SUMMARY (BY MANJEET YADAV)

This analysis is done for X Education to select the most promising leads, i.e., the leads that are most likely to convert into paying customers. This requires building a model wherein we need to assign a lead score to each of the leads such that the customers with higher lead score have a higher conversion chance and the customers with lower leads score and a lower conversion chance.

Approach for the Analysis and Logistic regression model building and evaluation:

- **Inspecting data:**
    - Shape of dataset.
    - Information of dataset.
    - Descriptive statistics of numeric columns.
- **Exploratory Data Analysis**:
    - **Data Wrangling**:
        - Checking and handling duplicate records.
        - Checking the null value percentage and dropping the variables having null value percentage more than 50%.
        - Dropping the variables which are not useful for analysis like Prospect ID and Lead Number.
        - Checking the null values in Lead Quality variable and imputing the null values with “Not Sure” category as this variable seems important.
        - Dropping the redundant columns like 'Asymmetrique Activity Index', 'Asymmetrique Profile Index', 'Asymmetrique Activity Score', 'Asymmetrique Profile Score’ as 45% of the records are missing and imputation won’t give promising result.
        - Similarly, checking all other variables one by one and imputation done accordingly.
        - Identified extreme outliers, that can potentially skew results when analysing and handled accordingly.
        - Identified discrepancies and either explained or removed them.
        - Considered “Select” as a null value.
        - Extracted insights from the data.
- **Data Pre-processing**:
    - Dummy encoding:
        - Transforming categorical columns to dummy variables.
    - Feature Scaling:
        - Normalizing numeric columns.
    - Train-Test Split:
        - Split dataset in the ratio 70:30.
- **Model Building**:
    - **Automated approach:**

        - Used Recursive Feature Elimination to get the top 15 relevant features first.
    - **Manual approach**:
        - Checked Variance Inflation Factor and p-values to further drop insignificant predictors.
- **Model Validation**:
    - The model includes statistically significant and important features.
    - The goodness of fit is measured by Log-likelihood and Pearson chi-squared measures.
- **Model Evaluation**:
    - Visualized Confusion Matrix.
    - Found the optimum cut-off threshold as 0.2, and plotted their respective accuracy, sensitivity and specificity of the model.
    - Metrics obtained on train dataset:
        - Sensitivity: 0.86
        - Specificity: 0.94
        - False Positive Rate: 0.05
        - Positive Predictive Value: 0.91
        - Negative Predictive Value: 0.91
        - F1 Score: 0.88
        - Accuracy: 0.91
    - Metrics obtained on test dataset:
        - Sensitivity: 0.84
        - Specificity: 0.95
        - False Positive Rate: 0.05
        - Positive Predictive Value: 0.90
        - Negative Predictive Value: 0.91
        - F1 Score: 0.87
        - Accuracy: 0.91
    - Plotted Receiver Operating Characteristic and calculated the Area Under Curve: 0.95 for both train and test dataset.
    - From the precision recall curve, 0.25 is the optimum point to take as a cutoff probability. We can check our accuracy using this cutoff too.
- **Assigning lead score:**
    - Lead Score = 100 * ConversionProbability
    - This needs to be calculated for all the leads from the original dataset (train + test)
- **Summary:**
    - Features having positive impact on conversion probability in decreasing order of impact:
        - Tags_Lost to EINS
        - Tags_Closed by Horizzon
        - Tags_Will revert after reading the email
        - Tags_Busy

    - Lead Source_Welingak Website
    - Last Notable Activity_SMS Sent
    - Lead Origin_Lead Add Form
  - Features having negative impact on conversion probability in decreasing order of impact:
    - Lead Quality_Worst
    - Lead Quality_Not Sure
    - Tags_switched off
    - Tags_Ringing
    - Do Not Email
  - Focusing on the above predictors, X Education can aim to select the most promising leads.