Lead Scoring Case Study

By

1. Sandesh
2. Rajani M
3. VinayKumar V
4. Srikar

About Problem Statement

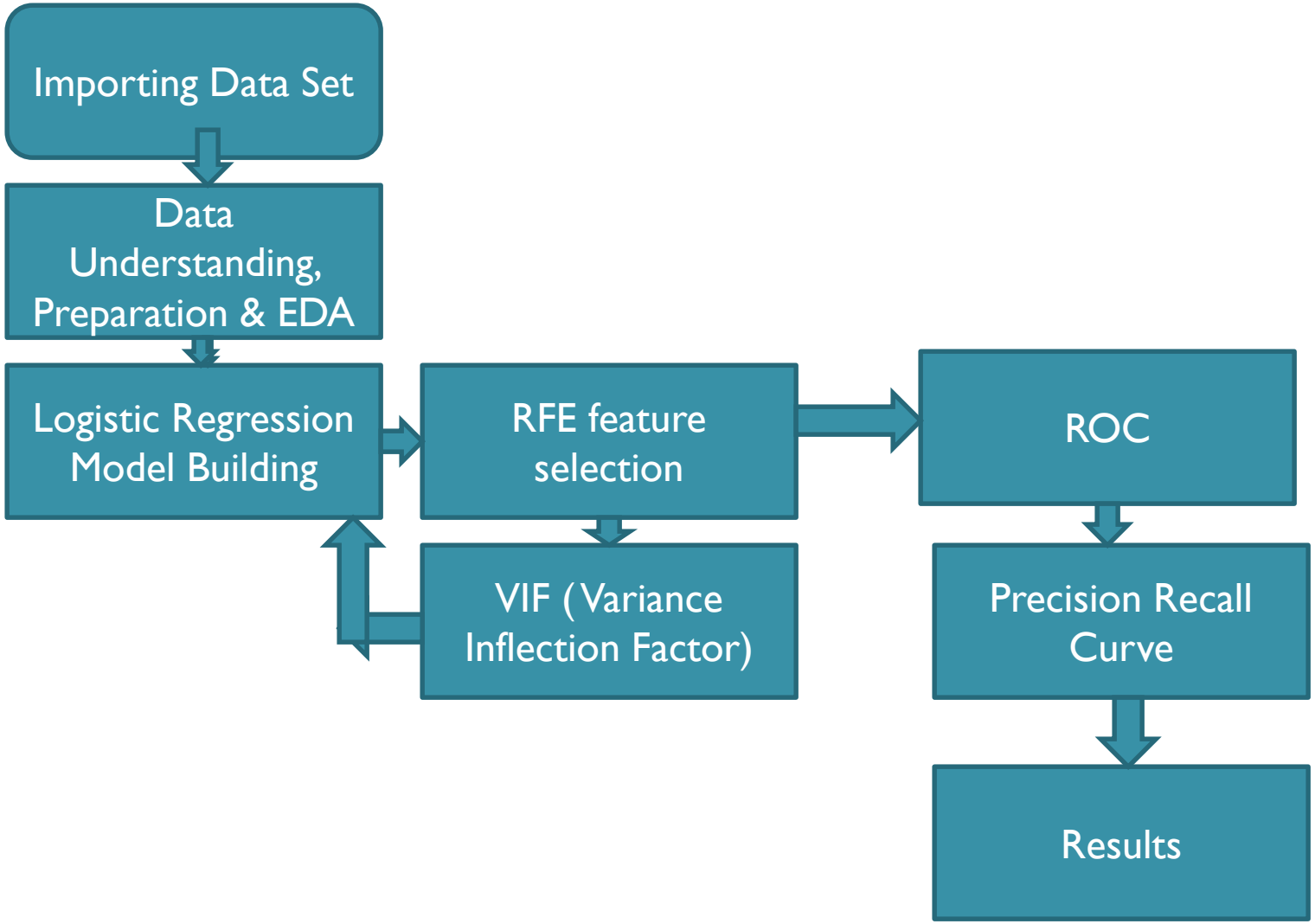
An instruction organization named X Education pitches online courses to industry experts. On some random day, numerous experts who are keen on the courses arrive on their site and peruse for courses. The organization showcases its seminars on a few sites and web indexes like Google. When these individuals arrive on the site, they may peruse the courses or top off a structure for the course or watch a few recordings. At the point when these individuals top off a structure giving their email address or telephone number, they are grouped to be a lead. In addition, the organization additionally gets leads through past referrals. When these leads are gained, workers from the business group begin making calls, composing messages, and so on. Through this procedure, a portion of the leads get changed over while most don't. The run of the mill lead transformation rate at X training is around 30%. The organization expects you to fabricate a model wherein you have to dole out a lead score to every one of the leads to such an extent that the clients with higher lead score have a higher change shot and the clients with lower lead score have a lower transformation possibility. The CEO, specifically, has given a ballpark of the objective lead change rate to be around 80%.

Business Objective

There are a lot of leads generated in the initial stage(let 100%) but only a few of them come out as paying customers from the bottom(30%), so we need to nurture the potential leads well in order to get a higher(80%) lead conversion.

Problem Solving Methodology

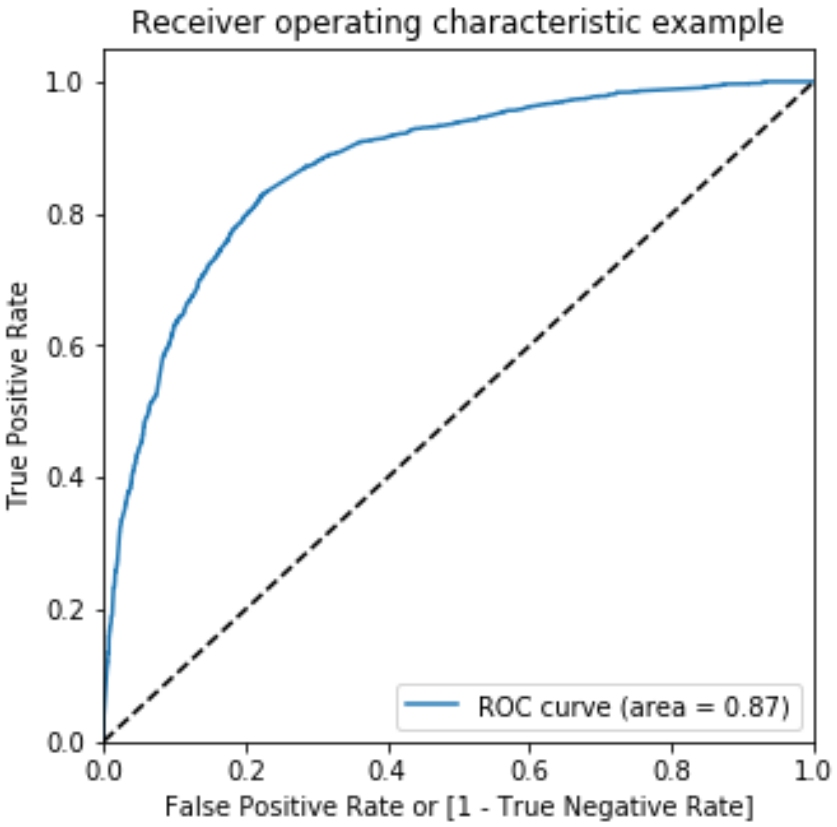
- Flow Chart :



Problem solving method

- **Step 1(Data Preparation):** Importing data set and understanding data, data cleaning & EDA
- **Step 2: (Logistic Regression Model Building):**
- - Logistic Regression Model building using RFE
- - using RFE we taken top 20 features
- - using GLM fitting the model based on Statistical measures and VIF index
- - Iterating the process till we get better accurate score on the model
- - predicting the futures
- **Step 3:** finding sensitivity and specificity of the model using confusion matrix
- **Step 4:** Using ROC curve defining new cut off based prediction of features again
- **Step 5:** Using precision and recall curve we will again taking new cut off point and again predicting the futures.
- **Step 6:** obtaining required results from the above process.

ROC - Curve

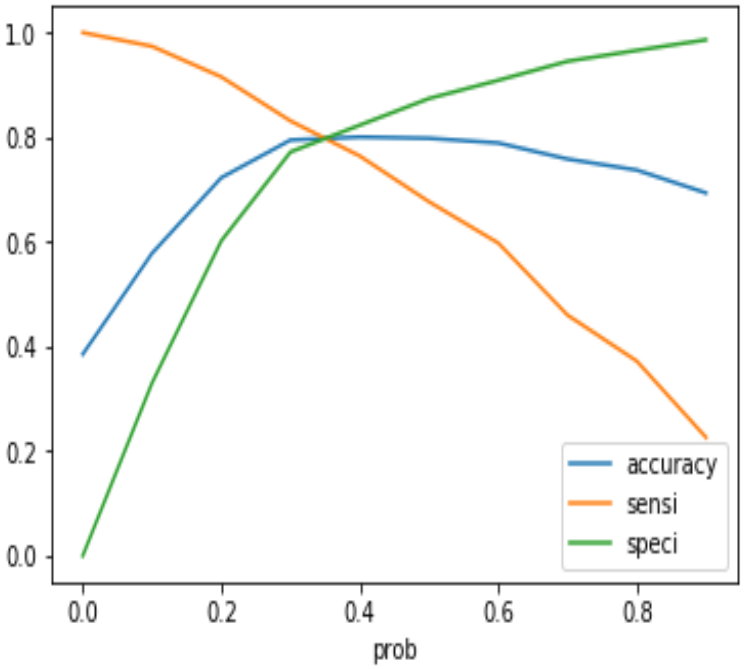


Here, receiver operating characteristic curve, or ROC curve, is a graphical plot that illustrates the diagnostic ability of a binary classifier system as its discrimination threshold is varied.

ROC curve value = 0.87

Above ROC curve is inclined towards and Y axis and area under curve is 0.87 which is pretty good.

ROC-Curve



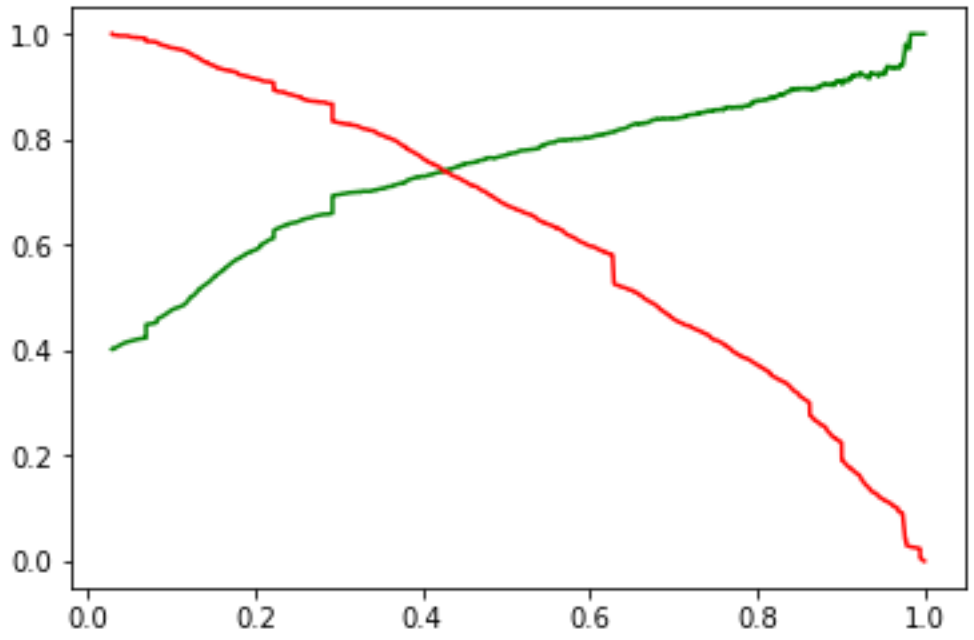
*We see at 0.33 all the three metrics meet each other.
But as per business requirement
*'Precision' should be at least 0.8.
So let's make the cut off as 0.62 at which our 'Precision' is 0.8.

Observations from above metrics

We see accuracy is good but sensitivity is little less.
We see False positive rate less which is good.
But Precision is 0.76 but it should be 0.8 as per business.
Let's plot ROC curve to find optimal cutoff point

Precision and Recall curve

The **precision-recall curve** shows the tradeoff between **precision and recall** for different threshold. A high area under the **curve** represents both high **recall** and high **precision**, where high **precision** relates to a low false positive rate, and high **recall** relates to a low false negative rate



***ROC Curve : 80.08**
***Precision & recall : 79.79**
***Precision Value :**
***Log RFE : 79.99**
***ROC Curve : 74.5**
***Precision & recall : 80%**

Results Obtained

- Based on above result formulated from RFE Logistic Regression , ROC Curve Model & precision & recall Model
- In all three cases we got accuracy score
- Log RFE 80.8
- ROC Curve : 80.08
- Precision & recall : 79.79
- Precision Value :
- Log RFE : 79.99
- ROC Curve : 74.5
- Precision & recall : 80%
- the top three variables in your model which contribute most towards the probability of a lead getting converted

  "Lead Origin , Total Time Spent on Website, Lead Profile"

Thank You !