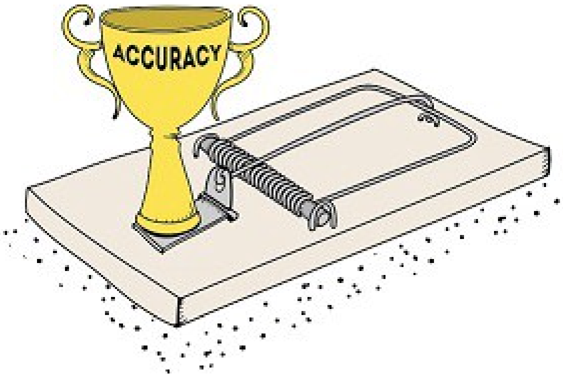
Accuracy Trap! Pay Attention to Recall, Precision, F-Score, AUC

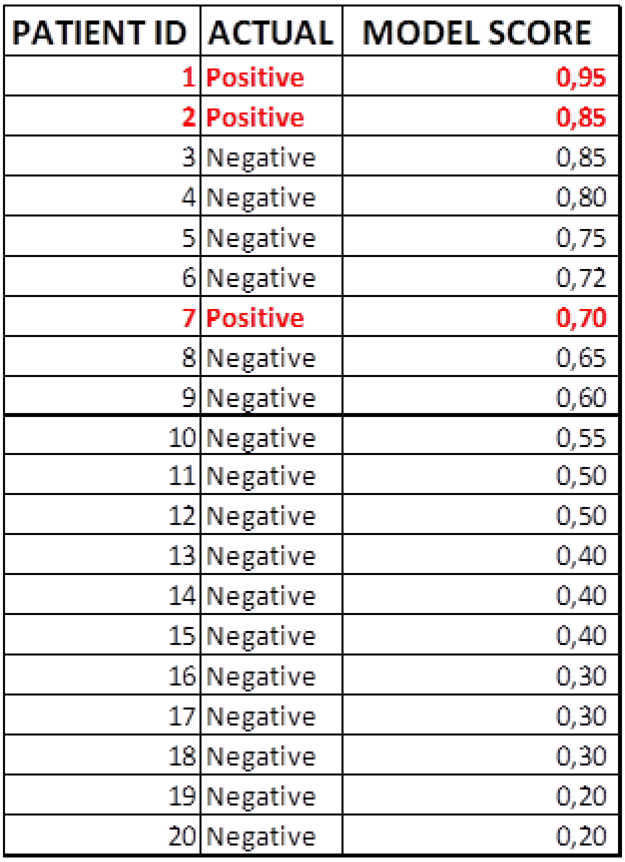
Haydar Özler Follow
Feb 25 · 6 min read

The article contains examples to explain accuracy, recall, precision, f-score, AUC concepts.



Assume you are working on a machine learning model to predict whether the person is HPV positive or not.

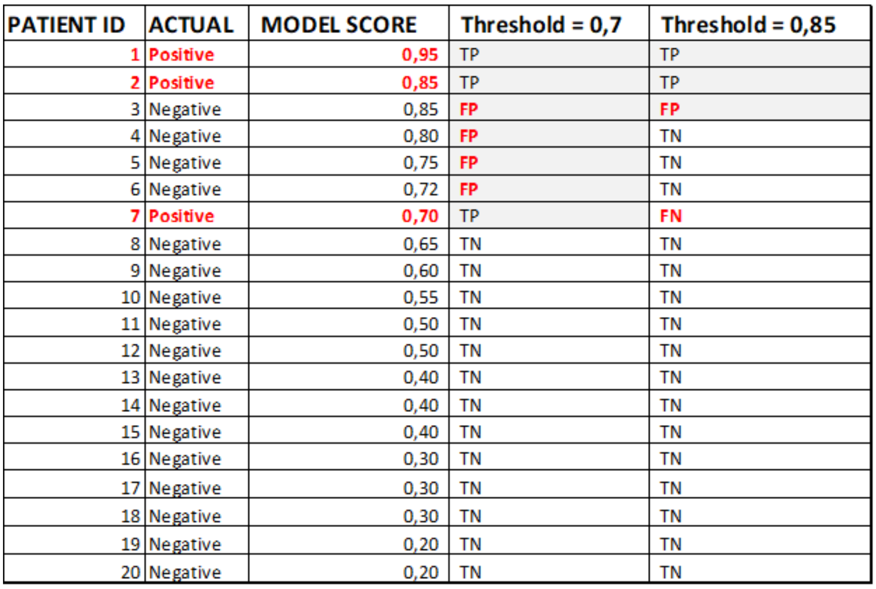
Test set is composed of 20 patients and 3 of them are positive (infected). Table-1 shows their actual status and the prediction score of the model.



PATIENT ID	ACTUAL	MODEL SCORE
1	Positive	0,95
2	Positive	0,85
3	Negative	0,85
4	Negative	0,80
5	Negative	0,75
6	Negative	0,72
7	Positive	0,70
8	Negative	0,65
9	Negative	0,60
10	Negative	0,55
11	Negative	0,50
12	Negative	0,50
13	Negative	0,40
14	Negative	0,40
15	Negative	0,40
16	Negative	0,30
17	Negative	0,30
18	Negative	0,30
19	Negative	0,20
20	Negative	0,20

Table-1 Test Set with Actuals and Prediction Scores

Before going live, you have to choose the threshold. Table-2 has two columns for threshold alternatives. These columns have true positive, true negative, false positive and false negative rows for the selected threshold values.



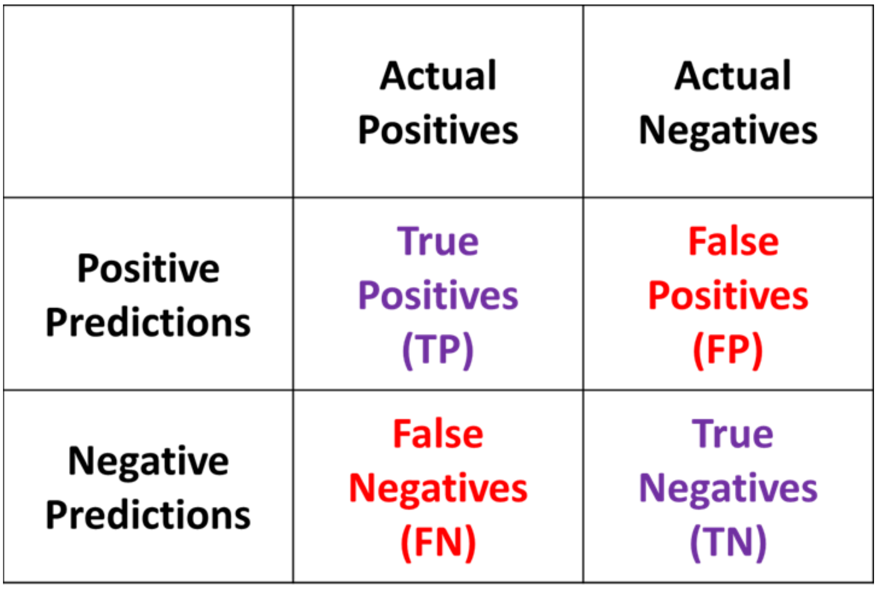
PATIENT ID	ACTUAL	MODEL SCORE	Threshold = 0,7	Threshold = 0,85
1	Positive	0,95	TP	TP
2	Positive	0,85	TP	TP
3	Negative	0,85	FP	FP
4	Negative	0,80	FP	TN
5	Negative	0,75	FP	TN
6	Negative	0,72	FP	TN
7	Positive	0,70	TP	FN
8	Negative	0,65	TN	TN
9	Negative	0,60	TN	TN
10	Negative	0,55	TN	TN
11	Negative	0,50	TN	TN
12	Negative	0,50	TN	TN
13	Negative	0,40	TN	TN
14	Negative	0,40	TN	TN
15	Negative	0,40	TN	TN
16	Negative	0,30	TN	TN
17	Negative	0,30	TN	TN
18	Negative	0,30	TN	TN
19	Negative	0,20	TN	TN
20	Negative	0,20	TN	TN

When you choose threshold = 0,7: 7 of 20 test result will be predicted as positive and these patients should take some other tests and 13 of 20 will be predicted as negative so they can leave hospital happy :). Accuracy is 0,80.

When you choose threshold = 0,85: 3 of 20 test result will be predicted as positive and these patients should take some other tests and 17 of 20 will be predicted as negative so they can leave hospital happy :). Accuracy is 0,90.

As a result is it possible to assume that the threshold should be 0,85 because the accuracy is higher? That would definitely be a mistake. Imagine an illness which affects 1 in 10.000 people. If our predictive model tells everybody is healthy, it is 99,99% accurate. Is it a good model? Absolutely no.

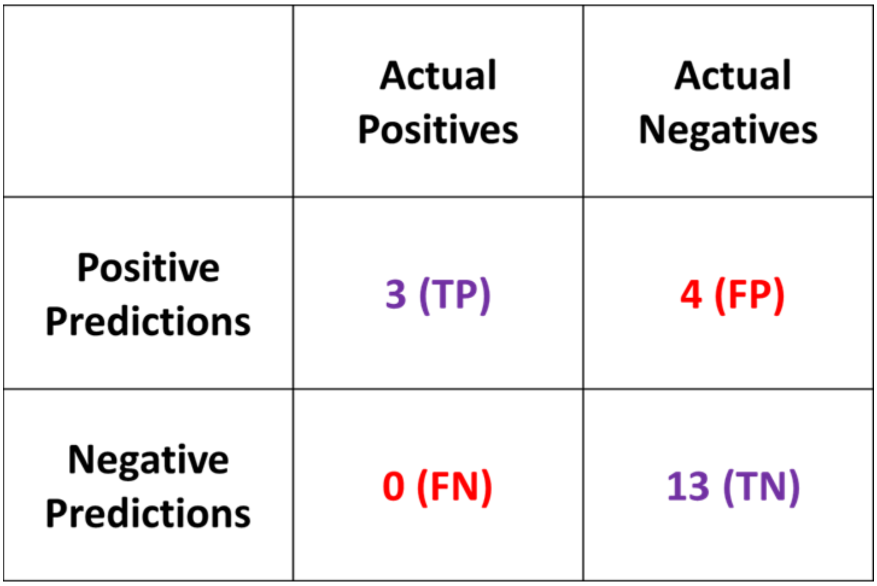
Lets check other concepts to make a better decision. Table-3 is an explanation for confusion matrix.



	Actual Positives	Actual Negatives
Positive Predictions	True Positives (TP)	False Positives (FP)
Negative Predictions	False Negatives (FN)	True Negatives (TN)

Table-3 Confusion Matrix Explained

Table-4 is confusion matrix for threshold = 0,7.



	Actual Positives	Actual Negatives
Positive Predictions	3 (TP)	4 (FP)
Negative Predictions	0 (FN)	13 (TN)

Table-4 Confusion Matrix for Threshold = 0,7

Lets make our calculations for threshold = 0,7:

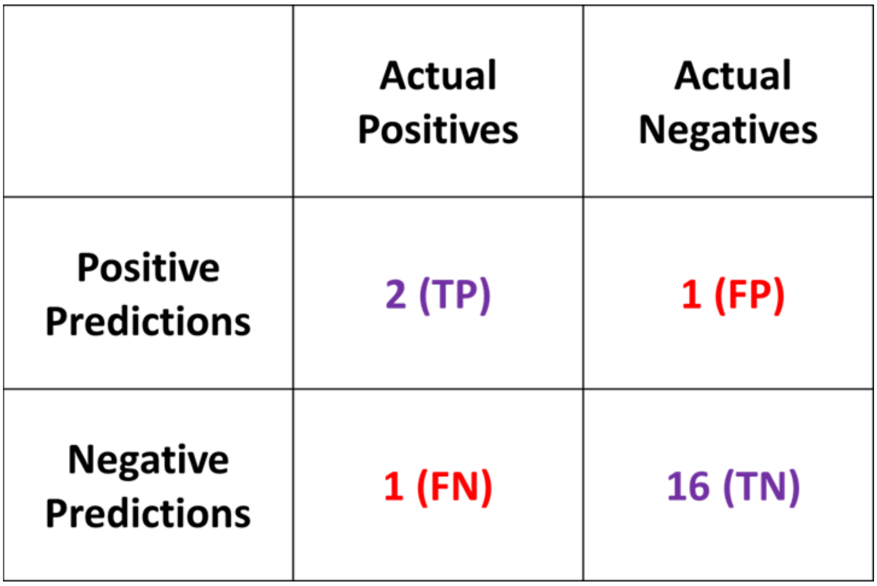
$$\text{Accuracy} = \frac{(TP+TN)}{(TP+TN+FP+FN)} = \frac{(3+13)}{(3+13+4+0)} = 0{,}80$$

$$\text{Recall} = \frac{TP}{(TP+FN)} = \frac{3}{(3+0)} = 1{,}00$$

$$\text{Precision} = \frac{TP}{(TP+FP)} = \frac{3}{(3+4)} = 0{,}43$$

$$\text{F-Score} = \frac{2*Recall*Precision}{(Recall+Precision)} = \frac{2*1*0{,}43}{(1+0{,}43)} = 0{,}60$$

Before going into details let’s make our calculations for threshold = 0,85. Thereupon these concepts will be reviewed by comparing two thresholds. Table-5 is confusion matrix for threshold = 0,85.



	Actual Positives	Actual Negatives
Positive Predictions	2 (TP)	1 (FP)
Negative Predictions	1 (FN)	16 (TN)

Table-5 Confusion Matrix for Threshold = 0,85

$$\text{Accuracy} = \frac{(TP+TN)}{(TP+TN+FP+FN)} = \frac{(2+16)}{(2+16+1+1)} = 0{,}90$$

$$\text{Recall} = \frac{TP}{(TP+FN)} = \frac{2}{(2+1)} = 0{,}67$$

$$\text{Precision} = \frac{TP}{(TP+FP)} = \frac{2}{(2+1)} = 0{,}67$$

$$\text{F-Score} = \frac{2*Recall*Precision}{(Recall+Precision)} = \frac{2*0{,}67*0{,}67}{(0{,}67+0{,}67)} = 0{,}67$$

If we don’t choose the model with higher accuracy, which one will we use to decide? Recall? Precision?

It depends on what kind of falses you can tolerate?

- Can you tolerate “false negatives”, which means telling ill people they are healthy?
- Can you tolerate “false positives”, meaning you’d have to tell healthy people they are ill?

In the first case people can die. In the second one, people will be worried and will take some extra tests to learn that they are healthy. Therefore “false negatives” are not tolerable here.

Recall tells us the prediction accuracy among only actual positives. It means how correct our prediction is among ill people. That matters in that case. That is why we have to minimize false negatives which means we are trying to maximize recall. It can cost us lower accuracies, which is still sufficient. That is why we choose threshold = 0,7 because it has a perfect recall.

Recall is considerable but in which cases we can go for precision then?

When “false positives” can not be tolerated, precision should be favoured. A model for spam detection serves as a great example for this.

- Can you tolerate “false negatives”? Which means you mark a “spam” mail as “not spam” and person will see a spam mail in his/her inbox.
- Can you tolerate “false positives”? Which means you mark a “non-spam” mail as “spam” and person won’t see this real mail in his/her inbox.

Of course, we can’t tolerate the second one. Since precision is the performance indicator about positive predictions, in such cases we try to maximize precision by decreasing number of false positives. It would also cost us a lower accuracy but it might be worthy.

What about f-score?

It is the harmonic mean of recall and precision. There might be two comments about it:

- It is a balance between recall and precision.
- It is an alternative to accuracy.

What AUC—ROC means?

ROC (Receiver Operating Characteristics) is the curve drawn by connecting the dots of x-axis = FPR (False Positive Rates) and y-axis = TPR (True Positive Rates) for different threshold values. It means you choose different threshold values for your model and calculates TPR and FPR for them and the draw the ROC curve and calculate the area under the curve. AUC (Area Under Curve) is the area under the ROC curve. Image-1 is an example for AUC-ROC curve (Ref-1).

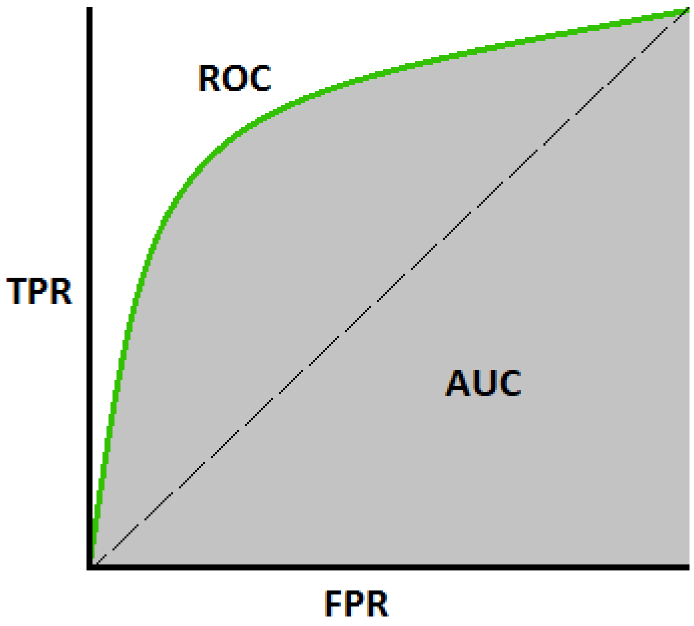


Image-1 Typical AUC-ROC Curve

Why do we need this curve?

Two main reasons are followings:

- It tells us how good our model is about seperating the two classes. For our case, classes are ill or healthy, positive or negative.
- It helps us about choosing the best threshold.

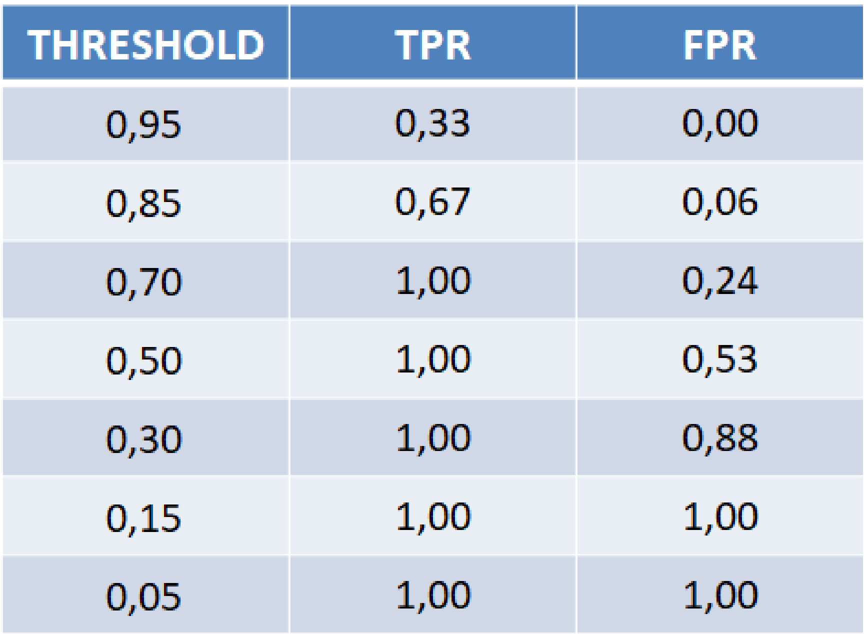
AUC = 0,5 means that your model seperates two possible outcomes randomly. AUC can be 1.0 maximum which means perfect seperation.

AUC-ROC Curve for Our Model

Here is the formula for TPR and FPR. Table-6 shows the values of TPR and FPR of different thresholds for our model. And Image-2 is the actual AUC-ROC Curve for our model.

TPR = Recall

$$\text{FPR} = \frac{FP}{(TN+FP)}$$



THRESHOLD	TPR	FPR
0,95	0,33	0,00
0,85	0,67	0,06
0,70	1,00	0,24
0,50	1,00	0,53
0,30	1,00	0,88
0,15	1,00	1,00
0,05	1,00	1,00

Table-6 TPR and FPR for different Thresholds

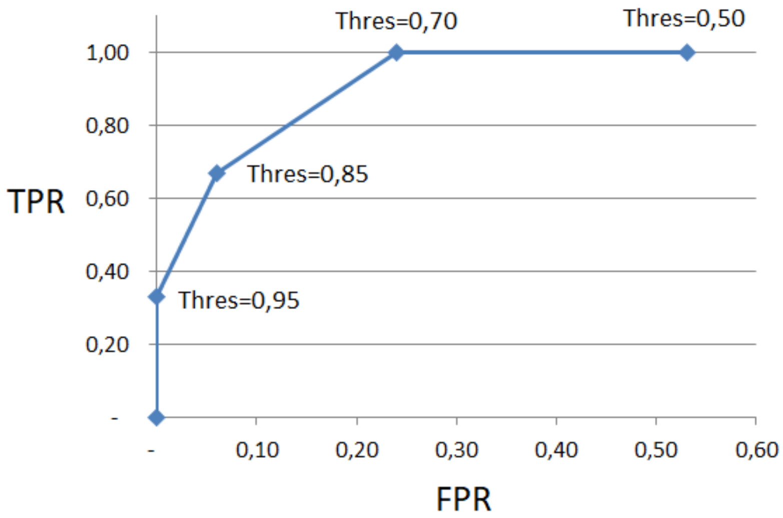


Image-2 AUC-ROC Curve for Our Model

In conclusion;

- For this model, threshold = 0,70 looks fine.
- It is wrong to evaluate your classification models with accuracy only.

- Choosing your evaluation parameters depending on the problem is important.
- Keep in mind that tests are never 100% accurate :).

If you have any further questions, please don’t hesitate to write: haydarozler@gmail.com

Ref-1: https://towardsdatascience.com/understanding-auc-roc-curve-68b2303cc9c5

. . .