Summary

The model building and prediction is being done for company X Education and to find ways to convert potential users. We will further understand and validate the data to reach a conclusion to target the correct group and increase conversion rate. Let us discuss steps followed:

1. **EDA:**
    - Quick check was done on % of null value and we dropped columns with more than 35% missing values.
    - Replacing the NaN values with 'not provided' for columns – Specialization, What matters most to you in choosing a course, Country, What is your current occupation.
    - For Country, imputing all not provided with India, since India was the most common occurrence among the non-missing values.
    - Removing ID columns since it is unique for all.
    - Performed Univariate Analysis: Categorical and Numerical variables.
    - Relating all categorical variables and categorizing b/w converted and non-converted.

2. **Train-Test split & Scaling :**

    - The split was done at 70% and 30% for train and test data respectively.
    - We have done min-max scaling on the columns - TotalVisits, Page Views Per Visit, Total Time Spent on Website.

3. **Model Building**

    - For Feature Selection, we have used RFE to attain top 15 relevant variables.
    - Then for 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 81.03%.

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 0.88.
            - After Plotting we found that optimum cutoff was **0.35** which gave

Accuracy 80.31%
Sensitivity 80.37 %
Specificity 80.28%.

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

    Accuracy 80.79%
    Sensitivity 81.03%
    Specificity 80.50%

- **Precision – Recall:**

  If we go with Precision – Recall Evaluation

  - On **Training Data**
    - With the cutoff of 0.35 we get the Precision & Recall of 78.86% & 69.58% 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 **0.41** which gave

      Accuracy 81.10%
      Precision 75.40%
      Recall 75.89%

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

      Accuracy 81.52%
      Precision 73.24%
      Recall 76.60%

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 **0.41**

CONCLUSION

TOP VARIABLE CONTRIBUTING TO CONVERSION:

- Total Visits
- Total Time Spent on Website
- Lead Origin lead add form
- Lead Source olark chat
- Lead Source welingak website
- Do Not Email yes
- Last Activity olark chat conversation
- Last Activity sms sent
- What is your current occupation other
- What is your current occupation student
- What is your current occupation unemployed
- What is your current occupation working professional
- Last Notable Activity unreachable

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.