How to Interpret and Improve Your F1 Score

Authored by

stats writer

December 3, 2025

The F1 score is widely recognized as a crucial metric for evaluating a classification model's performance and accuracy, particularly in scenarios involving imbalanced datasets. It serves as a measure that intelligently balances the often-conflicting needs of precision and recall. Mathematically, the F1 score is calculated as the harmonic mean of these two metrics. A perfect F1 score is 1.0, indicating flawless performance where the model achieves both perfect precision and perfect recall. Conversely, a score of 0.0 signifies that the model is entirely ineffective at classifying observations correctly.

The utility of the F1 score lies in its comprehensive nature. While simple accuracy metrics can be misleading when true negative cases vastly outnumber true positive cases, the F1 score provides a single, reliable value that penalizes models failing to achieve a balance between avoiding false positives (precision) and avoiding false negatives (recall). This balance is essential across various fields, from medical diagnostics to fraud detection, where the cost of different types of errors can vary dramatically.

The F1 Score: A Critical Metric for Model Evaluation

When developing and assessing models in machine learning, selecting the appropriate evaluation metric is paramount. While overall accuracy seems intuitive, it often fails when dealing with datasets where the classes are unevenly distributed, a situation known as class imbalance. This is precisely why the **F1 Score** is frequently adopted as the standard measure of performance. The F1 Score acts as the harmonic mean of precision and recall, providing a balanced evaluation of a classifier's ability to correctly identify positive cases while minimizing both false positives and false negatives.

The calculation of the F1 score is robust because the harmonic mean disproportionately penalizes models that achieve extremely high scores in one metric while performing poorly in the other. For instance, a model that simply predicts every observation as positive might achieve perfect recall but suffer terrible precision, and the resulting F1 score would be moderate, accurately reflecting the model's limited utility. The range of the F1 score is always between 0 and 1, where 1 represents perfect classification and 0 represents complete failure. Understanding this range is the first step in determining what constitutes a "good" F1 score in a given context.

The mathematical formulation of this metric is straightforward but crucial for understanding its derivation and application. We define the F1 Score using its constituent parts, Precision and Recall, which are themselves derived from the results summarized in a confusion matrix. The formula elegantly combines these metrics into a single, comprehensive value, making model comparison simpler and more reliable than relying on either precision or recall alone.

This metric is calculated as:

F1 Score = 2 * (Precision * Recall) / (Precision + Recall)

Understanding the Components: Precision and Recall

To fully appreciate the F1 score, we must first delve into its two fundamental components: Precision and Recall. These metrics quantify different aspects of a model's predictive capabilities. Precision focuses on the quality of the positive predictions, answering the question: "Of all the cases the model predicted as positive, how many were actually correct?" High precision is desirable when the cost of a false positive (Type I error) is high, such as mistakenly flagging a healthy patient as having a disease.

Conversely, Recall (also known as sensitivity or true positive rate) addresses the completeness of the positive predictions, answering: "Of all the actual positive cases in the dataset, how many did the model correctly identify?" High recall is critical when the cost of a false negative (Type II error) is high, for example, failing to detect a critical security threat or missing a critical manufacturing defect. In many real-world scenarios, there is an inherent trade-off between maximizing precision and maximizing recall, often dictated by the chosen classification threshold.

The relationship between these two metrics is often inversely proportional. Adjusting the model parameters to increase precision often leads to a decrease in recall, and vice versa. The F1 score resolves this inherent tension by providing a single metric that seeks to maximize both simultaneously. Therefore, a high F1 score ensures that the model is performing optimally across both dimensions, neither flooding the user with false alarms (low precision) nor failing to catch significant events (low recall).

The components are defined as follows:

Precision: The ratio of correct positive predictions (True Positives) relative to the total number of positive predictions made by the model (True Positives + False Positives).
Recall: The ratio of correct positive predictions (True Positives) relative to the total number of actual positive cases in the dataset (True Positives + False Negatives).

The Role of the Confusion Matrix

Before applying the F1 calculation, the model's performance must be summarized using a Confusion Matrix. This matrix is a fundamental tool in classification analysis, providing a complete breakdown of the predictions made by the model versus the actual outcomes. It allows us to distinguish between four crucial output states: True Positives (TP), True Negatives (TN), False Positives (FP), and False Negatives (FN).

A **True Positive** occurs when the model correctly predicts a positive outcome, while a **True**

Negative occurs when the model correctly predicts a negative outcome. The errors are represented by the 'False' categories: A **False Positive** is an instance where the model incorrectly predicts a positive outcome (Type I error), and a **False Negative** is an instance where the model incorrectly predicts a negative outcome when the actual result was positive (Type II error). All core evaluation metrics, including accuracy, precision, and recall, are derived directly from the counts within this matrix.

For example, suppose we are utilizing a logistic regression model to predict the outcome of a binary classification task--specifically, whether or not 400 different college basketball players will be drafted into the NBA. The Confusion Matrix is essential for visualizing how our model distributes its 400 predictions across the four possible outcome categories. It provides the necessary counts to calculate the precision and recall values that feed into the F1 score formula, allowing us to quantify the model's predictive power accurately.

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

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

Calculating the F1 Score: A Detailed Example

Using the data derived from the confusion matrix above, we can proceed with the step-by-step calculation of the F1 score. This process clearly illustrates how the raw counts translate into the final, aggregated performance metric. We begin by calculating the precision, focusing on the quality of the positive predictions (Drafted, in this case). According to the matrix, we have 120 True Positives (TP) and 70 False Positives (FP).

The calculation for Precision is as follows:

Precision = True Positive / (True Positive + False Positive) = 120 / (120 + 70) = **0.63157**

Next, we calculate Recall, focusing on the completeness of the predictions--how many of the actual drafted players did we correctly identify? We have 120 True Positives (TP) and 40 False Negatives (FN), meaning 40 players who were drafted were missed by the model. This value reflects the model's ability to cover all positive instances.

The calculation for Recall is as follows:

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

Finally, we combine these two intermediate results using the harmonic mean formula to derive the **F1 Score**. As shown, the F1 score balances the moderate precision (0.63157) and the stronger recall (0.75), yielding a unified performance metric. This result provides a clear, single figure that we can use to compare this model against other candidate models or established benchmarks.

The final calculation for the F1 Score is:

F1 Score = 2 * (0.63157 * 0.75) / (0.63157 + 0.75) = **0.6857**

Interpreting F1 Scores: What Defines "Good"?

The immediate question posed by anyone encountering this metric is: "What constitutes a 'good' F1 score?" In the simplest terms, higher scores are unequivocally better. Since the range is strictly bounded between 0 and 1, a score approaching 1.0 indicates superior classification performance. An F1 score of exactly 1.0 suggests a model that perfectly classifies every single observation, achieving 100% precision and 100% recall. Conversely, a score of 0.0 means the model is incapable of making any correct positive predictions.

However, interpreting what is "good" must always be contextualized. Unlike academic exercises where perfect scores are sometimes achievable, real-world data is noisy, complex, and often ambiguous. For highly sensitive tasks, such as rare disease detection or critical infrastructure failure prediction, even an F1 score of 0.95 might be deemed inadequate if the remaining 5% failure rate represents unacceptable risk. In contrast, for high-volume, low-stakes classification tasks, an F1 score of 0.75 might be perfectly acceptable, especially if the current industry standard or existing alternative methods perform significantly worse.

To illustrate the concept of a perfect score, consider a scenario where a second model is developed for the basketball drafting prediction, and it manages to achieve perfect separation of the classes. This ideal result is summarized in the following confusion matrix, demonstrating 240 True Positives, 160 True Negatives, and zero errors (False Positives and False Negatives).

		Predicted	
		Drafted = Yes	**Drafted = No**
Actual	**Drafted = Yes**	**240** (True Positive)	**0** (False Negative)
	Drafted = No	**0** (False positive)	**160** (True Negative)

The calculations for this ideal model clearly demonstrate why 1.0 is the upper bound:

Precision = True Positive / (True Positive + False Positive) = 240 / (240 + 0) = **1**

Recall = True Positive / (True Positive + False Negative) = 240 / (240 + 0) = **1**

F1 Score = 2 * (1 * 1) / (1 + 1) = **1**

This result confirms that an F1 score is equal to one because the model is able to perfectly classify each of the 400 observations, demonstrating both perfect precision and perfect recall. In practice, achieving this score is rare, and evaluation usually involves comparing relative scores rather than striving for theoretical perfection.

The Importance of Baseline Models

A crucial step in assessing the practical utility of any trained model is comparing its performance against a **baseline model**. A baseline model is typically a simple, often non-machine learning, predictor that establishes the minimum acceptable level of performance. This comparison allows us to ascertain whether the complexity introduced by our sophisticated classification algorithm actually yields meaningful improvement over trivial methods.

Consider a simple, yet necessary, baseline model for our basketball drafting prediction: a model that predicts every single player will be drafted, regardless of their statistics or attributes. While seemingly useless, this model provides an essential benchmark because it represents the performance floor. If our complex model does not outperform this naive baseline, then the complex model offers no real predictive value, despite its training complexity.

The results of this specific, highly biased baseline model are summarized in the following confusion matrix. Notice that since it always predicts positive, there are zero False Negatives (FN=0), but the number of False Positives (FP=240) is maximal:

		Predicted	
		Drafted = Yes	**Drafted = No**
Actual	**Drafted = Yes**	**160** (True Positive)	**0** (False Negative)
	Drafted = No	**240** (False positive)	**0** (True Negative)

We calculate the F1 score for this baseline model:

Precision = True Positive / (True Positive + False Positive) = 160 / (160 + 240) = **0.4**

Recall = True Positive / (True Positive + False Negative) = 160 / (160 + 0) = **1**

F1 Score = 2 * (0.4 * 1) / (0.4 + 1) = **0.5714**

The resulting F1 score of 0.5714 serves as our minimum threshold. Recall that our original, more complex model had an F1 score of **0.6857**. While 0.6857 is indeed higher than 0.5714, the margin of improvement is relatively small. This indicates that our complex model is more useful than the naive baseline, but the incremental gain suggests that significant optimization or feature engineering is still needed.

A Structured Approach to Model Selection

Since there is no specific universal value considered a "good" F1 score, model selection is inherently a comparative process. The utility of the metric is maximized when it is used to rank and select the best candidate from a pool of competing models. This systematic approach ensures that the final choice is robust, validated against simple benchmarks, and optimized for the specific classification goal.

In practice, data scientists typically follow a multi-step process for selecting the optimal classification model:

Fit a Baseline Model: Start by establishing the floor of performance. Fit a simple, non-learning model (e.g., a model that predicts the majority class or a random predictor) and calculate its F1 score. This score establishes the minimum performance bar that any viable model must exceed.
Fit and Evaluate Candidate Models: Train several different classification algorithms (such as support vector machines, random forests, or neural networks) on the dataset. For each candidate model, rigorously calculate the F1 score, ensuring consistency in the evaluation dataset (e.g., using a holdout test set or cross-validation).
Select the Best Performer: Choose the model that yields the highest **F1 Score** among all

candidates. Crucially, verify that this selected model produces a higher F1 score than the established baseline model. If the difference is negligible, the added complexity of the sophisticated model may not be justified.

By using this comparative methodology centered on the F1 score, we move beyond subjective interpretations of "goodness" toward an objective, data-driven selection process. The goal is not just a high F1 score in isolation, but the highest achievable F1 score relative to alternative models and, most importantly, relative to the specific context and requirements of the application.