# SUMMARY

## OBJECTIVE

Building a machine learning model for an Education company X, wherein a lead score is assigned to each of the leads such that the company can target potential leads, whereby the customers with higher lead score have a higher conversion chance and the customers with lower lead score have a lower conversion chance. Additionally, around 80% of the lead conversion rate is targeted by the CEO.

## STEPS

1. **Understanding the dataset**
   The dataset and the data dictionary were provided, whereby the data contained *9240* samples and *37* features.

2. **Data Cleaning**
   - **Handling missing values**
     The dataset contained some missing values, which were either in the form of a null value, or a text like *SELECT* which had to be converted to null value. Firstly, all features with more than 40% missing values were dropped. Thereafter the features with high data imbalance were dropped. For all the remaining features, the missing values were either imputed with the strategies like mean, median and mode, or the rows were dropped. For the column '*What is your current occupation*', a new category called 'Unspecified' was created.
   - **Handling outliers**
     Visualization tools like boxplot were used to identify the outliers, and these features were capped to some threshold values.

3. **EDA**
   Using the visualization tools, the univariate and bivariate analysis was done on the data to draw useful insights like
   - Data imbalance of 60-40 in the target variable.
   - Some correlation between some features like '*Total time spent of website*' and the target variable.
   - Some categorical features had large data imbalance, therefore were dropped.
   - '*Lead Source*' had a single category '*Google*' with another version '*google*'. Hence, they were clubbed.

4. **Data preprocessing**
   - **Creating Dummy variables**
     The categorical features were encoded by creating the dummy variables. Afterwards, for each column, the dummy variables which had the least amount of non-zero entries were dropped.
   - **Train-test Split**
     The features and the target were separated and split into 70% train data and 30% test data. Since the target variable had imbalance in distribution, henceforth, the splitting was done according to the stratified sampling of the target variable.

- **Feature Scaling**
  Standard scaling technique was used to scale the features.

5. **Model building**
   - RFE was used to reduce the number of features to 15.
   - Afterwards the model was repeatedly trained using the logistic regression in Statmodel api and thereby the features were manually eliminated based on the p-value and the VIF values, ultimately getting to a final model with 12 features.
   - An optimal cut-off threshold of *0.33* was found by plotting the curves for Sensitivity, Specificity and Accuracy against different threshold values.

6. **Model Evaluation**
   The final model was evaluated on the test set and the following results were obtained.
   **Accuracy: 81.58%**
   **Sensitivity: 80.36%**
   **Specificity: 82.34%**
   **Precision: 0.74**
   **Recall: 0.80**
   **F1-score: 0.77**

   The results are in accordance with the business requirement.

7. **Lead score assignment**
   Lead scores were assigned using the output probabilities obtained from the model.

$$Lead\ Score = Ouptut\ Probability * 100$$

**KEY LEARNINGS**

- Interpreting the business problems.
- Using ML models like Logistic Regression to provide the solutions.
- Dealing with the given data to make it interpretable for modelling.
- Provide insights to identify the drivers.

The top three driving features for this assignment are:

- Indicator variables
  - Last Notable Activity 'Had a Phone Conversation'
  - Lead Origin 'Lead Add Form'
  - Lead Source 'Welingak Website'