Summary

The analysis done for an X Education to find a ways to get more industry professionals who can 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:

1. **Cleaning data:**
   We cleaned the data partially for few null values and also for new values like selected we changed it to null. There were few null values that we changed to ‘not provided’ so not lose much data. Although they were later removed while making dummies. We can see a lot of people from India and less from outside, so we decided to changed it to ‘India’, ‘Outside India’ and ‘not provided’.

2. **EDA(Exploratory Data Analysis):**
   Done some EDA to check data condition. It was found that a lot of elements in the categorical variables were irrelevant. The numeric values seems good and no outliers were found.

3. **Dummy Variables:**
   The dummy variables were created and later on 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:**
   We did RFE attain the top 15 relevant variables. Later the rest of the variables were removed manually depending on the VIF values and p-value (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 ROC curve) was used to find the accuracy, sensitivity and specificity which came to be around 80% each.

7. **Prediction:**
   Prediction was done on the test data frame and with an optimum cut off as 0.35 with accuracy, sensitivity and specificity of 80%.

8. **Precision – Recall:**

   This method was also used to recheck and a cut off of 0.41 was found with Precision around 73% and recall around 75% on the test data frame. It was found that the variables that mattered the most in the potential buyers are (In descending order):

   The total time spend on the Website.

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