Lead Scoring Assignment

Anjali Paravannoor
Akshay Shetty

Steps Involved

Data Understanding

Fix Data Quality Issues

Checking and Handling for Missing Values

Visualize the Data

Outlier Detection and Treatment

Prepare the Data for Modeling

Modeling

Insights

Data Understanding

1. Read the Dataset

	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 curren occupatior
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
1	2a272436-5132-4136-86fa-dcc88c88f482	660728	API	Organic Search	No	No	0	5.0	674	2.5	Email Opened	India	Select	Select	Unemployed
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	Studen
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
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

2. Check distribution of Numerical Variables

	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
max	660737.000000	1.000000	251.000000	2272.000000	55.000000	18.000000	20.000000

3. Get Information on the Dataset

```
<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
```

We observe that there are many columns with missing values. We will analyze them in later steps.

Fix Data Quality Issues

Unique Values

Columns with just 1 value will not impact the analysis or the model, so removing the below columns:

- Magazine
- Update me on Supply Chain Content
- Receive More Updates About Our Courses
- Get updates on DM Content
- I agree to pay the amount through cheque

```
Prospect ID                                    9240
Lead Number                                    9240
Lead Origin                                       5
Lead Source                                      21
Do Not Email                                      2
Do Not Call                                       2
Converted                                         2
TotalVisits                                      41
Total Time Spent on Website                    1731
Page Views Per Visit                            114
Last Activity                                    17
Country                                          38
Specialization                                   19
How did you hear about X Education               10
What is your current occupation                   6
What matters most to you in choosing a course     3
Search                                            2
Magazine                                          1
Newspaper Article                                 2
X Education Forums                                2
Newspaper                                         2
Digital Advertisement                             2
Through Recommendations                           2
Receive More Updates About Our Courses            1
Tags                                             26
Lead Quality                                      5
Update me on Supply Chain Content                 1
Get updates on DM Content                         1
Lead Profile                                      6
City                                              7
Asymmetrique Activity Index                       3
Asymmetrique Profile Index                        3
Asymmetrique Activity Score                      12
Asymmetrique Profile Score                       10
I agree to pay the amount through cheque          1
A free copy of Mastering The Interview            2
Last Notable Activity                            16
dtype: int64
```

Select

Below columns have something like 'Select' as a value for the column which means the data is not available. Replacing with NaN for now and will handle in next segment

- Specialization
- How did you hear about X Education
- City
- Lead Profile

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

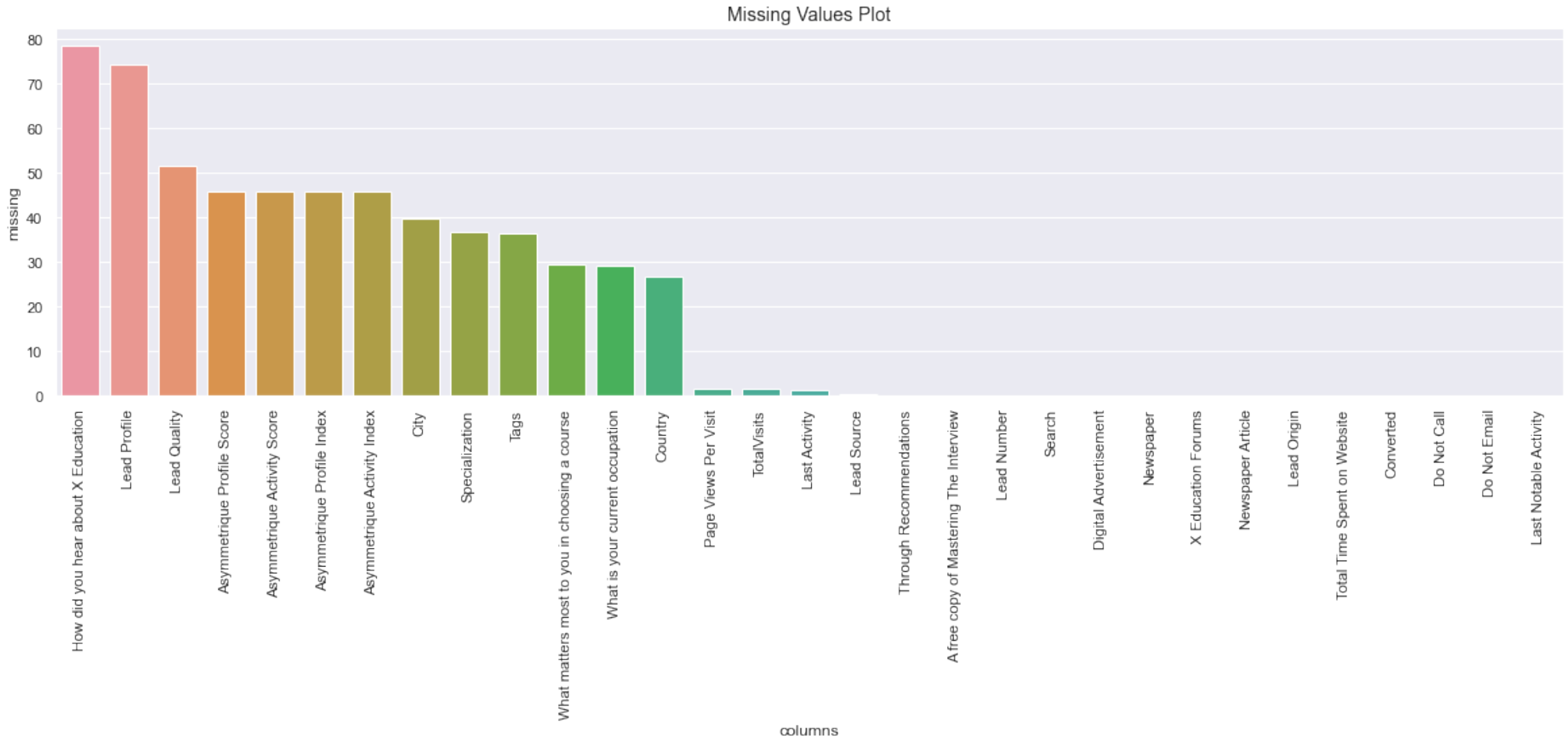
```
How did you hear about X Education
Select                   5043
Online Search             808
Word Of Mouth             348
Student of SomeSchool     310
Other                     186
Multiple Sources          152
Advertisements             70
Social Media               67
Email                      26
SMS                        23
Name: How did you hear about X Education, 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
Media and Advertising                 203
Travel and Tourism                    203
International Business                178
Healthcare Management                 159
Hospitality Management                114
E-COMMERCE                            112
Retail Management                     100
Rural and Agribusiness                 73
E-Business                             57
Services Excellence                    40
Name: Specialization, dtype: int64
```

Checking & Handling for Missing Values

1.Removing the following columns that have more than 40% missing values and also these columns are generated by Sales Team & do not come from Source system so we need not have them in the model.

- How did you hear about X Education
- Lead Profile
- Lead Quality
- Asymmetrique Profile Score
- Asymmetrique Activity Score
- Asymmetrique Profile Index
- Asymmetrique Activity Index



2. To handle the null values further, we update the values as 'Not Provided' for the categorical columns where ever they have missing values

- City
- Tags
- Specialization
- What matters most to you in choosing a course?
- What is your current occupation
- Country

3. Remaining below columns have very less percentage of null values. So, let us drop these records as we have good amount of data even after removing it.

- Total Visits
- Page Views Per Visit
- Last Activity
- Lead Source

Only 1.48% loss of records and 98.2% records are retained after null value records are removed.

4. We notice that the below columns have so many values/categories which hold very less percentage of entire data.

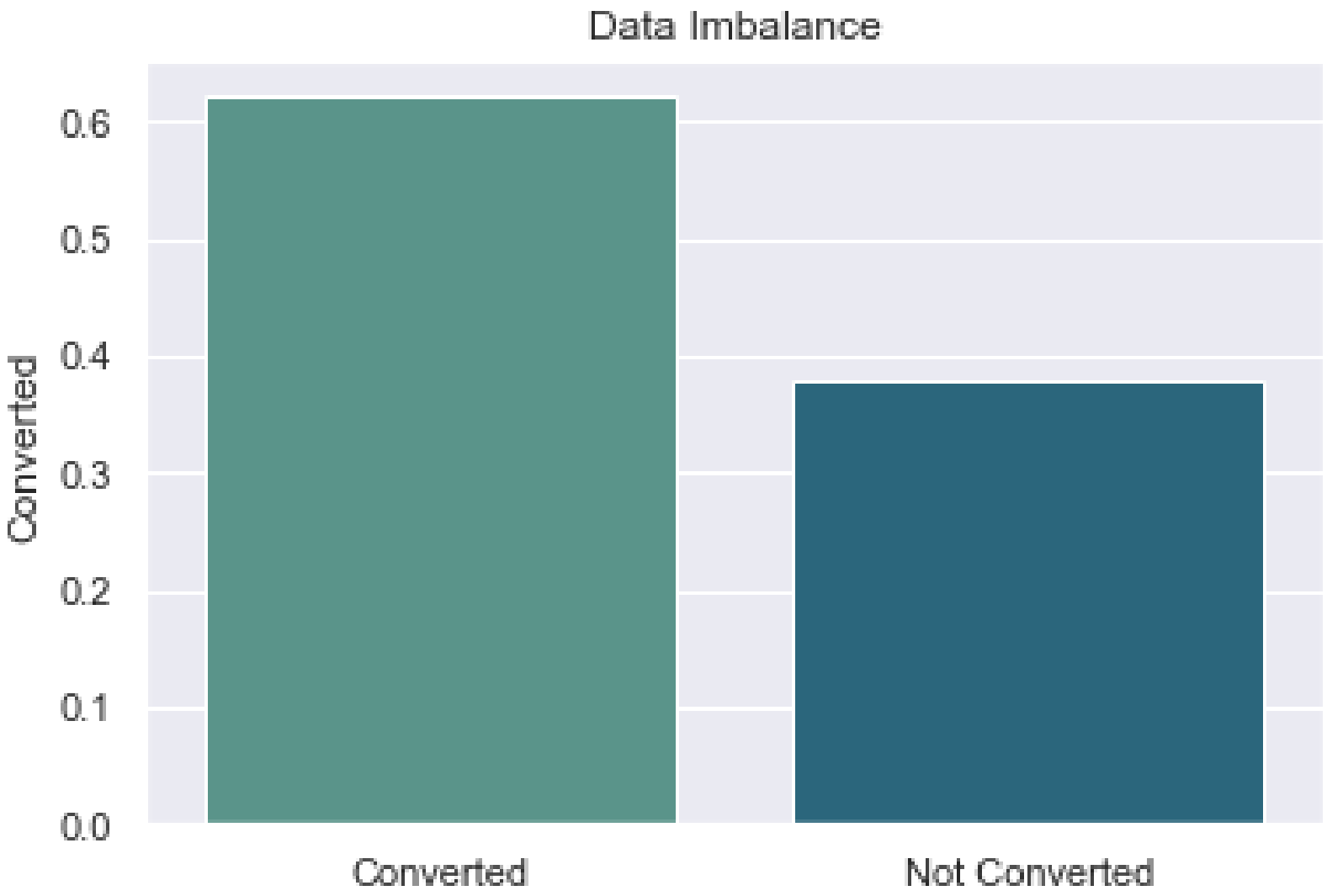
- Lead Source
- Specialization
- Country
- City

Let us combine the least percentage/ Low frequency categories into one single category, so that we have lesser dummy variables and easy to interprete model.

Visualize the data

Data Imbalance

We have currently 37% Lead Conversion Rate in the data. The data has 62:38 ratio which seems to be balanced

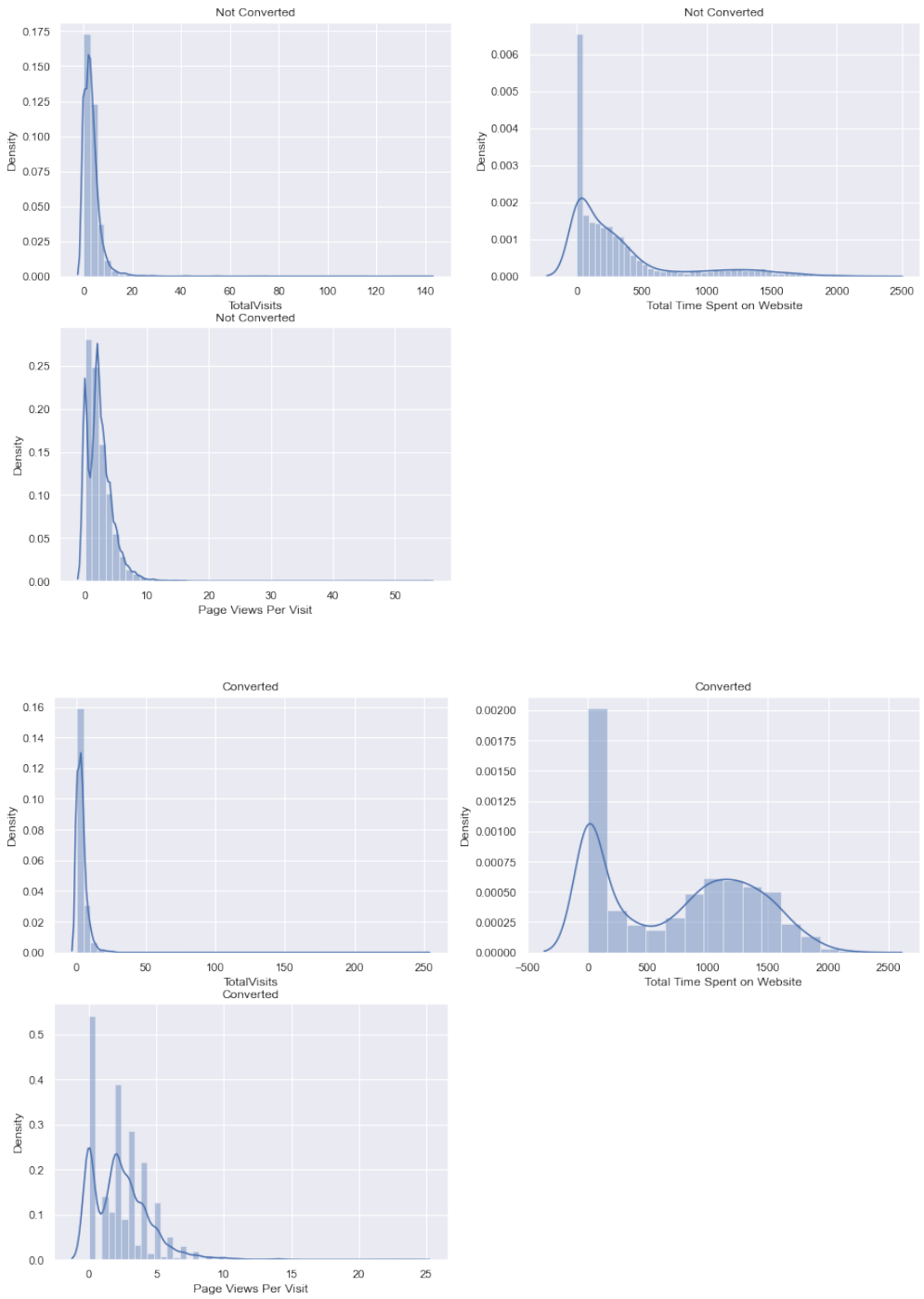


Univariate Analysis

Numerical Variables:

- Total Visits
- Total Time Spent on Website
- Page Views Per Visit

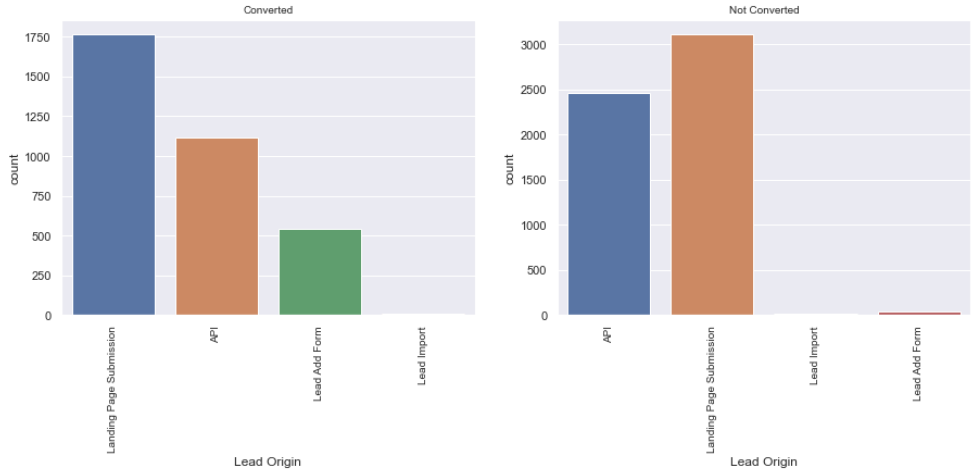
Total time spent on website shows a sharp increase for leads converted as compared to not converted. This makes sense considering interested students would spend more time on the website to understand various aspects of the course such as curriculum, specializations, professors, etc.

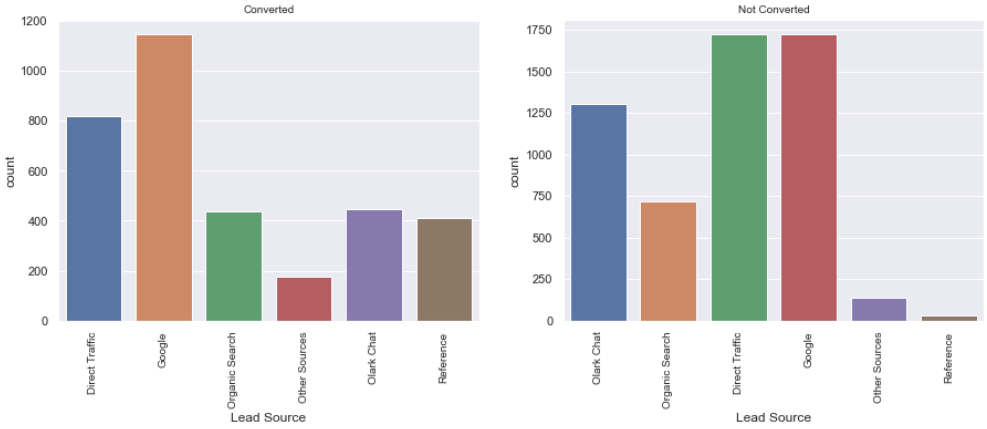


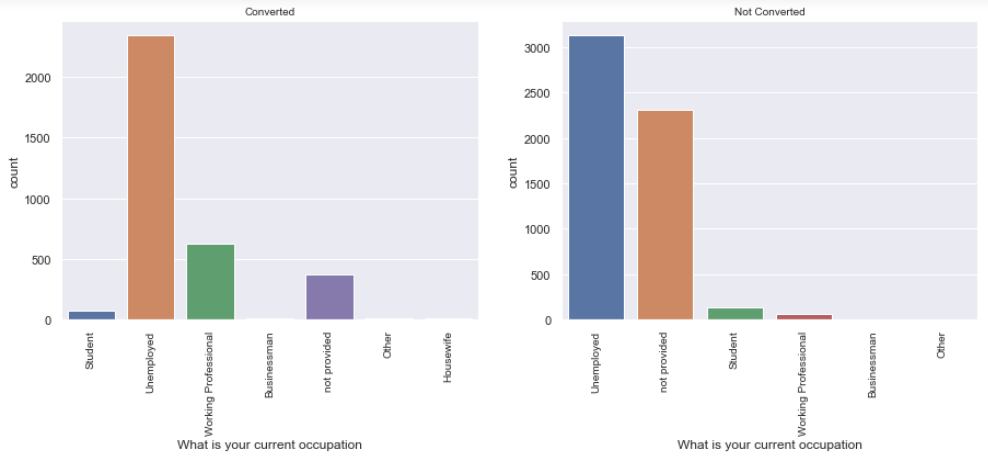
Univariate Analysis

Categorical Variables:

- **Current occupation**: Most number of unemployed people have been approached. But working professionals show a high conversion rate.
- **Lead Origin**: Lead Add Form shows the highest conversion rate.
- **Lead Source**: Leads coming from Reference have shown a higher conversion rate.



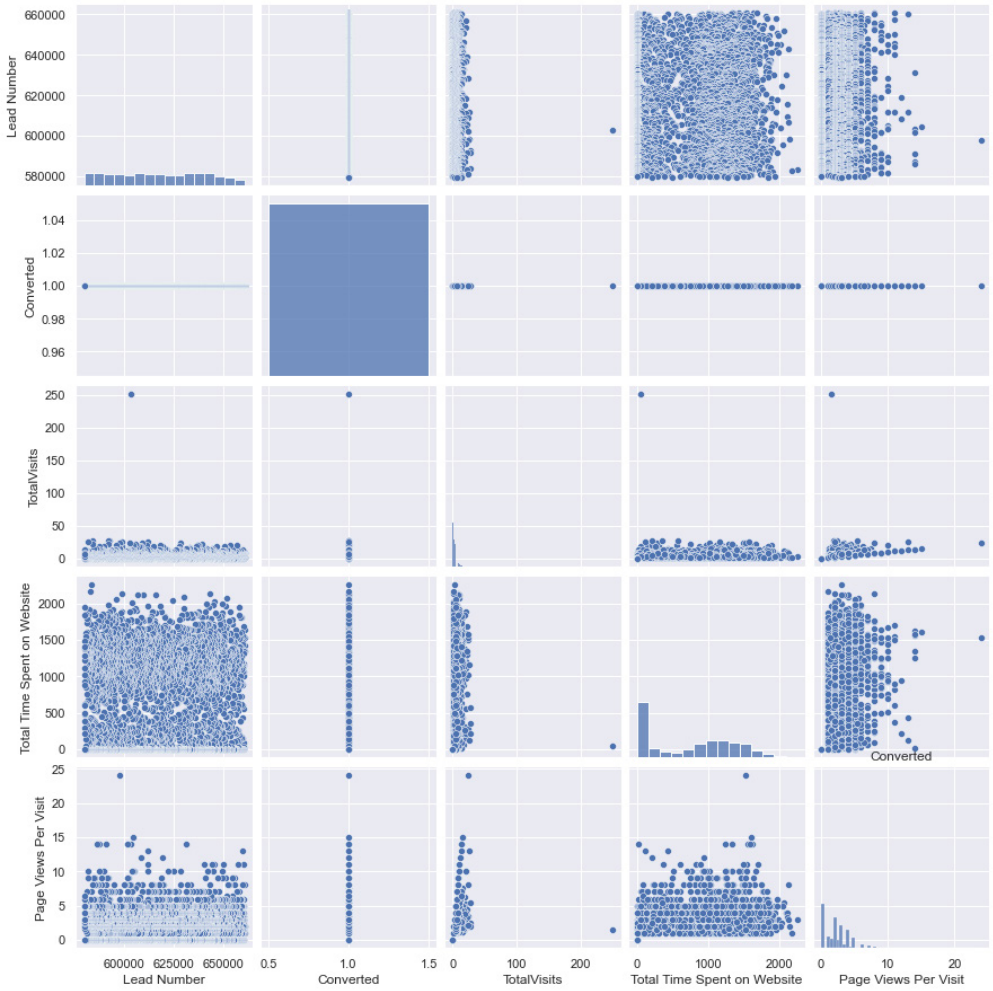


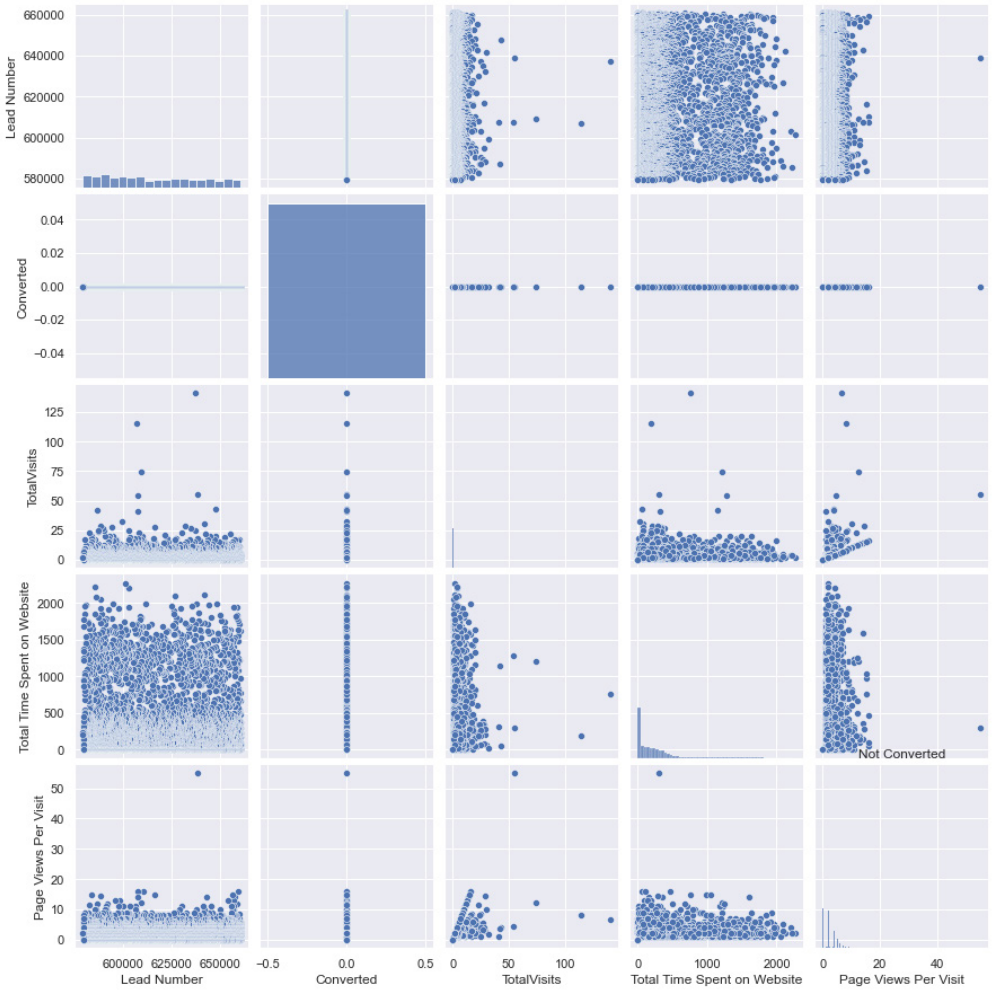


Bivariate Analysis

Numerical Variables:

We can see no discernable relationships or correlation between the numerical variables.



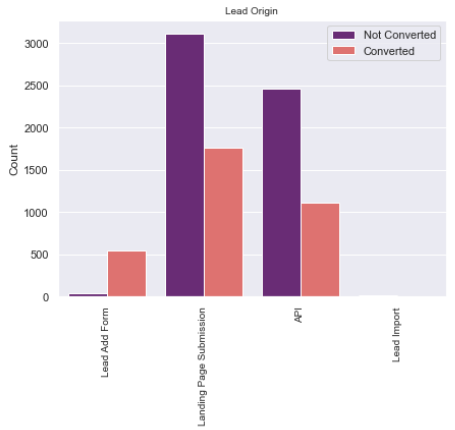


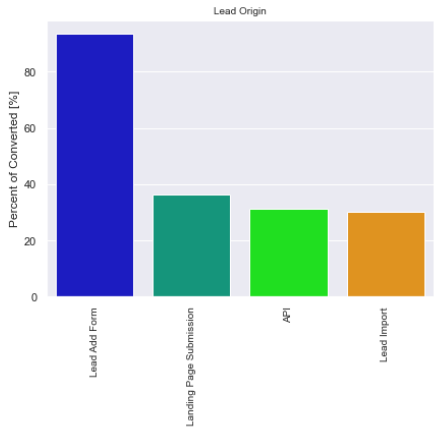
Bivariate Analysis

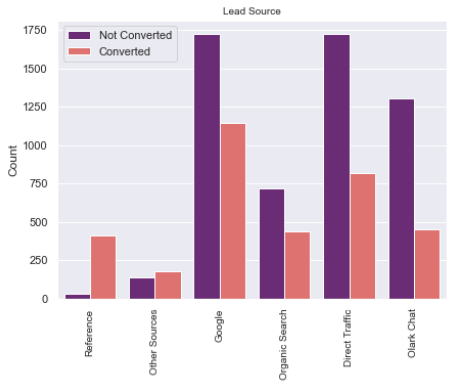
Categorical Variables:

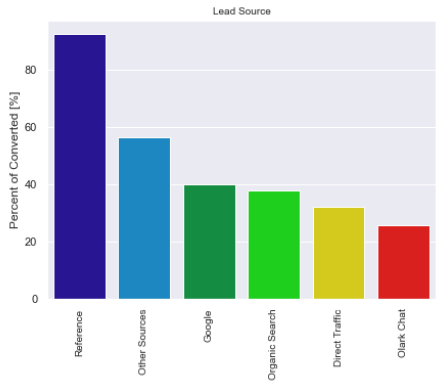
- **Current occupation**: Most number of unemployed people have been approached. But working professionals show a high conversion rate.
- **Lead Origin**: Lead Add Form shows the highest conversion rate.
- **Lead Source**: Leads coming from Reference have shown a higher conversion rate.

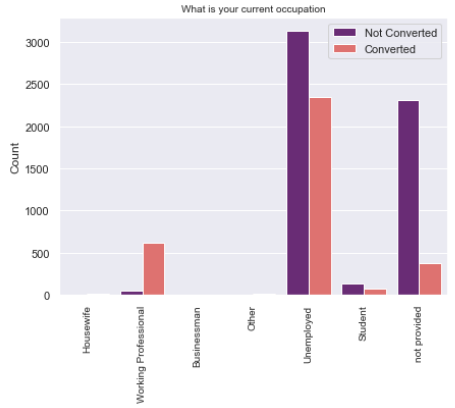
The rest of the categorical variables do not show any discernable patterns at this point.

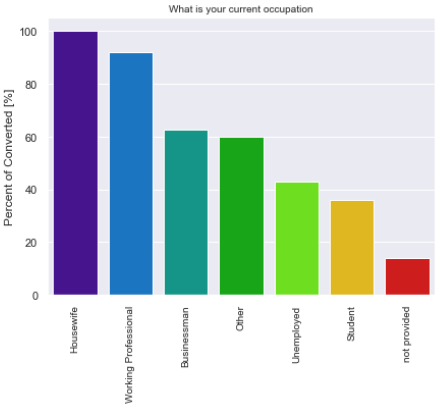






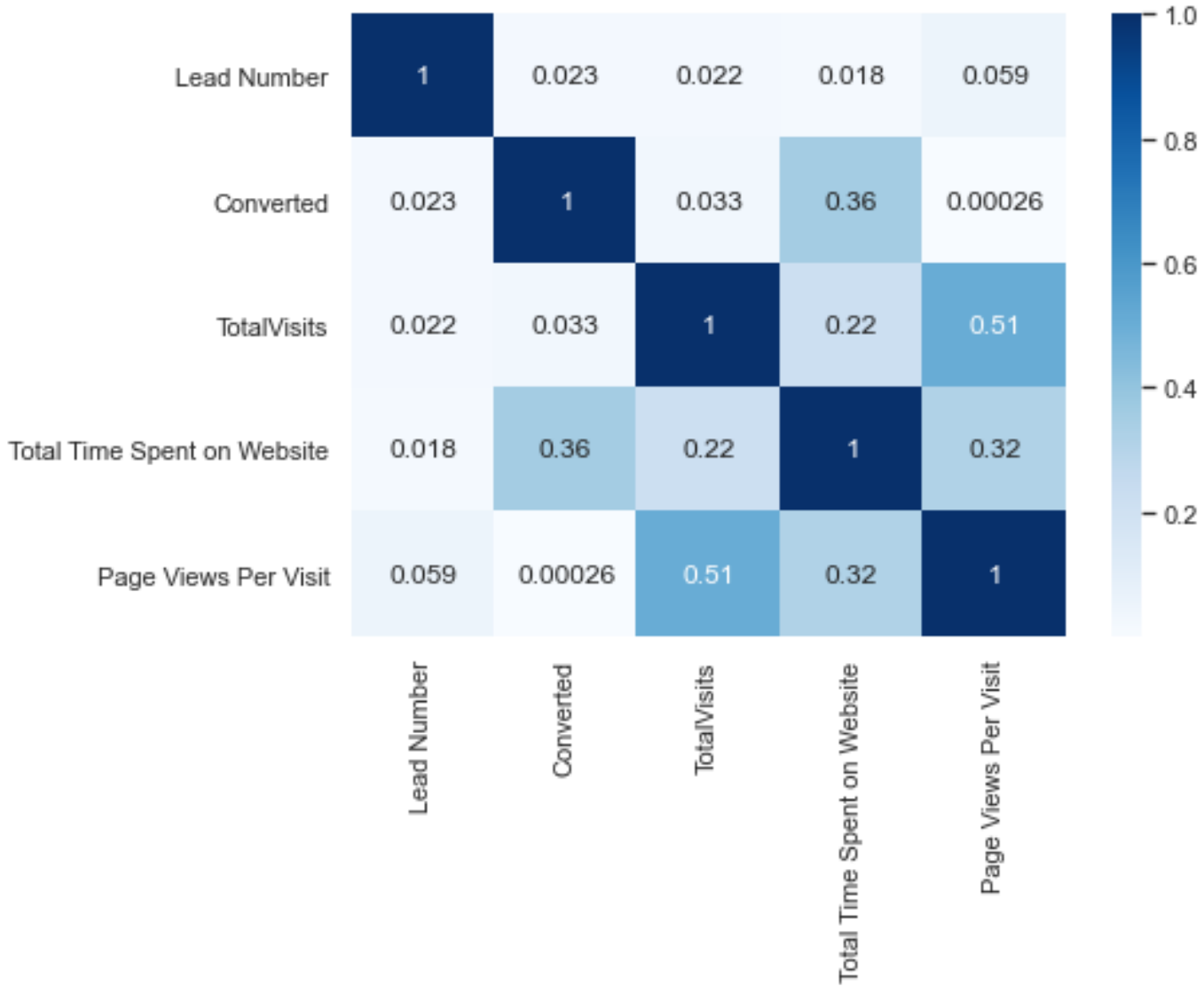






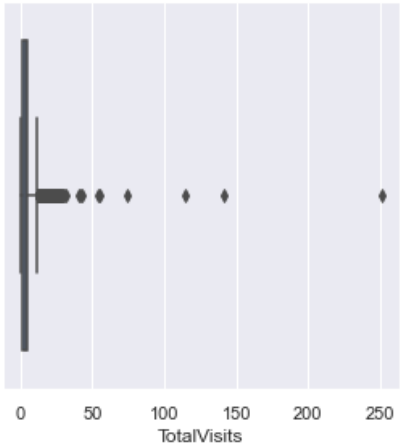
Multivariate Analysis

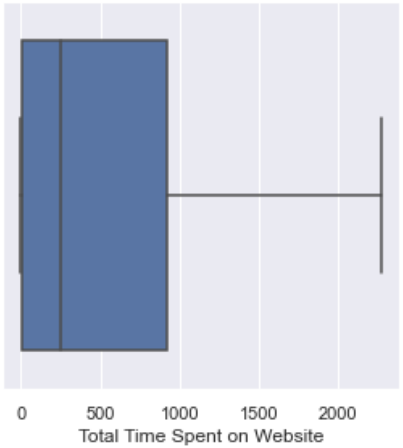
We see a moderate correlation between Total Visits and Page Views per Visit but no strong correlations.

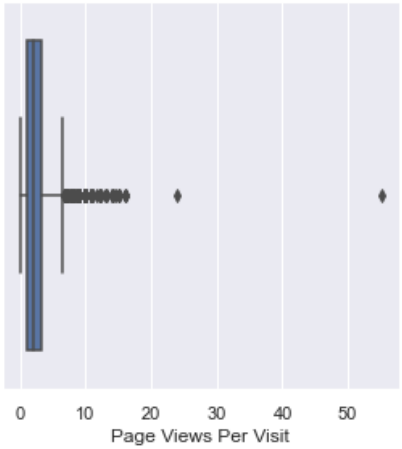


Outlier Detection & Treatment

2 columns with outliers identified:
Total Visits & Page Views Per Visit





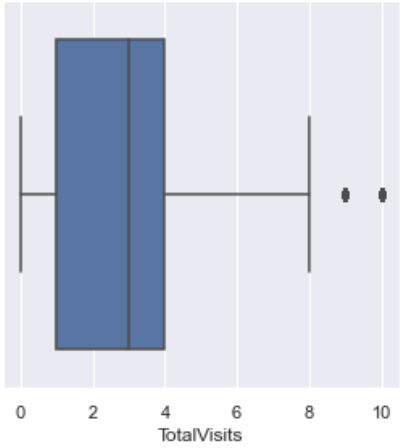


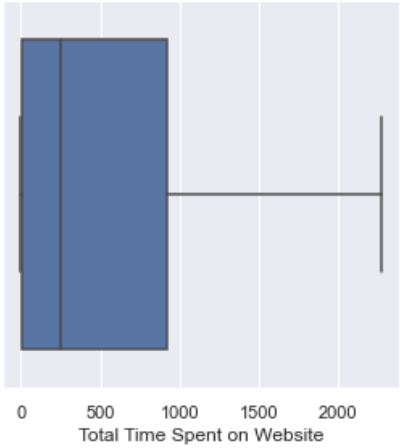
	index	Page Views Per Visit
0	count	9074.00
1	mean	2.37
2	std	2.16
3	min	0.00
4	10%	0.00
5	25%	1.00
6	50%	2.00
7	75%	3.20
8	90%	5.00
9	95%	6.00
10	99%	9.00
11	100%	55.00
12	max	55.00

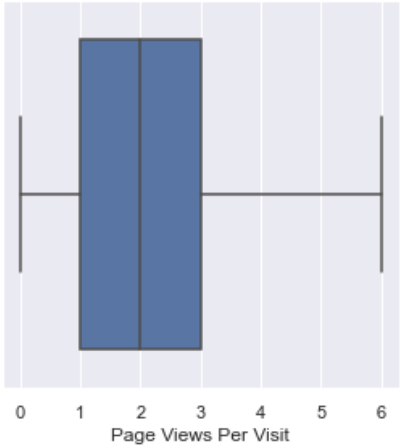
	index	TotalVisits
0	count	9074.00
1	mean	2.92
2	std	2.42
3	min	0.00
4	10%	0.00
5	25%	1.00
6	50%	3.00
7	75%	4.00
8	90%	6.00
9	95%	8.00
10	99%	10.00
11	100%	10.00
12	max	10.00

Statistical Summary for Total Visits and Page Views per Visit

- **Total Visits**: Replacing the outlier greater than Q3(95%) with Median value
- **Page Views Per Visit**: Replacing the outlier greater than Q4 (99%) with Median values







◀ After replacement of Outliers

Prepare the data for modelling

Data Generated by Sales Team

The objective of this exercise is to help the company understand what profile of customers make good leads so that the company can only call those customers. The following columns contain information updated by the sales team after making the initial calls.

- Lead Profile
- Lead Quality
- Asymmetrique Profile Score
- Asymmetrique Activity Score
- Asymmetrique Profile Index
- Asymmetrique Activity Index
- Tags
- Last Activity
- Last Notable Activity

First 6 were removed while handling missing values. So, now we remove the last 3 columns.

Dummy Variables

After removing the columns generated by the sales team we are left with the following categorical columns. We shall proceed to create dummy variables for the same.

- Lead Origin
- Lead Source
- Do Not Email
- Do Not Call
- Country
- Specialization
- What is your current occupation
- What matters most to you in choosing a course
- Search
- Newspaper Article
- X Education Forums
- Newspaper
- Digital Advertisement
- Through Recommendations
- City
- A free copy of Mastering The Interview

Test-Train Split

The **Converted** column is the target variable. Therefore we will separate this out from the features we need for modeling. We then use **train_test_split from sklearn library** to create a **70:30** split of training and test data. After splitting the dataset we have:

- The training set has 6351 rows and 39 columns.
- The test set has 2723 rows and 39 columns.

Scaling

We use the **StandardScaler** that follows Standard Normal Distribution (SND). Therefore, it makes mean = 0 and scales the data to unit variance. We have three numerical columns: Total Visits, Total Time Spent on Website, Page Views Per Visit

	TotalVisits	Total Time Spent on Website	Page Views Per Visit
3009	-0.385477	-0.160255	-0.076507
1012	-0.385477	-0.540048	-0.076507
9226	-1.206045	-0.888650	-1.295721
4750	-0.385477	1.643304	-0.076507
7987	0.845375	2.017593	0.228296

Modeling

Feature Selection using RFE

We have a large number of columns. Carrying out an entirely manual selection will be a time consuming process. Therefore we use Recursive Feature Selection to narrow down the number of features to 20. These are the selected features:

- Total Time Spent on Website
- Lead Origin_Landing Page Submission
- Lead Origin_Lead Add Form
- Lead Origin_Lead Import
- Lead Source_Google
- Lead Source_Olark Chat
- Lead Source_Reference
- Do Not Email_Yes
- Country_not provided
- Specialization_not provided
- What is your current occupation_Housewife
- What is your current occupation_Other
- What is your current occupation_Student
- What is your current occupation_Unemployed
- What is your current occupation_Working Professional
- What is your current occupation_not provided
- What matters most to you in choosing a course_not provided
- Newspaper Article_Yes
- Newspaper_Yes
- City_not provided

Model Building

Significance Level and p-value is the amount of change a feature will affect towards the final output i.e. how important is this feature and how much it affects the final output. We take **5%/0.05** significance level.

VIF is a measure of the amount of multicollinearity in a set of multiple regression variables. We have considered a **threshold of 5** for VIF.

After 11 iteration on manual selection we are left with the 10 variables that fulfill the above criterias for p-value and VIF:

- Lead Origin_Lead Add Form
- Lead Source_Reference
- Specialization_not provided
- Lead Source_Olark Chat
- Lead Origin_Landing Page Submission
- What matters most to you in choosing a course_not provided
- Lead Source_Google
- Total Time Spent on Website
- What is your current occupation_Working Professional
- Do Not Email_Yes
- What is your current occupation_Student

Generalized Linear Model Regression Results

Dep. Variable:	Converted	No. Observations:	6351
Model:	GLM	Df Residuals:	6340
Model Family:	Binomial	Df Model:	10
Link Function:	logit	Scale:	1.0000
Method:	IRLS	Log-Likelihood:	-2812.6
Date:	Wed, 14 Jul 2021	Deviance:	5625.2
Time:	10:32:42	Pearson chi2:	6.70e+03
No. Iterations:	6		
Covariance Type:	nonrobust		

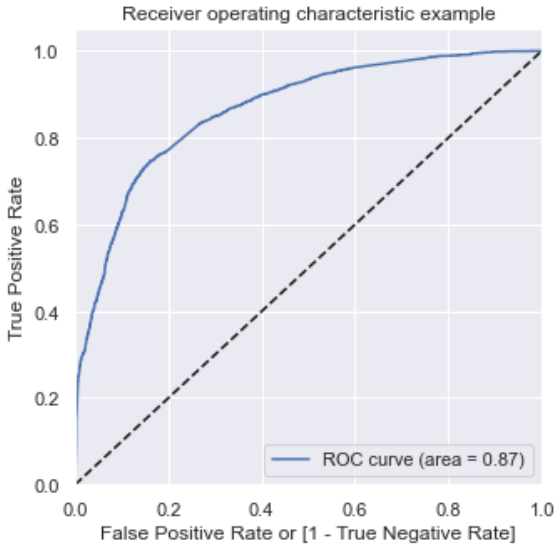
	coef	std err	z	P>|z|	[0.025	0.975]
const	-0.1027	0.127	-0.807	0.419	-0.352	0.147
Total Time Spent on Website	1.1125	0.039	28.457	0.000	1.036	1.189
Lead Origin_Landing Page Submission	-0.7316	0.124	-5.918	0.000	-0.974	-0.489
Lead Origin_Lead Add Form	5.3169	0.526	10.102	0.000	4.285	6.348
Lead Source_Google	0.2680	0.078	3.440	0.001	0.115	0.421
Lead Source_Olark Chat	1.1819	0.125	9.463	0.000	0.937	1.427
Lead Source_Reference	-1.7169	0.562	-3.056	0.002	-2.818	-0.616
Do Not Email_Yes	-1.4033	0.166	-8.458	0.000	-1.728	-1.078
Specialization_not provided	-0.8979	0.119	-7.517	0.000	-1.132	-0.664
What is your current occupation_Working Professional	2.3880	0.186	12.815	0.000	2.023	2.753
What matters most to you in choosing a course_not provided	-1.3142	0.084	-15.559	0.000	-1.480	-1.149

	Features	VIF
2	Lead Origin_Lead Add Form	4.41
5	Lead Source_Reference	4.25
7	Specialization_not provided	2.37
4	Lead Source_Olark Chat	2.01
1	Lead Origin_Landing Page Submission	1.63
9	What matters most to you in choosing a course_...	1.61
3	Lead Source_Google	1.58
0	Total Time Spent on Website	1.28
8	What is your current occupation_Working Profes...	1.18
6	Do Not Email_Yes	1.12

1. We make predictions on the training set using Model No.11.

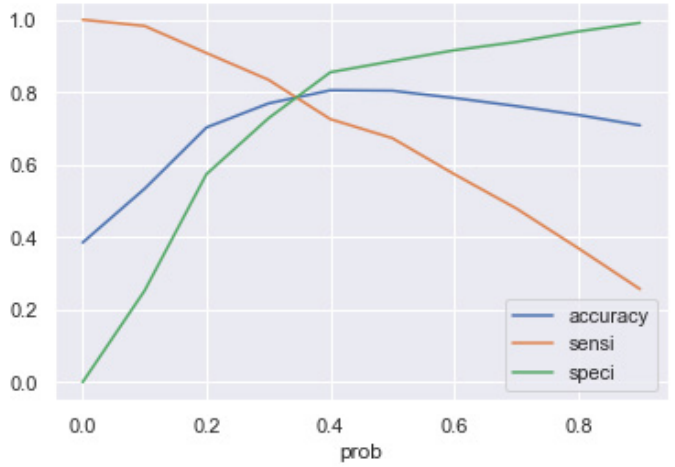
	Converted	Convert_Prob	Lead Number
0	0	0.088935	3009
1	0	0.065227	1012
2	0	0.308477	9226
3	1	0.420595	4750
4	1	0.803808	7987

2. ROC Curve: We use the convert_prob (predicted value) and the Converted (actual value) to make the ROC Curve. As observed the **area under the ROC curve is 0.87** which is a good value.



3. Let's calculate accuracy sensitivity and specificity for various probability cutoffs.

	prob	accuracy	sensi	speci
0.0	0.0	0.385136	1.000000	0.000000
0.1	0.1	0.533617	0.982829	0.252241
0.2	0.2	0.702724	0.908013	0.574136
0.3	0.3	0.769013	0.834424	0.728041
0.4	0.4	0.805228	0.725675	0.855058
0.5	0.5	0.803968	0.673344	0.885787
0.6	0.6	0.784128	0.573998	0.915749
0.7	0.7	0.761770	0.479558	0.938540
0.8	0.8	0.737364	0.369992	0.967478
0.9	0.9	0.708550	0.256746	0.991549



From the curve above, 0.3 is the optimum point to take it as a cutoff probability.

4. We now assign a value based on the train data with 0.3 as the cutoff probability.

After choosing the cutt-off as 0.3, with 10 variables model,the stats are as below:

- Accuracy: 76.9%
- Sensitivity: 83.44%
- Specificity: 72.8%

As required in the problem statement, the sensitivity is greater than 80%.

5. We now make predictions on the Test set and assign a Lead Score.

	Lead Number	Converted	Convert_Prob	Lead_Score	final_predicted
0	3271	0	0.062100	6	0
1	1490	1	0.974750	97	1
2	7936	0	0.052935	5	0
3	4216	1	0.924757	92	1
4	3830	0	0.057201	6	0

After choosing the cutt-off as 0.3, with 10 variables model,the stats are as below for the test data:

- Accuracy: 76.8%
- Sensitivity: 82.9%
- Specificity: 73.41%

The CEO had a requirement that the target lead conversion rate should be 80%. As seen in the results above, this condition is fulfilled.

Insights

Positive Influence of Lead Conversion:

The company should focus on the following features to generate higher lead converion rates.

- **What is your current occupation_Working Professional (2.388 coefficient)** - The company should focus on targeting working professionals. These courses seem to be designed for working professionals to upskill. Therefore students who are already enrolled in another course and unemployed people are less inclined to enroll in courses by X Education.
- **Total Time Spent on Website (1.115 coefficient) -** Potential leads that are genuinely interested in the course will spend more time on the website to research about the curriculum, professors, fees, and career prospects. The leads team should focus on these leads.
- **Lead Origin_Lead Add Form (5.319 coefficient) -** Leads generated via aff forms have shown high chances of being converted.