Summary Lead Scoring Case Study

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 different ways of interactions. Also, the conversion rate.

The following are the steps used:

1. **Importing libraries and Data Sourcing:** To perform all the necessary tasks. We had to first import basic libraries like: NumPy, pandas, matplotlib and seaborn. We imported the Leads.csv file and read the data. Also, as we were moving ahead in the process, we imported other libraries too.

2. **Data Inspection:** It required checking shape of the data and the data types of variables.

3. **Cleaning Data:** The data was partially clean except for a few null values and the option select had to be replaced with a null(NaN) value since it was not giving us much information. Although they were later removed while making dummies and we removed Prospect ID and Lead Number columns due to unique value in it.

   As there were many from India and few from outside, the elements were changed to 'India', 'Outside India'. Also, for some categorical variables (Lead Source, Specialization, tags, Last Activity, Last Notable Activity) were having data which was quite low in frequency hence clubbed such data together with in variable with different names (Others, Management_Specialization, Not Specified, Others, Other_Notable_Activity) respectively.

4. **Dummy Variables:** The dummy variables were created and later the dummies with clubbed variable names (Others, Management_Specialization, Not Specified, Others, Other_Notable_Activity) elements were removed. For numeric values we used the StandardScaler.

5. **Train-Test split:** The split was done at 70% and 30% for train and test data respectively.

6. **Model Building:** RFE was done to get the top 15 relevant variables. Later the rest of the variables were removed manually depending on the p-value and VIF values (The variables with VIF > 5 and p-value >0.05 were dropped).

7. **Model Evaluation:** A confusion matrix was made. Later on the optimum cut off value using ROC curve. ROC Curve area value came .88 and it was used to calculate the accuracy, sensitivity and specificity.

8. **Prediction**: Prediction was done on the test data frame and with an optimum cut off as 0.36 with accuracy, sensitivity and specificity of 80.12%, 77.87%, 81.47% respectively.

9. **Precision** – **Recall**: This method was also used to recheck and a cut off 0.4 was found with Precision around 71.66% and recall around 77.87% on the test data frame.