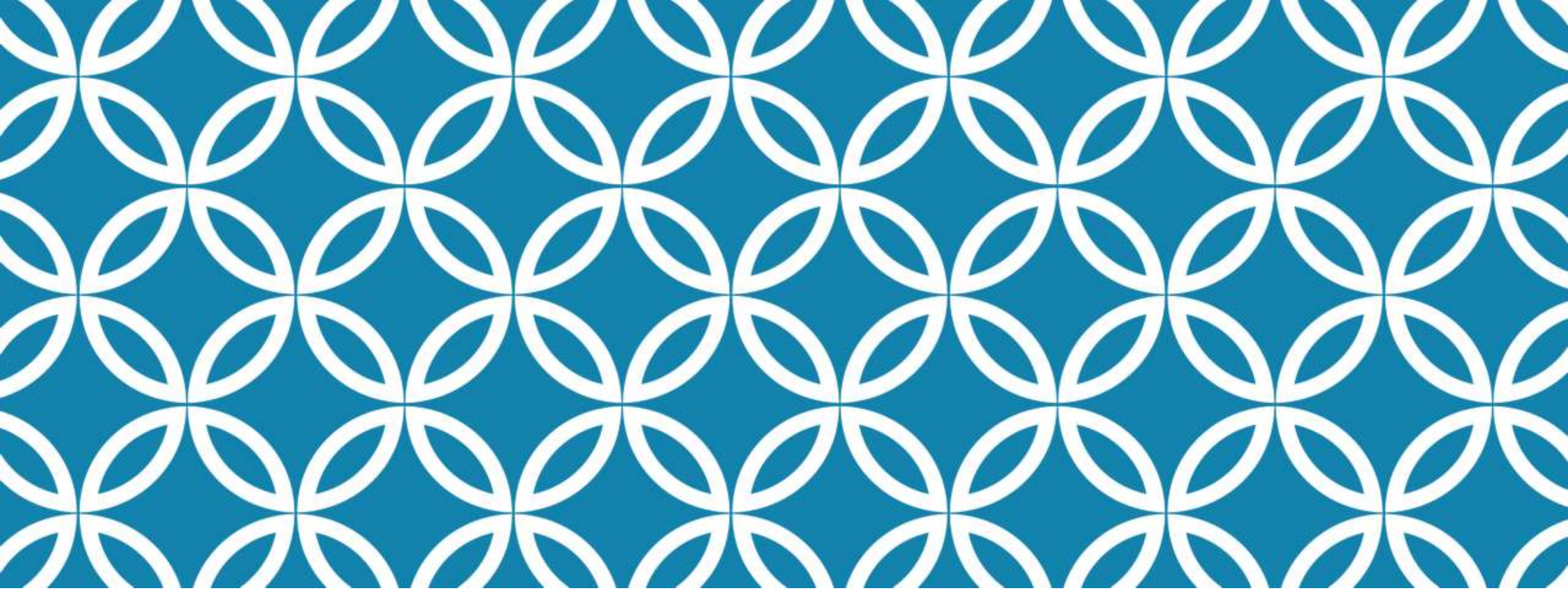


Lead Scoring Case Study

Dinesh Vemula
Yash Gupta
Nabeeg Akram

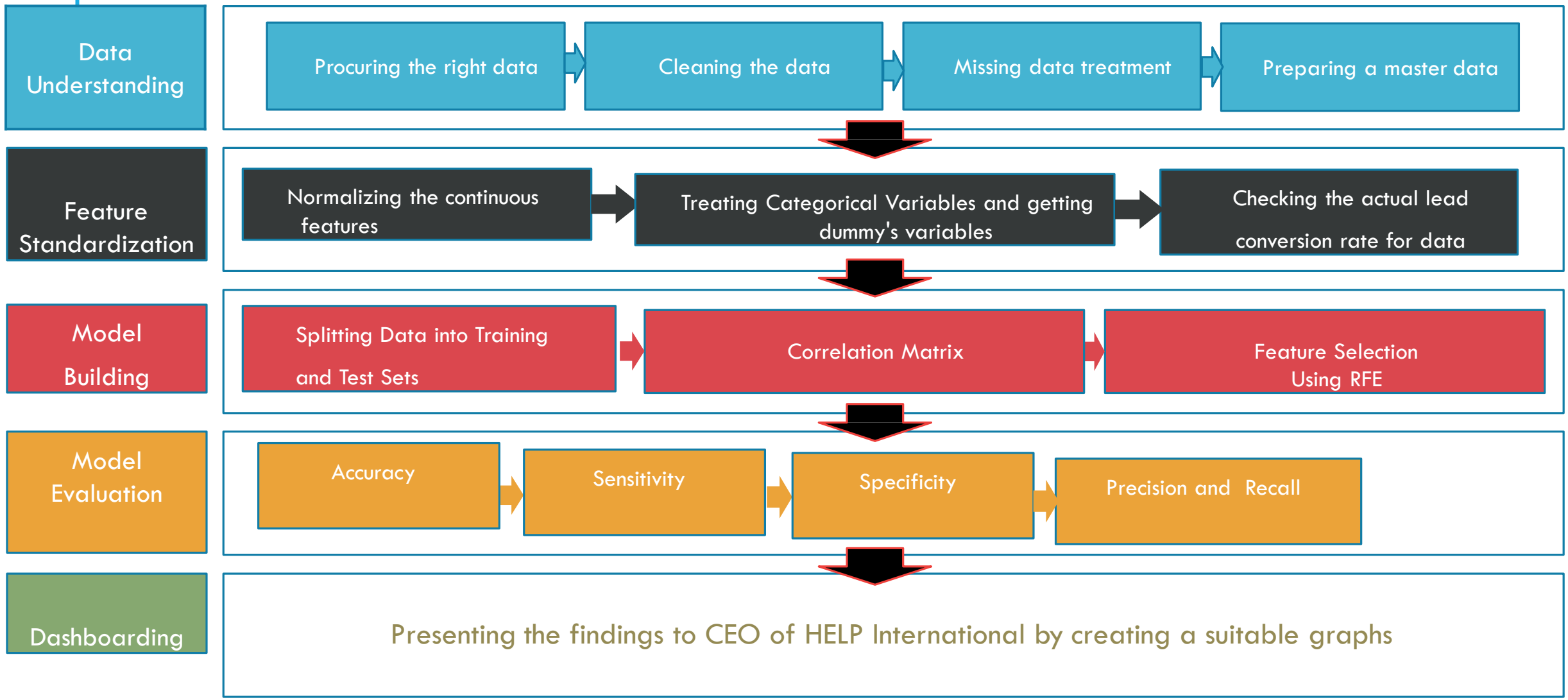
BUSINESS OBJECTIVE

Abstract

- Build a logistic regression model to assign a lead score between 0 and 100 to each of the leads which can be used to target potential leads. A higher score would mean that the lead is hot, i.e. is most likely to convert whereas a lower score would mean that the lead is cold and will mostly not get converted.

- To make this process more efficient, 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.

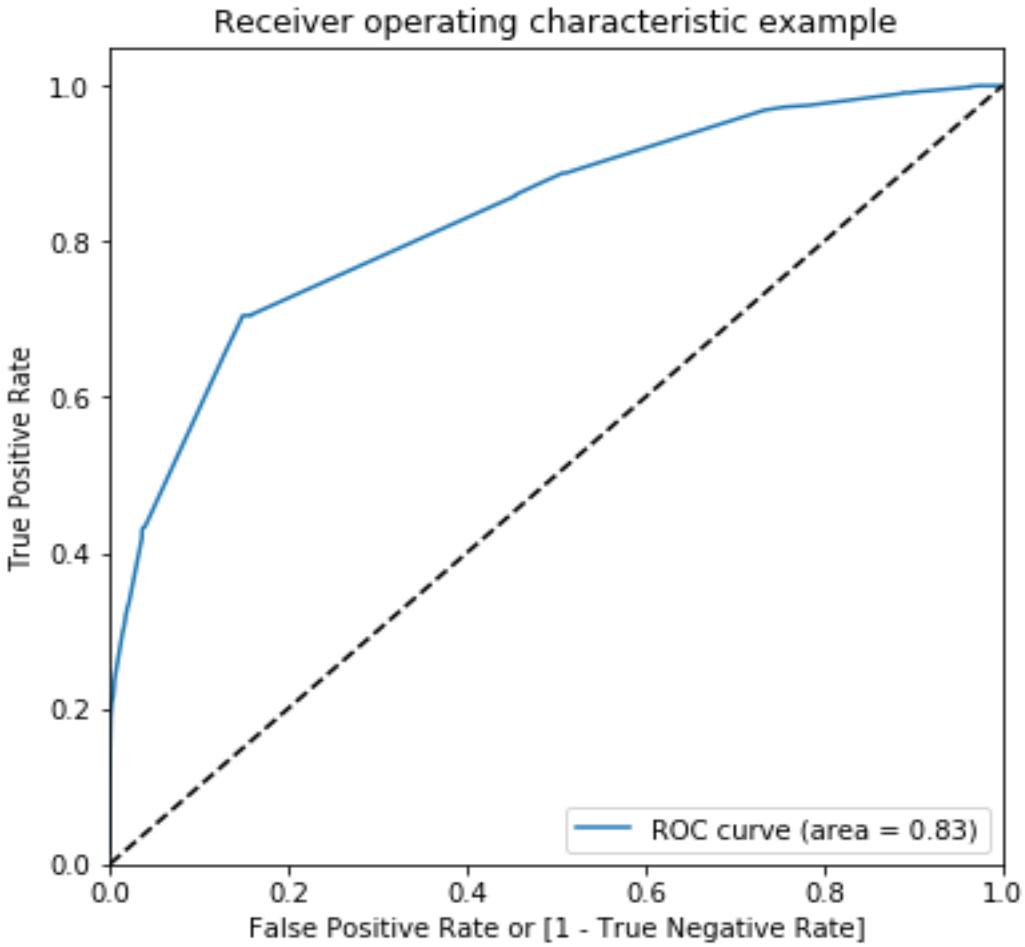
Problem Solving Methodology



DATA CLEANING

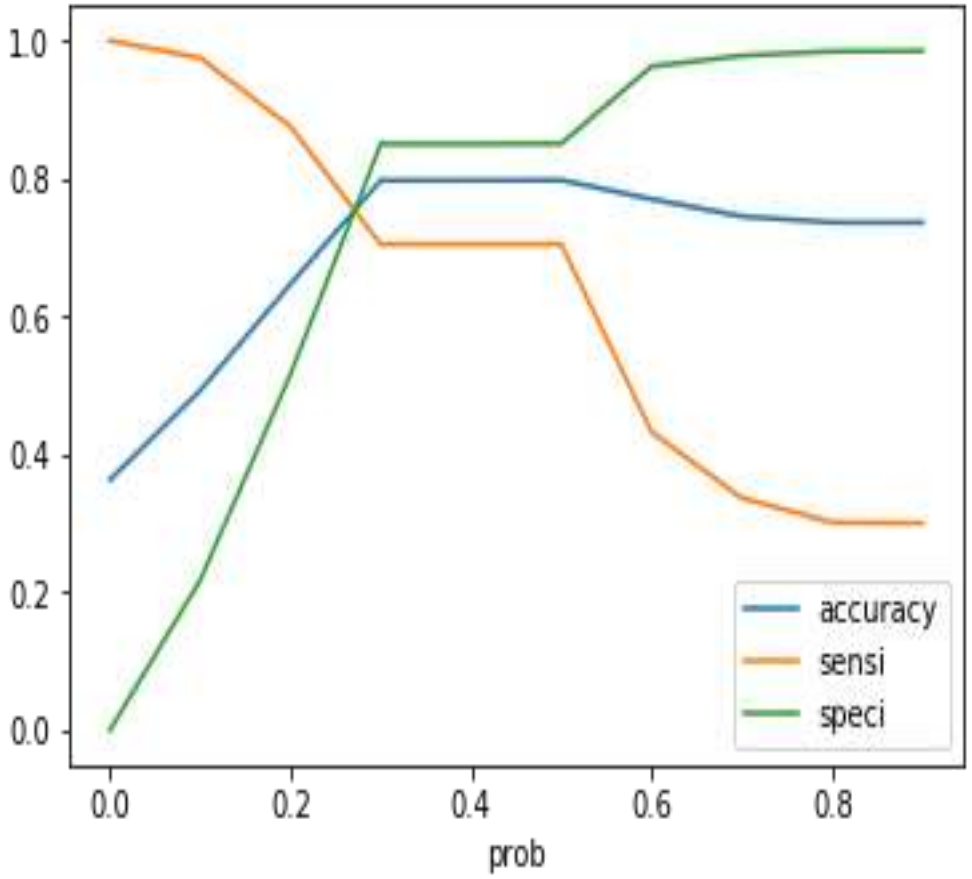
- Prospect ID'& 'Lead Number' is unique id for each unique lead. we can delete the Prospect ID & Lead Number
- 'Lead Quality' column has more than 50% of null values so we will be dropping that column
- 'Receive More Updates About Our Courses' column has unique value 'NO' so we will be dropping that column
- 'Magazine' column has unique value 'NO' so we will be dropping that column
- 'Update me on Supply Chain Content' column has unique value 'NO' so we will be dropping that column
- 'Get updates on DM Content' column has unique value 'NO' so we will be dropping that column
- 'I agree to pay the amount through cheque' column has unique value 'NO' so we will be dropping that column
- 'Tags' column is a comments section so we will be dropping that column
- 'Country' and 'City' column is not required since it is a online course everyone can access irrespective of country and city, so we will be dropping that column
- Asymmetrique Activity Index and Asymmetrique Profile Index scores has more than 45% of null values so we are dropping Index and scores

ROC CURVE



From the Roc curve it can be seen that the AUC is 0.83 which indicates the model is quite good.

Trade off Probability



From the curve it can be seen that optimum cut off for balancing sensitivity and specificity is 0.3 (Lead Score =30).

Model Result(1/2)

- Overall Accuracy : 0.7972 (80%)
- Confusion Matrix : [[1474 260]
  [292 697]]
- Sensitivity of the model : 0.7047 (70%)
- specificity of the model : 0.8500 (85%)
- Positive predictive value : 0.7283 (73%)
- Negative predictive value : 0.8346 (84%)
- *Precision* : 0.7283 (73%)
- *Recall* : 0.70475 (70%)

Model Result (2/2)

Top 3 top three variables in the model which contribute the most towards the probability of a lead getting converted.

1) Occupation
2) Lead Source
3) Lead Profile

Top 3 categorical/dummy variables in the model which should be focused the most on in order to increase the probability of lead conversion

1) Lead Add Form Welingak Website
2) Occupation Working Professional
3) Lead Add Form Reference

Recommendations (1/2)

- When The sales team, in particular, has around 10 interns allotted to them so high number of phone calls are affordable. we need to have tradeoff between sensitivity and specificity such a way that high sensitivity is desirable.

prob	accuracy	sensitivity	specificity
0.0	0.363202	1.000000	0.000000
0.1	0.492472	0.974722	0.217416
0.2	0.646713	0.875632	0.516148
0.3	0.797282	0.704752	0.850058
0.4	0.797282	0.704752	0.850058
0.5	0.797650	0.704752	0.850634
0.6	0.770107	0.432760	0.962514
0.7	0.745134	0.336704	0.978085
0.8	0.736320	0.301314	0.984429
0.9	0.736320	0.300303	0.985006

- From the above table, 0.20 (Lead Score = 20) is the optimum point to take it as a cutoff probability for this period of time so that almost all of the potential leads can be identified and contacted to make the lead conversion more aggressive.

Recommendations(2/2)

- Similarly, at times, the company reaches its target for a quarter before the deadline to make phone calls if its extremely necessary only, false positives must be minimized(this relates to specificity). We need to have tradeoff between sensitivity and specificity in such a way that high specificity is desirable.

prob	accuracy	sensitivity	specificity
0.0	0.363202	1.000000	0.000000
0.1	0.492472	0.974722	0.217416
0.2	0.646713	0.875632	0.516148
0.3	0.797282	0.704752	0.850058
0.4	0.797282	0.704752	0.850058
0.5	0.797650	0.704752	0.850634
0.6	0.770107	0.432760	0.962514
0.7	0.745134	0.336704	0.978085
0.8	0.736320	0.301314	0.984429
0.9	0.736320	0.300303	0.985006

- From the above table, 0.80 (lead score = 80) is the optimum point to take it as a cutoff probability for such a condition so that only the potential leads to be converted can be identified and hence thereafter only focused to avoid unnecessary phone calls.