LEAD SCORING CASE STUDY

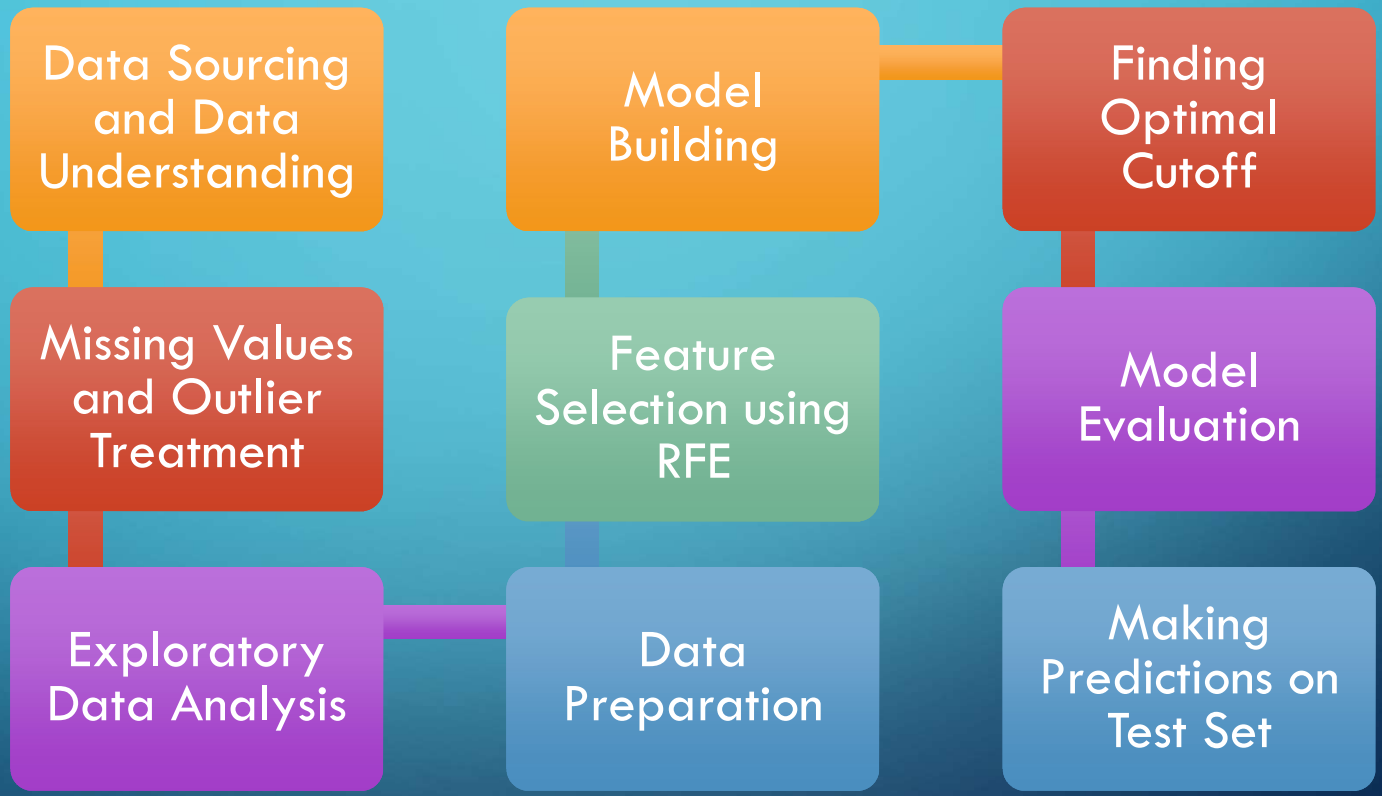
BUSINESS UNDERSTANDING

- X Education company 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%.

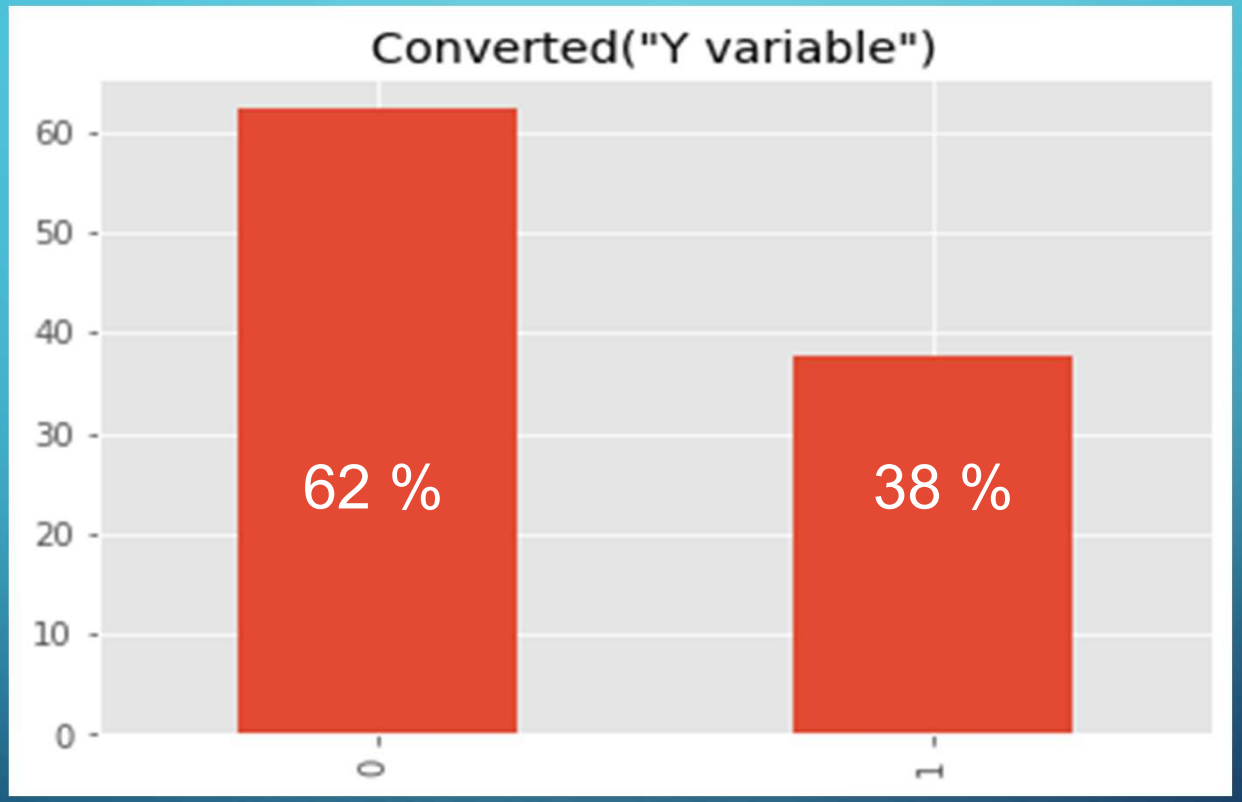
BUSINESS OBJECTIVE

- Although X Education gets a lot of leads, its lead conversion rate is very poor(30 %).
- To make this process more efficient, the company wishes to identify the most potential leads, also known as ‘Hot Leads’.
- 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%.

METHODOLOGY:

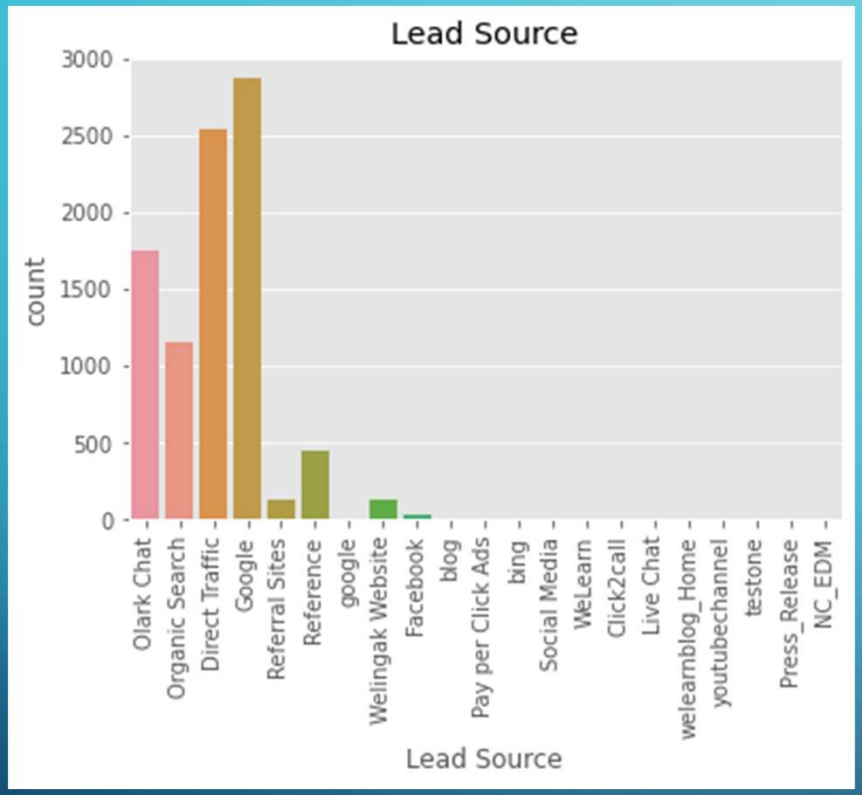


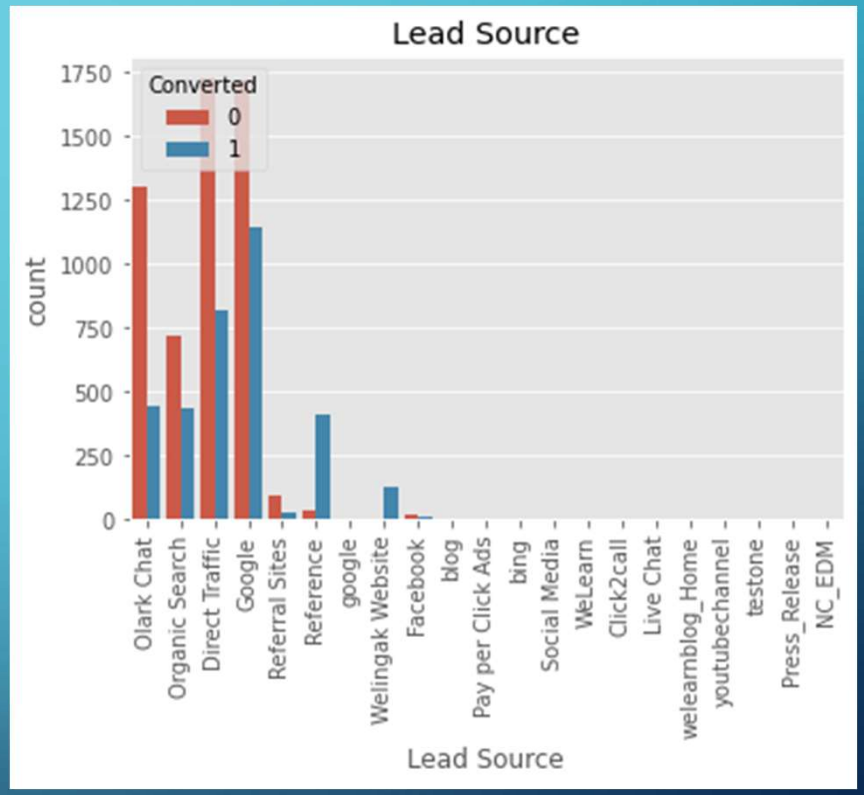
DATA IMBALANCE CHECK ON TARGET VARIABLE



Target variable has Data imbalance. 62% peoples are not converted while 38% people are converted.

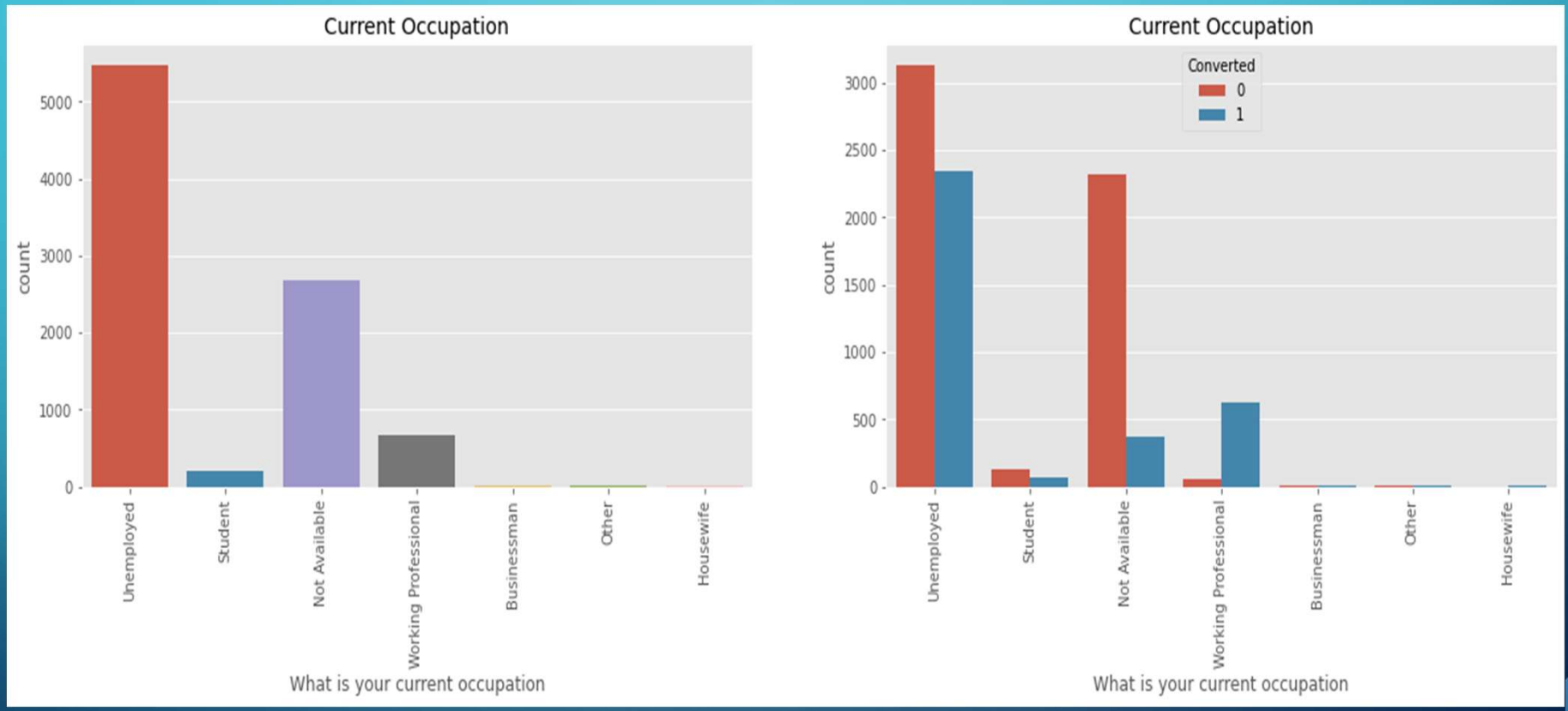
LEAD SOURCE ANALYSIS





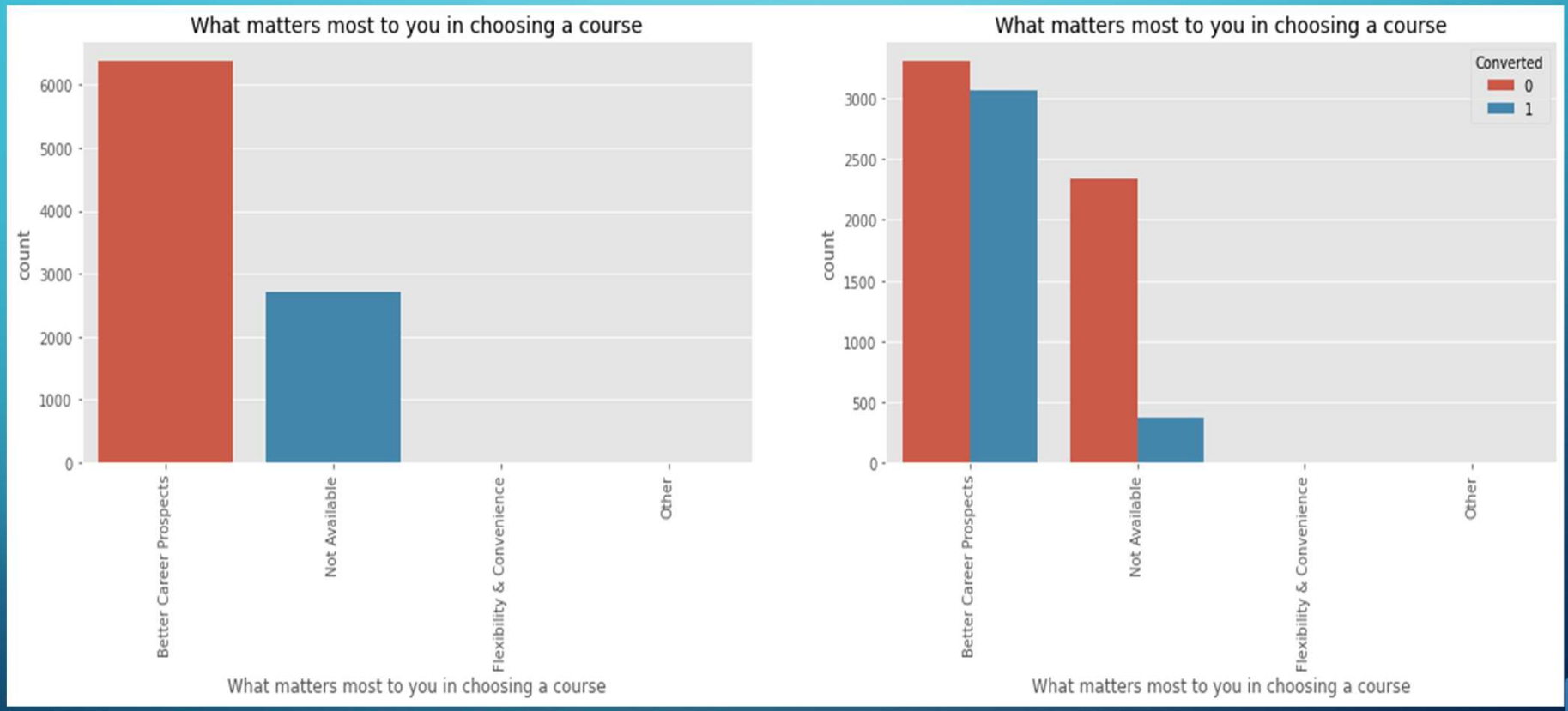
Most of the lead sources are either from Google or they are Direct Traffic. While Leads from Google and Organic Search have high conversion rate compared to others.

LEAD OCCUPATION ANALYSIS



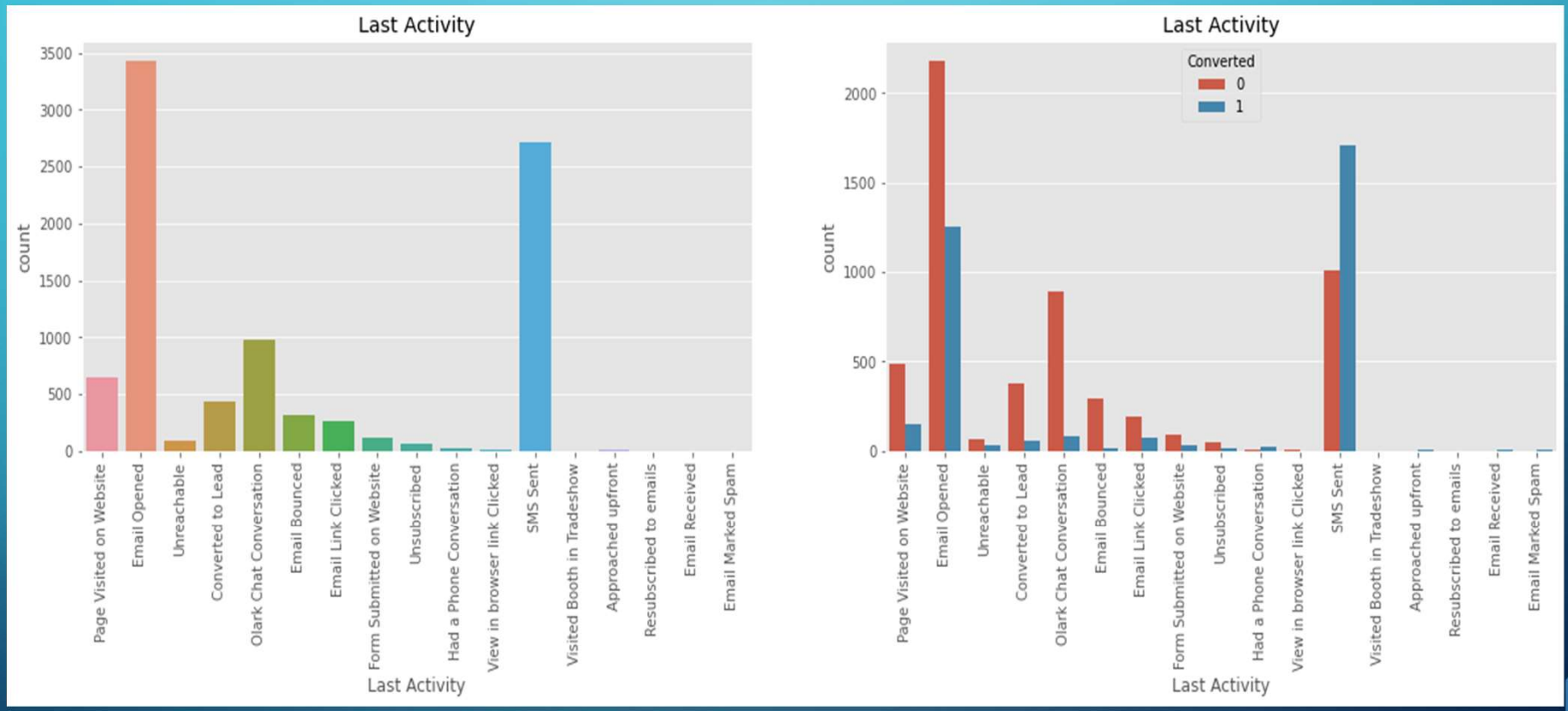
People who are Unemployed have shown most interest in the course. While working professional people have shown a greater number of conversions followed by unemployed people. X-education should target Working professional people more for higher lead conversion.

EXPECTATION FROM THE COURSE



People who have expectation of “Better Career Prospects” have largely shown interest in this course with almost 50% conversion. X-education should focus on providing more career opportunities along with course to get more lead conversions.

LAST ACTIVITY PERFORMED BY LEAD



People who have sent the SMS to X-education have shown huge conversion. X-education should focus on LEAD who have contacted X-Education through similar SMS.

TOP FEATURES SELECTED BY REF

a. 'Total Time Spent on Website'
b. 'Lead Origin_Lead Add Form'
c. 'Lead Source_Direct Traffic'
d. 'Lead Source_Welingak Website'
e. 'Do Not Email_Yes'
f. 'Last Activity_Olark Chat Conversation'
g. 'What is your current occupation_Working Professional',
h. 'Last Notable Activity_Email Link Clicked'
i. 'Last Notable Activity_Email Opened'
j. 'Last Notable Activity_Modified'
k. 'Last Notable Activity_Olark Chat Conversation'
l. 'Last Notable Activity_Page Visited on Website'

MODEL 1

	coef	std err	z	P>|z|	[0.025	0.975]
const	-0.0499	0.089	-0.558	0.577	-0.225	0.125
TotalVisits	1.3774	0.288	4.782	0.000	0.813	1.942
Total Time Spent on Website	4.2241	0.157	26.823	0.000	3.915	4.533
Page Views Per Visit	-3.3828	0.420	-8.061	0.000	-4.205	-2.560
Lead Origin_Lead Add Form	3.5518	0.252	14.068	0.000	3.057	4.047
Lead Source_Direct Traffic	-0.5590	0.078	-7.135	0.000	-0.713	-0.405
Lead Source_Welingak Website	2.4910	1.043	2.388	0.017	0.447	4.535
Do Not Email_Yes	-1.7753	0.177	-10.034	0.000	-2.122	-1.429
Last Activity_Olark Chat Conversation	-0.9804	0.191	-5.144	0.000	-1.354	-0.607
What is your current occupation_Housewife	22.2913	1.76e+04	0.001	0.999	-3.46e+04	3.46e+04
What is your current occupation_Working Professional	2.7197	0.189	14.384	0.000	2.349	3.090
Last Notable Activity_Email Link Clicked	-1.9390	0.268	-7.223	0.000	-2.465	-1.413
Last Notable Activity_Email Opened	-1.4133	0.089	-15.922	0.000	-1.587	-1.239
Last Notable Activity_Modified	-1.9367	0.097	-19.900	0.000	-2.127	-1.746
Last Notable Activity_Olark Chat Conversation	-1.6887	0.373	-4.530	0.000	-2.419	-0.958
Last Notable Activity_Page Visited on Website	-2.0669	0.212	-9.735	0.000	-2.483	-1.651

Model 1 which is build from features selected by RFE has high “p” value for “Housewife” occupation. Hence, this variable is dropped and model is rebuilt.

Generalized Linear Model Regression Results

Dep. Variable:	Converted	No. Observations:	6293
Model:	GLM	Df Residuals:	6277
Model Family:	Binomial	Df Model:	15
Link Function:	Logit	Scale:	1.0000
Method:	IRLS	Log-Likelihood:	-2626.1
Date:	Fri, 03 Mar 2023	Deviance:	5252.2
Time:	23:01:54	Pearson chi2:	6.39e+03
No. Iterations:	21	Pseudo R-squ. (CS):	0.3882
Covariance Type:	nonrobust		

	Features	VIF
2	Page Views Per Visit	5.26
0	TotalVisits	5.25
1	Total Time Spent on Website	2.02
12	Last Notable Activity_Modified	1.90
7	Last Activity_Olark Chat Conversation	1.71
11	Last Notable Activity_Email Opened	1.52
3	Lead Origin_Lead Add Form	1.50
4	Lead Source_Direct Traffic	1.46
5	Lead Source_Welingak Website	1.34
13	Last Notable Activity_Olark Chat Conversation	1.34
9	What is your current occupation_Working Profes...	1.17
14	Last Notable Activity_Page Visited on Website	1.16
6	Do Not Email_Yes	1.14
10	Last Notable Activity_Email Link Clicked	1.02
8	What is your current occupation_Housewife	1.01

MODEL 2

	coef	std err	z	P>|z|	[0.025	0.975]
const	-0.0509	0.089	-0.569	0.569	-0.226	0.124
TotalVisits	1.3677	0.288	4.749	0.000	0.803	1.932
Total Time Spent on Website	4.2290	0.157	26.861	0.000	3.920	4.538
Page Views Per Visit	-3.3734	0.419	-8.042	0.000	-4.196	-2.551
Lead Origin_Lead Add Form	3.5766	0.252	14.188	0.000	3.083	4.071
Lead Source_Direct Traffic	-0.5582	0.078	-7.131	0.000	-0.712	-0.405
Lead Source_Welingak Website	2.4649	1.043	2.363	0.018	0.421	4.509
Do Not Email_Yes	-1.7777	0.177	-10.045	0.000	-2.125	-1.431
Last Activity_Olark Chat Conversation	-0.9835	0.191	-5.161	0.000	-1.357	-0.610
What is your current occupation_Working Professional	2.7168	0.189	14.367	0.000	2.346	3.087
Last Notable Activity_Email Link Clicked	-1.9278	0.267	-7.227	0.000	-2.451	-1.405
Last Notable Activity_Email Opened	-1.4102	0.089	-15.892	0.000	-1.584	-1.236
Last Notable Activity_Modified	-1.9330	0.097	-19.873	0.000	-2.124	-1.742
Last Notable Activity_Olark Chat Conversation	-1.6847	0.373	-4.519	0.000	-2.415	-0.954
Last Notable Activity_Page Visited on Website	-2.0652	0.212	-9.726	0.000	-2.481	-1.649

Model 2 has “p” values below 0.05 but VIF values above 5. So “Page Views Per Visit” variable is dropped and model is rebuilt.

Generalized Linear Model Regression Results

Dep. Variable:	Converted	No. Observations:	6293
Model:	GLM	Df Residuals:	6278
Model Family:	Binomial	Df Model:	14
Link Function:	Logit	Scale:	1.0000
Method:	IRLS	Log-Likelihood:	-2629.4
Date:	Fri, 03 Mar 2023	Deviance:	5258.7
Time:	23:02:04	Pearson chi2:	6.41e+03
No. Iterations:	7	Pseudo R-squ. (CS):	0.3876
Covariance Type:	nonrobust		

	Features	VIF
2	Page Views Per Visit	5.26
0	TotalVisits	5.25
1	Total Time Spent on Website	2.02
11	Last Notable Activity_Modified	1.90
7	Last Activity_Olark Chat Conversation	1.71
10	Last Notable Activity_Email Opened	1.52
3	Lead Origin_Lead Add Form	1.49
4	Lead Source_Direct Traffic	1.46
5	Lead Source_Welingak Website	1.34
12	Last Notable Activity_Olark Chat Conversation	1.34
8	What is your current occupation_Working Profes...	1.17
13	Last Notable Activity_Page Visited on Website	1.16
6	Do Not Email_Yes	1.14
9	Last Notable Activity_Email Link Clicked	1.02

MODEL 3

	coef	std err	z	P>|z|	[0.025	0.975]
const	-0.2713	0.085	-3.186	0.001	-0.438	-0.104
TotalVisits	-0.2148	0.217	-0.992	0.321	-0.639	0.210
Total Time Spent on Website	4.0568	0.154	26.368	0.000	3.755	4.358
Lead Origin_Lead Add Form	3.7464	0.251	14.906	0.000	3.254	4.239
Lead Source_Direct Traffic	-0.5699	0.078	-7.347	0.000	-0.722	-0.418
Lead Source_Welingak Website	2.4634	1.043	2.362	0.018	0.419	4.507
Do Not Email_Yes	-1.7748	0.175	-10.154	0.000	-2.117	-1.432
Last Activity_Olark Chat Conversation	-0.8634	0.190	-4.548	0.000	-1.236	-0.491
What is your current occupation_Working Professional	2.6939	0.188	14.317	0.000	2.325	3.063
Last Notable Activity_Email Link Clicked	-1.8062	0.265	-6.819	0.000	-2.325	-1.287
Last Notable Activity_Email Opened	-1.3495	0.088	-15.399	0.000	-1.521	-1.178
Last Notable Activity_Modified	-1.8865	0.096	-19.597	0.000	-2.075	-1.698
Last Notable Activity_Olark Chat Conversation	-1.5240	0.365	-4.179	0.000	-2.239	-0.809
Last Notable Activity_Page Visited on Website	-1.7007	0.202	-8.421	0.000	-2.097	-1.305

Model 3 has VIF values below 5 but “p” value for variable “TotalVisits” suddenly increased making it insignificant. So “TotalVisits” variable is dropped and model is rebuilt.

Generalized Linear Model Regression Results

Dep. Variable:	Converted	No. Observations:	6293
Model:	GLM	Df Residuals:	6279
Model Family:	Binomial	Df Model:	13
Link Function:	Logit	Scale:	1.0000
Method:	IRLS	Log-Likelihood:	-2663.0
Date:	Fri, 03 Mar 2023	Deviance:	5326.0
Time:	23:02:07	Pearson chi2:	6.36e+03
No. Iterations:	7	Pseudo R-squ. (CS):	0.3810
Covariance Type:	nonrobust		

	Features	VIF
0	TotalVisits	2.37
1	Total Time Spent on Website	1.96
10	Last Notable Activity_Modified	1.86
6	Last Activity_Olark Chat Conversation	1.70
9	Last Notable Activity_Email Opened	1.50
2	Lead Origin_Lead Add Form	1.49
3	Lead Source_Direct Traffic	1.44
4	Lead Source_Welingak Website	1.34
11	Last Notable Activity_Olark Chat Conversation	1.34
7	What is your current occupation_Working Profes...	1.17
12	Last Notable Activity_Page Visited on Website	1.14
5	Do Not Email_Yes	1.13
8	Last Notable Activity_Email Link Clicked	1.02

MODEL 4 (FINAL MODEL)

	coef	std err	z	P>|z|	[0.025	0.975]
const	-0.3078	0.077	-4.005	0.000	-0.458	-0.157
Total Time Spent on Website	4.0171	0.148	27.096	0.000	3.727	4.308
Lead Origin_Lead Add Form	3.7810	0.249	15.190	0.000	3.293	4.269
Lead Source_Direct Traffic	-0.5694	0.077	-7.347	0.000	-0.721	-0.417
Lead Source_Welingak Website	2.4632	1.043	2.362	0.018	0.419	4.507
Do Not Email_Yes	-1.7703	0.175	-10.135	0.000	-2.113	-1.428
Last Activity_Olark Chat Conversation	-0.8498	0.189	-4.485	0.000	-1.221	-0.478
What is your current occupation_Working Professional	2.6943	0.188	14.313	0.000	2.325	3.063
Last Notable Activity_Email Link Clicked	-1.8043	0.265	-6.810	0.000	-2.324	-1.285
Last Notable Activity_Email Opened	-1.3518	0.088	-15.429	0.000	-1.524	-1.180
Last Notable Activity_Modified	-1.8822	0.096	-19.582	0.000	-2.071	-1.694
Last Notable Activity_Olark Chat Conversation	-1.5347	0.365	-4.201	0.000	-2.251	-0.819
Last Notable Activity_Page Visited on Website	-1.7355	0.199	-8.727	0.000	-2.125	-1.346

Model 4 has “p” values below 0.05 and VIF values below 5. Thus, this model is considered as final model

Generalized Linear Model Regression Results

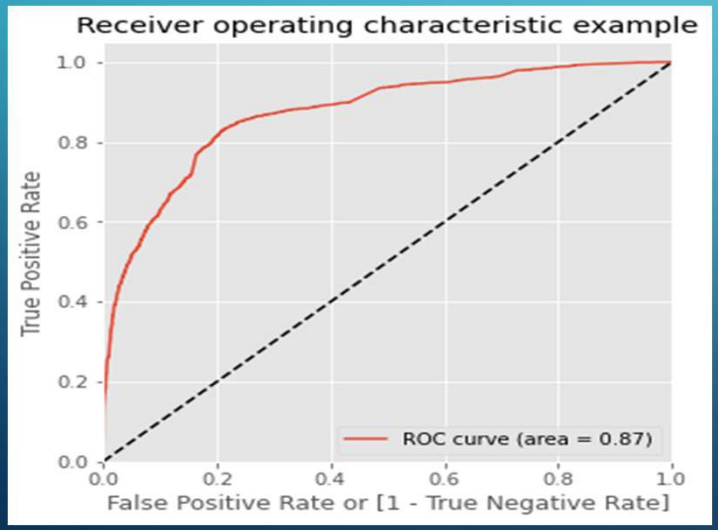
Dep. Variable:	Converted	No. Observations:	6293
Model:	GLM	Df Residuals:	6280
Model Family:	Binomial	Df Model:	12
Link Function:	Logit	Scale:	1.0000
Method:	IRLS	Log-Likelihood:	-2663.5
Date:	Fri, 03 Mar 2023	Deviance:	5327.0
Time:	23:02:17	Pearson chi2:	6.36e+03
No. Iterations:	7	Pseudo R-squ. (CS):	0.3809
Covariance Type:	nonrobust		

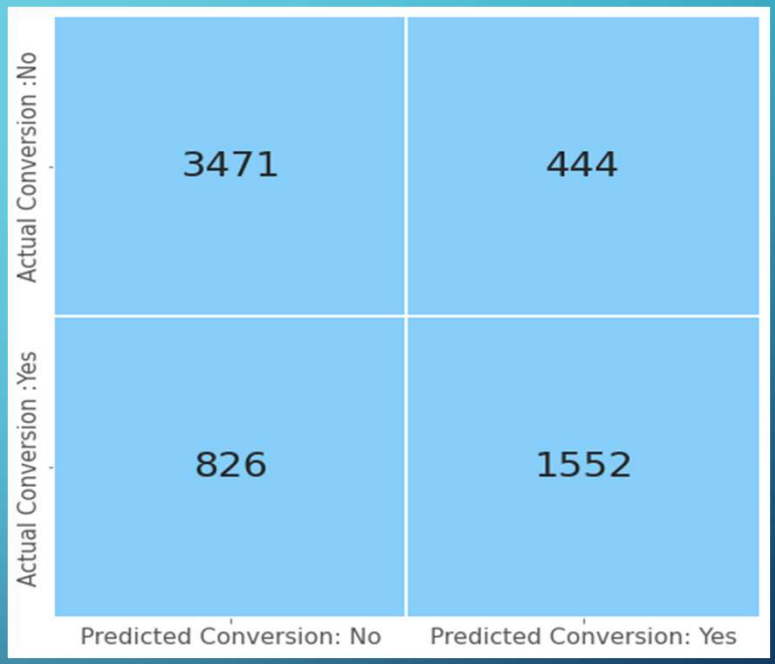
	Features	VIF
9	Last Notable Activity_Modified	1.73
5	Last Activity_Olark Chat Conversation	1.70
0	Total Time Spent on Website	1.57
1	Lead Origin_Lead Add Form	1.48
2	Lead Source_Direct Traffic	1.43
3	Lead Source_Welingak Website	1.34
10	Last Notable Activity_Olark Chat Conversation	1.33
8	Last Notable Activity_Email Opened	1.32
6	What is your current occupation_Working Profes...	1.17
4	Do Not Email_Yes	1.13
11	Last Notable Activity_Page Visited on Website	1.05
7	Last Notable Activity_Email Link Clicked	1.01

MODEL EVALUATION BASED ON MANUAL CUTOFF OF 0.5

Below model metrices are obtained using manual cutoff of 0.5:

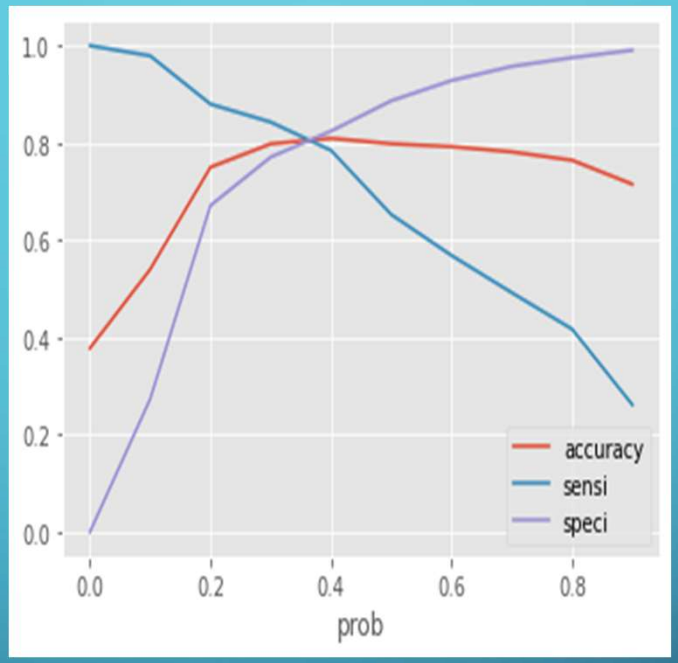
- Accuracy = 79.81 %
- Sensitivity = 65.26 %
- Specificity = 88.65 %
- ROC curve Area = 87.00 %

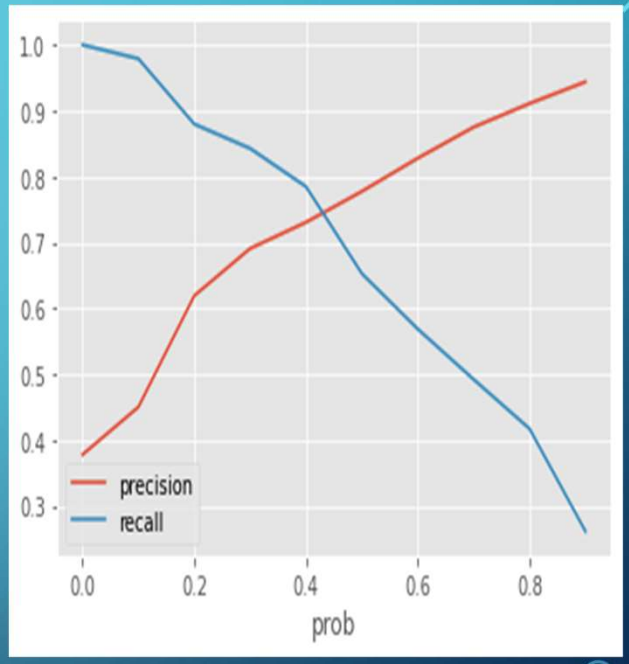




MODEL EVALUATION AT DIFFERENT CUTOFFS

	prob	accuracy	sensi	speci	precision	recall
0.0	0.0	0.377880	1.000000	0.000000	0.377880	1.000000
0.1	0.1	0.540124	0.978974	0.273563	0.450116	0.978974
0.2	0.2	0.750040	0.879731	0.671264	0.619118	0.879731
0.3	0.3	0.798030	0.842725	0.770881	0.690796	0.842725
0.4	0.4	0.809312	0.784693	0.824266	0.730619	0.784693
0.5	0.5	0.798188	0.652649	0.886590	0.777555	0.652649
0.6	0.6	0.792309	0.568545	0.928225	0.827924	0.568545
0.7	0.7	0.781503	0.492010	0.957344	0.875093	0.492010
0.8	0.8	0.764341	0.417157	0.975223	0.910927	0.417157
0.9	0.9	0.715080	0.261564	0.990549	0.943854	0.261564





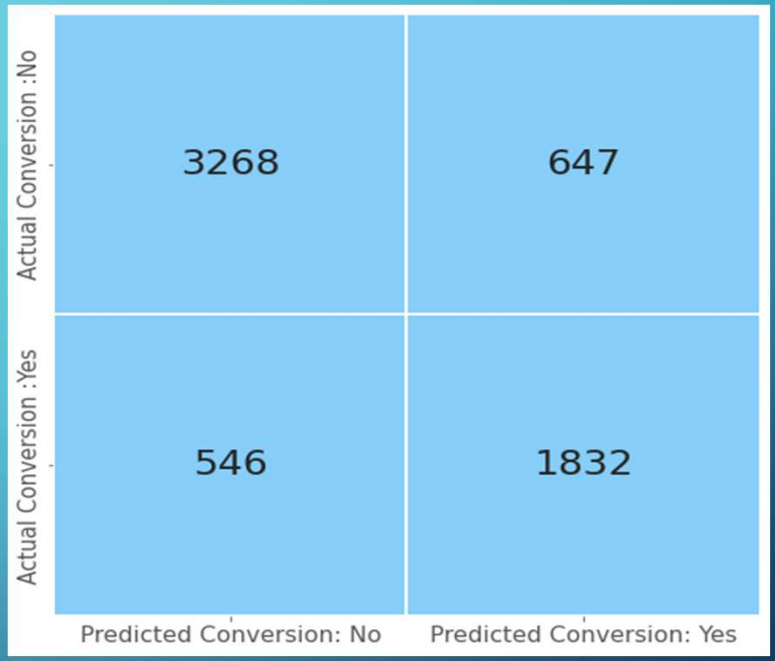
0.36 is obtained as optimal cutoff from Sensitivity-Specificity view.
While 0.44 optimal cutoff is obtained from Precision-Recall view.
Since, CEO has aimed for 80% conversion rate which means 80% of precision, we will select 0.44 as optimal cutoff.

MODEL EVALUATION BASED ON OPTIMAL CUTOFF OF 0.44

Below model metrices are obtained using optimal cutoff of 0.44:

A) On Train Set :
- Accuracy = 81.04 %
- Precision = 73.90 %
- Recall = 77.03 %

B) On Test Set :
- Accuracy = 79.05 %
- Precision = 73.36 %
- Recall = 70.12 %



BUSINESS INSIGHTS

- Target variable has Data imbalance. 62% peoples are not converted while 38% people are converted.
- Most of the lead sources are either from Google or they are Direct Traffic. While leads from Google and Organic Search have high conversion rate compared to others.
- People who are Unemployed have shown most interest in the course. While working professional people have shown a greater number of conversions followed by unemployed people. X-education should target Working professional people more for higher lead conversion.
- People who have expectation of “Better Career Prospects” have largely shown interest in this course with almost 50% conversion. X-education should focus on providing more career opportunities along with course to get more lead conversions.
- People who have sent the SMS to X-education have shown huge conversion. X-education should focus on LEAD who have contacted X-Education through similar SMS.