LEAD SCORING CASE STUDY

SUBMITTED BY:

ASHOK SINGH

ASHI JAISWAL

ARYA MADHU

PROBLEM STATEMENT

- An education company named X Education sells online courses. On any given day, many professionals who are interested in the courses land on their website and browse for courses.
- Once these people land on the website, they might browse the courses or fill up a form for the course or watch some videos. When these people fill up a form providing their email address or phone number, they are classified to be a lead, Employees from sales team start making calls, writing emails, etc. Through this process, some of the leads get converted while most do not.
- Although X Education gets a lot of leads, its lead conversion rate is very poor. The lead conversion rate is around 30%.

OBJECTIVE

- X Education wants the leads that are most likely to convert into paying customers. Build LOGISTIC REGRESSION model wherein need to assign a lead score to each of the leads such that the customers with higher lead score have higher conversion chance and vice versa.
- The CEO has given the target that lead conversion rate to be around 80%.

APPROACH

1. DATA IMPORTING & UNDERSTANDING
2. DATA PREPARATION
3. EDA
4. TRAIN-TEST SPLIT
5. MODEL BUILDING
6. FEATURE SELECTION USING RFE
7. MODEL EVALUATION
8. MAKING PREDICTION ON TEST SET
9. EVALUATION OF TEST SET
10. CONCLUSION

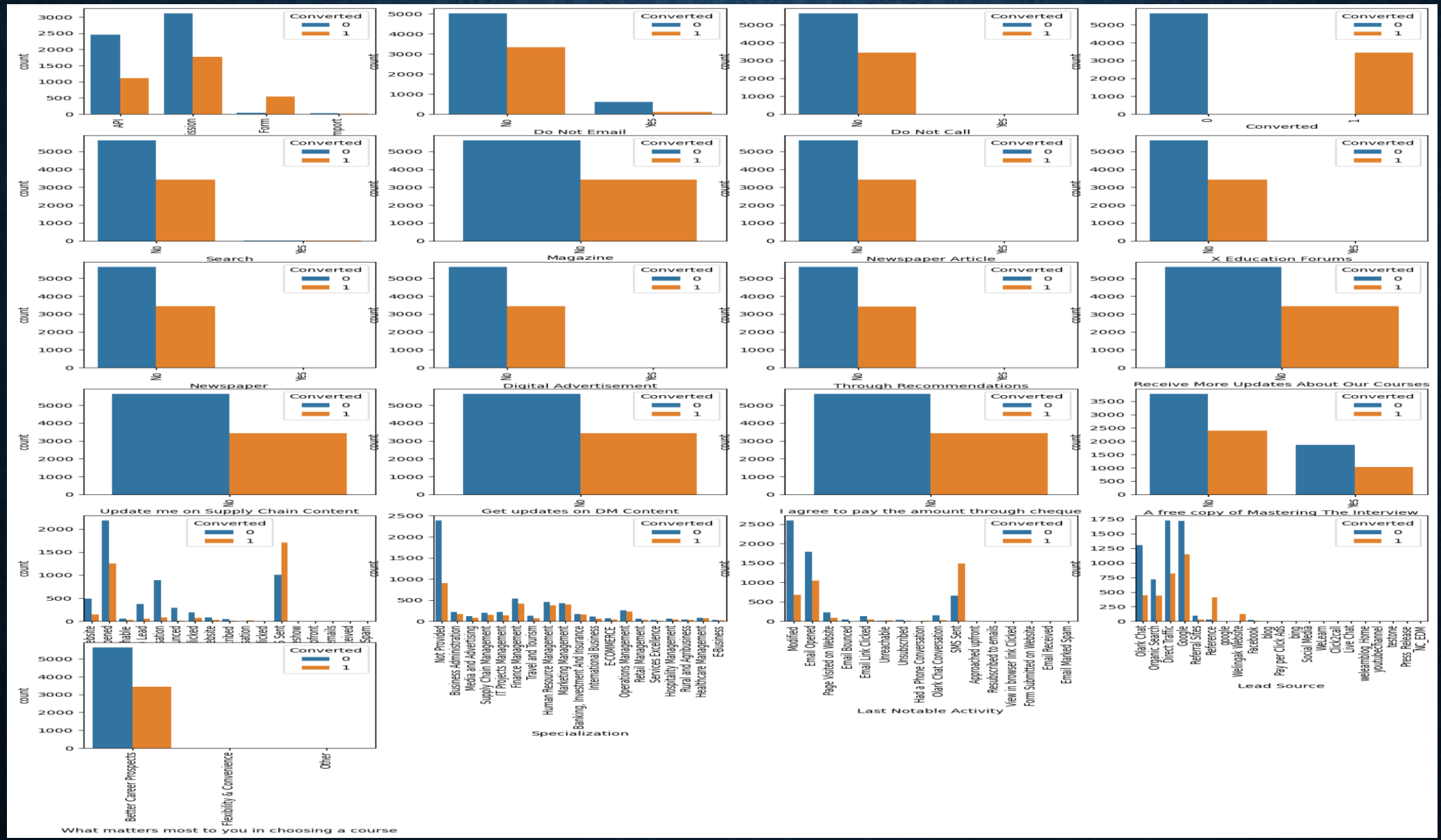
1. DATA IMPORTING & UNDERSTANDING

- Shape of the data :9240 rows ,37 columns
- There are missing values in most of the columns,
- We noticed that some columns contains large amount of ‘SELECT’ value which needs to be handled because it is as important as null value.
- There are some unnecessary columns which will not add any values to the model because some of them has single value and some has values like Id and s.no
- There is presence of outliers in some numerical variables

2. DATA PREPARATION

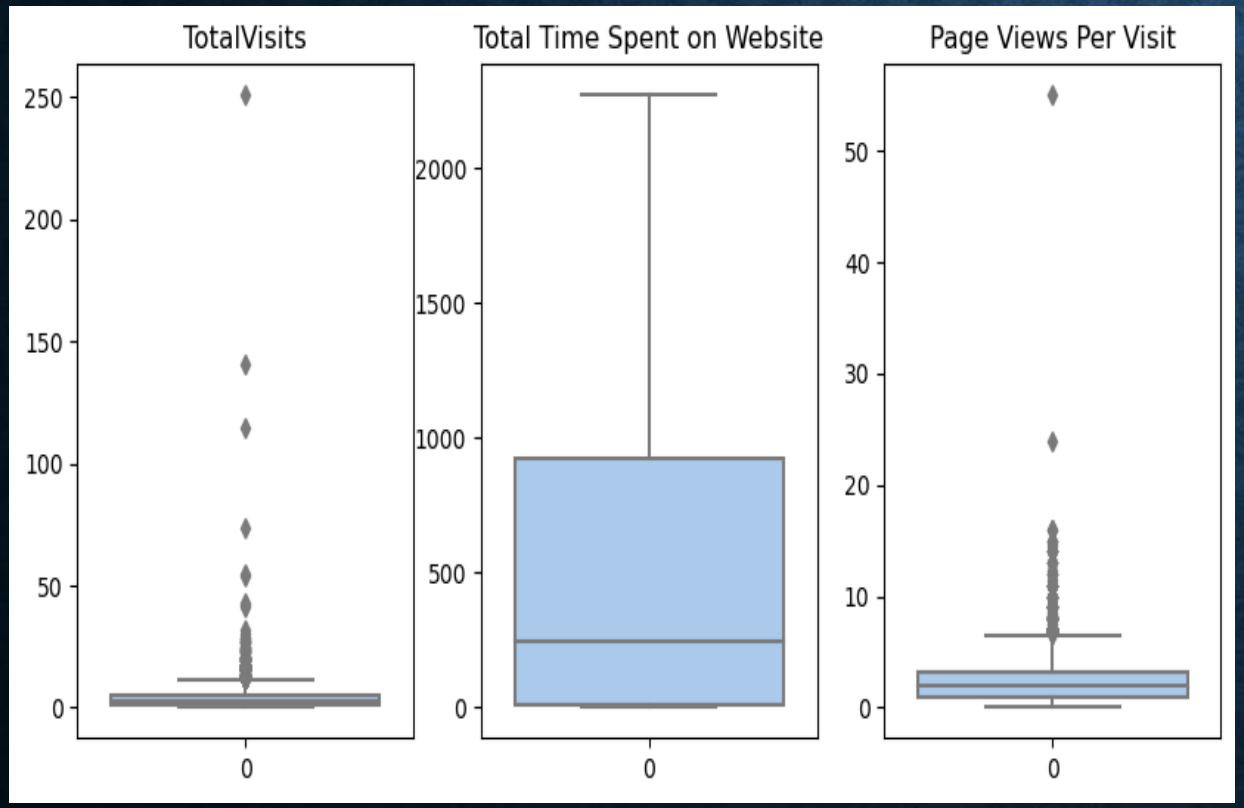
- . Converting the ‘SELECT’ values with ‘NaN’. It will be the best option as cannot assign any value.
- Removing the columns having missing values greater than 45% and we will modify the missing values less than 45%.
- Dropping ‘Country’, ’City ‘, ‘Tags’ columns as they have some category which has very high no of values. So , these will not add any value.
- Replacing missing of ('Specialization’, 'What is your current occupation’) and ('What matters most to you in choosing a course') with ‘Not Provided’ and mode() of the column respectively.
- Dropping the column having missing value in range of 1-3% as very small no. and dropping will be the best option.

- Visualising some unnecessary columns
- Some columns has single value and will be of no use and some columns has very less use in model

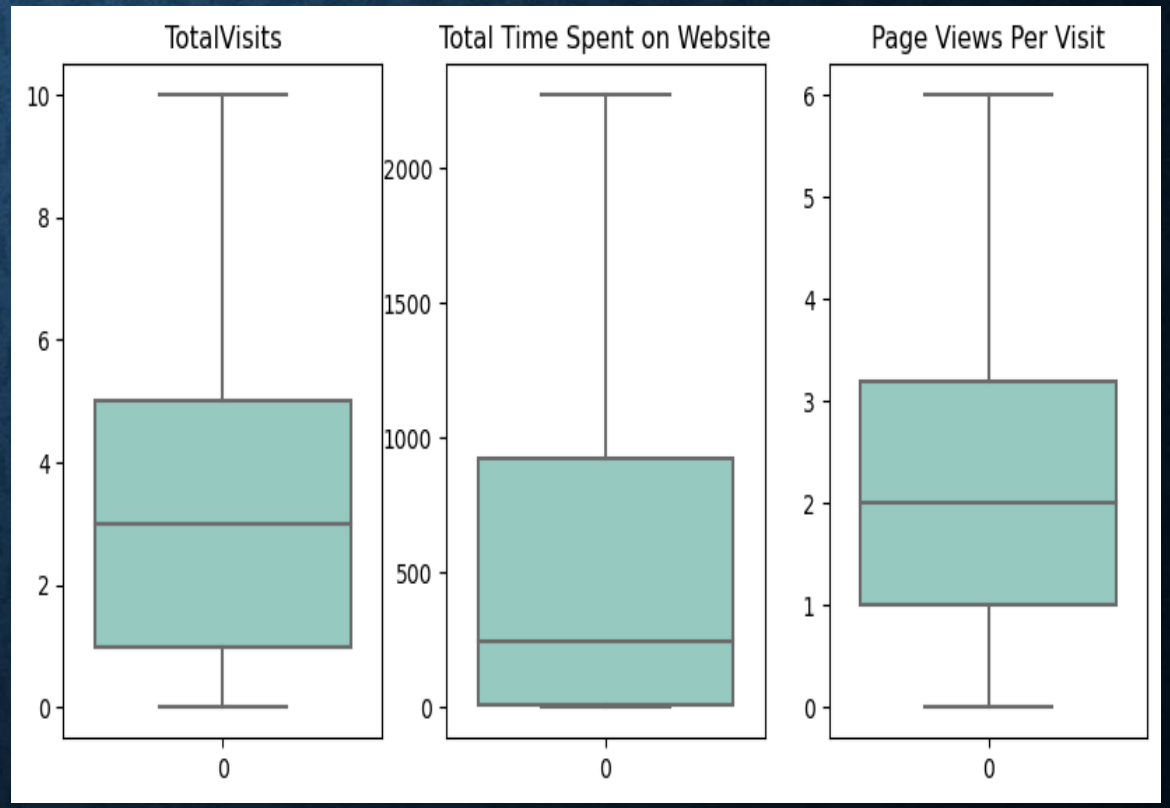


❑ CHECKING OUTLIERS

- Checking the Outliers on numerical variables.
- ‘Total Time Spent on Website’ has some outliers but they are under acceptable range.
- Capping the outliers within range 5% -95%
- Capping outliers from ‘TotalVisits’, ‘Page Views Per Visit’.



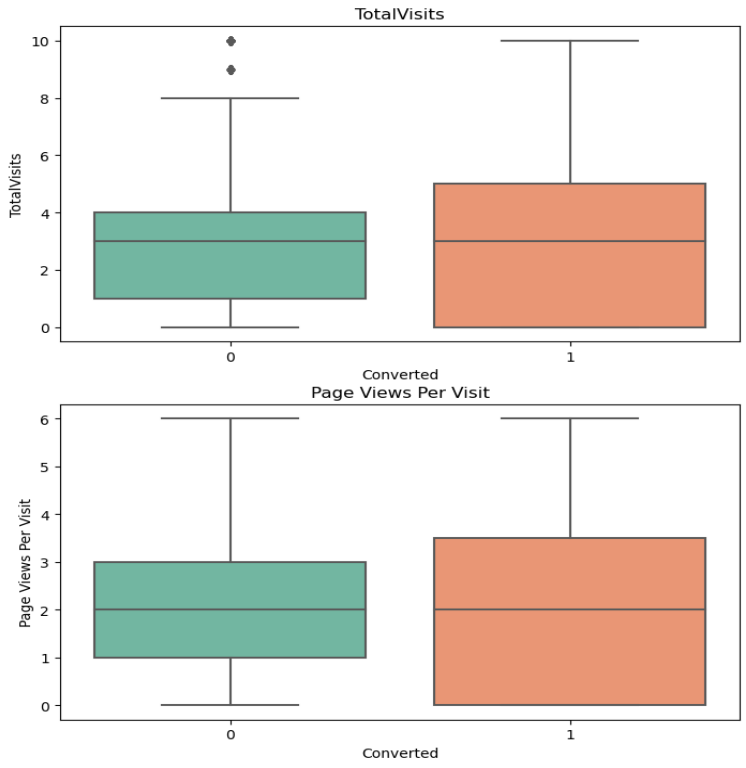
Before Capping

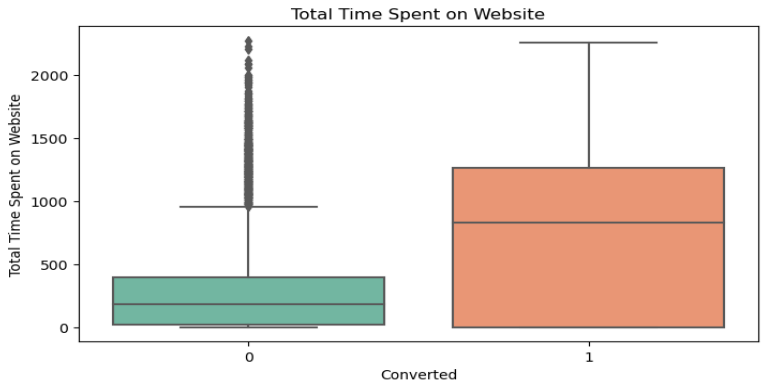


After Capping

3.EDA

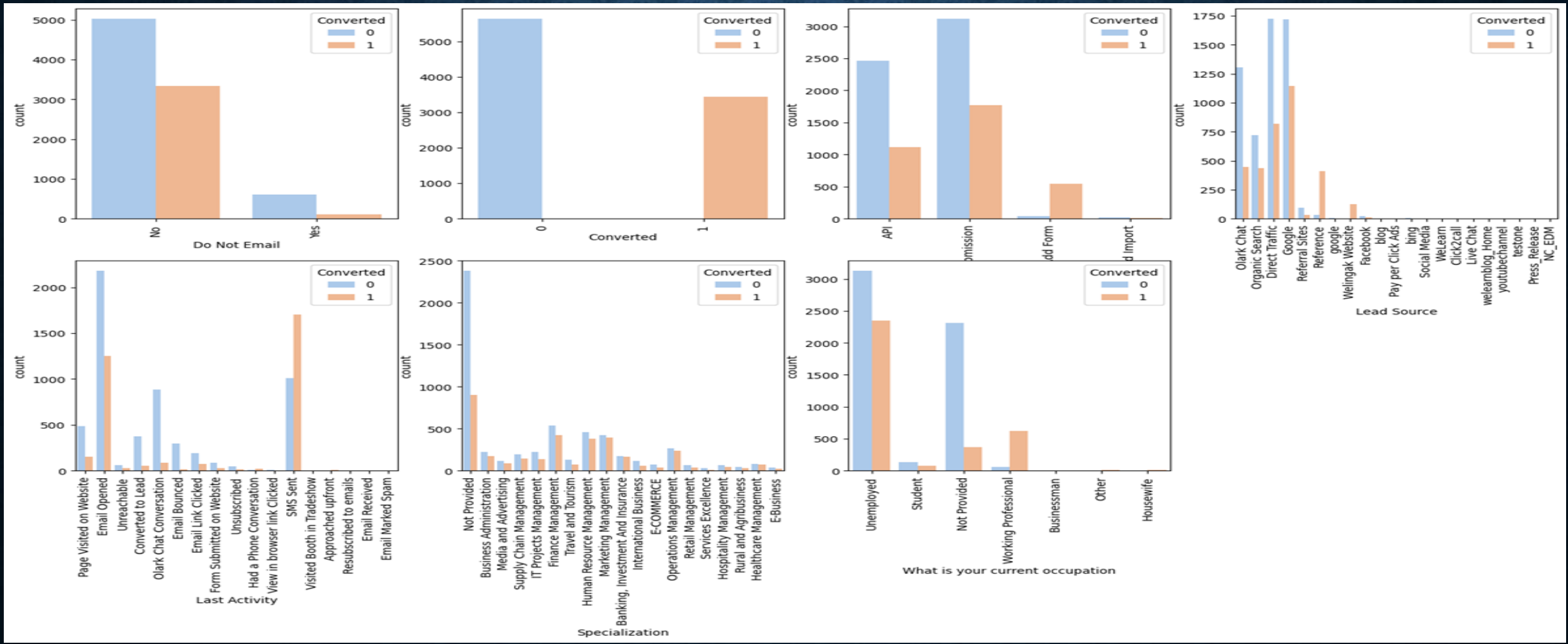
- NUMERICAL DATA:
- Checking the correlation between the numerical variables.
- Checking the correlation of numerical data with converted variables.
- Leads who spends more time on Website has highest no of converted Leads and vice versa.
- Leads visit the page also has good amount of lead conversion.





❑ CATEGORICAL DATA :

- Lead origin tell, Lead landing on the Admission page has highest conversion rate.
- Lead using the Google Lead Source has the highest conversion rate.
- Apart from the Unemployed occupation Working professional lead has highest rate of conversion
- Last activity of SMS Sent, Email Opened and Specialization in Marketing and Finance Management has highest conversion rate.

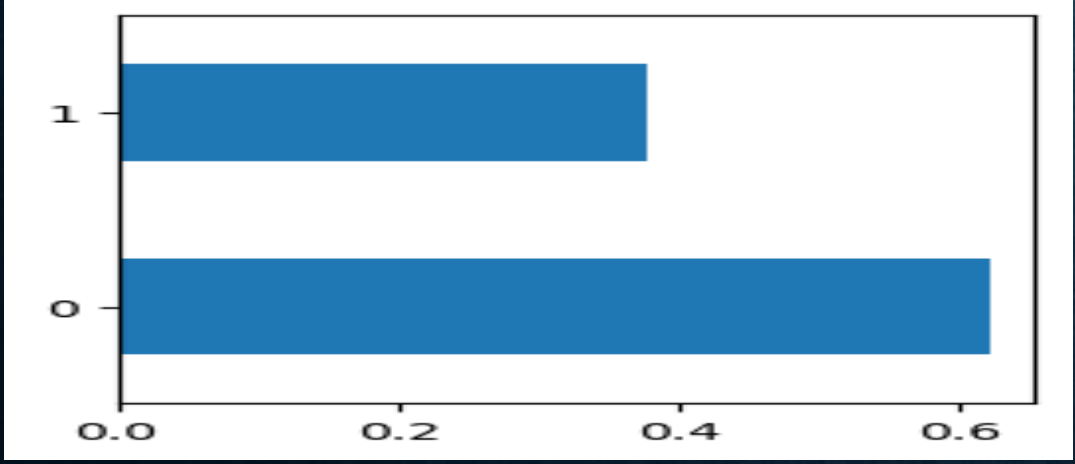


❑ MODIFYING & MAKING DUMMY VARIABLES

- Modify the column 'Lead Source’ and ‘Last Activity’ for dummy variable, we can merge the values which are less in no. ,has low frequency or has similar values meanings.
- Replacing(‘bing’, ‘ Click2call ’ , ‘Press_Release’ , ‘Live Chat’, ‘youtubechannel’, ‘ testone’, ‘ Pay per Click Ads’, 'welearnblog_Home’ , ‘WeLearn’, ‘blog’, ‘NC_EDM’) with ‘Other Platform’.
- Replacing (‘ Unreachable’ , ‘Unsubscribed’ , ‘Had a Phone Conversation’, ‘ View in browser link Clicked’ , ‘ Approached upfront','Email Received’ , ‘ Email Marked Spam’, ‘ Visited Booth in Tradeshow','Resubscribed to emails’) with ‘ Other Activity’.
- Converting ‘Yes’/ ‘No’ to 1/0 of ‘ Do Not Email’ column.
- Making dummy variables of columns (‘Lead Origin’ , ‘ Lead Source’, ‘Last Activity’ , ‘Specialization’ , ‘What is your current occupation’)

4. TRAIN-TEST SPLIT

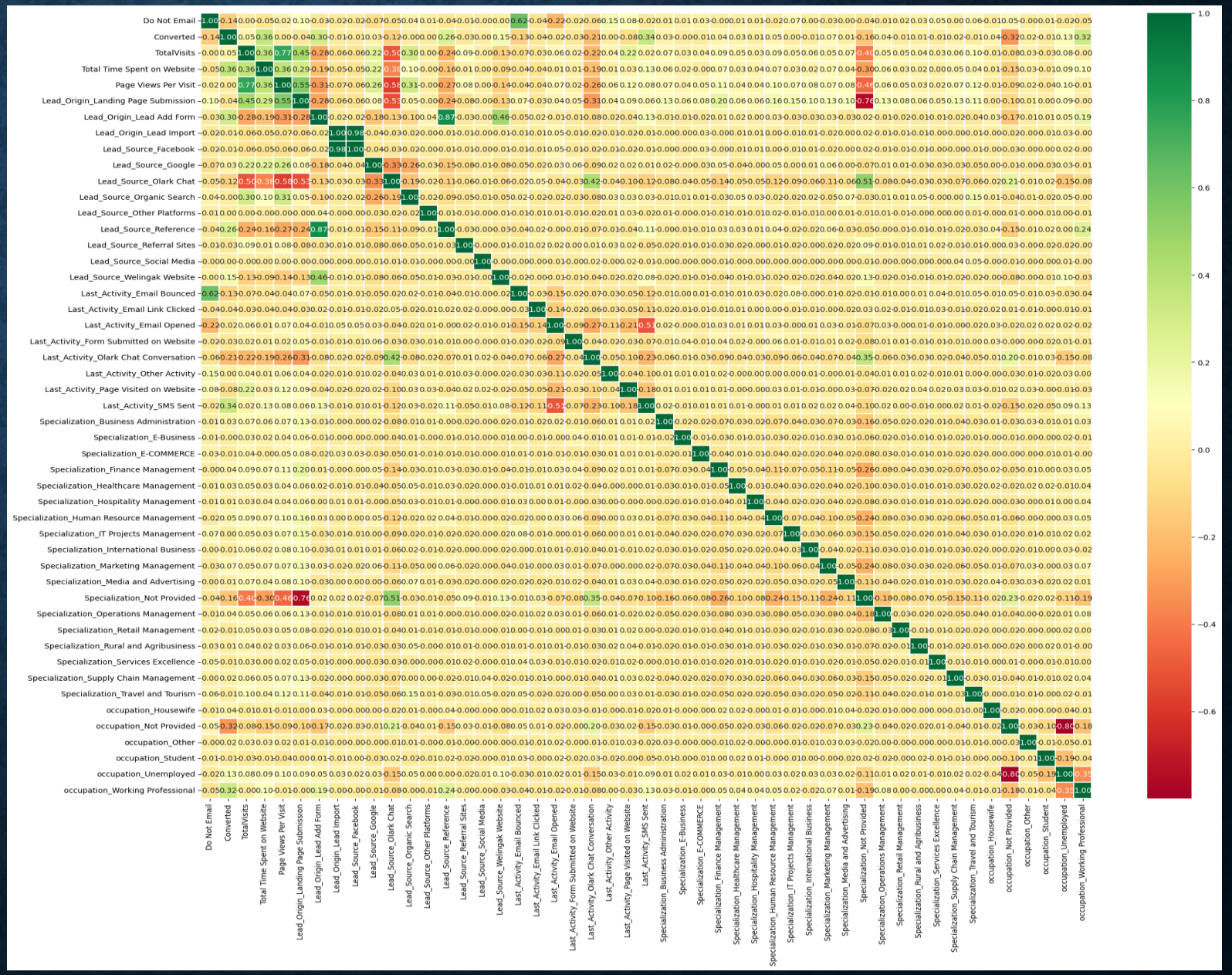
- Splitting the Train and Test data
- FEATURE SCALING: Rescaling the numerical data as to get balance between other columns
- Checking the Conversion Rate: Conversion rate is around 38%




```
0    62.144589
1    37.855411
Name: Converted, dtype: float64
```

❑ LOOKING THE CORRELATION

- Removing the column having high correlation with each other as it can hamper the model in further process model building and show variation in values.
- Removing (‘ Specialization_Not Provided’ , ‘ occupation_Not Provided’ , ‘Lead_Origin_Lead Import’) and checks the X_train correlation if there are more such columns to remove.



5. MODEL BUILDING

- We will build the Model of LOGISTIC REGRESSION using GLM

6. FEATURE SCALING

- Performed the RFE(RECURSIVE FEATURE ELIMINATION) and removing the column one by one after checking the p-value <0.05 and vif <5under 5.
- After getting the p-value and VIF under acceptable range We Predicted the y_train set

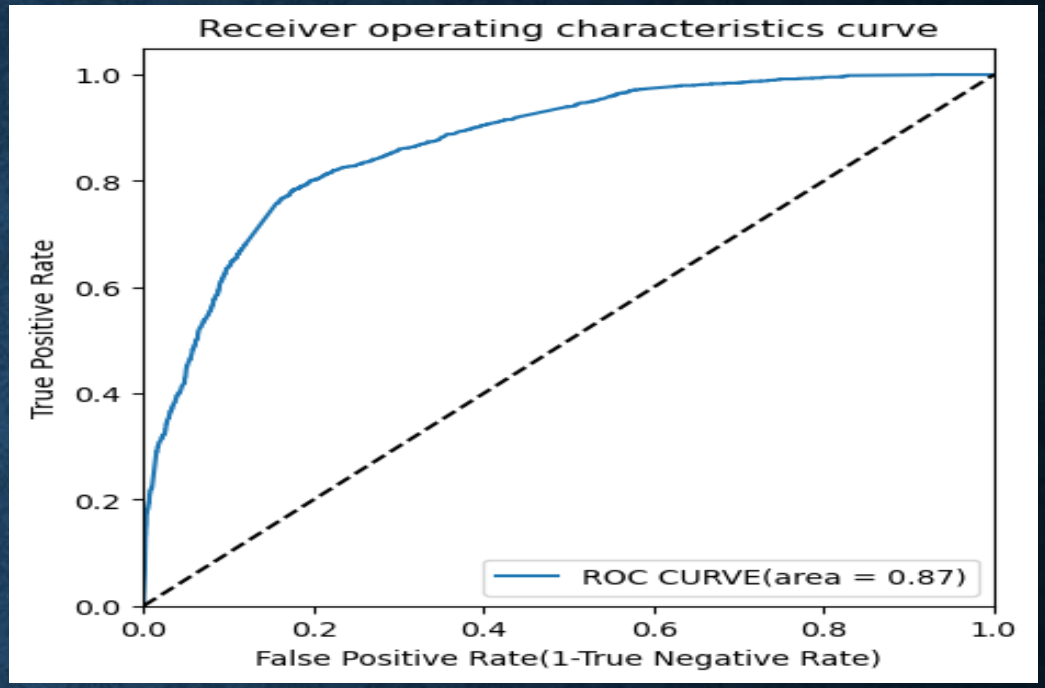
7. MODEL EVALUATION

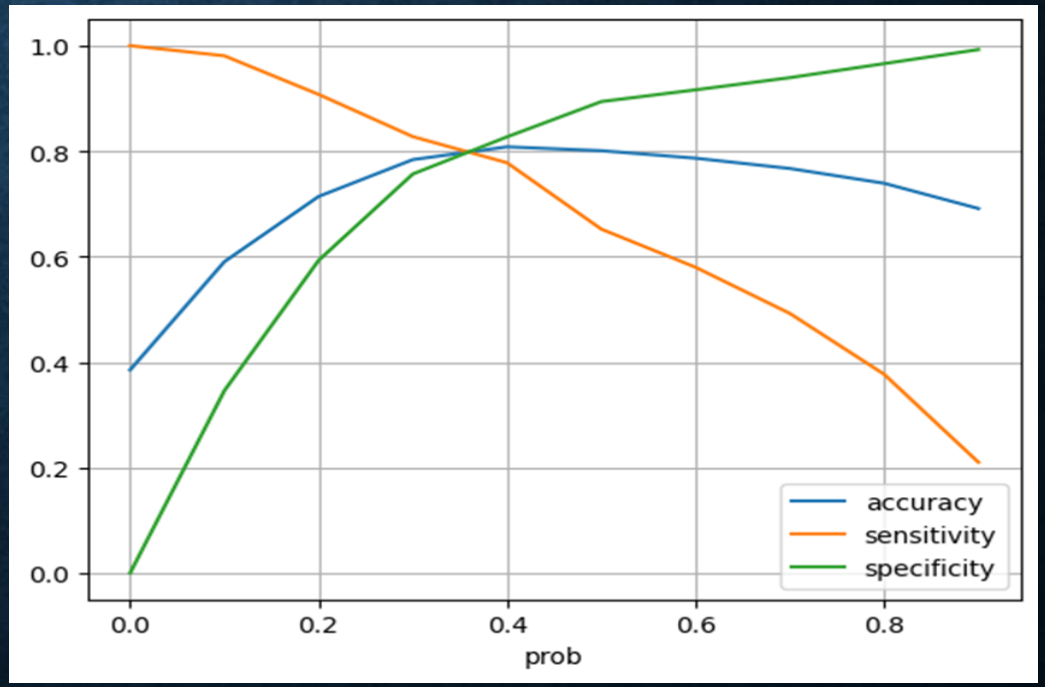
- Evaluating the model and getting the following details : using cutoff 0.35
  - ACCURACY : 80.22%
  - SENSITIVITY: 80.04%
  - SPECIFICITY: 80.03%

- Plotting the ROC(Receiver Operating Characteristic) curve: roc curve area= 0.87

- Finding the Optimal Cutoff using accuracy, sensitivity and specificity curve:

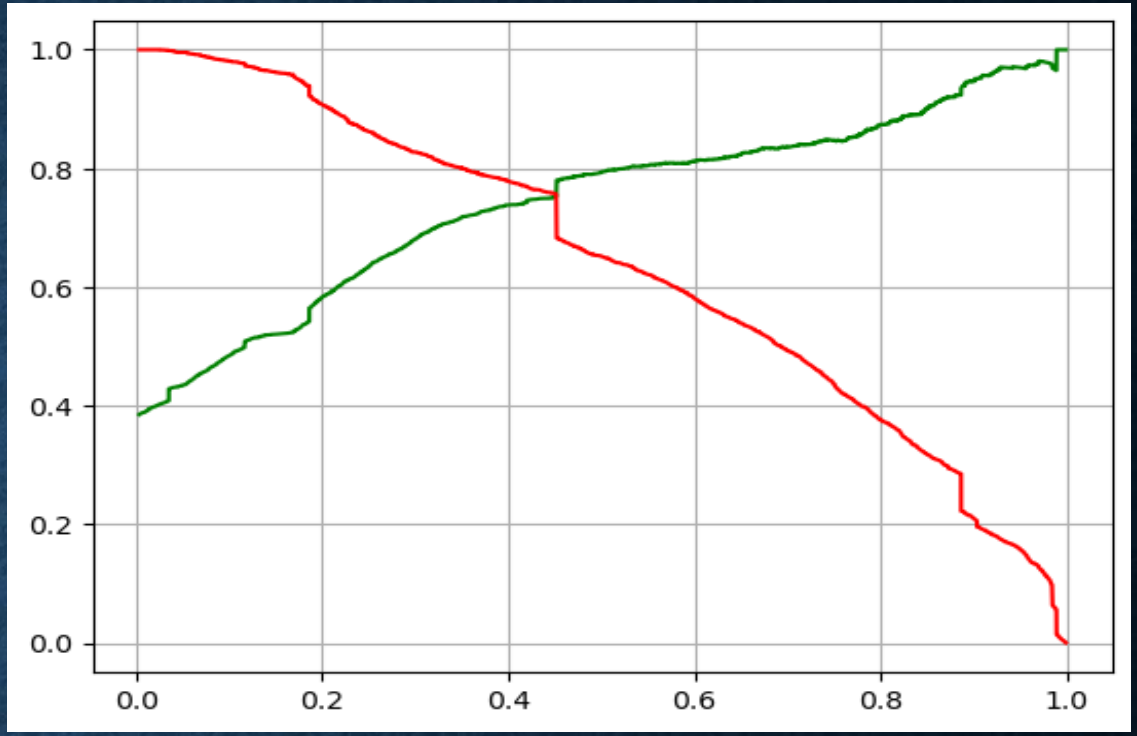
We took the optimal cutoff as 0.35





❑ Precision and Recall:

- Plotting a precision and recall trade off curve with cutoff 0.42
- ACCURACY: 81.02%
- PRECISION: 74.63%
- RECALL: 76.68

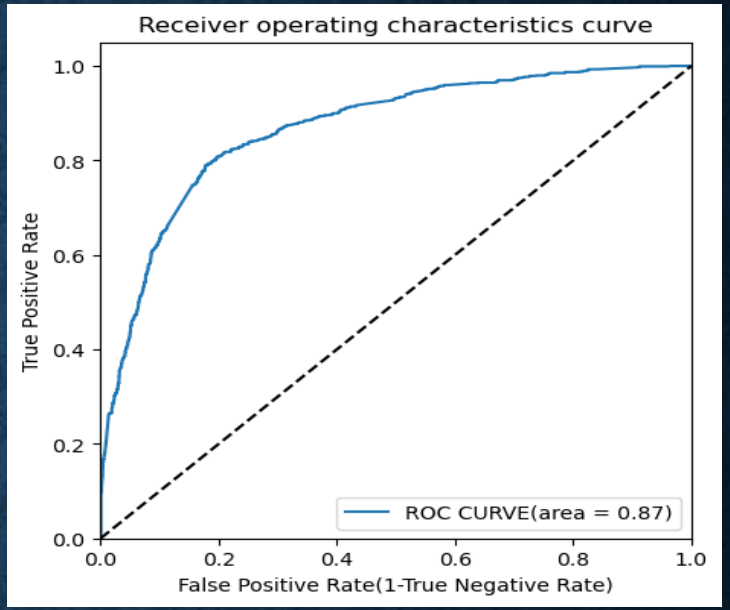


8. PREDICTION ON TEST SET

- Rescaled the Test set.
- Concatenating the predicted y set and y_test set in a data frame for further evaluation.

9. EVALUATION OF TEST SET

❑Plotting the ROC curve from the Test set



❑Result after evaluating the test set we get:

- ACCURACY: 80.27%
- SENSITIVITY: 80.78%
- SPECIFICITY: 79.98% =80%

❑ Features got from the final model:

occupation_Working Professional	3.709364
Lead_Origin_Lead Add Form	3.647871
Lead_Source_Welingak Website	2.068482
occupation_Other	1.803088
Lead_Source_Olark Chat	1.398972
occupation_Unemployed	1.279952
occupation_Student	1.150970
Total Time Spent on Website	1.141557
Last_Activity_Page Visited on Website	-0.879898
Do Not Email	-1.025525
Last_Activity_Form Submitted on Website	-1.087050
Last_Activity_Email Bounced	-1.744048
Last_Activity_Olark Chat Conversation	-1.830636
const	-1.858768

dtype: float64

10. CONCLUSION

- ❑ Optimal cutoff 0.35 is best for model prediction as well as for balancing accuracy, sensitivity and specificity
- ❑ The values of accuracy, sensitivity, specificity of test set are 80.27% ,80.78% and 79.98% respectively which are almost equal to the train set value.
- ❑ The sensitivity of the Train set is 80.04% and 80.78% for Test set.
- ❑ The Top 3 features responsible for conversion can be:

- occupation_Working Professional
- Lead_Origin_Lead Add Form
- Lead_Source_Welingak Website

- ❑ We achieved the target given by the CEO that the target lead conversion rate to be around 80%.

RECOMMENDATION

- Focus on leads with the occupation "Working Professional”.
- Give priority to leads from the "Lead Add Form" origin.
- Allocate additional resources to leads from the "Welingak Website" source.
- Utilize SMS as a communication channel for leads.
- Craft compelling email content for leads who have opened emails.
- Enhance engagement with leads who spend more time on the website.
- Regularly monitor the model's performance and update it as needed.
- Incorporate feedback from the sales team for model refinement.
- Allocate resources based on lead scores for efficient resource utilization.
- Train the sales team to utilize lead scores and engage effectively with high-potential leads.

THANK YOU