SUMMARY

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 wit higher lead score have a higher conversion chance and the customers with lower leads score and a lower conversion chance.

Framework of approach to perform analysis:

- <u>Inspecting data:</u>
    - Shape of dataset.
    - Information of dataset.
    - Descriptive statistics of numeric columns.
- <u>Exploratory Data Analysis:</u>
    - <u>Data Wrangling:</u>
        - Changing null values and standardizing the data.
        - Identified extreme outliers, that can potentially skew results when analysing and handled accordingly.
        - Identified discrepancies and either explained or removed them.
        - Visualized columns with imbalanced categories and dropped them.
        - Considered "Select" as a null value.
        - Extracted insights from the data.
- <u>Data Pre-processing:</u>
    - Dummy encoding:
        - Transforming categorical columns to dummy variables.
    - Feature Scaling:
        - Normalizing numeric columns.
    - Train-Test Split:
        - Split dataset in the ratio 70:30.
- <u>Model Building:</u>
    - Automated approach:
        - Used Recursive Feature Elimination to attain the top 15 relevant features.
    - Manual approach:
        - Checked Variance Inflation Factor and p-values to further drop insignificant predictors.
- <u>Model Validation:</u>
    - The model includes statistically significant and important features.
    - The goodness of fit is measured by Log-likelihood and Pearson chi-squared measures.
    - Analysed residual deviance and studentized Pearson residuals with respect to the fitted-values and visualized the plot has parallel lines with zero intercept which indicates that there isn't significant model inadequacy.
- <u>Model Evaluation:</u>
    - Visualized Confusion Matrix.
    - Found the optimum cut-off threshold as 0.3, and plotted their respective accuracy, sensitivity and specificity of the model.
    - Metrics obtained on train dataset:
        - Precision: 0.87
        - Recall: 0.90
        - Specificity: 0.92

        - Accuracy: 0.91
        - False Positive Rate: 0.04
        - F1 Score: 0.91
    - Metrics obtained on test dataset:
        - Precision: 0.87
        - Recall: 0.90
        - Specificity: 0.91
        - Accuracy: 0.91
        - F1 Score: 0.91
    - Plotted Receiver Operating Characteristic and calculated the Area Under Curve: 0.96 for both train and test dataset.
- Summary:
    - The significant predictors are obtained as:
        - Total Time spent on website
        - Lead Origin_Lead Add form
        - Lead Source_Welingak Website
        - Do Not Email_Yes
        - Tags_Closed by Horizzon
        - Tags_Lost to EINS
        - Tags_Ringing
        - Tags_Will revert after reading the email
        - Tags_Switched off
        - Last Notable Activity_Misc_Last Notable Activity
        - Last Notable Activity_SMS Sent
    - Focusing on the above predictors, X Education can aim to select the most promising leads.