A Mini Project Report

On

CREDIT CARD FRAUD DETECTION USING MACHINE LEARNING

Submitted in partial fulfillment of the requirement of
University of Mumbai for the Degree of

Bachelor of Engineering
In
Computer Engineering

Submitted By
Nikhil Mathapati
Ashwin Nair
Deepanshu Pandita
Yuvraj Patel

Supervisor
Prof. Payel Thakur

Department of Computer Engineering

PILLAI COLLEGE OF ENGINEERING

New Panvel – 410 206

UNIVERSITY OF MUMBAI

Academic Year 2020-21

DEPARTMENT OF COMPUTER ENGINEERING
Pillai College of Engineering
New Panvel – 410 206

CERTIFICATE

This is to certify that the requirements for the TE Mini Project Report entitled '**CREDIT CARD FRAUD DETECTION USING MACHINE LEARNING**' have been successfully completed by the following students:

Name	**Roll No.**
Nikhil Mathapati	A631
Ashwin Nair	A638
Deepanshu Pandita	A643
Yuvraj Patel	A648

in partial fulfillment of Bachelor of Engineering of Mumbai University in the Department of Computer Engineering, Pillai College of Engineering, New Panvel – 410 206 during the Academic Year 2020– 2021.

_____

Supervisor

(Prof. Payel Thakur)

_____ _____

Head of Department **Principal**

Dr. Sharvari Govilkar **Dr. Sandeep M. Joshi**

DEPARTMENT OF COMPUTER ENGINEERING
Pillai College of Engineering
New Panvel – 410 206

PROJECT REPORT APPROVAL FOR T.E

This TE Mini Project Report entitled "**CREDIT CARD FRAUD DETECTION USING MACHINE LEARNING**" by **Nikhil Mathapati, Ashwin Nair, Deepanshu Pandita and Yuvraj Patel** are approved for the degree of T.E. in **Computer Engineering**.

Examiners:

1. _____

2. _____

Supervisors:

1. _____

2. _____

Chairman:

1. _____

Date:

Place:

Declaration

We declare that this written submission for TE Mini Project Report Declaration entitled "**CREDIT CARD FRAUD DETECTION USING MACHINE LEARNING**" represents our ideas in our own words and where others' ideas or words have been included. We have adequately cited and referenced the original sources. We also declared that we have adhered to all principles of academic honesty and integrity and have not misrepresented or fabricated or falsified any ideas / data / fact / source in our submission. We understand that any violation of the above will cause disciplinary action by the institute and also evoke penal action from the sources which have thus not been properly cited or from whom paper permission have not been taken when needed.

Project Group Members:

Nikhil Mathapati & Sign: _____

Ashwin Nair & Sign: _____

Deepanshu Pandita & Sign: _____

Yuvraj Patel & Sign: _____

Date:

Place:

Table Contents

Abstract

The advancement of new technologies and the fast growing of technological development have generated new possibilities as well as imposing new challenges. Fraud, the biggest challenges for business and organization, emerge with new technologies to take new and distinctive forms that are hidden and tougher to identify than the conventional forms of this crime. Credit card frauds also grow up along with growing in technology. Credit card fraud detection is presently the most frequently occurring problem in the present world. This is due to the rise in both online transactions and e-commerce platforms. Credit card fraud generally happens when the card was stolen for any of the unauthorized purposes or even when the fraudster uses the credit card information for his use. In the present world, we are facing a lot of credit card problems. To detect the fraudulent activities the credit card fraud detection system was introduced. In this project, machine learning algorithms are used to detect credit card fraud. In this project,we are trying to balance the dataset then machine learning algorithms are used to detect credit card fraud. Standard models are used like the Logistic regression, decision tree,Random Forest algorithm. .To evaluate the model efficacy, a publicly available credit card data set is used. Then, a real-world credit card data set from a financial institution is analyzed. In addition, noise is added to the data samples to further assess the robustness of the algorithms.

List of Figure

Chapter 1

Introduction

Financial fraud is a growing concern with far reaching consequences in the government, corporate organizations, finance industry, In Today's world high dependency on internet technology has enjoyed increased credit card transactions but credit card fraud had also accelerated as online and offline transaction. As credit card transactions become a widespread mode of payment, focus has been given to recent computational methodologies to handle the credit card fraud problem.. Data mining technique is one notable and popular methods used in solving credit fraud detection problem. It is impossible to be sheer certain about the true intention and rightfulness behind an application or transaction. In reality, to seek out possible evidences of fraud from the available data using mathematical algorithms is the best effective option. Fraud detection in credit card is the truly the process of identifying those transactions that are fraudulent into two classes of legit class and fraud class transactions, several techniques are designed and implemented to solve to credit card fraud detection such as comparative analysis of logistic regression,decision tree and random forest is carried out. Credit card fraud detection is a very popular but also a difficult problem to solve. Firstly, due to the issue of having only a limited amount of data, credit cards make it challenging to match a pattern for a dataset. Secondly, there can be many entries in the dataset with truncations of fraudsters which also fit a pattern of legitimate behavior. Also the problem has many constraints. Firstly, data sets are not easily accessible for public and the results of researches are often hidden and censored, making the results inaccessible and due to this it is challenging to benchmarking for the models built. Datasets in previous researches with real data in the literature are nowhere mentioned. Secondly, the improvement of methods is more difficult by the fact that the security concern imposes an limitation to exchange of ideas and methods in fraud detection, and especially in credit card fraud detection. Lastly, the data sets are continuously evolving and changing making the profiles of normal and fraudulent behaviors always different. That is the legit transaction in the past may be a fraud in present or vice versa. Decision tree,Logistic regression and random forests and then a collative comparison is made to evaluate which model performed best. Credit card transaction datasets are rarely available, highly imbalanced and skewed. Optimal feature

(variables) selection for the models, suitable metric is the most important part of data mining to evaluate performance of techniques on skewed credit card fraud data.

1.1 Principle

The purpose of this project is to detect the credit card frauds as 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.

1.2 Objective

1. The objectives of the project is to implement machine learning algorithms to detect credit card fraud detection with respect to time and amount of transaction.
2. We can find the most accurate detection using this technique which reduces the tedious work of an employee in the bank .

1.3 Scope

The system prevents fraudulent users from misusing the details of the credit-card of the genuine users for their personal gain. The spending habits of the credit-card owner is to detect the fraud. As the fake user might not be aware of the spending habits of the owner, there will be an irregularity in the spending pattern, which the system will detect.. Thus, the system protects legitimate users from financial loss.

The system helps in making electronic payment safer and more reliable. The principles in the proposed system can also be adopted and implemented in other electronic payment services such as online banking facility and payment gateways.

Chapter 2

Literature Survey

In [1] This paper represents an research about a case study involving credit card fraud detection, where data normalization is applied before Cluster Analysis and with results obtained from the use of Cluster Analysis and Artificial Neural Networks on fraud detection has shown that by clustering attributes neuronal inputs can be minimized. And promising results can be obtained by using normalized data and data should be MLP trained. This research was based on unsupervised learning. Significance of this paper was to find new methods for fraud detection and to increase the accuracy of results.

In [2] In this paper a new collative comparison measure that reasonably represents the gains and losses due to fraud detection is proposed. A cost sensitive method which is based on Bayes minimum risk is presented using the proposed cost measure. Improvements up to 23% is obtained when this method and other state of art algorithms are compared. The data set for this paper is based on real life transactional data by a large European company and personal details in data is kept confidential., accuracy of an algorithm is around 50%. Significance of this paper was to find an algorithm and to reduce the cost measure. The result obtained was by 23% and the algorithm they find was Bayes minimum risk.

In [3] Various modern techniques based on Sequence Alignment, Machine learning, Artificial Intelligence, Genetic Programming, Data mining etc. has been evolved and is still evolving to detect fraudulent transactions in credit card. A sound and clear understanding on all these approaches is needed that will certainly lead to an efficient credit card fraud detection system. Survey of various techniques used in credit card fraud detection mechanisms has been Shown in this paper along with evaluation of each methodology based on certain design criteria. Analysis on Credit Card Fraud Detection Methods has been done. The survey in this paper was purely based to detect the efficiency and transparency of each method. Significance of this paper was conduct a survey to compare different credit card fraud detection algorithm to find the most suitable algorithm to solve the problem. International Journal of Pure and Applied Mathematics Special Issue 828

In [4] A comparison has been made between models based on artificial intelligence along with general description of the developed fraud detection system are given in this paper such as the Naive Bayesian Classifier and the model based on Bayesian Networks, the clustering model. And in the end conclusions about results of models' evaluative testing are made. Number of legal truncations was determined greater or equal to 0.65 that is their accuracy was 65% using Bayesian Network. Significance of this paper is to compare between models based on artificial intelligence along with general description of the developed system and to state the accuracy of each model along with the recommendation to make the better model.

Chapter 3

Implemented System

3.1 Overview

The credit card fraud detection features uses user behavior and location scanning to check for unusual patterns. These patterns include user characteristics such as user spending patterns as well as usual user geographic locations to verify his identity. If any unusual pattern is detected, the system requires revivification.

The system analyses user credit card data for various characteristics. These characteristics include user country, usual spending procedures. Based upon previous data of that user the system recognizes unusual patterns in the payment procedure. So now the system may require the user to login again or even block the user for more than 3 invalid attempts.

Core Features:

1. The system stores previous transaction patterns for each user.
2. Based upon the user spending ability and even country, it calculates the user's characteristics.
3. More than 20 -30 %deviation of users transaction(spending history and operating country) is considered as an invalid attempt and the system takes action.

3.1.1 Proposed System

The proposed techniques are used in this paper, for detecting the frauds in credit card System. The comparison are made for different machine learning algorithms such as Logistic Regression, Decision Trees, Random Forest, to determine which algorithm gives suits best and can be adapted by credit card merchants for identifying fraud transactions.

The processing steps are discussed below to detect the best algorithm for the given dataset

Algorithm steps:

Step 1:Read the dataset.

Step 2: Random Sampling is done on the data set to make it balanced.

Step 3: Divide the dataset into two parts i.e., Train dataset and Test dataset.

Step 4: Feature selection is applied for the proposed models.

Step 5:Accuracy and performance metrics has been calculated to know the efficiency for different algorithms.

Step6: Then retrieve the best algorithm based on efficiency for the given dataset

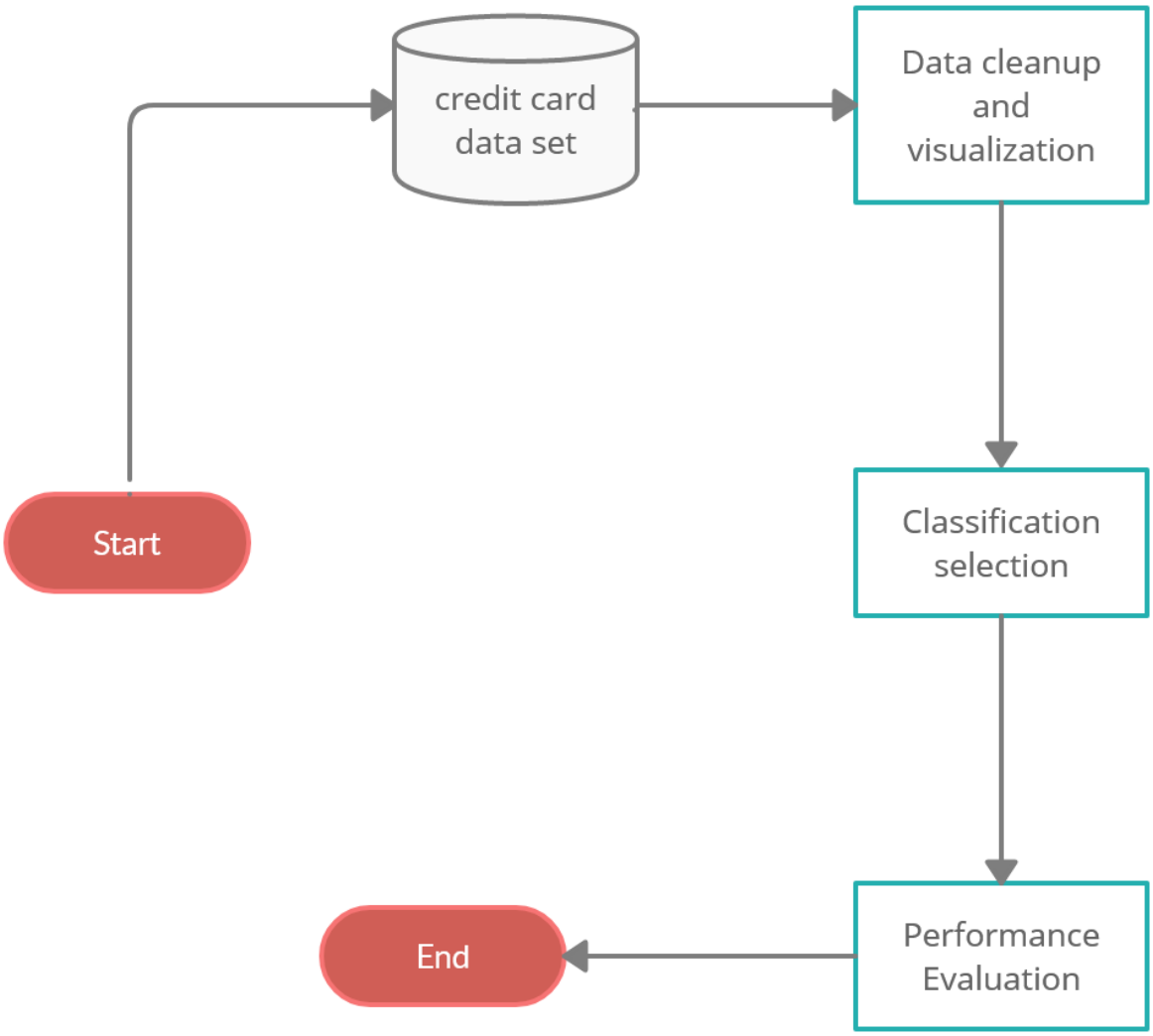

Fig 3.1.2: System Architecture

3.2 Implementation Details

3.2.1 Algorithms

3.2.1.1 Random Sampling

Random sampling is a part of the sampling technique in which each sample has an equal probability of being chosen. A sample chosen randomly is meant to be an unbiased representation of the total population. If for some reasons, the sample does not represent the population, the variation is called a sampling error.

Random sampling is one of the simplest forms of collecting data from the total population. Under random sampling, each member of the subset carries an equal opportunity of being chosen as a part of the sampling process. For example, the total workforce in organisations is 300 and to conduct a survey, a sample group of 30 employees is selected to do the survey. In this case, the population is the total number of employees in the company and the sample group of 30 employees is the sample. Each member of the workforce has an equal opportunity of being chosen because all the employees which were chosen to be part of the survey were selected randomly. But, there is always a possibility that the group or the sample does not represent the population as a whole, in that case, any random variation is termed as a sampling error.

3.2.1.2 Logistic Regression

Logistic regression is a supervised learning classification algorithm used to predict the probability of a target variable. The nature of the target or dependent variable is dichotomous, which means there would be only two possible classes. In simple words, the dependent variable is binary in nature having data coded as either 1 (stands for success/yes) or 0 (stands for failure/no).Mathematically, a logistic regression model predicts P(Y=1) as a function of X. It is one of the simplest ML algorithms that can be used for various classification problems such as spam detection, Diabetes prediction, cancer detection etc.

Types of Logistic Regression

Generally, logistic regression means binary logistic regression having binary target variables, but there can be two more categories of target variables that can be predicted by it. Based on those number of categories, Logistic regression can be divided into following types −

Binary or Binomial

In such a kind of classification, a dependent variable will have only two possible types either 1 and 0. For example, these variables may represent success or failure, yes or no, win or loss etc.

Multinomial

In such a kind of classification, dependent variable can have 3 or more possible unordered types or the types having no quantitative significance. For example, these variables may represent “Type A” or “Type B” or “Type C”.

Ordinal

In such a kind of classification, dependent variables can have 3 or more possible ordered types or the types having a quantitative significance. For example, these variables may represent “poor” or “good”, “very good”, “Excellent” and each category can have scores like 0,1,2,3.

3.2.1.3 Decision Tree

Decision trees are the most powerful and popular tool for classification and prediction. A Decision tree is a flowchart like tree structure, where each internal node denotes a test on an attribute, each branch represents an outcome of the test, and each leaf node (terminal node) holds a class label.

Construction of Decision Tree :

A tree can be “learned” by splitting the source set into subsets based on an attribute value test. This process is repeated on each derived subset in a recursive manner called recursive partitioning. The recursion is completed when the subset at a node all has the same value of the target variable, or when splitting no longer adds value to the predictions. The construction of a decision tree classifier does not require any domain knowledge or parameter setting, and therefore is appropriate for exploratory knowledge discovery. Decision trees can handle high dimensional data. In general, decision tree classifiers have good accuracy. Decision tree induction is a typical inductive approach to learn knowledge on classification

Decision Tree Representation :

Decision trees classify instances by sorting them down the tree from the root to some leaf node, which provides the classification of the instance. An instance is classified by starting at the root node of the tree,testing the attribute specified by this node,then moving down the tree branch corresponding to the value of the attribute as shown in the above figure.This process is then repeated for the subtree rooted at the new node.

3.2.1.4 Random Forest Algorithm

Random Forest is a popular machine learning algorithm that belongs to the supervised learning technique. It can be used for both Classification and Regression problems in ML. It is based on the concept of ensemble learning, which is a process of combining multiple classifiers to solve a complex problem and to improve the performance of the model.

As the name suggests, "Random Forest is a classifier that contains a number of decision trees on various subsets of the given dataset and takes the average to improve the predictive accuracy of that dataset." Instead of relying on one decision tree, the random forest takes the prediction from each tree and based on the majority votes of predictions, and it predicts the final output.

The greater number of trees in the forest leads to higher accuracy and prevents the problem of overfitting.

The below diagram explains the working of the Random Forest algorithm:

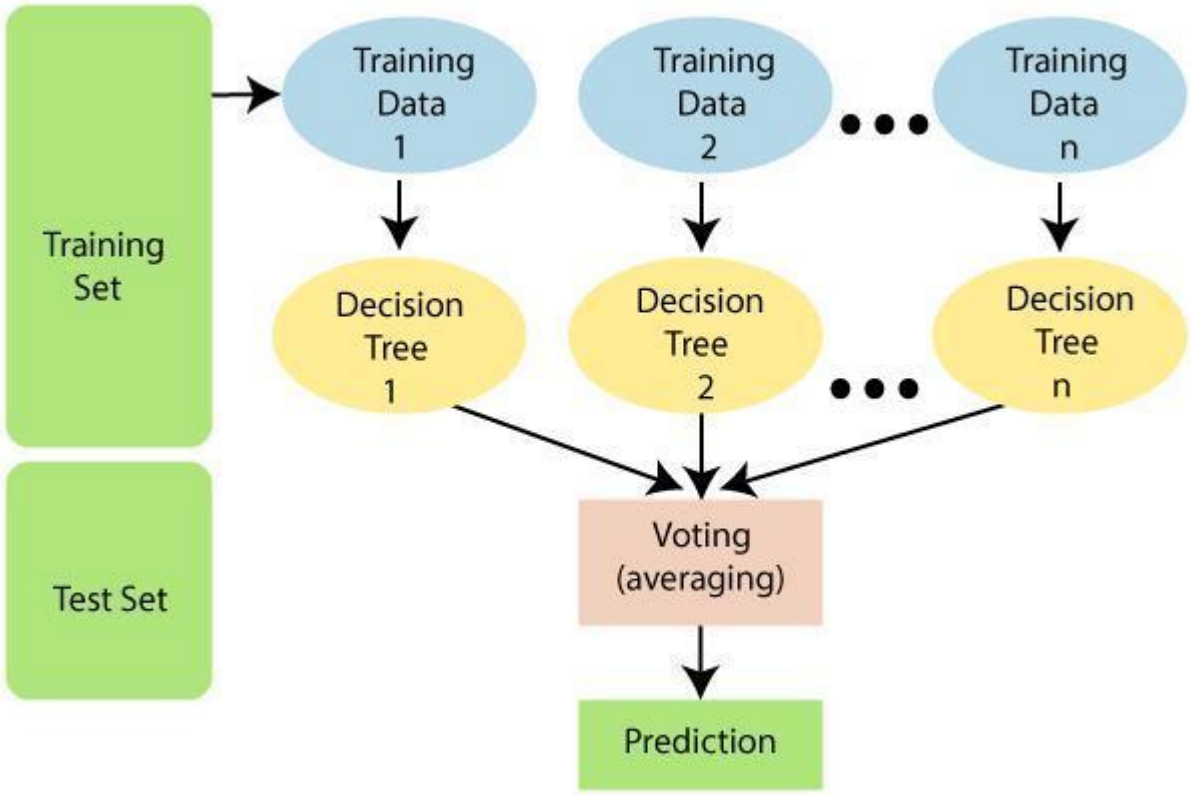

Fig 3.2.1.4:Working Of Random Forest Algorithm

3.2 Hardware and Software Specification

Software Requirements

Language : Python

OS : Windows 10 (64 bit)

IDE : jupyter notebook

Hardware Requirements

Processor : Above 1.5GHZ

Hard Disk : 80GB

RAM : 2G

Chapter 4

Result and Discussion

4.1 Testing

Software testing is an investigation conducted to provide stakeholders with information about thequality of the product or service under test. Software testing can also provide an objective,independent view of the software to allow the business to appreciate and understand the risks of

software implementation. Test techniques include the process of executing a program or application with the intent of finding software bugs (errors or other defects). Software testing involves the execution of a software component or system component to evaluate one or more properties of interest. In general, these properties indicate the extent to which the component or system under test:

1]meets the requirements that guided its design and development,

2]responds correctly to all kinds of inputs,

3]performs its functions within an acceptable time,

4]is sufficiently usable,

5]can be installed and run in its intended environments, and

6]Achieves the general result its stakeholder‟s desire.

As the number of possible tests for even simple software components is practically infinite, all

software testing uses some strategy to select tests that are feasible for the available time and

resources. As a result, software testing typically (but not exclusively) attempts to execute a program or application with the intent of finding software bugs (errors or other defects)[7]. The job of testing is an iterative process as when one bug is fixed; it can illuminate other, deeper

bugs, or can even create new ones. Software testing can provide objective, independent information about the quality of software and risk of its failure to users and/or sponsors. Software testing can be conducted as soon as executable software (even if partially complete) exists. The overall approach to software development often determines when and how testing is conducted. For example, in a phased process, most testing occurs after system requirements have been defined and then implemented in testable programs. In contrast, under an Agile approach, requirements, programming, and testing are often done concurrently.

4.2 Snapshots

	Time	V1	V2	V3	V4	V5	V6	V7	V8	V9	...	V21	V22	V23	V24	
0	0.0	-1.359807	-0.072781	2.536347	1.378155	-0.338321	0.462388	0.239599	0.098698	0.363787	...	-0.018307	0.277838	-0.110474	0.066928	0.1
1	0.0	1.191857	0.266151	0.166480	0.448154	0.060018	-0.082361	-0.078803	0.085102	-0.255425	...	-0.225775	-0.638672	0.101288	-0.339846	0.1
2	1.0	-1.358354	-1.340163	1.773209	0.379780	-0.503198	1.800499	0.791461	0.247676	-1.514654	...	0.247998	0.771679	0.909412	-0.689281	-0.3
3	1.0	-0.966272	-0.185226	1.792993	-0.863291	-0.010309	1.247203	0.237609	0.377436	-1.387024	...	-0.108300	0.005274	-0.190321	-1.175575	0.6
4	2.0	-1.158233	0.877737	1.548718	0.403034	-0.407193	0.095921	0.592941	-0.270533	0.817739	...	-0.009431	0.798278	-0.137458	0.141267	-0.2
5	2.0	-0.425966	0.960523	1.141109	-0.168252	0.420987	-0.029728	0.476201	0.260314	-0.568671	...	-0.208254	-0.559825	-0.026398	-0.371427	-0.2
6	4.0	1.229658	0.141004	0.045371	1.202613	0.191881	0.272708	-0.005159	0.081213	0.464960	...	-0.167716	-0.270710	-0.154104	-0.780055	0.7
7	7.0	-0.644269	1.417964	1.074380	-0.492199	0.948934	0.428118	1.120631	-3.807864	0.615375	...	1.943465	-1.015455	0.057504	-0.649709	-0.4
8	7.0	-0.894286	0.286157	-0.113192	-0.271526	2.669599	3.721818	0.370145	0.851084	-0.392048	...	-0.073425	-0.268092	-0.204233	1.011592	0.3
9	9.0	-0.338262	1.119593	1.044367	-0.222187	0.499361	-0.246761	0.651583	0.069539	-0.736727	...	-0.246914	-0.633753	-0.120794	-0.385050	-0.0

10 rows × 31 columns

V25	V26	V27	V28	Amount	Class
0.128539	-0.189115	0.133558	-0.021053	149.62	0
0.167170	0.125895	-0.008983	0.014724	2.69	0
-0.327642	-0.139097	-0.055353	-0.059752	378.66	0
0.647376	-0.221929	0.062723	0.061458	123.50	0
-0.206010	0.502292	0.219422	0.215153	69.99	0
-0.232794	0.105915	0.253844	0.081080	3.67	0
0.750137	-0.257237	0.034507	0.005168	4.99	0
-0.415267	-0.051634	-1.206921	-1.085339	40.80	0
0.373205	-0.384157	0.011747	0.142404	93.20	0
-0.069733	0.094199	0.246219	0.083076	3.68	0

Fig 4.2.1 Dataset

In these ten rows, based on the context of the problem, we can see that there are 28 main components (the V's), a time column (Time), a quantity attribute (Amount), which refers to the amount of the transaction and the column to predict (Class).The dataset does not contain empty values and, except for the Class column, which is of type integer, all other attributes are of type float. It is observed that the principal components have mean 0 and different standard deviations.

```
0    284315
1       492
Name: Class, dtype: int64
```

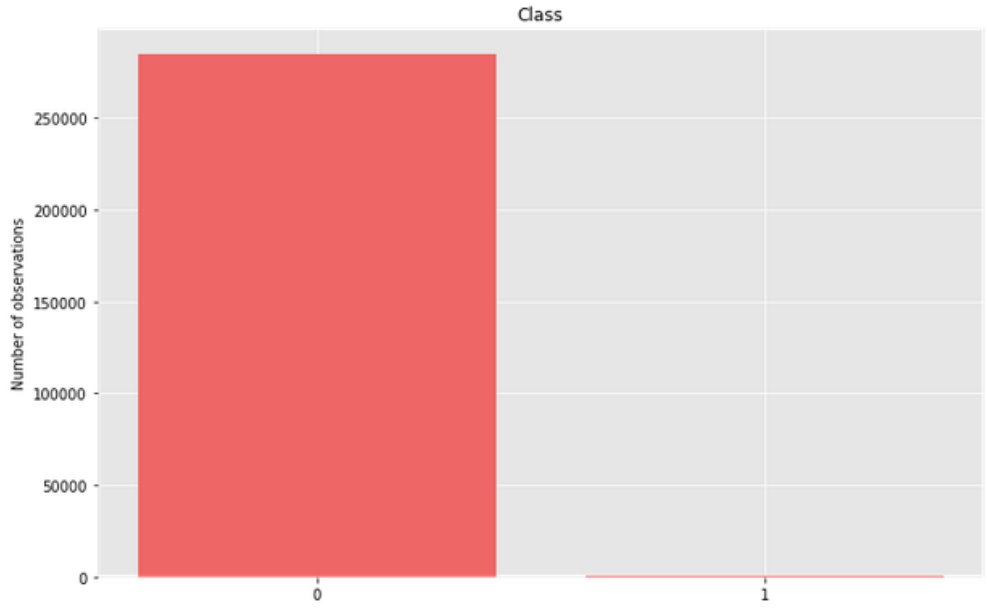

Fig 4.2.2: Data Imbalance

There is only 0.17% fraud. The imbalance is notoriously high!

Let's now look at some correlations:

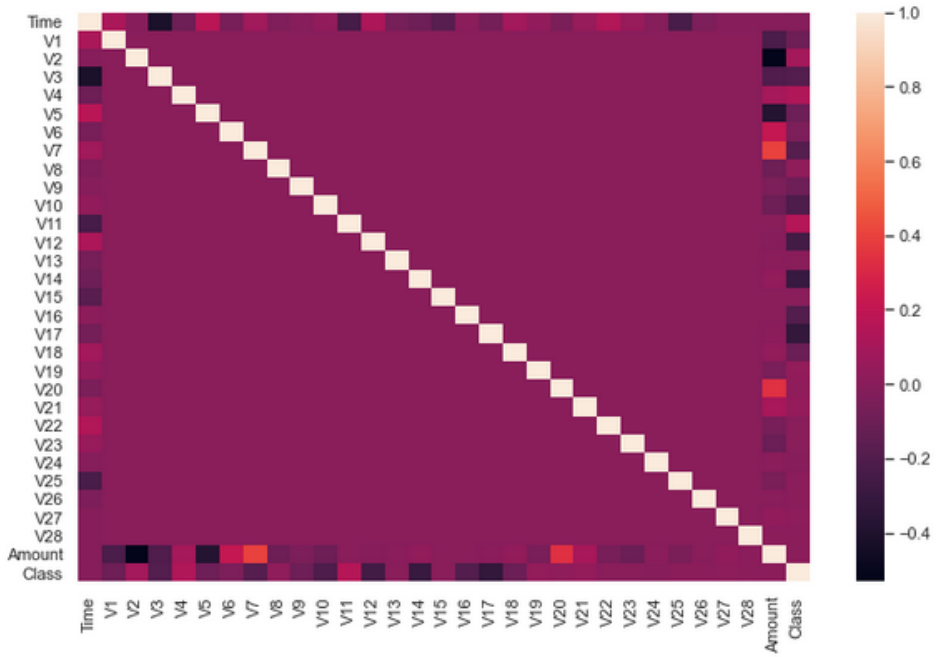

Fig4.2.3: Correlations

There are no strong linear correlations, it may be that there are other types of correlations. However, we will apply a classic data transformation: multiply some attributes and see if there is a stronger linear relationship.

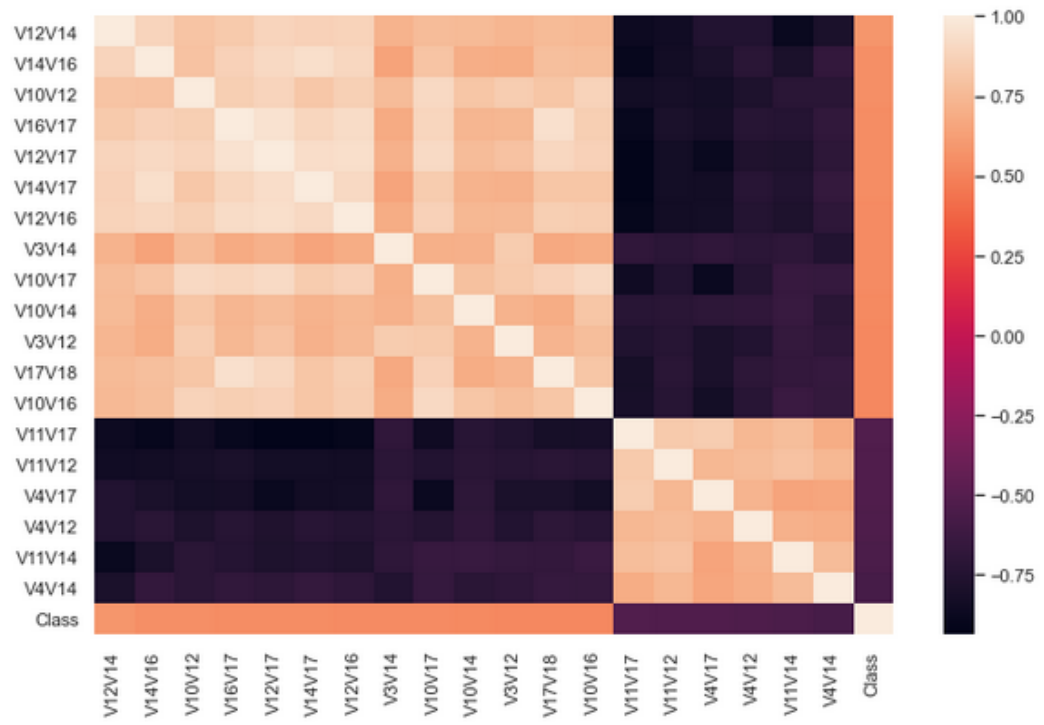

Fig 4.2.4: Linear correlations increased

Linear correlations increased! It should be mentioned that multiplication with decimal numbers tends to decrease values. For this reason, the correlation with respect to the Class attribute increases. If we carry out multiplication with three attributes, the results will continue to decrease, increasing the correlation. That is why only binary multiplication was chosen.

Preparing the data for Machine Learning algorithms

Using Random sampling, We have a new set of attributes with better linear correlation with respect to the Class attribute. In the next part a quantity of data is randomly extracted which is called "external_set". This data is never with the training set nor is it part of the test set. They are "new data" to verify the effectiveness of the algorithms. The training and test sets are obtained from the remaining set called "analysis_set".

```
Train set
 0    204724
1       338
Name: Class, dtype: int64
Test set
 0    51157
1      108
Name: Class, dtype: int64
External set
 0    28434
1       46
Name: Class, dtype: int64
```

Fig 4.2.5: Preparing the data for Machine Learning algorithms

Logistic Regression Model

Training a Logistic Regression Model

```
Train set
 [[201766   2958]
 [    33    305]]
              precision    recall  f1-score   support

     class 0      1.000     0.986     0.993    204724
     class 1      0.093     0.902     0.169       338

    accuracy                         0.985    205062
   macro avg      0.547     0.944     0.581    205062
weighted avg      0.998     0.985     0.991    205062
```

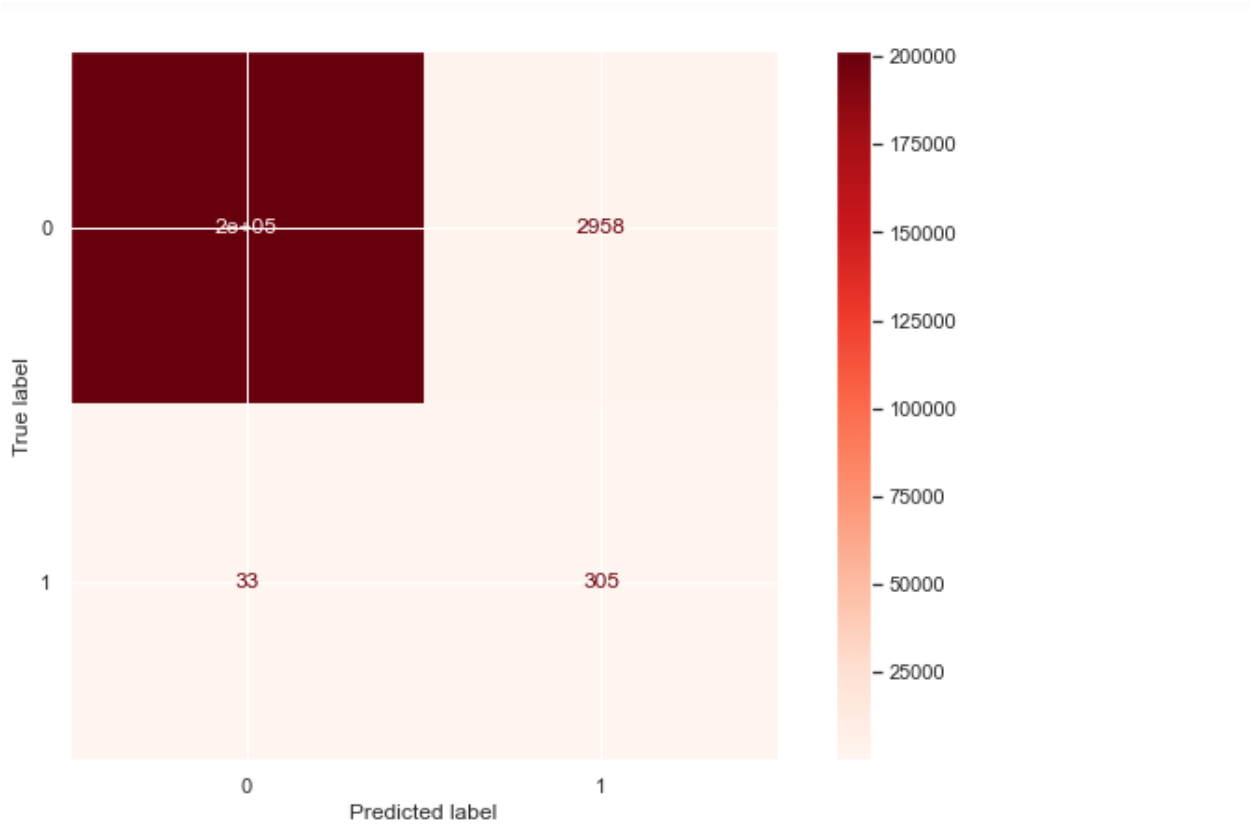

Fig 4.2.6:Training Of Logistic Regression Model

Here 33 valid transactions were detected fraudulent transactions and 2958 fraudulent transactions were detected valid transactions with recall value of 0.90 **[here recall means true positives/(true positives + false positives)]** on fraudulent transactions.

Predictions test

```
Test set
 [[50467   690]
 [   13    95]]
              precision    recall  f1-score   support

     class 0      1.000     0.987     0.993     51157
     class 1      0.121     0.880     0.213       108

    accuracy                          0.986     51265
   macro avg      0.560     0.933     0.603     51265
weighted avg      0.998     0.986     0.991     51265
```

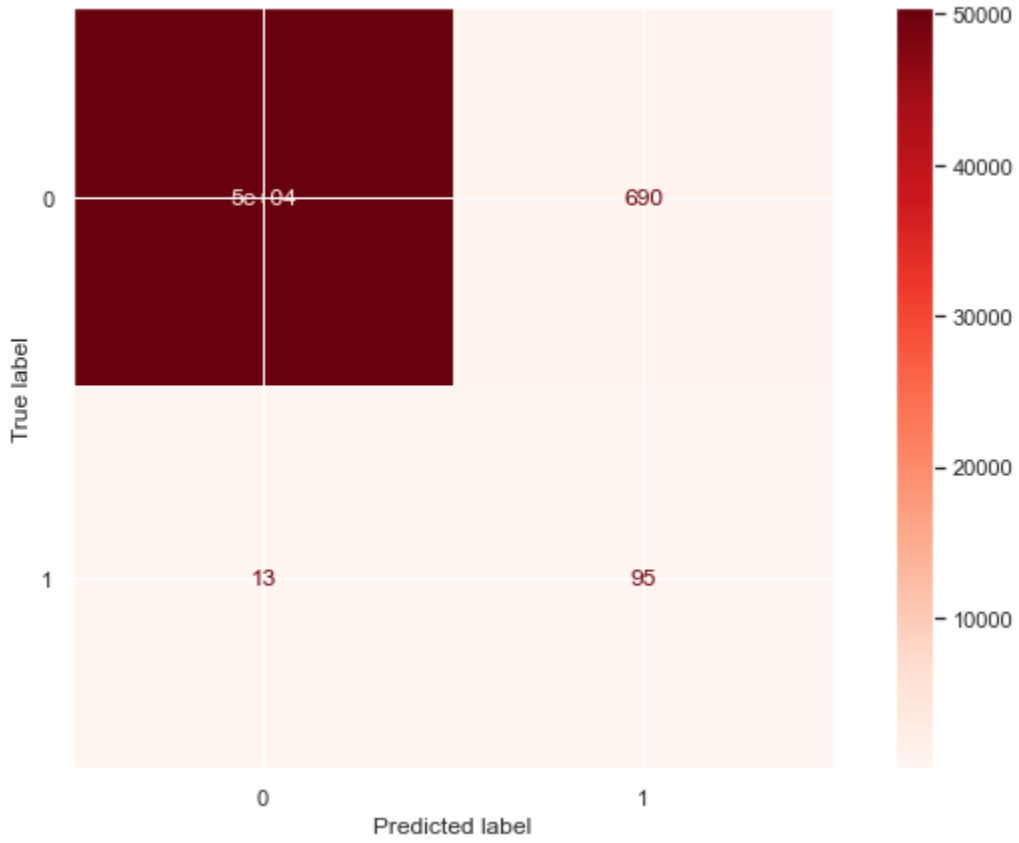

Fig 4.2.7:Predictive test Of Logistic Regression Model

Here 13 valid transactions were detected fraudulent transactions and 690 fraudulent transactions were detected valid transactions with recall value of 0.88 on fraudulent transactions.

External Data Set

```
External set
 [[28045   389]
 [    5    41]]
              precision    recall  f1-score   support

     class 0      1.000     0.986     0.993     28434
     class 1      0.095     0.891     0.172        46

    accuracy                         0.986     28480
   macro avg      0.548     0.939     0.583     28480
weighted avg      0.998     0.986     0.992     28480
```

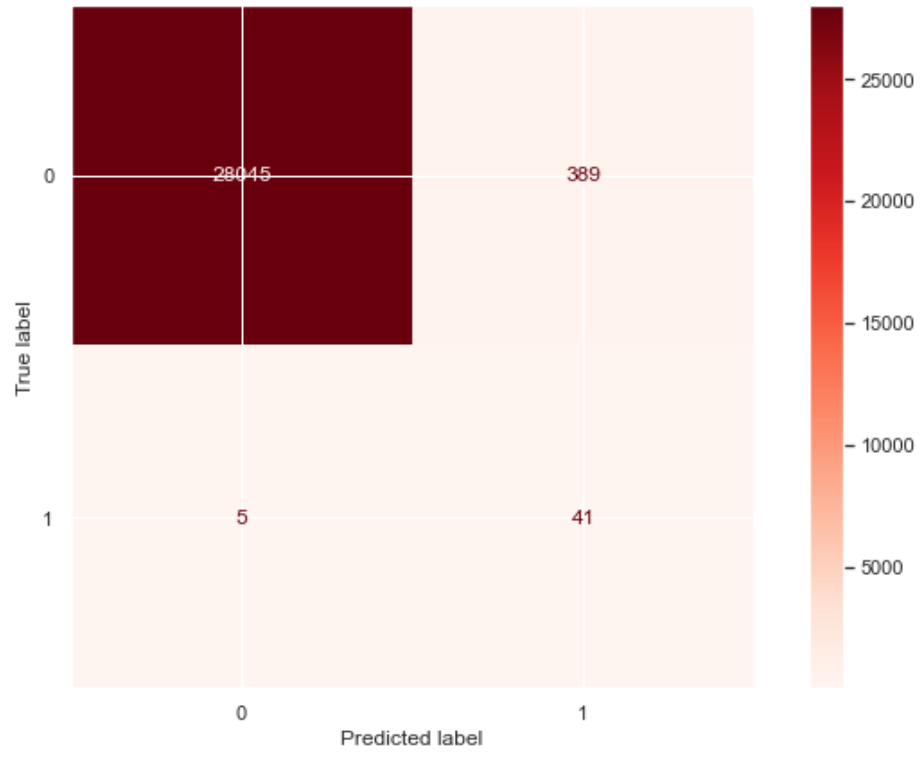

Fig 4.2.8: External data test Of Logistic Regression Model

Here 5 valid transactions were detected fraudulent transactions and 389 fraudulent transactions were detected valid transactions with recall value of 0.89 on fraudulent transactions.

Now if we average the recall values of train set , test set and external set we get the 0.89 recall value which is the accuracy of the logistic regression model which is quite good.

Decision Tree Classifier Model

Training a Decision Tree Classifier Model

```
Train set
 [[204724      0]
 [     0    338]]
              precision    recall  f1-score   support

     class 0      1.000     1.000     1.000    204724
     class 1      1.000     1.000     1.000       338

    accuracy                         1.000    205062
   macro avg      1.000     1.000     1.000    205062
weighted avg      1.000     1.000     1.000    205062
```

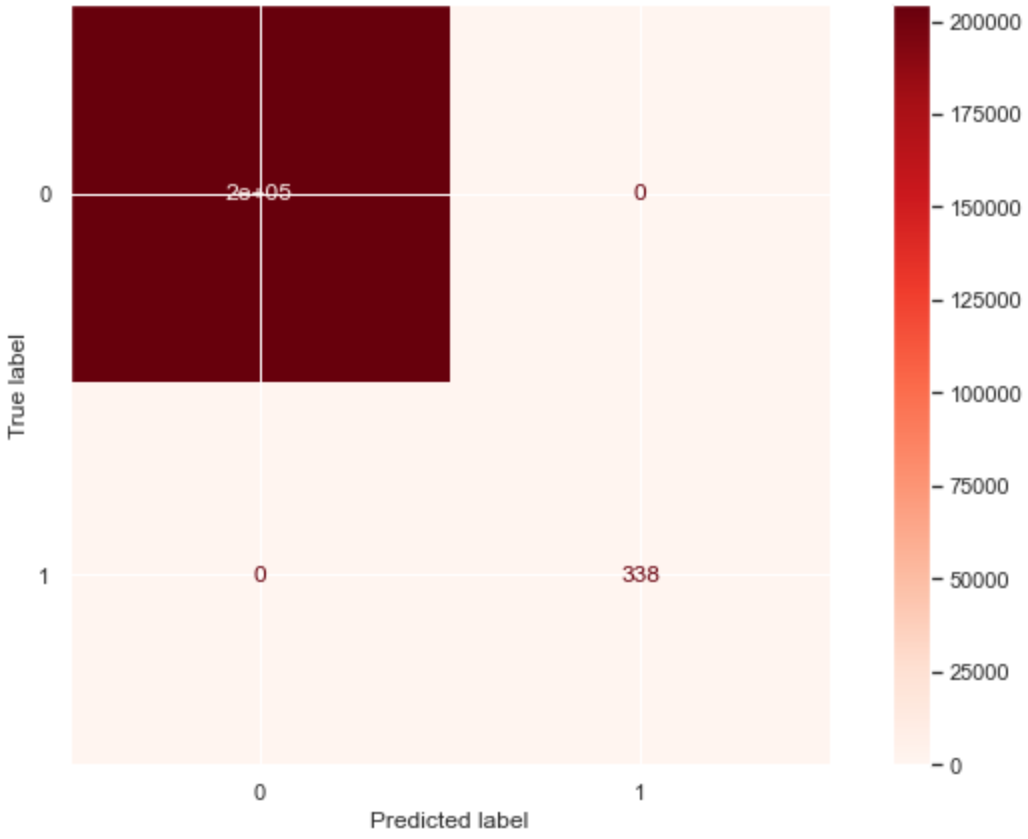

Fig 4.2.9:Training Of Decision Tree Classifier Model

Here 0 valid transactions were detected fraudulent transactions and 0 fraudulent transactions were detected valid transactions with recall value of 1.00 on fraudulent transactions.

Predictions test

```
Test set
 [[51143    14]
 [   27    81]]
              precision    recall  f1-score   support

     class 0      0.999     1.000     1.000     51157
     class 1      0.853     0.750     0.798       108

    accuracy                          0.999     51265
   macro avg      0.926     0.875     0.899     51265
weighted avg      0.999     0.999     0.999     51265
```

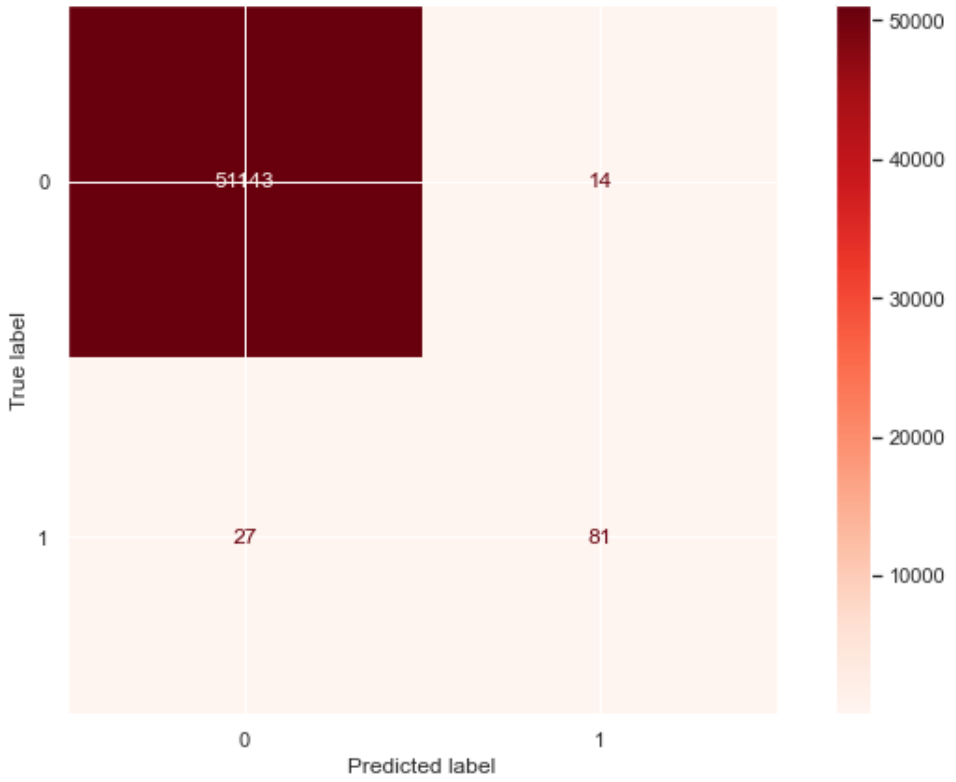

Fig 4.2.10 Predictive test Of Decision Tree Classifier Model

Here 27 valid transactions were detected fraudulent transactions and 14 fraudulent transactions were detected valid transactions with recall value of 0.75 on fraudulent transactions which is quite low from training set.

External Data Set

```
External set
 [[28427     7]
 [   13    33]]
              precision    recall  f1-score   support

     class 0      1.000     1.000     1.000     28434
     class 1      0.825     0.717     0.767        46

    accuracy                          0.999     28480
   macro avg      0.912     0.859     0.884     28480
weighted avg      0.999     0.999     0.999     28480
```

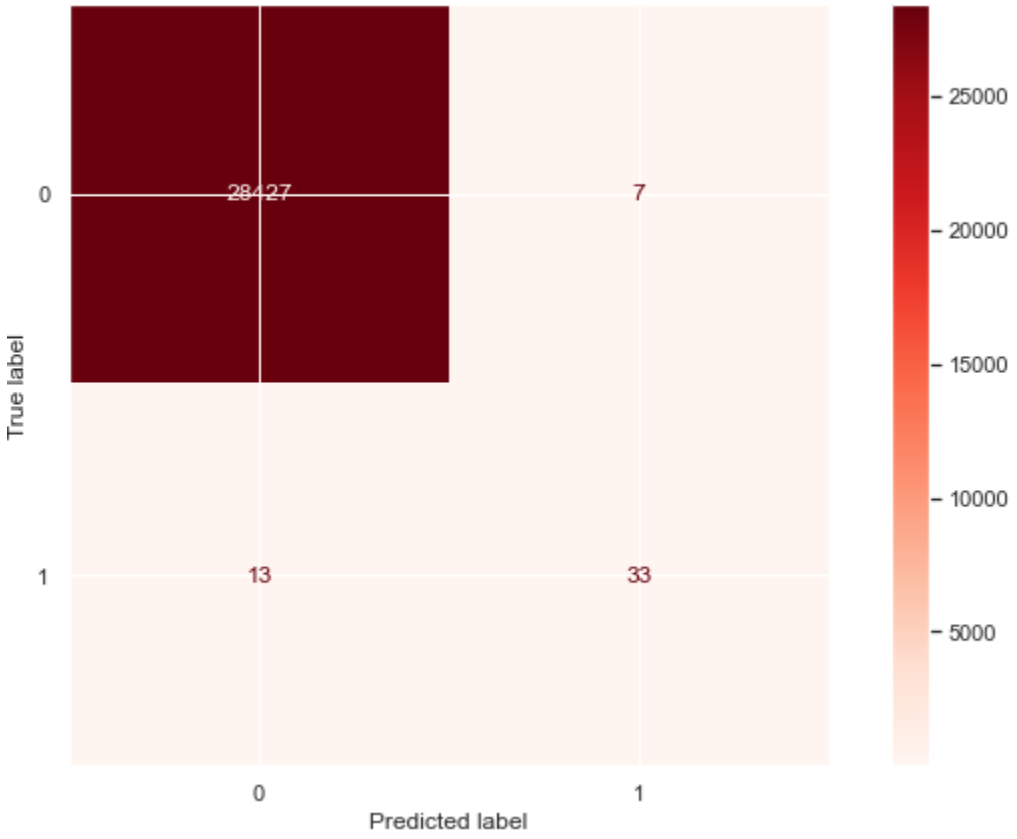

Fig 4.2.11:External data test Of Decision Tree Classifier Model

Here 13 valid transactions were detected fraudulent transactions and 7 fraudulent transactions were detected valid transactions with recall value of 0.71 on fraudulent transactions.

Now if we average the recall values of train set , test set and external set we get the 0.82 recall value which is the accuracy of the Decision Tree Classifier model which is not as good as logistic regression.

Random forest Classifier Model

Training a Random forest Classifier Model

```
Train set
 [[204724      0]
 [     0    338]]
              precision    recall  f1-score   support

     class 0      1.000     1.000     1.000    204724
     class 1      1.000     1.000     1.000       338

    accuracy                         1.000    205062
   macro avg      1.000     1.000     1.000    205062
weighted avg      1.000     1.000     1.000    205062
```

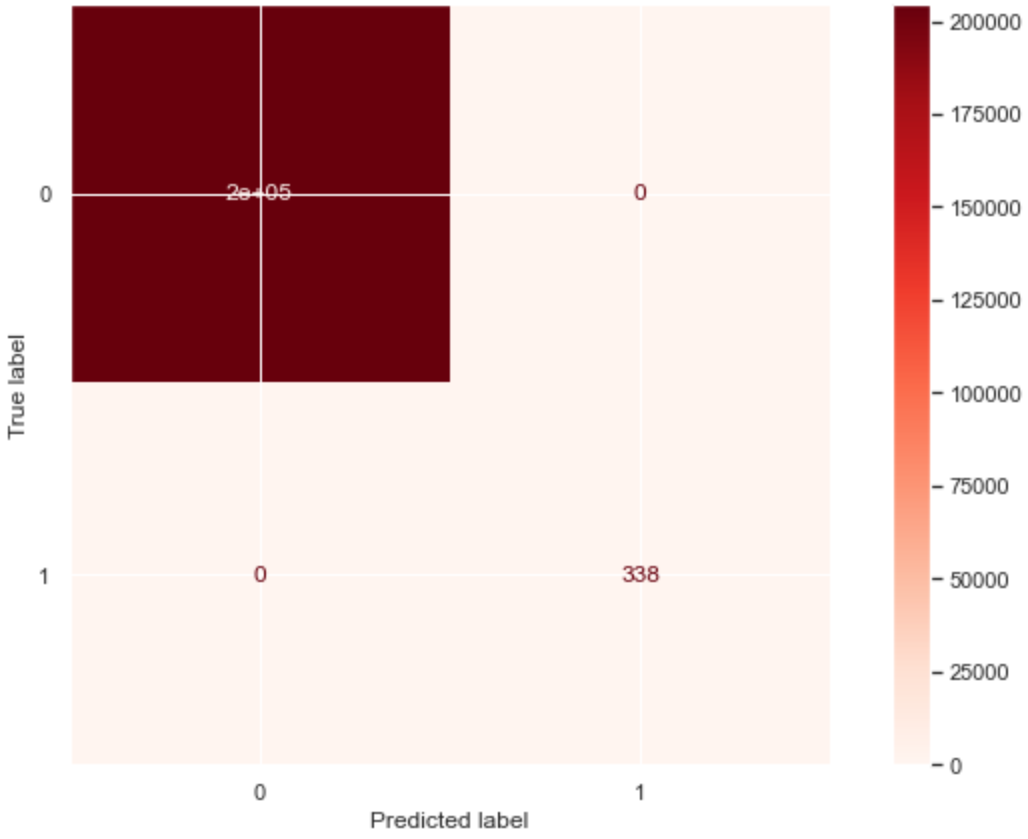

Fig 4.2.12:Training Of Random forest Classifier Model

Here 0 valid transactions were detected fraudulent transactions and 0 fraudulent transactions were detected valid transactions with recall value of 1.00 on fraudulent transactions.

Predictions test

```
Test set
 [[51152     5]
 [   23    85]]
              precision    recall  f1-score   support

     class 0      1.000     1.000     1.000     51157
     class 1      0.944     0.787     0.859       108

    accuracy                          0.999     51265
   macro avg      0.972     0.893     0.929     51265
weighted avg      0.999     0.999     0.999     51265
```

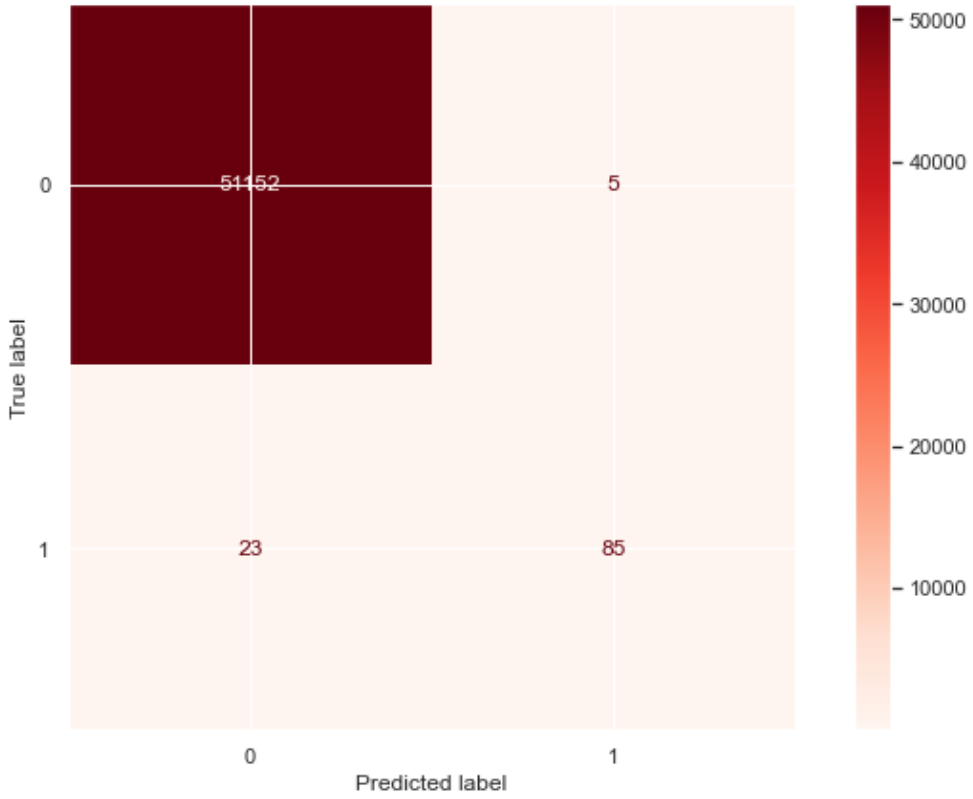

Fig 4.2.13:Predictive test Of Random forest Classifier Model

In this test set, the recall 0.79 (for fraud) is greater than that obtained in the decision tree model but less than that obtained in the logistic regression.

External Data Set

```
External set
 [[28434     0]
 [   11    35]]
              precision    recall  f1-score   support

     class 0      1.000     1.000     1.000     28434
     class 1      1.000     0.761     0.864        46

    accuracy                          1.000     28480
   macro avg      1.000     0.880     0.932     28480
weighted avg      1.000     1.000     1.000     28480
```

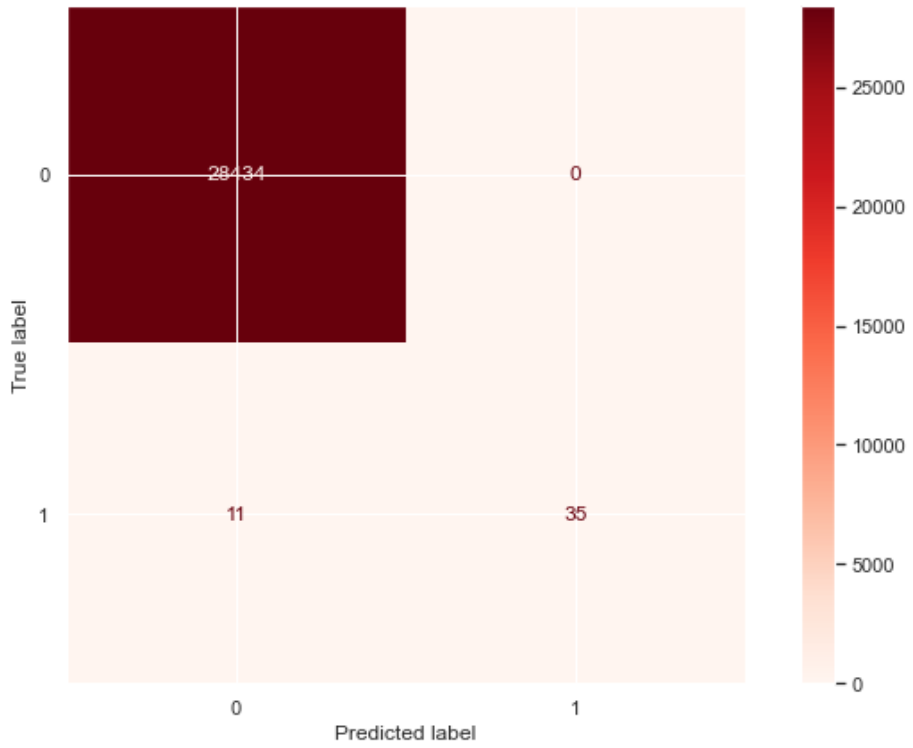

Fig 4.2.14 External data test Of Random forest Classifier Model

Here 11 valid transactions were detected fraudulent transactions and 0 fraudulent transactions were detected valid transactions with recall value of 0.76 on fraudulent transactions.

Now if we average the recall values of train set , test set and external set we get the 0.84 recall value which is the accuracy of the Random forest classifier model which is greater than Decision Tree Classifier model but not as good as logistic regression.

Chapter 5

Conclusion And Future Scope

5.1 Conclusion

In the above implementation we found out A function was created to carry out binary combinations (multiplication of attributes) by returning a dataframe with the combinations that meet the required value of correlation interest. In this case, those combinations that present a Pearson correlation greater than 0.5 were sealed. From the principal components, binary combinations were carried out to improve the linear correlation with respect to the Class attribute. In the logistic regression, a recall for fraud detection of 0.88 is obtained with the test set; a 0.89 recall with the External set, and a 0.90 recall on training with average of 0.89 .In the Decision tree classifier model , a recall for fraud detection of 0.75 is obtained with the test set; a 0.71 recall with the External set, and a 1.00 recall on training with average of 0.82. In the random Forest Classifier , a recall for fraud detection of 0.78 is obtained with the test set; a 0.76 recall with the External set, and a 1.00 recall on training with average of 0.84 Of the three classification models presented, the logistic regression model presents better results, secondly the random forest model and, lastly, the decision tree model.

5.2 Future Scope

- The very nature of this project allows for multiple algorithms to be integrated together.
- Addition of more algorithms can make it more efficient.
- This provides a great degree of modularity and versatility to the project.
- More room for improvement can be found in the dataset.

References

[1] Raj S.B.E., Portia A.A., Analysis on credit card fraud detection methods, Computer, Communication and Electrical Technology International Conference on (ICCCET) (2011), 152-156.

[2] Jain R., Gour B., Dubey S., A hybrid approach for credit card fraud detection using rough set and decision tree technique, International Journal of Computer Applications 139(10) (2016).

[3] Dermala N., Agrawal A.N., Credit card fraud detection using SVM and Reduction of false alarms, International Journal of Innovations in Engineering and Technology (IJIET) 7(2) (2016).

[4] Phua C., Lee V., Smith, Gayler K.R., A comprehensive survey of data mining-based fraud detection research. arXiv preprint arXiv:1009.6119 (2010). [5] Bahnsen A.C., Stojanovic A., Aouada D., Ottersten B., Cost sensitive credit card fraud detection using Bayes minimum risk. 12th International Conference on Machine Learning and Applications (ICMLA) (2013), 333-338.

Acknowledgement

We would like to express our Supervisor madam Prof.Payel Thakur for providing their invaluable guidance,comments and suggestions throughout the course of the project.we are also thankful for constantly motivating us to work harder.

We are extremely thankful and pay gratitude to our Head OF Department madam Dr.Sharvari Govilkar for her valuable guidance and support on completion of this project presently.

We would also like to extend our gratitude to the principal Sir Dr. Sandeep M. Joshi for providing all the facilities that were required.

Nikhil Mathapati

Yuvraj Patel

Deepanshu Pandita

Ashwin Nair