Lead Score Case Study

DHAVAL DAMANIA
HOANG NGOC TIEN
AJINKYA KORDE

Problem Statement:

- An X Education company has a very poor lead conversion rate.
- The company wishes to identify the most potential leads, also known as ‘Hot Leads’, so that it can improve the conversion rate.
- The company requires us to build a model wherein you need to assign a lead score to each of the leads such that the customers with higher lead score have a higher conversion chance and the customers with lower lead score have a lower conversion chance.
- The CEO, in particular, has given a ballpark of the target lead conversion rate to be around 80%.

Business Goals & Objective :

There are quite a few goals for this case study.

- Build a logistic regression model to assign a lead score between 0 and 100 to each of the leads which can be used by the company to target potential leads.
- A higher score would mean that the lead is hot, i.e. is most likely to convert into a paying customer whereas a lower score would mean that the lead is cold and will mostly not get converted.
- The target lead conversion rate need to be around 80%.

Solution Methodology :

- Data cleaning and data manipulation.

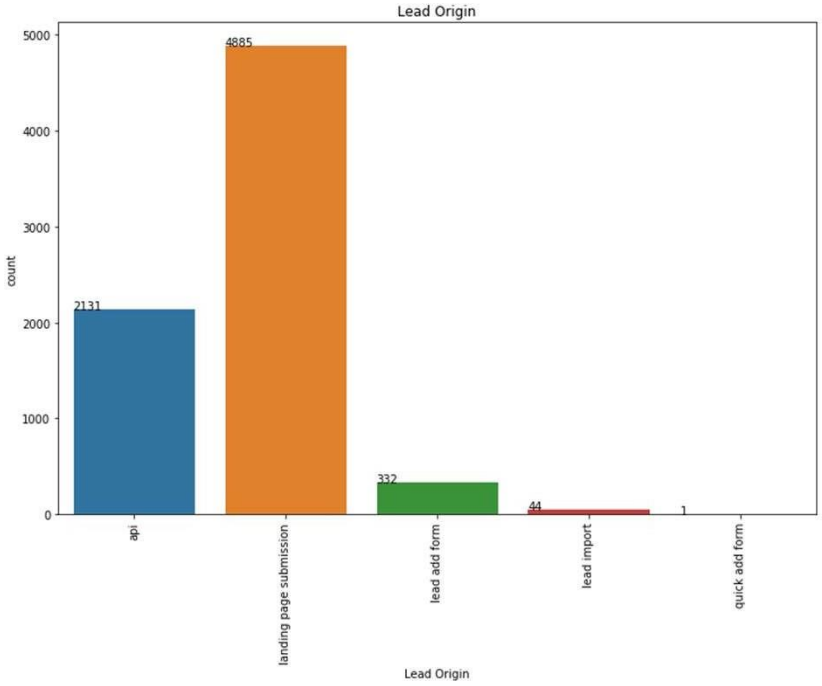
1. Drop the duplicate records.
2. Drop columns, if it contains large amount of missing values and not useful for the analysis.
3. Handle the missing values. Imputation of the values, if necessary.
4. Check and handle outliers in data.

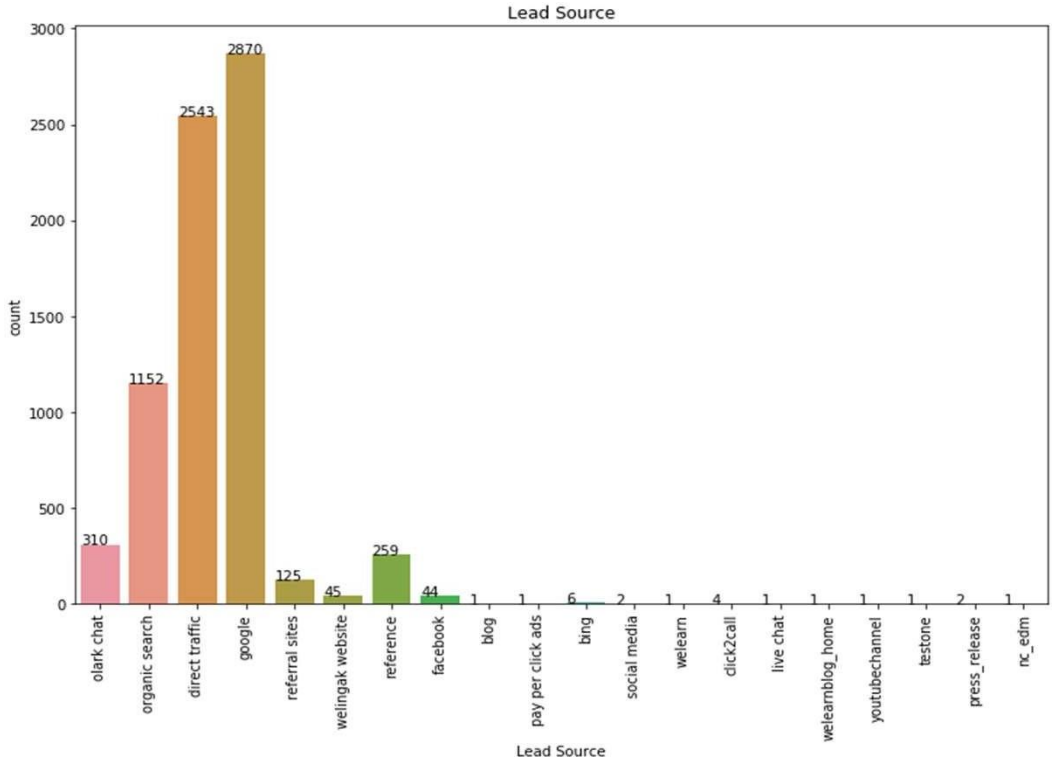
- EDA
  1. **Univariate data analysis: value count, data distribution**
  2. **Segmented Univariate data analysis.**
  3. **Bivariate data analysis: correlation coefficients and pattern between the variables etc.**
  4. **Multivariate Data Analysis**
- Dummy Variables Creations Where needed
- Scaling , Feature Selection and encoding of the data.
- Classification technique: logistic regression used for the model making and prediction.
- Validation of the model.
- Model presentation.
- Conclusions and recommendations.

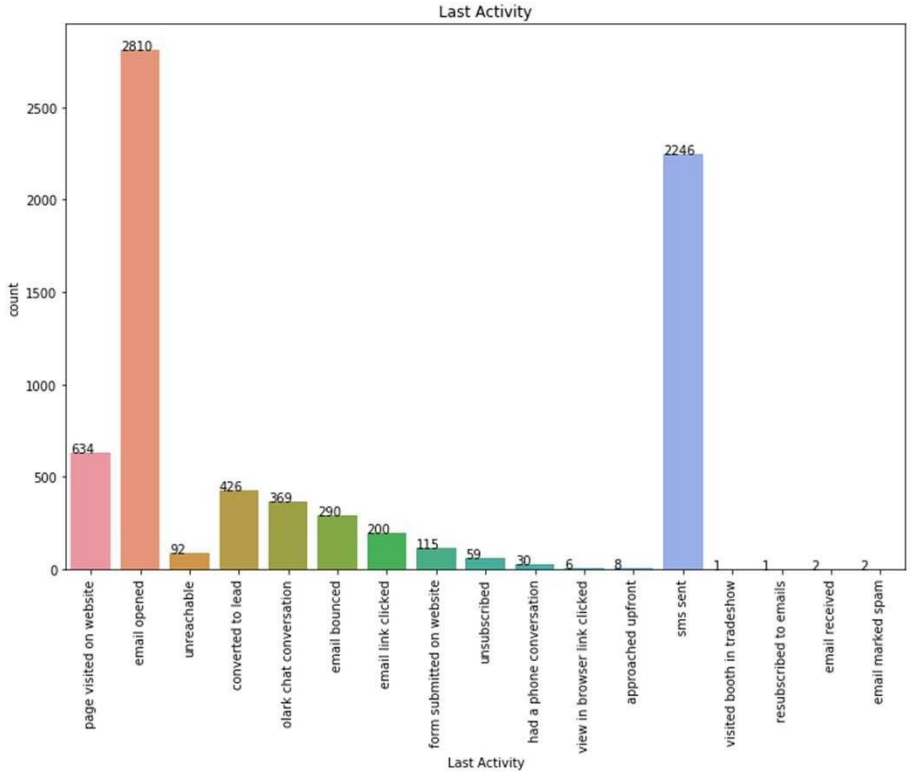
Data Manipulation :

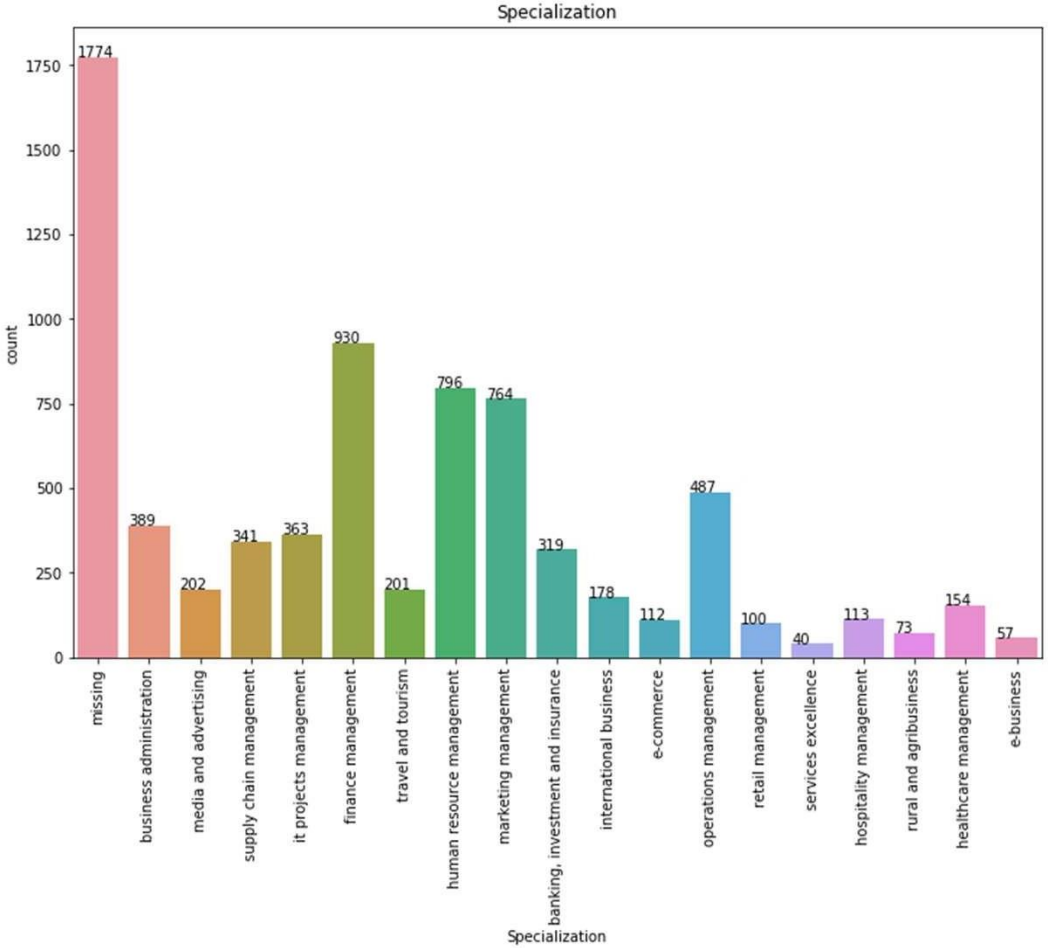
- Total Number of Rows = 9240, Total Number of Columns = 37
- Single value features like “Magazine”, “Receive More Updates About Our Courses”, “Update me on Supply”
- Chain Content”, “Get updates on DM Content”, “I agree to pay the amount through cheque” etc. have been dropped.
- Removing the “Prospect ID” and “Lead Number” which is not necessary for the analysis
- Dropping the columns having more than 30% as missing value such as ‘How did you hear about X Education’ and ‘Lead Profile’.

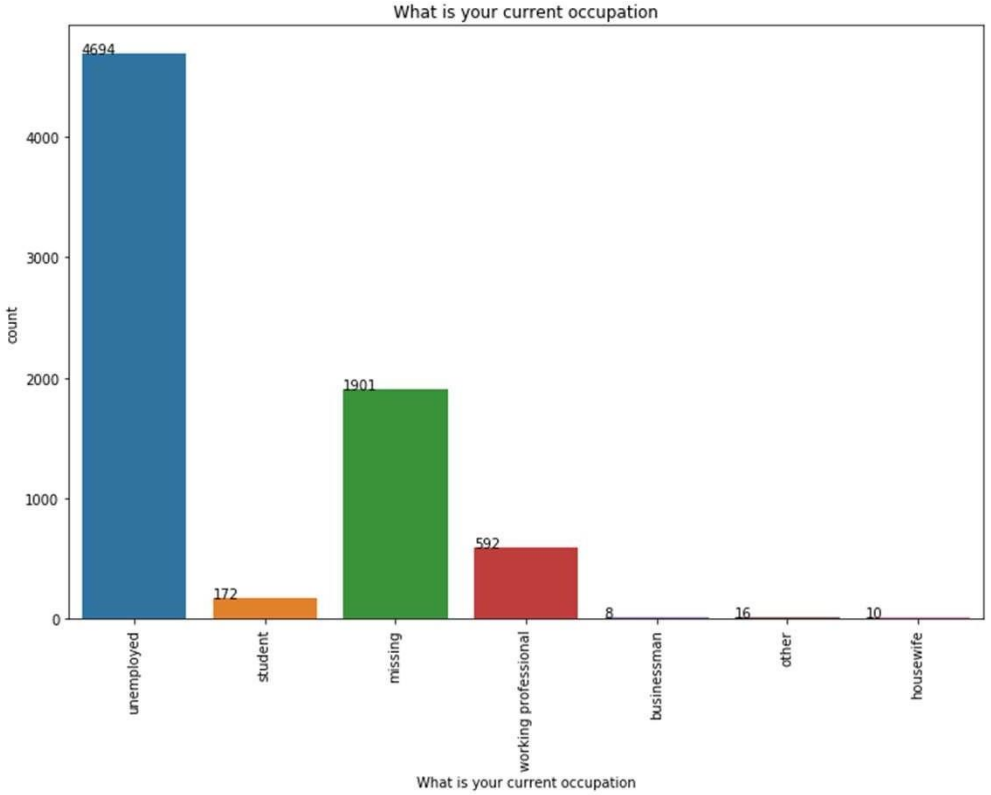
EDA :

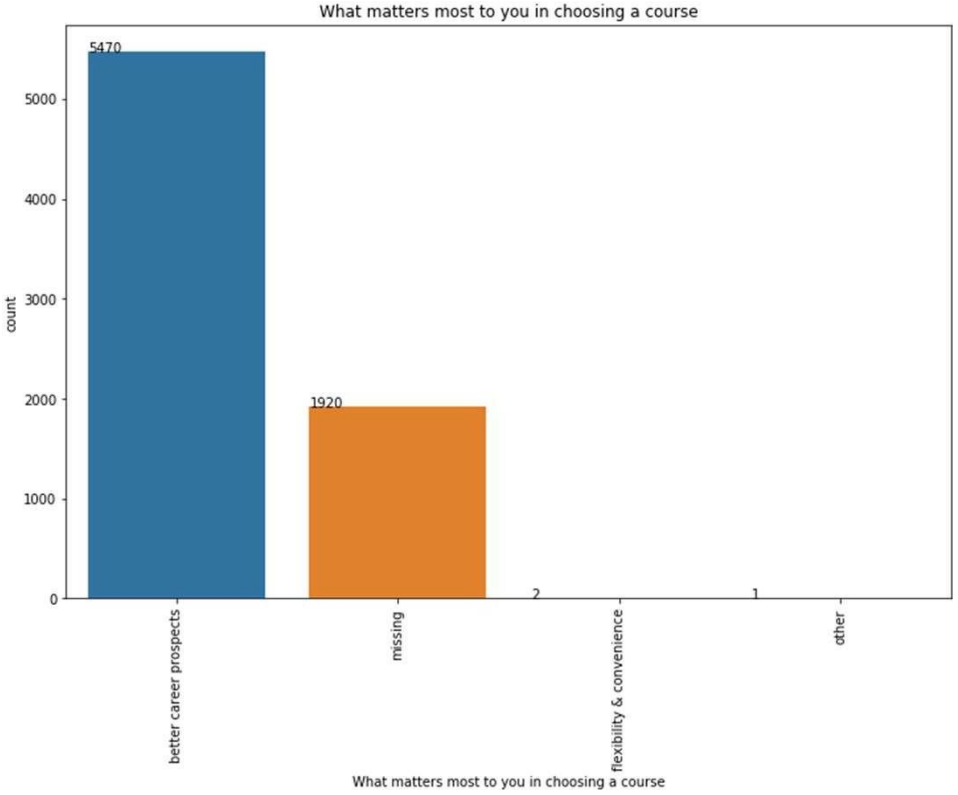

Lead Origin
count
5000
4000
3000
2000
1000
0
2131
4885
332
44
1
api
landing page submission
lead add form
lead import
quick add form
Lead Origin

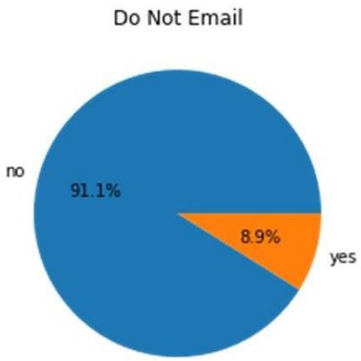

Lead Source
count
3000
2500
2000
1500
1000
500
0
310
1152
2543
2870
125
45
259
44
1
1
6
2
1
4
1
1
1
1
2
1
olark chat
organic search
direct traffic
google
referral sites
welingak website
reference
facebook
blog
pay per click ads
bing
social media
welearn
click2call
live chat
welearnblog_home
youtubechannel
testone
press_release
nc_edm
Lead Source

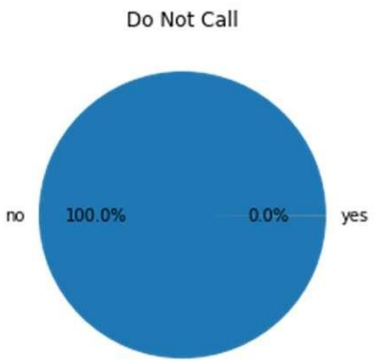

Last Activity
count
Last Activity
page visited on website
email opened
unreachable
converted to lead
olark chat conversation
email bounced
email link clicked
form submitted on website
unsubscribed
had a phone conversation
view in browser link clicked
approached upfront
sms sent
visited booth in tradeshow
resubscribed to emails
email received
email marked spam
634
2810
92
426
369
290
200
115
59
30
6
8
2246
1
1
2
2

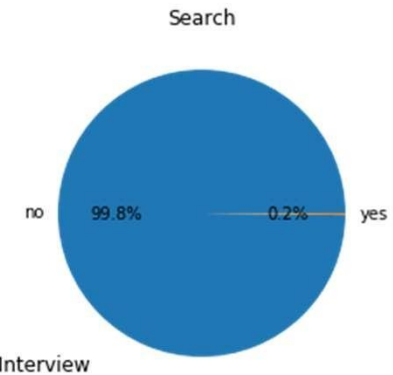

Specialization
count
Specialization
missing
business administration
media and advertising
supply chain management
it projects management
finance management
travel and tourism
human resource management
marketing management
banking, investment and insurance
international business
e-commerce
operations management
retail management
services excellence
hospitality management
rural and agribusiness
healthcare management
e-business
1774
389
202
341
363
930
201
796
764
319
178
112
487
100
40
113
73
154
57

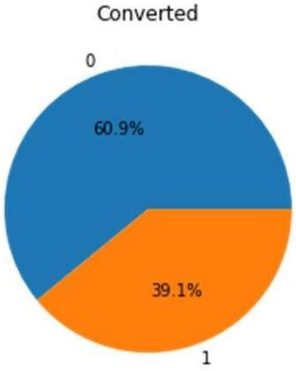

What is your current occupation
count
4694
172
1901
592
8
16
10
unemployed
student
missing
working professional
businessman
other
housewife
What is your current occupation

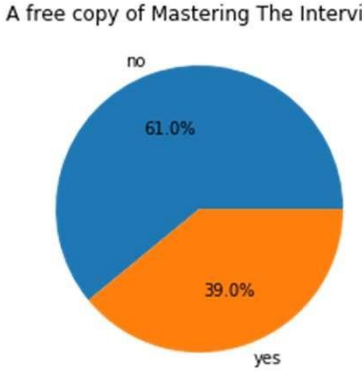

What matters most to you in choosing a course
count
5470
1920
2
1
better career prospects
missing
flexibility & convenience
other
What matters most to you in choosing a course

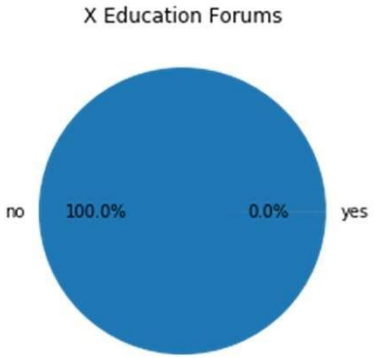

Do Not Email
no
91.1%
8.9%
yes


Do Not Call
no
100.0%
0.0%
yes

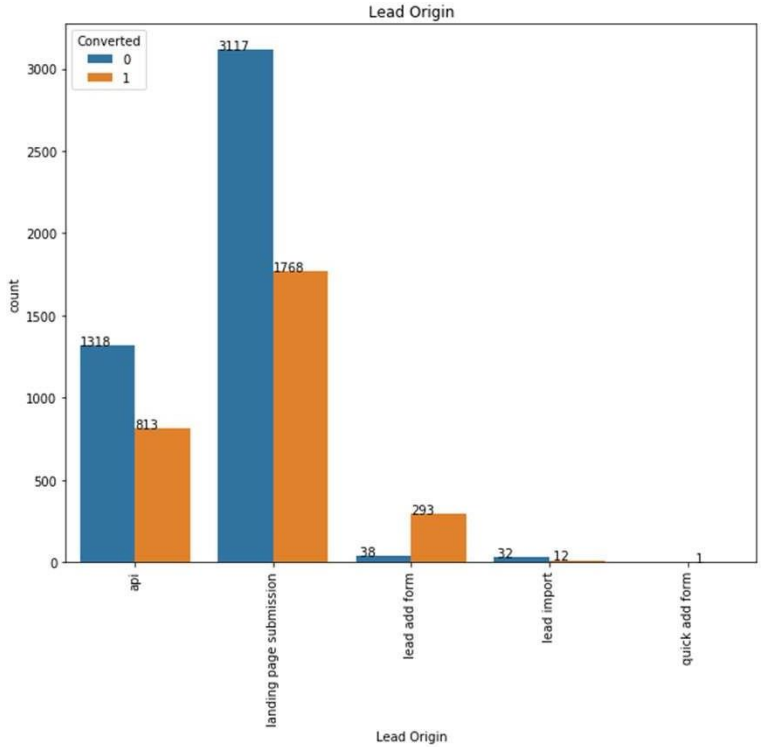

Search
no
99.8%
0.2%
yes

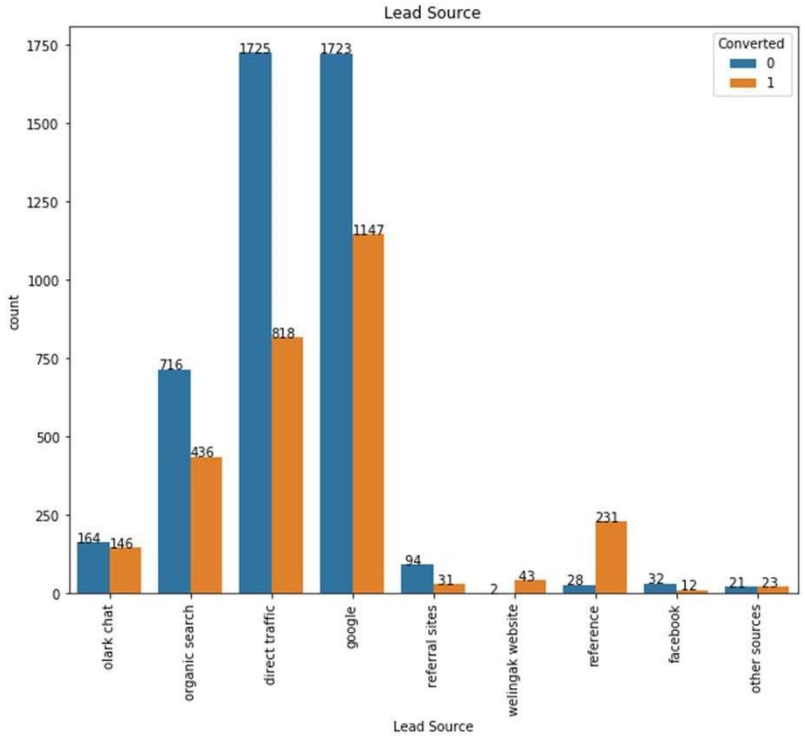

Converted
0
60.9%
39.1%
1

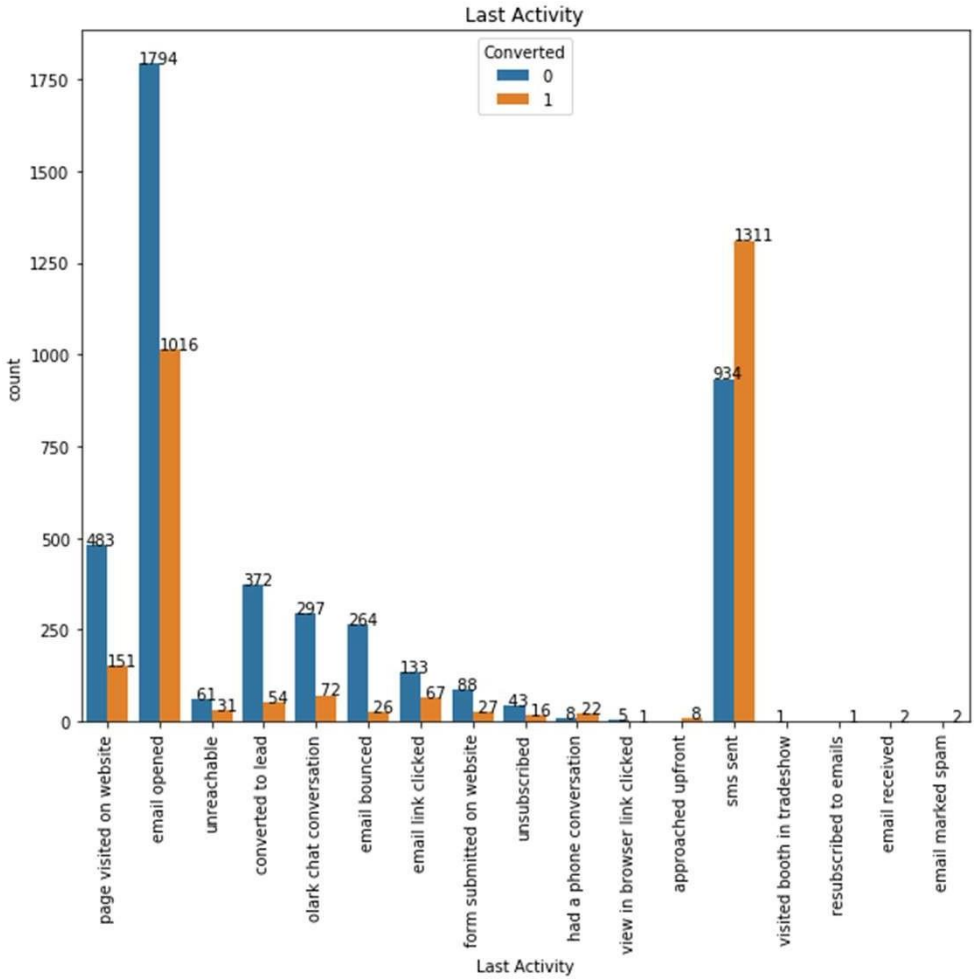

A free copy of Mastering The Interview
no
61.0%
39.0%
yes

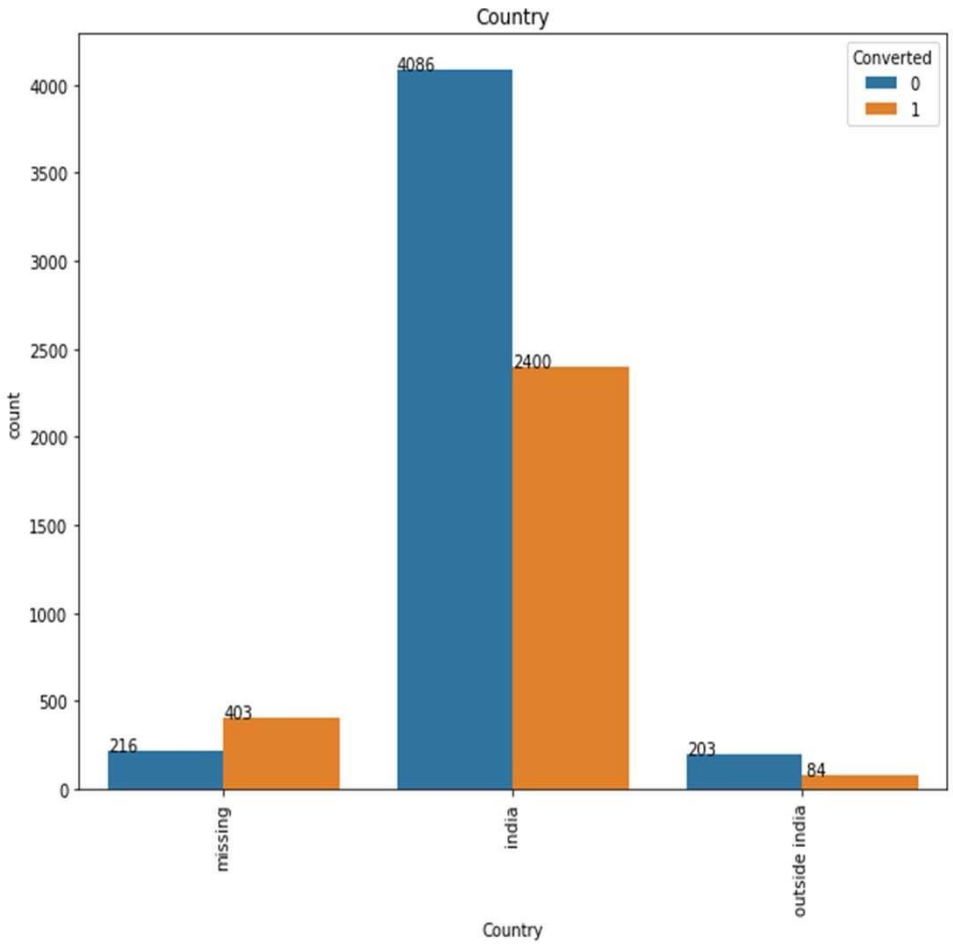

X Education Forums
no
100.0%
0.0%
yes

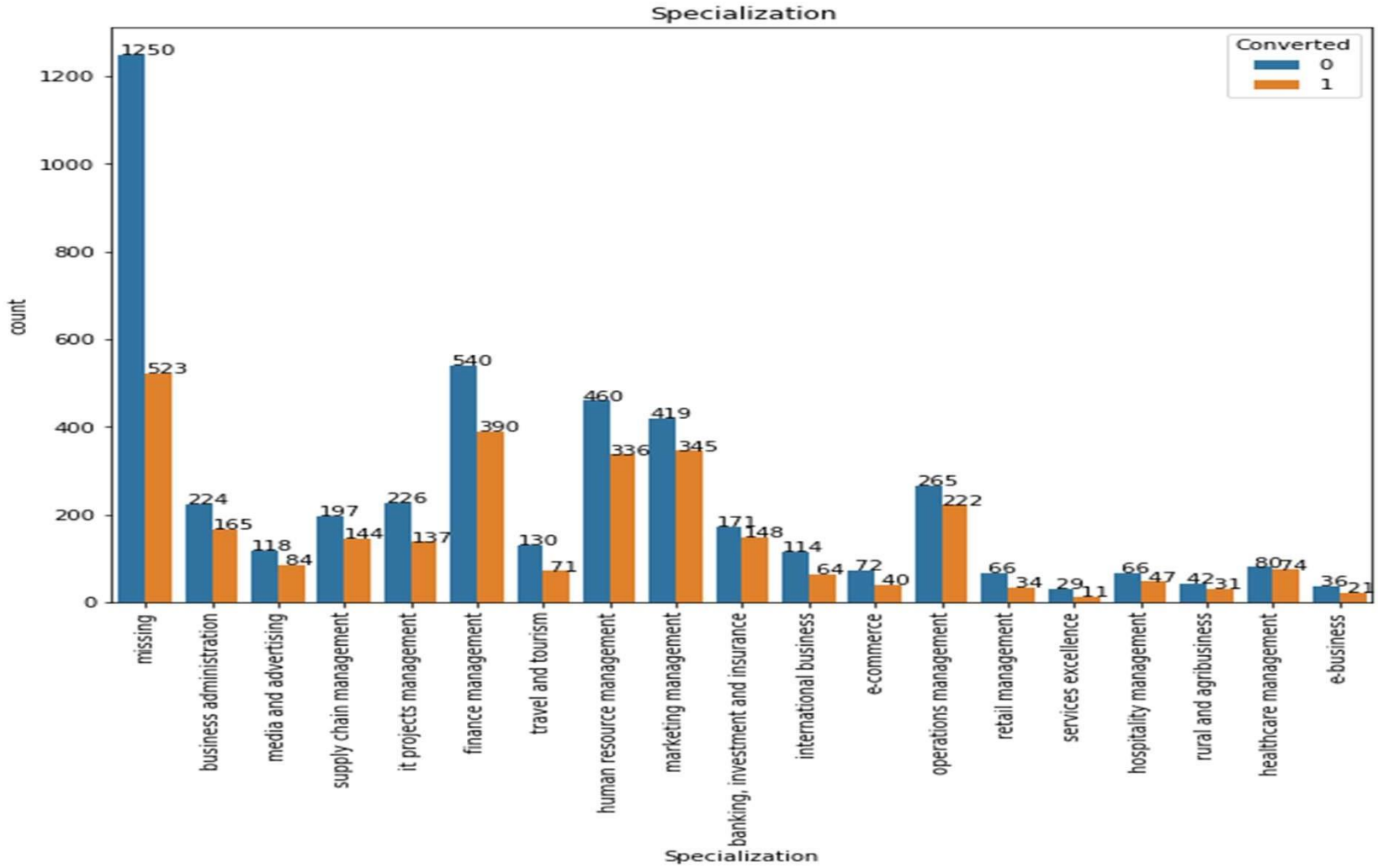
Categorical Variable Relation :

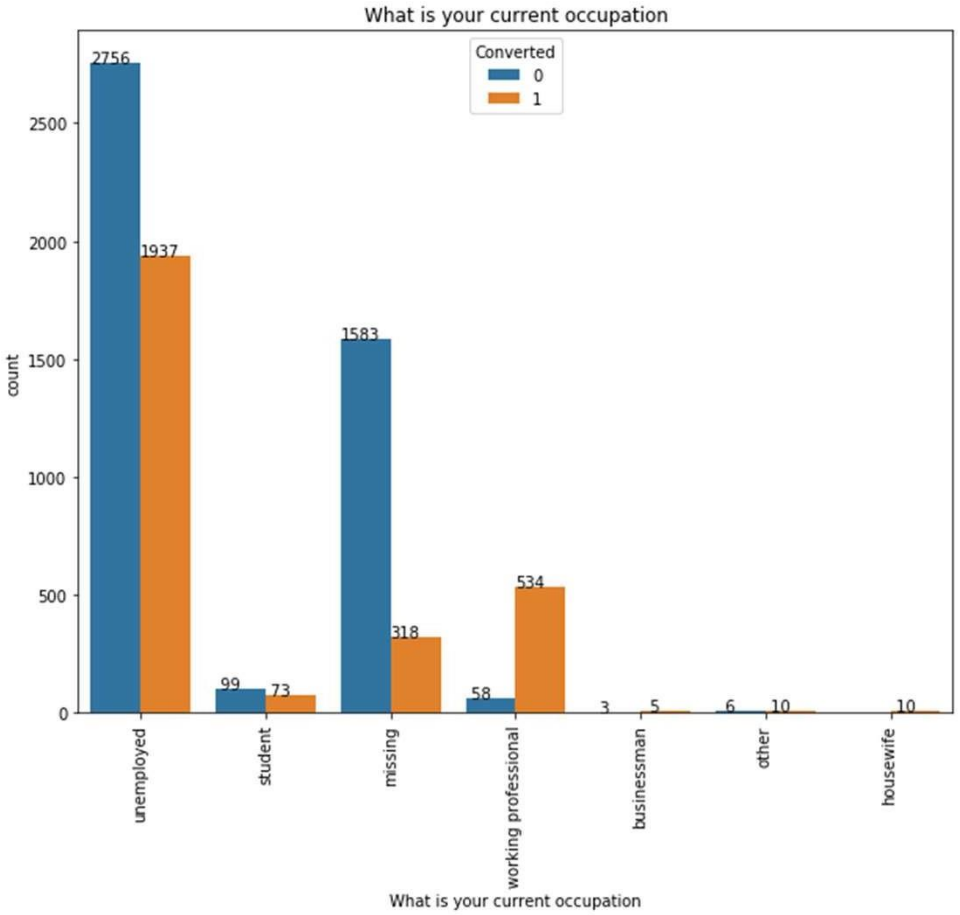


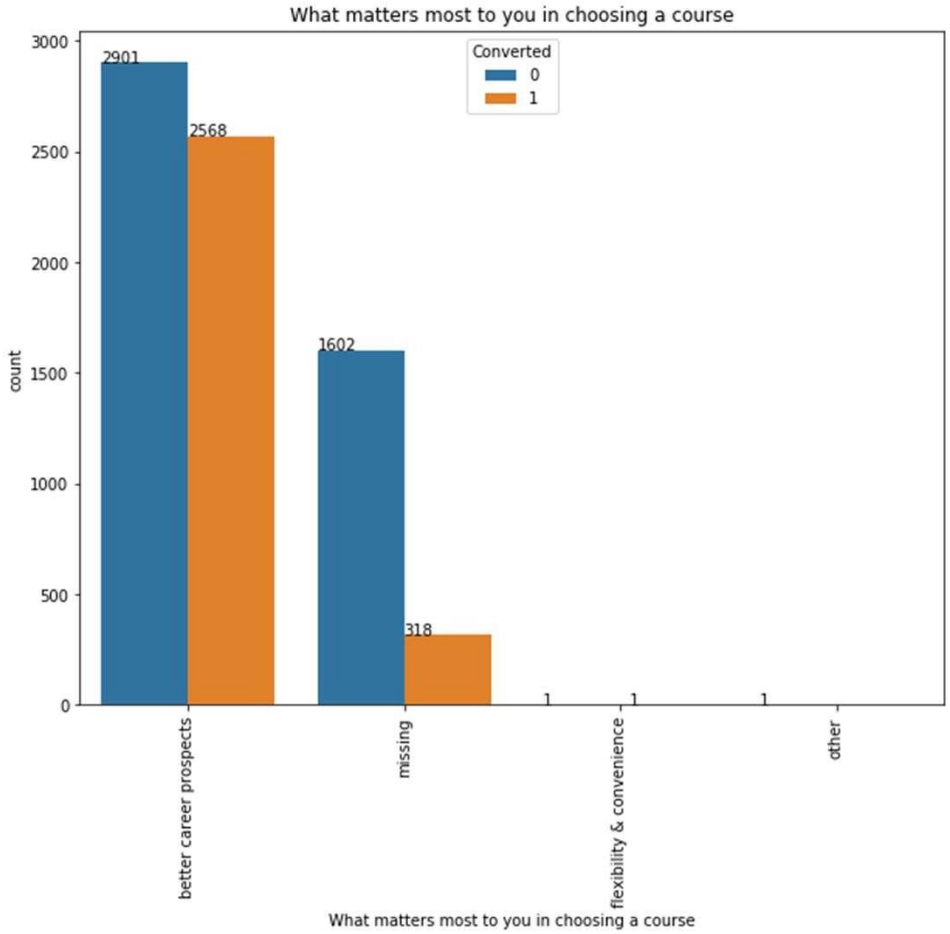



Last Activity
Converted
0
1
count
Last Activity
page visited on website
email opened
unreachable
converted to lead
olark chat conversation
email bounced
email link clicked
form submitted on website
unsubscribed
had a phone conversation
view in browser link clicked
approached upfront
sms sent
visited booth in tradeshow
resubscribed to emails
email received
email marked spam
483
151
1794
1016
61
31
372
54
297
72
264
26
133
67
88
27
43
16
8
22
5
1
8
934
1311
1
1
2
2

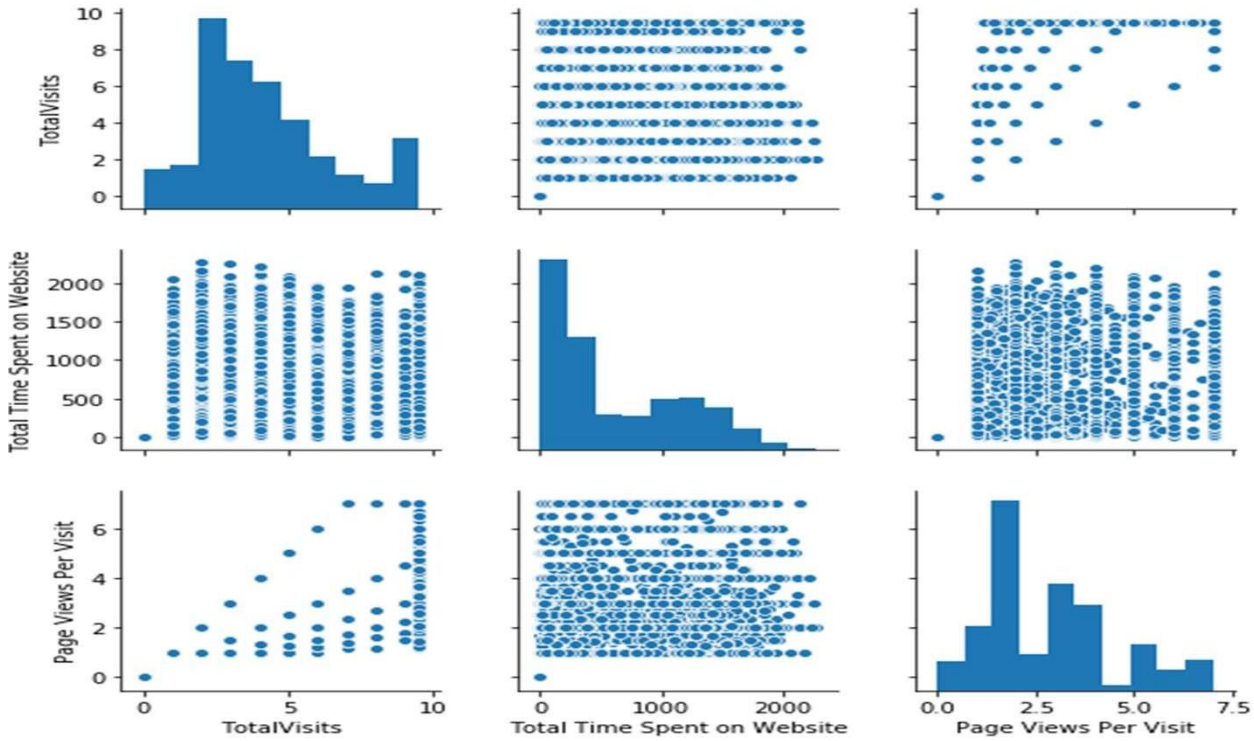

Country
Converted
0
1
count
Country
missing
india
outside india
216
403
4086
2400
203
84


Specialization
Converted
0
1
count
1250
523
224
165
118
84
197
144
226
137
540
390
130
71
460
336
419
345
171
148
114
64
72
40
265
222
66
34
29
11
66
47
42
31
80
74
36
21
missing
business administration
media and advertising
supply chain management
it projects management
finance management
travel and tourism
human resource management
marketing management
banking, investment and insurance
international business
e-commerce
operations management
retail management
services excellence
hospitality management
rural and agribusiness
healthcare management
e-business
Specialization


What is your current occupation
Converted
0
1
count
2756
1937
1583
318
99
73
58
534
3
5
6
10
10
unemployed
student
missing
working professional
businessman
other
housewife
What is your current occupation


What matters most to you in choosing a course
Converted
0
1
count
2901
2568
1602
318
1
1
1
better career prospects
missing
flexibility & convenience
other
What matters most to you in choosing a course

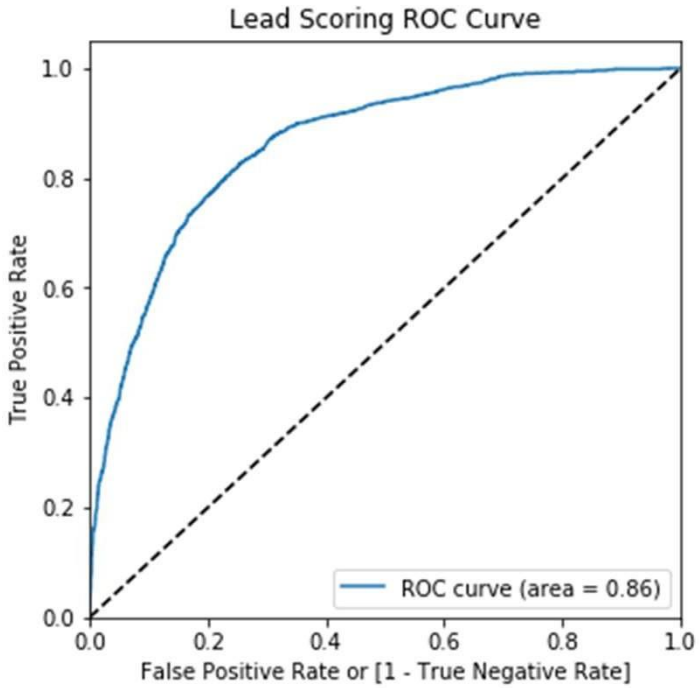
Multivariate Analysis:

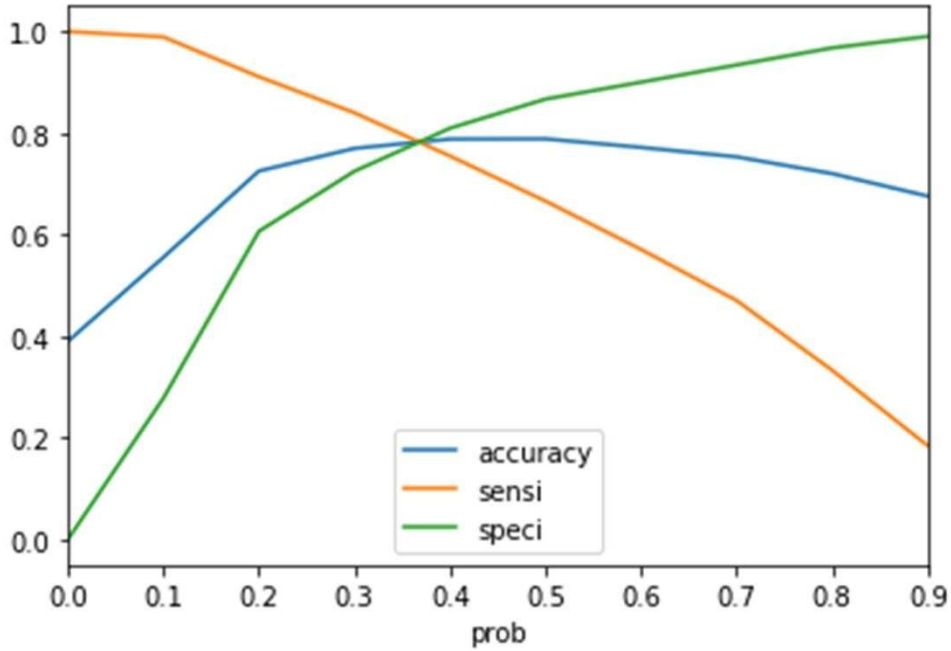


Model Building :

- Splitting the Data into Training and Testing Sets
- The first basic step for regression is performing a train-test split, we have chosen 70:30 ratio
- Use RFE for Feature Selection
- Running RFE with 15 variables as output
- Predictions on test data set
- Overall accuracy 80%

Optimize Cut Off (ROC Curve) :





Conclusion :

It was found that the variables that mattered the most in the potential leads are:

- The total time spend on the Website.
- Page Views Per Visit.
- Total Visits
- When the Last Notable Activity is “Had a phone conversation”
- When the last Activity is “ SMS Sent”
- When the Load Origin is Load add Format
- When the Leads Current Profession is “Working Professionals”.