Summary Report -

1. Import and Inspect the input data: The lead datasets consist of 9240 records with 37 attributes which include categorical and numerical columns are available.

2. Data Cleansing / Reduction:

- Dropped most of the variables with more than 15% missing value. Retained couple of variables for assessment based on the business understanding though >15% data was missing. Data imputation performed for the missing values based on standard techniques.
- As 'Select' is not a valid class and treated as default value set in the form dropdowns, the same is considered as null values and leveraged as the key point to drop couple of columns.
- Columns with only one unique value with no variance are dropped from assessment.
- Created new bins for the categorical variables with multiple levels to fewer values.
- Datatype Formatting and variable renaming perform for ease of reference for analysis.

3. Data Preparation and Visualization:

- Outlier Analysis and Treatment: Box plots
- Data Value Imputation (Statistical Imputation) : Calculated median, mode on based on the variable data types
- Count Plot of different categorical variables with Label = 'Converted' to visualize the data distribution
- Perform encoding and dummy features for the categorical variables
- Train-Test Split: Dataset has been split into Train and Test in 70:30 ratio
- Performed MinMax Scaling on Train data
- Created pair plots, heatmaps to analyse the correlations between the variables

5. Feature Selection and Model Building:

- Used automated RFE technique followed by manually elimination of features one by one.
- Total 2 models were built and after each model building p-values of all beta-coefficients and VIFs have been checked simultaneously. Accepted p-value is lower than .05 and VIF < 5.
- Checked Overall model accuracy, Confusion Matrix after each new model to analyse the model performance.

6. Model Evaluation and Prediction:

- Applied random Probability cutoff on Train Data
- Calculated specificity, sensitivity, and accuracy of the model for various probability cut-off based on the specificity, sensitivity, and accuracy intersect.
- Applying the Optimal Probability cutoff & Evaluating on Train Data followed by Test data evaluation to predict the Target Variable as 0 or 1. Additionally, created Lead Score variable to indicate if the leads are HOT or COLD
- The metrics seem to hold on the test dataset as well. So, it looks like the decent model being created for the Leads dataset as the metrics are decent for both the training and test datasets

7. Observations:

- Top 3 variables which contribute the most towards the probability of a lead getting converted
- Top 3 categorical variables which contribute the most towards the probability of a lead getting converted
- X Education is suggested to consider recommended variables given the influence on the end results and plan accordingly to increase the conversions. Additionally, we can definitely explore more options in adding or deleting the features to get better results than the already existing Model 2 which is overall a good model based on the output generated