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

Thomas Verbraken[1], Cristián Bravo[†2], Richard Weber[3], and Bart Baesens[1,4]

[1]Dept. of Decision Sciences and Information Management, KU Leuven, Belgium
[2]Depto. de Modelamiento y Gestión Industrial, Universidad de Talca, Chile
[3]Dept. of Industrial Engineering, Universidad de Chile, Chile
[4]School of Management, University of Southampton, United Kingdom

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

*NOTICE: this is the author's version of a work that was accepted for publication in the European Journal of Operational Research. Changes resulting from the publishing process, such as peer review, editing, corrections, structural formatting, and other quality control mechanisms may not be reflected in this document. Changes may have been made to this work since it was submitted for publication. Please cite this paper as follows: Verbraken, T., Bravo, C., Weber, R. and Baesens, B. (2014) Development and application of consumer credit scoring models using profit-based classification measures. European Journal of Operational Research. In Press. Available Online: `http://www.sciencedirect.com/science/article/pii/S0377221714003105`

[†]Email Addresses: thomas.verbraken@kuleuven.be, crbravo@utalca.cl (corresponding), rweber@dii.uchile.cl, bart.baesens@kuleuven.be

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.
Keywords: Data Analytics, Credit Scoring, Classification, Performance Measurement, Cutoff point.

1 Introduction

Credit scoring is a very important application in statistical modeling, and concerns distinguishing *good* from *bad* loan applicants (Thomas et al., 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 and 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 et al., 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.

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 $s \geq t$

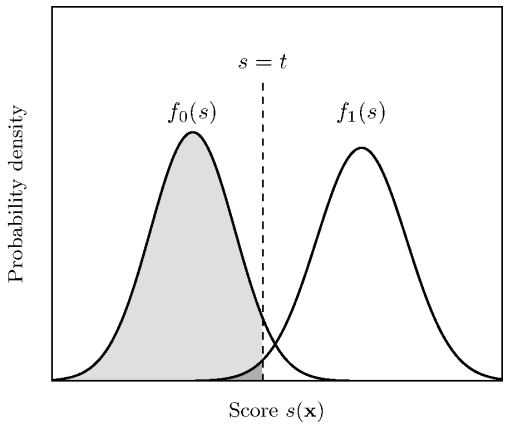


Figure 1: Example of score distributions and the classification process.

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 Figure 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.

By using a credit scoring system, the financial institution will be able to

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

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

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 $b_0, c_1 \geq 0$ 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 et al. (2000). In the data mining community, the

best-known measures include (Hand, 2009):

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

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 et al. (2005) developed an ontology-based approach for cost-sensitive classification, Zhou and Liu (2006) used over- and undersampling and threshold moving (and an ensemble of these methods) for cost-sensitive learning with neural networks, and Hand (2009) introduced the H-measure, which takes misclassification costs into account. However, this paper looks at the incremental profit generated by employing a classification model in a business context.

3 The Expected Maximum Profit measure for credit scoring

This section presents the application of the Expected Maximum Profit (EMP) measure, a general profit-based performance measure, to the particular case of credit scoring. Section 3.1 explains this general framework for classification performance.

Its application in a particular setting requires determining the respective cost and benefit parameters, which is discussed in Section 3.2. Section 3.3 presents how the EMP measure can be estimated empirically. Its relationship with AUC is analyzed in Section 3.4.

3.1 Profit-based performance measurement

The general framework starts by defining the **average classification profit per borrower**, generated by employing a classifier, which is calculated as follows:

$$P(t; b_0, c_1, c^*) = (b_0 - c^*)\pi_0 F_0(t) - (c_1 + c^*)\pi_1 F_1(t). \tag{1}$$

Optimizing the average profit which depends on the cutoff value t leads to the **maximum profit measure**, introduced by Verbeke et al. (2012):

$$\text{MP} = \max_{\forall t} P(t; b_0, c_1, c^*) = P(T; b_0, c_1, c^*), \tag{2}$$

with T the optimal cutoff value under the given circumstances:

$$T = \arg\max_{\forall t} P(t; b_0, c_1, c^*). \tag{3}$$

The optimal cutoff value T satisfies the first order condition:

$$\frac{f_0(T)}{f_1(T)} = \frac{\pi_1(c_1 + c^*)}{\pi_0(b_0 - c^*)} = \frac{\pi_1}{\pi_0}\theta, \tag{4}$$

where θ is the cost-benefit ratio, introduced for notational convenience:

$$\theta = \frac{c_1 + c^*}{b_0 - c^*}. \tag{5}$$

Note that the right-hand side of (4) only contains priors and cost and benefit parameters. The left-hand side is a ratio of the probability density functions evaluated at cutoff T and corresponds to a certain slope on the receiver operating characteristic (ROC) curve. Thus, varying θ from zero to infinity corresponds to a translation over the ROC curve (a more detailed derivation can be found in (Verbraken et al., 2013)). As argued by Verbeke et al. (2012), MP in itself could be used as a classification performance measure which allows to select the model with the highest incremental profit. Moreover, contrary to traditional performance measures, the optimal cutoff is clearly defined and the fraction of the customer base towards which the action should be undertaken is equal to:

$$\bar{\eta}_{\text{mp}} = \pi_0 F_0(T) + \pi_1 F_1(T). \tag{6}$$

The maximum profit measure was further refined by Verbraken et al. (2013), assuming that the cost and benefit parameters, c_1 and b_0, are not always exactly known but follow a probability distribution. This assumption generalizes the profit model shown in equation (1), and allows explicitly considering randomness in costs and benefits across the observed sample. The **expected maximum profit measure (EMP)** is defined as follows:

$$\text{EMP} = \int_{b_0} \int_{c_1} P(T(\theta); b_0, c_1, c^*) \cdot h(b_0, c_1) \mathrm{d}c_1 \mathrm{d}b_0, \tag{7}$$

with $h(b_0, c_1)$ the joint probability density of the classification costs.

It has been shown (Verbraken et al., 2013) that EMP corresponds to an integration over a range of the ROC curve, and it is an upper bound to the profit a company can achieve by applying the classifier. Analogously to the deterministic optimal fraction $\bar{\eta}_{\text{mp}}$, the expected profit maximizing fraction, $\bar{\eta}_{\text{emp}}$, is the fraction of cases towards which an action is undertaken:

$$\bar{\eta}_{\text{emp}} = \int_{b_0} \int_{c_1} \left[\pi_0 F_0\left(T(\theta)\right) + \pi_1 F_1(T\left(\theta\right))\right] \cdot h(b_0, c_1) \mathrm{d}c_1 \mathrm{d}b_0. \tag{8}$$

3.2 Cost and benefit parameters

To apply the general EMP framework to the case of credit scoring, the conditions to determine the optimal cutoff value (4) have to be adapted. This requires specifying the parameters b_0, c_1, and c^* as well as the probability distribution $h(b_0, c_1)$ in Equation (7). Next, we will use the methodology developed by Bravo et al. (2013) to calculate each of these parameters.

Parameter b_0 is the benefit of correctly identifying a defaulter, more precisely it is the fraction of the loan amount which is lost after default:

$$b_0 = \frac{\text{LGD} \cdot \text{EAD}}{A} = \lambda, \tag{9}$$

with $\lambda \in [0, 1]$ for notational convenience. A is the principal, LGD is the loss given default, and EAD is the exposure at default (Mays and Nuetzel, 2004).

Parameter c_1 is the cost of incorrectly classifying a good applicant as a defaulter and is equal to the return on investment (ROI) of the loan, which considers the cost of the funds and all operational costs, i.e. $c_1 = \text{ROI}$. For any loan with instalments p to pay, the borrower-requested maturity M and principal A can be used to estimate the return, considering the interest rate r

typically offered at that term and principal level. Under those assumptions, the ROI can be estimated using the well-known total interest (I) formulas (Broverman, 2010):

$$\begin{aligned} I &= pM - A \\ p &= \frac{Ar}{1-(1+r)^{-M}} \\ ROI &= \frac{I}{A} = \frac{rM}{1-(1+r)^{-M}} - 1 \end{aligned} \tag{10}$$

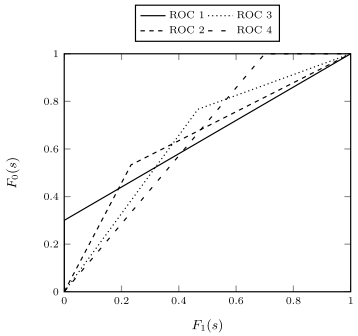
Parameter c^* is the cost of the action. Since rejecting a customer does not generate costs, we assume $c^* = 0$. Note that there is a cost involved with building the credit scoring model, but this cost is not related to a particular individual (i.e. it is not a variable cost). Therefore, in the long run, it is marginal for large portfolios – as is usually the case in consumer credit scoring (Edelberg, 2006) – and can be omitted.

Finally, we have to specify the probability distribution $h(b_0, c_1)$. We assume the ROI (c_1) to be constant for all loans, as is usually the case in consumer credit scoring. Equation (10) shows that at fixed terms the ROI depends on the interest rate. Furthermore, Edelberg (2006) noticed that the interest rates varied between 3% and 5% p.a. in over a decade. This justifies our assumption of a constant ROI in portfolios with similar terms.

Parameter λ (b_0), however, is much more uncertain, since recovery rates may vary between 0% and 100% of the total loan amount, and several distributions may arise (Somers and Whittaker, 2007). The empirical cumulative distribution $H(\lambda)$ for the three datasets used in this paper has three parts: a large part of the probability mass is situated in $\lambda = 0$, i.e. complete recovery of the loan amount. Another, smaller probability is observed for $\lambda = 1$ (i.e. complete loss). The remainder of the probability is spread out roughly evenly between zero and one. Thus, to calculate EMP, for each defaulter it is assumed that:

- $\lambda = 0$ with probability p_0, i.e. the customer pays back the entire loan,
- $\lambda = 1$ with probability p_1, i.e. the customer defaults on the entire loan,
- λ follows a uniform distribution in $(0, 1)$, with $h(\lambda) = 1 - p_0 - p_1$,

with p_0 and p_1 parameters specifying $h(\lambda)$, thereby providing flexibility to adjust it to the specific situation in a given company.



(a) Four ROC curves with equal AUC.

ROC Curve	AUC	EMP	Fraction rejected
1	0.65	1.65	2.7%
2	0.65	1.11	8.45%
3	0.65	0.89	11.83%
4	0.65	0.76	14.51%

(b) Comparison of AUC and EMP.

Figure 2: Four synthetic ROC curves and their EMP.

With these elements, the EMP measure, as introduced in (7), becomes:

$$\text{EMP} = \int_0^1 P(T(\theta); \lambda, \text{ROI}) \cdot h(\lambda) \text{d}\lambda, \tag{11}$$

with

$$P(t; \lambda, \text{ROI}) = \lambda \cdot \pi_0 F_0(t) - \text{ROI} \cdot \pi_1 F_1(t) \tag{12}$$

and $\theta = \text{ROI}/\lambda$ (since $c^* = 0$); see (5). Note that the cost-benefit ratio θ ranges from ROI (for $\lambda = 1$) to $+\infty$ (for $\lambda \to 0$). This means that the EMP integration does not cover the entire ROC curve, since the slope of the ROC curve varies from $+\infty$ (in the origin) to 0 (in $(1, 1)$). As a consequence, different ROC curves with the same AUC can lead to different EMP values. This is illustrated by the four ROC curves shown in Figure 2a. All four curves have the same AUC, but different EMP values, as shown in Figure 2b (for this calculation we assumed $\pi_0 = 0.20$ and $\text{ROI} = 0.2644$).

3.3 Empirical Estimation of EMP

For theoretical derivations, it is usually assumed that ROC curves are smooth. An empirical ROC curve, however, is stepwise constant with diagonal elements if there are ties. Furthermore, Fawcett (2006) showed that the points on the convex hull of a ROC curve are the set of optimal operational points. Also Verbraken et al. (2013) use the convex hull to calculate the EMP measure for customer churn models. This section will derive an analogous algorithm to calculate the EMP measure for credit scoring models based on the convex hull of the ROC curve.

Assume the convex hull of the ROC curve consists of m segments, and let (r_{1i}, r_{0i}) be the end point of segment i $(i = 1, \ldots m)$ with $(r_{10}, r_{00}) := (0, 0)$.

A score $s \in [r_{1i}, r_{1i+1}]$, will be the optimal cutoff value for the following value of λ (due to Equation (4)):

$$\lambda_{i+1} = \frac{\pi_1 \left(r_{1(i+1)} - r_{1i}\right)}{\pi_0 (r_{0(i+1)} - r_{0i})} \cdot \text{ROI} \qquad (i = 0, \ldots, m-1). \tag{13}$$

We define $\lambda_0 := 0$. The values λ are not bounded by 1 along the ROC curve. When approaching the point $(1, 1)$, λ becomes infinitely large.

Then, when calculating EMP, one replaces the series $\{\lambda_i | i = 0, \ldots, m\}$ by $\{\lambda_i | i = 0, \ldots, k+1\}$, with $k := \max\{i | \lambda_i < 1\}$, and $\lambda_{k+1} := 1$. Based on Equation (11) and (12), the EMP can be estimated by:

$$\begin{aligned} \text{EMP} = &[\lambda_0 \cdot \pi_0 \cdot r_{00} \cdot p_0 - \text{ROI} \cdot \pi_1 \cdot r_{10} \cdot p_0] \\ &+ \sum_{i=0}^{k} \int_{\lambda_i}^{\lambda_{i+1}} \lambda \cdot \pi_0 r_{0i} \cdot h(\lambda) \mathrm{d}\lambda - \sum_{i=0}^{k} \int_{\lambda_i}^{\lambda_{i+1}} \text{ROI} \cdot \pi_1 r_{1i} \cdot h(\lambda) \mathrm{d}\lambda \\ &+ \left[\lambda_{k+1} \cdot \pi_0 \cdot r_{0(k+1)} \cdot p_1 - \text{ROI} \cdot \pi_1 \cdot r_{1(k+1)} \cdot p_1\right]. \end{aligned} \tag{14}$$

The contributions in the square brackets are the probability masses for $\lambda = 0$ and $\lambda = 1$, respectively. Since λ is constant over the segments, r_{0i} and r_{1i} are constant in the end points of the segments, and $h(\lambda) = 1 - p_0 - p_1$, this can be written as:

$$\begin{aligned} \text{EMP} = &(1 - p_0 - p_1) \sum_{i=0}^{k} \left[\frac{\pi_0 r_{0i}}{2} (\lambda_{i+1}^2 - \lambda_i^2) - \text{ROI} \cdot \pi_1 r_{1i} (\lambda_{i+1} - \lambda_i)\right] \\ &+ \left[\pi_0 \cdot r_{0(k+1)} \cdot p_1 - \text{ROI} \cdot \pi_1 \cdot r_{1(k+1)} \cdot p_1\right]. \end{aligned} \tag{15}$$

Note that the contribution for $\lambda = 0$ vanishes since $r_{00} = r_{10} = 0$, and that $\lambda_{k+1} = 1$. Since the upper bound for λ is equal to 1, the integration does not cover the entire ROC curve.

3.4 Relationship Between EMP and AUC

AUC is a standard measure when evaluating binary classification models, and it has been related to most other measures by Hernández-Orallo et al.

(2012). We study the relationship between AUC and EMP, in order to enable the comparison to all other common measures.

We start from the definition of EMP (see (12)) and will carry out a variable transformation from λ to T, the optimal cutoff value. From Equations (4) and (9) we know that:

$$\lambda = \frac{\text{ROI} \cdot \pi_1 f_1(T)}{\pi_0 f_0(T)} = \frac{\text{ROI} \cdot \pi_1}{\pi_0 \cdot S_{ROC}}, \tag{16}$$

with $S_{ROC} = \frac{f_0(T)}{f_1(T)}$ the slope of the ROC curve for $s = T$. Substituting λ by T, leads to:

$$\text{EMP} = \int_{-\infty}^{T_{max}} \frac{\text{ROI}\pi_0\pi_1}{\pi_0 f(T)} \left[f_1(T)F_0(T) - f_0(T)F_1(T) \right] w(T)\mathrm{d}T, \tag{17}$$

where $w(T)$ is $h(\lambda)$ which absorbed the Jacobian of the variable transformation. Note that the integration bounds have changed due to the variable transformation. The lower bound ($\lambda = 0$) corresponds to a cutoff at $-\infty$. The upper bound ($\lambda = 1$), however, is linked to a finite cutoff $T = T_{max}$ due to the fact that for $\lambda = 1$, the slope of the ROC curve is $S_{ROC} = \text{ROI} \cdot \pi_1/\pi_0$. Note that we also assumed a one-to-one relationship between T and λ, which is valid when the ROC curve is bijective. Since the convex hull of the ROC curve is used, this condition is met.

If we assume that $w(T) = \pi_0 f_0(T)$, EMP can be written as:

$$\text{EMP} = \text{ROI} \cdot \pi_0\pi_1 \left\{ \underbrace{\int_{-\infty}^{T_{max}} f_1(T)F_0(T)\mathrm{d}T}_{\text{(a)}} - \underbrace{\int_{-\infty}^{T_{max}} f_0(T)F_1(T)\mathrm{d}T}_{\text{(b)}} \right\} \tag{18}$$

Since AUC is equal to:

$$\text{AUC} = \int_{-\infty}^{+\infty} f_1(T)F_0(T)\mathrm{d}T, \tag{19}$$

element (a) in Equation (18) is a part of the area under the ROC curve, from the origin up to T_{max}. Element (b) can be worked out as follows:

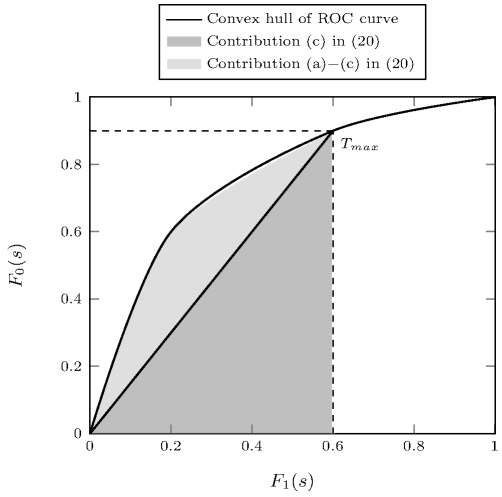


Figure 3: Relation between AUC and EMP illustrated on a ROC curve.

$$\int_{-\infty}^{T_{max}} f_0(T)F_1(T)\mathrm{d}T = \int_{-\infty}^{T_{max}} F_1(T)\mathrm{d}F_0(T)$$
$$= F_0(T_{max})F_1(T_{max}) - \underbrace{\int_{-\infty}^{T_{max}} f_1(T)F_0(T)\mathrm{d}T}_{\text{(a)}}$$

Replacing this result in Equation (18), we obtain:

$$\mathrm{EMP} = 2 \cdot \mathrm{ROI} \cdot \pi_0\pi_1 \left\{ \underbrace{\int_{-\infty}^{T_{max}} f_1(T)F_0(T)\mathrm{d}T}_{\text{(a)}} - \underbrace{\frac{1}{2}F_0(T_{max})F_1(T_{max})}_{\text{(c)}} \right\} \quad (20)$$

This has been illustrated on the graph in Figure 3, and shows how EMP measures the area under the ROC curve which is relevant for decision making.

It is straightforward to see that, for $T_{max} \to +\infty$ (i.e. moving the point corresponding to T_{max} to $(1, 1)$), the EMP measure becomes a linear combination of AUC:

$$\mathrm{EMP} \to 2 \cdot \mathrm{ROI} \cdot \pi_0\pi_1 \left\{ \mathrm{AUC} - \frac{1}{2} \right\}. \quad (21)$$

We study next under which conditions AUC and EMP are equivalent. This will be so when:

1. $w(T) = \pi_0 f_0(T)$,
2. $T_{max} \to +\infty$

Assumption 1 means that the probability distribution of λ depends on the score distribution, i.e. the output of a classifier. This clearly is not desirable since the costs are then depending on classification results themselves (as pointed out by Hand (2009)). Assumption 2 occurs when either $\pi_1 \to 0$ or ROI $\to 0$, which is unlikely to happen. Furthermore, in this case, the EMP measure vanishes to zero.

We argue that, in a business context, EMP is a better measure than AUC, since it analyzes the segments of the ROC curve which will indeed be used for decision-making.

4 Experimental Setting and Results

In this section, we demonstrate the potential of the proposed profit-based performance measure for credit scoring using real-world data. We focus on two widely used performance measures for comparison: AUC and accuracy (ACC). Our experimental procedure compares the use of each of these metrics, while focusing on two important aspects: parameter tuning and cutoff point determination. Section 4.1 describes the dataset, after which the experimental setup is discussed in Section 4.2. Section 4.3 and Section 4.4 respectively address the results with regards to parameter selection and to cutoff point determination. In Section 4.5 we perform a sensitivity analysis regarding some of the parameters used.

4.1 Consumer credit dataset

For our experiments, we use two datasets composed of loans for micro-entrepreneurs granted by a government organization between 1997 and 2007. The dataset characteristics are:

- New Borrowers: The first dataset consists of 37,492 loans granted to borrowers with no previous credit history with the organization. Each

loan is described by 16 variables, such as socio-demographic descriptors (age, employment, etc.) and an economic profile (ownership of properties, goods relevant for the application of the loan, etc.). The mean loan value is 1,123 EUR, the mean term is 2.5 years, and the dataset presents a default rate of 30.56%.

- Returning Borrowers: This dataset is formed by 103,466 loans granted to borrowers that already had a loan with the institution, i.e. there was credit history available. The variables presented before are complemented by eight credit history variables, such as the total number of past and current loans, maximum and mean arrears in previous loans, total amount granted, etc. The dataset has an average loan value of $1,150$ EUR, with a mean term of 2.4 years and a default rate of 20.47%.

Both datasets have already been used to develop credit scoring models (Bravo et al. (2013)), where the before mentioned variables turned out to be statistically significant. Furthermore, additional information was captured while the loan was being repaid (in particular information necessary to estimate the exposure and the loss, i.e. repayments made after default, total number of payments, collateral value, and recovery percentage at default). The EAD and LGD of defaulted loans are used to estimate the perceived loss. The granted amount is used to estimate each loan's profit; see Section 3.

4.2 Experimental setup

For our experiments we chose logistic regression and artificial neural networks (ANN) using logistic output transfer functions and one hidden layer. The reasons for these models are that logistic regression is by far the most commonly used method for credit scoring according to Thomas et al. (2002). ANN, however, gave best results on a large number of datasets (Baesens et al., 2003). The problem with ANN is that their black-box nature goes against Basel II/III regulations which require transparency in the loan granting process. Nevertheless, we use ANN as a benchmark to obtain best-case results.

Both datasets were divided into three subsets: validation set (20% of observations), used to vary the parameters, a training set (60% of observations) for training the model once the optimal parameters were found, and an independent test set (20% of observations) which is used for reporting results. The test set is the same across all experiments, so results are comparable throughout the paper.

Table 2: Results of parameter selection for "New Borrowers".

Performance Measure	Iters.	Hidden Layer Size	Value of PM	Optimal Fraction
Accuracy	450	29	0.6772	N/A
AUC	150	32	0.6834	N/A
EMP	150	32	0.0301	17.56%

Table 3: Results of parameter selection for "Returning Borrowers".

Performance Measure	Iters.	Hidden Layer Size	Value of PM	Optimal Fraction
Accuracy	50	25	0.768	N/A
AUC	250	26	0.827	N/A
EMP	400	21	0.023	10.16%

During the model building step of ANN, certain parameters, such as the number of hidden neurons and the number of training iterations need to be tuned, as will be shown in Section 4.3. Once the model is obtained, a decision has to be made regarding the classification of any given loan applicant. That decision is made by setting a cutoff point, which transforms the continuous score into a binary output; see Section 4.4 for more details. To assess the quality of the resulting credit scoring model, we compare three different measures: accuracy, total profit, and the average profit per accepted loan.

4.3 Parameter selection

For the ANN we determined two parameters: the number of training iterations and the number of hidden neurons. We conducted a grid search over a set of candidate parameters. The number of hidden neurons was chosen from the interval $[\frac{V}{2}, 2V]$, with V the number of input variables. The number of iterations was chosen from the interval $[50, 1000]$ in steps of 50 iterations.

In order to select the best set of parameters, a performance measure is needed. We will contrast AUC, accuracy, and EMP. To estimate EMP, we have programmed an R package, publicly available at CRAN (Bravo and Verbraken, 2014) with the necessary functions.

A model was trained in each of the grid elements for the corresponding parameter set. The best model for each measure (AUC, ACC, EMP) was then trained using the respective parameter set. Subsequently, the resulting models were applied to the test sets. Table 2 and Table 3 show the results for each of the datasets.

For "New Borrowers", both AUC and EMP select the same configuration: 150 iterations and 32 neurons, whereas accuracy expands the training time (450 iterations), but reduces the network's complexity (29 hidden neurons). For the dataset of returning borrowers, EMP selects 400 training iterations, but only 21 neurons, versus 250 iterations and 26 neurons for AUC.

The last two columns of Table 2 and Table 3 show the value of the respective performance measure (PM) and the optimal fraction (see Equation (8)). The performance in terms of AUC and accuracy is better for the returning borrowers, as one would expect given the richer data. This is not true, however, for EMP, where performance decreases from 3% to 2.3%. This seems counter-intuitive considering the richer data for "Returning Borrowers", but is explained by the fact that the dataset of new borrowers contains more defaulters (30.56%) than the dataset of returning borrowers (20.47%). Remember that EMP measures the *incremental* profit as compared to not building a credit scoring model, expressed as a percentage of the total loan amount. The more defaulters there are in a dataset, the easier it is to increase the profitability by building a credit scoring model, even with less data available. This also means that it is worthwhile to reject more applicants for "New Borrowers" (17.56%) as compared to "Returning Borrowers" (10.16%). Note that AUC and accuracy do not provide information about the profitability, one of the major strengths of EMP besides the optimal fraction, as will be discussed in the next section.

4.4 Cutoff point determination and results

After having trained a model, the cutoff point has to be determined. According to Bravo et al. (2013), there are two methods to take that decision (without using the EMP measure): (1) focusing on the cost of the operation, or (2) using the accuracy to define the optimal cutoff. The EMP measure, however, gives the optimal fraction of cases that should be rejected, which can then be transformed to the corresponding cutoff point. This characteristic is unique among all methods compared. For our benchmark we chose two approaches: if a model was built using accuracy as performance measure, then accuracy is also used to determine the cutoff point (maximum accuracy in training set). For AUC, we use the cutoff in which the derivative of the ROC curve tangent is equal to the ratio of the error costs, estimating for each point in the ROC curve the total loss perceived and the total utility lost. The cutoff value has been determined for the different performance

Table 4: Cutoff selection for each measure, ANN, new borrowers.

Model	Cutoff	Test Accuracy	Total Profit [EUR]	Profit/Loan [EUR]	Number of granted loans
No Model	N/A	69.48%	671,712	17.92	37,492
Accuracy-based	0.80	70.32%	718,304	104.22	6,892
AUC-based	0.60	69.21%	719,754	129.92	5,540
EMP-based	0.67	70.48%	764,680	124.84	6,125

Table 5: Cutoff selection for each measure, ANN, returning borrowers.

Model	Cutoff	Test Accuracy	Total Profit [EUR]	Profit/Loan [EUR]	Number of granted loans
No Model	N/A	79.83%	3,375,666	32.63	103,466
Accuracy-based	0.80	83.63%	3,751,123	209.71	17,887
AUC-based	0.70	82.49%	3,662,233	219.98	16,648
EMP-based	0.84	83.74%	3,781,266	204.81	18,462

Table 6: Cutoff selection for each measure, logistic regression, new borrowers.

Model	Cutoff	Test Accuracy	Total Profit [EUR]	Profit/Loan [EUR]	Number of granted loans
No Model	N/A	69.48%	671,712	17.92	37,492
Accuracy-based	0.60	69.77%	691,468	117.62	5,879
AUC-based	0.60	69.77%	691,468	117.62	5,879
EMP-based	0.61	69.81%	691,485	115.02	6,012

measures, after which the performance of the model has been assessed using the test set. Table 4 and Table 5 present the results for the ANN, using the parameters determined in Section 4.3.

The results employing logistic regression are shown in Table 6 and Table 7. From these tables, the advantages of using a profit-driven measure are evident. Considering the total profit, EMP brings the highest value among all combinations, with differences of up to 12% as compared to the scenario where no model is used.

EMP achieves best results regarding both criteria, accuracy as well as profit. The EMP-based model also selects the highest number of loans to be granted (except Table 4). This shows that the EMP-based model ensures better total rewards across the granted loans, even though some riskier loans could be accepted that might end in default.

Since logistic regression models do not have parameters to be tuned, the variation in the results is entirely attributable to the cutoff value. Hence,

Table 7: Cutoff selection for each measure, logistic regression, returning borrowers.

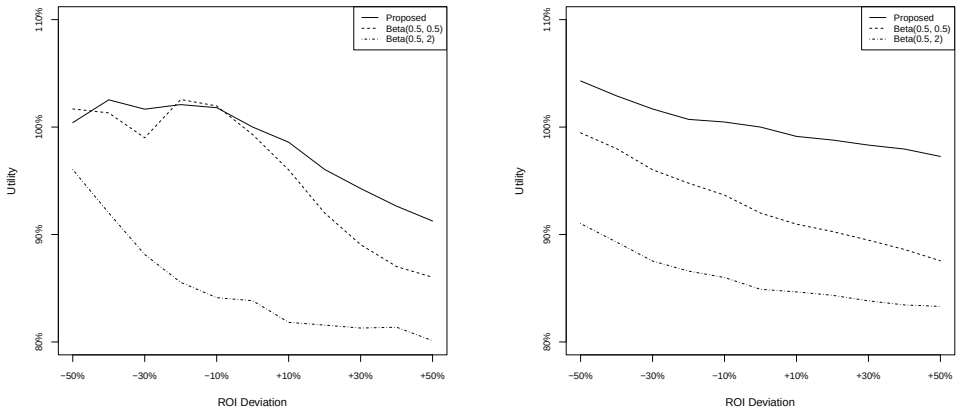
Model	Cutoff	Test Accuracy	Total Profit [EUR]	Profit/Loan [EUR]	Number of granted loans
No Model	N/A	79.83%	3,375,666	32.63	103,466
Accuracy-based	0.80	83.20%	3,648,778	201.00	18,153
AUC-based	0.70	82.27%	3,528,172	211.29	16,698
EMP-based	0.82	83.20%	3,687,437	199.81	18,455

these results underline the importance of an adequate cutoff point determination. For the new borrowers model, the difference between EMP-based and ACC-based cutoff value is very small in terms of accuracy, and it is non-existent in the case of returning borrowers. However, there are again differences in the total profit and the average profit per loan. This once again illustrates that the EMP-based cutoff determination is a better overall choice, resulting in the best accuracy and a significant, although lower, improvement in the monetary gain. The results are also consistent regarding the average profit per loan: EMP leads to the lowest one among the three models, and in this case AUC is the one with highest average profit. The reason is that the AUC model is much more restrictive, since we reproduce an already high default rate, with a cutoff of 0.60 for the first dataset and 0.70 for the second one, so it takes a much more conservative approach than the other two measures.

As shown in Bravo et al. (2013), a cutoff purely based on accuracy is too lenient to be used on its own, mostly because there usually is a much higher number of good borrowers than bad borrowers in a dataset. On the other hand, a cutoff based solely on average cost, or the proportion of them as is the case of AUC, is too restrictive, since this implies rejecting too many loans as each loan represents a risk. The use of a profit-oriented performance measure such as EMP has the advantage of achieving an excellent trade-off between both criteria, when just one cutoff is to be determined.

4.5 Sensitivity Analysis

The final question to be answered is how relevant the parameter selection in the EMP application process is. There are two decisions to be made: the ROI to be used and the distribution of λ, i.e. the losses perceived. In the previous sections, λ is determined using an ad-hoc distribution estimated from data.



(a) New borrowers (b) Returning borrowers

Figure 4: Sensitivity of ROI and λ distribution selection.

The ROI is the actual utility that each loan brings. To adjust this decision, Loterman et al. (2012) show that the LGD distribution follows, in most cases, either a distribution that decays exponentially, or a U-shaped distribution; both can be modeled using a Beta distribution with parameters $(2, 0)$ and $(0.5, 0.5)$, respectively. Additionally, we perturbed the obtained ROI value by steps of 5% in the range $[-50\%, +50\%]$, accounting for potential errors or divergences in this value. Then we measured the change in utility perceived if the model was applied with these quantities to the test set. The results for both universes (new and returning borrowers) are depicted in Figures 4a and 4b.

The graphs show that determining correctly both the distribution and the ROI is paramount to obtaining a maximum-profit result, as it is to be expected. The most critical parameter is the distribution of λ, since there can be a rapid decay in the obtained utility if a theoretical distribution is used. For example, in Figure 4b we observe a 10% drop in utility. The most likely explanation for this comes from the nature of the loss distribution: it can vary greatly among institutions, so using an approximation that is closer to reality, such as the one proposed in this paper, can improve significantly the final result.

For the ROI little deviation in utility close to the parameter's true value is observed. However, in both datasets there is a significant drop in utility if the

ROI is over-estimated, i.e., the estimated utility per granted Euro is larger than it actually is. By over-estimating the ROI, the institution is more prone to take riskier positions, accepting more bad borrowers to take advantage of the larger utility that each good borrower seemingly brings, and thus causing a larger loss than by using the correct ROI. In turn, under-estimating ROI may improve the results by a small percentage, which can be explained by particular loans that cause a large loss.

5 Conclusions and future work

This paper presents a profit-based performance measure based on EMP, a recently proposed general classification performance measure. Our contribution is to adapt this general approach to the specific case of consumer credit scoring. This performance measure accounts for the benefits generated by healthy loans and the costs caused by loan defaults. As a result, the profit-based measure allows for profit-driven model selection, i.e. it allows practitioners identifying the credit scoring model which increases profitability most. Furthermore, the proposed measure provides the optimal cutoff value, which is required in order to transform the continuous score from a credit scoring model into a binary decision. This feature which other performance measures do not have is a major advantage of the EMP measure.

The results of our experiments indicate that using the EMP measure for model selection leads to more profitable credit scoring models. Moreover, employing the EMP-based cutoff value further increases the profitability by granting more loans than traditional approaches. Besides, the lender gains insight in the monetary reward of implementing a credit scoring model, which improves its practical use.

This paper focuses on profit-based model performance measurement. An interesting venue for future research is to incorporate the profitability criterion into the model building step. Currently, models typically optimize a statistical criterion, such as e.g. maximum likelihood. A focus shifted to profitability may provide further opportunities for improving credit scoring practices. A second opportunity for future research is to apply this measure to other types of credit. The EMP parameter values determined in this paper are tailored to consumer credits. It would be interesting to determine the respective cost distributions for other types of credit, such as e.g. mortgages.

Acknowledgements

The authors acknowledge support from the "Instituto Sistemas Complejos de Ingeniería" (ICM: P-05-004-F, CONICYT: FBO16), the Flemish Research Council (FWO, Odysseus grant B.0915.09), Conicyt's Becas Chile Program (PD-74140041), and the Explorative Scientific Co-operation Programme 2012-2013 which funded the project "Development of rule-based classification models using profit maximization" (BIL 12/01).

References

Ali, S., Smith, K., 2006. On learning algorithm selection for classification. Applied Soft Computing 6 (2), 119–138.

Baesens, B., Van Gestel, T., Viaene, S., Stepanova, M., Suykens, J., Vanthienen, J., 2003. Benchmarking state-of-the-art classification algorithms for credit scoring. Journal of the Operational Research Society 54 (6), 627–635.

Baldi, P., Brunak, S., Chauvin, Y., Andersen, C., Nielsen, H., 2000. Assessing the accuracy of prediction algorithms for classification: an overview. Bioinformatics 16 (5), 412–424.

Bernstein, A., Provost, F., Hill, S., 2005. Toward intelligent assistance for a data mining process: An ontology-based approach for cost-sensitive classification. IEEE Transactions on Knowledge and Data Engineering 17 (4), 503–518.

Bravo, C., Maldonado, S., Weber, R., 2013. Granting and managing loans for micro-entrepreneurs: New developments and practical experiences. European Journal of Operational Research 227 (2), 358–366.

Bravo, C., Verbraken, T., 2014. EMP: Expected Maximum Profit for Credit Scoring. R package version 1.0. URL: `http://CRAN.R-project.org/package=EMP`

Broverman, S. A., 2010. Mathematics of investment and credit. Actex Publications, Winsted, USA.

Domingos, P., 1999. Metacost: A general method for making classifiers cost-sensitive. In: Proceedings of the fifth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. ACM, San Diego, CA, pp. 155–164.

Edelberg, W., 2006. Risk-based pricing of interest rates for consumer loans. Journal of Monetary Economics 53 (8), 2283–2298.

Fawcett, T., 2006. An introduction to ROC analysis. Pattern Recognition Letters 27 (8), 861–874.

Finlay, S., 2010. Credit scoring for profitability objectives. European Journal of Operational Research 202 (2), 528–537.

Hand, D., 2005. Good practice in retail credit scorecard assessment. Journal of the Operational Research Society 56 (9), 1109–1117.

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

Hernández-Orallo, J., Flach, P., Ferri, C., 2012. A unified view of performance metrics: translating threshold choice into expected classification loss. Journal of Machine Learning Research 13, 2813–2869.

Loterman, G., Brown, I., Martens, D., Mues, C., Baesens, B., 2012. Benchmarking regression algorithms for loss given default modeling. International Journal of Forecasting 28 (1), 161–170.

Mays, E., Nuetzel, P., 2004. Credit Scoring for Risk Managers: The Handbook for Lenders. South-Western Publishing, Mason, OH, Ch. Scorecard Monitoring Reports, pp. 201–217.

Provost, F., Fawcett, T., 2001. Robust classification for imprecise environments. Machine Learning 42 (3), 203–231.

Somers, M., Whittaker, J., 2007. Quantile regression for modelling distributions of profit and loss. European Journal of Operational Research 183 (3), 1477–1487.

Thomas, L. C., 2009. Consumer Credit Models: Pricing, Profit and Portfolios. Oxford University Press, New York.

Thomas, L. C., Crook, J. N., Edelman, D. B., 2002. Credit Scoring and its Applications. SIAM, Philadelphia.

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. European Journal of Operational Research 218 (1), 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 25 (5), 961–973.

Zhou, Z., Liu, X., 2006. Training cost-sensitive neural networks with methods addressing the class imbalance problem. IEEE Transactions on Knowledge and Data Engineering 18, 63–77.