Summary

The purpose of the model construction and prediction work for company X Education is to identify strategies for converting prospective users. To determine how to target the right group and raise the conversion rate, we further analysed and validate the data.

Here is a summary of the actions taken:

- **EDA:**
  - We performed a fast check on the percentage of null values and eliminated columns that had more than 45% missing values.
  - We also observed that the null values in the rows were significant columns that would cost us a lot of data. Therefore, we changed the NaN values to "not provided" instead.
  - We assumed that all of the missing data came from India because it was the most frequently occurring value among the non-missing values.
  - After that, we noticed that India had a large number of values—nearly 97% of the data—so we removed this field.
  - We also worked on numerical variables, outliers and dummy variables.
- **Train-Test split & Scaling:**
  - For the train and test data, the split was performed at 70% and 30%, respectively.
  - For the variables ["TotalVisits," "Page Views Per Visit," and "Total Time Spent on Website," we used min-max scaling.
- **Model Building:**
  - RFE was employed to select features, and it was then used to identify the top 15 significant variables.
  - Afterwards, based on the p-value and VIF values, the remaining variables were manually eliminated.

- After creating a confusion matrix, the overall accuracy was determined to be 80.91%.

- **Model Evaluation:**
    - Sensitivity – Specificity evaluation :
        - On Training Data
            - The ROC curve was used to determine the ideal cut off value. There was 0.88 area under the ROC curve.
            - Plotting revealed that the ideal cutoff was 0.35, which provided
                - Accuracy 80.91%
                - Sensitivity 79.94%
                - Specificity 81.50%.
        - On Test Data
                - Accuracy 80.02%
                - Sensitivity 79.23%
                - Specificity 80.50%

    - Precision – Recall Evaluation**:**
        - On Training Data
            - The precision and recall with the cutoff of 0.35 are 79.29% and 70.22%, respectively.
            - Therefore, we must adjust the cut off value in order to raise the above percentage. Plotting revealed that the ideal cutoff value was 0.44, which provided
                - Accuracy 81.80%
                - Precision 75.71%
                - Recall 76.32%
        - On Test Data
                - Accuracy 80.57%
                - Precision 74.87%
                - Recall 73.26%

    Accordingly, the ideal cut off values for Sensitivity-Specificity Evaluation and Precision-Recall Evaluation, respectively, would be 0.35 and 0.44.

CONCLUSION

Top Variables Contributing to Conversion:

1. Lead Source:
    a. Total Visits
    b. Total Time Spent on Website
2. Lead Origin:
    a. Lead Add Form
3. Lead source:
    a. Direct traffic
    b. Google
    c. Welingak website
    d. Organic search
    e. Referral Sites
4. Last Activity:
    a. Do Not Email_Yes
    b. Last Activity_Email Bounced
    c. Olark chat conversation

The Model seems to predict the Conversion Rate very well and we should be able to give the Company confidence in making good calls based on this model.