Summary

This analysis is done for X Education and to find ways to get more leads 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 a few null values and the option select had to be replaced with a null value since it did not give us much information. Few of the null values were changed to 'not available' so as to not lose much data. Although they were later removed while making dummies. Since there were many from India and few from outside, the elements were changed to 'India', 'Outside India' and 'not provided'.

2. EDA:
A quick EDA was done to check the condition of our data. Univariate analysis was done on categorical variables and inferences was drawn where possible. It was found that a lot of elements in the categorical variables were irrelevant. The numeric values required outlier treatment

3. **Data preparation for model building:**
The dummy variables were created on categorical For numeric values we used the MinMaxScaler. Train-Test split was done at 70% and 30% for train and test data respectively.

5. **Model Building:**
Initial feature selection through RFE. Python libraries to run the logistic models. Iteratively remove attributes and check p values & VIF.(The variables with VIF < 5 and p-value < 0.05 were kept).

6. **Model Evaluation and Optimisation:**
A confusion matrix was made 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 72% and recall around 78% on the test data frame.

Attributes that mattered the most for lead conversions –

1. TotalVisits
2. Total Time Spent on Website
3. Last Notable Activity_modified
4. Last Activity_olark chat conversation
5. Last Notable Activity_email opened

6. Lead Origin_lead add form
7. Lead Source_direct traffic
8. Lead Source_welingak website
9. Last Notable Activity_olark chat conversation
10. What is your current occupation_working profession
11. Last Notable Activity_page visited on website
12. Do Not Email_yes
13. Last Notable Activity_email link clicked

Recommendation:

1. Using such parameters mention above , the company could actually set up their business model in order to have more leads converted

2. Targeted marketing/customized telephonic conversations could be made to the people who have done the following:

- Spent a lot of time in the website & keep returning back to the website multiple times
- Last activity through SMS or Olarkchat
- Targeted advertisements for working professionals
- Discount coupon codes for who has at least initiated the chat or sent SMS