Lead Score - Case Study

Problem Statement

An education company named X Education sells online courses to industry professionals. On any given day, many professionals who are interested in the courses land on their website and browse for courses.

The company markets its courses on several websites and search engines like Google. 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. Moreover, the company also gets leads through past referrals. Once these leads are acquired, employees from the sales team start making calls, writing emails, etc. Through this process, some of the leads get converted while most do not. The typical lead conversion rate at X education is around 30%.

There are a lot of leads generated in the initial stage, but only a few of them come out as paying customers. In the middle stage, you need to nurture the potential leads well (i.e. educating the leads about the product, constantly communicating etc.) in order to get a higher lead conversion.

X Education has appointed you to help them select the most promising leads, i.e. the leads that are most likely to convert into paying customers. The company requires you 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%.

Goals of the 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 whereas a lower score would mean that the lead is cold and will mostly not get converted.

All the outcomes and understandings are written in BLUE

```python
# Supress Warnings
import warnings
warnings.filterwarnings('ignore')
#Importing required packages
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns
```

1 : Loading and Cleaning Data

1.1 Import Data

```
# Loading the data using Pandas
df = pd.read_csv('C:\Data science Project\Lead+Scoring+Case+Study\Lead
Scoring Assignment\Leads.csv')
df
```

```
                               Prospect ID  Lead Number  \
0     7927b2df-8bba-4d29-b9a2-b6e0beafe620       660737
1     2a272436-5132-4136-86fa-dcc88c88f482       660728
2     8cc8c611-a219-4f35-ad23-fdfd2656bd8a       660727
3     0cc2df48-7cf4-4e39-9de9-19797f9b38cc       660719
4     3256f628-e534-4826-9d63-4a8b88782852       660681
5     2058ef08-2858-443e-a01f-a9237db2f5ce       660680
6     9fae7df4-169d-489b-afe4-0f3d752542ed       660673
7     20ef72a2-fb3b-45e0-924e-551c5fa59095       660664
8     cfa0128c-a0da-4656-9d47-0aa4e67bf690       660624
9     af465dfc-7204-4130-9e05-33231863c4b5       660616
10    2a369e35-ca95-4ca9-9e4f-9d27175aa320       660608
11    9bc8ce93-6144-49e0-9f9d-080fc980f83c       660570
12    8bf76a52-2478-476b-8618-1688e07874ad       660562
13    88867067-3750-4753-8d33-1c7d1db53b5e       660558
14    a8531c22-fcf1-48f8-a711-fb5abf98ad87       660553
15    25f4ac14-ff4b-4cd2-9c61-b44c85e19c8f       660547
16    3abb7c77-1634-4083-9a9f-861068220611       660540
17    e5c3beca-a0b6-4b3f-8c01-0919fb9ca3f2       660534
18    82cb5fb0-2d97-4a39-a630-ab5fe2e7f18c       660522
19    4512c16a-e96a-4459-b9ec-c7d8fe8c4880       660509
20    c4419c99-b002-408b-a6fd-fa100716592c       660479
21    fd71ab5b-53b8-4105-9960-efedc44962fa       660478
22    8fd38b83-5c32-4277-bcfb-499f34a01c56       660471
23    ecbc6e69-29a9-44bf-804a-13079ef301bc       660461
24    ecd117ca-375f-49ea-afd6-b52b84d00c69       660458
25    31c326f0-4a9b-43a6-9006-99d3830fbcae       660447
26    c494aca4-8c8e-4081-9784-41eb6346015e       660432
27    6d143c0e-abae-425f-a2c0-52c2946cbd45       660424
28    8247051c-f838-4a41-b39c-1f0b44c3d5e6       660423
29    b3455e2e-8236-478a-b1aa-666ad3381722       660410
...                                    ...          ...
9210  14ac6418-af18-4acd-b464-02f6e0fefa1c       579833
9211  8458b410-48fe-4bcd-aecf-5813b6006ee2       579832
9212  0c15052a-9f8a-47c4-9fc3-eb20c84ffd74       579830
9213  d4587acb-02d1-4c5e-9110-6032d829bac1       579822
9214  479a8b1c-d410-4220-a24f-854a376be43d       579808
9215  06334ac1-64a8-444c-92a7-117dcd26dea5       579802
9216  6da5be9f-3f34-4dc7-9e30-7c26d030372e       579799
9217  b8872c12-7534-498d-8f4a-e79a19516db1       579786
```

```
9218  eee466be-b98c-4126-9220-fc406093b9ce       579784
9219  9c970d5c-2748-4f61-90a6-eafd9ad5a242       579778
9220  679ab5f9-0f85-4f16-a903-821ecd82e731       579769
9221  b92509cd-7f4c-414e-a8af-eb9cf0c89da7       579767
9222  68e53bdc-b66d-48ef-8592-973a8a65377e       579764
9223  c55de92b-9295-40e1-90e8-a628c349c292       579755
9224  18930f11-41cd-42d1-96d7-34ac870174cb       579753
9225  787ab5f4-6f09-41c0-b083-55521ca23f8a       579744
9226  c3bb1471-53d5-4244-b2e5-4bbb543835c1       579735
9227  ac95586a-506a-4222-9967-17dfe9f82524       579728
9228  40d3b3cf-d939-49ff-bea5-60e8d4025104       579717
9229  5cfdd915-d5a0-4976-b38d-e5f72ec55526       579712
9230  d11c15b7-8056-45a6-8954-771c0d0495fe       579701
9231  4aeae36b-2b57-494f-bdab-dd58844286b4       579697
9232  2d0109e9-dfb2-4664-83de-c2ea75ec7516       579642
9233  3f715465-2546-47cd-afa8-8b8dc63b8b43       579622
9234  c0b25922-511f-4c56-852e-ced210a45447       579615
9235  19d6451e-fcd6-407c-b83b-48e1af805ea9       579564
9236  82a7005b-7196-4d56-95ce-a79f937a158d       579546
9237  aac550fe-a586-452d-8d3c-f1b62c94e02c       579545
9238  5330a7d1-2f2b-4df4-85d6-64ca2f6b95b9       579538
9239  571b5c8e-a5b2-4d57-8574-f2ffb06fdeff       579533

                  Lead Origin     Lead Source Do Not Email Do Not Call  \
0                         API      Olark Chat           No          No

1                         API  Organic Search           No          No

2     Landing Page Submission  Direct Traffic           No          No

3     Landing Page Submission  Direct Traffic           No          No

4     Landing Page Submission          Google           No          No

5                         API      Olark Chat           No          No

6     Landing Page Submission          Google           No          No

7                         API      Olark Chat           No          No

8     Landing Page Submission  Direct Traffic           No          No

9                         API          Google           No          No

10    Landing Page Submission  Organic Search           No          No

11    Landing Page Submission  Direct Traffic           No          No

12                        API  Organic Search           No          No
```

13	Landing Page Submission	Organic Search	No	No
14	Landing Page Submission	Direct Traffic	Yes	No
15	API	Organic Search	No	No
16	API	Olark Chat	No	No
17	API	Referral Sites	No	No
18	Landing Page Submission	Google	No	No
19	API	Organic Search	No	No
20	Landing Page Submission	Google	No	No
21	API	Google	No	No
22	Landing Page Submission	Google	No	No
23	Landing Page Submission	Google	No	No
24	API	Google	No	No
25	Landing Page Submission	Google	No	No
26	Landing Page Submission	Organic Search	No	No
27	Landing Page Submission	Google	No	No
28	Landing Page Submission	Direct Traffic	No	No
29	API	Google	No	No
...	...	...	...	...
9210	Landing Page Submission	Direct Traffic	No	No
9211	Landing Page Submission	Direct Traffic	No	No
9212	Landing Page Submission	Google	Yes	No
9213	Landing Page Submission	Direct Traffic	Yes	No
9214	API	Organic Search	No	No
9215	Landing Page Submission	Organic Search	No	No
9216	Landing Page Submission	Direct Traffic	Yes	No
9217	API	Olark Chat	No	No

```
9218  Landing Page Submission           Google           Yes            No
9219  Landing Page Submission  Direct Traffic             No            No
9220  Landing Page Submission  Direct Traffic             No            No
9221  Landing Page Submission           Google             No            No
9222                      API           Google             No            No
9223                      API  Organic Search             No            No
9224  Landing Page Submission           Google             No            No
9225  Landing Page Submission  Direct Traffic            Yes            No
9226                      API      Olark Chat             No            No
9227  Landing Page Submission           Google             No            No
9228  Landing Page Submission           Google             No            No
9229  Landing Page Submission  Organic Search             No            No
9230  Landing Page Submission           Google             No            No
9231  Landing Page Submission           Google             No            No
9232  Landing Page Submission  Direct Traffic             No            No
9233                      API  Direct Traffic             No            No
9234  Landing Page Submission  Direct Traffic             No            No
9235  Landing Page Submission  Direct Traffic            Yes            No
9236  Landing Page Submission  Direct Traffic             No            No
9237  Landing Page Submission  Direct Traffic            Yes            No
9238  Landing Page Submission           Google             No            No
9239  Landing Page Submission  Direct Traffic             No            No

      Converted  TotalVisits  Total Time Spent on Website  \
0             0          0.0                            0
1             0          5.0                          674
2             1          2.0                         1532
3             0          1.0                          305
4             1          2.0                         1428
```

5	0	0.0	0
6	1	2.0	1640
7	0	0.0	0
8	0	2.0	71
9	0	4.0	58
10	1	8.0	1351
11	1	8.0	1343
12	1	11.0	1538
13	0	5.0	170
14	0	1.0	481
15	1	6.0	1012
16	0	0.0	0
17	0	6.0	973
18	1	6.0	1688
19	0	3.0	98
20	0	1.0	233
21	0	4.0	377
22	1	1.0	1013
23	0	4.0	771
24	1	6.0	1137
25	1	3.0	1068
26	1	4.0	1000
27	1	6.0	1315
28	0	5.0	182
29	1	3.0	78
...	...	...	...
9210	1	4.0	927
9211	1	4.0	1112
9212	0	5.0	78
9213	0	5.0	234
9214	1	2.0	881
9215	0	8.0	397
9216	0	6.0	1679
9217	0	0.0	0
9218	0	1.0	149
9219	1	6.0	1389
9220	0	5.0	20
9221	0	4.0	1347
9222	0	6.0	228
9223	0	7.0	142
9224	0	4.0	455
9225	0	2.0	74
9226	0	0.0	0
9227	1	5.0	1283
9228	1	4.0	1944
9229	1	13.0	1226
9230	0	2.0	870
9231	1	8.0	1016
9232	0	2.0	1770

```
9233          1         13.0                             1409
9234          1          5.0                              210
9235          1          8.0                             1845
9236          0          2.0                              238
9237          0          2.0                              199
9238          1          3.0                              499
9239          1          6.0                             1279

      Page Views Per Visit           ...            Get updates on DM
Content  \
0                     0.00           ...
No
1                     2.50           ...
No
2                     2.00           ...
No
3                     1.00           ...
No
4                     1.00           ...
No
5                     0.00           ...
No
6                     2.00           ...
No
7                     0.00           ...
No
8                     2.00           ...
No
9                     4.00           ...
No
10                    8.00           ...
No
11                    2.67           ...
No
12                   11.00           ...
No
13                    5.00           ...
No
14                    1.00           ...
No
15                    6.00           ...
No
16                    0.00           ...
No
17                    6.00           ...
No
18                    3.00           ...
No
19                    3.00           ...
```

```
No
20                    1.00          ...
No
21                    1.33          ...
No
22                    1.00          ...
No
23                    4.00          ...
No
24                    1.50          ...
No
25                    3.00          ...
No
26                    2.00          ...
No
27                    6.00          ...
No
28                    5.00          ...
No
29                    3.00          ...
No
...                    ...          ...
...
9210                  4.00          ...
No
9211                  4.00          ...
No
9212                  5.00          ...
No
9213                  2.50          ...
No
9214                  2.00          ...
No
9215                  8.00          ...
No
9216                  6.00          ...
No
9217                  0.00          ...
No
9218                  1.00          ...
No
9219                  6.00          ...
No
9220                  2.50          ...
No
9221                  2.00          ...
No
9222                  6.00          ...
No
```

```
9223                       7.00            ...
No
9224                       4.00            ...
No
9225                       2.00            ...
No
9226                       0.00            ...
No
9227                       1.67            ...
No
9228                       2.00            ...
No
9229                       6.50            ...
No
9230                       2.00            ...
No
9231                       4.00            ...
No
9232                       2.00            ...
No
9233                       2.60            ...
No
9234                       2.50            ...
No
9235                       2.67            ...
No
9236                       2.00            ...
No
9237                       2.00            ...
No
9238                       3.00            ...
No
9239                       3.00            ...
No

        Lead Profile                City Asymmetrique Activity
Index  \
0             Select              Select                     02.Medium

1             Select              Select                     02.Medium

2     Potential Lead              Mumbai                     02.Medium

3             Select              Mumbai                     02.Medium

4             Select              Mumbai                     02.Medium

5                NaN                 NaN                       01.High

6     Potential Lead              Mumbai                     02.Medium
```

7	NaN	NaN	02.Medium
8	NaN	Thane & Outskirts	02.Medium
9	NaN	Mumbai	02.Medium
10	Select	Other Metro Cities	02.Medium
11	Select	Thane & Outskirts	02.Medium
12	Potential Lead	Select	01.High
13	Select	Thane & Outskirts	02.Medium
14	Select	Select	01.High
15	Select	Select	02.Medium
16	NaN	NaN	01.High
17	Select	Select	02.Medium
18	Select	Mumbai	02.Medium
19	Select	Select	02.Medium
20	Select	Mumbai	02.Medium
21	Potential Lead	Select	02.Medium
22	Potential Lead	Mumbai	02.Medium
23	Select	Mumbai	02.Medium
24	Potential Lead	Mumbai	02.Medium
25	Select	Mumbai	02.Medium
26	Potential Lead	Other Cities	03.Low
27	Potential Lead	Mumbai	02.Medium
28	Select	Mumbai	02.Medium
29	Potential Lead	Mumbai	01.High
...	...	...	...
9210	Potential Lead	Mumbai	02.Medium
9211	Other Leads	Mumbai	02.Medium

9212	Potential Lead	Mumbai	02.Medium
9213	NaN	Mumbai	01.High
9214	NaN	NaN	02.Medium
9215	NaN	Thane & Outskirts	02.Medium
9216	Other Leads	Mumbai	01.High
9217	Potential Lead	Select	02.Medium
9218	NaN	Mumbai	02.Medium
9219	Potential Lead	Other Metro Cities	02.Medium
9220	Potential Lead	Thane & Outskirts	02.Medium
9221	Select	Mumbai	NaN
9222	Potential Lead	Other Cities	02.Medium
9223	Potential Lead	Mumbai	02.Medium
9224	Potential Lead	Mumbai	03.Low
9225	Potential Lead	Mumbai	03.Low
9226	Select	Select	01.High
9227	Potential Lead	Mumbai	02.Medium
9228	Select	Mumbai	NaN
9229	Potential Lead	Mumbai	02.Medium
9230	Potential Lead	Mumbai	02.Medium
9231	Potential Lead	Mumbai	02.Medium
9232	Potential Lead	Mumbai	02.Medium
9233	Select	Select	NaN
9234	Potential Lead	Mumbai	02.Medium
9235	Potential Lead	Mumbai	02.Medium
9236	Potential Lead	Mumbai	02.Medium
9237	Potential Lead	Mumbai	02.Medium

```
9238             NaN  Other Metro Cities                       02.Medium

9239  Potential Lead       Other Cities                       02.Medium


     Asymmetrique Profile Index Asymmetrique Activity Score  \
0                     02.Medium                        15.0
1                     02.Medium                        15.0
2                       01.High                        14.0
3                       01.High                        13.0
4                       01.High                        15.0
5                     02.Medium                        17.0
6                       01.High                        14.0
7                     02.Medium                        15.0
8                     02.Medium                        14.0
9                     02.Medium                        13.0
10                    02.Medium                        15.0
11                      01.High                        14.0
12                    02.Medium                        16.0
13                      01.High                        14.0
14                      01.High                        16.0
15                    02.Medium                        14.0
16                    02.Medium                        17.0
17                    02.Medium                        13.0
18                      01.High                        15.0
19                    02.Medium                        14.0
20                      01.High                        13.0
21                    02.Medium                        15.0
22                      01.High                        15.0
23                      01.High                        14.0
24                      01.High                        14.0
25                    02.Medium                        14.0
26                      01.High                        11.0
27                      01.High                        15.0
28                      01.High                        13.0
29                      01.High                        16.0
...                         ...                         ...
9210                    01.High                        14.0
9211                  02.Medium                        15.0
9212                    01.High                        13.0
9213                  02.Medium                        17.0
9214                  02.Medium                        15.0
9215                    01.High                        13.0
9216                    01.High                        16.0
9217                  02.Medium                        15.0
9218                    01.High                        13.0
9219                    01.High                        15.0
9220                    01.High                        13.0
9221                        NaN                         NaN
```

```
9222                        02.Medium                           15.0
9223                          01.High                           13.0
9224                          01.High                           12.0
9225                          01.High                           12.0
9226                        02.Medium                           16.0
9227                          01.High                           15.0
9228                              NaN                            NaN
9229                          01.High                           15.0
9230                          01.High                           13.0
9231                          01.High                           15.0
9232                          01.High                           14.0
9233                              NaN                            NaN
9234                          01.High                           14.0
9235                          01.High                           15.0
9236                          01.High                           14.0
9237                          01.High                           13.0
9238                        02.Medium                           15.0
9239                          01.High                           15.0

     Asymmetrique Profile Score I agree to pay the amount through
cheque  \
0                           15.0
No
1                           15.0
No
2                           20.0
No
3                           17.0
No
4                           18.0
No
5                           15.0
No
6                           20.0
No
7                           15.0
No
8                           14.0
No
9                           16.0
No
10                          14.0
No
11                          17.0
No
12                          16.0
No
13                          17.0
No
```

```
14                                        17.0
No
15                                        15.0
No
16                                        15.0
No
17                                        13.0
No
18                                        18.0
No
19                                        15.0
No
20                                        17.0
No
21                                        16.0
No
22                                        20.0
No
23                                        18.0
No
24                                        18.0
No
25                                        16.0
No
26                                        18.0
No
27                                        19.0
No
28                                        18.0
No
29                                        18.0
No
...                                        ...                                                    .
..
9210                                      20.0
No
9211                                      15.0
No
9212                                      20.0
No
9213                                      15.0
No
9214                                      15.0
No
9215                                      17.0
No
9216                                      18.0
No
9217                                      16.0
```

```
No
9218                           18.0
No
9219                           18.0
No
9220                           19.0
No
9221                            NaN
No
9222                           16.0
No
9223                           18.0
No
9224                           20.0
No
9225                           20.0
No
9226                           15.0
No
9227                           20.0
No
9228                            NaN
No
9229                           20.0
No
9230                           20.0
No
9231                           20.0
No
9232                           20.0
No
9233                            NaN
No
9234                           20.0
No
9235                           17.0
No
9236                           19.0
No
9237                           20.0
No
9238                           16.0
No
9239                           18.0
No

     A free copy of Mastering The Interview    Last Notable Activity
0                                         No                Modified
1                                         No            Email Opened
```

2	Yes	Email Opened
3	No	Modified
4	No	Modified
5	No	Modified
6	No	Modified
7	No	Modified
8	Yes	Email Opened
9	No	Email Opened
10	Yes	Email Opened
11	Yes	Page Visited on Website
12	No	Modified
13	Yes	Email Opened
14	No	Email Bounced
15	No	Email Opened
16	No	Modified
17	No	Modified
18	No	Page Visited on Website
19	No	Modified
20	No	Modified
21	No	Modified
22	No	Modified
23	No	Email Link Clicked
24	Yes	Email Opened
25	No	Modified
26	Yes	Email Opened
27	No	Email Opened
28	No	Email Opened
29	No	Unreachable
...	...	...
9210	No	Modified
9211	No	SMS Sent
9212	Yes	Unsubscribed
9213	No	Modified
9214	No	SMS Sent
9215	Yes	Email Opened
9216	Yes	Modified
9217	No	SMS Sent
9218	No	Modified
9219	Yes	Email Opened
9220	Yes	Modified
9221	Yes	SMS Sent
9222	No	Modified
9223	Yes	Modified
9224	No	Modified
9225	Yes	Modified
9226	No	Modified
9227	No	Email Opened
9228	Yes	Modified
9229	Yes	Modified

```
9230                                                     No          Email Opened
9231                                                     No          Email Opened
9232                                                    Yes              SMS Sent
9233                                                     No              SMS Sent
9234                                                     No              Modified
9235                                                     No     Email Marked Spam
9236                                                    Yes              SMS Sent
9237                                                    Yes              SMS Sent
9238                                                     No              SMS Sent
9239                                                    Yes              Modified

[9240 rows x 37 columns]
```

1.2 Inspect the dataframe

This helps to give a good idea of the dataframes.

```
# The .info() code gives almost the entire information that needs to
be inspected, so let's start from there
df.info()

<class 'pandas.core.frame.DataFrame'>
RangeIndex: 9240 entries, 0 to 9239
Data columns (total 37 columns):
Prospect ID                                      9240 non-null object
Lead Number                                      9240 non-null int64
Lead Origin                                      9240 non-null object
Lead Source                                      9204 non-null object
Do Not Email                                     9240 non-null object
Do Not Call                                      9240 non-null object
Converted                                        9240 non-null int64
TotalVisits                                      9103 non-null float64
Total Time Spent on Website                      9240 non-null int64
Page Views Per Visit                             9103 non-null float64
Last Activity                                    9137 non-null object
Country                                          6779 non-null object
Specialization                                   7802 non-null object
How did you hear about X Education               7033 non-null object
What is your current occupation                  6550 non-null object
What matters most to you in choosing a course    6531 non-null object
Search                                           9240 non-null object
Magazine                                         9240 non-null object
Newspaper Article                                9240 non-null object
X Education Forums                               9240 non-null object
Newspaper                                        9240 non-null object
Digital Advertisement                            9240 non-null object
Through Recommendations                          9240 non-null object
Receive More Updates About Our Courses           9240 non-null object
Tags                                             5887 non-null object
```

```
Lead Quality                                        4473 non-null object
Update me on Supply Chain Content                   9240 non-null object
Get updates on DM Content                           9240 non-null object
Lead Profile                                        6531 non-null object
City                                                7820 non-null object
Asymmetrique Activity Index                         5022 non-null object
Asymmetrique Profile Index                          5022 non-null object
Asymmetrique Activity Score                         5022 non-null float64
Asymmetrique Profile Score                          5022 non-null float64
I agree to pay the amount through cheque            9240 non-null object
A free copy of Mastering The Interview              9240 non-null object
Last Notable Activity                               9240 non-null object
dtypes: float64(4), int64(3), object(30)
memory usage: 2.6+ MB
```

```
#To get the idea of how the table looks like we can use .head()
or .tail() command
df.head()
```

```
                            Prospect ID  Lead Number              Lead
Origin  \
0  7927b2df-8bba-4d29-b9a2-b6e0beafe620       660737
API
1  2a272436-5132-4136-86fa-dcc88c88f482       660728
API
2  8cc8c611-a219-4f35-ad23-fdfd2656bd8a       660727  Landing Page
Submission
3  0cc2df48-7cf4-4e39-9de9-19797f9b38cc       660719  Landing Page
Submission
4  3256f628-e534-4826-9d63-4a8b88782852       660681  Landing Page
Submission

      Lead Source Do Not Email Do Not Call  Converted  TotalVisits  \
0      Olark Chat           No          No          0          0.0
1  Organic Search           No          No          0          5.0
2  Direct Traffic           No          No          1          2.0
3  Direct Traffic           No          No          0          1.0
4          Google           No          No          1          2.0

   Total Time Spent on Website  Page Views Per Visit          ...
\
0                            0                   0.0          ...

1                          674                   2.5          ...

2                         1532                   2.0          ...

3                          305                   1.0          ...

4                         1428                   1.0          ...
```

```
  Get updates on DM Content    Lead Profile    City  \
0                        No          Select  Select
1                        No          Select  Select
2                        No  Potential Lead  Mumbai
3                        No          Select  Mumbai
4                        No          Select  Mumbai

  Asymmetrique Activity Index Asymmetrique Profile Index  \
0                   02.Medium                  02.Medium
1                   02.Medium                  02.Medium
2                   02.Medium                    01.High
3                   02.Medium                    01.High
4                   02.Medium                    01.High

  Asymmetrique Activity Score Asymmetrique Profile Score  \
0                        15.0                       15.0
1                        15.0                       15.0
2                        14.0                       20.0
3                        13.0                       17.0
4                        15.0                       18.0

  I agree to pay the amount through cheque  \
0                                       No
1                                       No
2                                       No
3                                       No
4                                       No

  A free copy of Mastering The Interview Last Notable Activity
0                                     No             Modified
1                                     No         Email Opened
2                                    Yes         Email Opened
3                                     No             Modified
4                                     No             Modified

[5 rows x 37 columns]
```

```
# The .shape code gives the no. of rows and columns
df.shape
```

```
(9240, 37)
```

```
#To get an idea of the numeric values, use .describe()
df.describe()
```

```
         Lead Number    Converted  TotalVisits  Total Time Spent on
Website  \
count    9240.000000  9240.000000  9103.000000
9240.000000
```

```
mean   617188.435606     0.385390     3.445238
487.698268
std     23405.995698     0.486714     4.854853
548.021466
min    579533.000000     0.000000     0.000000
0.000000
25%    596484.500000     0.000000     1.000000
12.000000
50%    615479.000000     0.000000     3.000000
248.000000
75%    637387.250000     1.000000     5.000000
936.000000
max    660737.000000     1.000000   251.000000
2272.000000

       Page Views Per Visit  Asymmetrique Activity Score  \
count           9103.000000                  5022.000000
mean               2.362820                    14.306252
std                2.161418                     1.386694
min                0.000000                     7.000000
25%                1.000000                    14.000000
50%                2.000000                    14.000000
75%                3.000000                    15.000000
max               55.000000                    18.000000

       Asymmetrique Profile Score
count                 5022.000000
mean                    16.344883
std                      1.811395
min                     11.000000
25%                     15.000000
50%                     16.000000
75%                     18.000000
max                     20.000000
```

1.3 Cleaning the dataframe

```
# Converting all the values to lower case
df = df.applymap(lambda s:s.lower() if type(s) == str else s)

# Replacing 'Select' with NaN (Since it means no option is selected)
df = df.replace('select',np.nan)

# Checking if there are columns with one unique value since it won't
affect our analysis
df.nunique()

Prospect ID                                               9240
Lead Number                                               9240
Lead Origin                                                  5
```

```
Lead Source                                        20
Do Not Email                                        2
Do Not Call                                         2
Converted                                           2
TotalVisits                                        41
Total Time Spent on Website                      1731
Page Views Per Visit                              114
Last Activity                                      17
Country                                            38
Specialization                                     18
How did you hear about X Education                  9
What is your current occupation                     6
What matters most to you in choosing a course       3
Search                                              2
Magazine                                            1
Newspaper Article                                   2
X Education Forums                                  2
Newspaper                                           2
Digital Advertisement                               2
Through Recommendations                             2
Receive More Updates About Our Courses              1
Tags                                               26
Lead Quality                                        5
Update me on Supply Chain Content                   1
Get updates on DM Content                           1
Lead Profile                                        5
City                                                6
Asymmetrique Activity Index                         3
Asymmetrique Profile Index                          3
Asymmetrique Activity Score                        12
Asymmetrique Profile Score                         10
I agree to pay the amount through cheque            1
A free copy of Mastering The Interview              2
Last Notable Activity                              16
dtype: int64
```

```python
# Dropping unique valued columns
df1= df.drop(['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'],axis=1)

# Checking the percentage of missing values
round(100*(df1.isnull().sum()/len(df1.index)), 2)
```

```
Prospect ID                                      0.00
Lead Number                                      0.00
Lead Origin                                      0.00
Lead Source                                      0.39
Do Not Email                                     0.00
Do Not Call                                      0.00
```

```
Converted                                         0.00
TotalVisits                                       1.48
Total Time Spent on Website                       0.00
Page Views Per Visit                              1.48
Last Activity                                     1.11
Country                                          26.63
Specialization                                   36.58
How did you hear about X Education               78.46
What is your current occupation                  29.11
What matters most to you in choosing a course    29.32
Search                                            0.00
Newspaper Article                                 0.00
X Education Forums                                0.00
Newspaper                                         0.00
Digital Advertisement                             0.00
Through Recommendations                           0.00
Tags                                             36.29
Lead Quality                                     51.59
Lead Profile                                     74.19
City                                             39.71
Asymmetrique Activity Index                      45.65
Asymmetrique Profile Index                       45.65
Asymmetrique Activity Score                      45.65
Asymmetrique Profile Score                       45.65
A free copy of Mastering The Interview            0.00
Last Notable Activity                             0.00
dtype: float64
```

```
# Removing all the columns that are no required and have 35% null values
df2 = df1.drop(['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'],axis=1)
df2.head()
```

```
                            Prospect ID              Lead Origin  \
0  7927b2df-8bba-4d29-b9a2-b6e0beafe620                      api   
1  2a272436-5132-4136-86fa-dcc88c88f482                      api   
2  8cc8c611-a219-4f35-ad23-fdfd2656bd8a  landing page submission   
3  0cc2df48-7cf4-4e39-9de9-19797f9b38cc  landing page submission   
4  3256f628-e534-4826-9d63-4a8b88782852  landing page submission   

      Lead Source Do Not Email Do Not Call  Converted  TotalVisits  \
0      olark chat           no          no          0          0.0   
1  organic search           no          no          0          5.0   
2  direct traffic           no          no          1          2.0   
3  direct traffic           no          no          0          1.0   
4          google           no          no          1          2.0   
```

```
   Total Time Spent on Website  Page Views Per Visit            Last
Activity  \
0                            0                   0.0  page visited on
website
1                          674                   2.5             email
opened
2                         1532                   2.0             email
opened
3                          305                   1.0
unreachable
4                         1428                   1.0       converted
to lead

           ...          What is your current occupation  \
0          ...                               unemployed
1          ...                               unemployed
2          ...                                  student
3          ...                               unemployed
4          ...                               unemployed

  What matters most to you in choosing a course Search Newspaper
Article  \
0                       better career prospects     no
no
1                       better career prospects     no
no
2                       better career prospects     no
no
3                       better career prospects     no
no
4                       better career prospects     no
no

  X Education Forums Newspaper Digital Advertisement Through
Recommendations  \
0                 no        no                    no
no
1                 no        no                    no
no
2                 no        no                    no
no
3                 no        no                    no
no
4                 no        no                    no
no

  A free copy of Mastering The Interview Last Notable Activity
0                                     no              modified
1                                     no          email opened
2                                    yes          email opened
```

```
3                                                    no                modified
4                                                    no                modified

[5 rows x 22 columns]

# Rechecking the percentage of missing values
round(100*(df2.isnull().sum()/len(df2.index)), 2)

Prospect ID                                       0.00
Lead Origin                                       0.00
Lead Source                                       0.39
Do Not Email                                      0.00
Do Not Call                                       0.00
Converted                                         0.00
TotalVisits                                       1.48
Total Time Spent on Website                       0.00
Page Views Per Visit                              1.48
Last Activity                                     1.11
Country                                          26.63
Specialization                                   36.58
What is your current occupation                  29.11
What matters most to you in choosing a course    29.32
Search                                            0.00
Newspaper Article                                 0.00
X Education Forums                                0.00
Newspaper                                         0.00
Digital Advertisement                             0.00
Through Recommendations                           0.00
A free copy of Mastering The Interview            0.00
Last Notable Activity                             0.00
dtype: float64
```

There is a huge value of null variables in 4 columns as seen above. But removing the rows with the null value will cost us a lot of data and they are important columns. So, instead we are going to replace the NaN values with 'not provided'. This way we have all the data and almost no null values. In case these come up in the model, it will be of no use and we can drop it off then.

```
df2['Specialization'] = df2['Specialization'].fillna('not provided')
df2['What matters most to you in choosing a course'] = df2['What matters most to you in choosing a course'].fillna('not provided')
df2['Country'] = df2['Country'].fillna('not provided')
df2['What is your current occupation'] = df2['What is your current occupation'].fillna('not provided')
df2.info()

<class 'pandas.core.frame.DataFrame'>
RangeIndex: 9240 entries, 0 to 9239
Data columns (total 22 columns):
Prospect ID                                      9240 non-null object
```

```
Lead Origin                                      9240 non-null object
Lead Source                                      9204 non-null object
Do Not Email                                     9240 non-null object
Do Not Call                                      9240 non-null object
Converted                                        9240 non-null int64
TotalVisits                                      9103 non-null float64
Total Time Spent on Website                      9240 non-null int64
Page Views Per Visit                             9103 non-null float64
Last Activity                                    9137 non-null object
Country                                          9240 non-null object
Specialization                                   9240 non-null object
What is your current occupation                  9240 non-null object
What matters most to you in choosing a course    9240 non-null object
Search                                           9240 non-null object
Newspaper Article                                9240 non-null object
X Education Forums                               9240 non-null object
Newspaper                                        9240 non-null object
Digital Advertisement                            9240 non-null object
Through Recommendations                          9240 non-null object
A free copy of Mastering The Interview           9240 non-null object
Last Notable Activity                            9240 non-null object
dtypes: float64(2), int64(2), object(18)
memory usage: 1.6+ MB
```

```
# Rechecking the percentage of missing values
round(100*(df2.isnull().sum()/len(df2.index)), 2)
```

```
Prospect ID                                      0.00
Lead Origin                                      0.00
Lead Source                                      0.39
Do Not Email                                     0.00
Do Not Call                                      0.00
Converted                                        0.00
TotalVisits                                      1.48
Total Time Spent on Website                      0.00
Page Views Per Visit                             1.48
Last Activity                                    1.11
Country                                          0.00
Specialization                                   0.00
What is your current occupation                  0.00
What matters most to you in choosing a course    0.00
Search                                           0.00
Newspaper Article                                0.00
X Education Forums                               0.00
Newspaper                                        0.00
Digital Advertisement                            0.00
Through Recommendations                          0.00
A free copy of Mastering The Interview           0.00
Last Notable Activity                            0.00
dtype: float64
```

```python
df2["Country"].value_counts()
```

```
india                     6492
not provided              2461
united states               69
united arab emirates        53
singapore                   24
saudi arabia                21
united kingdom              15
australia                   13
qatar                       10
bahrain                      7
hong kong                    7
france                       6
oman                         6
unknown                      5
nigeria                      4
south africa                 4
kuwait                       4
germany                      4
canada                       4
sweden                       3
bangladesh                   2
belgium                      2
philippines                  2
ghana                        2
netherlands                  2
italy                        2
china                        2
asia/pacific region          2
uganda                       2
tanzania                     1
denmark                      1
switzerland                  1
malaysia                     1
russia                       1
sri lanka                    1
vietnam                      1
kenya                        1
liberia                      1
indonesia                    1
Name: Country, dtype: int64
```

```python
def slots(x):
    category = ""
    if x == "india":
        category = "india"
    elif x == "not provided":
        category = "not provided"
    else:
```

```
        category = "outside india"
    return category

df2['Country'] = df2.apply(lambda x:slots(x['Country']), axis = 1)
df2['Country'].value_counts()
```

```
india            6492
not provided     2461
outside india     287
Name: Country, dtype: int64
```

```
# Rechecking the percentage of missing values
round(100*(df2.isnull().sum()/len(df2.index)), 2)
```

```
Prospect ID                                      0.00
Lead Origin                                      0.00
Lead Source                                      0.39
Do Not Email                                     0.00
Do Not Call                                      0.00
Converted                                        0.00
TotalVisits                                      1.48
Total Time Spent on Website                      0.00
Page Views Per Visit                             1.48
Last Activity                                    1.11
Country                                          0.00
Specialization                                   0.00
What is your current occupation                  0.00
What matters most to you in choosing a course    0.00
Search                                           0.00
Newspaper Article                                0.00
X Education Forums                               0.00
Newspaper                                        0.00
Digital Advertisement                            0.00
Through Recommendations                          0.00
A free copy of Mastering The Interview           0.00
Last Notable Activity                            0.00
dtype: float64
```

```
# Checking the percent of lose if the null values are removed
round(100*(sum(df2.isnull().sum(axis=1) > 1)/df2.shape[0]),2)
```

```
1.48
```

```
df3 = df2[df2.isnull().sum(axis=1) <1]
```

```
# Code for checking number of rows left in percent
round(100*(df3.shape[0])/(df.shape[0]),2)
```

```
98.2
```

```python
# Rechecking the percentage of missing values
round(100*(df3.isnull().sum()/len(df3.index)), 2)
```

```
Prospect ID                                      0.0
Lead Origin                                      0.0
Lead Source                                      0.0
Do Not Email                                     0.0
Do Not Call                                      0.0
Converted                                        0.0
TotalVisits                                      0.0
Total Time Spent on Website                      0.0
Page Views Per Visit                             0.0
Last Activity                                    0.0
Country                                          0.0
Specialization                                   0.0
What is your current occupation                  0.0
What matters most to you in choosing a course    0.0
Search                                           0.0
Newspaper Article                                0.0
X Education Forums                               0.0
Newspaper                                        0.0
Digital Advertisement                            0.0
Through Recommendations                          0.0
A free copy of Mastering The Interview           0.0
Last Notable Activity                            0.0
dtype: float64
```

```python
# To familiarize all the categorical values
for column in df3:
    print(df3[column].astype('category').value_counts())

print('----------------------------------------------------------------------------------------')
```

```
fffb0e5e-9f92-4017-9f42-781a69da4154    1
539366d9-f633-455a-99e4-dbc5907db28e    1
53ac14bd-2bb2-4315-a21c-94562d1b6b2d    1
53aabd84-5dcc-4299-bbe3-62f3764b07b1    1
539ffa32-1be7-4fe1-b04c-faf1bab763cf    1
539eb309-df36-4a89-ac58-6d3651393910    1
5398e7ff-74db-4074-89fb-4fd9a603f521    1
53953744-234a-4cb9-9af4-bcc47eb472f4    1
5390c5fe-b12c-4f6e-ae92-908672abb0a1    1
53dbb914-71e7-458a-9749-cfb4d655eac2    1
5379ee79-64b7-44f8-8c56-0e1ca2d5b887    1
537963c8-22d9-459d-8aae-ddac40580ffb    1
53744d5a-0483-42c0-80b0-8990a4d2356d    1
53715ab1-2106-4c4e-8493-81cc465eb9ce    1
536cdc6b-f4c1-449d-bfd8-9ef0ac912dbb    1
53690d88-52f0-4ce5-b6b8-a13570a6db35    1
```

```
53c4e210-3344-4737-813f-74ef9a747ab6    1
53dd16bd-8201-448d-8e20-97de1cf44a7f    1
5464e56f-d39b-49a4-881c-8c6f75f2bbc7    1
54170a0f-0470-4612-b284-3ea12d3a9ea0    1
543892e8-5b9a-4552-99b9-87d57f40552a    1
5434ccf3-9de6-4c72-8dd6-66c2829d0ee2    1
542a0891-2e52-40ba-ab42-e468b9636322    1
54238b21-65ce-4304-98c6-0f8a6b9671e3    1
5420238f-2224-4472-8041-d127c8a5533f    1
5418151f-a055-4e26-b56f-6f1726638b68    1
541325bd-15bb-4b52-8ad9-3fdf3cb1dd55    1
53e64fef-c5c6-4d03-b07a-8ccde69a6218    1
54113bf6-465b-4f6c-b0ee-2a582d37323e    1
540e2e23-517c-4470-b163-6ad9e89b8890    1
                                       ..
aa503b9c-f853-497f-a1cc-97d6b13312d1    1
aa4f0ba5-5985-469f-8cd7-98f7b20d27ea    1
aa4180a5-84f1-4e67-8d90-0c8403070a59    1
aa405742-17ac-4c65-b19e-ab91c241cc53    1
aa27a0af-eeab-4007-a770-fa8a93fa53c8    1
aabadcb8-fe4f-4456-b3b5-16e937cef311    1
aa1edcad-f74f-426c-881a-5bbaa5ce717d    1
aa02cd65-92f9-447c-8cc2-44b7b6f817fe    1
a9fab024-c486-4a99-a05d-aba8c6252dc8    1
a9f12b1c-c158-4347-a695-9565a947fd55    1
a9ecd64e-dc3e-4058-8637-fefd2cd72768    1
a9ea3237-c91c-4a93-b7e8-f6550511bff1    1
aa5fb614-bf24-408d-9c89-e97b91d9479d    1
aa5ff9e9-bd5c-4a6e-bc03-e19552725635    1
aa613715-ff22-429d-9fbb-92da56b827aa    1
aa6fc8ca-ae09-4c9e-bae0-0427f5f56a70    1
aa708f29-9cb7-4959-a251-8aff9613b024    1
aa7e4871-e2f5-4c6a-887a-040c3a7b80bb    1
aa7f5fc5-f49a-44a7-b870-e7abfbd0fe76    1
aa897134-688c-45b9-ba5c-33c952dc0199    1
aa978022-96be-45b7-bf9c-e00fec32734e    1
aa994ac7-bf38-4b47-85cd-afbdd9c556b8    1
aa9b208a-31f7-456f-8968-beee2b2ab2c7    1
aaa762ef-af82-45b3-aa72-279403f1dbfd    1
aaa8345c-314b-4a24-aafb-aeb28f65c7ad    1
aaaaf89c-20bc-4974-8d0d-e31f1dc4f562    1
aab11d65-90a3-4f8a-98ac-58cfa19475ba    1
aab516e2-9881-4f4f-901a-cde597f7f9e9    1
aab6143a-424d-4a19-993e-03065412c420    1
000104b9-23e4-4ddc-8caa-8629fe8ad7f4    1
Name: Prospect ID, Length: 9074, dtype: int64
----------------------------------------------------------------------
------------------
landing page submission    4885
```

```
api                             3578
lead add form                    581
lead import                       30
Name: Lead Origin, dtype: int64
----------------------------------------------------------------------
------------------
google                 2873
direct traffic         2543
olark chat             1753
organic search         1154
reference               443
welingak website        129
referral sites          125
facebook                 31
bing                      6
click2call                4
press_release             2
social media              2
live chat                 2
pay per click ads         1
nc_edm                    1
testone                   1
welearn                   1
welearnblog_home          1
blog                      1
youtubechannel            1
Name: Lead Source, dtype: int64
----------------------------------------------------------------------
------------------
no     8358
yes     716
Name: Do Not Email, dtype: int64
----------------------------------------------------------------------
------------------
no     9072
yes       2
Name: Do Not Call, dtype: int64
----------------------------------------------------------------------
------------------
0    5639
1    3435
Name: Converted, dtype: int64
----------------------------------------------------------------------
------------------
0.0      2161
2.0      1679
3.0      1306
4.0      1120
5.0       783
```

```
6.0       466
1.0       395
7.0       309
8.0       224
9.0       164
10.0      114
11.0       86
13.0       48
12.0       45
14.0       36
16.0       21
15.0       18
17.0       16
18.0       15
20.0       12
19.0        9
21.0        6
23.0        6
24.0        5
25.0        5
27.0        5
22.0        3
29.0        2
26.0        2
28.0        2
43.0        1
115.0       1
74.0        1
55.0        1
54.0        1
141.0       1
42.0        1
41.0        1
32.0        1
30.0        1
251.0       1
Name: TotalVisits, dtype: int64
-----------------------------------------------------------------------------
------------------
0       2165
60        19
127       18
75        18
234       17
87        17
74        17
62        17
157       17
69        16
```

```
213         16
32          16
96          16
12          15
176         15
68          15
94          15
71          15
33          15
247         15
78          14
63          14
139         14
49          14
36          14
2           14
129         14
151         14
14          14
100         14
        ...
546          1
544          1
1214         1
460          1
1253         1
1251         1
1249         1
468          1
1235         1
1233         1
483          1
484          1
1229         1
486          1
495          1
509          1
1193         1
511          1
512          1
513          1
514          1
1206         1
522          1
523          1
524          1
528          1
530          1
1197         1
```

```
532         1
2272        1
Name: Total Time Spent on Website, Length: 1717, dtype: int64
----------------------------------------------------------------------
------------------
0.00     2161
2.00     1794
3.00     1196
4.00      896
1.00      651
5.00      517
1.50      306
6.00      244
2.50      241
7.00      133
3.50       94
8.00       86
1.33       66
1.67       60
2.33       59
2.67       54
9.00       45
4.50       43
1.75       28
3.33       27
10.00      25
1.25       23
5.50       21
2.25       19
11.00      18
3.67       16
6.50       13
1.80       13
2.75       12
1.40       11
         ...
1.31        1
1.27        1
1.21        1
8.21        1
1.63        1
3.91        1
4.17        1
2.63        1
24.00       1
2.57        1
2.56        1
2.86        1
2.45        1
```

```
2.90        1
2.38        1
3.17        1
2.29        1
3.29        1
3.38        1
3.43        1
2.14        1
2.13        1
3.57        1
2.08        1
1.93        1
1.86        1
3.80        1
3.82        1
3.83        1
55.00       1
Name: Page Views Per Visit, Length: 114, dtype: int64
----------------------------------------------------------------------
------------------
email opened                    3432
sms sent                        2716
olark chat conversation          972
page visited on website          640
converted to lead                428
email bounced                    312
email link clicked               267
form submitted on website        116
unreachable                       90
unsubscribed                      59
had a phone conversation          25
view in browser link clicked       6
approached upfront                 5
email marked spam                  2
email received                     2
resubscribed to emails             1
visited booth in tradeshow         1
Name: Last Activity, dtype: int64
----------------------------------------------------------------------
------------------
india            6491
not provided     2296
outside india     287
Name: Country, dtype: int64
----------------------------------------------------------------------
------------------
not provided                         3282
finance management                    959
human resource management             837
```

```
marketing management                       823
operations management                      499
business administration                    399
it projects management                     366
supply chain management                    346
banking, investment and insurance          335
media and advertising                      202
travel and tourism                         202
international business                     176
healthcare management                      156
hospitality management                     111
e-commerce                                 111
retail management                          100
rural and agribusiness                      73
e-business                                  57
services excellence                         40
Name: Specialization, dtype: int64
----------------------------------------------------------------------
------------------
unemployed              5476
not provided            2683
working professional     677
student                  206
other                     15
housewife                  9
businessman                8
Name: What is your current occupation, dtype: int64
----------------------------------------------------------------------
------------------
better career prospects      6370
not provided                 2702
other                           1
flexibility & convenience       1
Name: What matters most to you in choosing a course, dtype: int64
----------------------------------------------------------------------
------------------
no     9060
yes      14
Name: Search, dtype: int64
----------------------------------------------------------------------
------------------
no     9072
yes       2
Name: Newspaper Article, dtype: int64
----------------------------------------------------------------------
------------------
no     9073
yes       1
Name: X Education Forums, dtype: int64
```

```
--------------------------------------------------------------------
------------------
no     9073
yes       1
Name: Newspaper, dtype: int64
--------------------------------------------------------------------
------------------
no     9070
yes       4
Name: Digital Advertisement, dtype: int64
--------------------------------------------------------------------
------------------
no     9067
yes       7
Name: Through Recommendations, dtype: int64
--------------------------------------------------------------------
------------------
no     6186
yes    2888
Name: A free copy of Mastering The Interview, dtype: int64
--------------------------------------------------------------------
------------------
modified                        3267
email opened                    2823
sms sent                        2152
page visited on website          318
olark chat conversation          183
email link clicked               173
email bounced                     60
unsubscribed                      45
unreachable                       32
had a phone conversation          14
email marked spam                  2
view in browser link clicked       1
resubscribed to emails             1
form submitted on website          1
email received                     1
approached upfront                 1
Name: Last Notable Activity, dtype: int64
--------------------------------------------------------------------
------------------
```

```python
# Removing Id values since they are unique for everyone
df_final = df3.drop('Prospect ID',1)
df_final.shape
```

```
(9074, 21)
```

2. EDA

2.1. Univariate Analysis

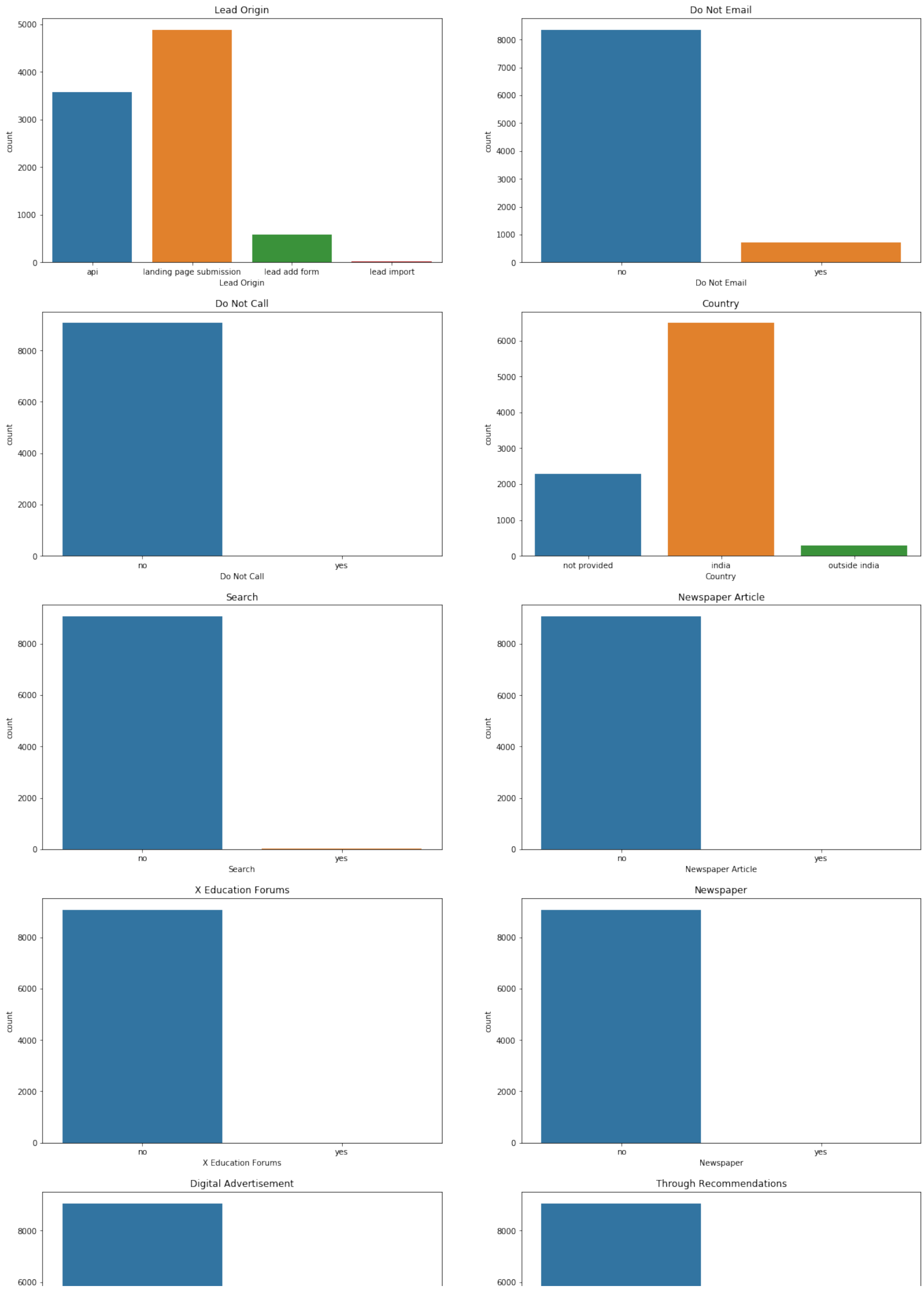
2.1.1. Categorical Variables

```
df_final.info()
```

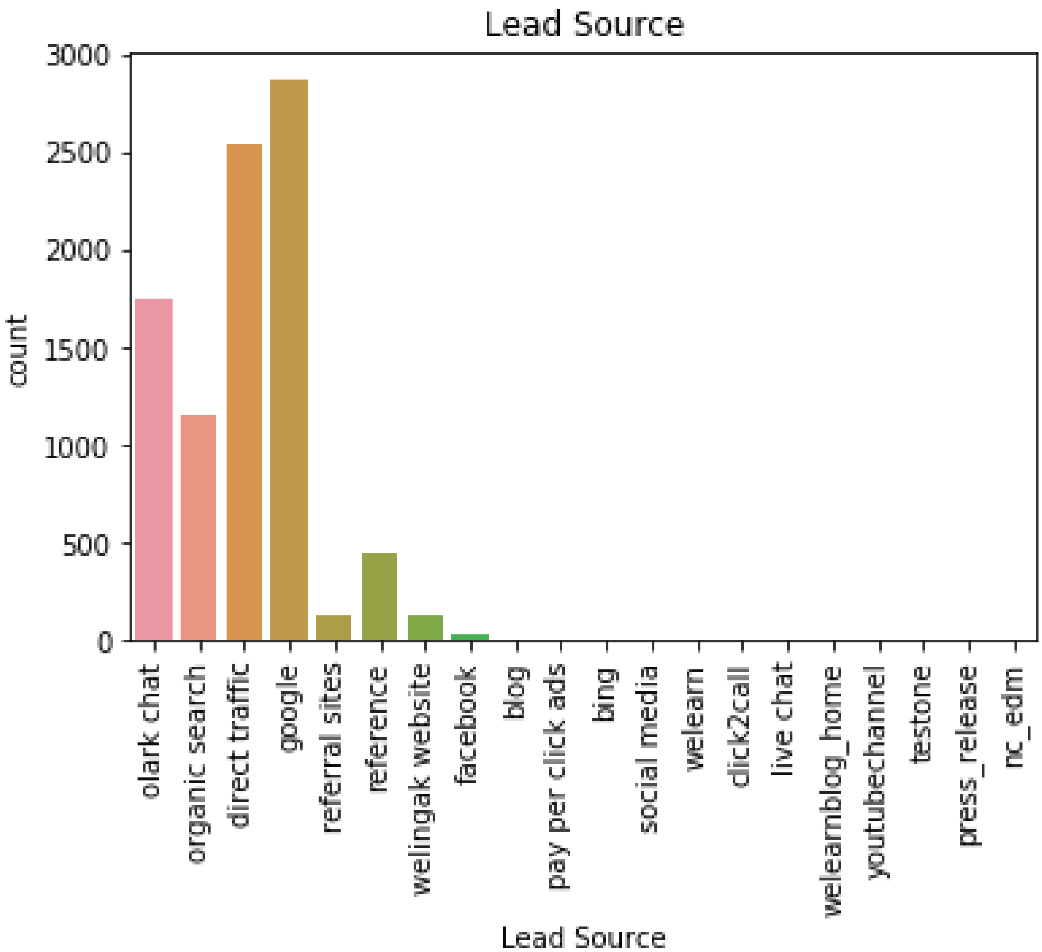
```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 9074 entries, 0 to 9239
Data columns (total 21 columns):
Lead Origin                                      9074 non-null object
Lead Source                                      9074 non-null object
Do Not Email                                     9074 non-null object
Do Not Call                                      9074 non-null object
Converted                                        9074 non-null int64
TotalVisits                                      9074 non-null float64
Total Time Spent on Website                      9074 non-null int64
Page Views Per Visit                             9074 non-null float64
Last Activity                                    9074 non-null object
Country                                          9074 non-null object
Specialization                                   9074 non-null object
What is your current occupation                  9074 non-null object
What matters most to you in choosing a course    9074 non-null object
Search                                           9074 non-null object
Newspaper Article                                9074 non-null object
X Education Forums                               9074 non-null object
Newspaper                                        9074 non-null object
Digital Advertisement                            9074 non-null object
Through Recommendations                          9074 non-null object
A free copy of Mastering The Interview           9074 non-null object
Last Notable Activity                            9074 non-null object
dtypes: float64(2), int64(2), object(17)
memory usage: 1.5+ MB
```

```
plt.figure(figsize = (20,40))

plt.subplot(6,2,1)
sns.countplot(df_final['Lead Origin'])
plt.title('Lead Origin')

plt.subplot(6,2,2)
sns.countplot(df_final['Do Not Email'])
plt.title('Do Not Email')

plt.subplot(6,2,3)
sns.countplot(df_final['Do Not Call'])
plt.title('Do Not Call')

plt.subplot(6,2,4)
```

```
sns.countplot(df_final['Country'])
plt.title('Country')

plt.subplot(6,2,5)
sns.countplot(df_final['Search'])
plt.title('Search')

plt.subplot(6,2,6)
sns.countplot(df_final['Newspaper Article'])
plt.title('Newspaper Article')

plt.subplot(6,2,7)
sns.countplot(df_final['X Education Forums'])
plt.title('X Education Forums')

plt.subplot(6,2,8)
sns.countplot(df_final['Newspaper'])
plt.title('Newspaper')

plt.subplot(6,2,9)
sns.countplot(df_final['Digital Advertisement'])
plt.title('Digital Advertisement')

plt.subplot(6,2,10)
sns.countplot(df_final['Through Recommendations'])
plt.title('Through Recommendations')

plt.subplot(6,2,11)
sns.countplot(df_final['A free copy of Mastering The Interview'])
plt.title('A free copy of Mastering The Interview')

plt.subplot(6,2,12)
sns.countplot(df_final['Last Notable Activity']).tick_params(axis='x',
rotation = 90)
plt.title('Last Notable Activity')


plt.show()
```

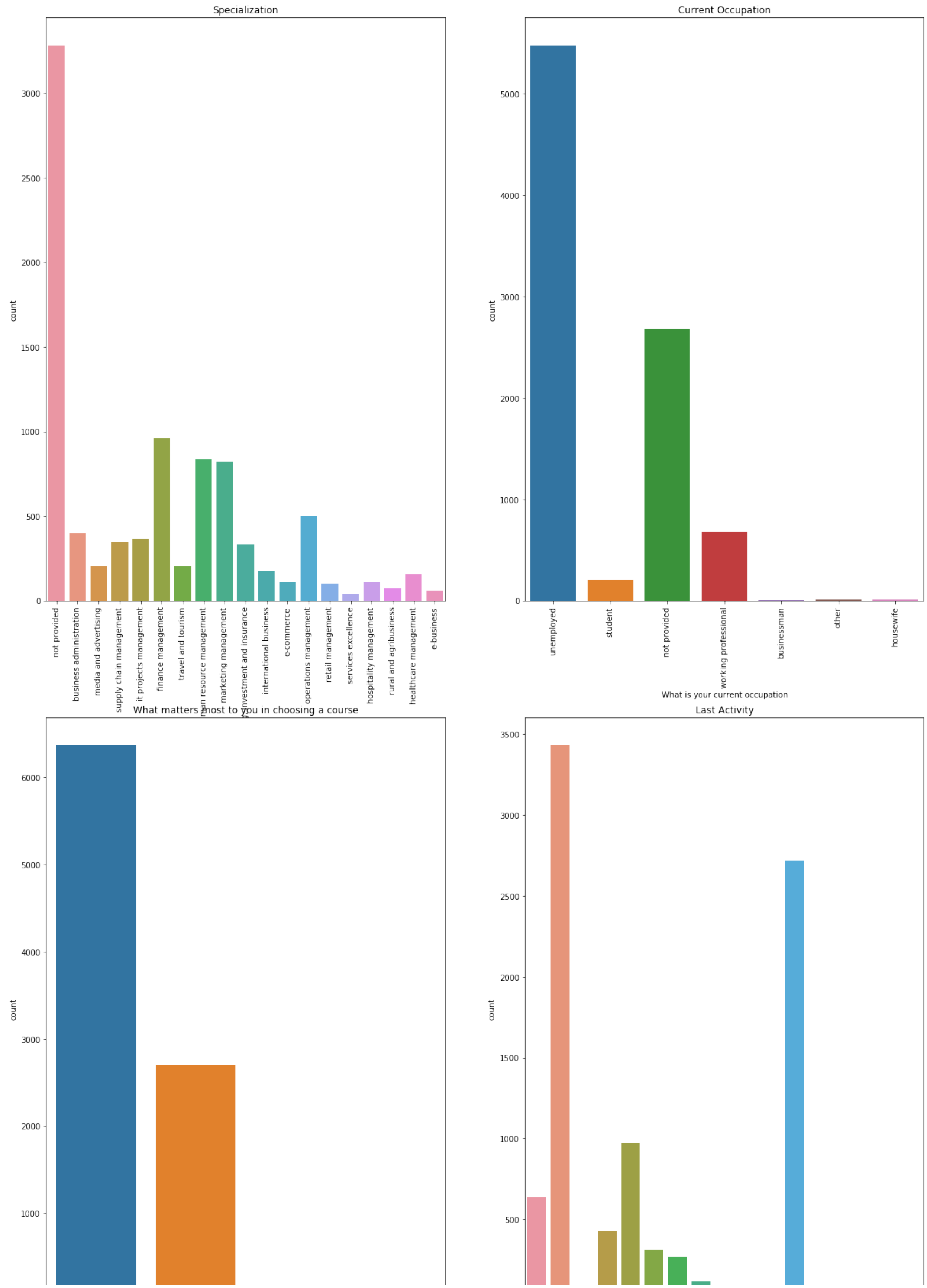

Lead Origin
count
0
1000
2000
3000
4000
5000
api
landing page submission
lead add form
lead import
Lead Origin
Do Not Email
count
0
1000
2000
3000
4000
5000
6000
7000
8000
no
yes
Do Not Email
Do Not Call
count
0
2000
4000
6000
8000
no
yes
Do Not Call
Country
count
0
1000
2000
3000
4000
5000
6000
not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

```
sns.countplot(df_final['Lead Source']).tick_params(axis='x', rotation
= 90)
plt.title('Lead Source')
plt.show()
```



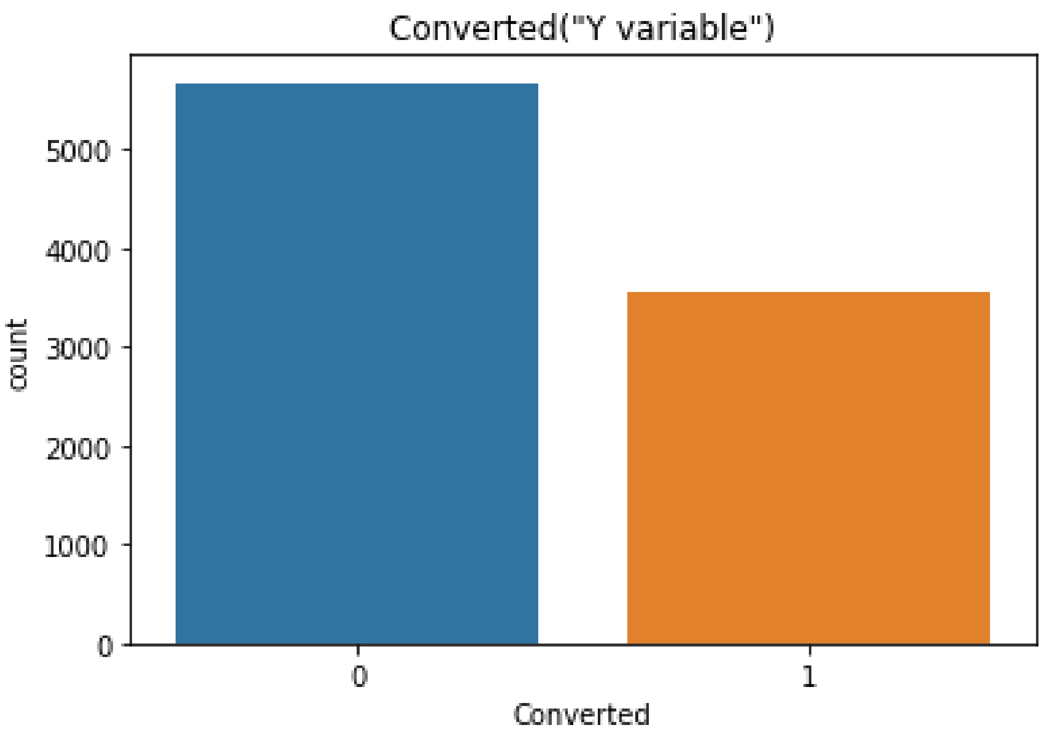

```
plt.figure(figsize = (20,30))
plt.subplot(2,2,1)
sns.countplot(df_final['Specialization']).tick_params(axis='x',
rotation = 90)
plt.title('Specialization')
plt.subplot(2,2,2)
sns.countplot(df_final['What is your current
occupation']).tick_params(axis='x', rotation = 90)
plt.title('Current Occupation')
plt.subplot(2,2,3)
sns.countplot(df_final['What matters most to you in choosing a
course']).tick_params(axis='x', rotation = 90)
plt.title('What matters most to you in choosing a course')
plt.subplot(2,2,4)
```

```
sns.countplot(df_final['Last Activity']).tick_params(axis='x',
rotation = 90)
plt.title('Last Activity')
plt.show()
```



Specialization
count
0
500
1000
1500
2000
2500
3000
not provided
business administration
media and advertising
supply chain management
it projects management
finance management
travel and tourism
man resource management
marketing management
investment and insurance
international business
e-commerce
operations management
retail management
services excellence
hospitality management
rural and agribusiness
healthcare management
e-business
Current Occupation
count
0
1000
2000
3000
4000
5000
unemployed
student
not provided
working professional
businessman
other
housewife
What is your current occupation
What matters most to you in choosing a course
count
1000
2000
3000
4000
5000
6000
Last Activity
count
500
1000
1500
2000
2500
3000
3500

```
sns.countplot(df['Converted'])
plt.title('Converted("Y variable")')
plt.show()
```

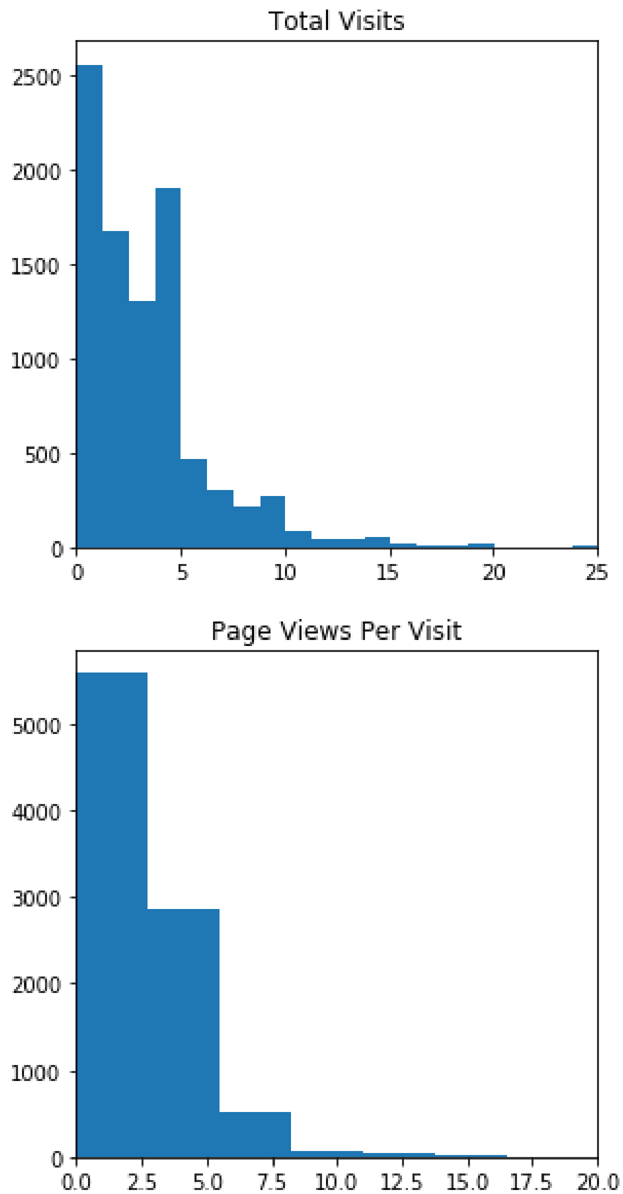


2.1.1. Numerical Variables

```
df_final.info()
```

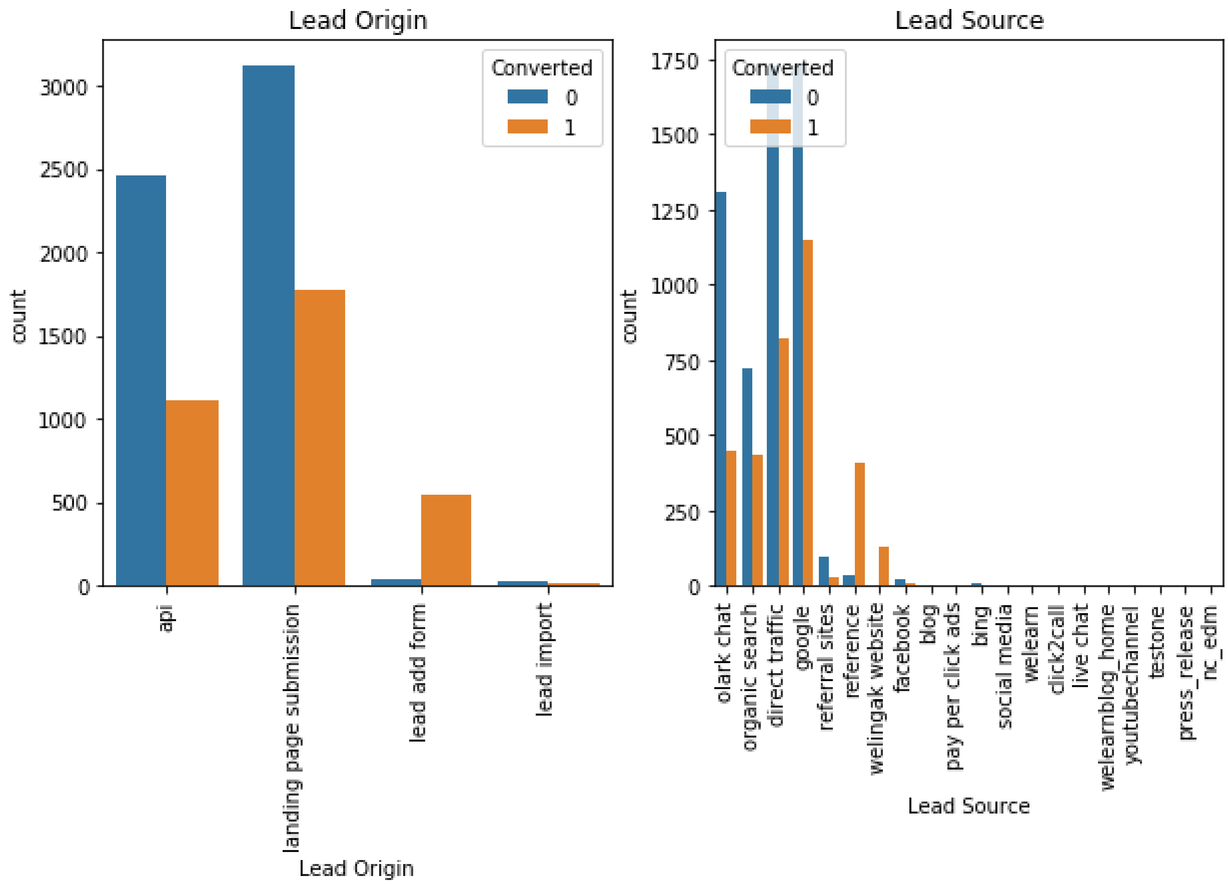
```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 9074 entries, 0 to 9239
Data columns (total 21 columns):
Lead Origin                                      9074 non-null object
Lead Source                                      9074 non-null object
Do Not Email                                     9074 non-null object
Do Not Call                                      9074 non-null object
Converted                                        9074 non-null int64
TotalVisits                                      9074 non-null float64
Total Time Spent on Website                      9074 non-null int64
Page Views Per Visit                             9074 non-null float64
Last Activity                                    9074 non-null object
Country                                          9074 non-null object
Specialization                                   9074 non-null object
What is your current occupation                  9074 non-null object
What matters most to you in choosing a course    9074 non-null object
Search                                           9074 non-null object
Newspaper Article                                9074 non-null object
X Education Forums                               9074 non-null object
```

```
Newspaper                                              9074 non-null object
Digital Advertisement                                  9074 non-null object
Through Recommendations                                9074 non-null object
A free copy of Mastering The Interview                 9074 non-null object
Last Notable Activity                                  9074 non-null object
dtypes: float64(2), int64(2), object(17)
memory usage: 1.8+ MB

plt.figure(figsize = (10,10))
plt.subplot(221)
plt.hist(df_final['TotalVisits'], bins = 200)
plt.title('Total Visits')
plt.xlim(0,25)

plt.subplot(222)
plt.hist(df_final['Total Time Spent on Website'], bins = 10)
plt.title('Total Time Spent on Website')

plt.subplot(223)
plt.hist(df_final['Page Views Per Visit'], bins = 20)
plt.title('Page Views Per Visit')
plt.xlim(0,20)
plt.show()
```

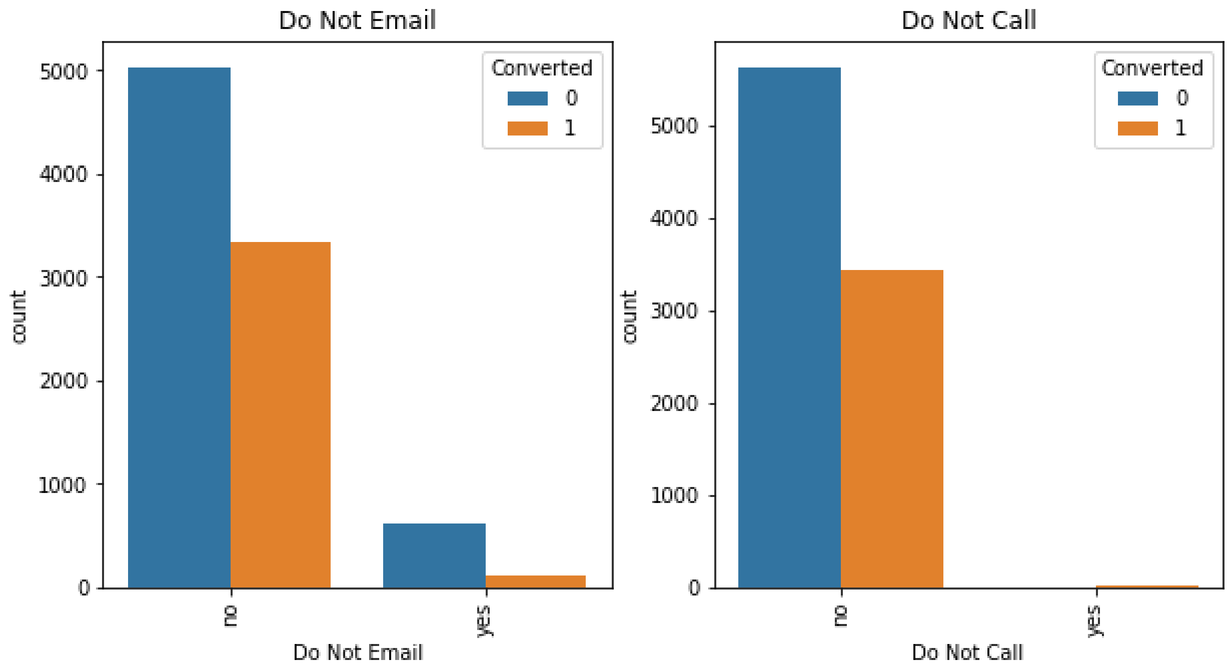


2.1. Relating all the categorical variables to Converted

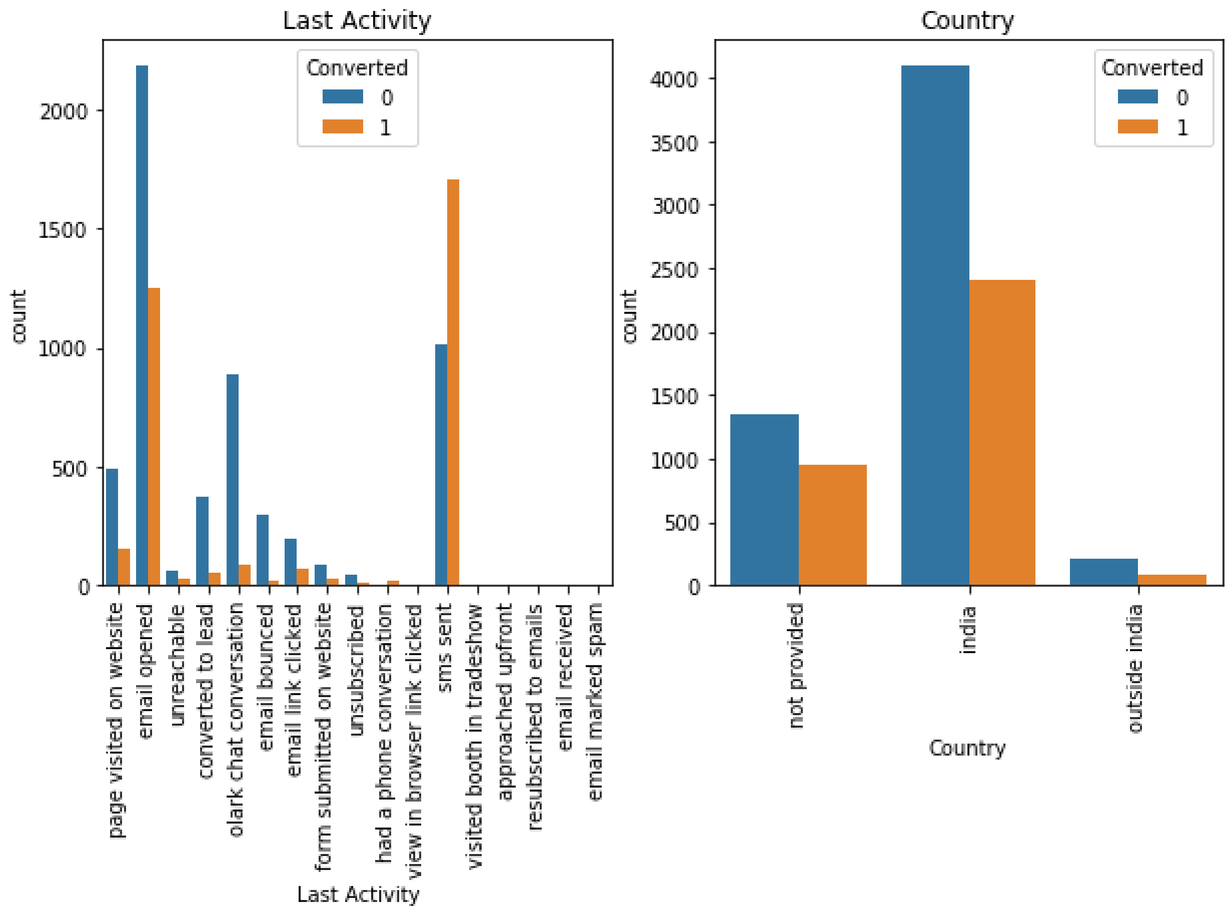
```
plt.figure(figsize = (10,5))

plt.subplot(1,2,1)
sns.countplot(x='Lead Origin', hue='Converted', data=
df_final).tick_params(axis='x', rotation = 90)
plt.title('Lead Origin')

plt.subplot(1,2,2)
sns.countplot(x='Lead Source', hue='Converted', data=
df_final).tick_params(axis='x', rotation = 90)
plt.title('Lead Source')
plt.show()
```



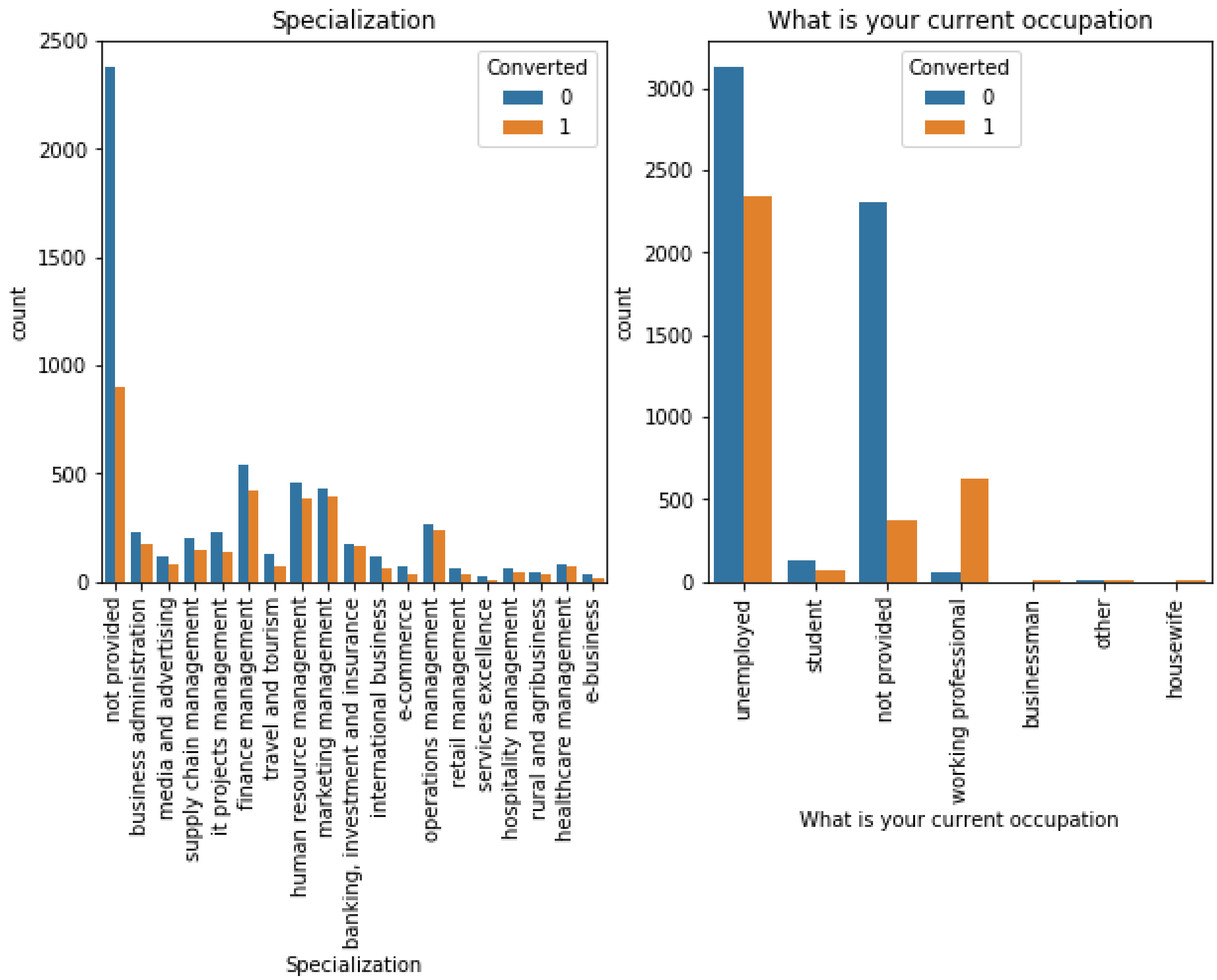

```
plt.figure(figsize = (10,5))

plt.subplot(1,2,1)
sns.countplot(x='Do Not Email', hue='Converted', data=
df_final).tick_params(axis='x', rotation = 90)
plt.title('Do Not Email')

plt.subplot(1,2,2)
sns.countplot(x='Do Not Call', hue='Converted', data=
df_final).tick_params(axis='x', rotation = 90)
plt.title('Do Not Call')
plt.show()
```



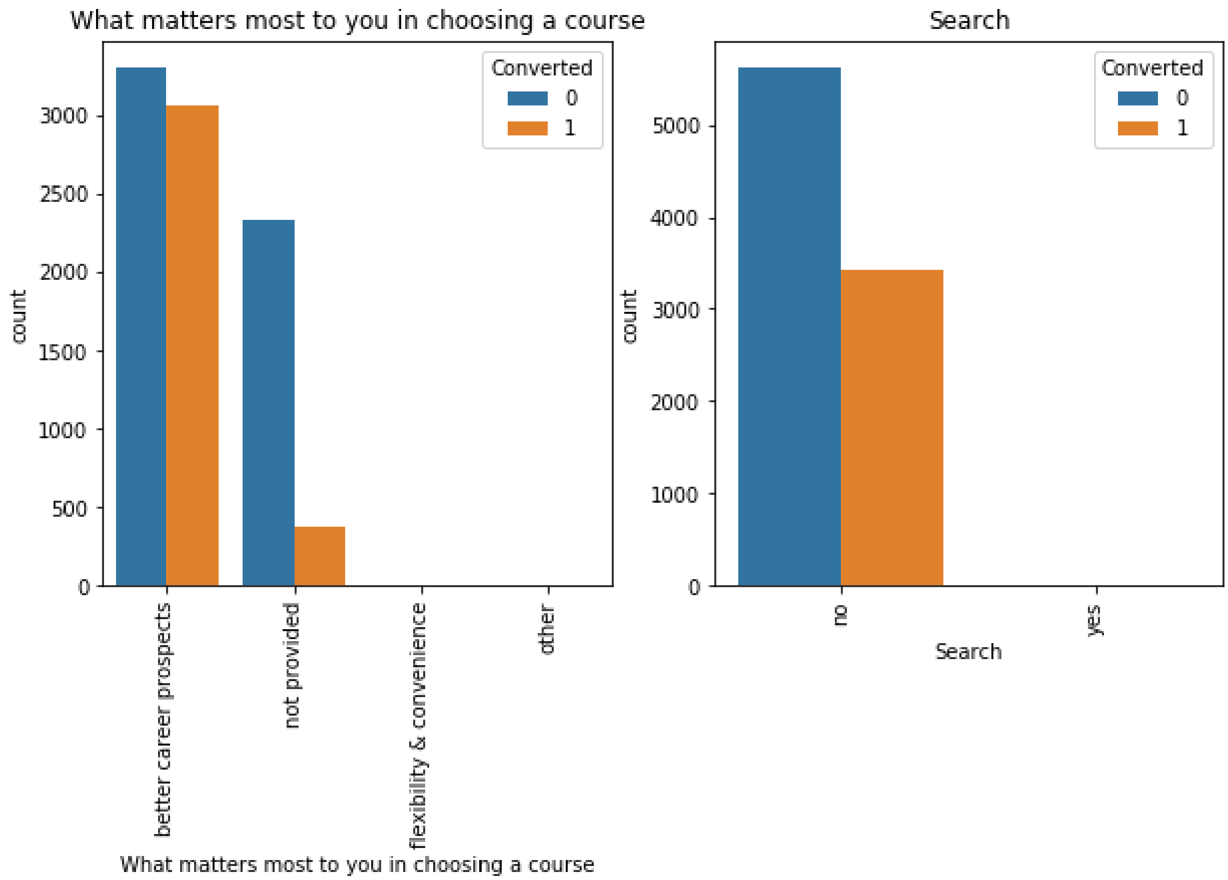

```
plt.figure(figsize = (10,5))

plt.subplot(1,2,1)
sns.countplot(x='Last Activity', hue='Converted', data=
df_final).tick_params(axis='x', rotation = 90)
plt.title('Last Activity')

plt.subplot(1,2,2)
sns.countplot(x='Country', hue='Converted', data=
df_final).tick_params(axis='x', rotation = 90)
plt.title('Country')
plt.show()
```



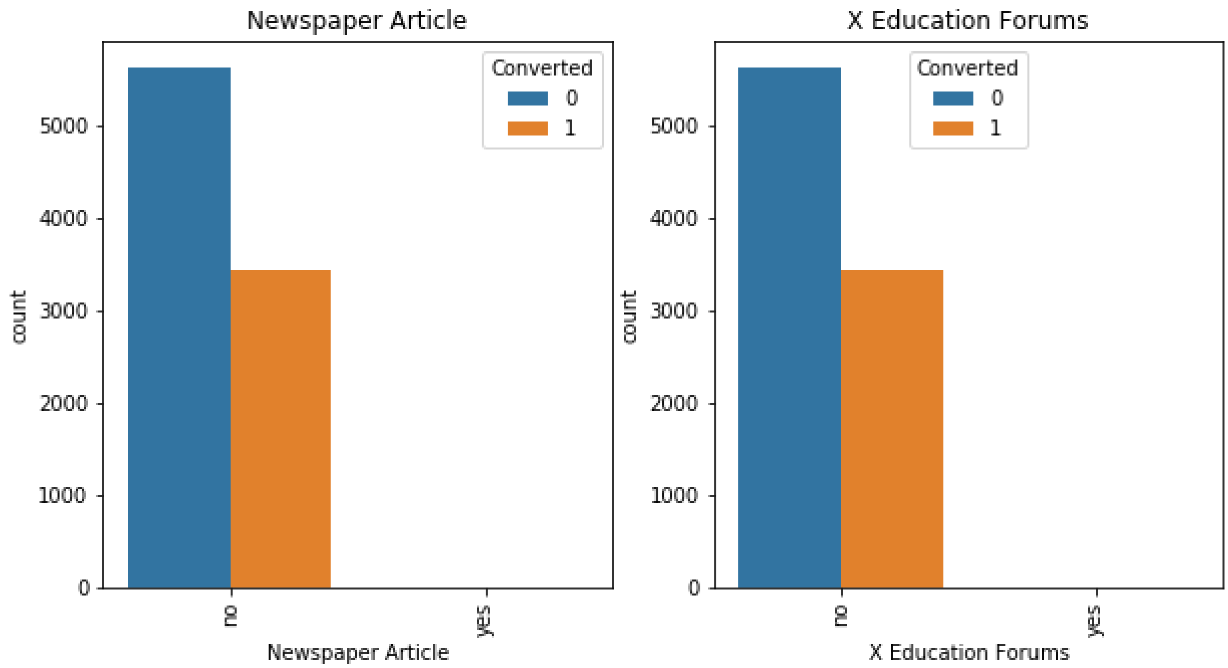

```
plt.figure(figsize = (10,5))

plt.subplot(1,2,1)
sns.countplot(x='Specialization', hue='Converted', data=
df_final).tick_params(axis='x', rotation = 90)
plt.title('Specialization')

plt.subplot(1,2,2)
sns.countplot(x='What is your current occupation', hue='Converted',
data= df_final).tick_params(axis='x', rotation = 90)
plt.title('What is your current occupation')
plt.show()
```



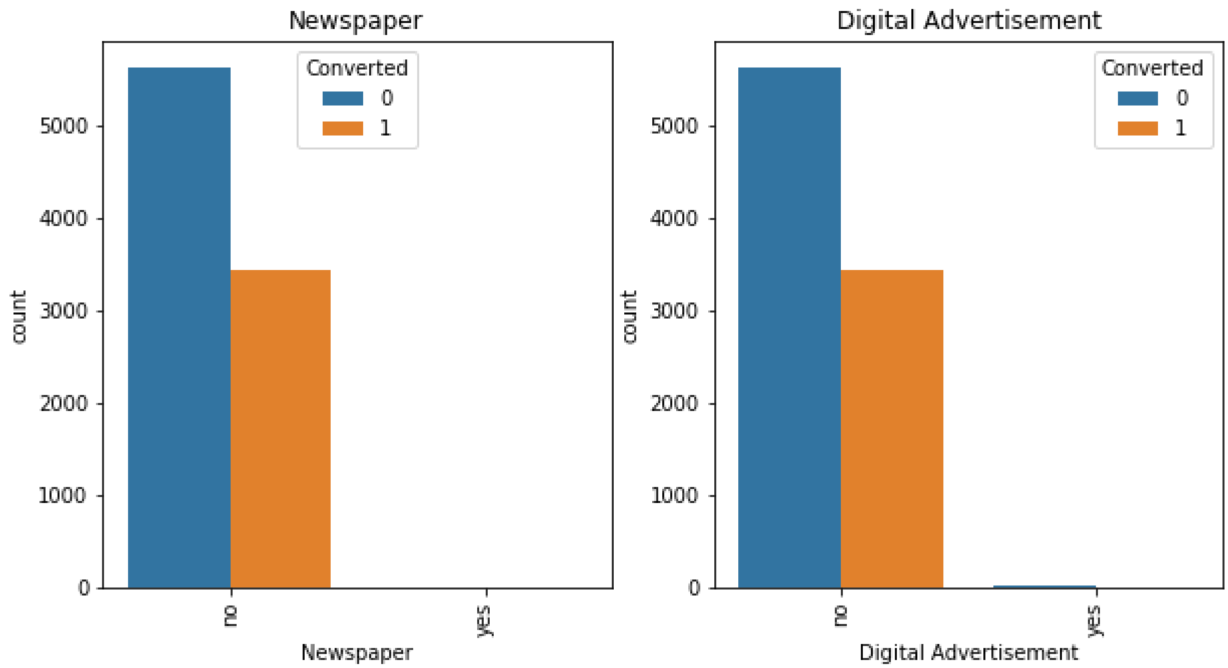

```
plt.figure(figsize = (10,5))

plt.subplot(1,2,1)
sns.countplot(x='What matters most to you in choosing a course',
hue='Converted', data= df_final).tick_params(axis='x', rotation = 90)
plt.title('What matters most to you in choosing a course')

plt.subplot(1,2,2)
sns.countplot(x='Search', hue='Converted', data=
df_final).tick_params(axis='x', rotation = 90)
plt.title('Search')
plt.show()
```



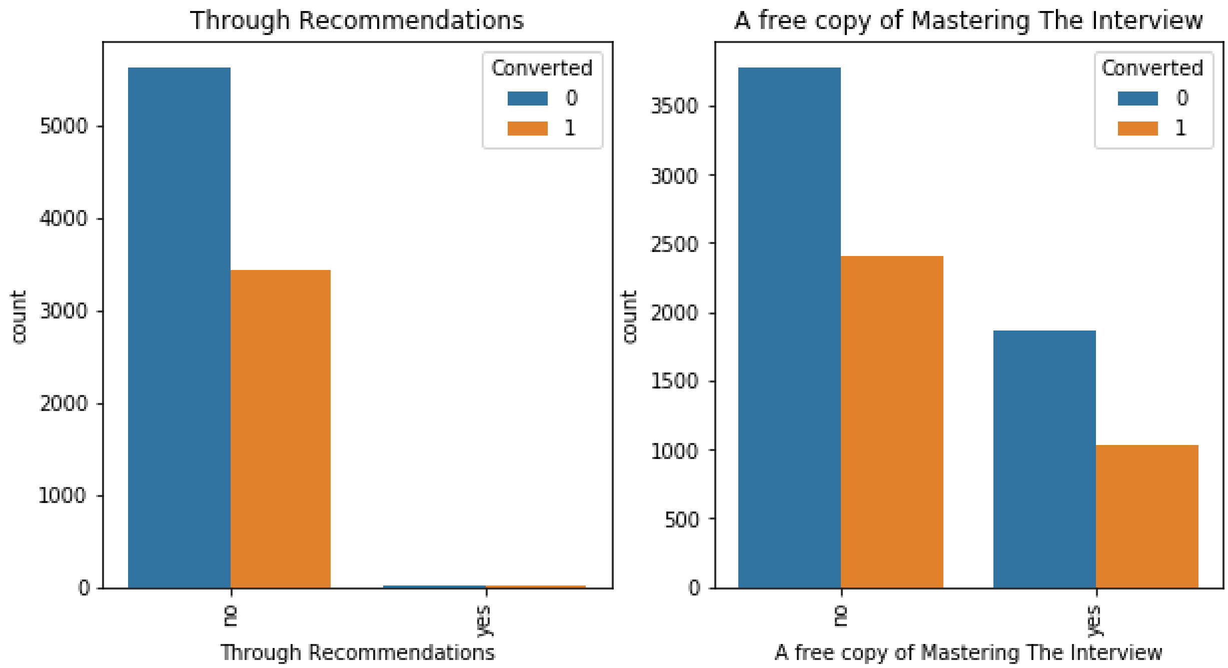

```
plt.figure(figsize = (10,5))

plt.subplot(1,2,1)
sns.countplot(x='Newspaper Article', hue='Converted', data=
df_final).tick_params(axis='x', rotation = 90)
plt.title('Newspaper Article')

plt.subplot(1,2,2)
sns.countplot(x='X Education Forums', hue='Converted', data=
df_final).tick_params(axis='x', rotation = 90)
plt.title('X Education Forums')
plt.show()
```



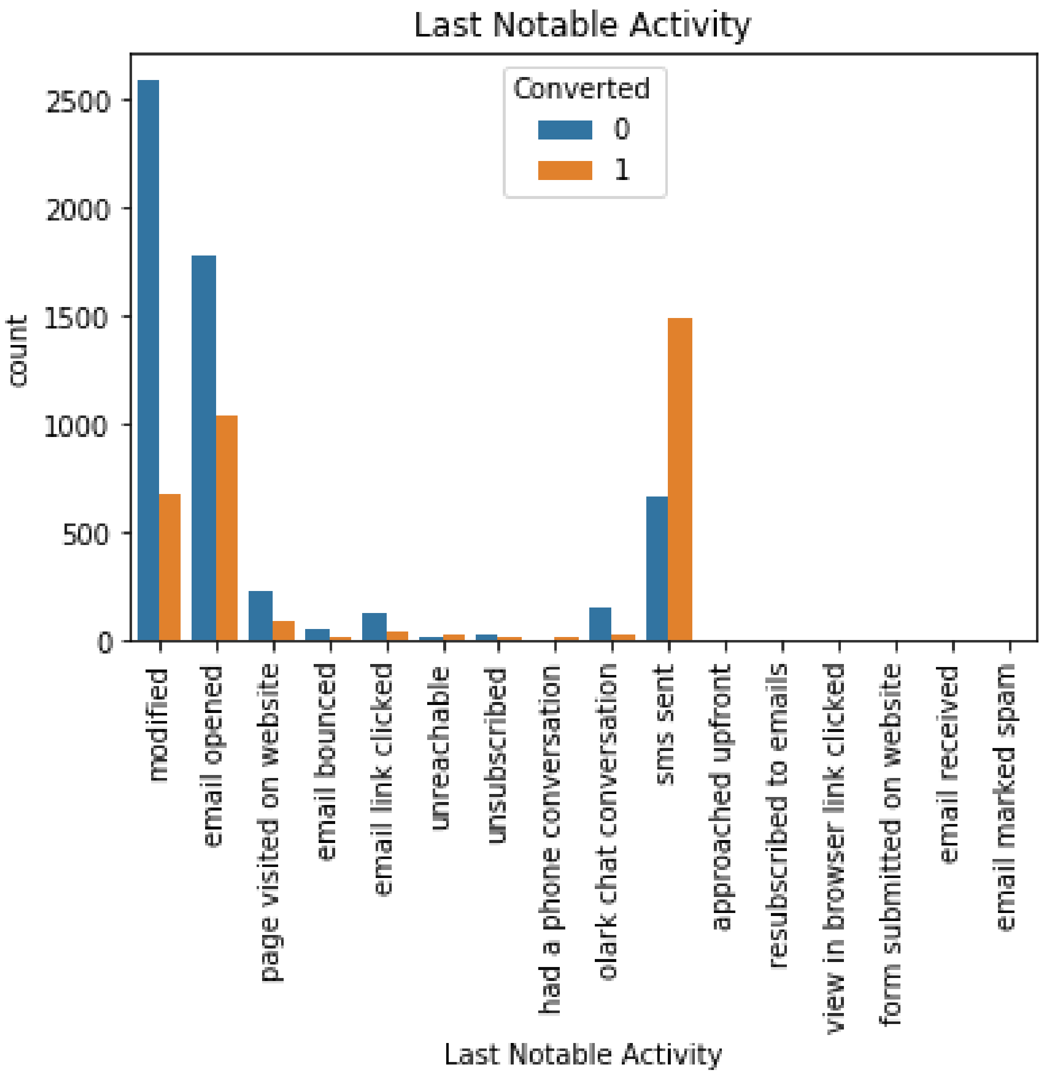

```
plt.figure(figsize = (10,5))

plt.subplot(1,2,1)
sns.countplot(x='Newspaper', hue='Converted', data=
df_final).tick_params(axis='x', rotation = 90)
plt.title('Newspaper')

plt.subplot(1,2,2)
sns.countplot(x='Digital Advertisement', hue='Converted', data=
df_final).tick_params(axis='x', rotation = 90)
plt.title('Digital Advertisement')
plt.show()
```



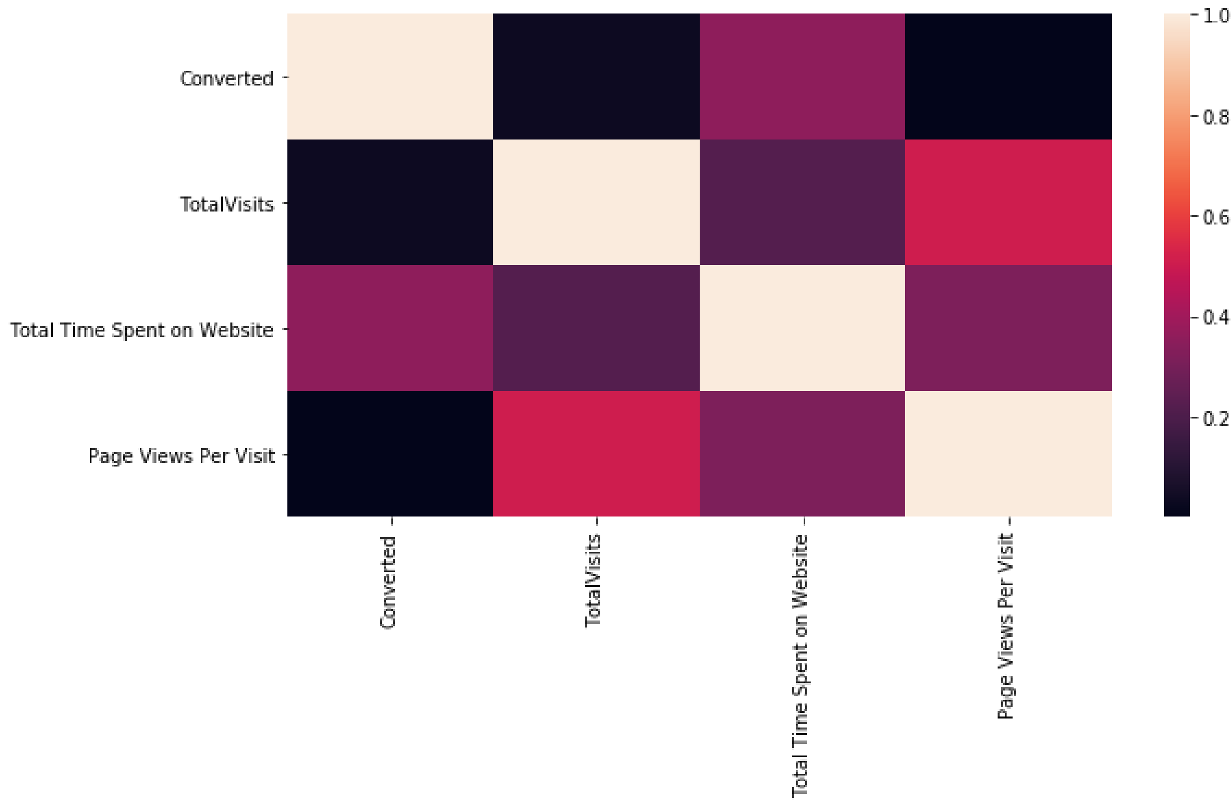

```
plt.figure(figsize = (10,5))

plt.subplot(1,2,1)
sns.countplot(x='Through Recommendations', hue='Converted', data=
df_final).tick_params(axis='x', rotation = 90)
plt.title('Through Recommendations')

plt.subplot(1,2,2)
sns.countplot(x='A free copy of Mastering The Interview',
hue='Converted', data= df_final).tick_params(axis='x', rotation = 90)
plt.title('A free copy of Mastering The Interview')
plt.show()
```




```
sns.countplot(x='Last Notable Activity', hue='Converted', data=
df_final).tick_params(axis='x', rotation = 90)
plt.title('Last Notable Activity')
plt.show()
```




```
# To check the correlation among varibles
plt.figure(figsize=(10,5))
sns.heatmap(df_final.corr())
plt.show()
```



It is understandable from the above EDA that there are many elements that have very little data and so will be of less relevance to our analysis.

```
numeric = df_final[['TotalVisits','Total Time Spent on Website','Page Views Per Visit']]
numeric.describe(percentiles=[0.25,0.5,0.75,0.9,0.99])
```

	TotalVisits	Total Time Spent on Website	Page Views Per Visit
count	9074.000000	9074.000000	9074.000000
mean	3.456028	482.887481	2.370151
std	4.858802	545.256560	2.160871
min	0.000000	0.000000	0.000000
25%	1.000000	11.000000	1.000000
50%	3.000000	246.000000	2.000000
75%	5.000000	922.750000	3.200000
90%	7.000000	1373.000000	5.000000
99%	17.000000	1839.000000	9.000000
max	251.000000	2272.000000	55.000000

There aren't any major outliers, so moving on to analysis

3. Dummy Variables

```
df_final.info()
```

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 9074 entries, 0 to 9239
Data columns (total 21 columns):
Lead Origin                                      9074 non-null object
Lead Source                                      9074 non-null object
Do Not Email                                     9074 non-null object
Do Not Call                                      9074 non-null object
Converted                                        9074 non-null int64
TotalVisits                                      9074 non-null float64
Total Time Spent on Website                      9074 non-null int64
Page Views Per Visit                             9074 non-null float64
Last Activity                                    9074 non-null object
Country                                          9074 non-null object
Specialization                                   9074 non-null object
What is your current occupation                  9074 non-null object
What matters most to you in choosing a course    9074 non-null object
Search                                           9074 non-null object
Newspaper Article                                9074 non-null object
X Education Forums                               9074 non-null object
Newspaper                                        9074 non-null object
Digital Advertisement                            9074 non-null object
Through Recommendations                          9074 non-null object
A free copy of Mastering The Interview           9074 non-null object
Last Notable Activity                            9074 non-null object
dtypes: float64(2), int64(2), object(17)
memory usage: 1.8+ MB
```

```
df_final.loc[:, df_final.dtypes == 'object'].columns
```

```
Index(['Lead Origin', 'Lead Source', 'Do Not Email', 'Do Not Call',
       'Last Activity', 'Country', 'Specialization',
       'What is your current occupation',
       'What matters most to you in choosing a course', 'Search',
       'Newspaper Article', 'X Education Forums', 'Newspaper',
       'Digital Advertisement', 'Through Recommendations',
       'A free copy of Mastering The Interview', 'Last Notable
Activity'],
      dtype='object')
```

```
# Create dummy variables using the 'get_dummies'
dummy = pd.get_dummies(df_final[['Lead Origin','Specialization' ,'Lead
Source', 'Do Not Email', 'Last Activity', 'What is your current
occupation','A free copy of Mastering The Interview', 'Last Notable
Activity']], drop_first=True)
# Add the results to the master dataframe
df_final_dum = pd.concat([df_final, dummy], axis=1)
df_final_dum
```

```
                  Lead Origin     Lead Source Do Not Email Do Not Call
\
```

0	api	olark chat	no	no
1	api	organic search	no	no
2	landing page submission	direct traffic	no	no
3	landing page submission	direct traffic	no	no
4	landing page submission	google	no	no
5	api	olark chat	no	no
6	landing page submission	google	no	no
7	api	olark chat	no	no
8	landing page submission	direct traffic	no	no
9	api	google	no	no
10	landing page submission	organic search	no	no
11	landing page submission	direct traffic	no	no
12	api	organic search	no	no
13	landing page submission	organic search	no	no
14	landing page submission	direct traffic	yes	no
15	api	organic search	no	no
16	api	olark chat	no	no
17	api	referral sites	no	no
18	landing page submission	google	no	no
19	api	organic search	no	no
20	landing page submission	google	no	no
21	api	google	no	no
22	landing page submission	google	no	no
23	landing page submission	google	no	no
24	api	google	no	no
25	landing page submission	google	no	no

26	landing page submission	organic search	no	no
27	landing page submission	google	no	no
28	landing page submission	direct traffic	no	no
29	api	google	no	no
...	...	...	...	...
9210	landing page submission	direct traffic	no	no
9211	landing page submission	direct traffic	no	no
9212	landing page submission	google	yes	no
9213	landing page submission	direct traffic	yes	no
9214	api	organic search	no	no
9215	landing page submission	organic search	no	no
9216	landing page submission	direct traffic	yes	no
9217	api	olark chat	no	no
9218	landing page submission	google	yes	no
9219	landing page submission	direct traffic	no	no
9220	landing page submission	direct traffic	no	no
9221	landing page submission	google	no	no
9222	api	google	no	no
9223	api	organic search	no	no
9224	landing page submission	google	no	no
9225	landing page submission	direct traffic	yes	no
9226	api	olark chat	no	no
9227	landing page submission	google	no	no
9228	landing page submission	google	no	no
9229	landing page submission	organic search	no	no
9230	landing page submission	google	no	no

```
9231  landing page submission          google            no           no

9232  landing page submission  direct traffic            no           no

9233                      api  direct traffic            no           no

9234  landing page submission  direct traffic            no           no

9235  landing page submission  direct traffic           yes           no

9236  landing page submission  direct traffic            no           no

9237  landing page submission  direct traffic           yes           no

9238  landing page submission          google            no           no

9239  landing page submission  direct traffic            no           no


      Converted  TotalVisits  Total Time Spent on Website  \
0             0          0.0                            0
1             0          5.0                          674
2             1          2.0                         1532
3             0          1.0                          305
4             1          2.0                         1428
5             0          0.0                            0
6             1          2.0                         1640
7             0          0.0                            0
8             0          2.0                           71
9             0          4.0                           58
10            1          8.0                         1351
11            1          8.0                         1343
12            1         11.0                         1538
13            0          5.0                          170
14            0          1.0                          481
15            1          6.0                         1012
16            0          0.0                            0
17            0          6.0                          973
18            1          6.0                         1688
19            0          3.0                           98
20            0          1.0                          233
21            0          4.0                          377
22            1          1.0                         1013
23            0          4.0                          771
24            1          6.0                         1137
25            1          3.0                         1068
26            1          4.0                         1000
27            1          6.0                         1315
28            0          5.0                          182
29            1          3.0                           78
```

```
...          ...          ...                            ...
9210           1          4.0                            927
9211           1          4.0                           1112
9212           0          5.0                             78
9213           0          5.0                            234
9214           1          2.0                            881
9215           0          8.0                            397
9216           0          6.0                           1679
9217           0          0.0                              0
9218           0          1.0                            149
9219           1          6.0                           1389
9220           0          5.0                             20
9221           0          4.0                           1347
9222           0          6.0                            228
9223           0          7.0                            142
9224           0          4.0                            455
9225           0          2.0                             74
9226           0          0.0                              0
9227           1          5.0                           1283
9228           1          4.0                           1944
9229           1         13.0                           1226
9230           0          2.0                            870
9231           1          8.0                           1016
9232           0          2.0                           1770
9233           1         13.0                           1409
9234           1          5.0                            210
9235           1          8.0                           1845
9236           0          2.0                            238
9237           0          2.0                            199
9238           1          3.0                            499
9239           1          6.0                           1279

      Page Views Per Visit              Last Activity
Country  \
0                     0.00    page visited on website   not provided

1                     2.50               email opened          india

2                     2.00               email opened          india

3                     1.00                unreachable          india

4                     1.00          converted to lead          india

5                     0.00    olark chat conversation   not provided

6                     2.00               email opened          india

7                     0.00    olark chat conversation   not provided
```

8	2.00	email opened	india
9	4.00	email opened	india
10	8.00	email opened	india
11	2.67	page visited on website	india
12	11.00	email opened	india
13	5.00	email opened	india
14	1.00	email bounced	outside india
15	6.00	email opened	india
16	0.00	olark chat conversation	not provided
17	6.00	email link clicked	india
18	3.00	page visited on website	india
19	3.00	page visited on website	india
20	1.00	unreachable	india
21	1.33	page visited on website	india
22	1.00	converted to lead	india
23	4.00	email link clicked	india
24	1.50	email opened	india
25	3.00	form submitted on website	india
26	2.00	email opened	india
27	6.00	email opened	india
28	5.00	email opened	india
29	3.00	unreachable	india
...	...	...	...
9210	4.00	email link clicked	india
9211	4.00	sms sent	india
9212	5.00	unsubscribed	india

9213	2.50	page visited on website	india
9214	2.00	sms sent	india
9215	8.00	email opened	india
9216	6.00	page visited on website	india
9217	0.00	sms sent	not provided
9218	1.00	email bounced	india
9219	6.00	email opened	india
9220	2.50	sms sent	india
9221	2.00	sms sent	india
9222	6.00	sms sent	india
9223	7.00	email opened	india
9224	4.00	form submitted on website	india
9225	2.00	email bounced	outside india
9226	0.00	sms sent	not provided
9227	1.67	email opened	india
9228	2.00	sms sent	india
9229	6.50	sms sent	india
9230	2.00	email opened	india
9231	4.00	email opened	india
9232	2.00	sms sent	india
9233	2.60	sms sent	india
9234	2.50	sms sent	india
9235	2.67	email marked spam	outside india
9236	2.00	sms sent	india
9237	2.00	sms sent	india
9238	3.00	sms sent	india

```
9239                  3.00                  sms sent  outside india


                            ...                         \
0                           ...
1                           ...
2                           ...
3                           ...
4                           ...
5                           ...
6                           ...
7                           ...
8                           ...
9                           ...
10                          ...
11                          ...
12                          ...
13                          ...
14                          ...
15                          ...
16                          ...
17                          ...
18                          ...
19                          ...
20                          ...
21                          ...
22                          ...
23                          ...
24                          ...
25                          ...
26                          ...
27                          ...
28                          ...
29                          ...
...                         ...
9210                        ...
9211                        ...
9212                        ...
9213                        ...
9214                        ...
9215                        ...
9216                        ...
9217                        ...
9218                        ...
9219                        ...
9220                        ...
9221                        ...
9222                        ...
9223                        ...
```

```
9224                                ...
9225                                ...
9226                                ...
9227                                ...
9228                                ...
9229                                ...
9230                                ...
9231                                ...
9232                                ...
9233                                ...
9234                                ...
9235                                ...
9236                                ...
9237                                ...
9238                                ...
9239                                ...

     Last Notable Activity_form submitted on website  \
0                                                   0
1                                                   0
2                                                   0
3                                                   0
4                                                   0
5                                                   0
6                                                   0
7                                                   0
8                                                   0
9                                                   0
10                                                  0
11                                                  0
12                                                  0
13                                                  0
14                                                  0
15                                                  0
16                                                  0
17                                                  0
18                                                  0
19                                                  0
20                                                  0
21                                                  0
22                                                  0
23                                                  0
24                                                  0
25                                                  0
26                                                  0
27                                                  0
28                                                  0
29                                                  0
...                                               ...
```

```
9210                                              0
9211                                              0
9212                                              0
9213                                              0
9214                                              0
9215                                              0
9216                                              0
9217                                              0
9218                                              0
9219                                              0
9220                                              0
9221                                              0
9222                                              0
9223                                              0
9224                                              0
9225                                              0
9226                                              0
9227                                              0
9228                                              0
9229                                              0
9230                                              0
9231                                              0
9232                                              0
9233                                              0
9234                                              0
9235                                              0
9236                                              0
9237                                              0
9238                                              0
9239                                              0

     Last Notable Activity_had a phone conversation  \
0                                                0
1                                                0
2                                                0
3                                                0
4                                                0
5                                                0
6                                                0
7                                                0
8                                                0
9                                                0
10                                               0
11                                               0
12                                               0
13                                               0
14                                               0
15                                               0
16                                               0
```

```
17                                                  0
18                                                  0
19                                                  0
20                                                  0
21                                                  0
22                                                  0
23                                                  0
24                                                  0
25                                                  0
26                                                  0
27                                                  0
28                                                  0
29                                                  0
...                                               ...
9210                                                0
9211                                                0
9212                                                0
9213                                                0
9214                                                0
9215                                                0
9216                                                0
9217                                                0
9218                                                0
9219                                                0
9220                                                0
9221                                                0
9222                                                0
9223                                                0
9224                                                0
9225                                                0
9226                                                0
9227                                                0
9228                                                0
9229                                                0
9230                                                0
9231                                                0
9232                                                0
9233                                                0
9234                                                0
9235                                                0
9236                                                0
9237                                                0
9238                                                0
9239                                                0

     Last Notable Activity_modified  \
0                                 1
1                                 0
2                                 0
3                                 1
```

4	1
5	1
6	1
7	1
8	0
9	0
10	0
11	0
12	1
13	0
14	0
15	0
16	1
17	1
18	0
19	1
20	1
21	1
22	1
23	0
24	0
25	1
26	0
27	0
28	0
29	0
...	...
9210	1
9211	0
9212	0
9213	1
9214	0
9215	0
9216	1
9217	0
9218	1
9219	0
9220	1
9221	0
9222	1
9223	1
9224	1
9225	1
9226	1
9227	0
9228	1
9229	1
9230	0
9231	0

```
9232                                0
9233                                0
9234                                1
9235                                0
9236                                0
9237                                0
9238                                0
9239                                1

     Last Notable Activity_olark chat conversation  \
0                                                0
1                                                0
2                                                0
3                                                0
4                                                0
5                                                0
6                                                0
7                                                0
8                                                0
9                                                0
10                                               0
11                                               0
12                                               0
13                                               0
14                                               0
15                                               0
16                                               0
17                                               0
18                                               0
19                                               0
20                                               0
21                                               0
22                                               0
23                                               0
24                                               0
25                                               0
26                                               0
27                                               0
28                                               0
29                                               0
...                                            ...
9210                                             0
9211                                             0
9212                                             0
9213                                             0
9214                                             0
9215                                             0
9216                                             0
9217                                             0
```

```
9218                                              0
9219                                              0
9220                                              0
9221                                              0
9222                                              0
9223                                              0
9224                                              0
9225                                              0
9226                                              0
9227                                              0
9228                                              0
9229                                              0
9230                                              0
9231                                              0
9232                                              0
9233                                              0
9234                                              0
9235                                              0
9236                                              0
9237                                              0
9238                                              0
9239                                              0

     Last Notable Activity_page visited on website  \
0                                                 0
1                                                 0
2                                                 0
3                                                 0
4                                                 0
5                                                 0
6                                                 0
7                                                 0
8                                                 0
9                                                 0
10                                                0
11                                                1
12                                                0
13                                                0
14                                                0
15                                                0
16                                                0
17                                                0
18                                                1
19                                                0
20                                                0
21                                                0
22                                                0
23                                                0
24                                                0
```

```
25                                                  0
26                                                  0
27                                                  0
28                                                  0
29                                                  0
...                                               ...
9210                                                0
9211                                                0
9212                                                0
9213                                                0
9214                                                0
9215                                                0
9216                                                0
9217                                                0
9218                                                0
9219                                                0
9220                                                0
9221                                                0
9222                                                0
9223                                                0
9224                                                0
9225                                                0
9226                                                0
9227                                                0
9228                                                0
9229                                                0
9230                                                0
9231                                                0
9232                                                0
9233                                                0
9234                                                0
9235                                                0
9236                                                0
9237                                                0
9238                                                0
9239                                                0

     Last Notable Activity_resubscribed to emails  \
0                                                0
1                                                0
2                                                0
3                                                0
4                                                0
5                                                0
6                                                0
7                                                0
8                                                0
9                                                0
10                                               0
```

11	0
12	0
13	0
14	0
15	0
16	0
17	0
18	0
19	0
20	0
21	0
22	0
23	0
24	0
25	0
26	0
27	0
28	0
29	0
...	...
9210	0
9211	0
9212	0
9213	0
9214	0
9215	0
9216	0
9217	0
9218	0
9219	0
9220	0
9221	0
9222	0
9223	0
9224	0
9225	0
9226	0
9227	0
9228	0
9229	0
9230	0
9231	0
9232	0
9233	0
9234	0
9235	0
9236	0
9237	0
9238	0

```
9239                                            0

     Last Notable Activity_sms sent Last Notable Activity_unreachable  \
0                                 0                                 0
1                                 0                                 0
2                                 0                                 0
3                                 0                                 0
4                                 0                                 0
5                                 0                                 0
6                                 0                                 0
7                                 0                                 0
8                                 0                                 0
9                                 0                                 0
10                                0                                 0
11                                0                                 0
12                                0                                 0
13                                0                                 0
14                                0                                 0
15                                0                                 0
16                                0                                 0
17                                0                                 0
18                                0                                 0
19                                0                                 0
20                                0                                 0
21                                0                                 0
22                                0                                 0
23                                0                                 0
```

24	0	0
25	0	0
26	0	0
27	0	0
28	0	0
29	0	1
...	...	...
9210	0	0
9211	1	0
9212	0	0
9213	0	0
9214	1	0
9215	0	0
9216	0	0
9217	1	0
9218	0	0
9219	0	0
9220	0	0
9221	1	0
9222	0	0
9223	0	0
9224	0	0
9225	0	0
9226	0	0
9227	0	0
9228	0	0

```
9229                                    0                                       0
9230                                    0                                       0
9231                                    0                                       0
9232                                    1                                       0
9233                                    1                                       0
9234                                    0                                       0
9235                                    0                                       0
9236                                    1                                       0
9237                                    1                                       0
9238                                    1                                       0
9239                                    0                                       0

     Last Notable Activity_unsubscribed  \
0                                     0
1                                     0
2                                     0
3                                     0
4                                     0
5                                     0
6                                     0
7                                     0
8                                     0
9                                     0
10                                    0
11                                    0
12                                    0
13                                    0
14                                    0
15                                    0
16                                    0
17                                    0
18                                    0
19                                    0
20                                    0
21                                    0
22                                    0
23                                    0
24                                    0
25                                    0
```

```
26                                    0
27                                    0
28                                    0
29                                    0
...                                 ...
9210                                  0
9211                                  0
9212                                  1
9213                                  0
9214                                  0
9215                                  0
9216                                  0
9217                                  0
9218                                  0
9219                                  0
9220                                  0
9221                                  0
9222                                  0
9223                                  0
9224                                  0
9225                                  0
9226                                  0
9227                                  0
9228                                  0
9229                                  0
9230                                  0
9231                                  0
9232                                  0
9233                                  0
9234                                  0
9235                                  0
9236                                  0
9237                                  0
9238                                  0
9239                                  0

     Last Notable Activity_view in browser link clicked
0                                                    0
1                                                    0
2                                                    0
3                                                    0
4                                                    0
5                                                    0
6                                                    0
7                                                    0
8                                                    0
9                                                    0
10                                                   0
11                                                   0
```

```
12      0
13      0
14      0
15      0
16      0
17      0
18      0
19      0
20      0
21      0
22      0
23      0
24      0
25      0
26      0
27      0
28      0
29      0
...    ..
9210    0
9211    0
9212    0
9213    0
9214    0
9215    0
9216    0
9217    0
9218    0
9219    0
9220    0
9221    0
9222    0
9223    0
9224    0
9225    0
9226    0
9227    0
9228    0
9229    0
9230    0
9231    0
9232    0
9233    0
9234    0
9235    0
9236    0
9237    0
9238    0
9239    0
```

```
[9074 rows x 100 columns]

df_final_dum = df_final_dum.drop(['What is your current occupation_not
provided','Lead Origin', 'Lead Source', 'Do Not Email', 'Do Not
Call','Last Activity', 'Country', 'Specialization',
'Specialization_not provided','What is your current occupation','What
matters most to you in choosing a course', 'Search','Newspaper
Article', 'X Education Forums', 'Newspaper','Digital Advertisement',
'Through Recommendations','A free copy of Mastering The Interview',
'Last Notable Activity'], 1)
df_final_dum
```

```
      Converted  TotalVisits  Total Time Spent on Website  \
0             0          0.0                            0
1             0          5.0                          674
2             1          2.0                         1532
3             0          1.0                          305
4             1          2.0                         1428
5             0          0.0                            0
6             1          2.0                         1640
7             0          0.0                            0
8             0          2.0                           71
9             0          4.0                           58
10            1          8.0                         1351
11            1          8.0                         1343
12            1         11.0                         1538
13            0          5.0                          170
14            0          1.0                          481
15            1          6.0                         1012
16            0          0.0                            0
17            0          6.0                          973
18            1          6.0                         1688
19            0          3.0                           98
20            0          1.0                          233
21            0          4.0                          377
22            1          1.0                         1013
23            0          4.0                          771
24            1          6.0                         1137
25            1          3.0                         1068
26            1          4.0                         1000
27            1          6.0                         1315
28            0          5.0                          182
29            1          3.0                           78
...         ...          ...                          ...
9210          1          4.0                          927
9211          1          4.0                         1112
9212          0          5.0                           78
9213          0          5.0                          234
9214          1          2.0                          881
```

9215	0	8.0	397
9216	0	6.0	1679
9217	0	0.0	0
9218	0	1.0	149
9219	1	6.0	1389
9220	0	5.0	20
9221	0	4.0	1347
9222	0	6.0	228
9223	0	7.0	142
9224	0	4.0	455
9225	0	2.0	74
9226	0	0.0	0
9227	1	5.0	1283
9228	1	4.0	1944
9229	1	13.0	1226
9230	0	2.0	870
9231	1	8.0	1016
9232	0	2.0	1770
9233	1	13.0	1409
9234	1	5.0	210
9235	1	8.0	1845
9236	0	2.0	238
9237	0	2.0	199
9238	1	3.0	499
9239	1	6.0	1279

	Page Views Per Visit	Lead Origin_landing page submission \
0	0.00	0
1	2.50	0
2	2.00	1
3	1.00	1
4	1.00	1
5	0.00	0
6	2.00	1
7	0.00	0
8	2.00	1
9	4.00	0
10	8.00	1
11	2.67	1
12	11.00	0
13	5.00	1
14	1.00	1
15	6.00	0
16	0.00	0
17	6.00	0
18	3.00	1
19	3.00	0
20	1.00	1
21	1.33	0

```
22                     1.00                                      1
23                     4.00                                      1
24                     1.50                                      0
25                     3.00                                      1
26                     2.00                                      1
27                     6.00                                      1
28                     5.00                                      1
29                     3.00                                      0
...                     ...                                    ...
9210                   4.00                                      1
9211                   4.00                                      1
9212                   5.00                                      1
9213                   2.50                                      1
9214                   2.00                                      0
9215                   8.00                                      1
9216                   6.00                                      1
9217                   0.00                                      0
9218                   1.00                                      1
9219                   6.00                                      1
9220                   2.50                                      1
9221                   2.00                                      1
9222                   6.00                                      0
9223                   7.00                                      0
9224                   4.00                                      1
9225                   2.00                                      1
9226                   0.00                                      0
9227                   1.67                                      1
9228                   2.00                                      1
9229                   6.50                                      1
9230                   2.00                                      1
9231                   4.00                                      1
9232                   2.00                                      1
9233                   2.60                                      0
9234                   2.50                                      1
9235                   2.67                                      1
9236                   2.00                                      1
9237                   2.00                                      1
9238                   3.00                                      1
9239                   3.00                                      1

      Lead Origin_lead add form  Lead Origin_lead import  \
0                             0                        0
1                             0                        0
2                             0                        0
3                             0                        0
4                             0                        0
5                             0                        0
6                             0                        0
7                             0                        0
```

8	0	0
9	0	0
10	0	0
11	0	0
12	0	0
13	0	0
14	0	0
15	0	0
16	0	0
17	0	0
18	0	0
19	0	0
20	0	0
21	0	0
22	0	0
23	0	0
24	0	0
25	0	0
26	0	0
27	0	0
28	0	0
29	0	0
...	...	...
9210	0	0
9211	0	0
9212	0	0
9213	0	0
9214	0	0
9215	0	0
9216	0	0
9217	0	0
9218	0	0
9219	0	0
9220	0	0
9221	0	0
9222	0	0
9223	0	0
9224	0	0
9225	0	0
9226	0	0
9227	0	0
9228	0	0
9229	0	0
9230	0	0
9231	0	0
9232	0	0
9233	0	0
9234	0	0
9235	0	0

```
9236                              0                          0
9237                              0                          0
9238                              0                          0
9239                              0                          0

      Specialization_business administration  Specialization_e-
business  \
0                                          0
0
1                                          0
0
2                                          1
0
3                                          0
0
4                                          0
0
5                                          0
0
6                                          0
0
7                                          0
0
8                                          0
0
9                                          0
0
10                                         0
0
11                                         0
0
12                                         0
0
13                                         1
0
14                                         1
0
15                                         0
0
16                                         0
0
17                                         0
0
18                                         0
0
19                                         0
0
20                                         0
0
```

```
21                                          0
0
22                                          0
0
23                                          0
0
24                                          0
0
25                                          0
0
26                                          0
0
27                                          0
0
28                                          0
0
29                                          0
0
...                                       ...                        ..
.
9210                                        0
0
9211                                        0
0
9212                                        0
0
9213                                        0
0
9214                                        0
0
9215                                        1
0
9216                                        1
0
9217                                        0
0
9218                                        0
0
9219                                        0
0
9220                                        0
0
9221                                        0
0
9222                                        0
0
9223                                        0
0
9224                                        0
```

```
0
9225                                        0
0
9226                                        0
0
9227                                        0
0
9228                                        0
0
9229                                        0
0
9230                                        0
0
9231                                        0
0
9232                                        0
0
9233                                        0
0
9234                                        1
0
9235                                        0
0
9236                                        0
0
9237                                        1
0
9238                                        0
0
9239                                        0
0

      Specialization_e-commerce  \
0                             0
1                             0
2                             0
3                             0
4                             0
5                             0
6                             0
7                             0
8                             0
9                             0
10                            0
11                            0
12                            0
13                            0
14                            0
15                            0
```

```
16                             0
17                             0
18                             0
19                             0
20                             0
21                             0
22                             0
23                             0
24                             0
25                             0
26                             0
27                             0
28                             0
29                             0
...                          ...
9210                           0
9211                           0
9212                           0
9213                           1
9214                           0
9215                           0
9216                           0
9217                           0
9218                           0
9219                           0
9220                           0
9221                           0
9222                           0
9223                           0
9224                           0
9225                           0
9226                           0
9227                           0
9228                           0
9229                           0
9230                           0
9231                           0
9232                           0
9233                           0
9234                           0
9235                           0
9236                           0
9237                           0
9238                           0
9239                           0

                              ...                          \
0                             ...
1                             ...
```

```
2       ...
3       ...
4       ...
5       ...
6       ...
7       ...
8       ...
9       ...
10      ...
11      ...
12      ...
13      ...
14      ...
15      ...
16      ...
17      ...
18      ...
19      ...
20      ...
21      ...
22      ...
23      ...
24      ...
25      ...
26      ...
27      ...
28      ...
29      ...
...     ...
9210    ...
9211    ...
9212    ...
9213    ...
9214    ...
9215    ...
9216    ...
9217    ...
9218    ...
9219    ...
9220    ...
9221    ...
9222    ...
9223    ...
9224    ...
9225    ...
9226    ...
9227    ...
9228    ...
9229    ...
```

```
9230                          ...
9231                          ...
9232                          ...
9233                          ...
9234                          ...
9235                          ...
9236                          ...
9237                          ...
9238                          ...
9239                          ...

      Last Notable Activity_form submitted on website  \
0                                                   0
1                                                   0
2                                                   0
3                                                   0
4                                                   0
5                                                   0
6                                                   0
7                                                   0
8                                                   0
9                                                   0
10                                                  0
11                                                  0
12                                                  0
13                                                  0
14                                                  0
15                                                  0
16                                                  0
17                                                  0
18                                                  0
19                                                  0
20                                                  0
21                                                  0
22                                                  0
23                                                  0
24                                                  0
25                                                  0
26                                                  0
27                                                  0
28                                                  0
29                                                  0
...                                               ...
9210                                                0
9211                                                0
9212                                                0
9213                                                0
9214                                                0
9215                                                0
```

```
9216                                                      0
9217                                                      0
9218                                                      0
9219                                                      0
9220                                                      0
9221                                                      0
9222                                                      0
9223                                                      0
9224                                                      0
9225                                                      0
9226                                                      0
9227                                                      0
9228                                                      0
9229                                                      0
9230                                                      0
9231                                                      0
9232                                                      0
9233                                                      0
9234                                                      0
9235                                                      0
9236                                                      0
9237                                                      0
9238                                                      0
9239                                                      0

      Last Notable Activity_had a phone conversation  \
0                                                    0
1                                                    0
2                                                    0
3                                                    0
4                                                    0
5                                                    0
6                                                    0
7                                                    0
8                                                    0
9                                                    0
10                                                   0
11                                                   0
12                                                   0
13                                                   0
14                                                   0
15                                                   0
16                                                   0
17                                                   0
18                                                   0
19                                                   0
20                                                   0
21                                                   0
22                                                   0
23                                                   0
```

```
24                                                      0
25                                                      0
26                                                      0
27                                                      0
28                                                      0
29                                                      0
...                                                   ...
9210                                                    0
9211                                                    0
9212                                                    0
9213                                                    0
9214                                                    0
9215                                                    0
9216                                                    0
9217                                                    0
9218                                                    0
9219                                                    0
9220                                                    0
9221                                                    0
9222                                                    0
9223                                                    0
9224                                                    0
9225                                                    0
9226                                                    0
9227                                                    0
9228                                                    0
9229                                                    0
9230                                                    0
9231                                                    0
9232                                                    0
9233                                                    0
9234                                                    0
9235                                                    0
9236                                                    0
9237                                                    0
9238                                                    0
9239                                                    0

      Last Notable Activity_modified  \
0                                  1
1                                  0
2                                  0
3                                  1
4                                  1
5                                  1
6                                  1
7                                  1
8                                  0
9                                  0
```

10	0
11	0
12	1
13	0
14	0
15	0
16	1
17	1
18	0
19	1
20	1
21	1
22	1
23	0
24	0
25	1
26	0
27	0
28	0
29	0
...	...
9210	1
9211	0
9212	0
9213	1
9214	0
9215	0
9216	1
9217	0
9218	1
9219	0
9220	1
9221	0
9222	1
9223	1
9224	1
9225	1
9226	1
9227	0
9228	1
9229	1
9230	0
9231	0
9232	0
9233	0
9234	1
9235	0
9236	0
9237	0

```
9238                                   0
9239                                   1

      Last Notable Activity_olark chat conversation  \
0                                                  0
1                                                  0
2                                                  0
3                                                  0
4                                                  0
5                                                  0
6                                                  0
7                                                  0
8                                                  0
9                                                  0
10                                                 0
11                                                 0
12                                                 0
13                                                 0
14                                                 0
15                                                 0
16                                                 0
17                                                 0
18                                                 0
19                                                 0
20                                                 0
21                                                 0
22                                                 0
23                                                 0
24                                                 0
25                                                 0
26                                                 0
27                                                 0
28                                                 0
29                                                 0
...                                              ...
9210                                               0
9211                                               0
9212                                               0
9213                                               0
9214                                               0
9215                                               0
9216                                               0
9217                                               0
9218                                               0
9219                                               0
9220                                               0
9221                                               0
9222                                               0
9223                                               0
```

```
9224                                             0
9225                                             0
9226                                             0
9227                                             0
9228                                             0
9229                                             0
9230                                             0
9231                                             0
9232                                             0
9233                                             0
9234                                             0
9235                                             0
9236                                             0
9237                                             0
9238                                             0
9239                                             0

      Last Notable Activity_page visited on website  \
0                                                0
1                                                0
2                                                0
3                                                0
4                                                0
5                                                0
6                                                0
7                                                0
8                                                0
9                                                0
10                                               0
11                                               1
12                                               0
13                                               0
14                                               0
15                                               0
16                                               0
17                                               0
18                                               1
19                                               0
20                                               0
21                                               0
22                                               0
23                                               0
24                                               0
25                                               0
26                                               0
27                                               0
28                                               0
29                                               0
...                                            ...
```

```
9210                                                    0
9211                                                    0
9212                                                    0
9213                                                    0
9214                                                    0
9215                                                    0
9216                                                    0
9217                                                    0
9218                                                    0
9219                                                    0
9220                                                    0
9221                                                    0
9222                                                    0
9223                                                    0
9224                                                    0
9225                                                    0
9226                                                    0
9227                                                    0
9228                                                    0
9229                                                    0
9230                                                    0
9231                                                    0
9232                                                    0
9233                                                    0
9234                                                    0
9235                                                    0
9236                                                    0
9237                                                    0
9238                                                    0
9239                                                    0

      Last Notable Activity_resubscribed to emails  \
0                                                   0
1                                                   0
2                                                   0
3                                                   0
4                                                   0
5                                                   0
6                                                   0
7                                                   0
8                                                   0
9                                                   0
10                                                  0
11                                                  0
12                                                  0
13                                                  0
14                                                  0
15                                                  0
16                                                  0
```

```
17                                                   0
18                                                   0
19                                                   0
20                                                   0
21                                                   0
22                                                   0
23                                                   0
24                                                   0
25                                                   0
26                                                   0
27                                                   0
28                                                   0
29                                                   0
...                                                ...
9210                                                 0
9211                                                 0
9212                                                 0
9213                                                 0
9214                                                 0
9215                                                 0
9216                                                 0
9217                                                 0
9218                                                 0
9219                                                 0
9220                                                 0
9221                                                 0
9222                                                 0
9223                                                 0
9224                                                 0
9225                                                 0
9226                                                 0
9227                                                 0
9228                                                 0
9229                                                 0
9230                                                 0
9231                                                 0
9232                                                 0
9233                                                 0
9234                                                 0
9235                                                 0
9236                                                 0
9237                                                 0
9238                                                 0
9239                                                 0

      Last Notable Activity_sms sent  Last Notable
Activity_unreachable  \
0                                   0
0
```

```
1                                    0
0
2                                    0
0
3                                    0
0
4                                    0
0
5                                    0
0
6                                    0
0
7                                    0
0
8                                    0
0
9                                    0
0
10                                   0
0
11                                   0
0
12                                   0
0
13                                   0
0
14                                   0
0
15                                   0
0
16                                   0
0
17                                   0
0
18                                   0
0
19                                   0
0
20                                   0
0
21                                   0
0
22                                   0
0
23                                   0
0
24                                   0
0
25                                   0
```

```
0
26                                   0
0
27                                   0
0
28                                   0
0
29                                   0
1
...                                ...                                   ..
.
9210                                 0
0
9211                                 1
0
9212                                 0
0
9213                                 0
0
9214                                 1
0
9215                                 0
0
9216                                 0
0
9217                                 1
0
9218                                 0
0
9219                                 0
0
9220                                 0
0
9221                                 1
0
9222                                 0
0
9223                                 0
0
9224                                 0
0
9225                                 0
0
9226                                 0
0
9227                                 0
0
9228                                 0
0
```

```
9229                                   0
0
9230                                   0
0
9231                                   0
0
9232                                   1
0
9233                                   1
0
9234                                   0
0
9235                                   0
0
9236                                   1
0
9237                                   1
0
9238                                   1
0
9239                                   0
0

      Last Notable Activity_unsubscribed  \
0                                         0
1                                         0
2                                         0
3                                         0
4                                         0
5                                         0
6                                         0
7                                         0
8                                         0
9                                         0
10                                        0
11                                        0
12                                        0
13                                        0
14                                        0
15                                        0
16                                        0
17                                        0
18                                        0
19                                        0
20                                        0
21                                        0
22                                        0
23                                        0
24                                        0
```

```
25                                      0
26                                      0
27                                      0
28                                      0
29                                      0
...                                   ...
9210                                    0
9211                                    0
9212                                    1
9213                                    0
9214                                    0
9215                                    0
9216                                    0
9217                                    0
9218                                    0
9219                                    0
9220                                    0
9221                                    0
9222                                    0
9223                                    0
9224                                    0
9225                                    0
9226                                    0
9227                                    0
9228                                    0
9229                                    0
9230                                    0
9231                                    0
9232                                    0
9233                                    0
9234                                    0
9235                                    0
9236                                    0
9237                                    0
9238                                    0
9239                                    0

      Last Notable Activity_view in browser link clicked
0                                                      0
1                                                      0
2                                                      0
3                                                      0
4                                                      0
5                                                      0
6                                                      0
7                                                      0
8                                                      0
9                                                      0
10                                                     0
```

```
11      0
12      0
13      0
14      0
15      0
16      0
17      0
18      0
19      0
20      0
21      0
22      0
23      0
24      0
25      0
26      0
27      0
28      0
29      0
...    ...
9210    0
9211    0
9212    0
9213    0
9214    0
9215    0
9216    0
9217    0
9218    0
9219    0
9220    0
9221    0
9222    0
9223    0
9224    0
9225    0
9226    0
9227    0
9228    0
9229    0
9230    0
9231    0
9232    0
9233    0
9234    0
9235    0
9236    0
9237    0
9238    0
9239    0
```

```
[9074 rows x 81 columns]
```

4. Test-Train Split

```python
# Import the required library
from sklearn.model_selection import train_test_split

X = df_final_dum.drop(['Converted'], 1)
X.head()
```

```
   TotalVisits  Total Time Spent on Website  Page Views Per Visit  \
0          0.0                            0                   0.0
1          5.0                          674                   2.5
2          2.0                         1532                   2.0
3          1.0                          305                   1.0
4          2.0                         1428                   1.0

   Lead Origin_landing page submission  Lead Origin_lead add form  \
0                                    0                          0
1                                    0                          0
2                                    1                          0
3                                    1                          0
4                                    1                          0

   Lead Origin_lead import  Specialization_business administration  \
0                        0                                       0
1                        0                                       0
2                        0                                       1
3                        0                                       0
4                        0                                       0

   Specialization_e-business  Specialization_e-commerce  \
0                          0                          0
1                          0                          0
2                          0                          0
3                          0                          0
4                          0                          0

   Specialization_finance management  \
0                                  0
1                                  0
2                                  0
3                                  0
4                                  0

                          ...                          \
0                         ...
1                         ...
2                         ...
```

```
3                         ...
4                         ...

   Last Notable Activity_form submitted on website  \
0                                                 0
1                                                 0
2                                                 0
3                                                 0
4                                                 0

   Last Notable Activity_had a phone conversation  \
0                                                0
1                                                0
2                                                0
3                                                0
4                                                0

   Last Notable Activity_modified  \
0                               1
1                               0
2                               0
3                               1
4                               1

   Last Notable Activity_olark chat conversation  \
0                                               0
1                                               0
2                                               0
3                                               0
4                                               0

   Last Notable Activity_page visited on website  \
0                                               0
1                                               0
2                                               0
3                                               0
4                                               0

   Last Notable Activity_resubscribed to emails  \
0                                              0
1                                              0
2                                              0
3                                              0
4                                              0

   Last Notable Activity_sms sent  Last Notable
Activity_unreachable  \
0                               0                                   0

1                               0                                   0
```

```
2                              0                                  0

3                              0                                  0

4                              0                                  0


   Last Notable Activity_unsubscribed  \
0                                    0
1                                    0
2                                    0
3                                    0
4                                    0

   Last Notable Activity_view in browser link clicked
0                                                   0
1                                                   0
2                                                   0
3                                                   0
4                                                   0

[5 rows x 80 columns]
```

```
# Putting the target variable in y
y = df_final_dum['Converted']
y.head()
```

```
0    0
1    0
2    1
3    0
4    1
Name: Converted, dtype: int64
```

```
# Split the dataset into 70% and 30% for train and test respectively
X_train, X_test, y_train, y_test = train_test_split(X, y,
train_size=0.7, test_size=0.3, random_state=10)
```

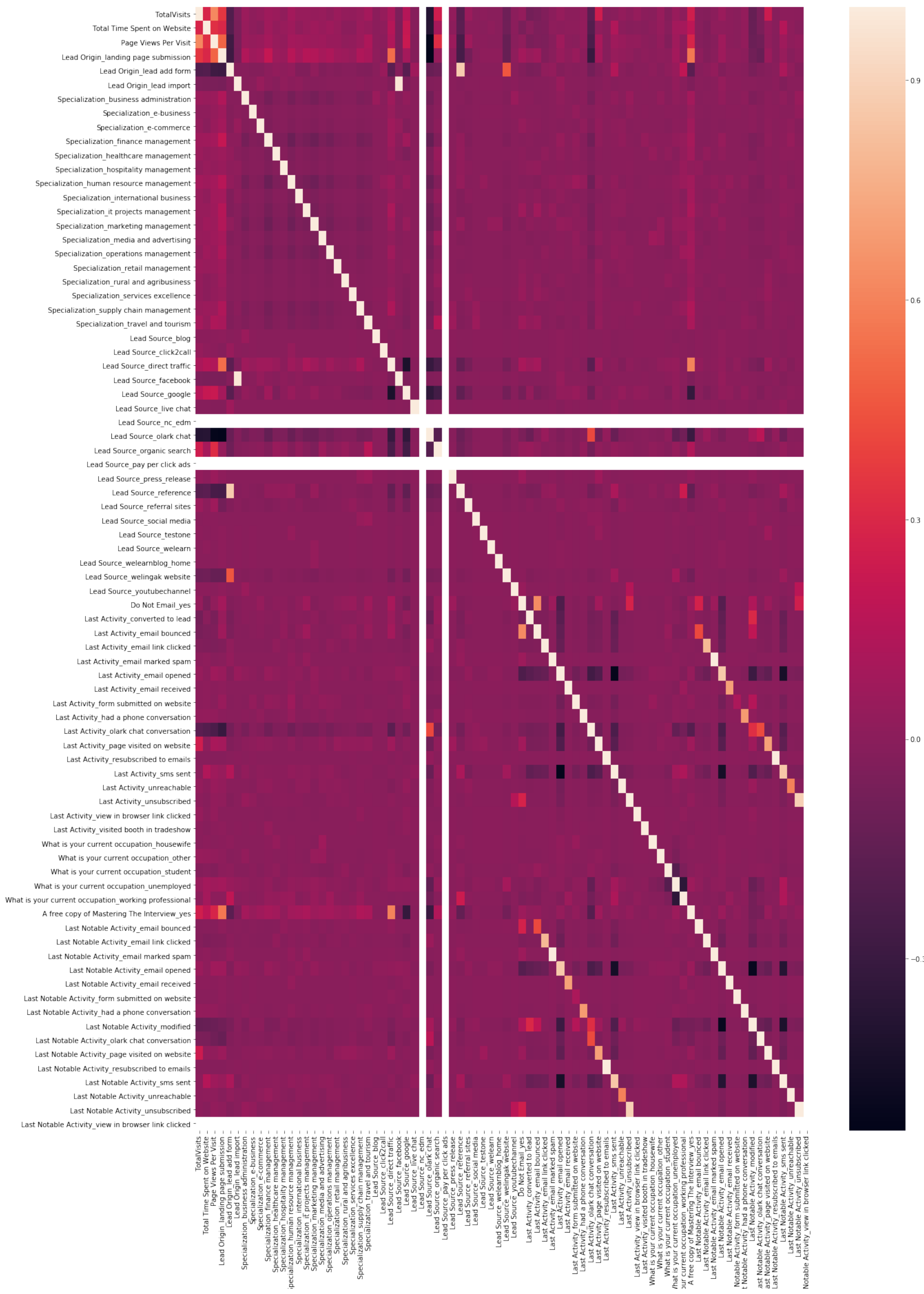
```
# Import MinMax scaler
from sklearn.preprocessing import MinMaxScaler
# Scale the three numeric features
scaler = MinMaxScaler()
X_train[['TotalVisits', 'Page Views Per Visit', 'Total Time Spent on
Website']] = scaler.fit_transform(X_train[['TotalVisits', 'Page Views
Per Visit', 'Total Time Spent on Website']])
X_train.head()
```

```
      TotalVisits  Total Time Spent on Website  Page Views Per
Visit  \
1289     0.014184                     0.612676              0.083333
```

```
3604    0.000000                           0.000000                0.000000

5584    0.042553                           0.751761                0.250000

7679    0.000000                           0.000000                0.000000

7563    0.014184                           0.787852                0.083333


      Lead Origin_landing page submission  Lead Origin_lead add
form  \
1289                                    1                          0

3604                                    0                          0

5584                                    1                          0

7679                                    0                          0

7563                                    1                          0


      Lead Origin_lead import  Specialization_business administration
\
1289                        0                                       0

3604                        0                                       0

5584                        0                                       0

7679                        0                                       0

7563                        0                                       0


      Specialization_e-business  Specialization_e-commerce  \
1289                          0                          0
3604                          0                          0
5584                          0                          0
7679                          0                          0
7563                          0                          0

      Specialization_finance management  \
1289                                  1
3604                                  0
5584                                  0
7679                                  0
7563                                  0

                             ...                            \
```

```
1289                         ...
3604                         ...
5584                         ...
7679                         ...
7563                         ...

      Last Notable Activity_form submitted on website  \
1289                                                 0
3604                                                 0
5584                                                 0
7679                                                 0
7563                                                 0

      Last Notable Activity_had a phone conversation  \
1289                                                0
3604                                                0
5584                                                0
7679                                                0
7563                                                0

      Last Notable Activity_modified  \
1289                                0
3604                                0
5584                                0
7679                                0
7563                                1

      Last Notable Activity_olark chat conversation  \
1289                                               0
3604                                               0
5584                                               0
7679                                               0
7563                                               0

      Last Notable Activity_page visited on website  \
1289                                               0
3604                                               1
5584                                               0
7679                                               0
7563                                               0

      Last Notable Activity_resubscribed to emails  \
1289                                              0
3604                                              0
5584                                              0
7679                                              0
7563                                              0

      Last Notable Activity_sms sent  Last Notable
Activity_unreachable  \
```

```
1289                                    0
0
3604                                    0
0
5584                                    0
0
7679                                    0
0
7563                                    0
0

      Last Notable Activity_unsubscribed  \
1289                                        0
3604                                        0
5584                                        0
7679                                        0
7563                                        0

      Last Notable Activity_view in browser link clicked
1289                                                  0
3604                                                  0
5584                                                  0
7679                                                  0
7563                                                  0

[5 rows x 80 columns]

# To check the correlation among varibles
plt.figure(figsize=(20,30))
sns.heatmap(X_train.corr())
plt.show()
```


TotalVisits
Total Time Spent on Website
Page Views Per Visit
Lead Origin_landing page submission
Lead Origin_lead add form
Lead Origin_lead import
Specialization_business administration
Specialization_e-business
Specialization_e-commerce
Specialization_finance management
Specialization_healthcare management
Specialization_hospitality management
Specialization_human resource management
Specialization_international business
Specialization_it projects management
Specialization_marketing management
Specialization_media and advertising
Specialization_operations management
Specialization_retail management
Specialization_rural and agribusiness
Specialization_services excellence
Specialization_supply chain management
Specialization_travel and tourism
Lead Source_blog
Lead Source_click2call
Lead Source_direct traffic
Lead Source_facebook
Lead Source_google
Lead Source_live chat
Lead Source_nc_edm
Lead Source_olark chat
Lead Source_organic search
Lead Source_pay per click ads
Lead Source_press_release
Lead Source_reference
Lead Source_referral sites
Lead Source_social media
Lead Source_testone
Lead Source_welearn
Lead Source_welearnblog_home
Lead Source_welingak website
Lead Source_youtubechannel
Do Not Email_yes
Last Activity_converted to lead
Last Activity_email bounced
Last Activity_email link clicked
Last Activity_email marked spam
Last Activity_email opened
Last Activity_email received
Last Activity_form submitted on website
Last Activity_had a phone conversation
Last Activity_olark chat conversation
Last Activity_page visited on website
Last Activity_resubscribed to emails
Last Activity_sms sent
Last Activity_unreachable
Last Activity_unsubscribed
Last Activity_view in browser link clicked
Last Activity_visited booth in tradeshow
What is your current occupation_housewife
What is your current occupation_other
What is your current occupation_student
What is your current occupation_unemployed
What is your current occupation_working professional
A free copy of Mastering The Interview_yes
Last Notable Activity_email bounced
Last Notable Activity_email link clicked
Last Notable Activity_email marked spam
Last Notable Activity_email opened
Last Notable Activity_email received
Last Notable Activity_form submitted on website
Last Notable Activity_had a phone conversation
Last Notable Activity_modified
Last Notable Activity_olark chat conversation
Last Notable Activity_page visited on website
Last Notable Activity_resubscribed to emails
Last Notable Activity_sms sent
Last Notable Activity_unreachable
Last Notable Activity_unsubscribed
Last Notable Activity_view in browser link clicked
0.9
0.6
0.3
0.0
−0.3

Since there are a lot of variables it is difficult to drop variable. We'll do it after RFE

5. Model Building

```
# Import 'LogisticRegression'
from sklearn.linear_model import LogisticRegression
logreg = LogisticRegression()

# Import RFE
from sklearn.feature_selection import RFE

# Running RFE with 15 variables as output
rfe = RFE(logreg, 15)
rfe = rfe.fit(X_train, y_train)

# Features that have been selected by RFE
list(zip(X_train.columns, rfe.support_, rfe.ranking_))
```

```
[('TotalVisits', True, 1),
 ('Total Time Spent on Website', True, 1),
 ('Page Views Per Visit', False, 6),
 ('Lead Origin_landing page submission', False, 26),
 ('Lead Origin_lead add form', True, 1),
 ('Lead Origin_lead import', False, 46),
 ('Specialization_business administration', False, 33),
 ('Specialization_e-business', False, 32),
 ('Specialization_e-commerce', False, 23),
 ('Specialization_finance management', False, 30),
 ('Specialization_healthcare management', False, 25),
 ('Specialization_hospitality management', False, 43),
 ('Specialization_human resource management', False, 31),
 ('Specialization_international business', False, 37),
 ('Specialization_it projects management', False, 28),
 ('Specialization_marketing management', False, 22),
 ('Specialization_media and advertising', False, 40),
 ('Specialization_operations management', False, 27),
 ('Specialization_retail management', False, 63),
 ('Specialization_rural and agribusiness', False, 24),
 ('Specialization_services excellence', False, 21),
 ('Specialization_supply chain management', False, 29),
 ('Specialization_travel and tourism', False, 35),
 ('Lead Source_blog', False, 41),
 ('Lead Source_click2call', False, 61),
 ('Lead Source_direct traffic', True, 1),
 ('Lead Source_facebook', False, 45),
 ('Lead Source_google', True, 1),
 ('Lead Source_live chat', False, 48),
 ('Lead Source_nc_edm', False, 64),
 ('Lead Source_olark chat', False, 19),
 ('Lead Source_organic search', True, 1),
 ('Lead Source_pay per click ads', False, 65),
```

```
 ('Lead Source_press_release', False, 51),
 ('Lead Source_reference', False, 18),
 ('Lead Source_referral sites', False, 4),
 ('Lead Source_social media', False, 20),
 ('Lead Source_testone', False, 42),
 ('Lead Source_welearn', False, 49),
 ('Lead Source_welearnblog_home', False, 44),
 ('Lead Source_welingak website', True, 1),
 ('Lead Source_youtubechannel', False, 47),
 ('Do Not Email_yes', True, 1),
 ('Last Activity_converted to lead', False, 14),
 ('Last Activity_email bounced', False, 11),
 ('Last Activity_email link clicked', False, 54),
 ('Last Activity_email marked spam', False, 39),
 ('Last Activity_email opened', False, 58),
 ('Last Activity_email received', False, 56),
 ('Last Activity_form submitted on website', False, 36),
 ('Last Activity_had a phone conversation', False, 5),
 ('Last Activity_olark chat conversation', True, 1),
 ('Last Activity_page visited on website', False, 16),
 ('Last Activity_resubscribed to emails', False, 15),
 ('Last Activity_sms sent', True, 1),
 ('Last Activity_unreachable', False, 17),
 ('Last Activity_unsubscribed', False, 52),
 ('Last Activity_view in browser link clicked', False, 53),
 ('Last Activity_visited booth in tradeshow', False, 55),
 ('What is your current occupation_housewife', True, 1),
 ('What is your current occupation_other', True, 1),
 ('What is your current occupation_student', False, 2),
 ('What is your current occupation_unemployed', False, 3),
 ('What is your current occupation_working professional', True, 1),
 ('A free copy of Mastering The Interview_yes', False, 59),
 ('Last Notable Activity_email bounced', False, 50),
 ('Last Notable Activity_email link clicked', False, 10),
 ('Last Notable Activity_email marked spam', False, 34),
 ('Last Notable Activity_email opened', False, 13),
 ('Last Notable Activity_email received', False, 60),
 ('Last Notable Activity_form submitted on website', False, 57),
 ('Last Notable Activity_had a phone conversation', True, 1),
 ('Last Notable Activity_modified', False, 7),
 ('Last Notable Activity_olark chat conversation', False, 9),
 ('Last Notable Activity_page visited on website', False, 12),
 ('Last Notable Activity_resubscribed to emails', False, 8),
 ('Last Notable Activity_sms sent', False, 62),
 ('Last Notable Activity_unreachable', True, 1),
 ('Last Notable Activity_unsubscribed', False, 38),
 ('Last Notable Activity_view in browser link clicked', False, 66)]
```

```
# Put all the columns selected by RFE in the variable 'col'
col = X_train.columns[rfe.support_]
```

All the variables selected by RFE, next statistics part (p-values and the VIFs).

```
# Selecting columns selected by RFE
X_train = X_train[col]

# Importing statsmodels
import statsmodels.api as sm

X_train_sm = sm.add_constant(X_train)
logm1 = sm.GLM(y_train, X_train_sm, family = sm.families.Binomial())
res = logm1.fit()
res.summary()

<class 'statsmodels.iolib.summary.Summary'>
"""
                 Generalized Linear Model Regression Results

======================================================================
========
Dep. Variable:              Converted   No. Observations:
6351
Model:                            GLM   Df Residuals:
6335
Model Family:                Binomial   Df Model:
15
Link Function:                  logit   Scale:
1.0000
Method:                          IRLS   Log-Likelihood:
-2741.3
Date:            Mon, 10 Jun 2019   Deviance:
5482.6
Time:                    17:10:21   Pearson chi2:
6.64e+03
No. Iterations:                    22   Covariance Type:
nonrobust
======================================================================
==================================================
                                                           coef    std
err          z      P>|z|      [0.025      0.975]
----------------------------------------------------------------------
--------------------------------------------------
const                                                   -1.2524
0.081    -15.450      0.000      -1.411      -1.094
TotalVisits                                              4.5519
1.398      3.256      0.001       1.812       7.292
Total Time Spent on Website                              4.5660
0.162     28.101      0.000       4.248       4.884
Lead Origin_lead add form                                2.6773
0.225     11.916      0.000       2.237       3.118
Lead Source_direct traffic                              -1.4795
```

```
0.114    -12.979      0.000      -1.703      -1.256
Lead Source_google                                              -1.1705
0.109    -10.690      0.000      -1.385      -0.956
Lead Source_organic search                                      -1.2823
0.134     -9.541      0.000      -1.546      -1.019
Lead Source_welingak website                                     2.5984
1.033      2.515      0.012       0.573       4.624
Do Not Email_yes                                                -1.4076
0.168     -8.387      0.000      -1.737      -1.079
Last Activity_olark chat conversation                           -1.4678
0.165     -8.874      0.000      -1.792      -1.144
Last Activity_sms sent                                           1.3213
0.073     18.222      0.000       1.179       1.463
What is your current occupation_housewife                       24.4759
3.07e+04      0.001      0.999   -6.01e+04    6.01e+04
What is your current occupation_other                            1.4134
0.760      1.859      0.063      -0.077       2.904
What is your current occupation_working professional             2.8071
0.193     14.509      0.000       2.428       3.186
Last Notable Activity_had a phone conversation                  24.2053
2.18e+04      0.001      0.999   -4.28e+04    4.28e+04
Last Notable Activity_unreachable                                1.7029
0.610      2.790      0.005       0.507       2.899
======================================================================
==================================================
"""

# Importing 'variance_inflation_factor'
from statsmodels.stats.outliers_influence import
variance_inflation_factor

# Make a VIF dataframe for all the variables present
vif = pd.DataFrame()
vif['Features'] = X_train.columns
vif['VIF'] = [variance_inflation_factor(X_train.values, i) for i in
range(X_train.shape[1])]
vif['VIF'] = round(vif['VIF'], 2)
vif = vif.sort_values(by = "VIF", ascending = False)
vif

                                             Features   VIF
1                         Total Time Spent on Website  2.34
0                                         TotalVisits  2.28
4                                  Lead Source_google  2.04
3                          Lead Source_direct traffic  1.91
5                          Lead Source_organic search  1.60
9                              Last Activity_sms sent  1.49
2                           Lead Origin_lead add form  1.47
6                        Lead Source_welingak website  1.31
12  What is your current occupation_working profes...  1.17
```

```
7                                        Do Not Email_yes  1.10
8                  Last Activity_olark chat conversation  1.02
11                 What is your current occupation_other  1.01
14                     Last Notable Activity_unreachable  1.01
10             What is your current occupation_housewife  1.00
13       Last Notable Activity_had a phone conversation  1.00
```

The VIF values seem fine but the p-values aren't. So removing 'Last Notable Activity had a phone conversation'

```
X_train.drop('Last Notable Activity_had a phone conversation', axis = 
1, inplace = True)

# Refit the model with the new set of features
X_train_sm = sm.add_constant(X_train)
logm2 = sm.GLM(y_train, X_train_sm, family = sm.families.Binomial())
res = logm2.fit()
res.summary()

<class 'statsmodels.iolib.summary.Summary'>
"""
                 Generalized Linear Model Regression Results              

======================================================================
========
Dep. Variable:              Converted   No. Observations:                 
6351
Model:                            GLM   Df Residuals:                     
6336
Model Family:                Binomial   Df Model:                         
14
Link Function:                  logit   Scale:                            
1.0000
Method:                          IRLS   Log-Likelihood:                   
-2749.9
Date:                Mon, 10 Jun 2019   Deviance:                         
5499.7
Time:                        17:10:22   Pearson chi2:                     
6.64e+03
No. Iterations:                    20   Covariance Type:                  
nonrobust
======================================================================
==================================================
                                                          coef    std 
err          z      P>|z|      [0.025      0.975]
----------------------------------------------------------------------
--------------------------------------------------
const                                                  -1.2492     
0.081    -15.422      0.000      -1.408      -1.090
```

```
TotalVisits                                                     4.7231
1.410      3.349      0.001       1.959       7.488
Total Time Spent on Website                                     4.5511
0.162     28.089      0.000       4.234       4.869
Lead Origin_lead add form                                       2.6773
0.225     11.918      0.000       2.237       3.118
Lead Source_direct traffic                                     -1.4795
0.114    -12.987      0.000      -1.703      -1.256
Lead Source_google                                             -1.1600
0.109    -10.611      0.000      -1.374      -0.946
Lead Source_organic search                                     -1.2778
0.134     -9.510      0.000      -1.541      -1.014
Lead Source_welingak website                                    2.5990
1.033      2.515      0.012       0.574       4.624
Do Not Email_yes                                               -1.4113
0.168     -8.413      0.000      -1.740      -1.083
Last Activity_olark chat conversation                          -1.4730
0.165     -8.908      0.000      -1.797      -1.149
Last Activity_sms sent                                          1.3132
0.072     18.136      0.000       1.171       1.455
What is your current occupation_housewife                      22.4667
1.13e+04      0.002      0.998   -2.21e+04    2.21e+04
What is your current occupation_other                           1.4049
0.760      1.848      0.065      -0.085       2.895
What is your current occupation_working professional            2.8013
0.193     14.487      0.000       2.422       3.180
Last Notable Activity_unreachable                               1.6925
0.610      2.774      0.006       0.497       2.888
======================================================================
==================================================
"""

# Make a VIF dataframe for all the variables present
vif = pd.DataFrame()
vif['Features'] = X_train.columns
vif['VIF'] = [variance_inflation_factor(X_train.values, i) for i in
range(X_train.shape[1])]
vif['VIF'] = round(vif['VIF'], 2)
vif = vif.sort_values(by = "VIF", ascending = False)
vif
```

	Features	VIF
1	Total Time Spent on Website	2.34
0	TotalVisits	2.28
4	Lead Source_google	2.04
3	Lead Source_direct traffic	1.91
5	Lead Source_organic search	1.60
9	Last Activity_sms sent	1.49
2	Lead Origin_lead add form	1.47
6	Lead Source_welingak website	1.31

```
12  What is your current occupation_working profes...  1.17
7                                    Do Not Email_yes  1.10
8                 Last Activity_olark chat conversation  1.02
11                What is your current occupation_other  1.01
13                    Last Notable Activity_unreachable  1.01
10            What is your current occupation_housewife  1.00
```

The VIF values seem fine but the p-values aren't. So removing 'What is your current occupation housewife'

```
X_train.drop('What is your current occupation_housewife', axis = 1,
inplace = True)

# Refit the model with the new set of features
X_train_sm = sm.add_constant(X_train)
logm3 = sm.GLM(y_train, X_train_sm, family = sm.families.Binomial())
res = logm3.fit()
res.summary()

<class 'statsmodels.iolib.summary.Summary'>
"""
                 Generalized Linear Model Regression Results

======================================================================
========
Dep. Variable:              Converted   No. Observations:
6351
Model:                            GLM   Df Residuals:
6337
Model Family:                Binomial   Df Model:
13
Link Function:                  logit   Scale:
1.0000
Method:                          IRLS   Log-Likelihood:
-2755.4
Date:                Mon, 10 Jun 2019   Deviance:
5510.8
Time:                        17:10:22   Pearson chi2:
6.65e+03
No. Iterations:                     7   Covariance Type:
nonrobust
======================================================================
==================================================
                                                           coef    std
err          z      P>|z|      [0.025      0.975]
----------------------------------------------------------------------
--------------------------------------------------
const                                                   -1.2461
0.081    -15.396      0.000      -1.405      -1.087
```

```
TotalVisits                                                       4.6490
1.403      3.314      0.001       1.899       7.399
Total Time Spent on Website                                       4.5480
0.162     28.098      0.000       4.231       4.865
Lead Origin_lead add form                                         2.6841
0.224     11.957      0.000       2.244       3.124
Lead Source_direct traffic                                       -1.4736
0.114    -12.954      0.000      -1.697      -1.251
Lead Source_google                                               -1.1551
0.109    -10.580      0.000      -1.369      -0.941
Lead Source_organic search                                       -1.2633
0.134     -9.426      0.000      -1.526      -1.001
Lead Source_welingak website                                      2.5921
1.033      2.509      0.012       0.567       4.617
Do Not Email_yes                                                 -1.4146
0.168     -8.437      0.000      -1.743      -1.086
Last Activity_olark chat conversation                            -1.4765
0.165     -8.932      0.000      -1.800      -1.152
Last Activity_sms sent                                            1.3072
0.072     18.070      0.000       1.165       1.449
What is your current occupation_other                             1.4003
0.760      1.842      0.066      -0.090       2.890
What is your current occupation_working professional      2.7968
0.193     14.467      0.000       2.418       3.176
Last Notable Activity_unreachable                                 1.6871
0.610      2.766      0.006       0.492       2.883
======================================================================
==================================================
"""

# Make a VIF dataframe for all the variables present
vif = pd.DataFrame()
vif['Features'] = X_train.columns
vif['VIF'] = [variance_inflation_factor(X_train.values, i) for i in
range(X_train.shape[1])]
vif['VIF'] = round(vif['VIF'], 2)
vif = vif.sort_values(by = "VIF", ascending = False)
vif
```

	Features	VIF
1	Total Time Spent on Website	2.34
0	TotalVisits	2.28
4	Lead Source_google	2.04
3	Lead Source_direct traffic	1.91
5	Lead Source_organic search	1.60
9	Last Activity_sms sent	1.49
2	Lead Origin_lead add form	1.47
6	Lead Source_welingak website	1.31
11	What is your current occupation_working profes...	1.17
7	Do Not Email_yes	1.10

```
8                Last Activity_olark chat conversation  1.02
10               What is your current occupation_other  1.01
12                   Last Notable Activity_unreachable  1.01
```

The VIF values seem fine but the p-values aren't. So removing 'What is your current occupation other'

```
X_train.drop('What is your current occupation_other', axis = 1,
inplace = True)

# Refit the model with the new set of features
X_train_sm = sm.add_constant(X_train)
logm4 = sm.GLM(y_train, X_train_sm, family = sm.families.Binomial())
res = logm4.fit()
res.summary()
```

```
<class 'statsmodels.iolib.summary.Summary'>
"""
                 Generalized Linear Model Regression Results

======================================================================
========
Dep. Variable:              Converted   No. Observations:
6351
Model:                            GLM   Df Residuals:
6338
Model Family:                Binomial   Df Model:
12
Link Function:                  logit   Scale:
1.0000
Method:                          IRLS   Log-Likelihood:
-2757.3
Date:                Mon, 10 Jun 2019   Deviance:
5514.5
Time:                        17:10:22   Pearson chi2:
6.65e+03
No. Iterations:                     7   Covariance Type:
nonrobust
======================================================================
==================================================
                                                           coef    std
err          z      P>|z|      [0.025      0.975]
----------------------------------------------------------------------
--------------------------------------------------
const                                                   -1.2466
0.081    -15.398      0.000      -1.405      -1.088
TotalVisits                                              4.7586
1.410      3.375      0.001       1.995       7.522
Total Time Spent on Website                              4.5539
```

```
0.162     28.136      0.000       4.237       4.871
Lead Origin_lead add form                                     2.6860
0.224     11.966      0.000       2.246       3.126
Lead Source_direct traffic                                   -1.4706
0.114    -12.929      0.000      -1.694      -1.248
Lead Source_google                                           -1.1564
0.109    -10.588      0.000      -1.370      -0.942
Lead Source_organic search                                   -1.2631
0.134     -9.416      0.000      -1.526      -1.000
Lead Source_welingak website                                  2.5923
1.033      2.509      0.012       0.567       4.617
Do Not Email_yes                                             -1.4186
0.168     -8.461      0.000      -1.747      -1.090
Last Activity_olark chat conversation                        -1.4717
0.165     -8.909      0.000      -1.796      -1.148
Last Activity_sms sent                                        1.3038
0.072     18.031      0.000       1.162       1.445
What is your current occupation_working professional          2.7934
0.193     14.449      0.000       2.414       3.172
Last Notable Activity_unreachable                             1.6837
0.610      2.761      0.006       0.488       2.879
======================================================================
==================================================
"""

# Make a VIF dataframe for all the variables present
vif = pd.DataFrame()
vif['Features'] = X_train.columns
vif['VIF'] = [variance_inflation_factor(X_train.values, i) for i in range(X_train.shape[1])]
vif['VIF'] = round(vif['VIF'], 2)
vif = vif.sort_values(by = "VIF", ascending = False)
vif
```

	Features	VIF
1	Total Time Spent on Website	2.33
0	TotalVisits	2.28
4	Lead Source_google	2.04
3	Lead Source_direct traffic	1.91
5	Lead Source_organic search	1.60
9	Last Activity_sms sent	1.49
2	Lead Origin_lead add form	1.47
6	Lead Source_welingak website	1.31
10	What is your current occupation_working profes...	1.17
7	Do Not Email_yes	1.10
8	Last Activity_olark chat conversation	1.02
11	Last Notable Activity_unreachable	1.01

All the VIF values are good and all the p-values are below 0.05. So we can fix model.

6. Creating Prediction

```
# Predicting the probabilities on the train set
y_train_pred = res.predict(X_train_sm)
y_train_pred[:10]
```

```
1289    0.611739
3604    0.223294
5584    0.425011
7679    0.223294
7563    0.432202
7978    0.732762
7780    0.130274
7863    0.982565
838     0.779231
708     0.132990
dtype: float64
```

```
# Reshaping to an array
y_train_pred = y_train_pred.values.reshape(-1)
y_train_pred[:10]
```

```
array([0.61173868, 0.22329356, 0.42501069, 0.22329356, 0.43220183,
       0.73276232, 0.13027447, 0.9825646 , 0.77923117, 0.13298976])
```

```
# Data frame with given convertion rate and probablity of predicted
ones
y_train_pred_final = pd.DataFrame({'Converted':y_train.values,
'Conversion_Prob':y_train_pred})
y_train_pred_final.head()
```

	Converted	Conversion_Prob
0	1	0.611739
1	0	0.223294
2	0	0.425011
3	0	0.223294
4	0	0.432202

```
# Substituting 0 or 1 with the cut off as 0.5
y_train_pred_final['Predicted'] =
y_train_pred_final.Conversion_Prob.map(lambda x: 1 if x > 0.5 else 0)
y_train_pred_final.head()
```

	Converted	Conversion_Prob	Predicted
0	1	0.611739	1
1	0	0.223294	0
2	0	0.425011	0
3	0	0.223294	0
4	0	0.432202	0

7. Model Evaluation

```
# Importing metrics from sklearn for evaluation
from sklearn import metrics

# Creating confusion matrix
confusion = metrics.confusion_matrix(y_train_pred_final.Converted,
y_train_pred_final.Predicted )
confusion

array([[3403,  492],
       [ 729, 1727]], dtype=int64)

# Predicted     not_churn    churn
# Actual
# not_churn        3403       492
# churn             729      1727

# Check the overall accuracy
metrics.accuracy_score(y_train_pred_final.Converted,
y_train_pred_final.Predicted)

0.807746811525744
```

That's around 81% accuracy with is a very good value

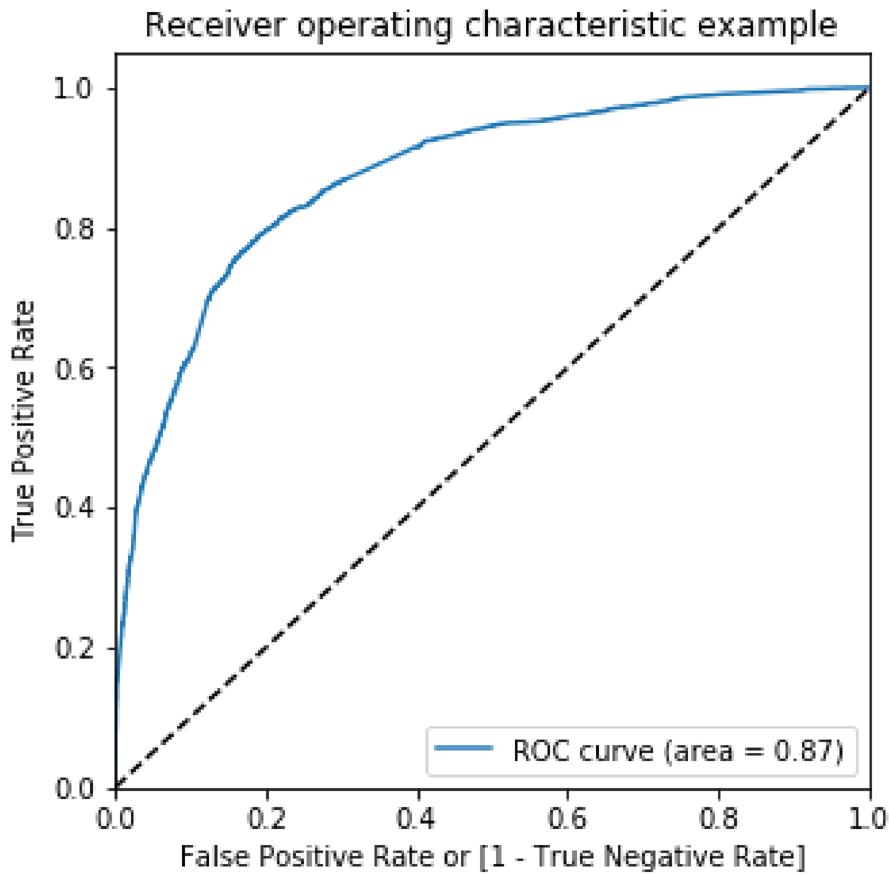
```
# Substituting the value of true positive
TP = confusion[1,1]
# Substituting the value of true negatives
TN = confusion[0,0]
# Substituting the value of false positives
FP = confusion[0,1]
# Substituting the value of false negatives
FN = confusion[1,0]

# Calculating the sensitivity
TP/(TP+FN)

0.7031758957654723

# Calculating the specificity
TN/(TN+FP)

0.8736842105263158
```

With the current cut off as 0.5 we have around 81% accuracy, sensitivity of around 70% and specificity of around 87%.

7. Optimise Cut off (ROC Curve)

The previous cut off was randomely selected. Now to find the optimum one

```python
# ROC function
def draw_roc( actual, probs ):
    fpr, tpr, thresholds = metrics.roc_curve( actual, probs,
                                              drop_intermediate =
False )
    auc_score = metrics.roc_auc_score( actual, probs )
    plt.figure(figsize=(5, 5))
    plt.plot( fpr, tpr, label='ROC curve (area = %0.2f)' % auc_score )
    plt.plot([0, 1], [0, 1], 'k--')
    plt.xlim([0.0, 1.0])
    plt.ylim([0.0, 1.05])
    plt.xlabel('False Positive Rate or [1 - True Negative Rate]')
    plt.ylabel('True Positive Rate')
    plt.title('Receiver operating characteristic example')
    plt.legend(loc="lower right")
    plt.show()

    return None

fpr, tpr, thresholds =
metrics.roc_curve( y_train_pred_final.Converted,
y_train_pred_final.Conversion_Prob, drop_intermediate = False )

# Call the ROC function
draw_roc(y_train_pred_final.Converted,
y_train_pred_final.Conversion_Prob)
```



The area under ROC curve is 0.87 which is a very good value.

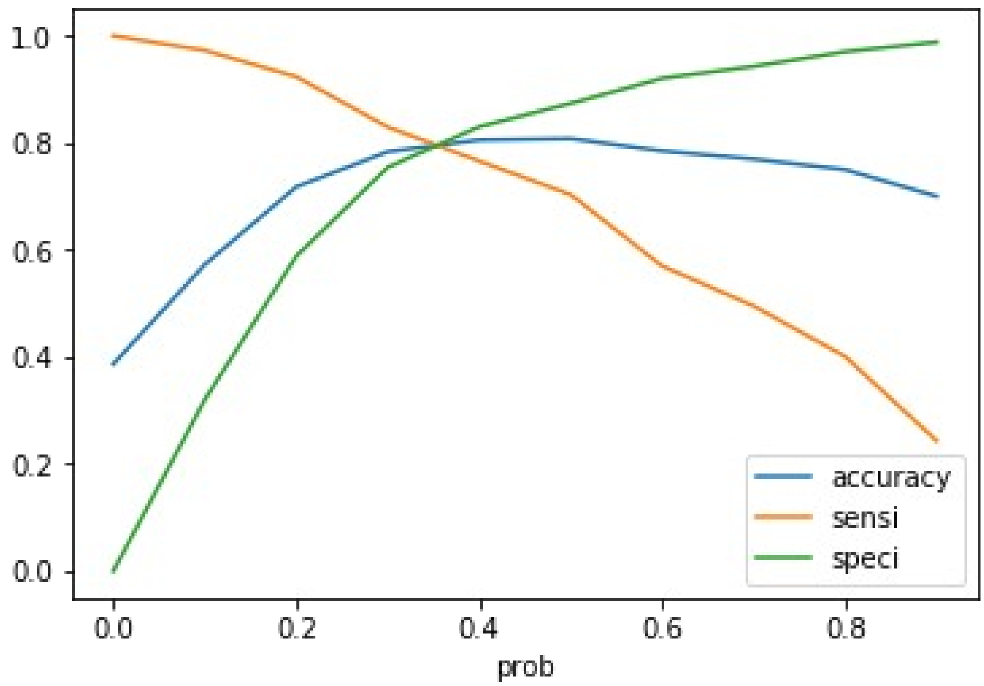
```
# Creating columns with different probability cutoffs
numbers = [float(x)/10 for x in range(10)]
for i in numbers:
    y_train_pred_final[i]=
y_train_pred_final.Conversion_Prob.map(lambda x: 1 if x > i else 0)
y_train_pred_final.head()

   Converted  Conversion_Prob  Predicted  0.0  0.1  0.2  0.3  0.4  0.5
0.6  \
0          1         0.611739          1    1    1    1    1    1    1
1
1          0         0.223294          0    1    1    1    0    0    0
0
2          0         0.425011          0    1    1    1    1    1    0
0
3          0         0.223294          0    1    1    1    0    0    0
0
4          0         0.432202          0    1    1    1    1    1    0
0

   0.7  0.8  0.9
```

0	0	0	0
1	0	0	0
2	0	0	0
3	0	0	0
4	0	0	0

```
# Creating a dataframe to see the values of accuracy, sensitivity, and
specificity at different values of probabiity cutoffs
cutoff_df = pd.DataFrame( columns =
['prob','accuracy','sensi','speci'])
# Making confusing matrix to find values of sensitivity, accurace and
specificity for each level of probablity
from sklearn.metrics import confusion_matrix
num = [0.0,0.1,0.2,0.3,0.4,0.5,0.6,0.7,0.8,0.9]
for i in num:
    cm1 = metrics.confusion_matrix(y_train_pred_final.Converted,
y_train_pred_final[i] )
    total1=sum(sum(cm1))
    accuracy = (cm1[0,0]+cm1[1,1])/total1

    speci = cm1[0,0]/(cm1[0,0]+cm1[0,1])
    sensi = cm1[1,1]/(cm1[1,0]+cm1[1,1])
    cutoff_df.loc[i] =[ i ,accuracy,sensi,speci]
cutoff_df
```

	prob	accuracy	sensi	speci
0.0	0.0	0.386711	1.000000	0.000000
0.1	0.1	0.572508	0.972720	0.320154
0.2	0.2	0.717840	0.923453	0.588190
0.3	0.3	0.783341	0.829397	0.754300
0.4	0.4	0.805228	0.765879	0.830039
0.5	0.5	0.807747	0.703176	0.873684
0.6	0.6	0.784758	0.569625	0.920411
0.7	0.7	0.769643	0.495114	0.942747
0.8	0.8	0.749961	0.400651	0.970218
0.9	0.9	0.700205	0.243485	0.988190

```
# Plotting it
cutoff_df.plot.line(x='prob', y=['accuracy','sensi','speci'])
plt.show()
```



From the graph it is visible that the optimal cut off is at 0.35.

```
y_train_pred_final['final_predicted'] =
y_train_pred_final.Conversion_Prob.map( lambda x: 1 if x > 0.35 else
0)
y_train_pred_final.head()

   Converted  Conversion_Prob  Predicted  0.0  0.1  0.2  0.3  0.4  0.5
0.6  \
0          1         0.611739          1    1    1    1    1    1    1
1
1          0         0.223294          0    1    1    1    0    0    0
0
2          0         0.425011          0    1    1    1    1    1    0
0
3          0         0.223294          0    1    1    1    0    0    0
0
4          0         0.432202          0    1    1    1    1    1    0
0

   0.7  0.8  0.9  final_predicted
0    0    0    0                1
1    0    0    0                0
2    0    0    0                1
3    0    0    0                0
4    0    0    0                1
```

```
# Check the overall accuracy
metrics.accuracy_score(y_train_pred_final.Converted,
y_train_pred_final.final_predicted)
```

```
0.7967249252086286
```

```
# Creating confusion matrix
confusion2 = metrics.confusion_matrix(y_train_pred_final.Converted,
y_train_pred_final.final_predicted )
confusion2
```

```
array([[3097,  798],
       [ 493, 1963]], dtype=int64)
```

```
# Substituting the value of true positive
TP = confusion2[1,1]
# Substituting the value of true negatives
TN = confusion2[0,0]
# Substituting the value of false positives
FP = confusion2[0,1]
# Substituting the value of false negatives
FN = confusion2[1,0]
```

```
# Calculating the sensitivity
TP/(TP+FN)
```

```
0.7992671009771987
```

```
# Calculating the specificity
TN/(TN+FP)
```

```
0.7951219512195122
```

With the current cut off as 0.35 we have accuracy, sensitivity and specificity of around 80%.

8. Prediction on Test set

```
# Scaling numeric values
X_test[['TotalVisits', 'Page Views Per Visit', 'Total Time Spent on
Website']] = scaler.transform(X_test[['TotalVisits', 'Page Views Per
Visit', 'Total Time Spent on Website']])
```

```
# Substituting all the columns in the final train model
col = X_train.columns
```

```
# Select the columns in X_train for X_test as well
X_test = X_test[col]
# Add a constant to X_test
X_test_sm = sm.add_constant(X_test[col])
X_test_sm
X_test_sm
```

	const	TotalVisits	Total Time Spent on Website	\
8308	1.0	0.035461	0.416813	
7212	1.0	0.028369	0.001320	
2085	1.0	0.000000	0.000000	
4048	1.0	0.028369	0.617077	
4790	1.0	0.028369	0.005282	
8552	1.0	0.063830	0.552817	
2232	1.0	0.021277	0.496919	
5259	1.0	0.000000	0.000000	
2399	1.0	0.028369	0.639085	
8018	1.0	0.000000	0.000000	
3221	1.0	0.014184	0.394366	
1226	1.0	0.000000	0.000000	
8914	1.0	0.014184	0.000880	
765	1.0	0.007092	0.041373	
2973	1.0	0.028369	0.199384	
3917	1.0	0.021277	0.615757	
2201	1.0	0.035461	0.056778	
8088	1.0	0.000000	0.000000	
3192	1.0	0.014184	0.462588	
6636	1.0	0.035461	0.745599	
2542	1.0	0.035461	0.532130	
6095	1.0	0.028369	0.783011	
9217	1.0	0.000000	0.000000	
5664	1.0	0.035461	0.700704	
4967	1.0	0.014184	0.097271	
5889	1.0	0.063830	0.437940	
4758	1.0	0.028369	0.476232	
4999	1.0	0.028369	0.134243	
2734	1.0	0.000000	0.000000	
653	1.0	0.028369	0.093750	
...	...	...	...	
124	1.0	0.028369	0.720951	
2172	1.0	0.085106	0.365317	
8016	1.0	0.049645	0.042254	
1681	1.0	0.000000	0.000000	
1593	1.0	0.014184	0.641725	
7103	1.0	0.000000	0.000000	
2603	1.0	0.028369	0.552377	
8331	1.0	0.007092	0.068662	
2711	1.0	0.021277	0.224472	
3141	1.0	0.000000	0.000000	
3847	1.0	0.049645	0.478433	
301	1.0	0.028369	0.261884	
7883	1.0	0.014184	0.308099	
4182	1.0	0.014184	0.227113	
3071	1.0	0.028369	0.204225	
6790	1.0	0.078014	0.140405	
5404	1.0	0.021277	0.132042	
1411	1.0	0.028369	0.555018	

```
2141    1.0     0.014184                             0.190581
97      1.0     0.000000                             0.000000
7796    1.0     0.035461                             0.100352
2453    1.0     0.035461                             0.169014
8639    1.0     0.035461                             0.084067
4039    1.0     0.000000                             0.000000
7311    1.0     0.014184                             0.186180
3261    1.0     0.000000                             0.000000
8179    1.0     0.170213                             0.148768
6236    1.0     0.000000                             0.000000
5240    1.0     0.078014                             0.458627
7243    1.0     0.035461                             0.499560

      Lead Origin_lead add form  Lead Source_direct traffic  \
8308                          0                           1
7212                          0                           0
2085                          1                           0
4048                          0                           1
4790                          0                           1
8552                          0                           1
2232                          0                           0
5259                          0                           0
2399                          0                           0
8018                          0                           0
3221                          0                           0
1226                          0                           0
8914                          0                           1
765                           0                           0
2973                          0                           1
3917                          0                           0
2201                          0                           1
8088                          1                           0
3192                          0                           0
6636                          0                           1
2542                          0                           0
6095                          0                           0
9217                          0                           0
5664                          0                           0
4967                          0                           0
5889                          0                           0
4758                          0                           0
4999                          0                           0
2734                          1                           0
653                           0                           0
...                         ...                         ...
124                           0                           0
2172                          0                           0
8016                          0                           0
1681                          0                           0
1593                          0                           0
```

```
7103                            0                            0
2603                            0                            0
8331                            0                            1
2711                            0                            0
3141                            0                            0
3847                            0                            0
301                             0                            0
7883                            0                            0
4182                            0                            1
3071                            0                            0
6790                            0                            0
5404                            0                            1
1411                            0                            0
2141                            0                            0
97                              0                            0
7796                            0                            0
2453                            0                            0
8639                            0                            0
4039                            0                            0
7311                            0                            1
3261                            0                            0
8179                            0                            0
6236                            0                            0
5240                            0                            0
7243                            0                            0

      Lead Source_google  Lead Source_organic search  \
8308                   0                           0
7212                   0                           1
2085                   0                           0
4048                   0                           0
4790                   0                           0
8552                   0                           0
2232                   0                           1
5259                   0                           0
2399                   1                           0
8018                   0                           0
3221                   1                           0
1226                   0                           0
8914                   0                           0
765                    1                           0
2973                   0                           0
3917                   1                           0
2201                   0                           0
8088                   0                           0
3192                   1                           0
6636                   0                           0
2542                   1                           0
6095                   1                           0
9217                   0                           0
```

```
5664                     1                            0
4967                     1                            0
5889                     0                            1
4758                     0                            1
4999                     1                            0
2734                     0                            0
653                      0                            0
...                    ...                          ...
124                      1                            0
2172                     1                            0
8016                     0                            1
1681                     0                            0
1593                     1                            0
7103                     0                            0
2603                     0                            1
8331                     0                            0
2711                     1                            0
3141                     0                            0
3847                     0                            1
301                      1                            0
7883                     1                            0
4182                     0                            0
3071                     1                            0
6790                     0                            1
5404                     0                            0
1411                     1                            0
2141                     1                            0
97                       0                            0
7796                     1                            0
2453                     1                            0
8639                     1                            0
4039                     0                            0
7311                     0                            0
3261                     0                            0
8179                     1                            0
6236                     0                            0
5240                     1                            0
7243                     0                            1

      Lead Source_welingak website  Do Not Email_yes  \
8308                             0                 0
7212                             0                 0
2085                             1                 0
4048                             0                 0
4790                             0                 0
8552                             0                 0
2232                             0                 1
5259                             0                 0
2399                             0                 1
8018                             0                 0
```

3221	0	0
1226	0	0
8914	0	0
765	0	0
2973	0	0
3917	0	0
2201	0	0
8088	0	0
3192	0	0
6636	0	0
2542	0	0
6095	0	0
9217	0	0
5664	0	0
4967	0	0
5889	0	0
4758	0	0
4999	0	0
2734	0	0
653	0	0
...	...	...
124	0	0
2172	0	0
8016	0	0
1681	0	0
1593	0	0
7103	0	0
2603	0	0
8331	0	1
2711	0	0
3141	0	0
3847	0	0
301	0	0
7883	0	0
4182	0	0
3071	0	0
6790	0	1
5404	0	1
1411	0	0
2141	0	0
97	0	0
7796	0	1
2453	0	0
8639	0	0
4039	0	0
7311	0	0
3261	0	0
8179	0	0
6236	0	0
5240	0	0

```
7243                                  0                  0

      Last Activity_olark chat conversation  Last Activity_sms sent  \
8308                                      0                       0
7212                                      0                       1
2085                                      0                       0
4048                                      0                       1
4790                                      0                       0
8552                                      0                       1
2232                                      0                       0
5259                                      0                       0
2399                                      0                       1
8018                                      1                       0
3221                                      0                       1
1226                                      1                       0
8914                                      0                       0
765                                       0                       0
2973                                      0                       0
3917                                      0                       1
2201                                      1                       0
8088                                      0                       1
3192                                      0                       0
6636                                      0                       0
2542                                      0                       0
6095                                      0                       0
9217                                      0                       1
5664                                      0                       1
4967                                      1                       0
5889                                      0                       0
4758                                      0                       0
4999                                      0                       0
2734                                      0                       0
653                                       0                       0
...                                     ...                     ...
124                                       0                       0
2172                                      0                       0
8016                                      0                       1
1681                                      0                       0
1593                                      0                       0
7103                                      0                       0
2603                                      0                       1
8331                                      0                       0
2711                                      0                       0
3141                                      0                       0
3847                                      0                       1
301                                       0                       0
7883                                      0                       0
4182                                      0                       0
3071                                      0                       1
6790                                      0                       0
```

```
5404                                      0                        1
1411                                      0                        0
2141                                      0                        0
97                                        1                        0
7796                                      0                        1
2453                                      0                        0
8639                                      0                        0
4039                                      0                        0
7311                                      0                        0
3261                                      1                        0
8179                                      0                        1
6236                                      0                        0
5240                                      0                        1
7243                                      0                        0

      What is your current occupation_working professional  \
8308                                                    0
7212                                                    1
2085                                                    0
4048                                                    0
4790                                                    0
8552                                                    0
2232                                                    0
5259                                                    0
2399                                                    0
8018                                                    0
3221                                                    0
1226                                                    0
8914                                                    0
765                                                     0
2973                                                    0
3917                                                    0
2201                                                    0
8088                                                    1
3192                                                    1
6636                                                    0
2542                                                    1
6095                                                    0
9217                                                    0
5664                                                    0
4967                                                    1
5889                                                    0
4758                                                    0
4999                                                    0
2734                                                    0
653                                                     0
...                                                   ...
124                                                     0
2172                                                    0
8016                                                    0
```

1681	0
1593	0
7103	0
2603	0
8331	0
2711	0
3141	0
3847	0
301	0
7883	1
4182	1
3071	0
6790	0
5404	0
1411	0
2141	0
97	0
7796	0
2453	0
8639	0
4039	0
7311	0
3261	0
8179	0
6236	0
5240	0
7243	0

	Last Notable Activity_unreachable
8308	0
7212	0
2085	0
4048	0
4790	0
8552	0
2232	0
5259	0
2399	0
8018	0
3221	0
1226	0
8914	0
765	0
2973	0
3917	0
2201	0
8088	0
3192	0
6636	0
2542	0

```
6095                                       0
9217                                       0
5664                                       0
4967                                       0
5889                                       0
4758                                       0
4999                                       0
2734                                       0
653                                        0
...                                      ...
124                                        0
2172                                       0
8016                                       0
1681                                       0
1593                                       0
7103                                       0
2603                                       0
8331                                       0
2711                                       0
3141                                       0
3847                                       0
301                                        0
7883                                       0
4182                                       0
3071                                       0
6790                                       0
5404                                       0
1411                                       0
2141                                       0
97                                         0
7796                                       0
2453                                       0
8639                                       0
4039                                       0
7311                                       0
3261                                       0
8179                                       0
6236                                       0
5240                                       0
7243                                       0

[2723 rows x 13 columns]

# Storing prediction of test set in the variable 'y_test_pred'
y_test_pred = res.predict(X_test_sm)
# Coverting it to df
y_pred_df = pd.DataFrame(y_test_pred)
# Converting y_test to dataframe
y_test_df = pd.DataFrame(y_test)
# Remove index for both dataframes to append them side by side
```

```
y_pred_df.reset_index(drop=True, inplace=True)
y_test_df.reset_index(drop=True, inplace=True)
# Append y_test_df and y_pred_df
y_pred_final = pd.concat([y_test_df, y_pred_df],axis=1)
# Renaming column
y_pred_final= y_pred_final.rename(columns = {0 : 'Conversion_Prob'})
y_pred_final.head()
```

	Converted	Conversion_Prob
0	0	0.342925
1	1	0.849219
2	1	0.982565
3	1	0.822258
4	0	0.071883

```
# Making prediction using cut off 0.35
y_pred_final['final_predicted'] =
y_pred_final.Conversion_Prob.map(lambda x: 1 if x > 0.35 else 0)
y_pred_final
```

	Converted	Conversion_Prob	final_predicted
0	0	0.342925	0
1	1	0.849219	1
2	1	0.982565	1
3	1	0.822258	1
4	0	0.071883	0
5	1	0.803423	1
6	0	0.173071	0
7	1	0.223294	0
8	1	0.628924	1
9	0	0.061901	0
10	1	0.682271	1
11	0	0.061901	0
12	0	0.066257	0
13	0	0.101488	0
14	0	0.157866	0
15	1	0.858899	1
16	0	0.022718	0
17	1	0.996075	1
18	1	0.928541	1
19	1	0.699933	1
20	1	0.951774	1
21	1	0.785467	1
22	0	0.514295	1
23	0	0.905554	1
24	0	0.361032	1
25	1	0.447308	1
26	1	0.448702	1
27	0	0.160215	0
28	1	0.808364	1

```
29           0        0.335224                0
...        ...             ...              ...
2693         0        0.734036                1
2694         1        0.417184                1
2695         0        0.314921                0
2696         0        0.223294                0
2697         0        0.642657                1
2698         0        0.223294                0
2699         0        0.809167                1
2700         0        0.022112                0
2701         0        0.217639                0
2702         0        0.223294                0
2703         1        0.770143                1
2704         1        0.254385                0
2705         1        0.865408                1
2706         0        0.764576                1
2707         0        0.491460                1
2708         0        0.051292                0
2709         0        0.106272                0
2710         1        0.564526                1
2711         0        0.187313                0
2712         0        0.061901                0
2713         0        0.131017                0
2714         0        0.187775                0
2715         0        0.135711                0
2716         0        0.223294                0
2717         0        0.141629                0
2718         1        0.061901                0
2719         0        0.595864                1
2720         0        0.223294                0
2721         1        0.795858                1
2722         1        0.483521                1

[2723 rows x 3 columns]
```

```python
# Check the overall accuracy
metrics.accuracy_score(y_pred_final['Converted'],
y_pred_final.final_predicted)
```

```
0.8005875872199779
```

```python
# Creating confusion matrix
confusion2 = metrics.confusion_matrix(y_pred_final['Converted'],
y_pred_final.final_predicted )
confusion2
```

```
array([[1394,  350],
       [ 193,  786]], dtype=int64)
```

```
# Substituting the value of true positive
TP = confusion2[1,1]
# Substituting the value of true negatives
TN = confusion2[0,0]
# Substituting the value of false positives
FP = confusion2[0,1]
# Substituting the value of false negatives
FN = confusion2[1,0]

# Calculating the sensitivity
TP/(TP+FN)

0.8028600612870276

# Calculating the specificity
TN/(TN+FP)

0.7993119266055045
```

With the current cut off as 0.35 we have accuracy, sensitivity and specificity of around 80%.

9. Precision-Recall

```
confusion = metrics.confusion_matrix(y_train_pred_final.Converted,
y_train_pred_final.Predicted )
confusion

array([[3403,  492],
       [ 729, 1727]], dtype=int64)

# Precision = TP / TP + FP
confusion[1,1]/(confusion[0,1]+confusion[1,1])

0.7782785038305543

#Recall = TP / TP + FN
confusion[1,1]/(confusion[1,0]+confusion[1,1])

0.7031758957654723
```

With the current cut off as 0.35 we have Precision around 78% and Recall around 70%

9.1. Precision and recall tradeoff

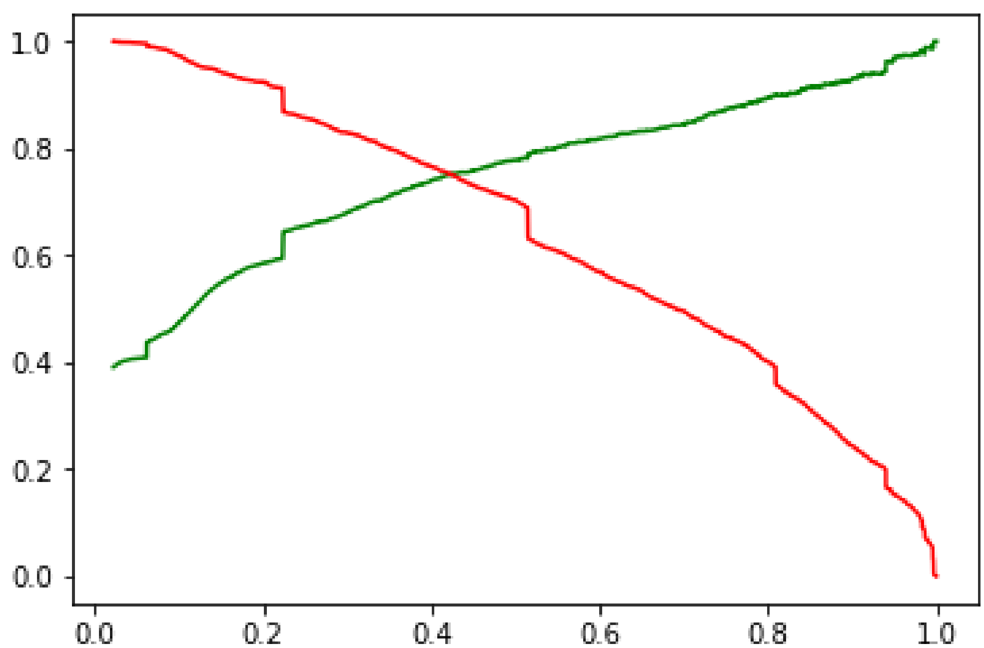
```
from sklearn.metrics import precision_recall_curve

y_train_pred_final.Converted, y_train_pred_final.Predicted

(0        1
 1        0
 2        0
 3        0
```

```
4       0
5       1
6       1
7       1
8       1
9       0
10      1
11      1
12      1
13      0
14      1
15      0
16      0
17      1
18      1
19      0
20      1
21      1
22      0
23      1
24      0
25      0
26      1
27      0
28      0
29      0
       ..
6321    1
6322    1
6323    0
6324    0
6325    0
6326    0
6327    0
6328    1
6329    1
6330    1
6331    1
6332    0
6333    0
6334    1
6335    0
6336    0
6337    1
6338    0
6339    0
6340    1
6341    0
6342    0
```

```
6343    0
6344    1
6345    1
6346    0
6347    0
6348    0
6349    0
6350    1
Name: Converted, Length: 6351, dtype: int64, 0       1
1       0
2       0
3       0
4       0
5       1
6       0
7       1
8       1
9       0
10      1
11      1
12      0
13      0
14      0
15      0
16      0
17      1
18      0
19      0
20      0
21      1
22      0
23      0
24      0
25      0
26      1
27      0
28      0
29      0
       ..
6321    1
6322    1
6323    0
6324    0
6325    0
6326    0
6327    0
6328    0
6329    0
6330    0
```

```
6331    0
6332    0
6333    0
6334    1
6335    0
6336    0
6337    1
6338    0
6339    0
6340    1
6341    0
6342    0
6343    0
6344    0
6345    1
6346    0
6347    0
6348    0
6349    0
6350    0
Name: Predicted, Length: 6351, dtype: int64)
```

```
p, r, thresholds =
precision_recall_curve(y_train_pred_final.Converted,
y_train_pred_final.Conversion_Prob)

plt.plot(thresholds, p[:-1], "g-")
plt.plot(thresholds, r[:-1], "r-")
plt.show()
```



```
y_train_pred_final['final_predicted'] =
y_train_pred_final.Conversion_Prob.map(lambda x: 1 if x > 0.41 else 0)
y_train_pred_final.head()
```

```
   Converted  Conversion_Prob  Predicted  0.0  0.1  0.2  0.3  0.4  0.5
0.6  \
0          1         0.611739          1    1    1    1    1    1    1
1
1          0         0.223294          0    1    1    1    0    0    0
0
2          0         0.425011          0    1    1    1    1    1    0
0
3          0         0.223294          0    1    1    1    0    0    0
0
4          0         0.432202          0    1    1    1    1    1    0
0

   0.7  0.8  0.9  final_predicted
0    0    0    0                1
1    0    0    0                0
2    0    0    0                1
3    0    0    0                0
4    0    0    0                1
```

```
# Accuracy
metrics.accuracy_score(y_train_pred_final.Converted,
y_train_pred_final.final_predicted)
```

```
0.8060148008187688
```

```
# Creating confusion matrix again
confusion2 = metrics.confusion_matrix(y_train_pred_final.Converted,
y_train_pred_final.final_predicted )
confusion2
```

```
array([[3256,  639],
       [ 593, 1863]], dtype=int64)
```

```
# Substituting the value of true positive
TP = confusion2[1,1]
# Substituting the value of true negatives
TN = confusion2[0,0]
# Substituting the value of false positives
FP = confusion2[0,1]
# Substituting the value of false negatives
FN = confusion2[1,0]
```

```
# Precision = TP / TP + FP
TP / (TP + FP)
```

```
0.7446043165467626
```

```
#Recall = TP / TP + FN
TP / (TP + FN)

0.7585504885993485
```

With the current cut off as 0.41 we have Precision around 74% and Recall around 76%

10. Prediction on Test set

```
# Storing prediction of test set in the variable 'y_test_pred'
y_test_pred = res.predict(X_test_sm)
# Coverting it to df
y_pred_df = pd.DataFrame(y_test_pred)
# Converting y_test to dataframe
y_test_df = pd.DataFrame(y_test)
# Remove index for both dataframes to append them side by side
y_pred_df.reset_index(drop=True, inplace=True)
y_test_df.reset_index(drop=True, inplace=True)
# Append y_test_df and y_pred_df
y_pred_final = pd.concat([y_test_df, y_pred_df],axis=1)
# Renaming column
y_pred_final= y_pred_final.rename(columns = {0 : 'Conversion_Prob'})
y_pred_final.head()
```

	Converted	Conversion_Prob
0	0	0.342925
1	1	0.849219
2	1	0.982565
3	1	0.822258
4	0	0.071883

```
# Making prediction using cut off 0.41
y_pred_final['final_predicted'] =
y_pred_final.Conversion_Prob.map(lambda x: 1 if x > 0.41 else 0)
y_pred_final
```

	Converted	Conversion_Prob	final_predicted
0	0	0.342925	0
1	1	0.849219	1
2	1	0.982565	1
3	1	0.822258	1
4	0	0.071883	0
5	1	0.803423	1
6	0	0.173071	0
7	1	0.223294	0
8	1	0.628924	1
9	0	0.061901	0
10	1	0.682271	1
11	0	0.061901	0
12	0	0.066257	0

13	0	0.101488	0
14	0	0.157866	0
15	1	0.858899	1
16	0	0.022718	0
17	1	0.996075	1
18	1	0.928541	1
19	1	0.699933	1
20	1	0.951774	1
21	1	0.785467	1
22	0	0.514295	1
23	0	0.905554	1
24	0	0.361032	0
25	1	0.447308	1
26	1	0.448702	1
27	0	0.160215	0
28	1	0.808364	1
29	0	0.335224	0
...	...	...	...
2693	0	0.734036	1
2694	1	0.417184	1
2695	0	0.314921	0
2696	0	0.223294	0
2697	0	0.642657	1
2698	0	0.223294	0
2699	0	0.809167	1
2700	0	0.022112	0
2701	0	0.217639	0
2702	0	0.223294	0
2703	1	0.770143	1
2704	1	0.254385	0
2705	1	0.865408	1
2706	0	0.764576	1
2707	0	0.491460	1
2708	0	0.051292	0
2709	0	0.106272	0
2710	1	0.564526	1
2711	0	0.187313	0
2712	0	0.061901	0
2713	0	0.131017	0
2714	0	0.187775	0
2715	0	0.135711	0
2716	0	0.223294	0
2717	0	0.141629	0
2718	1	0.061901	0
2719	0	0.595864	1
2720	0	0.223294	0
2721	1	0.795858	1
2722	1	0.483521	1

```
[2723 rows x 3 columns]

# Check the overall accuracy
metrics.accuracy_score(y_pred_final['Converted'],
y_pred_final.final_predicted)

0.808666911494675

# Creating confusion matrix
confusion2 = metrics.confusion_matrix(y_pred_final['Converted'],
y_pred_final.final_predicted )
confusion2

array([[1465,  279],
       [ 242,  737]], dtype=int64)

# Substituting the value of true positive
TP = confusion2[1,1]
# Substituting the value of true negatives
TN = confusion2[0,0]
# Substituting the value of false positives
FP = confusion2[0,1]
# Substituting the value of false negatives
FN = confusion2[1,0]

# Precision = TP / TP + FP
TP / (TP + FP)

0.7253937007874016

#Recall = TP / TP + FN
TP / (TP + FN)

0.7528089887640449
```

With the current cut off as 0.41 we have Precision around 73% and Recall around 75%

Conclusion

It was found that the variables that mattered the most in the potential buyers are (In descending order) :

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

1. When the last activity was:

a. SMS

b. Olark chat conversation

1. When the lead origin is Lead add format.
2. When their current occupation is as a working professional. Keeping these in mind the X Education can flourish as they have a very high chance to get almost all the potential buyers to change their mind and buy their courses.