Lead Scoring Case Study

By

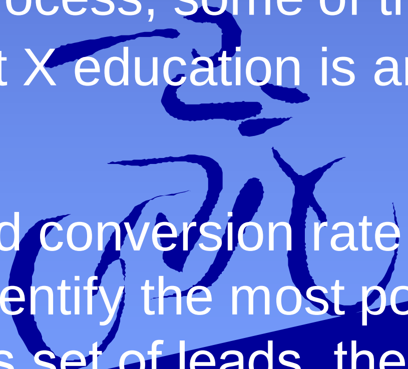
Rahul Kumar Gupta

Introduction

- This assignment aims to give an idea of applying Logistic Regression Model in a real business scenario. In this assignment, a basic understanding of marketing and customer behaviour is also used to understand how data can be used to increase the number of customers.

Problem Statement

- An education company named X Education sells online courses to industry professionals.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. 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.



Business Goal

- X Education has appointed you to help them select the most promising leads, i.e. the leads that are most likely to convert into paying customers.

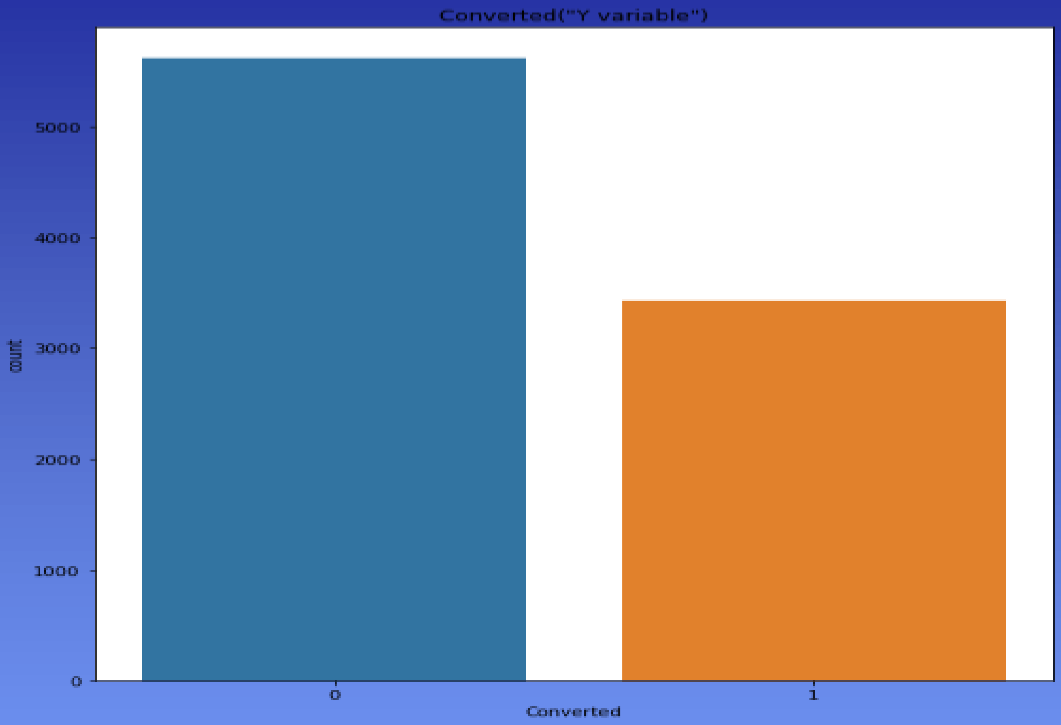
- The company requires you to build a model wherein you need to assign a lead score to each of the leads such that the customers with a higher lead score have a higher conversion chance and the customers with a lower lead score have a lower conversion chance. The CEO, in particular, has given a ballpark of the target lead conversion rate to be around 80%.

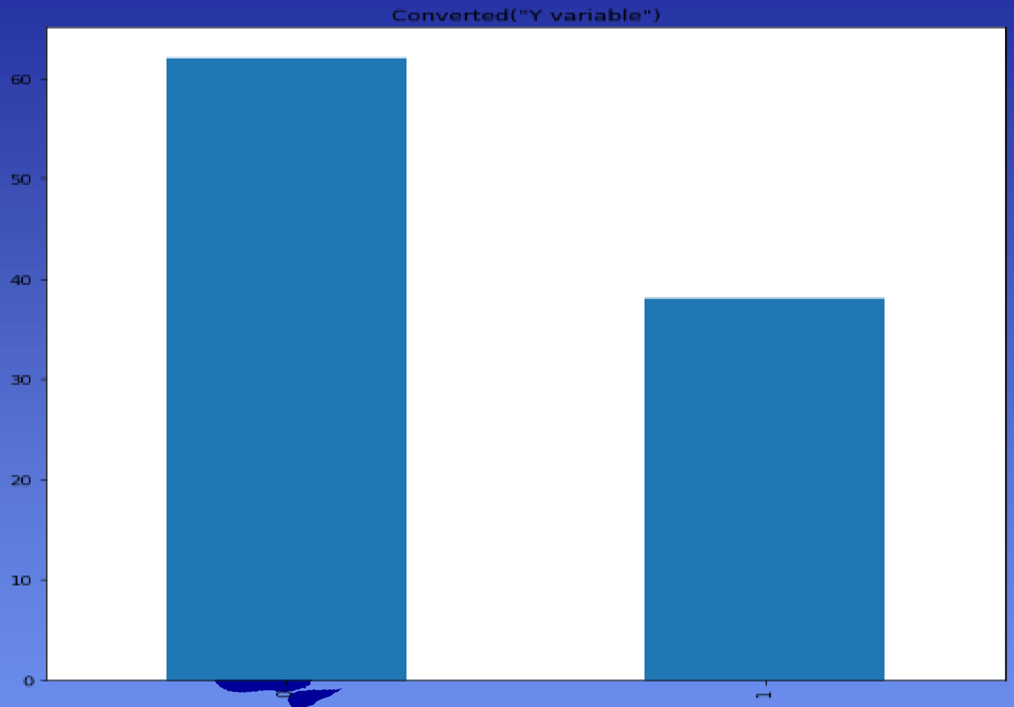
STRATEGY

1. EDA
2. MODEL BUILDING
3. EVALUATION OF MODEL
4. TESTING OF MODEL ON TEST DATASET
5. FINAL EVALUATION
6. INSIGHTS

Analysis of the Leads(EDA)

Analysis of the Target variable





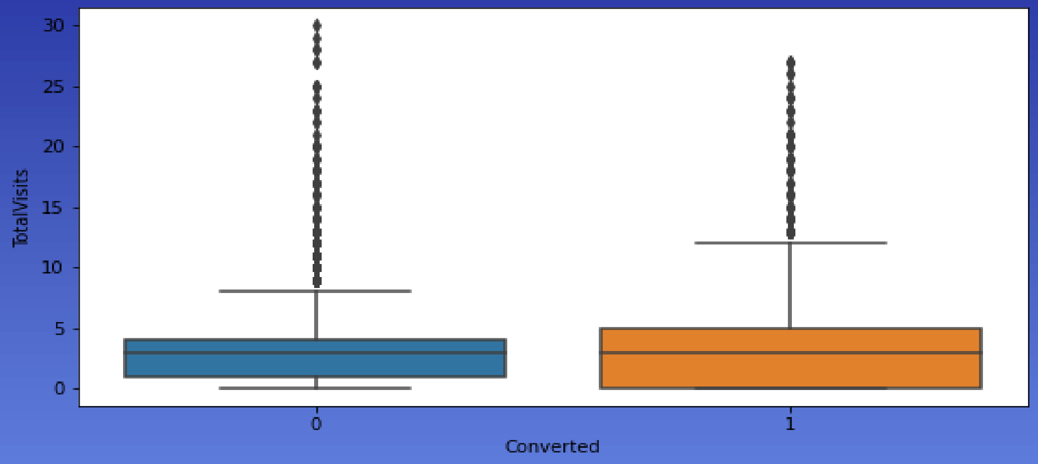
```
0    5620
1    3426
Name: Converted
```

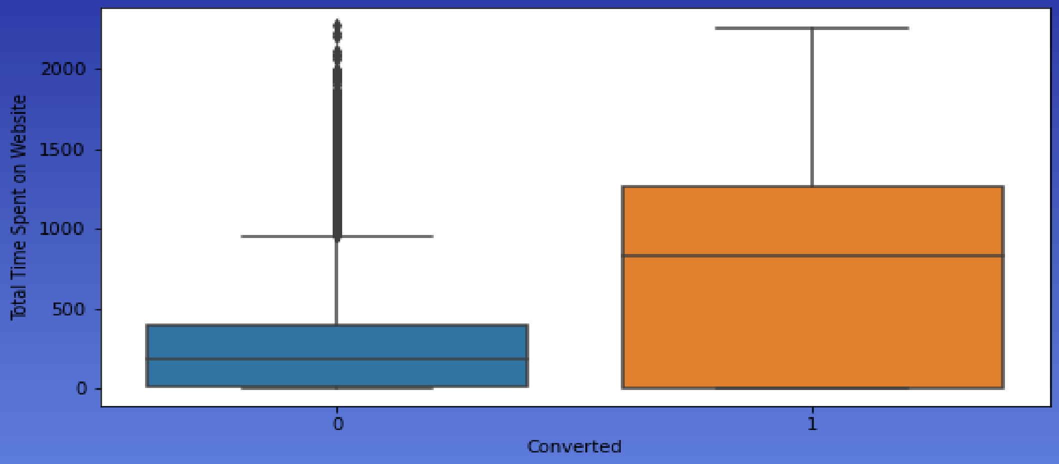
```
0    62%
1    37%
Name: Converted
```

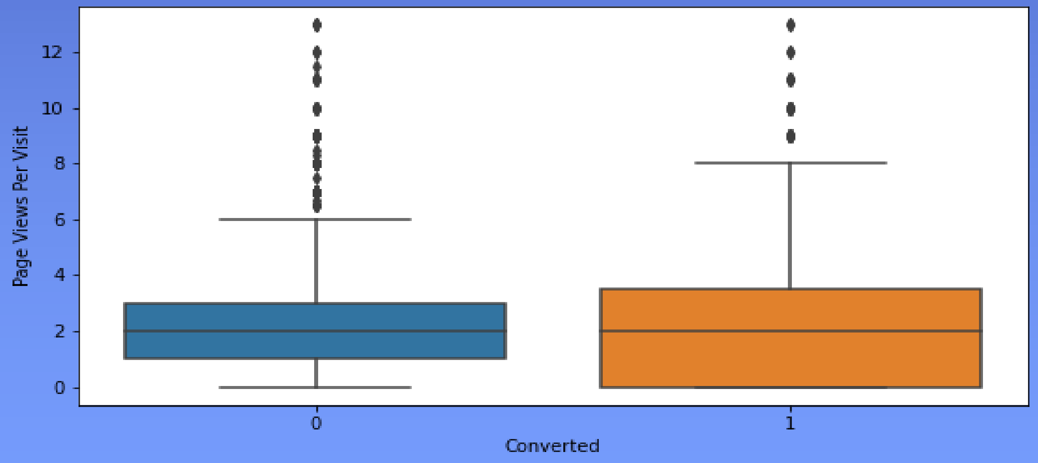
We can see that 37% are Converted while 67% customers are not converted

Numeric variables

TotalVisits, Total Time Spent on Website, Page Views Per Visit

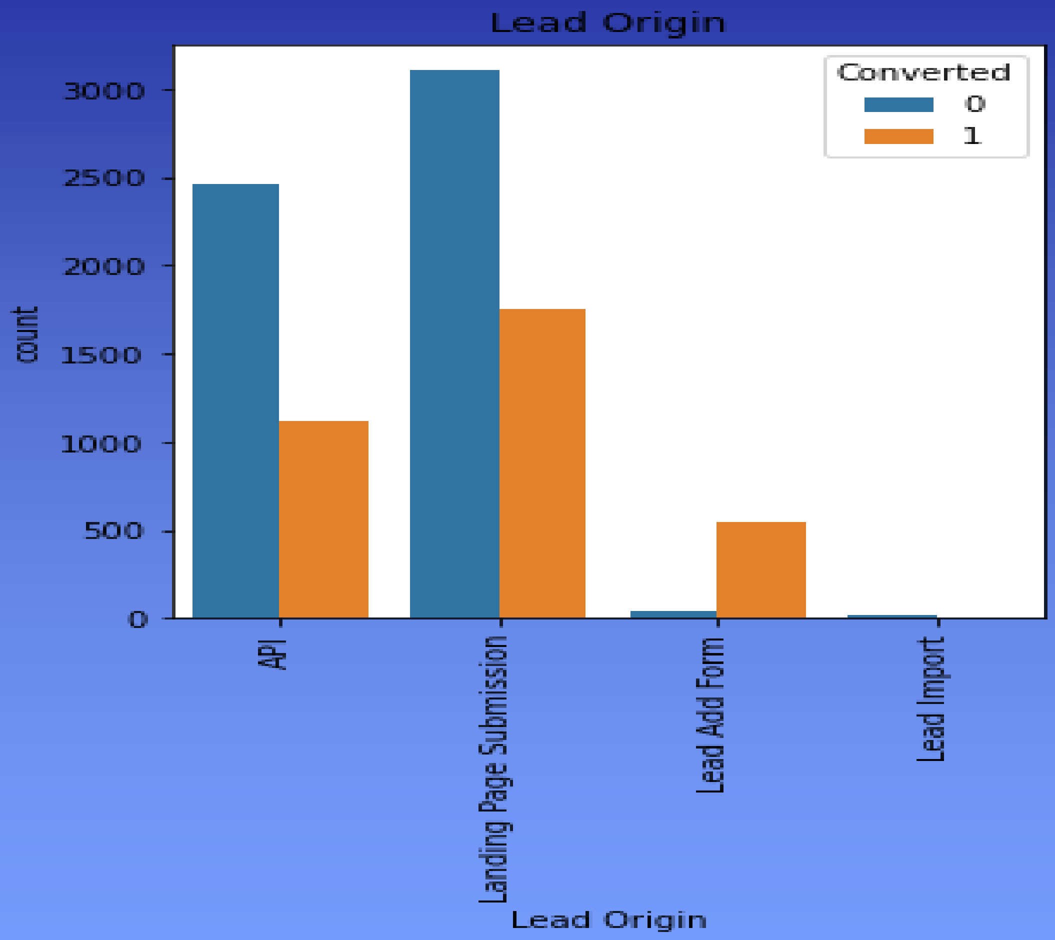


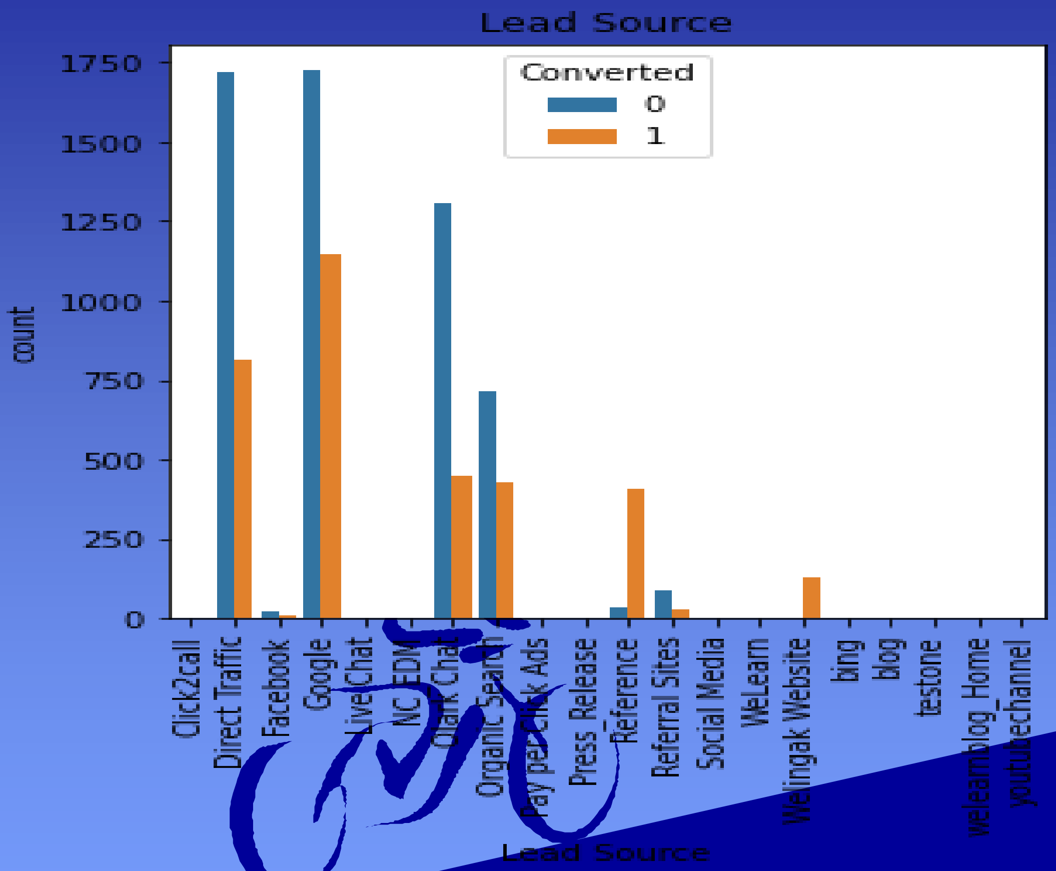




We can see that the max for Converted is higher than not converted
which means people spendingmore time on website are more likely to be converted

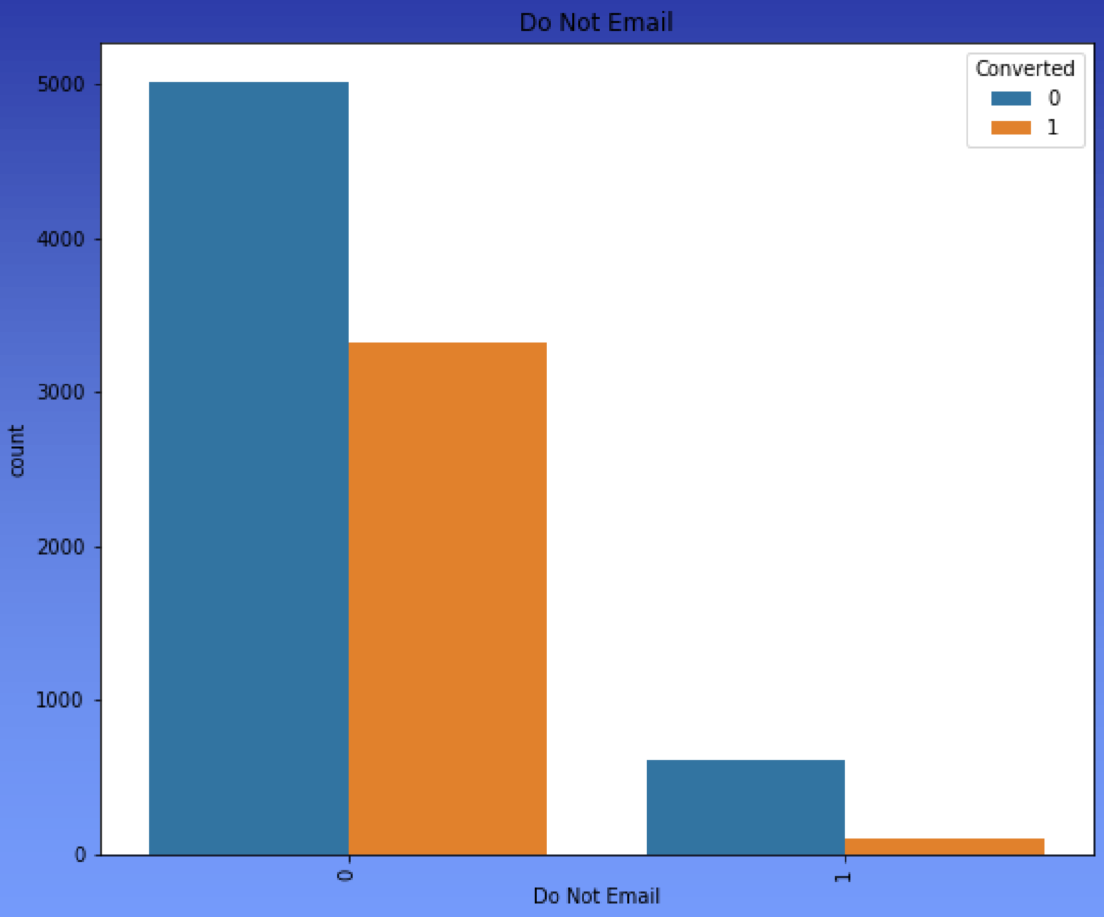
Lead Origin & Lead Source

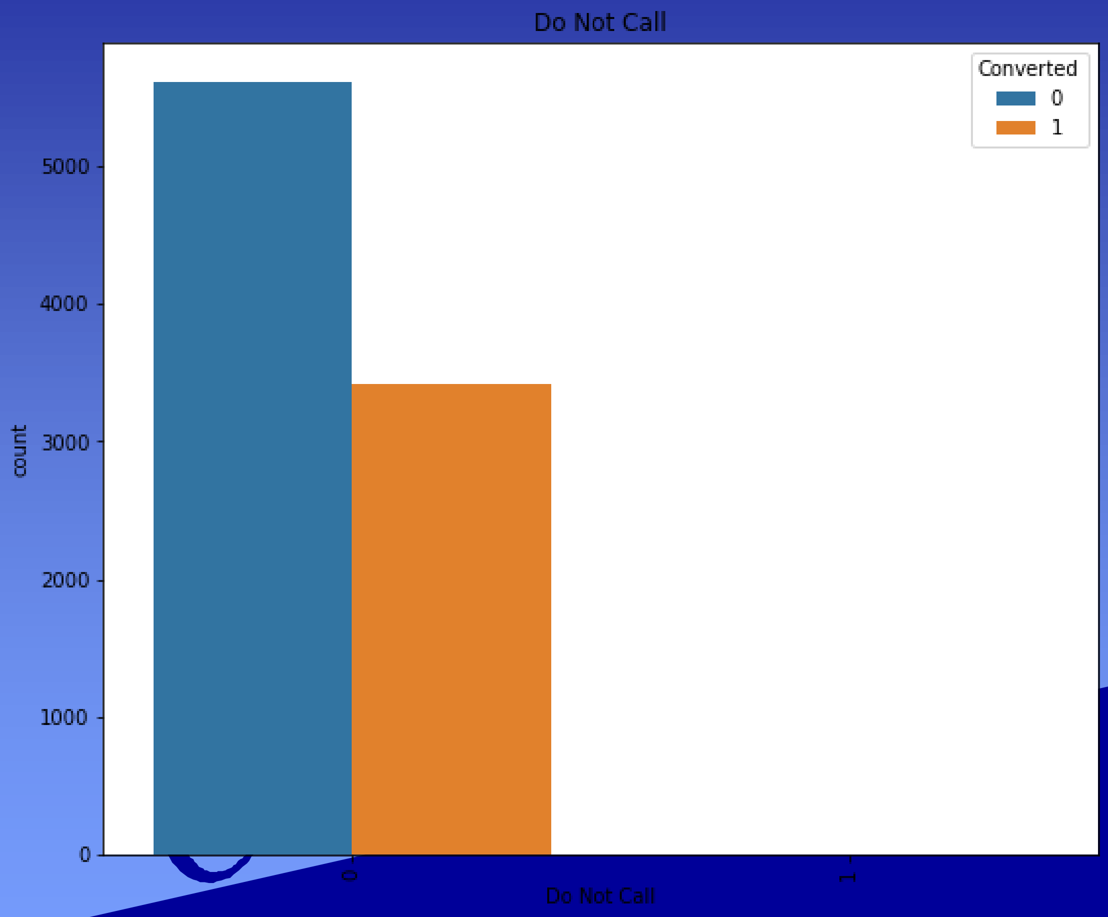




We can see that lead identified from lead add form have more converted rate than non conversion.
We can also see Google, Direct Traffic & Olark chat bringing many customers who converted.

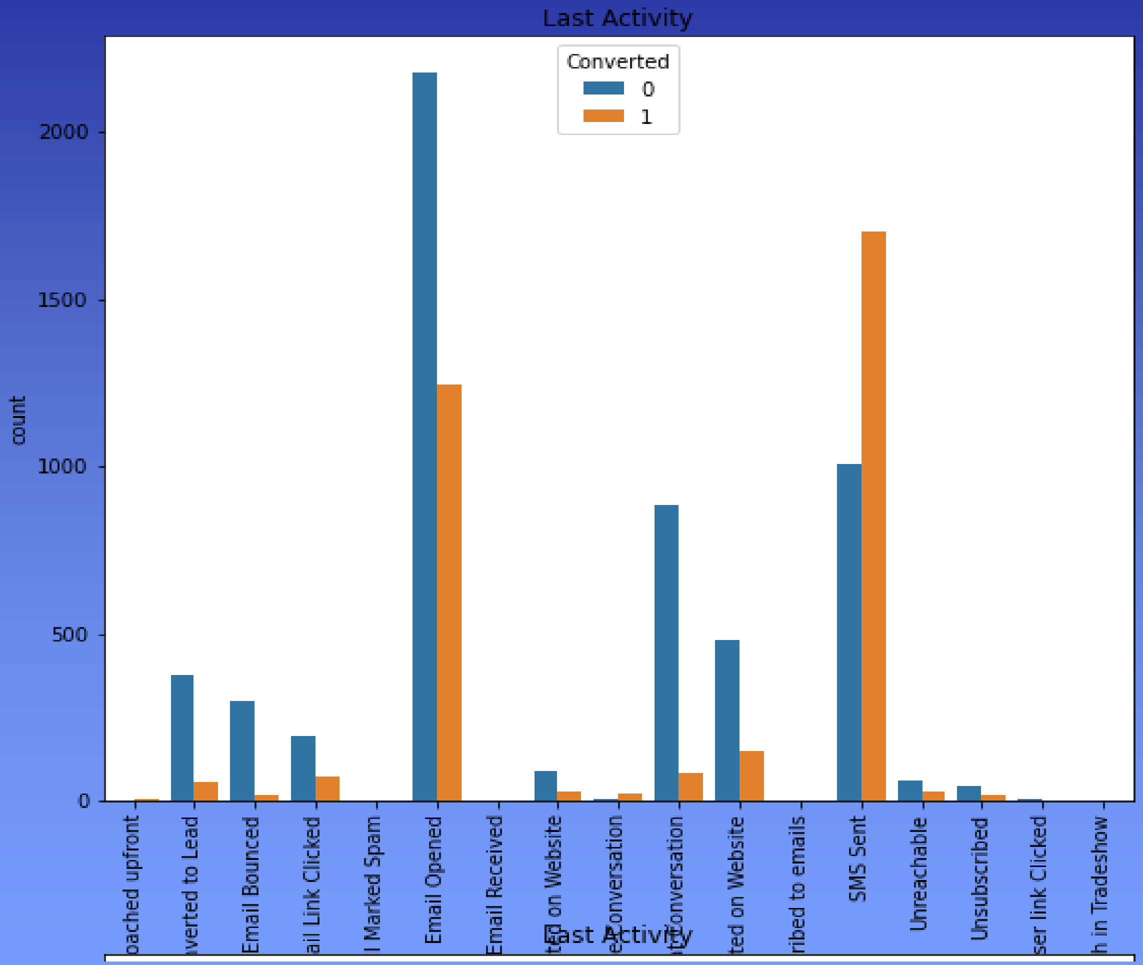
Do not Email & Call

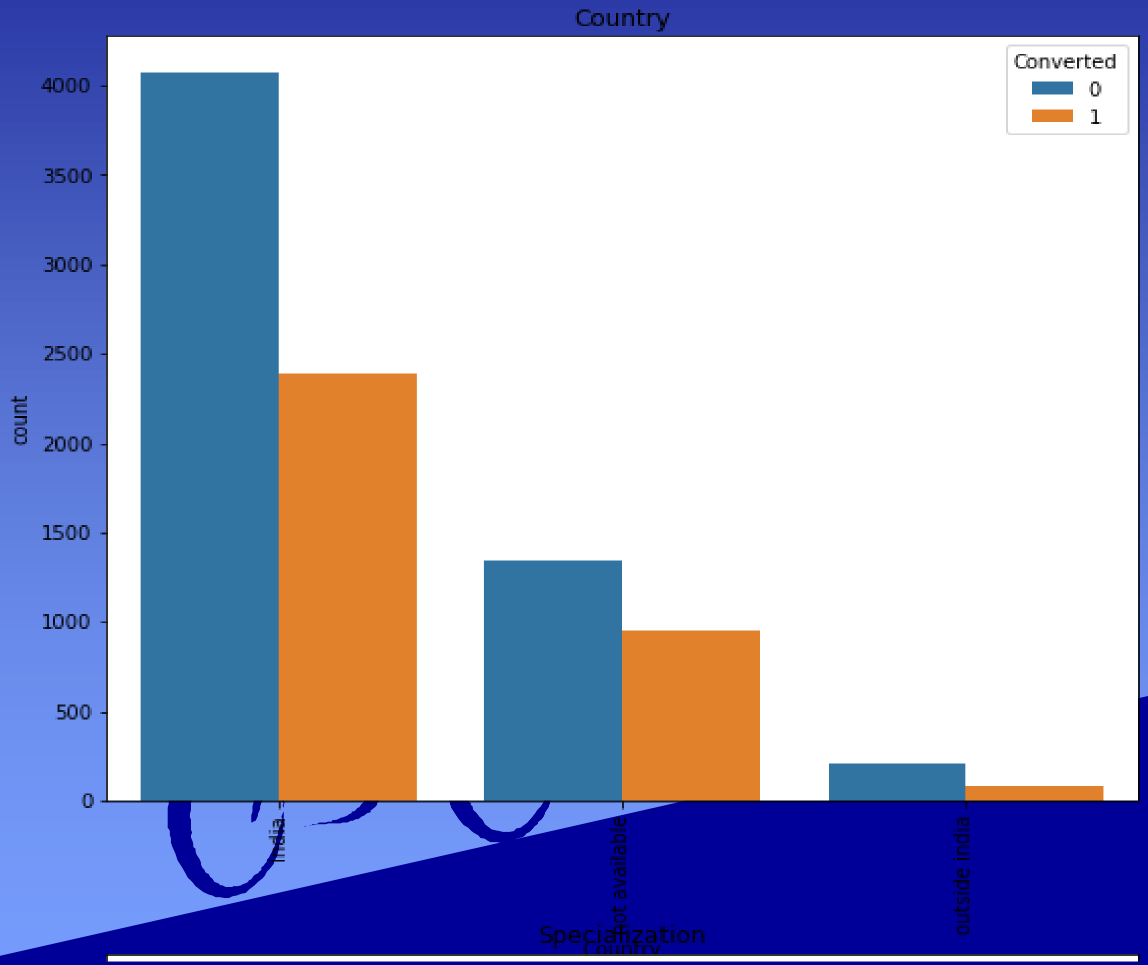




We can see that customer do not want to be called or e-mailed.

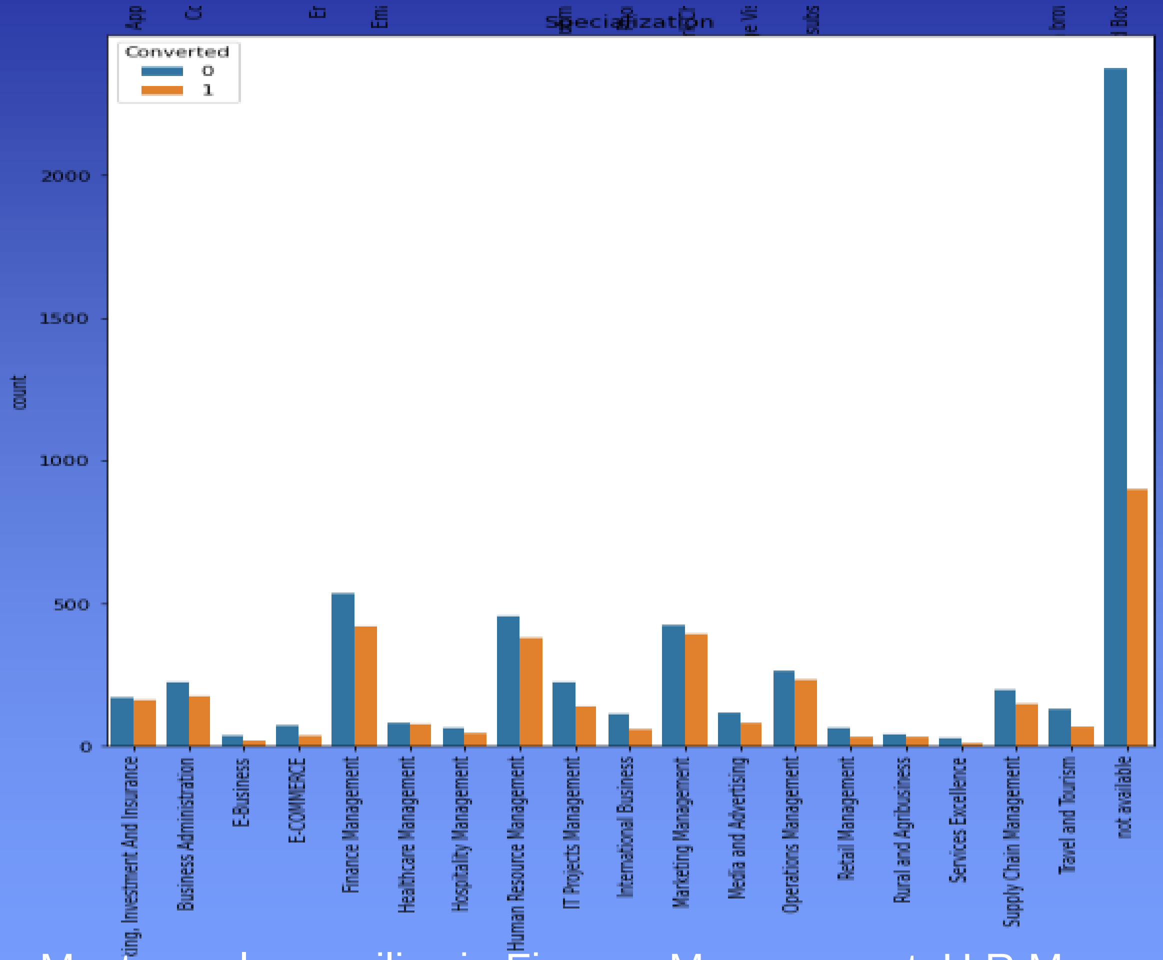
Last Activity & Country

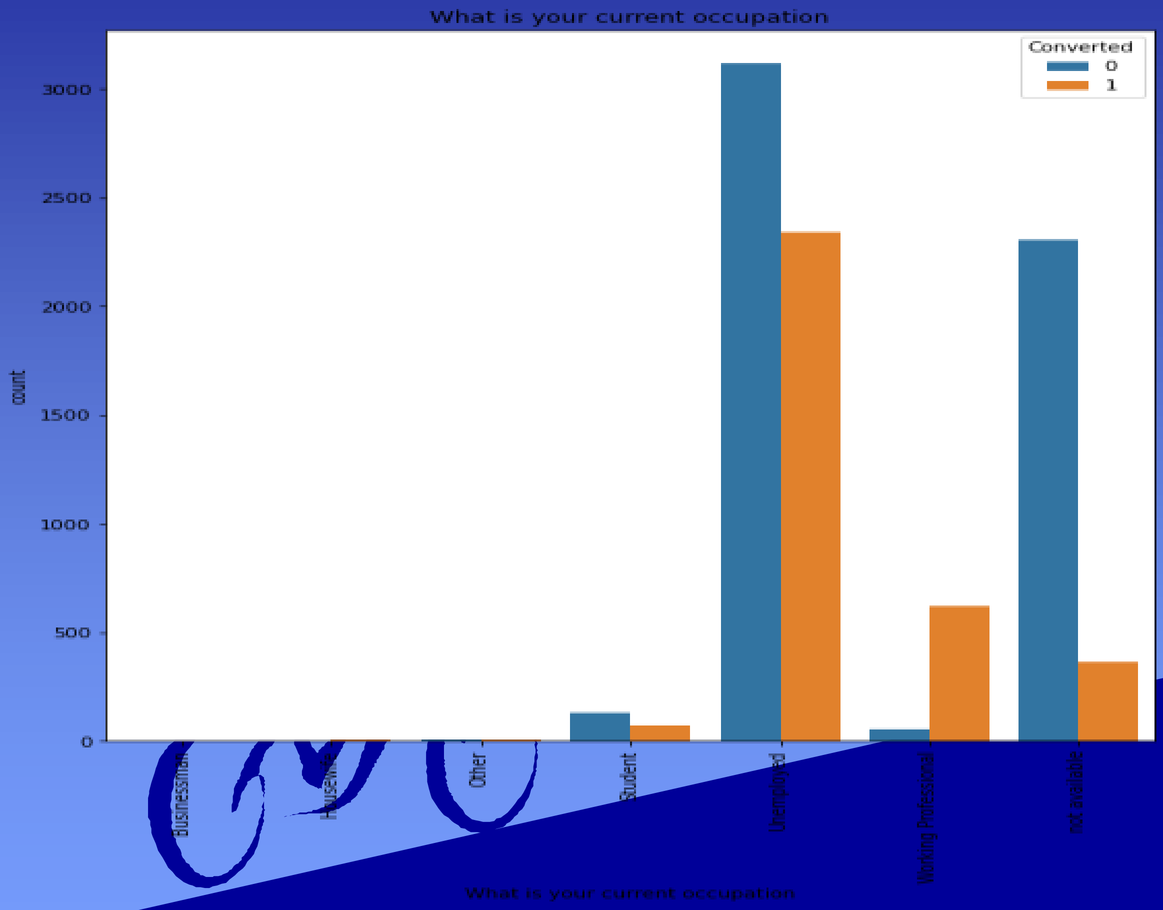




We see most Customers are from India.
The most Last activity are Email opened & SMS Sent.

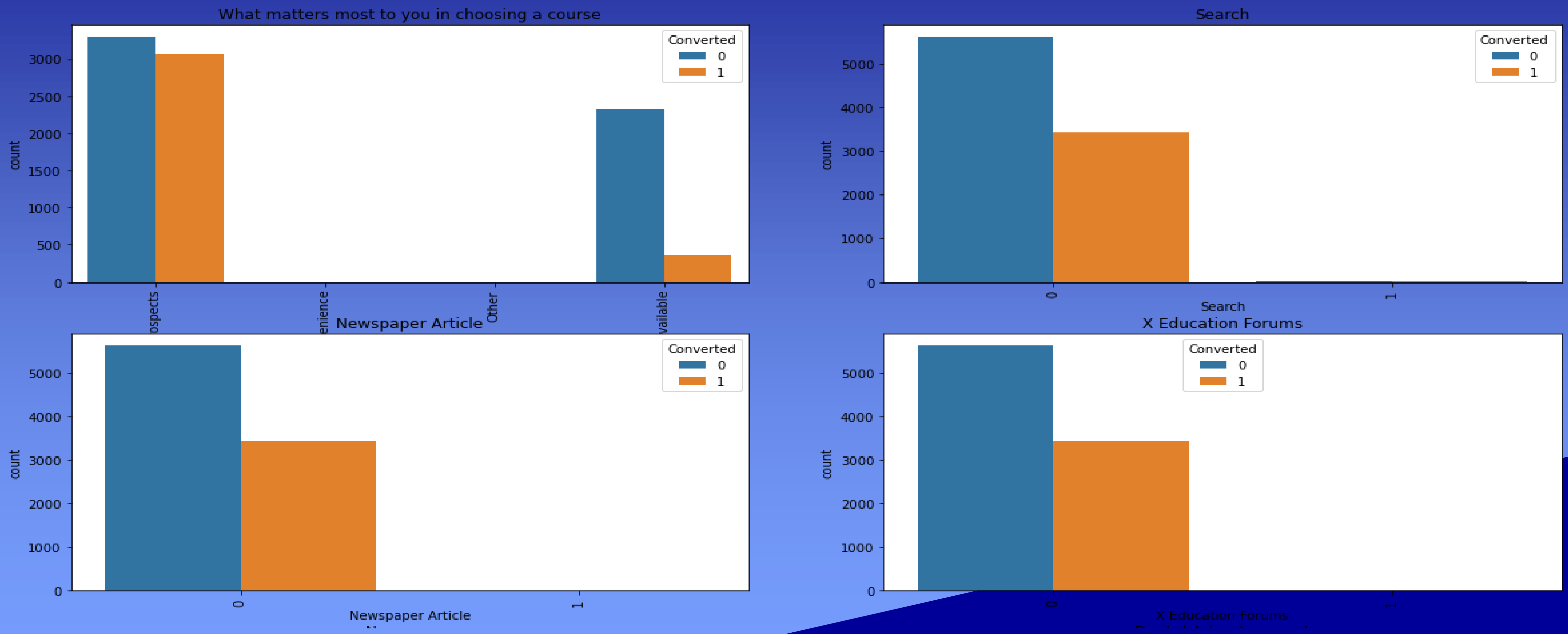
Specilization & Current Occupation





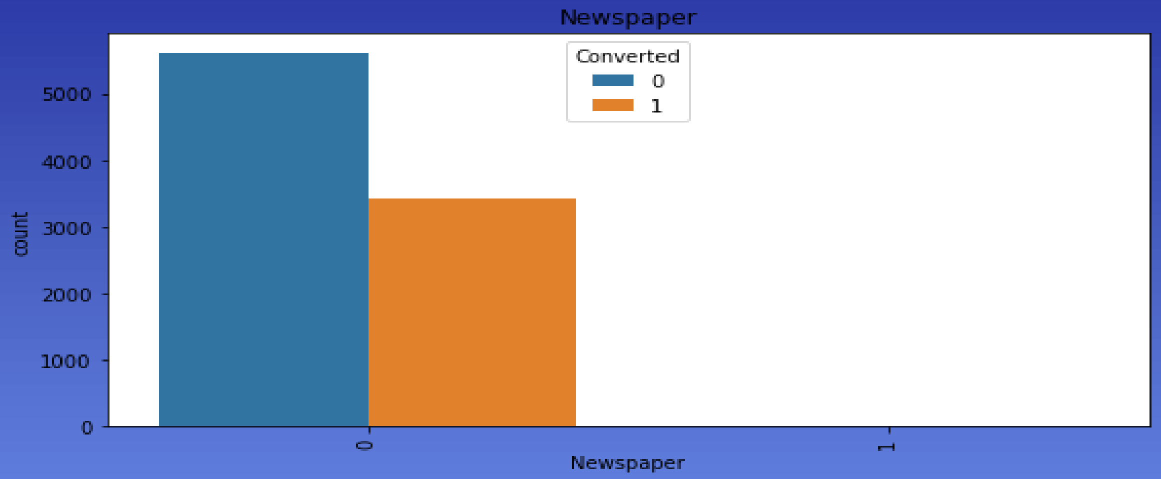
Most people specilize in Finance Management, H.R Management, Marketing Management and Operations Management.
Most people looking for the courses are Unemployed then next are Working Professional and some are students.

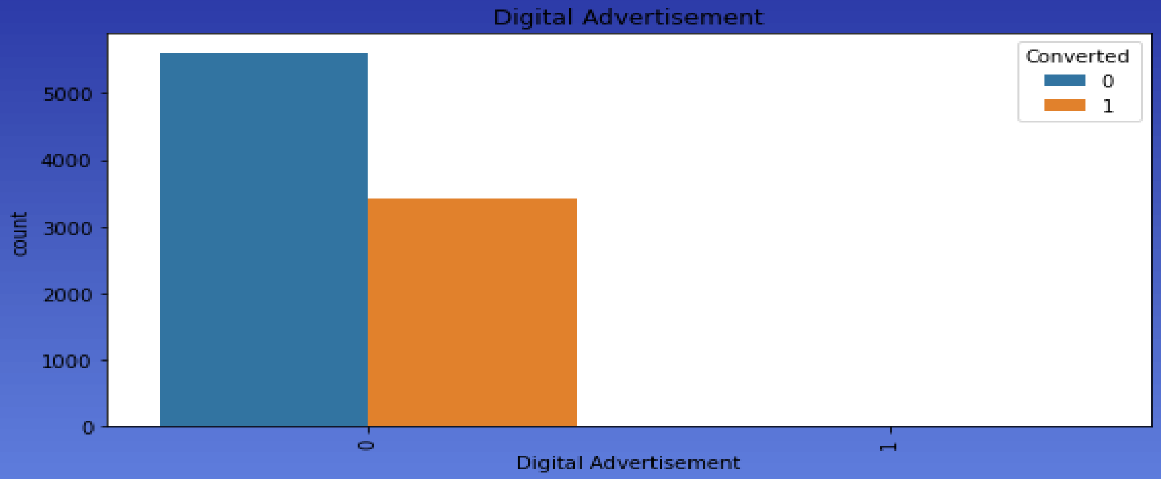
Reason for the Course & ADs

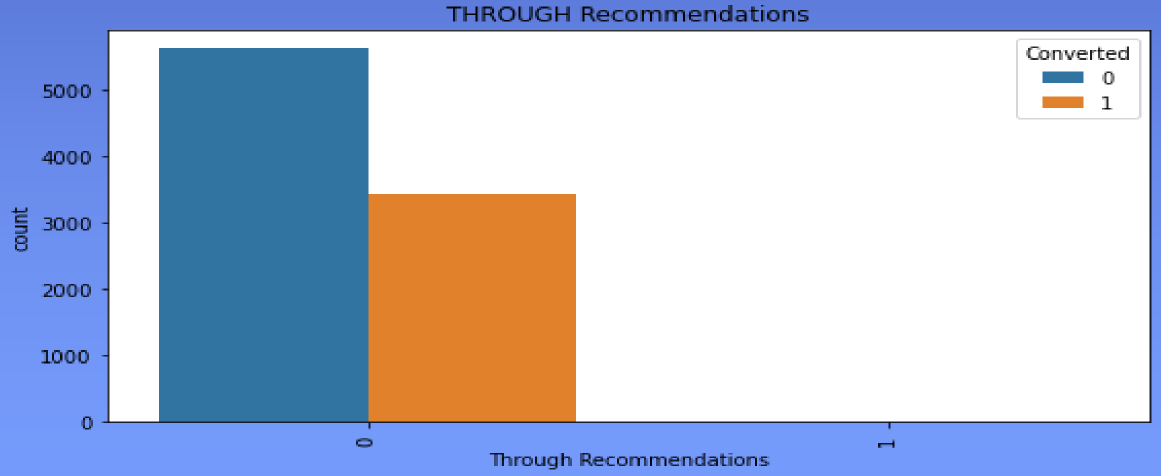


We see that all people are looking for better prospects.

ADs

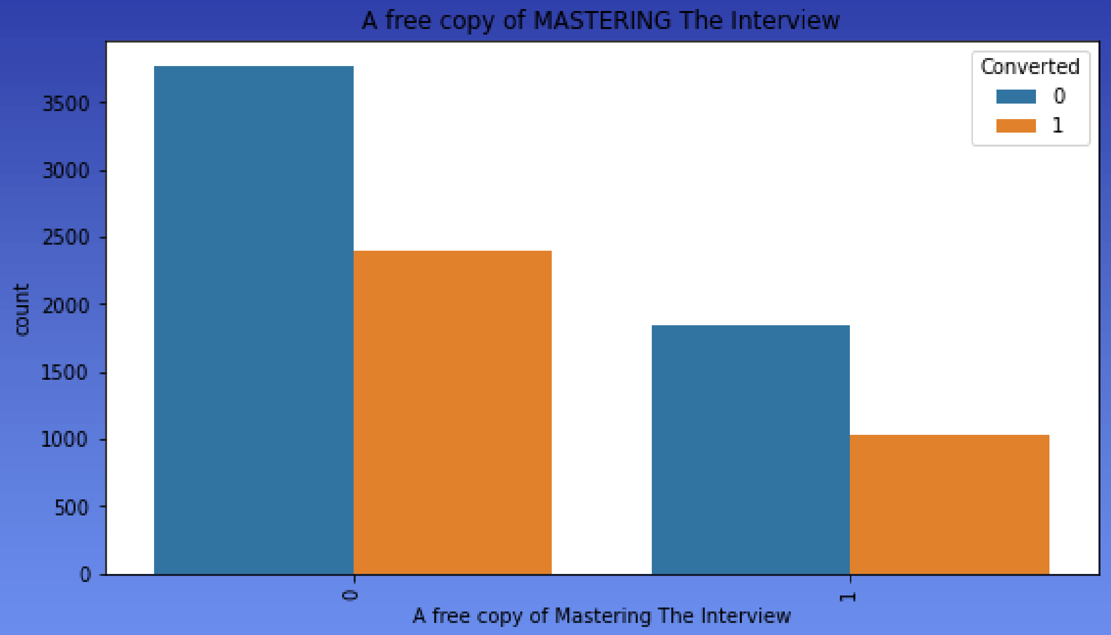


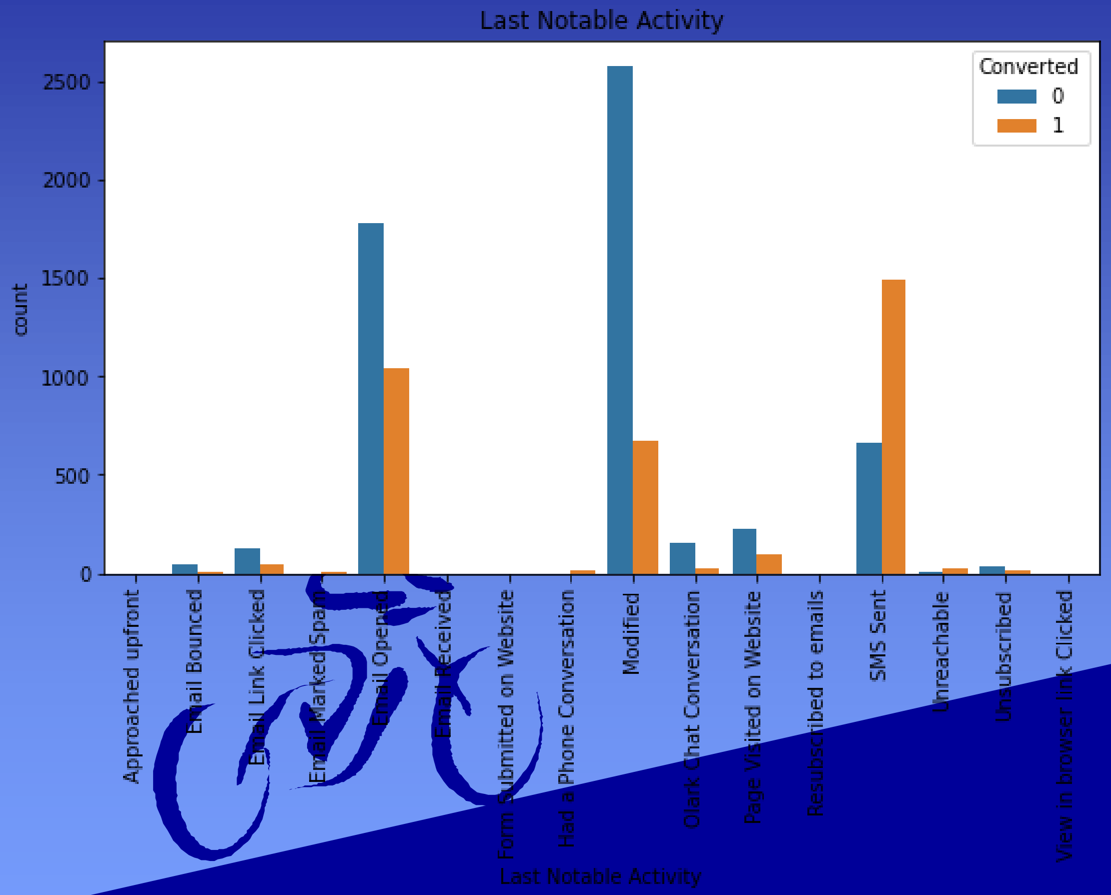




Through previous & this graph, we see that all people have reported that they have not seen any ads before coming.

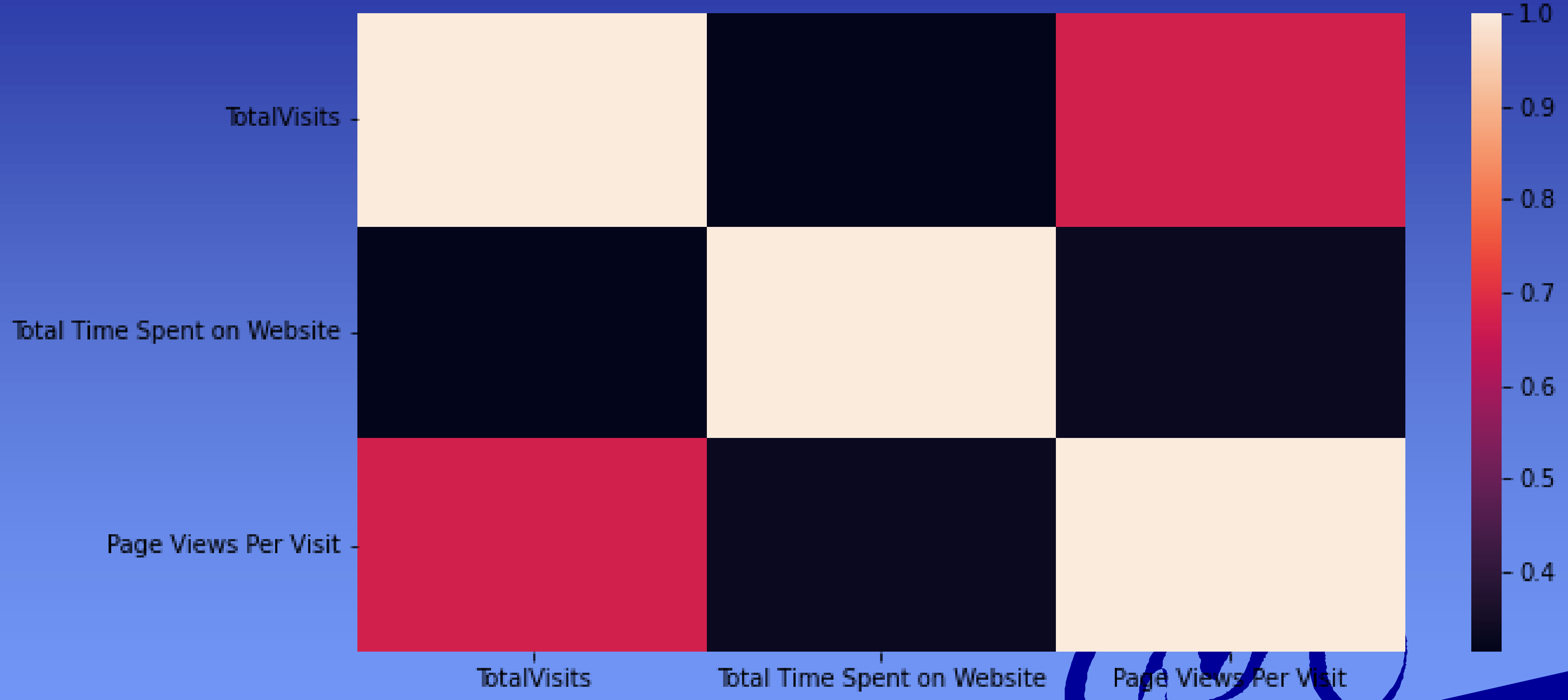
Free copy of Mastering the Interview
Last Notable Activity





We can see that people like a free copy of mastering the Interview.
We can see that the last notable activity are Modified,Email opened and SMS Sent.

Correlation

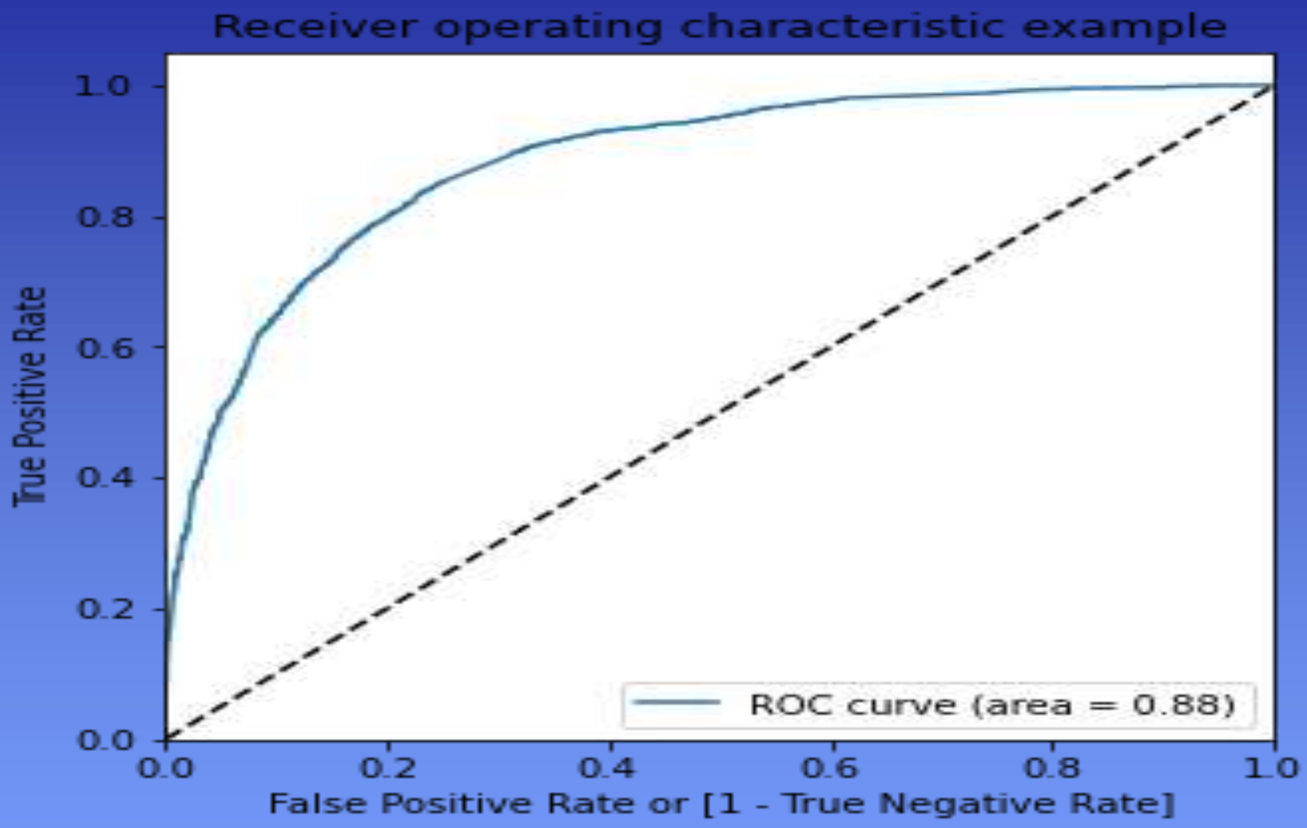


We can see that TotalVisits and Page Views Per Visits are correlated.

MODEL BUILDING

1. Splitting the data into Train & Test set
2. RFE
3. Scaling the variables
4. Building the model
5. Predicting on Test set
6. Evaluating the model

MODEL EVALUATION (TRAIN)

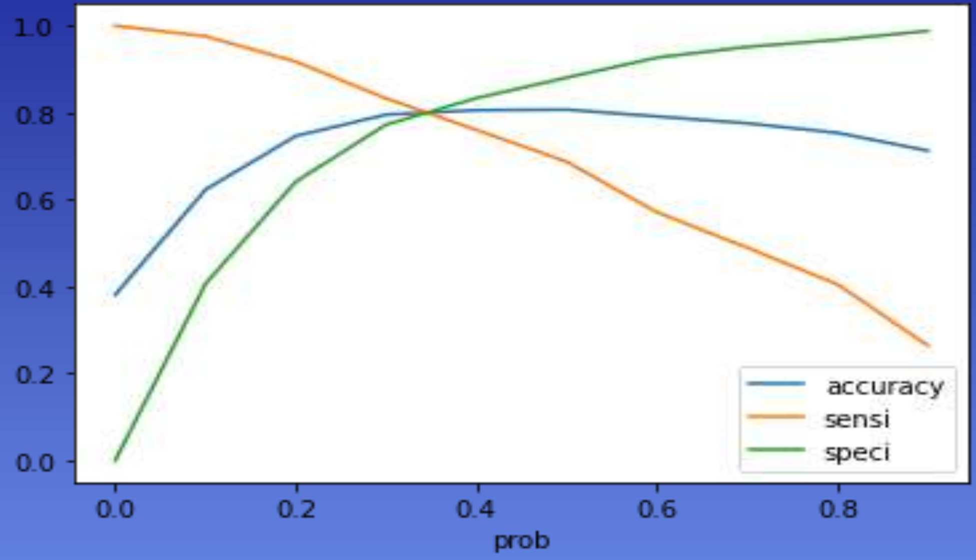


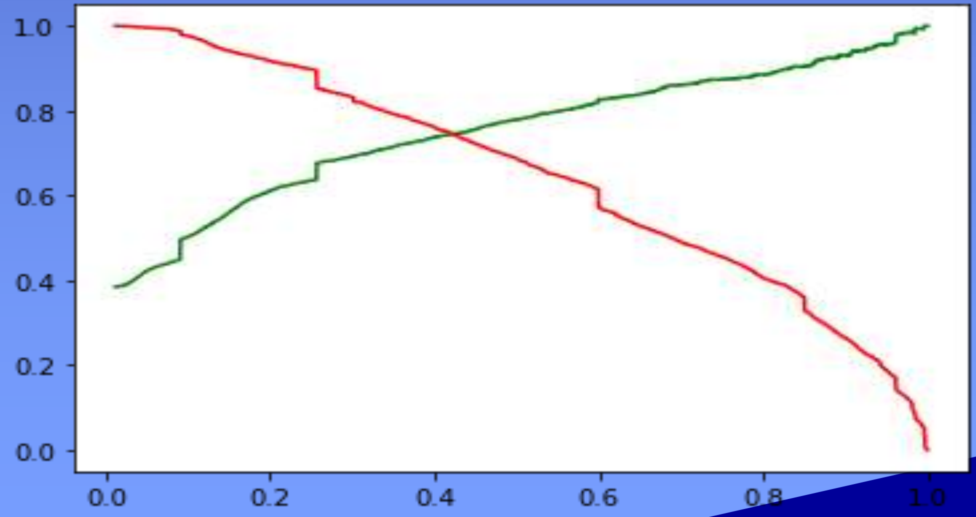
Confusion Matrix

[3161, 761]
[502, 1908]

The area under ROC Curve is 0.88 which is good.

MODEL EVALUATION (TRAIN)





WITH CUT-OFF AS 0.35

ACCURACY, SENSITIVITY & SPECIFICITY

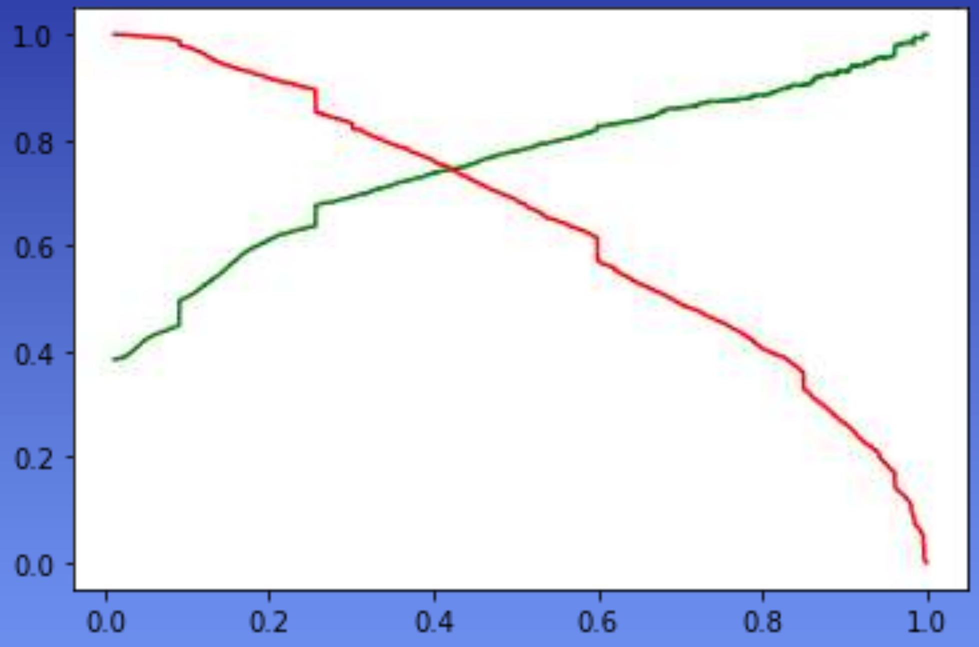
[3161, 761]
[502, 1908]

- 80% Accuaracy
- 79% Sensitivity
- 80% Specificity

PRECISION & RECALL

- 77% Precision
- 68% Recall

MODEL EVALUATION (TEST)



PRECISION & RECALL

74% Precision
77% Recall

Test cut-off has been set as 0.41

ACCURACY, SENSITIVITY & SPECIFICITY

```
[1425,  273]
[ 232,   784]
```

81% Accuracy
77% Sensitivity
83% Specificity

CONCLUSION

EDA:

- People spending more than average time are promising leads.
- Landing page submissions can be helpful for more leads.
- SMS can help in attracting and coverting leads.
- Finance Management, H.R Management, Marketing Management and Operations Management can be targeted more as they Convert more.
- Unemployed people are the most Customers then next are Working Professional and some are students.

Logistic Regression Model:

- The model shows the following data on the test set:

1. Accuracy: 81%
2. Sesitivity: 77%
3. Specificity: 83%
4. Precision: 74%
5. Recall Score: 77%

- The model correctly predicts leads who can convert and those who will not.
- Overall, the model is working correctly.