LEAD SCORING

CASE STUDY

BY
Manoj
Sandhya
Sania

- Problem Statement :

1. X Education sells online courses to industry professionals.
2. The typical lead conversion rate at X education is around 30%.
3. Despite getting lot of leads, its lead conversion rate is very poor.
4. To improve the rate, the company wishes to identify the best potential leads, also known as ‘Hot Leads’.
5. By Doing so, the sales team can focus more on communicating with the potential leads rather than making calls to everyone.

- Business Objective:

1. X education wants to Identify the most promising leads.
2. They want a Model which identifies the hot leads, and to assign a lead score to each of the leads such that the customers with a higher lead score have a higher conversion chance and the customers with a lower lead score have a lower conversion chance.
3. The CEO, in particular, has given a ballpark of the target lead conversion rate to be around 80%.

Solution Methodology

- **<u>Data cleaning and data manipulation.</u>**

  1. Checking and handling duplicate data.

  2. Check and handle NA values and missing values.

  3. Dropping columns, if it has large number of missing values and if the column is not useful for analysis.

  4. Imputation of the values, if and when necessary.

  5. Check and handle outliers in data.

- <u>EDA</u>

  Univariate data analysis: value count, distribution of variable etc.

- Feature Scaling & Dummy Variables and encoding of the data.

- **<u>Classification technique:</u>**

  1.logistic regression used for the model making and prediction.

  2.Validation of the model.

  3.Model presentation.

  4.Conclusions and recommendations.

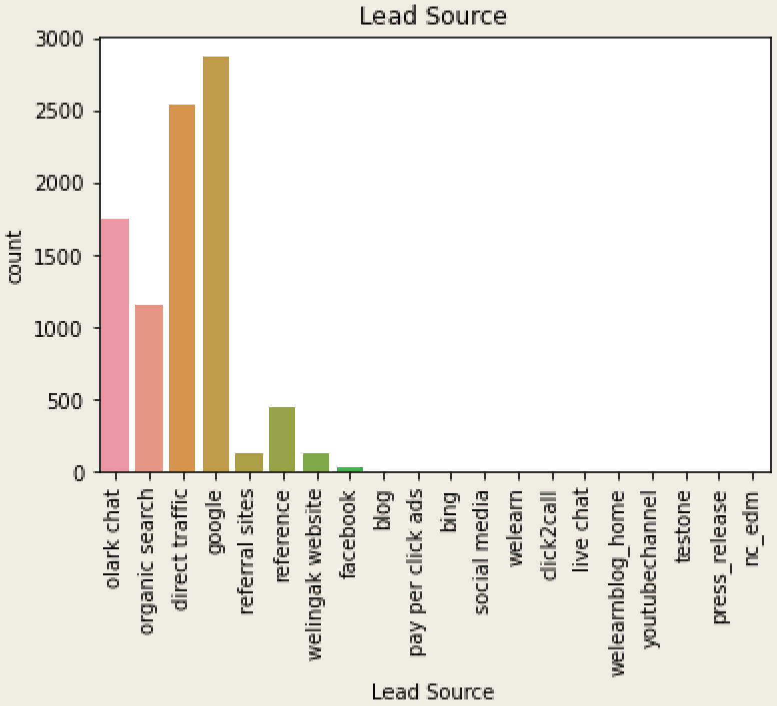
Data Manipulation:

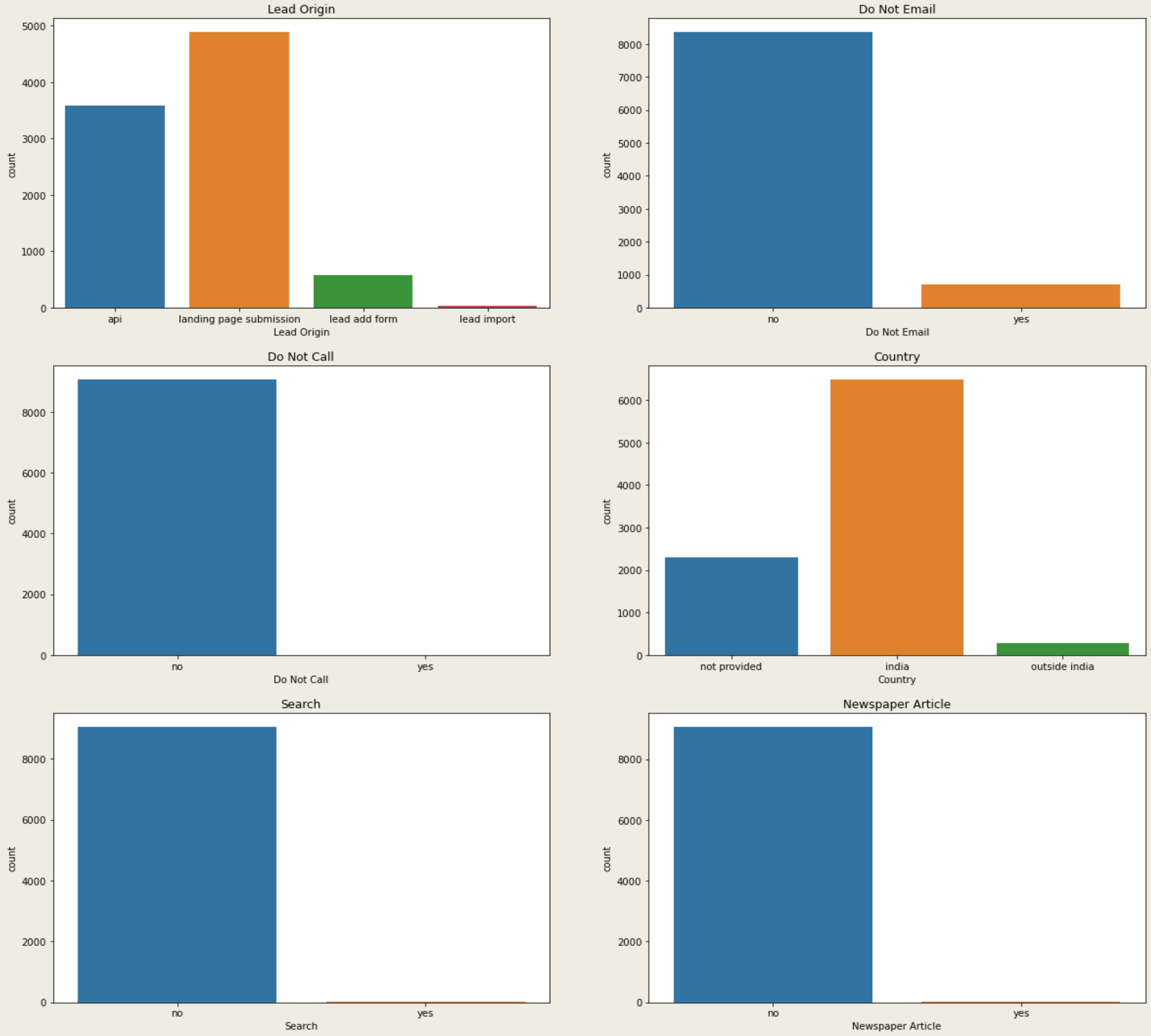
- The data set has 37 variables(rows) and 9240 records(columns).

- Dropping unique valued columns like , 'Magazine','Receive More Updates About Our Courses','I agree to pay the amount through cheque','Get updates on DM Content','Update me on Supply Chain Content’.

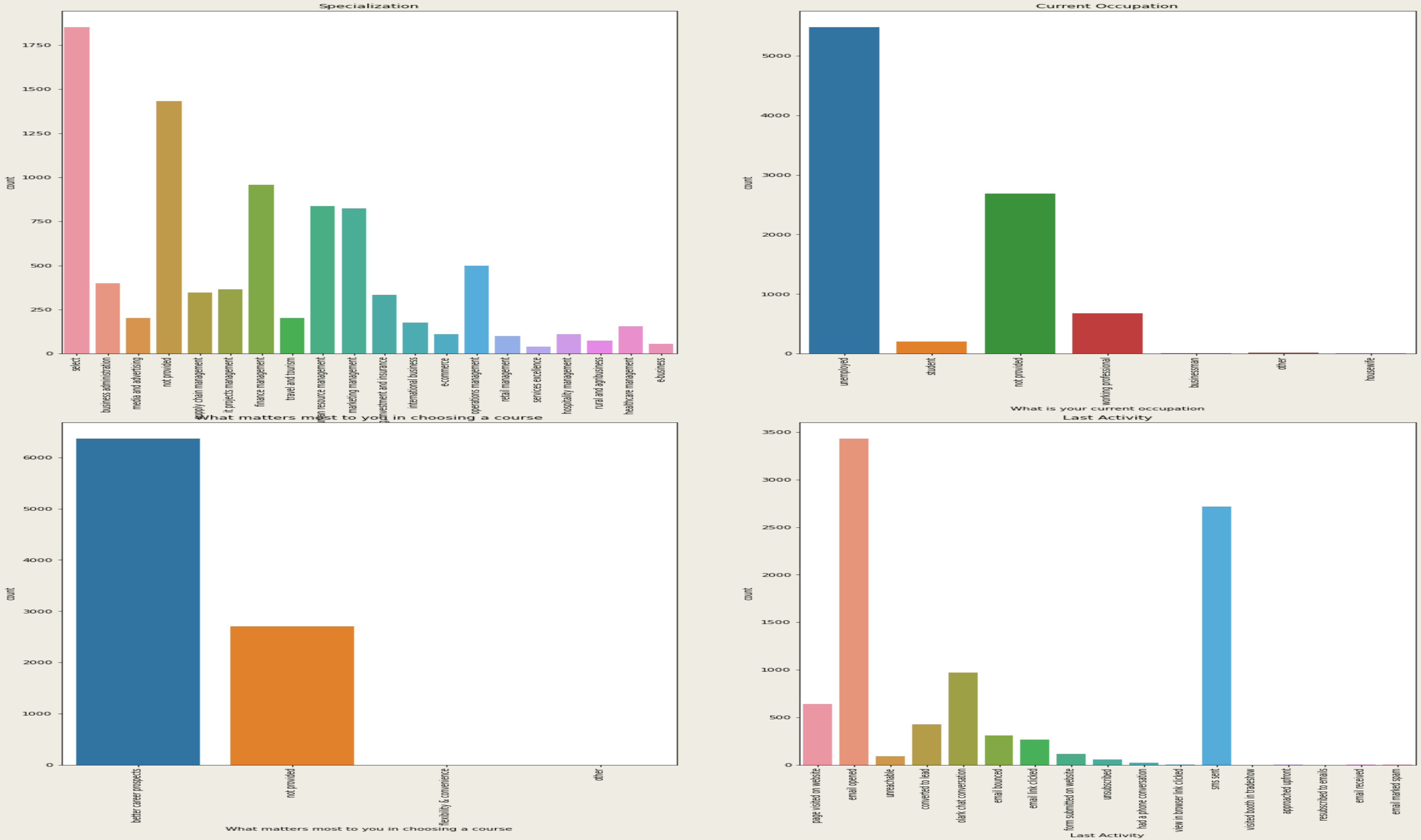
- Dropping columns that are not relevant and has more than 35% Null Values such as 'Asymmetrique Profile Index','Asymmetrique Activity Index','Asymmetrique Activity Score’,'Asymmetrique Profile Score','Lead Profile','Tags','Lead Quality’,'How did you hear about X Education','City','Lead Number’.

EDA

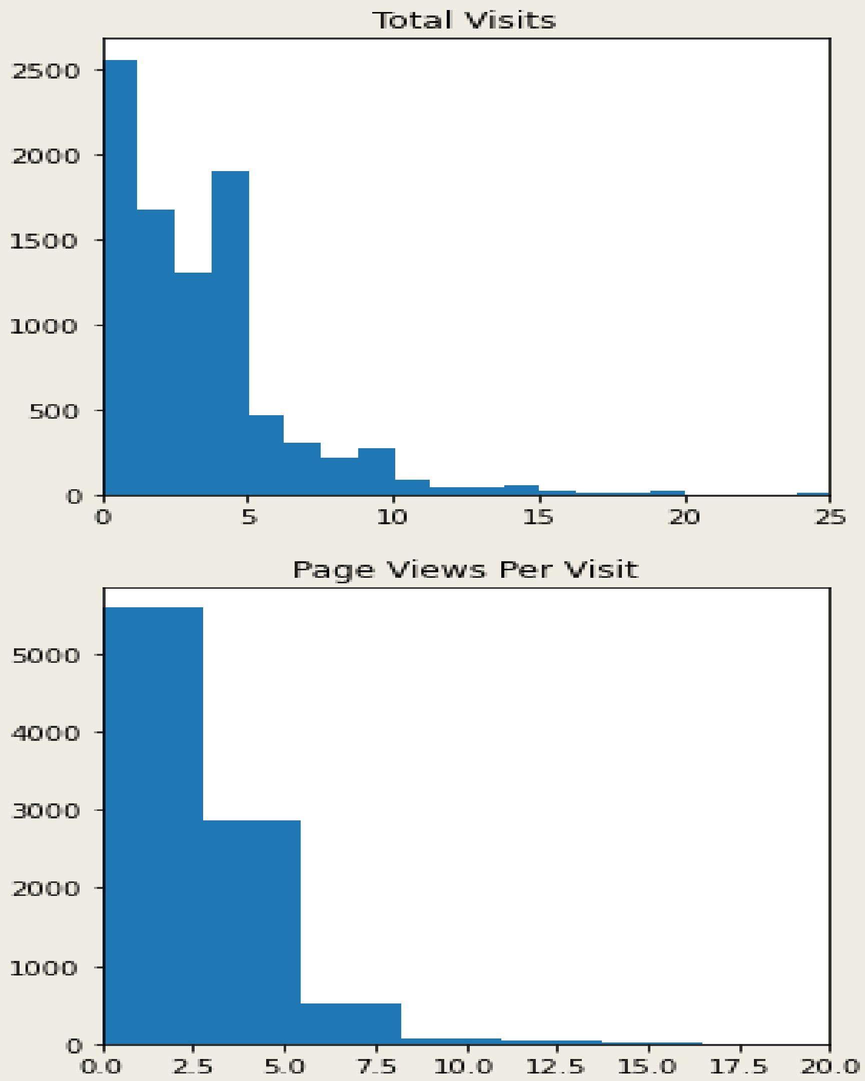

Lead Source
count
3000
2500
2000
1500
1000
500
0
olark chat
organic search
direct traffic
google
referral sites
reference
welingak website
facebook
blog
pay per click ads
bing
social media
welearn
click2call
live chat
welearnblog_home
youtubechannel
testone
press_release
nc_edm
Lead Source


Lead Origin
count
5000
4000
3000
2000
1000
0
api
landing page submission
lead add form
lead import
Lead Origin
Do Not Email
count
8000
7000
6000
5000
4000
3000
2000
1000
0
no
yes
Do Not Email
Do Not Call
count
8000
6000
4000
2000
0
no
yes
Do Not Call
Country
count
6000
5000
4000
3000
2000
1000
0
not provided
india
outside india
Country
Search
count
8000
6000
4000
2000
0
no
yes
Search
Newspaper Article
count
8000
6000
4000
2000
0
no
yes
Newspaper Article

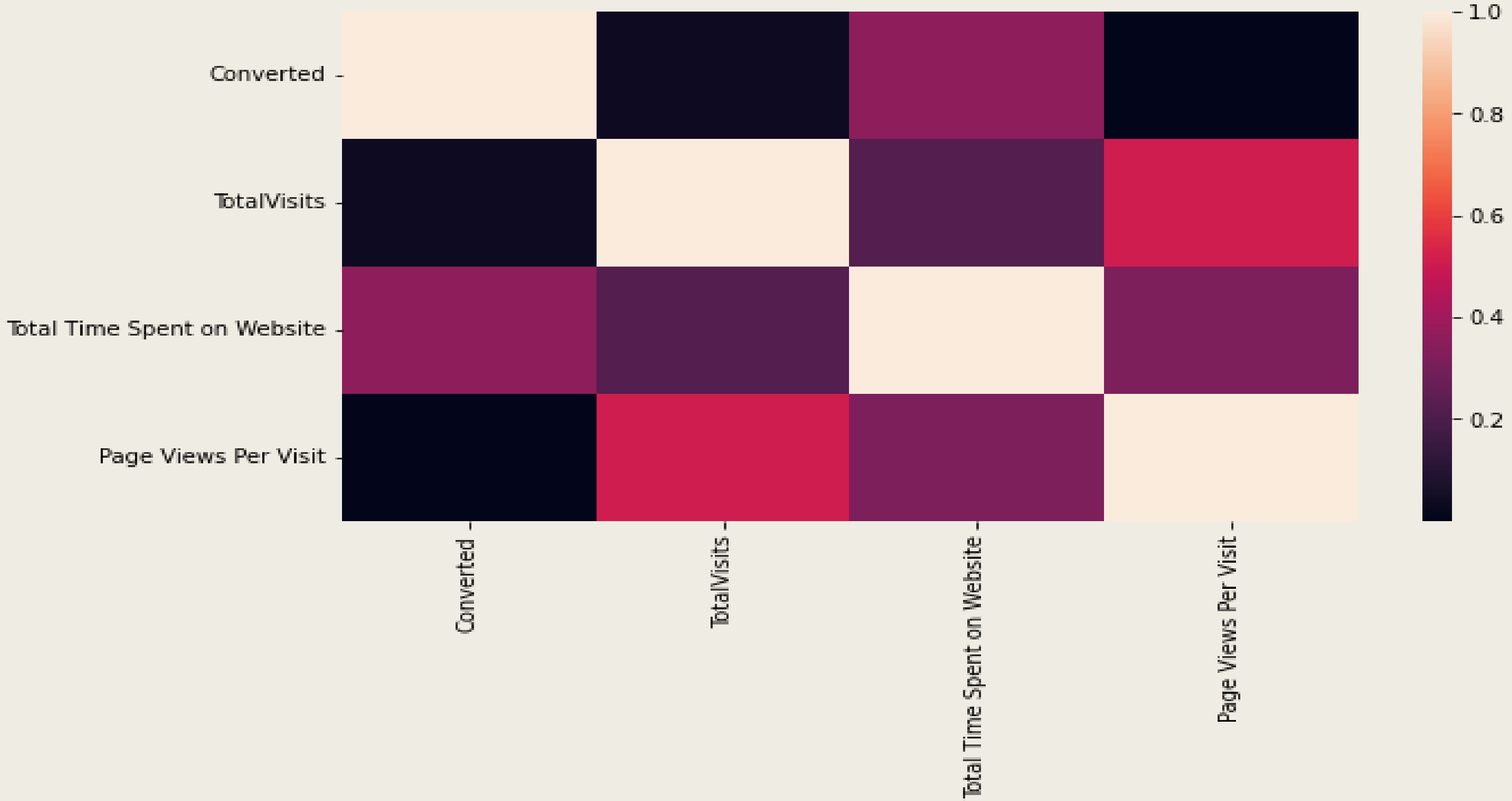
Categorical features



Numerical features



Numerical Features



- Data Conversion:
  1. Normalising numerical variables
  2. Created dummy variables for object type variables.

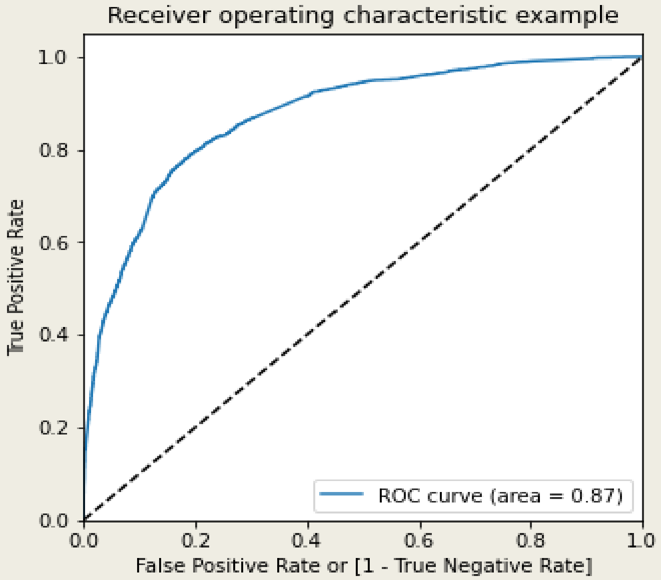
- Model Building:
  1. Splitting the Data into Training and Testing Sets

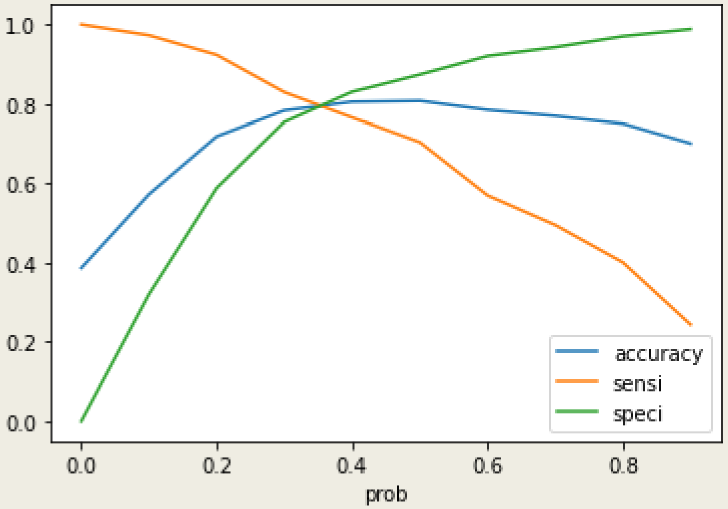
  2. To perform regression creating a train-test split, we have chosen 70:30 ratio.

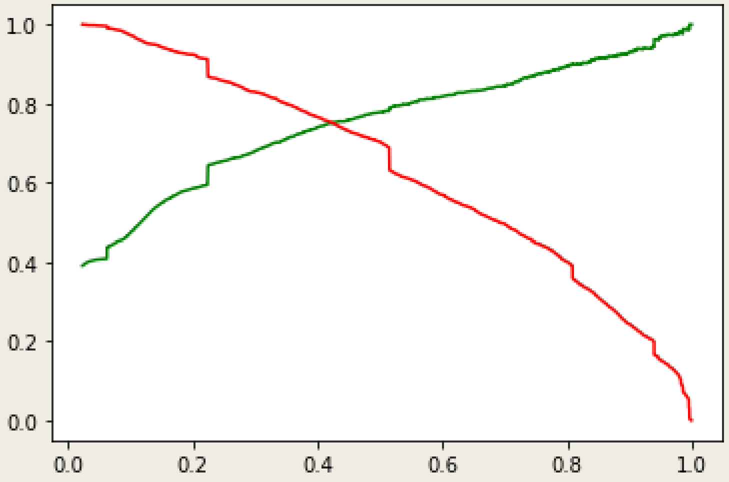
  3. Use RFE for Feature Selection

  4. Building Model by removing the variable whose p-value is greater than 0.05 and vif value is greater than 5
  5. Overall accuracy 80.05% on test data, and 79.67% on train data set.

ROC curve







- Finding Optimal Cut off Point
- Optimal cut off probability: is the Probability where we get balanced sensitivity and specificity.
- The second graph shows the optimal cut off is at 0.35.
- Precision and recall chart shows suitable threshold is 0.41
- Although the values are same at both cutoff 0.35 and 0.41, so kept 0.35.

Conclusion

It was found that, features mattered the most in the potential buyers are :

1. The total time spend on the Website.
2. Total number of visits.
3. When the lead source was: a. Google b. Direct traffic c. Organic search d.Welingak website
4. When the last activity was: a. SMS b. Olark chat conversation
5. When the lead origin is Lead add format.
6. When their current occupation is as a working professional.