
Machine Learning project report on

# Telemarketing in Banking Sector

*Submitted by:*

Dare Devils

Aishwarya Suresh
Bhimesh
Pavankumar Kotha
Elizabeth Gunde
Likith Reddy

# Acknowledgement

We sincerely thank **Smart Bridge** and **IBM** for providing us a platform to develop our skills in the domain of Machine Learning. We convey our heartfelt thanks to **Mr. P. Mohan**, our trainer for making every session interactive and interesting. We also thank the mentors, **Mr. Sai Radha Krishna**, **Mrs. G. Moulika**, **Mr. Gokul Harikumar**, **Mr. Rahul Pavithran**, **Mr.Bharath Nandan** and **Mr. Anil Choudary** for being patient and guiding us all through the program.

We thank our respective institutions for permitting us to attend this program which is sure to play a significant part in the interviews that we are to face soon. I thank the Jawaharlal Nehru Technological University Hyderabad for providing us with necessary infrastructure to do my work.

Dare Devils

# Certificate

This is to certify that , enrolled in the B.tech degree programme ( ) of the has successfully completed the four week internship cum hands-on training program conducted by Smartbridge at Jawaharlal Nehru Technological University in Introduction to Machine Learning using Python 3' during the time period from 6th May, 2019 to 29th May, 2019 under the guidance of Mr. P Mohan, Senior Data Consultant. During this period of internship with us she was found punctual, hardworking and inquisitive.

_____

Mentor

# Contents

# 1.1 Introduction

Marketing is one of the most important department in every industry and banking is no exception to this. It was in the 1970s that the banking sector started to impart marketing to stand the increasing competition the sector. Banking marketing is applying marketing theories and practices in the banking sector.

Marketing in the banking sector can be done in multiple ways using tools like telecommunications, advertisements, social media platforms like face book, twitter, mails, etc.,
All of these aim at getting the customer familiarized with the existing and upcoming plans that are the bank offers. This not only brings in new customers but also aids in increasing the profits to the bank.

Different Strategies are used by banks to expand their market like Digital marketing, telemarketing and many more.
The given data set is based on direct marketing campaigns (phone calls) of a Portuguese banking institution.

# 1.2 Objectives of Research:

The given data set is that of the direct marketing campaign conducted by a Portuguese banking institution. The given data is to analyzed. Various objectives of the given data set are as follows:

a) To study the given data

b) To apply data cleaning methods to remove unknown data from the given data set

c) To apply various classification models on the given data set to produce a model with maximum efficiency.

d) Test the designed model's working

e) Draw conclusions from the developed model

f) Predict whether the plan (product) will be subscribed by the customer or not.

# 1.3 Problem Statement

The data is related with direct marketing campaigns of a Portuguese banking institution. The marketing campaigns were based on phone calls. Often, more than one contact to the same client was required, in order to access if the product (bank term deposit) would be ('yes') or not ('no') subscribed. It is to be predicted whether a customer will subscribe to the product or not.

# 1.4 Industry Profile

A financial institution that provides banking and other financial services to customers is called a bank. It is generally understood as an institution that provides fundamental banking services like accepting deposits and providing loans. A banking system is also referred as a system provided by the bank which offers cash management services for customers, reporting the transactions of their accounts and portfolios, throughout the day.

Banking sector offers a wide range of facilities and opportunities to their customers. It safeguards the money and valuables and provide loans, credit, and payment services, such as checking accounts, money orders, and cashier's cheques. It also offers investment and insurance products. Though a variety of models for cooperation and integration among finance industries have emerged diminishing the distinctions between banks, insurance companies, and securities firms, banks continue to maintain and perform their primary role, which is accepting deposits and lending funds from these deposits.

Banks differ in the number of services and the clients they serve. Commercial banks, dominate the industry as they offer a full range of services for individuals, businesses, and governments. They come in a wide range of sizes, from large global banks to regional and community banks.

Global banks engage in international lending and foreign currency trading, and other high end services. Regional banks are numerous with Automated Teller Machine (ATM) located at various places providing banking services to individuals.

Lately banks have become more oriented toward marketing and sales. This in turn creates a necessity for the employees to know about all types of products and services offered by banks. Community banks are preferred by most individuals and small businesses as they are based locally and offer more personal attention. Recently online banks have emerged providing all banking services entirely over the Internet. This has made banking much easier, avoiding long queues and bank timings. Traditional banks have also expanded to offer online banking. There are also Internet-only banks, which provide all services only through the internet.

# 2. Review of Literature

A marketing strategy brings product development, promotion, distribution, pricing, relationship management and other elements together; identifies the firm's marketing goals and explains how they can be achieved within the stipulated timeframe. It determines the choice of target market segments, positioning, marketing mix, and allocation of resources. It is one of the most important areas that needs to be carefully examined by the policy makers of banks. Improvement in the performance and ensuring sustainable growth of banks has become a necessity as competition in the banking industry intensifies.

A sound marketing strategy not only operates in an environment where service quality and financial returns are perceived as the essential criteria from customers' viewpoint, but also competes with conventional banks which are known to have better experience and expertise in the banking business. Banks now have a firm belief that effective marketing strategies applied in the bank reducing the cost of services provided to customers, and raise the quality of banking services provided, and to influence the response to the client alone can assure the future of banking business. The paper is review of marketing strategies prevalent in Banking Sector.

In the present era where mature and intense competitive pressures are persistent, it is essential that banks maintain a loyal customer base. Thus banks should exert more focus on the relation with the consumer and market needs, and the consequent marketing strategies. As the customer is the source of a company's revenue, marketing strategy is closely linked with sales.

In the present data set the marketing strategy used is **telemarketing** which comes under direct marketing. Telemarketing is a method of selling products and services over the telephone. In the banking sector the product is nothing but the plans or packages introduced by the bank. It has both advantages and disadvantages. The advantages are it is easy to reach out to customers and it is cost effective if done successfully. The disadvantages are that it has a bad reputation and some of the startup costs are expensive.

Telemarketing may either by carried out by telemarketers from call centers, or increasingly, by automated telephone calls or robocalls. The intrusive nature of telemarketing, as well as reports of scams and fraud over the telephone, has developed a growing hurdle against this direct marketing practice. Telemarketing may also be referred to as **telesales** or **inside sales.**

# 3. Data Collection

The data set given is related to direct marketing campaigns (telemarketing) of a Portuguese banking institution. It was taken from the website **data.world**. The website provides various datasets from various domains.

The following link will lead to the description of the data set
https://data.world/data-society/bank-marketing-data

The given data set consists of 21 features and 41,188 cases. The features are as follows:
1. age (numeric)
2. job : type of job (categorical: 'admin.', 'blue-collar', 'entrepreneur', 'housemaid', 'management', 'retired', 'self-employed', 'services', 'student', 'technician', 'unemployed', 'unknown')
3. marital : marital status (categorical: 'divorced', 'married', 'single', 'unknown'; note: 'divorced' means divorced or widowed)
4. education (categorical: 'basic.4y', 'basic.6y', 'basic.9y', 'high.school', 'illiterate', 'professional.course', 'university.degree', 'unknown')
5. default: has credit in default? (categorical: 'no', 'yes', 'unknown')
6. housing: has housing loan? (categorical: 'no', 'yes',

'unknown')
7. loan: has personal loan? (categorical: 'no', 'yes', 'unknown')

## Related with the last contact of the current campaign:

8. contact: contact communication type (categorical: 'cellular', 'telephone')
9. month: last contact month of year (categorical: 'jan', 'feb', 'mar', ..., 'nov', 'dec')
10. day_of_week: last contact day of the week (categorical: 'mon', 'tue', 'wed', 'thu', 'fri')
11. duration: last contact duration, in seconds (numeric). Important note: this attribute highly affects the output target (e.g., if duration=0 then y='no'). Yet, the duration is not known before a call is performed. Also, after the end of the call y is obviously known. Thus, this input should only be included for benchmark purposes and should be discarded if the intention is to have a realistic predictive model.

## Other attributes:

12. campaign: number of contacts performed during this campaign and for this client (numeric, includes last contact)
13. pdays: number of days that passed by after the client was last contacted from a previous campaign (numeric; 999 means client was not previously contacted)
14. previous: number of contacts performed before this campaign and for this client (numeric)

15. poutcome: outcome of the previous marketing campaign (categorical: 'failure','nonexistent','success')

# Social and economic context attributes:

16. emp.var.rate: employment variation rate - quarterly indicator (numeric)
17. cons.price.idx: consumer price index - monthly indicator (numeric)
18. cons.conf.idx: consumer confidence index - monthly indicator (numeric)
19. euribor3m: euribor 3 month rate - daily indicator (numeric)
20. nr.employed: number of employees - quarterly indicator (numeric)

Output variable (desired target):
21. y - has the client subscribed a term deposit? (binary: 'yes', 'no')

Out of the given features the output is given by the 21$^{st}$ feature is the output feature and the remaining are input features. The output feature **y** tells whether the customer will subscribe the given plan or not.

# 4. Methodology

The given data set consists of 41,188 cases and 21 features out of which one is output and the remaining 20 are input features. As the first step the a few essential packages like **pandas, numpy, matplotlib, seaborn** and **os** are imported into the Jupiter notebook. Now the given data set is imported to the Jupiter note book using pandas. It can be observed that most of the data in the given data set is categorical.

The second step is to check if there is any missing data in the given data set. On applying **isnull** function we observe that no visible missing data is present. But on further analyzations it can be observed that six features have **unknown** values which have to be sorted before applying classification models on the given data. This process is called data cleaning. As the data is categorical in nature, the unknown values are replaced by the mode category of the feature.

Once the data is cleaned, different classification models like **Knn, Longistic Regression, Support Vector Classification, Decission Tree,** and **Random Forest** are applied on the cleaned data. The various metrics of classification models like **accuracy, ROC score, AUC, Recall score, ROC curve and Confusion matrix** of the different classification models are

compared for various splits of training and testing data and the model with highest efficiency is chosen to draw conclusions regarding the given dataset.

# 4.1 Exploratory Data Analysis:

During the process of data cleaning we can observe that the features job, marital, education, default, housing, and loan having values labeled as unknown. As the data in these features is categorical in nature, the unknown values are to be replace by the mode category in general. But applying this method directly will draw contradictions. To avoid there a few direct imputations are made to the dataset based on the observations made from the cross tables drawn between the features that contain unknown values.

In [7]:
```
df1.loc[(df1['age']>60) & (df1['job']=='unknown'), 'job'] = 'retired'
df1.loc[(df1['education']=='unknown') & (df1['job']=='management'), 'education'] = 'university.degree'
df1.loc[(df1['education']=='unknown') & (df1['job']=='services'), 'education'] = 'high.school'
df1.loc[(df1['education']=='unknown') & (df1['job']=='housemaid'), 'education'] = 'basic.4y'
df1.loc[(df1['job'] == 'unknown') & (df1['education']=='basic.4y'), 'job'] = 'blue-collar'
df1.loc[(df1['job'] == 'unknown') & (df1['education']=='basic.6y'), 'job'] = 'blue-collar'
df1.loc[(df1['job'] == 'unknown') & (df1['education']=='basic.9y'), 'job'] = 'blue-collar'
df1.loc[(df1['job']=='unknown') & (df1['education']=='professional.course'), 'job'] = 'technician'
df1.loc[(df1['job'] == 'unknown') & (df1['education']=='high.school'), 'job'] = 'admin.'
df1.loc[(df1['job'] == 'unknown') & (df1['education']=='university.degree'), 'job'] = 'management'
df1.loc[(df1['education']=='unknown') & (df1['job']=='admin.'), 'education'] = 'university.degree'
df1.loc[(df1['education']=='unknown') & (df1['job']=='blue-collar'), 'education'] = 'basic.9y'
df1.loc[(df1['education']=='unknown') & (df1['job']=='entrepreneur'), 'education'] = 'university.degree'
df1.loc[(df1['education']=='unknown') & (df1['job']=='retired'), 'education'] = 'basic.4y'
df1.loc[(df1['education']=='unknown') & (df1['job']=='self-employed'), 'education'] = 'university.degree'
df1.loc[(df1['education'] == 'unknown') & (df1['job']=='student'), 'education'] = 'high.school'
df1.loc[(df1['education'] == 'unknown') & (df1['job']=='technician'), 'education'] = 'professional.course'
df1.loc[(df1['education'] == 'unknown') & (df1['job']=='unemployed'), 'education'] = 'university.degree'
df1['job'].replace(['unknown'],['admin.'],inplace=True)
df1['education'].replace(['unknown'],['university.degree'],inplace=True)
```

After the direct imputations the remaining unknown data was replaced by mode categories of each feature respectively.

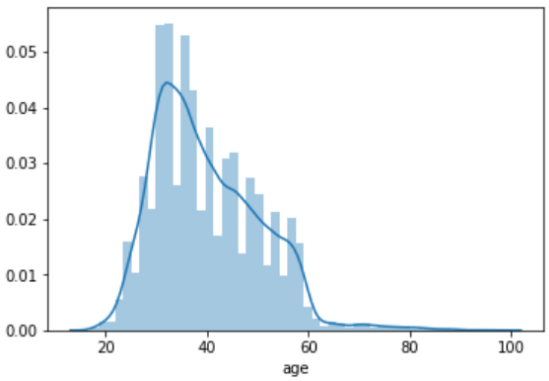
Later when the data is free from unknown data, the categorical data is converted to continuous using the **replace** function.

# 4.1.1 Distribution Plots

Now visualization technique of using distribution plots are plotted. Here is the distribution plot for the feature

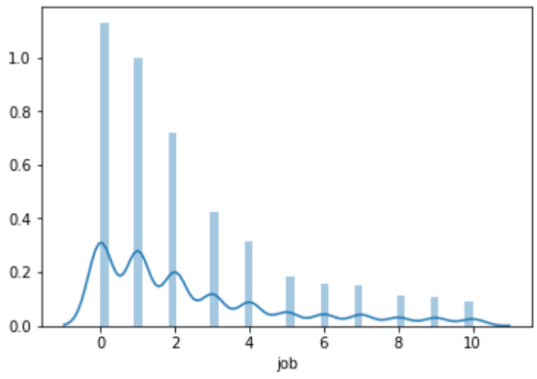
**a) age:**

```
Out[59]: <matplotlib.axes._subplots.AxesSubplot at 0x1da4fb793c8>
```



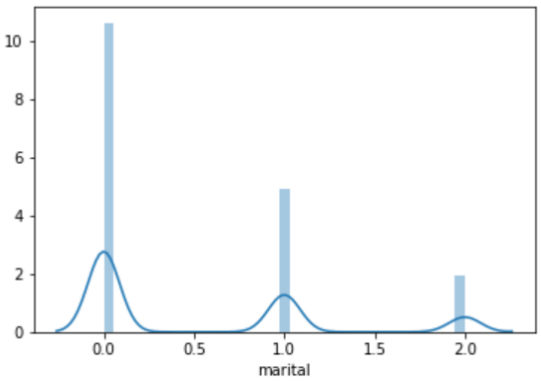

**b) job:**

```
Out[60]: <matplotlib.axes._subplots.AxesSubplot at 0x1da4fc17ef0>
```



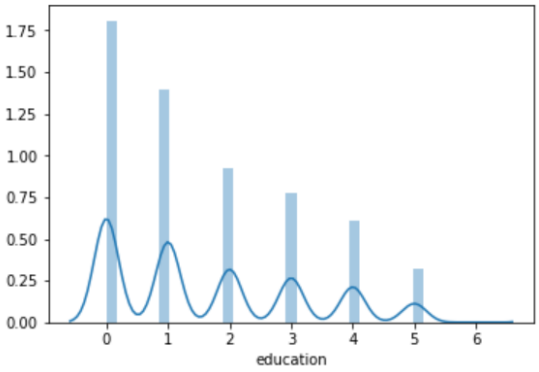
## c) marital:

```
Out[62]: <matplotlib.axes._subplots.AxesSubplot at 0x1da4fbd2240>
```



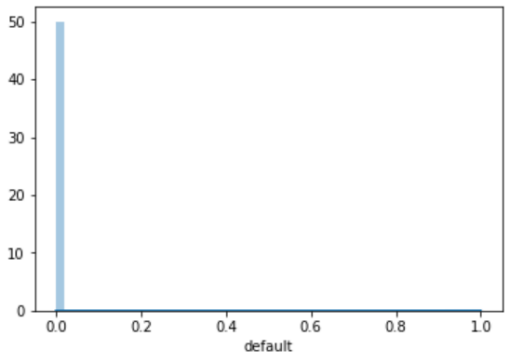

## d) education:

```
Out[63]: <matplotlib.axes._subplots.AxesSubplot at 0x1da50932cf8>
```



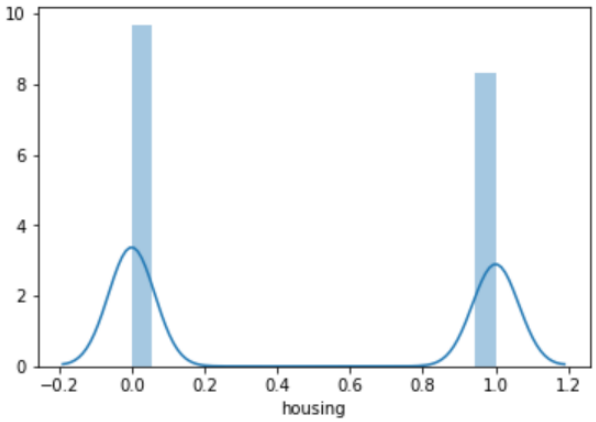

## e) default:

```
Out[64]: <matplotlib.axes._subplots.AxesSubplot at 0x1da508d7978>
```



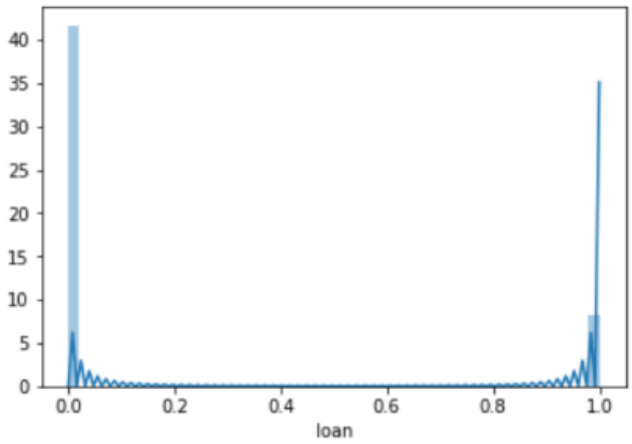
## f) housing :

```
Out[47]: <matplotlib.axes._subplots.AxesSubplot at 0x1da4f65b278>
```



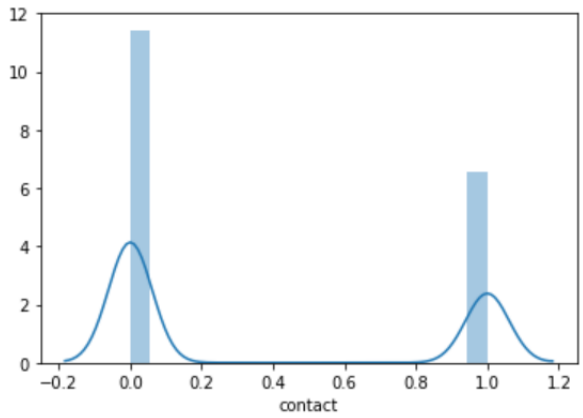

## g) loan:

```
Out[65]: <matplotlib.axes._subplots.AxesSubplot at 0x1da4fc
```



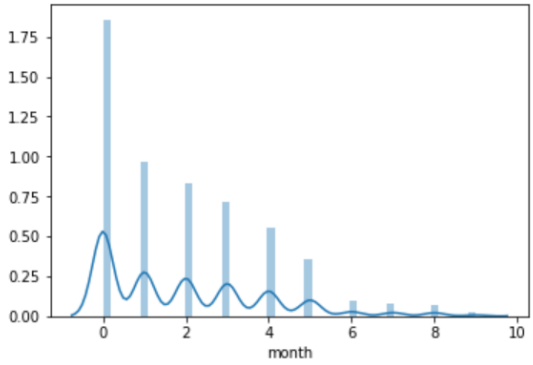

## h) contact:

```
Out[66]: <matplotlib.axes._subplots.AxesSubplot at 0x1da4f7b40f0>
```



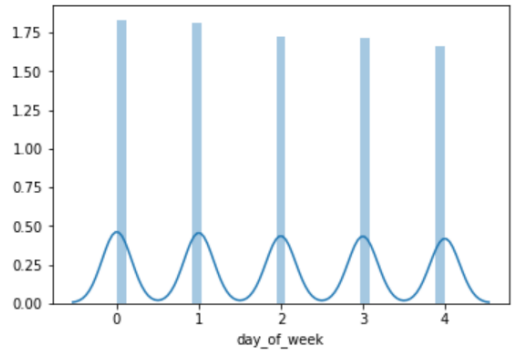
## i) month:

```
Out[67]: <matplotlib.axes._subplots.AxesSubplot at 0x1da50bb3da0>
```



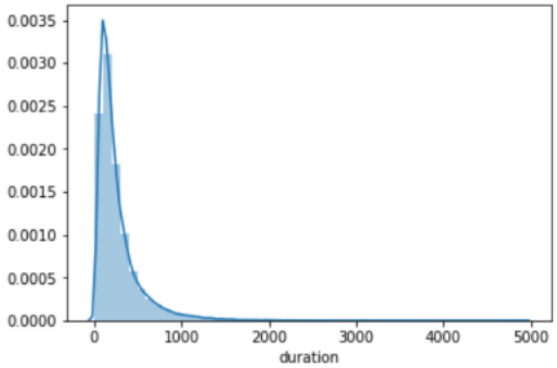

## j) day_of_week:

```
Out[68]: <matplotlib.axes._subplots.AxesSubplot at 0x1da50cdd470>
```



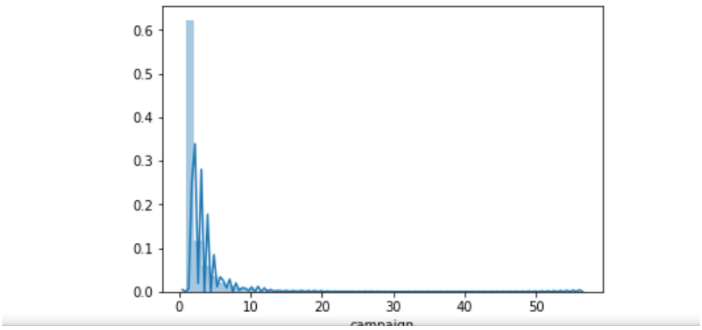

## k) duration:

```
Out[69]: <matplotlib.axes._subplots.AxesSubplot at 0x1da50d5b828>
```



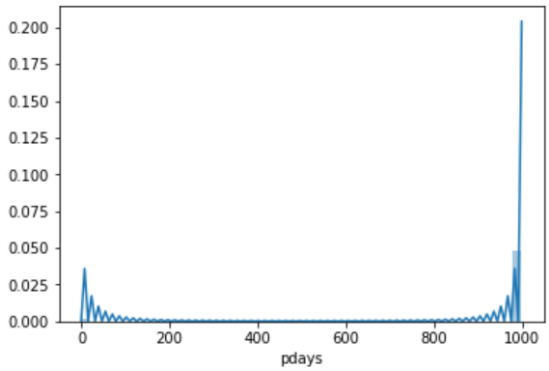
## l) campaign:

```
Out[70]: <matplotlib.axes._subplots.AxesSubplot at 0x1da50dfc710>
```



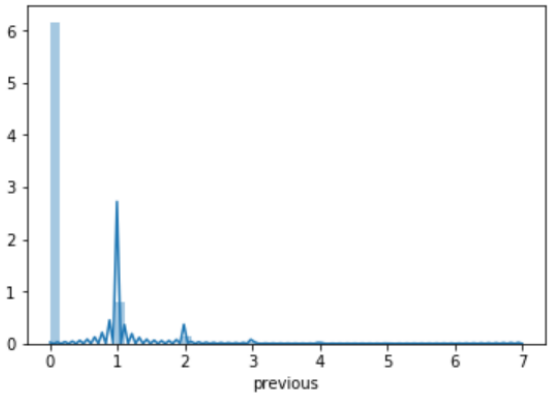

## m) pdays:

```
Out[72]: <matplotlib.axes._subplots.AxesSubplot at 0x1da4fb5cd68>
```



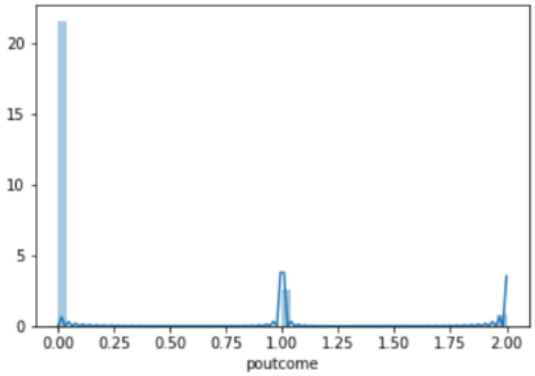

## n) previous:

```
Out[73]: <matplotlib.axes._subplots.AxesSubplot at 0x1da50f8c0f0>
```



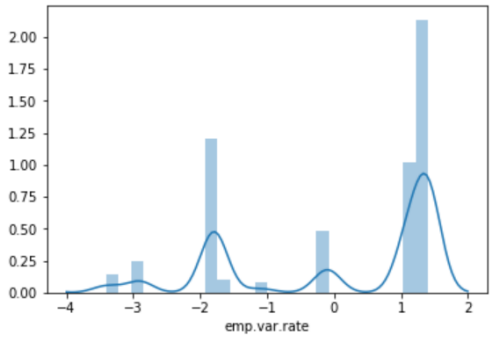
## o) poutcome:

```
Out[74]: <matplotlib.axes._subplots.AxesSubplot at 0x1da52036c18>
```



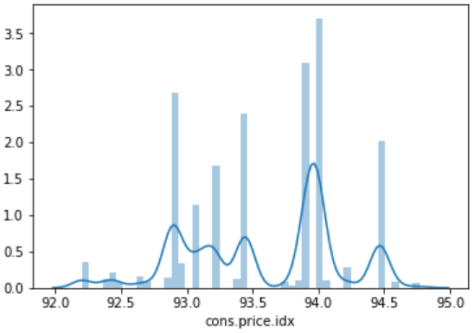

## p) emp.var.rate:

```
Out[75]: <matplotlib.axes._subplots.AxesSubplot at 0x1da52156080>
```



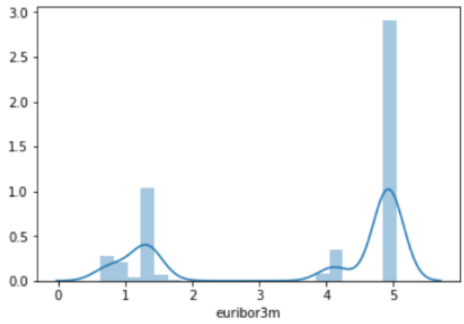

## q) cosn.price.idx:

```
Out[76]: <matplotlib.axes._subplots.AxesSubplot at 0x1da521c4828>
```



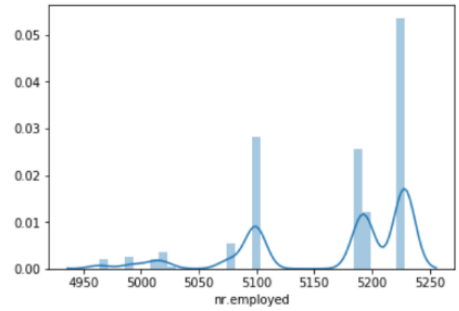
## r) eruribor3m:

```
Out[78]: <matplotlib.axes._subplots.AxesSubplot at 0x1da5235ddd8>
```



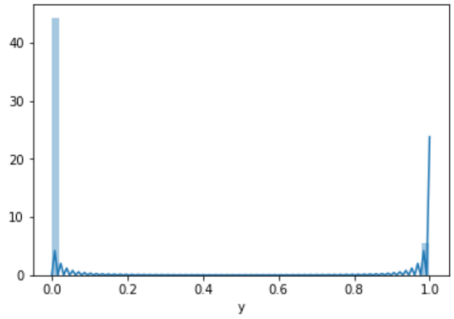

## s) nr.employed:

```
Out[48]: <matplotlib.axes._subplots.AxesSubplot at 0x1da4f72ec18>
```




## t) y:

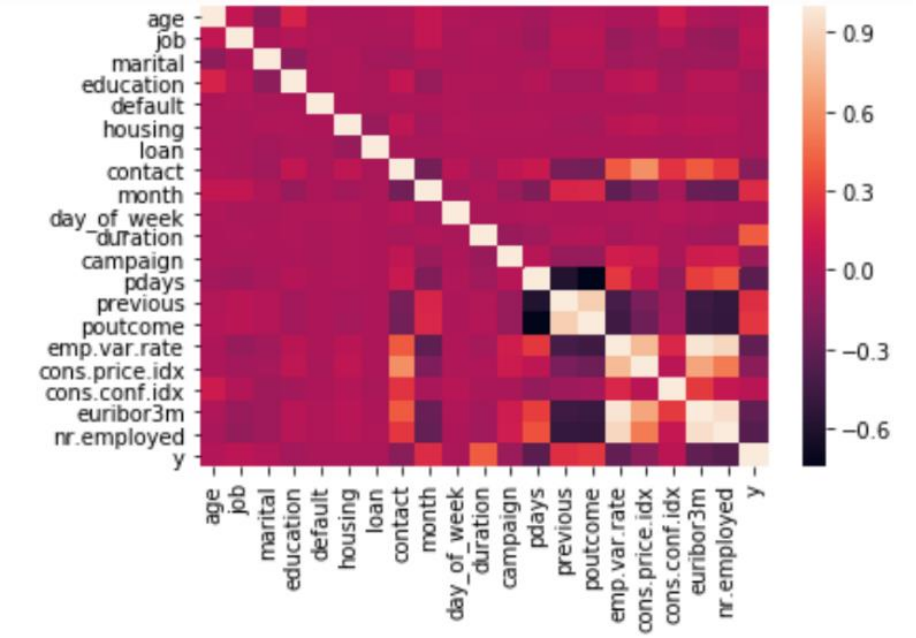
```
Out[79]: <matplotlib.axes._subplots.AxesSubplot at 0x1da52411748>
```



From the above distribution plots, it can be observed that no plot is completely normalized.

## 4.1.2 Heat map

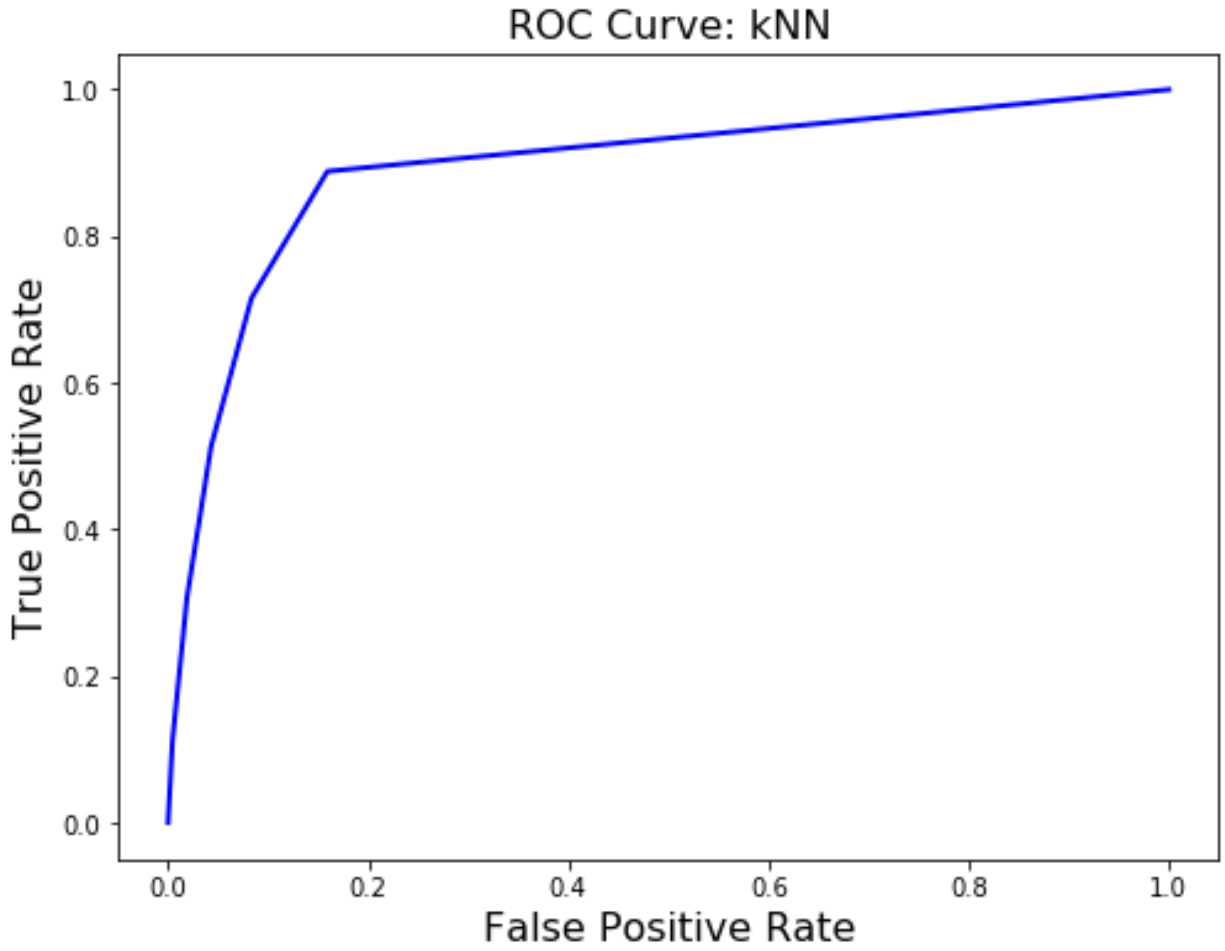
The following figure represents the heat map plotted between the features of the given data set.



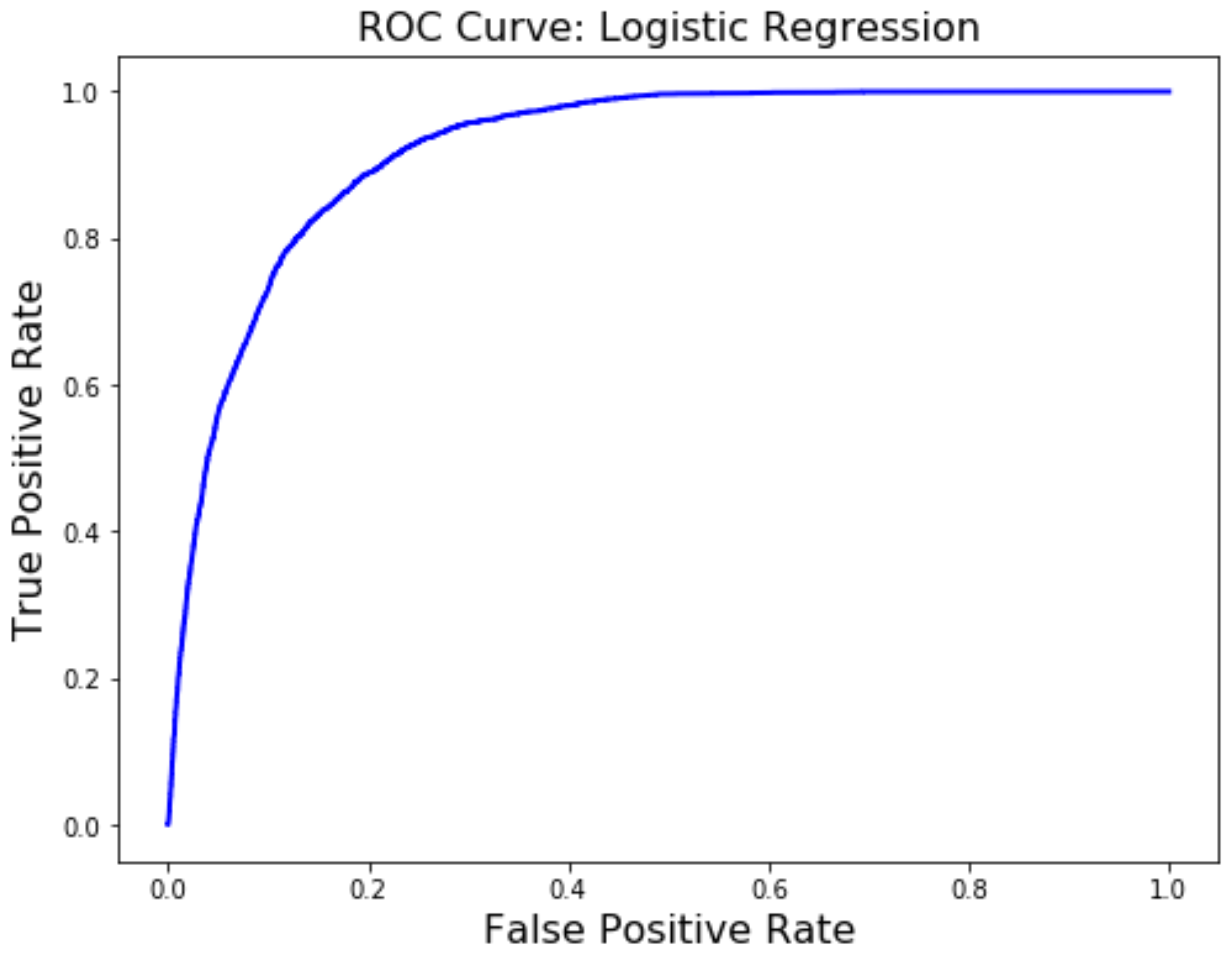

It can be observed from the above heatmap that the features duration, cons.conf.idx, euribor3m, emp.var.rate, poutcome campaign and contact have greater influence on the output variable.

# 4.1.3 ROC curves

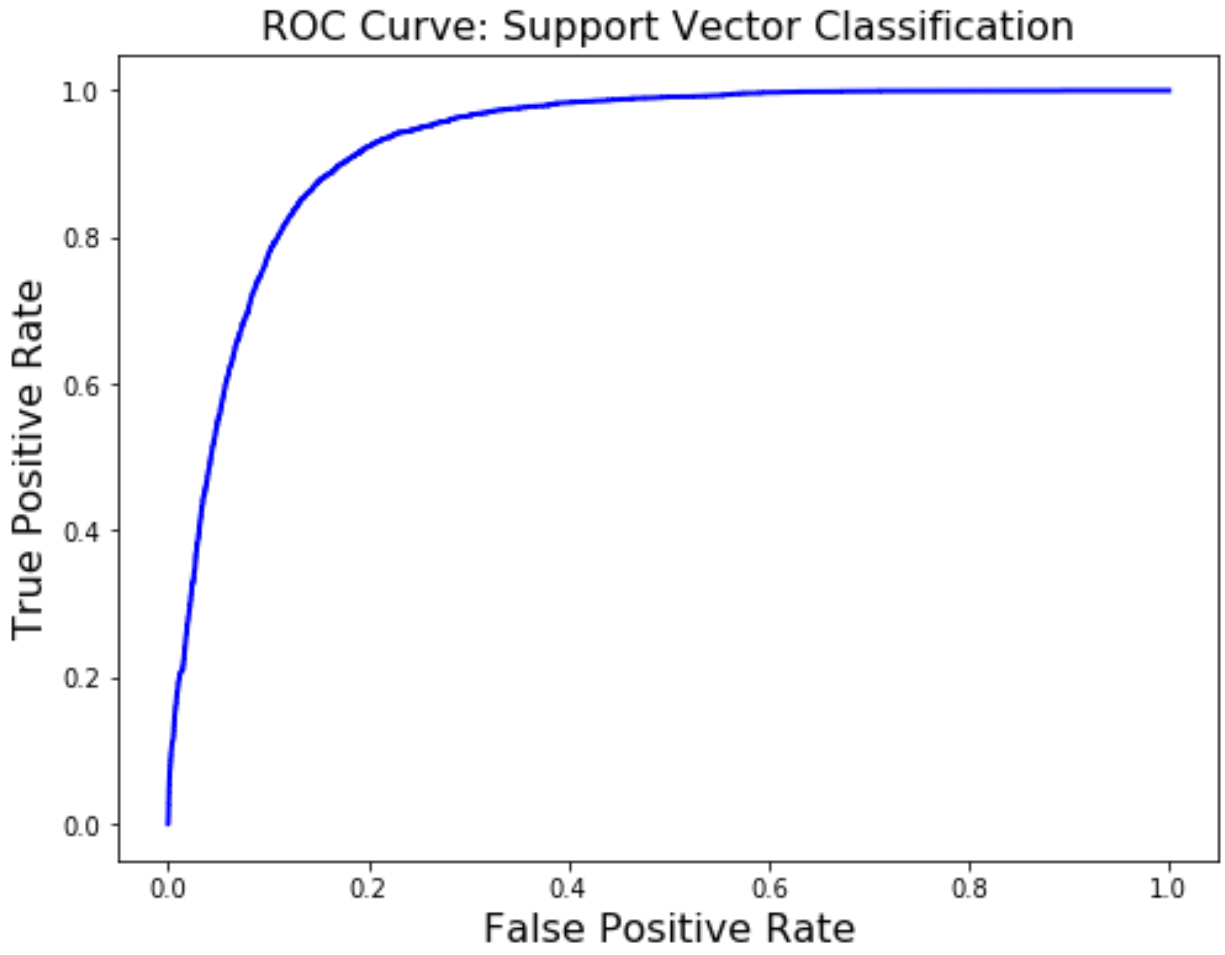
```
AUC Score (kNN): 0.89
```



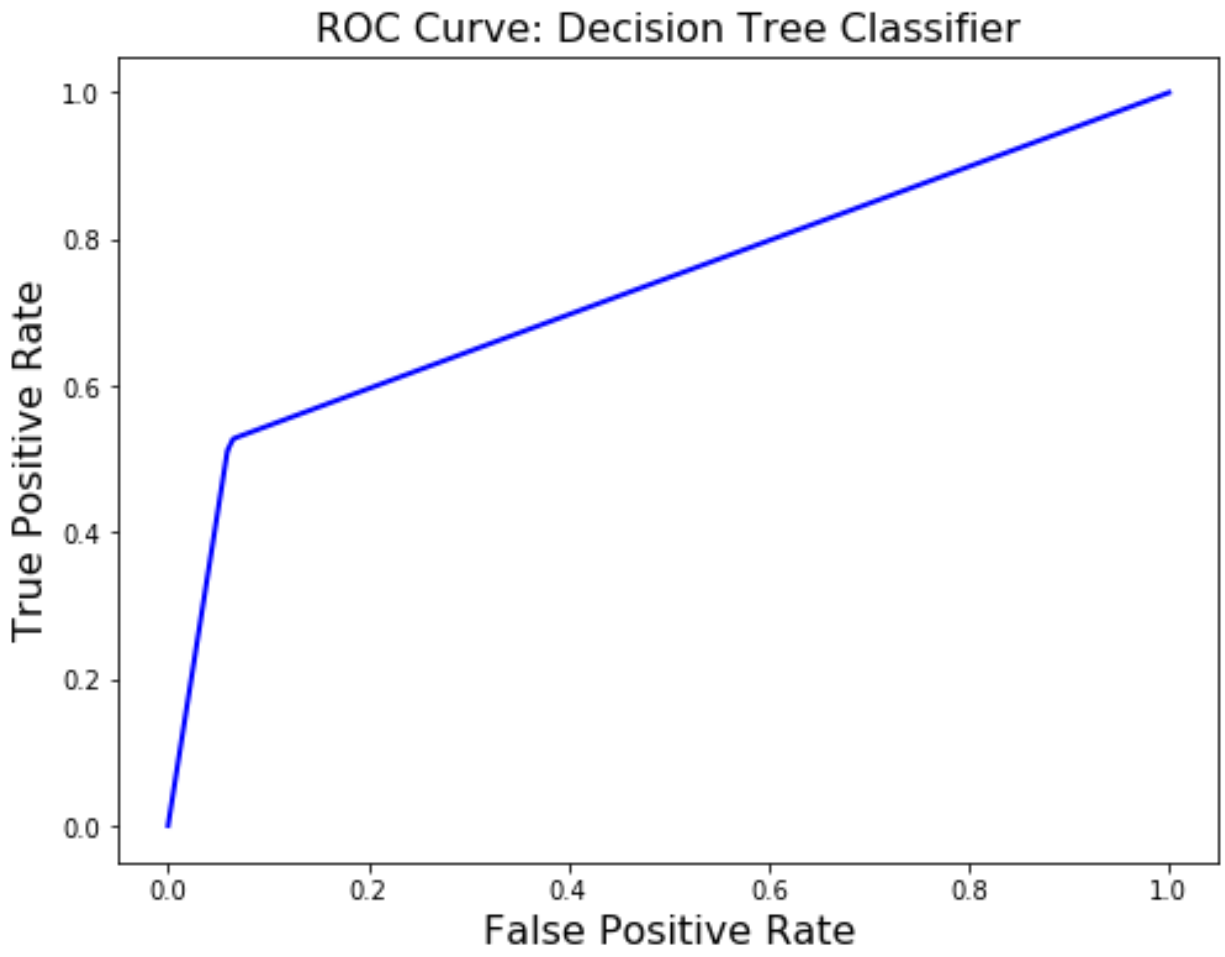
AUC Score (Logistic Regression): 0.92



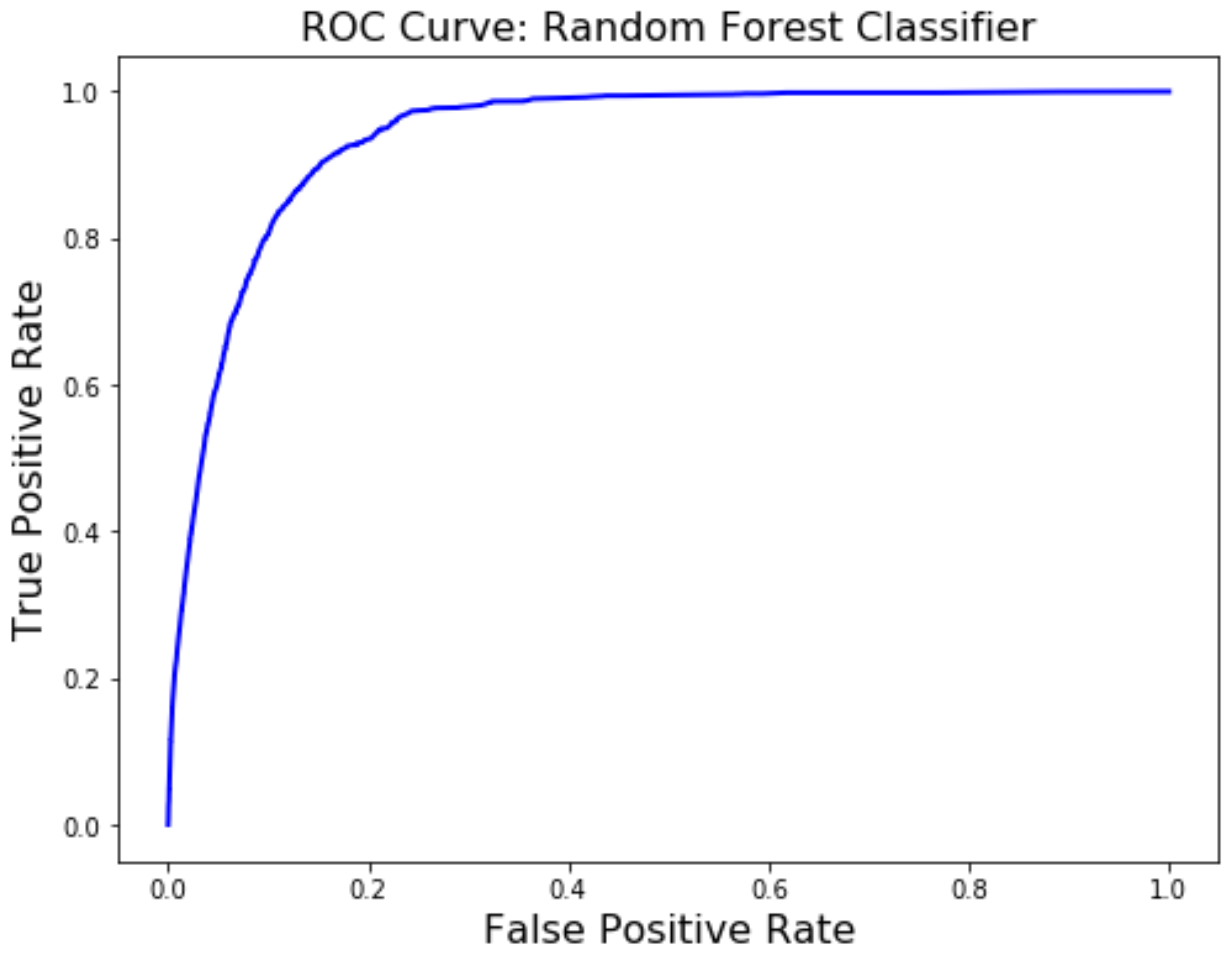
AUC Score (Support Vector Classification): 0.93



AUC Score (Decision Tree Classifier): 0.73



AUC Score (Random Forest Classifier): 0.94



# 5. Findings and Suggestions

Through Exploratory Data Analysis, five classification models each were built for data split in 80-20 and 90-10 partition.

**80-20 split**

- The Accuracy for Support vector classification is 89.65%. It had a recall score of 28.12 and Roc score 62.68. It's confusion matrix is as follows [[7132 203]
  [ 649 254]]
- The Accuracy for Random Forest classification is 91.6%. It had a recall score of 51.71 and Roc score 74.13. It's confusion matrix is as follows [[7082 253]
  [ 436 467]]
- The Accuracy for Decision Tree classification is 89.37%. It had a recall score of 49.39 and Roc score 71.84. It's confusion matrix is as follows [[6917 418]
  [ 457 446]]
- The Accuracy for Logistic Regression is 91.00%. It had a recall score of 38.98 and Roc score 68.19. It's confusion matrix is as follows [[7145 190]
  [ 551 352]]

- The Accuracy for kNN classification is 90.2%. It had a recall score of 46.5 and Roc score 71.09. It's confusion matrix is as follows `[[7018 317]`
  `[ 483 420]]`

**90-10 split**

- The Accuracy for Support vector classification is 89.65%. It had a recall score of 28.94 and Roc score 63.06. It's confusion matrix is as follows `[[3560 103]`
  `[ 324 132]]`
- The Accuracy for Random Forest classification is 91.6%. It had a recall score of 90.94 and Roc score 68.79. It's confusion matrix is as follows `[[3562 101]`
  `[ 272 184]]`
- The Accuracy for Decision Tree classification is 89.14%. It had a recall score of 40.35 and Roc score 71.33. It's confusion matrix is as follows `[[3451 212]`
  `[ 235 221]]`
- The Accuracy for Logistic Regression is 91.04%. It had a recall score of 37.28 and Roc score 67.50. It's confusion matrix is as follows `[[3580 83]`
  `[ 286 170]]`

- The Accuracy for kNN classification is 90.26%. It had a recall score of 45.39 and Roc score 70.62. It's confusion matrix is as follows `[[3511 152]`
  `[ 249 207]]`

# 6. Conclusion

Out of all the classification models built using different data splits, the classification model built using 80-20 data split with Random Forest classification showed comparatively more accuracy, recall score, ROC score and less number of false negatives in the confusion matrix.

From the above parameters it can be concluded that Random forest classification model with 80-20 data split is a better model to analyze the given data set.

# 7. Bibliography and Reference


1. https://data.world/data-society/bank-marketing-data

2. www.quora.com

3.https://www.slideshare.net/pratikaloni/marketing-in-banking-sector

4.https://thefinancialbrand.com/57325/ten-financial-marketing-priorities-trends/