Summary Report

Objective

To help the X education select the most promising leads, i.e. the leads that are most likely to convert into paying customers. The requirement is 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.

Build a logistic regression model such that we are able 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.

Method Implemented

1. We have been provided with a leads dataset from the past with 37 columns containing the data points. This dataset consists of various attributes such as Lead Source, Total Time Spent on Website, Total Visits, Last Activity, etc. which may or may not be useful in ultimately deciding whether a lead will be converted or not.

2. As a first step, exploratory data analysis is conducted to figure out the trends among the dataset, and to find the outliers on the data. Columns with high null values are discarded and columns with less percentage are imputed to suit the model creation.

3. Out of all the variables created, Recursive Feature Elimination (RFE) is done to reduce the feature set to 15. Checking Variance Inflation Factor (VIF) we find that none of the left-over columns for analysis have a high value, thus are not correlated to each other.

4. Logistic regression model is then created and checked for the p-value (using stats model's General linear model) to check if any more columns have a high value of p.

5. Receiver operating characteristic (ROC) curve, Accuracy-Sensitivity-Specificity graph, and Precision- Recall plots are created. The optimal value is found to be 0.3 and 0.4 from the latter two graphs.

If the probability of conversion is greater than 0.3 (cutoff value), it is considered that the person is a suitable lead for conversion into a possible client to buying the program.

Lead score is calculated by expressing the conversion probability as a percentage.