Summary Report

OBJECTIVE:
X Education sells online courses to industry professionals. The typical lead conversion rate at X education is around 30%
The **GOAL** of this project is to build a logistic regression model based on the past leads dataset to assign a lead score between 0 and 100 to each of the leads which can be used by the Sales Team in the company to target high potential leads and increase the lead conversion rate.

SOLUTION STEPS:

1. **Data Inspection, Cleaning:**
   a. Inspected the dataset to look at all the features in the provided dataset to identify the ones that would be useful from the business point of view. It contained the system generated columns along with few columns added by the Sales Team later. We only kept the system generated columns.
   b. Next were cleaning steps like removing columns with high percentage (>40) of missing values and imputing the others.
2. **Exploratory Data Analysis:**
   a. Analyzed features to see its impact on conversion and observed - leads that come through Reference have the highest conversion rate as compared to other Lead Sources.
   b. Another inference - leads that chose ‘Working Professionals’ as their occupation have higher chance of converting.
3. **Data Preparation:**
   a. Converted all the binary variables from Yes/No to 1/0
   b. Created dummy features for categorical variables with multiple levels.
   c. Next we split the provided dataset into Train and Test
   d. For the continuous numerical variables, we performed feature scaling.
4. **Hybrid Feature Selection (RFE & Manual):**
   a. We first performed RFE with 15 features as its output on the train data set.
   b. Next we checked for the VIF values to check for multicollinearity and remove the ones that are higher than 5.
5. **Model Building & Evaluation:**
   a. With GLM model building technique we built and iterated the model, manually removed the variables which had a high P value till the point all the necessary variables are in check and are suitable for our final model.
   b. Performed the confusion matrix on the train data and plotted the ROC curve and assessed the optimum cutoff probability point (0.22).
   c. For **train data**, we achieved Accuracy = 76%, **Sensitivity = 82%,** Specificity = 73%
   d. Using the model to predict on the **test data**, we observed Accuracy = 75%, **Sensitivity = 82%,** Specificity =71%
   e. We also added a column ‘Lead Score’ that provides a score to each lead between 0-100, with higher score meaning higher probability of converting.

CONCLUSION:

We successfully built a model for X Education to better predict whether a lead will convert based on the ‘Lead Score’ assigned to them. This can by leveraged by the Sales Team to reach more high potential leads first.

Though the model has good sensitivity which is what we need here to make sure we don’t miss out on high potential leads, it’s imperative that X Education continues to maintain and enhance the model frequently to keep up its accuracy.