Lead_Scoring_Assignment_-_Logistic_Regression

June 25, 2024

1 Problem Statement

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

The company markets its courses on several websites and search engines like Google. Once these people land on the website, they might browse the courses or fill up a form for the course or watch some videos. When these people fill up a form providing their email address or phone number, they are classified to be a lead. Moreover, the company also gets leads through past referrals. Once these leads are acquired, employees from the sales team start making calls, writing emails, etc. Through this process, some of the leads get converted while most do not. The typical lead conversion rate at X education is around 30%.

Now, although X Education gets a lot of leads, its lead conversion rate is very poor. For example, if, say, they acquire 100 leads in a day, only about 30 of them are converted. To make this process more efficient, the company wishes to identify the most potential leads, also known as 'Hot Leads'. If they successfully identify this set of leads, the lead conversion rate should go up as the sales team will now be focusing more on communicating with the potential leads rather than making calls to everyone. A typical lead conversion process can be represented using the following funnel:

Lead Conversion Process - Demonstrated as a funnel

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

X Education has appointed you to help them select the most promising leads, i.e. the leads that are most likely to convert into paying customers. The company requires you to build a model wherein you need to assign a lead score to each of the leads such that the customers with higher lead score have a higher conversion chance and the customers with lower lead score have a lower conversion chance. The CEO, in particular, has given a ballpark of the target lead conversion rate to be around 80%.

Data You have been provided with a leads dataset from the past with around 9000 data points. This dataset consists of various attributes such as Lead Source, Total Time Spent on Website, Total Visits, Last Activity, etc. which may or may not be useful in ultimately deciding whether a lead will be converted or not. The target variable, in this case, is the column 'Converted' which tells whether a past lead was converted or not wherein 1 means it was converted and 0 means it

wasn't converted. You can learn more about the dataset from the data dictionary provided in the zip folder at the end of the page. Another thing that you also need to check out for are the levels present in the categorical variables. Many of the categorical variables have a level called 'Select' which needs to be handled because it is as good as a null value (think why?).

2 Goals of the Case Study

2.1 There are quite a few goals for this case study.

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

3 Results Expected

1. A well-commented Jupyter note with at least the logistic regression model, the conversion predictions and evaluation metrics.
2. The word document filled with solutions to all the problems.
3. The overall approach of the analysis in a presentation
    1. Mention the problem statement and the analysis approach briefly
    2. Explain the results in business terms
    3. Include visualisations and summarise the most important results in the presentation
4. A brief summary report in 500 words explaining how you proceeded with the assignment and the learnings that you gathered.

You need to submit the following four components:

- Python commented file: Should include detailed comments and should not contain unnecessary pieces of code.
- Word File: Answer all the questions asked by the company in the word document provided.
- Presentation: Make a presentation to present your analysis to the chief data scientist of your company (and thus you should include both technical and business aspects). The presentation should be concise, clear, and to the point. Submit the presentation after converting it into PDF format.
- PDF File: Write the summary report in a word file and submit it as a PDF.

4 1. Data Reading and Understanding

1. importing neccessary libraries
2. reading the dataset provided
3. checking basic, statistical and value count info of all the columns to find various insights

[4]:
```
# importing required libraries for Data reading and understanding purpose
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns

# importing libraries for suppressing warnings
import warnings
warnings.filterwarnings('ignore')
```

[5]:
```
# notebook setting to display all the rowns and columns to have better clearity␣
↪on the data.

pd.set_option('display.max_rows', 500)
pd.set_option('display.max_columns', 500)
pd.set_option('display.width', 1000)
pd.set_option('display.expand_frame_repr', False)
```

[135]:
```
# reading the csv file
leads = pd.read_csv('Leads.csv')

leads.head()
```

[135]:
```
                            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
0  7927b2df-8bba-4d29-b9a2-b6e0beafe620       660737                       API
Olark Chat           No          No          0          0.0
0                   0.0  Page Visited on Website     NaN
Select                             Select                       Unemployed
Better Career Prospects     No       No                No                  No
No                    No                      No
No          Interested in other courses  Low in Relevance
No                          No          Select  Select                   02.Medium
02.Medium                           15.0                         15.0
No                                     No              Modified
1  2a272436-5132-4136-86fa-dcc88c88f482       660728                       API
Organic Search           No          No          0          5.0
```

```
674                    2.5             Email Opened   India
Select                               Select                       Unemployed
Better Career Prospects     No       No                No                  No
No                     No                        No
No                                Ringing               NaN
No                          No          Select  Select                    02.Medium
02.Medium                          15.0                        15.0
No                                      No          Email Opened
2  8cc8c611-a219-4f35-ad23-fdfd2656bd8a       660727  Landing Page Submission
Direct Traffic           No          No          1          2.0
1532                    2.0             Email Opened   India  Business
Administration                              Select
Student                         Better Career Prospects     No       No
No                  No        No                    No                        No
No  Will revert after reading the email          Might be
No                          No  Potential Lead  Mumbai                    02.Medium
01.High                           14.0                         20.0
No                                      Yes          Email Opened
3  0cc2df48-7cf4-4e39-9de9-19797f9b38cc       660719  Landing Page Submission
Direct Traffic           No          No          0          1.0
305                     1.0             Unreachable   India    Media and
Advertising                          Word Of Mouth                       Unemployed
Better Career Prospects     No       No                No                  No
No                     No                        No
No                                Ringing          Not Sure
No                          No          Select  Mumbai                    02.Medium
01.High                           13.0                         17.0
No                                      No              Modified
4  3256f628-e534-4826-9d63-4a8b88782852       660681  Landing Page Submission
Google           No          No          1          2.0
1428                    1.0        Converted to Lead   India
Select                                 Other                        Unemployed
Better Career Prospects     No       No                No                  No
No                     No                        No
No  Will revert after reading the email          Might be
No                          No          Select  Mumbai                    02.Medium
01.High                           15.0                         18.0
No                                      No              Modified
```

[7]:
```
# checking shape of the data set
leads.shape
```

[7]:
```
(9240, 37)
```

[8]:
```
# checking all the info of the data set like column data types, total entries
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
```

- there 9240 rows and 37 columns
- from the above the information we can see that there are 7 numerical variables and rest 30 are categorical variables

[10]:
```
# checking all the statistical information of numerical variables
leads.describe(percentiles=[0.25,0.50,0.75,0.90,0.95,0.97,0.99])
```

[10]:

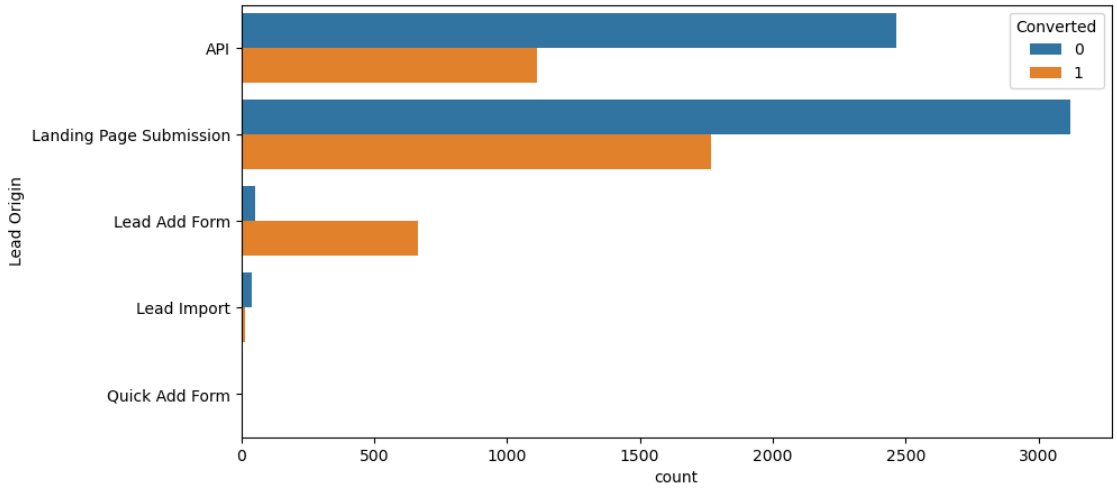
	Lead Number	Converted	TotalVisits	Total Time Spent on Website	Page Views Per Visit	Asymmetrique Activity Score	Asymmetrique Profile Score
count	9240.000000	9240.000000	9103.000000	9240.000000	9103.000000	5022.000000	5022.000000
mean	617188.435606	0.385390	3.445238	487.698268	2.362820	14.306252	16.344883
std	23405.995698	0.486714	4.854853	548.021466	2.161418	1.386694	1.811395
min	579533.000000	0.000000	0.000000	0.000000	0.000000	7.000000	11.000000
25%	596484.500000	0.000000	1.000000	12.000000	1.000000	14.000000	15.000000
50%	615479.000000	0.000000	3.000000	248.000000	2.000000	14.000000	16.000000
75%	637387.250000	1.000000	5.000000	936.000000	3.000000	15.000000	18.000000
90%	650506.100000	1.000000	7.000000	1380.000000	5.000000	16.000000	19.000000
95%	655404.050000	1.000000	10.000000	1562.000000	6.000000	17.000000	20.000000
97%	657466.940000	1.000000	11.000000	1660.000000	7.000000	17.000000	20.000000
99%	659592.980000	1.000000	17.000000	1840.610000	9.000000	17.000000	20.000000
max	660737.000000	1.000000	251.000000	2272.000000	55.000000	18.000000	20.000000

from above statistical info, we can see that there are outlier present mostly in 'totalVisits','Total Time Spent on Website' and 'Page Views Per Visit'

4.0.1 Exploratory Data Analysis

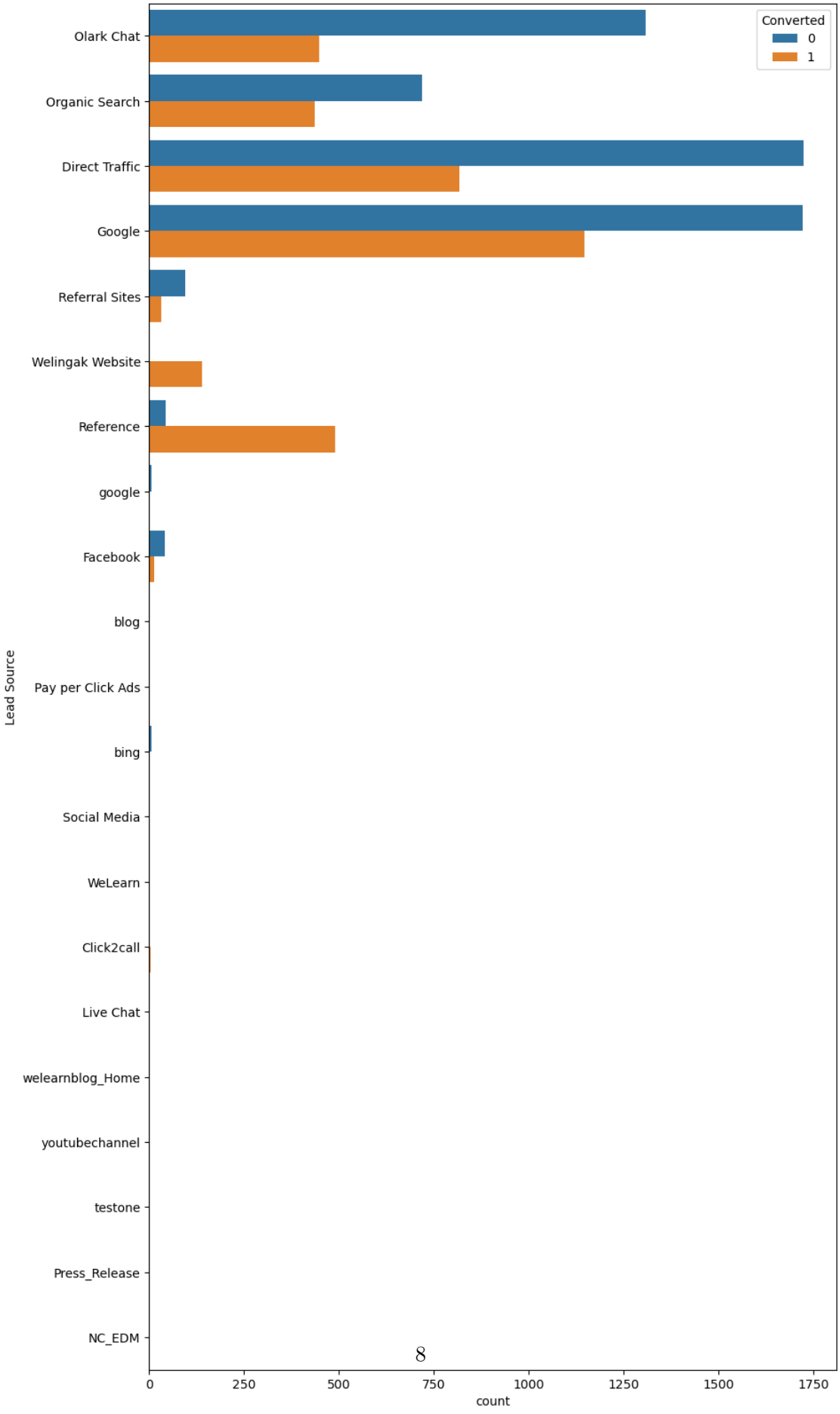
Let us try and understand the data now based on each columns effect on the conversion rates

[14]:
```
#Understanding Lead Conversion and Lead Origin
plt.figure(figsize=(10, 5))
sns.countplot(y="Lead Origin", hue="Converted", data=leads)
plt.show()
```



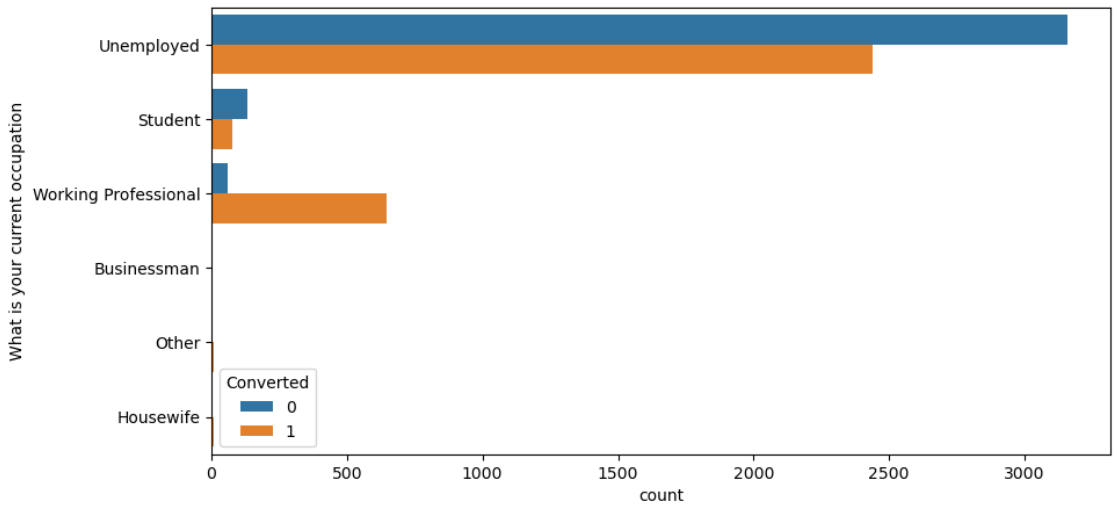
From Lead Origin finding, maximum lead conversion happened from Landing Page Submission.

[16]:
```
#Understanding Lead Conversion and Lead Source
plt.figure(figsize=(10, 20))
sns.countplot(y="Lead Source", hue="Converted", data=leads)
plt.show()
```


Converted
0
1
Olark Chat
Organic Search
Direct Traffic
Google
Referral Sites
Welingak Website
Reference
google
Facebook
blog
Pay per Click Ads
bing
Social Media
WeLearn
Click2call
Live Chat
welearnblog_Home
youtubechannel
testone
Press_Release
NC_EDM
Lead Source
0
250
500
750
1000
1250
1500
1750
count

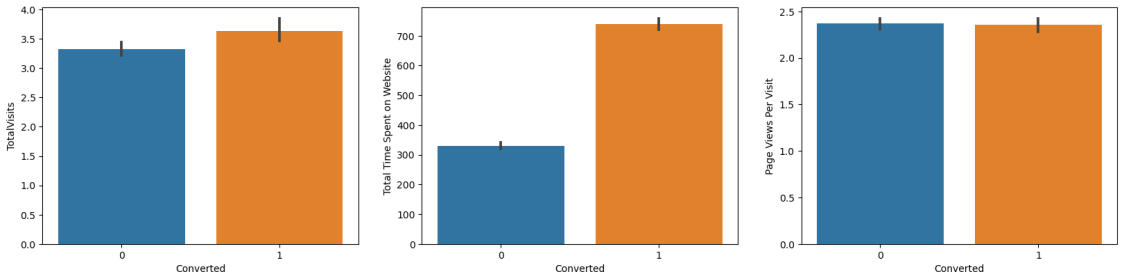
From the above graph, major lead conversion in the lead source is from 'Google'

[18]:
```
#Understanding Lead Conversion and Current Occupation
plt.figure(figsize=(10, 5))
sns.countplot(y="What is your current occupation", hue="Converted", data=leads)
plt.show()
```



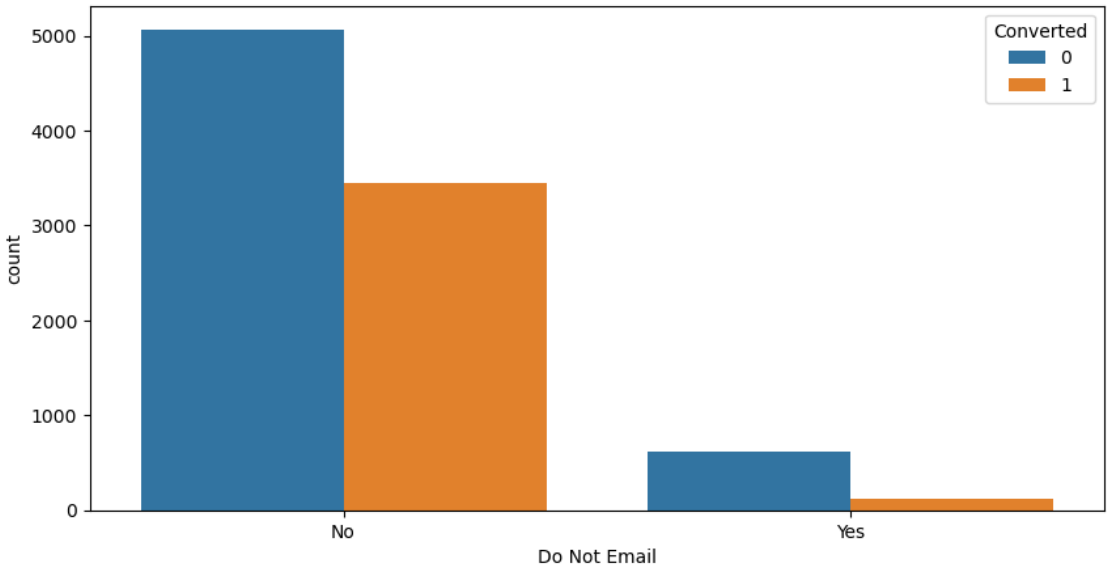
From the above graph, major lead conversion is from the Unemployed Group

[20]:
```
#Understanding the Lead Conversion on TotalVisits, Total Time Spent on Website,␣
 ↪Page Views Per Visit
plt.figure(figsize=(20, 15))
plt.subplot(3,3,1)
sns.barplot(x = 'Converted', y = 'TotalVisits', data = leads)

plt.subplot(3,3,2)
sns.barplot(x = 'Converted', y = 'Total Time Spent on Website',data = leads)

plt.subplot(3,3,3)
sns.barplot(x = 'Converted', y = 'Page Views Per Visit',data = leads)

plt.show()
```



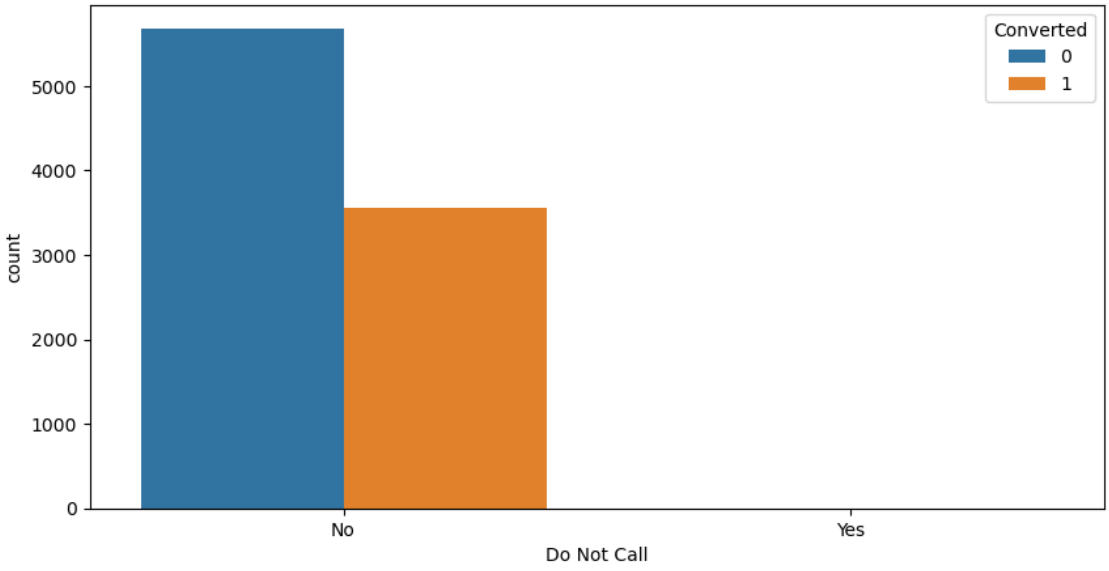
From the above graph, we have mojor lead conversion from TotalVisits, Total Time Spent on Website, Page Views Per Visit

[22]:
```
#Understanding Lead Conversion and Do Not Email
plt.figure(figsize=(10, 5))
sns.countplot(x="Do Not Email", hue="Converted", data=leads)
plt.show()
```



Based on the above graph, major lead conversion has happend from the emails that have been sent

[24]:
```
#Understanding Lead Conversion and Do Not Call
plt.figure(figsize=(10, 5))
sns.countplot(x="Do Not Call", hue="Converted", data=leads)
plt.show()
```



Observations from EDA Process -

- Maximum lead conversion happened from Landing Page Submission.
- Major lead conversion in the lead source is from 'Google'
- Major lead conversion is from the Unemployed Group
- Major lead conversion from TotalVisits, Total Time Spent on Website, Page Views Per Visit
- Major conversion has happened from the emails that have been sent

[26]:
```
# checking the value counts of each variables to find insights

for i in leads.columns:
    print(leads[i].value_counts())
    print('----------------------------------------')
```

```
Prospect ID
7927b2df-8bba-4d29-b9a2-b6e0beafe620    1
22e9d4ef-d294-4ebf-81c7-7c7a1105aeea    1
46befc49-253a-419b-abea-2fd978d2e2b1    1
9d35a2c2-09d8-439f-9875-0e8bbf267f5a    1
f0de9371-4dc2-48c2-9785-a08d6fc4fcb5    1
                                       ..
ff1f7582-cb7b-4b94-9cdc-3d0d0afdd9a3    1
644099a2-3da4-4d23-9546-7676340a372b    1
2a093175-415b-4321-9e69-ed8d9df65a3c    1
c66249a3-8500-4c66-a511-312d914573de    1
571b5c8e-a5b2-4d57-8574-f2ffb06fdeff    1
Name: count, Length: 9240, dtype: int64
----------------------------------------
```

```
Lead Number
660737    1
603303    1
602561    1
602557    1
602540    1
         ..
630422    1
630405    1
630403    1
630390    1
579533    1
Name: count, Length: 9240, dtype: int64
--------------------------------------
Lead Origin
Landing Page Submission    4886
API                        3580
Lead Add Form               718
Lead Import                  55
Quick Add Form                1
Name: count, dtype: int64
--------------------------------------
Lead Source
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
youtubechannel          1
testone                 1
Pay per Click Ads       1
welearnblog_Home        1
WeLearn                 1
blog                    1
NC_EDM                  1
Name: count, dtype: int64
--------------------------------------
Do Not Email
No     8506
```

```
Yes     734
Name: count, dtype: int64
---------------------------------------
Do Not Call
No     9238
Yes       2
Name: count, dtype: int64
---------------------------------------
Converted
0    5679
1    3561
Name: count, dtype: int64
---------------------------------------
TotalVisits
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
21.0        6
23.0        6
24.0        5
25.0        5
27.0        5
22.0        3
29.0        2
28.0        2
26.0        2
141.0       1
55.0        1
30.0        1
43.0        1
```

```
74.0        1
41.0        1
54.0        1
115.0       1
251.0       1
32.0        1
42.0        1
Name: count, dtype: int64
---------------------------------------
Total Time Spent on Website
0       2193
60        19
74        18
75        18
127       18
        …
1701       1
1952       1
1229       1
1743       1
927        1
Name: count, Length: 1731, dtype: int64
---------------------------------------
Page Views Per Visit
0.00     2189
2.00     1795
3.00     1196
4.00      896
1.00      651
5.00      517
1.50      306
6.00      244
2.50      241
7.00      133
3.50       94
8.00       86
1.33       66
1.67       60
2.33       59
2.67       54
9.00       45
4.50       43
1.75       28
3.33       27
10.00      25
1.25       23
5.50       21
2.25       19
```

11.00	18
3.67	16
6.50	13
1.80	13
2.75	12
1.40	11
2.80	9
14.00	9
4.33	9
2.20	9
2.17	8
1.60	8
3.25	8
2.40	6
13.00	6
12.00	5
1.20	5
1.83	4
3.40	4
2.60	4
1.43	4
15.00	4
1.71	4
1.78	3
4.25	3
16.00	3
4.75	3
5.67	3
1.57	3
1.38	3
3.60	2
1.23	2
1.56	2
5.40	2
2.22	2
7.50	2
1.14	2
5.25	2
2.09	2
3.20	2
3.75	2
5.33	2
2.83	2
2.71	2
1.22	2
2.13	1
1.54	1
6.67	1

```
24.00       1
2.14        1
2.45        1
3.29        1
1.48        1
3.82        1
4.17        1
1.63        1
3.38        1
1.17        1
14.50       1
3.80        1
1.19        1
3.17        1
1.93        1
11.50       1
8.33        1
55.00       1
4.40        1
8.21        1
8.50        1
2.63        1
1.27        1
2.57        1
2.86        1
3.91        1
6.71        1
3.57        1
1.31        1
2.90        1
3.83        1
1.45        1
2.38        1
1.86        1
2.29        1
1.21        1
12.33       1
3.43        1
2.56        1
6.33        1
1.64        1
2.08        1
Name: count, dtype: int64
--------------------------------------
Last Activity
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
Visited Booth in Tradeshow         1
Resubscribed to emails             1
Name: count, dtype: int64
---------------------------------------
Country
India                   6492
United States             69
United Arab Emirates      53
Singapore                 24
Saudi Arabia              21
United Kingdom            15
Australia                 13
Qatar                     10
Hong Kong                  7
Bahrain                    7
Oman                       6
France                     6
unknown                    5
South Africa               4
Nigeria                    4
Germany                    4
Kuwait                     4
Canada                     4
Sweden                     3
China                      2
Asia/Pacific Region        2
Uganda                     2
Bangladesh                 2
Italy                      2
Belgium                    2
Netherlands                2
Ghana                      2
Philippines                2
Russia                     1
Switzerland                1
Vietnam                    1
```

```
Denmark                    1
Tanzania                   1
Liberia                    1
Malaysia                   1
Kenya                      1
Sri Lanka                  1
Indonesia                  1
Name: count, dtype: int64
---------------------------------------
Specialization
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: count, dtype: int64
---------------------------------------
How did you hear about X Education
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: count, dtype: int64
---------------------------------------
What is your current occupation
Unemployed              5600
Working Professional     706
Student                  210
```

```
Other                     16
Housewife                 10
Businessman                8
Name: count, dtype: int64
---------------------------------------
What matters most to you in choosing a course
Better Career Prospects      6528
Flexibility & Convenience       2
Other                           1
Name: count, dtype: int64
---------------------------------------
Search
No     9226
Yes      14
Name: count, dtype: int64
---------------------------------------
Magazine
No    9240
Name: count, dtype: int64
---------------------------------------
Newspaper Article
No     9238
Yes       2
Name: count, dtype: int64
---------------------------------------
X Education Forums
No     9239
Yes       1
Name: count, dtype: int64
---------------------------------------
Newspaper
No     9239
Yes       1
Name: count, dtype: int64
---------------------------------------
Digital Advertisement
No     9236
Yes       4
Name: count, dtype: int64
---------------------------------------
Through Recommendations
No     9233
Yes       7
Name: count, dtype: int64
---------------------------------------
Receive More Updates About Our Courses
No    9240
Name: count, dtype: int64
```

```
--------------------------------------
Tags
Will revert after reading the email                  2072
Ringing                                               1203
Interested in other courses                            513
Already a student                                      465
Closed by Horizzon                                     358
switched off                                           240
Busy                                                   186
Lost to EINS                                           175
Not doing further education                            145
Interested  in full time MBA                           117
Graduation in progress                                 111
invalid number                                          83
Diploma holder (Not Eligible)                           63
wrong number given                                      47
opp hangup                                              33
number not provided                                     27
in touch with EINS                                      12
Lost to Others                                           7
Still Thinking                                           6
Want to take admission but has financial problems        6
In confusion whether part time or DLP                    5
Interested in Next batch                                 5
Lateral student                                          3
Shall take in the next coming month                      2
University not recognized                                2
Recognition issue (DEC approval)                         1
Name: count, dtype: int64
--------------------------------------
Lead Quality
Might be              1560
Not Sure              1092
High in Relevance      637
Worst                  601
Low in Relevance       583
Name: count, dtype: int64
--------------------------------------
Update me on Supply Chain Content
No    9240
Name: count, dtype: int64
--------------------------------------
Get updates on DM Content
No    9240
Name: count, dtype: int64
--------------------------------------
Lead Profile
Select                         4146
```

```
Potential Lead                1613
Other Leads                    487
Student of SomeSchool          241
Lateral Student                 24
Dual Specialization Student     20
Name: count, dtype: int64
--------------------------------------
City
Mumbai                        3222
Select                        2249
Thane & Outskirts              752
Other Cities                   686
Other Cities of Maharashtra    457
Other Metro Cities             380
Tier II Cities                  74
Name: count, dtype: int64
--------------------------------------
Asymmetrique Activity Index
02.Medium    3839
01.High       821
03.Low        362
Name: count, dtype: int64
--------------------------------------
Asymmetrique Profile Index
02.Medium    2788
01.High      2203
03.Low         31
Name: count, dtype: int64
--------------------------------------
Asymmetrique Activity Score
14.0    1771
15.0    1293
13.0     775
16.0     467
17.0     349
12.0     196
11.0      95
10.0      57
9.0        9
18.0       5
8.0        4
7.0        1
Name: count, dtype: int64
--------------------------------------
Asymmetrique Profile Score
15.0    1759
18.0    1071
16.0     599
```

```
17.0     579
20.0     308
19.0     245
14.0     226
13.0     204
12.0      22
11.0       9
Name: count, dtype: int64
--------------------------------------
I agree to pay the amount through cheque
No    9240
Name: count, dtype: int64
--------------------------------------
A free copy of Mastering The Interview
No     6352
Yes    2888
Name: count, dtype: int64
--------------------------------------
Last Notable Activity
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
Resubscribed to emails           1
View in browser link Clicked     1
Form Submitted on Website        1
Email Received                   1
Name: count, dtype: int64
--------------------------------------
```

from the above the information we will make respective insights - we will be removing these variables - Prospect ID - not required - Lead Number - not required - Country- not required - Receive More Updates About Our Courses- column only has 'No' doesn't makes sense to keep it. - Update me on Supply Chain Content - column only has 'No' doesn't makes sense to keep it - Get updates on DM Content - column only has 'No' doesn't makes sense to keep it - I agree to pay the amount through cheque -column only has 'No' doesn't makes sense to keep it - Magazine - column only has 'No' doesn't makes sense to keep it

- We will transform below columns of yes/no category to 1/0:
  - Do Not Email
  - Do Not Call

- Search
- Newspaper Article
- X Education Forums
- Newspaper
- Digital Advertisement
- Through Recommendations
- a free copy of Mastering The Interview

---

5 2. Data Cleaning

1. Cleaning the dataset by removing the redundant variables/features.
2. After removing the redundant columns, we found that some columns are having label as 'Select' which means customer chose to not answer this question. Thus we would label null value to 'select' label.
3. Remove columns having more than 40% null values
4. Imputing missing values as per column data available

[30]:
```
#dropping redundant columns from above insights

leads = leads.drop(['Prospect ID', 'Lead Number', 'Country', 'Receive More␣
↪Updates About Our Courses',
                    'Update me on Supply Chain Content', 'Get updates on DM Content',␣
↪'City',
                    'I agree to pay the amount through cheque','Magazine'], axis = 1 )
```

[31]:
```
leads.shape
```

[31]:
```
(9240, 28)
```

now we have noticed that there are columns which have 'select' category which means customer did not select any of the options. they eventually act as null values, thus we will make them null.

dealing with 'Select' label

[33]:
```
# Creating a for loop and listing the columns having 'Select'
have_select = []
for i in leads.columns:
    if len(leads[i].isin(['Select']).unique())>1:
        have_select.append(i)

have_select # Columns having Select option
```

[33]:
```
['Specialization', 'How did you hear about X Education', 'Lead Profile']
```

[34]:
```
# now replacing 'Select' category with null values
```

```
for i in have_select:
    leads[i] = leads[i].replace('Select',np.NaN)

leads.head()
```

[34]:

```
              Lead Origin     Lead Source Do Not Email Do Not Call  Converted
TotalVisits  Total Time Spent on Website  Page Views Per Visit            Last
Activity           Specialization How did you hear about X Education What is
your current occupation What matters most to you in choosing a course Search
Newspaper Article X Education Forums Newspaper Digital Advertisement Through
Recommendations                                    Tags      Lead Quality    Lead
Profile Asymmetrique Activity Index Asymmetrique Profile Index  Asymmetrique
Activity Score  Asymmetrique Profile Score A free copy of Mastering The
Interview Last Notable Activity
0                       API      Olark Chat           No          No          0
0.0                           0                   0.0  Page Visited on Website
NaN                               NaN                       Unemployed
Better Career Prospects     No                No                  No        No
No                        No          Interested in other courses  Low in
Relevance             NaN                    02.Medium                  02.Medium
15.0                          15.0                                      No
Modified
1                       API  Organic Search           No          No          0
5.0                            674                   2.5             Email Opened
NaN                               NaN                       Unemployed
Better Career Prospects     No                No                  No        No
No                        No                                Ringing
NaN             NaN                   02.Medium                  02.Medium
15.0                          15.0                                      No
Email Opened
2  Landing Page Submission  Direct Traffic           No          No          1
2.0                           1532                  2.0             Email Opened
Business Administration                                  NaN
Student                        Better Career Prospects     No                No
No        No                    No                        No  Will revert after
reading the email          Might be  Potential Lead                  02.Medium
01.High                           14.0                        20.0
Yes          Email Opened
3  Landing Page Submission  Direct Traffic           No          No          0
1.0                            305                   1.0             Unreachable
Media and Advertising                      Word Of Mouth
Unemployed                         Better Career Prospects     No
No                  No        No                    No                        No
Ringing          Not Sure             NaN                   02.Medium
01.High                          13.0                        17.0
No              Modified
4  Landing Page Submission          Google           No          No          1
```

```
2.0                            1428                    1.0        Converted to Lead
NaN                                 Other                        Unemployed
Better Career Prospects     No                 No                    No        No
No                        No  Will revert after reading the email          Might
be             NaN                    02.Medium                      01.High
15.0                           18.0                                     No
Modified
```

[35]:
```
# Checking percentage of missing values after removing the imputing 'Select'
↪with Null values

round(100*(leads.isnull().sum()/len(leads.index)), 2)
```

[35]:
```
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                                36.58
How did you hear about X Education            78.46
What is your current occupation               29.11
What matters most to you in choosing a course 29.32
Search                                         0.00
Newspaper Article                              0.00
X Education Forums                             0.00
Newspaper                                      0.00
Digital Advertisement                          0.00
Through Recommendations                        0.00
Tags                                          36.29
Lead Quality                                  51.59
Lead Profile                                  74.19
Asymmetrique Activity Index                   45.65
Asymmetrique Profile Index                    45.65
Asymmetrique Activity Score                   45.65
Asymmetrique Profile Score                    45.65
A free copy of Mastering The Interview         0.00
Last Notable Activity                          0.00
dtype: float64
```

from above we see there are columns having more than 40% missing values, so it is better to remove these columns as it imputing them could lead to bias predictions.

dropping columns having missing values above 40%

[37]:
```
#dropping columns having missing values more than 40%
above_40 = list(round(100*(leads.isnull().sum()/len(leads.index)),␣
 ↪2)[round(100*(leads.isnull().sum()/len(leads.index)), 2) > 40].index)
leads = leads.drop(above_40, axis =1)
leads.head()
```

[37]:
```
              Lead Origin     Lead Source Do Not Email Do Not Call  Converted
TotalVisits  Total Time Spent on Website  Page Views Per Visit            Last
Activity           Specialization What is your current occupation What matters
most to you in choosing a course Search Newspaper Article X Education Forums
Newspaper Digital Advertisement Through Recommendations
Tags A free copy of Mastering The Interview Last Notable Activity
0                       API      Olark Chat           No          No          0
0.0                           0                   0.0  Page Visited on Website
NaN                       Unemployed                         Better Career
Prospects     No                No                  No        No
No                        No          Interested in other courses
No              Modified
1                       API  Organic Search           No          No          0
5.0                           674                   2.5            Email Opened
NaN                       Unemployed                         Better Career
Prospects     No                No                  No        No
No                        No                               Ringing
No          Email Opened
2  Landing Page Submission  Direct Traffic           No          No          1
2.0                           1532                  2.0            Email Opened
Business Administration                          Student
Better Career Prospects     No                No                  No        No
No                        No  Will revert after reading the email
Yes          Email Opened
3  Landing Page Submission  Direct Traffic           No          No          0
1.0                            305                  1.0             Unreachable
Media and Advertising                        Unemployed
Better Career Prospects     No                   No                  No        No
No                        No                                 Ringing
No              Modified
4  Landing Page Submission          Google           No          No          1
2.0                           1428                  1.0        Converted to Lead
NaN                       Unemployed                         Better Career
Prospects     No                No                  No        No
No                        No  Will revert after reading the email
No              Modified
```

[38]:
```
#checking shape of data set after removing columns
leads.shape
```

[38]:
```
(9240, 21)
```

Missing Values Imputation Now we will impute values for columns having missing values less than 40%

[40]:
```
# finding columns having missing values above 0 and below 40 %

below_40 = list(round(100*(leads.isnull().sum()/len(leads.index)),␣
 ↪2)[round(100*(leads.isnull().sum()/len(leads.index)), 2) > 0].index)

below_40
```

[40]:
```
['Lead Source',
 'TotalVisits',
 'Page Views Per Visit',
 'Last Activity',
 'Specialization',
 'What is your current occupation',
 'What matters most to you in choosing a course',
 'Tags']
```

Note:- from the problem statement we get that the columns above, 'Last Activity', 'Tags' are provided by sales team. We will remove them before model building as the we don't a model having these features.

[42]:
```
# 1. Dealing Lead Source

print(leads['Lead Source'].value_counts())
print('-----------')
print('Missing values count --->', leads['Lead Source'].isna().sum())
print('====================================================================')
```

```
Lead Source
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
youtubechannel          1
testone                 1
Pay per Click Ads       1
```

```
welearnblog_Home        1
WeLearn                 1
blog                    1
NC_EDM                  1
Name: count, dtype: int64
-----------
Missing values count ---> 36
==============================================================
```

- data is skwed, we are going to replace these labels (Facebook, bing, Click2call, Live Chat,Press_Release, Social Media, testone, WeLearn, blog, Pay per Click Ads, welearnblog_Home, youtubechannel, NC_EDM) in one label as 'Others'.
- we will deal with missing values by imputing missing values with max occuring label

[44]:
```
leads['Lead Source'] = leads['Lead Source'].replace(['Facebook', 'bing',␣
 ↪'Click2call', 'Live Chat','Press_Release', 'Social Media',
                          'testone', 'WeLearn', 'blog','Pay per Click Ads',␣
 ↪'welearnblog_Home',
                          'youtubechannel', 'NC_EDM','Welingak␣
 ↪Website','Referral Sites'],'Other')
leads['Lead Source'] = leads['Lead Source'].replace('google','Google')

print(leads['Lead Source'].value_counts())
print('-----------')
print('Missing values count --->', leads['Lead Source'].isna().sum())
print('==============================================================')
```

```
Lead Source
Google            2873
Direct Traffic    2543
Olark Chat        1755
Organic Search    1154
Reference          534
Other              345
Name: count, dtype: int64
-----------
Missing values count ---> 36
==============================================================
```

[45]:
```
# imputing missing values to max occuring label i.e. Google

leads['Lead Source'] = leads['Lead Source'].replace(np.NaN,'Google')

print(leads['Lead Source'].value_counts())
print('-----------')
print('Missing values count --->', leads['Lead Source'].isna().sum())
print('==============================================================')
```

```
Lead Source
Google              2909
Direct Traffic      2543
Olark Chat          1755
Organic Search      1154
Reference            534
Other                345
Name: count, dtype: int64
-----------
Missing values count ---> 0
==============================================================
```

[46]:
```
# 2. Dealing with Specialization

print(leads['Specialization'].value_counts())
print('-----------')
print('Missing values count --->', leads['Specialization'].isna().sum())
print('==============================================================')
```

```
Specialization
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: count, dtype: int64
-----------
Missing values count ---> 3380
==============================================================
```

- here we will create another category for missing values as the count is very high and imputing missing values with median can lead to misleading results

[48]:
```
# replacing missing values with label 'Missing'
```

```
leads['Specialization'] = leads['Specialization'].replace(np.NaN,'Missing')
leads['Specialization'].value_counts()
```

[48]:
```
Specialization
Missing                              3380
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: count, dtype: int64
```

[49]:
```
# 3. 'What is your current occupation'

print(leads['What is your current occupation'].value_counts())
print('-----------')
print('Missing values count --->', leads['What is your current occupation'].
  ↪isna().sum())
print('====================================================================')
```

```
What is your current occupation
Unemployed              5600
Working Professional     706
Student                  210
Other                     16
Housewife                 10
Businessman                8
Name: count, dtype: int64
-----------
Missing values count ---> 2690
====================================================================
```

- here also we will create another category for missing values as the count is very high and imputing missing values with median can lead to misleading results.

[51]:
```
# replacing missing values with label 'Missing'

leads['What is your current occupation'] = leads['What is your current␣
 ↪occupation'].replace(np.NaN,'Missing')
leads['What is your current occupation'].value_counts()
```

[51]:
```
What is your current occupation
Unemployed              5600
Missing                 2690
Working Professional     706
Student                  210
Other                     16
Housewife                 10
Businessman                8
Name: count, dtype: int64
```

[52]:
```
# 4. 'What matters most to you in choosing a course'

print(leads['What matters most to you in choosing a course'].value_counts())
print('-----------')
print('Missing values count --->', leads['What matters most to you in choosing␣
 ↪a course'].isna().sum())
print('====================================================================')
```

```
What matters most to you in choosing a course
Better Career Prospects      6528
Flexibility & Convenience       2
Other                           1
Name: count, dtype: int64
-----------
Missing values count ---> 2709
====================================================================
```

- the data is higly skewed for this column and the missing values are also high, it's better we drop the column.

[54]:
```
leads = leads.drop('What matters most to you in choosing a course', axis=1)
leads.head()
```

[54]:
```
              Lead Origin     Lead Source Do Not Email Do Not Call  Converted
TotalVisits  Total Time Spent on Website  Page Views Per Visit            Last
Activity           Specialization What is your current occupation Search
Newspaper Article X Education Forums Newspaper Digital Advertisement Through
Recommendations                                 Tags A free copy of Mastering
The Interview Last Notable Activity
0                      API      Olark Chat           No          No          0
0.0                          0                   0.0  Page Visited on Website
Missing                      Unemployed     No                No
```

```
No        No                    No                       No          Interested
in other courses                                    No              Modified
1                      API  Organic Search           No          No          0
5.0                     674                   2.5             Email Opened
Missing                 Unemployed     No                No
No        No                    No                       No
Ringing                                    No           Email Opened
2  Landing Page Submission  Direct Traffic           No          No          1
2.0                     1532                  2.0             Email Opened
Business Administration                         Student     No                No
No        No                    No                       No  Will revert after
reading the email                                   Yes          Email Opened
3  Landing Page Submission  Direct Traffic           No          No          0
1.0                     305                   1.0             Unreachable
Media and Advertising                      Unemployed     No                No
No        No                    No                       No
Ringing                                     No              Modified
4  Landing Page Submission          Google           No          No          1
2.0                     1428                  1.0        Converted to Lead
Missing                 Unemployed     No                No
No        No                    No                       No  Will revert after
reading the email                                    No              Modified
```

[55]:
```
# Dealing with  5. 'TotalVisits', 6. 'Page Views Per Visit', 7. 'Last␣
 ↪Activity', 8. 'Tags'

# checking value counts and missing values count for all the columns

miss_max = ['TotalVisits','Page Views Per Visit','Last Activity','Tags'] #␣
 ↪assigning them in a list

for i in leads[miss_max].columns:
    print(leads[i].value_counts())
    print('-----------')
    print('Missing values count --->', leads[i].isna().sum())
    print('======================================================================')
```

```
TotalVisits
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
21.0        6
23.0        6
24.0        5
25.0        5
27.0        5
22.0        3
29.0        2
28.0        2
26.0        2
141.0       1
55.0        1
30.0        1
43.0        1
74.0        1
41.0        1
54.0        1
115.0       1
251.0       1
32.0        1
42.0        1
Name: count, dtype: int64
-----------
Missing values count ---> 137
================================================================
Page Views Per Visit
0.00     2189
2.00     1795
3.00     1196
4.00      896
1.00      651
5.00      517
1.50      306
6.00      244
2.50      241
7.00      133
3.50       94
```

8.00	86
1.33	66
1.67	60
2.33	59
2.67	54
9.00	45
4.50	43
1.75	28
3.33	27
10.00	25
1.25	23
5.50	21
2.25	19
11.00	18
3.67	16
6.50	13
1.80	13
2.75	12
1.40	11
2.80	9
14.00	9
4.33	9
2.20	9
2.17	8
1.60	8
3.25	8
2.40	6
13.00	6
12.00	5
1.20	5
1.83	4
3.40	4
2.60	4
1.43	4
15.00	4
1.71	4
1.78	3
4.25	3
16.00	3
4.75	3
5.67	3
1.57	3
1.38	3
3.60	2
1.23	2
1.56	2
5.40	2
2.22	2

7.50	2
1.14	2
5.25	2
2.09	2
3.20	2
3.75	2
5.33	2
2.83	2
2.71	2
1.22	2
2.13	1
1.54	1
6.67	1
24.00	1
2.14	1
2.45	1
3.29	1
1.48	1
3.82	1
4.17	1
1.63	1
3.38	1
1.17	1
14.50	1
3.80	1
1.19	1
3.17	1
1.93	1
11.50	1
8.33	1
55.00	1
4.40	1
8.21	1
8.50	1
2.63	1
1.27	1
2.57	1
2.86	1
3.91	1
6.71	1
3.57	1
1.31	1
2.90	1
3.83	1
1.45	1
2.38	1
1.86	1
2.29	1

```
1.21        1
12.33       1
3.43        1
2.56        1
6.33        1
1.64        1
2.08        1
Name: count, dtype: int64
-----------
Missing values count ---> 137
================================================================
Last Activity
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
Visited Booth in Tradeshow         1
Resubscribed to emails             1
Name: count, dtype: int64
-----------
Missing values count ---> 103
================================================================
Tags
Will revert after reading the email                2072
Ringing                                             1203
Interested in other courses                          513
Already a student                                    465
Closed by Horizzon                                   358
switched off                                         240
Busy                                                 186
Lost to EINS                                         175
Not doing further education                          145
Interested  in full time MBA                         117
Graduation in progress                               111
invalid number                                        83
Diploma holder (Not Eligible)                         63
wrong number given                                    47
```

```
opp hangup                                                    33
number not provided                                           27
in touch with EINS                                            12
Lost to Others                                                 7
Still Thinking                                                 6
Want to take admission but has financial problems              6
In confusion whether part time or DLP                          5
Interested in Next batch                                       5
Lateral student                                                3
Shall take in the next coming month                            2
University not recognized                                      2
Recognition issue (DEC approval)                               1
Name: count, dtype: int64
-----------
Missing values count ---> 3353
==================================================================
```

we will impute all the missing values with label having max occurences

[57]:
```
# imputing all the missing values with label having max occurences

for i in leads[miss_max].columns:
    max_str = leads[i].value_counts()[leads[i].value_counts() == leads[i].
 ↪value_counts().max()].index[0]
    leads[i] = leads[i].fillna(value=max_str)
    print(leads[i].value_counts())
    print('-----------')
    print('Missing values count --->', leads[i].isna().sum())
    print('==================================================================')
```

```
TotalVisits
0.0      2326
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
21.0        6
23.0        6
24.0        5
25.0        5
27.0        5
22.0        3
29.0        2
28.0        2
26.0        2
141.0       1
55.0        1
30.0        1
43.0        1
74.0        1
41.0        1
54.0        1
115.0       1
251.0       1
32.0        1
42.0        1
Name: count, dtype: int64
-----------
Missing values count ---> 0
================================================================
Page Views Per Visit
0.00     2326
2.00     1795
3.00     1196
4.00      896
1.00      651
5.00      517
1.50      306
6.00      244
2.50      241
7.00      133
3.50       94
8.00       86
1.33       66
1.67       60
2.33       59
2.67       54
9.00       45
4.50       43
1.75       28
```

3.33	27
10.00	25
1.25	23
5.50	21
2.25	19
11.00	18
3.67	16
6.50	13
1.80	13
2.75	12
1.40	11
2.80	9
14.00	9
4.33	9
2.20	9
2.17	8
1.60	8
3.25	8
2.40	6
13.00	6
12.00	5
1.20	5
1.83	4
3.40	4
2.60	4
1.43	4
15.00	4
1.71	4
1.78	3
4.25	3
16.00	3
4.75	3
5.67	3
1.57	3
1.38	3
3.60	2
1.23	2
1.56	2
5.40	2
2.22	2
7.50	2
1.14	2
5.25	2
2.09	2
3.20	2
3.75	2
5.33	2
2.83	2

```
2.71        2
1.22        2
2.13        1
1.54        1
6.67        1
24.00       1
2.14        1
2.45        1
3.29        1
1.48        1
3.82        1
4.17        1
1.63        1
3.38        1
1.17        1
14.50       1
3.80        1
1.19        1
3.17        1
1.93        1
11.50       1
8.33        1
55.00       1
4.40        1
8.21        1
8.50        1
2.63        1
1.27        1
2.57        1
2.86        1
3.91        1
6.71        1
3.57        1
1.31        1
2.90        1
3.83        1
1.45        1
2.38        1
1.86        1
2.29        1
1.21        1
12.33       1
3.43        1
2.56        1
6.33        1
1.64        1
2.08        1
Name: count, dtype: int64
```

```
-----------
Missing values count ---> 0
================================================================
Last Activity
Email Opened                    3540
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
Visited Booth in Tradeshow         1
Resubscribed to emails             1
Name: count, dtype: int64
-----------
Missing values count ---> 0
================================================================
Tags
Will revert after reading the email                5425
Ringing                                            1203
Interested in other courses                         513
Already a student                                   465
Closed by Horizzon                                  358
switched off                                        240
Busy                                                186
Lost to EINS                                        175
Not doing further education                         145
Interested  in full time MBA                        117
Graduation in progress                              111
invalid number                                       83
Diploma holder (Not Eligible)                        63
wrong number given                                   47
opp hangup                                           33
number not provided                                  27
in touch with EINS                                   12
Lost to Others                                        7
Still Thinking                                        6
Want to take admission but has financial problems     6
In confusion whether part time or DLP                 5
Interested in Next batch                              5
```

```
Lateral student                                    3
Shall take in the next coming month                2
University not recognized                          2
Recognition issue (DEC approval)                   1
Name: count, dtype: int64
-----------
Missing values count ---> 0
==============================================================
```

5.0.1 Checking missing values in rows

[59]:
```
# checking rows having missing values more than 40%

missing_row_count = leads.apply(lambda x: round(100*(sum(x.isnull().values)/
  ↪len(leads.index)),2), axis = 1)
missing_row_count[missing_row_count > 40]
```

[59]:
```
Series([], dtype: float64)
```

there are no row with missing values more than 40%

[61]:
```
# checking if any more null values in any columns

leads.isna().sum()
```

[61]:
```
Lead Origin                              0
Lead Source                              0
Do Not Email                             0
Do Not Call                              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
Search                                   0
Newspaper Article                        0
X Education Forums                       0
Newspaper                                0
Digital Advertisement                    0
Through Recommendations                  0
Tags                                     0
A free copy of Mastering The Interview   0
Last Notable Activity                    0
dtype: int64
```

Now all the data we have is clear of missing values and cleaned up. Now we will proceed with data

transformation for some columns having yes/no labels and will convert columns with numerical data to categorical data.

---

6 3. Data Transformation

1. Converting yes/no category columns to binary form 1/0.
2. to deal with columns having outliers will create bins for them.
3. will remove all the redundant and repeated columns.
4. create dummy variables

[64]:
```
#checking data set information to check the columns yes/no labels
leads.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 9240 entries, 0 to 9239
Data columns (total 20 columns):
 #   Column                                  Non-Null Count  Dtype
---  ------                                  --------------  -----
 0   Lead Origin                             9240 non-null   object
 1   Lead Source                             9240 non-null   object
 2   Do Not Email                            9240 non-null   object
 3   Do Not Call                             9240 non-null   object
 4   Converted                               9240 non-null   int64
 5   TotalVisits                             9240 non-null   float64
 6   Total Time Spent on Website             9240 non-null   int64
 7   Page Views Per Visit                    9240 non-null   float64
 8   Last Activity                           9240 non-null   object
 9   Specialization                          9240 non-null   object
 10  What is your current occupation         9240 non-null   object
 11  Search                                  9240 non-null   object
 12  Newspaper Article                       9240 non-null   object
 13  X Education Forums                      9240 non-null   object
 14  Newspaper                               9240 non-null   object
 15  Digital Advertisement                   9240 non-null   object
 16  Through Recommendations                 9240 non-null   object
 17  Tags                                    9240 non-null   object
 18  A free copy of Mastering The Interview  9240 non-null   object
 19  Last Notable Activity                   9240 non-null   object
dtypes: float64(2), int64(2), object(16)
memory usage: 1.4+ MB
```

[65]:
```
# creating a variable and storing the columns names for run in a loop

yes_no = ['Do Not Email','Do Not Call','Search','Newspaper Article','X␣
 ↪Education Forums',
```

```
'Newspaper','Digital Advertisement','Through Recommendations','A free copy of␣
 ↪Mastering The Interview']


# creating dictionary for two categories where; Yes : 1 , No : 0

category={"No":0,"Yes":1}

for i in yes_no:
    leads[i]=leads[i].map(category)


leads.head()
```

[65]:

```
              Lead Origin     Lead Source  Do Not Email  Do Not Call  Converted
TotalVisits  Total Time Spent on Website  Page Views Per Visit            Last
Activity           Specialization What is your current occupation  Search
Newspaper Article  X Education Forums  Newspaper  Digital Advertisement  Through
Recommendations                                  Tags  A free copy of Mastering
The Interview Last Notable Activity
0                      API      Olark Chat             0            0          0
0.0                            0                   0.0  Page Visited on Website
Missing                      Unemployed       0                  0
0          0                       0                          0          Interested
in other courses                                          0              Modified
1                      API  Organic Search             0            0          0
5.0                           674                   2.5             Email Opened
Missing                      Unemployed       0                  0
0          0                       0                          0
Ringing                                          0          Email Opened
2  Landing Page Submission  Direct Traffic             0            0          1
2.0                          1532                  2.0             Email Opened
Business Administration                        Student       0
0                    0          0                       0                        0
Will revert after reading the email                                          1
Email Opened
3  Landing Page Submission  Direct Traffic             0            0          0
1.0                           305                  1.0             Unreachable
Media and Advertising                     Unemployed       0                  0
0          0                       0                          0
Ringing                                          0              Modified
4  Landing Page Submission          Google             0            0          1
2.0                          1428                  1.0        Converted to Lead
Missing                      Unemployed       0                  0
0          0                       0                          0  Will revert after
reading the email                                      0              Modified
```

Checking for Outliers

[67]:
```
# checking the statistical data
leads.describe(percentiles=[0.25,0.50,0.75,0.90,0.95,0.97,0.99])
```

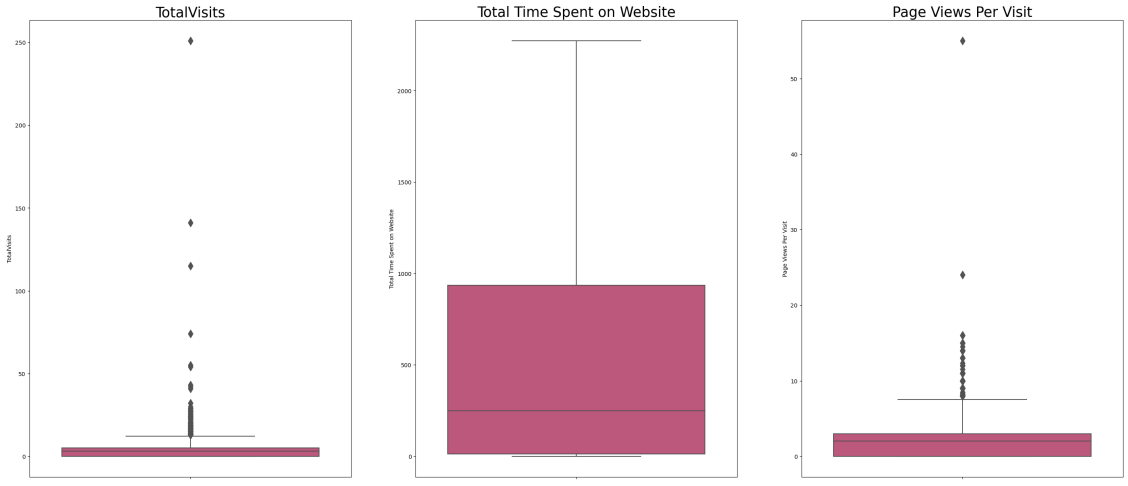
[67]:
```
       Do Not Email  Do Not Call    Converted  TotalVisits  Total Time Spent on
Website  Page Views Per Visit       Search  Newspaper Article  X Education
Forums    Newspaper  Digital Advertisement  Through Recommendations  A free copy
of Mastering The Interview
count   9240.000000  9240.000000  9240.000000  9240.000000
9240.000000           9240.000000  9240.000000        9240.000000
9240.000000  9240.000000            9240.000000             9240.000000
9240.000000
mean       0.079437     0.000216     0.385390     3.394156
487.698268             2.327787     0.001515          0.000216
0.000108     0.000108               0.000433                0.000758
0.312554
std        0.270435     0.014711     0.486714     4.836682
548.021466             2.164258     0.038898          0.014711
0.010403     0.010403               0.020803                0.027515
0.463559
min        0.000000     0.000000     0.000000     0.000000
0.000000              0.000000     0.000000          0.000000
0.000000     0.000000               0.000000                0.000000
0.000000
25%        0.000000     0.000000     0.000000     0.000000
12.000000             0.000000     0.000000          0.000000
0.000000     0.000000               0.000000                0.000000
0.000000
50%        0.000000     0.000000     0.000000     3.000000
248.000000             2.000000     0.000000          0.000000
0.000000     0.000000               0.000000                0.000000
0.000000
75%        0.000000     0.000000     1.000000     5.000000
936.000000             3.000000     0.000000          0.000000
0.000000     0.000000               0.000000                0.000000
1.000000
90%        0.000000     0.000000     1.000000     7.000000
1380.000000             5.000000     0.000000          0.000000
0.000000     0.000000               0.000000                0.000000
1.000000
95%        1.000000     0.000000     1.000000    10.000000
1562.000000             6.000000     0.000000          0.000000
0.000000     0.000000               0.000000                0.000000
1.000000
97%        1.000000     0.000000     1.000000    11.000000
1660.000000             7.000000     0.000000          0.000000
```

```
0.000000      0.000000                0.000000                  0.000000
1.000000
99%         1.000000      0.000000      1.000000     17.000000
1840.610000                9.000000      0.000000             0.000000
0.000000      0.000000                0.000000                  0.000000
1.000000
max         1.000000      1.000000      1.000000    251.000000
2272.000000               55.000000      1.000000             1.000000
1.000000      1.000000                1.000000                  1.000000
1.000000
```

As we can see there are outliers in 2 variables 'TotalVisits' and 'Page Views Per Visit'.

Let's visualize the outliers using boxplot to understand the outliers.

[69]:
```
# setting figure size
plt.figure(figsize=(35,50))

# Title names for the columns in the dataset

col_title={0:'TotalVisits',1:'Total Time Spent on Website',2:'Page Views Per␣
 ↪Visit'}

# Visualising the outliers with boxplot for all the variables

for i in range(3):
    plt.subplot(3,3,i+1)
    plt.title(col_title[i],fontsize=25)
    sns.boxplot(y=leads[col_title[i]],data=leads,palette='plasma',fliersize=10)
```



From the above boxplots we can now confirm that we have two outlier variables in our dataset ('TotalVisits' and 'Page Views Per Visit'). Now as per business requirement we cannot drop these

outliers because it may impact our analysis/model so we will create bins for these two outliers.

Creating Bins

[72]:
```
# 1. dealing with 'TotalVisits' variable
# As we have range from 0 to 251 we will create buckets as per need

# creating labels
TotalVisits_labels = ['TotalVisits_0',␣
 ↪'TotalVisits_1_2','TotalVisits_3_4','TotalVisits_5_6','TotalVisits_7_8','TotalVisits_9_10',
       'TotalVisits_11_12','TotalVisits_12_15','TotalVisits_above_15']

# creating bins for TotalVisits__labels
leads['TotalVisits'] = pd.cut(leads['TotalVisits'], bins=[-1,0.
 ↪0,2,4,6,8,10,12,15,251], labels = TotalVisits_labels)
```

[73]:
```
leads['TotalVisits'].value_counts()
```

[73]:
```
TotalVisits
TotalVisits_3_4         2426
TotalVisits_0           2326
TotalVisits_1_2         2075
TotalVisits_5_6         1249
TotalVisits_7_8          533
TotalVisits_9_10         278
TotalVisits_11_12        131
TotalVisits_above_15     120
TotalVisits_12_15        102
Name: count, dtype: int64
```

[74]:
```
# 2. dealing with 'Page Views Per Visit' variable
# As we have range from 0 to 55 we will create buckets as per need

# creating labels
pvpv_labels=['Page_Views_Per_Visit_0','Page_Views_Per_Visit_1_2','Page_Views_Per_Visit_3_4',
             ␣
 ↪'Page_Views_Per_Visit_5_6','Page_Views_Per_Visit_7_8','Page_Views_Per_Visit_9_10','Page_Vie

# creating bins for 'Page Views Per Visit'
leads['Page Views Per Visit'] = pd.cut(leads['Page Views Per Visit'],␣
 ↪bins=[-1,0,2,4,6,8,10,60], labels = pvpv_labels)
```

[75]:
```
leads['Page Views Per Visit'].value_counts()
```

[75]:
```
Page Views Per Visit
Page_Views_Per_Visit_1_2         3007
Page_Views_Per_Visit_3_4         2696
Page_Views_Per_Visit_0           2326
```

```
Page_Views_Per_Visit_5_6          851
Page_Views_Per_Visit_7_8          237
Page_Views_Per_Visit_9_10          73
Page_Views_Per_Visit_above_10      50
Name: count, dtype: int64
```

[76]:
```
# checking data set after creating bins

leads.head()
```

[76]:
```
              Lead Origin     Lead Source  Do Not Email  Do Not Call  Converted
TotalVisits  Total Time Spent on Website      Page Views Per Visit
Last Activity           Specialization What is your current occupation  Search
Newspaper Article  X Education Forums  Newspaper  Digital Advertisement  Through
Recommendations                                  Tags  A free copy of Mastering
The Interview Last Notable Activity
0                     API      Olark Chat             0            0          0
TotalVisits_0                            0    Page_Views_Per_Visit_0  Page
Visited on Website                   Missing                       Unemployed
0                  0                   0          0                        0
0          Interested in other courses                                      0
Modified
1                     API  Organic Search             0            0          0
TotalVisits_5_6                        674  Page_Views_Per_Visit_3_4
Email Opened                  Missing                        Unemployed       0
0                  0          0                       0                        0
Ringing                                          0          Email Opened
2  Landing Page Submission  Direct Traffic             0            0          1
TotalVisits_1_2                       1532  Page_Views_Per_Visit_1_2
Email Opened  Business Administration                          Student       0
0                  0          0                       0                        0
Will revert after reading the email                                          1
Email Opened
3  Landing Page Submission  Direct Traffic             0            0          0
TotalVisits_1_2                        305  Page_Views_Per_Visit_1_2
Unreachable    Media and Advertising                      Unemployed       0
0                  0          0                       0                        0
Ringing                                          0              Modified
4  Landing Page Submission          Google             0            0          1
TotalVisits_1_2                       1428  Page_Views_Per_Visit_1_2
Converted to Lead                  Missing                       Unemployed
0                  0                   0          0                        0
0  Will revert after reading the email                                       0
Modified
```

After creating bins we removed the outliers and are now good to go. Before creating the dummy variables let's remove redundant columns/varaibles.

Also from above we know columns : 'Last Activity', 'Tags', 'Last Notable Activity' activity columns came from sales team, thus we will drop these redundant columns.

[78]:
```
# dropping redundant column

redundant=['Last Activity', 'Tags', 'Last Notable Activity']

leads=leads.drop(redundant,axis=1)

leads.head()
```

[78]:
```
              Lead Origin     Lead Source  Do Not Email  Do Not Call  Converted
TotalVisits  Total Time Spent on Website      Page Views Per Visit
Specialization What is your current occupation  Search  Newspaper Article  X
Education Forums  Newspaper  Digital Advertisement  Through Recommendations  A
free copy of Mastering The Interview
0                       API      Olark Chat             0            0          0
TotalVisits_0                              0    Page_Views_Per_Visit_0
Missing                       Unemployed       0                   0
0          0                       0                          0
0
1                       API  Organic Search             0            0          0
TotalVisits_5_6                          674  Page_Views_Per_Visit_3_4
Missing                       Unemployed       0                   0
0          0                       0                          0
0
2  Landing Page Submission  Direct Traffic             0            0          1
TotalVisits_1_2                          1532  Page_Views_Per_Visit_1_2  Business
Administration                          Student       0                   0
0          0                       0                          0
1
3  Landing Page Submission  Direct Traffic             0            0          0
TotalVisits_1_2                          305  Page_Views_Per_Visit_1_2    Media
and Advertising                       Unemployed       0                   0
0          0                       0                          0
0
4  Landing Page Submission          Google             0            0          1
TotalVisits_1_2                             1428  Page_Views_Per_Visit_1_2
Missing                       Unemployed       0                   0
0          0                       0                          0
0
```

[79]:
```
leads.shape
```

[79]:
```
(9240, 17)
```

Next, we will create dummy variables for mutiple levels of categories

Creating Dummy Variables

[81]:
```
#Creating a dummy variables for 4 categories and dropping the first level.

cat = ['Lead Origin', 'Lead Source','Specialization','What is your current␣
 ↪occupation','TotalVisits','Page Views Per Visit']

#creating dummy variables data set
dummy = pd.get_dummies(leads[cat], drop_first=True)

# Adding these dummies to our original dataset
leads = pd.concat([leads,dummy],axis=1)

#dropping the duplicate columns
leads = leads.drop(cat, axis=1)

#viewing the dataset
leads.head()
```

[81]:
```
   Do Not Email  Do Not Call  Converted  Total Time Spent on Website  Search
Newspaper Article  X Education Forums  Newspaper  Digital Advertisement  Through
Recommendations  A free copy of Mastering The Interview  Lead Origin_Landing
Page Submission  Lead Origin_Lead Add Form  Lead Origin_Lead Import  Lead
Origin_Quick Add Form  Lead Source_Google  Lead Source_Olark Chat  Lead
Source_Organic Search  Lead Source_Other  Lead Source_Reference
Specialization_Business Administration  Specialization_E-Business
Specialization_E-COMMERCE  Specialization_Finance Management
Specialization_Healthcare Management  Specialization_Hospitality Management
Specialization_Human Resource Management  Specialization_IT Projects Management
Specialization_International Business  Specialization_Marketing Management
Specialization_Media and Advertising  Specialization_Missing
Specialization_Operations Management  Specialization_Retail Management
Specialization_Rural and Agribusiness  Specialization_Services Excellence
Specialization_Supply Chain Management  Specialization_Travel and Tourism  What
is your current occupation_Housewife  What is your current occupation_Missing
What is your current occupation_Other  What is your current occupation_Student
What is your current occupation_Unemployed  What is your current
occupation_Working Professional  TotalVisits_TotalVisits_1_2
TotalVisits_TotalVisits_3_4  TotalVisits_TotalVisits_5_6
TotalVisits_TotalVisits_7_8  TotalVisits_TotalVisits_9_10
TotalVisits_TotalVisits_11_12  TotalVisits_TotalVisits_12_15
TotalVisits_TotalVisits_above_15  Page Views Per Visit_Page_Views_Per_Visit_1_2
Page Views Per Visit_Page_Views_Per_Visit_3_4  Page Views Per
Visit_Page_Views_Per_Visit_5_6  Page Views Per Visit_Page_Views_Per_Visit_7_8
Page Views Per Visit_Page_Views_Per_Visit_9_10  Page Views Per
Visit_Page_Views_Per_Visit_above_10
0             0            0          0                              0       0
```

0 0 0 0 0
0 False False
False False False True
False False False
False False False
False False
False False
False False
False False True
False False False
False False False
False False
False False
True False
False False False
False False False
False False
False False
False False
False False
1 0 0 0 674 0
0 0 0 0 0
0 False False
False False False False
True False False
False False False
False False
False False
False False
False False True
False False False
False False False
False False
False False
True False
False False True
False False False
False False
False True
False False
False False
2 0 0 1 1532 0
0 0 0 0 0
1 True False
False False False False
False False False
True False False

False False
False False
False False
False False False
False False False
False False False
False False
False True
False False
True False False
False False False
False False
True False
False False
False False
3 0 0 0 305 0
0 0 0 0 0
0 True False
False False False False
False False False
False False False
False False
False False
False False
False True False
False False False
False False False
False False
False False
True False
True False False
False False False
False False
True False
False False
False False
4 0 0 1 1428 0
0 0 0 0 0
0 True False
False False True False
False False False
False False False
False False
False False
False False
False False True
False False False

```
False                                          False                                          False
False                                           False
False                                           False
True                                                    False
True                                 False                          False
False                                  False                              False
False                                      False
True                                                 False
False                                                 False
False                                                     False
```

[82]:
```
#checking statistical data
leads.describe()
```

[82]:
```
       Do Not Email  Do Not Call    Converted  Total Time Spent on Website
Search  Newspaper Article  X Education Forums    Newspaper  Digital
Advertisement  Through Recommendations  A free copy of Mastering The Interview
count   9240.000000  9240.000000  9240.000000                   9240.000000
9240.000000         9240.000000          9240.000000  9240.000000
9240.000000               9240.000000                               9240.000000
mean       0.079437     0.000216     0.385390                    487.698268
0.001515           0.000216            0.000108     0.000108
0.000433                  0.000758                                0.312554
std        0.270435     0.014711     0.486714                    548.021466
0.038898           0.014711            0.010403     0.010403
0.020803                  0.027515                                0.463559
min        0.000000     0.000000     0.000000                     0.000000
0.000000           0.000000            0.000000     0.000000
0.000000                  0.000000                                0.000000
25%        0.000000     0.000000     0.000000                    12.000000
0.000000           0.000000            0.000000     0.000000
0.000000                  0.000000                                0.000000
50%        0.000000     0.000000     0.000000                   248.000000
0.000000           0.000000            0.000000     0.000000
0.000000                  0.000000                                0.000000
75%        0.000000     0.000000     1.000000                   936.000000
0.000000           0.000000            0.000000     0.000000
0.000000                  0.000000                                1.000000
max        1.000000     1.000000     1.000000                  2272.000000
1.000000           1.000000            1.000000     1.000000
1.000000                  1.000000                                1.000000
```

[83]:
```
#checking shape of the data set
leads.shape
```

[83]:
```
(9240, 58)
```

from above tables we now see that all columns are converted to numerical data

---

7 4. Data Preparation

1. Split the dataset into train and test dataset and scaled the datasets.
2. After this, we plot a heatmap to check the correlations among the variables.
3. check heatmap for highly correlated features

train_test split

[87]:
```
# Importing train-test-split method from sklearn - model selection

from sklearn.model_selection import train_test_split
```

[88]:
```
# putting feature varibales in "X" and target variable in "y"

y=leads['Converted']
X=leads.drop('Converted',1)

y.head()
```

```
---------------------------------------------------------------------------
TypeError                                 Traceback (most recent call last)
Cell In[88], line 4
      1 # putting feature varibales in "X" and target variable in "y"
      3 y=leads['Converted']
----> 4 X=leads.drop('Converted',1)
      6 y.head()

TypeError: DataFrame.drop() takes from 1 to 2 positional arguments but 3 were␣
 ↪given
```

[ ]:
```
X.head()
```

[ ]:
```
# Splitting the datset into train and test dataset

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

[ ]:
```
#looking all the X, y train and test sets
print('X_train:',X_train.shape)
print('X_test:',X_test.shape)
print('y_train:',y_train.shape)
print('y_test:',y_test.shape)
```

Feature Standardization

[ ]:
```
# Importing Standard Scaler method from sklearn - preprocessing library

from sklearn.preprocessing import StandardScaler

scaler=StandardScaler()  # Creating an object
```

[ ]:
```
# Now, Scaling  'Total Time Spent on Website' variables with standard scaler␣
 ↪and fitting on X_train dataset

X_train[['Total Time Spent on Website']]=scaler.fit_transform(X_train[['Total␣
 ↪Time Spent on Website']])
X_train.describe()
```

[ ]:
```
## Checking the conversion rate from 'converted' column as it denotes the␣
 ↪target variable

print('Current Conversion Rate:',round((sum(y)/len(y.index))*100,2))
```

Correlation in the dataset

[ ]:
```
# setting the figure size
plt.figure(figsize=(55,35))

# Plotting a heatmap

sns.heatmap(leads.corr(method='spearman'))
plt.title('Correlations', fontsize =35)
plt.yticks(fontsize=30)
plt.xticks(fontsize=30)
plt.show()
```

Correlation is shown above by heatmap, from above we couldn't find much which features are highly correlated and to drop thus we will now proceed with building our model and based on the p-values and VIFs, we will again check for correlation.

8 5. Building a Model

[ ]:
```
# importing statmodels library for statistical summary and model creation

import statsmodels.api as sm
```

We are going to use hybrid model creation using RFE and manual features selection

8.0.1 feature selection using RFE

[ ]:
```
# Importing RFE and logistic regression libraries from scikit learn

from sklearn.linear_model import LogisticRegression
from sklearn.feature_selection import RFE

# creating an object

logreg = LogisticRegression()
```

[ ]:
```
# RFE model with 15 variables

rfe = RFE(logreg,15)

# fitting the model

rfe = rfe.fit(X_train,y_train)
```

[ ]:
```
#listing which all columns are selected(True) by RFE and which all are␣
 ↪rejected(False)

list(zip(X_train.columns, rfe.support_, rfe.ranking_))
```

[ ]:
```
# storing selected(True) columns by RFE in a list
rfe_col = X_train.columns[rfe.support_]

# listing features removed by RFE feature selection
X_train.columns[~rfe.support_]
```

[ ]:
```
# Creating new train dataframe with RFE selected features

X_train_rfe = X_train[rfe_col]
X_train_rfe.head()
```

Model 1

[ ]:
```
# creating 1st model

# Adding a constant
X_train_1=sm.add_constant(X_train_rfe)

# creating a model and fitting it.

logr1=sm.GLM(y_train,X_train_1,family=sm.families.Binomial()).fit()    # Using␣
 ↪GLM for creating model and fitting it
logr1.summary()                                                        #viewing␣
 ↪summary of the model
```

Now, From the above summary presented there are some features having high p -values, we will drop features having insignificant values one by one and create new models until all the features attain significant p-value<0.05 and vif-values < 4.

Calculating VIF

[ ]:
```
# importing VIFs library

from statsmodels.stats.outliers_influence import variance_inflation_factor

# Creating vif dataframe

vif=pd.DataFrame()

# adding same features as the x_train dataset have

vif['Features']=X_train_rfe[rfe_col].columns

# Caculating VIFs

vif['VIF']=[variance_inflation_factor(X_train_rfe[rfe_col].values,i) for i in␣
 ↪range(X_train_rfe[rfe_col].shape[1])]

# Rounding the vif values

vif['VIF']=round(vif['VIF'],2)

# Sorting the vif values

vif=vif.sort_values(by='VIF',ascending=False)
vif   # Viewing the dataset
```

Model 2

- for this we are dropping 'const','What is your current occupation_Housewife' due to high p-value

[ ]:
```
# Dropping the most insignificant values 'What is your current␣
 ↪occupation_Housewife' and constant

X_train_rfe2 = X_train_1.drop(['const','What is your current␣
 ↪occupation_Housewife'],1)

# Creating a new model 2

X_train_2=sm.add_constant(X_train_rfe2)                                    #␣
 ↪Adding constant
```

```
logr2=sm.GLM(y_train,X_train_2,family=sm.families.Binomial()).fit()     # Using␣
 ↪GLM for creating model and fitting it
logr2.summary()                                                          ␣
 ↪#viewing summary of the model
```

[ ]:
```
# Check for the VIF values of the feature variables.

# Create a dataframe that will contain the names of all the feature variables␣
 ↪and their respective VIFs
vif=pd.DataFrame()
vif['Features']=X_train_rfe2.columns
vif['VIF']=[variance_inflation_factor(X_train_rfe2.values,i) for i in␣
 ↪range(X_train_rfe2.shape[1])]
vif['VIF']=round(vif['VIF'],2)
vif=vif.sort_values(by='VIF',ascending=False)
vif
```

Model 3

- for this we are dropping 'const','Specialization_Missing' as this is has no information.

[ ]:
```
# Dropping the most insignificant values 'Specialization_Missing' and constant

X_train_rfe3 = X_train_2.drop(['const','Specialization_Missing'],1)

# Creating a new model 3

X_train_3=sm.add_constant(X_train_rfe3)                                  #␣
 ↪Adding constant
logr3=sm.GLM(y_train,X_train_3,family=sm.families.Binomial()).fit()     # Using␣
 ↪GLM for creating model and fitting it
logr3.summary()                                                          ␣
 ↪#viewing summary of the model
```

[ ]:
```
# Check for the VIF values of the feature variables.

# Create a dataframe that will contain the names of all the feature variables␣
 ↪and their respective VIFs
vif=pd.DataFrame()
vif['Features']=X_train_rfe3.columns
vif['VIF']=[variance_inflation_factor(X_train_rfe3.values,i) for i in␣
 ↪range(X_train_rfe3.shape[1])]
vif['VIF']=round(vif['VIF'],2)
vif=vif.sort_values(by='VIF',ascending=False)
vif
```

[ ]:
```
# checking all the coefficients
logr2.params.sort_values(ascending=False)
```

Now we have good amount of features having significant p-values and VIF-values<4. We will consider model 3 as our final model

8.0.2 Predicting the train dataset with our final model

[ ]:
```
#predicting train dataset with final model

y_train_pred=logr3.predict(X_train_3)

# Creating a new dataset and saving predicted values in it

y_train_pred_final=pd.DataFrame({'Converted':y_train.
 ↪values,'Converted_probability':y_train_pred,'ID':y_train.index})

y_train_pred_final.head()   # viewing first 5 rows
```

8.0.3 ROC Curve Plotting

- ROC curve shows the trade off between True postive rate and False positive rate - means if sensitivity increases specificity will decrease.
- The curve closer to the left side border then right side of the border is more accurate.
- The curve closer to the 45-degree diagonal of the ROC space is less accurate.

[ ]:
```
# Importing necessary libraries for roc curve

from sklearn.metrics import roc_curve
from sklearn.metrics import roc_auc_score

# Creating a function to plot roc curve with auc score

def draw_roc(actual, probability):

    # Creating roc curve to get true positive rate, false positive rate and␣
 ↪threshold

    fpr, tpr, thresholds = roc_curve( actual, probability,drop_intermediate =␣
 ↪False )

    # Calculating the auc score(area under the curve)

    auc_score = roc_auc_score( actual, probability )

    # Setting the figure size
```

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

    # Plotting the roc curve

    plt.plot( fpr, tpr, label='ROC curve (area = %0.2f)' % auc_score )

    # Plotting the 45% dotted line
    plt.plot([0, 1], [0, 1], 'r--')

    # Setting the x axis limit

    plt.xlim([0.0, 1.0])

    # Setting the y axis limit

    plt.ylim([0.0, 1.05])

    # Setting the x axis label
    plt.xlabel('False Positive Rate')

    # Setting the y axis label

    plt.ylabel('True Positive Rate')

    # Setting the title

    plt.title('Receiver operating characteristic')

    # Setting the legend on the left below to show the value of auc

    plt.legend(loc="lower right")

    # Showing the plot

    plt.show()

    return None   # no return
```

ROC CURVE

[ ]:
```
# Calling the roc curve function for plotting

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

[ ]:
```
# creating 10 points of probabilities to find the optimal point cutoff
```

```
numbers=[float(x)/10 for x in range(10)]  # from 0 to 0.9 with set size 0.1

for i in numbers:
    y_train_pred_final[i]=y_train_pred_final['Converted_probability'].
 ↪map(lambda x:1 if x > i else 0)   #  Mapping the probablities for each 10␣
 ↪points
y_train_pred_final.head()   # Viewing the first 5 rows
```

Accurary, Sensitivity, Specificity

[ ]:
```
# Calculating accuracy, sensitivity and specificity with probability cutoffs

# importing necessary library

from sklearn.metrics import confusion_matrix

# Creating a dataframe to store all the values to be created

df_cutoffs=pd.
 ↪DataFrame(columns=['Probability','Accuracy','Sensitvity','Specificity'])

# from 0 to 0.9 with set size 0.1

var=[0.0,0.1,0.2,0.3,0.4,0.5,0.6,0.7,0.8,0.9]

for i in var:
   ␣
 ↪cm_matrix=confusion_matrix(y_train_pred_final['Converted'],y_train_pred_final[i])␣
 ↪ # creating confusion matrix
    total=sum(sum(cm_matrix))                                                ␣
 ↪       # Taking the sum of the matrix
    accuracy=(cm_matrix[0,0]+cm_matrix[1,1])/total                           ␣
 ↪       # Storing Accuracy Data
    sensitivity=cm_matrix[1,1]/(cm_matrix[1,0]+cm_matrix[1,1])               ␣
 ↪       # Storing Sensitivity Data
    specificity=cm_matrix[0,0]/(cm_matrix[0,0]+cm_matrix[0,1])               ␣
 ↪       # Storing Specificity Data
    df_cutoffs.loc[i]=[i, accuracy, sensitivity, specificity]                ␣
 ↪       # Inserting all the data into the dataframe created earlier
print(df_cutoffs)   # Printing the data
```

PLotting Accuracy, Sensitivity and Specificity

[ ]:
```
# Ploting 'Accuracy' , 'Sensitivity' and 'Specificity' for various␣
 ↪probabilities(0.0 to 0.9).
```

```
df_cutoffs.plot.line(x='Probability',␣
 ↪y=['Accuracy','Sensitvity','Specificity'], figsize=(12,8))    # line plotting
plt.show()
```

[ ]:
```
# Predicting the outcomes with probability cutoff as 0.3 by creating new␣
 ↪columns in the final dataset

y_train_pred_final['Predicted']=y_train_pred_final['Converted_probability'].
 ↪map(lambda x:1 if x >0.3 else 0 )  # Predicted value

y_train_pred_final.head()
```

[ ]:
```
# Creating confusion matrix to find all the metrics

confusion_pr_train=confusion_matrix(y_train_pred_final.
 ↪Converted,y_train_pred_final.Predicted)
confusion_pr_train
```

[ ]:
```
#Sensitivity score
Sensitivity_train =round((confusion_pr_train[1,1]/
 ↪(confusion_pr_train[1,0]+confusion_pr_train[1,1])*100),2)

#specificity score
Specificity_train =round((confusion_pr_train[0,0]/
 ↪(confusion_pr_train[0,0]+confusion_pr_train[0,1])*100),2)

#print both
print('Sensitivity:',Sensitivity_train)
print('Specificity:',Specificity_train)
```

[ ]:
```
#  Pecision score
Precision_train = round((confusion_pr_train[1,1]/
 ↪(confusion_pr_train[0,1]+confusion_pr_train[1,1])*100),2)

#  Recall score
Recall_train = round((confusion_pr_train[1,1]/
 ↪(confusion_pr_train[1,0]+confusion_pr_train[1,1])*100),2)

#print both
print('Precision:',Precision_train)
print('Recall:',Recall_train)
```

[ ]:
```
# Checking accuracy for train dataset
from sklearn import metrics
```

```
Accuracy_train = round(metrics.accuracy_score(y_train_pred_final.
 ↪Converted,y_train_pred_final.Predicted)*100,2)
print('Train set Accuracy:',Accuracy_train)
```

[ ]:
```
# importing precision recall curve from sklearn library for train set

from sklearn.metrics import precision_recall_curve, f1_score

# Creating precision recall curve by creating three points and plotting

p ,r, thresholds=precision_recall_curve(y_train_pred_final.
 ↪Converted,y_train_pred_final.Converted_probability)
plt.title('Precision vs Recall tradeoff on Train set')
plt.plot(thresholds, p[:-1], "g-")    # Plotting precision
plt.plot(thresholds, r[:-1], "r-")    # Plotting Recall
plt.show()
```

---

9 Prediction on the test dataset

9.0.1 Scaling the test dataset

[ ]:
```
# Scalling the variables 'Total Time Spent on Website' with standard scaler and␣
 ↪tranforming the X_test dataset

X_test[['Total Time Spent on Website']]=scaler.transform(X_test[['Total Time␣
 ↪Spent on Website']])
```

[ ]:
```
# Predicting the test dataset with our final model

test_cols=X_train_3.columns[1:]              # Taking the same column train set␣
 ↪has
X_test_final=X_test[test_cols]               # Updating it in the final test set
X_test_final=sm.add_constant(X_test_final)   # Adding constant to the final set␣
 ↪set
y_pred_test=logr3.predict(X_test_final)      # Predicting the final test set
```

[ ]:
```
# Creating a new dataset and saving the prediction values in it

y_test_pred_final=pd.DataFrame({'Converted':y_test.
 ↪values,'Converted_Probability':y_pred_test,'ID':y_test.index})

y_test_pred_final.head()   # viewing first 5 rows
```

[ ]:
```
# Calling the roc curve function for plotting

draw_roc(y_test_pred_final.Converted, y_test_pred_final.Converted_Probability)
```

9.1 Model Evaluation

[ ]:
```
# Predicting the outcomes with probability cutoff as 0.3 by creating new␣
 ↪columns in the final test dataset

y_test_pred_final['Predicted']=y_test_pred_final['Converted_Probability'].
 ↪map(lambda x:1 if x >0.3 else 0 )  # Predicted value

y_test_pred_final.head()
```

[ ]:
```
# Creating confusion matrix to find precision and recall score

confusion_pr_test=confusion_matrix(y_test_pred_final.
 ↪Converted,y_test_pred_final.Predicted)
confusion_pr_test
```

[ ]:
```
#Sensitivity score
Sensitivity_test =round((confusion_pr_test[1,1]/
 ↪(confusion_pr_test[1,0]+confusion_pr_test[1,1])*100),2)

#specificity score
Specificity_test =round((confusion_pr_test[0,0]/
 ↪(confusion_pr_test[0,0]+confusion_pr_test[0,1])*100),2)

#print both
print('Sensitivity:',Sensitivity_test)
print('Specificity:',Specificity_test)
```

[ ]:
```
#  Pecision score
Precision_test = round((confusion_pr_test[1,1]/
 ↪(confusion_pr_test[0,1]+confusion_pr_test[1,1])*100),2)

#  Recall score
Recall_test = round((confusion_pr_test[1,1]/
 ↪(confusion_pr_test[1,0]+confusion_pr_test[1,1])*100),2)

#print both
print('Precision:',Precision_test)
print('Recall:',Recall_test)
```

[ ]:
```
# Checking test set accuracy
```

```
Accuracy_test = round(metrics.accuracy_score(y_test_pred_final.
 ↪Converted,y_test_pred_final.Predicted)*100,2)
print('Test set Accuracy:',Accuracy_test)
```

[ ]:
```
# Creating precision recall curve by crreating three points and plotting

p ,r, thresholds=precision_recall_curve(y_test_pred_final.Converted,␣
 ↪y_test_pred_final.Converted_Probability)
plt.title('Precision vs Recall tradeoff on test set')
plt.plot(thresholds, p[:-1], "g-")    # Plotting precision
plt.plot(thresholds, r[:-1], "r-")    # Plotting Recall
plt.show()
```

[ ]:
```
print('F1_Score: ',f1_score(y_test_pred_final.Converted, y_test_pred_final.
 ↪Predicted)*100)
```

9.1.1 Metrics Comparison between Train data set and Test data set

[ ]:
```
print('Train Data Set metrics:')
print()
print('Sensitivity:',Sensitivity_train)
print('Specificity:',Specificity_train)
print('Precision:',Precision_train)
print('Recall:',Recall_train)
print('Accuracy:',Accuracy_train)
print()
print('Test Data Set metrics:')
print()
print('Sensitivity:',Sensitivity_test)
print('Specificity:',Specificity_test)
print('Precision:',Precision_test)
print('Recall:',Recall_test)
print('Accuracy:',Accuracy_test)
```

9.1.2 Assigning a Lead Score to the Predicted values based on Lead Number

[ ]:
```
# Creating new columns for lead number and lead score
# lead score indicates higher score are hotter the leads and lower score are␣
 ↪colder the leads.

y_test_pred_final['Lead Number']=leads.iloc[y_test_pred_final['ID'],1]

y_test_pred_final['Lead Score']=y_test_pred_final['Converted_Probability'].
 ↪apply(lambda x:round(x*100))

y_test_pred_final.head(20)
```

9.2 Conclusion

9.2.1 Valuable Insights -

- The Sensitivity and Specificity, Accuracy, Precision and Recall score we got from test set are almost accurate.
- We have high recall score than precision score which is a sign of good model.
- In business terms, this model has an ability to adjust with the company's requirements in coming future.
- This concludes that the model is in stable state.
- Important features responsible for good conversion rate or the ones' which contributes more towards the probability of a lead getting converted are :
  - Lead Origin_Lead Add Form
  - Total Time Spent on Website
  - What is your current occupation_Working Professional

[ ]: