SUMMARY

This analysis is done for X Education in an effort to increase the number of industry experts enrolled in their courses. We learned a lot about potential consumers' website visits, their duration there, and their subsequent actions from the dataset that was made available. got to the website and the conversion percentage.

The following technical steps are used:-

1. DATA CLEANING:
   - First step to clean the dataset we choose to remove the redundant variables / features.
   - The data set was partially clean except for a few null values and the option ‘Select’ has to replace with a null value since it did not give us much information.
   - Dropped the high percentage of Null values more than 40%.
   - Checked the number of unique categories for all categorical columns.
   - From that identify the highly skewed columns and dropped them.
   - Treated the missing values by imputing the favourable aggregate function like Mean, Median, and Mode.
   - Detected the Outliers.

2. EXPLORATORY DATA ANALYSIS:
   - A quick EDA was done to check the condition of our data. It was found that a lot of elements in the categorical variables were irrelevant. The numeric values seems good but found the outliers.
   - Performed Univariate Analysis for both Continuous and Categorical variables.
   - Performed Bivariate Analysis with respect to Target variable.

3. DUMMY VARIABLES:
   - The dummy variables are created for all the categorical columns.

4. SCALING:
   - Used Standard scalar to scale the data for continuous variable.

5. TRAIN-TEST SPLIT:
   - For the train and test sets of data, the split was performed at 70% and 30%, respectively.

6. MODEL BUILDING:
    - using the 20 variables that are supplied and RFE. By using RFE with provided 20 variables. It provides the top 20 pertinent variables. Subsequently, the p-value and VIF values were used to manually remove the unnecessary characteristics (the variables with a p-value of 0.05 and a VIF of less than five were retained).

7. MODEL EVALUATION:
    - A confusion matrix was made. Subsequently, the accuracy, sensitivity, and specificity were determined by utilizing the ROC curve to determine the ideal cut-off value, which was approximately 80%.

8. PREDICTION:
    - On the test data frame, a prediction was made with an optimal cut-off of 0.37 with accuracy, sensitivity, and specificity of 80%.

9. PRECISION-RECALL:
    - The method was also used to recheck, and we got a cut-off is 0.41.

10. CONCLUSION:

    We have observed that the following factors are the most crucial to prospective buyers:

    - The total time spent on the Website.
    - Total number of visits.
    - When the lead source was:
      - Olark Chat
    - When the last activity was:
      - SMS
      - Olark chat conversation