Lead Scoring Case Study Summary

1.) Reading and Understanding Data.
- ➔ Read and analyze the data.

2.) Data Cleaning
- ➔ We have dropped the redundant variables that are not useful for prediction. We have dropped the columns containing null values/missing values which has(>40%)
- ➔ In case of categorical variable we have imputed the missing values with model(most occurrence)
- ➔ In the case of numerical variables we have handled the outliers with quantile value.

3.) Data Analysis
- ➔ Then we started with the Exploratory Data Analysis of the data set to derive further insight using visualization and analysis of predictor variables with the target variable.
- ➔ We have also done binning as required.
- ➔ We have checked the correlation among different variables.

4.) Test Train Split:
- ➔ We have divided the data set into test and train sections with a proportion of 70-30% values.

5.) Feature Scaling
- ➔ We used the Standard Scaling approach to scale the variables.

6.) Feature Selection and Model building
- ➔ We have selected the predictor variables using RFE.
- ➔ We have started building the model based on the features selected by RFE and the variables for the final model were selected based on p value(<0.05) and VIF(<5).

7.) Plotting the ROC Curve
We generated the ROC curve to determine the optimal cut-off point

8.) Finding the Optimal Cut-off Point
Based on the ROC curve and other evaluation metrics such as 'Accuracy,' 'Sensitivity,' and 'Specificity,' we determined that the optimal cut-off value is 0.4 in this case.

9.) Model Evaluation
We constructed a confusion matrix and calculated the model's accuracy. Additionally, we computed specificity, sensitivity, and the F1 score to assess the model's reliability.

Comparison between train and test metrics

Evaluation metric	Train Data Set	Test Data Set
Sensitivity	69.46	69.67
Specificity	96.23	95.95
Precision	91.85	91.95
Recall	69.46	69.67
Accuracy	86.08	85.46
F1 Score	79.1013	79.2746

In both the train and test data sets, the model demonstrates strong performance, with high accuracy, specificity, and precision. Sensitivity and F1 score are also relatively high, indicating a well-balanced model that can effectively classify positive and negative cases.

Conclusion:

- **Effective Model Generalization:** The model displays robust performance with consistency in evaluation metrics between the training and test sets, indicating it does not suffer from overfitting. This suggests the model's ability to generalize well to new data.

- **Room for Sensitivity Improvement:** While the model showcases good stability and overall performance, there is scope for enhancement in predicting lead conversion. Specifically, a higher sensitivity would better identify leads with a higher likelihood of conversion, aligning with the company's goal of targeting more promising leads.

- **Targeted Lead Scoring:** The logistic regression model successfully assigns lead scores that can be used to distinguish between "hot" and "cold" leads. The model's output can assist the company in more efficiently allocating resources and increasing lead conversion rates while closely aligning with the CEO's target of around 80% lead conversion.

##