Which metrics to choose?

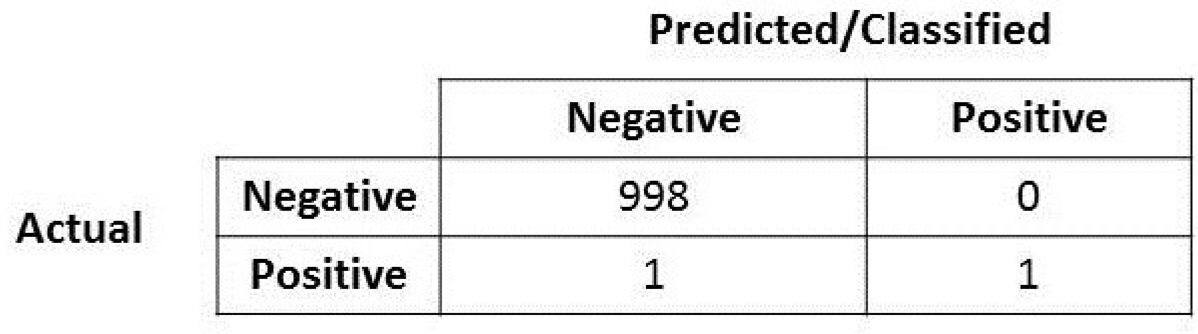
Accuracy, Precision, Recall or F1?

Koo Ping Shung Follow
Mar 15, 2018 · 5 min read

Often when I talk to organizations that are looking to implement data science into their processes, they often ask the question, "How do I get the most accurate model?". And I asked further, "What business challenge are you trying to solve using the model?" and I will get the puzzling look because the question that I posed does not really answer their question. I will then need to explain why I asked the question before we start exploring if Accuracy is the be-all and end-all model metric that we shall choose our “best” model from.

So I thought I will explain in this blog post that Accuracy need not necessary be the one-and-only model metrics data scientists chase and include simple explanation of other metrics as well.

Firstly, let us look at the following confusion matrix. What is the accuracy for the model?



		Predicted/Classified	
		Negative	Positive
Actual	Negative	998	0
	Positive	1	1

Very easily, you will notice that the accuracy for this model is very very high, at 99.9%!! Wow! You have hit the jackpot and holy grail (*scream and run around the room, pumping the fist in the air several times*)!

But….(well you know this is coming right?) what if I mentioned that the positive over here is actually someone who is sick and carrying a virus that can spread very quickly? Or the positive here represent a fraud case? Or the positive here represents terrorist that the model says its a non-terrorist? Well you get the idea. The costs of having a mis-classified actual positive (or false negative) is very high here in these three circumstances that I posed.

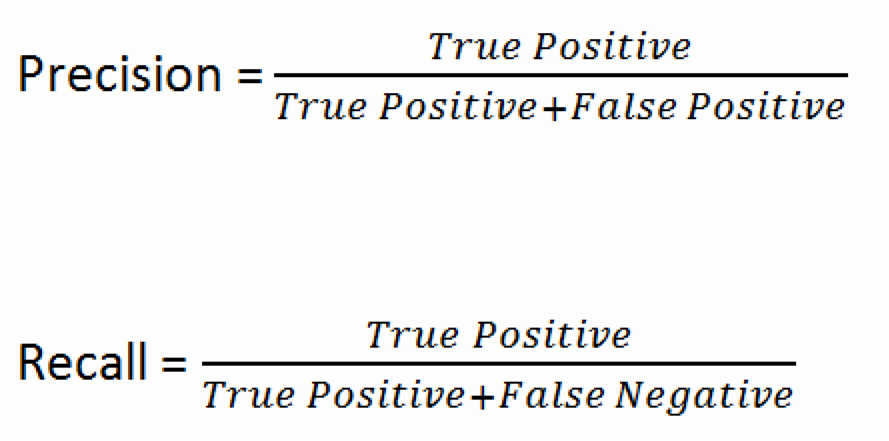
OK, so now you realized that accuracy is not the be-all and end-all model metric to use when selecting the best model…now what?

Precision and Recall

Let me introduce two new metrics (if you have not heard about it and if you do, perhaps just humor me a bit and continue reading? :D)

So if you look at Wikipedia, you will see that the the formula for calculating Precision and Recall is as follows:

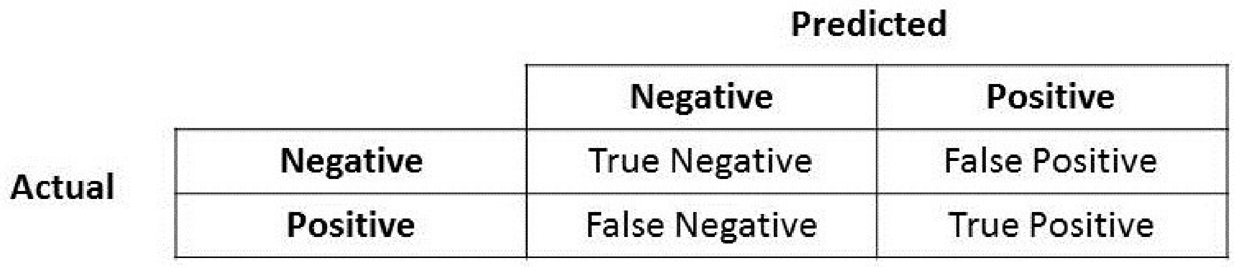
Let me put it here for further explanation.



$$\text{Precision} = \frac{True\ Positive}{True\ Positive + False\ Positive}$$

$$\text{Recall} = \frac{True\ Positive}{True\ Positive + False\ Negative}$$

Let me put in the confusion matrix and its parts here.

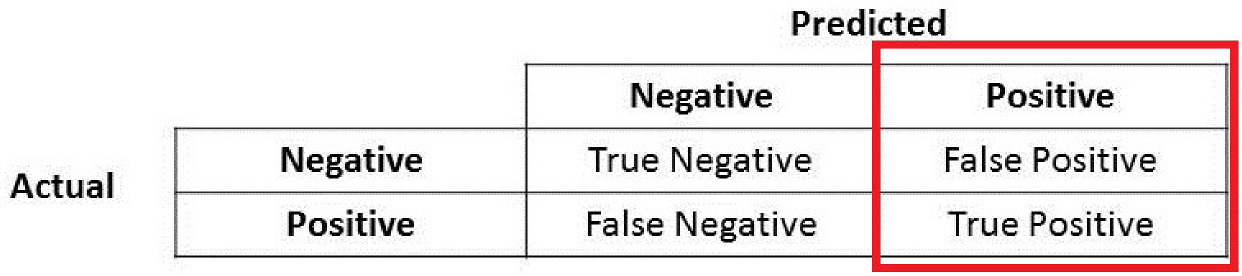


		Predicted	
		Negative	**Positive**
Actual	**Negative**	True Negative	False Positive
	Positive	False Negative	True Positive

Precision

Great! Now let us look at Precision first.

What do you notice for the denominator? The denominator is actually the Total Predicted Positive! So the formula becomes



		Predicted	
		Negative	Positive
Actual	Negative	True Negative	False Positive
	Positive	False Negative	True Positive

True Positive + False Positive = Total Predicted Positive

$$\text{Precision} = \frac{\mathit{True\ Positive}}{\mathit{True\ Positive} + \mathit{False\ Positive}}$$

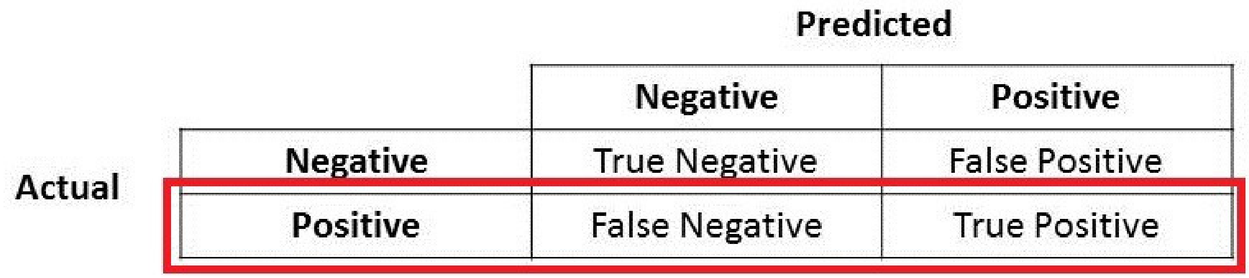
$$= \frac{\mathit{True\ Positive}}{\mathit{Total\ Predicted\ Positive}}$$

Immediately, you can see that Precision talks about how precise/accurate your model is out of those predicted positive, how many of them are actual positive.

Precision is a good measure to determine, when the costs of False Positive is high. For instance, email spam detection. In email spam detection, a false positive means that an email that is non-spam (actual negative) has been identified as spam (predicted spam). The email user might lose important emails if the precision is not high for the spam detection model.

Recall

So let us apply the same logic for Recall. Recall how Recall is calculated.



		Predicted	
		Negative	Positive
Actual	Negative	True Negative	False Positive
	Positive	False Negative	True Positive

True Positive + False Negative = Actual Positive

There you go! So Recall actually calculates how many of the Actual Positives our model capture through labeling it as Positive (True Positive). Applying the same understanding, we know that Recall shall be the model metric we use to select our best model when there is a high cost associated with False Negative.

For instance, in fraud detection or sick patient detection. If a fraudulent transaction (Actual Positive) is predicted as non-fraudulent (Predicted Negative), the consequence can be very bad for the bank.

Similarly, in sick patient detection. If a sick patient (Actual Positive) goes through the test and predicted as not sick (Predicted Negative). The cost associated with False Negative will be extremely high if the sickness is contagious.

F1 Score

Now if you read a lot of other literature on Precision and Recall, you cannot avoid the other measure, F1 which is a function of Precision and Recall. Looking at Wikipedia, the formula is as follows:

F1 Score is needed when you want to seek a balance between Precision and Recall. Right…so what is the difference between F1 Score and Accuracy then? We have previously seen that accuracy can be largely contributed by a large number of True Negatives which in most business circumstances, we do not focus on much whereas False Negative and False Positive usually has business costs (tangible & intangible) thus F1 Score might be a better measure to use if we need to seek a balance between Precision and Recall AND there is an uneven class distribution (large number of Actual Negatives).

I hope the explanation will help those starting out on Data Science and working on Classification problems, that Accuracy will not always be the metric to select the best model from.

I wish all readers a FUN Data Science learning journey and if you have liked this blog posts, do give a clap or two. Please visit my other blog posts and LinkedIn profile.

Machine Learning Data Science

$$\text{Precision} = \frac{True\ Positive}{True\ Positive + False\ Positive}$$

$$\text{Recall} = \frac{True\ Positive}{True\ Positive + False\ Negative}$$

$$= \frac{True\ Positive}{Total\ Actual\ Positive}$$

$$\mathrm{F1} = 2 \text{ x } \frac{Precision * Recall}{Precision + Recall}$$