SUMMARY

A brief report explaining the approach of the Case Study and insights gathered.

1. **Data Reading & Understanding:**
    - First data set is imported.
    - Then checking data types, shape and information of data set.
2. **Data Cleaning:**
    - Check null values, outliers & removing redundant variables.
    - There is a category 'Select' which is equal to null because it is category which is left by the customers, so replace SELECT with NaN using np.nan function.
    - Then delete all variables that have null values more than 40% as it would of no use.
    - Then imputing rest null values by the mode.
    - Merge some categories into one that has same meaning or can be counted in 'Others' type.
    - Lastly remove rows which has NaN values using dropna function.
    - Delete all the variables whose value counts is skewed towards one category or only have one category as they are of no use in building model.
    - Delete other redundant variables like prospect id and lead number.
    - Check for outliers in continuous variables. By checking we see two columns have outliers in 99^{th} percentile so delete 1 percentile from top and bottom to remove outliers.
    - Perform EDA on the variables left and gather insights.
3. **Data Preparation:**
    - Map binary variables into 0 and 1.
    - Create dummy variables for all the categorical variables. Check correlation matrix to check correlations between various variables.
    - Split the data set into train and test data set for model building.
    - Scale the continuous numerical features using Standard Scalar of train set.

4. **Model Building:**
    - Using RFE, we reduced variables to 20 so that it is easy to check variables p-values and VIF values for building a good model.
    - Then we eliminate variables one by one by checking their p-values and VIF values and build an efficient model with p-value < 0.05 and VIF value < 5.
5. **Model Evaluation:**
    - Predict the probability of conversion by choosing random cutoff of 0.5. Check various metric like Accuracy, Sensitivity, Specificity, Precision and Recall.
    - Draw ROC curve.
    - Draw sensitivity, specificity and accuracy graph to find optimal cutoff for prediction.
    - Draw precision and recall graph to find optimal cutoff for prediction.
    - By choosing sensitivity, specificity and accuracy graph, we choose optimal cutoff of 0.3.
    - Again, check various metric like Accuracy, Sensitivity, Specificity, Precision and Recall. We got a good value for the same.
6. **Predictions on Test Data:**
    - Scale continuous numerical features using standard scalar.
    - Then predict the probability of conversion.
    - Check various metric like Accuracy, Sensitivity, Specificity, Precision and Recall. We see that these scores are in acceptable region and our model is good so far.
7. **Conclusion:**

    Train set:

    - Accuracy: 87.88%
    - Sensitivity: 88.21%
    - Specificity: 87.68%
    - Precision: 81.48%
    - Recall: 88.21%

    Test Set:

        - Accuracy: 87.34%
        - Sensitivity: 86.13%
        - Specificity: 88.06%
        - Precision: 81.30%
        - Recall: 86.13%
    - Top features are:
        - Tags_Will revert after reading the email
        - Lead Origin_Lead Add Form
        - Tags_Other
        - Tags_Not Specified
        - Lead Source_Welingak Website
    - Overall model seems to be good and able to adjust with company's required changes made in coming future.