A PROJECT REPORT

on

**CREDIT CARD FRAUD DETECTION**

Using

MACHINE LEARNING ALGORITHMS

Submitted in partial fulfillment for the requirement of the award of

TRAINING

IN

Data Analytics, Machine Learning and AI using Python



Submitted by

**ANUPAM SHAW** (B. P. PODDAR INSTITUTE OF MANAGEMENT & TECHNOLOGY, Kolkata)

(CSE, 2nd Year)

Under the guidance of

**Mr. Bipul Shahi**

**ACKNOWLEDGEMENT**

My sincere gratitude and thanks towards my project paper guide Mr. Bipul Shahi, Corporate Trainer, Developer, IOT, Artificial Intelligence, Robotics, Cloud Computing, Android Apps !!!

It was only with his backing and support that I could complete the report. He provided me all sorts of help and corrected me if I ever seemed to make mistakes. I have no such words to express my gratitude. I acknowledge my sincere gratitude to the HOD of CSE, BPPIMT, Kolkata. She gave me the permission to do the project work. Without her support I couldn’t even start the work. So, I am grateful to her. I acknowledge my sincere gratitude to the lecturers, research scholars and the lab technicians for their valuable guidance and helping attitude even in their very busy schedule.

And at last but not the least, I acknowledge my dearest parents for being such a nice source of encouragement and moral support that helped me tremendously in this aspect. I also declare to the best of my knowledge and belief that my Project Work has not been submitted anywhere else.

**INTRODUCTION**

Credit card frauds are easy and friendly targets. E-commerce and many other online sites have increased the online payment modes, increasing the risk for online frauds. Increase in fraud rates, researchers started using different machine learning methods to detect and analyse frauds in online transactions.

The main aim of the project is to design and develop a novel fraud detection method for Streaming Transaction Data, with an objective, to analyse the past transaction details of the customers and extract the behavioural patterns.

**OBJECTIVE**

The objective in this project is to build machine learning models to classify or identify fraudulent card transactions from a given card transactions data.

**TECHNOLOGY & CONCEPTS**

**Machine Learning :**

Learning algorithms are widely used in computer vision applications. Before considering image related tasks, we are going to have a brief look at basics of machine learning.

Machine learning has emerged as a useful tool for modelling problems that are otherwise difficult to formulate exactly. Classical computer programs are explicitly programmed by hand to perform a task. With machine learning, some portion of the human contribution is replaced by a learning algorithm. As availability of computational capacity and data has increased, machine learning has become more and more practical over the years, to the point of being almost ubiquitous.

It can be used in two ways:

* *Supervised Learning*
* *Unsupervised Learning*

**Naive Bayes Classifier :**

A Naive Bayes classifier is a probabilistic machine learning model that’s used for classification task. The crux of the classifier is based on the Bayes theorem. They are fast and easy to implement but their biggest disadvantage is that the requirement of predictors to be independent. In most of the real life cases, the predictors are dependent, this hinders the performance of the classifier.

**Random Forest Classifier :**

A random forest is a meta estimator that fits a number of decision tree classifiers on various sub-samples of the dataset and uses averaging to improve the predictive accuracy and control over-fitting. The sub-sample size is controlled with the ‘max\_samples’ parameter if ‘bootstrap=True’ (default), otherwise the whole dataset is used to build each tree.

**Logistic Regression :**

Logistic Regression is a Machine Learning algorithm which is used for the classification problems, it is a predictive analysis algorithm and based on the concept of probability. We can call a Logistic Regression a Linear Regression model but the Logistic Regression uses a more complex cost function, this cost function can be defined as the ‘**Sigmoid function**’ or also known as the ‘logistic function’ instead of a linear function. The hypothesis of logistic regression tends it to limit the cost function between 0 and 1. Therefore linear functions fail to represent it as it can have a value greater than 1 or less than 0 which is not possible as per the hypothesis of logistic regression.

### **Description of Dataset**

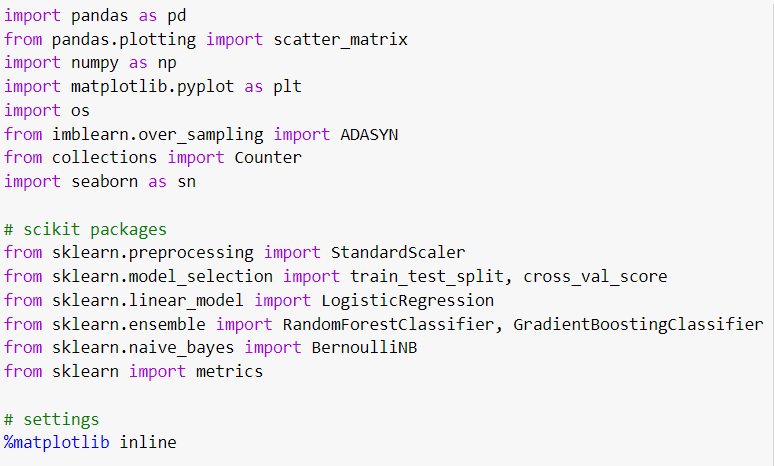
The [dataset](https://www.kaggle.com/mlg-ulb/creditcardfraud/data) contains two-days credit card transactions made in September 2013 by European cardholders. The dataset is highly unbalanced with a low percentage of fraudulent transactions within several records of normal transactions. The positive class (frauds) account for 0.172% (492 frauds out of 284,807 transactions) of all transactions.

Features V1, V2, ... V28 are the principal components obtained with PCA, the only features which have not been transformed with PCA are 'Time' and 'Amount'. Feature 'Time' contains the seconds elapsed between each transaction and the first transaction in the dataset. Feature 'Class' is the target variable with value 1 in case of fraud and 0 otherwise.

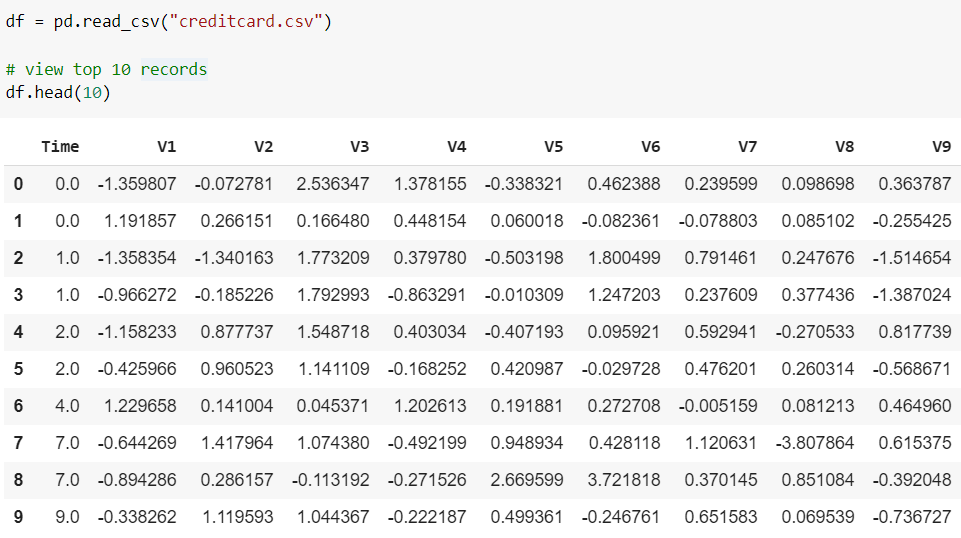
The dataset has been collected and analysed during a research collaboration of Worldline and the Machine Learning Group ([http://mlg.ulb.ac.be](http://mlg.ulb.ac.be/)) of ULB (Université Libre de Bruxelles) on big data mining and fraud detection.

**METHODOLOGY**

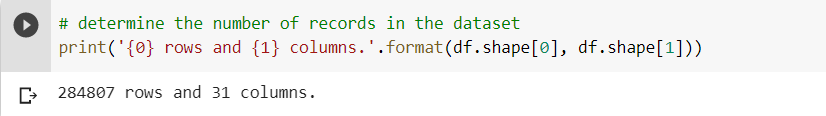
**Step-1: Import Libraries**

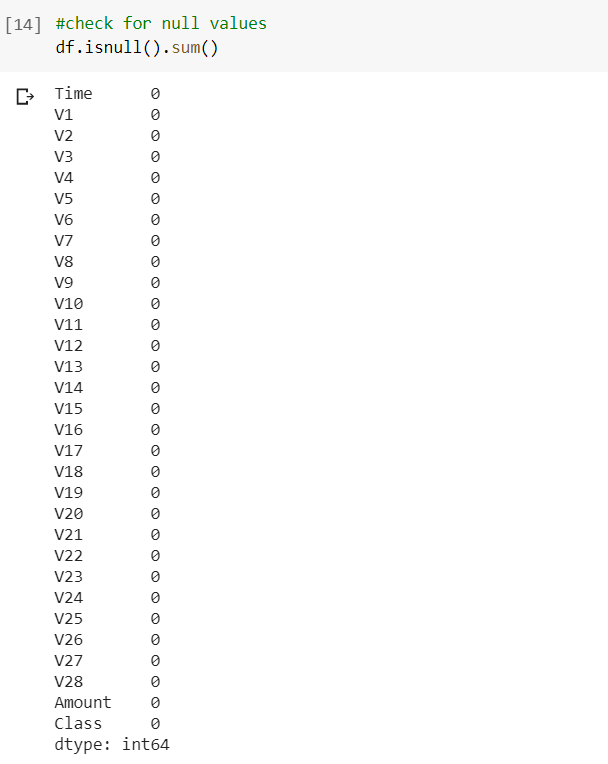


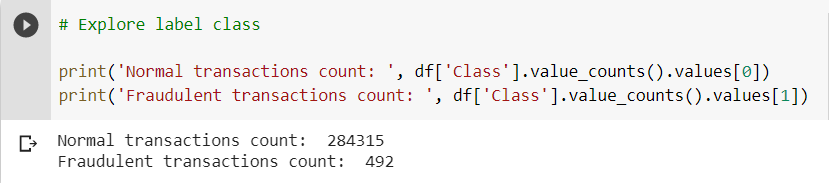
**Step-2: Import the dataset**

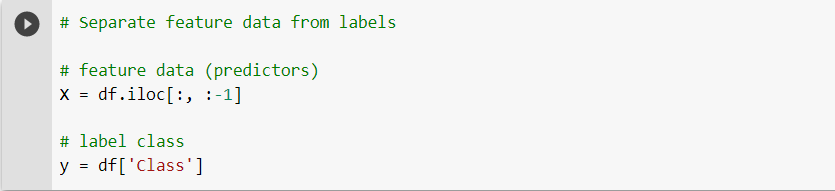


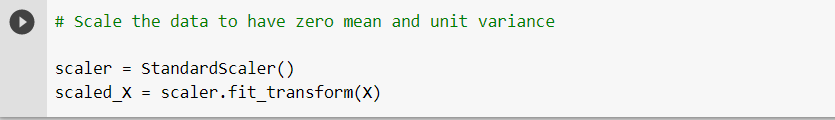
**Step-3: Exploring the Dataset**



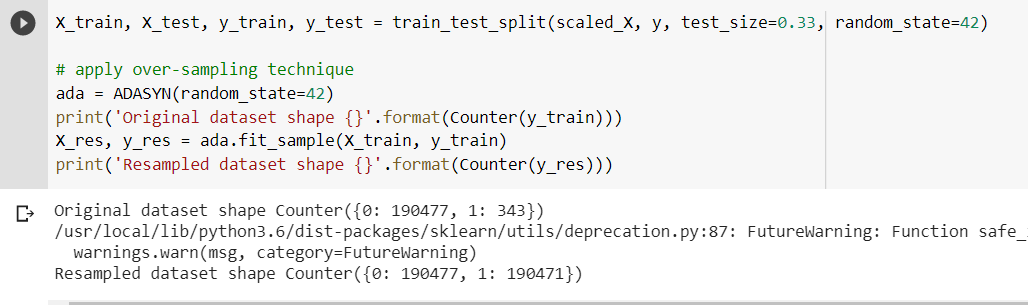




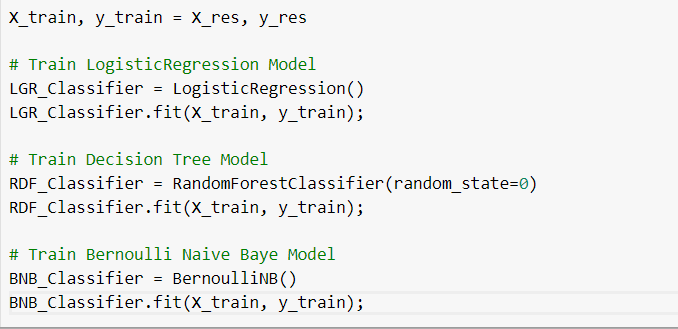




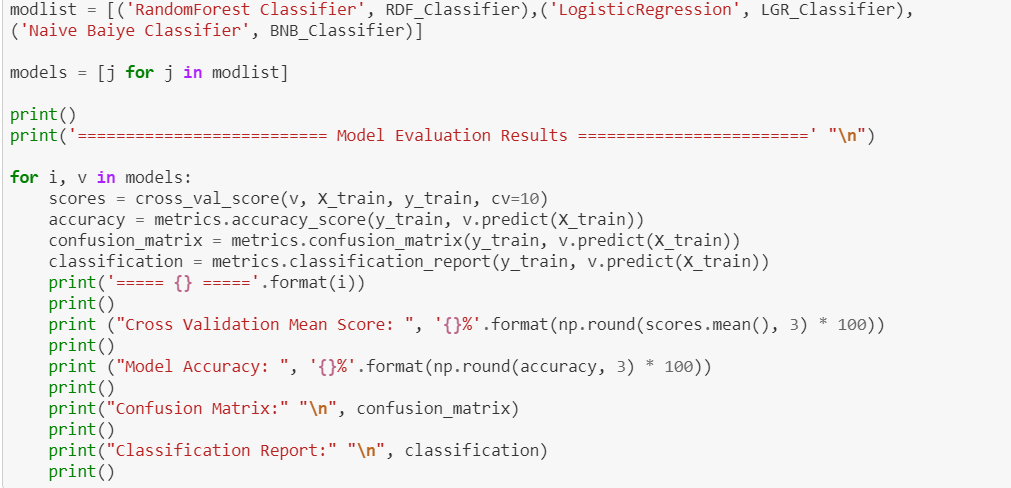
**Step-4: Splitting data for training and testing, & over-sampling data**



**Step-5: Training our model**

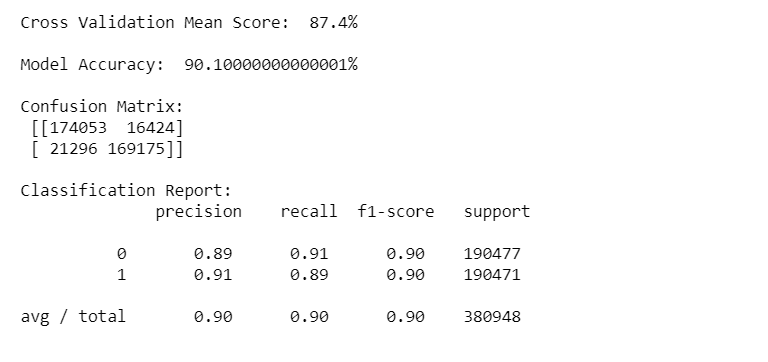


**Step-6: Evaluating our model**

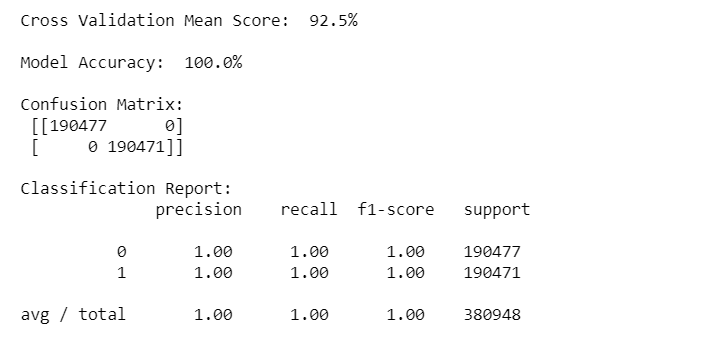


**Evaluation Output:**

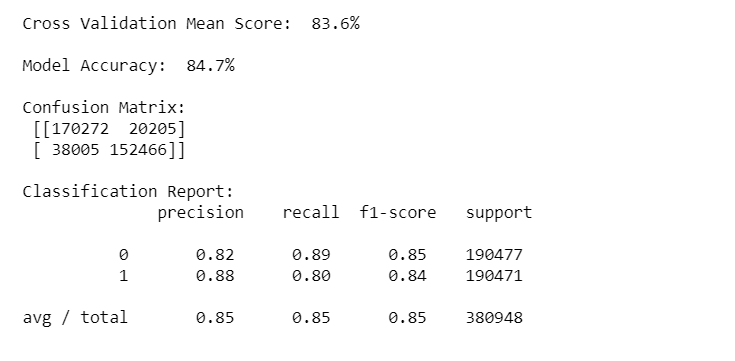
Using Logistic Regression Classifier:



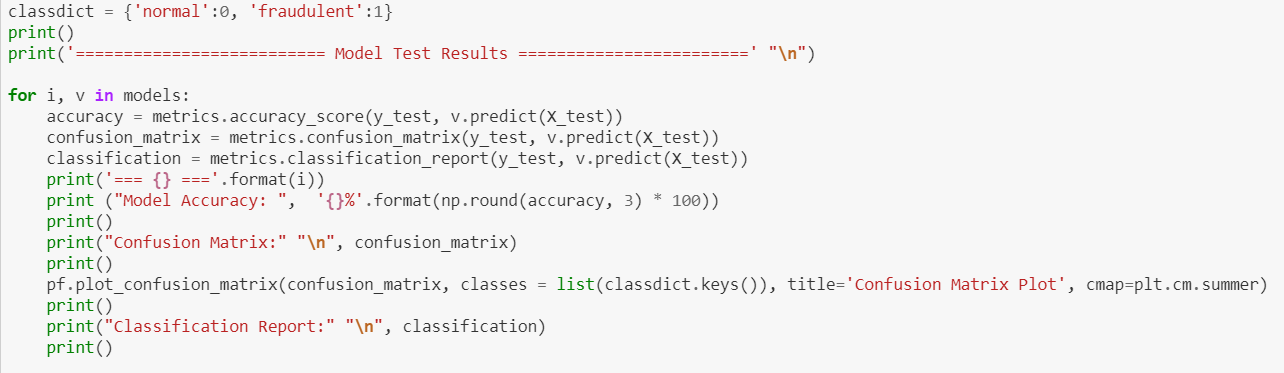
Using Random Forest Classifier:



Using Naïve-Bayes Classifier:

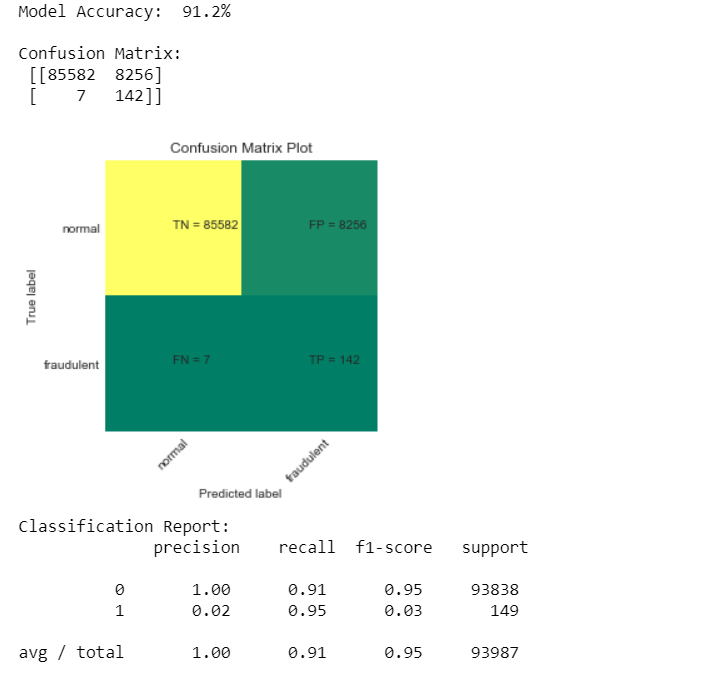


**Step-7: Testing our Model**

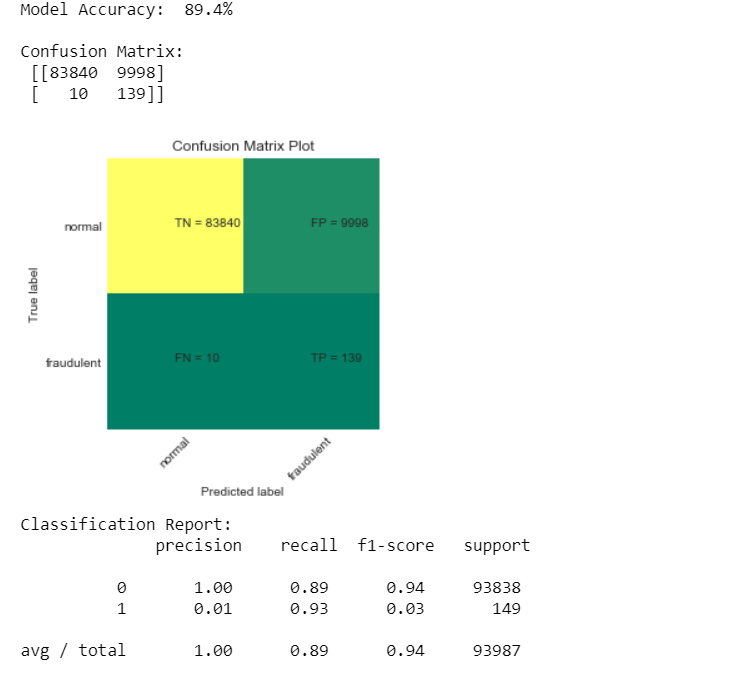


**Test Results:**

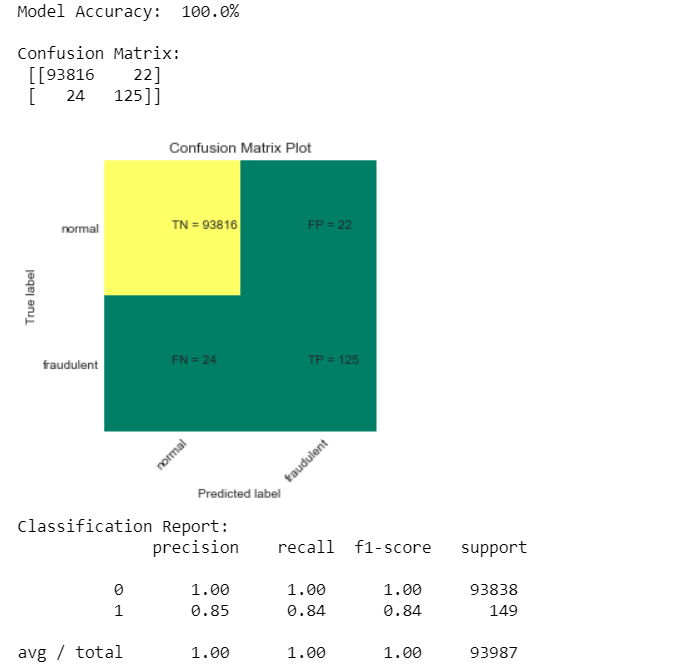
Using Logistic Regression classifier:



Using Naïve-Bayes Classifier:

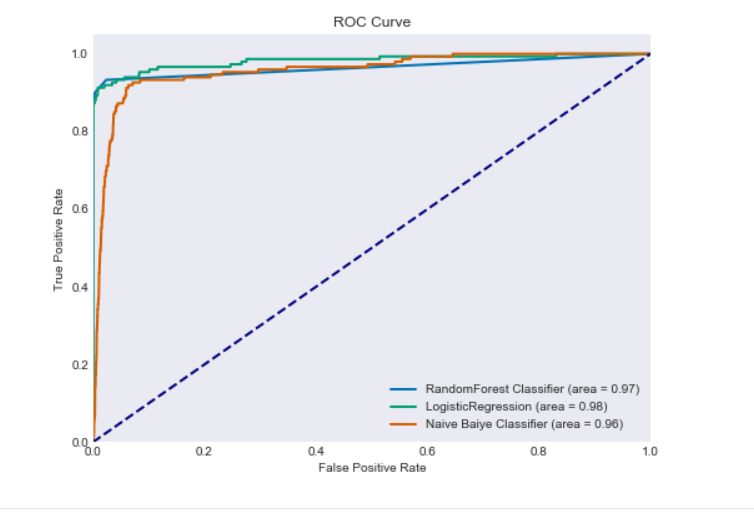


Using Random-Forest Classifier:



ROC Curve:





**CONCLUSION**

I have used three types ML algorithms for this project- Random Forest Classifier, Naïve Bayes Classifier & Logistic Regression Classifier. I have plotted the data in form of confusion matrix and got a very good model accuracy for Random Forest algorithm. I have used the ROC curve (if value close to 1, then model is good), and found Random forest & Logistic Regression algorithm model ROC value very close to 1.

Hence, we can say that the Random forest algorithm proved to have more descent success rate than Naïve Bayes and Logistic Regression for this dataset.

**BIBLIOGRAPHY**

* <https://www.kaggle.com/>
* <https://towardsdatascience.com/>
* Hands-on Machine Learning with Scikit-learn, keras and Tensorflow; by Aurelien Geron