# Logistic Regression



**Instructions:**

Please share your answers filled in-line in the word document. Submit code separately wherever applicable.

Please ensure you update all the details:

**Name: DHEERAJ MISHRA Batch ID:**  DS\_01072021

**Topic: Logistic Regression**

**Grading Guidelines:**

**1. An assignment submission is considered complete only when correct and executable code(s) are submitted along with the documentation explaining the method and results. Failing to submit either of those will be considered an invalid submission and will not be considered for evaluation.**

**2. Assignments submitted after the deadline will affect your grades.**

**Grading:**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Ans** | **Date** |  |  | **Ans** | **Date** |
| Correct | On time | A | 100 |  |  |
| 80% & above | On time | B | 85 | Correct | Late |
| 50% & above | On time | C | 75 | 80% & above | Late |
| 50% & below | On time | D | 65 | 50% & above | Late |
|  |  | E | 55 | 50% & below |  |
| Copied/No Submission |  | F | 45 |  |  |

* **Grade A: (>= 90):** When all assignments are submitted on or before the given deadline.
* **Grade B: (>= 80 and < 90):** 
  + When assignments are submitted on time but less than 80% of problems are completed.

(OR)

* + All assignments are submitted after the deadline.
* **Grade C: (>= 70 and < 80):** 
  + When assignments are submitted on time but less than 50% of the problems are completed.

(OR)

* + Less than 80% of problems in the assignments are submitted after the deadline.
* **Grade D: (>= 60 and < 70):**
  + Assignments submitted after the deadline and with 50% or less problems.
* **Grade E: (>= 50 and < 60):** 
  + Less than 30% of problems in the assignments are submitted after the deadline.

(OR)

* + Less than 30% of problems in the assignments are submitted before the deadline.
* **Grade F: (< 50):** No submission (or) malpractice.

**Hints:**

1. **Business Problem**
   1. **What is the business objective?**
   2. **Are there any constraints?**
2. **Work on each feature of the dataset to create a data dictionary as displayed in the below image:**



**2.1 Make a table as shown above and provide information about the features such as its data type and its relevance to the model building. And if not relevant, provide reasons and a description of the feature.**

**Using R and Python codes perform:**

1. **Data Pre-processing**

**3.1 Data Cleaning, Feature Engineering, etc.**

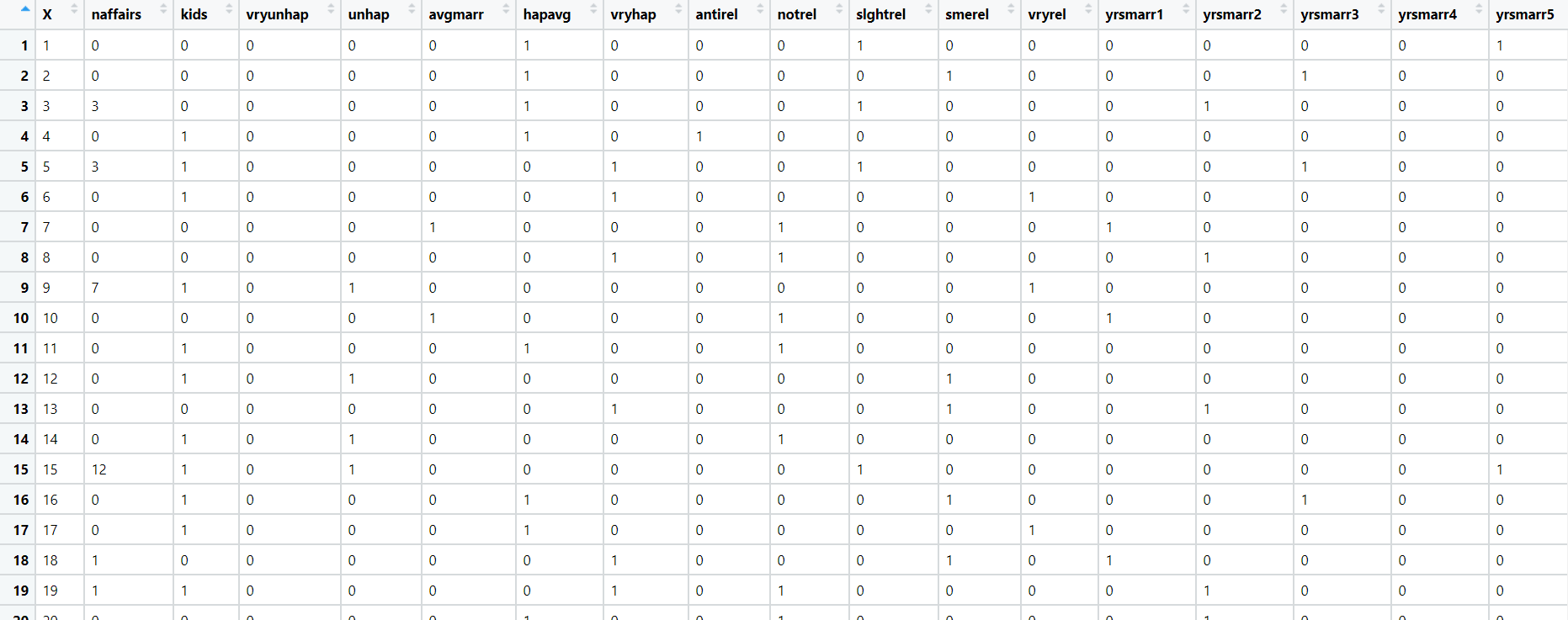
**3.2 Outlier Treatment.**

1. **Exploratory Data Analysis (EDA):**
   1. **Summary.**
   2. **Univariate analysis.**
   3. **Bivariate analysis.**
2. **Model Building**
   1. **Build the model on the scaled data (try multiple options).**
   2. **Build a Logistic Regression model.**
   3. **Train and test the model and compare accuracies by building a confusion matrix, plotting ROC and AUC curves.**
   4. **Briefly explain the model output in the documentation.**
3. **Write about the benefits/impact of the solution - in what way does the business (client) benefit from the solution provided?**

A screenshot of a cell phone

Description automatically generatedProblem Statement: -

1. A psychological study has been conducted by a team of students at a university on married couples to determine the cause of having an extra marital affair. They have surveyed and collected a sample of data on which they would like to do further analysis. Apply Logistic Regression on the data to correctly classify whether a given person will have an affair or not given the set of attributes. Convert the naffairs column to discrete binary type before proceeding with the algorithm.



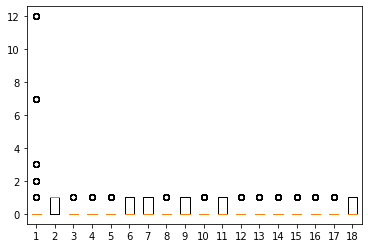
1. BUSINESS OBJECTIVE:-

Maximize relationship between multivariable in binary form

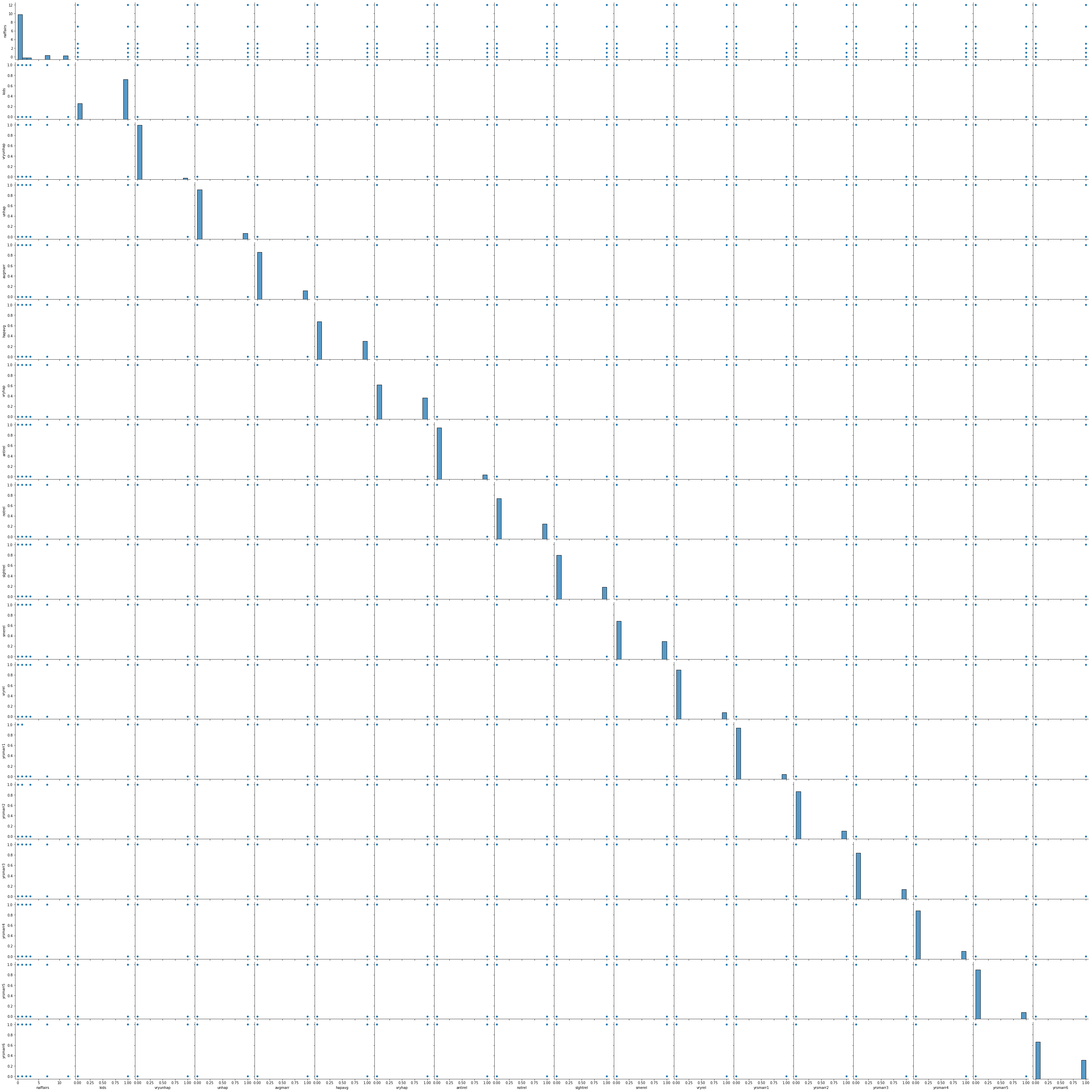
1. DATA UNDERSTANDING:-

|  |  |  |  |
| --- | --- | --- | --- |
| NAME OF FEATURE | DESCRIPTION | TYPE | RELEVANCE |
| Id | Serial no | Discrete | Not Relevant |
| naffairs | No of affairs | Discrete | Relevant |
| kids | Kids or not | Binary | Relevant |
| vryunhap | Very unhappy | Binary | Relevant |
| unhap | Total profit | Binary | Relevant |
| avgmarr | Average marriage | Binary | Relevant |
| hapavg | Happy marriage | Binary | Relevant |
| vryhap | Very happy | Binary | Relevant |
| antirel | Anti religious | Binary | Relevant |
| notrel | Not religious | Binary | Relevant |
| slghtrel | Slightly religious | Binary | Relevant |
| smerel | Somewhat religious | Binary | Relevant |
| vryrel | Very religious | Binary | Relevant |
| yrsmarr1 | Years of maariage 1 | Binary | Relevant |
| yrsmarr2 | Years of maariage 2 | Binary | Relevant |
| Yrsmarr3 | Years of maariage 3 | Binary | Relevant |
| Yrsmarr4 | Years of maariage 4 | Binary | Relevant |
| Yrsmarr5 | Years of maariage 5 | Binary | Relevant |
| Yrsmarr6 | Years of maariage 6 | Binary | Relevant |

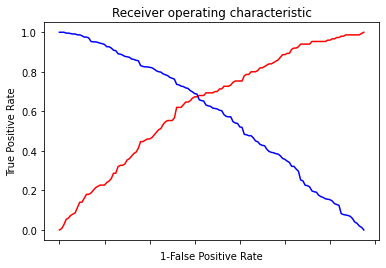
1. DATA CLEANSING :-
2. Dataset consists of 19 colums and 601 rows
3. Duplicate row exists and retained
4. All data types are of form int64
5. No null values found in each column
6. From describe function mean , median and standard deviation obtained
7. Outliers present and retained
8. Corelation coefficients calculated
9. There is collinearity problem
10. Some are positively skewed and some are negatively skewed
11. Columns name changed
12. Output variable converted to binary form
13. EDA:-
14. From box plot



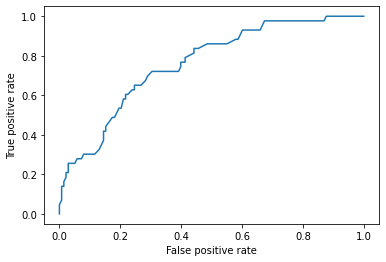
1. From pair plot



1. MODEL BUILDING:-
2. Model builded
3. Threshold value = 0.2521571570135329
4. Plot tpr vs 1-fpr



1. Area under the curve = 0.720880
2. Splitting data to training = 70% and testing = 30 %
3. Test accuracy = 0.64
4. AUC curve



1. Roc\_auc\_test = 0.7607853050219077
2. Train accuracy = 0.70

OUTPUT:-

It is right fit model

1. BENEFITS :-

From above information we can predict for affairs against all input variables

1. In this time and age of widespread internet usage, effective and targeted marketing plays a vital role. A marketing company would like to develop a strategy by analyzing their customer data. For this, data like age, location, time of activity, etc. has been collected to determine whether a user will click on an ad or not. Perform Logistic Regression on the given data to predict whether a user will click on an ad or not.

A screenshot of a cell phone

Description automatically generated

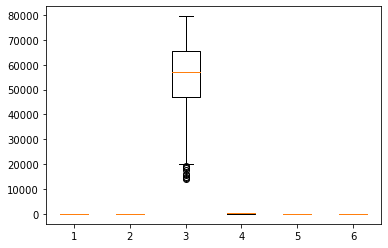
1. BUSINESS OBJECTIVE:-

Maximize relationship between multivariable in binary form

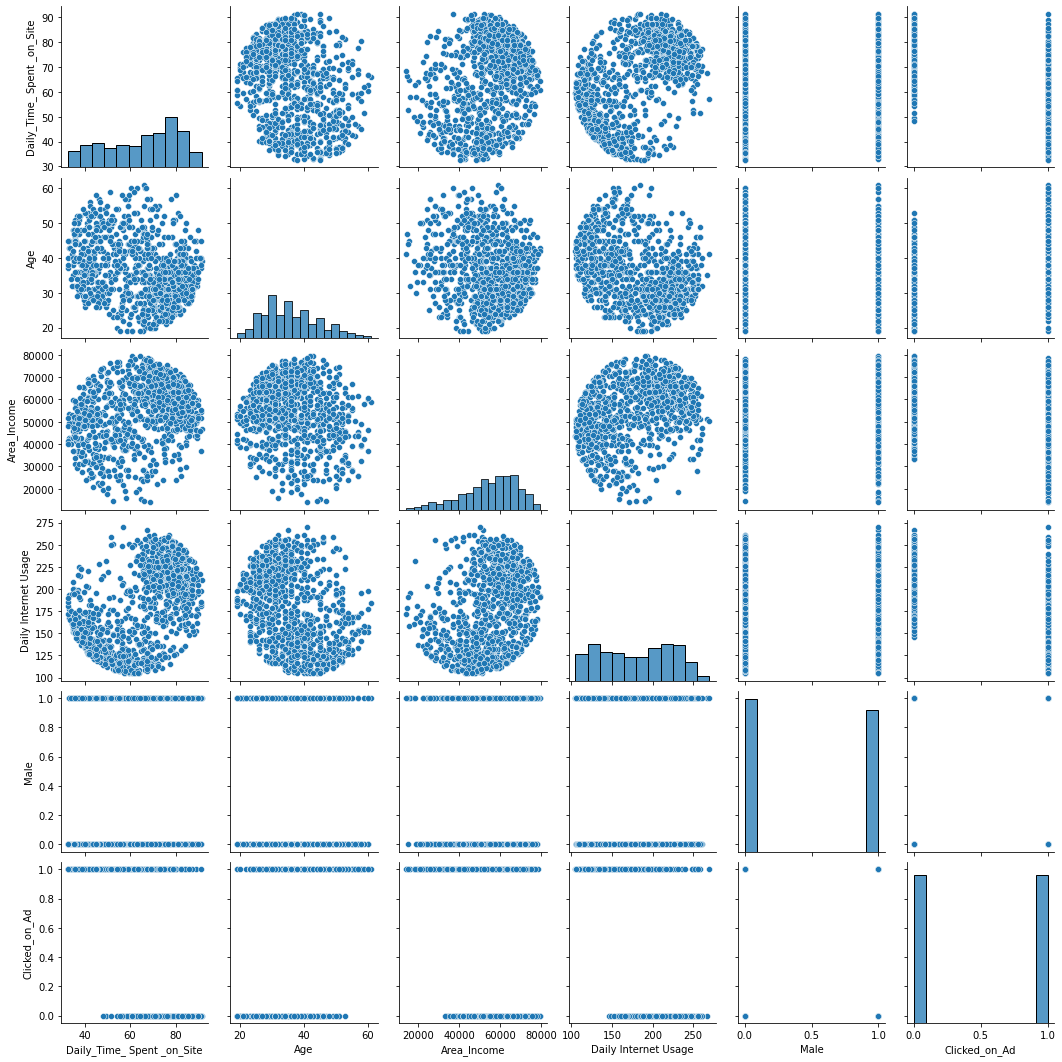
1. DATA UNDERSTANDING:-

|  |  |  |  |
| --- | --- | --- | --- |
| NAME OF FEATURE | DESCRIPTION | TYPE | RELEVANCE |
| Daily\_Time\_ Spent \_on\_Site | Daily time spent on site | Continuous | Relevant |
| Age | Age of customer | Discrete | Relevant |
| Area\_Income | Income of area | Continuous | Relevant |
| Daily Internet Usage | Daily internet usage | Continuous | Relevant |
| Ad\_Topic\_Line | Advertisement | Text | Not Relevant |
| City | City | Text | Not elevant |
| Male | Gender | Binary | Relevant |
| Country | Country name | Text | Not Relevant |
| Timestamp | Day and date | Time series | Not Relevant |
| Clicked\_on\_Ad | Result | Binary | Binary |

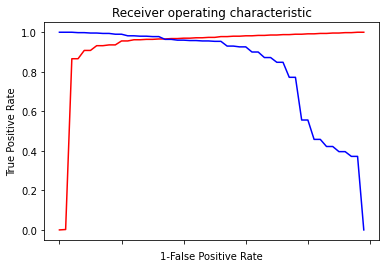
1. DATA CLEANSING :-
2. Dataset consists of 10 colums and 1000 rows
3. Duplicate row does not exists
4. All data types are of form int64 , float64 and object
5. No null values found in each column
6. From describe function mean , median and standard deviation obtained
7. Outliers present and retained
8. Corelation coefficients calculated
9. There is collinearity problem
10. Some are positively skewed and some are negatively skewed
11. Columns name changed
12. Normalization for scaling
13. Output variable converted to binary form
14. EDA:-
15. From box plot



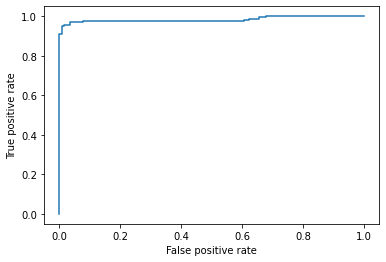
1. From pair plot



1. MODEL BUILDING:-
2. Model builded
3. Threshold value = 0.6326111998951834
4. Plot tpr vs 1-fpr



1. Area under the curve = 0.991800
2. Splitting data to training = 70% and testing = 30 %
3. Test accuracy = 0.9766666666666667
4. AUC curve



1. Roc\_auc\_test = 0.9826785714285714
2. Train accuracy = 0.97

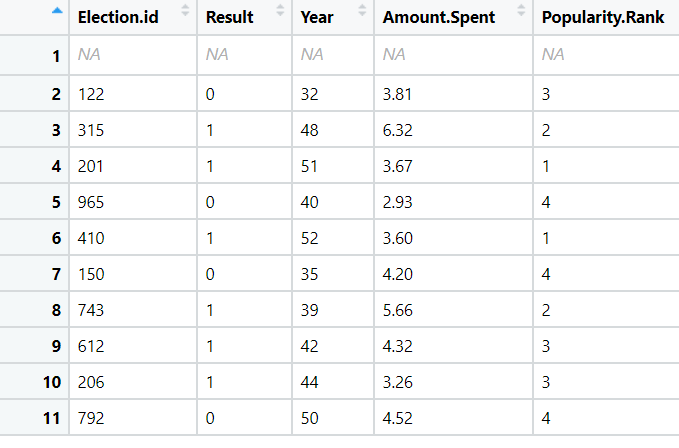
OUTPUT:-

It is right fit model

1. BENEFITS :-

From above information we can predict for clicked on advertisement against all input variables

1. Perform Logistic Regression on the dataset to predict whether a candidate will win or lose the election based on factors like amount of money spent and popularity rank.



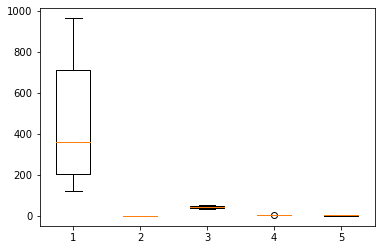
1. BUSINESS OBJECTIVE:-

Maximize relationship between multivariable in binary form

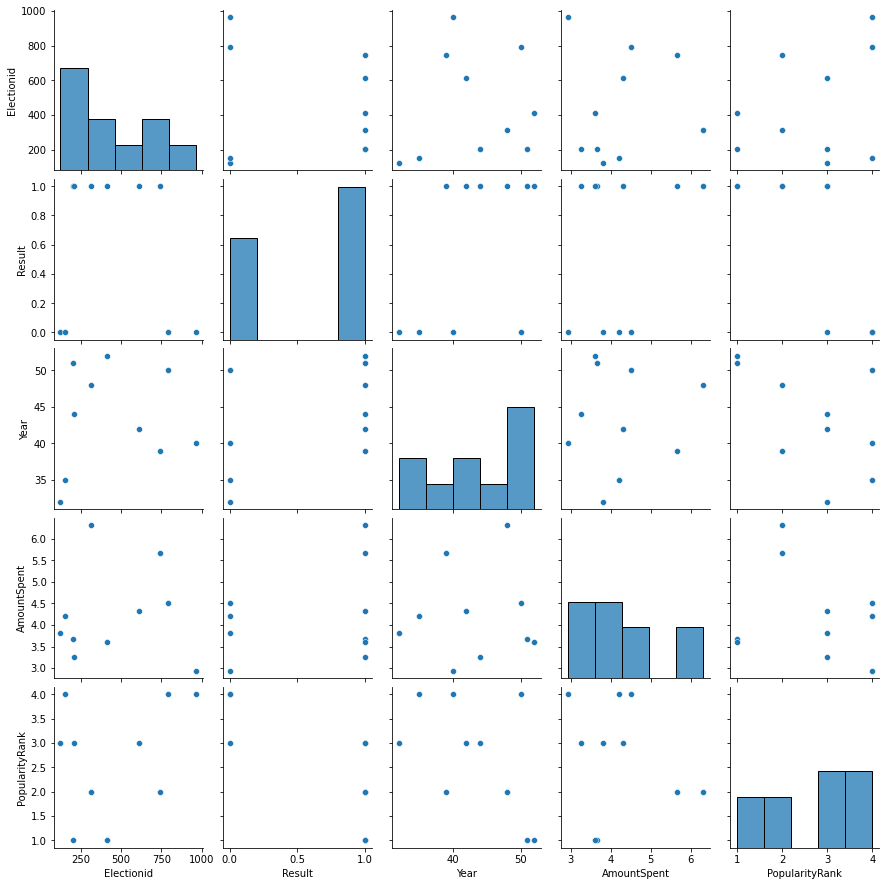
1. DATA UNDERSTANDING:-

|  |  |  |  |
| --- | --- | --- | --- |
| NAME OF FEATURE | DESCRIPTION | TYPE | RELEVANCE |
| Electionid | Id for election | Discrete | Relevant |
| Result | Result of election | Binary | Relevant |
| Year | Years for election | Discrete | Relevant |
| AmountSpent | Expenditure | Continuous | Relevant |
| PopularityRank | Popularity | Discrete | Relevant |

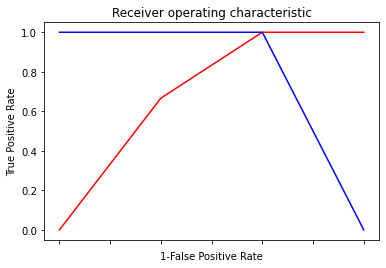
1. DATA CLEANSING :-
2. Dataset consists of 5 colums and 11 rows
3. Duplicate row does not exists
4. All data types are of form int64 , float64 and object
5. No null values found in each column
6. From describe function mean , median and standard deviation obtained
7. Outliers present and retained
8. Corelation coefficients calculated
9. There is collinearity problem
10. Some are positively skewed and some are negatively skewed
11. Columns name changed
12. Normalization for scaling
13. EDA:-
14. From box plot



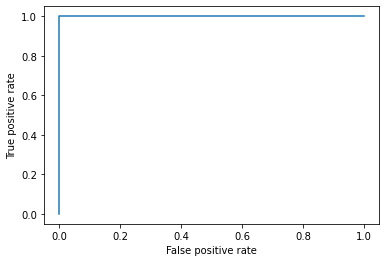
1. From pair plot



1. MODEL BUILDING:-
2. Model builded
3. Threshold value = 0.9999999999729929
4. Plot tpr vs 1-fpr



1. Area under the curve = 1
2. Splitting data to training = 80% and testing = 20 %
3. Test accuracy = 1
4. AUC curve



1. Roc\_auc\_test = 1
2. Train accuracy = 0.875

OUTPUT:-

It is right fit model

1. BENEFITS :-

From above information we can predict for result against all input variables

1. It is vital for banks that customers put in long term fixed deposits as they use it to pay interest to customers and it is not viable to ask every customer if they will put in a long-term deposit or not. So, build a Logistic Regression model to predict whether a customer will put in a long-term fixed deposit or not based on the different variables given in the data. The output variable in the dataset is Y which is binary. Snapshot of the dataset is given below.



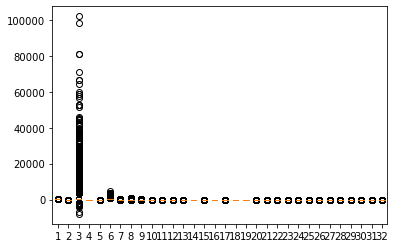
1. BUSINESS OBJECTIVE:-

Maximize relationship between multivariable in binary form

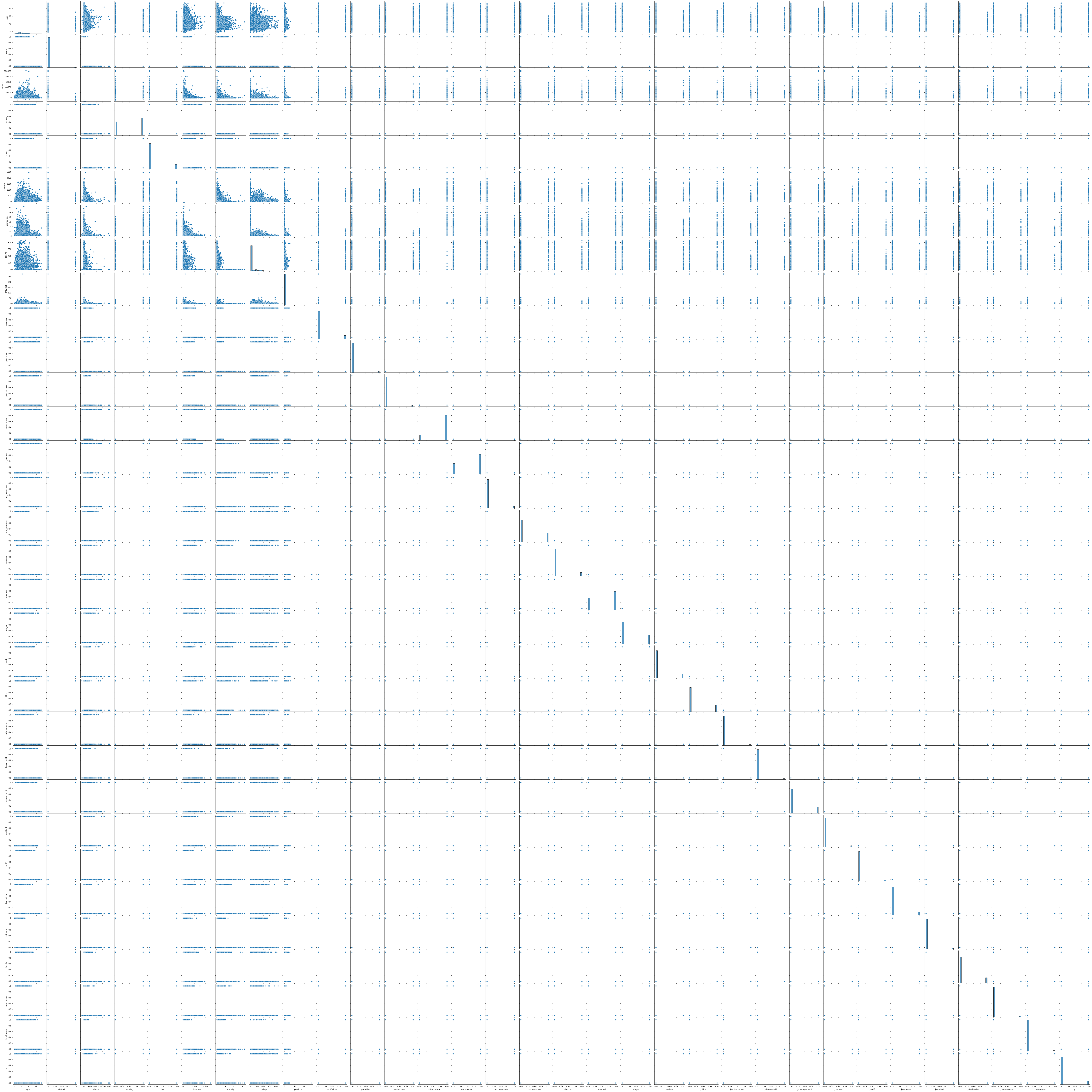
1. DATA UNDERSTANDING:-

|  |  |  |  |
| --- | --- | --- | --- |
| NAME OF FEATURE | DESCRIPTION | TYPE | RELEVANCE |
| age | Age of customer | Discrete | Relevant |
| default | Default or not | Binary | Relevant |
| balance | Account balance | Discrete | Relevant |
| housing | House or not | Binary | Relevant |
| loan | Loan or not | Binary | Relevant |
| duration | Time | Discrete | Relevant |
| campaign | Campaign | Discrete | Relevant |
| pdays | P days | Discrete | Relevant |
| previous | Previous or not | Discrete | Relevant |
| poutfailure | Failure | Binary | Relevant |
| poutother | Ohers | Binary | Relevant |
| poutsuccess | Success | Binary | Relevant |
| poutunknown | Unknown | Binary | Relevant |
| con\_cellular | Cellular | Binary | Relevant |
| con\_telephone | Telephone | Binary | Relevant |
| con\_unknown | Unknown | Binary | Relevant |
| divorced | Divorced or not | Binary | Relevant |
| married | Married or not | Binary | Relevant |
| single | Single or not | Binary | Relevant |
| joadmin | Admin job | Binary | Relevant |
| joblue | Blue job | Binary | Relevant |
| joentrepreneur | Entrepreneurs job | Binary | Relevant |
| johousemaid | Housemaid job | Binary | Relevant |
| jomanagement | Management job | Binary | Relevant |
| joretired | Retired | Binary | Relevant |
| joself | Self employed | Binary | Relevant |
| joservices | Services | Binary | Relevant |
| jostudent | Student | Binary | Relevant |
| jotechnician | Technician | Binary | Relevant |
| jounemployed | Unemployed | Binary | Relevant |
| jounknown | Job unknown | Binary | Relevant |
| y | Result | Binary | Relevant |

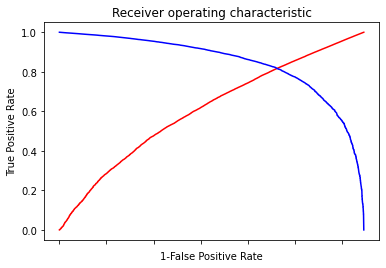
1. DATA CLEANSING :-
2. Dataset consists of 32 colums and 45211 rows
3. Duplicate row exists and retained
4. All data types are of form int64
5. No null values found in each column
6. From describe function mean , median and standard deviation obtained
7. Outliers present and retained
8. Corelation coefficients calculated
9. There is collinearity problem
10. Some are positively skewed and some are negatively skewed
11. Normalization for scaling
12. EDA:-
13. From box plot



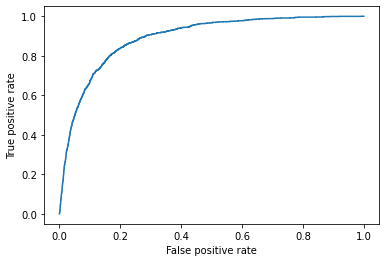
1. From pair plot



1. MODEL BUILDING:-
2. Model builded
3. Threshold value = 0.11474220787661779
4. Plot tpr vs 1-fpr



1. Area under the curve = 0.890843
2. Splitting data to training = 70% and testing = 30 %
3. Test accuracy = 0.8232822176349159
4. AUC curve



1. Roc\_auc\_test = 0.8913588988624107
2. Train accuracy = 0.8194457610516005

OUTPUT:-

It is right fit model

1. BENEFITS :-

From above information we can predict for result against all input variables