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.
    - Replaced ‘Select’ with Null value.
    - 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'.
    - We also worked on numerical variable, removed 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 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 came out to be approx. 81%. Sensitivity came out to be 71% and specificity was 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 0.88.
  - After Plotting we found that optimum cutoff was **0.35** which gave

    Accuracy = 80%
    Sensitivity = 80.4%
    Specificity = 80.2%.

- Prediction on **Test Data** with cutoff – 0.35

  - We get

    Accuracy = 81%
    Sensitivity = 81.5%
    Specificity = 80.6%

- **Precision – Recall:**

  If we go with Precision – Recall Evaluation

  - On **Training Data**

    - With the cutoff of 0.35 we get the Precision & Recall of 78.8% & 71% 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%
      Precision = 75%
      Recall = 75%

- Prediction on **Test Data**

  - We get

    Accuracy = 81.4%
    Precision = 73%
    Recall = 76%

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:

- LEAD SOURCE:
  - Total Visits
  - Total Time Spent on Website
- Lead Origin:
  - Lead Add Form
- Lead source:
  - Welingak website
  - Olark Chat
- Occupation:
  - Working Professional
  - Students