LEAD SCORE CASE STUDY

Deekshith Jagithyala

Crysl Lobo

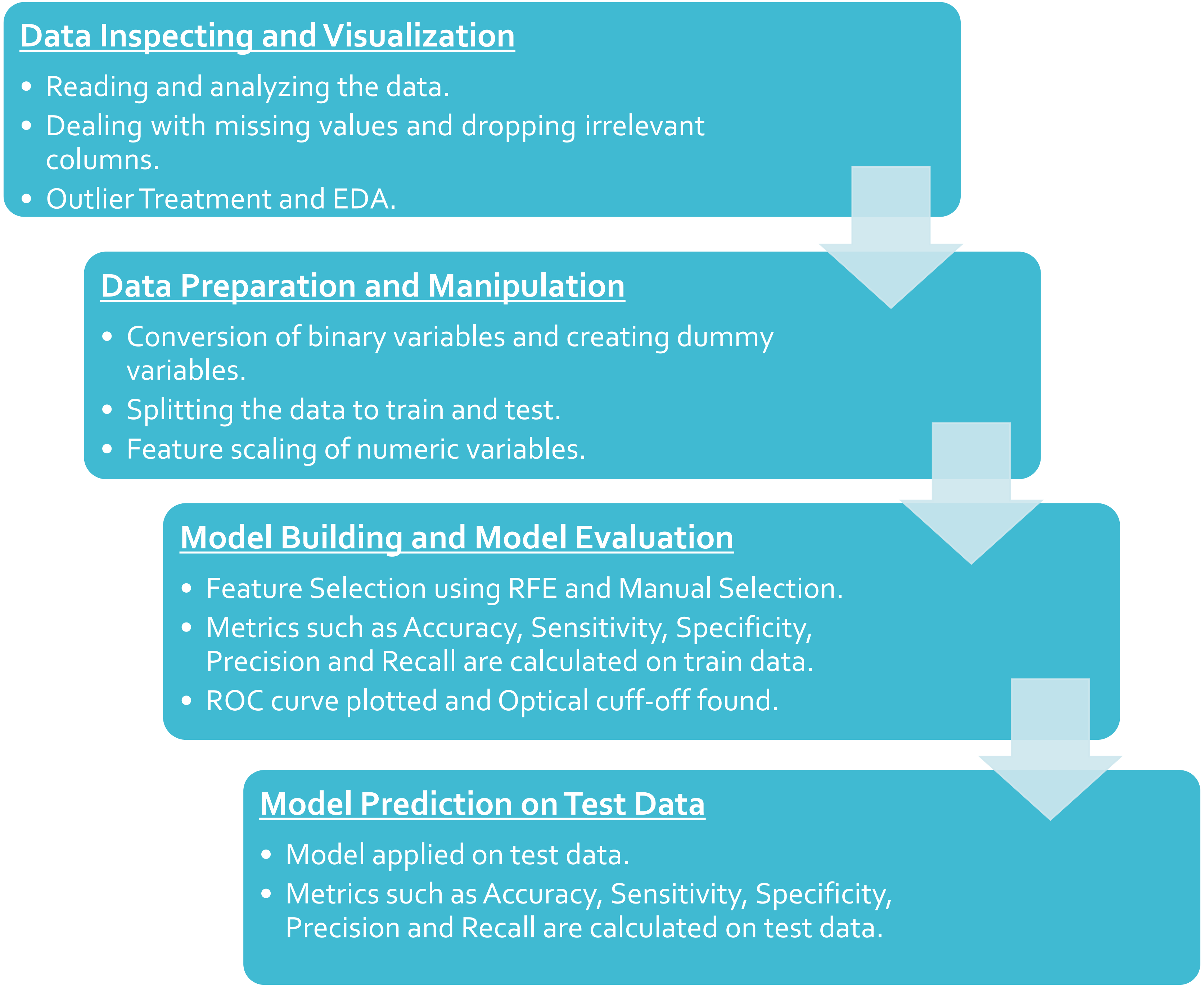
Problem Statement

- 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.
- The company markets its courses on several websites and search engines like Google. Once the leads are acquired from various sources, 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 Goal:

- X Education requires you to build a model wherein you need to 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, has given a ballpark of the target lead conversion rate to be around 80%.

Problem Solving Methodology



Data Inspecting and Visualization

- Reading and analyzing the data.
- Dealing with missing values and dropping irrelevant columns.
- Outlier Treatment and EDA.

Data Preparation and Manipulation

- Conversion of binary variables and creating dummy variables.
- Splitting the data to train and test.
- Feature scaling of numeric variables.

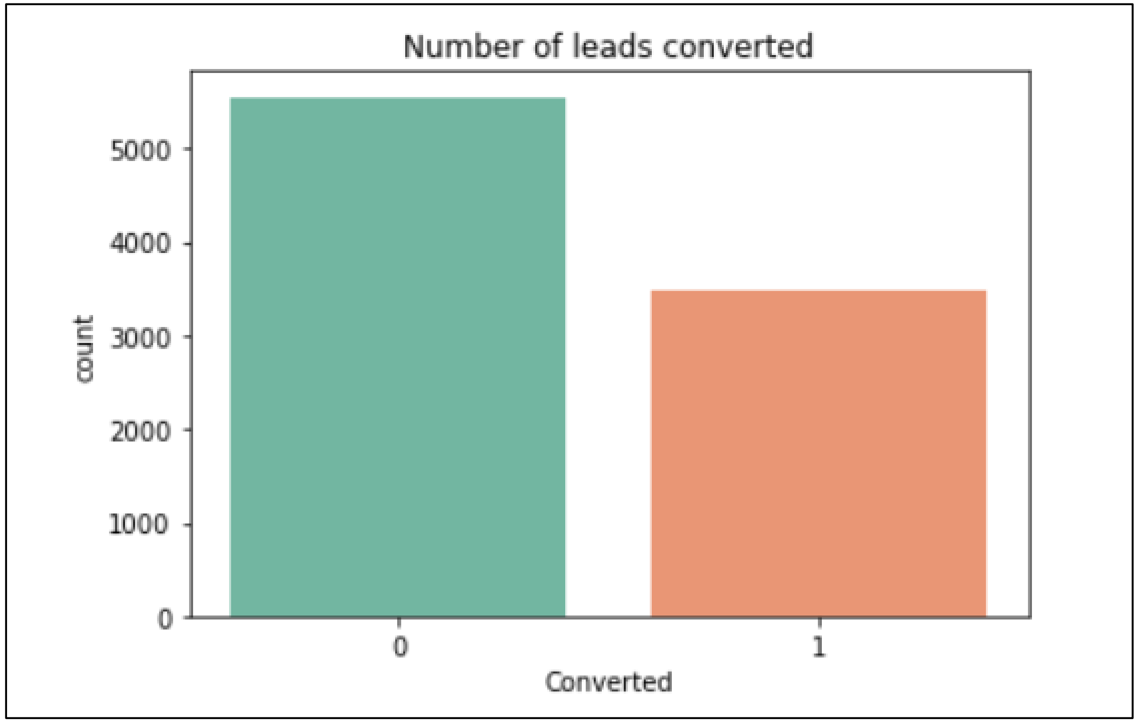
Model Building and Model Evaluation

- Feature Selection using RFE and Manual Selection.
- Metrics such as Accuracy, Sensitivity, Specificity, Precision and Recall are calculated on train data.
- ROC curve plotted and Optical cuff-off found.

Model Prediction on Test Data

- Model applied on test data.
- Metrics such as Accuracy, Sensitivity, Specificity, Precision and Recall are calculated on test data.

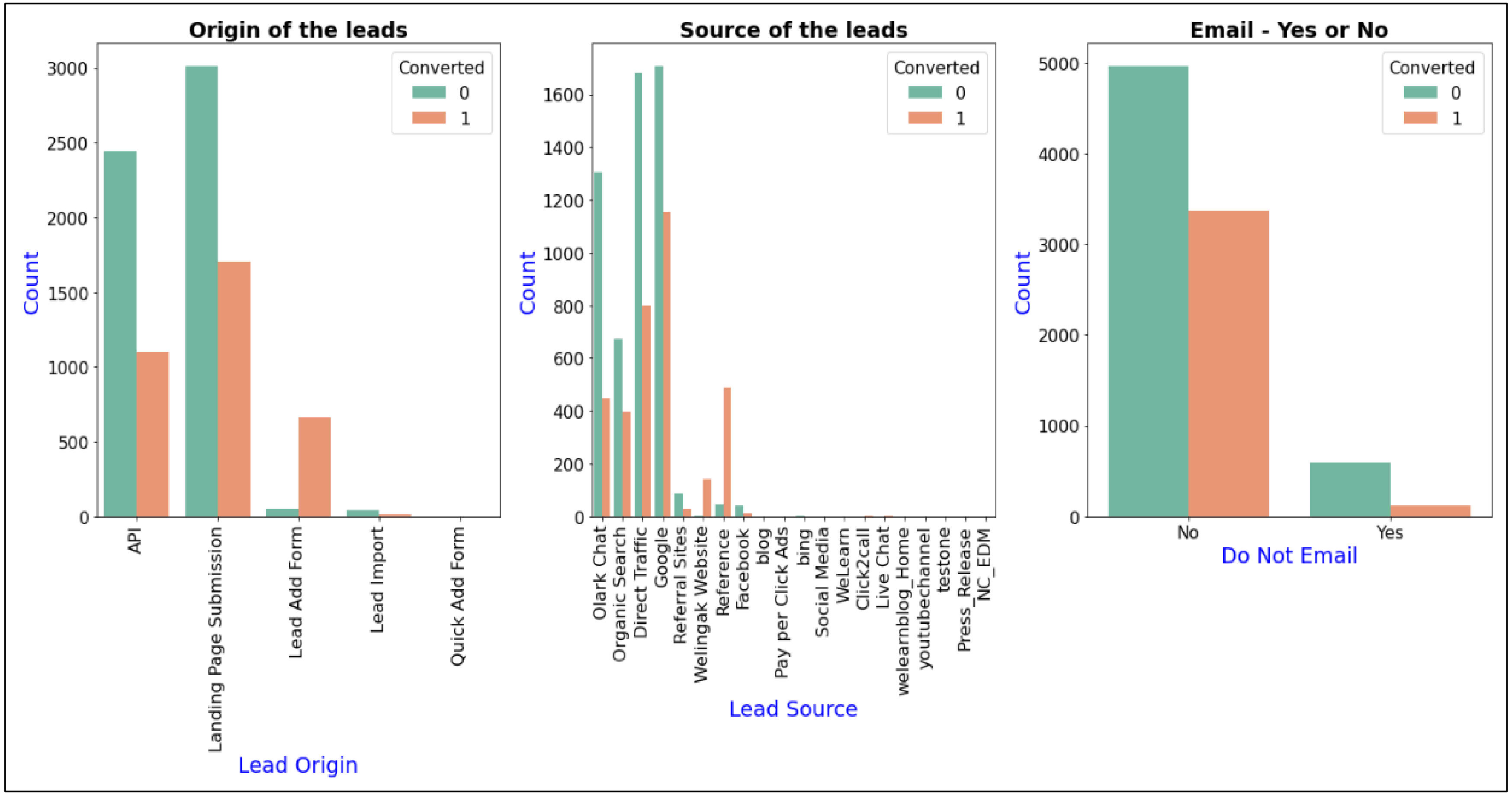
Lead Conversion Rate



Lead Conversion Rate is 39% (Before Model Building).

EDA

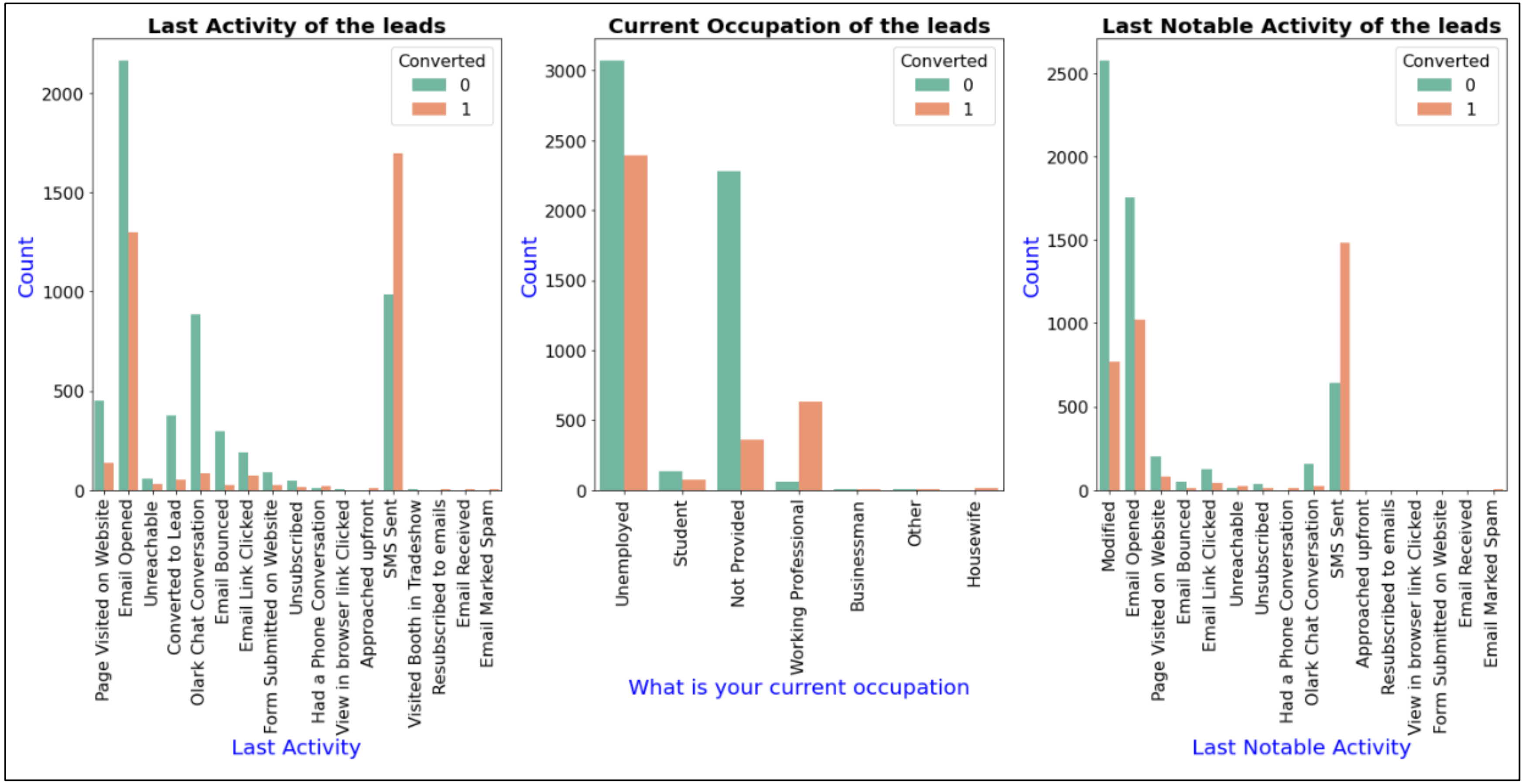
Categorical Variables



1. Lead Origin – Maximum conversion is from Landing Page Submission.
2. Lead Source – Google is the main source of conversion.

EDA

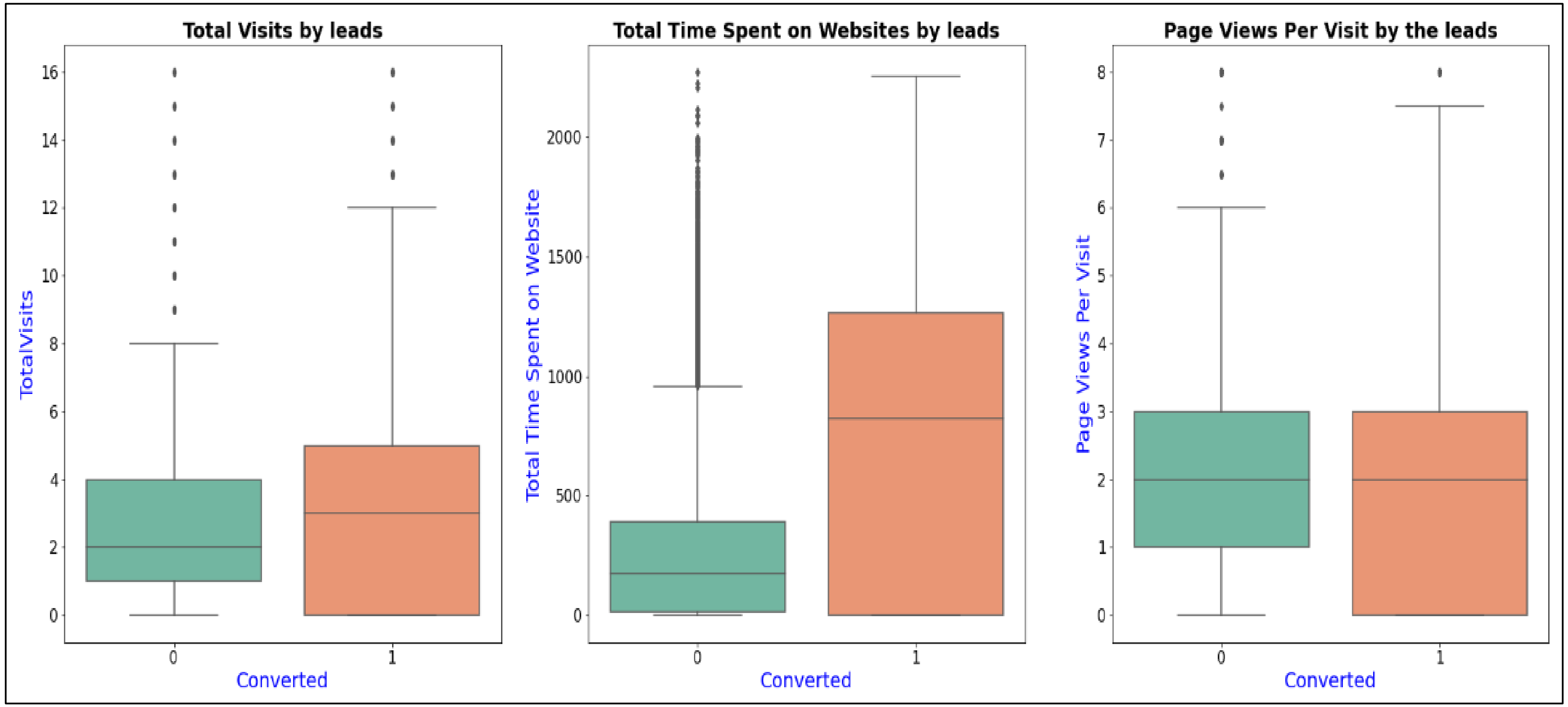
Categorical Variables



1. Last Activity, Last Notable Activity – SMS sent has the highest conversion.
2. What is your current occupation – Unemployed has the major conversion

EDA

Continuous Variables



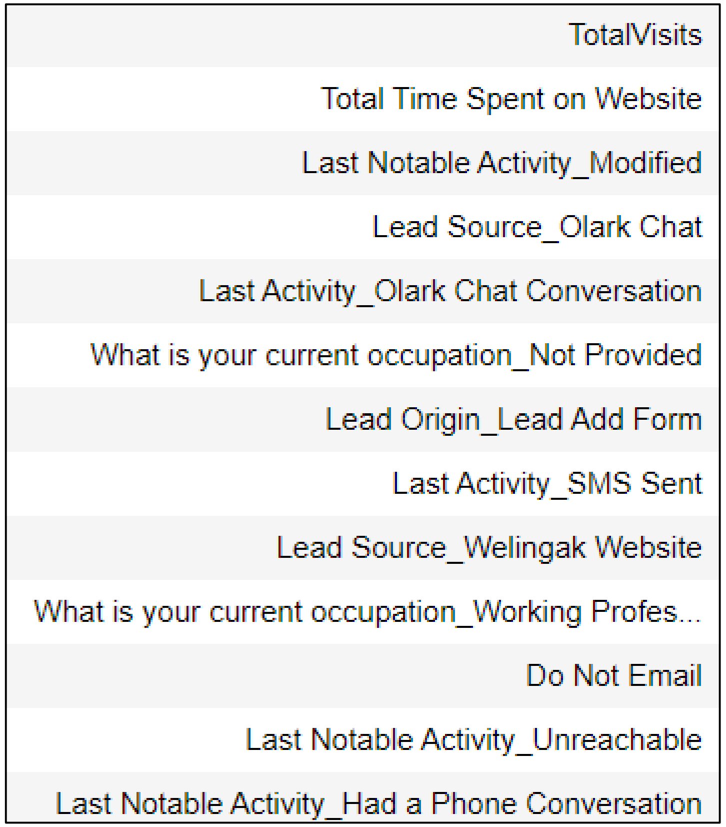
From the above graphs when the count of total visits, total time spent on websites and page views per visit are more, the conversion rate are also more.

Data Preparation and Manipulation

- The binary variables ‘Do not Email’ and ‘A free copy of Mastering The Interview’ were mapped to 1s and 0s.
- One-Hot Encoding or Dummy Variables were created for the categorical variables of ‘Lead Origin’, ‘Lead Source’, ‘Last Activity’, ‘What is your Current Occupation’ and ‘Last Notable Activity’.
- The data was then split into Train and Test Sets in the ratio of 7:3.
- Feature Scaling using MinMax Scaler was conducted on the continuous variables of ‘TotalVisits’ , 'Total Time Spent on Website’ and 'Page Views Per Visit’.

Model Building

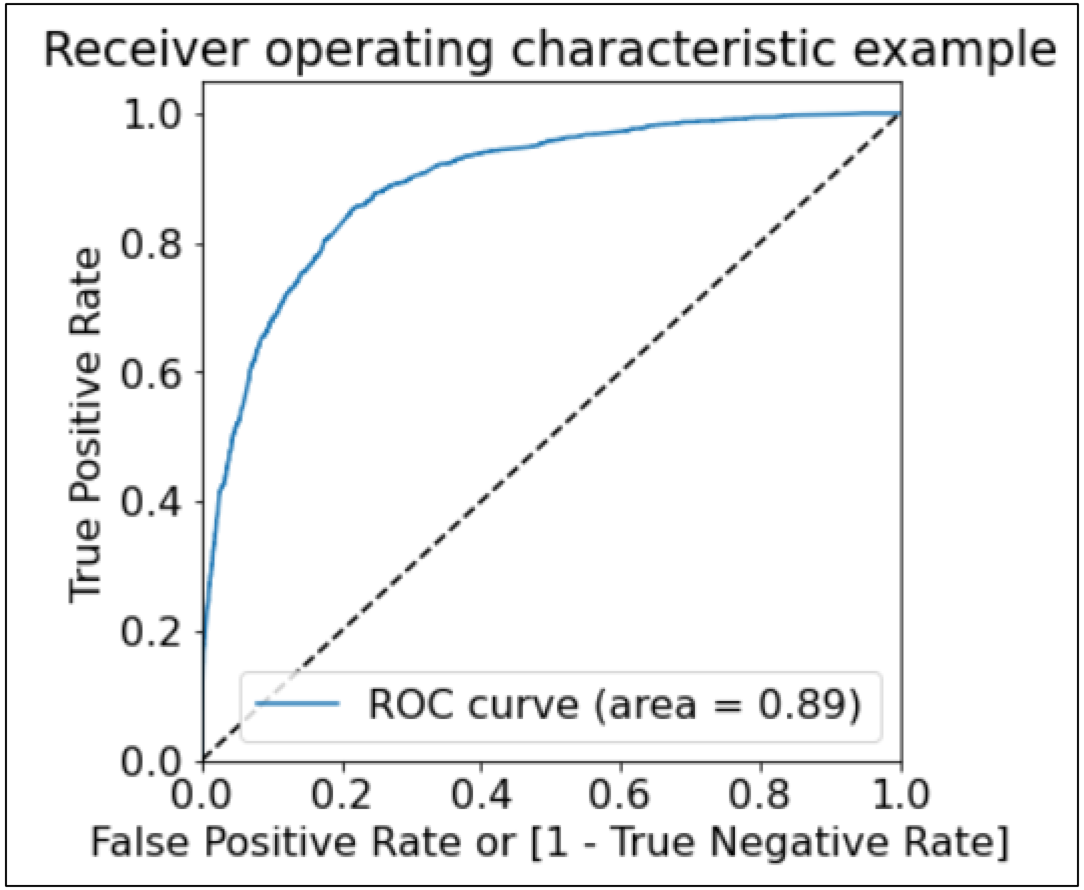
- Using Recursive Feature Elimination, 15 variables were selected for the model.
- Further using Manual Selection and after checking VIF(<5) and p-values(<0.05), 2 more variables were dropped.
- Final model consists of 13 variables.



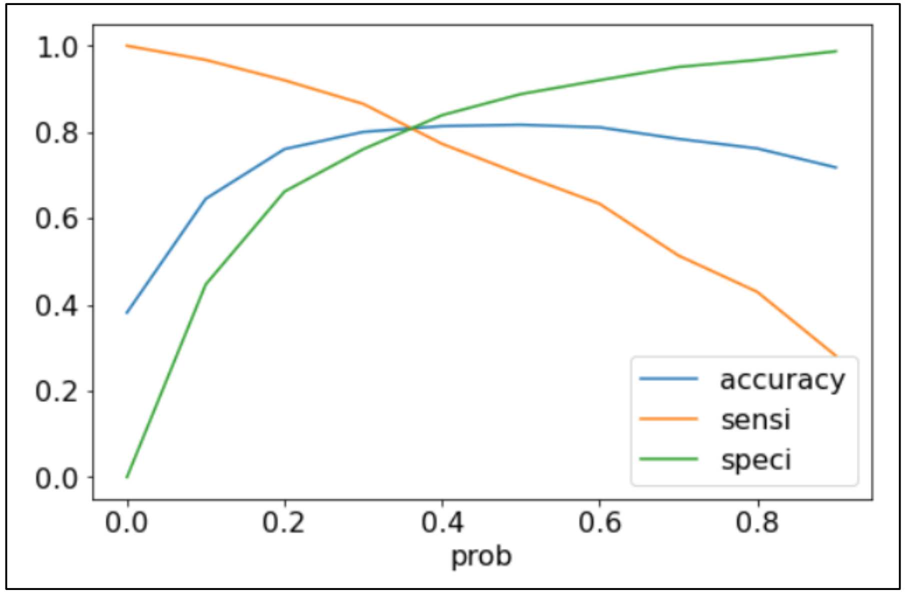
TotalVisits
Total Time Spent on Website
Last Notable Activity_Modified
Lead Source_Olark Chat
Last Activity_Olark Chat Conversation
What is your current occupation_Not Provided
Lead Origin_Lead Add Form
Last Activity_SMS Sent
Lead Source_Welingak Website
What is your current occupation_Working Profes...
Do Not Email
Last Notable Activity_Unreachable
Last Notable Activity_Had a Phone Conversation

Model Evaluation on Train Dataset

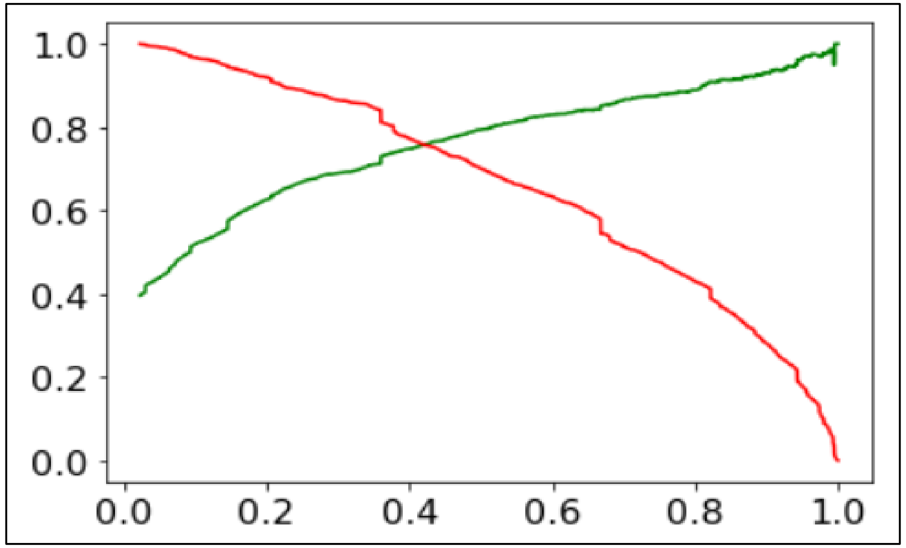
ROC curve – Area under the curve is 0.89



Model Evaluation on Train Dataset



From the curve, 0.37 is the optimum point to take it as a cutoff probability.



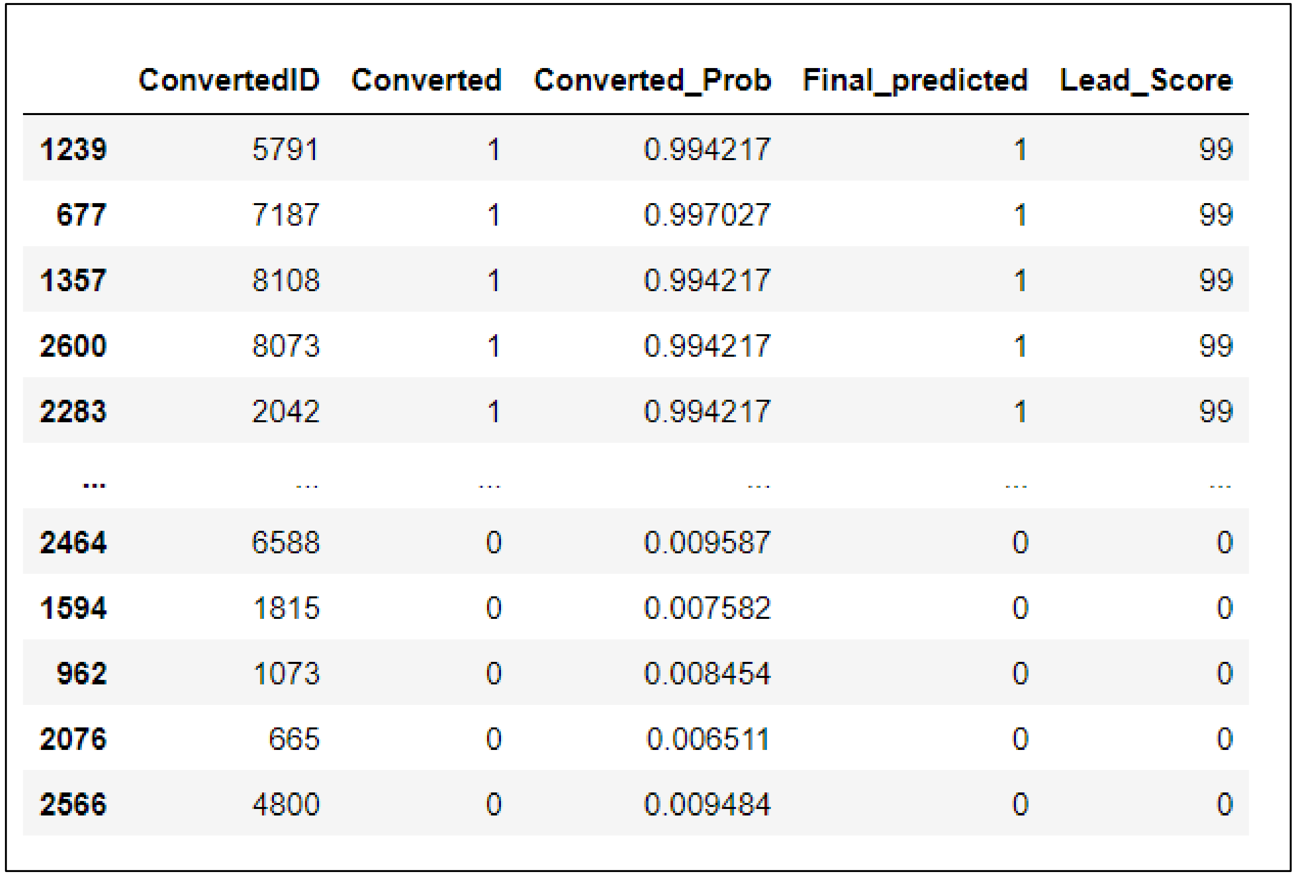
Precision and Recall Tradeoff

Accuracy score - 81.52%
Sensitivity score - 80.63%
Specificity score - 82.07%
Precision score – 73.50%
Recall Score – 80.63%

Model Evaluation on Test Dataset

- The model built was applied on the test dataset.
- The Evaluation metrics were checked on the test dataset.
- Accuracy score – 81.25%
- Sensitivity score – 80.86%
- Specificity score – 81.50%
- Precision score – 73.93%
- Recall score – 80.86%

Lead Score



	ConvertedID	Converted	Converted_Prob	Final_predicted	Lead_Score
1239	5791	1	0.994217	1	99
677	7187	1	0.997027	1	99
1357	8108	1	0.994217	1	99
2600	8073	1	0.994217	1	99
2283	2042	1	0.994217	1	99
...	...	...	...	...	...
2464	6588	0	0.009587	0	0
1594	1815	0	0.007582	0	0
962	1073	0	0.008454	0	0
2076	665	0	0.006511	0	0
2566	4800	0	0.009484	0	0

Lead score assigned to the leads which can be used by the company to target potential leads. 99 indicates high conversion and 0 indicates low conversion.

Conclusion

- The conversion rate before building the model was at 39%.
- The conversion rate as per the metrics of Accuracy, Sensitivity and Specificity on the test data is around 80%.
- The top three variables that contribute towards lead conversion are Total Time Spent on Website, Lead Origin - Lead Add Form and Last Notable Activity - Had a Phone Conversation.
- Overall, the model seems to be well built.

Recommendations

The strategy that X Education should employ at this stage can be as follows:

- They should make the website more interesting since the total time spent on the websites by the leads is the main variable that will help in conversion.
- The origin of the leads from Lead Add Form should be given more priority since they have higher chances of conversion as compared to the other lead origins.
- The leads having their last notable activity as a Phone Conversation should be given more preference as they have a better conversion rate.
- Priority should also be given to those leads who are working professionals since they have better conversion rates compared to the leads of other occupations.