Folia Oeconomica
Acta Universitatis Lodziensis
ISSN 0208-6018 e-ISSN 2353-7663

www.czasopisma.uni.lodz.pl/foe/
4(343) 2019

DOI: http://dx.doi.org/10.18778/0208-6018.343.02

Adam Piotr Idczak
University of Łódź, Faculty of Economics and Sociology, Department of Statistical Methods
adam.idczak@uni.lodz.pl

Remarks on Statistical Measures for Assessing Quality of Scoring Models

Abstract: Granting a credit product has always been at the heart of banking. Simultaneously, banks are obligated to assess the borrower's credit risk. Apart from creditworthiness, to grant a credit product, banks are using *credit scoring* more and more often. *Scoring models*, which are an essential part of credit scoring, are being developed in order to select those clients who will repay their debt. For lenders, high effectiveness of selection based on the scoring model is the primary attribute, so it is crucial to gauge its statistical quality.

Several textbooks regarding assessing statistical quality of scoring models are available, there is however no full consistency between names and definitions of particular measures. In this article, the most common statistical measures for assessing quality of scoring models, such as the pseudo Gini index, Kolmogorov-Smirnov statistic, and concentration curve are reviewed and their statistical characteristics are discussed. Furthermore, the author proposes the application of the well-known distribution similarity index as a measure of discriminatory power of scoring models. The author also attempts to standardise names and formulas for particular measures in order to finally contrast them in a comparative analysis of credit scoring models.

Keywords: credit scoring, scoring model quality, Lorenz and concentration curve, Gini index

JEL: C52

1. Introduction

Very rapid evolution of technology in recent years has meant that collection and processing of large volume datasets at a very low aggregation level has become available even for small companies. By means of statistical methods, these companies can derive valuable information from the data and be more competitive, make better decisions and reduce costs. In the banking sector, lenders may want to know if a given borrower will repay his or her debt. The answer to that is *credit scoring* which allows them to assess the borrower's risk. *Credit scoring* is simply "the use of statistical models to transform relevant data into numerical measures that guide credit decisions" (Anderson, 2007: 6). These numerical measures are called *scores* and they rank clients with respect to their credit risk. As for statistical models, it seems that logistic regression is the most widely used method for modelling credit risk. A number of various techniques, such as linear regression, discriminant analysis, mathematical programming, neutral networks, or decision trees, are also available.

By means of credit scoring, lenders can grant credit to new applicants or existing clients (cross-sell) and expand their business much more. Moreover, credit scoring is used to calculate the PD parameter which is an important part of calculating capital requirements in the advanced internal ratings-based approach.

When developing a *scoring model* (also known as a *scorecard*), it is crucial to evaluate its statistical quality. In other words, one needs to know how good a scoring model really is in a sense of its performance which is represented by its discriminatory power (i.e. the ability to distinguish those clients who will repay their debt and those who will not repay it). To measure that, there are several methods used in order to evaluate the performance of the scorecard and compare alternative models at the stage of the developing process or to evaluate performance of the scorecard as a part of its maintaining process.

Despite its relatively short history (dating back roughly to the 1950s), credit scoring has rapidly expanded in the field of finance during the last few decades (Abdou, Pointon, 2011). There exist a few books where statistical measures for assessing scoring models quality can be found (see Anderson, 2007; Crook et al., 2007; Finlay, 2010; Rezac, Kolacek, 2012; Siddiqi, 2017), but there are differences in names or symbols for particular statistics (such as the Kolmogorov-Smirnov statistic) to deal with. Names for particular curves also differ across publications (e.g.: the concentration curve). Moreover, these curves often vary in axes, so they are not equivalent (see: the Lorenz curve in Siddiqi, 2017).

The purpose of this paper is to review the most widely used statistical methods for gauging quality of a scoring model, standardise the above-mentioned differences in names and definitions of particular measures and discuss their main characteristics. Special attention has been paid to the measures related to the Lorenz curve and the Gini index. In addition, it is discussed what features are important in the process of developing a scoring model. In section 2, logistic regression is introduced as the most common approach to modelling credit risk. The next two paragraphs focus on graphical (section 3) and numerical (section 4) methods for assessing quality of a scoring model. These sections contain typical measures of discriminatory power of scoring models where the pseudo Gini index, Kolmogorov-Smirnov statistic, divergence and other methods are pointed out. In section 5, a case study based on a comparison of three scoring models is conducted.

2. Assessing the discriminatory power

First of all, it is crucial to outline what is actually modelled. Lenders are interested in the identification of those clients who will repay their debt and those who will not (a state called *default*). Let *Y* be the Bernoulli random variable that can take one of two values for each ($k = 1, \ldots, K$) observation:

$$Y_k = \begin{cases} 1, & \text{when default occurs} \\ 0, & \text{otherwise.} \end{cases} \tag{1}$$

The final scorecard will vary across a given *default* definition, so it is very important to be cautious when setting the dependent variable *Y*. In practice, the definition depends on the days past due (DPD) and the amount past due. That being said, a client is marked as "bad" (*default* occurred) when the DPD and the amount past due exceed a given threshold at a given time horizon, otherwise he/she is marked as "good" (*non-default*).

Logistic regression, which is the most common technique to assess client credit risk, is given by equation (Hosmer, Lemeshow, Sturdivant, 2013):

$$Logit(p_k) = \beta_0 + \hat{\boldsymbol{a}}^{\boldsymbol{T}} \boldsymbol{x}_k, \tag{2}$$

where:

$Logit(p_k) = ln\left(\frac{p_k}{1-p_k}\right)$ is the natural logarithm of odds ratio also called *score* and

p_k is the probability that the case k is a good client, i.e. $p_k(Y_k = 0)$,
β_0 is an intercept and β is a vector of estimated parameters,
$\boldsymbol{x}_k$ is a vector of explanatory variables with values for the case k.

Developing a scoring model, one has to keep in mind two main properties: discriminatory power (the ability of the model to distinguish good and bad clients) and accuracy (the ability of the model to predict default probabilities of clients).

The first stands for the degree of ranking ability, while the latter focuses on the model fitting to observed values of dependent variable. For lenders, power is the primary attribute, i.e. distinguishing good and bad clients, whereas accuracy may be secondary, and it can be attained through calibration (see: Anderson, 2007). Measuring power and accuracy should be part of any developing and maintaining process of a credit scoring model.

3. Graphical methods for assessing quality of scoring models

Lorenz and concentration curves

The Lorenz curves and concentration curves are widely used tools for the analysis of economic inequality and redistribution. The Lorenz curve (LC) was first introduced by Lorenz (1905) as a method of measuring the concentration of wealth. Let Y be a non-negative random variable, $f(y)$ its probability density function and $F(y)$ the cumulative distribution function of Y. Moreover, let $Q_y(p) = F_y^{-1}(p) = \inf\{y \mid F_Y(y) \geq p\}, p \, \varepsilon \, \langle 0;1 \rangle$ denote the quantile function (the inverse cumulative distribution function). The Lorenz function can be given by the following formula (see e.g.: Cowell, 2000):

$$L_Y(p) = \frac{\int_{-\infty}^{Q_y(p)} y \, dF_Y(y)}{\int_{-\infty}^{\infty} y \, dF_Y(y)}. \tag{3}$$

The above-mentioned formula can be applied when the theoretical probability distribution of Y is known and can be estimated from the data. In practice, we usually obtain the Lorenz curve directly using the finite population form of $L_Y(p)$ which is given as:

$$\widehat{L_Y}(p) = \frac{\sum_{i=1}^{N} y_i I\{y_i \leq Q_y(p)\}}{\sum_{i=1}^{N} y_i} \tag{4}$$

with $I\{.\}$ as an indicator function being equal to 1 if "." is true and 0 otherwise.

The Lorenz function of the variable Y refers to cumulative outcome proportions of population members ranked by the values of the same variable Y. Using another ranking variable X, while still measuring the outcome in terms of Y, leads to the so-called concentration curve (see e.g.: Cowell, 2000) which is often wrongly called the Lorenz curve.

$$L_{XY}(p)=\frac{\int_{-\infty}^{Q_x(p)}\int_{-\infty}^{\infty} yf_{XY}(xy)\,dy\,dx}{\int_{-\infty}^{\infty} y\,dF_Y(y)}, \tag{5}$$

where: $f_{XY}(x, y)$ is the density of the joint distribution of X and Y (see e.g.: Bishop, Chow, Formby, 1994). For the finite population of size N, formula (4) can be simplified to:

$$\hat{L}_{XY}(p)=\frac{\sum_{i=1}^{N} y_i I\{x_i \le Q_X(p)\}}{\sum_{i=1}^{N} y_i}. \tag{6}$$

In the credit scoring context, the empirical concentration curve given by formula (6) is applied with the scores S playing the role of the ranking variable X, while the variable of interest Y is binary (bad or good client). The graphical presentation of (6) takes the form of a plot with the empirical cumulative distribution function (ECDF) of bad clients $F_{Bad}(s)$ on the horizontal axis and the empirical cumulative distribution function of good clients $F_{Good}(s)$ on the vertical axis (see e.g.: Rezac, Kolacek, 2012). It is used to present the discrimination power of a given scoring model at any score value (Figure 1).

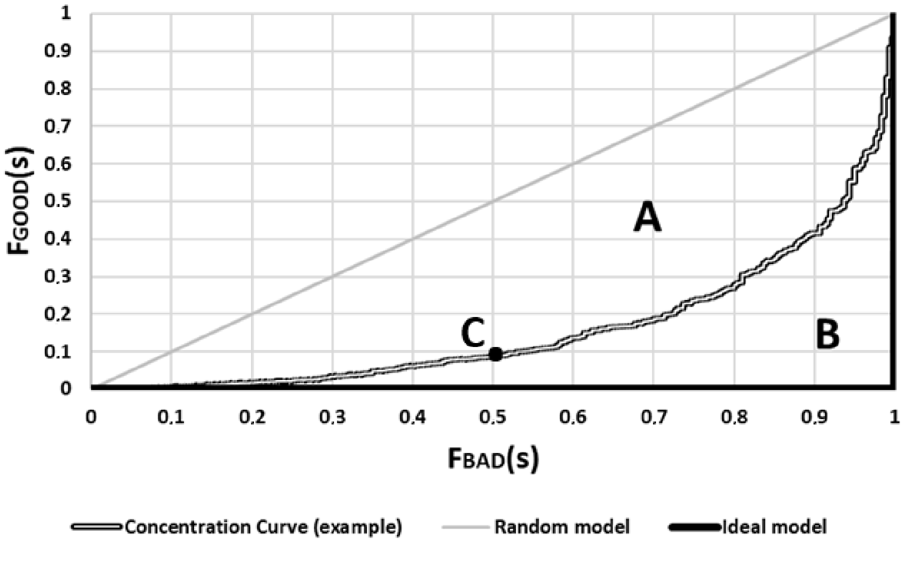

Figure 1. Concentration curve

Source: own elaboration

By means of the Lorenz and related curves, one can analyse the performance of scoring models at any value of the score. The diagonal line shows the performance of a random model (the model which randomly assigns score to good and bad clients); on the other hand, an ideal model assigns higher score values only to good clients (perfect separation between distributions of good and bad clients). In practice, the lower or upper values of the score are often investigated where a threshold (a particular value of the score below which all clients are classified as bad clients) is expected, e.g.: if 50% of bad clients is rejected, also roughly 10% of good clients is rejected at a given score value (see: point C in Figure 1).

The same curve, but with a reversed axis, called the Receiver Operating Characteristic (ROC), one can find in Anderson (2007), Finlay (2010), Hosmer, Lemeshow, Sturdivant (2013), Siddiqi (2017).

Cumulative Accuracy Profile

Another graphical way to assess quality of a scoring model is the Cumulative Accuracy Profile (CAP). This figure contains the cumulative distribution function of all clients on the horizontal axis and the cumulative distribution function of bad clients on the vertical axis. The CAP curve easily shows repercussions of rejecting the proportion of bad clients in terms of rejecting all clients (at any score value). An example of Cumulative Accuracy Profile is presented in Figure 2.

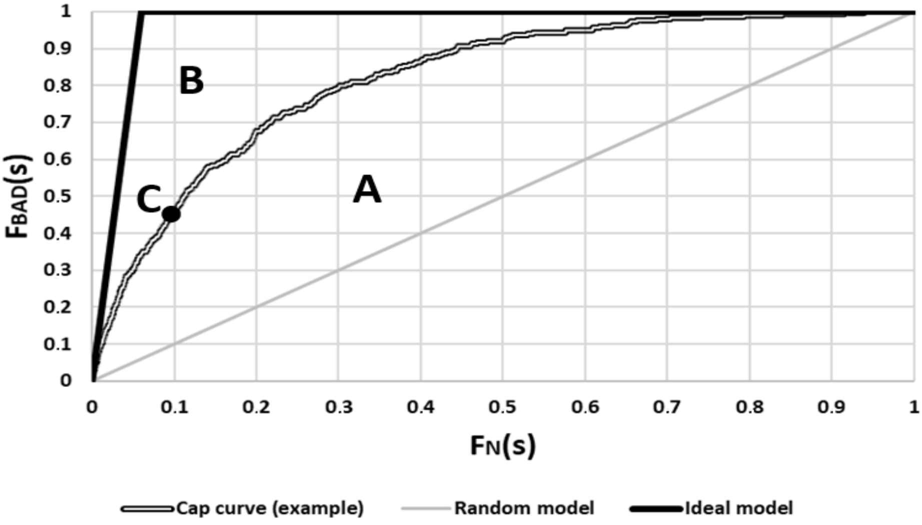

Figure 2. CAP curve

Source: own elaboration

A random model means that the model randomly assigns score to good and bad clients. The curve for an ideal model goes from point (0,0) through point (w_B,1) to point (1,1), where w_B is the fraction of all bad clients. The closer the CAP curve (for a given model) is to those for the ideal model, the better the scoring model is. Considering point C (0.1,0.46) as the threshold value, we reject 10% population and also get rid of 46% of bad clients.

Fish-eye graph

The fish-eye graph (also called the D_n curve) is a convenient method for investigating quality of a scoring model. It consists of plotting the empirical cumulative distribution function (ECDF) for both good and bad clients with respect to the score value. By means of the Fish-eye graph, one can analyse disproportions between fractions of good and bad clients. The greater the disproportions, the better the scoring model is. In fact, this method is connected with D_n statistic, which is defined as the maximum difference between empirical cumulative distributions (see: section 4).

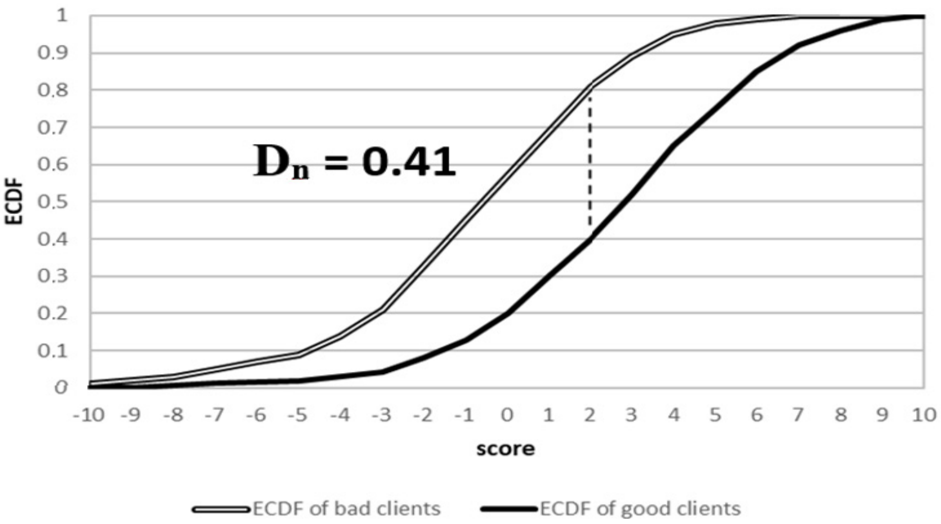

Figure 3. Fish-eye graph

Source: own elaboration

For the score that holds the maximum absolute value we have got the greatest disproportion between the fraction of bad and good clients. For example, at the score equal to 2 or smaller, there is a subset of population which consists of 81% of the empirical distribution of bad clients and 40% of the empirical distribution of good clients (Figure 3).

4. Numerical characteristics for assessing quality of scoring models

Gini index and related measures

The popular Gini index of inequality (Gini, 1912; 1914) was first proposed in 1912 but it became known after the publication from 1914 indicating the relation with the Lorenz curve. The Gini index can be described by several mathematical representations – each of them can be given its own interpretation and naturally leads to different estimator formulas. Among these formulas, the most popular is the geometric approach based on the Lorenz function (3) where the Gini index is defined as double the area between this function and the diagonal called the *line of equal shares*, as described in Figure 1:

$$G = \frac{A}{A+B} = 2A = 1 - 2B, \tag{7}$$

where: A is the area between the diagonal and Lorenz curve and B stands for the area under the Lorenz curve.

Another popular representation of the Gini index, proposed by Gini in 1912, is based on the absolute mean difference Δ, known as the *Gini mean difference* (GMD). This measure is a result of dividing the value of the absolute mean difference by the doubled expected value of Y:

$$G = \frac{\Delta}{2\mu}, \tag{8}$$

where $\Delta = E\,|Y_i - Y_j|$ is the expected value of the differences between the random variables Y_i and Y_j which come from the same distribution and represents variability of Y. The formula (8) enables the interpretation of the Gini index in terms of relative variability so it represents the so-called statistical approach.

When the form of the theoretical distribution of the random variable Y is known, we can utilise formulas (7) and (8) to determine the parametric estimates of the Gini index, as is usually the case in income studies. When we want to evaluate the Gini index directly from the data, we can apply several finite population representations of the Gini index (Jędrzejczak, 2010). In particular, formula (8) takes the form:

$$\hat{G} = \frac{\frac{1}{N^2}\sum_{i=1}^{N}\sum_{j=1}^{N}\left|y_i - y_j\right|}{2\mu}. \tag{9}$$

One can equivalently apply the following formula based on cumulative distributions functions:

$$\hat{G} = \frac{2}{\hat{\mu}} \int_{0}^{\infty} y\hat{F}(y)\, d\hat{F}(y) - 1, \tag{10}$$

where $\widehat{F}(y)$ is the empirical cumulative distribution function and $\hat{\mu}$ is the mean for the empirical values of the random variable *Y*. It is worth noting that formula (9) can lead to numerous problems when dealing with large datasets, especially in credit scoring where there are at least thousands of observations, while the results obtained by means of formula (10) can be ambiguous and lead to different values, depending on whether ECDF(y) is left or right continuous. One solution for that is data aggregation, though it usually leads to underestimating the true value of Gini index (Gastwirth, 1972).

To handle that, one may transform geometric formula (7) for the empirical values of the random variable *Y*, incorporating the so-called trapezium rule:

$$\hat{G} = 1 - \sum_{i=2}^{N} \left[\left(F_i - F_{i-1} \right) \times \left(L_i + L_{i-1} \right) \right], \tag{11}$$

where:

F_i is the cumulative distribution function of *Y*,

L_i is the value of the Lorenz curve for the *i*-th observation.

For purposes of gauging the discriminatory power of a scoring model, one can apply a modification of the Gini inequality index called the *concentration index* or *pseudo Gini index* (also called the *Gini*, *Gini statistic*, *Gini index*, *Gini coefficient*, see: Siddiqi, 2017; Finlay, 2010). It is based, contrary to the classical Gini index, on the concentration curve given by (4) and (5). It can be calculated from the data using the modification of formula (10) called the *Brown formula* (Finlay, 2010):

$$\tilde{G} = 1 - \sum_{i=2}^{N} \left[\left(F_{Bad_i}(s) - F_{Bad_{i-1}}(s) \right) \times \left(F_{Good_i}(s) + F_{Good_{i-1}}(s) \right) \right], \tag{12}$$

where:

$F_{Bad_i}(s)$ – is the empirical cumulative distribution function of bad clients' scores for the *i*-th observation,

$F_{Good_i}(s)$ – is the empirical cumulative distribution function of good clients' scores for the *i*-th observation.

The pseudo Gini index is very widely used to evaluate the discriminatory power of a scoring model. The classical Gini index measures the degree of inequality and takes values from <0; 1>, whereas the pseudo Gini index takes values from <–1; 1> and measures the concentration of good and bad clients, moreover, it gauges the direction of the relationship between scores and the dependent variable. Positive values mean that there is a positive relationship between the score and the dependent variable (the higher the score, the better the client), negative values mean that there is a negative relationship (the lower the score, the better the client), and value 0 means the model randomly assigns the score to the predicted variable. Absolute value 1 means an ideal model (the case when the distributions of good and bad clients are perfectly separated). The pseudo Gini index is connected with *c* statistics by the following relation:

$$c = \frac{1 + Gini}{2}, \tag{13}$$

where (13) is treated as the probability that a randomly selected observation from the distribution of bad clients has the score lower than a randomly selected observation marked as a good one:

$$c = P(s_1 > s_2 \mid Y_1 = 0 \wedge Y_2 = 1). \tag{14}$$

The minimum value of *c* equals 0.5, which means that the model randomly assigns scores to clients, on the other hand, the maximum value 1 means a perfect separation of good and bad observations.

In fact, the pseudo Gini index defined by formula (12) is a special case of Somers' D statistics with a discrete variable (Newson, 2006; Thomas, 2009). Moreover, there is a measure called the Accuracy Rate which is always equal to the value of the pseudo Gini for any scoring model.

The Accuracy Rate (*AR*) is a measure based on the CAP curve and it is calculated as:

$$AR = \frac{A}{A + B} = \frac{A}{0.5 \times (1 - w_B)}, \tag{15}$$

where:
A is the area between the CAP curve and the diagonal,
B is the area between the ideal model's CAP and the diagonal,
w_B is the fraction of all bad clients in the population of all clients.

Kolmogorov-Smirnov statistic

A measure of separation frequently used in the USA is the well-known Kolmogorov-Smirnov (D_n) statistic (Kolmogorov, 1933; Smirnov, 1936), which was originally proposed as a consistency test (see: e.g.: Domański, 1979). The Kolmogorov-Smirnov test is based on the comparison of empirical and theoretical cumulative distribution functions and verifies the hypothesis if a sample comes from a population with a specific (continuous) distribution, i.e.:

$$\begin{aligned} H_0 &: F(x) = F^*(x) \\ H_1 &: F(x) \neq F^*(x) \end{aligned} \tag{16}$$

where: $F(x)$ is the empirical distribution function based on the sample, $F^*(x)$ is the theoretical cumulative distribution function with known parameters. The D_n test statistic is defined as (Domański, 1979):

$$D_n = max\left|F^*(x) - F(x)\right|, \tag{17}$$

which is the maximum absolute difference between the empirical cumulative distribution function estimated on a random sample and the theoretical cumulative distribution function.

The D_n test applied in credit scoring is a statistic defined as the maximum difference between the cumulative distribution function of bad clients and the cumulative distribution of good clients:

$$D_n = max\left|F_{Bad}(s) - F_{Good}(s)\right|. \tag{18}$$

Firstly, the main disadvantage of D_n statistic is that it often chooses the score value that is too high or too low (usually D_n is obtained somewhere in the middle of the score range) for the scorecard threshold. Secondly, D_n statistic only tells us the maximum disproportion between the fractions of good and bad clients at some score value, hence it quite poorly describes quality of a scoring model as a whole. Thus, it is important to analyse the fish-eye graph and D_n statistic together and use them in conjunction with other measures as well.

Divergence

Divergence is a simple measure of separation of two groups. The measure can be easily obtained as the squared difference between the average scores of good and bad clients divided by their average variance (Siddiqi, 2017):

$$D^2 = \frac{\left(\pi_{Good} - \pi_{Bad}\right)^2}{\left(\sigma^2_{Good} + \sigma^2_{Good}\right)/2}. \tag{19}$$

Divergence is a parametric statistic which assumes that scores are normally distributed – this is an important limitation in the context of applicability because in practice the distribution of scores can often differ from the normal distribution.

Lift

Lift is a useful measure to assess the predictive power of a scorecard in each score interval. One can gauge how a model performs in a chosen range of the score – in a particular range where a threshold value is expected. Lift is defined as the ratio of the cumulative distribution function of bad clients and the cumulative distribution function of all clients (Rezac, Kolacek, 2012):

$$Lift(a) = \frac{F_{Bad}(a)}{F_{All}(a)}. \tag{20}$$

The presented measure indicates the number of times that the considered scoring model is better than the random model in a range of the score [$s_{\min}$; a]. Intuitively the higher the value of Lift, the better the scoring model is. Value 1 corresponds to a random model.

Distribution similarity index

The distribution similarity index (SI) was first proposed by the Polish statistician Egon Vielrose (1960) and has been well-known in economic research since then, especially in the field of income distribution analysis. This technique can also be applied to the evaluation of credit scoring systems and is described by the following equation (Domański, 2001):

$$SI = \sum_{i=1}^{K} \min\left(w_{Bi}, w_{Gi}\right), \tag{21}$$

where:

$w_{Bi} = \frac{b_i}{b}$ is the fraction of bad observations in the i-th score interval in the total number of bad observations,

$w_{Gi} = \frac{g_i}{g}$ is the fraction of good observations in the i-th score interval in the total number of bad observations, min(.) is a function which returns the smallest value from two arguments. The SI takes values from 0 to 1, where 0 means that the distributions are disjoint (perfect situation) and 1 means that structures of considered distributions are the same (random assignment of scores to clients). Obviously, the smaller the value of SI, the better the scoring model is.

5. Case study

As the illustration of the behaviour of the methods mentioned above, we investigate the quality of three different scoring models which can be applied to the same group of clients. Compared scorecards were developed to distinguish clients who are likely to repay their debts and those who are not. Basic information about the scorecards[1] is given in Table 1.

Table 1. Models comparison – basic statistics

	Model I	Model II	Model III
Number of predictors*	6	6	7
Min. score	–9.22	–10.44	–9.78
Avg. score	7.90	7.45	7.77
Max. score	11.45	9.24	10.34

* All of the estimated parameters are significant at 5% level.

Source: own calculations

The performance of these models has been examined on the basis of a dataset which consists of 5000 credit clients (non-mortgage loans) and 300 of them were bad clients[2]. In Table 2, the pseudo Gini index, D_n statistic, Divergence and Similarity Index (SI) for each scoring model have been presented.

Table 2. Pseudo Gini, D_n, Divergence and SI statistics

	Model I	Model II	Model III
Pseudo Gini	74.53%	73.26%	68.20%
D_n	61.55%	60.38%	53.09%
Divergence	2.58	2.41	1.95
SI	0.39	0.40	0.48

Source: own calculations

All of the computed global statistics outlined in Table 2 show that Model I has the highest discriminatory power, but Model II is almost as good as Model I. On the other hand, Model III seems to have the lowest discriminatory power. Given these measures, one can say that Model I performs only slightly better than Model II, but all of these values are at an acceptable level. According to the pseudo Gini index (74.53% vs 73.26%), D_n statistic (61.55% vs 60.38%), Divergence (2.58 vs 2.41) and

[1] Due to privacy policy, the structures of the models are not provided. All of the scorecards are acceptable from the statistical point of view (i.e. assumptions, the significance of the estimated parameters).

[2] A client was marked as bad when the days past due and the amount past due exceeded 90 days and 120 EUR respectively.

Similarity Index (0.39 vs 0.40), both models have similar discriminatory power. Because the pseudo Gini and the other measures cannot recognise a model which is significantly better than its competitors, further examination is necessary.

By means of Lift (eq. 20), it is possible to investigate performance of the models in particular score intervals, obtained by dividing all clients into decile groups (see Table 3).

Table 3. Lift values

Decile	Obs*	Model I		Model II		Model III	
		Bad obs**	LIFT	Bad obs	LIFT	Bad obs	LIFT
1	500	125	4.17	180	6.00	134	4.47
2	500	92	3.62	48	3.80	69	3.38
3	500	44	2.90	19	2.74	35	2.64
4	500	23	2.37	15	2.18	22	2.17
5	500	13	1.98	7	1.79	16	1.84
6	500	2	1.66	11	1.56	9	1.58
7	500	1	1.43	10	1.38	9	1.40
8	500	0	1.25	5	1.23	3	1.24
9	500	0	1.11	1	1.10	1	1.10
10	500	0	1.00	4	1.00	2	1.00
ALL	5000	300		300		300	

*The number of clients in each decile.
** The number of clients marked as bad in each decile.

Source: own calculations

It turns out that values of Lift differ across given models, especially in the first decile group, in favour of Model II and Model III. Model I is roughly 4 times better than a random model, whereas Models II and III are 6 times and roughly 4.5 times better than random selection, respectively. In this case, the maximum value of Lift would be equal to 10, hence Model II performs much better than the remaining ones in lower score intervals (the first and second decile).

Analysing the CAP curve (Figure 4), one can say that Model I performs better in higher values of scores (better separates the best clients from good clients) and Model II performs better in lower values of scores (better separates the worst clients from bad clients), whereas Model III seems more balanced (separates the best clients from good clients and the worst clients from bad clients with similar discriminatory power).

Discriminatory power can also be visualised by plotting concentration curves for each model. Curves show concentration of bad and good observation across all possible score values (see Figure 5). On the basis of the concentration curves (Figure 5), it can be noted that Model II is much better than Model I and Model III in lower score values, where a threshold is usually expected. For example, setting the threshold at a particular score value, for Model I, 60% of bad clients is rejected and also 12.4% of good clients is rejected (see: point B in Figure 5), for Model II 60% of bad clients is rejected and also 6.8% of good clients is rejected (see: point A in Figure 5) and for Model III 60% of bad clients is rejected and also 13.6% (see: point C in Figure 5) of good clients is rejected.

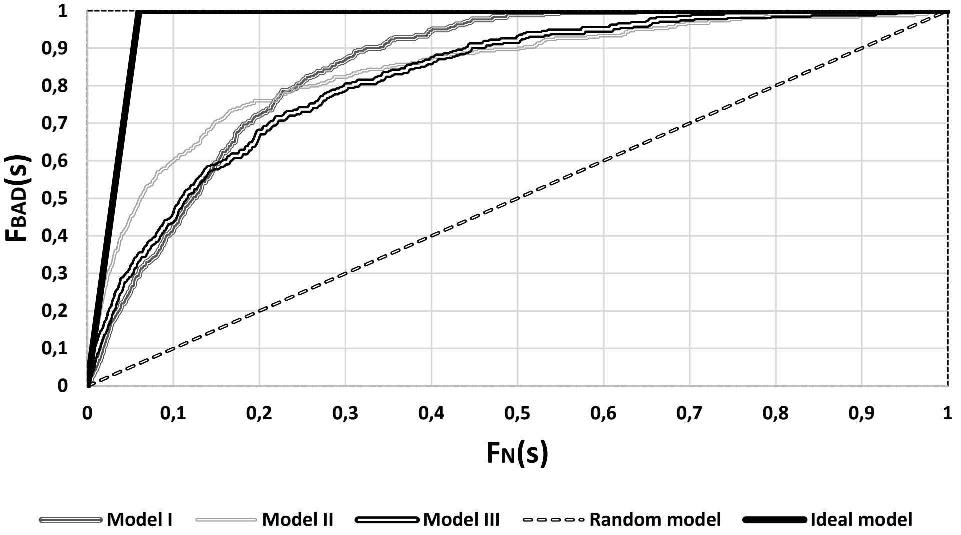

Figure 4. CAP curves for Model I, Model II and Model III

Source: own elaboration

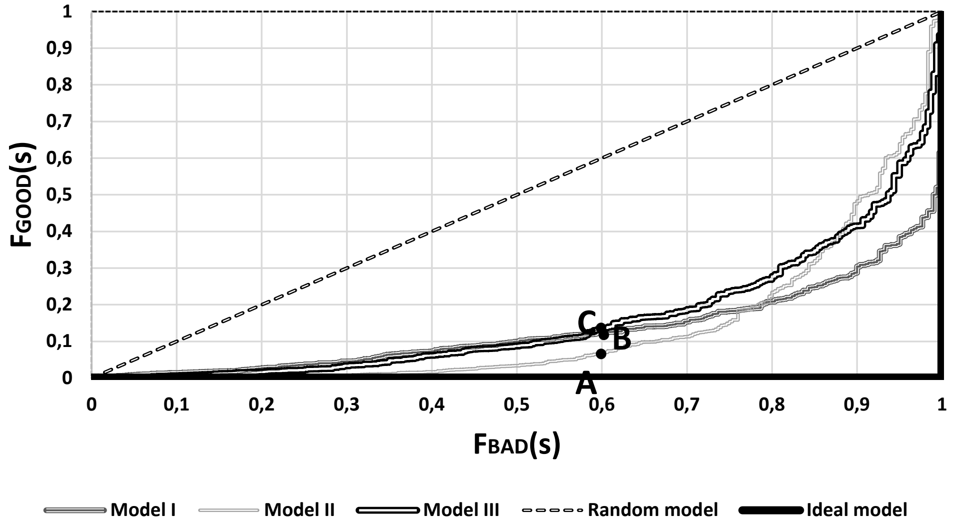

Figure 5. Concentration curves for Model I, Model II and Model III

Source: own elaboration

To sum up, all presented graphical methods and numerical characteristics for assessing quality of a scoring model lead to a particular choice which is Model II. Despite similar quality of the models based on global measures (taking into account all possible score values, see: Table 2), it has turned out that they perform quite differently in particular ranges of scores. In this case, the second model is the most reasonable for lenders in terms of quality due to its high discriminatory power in the lower range of the score where a threshold value is often expected to be set.

6. Conclusions

In a rapidly changing economic environment, rich in large volume data available at a low aggregation level, it becomes crucial for lenders to extract information about their customers in order to explore the market, make smart decisions and manage credit risk properly. These activities can be facilitated by the use of scoring models which produce *scores* based on consumer credit data. Scores are measures which rank clients with respect to their credit risk. The ability to gauge quality of a *scoring model* plays a key role in its developing or maintaining processes.

In this article, the most common methods (such as the pseudo Gini index*,* Kolmogorov-Smirnov statistic) for measuring quality of a *scoring model* were presented, simultaneously the author standardises the names of these methods (e.g.: the pseudo Gini index, concentration curve). It turns out that particular measures are named in various ways (often incorrectly), probably due to high contribution of practitioners in the development of credit scoring. Also, a case study was conducted which contained application of statistical measures for assessing quality of a scoring model in comparison analysis between three scoring models. It was shown that global measures should be analysed in conjunction with the local measure called *Lift* and graphs such as the concentration curve and Cumulative Accuracy Profile (CAP), especially when at first glance the models are not essentially different.

References

Abdou H., Pointon J. (2011), *Credit scoring, statistical techniques and evaluation criteria: a review of literature*, "Intelligent Systems in Accounting Finance & Management", vol. 18, no. 2–3, pp. 59–88.

Anderson R. (2007), *The credit scoring toolkit*, Oxford University Press, New York.

Bishop J.A., Chow K.V., Formby J.P. (1994), *Testing for Marginal Changes in Income Distributions with Lorenz and Concentration Curves*, "International Economic Review", vol. 35, no. 2, pp. 479–488.

Cowell F.A. (2000), *Measurement of Inequality*, [in:] A.B. Atkinson, F. Bourguignon (eds.), *Handbook of Income Distribution*, vol. 1, Elsevier, Amsterdam, pp. 87–166.

Crook J.N., Edelman D.B., Thomas L.C. (2007), *Recent developments in consumer credit risk assessment*, "European Journal of Operational Research", no. 183, pp. 1447–1465.

Domański C. (1979), *Statystyczne testy nieparametryczne*, Państwowe Wydawnictwo Ekonomiczne, Warszawa.

Domański C. (ed.) (2001), *Metody statystyczne. Teoria i zadania*, Wydawnictwo Uniwersytetu Łódzkiego, Łódź.

Finlay S. (2010), *Credit Scoring, Response Modelling and Insurance Rating: a practical guide to forecasting consumer behaviour*, Palgrave Macmillan, New York.

Gastwirth J. (1972), *The Estimation of the Lorenz Curve and Gini index*, "Review of Economics and Statistics", vol. 54, no. 3, pp. 306–316.

Gini C. (1912), *Variabilità e Mutuabilità. Contributoallo Studio delle Distribuzioni e delle Relazioni Statistiche,* C. Cuppini, Bologna.

Gini C. (1914), *Sulla misuradellaconcentrazione e dellavariabilitàdeicaratteri*, "Atti R. 1st. Veneto Sci. Lett. Arti", vol. LXXIII(II), pp. 1203–1248.

Hosmer D.W., Lemeshow S., Sturdivant R.X. (2013), *Applied Logistic Regression*, 3rd ed., John Wiley & Sons, New Jersey.

Jędrzejczak A. (2010), *Metody analizy rozkładu dochodów i ich koncentracji*, Wydawnictwo Uniwersytetu Łódzkiego, Łódź.

Kolmogorov A. (1933), *Sulla determinazioneempirica di unalegge di distribuzionc*, "Instituto Italiano degli Attuari", no. 4, pp. 1–11.

Lorenz M.O. (1905), *Methods of Measuring the Concentration of Wealth*, "Publications of the American Statistical Association", vol. 9, no. 70, pp. 209–219.

Newson R. (2006), *Confidence intervals for rank statistics: Somers' D and extensions*, "The Stata Journal", vol. 6, no. 3, pp. 309–334.

Rezac M., Kolacek J. (2012), *List-based quality indexes for credit scoring models as an alternative to Gini and KS*, "Journal of Statistics: Advances in Theory and Applications", vol. 7, no. 1, pp. 1–23.

Siddiqi N. (2017), *Intelligent credit scoring. Building and Implementing Better Credit Risk Scorecards*, 2nd ed., John Wiley & Sons, New Jersey.

Smirnov N.V. (1936), *Sur la distribution de w2 (criterium de M.R. von Mises)*, "Comptes rendus de l'Académie des Sciences", no. 202, pp. 449–452 [paper with the same title in Russian "Recueil Math" 1937, no. 2, pp. 973–993].

Thomas L.C. (2009), *Consumer Credit Models: Pricing, Profit, and Portfolio*, Oxford University Press, Oxford.

Vielrose E. (1960), *Rozkład dochodów według wielkości*, Polskie Wydawnictwo Gospodarcze, Warszawa.

Uwagi na temat statystycznych miar oceny jakości modelu scoringowego

Streszczenie: Jednym z podstawowych zadań banków jest udzielanie kredytów i pożyczek pieniężnych. Z punktu widzenia kredytodawcy w procesie kredytowaniem niezwykle istotna jest ocena ryzyka zaniechania płatności zobowiązań potencjalnego kredytobiorcy. W celu selekcji klientów, obok oceny ich zdolności kredytowej, coraz częściej wykorzystuje się modele scoringowe wchodzące w skład metodologii tzw. scoringu kredytowego (*creditscoring*). W podejściu tym z punktu widzenia kredytodawcy kluczowa jest jakość doboru jednostek, którym kredyt zostanie przyznany. To, czy klasyfikacja dokonywana na podstawie modelu scoringowego jest dobra, może być opisane za pomocą statystycznych miar oceny jakości.

Mimo coraz większej popularności metod scoringowych w praktyce gospodarczej literatura dotycząca statystycznych metod oceny ich jakości jest w dalszym ciągu stosunkowo uboga. Ponadto w publikacjach na ten temat często występują rozbieżności w zakresie nazewnictwa oraz konstrukcji poszczególnych miar. W artykule przedstawiono charakterystykę najczęściej stosowanych statystycznych miar oceny jakości modelu scoringowego (m.in. indeksu pseudo Giniego, statystyki Kolmogorova-Smirnova, krzywej koncentracji), a także podjęto próbę standaryzacji nazewnictwa oraz postaci samych miar jakości modelu scoringowego. Ponadto przedstawione zostało studium przypadku, w którym dokonano analizy porównawczej trzech modeli scoringowych w kontekście ich jakości klasyfikacyjnej.

Słowa kluczowe: scoring kredytowy, jakość modelu scoringowego, krzywa Lorenza, krzywa koncentracji, współczynnik Giniego

JEL: C52

Received: 2018-06-16; verified: 2019-02-15. Accepted: 2019-07-24