ISSN: 2281-1346

UNIVERSITÀ DI PAVIA
**Department of Economics
and Management**

DEM Working Paper Series

**A rank graduation accuracy
measure**

Arianna Agosto
(University of Pavia)

Paolo Giudici
(University of Pavia)

Emanuela Raffinetti
(University of Milan)

179 (01-20)

Via San Felice, 5
I-27100 Pavia

economieweb.unipv.it

A rank graduation accuracy measure

Arianna Agosto, Paolo Giudici

Department of Economics and Management, University of Pavia
and

Emanuela Raffinetti

Department of Economics, Management and Quantitative Methods,
University of Milan

January 28, 2020

Abstract

A key point in the application of data science models is the evaluation of their accuracy. Statistics and machine learning have provided, over the years, a number of summary measures aimed at measuring the accuracy of a model in terms of its predictions, such as the Area under the ROC curve and the Somers' coefficient. Our aim is to present an alternative measure, based on the distance between the predicted and the observed ranks of the response variable, which can improve model accuracy in challenging real world applications.

Keywords: Predictive accuracy, Concordance measures, Credit Scoring
JEL: C01, C18, C31, C52, G32

1 Introduction

One of the most important aspects in the evaluation of a statistical model is the construction of diagnostic measures which evaluate its predictive accuracy. Such measures may mitigate the risk of taking wrong actions, which can have a severe impact. For a review, see for example [8].

Indeed, the increasing availability of computational power has allowed to implement accuracy measures in statistical softwares. This has allowed to compare, on the same data, different types of predictive models of a response variable which, while belonging to the same general class, may differ substantially in terms of the employed explanatory variables.

Such predictive models include linear models, which may differ in the number of explanatory variables; graphical models, which differ in the number of conditional dependencies (edges in the graph); tree models, which differ in the number of final classes, and feed forward neural networks, which differ in terms of the number of hidden strata and nodes.

Once a class of models has been established, the problem is how to choose the “best” model from it. The traditional paradigm compares statistical models within the theory of statistical hypotheses testing, in which a model is chosen through a sequence of pairwise comparisons. These criteria are generally not applicable to machine learning models, which do not necessarily have an underlying probabilistic model and, therefore, do not allow the application of statistical hypotheses testing theory. Furthermore, it is often possible to use more than one model class. For example, in predictive classification problems it is possible to use both logistic regression and tree models as well as neural networks. In addition, model specification and therefore model choice is determined by the type of variables used. The variables can also be the result of transformations or of the elimination of observations, following an exploratory analysis. We then need to compare models based on different sets

of variables present at the start. For example, how do we compare a linear model with the original explanatory variables with one with principle components as explanatory variables?

The above considerations suggest that classical statistical model comparison is not sufficient to evaluate the predictive accuracy of a statistical model. This explains why the last years have witnessed the growing importance of model comparison methods based on the direct calculation of the predictive accuracy of a model, through cross-validation. In the cross-validation process, the data is split in two or more datasets, with training datasets used to fit a model and validation datasets used to compare the predictions made by the fitted model with the actual observed values.

In the cross-validation process, model comparison methods are not universal but differ, depending on the type of response variable to be predicted. Our work is motivated by the large body of applications of data science methods in credit scoring (for a review, see e.g. [14] and the references therein). In credit scoring applications, the response variable is usually binary. In this case, predictive accuracy can be evaluated in terms of false positive and false negative predictions, giving rise, under the assumption of a given set of cut-off values (such as the deciles), to the Receiver Operating Curve, and the area under it (AUROC), as a main model accuracy measure (see e.g. [8] and [14]). Wider discussions on the ROC curve and the associated AUROC summary measure as tools for the evaluation of predictive classifications may be found in several research papers over time, such as, for instance, [4], [9], [10] and [11], among others.

Although extensively employed, the ROC and the AUROC depend on the choice of the cut-off points. To overcome this limitation, [14] suggests to employ the Somers' D measure (see [15]) in the context of credit rating accuracy measurement.

Somers' D is a conditional version of the Kendall's coefficient which maps each combi-

nation between one observed value and one predicted value into a simple binary measure: -1 in case of a discordant pair and +1 in case of a concordant pair.

Although a useful step in the correct measurement of predictive accuracy, being independent on the choice of cut-off points, Somers' D has two main drawbacks. First it is highly computational intensive as, for a set of N observations to be predicted, it calculates $\binom{N}{2}$ combinations. Second, it employs a rather crude binary summary which does not take into account the actual distance between each combination pair. This can be a serious drawback, especially in the light of a possible generalisation to applications that involve ordinal or continuous response variables.

In this paper we propose a new predictive classification accuracy measure that attempts to solve the above drawbacks building up on previous research contributions in the field of dependence analysis (see e.g. [5]). More precisely, the proposed measure will be built on the Lorenz curve, obtained by reordering the observed response variable values in non-decreasing sense and on the concordance curve, obtained by reordering the observed response variable values re-arranged with respect to the corresponding predicted values. In other words, our proposal is based on a comparison between the observed and the predicted response variable ranks, as in Somers' D coefficient but using, rather than the ranks themselves, the actual values of the response variable corresponding to both ranks. The result is a measure which, although requiring only an order among the values of the response variable, takes into account not only the concordance but also the distance between observed and predicted pairs.

We remark that taking distances into account can allow the application of our measure not only to binary classification problem but to all predictive problems, independently of the nature of the considered response variable. In the literature, some attempts have indeed

tried to unify model predictive accuracy criteria across different types of response variable: see e.g. [13] and [1]. Our work offers a contribution that can also be generalised but in a different way, at the ordinal level rather than at the continuous level.

The paper is organized as follows. Section 2 introduces the predictive accuracy measures routinely used. Section 3 introduces our proposal, and its comparison with Somers' coefficient. Section 4 illustrates a statistical test for predictive accuracy, based on the proposed measure. Section 5 presents the results obtained applying the developed methods to the credit scoring context. Section 6 presents the results obtained applying the measure to a simulated dataset. We conclude with a final discussion.

2 Background

One of the most commonly used tool for assessing the predictive accuracy of a scoring model is the Receiver Operating Characteristic (ROC) curve. In the context of credit scoring, suppose to consider n borrowers, such that n equals to $n_D + n_{ND}$, where D and ND represent the sets of defaulting and non-defaulting borrowers, respectively. Let S_D denote the distribution of the credit scores of the defaulting borrowers and S_{ND} denote the distribution for the non-defaulting borrowers. For any cut-off value C , we define $F_D(C) = P(S_D \leq C)$ and $F_{ND}(C) = P(S_{ND} \leq C)$. Let $F_D(C)$ and $F_{ND}(C)$ correspond to the sensitivity (true positive rate) and 1-specificity (false positive rate) of a rating model based on the cut-off value C (see e.g. [3]).

For the cut-off values, the ROC curve is the plot of $F_D(C)$ versus $F_{ND}(C)$ characterised by the set of points (G_{ND_i}, G_{D_i}) , where $G_{ND_i} = \sum_{j=1}^n p_{ND_j}$, $G_{D_i} = \sum_{j=1}^n p_{D_j}$, $p_{ND_i} = P(S_{ND_i} = s_i)$, $p_{D_i} = P(S_{D_i} = s_i)$ and $i = 1, \dots, n$. The area under the ROC curve

(AUROC) is usually considered as a summary indicator of predictive performance. It is computed as:

$$AUROC = \frac{1}{2} \sum_{i=1}^n (G_{D_i} + G_{D_{i-1}})(G_{ND_i} - G_{ND_{i-1}}).$$

Note that the AUROC equals 0.5 for a random model without any predictive accuracy and 1 for a perfect model. In the intermediate situations of reasonable model predictive accuracy, AUROC takes values in the range (0.5, 1). Formally, it can be shown that the area under the ROC curve can also be calculated using the following formula (see e.g. [14])

$$AUROC = \frac{1}{n_D n_{ND}} \sum_{i=1}^{n_D} \sum_{j=1}^{n_{ND}} c_{ij}, \quad (1)$$

where

$$c_{ij} = \begin{cases} 1, & \text{if } S_{N_i} < S_{ND_j} \\ \frac{1}{2}, & \text{if } S_{N_i} = S_{ND_j} \\ 0, & \text{if } S_{N_i} > S_{ND_j}. \end{cases} \quad (2)$$

Another summary indicator of predictive performance is the Accuracy Ratio (AR). It can be derived using equations (1) and (2) replacing 1, $\frac{1}{2}$ and 0 with 1, 0 and -1. As illustrated by [3], the relationship $AR = 2AUROC - 1$ holds.

A generalization of the Accuracy Ratio is Somers' D measure (see [15]). The Somers' D measure (henceforth D_{XY}) requires to specify the role of the two involved variables, i.e. the target variable and the predictor. Following [14], let Y be the response variable and X be the predictor variable. Let n be the sample size including both the set of defaulting and non-defaulting borrowers. For the sake of simplicity, sort the values of Y in increasing sense, such that $Y_i \leq Y_j$ for $i < j$. Equation in (2) can be re-expressed as follows:

$$c_{ij} = \begin{cases} 1, & \text{if } X_i < X_j, Y_i < Y_j \\ -1, & \text{if } X_i > X_j, Y_i < Y_j \\ 0, & \text{else.} \end{cases} \quad (3)$$

Somers' D_{XY} can then be computed as

$$D_{XY} = \frac{1}{n_u} \sum_{i=1}^n \sum_{j>i} c_{ij}, \text{ with } n_u = \sum_{i=1}^n \sum_{j>i} 1_{[Y_i \neq Y_j]}. \quad (4)$$

Remark 1 Differently from the AUROC measure, the AR and D_{XY} measures take values in the closed range $[-1, +1]$. In addition, while the AUROC depends on the arbitrary choice of different cut-off points, the AR and Somers' D_{XY} measures provide the same information but are computed using the concordance and discordance between pairs of observations, and do not require the choice of cut-off points.

3 Proposal

3.1 Methodology

Let $\mathbf{y}$ be a vector of observed values to be predicted and let $\hat{\mathbf{y}}$ be the vector of the corresponding predicted values, computed through a specific model $f(\mathbf{X})$, where $\mathbf{X}$ is the matrix containing the observations on the explanatory variables.

Our goal is to compare different models: $\hat{y} = f^1(\mathbf{X})$, $\hat{y} = f^2(\mathbf{X})$, $\dots$, using a general methodology that applies to all the response variable measurement scales.

3.1.1 The concordance curve

Let Y be a quantitative target variable and let $X_1, X_2, \dots, X_p$ be a set of p explanatory variables.

Let D be the available data, which we assume, for the time being, formed by a matrix with $P + 1$ columns: the p explanatory variables plus the response variable; and $m = t + n$ rows, corresponding to all the joint observations of Y and $X_1, X_2, \dots, X_p$, divided into a training set D_t , of dimension $(p + 1) \times t$, on which the parameters of the proposed statistical models are estimated; and a validation set D_n , of dimension $(p + 1) \times n$, which will be used to evaluate the performance of the model.

A statistical predictive model will be evaluated comparing, in the test set, the observed values of a response variable Y with the predicted values for the same variable, obtained applying the model built on D_t , to the values of the explanatory variables $X_1, X_2, \dots, X_p$ in D_n . For example, if the assumed model is a multiple linear regression model, the predicted values are obtained as $\hat{Y} = E(Y|X_1, \dots, X_p) = \hat{\alpha} + \hat{\beta}_1 X_1 + \dots + \hat{\beta}_p X_p$.

As a result, we will obtain an n -dimensional vector $\hat{Y}$ of n predicted values for the response variable which can be paired to the n -dimensional vector Y of the observed values, to form a measure of predictive accuracy.

[6] proposed a comparison between Y and $\hat{Y}$ which, based on an ordinal scale, is potentially useful for our scope. We review below the main steps of their proposal.

The Y values can be used to build the Y Lorenz curve (see e.g. [12]), denoted with L_Y , re-ordering them in non-decreasing sense. More formally, the curve is characterised by the following pairs: $(i/n, \sum_{j=1}^i y_{r_j})$, for $i = 1, \dots, n$, where r_i indicates the (non-decreasing) ranks of Y .

The same Y values can also be used to build the Y dual Lorenz curve, denoted with

L'_Y , obtained by re-ordering the Y variable values in a non-increasing sense. More formally, the curve is characterised by the following pairs: $(i/n, \sum_{j=1}^i y_{d_j})$, for $i = 1, \dots, n$, where d_i indicates the (non-increasing) ranks of Y .

The predicted $\hat{Y}$ values can also be re-ordered, in a non-decreasing sense. Let $\hat{r}_i$, for $i = 1, \dots, n$, indicate the (non-decreasing) ranks of $\hat{Y}$. [6] suggested to build a concordance curve C , based on ordering the Y values with respect to the ranks of the predicted $\hat{Y}$ values, $\hat{r}_i$. Formally, the concordance curve is characterised by the pairs: $(i/n, \sum_{j=1}^i y_{\hat{r}_j})$, for $i = 1, \dots, n$, where $\hat{r}_i$ indicates the (non-decreasing) ranks of $\hat{Y}$.

To illustrate the previous concept, Figure 1 reports, for a given set of test values Y and the corresponding predictions $\hat{Y}$: the Lorenz curve, the dual Lorenz curve and the concordance curve, together with the bisector curve $(i/n, i/n)$, for $i = 1, \dots, n$. To ease the illustration, all values have been normalised using the sum of all Y values: $(n\bar{y})$, where $\bar{y}$ indicates the mean of Y .

From Figure 1 note that the Lorenz curve and its dual are symmetric around the bisector curve, and that the concordance curve lies between them.

Note also that, when $\hat{r}_i = r_i$, for all $i = 1, \dots, n$, the concordance curve is equal to the Lorenz curve, and a perfect concordance between the Y values and the corresponding predictions arises. On the other hand, when $\hat{r}_i = d_i$, the concordance curve is equal to the dual Lorenz curve and a perfect discordance between the Y values and the corresponding predictions emerges. In general, for any given point, a discrepancy between the Lorenz curve and the concordance curve arises only when the predicted rank is different from the observed one.

We finally remark that, when the $\hat{Y}$ values are all equal each other, the concordance C curve perfectly overlaps with the bisector curve. In this case, the model has no predictive

Figure 1: The L_Y and L_Y' Lorenz curves and the C concordance curve, normalised.

capability, as it coincides with a random prediction of the Y values.

3.1.2 The RGA index

The concordance curve, and its relationship with the Lorenz and the dual Lorenz curve can be exploited to summarise the “distance” between the Y and the $\hat{Y}$ values, in terms of the “discrepancy” between their corresponding ranks. A summary index, named C as the concordance curve, can be defined as follows:

$$C = \frac{\sum_{i=1}^n \{i/n - (1/(n\bar{y})) \sum_{j=1}^i y_{\hat{r}_j}\}}{\sum_{i=1}^n \{i/n - (1/(n\bar{y})) \sum_{j=1}^i y_{r_j}\}}, \quad (5)$$

where y_{r_j} are the Y variable values ordered according to the ranks r_j ; $y_{\hat{r}_j}$ are the same values but ordered according to the ranks $\hat{r}_j$.

From equation (5) note that the C index is a function of the y -axis values of the points lying on the concordance C curve and of the y -axis values of the points lying on the Lorenz curve L_Y . Indeed the numerator of the index in (5) compares the distance between the set of points lying on the bisector curve and the set of points lying on the concordance C curve, while the denominator compares the distance between the set of points lying on the bisector curve and the set of points lying on the Lorenz curve L_Y .

Consistently with the index being a summary of the concordance curve, it can be shown to have the following properties:

- $-1 \leq C_{Y,X_1,X_2,\dots,X_p} \leq +1$: specifically, when $0 < C_{Y,X_1,X_2,\dots,X_p} \leq +1$, Y and $\hat{Y}$ are concordant and when $-1 \leq C_{Y,X_1,X_2,\dots,X_p} < 0$ they are discordant;
- $C_{Y,X_1,X_2,\dots,X_p} = +1$ if and only if $C = L_Y$ (full concordance): the concordance C curve overlaps with the Lorenz curve L_Y ;
- $C_{Y,X_1,X_2,\dots,X_p} = -1$ if and only if $C = L'_Y$ (full discordance): the concordance C curve overlaps with the dual Lorenz curve L'_Y .

Remark 2 Note that, when some of the $\hat{Y}$ values are equal to each other, the original Y values associated with the equal $\hat{Y}$ values can be substituted by their mean, as suggested by [5]. This adjustment is coherent with the definition of a model without predictive capability. To illustrate this point, suppose to consider a general model $f(X)$ with only one explanatory variable, such that $\hat{Y} = E(Y|X) = E(Y) = \bar{y}$ holds for any value of X . Since a re-ordering problem arises if the response variable values are associated with

equal estimated values, the response variable values corresponding to the same estimated values are replaced by their mean. As a result, the resulting concordance curve C overlaps with the bisector curve, whose co-ordinates are given by the set of pairs $(i/n, i/n)$. This can be easily shown considering the normalised set of pairs characterising the C concordance curve $(i/n, \sum_{j=1}^i y_{\hat{r}_j}/n\bar{y})$. In the case in which $\hat{y}_i = \bar{y}, \forall i = 1, \dots, n$, we obtain $(i/n, \sum_{j=1}^i y_{\hat{r}_j}/n\bar{y}) = (i/n, \sum_{j=1}^i \bar{y}/n\bar{y}) = (i/n, i\bar{y}/n\bar{y}) = (i/n, i/n)$.

Looking more closely at equation (5) note that, when different models are compared, the denominator does not change, while the numerator does. It is therefore intuitive to compare models in terms of differences between the distances expressed by the numerator of formula (5), leading to the following:

$$C_{num} = \sum_{i=1}^n \left\{ i/n - (1/(n\bar{y})) \sum_{j=1}^i y_{\hat{r}_j} \right\}. \quad (6)$$

The above measure suffers from a drawback: positive values of the index may be compensated by negative values, as illustrated by the example in Figure 2 below.

To overcome the compensation issue illustrated in Figure 2, we can resort to the squared distance between the set of points lying on the C concordance curve and the set of points lying on the bisector curve. Indeed, as the bisector curve defines the situation of a random, non predictive model, for which the Y values are independent on the $\hat{Y}$, we can interpret the squared distance as the difference between the observed and the expected concordance values of Y , where by expected we mean the concordance values that we would have with a random model. If we divide the difference by the expected values themselves, we obtain a new index, that can be called RGA (Rank Graduation Accuracy) index, defined as:

Figure 2: The C concordance curve in the compensation case

$$RGA = \sum_{i=1}^n \frac{\left\{ (1/(n\bar{y})) \sum_{j=1}^i y_{\hat{r}_j} - i/n \right\}^2}{i/n}. \quad (7)$$

Through some manipulations, an equivalent version of (7) can be further derived as

$$RGA = \sum_{i=1}^n \frac{\{C(y_{\hat{r}_i}) - i/n\}^2}{i/n}, \quad (8)$$

which emphasises the role of the quantity $C(y_{\hat{r}_j}) = \sum_{j=1}^i y_{\hat{r}_j} / \sum_{i=1}^n y_{r_i}$, that represents the cumulative values of the (normalised) response variable.

To better illustrate the RGA index, we now introduce two examples to clarify its computation. in the first one, there are no tied observations; in the second there are.

Example 1: let Y and X be two variables, whose observed values are $y = \{15, 10, 26, 21, 32, 45\}$

and $x = \{1.5, 2.3, 1.7, 3.2, 3.8, 2.1\}$ and suppose to consider a linear regression model. Through the least squares method, the fitted line is $\hat{y} = 18.476 + 2.613x$ and the corresponding predicted values are $\hat{y} = \{22.39, 24.48, 22.92, 26.84, 28.4, 23.96\}$. If we order the response variable values according to the ranks of the predictions we obtain that $y_{\hat{r}} = \{15, 26, 45, 10, 21, 32\}$. Thus, being $\sum_{i=1}^6 y_i = \sum_{i=1}^6 y_{r_i} = 149$ it follows that

$$RGA = \frac{\left(\frac{15}{149} - \frac{1}{6}\right)^2}{\frac{1}{6}} + \frac{\left(\frac{41}{149} - \frac{2}{6}\right)^2}{\frac{2}{6}} + \frac{\left(\frac{86}{149} - \frac{3}{6}\right)^2}{\frac{3}{6}} + \frac{\left(\frac{96}{149} - \frac{4}{6}\right)^2}{\frac{4}{6}} + \frac{\left(\frac{117}{149} - \frac{5}{6}\right)^2}{\frac{5}{6}} + \frac{\left(\frac{149}{149} - \frac{6}{6}\right)^2}{\frac{6}{6}} \approx 0.052.$$

Example 2: let Y be the same as in Example 1 and X a variable with tied values: $x = \{1.5, 2.3, 1.5, 3.2, 3.2, 2.3\}$. The fitted line is $\hat{y} = 16.823 + 3.433x$ and the corresponding predicted values are $\hat{y} = \{21.97, 24.72, 21.97, 27.81, 27.81, 24.72\}$. Since $\hat{y}$ presents three pairs of equal values, the corresponding y values are substituted according to their mean values, so that $y = \{20.5, 27.5, 20.5, 26.5, 26.5, 27.5\}$, with $20.5 = (15 + 26)/2$, $27.5 = (10 + 45)/2$ and $26.5 = (21 + 32)/2$. The response variable values can then be ordered according to the corresponding predicted values, to compute the RGA index as:

$$RGA = \frac{\left(\frac{20.5}{149} - \frac{1}{6}\right)^2}{\frac{1}{6}} + \frac{\left(\frac{41}{149} - \frac{2}{6}\right)^2}{\frac{2}{6}} + \frac{\left(\frac{68.5}{149} - \frac{3}{6}\right)^2}{\frac{3}{6}} + \frac{\left(\frac{96}{149} - \frac{4}{6}\right)^2}{\frac{4}{6}} + \frac{\left(\frac{122.5}{149} - \frac{5}{6}\right)^2}{\frac{5}{6}} + \frac{\left(\frac{149}{149} - \frac{6}{6}\right)^2}{\frac{6}{6}} \approx 0.019.$$

Note that the RGA index takes values between 0 and RGA_{max} . Its maximum value is obtained when the predicted ranks order the response variable values in full concordance (or full discordance) with the observed ranks. It can be used to normalise the values of the

RGA index, obtaining a measure that is bounded between 0 and 1. It is worth remarking that all models with the same predicted ranks provide the same value of the RGA index. Note also that the RGA proposal is not sensitive to the nature of the response variable and, therefore, could be employed also when the response variable is ordinal or continuous. Here we focus on credit scoring, which involves a binary response variable.

Given a set of p explanatory variables, in credit scoring applications the response variable Y takes one of two possible outcomes, according to the presence (typically denoted with value 1) or the absence (typically denoted with value 0) of the attribute of interest. In such case, one typically resorts to binary response models (for instance, logistic regression models and classification trees) to directly model the response probabilities $P(y_i = 1)$ of the dependent variable Y . In terms of our diagnostic measure, the response variable Y values can be re-ordered according to the predicted values $P(y_i = 1)$, which indeed take real values.

The possible behaviours of the concordance curve is illustrated in Figure 3.

Figure 3 illustrates the three alternative scenarios that can arise, in which Y and $\hat{Y}$ are: a) perfectly concordant, b) perfectly discordant and c) partially concordant (discordant).

Looking more closely at Figure 3 note that the C concordance curve has a behavior which is similar to the Receiver Operating Characteristic (ROC) curve (see e.g. [2]), typically used as a predictive accuracy diagnostic for binary response variables. However, while the ROC curve is built ordering cut-off points in an arbitrary way, the C concordance overcomes this subjectivity issue, as the ordering is based on the predicted values themselves. More precisely, the number of points on which the concordance curve is constructed is equal to the number of observations. For each observation the RGA index compares the values of the actual response, which in the binary case can be either 0 or 1, ordered in one

Figure 3: The concordance C curve behavior in the case of binary response

case according to the ranks of the observed response, in the other according to the ranks of the predicted response. We have perfect concordance (Figure 3 a)) when the ranks coincide on all observations; perfect discordance (Figure 3 b)) when the ranks are in reverse correspondence.

We remark that both the RGA and D_{XY} resort to the notions of concordance and discordance and do not depend on the choice of cut-off points. However, the notions of concordance and discordance come into play in a different manner. D_{XY} considers concordance and discordance between pairs of observations from the response variable Y and the predictor X , while the RGA measure focuses on the concordance or discordance between the ranks of the observed target variable Y values and those of the corresponding predicted $\hat{Y}$ values. In addition, while D_{XY} is computed by assigning value +1 to the term c_{ij} in equation (4), if a pair is concordant, and -1, if a pair is discordant, the RGA measure calculates a more general quantitative distance between the non-decreasing ordering of

the observed response variable Y values with respect to their re-ordering based on the corresponding predicted $\hat{Y}$ values.

To compare the two measures in real applications, we need to redefine Somers' coefficient. D_{XY} should be re-expressed by setting $X = \hat{Y}$, where the $\hat{Y}$ values are the model predictions. The focus is then on the count of the concordant and discordant pairs of the predicted $\hat{Y}$ values and the observed Y values. This version, henceforth denoted with $D_{\hat{Y}Y}$, is not penalized against ties on Y so that the denominator in (4) refers to the total number of pairs. Moreover, $D_{\hat{Y}Y}$ has to be re-scaled in the closed range $[0, 1]$. For comparison purposes, we consider $|D_{\hat{Y}Y}|$, being the sign an indicator of the direct or inverse relationship between Y and $\hat{Y}$.

3.1.3 Testing predictive accuracy

In this subsection we derive the statistical distribution of the RGA index, with the aim of building a statistical test that can detect whether a given model is significantly more accurate than a random model.

Let us denote with t' the statistics computed on the sample data and corresponding to the RGA index expression defined in equation (7), i.e.

$$t' = \sum_{i=1}^n \frac{\left\{ (1/(n\bar{y})) \sum_{j=1}^i y_{\hat{r}_j} - i/n \right\}^2}{i/n}. \quad (9)$$

As a test statistics, we consider a transformation of the t' statistics in (9). Let t indicate the new statistics provided by

$$t = n\bar{y}t' = n\bar{y} \sum_{i=1}^n \frac{\left\{ (1/(n\bar{y})) \sum_{j=1}^i y_{\hat{r}_j} - i/n \right\}^2}{i/n}. \quad (10)$$

Remark 3 The statistics t in (10) is equivalent to

$$t = \sum_{i=1}^n \frac{(\sum_{j=1}^i y_{\hat{r}_j} - i\bar{y})^2}{i\bar{y}}. \quad (11)$$

Proof It is worth noting that the term i/n , appearing both at numerator and denominator of equation (10), can be written as $(1/(n\bar{y})) \sum_{j=1}^i y_{\hat{r}_j}^e$, where $y_{\hat{r}_j}^e$ are the expected concordance values associated with a random model. Thus,

$$\begin{aligned} t = n\bar{y}t' &= n\bar{y} \sum_{i=1}^n \frac{\left\{ (1/(n\bar{y})) \sum_{j=1}^i y_{\hat{r}_j} - (1/(n\bar{y})) \sum_{j=1}^i y_{\hat{r}_j}^e \right\}^2}{(1/(n\bar{y})) \sum_{j=1}^i y_{\hat{r}_j}^e} \\ &= \sum_{i=1}^n \frac{\left\{ \sum_{j=1}^i y_{\hat{r}_j} - \sum_{j=1}^i y_{\hat{r}_j}^e \right\}^2}{\sum_{j=1}^i y_{\hat{r}_j}^e}. \end{aligned} \quad (12)$$

Since according to the definition of the random model scenario it results that $y_{\hat{r}_j}^e = \bar{y}$, $\forall j = 1, \dots, i$, through some manipulations the new expression of t is derived as follows:

$$t = \sum_{i=1}^n \frac{\left\{ \sum_{j=1}^i y_{\hat{r}_j} - \sum_{j=1}^i \bar{y} \right\}^2}{\sum_{j=1}^i \bar{y}} = \sum_{i=1}^n \frac{(\sum_{j=1}^i y_{\hat{r}_j} - i\bar{y})^2}{i\bar{y}}. \quad \blacksquare$$

Consider the test statistics T provided by

$$T = \sum_{i=1}^n \frac{\left\{ \sum_{j=1}^i Y_{\hat{r}_j} - \sum_{j=1}^i Y_{\hat{r}_j}^e \right\}^2}{\sum_{j=1}^i Y_{\hat{r}_j}^e}, \quad (13)$$

whose sample version is defined in (12). We can thus obtain the distribution of the test statistics, as follows.

Remark 4 Let $Y_{\hat{r}_1}, \dots, Y_{\hat{r}_i}, \dots, Y_{\hat{r}_n}$ be n independent and identically distributed random variables, with $i = 1, \dots, n$. When the response variable is binary, we can assume that $Y_{\hat{r}_i} \sim \text{Poisson}(\mu)$ with $E(Y_{\hat{r}_i}) = \mu$ and $\text{Var}(Y_{\hat{r}_i}) = \mu$, $\forall i = 1, \dots, n$. Under the null hypothesis and without loss of generality, $Y_{\hat{r}_j}^e = \mu_0$, for all $j = 1, \dots, i$, with μ_0 corresponding to a known value. As $\sum_{j=1}^i Y_{\hat{r}_j} \sim \text{Poisson}(i\mu)$, for i sufficiently large $Z = (\sum_{j=1}^i Y_{\hat{r}_j} - i\mu_0)/(\sqrt{i\mu_0})$ becomes a standard normal distribution. It follows that the test statistics $T = \sum_{i=1}^n \left\{ (\sum_{j=1}^i Y_{\hat{r}_j} - i\mu_0)/(\sqrt{i\mu_0}) \right\}^2$ can be written as $T = \sum_{i=1}^n Z_i^2 \sim \chi_n^2$.

Note that the previous test can be generalised to the case in which the response variable is continuous. We can categorize a continuous variable into k classes, for example according to equally spaced intervals, so that all classes have different frequencies. Doing so, one can resort to the test statistics T specified above. We remark that this is not a great loss of information, as discretisation is applied only after the calculation of the concordance curve, and only for the purpose of the test.

4 Application

In this section we apply the RGA index to measure the predictive accuracy of alternative logistic regression models employed in credit scoring applications. In particular, we consider data supplied by a European External Credit Assessment Institution (ECAI) specialized in

credit scoring for P2P platforms focused on SME commercial lending. The analyzed dataset is composed of end-of-year 2015 official financial information (balance-sheet variables) for 15045 South-European SMEs, for which the information about the status (0 = active, 1 = defaulted) one year later (2016) is also provided. Many of the companies included in the sample have less than 20 employees and operate in the manufacturing sector. More details about the data can be found in [7].

Table 1 lists the financial ratios included in our dataset, together with their type (continuous or dichotomous).

The proportion of defaulted companies within the dataset is 10.9%. To show how the RGA index can contribute to a model selection procedure, we perform the following analysis. We split the dataset into a training and a test subsample, corresponding to 70% and 30% of the whole sample respectively. We then perform a stepwise logistic regression on the training dataset. The R output of the implemented stepwise procedure is reported in Figure 4. From the variables included in the model selected by the stepwise algorithm, we remove those which are not significant at a level of 1% or less. We end up with a model that includes 9 regressors from the original 23 reported in Table 1.

Then, we estimate all possible models obtained by combining subsets of the 9 predictors on the test dataset. In particular, we consider subsets with a number of predictors from 1 to 8. For each model we calculate the RGA index, the Somers' D and the AUROC based on the fitted values on the training dataset. The boxplots in Figure 5 and 6 represent the distribution of the three measures for each considered cardinality of predictors.

The boxplots in Figures 5 and 6 show that the variability of the RGA index across the estimated models is larger than that associated with the other measures, except for the case when only one predictor is considered. This finding can be interpreted as a major

ID	Formula or Description	Type
1	Total Assets/Equity	Continuous
2	(Long term debt + Loans)/Shareholders Funds	Continuous
3	Total Assets/Total Liabilities	Continuous
4	Current Assets/Current Liabilities	Continuous
5	(Current assets - Current assets: stocks)/Current liabilities	Continuous
6	Shareholders Funds + Non current liabilities)/Fixed assets	Continuous
7	EBIT/interest paid	Continuous
8	(Profit or Loss before tax + Interest paid)/Total assets	Continuous
9	Return on Equity (ROE)	Continuous
10	Operating revenues/Total assets	Continuous
11	Sales/Total assets (Activity Ratio)	Continuous
12	Interest paid/(Profit before taxes + Interest paid)	Continuous
13	EBITDA/interest paid (Solvency ratio)	Continuous
14	EBITDA/Operating revenues	Continuous
15	EBITDA/Sales	Continuous
16	EBIT Dummy (=1 if EBIT<0, 0 otherwise)	Dichotomous
17	Profit before tax Dummy (=1 if Profit before tax<0, 0 otherwise)	Dichotomous
18	Financial Profit Dummy (=1 if Financial Profit<0, 0 otherwise)	Dichotomous
19	Net Profit Dummy (=1 if Net Profit<0, 0 otherwise)	Dichotomous
20	Trade Payables/Operating Revenues	Continuous
21	Trade Receivables/Operating Revenues	Continuous
22	Inventories/Operating Revenues	Continuous
23	Turnover	Continuous

Table 1: List of financial ratios used as independent variables.

```

Coefficients:
      Estimate Std. Error z value Pr(>|z|)
(Intercept) -1.768e+00  1.715e-01 -10.309 < 2e-16 ***
id_1         3.212e-03  1.521e-03   2.112  0.03465 *
id_3        -5.538e-01  1.144e-01 -4.839  1.30e-06 ***
id_4        -3.351e-01  6.943e-02 -4.826  1.39e-06 ***
id_6         4.775e-03  1.584e-03   3.015  0.00257 **
id_7         3.418e-03  1.645e-03   2.078  0.03773 *
id_8        -2.829e+00  3.588e-01 -7.884  3.18e-15 ***
id_9        -6.001e-02  4.175e-02 -1.438  0.15058
id_10       -2.745e-01  1.615e-01 -1.699  0.08923 .
id_11        3.717e-01  1.602e-01   2.320  0.02034 *
id_13       -3.029e-03  1.347e-03 -2.249  0.02451 *
id_15       -4.749e-01  1.973e-01 -2.407  0.01608 *
id_20        1.849e-03  2.954e-04   6.260  3.85e-10 ***
id_21        7.038e-04  2.681e-04   2.625  0.00867 **
id_23       -2.245e-05  7.658e-06 -2.932  0.00337 **
id_17        4.770e-01  1.631e-01   2.924  0.00345 **
id_19        6.236e-01  1.538e-01   4.055  5.01e-05 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

```

Figure 4: Logistic regression output for the model selected through the R stepwise procedure.

capability of the RGA index in discriminating between different model specifications based on the measured accuracy. While Somers' D works better in an exploratory low dimensional approach (e.g. when one predictor is considered), the RGA measure is better in all other cases. Note that the relative advantage of the RGA decreases as the number of predictors increase, and that the three measures converge when the number of predictors approaches that of the optimal model.

We underline that the superior predictive accuracy of the RGA index, with respect to the AUROC summary, stems from the different number of evaluation points. While the AUROC is calculated at a selected set of cut-off points, the RGA is calculated at all observed response values. This makes it more sensible to model variations. On the other hand, Somers' D is based on the binarisation of model errors, which makes it less sensible than the RGA.

To further assess the impact of an increasing model complexity on the three measures,

Figure 5: Distribution of RGA index, Somers' D and AUROC over the models estimated on credit rating data (topleft: 1 predictor; topright: 2 predictors; bottomleft: 3 predictors; bottomright: 4 predictors)

we consider their trend with respect to the number of regressors. In particular, Figure 7 shows the value of RGA index, Somers' D and AUROC calculated on the best model - the one for which the analyzed measure is maximum - for each cardinality of predictors from 1 to 8 (the values have been previously normalized so as to represent them on a 0-1 scale). We stop at 8 predictors, as the best model contains 9 predictors, and we would like to understand which predictive accuracy measure approaches faster such best model.

To help understanding the best performing measure, Figure 8 shows the elasticity (relative change) of the measures with respect to an increase in the number of regressors.

Figure 7 shows that the AUROC becomes nearly flat starting from a number of regres-

Figure 6: Distribution of RGA index, Somers' D and AUROC over the models estimated on credit rating data (topleft: 5 predictors; topright: 6 predictors; bottomleft: 7 predictors; bottomright: 8 predictors)

sors equal to 4. The behaviour of the RGA and Somers' D is quite similar. However, when moving from a 1 to a 2-regressor specification, the increase in the RGA is the largest. This is confirmed by Figure 8, which also shows that the RGA index dominates the others in terms of elasticity for all the considered cardinalities.

We can thus conclude that the RGA measure allows a more rapid convergence towards the best model, in the sense that it provides the best approximation to it, when a lower dimensional predictor space is considered.

Figure 7: RGA index, Somers' D and AUROC for increasing number of predictors

Figure 8: Relative change of RGA index, Somers' D and AUROC for increasing number of predictors

5 Simulation study

In this section we present a simulation study aimed at evaluating the robustness of the RGA, AUROC and Somer's D to misspecification. Specifically, using three of the financial indicators included in our dataset, we draw 1500 realizations of a 0-1 variable, with probability of the "1" occurrence given by:

$$PD = \frac{\exp(\beta_0 + \beta_1 \text{ROE} + \beta_2 \text{Activity Ratio} + \beta_3 \text{Solvency Ratio})}{1 + \exp(\beta_0 + \beta_1 \text{ROE} + \beta_2 \text{Activity Ratio} + \beta_3 \text{Solvency Ratio})}$$

where $\beta_0 = -1.8$ is a constant term, $\beta_1 = -0.7$, $\beta_2 = -0.2$, $\beta_3 = -0.06$ are the coefficient associated with ROE (id9), Activity Ratio (d11) and Solvency Ratio (d13) respectively. The set values of the regression coefficients lead to a simulated sample of companies for which the percentage of defaulted companies is 12.1%.

As, in the simulated dataset, the "true" model contains three regressors we estimate all possible models obtained by combining 3 of the the 23 available predictors, on the same sample. After calculating the values of RGA Index, Somers' D and AUROC for each model, we consider the measured accuracy of the correctly specified model with respect to that of misspecified models. The obtained results are represented in Figure 9.

Figure 9 clearly shows that the variability of the RGA index is larger than that observed for the other predictive accuracy metrics.

Indeed, the ratio between the mean RGA index for the misspecified models and the RGA value for the correct model is 0.58, while the corresponding ratio is 0.72 for Somers' D and 0.88 for AUROC. According to these results and coherently with the findings of Section 4, the RGA index turns out to be the most capable to discriminate between alternative specifications among the three considered measures.

Figure 9: Distribution of RGA index, Somers’ D and AUROC over the models estimated on the simulated sample

6 Discussion

In the paper we have proposed a new tool to evaluate the predictive accuracy of data science models.

The motivation of our proposal is to develop a measure that does not depend on thresholds or cut-off points, and that is potentially extendable to any type of model and response variable.

To fulfill this goal, we have proposed the *RGA* index, which is based on the calculation of the cumulative values of the response variable, according to the ranks of the values predicted by a given model.

The proposed measure, besides being mathematically elegant, and relatively easy to implement, has been found quite effective in a real credit scoring application. It overperforms the AUROC and Somers' D in discriminating between "good" and "bad" models.

Future research extensions should consider the application of the measure to other real examples, possibly involving ordered or continuous response variables.

We believe that the main beneficiaries of our results may be academicians but also policy makers, regulators and supervisors, interested in weighing the benefits of data science applications against their potential inaccuracies.

References

- Agresti A. and Tarantola, C. (2018). Simple ways to interpret effects in modeling ordinal categorical data. *Statistica Neerlandica*, 72, 210-223
- DeLong, E.R., DeLong, D.M. and Clarke-Pearson, D.L. (1988). Comparing the Areas under Two or More Correlated Receiver Operating Characteristic Curves: A Nonparametric Approach. *Biometrics*, 44(3), 837-845
- Engelmann, B. (2006). Measures of a Rating's Discriminative Power-Applications and Limitations. *The Basel II Risk Parameters*, Springer
- Fawcett T. (2006). An introduction to ROC analysis, *Pattern Recognition Letters*, 27, 861-874
- Ferrari, P.A. and Raffinetti, E. (2015). A Different Approach to Dependence Analysis. *Multivariate Behavioral Research*, 50(2), 248-264

- Giudici, P. and Raffinetti, E. (2011). On the Gini measure decomposition. *Statistics & Probability Letters*, 81(1), 133-139.
- Giudici, P., Hadji-Misheva, B. and Spelta, A. (2019). Network based credit scoring. *Quality Engineering*, to appear.
- Hand, D., Mannila, H., and Smyth, P. (2001). *Principles of data mining*. Adaptive Computation and Machine Learning Series. MIT Press
- Hanley, J.A. (1989). Receiver operating characteristic (ROC) methodology: the state of the art, *Critical Reviews in Diagnostic Imaging*, 29(3), 307-335
- J.A. Hanley, J.A. and McNeil, B.J. (1982). The meaning and use of the area under an ROC curve, *Radiology*, 143, 29-36
- Krzanowski, W.J. and Hand, D.J. (2009). *ROC curves for continuous data*. Chapman and Hall, London
- Lorenz, M.O. (1905). Methods of Measuring the Concentration of Wealth. *Journal Publications of the American Statistical Association*, 9(70), 209-219
- McFadden, D. (1974). Conditional logit analysis of qualitative choice behavior. In *Frontiers in Econometrics*, ed. P. Zarembka, New York: Academic Press, 105-42
- Orth, W. (2012). The predictive accuracy of credit ratings: Measurement and statistical inference, *International Journal of Forecasting* 28, 288-296
- Somers, R.H. (1962). A new asymmetric measure of association for ordinal variables, *American Sociological Review* 27(6), 799-811.