Summary

This analysis is done for X Education to find ways to get more industry professionals to join their courses. The basic data provided gave us a lot of information about the Leads.

The following are the steps used:

1. **Cleaning data:**
   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. A few of the null values were changed to 'not provided' to not lose much data. Although they were later removed while making dummies. Since there were many from India and few from outside, the elements were changed to 'India', 'Outside India', and 'not provided'.

2. **EDA:**
   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 seem good and no outliers were found.

3. **Dummy Variables:**
   The dummy variables were created and later the dummies with 'not provided' elements were removed. For numeric values we used the, MinMaxScaler.

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

5. **Model Building:**
   As there were a lot of variables present in the dataset which we cannot deal with. So, the best way to approach this is to select a small set of features from this pool of variables using RFE. It 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 made. Later on the optimum cut-off value (using the ROC curve) was used to find the accuracy, sensitivity, and specificity which came to be as follows, 78%, 73.9%, and 83%.
   .

7. **Prediction:**
   The prediction was done on the test data frame with an optimum cut-off as 0.42 with accuracy, sensitivity, and specificity of around 80%.

8. **Precision – Recall:**

This method was also used to recheck and a cut-off of 0.41 was found with a Precision of around 78.28% and recall of around 76.74% on the test data frame.
It was found that the variables that mattered the most in the potential buyers are (In descending order):

1. The total time spend on the Website.
2. Total number of visits.
3. When the lead source was:
    a. Google
    b. Direct traffic
    c. Organic search
    d. Welingak website
4. When the last activity was:
    a. SMS
    b. Phone conversation
5. When the lead origin is Lead add format.
6. When their current occupation is as a working professional.

Keeping these in mind ‘X Education’ can flourish as they have a very high chance to get almost all the potential buyers to change their mind and buy their courses.