## Summary Of the Assignment

The given problem Statement was :

X Education sells online courses to industry professionals. X Education needs help in selecting the most promising leads, i.e. the leads that are most likely to convert into paying customers. The company needs a model where a lead score is 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.

The CEO needs of the target lead conversion rate to be around 80%

Approach followed to solve this problem :-

1)**Cleaning and Handling Data**

1) We have taken the dataset provided by the company which had about 9240 past users records and 37 columns.
2) We had dropped some of the sales team provided columns because they will be generated after a lead is converted. So these are of no use to us. So those were discarded
3) We replaced all the Select values in columns with NaNs. Also missing values treatment is done for all the categorical columns which has more than 2 unique values.
4) We have performed outlier treatments for numerical columns by removing 1% both bottom and top Outliers
5) We have also plotted the columns to visualise the conversion data.
6) Created dummy variables for categorical data and dropped the columns which are not necessary.

2) **Model Building**

1) We have divided the dataset into Train and Test data with 70-30 ratio.

2) Performed Scaling on Numerical Columns

3) Built the model using Logistic Regression.

4) Selected top 20 columns from RFE method and dropped some of the columns which has p value more than 0.05 and VIF more than 5 to solve multicollinearity issues.

5) The columns left are used to get confusion metrics and thus to calculate Accuracy, Sensitivity, Specificity.

6) Also visualised this to get optimal cut-off point.

7) Plotted ROC curve and got area under curve as about 83%

8) Calculated precision and recall scores.

3**) Model Building for Test Dataset**

Same thing is performed on test dataset as well

**Conclusion:**

We have got following numbers for Train dataset

- Sensitivity – 77.47%
- Specificity – 80.40%
- Accuracy – 79.47%

Test Dataset :-

- Sensitivity – 77.66%
- Specificity – 80.04%
- Accuracy - 79.57%

The conversion rate of this model is about 79% which is very much close to the requested by the CEO.