LEAD SCORING CASE STUDY

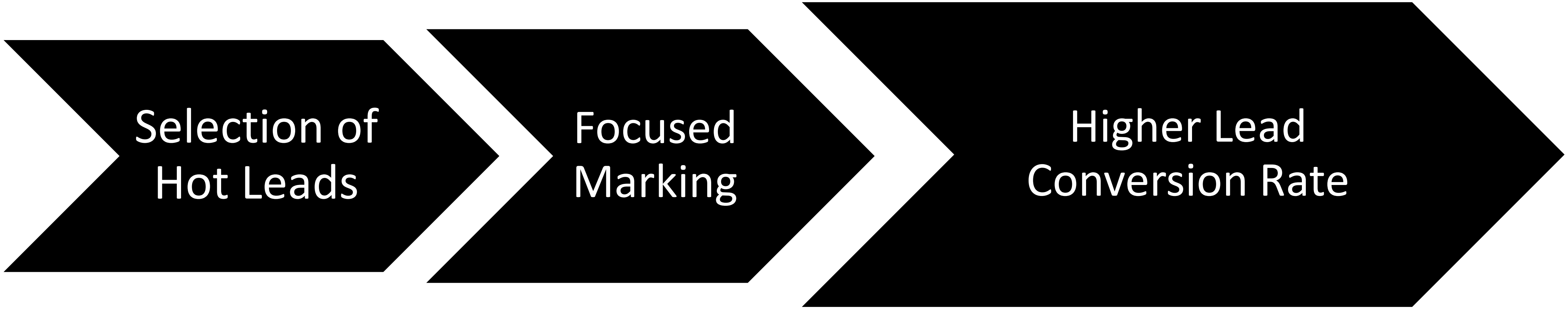
Heena Koushar
Prince Bhatt
Suraj Arun Ghag

PROBLEM STATEMENT

- X Education sell online courses to industry professional.
- X Education gets a lots of leads ,its conversion rate is very poor. For example, if, say they acquire 100 leads in a day, only about 30 of them are converted.
- In order to increase the lead conversion rate, the company first should identify the most potential leads, also knows as ‘Hot Leads’.
- If they successfully identify this set of lead conversion rate should rate go up as the sales team will now be focusing more on communicating with the potential leads rather than marking call to everyone.

Business Objective

To help X Education select most promising leads(Hots lead), i.e the leads that are most likely to convert into paying customer.



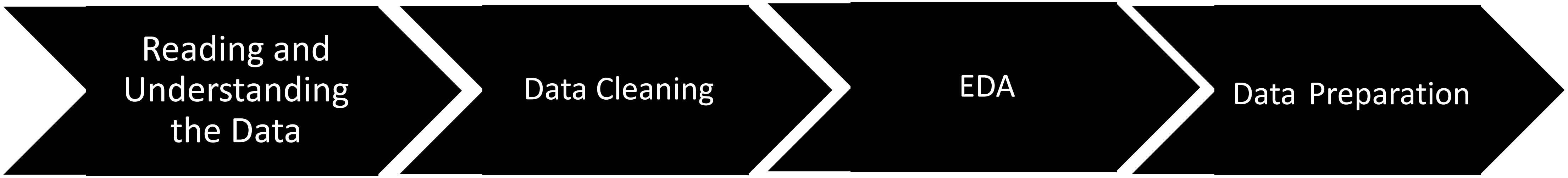
Selection of Hot Leads

Focused Marking

Higher Lead Conversion Rate

SOLUTION METHODOLOGY

- To build a Logistic Regression model that assigns lead scores to all leads such that the customers with higher lead score have a higher conversion chance and vice versa.
- Target Lead conversion Rate~80%



Importing and observing the data

Reading and Understanding the Data

Data Cleaning

Missing value imputation, removing data and other redundancy
Outlier treatment

Univariate and Bivariate Analysis

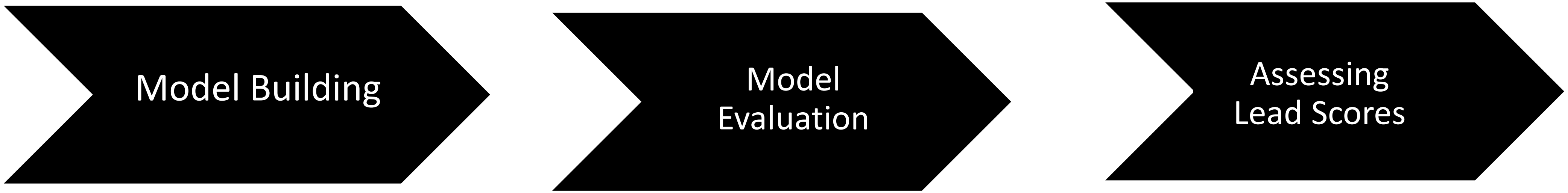
EDA

Data Preparation

- Dummy variable creation
- Train-Test split
- Scaling

- Feature selection using RFE.
- Manual feature elimination based on P-value and VIFs

- Finalizing the first model
- Using predicted probabilities to calculate Lead Scores.
  Lead Score = probability*100



Model Building

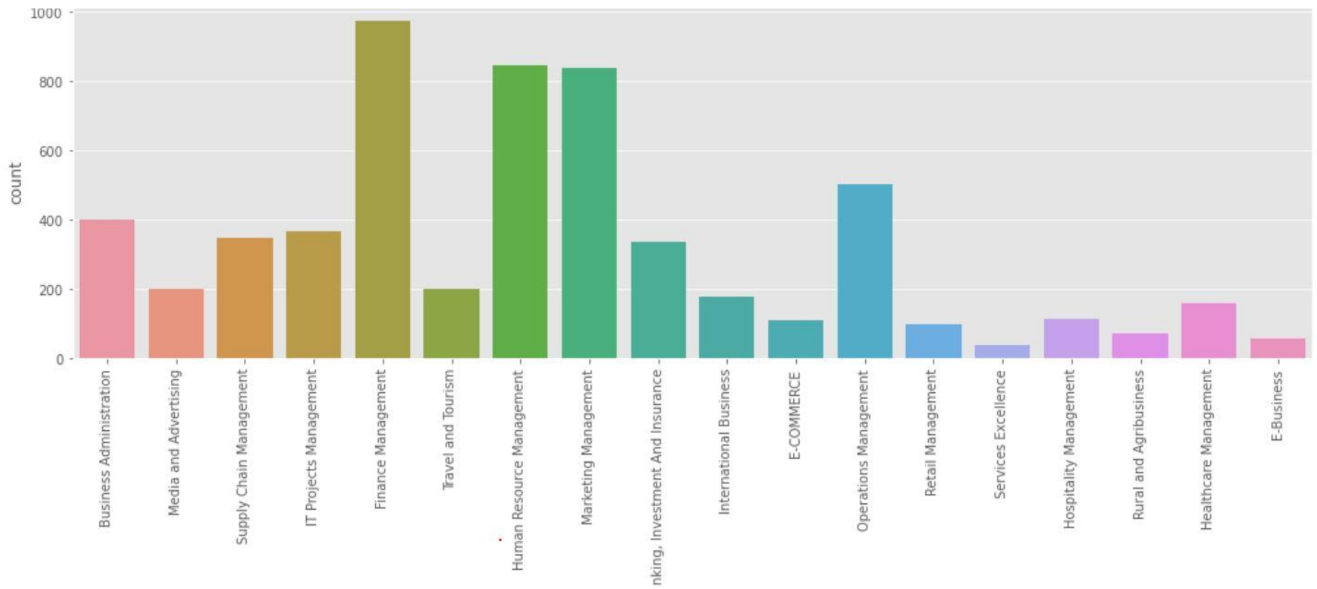
Model Evaluation

Assessing Lead Scores

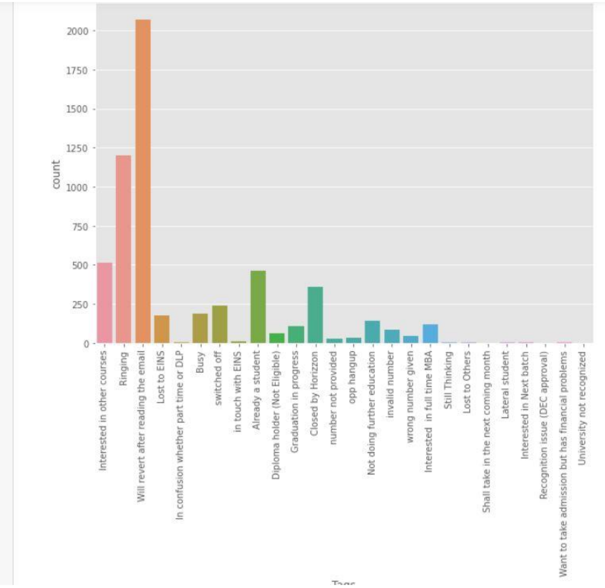
- Evaluation model based on various evaluation metrics.
- Finding the optimal probability threshold

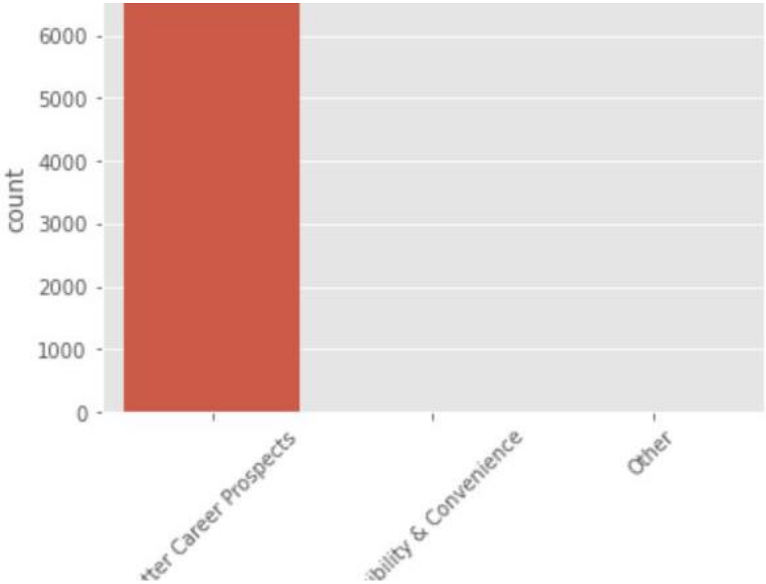
Specialization

There is 37% missing values present in the Specialization column .It may be possible that the lead may leave this column blank if he may be a student or not having any specialization or his specialization is not there in the options given. So we can create another category 'Others' for this.



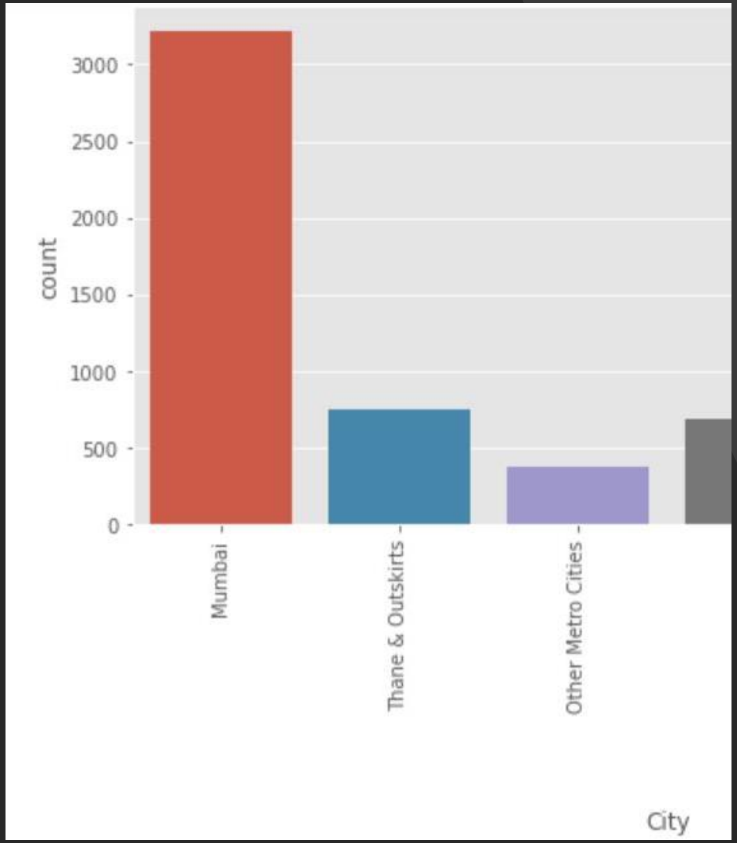
Visualizing Column



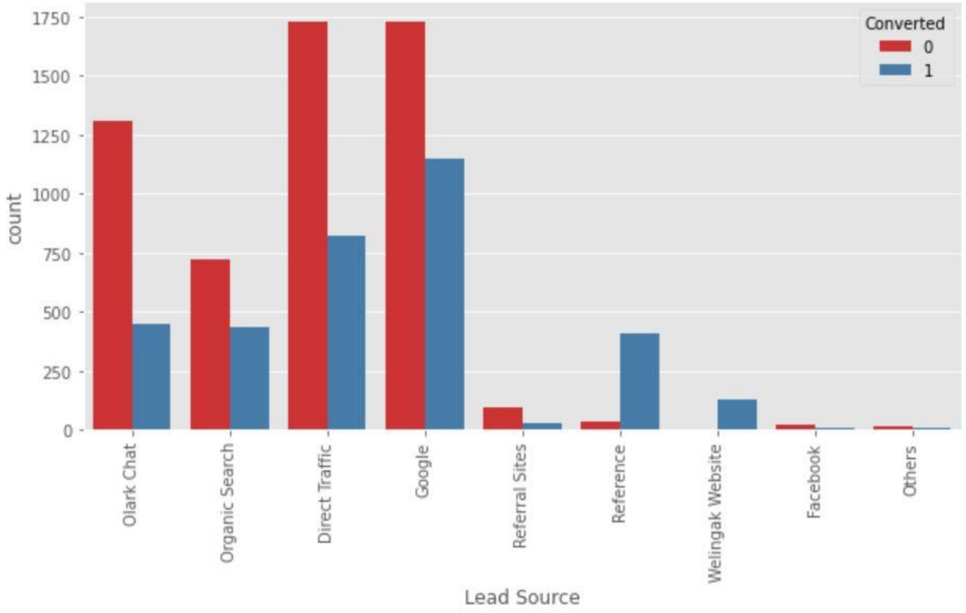


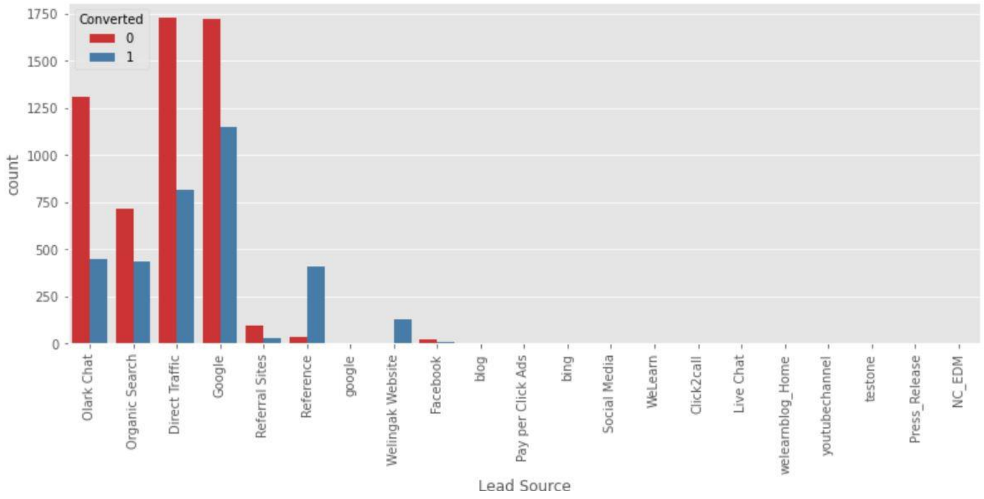
Exploratory Data Analysis

This City column has 40% missing value



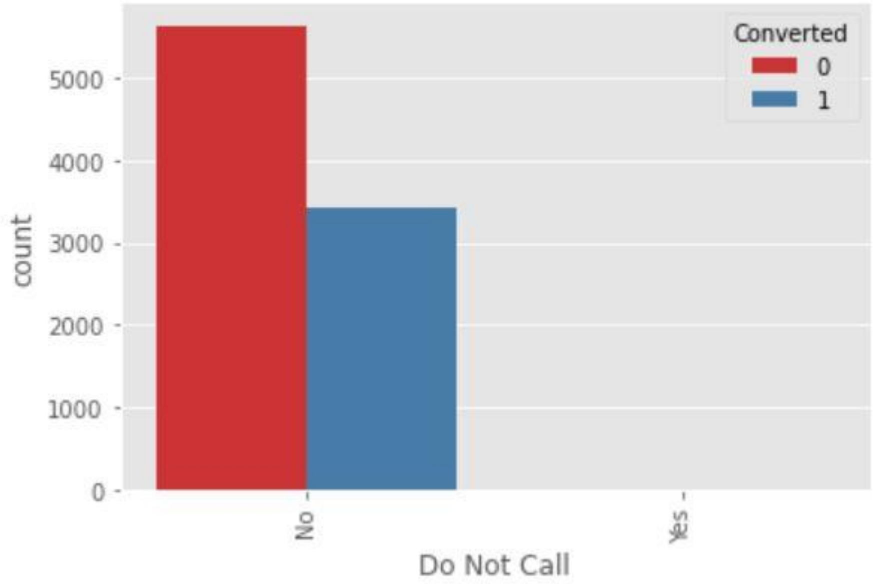
Lead Source

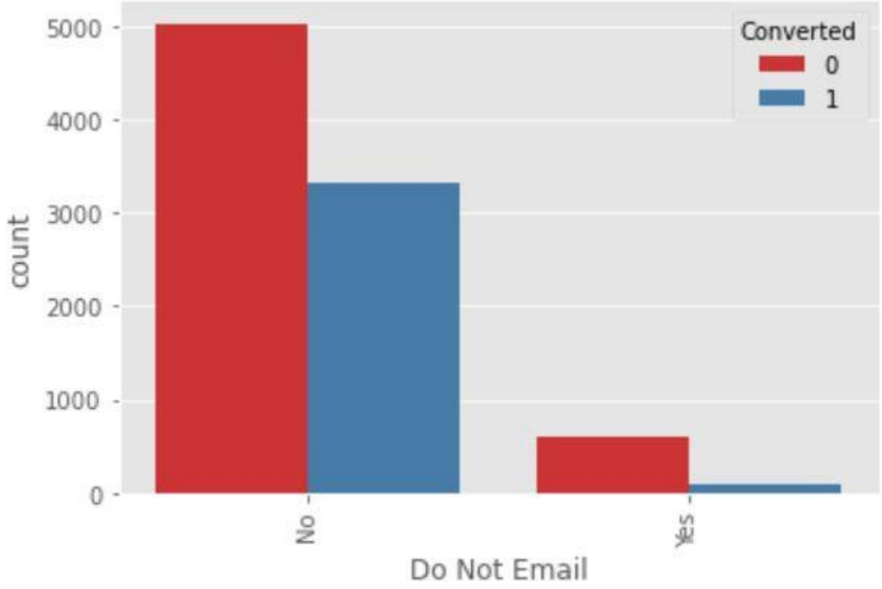

Converted
0
1
count
1750
1500
1250
1000
750
500
250
0
Olark Chat
Organic Search
Direct Traffic
Google
Referral Sites
Reference
Welingak Website
Facebook
Others
Lead Source


Converted
0
1
count
1750
1500
1250
1000
750
500
250
0
Olark Chat
Organic Search
Direct Traffic
Google
Referral Sites
Reference
google
Welingak Website
Facebook
blog
Pay per Click Ads
bing
Social Media
WeLearn
Click2call
Live Chat
welearnblog_Home
youtubechannel
testone
Press_Release
NC_EDM
Lead Source

Do not call

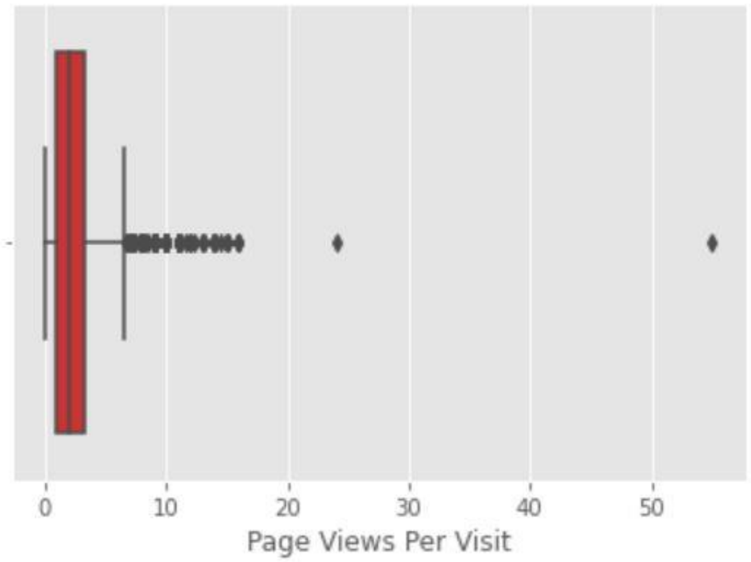
Do not Email

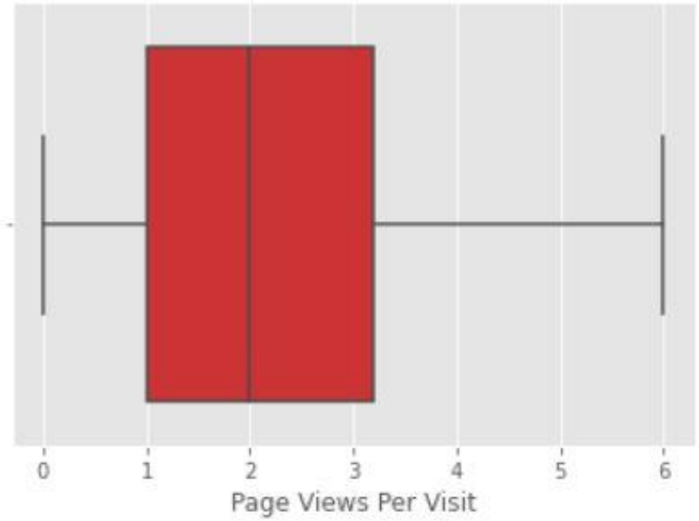




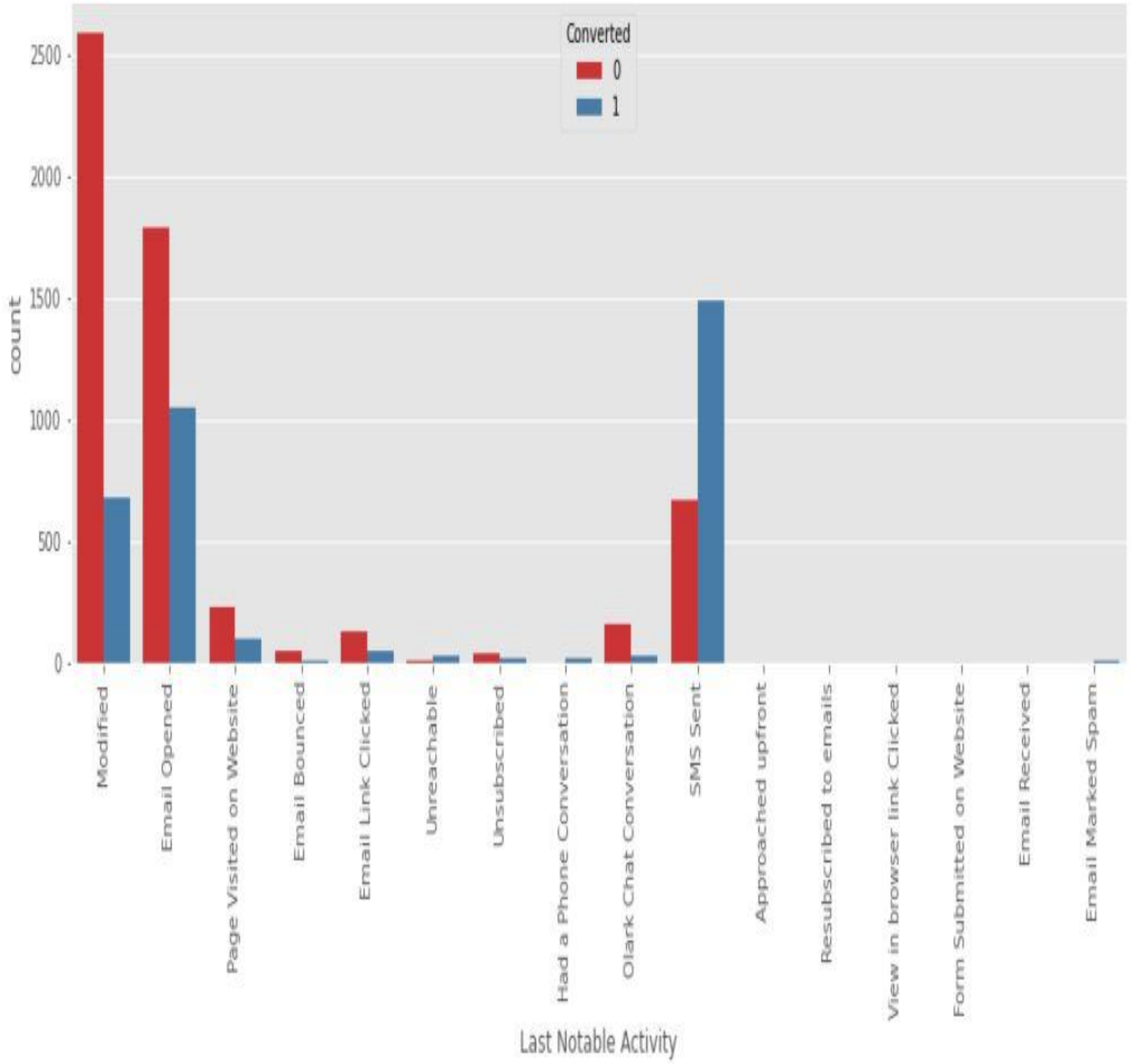
Page View Per Visit

As we can see there are a number of outliers in the data. We will cap the outliers to 95% value for analysis.

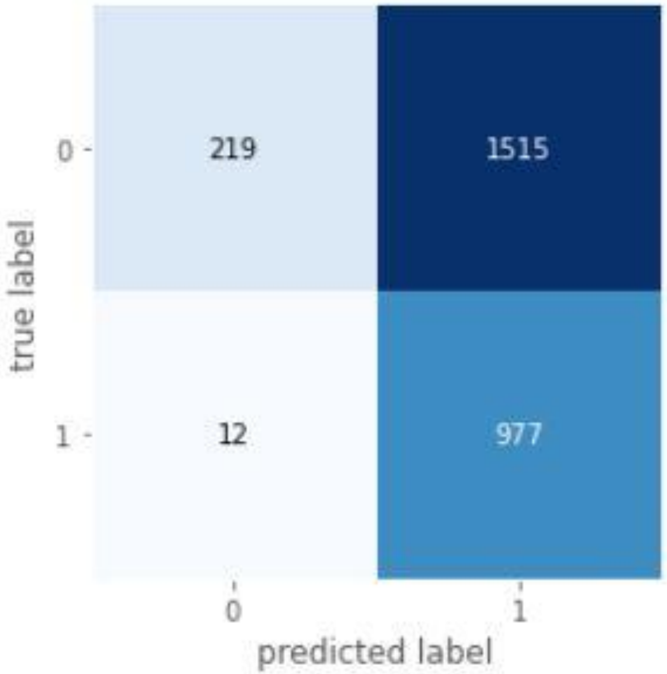




Last Notable Activity



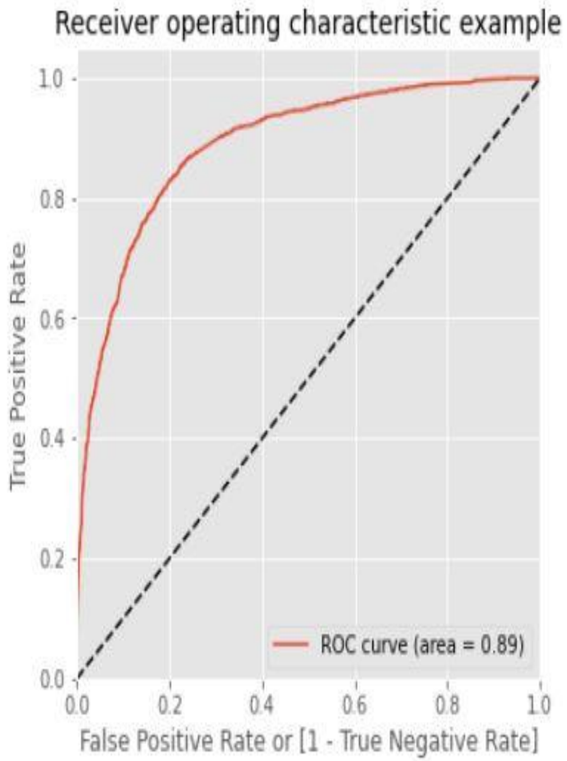
Confusion Matrix



true label \ predicted label	0	1
0	219	1515
1	12	977

The ROC Curve

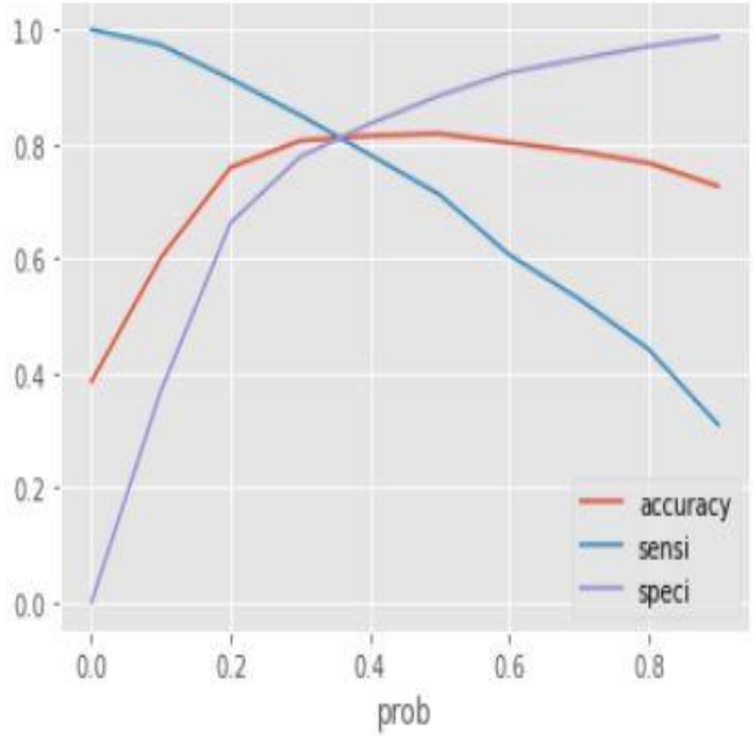
It shows the tradeoff between sensitivity and specificity (any increase in sensitivity will be accompanied by a decrease in specificity). The closer the curve follows the left-hand border and then the top border of the ROC space, the more accurate the test. The closer the curve comes to the 45-degree diagonal of the ROC space, the less accurate the test.



Since we have higher (0.89) area under the ROC curve , therefore our model is a good one.

Optimal Cutoff Point

Above we had chosen an arbitrary cut-off value of 0.5. We need to determine the best cut-off value and the below section deals with that. Optimal cutoff probability is that prob where we get balanced sensitivity and specificity



From the curve above, 0.34 is the optimum point to take it as a cutoff probability.

Lead Score

Train Data

- Accuracy : 81.1 %
- Sensitivity : 82.5 %
- Specificity : 80.2 %

Test Data

- Accuracy : 80.1 %
- Sensitivity : 80.6 %
- Specificity : 79.8 %

Recommendation

The company should make calls to the leads coming from the lead sources "Welingak Websites" and "Reference" as these are more likely to get converted. The company should make calls to the leads who are the "working professionals" as they are more likely to get converted. The company should make calls to the leads who spent "more time on the websites" as these are more likely to get converted. The company should make calls to the leads coming from the lead sources "Olark Chat" as these are more likely to get converted. The company should make calls to the leads whose last activity was SMS Sent as they are more likely to get converted.

The company should not make calls to the leads whose last activity was "Olark Chat Conversation" as they are not likely to get converted.

The company should not make calls to the leads whose lead origin is "Landing Page Submission" as they are not likely to get converted. The company should not make calls to the leads whose Specialization was "Others" as they are not likely to get converted. The company should not make calls to the leads who chose the option of "Do not Email" as "yes" as they are not likely to get converted.