Summary

The analysis is 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 and so on.

Steps followed are:

1. **Cleaning data:**

The data underwent partial cleaning, with only a few null values remaining. To address the ambiguity of the "option select" entries, they were replaced with null values, as they didn't provide meaningful information. Additionally, some null values were replaced with "not provided" to prevent data loss, although these entries were later excluded during the creation of dummy variables. Furthermore, due to the predominance of entries from India and a limited number from other locations, the categories were simplified to 'India', 'Outside India', and 'not provided'.

2. EDA:

It was found that a lot of elements in the categorical variables were irrelevant. The numeric values seems good and no outliers were found.

3. **Dummy Variables:**

The dummy variables were created and later on the dummies 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 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).

6. **Model Evaluation:**

A confusion matrix was made. Later the optimum cut off value (using ROC curve) was used to find the accuracy, sensitivity and specificity which came to be around 81% 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 81%.

8. **Precision – Recall:**

This method was also used to recheck and a cut off of 0.41 was found with Precision around 73% and recall around 76% on the test data frame.

9.**Conclusion:**

It was found that the variables that mattered the most in the potential buyers are (In descending order):
1. The total time spend on the Website.
2. Total number of visits.
3. When the lead source was:
a. Google
b. Direct traffic
c. Organic search
d. Welingak website
4. When the last activity was:
a. SMS
b. Olark chat conversation
5. When the lead origin is Lead add format.
6. When their current occupation is as a working professional.

From the above points, the X Education can have a very high chance to get almost all the potential buyers to change their mind and buy their courses.