Lead Scoring Case Study

By
Priti Satpute
&
Deepa B.

Introduction

- We are going to solve the business problem “X Education Company” using Logistic Regration.

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.
- 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 higher lead score have a higher conversion chance and the customers with lower lead score have a lower conversion chance. The CEO has given a ballpark of the target lead conversion rate to be around 80%.

Data

- We have been provided with a lead's dataset from the past with around 9000 data points. This dataset consists of various attributes such as Lead Source, Total Time Spent on Website, Total Visits, Last Activity, etc. which may or may not be useful in ultimately deciding whether a lead will be converted or not. The target variable, in this case, is the column ‘Converted’ which tells whether a past lead was converted or not wherein 1 means it was converted and 0 means it wasn’t converted.

Data Cleaning & Preparation

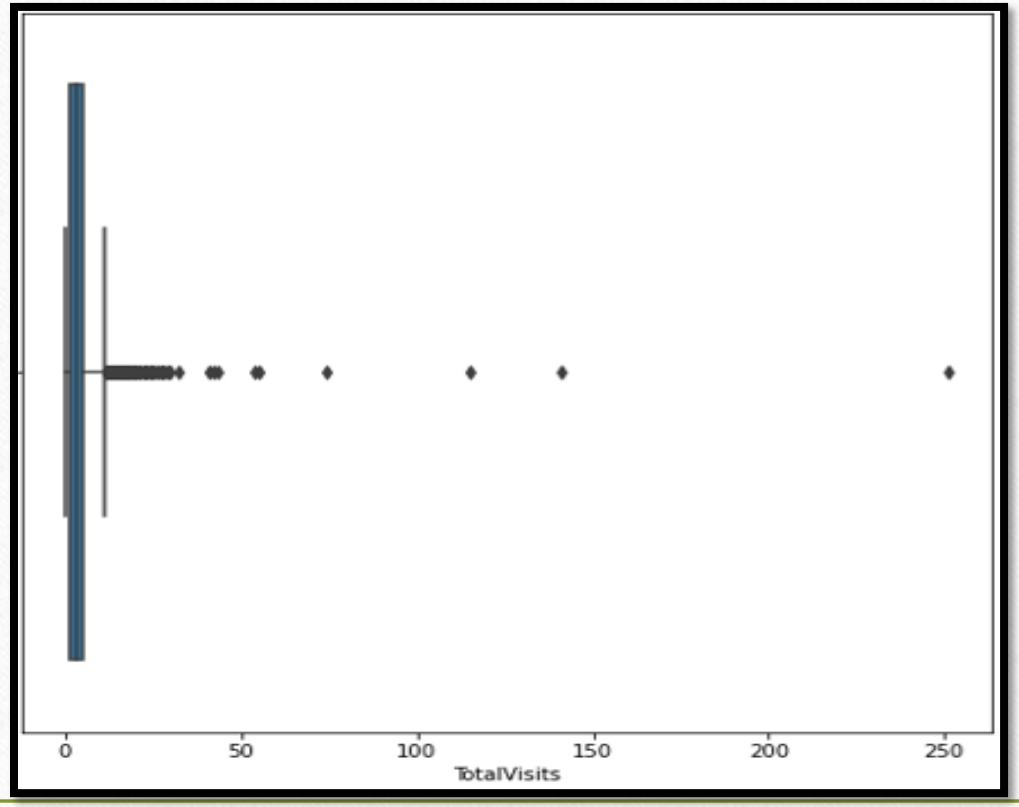
- We dropped columns having higher null values like

  Lead quality, Asymmetrique Activity Index, Asymmetrique Profile Index,

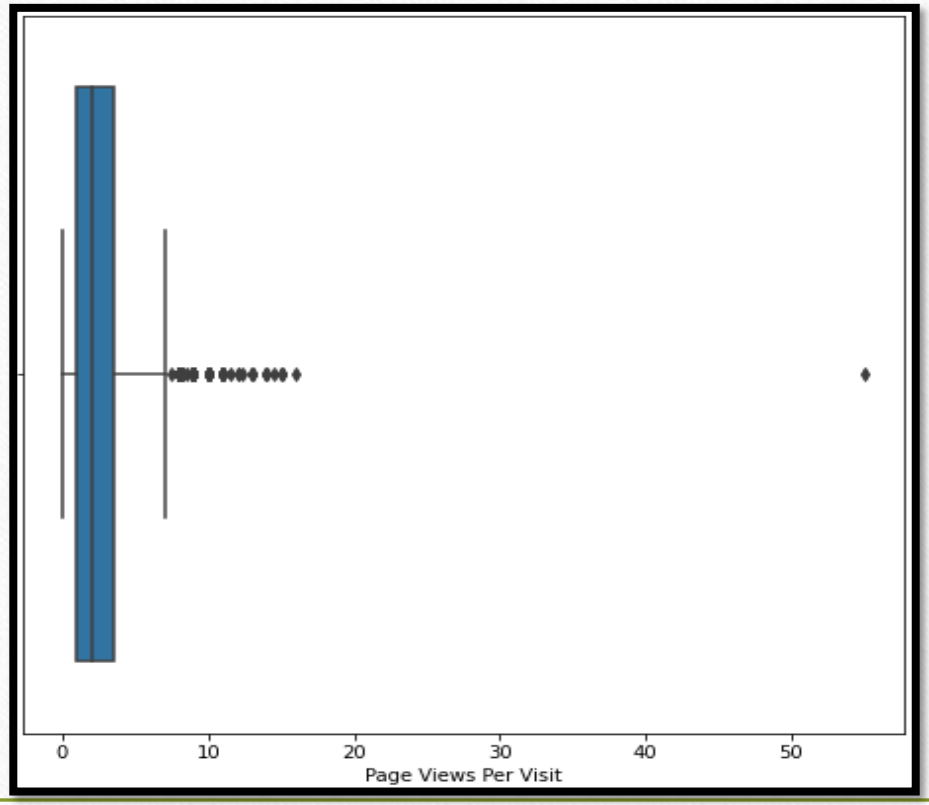
  Asymmetrique Activity Score, Asymmetrique Profile Score, etc.
- Some columns like “Magazine” have high data imbalance hence we dropped that column
- Column like “Last Activity” have less frequent category hence we grouped them together one category shown as “Others”

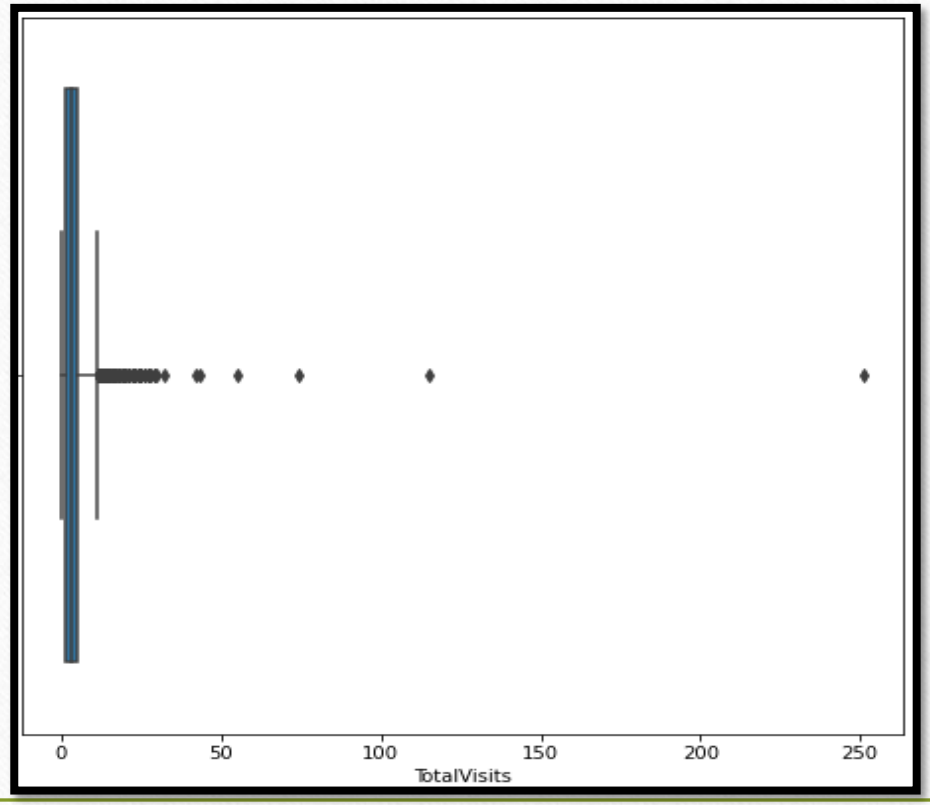
Data Cleaning & Preparation

- Column total visits have many outliers, we impute the missing values using Median.



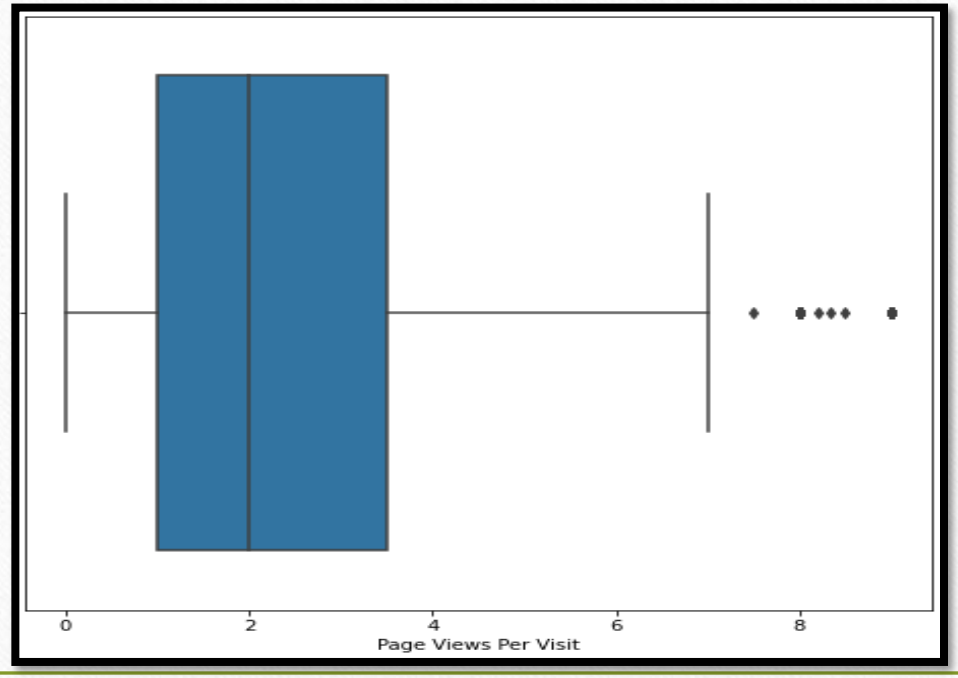
TREATING THE OUTLIERS

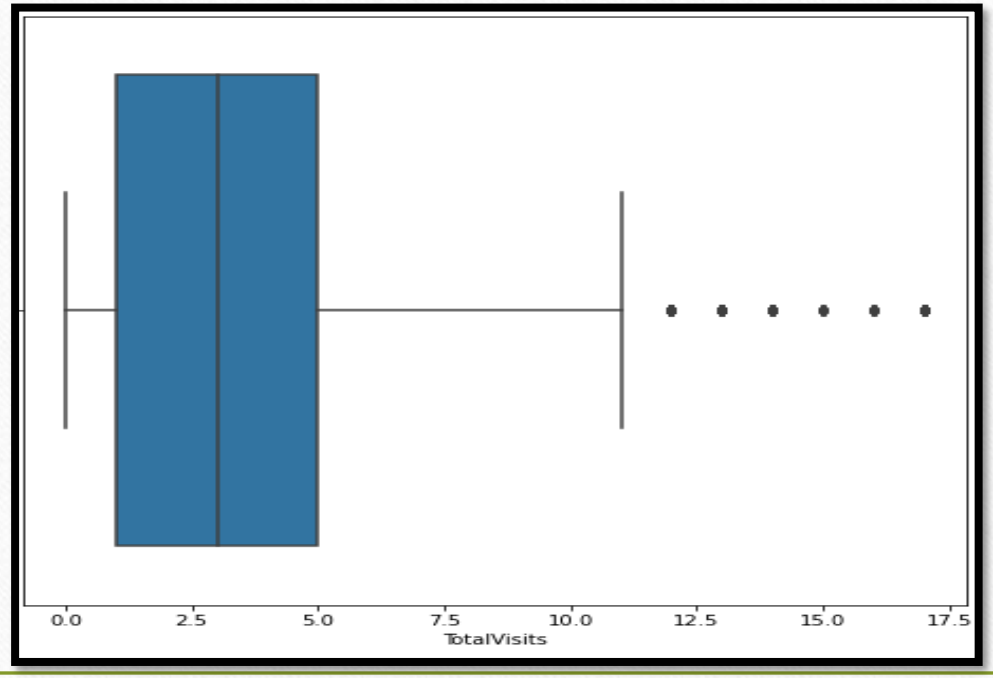




TREATING THE OUTLIERS

- From the above graph we can see that there are outliers with extremely high values which can affect our model. Hence, we treat them by capping them at 99 percentile

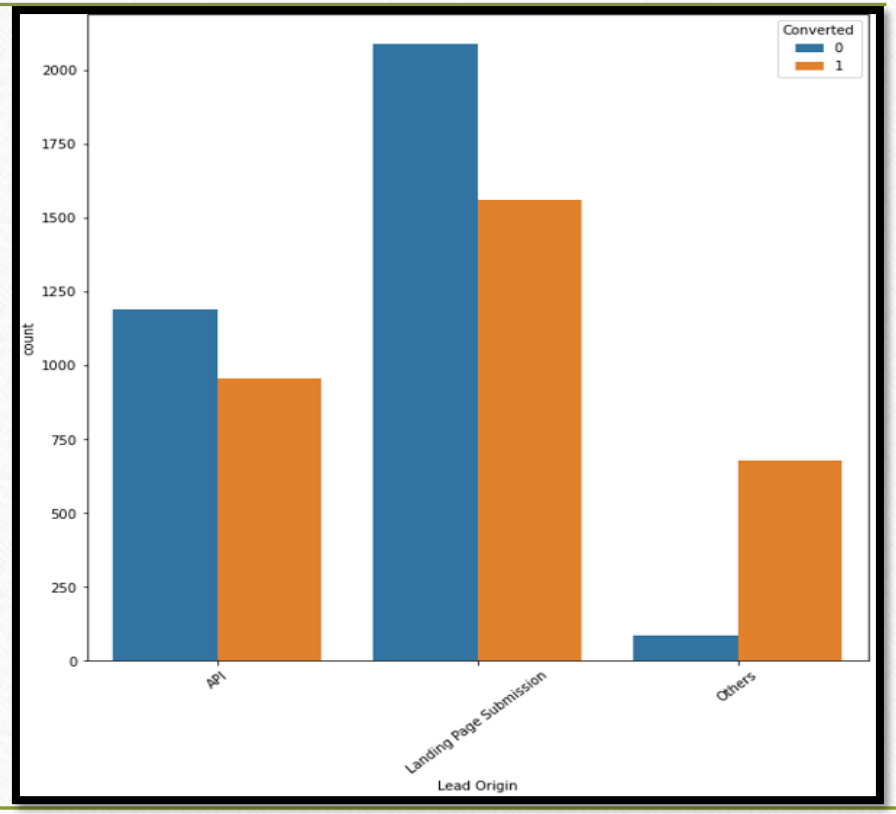




Exploratory Data Analysis

- Lead Origin:

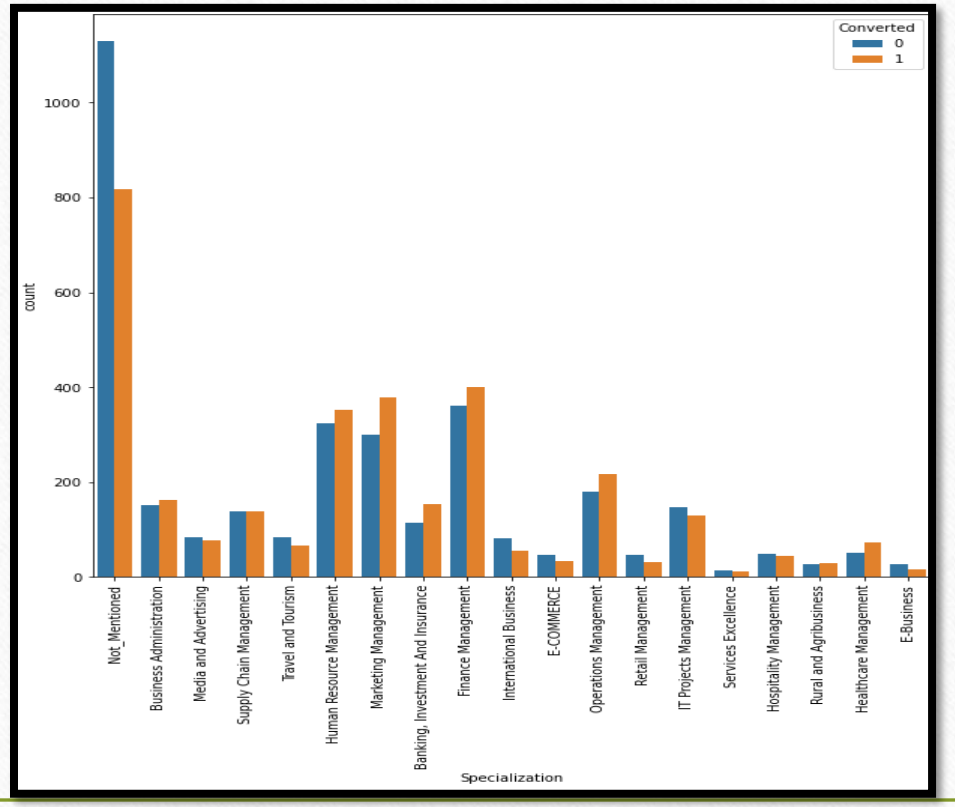
  From the graph we noticed that Landing Test Submission and API have higher conversion rate therefore we should improve in these areas.



Exploratory Data Analysis

- Specialization:

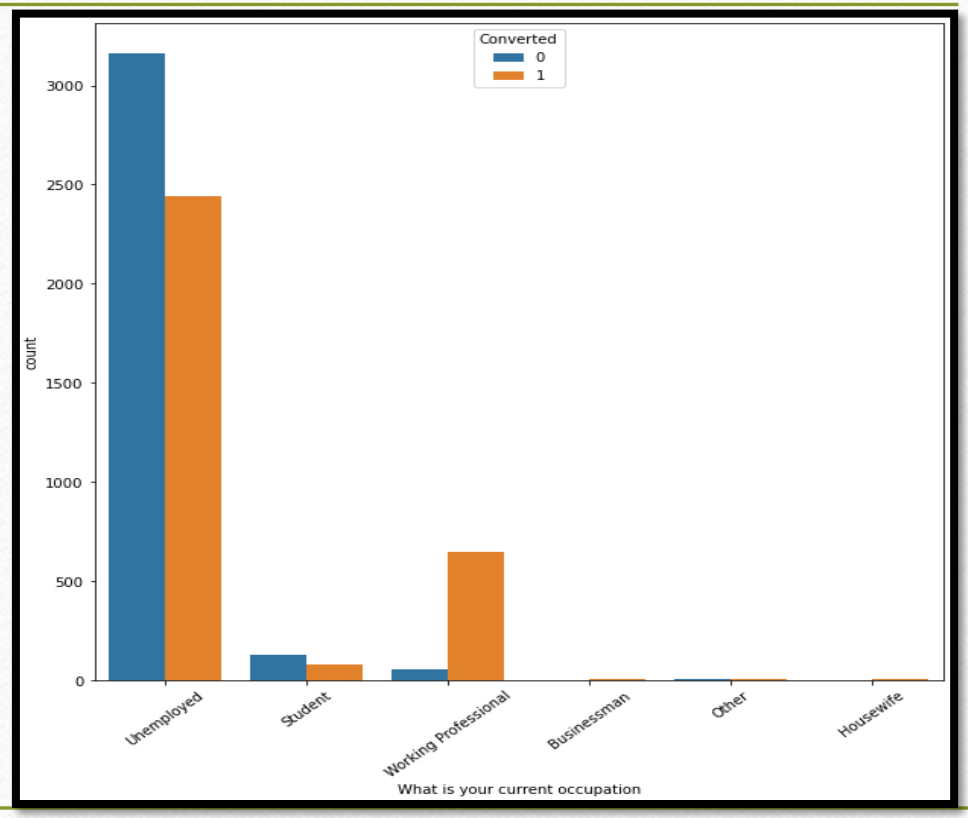
  From the above graph, we can see that people who have not mentioned any specialization are more likely to get converted.



Exploratory Data Analysis

- What is your current occupation

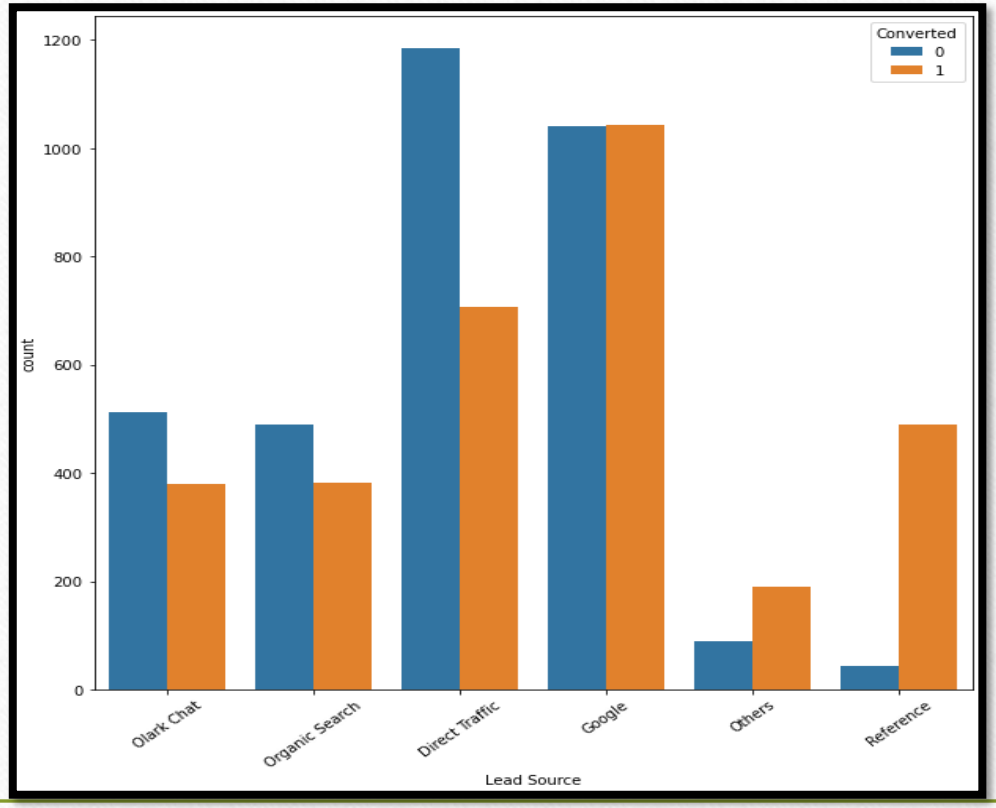
  From the above graph we can see that unemployed people are more likely to get converted into paying customers, but these are the same people who are more not likely to get converted



Exploratory Data Analysis

- Lead source

  From the above graph ,we can see that people who either searched for the website or recommended to them through browser bookmarks have the higher conversion rate



Creating Dummy Variables

- We will create dummy variables for

1. Lead Origin
2. Lead Source
3. Do Not E-mail
4. Last Activity
5. What is your current occupation
6. Specialization
7. Tags
8. Last Notable Activity

Splitting the Data and Model Building

- We split the data in Training and Testing in the ratio of 70:30
- Will build Logistic Regration Model

	coef	std err	z	P>|z|	[0.025	0.975]
const	-4.3091	0.300	-14.365	0.000	-4.897	-3.721
TotalVisits	2.1758	0.557	3.908	0.000	1.085	3.267
Total Time Spent on Website	4.0965	0.353	11.597	0.000	3.404	4.789
Page Views Per Visit	-1.6486	0.538	-3.067	0.002	-2.702	-0.595
Lead Origin_Others	1.6885	0.311	5.428	0.000	1.079	2.298
Lead Source_Olark Chat	0.9773	0.257	3.805	0.000	0.474	1.481
Do Not Email_Yes	-0.8198	0.364	-2.249	0.025	-1.534	-0.105
Last Activity_Email Bounced	-0.9336	0.572	-1.631	0.103	-2.056	0.188
Last Activity_SMS Sent	1.7286	0.181	9.555	0.000	1.374	2.083
Specialization_Travel and Tourism	-0.9404	0.553	-1.699	0.089	-2.025	0.144
Tags_Busy	2.8864	0.292	9.899	0.000	2.315	3.458
Tags_Closed by Horizzon	8.4440	0.762	11.086	0.000	6.951	9.937
Tags_Interested in full time MBA	-20.5372	1.29e+04	-0.002	0.999	-2.54e+04	2.53e+04
Tags_Lost to EINS	7.4668	0.646	11.558	0.000	6.201	8.733
Tags_Not doing further education	-0.6886	1.048	-0.657	0.511	-2.742	1.365
Tags_Ringing	-1.2710	0.307	-4.145	0.000	-1.872	-0.670
Tags_Unknown	3.8843	0.243	15.968	0.000	3.408	4.361
Tags_Will revert after reading the email	6.4085	0.269	23.782	0.000	5.880	6.937
Tags_switched off	-1.7983	0.639	-2.815	0.005	-3.051	-0.546

- As we observe the model has high "p value" (Greater than 0.05) and high VIF (Greater than 5)

	Features	VIF
2	Page Views Per Visit	5.32
0	TotalVisits	4.21
16	Tags_Will revert after reading the email	2.83
1	Total Time Spent on Website	2.59
7	Last Activity_SMS Sent	1.82
3	Lead Origin_Others	1.78
5	Do Not Email_Yes	1.69
6	Last Activity_Email Bounced	1.66
14	Tags_Ringing	1.63
18	Last Notable Activity_Modified	1.56
15	Tags_Unknown	1.55
10	Tags_Closed by Horizzon	1.55
4	Lead Source_Olark Chat	1.43
9	Tags_Busy	1.17
17	Tags_switched off	1.14
13	Tags_Not doing further education	1.12
12	Tags_Lost to EINS	1.11
11	Tags_Interested in full time MBA	1.06
8	Specialization_Travel and Tourism	1.04
19	Last Notable Activity_Olark Chat Conversation	1.04

Splitting the Data and Model Building

- As “page views” per visit high VIF value we dropped the column and again run the module
- We run the module till we get all “p values” less than 0.05 and VIF values less than 5

Model Evaluation

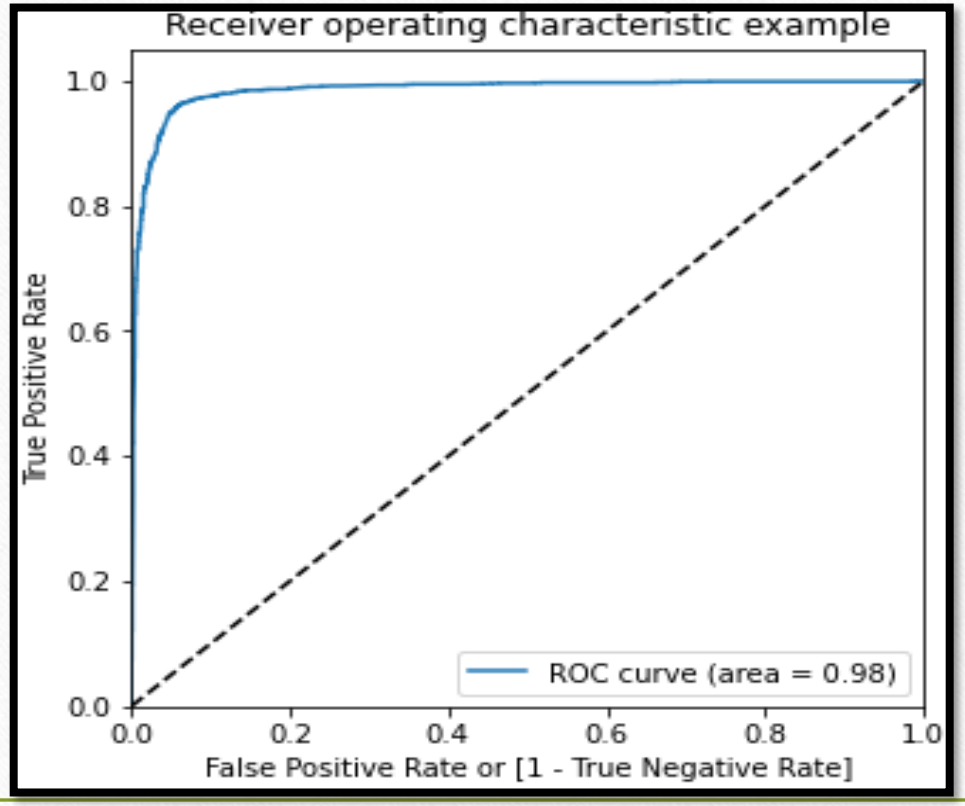
- We predict the probability on train set and find the conversion probability
- By creating confusion matrix will received

1. Accuracy 95%
2. Sensitivity 95%
3. Specificity 94%

Model Evaluation

- Plotting the ROC curve to find optimal cut off:

  Optimal cut off probability is that probability where we get balanced sensitivity & specificity.



Model Evaluation

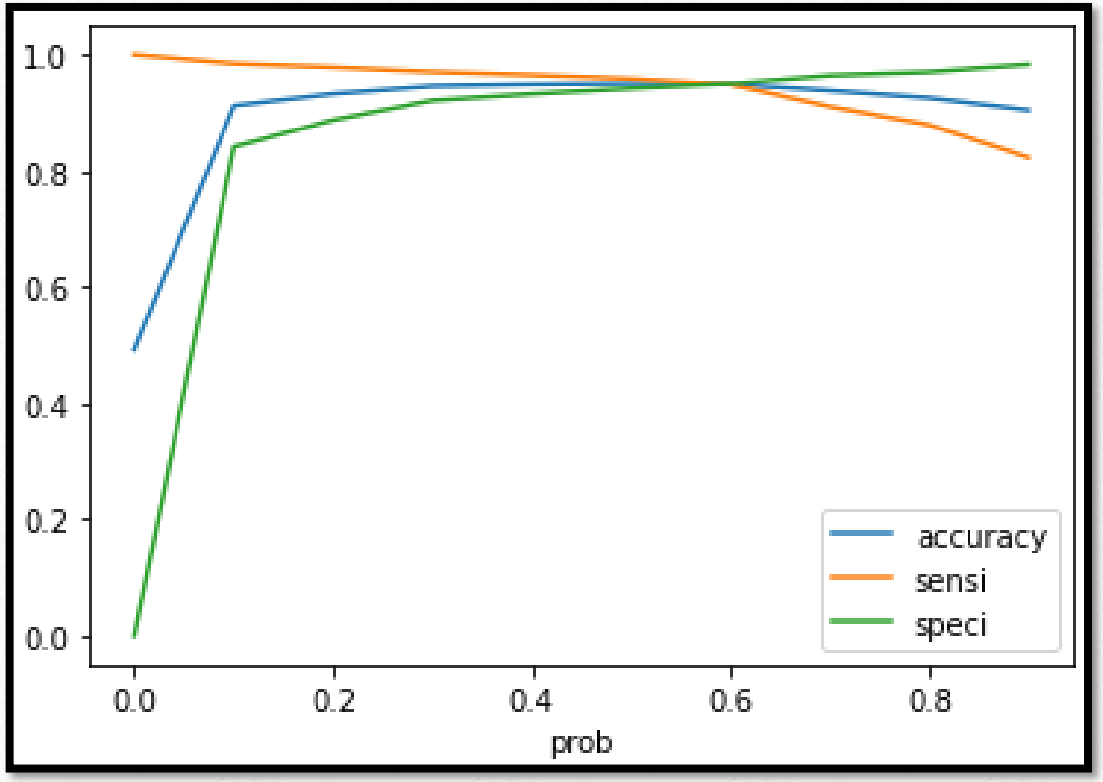
- Finding Optimal Cut-off Value which balance the Accuracy, Sensitivity, Specificity

	prob	accuracy	sensi	speci
0.0	0.0	0.493130	1.000000	0.000000
0.1	0.1	0.912323	0.984962	0.841652
0.2	0.2	0.932606	0.978770	0.887694
0.3	0.3	0.945692	0.970809	0.921256
0.4	0.4	0.948746	0.965502	0.932444
0.5	0.5	0.950273	0.958868	0.941910
0.6	0.6	0.949400	0.948695	0.950086
0.7	0.7	0.938059	0.910659	0.964716
0.8	0.8	0.925627	0.879257	0.970740
0.9	0.9	0.904689	0.823529	0.983649

Model Evaluation

- We plot the Accuracy, Sensitivity, Specificity:

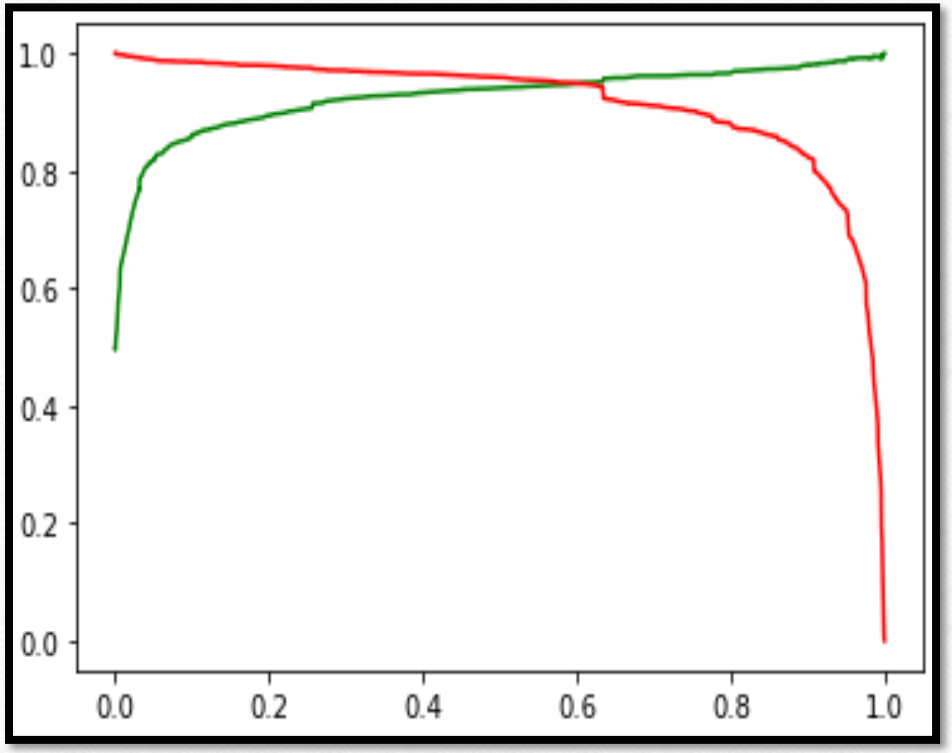
  From the graph we observed that threshold value is around 0.57



Model Evaluation

- By applying the model on "Test set" we get the following Precision recall curve:

  From the graph we observed that threshold value is around 0.60



Model Evaluation

- By making the predictions on “Test set” using 0.60 as cut off will get

1. Accuracy 95%
2. Precision 95%
3. Recall 94%

Conclusion

- In order to achieve higher conversion rate company should target people who

  1.Unemployed/Freshers & Working Professionals

  2.Have given contact information

  3.They where Lead Source from google & Spend lot of time on website

Thank You !