Lead Score Case Study

- by Sachin Sasy & Tanya Chaudhary

Problem statement:

- A Score needs to be assigned to the prospective leads of a company that offers online educational courses to measure their chances of enrolling for a course

Analysis approach:

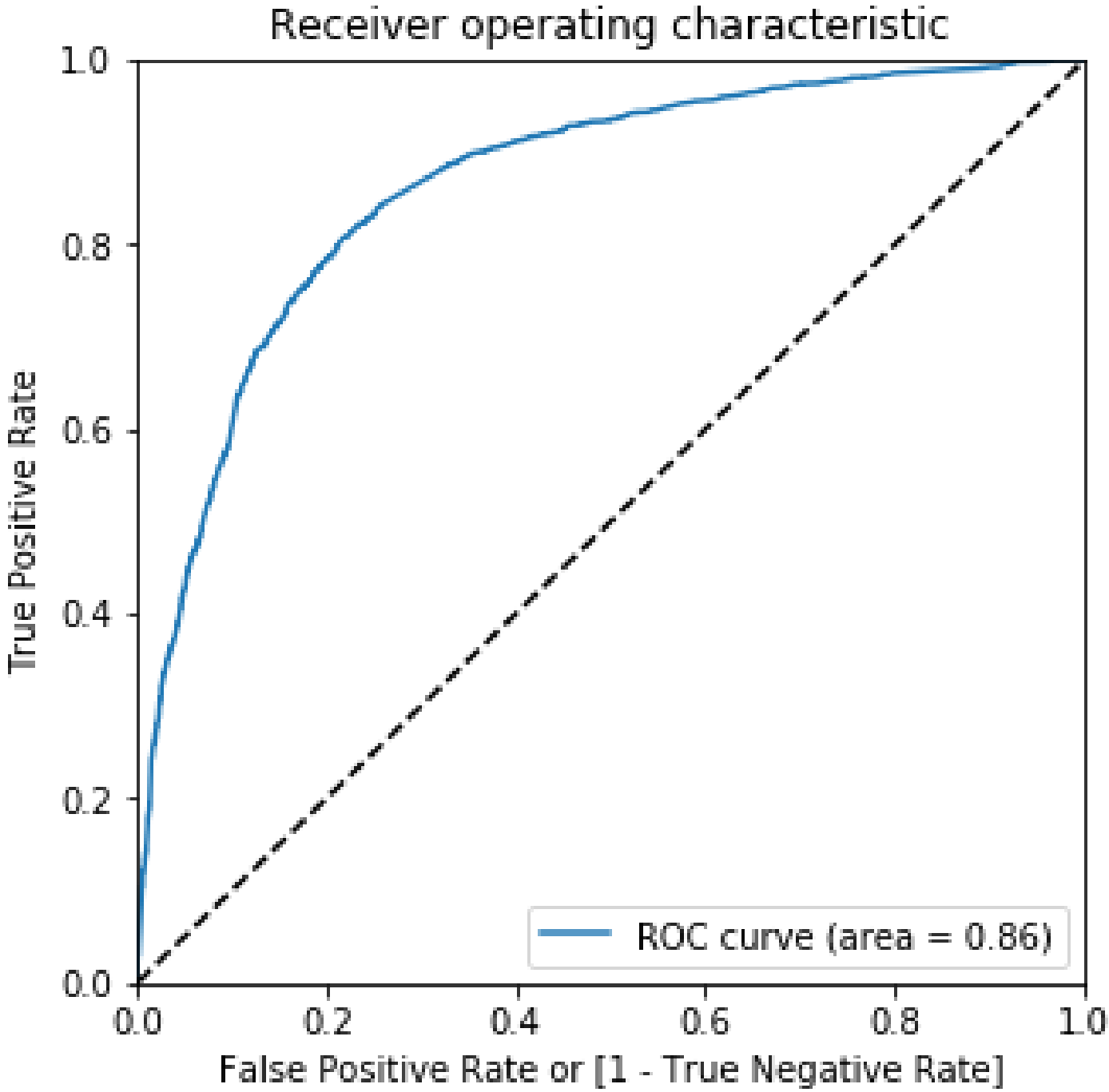
- The dataset provided was analyzed and cleaned and then split into train and test sets
- A regression model was built on train set using RFE and iteratively removing parameters to remove multi collinearity among the parameters.
- The calculated VIF for the resultant model was acceptable and the model was executed on the test data set. An optimal cutoff probability of 0.415 was arrived at using Precision vs Recall graph.
- An accuracy of 80% was achieved in both the train and test sets and a lead score based on the probability of conversion was calculated and appended back to the original dataset.

ROC Curve

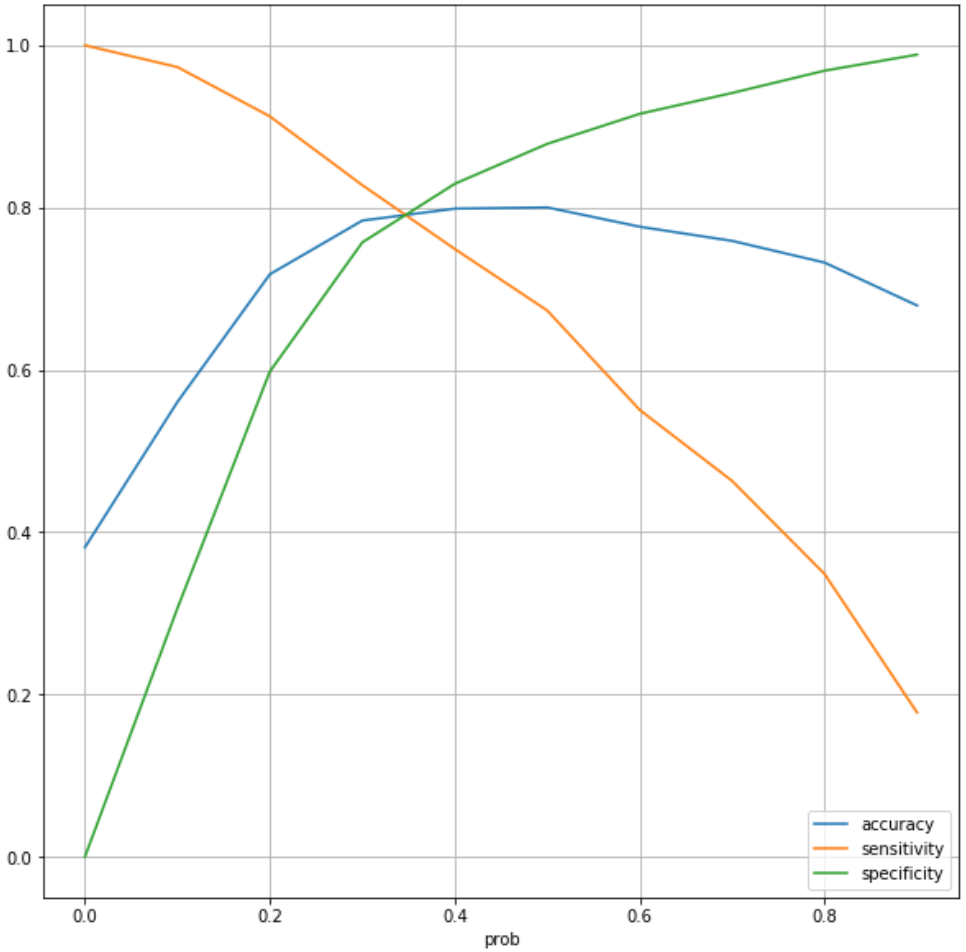
The ROC curve plotted for the model had an area of 0.86 under the curve. The curve can also be seen to be considerably closer to the top left corner thereby confirming that it is accurate.

The confusion matrix for the model was validated and the following metrics were calculated:

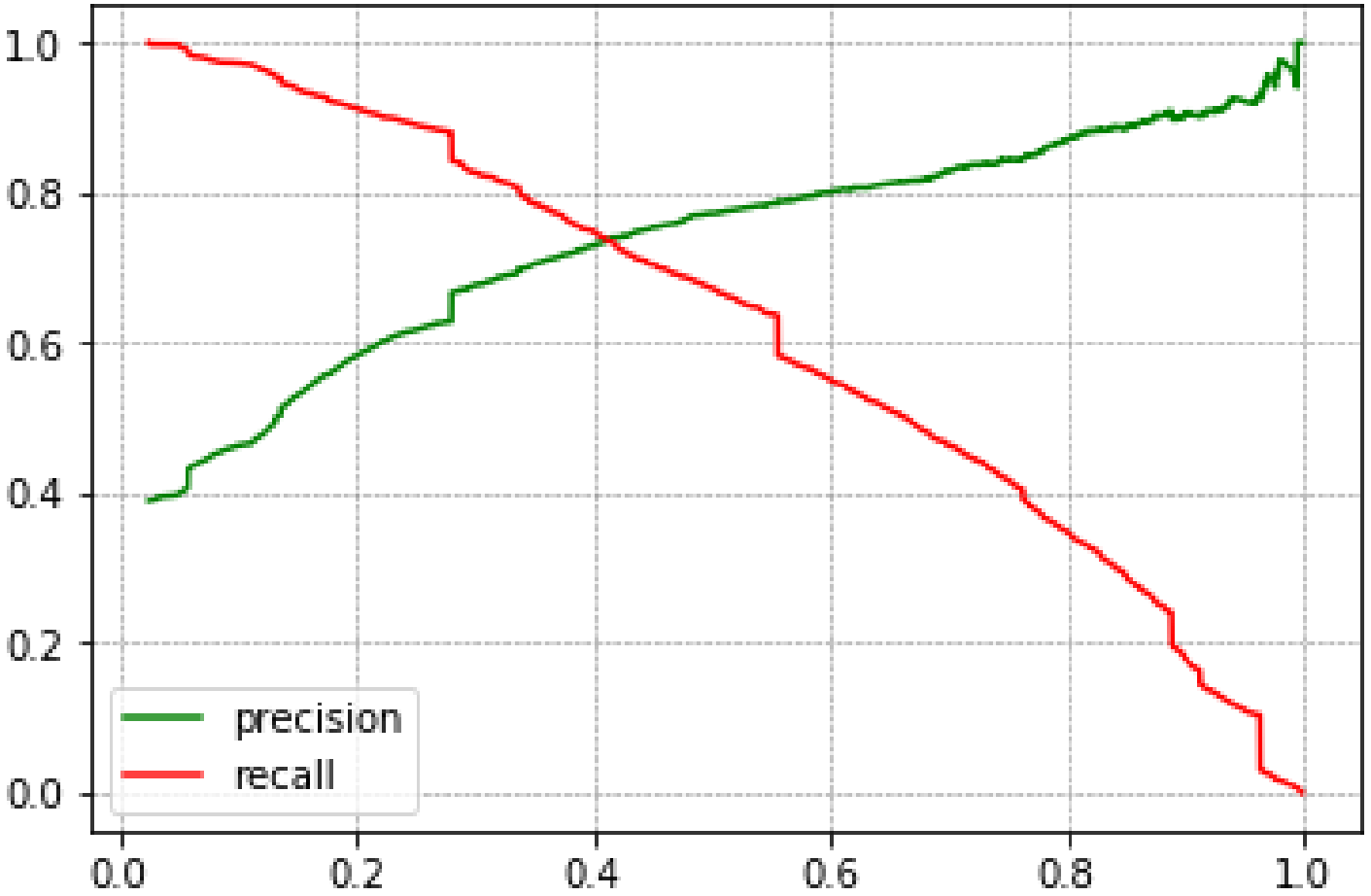
Accuracy = 80 %.
Sensitivity = 67 %
Specificity = 88 %
Precision = 77 %



The optimal probability cutoff for the model was found to be 0.36 by plotting the calculated accuracy, sensitivity and specificity of the model



The precision vs recall graph was then plotted, and the optimal cutoff probability was found to be 0.415



Logistic Regression model using RFE

The table to the right shows the final set of variables selected for making the model and the p-values and coefficient values for each of the model attributes. The attributes were arrived at by iteratively building the model removing the attribute with the most p-value at each iteration and then analyzing the variation in the other attributes. The final Logistic regression model equation is concluded as below.

	coef	std err	z	P>|z|	[0.025	0.975]
const	-1.0769	0.053	-20.324	0.000	-1.181	-0.973
Lead Origin_Lead Add Form	4.1392	0.178	23.303	0.000	3.791	4.487
Lead Source_Olark Chat	1.1286	0.098	11.493	0.000	0.936	1.321
Last Activity_Email Bounced	-1.5524	0.292	-5.309	0.000	-2.126	-0.979
Last Activity_Had a Phone Conversation	2.3856	0.618	3.859	0.000	1.174	3.597
Last Activity_Olark Chat Conversation	-0.9434	0.165	-5.711	0.000	-1.267	-0.620
Last Activity_SMS Sent	1.1682	0.070	16.615	0.000	1.030	1.306
Last Notable Activity_Modified	-0.9008	0.077	-11.710	0.000	-1.052	-0.750
Last Notable Activity_Unreachable	1.5570	0.494	3.149	0.002	0.588	2.526
Total Time Spent on Website	1.1254	0.038	29.520	0.000	1.051	1.200

Conversion probability = -1.0769 + (4.1392 x Lead Origin_Lead Add Form) + (1.1286 x Lead Source_Olark Chat)
- (1.5524 x Last Activity_Email Bounced) + (2.3856 x Last Activity_Had a Phone Conversation)
- (0.9434 x Last Activity_Olark Chat Conversation) + (1.1682 x Last Activity_SMS Sent)
- (0.9008 x Last Notable Activity_Modified) + (1.5570 x Last Notable Activity_Unreachable)
+ (1.1254 x Total Time Spent on Website)

The Lead score to be assigned to the leads was calculated from the conversion probability received from the regression model as:

Lead Score = round(Conversion probability x 100)

Conclusions / Insights:

A high lead score can be achieved if the lead,

1. originated from the Lead Add Form,
2. when the last activity corresponding to the lead was a phone conversation and
3. when the lead source was an olark chat

It can also be noted that the lead score is low when no further action was taken after

1. emails to the leads have bounced or
2. an olark chat conversation took place

In all the above situations - both high and low lead scores – the score can be increased further by setting up and follow up calls since having a phone conversation as the Last activity increases the lead score.