Summary

The model building and prediction are 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 the conversion rate.

Step 1: EDA

- A quick check was done on % of null values and we dropped columns with more than 30% 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 replacing we remove the row containing null values.
- Since we see the most common occurrence of select words in some columns so after categorical attributes analysis we drop that column
- We also worked on categorical variables, outliers, and dummy variables.

Step 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 ['Total Visits', 'Page Views Per Visit, 'Total Time Spent on Website']

Step 3: Model Building

- RFE was used for feature selection.
- Then RFE was done to attain the top 15 relevant variables.
- Later, the variables were removed manually depending on the VIF values and p-value.

Step 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 the ROC curve. The area under the ROC curve was 0.87.
        - After Plotting we found that the optimum cutoff was **0.42** which gave

          Accuracy 78.10%
          Sensitivity 73.98%
          Specificity 83.86%.

    - Prediction of **Test Data**

        - We get

          Accuracy 78.66%
          Sensitivity 78.15%
          Specificity 76.96%

- **Precision-Recall:**

  If we go with Precision–Recall Evaluation

    - On **Training Data**

        - With the cutoff of 0.42, the value increases the above percentage. After plotting we found that it gives

          Accuracy 78.99%
          Precision 78.73%
          Recall 79.23%

    - Prediction of **Test Data**

        - We get

          Accuracy 79.17%
          Precision 78.85%
          Recall 77.57%

So if we go with Sensitivity-Specificity Evaluation the optimal cut-off value would be **0.**42. And if we go with Precision – Recall Evaluation the optimal cut-off value would be **0.44**

CONCLUSION

TOP VARIABLE CONTRIBUTING TO CONVERSION:

Total Time Spent on Website

Lead Origin_Lead Add Form

Lead Source_Olark Chat

Lead Source_Welingak Website

Do Not Email_Yes

Last Activity_Had a Phone Conversation

Last Activity_SMS Sent

What is your current occupation_Student

What is your current occupation_Unemployed

Last Notable Activity_Modified

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.