Lead Scoring Case Study Summary

Problem Statement:

This analysis is done for X Education and to find ways to get more industry professionals to join their courses. The basic data provided gave us a lot of information about how the potential customers visit the site, the time they spend there, how they reached the site and the conversion rate.

Solution Summary:

Data Cleaning:

The data was partially clean except for a few null values and the option select had to be replaced with a null value since it did not give us much information. Few of the null values were changed to 'not provided' so as to not lose much data. Although they were later removed while making dummies. Since there were many from India and few from outside, the elements were changed to 'India', 'Outside India' and 'not provided'.

Dummy Variables Creation:

The dummy variables were created and later on the dummies with 'not provided' elements were removed. For numeric values we used the Min Max Scalier.

Test Train Split:

The next step was to divide the data set into test and train sections with a proportion of 70-30% value.

Recall:

We used the Min Max Scaling to scale the original numerical variables. The, we plot the heatmap to check the correlations among the variables. Dropped the highly correlated dummy variables.

Model Building:

Firstly, RFE was done to attain the top 15 relevant variables. Later the rest of the variables were removed manually depending on the VIF values and p-value.

Conclusion:

- The lead score calculated in the test set of data shows the conversion rate of 83% on the final predicted model which clearly meets the expectation of CEO has given a ballpark of the target lead conversion rate to be around 80%.
- Good value of sensitivity of our model will help to select the most promising leads.

Features which contribute more towards the probability of a lead getting converted are:

i. Lead Origin Lead Add Form
ii. What is your current occupation Working Professional
iii. Total Time Spent on Website