Lead Scoring

By
- Prema Halvi
- Ruchi Varma

Business Problem:

X Education has appointed our team to help them select the most promising leads, i.e. the leads that are most likely to convert into paying customers. The company requires us 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:

Understanding the Problem Statement

Data Collection

Data Pre-processing and Analysis

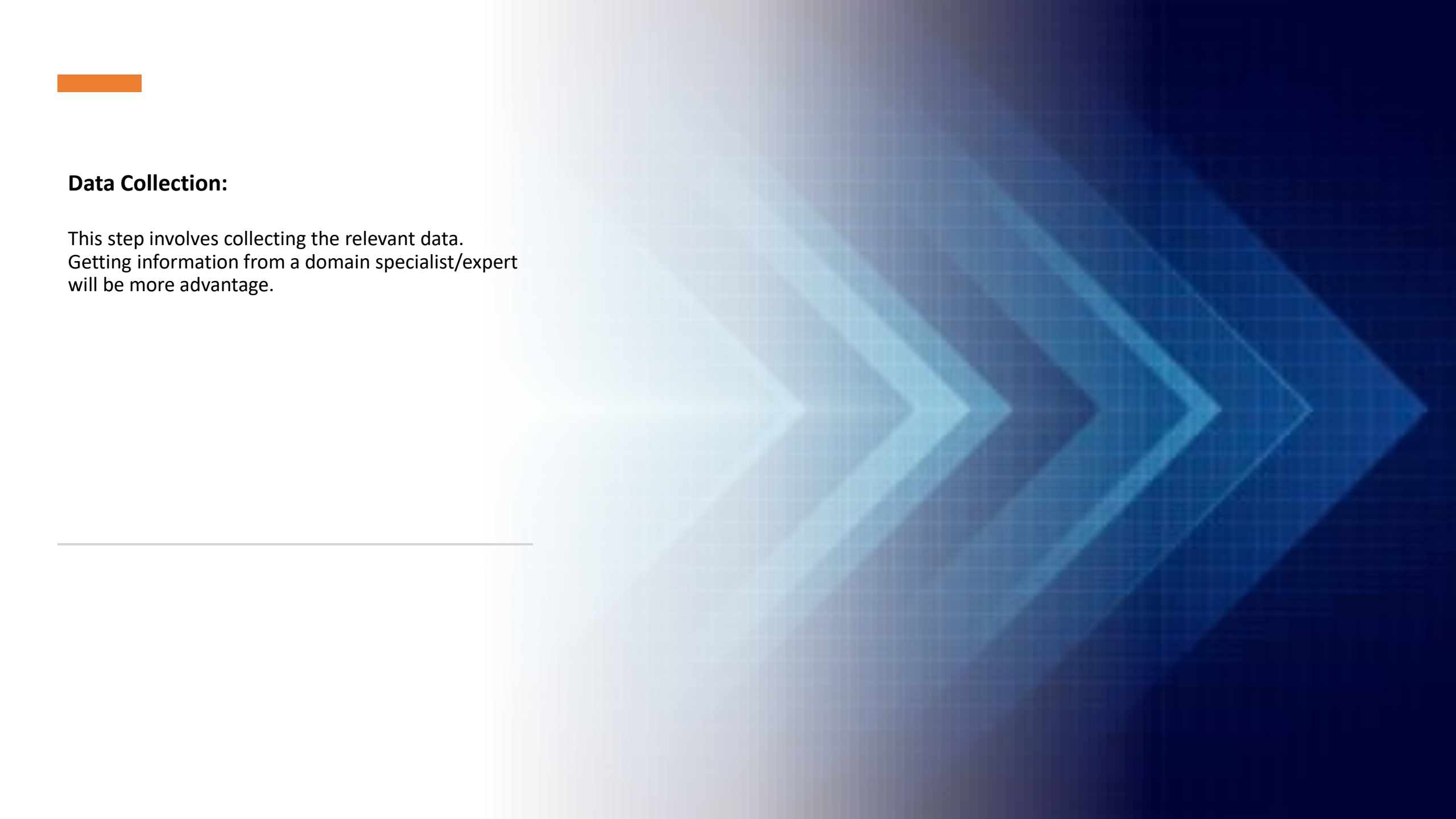
Creating and Testing Model

Performance Evaluation

Model Improvements

Data Collection:

This step involves collecting the relevant data. Getting information from a domain specialist/expert will be more advantage.



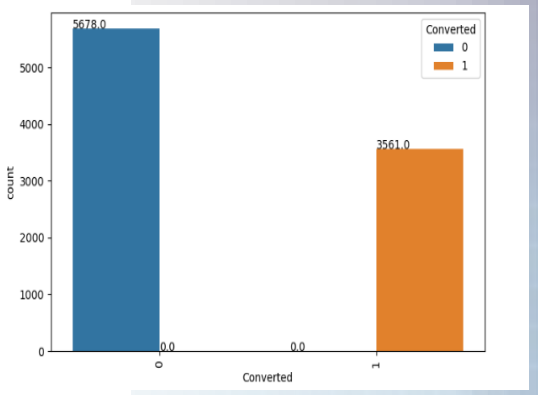
Data Exploration:

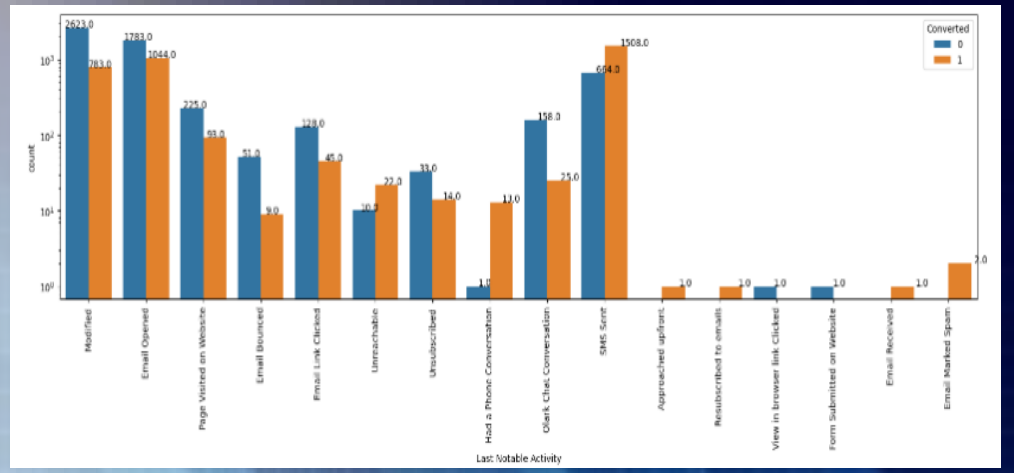
Begin by exploring the dataset related to leads. This includes variables such as browsing behaviour, past interactions, etc. Understand the meaning and significance of each variable in the context of lead scoring.

When we examined the number of rows and columns and the data types. This helps to understand the size of the dataset and the nature of the variables involved.

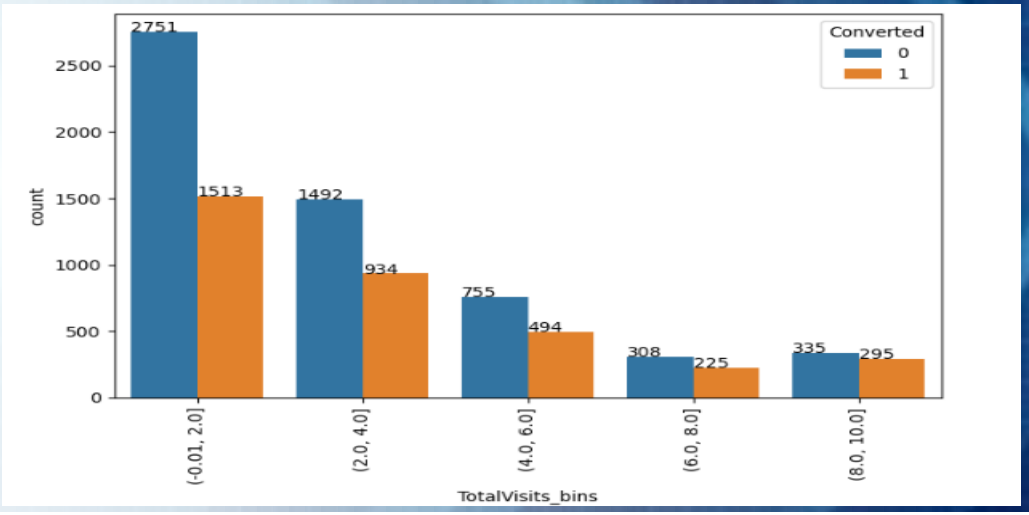
Data Analysis:

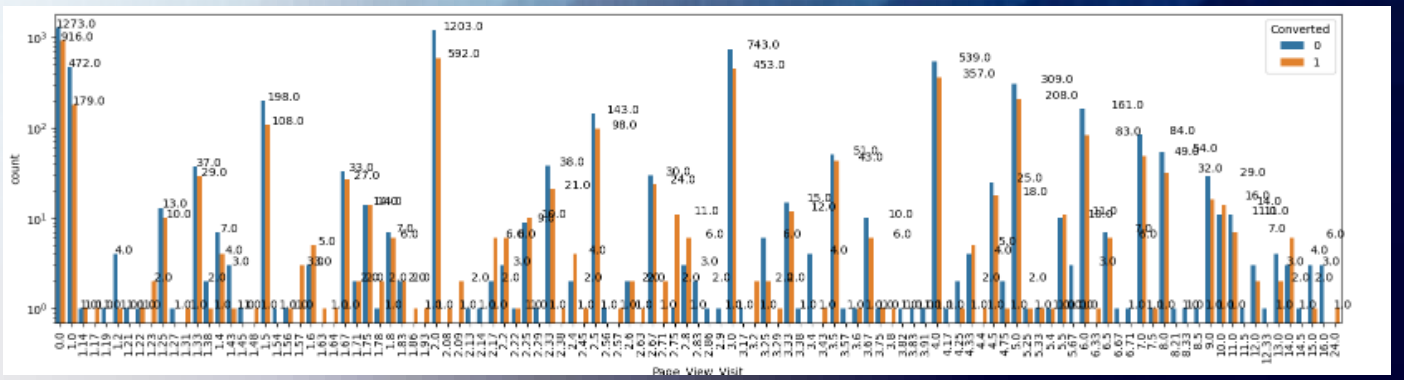
Total Conversion Rate



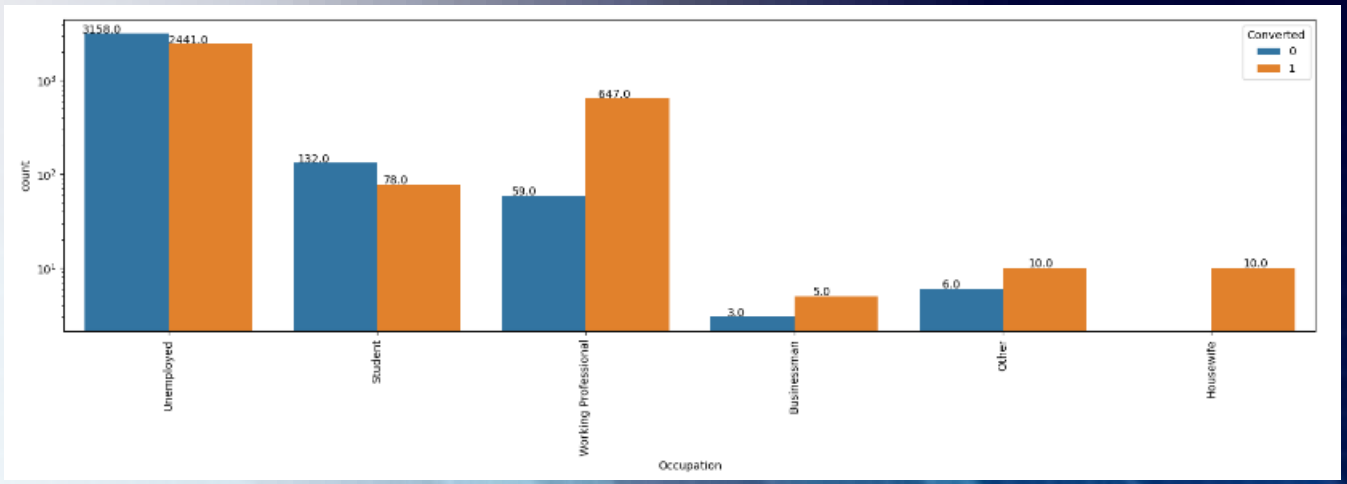


Conversion rates were high for the following features
SMS sent,
Had a phone Conversation
Last notable activity
TotalVisits
Page_View_Visit

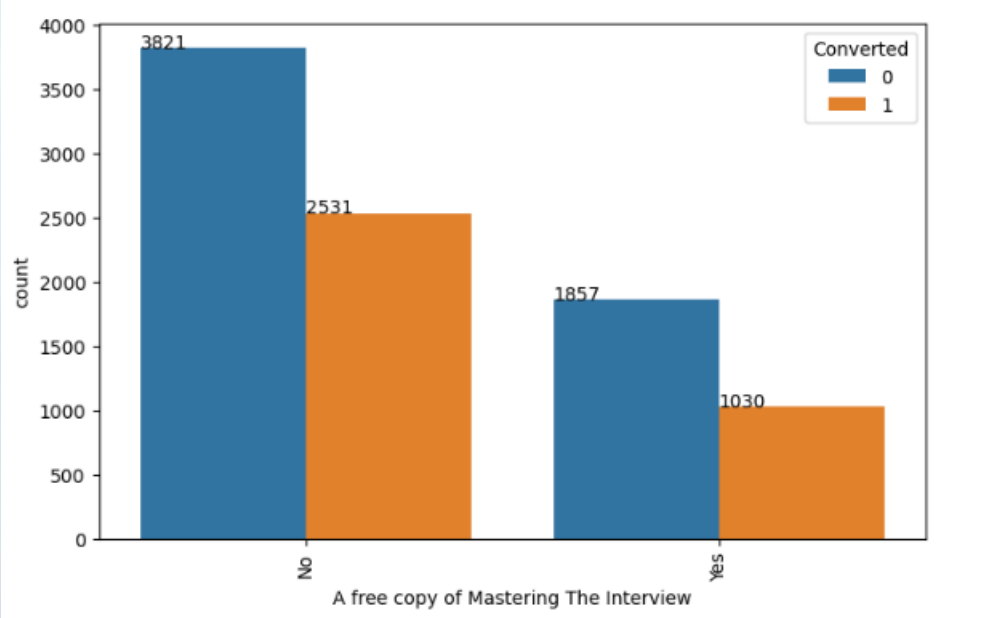




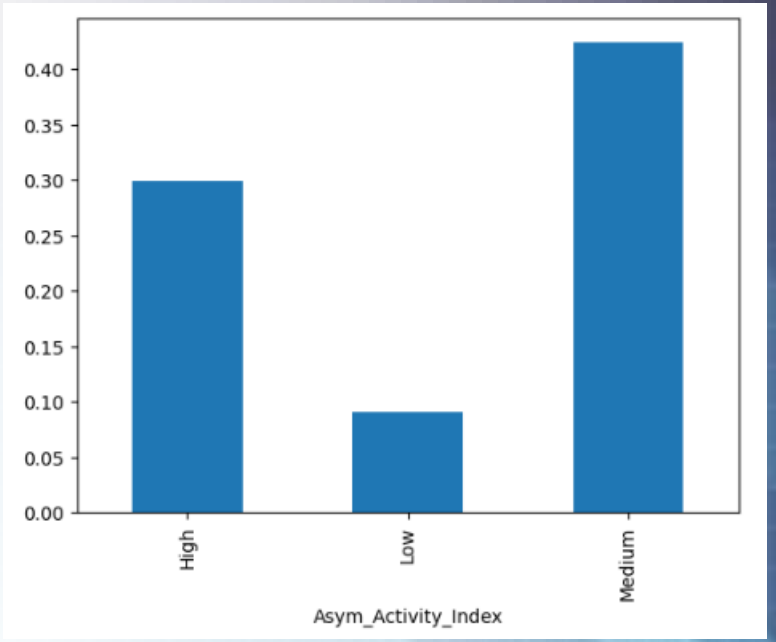
Occupation has better conversion for Unemployed and working professionals

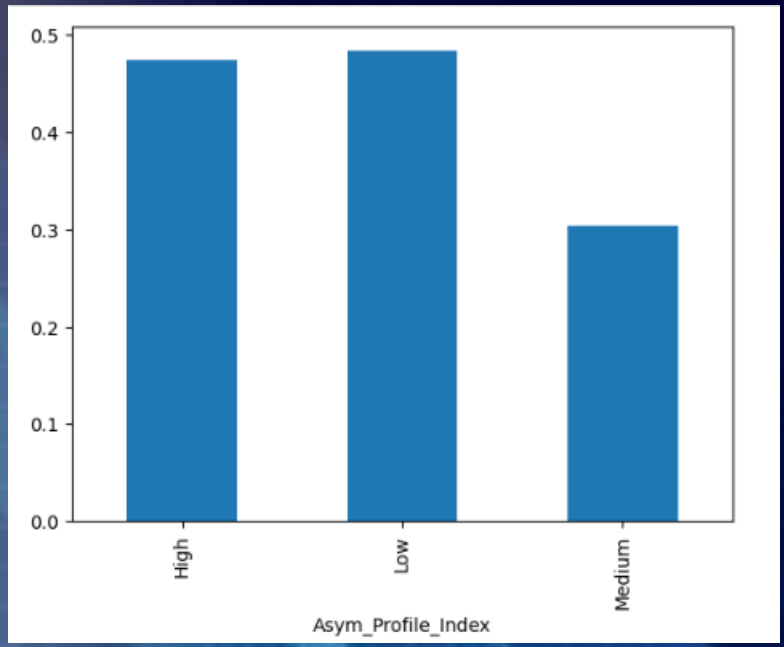


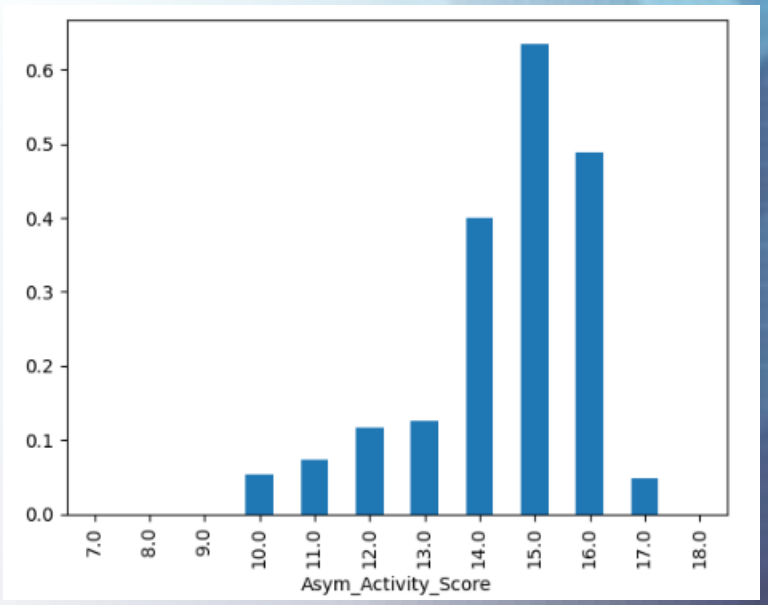
Free Copy of Interview has better conversion but most of the values are “No”

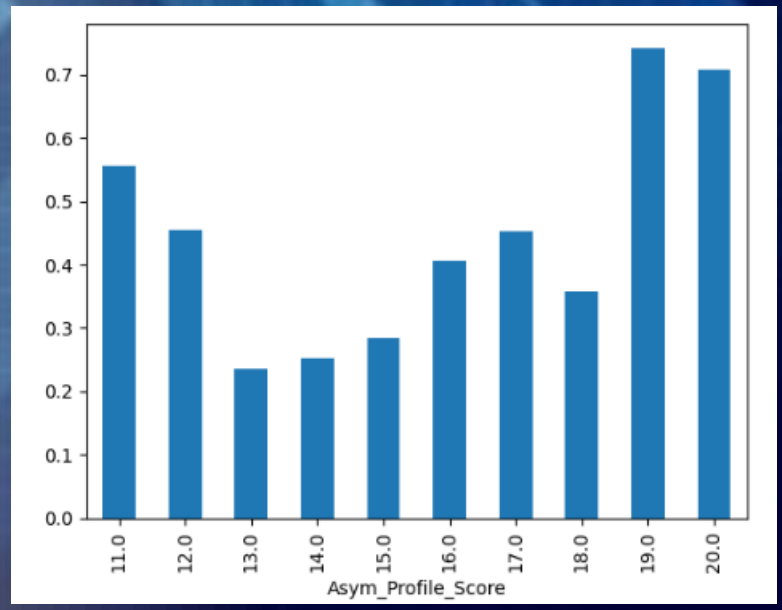


Asymmetrique Activity Index,
Asymmetrique Profile Index ,
Asymmetrique Activity Score ,
Asymmetrique Profile Score does not have significant impact on model , also there were almost 42% of missing values, so they needed to be dropped





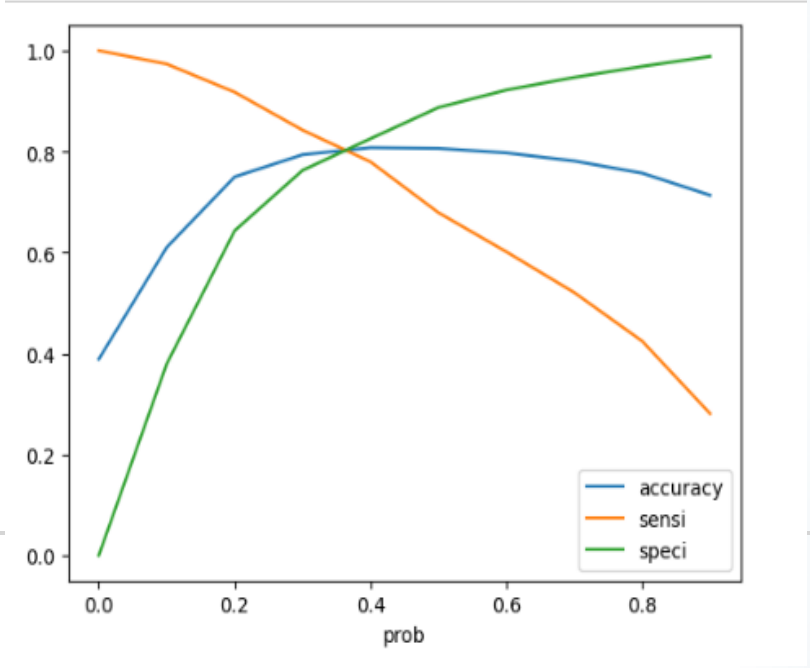




Features Impacting the Conversion Rate:

Unemployed
Lead_Origin_Landing_Page_Submission
Modified
SMS_Sent
Working_Professional
Landing_Origin_Lead_Add_Form
Time_Spent_Website
Student
Hear_About_XEducation_Multiple Sources
Chat_Conversation
Email_Bounced
Lead_Origin_Lead_Import
Hear_About_XEducation_Email
Hear_About_XEducation_SMS
Unreachable
Housewife
Other
Phone_Conversation
Resubscribed_emails

Model Evaluation on Train Data Set:

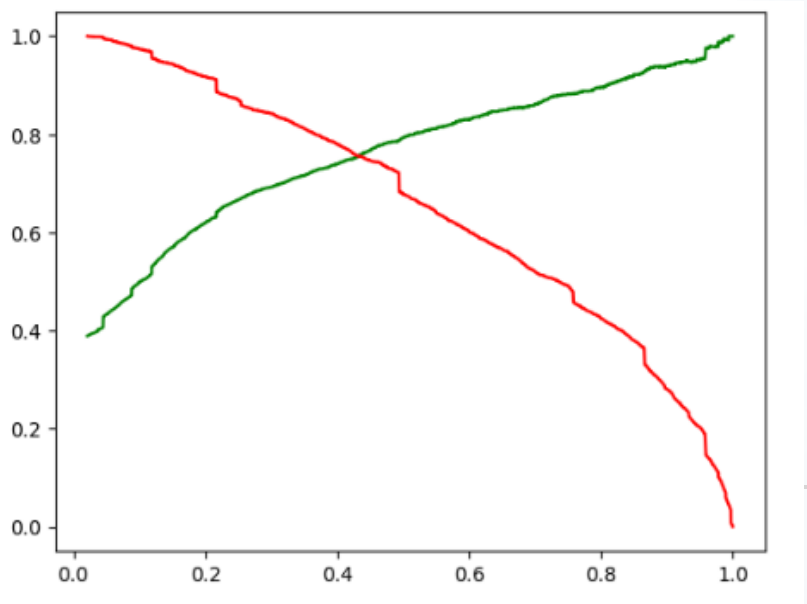


Accuracy - 80%
Sensitivity - 67%
Specificity – 88%
False positive rate –11%
positive predictive value- 79%
Negative predictive value- 81%

Confusion Matrix

Precision – 79%
Recall – 67%

Model Evaluation –Precision and Recall on Train Dataset:



Model Evaluation on Test Data Set:

Accuracy - 81%
Sensitivity - 77%
Specificity – 84%

Final Prediction Head:

	Converted	Lead_Number	Converted_Prob	final_predicted
0	0	648	0.055601	0
1	0	6077	0.313777	0
2	1	1877	0.446991	1
3	0	658	0.479273	1
4	0	6359	0.366784	0
5	0	6250	0.044805	0
6	1	3803	0.448614	1
7	1	2061	0.854604	1
8	0	6143	0.153547	0
9	0	7001	0.088013	0

Confusion Matrix

Conclusion:

The top variables are –
a. Time_Spent_Website
b. Occupation like whether the customer is Unemployed, Student or Working Professional
c. SMS Sent

Top 3 Categorical variables are-
a. Occupation like whether the customer is Unemployed, Student or Working Professional
b. Last Activity like SMS_Sent
c. Lead_Origin like Landing_Page_Submission

Suggestions may include the following

a. Target potential customers who spend a lot of time on the platform
b. Such leads can be followed up by ending SMS's
c. Anyone spending vast amount of time on the portal might be comparing prices and course content. Highlight these aspects when the lead is on the portal