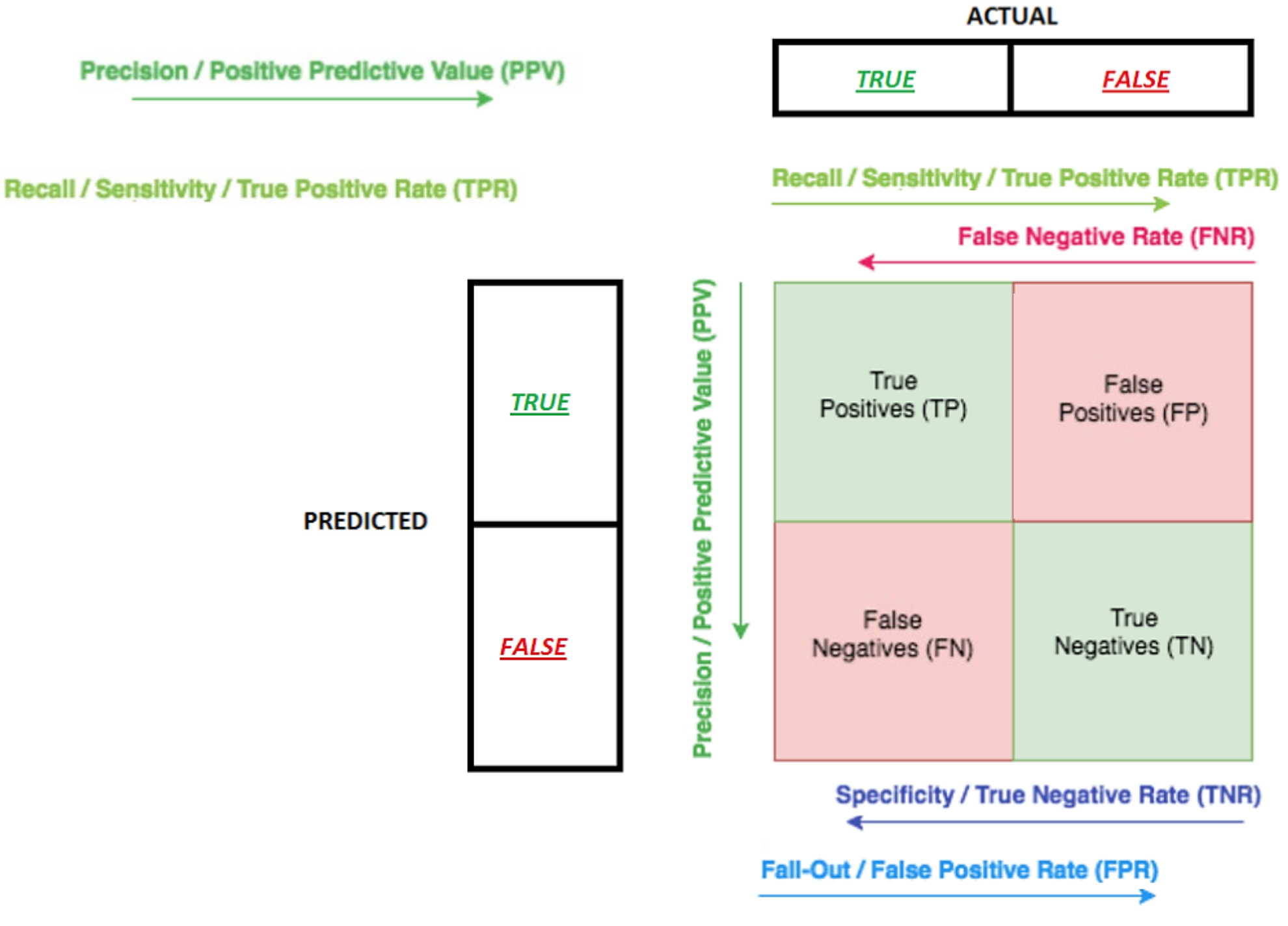
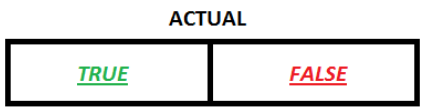
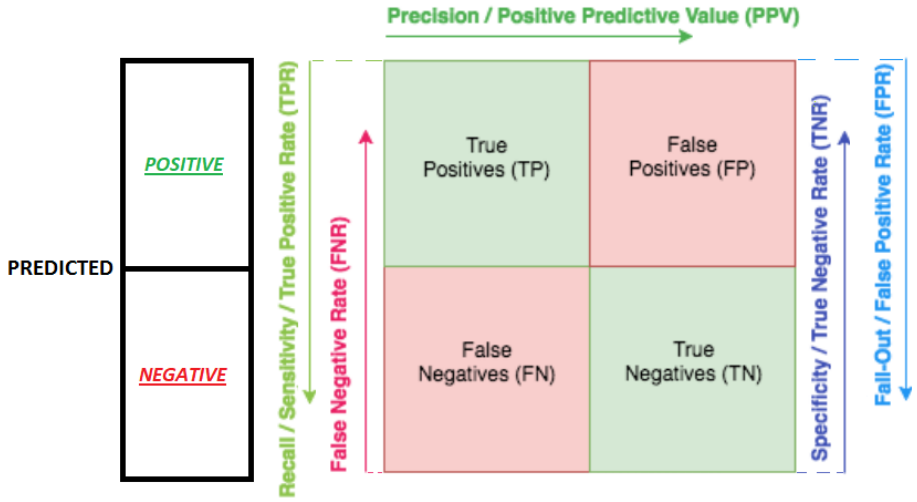

ACTUAL
TRUE
FALSE
Precision / Positive Predictive Value (PPV)
Recall / Sensitivity / True Positive Rate (TPR)
Recall / Sensitivity / True Positive Rate (TPR)
False Negative Rate (FNR)
PREDICTED
TRUE
FALSE
Precision / Positive Predictive Value (PPV)
True Positives (TP)
False Positives (FP)
False Negatives (FN)
True Negatives (TN)
Specificity / True Negative Rate (TNR)
Fall-Out / False Positive Rate (FPR)

Just Remember, we describe:
PREDICTED values as Positive and Negative
ACTUAL values as True and False





Each ray in the visualization above specifies the name of the metric.

Start point of each ray represents the numerator of that metric and the span of the ray represents the summation of the adjacent terms.

Note that each metric is essentially a fraction.

sensitivity, recall, hit rate, or true positive rate (TPR)

$$\mathrm{TPR} = \frac{\mathrm{TP}}{P} = \frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FN}} = 1 - \mathrm{FNR}$$

specificity, selectivity or true negative rate (TNR)

$$\mathrm{TNR} = \frac{\mathrm{TN}}{N} = \frac{\mathrm{TN}}{\mathrm{TN}+\mathrm{FP}} = 1 - \mathrm{FPR}$$

precision or positive predictive value (PPV)

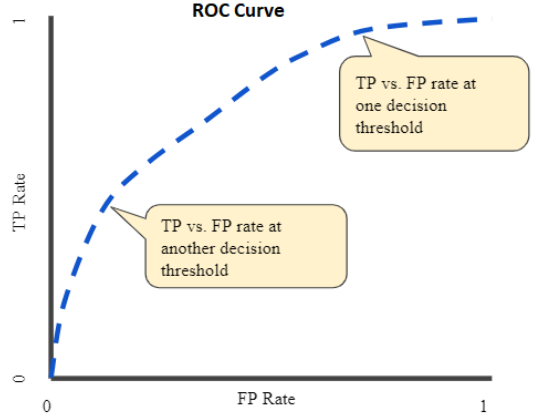
$$\mathrm{PPV} = \frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FP}} = 1 - \mathrm{FDR}$$

miss rate or false negative rate (FNR)

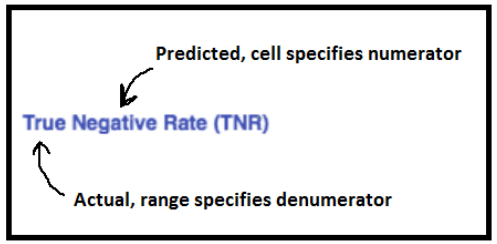
$$\mathrm{FNR} = \frac{\mathrm{FN}}{P} = \frac{\mathrm{FN}}{\mathrm{FN}+\mathrm{TP}} = 1 - \mathrm{TPR}$$

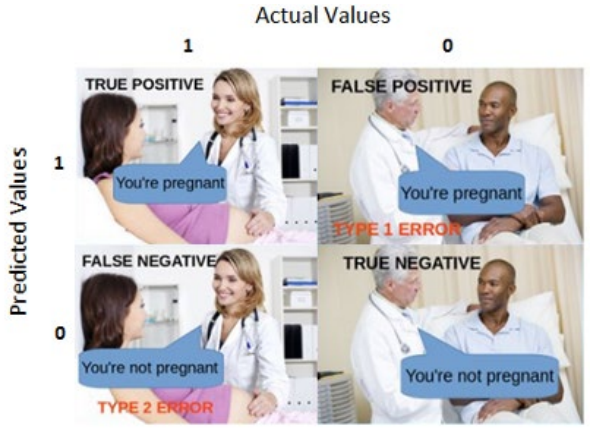
fall-out or false positive rate (FPR)

$$\mathrm{FPR} = \frac{\mathrm{FP}}{N} = \frac{\mathrm{FP}}{\mathrm{FP}+\mathrm{TN}} = 1 - \mathrm{TNR}$$



True Positive Rate:
num: #TP
den: #TP + #FN
NOTE:
False Negative is an inversion of True Positive





RECALL (SENSITIVITY): Out of all the positive instances, how much we predicted correctly. It should be high as possible. A measure of a classifiers completeness.

100% sensitive model -> it did NOT miss any True Positive, in other words, there were NO False Negatives. But there is a risk of having a lot of False Positives.

SPECIFICITY: 100% specific model -> it did NOT miss any True Negative, in other words, there were NO False Positives. But there is a risk of having a lot of False Negatives.

PRECISION: Out of all positive instances that we have predicted correctly, how many are actually positive. Measure of a classifiers exactness.

100% precise model -> it could catch all True positive but there were NO False Positive.

Recall gives us information about a classifier's performance with respect to false negatives.

(how many did we miss)

Precision gives us information about its performance with respect to false positives (how many did we catch)

As a rule of thumb, if the cost of having False negative is high, we want to increase the model **recall** (**sensitivity**).

For instance, in fraud detection or sick patient detection, we don't want to label/predict a fraudulent transaction (True Positive) as non-fraudulent (False Negative). Also, we don't want to label/predict a contagious sick patient (True Positive) as not sick (False Negative).
This is because the consequences will be worse than a False Positive (incorrectly labelling a harmless transaction as fraudulent or a non-contagious patient as contagious).

On the other hand, if the cost of having False Positive is high, then we want to increase the model **specificity** and **precision**.

For instance, in email spam detection, we don't want to label/predict a non-spam email (True Negative) as spam (False Positive). On the other hand, failing to label a spam email as spam (False Negative) is less costly.

$$F\text{ - measure} = \frac{2*\text{Recall}*\text{Precision}}{\text{Recall} + \text{Precision}}$$

It is difficult to compare two models with low **Precision** and high **Recall** or vice versa. So to make them comparable, we use **F-Score**. **F-score** helps to measure **Recall** and **Precision** at the same time.

It uses Harmonic Mean in place of Arithmetic Mean by punishing the extreme values more:
Recall=0.7 ; **Precision**=0.6 -> **F-score** = 0.65
Recall=0.95 ; **Precision**=0.01 -> **F-score** = 0.02 (but mean = 0.48)

It is multiplied by 2 so that when **Recall** = **Precision** = 1 -> **F-score** = 1

As a rule of thumb, every time you want to compare **ROC AUC** vs **F1 Score**, think about it as if you are comparing your model performance based on:

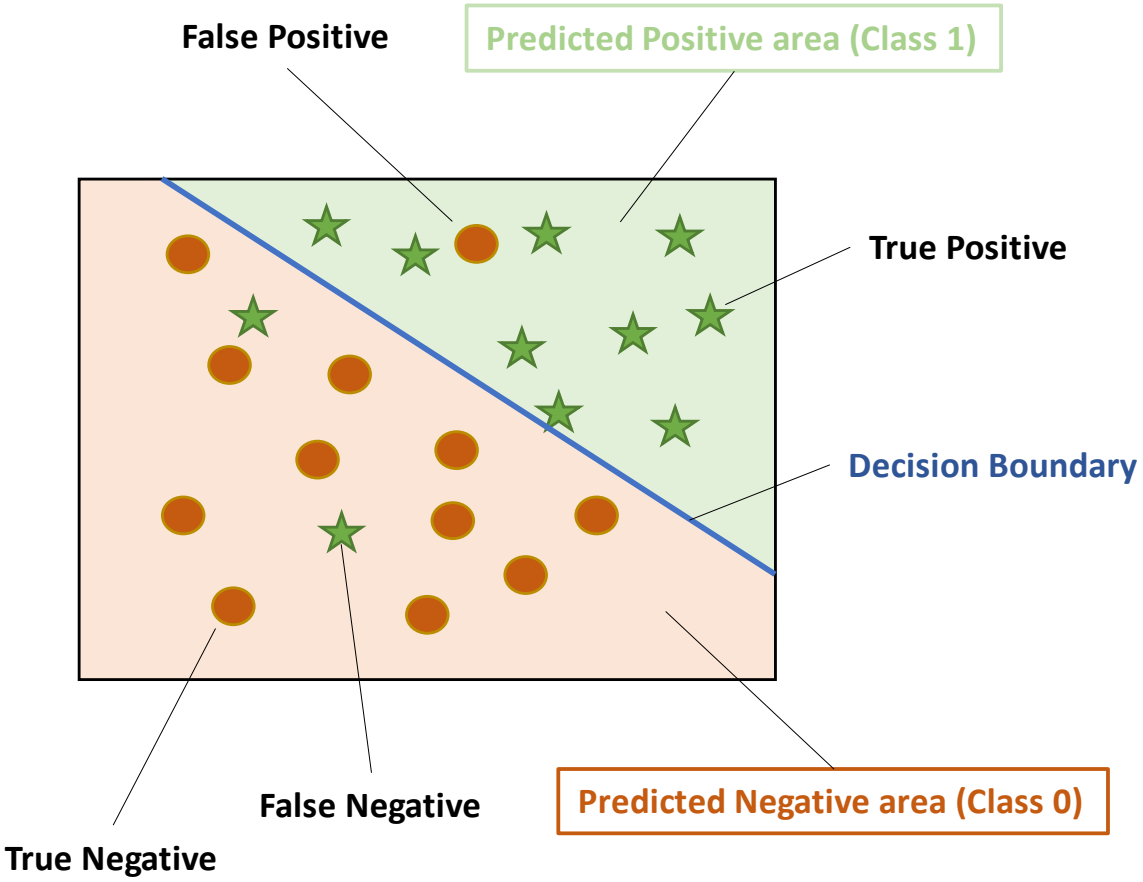
[Sensitivity vs (1-Specificity)] VS [Precision vs Recall]
or
[TPR vs FPR] VS [Precision vs TPR]

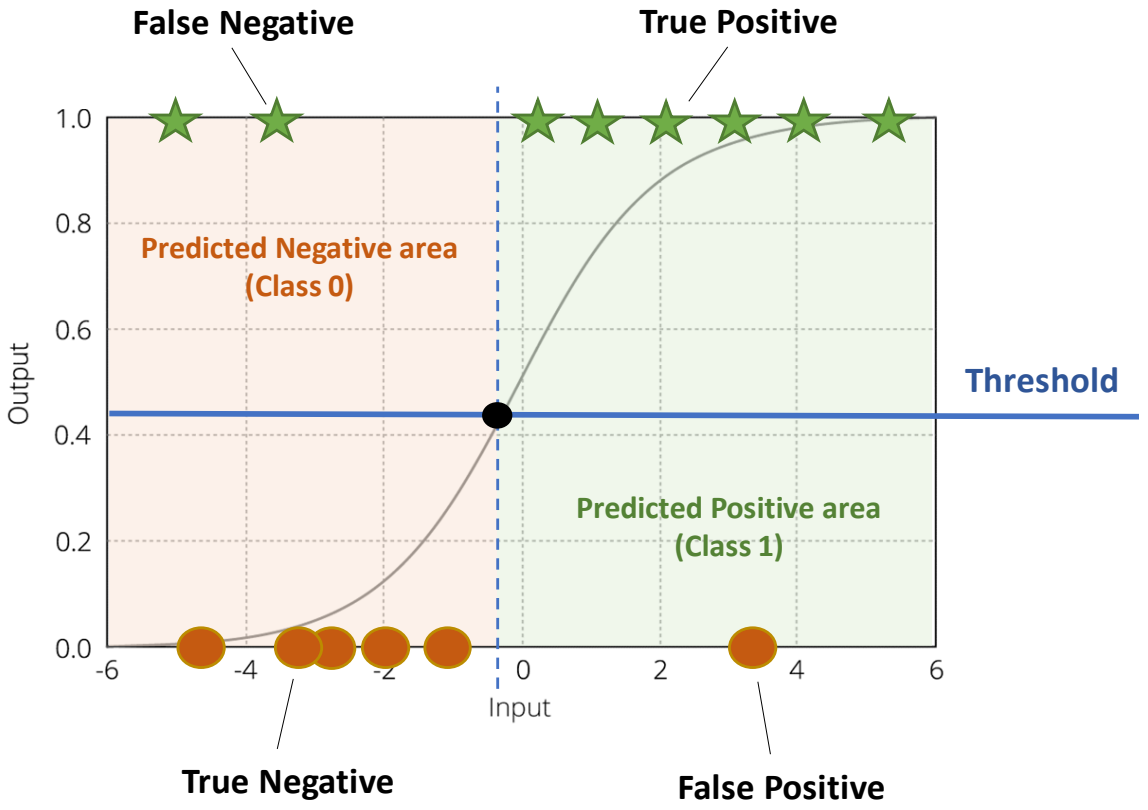
In general, the ROC is for many different levels of thresholds and thus it has many F score values. F1 score is applicable for any particular point on the ROC curve. You may think of it as a measure of precision and recall at a particular threshold value whereas AUC is the area under the ROC curve.
Consequently, when you have a data **imbalance** between positive and negative samples, you should always use F1-score because ROC **averages** over all possible thresholds!

If you look at the definitions, you can that both AUC and F1-score optimize "something" together with the fraction of the sample labeled "positive" that is actually true positive. This "something" is:

•For the AUC, the specificity, which is the fraction of the negatively labeled sample that is correctly labeled. You're not looking at the fraction of your positively labeled samples that is correctly labeled.
•Using the F1 score, it's precision: the fraction of the positively labeled sample that is correctly labeled. And using the F1-score you don't consider the purity of the sample labeled as negative (the specificity).

The difference becomes important when you have highly unbalanced or skewed classes: For example there are many more true negatives than true positives. Suppose you are looking at data from the general population to find people with a rare disease. There are far more people "negative" than "positive", and trying to optimize how well you are doing on the positive and the negative samples simultaneously, using AUC, is not optimal. You want the positive sample to include all positives if possible and you don't want it to be huge, due to a high false positive rate. So in this case you use the F1 score.





For a linear classifier (e.g. simple Logistic Regression) the decision boundary is linear.

In the real world scenario is almost always impossible to perfectly separate the data with the line. Thus, we will always have either False Positives, or False Negatives or both.

By adjusting the threshold, we will shift decision boundary.

Decreasing Threshold:
- # True Positives ↑
- # False Negatives ↓
- # True Negatives ↓
- # False Positives ↑

Recall (TPR) & FPR ↑
Specificity (TNR) & FNR ↓
Precision – can do both

For TPR = TP / (TP + FN) change is determined by TP/FN ratio

Middle letter defines the opposite direction of metric change depending on threshold direction:
Positive -> opposite direction from Threshold; Negative -> same direction as Threshold
Threshold ↓ (negative direction) -> TPR and FPR ↑ ; Specificity (TNR) & FNR ↓
Threshold ↑ (positive direction) -> TPR and FPR ↓ ; Specificity (TNR) & FNR ↑