# SUMMARY

- **PROBLEM STATEMENT:**
  X is an education company which is selling their online courses to the industrial professionals. The CEO of the company want target leads conversion to be 80% for which they are asking to predict higher lead score which have a higher conversion chance and the customers with a lower lead score have a lower conversion chance.

- **STEPS INVOLVED:**
  After clearly understanding we need to build a logistic regression model to assign the lead score to each of the leads for the potential leads which are achieved by below steps:

  1. **IMPORTING AND INSPECTING DATASET:**

     1. Imported **required libraries** and **Leads.csv** file using pandas library and converted it into **leadf** data frame.
     2. Inspected leadf dataset using **shape, info(), and describe()** for structural understanding.
     3. Checked duplicate value presence inside the **ProspectID and Lead Number**
     4. **Preserved the Lead Number** into another variable Lead_Number for assigning lead score further.

  2. **DATA CLEANING:**
     As stated, select values are to be consider as null values so assigned them as missing values so replaced them with NaN values using **np.nan**.
     1. Dropped around **7 variables** having **more than 40% null values.**
     2. Used **nunique()** to remove all the **variables having 1 value_counts** in them and checked the uniqueness of the **variables having 2 value_counts** and dropped them also if required.
     3. By imputation with higher values or other keywords replaced the null values left

     After performing these imputations and dropping we were left with **13 variables.**

  3. **EDA AND DATA PREPARATION:**
     1. **Checked Imbalance of the dataset to be 38%** according to its Target variable.
     2. **Performed univariate analysis** on numerical and categorical variables and **bivariate analysis**

    **for numerical variables** in regards to Target variable 'Converted'
3. **Performed Outlier Analysis** and treatment for numerical variables.
4. **Dummy Variable creation** for categorical variables and **conversion of some binary variable** to 0/1

After dummy variable creation we were having **58 variables** for model building.

4. **<u>MODEL BUILDING:</u>**
    1. Performed **train-test split at 70% and 30%** resepectively.
    2. **StandardScaler()** for numerical variables feature scaling
    3. **Performed RFE for feature selection** to attain top **20 features.**
    4. **MODEL 4** was the ideal one with VIF<5 and all p-values <0.05.
5. **<u>MODEL EVALUATION:</u>**
    1. After performed Prediction on train model we got **ROC Curve which gave threshold value of 0.96, and also got optimal cut-off point to be 0.3 metric calculated**
        - **Accuracy – 89.54%**
        - **Sensitivity – 89.21%**
        - **Specificity – 89.75%**
        - **Precision-84.29%**
        - **Recall – 89.21%**
    2. After Performed Prediction on test model calculated metrics beyond accuracies:
        - **Accuracy - 89.61%**
        - **Sensitivity – 89.68%**
        - **Specificity – 89.56%**
        - **Precision – 84.87%**
        - **Recall – 89.68%**

6. **<u>CONCLUSION:</u>**
    1. **Calculated Lead Score** and assigned it with Lead number to leadscore dataset having all the features on which our ideal model was built.
    2. Also provided the dataset of HotLeads which are having leadscore more than 80%.

- **RECOMMENDATIONS:**

  Important features responsible for good conversion rate are :

  a. Tags_closed by horizzon
  b. Tags_willrevert after reading the email
  c. Lead Source_welingakwebsite