How to Choose Between F1 Score and Accuracy for Model Evaluation

Authored by

stats writer

December 3, 2025

Choosing the correct evaluation metric is perhaps one of the most critical decisions in building and validating a robust classification model in machine learning. While simple performance measures like Accuracy are intuitive and easy to grasp, they often fail spectacularly when applied to real-world datasets characterized by significant class disparities. This disparity necessitates the use of more sophisticated metrics, such as the F1 Score, which provide a balanced assessment of a model's predictive capabilities across all classes.

The F1 Score is demonstrably superior when the goal is to achieve a fine balance between precision and recall, especially in scenarios where misclassification costs are asymmetrical. Conversely, Accuracy remains appropriate primarily for datasets where the distribution of classes is roughly equal, and where the costs associated with false positives and false negatives are considered negligible or identical. Understanding the mathematical foundation and practical implications of each metric is vital for effective model validation and deployment.

This comprehensive guide delves into the mechanisms of both Accuracy and the F1 Score, demonstrating their calculation using a practical example based on a classification task. Furthermore, we will establish clear guidelines for when to prioritize one metric over the other, particularly focusing on the crucial factor of imbalanced classes, which dramatically influences the utility of simple predictive measures.

The Core Metrics of Classification Performance

In classification problems, the effectiveness of a statistical or machine learning model is measured by its ability to correctly assign data points to predefined categories. Two foundational metrics frequently employed for this assessment are the F1 Score and Accuracy. While both aim to quantify model performance, they capture distinct aspects of predictive success, making the choice between them highly context-dependent. A higher value in either metric generally indicates a more capable model in separating observations into their respective classes.

Despite their shared goal of performance quantification, the underlying formulae and the sensitivities of these metrics differ significantly. Accuracy offers a straightforward interpretation of overall correctness, while the F1 Score provides a nuanced view that accounts for conditional errors. The inherent divergence in their construction means that relying solely on one metric without understanding the data context can lead to misleading conclusions about a model's true utility, particularly in critical applications like medical diagnostics or fraud detection.

Before proceeding to a detailed comparison, it is essential to establish how these metrics are derived from the foundational elements of model output: the true positives, true negatives, false positives, and false negatives. These four outcomes, organized within the Confusion Matrix, are the building blocks for nearly every advanced classification metric. The subsequent sections will first illustrate the practical calculation of these metrics before examining their inherent advantages

and disadvantages in various modeling scenarios.

Deconstructing the Confusion Matrix: The Foundation

The performance evaluation of any binary classification model begins with the construction of the Confusion Matrix. This matrix is a two-by-two table that summarizes the results of the prediction model, showing the counts of correct and incorrect predictions broken down by each class. Understanding the components of this matrix--True Positives (TP), True Negatives (TN), False Positives (FP), and False Negatives (FN)--is indispensable for calculating metrics like precision and recall, which ultimately feed into the F1 Score.

A True Positive occurs when the model correctly predicts the positive class, while a True Negative is a correct prediction of the negative class. The errors are represented by False Positives (Type I Error), where the model incorrectly predicts positive when the actual class is negative, and False Negatives (Type II Error), where the model incorrectly predicts negative when the actual class is positive. The relative importance of minimizing FP versus FN depends entirely on the application; for example, minimizing FNs is crucial in disease screening, whereas minimizing FPs might be more important in spam filtering.

The metrics of precision and recall provide orthogonal views of model performance derived directly from the confusion matrix. Precision quantifies the accuracy of positive predictions, answering the question: "Of all the cases the model predicted as positive, how many were actually positive?" Recall, conversely, measures the model's ability to find all positive samples, answering: "Of all the actual positive cases, how many did the model correctly identify?" The synergy between these two metrics is essential for developing a truly effective classification system.

Example: Calculating F1 Score & Accuracy in Practice

To solidify the understanding of these concepts, consider a classification scenario where we employ a logistic regression model to predict the outcome for 400 college basketball players--specifically, whether or not they will be drafted into the NBA. In this context, being drafted is the positive class, and not being drafted is the negative class. The resulting predictions allow us to construct a detailed confusion matrix, summarizing the model's performance on this specific dataset.

The following illustration represents the resulting confusion matrix, which forms the basis for all subsequent metric calculations. This matrix clearly lays out the 120 players who were correctly predicted as drafted (TP), the 170 players correctly predicted as undrafted (TN), and the errors: the 70 false positives (FP) and 40 false negatives (FN). These raw counts are critical for quantifying the model's performance dimensions.

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

Using the values derived from this matrix (TP=120, TN=170, FP=70, FN=40), we can now proceed to calculate the standard classification metrics. These calculations demonstrate how precision and recall are first established, followed by the derivation of Accuracy and the F1 Score, illustrating the mathematical relationships between them.

Precision: Represents the ratio of correct positive predictions relative to the total number of instances classified as positive by the model.

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

Recall: Measures the ratio of correct positive predictions relative to all actual positive instances in the dataset (i.e., the sensitivity).

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

Accuracy: Quantifies the percentage of all observations that were correctly classified, regardless of their class label.

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

F1 Score: Calculated as the harmonic mean of precision and recall, providing a single metric that balances both error types.

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

The Pitfalls of Accuracy in Imbalanced Classes

While Accuracy is easily the most intuitive metric, its simplicity hides a critical vulnerability: its susceptibility to distortion in datasets featuring imbalanced classes. Class imbalance occurs when one class (the majority class) significantly outnumbers the other (the minority class). In such scenarios, a model can achieve deceptively high accuracy simply by prioritizing the prediction of the majority class, effectively ignoring the minority class which is often the class of interest.

Consider the basketball drafting example again, but assume a highly skewed distribution where 95% of players are never drafted (negative class) and only 5% are drafted (positive class). If a naive model predicts that absolutely no player will ever be drafted, its overall Accuracy would still be 95%. This high score suggests excellent performance, yet the model is completely useless for the critical task of identifying the drafted players. Such a scenario underscores why relying solely on accuracy in domains like anomaly detection or medical diagnostics is highly dangerous, as the cost of missing a true positive (a False Negative) is often catastrophic.

The primary advantage of Accuracy is its interpretability: stating that a model is 90% accurate is immediately understood by non-technical stakeholders. However, this clarity dissolves when dealing with imbalanced classes, as the metric fails to account for how the predictions are distributed across the classes. This lack of sensitivity to the crucial minority class makes accuracy an unreliable indicator of a model's true generalization capability when faced with skewed data distributions, mandating the use of alternative, balance-aware metrics.

The Strength of the F1 Score: Balancing Precision and Recall

The F1 Score addresses the critical limitations of accuracy by integrating precision and recall into a single, comprehensive measure. Since the F1 Score is the harmonic mean, it assigns more weight to lower values. This mathematical property ensures that a model cannot achieve a high F1 score unless both precision and recall are reasonably high. If a model exhibits extremely high precision but poor recall (or vice-versa), the F1 score will be heavily penalized, offering a much more realistic view of overall performance than accuracy would.

The fundamental strength of the F1 Score lies in its robustness when evaluating models built upon imbalanced classes. Because both precision (which concerns False Positives) and recall (which concerns False Negatives) must be optimized simultaneously, the F1 Score intrinsically accounts for the distributional characteristics of the data. For instance, in the previous example where 95% of players are not drafted, a model that simply predicts 'not drafted' for everyone would have 0% recall for the drafted class, resulting in an F1 Score of zero, correctly signaling the model's failure to capture the positive class.

However, the F1 Score is not without its drawbacks, primarily concerning interpretation. Because it

is a synthesized metric--a blend of two other ratios--the F1 Score is inherently less straightforward for non-experts to interpret compared to simple percentage Accuracy. While a high F1 Score indicates a balance between minimizing False Positives and False Negatives, explaining the precise trade-off captured by a score of 0.75 requires a deeper understanding of precision and recall definitions. Despite this minor interpretability challenge, its utility in complex, real-world classification problems far outweighs this limitation.

Final Recommendations: Choosing the Right Metric

The selection between the F1 Score and Accuracy ultimately hinges on two critical factors: the balance of the dataset classes and the specific costs associated with prediction errors (False Positives vs. False Negatives). A useful rule of thumb provides clarity for most modeling situations: utilize Accuracy when class distributions are roughly balanced, and when the minimization of overall error is the singular focus, meaning there is no major differential downside to predicting a False Negative versus a False Positive.

Conversely, the F1 Score should be the default metric when the underlying class distributions are significantly imbalanced classes, or when there is a severe penalty associated with one type of error, most commonly False Negatives. A classic example illustrating this necessity is the detection of life-threatening conditions, such as using a model to predict whether or not someone has cancer. In this scenario, a False Negative (telling a person they are healthy when they actually have cancer) is profoundly worse than a False Positive (a potential false alarm requiring further testing).

Because the F1 Score heavily penalizes models that suffer from poor recall--the inability to find positive cases--it inherently drives the model toward minimizing those costly False Negatives. In high-stakes applications where the identification of the minority positive class is paramount, the F1 Score offers the most reliable assessment of a model's fitness for purpose. Therefore, data scientists must always conduct a thorough domain analysis to determine the relative costs of errors before settling on a single performance metric.

Summary of Pros and Cons

The following summarized list encapsulates the main arguments for and against using Accuracy versus the F1 Score.

Accuracy:

Pro: It is exceptionally easy to calculate and interpret. A 90% accuracy rate clearly signifies that 90% of all observations were correctly classified, making it ideal for communicating overall performance to non-technical audiences.

Con: It is highly misleading when datasets exhibit class imbalance. Accuracy fails to account for the actual distribution of data, allowing models to appear highly successful even when they completely fail to identify the critical minority class.

F1 Score:

Pro: It provides a superior assessment of model performance in scenarios involving imbalanced classes. By requiring both high precision and high recall, it ensures the model is effective across all classes, particularly the rare, critical ones.

Con: It is inherently harder to interpret than accuracy. Being the harmonic mean of two distinct ratios, the F1 Score requires domain knowledge to fully understand the specific performance trade-offs it encapsulates.