LEAD SCORE CASE STUDY

Members:

1. NISHCHAY YADAV
2. AKSHAY SIRASWAR

PROBLEM:

- To build a model wherein the customers with higher lead score have a higher conversion chance and the customers with lower lead score have a lower conversion chance.
- To estimate the target lead conversion rate to be around 80%.

Approach of analysis:

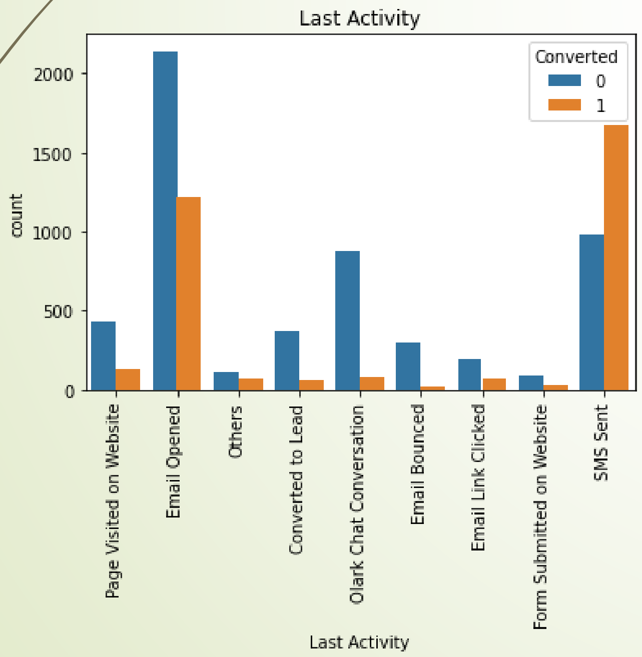
1. Read and understand the data
2. Clean the data
3. EDA
4. Prepare the data for Model Building
5. Model Building
6. Model Evaluation
7. Optimizing cutoff (ROC curve)
8. Making Predictions on the Test Set
9. Precision-Recall
10. Prediction on the Test Set

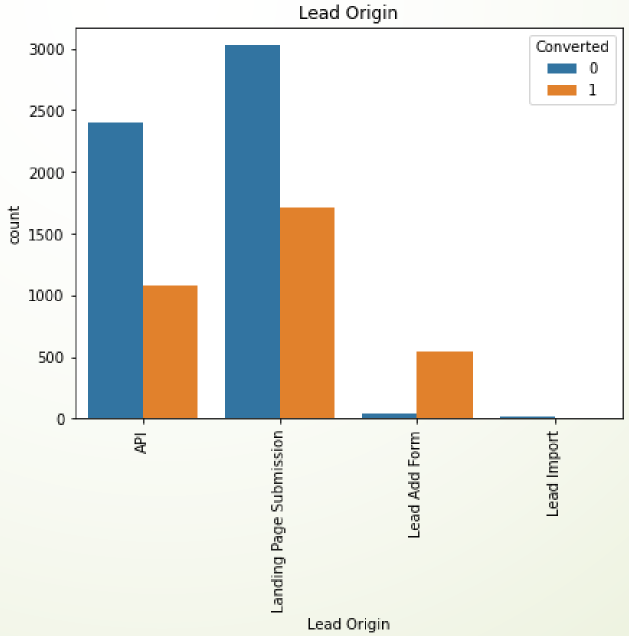
Reading and cleaning the Dataset:

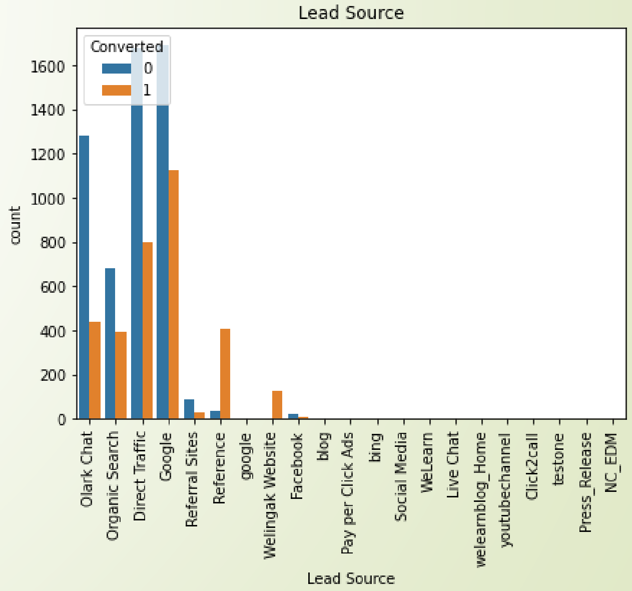
- Dataset contain 9240 rows and 37 columns.
- There are many columns need to be dropped based on the high null values like Asymmetrique Activity Index, Asymmetrique Profile Index, Asymmetrique Activity Score, Asymmetrique Profile Score
- Removed some of the redundant columns like How did you hear about X Education, Magazine".
- Certain columns' null values are imputed after categorizing them as these columns can not be dropped.
- Rows of the low null values columns are dropped.

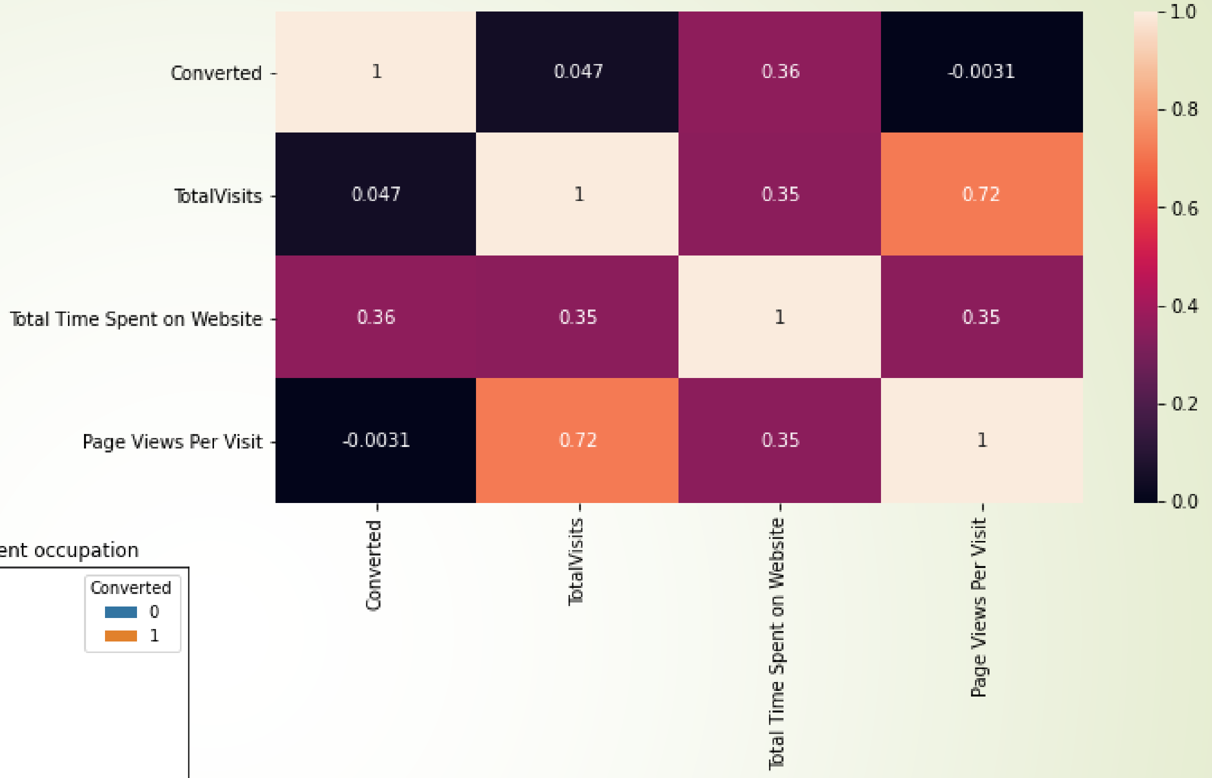
EDA

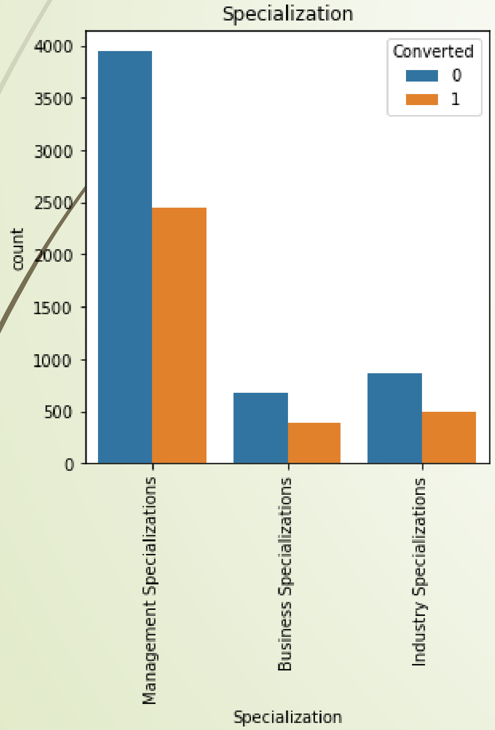
- We performed some EDA techniques on our dataset to get a better understanding of the variables.
- We drew heat map for checking the multicollinearity.

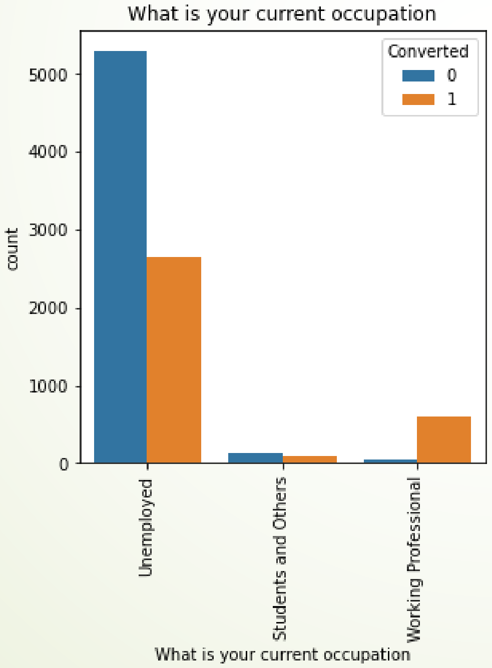







Converted
TotalVisits
Total Time Spent on Website
Page Views Per Visit
1
0.047
0.36
-0.0031
0.047
1
0.35
0.72
0.36
0.35
1
0.35
-0.0031
0.72
0.35
1
Converted
TotalVisits
Total Time Spent on Website
Page Views Per Visit
1.0
0.8
0.6
0.4
0.2
0.0


Specialization
Converted
0
1
count
4000
3500
3000
2500
2000
1500
1000
500
0
Management Specializations
Business Specializations
Industry Specializations
Specialization


What is your current occupation
Converted
0
1
count
5000
4000
3000
2000
1000
0
Unemployed
Students and Others
Working Professional
What is your current occupation

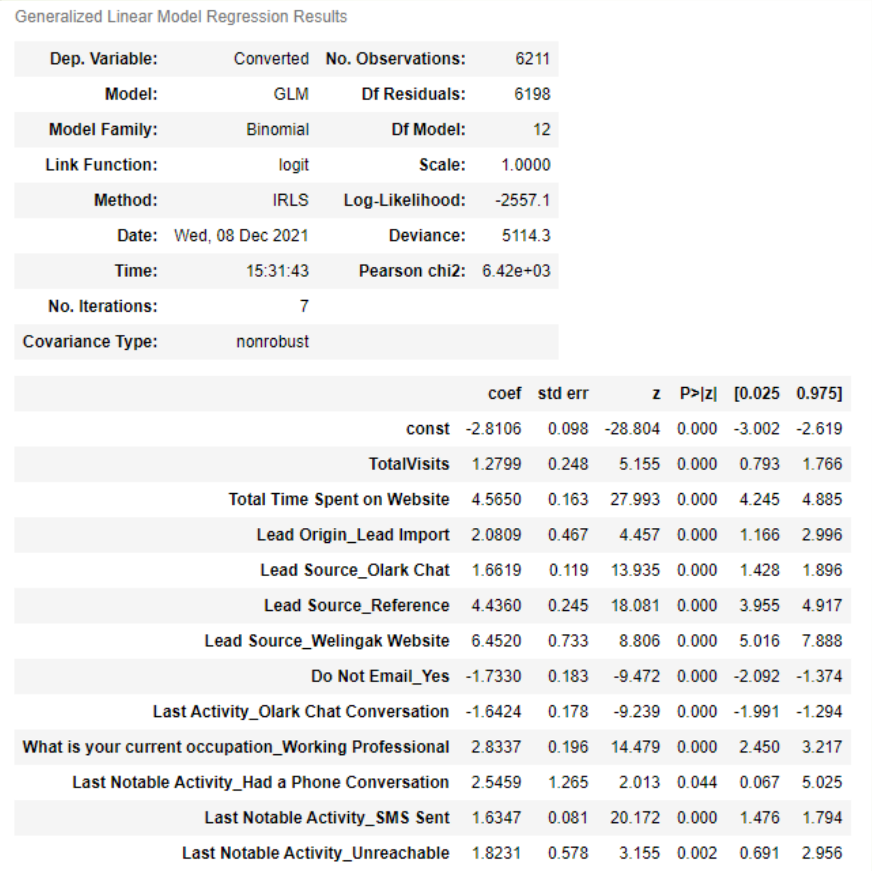
Preparing the data for Model Building

- After dummy creation we proceed with our next step of analysis.
- a) We split the dataset into train and test set.
- b) Standardization is required in order to keep all the variables in same scale which will help us in computation in more efficient way.
- c) Checked the correlation of the dataset through heat map where using RFE approach we further dropped the highly correlated features.

Model Building :

- After splitting the Data into Training and Testing Sets we used RFE feature selection technique to eliminate the insignificant features available in the data.
- We ran RFE count with 15 variables as the significant variables.
- After that we started building the models by removing the variable whose p-value is greater than 0.05 and VIF value is greater than 5.
- With different models we were encountering with different p-values and VIF values.
- On getting the P-values and VIF values under the range we stopped our model building and chose it our final model.

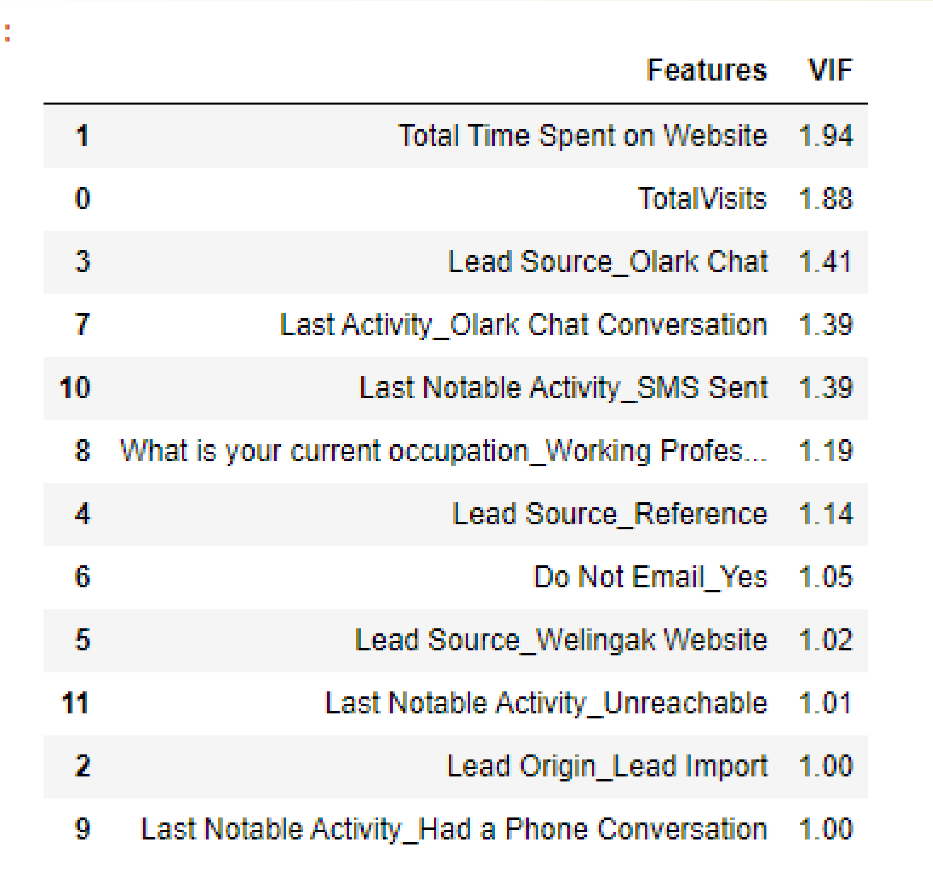
Final Model statistics:



Generalized Linear Model Regression Results

Dep. Variable:	Converted	No. Observations:	6211
Model:	GLM	Df Residuals:	6198
Model Family:	Binomial	Df Model:	12
Link Function:	logit	Scale:	1.0000
Method:	IRLS	Log-Likelihood:	-2557.1
Date:	Wed, 08 Dec 2021	Deviance:	5114.3
Time:	15:31:43	Pearson chi2:	6.42e+03
No. Iterations:	7		
Covariance Type:	nonrobust		

	coef	std err	z	P>|z|	[0.025	0.975]
const	-2.8106	0.098	-28.804	0.000	-3.002	-2.619
TotalVisits	1.2799	0.248	5.155	0.000	0.793	1.766
Total Time Spent on Website	4.5650	0.163	27.993	0.000	4.245	4.885
Lead Origin_Lead Import	2.0809	0.467	4.457	0.000	1.166	2.996
Lead Source_Olark Chat	1.6619	0.119	13.935	0.000	1.428	1.896
Lead Source_Reference	4.4360	0.245	18.081	0.000	3.955	4.917
Lead Source_Welingak Website	6.4520	0.733	8.806	0.000	5.016	7.888
Do Not Email_Yes	-1.7330	0.183	-9.472	0.000	-2.092	-1.374
Last Activity_Olark Chat Conversation	-1.6424	0.178	-9.239	0.000	-1.991	-1.294
What is your current occupation_Working Professional	2.8337	0.196	14.479	0.000	2.450	3.217
Last Notable Activity_Had a Phone Conversation	2.5459	1.265	2.013	0.044	0.067	5.025
Last Notable Activity_SMS Sent	1.6347	0.081	20.172	0.000	1.476	1.794
Last Notable Activity_Unreachable	1.8231	0.578	3.155	0.002	0.691	2.956



	Features	VIF
1	Total Time Spent on Website	1.94
0	TotalVisits	1.88
3	Lead Source_Olark Chat	1.41
7	Last Activity_Olark Chat Conversation	1.39
10	Last Notable Activity_SMS Sent	1.39
8	What is your current occupation_Working Profes...	1.19
4	Lead Source_Reference	1.14
6	Do Not Email_Yes	1.05
5	Lead Source_Welingak Website	1.02
11	Last Notable Activity_Unreachable	1.01
2	Lead Origin_Lead Import	1.00
9	Last Notable Activity_Had a Phone Conversation	1.00

Model Evaluation:

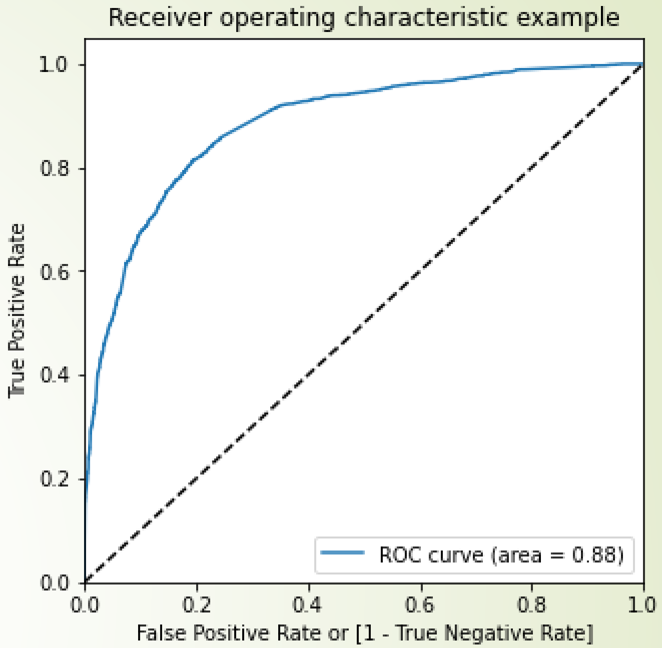
1. After the model building we performed some metrics analysis like accuracy, specificity and sensitivity.

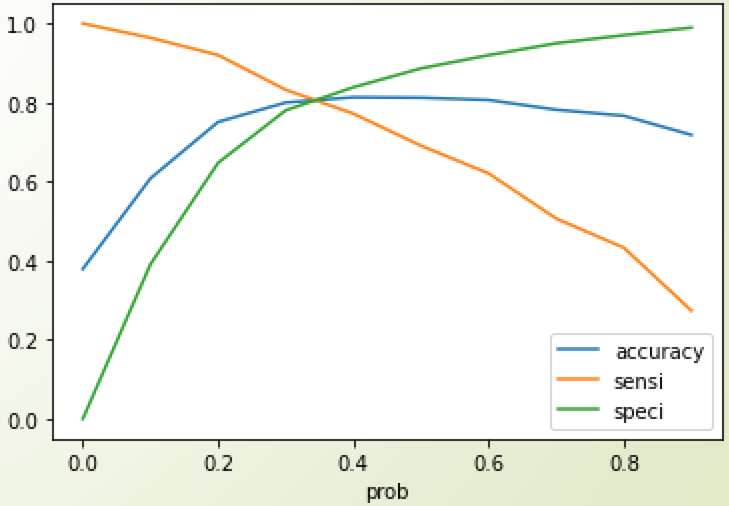
2. After taking random cutoff we drew the ROC curve and it was showing an area of 88% which is good for model.

3. Then we optimized our cutoff and calculated the various metrics like accuracy, sensitivity and specificity.

4. With the current cut off as 0.35 we have accuracy around 81%, sensitivity around 80% and specificity of around 81%.

5. After that we performed business oriented metrics approach by using precision and recall.





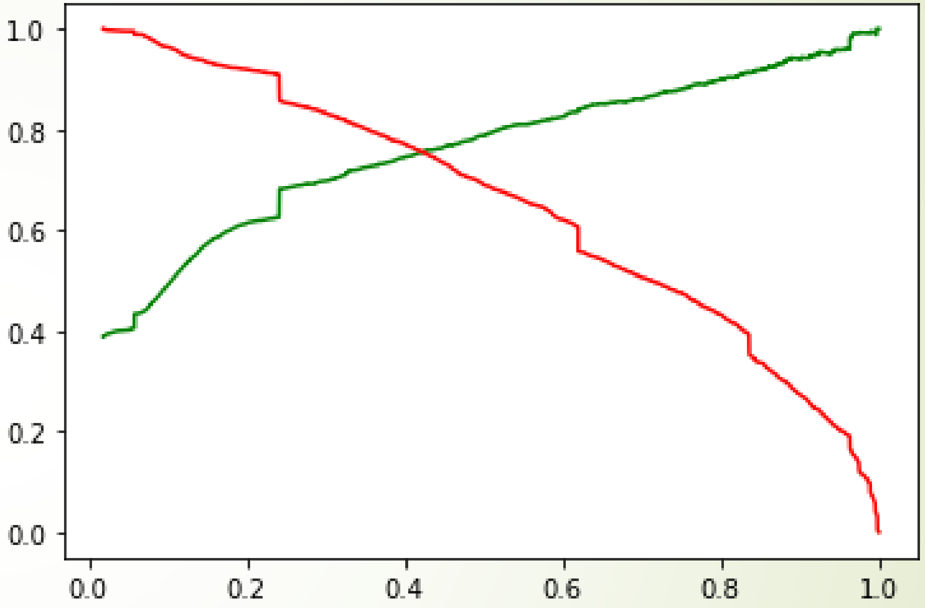
Precision and Recall:

1. For the business perpective we evaluated the precision and Recall metrics.

2. Precision and Recall plays very important role in building the model more business oriented

3. With the current cut off as 0.42 we have Precision around 75% and Recall around 76%.

4. With accuracy(81.4%)and Recall(76%) in acceptable range we can consider our model to be effective to be good for model.



Prediction on test set:

- We followed with the same steps like scaling and evaluating the metrics as we did for the train set.

- After this we did model evaluation i.e. finding the accuracy, precision and recall.

- Our test prediction is having accuracy of 80%, with 72% precision and 74% recall score which shows our model is stable and good for analysis.

Conclusion:

- The Accuracy, Precision and Recall are showing similar scores in test set which is as expected after looking the same in train set evaluation steps.
- Recall metrics further enhance the stability of our model with value 74%.
- Important features responsible for good conversion rate are :

Total Time Spent on Website, TotalVisits, Lead Source as Olark chat, Reference, wellingak website and When their current occupation is as a working professional.

- We should contact the person who spend more time on website.
- Special attention on the people who have lead source as Olark chat, reference, and wellingak website by mentoring them for the course.
- Emphasis should be on working professionals and unemployed people.
- People with last activity as Olark chat conversation can be potential leads.
- X Education can increase the conversion rate by keeping the above variables in mind and can sustain well for future.