Lead Scoring – Group Case Study

X-Education

Online Courses

SUBMISSION

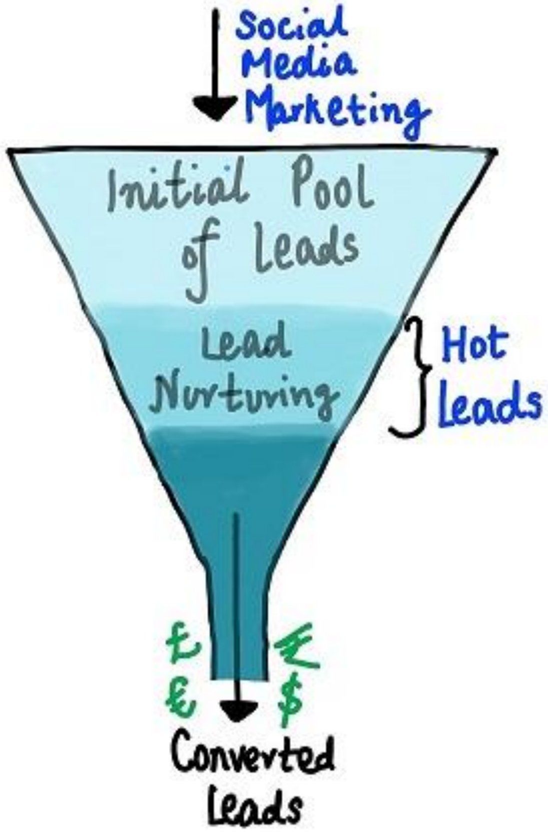
By :

- Anindo Mazumdar
- Mukundan A P

Overview :

X Education sells online courses to industry professionals. The company markets its courses on several websites and search engine. 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.

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.



Overview :

Problem : Lead Conversion

The typical lead conversion rate at X education is around 30%.

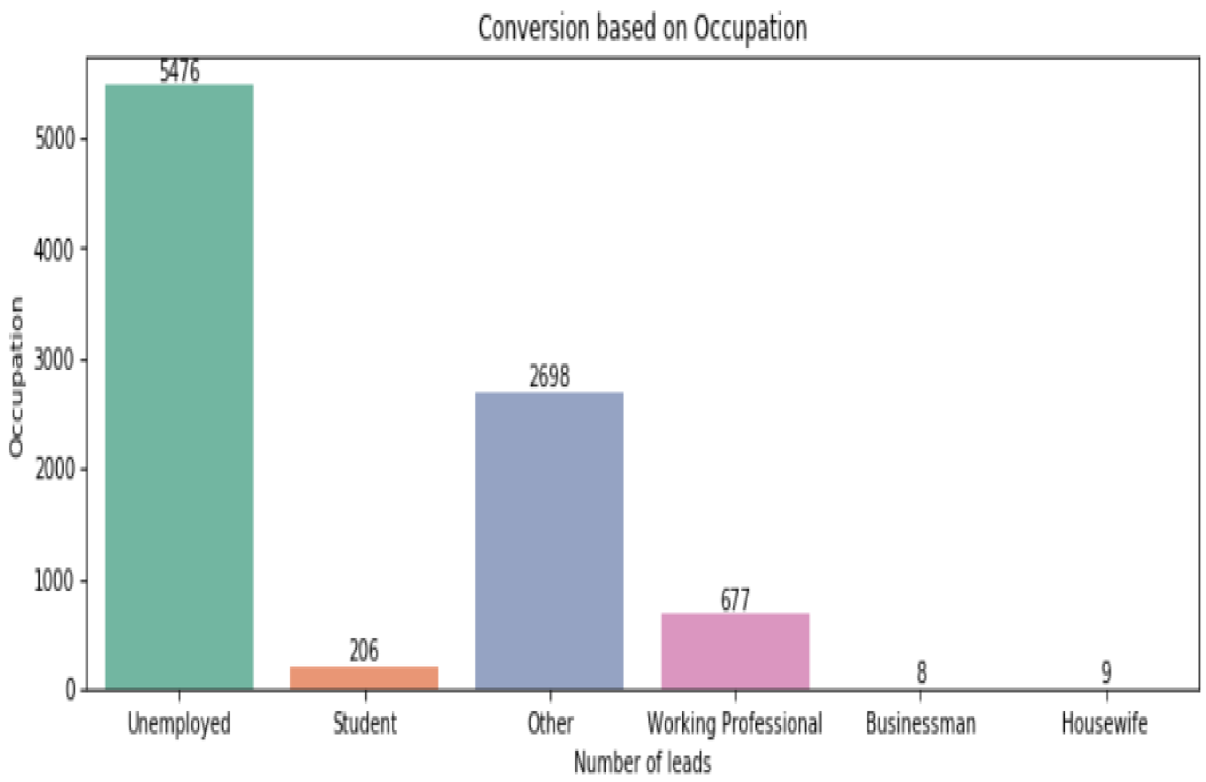
Objectives :

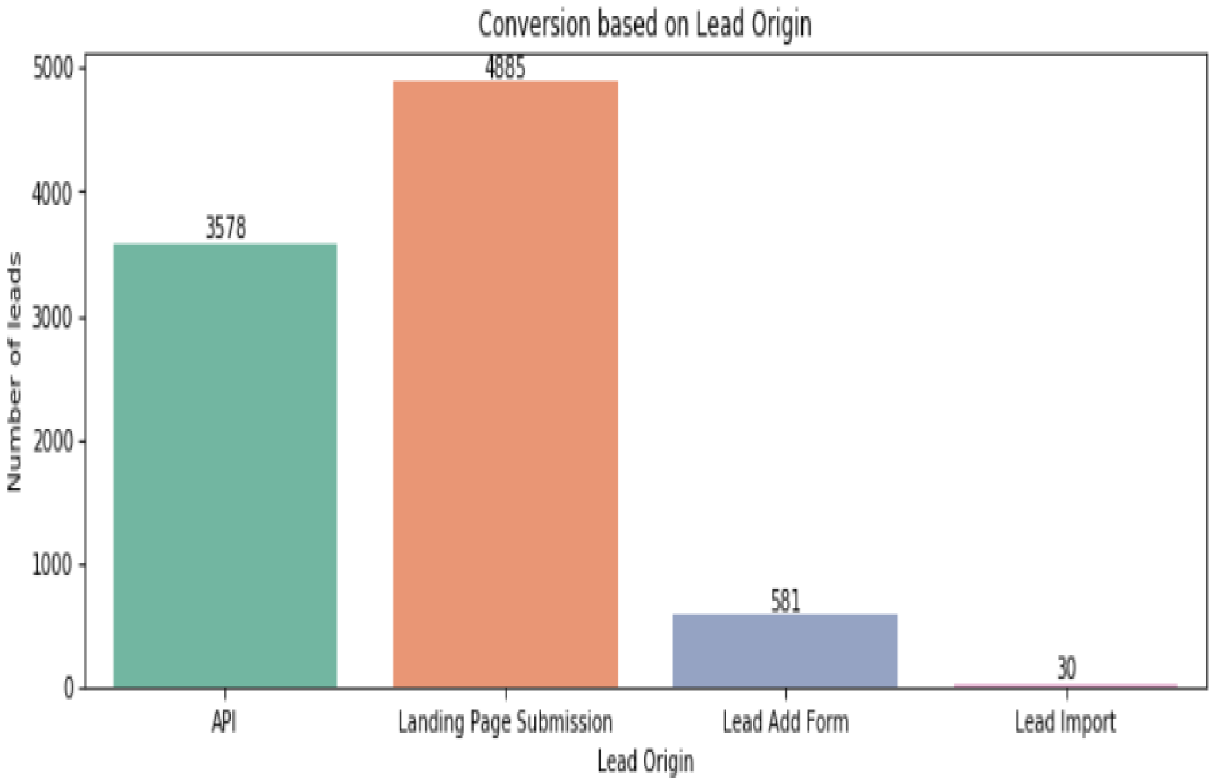
- Build a model to using available lead data, assign a lead score to each of the leads such that the customers with higher lead score have a higher conversion chance and lower score with lower conversion chance.
- Business Recommendation - Target lead conversion rate to be around 80%.
- Additional asks - Provide answers to business questions on lead and its features

Implementation Details

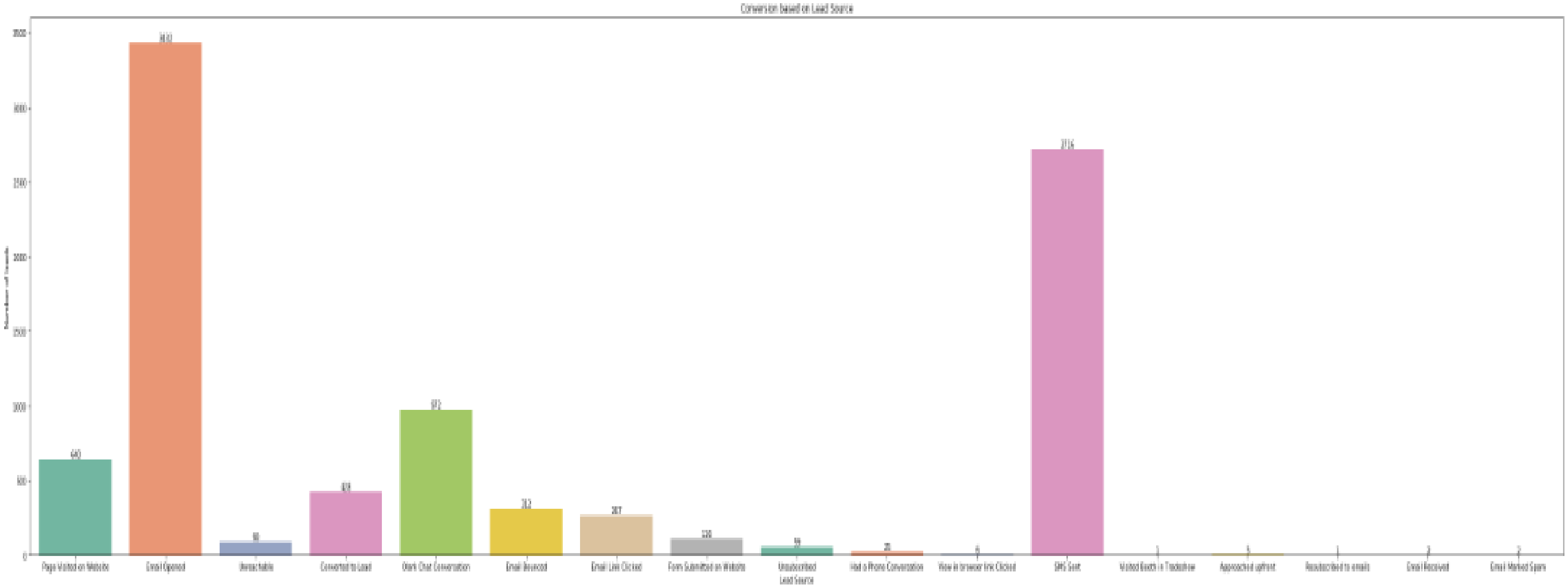
- We started off by looking at base data and identifying relevant/ non relevant columns
- Since the objective of the exercise was to create a model which defines a lead conversion (i.e. positive or negative) , we developed a **“Logistic Regression Model”**
- During this exercise, we were able to ***narrow down 37 predictor variables to 13 predictor variables***, which highlights both negative and positive influencers on the Lead Conversion
- We were able to achieve an ***accuracy of 82% using this model***.
- Some implementation details are as follows –
    - Variable elimination through null value imputation and number of distinct values for a column.
    - Outlier treatment to prevent skewedness in model
    - Creating dummy variables for categorical variable treatment
    - Feature selection using RFE
    - Variable elimination through VIF and statsmodel summary
    - Adjusting cutoff of predicted value to optimize Accuracy, Precision and Recall.

Leads Conversion : Occupation Type & Lead Origin

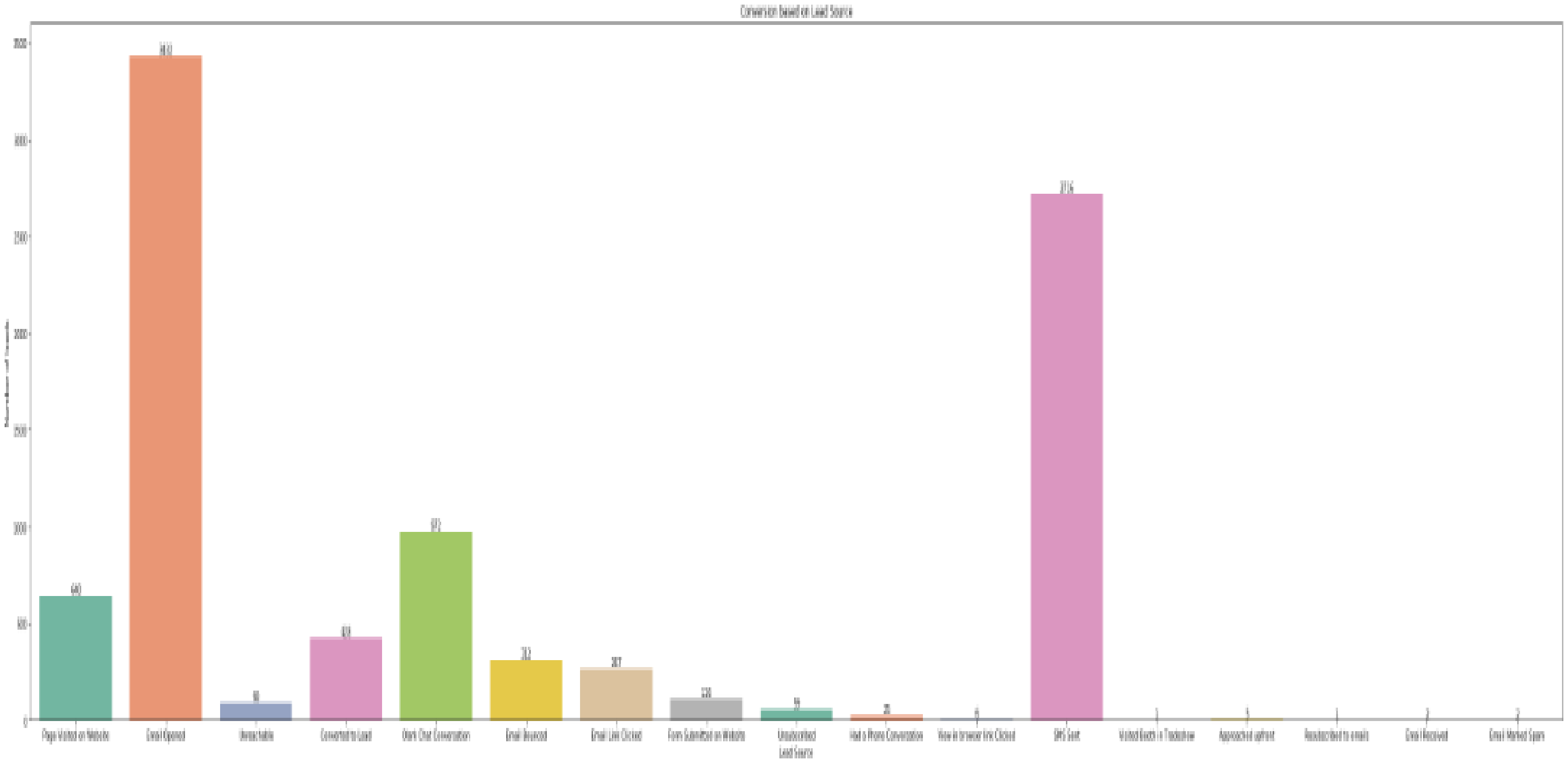




Leads Conversion : Lead Source

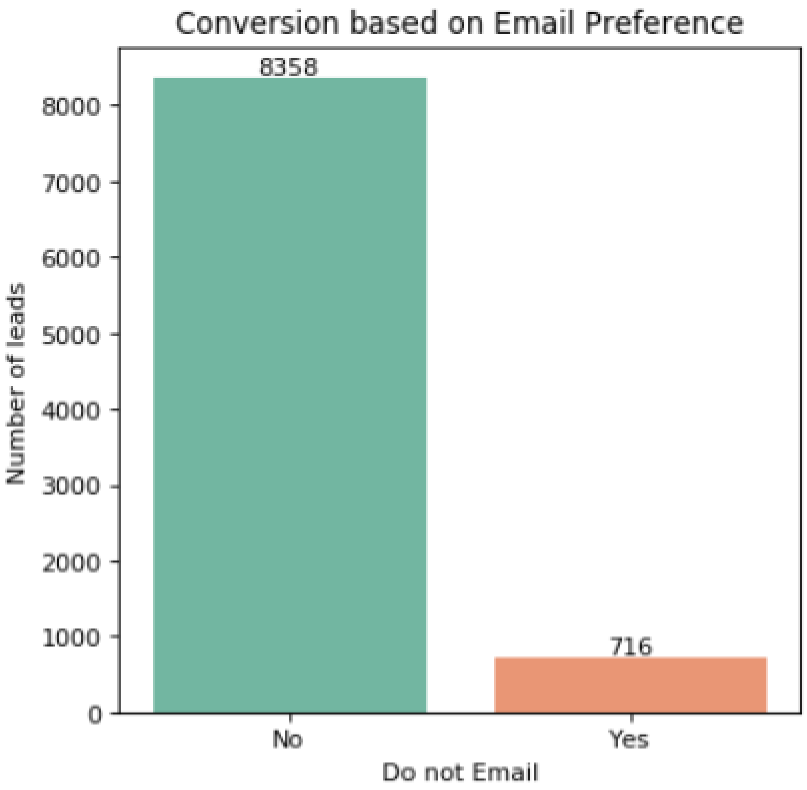


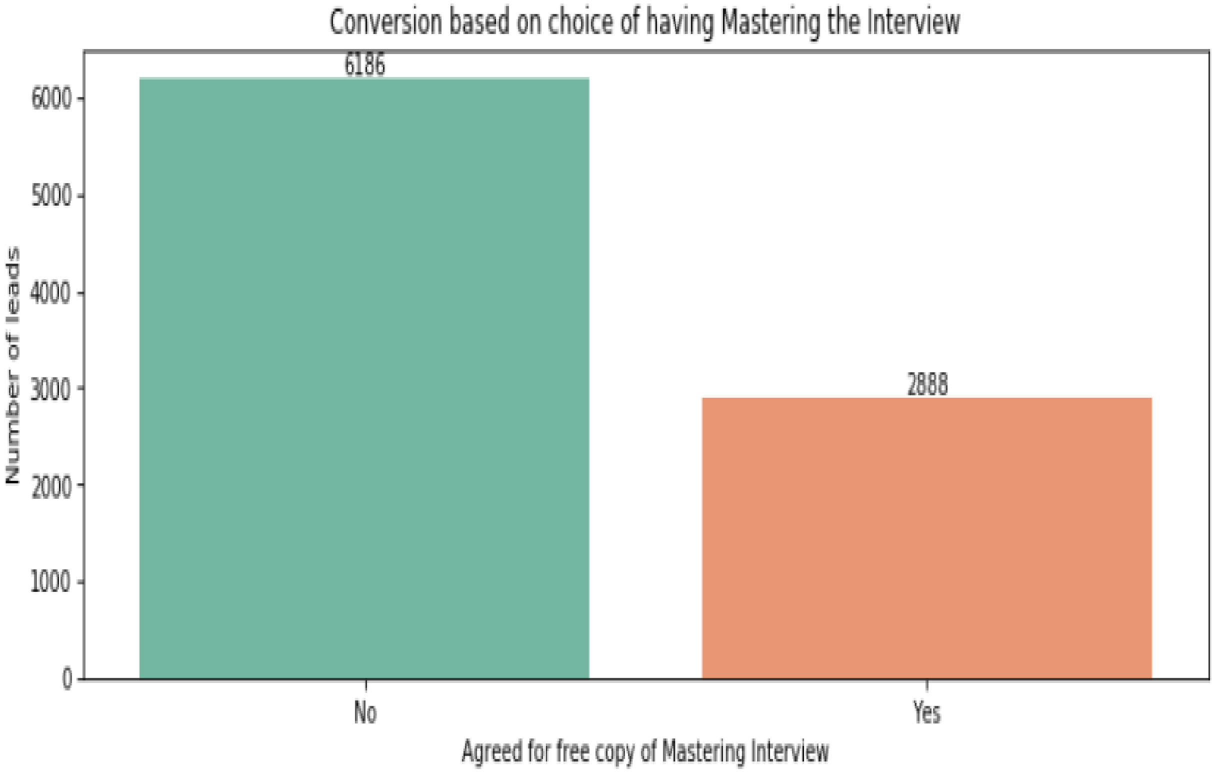
Leads Conversion : Lead Activity



	Last Activity	Converted	Count
21	SMS Sent	1	1705
9	Email Opened	1	1250
18	Page Visited on Website	1	151
16	Olark Chat Conversation	1	84
6	Email Link Clicked	1	73
2	Converted to Lead	1	54
23	Unreachable	1	29
12	Form Submitted on Website	1	28
14	Had a Phone Conversation	1	20
4	Email Bounced	1	16

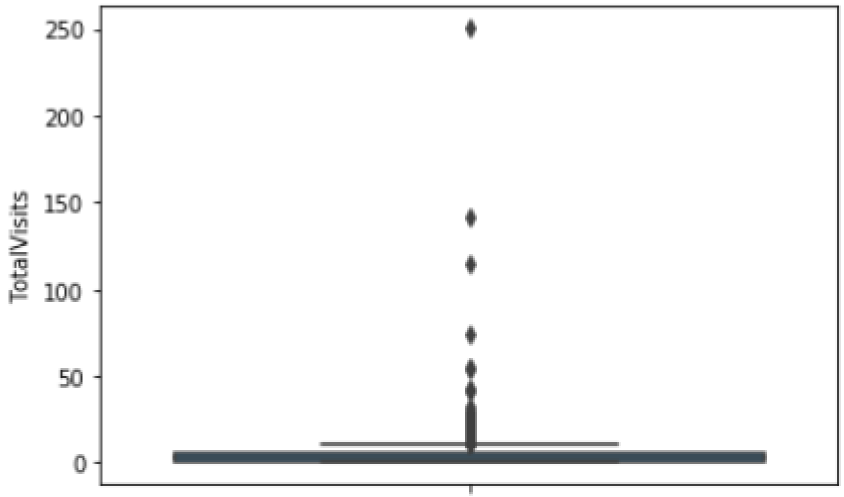
Leads Conversion : Email Preference & Mastering the Interview



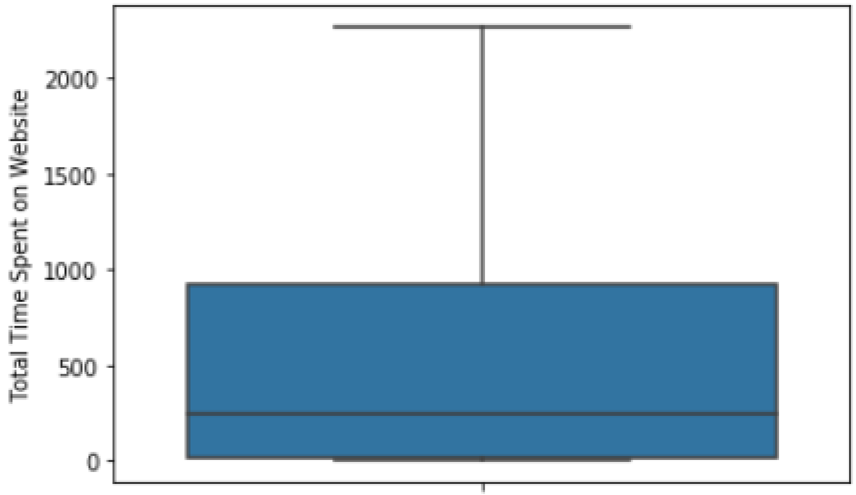


Outliers :

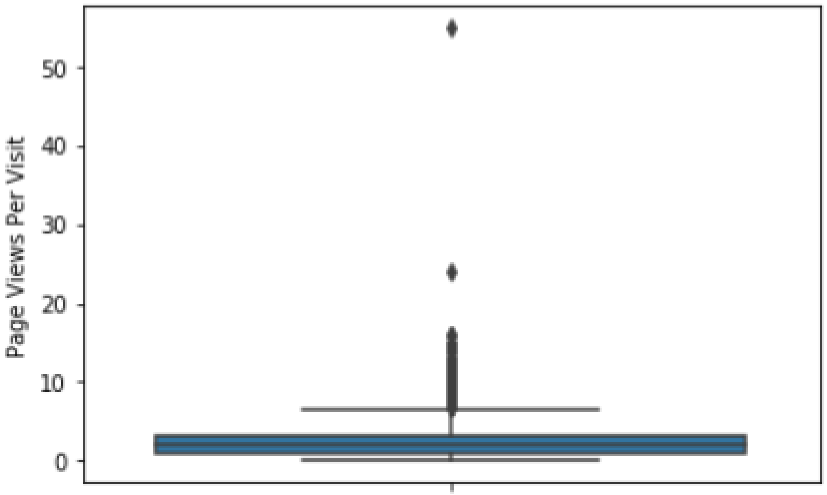
TOTAL VIEWS



TOTAL TIME SPENT ON THE WEBSITE

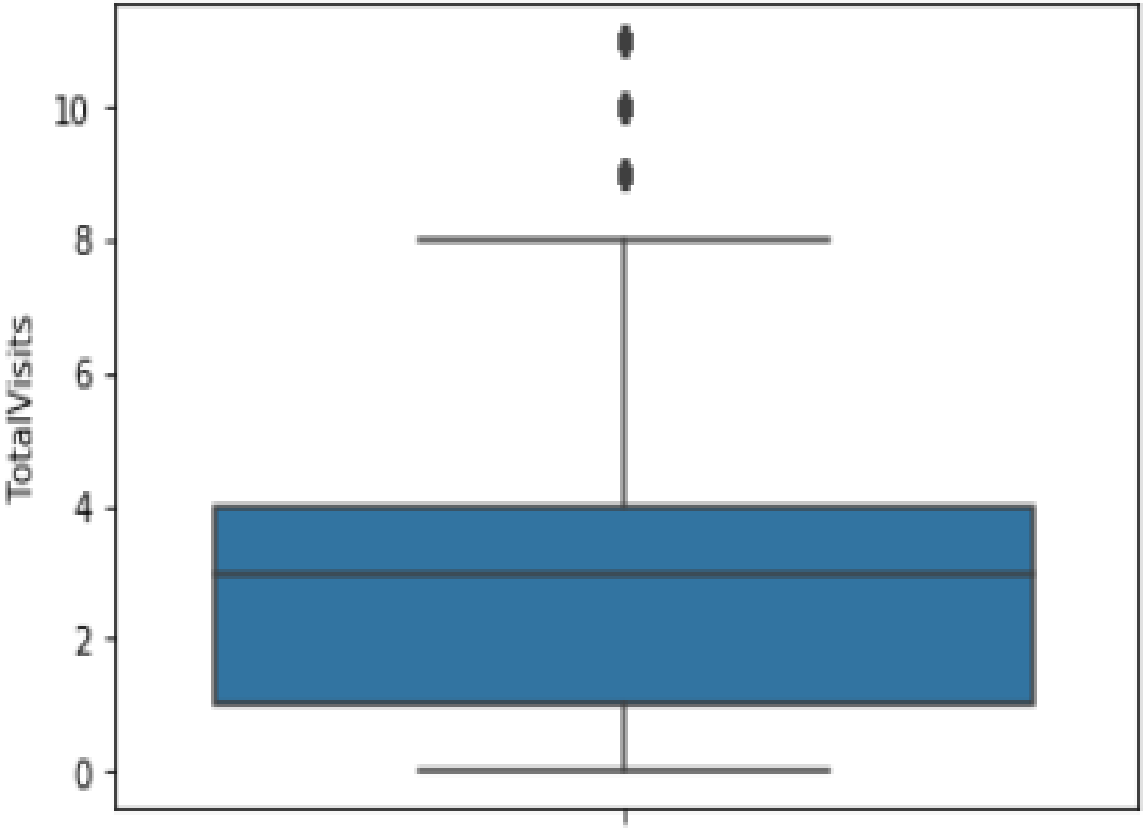


PAGE VIEWS PER VISIT

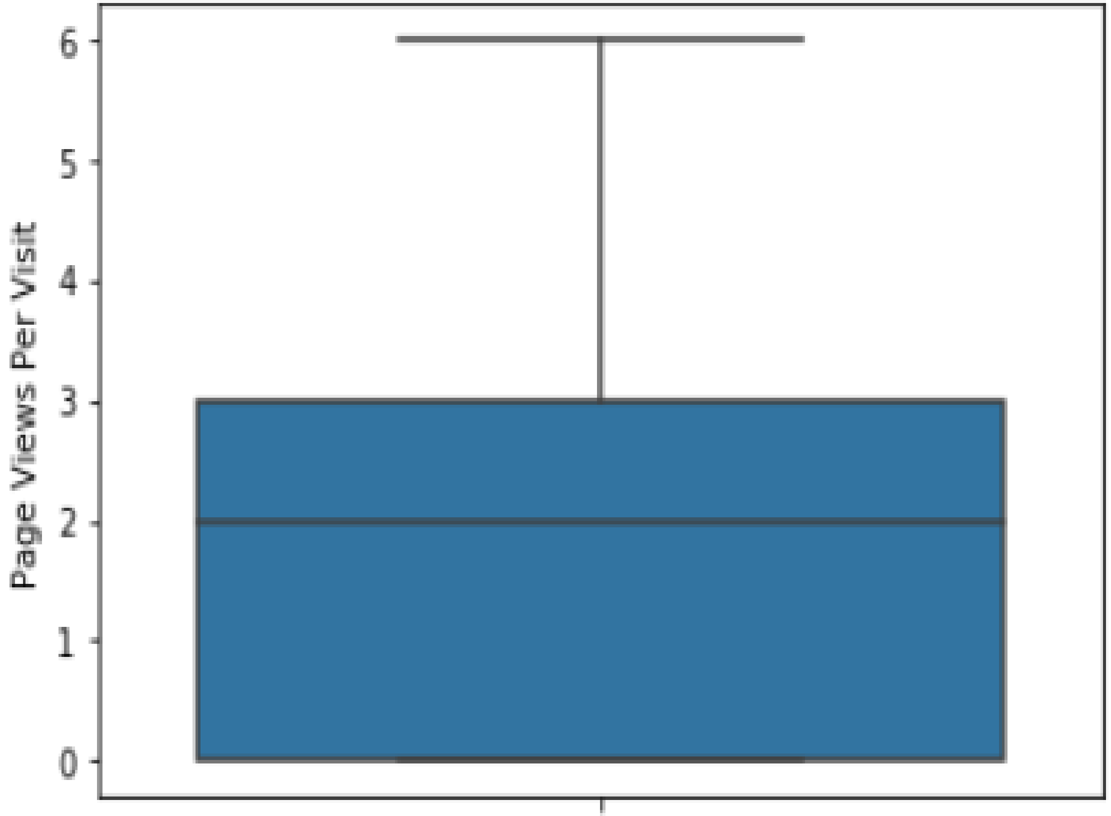


Outlier Treatment :

TOTAL VIEWS



PAGE VIEWS PER VISIT

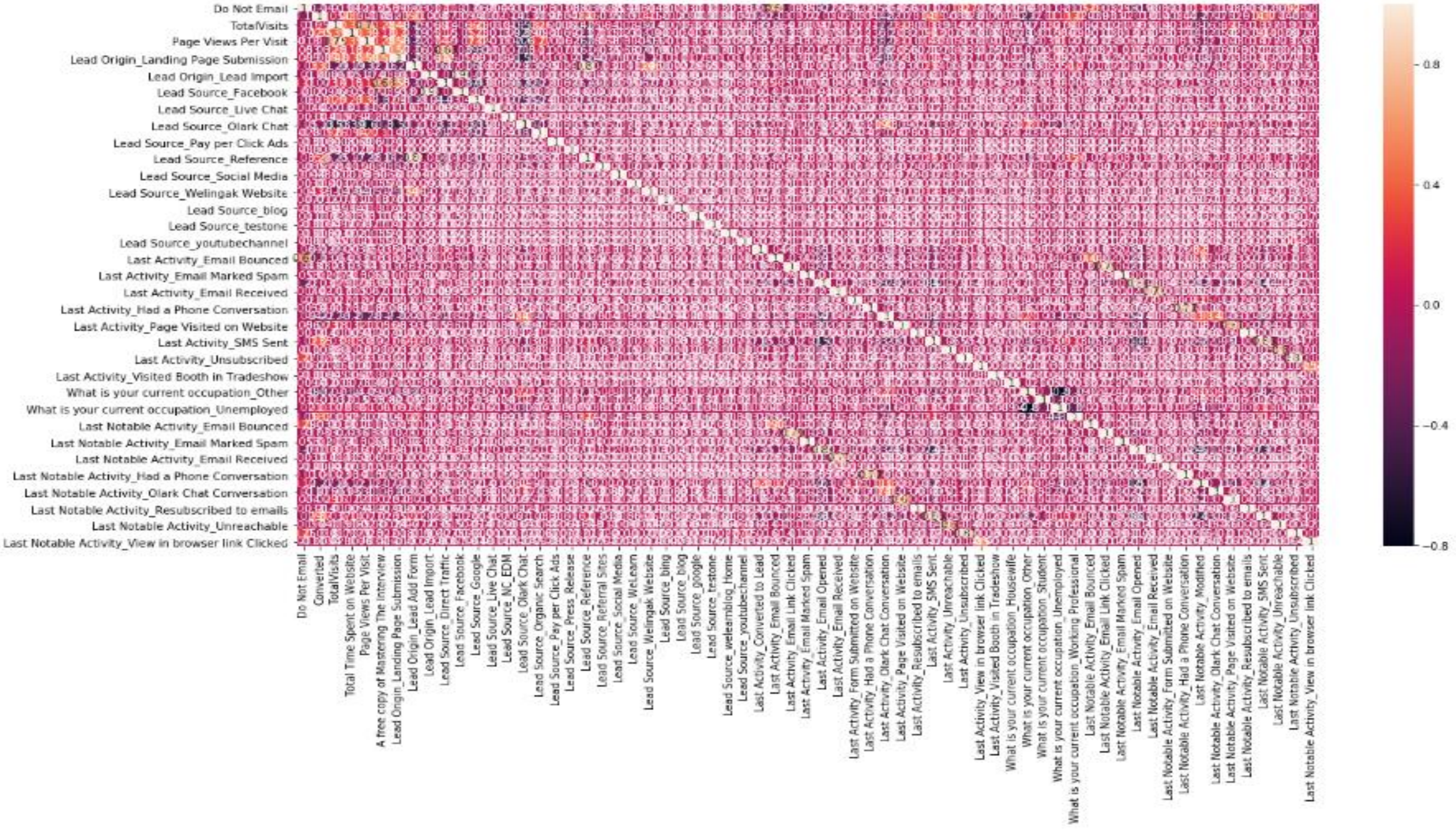


Data Frame after Treating Outliers :

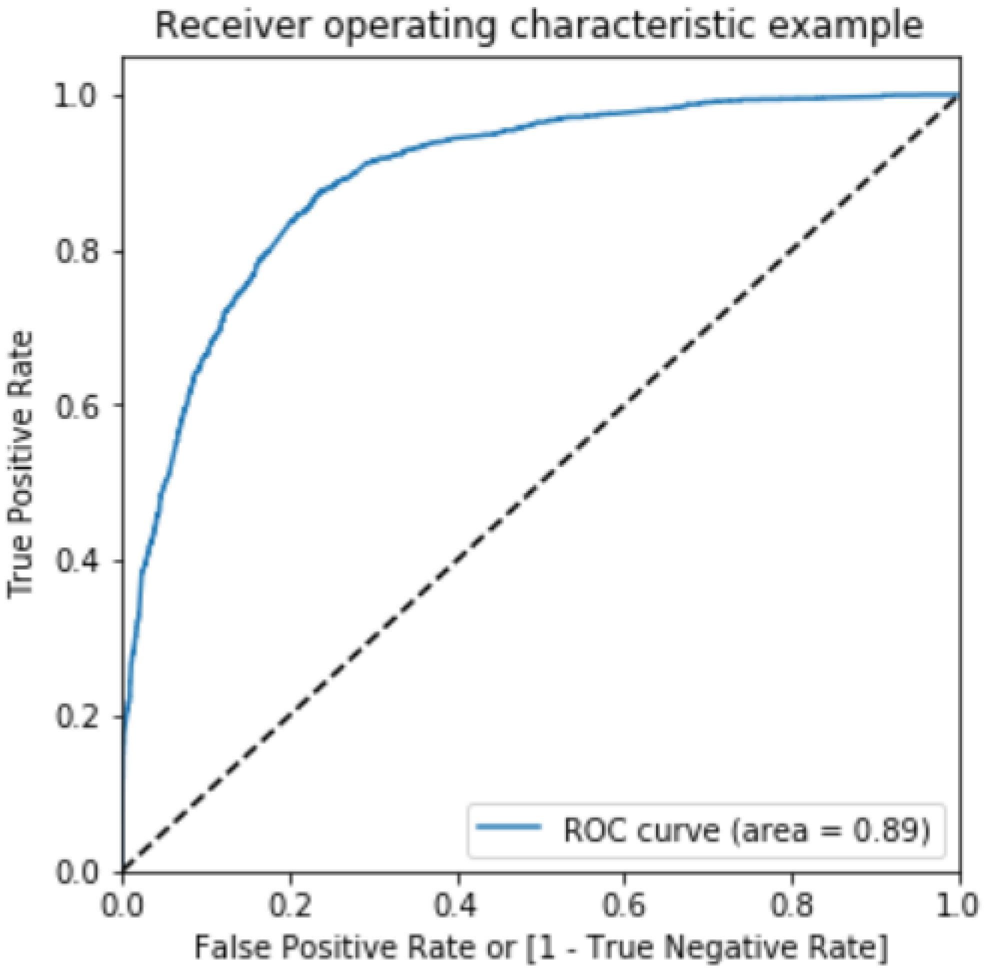
	Lead Origin	Lead Source	Do Not Email	Converted	Total Visits	Total Time Spent on Website	Page Views Per Visit	Last Activity	What is your current occupation	A free copy of Mastering The Interview	Last Notable Activity
0	API	Olark Chat	0	0	0	0	0	Page Visited on Website	Unemployed	0	Modified
1	API	Organic Search	0	0	5	674	2.5	Email Opened	Unemployed	0	Email Opened
2	Landing Page Submission	Direct Traffic	0	1	2	1532	2	Email Opened	Student	1	Email Opened
3	Landing Page Submission	Direct Traffic	0	0	1	305	1	Unreachable	Unemployed	0	Modified
4	Landing Page Submission	Google	0	1	2	1428	1	Converted to Lead	Unemployed	0	Modified

- *Binary mapping is done for the columns: **'Do Not Email‘** and **'A free copy of Mastering The*** **Interview‘.**
- *Dummy variables are created for the rest of the categorical variables: **'Lead Origin'**, **'Lead Source'**, **'Last Activity'**, **'What is your current occupation‘** and **'Last Notable Activity‘**.*
- *The final shape of the dataset comes to **(8513,65).***

Correlation Matrix :

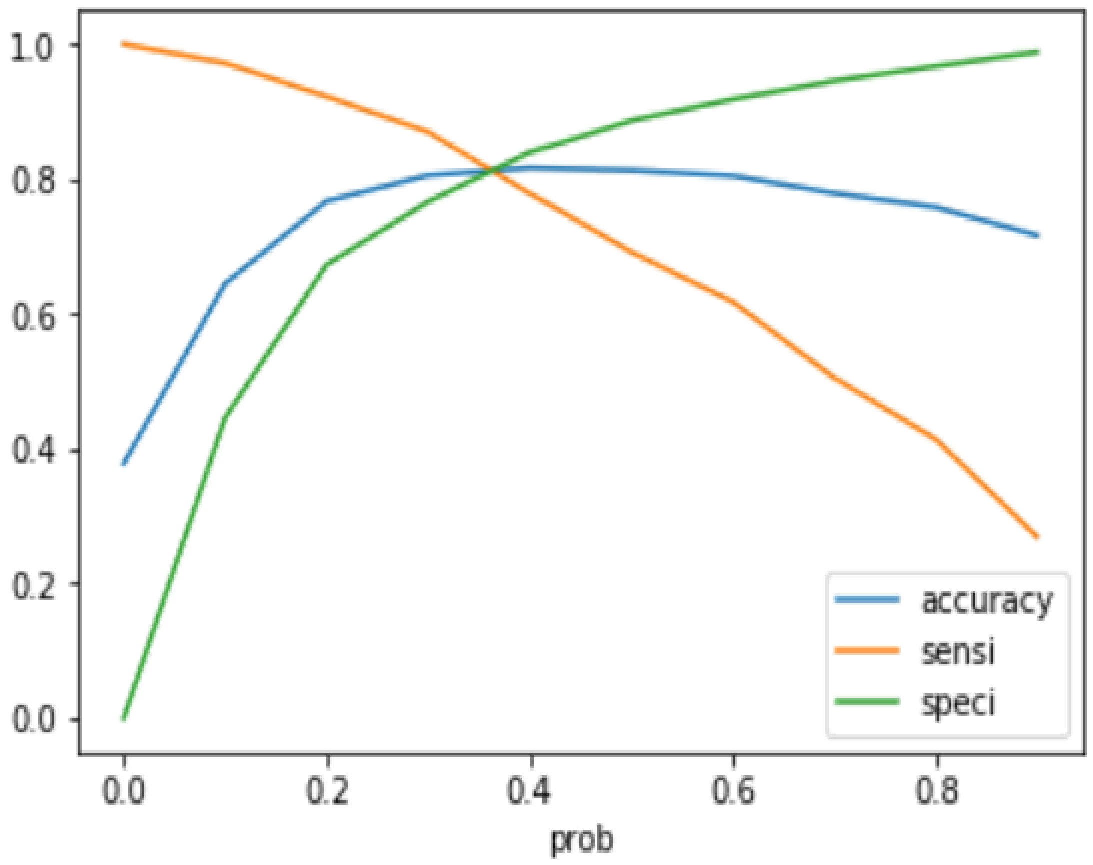


ROC Curve :



Receiver operating characteristic example
True Positive Rate
0.0
0.2
0.4
0.6
0.8
1.0
False Positive Rate or [1 - True Negative Rate]
0.0
0.2
0.4
0.6
0.8
1.0
ROC curve (area = 0.89)

ROC Curve :



prob	accuracy	sensi	speci
0.0	0.378084	1.000000	0.000000
0.1	0.644739	0.972037	0.445764
0.2	0.767075	0.922326	0.672693
0.3	0.805336	0.869951	0.766055
0.4	0.816412	0.778961	0.839180
0.5	0.812888	0.691522	0.886670
0.6	0.804665	0.617843	0.918241
0.7	0.778990	0.505104	0.945494
0.8	0.758181	0.414115	0.967350
0.9	0.716395	0.269419	0.988127

Observations :

- Variables “*Lead Origin*”, “*Last Activity(telephonic conversation)*” and “*Current Occupation(working)”* we identified as 3 most positive critical variables for lead conversion
- Variables – “*Do Not Email*” and “*Current Occupation (Others)”* were identified as negative correlating factors in the model
- Based on above observation we suggest –
- **FOCUS ON** – Working professionals , who are calling in / requesting a call back for further discussions.
- **LOOKOUT FOR** – Instances where people aren’t willing to give their personal details (i.e. do not email) or are using online chat services to get more details around the program rather than dialing in to call or requesting for call back.
- Suggest to have regular look at the accuracy of the data model to ensure high levels of predictive power/ relevance

Recommendations :

- With 80% conversion rate, the model has been built with the conversion probability as 0.52 and taking Lead Origin , Last Activity and Current Occupation of the lead as the most important factors behind the conversion.

- If X Education focusses most on these factors, they will be able to increase their hot leads count as also shown in the EDA they would will be able to meet their 80% conversion rate successfully

- By marketing more in the most trending Lead Origin (ie Google), or communicating more in the SMS mode amongst Unemployed people will help their company grow more.

- The approach should be made more to the unemployed people with exciting deals so that the course gets it worth.

- The advertisement should be made in a tempting way in Google more so that people feel more likely to click on it and register on the landing page.

- The exciting deals can be communicated via SMS, so that they can go through it in their leisure time. By calls,
- mostly people feel reluctant to pay any attention and can go out of opportunity.