Problem Summary

Step1: Reading and Understanding Data:

- Read and inspected the data.

Step2: Data Cleaning:

- The first step to cleaning the dataset we chose was to drop the variables having unique values.
- Then, there were a few columns with the value 'Select' which means the leads did not choose any given option. We changed those values to Null values.
- We dropped the columns having NULL values greater than 45%.
- Next, we removed the imbalanced and redundant variables. This step also included imputing the missing values as and where required with median values in case of numerical variables and the creation of new classification variables in case of categorical variables.
- We did bucket for a few columns having very less value count.
- Also dropped few columns which have more than 99% same values.
- All sales team-generated variables were removed to avoid any ambiguity in final
- solution.

Step3: EDA

- We did some univariant and bivariant data analysis using some visualization(I have used count plot and box plot)
- Outlier findings and treatment with the help of box plot visualization (treatment has been done with the percentile range of 0.05-.95)
- Also done heatmap to find the correlation and dropped highly correlated columns

Step4: Dummy Variables Creation:

- We created dummy variables for the categorical variables.
- Removed all the repeated and redundant variables

Step5: Test Train Split:

The next step was to divide the data set into test and train sections with a proportion of 70-30% values.

Step 6: Model Building:

- Scale the continuous variable columns
- Build our model using RFE (I have chosen 15 feature count)
- We should find the best model by using p_value < 0.5 and VIF < 5

Step 8: Plotting the ROC Curve and optimal cut-off point (in this case it is 0.4)

Step 7: Check the confusion matrix and evaluate the accuracy, recall, specificity, and precision