Lead Scoring Case Study

Date: 30-03-2024

Pavan Kumar N
July 2023 batch
IIITB
9901682405

Step 1: Lets import all the necessary libraries and look at the data sample

[13]:
```
#lets check how the data looks
leadsdata.head()
```

[13]:

	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	m m ch a
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	Pro
1	2a272436-5132-4136-86fa-dcc88c88f482	660728	API	Organic Search	No	No	0	5.0	674	2.5	Email Opened	India	Select	Select	Unemployed	Pro
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	Pro
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	Pro
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	Pro

[20]:
```
leadsdata.describe()
```

[20]:

	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

[21]:
```
leadsdata.shape
```

[21]: (9240, 37)

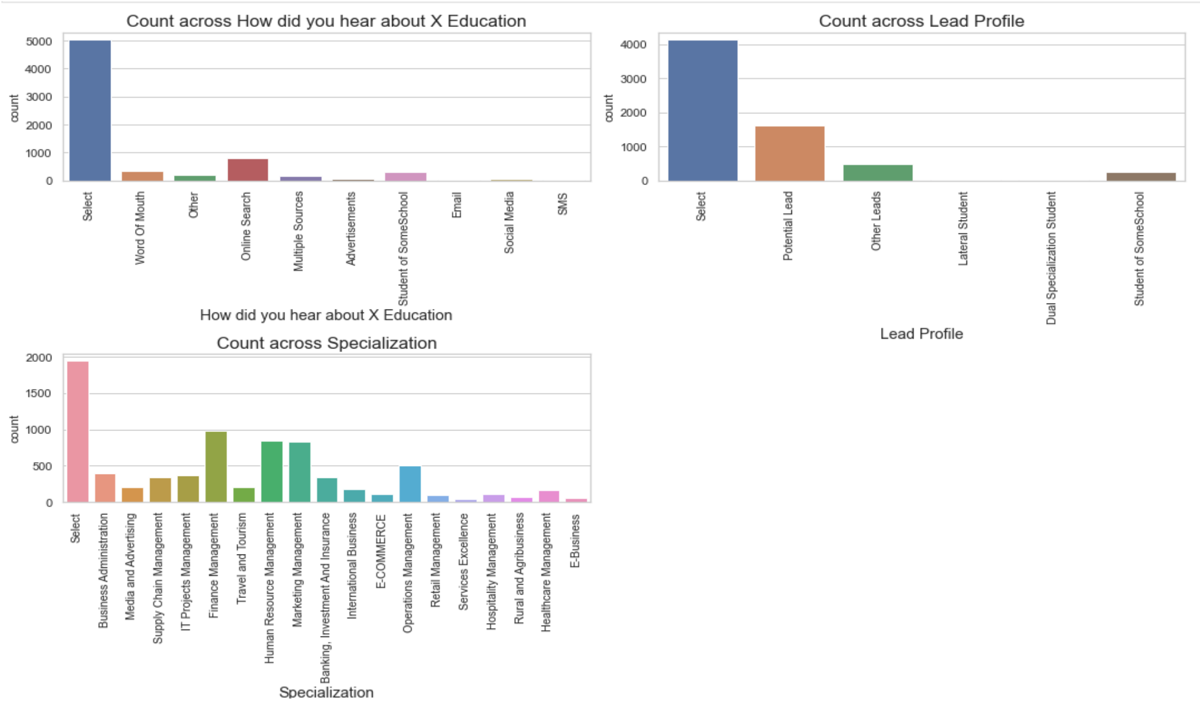
Step 2: Data cleaning and preparation

Lets look at all the columns having null values and drop few columns accordingly

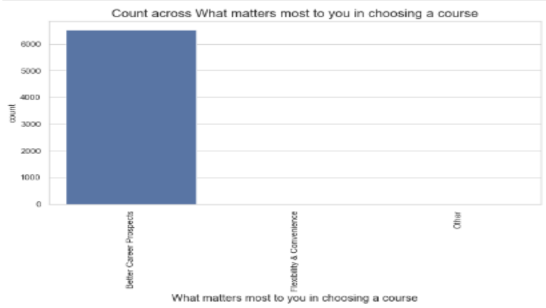
```
[9]: # Checking the number of missing values in each column
leadsdata.isnull().sum().sort_values(ascending=False)

Last Activity                                  103
Lead Source                                     36
Receive More Updates About Our Courses           0
I agree to pay the amount through cheque         0
Get updates on DM Content                        0
Update me on Supply Chain Content                0
A free copy of Mastering The Interview           0
Prospect ID                                      0
Newspaper Article                                0
Through Recommendations                          0
Digital Advertisement                            0
Newspaper                                        0
X Education Forums                               0
Lead Number                                      0
Magazine                                         0
Search                                           0
Total Time Spent on Website                      0
Converted                                        0
```

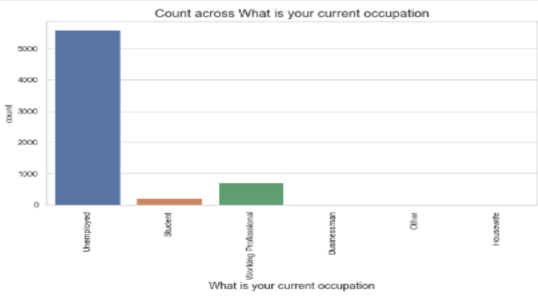
After dropping few columns, we look at the number of missing values in columns and drop some irrelevant columns.

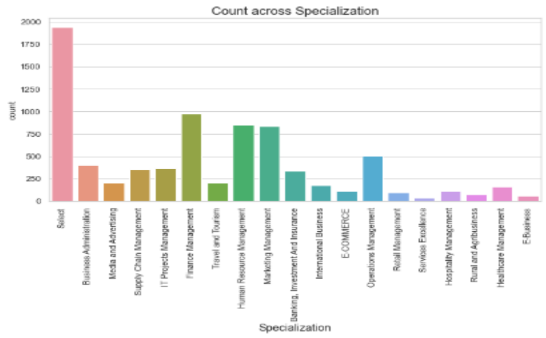
Lets plot values using countplot for further analysis on data preparation



After modifying and cleaning some data, lets plot values again using countplot





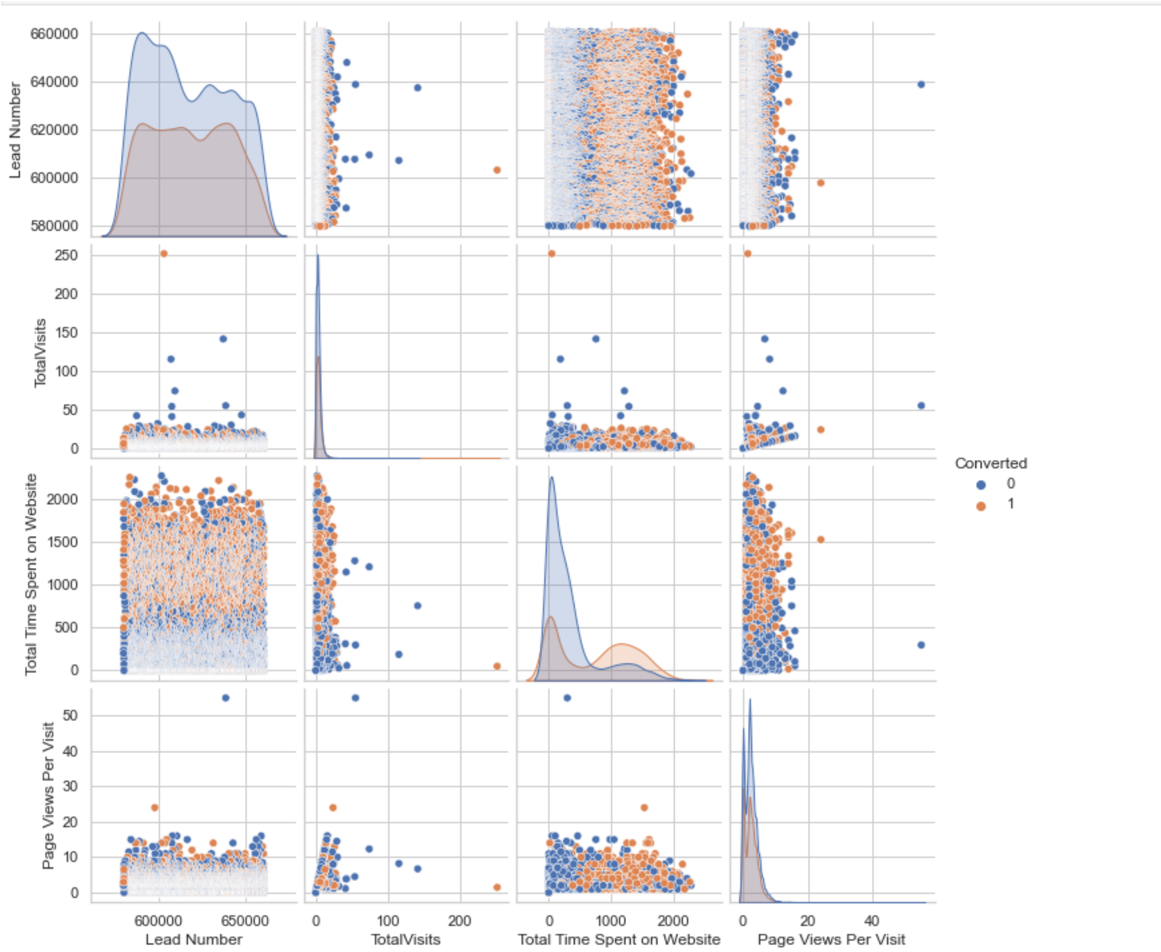


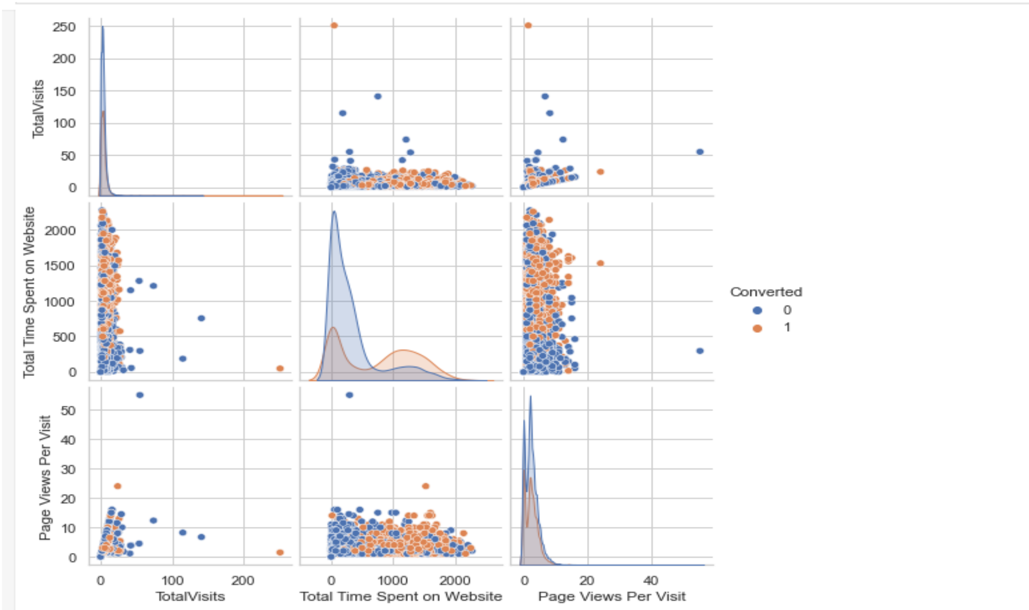
As it can be seen that the levels of "Lead Profile" and "How did you hear about X Education" have a lot of rows which have the value Select which is of no use to the analysis. So it's best that we drop them.

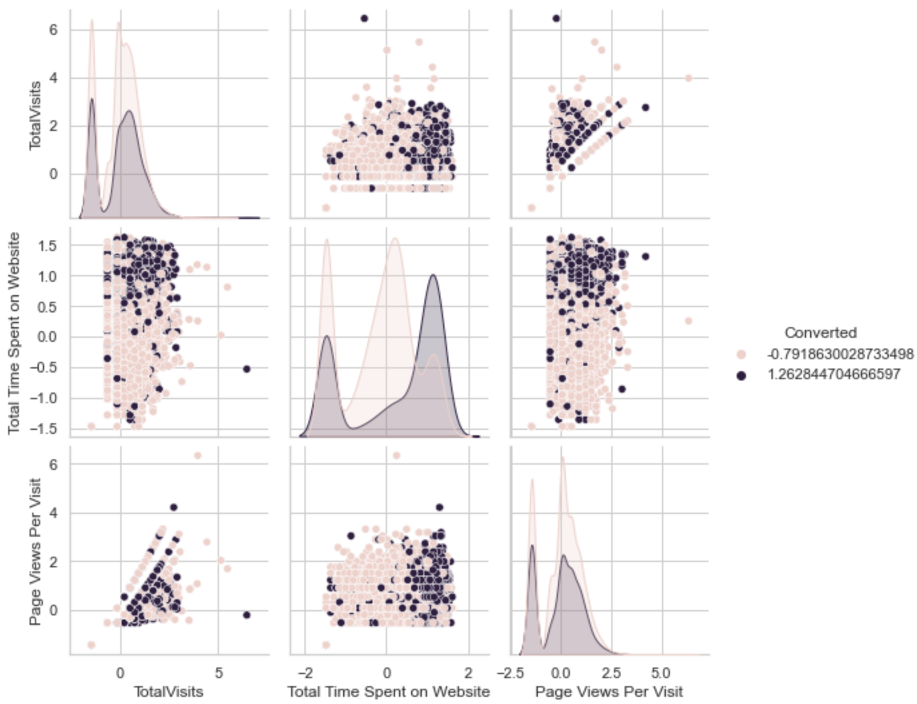
Step 3: Data visualization

In this section, we visualize data. Also, based on the outcome, we may need to perform additional data cleaning and preparation activities here

Lets use pairplot and visualize



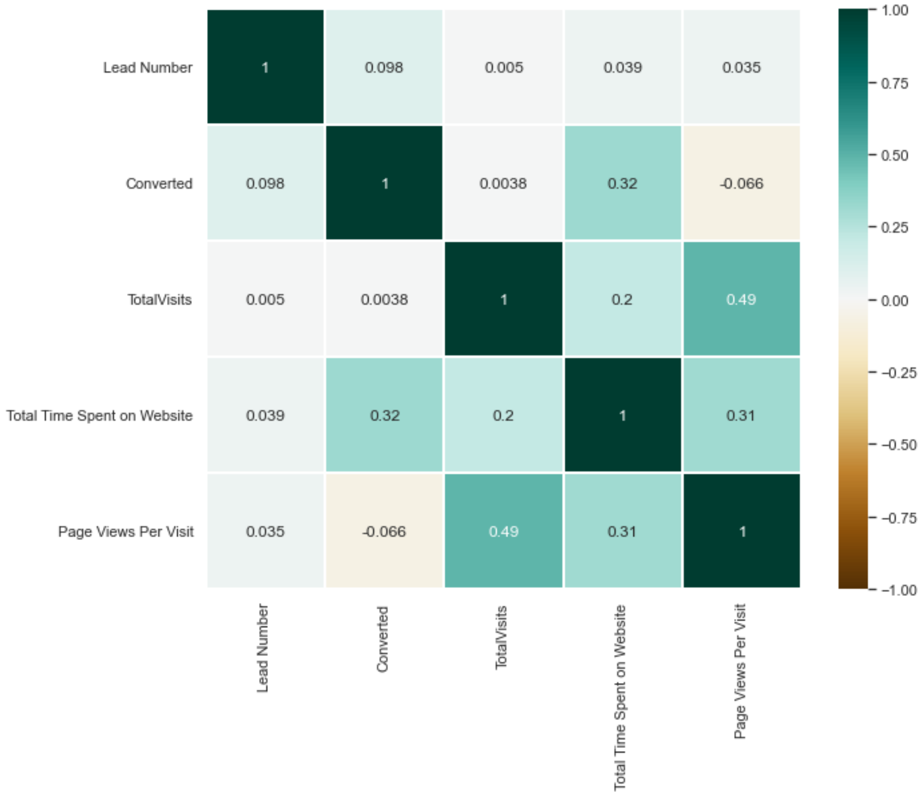




The variable What matters most to you in choosing a course has the level Better Career Prospects 6528 times while the other two levels appear once twice and once respectively.
So we should dropping this column as well.
Now, there's the column What is your current occupation which has a lot of null values. Now you can drop the entire row but since we have already lost so many feature variables, we choose not to drop it as it might turn out to be significant in the analysis. So let's just drop the null rows for the column What is you current occupation.

Step 4: Correlation

Lets use heatmap to visualize the data and understand correlation



We perform additional data cleaning and preparation by dropping few more irrelevant columns.

```
[37]: # Checking the number of null values again
      leadsdata.isnull().sum().sort_values(ascending=False)

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

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

```
[43]: # Checking the number of null values again
leadsdata.isnull().sum().sort_values(ascending=False)

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

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

We further drop Prospect ID and Lead Number as these wont help our analysis.

Step 5:

a. Dummy variable creation and mapping

The next step is to dealing with the categorical variables in the dataset and creating dummy variables for the same for mapping purpose.

After dropping original columns, lets look at dataset

```
leadsdata.head()
```

[56]:

	Converted	TotalVisits	Total Time Spent on Website	Page Views Per Visit	Lead Origin_Landing Page Submission	Lead Origin_Lead Add Form	Lead Origin_Lead Import	Lead Source_Direct Traffic	Lead Source_Facebook	Lead Source_Google	Lead Source_Live Chat	Lead Source_Olark Chat	Sourc
0	0	0.0	0	0.0	0	0	0	0	0	0	0	1	
1	0	5.0	674	2.5	0	0	0	0	0	0	0	0	
2	1	2.0	1532	2.0	1	0	0	1	0	0	0	0	
3	0	1.0	305	1.0	1	0	0	1	0	0	0	0	
4	1	2.0	1428	1.0	1	0	0	0	0	1	0	0	

b. Scaling

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

After scaling dataset looks like below:

[66]:

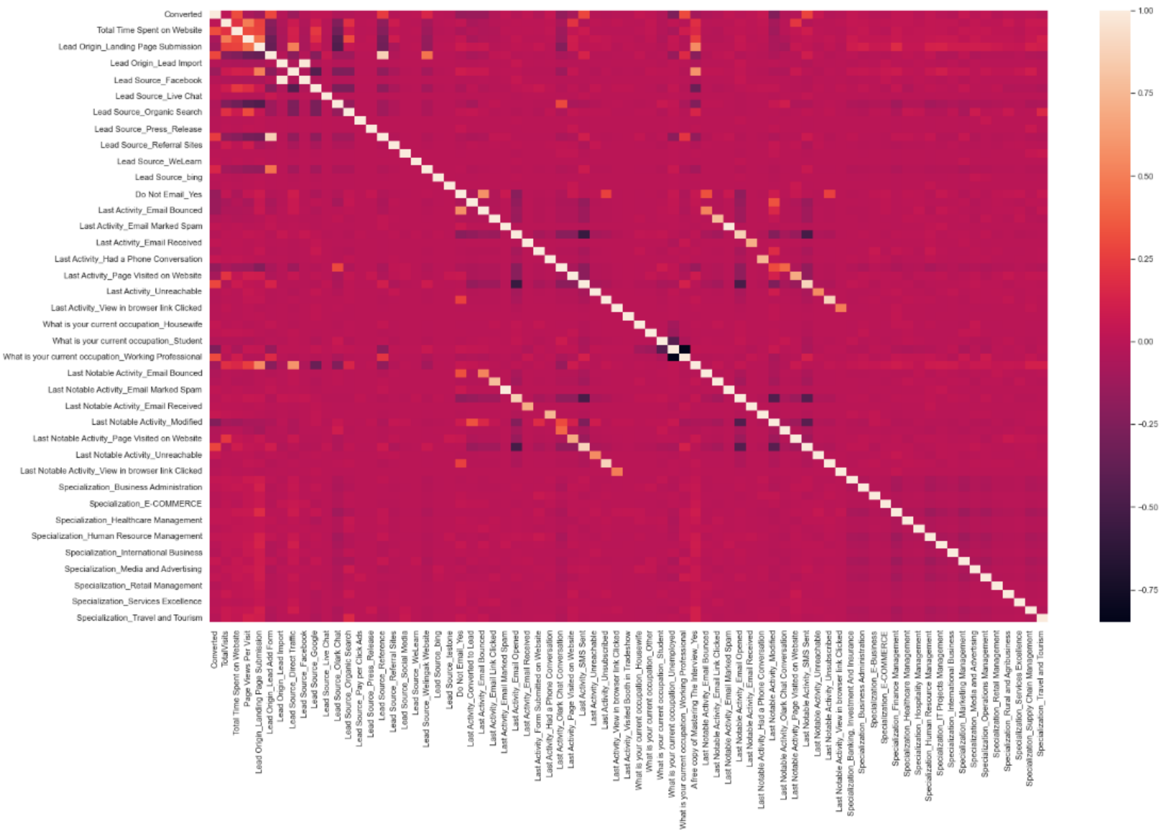
	TotalVisits	Total Time Spent on Website	Page Views Per Visit	Lead Origin_Landing Page Submission	Lead Origin_Lead Add Form	Lead Origin_Lead Import	Lead Source_Direct Traffic	Lead Source_Facebook	Lead Source_Google	Lead Source_Live Chat	Lead Source_Olark Chat	Lea Source_Organ Searc
8003	0.015936	0.029489	0.125	1	0	0	1	0	0	0	0	
218	0.015936	0.082306	0.250	1	0	0	1	0	0	0	0	
4171	0.023904	0.034331	0.375	1	0	0	1	0	0	0	0	
4037	0.000000	0.000000	0.000	0	0	0	0	0	0	0	1	
3660	0.000000	0.000000	0.000	0	1	0	0	0	0	0	0	

Step 6: Test-Train Split

Lets split the data into training and testing data. After Splitting the dataset into 70% train and 30% test, lets check the shape of data.

```
[67]: #lets check the shape
      print("X_train Size", X_train.shape)
      print("y_train Size", y_train.shape)

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

When we do a heat map again for correlation, it looks like below:



Step 7: Model building

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

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

[76]:

Generalized Linear Model Regression Results

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

	coef	std err	z	P>|z|	[0.025	0.975]
const	-1.0061	0.600	-1.677	0.094	-2.182	0.170
TotalVisits	11.3439	2.682	4.230	0.000	6.088	16.600
Total Time Spent on Website	4.4312	0.185	23.924	0.000	4.068	4.794
Lead Origin_Lead Add Form	2.9483	1.191	2.475	0.013	0.614	5.283
Lead Source_Olark Chat	1.4584	0.122	11.962	0.000	1.219	1.697
Lead Source_Reference	1.2994	1.214	1.070	0.285	-1.080	3.679
Lead Source_Welingak Website	3.4159	1.558	2.192	0.028	0.362	6.470
Do Not Email_Yes	-1.5053	0.193	-7.781	0.000	-1.884	-1.126
Last Activity_Had a Phone Conversation	1.0397	0.983	1.058	0.290	-0.887	2.966
Last Activity_SMS Sent	1.1827	0.082	14.362	0.000	1.021	1.344
What is your current occupation_Housewife	22.6492	2.45e+04	0.001	0.999	-4.8e+04	4.8e+04
What is your current occupation_Student	-1.1544	0.630	-1.831	0.067	-2.390	0.081
What is your current occupation_Unemployed	-1.3395	0.594	-2.254	0.024	-2.505	-0.175
What is your current occupation_Working Professional	1.2743	0.623	2.045	0.041	0.053	2.496
Last Notable Activity_Had a Phone Conversation	23.1932	2.08e+04	0.001	0.999	-4.08e+04	4.08e+04
Last Notable Activity_Unreachable	2.7868	0.807	3.453	0.001	1.205	4.369

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

]:

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

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

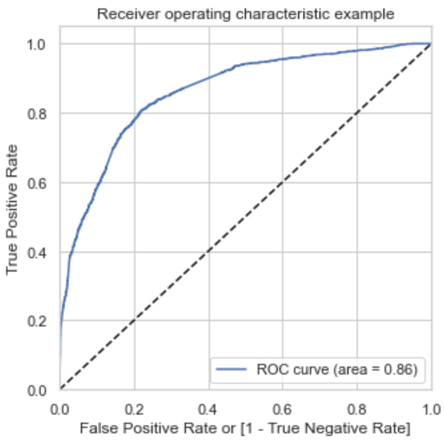
Let's first drop the variable Lead Source_Reference since it has a high p-value as well as a high VIF.

We build further models and perform analysis similarly.

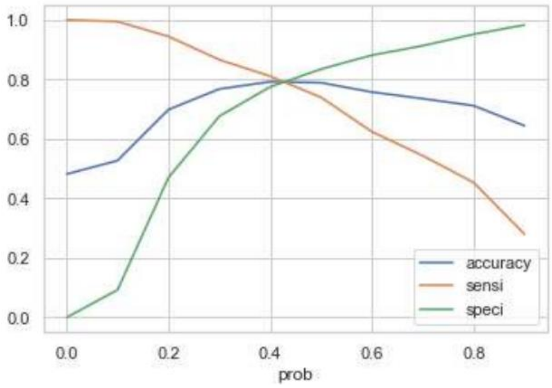
Step 8: Model evaluation

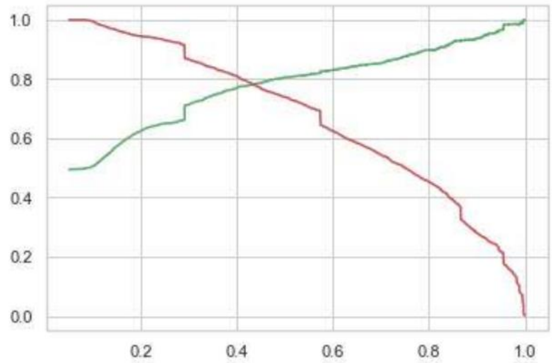
Now, both the p-values and VIFs seem decent enough for all the variables. So let's go ahead and make predictions using this final set of features.
We also create confusion matrix. Using this, we understand overall accuracy, sensitivity and specificity

In order to get good results, we need to optimise the threshold. So first let's plot an ROC curve to see what AUC we get.



Trade-off between Precision and Recall is 0.42





Final Outcomes:

Train Data:

- Accuracy : 80%
- Sensitivity : 77%
- Specificity : 80%

Test Data:

- Accuracy : 80%
- Sensitivity : 77%
- Specificity : 80%

Important columns

- Specialization_Others
- Lead Origin_Lead Add Form
- Total Time Spent on Website
- Lead Origin_Landing Page Submission
- Do Not Email
- Lead Source_Welingak Website
- Lead Source_Olark Chat
- What is your current occupation_Working Professionals

Final Observations

- ✓ The maximum leads are generated by customers using google and by direct traffic.
- ✓ Probability of converting is more when users spend more time on website.
- ✓ Probability of conversion is more with working professionals.
- ✓ Most common last activity is email opened.