Summary

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:

Step 1: **Reading and Understanding Data:**
First of all, we import the "Lead.csv" file using Pandas library and read and analyse the whole dataset.

Step 2: **Data Cleaning**:
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. Few of the null values were changed to 'not provided' so as to not lose much data. We dropped the variables that had high percentage of NULL values in them. This step also included imputing the missing values as and where required with median values in case of numerical variables and creation of new classification variables in case of categorical variables. The outliers were identified and removed.

Step 3: **Exploratory Data Analysis:**
Then we started with the Exploratory Data Analysis of the data set to get a feel of how the data is oriented. In this step, there were around 3 variables that were identified to have only one value in all rows. These variables were dropped.

Step 4: **Creating Dummy Variables:**
we went on with creating dummy data for the categorical variables and later on, the dummies with 'not provided' elements were removed.

Step 5: **Test Train Split**:
The next step was to divide the data set into 70% and 30% for train and test data respectively.

Step 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).

Step 7: **Model Evaluation:**

A confusion matrix was made. Later on, the optimum cut off value (using ROC curve) was used to find the accuracy of 81%, sensitivity of 80% and specificity which came to be around 82%.

Step 8: Prediction:

Prediction was done on the test data frame and with an optimum cut off as 0.37 with accuracy, sensitivity and specificity of 81%.

Step 9: Precision – Recall:

This method was also used to recheck and a cut off of 0.42 was found with Precision around 74% and recall around 81% on the test data frame.