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 reached the site and the conversion rate.

The following are the steps used:

1. **Cleaning data:**

The columns in data having more than 35 % null values are dropped.
After that we have drop the rows (for the columns have more null values) and cleaned the data for better readability.
After the above step we have dropped the columns which are not expressing some concrete output . for example columns like Do Not Call', 'Search', 'Magazine' etc.

2. EDA:

A quick EDA was done to check the condition of our data.
Firstly, we had done univariate analysis with checking each and every variable with its value count to see if that is required or not.
Secondly, we have check correlation Metrix as part of multivariate analysis.

3. **Dummy Variables:**

The dummy variables were created and later the dummies with ‘not provided’ elements were removed.

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 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).
We have removed columns like ‘What is your current occupation_Housewife','Last Notable Activity_Had a Phone Conversation', 'What is your current occupation_Working Professional
As these are having high p values, after removing these we could not find any variable which has more VIF.

6. **Model Evaluation:**

A confusion matrix was made. Later on, the optimum cut off value (using ROC curve which comes as 0.86 which is quite good) was used to find the accuracy, sensitivity and specificity which came to be around 79% each.

7. **Prediction:**

Prediction was done on the test data frame and with an optimum cut off as 0.41 with accuracy, sensitivity, and specificity of 80%.

8. **Precision – Recall and test data set :**

This method was also used to recheck and a cut off of 0.41 was found with Precision around 76% and recall around 79% on the test data frame.

Conclusion:

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

Total Visits

Total Time Spent on Website

Lead Origin_Lead Add Form

Last Notable Activity_Unreachable

Last Activity_Had a Phone Conversation

What is your current occupation_Unemployed

Keeping these in mind the 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.