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 have replaced those values to India.
    - We have done unvarient analysis for some attributes in that some of the attributes contain only one type of value those column won't give More insights so we dropped those column
    - 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 have done StandardScaler on the variables ['Total Visits', '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 features.
    - 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 81.29%.

4.Model Evaluation

Sensitivity – Specificity

If we go with Sensitivity- Specificity Evaluation. We will get :

On **Training Data**

- With the current cut off as 0.5 we have around 82% accuracy, sensitivity of around 70% and specificity of around 89%.

- 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 **0.37** which gave

  Accuracy 81.1%
  Sensitivity 79.7%
  Specificity 82%.

Prediction on **Test 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 **0.37** which gave

- We get

  Accuracy 80.9%
  Sensitivity 78.7%
  Specificity 82%

Precision – Recall:

On **Training Data**

- We get

  Precision 73.5%
  Recall 79.7%

Prediction on **Test Data**

- We get

  Precision 71.5%
  Recall 78.7%

CONCLUSION

The variables that mattered the most in the potential buyers are :

1 The total time spend on the website.

2 When the lead source is:

A. Google

B. Direct traffic

C. Organic search

D. Welingak website

E. Referral sites

3 When the last activity is

A. Olark chat conversation

4 When the lead origin is lead add format.

5 Current occupation as a working professional students and unemployed candidates

6 Notable activity is email link clicked and email opened.

These variable X Education can consider and achieve high chance to convert their all the potential buyers .