

Article

A Probability-Based Models Ranking Approach: An Alternative Method of Machine-Learning Model Performance Assessment

Stanisław Gajda and Marcin Chlebus * 

Faculty of Economic Sciences, University of Warsaw, Długa Street 44/50, 00-241 Warsaw, Poland

* Correspondence: mchlebus@wne.uw.edu.pl

Abstract: Performance measures are crucial in selecting the best machine learning model for a given problem. Estimating classical model performance measures by subsampling methods like bagging or cross-validation has several weaknesses. The most important ones are the inability to test the significance of the difference, and the lack of interpretability. Recently proposed Elo-based Predictive Power (EPP)—a meta-measure of machine learning model performance, is an attempt to address these weaknesses. However, the EPP is based on wrong assumptions, so its estimates may not be correct. This paper introduces the Probability-based Ranking Model Approach (PMRA), which is a modified EPP approach with a correction that makes its estimates more reliable. PMRA is based on the calculation of the probability that one model achieves a better result than another one, using the Mixed Effects Logistic Regression model. The empirical analysis was carried out on a real mortgage credits dataset. The analysis included a comparison of how the PMRA and state-of-the-art k-fold cross-validation ranked the 49 machine learning models, an example application of a novel method in hyperparameters tuning problem, and a comparison of PMRA and EPP indications. PMRA gives the opportunity to compare a newly developed algorithm to state-of-the-art algorithms based on statistical criteria. It is the solution to select the best hyperparameters configuration and to formulate criteria for the continuation of the hyperparameters space search.

Keywords: machine learning; model performance assessment; model selection; hyperparameters tuning; model performance measures; Elo-based Predictive Power; mixed effects logistic regression



Citation: Gajda, S.; Chlebus, M. A Probability-Based Models Ranking Approach: An Alternative Method of Machine-Learning Model Performance Assessment. *Sensors* **2022**, *22*, 6361. <https://doi.org/10.3390/s22176361>

Academic Editors: Aris Leivadreas, Vasileios Karyotis and Dimitrios Dechouniotis

Received: 21 July 2022

Accepted: 18 August 2022

Published: 24 August 2022

Publisher's Note: MDPI stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.



Copyright: © 2022 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

1. Introduction and Literature Overview

Model performance assessment plays a crucial role in machine learning. The applied comparison procedure matters during model selection for a given problem, in hyperparameters tuning, and while testing the newly proposed algorithm. Depending on the characteristic of the problem and expected properties, different measures are applied [1]. Among those most commonly used in recent years are: Accuracy [2], F1 [3], cross-entropy [4], AUC [5] for classification and RMSE [6], MSE [7], MAE [8] for regression. The classic approach divides available datasets into train, validation, and test sets. A disadvantage of this method is that obtained measure estimate has a high variance. In order to reduce it, subsampling methods are applied. Cross-validation and bootstrap aggregating (bagging) are most often used. In [9], it is claimed that many commonly used subsampling schemes suffer from substantial negative bias when considering AUC in the small-sample setting. The variance of the obtained results is also an important issue. In [10], it is shown that there exists no universal (valid under all distributions) estimator of the variance of cross-validation.

In [11], the authors described the two most often used approaches to calculate AUC and different measures over folds: *pooling* and *averaging*. In pooling, the results obtained in each fold are joined into one set on which the AUC measure is calculated. Averaging assumes the calculation of the mean value of AUC measures obtained in all folds. Parker et al. show in their study on low-signal simulated data and Van't Veer breast cancer

dataset that pooling may cause substantial biases due to comparing observations from different test samples [12]. They explain why using k-fold cross-validation modifications: *Balanced cv*, *Stratified cv*, *Balanced leave-one-out cv* reduce the bias for small datasets. In [13], Varoquaux states that the use of cross-validation on small sample datasets leads to large error bars. The simulation that was carried out in a biological context on small sample data showed that the cross-validation approach provides inaccurate results—with, e.g., a 10% error bar for accuracy measure on a sample with 100 observations. In [14], Krzanowski proposed an approach that combines the advantages of pooling and averaging—*Leave-pair-out cross-validation* (LPOCV). It involves calculating AUC values for each positive-negative pair. The final AUC value is obtained by averaging through all of these pairs. Efficient LPOCV implementation was developed by Pahikkala et al. [15]. The *Leave-one-out cross-validation* (LOOCV) is equivalent to LPO for measures other than AUC. The limitation of both algorithms is their complexity—the required number of training rounds is of the order: $O(n^2)$, $O(n)$ for LPOCV and LOOCV respectively, where n is a number of observations. Therefore these approaches may be applied only to small datasets. LPOCV and LOOCV, similarly to pooling, make maximal use of the available dataset. However, these methods are not perfect—in [16], Gronau et al. show their limitations.

An essential aspect of the model selection process is testing the significance of the performance difference between two models. In [17], the authors conducted an extensive review of the statistical methods used for model selection. They described several methods to compare algorithms, including fitting and evaluating models via cross-validation. The first of the presented methods to test performance difference significance is the difference of the proportions test. This involves comparing confidence intervals of two accuracies using a z-score. However, Dietterich in his simulation study states that this test tends to have a high false positive rate [18]. Instead, Dietterich recommends using McNemar's test [19]. The test statistic is calculated from the number of observations for which models' predictions were different. p -values are calculated based on the Chi-square distribution. When the number of differences in predictions is relatively small, the Chi-squared distribution may be imprecise. In this case, the binomial test is recommended.

Methods mentioned above: *Difference of proportions test*, *McNemar's test* and *Binomial test* all assume specific threshold level, which determines the models' confusion matrix. What is more, these methods compare two models. Comparison of multiple models' performance can be made using *Cochran's Q test*. It verifies the null hypothesis that accuracy of n models is equal, but it does not indicate which models differ [20].

Resampling-based testing procedures are also of practical importance for the model selection problem. One of them is Dietterich's *5 × 2-Fold Cross-Validated Paired t-Test* that assumes performing 5 iterations of 2-fold cross-validation [18]. The more robust alternative: *Combined 5 × 2 cv F-test* was proposed by Alpaydm [21]. In [22], the authors presented machine learning models comparison methods, which are most commonly used in practice. Beyond the above-mentioned *5 × 2-Fold Cross-Validation Paired t-Test* and *McNemar's test* there is *paired Student's t-test* described with further correction proposed in [23]. Although the approaches discussed above give accurate results, they are not often used. They are an addition to the classical approach of averaging and are not directly embedded in it.

Elo-based Predictive Power (EPP)—a measure introduced in [24] is the answer to these drawbacks of classical approaches to model performance assessment. It assumes that the performance of every pair of models is compared. The outcome of the comparison is the input to the logit model. The logit model coefficients indicate the predictive ability of the analyzed models. This approach is novel as it considers all models at once in comparison. Furthermore, the estimated values of the measure are interpretable and it is possible to use a statistical test to investigate the significance of the difference between two models' performance. However, the EPP does not consider the lack of independence of comparisons made on the same subsample. Therefore, its estimates are based on fundamentally flawed assumptions, underestimating results variance, which makes its indications potentially unreliable. EPP aggregates relative information about models' pairs comparison and turns

it into one value per model describing their predictive power. The different approach to decision making in context of subjective information is presented in [25,26].

The aim of this article is to introduce the Probability-based Models Ranking Approach (PMRA), a method of evaluating the performance of a machine learning model, which makes it possible to interpret the difference in the performance of 2 models and to test the statistical significance of this difference, the method is also more reliable than the state-of-the-art averaging approach. PMRA is a modification of the EPP that retains its advantages over the state-of-the-art approach and, at the same time, addresses its main disadvantage. Using a mixed effects model gives a more realistic estimate of the variance, which allows greater confidence in indications of statistical tests stating whether two models' performances are significantly different or not. In the study, PMRA is practically applied to the results of various machine learning models obtained on the Multifamily mortgages dataset describing real credits acquired by Fannie Mae in the years 2000–2020.

The paper is organized as follows: in the next section, the main limitations of the subsampling approach to model performance assessment are described. In Section 3, EPP measure construction, based on the logit model, is presented with its weaknesses. Then conception of the Probability-based Models Ranking Approach (PMRA) using mixed effects logistic regression is introduced. There is a discussion of statistical tests that can be used in the PMRA—Wald and Likelihood-ratio tests. Then we described possible practical applications of this method in hyperparameters tuning and in models' ranking creation. Section 4 contains empirical analysis. It begins with the used dataset description, then PMRA and state-of-the-art averaging approaches are compared in the most common practical applications. Finally, the EPP and PMRA approaches are compared, showing the practical advantages of PMRA over EPP.

2. Weaknesses of Classical Model Assessment Methods Based on Cross-Validation

The most commonly used subsampling methods of evaluating model performance are cross-validation and bootstrap aggregating (bagging). The idea behind both of them is to generate from whole dataset (X, Y) n subsamples $(X_1, Y_1), \dots, (X_n, Y_n)$. Cross-validation assumes dividing (X, Y) into n equal parts and placing in each subsample all of them except one [27]. Bagging involves generating new training datasets, by sampling from (X, Y) uniformly and with replacement (some observation may be chosen multiple times) [28]. Then the algorithm is trained on each dataset and tested on observations not belonging to it. Values r_1, \dots, r_n of one of classical measures (e.g., AUC , F_1 for classification, MSE , MAE for regression) are obtained and then averaged.

$$\bar{r} = \frac{r_1 + \dots + r_n}{n} \quad (1)$$

The resulting value \bar{r} stands for model performance over a dataset. Inequality between two models' average results is the basis of the statement that one model can learn dependencies between independent and dependent variables better than the other one [29]. In this section, we will present four main weaknesses of using this classical approach for the AUC measure example.

2.1. Lack of Interpretability

Most commonly used performance measures are interpretable. For example, AUC represents the probability that a model assigned a higher probability of belonging to a positive class to a randomly chosen positive example than to a randomly chosen negative example [30]. Therefore, the difference between the two models' performances may be interpreted. However, for many classic measures, the subsampling approach, assuming averaging measure values obtained in specific samples, suffers from biases [31]. This problem is particularly relevant for small sample problems [9]. The reasons mentioned make it impossible to rely on the interpretation of the results.

2.2. Lack of Statistical Tools to Measure the Statistical Significance of Results Differences

For many performance measures, it is possible to test the significance of the difference between two models' performances. However, making conclusions about the significance of performance difference based only on the obtained average value of this measure is a more complex issue. For example for AUC, there are statistical tests to conclude about equality of two AUC results [32,33]. However, their use to compare mean AUC would be incorrect as they are based on ROC curve analysis. As described in the literature overview—the most commonly used methods (Paired *t*-test, 5×2 folds *cv*) are imperfect due to their pairwise character.

2.3. Sensitivity on Outliers

Conclusions based on cross-validation average AUC can also be misleading when one of the model's obtained AUC scores stands out greatly from the others. It may result in having such a strong impact on mean AUC that it does not reflect the expected relation between the two models' results. The model which performs worse in most cases may be considered better because of just one outlying observation. Figure 1 presents results obtained by 2 models in 5-fold cross-validation. *M1* model's results are better in 4 of 5 folds. The extraordinarily high result of model *M2* on the fourth fold overstates its mean AUC so that it is higher than the mean AUC of *M1*. Similarly, in the cross-validation, presented in Figure 2, the surprisingly low score of model *M1* in the fourth fold underscores its mean AUC so that it is lower than the mean AUC of the *M2* model.

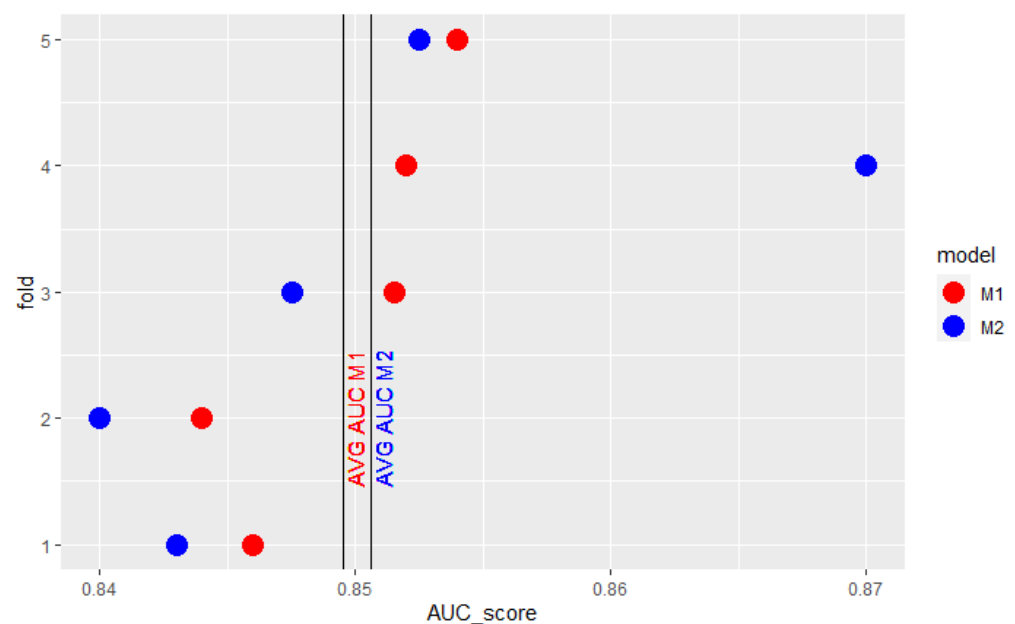


Figure 1. Example AUC scores of 2 models obtained in 5-fold cross-validation.

2.4. Ignoring in Analysis Variance of Outcomes in Specific Folds

Measuring the model's performance using cross-validation assumes simple averaging of obtained scores from all test folds—the result from each fold is taken with equal weight. However, distributions of models' results may differ between folds. Variance in one fold may be several times higher than in the other. The omission of this factor in the analysis may distort the results. Figure 3 presents the diversity of 49 models AUC results among 10 folds. Results were calculated on the credit-risk dataset. The highest standard deviation (8th fold) is over 3 times higher than the lowest (5th fold).

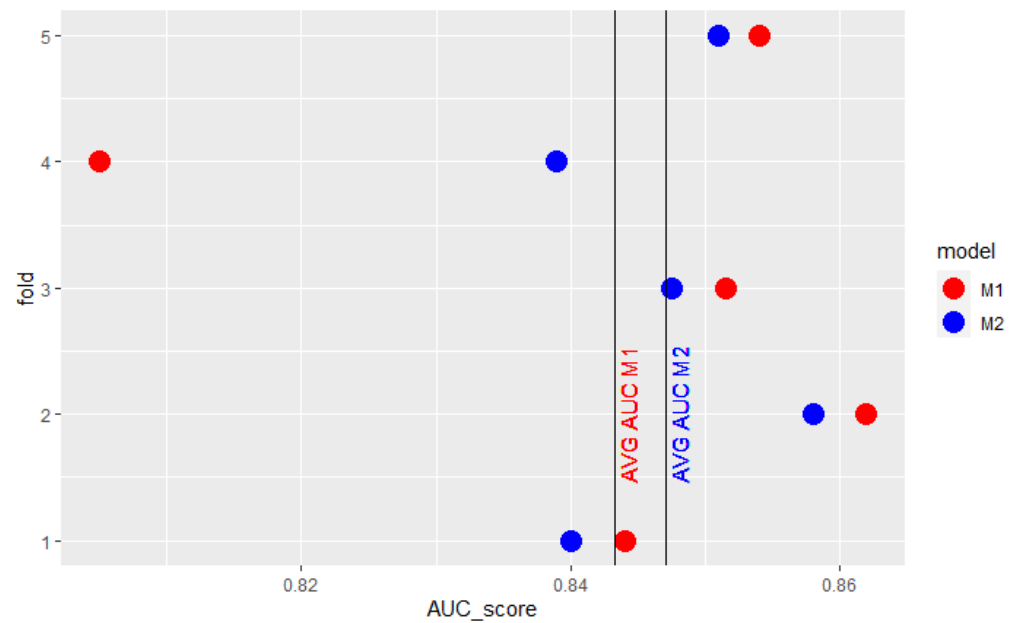


Figure 2. Example AUC scores of 2 models obtained in 5-fold cross-validation.

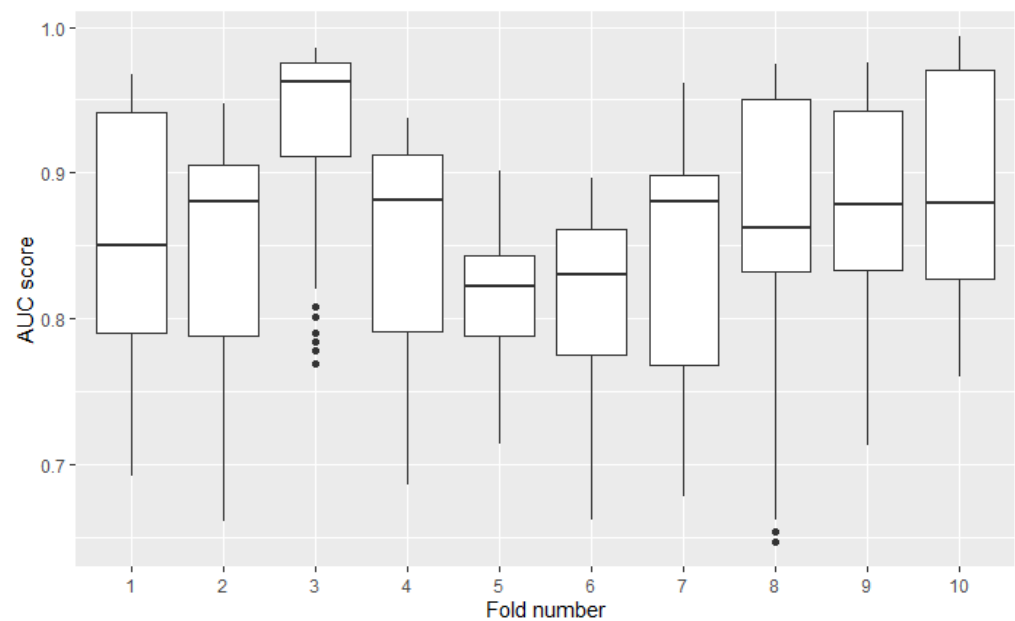


Figure 3. Distribution of 49 models AUC performances among folds obtained from 10-fold cross-validation on credit risk dataset.

Not taking diversity of other models' performances in the fold may lead to a paradoxical situation, where 2 models ranked similarly get considerably different mean AUC values, depending on whether the folds where they scored well had high or low variance. Quantile and AUC results of 2 models presented in Table 1 illustrate this situation. In 2 folds of cross-validation M1 and M2 placed in 60 and 75 percentile. M1 and M2 performed the same, comparing their results to other models. However, their average AUC values differ. The reason for that is different variance in 1st and 2nd fold.

Table 1. Example AUC and percentile results of 2 models in 2 folds.

Model	Fold	AUC	Result Percentile
Model 1	1	0.75	60
	2	0.8	75
Model 2	1	0.8	75
	2	0.7	60

Listed weaknesses of classical subsampling performance assessment methods mean one cannot be sure of their reliability in ordering models. Moreover, these methods do not provide sufficient statistical tools to distinguish between models with better and those with poorer performance in practical applications. The answer to the described challenges are measures based on the probability of win—EPP and PMRA presented in another section.

3. Probability Measures of Model Performance

In order to resolve classic subsampling approach weaknesses mentioned in Section 2, it is recommended to use measures based on the estimation of the probability that one model will achieve better results than another. This section begins by introducing Elo-based Predictive Power measure proposed in [24], which initiated a probability-based approach to the model performance assessment. Further, it is shown how the EPP measure addresses four disadvantages of classic performance measures listed in Section 2. Then Probability-based Models Ranking Approach is introduced, which is the EPP extension resolving its crucial weaknesses. The section concludes with a description of the possible applications of probability model performance measures.

3.1. Elo-Based Predictive Power (EPP)

Assume $\mathcal{M} = \{M_1, \dots, M_m\}$ is a set of machine learning algorithms with fixed hyperparameters, which were evaluated in k-fold cross-validation. The EPP approach is to assign them scores β_1, \dots, β_m representing their predictive power. The higher score, the better model is assessed. This measure is constructed so that difference between 2 models' scores is interpretable. From value of $\beta_i - \beta_j$ difference of M_i, M_j models' scores probability that model M_i performance expressed in classic measure value beats M_j performance, can be calculated. EPP scores can be estimated as coefficients of a following logistic regression model with intercept $\beta_0 = 0$.

$$\ln\left(\frac{p_{i,j}}{1-p_{i,j}}\right) = \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_m x_m \quad (2)$$

where:

$$x_a = 1_{a=i} - 1_{a=j}$$

$p_{i,j}$ —probability, that model M_i gets better result than model M_j .

Logistic Regression Model is estimated on cross-validation results dataset transformed, so that it reflects effect of comparison of every pair of models. Details about dataset transformation are described in Section 3.3 From the model's formula we may obtain the exact probability value:

$$p_{i,j} = \frac{e^{\beta_i - \beta_j}}{e^{\beta_i - \beta_j} + 1}. \quad (3)$$

Such constructed EPP score addresses four classic model performance measures weaknesses listed in Section 2.

1. **Lack of interpretability:** EPP score is interpretable. It gives a clear answer about the probability of one model obtaining better results than another.
2. **Lack of statistical tools to measure statistical significance of results differences:** EPP scores are logistic regression coefficients estimates. To check if two models'

performance difference is significant we can use Wald or Likelihood Ratio tests. They are able to verify linear null hypotheses like: $H_0 : \beta_i - \beta_j = 0$, which implies $p_{i,j} = \frac{1}{2}$.

3. **Sensitivity on outliers:** The outstanding result of the model in one of the folds, of course, will positively impact the assessment of its performance. Nevertheless, the decision will be based on the relation observed in the majority of folds.
4. **Omitting in analysis variance differences between folds:** Data from every fold have an equal impact on fixed effects estimation. The key impact on the EPP score has not the exact value of the measure, but the order of results, those models obtained in specific folds.

The EPP approach is innovative and appears to improve our ability to analyze the model's performance. However, it has one major drawback. One of the logistic regression assumptions is the independence of observations. This assumption is not satisfied because all $\frac{n(n-1)}{2}$ observations expressing models' comparisons made in one fold are not independent. As a result, variance-covariance matrix estimates may be inconsistent [34], which results in underestimated p -value in Wald and LR tests. This may cause one model's performance to be mistakenly considered better than another. As a result, the use of specific algorithms may be rejected, even though they are not significantly weaker.

3.2. Probability-Based Models Ranking Approach

As a solution to the main EPP problem, we propose a Probability-based Models Ranking Approach, which is a modification of the EPP approach. The probability of model M_i overcoming model M_j ($i < j$) is estimated using generalized mixed effects model with logit as a link function in version defined by formula [35]:

$$\ln\left(\frac{p_{i,j}}{1 - p_{i,j}} \mid \text{fold} = k\right) = \beta_0 + \sum_{s=1}^n \beta_s * X_s + u_k + \epsilon \quad (4)$$

where:

$x_a = 1_{a=i} - 1_{a=j}$;

$p_{i,j}$ —probability, that model M_i gets better result than model M_j ;

β_0, \dots, β_n —fixed effects coefficients;

u_k —random intercept depending on number of fold;

ϵ —random error.

$$u_1, \dots, u_k \sim \mathcal{N}(0, \Sigma_{intercept}^2).$$

$$\epsilon \sim \mathcal{N}(0, \sigma^2).$$

From such a formulated model, we can easily calculate probabilities of results between models M_i and M_j . We calculate the expected value of both sides of the formula.

$$\exp\left(\ln\left(\frac{p_{i,j}}{1 - p_{i,j}} \mid \text{fold} = k\right)\right) = \exp\left(\beta_0 + \sum_{s=1}^n \beta_s * X_s + u_k + \epsilon\right) \quad (5)$$

$$\ln\left(\frac{p_{i,j}}{1 - p_{i,j}}\right) = \beta_0 + \beta_i - \beta_j \quad (6)$$

$$p_{i,j} = \frac{\exp(\beta_0 + \beta_i - \beta_j)}{1 + \exp(\beta_0 + \beta_i - \beta_j)} \quad (7)$$

3.2.1. Statistical Tests

Two statistical tests may be helpful to verify linear hypotheses on the values of fixed effects—Likelihood Ratio Test (LR test) and Wald Test. They both allow us to test whether a set of s independent linear hypotheses is true. The null and alternative hypotheses for both of them are:

$$H_0 : K\beta = c \quad (8)$$

$$H_1 : K\beta \neq c \quad (9)$$

where:

K —known, full rank matrix, $\text{rank}(K) = s$;

β —vector of fixed effects.

The Likelihood Ratio Test (LRT) compares two nested models—general and restricted. They are built on the same data, but some fixed effects are omitted in the restricted one. Therefore, the model without the considered effect is a special case of the model with this effect [36]. Test may be applied when both models were fitted using maximum likelihood estimation. The LR test statistic is as follows [37]:

$$LR = -2(l(\tilde{\beta}, \tilde{\Sigma}|y) - l(\hat{\beta}, \hat{\Sigma}|y)) \quad (10)$$

where $(\tilde{\beta}, \tilde{\Sigma})$ are estimations of fixed effects and their variance-covariance matrix in restricted model. $(\hat{\beta}, \hat{\Sigma})$ are estimations in the general model.

The Wald test statistic w given in [37]:

$$[K\hat{\beta} - c]^T [KF_{\hat{\beta}}^{-1}K^T][K\hat{\beta} - c] \quad (11)$$

$F_{\hat{\beta}}$ can be either observed and expected information matrix evaluated in $\hat{\beta}$:

$$F_{\hat{\beta}} = -\frac{\partial^2 l(\hat{\beta}, \hat{\Sigma}|y)}{\partial \beta \partial \beta^T} \Big|_{\beta=\hat{\beta}} \quad (12)$$

or

$$F_{\hat{\beta}} = E\left(-\frac{\partial^2 l(\hat{\beta}, \hat{\Sigma}|y)}{\partial \beta \partial \beta^T} \Big|_{\beta=\hat{\beta}}\right) \quad (13)$$

Both tests statistics w and LR are chi-square distributed with $\text{rank}(K)$ degrees of freedom. The Wald test is a quadratic approximation to the log-likelihood function by a second-order Taylor expansion and is an approximation to the LR statistic. p -values obtained for the LR test are more exact than the p -values obtained for the Wald test. The disadvantage of the LR test is that both the restricted and the unrestricted models have to be fitted, whereas, for the Wald test, only the unrestricted model is needed. With increasing sample size, the log-likelihood becomes approximately quadratic so that both tests are asymptotically equivalent [37].

In the Probability-based Models Ranking Approach, either Wald or LR tests can be applied to test hypotheses of two kinds. The first one has the following null and alternative hypotheses:

$$H_0 : \beta_0 + \beta_i - \beta_j = 0$$

$$H_1 : \beta_0 + \beta_i - \beta_j \neq 0$$

what is equivalent to:

$$H_0 : p_{i,j} = \frac{1}{2}$$

$$H_1 : p_{i,j} \neq \frac{1}{2}$$

It allows us to verify hypotheses about the equality of two models' performances. Another application of LR and Wald tests in PMRA includes testing the following null and alternative hypotheses:

$$H_0 : \beta_{i_1} = \dots = \beta_{i_l} = 0$$

$$H_1 : \exists k \in \{1, \dots, l\} : \beta_k \neq 0$$

It allows us to test whether a group of fixed effects is significant or can be deleted from the model. Removing the insignificant fixed effect corresponding to a given model from the mixed effects model causes a slight change in the estimate of the performance of that model.

In return, reducing the complexity improves the precision of estimates of fixed effects and variance for the best models (which estimates are crucial in most practical applications).

3.2.2. Model Estimation

The described mixed effects model may be estimated using the lme4 package [38]. Coefficients are estimated with maximum likelihood fitting. The approximation algorithm assumes that the Penalized Least Squares method has been used multiple times.

Data matrix from transformed dataset described in Section 3.3 does not have a full columnar rank because the sum of all fixed effects columns is equal to the zero vector. Hence, it is necessary to choose $i_{zero} \in \{1, \dots, n\}$ for which $\beta_{i_{zero}} = 0$ is assumed. This does not affect the results because the coefficients' values are relative. Using the weakest model as model zero is recommended. Then, the subsequently applied procedure eliminating insignificant fixed effect has no negative impact on the estimates of the best models. After the first model estimation, fixed effects, whose coefficients are statistically equal to zero, should be eliminated. There are several methods of choosing a set of insignificant variables. They differ in their accuracy and computation time. The algorithms below are ordered from the slowest but most accurate to the fastest but least accurate.

1. Rigorous backward approach

- (a) sort fixed effects of current model by coefficients' p -value obtained in Wald test ($H_0 : \beta_i = 0$) in descending order.
 - (b) try to eliminate any of fixed effects from current model, beginning from these with highest p -value. Use Likelihood Ratio Test to test joint significance of fixed effects: $X_{i_1}, \dots, X_{i_l}, X_k$, where X_{i_1}, \dots, X_{i_l} are already eliminated fixed effects and X_k is currently inspected fixed effect ($H_0 : \beta_{i_1} = \dots = \beta_{i_l} = \beta_k = 0$)
 - (c) if any fixed effect was eliminated get back to point (a) with restricted model, finish otherwise.
2. **Accelerated backward approach:** The same as Rigorous backward approach, but in point (b), consider eliminating only fixed effects whose p -value is above a certain assumed confidence level α .
 3. **Heuristic backward approach:** Until we get the model for which all coefficients are significant, eliminate the fixed effect with the highest Wald Test p -value.

The Accelerated backward approach seems to be a good compromise between speed and accuracy.

3.3. Data Preparation

Data on which probability measures of model performance are based is the result of a cross-validation process with itemized test folds numbers on which assessment was made. The form of the data is as follows:

$$(Model, Fold, AUC) \quad (14)$$

Example input dataset is presented in Table 2. In the preparation process, the form of the data must be changed to fit the EPP and PMRA models. The new form does not contain exact information about the AUC value. It only provides information for each pair of models, which one performed better (has a higher AUC value). If in k -th fold models (M_i, M_j) ($i < j$) scored $(AUC_{M_i,k}, AUC_{M_j,k})$ the following observation will be added to new dataset.

$$(X_1, X_2, \dots, X_m, Fold, Result) \quad (15)$$

where:

$$X_s = \begin{cases} 1 & \text{if } s = i \\ -1 & \text{if } s = j \\ 0 & \text{otherwise} \end{cases} \quad (16)$$

$$result = \begin{cases} 1 & \text{if } AUC_{M_i,k} > AUC_{M_j,k} \\ 0 & \text{otherwise} \end{cases} \tag{17}$$

$$Fold = k \tag{18}$$

Table 3 shows example dataset from Table 2 after transformation. In first fold occurs:

$$AUC_{M_1,1} > AUC_{M_2,1} > AUC_{M_3,1}$$

therefore, the first 3 rows of the transformed table encode the following information:

$$\begin{aligned} AUC_{M_1,1} &> AUC_{M_2,1} \\ AUC_{M_1,1} &> AUC_{M_3,1} \\ AUC_{M_2,1} &> AUC_{M_3,1} \end{aligned}$$

If the input data are results of n models obtained in k-fold cross validation, then the table we have at the beginning is $(m * k) \times 3$ and after-transformation table is $\frac{m(m-1)}{2} * k \times (m + 2)$.

Table 2. Input dataset.

Model	Fold	AUC
M_1	1	0.785
M_2	1	0.743
M_3	1	0.721
M_1	2	0.727
M_2	2	0.672
M_3	2	0.746

Table 3. Transformed dataset.

X1	X2	X3	Fold	Result
1	-1	0	1	1
1	0	-1	1	1
0	1	-1	1	1
1	-1	0	2	1
1	0	-1	2	0
0	1	-1	2	0

3.4. Comparison of PMRA to EPP

The Probability-based Models Ranking Approach is a modification of the innovative Elo-based Predictive Power (EPP) approach introduced in [24]. Modified model loses some of EPP’s features to address its weaknesses and improve its efficiency in some of the applications.

The main problem with EPP is that one of the logistic regression assumptions— independence of observations—is not satisfied. The results of n models in each fold of input dataset generates $\frac{n(n-1)}{2}$ observations expressing model comparisons. These observations are not independent. As a result, the estimated variance-covariance matrix is underestimated. This makes the EPP tend consider models that have statistically equal performances as different. An attempt to solve this problem is implementing random effects in the Probability-based Models Ranking Approach. Random effects cluster the results obtained in the same fold. As a result, the estimate of the variance-covariance matrix of fixed effects is more accurate. PMRA allows us to rely on statistical tests’ p -values and make decisions in the model selection process based on statistical criteria rather than arbitrarily established conditions.

Another difference is that PMRA contains the procedure of eliminating fixed effects corresponding to models statistically equal to model zero. This General to Specific approach may distort estimates of the probability of a win for models corresponding to deleted fixed effects. However, the model becomes less complex, and the coefficients' variance decreases. PMRA enables better estimation of the probability of win and equal performance tests p -values for models which rated best (in most applications, we are usually more interested in proper performance evaluation of models at the top of the ranking than those from the bottom).

The different design of the Probability-based Models Ranking Approach means that, unlike EPP, it does not have the convenient property that performance can be specified by a single number. Moreover, it cannot be used to compare results from different datasets. What PMRA gives in return is a better estimation of the variance-covariance matrix resulting in better estimates of p -values in tests for equality of two models. This makes PMRA equality of performance of two models reliable in contrast to EPP. In summary, the PMRA is a modification of the EPP lacking its universal properties. In return, its indications are reliable, what enables using it in practical applications.

3.5. Probability Measures Applications

Thanks to the described advantages over EPP, PMRA is suitable for practical applications. Its interpretability and statistical properties constitute advantages over the classical subsampling approach in problems of hyperparameter tuning and creating a ranking of models.

3.5.1. Application in Models Ranking Building

Result of Probability-based Models Ranking Approach estimation may be basis for creation of $\{M_1, \dots, M_m\}$ models ranking in 2 steps.

$$1. \quad \text{Probability matrix calculation} \quad \begin{matrix} & M_1 & M_2 & \dots & M_m \\ \begin{bmatrix} p_{1,1} & p_{1,2} & \dots & p_{1,m} \\ p_{2,1} & p_{2,2} & \dots & p_{2,m} \\ \dots & \dots & \dots & \dots \\ p_{m,1} & p_{m,2} & \dots & p_{m,m} \end{bmatrix} & M_1 \\ & M_2 \\ & \dots \\ & M_m \end{matrix}$$

Probabilities values can be estimated as described in Equation (7). In case fixed effect corresponding to i -th model was eliminated, assume $\beta_i = 0$

2. Assigning places $1, 2, \dots, m$ in ranking to models $M_{i_1}, M_{i_2}, \dots, M_{i_m}$ so that:

for each $k, l \in \{1, 2, \dots, m\} : k < l$ occurs: $p_{i_k, i_l} > \frac{1}{2}$

Theoretically, it may happen that models cannot be ranked in this way because several of them form a cycle in which each model beats the next with a probability higher than 0.5 e.g., $p_{i_k, i_l} > \frac{1}{2}, p_{i_l, i_m} > \frac{1}{2}, p_{i_m, i_k} > \frac{1}{2}$. However, such a case rarely occurs in practice and usually means that the models are insignificantly different. Therefore in this case, these models get the same ranking number.

3.5.2. Application in the Process of Hyperparameters Tuning

There are many hyperparameters tuning methods like: *grid search*, *random search*, *bayesian optimization* [39] or *gradient-based optimization* [40]. Probability-based Models Ranking Approach may provide valuable information supporting model selection decisions during these procedures. Regarding the example of the iterated grid search [41]: while tuning 2 hyperparameters α, γ whose values come from the interval $[0, 1]$ the grid Ω of $m * n$ potential parameters sets is created:

$$\Omega_1 = A \times B \quad (19)$$

where A, B are sets of potential values of α, γ respectively. For each hyperparameters set a model performance is calculated using cross-validation and the best one (a, b) is chosen.

After that, another grid of values in neighbourhood of best set is created and the process is repeated until we decide the result will not improve significantly by further tuning.

$$\Omega_2 = \{(a \pm \epsilon_i, b \pm \delta_j)\} \quad (20)$$

Probability-based Models Ranking Approach may be helpful in two ways:

1. It can indicate which hyperparameters set is the best of considered in the iteration and estimate the probability of overperforming other models by it. Using statistical tests, it can be determined which of the different hyperparameters' sets was not significantly worse than the best and should be included in another search iteration.
2. It can also be a basis for formulating stop conditions for a further grid search. If changes in parameters cause performance changes that are not statistically significant, a continuation of the search is pointless. For example, conditions of this kind may be used: *Stop further search if significantly better hyperparameters set was not found in last k iterations.*

4. Real Data Application

Probability-based Models Ranking Approach may have several applications in the selection of machine learning model. PMRA may be used to create an interpretable ranking of models solving a specific task. It allows us to check which algorithms are best suited to this particular problem. When the algorithm is chosen, PMRA may be helpful in the hyperparameters tuning process. Probability-based Models Ranking Approach gives statistical tools to select the set of optimal parameters and to set up stopping conditions, which will limit computation time to the necessary minimum. This section describes these applications using results obtained on a processed dataset containing real mortgage credit repayment history. Credits included in this dataset describe multifamily loans (5+ residential units) acquired by Fannie Mae (<https://multifamily.fanniemae.com/news-insights/multifamily-wire/fannie-mae-multifamily-reports-q1-2022-financial-results>, accessed date: 1 April 2021) in years 2000–2020.

4.1. Dataset Description

The dataset on which the probability-based models were built consisted of AUC measures of 49 different models obtained from 10-fold cross-validation. In the study, the following algorithms were used:

1. Ada Boost (10 hyperparameters sets),
2. K-Nearest Neighbors (9 hyperparameters sets),
3. Decision Tree (10 hyperparameters sets),
4. Random Forest (10 hyperparameters sets),
5. XGBoost (10 hyperparameters sets).

These algorithms were trained and tested on the Fannie Mae Multifamily Mortgages dataset. Fannie Mae is USA Government Sponsored Enterprise operating on the secondary mortgage market. After preprocessing, the dataset counted 224,988 observations of 62 features and default flag (dependent binary variable) describing 54,160 credits acquired in the period 2000–2020. The default rate was around 1%. Data included repayment history, property information, loan repayment terms, and information about the borrower's cash flow. Tables 4 and 5 present 10 variables with the highest discriminatory power.

Table 4. Top categorical variables.

Variable Name	Definition	Levels	Gini
payment_status	Status of the mortgage loan payment	Current 30–59 days delinquent 60–89 days delinquent	0.46
in_mbs	Information whether it belongs to MBS	YES NO	0.20

Table 5. Top numeric variables.

Variable Name	Definition	Min	Mean	Max	Gini
business_cycle	Business cycle indicator	2000.17	2012.83	2020.67	0.34
loan_acq_LTV	Unpaid principal balance to the value of all underlying properties.	0	64.81	276.20	0.35
90_days_DELIQ_6 M	How many months in the last 6 M was the credit delinquent at least 90 days	0	0.01	6	0.32
DSCR	A ratio of underwritten net cash flow to an actual debt	0.39	1.78	85.72	0.29
DTB_interest_rate	Difference between reference rates and current credit rates	−2.23	4.39	9.72	0.26
reference_rate_origin	Interbank credits rates at credit origination date	0.07	1.85	6.54	0.25
reference_rate	Reference interest rate at reporting date	0.05	1.10	6.54	0.22
months_on_book	Number of months since credit was granted	0	45.72	236	0.21

4.2. Application in Models Ranking Building

The crucial step in solving any problem with machine learning is selecting a model appropriate to the given problem. As described in Section 4.1, 5 algorithms, each with several hyperparameters sets were tested on a real mortgage credits dataset. Their results were analyzed using PMRA and averaging approach. The use of PMRA allowed us to obtain the probability of winning in the comparison of each pair of models, as well as statistically testing whether the models' performances are significantly different. The estimated probabilities and Wald test p -values for top 10 models comparisons are presented in Figures 4 and 5. Average AUC and Probability-based Models Ranking Approach rankings calculated on real data are presented in Tables 6 and 7 respectively. The top 2 models in both rankings are the same, but in further places, differences are visible. Model RF9 is best in both rankings. Nevertheless, applying the Wald test on 5% confidence level shows that XGBoost configurations, which took places 2–4, are statistically equal to it. The estimated probabilities confirm that. The probability of an RF9 win against XGB6 is only 0.505. This suggests that both algorithms, Random Forest and XGBoost, should be considered in further analysis. It is worth noticing that three different XGBoost's hyperparameters sets were recognized as equal to the best model. It indicates that XGBoost in this application will be a stable solution—its results will not deteriorate significantly when changing the hyperparameter values.

Table 6. Mean AUC ranking.

	Model	AUC
1	RF9	0.938
2	XGB6	0.933
3	RF8	0.932
4	XGB7	0.931
5	XGB9	0.929
6	XGB0	0.929
7	XGB3	0.926
8	XGB4	0.924
9	RF5	0.920
10	XGB5	0.920

Table 7. PMRA ranking.

	Model	P. of Win against Top Model	Wald <i>p</i> -Value
1	RF9	-	-
2	XGB6	0.495	0.948
3	XGB9	0.388	0.093
4	XGB7	0.386	0.088
5	RF8	0.355	0.031
6	XGB0	0.369	0.051
7	XGB3	0.309	0.003
8	RF2	0.276	<0.001
9	XGB4	0.286	0.001
10	RF5	0.231	<0.001

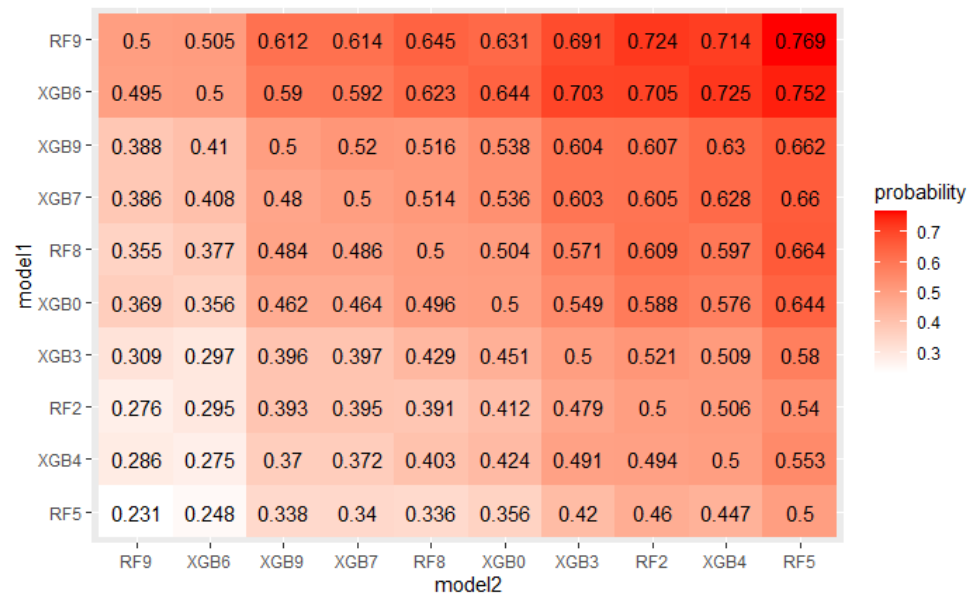


Figure 4. Probabilities of win estimated by PMRA for top 10 models.

Some models are ranked differently in terms of mean AUC and PMRA rankings. For example, model XGB5 has a slightly higher mean AUC compared to the RF2 model (0.920 vs. 0.918) while Probability-based Models Ranking Approach classified RF2 considerably higher than XGB5 ($p_{RF2,XGB5} \approx 0.605$) Wald Chi-square test rejects null hypothesis that models' performances are the same with p -value = 0.04 Comparison of the results models obtained in the specific folds, shown in the Figure 6, leads to the conclusion that in 6 out of 10 of them, RF2 performed better, but at the same time, the XGB5 average AUC is higher. This fact is a consequence of simple averaging approach weaknesses mentioned

in Section 2—outliers impact and not including standard deviation differences between folds in analysis. The most significant performance difference in favor of XGB5 occurred in fold 7. It can be considered an outlier because the AUC value obtained in this fold by RF2 significantly deviates from others obtained in most other folds. Moreover, the outstanding performance of RF2 in fold 5 was overshadowed by the lowest standard deviation of results in this fold, as presented in Figure 7. AUC difference is nearly 3 times lower than in the mentioned 7th fold, but performance measured with quantiles, which is presented in Figure 8 is comparable. The above example confirms the Probability-based Models Ranking Approach’s advantages over the classical averaging approach in crucial aspects listed in the Section 2.

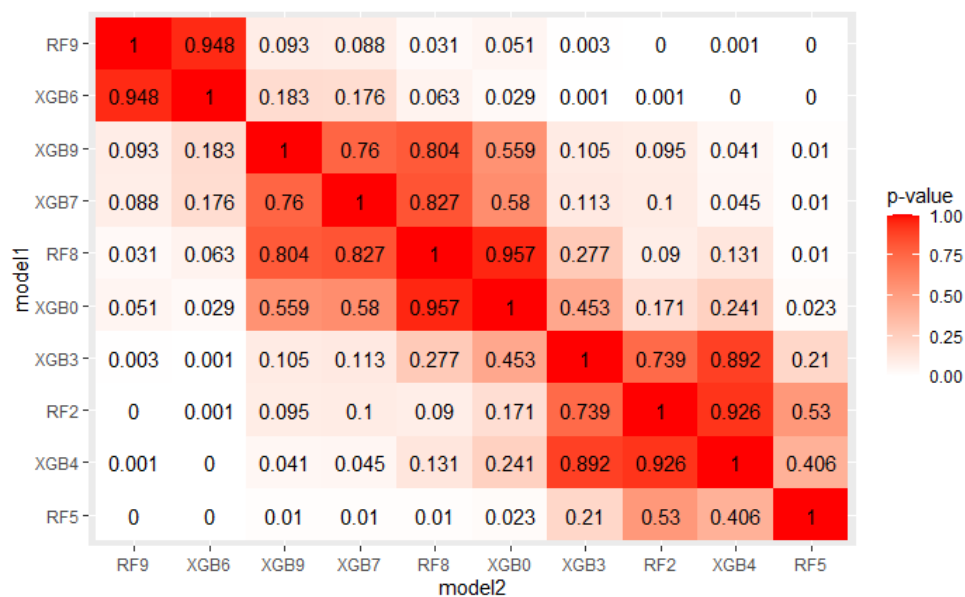


Figure 5. Wald test p-value, where null hypothesis is that two models have equal performances.

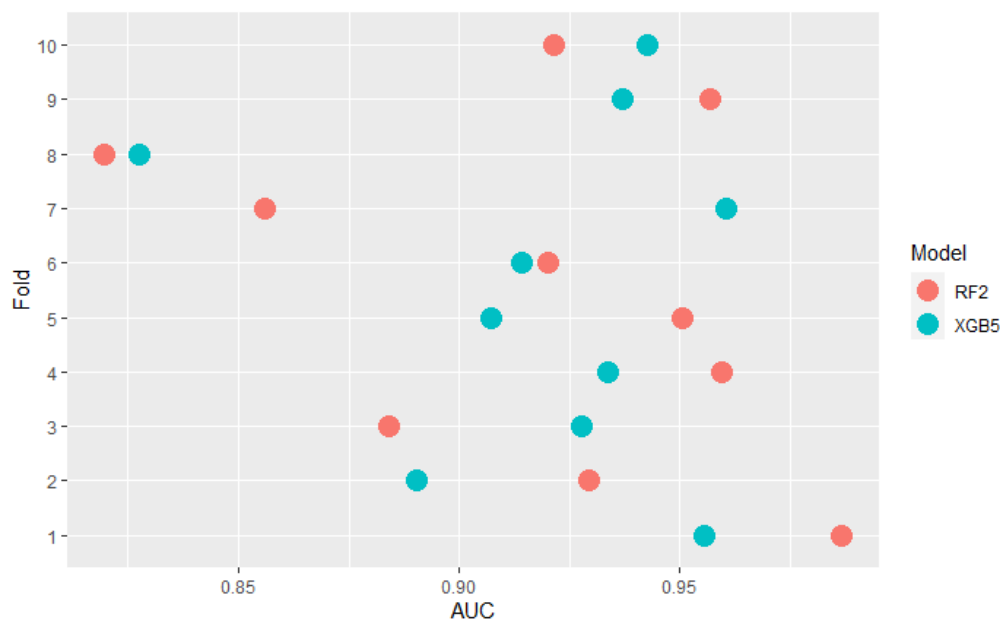


Figure 6. Comparison of RF2 and XGB5 models AUC performance across folds.

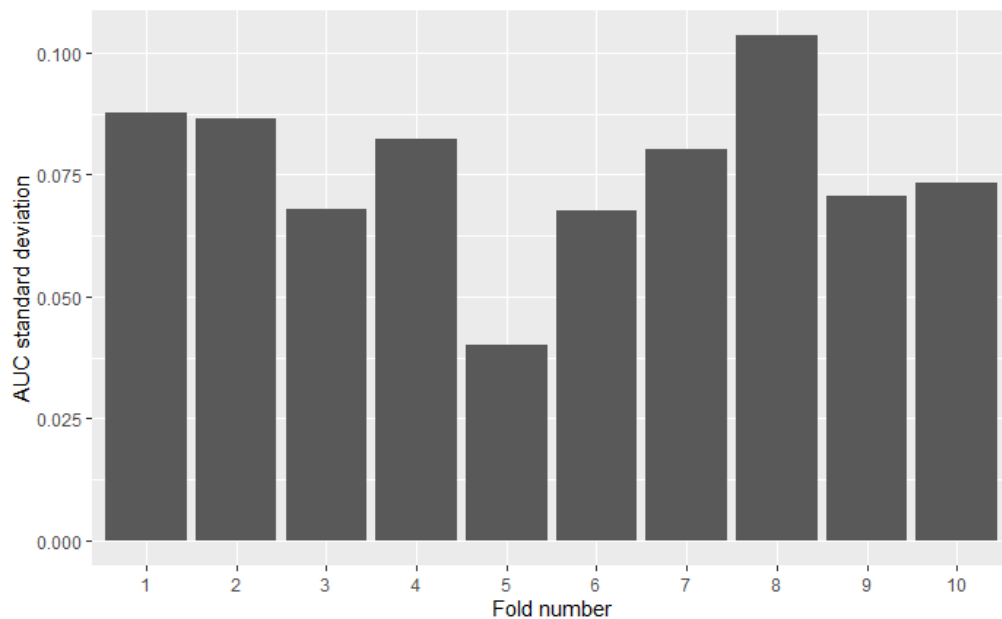


Figure 7. Standard deviation of the models’ performances across folds.

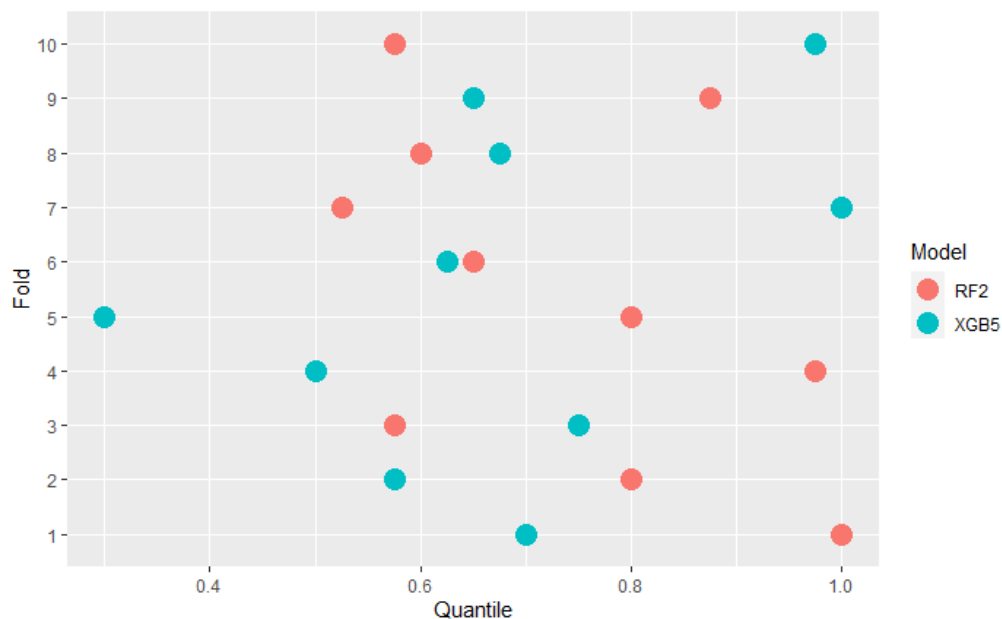


Figure 8. Comparison of RF2 and XGB5 models AUC performance quantiles across folds.

4.3. Application in Hyperparameter Tuning

After choosing the right algorithm for the problem, the next step is to tune its hyperparameters. The probability-based Models Ranking Approach may also be helpful in this. Parameters *max_depth* and *colsample* of XGBoost [42] algorithm were tuned. The initial grid was:

$$M \times C \tag{21}$$

where:

- $M = \{4, 8, 12, 16\}$ (*max_depth* parameter values);
- $C = \{0.25, 0.5, 0.75\}$ (*colsample* parameter values).

In each iteration, the algorithm was trained with the hyperparameters’ sets that were located around the best set and those that the Probability-based Models Ranking Approach recognized as statistically equal to the best one. In each successive iteration, the range from

which new sets of parameters were selected decreased twice. The Wald test was used with a 10% significance level to test the significance of the difference between the two models' performance. The assumed condition to stop further search was that the best model after $n + 2$ iterations is not significantly better than the best model after n iterations. The stop condition was fulfilled after the 4th iteration. Figures 9 and 10 present hyperparameters sets which were tested after 2 and 4 iterations. Those recognized as best or not significantly different from best are highlighted on the plot by their shape.

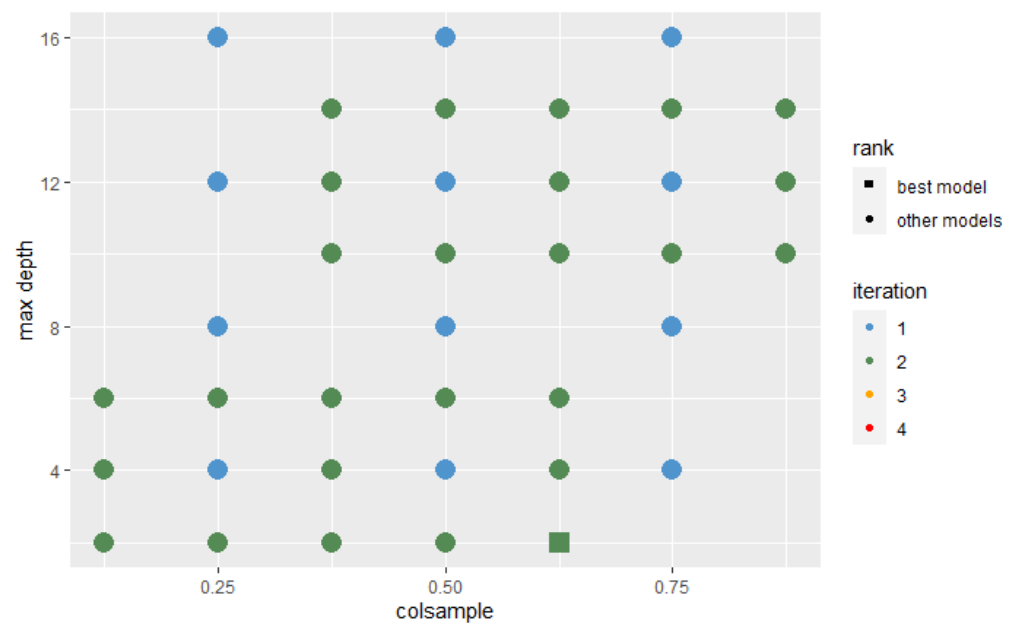


Figure 9. Hyperparameters sets analyzed in 2 iterations.

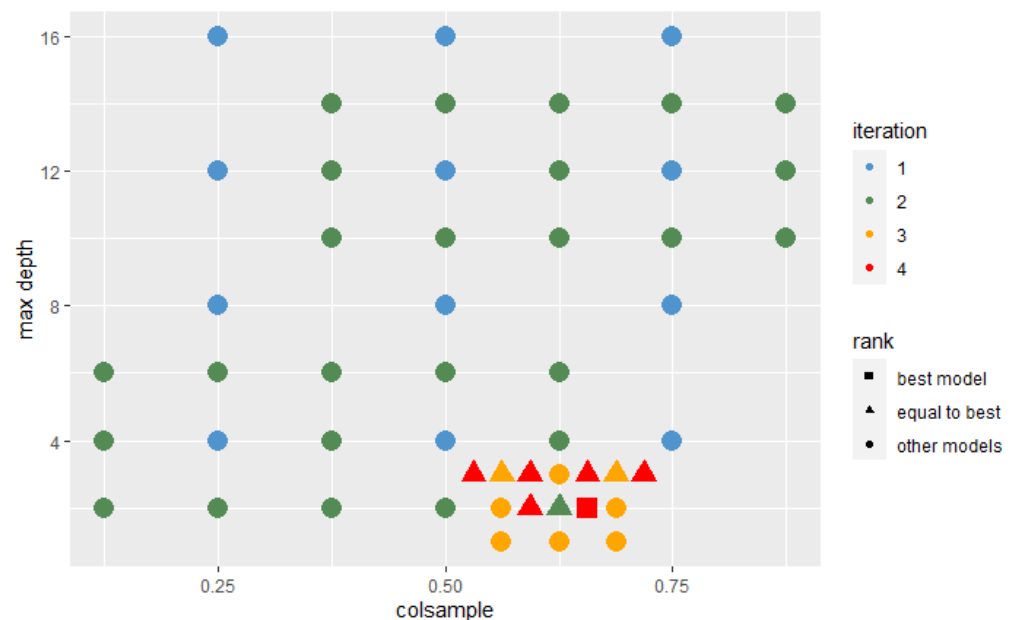


Figure 10. Hyperparameters sets analyzed in 4 iterations.

4.4. Comparison of Probability-Based Models Ranking Approach and EPP

The EPP and Probability-based Models Ranking Approach were estimated on a credit dataset described in Section 4.1 to check if the obtained probabilities of win and statistical tests p -values confirm expected differences between models. Figures 11 and 12 present probabilities of wins estimated in both approaches for the top 10 models. As expected, the differences are small.

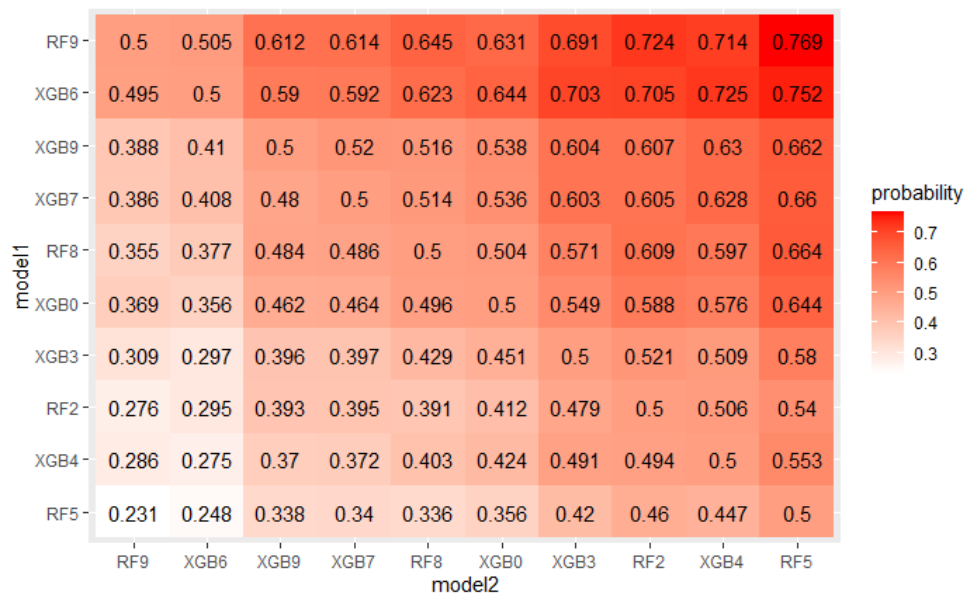


Figure 11. Probabilities of win estimated by PMRA for top 10 models.

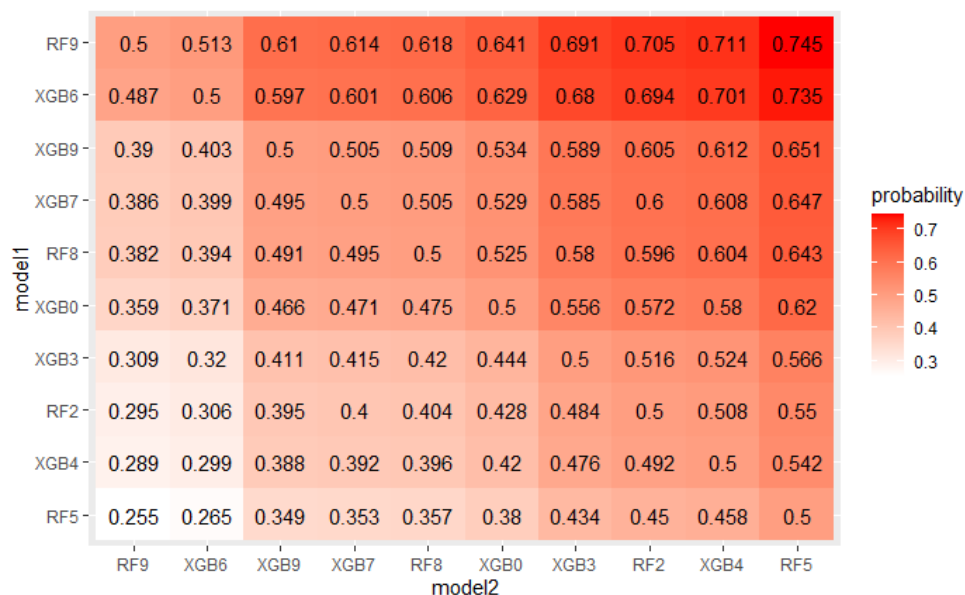


Figure 12. Probabilities of win estimated in EPP approach for the top 10 models.

The Wald test was performed for every pair from the top 25 models to verify the null hypothesis that their performance is equal. Figure 13 presents obtained p -values for both PMRA and EPP, depending on the estimated win probability. It can be observed that for pairs of models whose estimated probability is approximately equal to 0.5, p -values obtained from EPP are slightly higher than those from PMRA. Nevertheless, it does not affect model selection decisions because these p -values significantly exceed usually used confidence levels. When the win probability is more deviated from 0.5, EPP p -values are sev-

eral times higher than those obtained in the Probability-based Models Ranking Approach. These results are consistent with the theoretical expectations—the EPP model assumes independence of observations, which are, in fact, dependent and therefore underestimates the variance-covariance matrix. It results in the higher p -values obtained in statistical tests. Modification introduced in the Probability-based Models Ranking Approach allows for more realistic variances values. It means that the EPP approach may incorrectly reject the null hypothesis, that performance of the two models is equal, while PMRA indicates that there is no reason to reject it. Inference from EPP indications may lead to the false conclusion that it is not worth using a particular algorithm for the problem, even if it was not significantly worse. This unnecessarily limits the range of algorithms available.

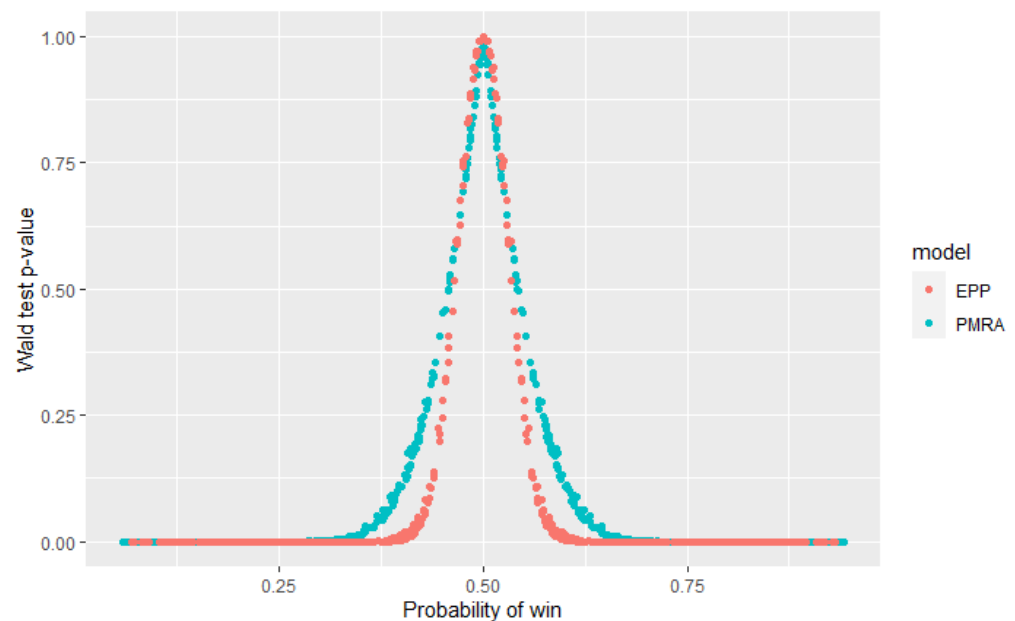


Figure 13. Distribution of the obtained p -values in the Wald test according to the probability of winning for PMRA and EPP.

5. Discussion

The subsampling methods of the model performance assessment have several weaknesses which make their indications insufficient for most practical applications: (i) they are not interpretable (ii) there is no possibility to statistically test the equality of 2 models' performance (iii) they are not outliers resistant (iv) they do not consider the diversity of different models' performances on a specific subsample. Our novel Probability-based Models Ranking Approach addressed all these issues. In recent years, many articles have been published showing alternative approaches to the assessment of model performance addressing these issues. As explained in Section 2 the main cause of (i) and (ii) is the bias of the value of the performance metric obtained in the subsampling averaging process. As studies [9,13] prove, bias is particularly large for small-sample problems.

Efforts of some of the newly proposed methods have been focused on reducing the bias of algorithm's estimated prediction error. Parker et al. show in their study how applying of sample balancing and stratification in methods like *Balanced cv*, *Stratified cv*, *Balanced leave-one-out cv* causes a reduction of averaging bias for small datasets. Another commonly used methods of model performance estimation are *Repeated cross-validation* [43], *Leave-pair-out cross-validation* (LPOCV) [14] and *Leave-one-out cross-validation* (LOOCV) [44]. These methods (especially LPOCV and LOOCV) appear to give the most reliable results possible, but they require fitting the model to data a great many times. This excludes them from applications where the number of calls may be large, such as tuning hyperparameters. The use of these methods is therefore limited to a narrow group of problems, where model fitting time is relatively short. Probability-based Models Ranking Approach's aim is not

to estimate the exact value of the metric of the model's performance. It is not a necessary feature for every performance assessment procedure (e.g., hyperparameter tuning). Instead, PMRA indicates models' relative performance based on statistical criteria. The specifics of the time-complexity of PMRA and the methods mentioned above are different. In the case of PMRA, the execution time increases significantly, primarily as the number of models participating in a comparison grows, while for bias reduction methods the factor having the strongest impact on computation time is the number of models' fits to data.

Other methods attempting to address the weaknesses of classical model performance measures are focused on the statistical aspect of models' comparison. There is a number of testing procedures determining whether sub-sampling results support the conclusion that the performance of the 2 models differs significantly. Among them *5 × 2-Fold Cross-Validation Paired t-Test* [22] and *Combined 5 × 2 cv F-test* [21] gained the most importance in practical applications. Results provided by these methods may be a valuable decision criterion in the model selection process. Although the above-mentioned methods are reliable, they do not use the full potential of subsampling results. The Probability-based Models Ranking Approach, when used for the creation of models' ranking or for hyperparameters tuning, compares not only the relative results of 2 analyzed models, but also the results of every other model, fitted to the same subsamples. Thanks to this, PMRA provides a precise and reliable estimate of the probability of win and *p*-value in the statistical test for two models' performance equality.

6. Conclusions and Further Development

In this paper, we identified the most critical weaknesses of subsampling methods, like cross-validation and bagging in the model performance assessment process. Then we presented Elo-based Predictive Power—an innovative performance measure that attempts to respond to the problems of the classical approach. It is interpretable and allows the significance of the difference between 2 models to be tested. Despite its considerable advantages, it has one crucial drawback—*p*-values calculation is based on incorrect assumptions. Further, we introduce the Probability-based Models Ranking Approach—a modified version of EPP. This approach loses EPP's universality but is based on correct assumptions, and, thanks to that, its indications are reliable.

The basic properties of the Probability-based Model Ranking Approach are its interpretability and the possibility of statistical testing of differences between models' performance. PMRA may be used to create an interpretable ranking of algorithms to select the one that is most appropriate to the problem. Ranking creation also enables verification of the performance of the newly developed model against a group of state-of-the-art algorithms. PMRA ranking allows us to check whether a new model brings any new quality or its improvement is insignificant.

After the algorithm selection, PMRA may be helpful in the hyperparameters tuning process. It indicates places in the hyperparameters space where it is worth continuing the search and allows to determine the stop conditions reducing the search time to the necessary minimum. PMRA properties may be used to create more complex criteria for tuning continuation depending on the probability of result improvement, statistical tests, and expected computation time. The application of PMRA in developing such criteria is a field for further research. Further research is also needed to examine a reasonable number of folds to apply the PMRA approach and its stability for various numbers of folds.

The results obtained in the study confirm the claims made about the advantages of PMRA over the state-of-the-art averaging approach. However, the results have their limitations. Firstly, the analysis considered multiple machine learning models, but they were trained on a single dataset. This does not negatively influence the validity of findings; however, consideration of a wider range of various datasets would have allowed us to determine in which context PMRA brings the most valuable insight. Secondly, the empirical analysis conducted in Section 4 focused on the qualitative factor. It examined individual cases where the indications of the PMRA measure differed from the state-of-the-

art averaging approach and confirmed that the differences were due to the advantages of the new approach over the benchmark subsampling. What was missing, however, was a quantitative study—an examination of how often the two approaches differ in their results—to rank the models differently when creating the models' ranking, and determine how often parameter tuning leads to similar results in both cases. Conducting the above analysis would have made it possible to determine in which conditions—sample size, number of fitted models—the potentially time-consuming PMRA offers added value over the benchmark approach in model performance assessment.

Author Contributions: Conceptualization, M.C.; methodology, S.G.; software, S.G.; validation, M.C.; formal analysis, M.C.; investigation, S.G.; resources, S.G.; data curation, S.G.; writing—original draft preparation, S.G.; writing—review and editing, S.G. and M.C.; visualization, S.G.; supervision, M.C.; project administration, M.C.; funding acquisition, S.G. All authors have read and agreed to the published version of the manuscript.

Funding: The article was prepared within the framework of a non-competitive project of the Ministry of Education “School of Eagles” based on the agreement between the Ministry of Education and the University of Warsaw, contract no. MNISW/2020/142/DIR/KH, signed on 26 May 2020. Project co-financed from the European Social Fund under the Operational Programme Knowledge Education Development 2014–2020.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Initial Fannie Mae dataset: <https://capitalmarkets.fanniemae.com/credit-risk-transfer/single-family-credit-risk-transfer/fannie-mae-single-family-loan-performance-data>; Code and processed data: <https://github.com/stagajda/PMRA> (accessed on 1 April 2021).

Conflicts of Interest: The authors declare no conflict of interest.

Abbreviations

The following abbreviations are used in this manuscript:

EPP	Elo-based Predictive Power
PMRA	Probability-based Models Ranking Approach

References

1. Garcia, S.; Fernandez, A.; Luengo, J.; Herrera, F. A study of statistical techniques and performance measures for genetics-based machine learning: Accuracy and interpretability. *Soft Comput.* **2008**, *13*, 959. [CrossRef]
2. Kowalik, S.J. Deep Learning in Double Dummy Bridge Problem. Ph.D. Thesis, Zakład Sztucznej Inteligencji i Metod Obliczeniowych, Warsaw, Poland, 2021.
3. Yadavendra; Chand, S. A comparative study of breast cancer tumor classification by classical machine learning methods and deep learning method. *Mach. Vis. Appl.* **2020**, *31*, 46. [CrossRef]
4. Boubezoul, A.; Paris, S.; Ouladsine, M. Application of the cross entropy method to the GLVQ algorithm. *Pattern Recognit.* **2008**, *41*, 3173–3178. [CrossRef]
5. Yala, A.; Lehman, C.; Schuster, T.; Portnoi, T.; Barzilay, R. A Deep Learning Mammography-based Model for Improved Breast Cancer Risk Prediction. *Radiology* **2019**, *292*, 60–66. [CrossRef] [PubMed]
6. Nguyen, H.V.; Naeem, M.A.; Wichitaksorn, N.; Pears, R. A smart system for short-term price prediction using time series models. *Comput. Electr. Eng.* **2019**, *76*, 339–352. [CrossRef]
7. Hu, Z.; Zhao, Y.; Khushi, M. A Survey of Forex and Stock Price Prediction Using Deep Learning. *Appl. Syst. Innov.* **2021**, *4*, 9. [CrossRef]
8. Cao, L.; Tay, F.E. Financial Forecasting Using Support Vector Machines. *Neural Comput. Appl.* **2001**, *10*, 184–192. [CrossRef]
9. Airola, A.; Pahikkala, T.; Waegeman, W.; De Baets, B.; Salakoski, T. An experimental comparison of cross-validation techniques for estimating the area under the ROC curve. *Comput. Stat. Data Anal.* **2011**, *55*, 1828–1844. [CrossRef]
10. Bengio, Y.; Grandvalet, Y. No unbiased estimator of the variance of k-fold cross-validation. *Adv. Neural Inf. Process. Syst.* **2003**. Available online: <https://proceedings.neurips.cc/paper/2003/hash/e82c4b19b8151ddc25d4d93baf7b908f-Abstract.html> (accessed on 8 May 2021).
11. Bradley, A.P. The use of the area under the ROC curve in the evaluation of machine learning algorithms. *Pattern Recognit.* **1997**, *30*, 1145–1159. [CrossRef]

12. Parker, B.J.; Günter, S.; Bedo, J. Stratification bias in low signal microarray studies. *BMC Bioinform.* **2007**, *8*, 326. [[CrossRef](#)] [[PubMed](#)]
13. Varoquaux, G. Cross-validation failure: Small sample sizes lead to large error bars. *NeuroImage* **2018**, *180*, 68–77. [[CrossRef](#)] [[PubMed](#)]
14. Krzanowski, W.; Hand, D. Assessing Error Rate Estimators: The Leave-One-Out Method Reconsidered. *Aust. J. Stat.* **1997**, *39*, 35–46. [[CrossRef](#)]
15. Pahikkala, T.; Airola, A.; Boberg, J.; Salakoski, T. Exact and efficient leave-pair-out cross-validation for ranking RLS. *Proc. AKRR* **2008**, *2008*, 1–8.
16. Gronau, Q.F.; Wagenmakers, E.J. Limitations of Bayesian Leave-One-Out Cross-Validation for Model Selection. *Comput. Brain Behav.* **2019**, *2*, 1–11. [[CrossRef](#)]
17. Raschka, S. Model evaluation, model selection, and algorithm selection in machine learning. *arXiv* **2018**, arXiv:1811.12808.
18. Dietterich, T.G. Approximate statistical tests for comparing supervised classification learning algorithms. *Neural Comput.* **1998**, *10*, 1895–1923. [[CrossRef](#)]
19. McNemar, Q. Note on the sampling error of the difference between correlated proportions or percentages. *Psychometrika* **1947**, *12*, 153–157. [[CrossRef](#)]
20. Cochran, W.G. The comparison of percentages in matched samples. *Biometrika* **1950**, *37*, 256–266. [[CrossRef](#)]
21. Alpaydm, E. Combined 5×2 cv F test for comparing supervised classification learning algorithms. *Neural Comput.* **1999**, *11*, 1885–1892. [[CrossRef](#)]
22. Brownlee, J. *Machine Learning Mastery Blog*. 2019. Available online: <https://machinelearningmastery.com/blog/> (accessed on 8 May 2021).
23. Nadeau, C.; Bengio, Y. Inference for the Generalization Error. *Mach. Learn.* **2003**, *52*, 239–281. [[CrossRef](#)]
24. Gosiewska, A.; Woznica, K.; Biecek, P. Interpretable Meta-Measure for Model Performance. *arXiv* **2020**, arXiv:2006.02293.
25. Wu, J.; Chen, J.; Liu, W.; Liu, Y.; Liang, C.; Cao, M. A Calibrated Individual Semantic Based Failure Mode and Effect Analysis and Its Application in Industrial Internet Platform. *Mathematics* **2022**, *10*, 2492. [[CrossRef](#)]
26. Cao, M.; Liu, Y.; Gai, T.; Zhou, M.; Fujita, H.; Wu, J. A Comprehensive Star Rating Approach for Cruise Ships Based on Interactive Group Decision Making with Personalized Individual Semantics. *J. Mar. Sci. Eng.* **2022**, *10*, 638. [[CrossRef](#)]
27. Berrar, D. Cross-Validation. 2019. Available online: https://www.researchgate.net/profile/Daniel-Berrar/publication/324701535_Cross-Validation/links/5cb4209c92851c8d22ec4349/Cross-Validation.pdf (accessed on 6 May 2021).
28. Ghojogh, B.; Crowley, M. The theory behind overfitting, cross validation, regularization, bagging, and boosting: Tutorial. *arXiv* **2019**, arXiv:1905.12787.
29. Browne, M.W. Cross-Validation Methods. *J. Math. Psychol.* **2000**, *44*, 108–132. [[CrossRef](#)]
30. Huang, J.; Ling, C.X. Using AUC and accuracy in evaluating learning algorithms. *IEEE Trans. Knowl. Data Eng.* **2005**, *17*, 299–310. [[CrossRef](#)]
31. Rosset, S. Model selection via the AUC. In Proceedings of the Twenty-First International Conference on Machine Learning, Banff, AB, Canada, 4–8 July 2004; p. 89.
32. DeLong, E.R.; DeLong, D.M.; Clarke-Pearson, D.L. Comparing the Areas under Two or More Correlated Receiver Operating Characteristic Curves: A Nonparametric Approach. *Biometrics* **1988**, *44*, 837–845. [[CrossRef](#)]
33. Hanley, J.A.; McNeil, B.J. A method of comparing the areas under receiver operating characteristic curves derived from the same cases. *Radiology* **1983**, *148*, 839–843. [[CrossRef](#)]
34. Gałęcki, A.; Burzykowski, T. Linear mixed-effects model. In *Linear Mixed-Effects Models Using R*; Springer: Berlin/Heidelberg, Germany, 2013; pp. 245–273.
35. Breslow, N.E.; Clayton, D.G. Approximate Inference in Generalized Linear Mixed Models. *J. Am. Stat. Assoc.* **1993**, *88*, 9–25.
36. Biecek, P. *Analiza Danych z Programem R: Modele Liniowe z Efektami Stałymi, Losowymi i Mieszanymi*; Wydawnictwo Naukowe PWN: Warsaw, Poland, 2013.
37. Tuerlinckx, F.; Rijmen, F.; Verbeke, G.; De Boeck, P. Statistical inference in generalized linear mixed models: A review. *Br. J. Math. Stat. Psychol.* **2006**, *59*, 225–255. [[CrossRef](#)] [[PubMed](#)]
38. Bates, D.; Mächler, M.; Bolker, B.; Walker, S. Fitting Linear Mixed-Effects Models Using lme4. *J. Stat. Softw.* **2015**, *67*, 1–48. [[CrossRef](#)]
39. Wu, J.; Chen, X.Y.; Zhang, H.; Xiong, L.D.; Lei, H.; Deng, S.H. Hyperparameter Optimization for Machine Learning Models Based on Bayesian Optimization. *J. Electron. Sci. Technol.* **2019**, *17*, 26–40. [[CrossRef](#)]
40. Paul, S.; Kurin, V.; Whiteson, S. Fast efficient hyperparameter tuning for policy gradient methods. *Adv. Neural Inf. Process. Syst.* **2019**. Available online: <https://proceedings.neurips.cc/paper/2019/hash/743c41a921516b04afde48bb48e28ce6-Abstract.html> (accessed on 6 May 2021).
41. Ataee, M.; Osanlou, M. Determination of Optimum Cutoff Grades of Multiple Metal Deposits by Iterated Grid Search Method. *Int. J. Eng. Sci.* **2003**, *14*, 79–88.
42. Chen, T.; He, T.; Benesty, M.; Khotilovich, V.; Tang, Y.; Cho, H.; Chen, K. Xgboost: Extreme gradient boosting. In *R Package Version 0.4-2*; 2015; Volume 1, pp. 1–4. Available online: <https://cran.microsoft.com/snapshot/2017-12-11/web/packages/xgboost/vignettes/xgboost.pdf> (accessed on 6 May 2021).

-
43. Peterson, R.A.; Cavanaugh, J.E. Ordered quantile normalization: A semiparametric transformation built for the cross-validation era. *J. Appl. Stat.* **2020**, *47*, 2312–2327. [[CrossRef](#)]
 44. Rad, K.R.; Maleki, A. A scalable estimate of the out-of-sample prediction error via approximate leave-one-out cross-validation. *J. R. Stat. Soc. Ser. B (Stat. Methodol.)* **2020**, *82*, 965–996. [[CrossRef](#)]