Summary

This analysis is done for X Education and to find ways to get more industry professionals to join their courses. The basic data provided gave us a lot of information about how the potential customers visit the site, the time they spend there,how they reached the site and the conversion rate.

Answer: The following steps can be used.

1. Data Cleaning: The data was mostly clean, with a few null values addressed and the 'select' option replaced with null for better information.
Null values were changed to 'not provided' to retain data, and later removed while creating dummy variables.
Categorized elements from India, outside India, and 'not provided' to streamline geographical information.

2. Exploratory Data Analysis (EDA): Conducted a quick EDA to assess the data's condition.
Identified irrelevant elements in categorical variables.
Numeric values appeared sound with no outliers detected.

3. Dummy Variables: Created dummy variables and removed those with 'not provided' elements.
Applied MinMaxScaler for numeric values.

4. Train-Test Split: Split the data into 70% for training and 30% for testing.

5. Model Building: Utilized Recursive Feature Elimination (RFE) to identify the top 15 relevant variables.
Manually removed additional variables based on VIF values and p-values (keeping VIF < 5 and p-value < 0.05).

6. Model Evaluation: Constructed a confusion matrix. Determined the optimal cut-off value using the ROC curve, resulting in approximately 80% accuracy, sensitivity, and specificity.

7. Prediction: Applied the model to the test dataset with an optimal cut-off of 0.35, achieving 80% accuracy, sensitivity, and specificity.

8. Precision-Recall: Used Precision-Recall analysis, determining a cut-off of 0.41 with precision around 73% and recall around 75% on the test dataset.

It was found that the variables that mattered the most in the potential buyers are (In descending order):

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