Lead Score

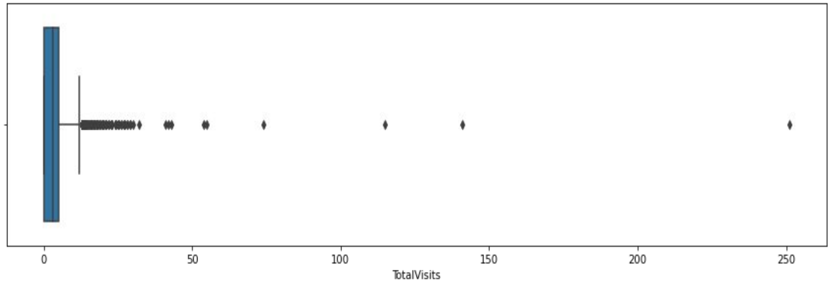
Problem Statement

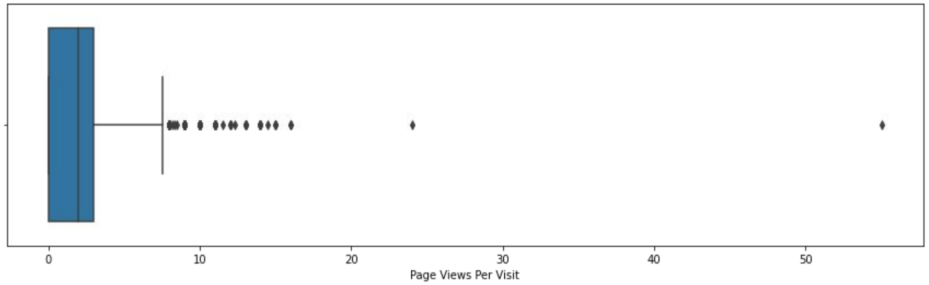
An education company named X Education sells online courses to industry professionals. On any given day, many professionals who are interested in the courses land on their website and browse for courses. Now, although X Education gets a lot of leads, its lead conversion rate is very poor. To make this process more efficient, the company wishes to identify the most potential leads, also known as ‘Hot Leads’.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%.

Treating Missing Values

1. After calculating the percentage of missing values for each variables. Drop the variables having missing value more than 45 percent.
2. There are variables in the dataset which are generated by the sales team for their convenience for contacting persons. Such variables are found and dropped.
3. As per the problem statement many of the categorical variables have a level called 'Select' which needs to be handled because it is as good as a null value.
4. While treating the missing values for all the columns, there are levels which are very less when compared to other levels. Hence these levels are capped to 'Other'.
5. For numerical variables fill the missing values with mode

Checking for Outliers

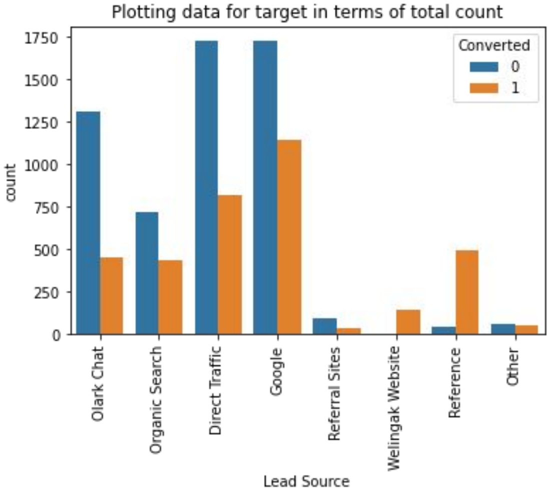


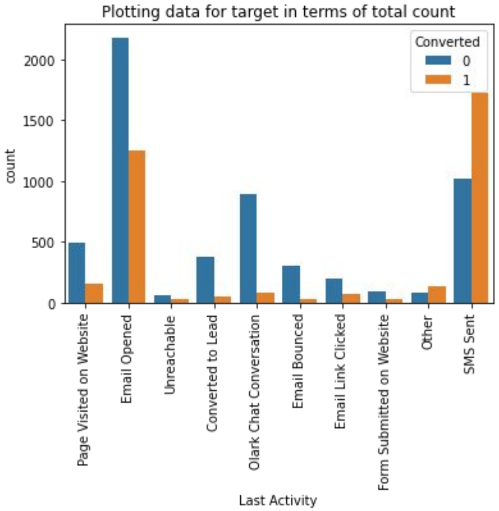


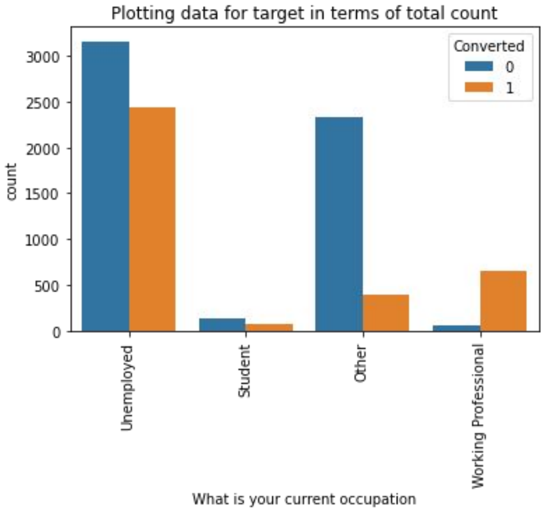
Numerical columns are plotted using boxplot to find any outliers in that particular variable

1. Variables such as ‘TotalVisits’ and 'Page Views Per Visit' contain outlier values.
2. Since the number of outliers are less compared to length of the dataset. We can replace those values with the 0.99 quantile value

Bivariate Analysis







1. Lead having values Reference and Welingak Website tend to join the coarse
2. Lead having values SMS Sent in Last Activity tend to join the coarse
3. Lead having values Working Residential in Occupation tend to join the coarse.

Data Preparation

1. Converting some binary variables (Yes/No) to 0/1
2. Creating dummy variables for the remaining categorical variables and dropping the level with big names.
3. Dropping Duplicate column
4. Adding the results to the master dataframe

Pre Model Building

1. Split the dataset into Train-Test dataset in the ratio of 70:30 by allotting feature variable to ‘X’ and response variable (Converted) to ‘y’
2. Scaling is done for the numerical columns such as ['TotalVisits', 'Total Time Spent on Website', 'Page Views Per Visit'].
3. The Converted Rate was found to be 38 percent were we can say that the dataset is pity much balanced
4. Dropping highly correlated dummy variables from correlation matrix

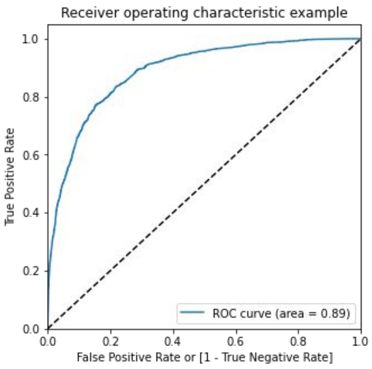
Model Building

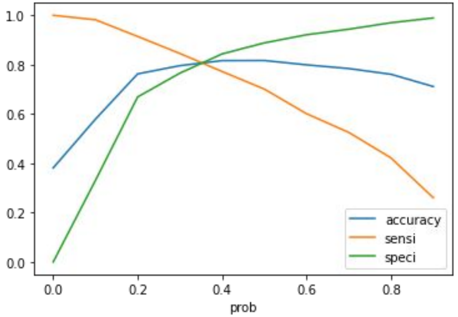
Using statsmodel we build a Logistic Regression Model:

1. From RFE and VIF method we select variables that defines the response variable
2. By Getting the predicted values on the train set, Create a new dataset which contains predicted probabilities.
3. Derive the Confusion matrix
4. Check the overall accuracy of the model

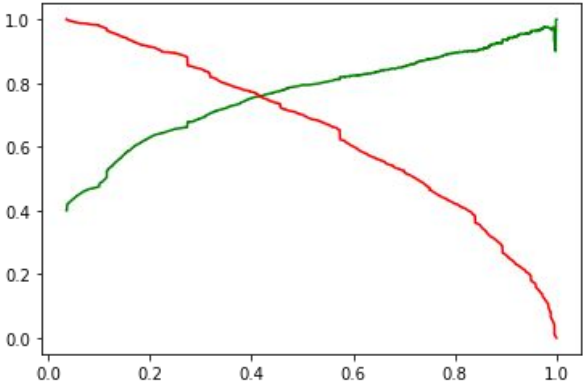
Model Evaluation

ROC





Precision and Recall



Plot a ROC curve to derive the optimal cut-off by creating different probability cutoffs. From the curve above, 0.35 is the optimum point to take it as a cutoff probability

This is another method to derive the optimal cut-off by importing libraries from scikit learn. From the above graph we get 0.42 as the optimum point for cutoff probability

Predictions

1. Taking the optimal cutoff of ROC we derive confusion matrix from the test dataset
2. From the matrix we calculated the percentage of sensitivity, specificity, Positive predictive value, Negative predictive value
3. As per the problem statement the predictive capability should be around 80 percent. This value is represented by the sensitivity
4. Sensitivity = 80.09%, Specificity = 81.81%, Positive predictive value = 74.19%, Negative predictive value = 86.28%

Thus the desired output is obtained

Inference

The parameters of the model are:

Do Not Email x -1.0708 + Total Time Spent on Website x 1.0722 + Lead Origin_Landing Page Submission x -0.8781 + Lead Origin_Lead Add Form x 2.9575 + Lead Source_Olark Chat x 1.1304 + Lead Source_Welingak Website x 2.2897 + Specialization_Hospitality Management x -0.9788 + Specialization_Other x -0.8012 + occupation_Student x 1.2309 + occupation_Unemployed x 1.0581 + occupation_Working Professional x 3.4898 + Last Activity_Email Bounced x -0.8511 + Last Activity_Olark Chat Conversation x -1.2143 + Last Activity_SMS Sent x 1.2714

Top 3 Variables are:

1. occupation_Working Professional
2. Lead Origin_Lead Add Form
3. Lead Source_Welingak Website