Summary Report

There are multiple ways of proceeding with an assignment, but we divided the assignment into following components, and followed step-by-step approach to achieve the desired results:

Step 1: Importing Data

We imported the dataset into our jupyter notebook.

Step 2: Inspecting the Data frame

We checked the shape, inspected column types, and looked at the numeric columns ranges.

Step 3: Data Preparation and looking at correlations

In this step, we prepared the data for subsequent modelling. Following actions were taken in this step:

1. Dropped 'Prospect ID' and 'Lead Number' columns, as it's not adding any value to the result.
2. Performed outlier treatment of numerical columns after plotting box plots and found out their correlations.
3. Different levels of categorical columns were observed. Misspelled levels were corrected, null was imputed for 'Select' value and multiple levels were clubbed together to form a single level. Categorical columns with 1 or few levels were dropped.
4. Frequency plots were plotted for categorical values.
5. Null values were imputed either with mean or mode depending on the type of column.
6. Pair plots were plotted, and dummy variables were imputed for categorical column levels.

Step 4: Test-Train Split

Data set was split into test, and train data set. Conversion rate of whole dataset was also checked.

Step 5: Feature Scaling

Applied feature scaling on test set, and correlation heat map was plotted.

Step 6: Logistic Regression Model

Initial model with all features was created

Step 7: Feature selection using RFE

Initial set of 15 features was selected through RFE, and then based on p-value, and VIF score of those features, few columns were removed from the selected lot.

Step 8: Plotting the ROC curve and Finding optimal cutoff point

Parameters such as Sensitivity, specificity, False positive rate, Positive predictive value were calculated. Also, ROC (Receiver operating characteristic) curve was plotted.

Also, specificity, sensitivity, and accuracy values over different lead score cut off(s) were plotted over a graph, and an optimal cutoff point was identified. Based on optimal cutoff point, model evaluation parameters were again calculated.

Precision and Recall values were calculated, and their trade off was plotted.

Finding the cutoff value of the lead score to identify a converted user

Step 9: Making predictions on the test set

Applied feature scaling to test set, selected the same columns identified by RFE, and applied Logistic regression model.

Model evaluation parameters were calculated for the test set and compared against the train set parameters.

Learnings from the Assignment

Parameters such as *Lead Source_Welingak Website, Tags_Busy, Tags_Closed by Horizzon, Tags_switched off, Tags_Lost to EINS, Asymmetrique Activity Index_03.Low, Lead Origin_Lead Add Form, Lead Quality_Worst, Lead Quality_Not Sure, Tags_Ringing, Tags_Will revert after reading the email, Last Notable Activity_SMS Sent* are immensely helpful in determining the conversion of a professional.

Sensitivity or True positive rate on test set is 86%, which is way over 80% requirement of the management. Also, changing the cut off value can change the sensitivity score. Thus, throttling cut off value, can provide more leads once internship period starts, so that company can do additional communication to increase revenues.