SUMMARY

PROBLEM STATEMENT:

X Education is an education services provider which is able to get only 30% conversion from her sales efforts. The company has consulted us to devise a predictive model which would classify their leads as hot or cold so that the sales team could direct their focus on the hot ones for higher conversion rates. The CEO has set a target of 80%.

Solution Approach:

In order to classify the leads, a simple Logistic Regression model was devised. The model was revised a number of times to ensure high accuracy and sensitivity.

1. The data presented has over 9240 rows and 37 columns.
2. The data was imported to a Pandas Dataframe and some basic sanity checks such as checking the shape, data types and statistical parameters were done.
3. It was found that the customers had not answered many questions and therefore the data had many 'Select' values. These were imputed with np.NaN to treat further.
4. The duplicate values were checked and we found none.
5. In the **Data Cleaning** process, the percentage of missing values was checked. The cut off was set at 70% missing values, in an attempt to retain maximum data. The columns which had more missing values than this cut off were dropped.
6. During **EDA**, univariate analysis on all the remaining columns were performed to understand data skewness and imputing missing values. Appropriate bar charts, box plots count plots were plotted. Also, we found some of the columns did not give much information to contribute to analysis. These were dropped.
7. **Data Preparation:**
    a. During this phase, the columns with Yes and No values were converted to binary values.
    b. The dummy variables were created and the data was standardised using the Standard Scaler.
    c. The data was split into two: Train-Test split with the train data with 70% as the train data to train the model.
        i. Two variables X and y were created with X having the data after dropping the target variable and Prospect ID while y was the target variable.
        ii. The X train data was rescaled.
    d. The RFE selection method was used to select the most relevant features
        i. We set a cut of 15 most relevant features for RFE to select.
8. **Modelling:**
    a. Using the RFE supported features, the initial model was devised.
    b. We made three models following the initial one, dropping two features after checking p-values>0.05.
    c. We found the accuracy to be 92% with 13 features.
    d. We checked the VIF scores and dropped 'What is your current occupation_Unemployed" due to high VIF score.
    e. We ran our fifth model and found the p-values low. We checked the VIF scores again and they were within acceptable ranges. We finalised this model with 12 features.
    f. The following results were obtained:

TRAIN DATA SET	
Overall accuracy after building model	0.92
Accuracy after VIF	0.92
Sensitivity	0.85
Specificity	0.96
False Positive Rate	0.038
Positive Predictive Value	0.93
Recall	0.91
Probability Threshold/Optimal Cut off	1.8

After the optimal cut off is chosen, the following results were obtained:

TRAIN DATA SET	
Overall accuracy after building model	0.92
Accuracy after VIF	0.92
Sensitivity	0.97
Specificity	0.57
False Positive Rate	0.42
Positive Predictive Value	0.58
Negatively Predicted Value	0.97

The results obtained on the test data were within the acceptable range of the train data set results.

TEST DATA SET	
Overall Accuracy	0.91
Sensitivity	0.84
Specificity	0.95
Precision	0.9
Recall	0.84
F1 Score	0.87