Summary Report

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

Following are the things that was done to achieve the objectives of the assignment:

1. **Data Cleaning**
    a. Removed redundant columns which was not necessary in our analysis.
    b. After removing the redundant columns, there were columns who had labels called "Select" which were present because the customer did not fill that entry. Hence, we had to be converted it into null values so that it can be dealt with later.
    c. Removed columns having missing values percentage more than 40%
    d. After removing, we checked the remaining columns for missing values and imputed them with the mode (Highest frequency value)
    e. Some of the columns were having biased values which means that some entries in that column had exceptionally low frequency (Both Converted and Non-converted cases) as compared to the high frequency values. Hence collating all those entries in those columns was best to do for further analysis.
    f. Removed those columns which had data imbalance.
    g. Then we checked for outliers and successfully removed them.

2. **EDA (Exploratory Data Analysis)**
    a. We did some bivariate analysis of the numerical variables and categorical variables and drew inferences out of the plots drawn.
    b. Also, we did multivariate analysis by plotting heat maps.

3. **Data Transformation**
    a. We created dummy variables for all categorical variables. Here we dropped a level of our choice and not the first column after creating dummy variables.
    b. We transformed the binary variables into 0 and 1.
    c. Removed all the redundant columns.

4. **Data Preparation**
    a. We split the data into train and test dataset.
    b. Then we used Standard Scaling on these datasets.
    c. After that we plotted a heatmap of all the variables to see the correlation of variables with each other.

5. **Data Modelling**
    a. We made a model using RFE with 20 variables in it. We made a total of 5 models and our final model had P-Values less than 5% and VIF less than 5 and calculated the accuracy of the train dataset.
    b. We created the confusion matrix and calculated various metrics like Sensitivity, Specificity, Positivity Predictive Value, Negative predictive value, False Positive Rate, and the True Positive Rate for the train dataset.
    c. We plotted the ROC Curve for having a gist of the predictive power of the model.
    d. We found the optimal cut off point and then came up with new outcomes. Then we repeated step b again. Then we plotted the Precision-Recall Curve.
    e. After this we scaled the test data set and did predictions on this test dataset.
    f. We came up with our final outcomes and repeated step b on this test dataset.
    g. We added a column called lead score which is a measure of a potential lead.

6. Conclusions

Following learnings were gathered:

a. Test dataset has a good Accuracy, Sensitivity (Recall) and Specificity.

b. Our final model has good metrics which only goes to say that it will cater to the company's requirement changes in the future, if there are any.

c. The top features in the model are:

   i. **Tags_Closed by Horizzon**
   ii. **Tags_Lost to EINS**
   iii. **Lead Origin_Lead Add Form**
   iv. **Tags_Will revert after reading the email**