# Summary

## Problem statement

An education company named X Education sells online courses to industry professionals. The company wishes to find the "Hot leads".

X Education focuses on to select the most promising leads, the leads that are most likely to convert into paying customers. The company requires you to build a model wherein you need to assign a lead score to each of the leads such that the customers with a higher lead score have a higher conversion chance and the customers with a 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%.

## Solution Summary

### Understanding the Data

Reading and understanding the data. And analyzing their data description and information.

### Data cleaning

➢ Columns with greater than 35% of null values has been dropped, which includes Imputing the missing values as and where required.
➢ The outliers were identified and removed.
➢ Numerical categorical data were imputed with mode and columns with only one unique response from customer were dropped.

### Data Analysis

➢ Data imbalance checked.
➢ Performed univariate and bivariate analysis for categorical and numerical variables. Provided the valuable insight on effect on target variable.
➢ Time spend on website shows positive impact on lead conversion.

## Data Preparation

- ➤ Created dummy features (one-hot encoded) for categorical variables.
- ➤ Splitting Train & Test Sets.
- ➤ Feature Scaling using Standardization.
- ➤ Dropped few columns, they were highly correlated with each other.

## Model Building

- ➤ Used RFE to reduce variables from 22 to 15. This will make data frame more manageable.
- ➤ Manual Feature Reduction process was used to build models by dropping variables with p – value.
- ➤ Total 4 models were built before reaching final Model 5 which was stable with (p-values < 0.05). No sign of multicollinearity with VIF < 5.
- ➤ Logm5 was selected as final model with 11 variables; we used it for making prediction on train and test set.

## Model Evaluation

- ➤ Confusion matrix was made and cut off point of 0.3 was selected based on accuracy, sensitivity and specificity plot.
- ➤ This cut off gave accuracy, specificity and precision all around 80%. Whereas precision recall view gave less performance metrics around 75%.
- ➤ As to solve business problem CEO asked to boost conversion rate to 80%, but metrics dropped when we took precision-recall view.
- ➤ So, we will choose sensitivity-specificity view for our optimal cut-off for final predictions.
- ➤ Lead score was assigned to train data using 0.345 as cut off.

## Making predictions on test data

- ➤ Making Predictions on Test: Scaling and predicting using final model.
- ➤ Evaluation metrics for train & test are very close to around 80%.
- ➤ Lead score was assigned.
- ➤ Top 3 features are:
    - ▪ Lead Source_Welingak Website
    - ▪ Lead Source_Reference
    - ▪ Current_occupation_Working Professional

.