Summary

This analysis was conducted for X Education with the aim of attracting more industry professionals to enroll in their courses. The initial data provided valuable insights into the visitors' site engagement, duration of their visits, referral sources, and conversion rates. The following are the steps used:

1. **Cleaning data:**
   The data was mostly clean, with only a few null values and the "Select" option being replaced with a null value due to its lack of informative value. Some null values were updated to 'not provided' to retain data integrity, though they were subsequently removed when creating dummy variables. Given the majority of respondents were from India and a few were from other countries, the categories were updated to 'India', 'Outside India', and 'not provided'.

2. **EDA:**
   A brief Exploratory Data Analysis (EDA) was conducted to assess the data quality. It was observed that many categories within the categorical variables were irrelevant. The numerical values appeared to be in good shape, with no outliers detected.

3. **Dummy Variables:**
   Dummy variables were generated, and subsequently, those containing 'not provided' elements were eliminated. For the numerical values, we applied MinMaxScaler for normalization.

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 constructed, followed by determining the optimal cut-off value using the ROC curve to ascertain the accuracy, sensitivity, and specificity, all of which were approximately 80%.

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 approach was further validated, revealing a cut-off of 0.41, with precision approximately at 73% and recall around 75% on the test dataset.

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.

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.

X-----X-----X------X