SUMMARY

This analysis was conducted for X Education with the goal of identifying strategies to attract more industry professionals to enrol in their courses. The dataset provided valuable insights into how potential customers interact with the website, the time they spend on it, their paths to reaching the site, and the conversion rates.

The following technical steps were undertaken:

1. **Data Cleaning:**
   - The first step in cleaning the dataset involved removing redundant variables and features.
   - Although the dataset was mostly clean, there were a few missing values. Specifically, the "Select" option was replaced with a null value, as it did not contribute meaningful information.
   - Variables with more than 40% missing data were dropped.
   - The number of unique categories in all categorical columns was checked.
   - Highly skewed columns were identified and removed.
   - Missing values were handled by imputing appropriate aggregate functions, such as mean, median, or mode.
   - Outliers were detected and addressed.
2. **Exploratory Data Analysis (EDA):**
   - A preliminary EDA was performed to assess the quality of the data. It was observed that many of the categorical variables contained irrelevant elements, though the numeric data appeared in good shape, aside from the presence of some outliers.
   - Univariate analysis was carried out on both continuous and categorical variables.
   - Bivariate analysis was conducted to examine relationships with the target variable.
3. **Dummy Variables:**
   - Dummy variables were created for all categorical columns to convert them into a usable format for the model.
4. **Scaling:**
   - Standard scaling was applied to continuous variables to normalize the data.
5. **Train-Test Split:**
   - The dataset was split into a 70% training set and a 30% test set.
6. **Model Building:**
   - Using Recursive Feature Elimination (RFE) with the provided 20 variables, the top 20 most relevant features were selected. Subsequently, irrelevant features were removed manually based on their Variance Inflation Factor (VIF) and p-values. Variables with VIF less than 5 and p-values greater than 0.05 were retained.

7. **Model Evaluation:**
   - A confusion matrix was created to evaluate the model's performance. The optimum cut-off value was determined using the ROC curve, which led to an accuracy, sensitivity, and specificity of approximately 80%.
8. **Prediction:**
   - Predictions were made on the test dataset using an optimal cut-off of 0.37, achieving an accuracy, sensitivity, and specificity of around 80%.
9. **Precision-Recall:**
   - A precision-recall method was also applied, with a revised cut-off value of 0.41, further validating the model's performance.

Conclusion:

- The total time spent on the website.
- The total number of visits.
- The lead source, specifically when it was from Olark Chat.
- The last activity, especially when it involved SMS or an Olark Chat conversation.