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.

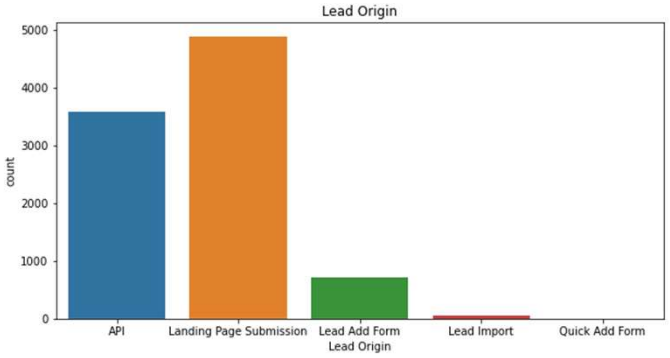
The company markets its courses on several websites and search engines like Google. Once these people land on the website, they might browse the courses or fill up a form for the course or watch some videos. When these people fill up a form providing their email address or phone number, they are classified 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:

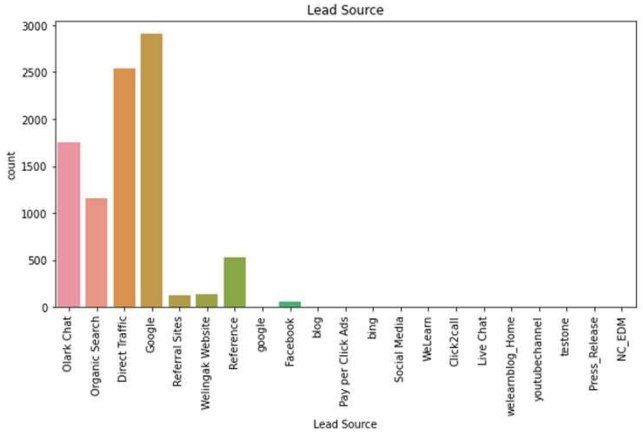
Goal

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.

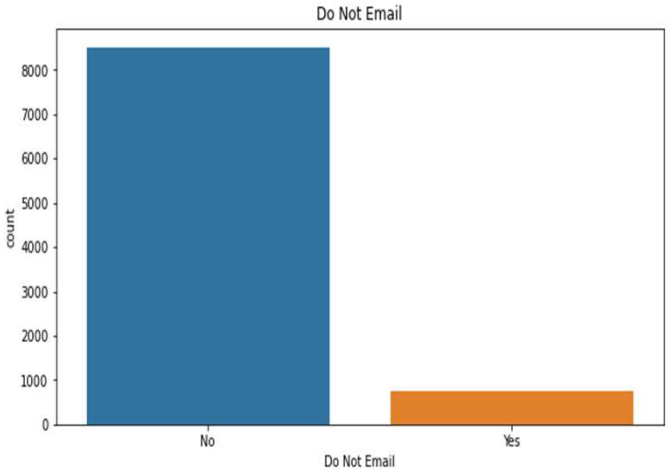
Univariate Analysis Categorical Variables



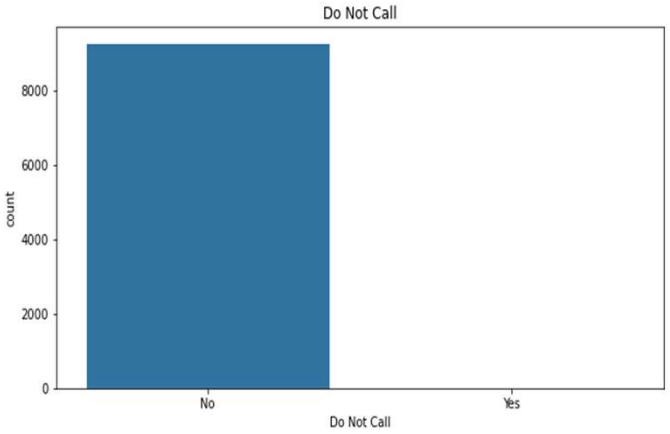
Landing Page Submission has highest count followed by API.



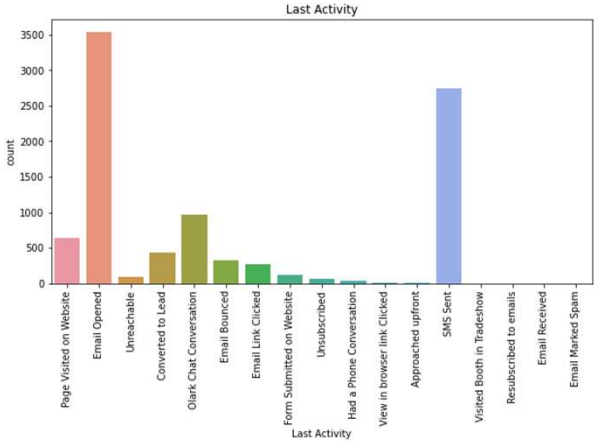
Most of the people came via Google followed by Direct Traffic source.



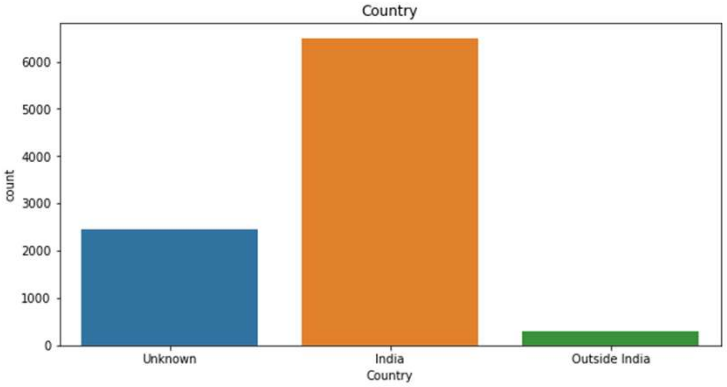
Most of the people have choose 'Don't Email' option from this dataset.



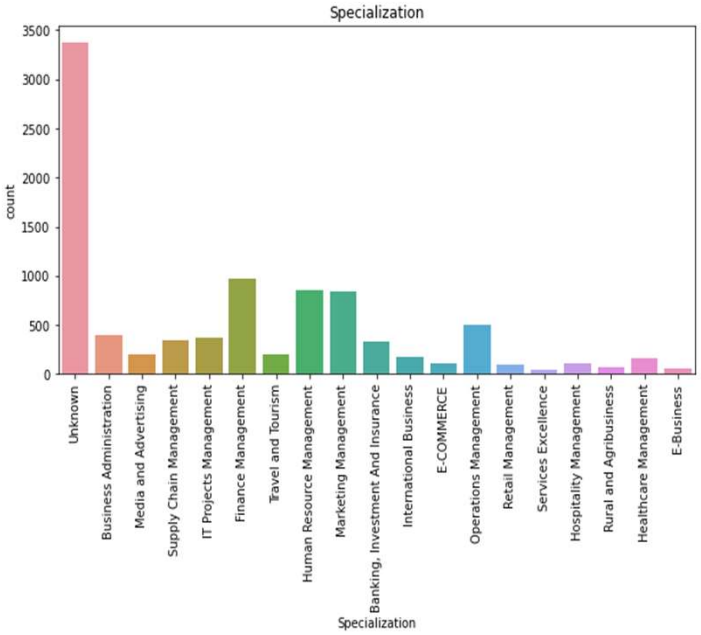
Most of the people have choose 'Don't Call' option from this dataset.



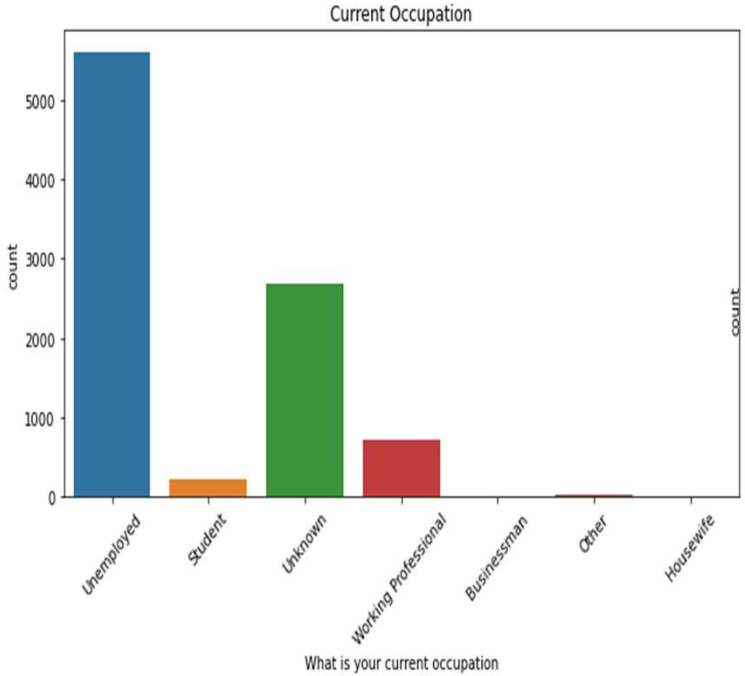
Majority of people have atleast opened the email or read the sms sent.



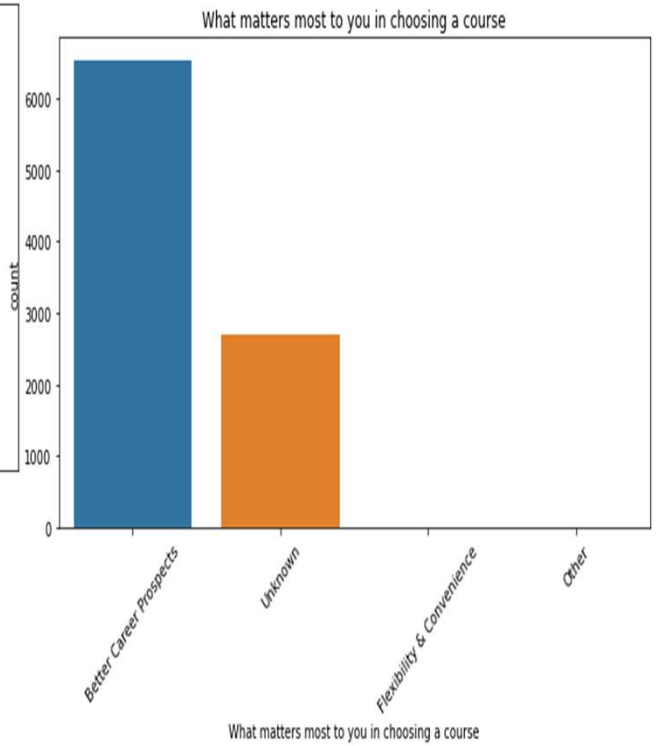
Most of the people are from India



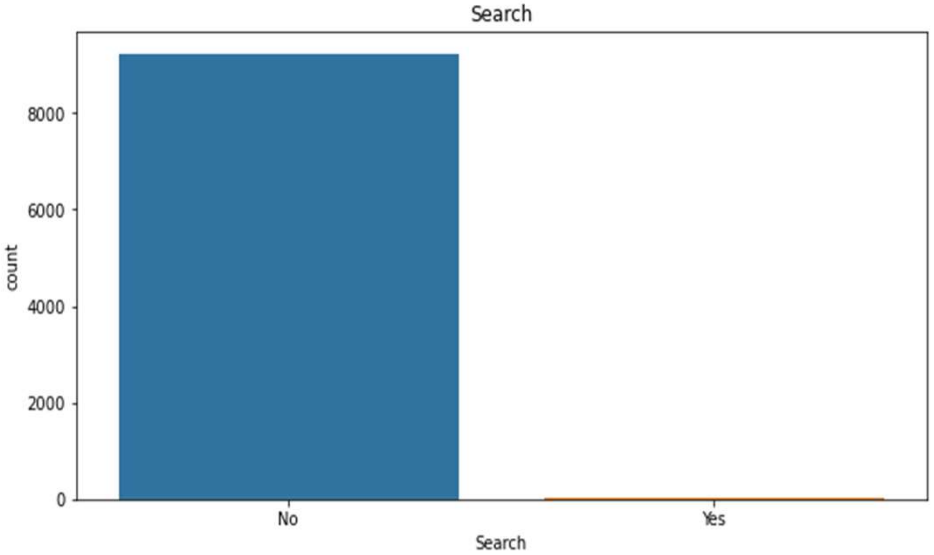
•People having management profession in any genre are more likely to be a lead.
• Most of the people didn't mentioned their specialization.



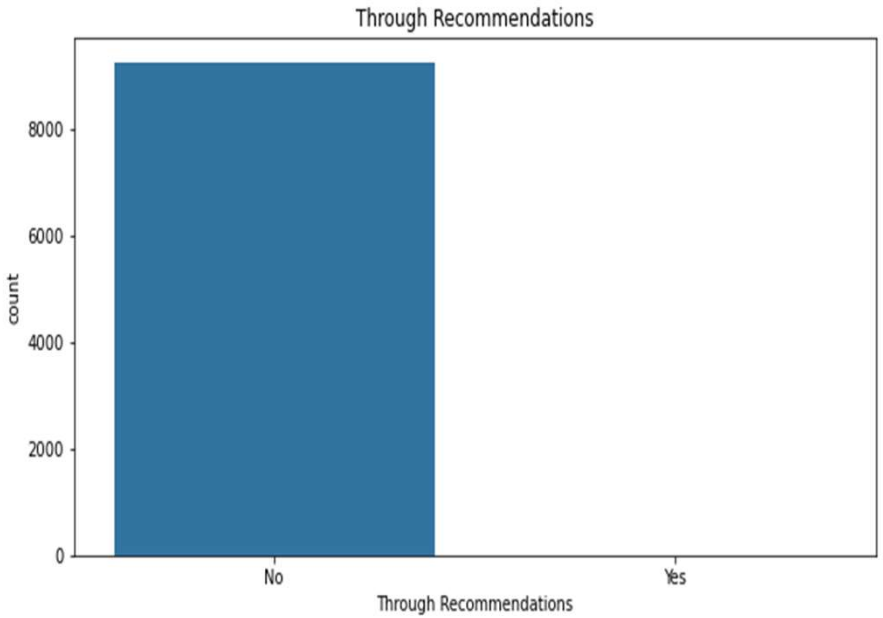
Most of the people are unemployed.



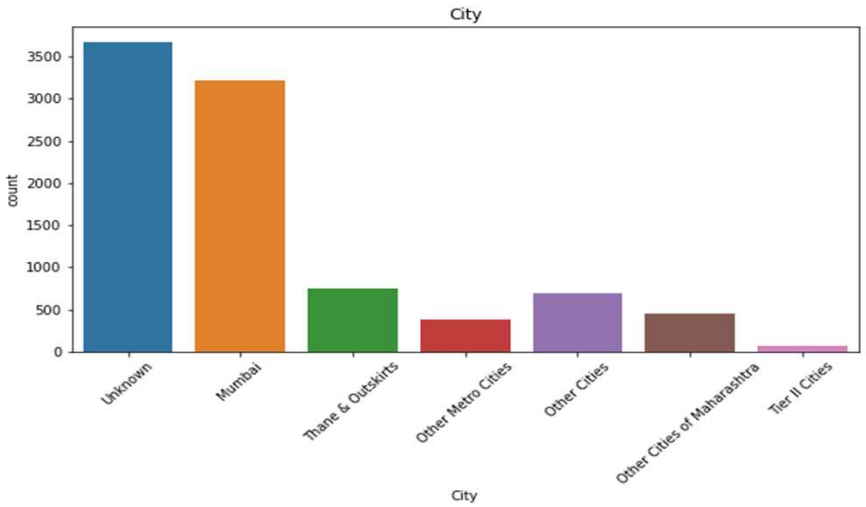
Most of the people wants Better Career Prospects from the chosen course



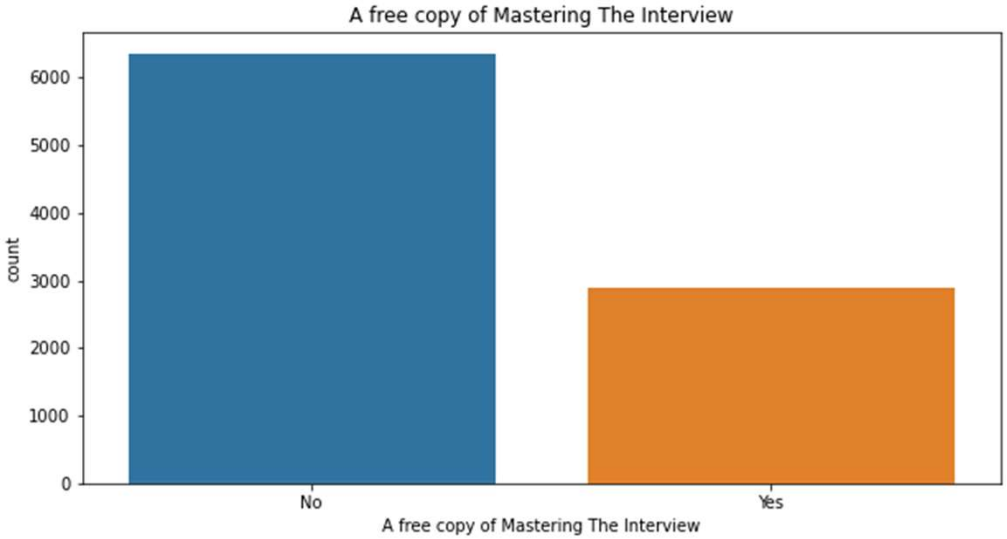
Majority of people haven't searched about any of the ads.



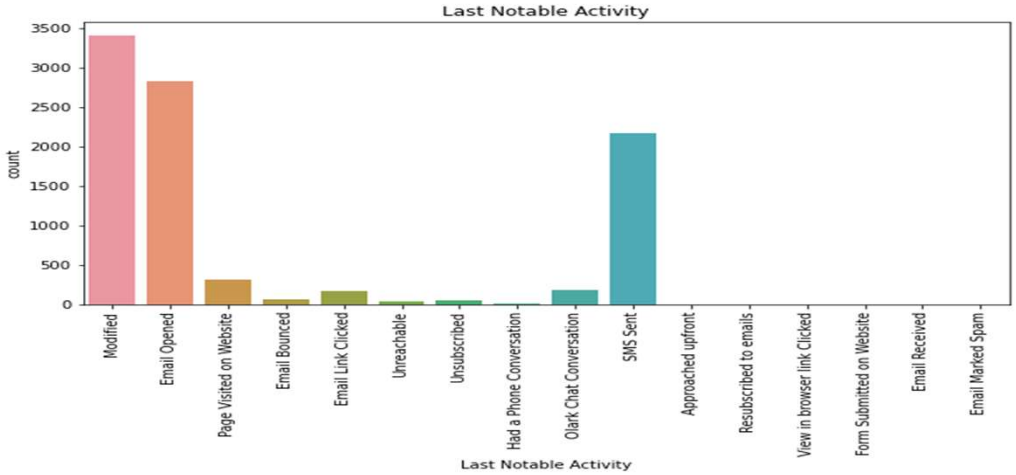
Majority of people haven't seen any ads regarding course in Newspaper Article



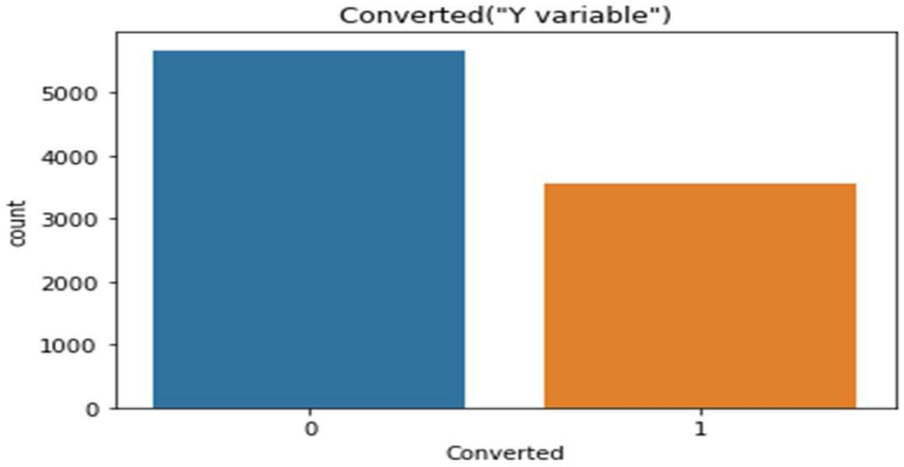
People from Mumbai are majority in number.



Most of the People wanted a free copy of Mastering the Interview from X education.

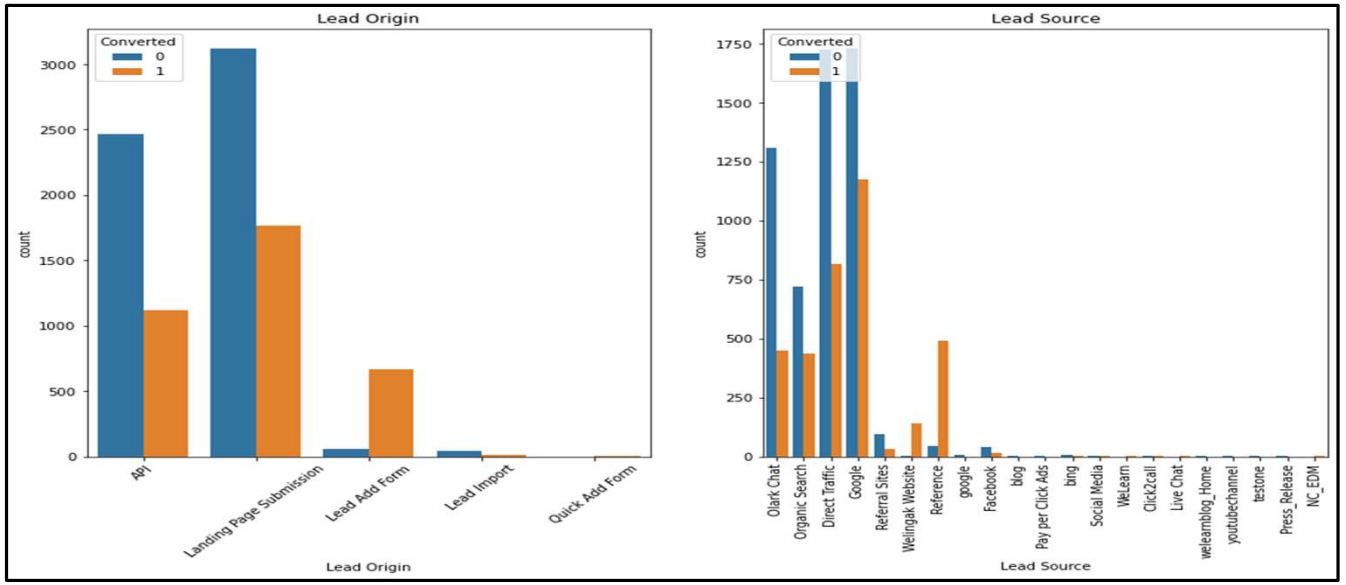


Counts of activities like Modified, Email Opened and sms Sent are high compared to other activities.



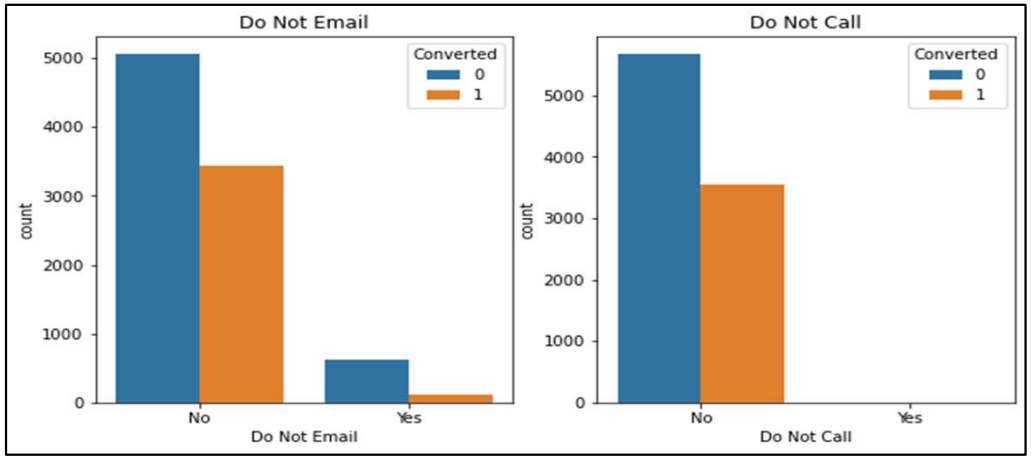
Around 5700 people who visited X education site haven't join any course whereas around 3500 people have got converted to leads by joining any course from X education.

Relating all the categorical variables to Converted



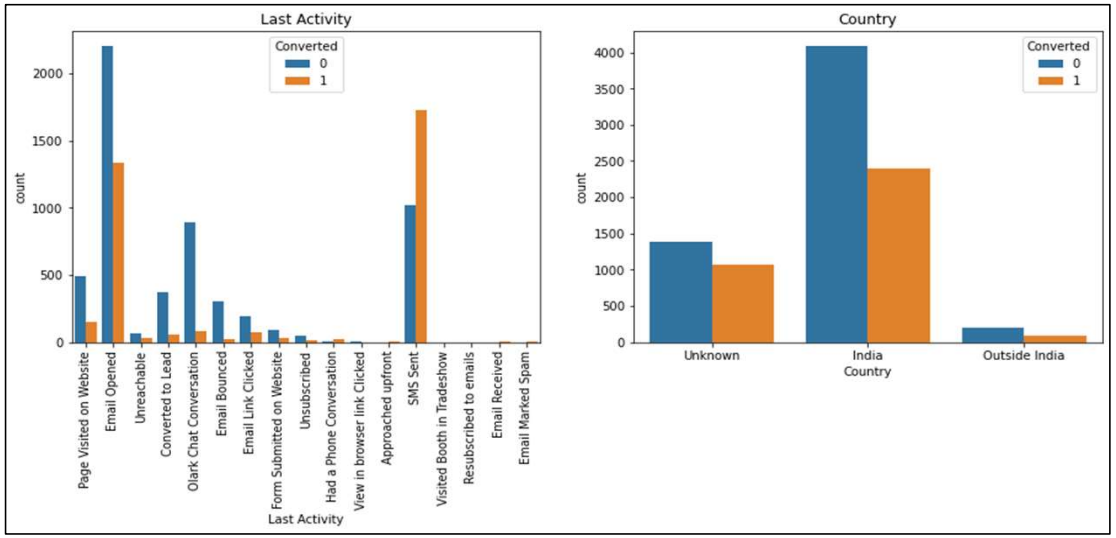
Insights

1. Added form is more effective way to convert people but it is significantly less in count.
2. Landing Page Submission has highest count of people who did'nt convert. Still it is second best effective way to convert people.
3. Reference helps most in converting people followed by Google.
4. Olark chat and refferal sites perform lowest in conversion of people.



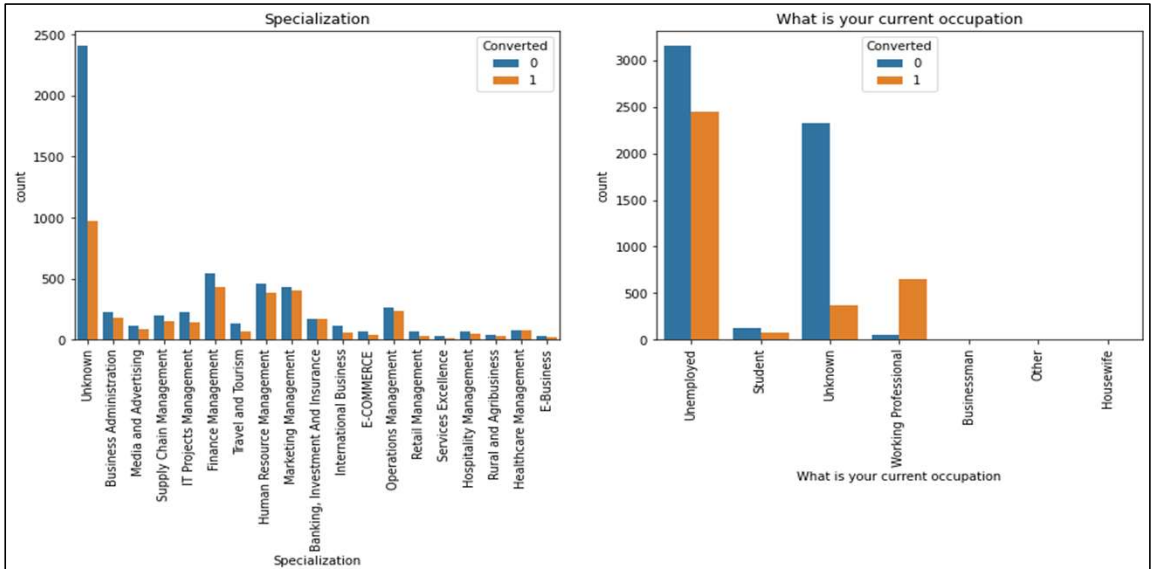
insights

People who optioned out for no email and no call are having high chances of getting converted to join any course.



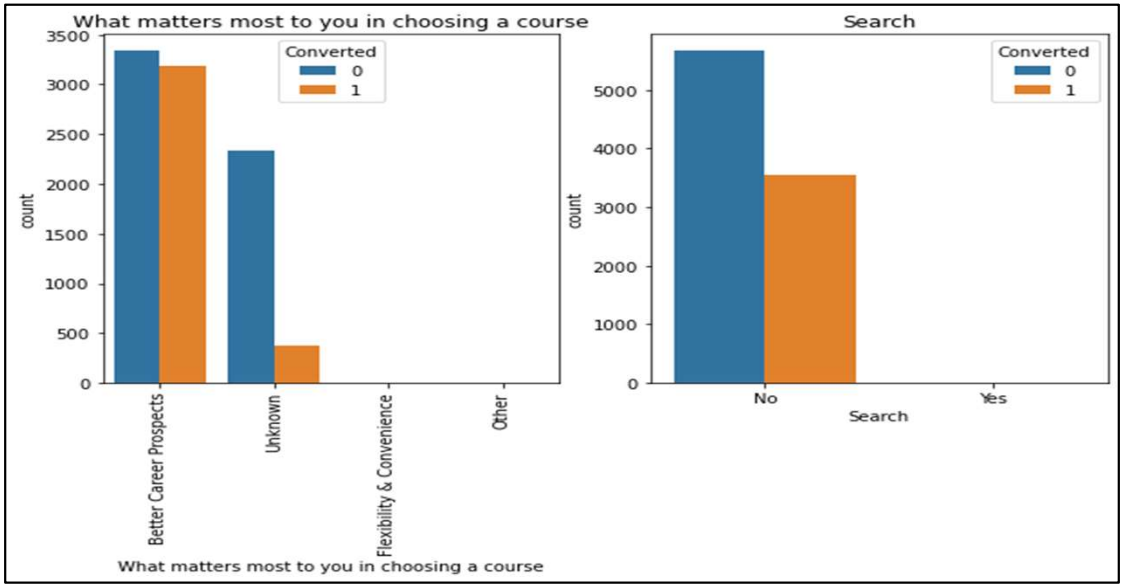
<u>Insights</u>

1. SMS sending have very good response from people which relfects in the conversion count. inference
2. Email opened activity has less but good reponse from people in conversion count.
3. Indian people are showing positive response in conversion count compared to out of India countries.

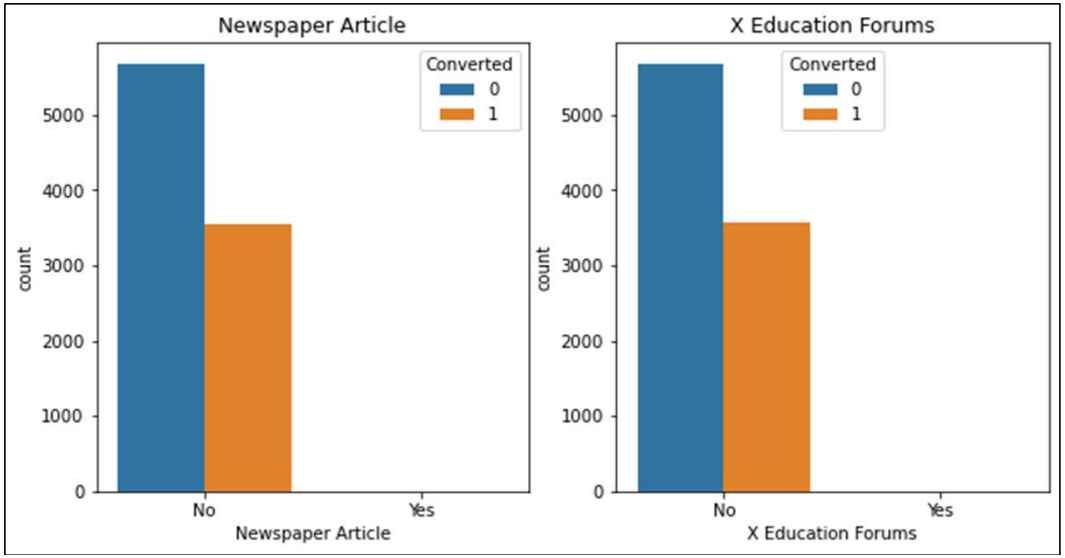


<u>Insights</u>

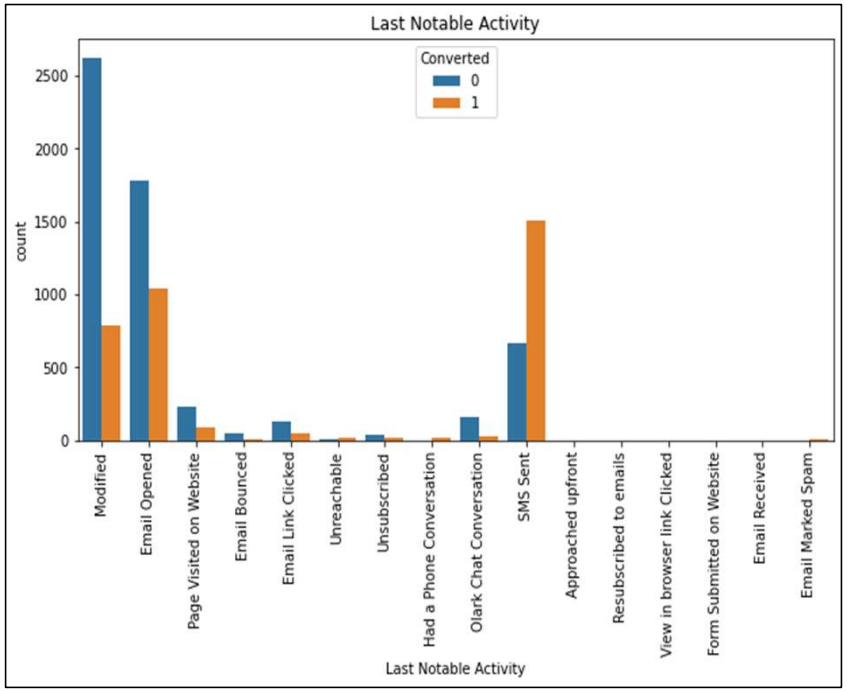
1. Management professions like Finance, HR, Marketing and Operations have very good count of conversion compared to other specializations.
2. Working profession shows excellent count of conversion whereas unemployed people have higher count for being converted.



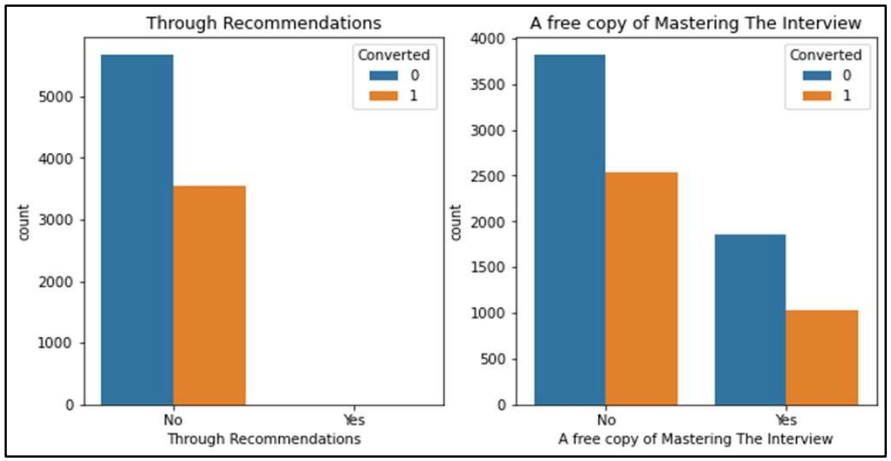
Insights
1. People asking for Better Career Prospects shows highly positive response in conversion.
2. People who didn't search about X Education courses are having good chances for conversion



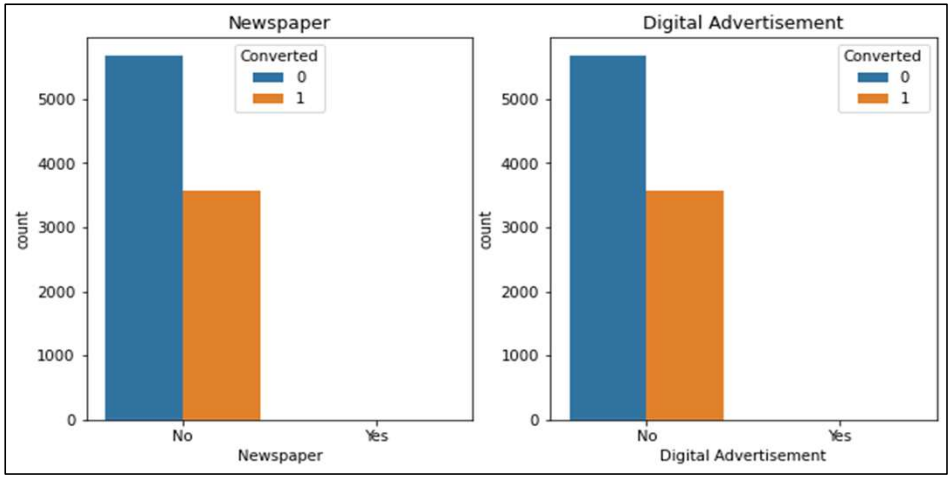
People who haven't seen ads on Newspaper Articles and X Education Forum has good conversion rate but still lower than non conversion rate.



Sms Sent have highest conversion count compared to other activities followed by Email Opened.

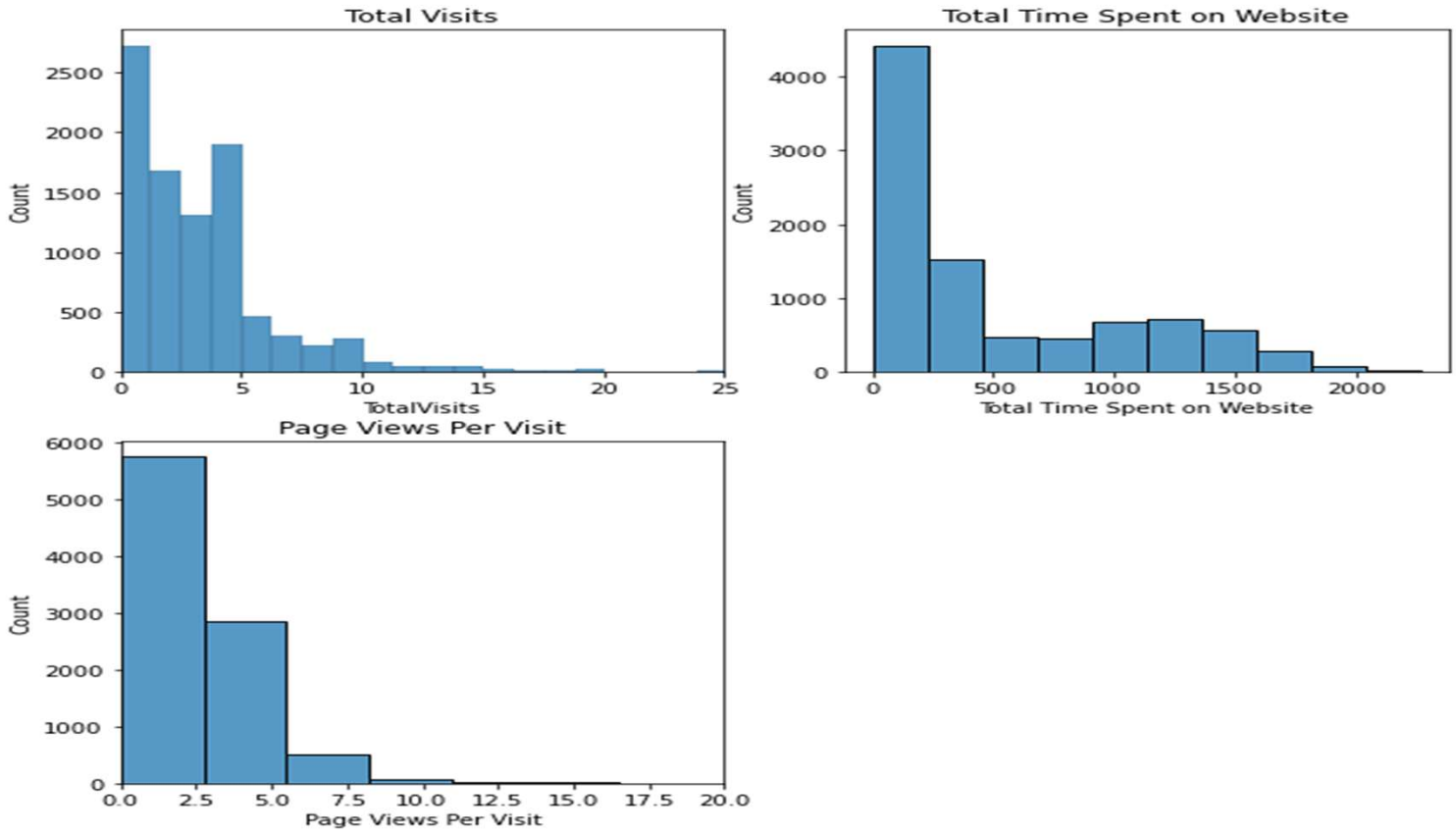


People who haven't seen any ads through Recommendation as well as didn't demand for a free copy of Mastering the Interview have good count of conversion above 2400.

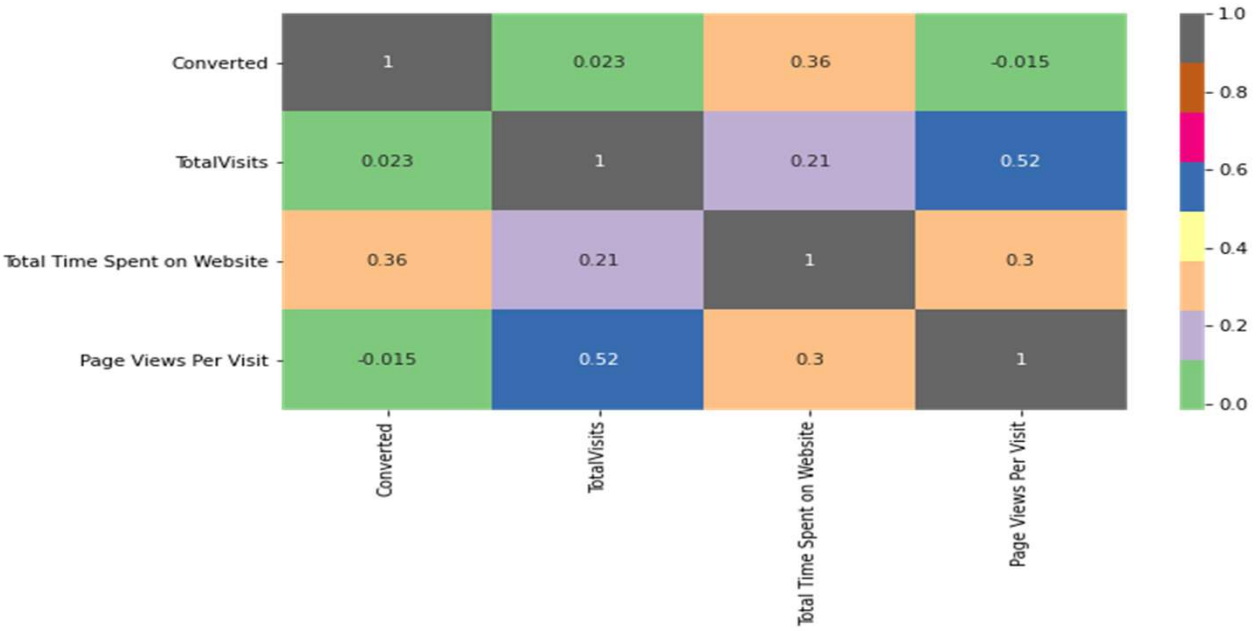


People who haven't seen ads in Newspaper and Digital Advertisement has good conversion rate but still lower than non conversion rate.

Numerical Variables

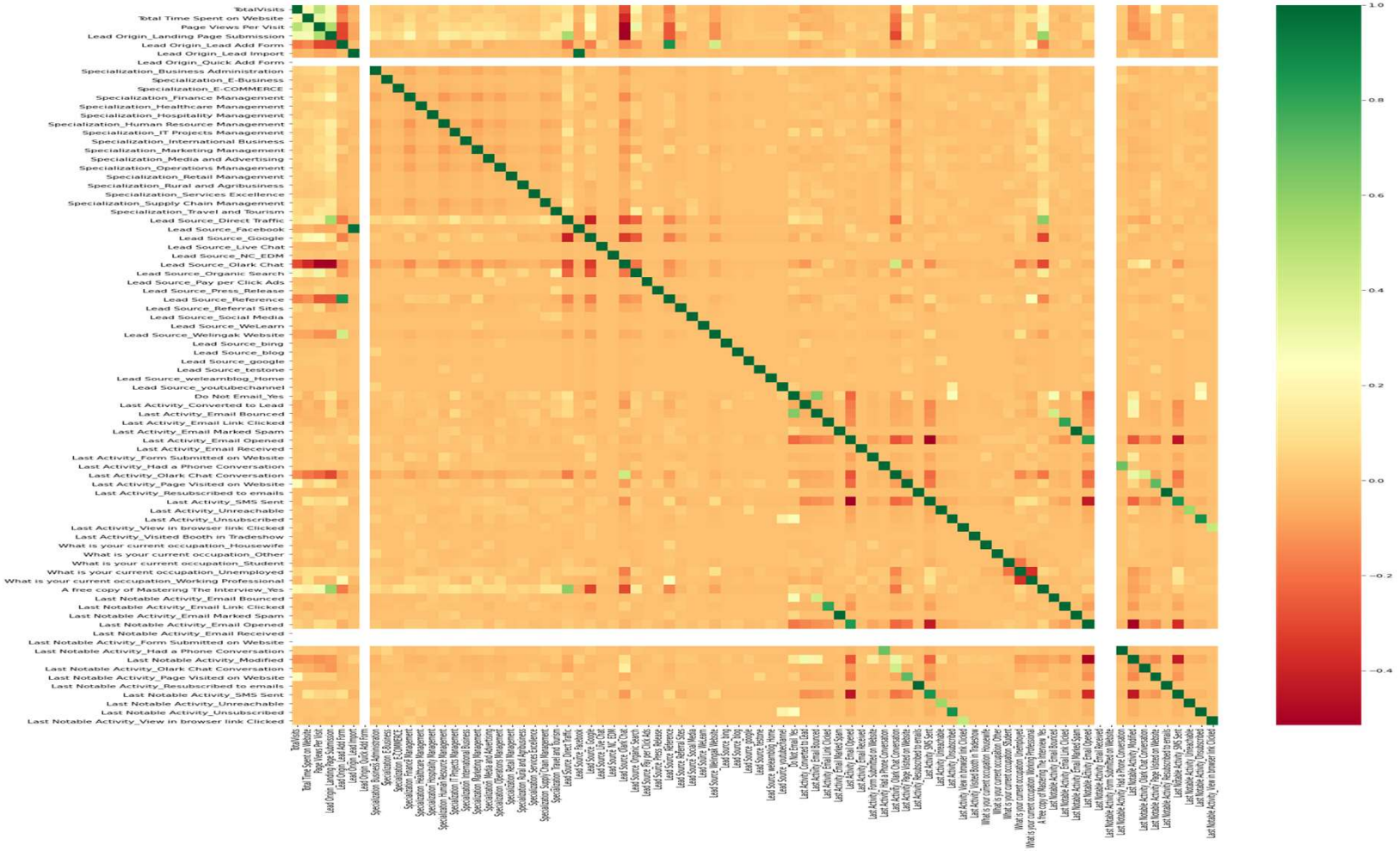


Multi Variate Analysis



•There is 0.36 correlation of "Total Time Spent on Website" with target variable "Converted".
•"Page Views Per Visit" have -0.015 correlation with target variable.

Train Test split - Correlation Among X-Variables



Top 10 Correlation for X_train

	VAR1	VAR2	Correlation_Value	Corr_abs
5706	Last Notable Activity_Email Marked Spam	Last Activity_Email Marked Spam	1.000000	1.000000
6369	Last Notable Activity_Resubscribed to emails	Last Activity_Resubscribed to emails	1.000000	1.000000
2055	Lead Source_Facebook	Lead Origin_Lead Import	0.972067	0.972067
6618	Last Notable Activity_Unsubscribed	Last Activity_Unsubscribed	0.868863	0.868863
2710	Lead Source_Reference	Lead Origin_Lead Add Form	0.858239	0.858239
6452	Last Notable Activity_SMS Sent	Last Activity_SMS Sent	0.856511	0.856511
5789	Last Notable Activity_Email Opened	Last Activity_Email Opened	0.837172	0.837172
5623	Last Notable Activity_Email Link Clicked	Last Activity_Email Link Clicked	0.801010	0.801010
6286	Last Notable Activity_Page Visited on Website	Last Activity_Page Visited on Website	0.704623	0.704623
6038	Last Notable Activity_Had a Phone Conversation	Last Activity_Had a Phone Conversation	0.670249	0.670249

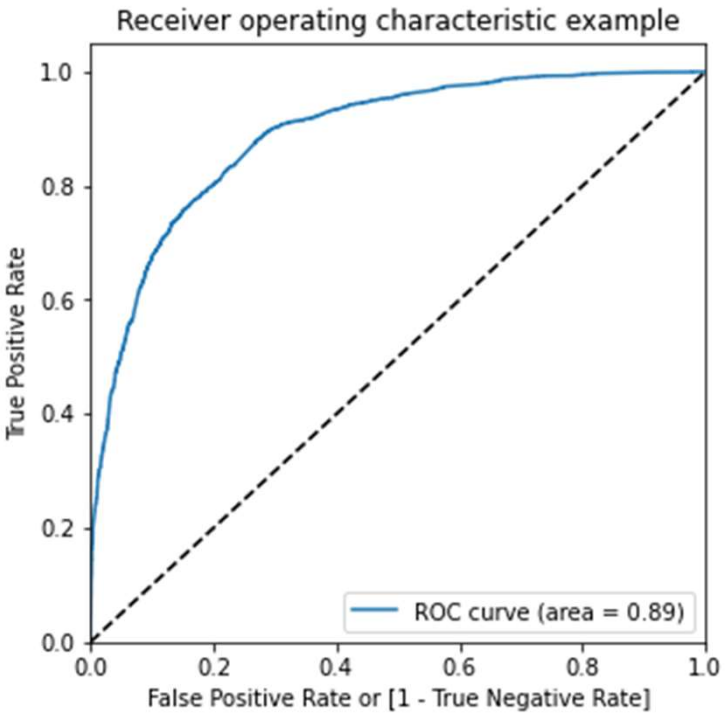
Model Building

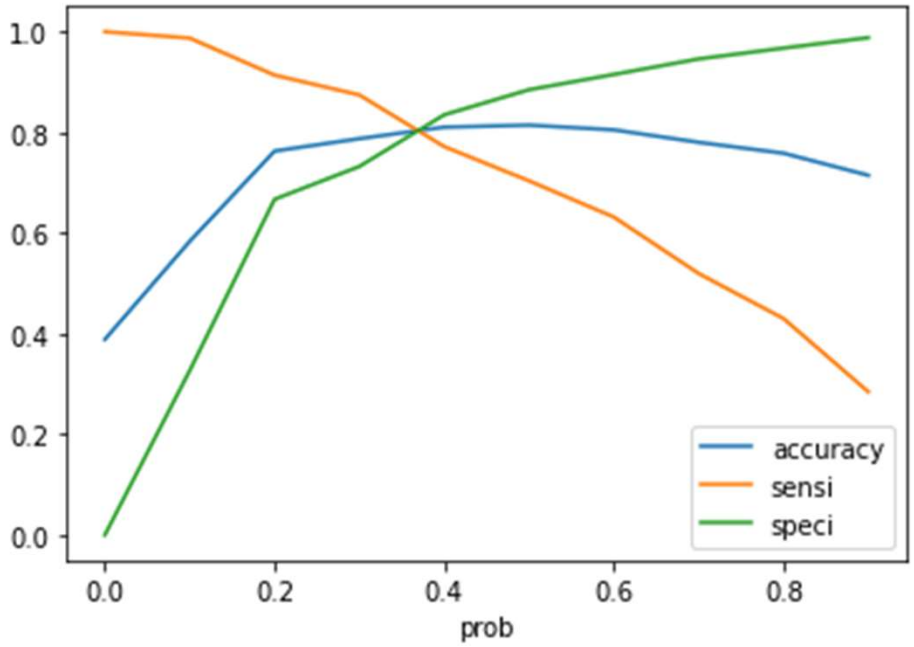
	Features	VIF
9	What is your current occupation_Unemployed	2.19
1	Total Time Spent on Website	1.91
2	Lead Origin_Lead Add Form	1.51
3	Lead Source_Olark Chat	1.50
0	TotalVisits	1.48
7	Last Activity_SMS Sent	1.48
6	Last Activity_Olark Chat Conversation	1.38
10	What is your current occupation_Working Profes...	1.33
4	Lead Source_Welingak Website	1.24
5	Do Not Email_Yes	1.06
8	What is your current occupation_Student	1.04
11	Last Notable Activity_Unreachable	1.00

Model Building

- RFE was used for feature selection.
- Then RFE was done to attain the top 12 relevant variables.
- Later the rest of the variables were removed manually depending on the VIF values and p-value.
- A confusion matrix was created, and overall accuracy was checked which came out to be 80.91%.

Model Evaluation





	Converted	Conversion_Prob	Predicted	0.0	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9
0	0	0.141154	0	1	1	0	0	0	0	0	0	0	0
1	0	0.200083	0	1	1	1	0	0	0	0	0	0	0
2	0	0.227713	0	1	1	1	0	0	0	0	0	0	0
3	0	0.169843	0	1	1	0	0	0	0	0	0	0	0
4	0	0.166704	0	1	1	0	0	0	0	0	0	0	0

	prob	accuracy	sensi	speci
0.0	0.0	0.394598	1.000000	0.000000
0.1	0.1	0.447377	0.994493	0.090769
0.2	0.2	0.634741	0.904013	0.459231
0.3	0.3	0.789351	0.761998	0.807179
0.4	0.4	0.798820	0.717545	0.851795
0.5	0.5	0.786247	0.648308	0.876154
0.6	0.6	0.769637	0.555468	0.909231
0.7	0.7	0.746818	0.448466	0.941282
0.8	0.8	0.721360	0.335956	0.972564
0.9	0.9	0.682862	0.206924	0.993077

Prediction on Test set-1

	const	Do Not Email	TotalVisits	Total Time Spent on Website	Lead Origin_Lead Add Form	Lead Source_Olark Chat	Lead Source_Reference	Lead Source_Welingak Website	Lead Source_google	Specialization_Hospitality Management	Specializ I
2400	1.0	0	0.028369	0.423856	0	0	0	0	0	0	
849	1.0	0	0.085106	0.029930	0	0	0	0	0	0	
7459	1.0	0	0.014184	0.058539	0	0	0	0	0	0	
6184	1.0	1	0.021277	0.233715	0	0	0	0	0	0	
4869	1.0	0	0.014184	0.581426	0	0	0	0	0	0	
...	...	...	...	...	...	...	...	...	...	...	
3640	1.0	0	0.028369	0.091549	0	0	0	0	0	0	
1051	1.0	0	0.028369	0.613556	0	0	0	0	0	0	
8707	1.0	0	0.049645	0.264085	0	0	0	0	0	0	
9103	1.0	0	0.007092	0.301937	0	0	0	0	0	0	
6153	1.0	0	0.028369	0.290493	0	0	0	0	0	0	

Prediction on Test set-2

```
y_test_pred[:10]
```

```
2400    0.489831
849     0.155418
7459    0.143642
6184    0.079679
4869    0.968269
2519    0.811714
5419    0.163886
2103    0.494985
3987    0.227713
3512    0.306841
dtype: float64
```

sensitivity

```
: # Calculate sensitivity
  TP / float(TP+FN)
```

```
: 0.7
```

Specificity

```
: # Calculate specificity
  TN / float(TN+FP)
```

```
: 0.8628668171557562
```

Precision-Recall

Precision

TP / TP + FP

```
confusion[1,1]/(confusion[0,1]+confusion[1,1])
```

```
0.7733458470201783
```

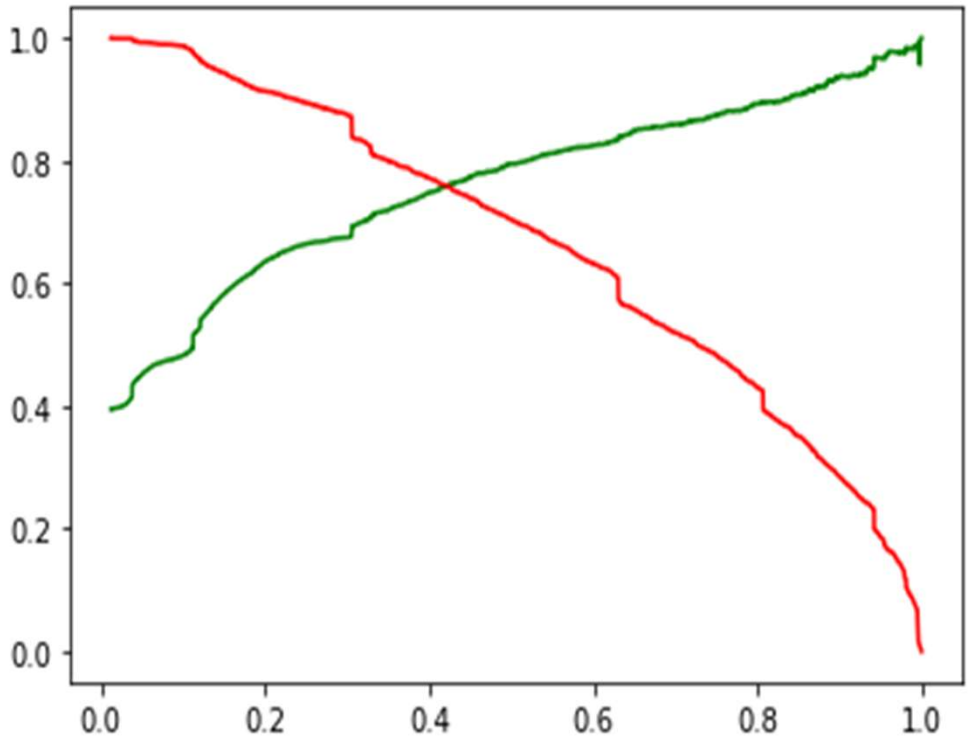
Recall

TP / TP + FN

```
confusion[1,1]/(confusion[1,0]+confusion[1,1])
```

```
0.6483084185680567
```

Precision recall curve



Threshold point is 0.3

Summary

- **After running the model on the Test Data these are the figures we obtain:**
- Accuracy = 81.10%
- Sensitivity = 77.08%
- Specificity = 83.54%
- **Final Observation**

Train Set:

- Accuracy = 81.03%
- Sensitivity = 77.17%
- Specificity = 83.49%

Test Set:

- Accuracy = 81.02%
- Sensitivity = 76.31%
- Specificity = 83.88%
- The Model seems to predict the Conversion Rate very well and we should be able to give the CEO confidence in making good calls based on this model

Conclusion

- TOP VARIABLE CONTRIBUTING TO CONVERSION:
- LEAD SOURCE:
    - Total Visits
    - Total Time Spent on Website
  - Lead Origin:
    - Lead Add Form
  - Lead source:
    - Direct traffic
    - Google
    - Welingak website
    - Organic search
    - Referral Sites
- Last Activity:
  - Do Not Email_Yes
  - Last Activity_Email Bounced
  - Olark chat conversation

- The Model seems to predict the Conversion Rate very well and we should be able to give the Company confidence in making good calls based on this model.