A Robust Profit Measure for Binary Classification Model Evaluation*

Franco Garrido[1], Wouter Verbeke[2], and Cristián Bravo[3]

[1]Programa de Magíster en Gestión de Operaciones, Universidad de Talca, fgarridoc@alumnos.utalca.cl
[2]Faculty of Economic and Social Sciences and Solvay Business School, Vrije Universiteit Brussel, Belgium, wouter.verbeke@vub.be
[3]Department of Decision Analytics and Risk, Southampton Business School, University of Southampton, c.bravo@soton.ac.uk

Abstract

Using profit-based evaluation measures is a necessity in business-oriented contexts, as they aid companies in making cost-optimal decisions. Among the measures that effectively include the true nature of costs and benefits in binary classification, the expected maximum profit (EMP) has been used successfully for churn prediction and credit scoring, and defined in general for binary classification problems. However, despite its competitive results against the most frequently used measures, the EMP relies on a fixed probability distribution of costs and benefits, the range of which in real applications is not entirely known. In this paper, we propose to extend this measure by adding random shocks to these distributions. We call this new measure the R-EMP, following the convention of the analogous EMP measure. Our metric adds a stochastic component to each point of

*NOTICE: this is the author's version of a work that was accepted for publication in Expert Systems with Applications in September 19, 2017, published online as a self-archive copy after the 24 month embargo period. 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: Franco Garrido, Wouter Verbeke, Cristin Bravo, A Robust Profit Measure for Binary Classification Model Evaluation, In Expert Systems with Applications, 2017, Accepted: Available Online https://doi.org/10.1016/j.eswa.2017.09.045.

the cost-benefit distributions, assuming that costs and benefits have a fixed probability, but its distribution range is subject to an external shock, which can be different for each cost or benefit. The experimental set-up is focused on a credit scoring application using a dataset of a Chilean financial institution, with the attribute selection for a logistic regression being accomplished using the AUC, EMP, H-measure, and R-EMP as the selection criteria. The results indicate that the R-EMP measure is the most robust metric for achieving the greatest profit for the company under uncertain external conditions.

Keywords: Supervised Binary Classification; Business analytics; Performance measures; Profit-driven analytics

1 Introduction

The development of performance measures for classification methods has become an important task in data analytics, given their critical role in operations management (Baesens et al., 2009). In many industries, information analysis has become the only method of differentiation (Davenport, 2006). McAfee and Brynjolfsson (2012) stated the following: *"You can't manage what you don't measure"*.

The most common predictive analytics problem, binary classification, has the goal of classifying elements into one of two classes. Most models that are used to solve this problem, such as logistic regression or neural networks, return a continuous score that indicates how likely each case is to belong to one of the two classes, and it is up to the practitioners to determine the threshold that defines the frontier between the two classes. There is a wide variety of measures available for evaluating the performances of algorithms, among which the receiver operating characteristic (ROC) curve and, especially, the area under the ROC curve (AUC) are the most frequently used (Bradley, 1997).

Researchers have demonstrated that measures like the AUC are not suitable for environments in which misclassification costs are different (Hand, 2009). There are measures that consider the true nature of costs effectively; among them, the H-measure (Hand, 2009), the maximum profit (MP) measure (Verbeke et al., 2012), and the expected maximum profit (EMP) measure (Verbraken et al., 2013) are some of the best known. The last two measures are designed as total profit measures; the former (MP) assumes certainty in

cost parameters, obtaining the maximum benefit and the optimal threshold, while the latter (EMP) is a stochastic version of the MP measure, in which cost and benefit parameters are described by a probability distribution, leading to the estimation of the expected maximum profit.

In real applications, the MP and EMP measures have contributed to the selection of models aligned with the nature of costs and benefits. Some applications of these measures include determining the optimal fraction of the consumers to be targeted for a churn prevention campaign at a telecommunications company (Verbeke et al., 2012) and estimating the maximum profit of credit scoring models (Verbraken et al., 2014a).

This paper presents a more robust version of the measure for the case in which the uncertainty comes not only from the profit parameters but also from external random shocks, which we call the **R**obust **E**xpected **M**aximum **P**rofit (R-EMP) measure. Our method is based on the rationale that random shocks can modify an originally rigid profit estimation; thus, if the distribution of these potential shocks can be known, then the R-EMP will fit the profit estimation, taking into account this information.

This paper is structured as follows: in Section 2, we describe the state-of-the-art profit-based performance measures for evaluating classification models. Section 3 shows the R-EMP formulation, specifying its structure and all the considerations regarding its implementation. Section 4 presents the experimental design of this work, which consists of a synthetic case (Section 5) and an empirical case using loan data (Section 6). This section includes both the benchmark of the R-EMP against other measures and a case in which we show the use of the measure as a decision-making tool. Finally, conclusions are presented in Section 7.

2 Evaluation Measures for Classification Models

Within the field of predictive analytics, a classification problem refers to the task of determining a class label for an element from a set of known labels. When the number of possible labels is only two, this task is known as binary classification. Data mining/analytics models for supervised classification allow determining the labels for new cases with unobserved labels, and binary classifiers usually return a probability of belonging to one of two classes, lead-

ing to the necessity of defining a threshold that separates the two classes, i.e., a cut-off value.

According to Ali and Smith (2006), there is no unique measure that can be used to find the best classification model. Baldi et al. (2000) showed that the most frequently used measures are percentages, different kinds of distances, correlation, entropy, mutual information, and ROC curves. Various authors have applied ROC curves in many applications and used the area under the ROC curve, i.e., the AUC, as the performance measure, mainly because the AUC does not depend on a cut-off value and is insensitive to class distribution (Bradley, 1997). The AUC is also easily implemented (Brown and Davis, 2006) and interpreted (Fawcett, 2006a). Fawcett (2006b) showed that ROC graphs are not a suitable reference when there are instance-varying, i.e., case-dependent, classification costs. To fix this problem, he developed a variant called the ROCIV. Hand (2009) detected an additional AUC weakness and proposed an alternative measure known as the H-measure, which is coherent and therefore should yield a more reliable indication of performance than the area under the ROC curve. Correa Bahnsen et al. (2014) indicated a significant need that exists for measures that are sensitive to classification costs. They proposed an algorithm for credit scoring that allows constructing a classifier while simultaneously taking into account the variable nature of costs. Several publications presented measures or techniques for evaluating classification models. Most of the proposed measures were compared to the AUC measure. Among these works, we find McDonald (2006) introducing a measure that has the characteristic of allowing an unbiased (with or without cost sensitivity) comparison between different classifiers; De Bock and Van den Poel (2011) presenting a methodology that considers a rotative evaluation of performance measures; and Aman et al. (2015) proposing a set of measures that allows comparing models in terms of independence, reliability, volatility, and cost. Later, Clemente-Císcar et al. (2014) proposed two measures, one based on benefits and another based on returns, with the objective of evaluating the performance of a customer retention campaign.

In this paper, we elaborate upon the MP framework as introduced by Verbeke et al. (2012) for supervised binary classification problems. The MP measure Verbeke et al. (2012), which is the first measure in the MP framework, considers the different costs of classification and at the same time facilitates the obtaining of the optimal cut-off value to be applied when operating the obtained classification model, which is a practical advantage when compared to alternative measures. Verbraken et al. (2013) developed a stochastic

version of the MP measure, called the EMP, which models each cost using a probability distribution, allowing the estimation of the expected value of the maximum profit. The MP and EMP measures have been implemented and adopted successfully in churn prediction (Verbraken et al., 2014b) and credit scoring (Verbraken et al., 2014a). Both the MP and EMP measures are discussed in more detail in the next section.

2.1 Profit-based Evaluation Measures

The MP measure is designed as a profit-based function, in which the parameters b_0 and c_0 (b_1 and c_1) are, respectively, the benefit and cost associated with correctly and incorrectly classifying a good (bad) case, $F_0(t)$ and $F_1(t)$ denote the cumulative fraction of, respectively, goods and bads, with a score assigned by the classifier below the variable cut-off t. The average classification profit per case resulting from adopting a threshold t is defined as follows:

$$P(t; b_0, c_0, b_1, c_1) = b_0\pi_0 F_0(t) + b_1\pi_1(1-F_1(t)) - c_0\pi_0(1-F_0(t)) - c_1\pi_1 F_1(t) \quad (1)$$

Since all parameters in this function, i.e., b_0, b_1, c_0, and c_1, are assumed to be positive, then it follows that the theoretical overall maximum profit can be attained when $F_0(t) = 1$ and $F_1(t) = 0$. This, however, only occurs when a classifier perfectly discriminates between goods and bads. The maximum profit that can be obtained for a non-perfect classifier is defined in Equation (2), where T is the optimal cut-off value that defines the threshold score separating the two classes.

$$MP = \max_{\forall t} P(t; b_0, c_0, b_1, c_1) = P(T; b_0, c_0, b_1, c_1) \quad (2)$$

The value of T can be obtained under the maximization of the profit function and satisfies the first-order condition for the maximization of the average profit, P:

$$\frac{f_0(t)}{f_1(t)} = \frac{\pi_1(b_1 + c_1)}{\pi_0(b_0 + c_0)} = \frac{\pi_1\theta}{\pi_0} \quad (3)$$

with π_0 and π_1 being the prior class probabilities and $\theta = \frac{b_1+c_1}{b_0+c_0}$ being the *cost-benefit ratio*. Hence, the optimal threshold T depends on the cost-benefit ratio θ. The MP measure has the merit of being oriented toward the

central business objective, i.e., profit maximization, and also the practical benefit of providing the optimal cut-off value.

More recently, the EMP measure has been proposed as an extension of the MP (Verbraken et al., 2013). This measure was designed considering that in real application settings, it is often difficult to estimate accurate values for benefit and cost parameters or costs and benefits may be case dependent; therefore, these parameters were modeled using a probability distribution. The EMP measure is presented in Equation (4) and accounts for the involved uncertainty, with $\omega(b_0, c_0, b_1, c_1)$ being the conjoint probability distribution of the cost and benefit parameters. For each possible combination of the cost and benefit parameters (b_0, c_0, b_1, c_1), the optimal threshold T is determined using Equation (3) as a function of the cost-benefit ratio θ.

$$EMP = \int_{b_0} \int_{c_0} \int_{b_1} \int_{c_1} P(T(\theta); b_0, c_0, b_1, c_1) \cdot \omega(b_0, c_0, b_1, c_1) db_0 dc_0 db_1 dc_1 \quad (4)$$

3 The R-EMP measure

In this article, we extend the EMP measure by acknowledging that in addition to the uncertainty regarding the cost and benefit parameters, as captured by $\omega(b_0, c_0, b_1, c_1)$ in the EMP measure, these parameters can change because of a *random shock*. Such a random shock can either be an external or internal event, or, despite the name, a steady evolution of the operational setting in which the classification model functions. For instance, changes in economic conditions or technological evolutions can have an impact on the operational setting and are examples of external shocks that affect profitability. On the other hand, changes in customer behavior, customer relationship management, business strategies and business processes are examples of internal shocks affecting profitability. Therefore, the presented R-EMP measure extends the EMP approach, which models the benefit and cost parameters using a probability distribution to capture either uncertainty in estimating the exact values or to account for inherent variability across cases, by superimposing a perturbation or uncertainty on top of these distributions to capture the effect of such random shocks on profitability. As such, we aim to achieve a more robust measure and, through the use of this measure, as will be explained in a later section, to obtain more robust classification models for improved decision-making under variable conditions. Thus, we include the

impact of external information, given by the random shock, and the potential correlation between the components of profit, as the random shock can affect each measure separately or the same shock can affect multiple parameters at once. For example, both the benefits and the costs of an application can be affected by inflation, which is both external to any intrinsic uncertainty and the same for all measures, thus creating correlation between the costs and benefits.

If Equation (4) is considered to give an estimation of the maximum profit but there is an external event η (that is out of our control) affecting this value, then we can extend the definition in Equation (4) to incorporate such a random shock. This leads to the definition of the extended, more robust R-EMP measure. For this purpose, the benefit and cost parameters are defined by a probability function and, in addition, by a random shock. As explained, these random shocks correspond to perturbations of benefits and costs. The extended expressions for the benefit and cost parameters are given in Equations (5), (6), (7) and (8).

$$b'_0 = f\left(b_0, \eta_{b_0}\right) \tag{5}$$

$$c'_0 = f\left(c_0, \eta_{c_0}\right) \tag{6}$$

$$b'_1 = f\left(b_1, \eta_{b_1}\right) \tag{7}$$

$$c'_1 = f\left(c_1, \eta_{c_1}\right) \tag{8}$$

Then, b'_0 represents the stochastic benefit of correctly classifying a case of class 0, which is a function of the *original* stochastic variable b_0, as adopted in the EMP measure, and additionally of a random variable (η_{b_0}), representing an external random shock. b'_1 is the benefit of correctly classifying a case of class 1 and is defined as a function of the stochastic variables b_1 and η_{b_1}. The cost variables c'_0 and c'_1 define the cost of classifying a case as class 0 that belongs to class 1 or vice versa, respectively; these variables have been defined in a manner similar to that of the benefit parameters that include a stochastic random shock.

If ω represents the joint distribution function of these parameters, then the R-EMP measure is defined by the following equation:

$$R-EMP=\int_{b_0'}\int_{c_0'}\int_{b_1'}\int_{c_1'} P(T(\theta');b_0',c_0',b_1',c_1')\cdot\omega(b_0',c_0',b_1',c_1')db_0'dc_0'db_1'dc_1' \tag{9}$$

Note that the definitions of the random shocks, η_j, allow the specification and inclusion of highly complex probability distributions for the cost and benefit parameters, incorporating both stochastic effects that are intrinsic to the operation and random shocks that are external to the user. For example, by defining the distribution d of each component η_j as $d(\eta_j)=d(\nu_j,\varepsilon)$, with ν_j being an internal random shock affecting only one parameter and ε being an external random shock affecting every parameter $j\in\{b_0',b_1',c_0',c_1'\}$, then each parameter will depend on internal and external stochastic effects.

3.1 R-EMP for Credit Scoring

The R-EMP measure proposed in the previous section is a generic profit measure that can be adapted towards application in any business setting that requires accounting for stochastic costs and benefits, which may be subject to shocks. In this section, we define the functional form of the R-EMP measure for credit scoring.

In credit risk management, one critical decision involves whether or not to grant a loan to a consumer. Credit scorecards, especially application scorecards, are classification models that are typically developed for making this decision in a data-driven manner (Thomas et al., 2002; Siddiqi, 2016). By defining the cost and benefit parameters and establishing the involved profit formula, Verbraken et al. (2014a) adapted the EMP measure as defined in Equation (4) to evaluate credit scorecards:

$$EMP^{CS}=\int_0^1 P(T(\theta);\lambda,ROI)\cdot h(\lambda)d\lambda \tag{10}$$

with

$$P(t,\lambda,ROI)=\lambda\cdot\pi_0F_0(t)-ROI\cdot\pi_1F_1(t) \tag{11}$$

The EMP measure for credit scoring defined in Equation (10) involves a single uncertain parameter, i.e., the loss fraction of a loan, λ, which is the benefit of correctly classifying a bad applicant (i.e., b_1). This fraction is defined in Equation (12) below as the amount owed in the case of default,

i.e., the exposure at default (EAD), multiplied by the loss after all collection measures have been exhausted, i.e., the loss given default (LGD), and divided by the original amount of the loan (A).

$$b_1 = \lambda = \frac{\text{LGD} \cdot \text{EAD}}{A} \tag{12}$$

Additionally, when an application scorecard wrongly classifies a good applicant as a bad applicant, an opportunity cost is incurred equal to the total return over the investment (ROI). This cost is considered relative to the amount of the requested loan (A) and will be assumed to be constant across cases (Verbraken et al., 2014a).

In defining the R-EMP measure for credit scoring, we adopt the same approach as for the EMP measure, except that the λ variable is replaced in the same way that b_1 is replaced in Equation (7), i.e., instead of λ, we now consider λ', which is expressed by $\lambda' = f(\lambda, \eta_\lambda)$. This function adds a random shock η_λ to the stochastic loss fraction λ defined in Equation (12), which impacts the potential losses. The distribution of the loss fractions can be impacted, for instance, by changing the economic conditions and collateral prices. The R-EMP measure for credit scoring is defined as follows:

$$R - EMP^{CS} = \int_0^1 \left(f(\lambda, \eta_\lambda) \cdot \pi_0 F_0(t) - ROI \cdot \pi_1 F_1(t)\right) \cdot h(\lambda') d\lambda' \tag{13}$$

In the following sections, we will extensively illustrate the use of this measure, using both a synthetic and a real credit scoring dataset.

4 Experimental Settings

An evaluation measure has practical use when it assists in decision-making during model development and operation. Common decisions that have to be made are, for example, choosing model parameters to maximize the classification performance, choosing the best attributes to include in a classification model, or choosing the cut-off point that is to be used for making a binary decision (e.g., to accept or reject a loan application) based upon the continuous score that is produced by the classification model. In the following sections, we will illustrate how the R-EMP measure supports and improves such decision-making in a business context.

For this purpose, we conducted experiments on a synthetic and on a real credit scoring dataset. The experiments on the synthetic dataset focus on pinpointing differences between the R-EMP and other commonly used evaluation measures, such as the AUC, H-measure and EMP, while the empirical case study focuses on the use of the R-EMP measure for practical decision-making.

In the empirical case, given that we have data spanning 12 years, we will show both a comparison of measures overall with an out-of-time benchmark and a year-by-year comparison. This allows observing the behavior of these measures both over short and long terms and to assess their robustness, which was the main objective in developing the R-EMP measure.

5 Synthetic Case

For the experiments reported in this section, a synthetic dataset has been created to compare the characteristics of a selection of measures often adopted to evaluate credit scorecards.

The goal of this experiment is to compare how the applied measures behave when we subject the profit to a higher level of uncertainty. For this, we built a synthetic dataset, with a binary target variable (default and non-default) following a binomial distribution with size $n = 1000$ and probability of success $p = 0.5$. We created 12 attributes originating from six distributions (binomial, exponential, normal, Poisson, uniform and Weibull). Each attribute had different parameters per class; thus, there was a slight overlap (15%) between the distributions for each class. This process resulted in attributes that do not allow for linearly separable classes but do allow the achievement of a very high predictive accuracy.

To construct the costs and benefits, we first created the loan amount A for each of the n cases, and then the EAD and the LGD were set based on this value. Benefits were defined as $b = LGD \cdot EAD$ and costs as $c = ROI \cdot A$, meaning that there are benefits when defaulters are identified correctly because we are avoiding the loss of b and that there are losses when rejecting good applicants because we are not earning c. Following this, we calculated the value of λ using Equation (12). The dataset was then randomly divided into a training and a test set. To introduce a shock, we perturbed the value of λ following Equation (14).

$$\lambda' = \lambda + N(\mu = 0, \sigma^2 = 0.2 \cdot \lambda^2) \tag{14}$$

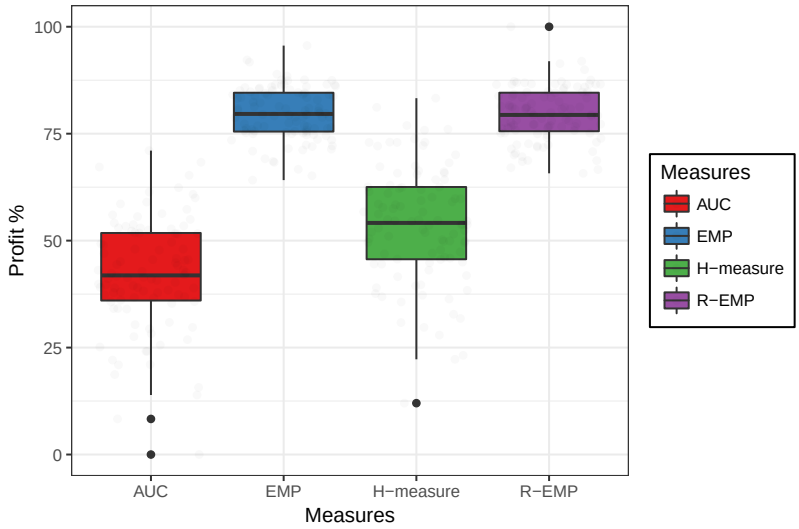
We want to study the performance of different measures when more noise is added to the dataset to study the behavior of different performance measures in this situation. Because each variable is simulated with an overlap, as more variables become available, the uncertainty (noise) in the model will increase. We selected random forest as our underlying classifier to extract as much information as possible from the variables, to filter random noise that can be easily eliminated by a model, and to focus on the impact of the uncertainty that cannot be filtered and the impact of the profit subject to a random shock.

The simulation starts with no variables, and in each iteration, we include in the model the attribute that maximizes each evaluation measure. Once the procedure converges, i.e., when no further improvements can be achieved by adding more attributes, the resulting profit is calculated for the test sample. This process is repeated 100 times, generating new attributes without varying the underlying distributions. After all profits are calculated, the maximum profit achieved across all 100 iterations is calculated and used to normalize the results. Hence, the experiment simulates different samples of the same population with the same statistical structure but also with different external shocks to the profit structure.

Figure 1 shows the profit in the test sample as a proportion of the maximum profit for that sample. Here, the power of the profit-driven measures is demonstrated. The AUC shows a high variability, with profit proportions ranging from 0% to 70% and an average of just 42%. The H-measure performs slightly better, with less variability and a somewhat higher average profit proportion being achieved (54%), but yields results much lower than those achieved by the EMP and R-EMP. Both of these measures yield an average profit of 80% (EMP: 79.8%, R-EMP: 79.6%). It can also be seen that the R-EMP achieves a smaller overall standard deviation for most experiments but that there are outliers, which occur when the maximum profit is achieved. This effect is consistent with the design of the measure: the model will generally select the most robust measure (small deviation), but that measure will be a maximum profit measure.

The robustness of the measure is shown in Table 1. In 57 out of 100 times, the R-EMP reaches the maximum value. The small increased standard deviation of the R-EMP occurs only because of the outlier value (without this

Figure 1: Synthetic Dataset - Profit Out-of-time by Measure



value, the R-EMP falls to 5.9%), and acknowledging some small difference, the means are basically the same. We can conclude that the R-EMP is equivalent to the EMP in that it presents better behavior when the noise in the sample is bigger; explicit information regarding this variability can be captured by adding a new exogenous variable.

Table 1: Results of synthetic experiments

Measure	Average	s.d.	Times best measure
R-EMP	79.6%	6.3%	57
EMP	79.8%	6.1%	43
AUC	42.2%	13.0%	0
H-Measure	53.8%	13.5%	0

6 Empirical Case

The dataset, consisting of loans for small businesses, that was used to evaluate the presented empirical case was provided by a Chilean financial institution. The dataset contains 9 attributes, which after preprocessing and

one-hot encoding result in 16 predictive variables, as shown in Table 2. The attributes can be grouped into three types:

- Loan variables: which describe the characteristics of the operation. Besides the amount and the term - in two forms, to account for different types of loans that might be either very short or very long term -, whether the borrower has collateral, or if the loan has a guarantor.
- Sociodemographic variables: These variables describe the owner of the business. It is composed of the age of the borrower, a grouping of the zip code where the borrower operates in terms of default rates, and the ownership status of their main plot of land. For the last variable, the borrower can either rent, own, have free use of the land, part-own, or have other arrangements, resulting in five categories.
- Business segment: These variables describe the business segment in which the borrower operates. The number of plots was segmented considering the default rates at each group, resulting in three categories: one, two or more than two plots. The second variable, the business segment, was divided following the same logic into four categories, each describing a macro-economic sector in the economy.

Table 2: Variables used in empirical experiment.

Variable	Description	Type
Guarantor	If the borrower had a guarantor for the loan	Loan
Collateral	If the borrower had collateral on the loan	Loan
Term (months)	Term of the loan (in months)	Loan
Term (year)	Term of the loan (in years)	Loan
Age	Age of the borrower	Sociodemographic
Housing	Ownership status	Sociodemographic
ZipGroup	Region of the country	Sociodemographic
Properties	Number of plots of land	Business
EconSector	Main economic sector	Business

The number of cases in the dataset is approximately $40,000$, with a default rate of 24%. The dataset includes loan applications from 1996 to 2008, but the cases are not distributed equally across years: 35% of the cases stem from the first two years and the last four years, with 13% of the total cases corresponding to the most recent year. The last four years are selected as

the out-of-time sample, and the cases occurring between 1996 and 2004 are randomly divided into a training and a test sample in a 70% versus 30% proportion, respectively. For more details regarding this dataset, one may refer to Bravo et al. (2013).

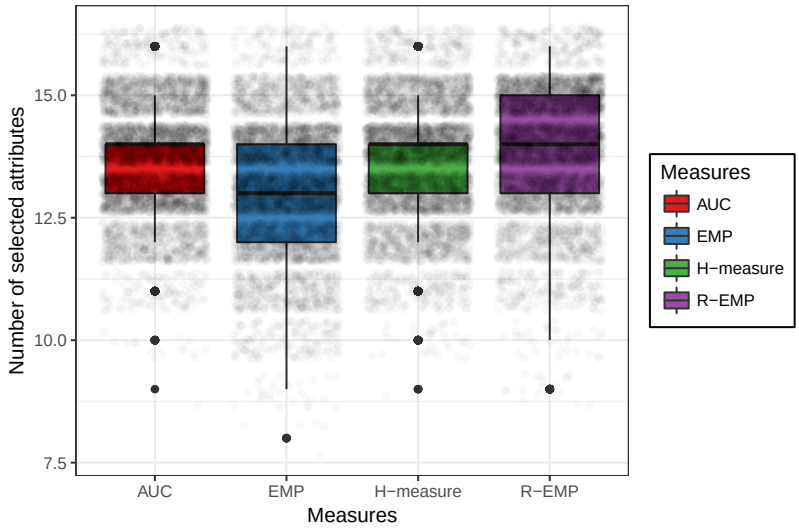
6.1 Using R-EMP as a Profit Measure

We first repeat the experimental setup applied in Section 5 for the synthetic dataset, allowing us to study how the R-EMP measure behaves when shocks affect a dataset with more complex data structures. As before, we want to choose the best model as more information becomes available, following a forward selection-like procedure. We stop adding information once there is no improvement in the measure and record the profit over the test set. As in the synthetic case, we assume a normal distribution to model the random shock (η_λ), and the number of replications (n_e) is again set to 100.

From Figure 2, we can see that the AUC and the H-measure generally lead to a similar number of attributes being selected; the EMP selects a smaller number of attributes; and the R-EMP selects the highest number. The R-EMP, AUC and H-measure select very similar average numbers of attributes, with the EMP being slightly off this mark. Again, as in the previous section, the R-EMP selects the models more robustly, with only small deviations with respect to the mean number of attributes, instead of either less or more attributes than average typically being selected, as with the other measures.

Using the test sample, which covers the same time period as the training sample, we obtain the results shown in Table 3. The main insight from this table is that the R-EMP outperforms the other measures in all years except for 1999. This result again exemplifies the goal of the R-EMP, i.e., to provide a measure that is resistant to random shocks and variability. The selected model exhibits a stable, high performance over the selected period of time, as opposed to exhibiting such performance only in *average* years, as observed for the EMP and the H-measure. Additionally, note that in 2002, which involved a severe economic downturn, the model that was developed using the R-EMP measure is the only model that yielded a positive profit. From 2002 onward, the R-EMP model clearly outperforms the other models. Upon calculating the total average profit, the R-EMP is found to achieve the highest profit, followed at some distance by the EMP, H-measure, and AUC. Finally, the total average standard deviation is calculated over the full set of experimental results. As expected, the R-EMP has the lowest standard

Figure 2: Empirical Dataset - Number of Selected Attributes by Measure



deviation with respect to profit.

A final comparison between the models developed using the different measures considers their predictive ability for an out-of-time test sample. The results are given in Figure 3. According to this figure, it is possible to observe that the density of profit obtained using the EMP and R-EMP is more concentrated than that obtained using the AUC and H-measure; also, the EMP and R-EMP yield less dispersion, again hinting at the robustness of the proposed measure. The average profit for the out-of-time test sample is only slightly different across the models. According to Table 4, the R-EMP yields the highest profit (51,930 EUR), with the EMP in second place, closely followed by the AUC and the H-measure.

Both Table 4 and Figure 3 indicate that even though the difference in average profit is small, the use of the R-EMP does consistently yield a higher profit model, which, importantly, is either less disperse or more robust. This outcome is exactly what the measure was designed to accomplish.

6.2 R-EMP as a decision-making tool

This section is devoted to showing how the R-EMP can be used as a decision-making tool. Therefore, two additional experiments are conducted: the first

Table 3: Empirical Dataset - Average Profit Year-by-year in EUR by Measure

Year	AUC	EMP	R-EMP	H-measure
1996	17,273	18,010	18,373	18,277
1997	6,391	5,043	7,877	6,511
1998	41,981	47,271	52,202	42,487
1999	121,923	121,625	109,097	123,311
2000	143,919	143,667	145,059	144,901
2001	103,714	103,375	104,923	103,987
2002	-3,683	-2,686	1,776	-2,856
2003	59,515	68,550	78,670	59,893
2004	41,235	50,974	59,539	40,548
Total average	532,269	555,827	577,515	537,059
Total standard deviation	217,210	216,181	209,535	218,179

Table 4: Empirical Dataset - Profit Out-of-time ± Standard Deviation in EUR by Measure

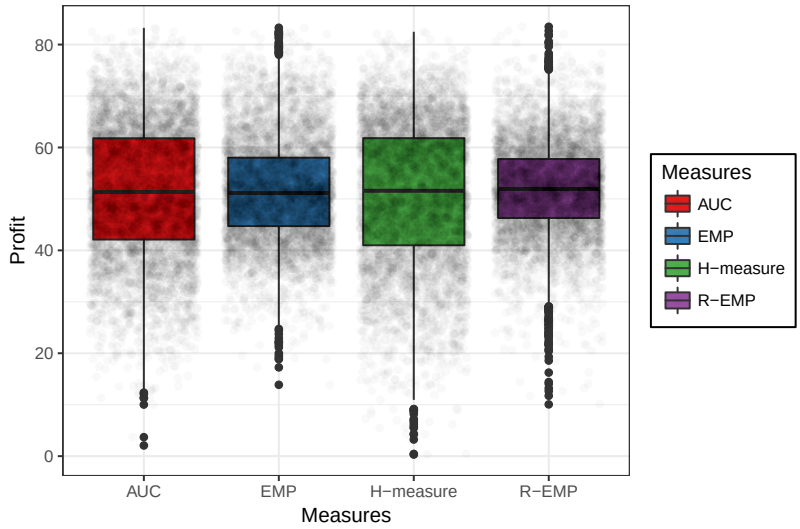
AUC	EMP	R-EMP	H-measure
51,270 ± 13,138	51,732 ± 10,162	51,930 ± 9,057	50,665 ± 13,936

experiment concerns parameter tuning, while the second experiment concerns determining a cut-off value.

For this purpose, and for illustrating the use of the developed profit driven evaluation measure for decision making, an artificial neural network is trained. More specifically, since often adopted in a business analytics context (Verbeke et al., 2012), a multilayer perceptron (MLP) with one hidden layer is trained, also given the importance of tuning the characteristics of an MLP, for which various evaluation measures can be adopted. As a result, we obtain an indication of the potential gain from a profit perspective that may be achieved when consistently adopting the proposed measure during development of a classification model, in this case, a credit risk model.

Note that the evaluation measure can be adopted in combination with any supervised learning approach, including alternative neural networks and deep learning approaches (Schmidhuber, 2015). A broad benchmarking study to compare various supervised learning techniques and performance evaluation measures, is beyond the scope of this paper, but is considered an important topic for further research.

Figure 3: Empirical Dataset - Profit Out-of-time in EUR (thousands) by Measure



An MLP requires a large number of hyperparameters to be tuned to function optimally. One typical method for setting parameter values is using a grid-search over various combinations of parameter values, thus limiting the infinite search space. The parameters tuned in this experiment are the number of units in the hidden layer of the network and the maximum number of iterations of the algorithm. By using a training sample that includes data from the real dataset (described in the previous section) for the years from 1996 to 2004 and an out-of-time test sample that includes data for the years from 2005 to 2008 to evaluate the performance, we select the best combination of parameters using the Accuracy, AUC and R-EMP measures. Given the strong correlation between the AUC and H-measure that was observed in the previous sections, we have omitted the H-measure from this experiment and instead included the Accuracy. The same reasoning was used for the EMP and R-EMP, focusing, of course, on the latter. Based on the number of attributes, the size of the hidden layer is tested from 8 neurons to 32 neurons in steps of 1, and the number of iterations is set from 50 iterations to 1000 iterations in steps of 50.

In Table 5, the optimal values of the parameters using the different measures are shown. The table shows the values for the iterations and hidden

layer size, the value of the performance measure (PM) at which that parameter combination occurred, and, for the R-EMP, the optimal fraction (cut-off value) suggested for the model.

Table 5: Parameter selection decision driven by different measures

Performance Measure	Iterations	Hidden layer size	Value of PM	Optimal fraction
Accuracy	800	17	0.6945	N/A
AUC	50	20	0.7364	N/A
R-EMP	50	30	0.0127	6.37%

To operate a credit scorecard in practice to decide whether to accept or reject loan applications, a cut-off value needs to be adopted. Setting a cut-off value also allows a straightforward comparison of the performances of the various models in terms of profitability. The R-EMP measure implies a cut-off to be used, reported as the optimal fraction in Table 5. However, for the Accuracy and AUC, we need to choose the cut-off, which for Accuracy is selected based on the score of the test sample for which the accuracy was maximal, while for the AUC, we select the score for which the tangent of the ROC curve is equal to the proportion between average costs for acceptance and rejection, following Hand (2009).

In Table 6, the behavior of the models based on the selected cut-off value and parameters is reported. The R-EMP, AUC and Accuracy measures are considered as alternatives, and a baseline scenario, in which no model is used to make a decision, i.e., loans are always granted, is reported as a reference. Under the baseline scenario, 4,566 loans are granted, leading to a total negative profit of -382,197 EUR. When using the Accuracy-based model and selecting the optimal cut-off for the validation sample, there is an improvement in terms of profit, resulting in a positive number of 11,567 EUR. AUC-based decision-making leads to an improvement in the total profit of up to 22,609 EUR. Note that the test accuracy is considerably lower when using the AUC and that the number of rejected loans is relatively larger. Finally, when adopting the R-EMP-based model, we further improve upon the AUC-based model profit, yielding 45,028 EUR and significantly reducing the number of rejected loans.

As reported in Table 6, the R-EMP achieves both the highest accuracy and profit for the test sample. The accuracy of the Accuracy-based model is very similar to the accuracy of the model using the R-EMP, with the marginal gain in accuracy when using the R-EMP probably due to the robustness of

Table 6: Cut-off value decision driven by different measures

Model	Cut-off	Test accuracy	Total profit (EUR)	Profit/loan (EUR)	Number of granted loans
No model	N/A	80.20%	-382,197	-83.70	4,566
Accuracy-based	0.78	57.60%	11,567	2.74	4,220
AUC-based	0.65	47.82%	22,609	6.89	3,283
R-EMP-based	0.75	58.18%	45,028	10.53	4,275

the measure, as it considers distributions over the sample as opposed to only the averages. The result in terms of the achieved profit is expected, as the R-EMP measure is designed to maximize profit over populations using distributions over samples, whereas the other measures are not. These results show that the R-EMP can be used with confidence to make decisions during model development and for model selection, supporting the method as a decision-making tool.

7 Conclusions

In business environments, it is imperative to strive for optimal and robust decision-making, taking into account risks and evolving conditions. This article contributes to achieving optimal and robust decision-making by proposing a novel performance metric for improving the decision-making process in developing classification models, i.e., by designing a variation of the EMP measure for evaluating classification performance when random shocks may affect the distribution of the profit parameters. The novel measure, dubbed the R-EMP, is experimentally evaluated using both a synthetic and real credit scoring dataset to assess its appropriateness and to compare its characteristics with those of the EMP measure as well as the AUC and H-measure.

The results of the experiments indicate that the R-EMP effectively outperforms the EMP, as well as the AUC and H-measure, when there are external factors affecting the profit variables, thus demonstrating that taking into consideration the impact of random shocks improves the quality of decisions in terms of profit. Additionally, the experiments provide evidence that the use of the EMP and R-EMP effectively leads to the selection of different models and yields better performance in terms of achieved profits. Moreover, using the R-EMP for decision-making results in reduced variability and therefore improved robustness. Moreover, the presented results show that the novel measure can be reliably used to select the best model and to define the cut-off point for future use.

The experiments on a real dataset confirm that the use of the R-EMP results in a more robust model than the use of the EMP, which leads to the conclusion that by adding perturbations to the profit variables, the EMP measure can effectively be improved. For credit scoring applications, the R-EMP measure leads to better decisions than the main measures reported in the literature. Hence, the R-EMP appears to be a robust measure for building predictive analytics models within highly variable business environments.

Acknowledgments

We acknowledge the support of Conicyt Fondecyt Initiation Into Research 11140264.

References

References

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

Aman, S., Simmhan, Y., Prasanna, V. K., 2015. Holistic measures for evaluating prediction models in smart grids. Transactions on Knowledge and Data Engineering, IEEE 27 (2), 475–488.

Baesens, B., Mues, C., Martens, D., Vanthienen, J., 2009. 50 years of data mining and OR: upcoming trends and challenges. Journal of the Operational Research Society 60 (1), S16–S23.

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

Bradley, A. P., 1997. The use of the area under the ROC curve in the evaluation of machine learning algorithms. Pattern Recognition 30 (7), 1145–1159.

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.

Brown, C. D., Davis, H. T., 2006. Receiver operating characteristics curves and related decision measures: A tutorial. Chemometrics and Intelligent Laboratory Systems 80 (1), 24–38.

Clemente-Císcar, M., San Matías, S., Giner-Bosch, V., 2014. A methodology based on profitability criteria for defining the partial defection of customers in non-contractual settings. European Journal of Operational Research 239 (1), 276–285.

Correa Bahnsen, A., Aouada, D., Ottersten, B., 2014. Example-dependent cost-sensitive logistic regression for credit scoring. In: International Conference on Machine Learning and Applications. p. 7.

Davenport, T. H., 2006. Competing on analytics. Harvard Business Review (84), 98–107.

De Bock, K. W., Van den Poel, D., 2011. An empirical evaluation of rotation-based ensemble classifiers for customer churn prediction. Expert Systems with Applications 38 (10), 12293–12301.

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

Fawcett, T., 2006b. ROC graphs with instance-varying costs. Pattern Recognition Letters 27 (8), 882–891.

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

McAfee, A., Brynjolfsson, E., 2012. Big data: the management revolution. Harvard Business Review (90), 60–6.

McDonald, R. A., 2006. The mean subjective utility score, a novel metric for cost-sensitive classifier evaluation. Pattern Recognition Letters 27 (13), 1472–1477.

Schmidhuber, J., 2015. Deep learning in neural networks: An overview. Neural networks 61, 85–117.

Siddiqi, N., 2016. Intelligent Credit Scoring: Building and Implementing Better Credit Risk Scorecards. John Wiley & Sons.

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

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., Bravo, C., Weber, R., Baesens, B., 2014a. Development and application of consumer credit scoring models using profit-based classification measures. European Journal of Operational Research 238 (2), 505–513.

Verbraken, T., Verbeke, W., Baesens, B., 2013. A novel profit maximizing metric for measuring classification performance of customer churn prediction models. Transactions on Knowledge and Data Engineering, IEEE 25 (5), 961–973.

Verbraken, T., Verbeke, W., Baesens, B., 2014b. Profit optimizing customer churn prediction with bayesian network classifiers. Intelligent Data Analysis 18 (1), 3–24.