Summary

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

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 for number of unique Categories for all Categorical columns.
   - From that Identified 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 variables.

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

6. **Model Building:**
   - By using RFE with provided 20 variables. It gives top 20 relevant variables. Later the irrelevant features was 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 by using ROC curve was used to find the accuracy, sensitivity and specificity which came to be around 80%.

8. **Prediction:**
   - Prediction was done on the test data frame an optimum cut-off as 0.37 with accuracy, sensitivity and Specificity of 80%.

9. **Precision-Recall:**
   - The method was also used to recheck and a cut-off of 0.41.

10. **Conclusion:**

We have noted that the variables that important the most in the potential buyers are:

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