Lead Scoring Case Study

The Lead Scoring case study is based on providing insights to a X Education Company is looking suggestions where higher lead score have a higher conversion chance and the customers with a lower lead score have a lower conversion chance.

There were total 9240 data points and 37 columns.

Following approach was followed for the Case Study

1. Reading the data

The data was provided in a CSV file. Using Pandas Library, the data was read. The data set had a shape of (9240, 37).

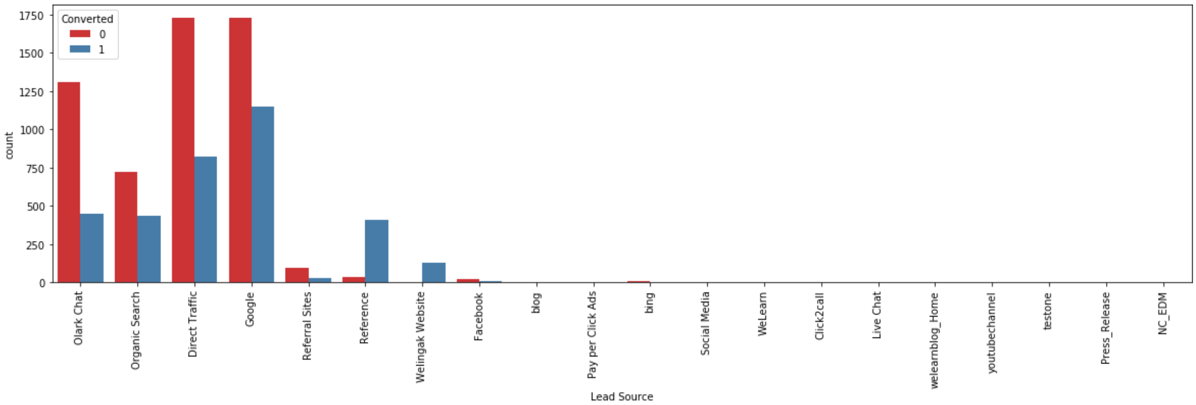
2. Inspecting the Data

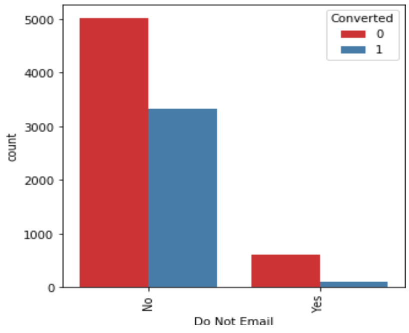
The data has NuLL values in 17 columns. In addition to NULL values there is SELECT which is also to be considered as NULL value.

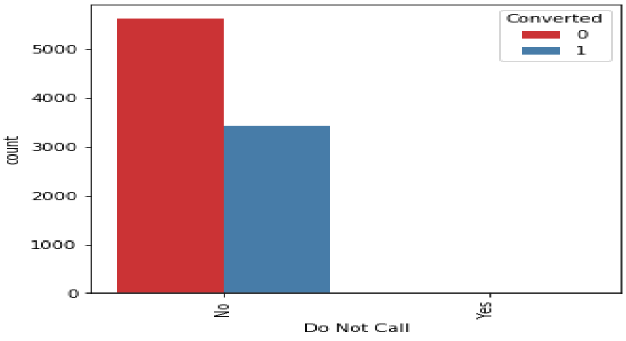
3. Data Cleaning

The 40% cut-off was used to delete column which had NULL values.
The outliers were handled were Q1 – 5% and Q3 – 95%.

4. Exploratory Data Analysis







Above are some of the Data Analysis which we done as part of EDA to understand the data.

Based on EDA, following are the recommendation

1] The majority of the leads' most recent action was **Email Opened**.

2] About 70% of leads with an **SMS sent** as their most recent activity converted.

5. Creating Dummy Variables

Dummy Variables were created for following columns

Lead Origin, Lead Source, Last Activity, Specialization, What is your current occupation, Last Notable Activity

6. Splitting the data set to Test and Train

train_size = 0.7, test_size = 0.3

7. Scaling of Data

Columns which had continuous data were scaled. StandardScaler techniques was used for scaling the parameters.

8. RFE

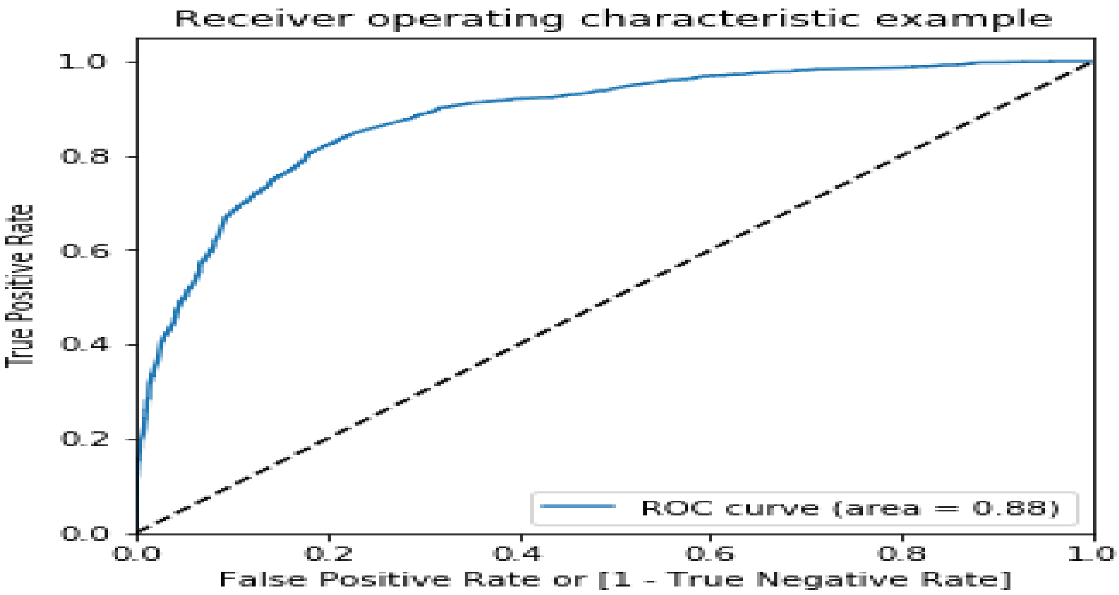
RFE was used for selecting the variable which can have the maximum impact in model building. This variable were later tested for P-Value and VIF, and accordingly dropped.

9. Model Building

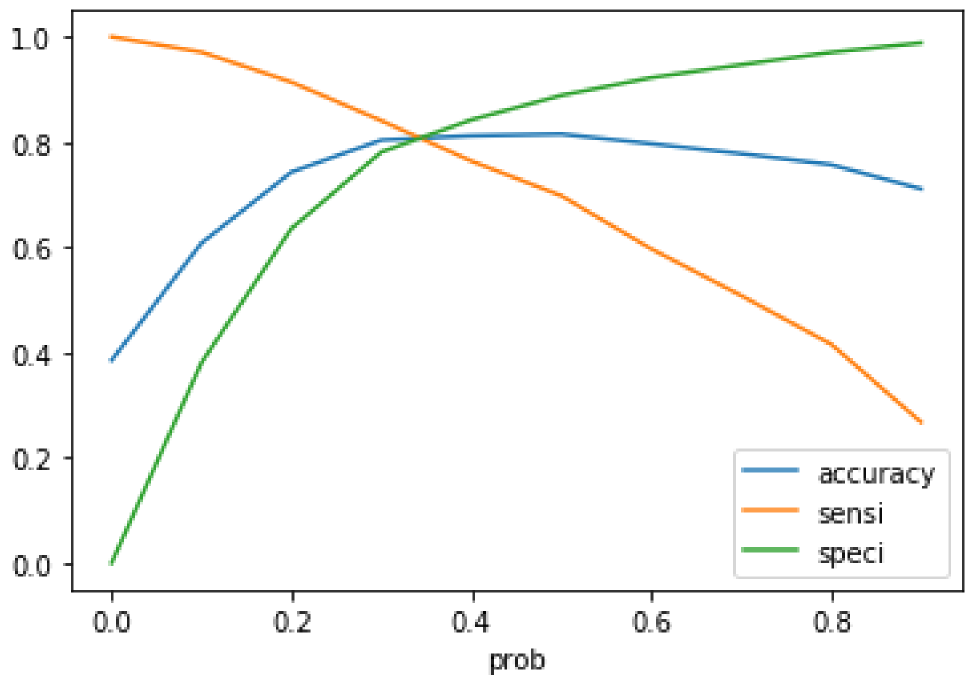
Using P-value and VIF variable were dropped and following is the list of variables which could fit in the model

Do Not Email
Total Time Spent on Website
Lead Source_Olark Chat

Lead Source_Reference
Lead Source_Welingak Website
Last Activity_Converted to Lead
Last Activity_Had a Phone Conversation
Last Activity_Olark Chat Conversation
Last Activity_SMS Sent
Specialization_Others
What is your current occupation_Working Professional
Last Notable Activity_Email Link Clicked
Last Notable Activity_Modified
Last Notable Activity_Olark Chat Conversation
Last Notable Activity_Page Visited on Website



Since area under the curve is 0.88, it indicates that most of the data points are covered by the model.



The intersection point of accuracy, sensitivity and specificity is used for finding the optimum cut-off point for the model. Based on this cut-off, the predicted_converted is calculated.

10. Model Evaluation

Comparing the values obtained for rain & rest: rain

Data:

Accuíacy : 81.24%

Sensitivity : 80.7%

Specificity : 81.58%

rest Data:Accuracy

:80.20%

Sensitivity : 81.09%

Specificity : 79.70%