Lead Scoring

OAISWARI SHAW

10/18/2022

Background

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.

Objective

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 customers. 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.

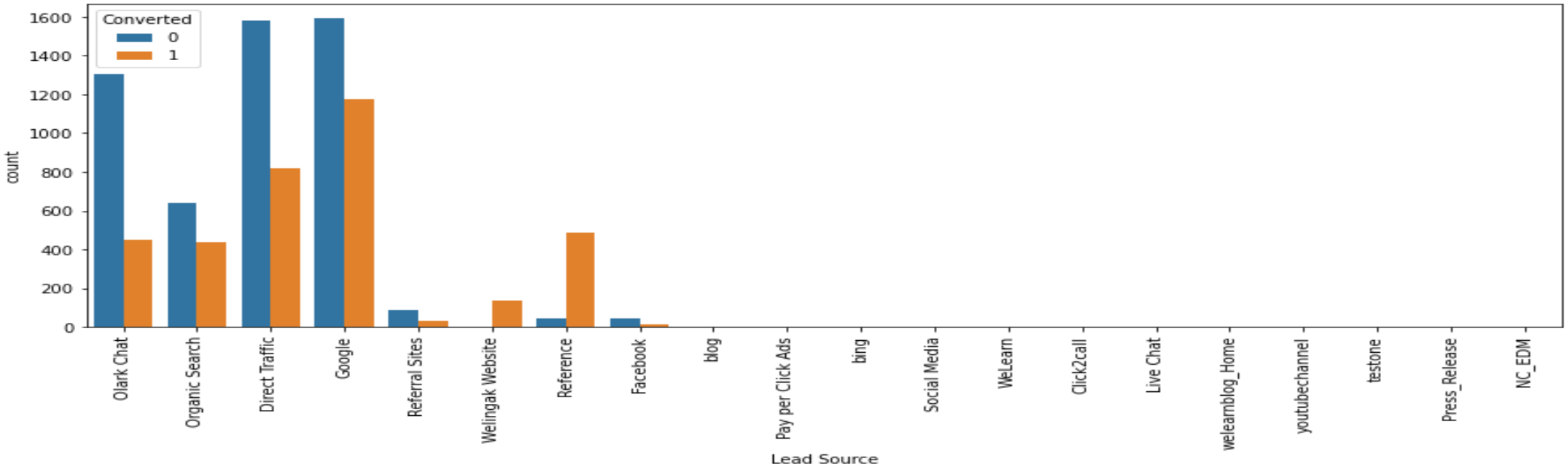
Data Understanding (1)

- A total of 9240 samples were given in input file with 37 features. Most of the features (30/37) are of categorical types.
- The columns **'Prospect ID**' and **'Lead Number**' are unique identifier of each row. Hence they can be dropped being included in training data.

```
In [4]: data.shape
Out[4]: (9240, 37)

In [67]: data['Prospect ID'].nunique()
Out[67]: 9240

In [68]: data['Lead Number'].nunique()
Out[68]: 9240

In [70]: categorical = data.select_dtypes('object')
         print("categorical columns: {}".format(len(categorical.columns)))
         print("numerical columns: {}".format(len(data.columns) - len(categorical.columns)))

         categorical columns: 30
         numerical columns: 7
```

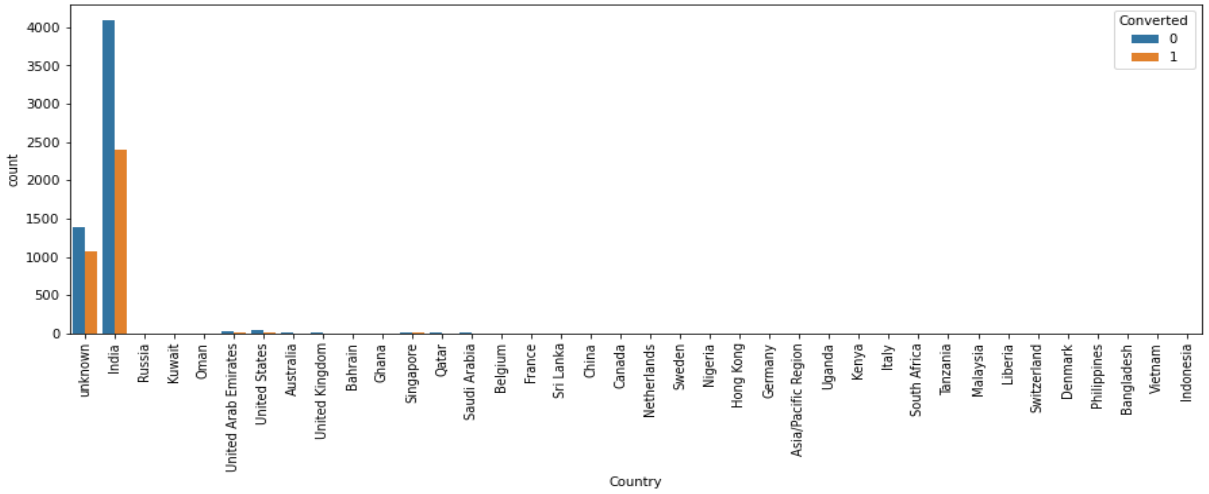
Data Understanding (2)

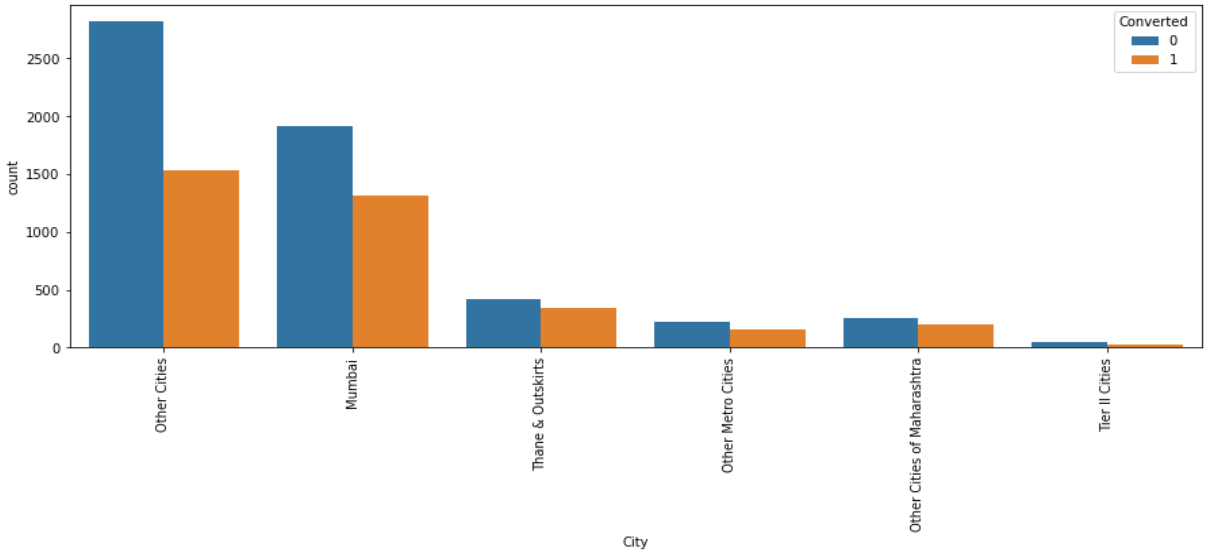
- Maximum Leads are generated by Google and Direct Traffic.
- Conversion rate of Reference leads and Welinkgak Website leads are also very high.



Data Understanding (3)

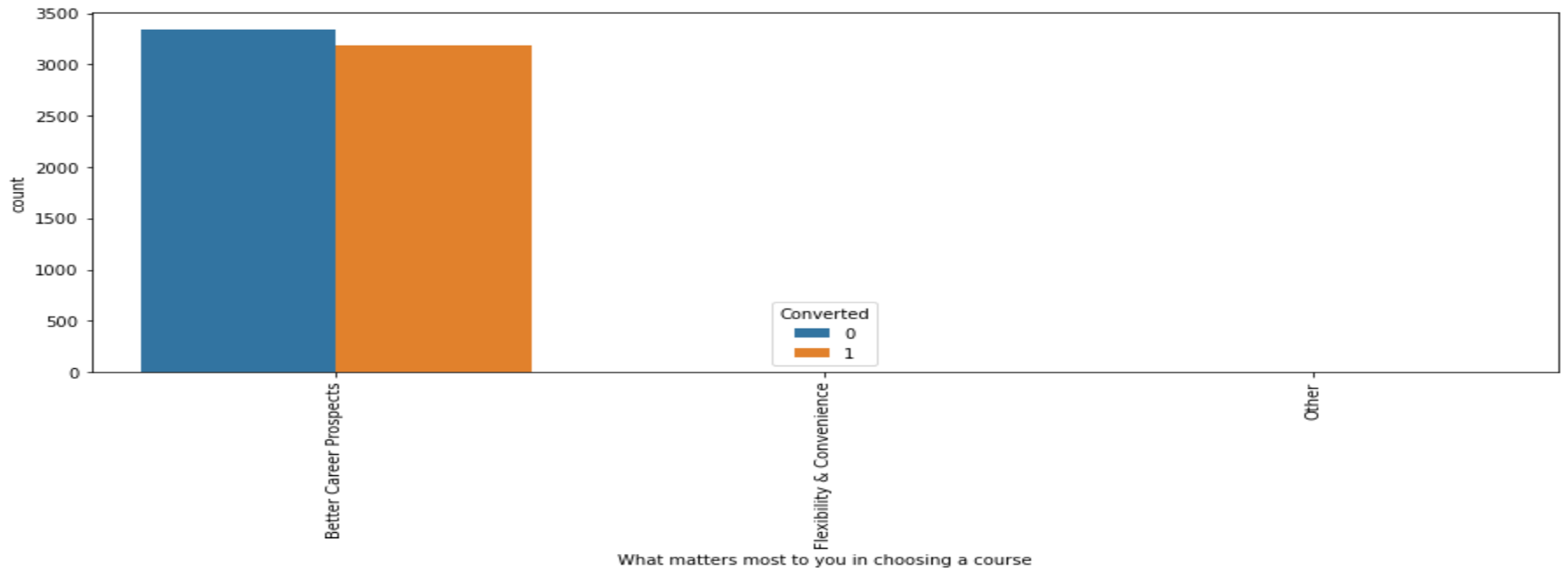
- Majority of the leads are from India.
- From the count plot of city, we can see that maximum leads are from Mumbai.





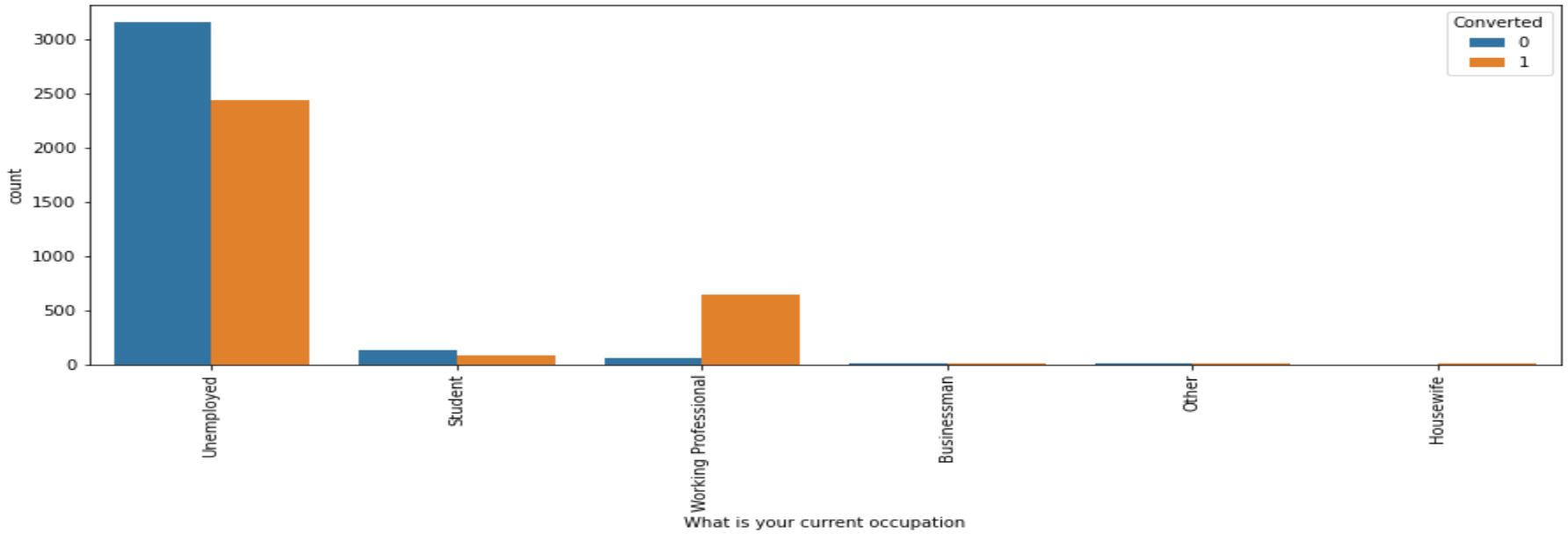
Data Understanding (4)

- For Most of the leads, having better career prospects, matters the most while choosing a course.



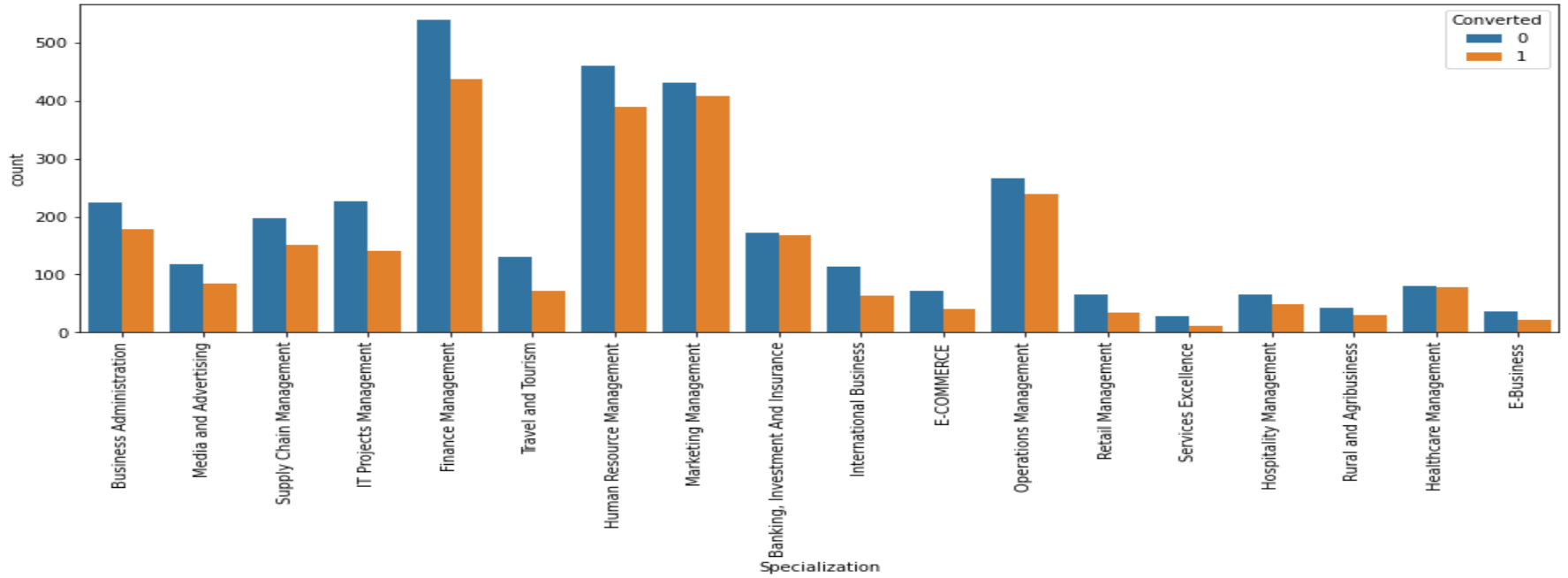
Data Understanding (5)

- The current occupation of maximum number of leads is “Unemployed” followed by “Working Professionals”. Hence they can be targeted.



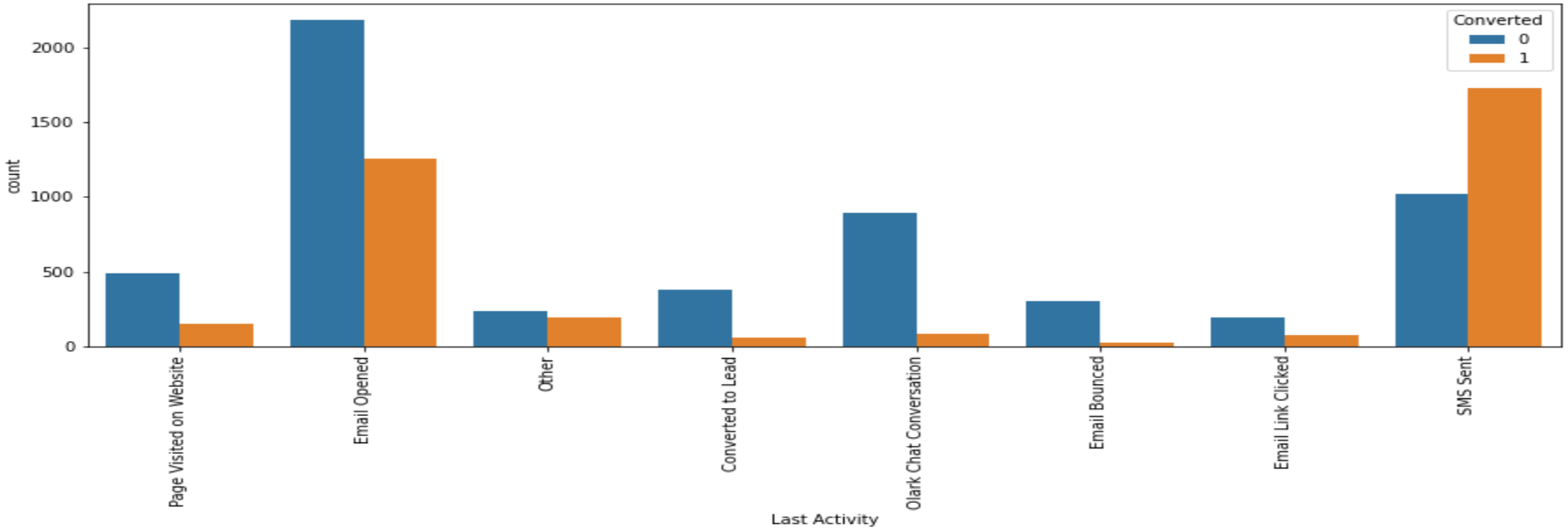
Data Understanding (6)

- Majority of the leads are specialised in Finance Management, followed by Human Resource & Manufacturing Management.



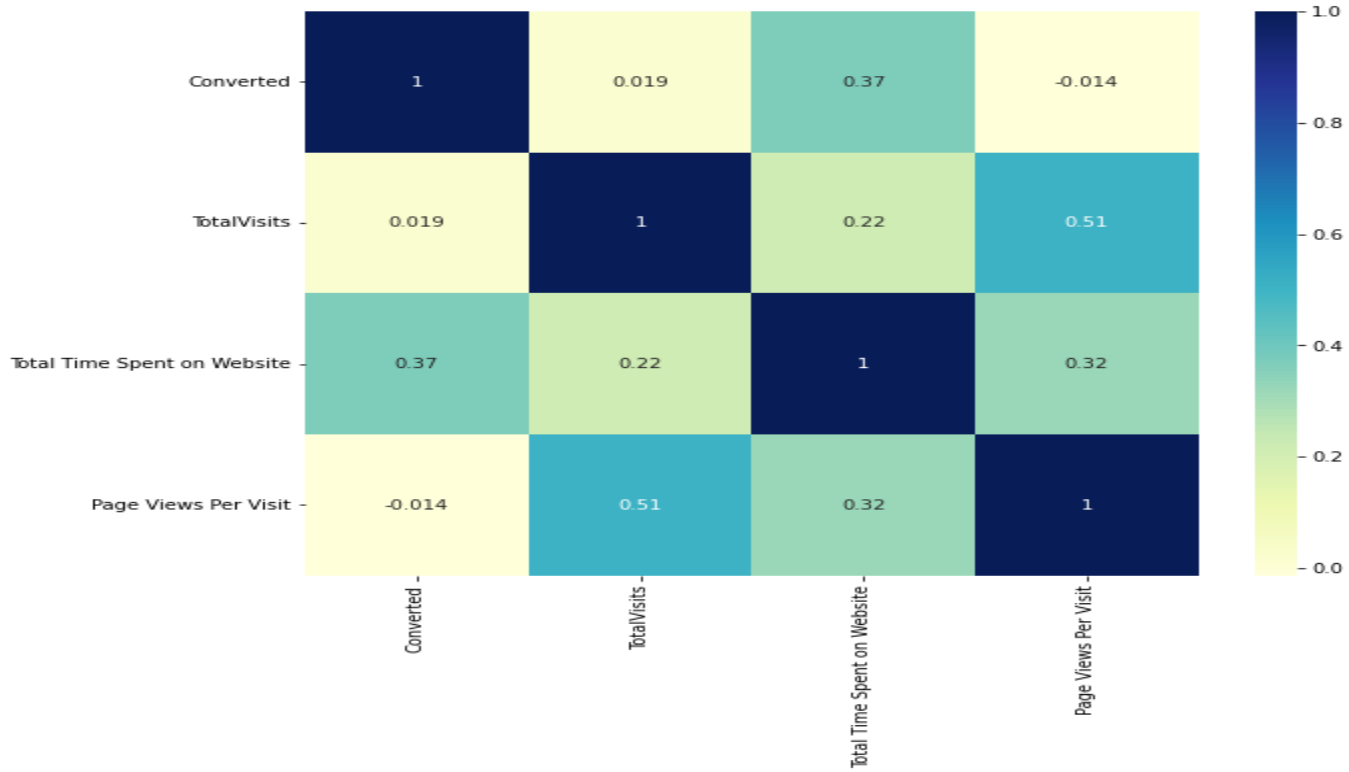
Data Understanding (7)

- Majority of the leads were the customers tagged as -will revert after reading the email.
- Maximum leads are generated having last activity as Email opened but conversion rate is not too good.
- SMS sent as last acitivity has high conversion rate.



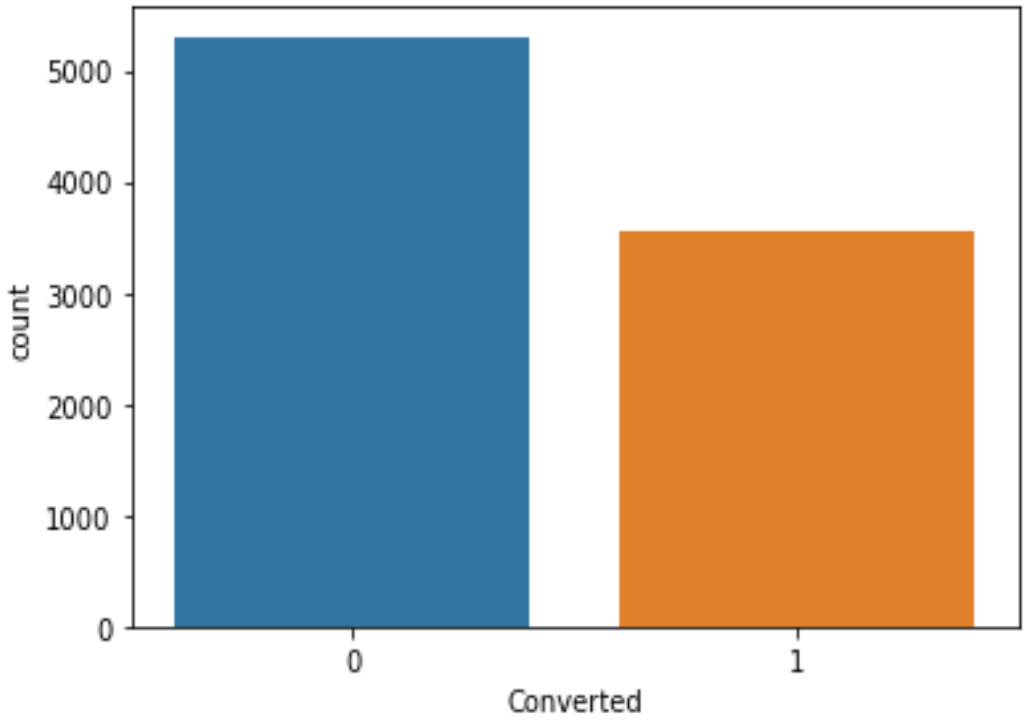
Data Understanding (8)

- Total Time Spent on Website is highly corelated with Converted. Thus website should be made more engaging to increase conversion rate.



Data Understanding (9)

- The data is fairly balanced for predictive analysis.



Model Finding

- The below features found to be most important

```
In [70]: #coefficient analysis
         df = pd.DataFrame()
         df["col"] = col
         df["coef"] = m1.coef_[0]
         df.sort_values("coef", ascending=False)
```

Out[70]:

	col	coef
14	Tags_Closed by Horizzon	5.737972
19	Tags_Lost to EINS	5.461410
24	Tags_Will revert after reading the email	2.581211
13	Tags_Busy	2.149123
2	Lead Origin_Lead Add Form	1.858813
4	Lead Source_Welingak Website	1.507185
9	Last Activity_SMS Sent	1.424546
18	Tags_Lateral student	1.280821
25	Tags_in touch with EINS	1.137689
0	Total Time Spent on Website	1.085640
10	Country_unknown	1.035403

Achievement

- Over 88% accuracy is achieved for both in-sample and out-sample prediction with very good precision, recall values.
- In other words, the trained model is able to produce the expected target set by the CEO (80%).

```
In [67]: from sklearn import metrics
         print(metrics.classification_report(in_sample, y_train))
```

	precision	recall	f1-score	support
0	0.90	0.90	0.90	3717
1	0.85	0.86	0.85	2489
accuracy			0.88	6206
macro avg	0.88	0.88	0.88	6206
weighted avg	0.88	0.88	0.88	6206

```
In [68]: from sklearn import metrics
         print(metrics.classification_report(out_sample, y_test))
```

	precision	recall	f1-score	support
0	0.91	0.90	0.91	1639
1	0.84	0.86	0.85	1021
accuracy			0.89	2660
macro avg	0.88	0.88	0.88	2660
weighted avg	0.89	0.89	0.89	2660