Summary

This analysis was done for X Education in order to find ways to get more industry professionals to join their courses. The basic data provided gave us a lot of information about how 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:

1. **Cleaning data:**
   The data was partially clean. There were a few null values, and the option select had to be replaced with a null value since it did not give us much information. We changed some of the null values to 'not provided' so as not to lose too much data, but we later removed them while making dummies. Since there were many responses from India and few from outside, we changed the elements to 'India', 'Outside India' and 'not provided'.

2. **EDA:**
   We did a quick check of our data to make sure everything was in good condition. We found that there were a lot of irrelevant elements in the categorical variables, but the numeric values were all good. We didn't find any outliers either.

3. **Dummy Variables:**
   Dummy variables were created and those with 'not provided' elements were removed. For numeric values, the MinMaxScaler was used.

4. **Train-Test split:**
   The data was split into 70% training data and 30% test data.

5. **Model Building:**
   In order to determine the 15 most relevant variables, we first used RFE. Then, we removed the rest of the variables 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, the optimum cut off value (using ROC curve) was used to find the accuracy, sensitivity and specificity which came to be around 80% each.

7. **Prediction:**
   The prediction was done on the test data frame. The optimum cut-off was 0.35. As a result, the accuracy, sensitivity and specificity were 80%.

8. **Precision – Recall:**
   This method was used to recheck the data and a cut off of 0.41 was found to be the most accurate. Precision was around 75% and recall was around 76%.

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. Olark chat conversation

5. When the lead origin is Lead add format.

6. The X Education can flourish as they have a very high chance of getting almost all potential buyers to change their minds and buy their courses. This is especially true for people who are currently working professionals.