LEADS SCORING CASE STUDY

MR. SHAHABAJ SHAIKH

MS. SNEHAL HOLE

MR. ANKIT CHATURVEDI

PROBLEM STATEMENT

- An education company named 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 these people land on the website, they might browse the courses or fill up a form for the course or watch some videos. When these people fill up a form providing their email address or phone number, they are classified to be a lead. Moreover, the company also gets leads through past referrals. Once these leads are acquired, 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%.
- Now, although X Education gets a lot of leads, its lead conversion rate is very poor. For example, if, say, they acquire 100 leads in a day, only about 30 of them are converted. To make this process more efficient, the company wishes to identify the most potential leads, also known as ‘Hot Leads’. If they successfully identify this set of leads, the lead conversion rate should go up as the sales team will now be focusing more on communicating with the potential leads rather than making calls toeveryone.
- The company 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, in particular, has given a ballpark of the target lead conversion rate to be around 80%.

Goal of the Case Study

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.

- **Steps Involved In Solving The Case Study**:
  - Importing and reading the dataset.
  - Inspecting the dataset.
  - Data preparation.
  - Dummy variable creation.
  - Train-Test split.
  - Feature scaling.
  - Model building.
  - Plotting ROC curve.
  - Finding optimal cutoff point.
  - Making predictions on the test data

Importing and Reading the Dataset:

Importing the necessary libraries.
Importing the “Leads.csv” file into python

Inspecting the Dataset:

- Checking the head of the data frame.
- Checking the shape of the data frame.
- Checking the statistical aspect of the data frame
- Checking the basic info of each column

DATA PREPARATION:

- Replacing the level “Select” in categorical variables with NaN values.
- Checking the percentage of missing values in each column.
- Dropping redundant columns and columns having missing value percentage greater than 40%.
- Replacing NaN values in certain columns with mode/new category values.
- Dropping Rows having Nan values in very small proportion.
- Dropping columns which had a particular level representing the categorical variable majorly.
- Checking for outliers using percentiles.
- Capping and flooring method used to treat outliers

DUMMY VARIABLE CREATION :

Mapping Binary variables with two levels(yes & no) to 1's and 0's.

Creating dummy variables for categorical variables using pd.get_dummies.

Dropping categorical columns for which dummies have been created.

TRAIN-TEST SPLIT :

Importing train_test_split from sklearn.model_selection.

Putting all feature variables in X.

Putting Target variable in y.

Splitting the dataset into train and test set in the ratio of 70:30.

DUMMY VARIABLE CREATION

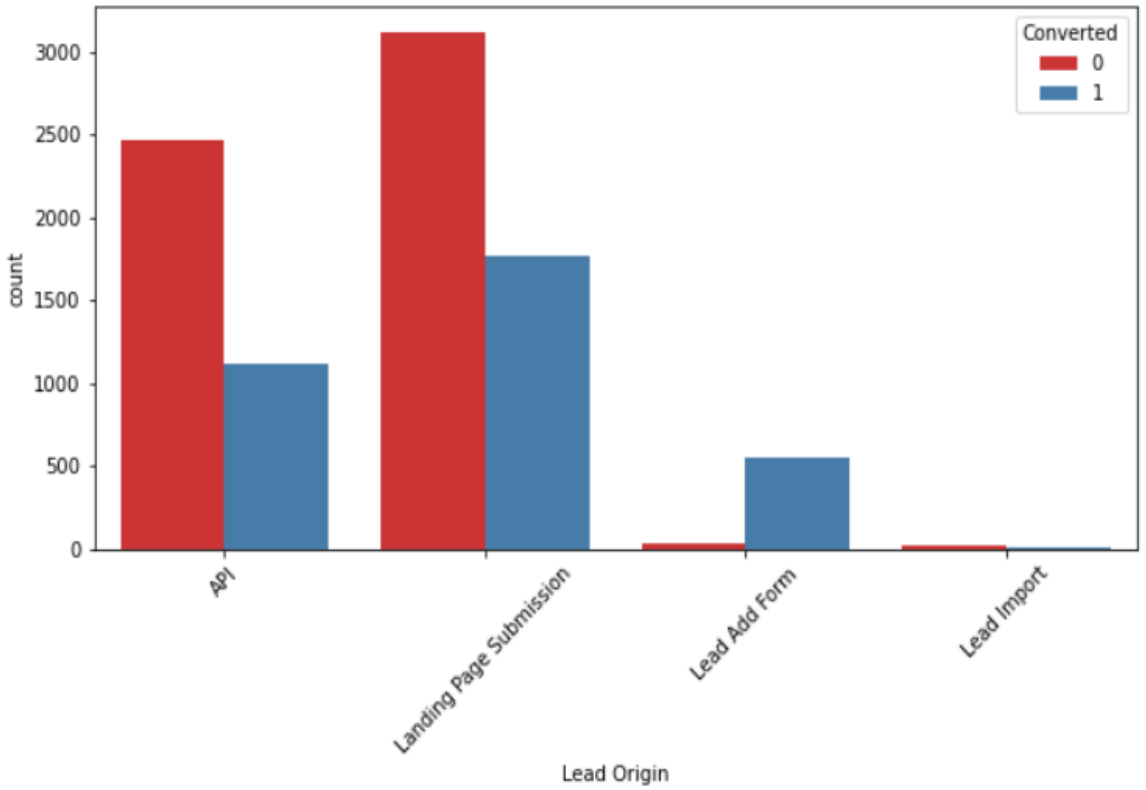
Using Standard scaler from sklearn.preprocessing.

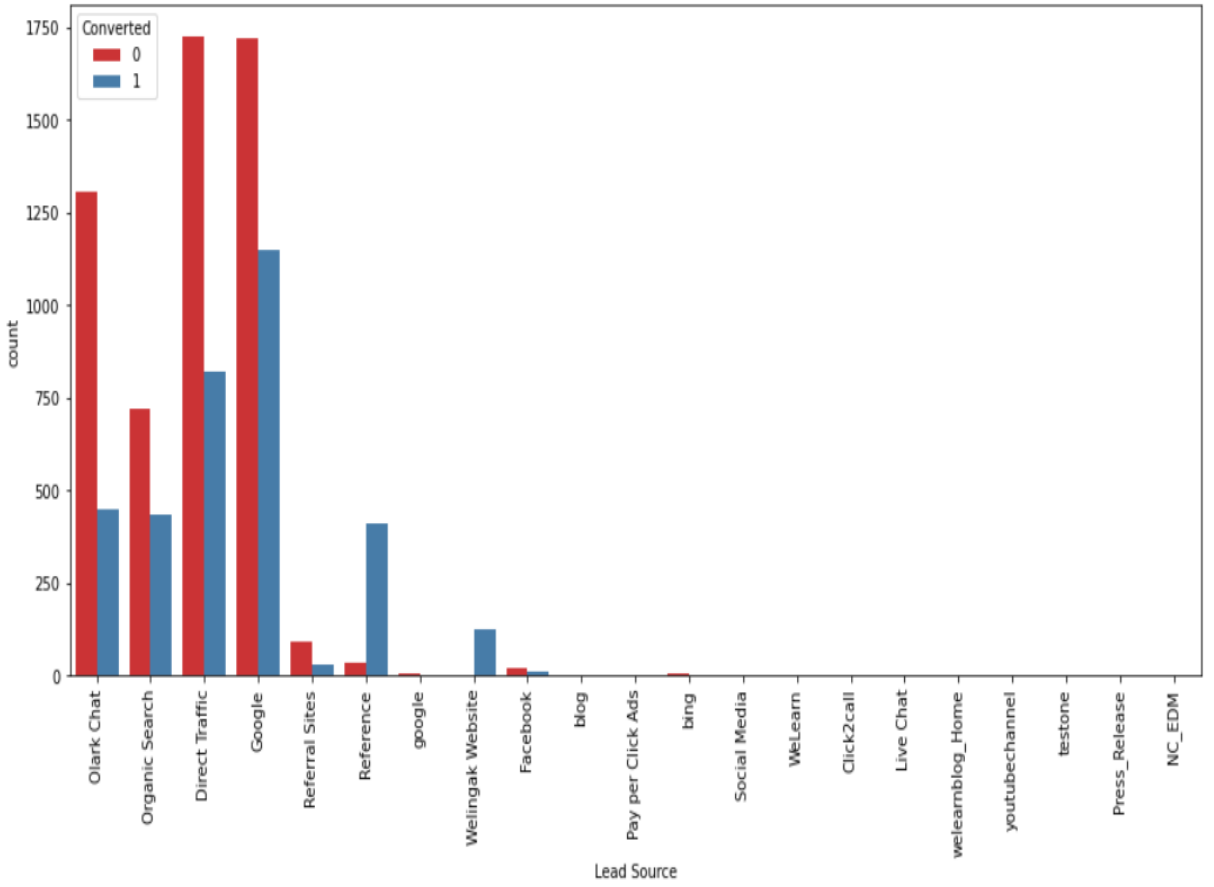
Scaling numerical features using Standard scaler

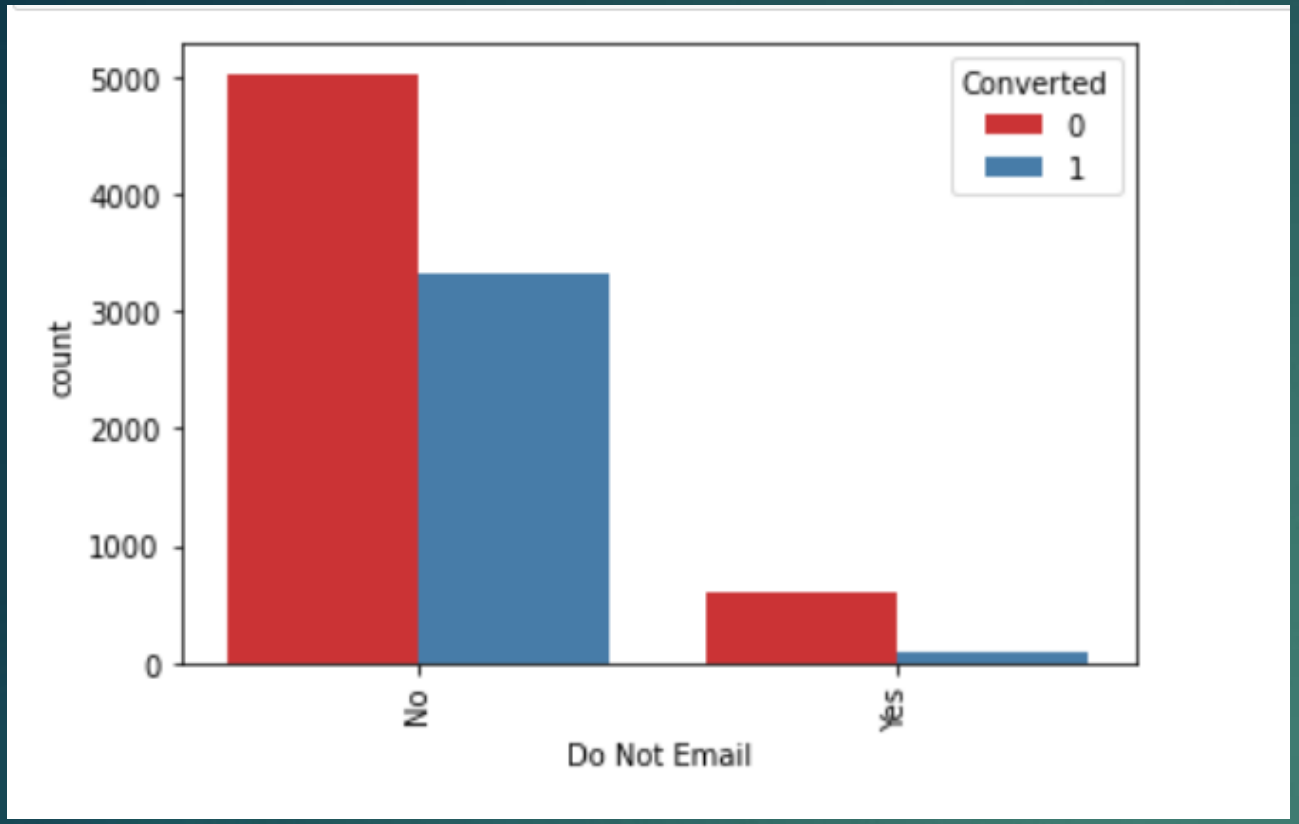
MODEL BUILDING :

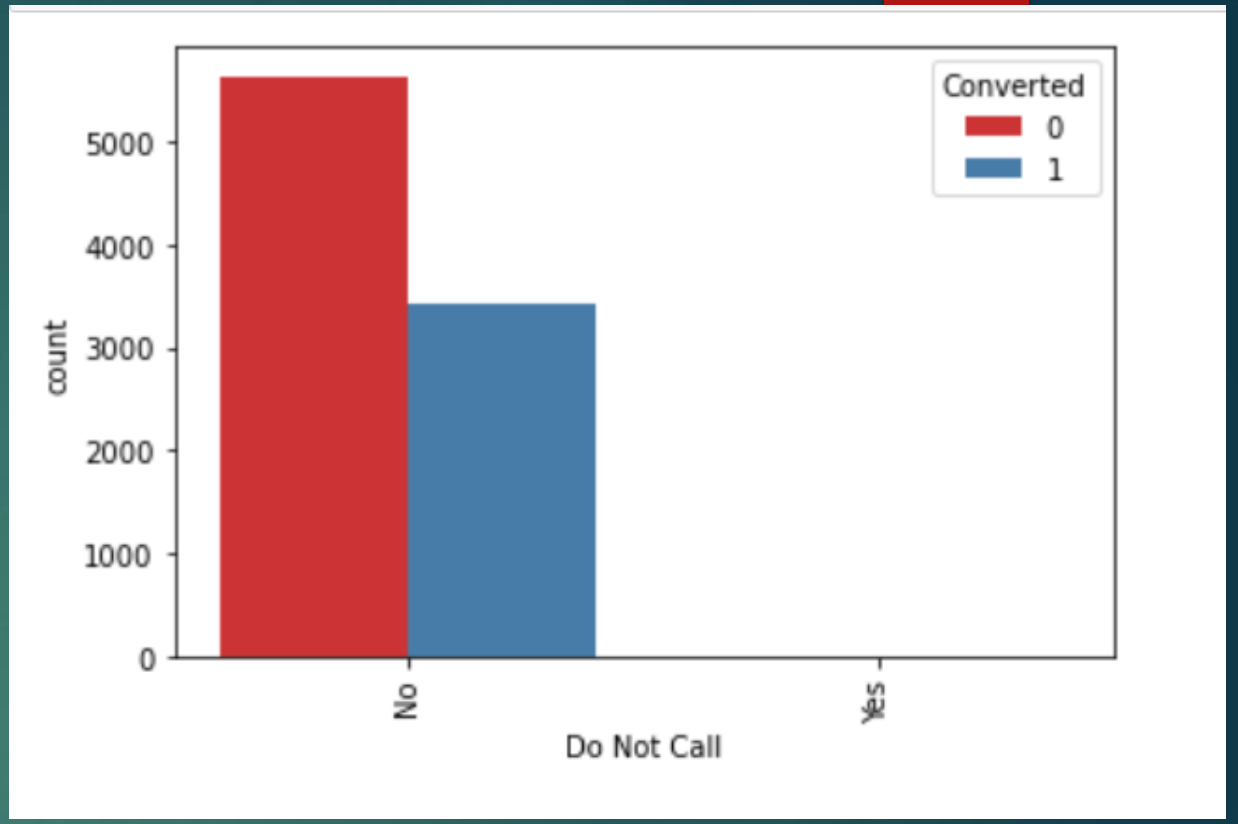
- Using RFE(recursive feature elimination) to select top 15 variables for our model.
- Building a model with the features selected by RFE using Stats Models.
- Dropping feature with a high p-value.
- Checking VIF value for features and dropping the feature with a high VIF value.
- We repeated the model building process till we arrived to a model with features having normal p-value and VIF values.
- Predicting values on the train set.
- Creating a confusion matrix.
- Checking accuracy, sensitivity and specificity.

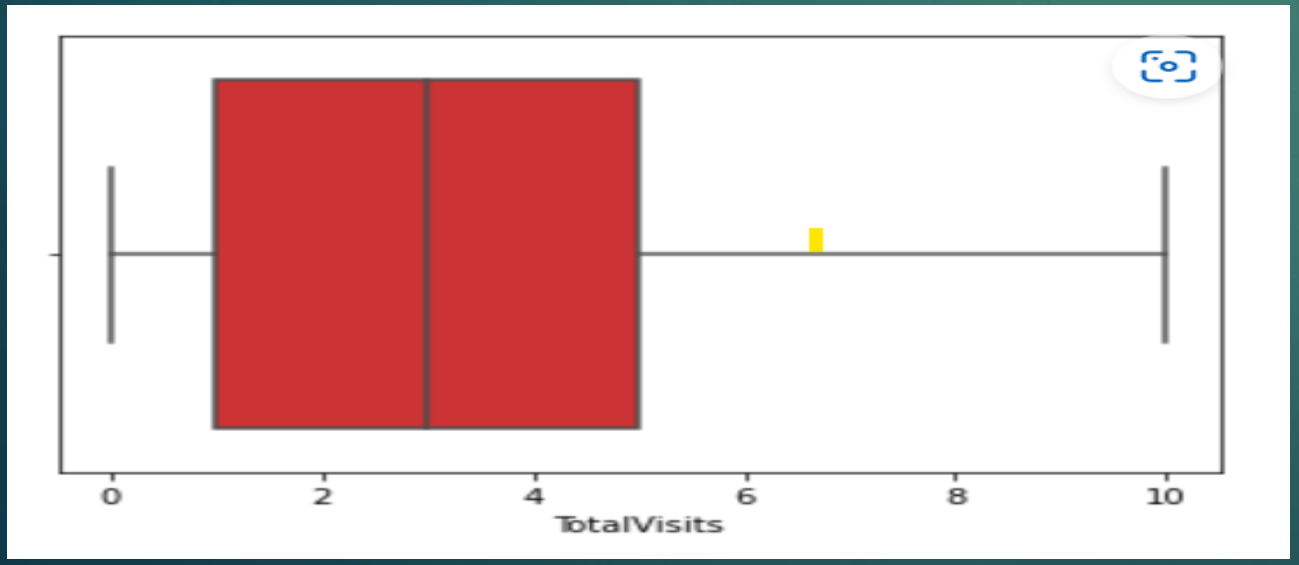
EDA

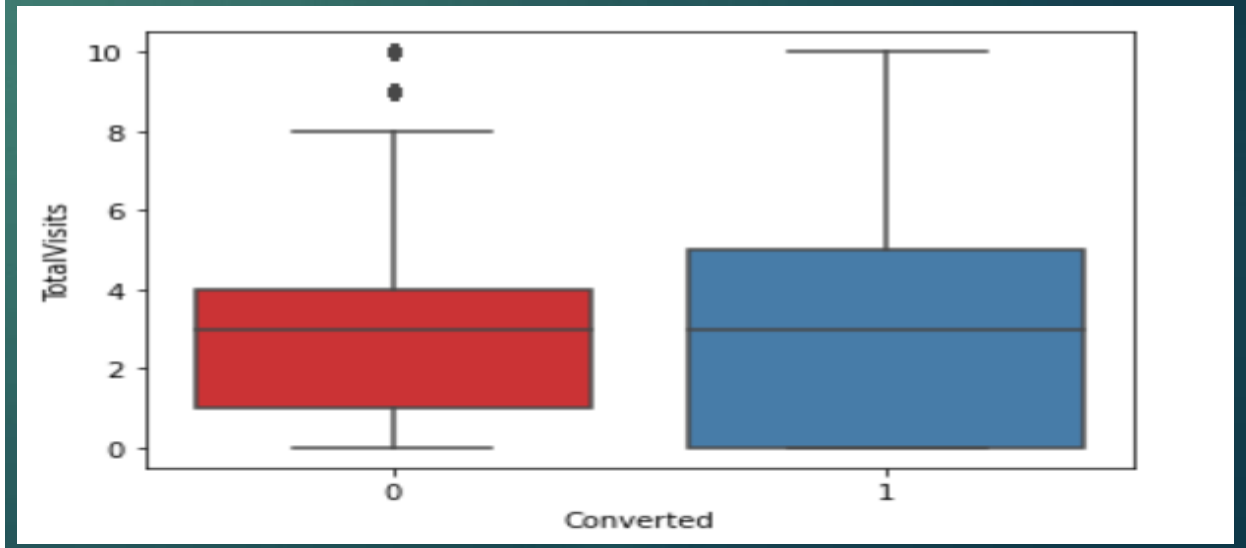

Converted
0
1
count
3000
2500
2000
1500
1000
500
0
API
Landing Page Submission
Lead Add Form
Lead Import
Lead Origin


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

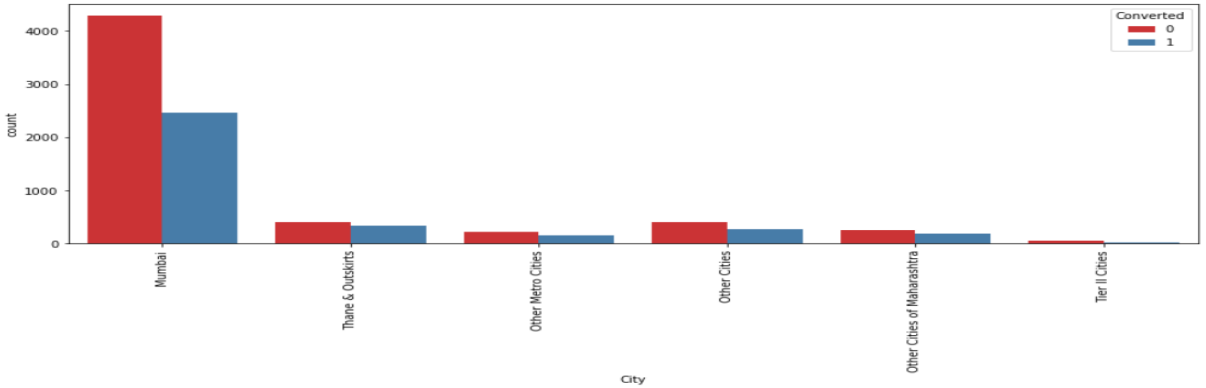

Converted
0
1
count
5000
4000
3000
2000
1000
0
No
Yes
Do Not Email

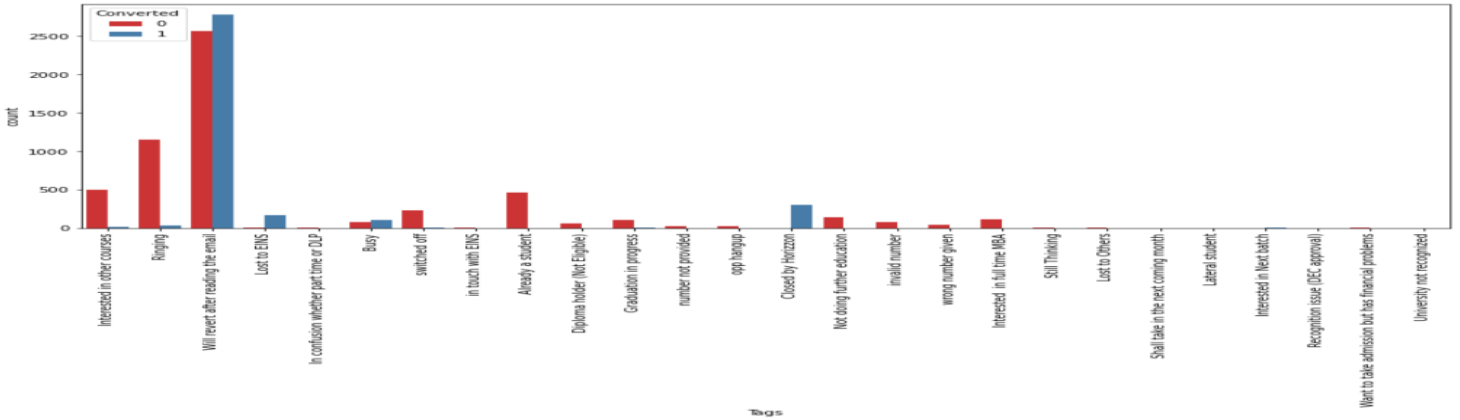

Converted
0
1
count
5000
4000
3000
2000
1000
0
No
Yes
Do Not Call

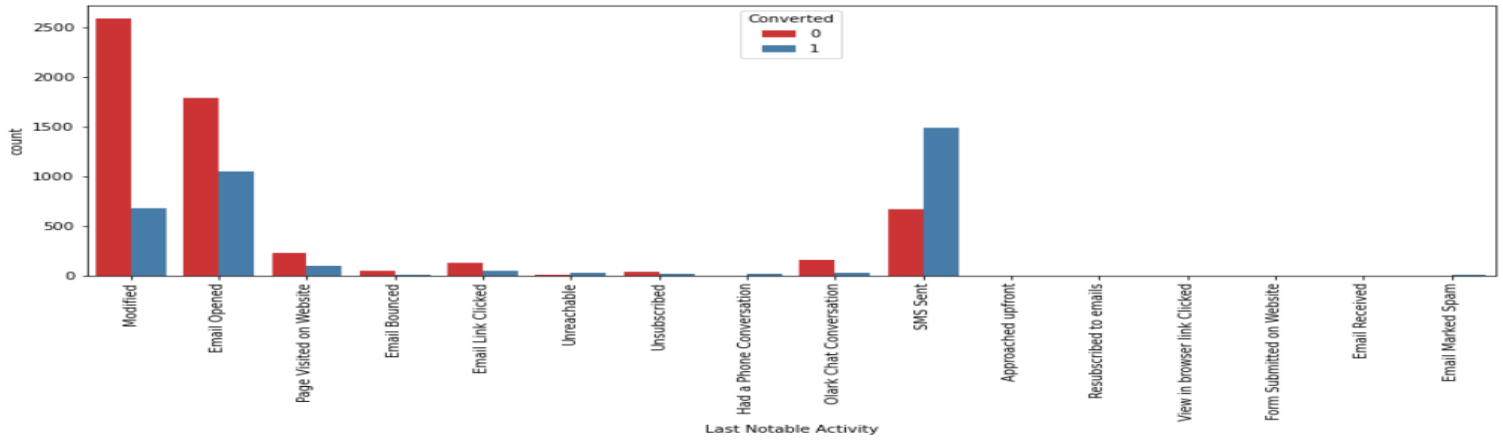

0
2
4
6
8
10
TotalVisits


TotalVisits
10
8
6
4
2
0
0
1
Converted

CATEGORICAL VARIABLE RELATION

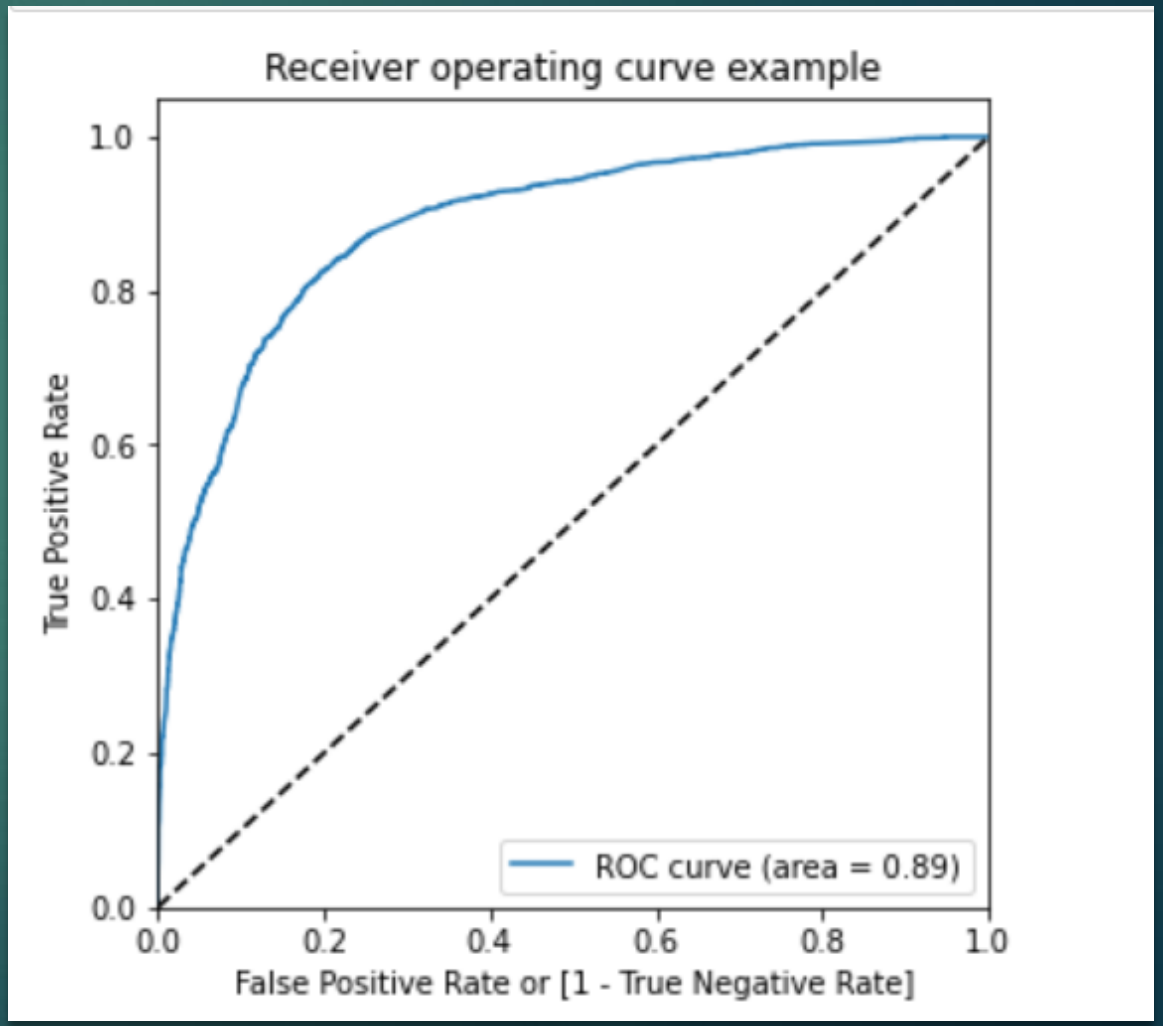






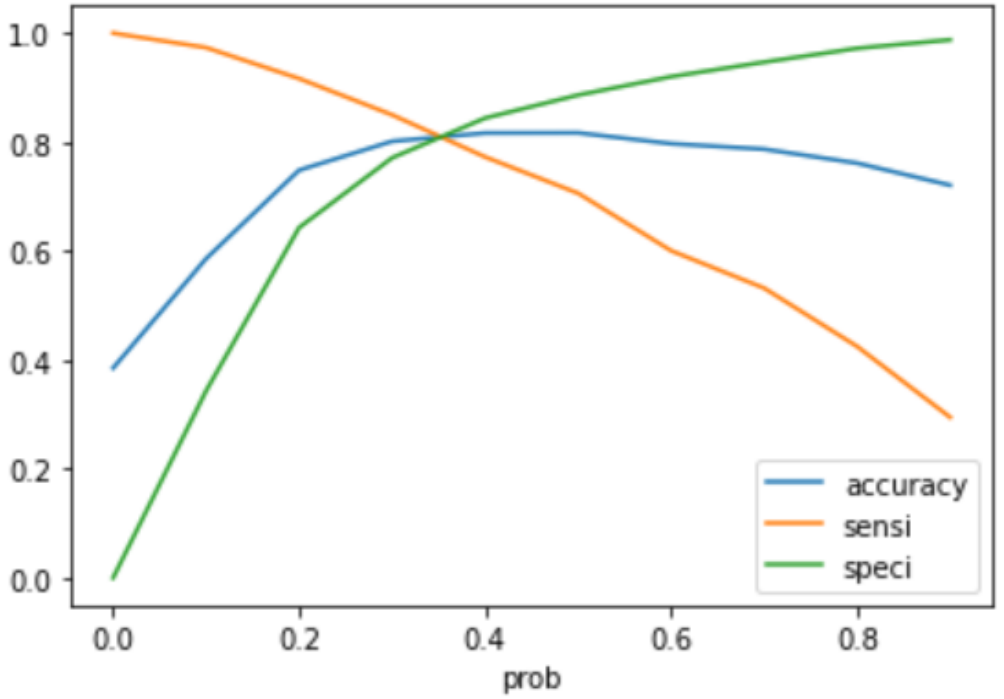
PLOTTING ROC CURVE:

- A Good ROC curve is the one which touches the upper left corner of the graph.
- we have a similar curve over here.
- Higher the area under the ROC curve the better is your model. The value of
- ROC curve should be closer to 1, we have the area under ROC curve = 0.89.



FINDING OPTIMAL CUT OFF POINT:

- From the plot we decided to take a cutoff point of 0.35.
- Checking accuracy after taking 0.35 cutoff.
- Creating a confusion matrix.
- Checking sensitivity, specificity and false positive rate.
- Checking precision and recall score



RESULT

- We have very close values for accuracy, sensitivity and specificity when comparing results from train and test sets.
- These values show that the model is performing well.

Train Data:

Accuracy : 81.0 %

Sensitivity : 81.7 %

Specificity : 80.6 %

Test Data:

Accuracy : 80.4 %

Sensitivity : 80.4 %

Specificity : 80.5 %

SUMMARY:

1.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.

2.The company should make calls to the leads who are the "working professionals" as they are more likely to get converted.

3.The company should make calls to the leads who spent "more time on the websites" as these are more likely to get converted.

4.The company should make calls to the leads coming from the lead sources "Olark Chat" as these are more likely to get converted.

5.The company should make calls to the leads whose last activity was SMS Sent as they are more likely to get converted.

THANK YOU