Summary

This analysis is done for X Education and to find ways to get more industry professionals to 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 reaches the site and the conversion rate.

The following were the steps used to score the leads:

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. There were many values as 'Select' in dataset, which is equivalent to null [As mentioned in problem statement]
   Hence, replacing all such values as null

2. ## EDA:
   A quick EDA was done to check 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 on the dummied 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:
   Firstly, RFE was done to attain the top 15 relevant variables. Later the rest of variables were removed manually depending on the VIF values and p-value.

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 0.37 with accuracy, sensitivity and specificity of 80%.

8. ## Precision-recall:
   This method was also used to recheck and a cutoff of 0.4 was found with Precision around 72.2% and recall around 78.0% on the test data frame.

It was found out that variables that mattered the most in the potential buyers are:

1. Total number of visits.
2. The total time spent on website.
3. The last notable activity

- o Had a phone conversation
- o Unreachable
4. When the lead source was :
  - o Wellingak website
5. When the last activity was :
  - o SMS
  - o Olark chat conversation
  - o Email_Bounced
6. When the lead origin is lead add form.
7. When the current occupation was :
  - o Student
  - o Unemployed
  - o Other

Recommendations:
Keeping the above mentioned points in mind the X Education can increase all the potential buyers to change their mind and buy there courses. Company shouldn't make call to the leads whose last activity was "Olark Chat Conversation", "Email bounced" as well as whose current occupation is "student", "NA", and "unemployed" as they are almost unlikely to convert.
There are a lot of leads generated in the initial stage (top) but only a few of them come out as paying customers from the bottom. So company need to nurture the potential leads well (i.e. educating the leads about the product, constantly communicating etc.)