Summary

For company X Education, model building and prediction are being done in order to find and convert potential leads. In order to target the potential leads and boost conversion rates, we will further analyze and validate the data.

Let's go over the subsequent steps:

1. **EDA:**
   - We performed a quick check on the percentage of null values and dropped columns that had more than 45% missing values.
   - We also discovered that rows with null values would cost us a lot of data. So, we replaced the Nan values with 'Others' instead for columns applicable
   - We replaced the select data with Nan to make it understandable
   - We imputed all missing values with India because it was the most common occurrence among the non-missing values.
   - Then we noticed that the number of values for India was quite high (nearly 97% of the data), so we dropped this column.
   - The same methodology was adopted for each column that followed the same characteristics.
   - Additionally, we worked on numerical variables, outliers, and dummy variables.

2. **Train-Test split & Scaling:**
   - For train and test data, the split was done at 70% and 30%, respectively.
   - Min-Max scaling was performed to the variables "Total Visits," "Page Views Per Visit," and "Total Time Spent on Website."

3. **Model Building**
   - RFE was used for feature selection and attain top 20 features for further processing
   - The remaining variables were then manually removed based on the VIF values and p-value.
   - A confusion matrix was created, and overall accuracy was checked which came out to be 80.88%.

4. **Model Evaluation**

   - **Sensitivity – Specificity**

     If we go with Sensitivity- Specificity Evaluation. We will get:

- On **Training Data**

    - The optimum cut off value was found using ROC curve. The area under ROC curve was 0.89.
    - After Plotting we found that optimum cutoff was **0.35** which gave

      Accuracy 80.88%
      Sensitivity 81.76%
      Specificity 80.34%.

- Prediction on **Test Data**

    - We get

      Accuracy 80.05%
      Sensitivity 80.38%
      Specificity 79.86%

- **Precision – Recall:**

  If we go with Precision – Recall Evaluation

    - On **Training Data**

        - With the cutoff of 0.35 we get the Precision & Recall of 78.83% & 70.56% respectively.
        - So, to increase the above percentage we need to change the cut off value. After plotting we found the optimum cut off value of **0.44** which gave

          Accuracy 81.60%
          Precision 76.19%
          Recall 75.07%

- Prediction on **Test Data**
    - We get

      Accuracy 80.58%
      Precision 74.58%
      Recall 73.65%

5. So, if we go with Sensitivity-Specificity Evaluation the optimal cut off value would be **0.35**
   &
   If we go with Precision – Recall Evaluation the optimal cut off value would be **0.44**

CONCLUSION

TOP VARIABLE CONTRIBUTING TO CONVERSION:

- Total Visits
- Total Time Spent on Website
- Lead Origin_Landing Page Submission

- Lead source:
    - Olark Chart
    - Welingak website
    - Referral Sites
- Last Activity:
    - Do Not Email
    - Last Activity_Email Opened
    - Olark chat conversation
    - SMS Sent

The Model appears to accurately predict the Conversion Rate, and we should be able to give the Company confidence in making good decisions based on this model.