## Averaging precision and recall

* The medical model has a precision of 55.7 % and a recall of 83.3%. It's supposed to be a high-recall model.
* The spam detector has a precision of 76.9 % and a recall of 37%. It's supposed to be a high-precision model.

Do we want to be carrying two numbers around all the time? We kind of want to only have one score. So the question is, how do we combine these two scores into one?

One answer is simply to take the average. Let's take the average of precision and recall.

* On the left, we get 69.5 percent.
* On the right, we get 56.95 percent.

And that's an okay metric, but not very different from accuracy. The way to see how this average is **not** the best idea is to try it in the extreme example.

## Credit card fraud example 1

### Precision

We have a bunch of good and fraudulent credit card transactions. Let's pick a terrible model one that says, "All transactions are good." What is the precision of this model? Well, the precision is, "out of the ones we classified as bad, how many of them are bad?" That's a question about numbers because we didn't label anything as fraudulent. So it's kind of zero divided by zero, which is undefined. But it makes sense to think of it as 100 percent since we made zero mistakes among the ones who predicted positive, which is what precision tries to measure. So let's say this model has 100 percent precision.

### Recall

The recall is, "how many of the fraudulent transactions did we catch?" Well, since we caught none, this number is zero.

### The average

The average between precision and recall is 50 percent since it's the average of 100 and zero. Now the question is, do I want to give this horrendous model of 50 percent? It seems like a really high score for such a lousy model. I kind of want to give it a much lower score, perhaps even zero.

## Credit card fraud example 2

### Precision

Now let's try the opposite. And let's try the model that says, "all transactions are fraudulent. " What is the precision of this model? Out of all the transactions that I said are fraudulent, 472 were actually fraudulent. The precision is 472 / 284,807, which is 0.16%.

### Recall

The Recall is actually pretty good because out of the 472 fraudulent transactions, I caught all of them. So the recall is 472 / 472, which is 100%.

### The average

The average of the two is the average of 0.16 and 100, which is 50.08%, a very high score for a really lousy model. So we want to give it a lower score or maybe something close to zero.

## Averaging conclusion

Averaging is **not** the greatest thing in principle if either precision or recall is very low. We want the number to be low even if the other one is high.

## Harmonic mean

There's another type of average called the harmonic mean, and it works as follows. Imagine we have two numbers, X and Y, with X < Y. And we have their arithmetic mean, which is the average, (X+Y)/2. And we have something called the harmonic mean which is defined by

2*xy / x*+*y*

​

It's kind of an average of the two in the sense that if the two numbers are equal, we get X or Y, and it always lies between X and Y.

It's a mathematical fact that the harmonic mean is always less than the arithmetic mean.

So it's closer to the smaller number than to the higher number.

### F1 score

If the precision is one and the recall is zero, the average is 0.5, but the harmonic mean is, if we plug in the formula, zero. Another example, if the precision is 0.2 and the recall is 0.8, then the arithmetic mean is 0.5, but the harmonic mean is 0.32. So it's closer to the lower number.

From now on, we will not be using the average or arithmetic mean, but we'll be using the harmonic mean, and that's going to be called the **F1 score**, which is closer to the smallest value between precision and recall. If one of them is particularly low, the F1 score can raise a flag.

*Fβ* = (1+*β*2) \* Precision⋅Recall / (*β*2 \* Precision)+Recall

​

You can see that the *β*(beta) parameter controls the degree to which precision is weighed into the F score, which allows precision and recall to be considered simultaneously. The most common value for beta is 1, as this is where you are finding the harmonic average between precision and recall.

### **Quiz Question**

Which of the following is not true about an F1 score?

1. It is the harmonic mean of the dataset
2. It will always be closer to the smaller result of precision and recall values
3. It will always be closer to the larger result of precision and recall values

**Credit card fraud example**

Go back to the credit card fraud example and calculate the F1 score. Since it's going to be the harmonic mean between the precision, which is 100%, and the recall, which is 0, we can plug in the formula and actually get an F1 score of zero. This is much closer to what the models should score.

So, in the exercise below, we'll let you calculate the F1 score of the medical model and the spam email model.

## F1 Score Quiz

For the following, remember that the formula for F1 Score is:

F1 Score = 2 ⋅ Precision ∗ Recall / (Precision+Recall)

### **Precision Quiz**

If the Precision of the medical model is **55.6%**, and the Recall is **83.3%**, what is the F1 Score? (Please write your answer as a percentage, and round it to 1 decimal point.)

Answer: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

Precision (P) = 55.6% → 0.556

Recall (R) = 83.3% → 0.833

Final Answer:

66.6

​