Lead Scoring Case Study
Assignment DS C46

Group members:

Mayur Balasaheb Pathak

Umesh Sati

Nagarjuna D

Problem Statement

In this assignment we will be working on the data provided by the Education Institute who sells the online courses to industry professionals.

- 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.
- In order to increase the lead conversion rate, the company first should 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.

Business Objective:

- X education wants to know most promising leads.
- For that they want to build a Model which identifies the hot leads.
- Deployment of the model for the future use

Solutions Methodologies

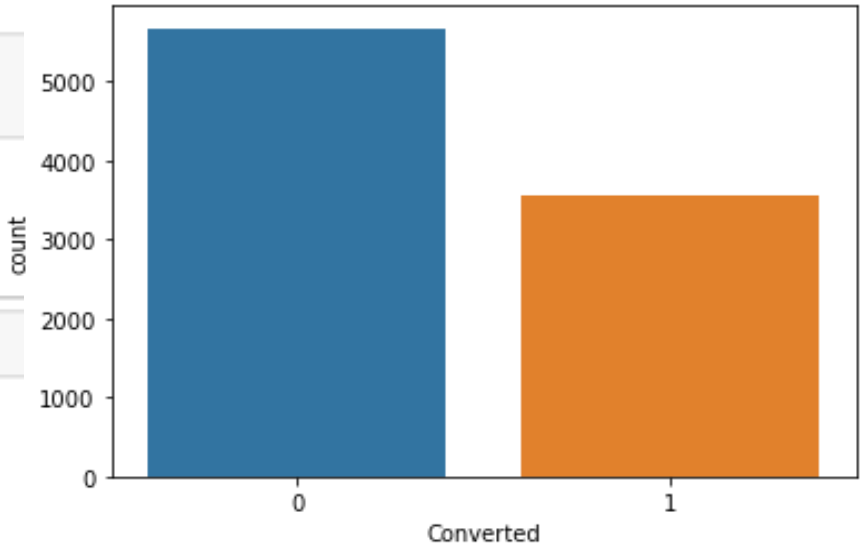
- We will be following the below steps for data cleaning and data manipulation.
  - Checked and handle the missing values
  - Drop the columns having missing values more than 40% and not required for the analysis.
  - Handle the duplicate data
  - Imputation of columns wherever required
  - Uni-variate and Bivariate data analysis
  - Scaling, dummification and encoding of the data.
  - Logistic regression used for the model making and prediction.

EDA
Analysing the Target Column

```
]: 1 #Checking Converted
   2 lead_df['Converted'].value_counts(dropna=False)
```

```
]: 0    5679
   1    3561
   Name: Converted, dtype: int64
```

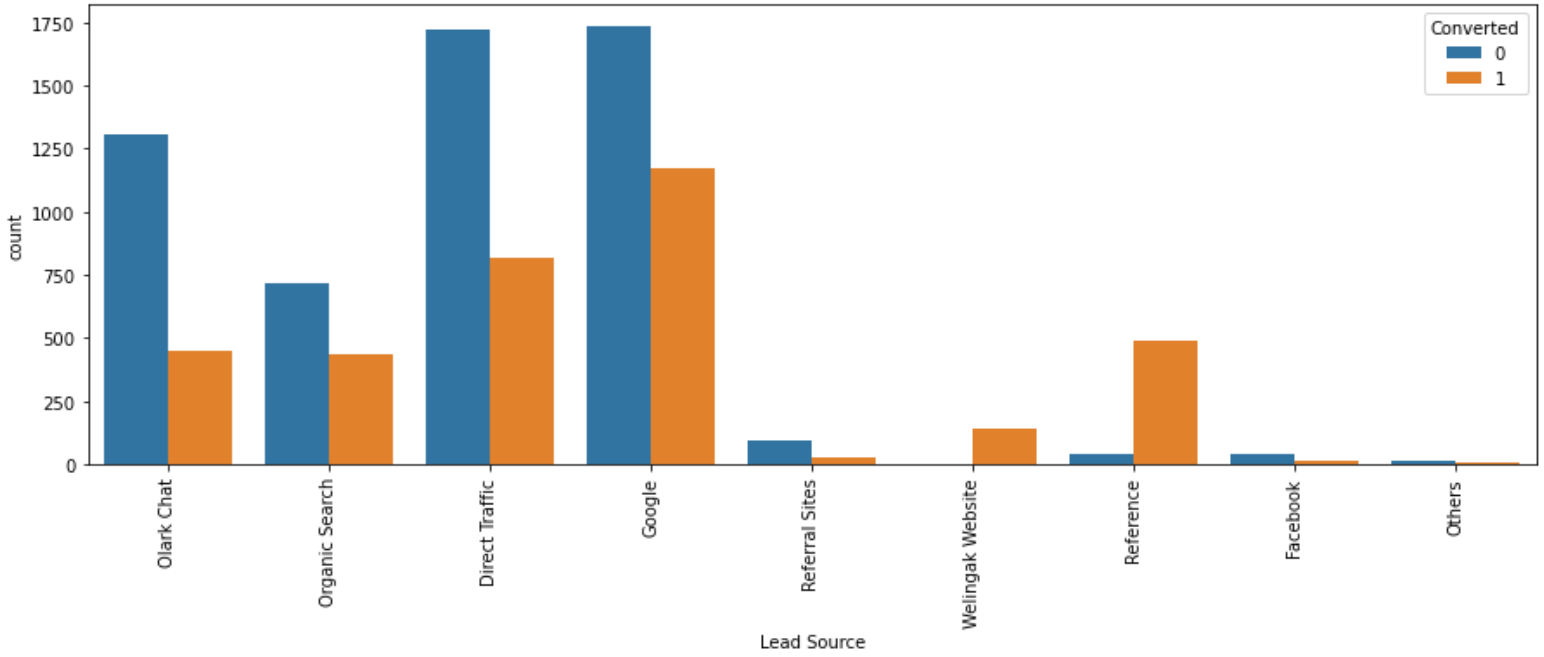
```
]: 1 sns.countplot(lead_df['Converted'])
```



The converted column is divided into two parts. One is converted leads and another one is not converted leads so we are denoting:

- 1 - denotes to the converted leads
- 0 - denotes to the not converted leads

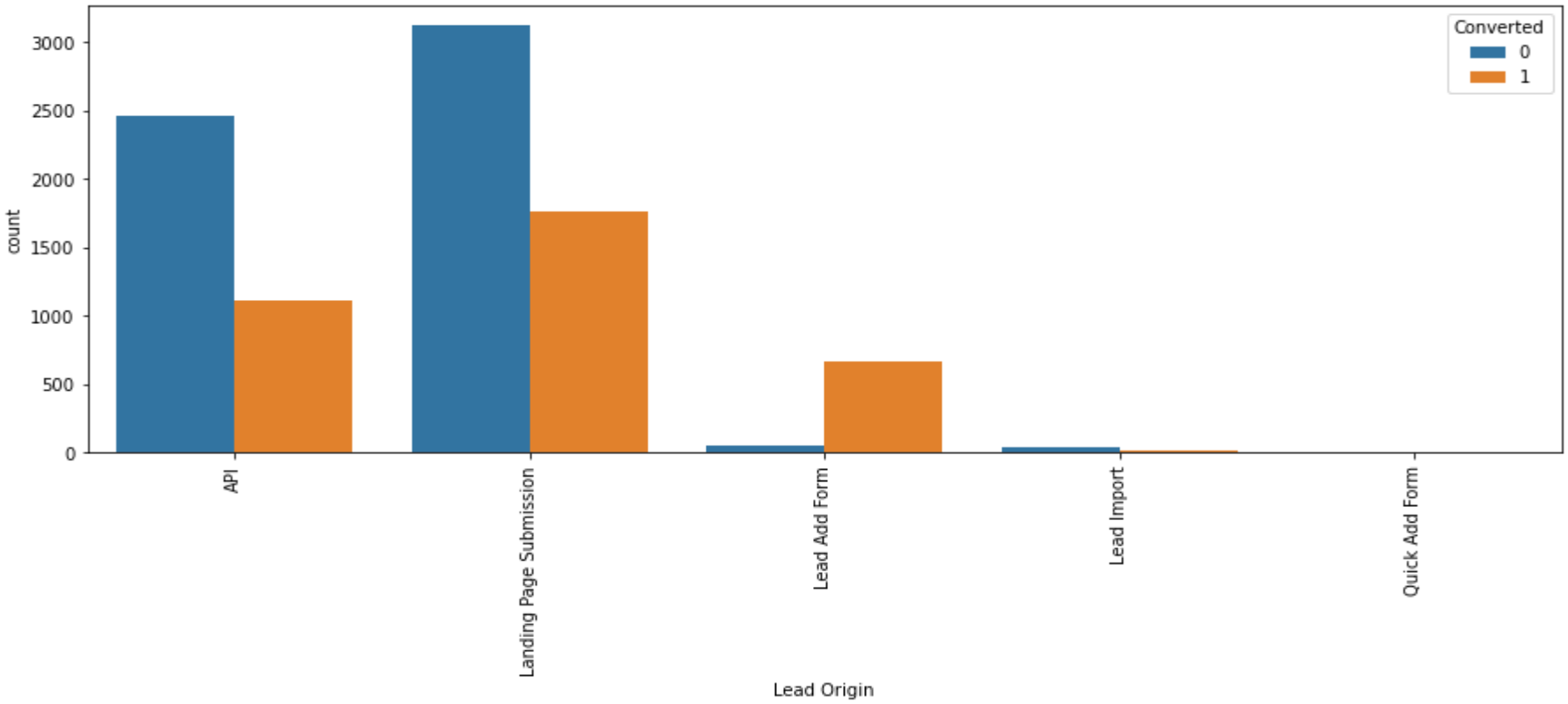
Analysis of Lead Source Column on the basis of Converted and Non Converted data



Insights:

1. From the above chart we can see that Lead generation from Direct Traffic and Google is more comparatively to others
2. Lead conversion rate and number of leads from References and Welingak Website is more than others

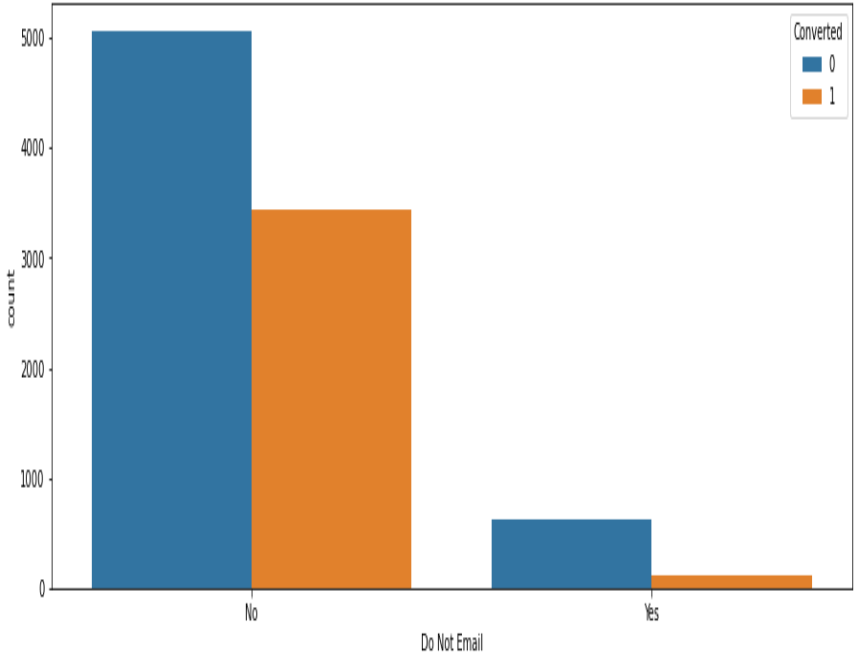
Analysis of Lead Source Column on the basis of Converted and Non Converted data

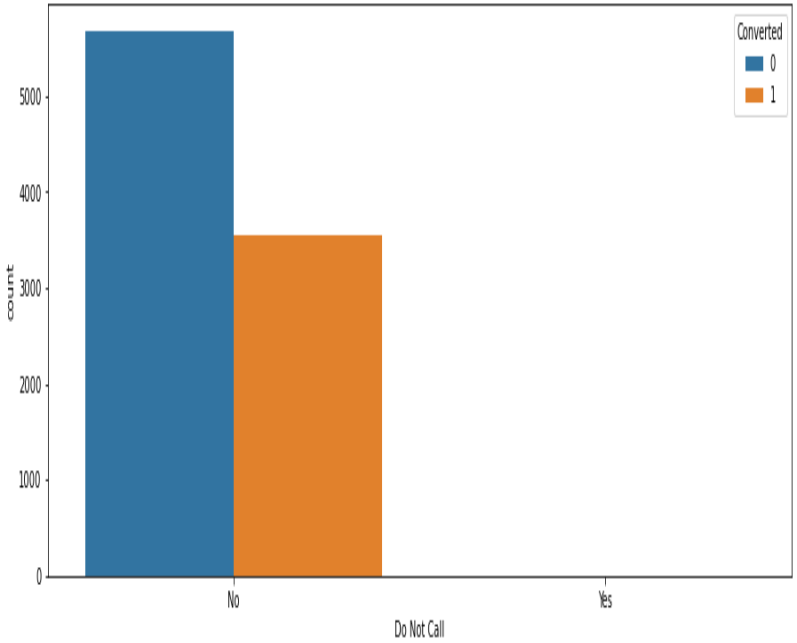


Insights:

1. API and Landing Page Submission required has more lead origins
2. Leads converted more from Lead Add Form

Analysis of Do Not Email and Do Not Call column on the basis of Converted and Non Converted data





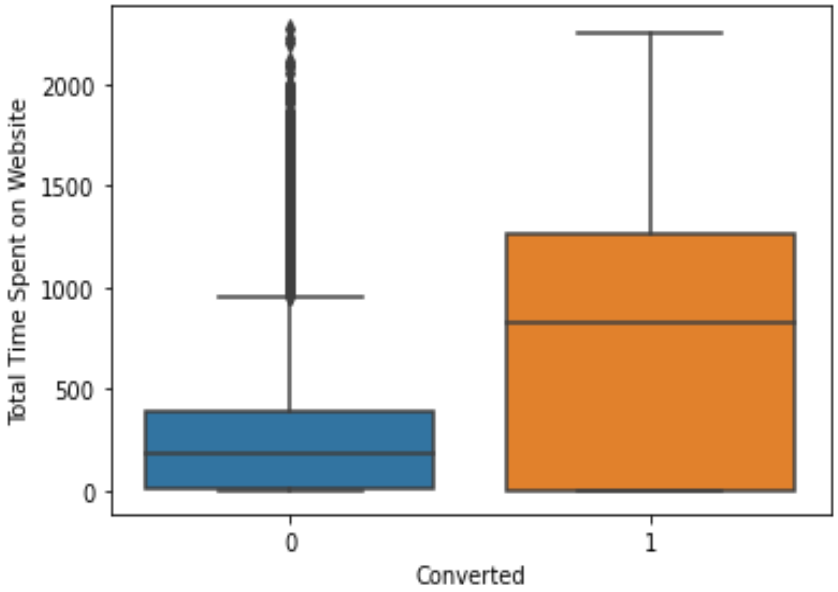
Insights:

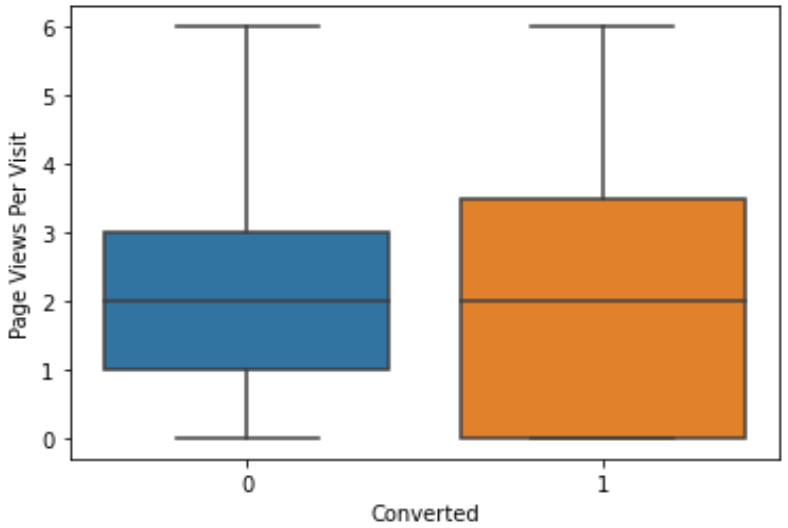
- From the above chart we can come to conclusion that email has higher count for creating leads and conversion also.
- So we can recommend to focus more on emails.

Insights:

- Do Not Call data is highly skewed column.
- So we are going to drop this column to make the data better

Analysis of Total Time Spent & Page Views Per Visit on the basis of Converted and Non Converted data





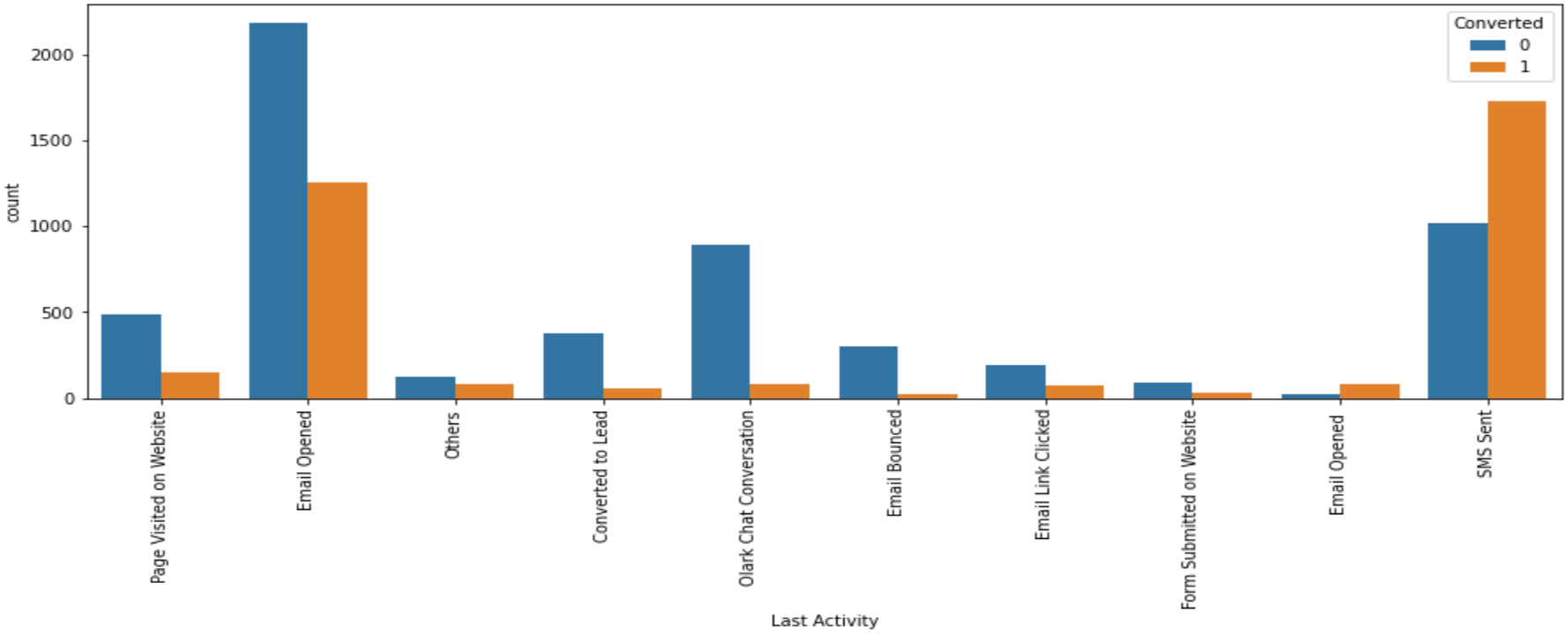
Insights:

- In the above box plot we analyzed the Total Time Spent column on the basis of Converted and Non Converted data and we can see that, who spent more time on website are likely to convert so website interface is more user friendly and informative.

Insights:

- Here we analyzed the pages views per visits column on the basis of Converted and Non Converted data and we can see that, median for converted and non converted is same so Page Views Per Visit column is non conclusive.

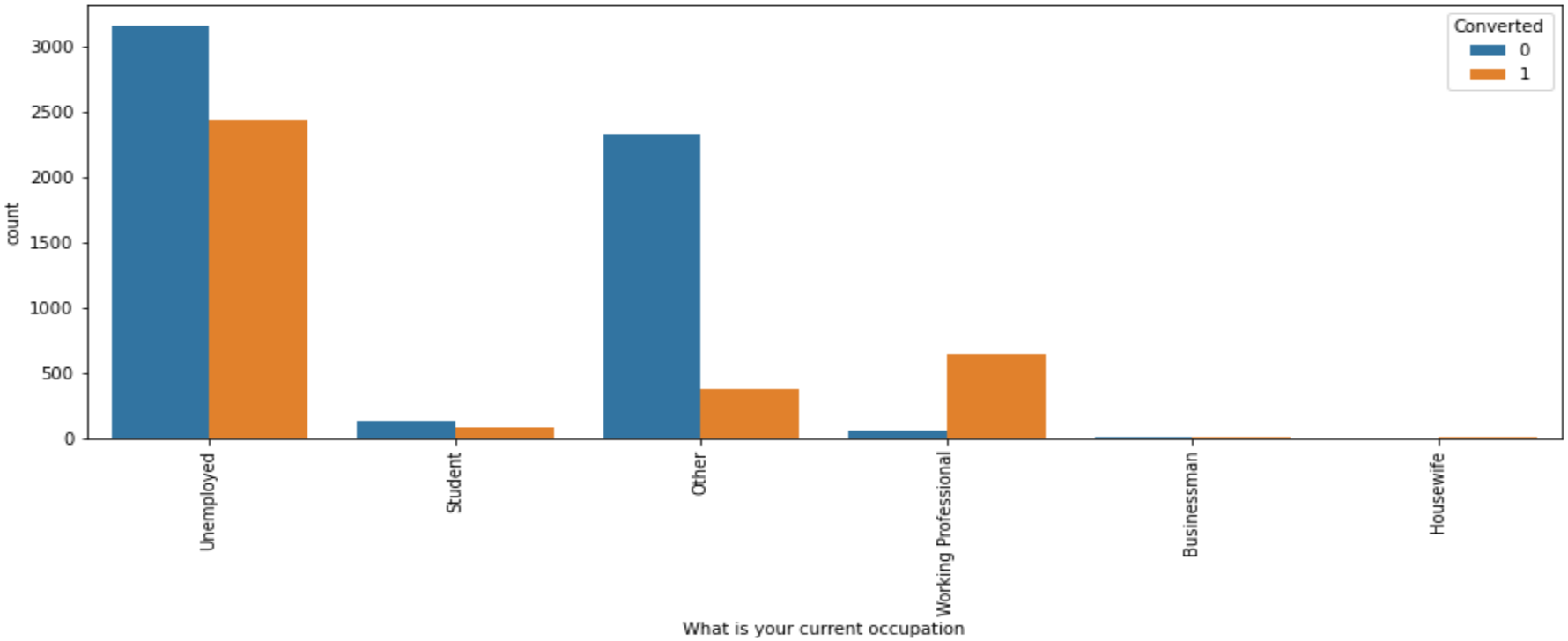
Analysis of Last Activity column on the basis of Converted and Non Converted data



Insights:

- We can observe that Email opened has highest leads as compare to others.
- Whereas the lead conversion is more with SMS Sent

Analysis of “What is your current occupation” column on the basis of Converted and Non Converted data



Insights:

- We can observe that Unemployed has generated more leads and conversion is also good
- The conversion is more with working professional and lead generations is less so the focus should be more on this category.

Model Building

Dep. Variable:	Converted	**No. Observations:**	6372
Model:	GLM	**Df Residuals:**	6359
Model Family:	Binomial	**Df Model:**	12
Link Function:	Logit	**Scale:**	1.0000
Method:	IRLS	**Log-Likelihood:**	-2727.7
Date:	Mon, 23 Jan 2023	**Deviance:**	5455.5
Time:	16:51:48	**Pearson chi2:**	6.56e+03
No. Iterations:	7	**Pseudo R-squ. (CS):**	0.3774
Covariance Type:	nonrobust		

	coef	std err	z	P>|z|	[0.025	0.975]
const	-0.4752	0.135	-3.528	0.000	-0.739	-0.211
Do Not Email	-1.6752	0.184	-9.107	0.000	-2.036	-1.315
Total Time Spent on Website	1.0858	0.039	27.586	0.000	1.009	1.163
Lead Origin_Landing Page Submission	-0.9611	0.128	-7.495	0.000	-1.212	-0.710
Lead Origin_Lead Add Form	3.7027	0.229	16.144	0.000	3.253	4.152
Lead Source_Olark Chat	0.9847	0.117	8.437	0.000	0.756	1.213
Lead Source_Welingak Website	2.3053	1.038	2.220	0.026	0.270	4.340
Last Activity_Email Opened	0.5710	0.083	6.853	0.000	0.408	0.734
What is your current occupation_Other	-1.2577	0.088	-14.332	0.000	-1.430	-1.086
What is your current occupation_Student	-0.1123	0.223	-0.503	0.615	-0.550	0.325
City_Others	-0.9570	0.125	-7.638	0.000	-1.203	-0.711
Last Notable Activity_Others	2.0171	0.276	7.297	0.000	1.475	2.559
Last Notable Activity_SMS Sent	1.9972	0.093	21.487	0.000	1.815	2.179

- In the above images we can see the Important factors for lead conversion

Confusion matrix for train

```
[[3412  526]
 [ 730 1704]]
```

Accuracy of train:80.39%

Confusion matrix for test

```
array([[1400,  304],
       [ 215,  812]]
```

Accuracy of train:80.99%

- **These are top three variables that contributed to the lead conversion:**
    1. Lead Origin_Lead Add Form
    2. Lead Source_Welingak Website
    3. Last Notable Activity_Others

- **These are the factors that we look for higher conversion rates:**
    1. 1. Refererals and Welingak Website
    2. 2. Lead Origin_Lead Add Form
    3. 3. Working Professional
    4. 4. Last_Activity SMS sent
    5. 5. From city like Mumbai
    6. 6. Last Notable Activity_Others