Summary

This analysis is conducted for X Education with the aim of attracting more industry professionals to enroll in their courses. The initial dataset provided valuable insights into how potential customers interact with the site, their duration of visit, referral sources, and conversion rates.

The steps involved in the analysis are as follows:

Data Cleaning:

- The dataset was mostly clean, with a few null values and 'select' options that were replaced with null values as they didn't provide significant information.
- Some null values were handled to minimize data loss, although they were eventually removed during the creation of dummy variables.

Exploratory Data Analysis (EDA):

- A brief EDA was conducted to assess the data quality. Irrelevant elements in categorical variables were identified and addressed.
- Numeric values were deemed satisfactory, and outliers were managed individually for each column.

Dummy Variables:

- Dummy variables were generated, and those containing 'not specified' elements were subsequently removed.
- StandardScaler was employed for numeric values.

Train-Test Split:

- The dataset was split into 70% for training and 30% for testing.

Model Building:

- Recursive Feature Elimination (RFE) was initially performed to identify the top 15 relevant variables.
- Further variable reduction was executed manually based on VIF (Variance Inflation Factor) values and p-values, retaining variables with VIF < 5 and p-value < 0.05.

Model Evaluation:

- A confusion matrix was generated, and the optimal cutoff value was determined using the ROC curve, resulting in an ROC curve value of 0.97, signifying excellent performance.
- The following metrics were obtained for the training data:
    - Accuracy: 92.29%
    - Sensitivity: 91.70%
    - Specificity: 92.66%

Prediction:

- Predictions were made on the test dataset using an optimal cutoff of 0.3.
- The following metrics were achieved:
    - Accuracy: 92.78%
    - Sensitivity: 91.98%
    - Specificity: 93.26%

Precision-Recall:

- Precision-recall analysis was conducted, and a cutoff of 0.3 yielded precision around 89% and recall around 91% on the test dataset.
- Overall, the model demonstrates strong predictive capabilities for the conversion rate, instilling confidence in the CEO to make informed decisions based on its insights.