Lead Scoring Case Study

- Archana Krishna
- Ritika Shelkar
- Shanu Gautam

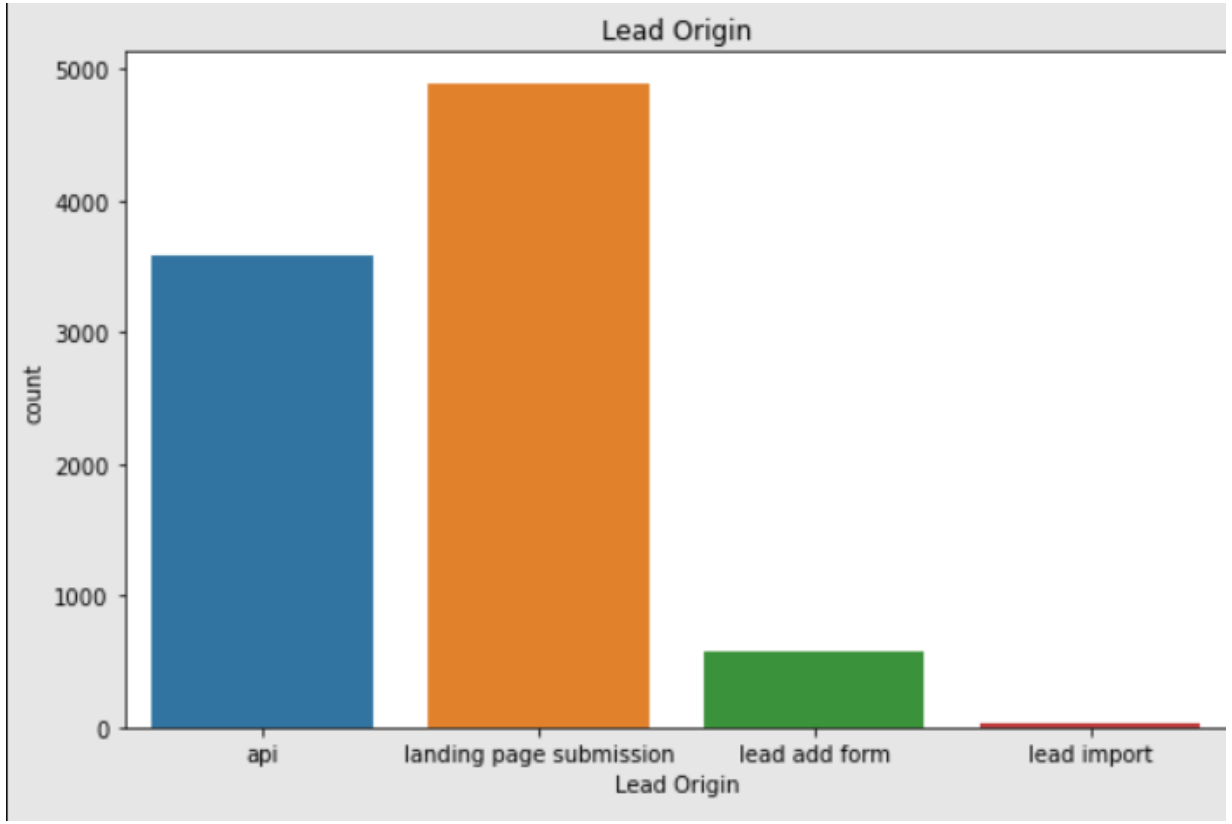
Problem Statement

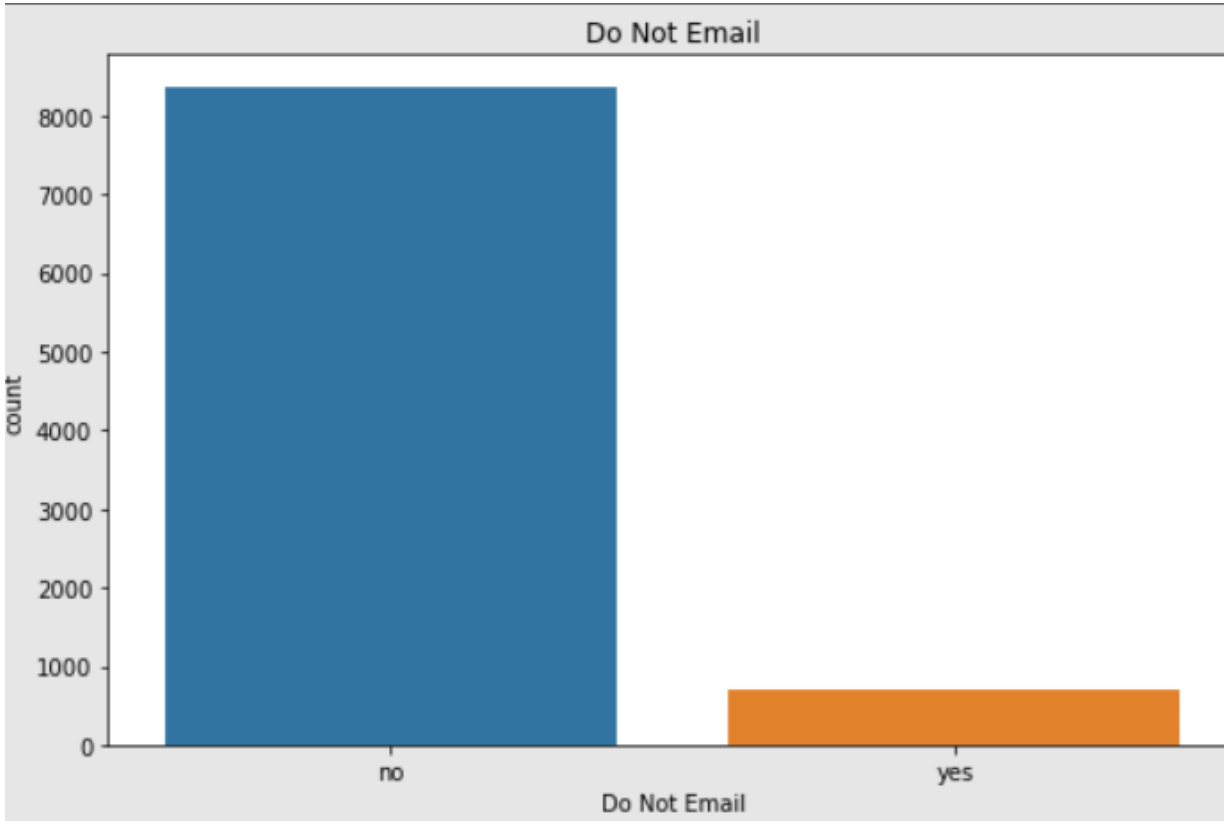
- An X Education need help to select the most promising leads, i.e. the leads that are most likely to convert into paying customers.
- The company requires us to build a model wherein we need to assign a lead score to each of the leads such that
- the customers with higher lead score have a higher conversion chance and the customers with lower lead score have a lower conversion chance.
- The CEO, in particular, has given a ballpark of the target lead conversion rate to be around 80%.

Approach

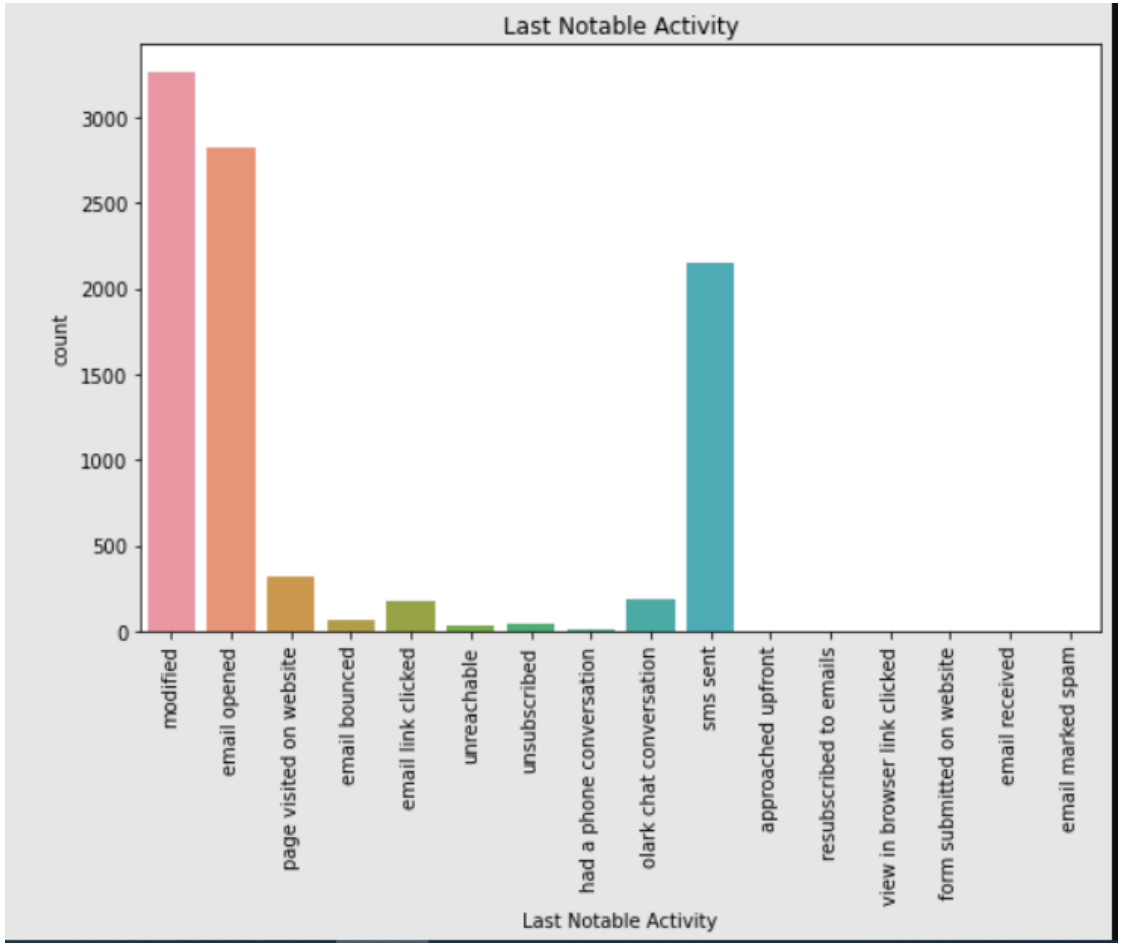
- We will initially start with analyzing the data with the basic things like the shape of the dataset, number of columns, count of null values in a columns, checking the number of unique values in a column etc.
- Then we will do the data cleaning of the dataset with removing the unwanted columns from the dataset, followed with data preparation by imputing the null value, standardizing the data, remove the columns with more than 35% of null values.
- Perform Exploratory Data Analysis on the dataset and select the columns that will contribute to the model using RFE and checking the p-value and VIF.
- Then we will split the data with 70% of dataset as train dataset and 30% as test dataset. We will then train our model with the &70% of the dataset and find the accuracy score and the precision and the recall.
- Finally, we will evaluate our model with the test dataset.

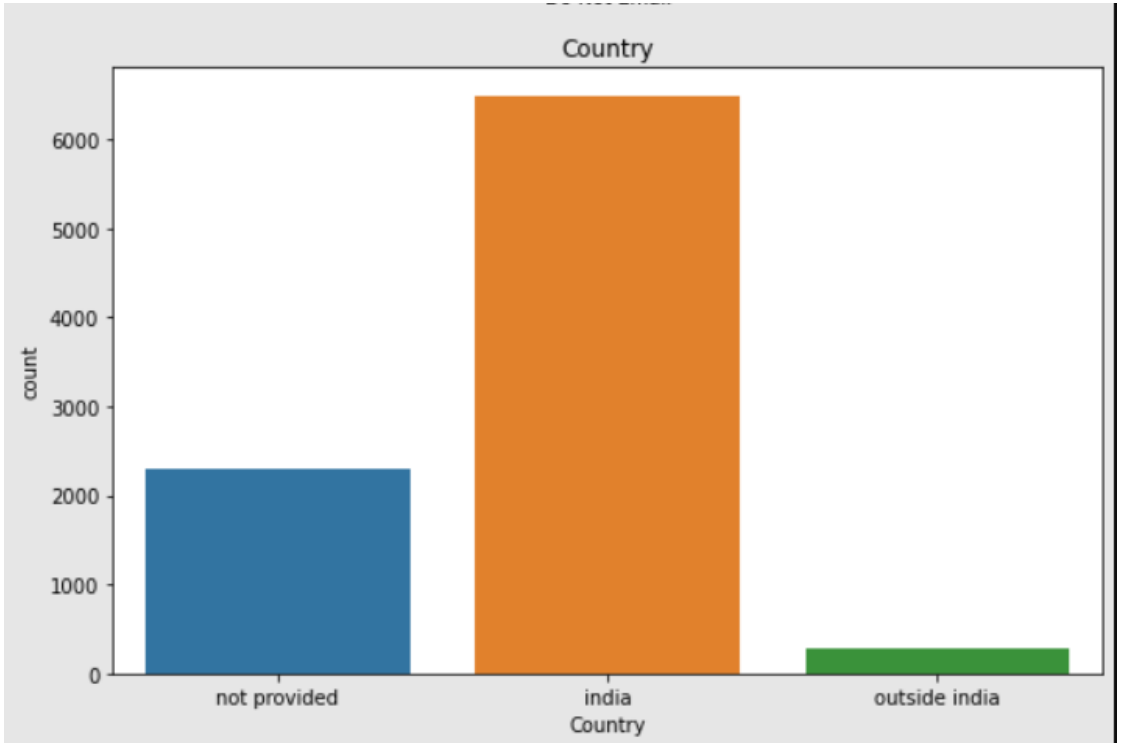
EDA Results





- From the above graphs we can see that the origin identifier with which the customer was identified to be a lead is the landing page submissions followed by api aand lead add form.
- From the second bar graph we can see that most of the customers opted receiving mails regarding the courses.





- From the first bar graph we can see that the last activity of most of the students was modified, followed by email opened and sms sent.
- From the second graph we can see that most of the students are from India followed.

Fig 1

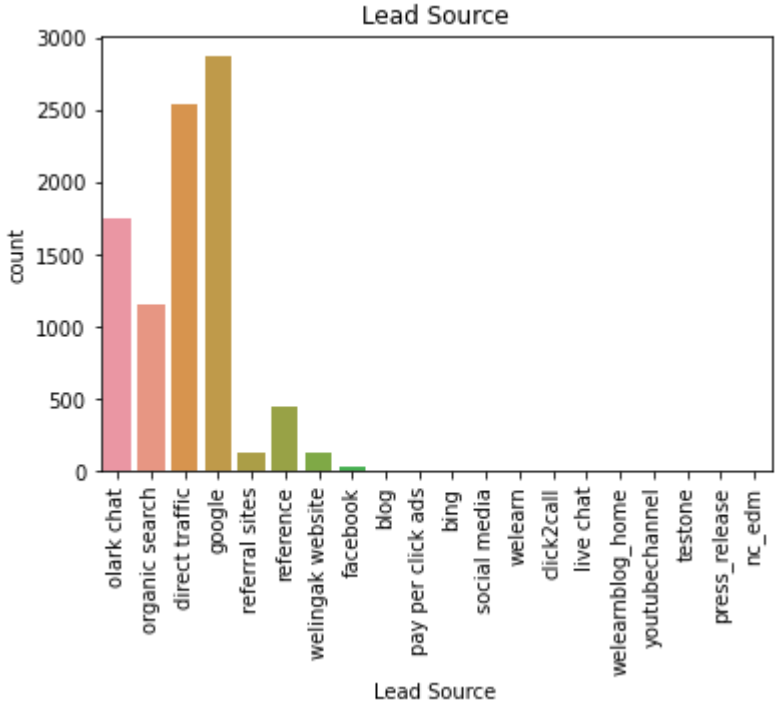
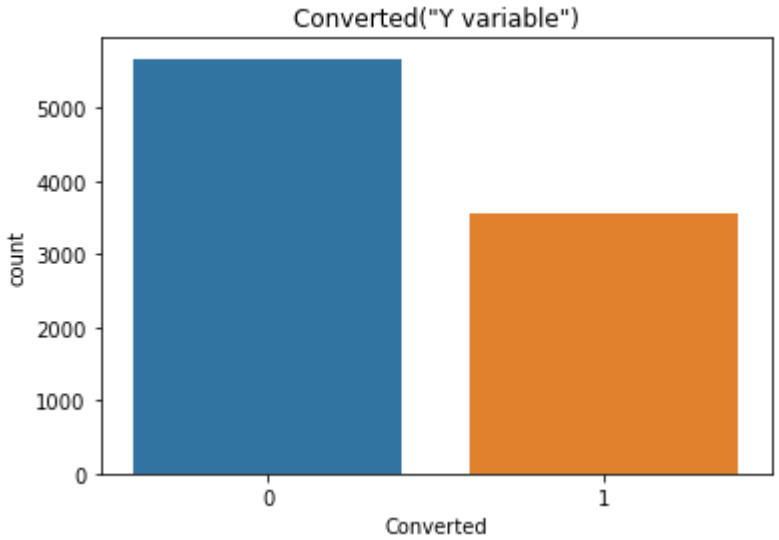


Fig 3



- Fig 1: We can see that most of the leads were obtained from google, followed by direct traffic and olark chat.
- Fig 2: We can see that most of the students have specialization in finance management followed by HR and marketing and operations management.
- Fig 3: The converted students count present are around 3100 in the dataset out of 9240 entries.

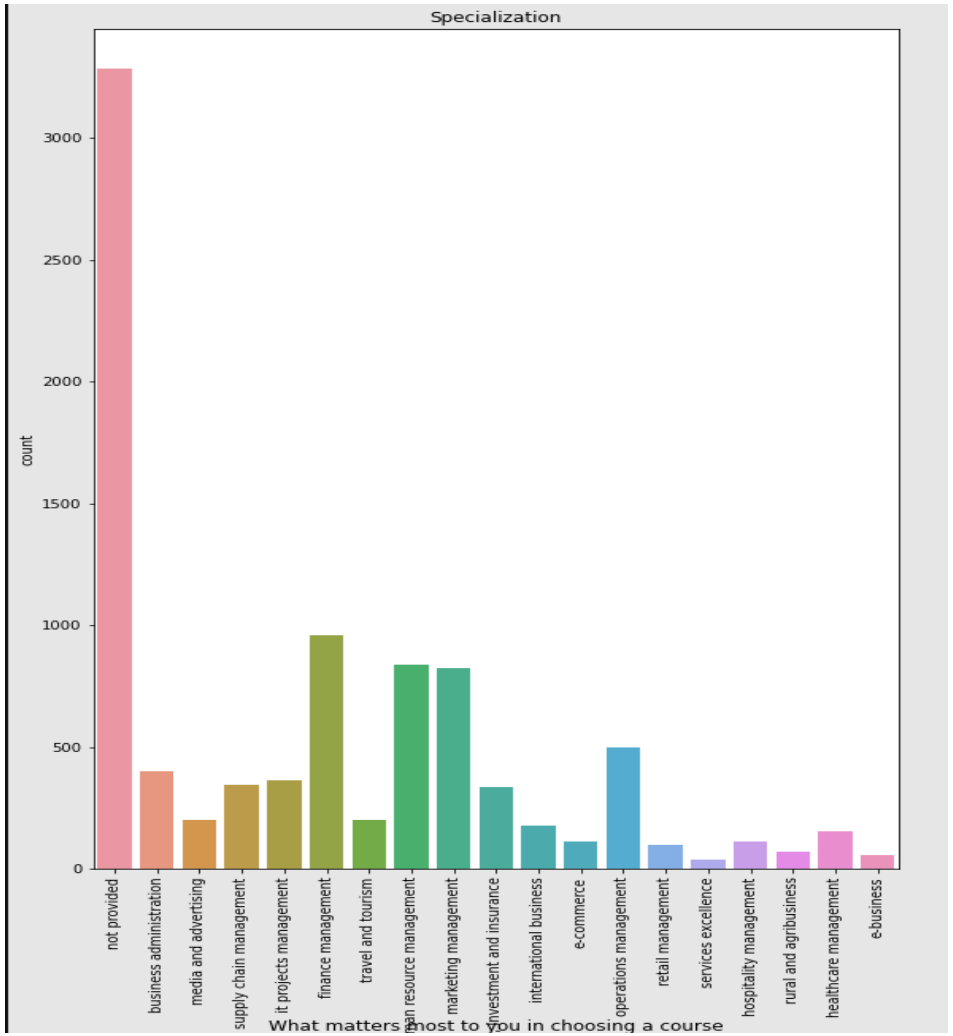


Fig 2

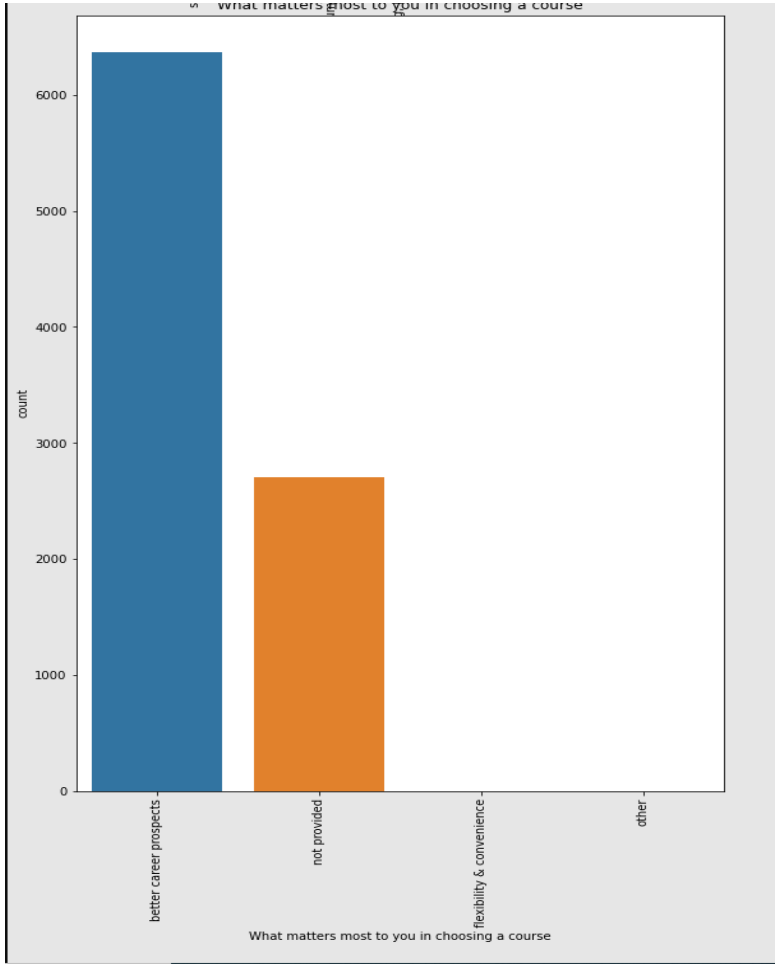


Fig 1

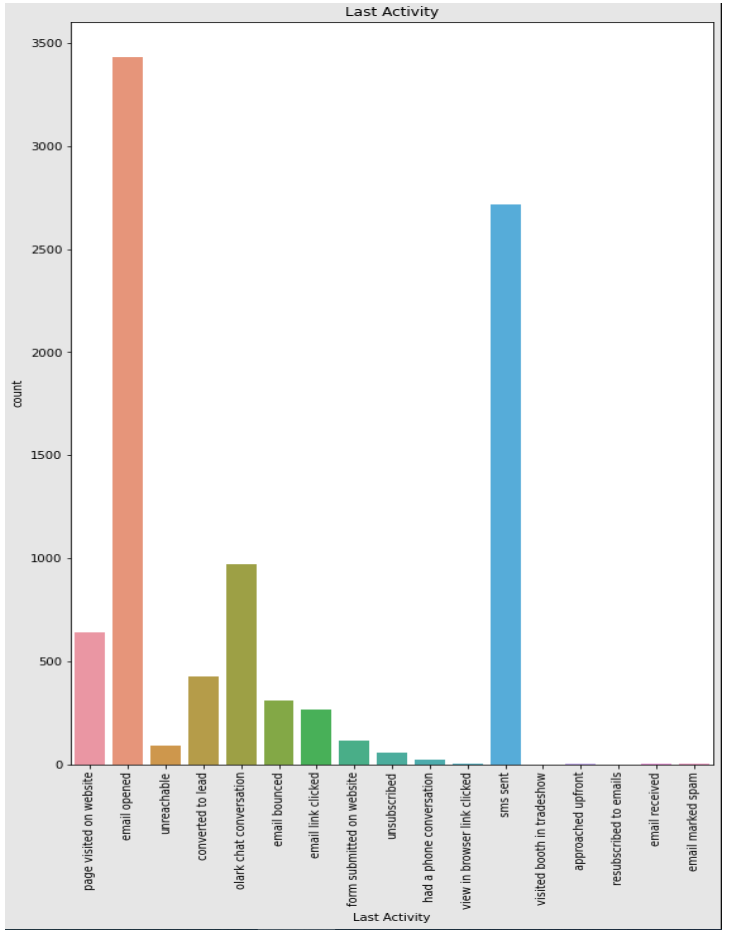


Fig 2

Fig 1: We can see that students usually look for a course that have a better career prospects.
Fig 2: The last activity of most of the students is email opened followed by sms sent.

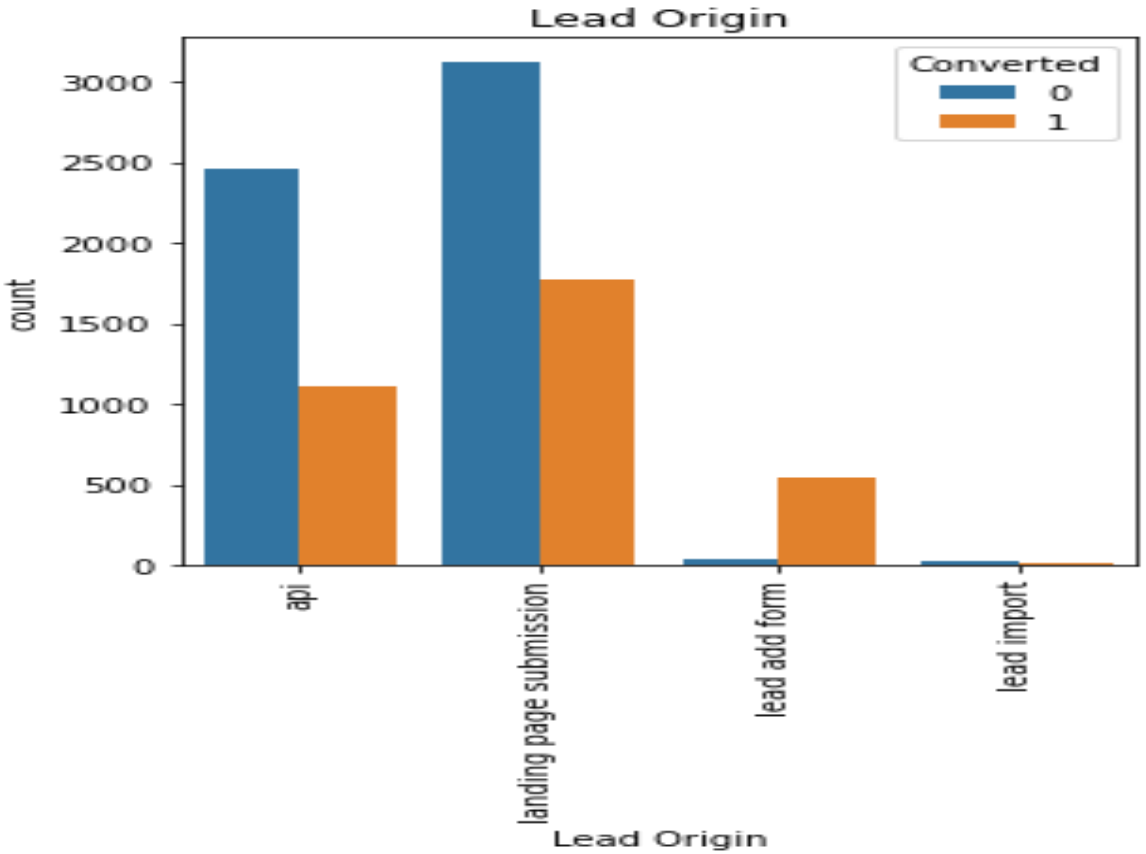


Fig 1

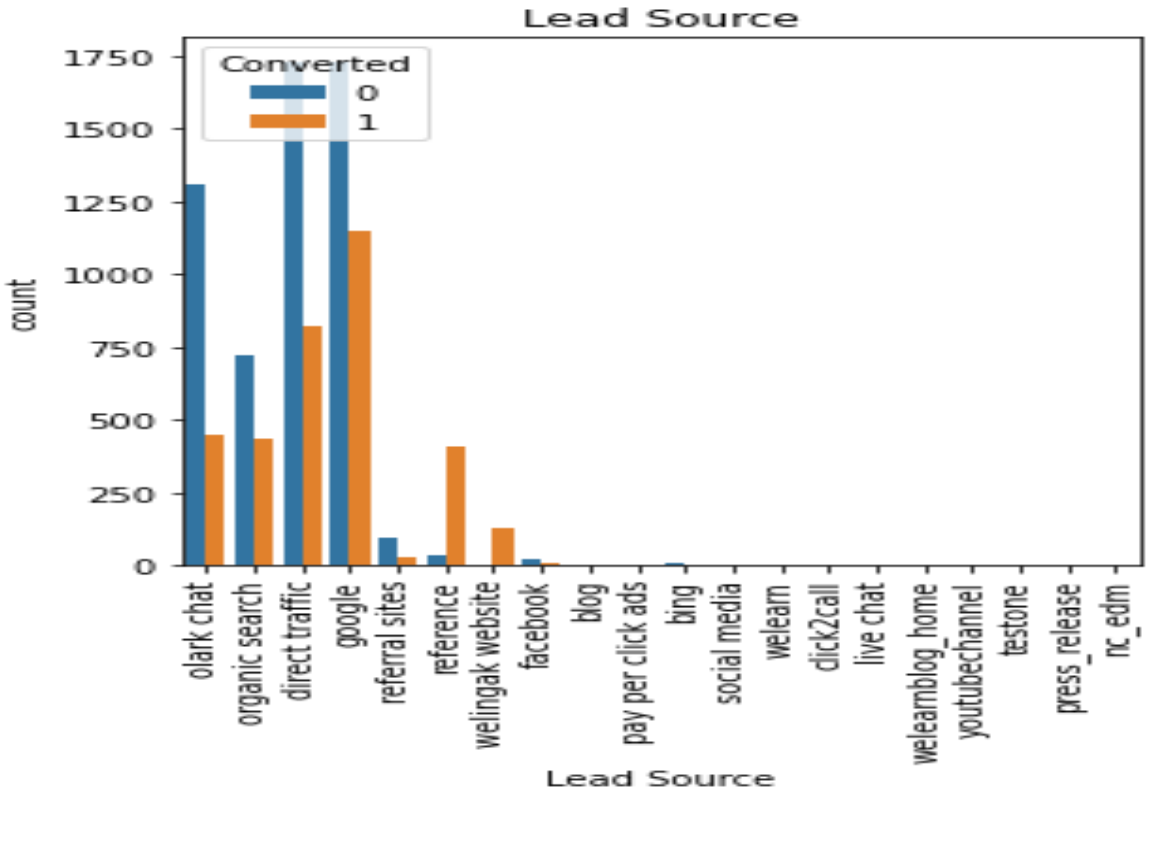
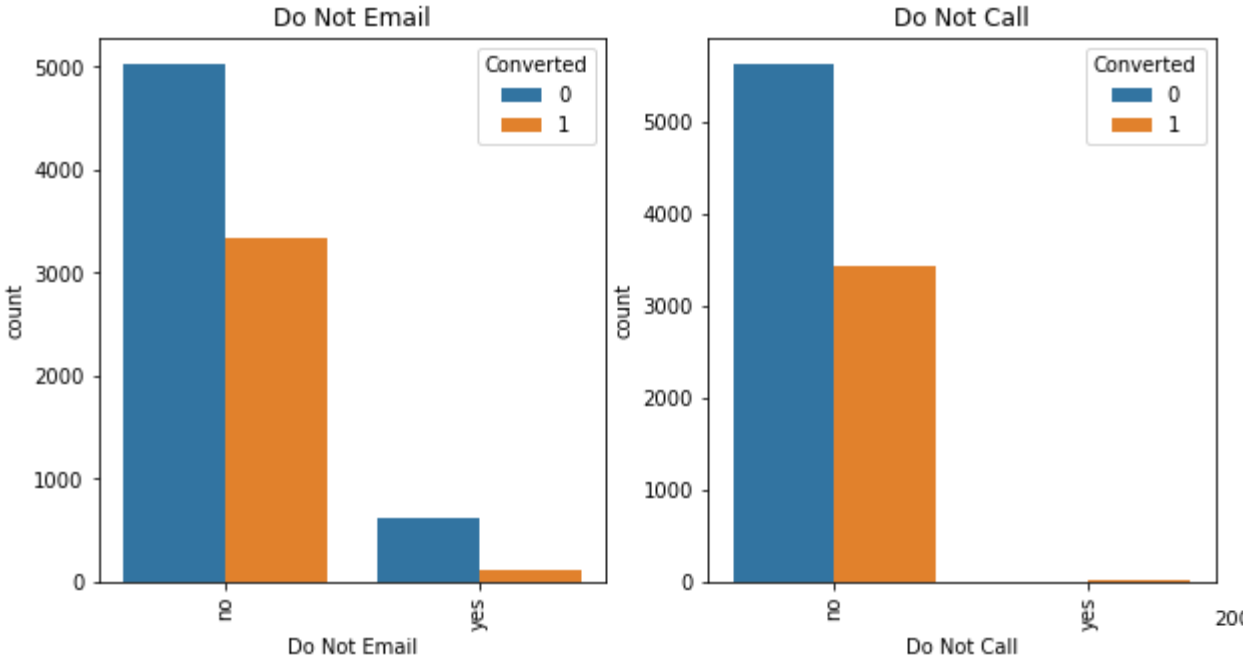


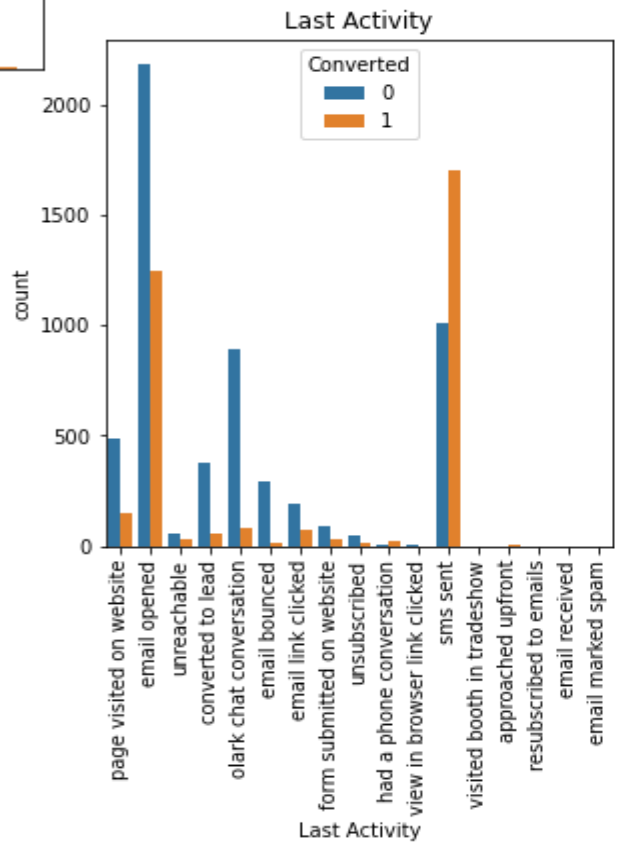
Fig 2

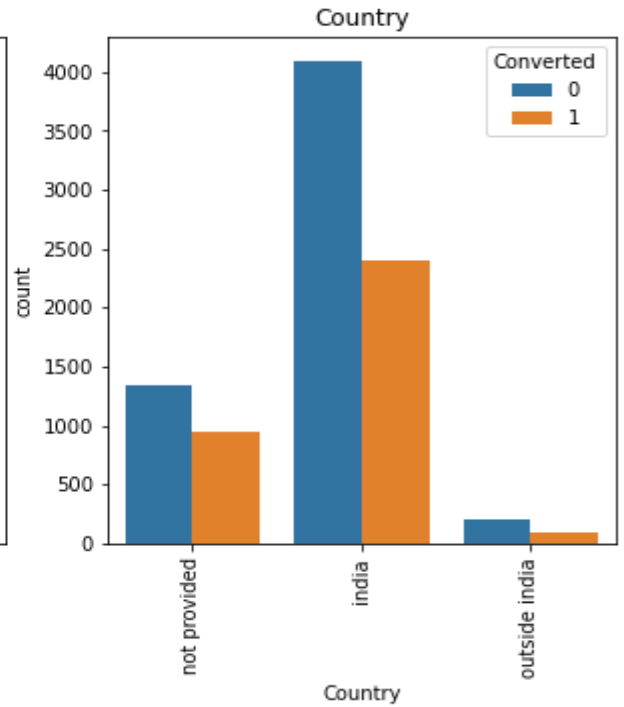
- Fig 1: We can see that most of the students that were converted had an origin of landing page submission followed by api.
- Fig 2: The Lead sources that got converted were from google followed by direct traffic and olark chat.

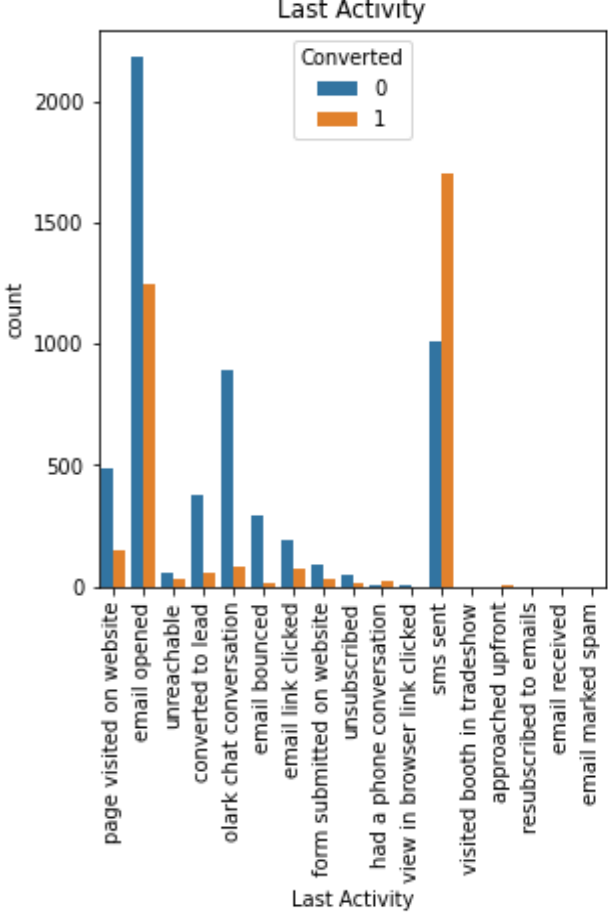


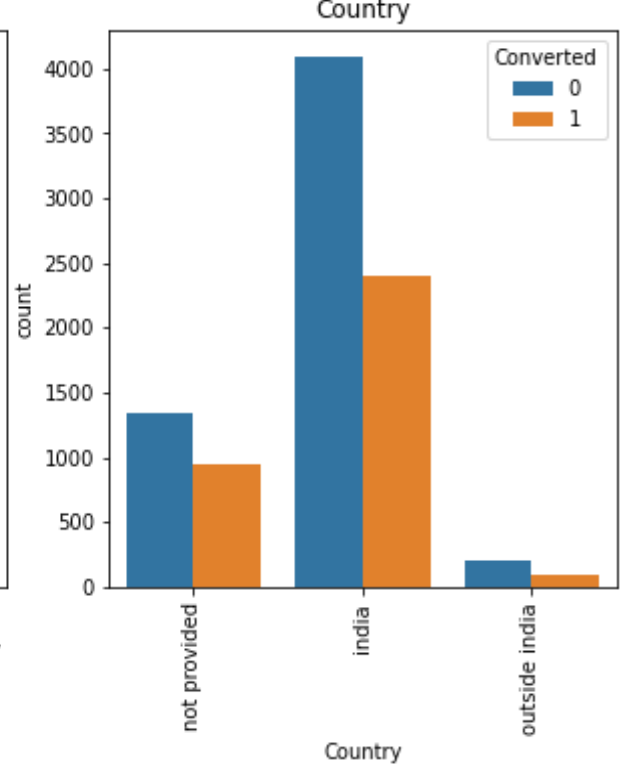
The students who opted for receiving the mails and call were converted to take the course.

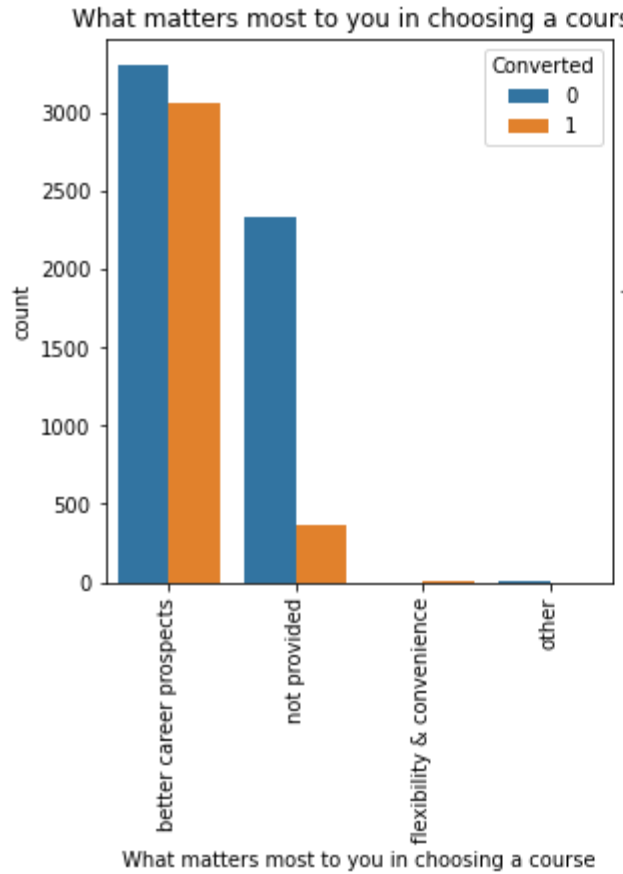
- From the last activity graph we can see that the students who received an sms were converted followed by email opened.
- From the country graph we can see that most of the converted students were from India.

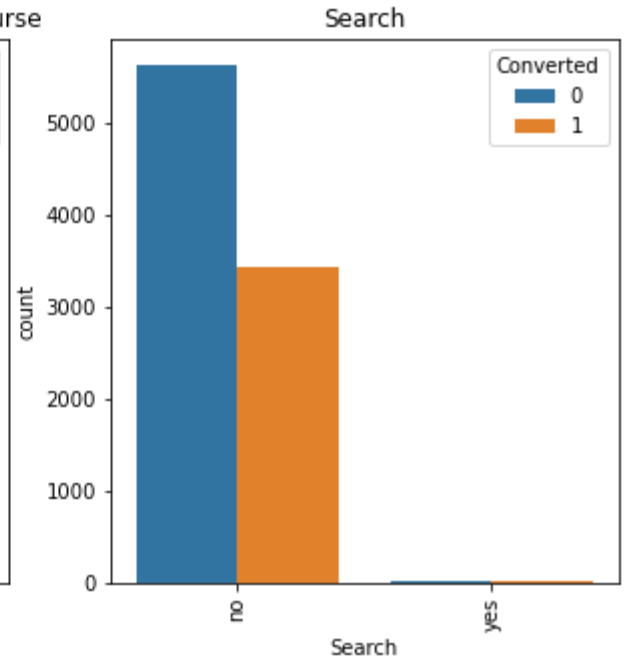


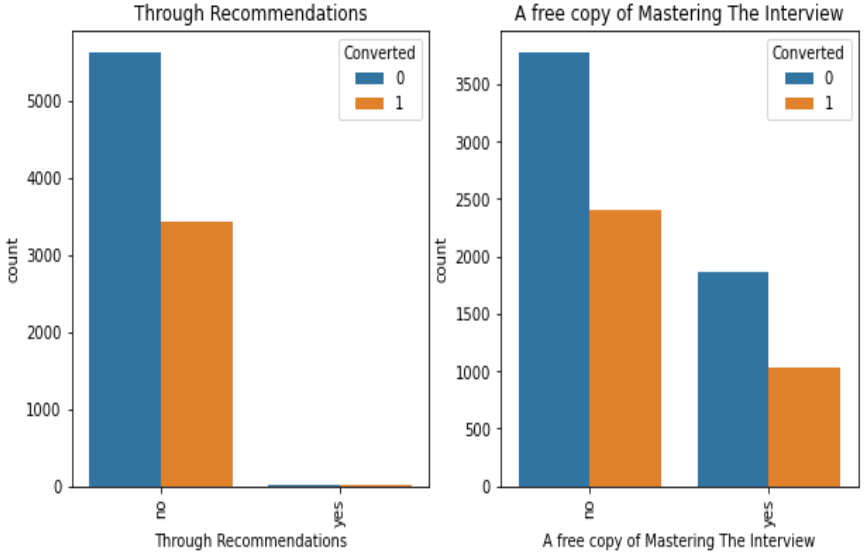


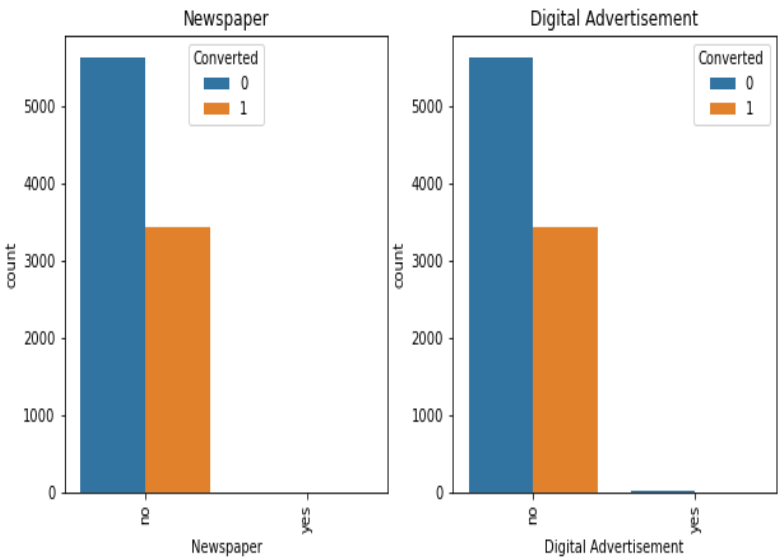

Last Activity
Converted
0
1
count
2000
1500
1000
500
0
page visited on website
email opened
unreachable
converted to lead
olark chat conversation
email bounced
email link clicked
form submitted on website
unsubscribed
had a phone conversation
view in browser link clicked
sms sent
visited booth in tradeshow
approached upfront
resubscribed to emails
email received
email marked spam
Last Activity

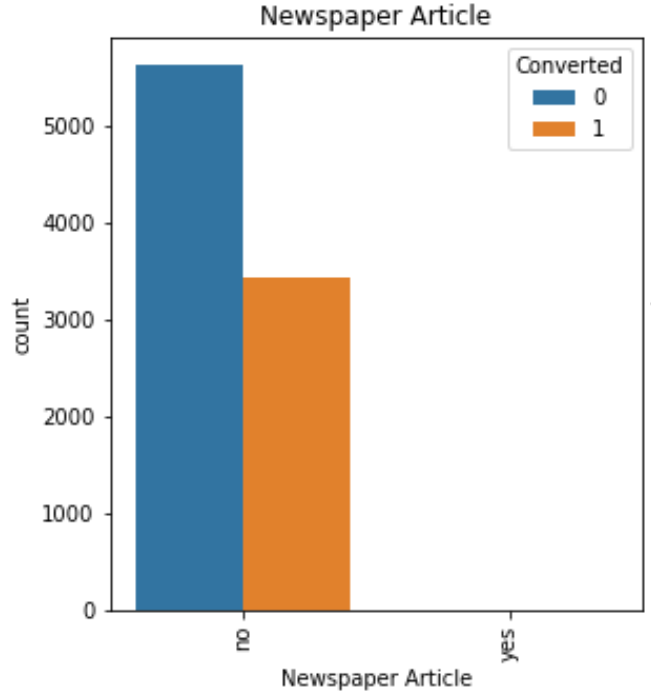

Country
Converted
0
1
count
4000
3500
3000
2500
2000
1500
1000
500
0
not provided
india
outside india
Country

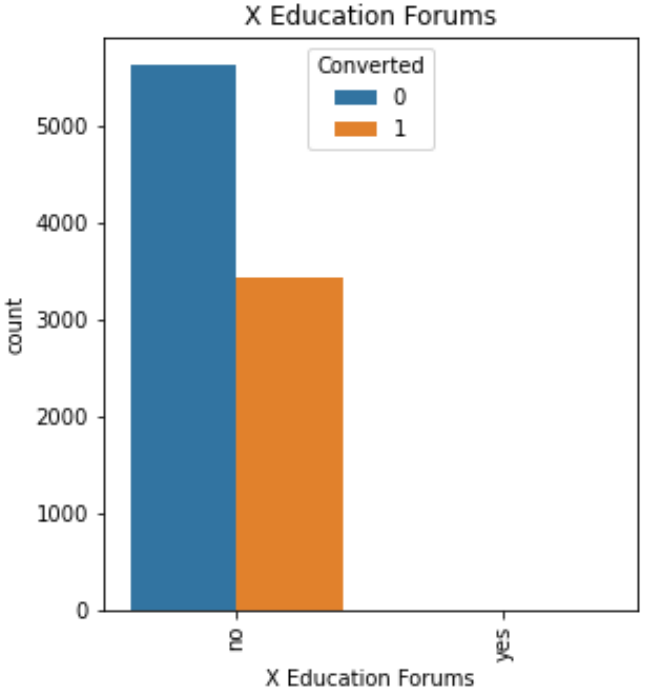

What matters most to you in choosing a course
Converted
0
1
count
3000
2500
2000
1500
1000
500
0
better career prospects
not provided
flexibility & convenience
other
What matters most to you in choosing a course


Search
Converted
0
1
count
5000
4000
3000
2000
1000
0
no
yes
Search

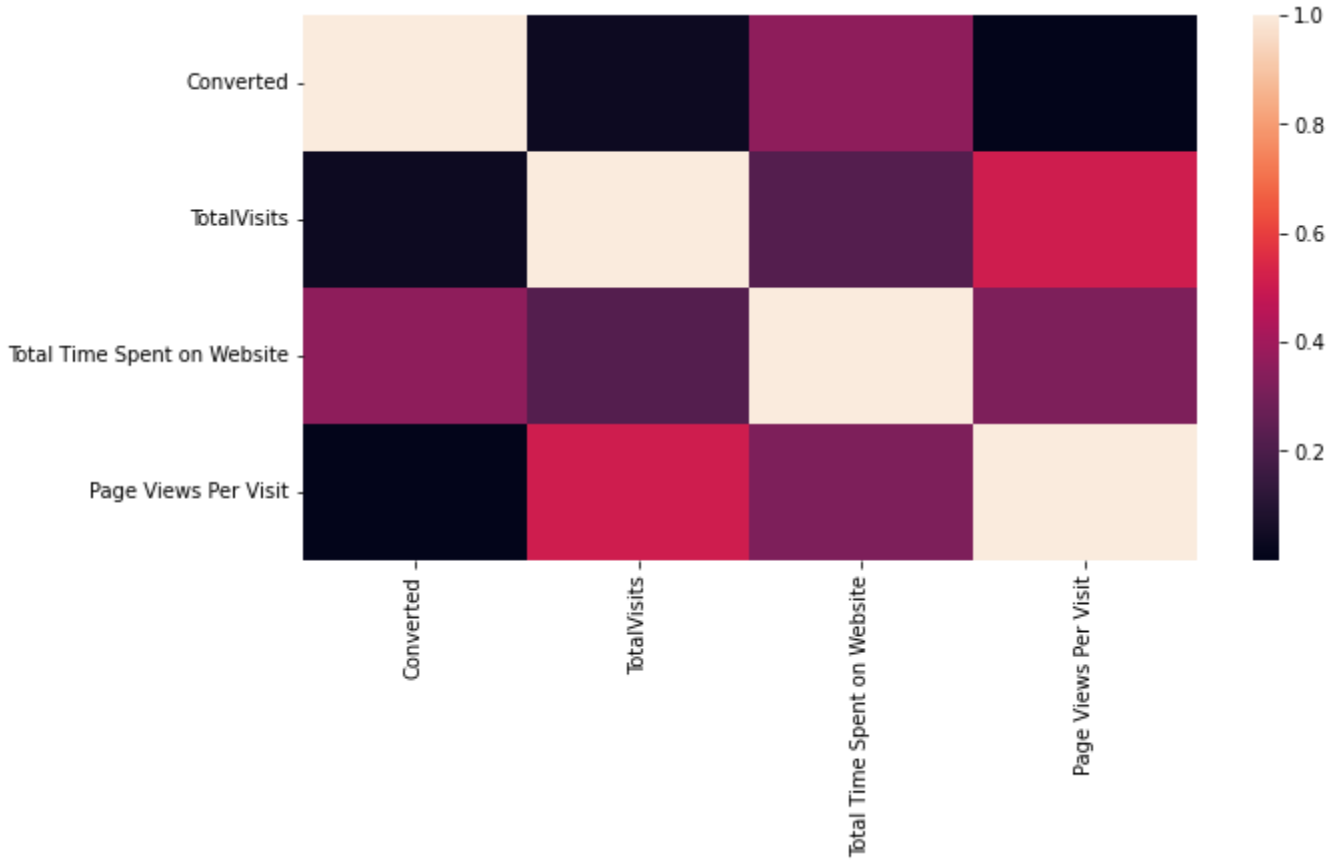








The above graphs show the plots for various sources and the conversion ratio for the same.



The following is a correlation matrix which shows that most of the correlation is between total visits and page views per visit. And page views per visit and total time spent on website.

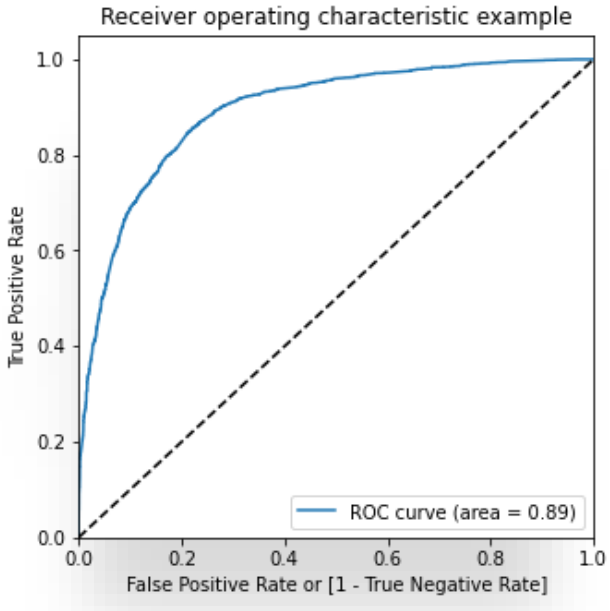


Fig 1

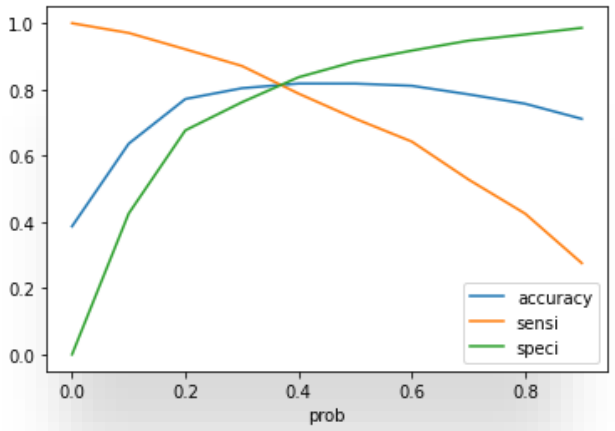


Fig 2

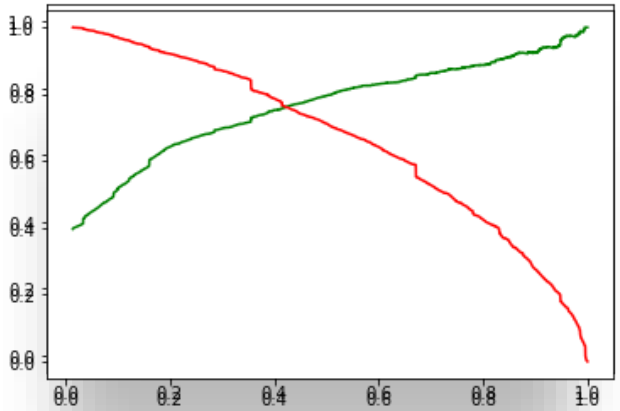


Fig 3

Results

- Fig 1: Shows the ROC curve with an area of 0.89. Since out ROC curve touches the upper-left corner of the graph it shows that are model is a good model.
- Fig 2: Shows the trade-off between accuracy, sensitivity and specificity. We can see that these 3 intersect at around 0.4.
- Fig 3: Shows the trade-off between Precision and recall which also intersect at 0.4.

In [118]:
```
# Creating confusion matrix again
Confusion2 = metrics.confusion_matrix(y_train_pred_final.Converted, y_train_pred_final.final_predicted )
Confusion2
```

Out[118]:
```
array([[3273,  622],
       [ 539, 1917]], dtype=int64)
```

In [120]:
```
# Precision = TP / TP + FP
TP / (TP + FP)
```

Out[120]:
```
0.7160578911095796
```

In [121]:
```
#Recall = TP / TP + FN
TP / (TP + FN)
```

Out[121]:
```
0.8460912052117264
```

These are figures that show the confusion matrix, precision and recall for the trained data.

In [127]:
```
# Creating confusion matrix
confusion2 = metrics.confusion_matrix(y_pred_final['Converted'], y_pred_final.final_predicted )
confusion2
```

Out[127]:
```
array([[1476,  268],
       [ 213,  766]], dtype=int64)
```

In [128]:
```
# Substituting the value of true positive
TP = confusion2[1,1]
# Substituting the value of true negatives
TN = confusion2[0,0]
# Substituting the value of false positives
FP = confusion2[0,1]
# Substituting the value of false negatives
FN = confusion2[1,0]
```

In [129]:
```
# Precision = TP / TP + FP
TP / (TP + FP)
```

Out[129]:
```
0.7408123791102514
```

In [130]:
```
#Recall = TP / TP + FN
TP / (TP + FN)
```

Out[130]:
```
0.7824310520939735
```

These are figures that show the confusion matrix, precision and recall for the test data.

We can see that precision is around 0.74 and Recall is around 0.78

Recommendations

- Thus, we can say that the company must focus on students that have specialization in finance management, HR, marketing and operations management.
- Also, we saw that most of the students that were converted had sources from google, direct traffic and olark chat.
- Also, we should focus on converting on students from India the most for the course. Also, should keep in mind that students usually look for courses with better career aspects.
- Also, the students mostly reply to emails and sms sent so should communicate mostly through these mediums.