Summary

A machine learning model has been built for X education company which gave us a lot of information about how the potentials customers visit the site, the time they spend over there, then how they reached the site and the conversion rate.

The following technical steps are employed:

1. Data Reading & Understanding:

   - ✓ Initially dataset was loaded to notebook & analysed for basic checks. Further steps & operations were devised for this dataset.

2. Data Wrangling:

   - ✓ First step to clean the dataset we choose to remove the redundant variables/features.
   - ✓ The data set was partially clean except for a few null/ NAN values and the option 'Select' has to replace with a null value since it did not give us much information.
   - ✓ Dropped the features which had little to no information about the leads in elementary phase.
   - ✓ Checked for duplicated values in each Categories for all Categorical column.
   - ✓ Checked for number of unique Categories for all Categorical columns.
   - ✓ Treated the missing values by imputing the favourable aggregate function like mode for categorical variables & median for continuous variables.

3. Data Visualization:

   - ✓ Dataset was checked for data imbalance at first.
   - ✓ A quick EDA was done to check the condition of given dataset. It was found that a lot of elements in the categorical variables were irrelevant. The numeric values seem good but had outliers.
   - ✓ Performed Univariate Analysis for both Continuous and Categorical variables.
   - ✓ Bivariate Analysis with respect to Target variable.
   - ✓ Pair plot was made for quick check of trend & corelation.
   - ✓ As per the findings few columns were dropped before moving to next phase.

4. Data Preparation:
   - ✓ Conversion of some binary variables were performed from Yes/No to 1/0
   - ✓ The dummy variables are created for all the categorical columns. (One-hot-encoding)

5. Model Building:
   - ✓ The Spit was done at 70% and 30% for train and test the data respectively with random state value as 100
   - ✓ For feature scaling was done for Continuous variables.

- ✓ By using RFE with provided 15 variables. It gives top 15 relevant variables
- ✓ A total of 6 iteration was done to reach the final model with all values being in limit for p-values & VIF. (The variables with VIF < 5 and p-value 0.05 were kept).

6. Model Evaluation & Prediction:

- ✓ 1st Train dataset split prediction was performed for converted variable.
- ✓ Later on, the optimum cut-off value by using ROC curve (area = 0.96) was used to find the accuracy, sensitivity and specificity which came to be around 90%.
- ✓ The optimal cut off is at 0.195 & was used to calculate metrics again.
- ✓ Afterwards with this value prediction was done the test dataset split with previous selected features & cut-off value.

7. Final Observation:

✓ The values obtained for Train & Test dataset with the cut-off as **0.195**:

Metrics	Train Data	Test Data
Accuracy (%)	90.16	90.25
Sensitivity (%)	88.48	89.40
Specificity (%)	91.20	90.81