# SUMMARY

This case study is for X Education company to help them select the most promising leads, i.e., the leads that are most likely to convert into paying customers.
The company requires a model wherein lead score needs to be assigned to each of the leads such that the customers with higher lead score have a higher conversion chance and the customers with lower lead score have a lower conversion chance.

We were given a dataset having 37 fields which helped us in building relevant model for the company.

## Steps Followed:

### 1.Data Understanding :

Here we went through the given dataset. Verified the categorical and continuous variables. Checked the statistics of continuous variables.

### 2.Data Cleaning:

We did basic cleaning of dataset. Dropped the features having more than 30% missing values.

### 3.Data Preparation:

Here we took care of missing values, outliers. Created dummy variables from the categorical variable.

### 4.Exploratory Data Analysis:

We did univariate, bivariate and multivariate analysis here. Plotted few important features with target convert variable.
Performed Train Test split on the dataset with 70% trained data and 30% test data.
Checked the correlation between each variables using heatmap.

### 5.Model Building:

We tried to evaluate important features for our model using RFE. After that we created stats model and validated the p value. Made sure the p value not to exceed 0.05 and VIF score to fall below 5.VIF score determined the correlation between each independent variables. Examined this by dropping single feature at a time based on p value and VIF value. Eventually stopped when both the conditions satisfied($p<0.05$ and $VIF<5$).

### 6.Model Evaluation:

We determined the threshold based on ROC curve. Evaluated respective accuracy, sensitivity, specificity, precision and recall. Made sure the model to around 80% accurate.

### 7.Predictions on the Test Dataset:

Based on determined threshold from trained model we predicted the test data set which came out to be pretty accurate with 78.45% accuracy.
Sensitivity came to be 77.94%, Specificity was 78.91%.

## Conclusion:

- 'TotalVisits' , 'Total Time Spent on Website' , 'Page Views Per Visit' features contribute most towards the probability of a lead getting converted. So, company needs to focus more on this.
- The model is able to capture correct leads having high probability of conversions also leads having low probability.
- The accuracy of the model is close to 80% with high sensitivity and specificity which will help the company reaching target customers.