Which metric should be used: F1 Score or Accuracy?

Authored by

stats writer

May 7, 2024

The decision to use either F1 Score or Accuracy as a metric depends on the specific goals and requirements of the task at hand. F1 Score is a measure of the overall accuracy of a classification model, taking into account both precision and recall. On the other hand, Accuracy simply measures the percentage of correctly classified instances. F1 Score is particularly useful in situations where there is a class imbalance, as it takes into account both false positives and false negatives. However, if the goal is to simply maximize the number of correctly classified instances, then Accuracy may be a more appropriate metric. Ultimately, the choice between F1 Score and Accuracy should be based on the specific objectives and needs of the project.

F1 Score vs. Accuracy: Which Should You Use?

When using in machine learning, two metrics we often use to assess the quality of the model are F1 Score and Accuracy.

For both metrics, the higher the value the better a model is able to classify observations into classes.

However, each metric is calculated using a different formula and there are pros and cons to using each.

The following example shows how to calculate each metric in practice.

Example: Calculating F1 Score & Accuracy

Suppose we use a logistic regression model to predict whether or not 400 different college basketball players get drafted into the NBA.

The following confusion matrix summarizes the predictions made by the model:

		Predicted	
		Drafted = Yes	**Drafted = No**
Actual	**Drafted = Yes**	**120** (True Positive)	**40** (False Negative)
	Drafted = No	**70** (False positive)	**170** (True Negative)

Here is how to calculate various metrics for the confusion matrix:

Precision: Correct positive predictions relative to total positive predictions

Precision = True Positive / (True Positive + False Positive)Precision = 120 / (120 + 70)Precision = 0.63

Recall: Correct positive predictions relative to total actual positives

Recall = True Positive / (True Positive + False Negative)Recall = 120 / (120 + 40)Recall = 0.75

Accuracy: Percentage of all correctly classified observations

Accuracy = (True Positive + True Negative) / (Total Sample Size)Accuracy = (120 + 170) / (400)Accuracy = 0.725

F1 Score: Harmonic mean of precision and recall

F1 Score = 2 * (Precision * Recall) / (Precision + Recall)F1 Score = 2 * (0.63 * 0.75) / (0.63 + 0.75)F1 Score = 0.685

When to Use F1 Score vs. Accuracy

There are pros and cons to using F1 score and accuracy.

Accuracy:

Pro: Easy to interpret. If we say that a model is 90% accurate, we know that it correctly classified 90% of observations.

Con: Does not take into account how the data is distributed. For example, suppose 90% of all players do not get drafted into the NBA. If we have a model that simply predicts every player to not get drafted, the model would correctly predict the outcome for 90% of

the players. This value seems high, but the model is actually unable to correctly predict any player who gets drafted.

F1 Score:

Pro: Takes into account how the data is distributed. For example, if the data is highly imbalanced (e.g. 90% of all players do not get drafted and 10% do get drafted) then F1 score will provide a better assessment of model performance.

Con: Harder to interpret. The F1 score is a blend of the precision and recall of the model, which makes it a bit harder to interpret.

As a rule of thumb:

We often use accuracy when the classes are balanced and there is no major downside to predicting false negatives.

We often use F1 score when the classes are imbalanced and there is a serious downside to predicting false negatives.

For example, if we use a logistic regression model to predict whether or not someone has cancer, false negatives are really bad (e.g. predicting that someone does not have cancer when they actually do) so F1 score will penalize models that have too many false negatives more than accuracy will.