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. Data Cleaning:
    - Quick check was done on % of null value and we dropped columns with more than 30% missing values.
    - We have put the most common value on which we were having <5% missing value.
    - 100% people don't want to pay through cheque, don't want updates on DM content, supply chain content, more updates on courses, so no help in analysis so drop it.
    - 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 'Missing'.
    - We also worked on numerical variable, categorical variables, outliers and dummy variables which is having more than 2 unique drops down.

2. EDA
    - Data imbalance checked- only 38.5% leads converted
    - We have created the univariate, bivariate analysis on both categorical and numerical columns, also we have checked the plot numerical vs converted columns and categorical vs converted columns where we have understood which columns is important for further analysis even, we have 29 % of missing value.
    - Maximum lead conversion happened from Landing Page Submission.
    - Max. lead conversion in the lead source is from 'Google'
    - Major lead conversion has happened from the emails that have been sent.
    - Major lead conversion has happened from the calls they are doing.
    - Maximum conversion to those customers who has opened their mails and whom SMS being sent, and those customers who has modified.
    - Major lead conversion is from the Unemployed Group
    - Max. lead conversion of those customer who has spent max time on the website.

3. 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'

4. 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 <5 and p-value < 0.05.
    - Till 3rd model we have got both p value and VIF value under control.

5. Model Evaluation

    - A confusion matrix was created, and overall accuracy was checked which came out to be 81.88 %.

- 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 0.90.
    - After Plotting we found that optimum cutoff was 0.35 which gave

      Accuracy 81.05%
      Sensitivity 82.31%
      Specificity 80.26%.

- Precision – Recall:

  If we go with Precision – Recall Evaluation

  - On Training Data

    - With the cutoff of 0.35 we get the Precision & Recall of 79.47% & 71.46% 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 0.42 which gave

      Accuracy 81.66%
      Precision 75.47%
      Recall 77.77%

  - Prediction on Test Data

    Accuracy 81.51%
    Precision 75.0%
    Recall 77.7%

CONCLUSION

- The Sensitivity and Specificity, Accuracy, Precision and Recall score we got from test set are almost accurate.
- Important features responsible for good conversion rate or the ones which contributes more towards the probability of a lead getting converted in decreasing order:
  - Lead Origin_Lead Add Form
  - Total Time Spent on Website
  - What is your current occupation_Working Professional
  - Last Notable Activity_Other
  - Last Notable Activity_SMS Sent
  - Lead Source_Olark Chat
  - TotalVisits