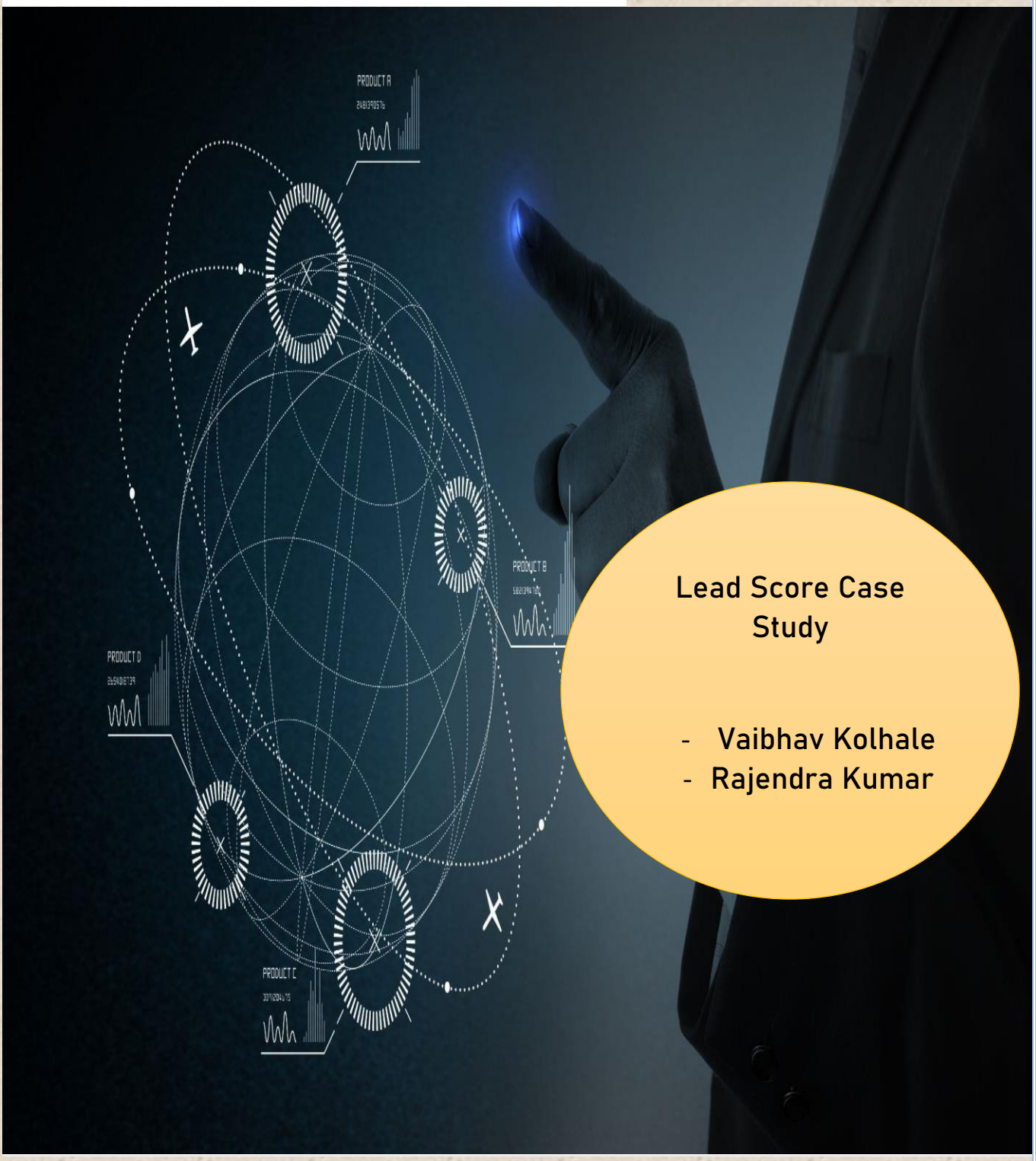


Lead Score Case Study

- Vaibhav Kolhale
- Rajendra Kumar

Problem Statement:-

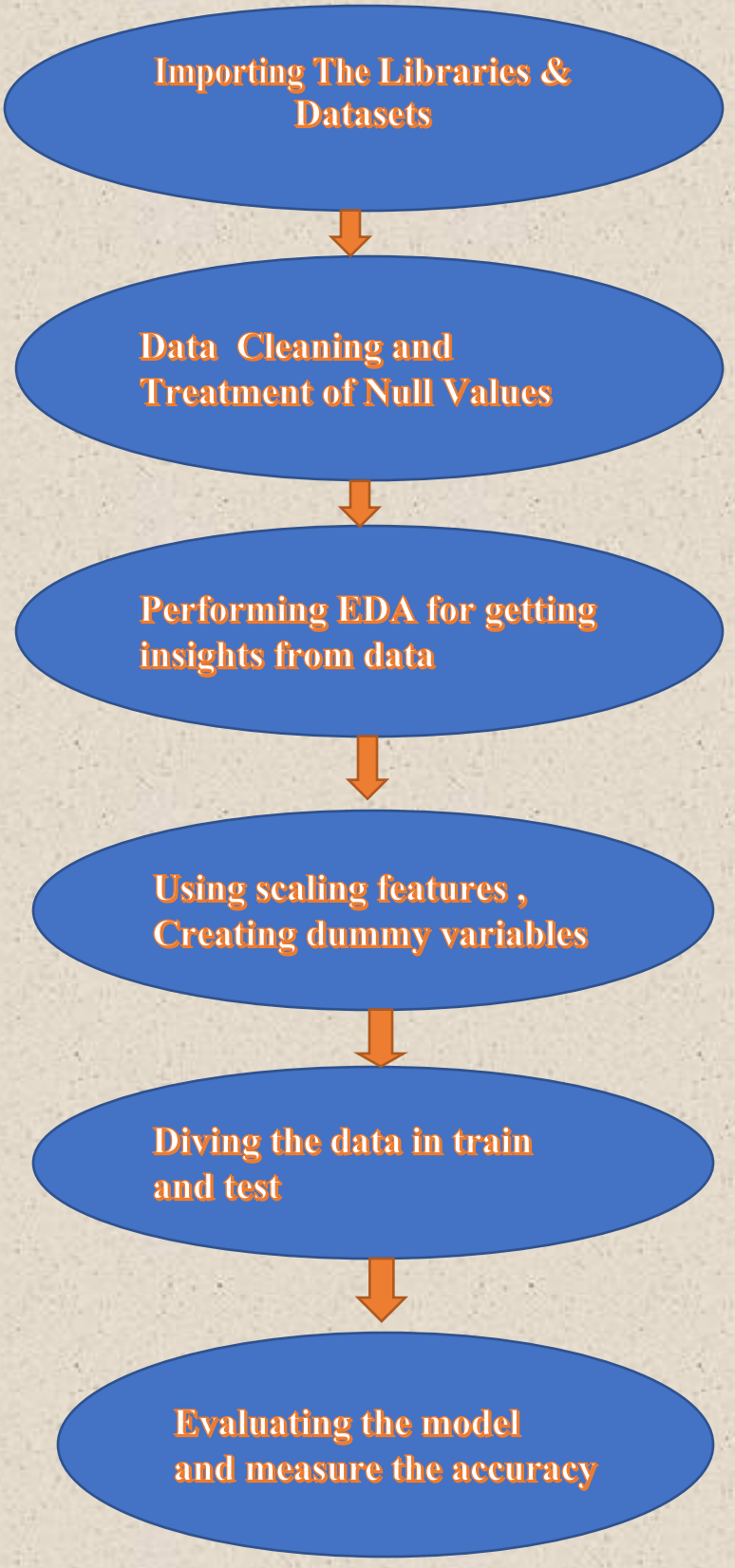
An education company named X Education sells online courses to industry professionals. On any given day, many
professionals who are interested in the courses land on their website and browse for courses.
X Education wants to select the most promising leads. The company market the course on several website . The
contact details of leads obtain through the form they filled on the website. 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%.

X education wants to develop the model which will give them the most promising leads .

Business Goal :-

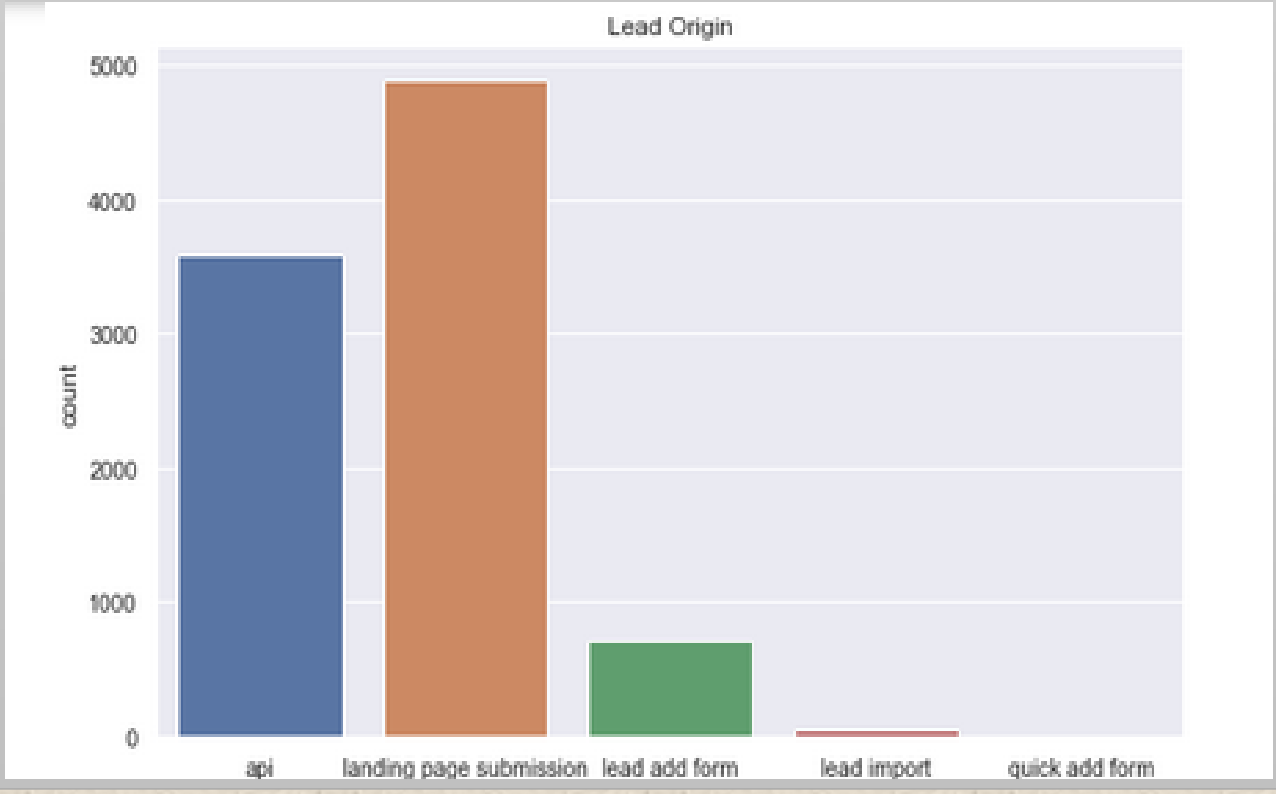
Build a logistic regression model to assign a lead score between 0 and 100 to each of the leads which can be used
by the company to target potential leads. A higher score would means Promising Leads
There are some more problems presented by the company which your model should be able to adjust to if the
company's requirement changes in the future so you will need to handle these as well. These problems are
provided in a separate doc file. Please fill it based on the logistic regression model you got in the first step. Also,
make sure you include this in your final PPT where you'll make recommendations.

STEPS TAKEN IN DEVELOPING THE MODEL:



EDA Finding:

It has been observed that the most of leads we are getting through the landing page submission.



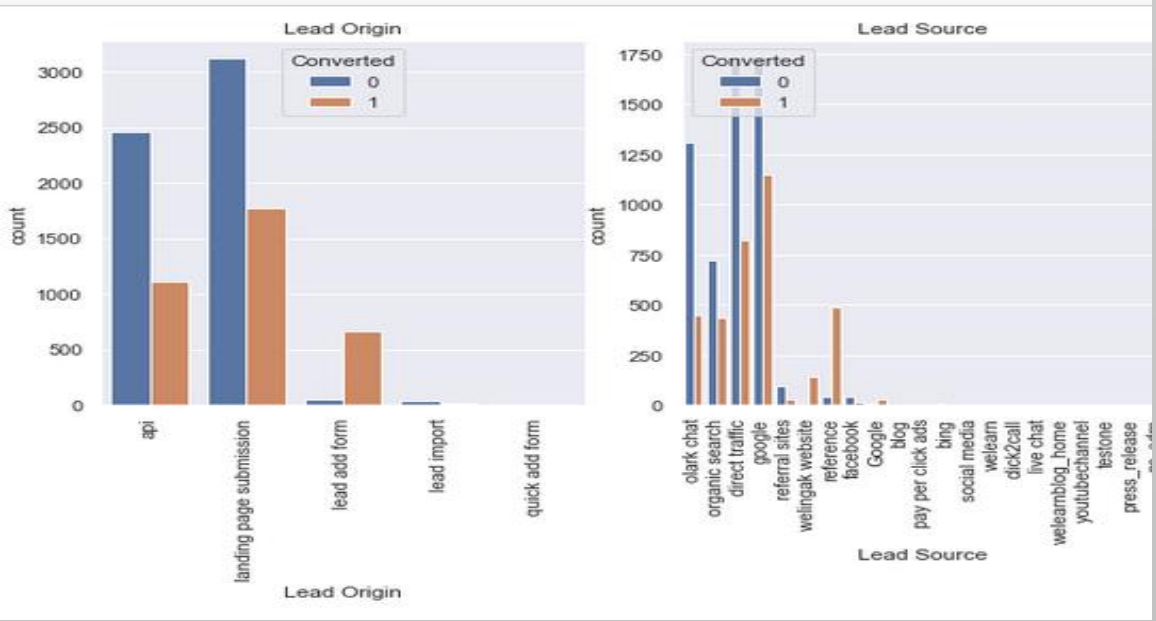
Lead origin against the converted :-

• It has been observed that the most promising leads are going to get through landing page submission.

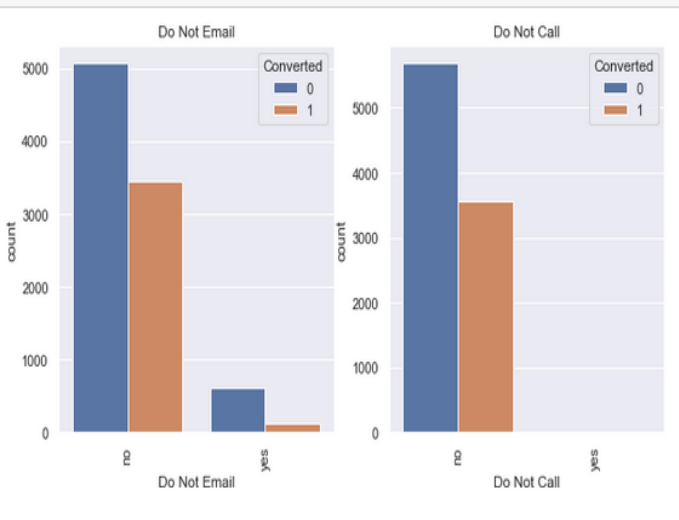
• The second observation we can draw from fig is organic search, direct traffic and google can lead us to most promising customers.

Represent the Customer said No

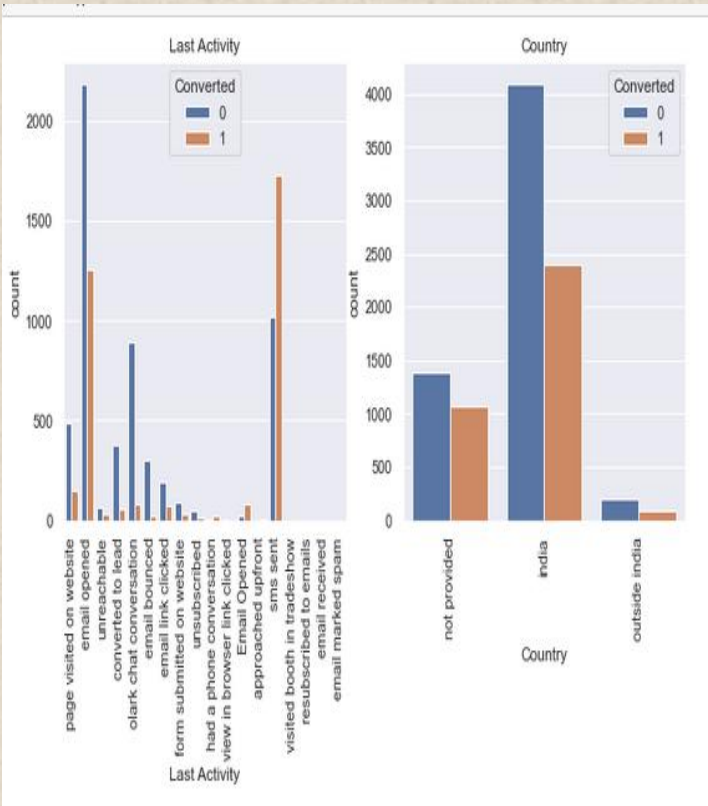
Represent the Customer said Yes



Distribution Of Cash Loans And Revolving Loans :-



From the diagram figure we can say the customer who is opting for email or call and it has been observed that customer who is opting for no email and call more like to enroll for the course

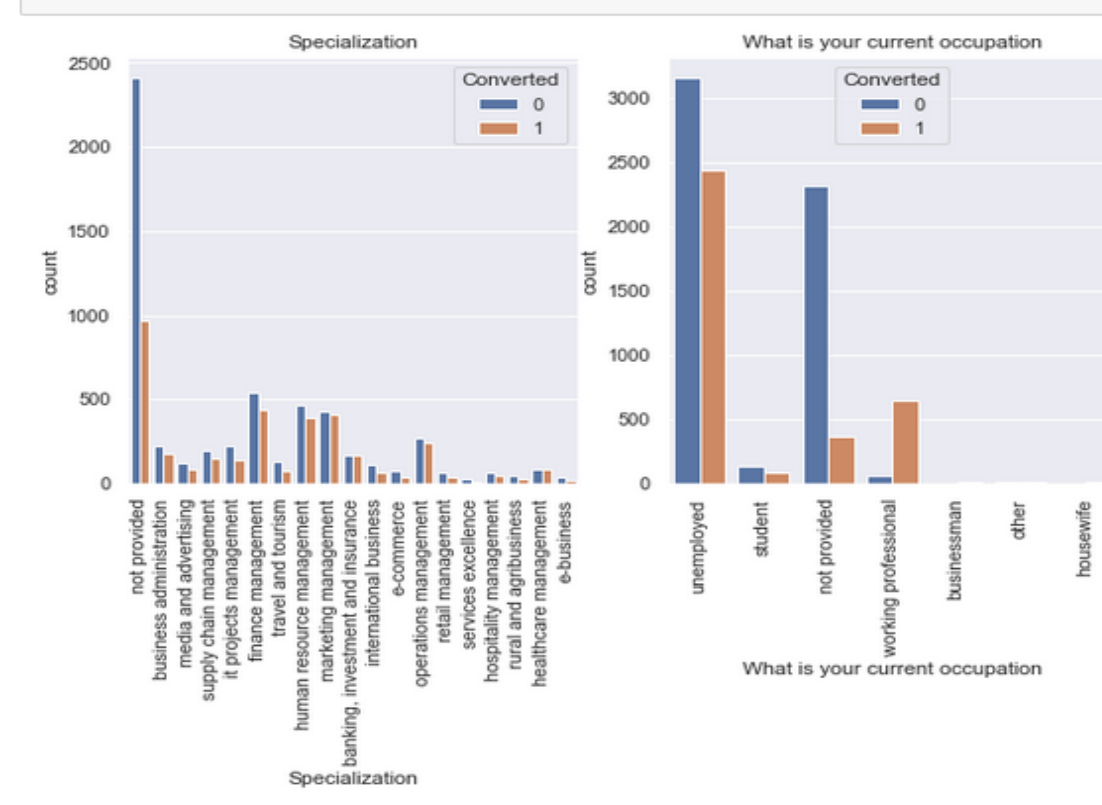


In the diagram fig. we are observing that the enrolling customer are from the India and last activity is sms sent or email opened.

- 0 = Not Promising Lead
- 1 = Promising Lead

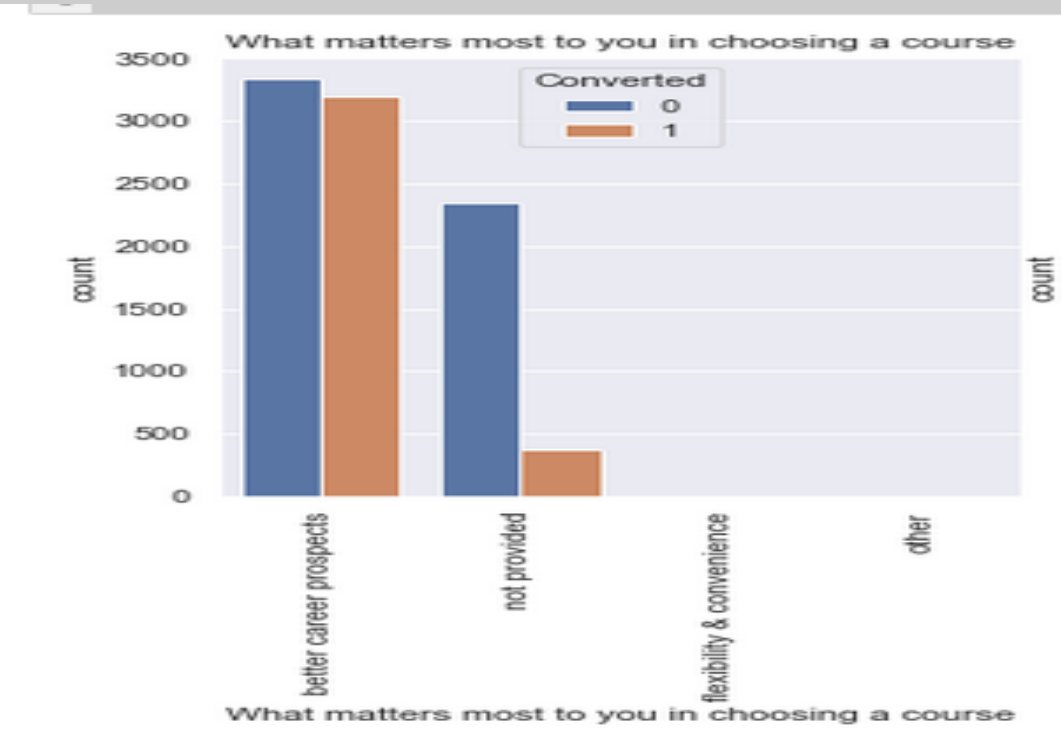
Specialization and Status of employment

It has been observed that the person who is not having any specialization or person who is unemployed are more likely to enroll to the course.



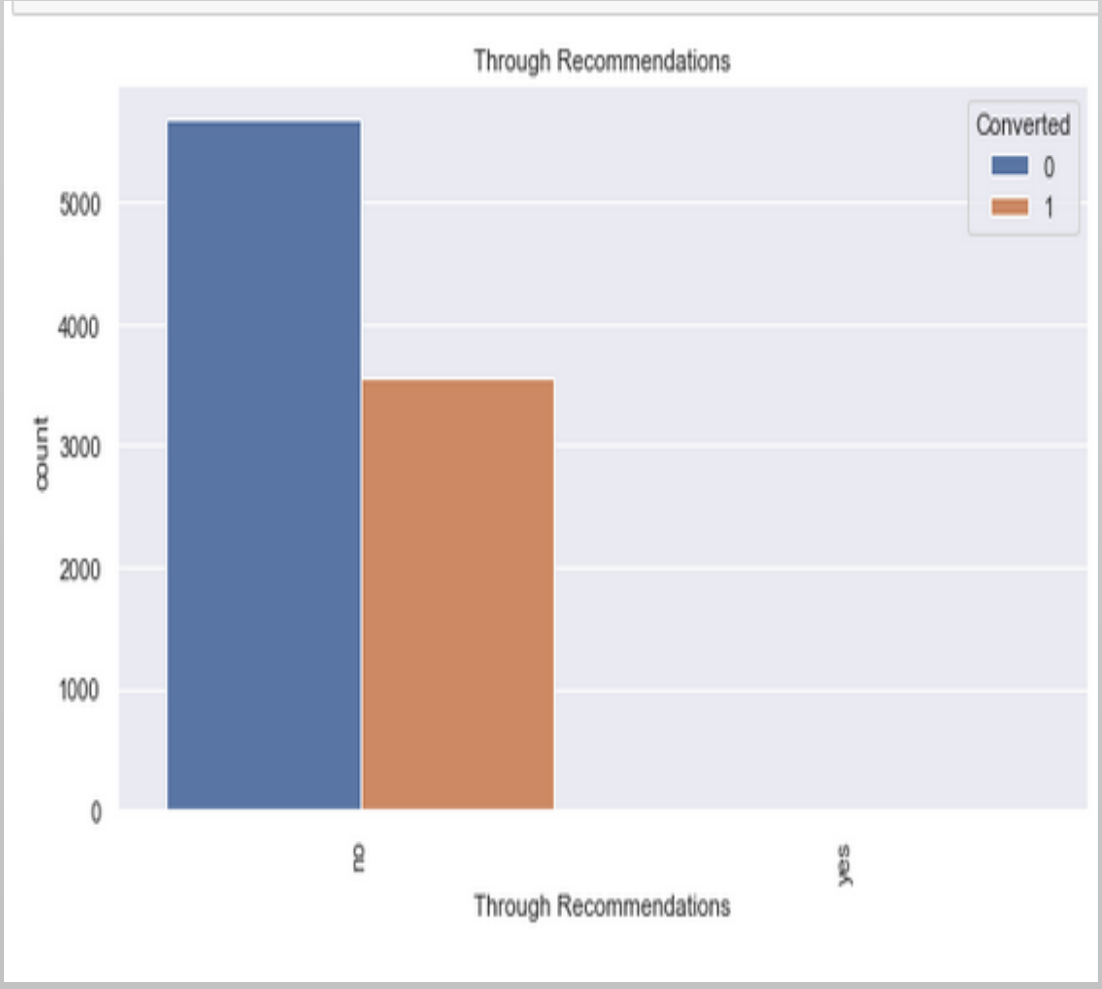
What matters the most

The candidate who are more drawn to career and looking for a career having better career prospects are most likely to join the course or we can say our promising leads



Recommendation :-

It has been observed that majority people are coming without recommendation



Modelling and VIF :-

This is the final model which has been obtain by us and this models are having p value less than 0.5 and VIF is less than 5 and hence we can conclude that there is no sign of multicollinearity.

60]:

Generalized Linear Model Regression Results

Dep. Variable:	Converted	No. Observations:	6468
Model:	GLM	Df Residuals:	6452
Model Family:	Binomial	Df Model:	15
Link Function:	Logit	Scale:	1.0000
Method:	IRLS	Log-Likelihood:	-2616.6
Date:	Tue, 18 Oct 2022	Deviance:	5233.1
Time:	14:38:42	Pearson chi2:	6.78e+03
No. Iterations:	22	Pseudo R-squ. (CS):	0.4056
Covariance Type:	nonrobust		

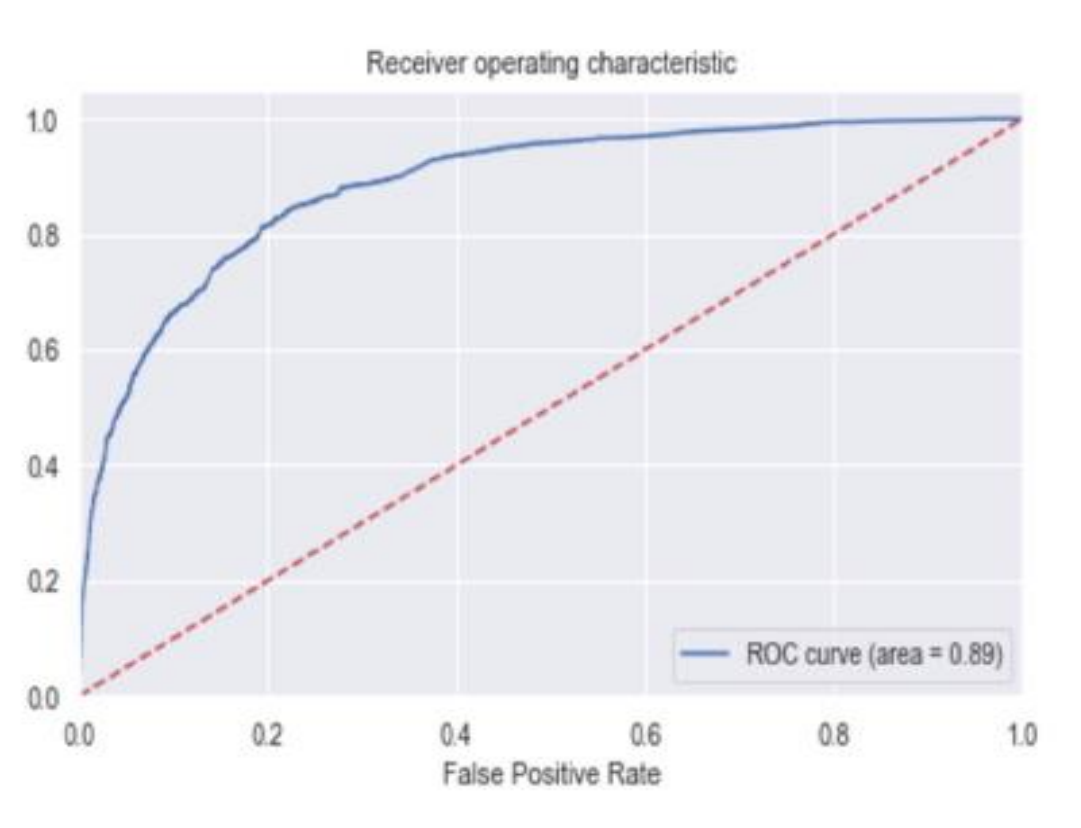
	coef	std err	z	P>|z|	[0.025	0.975]
const	-0.5025	0.121	-4.166	0.000	-0.739	-0.266
Total Time Spent on Website	4.0031	0.153	26.237	0.000	3.704	4.302
Lead Origin_landing page submission	-1.3981	0.121	-11.561	0.000	-1.635	-1.161
Lead Origin_lead add form	2.4668	0.194	12.715	0.000	2.087	2.847
Specialization_not provided	-0.9689	0.123	-7.879	0.000	-1.210	-0.728
Lead Source_welingak website	2.3725	0.745	3.186	0.001	0.913	3.832
Do Not Email_yes	-1.1638	0.167	-6.983	0.000	-1.490	-0.837
Last Activity_approached upfront	24.1266	2.73e+04	0.001	0.999	-5.35e+04	5.36e+04
Last Activity_had a phone conversation	1.1573	0.951	1.217	0.223	-0.706	3.021
Last Activity_sms sent	1.2882	0.074	17.370	0.000	1.143	1.434
What is your current occupation_housewife	23.9165	2.23e+04	0.001	0.999	-4.37e+04	4.37e+04
What is your current occupation_not provided	-0.9795	0.088	-11.190	0.000	-1.151	-0.808
What is your current occupation_working professional	2.4380	0.191	12.765	0.000	2.064	2.812
Last Notable Activity_had a phone conversation	2.0267	1.472	1.377	0.169	-0.858	4.912
Last Notable Activity_modified	-0.9143	0.080	-11.443	0.000	-1.071	-0.758
Last Notable Activity_unreachable	1.6091	0.544	2.961	0.003	0.544	2.674

ROC Curve:-

It has been observed that

1. The curve is closer to the left side of the border than to the right side hence our model is having great accuracy.

2. The area under the curve is 89% of the total area.

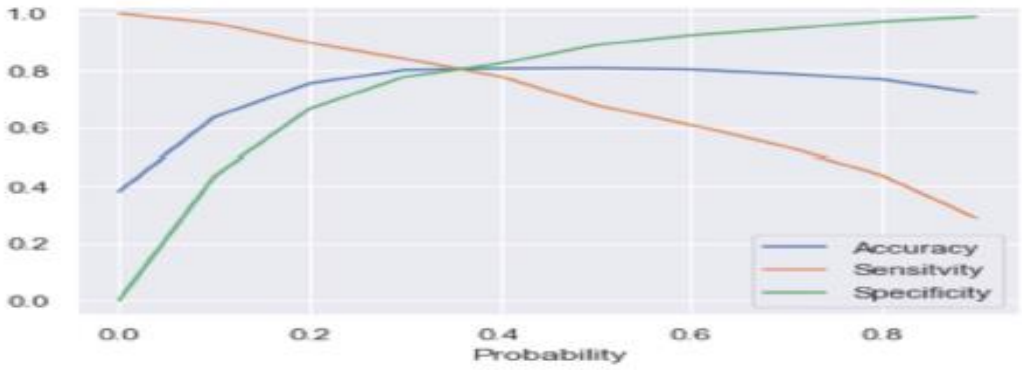


Optimal Probability cut off :-

After creating series of points let's check the possibilities of choosing any one points from 0 to 0.9. We will do this by finding 'Accuracy', 'Sensitivity' and 'Specificity' for each points. These three methods will tell us how our model is - whether it is having low accuracy or high or number of relevance data points is high or low etc

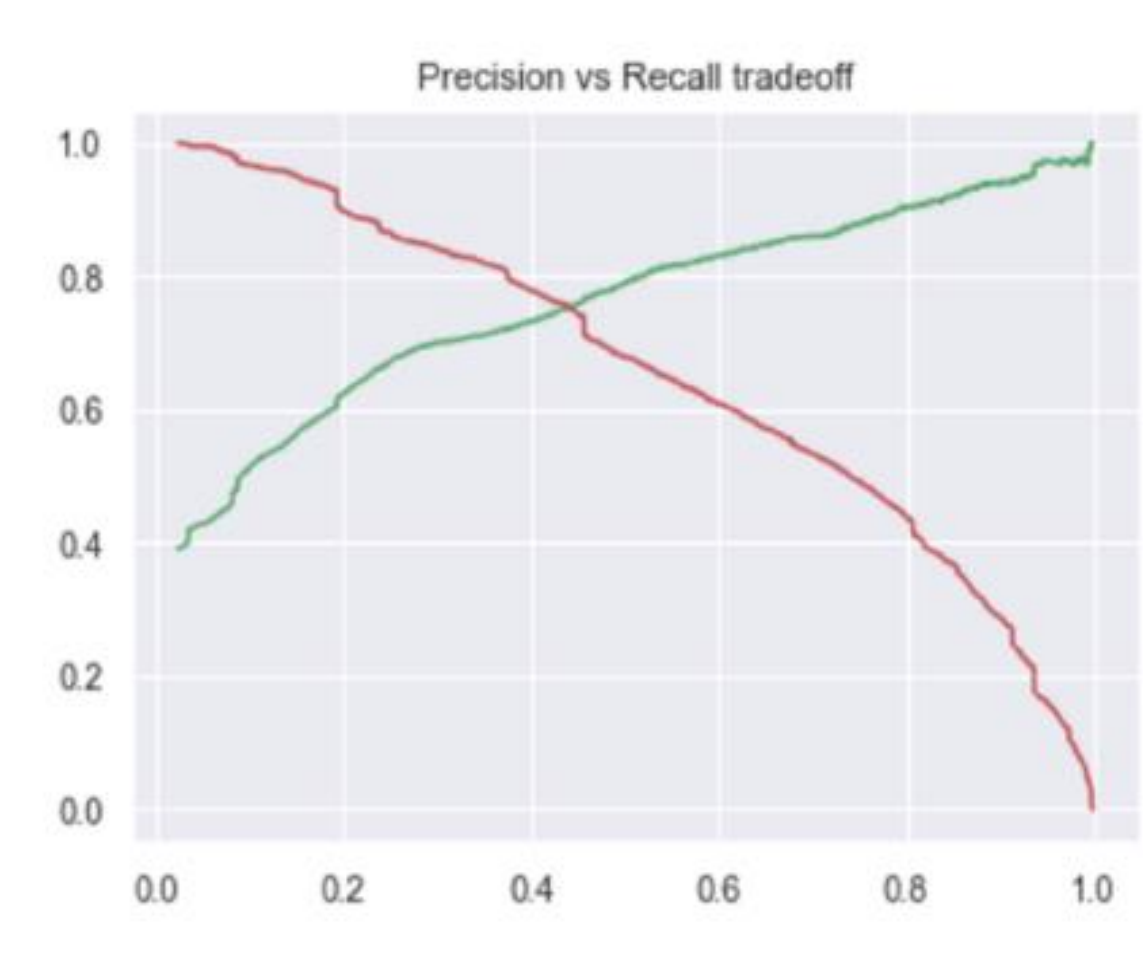
From the above curve, 0.4 is the optimum point for taking probability cutoff as the meeting point is slightly before from 0.4 hence final cutoff we choose is 0.40. Also we can see that there is a trade off between sensitivity and specificity.

	Converted	Converted_probability	ID	0.0	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9
1871	0	0.374573	1871	1	1	1	1	0	0	0	0	0	0
6795	0	0.127927	6795	1	1	0	0	0	0	0	0	0	0
3516	0	0.223629	3516	1	1	1	0	0	0	0	0	0	0
8105	0	0.826204	8105	1	1	1	1	1	1	1	1	1	0
3934	0	0.087507	3934	1	0	0	0	0	0	0	0	0	0



Precision and Recall :-

we can see that there is a trade off between Precision and Recall and the meeting point is nearly at 0.5.



CONCLUSION:

1. WE HAVE HIGH RECALL SCORE THAN PRECISION SCORE WHICH WE WERE EXACTLY LOOKING
FOR.

2. IN BUSINESS TERMS, THIS MODEL HAS AN ABILITY TO ADJUST WITH THE COMPANY'S
REQUIREMENTS IN COMING FUTURE.

3. THIS CONCLUDES THAT THE MODEL IS IN STABLE STATE.

IT WAS FOUND THAT THE VARIABLES THAT MATTERED THE MOST IN THE POTENTIAL BUYERS
ARE:
1. TOTAL TIME SPENT ON WEBSITE
2. WHAT IS YOUR CURRENT OCCUPATION_WORKING PROFESSIONAL
3. LAST ACTIVITY_SMS SENT
4. LEAD ORIGIN_LEAD ADD FORM

Thank You