# 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, in particular, has given a ballpark of the target lead conversion rate to be around 80%.

## Approach:

Below are the steps followed to solve this problem

1. **Data Reading, Understanding and Exploring:**
   Here we observed following things -
   - Number of rows and columns
   - Statistics and spread of the data
   - Missing and null values
   - Duplicate values
   - Data types of the columns

2. **Data Cleaning and Preprocessing:**

- Renaming columns to shorter name
- Treating missing values based on 3 criteria –
    - Remove columns if %(data) missing in the respective column > **approx. (30%)**
    - Remove rows if %(data) missing the respective column < **5%**
    - Impute the missing value if %(data) missing **>5% and approx. (<30%)**
- Deleting unnecessary columns
- Data Conversion
    - Changed the binary variables into '0' and '1'
- Sanity Check
- Fixing incorrect data types and index

3. **Exploratory Data Analysis and Visualization:**

- Univariate Analysis:
    - Visualized categorical variables using bar chart
    - Visualized numerical variables using bar chart
    - Outlier analysis and treatment using box plot
- Bivariate Analysis:
    - Visualized numerical and categorical variables using bar chart and clustermap

4. **Model Building:**
    - Created dummy variables for categorical variables
    - Removed all the repeated and redundant variables
    - Split the data into Train and Test Sets : 70:30 ratio
    - The original numerical variables were scaled using the Standard Scaling
    - Checked correlations among variables using heatmap and dropped the highly correlated columns
    - We selected the 20 most important features using the Recursive Feature Elimination method
    - We used manual feature elimination to build model by eliminating columns having p-value > 0.05 and VIF > 5.5
    - We created 6 models to arrive at the final model log_reg_v6. The p-value and VIF of all the features in the final model were also found to be good.
    - We are then left with 15 variables and 1 constant variable which has been able to predict the training data set at 80% accuracy
    - The same model has been applied to test data set after scaling the test data. We have observed accuracy of 80% and sensitivity of 81% of there as well.

**Recommendations**:

- The user experience can be enhanced and conversion rates can be raised by looking at the behaviour of customers who spend more time on the website. The company should put its attention on developing interesting content and easy-to-use navigation to encourage customers to spend more time on the website.
- As the website's ranking does not appear at the top of the search results, SEO optimization should be done to attract more students/prospects to the website.
- Company should focus on marketing using social media and search engines.
- Maximum identified prospects have not come through recommendations. Company should provide rewards/discounts to candidates for providing reference that convert to lead, it will encourage the student/candidate to provide more references.
- Prospects who opted for receiving Email and Phone calls were majorly not converted. There is potential for these to be improved.