Summary Report of Lead Scoring Case Study

Overall Approach:

Step 1: **Reading and Understanding Data**:
Read and understand the metadata and the data.

Step 2: **Data Cleaning**:
Handled the categorical columns that had 'Select' as one of their levels. Post which we dropped the variables that had high percentage of NULL values in them. The columns that had less percentage of missing values (less than 10 %) were imputed with either the median or creating a new level in numerical and categorical variables respectively. The outliers were identified and removed.

Step 3: **Data Analysis:**
Then we started with the Exploratory Data Analysis of the data set to get a feel of how the data is oriented. In this step, there were some variables that were identified to have only one value in all rows. These variables were dropped.

Step 4: **Data Preparation:**

1. Encoded binary categorical variables to either 1 or 0.
2. Created dummy data for the categorical variables.
3. Divided the data set into test and train sections with a proportion of 70-30% values.
4. Used the Min Max Scaling to scale the original numerical variables.

Step 5: Modelling:

1. Selected the top important features using RFE.
2. Using the statistics generated, we recursively tried looking at the P-values and the VIF in order to select the most significant features that should be present and dropped the insignificant features.

Step 6: Plotting the ROC Curve:
Plotted the ROC curve for the features and the curve came out be with an area coverage of 90%.

Step 7: Finding the Optimal Cutoff Point:
Plotted the probability graph for the 'Accuracy', 'Sensitivity', and 'Specificity' for different probability values. The intersecting point of the graphs was considered as the optimal probability cutoff point, which came out to be 0.4.

Based on the new value observed that close to 81% values were rightly predicted by the model. We observed the new values of the 'accuracy=81.7%, 'sensitivity=81.3%', 'specificity=81.95%', and 'precision=74.11%'

Step 8: **Making Predictions on Test Set:**

Predictions made on the test model and we calculated the conversion probability based on the Sensitivity and Specificity metrics and found out the accuracy value to be 81.88%; Sensitivity=80.14%; Specificity= 82.94%, and precision=74.06%.

Learnings:

The model can be made more robust and flexible regarding the business problems that can arise in the future. The problems such as,

1. Converting almost all the potential leads so that these leads can be approached in an efficient way. To achieve this, the cutoff obtained by the Sensitivity-Specificity graph can be reduced to predict the desirable leads.
2. Minimize the resources of the company in procuring the leads: This can be achieved by considering a higher cutoff thereby increasing the Specificity. This will ensure that only those leads will be approached who have higher chances of being converted.