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 data was partially clean except for some null values. Firstly, we removed those variables with more than 3000 missing values in it. Since Prospect Id, Lead Number, City & Country does not make any difference in the context of the given objective, we dropped them too. Further "select "option had to be dropped since it is good as null values did not give us much information. There were a few columns in which only one value was majorly present for all the data points Since practically providing single-value variables to model has no effect, it's best that we drop these columns as they won't help with our analysis.

2. **EDA:**
   A quick EDA was done to check the condition of our data. It was found that a lot of elements in the categorical variables were irrelevant. There were some outliers present in Total visits and Page views per visit so, we capped those two variables to its 99th Percentile

3. **Dummy Variables:**
   The dummy_dataframe were created for categorical variable. Then both categorical and numerical merged into one data frame for further processing.

4. **Train-Test split:**
   The split was done at 70% and 30% for train and test data respectively.

5. **Normalizing of training dataset:**
   For used MinMax scalar for normalization.

6. **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).

7. **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.

8. **Prediction:**
   Prediction was done on the train data frame and with an optimum cut off as 0.43 withaccuracy, sensitivity and specificity of 80%.

9. **Precision – Recall:**
   This method was also used to recheck and a cut off of 0.43 was found with Precision around 78% and recall around 79% on the test data frame.

It was found that the variables that mattered the most in the potential buyers are (In descending order):

1. Total Time Spent on Website
2. Lead Origin_Lead
3. What is your current occupation_Working Profes.
4. Lead Source_Welingak Website
5. Last Notable Activity_Email Bounced Lead
6. Source Olark Chat1
7. Last Activity_Email Bounced
8. Do Not Email
9. TotalVisits
10. Last Notable Activity
11. SMS Sent Last Activity_Olark
12. Chat Conversation

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.