# Summary

**Problem statement:** An education company named X Education sells online courses to industry professionals. X Education gets a lot of leads, its lead conversion rate is very poor. For example, if, say, they acquire 100 leads in a day, only about 30 of them are converted. To make this process more efficient, the company wishes to identify the most potential leads, also known as 'Hot Leads'. If they successfully identify this set of leads, the lead conversion rate should go up as the sales team will now be focusing more on communicating with the potential leads rather than making calls to everyone.

This analysis is done for X Education and build a logistic regression model to find ways to get more industry professionals to join their courses. The provided data contains a lot of information about how the potential customers visit the web site, the time they spend there, how they reached the web site and the conversion rate.

The followings are the steps to solve this case study:

1) **Data loading and cleaning:** In this step firstly, I have imported all the required libraries and load the data. Then I have inspected the data and find out the null values present in each feature. Secondly, I have dropped the features having more than 45% of null values.
2) **EDA(Exploratory Data Analysis):** In this step I have implemented univariate analysis, segmented univariate analysis on both categorical and numerical features. From that analysis I found some features have imbalance data so I dropped them. The numeric values seem good and no outliers were found.
3) **Create Dummy Variables:** In this step I have created dummy variables for the categorical features. Converted binary variables like Yes/No to 1/0.
4) **Train-Test Split:** In this step I have split the data into train (70%) and test (30%) data respectively.
5) **Scaling:** To scale the numerical features I used the MinMaxScaler.
6) **Model Building:** in this step a logistic regression model was build and RFE was used to select the top 15 relevant features. Later the rest of the variables were removed manually depending on the VIF values and p-value (The variables with VIF < 5 and p-value < 0.05 were kept).
7) **Model Evaluation:** In this step a confusion matrix was made. Then plotted a ROC curve to find out the optimal cutoff point and calculate the accuracy, sensitivity, and specificity.
8) **Prediction on test data:** Prediction was done on the test data frame and with an optimum cut off as 0.35. Then calculate accuracy, sensitivity and specificity again.
9) **Calculate Precision-Recall:** In this step the precision and recall were calculated. We got precision around 81% and recall around 84%.

## Conclusion:

It was found that the variables that mattered the most in the potential buyers are,

- Total time spent on website
- Total number of visits
- Last activity was SMS sent and Olark chat conversation
- Lead source was Google, Direct traffic, Organic search
- Lead origin is lead add form
- Current occupation is working professional