Problem Statement

X Education, an online education company faces a challenge with **low lead conversion rates** despite attracting numerous potential customers daily through website visits, form submissions, and referrals. The company seeks to enhance efficiency by identifying 'Hot Leads'—leads with higher conversion potential. Currently, **only about 30% of acquired leads are converted to paying customers**

Business Goal

X Education aims to **develop a lead scoring model** that assigns scores to leads based on their likelihood of conversion. This model is intended to aid the sales team in prioritizing communication efforts and focusing on leads that are more likely to convert, **potentially raising the overall lead conversion rate to the CEO's target of around 80%.**

Problem Approach Summary

Step 1: We read and understood the problem and data

Step 2: Data cleaning and preparation

We checked the shape and type of the data and then proceeded further to check any null values. We observed that multiple columns have null values. We decided to drop the columns which have more than 30 % null values (Select is converted as null values). Then we worked on the numerical column to check for outliers. Outlier values have been discarded.

Step 3: EDA

Conducted Univariate analysis and identified significant variables. Columns which has not provided any significant inference were dropped.

Step 4:Creating dummy variables

For categorical variables, dummy variables are created. Yes and No are converted to 1 and Zero and also one hot encoding has been applied for multilevel values.

Step 5: Test and Train Split

Data is split into train and test set with ratio 70% and 30 % .

Step 6:Feature scaling and RFE

We used the Standardscalar to scale the numerical variables.

Using the RFE we went ahead and selected the 20 top important features. We recursively tried looking at the P-values in order to select the most significant values that should be present and dropped the insignificant values. Features with high VIF value(>5%) also dropped

Step 7: Calculate Accuracy Sensitivity and Specificity.

Derived the Confusion Metrics and calculated the overall Accuracy of the model.
Calculated the '**Sensitivity**' and the '**Specificity**' matrices to understand how reliable the model

Step 8: Plotting the ROC curve

We plotted the ROC curve for the features and the curve came out with a good area coverage of 88% .

Step 9: Finding the optimal cut of point

Then we plotted the probability graph for the 'Accuracy', 'Sensitivity', and 'Specificity' for different probability values.. The cut-off point was found to be 0.3

We could also observe the new values of the 'accuracy=80.4%, 'sensitivity=80.8%', and' specificity=80.1%'.

Step 10 : Computing Precision and Recall metrics

Precision and Recall metrics values came out to be 79% and 69.3% respectively and cut off point as 0.4%

Step 11: **Making Predictions on Test Set**

Then we implemented the trained model to the test data and calculated the conversion probability based on the Sensitivity and Specificity metrics and found out the accuracy value to be 79%, Sensitivity=80.3%, and Specificity= 78.3% .