SUMMARY

The purpose of this analysis was to assist X Education in attracting more industry professionals to enroll in their courses. The initial dataset provided valuable information about customer behavior, including website visits, time spent on the site, referral sources and conversion rates.

The analysis followed these key steps:

1. **Data Cleaning:** The dataset was relatively clean, with only a few null values. The "option select" variable was replaced with a null value as it provided limited information. Some null values were replaced with "not provided" to retain data. Also, country categories were modified to "India," "Outside India," and "not provided."
2. **Exploratory Data Analysis (EDA):** A brief EDA was conducted to assess the dataset. It was observed that certain elements in categorical variables were irrelevant, while numeric values appeared to be reasonable, with no outliers detected.
3. **Dummy Variables:** Dummy variables were created, and subsequently, dummy variables with "not provided" elements were removed. Numeric variables were scaled using MinMaxScaler.
4. **Train-Test Split:** The dataset was split into a 70% training set and a 30% testing set.
5. **Model Building:** Initially, Recursive Feature Elimination (RFE) was employed to identify the top 15 relevant variables. Then, additional variables were manually removed based on their VIF values and p-values, retaining those with VIF < 5 and p-value < 0.05.
6. **Model Evaluation:** A confusion matrix was constructed, and the optimal cut-off value (determined using the ROC curve) was used to calculate accuracy, sensitivity, and specificity. The resulting values were approximately 80% for each metric.
7. **Prediction:** Predictions were made using the optimized cut-off value of 0.35 on the test dataset, yielding an accuracy, sensitivity, and specificity of 80%.
8. **Precision-Recall:** Precision-recall analysis was performed to validate the results, identifying a cut-off value of 0.41 with precision around 73% and recall around 75% on the test dataset.

Based on the analysis, the variables that had the most significant impact on potential buyers, in descending order, were:

1. Total time spent on the website.
2. Total number of visits.
3. Lead source: Google, Direct traffic, Organic search, Welingak website
4. Last activity: SMS, Olark chat conversation
5. Lead origin in the Lead add format
6. Current occupation as a working professional

Considering these findings, X Education is in a favorable position to thrive, as they have a high chance of convincing almost all potential buyers to change their minds and purchase their courses.