# 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.
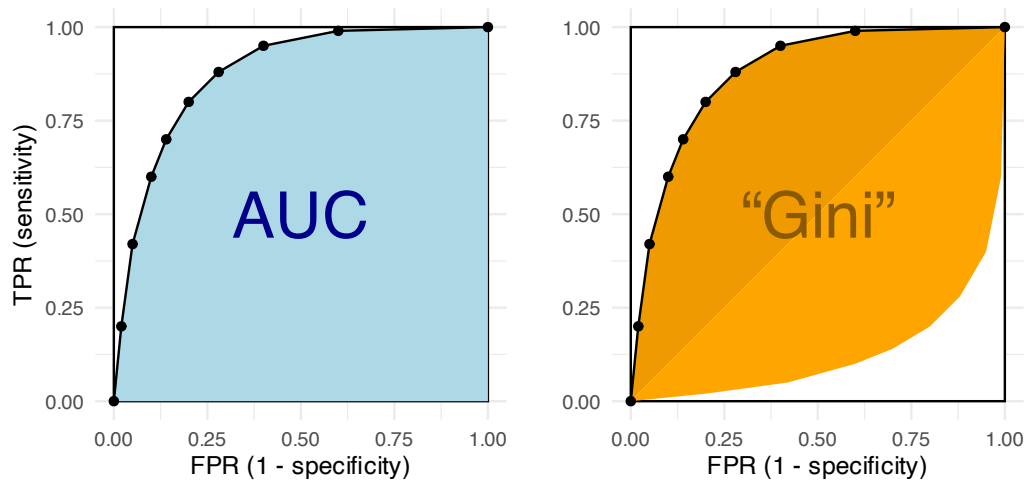
## 2 ROC curve and AUC

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 (Anderson, 2007; Řezáč and Řezáč, 2011). 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 (2010) denotes AUC by $\alpha$ and the Gini coefficient by $\Delta$, referring to both as measures of *stochastic superiority*. Cliff (1993) describes the effect size measure $d$ (or $\delta$ for the population), which adopts 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 Motivation for computationally efficient AUC calculation

A single calculation of the AUC is often insufficient for a comprehensive assessment of model quality. More advanced inference, validation, optimization, feature evaluation or simulation procedures require repeated computation of the AUC, placing substantial demands on the computational efficiency of the underlying algorithms.

Resampling methods are commonly used for inference on the area under the ROC curve. Confidence intervals for the AUC can be obtained using bootstrapping or cross-validation (Noma et al., 2021; Gu et al., 2008; LeDell et al., 2015). Permutation tests enable the assessment of statistical significance, either under the null hypothesis of no discriminatory power or when comparing the performance of two models (Pauly et al., 2016; Venkatraman, 2000; Bandos et al., 2006). Resampling procedures are essential for evaluating uncertainty, model stability, and statistical significance. However, despite their favorable statistical properties, including robustness (Simonoff et al., 1986; Pauly et al., 2016), these methods rely on repeated recalculation of the AUC and can therefore impose considerable computational burdens, particularly for large datasets or complex models (LeDell et al., 2015).

AUC maximization is another example of computational challenge that has recently gained importance in the development of ranking and classification algorithms. Directly optimizing AUC, rather than surrogate losses, leads to improved generalization and discrimination capabilities in large-scale settings (Yuan et al., 2021; Zhang and Xu, 2022; LeDell et al., 2016). Unlike many other objectives, the AUC function is non-decomposable over individual examples, which complicates the use of standard optimization techniques (Zhou et al., 2020). A comprehensive overview of the last 20+ years of AUC maximization research, emphasizing its relevance for imbalanced learning, large datasets, and deep learning applications has been provided by Yang and Ying (2023). Many, if not most, approaches to AUC maximization require repeated multiple computations of the AUC statistic (see, for example Calders and Jaroszewicz, 2007; Yang et al., 2020; Ataman et al., 2006; Gajowniczek and Ząbkowski, 2021).

Efficient AUC computation is also a prerequisite for many feature selection and assessment techniques. Permutation-based variable importance measures were, for example, proposed by Janitza et al. (2013) or Biecek (2018), while AUC-based feature selection algorithms were, inter alia, presented by Sun et al. (2017); Su et al. (2022), Xu and Suzuki (2014) or Ribeiro et al. (2021). Other applications of repeated AUC computation arise in simulation studies. For instance, Kochański (2021) introduced a simulation model for risk and pricing competition in which each bank optimizes its credit scoring strategy, requiring numerous AUC evaluations to assess performance under varying market conditions.

In summary, efficient and scalable computation of AUC has become a foundational requirement not only for model evaluation but also for advanced inference, optimization, feature selection, and simulation studies. As AUC-based methods continue to expand into areas such as deep models and large-scale applications, the computational complexity of repeated AUC calculations remains a critical challenge.

## 5 AUC computation implementation in R packages

Table 1 presents a selection of R packages and functions commonly used to calculate the area under the ROC curve (AUC). For each package, the table lists the primary function, typical usage syntax, the computational method employed, and the programming language(s) in which the core implementation is written. The computational methods range from rank-based approaches and pairwise comparisons to trapezoidal integration, reflecting different trade-offs between accuracy and efficiency. Some functions rely entirely on R, while others leverage compiled code in C++ for improved performance. It is important to note that certain functions perform additional checks or compute supplementary metrics beyond the AUC itself, which may reduce computational efficiency when only the AUC is needed. The practical impact of these differences will be examined empirically in the following section.

**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 |

Based on the reviewed functions, we identified three main types of algorithms for AUC computation. These approaches differ in how they process predicted scores and true labels, 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 $s$ 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 or fails to 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.

The three computational approaches described above can be expressed formally as follows:

**1. Trapezoidal integration over the ROC curve**

This approach computes the AUC as the area under the empirical ROC curve by applying the trapezoidal rule:

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

where $m$ is the number of distinct score thresholds from lowest to the highest, $TPR_k = TP_k/n_1$ is the true positive rate, $FPR_k = FP_k/n_0$ is the false positive Rate at threshold $k$, $TP_k$ is the number of true positives, $\text{FP}_k$ is the number of false positives, $n_1$ is the number of positive cases, $n_0$ is the number of negative cases.

**2. Optimized pairwise comparison**

In this formulation, the AUC is based on the probabilitic (pairwise comparison) definition, that is the probability that a randomly chosen positive instance receives a higher score than a randomly chosen negative instance, with 0.5 applied to ties.

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

**3. Rank-based formula**

The rank-based formula for AUC is directly based on the method developed for the Mann–Whitney U statistic computation (see Sheskin, 2020):

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

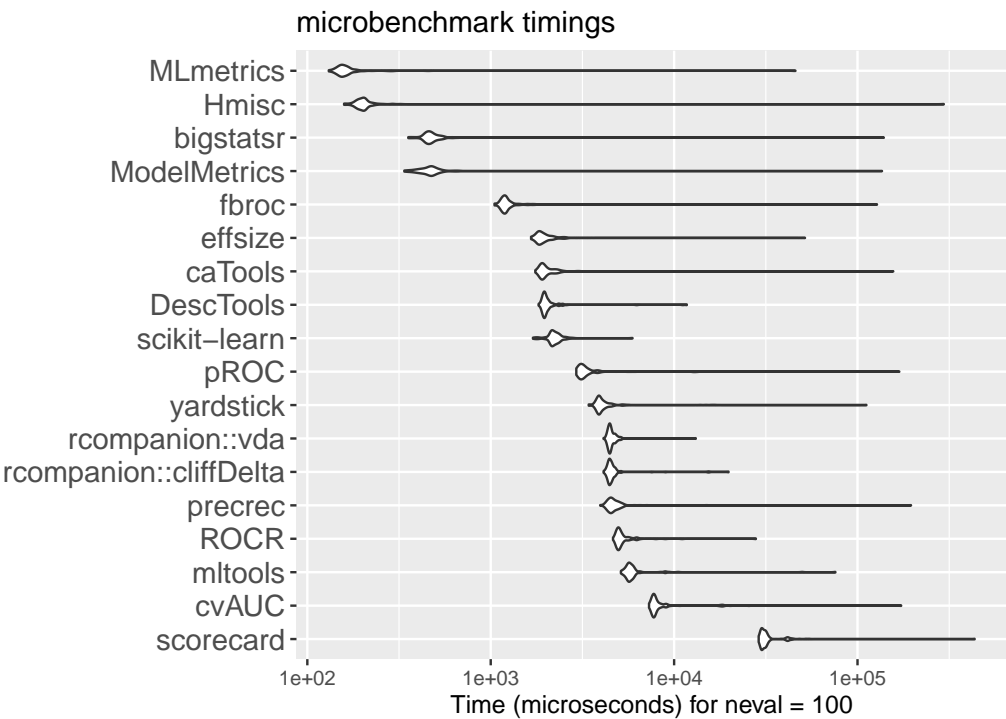where $\bar{R}_1$ is the mean rank for $s_i$ where $y_i = 1$:

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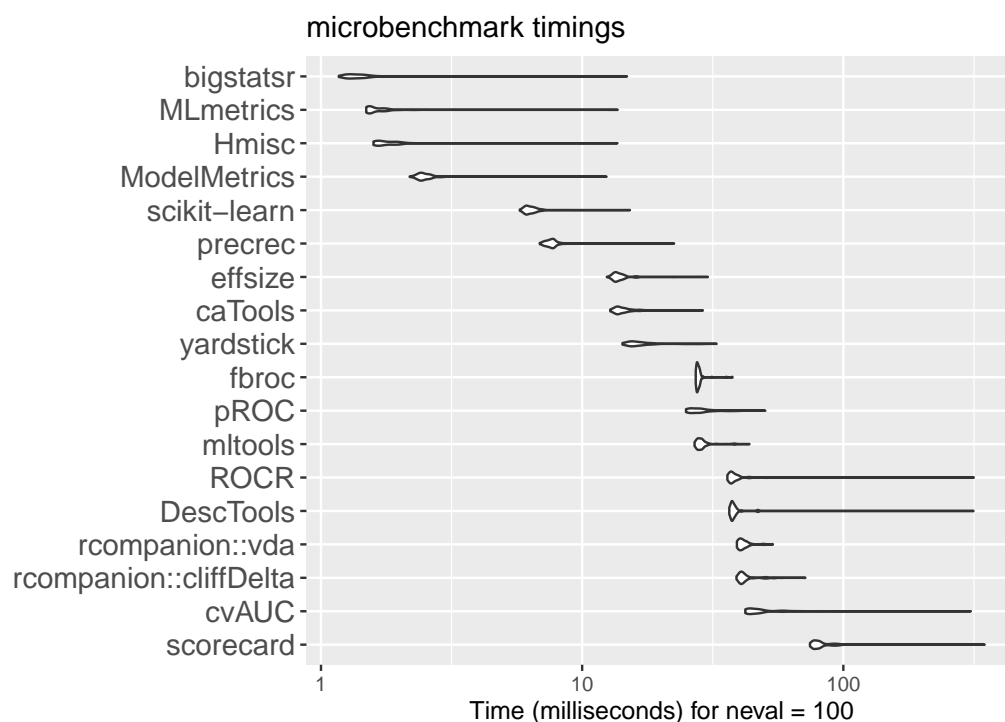
## 6 Efficiency study

**Performance of existing R Packages**

The computational efficiency of several R packages for AUC computation was evaluated through a series of benchmarks. In each benchmark, every AUC function was invoked 100 times. Experiments were conducted on three simulated datasets of increasing size, containing 1,000, 10,000, and 100,000 observations. The predictor variable *pred* was synthetically generated such that the first half of the observations followed a $\mathcal{N}(0,1)$ distribution, while the remaining observations followed a $\mathcal{N}(1,1)$ distribution. The corresponding response variable was binary, with the first half of the observations assigned label 1 and the second half assigned label 0.
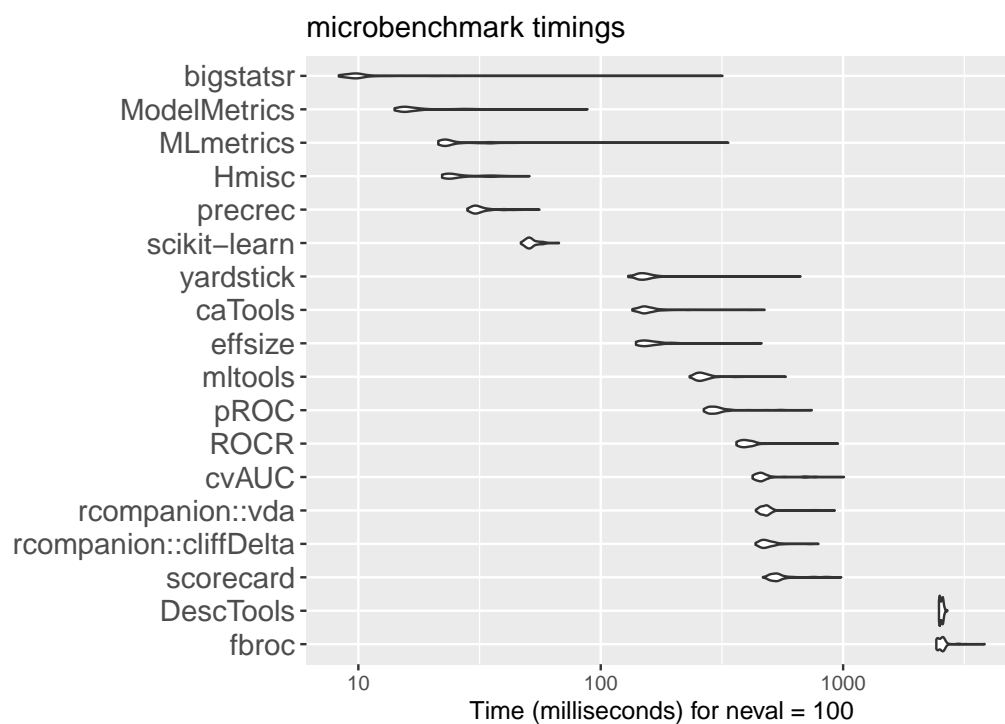
**First benchmark - 1000 observations**



**Second benchmark - 10000 observations**

**Third benchmark - 100000 observations**



The computational time required to evaluate the AUC metric varies substantially across the examined implementations, spanning several orders of magnitude. For instance, the *bigstatsr* implementation computes the AUC nearly two orders of magnitude faster than the corresponding function in the *scorecard* package. This discrepancy may be attributable to several factors, including additional computations performed internally by some functions (e.g., estimation of confidence intervals or simultaneous evaluation of multiple performance metrics), differences in algorithmic design, or a combination of these effects.
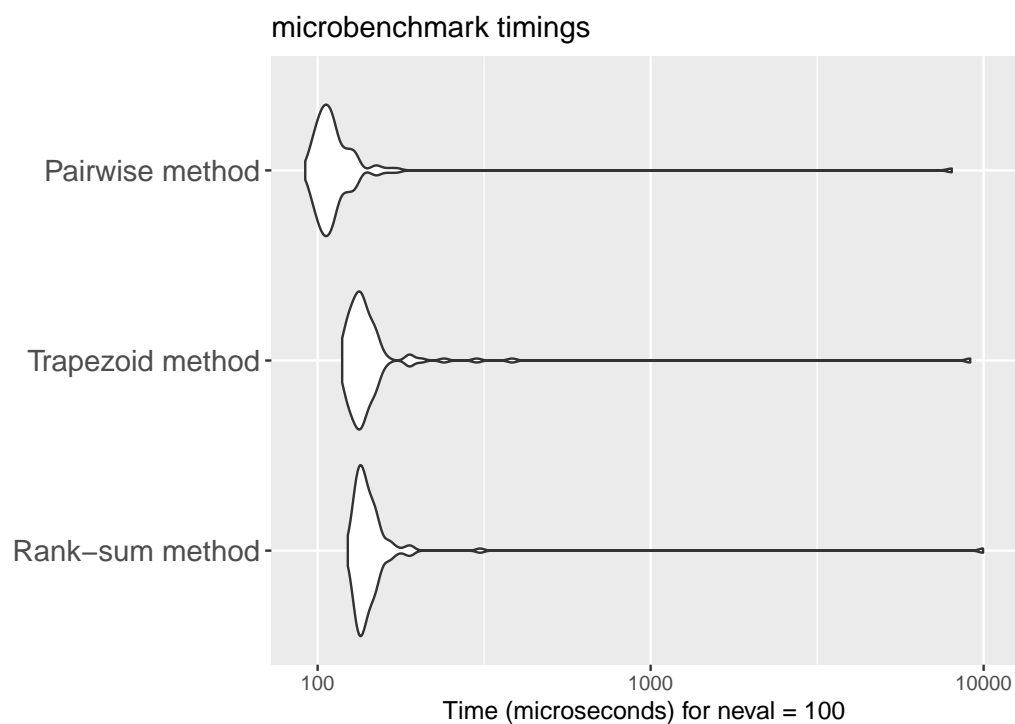
Overall, *bigstatsr* consistently provides the fastest AUC computation. Although *MLMetrics* and *Hmisc* exhibit slightly lower execution times on the smallest dataset, *bigstatsr* clearly outperforms these alternatives as the number of observations increases.

Consequently, when an existing R implementation is preferred, we recommend the use of the *AUC* function from the *bigstatsr* package for efficient computation of the AUC metric, particularly in large-scale settings.
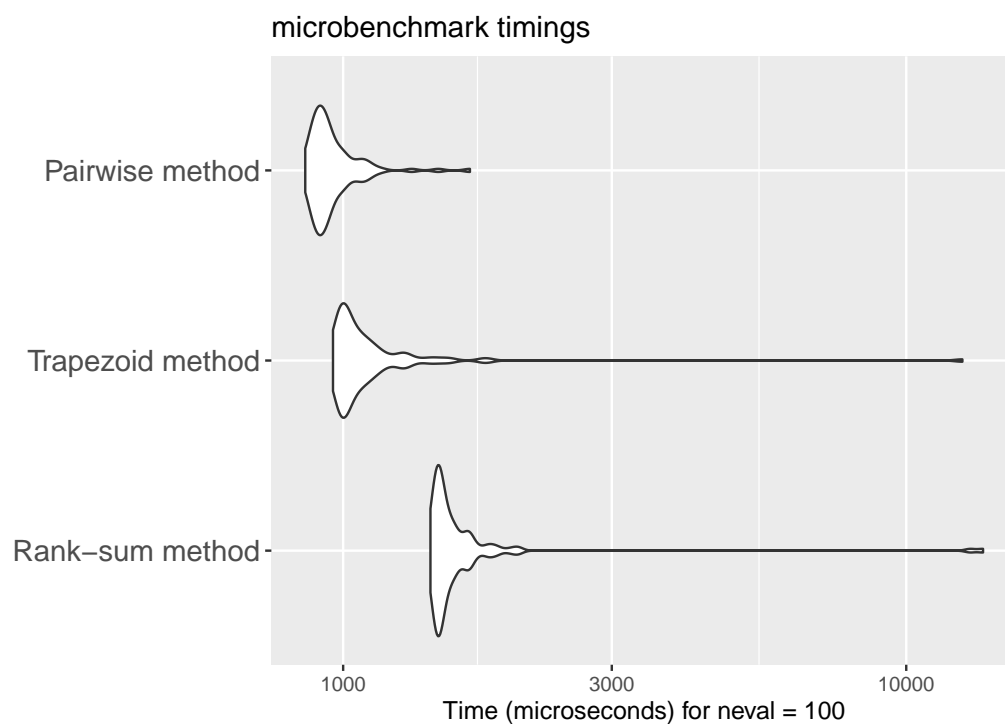
## Benchmarking custom R implementations

The three identified algorithms were evaluated through direct comparison using custom R implementations designed to be as computationally efficient as possible. Each implementation was restricted to the core AUC computation and deliberately omitted input validation and auxiliary checks in order to minimize overhead and isolate algorithmic performance. All benchmarks were conducted on the same datasets used in the package-level experiments.
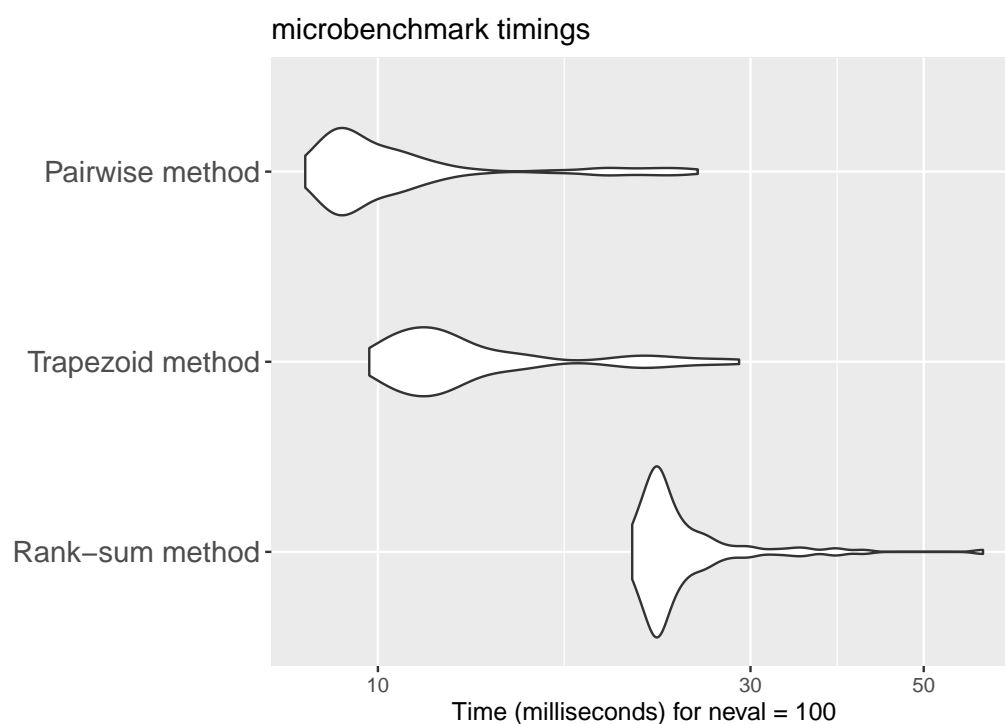
**First benchmark - 1000 observations**



**Second benchmark - 10000 observations**

microbenchmark timings

**Third benchmark - 100000 observations**



microbenchmark timings

   Across all benchmark settings, the optimized pairwise comparison method consistently exhibited the shortest execution time, with the trapezoidal method performing only marginally slower. In contrast, the rank-sum method showed a pronounced increase in computational cost as the number of observations grew.

   These results suggest that, when developing new R functions for AUC computation, the pairwise comparison approach provides the most efficient algorithmic foundation, particularly in large-scale scenarios.

# Bibliography

A. Agresti. *Analysis of Ordinal Categorical Data*. John Wiley & Sons, Apr. 2010. ISBN 978-0-470-08289-8. Google-Books-ID: VVIe4BPDR7kC. [p3]

R. Anderson. *The Credit Scoring Toolkit: Theory and Practice for Retail Credit Risk Management and Decision Automation*. OUP Oxford, Oxford, Aug. 2007. ISBN 978-0-19-922640-5. Google-Books-ID: 7LlGfPvOJLoC. [p3]

K. Ataman, W. N. Street, and Y. Zhang. Learning to rank by maximizing auc with linear programming. In *The 2006 IEEE International Joint Conference on Neural Network Proceedings*, page 123–129, 2006. doi: 10.1109/IJCNN.2006.246669. URL https://ieeexplore.ieee.org/document/1716080. [p4]

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]

T. Calders and S. Jaroszewicz. Efficient auc optimization for classification. In J. N. Kok, J. Koronacki, R. Lopez de Mantaras, S. Matwin, D. Mladenič, and A. Skowron, editors, *Knowledge Discovery in Databases: PKDD 2007*, Lecture Notes in Computer Science, page 42–53, Berlin, Heidelberg, 2007. Springer. ISBN 978-3-540-74976-9. doi: 10.1007/978-3-540-74976-9_8. [p4]

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]

K. Gajowniczek and T. Ząbkowski. Imbtreeauc: An r package for building classification trees using the area under the roc curve (auc) on imbalanced datasets. *SoftwareX*, 15:100755, 2021. ISSN 2352-7110. doi: 10.1016/j.softx.2021.100755. [p4]

J. Gu, S. Ghosal, and A. Roy. Bayesian bootstrap estimation of roc curve. *Statistics in Medicine*, 27(26):5407–5420, 2008. ISSN 1097-0258. doi: 10.1002/sim.3366. [p4]

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. [p4]

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. [p4]

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]

E. LeDell, M. J. van der Laan, and M. Peterson. Auc-maximizing ensembles through metalearning. *The international journal of biostatistics*, 12(1):203–218, May 2016. ISSN 2194-573X. doi: 10.1515/ijb-2015-0035. [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. [p4]

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]

G. Ribeiro, C. Gonçalves, P. Victor dos Santos, and R. Melgaço Barbosa. Feature selection method auc-based with estimation probability and smoothing. In *2021 International Conference on Engineering and Emerging Technologies (ICEET)*, page 1–8, Oct. 2021. doi: 10.1109/ICEET53442.2021.9659733. URL https://ieeexplore.ieee.org/abstract/document/9659733. [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]

D. J. Sheskin. *Handbook of Parametric and Nonparametric Statistical Procedures, Fifth Edition*. Chapman and Hall/CRC, New York, 5 edition, 2020. ISBN 978-0-429-18619-6. doi: 10.1201/9780429186196. [p5]

J. S. Simonoff, Y. Hochberg, and B. Reiser. Alternative estimation procedures for pr(x < y) in categorized data. *Biometrics*, 42(4):895–907, 1986. ISSN 0006-341X. doi: 10.2307/2530703. [p4]

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]

Y. Su, K. Du, J. Wang, J.-m. Wei, and J. Liu. Multi-variable auc for sifting complementary features and its biomedical application. *Briefings in Bioinformatics*, 23(2):bbac029, Mar. 2022. ISSN 1477-4054. doi: 10.1093/bib/bbac029. [p4]

L. Sun, J. Wang, and J. Wei. Avc: Selecting discriminative features on basis of auc by maximizing variable complementarity. *BMC Bioinformatics*, 18(3):50, Mar. 2017. ISSN 1471-2105. doi: 10.1186/s12859-017-1468-4. [p4]

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]

E. S. Venkatraman. A permutation test to compare receiver operating characteristic curves. *Biometrics*, 56(4):1134–1138, 2000. ISSN 1541-0420. doi: 10.1111/j.0006-341X.2000.01134.x. [p4]

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]

J.-W. Xu and K. Suzuki. Max-auc feature selection in computer-aided detection of polyps in ct colonography. *IEEE Journal of Biomedical and Health Informatics*, 18(2):585–593, Mar. 2014. ISSN 2168-2208. doi: 10.1109/JBHI.2013.2278023. [p4]

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]

Z. Yang, B. Zhou, Y. Lei, and Y. Ying. Stochastic hard thresholding algorithms for auc maximization. In *2020 IEEE International Conference on Data Mining (ICDM)*, page 741–750, Nov. 2020. doi: 10.1109/ICDM50108.2020.00083. URL https://ieeexplore.ieee.org/abstract/document/9338276. [p4]

Z. Yuan, Y. Yan, M. Sonka, and T. Yang. Large-scale robust deep auc maximization: A new surrogate loss and empirical studies on medical image classification. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, page 3020–3029, Montreal, QC, Canada, Oct. 2021. IEEE. ISBN 978-1-6654-2812-5. doi: 10.1109/ICCV48922.2021.00303. URL https://ieeexplore.ieee.org/document/9710509/. [p4]

X.-L. Zhang and M. Xu. Auc optimization for deep learning-based voice activity detection. *EURASIP Journal on Audio, Speech, and Music Processing*, 2022(1):27, Oct. 2022. ISSN 1687-4722. doi: 10.1186/s13636-022-00260-9. [p4]

B. Zhou, Y. Ying, and S. Skiena. Online auc optimization for sparse high-dimensional datasets. In *2020 IEEE International Conference on Data Mining (ICDM)*, page 881–890, Nov. 2020. doi: 10.1109/ICDM50108.2020.00097. URL https://ieeexplore.ieee.org/abstract/document/9338419. [p4]

M. Řezáč and F. Řezáč. How to measure the quality of credit scoring models. *Czech Journal of Economics and Finance (Finance a úvěr)*, 61(5):486–507, 2011. [p3]

*Błażej Kochański*




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




*Miriam Nieslona*

*Wiktor Galewski*

*Piotr Geremek*