Area under ROC curve – review and efficient computation in R

by Błażej Kochański, Przemysław Peplinski, Miriam Nieslona, Wiktor Galewski, and Piotr Geremek

Abstract The Area Under the Receiver Operating Characteristic Curve (AUC) is a widely used measure for evaluating the performance of binary classification models. In the literature and in practice, it appears under various names and is closely related to other performance measures. In this paper, we review these formulations and discuss the motivation for efficient AUC computation in empirical analysis. We survey R packages that provide implementations of AUC and describe the algorithms they employ. We then conduct a benchmarking study to compare the execution time of selected implementations. **A case study illustrates how the choice of AUC computation method can substantially reduce computation time and accelerate analytical workflows.**

1 Introduction

Assessing the quality of predictive models is a critical stage in machine learning and statistical inference processes. Among the many available metrics, the Area Under the Receiver Operating Characteristic curve (AUC) has established itself as a standard measure of discrimination, particularly in binary classification problems. The popularity of this metric stems from its independence from the chosen cut-off point and its robustness to class imbalance, making it useful in diverse fields such as medical diagnostics, psychology, and credit risk assessment.

Modern statistical applications require not only methodological correctness but also high computational efficiency. In the era of Big Data and increasing complexity of validation procedures, a single calculation of the AUC is often insufficient. Similar challenges are posed by simulation market models in banking, where generating ROC curves without historical data serves to estimate the impact of scoring models on portfolio profitability. In such scenarios, classical algorithm implementations can become a computational bottleneck, rendering procedures like the bootstrap impractical.

The goal of this paper is to review methods for calculating the AUC measure and to analyze the efficiency of available implementations within the R environment. In the first part of the work, we organize definitions and highlight the mathematical links between AUC and other measures, such as the Gini coefficient. Next, we discuss the main classes of algorithms used for AUC estimation: trapezoidal integration, optimized pairwise comparisons, and rank-based approaches. Finally, we present the results of performance benchmarks for popular R packages and propose our alternatives, identifying optimal solutions for different sizes of large datasets.

2 ROC curve and AUC

The ROC curve is one of the most widely used tools for assessing the performance of binary classifiers. Introduced in the 1940s for radar signal analysis, the ROC curve derives its name from "Receiver Operating Characteristic" (Junge and Dettori, 2018). Today, ROC curves are used across diverse fields—such as medicine, biotechnology, computer science, and banking—to evaluate scoring and classification models.

The ROC curve is plotted within the unit square, connecting points defined by sensitivity and specificity at various decision thresholds ("cut-offs"). In an ROC curve, the x-axis represents the false positive rate (FPR, or "1 - specificity"), the proportion of false positives among all actual negatives. The y-axis represents sensitivity (or "recall"), the proportion of correctly identified positives among all actual positives (true positive rate, TPR) (Fawcett, 2006).

The AUC (Area Under the ROC Curve) provides a single-number summary of the classification performance of binary models across all possible decision thresholds. A ROC curve that passes through the point (0, 1) corresponds to an ideal classifier, achieving zero false positives and perfect detection of positive cases. Curves lying closer to this point are therefore more desirable (Hanley and McNeil, 1982); they correspond to larger values of the Area Under the Curve (AUC). By contrast, the diagonal line connecting (0, 0) and (1, 1) represents a random classifier, for which the expected value of AUC equals 0.5.

In many application domains—most notably in credit scoring—the AUC is often transformed into a measure that tends to zero for a completely random classifier and equals one for a perfect classifier (it may also take negative values if the classifier's predictions are systematically reversed).

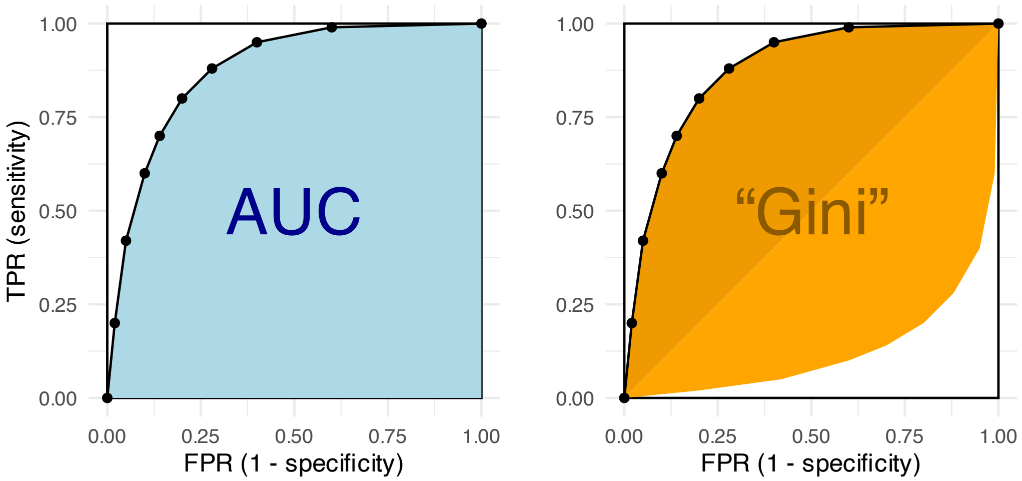


Figure 1: Geometric interpretation. The black piecewise linear curve represents the receiver operating characteristic (ROC) curve. The AUC (also referred to as AUROC, ROC score, C-statistic, or Vargha–Delaney A) is defined as the area under the ROC curve. A linear transformation of the AUC—known, among others, as the Gini coefficient, Cliff's delta, Somers' D, Accuracy Ratio, or the ROC skill score—equals twice the area between the diagonal line y=x and the ROC curve.

This transformation ($2 \times \text{AUC} - 1$) is commonly referred to as the Gini coefficient in the credit scoring literature. As we will see later, the same measure appears under different names in other fields.

Figure 1 illustrates an empirical ROC curve defined by eight discrete cut-off points, along with the areas corresponding to the AUC and the Gini coefficient.

3 AUC – alternative names and related measures

The Area Under the Receiver Operating Characteristic (ROC) curve appears in the literature under numerous aliases and mathematical formulations. Recognizing the equivalence of these measures is essential for correct interpretation, comparison of results across disciplines, and informed use of software implementations. As shown below, the AUC is also referred to as the *AUROC*, *ROC score*, *concordance statistic* (*C-statistic*), and the *Vargha-Delaney A statistic*. The AUC is closely related to the Mann-Whitney *U* statistic and is used as a measure of effect size in the context of the Mann-Whitney test and related procedures, such as the Brunner-Munzel test. In this context, it is often referred to as the *probability of superiority*, *relative treatment effect* or a *measure of stochastic superiority*.

The Gini coefficient, which is a simple linear transformation of the AUC, is also known as *Cliff's delta*, the *ROC skill score* (ROCSS), or the *Accuracy Ratio*. It is a special case of *Somers' D statistic* and is occasionally viewed as a variant of the *rank-biserial correlation*.

In medical statistics and epidemiology, particularly when evaluating logistic regression models with binary outcomes, the AUC is frequently referred to as the *C-statistic* or *concordance statistic* (Austin and Steyerberg, 2012). This term was introduced by Harrell et al. (1982) as representing the probability that a randomly selected positive instance receives a higher predicted score than a randomly selected negative instance.

This probabilistic interpretation serves as the bridge to non-parametric statistics. The U statistic in the Mann–Whitney test represents the number of correctly ordered pairs in all pairwise comparisons between the two groups, where one observation comes from each group, with ties typically counted as 0.5. It can be shown that the area under the ROC curve is equivalent to the Mann-Whitney U statistic divided by the product of the sample sizes of the two groups:

$$\text{AUC} = \frac{U}{n_1 n_0}$$

This equality allows confidence intervals for AUC to be derived from the distributional properties of the Mann–Whitney U statistic.

In the context of stochastic ordering tests, such as Mann-Whitney or Brunner–Munzel, the AUC can serve as a measure of effect size and be interpreted as the *probability of superiority*, as it represents the the likelihood that a randomly chosen positive case ranks higher than a randomly chosen negative case (Hanley and McNeil, 1982; Bamber, 1975). It is also be referred to as a *relative treatment effect* or *stochastic superiority statistic* in this context (Brunner and Munzel, 2000; Karch, 2021).

In the behavioral sciences and psychology, similar concepts have been developed to quantify effect sizes. The *Common Language Effect Size* (CL/CLES) (McGraw and Wong, 1992) defines an effect size as, again, the probability that a score sampled at random from one distribution exceeds a score sampled from another. The name reflects its purpose: to provide an intuitively interpretable metric for audiences without formal statistical training. While the original formulation by McGraw and Wong assumed normal distributions, the measure was later generalized and, under the name *A measure of stochastic superiority*, or *Vargha–Delaney A* (Vargha and Delaney, 2000), became mathematically equivalent to the AUC.

While the AUC ranges from 0.5 (random classification) to 1.0 (perfect classification), many disciplines prefer a measure scaled symmetrically around zero to express the strength and direction of association or discriminatory power. This motivates the use of the Gini coefficient (or Gini index). In classification settings, the Gini coefficient is a simple linear transformation of the AUC, defined as (Hand and Till, 2001):

$$\text{Gini} = 2 \times \text{AUC} - 1$$

The term "Gini coefficient" is used predominantly in the credit scoring literature. To distinguish it from the classical Gini index in economics, which measures income or wealth inequality, the name *pseudo-Gini* is sometimes employed (Idczak, 2019). In other application domains, mathematically equivalent quantities may appear under different names.

By construction, the Gini coefficient takes values in (-1, 1) with zero indicating no discriminatory power and negative values corresponding to systematically reversed predictions. Owing to this symmetry, it admits an interpretation analogous to that of a correlation coefficient. Indeed, some authors, such as Wendt (1972), have proposed the rank-biserial correlation in a form that is mathematically equivalent to the Gini coefficient, although this term is occasionally used for closely related but non-identical formulations (Rubia, 2022).

An alternative derivation, independent of ROC analysis, arises from the Cumulative Accuracy Profile (CAP) curve (Sobehart et al., 2000). The CAP plots the cumulative proportion of positive cases against the cumulative proportion of the population, sorted by predicted scores. From this curve, the *Accuracy Ratio* (*AR*) is defined as the ratio of the area between the model CAP and the random CAP to the area between the ideal CAP and the random CAP. It can be shown that

$$\text{AUC} = \frac{\text{AR} + 1}{2}$$

(Engelmann et al., 2003), demonstrating that although AR originates from the CAP curve rather than the ROC curve, it conveys exactly the same information as the Gini coefficient.

Just like AUC, the Gini coefficient admits a probabilistic interpretation. If we randomly sample one positive and one negative case and compare their scores, the Gini coefficient equals the difference between the probability that the positive case receives a higher score than the negative case and the probability of the reverse. Agresti (**?**) denotes AUC by α and the Gini coefficient by Δ, referring to both as measures of stochastic superiority. Cliff (1993) describes the effect size measure d, which advocates a similar probabilistic perspective on rank-based comparisons. This measure, identical to the Gini coefficient, is later sometimes referred to as *Cliff's delta* (Bais and Van Der Neut, 2022).

Somers' D is a measure of association between a predictor X and an ordinal dependent variable, based on the comparison of concordant and discordant pairs (Somers, 1962). When Y is binary, Somers' D reduces to 2AUC-1 (Newson, 2002). Thus, in binary classification settings, the Gini coefficient can be interpreted as Somers' D with a binary response variable.

It is worth mentioning that ROC curves (under the name of *relative* operating characteristics curves) are also used in meteorology to assess probabilistic forecasts for binary events. In this context, the AUC may be referred to as the ROC score, while the Gini coefficient is known as the ROC skill score (Kharin and Zwiers, 2003; Mason and Graham, 1999).

4 The need for computational efficiency in AUC estimation

In the context of modern statistical modeling and machine learning, a single calculation of the AUC metric is often insufficient for a comprehensive assessment of model quality. Complex validation, optimization, and simulation procedures require the AUC statistic to be computed repeatedly, imposing high demands on the computational efficiency of the underlying algorithms. The following sections outline key areas where the performance of AUC estimation is critical.

Uncertainty estimation and permutation tests

The estimation of confidence intervals and the verification of statistical hypotheses for the AUC often rely on resampling methods, such as bootstrapping or permutation tests. While necessary when simple parametric assumptions cannot be met, these methods generate significant computational loads.

In the case of massive datasets, the process of generating a single performance estimate can be computationally expensive (LeDell et al., 2015). The authors point out that when using complex prediction methods, even with relatively small datasets, cross-validation can consume a large amount of time, making the bootstrap a computationally intractable approach to variance estimation in many practical settings.

Similarly, in the context of meta-analysis of diagnostic accuracy studies, there are instances of bootstrap algorithms usage to determine confidence intervals for the AUC of the Summary ROC (SROC) curve (Noma et al., 2021). Their approach involves computing AUC estimates from a large number of bootstrap samples (e.g., $B = 1000$), which necessitates efficient calculation routines.

Permutation-based inference also demands repeated calculation. Bandos, Rockette, and Gur (2006) developed a permutation test for comparing ROC curves in multireader studies. An exact permutation test in this setting is formed by determining the frequency of the statistic—estimating the average difference in AUCs across all possible exchanges of reader ratings. For larger samples, this requires an asymptotic approach due to the computational intensity of calculating differences for every permutation (Bandos et al., 2006). Furthermore, Pauly, Asendorf, and Konietschke (2016) proposed rank-based studentized permutation methods for the nonparametric Behrens-Fisher problem, which corresponds to inference for the AUC. They demonstrated that the studentized permutation distribution ofthe Brunner-Munzel rank statistic is asymptotically standard normal, providing a theoretical foundation for consistent confidence intervals and tests, which rely on these intensive permutation procedures (Pauly et al., 2016).

Variable importance measures

Efficient AUC computation is a prerequisite for specific Variable Importance Measures (VIM). Therefore an AUC-based permutation variable importance measure for Random Forests was introduced, designed to be more robust to unbalanced classes than standard error-rate-based measures. This procedure requires the AUC to be computed for each tree in the forest, both before and after permuting the values of a given predictor. Given the number of trees in a forest and the number of predictors, this approach results in a vast number of AUC calculations, far exceeding the computational cost of standard single-pass metrics (Janitza et al., 2013).

Beyond model-specific methods, model-agnostic interpretability frameworks also rely heavily on repeated metric estimation. The DALEX package (Biecek, 2018) implements a permutation-based variable importance method applicable to any predictive model. This approach measures the change in model performance, such as the drop in AUC (represented as the loss function $L(y, \hat{y}) = 1 - AUC$), after permuting the values of a single predictor. To ensure stability of the importance estimates, this permutation process is typically repeated B times (e.g., $B = 10$) for every feature in the dataset. Consequently, for a model with p features, assessing global feature importance requires $p \times B$ independent AUC calculations, creating a linear dependency between the number of features and the computational cost.

AUC maximization algorithms

In the era of Big Data, learning algorithms that directly maximize AUC, rather than accuracy, are gaining importance but algorithms maximizing model accuracy do not necessarily maximize the AUC score. However, direct AUC maximization presents a computational challenge because the function is non-decomposable over individual examples. This has led to the development of stochastic AUC maximization methods for big data and "Deep AUC Maximization" (DAM) for deep learning, where optimization must be performed efficiently on large-scale datasets (Yang and Ying, 2023).

Simulation and curve modeling in credit risk

In credit risk management, theoretical ROC curve models are employed to simulate and assess the impact of scoring models when actual data is unavailable or limited. Kochański (2022) notes that fitting models such as the binormal or bigamma curves is helpful for "generating ROC curves without underlying data". This capability is essential for assessing the impact of a credit scorecard that is "yet to be built" (Kochański, 2022). Such simulation analyses often involve generating numerous curves and estimating their parameters to forecast portfolio quality, further justifying the need for computationally efficient AUC estimation methods.

Beyond single-curve fitting, Kochański (2021) extends the analysis to a dynamic market environment, proposing a simulation model for risk and pricing competition. This framework explores how the discrimination power of credit scoring models influences key business metrics, illustrating the "trade-off between profitability, market share, and credit loss rates" (Kochański, 2021). The study demonstrates that even marginal improvements in discrimination power can yield substantial benefits, a conclusion derived from simulation scenarios that model the interactions between lenders and borrowers. Such comprehensive market analyses require processing numerous scenarios to identify profit-maximizing strategies, further highlighting the role of performance metrics as fundamental parameters in complex economic models.

5 R packages

Table 1: Functions computing AUC.

Package	Function	Usage	Method	Language
bigstatsr	AUC	`AUC(-pred, target)`	Pairwise comparison (optimized)	C++, R
caTools	colAUC	`colAUC(-pred, target)[1, 1]`	Rank sum	R
cvAUC	AUC	`AUC(-pred, target)`	Trapezoidal rule	R
DescTools	Cstat	`Cstat(-pred, target)`	Pairwise comparison (optimized)	C++, R
effsize	VD.A	`VD.A(pred ~ factor(target))$estimate`	Rank sum	R
fbroc	boot.roc	`boot.roc(-pred, as.logical(target))$auc`	Trapezoidal rule	C++, R
Hmisc	somers2	`somers2(-pred, target)["C"]`	Rank sum	R
MLmetrics	AUC	`AUC(-pred, target)`	Rank sum	R
mltools	auc_roc	`auc_roc(-pred, target)`	Trapezoidal rule	R
ModelMetrics	auc	`auc(target, -pred)`	Rank sum	C++, R
precrec	evalmod	`evalmod(mmdata(scores = -pred, labels = target), mode="aucroc")$uaucs$aucs`	Rank sum	C++, R
pROC	auc	`auc(target, pred, lev=c('0', '1'), dir=">")`	Trapezoidal rule	R
rcompanion	vda	`vda(x = pred[target == 0], y = pred[target == 1], digits=100)`	Rank sum	R
ROCR	performance	`performance(prediction(-pred, target), "auc")@y.values[[1]]`	Trapezoidal rule	R
scikit-learn	roc_auc_score	`import("sklearn.metrics")$roc_auc_score(target, -pred)`	Trapezoidal rule	Python, C
scorecard	perf_eva	`perf_eva(-pred, target, binomial_metric = "auc", show_plot=FALSE)$binomial_metric$dat$AUC`	Trapezoidal rule	R
yardstick	roc_auc_vec	`roc_auc_vec(as.factor(target), pred)`	Trapezoidal rule	R
DescTools	SomersDelta	`SomersDelta(-pred, target)/2 + 1/2`	Pairwise comparison (unoptimized)	C++, R

We identified three main types of algorithms for AUC computation: (1) trapezoidal integration over the ROC Curve, (2) (optimized) pairwise comparison, (3) and rank-based (Mann–Whitney U statistic formulation).

Let (s_i, y_i) for $i = 1, \ldots, n$ denote the score assigned to an account and its corresponding true label. In line with standard practice in credit scoring, we assume that $y_i = 1$ indicates a "bad" account (e.g., one that defaults, doesn't repay the loan), while $y_i = 0$ represents a good account. A properly functioning scoring should assign lower scores to accounts with a higher predicted probability of being bad, and higher scores to those likely to be good.

Let $n_1 = \sum_{i=1}^{n} \mathbb{I}(y_i = 1)$ denote the number of bad accounts, and $n_0 = \sum_{i=1}^{n} \mathbb{I}(y_i = 0)$ – number of good accounts.

Trapezoidal integration over the ROC curve

$$\mathrm{AUC} = \sum_{k=1}^{m-1} (\mathrm{FPR}_{k+1} - \mathrm{FPR}_k) \cdot \frac{\mathrm{TPR}_{k+1} + \mathrm{TPR}_k}{2}$$

where m is the number of distinct score thresholds from lowest to the highest score, TPR_k is the True Positive Rate, FPR_k is the False Positive Rate. $TPR_k = TP_k/n_1$ and $FPR_k = FP_k/n_0$, where TP_k is the number of true positives, FP_k is the number of false positives, n_1 is the number of positive cases, n_0 is the number of negative cases.

Optimized pairwise comparison

AUC as the probability that a randomly chosen positive instance receives a higher score than a randomly chosen negative instance:

$$\mathrm{AUC} = \frac{1}{n_1 n_0} \sum_{i:y_i=1} \sum_{j:y_j=0} \left[\mathbb{I}(s_i < s_j) + \frac{1}{2} \mathbb{I}(s_i = s_j) \right]$$

Rank-based (Mann–Whitney U statistic formulation)

$$\text{AUC} = \frac{\bar{R}_1 - n_1(n_1+1)/2}{n_0}$$

where $\bar{R}_1$ is the mean rank for s_i where $y_i = 1$:

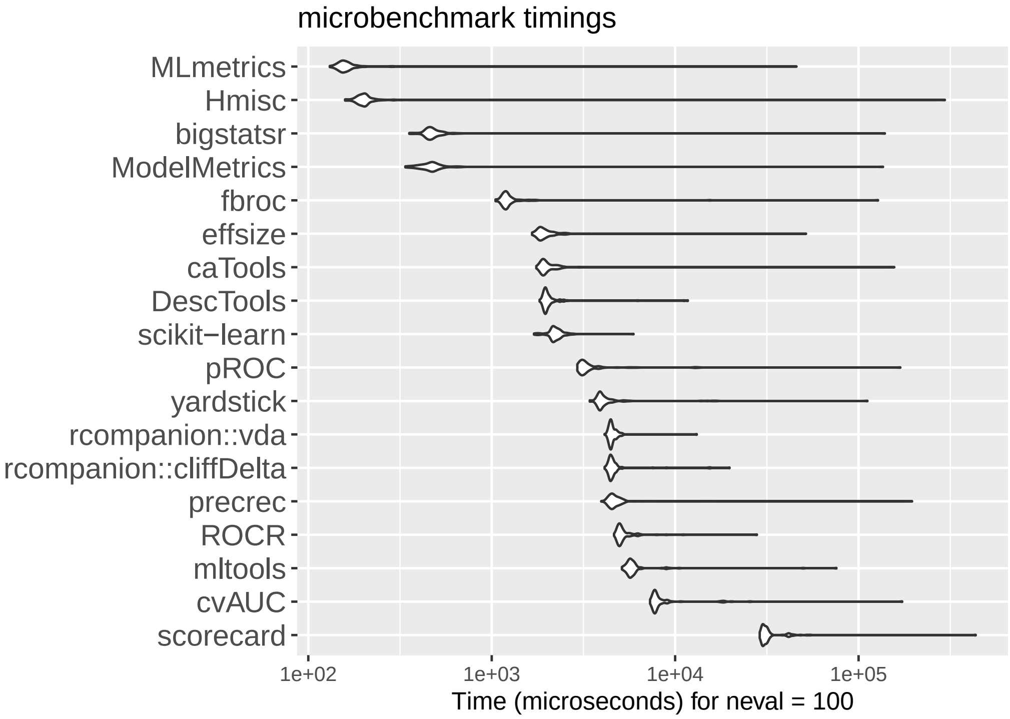
$$\bar{R}_1 = \frac{1}{n_1} \sum_{i:y_i=1} \text{Rank}(s_i)$$

6 Efficiency study

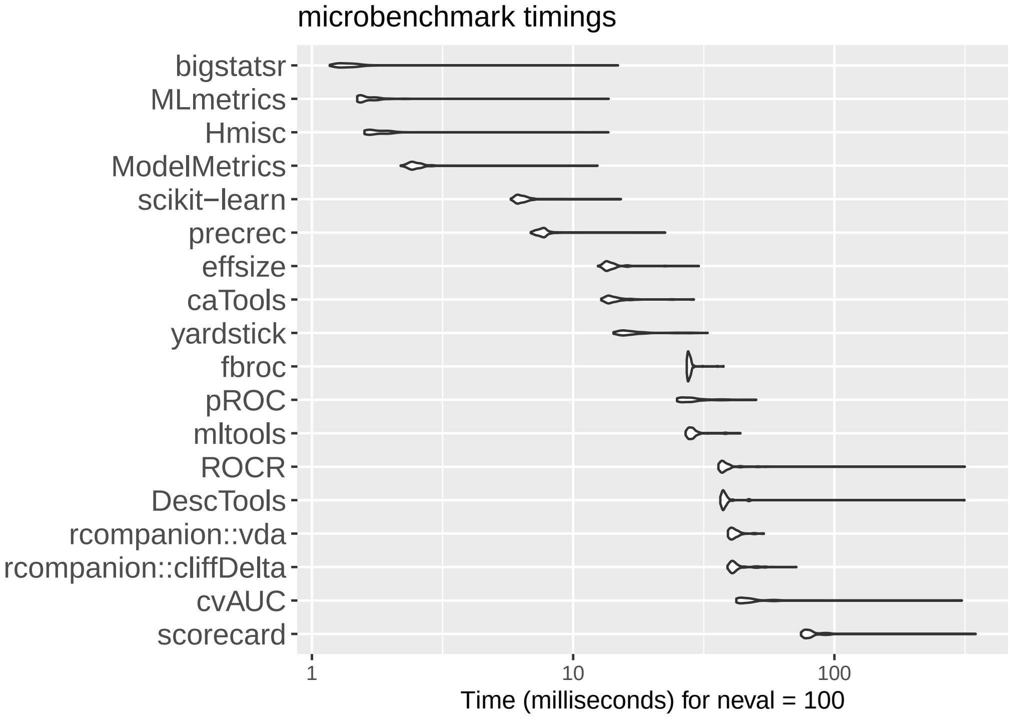
Benchmarking packages

All of the R packages that can be used to calculate the AUC measure have had their performance speed tested. In every benchmark, each function was called 100 times, and there are a total of three benchmarks, each with a different number of observations. The predictor *pred* has been generated from the normal distribution, with the first half of the observations using $N(0,1)$, and the second half being $N(1,1)$. The *target* variable is a vector, where the first half of values are all equal to 1 and the second half of values are all equal to 0.

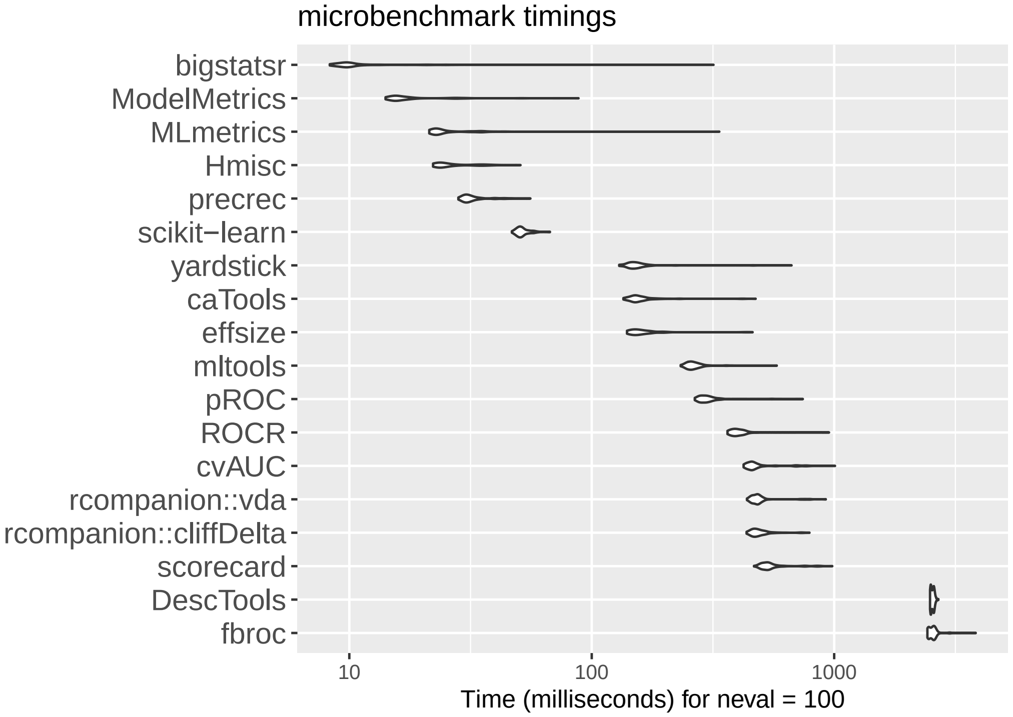
First benchmark - 1000 observations



Second benchmark - 10000 observations



Third benchmark - 100000 observations



The time needed to complete the calculations varies greatly and spans multiple orders of magnitude. For example, using *bigstatsr* to calculate the AUC metrics is almost 100 times faster than by using the *scorecard* package. There may be different causes to this, for example a function might additionally perform other calculations, like generating confidence intervals or also calculating other metrics, or otherwise due to the algorithms used being inefficient, or perhaps even both of these issues at once.

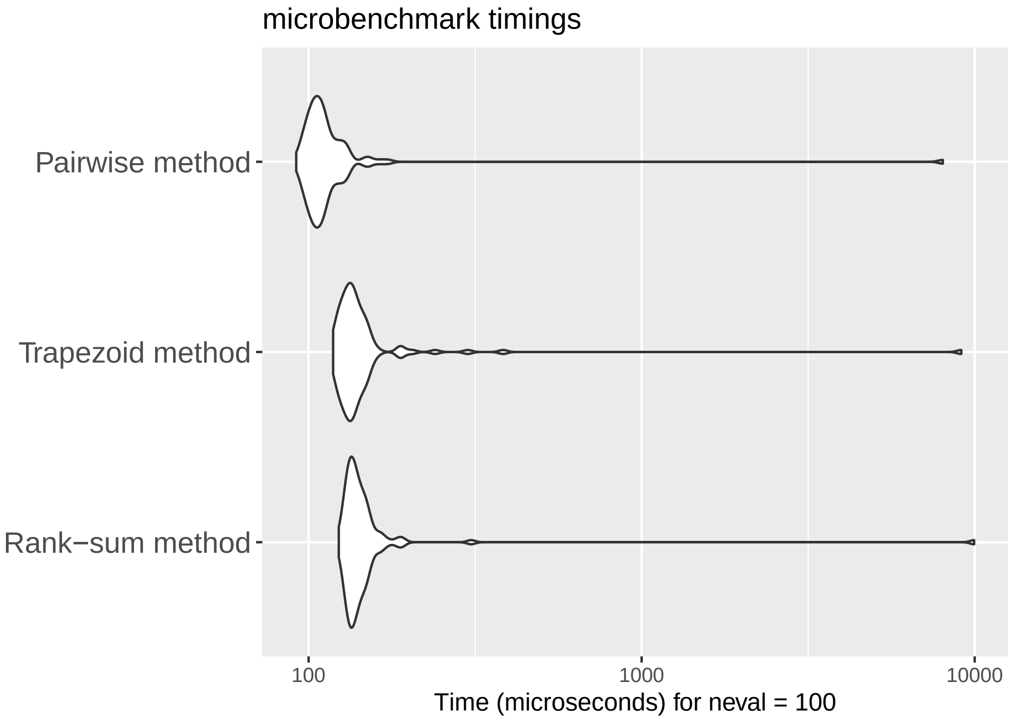
In general, *bigstatsr* offers the fastest way of computation overall. While *MLMetrics* and *Hmisc* are slightly faster on the smallest data set, *bigstatsr* noticeably outperforms them as the number of observations increases.

It is therefore recommended to use the *AUC* function from the *bigstatsr* package whenever calculating the AUC metric in R, whenever an already existing function is desired.

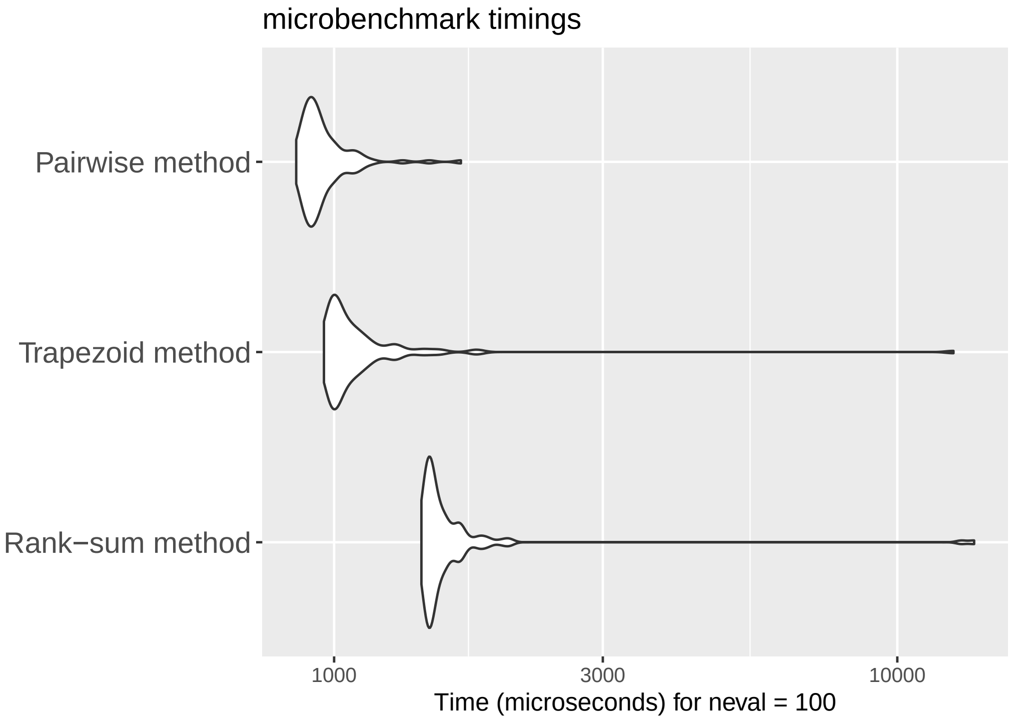
Benchmarking the algorithms

The three identified algorithms have been tested against each other, by creating as optimised functions implementing them as possible. They only contain the calculation, with no checks whether the inputs are correct, to minimise potential interference and isolate the algorithms. They have been performed on the same observations as the package benchmarks.

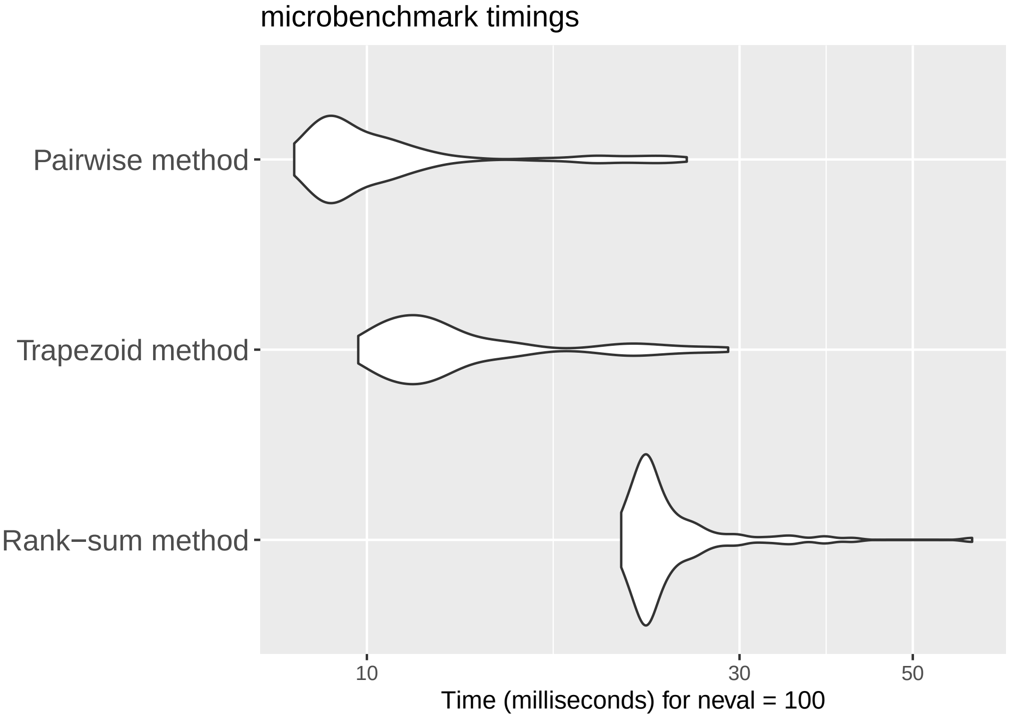
First benchmark - 1000 observations



Second benchmark - 10000 observations



Third benchmark - 100000 observations



In all the benchmarks, the optimised pairwise comparison method is the fastest, with the trapezoid method being only slightly slower. The rank-sum method becomes noticeably slower as the number of observations increases. As a result, whenever developing a new function to calculate the AUC metric, it is recommended to implement it using the pairwise comparison method.

Bibliography

P. C. Austin and E. W. Steyerberg. Interpreting the concordance statistic of a logistic regression model: relation to the variance and odds ratio of a continuous explanatory variable. *BMC Medical Research Methodology*, 12(1):82, Dec. 2012. ISSN 1471-2288. doi: 10.1186/1471-2288-12-82. [p2]

F. Bais and J. Van Der Neut. Adapting the robust effect size cliff's delta to compare behaviour profiles. *Survey Research Methods*, pages 329–352 Pages, Dec. 2022. doi: 10.18148/SRM/2022.V16I2.7908. [p3]

D. Bamber. The area above the ordinal dominance graph and the area below the receiver operating characteristic graph. *Journal of Mathematical Psychology*, 12(4):387–415, Nov. 1975. ISSN 00222496. doi: 10.1016/0022-2496(75)90001-2. [p3]

A. I. Bandos, H. E. Rockette, and D. Gur. A permutation test for comparing roc curves in multireader studies. *Academic Radiology*, 13(4):414–420, Apr. 2006. ISSN 10766332. doi: 10.1016/j.acra.2005.12.012. [p4]

P. Biecek. Dalex: Explainers for complex predictive models in r. *Journal of Machine Learning Research*, 19 (84):1–5, 2018. URL `https://jmlr.org/papers/v19/18-416.html`. [p4]

E. Brunner and U. Munzel. The nonparametric behrens-fisher problem: Asymptotic theory and a small-sample approximation. *Biometrical Journal*, 42(1):17–25, 2000. ISSN 1521-4036. doi: 10.1002/(SICI)1521-4036(200001)42:1\%3C17::AID-BIMJ17\%3E3.0.CO;2-U. [p3]

N. Cliff. Dominance statistics: Ordinal analyses to answer ordinal questions. *Psychological Bulletin*, 114 (3):494–509, 1993. ISSN 1939-1455. doi: 10.1037/0033-2909.114.3.494. [p3]

B. Engelmann, E. Hayden, and D. Tasche. Measuring the discriminative power of rating systems. *SSRN Electronic Journal*, 2003. ISSN 1556-5068. doi: 10.2139/ssrn.2793951. URL `https://www.ssrn.com/abstract=2793951`. [p3]

T. Fawcett. An introduction to roc analysis. *Pattern Recognition Letters*, 27(8):861–874, June 2006. ISSN 01678655. doi: 10.1016/j.patrec.2005.10.010. [p1]

D. J. Hand and R. J. Till. A simple generalisation of the area under the roc curve for multiple class classification problems. 2001. [p3]

J. A. Hanley and B. J. McNeil. The meaning and use of the area under a receiver operating characteristic (roc) curve. *Radiology*, 143(1):29–36, Apr. 1982. ISSN 0033-8419. doi: 10.1148/radiology.143.1.7063747. [p1, 3]

J. Harrell, Frank E., R. M. Califf, D. B. Pryor, K. L. Lee, and R. A. Rosati. Evaluating the yield of medical tests. *JAMA*, 247(18):2543–2546, May 1982. ISSN 0098-7484. doi: 10.1001/jama.1982.03320430047030. [p2]

A. P. Idczak. Remarks on statistical measures for assessing quality of scoring models. *Acta Universitatis Lodziensis. Folia Oeconomica*, 4(343343):21–38, Sept. 2019. ISSN 2353-7663. doi: 10.18778/0208-6018.343.02. [p3]

S. Janitza, C. Strobl, and A.-L. Boulesteix. An auc-based permutation variable importance measure for random forests. *BMC Bioinformatics*, 14(1):119, Dec. 2013. ISSN 1471-2105. doi: 10.1186/1471-2105-14-119. [p4]

M. R. J. Junge and J. R. Dettori. Roc solid: Receiver operator characteristic (roc) curves as a foundation for better diagnostic tests. *Global Spine Journal*, 8(4):424–429, June 2018. ISSN 2192-5682. doi: 10.1177/2192568218778294. [p1]

J. D. Karch. Psychologists should use brunner-munzel's instead of mann-whitney's u test as the default nonparametric procedure. *Advances in Methods and Practices in Psychological Science*, 4(2), 2021. ISSN 2515-2467. doi: 10.1177/2515245921999602. [p3]

V. V. Kharin and F. W. Zwiers. On the roc score of probability forecasts. *Journal of Climate*, 16(24): 4145–4150, Dec. 2003. ISSN 0894-8755, 1520-0442. doi: 10.1175/1520-0442(2003)016\%3C4145:OTRSOP\%3E2.0.CO;2. [p3]

B. Kochański. A simulation model for risk and pricing competition in the retail lending market. *Czech Journal of Economics and Finance*, 71(2):96–118, Oct. 2021. ISSN 2464-7683. doi: 10.32065/CJEF.2021.02.01. [p5]

B. Kochański. Which curve fits best: Fitting roc curve models to empirical credit-scoring data. *Risks*, 10(10):184, Sept. 2022. ISSN 2227-9091. doi: 10.3390/risks10100184. [p5]

E. LeDell, M. Petersen, and M. Van Der Laan. Computationally efficient confidence intervals for cross-validated area under the roc curve estimates. *Electronic Journal of Statistics*, 9(1), Jan. 2015. ISSN 1935-7524. doi: 10.1214/15-EJS1035. URL https://projecteuclid.org/journals/electronic-journal-of-statistics/volume-9/issue-1/Computationally-efficient-confidence-intervals-for-cross-validated-area-under-the/10.1214/15-EJS1035.full. [p4]

S. J. Mason and N. E. Graham. Conditional probabilities, relative operating characteristics, and relative operating levels. *Weather and Forecasting*, 14(5):713–725, Oct. 1999. ISSN 1520-0434, 0882-8156. doi: 10.1175/1520-0434(1999)014\%3C0713:CPROCA\%3E2.0.CO;2. [p3]

K. O. McGraw and S. P. Wong. A common language effect size statistic. *Psychological Bulletin*, 111(2): 361–365, 1992. ISSN 1939-1455. doi: 10.1037/0033-2909.111.2.361. [p3]

R. Newson. Parameters behind "nonparametric" statistics: Kendall's tau, somers' d and median differences. *The Stata Journal: Promoting communications on statistics and Stata*, 2(1):45–64, Mar. 2002. ISSN 1536-867X, 1536-8734. doi: 10.1177/1536867X0200200103. [p3]

H. Noma, Y. Matsushima, and R. Ishii. Confidence interval for the auc of sroc curve and some related methods using bootstrap for meta-analysis of diagnostic accuracy studies. *Communications in Statistics: Case Studies, Data Analysis and Applications*, 7(3):344–358, July 2021. ISSN 2373-7484. doi: 10.1080/23737484.2021.1894408. [p4]

M. Pauly, T. Asendorf, and F. Konietschke. Permutation-based inference for the auc: A unified approach for continuous and discontinuous data. *Biometrical Journal*, 58(6):1319–1337, Nov. 2016. ISSN 0323-3847, 1521-4036. doi: 10.1002/bimj.201500105. [p4]

J. M. d. l. Rubia. Note on rank-biserial correlation when there are ties. *Open Journal of Statistics*, 12(55): 597–622, Oct. 2022. doi: 10.4236/ojs.2022.125036. [p3]

J. R. Sobehart, S. C. Keenan, and R. M. Stein. *Benchmarking Quantitative Default Risk Models: A Validation Methodology*. 2000. URL https://www.researchgate.net/publication/243771528_Benchmarking_Quantitative_Default_Risk_Models_A_Validation_Methodology. [p3]

R. H. Somers. A new asymmetric measure of association for ordinal variables. *American Sociological Review*, 27(6):799, Dec. 1962. ISSN 00031224. doi: 10.2307/2090408. [p3]

A. Vargha and H. D. Delaney. A critique and improvement of the "cl" common language effect size statistics of mcgraw and wong. *Journal of Educational and Behavioral Statistics*, 25(2):101–132, 2000. ISSN 1076-9986. doi: 10.2307/1165329. [p3]

H. W. Wendt. Dealing with a common problem in social science: A simplified rank-biserial coefficient of correlation based on the u statistic. *European Journal of Social Psychology*, 2(4):463–465, 1972. ISSN 1099-0992. doi: 10.1002/ejsp.2420020412. [p3]

T. Yang and Y. Ying. Auc maximization in the era of big data and ai: A survey. *ACM Computing Surveys*, 55(8):1–37, Aug. 2023. ISSN 0360-0300, 1557-7341. doi: 10.1145/3554729. [p4]

Błażej Kochański

Przemysław Peplinski
Gdańsk University of Technology
Faculty of Management and Economics
Gdańsk

Miriam Nieslona

Wiktor Galewski

Piotr Geremek