Summary

This analysis is conducted for X Education to explore strategies for attracting more industry professionals to enroll in their courses. The initial data provided offered substantial insights into how potential customers interact with the site, the duration of their visits, their referral sources, and the conversion rate.

The following steps were employed:

1. **Data Cleaning:** The data was mostly clean, except for a few null values. The 'option select' entries were replaced with null values as they were not informative. Some null values were changed to 'not provided' to retain as much data as possible, although they were later excluded when creating dummy variables. Since there were many entries from India and a few from other countries, the elements were categorized as 'India', 'Outside India', and 'not provided'.

2. **Exploratory Data Analysis (EDA):** A brief EDA was performed to assess the data's condition. It was discovered that many elements in the categorical variables were irrelevant. The numeric values appeared satisfactory, with no outliers detected.

3. **Dummy Variables:** Dummy variables were created, and those with 'not provided' elements were subsequently removed. For numeric values, the MinMaxScaler was utilized.

4. **Train-Test Split:** The data was split into 70% for training and 30% for testing.

5. **Model Building:** Initially, Recursive Feature Elimination (RFE) was used to identify the top 15 relevant variables. Subsequently, the remaining variables were manually removed based on Variance Inflation Factor (VIF) values and p-values (variables with VIF < 5 and p-value < 0.05 were retained).

6. **Model Evaluation:** A confusion matrix was created. The optimal cut-off value (determined using the ROC curve) was used to calculate the accuracy, sensitivity, and specificity, which were approximately 80% each.

7. **Prediction:** Predictions were made on the test data frame, with an optimal cut-off of 0.35, resulting in accuracy, sensitivity, and specificity of 80%.

8. **Precision-Recall:** This method was also employed for verification, yielding a cut-off of 0.41 with a precision of around 73% and recall of approximately 75% on the test data frame.

Key Variables Influencing Potential Buyers (in descending order):

1. Total time spent on the website.

2. Total number of visits.

3. Lead source:
   - Google
   - Direct traffic
   - Organic search

4. Last activity:

1. SMS
2. Olark chat conversation
5. Lead origin as Lead add format.
6. Current occupation as a working professional.

By focusing on these factors, X Education can significantly increase the likelihood of converting potential buyers into course enrolees.