

CREDIT CARD FRAUD DETECTION

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

in

Computer Science and Engineering

By

Ritik Srivastava (181247)

Akshat Agarwal (181253)

Under the supervision of

Amit Jhakar



Department of Computer Science & Engineering and Information
Technology

Jaypee University of Information Technology Waknaghat,

Solan-173234, Himachal Pradesh

CERTIFICATE

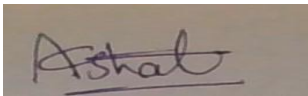
Candidate's Declaration

I hereby declare that the work presented in this report entitled “ **CREDIT CARD FRAUD DETECTION**” in partial fulfillment of the requirements for the award of the degree of **Bachelor of Technology in Computer Science and Engineering/Information Technology** submitted in the department of Computer Science & Engineering and Information Technology, Jaypee University of Information Technology Waknaghat is an authentic record of my own work carried out over a period from August 2021 to May 2022 under the supervision of **Amit Jhakkari** (Assistant Professor in Computer Science).

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




Ritik Srivastava, 181247.



Akshat Agarwal, 181253.

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



(Supervisor Signature)

Amit Kumar Jhakkari

Assistant Professor

Computer Science

Dated: 15/05/2022

ACKNOWLEDGEMENT

Firstly, I express my heartiest thanks and gratefulness to almighty God for His divine blessing makes us possible to complete the project work successfully.

I really grateful and wish my profound my indebtedness to Supervisor **Amit Kumar, Associate Professor** Department of CSE Jaypee University of Information Technology, Wazirpur. Deep Knowledge & keen interest of my supervisor in the field of Web Development to carry out this project. Her endless patience, scholarly guidance, continual encouragement, constant and energetic supervision, constructive criticism, valuable advice, reading many inferior drafts and correcting them at all stage have made it possible to complete this project.

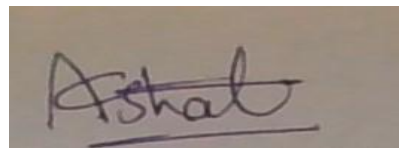
I would like to express my heartiest gratitude to **Amit Kumar** Department of CSE, for his kind help to finish my project.

I would also generously welcome each one of those individuals who have helped me straight forwardly or in a roundabout way in making this project a win. In this unique situation, I might want to thank the various staff individuals, both educating and non-instructing, which have developed their convenient help and facilitated my undertaking

Finally, I must acknowledge with due respect the constant support and patients of my parents.



Ritik Srivastava



Akshat Agarwal

TABLE OF CONTENTS

CERTIFICATE.....	I
ACKNOWLEDGEMENT.....	II
TABLE OF CONTENTS.....	III
LIST OF ABBREVIATIONS.....	IV
ABSTRACT.....	V
CHAPTER-01:	
INTRODUCTION.....	1
1.1 Introduction:	1
1.2 Problem Statement.....	2
1.3 Objectives.....	2
1.4 Methodology.....	3
1.5 Organization.....	3
CHAPTER-02:	
LITERATURE SURVEY	11
2.1 Literature Survey	11
CHAPTER-03:	
SYSTEM DEVELOPMENT.....	13
3.1 Analysis/Design/Development/Algorithm	13
3.2 Model Development	14
CHAPTER-04:	
PERFORMANCE ANALYSIS	21
4.1 Analysis of system developed	22
4.2 Output at various stages.....	23
4.3 Comparison of above results by at least two methods and justification for the differences or error in with theory or earlier published results	24
CHAPTER-05:	
CONCLUSIONS.....	26
5.1 Conclusions	26
5.2 Future Scope	29
5.3 Applications.....	29
REFERENCES	30
APPENDICES.....	31

LIST OF ABBREVIATIONS

- 1) ML: Machine Learning
- 2) CNN: Convolutional Neural Network
- 3) DL: Deep Learning
- 4) CSV: Comma Separated Values
- 5) API: Application Program Interface
- 6) FC: Fully connected
- 7) CLI: Command Line Interface
- 8) DNN: Deep Neural Network
- 9) XML: Extensible Markup Language
- 10) PCA: Principal Component Analysis
- 11) MAE: Mean Absolute Error

ABSTRACT

“Fraud detection is a series of existing activities taken to prevent money or property from being obtained by falsehood.” Fraud can also be executed in a variety of ways and in a wide range of industries. Credit card fraud simple is intended to be friendly. E-commerce and a lot of online sites have improved online payment methods, increasing the risk of online fraud. Increased fraud rates, researchers began using alternative machine learning ways to detect and analyse fraud in online trading. Credit card fraud usually occurs when the card was in existence stolen for any unauthorized purposes or even there The credit provider uses credit card information to operate it. Loss of money is lost due to credit card fraud every year. There is a lack of research studies in real-world analysis credit card data due to privacy issues. In this paper, machine learning algorithms used to obtain credit card fraud. To evaluate the performance of the model, a public Using the available credit card data set. System Forecast the rate and accuracy of detection of fraud does not exceed 100 accurate, therefore there is a chance to find fraud. Then, a real-world credit card data set from the financial institution is under scrutiny. In addition, sound added in data samples to further test the robustness of algorithms. Test results clearly show that most voting systems reach good levels of accuracy in to find credit card fraud charges.

CHAPTER-01: INTRODUCTION

1.1 Introduction

As we look at the digital world - cybersecurity becomes an integral part of our lives. If we talk about security in digital life then the biggest challenge is finding a unique job. If we make any purchase while buying any product online - a good number of people opt for credit cards. Credit limit on credit cards sometimes helps us make purchases even if we do not have the money at the time. but, on the other hand, these features are being misused by cybercriminals. To address this issue, we need a system that can reverse the transaction if it is caught. Here comes the need for a system that can track the pattern of all actions and if any pattern is abnormal then the action should be withdrawn.



Credit card fraud is unauthorized and does not require the use of an account by someone other than the owner of that account. The necessary precautionary measures can be taken to prevent this abuse and the behaviour of those fraudulent activities can be investigated to reduce and prevent similar situations in the future. another card for personal reasons while the owner and issuing authorities of the card are unaware of the fact that the card is in use.

Fraud detection involves monitoring the activities of the majority of users in order to measure, detect or avoid undesirable behaviour, including fraud, interference, and error. This is a very relevant problem that needs the attention of communities such as machine learning and data science where the solution to this problem can be automated. This problem is particularly acute in view of the learning process, as it is characterized by a variety of factors such as class inequality. The number of jobs allowed far exceeds that of a counterfeit. Also, practical patterns often change their mathematical features over time.

Today, we have many machine learning algorithms that can help us differentiate between extraordinary activities. The only requirement is past data and a suitable algorithm that can fit our data in a better way.

In this article, I will help you with a complete model training program - in the end, you will find the best model that can distinguish transactions into common and unusual types.

1.2 Credit card fraud types

As described by the fraud detection experts credit card fraud may be of 5 types:

Lost/stolen cards : Especially against the elderly cardholders, the fraudster gets a PIN code by reading on the shoulders and stealing card later. In this case the fraudster is a thief, a credit card does not go through a network of re-sales in organized crime.

Non received cards : Credit card stolen during the production or delivery of mail. To avoid this kind of fraud, banks can ask the customer to return his / her bank card card, or call them to unlock the card.

ID Theft : Card obtained using false or stolen ID documents

Counterfeint cards : Card copied during real-time card usage or during database theft and reproduction after which it is a fake plastic made by organized crime groups around the world. fraudster retrieve and reproduces magnetic line card data. This type of fraud has been

rampant in the past but has been partially resolved by EMV technology: only magnetic line terminals are no longer used The EU however remains in Asia and America. It is noteworthy that it is contactless payment is not as pleasurable to the impostor as lower payments only are allowed.

Card not present frauds : Most of credit card fraud occurs in e-commerce trading. Details (card number, expiration date and CVC) are usually returned during database and robberies organized by international criminal gangs and subsequently sold on the black web. British flights, Mariot hotels and tickets Master was exemplary victims of data breaches in 2018. Price for details depend on the location of the fraud (first digits) card numbers identify the bank and therefore blocking policy). Most retailers (90%) use 3D SECURE protective technology card holder with dual identification however some major retailer website like Ebay or Amazon does not protect its users with 3D SECURE. Another problem that prevents you from fighting the wrong card is fraud companies do not report any attacks that have caused data breaches because of the bad advertising that can cause it.

1.3 Problem Statement

A historical credit card model containing information of those who look to be fraudulent is used to solve the problem of finding a credit card fraud. This model is then utilised to determine whether or not a new employment is legitimate. Our objective is to detect 100% of fraudulent actions while reducing unjust categorisation. What exactly does a credit card fee entail?

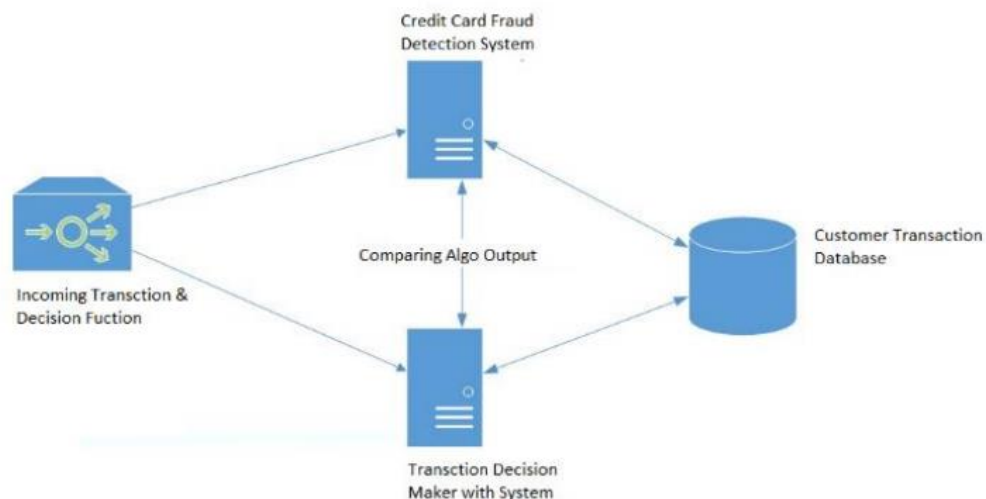
A historical credit card model containing information of those who look to be fraudulent is used to solve the problem of finding a credit card fraud. This model is then utilised to determine whether or not a new employment is legitimate.

1.4 Objective

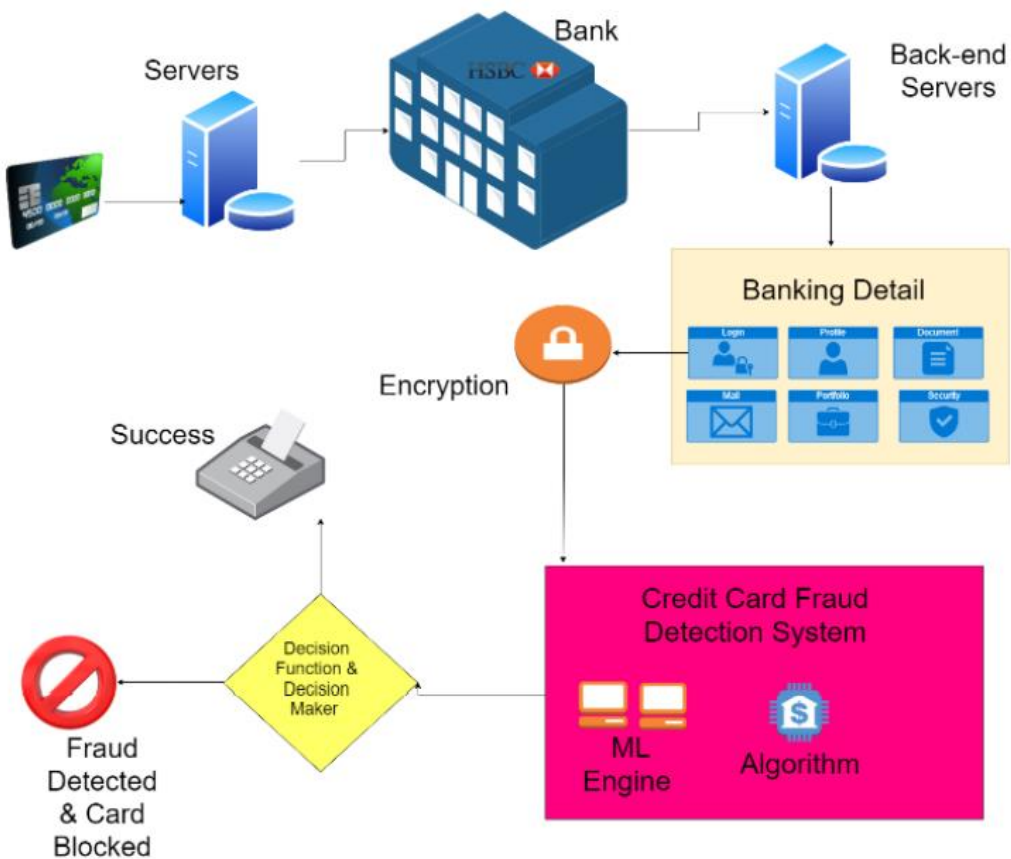
The major goal of this project is to train and model the Credit Card Fraud Detection Model, which combines previous credit card modelling with data from people who have been victims of fraud. This model is then utilised to determine whether or not a new employment is legitimate. Our objective is to catch all fraudulent acts while reducing unjust categorisation.

1.5 Methodology

The method proposed by this paper, uses the latest technology learning algorithms to find confusing tasks, called outsiders. A drawing of rough buildings can also be represented next image:



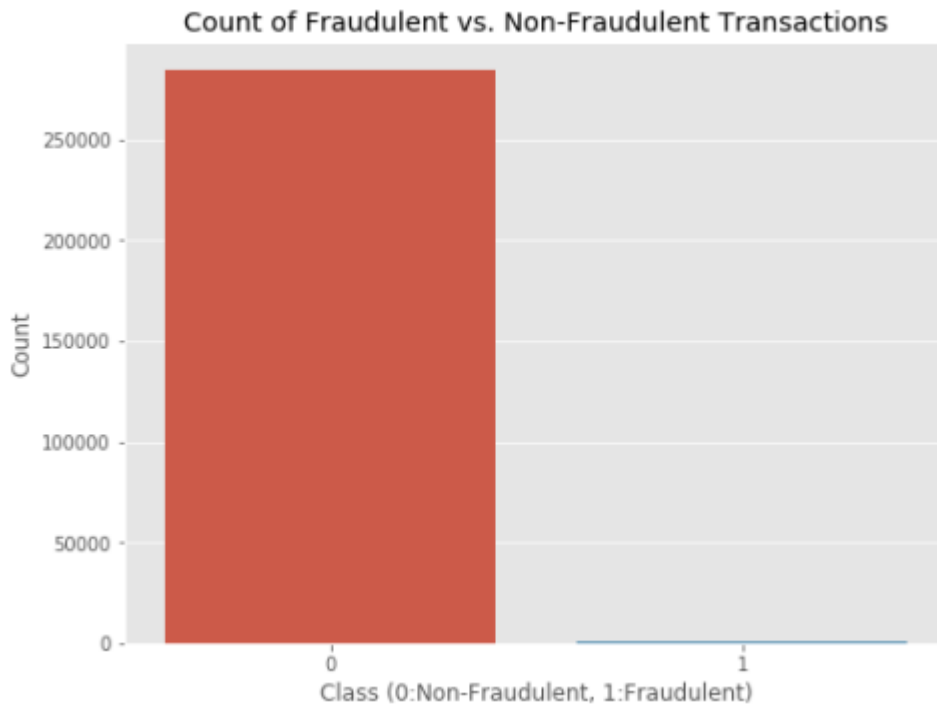
When viewed in detail on a large scale and in real life elements, a complete sketch of buildings can be represented as follows:



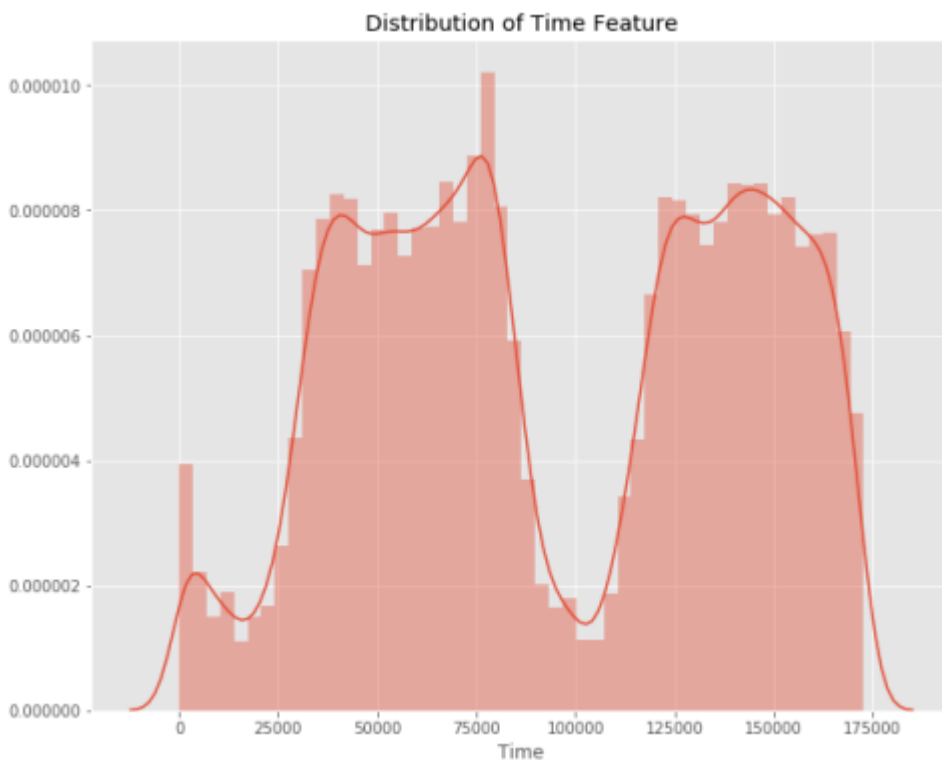
First, we looked for our database on Kaggle, a data analysis and database-sharing platform. To safeguard sensitive data, 31 of the 28 columns in this database are labelled v1-v28. Time, Value, and Class are all represented by certain columns. The time between the first action and the next one is indicated by time. The quantity of money made is the value. A genuine function is represented by section 0, whereas a fraud is represented by section 1.

We use several graphs to visually understand the dataset and check for discrepancies.

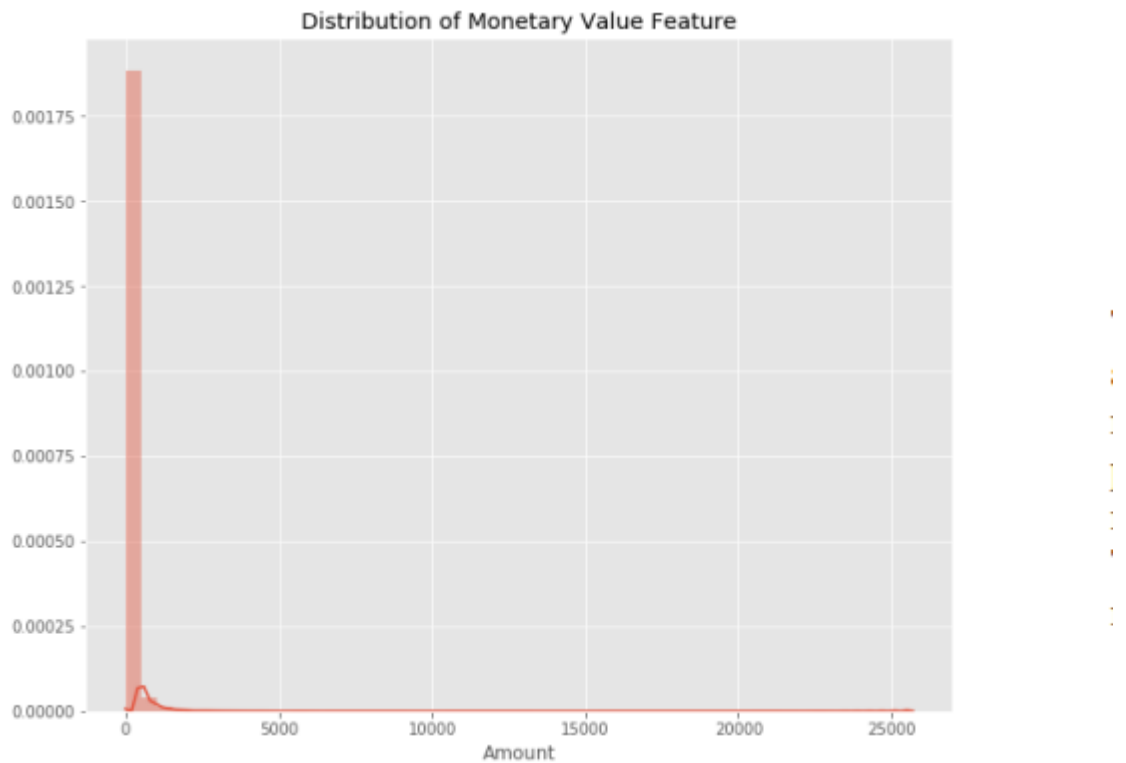
:



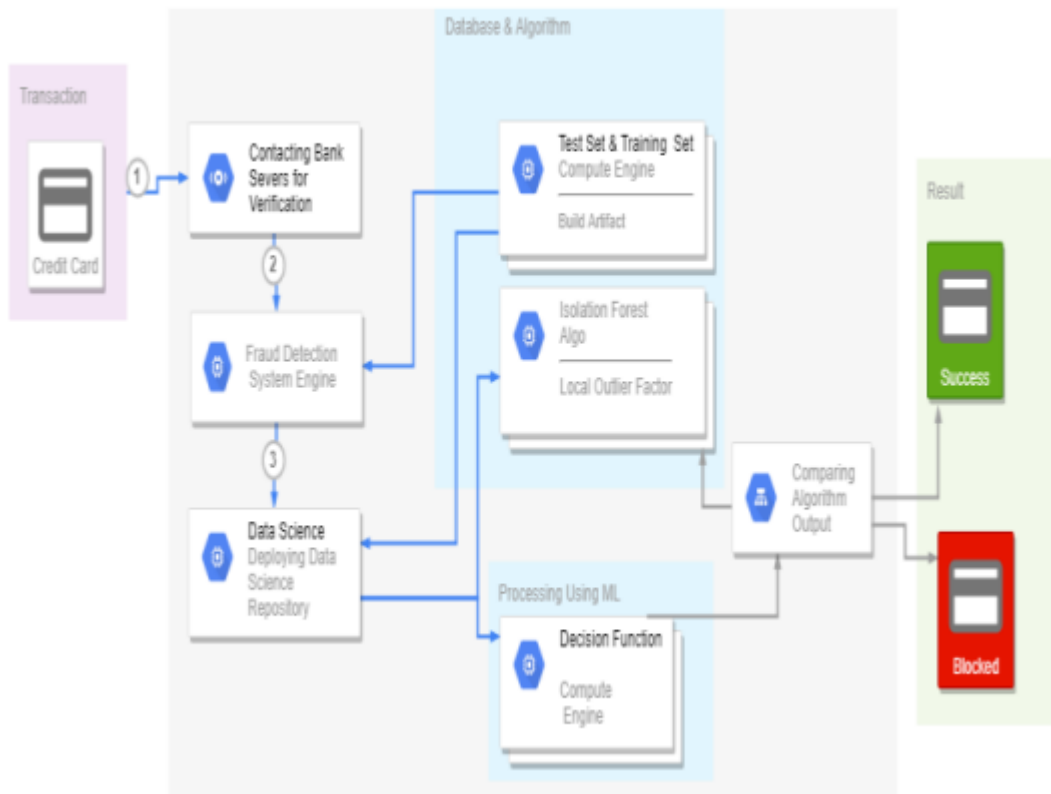
This graph shows that the number of fraudulent transactions is way more than the number of non fraudulent transactions.



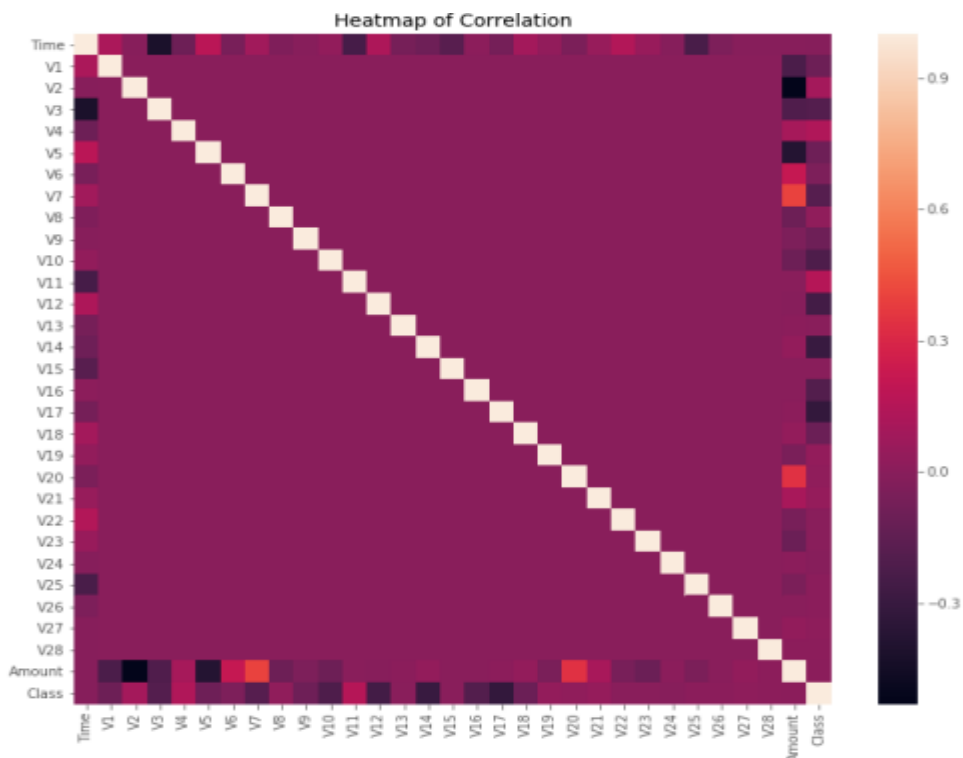
This graph shows the time when the transactions were made. It is clearly shown that the maximum number of transactions were made during day time.



This graph represents the value generated. A Most jobs are small and limited of which are closer to the maximum value of the product. After checking this database, we created a histogram for everyone column. This is done to get a symbolic representation of a database that can be used to ensure that nothing is missing any values in the database. This is done to make sure we do not they require any missing value and machine learning algorithms can process the database smoothly.



The heatmap of the given analysis is:



1.5 Organization

The report is divided into five modules, and a full description of each module for this project is provided for clarification and comprehension. Part 1: This module serves as the official introduction to the project and contains all the required information. In this section, we introduce the reader to many of the terms of the project while explaining the problem or motivation for doing the project from the beginning. In line with this, we begin again with the purpose of our project and the methods that will be used to achieve it. Part 2: This module contains a collection of current research studies conducted on our project. In this section, we place great emphasis on the method used by the articles. In addition, we look at the results of their various programs. Part 3: We will go through many stages of our development in this module, and we will learn about the design and operation of our algorithm. In this section, we will re-model and try to represent it with a variety of ideas, including analytical, mathematical, experimental, mathematical, and mathematical ideas, among others. Part 4: In this module, we will evaluate the effectiveness of our project and provide development recommendations. Part 5: This will be our last module, where we will discuss the results of our research and evaluate our findings and conclusions. In addition, we will explore the future scope of the project and any potential enhancements in the near future. In addition, we will describe some of the applications where the system may be profitable.

CHAPTER-02

LITERATURE SURVEY

2.1 Literature Survey

Fraud is defined as an illegal deception or a crime committed for financial or personal advantage. That is an intentional conduct that violates the law, law, or policy in order to get illegal financial advantage. Many books on unexplained discoveries or fraud have been written and are available for public use on this domain. Data mining applications, automatic fraud detection, and adversary recognition are among the strategy leases on this domain, according to Clifton Phua and his partners' extensive study. Suman, Scholar Research, GJUS & T at Hisar HCE, introduces credit card fraud detection methodologies such as Supervised and Unattended Learning in another work. Although these approaches and algorithms have had surprising success in some locations, they have failed to provide a long-term and consistent answer for fraud detection. Wen-Fang YU and Na Wang created the same research location where they run the In the simulated test of Credit card set data for particular trades the bank, the Outlier mines, Outlier discover mines, and Distance sum algorithms for accurate accuracy identify fraudulent activities. Outlier mining is a type of data analysis that is commonly utilised in financial forums and on the internet. It's all about getting stuff from the main system, such as false positives. They took the attributes of consumer behaviour and estimated the distance between the rental value of that attribute and its pre-determined value depending on their value. It is possible to see illegal occurrences on real card data sets using unusual techniques such as mixed data mining / complex network division algorithm, which is based on a network network algorithm that allows to create single model deviations from the reference group appears to be operating normally in the middle limited online transactions.

Efforts to enhance have also been made. A new feature built from the ground up. In the case of fraudulent behaviour, efforts have been undertaken to increase the interoperability of warning systems.

An authorised system will be alerted in the event of a fraudulent transaction, and a response to further rejection will be delivered transaction.

One of the most extensively utilised approaches, Artificial Genetic Algorithm, sheds new light on this sector, combating fraud from a different direction.

Detecting fraudulent transactions and limiting the amount of erroneous notifications sounded reasonable. However, there was a difficulty with variable separation and the expense of misalignment.

CHAPTER-03

SYSTEM DEVELOPMENT

3.1 Analysis/Design/Development/Algorithm

In this project, the challenge is to identify credit card fraud so that credit card companies can no longer be charged for purchases they have not made.

The major challenges involved in obtaining credit card fraud are:

Big Data is processed daily and model construction should be fast enough to respond to this scam early.

Non-Validated Data which means that most jobs (99.8%) are not fake which makes it very difficult to find fake ones.

Data availability as data is very private.

Incorrectly sorted data can be another major problem, as not all fraudulent activity has been detected and reported.

The familiar methods used by fraudsters against the model.

A. Ensemble Technique

Bagging and Boosting are the two types of methods used in ensemble tactics. "BOOTSTRAP AGGREGATION" is a shorthand for "FUNDRASING." Random Forest is one of Bagging's gimmicks. In the Bagging Technique, Using the "line sampling and transformation" approach, distinct sets of samples from the Data set are supplied to different types of models to be trained on that specific sample data. All of the different models are trained in those sample data sets at the same time. After the models have been trained, they may be put to the test with a collection of test data. The results of all the models are sent to the person who is voting in order to acquire the majority of the votes cast for the models, and those multiple votes are taken into account while analysing the survey data. This is how the bagging method functions.

B. Random Forest

We employ a decision tree in the models to be trained with selected data samples using the "line sampling and flexible sampling approach" and test models with test data and integrate all the findings. from models since the random forest is a Bagging method. Low Bias and High Variation are the two sections of the decision tree. When a decision tree is built to its full depth, it will be well-trained given a set of training data that has a very small training error. When fresh test data is utilised, these models have a higher tendency to produce a bigger number of mistakes. As a result, when a decision is taken in its entirety, it leads to overuse. The challenge of filling is alleviated by taking samples. Despite the fact that the samples' findings overlap, the Ensemble of all outcomes produces a precise result. Many decision trees are used in random forests, and each decision has a high degree of diversity, but when all decision trees are integrated into a majority vote, this high degree of diversity translates into a low variance. Each decision tree is processed in a different sample dataset and is specifically trained on this data. This takes into account the maximum effect between all decision trees and achieves the minimum difference. Therefore, changing a small part of the database does not affect its output and accuracy decisions. This happens to be the most important quality of the jungle. Most are output for classification and the sum of all results is output for regression

3.2 Model Development

1) Dataset Used

This dataset contains credit card transactions by European cardholders in September 2013. This dataset shows 492 of the 284,807 transactions executed in two days. The dataset is very imbalanced, with positive classes (illegal) accounting for 0.172% of all transactions. This includes only numeric input variables that are the result of the PCA conversion. Unfortunately, for confidentiality reasons, we are unable to provide the original functionality of the data and detailed background information. The properties V1, V2, ... V28 are the principal components obtained from the PCA, and the only properties that the PCA does not convert are "time" and "quantity". The time function contains the number of seconds elapsed from each transaction to the first transaction in the record. The "amount" characteristic is the transaction amount. This property can be used, for example, for dependent, cost-sensitive

learning. The characteristic "class" is a response variable that expects a value of 1 if it is invalid and a value of 0 otherwise. Considering the imbalance rate of the class, it is recommended to measure the accuracy in the area under the Precision Recall curve (AUPRC). The accuracy of the confusion matrix is meaningless for imbalanced classifications.

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

Data Cleaning and Preprocessing

To prepare the modeling data, you should avoid independent variables, check for anomalous entries, and remove the columns with many missing values. Similarly, if a row has many missing values, delete those specific rows. You can use the mouse algorithm to assign some missing values in R. Processing missing values is time consuming and an important process in data cleansing. Transform your data using normalization, discretization, logarithmic transformation, and feature selection. -> Build a Decision Tree

Decision tree is one of the most important models for segregation. It is famous for its openness. The branches of the tree of decision for each variation of the upper class division.

Gini Impurity:
$$I_G(p) = \sum_{i=1}^c p_i(1 - p_i) = 1 - \sum_{i=1}^c p_i^2$$

Information Gain or Entropy:
$$H(p) = - \sum_{i=1}^c p_i \log_2 p_i$$

When P_i has a chance to capture the attention of the class, we must first shuffle the data and divide it into two parts: the training set and the test set, with 80 percent of the training data being provided. To shuffle, we utilise the sample function. Rename the test data "Class" and resell it as a data framework. Because the classes are not distinct, we utilise the R-package

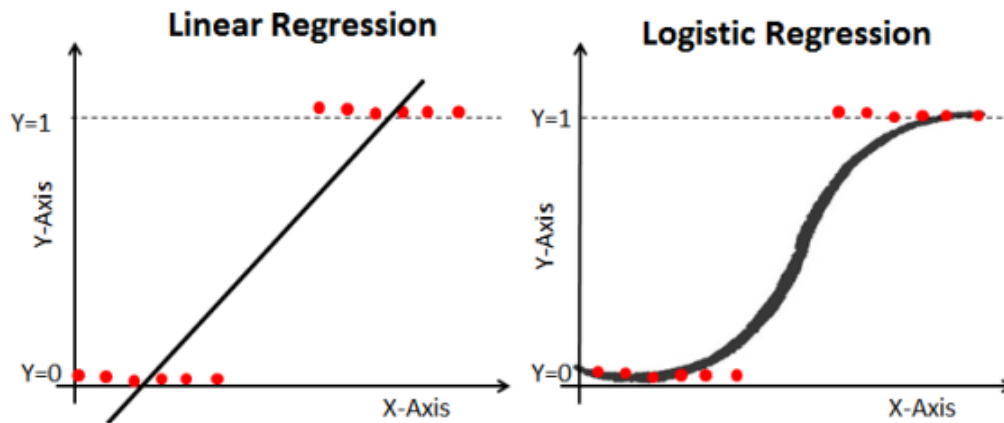
function to generate effective cutting trees when we use the class method to distinguish. We can acquire the true extent and intricacy of everything done by employing a two-column matrix and a confusion matrix. In the midst of the chaos, we discovered that 9 legitimate employment were disguised as scams, while 22 frauds were incorrectly categorised. We receive an average accuracy of 89.31%, which is a sensitivity and clarity measure. Using the function `rpart.plot`, we can visualise the decision tree with `drawings.plot`, this transparency provided by the decision tree.

Building the Classifier

In order to construct the classifier, logistic regression is used. In comparison to linear regression, logistic regression is more advanced. Because linear regression cannot categorise data which are widely distributed in a given space, this is the case.

For the goal of demonstrating this shortcoming, the graphic below displays a visual comparison of the linear regression and logistic regression methods. The advantages of logistic regression are as follows: 1) Logistic regression is more straightforward to implement than linear regression, and it is more faster to train. 2) No assumptions are made about class distributions in the feature space. 3) It's simple to expand to various classes (multinomial regression). 4) It's great for classifying unidentified records. The linear regression equation can be used to calculate the logistic regression equation. The following are the mathematical steps to obtain logistic regression equations: The equation for a straight line is as follows:

$y = a_0 + a_1 \times x_1 + a_2 \times x_2 + \dots + a_k \times x_k$ (1) In logistic regression, y can be between 0 and 1 only, so we divide the above equation by $(1 - y)$: $y / (1 - y) | 0 \text{ for } y = 0 \text{ and } \infty \text{ for } y = 1$ (2) As a result, the logistic regression equation is defined as: $\log [y / (1 - y)] = a_0 + a_1 \times x_1 + a_2 \times x_2 + \dots + a_k \times x_k$ (3)



Build Random Forest

Once you have 89 accuracy in the decision tree, build a decision tree to build a random forest model to make the result improved and the accurate. Using the complexity of the survey's resolution tree, the time complexity is equal to 'N³'. Where, 'N' is the reference number. Here n is 285000 and time=1 minute. Example: One tree takes 2 seconds, but 10 trees require 15 seconds, depending on the percentage of data. The model uses the RandomForest package and considers the corresponding parameters such as tree diameter 10-100, sample diameter 20% -80%, maxnodes 30-70, and so on. Once you have the code section displayed as a function of all variables, extract the data from the training set and set parameters such as ntree, samplesize, maxnodes and so on. Make predictions using a model of test data. By reviewing a matrix of two real operational confusions that were misunderstood as fraud and 20 misconduct that was misunderstood as true. With an accuracy of 90, Random Forest provides better final results compared to decision trees. Here, the emphasis is on high sensitivity or high specificity rather than reducing working hours. Comparing the confusion matrices in both decision trees and random forests reveals increased sensitivity and clear values. If you change the code with a nice new parameter = "1" for both the area under the curve and the confusion matrix, the area under the curve and the confusion matrix value if matched will remain in the same forest. Although you can get the decision tree sensitivity and clarity. In both cases, the tree trunks and unplanned forests have changed. You can also test performance by calculating additional steps such as accuracy and memory, Fscore from ML metrics, and Matthews correlation coefficient from mltools. You can see that using a

random forest with SMOTE (Synthetic Minority Oversampling Technique) and removing the Tomek link to get accurate points and storage values will significantly increase the end result. Random forests have proven that the sample reconstruction method works well when compared to diminishing conditions.

Pred	Ref	0	1
0	56856	20	
1		2	83

Accu	:	0.9996
Sens	:	1.0000
Spec	:	0.8058
B_Ac	:	0.9029

AUC : 0.9029

-> Imbalance in the data

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

Only 0.17% of all transactions are fraudulent. The data is terribly imbalanced. First, let's apply the model without balancing it. If you don't get enough accuracy, you can find a way to balance the dataset. However, do not implement the model first and balance the data only when needed. The amount details for Fraudulent Transaction are:

Amount details of the fraudulent transaction

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

The amount details for Normal Transaction are :

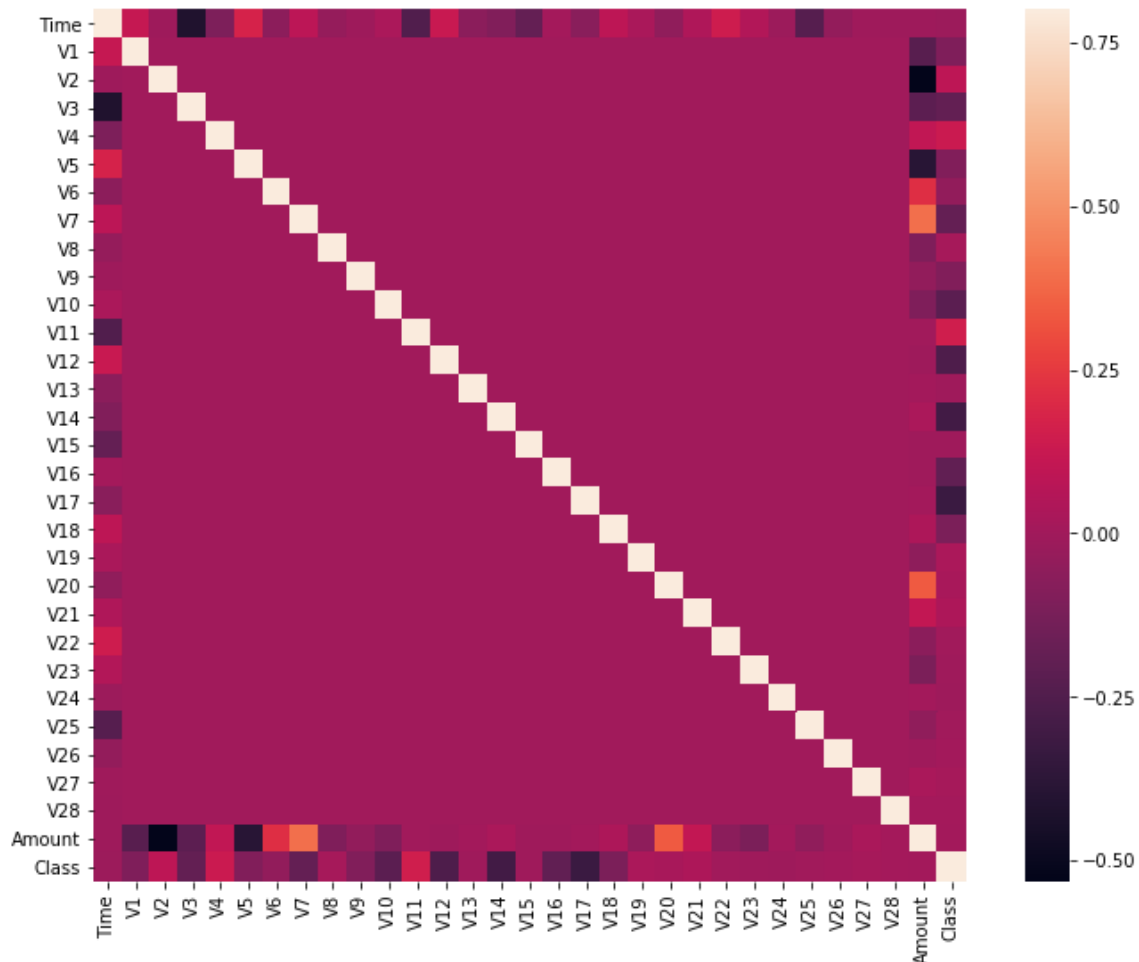
Amount details of valid transaction

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

As we can clearly see from this, the average amount of money laundering is high. This makes this problem very important to deal with.

Plotting the Correlation Matrix

The correlation matrix gives us an idea of how the features relate to each other and can help us predict which features are most appropriate for prediction.



In HeatMap we can clearly see that many features are not related to other features but there are some features that have a positive or negative relationship. For example, V2 and V5 are strongly associated with a feature called Value. We also see some connection with V20 and Price. This gives us a deeper understanding of available data.

Bifurcation of Training and Test Data

We will be dividing the database into two main groups. One for model training and the other for Testing of our trained model.

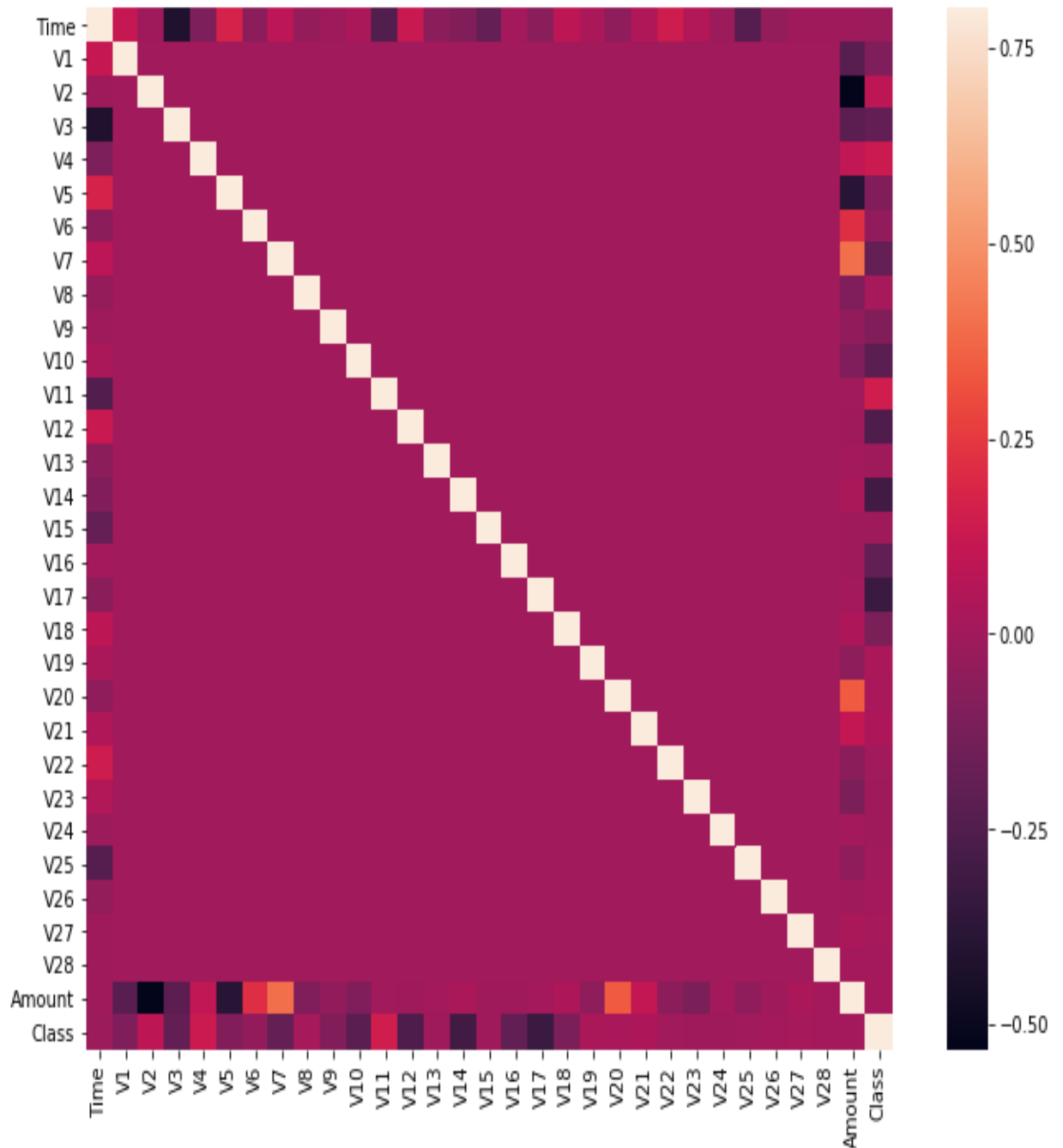
```
# Using Skicit-learn to split data into training and testing sets
from sklearn.model_selection import train_test_split
# Split the data into training and testing sets
xTrain, xTest, yTrain, yTest = train_test_split(
    xData, yData, test_size = 0.2, random_state = 42)
```

CHAPTER-04

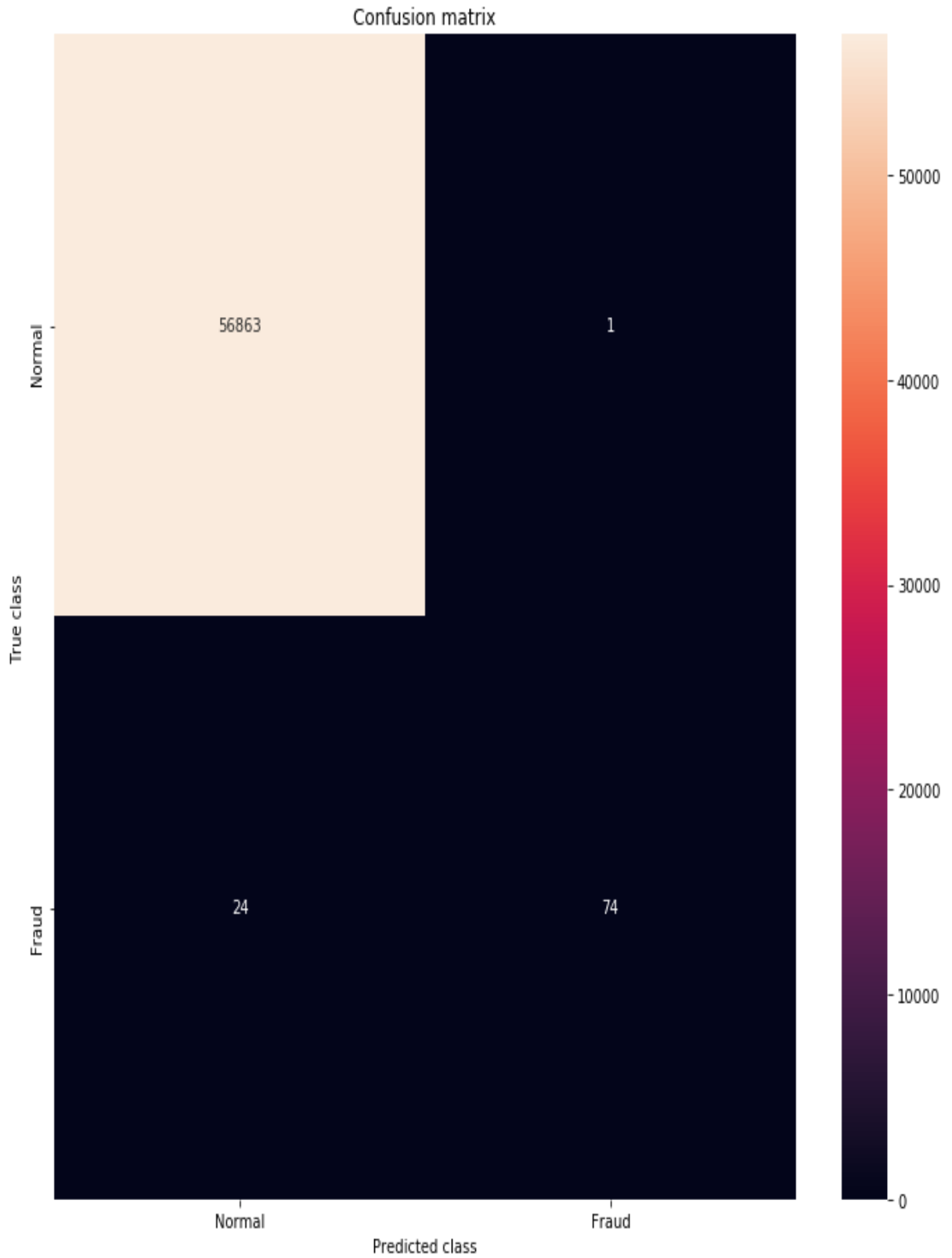
PERFORMANCE ANALYTICS

4.1 Outputs At Various Stages

Plotting the Correlation Matrix



Visualizing the Confusion Matrix



Comparing all the approaches for fraud detection

Markov Hidden Model (HMM)

The Hidden Markov Model is a stochastic model with a collection of border regions that measure the boundaries of transformative opportunities. The HMM Model is based on Markov's design, which states that future events are independent of prior regions and only dependent on current ones. Because of this structure, it can be used for predictive modelling and probable predictions. This model, as stated in, is used to detect fraud. They do this by modelling human behaviour based on cardholders' encryption patterns and the situation in the model type of purchase. Three visual signals are formed by only three l-low, m-medium, and h-high values. Allowing $l = (0, \$200)$, $m = [\$200, \$400]$, and $h = [\$400, \text{credit card limit}]$ as examples. The figures used are True Positive, TP-False ood, FP, and accuracy are the measures employed by this methodology. For any large or small inputs, the recommended accuracy is always near to 80%. However, if no information about the profile is provided, the performance of TP-FP measures will suffer. If there is a tiny discrepancy between real and cruel transactions, then FDS is facing a degradation in its functioning as a result of a decrease in TPs or a rise in FP.

Decision Tree

Supervised learning algorithm. Tree decision-making tree structure, which includes root node and some nodules are separated by binary or separate and often continue to be child nodules each tree uses its own algorithm to perform the separation process, until there is no longer a need to differentiate which will make a difference in our model, associating each element with an input value in relation to the method used as defined by Y. Sahin nd E. Duman. With the growth of the tree, there may be opportunities to complete the training data in a possible way incomprehensible to branches, certain errors or noise. Pruning is therefore used to improve the performance of tree sections by removing certain nodes. Easy to use, as well as compliance with the conditions provided by decision trees for the management of separate data types of attributes make them very popular.

Unplanned Forests

Instability in single trees and sensitivity to specific training data has led to the development of another random model forests. With each tree built without the other the computational efficiency of the random forest becomes comparable the best . Basically a group of receding and / or dividing trees that find differences between them trees are therefore easy to use due to the use of only two random resources or the boundaries that make up trees using trained data separates the bootstrap associated with the samples by considering only the random data attribute for each tree construct as specified.

Logistic Regression

It is an appropriate method that can be used in predictable analysis where the dependent variations are dyadic or binary . Since the division of labor is fraudulent is a double standard, this procedure can be applied. This he mathematical division based on probability detects fraud using the entry curve. From the importance of this the entry curve varies from 0 to 1, can be used to translate class membership opportunities. The database supplied as input to the model is broken down for training and model testing. Model post training, of course tested the minimum limit cut to predict. Then the most important variables are selected the model is configured correctly. Predictability accuracy is out of 70%. From a recession, based on other possibilities can split a plane using a single line and split data points into two straight circuits. Therefore, external factors are not well treated . It uses natural logarithmic function to calculate probability and to show that results fall under a certain category.

Support Vector Equipment

Vector-supporting machines or SVMs are a series of components as stated in high-performance because high size, non-linear function in the line becomes linear so this makes SVMs very useful in detection fraud. Because of its two most important features which is the kernel function to represent the dot partition function the output of the input data point, as well as the fact that it is trying to detect the hyperplane to increase the split classes while minimizing overcrowding of training data, provide very high productivity.

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

As you can see with our Random Forest Model we get a better result even remembering the hardest part.

CHAPTER-05 : CONCLUSIONS

5.1 Conclusions

The code outputs the number of bogus symbols received and compares them to actual values. This is used to compute the school's precision and accuracy algorithms. The percentage of data we used for quick testing was 10% of all databases. At the end, the entire database is utilised, and both results are printed. For each section, these findings and reports are provided. The algorithm is applied to the output in the following order: section 0 indicates that the transaction has been determined to work with 1 method, and section 1 indicates that the transaction has been judged to be fraudulent. This result was found to be good when compared to class values for false testing.

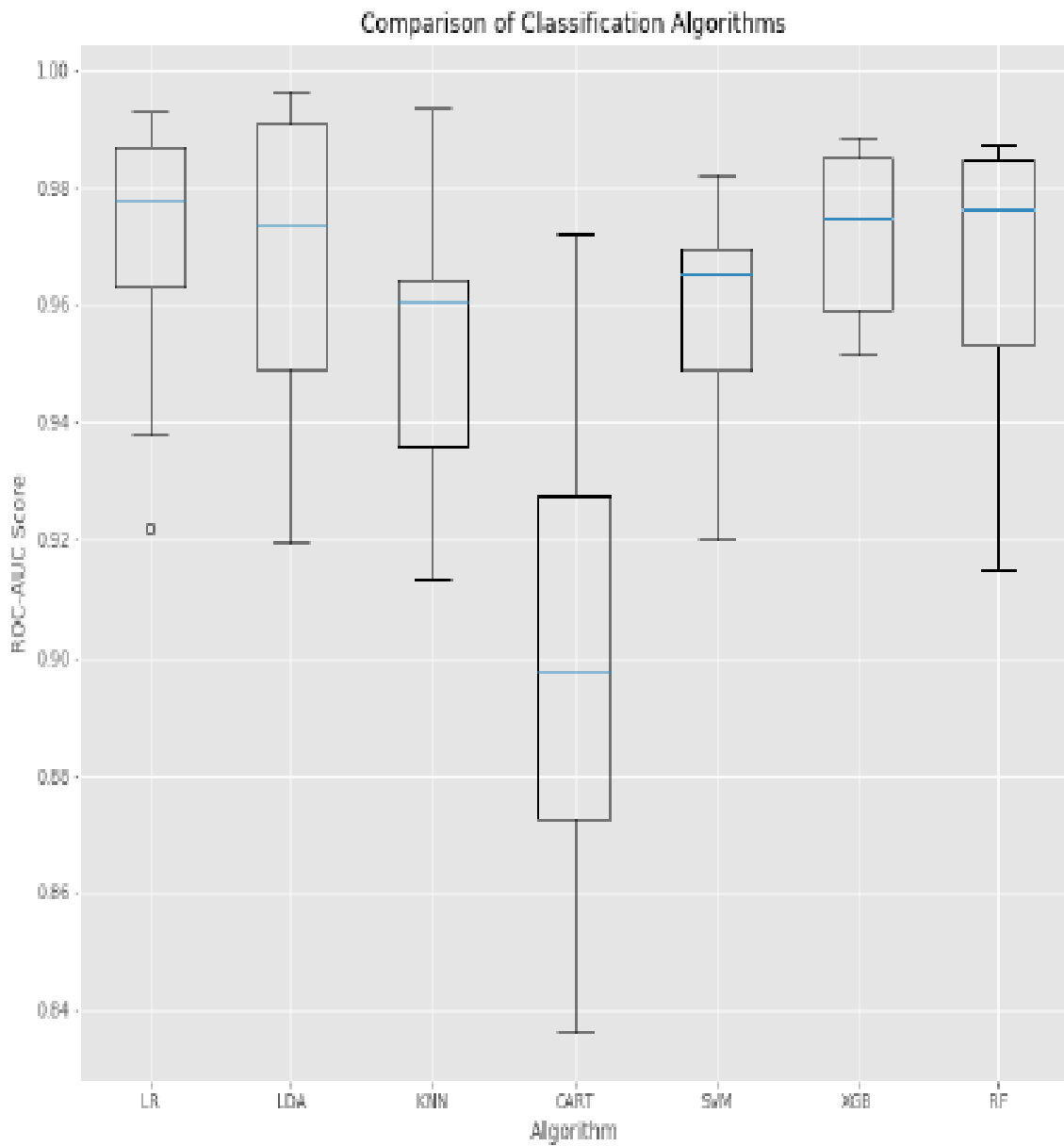
Random Forest is particularly good at maintaining a large number of less complicated datasets in a short period of time. When compared to algorithms like decision trees and vector support machines, the final findings reveal a considerable rise in credit card fraud. Due to their poor effectiveness in managing uneven data sets, they are used for retrofitting and other purposes. Using a pile of logged trees, the random forest addresses the problem. The informal forest is viewed as a means of resolving uneven partition. Due to transparency, accuracy, and the use of various re-sampling techniques such as SMOTE, tokek link removal, random sampling, Random sampling,, Random jungle plays an important role in the work cycle of large companies Institutions that use data science, Artificial Intelligence, and financial literacy save millions of dollars every day by preventing fraud in every manner possible.

Fit your model on k components before making predictions for k-1 folds (folds). The kth hold-out is folded. Then you go through the process again for each fold. Calculate the average of the resulting projections. To gain a better understanding, let's check which algorithm performs the best on our data. In a flash:ible method, look at some of the most common classification algorithms.

- Logistic Regression
- Linear Discriminant Analysis

- Classification trees
- Support Vector Classification
- Random Forest Classification

The results of the above can be visualized as follows:



As we can see, a few algorithms beat the rest by a large margin. As previously stated, the goal of this project was not only to achieve the maximum level of accuracy, but also to provide commercial value. As a result, selecting logistic regression over XGBoost may be an acceptable strategy for achieving a better level of comprehensiveness while only slightly lowering performance. Here's a representation of our Logistic regression model result that might be used to explain why a certain decision was made:

```
Accuracy Score

[ ] # accuracy on training data
    X_train_prediction = model.predict(X_train)
    training_data_accuracy = accuracy_score(X_train_prediction, Y_train)

▶ print('Accuracy on Training data : ', training_data_accuracy)

Accuracy on Training data : 0.9415501905972046

[ ] # accuracy on test data
    X_test_prediction = model.predict(X_test)
    test_data_accuracy = accuracy_score(X_test_prediction, Y_test)

[ ] print('Accuracy score on Test Data : ', test_data_accuracy)

Accuracy score on Test Data : 0.9390862944162437
```

5.2 Future Scope

We weren't able to achieve our target of 100 percent fraud detection accuracy, but with enough time and data, we were able to design a system that came pretty close. There is, like with every other endeavour of this nature, space for improvement. Because of the nature of this project, various algorithms can be used as modules, with the following results. Combine to increase the final result's accuracy. This model may be enhanced further. It contains an algorithm. These algorithms' effects, however, must be in the same format as the other algorithms. This is precisely how the problem is handled; modules are simply introduced in the same manner that they are in code. This allows for a lot of adaptability and project flexibility. The database contains positions for further development. As previously said, the larger the database, the more accurate the results.

As a result, the additional data improves the model's accuracy in identifying fraud. Reduce the amount of false positives by reducing the number of false positives. This, however, need the bank's formal backing.

5.3 Applications

According to a Bloomberg article, losses on credit card debit, debit, and prepaid loans issued globally totaled \$ 21.84 billion in 2015. Bloomberg estimates that by 2020, this will have increased by 45 percent on average.

According to our findings, both banks and retailers are considering AI solutions to protect their consumers.

Ecommerce website such as amazon, flipkart etc, also protect their customers using ai to detect fraudulent activities during purchase of their product.

REFERENCES

- [1] “Credit Card Fraud Detection Based on Transaction Behaviour -by John Richard D. Kho, Larry A. Veal” published by Proc. of the 2017 IEEE Region 10 Conference (TENCON), Malaysia, November 5-8, 2017
- [2] CLIFTON PHUA¹, VINCENT LEE¹, KATE SMITH¹ & ROSS GAYLER² “ A Comprehensive Survey of Data Mining-based Fraud Detection Research” published by School of Business Systems, Faculty of Information Technology, Monash University, Wellington Road, Clayton, Victoria 3800, Australia
- [3] “Survey Paper on Credit Card Fraud Detection by Suman” , Research Scholar, GJUS&T Hisar HCE, Sonapat published by International Journal of Advanced Research in Computer Engineering & Technology (IJARCET) Volume 3 Issue 3, March 2014
- [4] “Research on Credit Card Fraud Detection Model Based on Distance Sum – by Wen-Fang YU and Na Wang” published by 2009 International Joint Conference on Artificial Intelligence
- [5] “Credit Card Fraud Detection through Parenclitic Network AnalysisBy Massimiliano Zanin, Miguel Romance, Regino Criado, and SantiagoMoral” published by Hindawi Complexity Volume 2018, Article ID 5764370, 9 pages
- [6] “Credit Card Fraud Detection: A Realistic Modeling and a Novel Learning Strategy” published by IEEE TRANSACTIONS ON NEURAL NETWORKS AND LEARNING SYSTEMS, VOL. 29, NO. 8, AUGUST 2018
- [7] “Credit Card Fraud Detection-by Ishu Trivedi, Monika, Mrigya, Mridushi” published by International Journal of Advanced Research in Computer and Communication Engineering Vol. 5, Issue 1, January 2016
- [8] David J. Wetson, David J. Hand, M Adams, Whitrow and Piotr Juszczak “Plastic Card Fraud Detection using Peer Group Analysis” Springer, Issue 2008.

APPENDICES

```
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns
from matplotlib import gridspec
```

```
[ ] data = pd.read_csv("/content/drive/MyDrive/creditcard.csv")
```

```
data.head()
```

	Time	V1	V2	V3	V4	V5	V6	V7	V8	V9	...	V21	V22	V23	V24	V25	V26
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	0.128539	-0.189111
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	0.167170	0.125854
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	-0.327642	-0.139051
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	0.647376	-0.221922
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	-0.206010	0.502291

5 rows x 31 columns

```
print(data.shape)
print(data.describe())
```

```
(284807, 31)
```

	Time	V1	V2	V3	V4	V5	V6	V7	V8	V9
count	284807.000000	2.848070e+05	2.848070e+05	2.848070e+05	2.848070e+05	2.848070e+05	2.848070e+05	2.848070e+05	2.848070e+05	2.848070e+05
mean	94813.859575	3.918649e-15	5.682686e-16	-8.761736e-15	2.811118e-15	-1.552103e-15	2.040130e-15	-1.698953e-15	-1.893285e-16	-3.147640e-15
std	47488.145955	1.958696e+00	1.651309e+00	1.516255e+00	1.415869e+00	1.380247e+00	1.332271e+00	1.237094e+00	1.194353e+00	1.098632e+00
min	0.000000	-5.640751e+01	-7.271573e+01	-4.832559e+01	-5.683171e+00	-1.137433e+02	-2.616051e+01	-4.355724e+01	-7.321672e+01	-1.343407e+01
25%	54201.500000	-9.203734e-01	-5.985499e-01	-8.903648e-01	-8.486401e-01	-6.915971e-01	-7.682956e-01	-5.540759e-01	-2.086297e-01	-6.430976e-01
50%	84692.000000	1.810888e-02	6.548556e-02	1.798463e-01	-1.984653e-02	-5.433583e-02	-2.741871e-01	4.010308e-02	2.235804e-02	-5.142873e-02
75%	139320.500000	1.315642e+00	8.037239e-01	1.027196e+00	7.433413e-01	6.119264e-01	3.985649e-01	5.704361e-01	3.273459e-01	5.971390e-01
max	172792.000000	2.454930e+00	2.205773e+01	9.382558e+00	1.687534e+01	3.480167e+01	7.330163e+01	1.205895e+02	2.000721e+01	1.559499e+01

	...	V21	V22	V23	V24	\
count	...	2.848070e+05	2.848070e+05	2.848070e+05	2.848070e+05	
mean	...	1.473120e-16	8.042109e-16	5.282512e-16	4.456271e-15	
std	...	7.345240e-01	7.257016e-01	6.244603e-01	6.056471e-01	
min	...	-3.483038e+01	-1.093314e+01	-4.480774e+01	-2.836627e+00	
25%	...	-2.283949e-01	-5.423504e-01	-1.618463e-01	-3.545861e-01	
50%	...	-2.945017e-02	6.781943e-03	-1.119293e-02	4.097606e-02	
75%	...	1.863772e-01	5.285536e-01	1.476421e-01	4.395266e-01	
max	...	2.720284e+01	1.050309e+01	2.252841e+01	4.584549e+00	

	V25	V26	V27	V28	Amount	\
count	2.848070e+05	2.848070e+05	2.848070e+05	2.848070e+05	284807.000000	
mean	1.426896e-15	1.701640e-15	-3.662252e-16	-1.217809e-16	88.349619	
std	5.212781e-01	4.822270e-01	4.036325e-01	3.300833e-01	250.120109	
min	-1.029540e+01	-2.604551e+00	-2.256568e+01	-1.543008e+01	0.000000	
25%	-3.171451e-01	-3.269839e-01	-7.083953e-02	-5.295979e-02	5.600000	
50%	1.659350e-02	-5.213911e-02	1.342146e-03	1.124383e-02	22.000000	
75%	3.507156e-01	2.409522e-01	9.104512e-02	7.827995e-02	77.165000	
max	7.519589e+00	3.517346e+00	3.161220e+01	3.384781e+01	25691.160000	

	Class
count	284807.000000
mean	0.001727
std	0.041527
min	0.000000
25%	0.000000
50%	0.000000
75%	0.000000
max	1.000000

[8 rows x 31 columns]

```
▶ fraud = data[data['Class'] == 1]
  valid = data[data['Class'] == 0]
  outlierFraction = len(fraud)/float(len(valid))
  print(outlierFraction)
  print('Fraud Cases: {}'.format(len(data[data['Class'] == 1])))
  print('Valid Transactions: {}'.format(len(data[data['Class'] == 0])))
```

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

```
[ ] print('Amount details of the fraudulent transaction')
    fraud.Amount.describe()
```

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

```
▶ print('details of valid transaction')
  valid.Amount.describe()
```

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

Logistic Regression

```
[ ] model = LogisticRegression()
```

```
▶ # training the Logistic Regression Model with Training Data  
model.fit(X_train, Y_train)
```

```
⊙ LogisticRegression(C=1.0, class_weight=None, dual=False, fit_intercept=True,  
intercept_scaling=1, l1_ratio=None, max_iter=100,  
multi_class='auto', n_jobs=None, penalty='l2',  
random_state=None, solver='lbfgs', tol=0.0001, verbose=0,  
warm_start=False)
```

Accuracy Score

```
[ ] # accuracy on training data  
X_train_prediction = model.predict(X_train)  
training_data_accuracy = accuracy_score(X_train_prediction, Y_train)
```

```
▶ print('Accuracy on Training data : ', training_data_accuracy)
```

```
⊙ Accuracy on Training data : 0.9415501905972046
```

```
[ ] # accuracy on test data  
X_test_prediction = model.predict(X_test)  
test_data_accuracy = accuracy_score(X_test_prediction, Y_test)
```

```
[ ] print('Accuracy score on Test Data : ', test_data_accuracy)
```

```
Accuracy score on Test Data : 0.9390862944162437
```