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. **Data Cleaning and EDA:**
    - Percentage of missing values was checked and the columns with more than 40% missing values were dropped.
    - For columns having missing values less than 40% were replaced with others or the most common value.Eg, in the country column, since India is the most common occurrence among the non-missing values, we imputed all not provided values with India.
    - Univariate and bivariate analysis was done.
    - The columns have one unique value were checked and dropped.
    - We also worked on numerical variable, outliers and dummy variables.

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

    - The split was done at 70% and 30% for train and test data respectively.
    - Standard scaler was used on the variables ['TotalVisits', 'Page Views Per Visit', 'Total Time Spent on Website']

3. **Model Building**

    - RFE was used for feature selection.
    - Then RFE was done to attain the top 15 relevant variables.
    - Later 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 81.42%.

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 : 81.42%
    Sensitivity : 81.16 %
    Specificity : 81.59 %

  - Prediction on **Test Data**

    Accuracy : 80.87 %
    Sensitivity : 80.08 %
    Specificity : 81.31 %

- **Precision – Recall:**

  - On **Training Data**

    - With the cutoff of 0.35 we get the Precision & Recall of 79.57% & 70.68% 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 81.72%
      Precision 76.06%
      Recall 76.66%

  - Prediction on **Test Data**

      Accuracy 81.97%
      Precision 74.17%
      Recall 77.25%

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**

Recommendations:

The company should make calls to the leads coming from the Lead Origin_Lead Add Form, What is your current occupation_Working Professional and Lead Source_Welingak Website, those whose last activity is sms sent", who spent "more time on the websites" as they are more likely to get converted.

The company should not make calls to the leads whose last activity was "Olark Chat Conversation" , leads whose lead origin is "Landing Page Submission" , leads whose Specialization was "Others" , leads who chose the option of "Do not Email" as "yes" as they are not likely to get converted.

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.