SUMMARY

This analysis was conducted for X Education with the aim of devising strategies to attract more industry professionals to enroll in their courses. The initial dataset provided valuable insights into how prospective customers interact with the website, their time spent on it, the channels through which they arrived, and the conversion rate. The following steps were undertaken:

1. Data Cleaning:

- The dataset was mostly clean, but some null values needed attention.
- The "select" option was replaced with a null value as it did not provide useful information.
- A few null values were changed to 'not provided' to preserve data.
- To simplify the analysis, categories were consolidated into 'India,' 'Outside India,' and 'not provided' due to the high number of Indian users.

2. Exploratory Data Analysis (EDA):

- A brief EDA was conducted to assess the dataset's condition.
- Irrelevant elements within categorical variables were identified.
- Numeric values were checked for outliers, and none were found.

3. Dummy Variables:

- Dummy variables were created, and those with 'not provided' elements were subsequently removed.
- For numeric values, the MinMaxScaler was applied.

4. Train-Test Split:

- The data was divided into training (70%) and testing (30%) sets.

5. Model Building:

- Recursive Feature Elimination (RFE) was used to identify the top 15 relevant variables.
- Additional variables were manually removed based on VIF values and p-values (variables with VIF < 5 and p-value < 0.05 were retained).

6. Model Evaluation:

- A confusion matrix was generated.
- The optimal cutoff value (determined using the ROC curve) was used to calculate accuracy, sensitivity, and specificity, each of which reached approximately 80%.

7. Prediction:

- Predictions were made on the test dataset using an optimal cutoff of 0.35, resulting in an accuracy, sensitivity, and specificity of 80%.

8. Precision-Recall:

- Precision-recall analysis was conducted, revealing a cutoff of 0.41, with precision around 73% and recall around 75% on the test dataset.

Key findings indicate that the most influential factors among potential buyers, ranked in descending order, are as follows:

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

With these insights, X Education is well positioned to thrive, as they have a high probability of persuading nearly all potential buyers to reconsider and enroll in their courses.