Summary

The objective of this project is to help company X Education to increase their conversion rate by understanding and validating the data to target the correct group of potential users. The following steps were followed:

1. EDA:
- Columns with more than 45% missing values were dropped.
- NaN values were replaced with 'not provided'.
- Not provided values were imputed with India.
- Since India had nearly 97% of the data, the column was dropped.
- Numerical variables were analyzed for outliers and dummy variables were created.
2. Train-Test split & Scaling:
- The split was done at 70% and 30% for train and test data respectively.
- Min-max scaling was performed on the variables ['TotalVisits', 'Page Views Per Visit', 'Total Time Spent on Website']
3. Model Building:
- RFE was used for feature selection.
- The top 15 relevant variables were attained using RFE.
- The rest of the variables were removed manually depending on the VIF values and p-value.
- A confusion matrix was created, and overall accuracy was checked which came out to be 80.91%.
4. Model Evaluation:
- Sensitivity-Specificity and Precision-Recall evaluations were performed.
- The optimal cut-off value for Sensitivity-Specificity was found to be 0.35.
- The optimal cut-off value for Precision-Recall was found to be 0.44.
5. Conclusion:
- The top variables contributing to conversion were identified.
- The model seems to predict the conversion rate very well and can provide the company confidence in making good calls based on this model.

TOP VARIABLE CONTRIBUTING TO CONVERSION:

- Lead Source: Total Visits, 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.