A brief summary report in 500 words explaining how you proceeded with the assignment and the learnings that you gathered.

Step 1: Importing and understanding the data

The data was imported and checked for understanding. The data frame given had more than 30 columns and therefore, it was essential to figure out the ones that do not benefit the analysis and eliminate them.

Step 2: Data cleaning

The Prospect ID and lead number columns, for example, were individual for each row and would not help the classification. These columns were dropped.

Similarly, columns were checked for null values and those having more than or equal to 45% were dropped. Some columns like 'Magazine' and 'Get updates on DM Content' had skewed values, i.e. more than 50% instances of one value. These columns may affect the analysis and were dropped.

In some columns, the instances of some values were very less and these values were grouped together into 'Others'. These steps decreased the number of columns at our disposal without adversely affecting the analysis. Some columns with lengthy names were renamed for ease of coding.

Step 3: Data preparation

The data was largely in categorical form and needed to be converted into numerical data for building a logistic regression model. The variables with more than 2 values were allotted dummy variables and binary data was converted to values of 1 and 0.

After this was done, outliers had to be dealt with. The describe function gave an idea of the difference between the values at 25^{th} percentile, 75^{th} percentile, the median and the minimum and maximum values. Box plots were also used for visualization to get a better understanding of outliers. Based on the extent of outliers, the data was capped. This reduced the number of outliers which would help make the model more accurate.

Step 4: Visualizing the data

The data was visualized using bar graphs to understand the extent of conversion seen in various levels of a variable. This was done before converting data into numerical form.

Step 5: Looking at correlations

A heatmap was used to look at the correlation between the various variables. Variables with high collinearity were dropped depending on their role in the analysis. This would help in making the model more accurate.

Step 6: Model building

The data was split into test and train sets. The training set was scaled using fit-transform. The model was built based on the train set and run to check p values. RFE was used for feature selection. After RFE, VIF values were checked for manual elimination of variables. Since none

of the variables had a variance factor more than 5, they were retained in the model. The accuracy of the model was checked and the confusion matrix was built.
The ROC curve was plotted to check the threshold which would be the optimum probability cutoff. The precision and recall values were observed to check the trade-off.

Step 7: Making prediction on the test set

The model was run on the test set and various parameters like accuracy, sensitivity and specificity were checked to evaluate the model. Since the prediction seemed quite accurate, the model was finalized.