What is Balanced Accuracy? (Definition & Example)

Authored by

Mohammed looti

November 2, 2025

Understanding Classification Metrics and the Challenge of Imbalance

When building a classification model, evaluating its effectiveness requires robust metrics that accurately reflect its true performance. Many introductory machine learning projects rely solely on Overall accuracy, which measures the total proportion of correct predictions made across all classes.

However, this standard measure becomes misleading when the dataset exhibits significant class imbalance. In an imbalanced scenario, if 95% of the observations belong to Class A and only 5% belong to Class B, a simple model that always predicts Class A will still achieve 95% accuracy. This high score is deceptive because the model has failed entirely to learn anything about the minority class (Class B), which is often the class of most interest (e.g., fraud detection or disease diagnosis).

To overcome this fundamental limitation, data scientists turn to metrics that normalize performance across classes. The metric designed specifically for this task is **Balanced accuracy** (BA). BA provides a more truthful evaluation of the model's predictive power by accounting for the success rates on both the positive and negative classes, ensuring that the performance on the minority class is not overshadowed by the majority class.

By prioritizing the average recall achieved on each class, BA guarantees that a model must perform well on all categories--not just the most prevalent one--to achieve a high score. This makes it an indispensable tool in real-world applications where data distributions are rarely perfectly symmetrical.

Defining Balanced Accuracy (BA)

The **Balanced accuracy** metric is formally defined as the average of the recall obtained on each class, or, in the case of binary classification, the average of sensitivity and specificity. It is an intuitive and robust measure that effectively mitigates the bias introduced by unequal class sizes.

The core idea is to treat the positive and negative class prediction rates equally, regardless of how many instances of each class exist in the dataset. This equalization prevents the metric from being inflated by high performance solely on the dominant class.

For binary classification problems, the calculation is straightforward and relies on two foundational metrics: Sensitivity (True Positive Rate) and Specificity (True Negative Rate).

The formal calculation is expressed as:

Balanced accuracy = (Sensitivity + Specificity) / 2

where the components are defined by their rates of correct classification within their respective classes:

Sensitivity: Also known as the True Positive Rate (TPR). It measures the percentage of actual positive cases that the model correctly identifies. This is critical for capturing the minority class instances.
Specificity: Also known as the True Negative Rate (TNR). It measures the percentage of actual negative cases that the model correctly identifies.

The Core Components: Sensitivity and Specificity

To appreciate why balanced accuracy works so well, it is essential to understand the roles of Sensitivity and Specificity. These two metrics shift the focus from overall correct predictions to class-specific performance. They are derived from the outcomes summarized in a confusion matrix.

Sensitivity answers the question: "Of all the positive outcomes that truly occurred, how many did the model correctly flag?" A high sensitivity is vital when failing to detect a positive case (a False Negative) carries a high cost, such as missing a fraudulent transaction or a severe medical condition. Mathematically, it is calculated as True Positives divided by the total number of actual positives (True Positives + False Negatives).

Conversely, Specificity addresses: "Of all the negative outcomes that truly occurred, how many did the model correctly identify?" High specificity is important when incorrectly classifying a negative case as positive (a False Positive) is highly undesirable. It is calculated as True Negatives divided by the total number of actual negatives (True Negatives + False Positives).

In a severely imbalanced dataset, standard accuracy is dominated by high specificity, as the negative (majority) class provides most of the data points. **Balanced accuracy** forces the model evaluation to equally consider sensitivity (performance on the minority class) and specificity (performance on the majority class), providing a holistic and less biased view of the model's generalization capability.

Example: Calculating Balanced Accuracy in Practice

Consider a scenario where a sports analyst utilizes a statistical model to predict the outcome for 400 different college basketball players: whether they will or will not be drafted into the NBA. This scenario naturally features extreme class imbalance, as only a small fraction of college players are actually drafted.

The analyst runs the model, and the predictions are summarized in the following confusion matrix.

In this context, "Drafted" is the positive class, and "Not Drafted" is the negative class.

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

		Predicted	
		Drafted = Yes	**Drafted = No**
Actual	**Drafted = Yes**	**15** (True Positive)	**5** (False Negative)
	Drafted = No	**5** (False positive)	**375** (True Negative)

From the matrix, we observe that 20 players were drafted (15 True Positives + 5 False Negatives) and 380 players were not drafted (375 True Negatives + 5 False Positives). The imbalance ratio is 380:20, or 19:1.

To calculate the **Balanced accuracy** of the model, we must first calculate the sensitivity (True Positive Rate) and specificity (True Negative Rate):

Sensitivity (True Positive Rate): This measures the rate of correctly predicted drafted players.
Sensitivity = True Positives / (True Positives + False Negatives) = 15 / (15 + 5) = 15 / 20 = 0.75
Specificity (True Negative Rate): This measures the rate of correctly predicted undrafted players.
Specificity = True Negatives / (True Negatives + False Positives) = 375 / (375 + 5) = 375 / 380 ≈ 0.9868

We can then calculate the **Balanced accuracy** as the simple average of these two rates:

Balanced accuracy = (Sensitivity + Specificity) / 2
Balanced accuracy = (0.75 + 0.9868) / 2
Balanced accuracy = 0.8684

The balanced accuracy for the model is calculated to be **0.8684**. This score reflects a strong, but not perfect, performance across both the minority (drafted) and majority (not drafted) classes.

Why Balanced Accuracy is Essential for Imbalanced Datasets

The true utility of **Balanced accuracy** becomes evident when comparing it directly against the Overall accuracy metric in scenarios like the basketball drafting example, where classes are severely imbalanced.

Let's first calculate the overall accuracy for the same model:

Accuracy = (True Positives + True Negatives) / Total Observations
Accuracy = (15 + 375) / (15 + 375 + 5 + 5)
Accuracy = 390 / 400
Accuracy = 0.975

The overall accuracy of the model is **0.975** (97.5%). On the surface, this score sounds exceptionally high and suggests a near-perfect statistical model. However, this high score is heavily biased by the model's excellent performance on the 380 players who were not drafted (the majority class).

Consider a naive baseline model that simply predicts every single player will not get drafted. Since 380 out of 400 players genuinely were not drafted, this simplistic, non-predictive model would achieve an accuracy of 380 / 400 = **0.95** (95%). This result highlights the danger: our complex model only slightly outperformed a completely useless model in terms of overall accuracy (97.5% vs. 95%), obscuring the fact that it still missed 5 out of 20 actual drafted players.

The balanced accuracy score of **0.8684**, being significantly lower than the overall accuracy, gives a much more realistic and honest assessment. It captures the fact that while the model is very good at predicting non-drafted players (Specificity ≈ 0.987), its performance in identifying drafted players (Sensitivity = 0.75) is lower. By averaging these rates, BA reveals the true performance capability across both classes, preventing us from being fooled by the majority class dominance.

Interpreting the Balanced Accuracy Score

A **Balanced accuracy** score ranges from 0 to 1, where 1 signifies a perfect classification model that correctly identifies every instance of both the positive and negative classes. A score of 0.5 represents performance equivalent to random guessing, or the performance of a model that consistently predicts only the majority class.

In practical terms, the closer the balanced accuracy is to 1, the better the model is able to correctly classify observations across all categories, especially those rare events that are difficult to detect. A high BA score indicates strong generalization capabilities and minimal bias toward any single class.

If a model achieves high overall accuracy but low balanced accuracy, it is a clear indicator of class imbalance bias, suggesting the model is primarily learning to optimize for the majority class while neglecting the minority class. Conversely, if a model has a balanced accuracy score close to its overall accuracy score, it suggests the dataset is either balanced or the model is performing consistently well across all classes.

Ultimately, BA helps stakeholders make better decisions about model deployment. If the goal is

truly to identify players who will get drafted (the minority class), relying on the 0.975 overall accuracy is dangerous. The **0.8684** BA score provides the necessary insight, revealing that the model's ability to predict players who will get drafted *and* those who will not is acceptably high, yet still has room for improvement regarding the minority class.

Additional Resources for Classification Analysis

While balanced accuracy is a powerful metric, it is often used in conjunction with other evaluation tools to provide a complete picture of model performance. Metrics such as the F1-Score, Cohen's Kappa, and the Area Under the Receiver Operating Characteristic Curve (ROC AUC) also address class imbalance and provide different perspectives on the trade-offs between precision and recall.

To properly calculate and interpret balanced accuracy, it is essential to first generate the confusion matrix accurately. The following tutorials explain how to create a confusion matrix in different statistical software packages, which is the foundational step for calculating sensitivity, specificity, and ultimately, balanced accuracy: