Summary

This analysis is carried out for X Education to uncover strategies to attract more industry experts to their courses. The dataset supplied provides us with a lot of information about how potential clients access the site, how much time they spend there, and how they leave reached the site, as well as the conversion rate.

The following technological procedures are employed: -

1. Data Cleaning: The first stage in cleaning the dataset is to eliminate duplicated variables/features.
    a. The data set was mostly clean, save for a few null values, and the option 'Select' had to be replaced with a null value because it didn't provide us with any information.
    b. More than 40% of the high percentage of Null values were removed.
    c. Tested for the number of distinct Categories in all Categorical columns.
    d. The highly skewed columns were identified and discarded as a result.
    e. Missing data were treated by imputing a favourable aggregate function, such as (Mean, Median, and Mode).
    f. Outliers were identified.

2. Exploratory Data Analysis:
    a. A short EDA was performed to assess the state of our data. Several items in the category variables were shown to be irrelevant. The numerical figures appear to be fine, however outliers were discovered.
    b. Univariate analysis was performed on both continuous and categorical variables.
    c. Conducted Bivariate Analysis on the Target Variable
3. Dummy Variables:
    a. For each classified column, dummy variables are constructed.
4. Scaling:
    a. For continuous variables, we used a standard scalar to scale the data.

5. Train-Test Split:
   a. The Spit was set to 70% for train data and 30% for test data.
6. Model Building:
   a. Using RFE with the specified 20 variables. It provides the top 20 important factors.
   b. Afterwards, unnecessary characteristics were manually deleted based on VIF values and p-value (variables with VIF 5 and p-value 0.05 were maintained).
7. Model Evaluation:
   a. A confusion matrix was constructed. Eventually, the ROC curve was used to determine the best cut-off value for accuracy, sensitivity, and specificity, which came to be about 80%.
8. Prediction:
   a. In the test data frame, the best cut-off was 0.37, with an accuracy, sensitivity, and specificity of 80%.
9. Precision-Recall:
   a. The approach was also employed for rechecking, with a cut-off of 0.41.
10. Conclusion:
    a. We discovered that the most essential characteristics for potential purchasers are:
       i. the overall time spent on the Website.
       ii. Number of visitors in total.
       iii. When the source of the lead was: - Chat
       iv. When was the most recent activity: - SMS
       v. chat discussion