X EDUCATION

Case Study

Introduction

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.

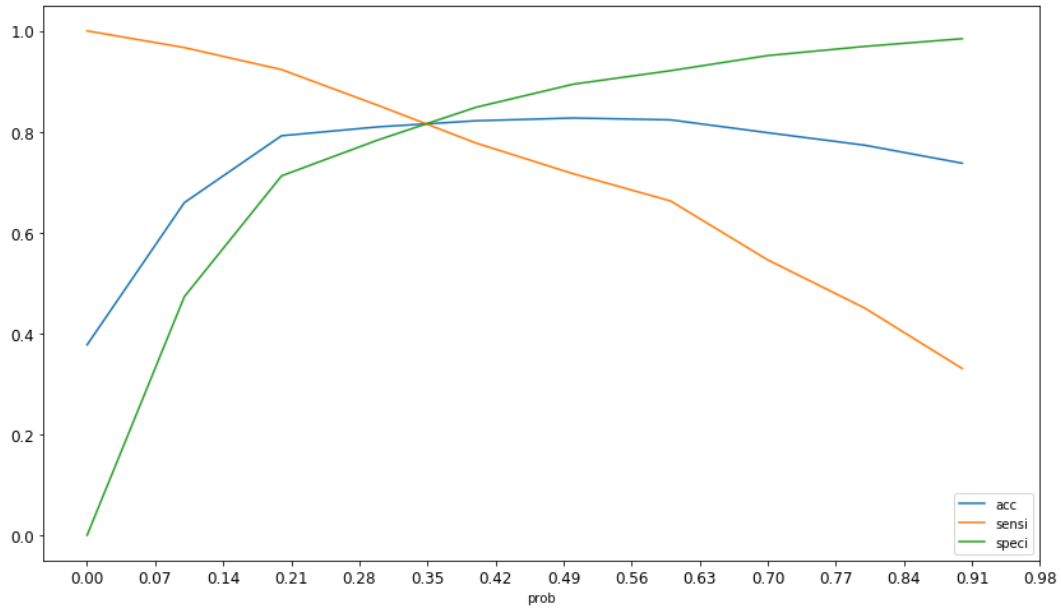
The company wishes to identify the most potential leads, also known as 'Hot Leads'. If they successfully identify this set of leads, the lead conversion rate should go up as the sales team will now be focusing more on communicating with the potential leads rather than making calls to everyone.

Problem Statement

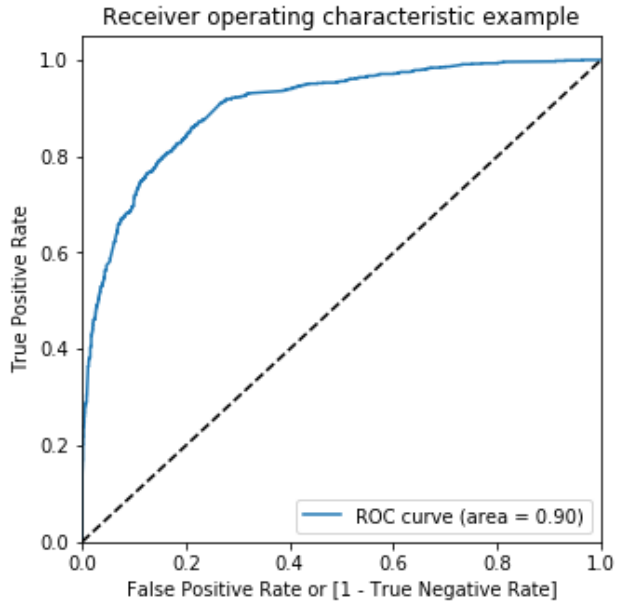
- We have been provided with a dataset from the past with around 9000 data points. This dataset consists of various attributes such as Lead Source, Total Time Spent on Website, Total Visits, Last Activity, etc. which may or may not be useful in ultimately deciding whether a lead will be converted or not.
- The target variable, in this case, is the column 'Converted' which tells whether a past lead was converted or not wherein 1 means it was converted and 0 means it wasn't converted.

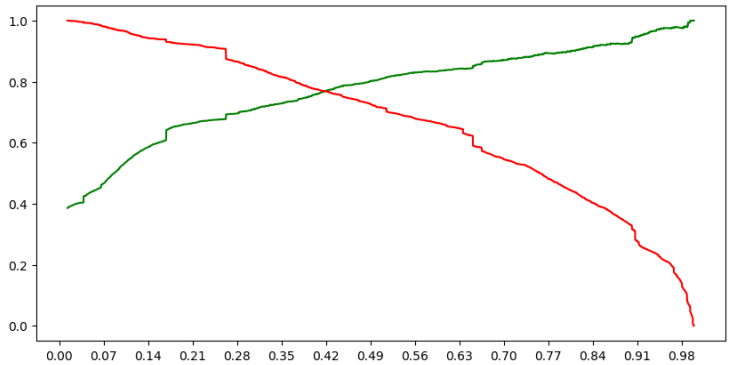
Analysis Approach

- We need to build a logistic regression model for the variable "Converted" such that the probability that we will get between 0-1 will be multiplied by 100 to get the "Lead Score".
- We need to remove columns with only 1 unique data, and nulls if any.
- We need to clean some data as it contains text like "Select" (because the data is likely collected from a form).
- We then do dummy variable creation.
- We need to then do RFE to select fewer columns as it contains over 100 columns.
- After RFE we start with 25 variables and we reduce them such as p-value is below 5% significance level and VIF values do not go above 3.
- As we go on reducing the number of variables, we arrive at 16 variables.



Sensitivity, Accuracy and Specificity





ROC Curve of the Model

Business Summary

- In our model we have final 16 variables, out of which 3 categorical variables play an important role.
  - Source of the Lead – Facebook, Welingak Website, Olark Chat etc,
  - Last Note – SMS Sent, Unsubscribed etc.
  - Lead Profile – Student, Potential Lead etc.
- The model has 80% Accuracy and True Positive Rate.
- The model has 10% False Positive Rate which is quite low and is a good result.