Summary

The purpose of this analysis is to help X Education attract more industry professionals to their courses. We learned a lot about the potential consumers' visitation patterns, length of stay, method of access, and conversion rate from the basic data provided.

The steps utilized are as follows:

1. **Data Cleaning:**
    1.1. The data was mostly clean, with a few null values. The 'option select' category was replaced with null values as it was not informative.
    1.2. Some null values were replaced with 'not provided' to retain as much data as possible, though these were removed when creating dummy variables.
    1.3. Location data was simplified into 'India', 'Outside India', and 'not provided' categories.

2. **Exploratory Data Analysis (EDA):**
    2.1. A quick EDA revealed that many categorical variable elements were irrelevant.
    2.2. Numeric values appeared clean with no outliers detected.

3. **Dummy Variables:**
    3.1. Dummy variables were created for categorical data, and those with 'not provided' elements were removed.
    3.2. Numeric data was normalized using MinMaxScaler.

4. **Train-Test Split:**
    4.1. The data was split into 70% for training and 30% for testing.

5. **Developing models:**
    5.1. Recursive Feature Elimination (RFE) was used to select the top 15 relevant variables.
    5.2. Variables were further refined manually based on Variance Inflation Factor (VIF) values and p-values, keeping those with VIF < 5 and p-value < 0.05.

6. **Evaluation of the Model:**
    6.1. A confusion matrix was created, and the optimal cut-off value was determined using the ROC curve.
    6.2. Our results show that we have about 81% accuracy, 70% sensitivity, and 88% specificity.

7. **Prediction for the Test set:**
    7.1. Predictions were made on the test data using an optimal cut-off of 0.35, yielding an accuracy, sensitivity, and specificity of 81%.

8. **Precision-Recall:**

   8.1. The precision-recall method was also used to validate the model.

   8.2. An optimal cut-off of 0.41 was identified, with precision and recall around 75% on the test data.

It turns out that the factors that mattered the most to potential customers were (in descending order):

1) The total time spend on the Website.
2) Total number of visits.
3) When the lead source was:
   - Google
   - Direct traffic
   - Organic search
   - Welingkar website
4) When the last activity was:
   - SMS
   - Olark chat conversation
5) When the lead origin is Lead add format.
6) When their current occupation is as a working professional.

If X Education keeps these in mind, they stand a very good chance of convincing nearly every prospective customer to change their mind and enroll in their courses.

The End