Understanding F1 Score and Accuracy: Choosing the Right Evaluation Metric for Classification Models

Authored by

Mohammed looti

November 2, 2025

The Dilemma of Model Evaluation in Classification

When developing predictive models in machine learning, particularly those designated for classification tasks, the selection of an appropriate evaluation metric is perhaps the most critical decision. Two metrics dominate the discussion surrounding model assessment: the **F1 Score** and **Accuracy**. Data scientists rely on these measures to quantify the quality and predictive power of an algorithm.

At a superficial level, both metrics aim to provide a single, quantifiable measure of model performance, with the intuitive goal that a higher value signifies superior function. However, despite their shared purpose, the underlying calculations and the contexts in which they are most effective diverge significantly. Understanding this divergence is essential for making informed decisions in real-world applications.

A thorough understanding of the specific calculation and the inherent biases of each metric is paramount to avoiding misleading results. While **Accuracy** offers a simple, holistic view of overall correctness across the entire dataset, the **F1 Score** provides a far more nuanced, robust measure, especially vital when analyzing complex or real-world datasets characterized by substantial class imbalance. We must look beyond mere percentages to determine which metric truly reflects the model's utility.

Deconstructing the F1 Score: Precision and Recall

To fully appreciate the distinction between the F1 Score and Accuracy, we must first examine the two fundamental metrics that form the basis of the F1 Score: Precision and Recall. These metrics quantify a model's performance by focusing specifically on the positive predictions made and the true positive cases that exist within the data. They move beyond simple overall correctness to assess the quality of the model's identification process.

Precision, often viewed as a measure of exactness or quality, answers the question: out of all the instances the model confidently predicted as positive, what proportion were genuinely positive outcomes? High precision means the model rarely makes a false positive error. Conversely, **Recall**, which measures completeness or sensitivity, addresses the question: out of all the instances that were truly positive in the dataset, how many did the model successfully manage to capture? High recall means the model avoids false negative errors.

These components are crucial because they highlight the fundamental trade-offs inherent in any classification algorithm. A model can easily achieve near-perfect precision by being highly conservative--only predicting positive outcomes when absolutely certain--but this often results in low recall, as many true positive cases are missed. The **F1 Score** is specifically engineered to harmonize this relationship, providing a single metric that rewards models only if they maintain a

healthy balance between both precision and recall. It is the harmonic mean of the two, ensuring neither metric can be maximized at the expense of the other.

Illustrating Performance: The Confusion Matrix Example

To demonstrate the practical application of these metrics and how their outcomes can diverge, let us consider a concrete classification scenario. Imagine we employ a logistic regression model tasked with predicting outcomes for 400 college basketball players--specifically, determining whether or not each player will be drafted into the NBA. This is a classic binary classification problem.

The performance of our model is systematically summarized using a **confusion matrix**. This matrix visually maps the model's predictions against the actual true outcomes, allowing us to quantify the four possible results: True Positives (TP), False Positives (FP), True Negatives (TN), and False Negatives (FN).

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

Based on the results summarized in this matrix, we have the following raw data points for our 400 observations: True Positive (TP) = 120, False Positive (FP) = 70, True Negative (TN) = 170, and False Negative (FN) = 40. Using these foundational counts, we can now proceed to calculate the derived performance metrics--Precision, Recall, Accuracy, and F1 Score--to analyze the model's effectiveness comprehensively.

Calculating Component and Aggregate Metrics

We begin the calculation process by determining the values for Precision and Recall, which are critical inputs necessary for computing the final F1 Score:

Precision: This metric calculates the proportion of correct positive predictions relative to all instances the model designated as positive.

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

Precision = **0.63**

Recall (Sensitivity): This metric calculates the proportion of correct positive predictions relative to the total number of actual positive cases present in the dataset.

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

Next, we determine the two aggregate performance metrics that are the subject of our comparison: Accuracy and the F1 Score.

Accuracy: This value represents the percentage of all observations (both positive and negative) that were correctly classified across the entire sample size.

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

F1 Score: This is the harmonic mean of precision and recall. By using the harmonic mean rather than the arithmetic mean, the F1 Score heavily penalizes models that exhibit a large discrepancy between precision and recall, ensuring a balanced assessment.

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

The Pitfall of Accuracy: Misleading Results with Imbalanced Data

The primary benefit of **Accuracy** is its inherent simplicity: it is easy to calculate, highly intuitive for non-technical audiences, and provides a quick overall snapshot of correctness. This simplicity, however, is also its greatest weakness, particularly when the underlying data distribution is skewed or uneven.

The main criticism of Accuracy is that it completely fails to account for class distribution, which makes it highly susceptible to producing misleadingly optimistic results in scenarios involving high class imbalance. Let us return to our NBA drafting example and hypothesize a severe imbalance: suppose 95% of college players are never drafted, meaning only 5% are positive cases. A naive model that simply predicts every single player will "not be drafted" (the majority class) achieves an impressive 95% accuracy.

While a 95% accuracy figure sounds outstanding, this model is functionally useless for its intended

purpose--identifying drafted players--because it has failed to identify a single true positive case. When classes are highly skewed, Accuracy masks the model's inability to perform on the minority class, providing an overly optimistic assessment that should be treated with extreme caution by data practitioners.

The Strength of F1 Score: Robustness in Imbalance

The **F1 Score** was specifically developed to overcome the limitations inherent in Accuracy, particularly when evaluating models trained on imbalanced datasets. Since the F1 Score demands high values for both Precision and Recall, it effectively penalizes models that achieve high overall accuracy by simply predicting the majority class. If a model fails to capture positive cases (low Recall) or makes too many false positive errors (low Precision), the resulting F1 Score will be low.

The primary advantage (Pro) of using the F1 Score is its robustness in assessing performance when class distribution is severely skewed. In the previous example where the naive model achieved 95% accuracy, its Recall for the critical 'drafted' class would be zero, instantly driving the F1 Score to zero. This zero result correctly signals the model's complete failure to perform the task of prediction, offering a far more honest evaluation than the 95% accuracy figure.

Conversely, the main drawback (Con) of the F1 Score is its relative complexity. Because it is a synthesized blend of two distinct metrics (Precision and Recall), it can be more challenging for business stakeholders or non-technical audiences to interpret compared to the simple, straightforward percentage offered by Accuracy. This complexity requires careful explanation to ensure stakeholders understand why a lower F1 Score might be preferred over a higher Accuracy in certain high-stakes contexts.

Choosing Your Metric: A Strategic Rule of Thumb

The ultimate decision regarding which metric to prioritize--F1 Score or Accuracy--must be guided by the specific business or domain context, particularly factoring in the financial or ethical cost associated with different types of errors (False Positives vs. False Negatives).

We generally prioritize **Accuracy** when:

The dataset exhibits classes that are relatively balanced (e.g., a 60/40 or 50/50 split), meaning the naive majority predictor strategy would not be effective.
The costs associated with a False Positive error and a False Negative error are roughly equivalent or equally acceptable within the operating context of the problem.

We must prioritize the **F1 Score** when:

The dataset classes are significantly imbalanced (e.g., a severe 95/5 or 99/1 split), requiring the

model to demonstrate competence on the rare minority class.
There is a serious, tangible downside to predicting **False Negatives**, compelling us to prioritize minimizing missed positive cases (i.e., achieving high recall).

For example, consider a medical screening scenario aimed at predicting a severe, life-threatening disease. Here, a False Negative (failing to diagnose an afflicted person) is catastrophic. Because the F1 Score severely penalizes models exhibiting low Recall, it provides a much safer and inherently more robust assessment of performance in such high-stakes scenarios where minimizing specific error types is critical, making it superior to simple overall Accuracy.

Additional Resources