# Lead Scoring – Summary Report

**Problem Statement:** An education company named X Education sells online courses to industry professionals. Leads are potentially generated through multiple sources as provided in the dataset. The typical lead conversion rate at X education is around 30%. The company is looking to make its process more efficient, whereby it can identify and focus on the most potential leads, with a target to achieve the Lead Conversion Rate to 80%

**Expected Outcome:** Develop a Logistic Regression Model to assign a lead score between 0 – 100 for each of the leads which can be used by the Company to target potential leads. Identify the key factors that drive convertibility

**Approach:**

First, the data was imported, which was present in a csv file having 9240 rows and 37 columns. After that **data preparation** was done, which involved the following steps:

1. Missing value imputation
   a. Handling Null values and 'Select' Values
   b. Handling values which had a very low count – Combining such smaller values into a single value
   c. Finding Relationship between columns
   d. Dropping the columns with a high percentage of missing values
   e. Dropping the columns which had either a single value OR binary values with <10 occurrences of the other value
   f. Converting the Remaining binary columns (From Yes/No to 1/0)

2. Outlier treatment
3. Dummy variable creation for categorical variables
4. Test-train split of the data : 70% Train and 30% Test
5. Standardisation of the scales of continuous variables - The variables are scaled in such a way that their mean is zero and standard deviation is one.

After all of this was done, a logistic regression model was built in Python using the function **GLM()** under statsmodel library. This model contained all the variables, some of which had insignificant coefficients. Hence, some of these variables were removed first based on an automated approach, i.e. RFE and then a manual approach based on the VIFs and p-values.

While doing VIF Analysis, discard variables with high VIFs (one by one). In our case, after discarding one variable, the VIFs obtained for the remaining were within acceptable range.

After this, the Confusion Matrix was created, and the Accuracy, Sensitivity and Specificity were determined.

Next, the ROC Curve and the Accuracy, Sensitivity, Specificity curve were plotted. From this, we concluded that the optimal cut-off for the model was around 0.4 and we chose this value to be our threshold and got decent values of all the three metrics – Accuracy (~81%), Sensitivity (~78%), and Specificity (~83%).

We then plotted a trade-off curve between precision and recall.

Finally, the Lead Score was determined for all the Leads and a recommendation was provided to classify Lead Score >= 56 as Hot Leads as they would result in a conversion rate of 80%