LEAD SCORE CASE STUDY

Submitted by:

- Najaf Naqvi
- Aditya B N S
- Pradeep
- Prathap Seesala

PROBLEM STATEMENT AND OBJECTIVES

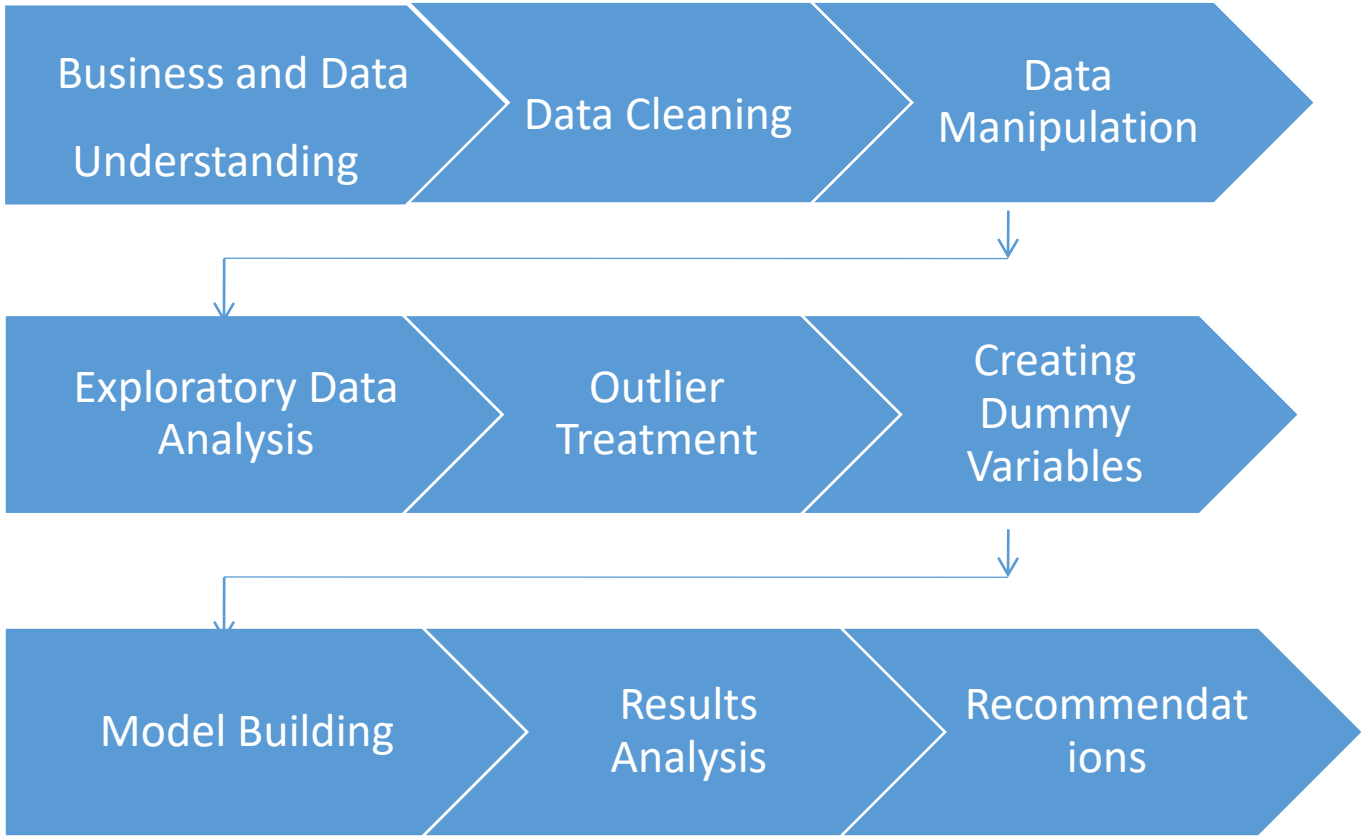
PROBLEM STATEMENT:

- X an Education company sells online courses to industry professionals. Although it gets a lot of leads, its lead conversion rate is very poor.
- To overcome this they want to focus more on communicating with the potential leads rather than making calls to everyone.

OBJECTIVE:

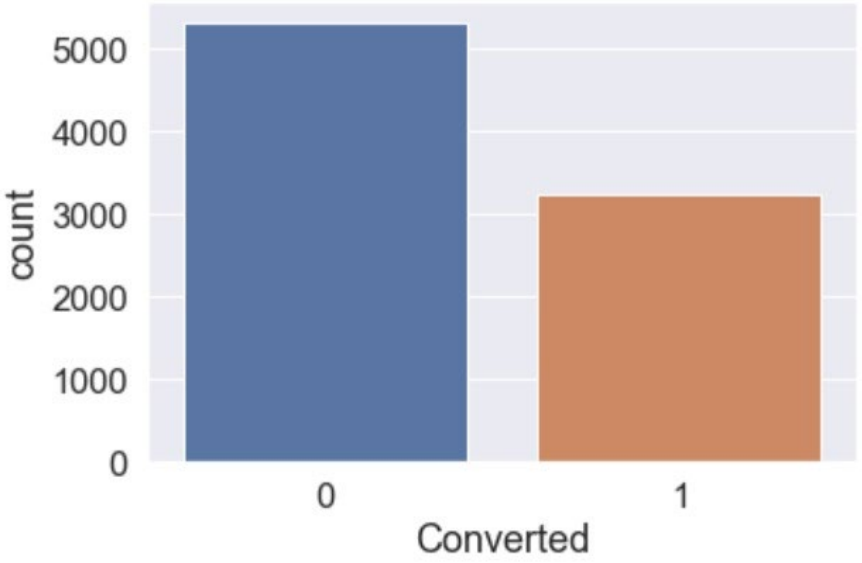
- To help company X in selecting the most promising leads, i.e. the leads that are most likely to convert into paying customers by building a model wherein we assign a lead score to each of the leads such that the customers with higher lead score have a higher conversion chance and the customers with lower lead score have a lower conversion chance.
- The CEO, in particular, has given us ballpark of the target lead conversion rate to be around 80%.

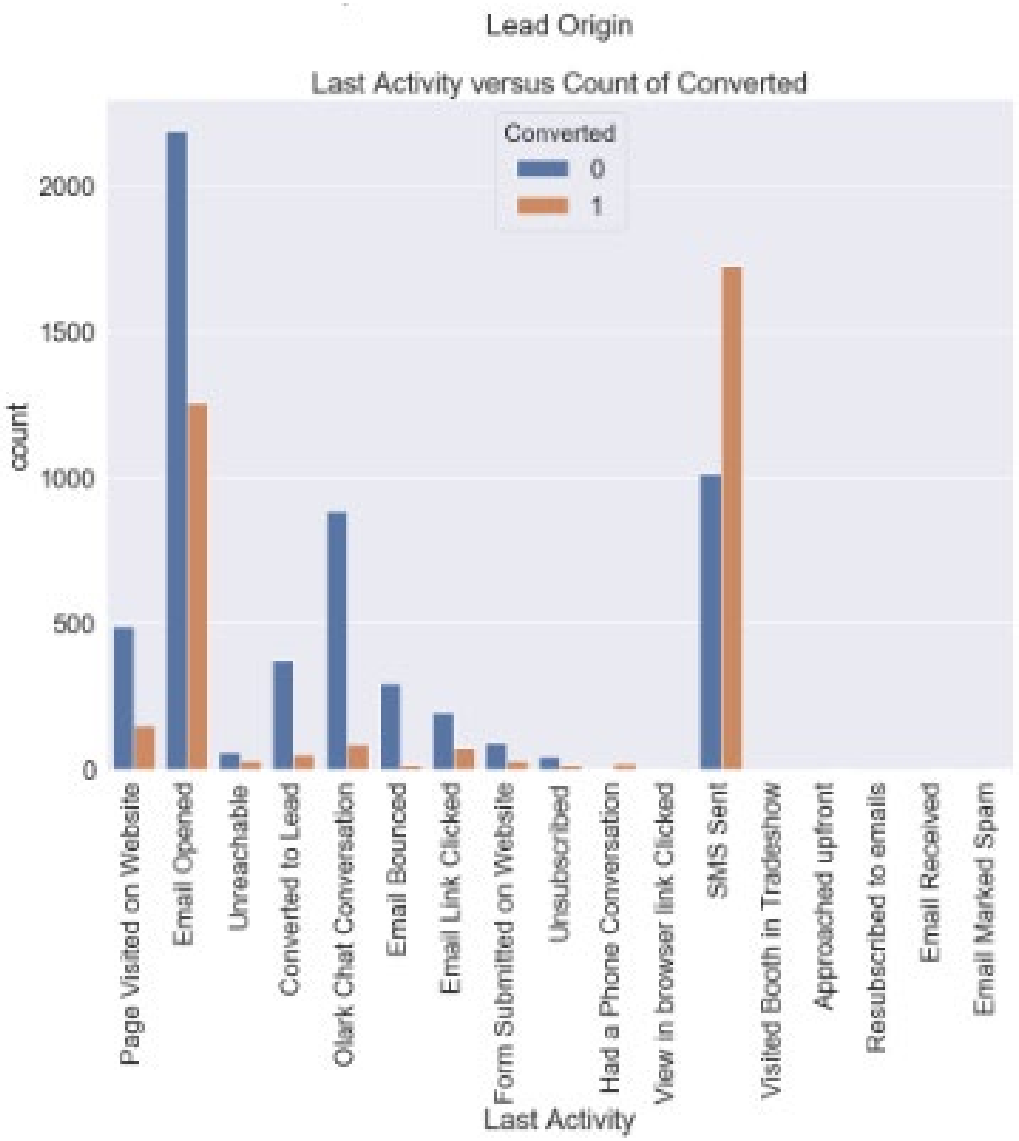
PROBLEM SOLVING METHODOLOGY:

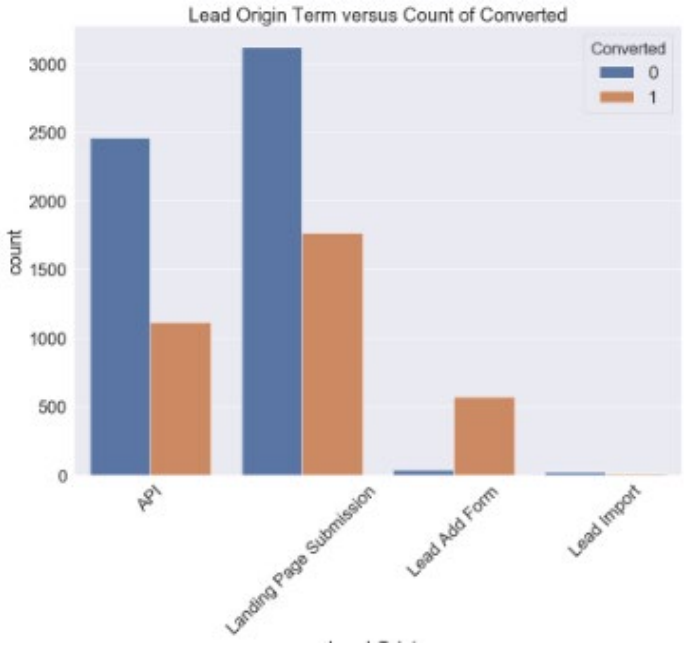


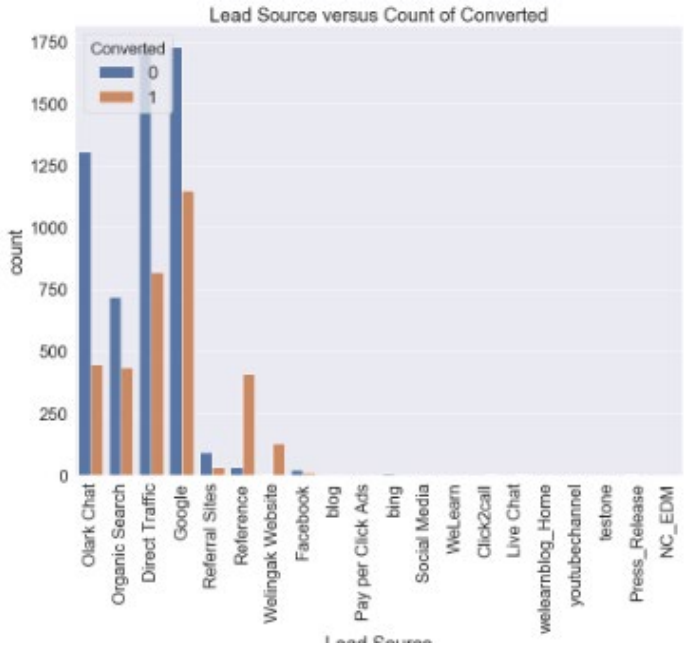
DATA MANIPULATION & EXPLORATORY DATA ANALYSIS

- We decided to remove the columns that have more than 35 % null values excluding 'specialization', 'Asymmetrique Activity Index', 'Asymmetrique Profile Index' as we believed that these will play a significancant role in analysis
- Critical Columns were kept and their null values were treated after dummy creation.
- The leads getting converted is less than ones not getting converted and the he leads for whom the SMS was sent and leads who open the E mail tend to convert more.





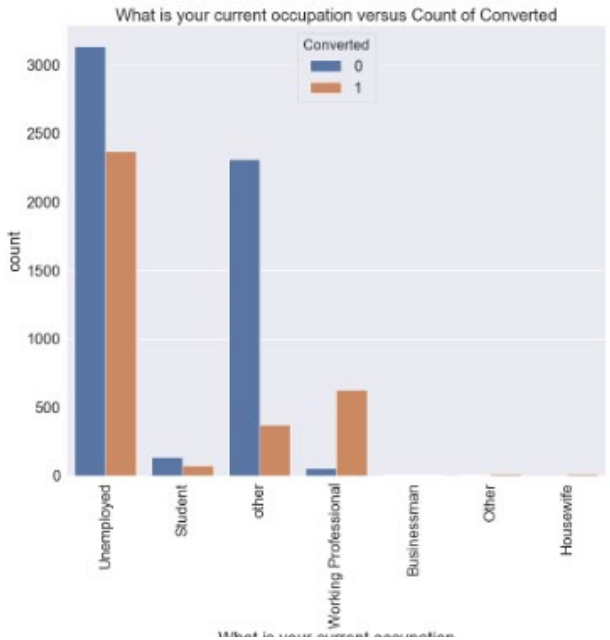


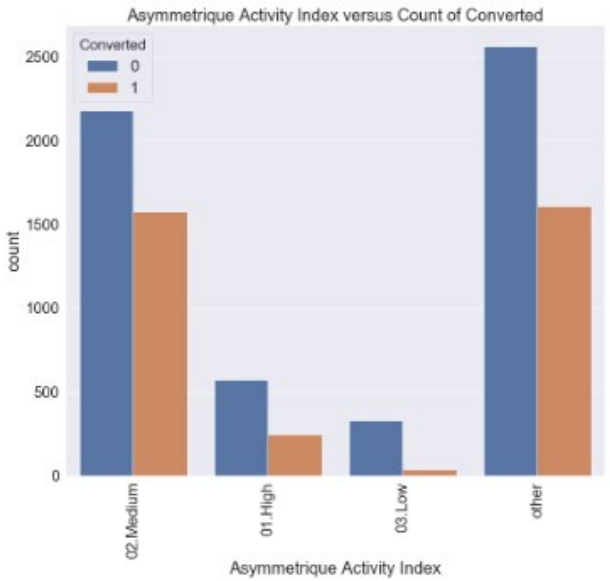


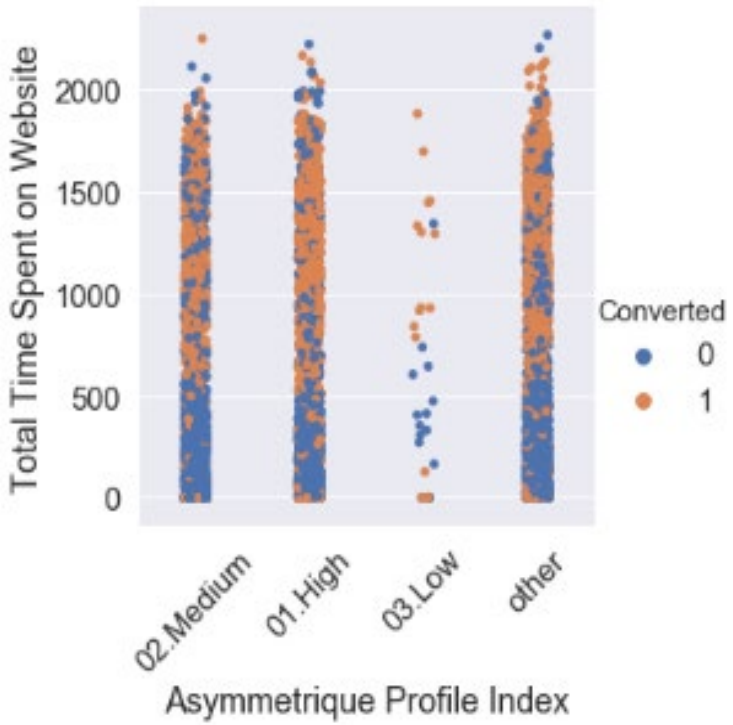
- Leads landing into submission pages and ones from Google and Direct Traffic are converting the most

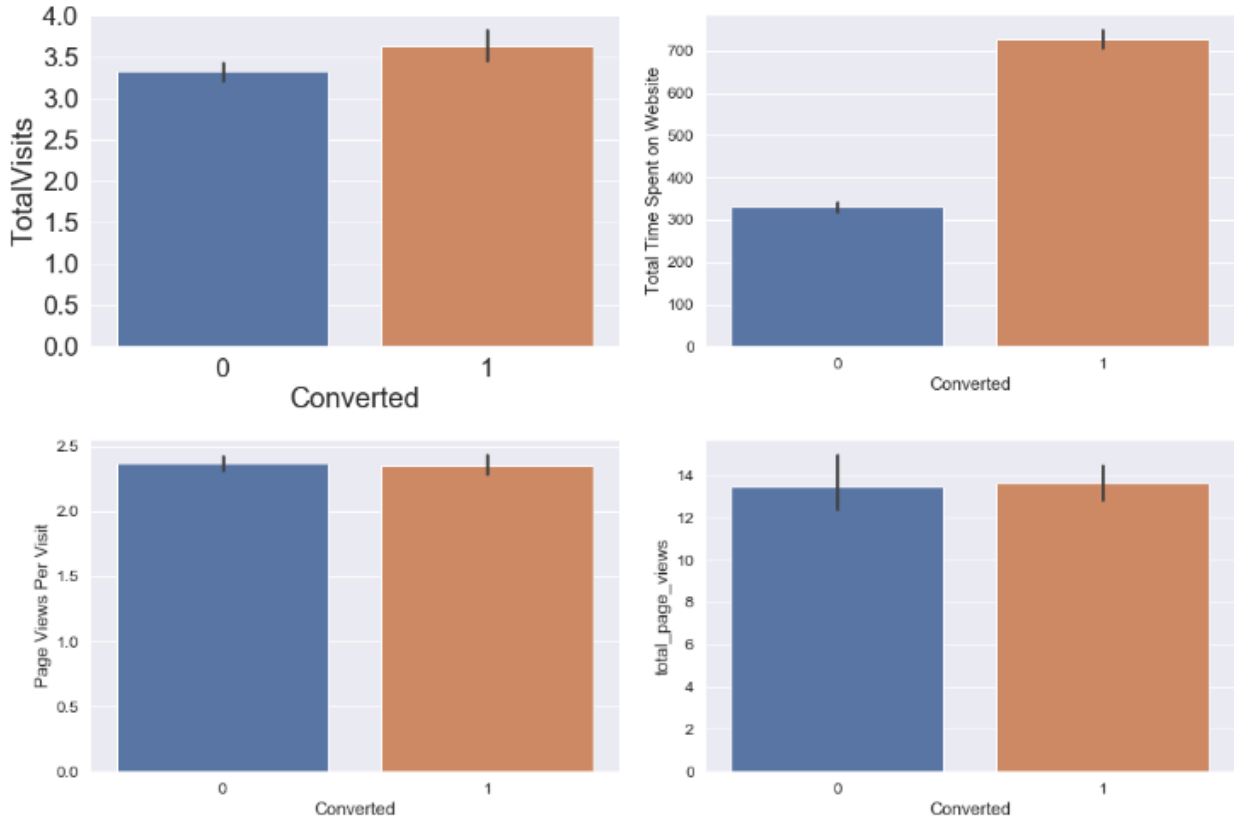
- The leads that are unemployed and medium asymmetric activity indexed tend to convert more

*Note: all the “other“ category is handled in the code





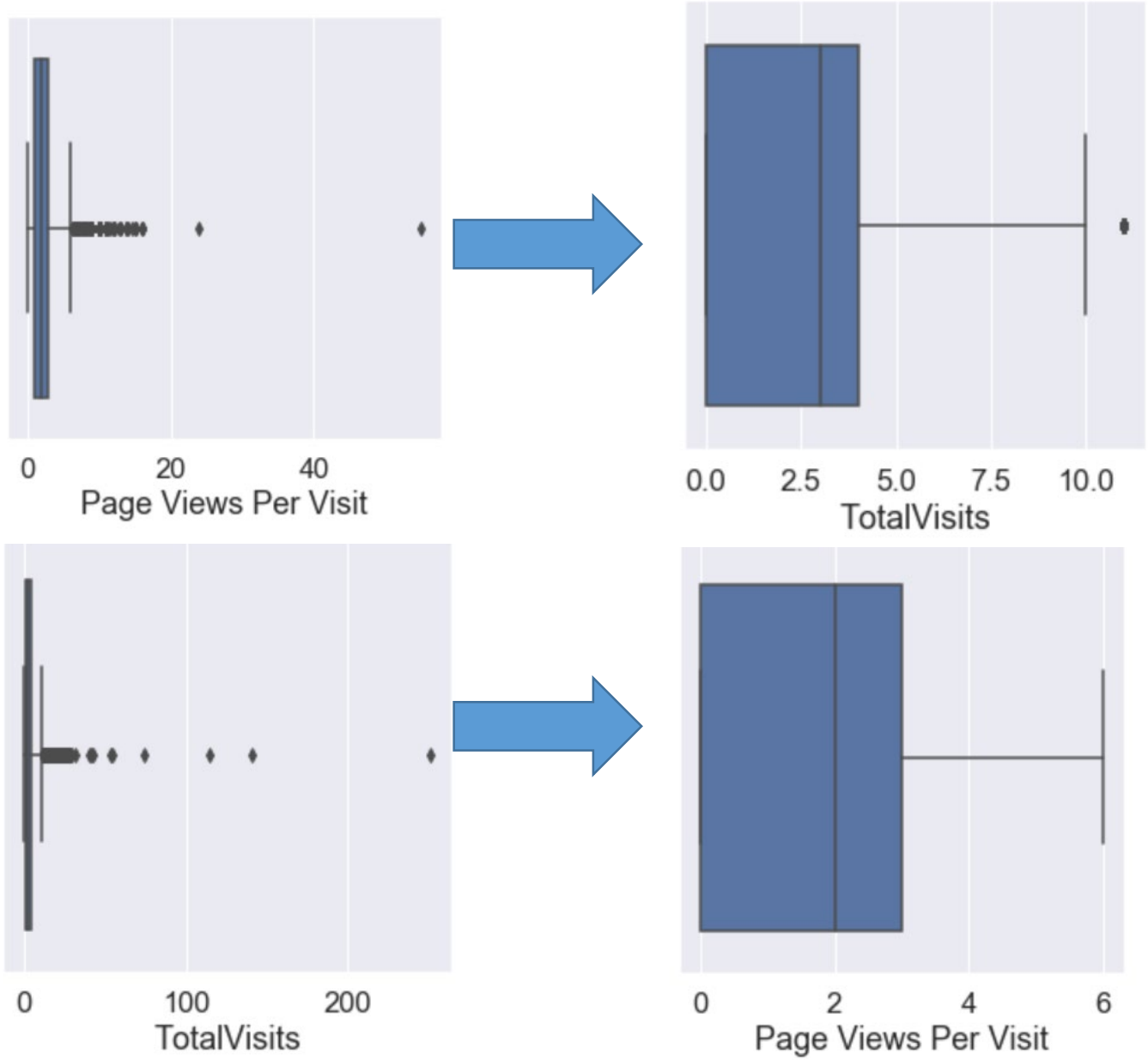




- The leads spending more time in website and more total visits convert more and the ones with high asymmetrique Profile Index convert with ease.

OUTLIER TREATMENT

- Outlier treatment for the numerical variables ‘Page Views Per Visit’ and ‘TotalVisits’.



MODEL EVALUATION

- Model built with 20 variables , using RFE and VIF methods variables have been finetuned.

- Final model has 14 variables with p-value below 0.05 and low VIF.

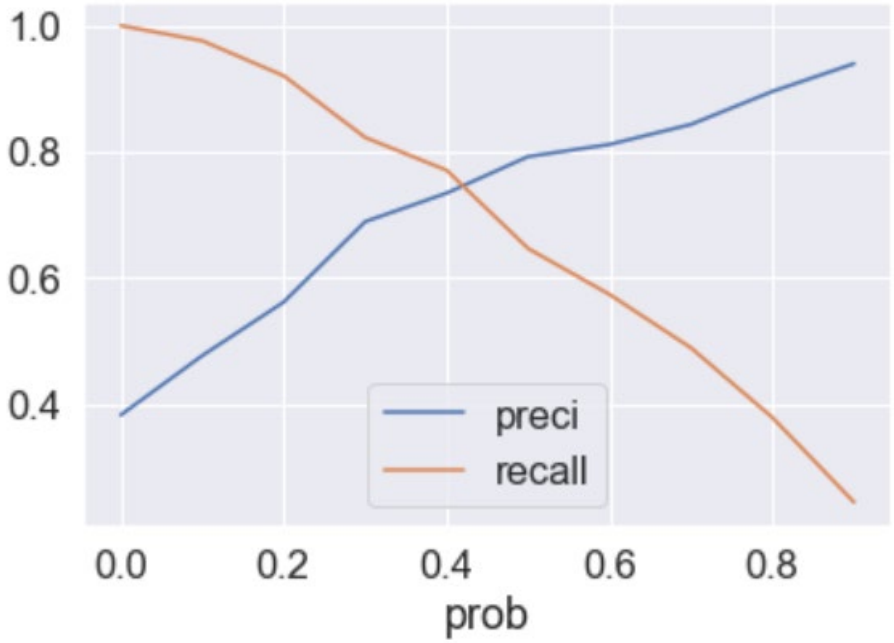
	coef	std err	z	P>|z|	[0.025	0.975]
const	0.5828	0.092	6.328	0.000	0.402	0.763
Lead_Origin_Lead Add Form	3.0620	0.222	13.790	0.000	2.627	3.497
Lead_Source_Direct Traffic	-1.3655	0.114	-11.980	0.000	-1.589	-1.142
Lead_Source_Google	-1.0419	0.109	-9.591	0.000	-1.255	-0.829
Lead_Source_Organic Search	-1.0356	0.136	-7.625	0.000	-1.302	-0.769
Lead_Source_Referral Sites	-1.3337	0.373	-3.575	0.000	-2.065	-0.603
Last_Activity_Converted to Lead	-2.0296	0.220	-9.236	0.000	-2.460	-1.599
Last_Activity_Email Bounced	-2.9327	0.377	-7.789	0.000	-3.671	-2.195
Last_Activity_Olark Chat Conversation	-2.4282	0.171	-14.219	0.000	-2.763	-2.094
Last_Notable_Activity_Email Link Clicked	-1.3791	0.285	-4.844	0.000	-1.937	-0.821
Last_Notable_Activity_Email Opened	-0.6436	0.076	-8.508	0.000	-0.792	-0.495
Last_Notable_Activity_Page Visited on Website	-1.0689	0.213	-5.027	0.000	-1.486	-0.652
current_occupation_Working Professional	2.8557	0.193	14.772	0.000	2.477	3.235
Asymmetrique_Activity_Index_03.Low	-1.9420	0.265	-7.324	0.000	-2.462	-1.422
Total Time Spent on Website	1.0942	0.040	27.052	0.000	1.015	1.173

METRICS INTERPRETATION

- Metrics for the logistic model (Accuracy , Sensitivity , Specificity , Precision and Recall) have been compared.

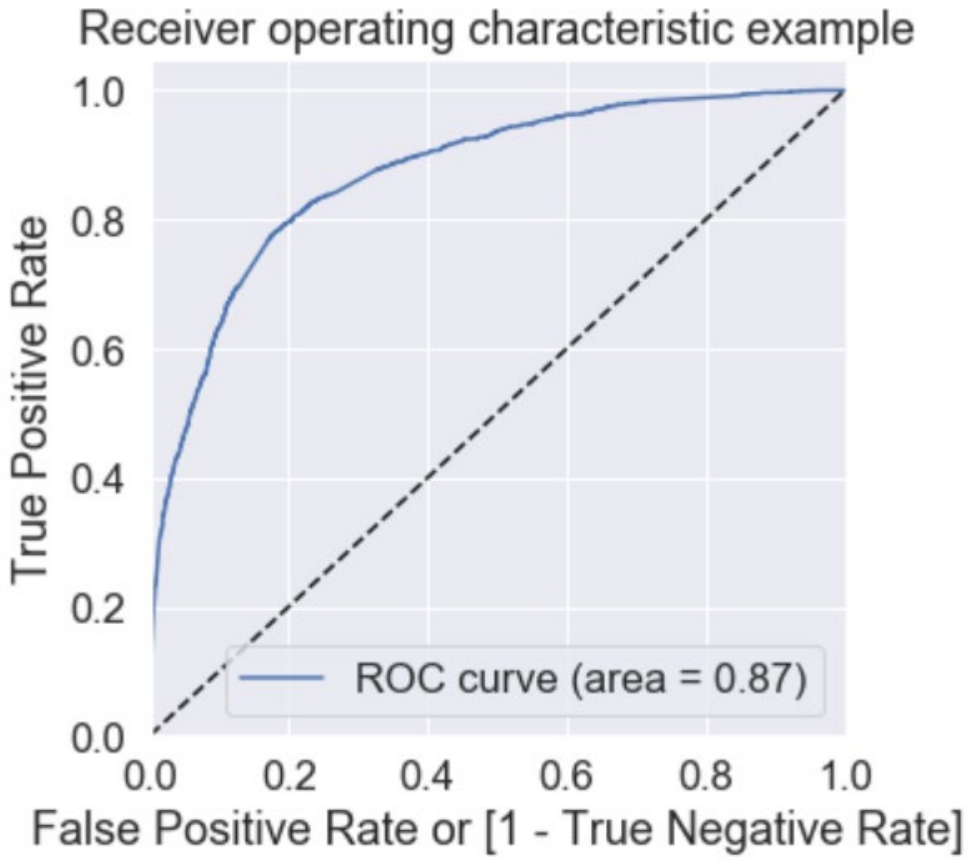
- Desired precision value is **80%** , which is achievable with cut off probability – **0.6**

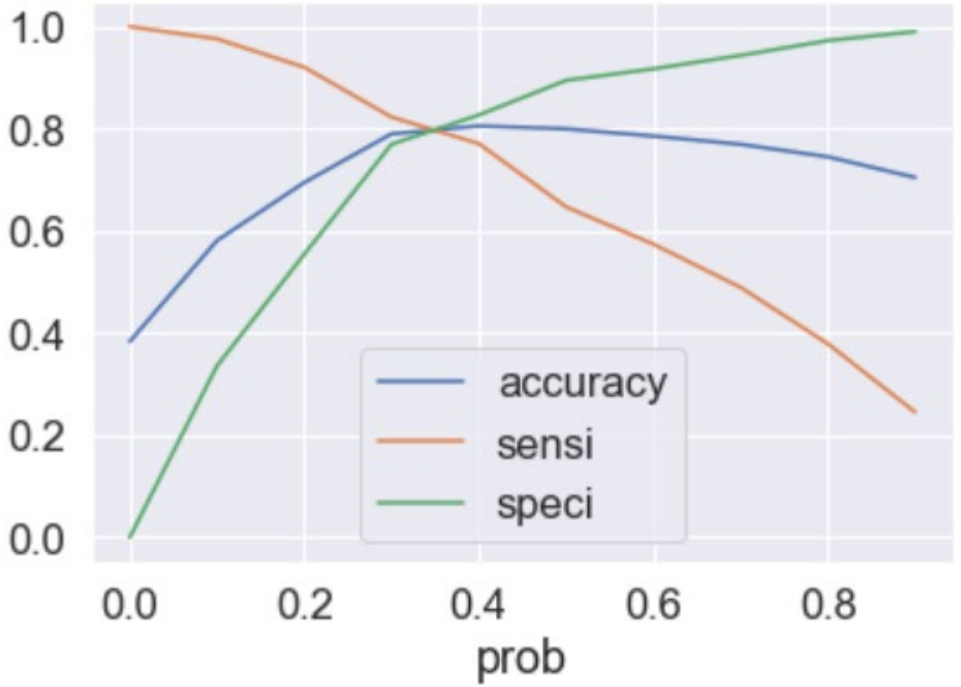
	prob	accuracy	sensi	speci	preci	recall
0.0	0.0	0.383342	1.000000	0.000000	0.383342	1.000000
0.1	0.1	0.581201	0.976003	0.335774	0.477379	0.976003
0.2	0.2	0.694765	0.920593	0.554380	0.562217	0.920593
0.3	0.3	0.790099	0.822862	0.769731	0.689580	0.822862
0.4	0.4	0.805486	0.770942	0.826960	0.734719	0.770942
0.5	0.5	0.799799	0.647033	0.894765	0.792624	0.647033
0.6	0.6	0.785750	0.574171	0.917277	0.811845	0.574171
0.7	0.7	0.769527	0.489529	0.943586	0.843609	0.489529
0.8	0.8	0.745108	0.379145	0.972606	0.895876	0.379145
0.9	0.9	0.704633	0.245201	0.990236	0.939799	0.245201



METRICS INTERPRETATION (CONT..)

- ROC Curve shows healthy accuracy of the model built.
- Sensitivity and Specificity values are 0.58 and 0. 92 with cut-off probability rate 0.6





OUTPUT ANALYSES

1. Probability cut off value has been decided as 0.6 to get the target precision value (80%).

2. Lead score calculated with the help of Converted_Prob i.e. Coverted_Prob*100.

```
# Reveiw the lead score
y_pred_final_lead_score.head()
```

	LeadID	Converted	Converted_Prob	Lead_score	predicted
0	1234	0	0.186653	18.67	0
1	1824	1	0.886884	88.69	1
2	3219	1	0.411167	41.12	0
3	5472	0	0.072334	7.23	0
4	1257	0	0.025021	2.50	0

OUTPUT ANALYSES (CONT..)

3. Top 3 variables in model which contributed towards the probability of lead getting converted

   - Lead_Origin_Lead Add Form
   - current_occupation_Working Professional
   - Total Time Spent on Website

4. Top 3 categorical variables in model which contributed to increase probability of lead conversion

   - Lead_Origin_Lead Add Form
   - current_occupation_Working Professional
   - Last_Notable_Activity_Email Opened

OUTPUT ANALYSES (CONT..)

5. Strategy to use interns to increase the lead conversion rate aggressively.

   Focus on the leads which have more than **lead score of 10**, this way all the potential leads can be contacted and converted. This is the best way to utilize the extra resources allocated.

	LeadID	Converted	Converted_Prob	Lead_score	predicted	0.0	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9
0	1234	0	0.186653	18.67	1	1	1	0	0	0	0	0	0	0	0
1	1824	1	0.886884	88.69	1	1	1	1	1	1	1	1	1	1	0
2	3219	1	0.411167	41.12	1	1	1	1	1	1	0	0	0	0	0
3	5472	0	0.072334	7.23	0	1	0	0	0	0	0	0	0	0	0
4	1257	0	0.025021	2.50	0	1	0	0	0	0	0	0	0	0	0

6. Strategy to reduce phone calls when target reached

   Focus on the leads which have more than **lead score of 90**, these are the top leads which has the best probability among all others. There will be less number of these top leads , hence number of phone calls needed will be less.