Summary

This analysis was conducted for X Education with the objective of attracting more industry professionals to enroll in their courses. The dataset provided valuable insights into how potential customers interacted with the website, the duration of their visits, the sources that led them to the site, and the conversion rate.

The following technical steps were undertaken:-

1. **Data Cleaning:**
   - Redundant variables/features were removed.
   - Null values were addressed, and the "Select" option was replaced with null values as it provided limited information.
   - Columns with more than 40% null values were dropped.
   - The number of unique categories for all categorical columns was examined.
   - Highly skewed columns were identified and removed.
   - Missing values were treated using appropriate aggregate functions (Mean, Median, Mode).
   - Outliers were detected.
2. **Exploratory Data Analysis:**
   - Quick exploratory data analysis (EDA) was performed to assess the dataset's condition
   - Irrelevant elements in categorical variables were identified, and outliers were found.
   - Univariate analysis was conducted for both continuous and categorical variables.
   - Bivariate analysis was performed with respect to the target variable.
3. **Dummy Variables:**
   - Dummy variables were created for all categorical columns.
4. **Scaling:**
   - Standard scaling was applied to the data.
5. **Train-Test Split:**
   - The dataset was split into training (70%) and testing (30%) sets.
6. **Model Building:**
   - Recursive Feature Elimination (RFE) was used to select the top 13 relevant variables.
   - Irrelevant features were not needed to be removed based on VIF values and p-values (variables with VIF < 5 and p-value < 0.05 were retained and High z values). As all variables displayed values well within acceptable limits.
7. **Model Evaluation:**
   - A confusion matrix was generated, and the optimal cut-off value was determined using the ROC curve. This yielded an accuracy, sensitivity, and specificity of approximately 80%.

8. **Prediction:**

- Predictions were made on the test dataset using an optimal cut-off value of 0.49, resulting in an accuracy, sensitivity, and specificity of 70% to 80%.

9. **Precision-Recall:**

- Precision-Recall analysis was conducted with a cut-off value of 0.49

10. **Conclusion :**

The most influential variables for potential buyers were identified as:

- Total Time Spent on Website
- When Last Notable Activity was:
  - SMS Sent
  - Email Opened
- Total number of visits.