SUMMARY REPORT-LEAD SCORING CASE STUDY

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

2. **Introduction-**
   In this report, we will describe the steps we took to build the model, the challenges we faced, and the insights we gained from the project.

3. **The Steps followed for Model Building are as follows-**

   - **Reading and Understanding_the_Data**
     We used the **.info()** and **.describe()** function do see the variables data types, null values presence, outliners and statistical measures of numerical attributes such mean, mode and std. deviation.

   - **Data_Cleaning & EDA-**
     We dropped column with more than 45% null values. Also, checked for duplicate values, dropped the variables with highly skewed data, created separate category for null values as "missing values" so that we do not loose much data and can analyze them later. We also merged categories that were showing similar trend and are present in very less proportion to reduce the no. of variables later. Dropped binary variables that has more than 95-96% representation of 1 value. Later we performed EDA on numerical variables are removed outliners.

   - **Data_Preparation-**
     After EDA, we prepared data for analysis by creating dummies (one-hot encoding) for categorical variables and "1/0" mapping for binary variables.

   - **Test-Train_Split-**
     Then we split the data into training and test set (0.7 & 0.3 respectively).

   - **Feature_Scaling-**
     Next, we used standardizing technique to scale the numerical variables.

   - **Checking_Correlations-**
     Next, we used heatmap to see correlation among independent variables and dropped variables having high correlation both -ve and +ve. (>0.7 & >-0.7)

   - **Model_Building-**
     After all preprocessing steps, we build our first model with all variables. Nut as the number of variables were too high, we decided to go for RFE automated feature elimination.

SUMMARY REPORT-LEAD SCORING CASE STUDY

- **Feature_Selection_Using_RFE-**
  We kept the threshold of 15 variables and used model to eliminate the less important variables.

- **Manual_Feature_Elimination-**
  After RFE we build our 1st model and got accuracy of 81% approx. But there were many variables which have high p-values (>0.05) and VIF (>5) and are insignificant & redundant for model. So, we dropped few and arrived at final model after 7 iterations with accuracy of 80% approx.

- **Plotting_ROC_Curve-**
  We used ROC curve to see model accuracy as area under ROC curve. We got 0.88, which was pretty good. (Maximum being 1)

- **Finding_Optimum_Cutoff_Point-**
  Found threshold value by curve between "Sensitivity, Specificity & Accuracy" for max. accuracy. We got 0.38 as optimum value and using that got model accuracy of 80.5 %.

- **Prediction_on_test_set-**
  Next, we made predictions on test set and got model accuracy of 79%.

- **Assigning_Lead_Score-**
  At last, we assigned the lead score from 0-100 to leads by multiply the predicted probability by 1.

4. **Insights and Learnings-**
   Firstly, we learned that data preprocessing is a crucial step in model building. We also learned that evaluating the model using multiple metrics provides a more comprehensive understanding of its performance than using a single metric.

5. **Conclusion-**
   The model was able to predict 79% of total lead conversion, meeting the CEO's target conversion rate.