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 contained a lot of null values, and 'Select' value in multiple columns. Few columns had Data imbalances as well. Each of these scenarios was analysed and appropriate handling technique was used. Example –

    - Columns with high null values (More than 50%) were dropped.
    - For few significant columns, null values were replaced with 'Not Provided'/'Others'.
    - Columns with data imbalances such as Country was dropped.

2. **EDA:** On the cleaned data, EDA was performed.
    - Univariate Analysis of Categorical and Numerical variables was performed.
    - Bivariate Analysis of important variables was performed with 'Converted' variable (Target Variable)
    - Based on graphs, less significant categories in few of the columns were clubbed into one.
    - Outliers observed during EDA were treated using 1.5 IQR Method.

3. **Data Pre-processing:** The following pre-processing steps were performed.
    - Binary Variables Yes/No were converted to 1/0
    - N-1 Dummy columns were created for given N categories for each categorical column.
    - Data was split into training and test dataset in the ratio of 70:30.
    - Feature Scaling was performed on continuous variables.

4. **Model Building:** Logistic Regression was performed on the training dataset using the following steps.
    - First RFE was done to attain top 15 relevant variables.
    - Using these 15 variables, model was built in iterative manner where VIF and p-values were observed for each model.
    - Variables with VIF > 5 or p-value > 0.05 were eliminated one by one and the model was rebuilt at every stage.

5. **Model Evaluation:**
    - Predicated values on the training dataset were obtained by using 0.5 as arbitrary cut-off, where in leads with conversion probability < 0.5 were tagged '0' and vice versa.
    - Confusion matrix was created using which accuracy(92%), sensitivity(86%), and specificity(95%) were calculated.

- ROC curve was plotted and optimal cut off was calculated to be around 0.2.
- Accuracy(92%), sensitivity(88%), and specificity(94%) were re-evaluated and Precision-Recall trade-off observed.

6. **Predictions:** Predictions on test data was made using the following steps.
    - Scaling was performed on continuous variables of test data.
    - Using the model built and cut-off fixed at 0.2, predictions were made on this dataset.
    - Confusion matrix was created using which accuracy(92%), sensitivity(88%), and specificity(94%) were calculated.
    - This helped us conclude that our model is performing well on unseen data.
    - Finally lead conversion score was given to each lead
      (Lead conversion score = conversion probability * 100)
    - Most important features that influence the conversion probability were noted.

The main learnings gathered from this assignment were:

1. Process of exploring data and handling missing values
2. Importance of performing EDA and Data pre-processing.
3. Approach for building model and feature selection and its impact on training and test dataset.
4. Finally, solving problem with team effort and playing by our strengths.