Summary Report: 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 the conversion rate.

The following are the steps used:

1. **Inspecting the data**: 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.

2. **EDA**: EDA was done on both categorical and numeric columns to check the condition of our data and have some intuition about it. Few of the null values were changed to 'missing' so as to not lose much data. Although they were later removed after making dummies. Since there were many from India and few from outside, the country column was not providing any insight and hence was dropped. Also, there were some categorical columns with only one unique value, since they won't provide any insight to the model hence, were also dropped. The numeric values 'Total Visit' and 'Page View per Visit' were having outliers.

3. **Binary and Dummy Variables creation:** Categorical variables with 2 unique status(Yes/No) were converted to 1/0 binary variables for ease of computation and the dummy variables were created using 'get_dummies' function and later on the dummies with '_Missing' elements were removed so we don't lose any valuable data.

4. **Feature Scaling:** For feature scaling of 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**: Firstly, RFE was done to attain the top 20 relevant variables. Later the rest of the variables were removed manually depending on the VIF values and p-value (The variables with VIF <= 5 and p-value <= 0.05 were kept).

7. **Model Evaluation**: A confusion matrix was made. Later on the optimum cut off value of 0.29 (using ROC curve) was used to find the accuracy, sensitivity and specificity of 79%, 77% and 81% respectively.

8. **Prediction**: Prediction was done on the test data frame and with an optimum cut off as 0.29 with accuracy, sensitivity and specificity of 78%, 76% and 80% respectively.

9. **Feature Importance:** It was found that the variables that mattered the most in the potential buyers are (In descending order):
    - Lead Origin: When it is from Lead Add Form
    - What is your current occupation: When it isWorking Professional
    - Lead Source: When it is from Welingak Website
    - Specialization: When it is Rural and Agribusiness or IT Projects Management