Lead Scoring Case study

Team:

Keerthana

Mahima

Kerel samuel

Problem statement

- An education company named X Education sells online courses to industry professionals.
- The company advertises about them in different platforms
- The people visiting their website and registering are called as ***leads***.
- The company wants to find the Most promising leads called ***Hot Leads***
- So, that they focus and spend their time on the hot leads

Business objective:

Wants to know the Hot Leads to expand the conversion rate that helps intern grow their business

Steps followed

- Imported the data
- Cleaned the data
  - Dropping the features null columns >35%

    Tags, Lead Quality, Lead Profile, City, Asymmetrique Activity Index, Asymmetrique Profile Index, Asymmetrique Activity Score, Asymmetrique Profile Score
  - Dropping the unique values=1 and unique values= total rows and Unwanted columns for analysis

    Magazine, Receive More Updates About Our Courses, Update me on Supply Chain Content, Prospect ID, Lead Number, Get updates on DM Content, I agree to pay the amount through cheque
  - No outliers are present in this dataset
  - Converted the Select as well as NAN as suggested

EDA

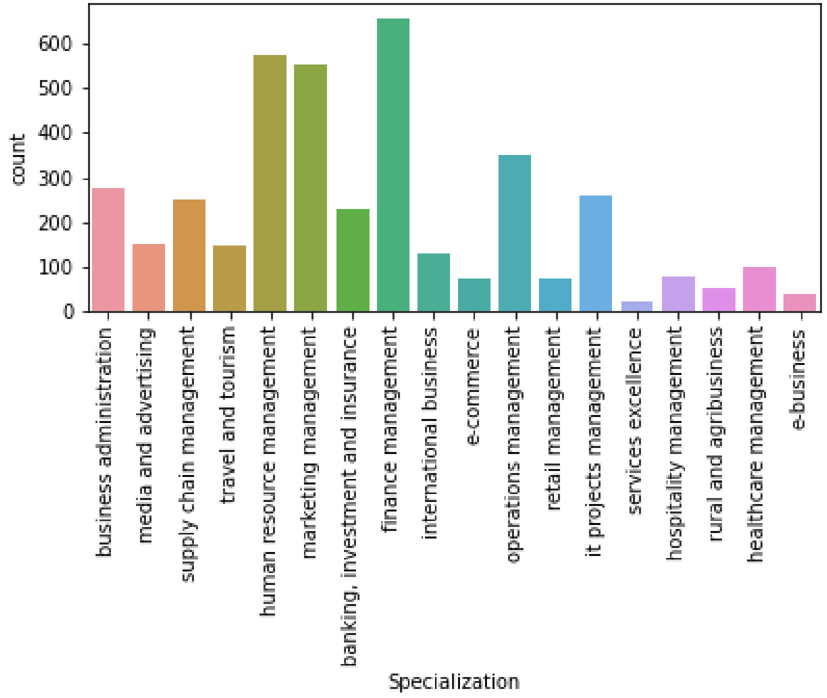
- Checking the categorical values and the numeric separately and visualizing them
- Heat map for the correlation with the Target variable converted.
- Categorical variables:

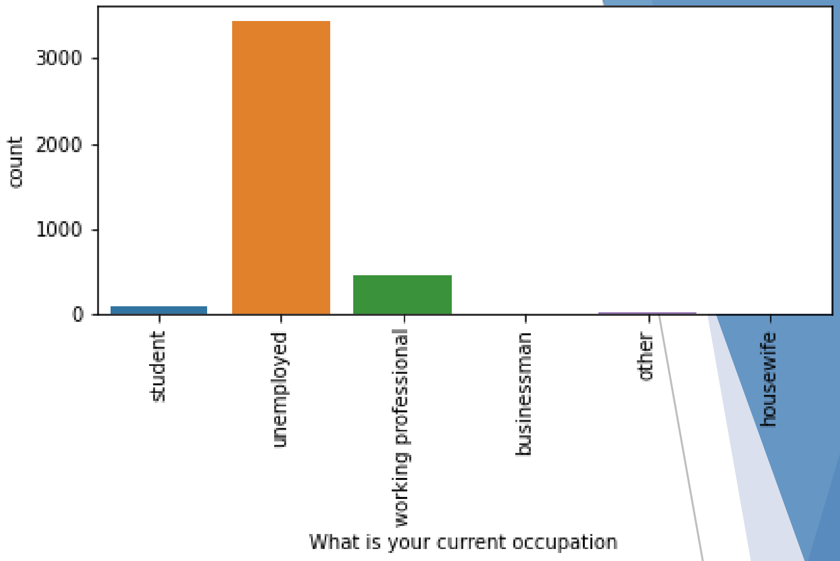
Lead Origin, Do Not Email, Do Not Call, Search, Newspaper Article, X Education Forums, Newspaper, Digital Advertisement, Through Recommendations, A free copy of Mastering The Interview, Specialization, What is your current occupation, What matters most to you in choosing a course, Last Activity, Country, Last Notable Activity

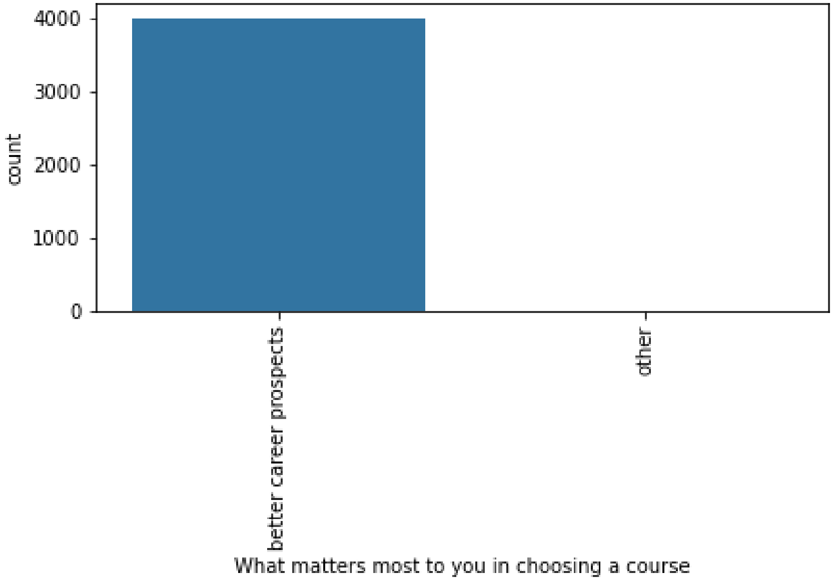
- Numeric Variables:

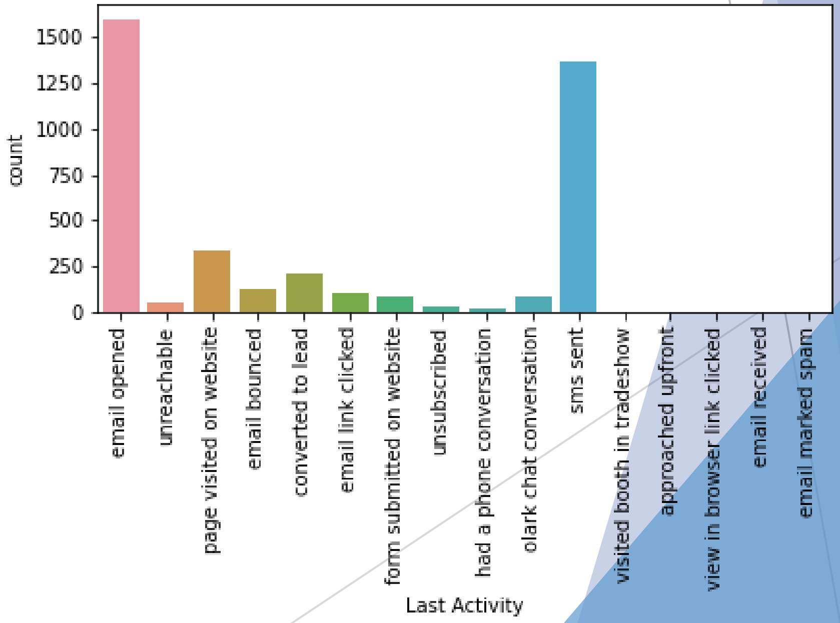
TotalVisits, Total Time Spent on Website, Page Views Per Visit

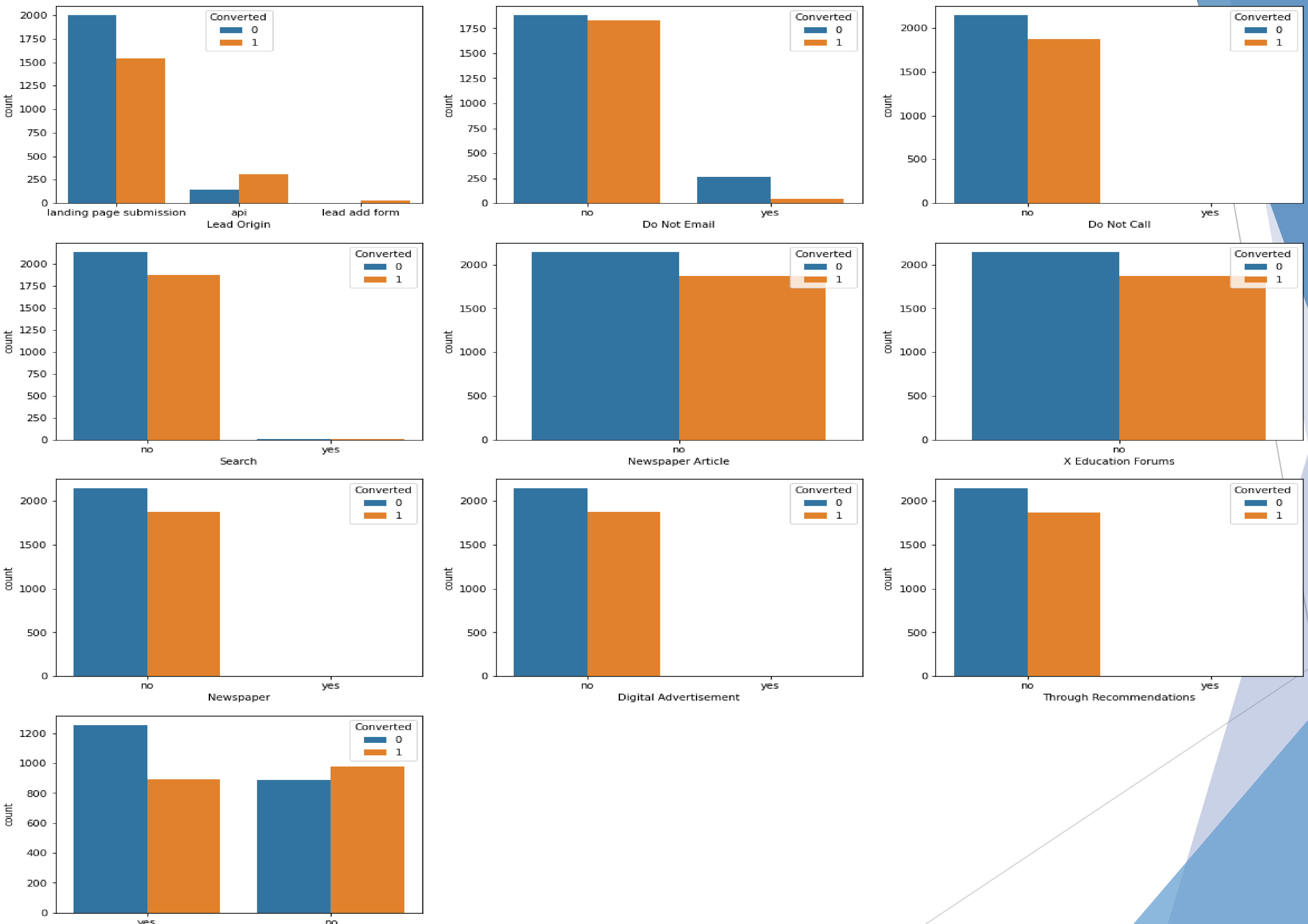
EDA categorical









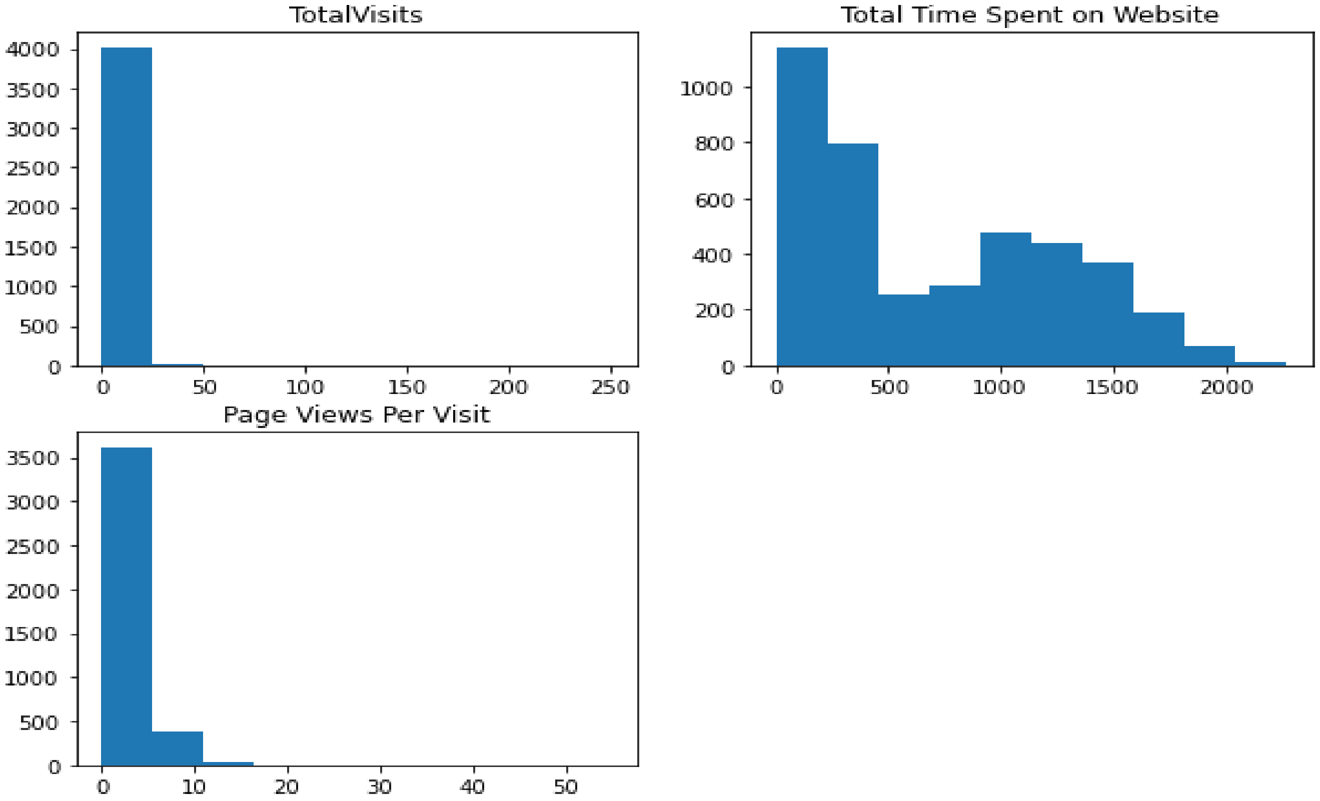


Converted
0
1
count
2000
1750
1500
1250
1000
750
500
250
0
landing page submission
api
lead add form
Lead Origin
no
yes
Do Not Email
Do Not Call
Search
Newspaper Article
X Education Forums
Newspaper
Digital Advertisement
Through Recommendations
1200
800
600
400
200
yes
no

- The News paper article, Digital Advertisement, Through Recommendations, Newspaper, search, Xeducation forums, Don’t call parameters have only NO as reply
- Equal no of people took and rejected the free copy of mastering the interview
- Last activity is through the email, SMS , pages visited on the website
- Lead origin alone seems to have seams importance
- lead source is mostly rely on the direct traffic , google and organic searches
- unemployed category are prominently checking out for the course
- Only goal of taking up this is to get better career opportunities

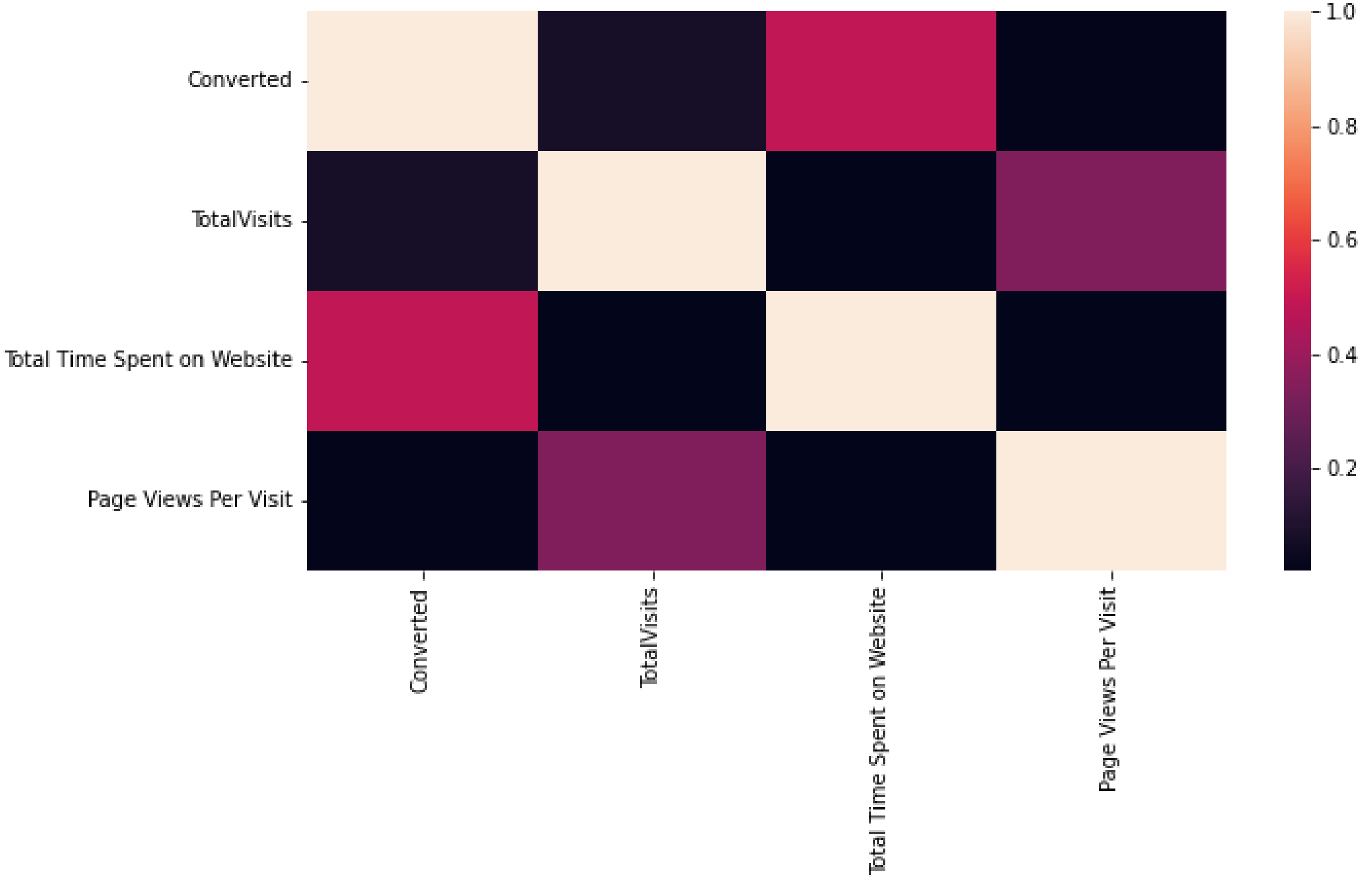
- Specialization with finance, HR, Marketing and health care , business shows more interest in the course
- API has more conversion rate following the landing page submission
- ones opted for Call and Email has equal conversion and rejection rate
- More people converted rejected the Free copy of mastering the interviews as a point to focus
- The most Conversion rate and rejection rate depends solely on INDIA , so, should be focusing more over that region
- High conversion are thru the SMS and then the Emails
- conversion rate is more than rejection in marketing and banking and health care sectors

Numeric EDA



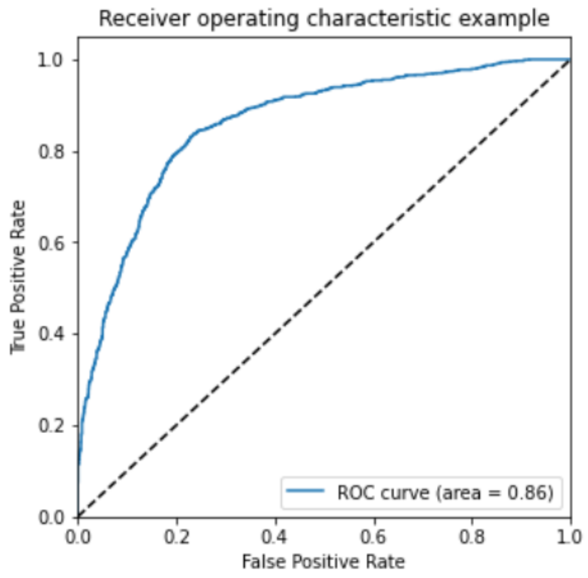
Total time spend gave more info about the conversion rate and the total time the person visited the website as well

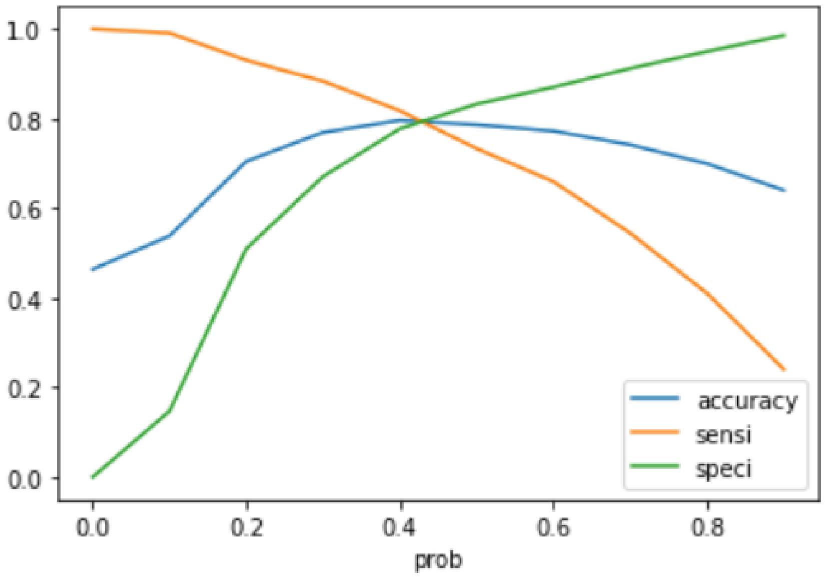
Heat MAP

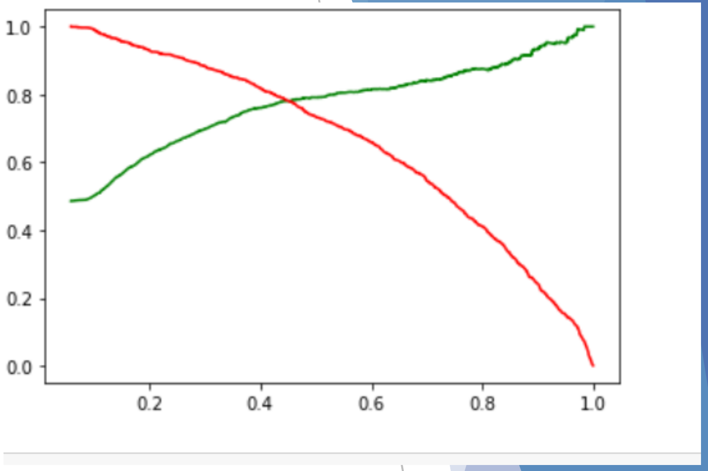


Total time spent on the website is an important parameter as it is very much correlated with the Conversion rate

- **Splitting the dataset to Training and testing**
  - 70% train and 30% test
- **Model building**
  - RFE with features are 15 is used to get the best 15 features
  - Using the Logistic regression model
- **Mode evaluation**
  - Using confusion matrix and the Accuracy, sensitivity, specificity
  - 79% accuracy is good percentage with sensitivity and specificity being 74% and 83%
  - ROC with 0.86 area
- **Prediction**
  - Predicting the test data set , both the percentages are very aligning meaning that the model is build properly
- **Precision and Recall**
  - Precision 73%
  - Recall 79%







The cutt of here is 0.4

Area under the ROC curve =0.86

The accuracy, sensitivity and specificity curve

Precision vs recall graph has cut off of **0.41**

CONCLUSION

- Parameters found to be useful are as follows:
  1. Total time spent on the website
  2. Lead source is mostly rely on the direct traffic, google and organic searches
  3. Last activity SMS and then the Emails
  4. What is your current occupation - unemployed
  5. Lead origin as API , Lead add form, landing page submission
  6. Country – India
  7. Specialization with marketing and banking and health care sectors