Summary

This analysis was conducted for X Education to explore strategies for attracting more industry professionals to their courses. The initial data provided insights into potential customer interactions with the website, including their visit duration, traffic sources, and conversion rates. The following steps were undertaken:

1. **Data Cleaning:**
   The dataset was mostly clean, but a few null values were present. The "option select" was replaced with a null value due to its limited information. Some null values were changed to "not provided" to retain data integrity, although they were removed during the creation of dummy variables. Given the majority of data came from India, entries were categorized as 'India', 'Outside India', and 'not provided'.

2. **Exploratory Data Analysis (EDA):**
   An initial EDA revealed that many elements in the categorical variables were irrelevant, while the numeric values appeared satisfactory with no outliers detected.

3. **Dummy Variables:**
   Dummy variables were created, and those containing "not provided" were subsequently eliminated. Numeric values were scaled using the MinMaxScaler.

4. **Train-Test Split:**
   The dataset was divided into training (70%) and testing (30%) subsets.

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

6. **Model Evaluation:**
   A confusion matrix was generated, and the optimum cutoff value was determined using the ROC curve, yielding accuracy, sensitivity, and specificity rates around 80%.

7. **Prediction:**
   Predictions were made on the test dataset, using an optimal cutoff of 0.35, which resulted in accuracy, sensitivity, and specificity rates of 80%.

8. **Precision – Recall:**
   The precision-recall method was applied to re-evaluate, identifying a cutoff of 0.41, with precision at approximately 73% and recall at around 75% on the test dataset.

The key variables influencing potential buyers were identified in descending order of importance:

1. Total time spent on the website.

2. Total number of visits.

3. Lead sources, notably: a. Google b. Direct traffic c. Organic search d. Welingak website

4. Last activity types, particularly: a. SMS b. Olark chat conversation

5. Lead origin as "Lead add format."

6. Current occupation as a working professional.

By focusing on these factors, X Education is well-positioned to convert a significant number of potential buyers into course enrolees.