


**By:**  
**Dipsikha Mudoi**  
**&**  
**Oshin Dubey**



# **Lead scoring case study**



## 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.



## Goals of the Case Study

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.



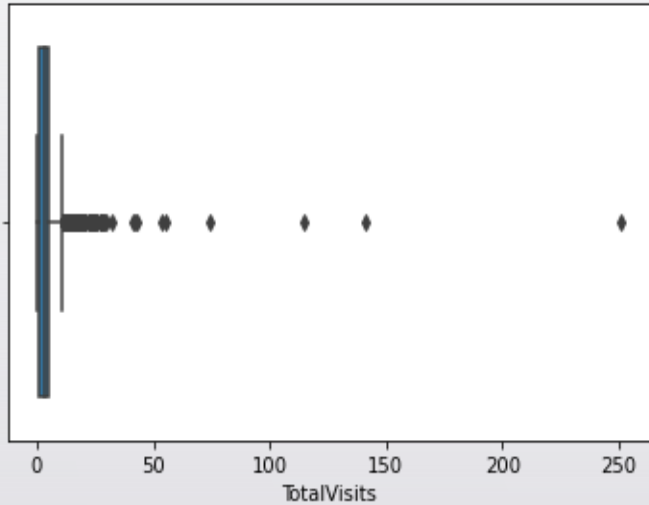
## Logistic Model Implementation

- Gathering Data : Source the data for Analysis.
- Reading & understanding the data.
- Data Cleaning : Duplicate removal, Missing value treatment etc.
- Data Preprocessing : Outliers Treatment, Removing unnecessary columns
- Performing EDA: Univariate, bivariate analysis, Multivariate (heat map )
- Prepare the Data for Modelling by dummy variable creation to the categorical columns, splitting Data into Train and Test set and scaling continuous feature in train and test data
- Building Model with Train data set.
- Making Predictions on the Train data set.



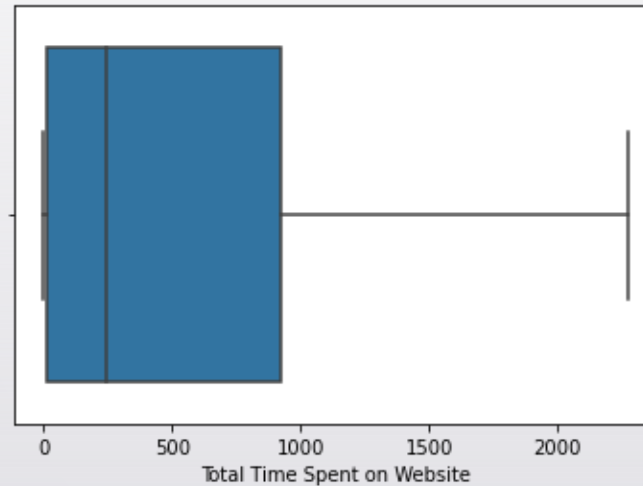
# Outlier Identification

Outlier in TotalVisits



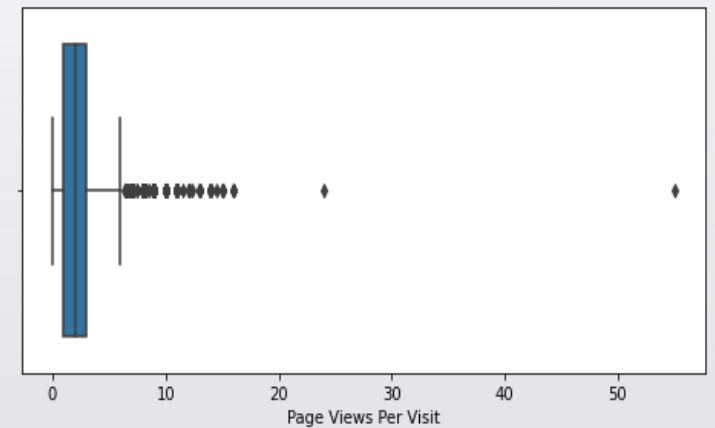
- Outlier present in column TotalVisits

Outlier in Total Time Spent on Website



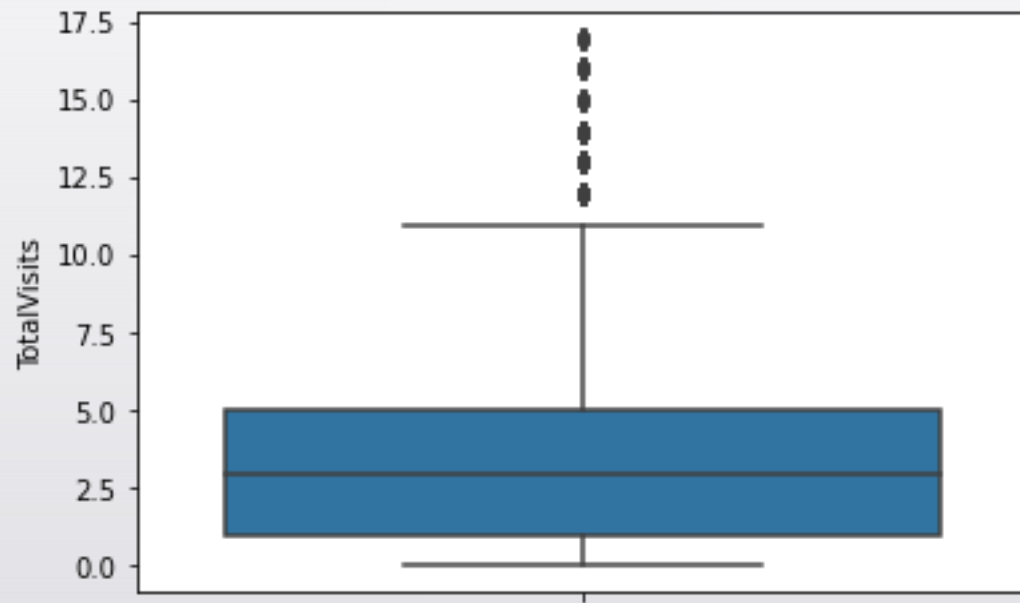
- No outlier present in Total Time Spent on Website. So no treatment is required.

Outlier in Page Views Per Visit

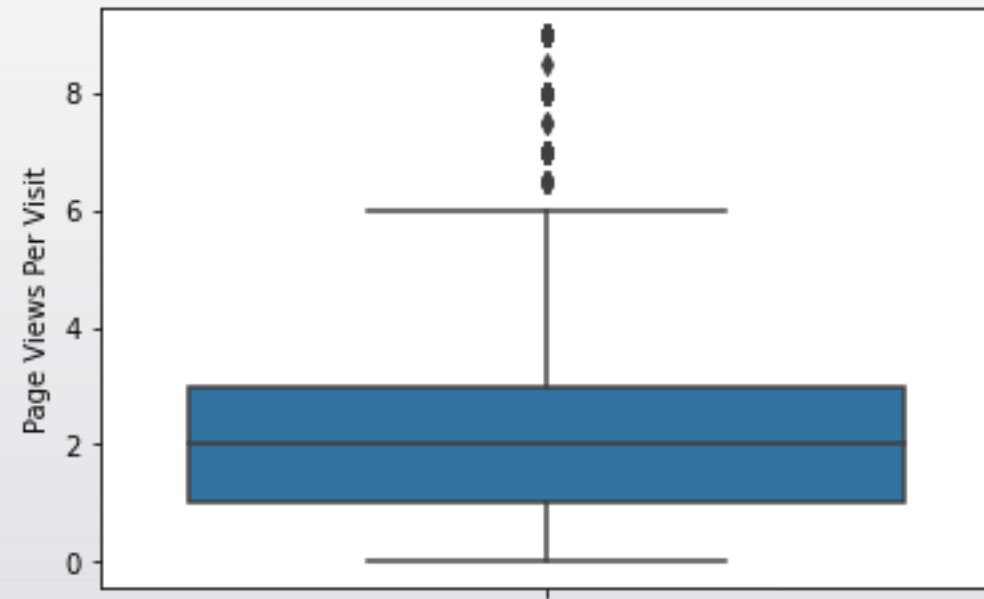


- Outlier present in Page Views Per Visit column.

Outlier Treatment:  
After Removing 1% top & bottom Outlier values

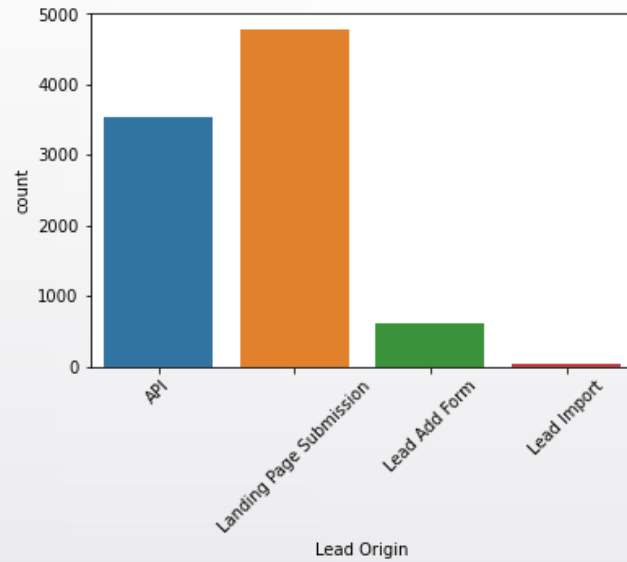


Total Values



Page Views per visit

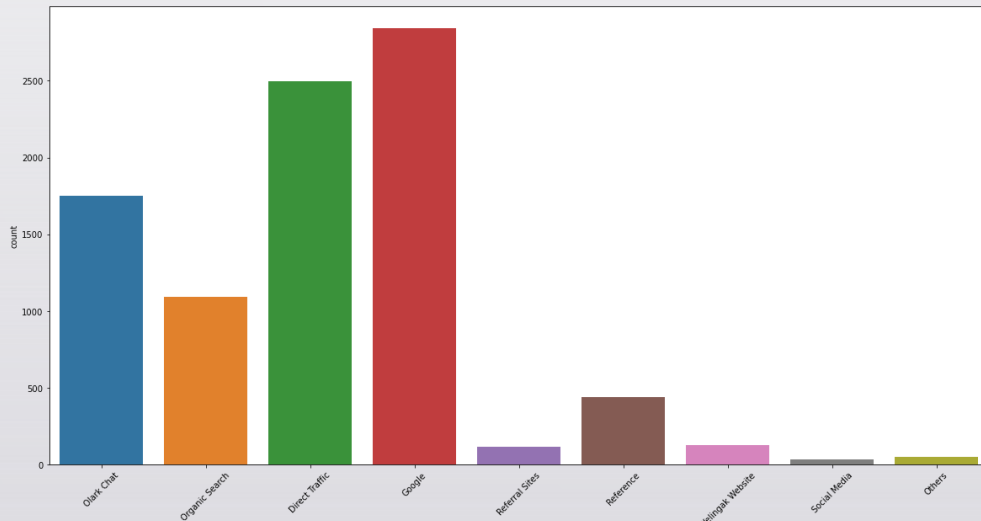
Lead Origin



## Univariate Analysis

--Based on Origin, highest lead origin is landing Page submission. Second is API.

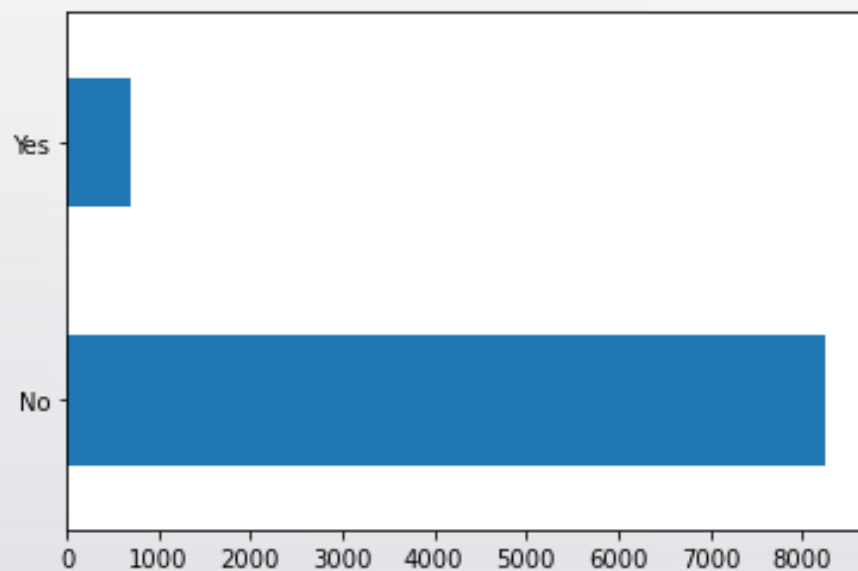
Lead Source



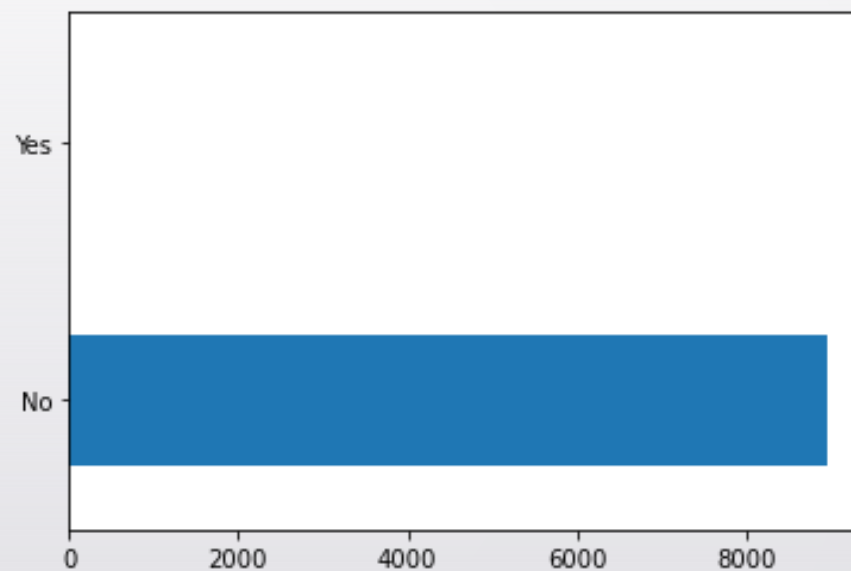
--Based on source ,Google has highest lead. Second highest lead source is direct traffic and third is Olark Chat.



Do Not Email



Do Not Call

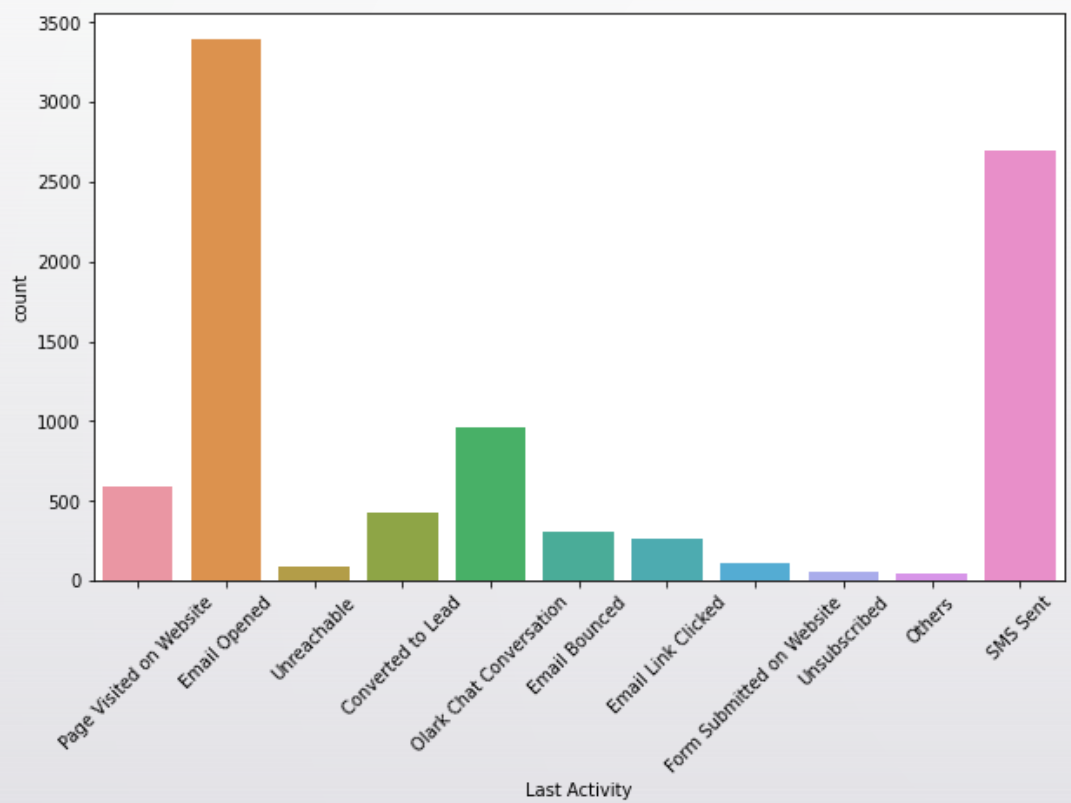


Data Imbalance present in column 'Do Not Call'



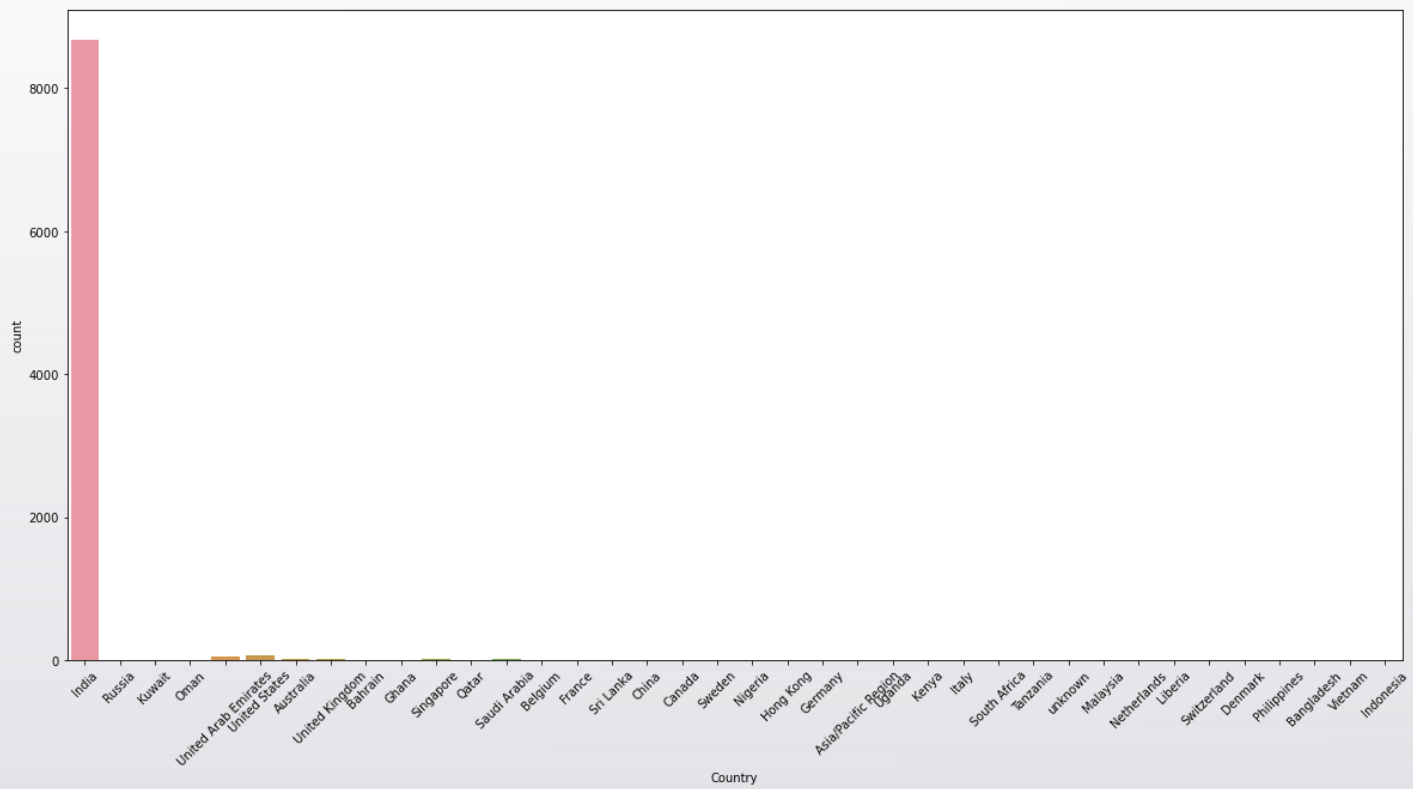


Last Activity



Highest last activity is Email opened,2nd is SMS sent,3rd is Olark Chat Conversation.

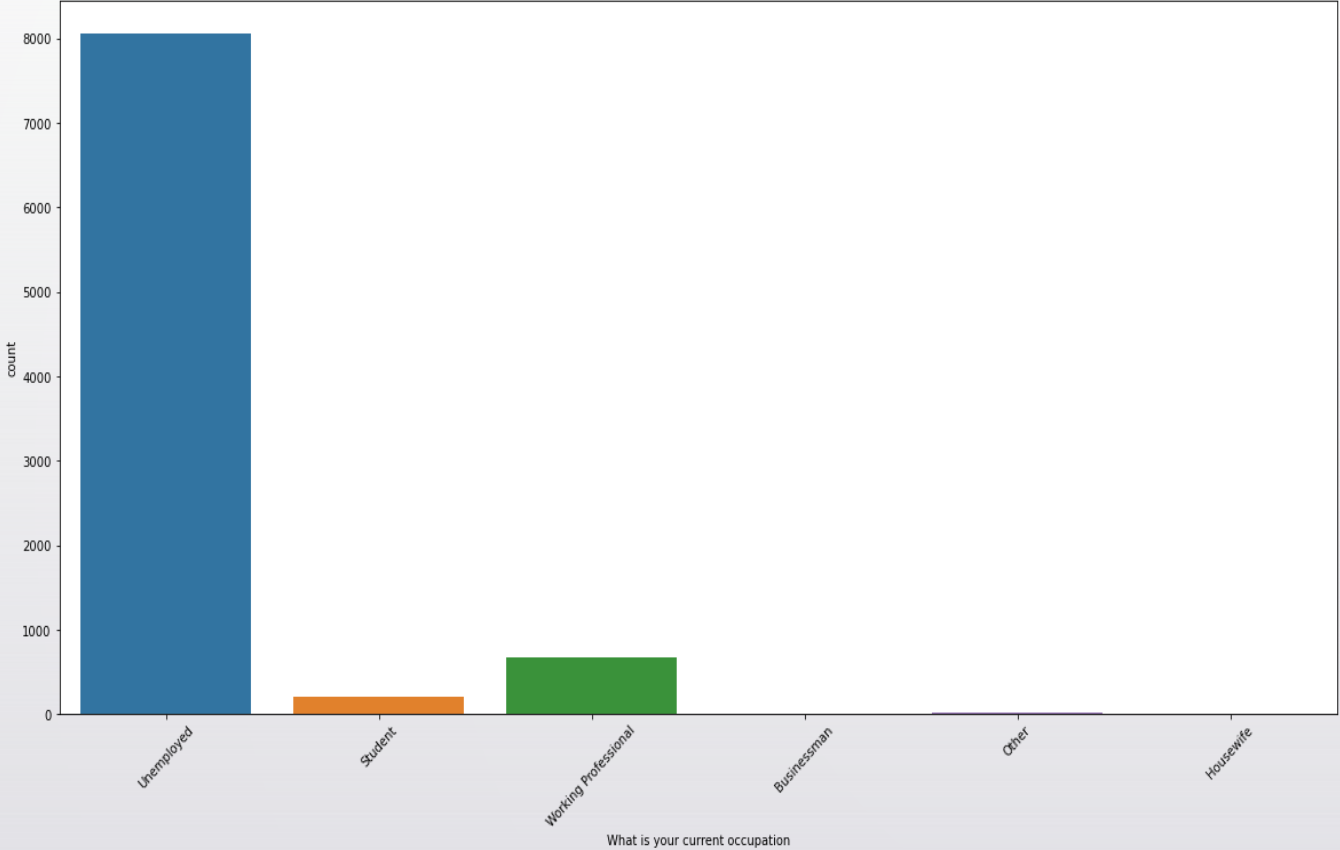
Country



Data imbalance is present in Country column because majority of data belongs to class 'India'

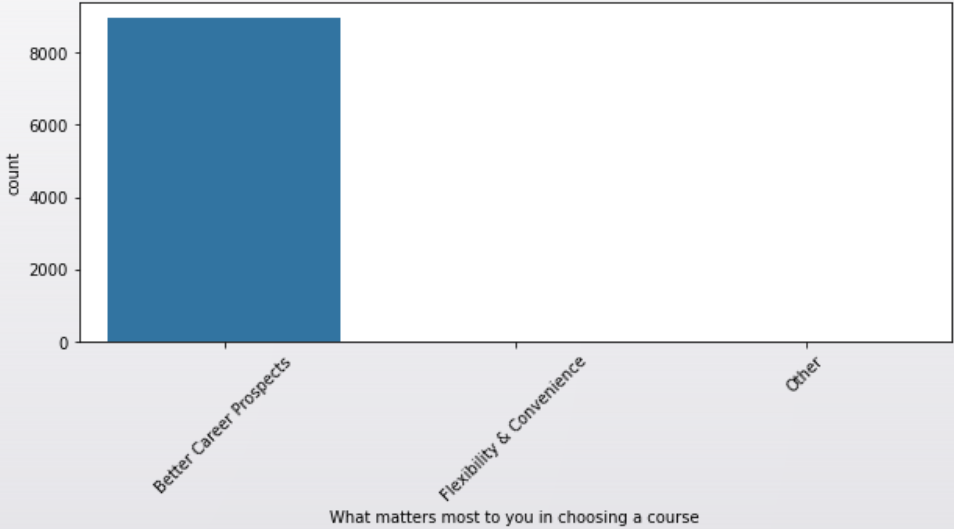


What is your current occupation



--Unemployed leads are highest in number.  
--Working Professionals second highest in number.

What matters most to you in choosing a course

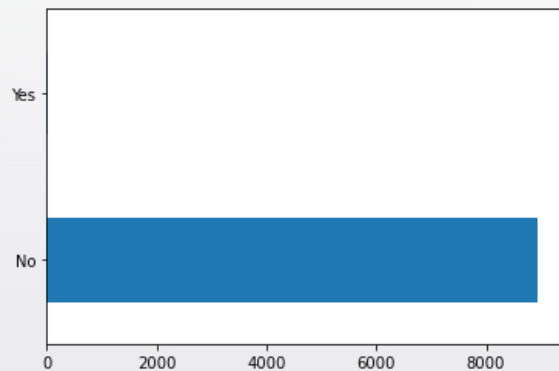


Will choose course according to Better career Prospects.

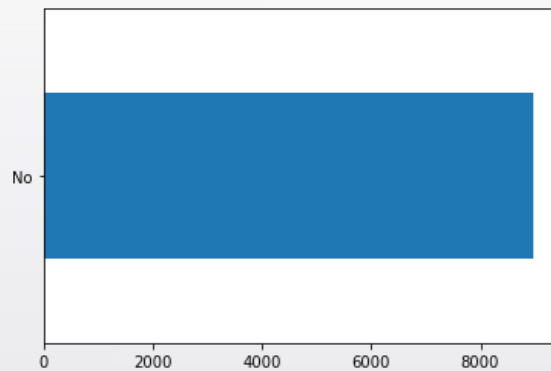


# Highly Imbalance data

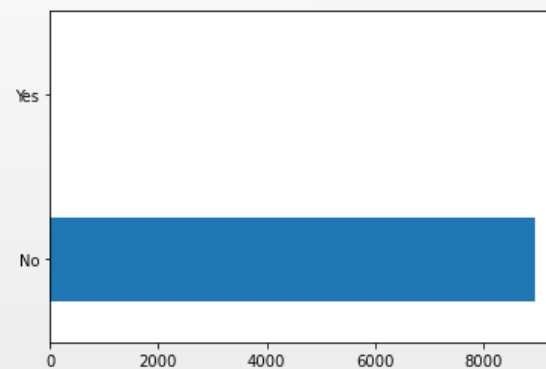
Search



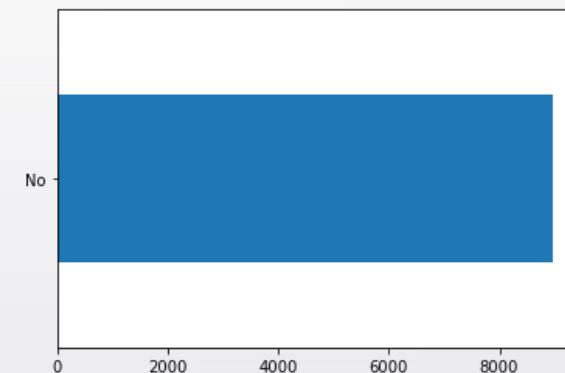
Magazine



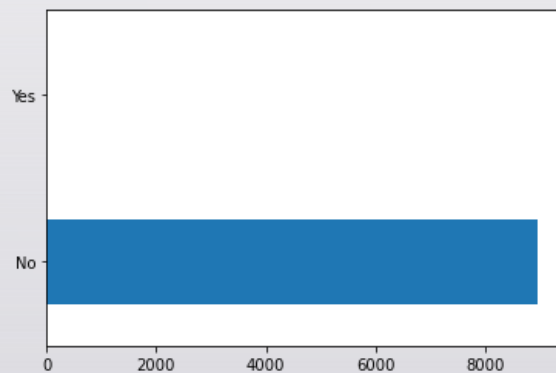
Newspaper Article



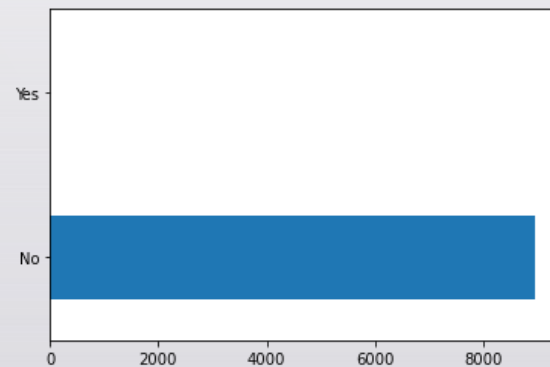
X Education Forums



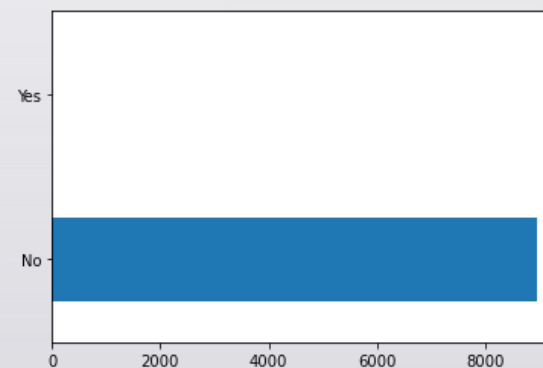
Newspaper



Digital Advertisement

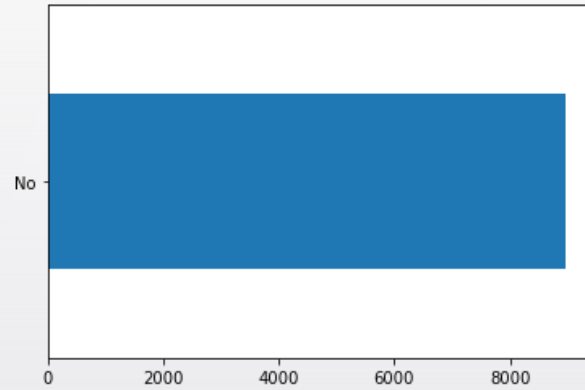


Through Recommendations

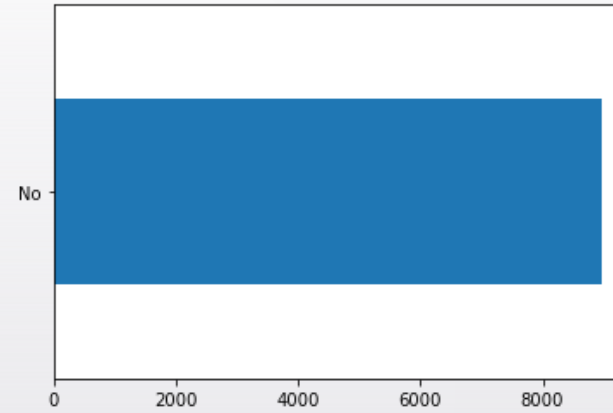




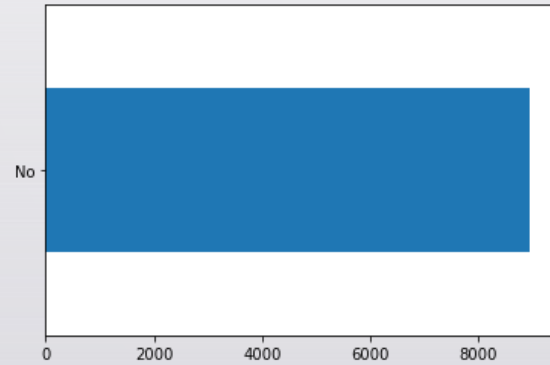
Receive More Updates About Our Courses



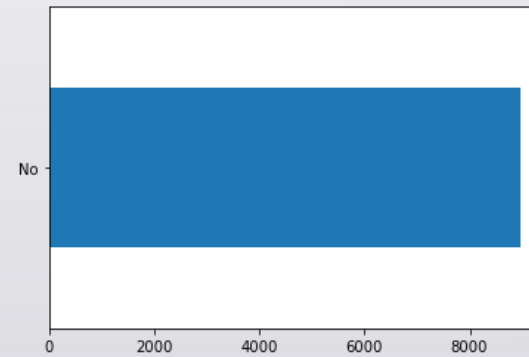
Update me on Supply Chain Content



I agree to pay the amount through cheque

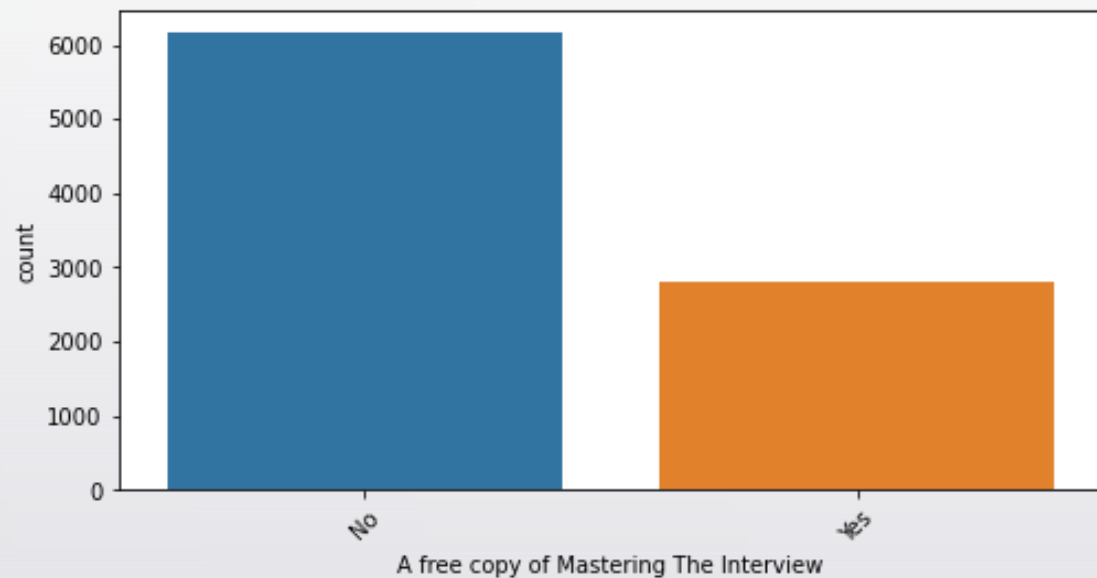


Get updates on DM Content

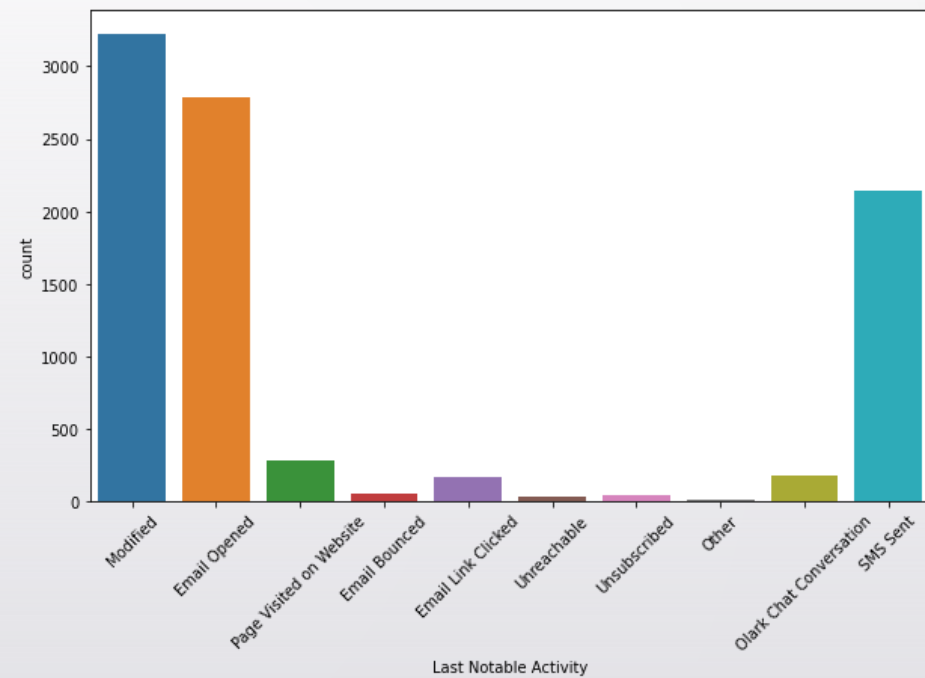




## A free copy of Mastering The Interview



## Last Notable Activity

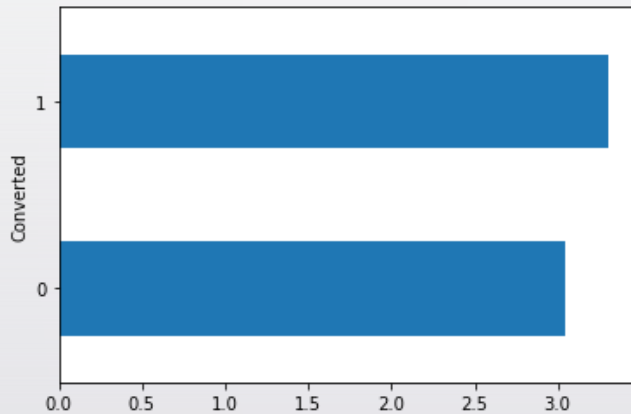


Modified, Email Opened, SMS sent are the three major last notable activity.



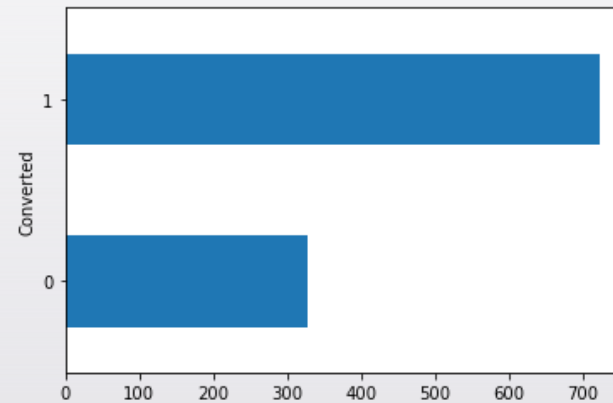
## Bivariate Analysis(Numerical- Categorical analysis)

Status of Total Visits  
Converted Vs Non-Converted



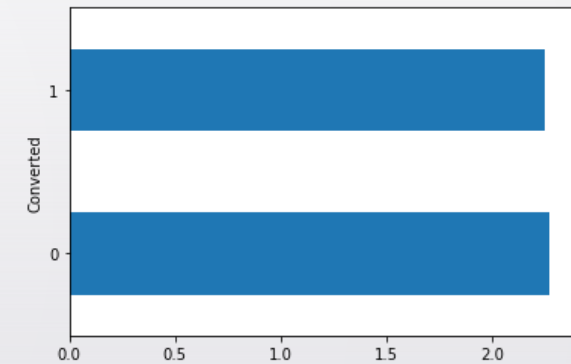
Out of all Visitors to the website, successfully converted lead is higher than non-converted one

Status of Total Time Spent on Website  
Converted Vs Non-Converted



Depending on Total Time Spent on Website, successfully converted lead is higher than non-converted one

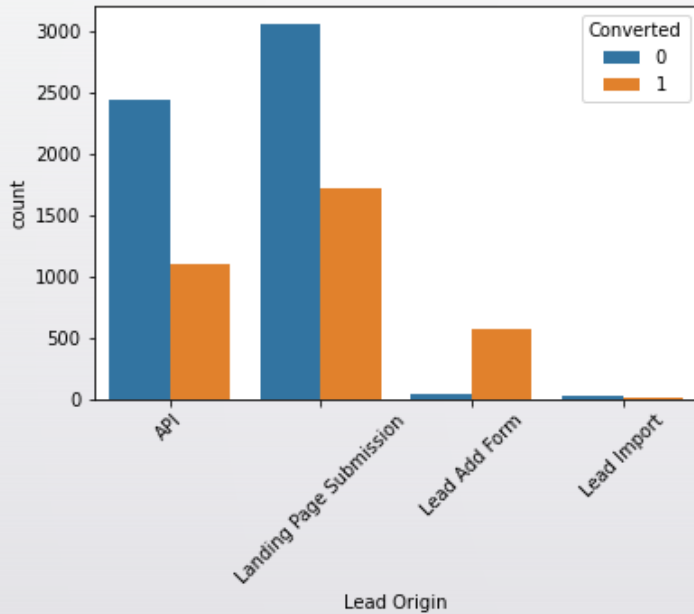
Status of Page Views Per Visit  
Converted Vs Non-Converted



There is no much difference between converted and non-converted lead based on page views per visits

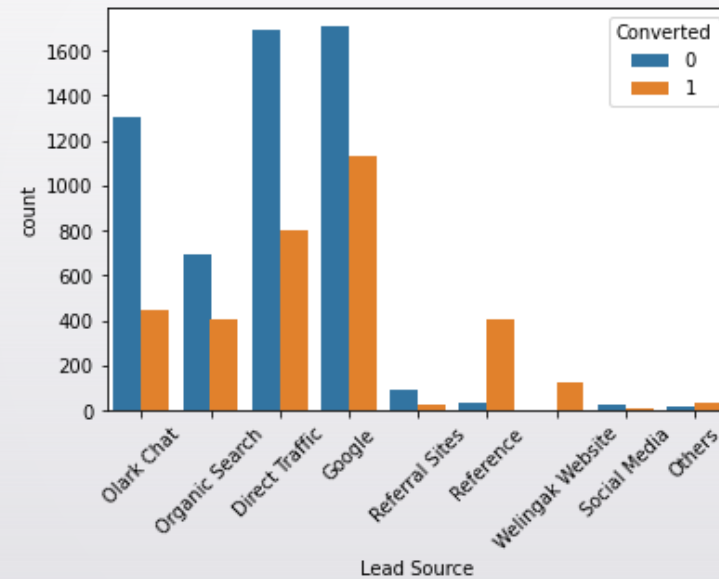
# Categorical – Categorical Analysis

Based on Lead Origin  
Converted Vs Non-Converted



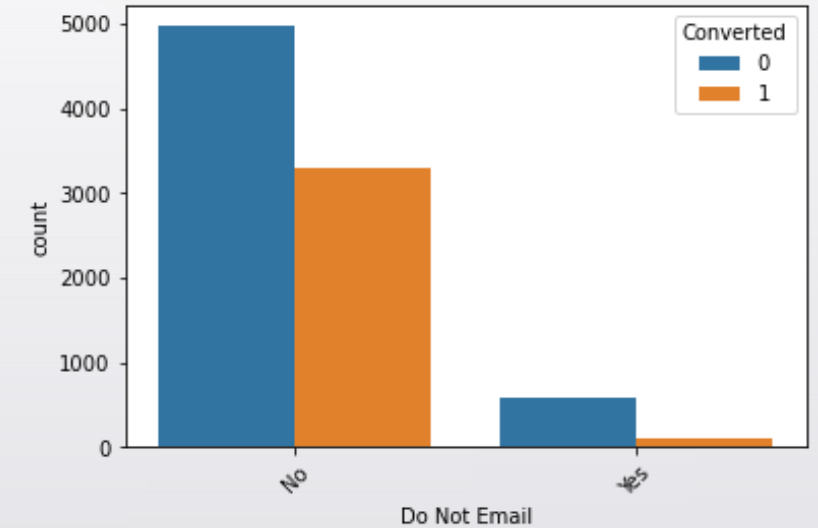
--Based on Lead Origin, converted leads are lower than non-converted leads  
--Highest converted leads are from Landing Page Submission.

Based on Lead Source  
Converted Vs Non-Converted



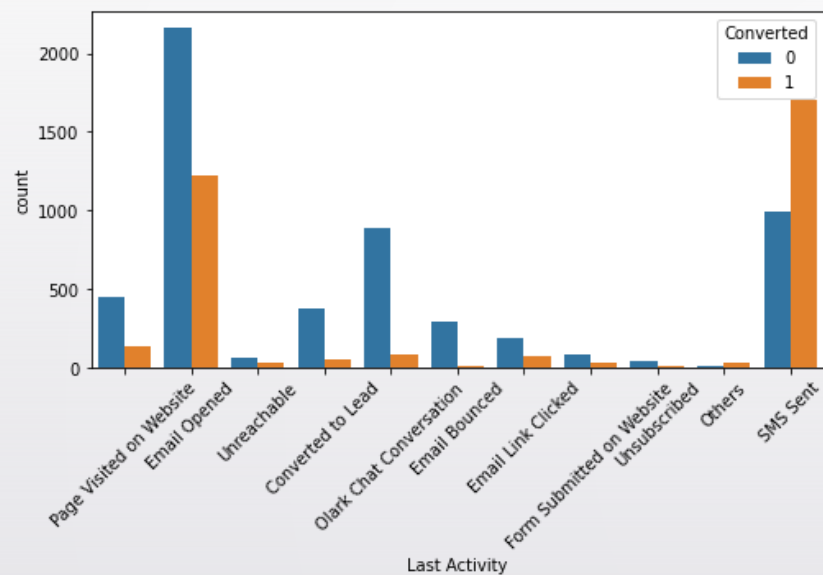
--Based on Lead Source, converted leads are lower than non-converted leads  
--Highest converted leads are from Google.

Based on 'Do Not Email' column  
Converted Vs Non-Converted



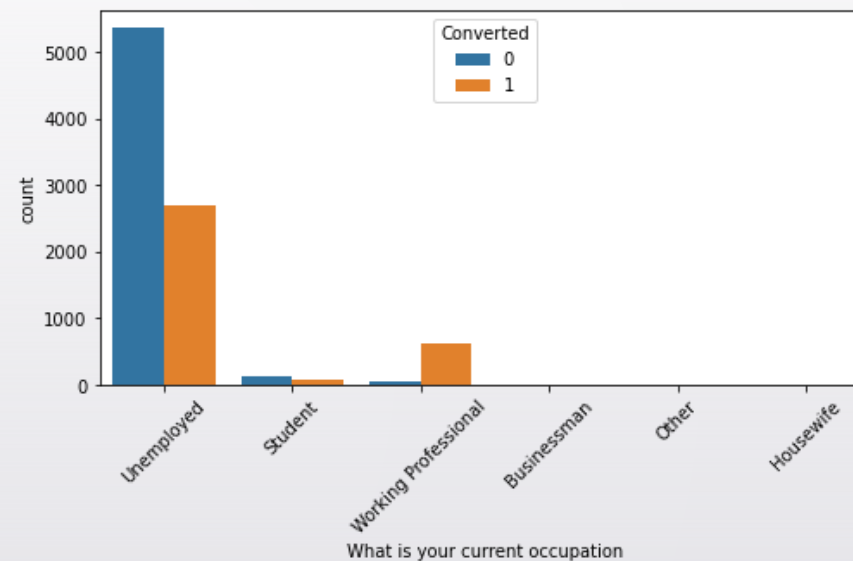
--Based on Do Not Email, converted leads are lower than non-converted leads

Based on 'Last Activity' column  
Converted Vs Non-Converted



- Last activity SMS sent has the highest converted leads
- Last activity email opened has 2nd highest converted leads

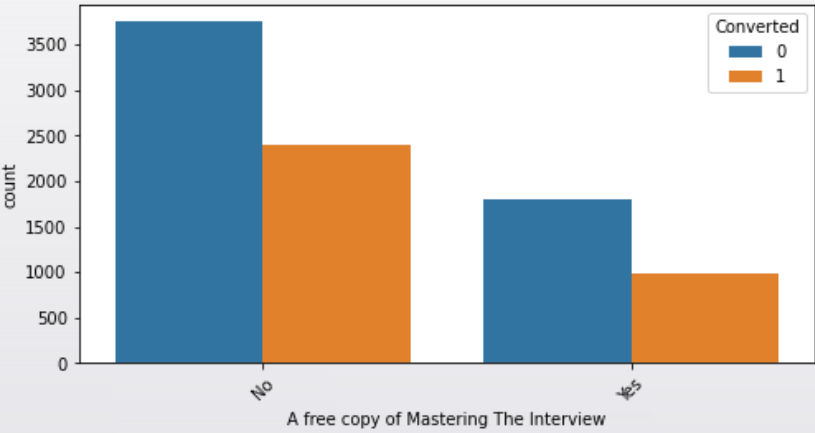
Based on 'What is your current occupation' column  
Converted Vs Non-Converted



Unemployed are the highest converted leads. Working Professional are in second possible leads.

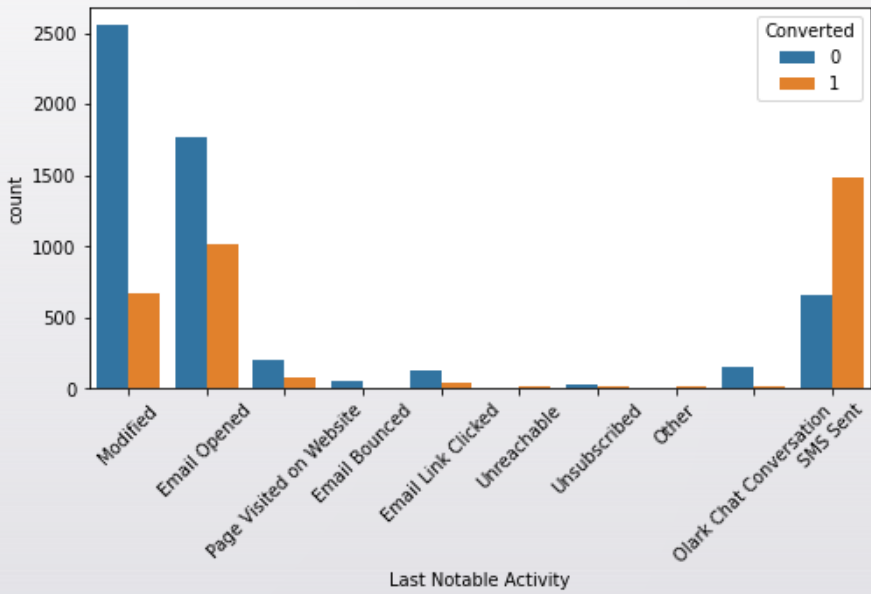


Based on Column 'A free copy of Mastering The Interview' column  
Converted Vs Non-Converted



Not preferring 'A free copy of mastering' has non-converted leads

Based on 'Last Notable Activity' column  
Converted Vs Non-Converted

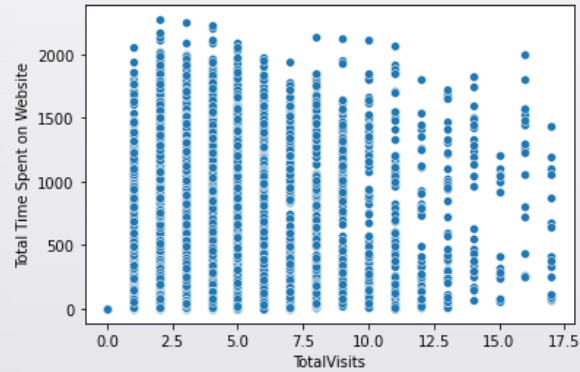


SMS sent has highest converted leads.



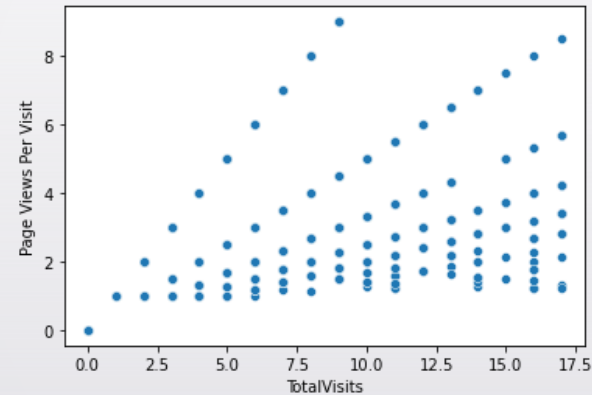
# Numerical- Numerical Analysis

Scatter Plot  
Total Visits And Total Time Spent on Website



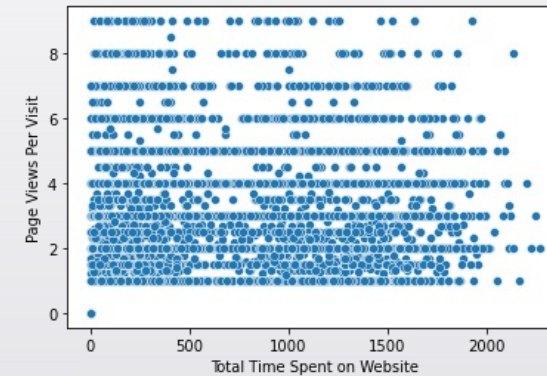
There is no correlation between Total Visits And Total Time Spent on Website

Scatter Plot  
Total Visits And Page Views Per Visit



There is very poor correlation between Total Visits And Page Views Per Visit

Scatter Plot  
Total Time Spent on Website And Page Views Per Visit



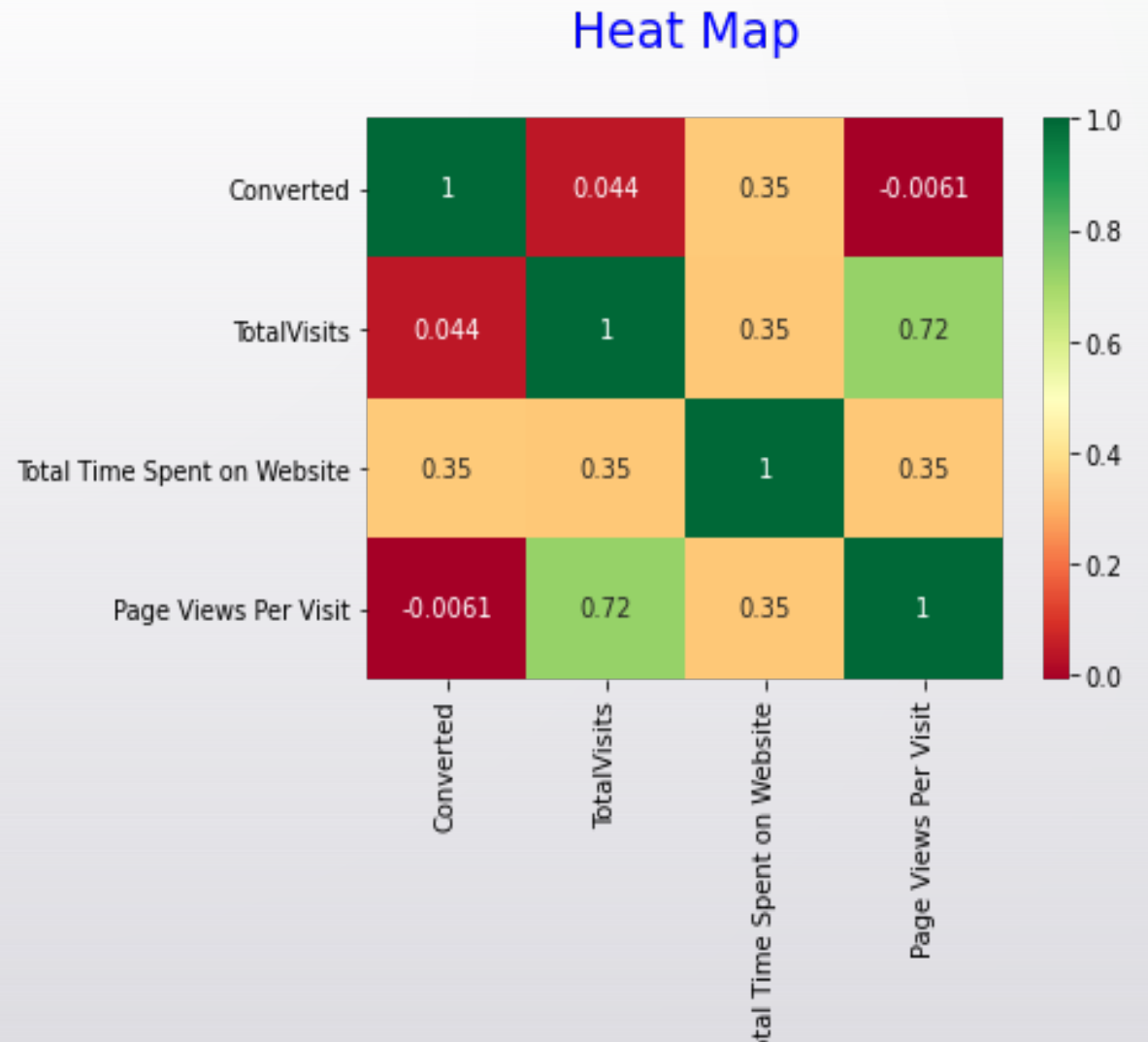
There is no correlation between Total Time Spent on Website And Page Views Per Visit.





## Multivariate Analysis

The target variable **Converted** has positive correlation with **Total Visits** and **Total Time Spent on Website**. **Page Views Per Visit** has a negative correlation with target column.





## Data Preparation

- Converting some binary yes/no to 1/0
- Dummy variable creation for non binary categorical variables
- Splitting the data into Train and Test
- Feature Scaling of the continuous variables for both Train and Test data.



## Model Building

- Logistic Regression Model Building using scikit-learn and Stats Model
- Creating the models with the train set
- Confusion Matrix  
[[3244, 638],  
[ 560, 1825]]
- Metrics

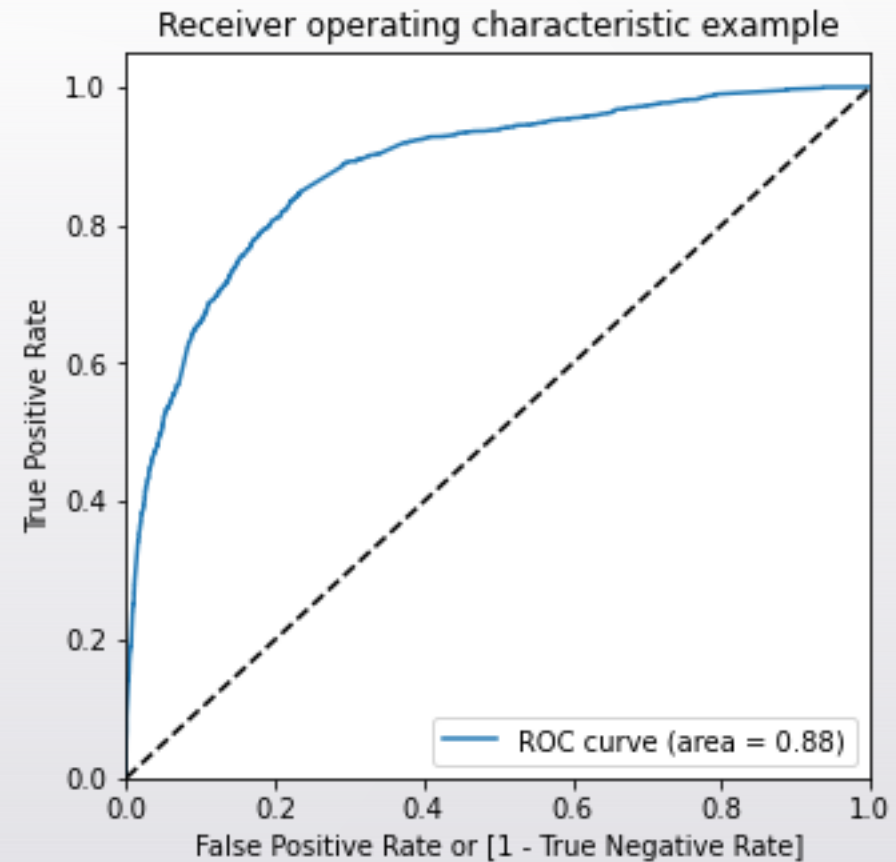
--Accuracy : 80.88%

--Sensitivity : 76.52%

--Specificity : 83.57%

## ROC Curve

The ROC Curve should be a value close to 1.  
We are getting a good value of 0.88 indicating  
a good predictive model.

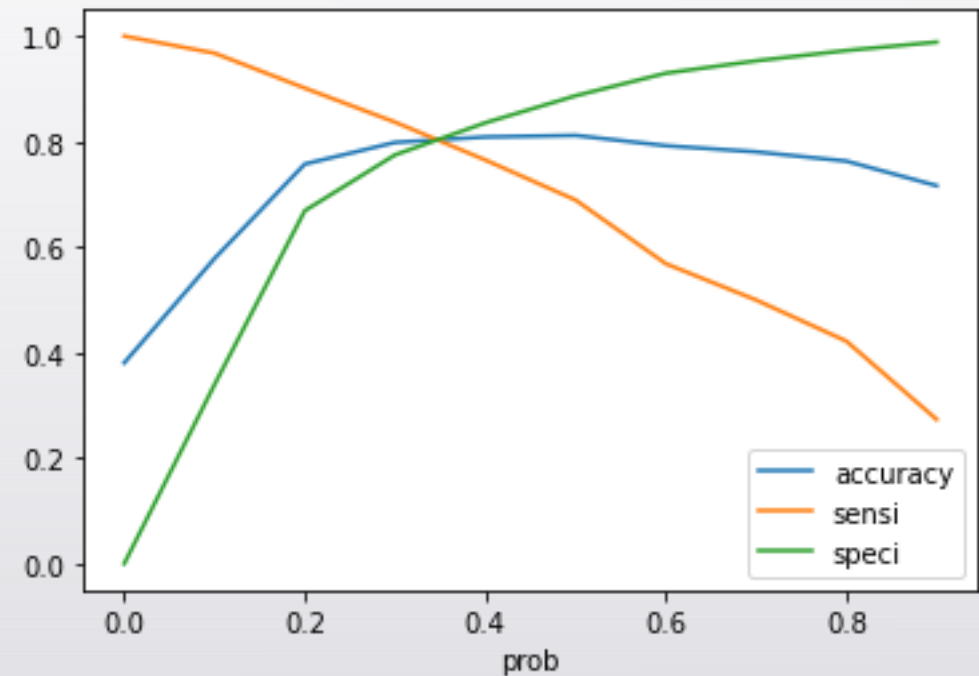




Plot for accuracy sensitivity and specificity for various probabilities.

### Optimal Cutoff Point

From the curve here, 0.4 is the optimum point to take it as a cutoff probability.

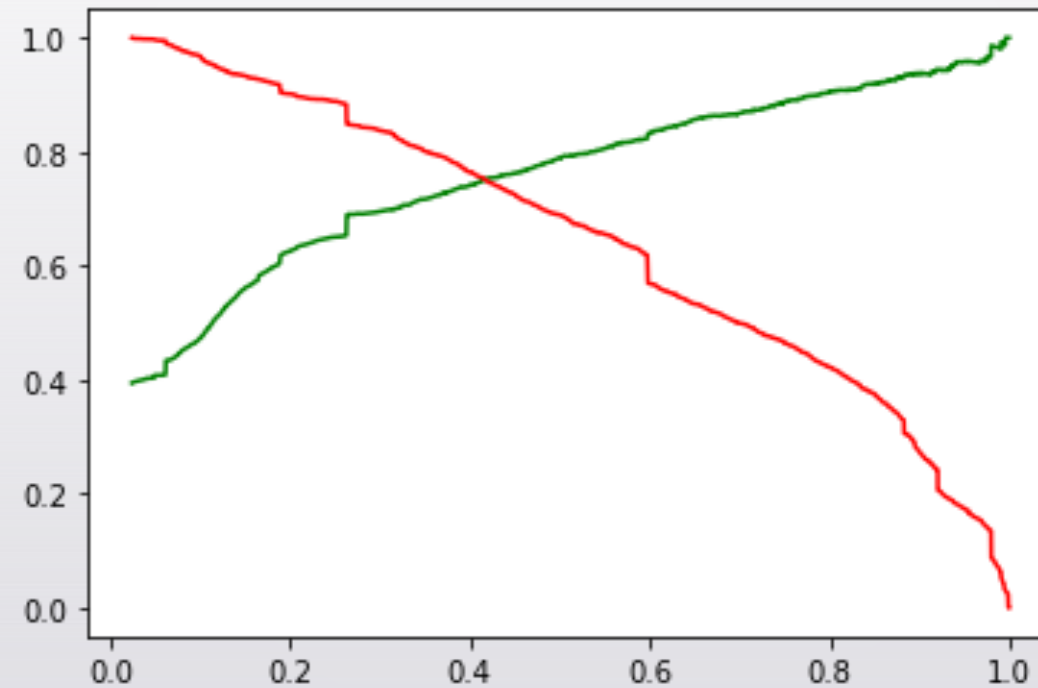






## Precision and Recall Tradeoff

Precision And Recall Tradeoff





## Model Prediction On the Test Set

- Accuracy : 80.67%
- Sensitivity : 77.72%
- Specificity : 82.46%



## Final Observation

Let us compare the values obtained for Train & Test:

### Train Data:

- Accuracy : 80.88%
- Sensitivity : 76.52%
- Specificity : 83.57%

### Test Data:

- Accuracy : 80.67%
- Sensitivity : 77.72%
- Specificity : 82.46%



## Conclusion

The hot lead customer are those having lead score  $>80$ . They can of following categories:

- Leads with add form
- Customer who spent highest time on website
- Working Professional
- Customer Active in Olark Chat Conversation