# Summary

This analysis was done on education domain and finding a way to get more people to join their courses. The basic data provided gave us a lot of information about how the

potential customers visit the site, the time they spend there, how they reached the site and the conversion rate.

1. Sanity check:

   In this dataset they gave partially good data but, in the dataset have some null values and unacceptable values. So, we worked on those values like some values removed and some replaced with mean, mode, median and some replaced with others, etc. few of the null values were changed to ‘not provided’ so as to not lose much data.

2. EDA:

   A quick EDA was done to check the condition of our data. It was found that a lot of elements in the categorical variables were irrelevant. The numeric values seems good and no outliers were found.

3. Creation of dummy variable:

   The dummy variables were created and later on we removed ‘Lead Number’ elements. For numeric values we used the MinMaxScaler.

### 4. Train-Test split:

The split was done at 70% and 30% for train and test data respectively.

### 5. Model Building:

Firstly, RFE was done to attain the top 15 relevant variables. 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 were kept).

### 6. Model Evaluation:

We created a confusion matrix to define the performance of a classification algorithm. Later on, the optimum cut off value (using ROC curve) was used to find the accuracy, sensitivity and specificity.

### 7. Prediction:

Prediction was done on the test data frame and with an optimum cut off as 0.28 with accuracy, sensitivity and specificity of 80%.

### 8. Precision – Recall:

This method was also used to recheck and cut-off of 0.32