Summary

In the context of our work with Company X Education, our primary goal is to enhance the conversion rate by identifying potential users and employing data-driven strategies. Here, we will elaborate on the key steps we followed throughout this data analysis process:

1. Exploratory Data Analysis (EDA):

- We initiated our analysis by assessing the extent of missing data. Columns with over 45% missing values were deemed less useful and were subsequently dropped.
- For columns with vital information and significant null values, we opted to replace the NaN entries with 'not provided' to preserve the data.
- Given that 'India' was the most prevalent entry among non-missing values, we filled the 'not provided' entries with 'India.'
- We observed that the 'India' category dominated the dataset (about 97%), rendering it less informative, so we decided to drop this column.
- Further data preprocessing involved handling numerical variables, addressing outliers, and creating dummy variables.

2. Train-Test Split & Scaling:

- We divided the dataset into training (70%) and testing (30%) subsets to facilitate model evaluation.
- To ensure the uniformity of the numerical variables, we performed min-max scaling on specific features, including 'TotalVisits,' 'Page Views Per Visit,' and 'Total Time Spent on Website.'

3. Model Building:

- We applied Recursive Feature Elimination (RFE) for feature selection, which helped us identify the top 15 most relevant variables.
- Additional variable selection was conducted by considering VIF (Variance Inflation Factor) values and p-values.
- Model performance was assessed through the creation of a confusion matrix, revealing an overall accuracy of 80.91%.

4. Model Evaluation:

- We explored model performance from the perspectives of Sensitivity-Specificity and Precision-Recall.

Sensitivity - Specificity Evaluation:

- On the training data, we found an optimum cut-off value using ROC curves, resulting in an area under the ROC curve of 0.88.
- The optimal cutoff value was determined to be 0.35, which led to an accuracy of 80.91%, sensitivity of 79.94%, and specificity of 81.50%.
- For predictions on the test data, we achieved an accuracy of 80.02%, sensitivity of 79.23%, and specificity of 80.50%.

Precision - Recall Evaluation:

- With a cutoff of 0.35 on the training data, we obtained a precision of 79.29% and recall of 70.22%.
- By adjusting the cutoff to 0.44, we improved accuracy to 81.80%, precision to 75.71%, and recall to 76.32%.
- In predictions on the test data, we achieved accuracy, precision, and recall values of 80.57%, 74.87%, and 73.26%, respectively.

5. Conclusion:

- In conclusion, our analysis pinpointed several crucial variables contributing to conversion:
  - Lead Source (Total Visits and Total Time Spent on Website).
  - Lead Origin (Lead Add Form).
  - Lead Source (Direct traffic, Google, Welingak website, Organic search, Referral Sites).
  - Last Activity (Do Not Email_Yes, Last Activity_Email Bounced, Olark chat conversation).
- Our model demonstrates the ability to predict conversion rates effectively, providing Company X Education with the confidence to make informed decisions based on data-driven insights.

This comprehensive analysis equips us with valuable information to guide our client in optimizing their conversion strategies and achieving their business objectives.