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 45% missing values.
    - We also saw that the rows with the null value would cost us a lot of data and they were important columns. So, instead we replaced the NaN values with 'not provided'.
    - Since India was the most common occurrence among the non-missing values, we imputed all not provided values with India.
    - Then we saw the Number of Values for India were quite high (nearly 97% of the Data), so this column was 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.
    - We will do min-max scaling 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 20 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 82.21%.

4. **Model Evaluation**

Sensitivity – Specificity:

If we go with Sensitivity- Specificity Evaluation. We will get as below:
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 81.0%
  Sensitivity 80.2%
  Specificity 81.50%

Prediction on Test Data:
We got as below:
Accuracy 80.4%
Sensitivity 80.03%
Specificity 80.67%

Precision – Recall:

If we go with Precision – Recall Evaluation

On Training Data:

- With the cutoff of 0.35 we get the Precision & Recall of 79.74% & 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 as below:

  Accuracy 81.67%
  Precision 75.04%
  Recall 76.84%

Prediction on Test Data: We got

Accuracy 80.65%
Precision 75.0%
Recall 74.12%

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:

- LEAD SOURCE:
  - Total Visits
  - Total Time Spent on Website
- Lead Origin:
  - Lead Add Form
- Lead source:
  - Google
  - Direct traffic
  - Organic search
  - Welingak website
  - Referral Sites

Last Activity:

- Olark chat conversation
- Do Not Email_Yes

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.

- - - - Thank You - - -