SUMMARY REPORT

The Approach followed in order to achieve the objective of the assignment:

1. **Data Cleaning and Analysis:**
    a. We first clean the categorical columns having just one type of data as it would not contribute to the analysis
    b. There were some categorical columns having 'Select' and 'unknown' as label which is equivalent to null values hence treated them as null values
    c. We then check the null percentage column wise and dropped the columns having more then 35% of null values
    d. Then the row wise null values were handled by dropping just the rows which is having null values as the count was less
    e. We used Countplot, Boxplot and Heatmap in order to perform data analysis

2. **Data Transformation and Preparation**:
    a. Once the data was analysed and clean, we mapped the variables having Yes and No values to 1 and 0
    b. We also create dummy variables for the categorical columns
    c. We then split the dataset into two sets train and test in the ratio 70:30
    d. We then perform fit and transform scaling on numeric variable on Train dataset

3. **Model Building**:
    a. We used RFE variable selection method and selected 15 most appropriate variable for our model
    b. We also calculated the corresponding VIF for the variables
    c. We then checked for insignificance and multi collinearity in the variable and handled them

4. **Model Evaluation and Prediction**:
    a. After getting the most appropriate features, we performed the prediction on the trainset, getting accuracy score as 81% at 0.5 as cutoff
    b. We then created ROC Curve and found 89% area under the curve

5. **Optimalization and Accuracy Score**:
    a. In order to perform better optimization to our model, we then plotted graph between Accuracy, Specificity and Sensitivity and got the Optimal Cutoff value to be 0.38.
    b. We also plotted precision and recall graph
    c. Considering 0.41 as the final Optimal Cutoff , we calculated the Accuracy , Precision and Recall score on the train and test data

Learning Gathered:

- We have created a column named Lead Score hence model could be run on the production data and then follow the Lead Score rating. Higher the score, higher the chances of getting it converted to **Hot Lead**
- The optimal cut off we found is 41%
- With this cut-off value we are getting 81.5% accuracy on train data and 80.9% accuracy on Test data which looks good.

- Also, we can see the Precision and Recall we got are as follows:
    - Train Data- Precision is **75%**, Recall is **77%**
    - Test Data- Precision is **73%**, Recall is **76%**
- Important features which contributes more towards the probability of a lead getting converted are:
    - Total Time Spent on Website
    - Lead Source_Welingak Website
    - What is your current occupation_Working Professional
    - Lead Origin_Lead Add Form
    - Last Notable Activity_SMS Sent
- Negatively impacting the conversion that should be avoided are:
    - Specialization_Not Provided
    - What is your current occupation_Not Provided
    - Do Not Email
    - Lead Origin _ Landing Page Submission