## Business Requirement

An education company sells online courses. When people fill the online forms, they are classified as Leads. Once these leads are acquired, employees from the sales team start making calls, writing emails, etc. Through this process, some of the leads get converted while most do not. The typical lead conversion rate is around 30%. To increase the conversion rate, the education institute wishes to identify the most potential leads. Once identified, the sales team could focus on communicating with the potential leads rather than making calls to everyone. Eventually, this would increase the lead conversion rate.

## Problem Statement

- We need to build a logistic regression model to identify the potential leads. Each lead will be assigned a lead score.
- Customers with high lead score will have high conversion chance and vice-versa.
- Based on the lead score the sales team can make call to the customers (high lead score) and can increase the conversion rate.

## Analysis and Approach

Read the csv file and perform all data cleaning process. Perform EDA – Univariate and Bivariate analysis. Treat the outliers using IQR. We need to maintain at least 70 – 80% of the data to proceed further. Convert all the categorical variables that have the value ‘Yes’ and ‘No’ to ‘1’ and ‘0’ and create dummy variables. Check the data imbalance in the target variable (**Converted** in our case). Remove the variables that are highly correlated which can be identified using heatmaps.

## Model Building (Logistic Regression)

Using RFE (Recursive Feature Elimination), choose at least 15 variables. Build the logistic regression model using Generalised Linear Models with the variables selected from RFE. Based on the model statistical summary (p-value) and VIF, we can drop the features based on the below conditions.

- The p-values < 0.0.5 for all the features and VIF < 5 have to be achieved.
- Using the model, predict the y-values which represent the probability of occurrence.
- Define a threshold probability and identify whether the particular customer has been converted or not.
- Use **ROC** curve to find the optimal threshold value of probability
    - Plot **Accuracy**, **Specificity** and **Sensitivity** (from ROC) and the point where they meet is the optimal threshold.
- Calculate the converted value using the threshold probability from the above plot.
- Calculate the accuracy, specificity, sensitivity and precision.
    - Accuracy, Specificity and Sensitivity should be **around 80%.**

## Model Evaluation – Test data

- Evaluate the model created using the test data.
- Predict the probability (**y_pred**)
- Use the same threshold used for train dataset and predict the converted value for the test dataset.
- Calculate the accuracy, sensitivity and specificity and these values should more or less same as our train dataset.

## Calculating Lead Score

- *Lead Score = Converted Probability * 100*
- High lead score is considered to have higher conversion rate and vice-versa.

## Business Recommendations

The sales team of the education company has to focus on the customers who have high lead score. By doing so, all potential leads could have been identified and the lead conversion rate could be increased.