Summary

X Education is an online course provider for industry professionals. Company markets its courses on multiple websites and search engines like google. The aim of company is to increase its lead conversion for potential leads.

Below is a brief summary of steps followed for the analysis performed:

1. **Cleaning data:**
   The data was partially clean except for a few null values and the option select had to be replaced with a null value since it did not give us much information. Few of the null values were changed to 'not provided' so as to not lose much data. Although they were later removed while making dummies. Since there were many from India and few from outside, the elements were changed to 'India', 'Outside India' and 'not provided'.

2. **Data Understanding and Exploration:**
   A quick EDA was done to check the condition of our data. It was found that a lot of elements in the categorical variables were irrelevant. The numeric values seems good and no outliers were found.

2. **Dummy Variables:**
   The dummy variables were created and later on the dummies with 'not provided' elements were removed. For numeric values we used the MinMaxScaler.

3. **Train-Test split:**
   The split was done at 70% and 30% for train and test data respectively.

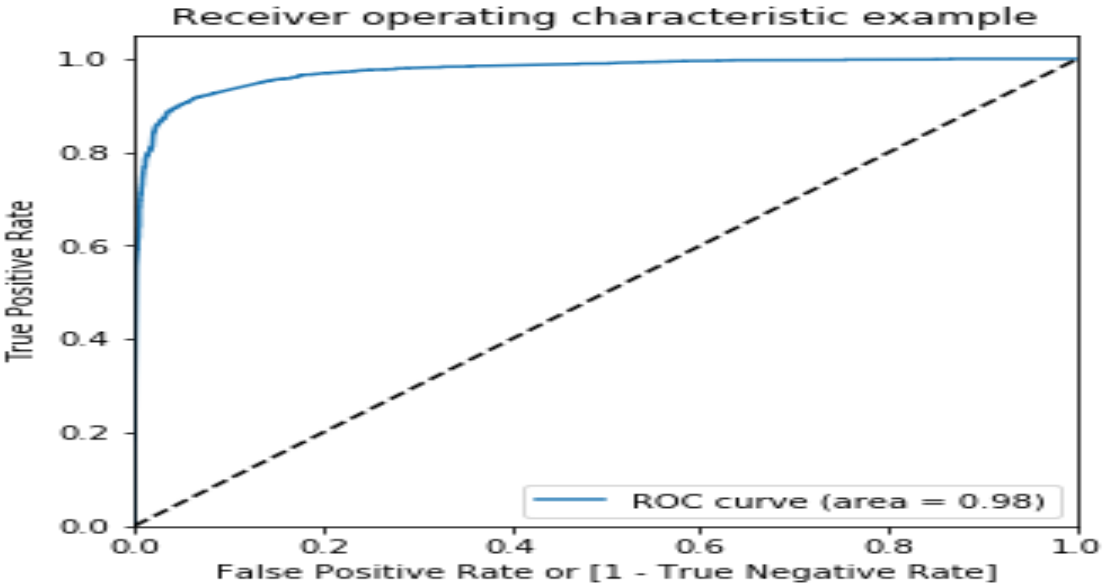
4. **Model Building and Evaluation:**

After data was cleaned, proceeded further to create the model. Since data contains categorical variables, performed the logistic regression.

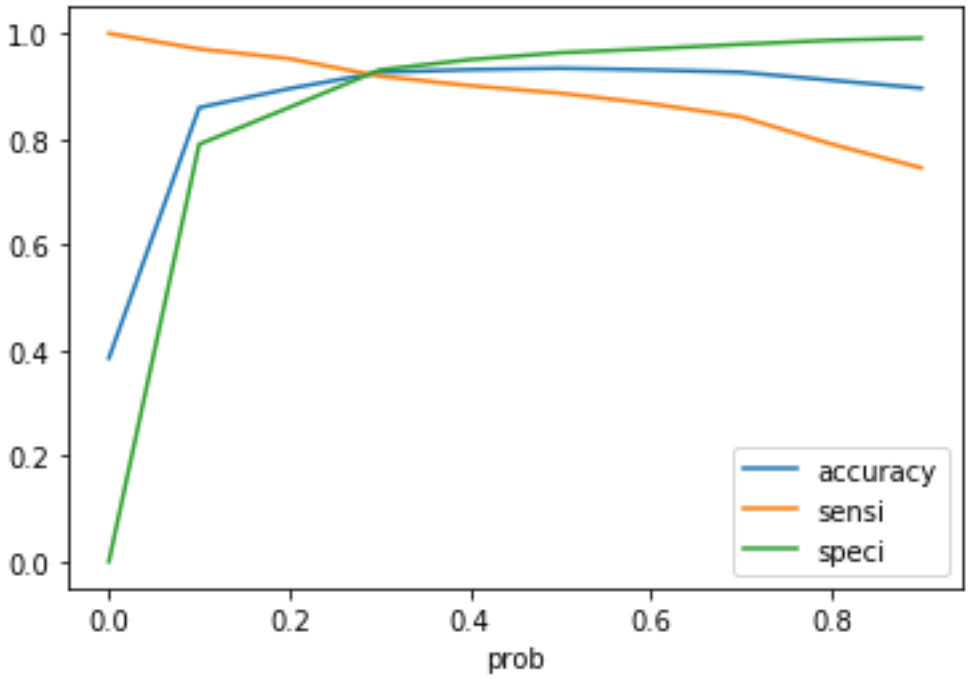
1. Scaling:
   Performed scaling on numeric columns so that data points are relatively dispersed.
2. RFE:
   Used RFE to reduce the number of features form 59 to 15.
3. Assessing the model with Stats Model:
   Used stats model for checking the summary of the model.
4. Calculated the sensitivity of the model on train data which turned out to be around 90%.

Plotting ROC Curve:

Earlier for predicting the values on train data, we used 0.5 as cut off value. Plotted ROC curve to get exact cut off value.



Finding Optimal Cut Off Value:



Calculated accuracy sensitivity and specificity for various probability cut-offs and plotted them to get the above graph.

All the variables intersect each other at 0.3, hence cut off value = 0.3

Making Predictions on Test Data:

When ran for Test Data, Model produced 92 % Sensitivity and 93% Accuracy as per the desired goal.

X-----X-----X ---- X