package Requirements

<u>Hardware Requirements</u>:
System: Intel i5 core
Hard Disk: 100GB
Monitor: 15” LED
Input devices: Keyboard, Mouse
RAM: 8 GB

2. LITERATURE SURVEY

Vimala Devi. J et al[1]. To detect counterfeit transactions, three machine-learning algorithms were presented and implemented. There are many measures used to evaluate the performance of classifiers or predictors, such as the Vector Machine, Random Forest, and Decision Tree.

R.R Popat and Chaudhary[2] supervised algorithms were presented Deep learning, Logistic Regression, Nave Bayesian, Support Vector Machine (SVM), Neural Network, Artificial Immune System, K Nearest Neighbour, Data Mining, Decision Tree, Fuzzy logic based System, and Genetic Algorithm are some of the techniques used. Credit card fraud detection algorithms identify transactions that have a high probability of being fraudulent. We compared machine-learning algorithms to prediction, clustering, and outlier detection.

Asha R B et al[3].have proposed a deep learning-based method for detecting fraud in credit card transactions. Using machine-learning algorithms such as support vector machine, k-nearest neighbour, and artificial neural network to predict the occurrence of fraud.

Sangeeta Mittal et al[4].To evaluate the underlying problems, some popular machine learning algorithms in the supervised and unsupervised categories were selected. A range of supervised learning algorithms, from classical to modern, have been considered. These include tree-based algorithms, classical and deep neural networks, hybrid algorithms and Bayesian approaches. The effectiveness of machine-learning algorithms in detecting credit card fraud has been assessed. On various metrics, a number of popular algorithms in the supervised, ensemble, and unsupervised categories were evaluated. It is concluded that unsupervised algorithms handle dataset skewness better and thus perform well across all metrics absolutely and in comparison to other techniques.

Borse, Suhas and Dhotre[5]. Machine learning's Naive Bayes classification was used to predict common or fraudulent transactions. The accuracy, recall, precision, F1score, AUC score of Naïve Bayes classifier are calculated.

Suman Arora[6] In this paper, many supervised machine learning algorithms apply on 70% training and 30% testing dataset. Random forest, stacking classifier, XGB classifier, SVM, Decision tree and KNN algorithms compare each other i.e. 94.59%,95.27%, 94.59%, 93.24%, 90.87%, 90.54% and 94.25% respectively. Summaries of this paper, SVM has the highest ranking with 0.5360 FPR, and stacking classifier has lowest ranking 0.0335.

Kosemani Temitayo Hafiz[7] in this paper, they describe flow chart of fraud detection process. i.e. data Acquisition, data pre-processing, Exploratory data analysis and methods or algorithms are in detail. Algorithms are K- nearest neighbour (KNN), random tree and Logistic regression accuracy are 96.91%, 94.32%, 57.73% and 98.24% respectively.

2.1 Existing System

In existing system, the K-means clustering model produced a low accuracy. Using K-means there were quite a few non-fraudulent activities, which wrongly got detected as frauds. Therefore, K-means would not be the preferred model, as it doesn’t correctly predict frauds and it also produced a lot of false positives. The Traditional detection method mainly depends on database system and the education of customers, which usually are delayed, inaccurate and not in-time. After that methods based on discreate analysis and mining algorithms are widely used which can detect fraud by credit rate for cardholders and credit card transaction. For a large amount of data it is not efficient.

2.2 Proposed System

The proposed system is a machine learning application to detect frauds in credit card transactions using Random Forest algorithm. These algorithms are used to classify the credit card data set and then regression is performed. The Random Forest Algorithm has been found to produce a good estimate of the generalization error and to be resistant to over fitting. This algorithm has been found to produce a good accuracy and precision. The proposed system overcomes the above mentioned issue in an efficient way. Using random forest and logistic algorithms the fraud is detected and the false alert is minimized and it produces an optimized result. Here the random forest and logistic algorithms are made where a set of interval valued parameters are optimized.

3. PROPOSED ARCHITECTURE

3.1 BLOCK DIAGRAM

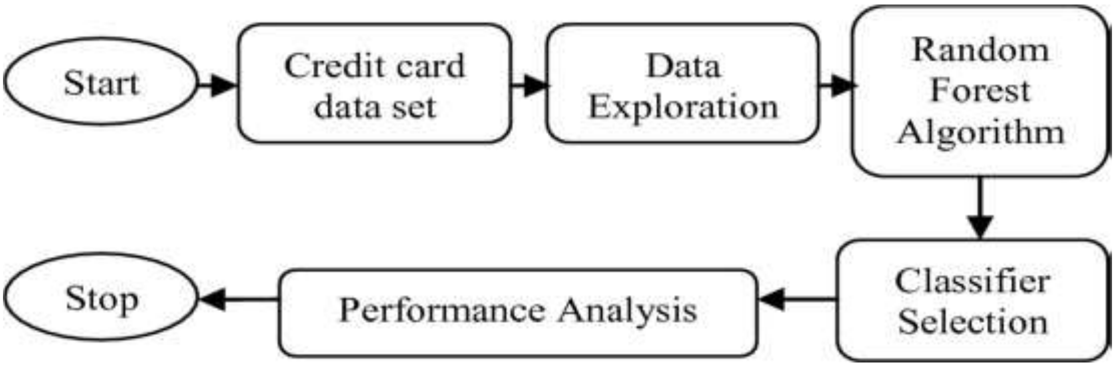


Fig 3.1.1: Proposed Architecture

Credit Card Data Set:

- Data Collection: Dataset used in this model are transactions made by credit cardholders.
- Data Pre-Processing: Pre-processing is the process of three important and common steps as follows:

Formatting: It is the process of putting the data in a legitimate way that it would be suitable to work with. Most recommended format is .csv files.

Cleaning: Data cleaning is a very important procedure in the path of data science as it constitutes the major part of the work. It includes removing missing data and complexity with naming category and so on.

Sampling: This is the technique of analyzing the subsets from whole large datasets, which could provide a better result and help in understanding the behaviour and pattern of data in an integrated way.

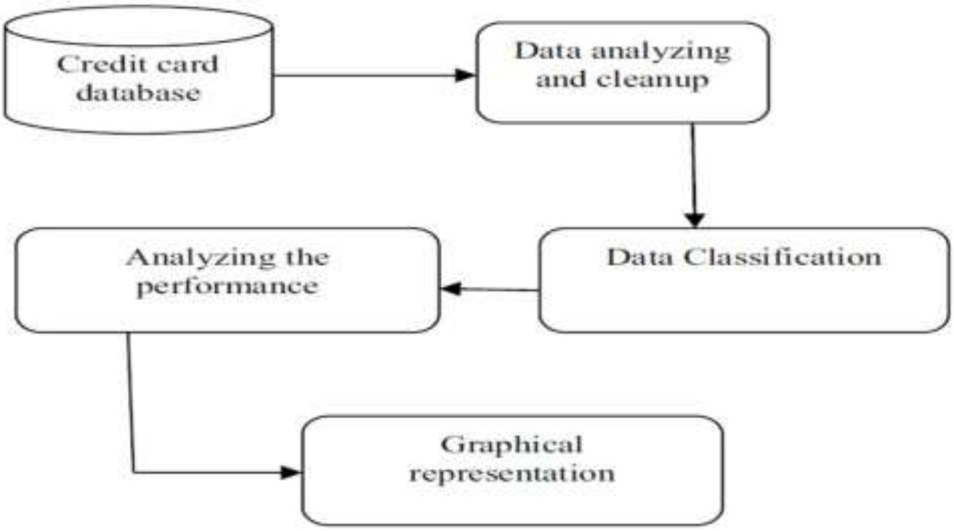


Fig 3.1.2 Steps involved in Database cleaning

- Feature Extraction: Feature extraction is the process of studying the behaviour and pattern of the analysed data and draw the features for further testing and training. Finally, our models are trained using the Classifier algorithm. Random forest algorithm was used to classify pre-processed data.

- Display Graph: The result will be in the visualized form. Representation of classified data in the form of graphs. Accuracy is well-defined as the proportion of precise predictions for the test data.

- It can be calculated easily by mathematical calculation i.e., dividing the number of correct prediction by the number of total prediction.

Random Forest Algorithm implementation in Credit Card Fraud Detection

In credit card fraud detection the Random Forest Algorithm gives better accuracy in results. First all the dataset will be collected and analysed. During analysis process all the duplicate values and also the null values will be removed from the dataset. Now the dataset will be pre-processed based on the amount and transaction time for finding the accuracy of the resultant dataset. After the pre-processing of dataset into amount and transaction time now the dataset will be divided into two categories. The dataset is classified in two categories as trained data and test dataset. Here for dataset classification we use a software called 'Scikit-learn'.

Scikit-learn is a free software for machine learning library in python where it contains features like classification, regression, Clustering algorithms and also various algorithms to interoperate with Python. After the pre-processing of the dataset now we apply the Random Forest Algorithm. By applying Random Forest Algorithm the pre processed dataset will be analysed again and then a confusion matrix will be obtained. In confusion matrix the dataset will be partitioned into four blocks as True Positive(TP), True Negative(TN), False Positive(FP) and False Negative(FN). Now the dataset will be partitioned continuously until all the data is validated. Now all these partitioned data will be evaluated and finally it will be represented as separate graphs. These separate graphs will give only less accuracy about the resultant dataset. So in order to obtain better accuracy we use Random Forest Algorithm where it takes all the graph values and give us only necessary values with better accuracy when compared with all other algorithms.

Performance Analysis:

- In this module first the dataset will be divided into two partitions as trained dataset and testing dataset.
- After the data partitions the Random Forest Algorithm is applied. After applying Random Forest Algorithm finally a confusion matrix is obtained.

Now the resultant data obtained in the form of confusion matrix can be evaluated by using graphical representation which gives better accuracy.

3.2 UML DIAGRAMS

The activity of the Uml diagram of credit card approval system which shows the flow between the activity of credit card, consumer, document, application, limits.

The main activity involved in this Uml activity diagram of credit card are as follows:

- Credit card activity
- Consumer activity
- Document activity
- Application activity
- Limits activity Features of the activity UML diagram of credit card

- Admin user can search credit card, view description of a selected credit card, add credit card, update credit card and delete credit card.
- It shows the activity flow of editing, adding and updating of consumer.
- User will be able to search and generate report of document, application, limits.
- All objects such as (credit card, consumer, limits) are interlinked.

It shows the full description and flow of credit card, application, limits, document, consumer.

Use-Case Diagram:

A use case diagram is a type of behavioral diagram defined by and created from a Use-case analysis. Its purpose is to present a graphical overview of the functionality provided by a system in terms of actors, their goals and any dependencies between those use cases. The below diagram figure shows the overall use case diagram for credit card fraud detection. A use case diagram is a type of behavioral diagram defined by the unified modeling language. The USE CASE diagram below describes the interaction between the customers and card issuers.

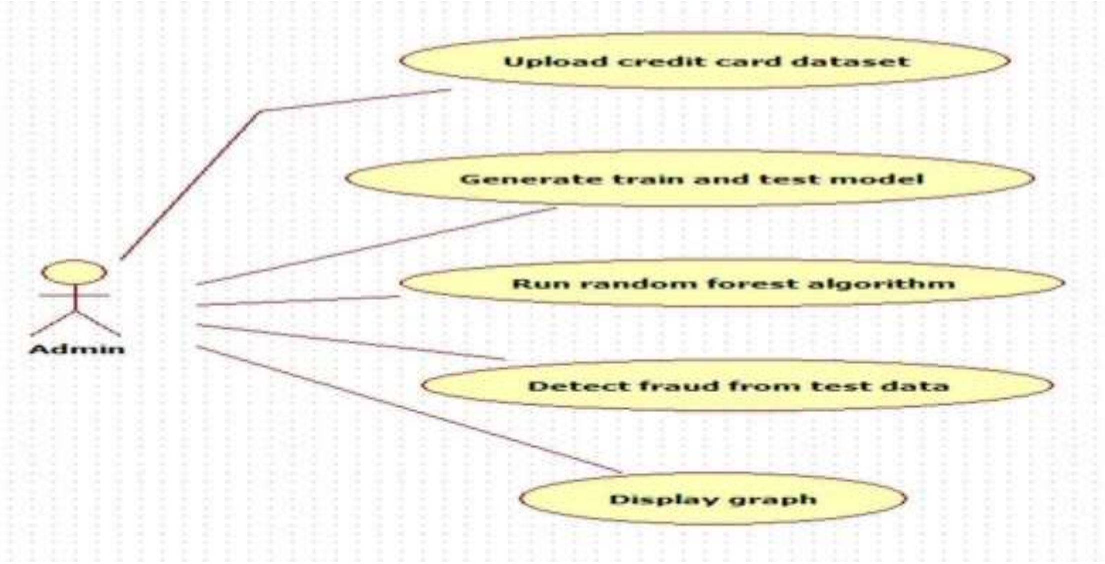
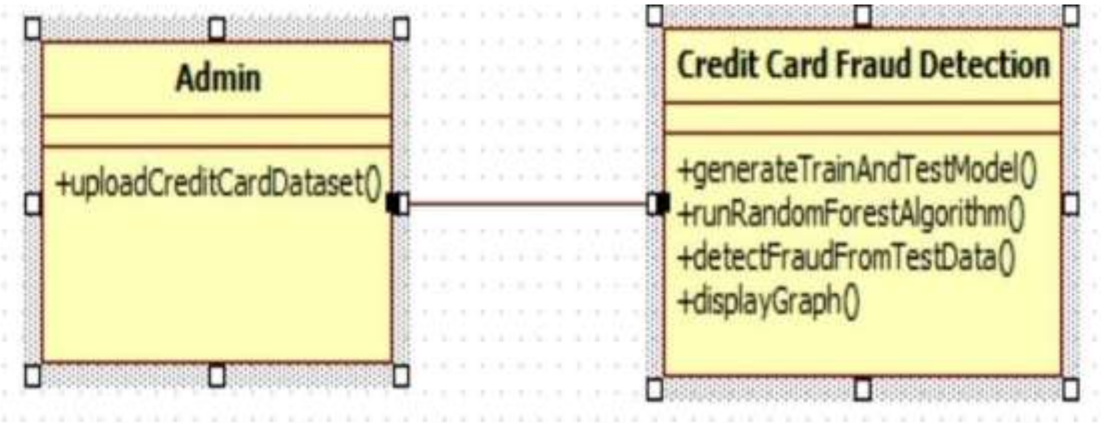

Upload credit card dataset
Generate train and test model
Run random forest algorithm
Detect fraud from test data
Display graph
Admin

Fig 3.2.1: Use-Case Diagram

Sequence Diagram:

Fig 3.2.2: Sequence Diagram

Class Diagram:


Admin
+uploadCreditCardDataset()
Credit Card Fraud Detection
+generateTrainAndTestModel()
+runRandomForestAlgorithm()
+detectFraudFromTestData()
+displayGraph()

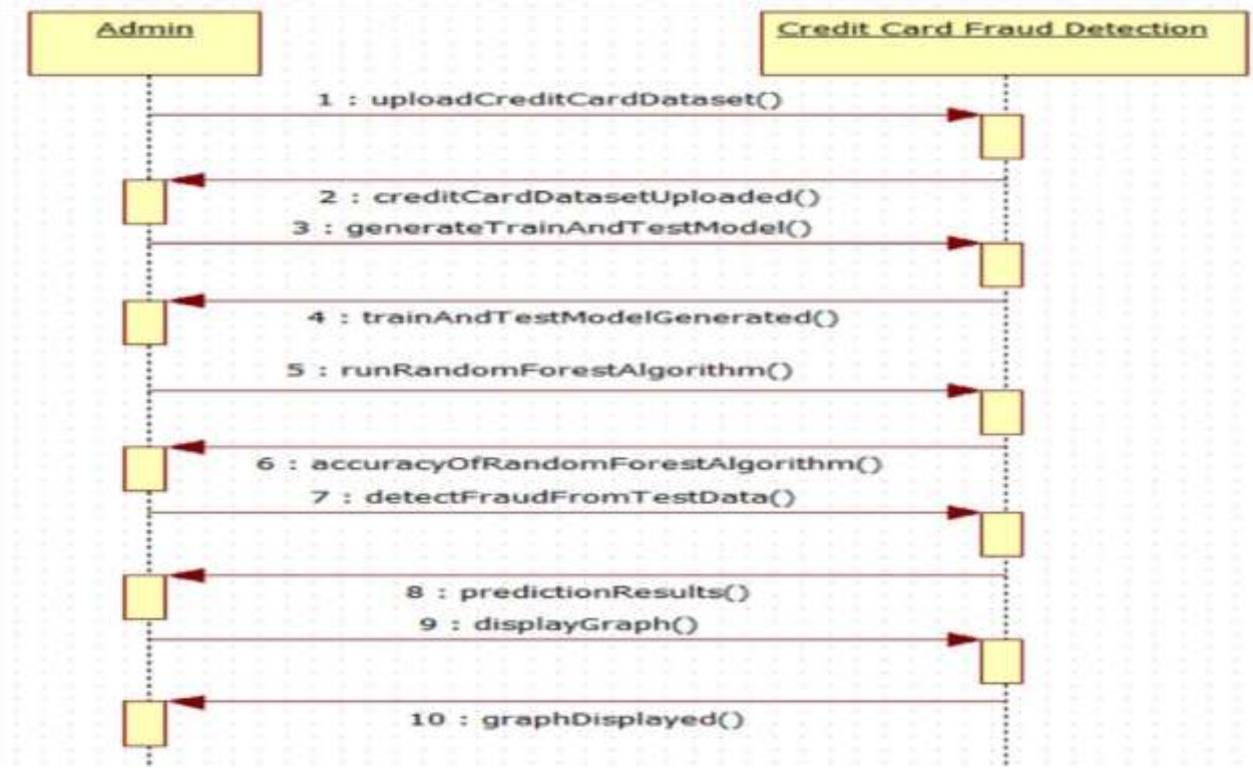

Admin
Credit Card Fraud Detection
1 : uploadCreditCardDataset()
2 : creditCardDatasetUploaded()
3 : generateTrainAndTestModel()
4 : trainAndTestModelGenerated()
5 : runRandomForestAlgorithm()
6 : accuracyOfRandomForestAlgorithm()
7 : detectFraudFromTestData()
8 : predictionResults()
9 : displayGraph()
10 : graphDisplayed()

Fig 3.2.3: Class Diagram

Activity Diagram:

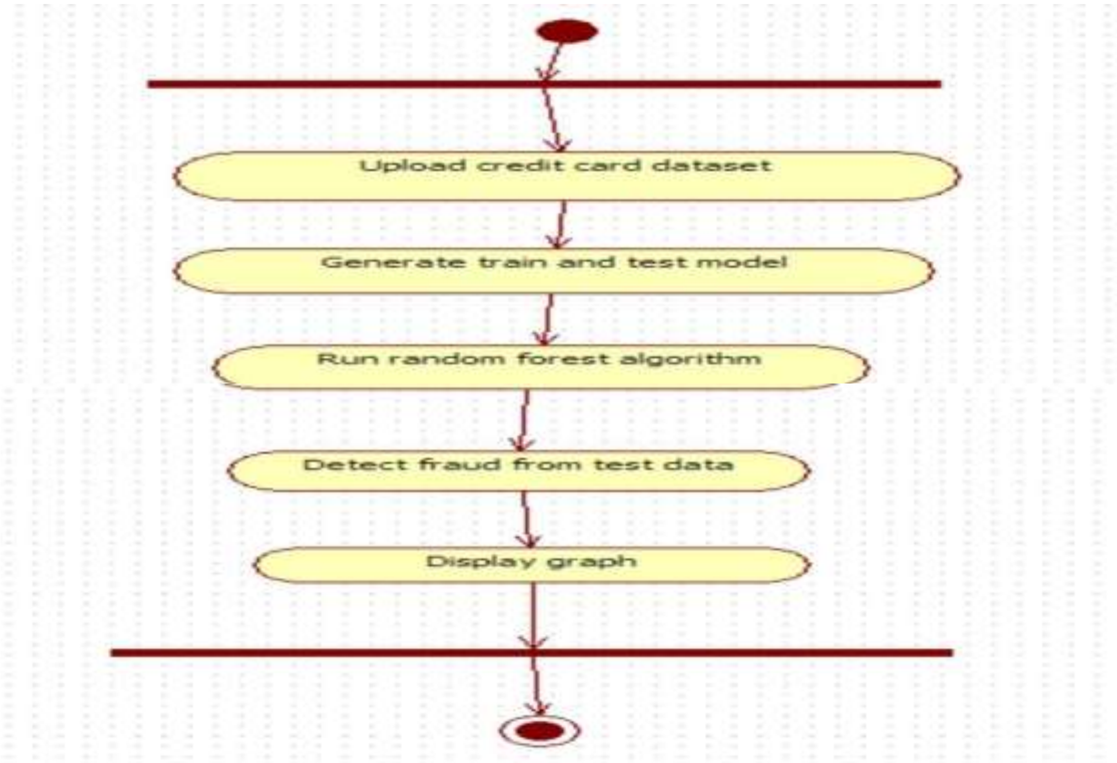

Upload credit card dataset
Generate train and test model
Run random forest algorithm
Detect fraud from test data
Display graph

Fig 3.2.4: Activity diagram

4. IMPLEMENTATION

Environmental Setup:

- We need to install and setup the IDE
- After installing we need to set the path in environmental variables
- The process for installing is as below

Steps Installing Anaconda:

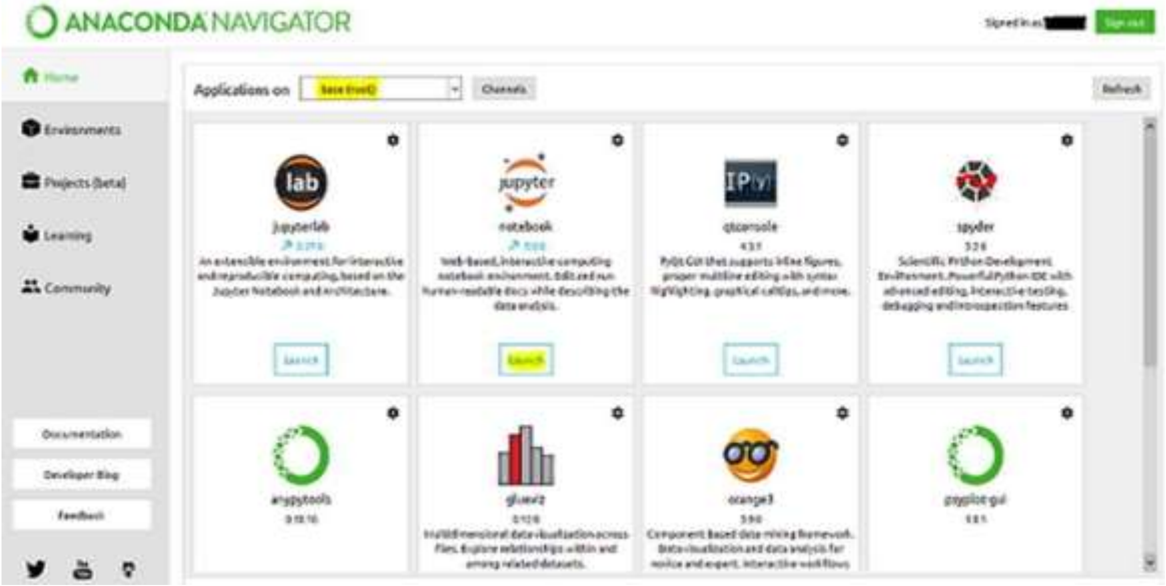
1.Downloads and install Anaconda from https://repo.anaconda.com/archive/Anaconda32021.05Windows-x86_64.exe.

2.After opening link u can see this download option

3.Click on the download option.
4.After downloading start installation.
5.Select the default options when prompted during the installation of Anaconda.
6.Ensure that the path to the folder where Anaconda is installed is added to your computer/system.

Start Jupyter Notebook:

1.Open anaconda navigator and the screen which is similar to below appears.



An extensible environment for interactive and reproducible computing, based on the Jupyter Notebook and Architecture.

Web-based, interactive computing notebook environment. Edit and run human-readable docs while describing the data analysis.

PyQt GUI that supports inline figures, proper multiline editing with syntax highlighting, graphical calltips, and more.

Scientific Python Development Environment. Powerful Python IDE with advanced editing, interactive testing, debugging and introspection features

Multidimensional data visualization across files. Explore relationships within and among related datasets.

Component based data mining framework. Data visualization and data analysis for novice and expert. Interactive workflows

Fig 4: Jupyter Notebook Screen

2.Open anaconda prompt to open jupyter notebook.

3.Now open jupyter new kernel.
4.Install required packages.

4.1 Algorithm

Random Forest Algorithm

Random Forest is also called as Random Decision Forest (RFA) which is used for Classification,Regression and other tasks that are performed by constructing multiple decision trees. This Random Forest Algorithm is based on supervised learning and the major advantage of this algorithm is that it can be used for both Classification and Regression. Random Forest Algorithm gives you better accuracy when compared with all other existing systems and this is most commonly used algorithm. In this paper the use of Random forest algorithm in credit card fraud detection can give you accuracy of about 90 to 95%.

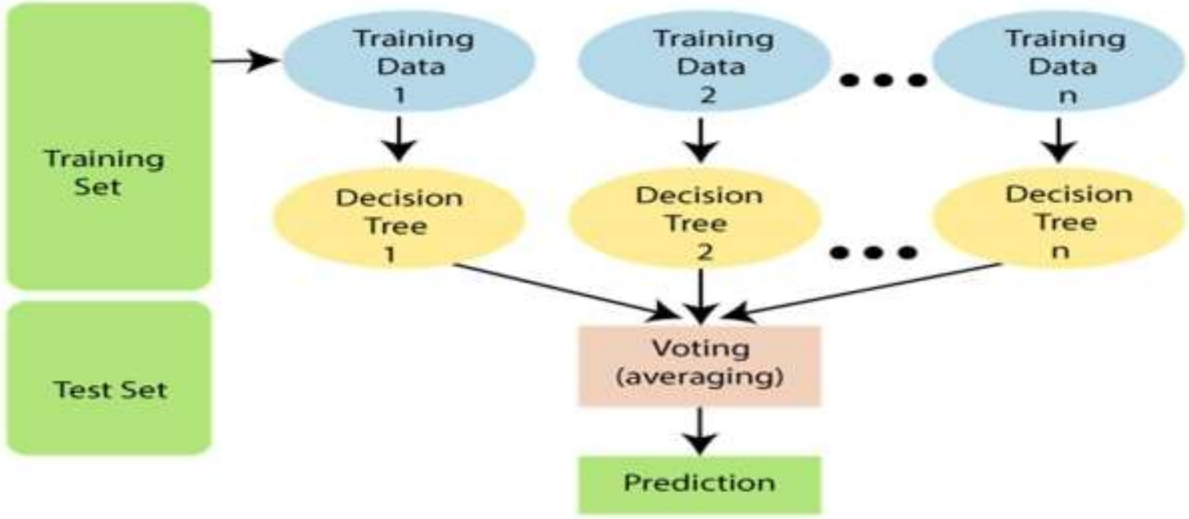


Fig 4.1.1: Random forest flowchart

Logistic Regression:

An algorithm that can be used for both regression and classification tasks, but it is most commonly used for classification.' Logistic Regression is used to predict categorical variables using dependent variables. Consider two classes, and a new data point is to be checked to see which class it belongs to. The algorithms then compute probability values ranging between (0) and (1). Logistic Regression employs a more complex cost function, this cost function is known as the Sigmoid Function or the Logistic Function'. LR also does not require independent variables to be linearly related, nor does it require equal variance within each group, making it a less stringent statistical analysis procedure. As a result, logistic regression was used to predict the likelihood of fraudulent credit.

We use Random Forest Algorithm and Logistic Regression for classification and regression of dataset. First we will collect the Credit Card dataset and analysis will be done on the collected dataset. After the analysis of dataset then cleaning of dataset is required. Generally in any dataset there will be many duplicate and null values will be present, so to remove all those duplicate and null values cleaning process is required. Then we have to split the dataset into two categories as Trained dataset and Testing dataset for comparing and analyzing the dataset. After dividing the dataset we have to apply the Logistic Regression and Random Forest Algorithm to compare which algorithm will give us the better accuracy about the credit card fraud transactions. By applying the Algorithm the dataset will be classified into four categories which will be obtained in the form of confusion matrix. Based on the above classification of data performance analysis will be done. In this analysis the accuracy of credit card fraud transactions can be obtained which will be finally represented in the form of graphical representation.

4.2 Code Implementation

```
#importing the necessary packages
import pandas as pd
import seaborn as sns
import numpy as np
import matplotlib.pyplot as plt
from matplotlib import gridspec
 #loading the data
data = pd.read_csv('credit card fraud.csv')
 #describing the data
data.head()
#describing the data
```

```
import pandas as pd
import seaborn as sns
import numpy as np
import matplotlib.pyplot as plt
from matplotlib import gridspec
#checking whether any null values are present
data.info()
#determining the fraudulent and not fraud transactions
Fraud = data[data['Class'] == 1]
notfraud = data[data['Class'] == 0]
outlier_fraction = len(Fraud)/float(len(notfraud))
print(outlier_fraction)
print('Fraud Cases: {}'.format(len(data[data['Class'] == 1])))
print('notfraud Transactions: {}'.format(len(data[data['Class'] == 0])))
#Correlation matrix
corrmat = data.corr()
fig=plt.figure(figsize=(12,9))
sns. heatmap(corrmat, vmax = .8, square = True)
plt.show()
#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)
X_data=X.values
Y_data=Y.values
#training and testing the data
from sklearn.model_selection import train_test_split
X_train, X_test, y_train, y_test = train_test_split(X_data, Y_data, train_size=0.70, test_size=0.30,
random_state=1)
```

Building logistic regression model using skicit learn

```
#Instantiate the model to an empty object
from sklearn.linear_model import LogisticRegression
model = LogisticRegression()
#Train the model using 'fit' method
model.fit(X_train, y_train)
#Test the model using 'predict' method
y_pred = model.predict(X_test)
#Print the classification report
From sklearn.metrics import classification_report,confusion_matrix,accuracy_score
print(classification_report(y_test, y_pred))
print(classification_report(y_test, y_pred))
print(confusion_matrix(y_test, y_pred))
print(accuracy_score(y_test, y_pred))
#printing the confusion matrix
LABELS = ['Notfraud', 'fraud']
conf_matrix = confusion_matrix(y_test, y_pred)
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()
#Building the Random Forest Classifier (RANDOM FOREST)
from sklearn.ensemble import RandomForestClassifier
```

```
# random forest model creation
rfc = RandomForestClassifier()
rfc.fit(X_train,y_train)
# predictions
y_pred = rfc.predict(X_test)
#Print the classification report
from sklearn.metrics import classification_report,confusion_matrix,accuracy_score
print(classification_report(ytest, y_pred))
print(confusion matrix(y_test, y_pred))
print(accuracy_score(y_test,y_pred))
#printing the confusion matrix
LABELS = ['Notfraud', 'fraud']
conf_matrix = confusion_matrix(y_test, y_pred)
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()
```

5. RESULT

Random Forest:

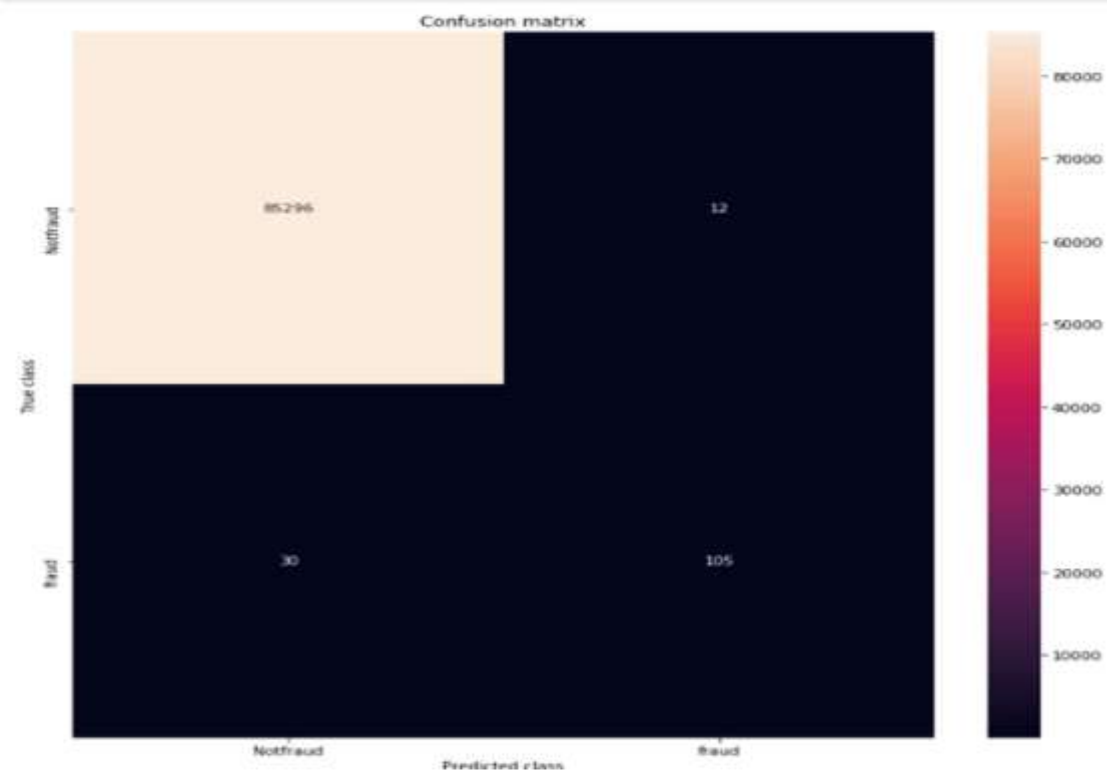


Fig 5.1: Random Forest Confusion Matrix

Logistic Regression:

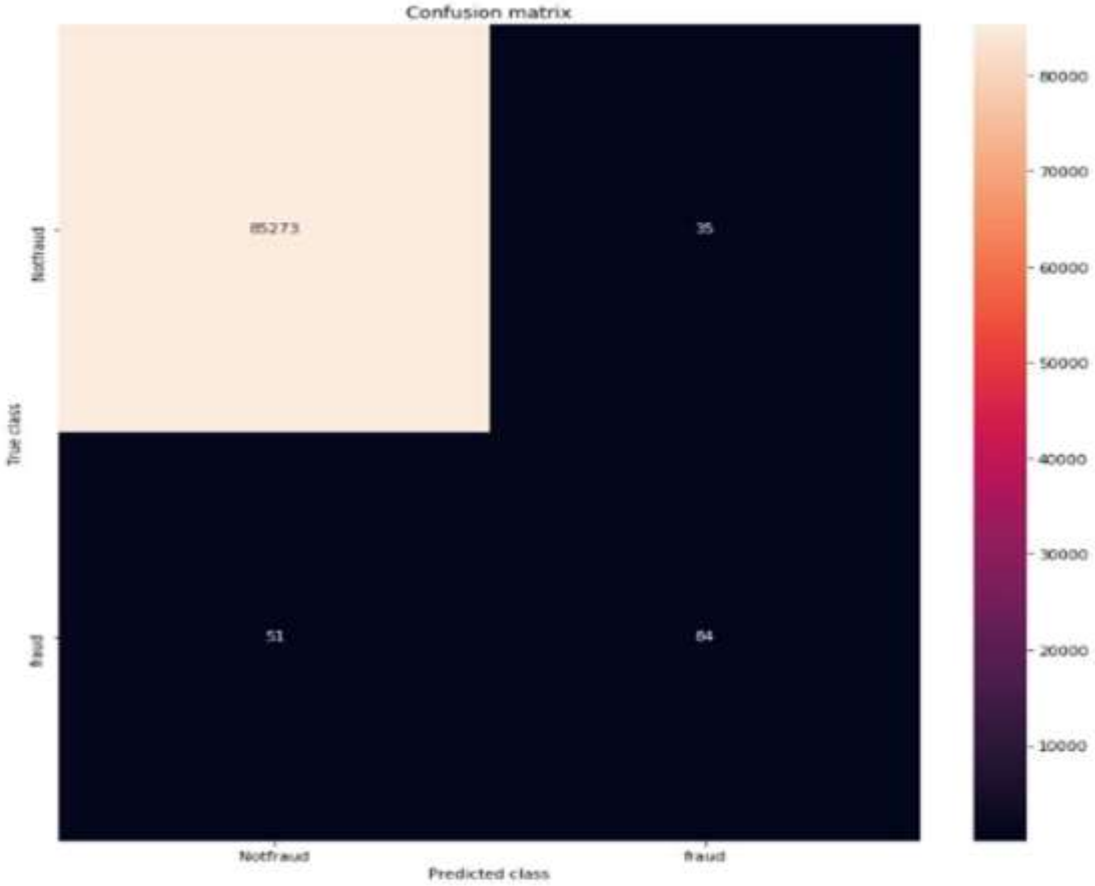
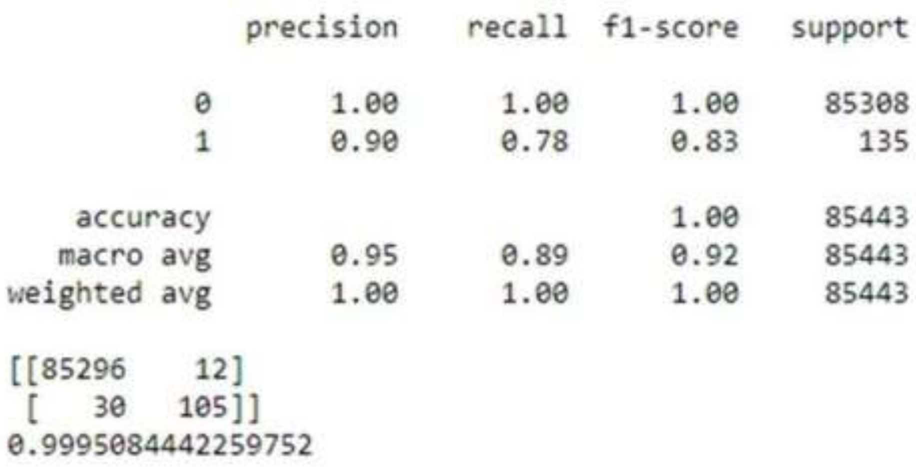


Fig 5.2: Logistic Regression Confusion Matrix

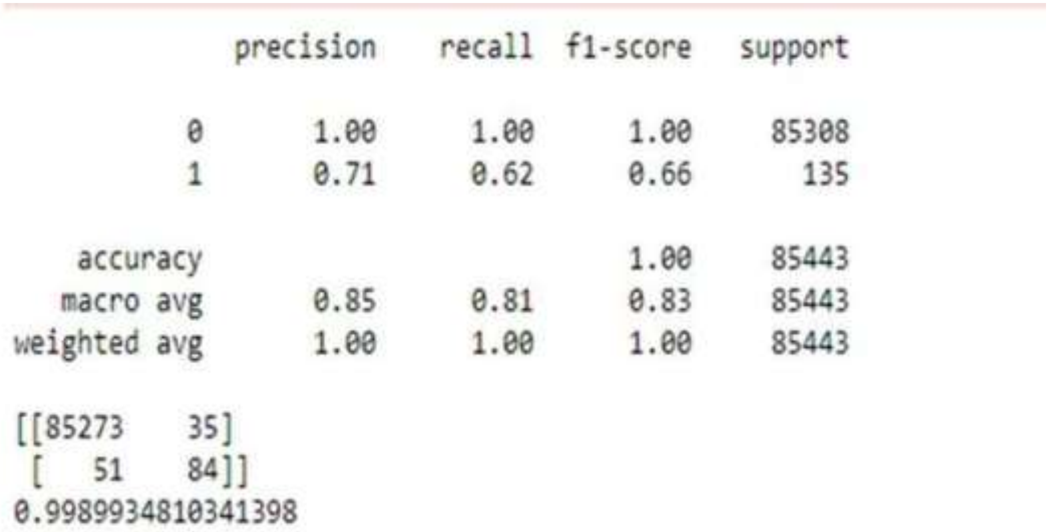
Random Forest:



```
              precision    recall  f1-score   support

           0       1.00      1.00      1.00     85308
           1       0.90      0.78      0.83       135

    accuracy                           1.00     85443
   macro avg       0.95      0.89      0.92     85443
weighted avg       1.00      1.00      1.00     85443

[[85296    12]
 [   30   105]]
0.9995084442259752
```

Fig 5.3: Random Forest Accuracy

Logistic Regression:



```
              precision    recall  f1-score   support

           0       1.00      1.00      1.00     85308
           1       0.71      0.62      0.66       135

    accuracy                           1.00     85443
   macro avg       0.85      0.81      0.83     85443
weighted avg       1.00      1.00      1.00     85443

[[85273    35]
 [   51    84]]
0.9989934810341398
```

Fig 5.4: Logistic Regression Accuracy

6. CONCLUSION

- This system is capable of providing most of the essential features required to detect fraudulent and legitimate transactions
- Although there are several fraud detection techniques available today but none is able to detect all frauds completely when they are actually happening, they usually detect it after the fraud has been committed. So we need a technology that can detect the fraudulent transaction when it is taking place so that it can be stopped then and there and that too in a minimum cost.
- The major drawback of all the techniques is that they are not guaranteed to give the same results in all environments. They give better results with a particular type of dataset and poor or unsatisfactory results with other.
- The Random forest algorithm will perform better with a larger number of training data.

7. FUTURE SCOPE

Further enhancement can be made by making this system secure by using both merchant and customer certificates and by adding new checks as technology changes can be added to know the pattern of fraudulent transactions and alert the respective card holders and bankers when fraud activity is identified. The dataset available on day to day processing may become outdated, it is compulsory to have updated data for effective fraud behaviour identification. To this extent, the incremental approach is necessary in making the system to learn from past as well as present data and capable of handling the both. Fraudster uses different new techniques that are instantaneously growing along with new technology makes it difficult for detection. Also the nature of access pattern may vary from one geographical location to another (such as urban and rural areas) that may result in a false positive detection. In such a case a future enhancement may be based on new multiple models with varying access pattern needs attention to improve the effectiveness. Privacy preserving techniques applied in distributed environment resolves the security related issues preventing private data access**.**

8. REFERENCES

1. [1] J. Vimala Devi and K. S. Kavitha, ―Fraud Detection in Credit Card Transactions by using Classification Algorithms,‖ Int. Conf. Curr. Trends Comput. Electr. Electron. Commun. CTCEEC 2017, pp. 125–131, 2018, doi: 10.1109/CTCEEC.2017.8455091.
2. [2]R. R. Popat and J. Chaudhary, ―A Survey on Credit Card Fraud Detection Using Machine Learning,‖ Proc. 2nd Int. Conf. Trends Electron. Informatics, ICOEI 2018, no. Icoei, pp. 1120–1125,2018, DOI: 10.1109/ICOEI.2018.8553963.
3. [3]A. RB and S. K. KR, ―Credit Card Fraud Detection Using Artificial Neural Network,‖ Glob. Transitions Proc., pp. 0–8, 2021, doi: 10.1016/j.gltp.2021.01.006.
4. [4]S. Mittal and S. Tyagi, ―Performance evaluation of machine learning algorithms for credit card fraud detection,‖ Proc. 9th Int. Conf. Cloud Comput. Data Sci. Eng. Conflu. 2019, pp. 320–324, 2019, doi: 10.1109/CONFLUENCE.2019.8776925
5. [5]D. D. Borse, P. S. H. Patil, and S. Dhotre, ―Credit Card Fraud Detection Using Naïve Bayes and C4,‖ vol. 10, no. 1, pp. 423–429, 2021.
6. [6] Suman Arora , "Selection of Optimal Credit Card Fraud Detection Models Using a Coefficient Sum Approach" , International Conference on Computing, Communication and Automation (ICCCA2017), pp 482 - 487, 2017.
7. [7] Kosemani Temitayo Hafiz, Dr. Shaun Aghili and Dr. Pavol Zavarsky, “The Use of Predictive Analytics Technology to Detect Credit Card Fraud in Canada”.
8. Machine Learning Decision Tree Classification Algorithm
   - Javatpoint.‖
9. https://www.javatpoint.com/machinelearning-decision-tree-classification-algorithm
10. Machine Learning Random Forest Algorithm - Javatpoint.‖
    https://www.javatpoint.com/machinelearningrandom-forest-algorithm