1. Problem Statement:

An education company named X Education sells online courses to industry professionals. The company markets its courses on several platforms. Through this process, some of the leads get converted while most do not. The typical lead conversion rate at X education is around 30%. The company wants to increase its conversation rate by only following up with most probable leads.

2. Solution Approach:

2.1 Data Cleaning and Preparation:

'Select' value is considered as a placeholder in many categorical columns, thus these values are imputed to null values.

All the Columns with more than 30% null values are dropped as they do not give much meaning to the predictive model.

With columns which have less than 30% null values, calculate the dummy variables (one hot encoding) and drop the unknown value column.

Calculate dummy variables for the other categorical variables and drop the least significant one based on its occurrence.

Convert Yes/No to 1/0 for categorical values.

This is done so that only numerical features are fed to the logistic regression model.

2.2 Training and Test Split:

Train and Test Data is Split at 70:30 ratio.

2.3 Feature Scaling:

Numerical values are scaled to give optimal results.

2.4 Feature Selection with RFE

With RFE, get the initial dataset consisting of all the important features to be considered for the logistic regression model.

RFE ranks all the features with the **rfe.ranking_** parameter and computes whether it should be included or not with **rfe.support_.**

2.5 Manual Feature Elimination

The following steps are followed in sequence:

1. Build the Model

2. Inspect RFE's for MultiColinearity

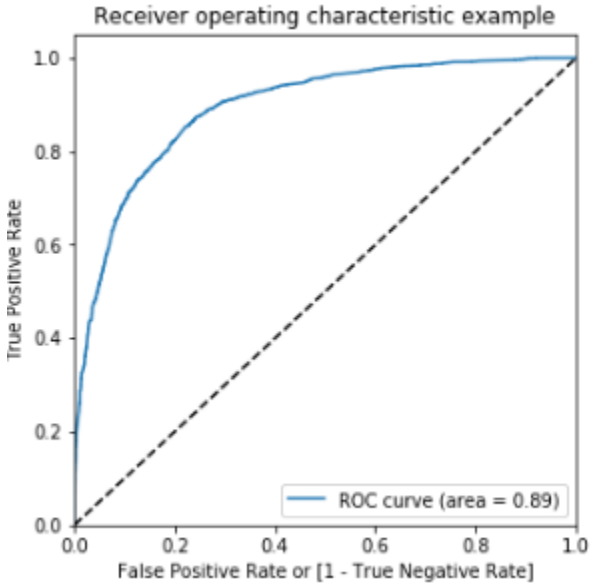
3. Drop variables with high VIF (VIF should be < 5 so that there is no co-relation between features)

4. Drop variables with high p-value (p-value should be lesser than 0.05 to be significant)

5. Update the model

6. Calculate the accuracy as the accuracy should not drop way too much

2.6 Check the accuracy with ROC Curve:

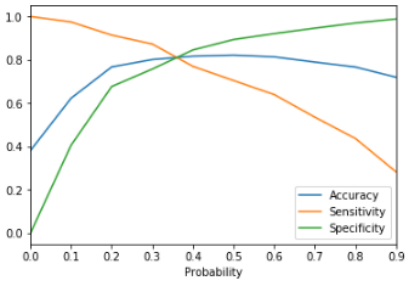


The curve obtained is closer to the left-hand border and then the top border of the ROC space as it is the tradeoff between sensitivity and specificity (any increase in sensitivity will be accompanied by a decrease in specificity), where

Sensitivity=Number of actual Yeses correctly predicted/ Total number of actual Yeses

Specificity=Number of actual Nos correctly predicted/Total number of actual Nos

2.7 Calculate the Optimum Threshold:



Optimum Threshold is given by the intersection of Accuracy, Sensitivity and Specificity. It is observed at 0.35. Hence, it is taken as the threshold for lead conversion probability.

2.8 Calculation of Lead Score:

The Lead Score is calculated based on the Probability computed by the Logistic Regression model based on the predicted lead conversions.

3.Inference:

- Lead Score is calculated based on the predicted probability of the converted leads.
- Conversion rates have been increased from 30% to 88.5% by following up with only these predicted leads.
- Lead Sources - Welingak website and by Reference have the major impact to get a lead converted and Working Professionals are the target audience as these are the top contributors in the model.