# Lead Score Assignment Summary

## Problem Description

An education company named X Education sells online courses to industry professionals. On any given day, many professionals who are interested in the courses land on their website and browse for courses. The company markets its courses on several websites and search engines like Google. Once these people land on the website, they might browse the courses or fill up a form for the course or watch some videos. When these people fill up a form providing their email address or phone number, they are classified to be a lead. Moreover, the company also gets leads through past referrals. Once these leads are acquired, employees from the sales team start making calls, writing emails, etc. Through this process, some of the leads get converted while most do not. The typical lead conversion rate at X education is around 30%. X Education needs help to select the most promising leads, i.e. the leads that are most likely to convert into paying customers. A model is required to be built wherein 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, in particular, has given a ballpark of the target lead conversion rate to be around 80%.

## Steps Followed:

### Data Reading and Understanding:

Here we get the look and feel of the data including:

* + Number of rows and columns
  + Data types of each columns
  + Checking first few rows how data looks
  + Checking how the data is spread.
  + Checking for duplicates, if any.

### **EDA**:

* Check for Null values.
* Treat Null Values by replacing them using appropriate techniques.
* Outlier detection in columns such as ‘Total\_Visits’
* Outlier Treatment using IQR method.

### **Train**-Test Split

* The split was done at 70% and 30% for train and test data respectively.

### **Feature Scaling**

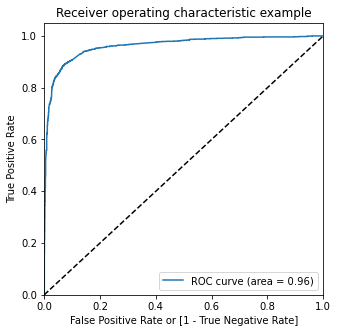
* Columns which have more than two levels were converted to dummies using pd.get\_dummies function.
* We will do min-max scaling on the variables ['TotalVisits', 'Page Views Per Visit', 'Total Time Spent on Website']
* Columns which have only two levels “Yes” and “No” were converted to numerical using binary mapping.
* RFE was used for feature selection to attain top 15 variables

### Model Building

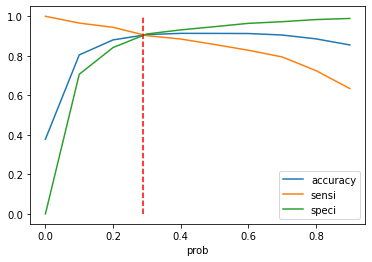
* Used to statsmodel library to create a logistic regression model
* A confusion matrix was created, and overall accuracy was checked which came out to be 91.3%.

### Model Evaluation

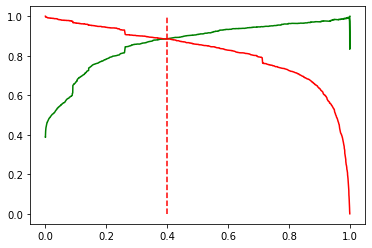
* Plotted the ROC Curve



* To make predictions on the train dataset, optimum cutoff of 0.29 was found from the intersection of sensitivity, specificity and accuracy as shown in below



* To make predictions on the test dataset, optimum cutoff was considered as obtained from Precision recall graph of the train dataset as shown below figure:



We can observe that 0.4 is the tradeoff between Precision and Recall. Thus, we can safely choose to consider any Prospect Lead with Conversion Probability higher than 40% to be a hot Lead.

### Predictions

* After running the model on the Train Dataset these are the figures we obtain:

Accuracy: 91.3%

Sensitivity: 85.70%

Specificity: 94.7%

* After running the model on the Test Dataset these are the figures we obtain:

Accuracy: 88.1%

Sensitivity: 92.76%

Specificity: 85.4%

### Conclusion

Top variables contributing to conversion:

* 1. Lead Source:
  + Total Visits
  + Total Time Spent on Website
  1. Lead Origin:
  + Lead Add Form
  1. Lead source:
  + Direct traffic
  + Google
  + Welingak website
  + Organic search
  + Referral Site
  1. Last Activity:
  + Do Not Email\_Yes
  + Last Activity\_Email Bounced
  + Olark chat conversation