# Lead Scoring Case Study Summary

## Problem Statement:

X Education sells online courses to industry professionals. X Education needs help in selecting the most promising leads, i.e. the leads that are most likely to convert into paying customers.

The company needs a model wherein you a lead score is assigned to each of the leads such that the customers with higher lead score have a higher conversion chance and the customers with lower lead score have a lower conversion chance.

The CEO has given a ballpark of the target lead conversion rate to be around 80%

## Overall Summary:

Step 1: Business Understanding

1. To understand the problem and analysis expected

Step 2: Data Understanding and Inspection

1. Read and analyze the given data set
2. Understood the variable definition from data dictionary

Step 3: Data Cleaning

1. The variables with >= 40% and rows with <=1% of null values were dropped
2. Checked for missing values and Imputed using mean/median/mode
3. We moved with 98.2% original rows available for EDA

Step 4: EDA (Exploratory Data Analysis)

1. Univariate Analysis to understand data spread
2. Bivariate Analysis to understand the variable vs impact on conversation rate
3. During analysis Outlier were handled using soft capping and irrelevant variables were dropped

Step 5: Data Preparation

1. Dummy variables created for categorical variables
2. The train-test split was done in the ratio of 80:20
3. Performed feature scaling using standard scaler for numerical variables
4. Initial Model with statsmodels gave a complete statistical view of all parameters

Step 6: Model Building on Train data set

1. RFE done to attain top 18 variables (~30% of total available features)

2. Validated Multi-Collinearity using VIF after each model creation
3. Variable with high VIF and/or p-value > 0.05 were removed in consecutive step
4. Final model was built with 11 variables with VIF and p-value in accepted range
5. Creating new column 'predicted' with 1 if Converted_Prob > 0.5 else 0
6. Generated confusion matrix with p=0.5 gave
    a. Accuracy – 89%
    b. Sensitivity – 82%
    c. Specificity – 94%

Step 7: Plotting ROC Curve

1. Plotted ROC gave curve area = 0.95 which further solidified our model

Step 8: Finding the optimal cut-off

1. Prob. Curve between accuracy, sensitivity and specificity for values between 0.1 to 0.9 gave 0.3 as the optimal cut-off
2. New matrix with p=0.3 gave accuracy, sensitivity, specificity of 88%

Step 9: Precision and Recall Trade off

1. Tradeoff value was nearly 0.3 and precision = 89%, recall = 88% and F1 Score = 86%

Step 10: Making prediction on test data

1. Prediction on test set gave accuracy, specificity of 88%, precision=83%, recall=85% and f1_score=84%

Step 11: Outcome

1. Top Predictors for potential leads:
    a. Tags
        i. Will revert after reading the email
        ii. Others
    b. Lead Source
        i. Welingak Website
        ii. Reference
        iii. Olark Chat
    c. Last Activity
        i. Others
        ii. SMS Sent
        iii. Olark Chat Conversation
    d. Total Time Spent on Website
    e. Specialization
        i. Travel and Tourism
    f. Do Not Email