Lead Score

Problem Statement

- X Education is an Edtech company which provides online courses for industry professionals.
- If any person shows intrest in their courses any provide X Education with their personal information, it is termed as leads.
- Company collects leads from different online plateforms and then its employees try to connect with these leads.
- In this process about 30% of the leads opt for any of the courses available any gets converted into a paying customer.
- Identify those leads who have high conversion probability.
- Provide lead score to them such that person with high lead score have higher chances to convert into a paying customer. while person with lower score have less chances to convert.

Goals and Objectives

There are quite a few goals for this case study.

Build a logistic regression model to assign a lead score between 0 and 100 to each of the leads which can be used by the company to target potential leads. A higher score would mean that the lead is hot, i.e. is most likely to convert whereas a lower score would mean that the lead is cold and will mostly not get converted. There are some more problems presented by the company which your model should be able to adjust to if the company's requirement changes in the future so you will need to handle these as well. These problems are provided in a separate doc file. Please fill it based on the logistic regression model you got in the first step. Also, make sure you include this in your final PPT where you'll make recommendations.

```
#Importing required packages
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns

from sklearn.model_selection import train_test_split
from sklearn.preprocessing import MinMaxScaler
import statsmodels.api as sm
from statsmodels.stats.outliers_influence import
variance_inflation_factor
from sklearn.linear_model import LogisticRegression
from sklearn.feature_selection import RFE
from sklearn import metrics
from sklearn.metrics import confusion_matrix
from sklearn.metrics import precision_score
from sklearn.metrics import precision_recall_curve

# Ignoring Warnings
import warnings
warnings.filterwarnings('ignore')
```

Importing and Cleaning Data

```
df_leads = pd.read_csv('Leads.csv')
df_leads.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]
```

```
df_leads.shape
```

```
(9240, 37)
```

```
df_leads.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 9240 entries, 0 to 9239
Data columns (total 37 columns):
 #   Column                                         Non-Null Count
Dtype
---  ------                                         --------------
-----
 0   Prospect ID                                    9240 non-null
object
 1   Lead Number                                    9240 non-null
int64
 2   Lead Origin                                    9240 non-null
object
 3   Lead Source                                    9204 non-null
object
 4   Do Not Email                                   9240 non-null
object
```

```
 5   Do Not Call                                    9240 non-null
object
 6   Converted                                      9240 non-null
int64
 7   TotalVisits                                    9103 non-null
float64
 8   Total Time Spent on Website                    9240 non-null
int64
 9   Page Views Per Visit                           9103 non-null
float64
 10  Last Activity                                  9137 non-null
object
 11  Country                                        6779 non-null
object
 12  Specialization                                 7802 non-null
object
 13  How did you hear about X Education             7033 non-null
object
 14  What is your current occupation                6550 non-null
object
 15  What matters most to you in choosing a course  6531 non-null
object
 16  Search                                         9240 non-null
object
 17  Magazine                                       9240 non-null
object
 18  Newspaper Article                              9240 non-null
object
 19  X Education Forums                             9240 non-null
object
 20  Newspaper                                      9240 non-null
object
 21  Digital Advertisement                          9240 non-null
object
 22  Through Recommendations                        9240 non-null
object
 23  Receive More Updates About Our Courses         9240 non-null
object
 24  Tags                                           5887 non-null
object
 25  Lead Quality                                   4473 non-null
object
 26  Update me on Supply Chain Content              9240 non-null
object
 27  Get updates on DM Content                      9240 non-null
object
 28  Lead Profile                                   6531 non-null
object
 29  City                                           7820 non-null
object
```

```
 30  Asymmetrique Activity Index                    5022 non-null
object
 31  Asymmetrique Profile Index                     5022 non-null
object
 32  Asymmetrique Activity Score                    5022 non-null
float64
 33  Asymmetrique Profile Score                     5022 non-null
float64
 34  I agree to pay the amount through cheque       9240 non-null
object
 35  A free copy of Mastering The Interview         9240 non-null
object
 36  Last Notable Activity                          9240 non-null
object
dtypes: float64(4), int64(3), object(30)
memory usage: 2.6+ MB
```

```
df_leads.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
```

Data clealing

```
# Identifying features with missing terms
round(100*(df_leads.isnull().sum()/len(df_leads.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                                   15.56
How did you hear about X Education               23.89
What is your current occupation                  29.11
What matters most to you in choosing a course    29.32
Search                                            0.00
Magazine                                          0.00
Newspaper Article                                 0.00
X Education Forums                                0.00
Newspaper                                         0.00
Digital Advertisement                             0.00
Through Recommendations                           0.00
Receive More Updates About Our Courses            0.00
Tags                                             36.29
Lead Quality                                     51.59
Update me on Supply Chain Content                 0.00
Get updates on DM Content                         0.00
Lead Profile                                     29.32
City                                             15.37
Asymmetrique Activity Index                      45.65
Asymmetrique Profile Index                       45.65
Asymmetrique Activity Score                      45.65
Asymmetrique Profile Score                       45.65
I agree to pay the amount through cheque          0.00
A free copy of Mastering The Interview            0.00
```

```
Last Notable Activity                             0.00
dtype: float64
```

```
df_leads.select_dtypes(include = 'object').info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 9240 entries, 0 to 9239
Data columns (total 30 columns):
 #   Column                                         Non-Null Count
Dtype
---  ------                                         --------------
-----
 0   Prospect ID                                    9240 non-null
object
 1   Lead Origin                                    9240 non-null
object
 2   Lead Source                                    9204 non-null
object
 3   Do Not Email                                   9240 non-null
object
 4   Do Not Call                                    9240 non-null
object
 5   Last Activity                                  9137 non-null
object
 6   Country                                        6779 non-null
object
 7   Specialization                                 7802 non-null
object
 8   How did you hear about X Education             7033 non-null
object
 9   What is your current occupation                6550 non-null
object
 10  What matters most to you in choosing a course  6531 non-null
object
 11  Search                                         9240 non-null
object
 12  Magazine                                       9240 non-null
object
 13  Newspaper Article                              9240 non-null
object
 14  X Education Forums                             9240 non-null
object
 15  Newspaper                                      9240 non-null
object
 16  Digital Advertisement                          9240 non-null
object
 17  Through Recommendations                        9240 non-null
object
 18  Receive More Updates About Our Courses         9240 non-null
object
 19  Tags                                           5887 non-null
```

```
object
 20  Lead Quality                                   4473 non-null
object
 21  Update me on Supply Chain Content              9240 non-null
object
 22  Get updates on DM Content                      9240 non-null
object
 23  Lead Profile                                   6531 non-null
object
 24  City                                           7820 non-null
object
 25  Asymmetrique Activity Index                    5022 non-null
object
 26  Asymmetrique Profile Index                     5022 non-null
object
 27  I agree to pay the amount through cheque       9240 non-null
object
 28  A free copy of Mastering The Interview         9240 non-null
object
 29  Last Notable Activity                          9240 non-null
object
dtypes: object(30)
memory usage: 2.1+ MB
```

```
# Replacing Select with nul values
df_leads = df_leads.replace('Select', np.nan)
```

```
df_leads.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 9240 entries, 0 to 9239
Data columns (total 37 columns):
 #   Column                                         Non-Null Count
Dtype
---  ------                                         --------------
-----
 0   Prospect ID                                    9240 non-null
object
 1   Lead Number                                    9240 non-null
int64
 2   Lead Origin                                    9240 non-null
object
 3   Lead Source                                    9204 non-null
object
 4   Do Not Email                                   9240 non-null
object
 5   Do Not Call                                    9240 non-null
object
 6   Converted                                      9240 non-null
int64
 7   TotalVisits                                    9103 non-null
```

```
float64
 8   Total Time Spent on Website                    9240 non-null
int64
 9   Page Views Per Visit                           9103 non-null
float64
 10  Last Activity                                  9137 non-null
object
 11  Country                                        6779 non-null
object
 12  Specialization                                 5860 non-null
object
 13  How did you hear about X Education             1990 non-null
object
 14  What is your current occupation                6550 non-null
object
 15  What matters most to you in choosing a course  6531 non-null
object
 16  Search                                         9240 non-null
object
 17  Magazine                                       9240 non-null
object
 18  Newspaper Article                              9240 non-null
object
 19  X Education Forums                             9240 non-null
object
 20  Newspaper                                      9240 non-null
object
 21  Digital Advertisement                          9240 non-null
object
 22  Through Recommendations                        9240 non-null
object
 23  Receive More Updates About Our Courses         9240 non-null
object
 24  Tags                                           5887 non-null
object
 25  Lead Quality                                   4473 non-null
object
 26  Update me on Supply Chain Content              9240 non-null
object
 27  Get updates on DM Content                      9240 non-null
object
 28  Lead Profile                                   2385 non-null
object
 29  City                                           5571 non-null
object
 30  Asymmetrique Activity Index                    5022 non-null
object
 31  Asymmetrique Profile Index                     5022 non-null
object
 32  Asymmetrique Activity Score                    5022 non-null
```

```
float64
 33  Asymmetrique Profile Score                     5022 non-null
float64
 34  I agree to pay the amount through cheque       9240 non-null
object
 35  A free copy of Mastering The Interview         9240 non-null
object
 36  Last Notable Activity                          9240 non-null
object
dtypes: float64(4), int64(3), object(30)
memory usage: 2.6+ MB
```

```
# From the data converting categorical features into binary values
for features in ['Do Not Email','Do Not
Call','Search','Magazine','Newspaper Article','X Education
Forums','Newspaper',
                 'Digital Advertisement','Through
Recommendations','Receive More Updates About Our Courses','Update me
on Supply Chain Content',
                 'Get updates on DM Content','I agree to pay the
amount through cheque','A free copy of Mastering The Interview']:
    df_leads[features] = df_leads[features].apply(lambda x : 1 if x ==
'Yes' else 0)


df_leads.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             0            0          0          0.0

1  Organic Search             0            0          0          5.0

2  Direct Traffic             0            0          1          2.0

3  Direct Traffic             0            0          0          1.0
```

```
4          Google                  0            0          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                         0             NaN     NaN   
1                         0             NaN     NaN   
2                         0  Potential Lead  Mumbai   
3                         0             NaN  Mumbai   
4                         0             NaN  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                                         0   
1                                         0   
2                                         0   
3                                         0   
4                                         0   

   A free copy of Mastering The Interview  Last Notable Activity
0                                       0               Modified
1                                       0           Email Opened
2                                       1           Email Opened
3                                       0               Modified
4                                       0               Modified

[5 rows x 37 columns]
```

```python
# Checking labels of other categorical features
for column in df_leads:
    print(df_leads[column].astype('category').value_counts())
print('---------------------------------------------------------------
-------------------------')
```

```
000104b9-23e4-4ddc-8caa-8629fe8ad7f4    1
a7a319ea-b6ae-4c6b-afc5-183b933d10b5    1
aa27a0af-eeab-4007-a770-fa8a93fa53c8    1
aa30ebb2-8476-41ce-9258-37cc025110d3    1
aa405742-17ac-4c65-b19e-ab91c241cc53    1
                                       ..
539eb309-df36-4a89-ac58-6d3651393910    1
539ffa32-1be7-4fe1-b04c-faf1bab763cf    1
53aabd84-5dcc-4299-bbe3-62f3764b07b1    1
53ac14bd-2bb2-4315-a21c-94562d1b6b2d    1
fffb0e5e-9f92-4017-9f42-781a69da4154    1
Name: Prospect ID, Length: 9240, dtype: int64
----------------------------------------------------------------------
------------------
579533    1
629593    1
630390    1
630403    1
630405    1
         ..
602534    1
602540    1
602557    1
602561    1
660737    1
Name: Lead Number, Length: 9240, dtype: int64
----------------------------------------------------------------------
------------------
Landing Page Submission    4886
API                        3580
Lead Add Form               718
Lead Import                  55
Quick Add Form                1
Name: Lead Origin, dtype: int64
----------------------------------------------------------------------
------------------
Google               2868
Direct Traffic       2543
Olark Chat           1755
Organic Search       1154
Reference             534
Welingak Website      142
Referral Sites        125
Facebook               55
bing                    6
google                  5
```

```
Click2call              4
Press_Release           2
Social Media            2
Live Chat               2
WeLearn                 1
Pay per Click Ads       1
NC_EDM                  1
blog                    1
testone                 1
welearnblog_Home        1
youtubechannel          1
Name: Lead Source, dtype: int64
----------------------------------------------------------------------
------------------
0    8506
1     734
Name: Do Not Email, dtype: int64
----------------------------------------------------------------------
------------------
0    9238
1       2
Name: Do Not Call, dtype: int64
----------------------------------------------------------------------
------------------
0    5679
1    3561
Name: Converted, dtype: int64
----------------------------------------------------------------------
------------------
0.0      2189
2.0      1680
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
```

```
23.0        6
21.0        6
24.0        5
25.0        5
27.0        5
22.0        3
26.0        2
28.0        2
29.0        2
54.0        1
141.0       1
115.0       1
74.0        1
55.0        1
30.0        1
43.0        1
42.0        1
41.0        1
32.0        1
251.0       1
Name: TotalVisits, dtype: int64
----------------------------------------------------------------------
------------------
0       2193
60        19
75        18
74        18
127       18
        ... 
1091       1
1088       1
1085       1
1084       1
2272       1
Name: Total Time Spent on Website, Length: 1731, dtype: int64
----------------------------------------------------------------------
------------------
0.0     2189
2.0     1795
3.0     1196
4.0      896
1.0      651
        ... 
3.57       1
3.8        1
3.82       1
3.83       1
55.0       1
Name: Page Views Per Visit, Length: 114, dtype: int64
----------------------------------------------------------------------
```

```
------------------
Email Opened                    3437
SMS Sent                        2745
Olark Chat Conversation          973
Page Visited on Website          640
Converted to Lead                428
Email Bounced                    326
Email Link Clicked               267
Form Submitted on Website        116
Unreachable                       93
Unsubscribed                      61
Had a Phone Conversation          30
Approached upfront                 9
View in browser link Clicked       6
Email Received                     2
Email Marked Spam                  2
Resubscribed to emails             1
Visited Booth in Tradeshow         1
Name: Last Activity, dtype: int64
----------------------------------------------------------------------
------------------
India                   6492
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
Kuwait                     4
Nigeria                    4
South Africa               4
Germany                    4
Canada                     4
Sweden                     3
Uganda                     2
Philippines                2
Asia/Pacific Region        2
Italy                      2
Ghana                      2
China                      2
Belgium                    2
Bangladesh                 2
Netherlands                2
Malaysia                   1
```

```
Liberia                    1
Russia                     1
Kenya                      1
Indonesia                  1
Sri Lanka                  1
Switzerland                1
Tanzania                   1
Denmark                    1
Vietnam                    1
Name: Country, dtype: int64
----------------------------------------------------------------------
------------------
Finance Management                   976
Human Resource Management            848
Marketing Management                 838
Operations Management                503
Business Administration              403
IT Projects Management               366
Supply Chain Management              349
Banking, Investment And Insurance    338
Media and Advertising                203
Travel and Tourism                   203
International Business               178
Healthcare Management                159
Hospitality Management               114
E-COMMERCE                           112
Retail Management                    100
Rural and Agribusiness                73
E-Business                            57
Services Excellence                   40
Name: Specialization, dtype: int64
----------------------------------------------------------------------
------------------
Online Search            808
Word Of Mouth            348
Student of SomeSchool    310
Other                    186
Multiple Sources         152
Advertisements            70
Social Media              67
Email                     26
SMS                       23
Name: How did you hear about X Education, dtype: int64
----------------------------------------------------------------------
------------------
Unemployed              5600
Working Professional     706
Student                  210
Other                     16
Housewife                 10
```

```
Businessman                8
Name: What is your current occupation, dtype: int64
----------------------------------------------------------------------
------------------
Better Career Prospects      6528
Flexibility & Convenience       2
Other                           1
Name: What matters most to you in choosing a course, dtype: int64
----------------------------------------------------------------------
------------------
0    9226
1      14
Name: Search, dtype: int64
----------------------------------------------------------------------
------------------
0    9240
Name: Magazine, dtype: int64
----------------------------------------------------------------------
------------------
0    9238
1       2
Name: Newspaper Article, dtype: int64
----------------------------------------------------------------------
------------------
0    9239
1       1
Name: X Education Forums, dtype: int64
----------------------------------------------------------------------
------------------
0    9239
1       1
Name: Newspaper, dtype: int64
----------------------------------------------------------------------
------------------
0    9236
1       4
Name: Digital Advertisement, dtype: int64
----------------------------------------------------------------------
------------------
0    9233
1       7
Name: Through Recommendations, dtype: int64
----------------------------------------------------------------------
------------------
0    9240
Name: Receive More Updates About Our Courses, dtype: int64
----------------------------------------------------------------------
------------------
Will revert after reading the email                  2072
Ringing                                              1203
```

```
Interested in other courses                            513
Already a student                                      465
Closed by Horizzon                                     358
switched off                                           240
Busy                                                   186
Lost to EINS                                           175
Not doing further education                            145
Interested  in full time MBA                           117
Graduation in progress                                 111
invalid number                                          83
Diploma holder (Not Eligible)                           63
wrong number given                                      47
opp hangup                                              33
number not provided                                     27
in touch with EINS                                      12
Lost to Others                                           7
Still Thinking                                           6
Want to take admission but has financial problems        6
Interested in Next batch                                 5
In confusion whether part time or DLP                    5
Lateral student                                          3
Shall take in the next coming month                      2
University not recognized                                2
Recognition issue (DEC approval)                         1
Name: Tags, dtype: int64
----------------------------------------------------------------------
------------------
Might be             1560
Not Sure             1092
High in Relevance     637
Worst                 601
Low in Relevance      583
Name: Lead Quality, dtype: int64
----------------------------------------------------------------------
------------------
0    9240
Name: Update me on Supply Chain Content, dtype: int64
----------------------------------------------------------------------
------------------
0    9240
Name: Get updates on DM Content, dtype: int64
----------------------------------------------------------------------
------------------
Potential Lead                 1613
Other Leads                     487
Student of SomeSchool           241
Lateral Student                  24
Dual Specialization Student      20
Name: Lead Profile, dtype: int64
----------------------------------------------------------------------
```

```
------------------
Mumbai                         3222
Thane & Outskirts               752
Other Cities                    686
Other Cities of Maharashtra     457
Other Metro Cities              380
Tier II Cities                   74
Name: City, dtype: int64
----------------------------------------------------------------------
------------------
02.Medium    3839
01.High       821
03.Low        362
Name: Asymmetrique Activity Index, dtype: int64
----------------------------------------------------------------------
------------------
02.Medium    2788
01.High      2203
03.Low         31
Name: Asymmetrique Profile Index, dtype: int64
----------------------------------------------------------------------
------------------
14.0    1771
15.0    1293
13.0     775
16.0     467
17.0     349
12.0     196
11.0      95
10.0      57
9.0        9
18.0       5
8.0        4
7.0        1
Name: Asymmetrique Activity Score, dtype: int64
----------------------------------------------------------------------
------------------
15.0    1759
18.0    1071
16.0     599
17.0     579
20.0     308
19.0     245
14.0     226
13.0     204
12.0      22
11.0       9
Name: Asymmetrique Profile Score, dtype: int64
----------------------------------------------------------------------
------------------
```

```
0    9240
Name: I agree to pay the amount through cheque, dtype: int64
----------------------------------------------------------------------
------------------
0    6352
1    2888
Name: A free copy of Mastering The Interview, dtype: int64
----------------------------------------------------------------------
------------------
Modified                        3407
Email Opened                    2827
SMS Sent                        2172
Page Visited on Website          318
Olark Chat Conversation          183
Email Link Clicked               173
Email Bounced                     60
Unsubscribed                      47
Unreachable                       32
Had a Phone Conversation          14
Email Marked Spam                  2
Approached upfront                 1
Email Received                     1
Form Submitted on Website          1
Resubscribed to emails             1
View in browser link Clicked       1
Name: Last Notable Activity, dtype: int64
----------------------------------------------------------------------
------------------
```

Handling data and the missing values

```
# Converting all the values to lower case as google is mentioned in
bot upper and lower case
df_leads = df_leads.applymap(lambda s:s.lower() if type(s) == str else
s)

# Checking if there are columns with one unique value since it won't
affect our analysis
df_leads.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
```

```
# Dropping unique valued columns
df_leads= df_leads.drop(['Magazine','Receive More Updates About Our
Courses','Update me on Supply Chain Content',
                         'Get updates on DM Content','I agree to pay
the amount through cheque'], axis =1)
```

```
# Checking the percentage of missing values
round(100*(df_leads.isnull().sum()/len(df_leads.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
df_2 = df_leads.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)
df_2.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             0            0          0          0.0   

1  organic search             0            0          0          5.0   

2  direct traffic             0            0          1          2.0   

3  direct traffic             0            0          0          1.0   

4          google             0            0          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      0
0
1                       better career prospects      0
0
2                       better career prospects      0
0
3                       better career prospects      0
0
4                       better career prospects      0
0

   X Education Forums  Newspaper  Digital Advertisement  \
0                   0          0                      0
1                   0          0                      0
2                   0          0                      0
3                   0          0                      0
4                   0          0                      0

   Through Recommendations  A free copy of Mastering The Interview  \
0                        0                                       0
1                        0                                       0
2                        0                                       1
3                        0                                       0
4                        0                                       0

   Last Notable Activity
0               modified
1           email opened
```

```
2           email opened
3               modified
4               modified

[5 rows x 22 columns]
```

```
# Rechecking the percentage of missing values
round(100*(df_2.isnull().sum()/len(df_2.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
```

```
# Replacing the remaing null values with not provided as removing
these values will result in huge data loss

df_2['Specialization'] = df_2['Specialization'].fillna('not provided')

df_2['What matters most to you in choosing a course'] = df_2['What
matters most to you in choosing a course'].fillna('not provided')
df_2['Country'] = df_2['Country'].fillna('not provided')
df_2['What is your current occupation'] = df_2['What is your current
occupation'].fillna('not provided')
df_2.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 9240 entries, 0 to 9239
Data columns (total 22 columns):
 #   Column                                         Non-Null Count
Dtype
---  ------                                         --------------
```

```
-----
 0   Prospect ID                                    9240 non-null
object
 1   Lead Origin                                    9240 non-null
object
 2   Lead Source                                    9204 non-null
object
 3   Do Not Email                                   9240 non-null
int64
 4   Do Not Call                                    9240 non-null
int64
 5   Converted                                      9240 non-null
int64
 6   TotalVisits                                    9103 non-null
float64
 7   Total Time Spent on Website                    9240 non-null
int64
 8   Page Views Per Visit                           9103 non-null
float64
 9   Last Activity                                  9137 non-null
object
 10  Country                                        9240 non-null
object
 11  Specialization                                 9240 non-null
object
 12  What is your current occupation                9240 non-null
object
 13  What matters most to you in choosing a course  9240 non-null
object
 14  Search                                         9240 non-null
int64
 15  Newspaper Article                              9240 non-null
int64
 16  X Education Forums                             9240 non-null
int64
 17  Newspaper                                      9240 non-null
int64
 18  Digital Advertisement                          9240 non-null
int64
 19  Through Recommendations                        9240 non-null
int64
 20  A free copy of Mastering The Interview         9240 non-null
int64
 21  Last Notable Activity                          9240 non-null
object
dtypes: float64(2), int64(11), object(9)
memory usage: 1.6+ MB
```

```
df_2["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
oman                       6
france                     6
unknown                    5
kuwait                     4
south africa               4
canada                     4
nigeria                    4
germany                    4
sweden                     3
philippines                2
uganda                     2
italy                      2
bangladesh                 2
netherlands                2
asia/pacific region        2
china                      2
belgium                    2
ghana                      2
kenya                      1
sri lanka                  1
tanzania                   1
malaysia                   1
liberia                    1
switzerland                1
denmark                    1
russia                     1
vietnam                    1
indonesia                  1
Name: Country, dtype: int64
```

```python
# Function to replace feature country into only 3 values
def slots(x):
    category = ""
    if x == "india":
        category = "india"
    elif x == "not provided":
        category = "not provided"
    else:
        category = "outside india"
    return category

df_2['Country'] = df_2.apply(lambda x:slots(x['Country']), axis = 1)
df_2['Country'].value_counts()
```

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

```
# Rechecking the percentage of missing values
round(100*(df_2.isnull().sum()/len(df_2.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
```

```
df3 = df_2[df_2.isnull().sum(axis=1) <1]
```

```
# 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
```

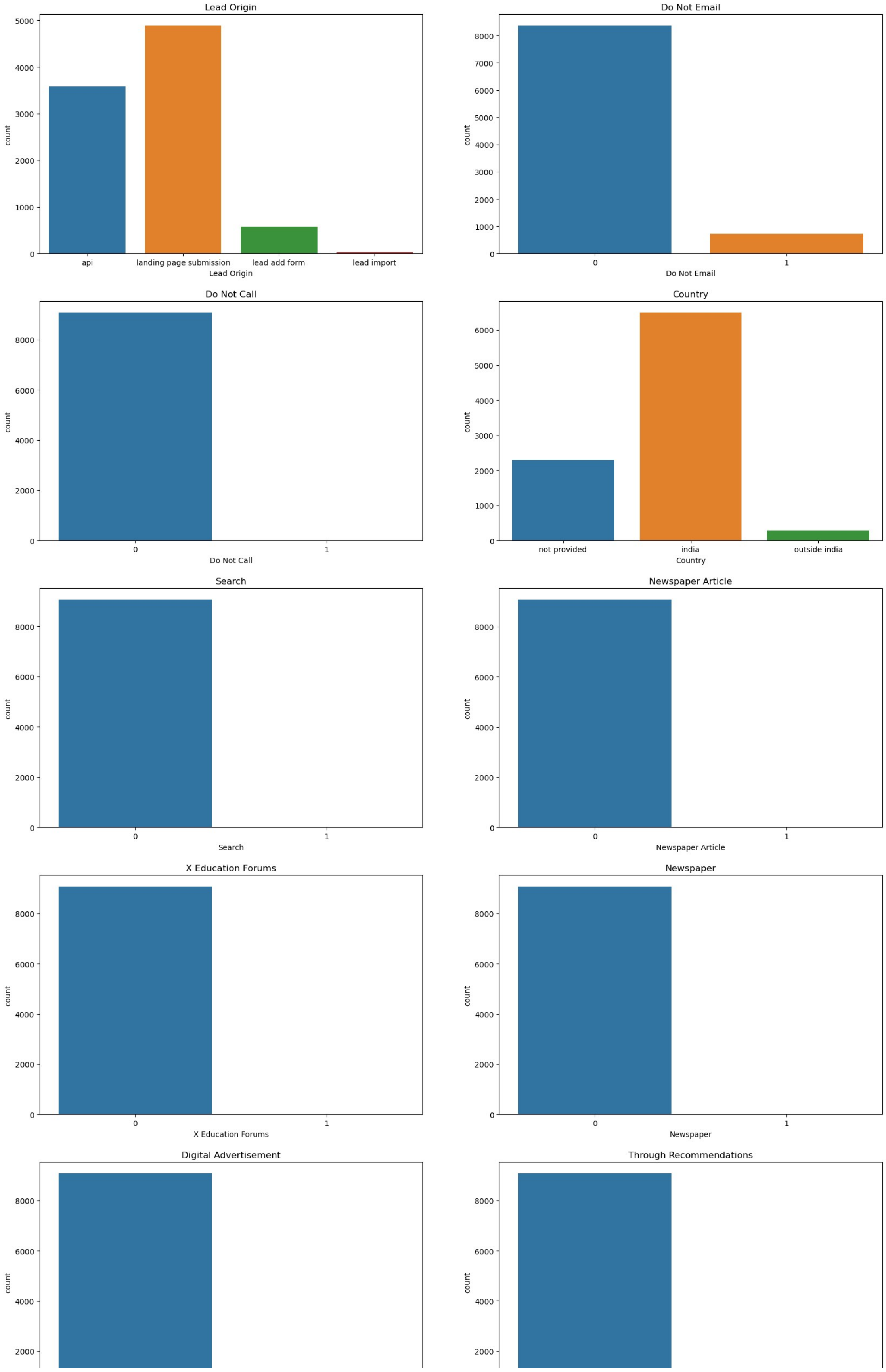
```
# Removing Id values since they are unique for everyone
df_final = df3.drop('Prospect ID',1)
df_final.shape
```

```
(9074, 21)
```

Univariate Analysis

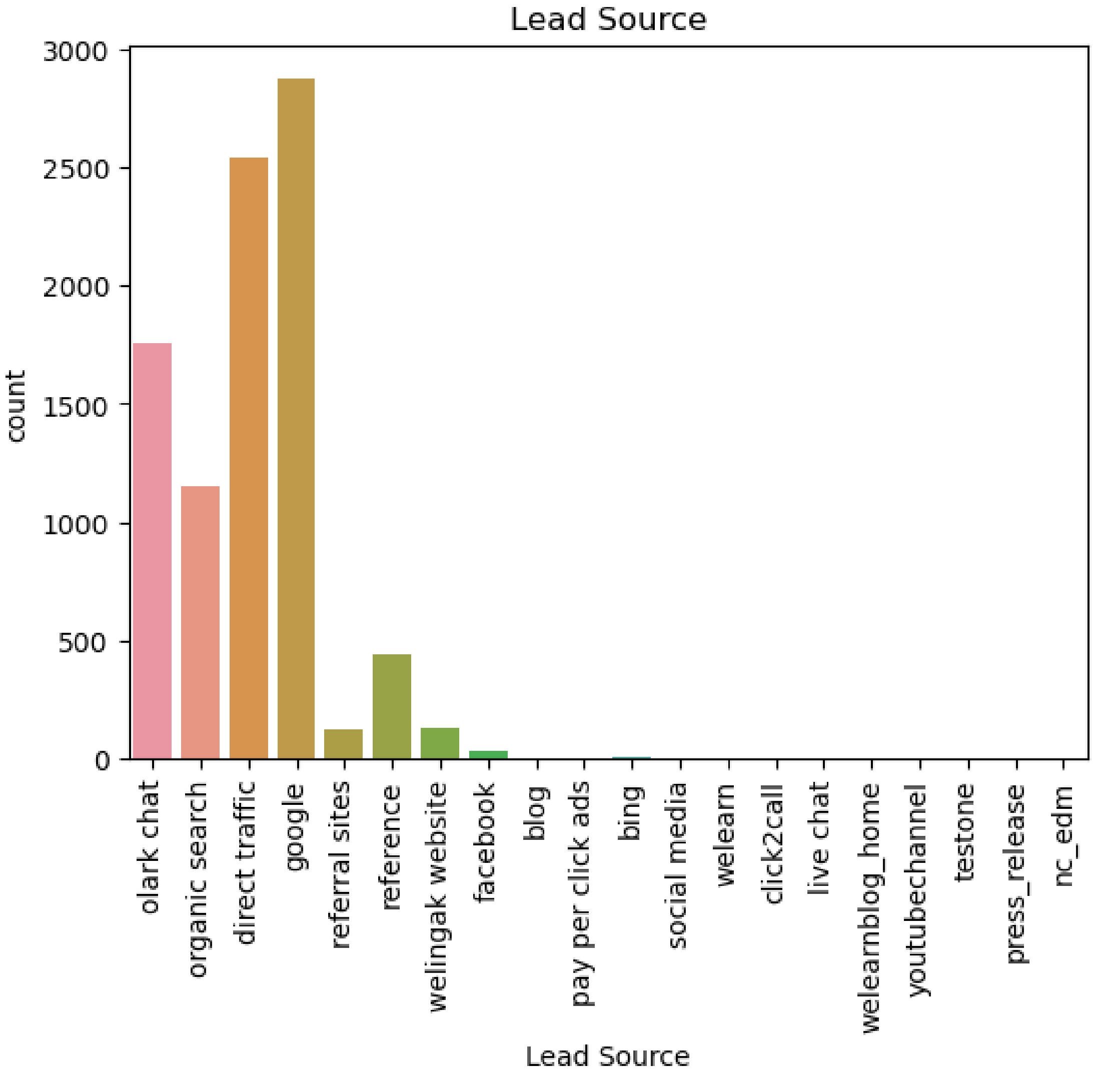
```
plt.figure(figsize = (20,40))

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')
```

```
Text(0.5, 1.0, 'Last Notable Activity')
```



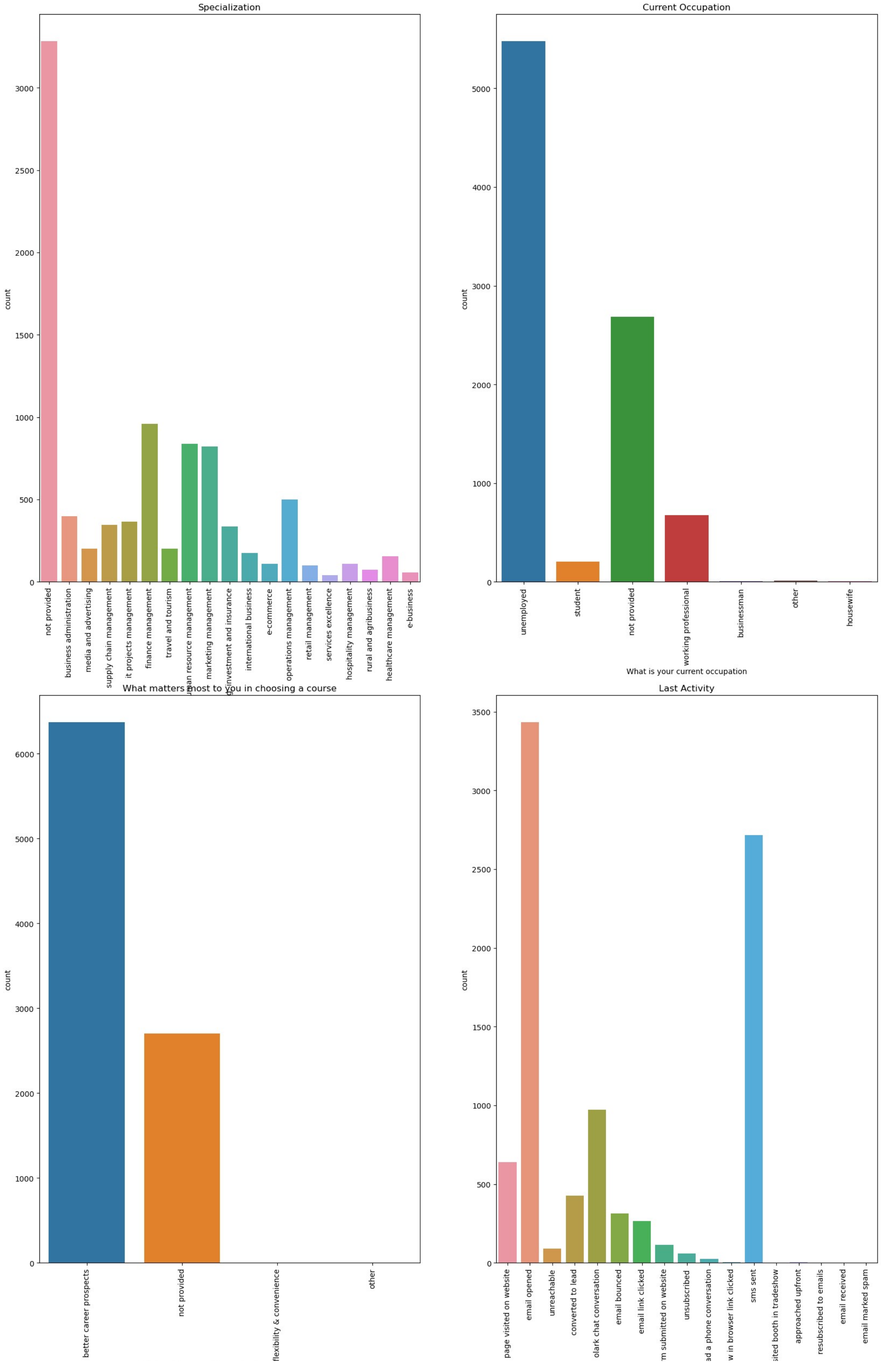
Lead Origin
count
0
1000
2000
3000
4000
5000
api
landing page submission
lead add form
lead import
Lead Origin
Do Not Email
count
0
1000
2000
3000
4000
5000
6000
7000
8000
0
1
Do Not Email
Do Not Call
count
0
2000
4000
6000
8000
0
1
Do Not Call
Country
count
0
1000
2000
3000
4000
5000
6000
not provided
india
outside india
Country
Search
count
0
2000
4000
6000
8000
0
1
Search
Newspaper Article
count
0
2000
4000
6000
8000
0
1
Newspaper Article
X Education Forums
count
0
2000
4000
6000
8000
0
1
X Education Forums
Newspaper
count
0
2000
4000
6000
8000
0
1
Newspaper
Digital Advertisement
count
2000
4000
6000
8000
Through Recommendations
count
2000
4000
6000
8000

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

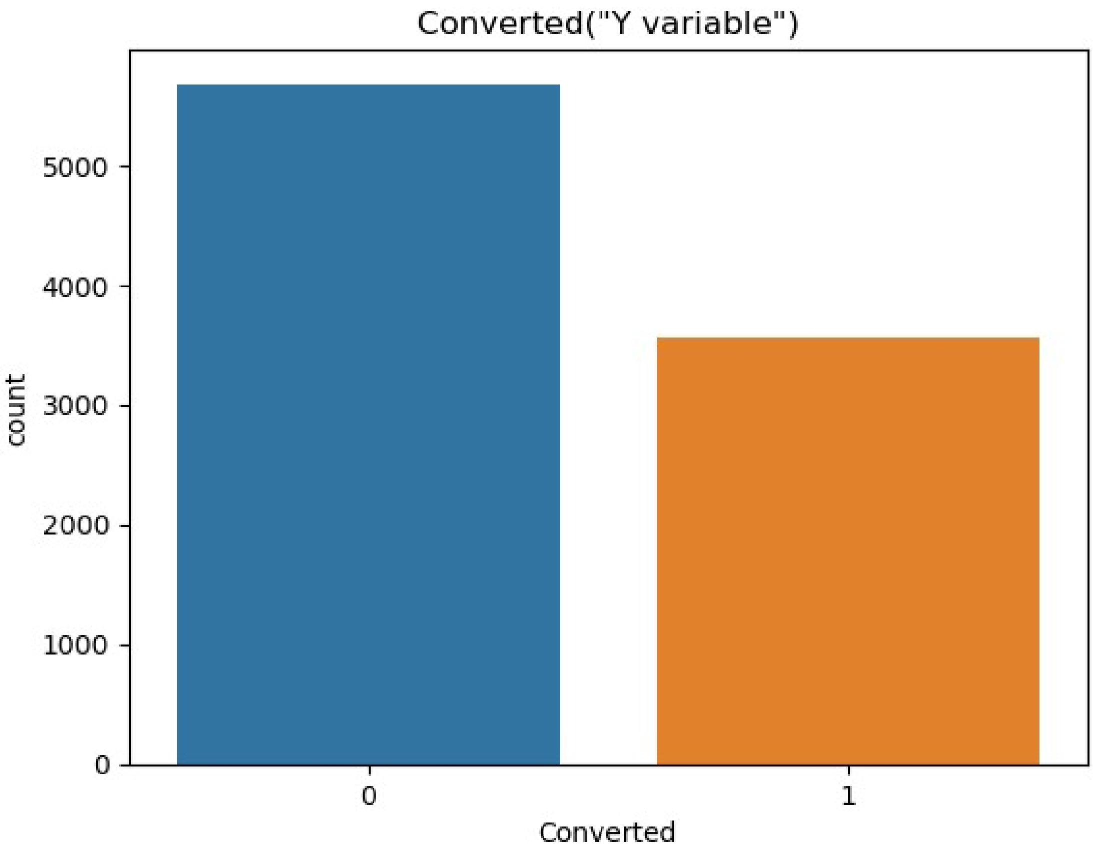


```
plt.figure(figsize = (20,30))
plt.subplot(2,2,1)
sns.countplot(df_final['Specialization']).tick_params(axis='x',
rotation = 90)
plt.title('Specialization')
plt.subplot(2,2,2)
sns.countplot(df_final['What is your current
occupation']).tick_params(axis='x', rotation = 90)
plt.title('Current Occupation')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

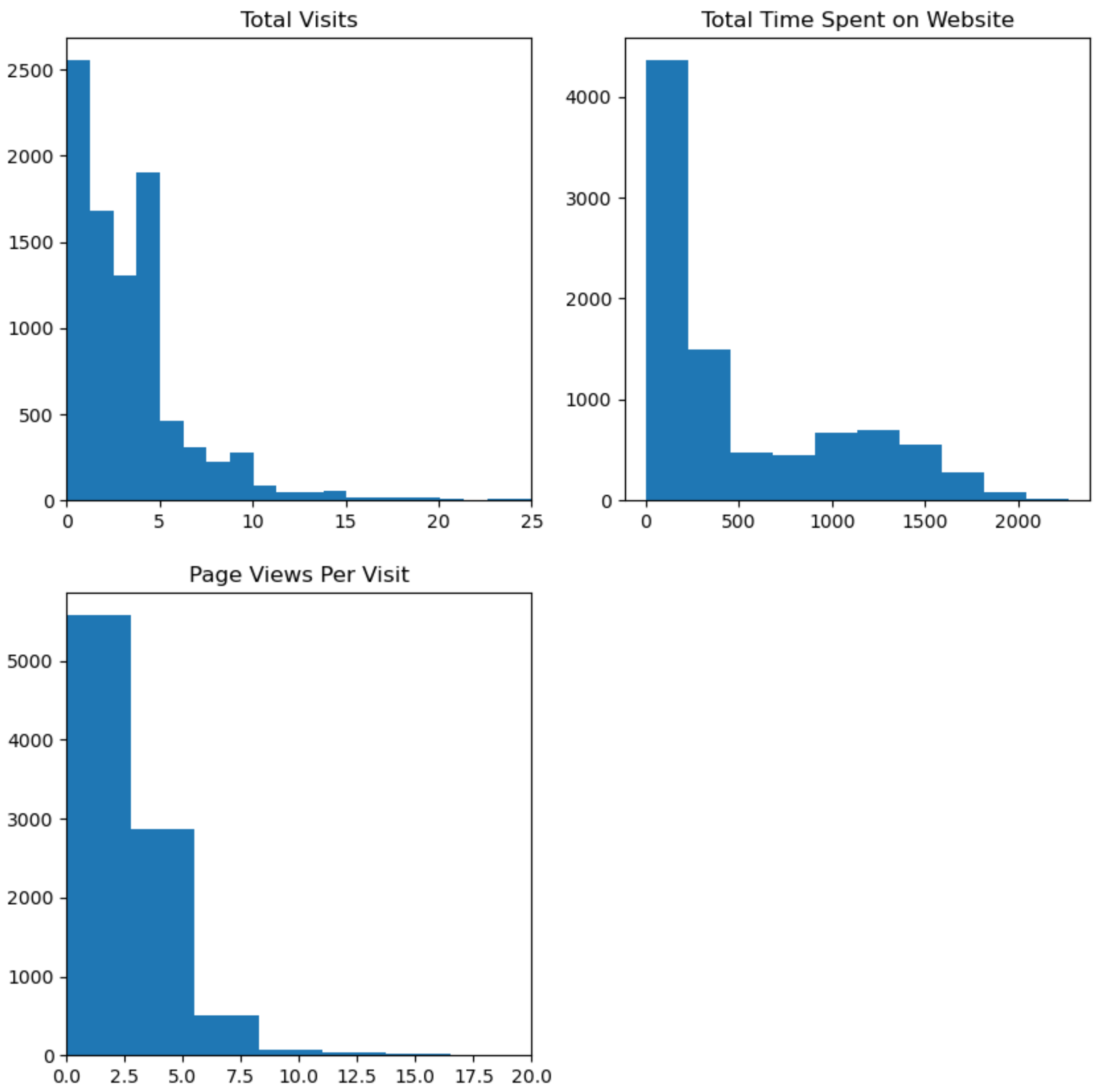
```
plt.title('What matters most to you in choosing a course')
plt.subplot(2,2,4)
sns.countplot(df_final['Last Activity']).tick_params(axis='x',
rotation = 90)
plt.title('Last Activity')
plt.show()
```


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
uman 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
0
1000
2000
3000
4000
5000
6000
better career prospects
not provided
flexibility & convenience
other
Last Activity
count
0
500
1000
1500
2000
2500
3000
3500
page visited on website
email opened
unreachable
converted to lead
olark chat conversation
email bounced
email link clicked
m submitted on website
unsubscribed
ad a phone conversation
w in browser link clicked
sms sent
sited booth in tradeshow
approached upfront
resubscribed to emails
email received
email marked spam

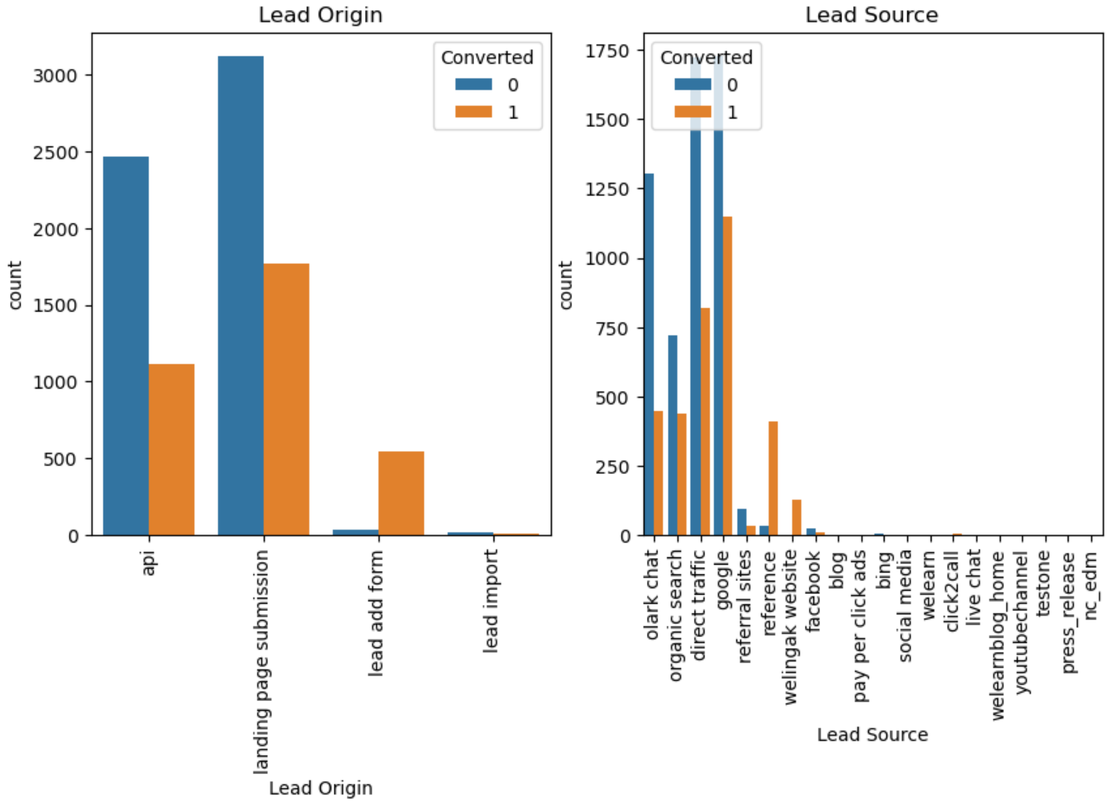
```
sns.countplot(df_leads['Converted'])
plt.title('Converted("Y variable")')
plt.show()
```



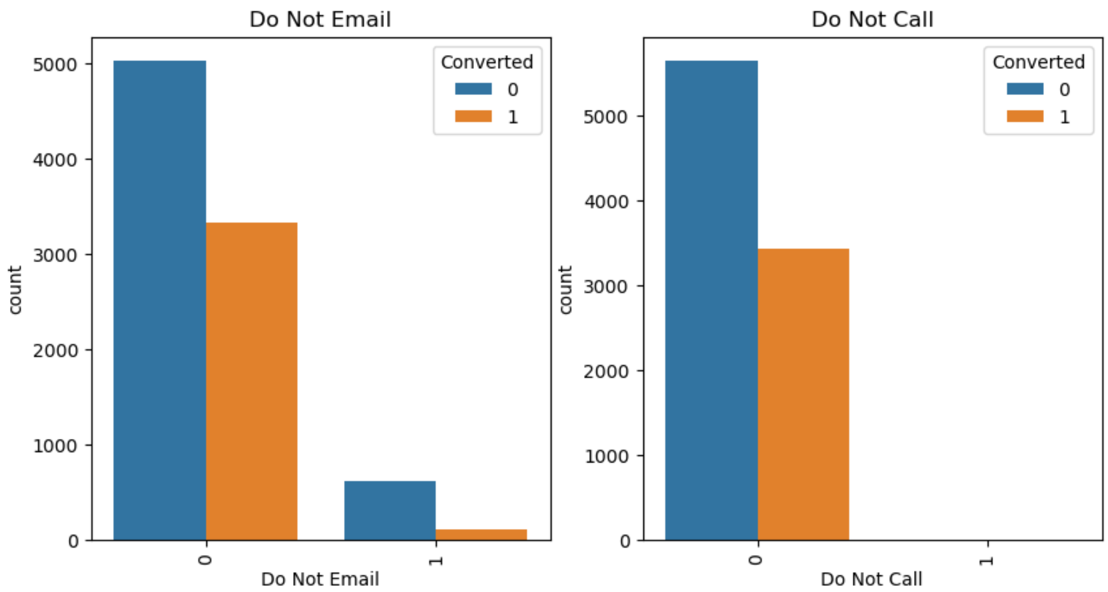
```
# Numerical variables
plt.figure(figsize = (10,10))
plt.subplot(221)
plt.hist(df_final['TotalVisits'], bins = 200)
plt.title('Total Visits')
plt.xlim(0,25)

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()
```



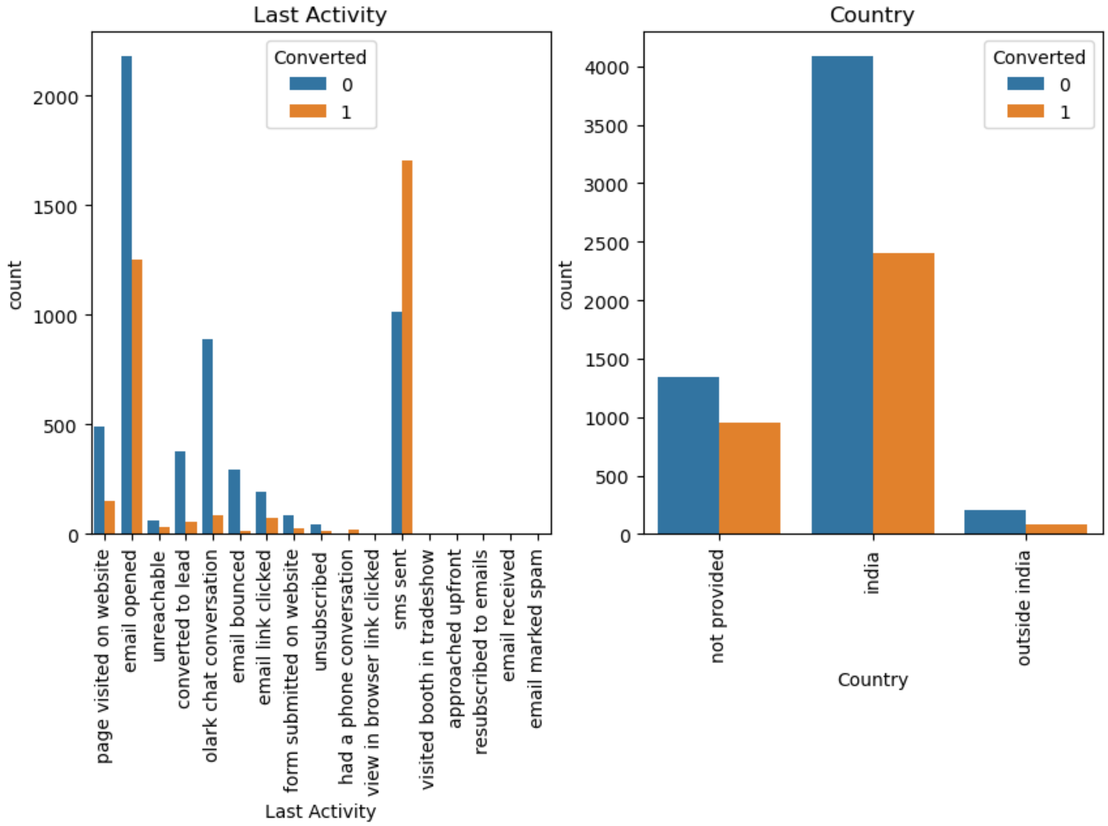
```
plt.figure(figsize = (10,5))

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()
```



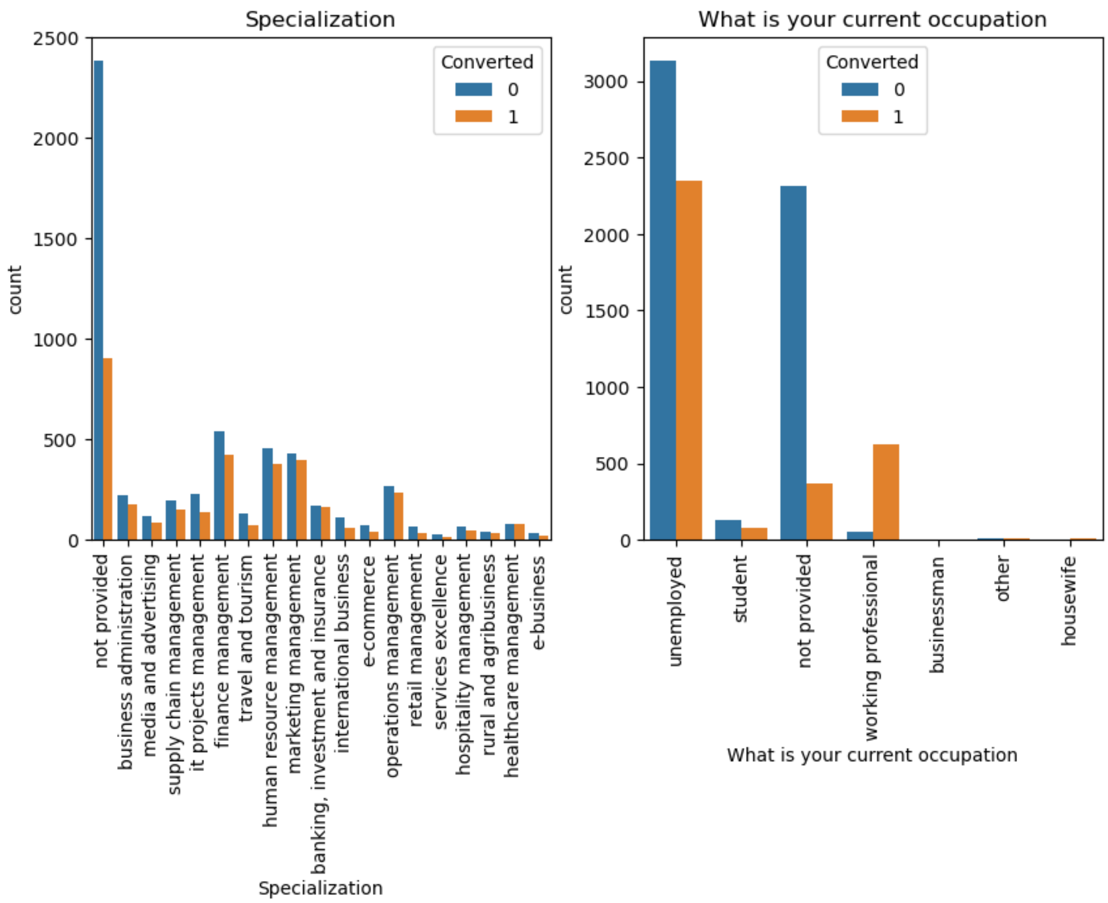
```
plt.figure(figsize = (10,5))

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()
```



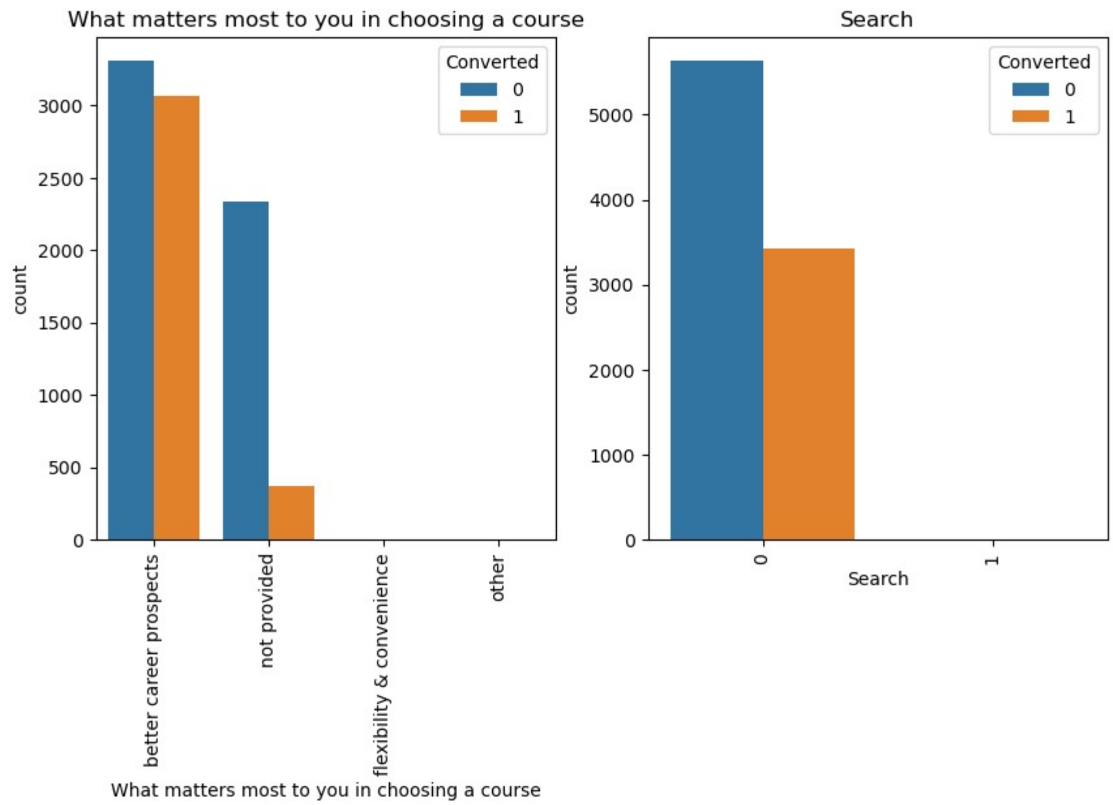
```
plt.figure(figsize = (10,5))

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()
```



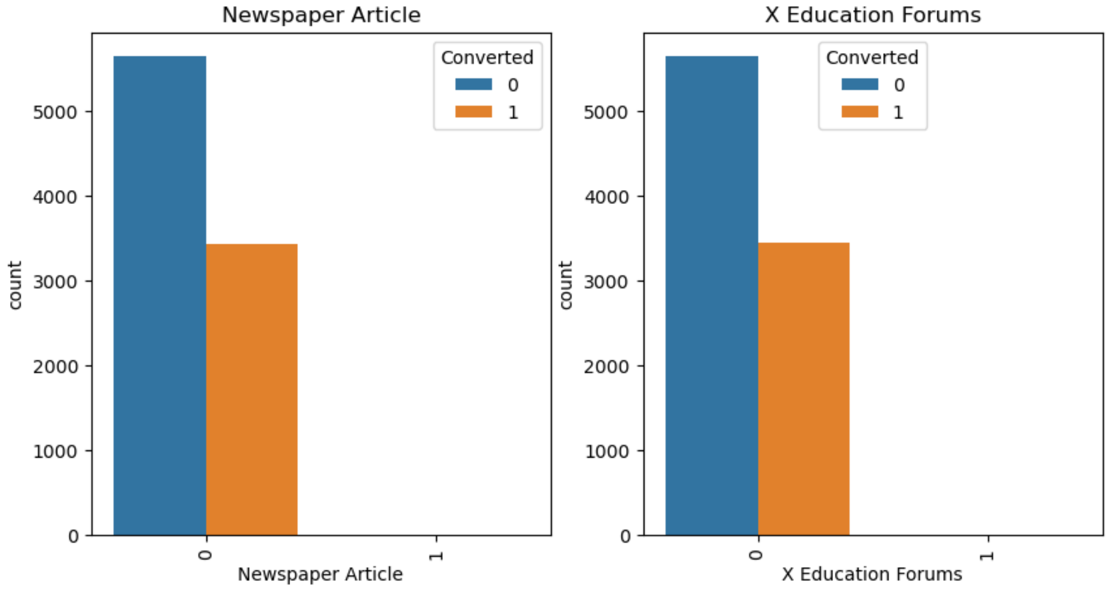
```
plt.figure(figsize = (10,5))

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()
```



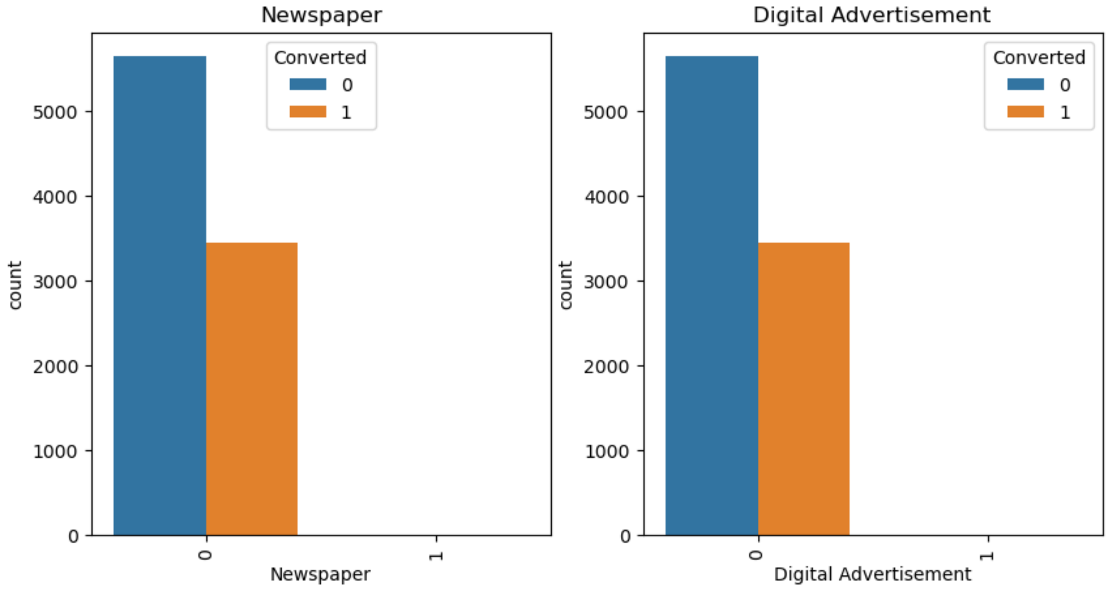
```
plt.figure(figsize = (10,5))

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()
```



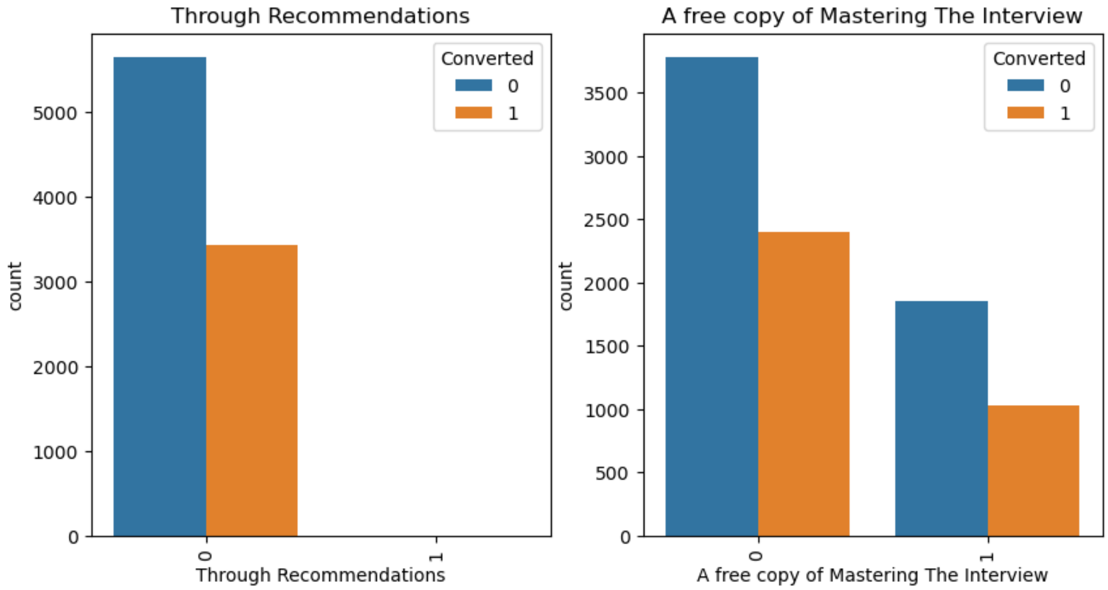
```
plt.figure(figsize = (10,5))

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()
```



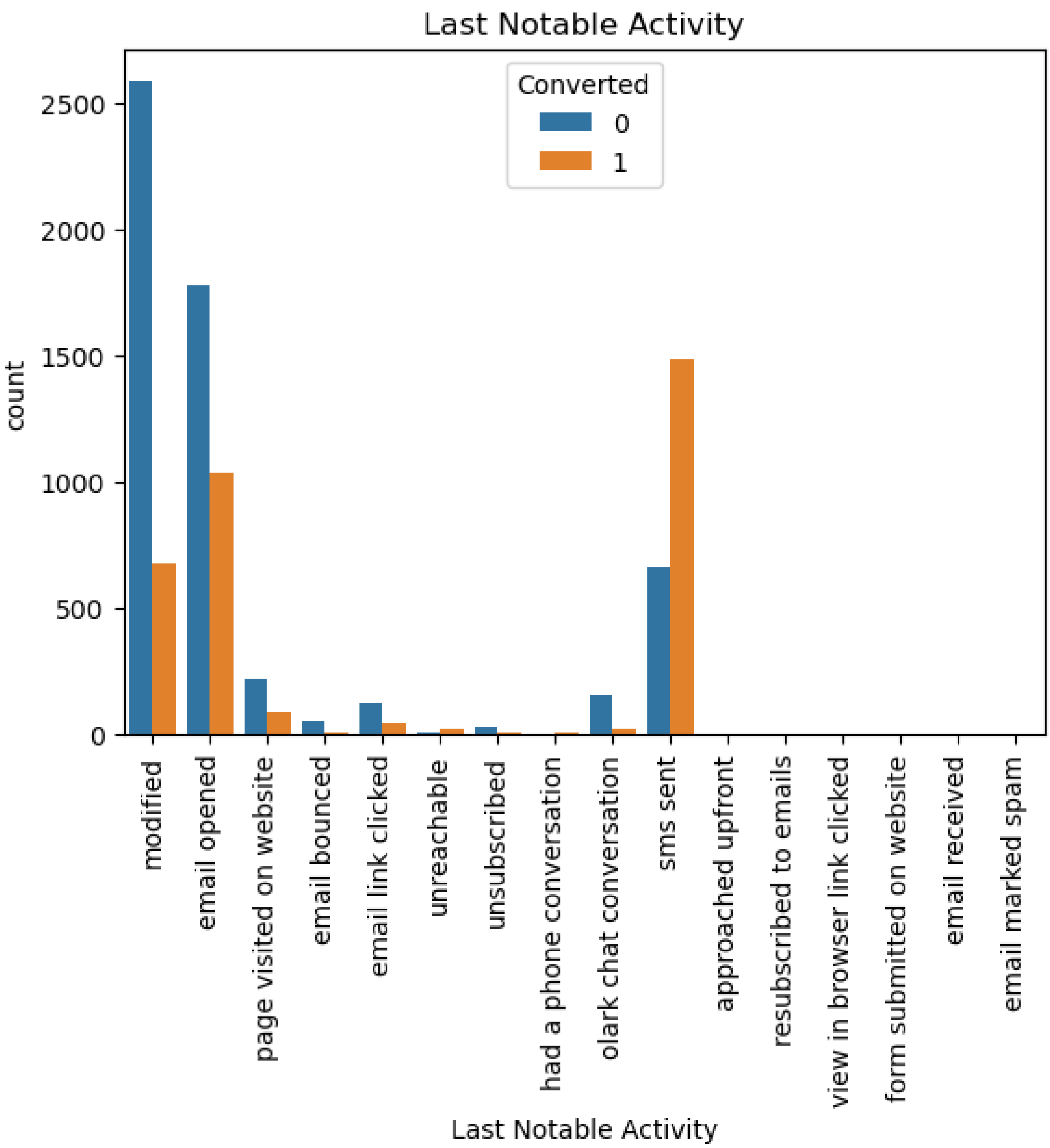
```
plt.figure(figsize = (10,5))

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()
```



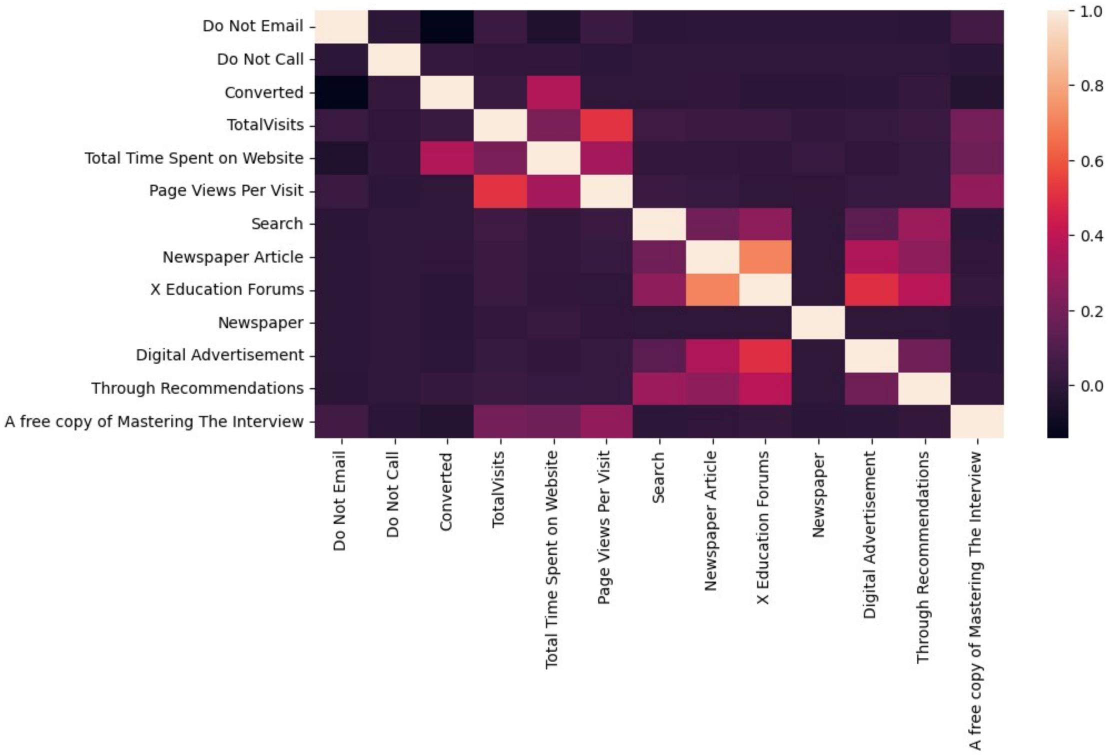
```
plt.figure(figsize = (10,5))

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()
```



```
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 are no major outliers in the data

Dummy variables

```
df_final.info()
```

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 9074 entries, 0 to 9239
Data columns (total 21 columns):
 #   Column                                     Non-Null Count  
Dtype
```

```
---  ------                                          --------------
-----
 0   Lead Origin                                     9074 non-null
object
 1   Lead Source                                     9074 non-null
object
 2   Do Not Email                                    9074 non-null
int64
 3   Do Not Call                                     9074 non-null
int64
 4   Converted                                       9074 non-null
int64
 5   TotalVisits                                     9074 non-null
float64
 6   Total Time Spent on Website                     9074 non-null
int64
 7   Page Views Per Visit                            9074 non-null
float64
 8   Last Activity                                   9074 non-null
object
 9   Country                                         9074 non-null
object
 10  Specialization                                  9074 non-null
object
 11  What is your current occupation                 9074 non-null
object
 12  What matters most to you in choosing a course  9074 non-null
object
 13  Search                                          9074 non-null
int64
 14  Newspaper Article                               9074 non-null
int64
 15  X Education Forums                              9074 non-null
int64
 16  Newspaper                                       9074 non-null
int64
 17  Digital Advertisement                           9074 non-null
int64
 18  Through Recommendations                         9074 non-null
int64
 19  A free copy of Mastering The Interview          9074 non-null
int64
 20  Last Notable Activity                           9074 non-null
object
dtypes: float64(2), int64(11), object(8)
memory usage: 1.5+ MB
```

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

```
Index(['Lead Origin', 'Lead Source', 'Last Activity', 'Country',
       'Specialization', 'What is your current occupation',
       'What matters most to you in choosing a course',
       '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             0
0
1                         api  organic search             0
0
2     landing page submission  direct traffic             0
0
3     landing page submission  direct traffic             0
0
4     landing page submission          google             0
0
...                       ...             ...           ...          .
..
9235  landing page submission  direct traffic             1
0
9236  landing page submission  direct traffic             0
0
9237  landing page submission  direct traffic             1
0
9238  landing page submission          google             0
0
9239  landing page submission  direct traffic             0
0

      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
...         ...          ...                          ...
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  ...
...                    ...                      ...            ...  ..
.
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  ...

     Last Notable Activity_form submitted on website  \
0                                                   0
1                                                   0
2                                                   0
3                                                   0
4                                                   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
...                                              ...
9235                                               0
9236                                               0
9237                                               0
9238                                               0
```

```
9239                                              0

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

```
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
...                                      ...                        ..
.
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  ...
...                         ...  ...
9235                          0  ...
9236                          0  ...
9237                          0  ...
9238                          0  ...
9239                          0  ...

      Last Notable Activity_form submitted on website  \
0                                                   0
1                                                   0
2                                                   0
3                                                   0
4                                                   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   
...                                              ...   
9235                                               0   
9236                                               0   
9237                                               0   
9238                                               0   
9239                                               0   

      Last Notable Activity_modified  \
0                                  1   
1                                  0   
2                                  0   
3                                  1   
4                                  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   
...                                             ...   
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
...                                              ...
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
...                                             ...
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
...                              ...                               ..
.
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
...                                  ...
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
...                                                  ...
9235                                                   0
9236                                                   0
9237                                                   0
9238                                                   0
9239                                                   0

[9074 rows x 79 columns]
```

Train test and 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  ...   

   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 78 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)

# 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  ...   

      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 78 columns]
```

Building the model

```
logreg = LogisticRegression()
# Running RFE with 15 variables as output
rfe = RFE(logreg,n_features_to_select=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, 4),
 ('Lead Origin_landing page submission', False, 25),
 ('Lead Origin_lead add form', True, 1),
 ('Lead Origin_lead import', False, 38),
 ('Specialization_business administration', False, 33),
 ('Specialization_e-business', False, 32),
 ('Specialization_e-commerce', False, 24),
 ('Specialization_finance management', False, 30),
 ('Specialization_healthcare management', False, 27),
 ('Specialization_hospitality management', False, 43),
 ('Specialization_human resource management', False, 31),
 ('Specialization_international business', False, 36),
 ('Specialization_it projects management', False, 29),
 ('Specialization_marketing management', False, 21),
 ('Specialization_media and advertising', False, 40),
 ('Specialization_operations management', False, 26),
 ('Specialization_retail management', False, 60),
 ('Specialization_rural and agribusiness', False, 23),
 ('Specialization_services excellence', False, 22),
 ('Specialization_supply chain management', False, 28),
 ('Specialization_travel and tourism', False, 35),
 ('Lead Source_blog', False, 41),
 ('Lead Source_click2call', False, 61),
 ('Lead Source_direct traffic', False, 16),
 ('Lead Source_facebook', False, 39),
 ('Lead Source_google', False, 18),
 ('Lead Source_live chat', False, 44),
 ('Lead Source_nc_edm', False, 63),
 ('Lead Source_olark chat', True, 1),
 ('Lead Source_organic search', False, 17),
 ('Lead Source_pay per click ads', False, 62),
 ('Lead Source_press_release', False, 34),
 ('Lead Source_reference', False, 2),
 ('Lead Source_referral sites', False, 19),
 ('Lead Source_social media', False, 20),
 ('Lead Source_testone', False, 42),
 ('Lead Source_welearn', False, 45),
 ('Lead Source_welearnblog_home', False, 46),
 ('Lead Source_welingak website', True, 1),
 ('Lead Source_youtubechannel', False, 48),
 ('Last Activity_converted to lead', False, 10),
 ('Last Activity_email bounced', True, 1),
 ('Last Activity_email link clicked', False, 56),
 ('Last Activity_email marked spam', False, 49),
 ('Last Activity_email opened', False, 37),
 ('Last Activity_email received', False, 52),
 ('Last Activity_form submitted on website', False, 51),
 ('Last Activity_had a phone conversation', False, 3),
 ('Last Activity_olark chat conversation', True, 1),
 ('Last Activity_page visited on website', False, 13),
 ('Last Activity_resubscribed to emails', False, 5),
 ('Last Activity_sms sent', True, 1),
 ('Last Activity_unreachable', False, 15),
 ('Last Activity_unsubscribed', False, 11),
 ('Last Activity_view in browser link clicked', False, 59),
 ('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', True, 1),
 ('What is your current occupation_unemployed', True, 1),
 ('What is your current occupation_working professional', True, 1),
 ('Last Notable Activity_email bounced', False, 53),
 ('Last Notable Activity_email link clicked', False, 9),
 ('Last Notable Activity_email marked spam', False, 47),
 ('Last Notable Activity_email opened', False, 12),
 ('Last Notable Activity_email received', False, 57),
 ('Last Notable Activity_form submitted on website', False, 58),
 ('Last Notable Activity_had a phone conversation', True, 1),
 ('Last Notable Activity_modified', False, 6),
 ('Last Notable Activity_olark chat conversation', False, 7),
 ('Last Notable Activity_page visited on website', False, 8),
 ('Last Notable Activity_resubscribed to emails', False, 14),
 ('Last Notable Activity_sms sent', False, 50),
 ('Last Notable Activity_unreachable', True, 1),
 ('Last Notable Activity_unsubscribed', False, 54),
 ('Last Notable Activity_view in browser link clicked', False, 64)]
```

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

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

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:
-2654.3
Date:                Thu, 15 Jun 2023   Deviance:
5308.6
Time:                        19:23:58   Pearson chi2:
6.59e+03
No. Iterations:                    22   Pseudo R-squ. (CS):
0.3926
Covariance Type:            nonrobust

=======================================================================
==================================================
                                                            coef    std
err          z      P>|z|      [0.025      0.975]
-----------------------------------------------------------------------
--------------------------------------------------
const                                                    -3.4921
0.114    -30.632      0.000      -3.716      -3.269
TotalVisits                                               4.4247
1.467      3.016      0.003       1.549       7.300
Total Time Spent on Website                               4.6634
0.166     28.038      0.000       4.337       4.989
Lead Origin_lead add form                                 3.6800
0.225     16.385      0.000       3.240       4.120
Lead Source_olark chat                                    1.6015
0.112     14.344      0.000       1.383       1.820
Lead Source_welingak website                              2.6284
1.036      2.537      0.011       0.597       4.659
Last Activity_email bounced                              -1.8713
0.337     -5.559      0.000      -2.531      -1.212
Last Activity_olark chat conversation                    -1.4071
0.167     -8.405      0.000      -1.735      -1.079
Last Activity_sms sent                                    1.2137
0.074     16.472      0.000       1.069       1.358
What is your current occupation_housewife                25.4295
3.09e+04      0.001      0.999   -6.05e+04    6.06e+04
What is your current occupation_other                     2.2360
0.756      2.959      0.003       0.755       3.717
What is your current occupation_student                   1.3091
0.226      5.798      0.000       0.867       1.752
What is your current occupation_unemployed                1.1793
0.086     13.747      0.000       1.011       1.347
What is your current occupation_working professional      3.7384
0.205     18.224      0.000       3.336       4.141
Last Notable Activity_had a phone conversation           24.0520
2.16e+04      0.001      0.999   -4.23e+04    4.24e+04
Last Notable Activity_unreachable                         1.8612
0.602      3.092      0.002       0.681       3.041
=======================================================================
==================================================
"""
```

```
# 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
11         What is your current occupation_unemployed  2.31
1                        Total Time Spent on Website  2.07
0                                        TotalVisits  1.82
2                            Lead Origin_lead add form  1.59
7                               Last Activity_sms sent  1.55
3                               Lead Source_olark chat  1.51
6                Last Activity_olark chat conversation  1.37
12  What is your current occupation_working profes...  1.32
4                          Lead Source_welingak website  1.31
10            What is your current occupation_student  1.05
5                          Last Activity_email bounced  1.03
9               What is your current occupation_other  1.01
14                 Last Notable Activity_unreachable  1.01
8           What is your current occupation_housewife  1.00
13     Last Notable Activity_had a phone conversation  1.00
```

```
# Revoming Last Notable Activity had a phone conversation as it has
high p value
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:
-2662.3
Date:                 Thu, 15 Jun 2023   Deviance:
5324.5
Time:                         19:26:29   Pearson chi2:
6.59e+03
No. Iterations:                     20   Pseudo R-squ. (CS):
0.3911
Covariance Type:             nonrobust

======================================================================
==================================================
                                                           coef    std
err          z      P>|z|      [0.025      0.975]
----------------------------------------------------------------------
--------------------------------------------------
const                                                   -3.4879
0.114    -30.606      0.000      -3.711      -3.265
TotalVisits                                              4.6014
1.477      3.115      0.002       1.707       7.496
Total Time Spent on Website                              4.6490
0.166     28.026      0.000       4.324       4.974
Lead Origin_lead add form                                3.6740
0.225     16.361      0.000       3.234       4.114
Lead Source_olark chat                                   1.5975
0.112     14.310      0.000       1.379       1.816
Lead Source_welingak website                             2.6282
1.036      2.536      0.011       0.597       4.659
Last Activity_email bounced                             -1.8760
0.336     -5.576      0.000      -2.535      -1.217
Last Activity_olark chat conversation                   -1.4115
0.167     -8.434      0.000      -1.740      -1.083
Last Activity_sms sent                                   1.2055
0.074     16.383      0.000       1.061       1.350
What is your current occupation_housewife               23.4238
1.14e+04      0.002      0.998   -2.22e+04    2.23e+04
What is your current occupation_other                    2.2289
0.755      2.950      0.003       0.748       3.710
What is your current occupation_student                  1.3076
0.226      5.795      0.000       0.865       1.750
What is your current occupation_unemployed               1.1845
0.086     13.821      0.000       1.017       1.352
What is your current occupation_working professional     3.7363
0.205     18.225      0.000       3.334       4.138
Last Notable Activity_unreachable                        1.8518
0.602      3.078      0.002       0.673       3.031
======================================================================
==================================================
"""
```

```
# 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
11         What is your current occupation_unemployed  2.30
1                        Total Time Spent on Website  2.07
0                                        TotalVisits  1.82
2                          Lead Origin_lead add form  1.59
7                             Last Activity_sms sent  1.55
3                             Lead Source_olark chat  1.51
6               Last Activity_olark chat conversation  1.37
12  What is your current occupation_working profes...  1.32
4                         Lead Source_welingak website  1.31
10            What is your current occupation_student  1.05
5                        Last Activity_email bounced  1.03
9               What is your current occupation_other  1.01
13                  Last Notable Activity_unreachable  1.01
8           What is your current occupation_housewife  1.00
```

```
# Removing What is your current occupation _housewife as it has high p
value
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:
-2670.9
Date:                Thu, 15 Jun 2023   Deviance:
5341.7
Time:                        19:27:54   Pearson chi2:
6.61e+03
No. Iterations:                     7   Pseudo R-squ. (CS):
0.3895
Covariance Type:            nonrobust

======================================================================
==================================================
                                                           coef    std
err          z      P>|z|      [0.025      0.975]
----------------------------------------------------------------------
--------------------------------------------------
const                                                   -3.4579
0.113    -30.555      0.000      -3.680      -3.236
TotalVisits                                              4.5335
1.472      3.080      0.002       1.649       7.418
Total Time Spent on Website                              4.6435
0.166     28.042      0.000       4.319       4.968
Lead Origin_lead add form                                3.6867
0.225     16.419      0.000       3.247       4.127
Lead Source_olark chat                                   1.5866
0.111     14.247      0.000       1.368       1.805
Lead Source_welingak website                             2.6112
1.036      2.520      0.012       0.580       4.642
Last Activity_email bounced                             -1.8831
0.336     -5.600      0.000      -2.542      -1.224
Last Activity_olark chat conversation                   -1.4171
0.167     -8.474      0.000      -1.745      -1.089
Last Activity_sms sent                                   1.1971
0.073     16.298      0.000       1.053       1.341
What is your current occupation_other                    2.2060
0.755      2.920      0.003       0.725       3.686
What is your current occupation_student                  1.2845
0.225      5.697      0.000       0.843       1.726
What is your current occupation_unemployed               1.1625
0.085     13.650      0.000       0.996       1.329
What is your current occupation_working professional     3.7125
0.205     18.134      0.000       3.311       4.114
Last Notable Activity_unreachable                        1.8421
0.601      3.063      0.002       0.663       3.021
======================================================================
==================================================
"""
```

```
# 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
10         What is your current occupation_unemployed  2.30
1                         Total Time Spent on Website  2.06
0                                         TotalVisits  1.82
2                            Lead Origin_lead add form  1.58
7                               Last Activity_sms sent  1.55
3                               Lead Source_olark chat  1.51
6                 Last Activity_olark chat conversation  1.37
11  What is your current occupation_working profes...  1.32
4                         Lead Source_welingak website  1.31
9              What is your current occupation_student  1.05
5                          Last Activity_email bounced  1.03
8                What is your current occupation_other  1.01
12                  Last Notable Activity_unreachable  1.01
```

```
# p value of what is your current occupation_other is not correct
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:
-2675.6
Date:                Thu, 15 Jun 2023   Deviance:
```

```
5351.2
Time:                        19:31:22   Pearson chi2:
6.61e+03
No. Iterations:                     7   Pseudo R-squ. (CS):
0.3886
Covariance Type:            nonrobust

======================================================================
==================================================
                                                             coef    std
err          z      P>|z|      [0.025      0.975]
----------------------------------------------------------------------
--------------------------------------------------
const                                                     -3.4394
0.113    -30.490      0.000      -3.660      -3.218
TotalVisits                                                4.7279
1.483      3.187      0.001       1.820       7.635
Total Time Spent on Website                                4.6530
0.166     28.107      0.000       4.328       4.977
Lead Origin_lead add form                                  3.6934
0.225     16.446      0.000       3.253       4.134
Lead Source_olark chat                                     1.5847
0.111     14.225      0.000       1.366       1.803
Lead Source_welingak website                               2.6117
1.036      2.520      0.012       0.581       4.643
Last Activity_email bounced                               -1.8882
0.336     -5.617      0.000      -2.547      -1.229
Last Activity_olark chat conversation                     -1.4128
0.167     -8.456      0.000      -1.740      -1.085
Last Activity_sms sent                                     1.1913
0.073     16.238      0.000       1.047       1.335
What is your current occupation_student                    1.2606
0.225      5.596      0.000       0.819       1.702
What is your current occupation_unemployed                 1.1385
0.085     13.472      0.000       0.973       1.304
What is your current occupation_working professional     3.6882
0.204     18.039      0.000       3.287       4.089
Last Notable Activity_unreachable                          1.8333
0.601      3.049      0.002       0.655       3.012
======================================================================
==================================================
"""
```

```
# 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
9           What is your current occupation_unemployed  2.30
1                              Total Time Spent on Website  2.06
0                                          TotalVisits  1.82
2                               Lead Origin_lead add form  1.58
7                                  Last Activity_sms sent  1.55
3                                  Lead Source_olark chat  1.51
6                   Last Activity_olark chat conversation  1.37
10  What is your current occupation_working profes...  1.32
4                              Lead Source_welingak website  1.31
8              What is your current occupation_student  1.05
5                             Last Activity_email bounced  1.03
11                     Last Notable Activity_unreachable  1.01
```

Now all the vifs and p values are good

Prediction

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

```
1289    0.649527
3604    0.135329
5584    0.164040
7679    0.135329
7563    0.387899
7978    0.758862
7780    0.155930
7863    0.982089
838     0.776544
708     0.146284
dtype: float64
```

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

```
array([0.64952699, 0.13532885, 0.16403992, 0.13532885, 0.38789903,
       0.75886225, 0.15593025, 0.98208925, 0.77654367, 0.14628394])
```

```
# 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         0.649527
1          0         0.135329
2          0         0.164040
3          0         0.135329
4          0         0.387899
```

```
# 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         0.649527          1
1          0         0.135329          0
2          0         0.164040          0
3          0         0.135329          0
4          0         0.387899          0
```

Evaluating the model

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

```
array([[3442,  453],
       [ 752, 1704]], dtype=int64)
```

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

```
0.810266099826799
```

```
# 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)
```

```
0.6938110749185668
```

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

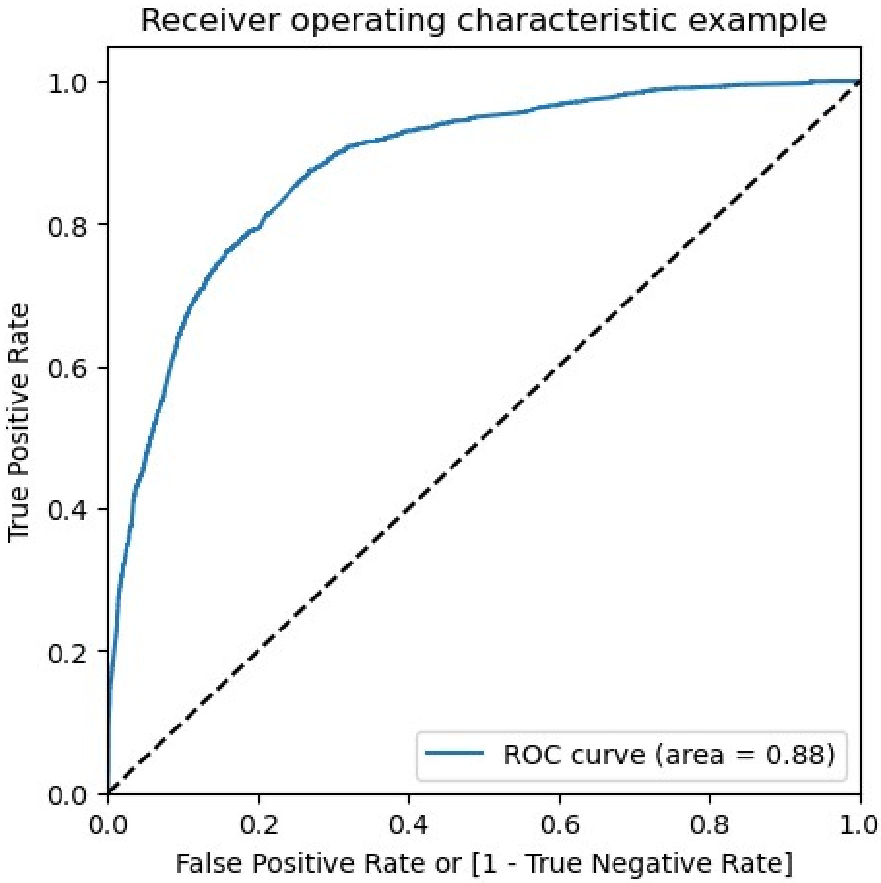
```
0.8836970474967908
```

ROC curve

```
# 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)
```

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



```
# 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()
```

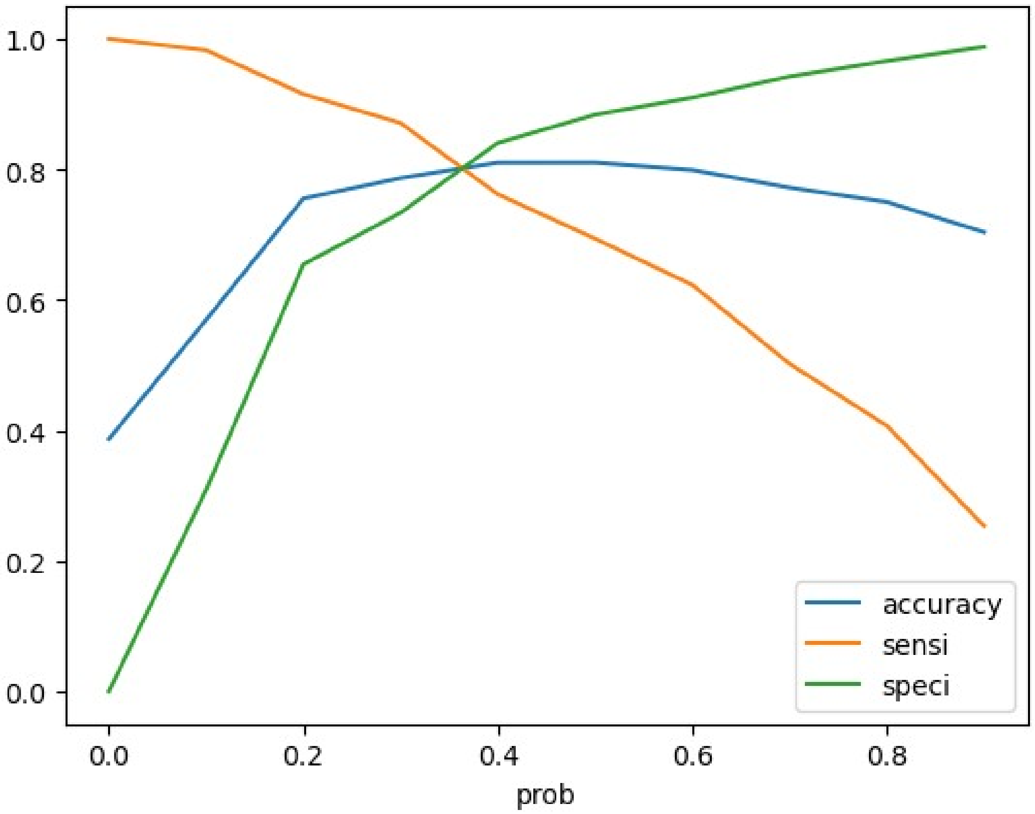
```
   Converted  Conversion_Prob  Predicted  0.0  0.1  0.2  0.3  0.4  0.5
0.6  \
0          1         0.649527          1    1    1    1    1    1    1
1
1          0         0.135329          0    1    1    0    0    0    0
0
2          0         0.164040          0    1    1    0    0    0    0
0
3          0         0.135329          0    1    1    0    0    0    0
0
4          0         0.387899          0    1    1    1    1    0    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  0.569359  0.982492  0.308858
0.2   0.2  0.755314  0.915309  0.654429
0.3   0.3  0.786648  0.870928  0.733504
0.4   0.4  0.810109  0.761808  0.840565
0.5   0.5  0.810266  0.693811  0.883697
0.6   0.6  0.798929  0.622964  0.909884
0.7   0.7  0.772004  0.502036  0.942234
0.8   0.8  0.749961  0.407166  0.966110
0.9   0.9  0.703826  0.253664  0.987677
```

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



```
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         0.649527          1    1    1    1    1    1    1
1
1          0         0.135329          0    1    1    0    0    0    0
0
2          0         0.164040          0    1    1    0    0    0    0
0
3          0         0.135329          0    1    1    0    0    0    0
0
4          0         0.387899          0    1    1    1    1    0    0
0

   0.7  0.8  0.9  final_predicted
0    0    0    0                1
1    0    0    0                0
2    0    0    0                0
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)
```

```
0.8001889466225791
```

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

```
array([[3135,  760],
       [ 509, 1947]], 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)
```

```
0.7927524429967426
```

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

```
0.8048780487804879
```

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
...     ...          ...                          ...
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_olark chat  \
8308                          0                       0
7212                          0                       0
2085                          1                       0
4048                          0                       0
4790                          0                       0
...                         ...                     ...
3261                          0                       1
8179                          0                       0
6236                          0                       1
5240                          0                       0
7243                          0                       0

      Lead Source_welingak website  Last Activity_email bounced  \
8308                             0                            0
7212                             0                            0
2085                             1                            0
4048                             0                            0
4790                             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
...                                     ...                     ...
3261                                      1                       0
8179                                      0                       1
6236                                      0                       0
5240                                      0                       1
7243                                      0                       0
```

```
      What is your current occupation_student  \
8308                                        0   
7212                                        0   
2085                                        0   
4048                                        0   
4790                                        0   
...                                       ...   
3261                                        0   
8179                                        0   
6236                                        0   
5240                                        0   
7243                                        0   

      What is your current occupation_unemployed  \
8308                                           1   
7212                                           0   
2085                                           1   
4048                                           1   
4790                                           1   
...                                          ...   
3261                                           1   
8179                                           0   
6236                                           0   
5240                                           1   
7243                                           1   

      What is your current occupation_working professional  \
8308                                                     0   
7212                                                     1   
2085                                                     0   
4048                                                     0   
4790                                                     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  
...                                 ...  
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         0.451705
1          1         0.829251
2          1         0.982089
3          1         0.869411
4          0         0.105066
```

```
# 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         0.451705                1
1             1         0.829251                1
2             1         0.982089                1
3             1         0.869411                1
4             0         0.105066                0
...         ...              ...              ...
2718          1         0.106317                0
2719          0         0.320571                0
2720          0         0.135329                0
2721          1         0.801105                1
2722          1         0.547662                1

[2723 rows x 3 columns]
```

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

```
0.8013220712449505
```

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

```
array([[1392,  352],
       [ 189,  790]], 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)
```

```
0.8069458631256384
```

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

```
0.7981651376146789
```

Presion and Recall

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

```
array([[3442,  453],
       [ 752, 1704]], dtype=int64)
```

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

```
0.7899860917941586
```

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

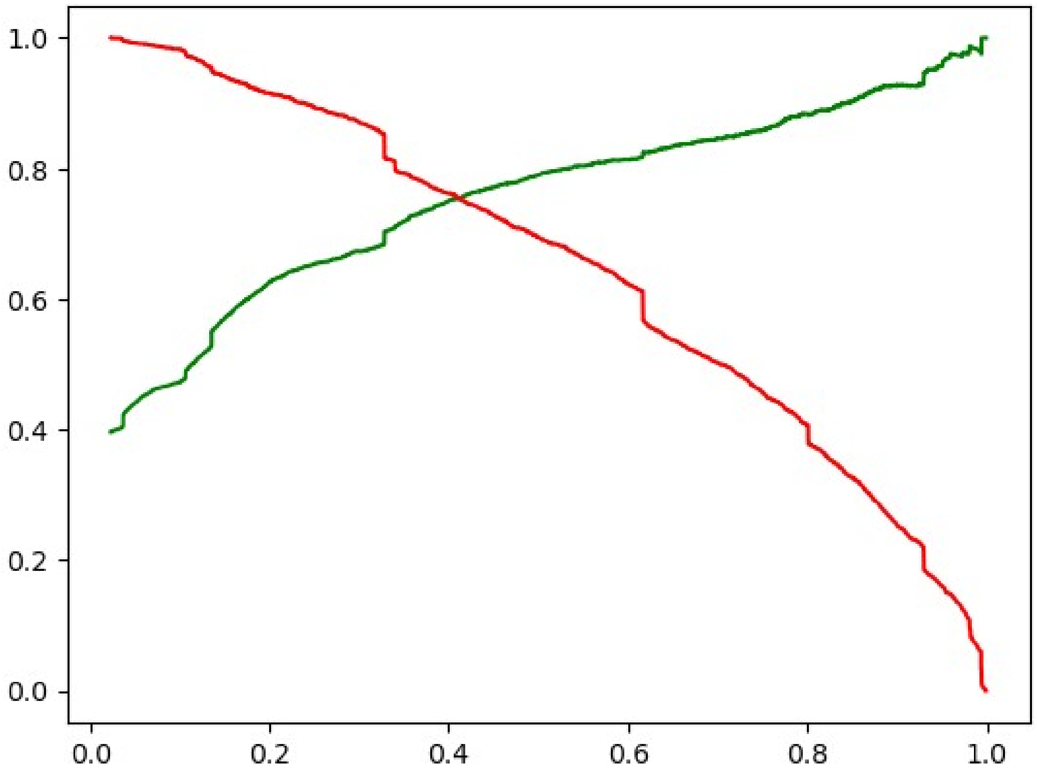
```
0.6938110749185668
```

```
y_train_pred_final.Converted, y_train_pred_final.Predicted
```

```
(0       1
 1       0
 2       0
 3       0
 4       0
```

```
        ..
 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
        ..
 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         0.649527          1    1    1    1    1    1    1
1
1          0         0.135329          0    1    1    0    0    0    0
0
2          0         0.164040          0    1    1    0    0    0    0
0
3          0         0.135329          0    1    1    0    0    0    0
0
4          0         0.387899          0    1    1    1    1    0    0
0

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

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

```
0.809951188789167
```

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

```
array([[3289,  606],
       [ 601, 1855]], 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)
```

```
0.753758634701341
```

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

```
0.7552931596091205
```

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         0.451705
```

```
1          1        0.829251
2          1        0.982089
3          1        0.869411
4          0        0.105066
```

```
# 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         0.451705                1
1             1         0.829251                1
2             1         0.982089                1
3             1         0.869411                1
4             0         0.105066                0
...         ...              ...              ...
2718          1         0.106317                0
2719          0         0.320571                0
2720          0         0.135329                0
2721          1         0.801105                1
2722          1         0.547662                1

[2723 rows x 3 columns]
```

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

```
0.8138082996694822
```

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

```
array([[1470,  274],
       [ 233,  746]], 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)
```

```
0.7313725490196078
```

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

```
0.7620020429009193
```

Conclusion

```
The variable that matter most are as follows in decending order
- The total time spend on website
- The total number of visits
- When the source of lead was
    - Google, Direct traffic, Organic search.
- When the last activity was
    - SMS, Olarck chat conversion
- When the Lead origin is Lead add format
- When their current occupation is waorking as professionals
```