Train In Data

Probability

Balanced dataset

Target
0
0
0
0
0
1
1
1
1
1

- The likelihood of receiving an observation of class 1 is 0.5

- If the model outputs a probability > 0.5 ➔ class 1
- If the model outputs a probability < 0.5 ➔ class 0

Imbalanced dataset

Target
0
0
0
0
0
0
0
0
0
1

- The likelihood of receiving an observation of class 1 is 0.1

- If the model outputs a probability > 0.1 ➔ class 1
- If the model outputs a probability < 0.1 ➔ class 0

- When we train models on imbalanced datasets, we tend to obtain lower probability values for the rare class.

Fine tune the probability threshold

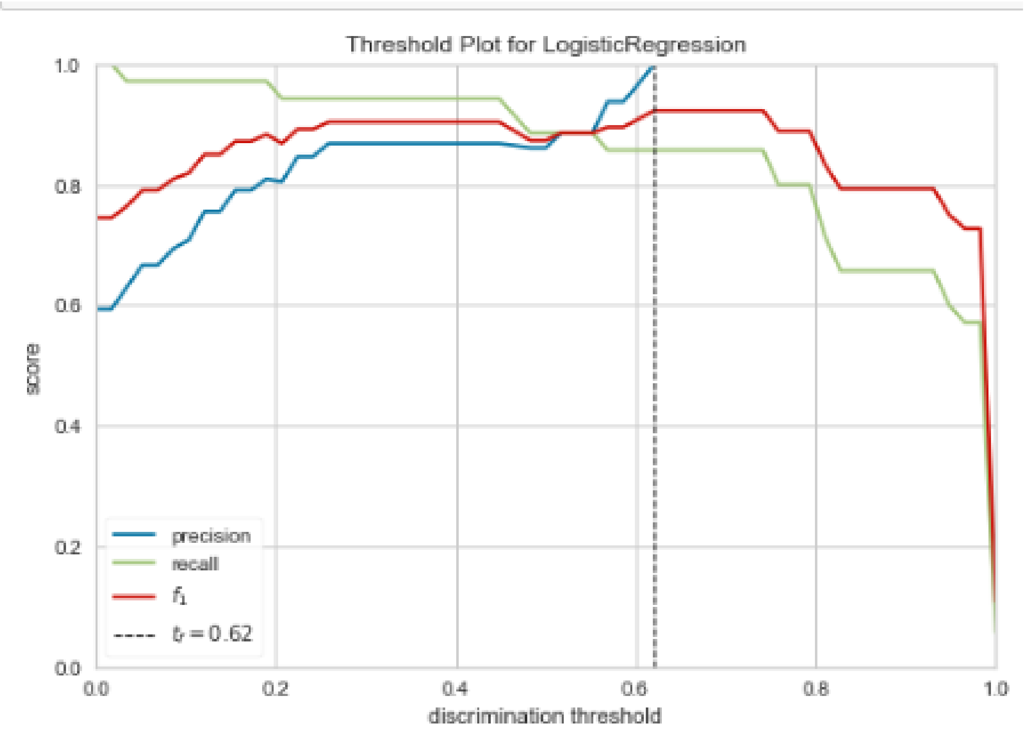
Using 0.5 as a default threshold does not make sense when we have imbalanced datasets.

We need to fine tune the threshold depending on what we want to optimise:

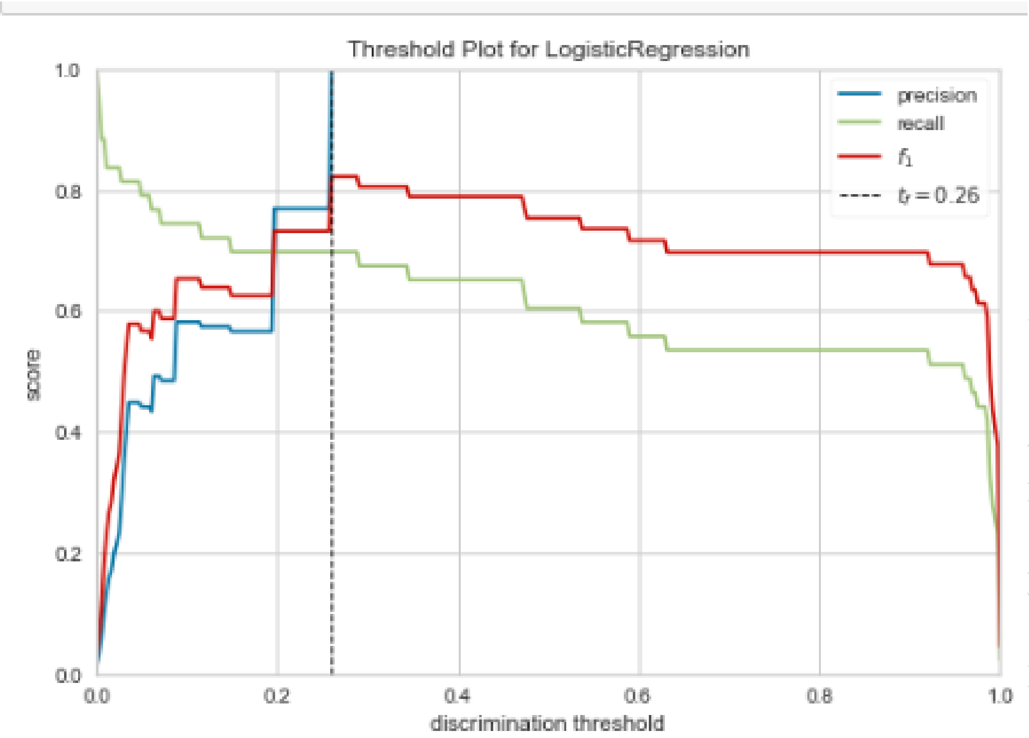
- Precision, recall or F-score
- False positive or false negative discovery rate

Probability and threshold

Balanced dataset



Imbalanced dataset



Train In Data

THANK YOU

www.trainindata.com