```python
# Supress unnecessary warnings

import warnings
warnings.filterwarnings('ignore')

# Importing the NumPy and Pandas packages

import numpy as np
import pandas as pd

#import visualization libraries
import matplotlib.pyplot as plt
import seaborn as sns
sns.set(style="whitegrid")

#import stats library
from scipy import stats
import statsmodels.api as sm

#import sklearn libraries
from sklearn.model_selection import train_test_split
from sklearn.preprocessing import MinMaxScaler
from sklearn.preprocessing import StandardScaler
from sklearn.linear_model import LogisticRegression
from sklearn.feature_selection import RFE
from statsmodels.stats.outliers_influence import 
variance_inflation_factor
from sklearn import metrics
from sklearn.metrics import 
classification_report,recall_score,roc_auc_score,roc_curve,accuracy_sc
ore,precision_score,precision_recall_curve,confusion_matrix
from sklearn.preprocessing import LabelEncoder

#import miscellaneous libraries
pd.set_option("display.max_columns",None)
pd.set_option("display.max_colwidth",200)
```

Importing the "Leads" Dataset

```python
# Read the dataset
leads = pd.read_csv("Leads.csv")

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            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

  Country           Specialization How did you hear about X Education
\
0     NaN                   Select                             Select

1   India                   Select                             Select

2   India  Business Administration                             Select

3   India    Media and Advertising                       Word Of Mouth

4   India                   Select                              Other


  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 Magazine  \
0                       Better Career Prospects     No       No
1                       Better Career Prospects     No       No
```

```
2                       Better Career Prospects     No       No
3                       Better Career Prospects     No       No
4                       Better Career Prospects     No       No

  Newspaper Article X Education Forums Newspaper Digital Advertisement
\
0                No                 No        No                     No

1                No                 No        No                     No

2                No                 No        No                     No

3                No                 No        No                     No

4                No                 No        No                     No


  Through Recommendations Receive More Updates About Our Courses  \
0                      No                                       No
1                      No                                       No
2                      No                                       No
3                      No                                       No
4                      No                                       No

                                   Tags      Lead Quality  \
0          Interested in other courses  Low in Relevance
1                               Ringing               NaN
2  Will revert after reading the email          Might be
3                               Ringing          Not Sure
4  Will revert after reading the email          Might be

  Update me on Supply Chain Content Get updates on DM Content    Lead
Profile  \
0                                No                       No
Select
1                                No                       No
Select
2                                No                       No
Potential Lead
3                                No                       No
Select
4                                No                       No
Select

     City Asymmetrique Activity Index Asymmetrique Profile Index  \
0  Select                   02.Medium                  02.Medium
1  Select                   02.Medium                  02.Medium
2  Mumbai                   02.Medium                    01.High
3  Mumbai                   02.Medium                    01.High
4  Mumbai                   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
```

```python
#Checking the Shape of dataset
leads.shape
```

```
(9240, 37)
```

```python
# Inspecting the different columns in the dataset

leads.columns
```

```
Index(['Prospect ID', 'Lead Number', 'Lead Origin', 'Lead Source',
       'Do Not Email', 'Do Not Call', 'Converted', 'TotalVisits',
       'Total Time Spent on Website', 'Page Views Per Visit', 'Last
Activity',
       'Country', 'Specialization', 'How did you hear about X
Education',
       'What is your current occupation',
       'What matters most to you in choosing a course', 'Search',
'Magazine',
       'Newspaper Article', 'X Education Forums', 'Newspaper',
       'Digital Advertisement', 'Through Recommendations',
       'Receive More Updates About Our Courses', 'Tags', 'Lead
Quality',
       'Update me on Supply Chain Content', 'Get updates on DM
Content',
       'Lead Profile', 'City', 'Asymmetrique Activity Index',
       'Asymmetrique Profile Index', 'Asymmetrique Activity Score',
       'Asymmetrique Profile Score',
       'I agree to pay the amount through cheque',
       'A free copy of Mastering The Interview', 'Last Notable
```

```
Activity'],
      dtype='object')

# Checking the summary of the dataset
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

# Checking the info to see the types of the feature variables and the
null values present
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
```

As it seems that there are quite a few categorical variables present in this dataset for which we will need to create dummy variables. Also, there are a lot of null values present as well, so we will need to treat them accordingly.

Step 1: Data Cleaning and Preparation

```
# Checking the number of missing values in each column
leads.isnull().sum().sort_values(ascending=False)
```

```
Lead Quality                                   4767
Asymmetrique Activity Index                    4218
Asymmetrique Profile Score                     4218
Asymmetrique Activity Score                    4218
Asymmetrique Profile Index                     4218
Tags                                           3353
```

```
Lead Profile                                     2709
What matters most to you in choosing a course    2709
What is your current occupation                  2690
Country                                          2461
How did you hear about X Education               2207
Specialization                                   1438
City                                             1420
Page Views Per Visit                              137
TotalVisits                                       137
Last Activity                                     103
Lead Source                                        36
Receive More Updates About Our Courses              0
I agree to pay the amount through cheque            0
Get updates on DM Content                           0
Update me on Supply Chain Content                   0
A free copy of Mastering The Interview              0
Prospect ID                                         0
Newspaper Article                                   0
Through Recommendations                             0
Digital Advertisement                               0
Newspaper                                           0
X Education Forums                                  0
Lead Number                                         0
Magazine                                            0
Search                                              0
Total Time Spent on Website                         0
Converted                                           0
Do Not Call                                         0
Do Not Email                                        0
Lead Origin                                         0
Last Notable Activity                               0
dtype: int64
```

As it is clearly seen there are a lot of columns which have high number of missing values. Clearly, these columns are not useful. Since, there are 9000 datapoints in our dataframe, let's eliminate the columns having greater than 3000 missing values as they are of no use to us.

```
# Droping all the columns in which greater than
for c in leads.columns:
    if leads[c].isnull().sum()>3000:
        leads.drop(c, axis=1,inplace=True)

leads.isnull().sum().sort_values(ascending=False)
```

```
What matters most to you in choosing a course    2709
Lead Profile                                     2709
What is your current occupation                  2690
Country                                          2461
How did you hear about X Education               2207
```

```
Specialization                                   1438
City                                             1420
Page Views Per Visit                              137
TotalVisits                                       137
Last Activity                                     103
Lead Source                                        36
Get updates on DM Content                           0
Newspaper                                           0
I agree to pay the amount through cheque            0
A free copy of Mastering The Interview              0
Update me on Supply Chain Content                   0
Receive More Updates About Our Courses              0
Through Recommendations                             0
Digital Advertisement                               0
Prospect ID                                         0
X Education Forums                                  0
Newspaper Article                                   0
Magazine                                            0
Search                                              0
Lead Number                                         0
Total Time Spent on Website                         0
Converted                                           0
Do Not Call                                         0
Do Not Email                                        0
Lead Origin                                         0
Last Notable Activity                               0
dtype: int64
```

```
#checking value counts of "City" column
leads['City'].value_counts(dropna=False)
```

```
Mumbai                          3222
Select                          2249
NaN                             1420
Thane & Outskirts                752
Other Cities                     686
Other Cities of Maharashtra      457
Other Metro Cities               380
Tier II Cities                    74
Name: City, dtype: int64
```

`Mumbai` has highest numbers of leads

As you might be able to interpret, the variable `City` won't be of any use in our analysis. So it's best that we drop it.

```
# dropping the "City" feature
leads.drop(['City'], axis = 1, inplace = True)
```

```python
#checking value counts of "Country" column
leads['Country'].value_counts(dropna=False)
```

```
India                   6492
NaN                     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
```

Highest number of leads from `INDIA`

```python
# dropping the "Country" feature
leads.drop(['Country'], axis = 1, inplace = True)
```

```python
#Now checking the percentage of missing values in each column

round(100*(leads.isnull().sum()/len(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
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
Update me on Supply Chain Content                 0.00
Get updates on DM Content                         0.00
Lead Profile                                     29.32
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
```

```python
# Checking the number of null values again
leads.isnull().sum().sort_values(ascending=False)
```

```
What matters most to you in choosing a course    2709
Lead Profile                                     2709
What is your current occupation                  2690
How did you hear about X Education               2207
Specialization                                   1438
TotalVisits                                       137
Page Views Per Visit                              137
Last Activity                                     103
Lead Source                                        36
Get updates on DM Content                           0
Update me on Supply Chain Content                   0
X Education Forums                                  0
```

```
I agree to pay the amount through cheque          0
A free copy of Mastering The Interview            0
Receive More Updates About Our Courses            0
Through Recommendations                           0
Digital Advertisement                             0
Newspaper                                         0
Prospect ID                                       0
Newspaper Article                                 0
Magazine                                          0
Search                                            0
Lead Number                                       0
Total Time Spent on Website                       0
Converted                                         0
Do Not Call                                       0
Do Not Email                                      0
Lead Origin                                       0
Last Notable Activity                             0
dtype: int64
```

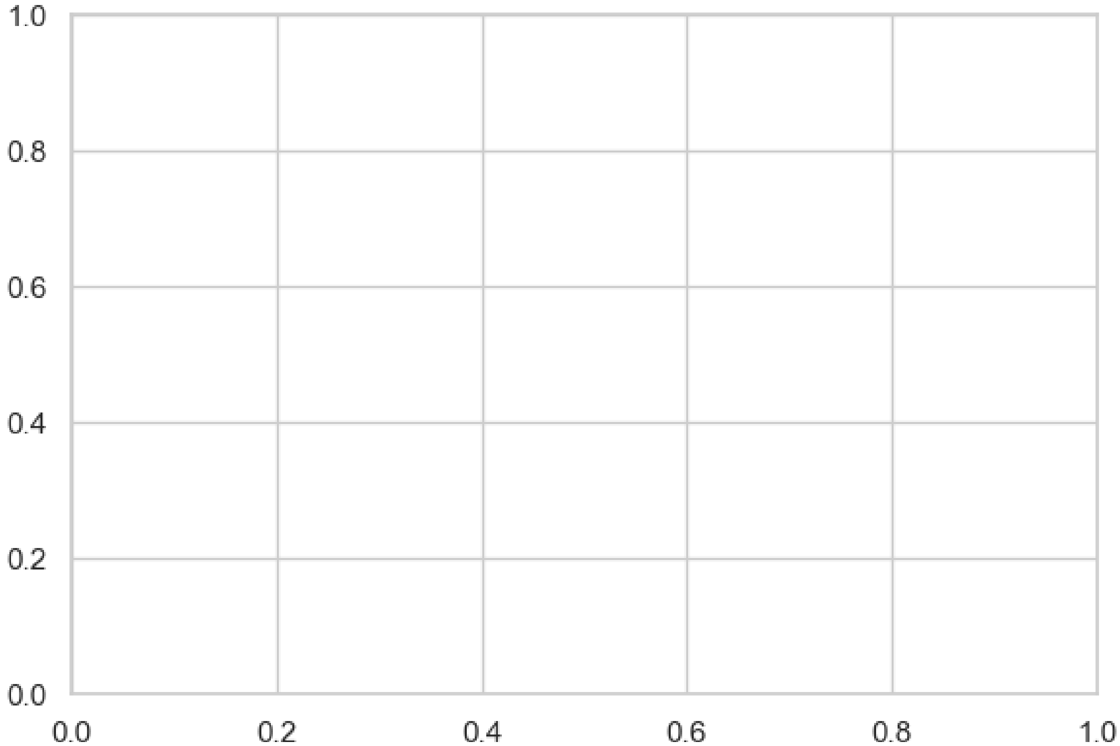
Visualizing the features with `Select` values

```
def countplot(x, fig):
    plt.subplot(2,2, fig)
    sns.countplot(leads[x])
    plt.title('Count across'+' '+ x, size = 16)
    plt.xlabel(x,size = 14)
    plt.xticks(rotation = 90)

plt.figure(figsize=(15,10))

countplot('How did you hear about X Education',1)
countplot('Lead Profile',2)
countplot('Specialization',3)



plt.tight_layout()
```

```
---------------------------------------------------------------------------
-----
ValueError                                Traceback (most recent call
last)
d:\Users\hp\Downloads\Lead Scoring Case Study SR_HN_JJ (1).ipynb Cell
26 line 1
      <a href='vscode-notebook-cell:/d%3A/Users/hp/Downloads/Lead
%20Scoring%20Case%20Study%20SR_HN_JJ%20%281%29.ipynb#X34sZmlsZQ%3D%3D?
line=5'>6</a>     plt.xticks(rotation = 90)
      <a href='vscode-notebook-cell:/d%3A/Users/hp/Downloads/Lead
%20Scoring%20Case%20Study%20SR_HN_JJ%20%281%29.ipynb#X34sZmlsZQ%3D%3D?
line=7'>8</a> plt.figure(figsize=(15,10))
```

```
---> <a href='vscode-notebook-cell:/d%3A/Users/hp/Downloads/Lead
%20Scoring%20Case%20Study%20SR_HN_JJ%20%281%29.ipynb#X34sZmlsZQ%3D%3D?
line=9'>10</a> countplot('How did you hear about X Education',1)
     <a href='vscode-notebook-cell:/d%3A/Users/hp/Downloads/Lead
%20Scoring%20Case%20Study%20SR_HN_JJ%20%281%29.ipynb#X34sZmlsZQ%3D%3D?
line=10'>11</a> countplot('Lead Profile',2)
     <a href='vscode-notebook-cell:/d%3A/Users/hp/Downloads/Lead
%20Scoring%20Case%20Study%20SR_HN_JJ%20%281%29.ipynb#X34sZmlsZQ%3D%3D?
line=11'>12</a> countplot('Specialization',3)

d:\Users\hp\Downloads\Lead Scoring Case Study SR_HN_JJ (1).ipynb Cell
26 line 3
      <a href='vscode-notebook-cell:/d%3A/Users/hp/Downloads/Lead
%20Scoring%20Case%20Study%20SR_HN_JJ%20%281%29.ipynb#X34sZmlsZQ%3D%3D?
line=0'>1</a> def countplot(x, fig):
      <a href='vscode-notebook-cell:/d%3A/Users/hp/Downloads/Lead
%20Scoring%20Case%20Study%20SR_HN_JJ%20%281%29.ipynb#X34sZmlsZQ%3D%3D?
line=1'>2</a>     plt.subplot(2,2, fig)
----> <a href='vscode-notebook-cell:/d%3A/Users/hp/Downloads/Lead
%20Scoring%20Case%20Study%20SR_HN_JJ%20%281%29.ipynb#X34sZmlsZQ%3D%3D?
line=2'>3</a>     sns.countplot(leads[x])
      <a href='vscode-notebook-cell:/d%3A/Users/hp/Downloads/Lead
%20Scoring%20Case%20Study%20SR_HN_JJ%20%281%29.ipynb#X34sZmlsZQ%3D%3D?
line=3'>4</a>     plt.title('Count across'+' '+ x, size = 16)
      <a href='vscode-notebook-cell:/d%3A/Users/hp/Downloads/Lead
%20Scoring%20Case%20Study%20SR_HN_JJ%20%281%29.ipynb#X34sZmlsZQ%3D%3D?
line=4'>5</a>     plt.xlabel(x,size = 14)

File c:\Users\hp\anaconda3\Lib\site-packages\seaborn\
categorical.py:2943, in countplot(data, x, y, hue, order, hue_order,
orient, color, palette, saturation, width, dodge, ax, **kwargs)
   2940 elif x is not None and y is not None:
   2941     raise ValueError("Cannot pass values for both `x` and
`y`")
-> 2943 plotter = _CountPlotter(
   2944     x, y, hue, data, order, hue_order,
   2945     estimator, errorbar, n_boot, units, seed,
   2946     orient, color, palette, saturation,
   2947     width, errcolor, errwidth, capsize, dodge
   2948 )
   2950 plotter.value_label = "count"
   2952 if ax is None:

File c:\Users\hp\anaconda3\Lib\site-packages\seaborn\
categorical.py:1530, in _BarPlotter.__init__(self, x, y, hue, data,
order, hue_order, estimator, errorbar, n_boot, units, seed, orient,
color, palette, saturation, width, errcolor, errwidth, capsize, dodge)
   1525 def __init__(self, x, y, hue, data, order, hue_order,
   1526              estimator, errorbar, n_boot, units, seed,
   1527              orient, color, palette, saturation, width,
```

```
   1528                errrcolor, errwidth, capsize, dodge):
   1529     """Initialize the plotter."""
-> 1530     self.establish_variables(x, y, hue, data, orient,
   1531                              order, hue_order, units)
   1532     self.establish_colors(color, palette, saturation)
   1533     self.estimate_statistic(estimator, errorbar, n_boot, seed)

File c:\Users\hp\anaconda3\Lib\site-packages\seaborn\
categorical.py:516, in _CategoricalPlotter.establish_variables(self,
x, y, hue, data, orient, order, hue_order, units)
    513     plot_data = data
    515 # Convert to a list of arrays, the common representation
--> 516 plot_data = [np.asarray(d, float) for d in plot_data]
    518 # The group names will just be numeric indices
    519 group_names = list(range(len(plot_data)))

File c:\Users\hp\anaconda3\Lib\site-packages\seaborn\
categorical.py:516, in <listcomp>(.0)
    513     plot_data = data
    515 # Convert to a list of arrays, the common representation
--> 516 plot_data = [np.asarray(d, float) for d in plot_data]
    518 # The group names will just be numeric indices
    519 group_names = list(range(len(plot_data)))

File c:\Users\hp\anaconda3\Lib\site-packages\pandas\core\
series.py:893, in Series.__array__(self, dtype)
    846 def __array__(self, dtype: npt.DTypeLike | None = None) ->
np.ndarray:
    847     """
    848     Return the values as a NumPy array.
    849
   (...)
    891             dtype='datetime64[ns]')
    892     """
--> 893     return np.asarray(self._values, dtype)

ValueError: could not convert string to float: 'Select'
```



there are a few columns in which there is a level called `'Select'` which basically means that the student had not selected the option for that particular column which is why it shows 'Select'. These values are as good as missing values and hence we need to identify the value counts of the level 'Select' in all the columns that it is present.

```
# checking the value counts of all the columns

for c in leads:
    print(leads[c].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
___________________________________________________
No     8506
Yes     734
Name: Do Not Email, dtype: int64
___________________________________________________
No     9238
Yes       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
___________________________________________________
Select                               1942
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
___________________________________________________
Select                     5043
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
___________________________________________________
No     9226
Yes      14
Name: Search, dtype: int64
___________________________________________________
No    9240
Name: Magazine, dtype: int64
___________________________________________________
No     9238
Yes       2
```

```
Name: Newspaper Article, dtype: int64
____________________________________________________
No     9239
Yes       1
Name: X Education Forums, dtype: int64
____________________________________________________
No     9239
Yes       1
Name: Newspaper, dtype: int64
____________________________________________________
No     9236
Yes       4
Name: Digital Advertisement, dtype: int64
____________________________________________________
No     9233
Yes       7
Name: Through Recommendations, dtype: int64
____________________________________________________
No    9240
Name: Receive More Updates About Our Courses, dtype: int64
____________________________________________________
No    9240
Name: Update me on Supply Chain Content, dtype: int64
____________________________________________________
No    9240
Name: Get updates on DM Content, dtype: int64
____________________________________________________
Select                           4146
Potential Lead                   1613
Other Leads                       487
Student of SomeSchool             241
Lateral Student                    24
Dual Specialization Student        20
Name: Lead Profile, dtype: int64
____________________________________________________
No    9240
Name: I agree to pay the amount through cheque, dtype: int64
____________________________________________________
No     6352
Yes    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
```

The following three columns now have the level 'Select'. Let's check them once again.

```
leads['Lead Profile'].astype('category').value_counts()
```

```
Select                               4146
Potential Lead                       1613
Other Leads                           487
Student of SomeSchool                 241
Lateral Student                        24
Dual Specialization Student            20
Name: Lead Profile, dtype: int64
```

```
leads['How did you hear about X Education'].value_counts()
```

```
Select                         5043
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
```

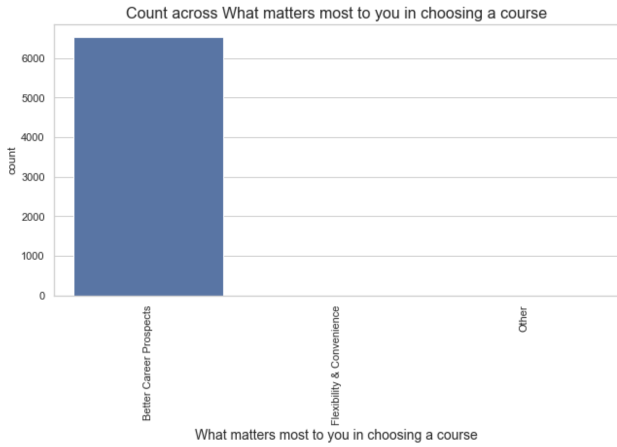
```
leads['Specialization'].value_counts()
```

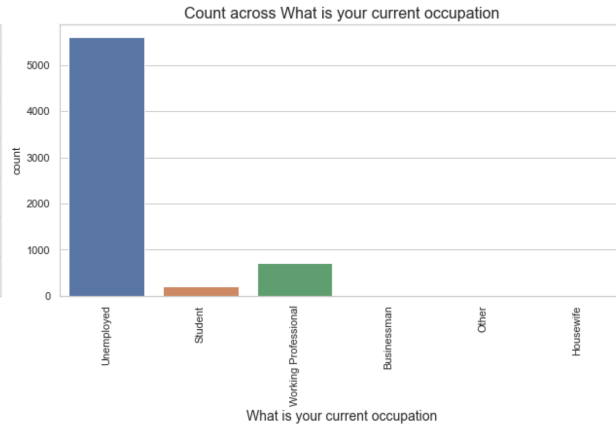
```
Select                                     1942
Finance Management                          976
Human Resource Management                   848
Marketing Management                        838
Operations Management                       503
Business Administration                     403
IT Projects Management                      366
Supply Chain Management                     349
Banking, Investment And Insurance           338
Travel and Tourism                          203
```

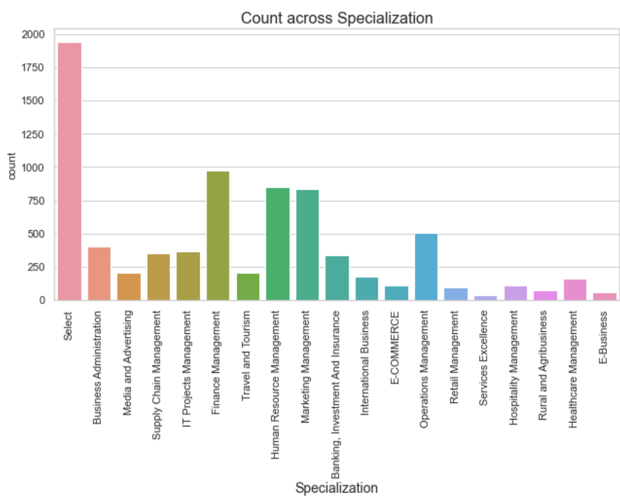
```
Media and Advertising                      203
International Business                     178
Healthcare Management                      159
Hospitality Management                     114
E-COMMERCE                                 112
Retail Management                          100
Rural and Agribusiness                      73
E-Business                                  57
Services Excellence                         40
Name: Specialization, dtype: int64
```

Visualizing the features

```
def countplot(x, fig):
    plt.subplot(4,2, fig)
    sns.countplot(leads[x])
    plt.title('Count across'+' '+ x, size = 16)
    plt.xlabel(x,size = 14)
    plt.xticks(rotation = 90)

plt.figure(figsize=(18,25))


countplot('What matters most to you in choosing a course',1)
countplot('What is your current occupation',2)
countplot('Specialization',3)

plt.tight_layout()
```







As it can be seen that the levels of `"Lead Profile"` and `"How did you hear about X Education"` have a lot of rows which have the value Select which is of no use to the analysis

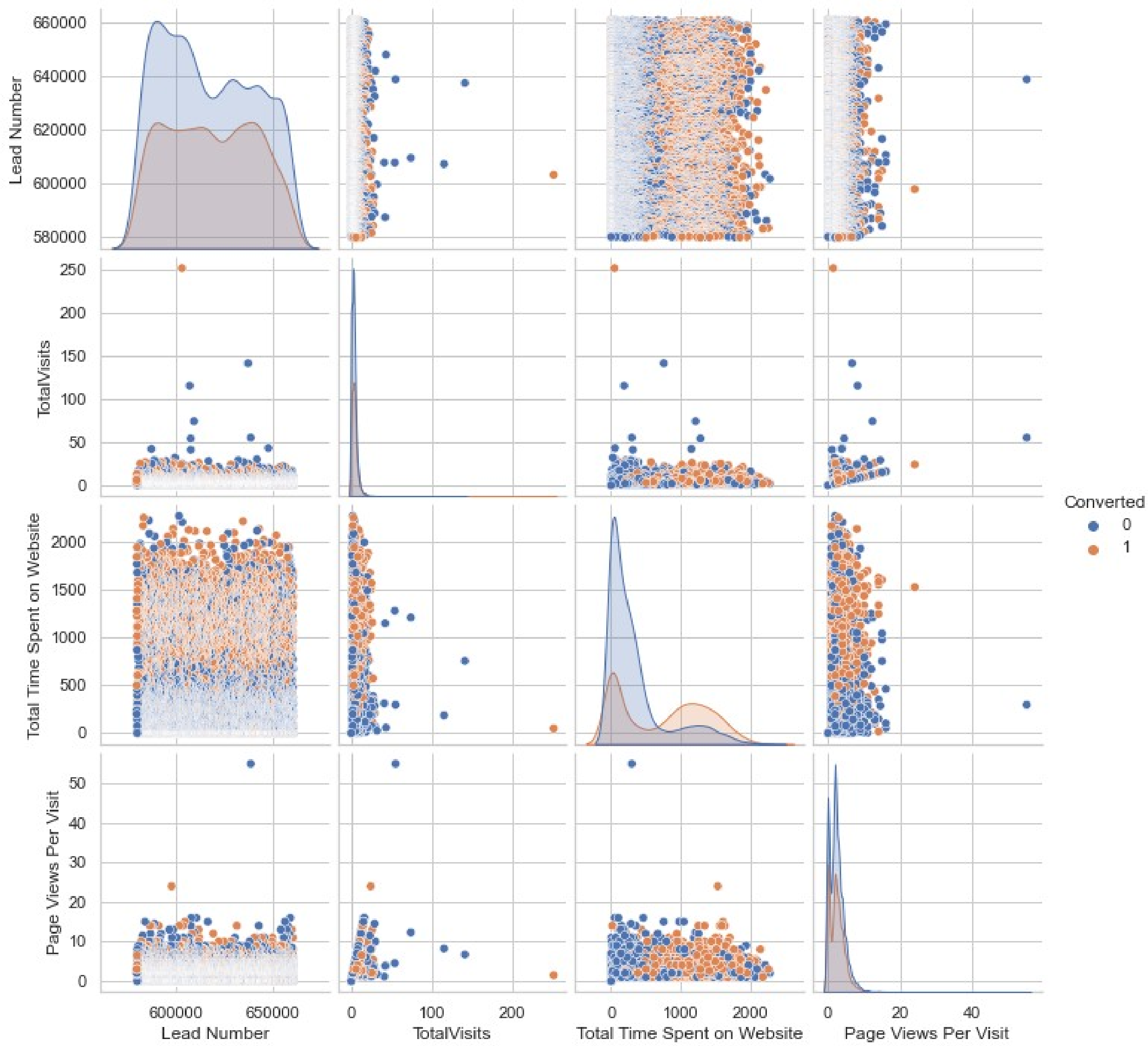
So it's best that we drop them.

```
# dropping Lead Profile and How did you hear about X Education cols
leads.drop(['Lead Profile', 'How did you hear about X Education'],
axis = 1, inplace = True)
```

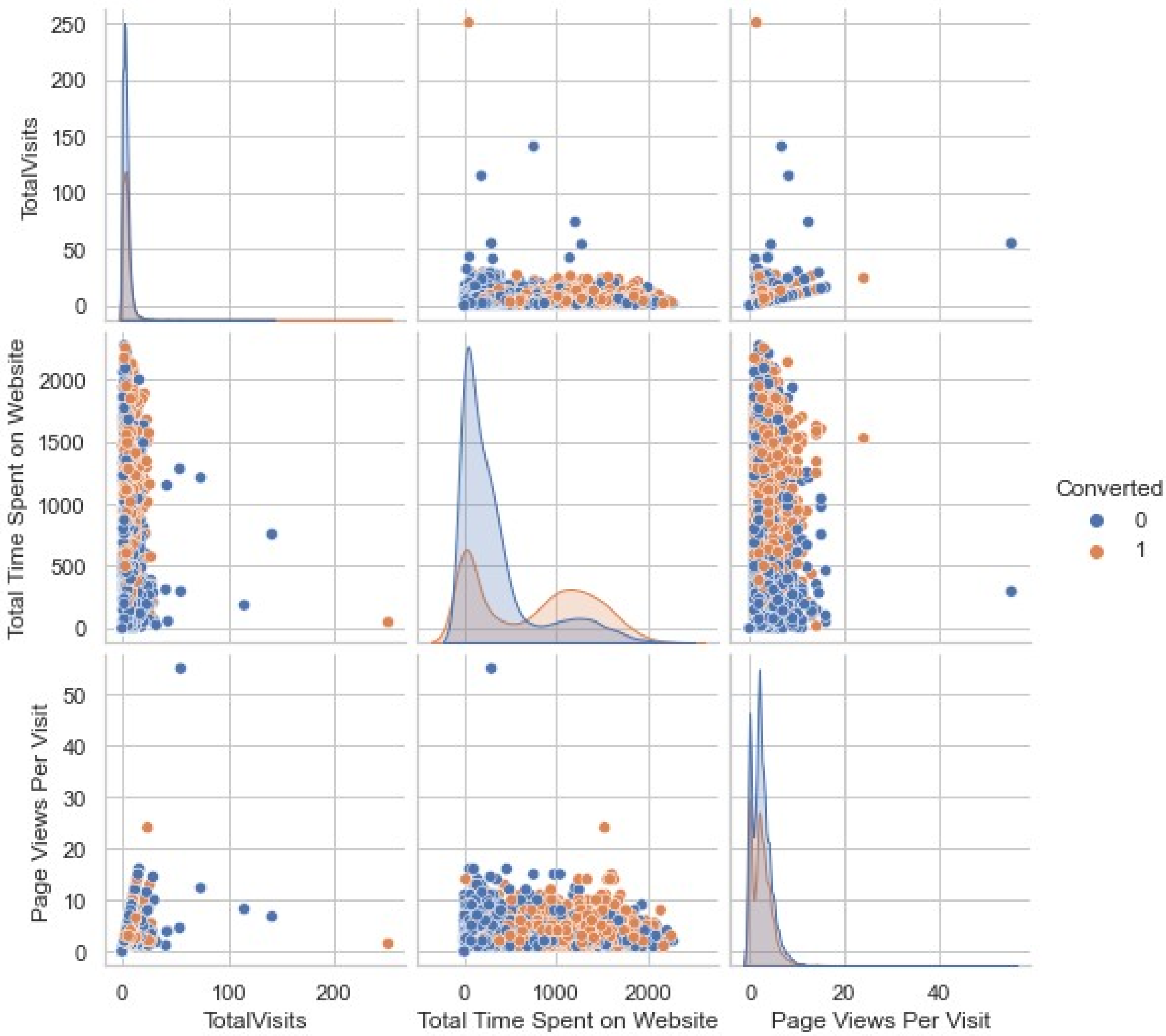
Also we notice that, when we got the value counts of all the columns, there were a few columns in which only one value was majorly present for all the data points. These include `Do Not Call`, `Search`, `Magazine`, `Newspaper Article`, `X Education Forums`, `Newspaper`, `Digital Advertisement`, `Through 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`. Since practically all of the values for these variables are `No`, it's best that we drop these columns as they won't help with our analysis.

```
from matplotlib import pyplot as plt
import seaborn as sns
```

```
sns.pairplot(leads,diag_kind='kde',hue='Converted')
plt.show()
```



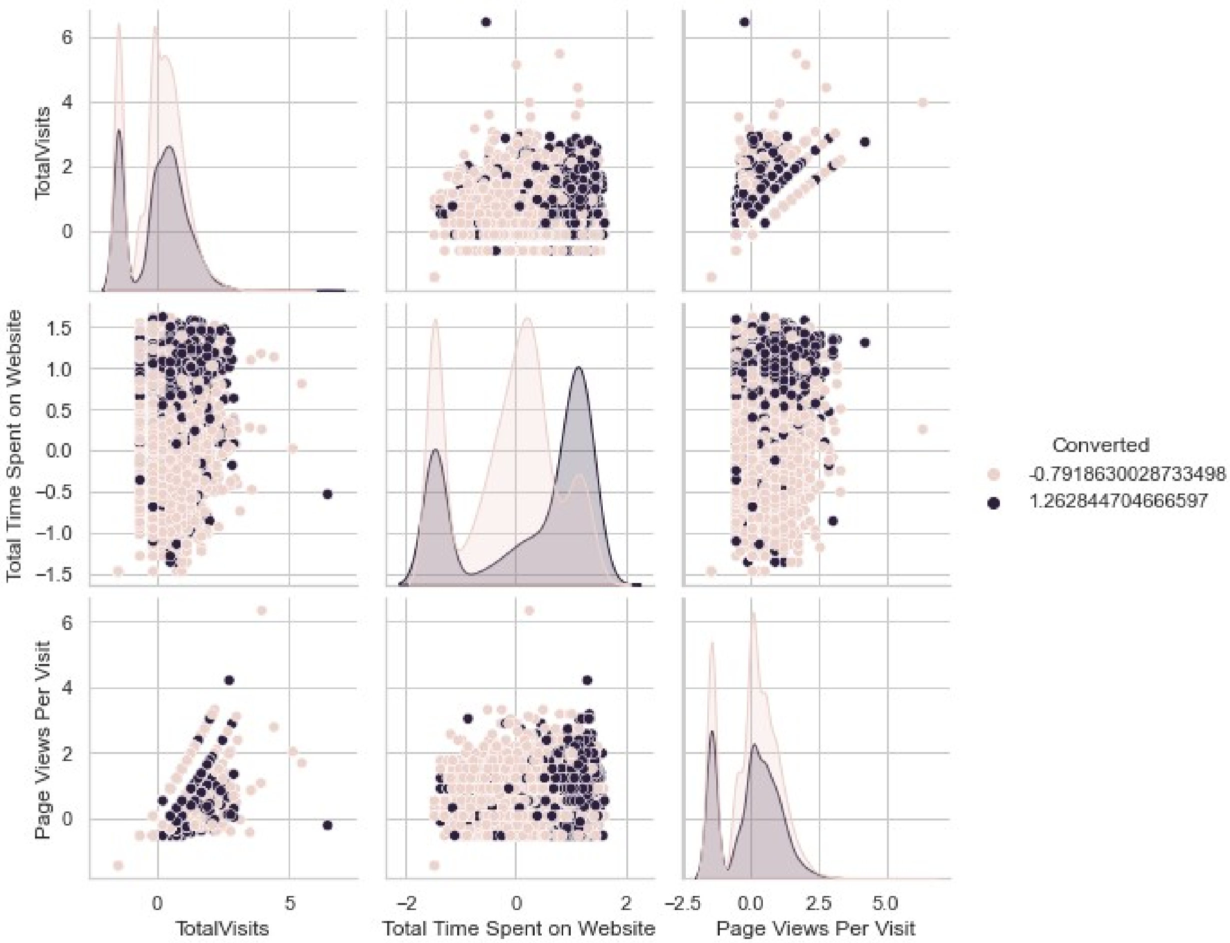

```
x_edu = leads[['TotalVisits','Total Time Spent on Website','Page Views Per Visit','Converted']]
sns.pairplot(x_edu,diag_kind='kde',hue='Converted')
plt.show()
```




```
from sklearn.preprocessing import PowerTransformer
pt = PowerTransformer()
transformedx_edu = pd.DataFrame(pt.fit_transform(x_edu))
transformedx_edu.columns = x_edu.columns
transformedx_edu.head()
```

```
   TotalVisits  Total Time Spent on Website  Page Views Per Visit
Converted
0    -1.457907                    -1.473767             -1.454706  -
0.791863
1     0.747918                     0.729628              0.308534  -
0.791863
2    -0.141636                     1.306093              0.065574
1.262845
3    -0.640428                     0.264936             -0.536967  -
0.791863
```

```
4    -0.141636                          1.252499             -0.536967
1.262845

sns.pairplot(transformedx_edu,diag_kind='kde',hue='Converted')
plt.show()
```




```
# Dropping the above columns

leads.drop(['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'], axis = 1,
inplace = True)

leads['What matters most to you in choosing a course'].value_counts()

Better Career Prospects      6528
Flexibility & Convenience       2
```

```
Other                                       1
Name: What matters most to you in choosing a course, dtype: int64
```

The variable `What matters most to you in choosing a course` has the `level Better Career Prospects` 6528 times while the other two levels appear once twice and once respectively.

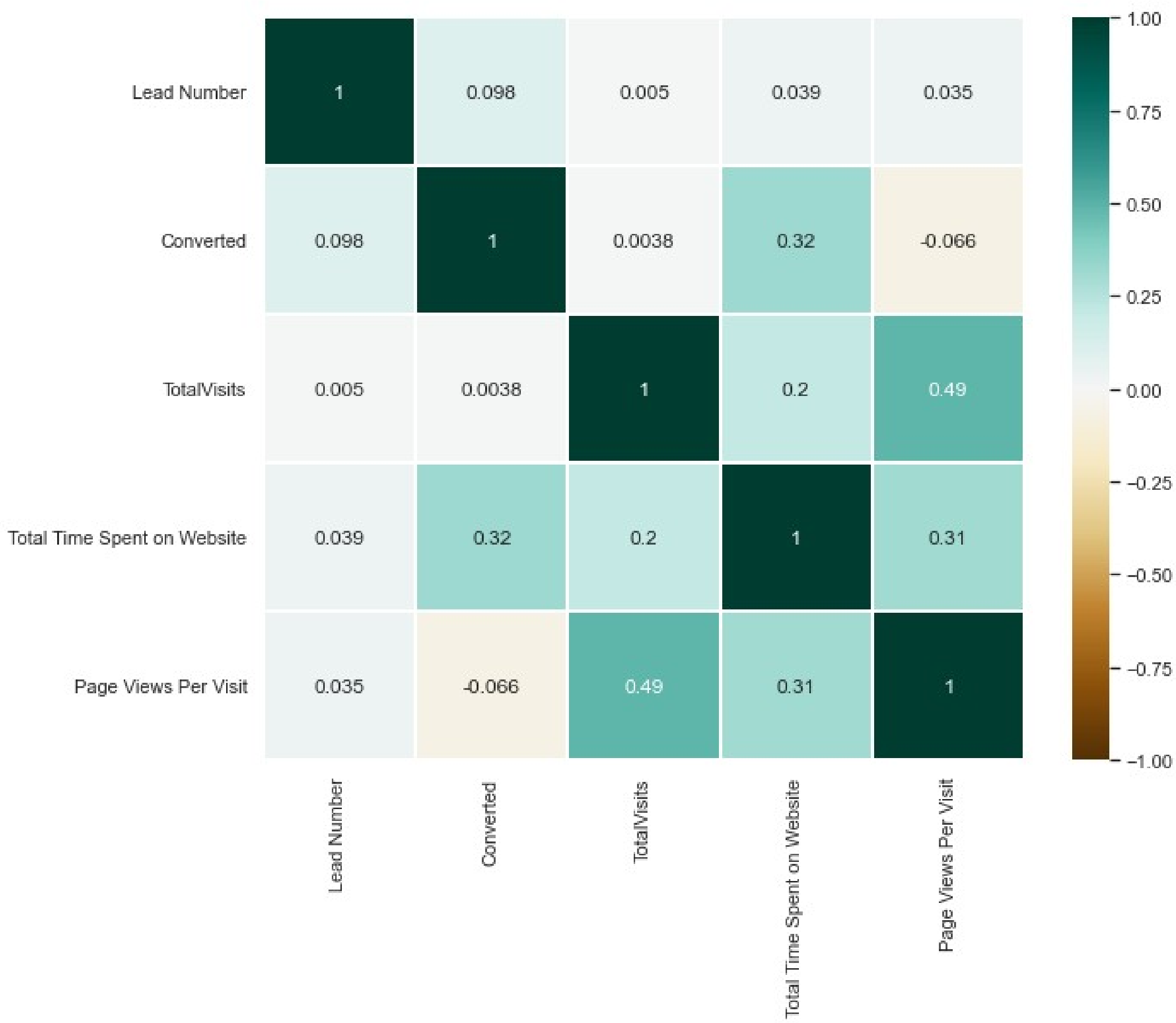
So we should dropping this column as well.

```
leads.drop(['What matters most to you in choosing a course'], axis = 1, inplace=True)

# Checking the number of null values again
leads.isnull().sum().sort_values(ascending=False)
```

```
What is your current occupation           2690
Specialization                            1438
TotalVisits                                137
Page Views Per Visit                       137
Last Activity                              103
Lead Source                                 36
Prospect ID                                  0
Lead Number                                  0
Lead Origin                                  0
Do Not Email                                 0
Converted                                    0
Total Time Spent on Website                  0
A free copy of Mastering The Interview       0
Last Notable Activity                        0
dtype: int64
```

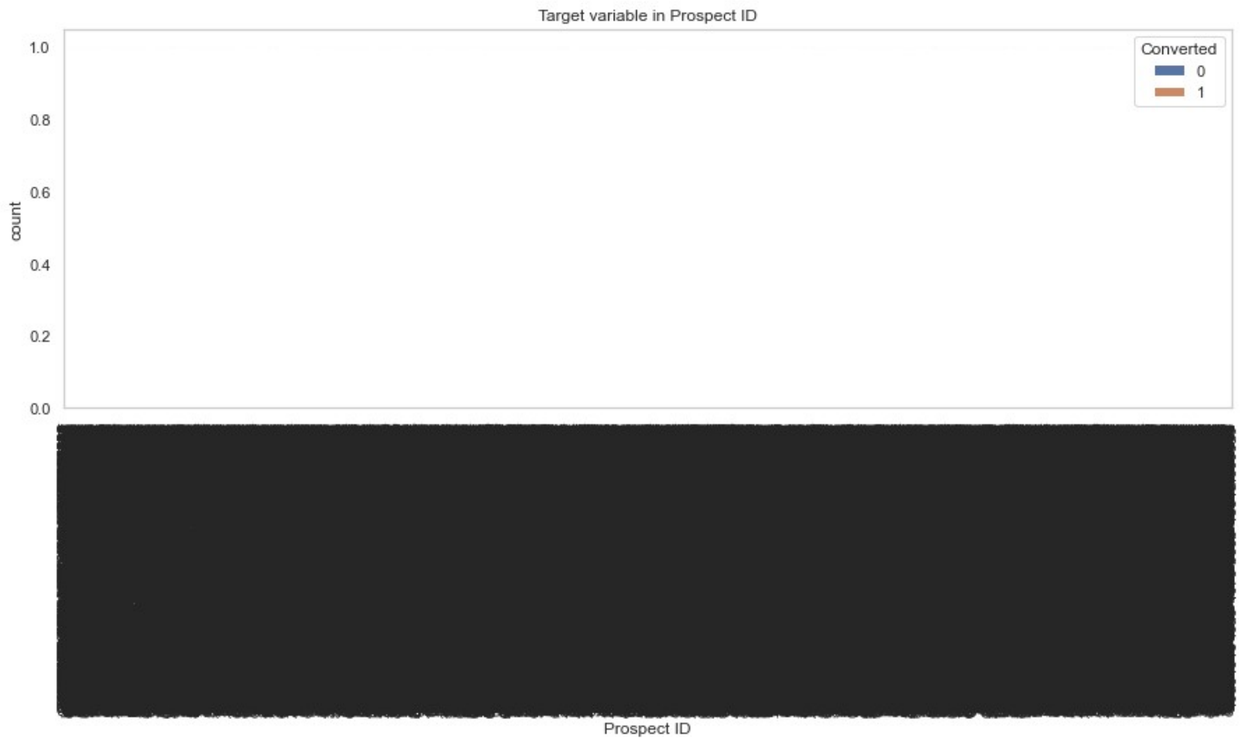
Now, there's the column `What is your current occupation` which has a lot of null values. Now you can drop the entire row but since we have already lost so many feature variables, we choose not to drop it as it might turn out to be significant in the analysis. So let's just drop the null rows for the column `What is you current occupation`.

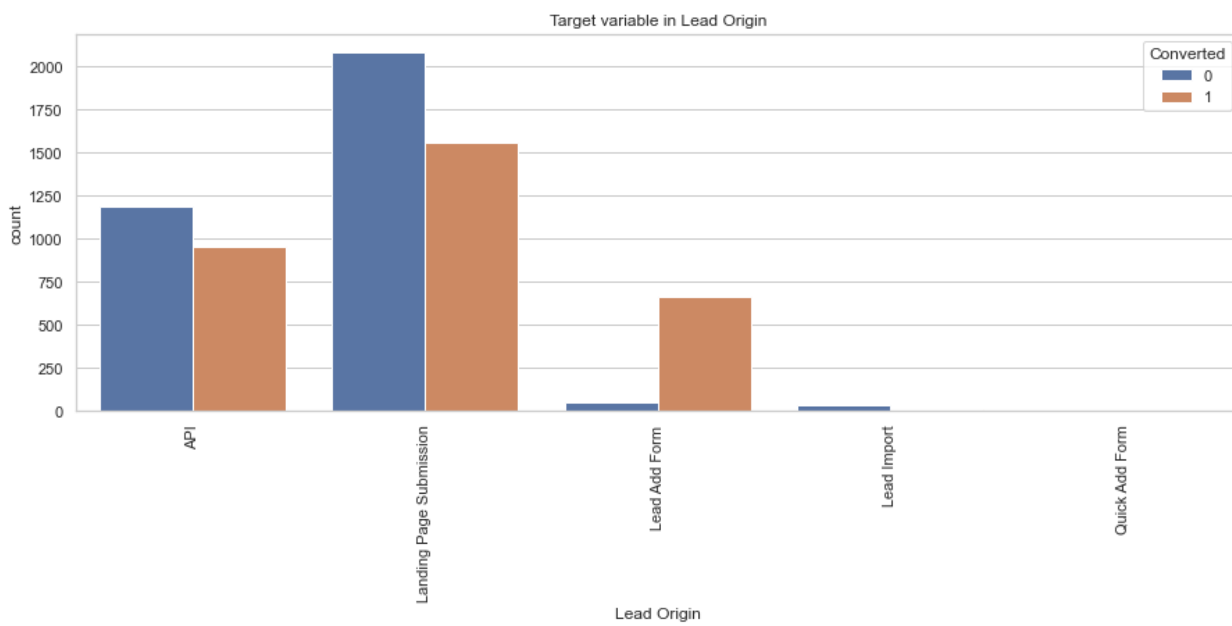
```
# Dropping the null values rows in the column 'What is your current occupation'

leads = leads[~pd.isnull(leads['What is your current occupation'])]

# Observing Correlation
# figure size
plt.figure(figsize=(10,8))

# heatmap
sns.heatmap(leads.corr(), annot=True,cmap="BrBG", robust=True,linewidth=0.1, vmin=-1 )
plt.show()
```

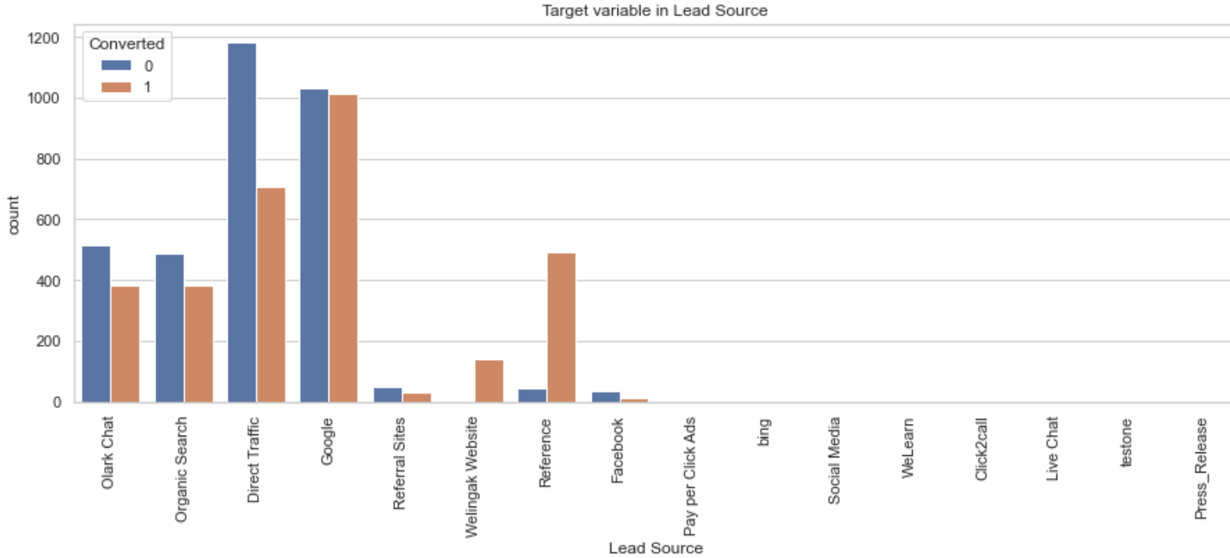


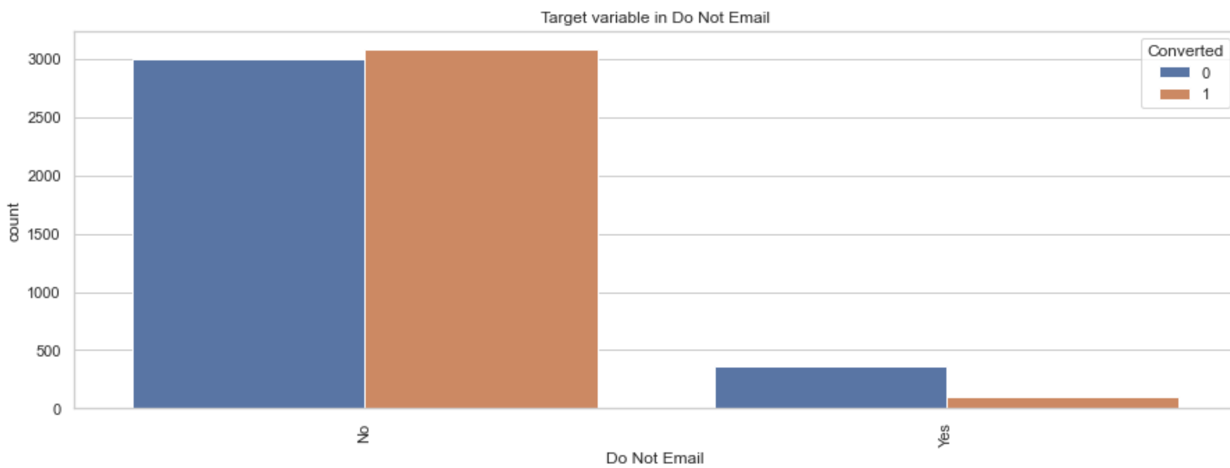
Analysing Categorical features

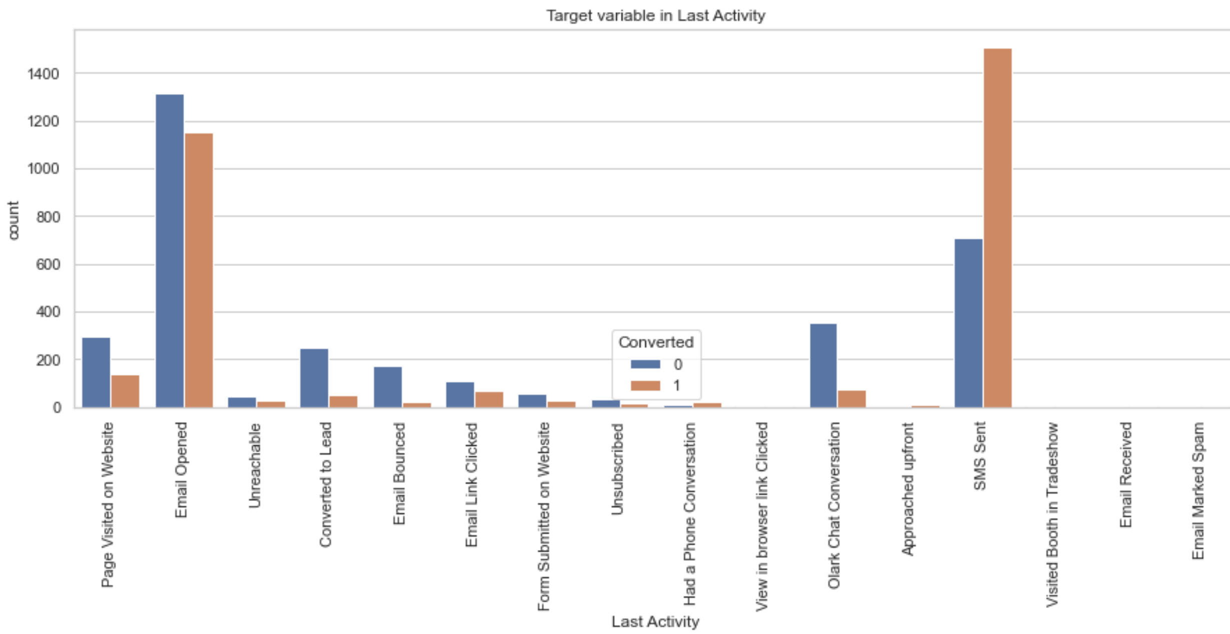
```
conv = leads.select_dtypes(include ="object").columns
for i in conv:

    plt.figure(figsize =(15,5))
    sns.countplot(leads[i], hue=leads.Converted)
    plt.xticks(rotation = 90)
    plt.title('Target variable in'+' '+ i)
    plt.xlabel(i)
    plt.show()
```


Target variable in Prospect ID
Converted
0
1
count
1.0
0.8
0.6
0.4
0.2
0.0
Prospect ID

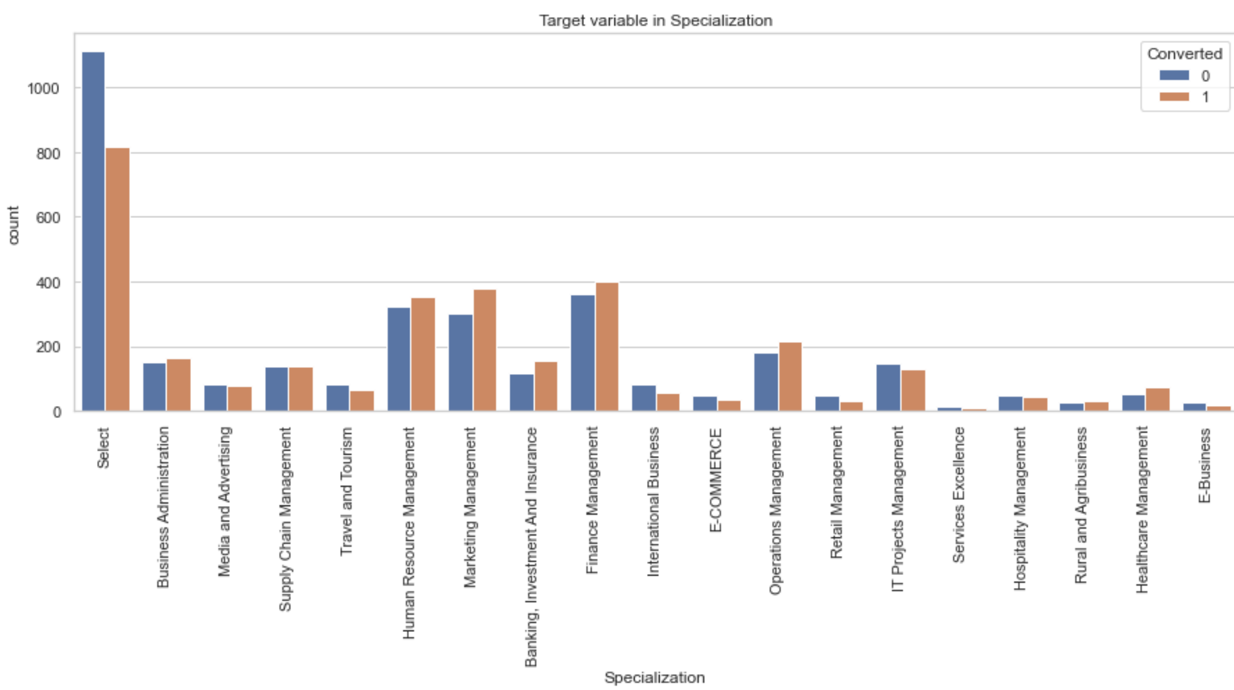

Target variable in Lead Origin
Converted
0
1
count
2000
1750
1500
1250
1000
750
500
250
0
API
Landing Page Submission
Lead Add Form
Lead Import
Quick Add Form
Lead Origin


Target variable in Lead Source
Converted
0
1
count
1200
1000
800
600
400
200
0
Olark Chat
Organic Search
Direct Traffic
Google
Referral Sites
Welingak Website
Reference
Facebook
Pay per Click Ads
bing
Social Media
WeLearn
Click2call
Live Chat
testone
Press_Release
Lead Source

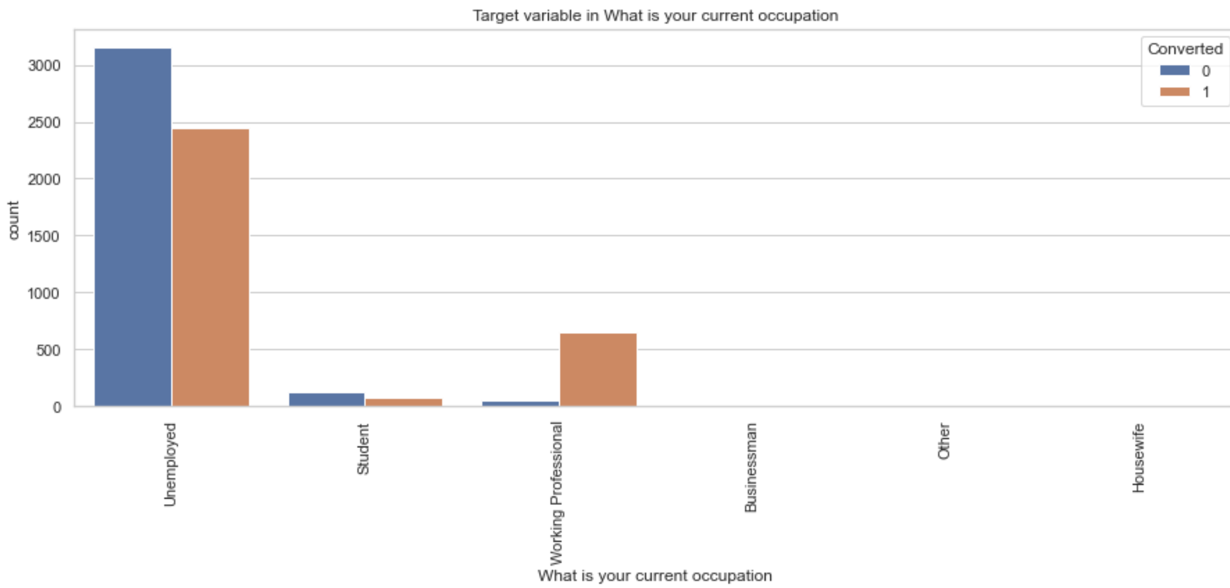

Target variable in Do Not Email
Converted
0
1
count
3000
2500
2000
1500
1000
500
0
No
Yes
Do Not Email

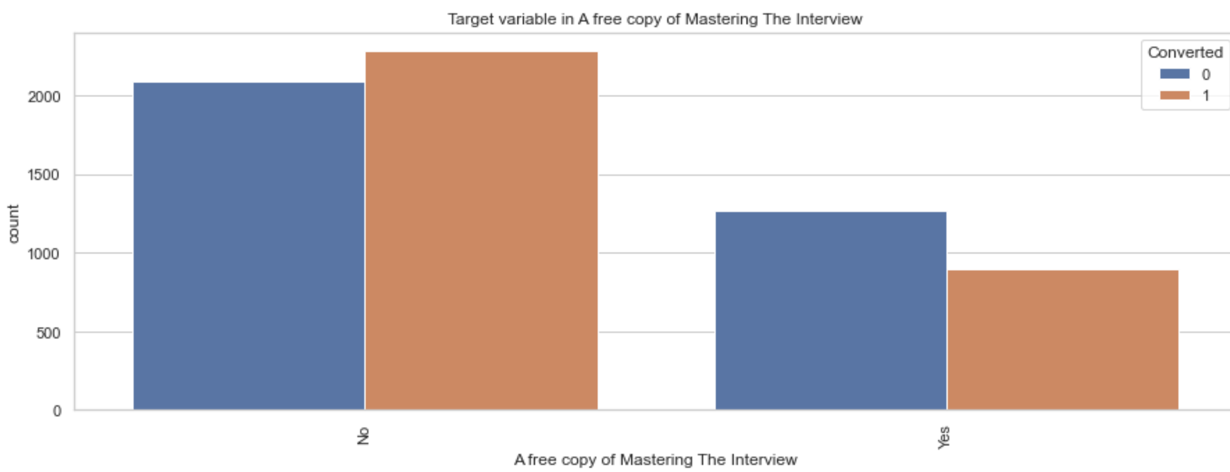


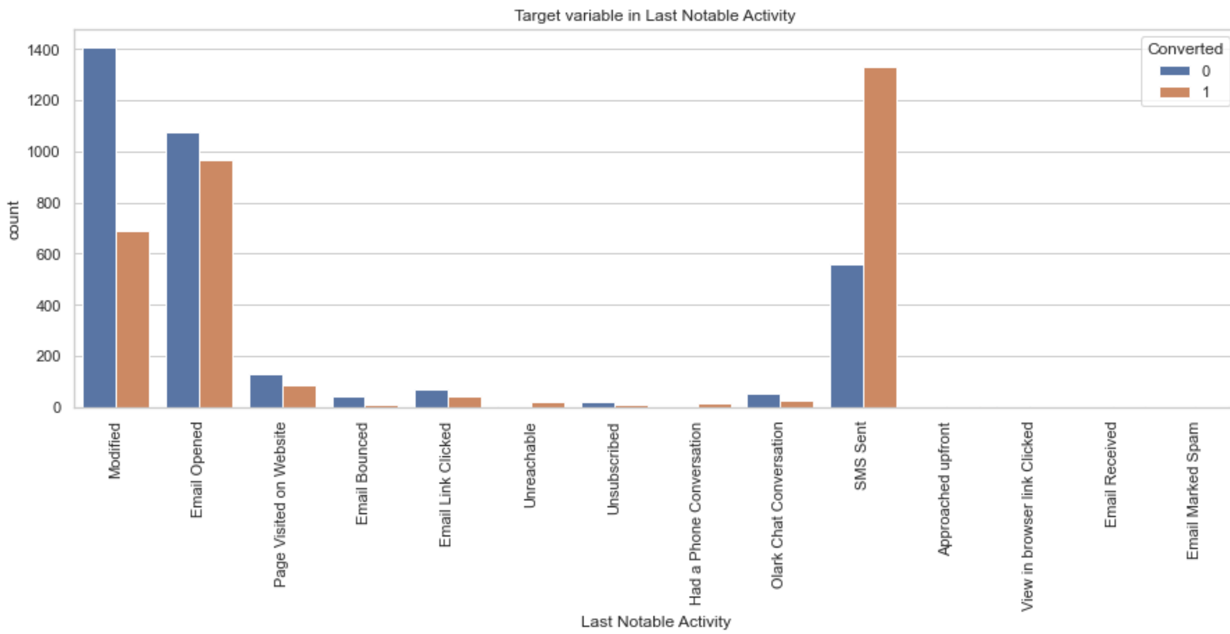
Target variable in Last Activity
count
1400
1200
1000
800
600
400
200
0
Converted
0
1
Page Visited on Website
Email Opened
Unreachable
Converted to Lead
Email Bounced
Email Link Clicked
Form Submitted on Website
Unsubscribed
Had a Phone Conversation
View in browser link Clicked
Olark Chat Conversation
Approached upfront
SMS Sent
Visited Booth in Tradeshow
Email Received
Email Marked Spam
Last Activity



Target variable in Specialization
count
1000
800
600
400
200
0
Converted
0
1
Select
Business Administration
Media and Advertising
Supply Chain Management
Travel and Tourism
Human Resource Management
Marketing Management
Banking, Investment And Insurance
Finance Management
International Business
E-COMMERCE
Operations Management
Retail Management
IT Projects Management
Services Excellence
Hospitality Management
Rural and Agribusiness
Healthcare Management
E-Business
Specialization


Target variable in What is your current occupation
Converted
0
1
count
3000
2500
2000
1500
1000
500
0
Unemployed
Student
Working Professional
Businessman
Other
Housewife
What is your current occupation


Target variable in A free copy of Mastering The Interview
Converted
0
1
count
2000
1500
1000
500
0
No
Yes
A free copy of Mastering The Interview




```
# Checking the number of null values again
leads.isnull().sum().sort_values(ascending=False)
```

```
TotalVisits                                 130
Page Views Per Visit                        130
Last Activity                               103
Lead Source                                  36
Specialization                               18
Prospect ID                                   0
Lead Number                                   0
Lead Origin                                   0
Do Not Email                                  0
Converted                                     0
Total Time Spent on Website                   0
What is your current occupation               0
A free copy of Mastering The Interview        0
Last Notable Activity                         0
dtype: int64
```

Since now the number of null values present in the columns are quite small we can simply drop the rows in which these null values are present.

```
# Dropping the null values rows in the column 'TotalVisits'

leads = leads[~pd.isnull(leads['TotalVisits'])]

# Checking the number of null values again
leads.isnull().sum().sort_values(ascending=False)
```

```
Lead Source                                 29
Specialization                              18
Prospect ID                                  0
Lead Number                                  0
Lead Origin                                  0
Do Not Email                                 0
Converted                                    0
TotalVisits                                  0
Total Time Spent on Website                  0
Page Views Per Visit                         0
Last Activity                                0
What is your current occupation              0
A free copy of Mastering The Interview       0
Last Notable Activity                        0
dtype: int64
```

```
# Dropping the null values rows in the column 'Lead Source'

leads = leads[~pd.isnull(leads['Lead Source'])]

# Checking the number of null values again
leads.isnull().sum().sort_values(ascending=False)
```

```
Specialization                              18
Prospect ID                                  0
Lead Number                                  0
Lead Origin                                  0
Lead Source                                  0
Do Not Email                                 0
Converted                                    0
TotalVisits                                  0
Total Time Spent on Website                  0
Page Views Per Visit                         0
Last Activity                                0
What is your current occupation              0
A free copy of Mastering The Interview       0
Last Notable Activity                        0
dtype: int64
```

```
# Drop the null values rows in the column 'Specialization'

leads = leads[~pd.isnull(leads['Specialization'])]

# Checking the number of null values again
leads.isnull().sum().sort_values(ascending=False)
```

```
Prospect ID                                 0
Lead Number                                 0
Lead Origin                                 0
Lead Source                                 0
Do Not Email                                0
```

```
Converted                                       0
TotalVisits                                     0
Total Time Spent on Website                     0
Page Views Per Visit                            0
Last Activity                                   0
Specialization                                  0
What is your current occupation                 0
A free copy of Mastering The Interview          0
Last Notable Activity                           0
dtype: int64
```

Now your data doesn't have any null values. Let's now check the percentage of rows that we have retained.

```
print(len(leads.index))
print(len(leads.index)/9240)

6373
0.6897186147186147
```

We still have around 69% of the rows which seems good enough.

```
# Let's look at the dataset again

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  Converted  TotalVisits  \
0      Olark Chat           No          0          0.0
1  Organic Search           No          0          5.0
2  Direct Traffic           No          1          2.0
3  Direct Traffic           No          0          1.0
4          Google           No          1          2.0

   Total Time Spent on Website  Page Views Per Visit            Last
Activity  \
0                            0                   0.0  Page Visited on
```

```
Website
1                           674                      2.5             Email
Opened
2                          1532                      2.0             Email
Opened
3                           305                      1.0
Unreachable
4                          1428                      1.0        Converted
to Lead

            Specialization What is your current occupation  \
0                   Select                      Unemployed
1                   Select                      Unemployed
2  Business Administration                         Student
3    Media and Advertising                      Unemployed
4                   Select                      Unemployed

  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
```

Now, clearly the variables `Prospect ID` and `Lead Number` won't be of any use in the analysis, so it's best that we drop these two variables.

```
# Dropping the "Prospect ID" and "Lead Number"
leads.drop(['Prospect ID', 'Lead Number'], 1, inplace = True)

leads.head()

               Lead Origin     Lead Source Do Not Email  Converted  \
0                      API      Olark Chat           No          0
1                      API  Organic Search           No          0
2  Landing Page Submission  Direct Traffic           No          1
3  Landing Page Submission  Direct Traffic           No          0
4  Landing Page Submission          Google           No          1

   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

             Last Activity           Specialization  \
0  Page Visited on Website                   Select
1             Email Opened                   Select
2             Email Opened  Business Administration
```

```
3              Unreachable    Media and Advertising
4        Converted to Lead                   Select

  What is your current occupation A free copy of Mastering The
Interview  \
0                      Unemployed
No
1                      Unemployed
No
2                         Student
Yes
3                      Unemployed
No
4                      Unemployed
No

  Last Notable Activity
0              Modified
1          Email Opened
2          Email Opened
3              Modified
4              Modified
```

Dummy variable creation

The next step is to dealing with the categorical variables present in the dataset. So first take a look at which variables are actually categorical variables.

```
# Checking the columns which are of type 'object'

temp = leads.loc[:, leads.dtypes == 'object']
temp.columns

Index(['Lead Origin', 'Lead Source', 'Do Not Email', 'Last Activity',
       'Specialization', 'What is your current occupation',
       'A free copy of Mastering The Interview', 'Last Notable
Activity'],
      dtype='object')

# Demo Cell
df = pd.DataFrame({'P': ['p', 'q', 'p']})
df

   P
0  p
1  q
2  p

pd.get_dummies(df)
```

```
   P_p  P_q
0    1    0
1    0    1
2    1    0
```

```
pd.get_dummies(df, prefix=['col1'])
```

```
   col1_p  col1_q
0       1       0
1       0       1
2       1       0
```

```
# Creating dummy variables using the 'get_dummies' command
dummy = pd.get_dummies(leads[['Lead Origin', '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
leads = pd.concat([leads, dummy], axis=1)

# Creating dummy variable separately for the variable 'Specialization'
since it has the level 'Select'
# which is useless so we
# drop that level by specifying it explicitly

dummy_spl = pd.get_dummies(leads['Specialization'], prefix =
'Specialization')
dummy_spl = dummy_spl.drop(['Specialization_Select'], 1)
leads = pd.concat([leads, dummy_spl], axis = 1)

# Dropping the variables for which the dummy variables have been
created

leads = leads.drop(['Lead Origin', 'Lead Source', 'Do Not Email',
'Last Activity',
                   'Specialization', 'What is your current
occupation',
                   'A free copy of Mastering The Interview', 'Last
Notable Activity'], 1)

# Let's take a look at the dataset again

leads.head()
```

```
   Converted  TotalVisits  Total Time Spent on Website  Page Views Per
Visit  \
0          0          0.0                            0
0.0
```

```
1          0          5.0                           674
2.5
2          1          2.0                          1532
2.0
3          0          1.0                           305
1.0
4          1          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  Lead Source_Direct Traffic  Lead
Source_Facebook  \
0                        0                           0
0
1                        0                           0
0
2                        0                           1
0
3                        0                           1
0
4                        0                           0
0

   Lead Source_Google  Lead Source_Live Chat  Lead Source_Olark
Chat  \
0                   0                      0                        1

1                   0                      0                        0

2                   0                      0                        0

3                   0                      0                        0

4                   1                      0                        0


   Lead Source_Organic Search  Lead Source_Pay per Click Ads  \
0                           0                              0
1                           1                              0
2                           0                              0
3                           0                              0
4                           0                              0

   Lead Source_Press_Release  Lead Source_Reference  \
```

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

   Lead Source_Referral Sites  Lead Source_Social Media  Lead
Source_WeLearn  \
0                           0                         0
0
1                           0                         0
0
2                           0                         0
0
3                           0                         0
0
4                           0                         0
0

   Lead Source_Welingak Website  Lead Source_bing  Lead Source_testone
\
0                             0                 0                    0

1                             0                 0                    0

2                             0                 0                    0

3                             0                 0                    0

4                             0                 0                    0


   Do Not Email_Yes  Last Activity_Converted to Lead  \
0                 0                                0
1                 0                                0
2                 0                                0
3                 0                                0
4                 0                                1

   Last Activity_Email Bounced  Last Activity_Email Link Clicked  \
0                            0                                 0
1                            0                                 0
2                            0                                 0
3                            0                                 0
4                            0                                 0

   Last Activity_Email Marked Spam  Last Activity_Email Opened  \
0                                0                           0
1                                0                           1
2                                0                           1
```

```
3                                0                           0
4                                0                           0

   Last Activity_Email Received  Last Activity_Form Submitted on
Website  \
0                             0
0
1                             0
0
2                             0
0
3                             0
0
4                             0
0

   Last Activity_Had a Phone Conversation  \
0                                       0
1                                       0
2                                       0
3                                       0
4                                       0

   Last Activity_Olark Chat Conversation  \
0                                      0
1                                      0
2                                      0
3                                      0
4                                      0

   Last Activity_Page Visited on Website  Last Activity_SMS Sent  \
0                                      1                       0
1                                      0                       0
2                                      0                       0
3                                      0                       0
4                                      0                       0

   Last Activity_Unreachable  Last Activity_Unsubscribed  \
0                          0                           0
1                          0                           0
2                          0                           0
3                          1                           0
4                          0                           0

   Last Activity_View in browser link Clicked  \
0                                           0
1                                           0
2                                           0
3                                           0
4                                           0
```

```
   Last Activity_Visited Booth in Tradeshow  \
0                                         0
1                                         0
2                                         0
3                                         0
4                                         0

   What is your current occupation_Housewife  \
0                                          0
1                                          0
2                                          0
3                                          0
4                                          0

   What is your current occupation_Other  \
0                                      0
1                                      0
2                                      0
3                                      0
4                                      0

   What is your current occupation_Student  \
0                                        0
1                                        0
2                                        1
3                                        0
4                                        0

   What is your current occupation_Unemployed  \
0                                           1
1                                           1
2                                           0
3                                           1
4                                           1

   What is your current occupation_Working Professional  \
0                                                     0
1                                                     0
2                                                     0
3                                                     0
4                                                     0

   A free copy of Mastering The Interview_Yes  \
0                                           0
1                                           0
2                                           1
3                                           0
4                                           0
```

```
   Last Notable Activity_Email Bounced  \
0                                    0
1                                    0
2                                    0
3                                    0
4                                    0

   Last Notable Activity_Email Link Clicked  \
0                                         0
1                                         0
2                                         0
3                                         0
4                                         0

   Last Notable Activity_Email Marked Spam  \
0                                        0
1                                        0
2                                        0
3                                        0
4                                        0

   Last Notable Activity_Email Opened  Last Notable Activity_Email
Received  \
0                                   0
0
1                                   1
0
2                                   1
0
3                                   0
0
4                                   0
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_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

   Specialization_Banking, Investment And Insurance  \
0                                                 0
1                                                 0
2                                                 0
3                                                 0
4                                                 0

   Specialization_Business Administration  Specialization_E-
Business  \
0                                       0                          0
```

```
1                                      0                          0

2                                      1                          0

3                                      0                          0

4                                      0                          0


   Specialization_E-COMMERCE  Specialization_Finance Management  \
0                          0                                  0
1                          0                                  0
2                          0                                  0
3                          0                                  0
4                          0                                  0

   Specialization_Healthcare Management  \
0                                     0
1                                     0
2                                     0
3                                     0
4                                     0

   Specialization_Hospitality Management  \
0                                      0
1                                      0
2                                      0
3                                      0
4                                      0

   Specialization_Human Resource Management  \
0                                         0
1                                         0
2                                         0
3                                         0
4                                         0

   Specialization_IT Projects Management  \
0                                      0
1                                      0
2                                      0
3                                      0
4                                      0

   Specialization_International Business  Specialization_Marketing
Management  \
0                                      0
0
1                                      0
0
```

```
2                                       0
0
3                                       0
0
4                                       0
0

   Specialization_Media and Advertising  Specialization_Operations
Management  \
0                                      0
0
1                                      0
0
2                                      0
0
3                                      1
0
4                                      0
0

   Specialization_Retail Management  Specialization_Rural and
Agribusiness  \
0                                  0
0
1                                  0
0
2                                  0
0
3                                  0
0
4                                  0
0

   Specialization_Services Excellence  Specialization_Supply Chain
Management  \
0                                    0
0
1                                    0
0
2                                    0
0
3                                    0
0
4                                    0
0

   Specialization_Travel and Tourism
0                                   0
1                                   0
2                                   0
```

```
3                                    0
4                                    0
```

Test-Train Split

The next step is to spliting the dataset into training an testing sets.

```
# Importing the `train_test_split` library

# Put all the feature variables in X

X = leads.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  Lead Source_Direct Traffic  Lead
Source_Facebook  \
0                        0                           0
0
1                        0                           0
0
2                        0                           1
0
3                        0                           1
0
4                        0                           0
0

   Lead Source_Google  Lead Source_Live Chat  Lead Source_Olark
Chat  \
0                   0                      0                        1

1                   0                      0                        0

2                   0                      0                        0

3                   0                      0                        0
```

```
4                    1                       0                        0


   Lead Source_Organic Search  Lead Source_Pay per Click Ads  \
0                           0                              0
1                           1                              0
2                           0                              0
3                           0                              0
4                           0                              0

   Lead Source_Press_Release  Lead Source_Reference  \
0                          0                      0
1                          0                      0
2                          0                      0
3                          0                      0
4                          0                      0

   Lead Source_Referral Sites  Lead Source_Social Media  Lead
Source_WeLearn  \
0                           0                         0
0
1                           0                         0
0
2                           0                         0
0
3                           0                         0
0
4                           0                         0
0

   Lead Source_Welingak Website  Lead Source_bing  Lead Source_testone
\
0                             0                 0                    0

1                             0                 0                    0

2                             0                 0                    0

3                             0                 0                    0

4                             0                 0                    0


   Do Not Email_Yes  Last Activity_Converted to Lead  \
0                 0                                0
1                 0                                0
2                 0                                0
3                 0                                0
4                 0                                1
```

```
   Last Activity_Email Bounced  Last Activity_Email Link Clicked  \
0                            0                                  0
1                            0                                  0
2                            0                                  0
3                            0                                  0
4                            0                                  0

   Last Activity_Email Marked Spam  Last Activity_Email Opened  \
0                                0                           0
1                                0                           1
2                                0                           1
3                                0                           0
4                                0                           0

   Last Activity_Email Received  Last Activity_Form Submitted on
Website  \
0                             0
0
1                             0
0
2                             0
0
3                             0
0
4                             0
0

   Last Activity_Had a Phone Conversation  \
0                                        0
1                                        0
2                                        0
3                                        0
4                                        0

   Last Activity_Olark Chat Conversation  \
0                                       0
1                                       0
2                                       0
3                                       0
4                                       0

   Last Activity_Page Visited on Website  Last Activity_SMS Sent  \
0                                       1                        0
1                                       0                        0
2                                       0                        0
3                                       0                        0
4                                       0                        0

   Last Activity_Unreachable  Last Activity_Unsubscribed  \
```

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

   Last Activity_View in browser link Clicked  \
0                                           0
1                                           0
2                                           0
3                                           0
4                                           0

   Last Activity_Visited Booth in Tradeshow  \
0                                         0
1                                         0
2                                         0
3                                         0
4                                         0

   What is your current occupation_Housewife  \
0                                          0
1                                          0
2                                          0
3                                          0
4                                          0

   What is your current occupation_Other  \
0                                      0
1                                      0
2                                      0
3                                      0
4                                      0

   What is your current occupation_Student  \
0                                        0
1                                        0
2                                        1
3                                        0
4                                        0

   What is your current occupation_Unemployed  \
0                                           1
1                                           1
2                                           0
3                                           1
4                                           1

   What is your current occupation_Working Professional  \
0                                                     0
```

```
1                                                      0
2                                                      0
3                                                      0
4                                                      0

   A free copy of Mastering The Interview_Yes  \
0                                            0
1                                            0
2                                            1
3                                            0
4                                            0

   Last Notable Activity_Email Bounced  \
0                                     0
1                                     0
2                                     0
3                                     0
4                                     0

   Last Notable Activity_Email Link Clicked  \
0                                          0
1                                          0
2                                          0
3                                          0
4                                          0

   Last Notable Activity_Email Marked Spam  \
0                                         0
1                                         0
2                                         0
3                                         0
4                                         0

   Last Notable Activity_Email Opened  Last Notable Activity_Email
Received  \
0                                    0
0
1                                    1
0
2                                    1
0
3                                    0
0
4                                    0
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_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
```

```
   Specialization_Banking, Investment And Insurance  \
0                                                  0
1                                                  0
2                                                  0
3                                                  0
4                                                  0

   Specialization_Business Administration  Specialization_E-
Business  \
0                                       0                          0

1                                       0                          0

2                                       1                          0

3                                       0                          0

4                                       0                          0


   Specialization_E-COMMERCE  Specialization_Finance Management  \
0                          0                                  0
1                          0                                  0
2                          0                                  0
3                          0                                  0
4                          0                                  0

   Specialization_Healthcare Management  \
0                                     0
1                                     0
2                                     0
3                                     0
4                                     0

   Specialization_Hospitality Management  \
0                                      0
1                                      0
2                                      0
3                                      0
4                                      0

   Specialization_Human Resource Management  \
0                                         0
1                                         0
2                                         0
3                                         0
4                                         0

   Specialization_IT Projects Management  \
```

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

   Specialization_International Business  Specialization_Marketing
Management  \
0                                     0
0
1                                     0
0
2                                     0
0
3                                     0
0
4                                     0
0

   Specialization_Media and Advertising  Specialization_Operations
Management  \
0                                    0
0
1                                    0
0
2                                    0
0
3                                    1
0
4                                    0
0

   Specialization_Retail Management  Specialization_Rural and
Agribusiness  \
0                                0
0
1                                0
0
2                                0
0
3                                0
0
4                                0
0

   Specialization_Services Excellence  Specialization_Supply Chain
Management  \
0                                  0
0
1                                  0
```

```
0
2                                    0
0
3                                    0
0
4                                    0
0

   Specialization_Travel and Tourism
0                                   0
1                                   0
2                                   0
3                                   0
4                                   0
```

```
y = leads['Converted']

y.head()
```

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

```
# Spliting the dataset into 70% train and 30% test

X_train, X_test, y_train, y_test = train_test_split(X, y,
train_size=0.7, test_size=0.3, random_state=100)

#lets check the shape
print("X_train Size", X_train.shape)
print("y_train Size", y_train.shape)
```

```
X_train Size (4461, 74)
y_train Size (4461,)
```

Scaling

Now there are a few numeric variables present in the dataset which have different scales. So let's go ahead and scale these variables.

```
# Importing the 'MinMax scaler' Library

# Scaling the three numeric features present in the dataset

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  \
8003     0.015936                     0.029489                0.125

218      0.015936                     0.082306                0.250

4171     0.023904                     0.034331                0.375

4037     0.000000                     0.000000                0.000

3660     0.000000                     0.000000                0.000


      Lead Origin_Landing Page Submission  Lead Origin_Lead Add
Form  \
8003                                    1                          0

218                                     1                          0

4171                                    1                          0

4037                                    0                          0

3660                                    0                          1


      Lead Origin_Lead Import  Lead Source_Direct Traffic  \
8003                        0                           1
218                         0                           1
4171                        0                           1
4037                        0                           0
3660                        0                           0

      Lead Source_Facebook  Lead Source_Google  Lead Source_Live Chat
\
8003                     0                   0                      0

218                      0                   0                      0

4171                     0                   0                      0

4037                     0                   0                      0

3660                     0                   0                      0


      Lead Source_Olark Chat  Lead Source_Organic Search  \
```

```
8003                      0                            0
218                       0                            0
4171                      0                            0
4037                      1                            0
3660                      0                            0

      Lead Source_Pay per Click Ads  Lead Source_Press_Release  \
8003                             0                          0
218                              0                          0
4171                             0                          0
4037                             0                          0
3660                             0                          0

      Lead Source_Reference  Lead Source_Referral Sites  \
8003                      0                           0
218                       0                           0
4171                      0                           0
4037                      0                           0
3660                      1                           0

      Lead Source_Social Media  Lead Source_WeLearn  \
8003                         0                    0
218                          0                    0
4171                         0                    0
4037                         0                    0
3660                         0                    0

      Lead Source_Welingak Website  Lead Source_bing  Lead
Source_testone  \
8003                             0                 0
0
218                              0                 0
0
4171                             0                 0
0
4037                             0                 0
0
3660                             0                 0
0

      Do Not Email_Yes  Last Activity_Converted to Lead  \
8003                 0                                0
218                  0                                0
4171                 0                                0
4037                 0                                0
3660                 0                                0

      Last Activity_Email Bounced  Last Activity_Email Link Clicked  \
8003                            0                                 0
218                             0                                 1
```

```
4171                             0                                   0
4037                             0                                   0
3660                             0                                   0

      Last Activity_Email Marked Spam  Last Activity_Email Opened  \
8003                                 0                            0
218                                  0                            0
4171                                 0                            0
4037                                 0                            1
3660                                 0                            0

      Last Activity_Email Received  Last Activity_Form Submitted on
Website  \
8003                              0
0
218                               0
0
4171                              0
0
4037                              0
0
3660                              0
0

      Last Activity_Had a Phone Conversation  \
8003                                        0
218                                         0
4171                                        0
4037                                        0
3660                                        0

      Last Activity_Olark Chat Conversation  \
8003                                       0
218                                        0
4171                                       0
4037                                       0
3660                                       0

      Last Activity_Page Visited on Website  Last Activity_SMS Sent  \
8003                                       0                        1
218                                        0                        0
4171                                       1                        0
4037                                       0                        0
3660                                       0                        1

      Last Activity_Unreachable  Last Activity_Unsubscribed  \
8003                           0                            0
218                            0                            0
4171                           0                            0
4037                           0                            0
```

```
3660                          0                           0

      Last Activity_View in browser link Clicked  \
8003                                           0
218                                            0
4171                                           0
4037                                           0
3660                                           0

      Last Activity_Visited Booth in Tradeshow  \
8003                                         0
218                                          0
4171                                         0
4037                                         0
3660                                         0

      What is your current occupation_Housewife  \
8003                                          0
218                                           0
4171                                          0
4037                                          0
3660                                          0

      What is your current occupation_Other  \
8003                                      0
218                                       0
4171                                      0
4037                                      0
3660                                      0

      What is your current occupation_Student  \
8003                                        0
218                                         0
4171                                        0
4037                                        0
3660                                        0

      What is your current occupation_Unemployed  \
8003                                           1
218                                            1
4171                                           1
4037                                           1
3660                                           1

      What is your current occupation_Working Professional  \
8003                                                     0
218                                                      0
4171                                                     0
4037                                                     0
3660                                                     0
```

```
      A free copy of Mastering The Interview_Yes  \
8003                                            1
218                                             1
4171                                            1
4037                                            0
3660                                            0

      Last Notable Activity_Email Bounced  \
8003                                     0
218                                      0
4171                                     0
4037                                     0
3660                                     0

      Last Notable Activity_Email Link Clicked  \
8003                                          0
218                                           0
4171                                          0
4037                                          0
3660                                          0

      Last Notable Activity_Email Marked Spam  \
8003                                         0
218                                          0
4171                                         0
4037                                         0
3660                                         0

      Last Notable Activity_Email Opened  \
8003                                    0
218                                     0
4171                                    0
4037                                    1
3660                                    0

      Last Notable Activity_Email Received  \
8003                                      0
218                                       0
4171                                      0
4037                                      0
3660                                      0

      Last Notable Activity_Had a Phone Conversation  \
8003                                                0
218                                                 0
4171                                                0
4037                                                0
3660                                                0
```

```
      Last Notable Activity_Modified  \
8003                               0
218                                1
4171                               1
4037                               0
3660                               0

      Last Notable Activity_Olark Chat Conversation  \
8003                                              0
218                                               0
4171                                              0
4037                                              0
3660                                              0

      Last Notable Activity_Page Visited on Website  \
8003                                              0
218                                               0
4171                                              0
4037                                              0
3660                                              0

      Last Notable Activity_SMS Sent  Last Notable
Activity_Unreachable  \
8003                               1
0
218                                0
0
4171                               0
0
4037                               0
0
3660                               1
0

      Last Notable Activity_Unsubscribed  \
8003                                   0
218                                    0
4171                                   0
4037                                   0
3660                                   0

      Last Notable Activity_View in browser link Clicked  \
8003                                                   0
218                                                    0
4171                                                   0
4037                                                   0
3660                                                   0

      Specialization_Banking, Investment And Insurance  \
8003                                                 0
```

```
218                                                     0
4171                                                    0
4037                                                    0
3660                                                    0

      Specialization_Business Administration  Specialization_E-
Business  \
8003                                       0
0
218                                        0
0
4171                                       0
0
4037                                       0
0
3660                                       0
0

      Specialization_E-COMMERCE  Specialization_Finance Management  \
8003                          0                                  0
218                           0                                  0
4171                          0                                  0
4037                          0                                  0
3660                          0                                  1

      Specialization_Healthcare Management  \
8003                                     0
218                                      0
4171                                     0
4037                                     0
3660                                     0

      Specialization_Hospitality Management  \
8003                                      0
218                                       0
4171                                      0
4037                                      0
3660                                      0

      Specialization_Human Resource Management  \
8003                                         0
218                                          1
4171                                         0
4037                                         0
3660                                         0

      Specialization_IT Projects Management  \
8003                                      1
218                                       0
4171                                      0
```

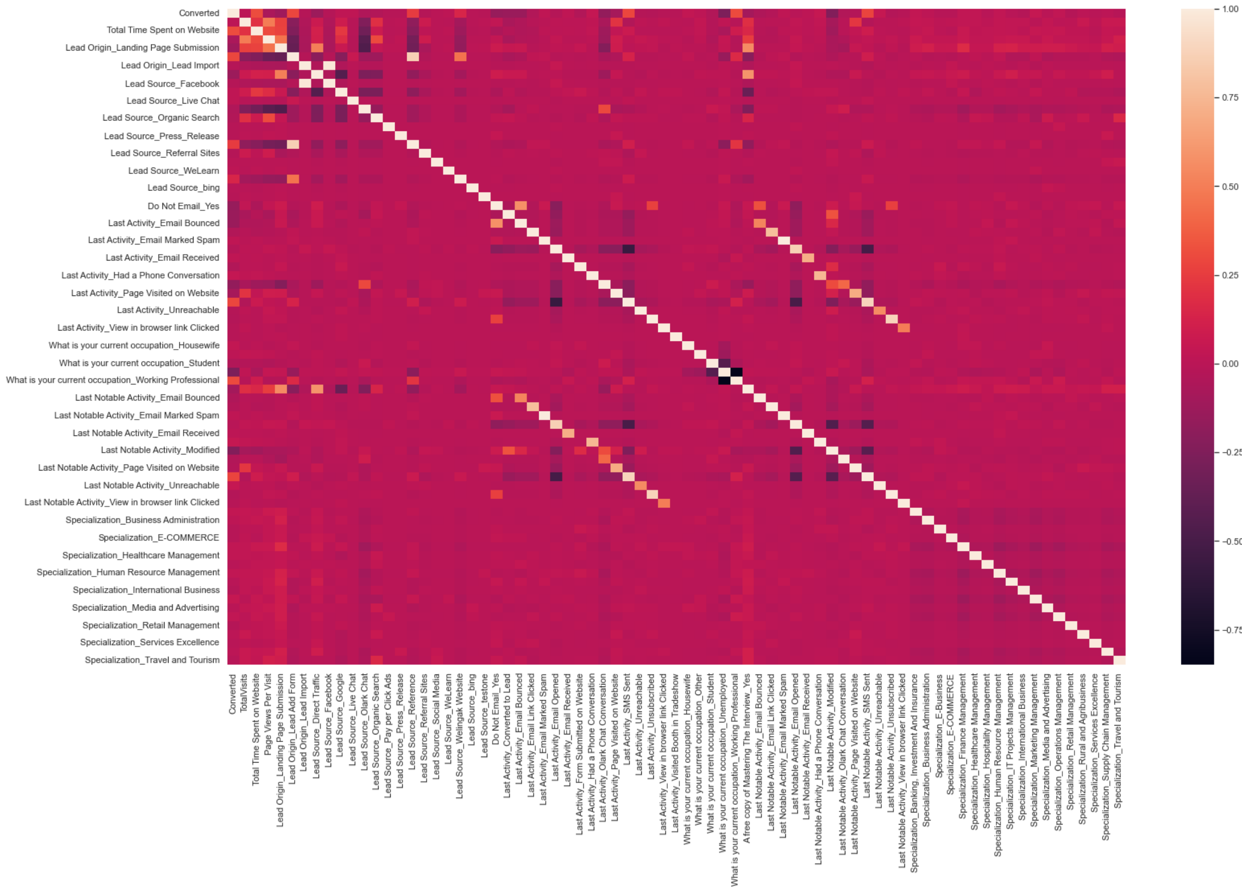
```
4037                                     0
3660                                     0

      Specialization_International Business  \
8003                                     0
218                                      0
4171                                     0
4037                                     0
3660                                     0

      Specialization_Marketing Management  \
8003                                   0
218                                    0
4171                                   0
4037                                   0
3660                                   0

      Specialization_Media and Advertising  \
8003                                    0
218                                     0
4171                                    0
4037                                    0
3660                                    0

      Specialization_Operations Management  Specialization_Retail
Management  \
8003                                    0
0
218                                     0
0
4171                                    0
0
4037                                    0
0
3660                                    0
0

      Specialization_Rural and Agribusiness  \
8003                                     0
218                                      0
4171                                     0
4037                                     0
3660                                     0

      Specialization_Services Excellence  \
8003                                  0
218                                   0
4171                                  0
4037                                  0
3660                                  0
```

```
      Specialization_Supply Chain Management  \
8003                                        0
218                                         0
4171                                        0
4037                                        0
3660                                        0

      Specialization_Travel and Tourism
8003                                   0
218                                    0
4171                                   1
4037                                   0
3660                                   0
```

Looking at the correlations

Let's now look at the correlations. Since the number of variables are pretty high, it's better that we look at the table instead of plotting a heatmap

```
# Looking at the correlation table
plt.figure(figsize = (25,15))
sns.heatmap(leads.corr())
plt.show()
```



Step 2: Model Building

Let's now move to model building. As you can see that there are a lot of variables present in the dataset which we cannot deal with. So the best way to approach this is to select a small set of features from this pool of variables using RFE.

```
# Importing the 'LogisticRegression' and creating a LogisticRegression object
logreg = LogisticRegression()

# Importing the 'RFE' and select 15 variables

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

# Let's take a look at which features 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, 23),
```

```
('Lead Origin_Landing Page Submission', False, 8),
('Lead Origin_Lead Add Form', True, 1),
('Lead Origin_Lead Import', False, 52),
('Lead Source_Direct Traffic', False, 24),
('Lead Source_Facebook', False, 51),
('Lead Source_Google', False, 36),
('Lead Source_Live Chat', False, 44),
('Lead Source_Olark Chat', True, 1),
('Lead Source_Organic Search', False, 35),
('Lead Source_Pay per Click Ads', False, 43),
('Lead Source_Press_Release', False, 53),
('Lead Source_Reference', True, 1),
('Lead Source_Referral Sites', False, 37),
('Lead Source_Social Media', False, 58),
('Lead Source_WeLearn', False, 42),
('Lead Source_Welingak Website', True, 1),
('Lead Source_bing', False, 33),
('Lead Source_testone', False, 38),
('Do Not Email_Yes', True, 1),
('Last Activity_Converted to Lead', False, 25),
('Last Activity_Email Bounced', False, 4),
('Last Activity_Email Link Clicked', False, 49),
('Last Activity_Email Marked Spam', False, 57),
('Last Activity_Email Opened', False, 41),
('Last Activity_Email Received', False, 54),
('Last Activity_Form Submitted on Website', False, 28),
('Last Activity_Had a Phone Conversation', True, 1),
('Last Activity_Olark Chat Conversation', False, 5),
('Last Activity_Page Visited on Website', False, 26),
('Last Activity_SMS Sent', True, 1),
('Last Activity_Unreachable', False, 47),
('Last Activity_Unsubscribed', False, 40),
('Last Activity_View in browser link Clicked', False, 34),
('Last Activity_Visited Booth in Tradeshow', False, 48),
('What is your current occupation_Housewife', True, 1),
('What is your current occupation_Other', False, 46),
('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),
('A free copy of Mastering The Interview_Yes', False, 50),
('Last Notable Activity_Email Bounced', False, 3),
('Last Notable Activity_Email Link Clicked', False, 20),
('Last Notable Activity_Email Marked Spam', False, 59),
('Last Notable Activity_Email Opened', False, 27),
('Last Notable Activity_Email Received', False, 60),
('Last Notable Activity_Had a Phone Conversation', True, 1),
('Last Notable Activity_Modified', False, 2),
('Last Notable Activity_Olark Chat Conversation', False, 32),
('Last Notable Activity_Page Visited on Website', False, 31),
```

```
 ('Last Notable Activity_SMS Sent', False, 45),
 ('Last Notable Activity_Unreachable', True, 1),
 ('Last Notable Activity_Unsubscribed', False, 39),
 ('Last Notable Activity_View in browser link Clicked', False, 29),
 ('Specialization_Banking, Investment And Insurance', False, 6),
 ('Specialization_Business Administration', False, 15),
 ('Specialization_E-Business', False, 11),
 ('Specialization_E-COMMERCE', False, 9),
 ('Specialization_Finance Management', False, 14),
 ('Specialization_Healthcare Management', False, 10),
 ('Specialization_Hospitality Management', False, 55),
 ('Specialization_Human Resource Management', False, 16),
 ('Specialization_IT Projects Management', False, 18),
 ('Specialization_International Business', False, 22),
 ('Specialization_Marketing Management', False, 12),
 ('Specialization_Media and Advertising', False, 21),
 ('Specialization_Operations Management', False, 19),
 ('Specialization_Retail Management', False, 30),
 ('Specialization_Rural and Agribusiness', False, 7),
 ('Specialization_Services Excellence', False, 56),
 ('Specialization_Supply Chain Management', False, 13),
 ('Specialization_Travel and Tourism', False, 17)]

# Putting all the columns selected by RFE in the variable 'col'

col = X_train.columns[rfe.support_]
```

Now we have all the variables selected by RFE and since we care about the statistics part, i.e. the p-values and the VIFs, let's use these variables to create a logistic regression model using statsmodels.

```
# Select only the columns selected by RFE

X_train = X_train[col]

# Importing 'statsmodels'
```

Model 1

```
# Fit a logistic Regression model on X_train after adding a constant
and output the summary

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:
4461
Model:                            GLM   Df Residuals:
4445
Model Family:                Binomial   Df Model:
15
Link Function:                  logit   Scale:
1.0000
Method:                          IRLS   Log-Likelihood:
-2072.8
Date:            Sun, 01 Jan 2023   Deviance:
4145.5
Time:                        14:18:37   Pearson chi2:
4.84e+03
No. Iterations:                    22

Covariance Type:            nonrobust

================================================================================
==================================================
                                                                    coef    std
err          z      P>|z|      [0.025      0.975]
--------------------------------------------------------------------------------
--------------------------------------------------
const                                                            -1.0061
0.600     -1.677      0.094      -2.182       0.170
TotalVisits                                                      11.3439
2.682      4.230      0.000       6.088      16.600
Total Time Spent on Website                                       4.4312
0.185     23.924      0.000       4.068       4.794
Lead Origin_Lead Add Form                                         2.9483
1.191      2.475      0.013       0.614       5.283
Lead Source_Olark Chat                                            1.4584
0.122     11.962      0.000       1.219       1.697
Lead Source_Reference                                             1.2994
1.214      1.070      0.285      -1.080       3.679
Lead Source_Welingak Website                                      3.4159
1.558      2.192      0.028       0.362       6.470
Do Not Email_Yes                                                 -1.5053
0.193     -7.781      0.000      -1.884      -1.126
Last Activity_Had a Phone Conversation                            1.0397
0.983      1.058      0.290      -0.887       2.966
Last Activity_SMS Sent                                            1.1827
0.082     14.362      0.000       1.021       1.344
What is your current occupation_Housewife                        22.6492
2.45e+04      0.001      0.999    -4.8e+04     4.8e+04
What is your current occupation_Student                          -1.1544
```

```
0.630     -1.831      0.067      -2.390       0.081
What is your current occupation_Unemployed                -1.3395
0.594     -2.254      0.024      -2.505      -0.175
What is your current occupation_Working Professional       1.2743
0.623      2.045      0.041       0.053       2.496
Last Notable Activity_Had a Phone Conversation            23.1932
2.08e+04      0.001      0.999   -4.08e+04    4.08e+04
Last Notable Activity_Unreachable                          2.7868
0.807      3.453      0.001       1.205       4.369
======================================================================
==================================================
"""
```

There are quite a few variable which have a p-value greater than 0.05. We will need to take care of them. But first, let's also look at the VIFs.

Checking `VIF`

```python
# Importing the 'variance_inflation_factor' library

# 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
2	Lead Origin_Lead Add Form	84.19
4	Lead Source_Reference	65.18
5	Lead Source_Welingak Website	20.03
11	What is your current occupation_Unemployed	3.65
7	Last Activity_Had a Phone Conversation	2.44
13	Last Notable Activity_Had a Phone Conversation	2.43
1	Total Time Spent on Website	2.38
0	TotalVisits	1.62
8	Last Activity_SMS Sent	1.59
12	What is your current occupation_Working Professional	1.56
3	Lead Source_Olark Chat	1.44
6	Do Not Email_Yes	1.09
10	What is your current occupation_Student	1.09
9	What is your current occupation_Housewife	1.01
14	Last Notable Activity_Unreachable	1.01

VIFs seem to be in a decent range except for three variables.

Let's first drop the variable `Lead Source_Reference` since it has a high p-value as well as a high VIF.

```
X_train.drop('Lead Source_Reference', axis = 1, inplace = True)
```

Model 2

```
# Refit the model with the new set of features

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

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

==============================================================================
========
Dep. Variable:              Converted   No. Observations:                 
4461
Model:                            GLM   Df Residuals:                     
4446
Model Family:                Binomial   Df Model:                         
14
Link Function:                  logit   Scale:                         
1.0000
Method:                          IRLS   Log-Likelihood:                
-2073.2
Date:                Sun, 01 Jan 2023   Deviance:                       
4146.5
Time:                        14:24:01   Pearson chi2:                 
4.82e+03
No. Iterations:                    22                                         

Covariance Type:            nonrobust                                         

==============================================================================
==================================================
                                                          coef    std 
err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
--------------------------------------------------
const                                                  -1.0057    
0.600     -1.677      0.094      -2.181       0.170
TotalVisits                                            11.3428    
2.682      4.229      0.000       6.086      16.599
Total Time Spent on Website                             4.4312    
0.185     23.924      0.000       4.068       4.794
Lead Origin_Lead Add Form                               4.2084    
```

```
0.259     16.277      0.000       3.702       4.715
Lead Source_Olark Chat                                          1.4583
0.122     11.960      0.000       1.219       1.697
Lead Source_Welingak Website                                    2.1557
1.037      2.079      0.038       0.124       4.188
Do Not Email_Yes                                               -1.5036
0.193     -7.779      0.000      -1.882      -1.125
Last Activity_Had a Phone Conversation                          1.0398
0.983      1.058      0.290      -0.887       2.966
Last Activity_SMS Sent                                          1.1827
0.082     14.362      0.000       1.021       1.344
What is your current occupation_Housewife                      22.6511
2.45e+04      0.001      0.999    -4.8e+04     4.8e+04
What is your current occupation_Student                        -1.1537
0.630     -1.830      0.067      -2.389       0.082
What is your current occupation_Unemployed                     -1.3401
0.594     -2.255      0.024      -2.505      -0.175
What is your current occupation_Working Professional      1.2748
0.623      2.046      0.041       0.053       2.496
Last Notable Activity_Had a Phone Conversation         23.1934
2.08e+04      0.001      0.999   -4.08e+04    4.08e+04
Last Notable Activity_Unreachable                               2.7872
0.807      3.454      0.001       1.205       4.369
======================================================================
==================================================
"""
```

Checking VIF

```python
# 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  3.65
6                 Last Activity_Had a Phone Conversation  2.44
12        Last Notable Activity_Had a Phone Conversation  2.43
1                            Total Time Spent on Website  2.38
2                              Lead Origin_Lead Add Form  1.71
0                                            TotalVisits  1.62
7                                 Last Activity_SMS Sent  1.59
11  What is your current occupation_Working Professional  1.56
3                                 Lead Source_Olark Chat  1.44
4                           Lead Source_Welingak Website  1.33
```

```
5                                          Do Not Email_Yes  1.09
9                       What is your current occupation_Student  1.09
8                     What is your current occupation_Housewife  1.01
13                            Last Notable Activity_Unreachable  1.01
```

The VIFs are now all less than 5. So let's drop the ones with the high p-values beginning with `Last Notable Activity_Had a Phone Conversation`.

```
X_train.drop('Last Notable Activity_Had a Phone Conversation', axis =
1, inplace = True)
```

Model 3

```
# Refit the model with the new set of features

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

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

======================================================================
========
Dep. Variable:              Converted   No. Observations:
4461
Model:                            GLM   Df Residuals:
4447
Model Family:                Binomial   Df Model:
13
Link Function:                  logit   Scale:
1.0000
Method:                          IRLS   Log-Likelihood:
-2076.1
Date:                Sun, 01 Jan 2023   Deviance:
4152.2
Time:                        14:30:54   Pearson chi2:
4.82e+03
No. Iterations:                    21

Covariance Type:            nonrobust

======================================================================
==================================================
                                                           coef    std
err          z      P>|z|      [0.025      0.975]
----------------------------------------------------------------------
--------------------------------------------------
```

```
const                                                              -1.0069
0.600     -1.679      0.093      -2.182       0.168
TotalVisits                                                        11.4551
2.686      4.265      0.000       6.191      16.720
Total Time Spent on Website                                         4.4237
0.185     23.900      0.000       4.061       4.787
Lead Origin_Lead Add Form                                           4.2082
0.259     16.276      0.000       3.701       4.715
Lead Source_Olark Chat                                              1.4581
0.122     11.958      0.000       1.219       1.697
Lead Source_Welingak Website                                        2.1557
1.037      2.079      0.038       0.124       4.188
Do Not Email_Yes                                                   -1.5037
0.193     -7.780      0.000      -1.882      -1.125
Last Activity_Had a Phone Conversation                              2.7502
0.802      3.430      0.001       1.179       4.322
Last Activity_SMS Sent                                              1.1826
0.082     14.364      0.000       1.021       1.344
What is your current occupation_Housewife                          21.6525
1.49e+04      0.001       0.999   -2.91e+04    2.91e+04
What is your current occupation_Student                            -1.1520
0.630     -1.828      0.068      -2.387       0.083
What is your current occupation_Unemployed                         -1.3385
0.594     -2.253      0.024      -2.503      -0.174
What is your current occupation_Working Professional                1.2743
0.623      2.045      0.041       0.053       2.495
Last Notable Activity_Unreachable                                   2.7862
0.807      3.453      0.001       1.205       4.368
======================================================================
==================================================
"""
```

Dropping the `What is your current occupation_Housewife` as having high P value

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

Model 4

```
# Refit the model with the new set of features

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

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

```
==============================================================================
========
Dep. Variable:              Converted   No. Observations:
4461
Model:                            GLM   Df Residuals:
4448
Model Family:                Binomial   Df Model:
12
Link Function:                  logit   Scale:
1.0000
Method:                          IRLS   Log-Likelihood:
-2078.3
Date:                Sun, 01 Jan 2023   Deviance:
4156.7
Time:                        14:33:01   Pearson chi2:
4.83e+03
No. Iterations:                     7

Covariance Type:            nonrobust

==============================================================================
==================================================
                                                                    coef    std
err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
--------------------------------------------------
const                                                            -0.4528
0.554     -0.818      0.413      -1.538       0.632
TotalVisits                                                      11.2586
2.672      4.214      0.000       6.023      16.495
Total Time Spent on Website                                       4.4217
0.185     23.898      0.000       4.059       4.784
Lead Origin_Lead Add Form                                         4.2057
0.258     16.274      0.000       3.699       4.712
Lead Source_Olark Chat                                            1.4530
0.122     11.930      0.000       1.214       1.692
Lead Source_Welingak Website                                      2.1541
1.037      2.078      0.038       0.122       4.186
Do Not Email_Yes                                                 -1.5063
0.193     -7.785      0.000      -1.886      -1.127
Last Activity_Had a Phone Conversation                            2.7515
0.802      3.432      0.001       1.180       4.323
Last Activity_SMS Sent                                            1.1823
0.082     14.362      0.000       1.021       1.344
What is your current occupation_Student                          -1.7017
0.588     -2.893      0.004      -2.855      -0.549
What is your current occupation_Unemployed                       -1.8879
0.550     -3.435      0.001      -2.965      -0.811
What is your current occupation_Working Professional              0.7246
```

```
0.581      1.248      0.212      -0.413      1.862
Last Notable Activity_Unreachable                        2.7834
0.807      3.448      0.001       1.201      4.365
======================================================================
==================================================
"""
```

Droppint hre `What is your current occupation_Working Professional` as having high P value

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

Model 4

```
# Refit the model with the new set of features

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

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

======================================================================
========
Dep. Variable:              Converted   No. Observations:
4461
Model:                            GLM   Df Residuals:
4449
Model Family:                Binomial   Df Model:
11
Link Function:                  logit   Scale:
1.0000
Method:                          IRLS   Log-Likelihood:
-2079.1
Date:                Sun, 01 Jan 2023   Deviance:
4158.1
Time:                        14:34:05   Pearson chi2:
4.80e+03
No. Iterations:                     7

Covariance Type:            nonrobust

======================================================================
========================================
                                                  coef    std err
```

```
z      P>|z|      [0.025      0.975]
----------------------------------------------------------------------
----------------------------------------
const                                                0.2040      0.196
1.043      0.297      -0.179       0.587
TotalVisits                                         11.1489      2.665
4.184      0.000       5.926      16.371
Total Time Spent on Website                          4.4223      0.185
23.899      0.000       4.060       4.785
Lead Origin_Lead Add Form                            4.2051      0.258
16.275      0.000       3.699       4.712
Lead Source_Olark Chat                               1.4526      0.122
11.934      0.000       1.214       1.691
Lead Source_Welingak Website                         2.1526      1.037
2.076      0.038       0.121       4.185
Do Not Email_Yes                                    -1.5037      0.193     -
7.774      0.000      -1.883      -1.125
Last Activity_Had a Phone Conversation               2.7552      0.802
3.438      0.001       1.184       4.326
Last Activity_SMS Sent                               1.1856      0.082
14.421      0.000       1.024       1.347
What is your current occupation_Student             -2.3578      0.281     -
8.392      0.000      -2.908      -1.807
What is your current occupation_Unemployed    -2.5445      0.186    -
13.699      0.000      -2.908      -2.180
Last Notable Activity_Unreachable                    2.7846      0.807
3.449      0.001       1.202       4.367
======================================================================
========================================
"""
```

Checking final VIF

```
# Making 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  2.82
1                  Total Time Spent on Website  2.00
0                                  TotalVisits  1.54
7                       Last Activity_SMS Sent  1.51
2                    Lead Origin_Lead Add Form  1.45
3                       Lead Source_Olark Chat  1.33
```

```
4                    Lead Source_Welingak Website  1.30
5                                Do Not Email_Yes  1.08
8      What is your current occupation_Student  1.06
6       Last Activity_Had a Phone Conversation  1.01
10            Last Notable Activity_Unreachable  1.01
```

Step 3: Model Evaluation

Now, both the p-values and VIFs seem decent enough for all the variables. So let's go ahead and make predictions using this final set of features.

```
# Use 'predict' to predict the probabilities on the train set

y_train_pred = res.predict(sm.add_constant(X_train))
y_train_pred[:10]
```

```
8003    0.300117
218     0.142002
4171    0.127629
4037    0.291558
3660    0.954795
207     0.194426
2044    0.178073
6411    0.949460
6498    0.075995
2085    0.982316
dtype: float64
```

```
# Reshaping it into an array

y_train_pred = y_train_pred.values.reshape(-1)
y_train_pred[:10]
```

```
array([0.30011695, 0.14200165, 0.12762885, 0.29155814, 0.95479546,
       0.19442563, 0.17807328, 0.94946006, 0.07599465, 0.98231619])
```

Creating a dataframe with the actual conversion flag and the predicted probabilities

```
# Creating a new dataframe containing the actual conversion flag and
the probabilities predicted by the model

y_train_pred_final = pd.DataFrame({'Converted':y_train.values,
'Conversion_Prob':y_train_pred})
y_train_pred_final.head()
```

```
   Converted  Conversion_Prob
0          0         0.300117
1          0         0.142002
```

```
2           1         0.127629
3           1         0.291558
4           1         0.954795
```

Creating new column 'Predicted' with 1 if Paid_Prob > 0.5 else 0

```
y_train_pred_final['Predicted'] =
y_train_pred_final.Conversion_Prob.map(lambda x: 1 if x > 0.5 else 0)

# Let's see the head
y_train_pred_final.head()
```

```
   Converted  Conversion_Prob  Predicted
0          0         0.300117          0
1          0         0.142002          0
2          1         0.127629          0
3          1         0.291558          0
4          1         0.954795          1
```

Now that you have the probabilities and have also made conversion predictions using them, it's time to evaluate the model.

```
# Importing the 'metrics' library from sklearn for evaluation
```

Creating the `Confusion matrix`

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

[[1929  383]
 [ 560 1589]]

# Let's check the overall accuracy

print(metrics.accuracy_score(y_train_pred_final.Converted,
y_train_pred_final.Predicted))

0.7886124187401928

# Let's evaluate the other metrics as well

TP = confusion[1,1] # true positive
TN = confusion[0,0] # true negatives
FP = confusion[0,1] # false positives
FN = confusion[1,0] # false negatives

# Calculating the 'sensitivity'

TP/(TP+FN)
```

```
0.739413680781759
```

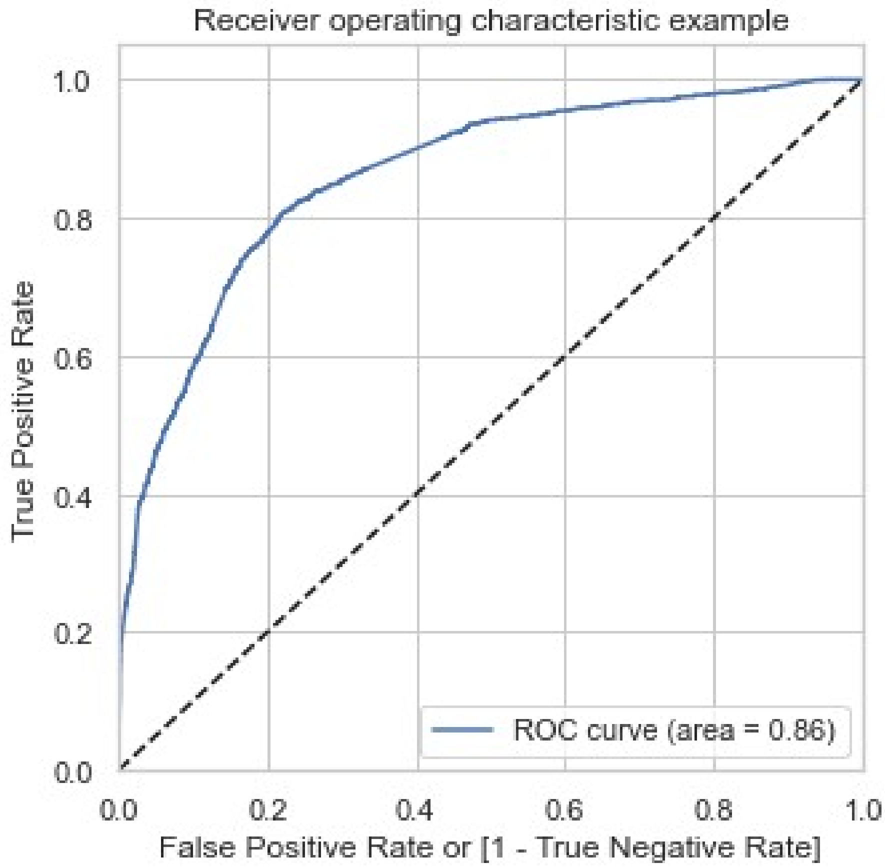
```
# Calculating the 'specificity'

TN/(TN+FP)
```

```
0.8343425605536332
```

Finding the Optimal Cutoff

Now 0.5 was just arbitrary to loosely check the model performace. But in order to get good results, you need to optimise the threshold. So first let's plot an ROC curve to see what AUC we get.

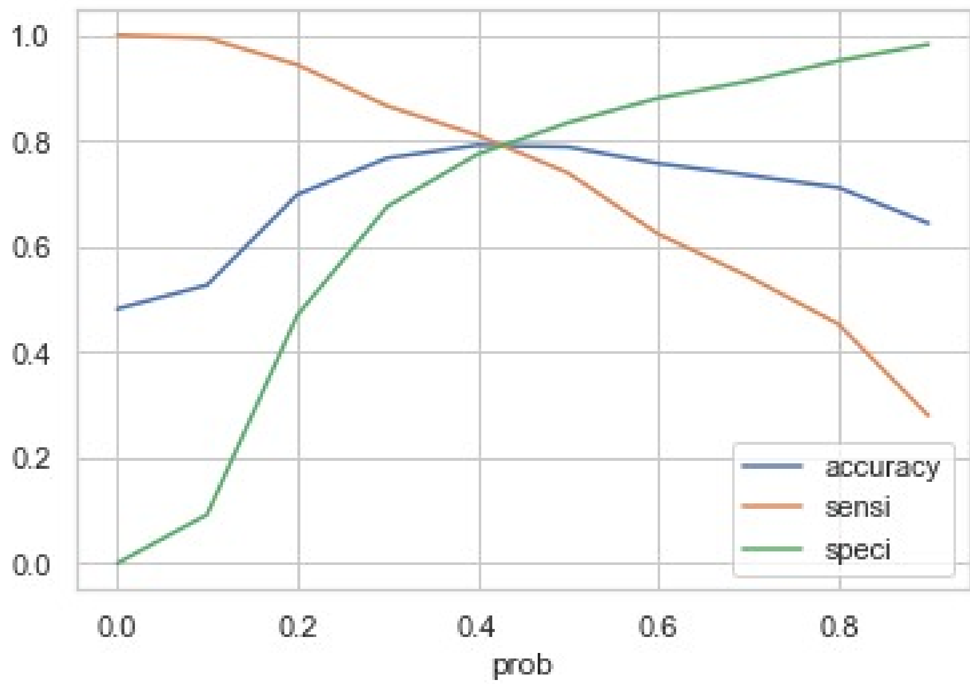
```
# ROC function

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

    return None

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

# Importing the 'matplotlib'  to plot the ROC curve`

# Calling the ROC function

draw_roc(y_train_pred_final.Converted,
y_train_pred_final.Conversion_Prob)
```



The area under the curve of the ROC is 0.86 which is quite good. So we seem to have a good model. Let's also check the sensitivity and specificity tradeoff to find the optimal cutoff point.

```
# Let's create 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          0         0.300117          0    1    1    1    1    0    0
0
1          0         0.142002          0    1    1    0    0    0    0
0
2          1         0.127629          0    1    1    0    0    0    0
0
3          1         0.291558          0    1    1    1    0    0    0
0
4          1         0.954795          1    1    1    1    1    1    1
1
```

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

# Let's create 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'])
from sklearn.metrics import confusion_matrix

# TP = confusion[1,1] # true positive
# TN = confusion[0,0] # true negatives
# FP = confusion[0,1] # false positives
# FN = confusion[1,0] # false negatives

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]
print(cutoff_df)

     prob  accuracy     sensi     speci
0.0   0.0  0.481731  1.000000  0.000000
0.1   0.1  0.527012  0.994416  0.092561
0.2   0.2  0.698274  0.944160  0.469723
0.3   0.3  0.767541  0.865984  0.676038
0.4   0.4  0.791975  0.810610  0.774654
0.5   0.5  0.788612  0.739414  0.834343
0.6   0.6  0.757229  0.624011  0.881055
0.7   0.7  0.735037  0.543509  0.913062
0.8   0.8  0.711500  0.453234  0.951557
0.9   0.9  0.644026  0.279665  0.982699

# Let's plot it as well

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



As you can see that around `0.42`, you get the optimal values of the three metrics. So let's choose 0.42 as our cutoff now.

```
y_train_pred_final['final_predicted'] =
y_train_pred_final.Conversion_Prob.map( lambda x: 1 if x > 0.42 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          0         0.300117          0    1    1    1    1    0    0
0
1          0         0.142002          0    1    1    0    0    0    0
0
2          1         0.127629          0    1    1    0    0    0    0
0
3          1         0.291558          0    1    1    1    0    0    0
0
4          1         0.954795          1    1    1    1    1    1    1
1

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

```
# Let's checking the `accuracy` now

metrics.accuracy_score(y_train_pred_final.Converted,
y_train_pred_final.final_predicted)

0.7908540685944856

# Let's create the confusion matrix once again

confusion2 = metrics.confusion_matrix(y_train_pred_final.Converted,
y_train_pred_final.final_predicted )
confusion2

array([[1823,  489],
       [ 444, 1705]], dtype=int64)

# Let's evaluate the other metrics as well

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

# Calculating the 'Sensitivity'

TP/(TP+FN)

0.793392275476966

# Calculating the 'Specificity'

TN/(TN+FP)

0.7884948096885813
```

This cutoff point seems good to go!

Step 4: Making Predictions on the Test Set

Let's now make predicitons on the test set

```
# Scaling the test set as well using just 'transform'

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']])

# Selecting the columns in X_train for X_test as well

X_test = X_test[col]
X_test.head()
```

```
      TotalVisits  Total Time Spent on Website  Lead Origin_Lead Add
Form  \
4771     0.000000                     0.000000
1
6122     0.027888                     0.029049
0
9202     0.015936                     0.416813
0
6570     0.011952                     0.378961
0
2668     0.031873                     0.395246
0

      Lead Source_Olark Chat  Lead Source_Reference  \
4771                       0                      1
6122                       0                      0
9202                       0                      0
6570                       0                      0
2668                       0                      0

      Lead Source_Welingak Website  Do Not Email_Yes  \
4771                             0                 0
6122                             0                 0
9202                             0                 0
6570                             0                 1
2668                             0                 0

      Last Activity_Had a Phone Conversation  Last Activity_SMS
Sent  \
4771                                       0                        1

6122                                       0                        0

9202                                       0                        1

6570                                       0                        1

2668                                       0                        1


      What is your current occupation_Housewife  \
4771                                          0
6122                                          0
9202                                          0
6570                                          0
2668                                          0

      What is your current occupation_Student  \
4771                                        0
6122                                        0
9202                                        0
```

```
6570                                              0
2668                                              0

      What is your current occupation_Unemployed  \
4771                                           0
6122                                           1
9202                                           1
6570                                           1
2668                                           1

      What is your current occupation_Working Professional  \
4771                                                     1
6122                                                     0
9202                                                     0
6570                                                     0
2668                                                     0

      Last Notable Activity_Had a Phone Conversation  \
4771                                               0
6122                                               0
9202                                               0
6570                                               0
2668                                               0

      Last Notable Activity_Unreachable
4771                                  0
6122                                  0
9202                                  0
6570                                  0
2668                                  0
```

```
# Adding a constant to X_test

X_test_sm = sm.add_constant(X_test[col])

# Checking X_test_sm

X_test_sm
```

```
      const  TotalVisits  Total Time Spent on Website  \
4771    1.0     0.000000                     0.000000
6122    1.0     0.027888                     0.029049
9202    1.0     0.015936                     0.416813
6570    1.0     0.011952                     0.378961
2668    1.0     0.031873                     0.395246
...     ...          ...                          ...
5828    1.0     0.011952                     0.027289
6583    1.0     0.011952                     0.152289
5531    1.0     0.055777                     0.702025
3056    1.0     0.011952                     0.417694
4088    1.0     0.019920                     0.530370
```

```
      Lead Origin_Lead Add Form  Lead Source_Olark Chat  \
4771                          1                       0
6122                          0                       0
9202                          0                       0
6570                          0                       0
2668                          0                       0
...                         ...                     ...
5828                          0                       0
6583                          0                       0
5531                          0                       0
3056                          0                       0
4088                          0                       0

      Lead Source_Reference  Lead Source_Welingak Website  Do Not
Email_Yes  \
4771                      1                             0
0
6122                      0                             0
0
9202                      0                             0
0
6570                      0                             0
1
2668                      0                             0
0
...                     ...                           ...
...
5828                      0                             0
0
6583                      0                             0
0
5531                      0                             0
0
3056                      0                             0
1
4088                      0                             0
0

      Last Activity_Had a Phone Conversation  Last Activity_SMS
Sent  \
4771                                       0                      1

6122                                       0                      0

9202                                       0                      1

6570                                       0                      1

2668                                       0                      1
```

```
...                                        ...                      ...
5828                                         0                        1
6583                                         0                        1
5531                                         0                        0
3056                                         0                        1
4088                                         0                        0

      What is your current occupation_Housewife  \
4771                                          0
6122                                          0
9202                                          0
6570                                          0
2668                                          0
...                                         ...
5828                                          0
6583                                          0
5531                                          0
3056                                          0
4088                                          0

      What is your current occupation_Student  \
4771                                        0
6122                                        0
9202                                        0
6570                                        0
2668                                        0
...                                       ...
5828                                        0
6583                                        0
5531                                        0
3056                                        0
4088                                        0

      What is your current occupation_Unemployed  \
4771                                           0
6122                                           1
9202                                           1
6570                                           1
2668                                           1
...                                          ...
5828                                           1
6583                                           1
5531                                           1
```

```
3056                                               1
4088                                               1

      What is your current occupation_Working Professional  \
4771                                                     1
6122                                                     0
9202                                                     0
6570                                                     0
2668                                                     0
...                                                    ...
5828                                                     0
6583                                                     0
5531                                                     0
3056                                                     0
4088                                                     0

      Last Notable Activity_Had a Phone Conversation  \
4771                                               0
6122                                               0
9202                                               0
6570                                               0
2668                                               0
...                                              ...
5828                                               0
6583                                               0
5531                                               0
3056                                               0
4088                                               0

      Last Notable Activity_Unreachable
4771                                  0
6122                                  0
9202                                  0
6570                                  0
2668                                  0
...                                 ...
5828                                  0
6583                                  0
5531                                  0
3056                                  0
4088                                  0

[1912 rows x 16 columns]

# Dropping the required columns from X_test as well

X_test.drop(['Lead Source_Reference', 'What is your current 
occupation_Housewife', 
             'What is your current occupation_Working Professional', 
                     'Last Notable Activity_Had a Phone 
```

```
Conversation'], 1,
                                inplace = True)

# Make predictions on the test set and store it in the variable
'y_test_pred'

y_test_pred = res.predict(sm.add_constant(X_test))

y_test_pred[:10]

4771    0.996296
6122    0.129992
9202    0.703937
6570    0.299564
2668    0.720796
4233    0.792250
3368    0.704038
9091    0.464521
5972    0.282978
3631    0.786460
dtype: float64

# Converting y_pred to a dataframe

y_pred_1 = pd.DataFrame(y_test_pred)

# Let's see the head

y_pred_1.head()

             0
4771  0.996296
6122  0.129992
9202  0.703937
6570  0.299564
2668  0.720796

# 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_1.reset_index(drop=True, inplace=True)
y_test_df.reset_index(drop=True, inplace=True)

# Append y_test_df and y_pred_1

y_pred_final = pd.concat([y_test_df, y_pred_1],axis=1)
```

```python
# Check 'y_pred_final'

y_pred_final.head()
```

```
   Converted         0
0          1  0.996296
1          0  0.129992
2          0  0.703937
3          1  0.299564
4          1  0.720796
```

```python
# Rename the column

y_pred_final= y_pred_final.rename(columns = {0 : 'Conversion_Prob'})

# Let's see the head of y_pred_final

y_pred_final.head()
```

```
   Converted  Conversion_Prob
0          1         0.996296
1          0         0.129992
2          0         0.703937
3          1         0.299564
4          1         0.720796
```

```python
# Make predictions on the test set using 0.45 as the cutoff

y_pred_final['final_predicted'] =
y_pred_final.Conversion_Prob.map(lambda x: 1 if x > 0.42 else 0)

# Check y_pred_final

y_pred_final.head()
```

```
   Converted  Conversion_Prob  final_predicted
0          1         0.996296                1
1          0         0.129992                0
2          0         0.703937                1
3          1         0.299564                0
4          1         0.720796                1
```

```python
# Let's check the overall accuracy

metrics.accuracy_score(y_pred_final['Converted'],
y_pred_final.final_predicted)
```

```
0.7845188284518828
```

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

```
array([[786, 210],
       [202, 714]], dtype=int64)
```

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

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

```
0.7794759825327511
```

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

```
0.7891566265060241
```

Precision-Recall View

Let's now also build the training model using the precision-recall view

```
#Looking at the confusion matrix again

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

```
array([[1929,  383],
       [ 560, 1589]], dtype=int64)
```

Precision =

```
     TP / TP + FP
```

```
confusion[1,1]/(confusion[0,1]+confusion[1,1])
```

```
0.8057809330628803
```

Recall =

```
      TP / TP + FN
```

```
confusion[1,1]/(confusion[1,0]+confusion[1,1])
```

```
0.739413680781759
```

Precision and recall tradeoff

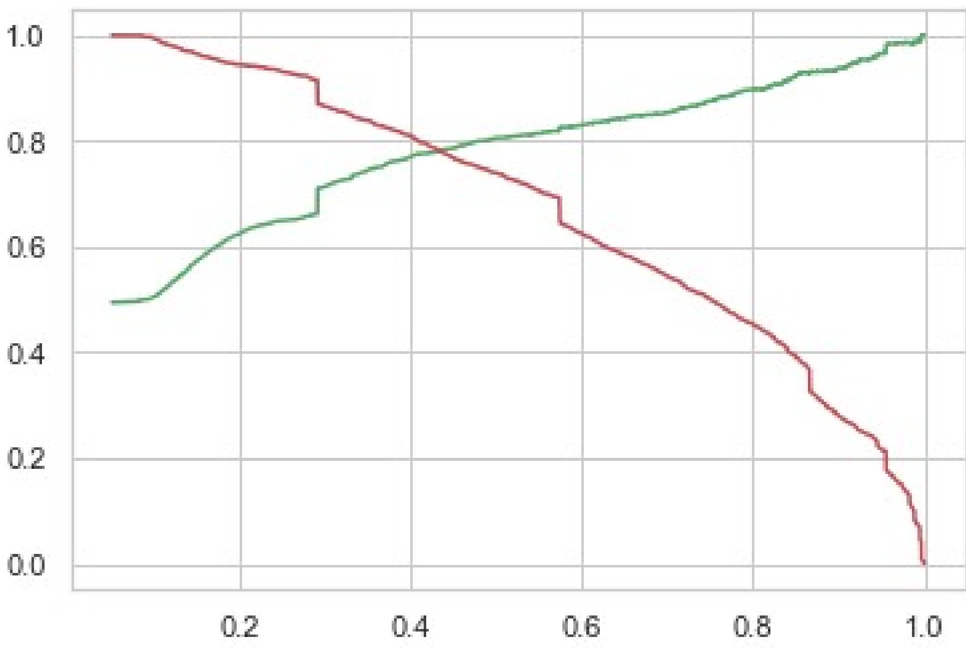
Importing the `Precision recall curve` library

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

```
(0       0
 1       0
 2       1
 3       1
 4       1
        ..
 4456    1
 4457    0
 4458    0
 4459    0
 4460    0
 Name: Converted, Length: 4461, dtype: int64,
 0       0
 1       0
 2       0
 3       0
 4       1
        ..
 4456    1
 4457    1
 4458    1
 4459    0
 4460    0
 Name: Predicted, Length: 4461, 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 > 0.44 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          0         0.300117          0    1    1    1    1    0    0
0
1          0         0.142002          0    1    1    0    0    0    0
0
2          1         0.127629          0    1    1    0    0    0    0
0
3          1         0.291558          0    1    1    1    0    0    0
0
4          1         0.954795          1    1    1    1    1    1    1
1

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

# Let's checking the `accuracy` now

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

```
0.7895090786819099
```

```
# Let's creating the confusion matrix once again

confusion2 = metrics.confusion_matrix(y_train_pred_final.Converted,
y_train_pred_final.final_predicted )
confusion2
```

```
array([[1852,  460],
       [ 479, 1670]], dtype=int64)
```

```
# Let's evaluate the other metrics as well

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

Precision

```
TP/(TP+FP)
```

```
0.784037558685446
```

Recall

```
TP/(TP+FN)
```

```
0.7771056305258259
```

This cutoff point seems good to go!

Step 5: Making Predictions on the Test Set

Let's now make predicitons on the test set.

```
# Making predictions on the test set and store it in the variable
'y_test_pred'

y_test_pred = res.predict(sm.add_constant(X_test))

y_test_pred[:10]
```

```
4771    0.996296
6122    0.129992
9202    0.703937
6570    0.299564
2668    0.720796
4233    0.792250
3368    0.704038
9091    0.464521
```

```
5972    0.282978
3631    0.786460
dtype: float64

# Converting y_pred to a dataframe

y_pred_1 = pd.DataFrame(y_test_pred)

# Let's see the head

y_pred_1.head()

             0
4771  0.996296
6122  0.129992
9202  0.703937
6570  0.299564
2668  0.720796

# Converting y_test to dataframe

y_test_df = pd.DataFrame(y_test)

# Removing index for both dataframes to append them side by side

y_pred_1.reset_index(drop=True, inplace=True)
y_test_df.reset_index(drop=True, inplace=True)

# Append y_test_df and y_pred_1

y_pred_final = pd.concat([y_test_df, y_pred_1],axis=1)

# Checking the 'y_pred_final'

y_pred_final.head()

   Converted         0
0          1  0.996296
1          0  0.129992
2          0  0.703937
3          1  0.299564
4          1  0.720796

# Rename the column

y_pred_final= y_pred_final.rename(columns = {0 : 'Conversion_Prob'})

# Let's see the head of y_pred_final

y_pred_final.head()
```

```
   Converted  Conversion_Prob
0          1         0.996296
1          0         0.129992
2          0         0.703937
3          1         0.299564
4          1         0.720796

# Making predictions on the test set using 0.44 as the cutoff

y_pred_final['final_predicted'] =
y_pred_final.Conversion_Prob.map(lambda x: 1 if x > 0.44 else 0)

# Checking y_pred_final

y_pred_final.head()

   Converted  Conversion_Prob  final_predicted
0          1         0.996296                1
1          0         0.129992                0
2          0         0.703937                1
3          1         0.299564                0
4          1         0.720796                1

# Let's checking the overall accuracy

metrics.accuracy_score(y_pred_final['Converted'],
y_pred_final.final_predicted)

0.7866108786610879

confusion2 = metrics.confusion_matrix(y_pred_final['Converted'],
y_pred_final.final_predicted )
confusion2

array([[801, 195],
       [213, 703]], dtype=int64)

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

# Calculating the Precision

TP/(TP+FP)

0.7828507795100222

# Calculating Recall

TP/(TP+FN)
```

```
0.767467248908297
```