Summary

This analysis was done for X Education to find ways to attract more industry professionals to their courses. The underlying data provided us with a wealth of information on how potential customers came to the site, how long they stayed there, how they got there, and the success rate.

The following are the steps used:

1. **Cleaning data:**
   The data is partially clean except for some null values, and the select option should be changed to null because it doesn't give us much information. Rarely empty values are changed to "not provided" to avoid losing large amounts of data. Although they were removed later when the dummy was made. Since there are many from India and few from outside, change the items to "India", "Outside India" and "Not Provided".

2. **EDA:**
   A quick EDA was performed to verify the status of our data. Many elements of the categorical variables were found to be irrelevant. The values look good, no outliers were found.

3. **Dummy Variables:**
   Dummy variables were created and dummy variables with "not provided" items were later deleted. For values, we used StandardScalar.

4. **Train-Test split:**
   The split was done at 70% and 30% for train and test data respectively.

5. **Model Building:**
   Firstly, RFE was done to attain the top 15 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).

6. **Model Evaluation:**
   A confusion matrix was created. Later, using optimal cut off values (using ROC curves) to find precision, sensitivity, and specificity, each around 80%.

7. **Prediction:**
   Predictions were made on the test data frame with an optimal threshold of 0.35 and an accuracy, sensitivity and specificity of 80%.

8. **Precision – Recall:**
   This approach was also double-checked and found to have a threshold of 0.41, an accuracy of around 73%, and a recall of around 75% on the test database.

The variables found to be most important to potential buyers are (in descending order):

- The total time spends on the Website.

- Total number of visits.
- When the lead source was:
    - Google
    - Direct traffic
    - Organic search
    - Welingak website
- When the last activity was:
    - SMS
    - Olark chat conversation
- When the lead origin is Lead add format.
- When their current occupation is as a working professional.

Keeping this in mind, X Education can thrive as it has a good chance of changing the mind of almost every potential buyer and buying their courses.