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. A typical lead conversion process can be represented using the following funnel:

Lead Conversion Process - Demonstrated as a funnel Lead Conversion Process - Demonstrated as a funnel As you can see, there are a lot of leads generated in the initial stage (top) but only a few of them come out as paying customers from the bottom. In the middle stage, you need to nurture the potential leads well (i.e. educating the leads about the product, constantly communicating etc.) in order to get a higher lead conversion.

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%.

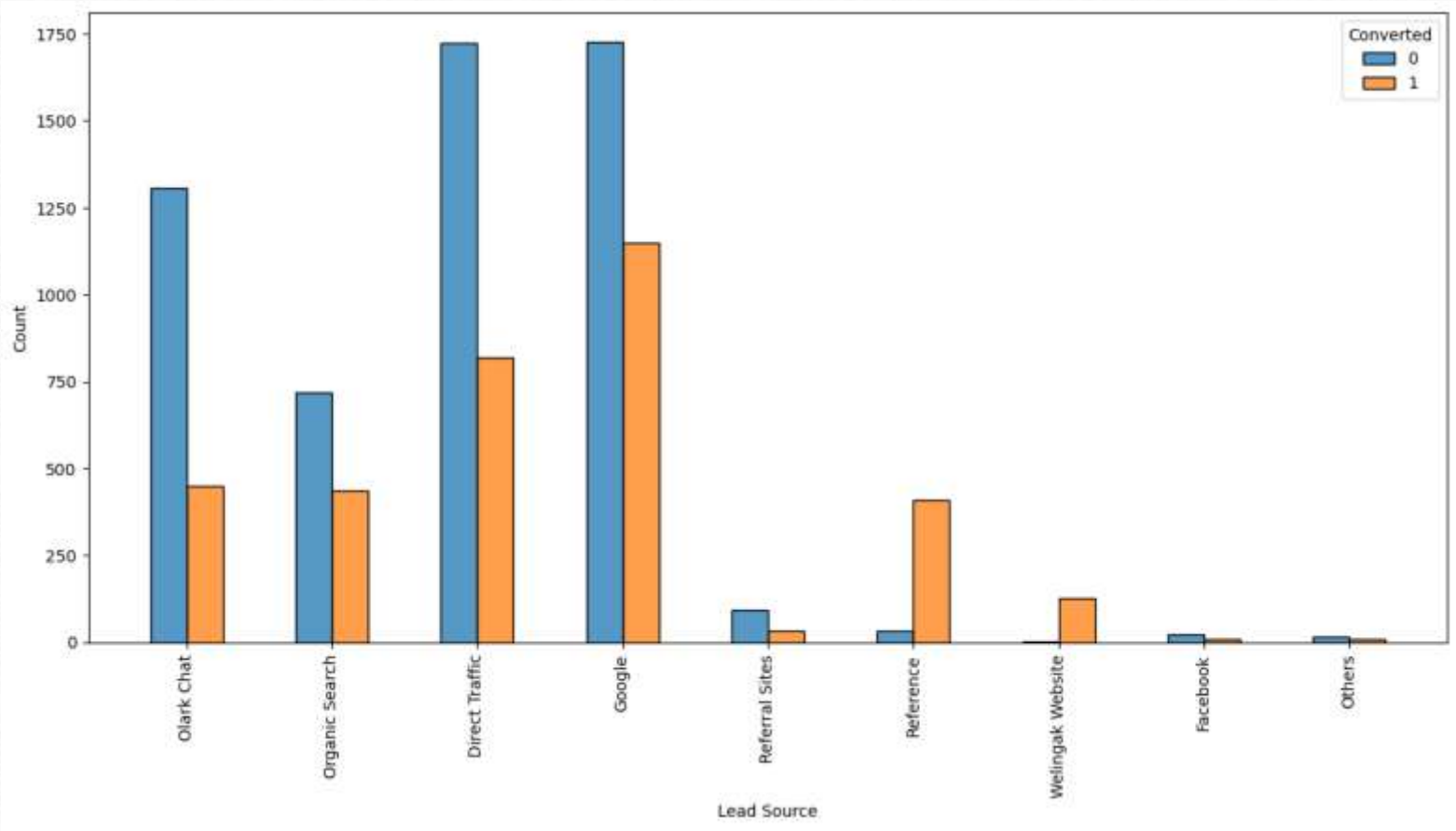
Data:

You have been provided with a leads 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. You can learn more about the dataset from the data dictionary provided in the zip folder at the end of the page. Another thing that you also need to check out are the levels present in the categorical variables. Many of the categorical variables have a level called 'Select' which needs to be handled because it is as good as a null value (think why?).

Steps to be followed:

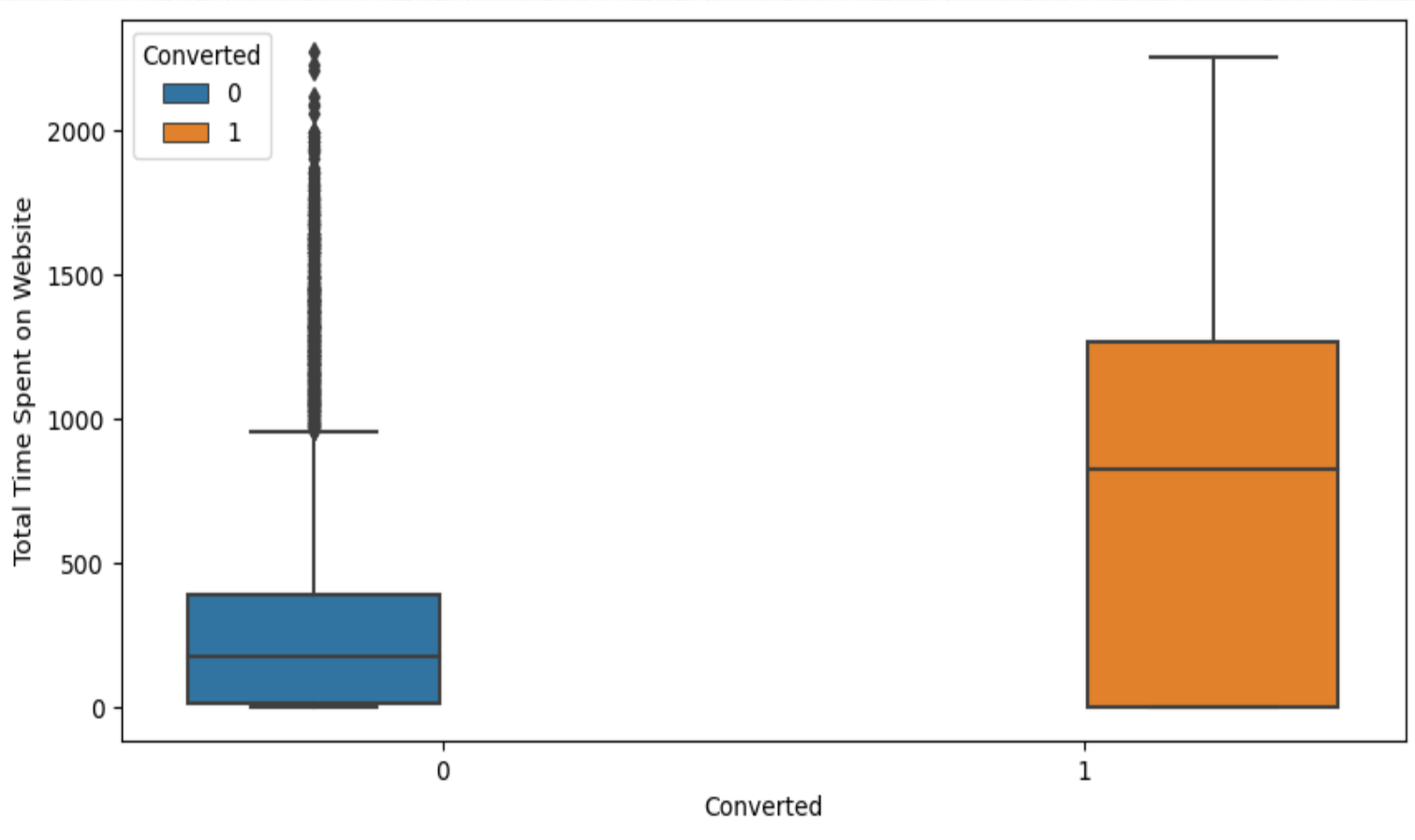
1.Data Cleaning
2.Exploratory Data Analysis
3.Data preparation for model building
4.Model building
5.Model Evalution

UNIVARIATE AND BIVARIATE ANALYSIS :
Lead Source Analysis :



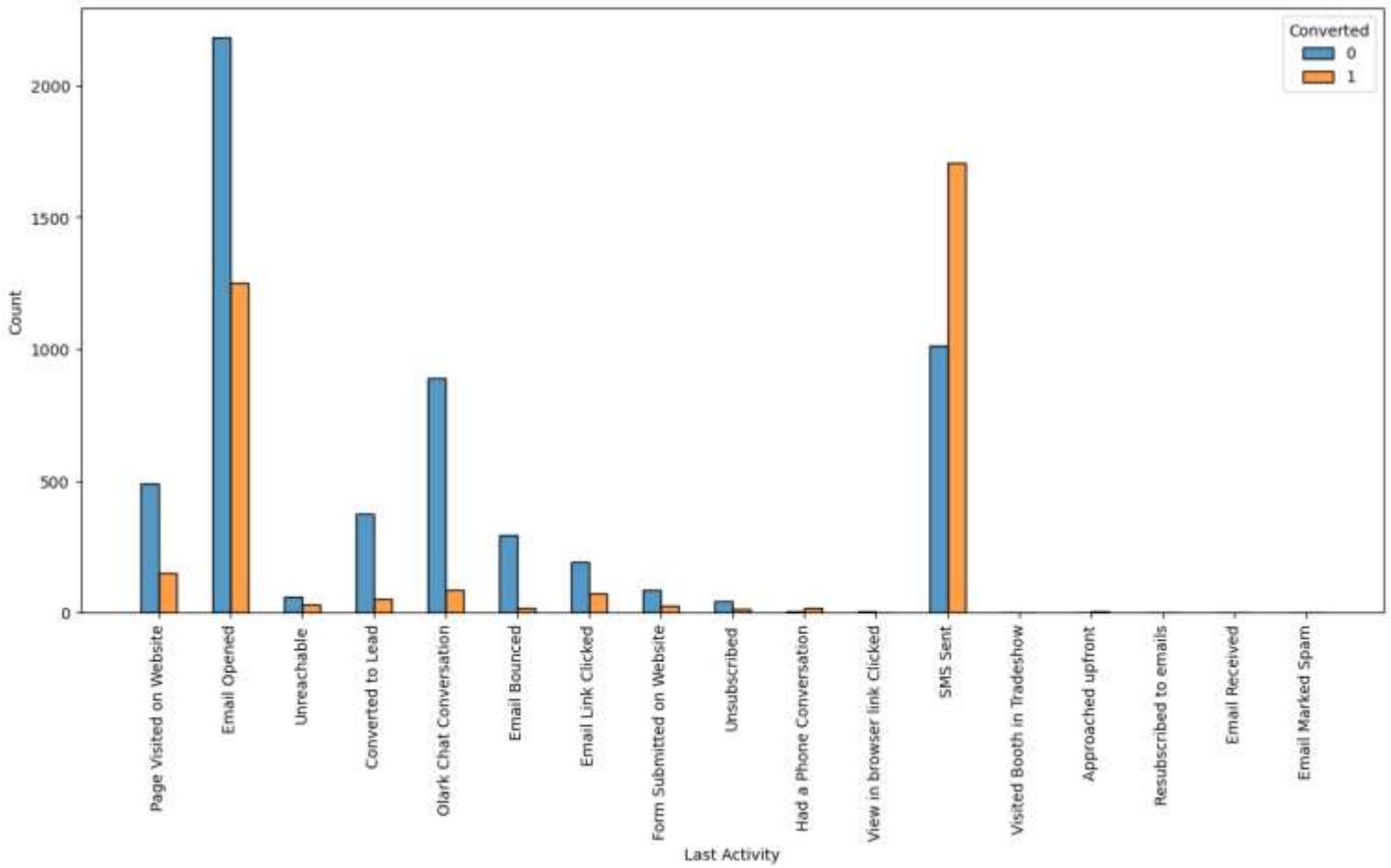
Inference:
1)Google and Direct traffic generates maximum number of leads.
2)Conversion Rate of reference leads and leads through welingak website is high.

Total Time Spent On Website :



Inference:
The candidate who are spending more time on website considered to be as leads and they are converted.

Last Activity :

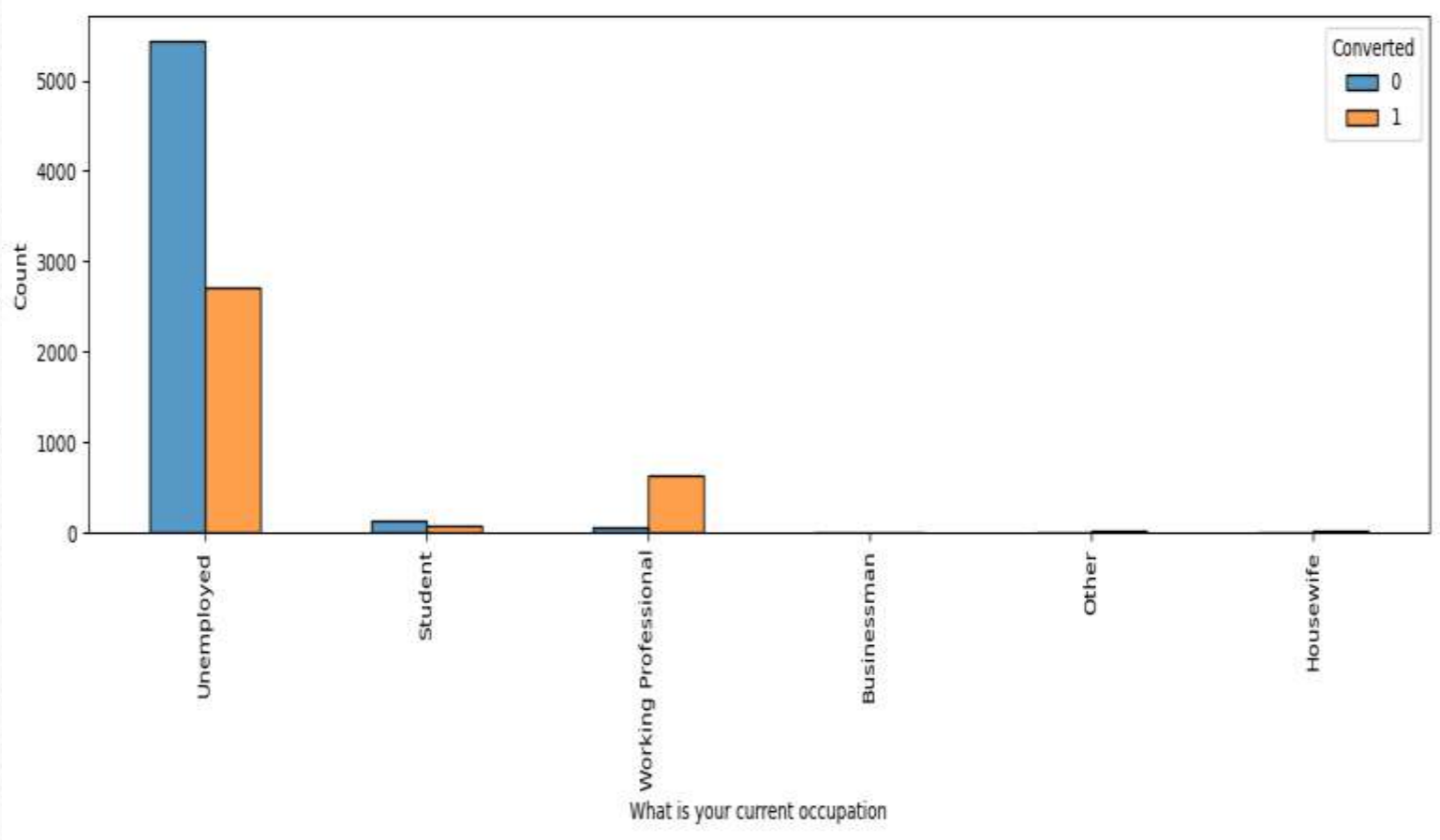


Inference:

1.Most of the lead have their Email opened as their last activity.

2.Conversion rate for leads with last activity as SMS Sent is more

What is your current occupation :

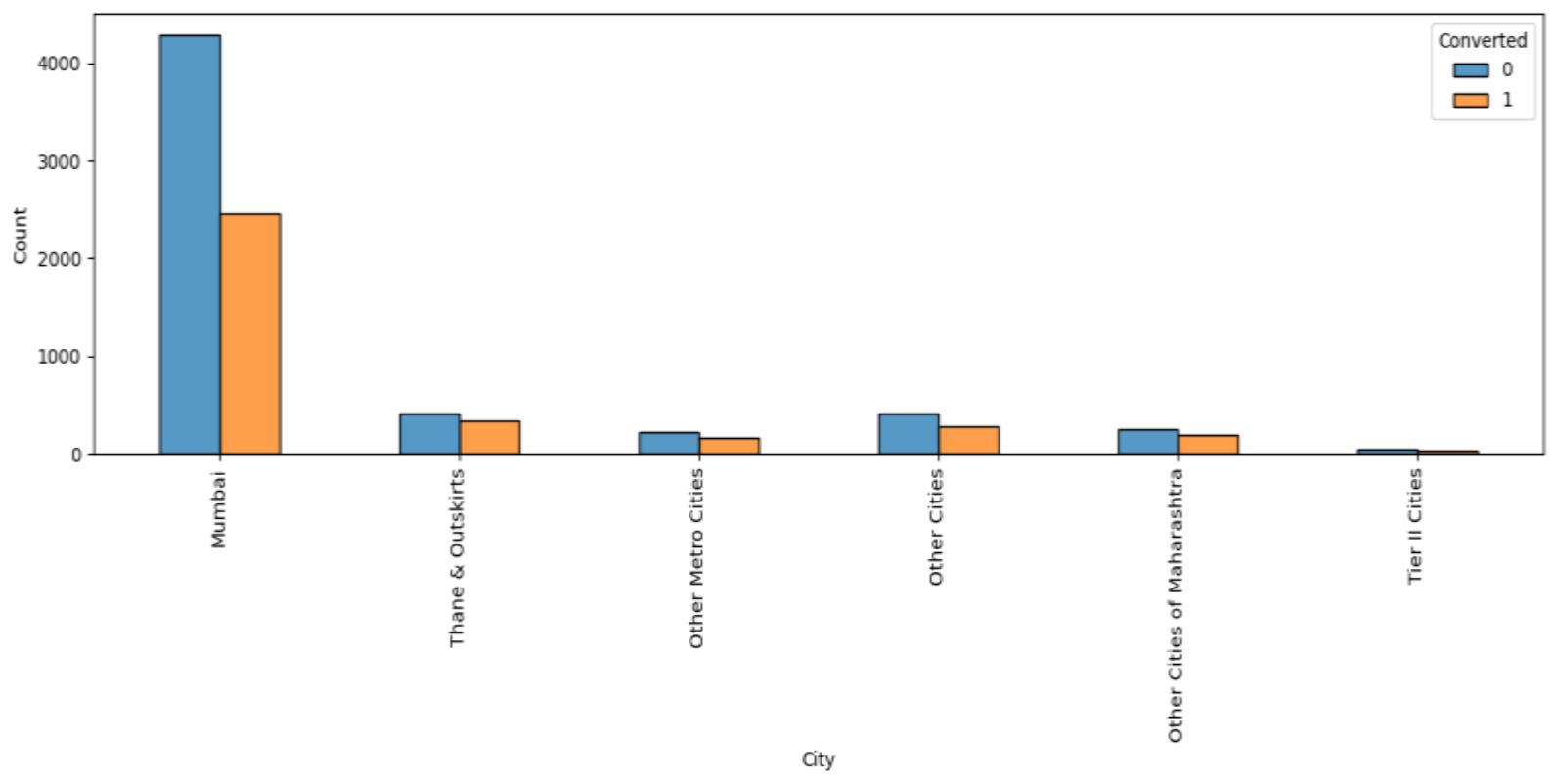


Inference:

1.Working Professional are having high chances of joining the course.

2.Unemployed leads also have chances of joining the course.

City :



Inference:
Most of the leads are from mumbai.

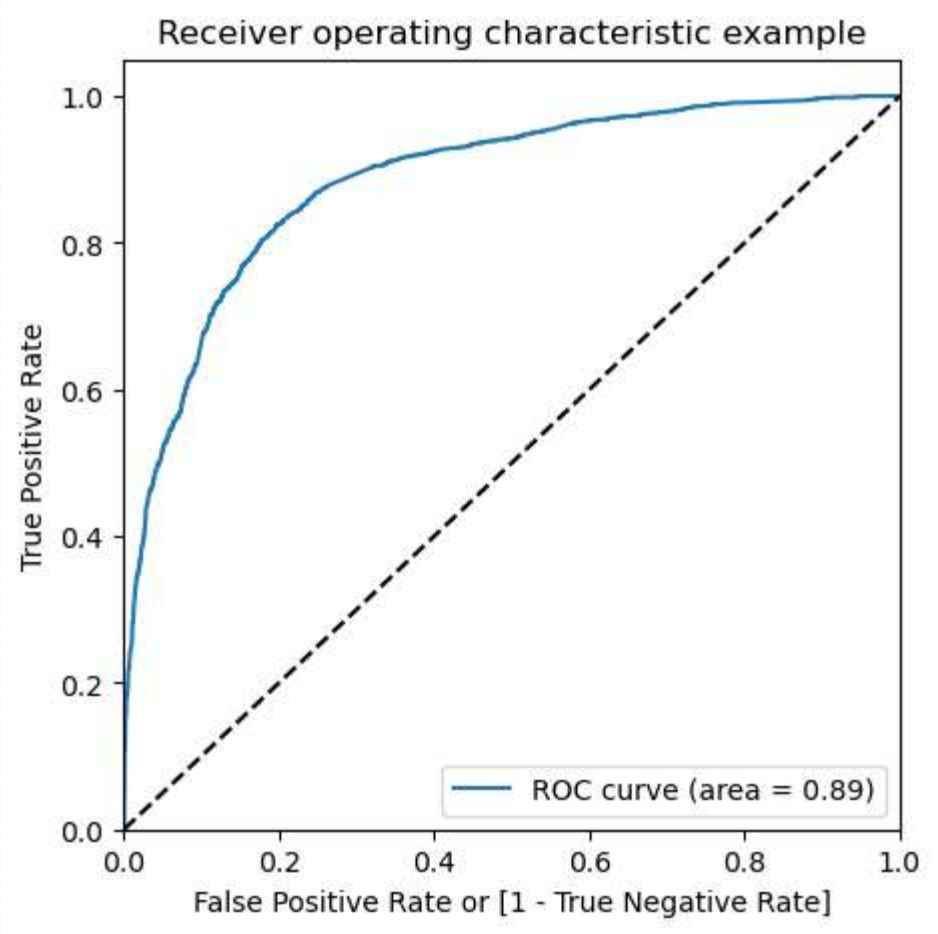
Model Building :

Generalized Linear Model Regression Results

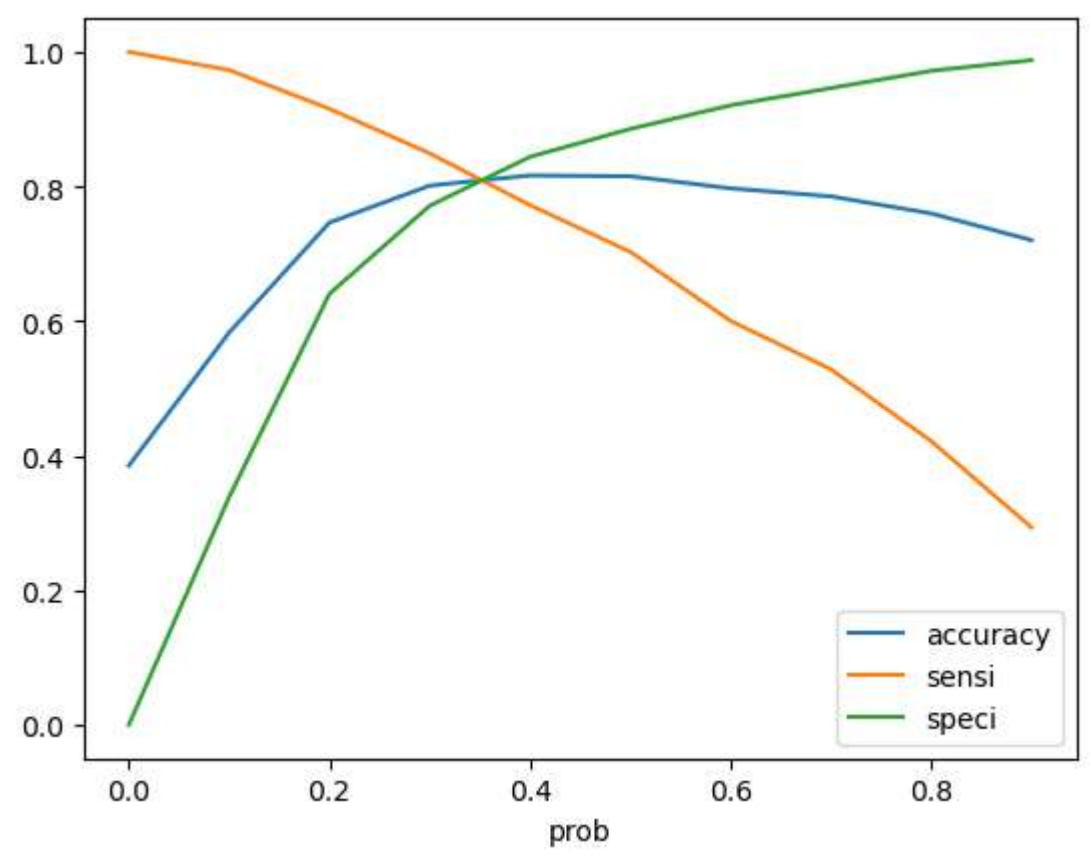
Dep. Variable:	Converted	No. Observations:	6351
Model:	GLM	Df Residuals:	6338
Model Family:	Binomial	Df Model:	12
Link Function:	Logit	Scale:	1.0000
Method:	IRLS	Log-Likelihood:	-2618.9
Date:	Mon, 19 Feb 2024	Deviance:	5237.9
Time:	21:29:26	Pearson chi2:	6.53e+03
No. Iterations:	7	Pseudo R-squ. (CS):	0.3985
Covariance Type:	nonrobust		

	coef	std err	z	P>|z|	[0.025	0.975]
const	-0.0188	0.125	-0.151	0.880	-0.264	0.226
Do Not Email	-1.5232	0.176	-8.637	0.000	-1.869	-1.178
Total Time Spent on Website	1.0938	0.040	27.249	0.000	1.015	1.173
Lead Origin_Landing Page Submission	-1.1996	0.127	-9.426	0.000	-1.449	-0.950
Lead Source_Olark Chat	1.0716	0.122	8.782	0.000	0.832	1.311
Lead Source_Reference	3.3024	0.241	13.701	0.000	2.830	3.775
Lead Source_Welingak Website	5.7975	0.728	7.963	0.000	4.371	7.224
Last Activity_Olark Chat Conversation	-0.9758	0.171	-5.698	0.000	-1.311	-0.640
Last Activity_Other Activity	1.6604	0.604	2.747	0.006	0.476	2.845
Last Activity_SMS Sent	1.2805	0.075	17.165	0.000	1.134	1.427
Specialization_Others	-1.2044	0.125	-9.620	0.000	-1.450	-0.959
What is your current occupation_Working Professional	2.6067	0.194	13.456	0.000	2.227	2.986
Last Notable Activity_Modified	-0.8947	0.081	-11.049	0.000	-1.053	-0.736

ROC Curve :


Receiver operating characteristic example
True Positive Rate
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)

Plot for knowing optimal cut off :



Inference :

From the curve above, 0.34 is the optimum point to take it as a cutoff probability.

Observations :

Train data:

Accuracy : 81.1%
Sensitivity : 81.8%
Specificity : 80.5%
Precision : 79.3%
Recall : 81.8%

Test data:

Accuracy : 80.4%
Sensitivity : 80.3%
Specificity : 80.5%
precision : 70.2%
Recall : 80.3%

Inferences from the model :

1.The company should make calls to the leads coming from the lead sources "Welingak Websites" and "Reference" as these are more likely to get converted.

2.The company should make calls to the leads who are the "working professionals" as they are more likely to get converted.

3.The company should make calls to the leads who spent "more time on the websites" as these are more likely to get converted.

4.The company should make calls to the leads coming from the lead sources "Olark Chat" as these are more likely to get converted.

5.The company should make calls to the leads whose last activity was SMS Sent as they are more likely to get converted.

6.The company should not make calls to the leads whose last activity was "Olark Chat Conversation" as they are not likely to get converted.

7.The company should not make calls to the leads whose lead origin is "Landing Page Submission" as they are not likely to get converted.

8.The company should not make calls to the leads whose Specialization was "Others" as they are not likely to get converted.

9.The company should not make calls to the leads who chose the option of "Do not Email" as "yes" as they are not likely to get converted.

THANK YOU