LEAD SCORE CASE STUDY

Himanshu Gupta
Neetu Sharma
Yogesh Bhardwaj
DSC47 – August 2022

Problem Statement

- X Education is an education company that offers online courses to professionals in various industries.
- Despite getting numerous leads from different sources, their lead conversion rate is low, with only 30 out of 100 leads getting converted.
- To improve efficiency, the company aims to identify the most promising leads so that their sales team can focus their efforts on them, instead of wasting time contacting everyone.

Business Objective

•X Education aims to identify the most promising leads.
•To achieve this goal, the company plans to build a model that can accurately identify the most promising leads.
•Once the model is developed, X Education intends to deploy it for future use.

Solution Methodology

Data Preparation

- Check and Handle Duplicate data.
- Check and Handle Missing data
- Drop Unnecessary Columns
- Imputation of values

Data Visualization and EDA

- Check and Handle Outlier
- Univariate Data Analysis
- Bivariate Data Analysis

Data Conversion

- Feature Scaling, dummy variables and encoding of data
- Test and train split

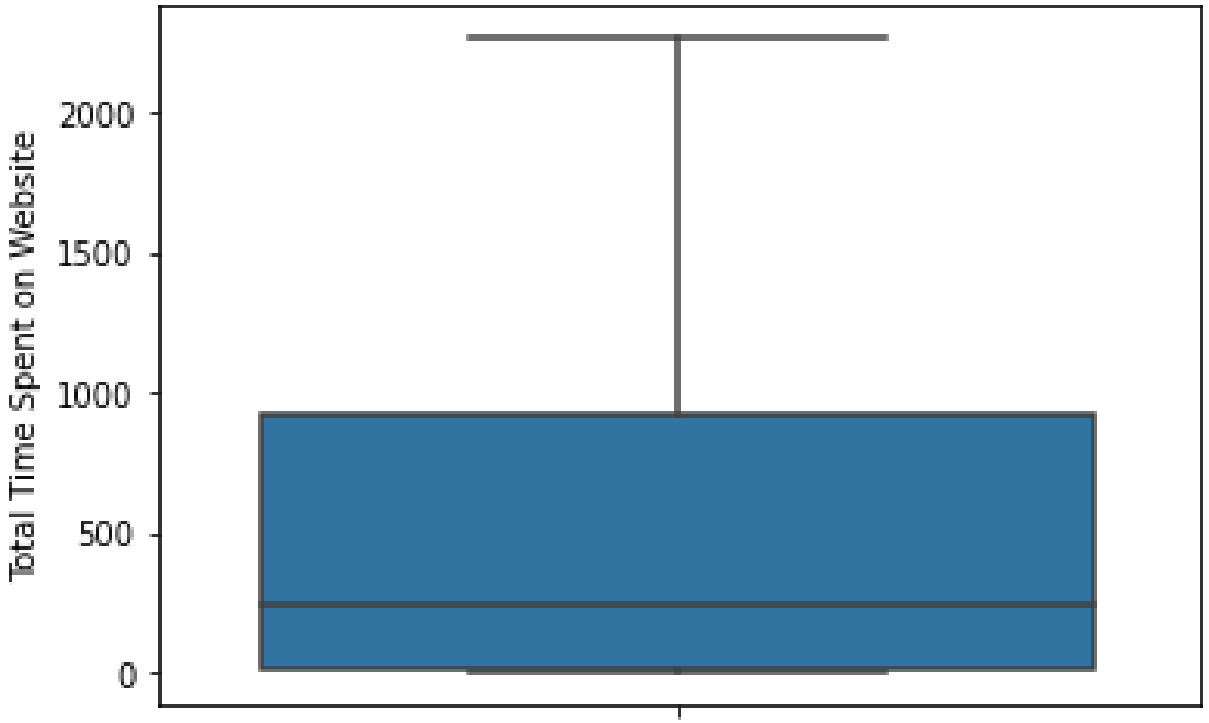
Model Building

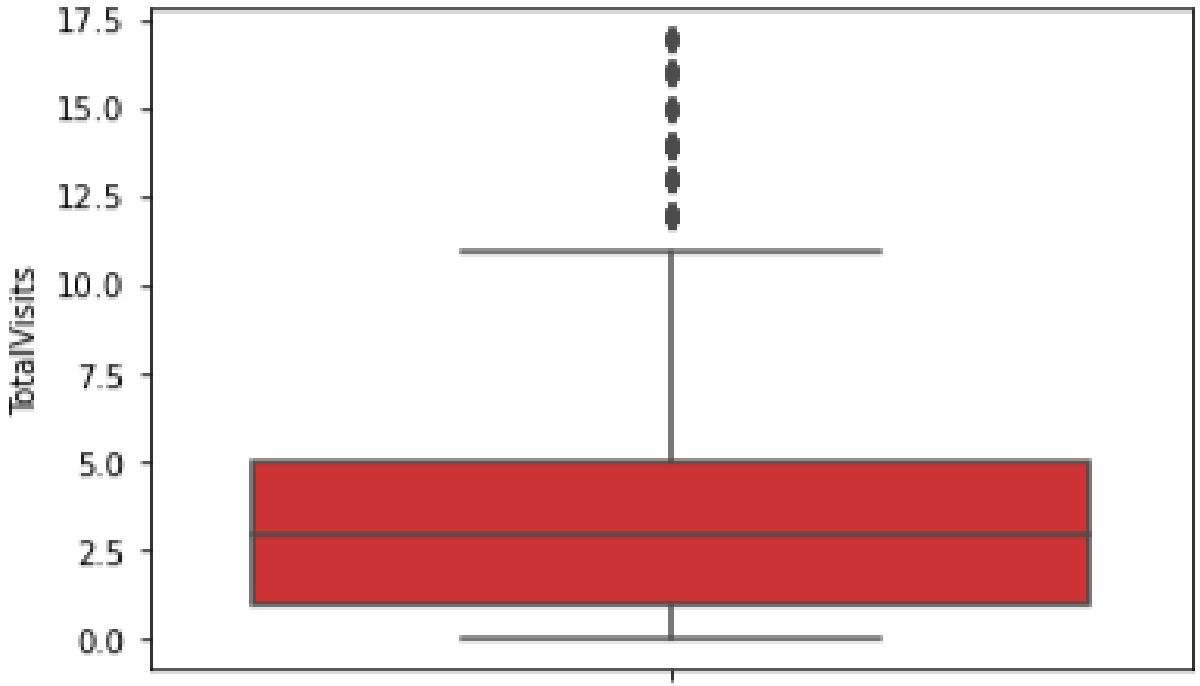
- Building the Logistic Model
- Drop the columns
- according to p-value and rebuild the model
- Finalize the model

Validation of Model

- Plotting the ROC Curve
- Making predictions on the test set

Data Visualization





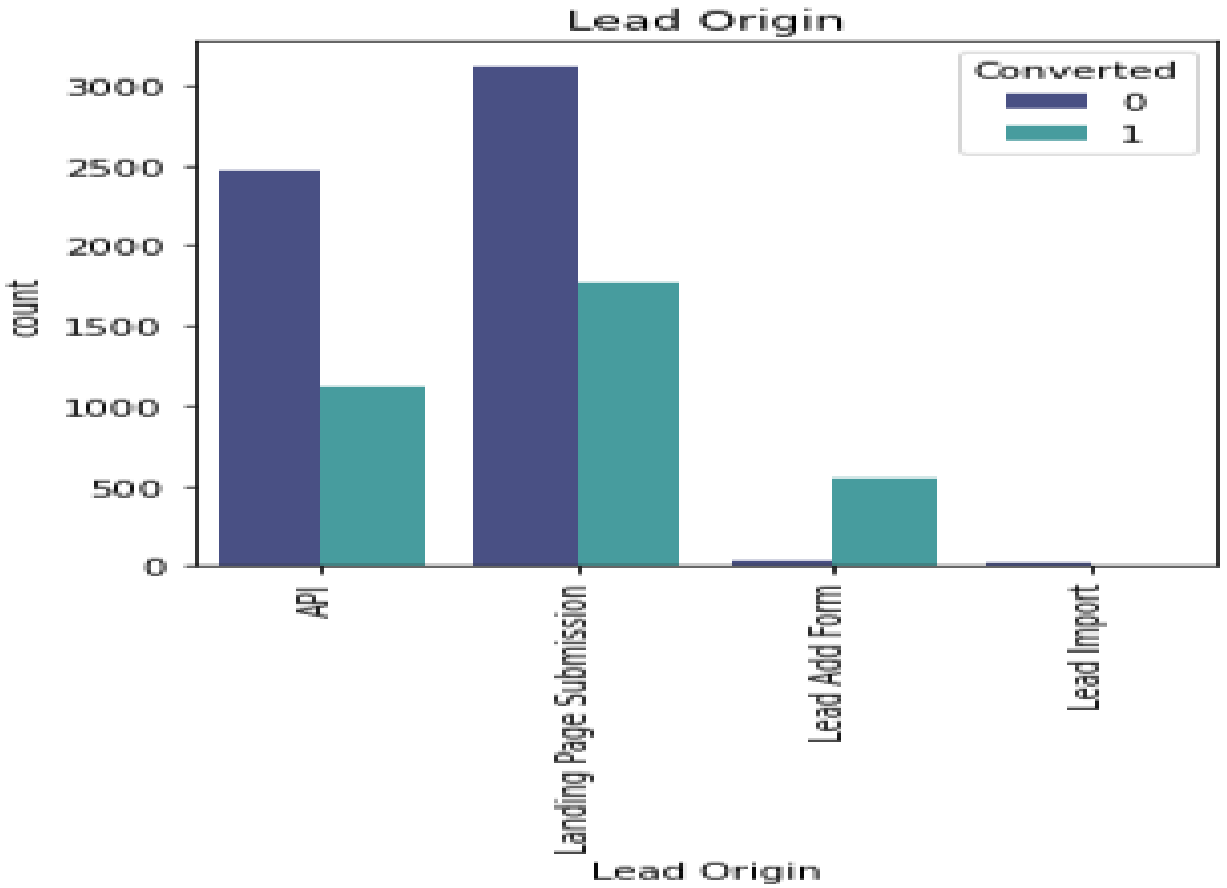
Total Time Spent on Website

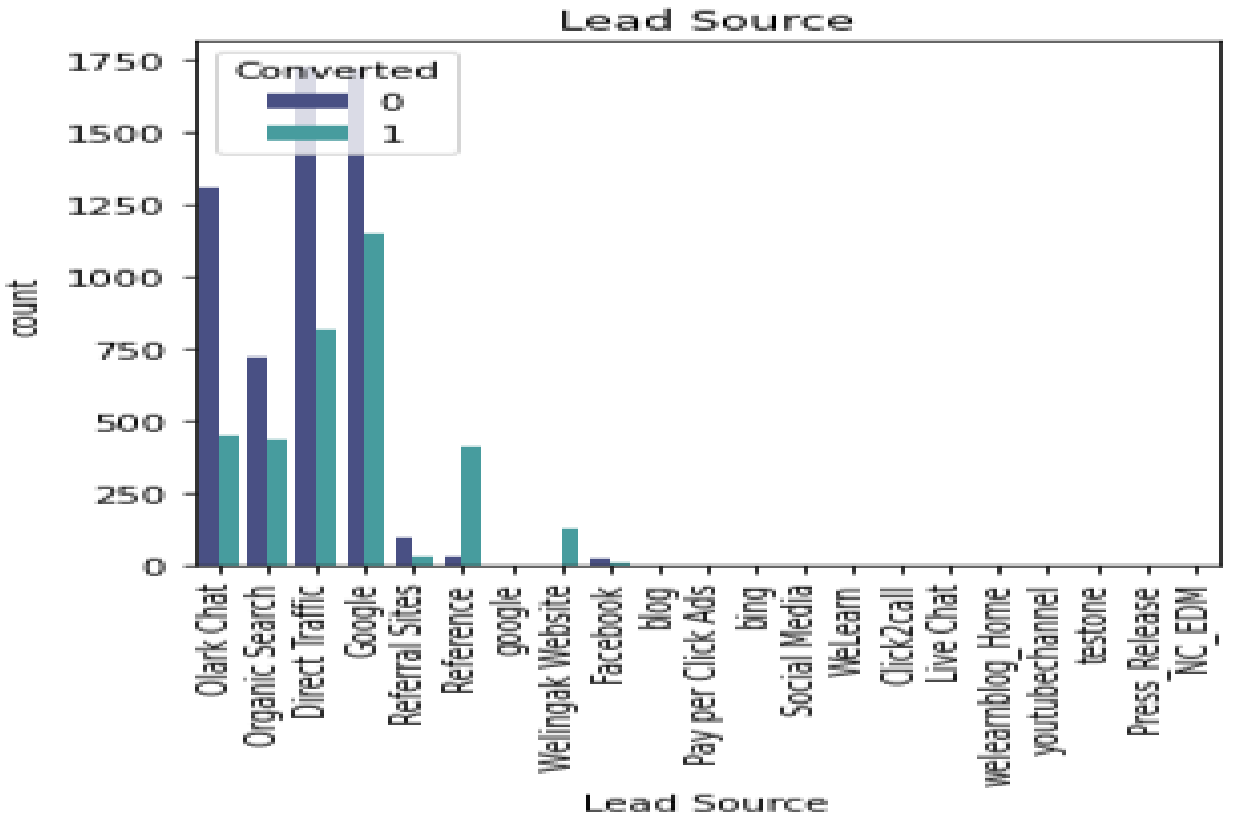
Leads spending more time on the Website are more likely to be converted.

Total visit

Nothing can be concluded on the basis of Total Visits

EDA





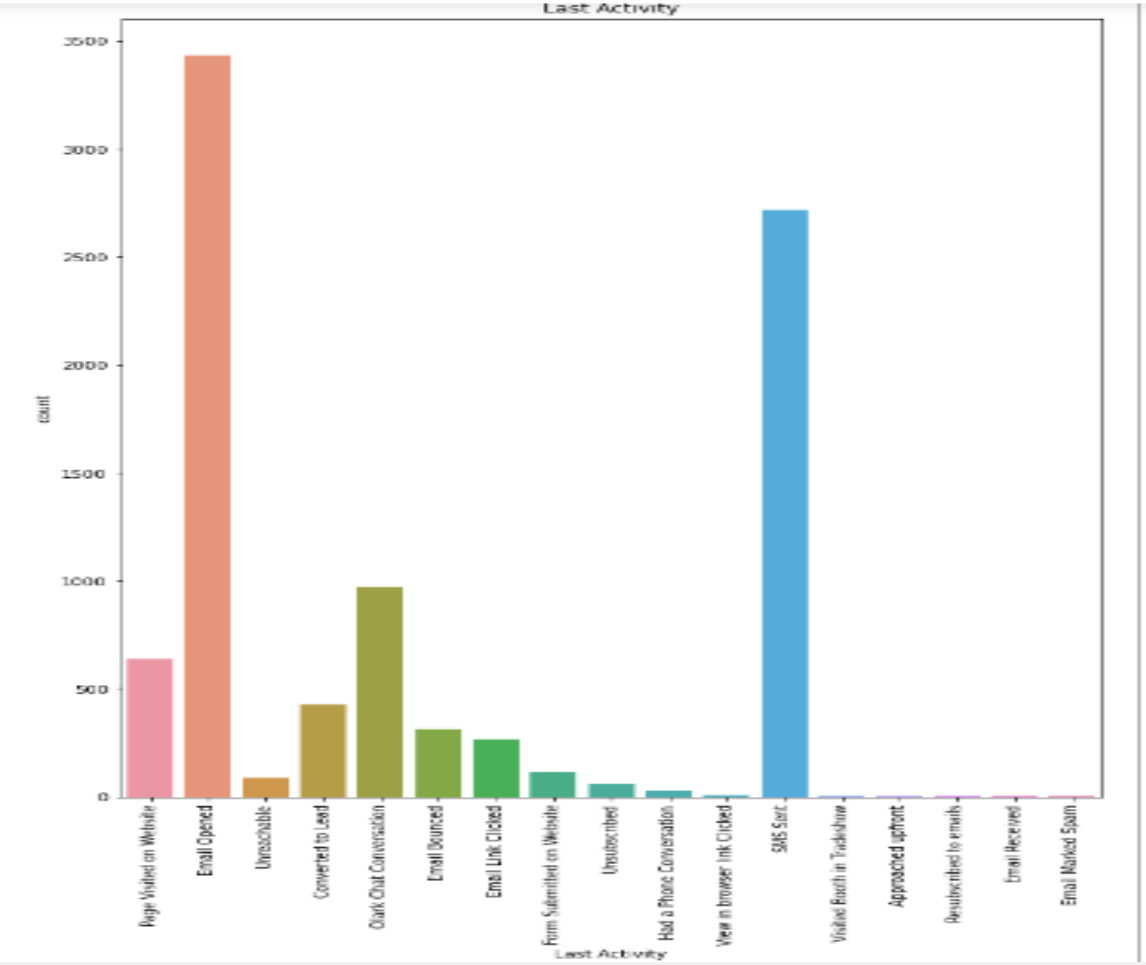
Lead Origin

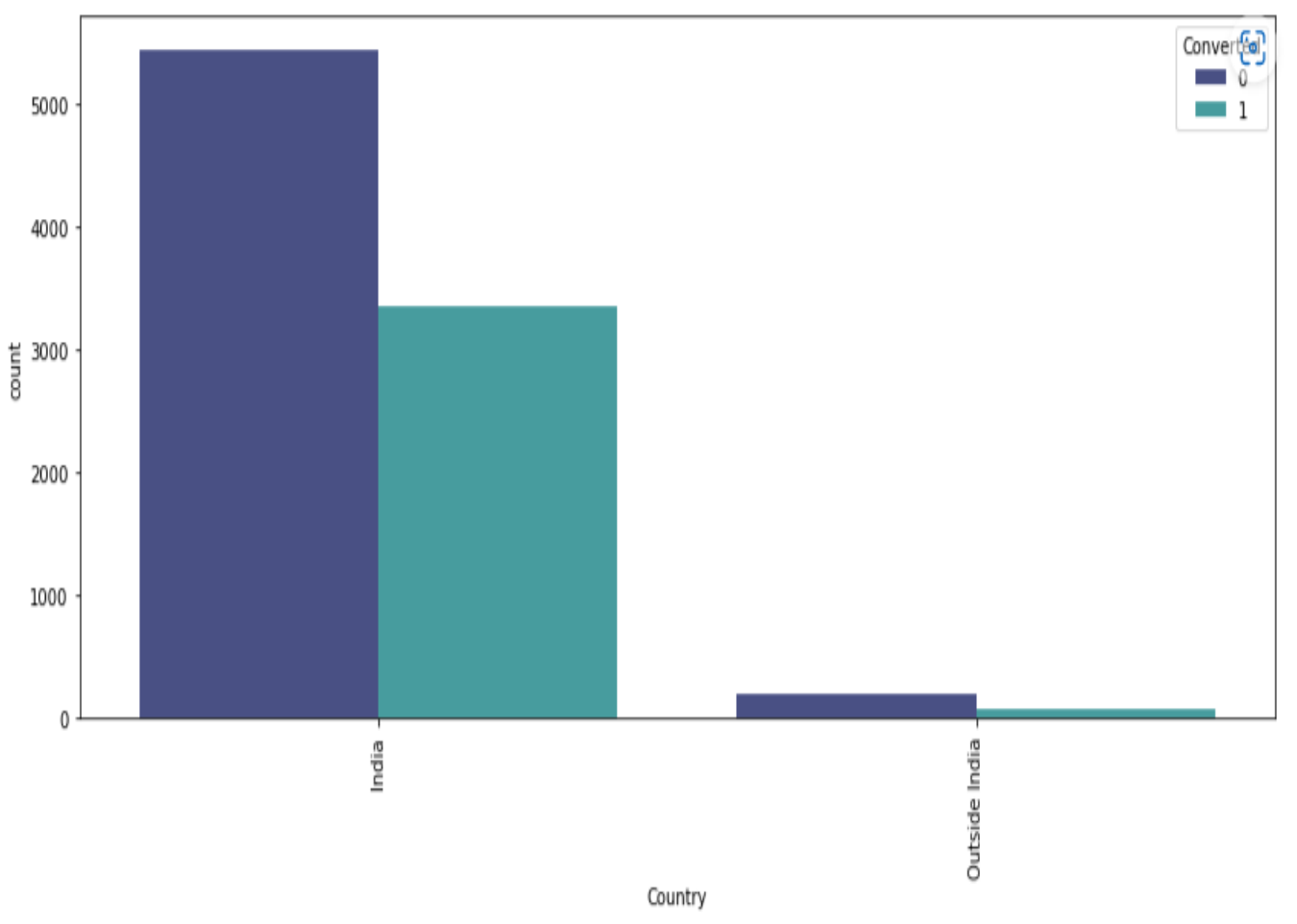
- Lead add form has maximum conversion rate but it has less number of leads.
- API and Landing page submission have less conversion rates but they have more number of leads.

Lead Source

- Google and Direct traffic have maximum number of leads.
- Conversion Rate of reference leads and welingak website is high.

EDA





Last Activity

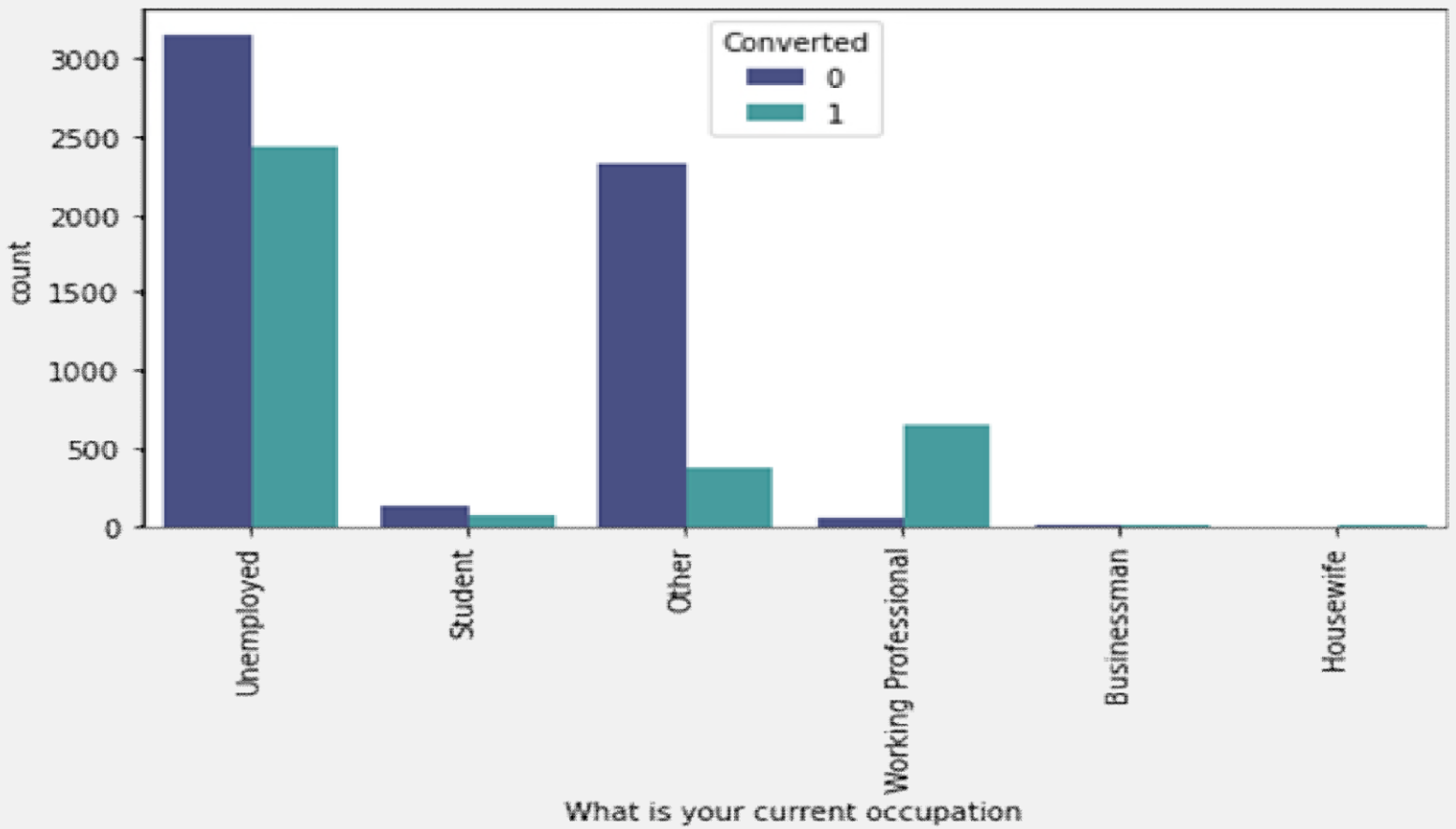
- Most of Leads have their last activity Email opened
- SMS sent last activity has good conversion rate

Country

We can observe that most of the values belong to India i.e. around 97%.

This shows that the country variable is imbalanced and hence can be dropped

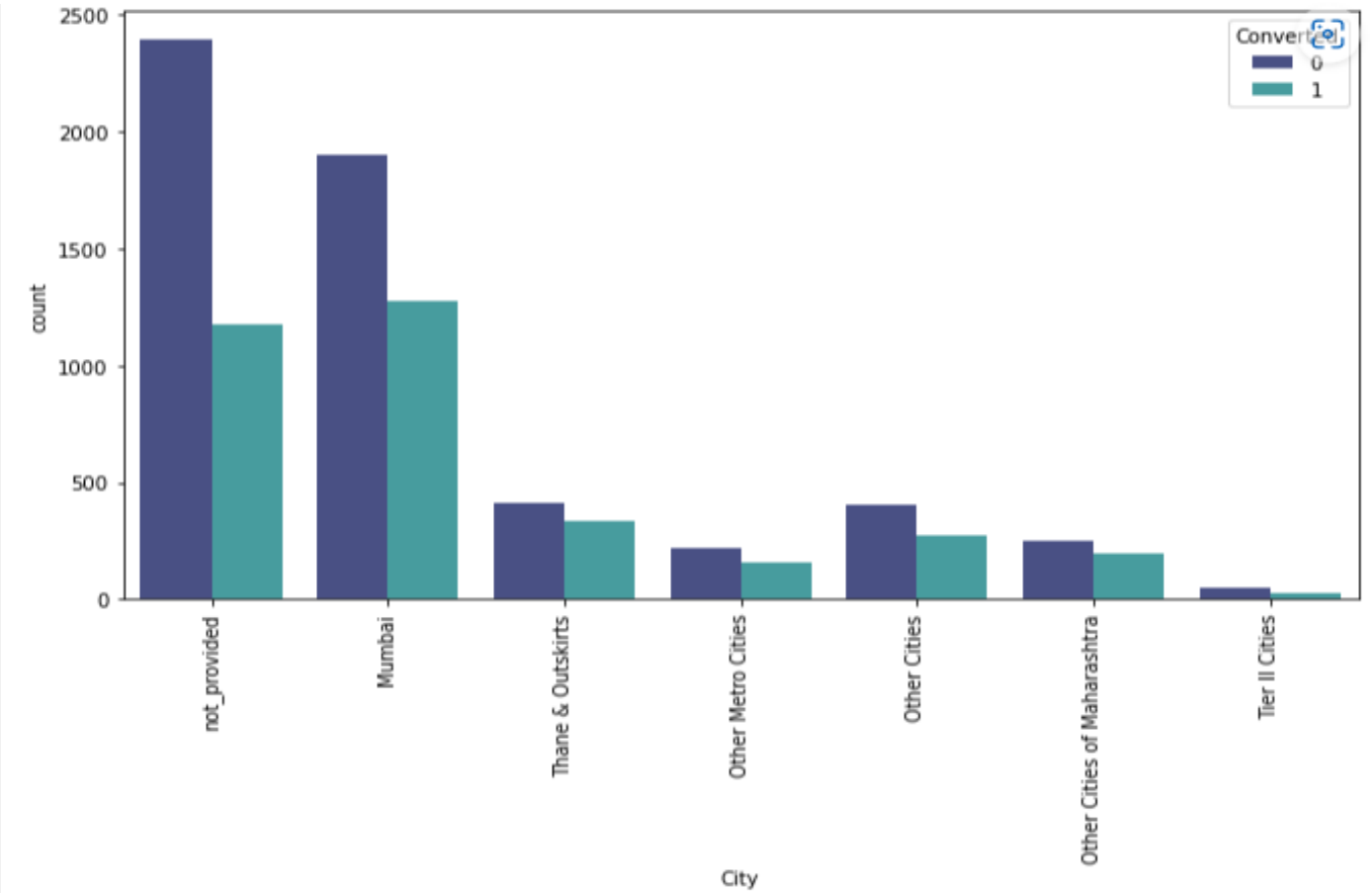
EDA



What is your current Occupation

- Working professional has high conversion rate
- Unemployed has Maximum number of leads but conversion rate is less

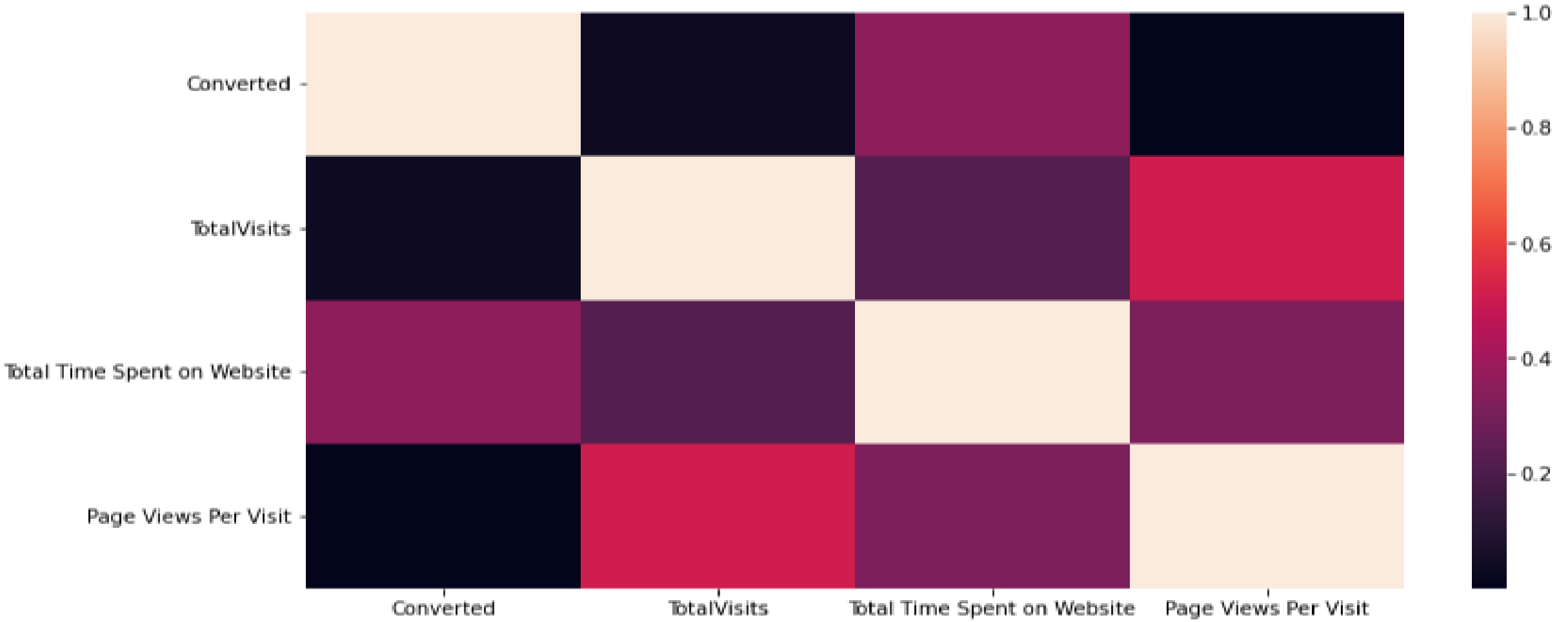
EDA



City

Mumbai has max leads from and has good conversion rate

EDA



Correlation Matrix

Data Conversion

- Numerical Variables Normalized
- Dummy Variables created for Object data type

Model Building

- Data Split into Test and Train Set with 70:30 Ratio
- Use RFE Feature Scaling
- Running RFE Feature Scaling with 15 Variables
- Model Building and removing Variable with high p-value and high VIF
- Predictions on test Set

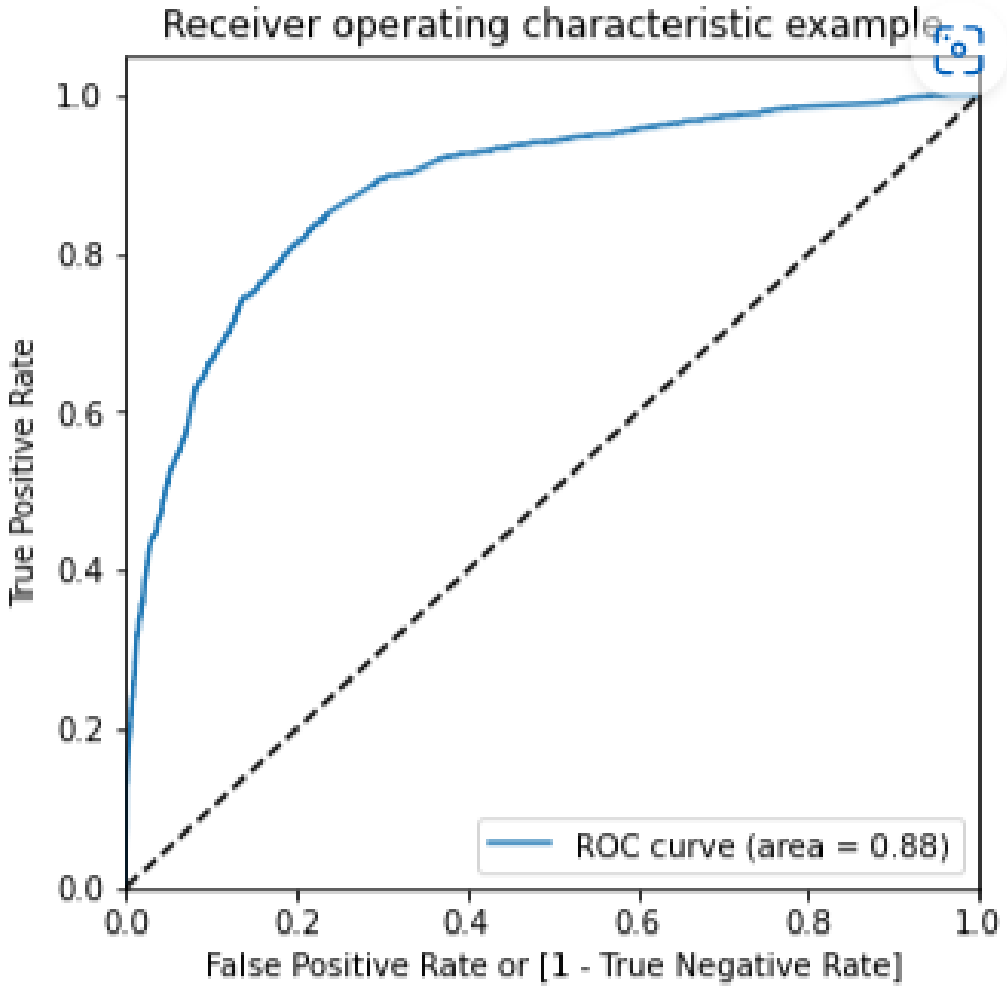
Result

Calculating VIF

Each value of VIF is good for further proceeding for making further predictions using this model.

	Features	VIF
0	TotalVisits	2.49
6	Last Activity_Olark Chat Conversation	1.94
10	Last Notable Activity_Modified	1.94
1	Total Time Spent on Website	1.92
9	Last Notable Activity_Email Opened	1.68
3	Lead Source_Olark Chat	1.67
2	Lead Origin_Lead Add Form	1.50
11	Last Notable Activity_Olark Chat Conversation	1.35
4	Lead Source_Welingak Website	1.34
7	What is your current occupation_Working Profes...	1.17
12	Last Notable Activity_Page Visited on Website	1.15
5	Do Not Email_Yes	1.11
8	Last Notable Activity_Email Link Clicked	1.06

ROC Curve



Area under the curve is 0.88 which is nice.

Classifiers that give curves closer to the top-left corner indicate a better performance.

The optimum cut off value in ROC curve is used to find the accuracy, sensitivity and specificity which came to be around 80% each.

Performance Measure

From this curve, 0.35 is the optimum point to take it as a cutoff probability.

Inference:

So the model seems to work correctly. And the ROC curve value is 0.88. We have the following values for the Train **Data**:

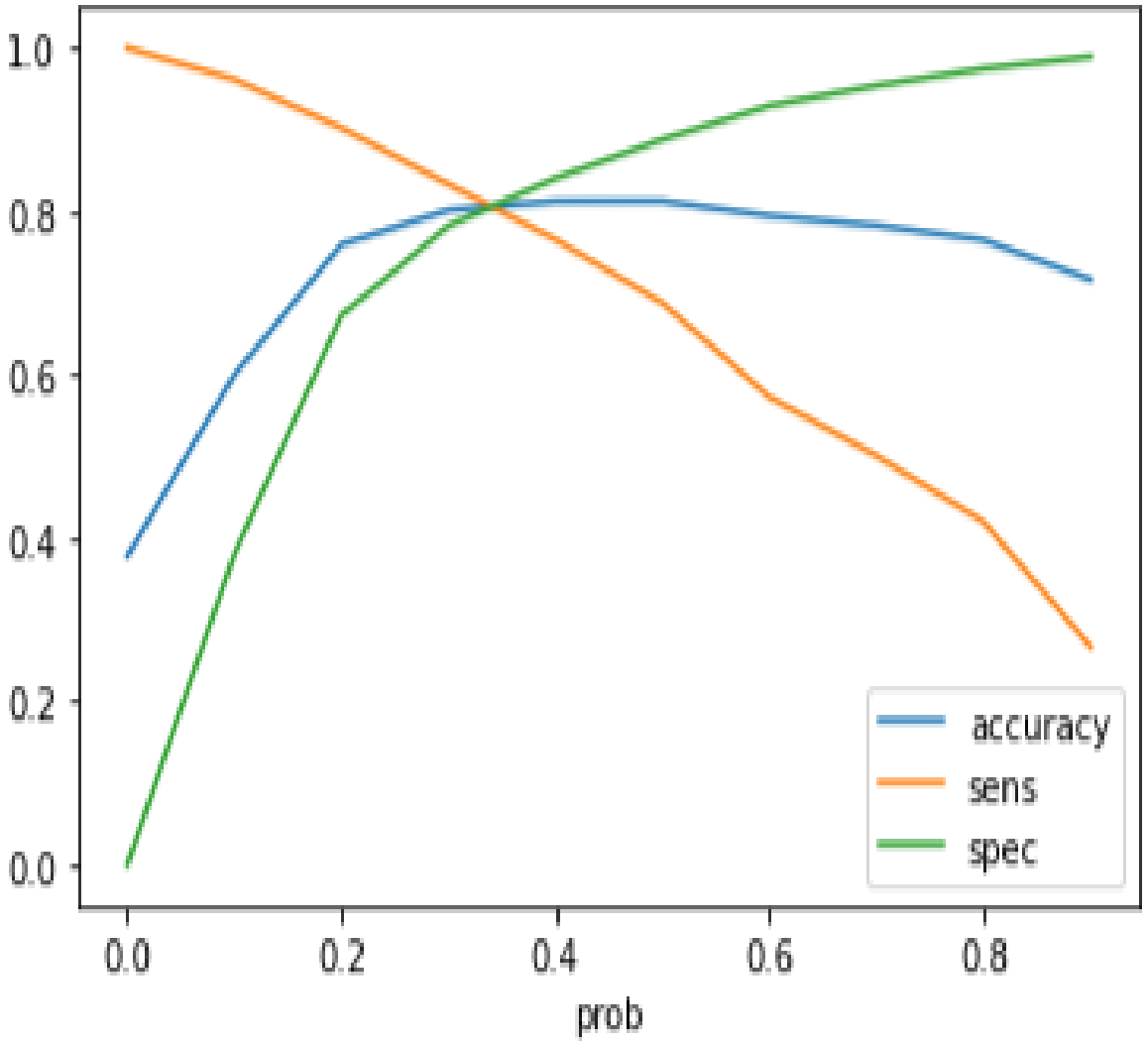
Accuracy : 80.74%

Sensitivity :79.98%

Specificity : 81.20%

Confusion Matrix:

3179	736
476	1902



Conclusion

Prediction:

Prediction was done on the test data frame and with an optimum cut off as 0.35 with accuracy, sensitivity and specificity all around 80%.

Precision – Recall:

This method was also used to recheck and a cut off of 0.4 was found with Precision around 73% and recall around 76% on the test data frame. The features that have affect the conversion rate at the maximum are as follows:

Conclusion

It was observed that the important variables in the selecting potential buyers are:

1. Total Visits
2. Last Activity_Olark Chat Conversation
3. Last Notable Activity_Modified
4. Total Time Spent on Website
5. Last Notable Activity_Email Opened
6. Lead Source_Olark Chat
7. Lead Origin_Lead Add Form
8. Last Notable Activity_Olark Chat Conversation¶
9. Lead Source_Welingak Website
10. What is your current occupation_Working Profes...
11. Last Notable Activity_Page Visited on Website
12. Do Not Email_Yes
13. Last Notable Activity_Email Link Clicked

Thank You