SUMMARY REPORT

X Education company sells their online courses to industry professionals. They wanted to do the analysis on the past data provided and build a model to identify the potential leads. They wanted to increase their lead conversion target to 80%. Since it is a classification prediction problem, we must build Logistic regression model.

Following steps are performed for EDA and Model build:

1. **Data Reading and Understanding:**
    - Imported the given data in Jupyter notebook.
    - Checked first few rows of a dataset.
    - Inspected no. of rows and columns present in the dataset and its datatypes.
2. **Data Cleaning:**
    - Dropped few columns which are unwanted for our analysis.
    - Dropped the columns which are having NULL values around 36% of given data.
    - Imputed other columns which are having NULL values with Unknown or 0.
    - Deleted duplicate records
    - Handled outliers for numerical variables with threshold as 99^{th} percentile.
3. **Data Visualization:**
    - Data analysis done for categorical variables (binary and multi-level) with respect to target variable (Converted) using count plot and found that some of them were useful, and others weren't.
    - Data analysis done for numerical variables with respect to target variable (Converted) using box plot and seems like there were some useful information.
    - Also checked correlation matrix to check the correlation between numerical variables.
4. **Data Preparation for Model Build:**
    - Based on EDA, dropped one of the categorical variables which has only one outcome and doesn't serve any purpose for further analysis.
    - Converted binary categorical variables with outcomes of 'No' and 'Yes' to 0 and 1 which will help for model building.
    - Created dummy variables for all multi-level categorical variables and dropped the original ones.
5. **Model Building:**
    - Splitted the dataset into train and test sets with proportion of 70% and 30% respectively.
    - Performed MinMaxScaler technique in train set to rescale and bring all the independent variables to same scale.
    - Used RFE (Recursive Feature Elimination) initially to select top 15 variables that are useful for model build.
    - Performed iterative operations of removing the features one by one until we achieved P-Value < 0.5 and VIF < 5 for all the features.

6. Model Evaluation on Train set:

- Performed MinMaxScaler technique in test set to rescale the data.
- Initially took 0.5 as cut-off value for prediction and validated metrics such as accuracy, sensitivity and specificity but found that the cut-off value is not good enough for prediction.
- Later we re-calculated cut-off values for the probabilities range between 0 and 0.9 and found that a cut-off value of 0.35 provides good sensitivity and specificity.
- We predicted train set with this cut-off value and looks good.

7. Prediction on Test set:

- After we predicted the test set based on the optimum cut-off value, below are the evaluation results.

Metrics	Train Set	Test Set
Model Accuracy	**79%**	**77.7%**
Sensitivity	**81.25%**	**82.02%**
Specificity	**77.64%**	**74.9%**

8. Conclusion:

- We can conclude that the final logistic regression model built isn't fit by chance and well generalized for prediction.