The project revolves around building a predictive model to help Company X Education identify potential users and convert them into active customers. The goal is to optimize conversion rates by targeting the correct group of users based on data insights. Below are the steps taken to build and evaluate the model, followed by a summary of the results:

1. Exploratory Data Analysis (EDA):

During the initial data exploration phase, a thorough check was performed on the data for missing values. Columns with more than 45% missing data were dropped to ensure a cleaner dataset. For columns that were crucial to the analysis but had missing values, the NaN entries were replaced with 'not provided' to retain data integrity. The 'Country' column, where India was the most frequent value, was imputed with India for missing values. However, as India accounted for nearly 97% of the data, this column was dropped as it lacked variability. The data cleaning also involved addressing numerical outliers and creating dummy variables for categorical data.

2. Train-Test Split & Scaling:

To assess the model's performance, the data was split into a training set (70%) and a test set (30%). To standardize the scale of numerical features, min-max scaling was applied to key variables, including 'Total Visits', 'Page Views Per Visit', and 'Total Time Spent on Website'. This step ensured that all features were on the same scale, which is critical for models like logistic regression.

3. Model Building:

For feature selection, Recursive Feature Elimination (RFE) was used to identify the top 15 variables that most significantly influenced conversion rates. Additional variables were manually removed based on their Variance Inflation Factor (VIF) and p-values. A confusion matrix was generated to evaluate model accuracy, which was found to be 80.91%. This indicated that the model performed well in distinguishing between converted and non-converted leads.

4. Model Evaluation:

Sensitivity-Specificity:

The model’s performance was evaluated using sensitivity and specificity metrics. After analyzing the ROC curve, the optimal cutoff value for predicting conversions was found to be 0.35. Using this cutoff:

- **Training Data**:
    - Accuracy: 80.91%
    - Sensitivity: 79.94%
    - Specificity: 81.50%
- **Test Data**:
    - Accuracy: 80.02%
    - Sensitivity: 79.23%
    - Specificity: 80.50%

Precision-Recall:

When evaluated using precision and recall, the model initially produced a precision of 79.29% and recall of 70.22% at the 0.35 cutoff. After optimizing the cutoff to 0.44:

- **Training Data**:
    - Accuracy: 81.80%
    - Precision: 75.71%
    - Recall: 76.32%
- **Test Data**:
    - Accuracy: 80.57%
    - Precision: 74.87%
    - Recall: 73.26%

5. Cutoff Value Conclusion:

The analysis suggests that there are two potential optimal cutoff values depending on the evaluation metrics:

- **Sensitivity-Specificity**: The optimal cutoff is 0.35.
- **Precision-Recall**: The optimal cutoff is 0.44.

Conclusion:

The model effectively predicts the conversion rate and provides actionable insights for Company X Education. Key variables contributing to conversion include:

- **Lead Source**: Total Visits and Total Time Spent on Website.
- **Lead Origin**: Lead Add Form.
- **Lead Source**: Direct Traffic, Google, Welingak Website, Organic Search, and Referral Sites.
- **Last Activity**: Do Not Email (Yes), Last Activity Email Bounced, and Olark Chat Conversation.

The model's high accuracy and ability to identify significant predictors make it a reliable tool for Company X Education to target the right users and improve their conversion strategies.