Economics Letters 145 (2016) 48–51

Contents lists available at ScienceDirect

Economics Letters

journal homepage: www.elsevier.com/locate/ecolet

Comparing the accuracy of default predictions in the rating industry for different sets of obligors☆

Walter Krämer *, Simon Neumärker [1]

Fakultät Statistik, Technische Universität Dortmund, Germany

HIGHLIGHTS

- We show how probability forecasts can be ranked for different sets of events.
- This ranking generalizes the refinement ordering which is only applicable to identical sets of events.
- This ranking provides a partial ordering which is consistent with popular skill scores used in practice.

ARTICLE INFO

Article history:
Received 11 February 2016
Received in revised form
20 May 2016
Accepted 21 May 2016
Available online 27 May 2016

JEL classification:
C4
G2

Keywords:
Moody's
S&P
Probability forecasts
Skill scores

ABSTRACT

We generalize the refinement ordering for well calibrated probability forecasts to the case were the debtors under consideration are not necessarily identical. This ordering is consistent with many well known skill scores used in practice. We also add an illustration using default predictions made by the leading rating agencies Moody's and S&P.

1. Introduction

Probability forecasting has a long tradition in many fields of application. In economics, the most popular ones are default predictions in the rating industry. According to the Basel-II and Basel-III accords for instance, banks have to attach predicted default probabilities to all outstanding loans. Although major rating agencies like Moody's or S&P are reluctant to identify their letter grades with predicted default probabilities, we will stick to this probability interpretation in what follows. Given two competing default predictors and the prevalence of split ratings in practice (see e.g. Hauck and Neyer, 2014), it is then natural to ask: Which one is better?

One option is to rely on some scalar measures of performance like the Brier Score. However, it is well known that different score functions might produce conflicting results (see e.g. Krämer and Güttler, 2008 for an example). The present paper therefore is concerned with partial orderings which, if valid, will imply identical rankings with respect to all members from some suitable class of scoring functions. It extends (Krämer, 2006), which covers only identical sets of debtors, to cases where the two debtors under considerations are not necessarily identical. It is not concerned with the equally important issue of how ratings are produced in the first place (see Lahiri and Yang, 2013 for an overview or Czarnitzki and Kraft, 2004 or Boumparis et al., 2015 for relevant discussions in the present journal).

Section 2 below introduces a novel partial ordering based on Generalized Lorenz curves and Section 3 provides an application to ten-year default predictions made by the leading rating agencies Moody's and S&P.

☆ Research supported by DFG-Sonderforschungsbereich 823. We are grateful by Peter Posch for providing the data for our example and to an unknown referee for most useful criticism and comments.

* Corresponding author. Tel.: +49 231 755 3125; fax: +49 231 755 5284.
E-mail addresses: walterk@statistik.tu-dortmund.de (W. Krämer), simon.neumaerker@tu-dortmund.de (S. Neumärker).

[1] Tel.: +49 231 755 3869; fax: +49 231 755 5284.

http://dx.doi.org/10.1016/j.econlet.2016.05.021
0165-1765/

2. Modified Lorenz dominance

Let $0 = a_1 < a_2 < \cdots < a_k = 1$ be a finite set of possible forecasts of default probabilities. Let $q^A(a_j)$ be the relative frequency with which the default probability a_j is predicted by forecaster A (similarly for B). This paper will only consider forecasts which are well calibrated, i.e. where

$$\mathbb{P}(\text{default } |a_j) = a_j \quad (j = 1, \ldots, k). \tag{1}$$

In addition, we will focus on theoretical distributions, i.e. we will not distinguish between relative default frequencies and default probabilities. Everything that follows will then depend only on the vectors $a := [a_1, \ldots, a_k]'$ and $q := [q(a_1), \ldots, q(a_k)]'$.

For the special case where A and B are rating the same set of debtors, DeGroot and Fienberg (1983) suggest the concept of refinement to discriminate between the two. If, by applying a randomization to the probability forecasts of A, one obtains a new probability forecast with the same distribution as B, then A is more refined than B. As shown by DeGroot and Eriksson (1985), this amounts to Lorenz-domination of the respective forecast distributions:

$$A \geq_L B \Leftrightarrow \underbrace{\frac{1}{p}\int_0^x F^{A^{-1}}(t)dt}_{=L^A(x)} \leq \underbrace{\frac{1}{p}\int_0^x F^{B^{-1}}(t)dt}_{=L^B(x)}, \quad (0 \leq x \leq 1) \tag{2}$$

where $L^A(x)$ and $L^B(x)$ are the respective Lorenz curves,

$$F^A(a) := \sum_{a_i \leq a} q^A(a_i) \tag{3}$$

is A's default forecast distribution and where

$$F^{A^{-1}}(t) := \inf\{a : F^A(a) \geq t\} \tag{4}$$

is the inverse of A's default forecast distribution (similarly for B). The overall default probability can then be expressed as

$$p = \int_0^1 F^{A^{-1}}(t)dt = \int_0^1 F^{B^{-1}}(t)dt \tag{5}$$

which equals the expectation of both F^A and F^B. In view of calibration, $p = \sum a_i q^A(a_i) = \sum a_i q^B(a_i)$. This expectation could as well be dropped in Eq. (2), as it appears on both sides of the inequality, and mainly sees to it that both Lorenz curves end in (1, 1).

Contrary to comparing income inequality, where Lorenz curves close to the diagonal are "good" (i.e. signal a more equal distribution of income), A is in the present application considered better than B if its Lorenz curve bends farther away from the diagonal, i.e. if its predicted default probabilities are more spread out. This is why we here, other than in the income distribution context, identify "domination" with a higher level of inequality. It can also easily be shown that the same ordering obtains if the ranking is based on predicted non-defaults:

$$\int_0^x F^{A^{-1}}(t)dt \leq \int_0^x F^{B^{-1}}(t)dt$$
$$\Leftrightarrow \quad \int_0^x \tilde{F}^{A^{-1}}(t)dt \leq \int_0^x \tilde{F}^{B^{-1}}(t)dt \tag{6}$$

for $0 \leq x \leq 1$, where $\tilde{F}(a) := \sum_{\tilde{a}_i \leq a} \tilde{q}(\tilde{a}_i)$ is the distribution function of the predicted survival probabilities $\tilde{a}_i := 1 - a_i$ and $\tilde{q}(\tilde{a}_i) := q(a_i)$.

If A and B are rating different (possibly overlapping) sets of debtors, the overall probability of default will in general differ between the respective sets, and the refinement concept does no longer apply. However, the Lorenz-ordering is still possible, by replacing the overall default probability $p = p_A = p_B$ in (2) with p_A and p_B, where appropriate. Other than in the case $p_A = p_B$, it now does matter whether we consider predicted default or predicted survival probabilities: It can be shown by simple counterexamples that A's Lorenz curve for predicted default probabilities is better and A's Lorenz curve for predicted survival probabilities is worse than that of B. Therefore the standard Lorenz order does not make much sense for nonidentical sets of debtors. Here is an extension:

Definition. A dominates B in the modified Lorenz sense ($A \geq_{ML} B$) if $A \geq_L B$ (i.e. (2) obtains with p_A and p_B in place of p) and in addition,

$$0.5 \geq p_A \geq p_B \quad (p_B < 0.5) \quad \text{or}$$
$$0.5 \leq p_A \leq p_B \quad (p_B > 0.5).$$

For $p_A = p_B$, this reduces to the standard refinement ordering. Without loss of generality, we will confine ourselves to the empirically more relevant case $p_B < 0.5$ in what follows. The inequality $p_A > p_B$ then implies that the generalized Lorenz curve (defined as p times standard Lorenz curve) of A is larger than that of B towards the right end of the [0, 1]-interval. Intuitively, this means that A's predictions are both more spread out and on average closer to 0.5 at the same time.

It is well known from the theory of proper scoring rules (see e.g. Winkler, 1996) that it becomes harder to obtain good results as the overall default probability approaches 0.5. The well known Brier score for instance, given by

$$B(a, q) := \sum_{i=1}^{k} q(a_i)a_i(1 - a_i) \tag{7}$$

whenever a forecaster is well calibrated, approaches its optimal value of 0 even for the trivial forecast $a_i = p\,\forall i$ whenever $p \to 0$ or $p \to 1$. And the trivial forecast is worst in the Brier sense if $p = 0.5$ (always assuming that p is among the available a_i's). Two additional scoring rules often used in application are the logarithmic score

$$L(a, q) := \sum_{i=1}^{k} q(a_i)\left(a_i \ln(a_i) + (1 - a_i)\ln(1 - a_i)\right)$$
$$(\text{with } 0\ln(0) := 0) \tag{8}$$

and the spherical score

$$S(a, q) := \sum_{i=1}^{k} q(a_i)\sqrt{a_i^2 + (1 - a_i)^2}, \tag{9}$$

which are likewise producing good results for the trivial forecasts as $p \to 0$ or $p \to 1$.

In order to compensate for this intrinsic difference in difficulty, it is common to rely on skill scores rather than on ordinary scoring rules whenever $p_A \neq p_B$ (see Lahiri and Yang, 2013 for additional motivation). Given any scoring rule $S(a, q)$, the corresponding skill score is given by

$$SS(a, q) := \frac{S(a, q) - S_t}{S_{opt} - S_t} \tag{10}$$

where S_t is the trivial score obtained for $a_i = p\ \forall i$ and S_{opt} is the optimal score where only $q(0)$ and/or $q(1)$ are different from zero (Winkler, 1996). A skill score then measures how close a forecaster is to the optimum. It takes its maximum value of 1 if defaults and non-defaults are both predicted with certainty; it takes the value zero for the trivial forecast, and it can even take on values less than zero if a forecaster is worse than the trivial forecast. For the Brier score, for instance, we have

$$BS(a, q) = \frac{B(a, q) - p(1 - p)}{-p(1 - p)}. \tag{11}$$