Summary Report:

The X Education currently has a lead conversion rate of around 30% and now they want to increase the lead conversion rate by giving more attention to potential leads with high probability of conversion. To achieve this business requirement, we built a logistic regression model that can assign a lead score between 0 to 100. Higher the score, higher the chances of conversion.

The dataset shared has a total of 9240 observations and 37 columns. We first started with the data understanding step where we looked into descriptive statistics and data cleaning. The data majorly contained categorical features and 3 numerical features. There were four columns with more than 45% of missing values and Tags column had around 36% missing values. All these columns were dropped initially, and Tags seemed to be filled by employeesafter analysis. So, we were confident to remove it. Another particularity about the dataset was the less or zero variability in data for 13 features. These features were removed as they won't add any new information to the model. In a similar scenario, city and country columns were also removed.During our analysis we found certain features having more than 60% values as category 'Select' which actually turned out to be missing values. So, we had to drop them as well. Since the dataset had sufficient number of observations (around 6K), we proceeded with dropping null valued rows approach than filling them. As there were only 5 rows with outlier values, they were also deleted. There was no class imbalance in the data.

Visualizations were done for bivariate analysis. Some major findings from this step are as follows. Even though most people from unemployed category came as leads, working professionals had a high conversion rate compared to all other occupation categories. Also, '**Total Time Spent**' clearly had a strong relation to the conversion. After this analysis, the data was divided into train and test set **(7:3 ratio)** for modelling. Before proceeding with modelling, dummy variables were created and Normalization was done for numeric features. Heatmap generated revealed multicollinearity among the dummy variables created. We haven't dropped them there as doing RFE and VIF analysis will anyway get rid of those highly correlated features.

The next step was feature elimination using RFE. After getting most important 15 features, we created a baseline model which gave nearly **78.9% accuracy** on train data. Because p values aren't trusted at this stage, VIF was used to eliminate correlated features. Models were rebuilt in iterations eliminating features with high VIF(>5) and high p value(> 0.05). The final model has 12 most significant features. By plotting accuracy, sensitivity and specificity for various probabilities we arrived at a **final cut off of 0.42** which gave acceptable value of around **79%** for all metrices.

Finally, model evaluation was done using the test data. Predicted probabilities were used to get lead scores. Most metrics values came around **77%** for test data and top features were identified with the model coefficient values.