Reported to Upgrad

7th Oct 19,

Bangalore

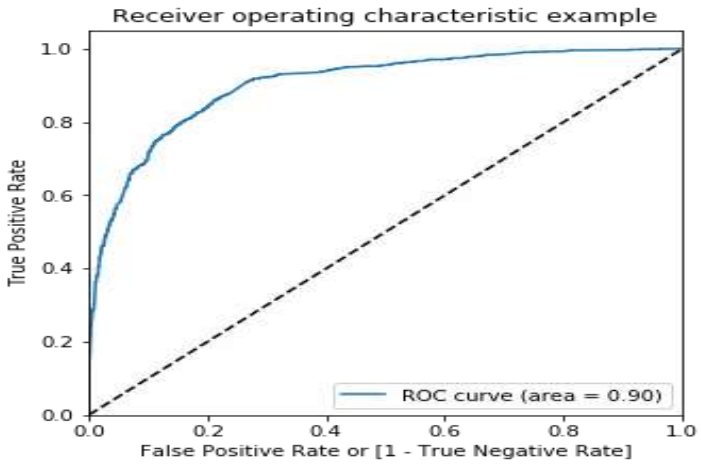
An education company named X Education sells online courses to industry professionals. On any given day, many professionals who are interested in the courses land on their website and browse for courses. To acquire the leads the sales team start making calls, writing emails, etc. Through this process, some of the leads get converted while most do not. The typical lead conversion rate at X education is around 30%.

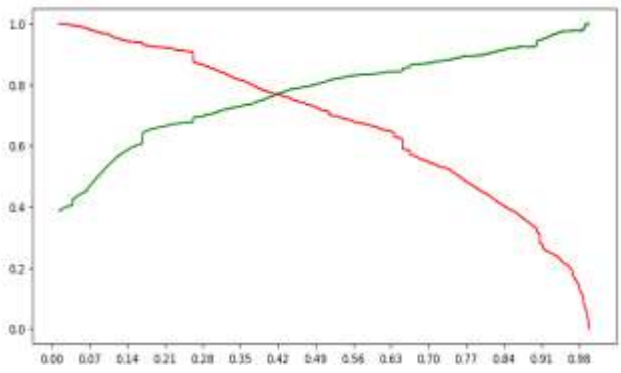
We have been provided with its dataset from the past with around 9000 data points. This dataset consists of various attributes such as Lead Source, Total Time Spent on Website, Total Visits, Last Activity, etc. which may or may not be useful in ultimately deciding whether a lead will be converted or not. We needs to build a model wherein we need to assign a lead score 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 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.

Steps followed to make some analysis.

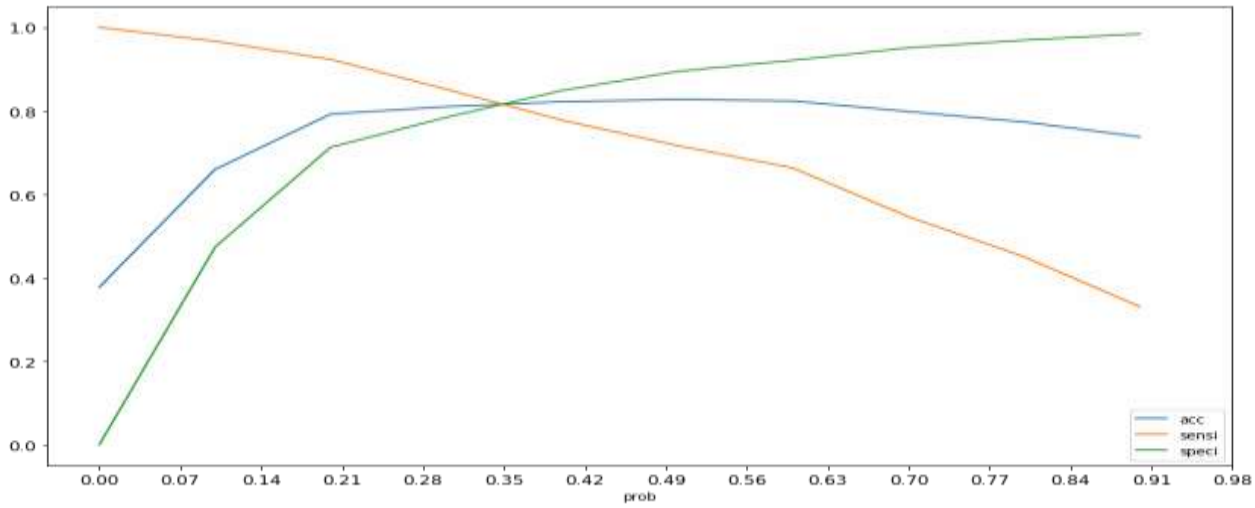
1. We need to build a logistic regression model for the variable "Converted" such that the probability that we will get between 0-1 will be multiplied by 100 to get the "Lead Score".
2. We need to remove columns with only 1 unique data, and nulls if any.
3. We need to clean some data as it contains text like "Select" (Because the data is likely collected from a form).
4. We then do dummy variable creation.
5. We need to then do RFE to select fewer columns as it contains over 100 columns.
6. 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.
7. As we go on reducing the number of variables, we arrive at 16 variables.

Below is the ROC curve of the model.



Below is the Sensitivity, Accuracy and Specificity.




- 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 90% Accuracy and True Positive Rate.
- The model has 10% False Positive Rate which is quite low and is a good result.

Respectfully submitted by:

Nikhil Singh

Neha Kumari