Classification: Precision and Recall

Source: Google Developers

Precision

Precision attempts to answer the following question:

> What proportion of positive identifications was actually correct?

Precision is defined as follows:

$$\text{Precision} = \frac{TP}{TP+FP}$$

Note: A model that produces no false positives has a precision of 1.0.

True Positives (TPs): 1	**False Positives (FPs): 1**
False Negatives (FNs): 8	**True Negatives (TNs): 90**

$$\text{Precision} = \frac{TP}{TP+FP} = \frac{1}{1+1} = 0.5$$

Our model has a precision of 0.5—in other words, when it predicts a tumor is malignant, it is correct 50% of the time.

Recall

Recall attempts to answer the following question:

> What proportion of actual positives was identified correctly?

Mathematically, recall is defined as follows:

$$\text{Recall} = \frac{TP}{TP + FN}$$

Note: A model that produces no false negatives has a recall of 1.0.

Let's calculate recall for our tumor classifier:

True Positives (TPs): 1	**False Positives (FPs): 1**
False Negatives (FNs): 8	**True Negatives (TNs): 90**

$$\text{Recall} = \frac{TP}{TP + FN} = \frac{1}{1 + 8} = 0.11$$

Our model has a recall of 0.11—in other words, it correctly identifies 11% of all malignant tumors.

Precision and Recall: A Tug of War

To fully evaluate the effectiveness of a model, you must examine **both** precision and recall. Unfortunately, precision and recall are often in tension. That is, improving precision typically reduces recall and vice versa. Explore this notion by looking at the following figure, which shows 30 predictions made by an email classification model. Those to the right of the classification threshold are classified as "spam", while those to the left are classified as "not spam."

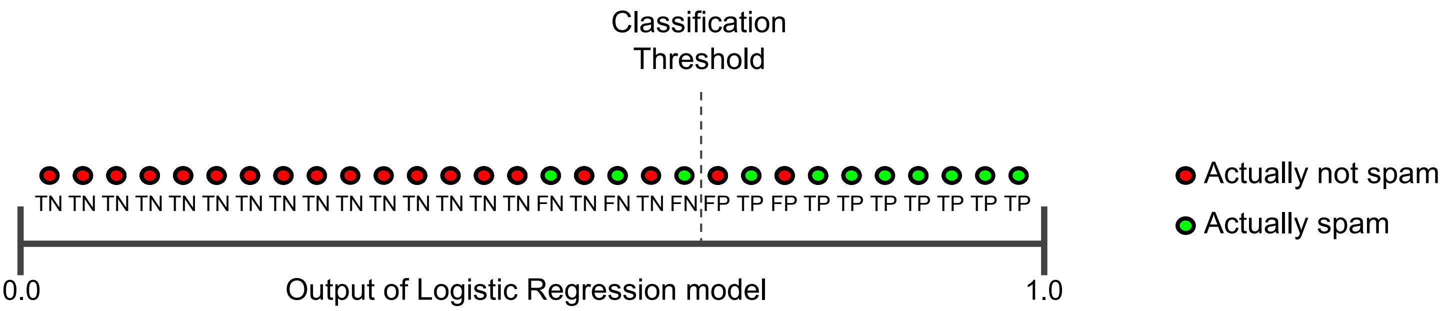
Classification Threshold
TN TN TN TN TN TN TN TN TN TN TN TN TN TN TN FN TN FN TN FN FP TP FP TP TP TP TP TP TP TP
0.0
Output of Logistic Regression model
1.0
Actually not spam
Actually spam

Figure 1. Classifying email messages as spam or not spam.

Let's calculate precision and recall based on the results shown in Figure 1:

True Positives (TP): 8	**False Positives (FP): 2**
False Negatives (FN): 3	**True Negatives (TN): 17**

Precision measures the percentage of **emails flagged as spam** that were correctly classified—that is, the percentage of dots to the right of the threshold line that are green in Figure 1:

$$\text{Precision} = \frac{TP}{TP+FP} = \frac{8}{8+2} = 0.8$$

Recall measures the percentage of **actual spam emails** that were correctly classified—that is, the percentage of green dots that are to the right of the threshold line in Figure 1:

$$\text{Recall} = \frac{TP}{TP+FN} = \frac{8}{8+3} = 0.73$$

Figure 2 illustrates the effect of increasing the classification threshold.

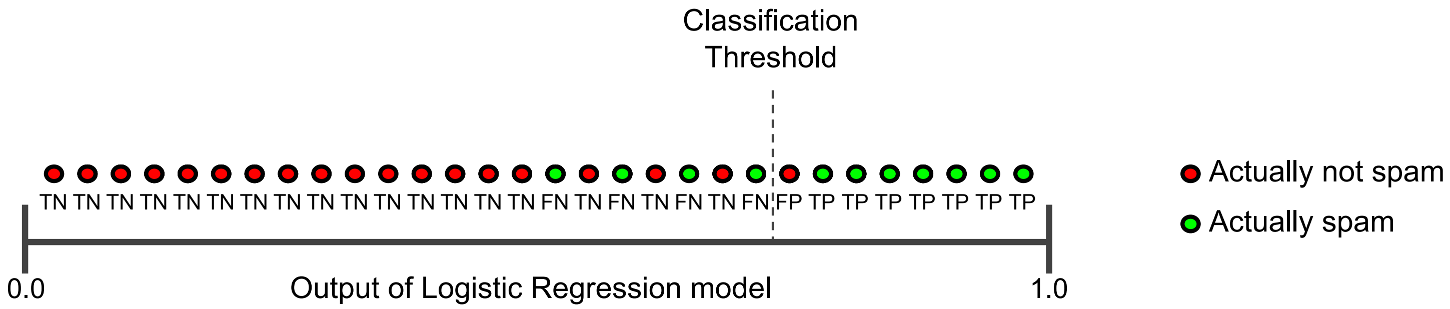
Classification Threshold
TN TN TN TN TN TN TN TN TN TN TN TN TN TN TN FN TN FN TN FN TN FN FP TP TP TP TP TP TP TP
0.0
Output of Logistic Regression model
1.0
Actually not spam
Actually spam

Figure 2. Increasing classification threshold.

The number of false positives decreases, but false negatives increase. As a result, precision increases, while recall decreases:

True Positives (TP): 7	**False Positives (FP):** 1
False Negatives (FN): 4	**True Negatives (TN):** 18

$$\text{Precision} = \frac{TP}{TP+FP} = \frac{7}{7+1} = 0.88$$

$$\text{Recall} = \frac{TP}{TP+FN} = \frac{7}{7+4} = 0.64$$

Conversely, Figure 3 illustrates the effect of decreasing the classification threshold (from its original position in Figure 1).

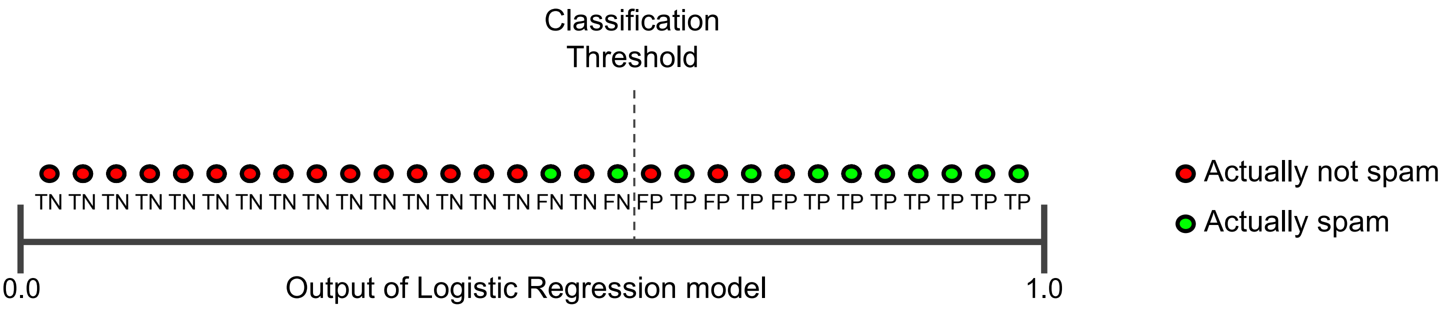
Classification Threshold
TN TN TN TN TN TN TN TN TN TN TN TN TN TN TN FN TN FN FP TP FP TP FP TP TP TP TP TP TP TP
0.0
Output of Logistic Regression model
1.0
Actually not spam
Actually spam

Figure 3. Decreasing classification threshold.

False positives increase, and false negatives decrease. As a result, this time, precision decreases and recall increases:

True Positives (TP): 9	**False Positives (FP): 3**
False Negatives (FN): 2	**True Negatives (TN): 16**

$$\text{Precision} = \frac{TP}{TP + FP} = \frac{9}{9+3} = 0.75$$

$$\text{Recall} = \frac{TP}{TP + FN} = \frac{9}{9+2} = 0.82$$

Various metrics have been developed that rely on both precision and recall. For example, see F1 score (https://wikipedia.org/wiki/F1_score).