Summary

We have been building and validating a model for Company X Education to identify ways to improve user conversion rates. Here’s a summary of the steps taken and findings:

1. **Exploratory Data Analysis (EDA):**
   - **Null Value Handling:** We assessed the percentage of missing values and removed columns with over 45% missing data. For columns with significant missing values, we replaced NaN values with 'not provided.' Since 'India' was the most frequent non-missing value, we imputed 'not provided' values with 'India.' Due to the high prevalence of 'India' (nearly 97% of the data), this column was subsequently dropped.
   - **Data Processing:** We addressed numerical variables, outliers, and created dummy variables.
2. **Train-Test Split & Scaling:**
   - We split the data into 70% training and 30% testing sets.
   - Min-max scaling was applied to the variables: ['TotalVisits', 'Page Views Per Visit', 'Total Time Spent on Website'].
3. **Model Building:**
   - Feature selection was performed using Recursive Feature Elimination (RFE), identifying the top 15 relevant variables.
   - Additional variables were manually removed based on Variance Inflation Factor (VIF) values and p-values.
   - A confusion matrix was created, showing an overall accuracy of 80.91%.
4. **Model Evaluation:**
   - **Sensitivity-Specificity:**
     - **Training Data:** The ROC curve identified an optimal cutoff value of 0.35, yielding:
       - Accuracy: 80.91%
       - Sensitivity: 79.94%
       - Specificity: 81.50%
     - **Test Data:** The results were:
       - Accuracy: 80.02%
       - Sensitivity: 79.23%
       - Specificity: 80.50%
   - **Precision-Recall:**
     - **Training Data:** At a cutoff of 0.35, Precision and Recall were 79.29% and 70.22%, respectively. Adjusting the cutoff to 0.44 improved the metrics to:
       - Accuracy: 81.80%
       - Precision: 75.71%
       - Recall: 76.32%
     - **Test Data:** The results were:
       - Accuracy: 80.57%
       - Precision: 74.87%
       - Recall: 73.26%

  - Based on Sensitivity-Specificity Evaluation, the optimal cutoff value is 0.35. For Precision-Recall Evaluation, the optimal cutoff value is 0.44.

Conclusion

Top Variables Contributing to Conversion:

- **Lead Source:**
  - Total Visits
  - Total Time Spent on Website
- **Lead Origin:**
  - Lead Add Form
- **Lead Source Channels:**
  - Direct Traffic
  - Google
  - Welingak Website
  - Organic Search
  - Referral Sites
- **Last Activity:**
  - Do Not Email_Yes
  - Last Activity_Email Bounced
  - Olark Chat Conversation

The model demonstrates strong predictive capability for conversion rates and provides valuable insights for Company X Education to make informed decisions