Summary

As a part of the Lead Scoring case study, we have been presented with the details how the company X Education pursues customer leads from various sources and tries to convert them to potential customers. The current conversion rate is quite low at 30%. So we have been tasked to analyze the data and come up with a model which can make predictions to the order to 80% Lead conversion.

Data Cleaning and EDA:

1. Quick check was done on % of null value and we dropped columns with more than 35% missing values.
2. 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'.
3. Since India was the most common occurrence among the non-missing values, we imputed all not provided values with India.
4. Then we saw the Number of Values for India were quite high, so this column was dropped.
5. We also worked on numerical variable, outliers and dummy variables.

Data Preparation (Train-Test Split & Scaling):

1. Created dummy features (one-hot encoded) for categorical variables.
2. Splitting Train & Test Sets: 70:30 ratio.
3. Feature Scaling using Standardization
4. We will do min-max scaling on the variables ['TotalVisits', 'Page Views Per Visit', 'Total Time Spent on Website']

Model Building:

1. RFE was used for feature selection.
2. Then RFE was done to attain the top 15 relevant variables.
3. Later the rest of the variables were removed manually depending on the VIF values and p-value.
4. A confusion matrix was created, and overall accuracy was checked which came out to be 80.24%.

Model Evaluation:

1. **Sensitivity – Specificity:**

    a. **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 = 80.24%
        - Sensitivity = 79.91%
        - Specificity = 80.45%.

    b. **On Test Data:**

        - Accuracy 80.95%

- Sensitivity 80.80%
- Specificity 81.04%.

2. **Precision – Recall:**

   a. **On Training Data:**

   - With the cutoff of 0.35 we get the Precision & Recall of 79.25% & 70.80% 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 0.40 which gave

     - Accuracy 80.76%
     - Precision 74.32%
     - Recall 77.16%

   b. **On Test Data:**

     - Accuracy 81.20%
     - Precision 73.95%
     - Recall 77.55%

3. So, if we go with Sensitivity-Specificity Evaluation the optimum cut off value would be 0.35 and if we go with Precision – Recall Evaluation the optimum cut off value would be 0.40.

Conclusion:

1. The Model seems to predict the Conversion Rate very well and we should be able to give the CEO confidence in making good calls based on this model
2. Important features responsible for good conversion rate or the ones' which contributes more towards the probability of a lead getting converted are:

   a. Total time spend on the Website
   b. Lead Origin_Lead Add Form
   c. Last Notable Activity_Had a Phone Conversation