Summary

The analysis is done for the company X Education with its aim to find various ways to increase the number of industry professionals joining their courses. The data that has been provided gives us a lot of information about how the potential customers visit the site, how they reached the site, how much time is being spent on the site and the conversion rate.

The steps followed are as below:

1. Reading the data.

2. Cleaning the data:
   During the data cleaning process it was found that the data was mostly clean except for a few null values. The option select was replaced with nulls since it did not impact our dataset much. Few null values were replaced by 'not provided'. The said values were later removed while creating the dummy variables. Furthermore countries were defined as 'India', 'outside India' and 'not provided'.

3. EDA:
   Exploratory data analysis was done to get a few insights about our dataset. There were no outliers and all numeric values were good, also a lot of data was irrelevant for our use.

4. Creating dummy variables:
   The dummy variables were created keeping in mind to remove the 'not provided' elements. MinMaxScaler was used for numeric values.

5. Splitting data into train and test data:
   The data was split 70% for training and 30% for testing data.

6. Building the model:
   We attained the top 15 most relevant variables using RFE (Recursive Feature Elimination). The rest of the variables were removed on the basis of the VIF (Variance inflation factor) values < 5, and p-value < 0.05 .

7. Evaluating the model:
   A confusion matrix was made and the optimum cut off value was used to find the accuracy, sensitivity and specificity using ROC curve, which was around 80% for each.

8. Predication:
   Prediction was done on the test data set with an optimum cut off = 0.35 and the accuracy, sensitivity and specificity at 80% each.

9. Precision-Recall:
   This method was used to recheck our model and the optimal cut off came to be 0.41 with precision to be around 73% and recall around 77% for the test data set.

The variables the most mattered for the potential buyers were as follows (Highest to lowest):

1. The total time spent on the website.
2. Total number of visits on the website.
3. Lead source:
    a. Google
    b. Direct traffic
    c. Organic search
    d. Welingak website
4. Last activity:
    a. SMS
    b. Olark chat conversation
5. Lead origin is Lead add format
6. Current occupation is Working professional