# Summary

## Problem Statement:

- X Education sells online courses to industry professionals.
- Although X Education gets a lot of leads, its lead conversion rate is very poor. The typical lead conversion rate at X education is around 30%.
- To make this process more efficient, the company wishes to identify the potential leads, also known as 'Hot Leads'.
- If they successfully identify this set of leads, the lead conversion rate should go up as the sales team will now be focusing more on communicating with the potential leads rather than making calls to everyone.

## Business Objective:

- X Education wants to identify the most promising leads.
- The company wants to build a model to identify the hot leads.
- Improve the lead conversion rate from 30% to 80%

## Solution Methodology

1. Read and understand Data
2. Data Cleaning
    a. Handle Missing values
    b. Handle Skewed Categorical Columns
    c. Handle Outliers
3. EDA
    a. Univariate Analysis of Categorical Variables with respect to Target Variable
    b. Bivariate Analysis on Numerical Columns
4. Data Preparation
    a. Dummy Variable Creation for the categorical variables with more than 2 levels
    b. For performing train-test split, we have chosen 70:30 ratio.
    c. After cleaning and dummy variable creation, total number of rows and columns are 8953 and 59 respectively.
    d. Feature Scaling
5. Modelling
    a. Used Logistic Regression Technique
    b. First used RFE method to come up with 15 important variables
    c. Then used manual method for dropping variables to make the model significant by looking at p-value and VIF. We tried to keep the variables which have p-value less than 0.05 and VIF less than 5.
    d. In our final model, all VIFs are low and p-values are below 0.05.
    e. Model Evaluation is done using accuracy, sensitivity and specificity.
    f. Used ROC to check the trade off between True Positive Rate and False Positive Rate. Area under ROC is 0.87, which is a very good value for a model.
    g. From the curve above, 0.35 is the optimum point to take it as a cutoff probability.

    h. Found the Optimal Cut-off point value as 0.35 to improve sensitivity
6. Model Validation and Analysis of Different Metrics
    a. Training Set:
    b. Accuracy:80.53%
    c. Sensitivity:81.38%
    d. Specificity:80.01%
    e. Test Set:
    f. Accuracy:80.15%
    g. Sensitivity:81.09%
    h. Specificity:79.59%
7. The model seems to predict the Lead Conversion Rate very well and we should be able to give the CEO confidence in making good calls based on this model.

8. Conclusions and Recommendation

    The variables that mattered most in the lead conversion are -

    - Lead Origin_Lead Add Form
    - What is your current occupation_Working Professional
    - Total Time Spent on Website
    - Lead Source_Organic Search
    - Lead Source_Direct Traffic
    - Last Activity_Olark Chat Conversation
    - Last Notable Activity_Email Opened
    - Do Not Email
    - Last Notable Activity_Page Visited on Website
    - Last Notable Activity_Email Link Clicked
    - Last Notable Activity_Modified
    - Last Notable Activity_Olark Chat Conversation

    X Education must focus on the people if –

    - The lead origin is through Lead Add Form
    - The prospect is a working professional
    - The prospect spends a lot of time on the website

As these parameters have positive coefficients, this will improve the lead score and in turn, these people will be ‘hot leads’ for X Education