Knowledge-Based Systems 000 (2017) 1–15

Contents lists available at ScienceDirect

Knowledge-Based Systems

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

Approaches for credit scorecard calibration: An empirical analysis

Artem Bequé[a,*], Kristof Coussement[b], Ross Gayler[c], Stefan Lessmann[a]

[a] School of Business and Economics, Humboldt-University of Berlin, Unter-den-Linden 6, Berlin 10099, Germany
[b] IESEG School of Management, Université Catholique de Lille, (LEM, UMR CNRS 9221), Department of Marketing, 3 Rue de la Digue, Lille F-59000, France
[c] Independent researcher, Melbourne, Australia

ARTICLE INFO

Article history:
Received 22 February 2017
Revised 23 July 2017
Accepted 25 July 2017
Available online xxx

Keywords:
Credit scoring
Classification
Calibration
Probability of default

ABSTRACT

Financial institutions use credit scorecards for risk management. A scorecard is a data-driven model for predicting default probabilities. Scorecard assessment concentrates on how well a scorecard discriminates good and bad risk. Whether predicted and observed default probabilities agree (i.e., calibration) is an equally important yet often overlooked dimension of scorecard performance. Surprisingly, no attempt has been made to systematically explore different calibration methods and their implications in credit scoring. The goal of the paper is to integrate previous work on probability calibration, to re-introduce available calibration techniques to the credit scoring community, and to empirically examine the extent to which they improve scorecards. More specifically, using real-world credit scoring data, we first develop scorecards using different classifiers, next apply calibration methods to the classifier predictions, and then measure the degree to which they improve calibration. To evaluate performance, we measure the accuracy of predictions in terms of the Brier Score before and after calibration, and employ repeated measures analysis of variance to test for significant differences between group means. Furthermore, we check calibration using reliability plots and decompose the Brier Score to clarify the origin of performance differences across calibrators. The observed results suggest that post-processing scorecard predictions using a calibrator is beneficial. Calibrators improve scorecard calibration while the discriminatory ability remains unaffected. Generalized additive models are particularly suitable for calibrating classifier predictions.

1. Introduction

Credit scoring helps to improve the efficiency of loan officers, reduce human bias in lending decisions, quantify expected losses, and, more generally, manage financial risks effectively and responsibly [13,21]. Today, almost all lenders rely upon scoring systems to assess financial risks [44]. In retail lending, for example, credit scoring is widely used to decide on applications for personal credit cards, consumer loans, and mortgages [26]. A lender employs data from past transactions to predict the chance of an applicant to default. To decide on the application, the lender then compares the predicted probability to default (PD) to a cut-off value; granting credit if the prediction is below the cut-off, and rejecting it otherwise [35].

Many techniques for scorecard development have been proposed and studied. Examples include artificial neural networks [2,51,55], support vector machines [16,36], multiple classifier systems [27], hybrid models [4,18], or genetic programming [1]. In general, any classification algorithm facilitates the construction of a scorecard and PD modeling in particular [15]. Logistic regression is the most widely used approach in industry [44], although other, more sophisticated classification algorithms have been shown to predict credit risks more accurately [6,45]. A fully-comprehensive review of 214 articles/books/theses on application credit scoring further supports [3] the view that more advanced techniques (e.g., genetic algorithms) outperform conventional models (e.g., logistic regression). However, the authors also report on studies that find similar performance in terms of predictive accuracy [3].

In addition to predictive accuracy, the suitability of a scorecard also depends on other dimensions such as comprehensibility and compliance [34] or defining key variables for classifiers by mitigating noise data and redundant attributes [69]. This paper, however, concentrates on one specific dimension of scorecard performance: *calibration*.

A well-calibrated scorecard is one which produces probabilistic forecasts that correspond with observed probabilities [20]. For example, consider one hundred loans in a band of predicted PD

* Corresponding author.
E-mail addresses: artemlive@live.com (A. Bequé), k.coussement@ieseg.fr (K. Coussement), r.gayler@gmail.com (R. Gayler), stefan.lessmann@hu-berlin.de (S. Lessmann).

http://dx.doi.org/10.1016/j.knosys.2017.07.034
0950-7051/

estimated to be ten percent by some scorecard. If the scorecard is well-calibrated, the actual number of eventually defaulting loans in this band should be close to ten.

Scorecard calibration is important for many reasons. Regulatory frameworks such as the Basel Accord require financial institutions to verify that their internal rating systems produce calibrated risk predictions. Poor calibration, therefore, is penalized with higher regulatory capital requirements [22]. Calibration is also relevant from a lending decision making point of view [19]. At a micro-level, well-calibrated risk predictions are essential to evaluate credit applications in economic terms (e.g., through calculating expected gains/losses), which is more relevant to the business than an evaluation in terms of statistical accuracy measures only [12,33]. At a macro-level, calibration is important for portfolio risk management and default rate estimation [64]. In particular, to forecast the default rate of a credit portfolio, one may adopt a classify-and-count strategy [10]. This approach derives the portfolio default rate forecast from individual level (single loan) risk predictions and thus benefits from calibration [65]. Furthermore, approaches to support managerial decisions have to account for the cognitive abilities and limitations of decision makers [50]. Although far from perfect, probabilities (rather than, say, log-odds) are a format to represent information that decision makers understand and process relatively well [43]. Thus, a credit analyst is likely to distil more information from a well-calibrated PD estimate. Last, scorecards are developed from loans granted in the past and used to forecast the risk of lending to novel applicants [37]. Due to changes in customer behavior, economic conditions, etc. default rates may differ across the corresponding distributions. Calibration is a way to account for the differences in prior probabilities [20].

The Institute of International Finance, Inc. and the International Swaps and Derivatives Association have called for a higher recognition of calibration when choosing among scorecards [40]. However, we find multiple studies that concentrate on e.g., balancing between accuracy and complexity [76], improving existing classifiers [69], offering new multiple classifier systems [6], but rarely studies devoted to calibration. In [3] the authors conclude that the receiver operating characteristic curve and the Gini coefficient are the most popular performance evaluation criteria in credit scoring. That is why we argue that the relevance of calibration is still not sufficiently reflected in the credit scoring literature. To further support this point, we consider a recent review of more than forty empirical credit scoring studies published between 2003 and 2014 [49]. Among the articles reviewed in [49], we find only one study [46] that explicitly raises the issue of calibration and uses suitable evaluation metrics such as Brier Score. More recent literature published after 2014 shows the same pattern. We find only two studies that use Brier Score to measure classifier performance [5,6]. However, both studies concentrate on developing novel classification systems, which are assessed in terms of the Brier Score, amongst others. Neither [5] nor [6] consider techniques to improve calibration, which supports the view that calibration methods have not been examined sufficiently in credit scoring; or, in the words of Van Hoorde et al. [68]: *calibration is often overlooked in risk modeling*.

There is ample evidence that especially advanced learning algorithms such as random forest, which enjoy much popularity in credit scoring, produce predictions that are poorly calibrated [47,54,60,74]. This suggests a trade-off between predictive accuracy and calibration. Calibration assumes that the relationship between the raw score, which a classification model produces, and the true PD is monotonic. Therefore, calibration consists of estimating a monotonic function to map raw scores to (calibrated) PDs. Given that the calibration function is monotonic, it maintains the ordering of the cases by raw score and consequently has no effect on the discriminative power of classifiers [71]. Examples of calibration techniques include isotonic regression or Platt scaling [54]. They promise to overcome the accuracy-calibration-trade-off and seem to have potential for credit scoring. To the best of our knowledge, no attempt has been made to systematically explore this potential in prior work in credit scoring.

The goal of this paper is to close this research gap. More specifically, we aim at examining the degree to which alternative algorithms for scorecard development suffer from poor calibration, evaluating techniques for improving calibration, and, thereby, contributing towards increasing the fit of advanced classifiers for real-world banking requirements. In pursuing these objectives, we make the following contributions. First, we establish the difference between accuracy and calibration measures. This helps to understand the conceptual differences between the two and to emphasize the need to address calibration in scorecard development. Second, we introduce several methods to improve calibration, subsequently called calibrators, to the credit scoring community and systematically assess their performance through empirical experimentation. Third, we examine the interaction between classifiers and calibrators. This allows us to identify synergies between the modeling approaches and to provide specific recommendations which techniques work well together. Last, relying upon reliability analysis and a decomposition of the Brier Score, we shed light on the determinants of calibrator effectiveness and provide insight into why and when calibrators work well.

The remainder of the paper is organized as follows: Section 2 introduces relevant methodology and the calibrators in particular. Section 3 describes the experimental design before empirical results are presented in Section 4. Section 5 concludes the paper.

2. Calibration methods

A classifier or a scorecard estimates a functional relationship between the probability distribution of a binary class label - good or bad risk - and a set of explanatory variables, which profile the applicant's characteristics and behavior. For example, bad risks are commonly defined as customers who miss three consecutive payments [66]. Calibration serves two purposes. First, some classification algorithms are *unable to produce probabilistic predictions*. For instance, support vector machines output a confidence score on the real interval $[-\infty; +\infty]$, whereby the sign of the prediction indicates the class assignment and the magnitude the confidence of the classifier in this assignment. For example, a positive confidence score might indicate that the classifier considers a credit applicant a bad risk, whereby a large (small) confidence score indicates that the classifier is certain (uncertain) about this prediction [58]. Second, some classifiers provide predictions in the interval [0; 1], which can be interpreted as probabilities, but suffer from biases and thus display poor calibration. Examples include the random forest classifier, the predictions of which habitually exhibit a characteristic sigmoid-shaped distortion [54]. Therefore, we define calibration as the process of converting the confidence scores or the raw (uncalibrated) probabilistic predictions – hereafter referred to as the credit risk output scores – to calibrated credit risk probabilities.

To demonstrate the technique behind calibration, Table 1 presents a theoretical example of credit risk output scores of a classifier *before* and *after* the application of calibration. Table 1 also gives the actual class. Values of 1 and 0 indicate default and non-default events, respectively. Recall that this example is only valid for a classifier that generates probabilistic predictions, meaning that the output of the classifier must be in the interval [0; 1].

Table 1 illustrates that calibration improves the quality of the probabilistic predictions of the theoretical classifier in a sense that the uncalibrated predictions of the classifier get closer to the true

Please cite this article as: A. Bequé et al., Approaches for credit scorecard calibration: An empirical analysis, Knowledge-Based Systems (2017), http://dx.doi.org/10.1016/j.knosys.2017.07.034