STATISTICS IN TRANSITION new series, June 2024
Vol. 25, No. 2, pp. 205–218, https://doi.org/10.59170/stattrans-2024-022
Received – 04.08.2023; accepted – 06.11.2023

The shape of an ROC curve in the evaluation of credit scoring models

Błażej Kochański[1]

Abstract

The AUC, i.e. the area under the receiver operating characteristic (ROC) curve, or its scaled version, the Gini coefficient, are the standard measures of the discriminatory power of credit scoring. Using binormal ROC curve models, we show how the shape of the curves affects the economic benefits of using scoring models with the same AUC. Based on the results, we propose that the shape parameter of the fitted ROC curve is reported alongside its AUC/Gini whenever the quality of a scorecard is discussed.

Key words: credit scoring, receiver operational characteristic, AUC.

1. Introduction

Credit scoring is probably the best known application of statistical methods in finance. Usually based primarily on a customer's credit history and application data, typically built using logistic regression, and less often using neural networks or other machine learning methods, a credit scoring model (a scorecard) returns a numerical score that allows the ranking of potential credit customers. Low scores typically indicate increased risk of default on the loan, while high scores imply lower credit risk.

In the past, the primary goal of a credit scorecard was to provide a single binary classification – to inform the decision maker when to accept or reject a loan. Today, the banking industry uses scorecards to rank customers for many other purposes. Credit institutions use scorecards to differentiate loan terms, such as interest rates (a practice referred to as "risk-based pricing"), loan amounts and periods. Credit limits are set and reset based on credit scores. Models drive cross-sell, up-sell and collection strategies. Credit scores feed into the capital adequacy and loss provisioning models.

The receiver operating characteristic (ROC) curve is a graphical plot that shows how well a binary classifier performs at various levels of the discrimination threshold. It is widely used in many domains, including signal processing, medical statistics, psychology, finance and machine learning applications in general. The ROC curve is

[1] Department of Statistics and Econometrics, Faculty of Management and Economics, Gdańsk University of Technology, Gdańsk, Poland. E-mail: blakocha@pg.edu.pl. ORCID: https://orcid.org/0000-0001-8502-931X.

plotted in the unit square with the false positive proportion (FPP) on the X-axis and the true positive proportion (TPP) on the Y-axis. It represents combinations of FPP and TPP for many possible thresholds (cut-off points). Credit scorecards are set up to identify potential non-payers ("defaults" or "bad" customers), therefore in this context, FPP is the share of non-defaulters ("goods") who scored below a given cut-off among all goods, and TPP is the share of bads who scored below the same given cut-off among all bads.

The AUC (area under the ROC curve) or its rescaled version, the Gini coefficient, are typical ways of assessing the discriminatory power of a credit scorecard. The AUC can be thought of as the probability of correctly assigning a worse score to a bad case, given one random good case and one random bad case (Hanley and McNeil, 1982). The AUC can vary from 0 to 1, where an uninformative scorecard has an AUC of 0.5, while a perfect scorecard has an AUC of 1. Credit scoring practitioners often prefer the Gini coefficient (equivalent to Somers' D statistic), which is a linearly scaled version of the AUC: uninformative scoring has a Gini of 0, while perfect scorecards have a Gini of 1.

Although the AUC is widely used by both practitioners and academics to assess the discriminatory power of credit scoring models, it has several serious drawbacks. Lobo, Jiménez-Valverde and Real (2008) summarise the drawbacks from a biostatistical perspective. From a credit scoring standpoint, the most important disadvantage of AUC can be described as "lift inconsistency". As Idczak (2019) and Řezáč and Koláček (2012) have shown, scorecards with the same AUC/Gini can have different efficiencies in reducing the default rate once a percentage of the worst applicants is rejected. This is related to Hand's (2009) broader claim that AUC is "fundamentally incoherent in terms of misclassification costs".

To address the shortcomings of the AUC, alternative measures are sometimes proposed. Řezáč and Koláček (2012) suggest using a summary of the lift curve at all possible cut-off points. Hand (2009) introduces an h-measure, which evaluates the model based on the distribution of likely values of the classification cost. Verbeke et al. (2012), Verbraken, Verbeke and Baesens (2013) and Garrido, Verbeke and Bravo (2018) introduce profit-based measures.

In this paper, we propose to take a different route. Often, if not always, practitioners will use the same credit scoring model for multiple purposes. Some objectives are better served by one scorecard, while other objectives are better served by another. Rather than deciding priorities for potential users or ranking different objectives, we suggest summarising the scoring model and the associated ROC curve by two numbers instead of one: (1) the AUC/Gini, which measures the area under the curve, and (2) a measure that summarises the shape of the curve. When discussing the quality of the model, two measures could be provided.

The remainder of the paper is organised as follows: in the next section, we present a motivating example of two intersecting ROC curves representing models that excel at different tasks. Then, a binormal ROC curve model is introduced, which produces

smooth, idealised ROC curves with only two parameters: AUC/Gini and the shape/symmetry parameter. The minimal distance fitting of binormal ROC curves to empirical ROC data is then described. This approach is subsequently demonstrated using two empirical examples from the literature. Finally, we discuss our conclusions.

2. When two ROC curves intersect – a motivating example

Several authors (e.g. Adams and Hand, 1999; Řezáč and Koláček, 2012) note that the problem with comparing the two AUCs becomes most apparent when the ROC curves of two scorecards intersect. Idczak (2019) and Tang and Chi (2015) provide empirical illustrations.

In Figure 1, we present a motivating (hypothetical) example of two scorecards (labelled I and II) with identical areas under their ROC curves (Gini = 0.7, AUC = 0.85) but different shapes. The intersecting ROC curves have been simulated using the binormal ROC formulas introduced in the next section.

We will now evaluate the scorecards based on how much the proportion of the customers who fail to repay their loans ("bad rate") in the loan portfolio will be lowered when the cut-off point is introduced and loan applications below cut-off are rejected. We assume here and in the examples below that the bad rate in the total scored population of loan applicants is 8%, which is a typical bad rate in the consumer finance market. However, analogous results can be obtained for population bad rates at other levels.

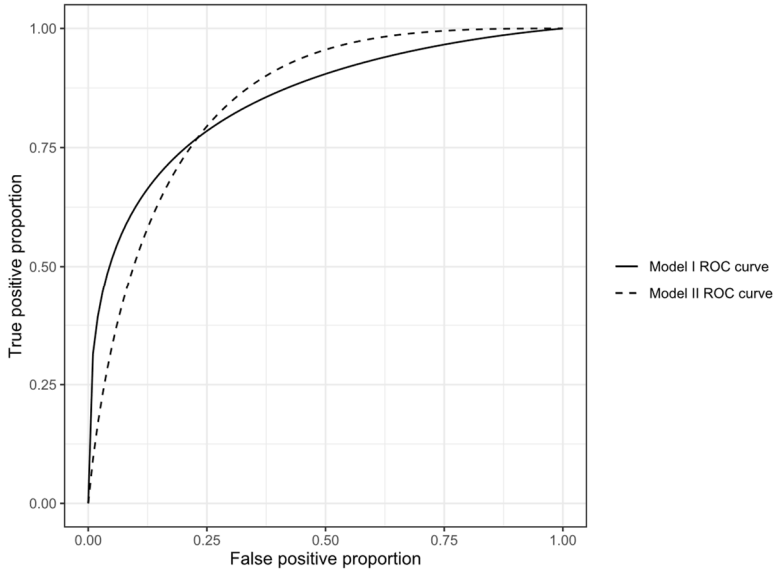

Figure 1: ROC curves for two scoring models with Gini=0.7. Model I is an example of a TPP-asymmetric ROC, and model II of a TNP-asymmetric ROC

If our strategy was to reject 10% of customers with the lowest scores, Model I would be preferred. It would have reduced the bad rate by 49.7% (after the introduction of the cut-off, and rejecting 10% of the applications, the share of defaulted customers would fall from 8% to 4.02%). Model II would have performed much worse in this setup, with a reduction of only 35.7% (from 8% to 5.15%).

The situation would be quite different if the portfolio manager were to select the top 10% of customers and offer them, for example, an increase in their credit limit. In such a case, Model II would prove to be much more effective. If selected using Model II, the top 10% of clients would have a default rate of only 0.04% (~200 times lower than the average). In comparison, the top 10% selected using Model I would have a default rate of 0.95% (only ~8 times lower than the average).

The difference between the curves is apparent on visual inspection. The ROC curve for Model I is more inclined to the OY axis, and the ROC curve for Model II is more inclined to the top line of the unit square (whose equation is y=1). The inclination indicates where the model performs best. Model I excels at identifying the worst of the bad cases (inclination to the OY axis), while Model II wins at picking the best of the good customers (inclination to the y=1 line). Outside of the field of credit scoring, such curves are referred to as asymmetric ROC curves (Killeen and Taylor, 2004, Bhattacharya and Hughes, 2015). Model I is an example of a TPP-asymmetric ROC and Model II of a TNP-asymmetric ROC, where TNP stands for "true negative proportion".

As this simple motivating example shows, two scorecards with the same AUC may be far from equivalent. Which of the models is most effective depends on the intended use in a particular case. Since credit scorecards have more than one specific application nowadays, they need to be assessed at many thresholds for various purposes. It would be virtually impossible to devise a single metric summarising their effectiveness. The h-measure – a popular alternative to AUC – is not a step forward in this case. The Monte Carlo simulation with R and the h-measure package (using the package's default assumptions about the distribution of possible cost values) showed that Model I has an h-measure of 0.325, while Model II has an h-measure of 0.193. The h-measure seems to indicate that Model II is inferior to Model I, which, as we have seen, is not necessarily true.

3. Modelling ROC curves with a binormal model

ROC curve models are mathematical formulae that allow idealised smooth ROC curves to be plotted that approximate real data. Examples of ROC curve models include binormal (Hanley, 1996), bilogistic (Walsh, 1997), bibeta (Chen and Hu, 2016) or bigamma (Dorfman et al., 1997) curves. Such models have 1-4 parameters that determine the course of the curve. ROC curve models are widely used in biostatistics

and signal processing, but their use in credit scoring research has so far been limited. Some prominent exceptions are Blöchlinger and Leippold (2006), Satchell and Xia (2008) and Kürüm et al. (2012), who applied ROC curve models to credit scoring problems. Kochański (2022) evaluates ROC curve models from a credit scoring perspective and examines how well they fit empirical data. He finds that the binormal model is the best fit for the largest number of empirical data sets and can be accommodated to have the Gini coefficient as a parameter. Following this result, we choose the binormal function from among the many ROC curve models. It can be expressed as the formula y=f(x) with two parameters: one representing the area under the ROC curve (or the Gini coefficient) and the other describing the shape (symmetry/inclination) of the curve.

The name "binormal" refers to the assumption that the scores of both good and bad customers are normally distributed (perhaps after a monotone transformation). The standard formula for the binormal ROC curve is as follows:

$$y = \Phi\left(a + b\Phi^{-1}(x)\right), \tag{1}$$

where x is a false positive proportion, y is a true positive proportion, Φ is the CDF of a standard normal variable, Φ^{-1} is its inverse, and a and b are parameters that can be interpreted as functions of the means and variances of the normally distributed scores of goods and bads (Bandos, Guo and Gur, 2017).

It can be shown that the AUC in this case is:

$$AUC = c = \Phi\left(\frac{a}{\sqrt{1+b^2}}\right), \tag{2}$$

which leads to an equivalent alternative parameterisation:

$$y = \Phi\left(\Phi^{-1}(c)\sqrt{1+b^2} + b\Phi^{-1}(x)\right) \tag{3}$$

or, using Gini g = $2c - 1$:

$$y = Bin_{b,g}(x) = \Phi\left(\Phi^{-1}\left(\frac{g+1}{2}\right)\sqrt{1+b^2} + b\Phi^{-1}(x)\right) \tag{4}$$

The ROC curve generated by the binormal model has a shape that depends on the parameter b. In the context of the binormal model construction, it can be thought of as the ratio between the standard deviations of the good and bad scores (which are assumed to follow normal distributions). Curves with b < 1 are TPP-asymmetric, curves with b > 1 exhibit TNP asymmetry. When b = 1, the ROC curve is symmetric about the negative diagonal of the unit square, i.e. the line connecting points (0,1) and (1,0).

Figure 2 illustrates how the parameter b affects the shape of the ROC curve in practice. The figure displays curves with two different levels of Gini and three

different values of the shape parameter *b*. The curves were generated using the binormal model as described by equation 4. Note that the curves in Figure 1 were also drawn using the binormal formula with parameters g = 0.7 (both models), b = 0.77 (Model I), b = 1.3 (Model II).

4. Shape of the ROC curves and efficiency of scoring models

As the motivating example showed, the reduction in the bad rate on accepted loans at the 10% cut-off appears to vary significantly depending on the shape of the ROC curve. Figure 3 extends this example by showing the reduction in the bad rate for a range of values of the shape parameter while maintaining Gini at g = 0.7. Since the empirical ROC curves have shapes corresponding to the range of 0.7 to 1.4 (Kochański, 2022), this graph shows this range. It appears that a 10% cut-off can reduce the bad rate on approved loans from 34% to 52%, depending on the asymmetry of the ROC curve.

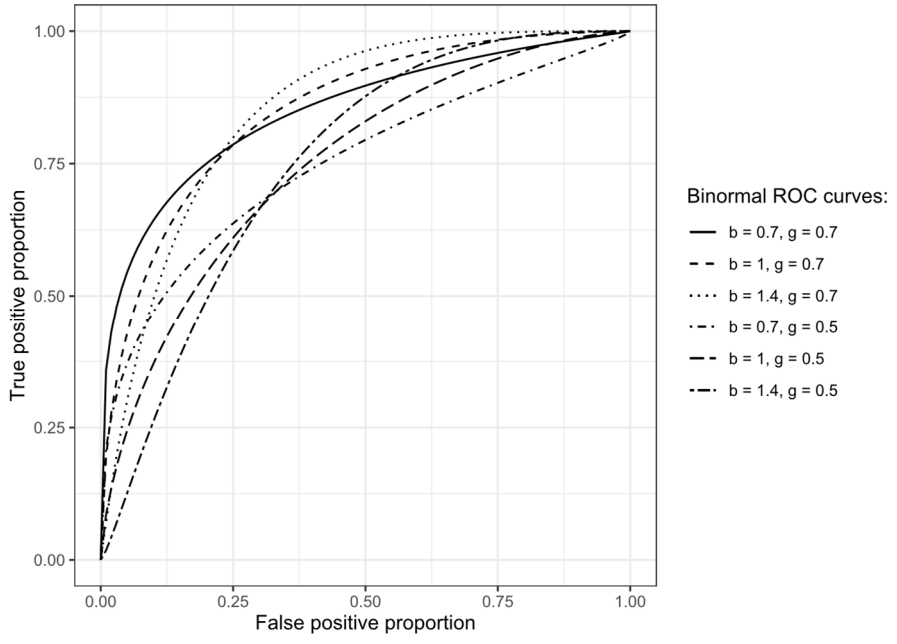

Figure 2: Binormal ROC curves with two Gini levels (0.5 and 0.7) and three values of b parameter

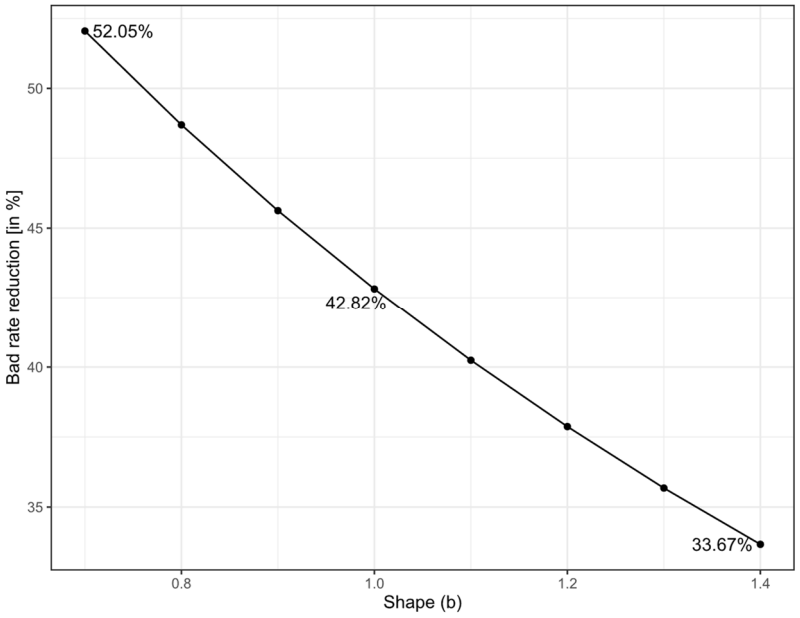

Figure 3: Bad rate reduction at 10% cut-off (90% approval rate) for a Gini=0.7 scorecard and various shapes of the ROC curve as modelled by the binormal model

Figure 4 illustrates this issue from a different perspective. We now fix the desired reduction in the bad rate (40% at the 10% cut-off) and see which Gini/shape combinations produce a reduction at this level. Again, the binormal model was used to answer this question. It turns out that, depending on the shape of the curve, such a reduction is possible with a Gini ranging from 0.56 to 0.75. Three such ROC curves are shown in Figure 5. As can be seen, they all intersect at one point, which corresponds to the cut-off point.

5. Minimal distance fitting of the shape parameter

The simulations presented show that information about the shape of the ROC curve can be crucial for credit decision makers and users of credit scoring. Therefore, we suggest that in business and research practice, in addition to reporting the Gini/AUC value, analysts should report a measure that describes the shape of the curve.

We suggest that ROC curve models could be used to describe the shape of the curve. To apply such a model, one must fit the empirical ROC data to a theoretical curve and find the parameters that give the best fit.

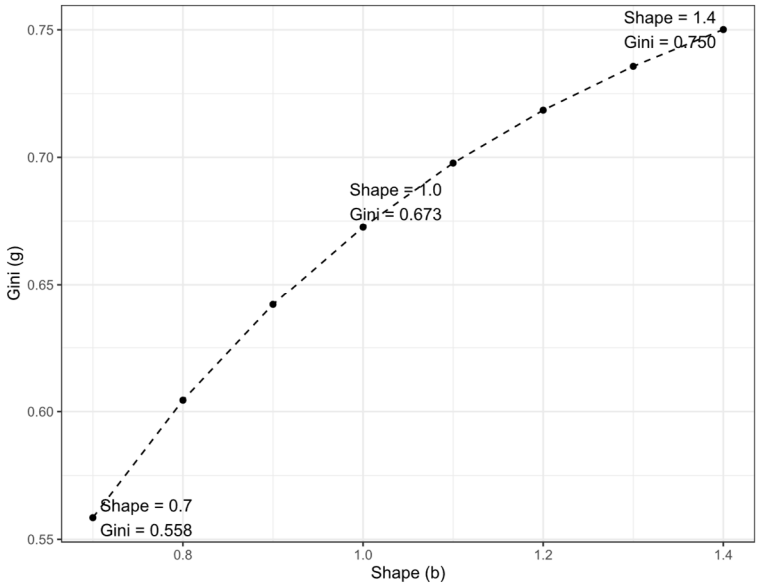

Figure 4: Shape (b) and Gini (g) combinations resulting in a bad rate reduction from 8.0% to 4.8% modelled with binormal ROC curves

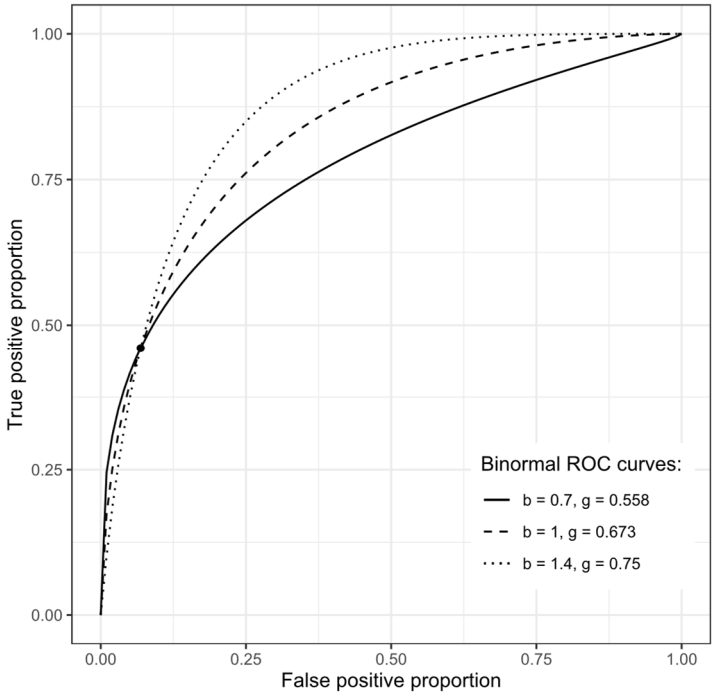

Figure 5: Three binormal ROC curves that reduce the bad rate by 40% when the cut-off point is set to reject 10% of loan applicants

The literature suggests many fitting methods, but for the problem at hand, minimal distance fitting, as described by Hsieh and Turnbull (1996), Davidov and Nov (2012), and Jokiel-Rokita and Topolnicki (2019), seems to be the best approach. For the purposes of this paper, we use the following procedure to fit an ROC curve where the two parameters are the Gini/AUC and the shape parameter:

(1) Empirical ROC data points are collected representing the TPP and FPP for a sufficient set of cut-off points on the ROC graph,

(2) The empirical ROC curve function is derived as a polygon (piecewise linear function) connecting the empirical ROC data points,

(3) The Gini coefficient (or AUC) is calculated using the usual trapezoid method,

(4) The shape parameter of the best fitting ROC curve is found after fixing the AUC.

Step (4) is performed using an appropriately adapted "minimum distance estimation" method. The aim is to minimise the L_2 distance between the empirical ROC curve y=ROC(x) and a theoretical ROC curve function $y = Bin_{b,g}(x)$, when g is already given, and the only value sought is the shape parameter b that gives the best-fitting curve.

Using the notation of Jokiel-Rokita and Topolnicki (2019), the minimised objective function is:

$$\|\xi(b)\| = \int_0^1 \xi^2(b,t)dt$$

where

$$\xi^2(b,t) = \left(\mathrm{ROC}(t) - Bin_{b,g}(t)\right)^2$$

6. Describing the shape of an ROC curve – examples

Rezáč and Rezáč (2011) and Hahm and Lee (2011) are rare examples of papers that provide data on empirical credit-scoring ROC curves in a form of a table. We use these examples to demonstrate the usefulness of describing the shape of ROC curves in assessing the quality of scoring models.

Figure 6 shows data points from Rezáč and Rezáč (2011) with polygon interpolation (dotted line) and fitted binormal function (Gini = 0.451, shape = 1.045). The shape parameter b is close to 1, which means that the curve is almost perfectly symmetric; it is similarly effective in both the high and low score ranges – a result that is usually desirable for a credit scorecard used for risk-based pricing.

The fit is quite good. The square root of $\|\xi(b)\|$ is 0.009, which means that the average (or, more precisely, the root mean square) vertical difference between the empirical and fitted ROC curves is less than one percentage point.

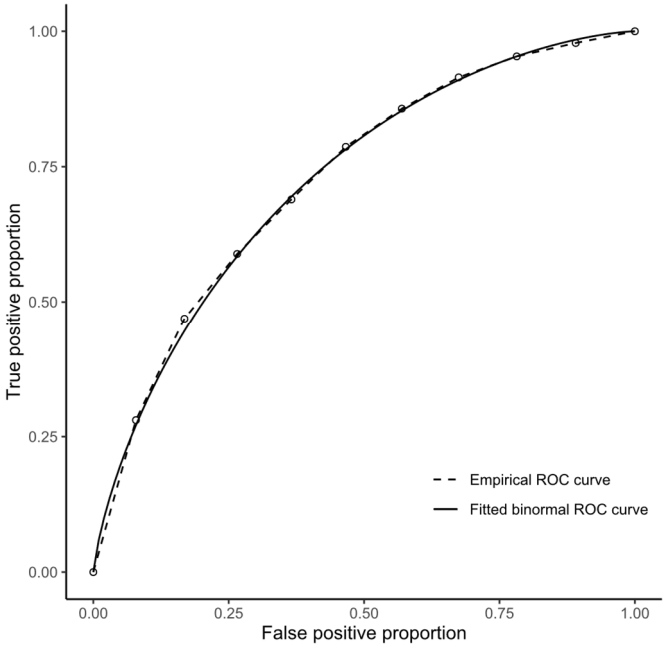

Figure 6: Rezáč and Rezáč (2011) data points with polygon interpolation (dotted line) and fitted binormal function ($g = 0.451$, $b = 1.045$, $\sqrt{\|\xi(b)\|}$=0.009)

The empirical and fitted ROC curves for models from Hahm and Lee (2011), who discuss scorecards developed using credit bureau data in a Korean bank, are shown in Figure 7. The fit of model A is quite good, but visibly worse than the previous case; the fit of model B is slightly better. Model A is a model that uses only negative credit bureau data (only default and late repayment data from other banks), while Model B is a model that also includes positive data (data on timely loan repayments). The inclusion of positive data increases the discriminatory power of the scorecard: AUC increases from 0.841 to 0.869 (Gini from 0.682 to 0.739). The shape parameter of the curve indicates that Model B is less asymmetric than Model A ($b = 0.785$ for Model A and 0.935 for Model B). The change in the shape of the curve implies that the inclusion of positive credit bureau data significantly improves the scorecard's ability to identify the best customers. Indeed, calculations show that both models reduce the default rate by about half when the 8% worst customers are rejected. If the rejection rate is set at 60%, banks can achieve a fivefold reduction using model A and about a tenfold reduction using model B. The apparent advantage of positive credit bureau data is that it allows the best customers to be identified much more easily, while negative data only filters out the worst customers.

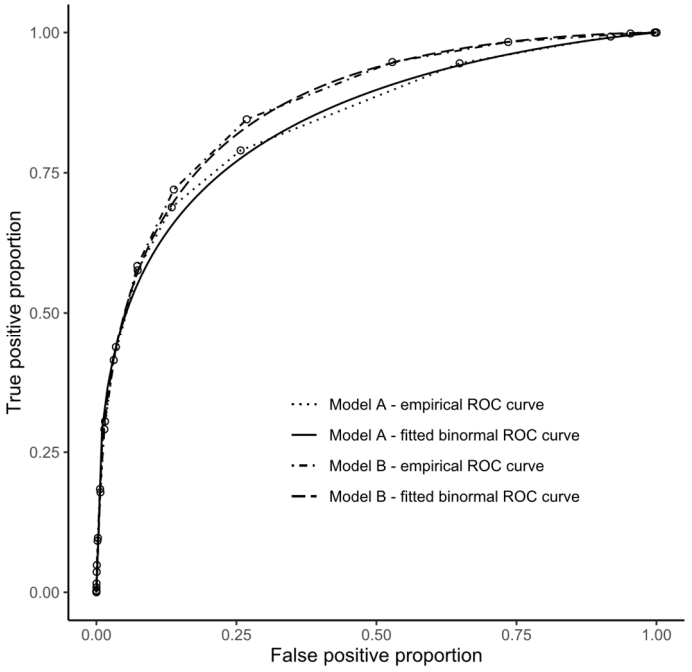

Figure 7: Hahm and Lee (2011) data points for models A and B with polygon interpolation and fitted binormal curves (model A: $g = 0.682$, $b = 0.785$, $\sqrt{\|\xi(b)\|} = 0.014$, model B: $g = 0.739$, $b = 0.935$, $\sqrt{\|\xi(b)\|} = 0.007$).

7. Conclusions

As it has been shown in the examples presented above, scorecards with the same AUC have different effectiveness in different score ranges. In this paper, we propose that when discussing the quality of a scorecard, we should not only provide the AUC/Gini, but also a figure that summarises the shape of the curve.

This approach can be useful when discussing the quality of existing or newly created scoring models. Knowing the shape of the curve allows one to assess which tasks the model is best suited for. Measuring the shape of the ROC curve can also be useful when evaluating individual component variables or component scorecards included in the combined master model. Reporting the shape of the ROC curve allows analysts to assess whether an individual variable or component scorecard owes its separation strength to its ability to find the best or worst customers, or – if the curve is symmetrical – it is equally efficient at both tasks.

In the current practice of lending businesses, the use of scoring models is multifaceted – they are used not only to identify the worst customers, but also to set the prices and terms, demand collateral, drive customer relationship and collection actions,

support loan provisioning procedures and capital adequacy calculations. Using only one number (AUC/Gini) or focusing only on one cut-off is simplistic and may not be sufficient to assess the separation power of a model. However, descriptive measures are needed to summarise the strength of the scorecard. Presenting two numbers instead of one seems to be a useful compromise between synthesis and attention to detail.

References

Adams, N. M., Hand, D. J., (1999). Comparing classifiers when the misallocation costs are uncertain. *Pattern Recognition*, Vol. 32(7), pp. 1139–1147.

Bandos, A. I., Guo, B., Gur, D., (2017). Estimating the area under ROC curve when the fitted Binormal Curves Demonstrate Improper Shape. *Academic Radiology*, Vol. 24(2), pp. 209–219.

Blöchlinger, A., Leippold, M., (2006). Economic benefit of powerful credit scoring. *Journal of Banking & Finance*, 30(3), pp. 851–873.

Chen, W., Hu, N., (2016). Proper bibeta ROC model: algorithm, software, and performance evaluation, in: *Medical Imaging 2016: Image Perception, Observer Performance, and Technology Assessment. Medical Imaging 2016: Image Perception, Observer Performance, and Technology Assessment*, SPIE, pp. 97–104.

Davidov, O., Nov, Y., (2012). Improving an estimator of Hsieh and Turnbull for the binormal ROC curve. *Journal of Statistical Planning and Inference*, Vol. 142(4), pp. 872–877.

Dorfman, D. D., Berbaum, K. S., Metz, C. E., Lenth, R. V., Hanley, J. A., Dagga, H. A., (1997). Proper receiver operating characteristic analysis: The bigamma model. *Academic Radiology*, 4(2), pp. 138–149.

England, W. L., (1988). An exponential model used for optimal threshold selection on ROC curves. *Medical Decision Making*, Vol. 8(2), pp. 120–131.

Garrido, F., Verbeke, W., Bravo, C., (2018). A Robust profit measure for binary classification model evaluation. *Expert Systems with Applications*, Vol. 92, pp. 154–160.

Hahm, J.-H., Lee, S., (2011). Economic effects of positive credit information sharing: the case of Korea. *Applied Economics*, Vol. 43(30), pp. 4879–4890.

Hand, D. J., (2009). Measuring classifier performance: A coherent alternative to the area under the ROC curve. *Machine Learning*, Vol. 77(1), pp. 103–123.

Hand, D. J., Anagnostopoulos, C., (2013). When is the area under the receiver operating characteristic curve an appropriate measure of classifier performance? *Pattern Recognition Letters*, Vol. 34(5), pp. 492–495.

Hanley, J. A., (1996). The Use of the "Binormal" Model for Parametric ROC Analysis of Quantitative Diagnostic Tests. *Statistics in Medicine*, 15(14), pp. 1575–1585.

Hanley, J. A., McNeil, B. J., (1982). The meaning and use of the area under a receiver operating characteristic (ROC) curve. *Radiology*, Vol. 143(1), pp. 29–36.

Hsieh, F., Turnbull, B. W., (1996). Nonparametric and semiparametric estimation of the receiver operating characteristic curve. *The Annals of Statistics*, Vol. 24(1), pp. 25–40.

Hughes, G., Bhattacharya, B., (2013). Symmetry Properties of Bi-Normal and Bi-Gamma Receiver Operating Characteristic Curves are Described by Kullback-Leibler Divergences. *Entropy*, Vol. 15(4), pp. 1342–1356.

Idczak, A. P., (2019). Remarks on statistical measures for assessing quality of scoring models. *Acta Universitatis Lodziensis. Folia Oeconomica*, Vol. 4(343), pp. 21–38.

Jokiel-Rokita, A., Topolnicki, R., (2019). Minimum distance estimation of the binormal ROC curve. *Statistical Papers*, Vol. 60(6), pp. 2161–2183.

Killeen, P. R., Taylor, T. J., (2004). Symmetric receiver operating characteristics. *Journal of Mathematical Psychology*, Vol. 48(6), pp. 432–434.

Kochański, B., (2022). Which curve fits best: fitting ROC curve models to empirical credit-scoring data. *Risks*, Vol. 10(10), p. 184.

Kürüm, E., Yildirak, K., Weber, G.-W., (2012). A classification problem of credit risk rating investigated and solved by optimisation of the ROC curve. *Central European Journal of Operations Research*, 20(3), pp. 529–557.

Lobo, J. M., Jiménez-Valverde, A., Real, R., (2008). AUC: a misleading measure of the performance of predictive distribution models. *Global Ecology and Biogeography*, Vol. 17(2), pp. 145–151.

Řezáč, M., Koláček, J., (2012). Lift-based quality indexes for credit scoring models as an alternative to Gini and KS. *Journal of Statistics: Advances in Theory and Applications*, Vol. 7(1), pp. 1–23.

Řezáč, M., Řezáč, F., (2011). How to Measure the Quality of Credit Scoring Models. *Czech Journal of Economics and Finance (Finance a úvěr)*, Vol. 61(5), pp. 486–507.

Satchell, S. Xia, W., (2008). Analytic models of the ROC Curve: Applications to credit rating model validation, in G. Christodoulakis and S. Satchell (eds). *The Analytics of Risk Model Validation*. Burlington: Academic Press (Quantitative Finance), pp. 113–133.

Tang, T.-C., Chi, L.-C., (2005). Predicting multilateral trade credit risks: comparisons of Logit and Fuzzy Logic models using ROC curve analysis. *Expert Systems with Applications*, Vol. 28(3), pp. 547–556.

Verbeke, W., Dejaeger, K., Martens, D., Hur, J., Baesens, B., (2012). New insights into churn prediction in the telecommunication sector: A profit driven data mining approach, Vol. 218(1), pp. 211–229.

Verbraken, T., Verbeke, W., Baesens, B., (2013). A Novel Profit Maximizing Metric for Measuring Classification Performance of Customer Churn Prediction Models. *IEEE Transactions on Knowledge and Data Engineering*, Vol. 25(5), pp. 961–973.

Walsh, S. J., (1997). Limitations to the Robustness of Binormal Roc Curves: Effects of Model Misspecification and Location of Decision Thresholds on Bias, Precision, Size and Power. *Statistics in Medicine*, 16(6), pp. 669–679.