UNIVERSIDAD DE TALCA

FACULTAD DE INGENIERÍA

A PROFIT MEASURE WITH DETERMINISTIC AND STOCHASTIC EFFECTS FOR DATA MINING MODELS

por

Franco Eduardo Garrido Céspedes

Tesis para optar al grado de

Magíster en Gestión de Operaciones

Prof. Supervisor Cristián Bravo Róman

Octubre, 2016

TALCA
UNIVERSIDAD
CHILE

CONSTANCIA

La Dirección del Sistema de Bibliotecas a través de su encargado Biblioteca Campus Curicó certifica que el autor del siguiente trabajo de titulación ha firmado su autorización para la reproducción en forma total o parcial e ilimitada del mismo.

Curicó, 2019

Abstract

Nowadays business environments are becoming more competitive, just those who take informed decisions remain successful, so it is imperative to take informed decisions to reach the businesses ultimate goal, achieve a profit as high as possible.

Business Analytics is an area which includes the use of Data Mining models to take informed business decisions. There is a wide variety of models, but there are few measures for those models that takes in consideration the costs and benefits associated to the decisions driven by them. In this context we aim to enrich the state-of-art on decision making tools by generating a new profit-driven measure. We identify a profit-driven measure and extent its domain aiming to model the variability of costs and benefits for highly-variable business environments, providing a measure able to model a wider number of business contexts. This new approach consist into adding random shocks to the cost-benefit distributions of a measure known as Expected Maximum Profit, the new measure is called R-EMP.

The research established a synthetic and empirical experiment in a context of credit scoring. The synthetic case was developed simulating a credit scoring data set, and the empirical case was based on Chilean financial institution dataset. For both experiments we used the Logistic Regression model to classify if a client fall in default or not, using as selecting criterion of attributes our measure (R-EMP), a commonly used measure known as AUC, a measure of losses known as H-measure and our primary measure known as EMP. Then, we selected our random shocks as random normally distributed information, replicating 5,000 times and simultaneously applied stratified sampling accordingly to the two imbalanced classes. The results of both experiments agrees that using R-EMP measure as selection criterion drives to the improvement of the total profit for the company.

As conclusion we validate the incorporation of random shocks to improve a decision making tool EMP measure. We recommended the use of R-EMP measure as selection criteria on highly-variable business environments.

For future research it would be interesting to incorporate random shocks to another business applications like churn prediction, also trying to test another variety of shocks, and also to capture a real business shock and incorporate this as input for the R-EMP measure.

Key Words: Classification, Business Analytics, Performance measures, Profit-driven Analytics.

Resumen

En la actualidad los entornos de negocios se están volviendo más competitivos, solo aquellos que toman decisiones informadas se mantienen exitosos, por lo tanto es imperativo tomar siempre decisiones bien informadas con el objetivo de lograr el fin último de los negocios, alcanzar la utilidad más alta posible.

Business Analytics es un área que incluye el uso de modelos de Data Mining para tomar decisiones de negocios de manera informada, pero a pesar de que existe una amplia variedad de modelos, existen pocas medidas para estos modelos que consideren los costos y beneficios asociados a las decisiones conducidas por éstos. En este contexto nosotros tratamos de enriquecer el estado del arte de la toma de decisiones a través de una mejora a una medida de utilidad, para lo cual identificamos una medida de utilidad y extendimos su dominio esperando modelar la variabilidad de los costos y beneficios en los ambientes de negocios con alta variabilidad, proporcionando una medida capaz de modelar un número más amplio de contextos de negocios. Este nuevo enfoque consiste en la adición de choques aleatorios a las distribuciones de costo y beneficio de una medida conocida como Expected Maximum Profit, la nueva medida es llamada R-EMP.

La investigación puso en marcha un experimento sintético y otro empírico en un contexto de credit scoring. El experimento sintético fue desarrollado simulando un conjunto de datos, y el caso empírico fue basado en un conjunto de datos de una empresa financiera de Chile. Para ambos experimentos se utilizó Regresión Logística para clasificar si un cliente cometió default o no, utilizando como criterio de selección nuestra medida (R-EMP), una medida frecuentemente usada conocida como AUC, una media de perdidas conocida como H-measure y nuestra medida primaria EMP. Luego seleccionamos nuestros choques aleatorios como información aleatoria normalmente distribuida, replicamos 5.000 veces y simultáneamente muestreando de forma estratificada las dos clases desbalanceadas, teniendo como resultado de ambos experimentos un acuerdo sobre la mejora en la utilidad cuando la medida R-EMP fue utilizada como criterio de selección.

Como conclusión validamos la incorporación de choques aleatorios para mejorar una herramienta para la toma de decisiones como lo es la medida EMP, entonces nosotros recomendamos el uso de la medida R-EMP como criterio de selección en ambientes de negocios muy variables.

Para investigaciones futuras sería interesante incorporar choques aleatorios en otras aplicaciones de negocios como la fuga de clientes, además de probar otras variedades de choques y además capturar choques asociados a contextos reales para incorporarlos como entrada a la medida R-EMP.

Palabras Claves: Clasificación, Análisis de negocio, Medidas de desempeño, Analítica orientada a las utilidades.

Index

1 Introduction

1.1 Background and problem

Data Mining is a powerful tool to find patterns in data, allows to enlighten researchers in the search of information hidden data, providing knowledge to support the decision making process. Among these tools, as part of predictive analytics, binary classification, which is basically an assignment of an element to one of two possible and know classes, is one of the most used algorithms. For those algorithms there are several measures to evaluate the performance, being the AUC measure the most frequently used (Bradley, 1997), but measures like this are not suitable when misclassification cost are different (Hand, 2009). There are measures that includes this kind of costs, among them we find H-Measure (Hand, 2009), but also there are measure like the Maximum Profit (MP) (Verbeke et al., 2012) and the Expected Maximum Profit (EMP) (Verbraken et al., 2013) which includes the benefit of taking a right decision. This latter approach is fundamental in business decisions, because minimizing costs in certain environments is not the same as profit maximization which is the business ultimate goal. Being aware of these useful measures, we identify a need of developing profit-driven measures because there is no a unique measure able to find the best classification model (Ali and Smith, 2006), and also the development of performance measures for classification methods has become an important task in Data Mining because of their critical role in operations management (Baesens et al., 2009).

1.2 Objectives of the study

Because of being align with business and also it successful results, we selected EMP measure and add random information in order to fit those to those environments with high variability on benefits and cost, we named the new measure as R-EMP. The research main goal is to:

- Propose and generate a general new profit-driven measure for binary classification.

Then, the particular objectives are:

- Set-up our measure for the credit scoring context.
- Evaluate its performance with other measures.

1.3 Contribution

Through this research we expect to contribute with a new measure for the decision making process when classification methods are used, expanding the set possibilities when evaluating the performance of those models, so our main goal is to check if the addition of random information could lead to a measure able to drive to better business decisions.

1.4 Methodology

In order to validate the measure we performed two experiments, the first is based in a totally random dataset, and the second belongs to a Chilean financial institution. For both experiments we used AUC, EMP, H-measure and R-EMP metrics to select attributes for a Logistic Regression, then we compared their results.

1.5 Main results and conclusions

Results indicates that R-EMP is suitable when classification benefits and cost suffer of highly variability, this measure provides the best results for the synthetic and empirical case. Then, the incorporation of random information effectively could improve the business decisions. Finally, as conclusion, we recommend the set-up of the measure for another common business scenarios like churn prediction.

1.6 Article estructure

In the first section of the article there is an introduction, followed with a section with a background on measures to evaluate classification models, then the third section provides the R-EMP definition, the fourth section defines the experimental settings, sections 5 and 6 describes, respectively, the synthetic and empirical case, and the last section provide the conclusions.

2 A profit measure with deterministic and stochastic effects for Data Mining models

Una medida de utilidad con efectos deterministicos y estocasticos para modelos de *Data Mining*

Abstract

There is a large variety of performance measures for evaluating Data Mining and Analytics models, but in a business-oriented environment there are but a few profit-based evaluation measures. Using such evaluation measures is a necessity in this context, since they aid companies in making cost-optimal decisions. Among the measures that effectively include the true nature of costs and benefits there is the Expected Maximum Profit (EMP), which has been used successfully for churn prediction and credit scoring, but, despite its competitive results against the most frequently used measures, relies on a fixed probability distribution of costs and benefits whose range 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 R-EMP, following the convention of the analogous EMP measure. Our metric adds a stochastic component to each point of the cost-benefit distributions, assuming that costs and benefits have a fixed probability, but its distribution range is affected by an external shock, which can be different for each cost or benefit. The experimental setup is focused on a credit scoring application on a dataset of a Chilean financial institution, with the variable selection for a logistic regression done using AUC, EMP, H-measure, and R-EMP as selection criteria. Results indicate that the R-EMP measure is the most suitable metric for achieving the greatest profit for the company.

Key Words: Classification, Business Analytics, Performance measures, Profit-driven Analytics.

2.1 Introduction

Data mining is a specialized area in knowledge discovery, which, through the use of statistics, is able to contribute to the generation of relevant knowledge. It has a sophisticated group of algorithms that serve different purposes, and, among those, classification (part of predictive analytics) is the main application in the field. In many applications we have to classify elements into one of two classes (binary classification). This kind of problem usually returns a continuous score that indicates how close each case is to one of the two classes, and it is a task for the practitioners to determine the threshold that defines the frontier between the two classes. There is a wide variety of measures available to evaluate the performance of algorithms, among which the receiver operating characteristic (ROC) curve and 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 where misclassification cost 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 measure (MP) (Verbeke et al., 2012) and the Expected Maximum Profit measure (EMP) (Verbraken et al., 2013). 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 MP where costs are described by a probability distribution, leading to estimation of the expected maximum profit. In real applications, the MP and EMP measures have contributed to select 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). This successful experience was replicated in the credit scoring field (Verbraken et al., 2014a).

This paper presents an attempt to determine if it is possible to generate a more robust version of the measure through the injection of external information. A new version, based on the EMP measure, is proposed, called the R-EMP. Our method is based on the rationale that profit estimation is rigid and may not be as it has been estimated. R-EMP fits the profit estimation based on the new information. In Section 2 of this paper we describe the state-of-the-art of performance measures. Section 3 shows the R-EMP, specifying its structure and all the considerations of its implementation. Section 4 shows the experimental design, and Sections 5 and 6 present experimentation results. Finally, conclusions are presented in Section 7.

2.2 Evaluation measures for classification models

The development of performance measures for classification methods has become an important task in data mining, 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, "*You can't manage what you don't measure*", and in this sense Clemente et al. (2010) indicate that for correct analysis of information it is necessary to generate measures capable of delivering clear messages.

A classification problem refers to the task of determining a class label for an element from a set of known labels. When the possible labels are just two, this is known as binary classification. Data mining/analytics models for classification allow determining the labels for new cases with unobserved labels, and usually binary classifiers return a probability of belonging to one of two classes, leading to the necessity of defining the threshold that separates the two classes, i.e., the cut-off value.

According to Ali and Smith (2006) there is no unique measure able to find the best classification model. Baldi et al. (2000) show that the most frequently used measures are percentages, different kinds of distances, correlation, entropy, mutual information, and ROC curves. Various authors in many applications have applied ROC curves, and the area under the ROC curve (AUC) as performance measures, mainly because of the fact that AUC does not depend on a cut-off value and is insensitive to class distribution (Bradley, 1997). It is also easily implemented (Brown and Davis, 2006) and interpreted

(Fawcett, 2006a). Fawcett (2006b) shows that ROC graphs are not a suitable reference when there are different classification costs, so in order to fix this problem he developed a variant called ROCIV. Hand (2009) also detected an AUC weakness, proposing an alternative new measure to overcome the problems he encountered. This measure is known as the H-measure, whose coherency and performance should yield better results than the area under the ROC curve. Correa Bahnsen et al. (2014) said that there is a huge need for developing measures sensitive to classification costs, and they proposed an algorithm for credit scoring that allows constructing a classifier, while simultaneously taking into account the variable nature of costs. There are many publications in which measures or techniques for evaluating classification models are proposed, and in most of them all those measures are compared to the AUC. Among those works we find McDonald (2006) who proposed his own measure that has the characteristic of allowing an unbiased (with or without cost-sensitive) comparison between different classifiers; Bock and Van den Poel (2011) who proposed a methodology that considers a rotative evaluation of performance measures, and Aman et al. (2015) which proposed 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 retaining campaign.

Verbeke et al. (2012) proposed a deterministic measure that allows estimating the maximum profit (MP) of a classification model. This measure considers the different costs of classification, and at the same time facilitates obtaining the optimal cut-off value to be applied in the model, which is a practical advantage when compared to alternative measures. Verbraken et al. (2013) developed a stochastic version of the MP measure, the Expected Maximum Profit (EMP), which models each cost through a probability distribution, and this fact allows the estimation of the expected value of the maximum benefit. The MP and EMP measures have been implemented successfully in churn prediction (Verbraken et al., 2014b) and credit scoring (Verbraken et al., 2014a).

2.2.1 Profit-based evaluation measures

In Verbeke et al. (2012) the MP measure is proposed, designed as a profit-based function in which the parameters b_0 and c_0 (b_1 and c_1) are the benefit and cost associated with good (bad) applicants, and t is the cut-off value that the classifier uses. It is defined as follows:

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

Since all parameters in this function, b_0, b_1, c_0, c_1 are supposed to be positive, then it follows that the maximum is attained when $F_0(t) = 1$ and $F_1(t) = 0$, and this occurs when a classifier performs perfectly. So the maximum profit is defined in eq. (2), where T is the optimal cut-off value that defines the limit between the two classes.

$$MP = \max_{\forall t} P(t; b_0, c_0, b_1, c_1) = P(T; b_0, c_0, b_1, c_1) \tag{1}$$

The value of T could be obtained under the maximization of the benefit function, in which the author identified that the threshold T depends on the relation of benefits and costs $\theta = \frac{b_1+c_1}{b_0+c_0}$. The MP measure has the merit of being oriented to the central business objective, i.e., profit maximization, and also the practical benefit of aiding to determine the optimal cut-off value.

Recently the EMP measure has been proposed, (Verbraken et al., 2013), as an extension of the MP. This measure was designed considering that in real contexts it is difficult to estimate values for benefit and cost parameters accurately, so those parameters could better be modeled through a probability distribution. The measure is presented in eq. (3), where $\omega(b_0, c_0, b_1, c_1)$ corresponds to the conjoint probability distribution of classification costs, and θ has to be determined for each parameter (b_0, c_0, b_1, c_1).

$$EMP = \int_{b_0} \int_{c_0} \int_{b_1} \int_{c_1} P(T(\theta); b_0, c_0, b_1, c_1) \cdot \omega(b_0, c_0, b_1, c_1) db_0 dc_0 db_1 dc_1 \tag{3}$$

2.3 The R-EMP measure

As mentioned above, in real business situations it is hard to estimate a value for benefits and costs, and that is why EMP assumes a probability distribution over them. In R-EMP it is assumed that the probability distributions exist, but now each benefit (b_0, b_1) and cost (c_0, c_1) is affected by a random shock, so every combination of these values will define a different profit from the one estimated by eq. (1). Therefore R-EMP focuses its attention on the scenario in which each benefit and cost probability remains the same, but each event could be affected by some kind of perturbation which leads to a different value from the maximum profit.

If eq.(1) is considered to give an estimation of the maximum profit, but there is an external event η which is out of our control affecting this value, we can extend the definition in eq. (3) to incorporate random shocks, leading to our measure, R-EMP. Now all the benefit and cost variables are defined by a deterministic and known value, and also by a random shock. These random shocks correspond to perturbations to benefits and costs. The new expressions for benefits and costs are given in equations 4, 5, 6 and 7.

$$b_0' = f(b_0, \eta_{b_0}) \tag{4}$$

$$c_0' = f(c_0, \eta_{c_0}) \tag{5}$$

$$b_1' = f(b_1, \eta_{b_1}) \tag{6}$$

$$c_1' = f(c_1, \eta_{c_1}) \tag{7}$$

Now b_0' is the benefit of correctly classifying a case of class 0, and it is defined by a function that depends on a random variable (η_{b_0}) , and a deterministic one (b_0). b_1' is the benefit of correctly classifying a case of class 1, and it is defined by a function that depends on variables b_1 and η_{b_1}, where the former is deterministic and the latter is random. Also we have cost variables c_0' and c_1', which define the cost of misclassification. The first is associated with classifying a case as class 0 when it is really class 1, and the second indicates the opposite. Both costs are defined in the same way as the benefit variables, defined by a function that depends on a deterministic and also a random variable.

The R-EMP measure now follows the equation given by:

$$R - EMP = \int_{b_0'} \int_{c_0'} \int_{b_1'} \int_{c_1'} P(T(\theta'); b_0', c_0', b_1', c_1') \cdot \omega(b_0', c_0', b_1', c_1') db_0' dc_0' db_1' dc_1' \qquad (8)$$

2.3.1 R-EMP for credit scoring

Up to this point R-EMP is a generic random profit measure that can be adjusted for each business-case, so now we will define its functional form in a real decision context.

In Finance and Banking, one of the main decisions is the granting of consumer credit to consumers. Credit Scoring, which is the set of models and underlying techniques involved in these decision making processes, (Thomas et al., 2002), aids in making a granting decision.

Verbraken et al. (2014a) developed the EMP approach for credit scoring, defined as follows:

$$\int_0^1 P(T(\theta); \lambda, ROI) h(\lambda) d\lambda \qquad (9)$$

In this approach there is just one variable involved, the loss fraction of each credit denoted by λ (b_1). There is a cost involved, the return over investment (ROI), but it is considered to be a constant parameter. This modeling indicates that the benefit is obtained when a defaulter is identified correctly because that fraction of the borrowed principal is not lost, so the company earns this amount. Also the model indicates that there is an amount that it is not earned per each wrongly rejected credit which is a constant percentage of the credit (the ROI).

Everything remains as in EMP except for the R-EMP approach except for the λ variable, in which λ is now replaced in the same way that b_1 is replaced in eq.(6), i.e., instead of λ, the $\lambda' = f(\lambda, \eta_\lambda)$. By making these changes, the R-EMP credit scoring approach is as follows:

$$\int_0^1 (\lambda' \cdot \pi_0 F_0(t) - ROI \cdot \pi_1 F_1(t)) \cdot h(\lambda')\, d\lambda' \qquad (10)$$

At this point the random profit measure has been inserted into the business decision context, the credit scoring field.

The remaining part of this article is devoted to describing the experiment, the comparison approach, and its results.

2.4 Experimental settings

Any evaluation measure is useful as long as it assists in decision-making. Common decisions that have to be made are, for example, deciding upon the parameters that a model requires to maximize classification capacity, or deciding on the best features to build a classifier. To illustrate how R-EMP supports decision-making in a business context, we will focus on the second problem, since it is common for all predictive models. The same process can be applied easily for most decisions that are required when building a predictive model.

To reach our goal, an experiment on two different data sets, a synthetic and a real one, was conducted. For the experiment the Logistic Regression was selected as the underlying technique because it is an industry standard; it is simple; and it performs well in credit scoring (Baesens et al., 2003). In order to assess the results, we drove a benchmarking process against the AUC, EMP, and H-measure.

For the experiment it is necessary to define a cut-off value, the threshold that separates the scores returned by the logistic regression to be used. When using AUC and H-measure it is up to the researcher to decide the approach to determine this value. This is because those measures do not provide information about this directly. So, in order to be fair, we are going to used the point of the Receiving Operating Characteristic (ROC) in which a tangent line, with slope based on benefits over costs, minimizes the distance to the optimal point $(0, 1)$ of the ROC. In the case of EMP the comparison is direct, because the former aids in finding the threshold, as does R-EMP also.

In the synthetic case the dataset is generated randomly, then divided into training, test, and out-of-time sets. By applying backward selection, i.e., starting with all the available attributes and deleting them one by one based on the measures to compare until all the remaining attributes contribute to improving profit, a set of attributes for each measure is obtained. Then, a logistic regression is trained with the final model, and the profit of the test and out-of-time sets is assessed. However, as the dataset is synthetic, the obtained profit will be scaled in terms of percentage, and after that, the process will be replicated in order to examine the stability of the measures. So each time we have different data sets, but all of them have the same statistical properties. In the empirical case, the methodology will be the same as in the synthetic case, but in this case the dataset is given. The goal of this experiment is to show the effectiveness of R-EMP as a decision making tool in a real-life scenario.

In both experiments results will show profit results, but a year-by-year comparison will be made first, and then, an out-of-time benchmark will be conducted. This way we are going to see how the measures used behave over both short and long terms.

The definition of the experiment, will be followed by a data description, separated by sections for each dataset, along with a presentation and discussion of the results.

2.5 Synthetic case

In this section a synthetic dataset is simulated with the objective of generating evidence of the decision of the different measures when an arbitrary dataset is used. The process for achieving this goal started with the generation of the target variable (default), for which a Binomial Distribution, with size n and probability of success p, was used. Then, for each of the n cases, a total of 30 fully random attributes was created using the following probability distributions (5 attributes for each one): Binomial, Exponential, Normal, Poisson, Uniform and Weibull. After the creation of the variables, completely random values for benefits and costs were generated. This process began with the selection of a maximum amount A for each of the n cases, then the Exposure at Default (EAD) and the Loss Given Default (LGD) were set from 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 losing b, and there are losses when rejecting good applicants because we are not earning c.

When defining the perturbation (η_λ) the fact that benefits or costs present stochastic behavior was taken into account, so based on its positive-negative support, we chose the Normal Distribution, expecting that its parameter of variance might help in representing the underlying perturbation. Also, since R-EMP has stochastic components, it would have different values in each computation, so we replicate its computation ($n_e = 100$) times, and then take the average value to represent the R-EMP result. Finally the eq. (11) represents the shape and magnitude of the random variable.

$$\lambda' = \lambda + N(\mu = 0, \sigma^2 = 0.2 \cdot \lambda^2) \tag{11}$$

Figure 1 shows that AUC and H-measure select a close number of attributes between 27 and 29, and also have a similar density. On the other hand, EMP selects the higher number of attributes, between 29 and 30, while R-EMP selects between 28 and 30. The main result here is that since R-EMP selects a different number of attributes than EMP, different attributes are selected, having different densities, so the selected models are different for both measures.

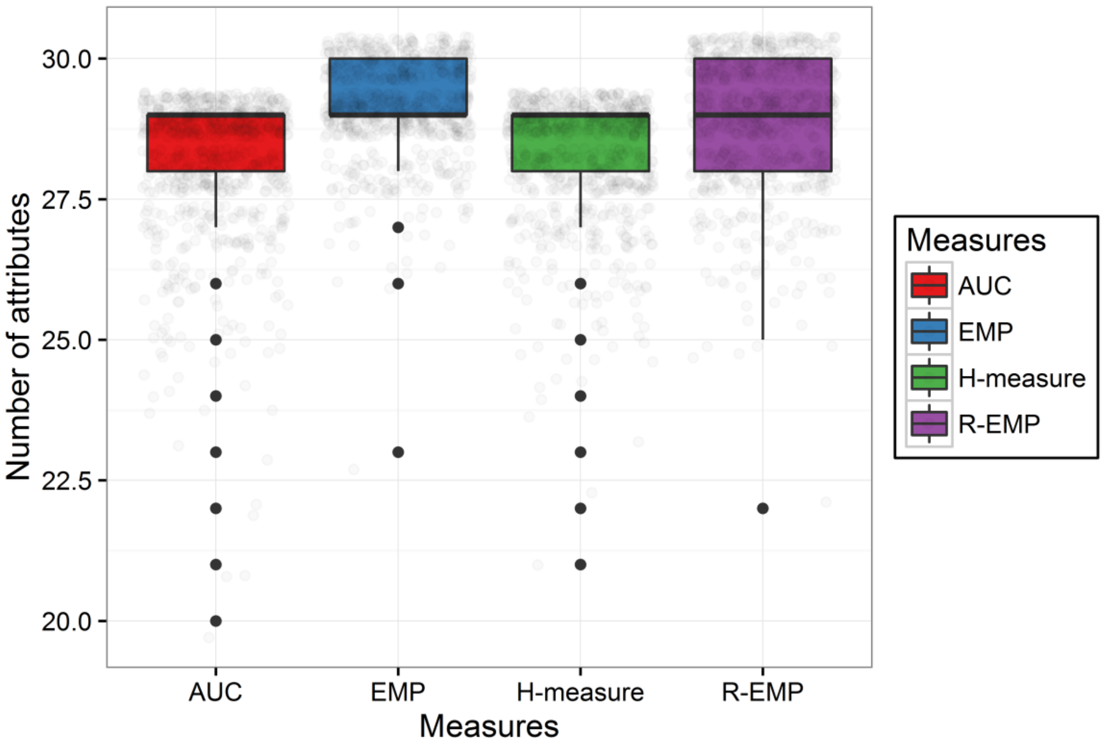

Figure 1: Synthetic Dataset - Number of selected attributes by measure.

Figure 2 shows that AUC with the out-of-time test data results has the higher variability, with profit percentages from almost 30% up to 100%. The H-measure is slightly better, being more stable, and the EMP and R-EMP are both concentrated at 100% profit with a very low dispersion. The density of profit obtained using EMP and R-EMP is more concentrated than with AUC and the H-measure, and the former measures have less dispersion.

Based on the results, it is possible to confirm that on a simulated credit scoring dataset the EMP and R-EMP measures outperform AUC and H-measure decisions. EMP and R-EMP do not select the same attributes since they select different models, but they get competitive results.

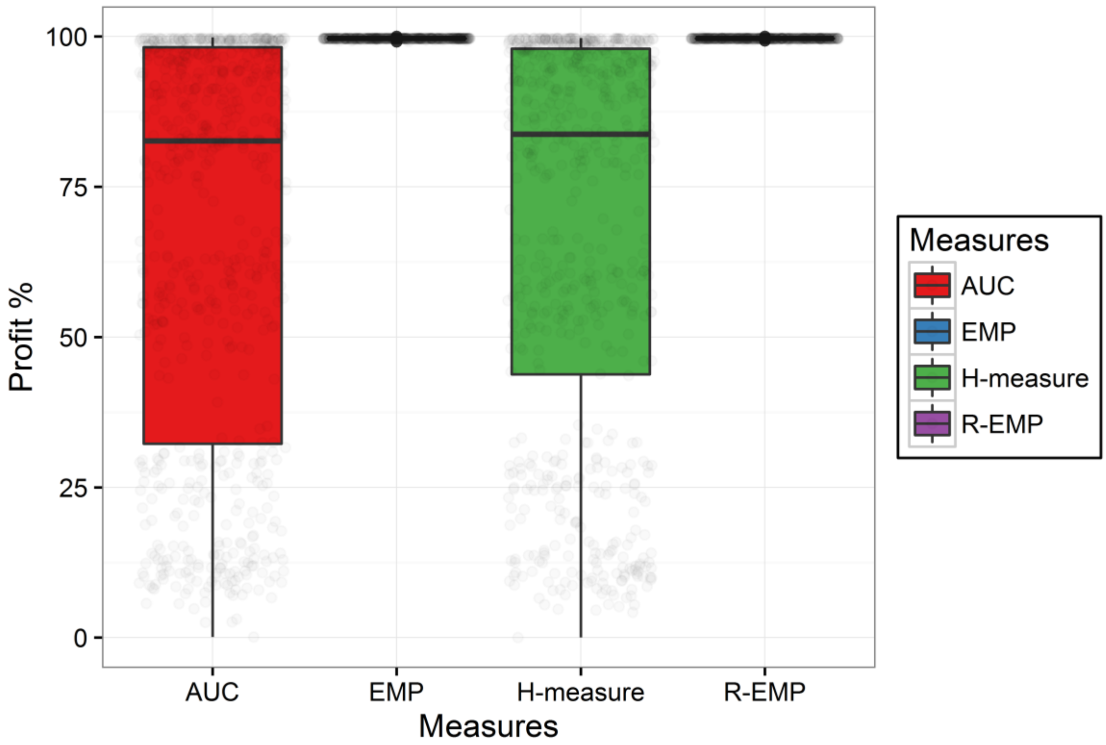

Figure 2: Synthetic Dataset - Profit out-of-time by measure.

2.6 Empirical case

The data set used for the experiment belongs to a Chilean financial institution. It is composed of a total of 16 attributes, of which most are demographic, such as the age of the customers and the economic zones in which they are located, and there is a small set of variables that describes the credit, such as its term. The size of the dataset is approximately 40,000 cases, with a default rate of 24%. The data belongs on a timeline between 1996 and 2008, but the data within each year is not balanced: 35% of the data is located in the first two years and the last four years, with the most recent, having just 13% of the total cases. These last four years are the ones that were selected to be the out-of-time sample, and the other cases were divided into training and test sets of 70% and 30% respectively.

As in the synthetic case, the perturbation (η_λ) is obtained from a normal distribution. Also the number of R-EMP replications (n_e) remains set to 100.

2.6.1 Results

Based on Figure 3 it is possible to identify that AUC and the H-measure select a close number of attributes; EMP selects a lesser number of attributes; and R-EMP selects the highest number. It is important to emphasize that as R-EMP selects a different number of attributes than EMP, the selected models are different for both measures. Table 1 shows the average number of attributes selected for each measure. The values are exactly the same as those in the former figure. R-EMP tends to select the greater number of attributes, and EMP the lesser.

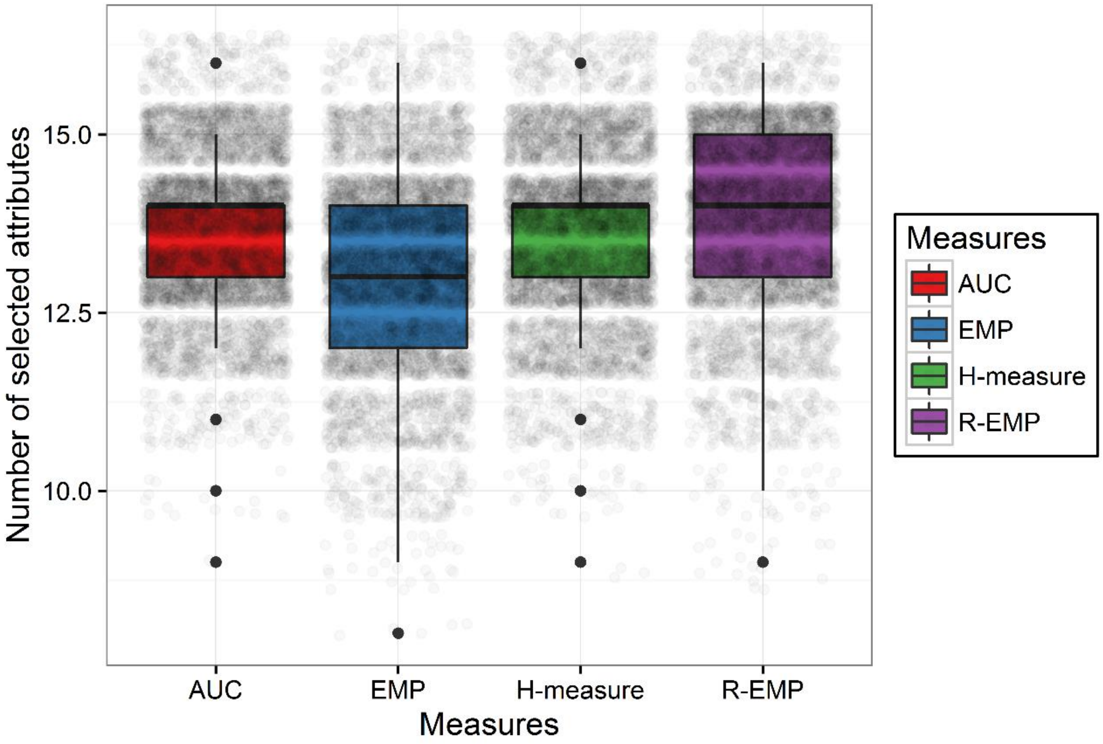

Figure 3: Empirical Dataset - Number of selected attributes by measure.

Table 1: Empirical case - Number of selected attributes ± standard deviation by measure.

AUC	H-measure	EMP	R-EMP
13.66±1.13	13.59±1.23	13.12±1.38	13.84±1.25

Using the test data on the same timeline as the training one, it was possible to obtain the results shown in Table 2. The main insight from this table is that R-EMP outperforms the other measures in all periods except for 1999.

It is also important to mention that in 2002 R-EMP was the only measure with positive profit, and from then onwards it has outperformed the other measures more clearly. When estimating the total average profit, the ranking of measures puts R-EMP at the top, followed by EMP, H-measure, and AUC. The total average standard deviation is calculated from the entire experiment. Its magnitude is explained by the experiments' including the profit of different years in which the profit magnitude varied significantly.

Table 2: Empirical Dataset - Average profit year-by-year in EUR by measure.

Year	AUC	H-measure	EMP	R-EMP
1996	17,273	18,277	18,01	18,373
1997	6,391	6,511	5,043	7,877
1998	41,981	42,487	47,271	52,202
1999	121,923	123,311	121,625	109,097
2000	143,919	144,901	143,667	145,059
2001	103,714	103,987	103,375	104,923
2002	-3,683	-2,856	-2,686	1,776
2003	59,515	59,893	68,55	78,67
2004	41,235	40,548	50,974	59,539
Total average	532,269	537,059	555,827	577,515
Total standard deviation	217,21	218,179	216,181	209,535

In order to test the predictive ability of models, we used the out-of-time test data, whose results are given in Figure 4. According to this figure it is possible to observe that the density of profit obtained using EMP and R-EMP is more concentrated than that of the AUC and H-measure, and also that the former measures have less dispersion. Average profit for this test set is just slightly different. According to Table 3, R-EMP has the higher profit (51,930 EUR), ahead of EMP which is in second place, followed by AUC and the H-measure.

Table 3: Empirical Dataset - Profit out-of-time ± standard deviation in EUR by measure.

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

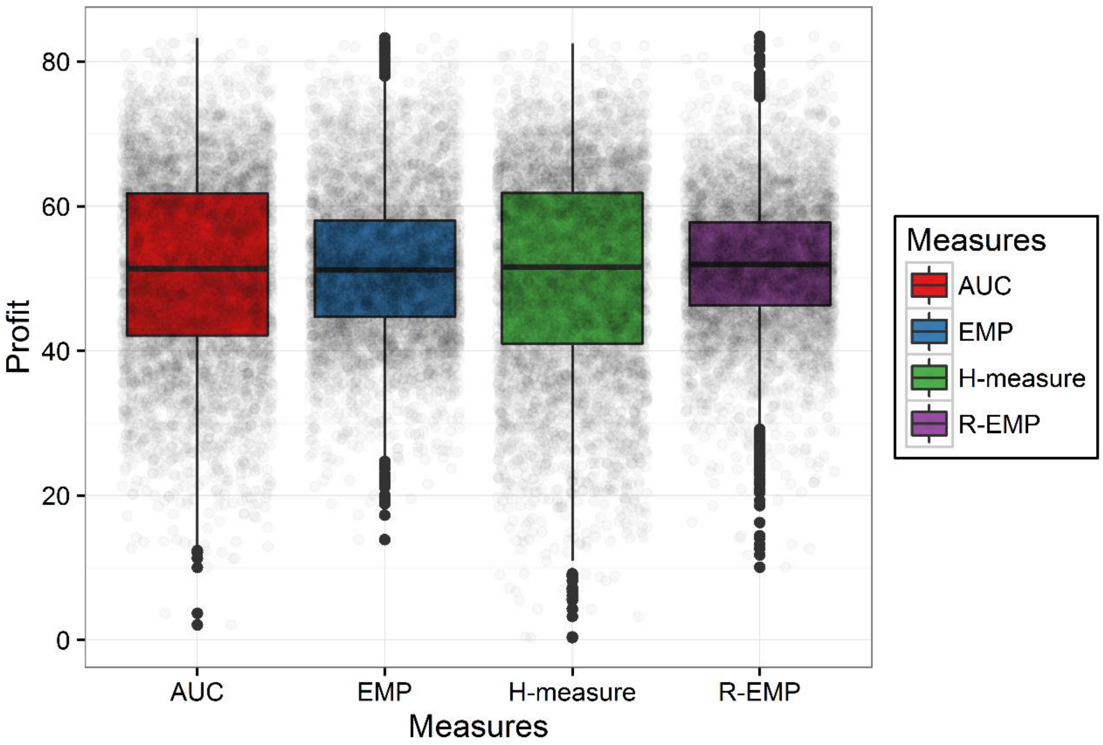

Figure 4: Empirical Dataset - Profit out-of-time in EUR thousands by measure.

Both Figures indicate that even though the difference in average profit might not be significant, R-EMP consistently delivers a higher profit model, that is also less disperse or more robust. This goal was exactly what the measure was designed to accomplish.

2.7 Conclusions

In business environments making continuously better decisions is imperative. To contribute a tool for this decison-making process, a stochastic variation of a profit driven measure is proposed in this article, which aids in evaluating Data Mining and Analytics models. R-EMP was developed and defined for the credit scoring field.

In this research it was possible to show that R-EMP effectively outperforms EMP, demonstrating that the addition of random shocks probably improves the quality of decisions, in terms of profit, in a business oriented context.

The results of the two experiments performed agree that EMP and R-EMP effectively: select different models, perform better than other measures both in synthetic and real credit data, and that the decisions made with their use have a lower standard deviation.

When performing the simulated experiment, the results for EMP and R-EMP were slightly different, but in the empirical case it was possible to confirm that R-EMP outperforms EMP, leading to the conclusion that by using external information, EMP can be improved. For credit scoring applications the R-EMP measure leads to better decisions than the main measures reported in the literature, so it is recommended to be used for selecting criteria within highly-variable business environments in order to build a scoring model and at the same time to determine the cut-off point.

2.8 Acknowledgments

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

2.9 References

Ali, Shawkat, y Kate A. Smith. «On learning algorithm selection for classification.» Applied Soft Computing (Elsevier) 6 (2006): 119-138.

Aman, Saima, Yogesh Simmhan, y Viktor K. Prasanna. «Holistic measures for evaluating prediction models in smart grids.» Transactions on Knowledge and Data Engineering, IEEE (IEEE) 27 (2015): 475-488.

Baesens, Bart, Christophe Mues, David Martens, y Jan Vanthienen. «50 years of data mining and OR: upcoming trends and challenges.» Journal of the Operational Research Society (Nature Publishing Group), 2009: S16--S23.

Baesens, Bart, Tony Van Gestel, Stijn Viaene, Maria Stepanova, Johan Suykens, y Jan Vanthienen. «Benchmarking state-of-the-art classification algorithms for credit scoring.» Journal of the Operational Research Society (Nature Publishing Group) 54 (2003): 627-635.

Baldi, Pierre, Søren Brunak, Yves Chauvin, Claus A. F. Andersen, y Henrik Nielsen. «Assessing the accuracy of prediction algorithms for classification: an overview.» Bioinformatics (Oxford Univ Press) 16 (2000): 412-424.

Bock, De and Koen, W., y Dirk Van den Poel. «An empirical evaluation of rotation-based ensemble classifiers for customer churn prediction.» Expert Systems with Applications (Elsevier) 38 (2011): 12293-12301.

Bradley, Andrew P. «The use of the area under the ROC curve in the evaluation of machine learning algorithms.» Pattern Recognition (Elsevier) 30 (1997): 1145-1159.

Brown, Christopher D., y Herbert T. Davis. «Receiver operating characteristics curves and related decision measures: A tutorial.» Chemometrics and Intelligent Laboratory Systems (Elsevier) 80 (2006): 24-38.

Clemente, M., V. Giner-Bosch, y S. San Matías. «Assessing classification methods for churn prediction by composite indicators.» Manuscript, Dept. of Applied Statistics, OR & Quality, Universitat Politècnica de València, Camino de Vera s/n 46022 (2010).

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

Correa Bahnsen, Alejandro, Djamila Aouada, y Björn Ottersten. «Example-Dependent Cost-Sensitive Logistic Regression for Credit Scoring.» International Conference on Machine Learning and Applications. 2014. 7.

Davenport, Thomas H. «Competing on analytics.» Harvard Business Review, 2006: 98-107.

Fawcett, Tom. «An introduction to ROC analysis.» Pattern Recognition Letters (Elsevier) 27 (2006): 861-874.

Fawcett, Tom. «ROC graphs with instance-varying costs.» Pattern Recognition Letters (Elsevier) 27 (2006): 882-891.

Hand, David J. «Measuring classifier performance: a coherent alternative to the area under the ROC curve.» Machine Learning (Springer) 77 (2009): 103-123.

McAfee, Andrew, y Erik Brynjolfsson. «Big data: the management revolution.» Harvard Business Review, 2012: 60-6.

McDonald, Ross A. «The mean subjective utility score, a novel metric for cost-sensitive classifier evaluation.» Pattern Recognition Letters (Elsevier) 27 (2006): 1472-1477.

Thomas, Lyn C., David B. Edelman, y Jonathan N. Crook. Credit scoring and its applications. Siam, 2002.

Verbeke, Wouter, Karel Dejaeger, David Martens, Joon Hur, y Bart Baesens. «New insights into churn prediction in the telecommunication sector: A profit driven data mining approach.» European Journal of Operational Research (Elsevier) 218 (2012): 211-229.

Verbraken, Thomas, Cristián Bravo, Richard Weber, y Bart Baesens. «Development and application of consumer credit scoring models using profit-based classification measures.» European Journal of Operational Research (Elsevier) 238 (2014): 505-513.

Verbraken, Thomas, Wouter Verbeke, y Bart Baesens. «A novel profit maximizing metric for measuring classification performance of customer churn prediction models.» Transactions on Knowledge and Data Engineering, IEEE (IEEE) 25 (2013): 961-973.

Verbraken, Thomas, Wouter Verbeke, y Bart Baesens. «Profit optimizing customer churn prediction with Bayesian network classifiers.» Intelligent Data Analysis (IOS Press) 18 (2014): 3-24.