Summary

The model building and prediction is done for an education company named X Education sells online courses to industry professionals to check ways for converting leads to potential customers. We will further understand and validate the data to reach a conclusion to get the correct group and increase conversion rate of the customers.

Steps involved:

1. Loading the data is done, there are 9240 rows and 37 columns in the given dataset.

2.Found there are many missing values in the dataset, so deleted columns with null values greater than 45%.

3. Then deleted the columns with one unique values as they won't effect the analysis.

4. From the visualization process came to that the lead source is mainly from google and direct traffic, and unemployed people are more interested in this online course.

5. Did outlier treatment on 'Totalvisits' variable.

6. Created dummy variables on categorical variables which resulted in 81 columns.

7. Split train dataset with 70% and test data set with 30% .

8. Used MinMaxScaling on the train dataset.

9. 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.

10.Model Evaluation:

- A confusion matrix was created, and overall accuracy was checked which came out to be 80%.
- The optimum cut off value was found using ROC curve. The area under ROC curve was 0.87.
- **On Training set:** After Plotting we found that optimum cutoff was **0.35** we got accuracy 80%, sensitivity 81.74% and specificity 78.18%.
- **On Test set:** we got accuracy 79.12%, sensitivity 80.34% and specificity 78.34%.

- **On Training set**: With the cutoff of 0.35 we get the Precision & Recall of 78.32% & 65.22% 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 accuracy 79.53%, precision 72.68% and recall 73.11%
- **On Test set:** We get accuracy 78.75%, precision 73.07% and recall 71.46%.

11.Conclusion:

Top variable contributing to conversion:

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