Summary of Analysis for X Education

Steps Taken:

1. Data Cleaning:

Initial data with 9240 records in leads.csv file has 37 columns which include 30 categorical and 7 numerical columns are available.

The option "select" was replaced with NaN values as it lacked meaningful information. Dropped the columns having more than 40% missing value.

2. Exploratory Data Analysis (EDA):

A quick EDA was performed to assess the data's condition. It was found that many elements in the categorical variables were irrelevant.

The numeric values appeared clean, with no significant outliers.

3. Creating Dummy Variables:

Dummy variables were created for categorical data. Dummies with "not provided" elements were removed.

The MinMaxScaler was used to scale numeric values.

4. Train-Test Split:

The data was split into 70% training and 30% testing sets.

5. Model Building:

Recursive Feature Elimination (RFE) was performed to select the top 1 relevant variables.

Variables were manually removed based on Variance Inflation Factor (VIF) values and p-values (keeping those with VIF < 5 and p-value < 0.05).

6. Model Evaluation:

A confusion matrix was created to evaluate the model.

The optimal cut-off value (determined using the ROC curve) was used to calculate accuracy around 80%, sensitivity around 78%, and specificity around 82%.

7. Prediction:

Predictions were made on the test dataset with an optimal cut-off of 0.45, achieving accuracy, sensitivity, and specificity of 79%, 78% and 81% respectively.

8. Precision-Recall Analysis:

The precision-recall method was also employed, identifying a cut-off of 0.45 with precision around 81% and recall around 73% on the test dataset.