LEAD SCORING CASE STUDY

(CAPSTONE PROJECT)

-ADARSH R

OBJECTIVE

To build a logistic regression model to assign a lead score between 0 and 100 to each of the leads which can be used by the company to target potential leads. A higher score would mean that the lead is hot, i.e. is most likely to convert whereas a lower score would mean that the lead is cold and will mostly not get converted.

APPROACH

Importing libraries

Reading and understanding the dataset

Data cleaning

Exploratory Data Analysis

Feature scaling

Splitting the data into train and test datasets

Data modelling

Model building

Feature selection using RFE

Plotting ROC curve

Model evaluation

EXPLORATORY DATA ANALYSIS

EDA is done for the following areas,

Lead origin

Lead source

Do not email and do not call

Total visits

Total time spent on the website

Page views per visit

Last activity

Country

Specialization

Occupation.

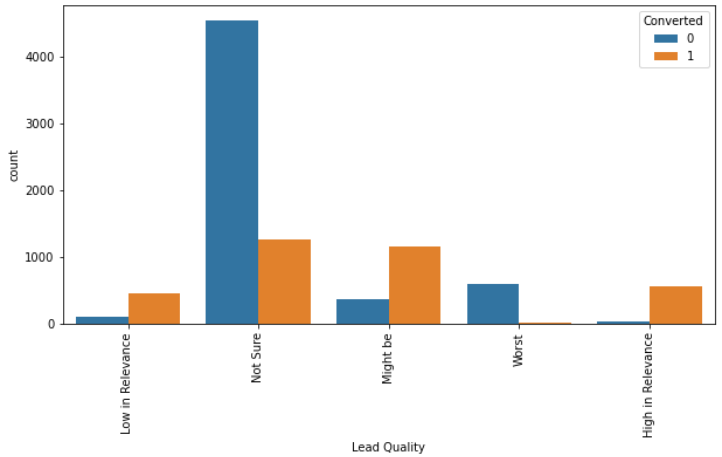
EXPLORATORY DATA ANALYSIS

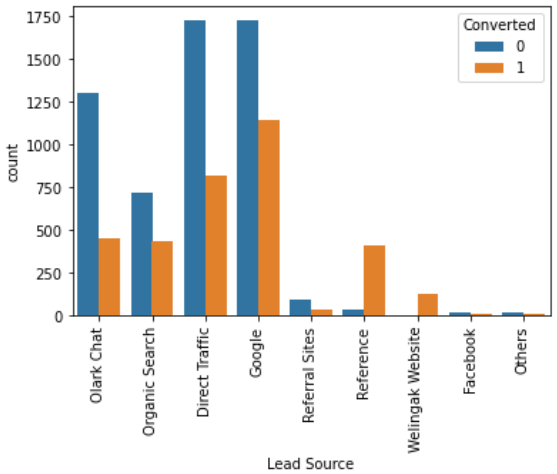
- Quick check was done on % of null value and dropped columns with more than 45% missing values.

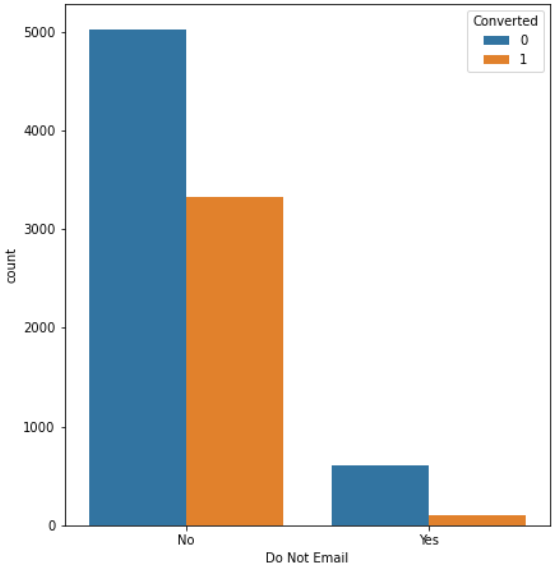
- The rows with the null value would cost us a lot of data and they were important columns. So, instead we replaced the NaN values with 'not provided'.

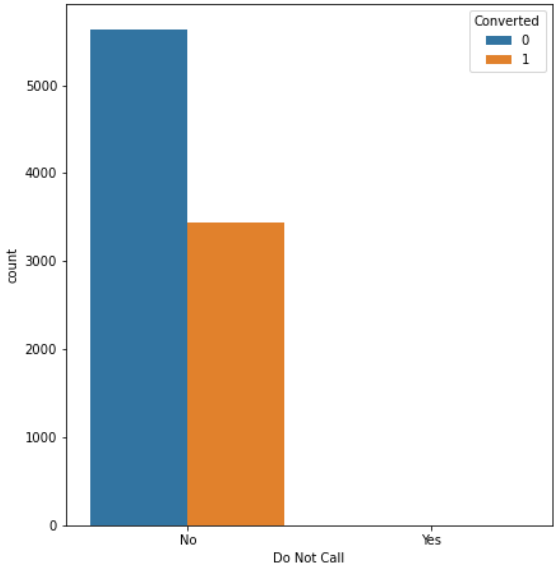
- Also worked on numerical variable, outliers and dummy variables.

LEAD QUALITY









TRAIN-TEST SPLIT AND SCALING

- The split was done at 70% and 30% for train and test data respectively.

- Did min-max scaling on the variables [‘Do not email and do not call’, 'Total Visits', 'Page Views Per Visit', 'Total Time Spent on Website', ‘Country’, ‘Last activity’]

MODEL BUILDING

- RFE was used for feature selection.

- Then RFE was done to attain the top 15 relevant variables.

- Later the rest of the variables were removed manually depending on the VIF values and p-value.

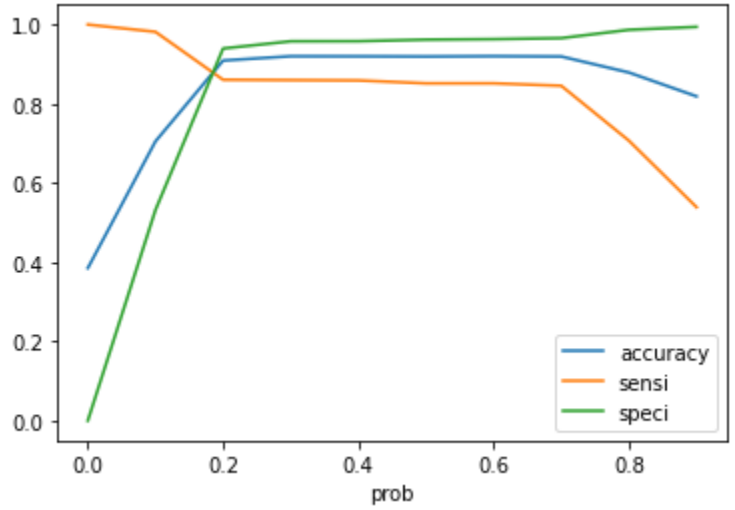
- A confusion matrix was created, and overall accuracy was checked which came out to be 91.93%.

MODEL EVALUATION

The optimum cut off value was found using ROC curve. The area under ROC curve was 0.95.

After Plotting it is found that optimum cutoff was 0.2 which gave

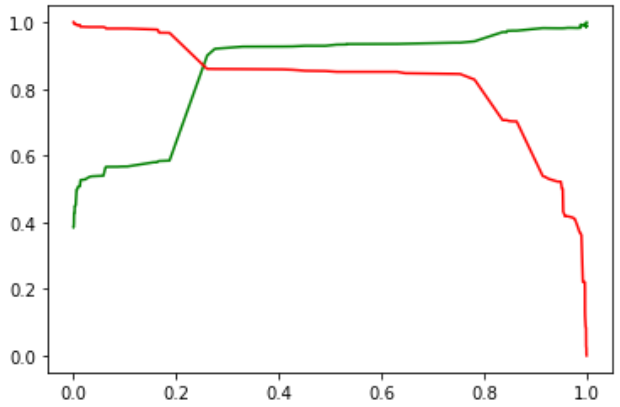
- Accuracy 91.93%
- Sensitivity 86.05%
- Specificity 93.95%.



MODEL EVALUATION

- With the cutoff of 0.35 got the Precision & Recall of 79.29% & 70.22% respectively.

- So, to increase the above percentage need to change the cut off value. After plotting found the optimum cut off value of 0.44 which gave
  - Accuracy 85.80%
  - Precision 93.32%
  - Recall 85.15%



So, with Sensitivity-Specificity Evaluation the optimal cut off value would be 0.25 and with Precision - Recall Evaluation the optimal cut off value would be 0.34

SUMMARY

X Education Company needs to focus on following key aspects to improve the overall conversion rate:

- Increase user engagement on their website since this helps in higher conversion.

- Increase on sending SMS notifications since this helps in higher conversion.

- Get Total Visits increased by advertising etc. since this helps in higher conversion.

- Checked with both sensitivity and specificity as well as precision and recall metrices and calculated the optimal cut off for the final prediction