

# How to compare two ROC curves?

Authored by  
**stats writer**

November 19, 2025

## RECOMMENDED CITATION

stats writer (2025). *How to compare two ROC curves?*. PSYCHOLOGICAL SCALES.  
Retrieved from <https://scales.arabpsychology.com/?p=96930>

In the field of machine learning and statistical modeling, determining the superior performance between two or more predictive models is a fundamental task. When dealing with binary outcomes, the ROC curve (Receiver Operating Characteristic curve) serves as an indispensable visualization tool for assessing the discriminatory power of a classification model.

The primary methods for comparing two ROC curves involve both quantitative statistical analysis and visual inspection. Quantitatively, researchers often rely on calculating the difference between the Area Under the Curve (AUC) values. If this difference crosses a statistically significant threshold, we can conclude that one model offers superior prediction capability. Visual comparison, while less rigorous, provides immediate insight by observing which curve is positioned closer to the ideal top-left corner of the plot, indicating higher overall accuracy across various threshold settings.

## Understanding the ROC Curve and Its Components

A ROC curve plots the True Positive Rate (also known as sensitivity) against the False Positive Rate (or  $1 - \text{specificity}$ ) across all possible classification thresholds. This visualization is crucial because it provides a comprehensive view of a model's performance, unlike metrics that rely on a single, arbitrarily chosen threshold. It allows practitioners to understand the trade-off between correctly identifying positive cases and incorrectly flagging negative cases as positive.

The structure of the ROC plot is standardized to represent these core performance measures. The **x-axis** represents the False Positive Rate (FPR), which is mathematically calculated as  $1 - \text{specificity}$ . The **y-axis** represents the True Positive Rate (TPR), commonly known as sensitivity. A perfectly random classifier would follow the diagonal line from (0, 0) to (1, 1), yielding an AUC of 0.5, while a perfect classifier would trace a path directly up the Y-axis and then across the top X-axis, achieving an ideal AUC of 1.0.

Understanding the components plotted on the ROC curve is vital for accurate model interpretation:

**Sensitivity (True Positive Rate)**: This is the probability that the classification model correctly identifies a positive outcome when the true outcome is indeed positive. High sensitivity is crucial in contexts like medical diagnostics where missing a positive case (a false negative) carries a high cost.

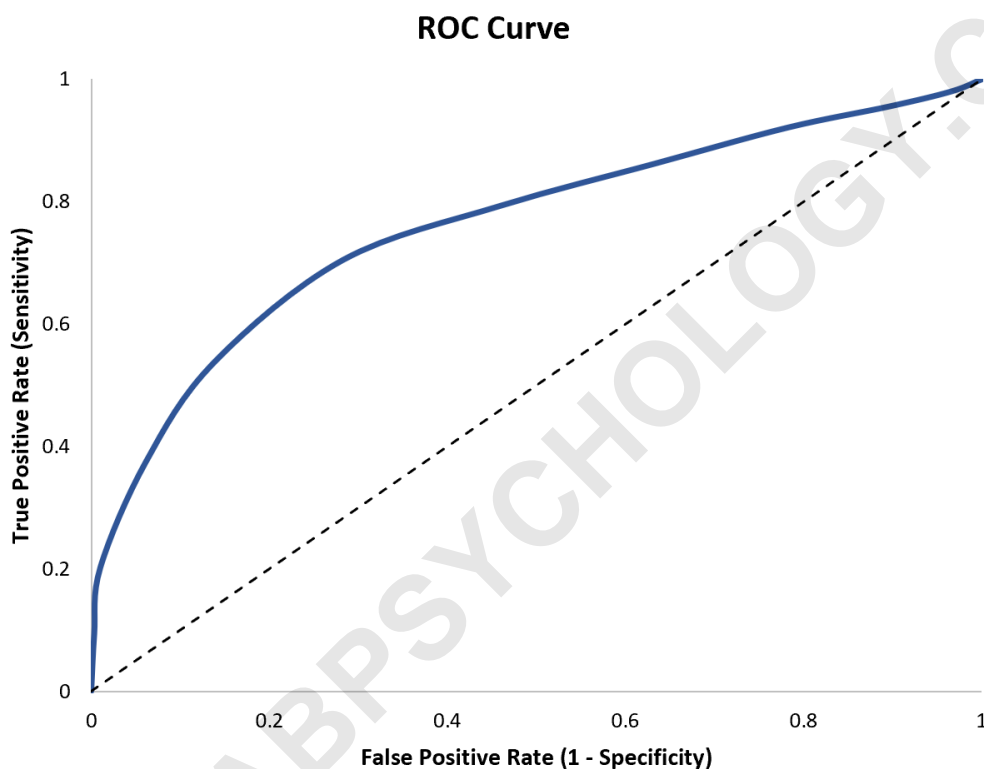
**Specificity (True Negative Rate)**: This represents the probability that the model correctly predicts a negative outcome when the actual outcome is negative. High specificity minimizes false alarms, which is important in contexts where false positives are costly or disruptive.

## Visualizing Model Performance on the ROC Plot

The curve illustrates how adjustments to the decision threshold simultaneously impact both sensitivity and specificity. Setting a threshold to be more lenient towards positive predictions

increases sensitivity but decreases specificity, causing the point on the curve to move toward the upper right of the graph. Conversely, enforcing a stricter threshold decreases sensitivity but increases specificity, pushing the point toward the lower left.

The overall shape and location of the curve are direct indicators of model quality. The more the ROC curve bends towards and hugs the upper-left corner--the region where sensitivity is maximized and FPR (1-specificity) is minimized--the better the model performs at separating the classes across all possible thresholds. This visual assessment provides an immediate, intuitive understanding of relative model quality.

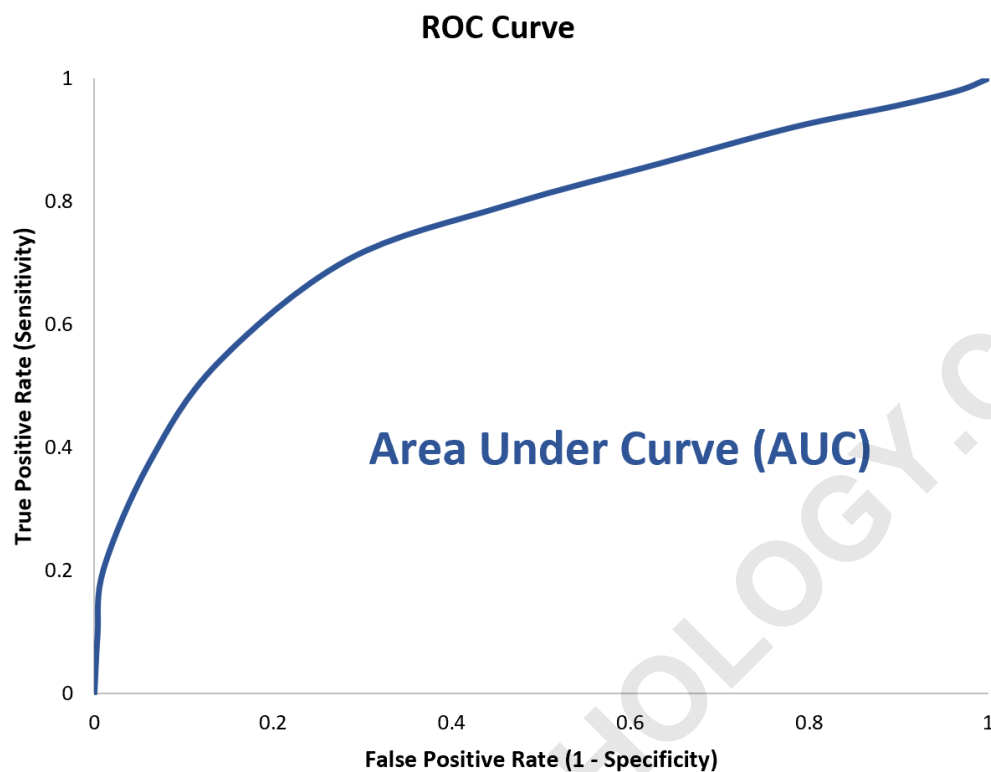


## Quantifying Performance with the Area Under the Curve (AUC)

While visual inspection offers a rapid qualitative assessment, the AUC provides a single, quantitative summary statistic of the model's overall performance, independent of the chosen classification threshold. It represents the entire two-dimensional area beneath the ROC curve. Critically, the AUC can be interpreted as the probability that the model will rank a randomly chosen positive instance higher than a randomly chosen negative instance.

The AUC ranges from 0 to 1. An AUC value hovering around 0.5 suggests that the model's predictions are no better than random guessing. Conversely, a value approaching 1.0 signifies highly accurate classification performance. For any serious model selection process, the objective

is inherently to maximize this metric, as it demonstrates superior discriminatory power across the full spectrum of operational points.



## Statistical Methods for Comparing Two ROC Curves

The simplest quantitative method for comparison is calculating the difference between the respective AUC values. The model with the higher AUC is provisionally considered superior. However, in rigorous analysis, simply observing a numerical difference is insufficient, as that difference might be attributable purely to sampling variation or chance.

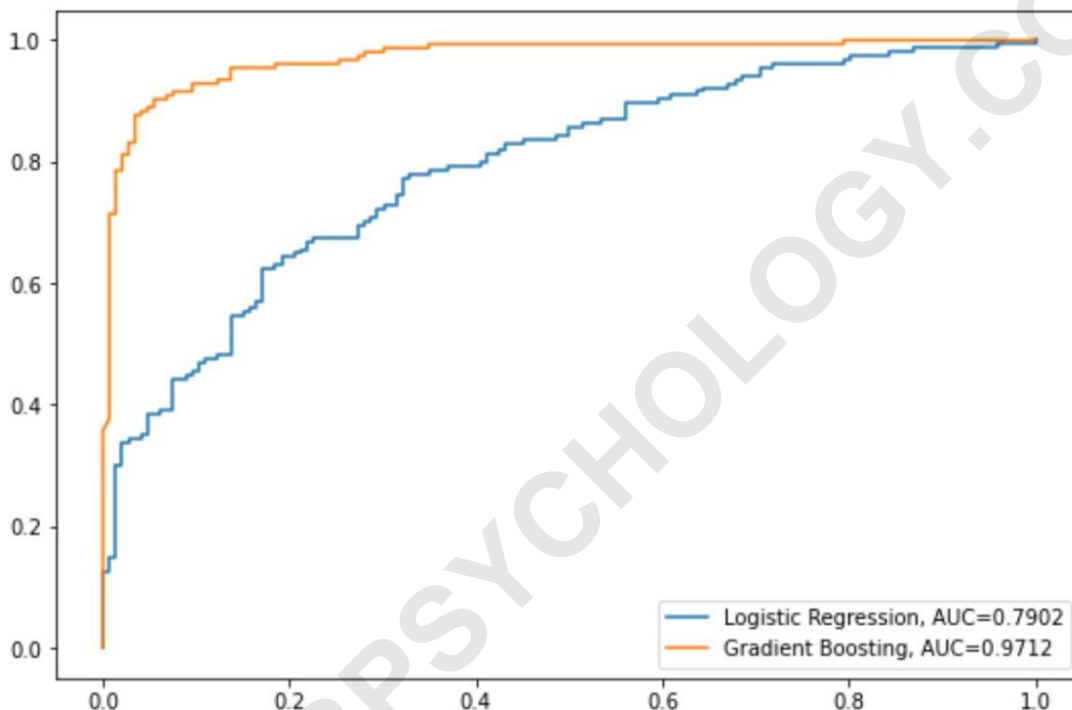
To definitively compare two models trained on the same data, specialized statistical tests are required to address the dependency between the paired ROC curves. Methods such as the Delong test or advanced bootstrapping techniques calculate confidence intervals and p-values for the difference in AUC. These tests allow researchers to determine if the observed difference is statistically significant, enabling a confident assertion that one model truly outperforms the other across various thresholds. If the confidence interval for the difference excludes zero, the difference is statistically significant, confirming the superiority of the model with the larger AUC.

This quantitative approach is paramount for robust academic reporting and critical decision-making in industry, ensuring that the selection of the best model is based on verifiable statistical evidence rather than mere visual bias or small, non-significant numerical advantages.

## Case Study: Comparing Logistic Regression and Gradient Boosting

To provide a concrete example of this comparative process, let us examine a scenario involving two distinct classification algorithms applied to the same dataset: a standard logistic regression model and a more complex gradient boosted model. The goal is to predict a binary outcome, and we generate the ROC curve for each to assess their relative predictive performance.

By plotting both curves on the same coordinate system, we can easily compare their performance characteristics visually:



In this visualization, the blue line corresponds to the performance of the logistic regression model, while the orange line represents the performance profile of the gradient boosted model. Visually, the orange curve is notably higher and maintains a closer proximity to the ideal upper-left boundary throughout its span, strongly suggesting that the gradient boosting approach is more effective.

## Interpreting the Quantitative Results for Model Selection

The visual observations are confirmed and quantified by calculating the specific Area Under the Curve (AUC) values for each classifier:

AUC of logistic regression model: **0.7902**

AUC of gradient boosted model: **0.9712**

The resulting difference in AUC is highly substantial (0.9712 versus 0.7902). This disparity indicates that the gradient boosted model possesses a significantly greater overall discriminatory capability than the simpler logistic regression model in this specific context. While an AUC of 0.7902 might be considered acceptable, 0.9712 is exceptional, demonstrating that the boosting algorithm excels at clearly separating the positive and negative classes with high precision across various decision thresholds.

### **Conclusion: Selecting the Optimal Classifier**

The methodology for comparing ROC curves is a foundational pillar in predictive analytics. While this case study focused on comparing only two models, the framework is readily scalable. Data scientists routinely evaluate numerous different classification models--such as Support Vector Machines, Random Forests, and Neural Networks--on the same dataset.

By plotting all their respective ROC curves and calculating their statistically compared AUCs, one can objectively and systematically identify the single best-performing model for a given prediction task. This process ensures that the selected classifier maximizes performance across critical metrics like sensitivity and specificity, guaranteeing the highest probability of correct classification for real-world deployment.