**Lead score case study summary**

**Problem Statement :-**

X Education sells online courses to industry professionals . X education needs to help in selecting promising lead i.e leadsconvert into paying customers.

The company needs a model wherein a lead score is assigned each of leads that the customers with high score lead have higher conversion and the customers with low score lead have low conversion chance.

The CEO, in particular, has given a ballpark of the target lead conversion rate to be around 80%.

**Solution Summary:-**

**Step 1:- Reading data and understating data**

We read data and check its data types its information of data and also check how many null value in dataset in which columns and rows .

**Step 2:-** **Data Cleaning**

In data cleaning step we will drop columns which columns have null values is greater than 70 and impute missing value and median values of numerical variables and creation classification variables in place of categorial variable .

**Step 3:- Exploratory Data Analysis**

In Exploratory Data Analysis data set feel how the data is oriented. There are three variables are identified to have only one variable in all rows and these rows were dropped.

**Step 4:- Data preparation / Dummy creation**

We will create dummy variables of categorical variables in binary form(‘yes’:1,’No’:0) and after we drop the categorical variable from original data .

**Step 5:- Train-Test Split**

In this step we split of data set in proportion of 70:30 values in random state 100.

**Step 6:- Feature Scaling**

In the feature scaling uses standard scaling for Continuous variables and in this process of normalizing the range of features in a dataset .

**Step 7:- Model Building**

In the Model building process we learn and generalize of training data then applying that acquired knowledge to new data .

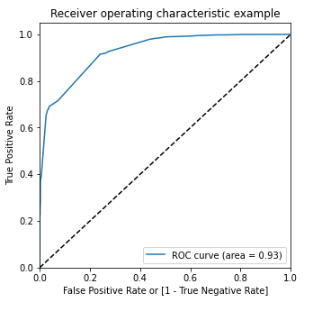
**Step 8:- Recursive Feature Elimination**

In the Recursive Feature Elimination we selected 20 important features to fit in the model and uses statistics . we remove the weakest features until the specified number of features is reached. We arrived 12 features and VIF of these features are also good and creating a dataframe of with values of actual and the predicated probabilities .

We derived confusion matrix of actual data and predicted data and we will check Accuracy, specificity, sensitivity, and recall and precision also.

**Step 9:- Ploting Roc curve**

ROC curve for the feature and the curve came out with pretty decent Area coverage of 93%.



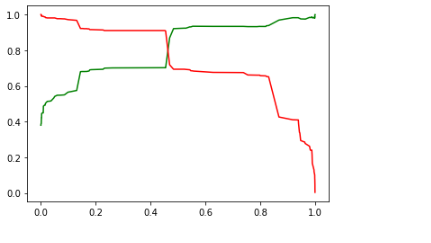
**Step 10 :- Finding optimum threshold**

We have plotted the probability graph of the ‘Accuracy’, ’Sensitivity’, ‘Specificity’ for different probability values. The intersecting point of graph is called cutoff point and the point is 0.42.

Also calculated the lead score and figured that the final predicted variables approximately gave a target lead prediction is 80%

**Step 11 :- Making prediction on test set**

To the test model and calculated the conversion probability based on sensitivity and specificity metrics and found out the Accuracy value is 81% , sensitivity value is 91%, specificity value is 76.4% , Precision value is 70.2%, Recall value is 91.6% .

****