European Journal of Operational Research 238 (2014) 505–513

Contents lists available at ScienceDirect

European Journal of Operational Research

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

Stochastics and Statistics

Development and application of consumer credit scoring models using profit-based classification measures

Thomas Verbraken [a], Cristián Bravo [b,*], Richard Weber [c], Bart Baesens [a,d]

[a] Dept. of Decision Sciences and Information Management, KU Leuven, Belgium
[b] Depto. de Modelamiento y Gestión Industrial, Universidad de Talca, Chile
[c] Dept. of Industrial Engineering, Universidad de Chile, Chile
[d] School of Management, University of Southampton, United Kingdom

ARTICLE INFO

Article history:
Received 26 March 2013
Accepted 3 April 2014
Available online 13 April 2014

Keywords:
Data analytics
Credit scoring
Classification
Performance measurement
Cutoff point

ABSTRACT

This paper presents a new approach for consumer credit scoring, by tailoring a profit-based classification performance measure to credit risk modeling. This performance measure takes into account the expected profits and losses of credit granting and thereby better aligns the model developers' objectives with those of the lending company. It is based on the Expected Maximum Profit (EMP) measure and is used to find a trade-off between the expected losses – driven by the exposure of the loan and the loss given default – and the operational income given by the loan. Additionally, one of the major advantages of using the proposed measure is that it permits to calculate the optimal cutoff value, which is necessary for model implementation. To test the proposed approach, we use a dataset of loans granted by a government institution, and benchmarked the accuracy and monetary gain of using EMP, accuracy, and the area under the ROC curve as measures for selecting model parameters, and for determining the respective cutoff values. The results show that our proposed profit-based classification measure outperforms the alternative approaches in terms of both accuracy and monetary value in the test set, and that it facilitates model deployment.

1. Introduction

Credit scoring is a very important application in statistical modeling, and concerns distinguishing *good* from *bad* loan applicants (Thomas, Crook, & Edelman, 2002). The main goal is to estimate the probability of default, i.e. the event of a customer not paying back a loan in a given period. For this task, a predictive model is developed which assigns a score to each loan applicant. Such a model is then put to practice, by defining a cutoff value. Each applicant with a score lower than this cutoff will be rejected, others will be granted a loan.

During the past decades, a myriad of classification techniques has been used for credit scoring (Baesens et al., 2003). Hence, performance measurement is essential for model selection, i.e. to identify the most suited classification technique as well as to tune the respective parameters (Ali & Smith, 2006). It has been shown that traditional performance measures such as the Gini coefficient, the KS statistic, and the AUC measure are inappropriate in many cases and may lead to incorrect conclusions (Hand, 2005, 2009), since they do not always properly take into account the business reality of credit scoring. Thus a guideline to select the most appropriate classification model as well as to calculate an adequate cutoff value is still missing if it comes to apply credit scoring in a profit-oriented setting, which has already been advocated by e.g. Thomas (2009) and Finlay (2010).

The main contribution of this paper is to establish an approach which tackles both requirements simultaneously. That is, we propose a profit-based classification performance measure, inspired by the EMP measure (Verbraken, Verbeke, & Baesens, 2013), that takes into account the business reality of credit scoring and allows to calculate the optimal cutoff value from a profitability perspective. In Section 2 of this paper we discuss the problem of classification and the respective performance measurement. Section 3 shows in detail how a profit-based performance measure can be implemented in a credit scoring context. Section 4 reports the experimental setup and the obtained results. Conclusions and future work are presented in Section 5.

* Corresponding author. Tel.: +56 75 220 1756.
E-mail addresses: thomas.verbraken@kuleuven.be (T. Verbraken), crbravo@utalca.cl (C. Bravo), rweber@dii.uchile.cl (R. Weber), bart.baesens@kuleuven.be (B. Baesens).

http://dx.doi.org/10.1016/j.ejor.2014.04.001
0377-2217/

2. Classification and its performance measurement

Classification is an important task in predictive modeling. A variety of performance measures has been proposed to assess classification models. Section 2.1 outlines the use of such models in a business context. Section 2.2 discusses statistically motivated classification performance measures.

2.1. Classification in a business context

We focus on binary classification and follow the convention that cases, i.e. the instances of interest such as e.g. the defaulters in credit scoring, belong to class 0, whereas the non-cases correspond to class 1. Note that in the literature several conventions have been adopted, such as class 1 for default cases (the opposite of this paper). In credit scoring, some authors assign the labels g (good) and b (bad) to non-defaulters and defaulters, respectively. The convention we opted for, however, offers the advantages that it simplifies notation and has also been adopted by Hand (2009), among others, which is relevant for this paper. The prior probabilities of class 0 and 1 are π_0 and π_1, respectively.

Typically, the output from a classification model serves as input for business decisions, such as e.g. accepting/rejecting a loan application in credit scoring. Generally, a classification model provides a continuous score, $s(\mathbf{x})$, which is a function of the attribute vector $\mathbf{x}$ of the respective instance. In this paper, it is assumed that the instances from class 0 have a lower score than those from class 1 (if not, for logistic regression models, simply multiply the beta coefficients by -1 before constructing the score).

The actual classification, i.e. the assignment of each instance to one of the two classes, is achieved by defining a cutoff value t, such that all instances with $s < t$ are classified as cases, whereas instances for which $s \geqslant t$ are classified as non-cases. Function $F_0(s)$ ($F_1(s)$) is the cumulative density function of the cases' (non-cases') scores s. Analogously, $f_0(s)$ ($f_1(s)$) is the probability density function of the cases' (non-cases') scores s; see Fig. 1. Cases for which $s < t$ (corresponding to the shaded area under $f_0(s)$) are correctly predicted. On the other hand, non-cases with $s < t$ (corresponding to the shaded area under $f_1(s)$) are incorrectly predicted.

The outcome of a classification model applied to N instances can be summarized in a confusion matrix, as displayed in Table 1, where the diagonal represents the correct predictions. The off-diagonal quadrants concern incorrect predictions, expressed as proportions. Varying the cutoff value t changes these proportions.

Each cell in the confusion matrix has related costs or benefits. In general, the cost or benefit $c(i|j)$ of classifying an instance from class j into class i (with $i, j \in \{0, 1\}$) can be different for each of the four cells. These costs and benefits should be measured against a base scenario, as mentioned by Verbraken et al. (2013). They propose taking as base scenario the situation where no classification occurs at all, and measuring costs and benefits in comparison to this scenario. In the case of credit scoring, the base scenario would be that all loans are granted. Obviously, this is not a realistic scenario, since every financial institution will have a credit scoring program in place. But comparing to the "*grant all loans*" base scenario, ensures consistency when evaluating different credit scoring models.

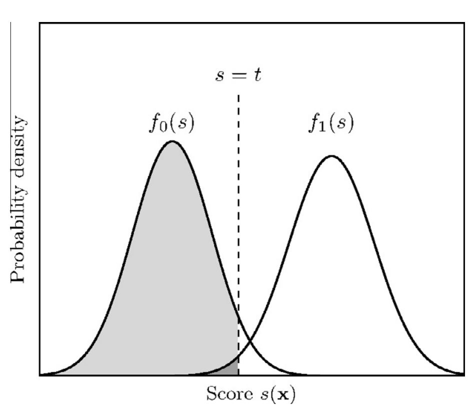

Fig. 1. Example of score distributions and the classification process.

Table 1
Confusion matrix with costs and benefits compared to base scenario.

True label	Predicted label	
	Case	Non-case
Case	$\pi_0 F_0(t)$ $[c(0 \mid 0) = b_0]$	$\pi_0(1 - F_0(t))$ $[c(1 \mid 0) = 0]$
Non-case	$\pi_1 F_1(t)$ $[c(0 \mid 1) = c_1]$	$\pi_1(1 - F_1(t))$ $[c(1 \mid 1) = 0]$
	↓ **Action** **@ cost** c^*	↓ **No** **Action**

By using a credit scoring system, the financial institution will be able to reject potentially harmful applications, hereby increasing its profit as compared to accepting all customers. Different models can thus be compared in terms of the extra profit they generate.

As a result, only costs and benefits corresponding to *predicted* cases (here: defaulters) are relevant (i.e. $c(1|0) = c(1|1) = 0$), since only predicted cases will experience an impact from the action undertaken – and hence will differ from the base scenario. For notational convenience, we define $b_0 := c(0|0)$ and $c_1 := c(0|1)$, where $b_0, c_1 \geqslant 0$ are a benefit and a cost, respectively. In general, the action undertaken by the company towards an individual case may come at a cost c^*. Finally, we should mention the fixed cost of building classification models, such as the cost of data collection, data preprocessing, model building, and model maintenance. However, these costs are irrelevant for model selection, as they will be approximately the same for all models.

2.2. Classification performance measurement

Several performance measures have been proposed to evaluate classification models; see e.g. Baldi, Brunak, Chauvin, Andersen, and Nielsen (2000). In the data mining community, the best-known measures include (Hand, 2009):

$$\text{Accuracy} = \pi_0 F_0(t) + \pi_1(1 - F_1(t)),$$
$$\text{Sensitivity} = F_0(t), \quad \text{Specificity} = 1 - F_1(t),$$
$$\text{AUC} = \int_{-\infty}^{\infty} F_0(s) f_1(s)\,\mathrm{d}s.$$

A classifier's accuracy measures the proportion of correctly classified observations. Sensitivity is the proportion of cases which are correctly classified, whereas specificity is the proportion of correctly predicted non-cases. The Area Under the receiver operating characteristic Curve (AUC) takes the entire range of possible cutoff values into account (Fawcett, 2006).

Most of these performance measures do not consider the misclassification costs, and are therefore only applicable when these costs are equal. Nevertheless, a lot of attention has been paid to cost-sensitive learning recently. Domingos (1999) proposed a general method to construct cost-sensitive classifiers, Provost and Fawcett (2001) combined ROC curve analysis with cost distribution information, Bernstein, Provost, and Hill (2005) developed an ontology-based approach for cost-sensitive classification, Zhou and Liu (2006) used over- and undersampling and threshold