Summary

The below analysis is performed for X Education which will tell us that which industry professionals will join their courses. The given dataset is having lot of information about the customers who 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:**
    - Remove the redundant variables/features.
    - We have replaced 'Select' with a null value since it did not give us much information.
    - Dropped null values having percentage higher than 40
    - Handled the missing values by imputing the max related value
    - Checked 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 favorable aggregate function like (Mean, Median, and Mode).
    - Detected the Outliers.
2. **Exploratory Data Analysis:**
    - We have performed a EDA on our data. It was found that a lot of elements in the categorical variables were irrelevant. The numeric values seem 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 Split was done at 70% and 30% for train and test the data respectively.
6. **Model Building:**
    - By using RFE. It gives top relevant variables.Later the irrelevant features 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 by using ROC curve was used to find the accuracy, sensitivity and specificity which came to be around 79%.

8. **Prediction:**
    - Prediction was done on the test data frame an optimum cut-off as 0.36 withaccuracy, sensitivity and Specificity of 78%.
9. **Conclusion:**

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

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