SUMMARY

Problem Statement

X Education is a company that provides online courses for industry professionals. This analysis was done for the company to find ways to select promising leads that can be converted into paying customers. Data was provided by the company where leads come through various modes like email, advertisements on websites, google searches, etc.

The company expects us to create a model in which we give a lead score to each lead so that customers with higher lead scores have a higher conversion chance and customers with lower lead scores have a lower conversion chance.

The following steps were used to analyze the data:

1. Data Understanding:

We read through the data and saw what details were available in the data that we received, the size of the data, trying to understand the requirements of the company and what needs to be done.

2. Data Cleaning:

The data provided was raw so we had to remove multiple columns containing data of a single value as these data do not contribute to the inference, we removed them from further analysis. Columns that had missing values over 40% were dropped as using such data will give skewed results. A few columns like 'TotalVisit', and 'Last Activity' were imputed with values of maximum occurrence. We checked if there were outliers, and data with yes/no variables were encoded with binary variables.

3. Exploratory Data Analysis:

Then we started with Exploratory Data Analysis where we checked which variables were showing major lead conversions, multiple graphs were drawn to make inferences.

4. Dummy variable and Train -Test split:

Dummy variables were created of all the remaining categorical variables and the first level was dropped. Then the data was splitted into two sections in the ratio of 70:30 as train and test data respectively, Data was scaled using a standard scaler.

5. Model Building:

Using RFE we selected 15 relevant variables, Multicollinearity was checked using VIF and in the final model, the p-values are below 0.05 and VIF values are below 5.

6. Prediction and Evaluation:

We evaluated the model using the ROC curve which gave us an area coverage of 0.85 and then we plotted the probability graph for Accuracy, Sensitivity, and Specificity using different probability values to find the optimal probability cut-off.

7. Final result:

Finally, we implemented our learning to make the predictions on the Test dataset and calculated the conversion probability based on the following factors:

Train Data Set metrics:

Specificity Score: 76.91

Precision Score: 68.14

Sensitivity/Recall Score: 80.13

Test Data Set metrics:

Specificity Score: 74.96

Precision Score: 67.62

Sensitivity/Recall Score: 80.09

Considering the above scores, we can conclude that our model performed well and predicted close to accurate results. The above model can be changed as per the company's requirements in the future.