# Summary Report

X Education - an education company sells online courses to industry professionals. The company needs to identify the most potential Leads from their online marketing channels. The goal of the case study is to identify the hot leads which have the maximum chances to take a course from X education.

**Step 1: Inspecting the Data**

A lot of information regarding the leads on their source, no. of visits and time spent on the company's website, occupational backgrounds, professional interests, city etc are provided in the data set.

**Step 2: Data Cleaning & Data Preparation**

Many columns have 'SELECT' as the values and replacing them with Nan values.

Dropping all the columns which has high % of Null values and dropping all the highly skewed categorical columns.

For numerical columns, we use MEAN/MEDIAN to replace the null values for the columns which have lesser % of missing values and we decide to keep it for the analysis.

For Categorical columns, we use MODE to replace the null values for the columns which have lesser % of missing values. Also, if no proper category is visible through mode, use "other" as a new category.

We have converted binary variables with values as Yes or No to 0's & 1's.

We are removing Tags field as it will not add much value to business(Just a status field).

**Step 3: EDA**

We have used countplots extensively throughout to explore the data distribution for various columns.

Visualizations on the categorical columns states how well the data is distributed and how can be the data get grouped for e.g. in case of the Lead Source, we grouped google & bing as 'Search Engines', facebook & Youtube as 'Social media' and many less used data points as 'others', this help in building a good model. It has also helped in identifying the skewed columns as in Country where India is most common occurrence >96% and this highlights the insights on the data and enables us to handle such scenarios.

For numerical columns, visualizations showed mostly the data is good and have helped in showcasing outliers in few cases. Eg: EDA on continuous variables has helped in removal of outliers in TotalVisits, PageViewPerVisit columns.

**Step 4: Creation of Dummy variables, Test-Train Split & Feature Scaling**

Dummy variables were created for Categorical variables, e.g LeadOrigin, Country, etc.

The Train-Test split was done as 70% for Train data and 30% for Test data.

StandardScaler technique is used for feature scaling.

### Step 5: Model Building

We use Recursive Feature Elimination (RFE) method to get the top 15 variables.

Then, we run the Generalized Linear Model Regression Results model multiple times and eliminate the variables manually depending on VIF value(<5) & p-value(<0.05) in each iteration.

### Step 6: Model Evaluation

A Confusion Matrix was created to find the sensitivity, specificity, false positive rate, positive & negative predictive value. We plotted the ROC curve and found the ROC value as 0.88 which states that our model has good prediction capacity.

### Step 7: Prediction

After plotting the Accuracy, sensitivity & specificity we found the optimal cut-off point for probability as 0.33.

Train Data:

Accuracy: 80.34%
Sensitivity: 84.31%
Specificity: 77.89%

Test Data:

Accuracy: 79.76% (~80%)
Sensitivity: 84.07%
Specificity: 76.99%

Sensitivity value of both Train and Test data are greater than 80%

Difference between Train and Test data in terms of Accuracy, Sensitivity & Specificity are less than 5%.

### Step 8: Precision & Recall

Recall expresses the ability to find all relevant instances in a dataset, precision expresses the proportion of the data points our model says was relevant were relevant. We calculated them for both train & test datasets.

Train Data:

Precision: 70.17%
Recall: 84.31%

Test Data:

Precision: 70.13%
Recall: 84.07 %

### Step 9: Assign Lead score based on converted Probability

We have converted probability in our model. By using the same, we have derived the lead score for both Train and Test dataset, and they behave the same way. They reflect the same pattern in the plotted visualizations.

### Report Conclusion:

**we have achieved more than the X education's CEO's expectations of 80% conversion rate.**