SUMMARY REPORT
(Lead Score Case Study)

Problem:

X Education wants to build a model where they assign a lead score to each lead such that the customers with a higher lead score have a higher conversion probability. The business requirement is to increase the lead conversion rate to around 80%.

Solution Approach:

1. **Data cleaning:** The data had numerous null values and multiple columns with the value "Select." A few columns also had data imbalances. Each of these situations was examined, and the proper management method was applied. As an illustration
    - More than 50% of the null values in a column were removed.
    - In a select number of significant columns, 'Not Provided' or 'Others' was used in place of null values.
    - Unbalanced data columns, such Country, were removed.

2. **EDA:** EDA was carried out on the cleansed data.
    - Categorical and numerical variables underwent a univariate analysis.
    - The 'Converted' variable (Target Variable) was used in bivariate analysis of significant variables.
    - Less important categories in a few of the columns were combined based on graphs.
    - Using the 1.5 IQR Method, outliers discovered during EDA were handled.

3. **Data Pre-processing:** The subsequent pre-processing actions were taken.
    - Yes/No binary variables were changed to 1/0.
    - For each categorical column, N-1 dummy columns were produced for the supplied N categories.
    - Data were divided into training and test datasets in a 70:30 ratio.
    - Continuous variables were used for feature scaling.

4. **Model Building:** On the training dataset, logistic regression was carried out using the methods listed below.
    - To identify the top 15 relevant factors, a first RFE was conducted.
    - A model was iteratively constructed using these 15 variables, and for each model, VIF and p-values were noted.
    - The model was rebuilt at each stage when variables with VIF > 5 or p-value > 0.05 were discarded one at a time.

5. **Model Evaluation:**

   - Predicted values on the training dataset were derived using an arbitrary cut-off of 0.5, where leads with conversion probabilities less than 0.5 were labelled with "0," and vice versa.

   - Accuracy (92%) sensitivity (86%) and specificity (95%) were assessed using a confusion matrix.

   - The ideal cut off was calculated to be around 0.2 after the ROC curve was plotted.

   - Precision-Recall trade-off was seen and accuracy (92%) sensitivity (88%) and specificity (94%) were all reevaluated.

6. **Predictions:** The following processes were used to make predictions based on test data.

   - Scaling was done on test data's continuous variables.
     On this dataset, predictions were made using the developed model with a fixed cut-off of 0.2.

   - Accuracy (92%) sensitivity (88%) and specificity (94%) were calculated using a confusion matrix. This allowed us to draw the conclusion that our model works well with unobserved data.

   - Each lead was then assigned a lead conversion score (lead conversion score = conversion probability * 100).

   - The most significant characteristics that affect the likelihood of conversion were recognized.

The main learnings gathered from this assignment were:

1. Data exploration and handling of missing values
2. The significance of pre-processing EDA and data.
3. Model construction strategy, feature selection, and the effect on training and test datasets.