SUMMARY REPORT AND RECOMMENDATIONS ON LEAD SCORING CASE STUDY

'X Education' is an EdTech company which sells online courses to industry professionals, and from this case study they want us to find the most promising leads who are most likely to convert to paying customers. The case study is performed on the **"Leads"** dataset and we achieved desired results from the following steps:

- ***Data Understanding:*** We started with understanding the data which is very important for any business problem solving. We found the shape, size and info of the data frame, followed by checking the presence of missing values and duplicated rows. Then we proceeded towards the most important stage- **data cleaning.**

- ***Data Cleaning:*** "Leads" data has some discrepancies like missing values, unique values, and a value called 'Select' which is as good as a null value since the customer has not selected any option for that category. We replaced **'Select'** with **'NaN'**, dropped columns which have extremely high null values (35% and above), dropped columns which have a unique value and have no significance in further analysis. For columns having <35% yet very high missing values, we have imputed those with a separate category as **'Unspecified'** in order to avoid a biased model. We have combined low frequency values for some categorical columns to obtain lesser number of dummy variables and dropped some rows having missing values.

- ***Exploratory Data Analysis:*** There is some data imbalance in the target column (**'Converted'**); 38.37% of the total leads have converted while 61.63% have not. We performed Univariate, Bivariate and Multivariate analysis on numerical and categorical columns separately with respect to the target variable. Some outliers are present but it is not significantly high as per distribution of data.

- ***Data Preparation:*** Prepared the data for model building by performing some **pre-processing steps** like conversion of binary variables (Yes/No) to 1/0, creating dummy variables for other categorical columns and dividing the data into 'X' and 'y' data frames. We split the data into 70:30 **Train** and **Test** ratio and used **'Stratified Sampling'** to handle the data imbalance in target column. We rescaled the features using **StandardScaler** so that the units of the coefficients obtained are all on the same scale.

- ***Model Building:*** Created a logistic regression model using **RFE** to get 15 features initially and dropped variables one by one manually on the basis of high P-value (>0.05) and high VIF (>5). Finally, we retained 12 variables in **'X_train'** data.

- ***Model Evaluation:*** We predicted the training data with an initial cut-off of 0.5, then plotted **ROC Curve** and evaluated an **AUC** of **0.89**. Based on trade-off between **'Precision'** and **'Recall'**, we have considered the optimal cut-off point as **0.42** and obtained the following metrics on the **Training data**:

Metric	Value (%)
Accuracy	81.48
Recall	76.29
Precision	75.65
F1-Score	75.97
Sensitivity	76.29
Specificity	84.71

- ***Prediction on Test Data:*** We made predictions on the **Test data** with an optimal cut-off of **0.42** and obtained these metrics:

Metric	Value (%)
Accuracy	81.06
Recall	75.47
Precision	75.25
F1-Score	75.36
Sensitivity	75.47
Specificity	84.54

- ***Summary and Recommendations:*** The following features were found to be useful and it is advisable for 'X Education' to concentrate more on these factors for higher percentage of leads conversion.

    - ***Occupation_Working Professional***
    - ***Lead Origin_Lead Add Form***
    - ***Last Notable Activity_SMS Sent***