Computational Statistics and Data Analysis 181 (2023) 107683

Contents lists available at ScienceDirect

Computational Statistics and Data Analysis

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

On the optimal binary classifier with an application

María Concepción López-Díaz [a,*], Miguel López-Díaz [b], Sergio Martínez-Fernández [c]

[a] Departamento de Matemáticas, Universidad de Oviedo, C/Federico García Lorca 18, E-33007 Oviedo, Spain
[b] Departamento de Estadística e I.O. y D.M., Universidad de Oviedo, C/Federico García Lorca 18, E-33007 Oviedo, Spain
[c] Unidad de Auditoría de Capital & Impairments, Banco Sabadell, C/Sena 12, P.I. Can Sant Joan, E-08174 Sant Cugat del Valles, Barcelona, Spain

ARTICLE INFO

Article history:
Received 21 January 2022
Received in revised form 28 September 2022
Accepted 17 December 2022
Available online 28 December 2022

Keywords:
Binary classifier
Conditional probability
Extended modelling vector
Optimal classifier
ROC and CAP curves and indexes

ABSTRACT

The alternative accumulated improvement curve stochastic order is a criterion for the comparison of the performance of classifiers that predict binary responses. An explicit optimal classifier for this criterion is obtained. That optimal classifier has the largest ROC and CAP curves and indexes, that is, it is also optimal for the criteria based on the comparison of such curves and indexes. An application of the results to the search of the best classifier to predict clients of a bank which will make a transaction in the future is developed.

1. Introduction

Statistical classification is devoted to identify which of a set of categories an individual belongs to. That is carried out by means of the values of variables or features observed at each individual. A system of classification is said to be binary if the number of categories is two. A binary classifier is a method to assign each individual to one of those categories. Very basically, it maps the data set of the observed variables at the individuals to one of such categories. Research on binary classification has experienced a significant growth in the last years because of the importance of the underlying problems.

Binary classification is usually performed by means of a bivariate random vector $X = (C_X, T_X)$, the so-called modelling vector, where C_X is said to be a classifier and T_X a target. A target T_X is a Bernoulli random variable with parameter q, where q denotes the true and unknown proportion of individuals which belong to a specific category. For instance, when a bank studies which potential clients will purchase a revolving card, q is the true and unknown proportion of clients which will acquire that card. The target takes on value 1 at an individual who will purchase the card, otherwise 0. The value of the target at each individual is unknown, and the purpose of any binary classification procedure is the estimation of that value. The estimation is carried out in terms of the value of the classifier at each individual. A classifier C_X is constructed by means of some observable features of the individuals. Each individual is assigned to one of the categories by means of a thresholding criterion on the values of C_X.

* Corresponding author.
E-mail addresses: cld@uniovi.es (M.C. López-Díaz), mld@uniovi.es (M. López-Díaz), martinezserg@bancsabadell.com (S. Martínez-Fernández).

https://doi.org/10.1016/j.csda.2022.107683

Different procedures have been created to construct classifiers using different tools, like for instance, neural networks, decision trees, random forest, Bayesian techniques, mesh methods or logistic regression, see, for instance, Breiman (2001), Wei and Chiu (2002), Hwang et al. (2004), Buckinx and Van den Poel (2005), Hung et al. (2006), Qi et al. (2009), Figini and Giudici (2010) or Güunther et al. (2014).

The comparison of the performance of the classifiers is essential for the identification of suitable classifiers to estimate a target. Quite a lot of effort has been placed on the development of comparison criteria of classifiers, see, for instance, Lloyd (1998), Lee (1999), Hand (2009), Hand and Zhou (2009), Hand (2010), Hand (2012), Hand and Anagnostopoulos (2012), Hand and Anagnostopoulos (2013) or Yousef (2013). The most common way to compare classifiers is using a summarised value for each classifier. The "identification" of a classifier with a single numeric value reduces all the probabilistic information to a unique number what entails a serious loss of information. Moreover, if classifiers are used to predict targets in different groups of the population, the rating needs to be reassessed for each group. To avoid those problems, two criteria for the comparison of the performance of classifiers were introduced in López-Díaz et al. (2017) and López-Díaz et al. (2019). Those systems have two clear advantages with respect to other methods. On one hand, they are defined by means of stochastic orders, which provide more detailed and efficient comparisons since they take into account the whole probability distributions of the classifiers and targets, instead of a single summarised value of them. On the other hand, the modification of the group size where classifiers are applied, does not entail the change of the appropriate classifier. Those orderings are the so-called accumulated improvement curve stochastic order and the alternative accumulated improvement curve stochastic order. The latter is a refinement and an improvement of the former.

An essential question in the theory of classification is the search of "the best" classifier which can be constructed with the available features of the individuals to predict a target T. Such information is modelled by a set of random variables, say $X_1, \ldots, X_n$. In this manuscript, we obtain classifiers $F(X_1, \ldots, X_n)$, such that the modelling vector $(F(X_1, \ldots, X_n), T)$ is optimal in the alternative accumulated improvement curve criterion, when we consider the set of modelling vectors $\{(H(X_1, \ldots, X_n), T) \mid H : \mathbb{R}^n \to \mathbb{R} \text{ is Borel measurable}\}$. Moreover, we will prove that those optimal classifiers are also optimal for the traditional criteria based on the ROC and CAP curves and indexes. The reader is referred, for instance, to Bamber (1975), Hsieh and Turnbull (1996), Zhou et al. (2002), Pepe (2003) or Krzanowski and Hand (2009) for an introduction to CAP and ROC curves and indexes criteria. Some manuscripts approaching the search of optimal classifiers for different optimality criteria under different frameworks are, for instance, Di Martino et al. (2013), Boughorbel et al. (2017), Zhu and Wang (2022) or Martínez-Camblor (2022).

A real application of the results of the manuscript will be developed to predict clients of a bank which will make a transaction in the future. The result shows the best performance of the method based on the aforementioned optimal classifier. The approached problem has been posed by Santander Bank. We will use the database provided by that bank for its analysis.

The structure of the paper is the following. Section 2 describes the aforementioned criteria. Section 3 contains the preliminaries of the manuscript. In Section 4, we obtain an optimal classifier in the alternative accumulated improvement curve criterion. Section 5 is devoted to prove that such a classifier is also optimal when the criteria based on the ROC and CAP curves and indexes are considered. We develop an application of the results to a real classification problem posed by a bank in Section 6. Conclusions of the manuscript are presented in Section 7.

2. The alternative accumulated improvement curve criterion

For ease of reading of the manuscript, the accumulated improvement curve and the alternative accumulated improvement curve criteria are described in this section.

Let $X = (C_X, T_X)$ be a modelling vector. The mapping $M_X : (0, 1) \to \mathbb{R}$, with

$$M_X(p) = P(T_X = 1 \mid C_X \geq F_{C_X}^{-1}(1 - p)) \text{ for all } p \in (0, 1),$$

is said to be the accumulated improvement curve of the modelling vector X ($F_{C_X}^{-1}$ stands for the quantile function of C_X).

The meaning of the mapping is the following. Given any $p \in (0, 1)$, we take the smallest group containing at least the $100p\%$ of the individuals with the largest values of the classifier C_X (that group is given by $(C_X \geq F_{C_X}^{-1}(1 - p))$), and we study the probability that the target takes on value 1 in such a group. That represents the probability of being right when we estimate the value of the target as 1 at those individuals.

Let $X = (C_X, T_X)$ and $Y = (C_Y, T_Y)$ be modelling vectors where T_X and T_Y follow Bernoulli distribution with parameter q (from an applied point of view, usually $T_X = T_Y$). The modelling vector X is said to be less than Y in the accumulated improvement curve stochastic order, if $M_X(p) \leq M_Y(p)$ for all $p \in (0, 1)$. It will be denoted by $X \preceq_M Y$. Let us analyse the meaning of $X \preceq_M Y$. Consider the smallest groups containing at least the $100p\%$ ($p \in (0, 1)$) of the individuals with the largest values of the corresponding classifiers C_X and C_Y. Basically, $X \preceq_M Y$ means that the probability of carrying out right classifications by estimating as 1 the value of the target at those individuals, is greater (at least the same) with classifier C_Y than with classifier C_X, whatever $p \in (0, 1)$.

A possible drawback of the above procedure is that when discrete classifiers are considered, the subgroups of the population given by $(C_X \geq F_{C_X}^{-1}(1 - p))$ and $(C_Y \geq F_{C_Y}^{-1}(1 - p))$ with $p \in (0, 1)$, could have different sizes. The comparison of classifiers in subgroups which are not similar in size is commonly eluded. To avoid that problem, a refinement of the above criterion was introduced in López-Díaz et al. (2019) as follows.

For any random variable W and $p \in (0,1]$, let p^W denote the real value which satisfies that

$$1 - p^W = F_W(F_W^{-1}(1-p))$$

(F_W denotes the distribution function of W). Note that $1-p^W$ is the first value greater or equal to $1-p$ which belongs to $Im(F_W)$, where Im denotes the image set.

The alternative accumulated improvement curve of a modelling vector $X = (C_X, T_X)$ is the mapping $\widetilde{M}_X : (0,1] \to \mathbb{R}$ given by

$$\widetilde{M}_X(p) = \frac{1}{p}\big(p^{C_X} P(T_X = 1 \mid F_{C_X}(C_X) > 1 - p^{C_X}) + (p - p^{C_X}) P(T_X = 1 \mid F_{C_X}(C_X) = 1 - p^{C_X})\big)$$

for all $p \in (0,1]$.

Let us clarify the meaning of this mapping. Given any $p \in (0,1]$, we aim to obtain the probability that the target assumes value 1 in the $100p\%$ of the individuals with the largest values of the classifier C_X. If there exists exactly a group with that percentage (in such a case, $p = p^{C_X}$ since $1 - p \in Im(F_{C_X})$), the above probability is calculated for those individuals. If such a subgroup does not exist (that is, $p \neq p^{C_X}$), on one hand, we take the biggest group with the largest values of the classifier whose size is lower than $100p\%$ (exactly $100p^{C_X}\%$, that group is given by $(F_{C_X}(C_X) > 1 - p^{C_X})$). On the other hand, that group is completed with the following group with the largest values of the classifier $(F_{C_X}(C_X) = 1 - p^{C_X})$, taking the appropriate weighting.

Let $X = (C_X, T_X)$ and $Y = (C_Y, T_Y)$ be modelling vectors. It will be said that X is less than Y in the alternative accumulated improvement curve stochastic order, if $\widetilde{M}_X(p) \leq \widetilde{M}_Y(p)$ for all $p \in (0,1]$. It will be denoted by $X \preceq_{\widetilde{M}} Y$. Consider the $100p\%$ ($p \in (0,1]$) of the individuals with the largest values of the corresponding classifiers C_X and C_Y. Fundamentally, $X \preceq_{\widetilde{M}} Y$ means that the probability of carrying out right classifications by estimating as 1 the value of the target at those individuals, is greater (at least the same) with classifier C_Y than with classifier C_X, whatever $p \in (0,1]$. Thus, modelling vector Y is better than X to classify individuals, whatever sizes of groups in which we compare those modelling vectors.

3. Preliminaries

A stochastic order is a pre-order relation on a set of probabilities associated with a measurable space. Basically, a stochastic order is a criterion to rank probabilities (see, for instance, Müller and Stoyan (2002), Shaked and Shanthikumar (2007) and Belzunce et al. (2016), for an introduction to the theory of stochastic orders).

Given a random vector or a random variable W, its distribution function will be denoted by F_W. When W is a random variable, the symbol F_W^{-1} will stand for the quantile function of W, that is, $F_W^{-1} : [0,1) \to \overline{\mathbb{R}}$, with $F_W^{-1}(u) = \inf\{x \in \mathbb{R} : F_W(x) \geq u\}$ for all $u \in [0,1)$. Note that $F_W^{-1}(0) = -\infty$. By agreement we define $F_W(-\infty) = 0$.

Let $(\Omega, \mathcal{F}, P)$ be a probability space. Let $W : \Omega \to \mathbb{R}^n$ be a random vector or a random variable. As usual, $\sigma(W)$ will stand for the σ-algebra generated by W on Ω. If $\mathcal{D}$ and $\mathcal{E}$ are σ-algebras on Ω, $\sigma(\mathcal{D}, \mathcal{E})$ will denote the σ-algebra generated by $\mathcal{D} \cup \mathcal{E}$. Moreover, $\sigma(W, \mathcal{D})$ will represent the σ-algebra generated by $\sigma(W)$ and $\mathcal{D}$, that is, $\sigma(\sigma(W), \mathcal{D})$.

Let $A \in \mathcal{F}$, we will denote by $P(A \mid W)$ the conditional probability of A given W, that is, $P(A \mid W) : \Omega \to \mathbb{R}$, $\sigma(W)$-measurable, with

$$P(A \cap C) = \int_C P(A \mid W)(\omega)\, dP$$

for all $C \in \sigma(W)$.

For each $w \in \mathbb{R}^n$, $P(A \mid W = w)$ will denote the conditional probability of A given that $W = w$. That is, $P(A \mid W = \cdot)$ is any Borel measurable mapping from $\mathbb{R}^n$ to $\mathbb{R}$, with

$$P(A \cap W^{-1}(B)) = \int_B P(A \mid W = w)\, dP_W$$

for all Borel sets $B \subset \mathbb{R}^n$.

When $\mathcal{D}$ is a sub-σ-algebra of $\mathcal{F}$, $E(W \mid \mathcal{D})$ will represent the conditional expectation of W given $\mathcal{D}$. That is, $E(W \mid \mathcal{D}) : \Omega \to \mathbb{R}$, $\mathcal{D}$-measurable, with

$$\int_C E(W \mid \mathcal{D})(\omega)\, dP = \int_C W(\omega)\, dP$$

for all $C \in \mathcal{D}$ (see, for instance, Ash (1972) or Billingsley (1995) for the above concepts).

The Bernoulli distribution with parameter $q \in (0,1)$ will denote by $\mathcal{B}(q)$. The uniform distribution on the interval $(0,1)$ will be indicated by $U_{(0,1)}$.

The symbol $\sim_{st}$ will mean the equality in distribution.

The following results appear in López-Díaz et al. (2019) (see Propositions 3.5 and 3.6 in such a reference). They will be applied in subsequent developments. Recall that for a random variable W and $p \in (0,1]$, p^W stands for the real value which satisfies that $1-p^W = F_W(F_W^{-1}(1-p))$, that is, $1-p^W$ is the first value greater or equal to $1-p$ which belongs to $Im(F_W)$.

Proposition 3.1. *Let W be a random variable and $p \in (0,1)$. It holds that $P(F_W(W) > 1-p^W) = p^W$.*

Proposition 3.2. *Let W be a random variable and $p \in (0,1]$. It holds that $p = p^W$ when $P(F_W(W) = 1-p^W) = 0$.*

4. On optimal classifiers in the alternative accumulated improvement curve criterion

In this section, we obtain optimal classifiers in the alternative accumulated improvement curve criterion to predict a target T when the set of classifiers is $\{H(X_1,\ldots,X_n) \mid H:\mathbb{R}^n \to \mathbb{R} \text{ is Borel measurable}\}$, where $X_1,\ldots,X_n$ stand for the observable random variables of the individuals.

In order to reach a solution to our problem, we introduce the concept of extended modelling vector.

Definition 4.1. A random vector $X = (C_X, T_X, Z_X)$ is said to be an extended modelling vector, if (C_X, T_X) is a modelling vector, $Z_X \sim_{st} U(0,1)$, and Z_X and (C_X, T_X) are independent.

Roughly speaking, the random variable Z_X will be the tool to select individuals of the group in which $(F_{C_X}(C_X) = 1-p^{C_X})$ when $p \neq p^{C_X}$, as we will check subsequently.

Let us define the alternative accumulated improvement curve of an extended modelling vector $X = (C_X, T_X, Z_X)$. As we will see, that will coincide with the alternative accumulated improvement curve of the modelling vector (C_X, T_X) (for ease of reading, and without ambiguity, we use the same name for both mappings).

Definition 4.2. Let $X = (C_X, T_X, Z_X)$ be an extended modelling vector. Let $\widetilde{\widetilde{M}}_X : (0,1] \to \mathbb{R}$ be the mapping given by

$$\begin{aligned}\widetilde{\widetilde{M}}_X(p) = \frac{1}{p}\big(&P(T_X = 1, F_{C_X}(C_X) > 1-p^{C_X}) \\ &+ P(T_X = 1, F_{C_X}(C_X) = 1-p^{C_X}, Z_X > k)\big),\end{aligned}$$

where $k = 1 - \frac{p-p^{C_X}}{P(F_{C_X}(C_X)=1-p^{C_X})}$ if $P(F_{C_X}(C_X) = 1-p^{C_X}) \neq 0$, otherwise $k = 1$. The mapping $\widetilde{\widetilde{M}}_X$ is said to be the alternative accumulated improvement curve of the extended modelling vector X.

Proposition 4.3. *Let $X = (C_X, T_X, Z_X)$ be an extended modelling vector. Then $\widetilde{\widetilde{M}}_X$ is equal to $\widetilde{M}_X$, that is, $\widetilde{\widetilde{M}}_X$ is equal to the alternative accumulated improvement curve of the modelling vector (C_X, T_X).*

Proof. Let us see that $\widetilde{\widetilde{M}}_X(p) = \widetilde{M}_X(p)$ for all $p \in (0,1]$. If $p \in (0,1]$ satisfies that $P(F_{C_X}(C_X) = 1-p^{C_X}) = 0$, by Proposition 3.2 $p = p^{C_X}$, and using Proposition 3.1, we obtain that

$$\begin{aligned}\widetilde{\widetilde{M}}_X(p) &= \frac{1}{p}\big(P(T_X = 1, F_{C_X}(C_X) > 1-p^{C_X})\big) \\ &= \frac{1}{p}\big(p^{C_X} P(T_X = 1 \mid F_{C_X}(C_X) > 1-p^{C_X})\big) \\ &= \widetilde{M}_X(p).\end{aligned}$$

If $p \in (0,1]$ satisfies that $P(F_{C_X}(C_X) = 1-p^{C_X}) > 0$, note that

$$\begin{aligned}P(T_X = 1, F_{C_X}(C_X) = 1-p^{C_X}, Z_X > k) &= P(T_X = 1, F_{C_X}(C_X) = 1-p^{C_X}) P(Z_X > k) \\ &= P(T_X = 1, F_{C_X}(C_X) = 1-p^{C_X}) \frac{p-p^{C_X}}{P(F_{C_X}(C_X) = 1-p^{C_X})} \\ &= P(T_X = 1 \mid F_{C_X}(C_X) = 1-p^{C_X})(p-p^{C_X}).\end{aligned}$$

On the other hand, by Proposition 3.1, we obtain that

$$P(T_X = 1, F_{C_X}(C_X) > 1 - p^{C_X}) = P(T_X = 1 \mid F_{C_X}(C_X) > 1 - p^{C_X}) p^{C_X},$$

which proves the result. □

Thus, we can extend the definition of the alternative accumulated improvement curve stochastic order to extended modelling vectors as follows.

Definition 4.4. Let $X = (C_X, T_X, Z_X)$ and $Y = (C_Y, T_Y, Z_Y)$ be extended modelling vectors. It will be said that X is less than Y in the alternative accumulated improvement curve stochastic order, if $\widetilde{\widetilde{M}}_X(p) \leq \widetilde{\widetilde{M}}_Y(p)$ for all $p \in (0, 1]$. It will be denoted by $X \preceq_{\widetilde{\widetilde{M}}} Y$.

Observe that $X \preceq_{\widetilde{\widetilde{M}}} Y$ when $(C_X, T_X) \preceq_{\widetilde{M}} (C_Y, T_Y)$.
From now on, $X \sim_{\widetilde{\widetilde{M}}} Y$ will mean that $X \preceq_{\widetilde{\widetilde{M}}} Y$ and $Y \preceq_{\widetilde{\widetilde{M}}} X$ hold simultaneously.
The proof of the following result appears tacitly in pages 465-466 of López-Díaz et al. (2019), and so, its proof is omitted.

Proposition 4.5. *Let $X = (C_X, T_X, Z_X)$ be an extended modelling vector. For all $p \in (0, 1]$, it holds that $\widetilde{\widetilde{M}}_X(p) = P(T_X = 1 \mid A_X^p)$, where*

$$A_X^p = \{F_{C_X}(C_X) > 1 - p^{C_X}\} \cup \{F_{C_X}(C_X) = 1 - p^{C_X}, Z_X > k\}$$

and $k = 1 - \frac{p - p^{C_X}}{P(F_{C_X}(C_X) = 1 - p^{C_X})}$ if $P(F_{C_X}(C_X) = 1 - p^{C_X}) \neq 0$, otherwise $k = 1$.

The following theorems solve the main issue of this manuscript, that is, the search of optimal classifiers to predict a target. To clarify the optimality criterion, the corresponding stochastic orders will be specified in the statements of the results.

Theorem 4.6. *Let $(X_1, X_2, \ldots, X_n, T, Z)$ be a random vector with $T \sim_{st} \mathcal{B}(q)$, $Z \sim_{st} U_{(0,1)}$, and $(X_1, X_2, \ldots, X_n, T)$ and Z independent. Let $\mathbb{X} = (X_1, X_2, \ldots, X_n)$ and $\psi_{\mathbb{X}} = P(T = 1 \mid X_1, X_2, \ldots, X_n)$. Then, $(\psi_{\mathbb{X}}, T, Z)$ is an optimal extended modelling vector in the order $\preceq_{\widetilde{\widetilde{M}}}$, when the set of classifiers is $\{H(X_1, X_2, \ldots, X_n) \mid H : \mathbb{R}^n \to \mathbb{R}$ is Borel measurable$\}$.*

Proof. Let $(\Omega, \mathcal{F}, P)$ be the probability space in which the random vector $(X_1, X_2, \ldots, X_n, T, Z)$ is defined. Let $\psi_{\mathbb{X}} = P(T = 1 \mid X_1, X_2, \ldots, X_n)$, that is, $\psi_{\mathbb{X}} : \Omega \to \mathbb{R}$, measurable with respect to $\sigma(X_1, X_2, \ldots, X_n)$, which satisfies that

$$P(T = 1 \cap C) = \int_C \psi_{\mathbb{X}}(\omega)\, dP$$

for all $C \in \sigma(X_1, X_2, \ldots, X_n)$.

In the same way, there exists $\tilde{\psi}_{\mathbb{X},Z} : \Omega \to \mathbb{R}$, measurable with respect to $\sigma(X_1, X_2, \ldots, X_n, Z)$, with

$$P(T = 1 \cap C) = \int_C \tilde{\psi}_{\mathbb{X},Z}(\omega)\, dP$$

for all $C \in \sigma(X_1, X_2, \ldots, X_n, Z)$, i.e., $\tilde{\psi}_{\mathbb{X},Z} = P(T = 1 \mid X_1, X_2, \ldots, X_n, Z)$.

Let A be the event $T = 1$. It holds that

$$E(I_A \mid \sigma(\mathbb{X})) = \psi_{\mathbb{X}} \text{ a.s. } [P]$$

since for all $B \in \sigma(\mathbb{X})$,

$$\int_B E(I_A \mid \sigma(\mathbb{X}))\, dP = \int_B I_A\, dP = P(A \cap B) = P(T = 1 \cap B) = \int_B \psi_{\mathbb{X}}(\omega)\, dP.$$

Similarly,

$$E(I_A \mid \sigma(\mathbb{X}, Z)) = \tilde{\psi}_{\mathbb{X},Z} \text{ a.s. } [P].$$

Let $\mathcal{D} = \sigma(\mathbb{X})$ and $\mathcal{E} = \sigma(Z)$. Observe that $\mathcal{E}$ is independent of $\sigma(\mathcal{D}, I_A)$. Thus,

$$E(I_A \mid \sigma(\mathcal{D}, \mathcal{E})) = E(I_A \mid \mathcal{D}) \text{ a.s. } [P].$$

On the other hand, $\sigma(\mathcal{D},\mathcal{E})=\sigma(\mathbb{X},Z)$, and so

$$E(I_A \mid \sigma(\mathcal{D},\mathcal{E}))=E(I_A \mid \sigma(\mathbb{X},Z)) \text{ a.s. } [P].$$

Therefore,

$$\psi_{\mathbb{X}}=\tilde{\psi}_{\mathbb{X},Z} \text{ a.s. } [P].$$

Now, let $\tilde{Y}=(\tilde{\psi}_{\mathbb{X},Z},T,Z)$ and $\tilde{X}=(H(\mathbb{X}),T,Z)$ be extended modelling vectors, where $H:\mathbb{R}^n\to\mathbb{R}$ is a Borel measurable mapping.

To prove the theorem, we should see that $\widetilde{\widetilde{M}}_{\tilde{X}}(p)\leq\widetilde{\widetilde{M}}_{\tilde{Y}}(p)$ for all $p\in(0,1]$. Proposition 4.5 says that $\widetilde{\widetilde{M}}_{\tilde{X}}(p)=P(T=1\mid A^p_{\tilde{X}})$ and $\widetilde{\widetilde{M}}_{\tilde{Y}}(p)=P(T=1\mid A^p_{\tilde{Y}})$ for all $p\in(0,1]$. By Propositions 3.1 and 3.2, it is immediately seen that $P(A^p_{\tilde{X}})=P(A^p_{\tilde{Y}})=p$. Then, we have that $\widetilde{\widetilde{M}}_{\tilde{X}}(p)\leq\widetilde{\widetilde{M}}_{\tilde{Y}}(p)$ if and only if $P(T=1\cap A^p_{\tilde{X}})\leq P(T=1\cap A^p_{\tilde{Y}})$.

Observe that $A^p_{\tilde{X}}$ and $A^p_{\tilde{Y}}$ belong to $\sigma(X_1,X_2,\ldots,X_n,Z)$. Therefore,

$$P(T=1\cap A^p_{\tilde{X}})=\int_{A^p_{\tilde{X}}}\tilde{\psi}_{\mathbb{X},Z}(\omega)\,dP \quad\text{and}\quad P(T=1\cap A^p_{\tilde{Y}})=\int_{A^p_{\tilde{Y}}}\tilde{\psi}_{\mathbb{X},Z}(\omega)\,dP.$$

Write the above integrals as

$$P(T=1\cap A^p_{\tilde{X}})=\int_{A^p_{\tilde{X}}\cap A^p_{\tilde{Y}}}\tilde{\psi}_{\mathbb{X},Z}(\omega)\,dP+\int_{A^p_{\tilde{X}}\cap\overline{A^p_{\tilde{Y}}}}\tilde{\psi}_{\mathbb{X},Z}(\omega)\,dP \quad\text{and}$$

$$P(T=1\cap A^p_{\tilde{Y}})=\int_{A^p_{\tilde{Y}}\cap A^p_{\tilde{X}}}\tilde{\psi}_{\mathbb{X},Z}(\omega)\,dP+\int_{A^p_{\tilde{Y}}\cap\overline{A^p_{\tilde{X}}}}\tilde{\psi}_{\mathbb{X},Z}(\omega)\,dP.$$

As a consequence, the result will be proved if we see that

$$\int_{A^p_{\tilde{X}}\cap\overline{A^p_{\tilde{Y}}}}\tilde{\psi}_{\mathbb{X},Z}(\omega)\,dP\leq\int_{A^p_{\tilde{Y}}\cap\overline{A^p_{\tilde{X}}}}\tilde{\psi}_{\mathbb{X},Z}(\omega)\,dP.$$

Since $P(A^p_{\tilde{X}})=P(A^p_{\tilde{Y}})=p$, we have that $P(A^p_{\tilde{X}}\cap\overline{A^p_{\tilde{Y}}})=P(\overline{A^p_{\tilde{X}}}\cap A^p_{\tilde{Y}})$, let us denote such a value by p'.

In accordance with the definition of $A^p_{\tilde{X}}$ and $A^p_{\tilde{Y}}$, we have that

$$\int_{A^p_{\tilde{X}}\cap\overline{A^p_{\tilde{Y}}}}\tilde{\psi}_{\mathbb{X},Z}(\omega)\,dP\leq p'F^{-1}_{\tilde{\psi}_{\mathbb{X},Z}}(1-p^{\tilde{\psi}_{\mathbb{X},Z}})=\int_{A^p_{\tilde{Y}}\cap\overline{A^p_{\tilde{X}}}}F^{-1}_{\tilde{\psi}_{\mathbb{X},Z}}(1-p^{\tilde{\psi}_{\mathbb{X},Z}})\,dP\leq\int_{A^p_{\tilde{Y}}\cap\overline{A^p_{\tilde{X}}}}\tilde{\psi}_{\mathbb{X},Z}(\omega)\,dP.$$

Note that the first inequality is because the first integral is over a subset of $\overline{A^p_{\tilde{Y}}}$ in which $F_{\tilde{\psi}_{\mathbb{X},Z}}(\tilde{\psi}_{\mathbb{X},Z})\leq 1-p^{\tilde{\psi}_{\mathbb{X},Z}}$, and so, $\tilde{\psi}_{\mathbb{X},Z}\leq F^{-1}_{\tilde{\psi}_{\mathbb{X},Z}}(1-p^{\tilde{\psi}_{\mathbb{X},Z}})$ *a.s.* $[P]$ since $1-p^{\tilde{\psi}_{\mathbb{X},Z}}\in Im(F_{\tilde{\psi}_{\mathbb{X},Z}})$. The second inequality of the formula is due to the fact that the second integral is over a subset of $A^p_{\tilde{Y}}$, and so $F_{\tilde{\psi}_{\mathbb{X},Z}}(\tilde{\psi}_{\mathbb{X},Z})\geq 1-p^{\tilde{\psi}_{\mathbb{X},Z}}$, equivalently, $F^{-1}_{\tilde{\psi}_{\mathbb{X},Z}}(1-p^{\tilde{\psi}_{\mathbb{X},Z}})\leq\tilde{\psi}_{\mathbb{X},Z}$ (see, for instance, Propositions 1 and 3 in Shorack and Wellner (1986)).

Thus, $\widetilde{\widetilde{M}}_{\tilde{X}}(p)\leq\widetilde{\widetilde{M}}_{\tilde{Y}}(p)$ for all $p\in(0,1]$. Therefore, $(\tilde{\psi}_{\mathbb{X},Z},T,Z)$ is an optimal extended modelling vector in the order $\preceq_{\widetilde{\widetilde{M}}}$, when we consider classifiers in the set $\{H(X_1,X_2,\ldots,X_n)\mid H:\mathbb{R}^n\to\mathbb{R}$ Borel measurable$\}$. Since $\tilde{\psi}_{\mathbb{X},Z}=\psi_{\mathbb{X}}$ *a.s.* $[P]$, $(\psi_{\mathbb{X}},T,Z)$ is optimal in the order $\preceq_{\widetilde{\widetilde{M}}}$. □

The above theorem permits to obtain optimal classifiers for the criterion given by the relation $\preceq_{\widetilde{M}}$.

Theorem 4.7. *Let $(X_1,X_2,\ldots,X_n,T)$ be a random vector with $T\sim_{st}\mathcal{B}(q)$. Let $\mathbb{X}=(X_1,X_2,\ldots,X_n)$ and $\psi_{\mathbb{X}}=P(T=1\mid X_1,X_2,\ldots,X_n)$. Then, $(\psi_{\mathbb{X}},T)$ is an optimal modelling vector in the order $\preceq_{\widetilde{M}}$, when the set of classifiers is $\{H(X_1,X_2,\ldots,X_n)\mid H:\mathbb{R}^n\to\mathbb{R}$ is Borel measurable$\}$.*

Proof. It follows from Theorem 4.6 and Proposition 4.3. □

Observe that Theorem 4.7 implies that there exists a "limit" in the discriminating capacity between classes of any classifier for some specific input variables $X_1, X_2, \ldots, X_n$.

Some consequences of the preceding theorems are included now.

Proposition 4.8. *Let $\{(X_1, X_2, \ldots, X_n, T, Z_n)\}_{n\in\mathbb{N}}$ be a sequence of random vectors where $T \sim_{st} \mathcal{B}(q)$, $\{Z_n\}_{n\in\mathbb{N}}$ are independent variables with distribution $U_{(0,1)}$, and Z_n and $(X_1, X_2, \ldots, X_n, T)$ are independent for any $n \in \mathbb{N}$. Let $\mathbb{X}_n = (X_1, X_2, \ldots, X_n)$ and $\psi_{\mathbb{X}_n} = P(T = 1 \mid X_1, X_2, \ldots, X_n)$ for each $n \in \mathbb{N}$. Then, $(\psi_{\mathbb{X}_n}, T, Z_n) \preceq_{\widetilde{M}} (\psi_{\mathbb{X}_{n+1}}, T, Z_{n+1})$.*

When additional features of the individuals are obtained, the preceding result means that the performance of the new optimal classifier is never worse than the performance of the former optimal classifier.

Proposition 4.9. *Let $\{(X_1, X_2, \ldots, X_n, T, Z_n)\}_{n\in\mathbb{N}}$ be a sequence of random vectors where $T \sim_{st} \mathcal{B}(q)$, $\{Z_n\}_{n\in\mathbb{N}}$ are independent variables with distribution $U_{(0,1)}$, and Z_n and $(X_1, X_2, \ldots, X_n, T)$ are independent for any $n \in \mathbb{N}$. Let $\mathbb{X}_n = (X_1, X_2, \ldots, X_n)$ and $\psi_{\mathbb{X}_n} = P(T = 1 \mid X_1, X_2, \ldots, X_n)$ for each $n \in \mathbb{N}$. If X_{n+1} is independent of $(X_1, X_2, \ldots, X_n, T)$, then it holds that $(\psi_{\mathbb{X}_n}, T, Z_n) \sim_{\widetilde{\widetilde{M}}} (\psi_{\mathbb{X}_{n+1}}, T, Z_{n+1})$.*

Proof. We have that $E(I_A \mid \sigma(X_1, X_2, \ldots, X_n)) = \psi_{\mathbb{X}_n}$ *a.s.* $[P]$ for each $n \in \mathbb{N}$, where A is the event $T = 1$. Note that

$$E(I_A \mid \sigma(X_1, X_2, \ldots, X_{n+1})) = E(I_A \mid \sigma(\sigma(X_1, X_2, \ldots, X_n), \sigma(X_{n+1}))),$$

and $\sigma(I_A, \sigma(X_1, X_2, \ldots, X_n))$ is independent of $\sigma(X_{n+1})$. Then,

$$E(I_A \mid \sigma(X_1, X_2, \ldots, X_{n+1})) = E(I_A \mid \sigma(X_1, X_2, \ldots, X_n)) \text{ a.s. } [P],$$

which implies that $\psi_{\mathbb{X}_n} = \psi_{\mathbb{X}_{n+1}}$ *a.s.* $[P]$, which leads to the result. □

When a new observable feature is independent of the target and of the other features, Proposition 4.9 ensures that the performances of the corresponding optimal classifiers are the same.

Proposition 4.10. *Let $X = (C_X, T_X)$ be a modelling vector. It holds that $(C_X, T_X) \preceq_{\widetilde{M}} (P(T_X = 1 \mid C_X), T_X)$.*

Consider a random vector $(X_1, X_2, \ldots, X_n, T)$ with $T \sim_{st} \mathcal{B}(q)$. Let $\mathbb{X} = (X_1, X_2, \ldots, X_n)$, $\psi_{\mathbb{X}} = P(T = 1 \mid X_1, X_2, \ldots, X_n)$ and $G(\mathbb{X})$ an optimal classifier. We cannot assure that $G(\mathbb{X}) = \psi_{\mathbb{X}}$ *a.s.* $[P]$. Note that $(C, T) \sim_{\widetilde{M}} (C + \lambda, T)$ for any $\lambda \in \mathbb{R}$. Therefore, $(\psi_{\mathbb{X}} + \lambda, T)$ is also an optimal element.

5. Optimality in the CAP and ROC curves and indexes criteria

In this section, we will prove that the optimal classifier based on the conditional probability $\psi_{\mathbb{X}}$ has the largest ROC and CAP curves and indexes. That is, $\psi_{\mathbb{X}}$ is also optimal when we compare classifiers by means of the ROC and CAP curves and indexes. As a consequence, the ROC and CAP curves and indexes of $\psi_{\mathbb{X}}$ can be viewed as models of reliable performance, when they are compared with the ROC and CAP indexes of other classifiers.

Let (C, T) be a modelling vector where C can take on only the values 0 and 1. The sensitivity of (C, T) is the proportion of actual positives that are correctly identified, that is, $P(C = 1 \mid T = 1)$.

Given (C, T) a modelling vector and $k \in \mathbb{R}$, let (C_k, T) be the modelling vector given by $C_k = 1$ when $C \geq k$, otherwise $C_k = 0$. For each (C_k, T) with $k \in \mathbb{R}$, consider its sensitivity, denoted by $Sens(k)$, that is, $P(C \geq k \mid T = 1)$. The CAP curve of (C, T) is defined by means of the set of points $\{(P(C \geq k), Sens(k)) \mid k \in \mathbb{R}\}$. Typically the CAP curve of (C, T) is considered when the classifier C is continuous, and so, the set $\{P(C \geq k) \mid k \in \mathbb{R}\}$ is the whole interval $(0, 1)$. For classifiers which are not continuous, different methods can be found in statistical literature to "fill" the interval $(0, 1)$, most of them based on some kind of interpolation. Our procedure to define the CAP curve will be based on the following result and on interpolation.

Proposition 5.1. *Let (C, T) be a modelling vector and let $p \in (0, 1)$. If there exists $k \in \mathbb{R}$ such that $P(C \geq k) = p$, the value of the CAP curve of (C, T) at p satisfies that $(P(C \geq k),\ Sens(k)) = (p, \widetilde{M}_{(C,T)}(p)\frac{p}{q})$.*

Proof. Let $p_k = P(C > k)$. Thus, $1 - p_k \in Im(F_C)$, which implies that $p_k = p^C$. Moreover, $p - p_k = P(C = k)$. Note that the events $(F_C(C) \geq 1 - p^C)$ and $(C \geq k)$ are equal *a.s.*, and the same holds with the events $(F_C(C) = 1 - p^C)$ and $(C = k)$. Therefore,

$$\begin{aligned}\widetilde{M}_{(C,T)}(p)\frac{p}{q} &= \frac{1}{q}(p^C P(T = 1 \mid F_C(C) > 1 - p^C) + (p - p^C)P(T = 1 \mid F_C(C) = 1 - p^C)) \\ &= \frac{1}{q}P(F_C(C) \geq 1 - p^C,\ T = 1)\end{aligned}$$

$$= \frac{1}{q} P(C \geq k, T = 1)$$
$$= P(C \geq k \mid T = 1)$$
$$= Sens(k).$$

Thus, $(P(C \geq k),\ Sens(k)) = (p, \widetilde{M}_{(C,T)}(p)\frac{p}{q})$. □

The above proposition proves that the CAP curve of a continuous classifier C is the set of points $\{(P(C \geq k), \widetilde{M}_{(C,T)}(P(C \geq k))\frac{P(C \geq k)}{q}) \mid k \in \mathbb{R}\} = \{(p, \widetilde{M}_{(C,T)}(p)\frac{p}{q}) \mid p \in (0,1)\}$.

Based on this result, we consider the CAP curve of any classifier C as the linear interpolation of the set of points

$$\{(P(C \geq k), \widetilde{M}_{(C,T)}(P(C \geq k))\frac{P(C \geq k)}{q}) \mid k \in \mathbb{R}\} \cup \{(P(C > k), \widetilde{M}_{(C,T)}(P(C > k))\frac{P(C > k)}{q}) \mid k \in \mathbb{R}\}.$$

Note that this is a natural extension of the CAP curve to the points whose abscissa is $P(C > k)$ for some $k \in \mathbb{R}$ because taking a decreasing sequence $\{k_n\}_n$ with $\lim_n k_n = k$, then

$$\lim_n (P(C \geq k_n),\ Sens(k_n)) = \lim_n (P(C \geq k_n), \widetilde{M}_{(C,T)}(P(C \geq k_n))\frac{P(C \geq k_n)}{q})$$
$$= (P(C > k),\ \widetilde{M}_{(C,T)}(P(C > k))\frac{P(C > k)}{q})$$

since the alternative accumulated improvement curve of any classifier is continuous (see Proposition 4.3 in López-Díaz et al. (2019)).

Lemma 5.2. *Let $X = (C, T)$ be a modelling vector. Let $S_C = \{p \in (0,1) \mid$ there is not $k \in \mathbb{R}$ with $p = P(C \geq k)$ or $p = P(C > k)\}$. If $p \in S_C$, there are $p_1, p_2 \in (0,1)$ with $p_1, p_2 \notin S_C$ such that $p_1 < p < p_2$ and for all $p' \in (p_1, p_2)$, it holds that $p' \in S_C$. Namely, $p_1 = p^C$ and $p_2 = p^C + P(F_C(C) = 1 - p^C)$.*

Proof. Let $p \in S_C$. Note that $p \neq 1 - F_C(k_1)$ and $p \neq 1 - F_C(k_2^-)$ for any $k_1, k_2 \in \mathbb{R}$. Hence, $1 - p^C > 1 - p$ since $1 - p \notin Im(F_C)$. Take $p_1 = p^C$. On the other hand, let $m = P(F_C(C) = 1 - p^C)$. Since $p \in S_C$, $m > 0$. Let $p_2 = p^C + m$.

Observe that $1 - p^C = F_C(F_C^{-1}(1-p))$, and so $p^C = P(C > F_C^{-1}(1-p))$, thus $p_1 \notin S_C$. Notice that

$$1 - p_2 = 1 - p^C - m = P(C \leq F_C^{-1}(1-p)) - P(C = F_C^{-1}(1-p^C))$$
$$= P(C < F_C^{-1}(1-p^C)).$$

Therefore, $p_2 = P(C \geq F_C^{-1}(1-p))$, which implies that $p_2 \notin S_C$ and $1 - p_2 < 1 - p$, and so, $p < p_2$.

Let $p' \in (p_1, p_2)$. Note that $p'^C = p^C$ and $p' \neq p^C + m$. This implies that there is not $k \in \mathbb{R}$ with $p' = P(C \geq k)$ or $p' = P(C > k)$, thus, $p' \in S_C$. □

Proposition 5.3. *Let $X = (C, T)$ be a modelling vector. Consider the curve given by the set $\Lambda = \{(p,\ \widetilde{M}_{(C,T)}(p)\frac{p}{q}) \mid p \in (0,1)\}$. Then, the CAP curve of X is the set Λ.*

Proof. Let $p \in (0,1)$. If $p \notin S_C$, the result is clear by the definition of the CAP curve. Let $p \in S_C$. Let us see that $\widetilde{M}_{(C,T)}(p)\frac{p}{q}$ is the value at p of the linear interpolation constructed with the points $(p^C, \widetilde{M}_{(C,T)}(p^C)\frac{p^C}{q})$ and $(p^C + m, \widetilde{M}_{(C,T)}(p^C + m)\frac{p^C + m}{q})$, where $m = P(F_C(C) = 1 - p^C)$. Observe that by Lemma 5.2, p^C, $p^C + m \notin S_C$, $p^C < p < p^C + m$ and for all $p' \in (p_1, p_2)$, we have that $p' \in S_C$.

Note that $p^{CC} = p^C$ and $(p^C + m)^C = p^C$. Thus,

$$\widetilde{M}_{(C,T)}(p^C)\frac{p^C}{q} = \frac{p^C}{q} P(T = 1 \mid F_C(C) > 1 - p^C) \text{ and}$$
$$\widetilde{M}_{(C,T)}(p^C + m)\frac{p^C + m}{q} = \frac{1}{q}(p^C P(T = 1 \mid F_C(C) > 1 - p^C) + m P(T = 1 \mid F_C(C) = 1 - p^C)).$$

As a consequence,

$$\widetilde{M}_{(C,T)}(p)\frac{p}{q} = (1 - \frac{p - p^C}{m})\widetilde{M}_{(C,T)}(p^C)\frac{p^C}{q} + \frac{p - p^C}{m}\widetilde{M}_{(C,T)}(p^C + m)\frac{p^C + m}{q},$$

which proves the result. □

Recall that the CAP index of (C,T) is defined as the area under the CAP curve.

Proposition 5.4. *Let $X=(C_X,T)$ and $Y=(C_Y,T)$ be modelling vectors such that $X \preceq_{\widetilde{M}} Y$. Let $CAP_X(p)$ and $CAP_Y(p)$ stand for the CAP curves of X and Y at the point p, respectively. Then,*

i) $CAP_X(p) \leq CAP_Y(p)$ for all $p \in (0,1)$,
ii) the CAP index of X is lower than or equal to the CAP index of Y.

Proof. In relation to $i)$, Proposition 5.3 ensures that for all $p \in (0,1)$, it holds that $CAP_X(p) = \frac{p}{q}\widetilde{M}_{(C_X,T)}(p)$ and $CAP_Y(p) = \frac{p}{q}\widetilde{M}_{(C_Y,T)}(p)$. The condition $X \preceq_{\widetilde{M}} Y$ leads to the conclusion.

Regarding $ii)$, recall that the CAP index of a modelling vector is the area under its CAP curve, and so the result follows from $i)$. □

Let (C,T) be a modelling vector with C taking only the values 0 and 1. The specificity of (C,T) is defined as the proportion of actual negatives that are correctly identified as such, that is, $P(C=0 \mid T=0)$.

Let (C,T) be a modelling vector, $k \in \mathbb{R}$ and let (C_k,T) be the modelling vector given by $C_k = 1$ when $C \geq k$, otherwise $C_k = 0$.

The specificity of (C_k,T) with $k \in \mathbb{R}$, denoted by $Spec(k)$, is $P(C<k \mid T=0)$. The ROC curve of the modelling vector (C,T) is defined by means of the set of points $\{(1-Spec(k), Sens(k)) \mid k \in \mathbb{R}\}$.

Note that $1-Spec(K) = \frac{P(C\geq k)-qSens(k)}{1-q} = \frac{P(C\geq k)-P(C\geq k)\widetilde{M}_{(C,T)}(P(C\geq k))}{1-q}$. That is, the ROC curve is given by the set of points

$$\{(\frac{P(C\geq k)-P(C\geq k)\widetilde{M}_{(C,T)}(P(C\geq k))}{1-q}, \widetilde{M}_{(C,T)}(P(C\geq k))\frac{P(C\geq k)}{q}) \mid k \in \mathbb{R}\}.$$

By analogy with the procedure considered in the case of the CAP curve, the ROC curve is defined as the linear interpolation of the points of the set

$$\{(\frac{P(C\geq k)-P(C\geq k)\widetilde{M}_{(C,T)}(P(C\geq k))}{1-q}, \widetilde{M}_{(C,T)}(P(C\geq k))\frac{P(C\geq k)}{q}) \mid k \in \mathbb{R}\}$$
$$\cup\{(\frac{P(C>k)-P(C>k)\widetilde{M}_{(C,T)}(P(C>k))}{1-q}, \widetilde{M}_{(C,T)}(P(C>k))\frac{P(C>k)}{q}) \mid k \in \mathbb{R}\}$$

The ROC index of (C,T) is the area under the ROC curve.

Proposition 5.5. *Let $X=(C,T)$ be a modelling vector. Consider the curve given by the set $\Gamma = \{(\frac{p-p\widetilde{M}_{(C,T)}(p)}{1-q}, \widetilde{M}_{(C,T)}(p)\frac{p}{q}) \mid p \in (0,1)\}$. Then, the ROC curve of X is the set Γ.*

Proof. Let $p \in (0,1)$. If $p \notin S_C$, the result follows from the definition of the ROC curve. Let $p \in S_C$. Recall that $p^C < p < p^C + m$ with $m = P(F_C(C) = 1-p^C)$, $p^C, p^C+m \notin S_C$ and for all $p' \in (p^C, p^C+m)$, we have that $p' \in S_C$ (see Lemma 5.2).

Note that

$$(\frac{p^C - p^C\widetilde{M}_{(C,T)}(p^C)}{1-q}, \widetilde{M}_{(C,T)}(p^C)\frac{p^C}{q})$$

is the point which is assigned to p^C by the ROC curve, and

$$(\frac{p^C+m-(p^C+m)\widetilde{M}_{(C,T)}(p^C+m)}{1-q}, \widetilde{M}_{(C,T)}(p^C+m)\frac{p^C+m}{q})$$

is the point which is assigned to p^C+m by such a curve.

Then, it is not hard to see that

$$(\frac{p-p\widetilde{M}_{(C,T)}(p)}{1-q}, \widetilde{M}_{(C,T)}(p)\frac{p}{q})$$

is the point which corresponds to p by linear interpolation, which proves the result. □

Proposition 5.6. *Let $X=(C_X,T)$ and $Y=(C_Y,T)$ be modelling vectors such that $X \preceq_{\widetilde{M}} Y$. Let $ROC_X(x)$ and $ROC_Y(x)$ stand for the ROC curves of X and Y at the point x, respectively. Then,*

i) $ROC_X(x) \leq ROC_Y(x)$ *for all* $x \in (0,1)$,
ii) *the ROC index of* X *is lower than or equal to the ROC index of* Y.

Proof. The relation $X \preceq_{\widetilde{M}} Y$ means that $\widetilde{M}_X(p) \leq \widetilde{M}_Y(p)$ for all $p \in (0,1)$. As a consequence,

$$\frac{p - p\widetilde{M}_X(p)}{1-q} \geq \frac{p - p\widetilde{M}_Y(p)}{1-q} \quad \text{and} \quad \widetilde{M}_X(p)\frac{p}{q} \leq \widetilde{M}_Y(p)\frac{p}{q}$$

for all $p \in (0,1)$. Since ROC curves are increasing, we immediately derive statement *i*) and thus *ii*). □

The following result solves the aim of this section.

Theorem 5.7. *Let* $(X_1, X_2, \ldots, X_n, T)$ *be a random vector with* $T \sim_{st} \mathcal{B}(q)$. *Let* $\mathbb{X} = (X_1, X_2, \ldots, X_n)$ *and* $\psi_{\mathbb{X}} = P(T = 1 \mid X_1, X_2, \ldots, X_n)$. *Then,* $(\psi_{\mathbb{X}}, T)$ *has the largest ROC and CAP curves and indexes, when the set of classifiers is* $\{H(X_1, X_2, \ldots, X_n) \mid H : \mathbb{R}^n \to \mathbb{R}$ *is Borel measurable*$\}$.

Proof. The result follows from Theorem 4.7, Proposition 5.4 and Proposition 5.6. □

6. An application to the identification of clients of a bank which will make a transaction

This section approaches the comparison of the performance of some classifiers for the identification of clients of a bank which will make a transaction in the future. The procedure based on the optimal classifier given by the conditional probability shows the best behaviour among those classifiers.

Our study analyses the problem posed by Santander Bank in the web page https://www.kaggle.com/c/santander-customer-transaction-prediction entitled Santander Customer Transaction Prediction.

The description of the problem is as follows (taken from such a page) "*At Santander our mission is to help people and businesses prosper. We are always looking for ways to help our customers understand their financial health and identify which products and services might help them achieve their monetary goals. Our data science team is continually challenging our machine learning algorithms, working with the global data science community to make sure we can more accurately identify new ways to solve our most common challenge, binary classification problems such as: is a customer satisfied? Will a customer buy this product? Can a customer pay this loan? In this challenge, we invite Kagglers to help us identify which customers will make a specific transaction in the future, irrespective of the amount of money transacted. The data provided for this competition has the same structure as the real data we have available to solve this problem.*"

The challenge proposed by Santander Bank tries to estimate the target variable $T =$ "the client will make a transaction in the future".

For that analysis, Santander Bank makes available the database train.csv (https://www.kaggle.com/c/santander-customer-transaction-prediction/data). This database contains 200 input variables (from Var_0 to Var_199) of 200.000 clients, as well as the value of the target variable of each client.

To approach the problem, we have generated 18 random samples of clients in the database train.csv, each of them with 100.000 observations, generating 1.800.000 observation in total.

Those observations are divided at random into two groups, a training group containing 1.500.000 observations, and a testing group with 300.000 individuals. The training group is used for the construction of the classifiers, whereas the testing group is used for the comparison of the performance of the classifiers.

For the development of the different classifiers, and because of its high computational cost, we have considered four input variables, those with the largest predictability measured by the accuracy rates provided by Santander Bank. Namely, those variables are Var_6, Var_12, Var_81 and Var_139. Among them, Var_81 was taken for the design of all the classifiers since it has the largest predictability.

The classifiers C_i are constructed taking three of the above variables, as

$$C_i = H_i(Var\_a, Var\_b, Var\_c),$$

where $a, b, c \in \{6, 12, 81, 139\}$ are different values and H_i is developed using different techniques, namely, by means of logistic regression, decision trees and conditional probabilities. We have considered the following classifiers:

Classifier C_1: constructed by means of logistic regression with the variables Var_12, Var_81 and Var_139 and stepwise selection method,

Classifier C_2: constructed by means of logistic regression with the variables Var_6, Var_81 and Var_139 and stepwise selection method,

Classifier C_3: constructed by means of logistic regression with the variables Var_6, Var_12 and Var_81 and stepwise selection method,

Classifier C_4: developed with a decision tree with the variables Var_12, Var_81 and Var_139, 2 branches per division, maximum depth of 6,

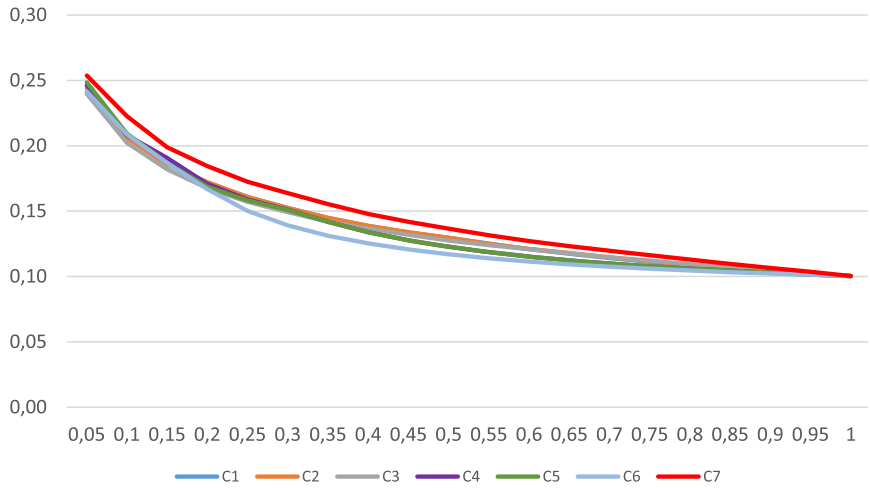

Fig. 1. Sample alternative accumulated improvement curves of the classifiers. Depicted by means of interpolation with 20 points (p) of the interval $(0, 1)$, from 0.05 to 1 with a step of 0.05. Horizontal axis for the values of p, vertical axis for the values of the sample counterpart of $\widetilde{M}_{(C_i,T)}(p), 1 \leq i \leq 7$. (For interpretation of the colours in the figure, the reader is referred to the web version of this article.)

Classifier C_5: developed with a decision tree with the variables Var_6, Var_81 and Var_139, 2 branches per division, maximum depth of 6,

Classifier C_6: developed with a decision tree with the variables Var_6, Var_12 and Var_81, 2 branches per division, maximum depth of 6,

Classifier C_7: based on the conditional probability of the target given the three variables with the largest predictability, specifically, Var_12, Var_81 and Var_139. For the development of this classifier, a mesh technique was used. The training group was divided by means of the deciles of the three variables (initially, 10^3 groups). Those groups whose sizes were less than 100 were joined to groups with at least 100 customers and with the nearest standard centroids. The classifier was calculated with the proportion of the event $T = 1$.

Propositions 4.3 and 4.5 say that given an extended modelling vector $X = (C, T, Z)$, we have that $\widetilde{M}_{(C,T)}(p) = P(T = 1 \mid A^p)$ for all $p \in (0, 1]$, where $A^p = \{F_C(C) > 1 - p^C\} \cup \{F_C(C) = 1 - p^C, Z > k\}$ and $k = 1 - \frac{p - p^C}{P(F_C(C) = 1 - p^C)}$ if $P(F_C(C) = 1 - p^C) \neq 0$, otherwise $k = 1$.

On one hand, this result says that $\widetilde{M}_{(C,T)}(p)$ is a population proportion, namely, $P(T = 1 \mid A^p)$ with T a Bernoulli random variable. On the other hand, it permits to calculate the sample version of $\widetilde{M}_{(C,T)}(p)$ as the proportion of ones of the variable target T in the sample counterpart of A^p.

The sample version of the set A^p is given by those individuals in which the value of the classifier at the empirical distribution function of C is greater than one minus the sample version of p^C (this corresponds to $\{F_C(C) > 1 - p^C\}$ in the definition of A^p). For those individuals in which the equality holds (that corresponds to $\{F_C(C) = 1 - p^C\}$), they are included or not in the sample version of A^p using the value of the random variable Z. Namely, they are included if Z is greater than the sample value of k.

That allows the inferential comparison of $\widetilde{M}_{(C_i,T)}(p)$ and $\widetilde{M}_{(C_j,T)}(p)$ for any two classifiers C_i and C_j and any value $p \in (0, 1]$, since that inferential procedure is reduced to a simple comparison of two proportions.

The graphical representation of the sample alternative accumulated improvement curves appears in Fig. 1. The whole testing group was taken for that representation. The graphic is depicted by means of interpolation with the sample values of the alternative accumulated improvement curves at the values of p from 0.05 to 1, with a step of 0.05 (20 points).

Such a representation suggests that classifier C_7 could be better than the remaining classifiers to predict customers which will make a transaction, since the corresponding curve (in red colour) seems to be greater than the curves of the other classifiers.

For the comparison of any two classifiers C_i and C_j, we will study

i) if C_i and C_j are equally efficient to estimate clients which will make a transaction, equivalently, $(C_i, T) \sim_{\widetilde{M}} (C_j, T)$, that is, $\widetilde{M}_{(C_i,T)} = \widetilde{M}_{(C_j,T)}$,

ii) if C_i is more efficient than C_j to predict clients which will make a transaction, which is the same as $(C_j, T) \preceq_{\widetilde{M}} (C_i, T)$, or $\widetilde{M}_{(C_j,T)} \leq \widetilde{M}_{(C_i,T)}$ (in the event of rejection of the same efficiency).

To compare any two classifiers C_i and C_j, the testing group was divided at random one hundred times into two groups of size 150.000. Each part was assigned at random to one classifier. This procedure tries to avoid the possible influence of a particular division in the conclusions.

Table 1
For each row (C_i) and each column (C_j), the median p-value×20 (Bonferroni correction) for the null hypothesis $H_0 : \widetilde{M}_{(C_j,T)} = \widetilde{M}_{(C_i,T)}$ is displayed.

	C_1	C_2	C_3	C_4	C_5	C_6	C_7
C_1	–	1.00	1.00	0.00	0.00	0.00	0.00
C_2	1.00	–	1.00	0.00	0.00	0.00	0.00
C_3	1.00	1.00	–	0.00	0.00	0.00	0.00
C_4	0.00	0.00	0.00	–	1.00	0.00	0.00
C_5	0.00	0.00	0.00	1.00	–	0.00	0.00
C_6	0.00	0.00	0.00	0.00	0.00	–	0.00
C_7	0.00	0.00	0.00	0.00	0.00	0.00	–

Table 2
For each row (C_i) and each column (C_j), the median p-value×20 (Bonferroni correction) for the null hypothesis $H_0 : \widetilde{M}_{(C_j,T)} \leq \widetilde{M}_{(C_i,T)}$ is displayed.

	C_1	C_2	C_3	C_4	C_5	C_6	C_7
C_1	–	1.00	1.00	1.00	1.00	1.00	0.00
C_2	1.00	–	1.00	1.00	1.00	1.00	0.00
C_3	0.76	0.99	–	0.22	0.96	1.00	0.00
C_4	0.00	0.00	0.00	–	1.00	1.00	0.00
C_5	0.00	0.00	0.00	1.00	–	1.00	0.00
C_6	0.00	0.00	0.00	0.00	0.00	–	0.00
C_7	1.00	1.00	1.00	1.00	1.00	1.00	–

For each division of the testing group, the sample alternative accumulated improvement curve of each classifier was calculated in a mesh of twenty points (20 values of p), from 0.05 to 1 with a step of 0.05, using the corresponding part of the division. For each classifier C and each of the values of p, the sample alternative accumulated improvement curve was obtained using the extended modelling vector, by means of the percentage of clients which made a transaction in the corresponding subgroup.

Regarding the same efficiency problem, we considered tests for the equality of the alternative accumulated improvement curves of both classifiers at the corresponding 20 values of p. The null hypothesis of those tests is $H_0 : \widetilde{M}_{(C_i,T)}(p) = \widetilde{M}_{(C_j,T)}(p)$. Note that this is a test for the equality of two independent proportions.

We obtained 20 p-values associated with the 20 values of p. As a representative p-value of the hypothesis $\widetilde{M}_{(C_i,T)} = \widetilde{M}_{(C_j,T)}$, we took the smallest one among the 20 p-values, that is, the p-value showing more evidence that the relation $\widetilde{M}_{(C_i,T)} = \widetilde{M}_{(C_j,T)}$ is false. A Bonferroni correction to reduce the chances of obtaining false positive results was considered. Taking the usual level of significance $\alpha = 0.05$, p-values should be compared with 0.05/20.

Since this procedure was repeated 100 hundred times (100 divisions of the testing group for any pair of classifiers), 100 p-values for the null hypothesis $H_0 : \widetilde{M}_{(C_i,T)} = \widetilde{M}_{(C_j,T)}$ were obtained. We took as a summarised p-value, the median of those values because of the robustness of that measure.

Table 1 shows the inferential conclusions. For ease of reading, it contains summarised p-values multiplied by 20 (to be compared with 0.05 because of Bonferroni correction). For each row (C_i) and each column (C_j), the table displays the summarised p-value×20 for the null hypothesis $H_0 : \widetilde{M}_{(C_j,T)} = \widetilde{M}_{(C_i,T)}$.

We conclude that classifiers C_1, C_2 and C_3 are equally efficient to predict clients which will make a transaction in the future. The same happens with classifiers C_4 and C_5. Note that the classifier based on the conditional probability (C_7) shows a performance different from any other classifiers, that also occurs with classifier C_6.

In relation to the more efficient classifier question, that is, the tests with null hypothesis $H_0 : \widetilde{M}_{(C_j,T)} \leq \widetilde{M}_{(C_i,T)}$, we followed the same steps. In this case, we used a test for the comparison $\leq$ of two independent proportions at each value of p.

Table 2 contains the results. For each row (C_i) and each column (C_j), we have the summarised p-value multiplied by 20 (to be compared with 0.05 because of Bonferroni correction) for the null hypothesis $H_0 : \widetilde{M}_{(C_j,T)} \leq \widetilde{M}_{(C_i,T)}$.
The conclusions on the performance of the classifiers can be summarised as follows

$$(C_6, T) \preceq_{\widetilde{M}} (C_5, T) \sim_{\widetilde{M}} (C_4, T) \preceq_{\widetilde{M}} (C_3, T)$$
$$\sim_{\widetilde{M}} (C_2, T) \sim_{\widetilde{M}} (C_1, T) \preceq_{\widetilde{M}} (C_7, T).$$

That corroborates the theoretical conclusions of the manuscript on the optimal classifier based on the conditional probability. The performance of such a classifier is better than the performance of any other classifiers to predict clients which will make a transaction in the future. Moreover, by the results in Section 5, we conclude that classifier C_7 has the largest ROC and CAP curves and indexes among the set of classifiers considered.

In this application, it has been seen how the optimal classifier improves the ability to distinguish between classes of the remaining classifiers. Fig. 1 provides a visual idea in terms of the area under the sample alternative accumulated improvement curve of a classifier, of its "distance" to the optimal classifier. This can serve to analyse if a classifier is far away from the optimal one, and so, there is room for improvement with the current input variables. If the classifier is close to the optimal, new input variables should be added to the model to achieve better results.

7. Conclusions

Binary classification is a relevant problem in multiple applied fields like medical diagnosis, biological classification, design of marketing strategies, credit scoring, insurance, commercial banking, etc. The comparison of the performance of classifiers is crucial since it permits the identification of appropriate classifiers, and so reaching more reliable results in the classification of individuals. Common methods for the comparison of classifiers are based on the comparison of a single global value for each of them. That entails a noteworthy loss of information. The alternative accumulated improvement curve stochastic order aims to compare classifiers. That criterion has important advantages with respect to traditional systems. Significant probabilistic information of classifiers and targets is used for the rating instead of a unique summarised value. Moreover, the modification of the sizes of the groups of the population where classifiers are applied, does not entail the change of the suitable classifier. The manuscript provides solution to a relevant problem in statistical classification theory. An explicit optimal classifier for the above comparison criterion is obtained when classifiers are constructed by means of a set of random variables. Moreover, we have proved that such an optimal classifier has the largest ROC and CAP curves and indexes among the above set of classifiers, that is, it is also optimal for the criteria based on the comparison of such curves and indexes. It is interesting to note that the results of the manuscript do not require classifiers or random variables to be continuous or discrete since they may have any kind of distribution. An estimation of clients of a bank which will make a transaction in the future is developed. The estimation given by the optimal classifier shows the best performance among the considered classifiers.

Funding

This work was supported by the Spanish Ministry of Science and Innovation [grant numbers MTM2017-83506-C2-2-P, MTM-PID2019-104486GB-I00]; and Principado de Asturias Government [grant number AYUD/2021/50897].

Acknowledgements

The authors want to thank the Reviewers and the Associate Editor for their interesting comments and suggestions, which have contributed to a substantial improvement of the manuscript.

References

Ash, R.B., 1972. Real Analysis and Probability. Probability and Mathematical Statistics, vol. 11. Academic Press, New York-London.
Bamber, D., 1975. The area above the ordinal dominance graph and the area below the receiver operating characteristic graph. J. Math. Psychol. 12, 387–415.
Belzunce, F., Martínez-Riquelme, C., Mulero, J., 2016. An Introduction to Stochastic Orders. Elsevier/Academic Press, Amsterdam.
Billingsley, P., 1995. Probability and Measure, third edition. Wiley Series in Probability and Mathematical Statistics. A Wiley-Interscience Publication. John Wiley & Sons, Inc., New York.
Boughorbel, S., Jarray, F., El-Anbari, M., 2017. Optimal classifier for imbalanced data using Matthews Correlation Coefficient metric. PLoS ONE 12 (6), e0177678.
Breiman, L., 2001. Random forests. Mach. Learn. 45, 5–32.
Buckinx, W., Van den Poel, D., 2005. Customer base analysis: partial defection of behaviourally loyal clients in a non-contractual FMCG retail setting. Eur. J. Oper. Res. 164, 252–268.
Di Martino, M., Hernández, G., Fiori, M., Fernández, A., 2013. A new framework for optimal classifier design. Pattern Recognit. 46, 2249–2255.
Figini, S., Giudici, P., 2010. Bayesian churn models. Adv. Appl. Stat. Sci. 1, 285–310.
Güunther, C.C., Tvete, I.F., Aas, K., Sandnes, G.I., Borgan, O., 2014. Modelling and predicting customer churn from an insurance company. Scand. Actuar. J. 1, 58–71.
Hand, D.J., 2009. Measuring classifier performance: a coherent alternative to the area under the ROC curve. Mach. Learn. 77, 103–123.
Hand, D.J., 2010. Evaluating diagnostic tests: the area under the ROC curve and the balance of errors. Stat. Med. 29, 1502–1510.
Hand, D.J., 2012. Assessing the performance of classification methods. Int. Stat. Rev. 80, 400–414.
Hand, D.J., Anagnostopoulos, C., 2012. A better Beta for the H measure of classification performance. Pattern Recognit. Lett. 40, 41–46.
Hand, D.J., Anagnostopoulos, C., 2013. When is the area under the receiver characteristic curve an appropriate measure of classifier performance? Pattern Recognit. 80, 400–414.
Hand, D.J., Zhou, F., 2009. Evaluating models for classifying customers in retail baking collections. J. Oper. Res. Soc. 61, 1540–1547.
Hsieh, F., Turnbull, B.W., 1996. Nonparametric and semiparametric estimation of the receiver operating characteristic curve. Ann. Stat. 24, 25–40.
Hung, S., Yen, D.C., Wang, H., 2006. Applying data mining to telecom churn management. Expert Syst. Appl. 31, 515–524.
Hwang, H., Jung, T., Su, E., 2004. An LTV model and customer segmentation based on customer value: a case study on the wireless telecommunication industry. Expert Syst. Appl. 26, 181–188.
Krzanowski, W.J., Hand, D.J., 2009. ROC Curves for Continuous Data. Chapman & Hall/CRC, Boca Ratón.
Lee, W.C., 1999. Probabilistic analysis of global performances of diagnostic tests: interpreting the Lorenz curve-based summary measures. Stat. Med. 18, 455–471.
Lloyd, C.J., 1998. Using smoothed receiver operating characteristic curves to summarize and compare diagnostic systems. J. Am. Stat. Assoc. 93, 1356–1364.

López-Díaz, M.C., López-Díaz, M., Martínez-Férnandez, S., 2017. A stochastic comparison of customer classifiers with an application to customer attrition in commercial banking. Scand. Actuar. J. 7, 606–627.
López-Díaz, M.C., López-Díaz, M., Martínez-Férnandez, S., 2019. A criterion for the comparison of binary classifiers based on a stochastic dominance with an application to the sale of home insurances. Scand. Actuar. J. 6, 453–477.
Martínez-Camblor, P., 2022. The fundamental role of density functions in the binary classification problem. J. Stat. Comput. Simul. 92, 2846–2861.
Müller, A., Stoyan, D., 2002. Comparison Methods for Stochastic Models and Risks. John Wiley & Sons, Chichester.
Pepe, M.S., 2003. The Statistical Evaluation of Medical Tests for Classification and Prediction. Oxford University Press, Oxford.
Qi, J., Zhang, L., Liu, Y., Li, L., Zhou, Y., Shen, Y., Liang, L., Li, H., 2009. ADTreesLogit model for customer churn prediction. Ann. Oper. Res. 168, 247–265.
Shaked, M., Shanthikumar, J.G., 2007. Stochastic Orders. Springer, New York.
Shorack, G.R., Wellner, J.A., 1986. Empirical Processes with Applications to Statistics. Wiley Series in Probability and Mathematical Statistics. John Wiley & Sons, Inc., New York.
Wei, C., Chiu, I., 2002. Turning telecommunications call details to churn prediction: a data mining approach. Expert Syst. Appl. 23, 103–112.
Yousef, W.A., 2013. Assessing classifiers in terms of the partial area under the ROC curve. Comput. Stat. Data Anal. 64, 51–70.
Zhou, X.H., McClish, D.K., Obuchowski, N.A., 2002. Statistical Methods in Diagnostic Medicine. Wiley Series in Probability and Statistics. Wiley.
Zhu, Y., Wang, M.C., 2022. Obtaining optimal cutoff values for tree classifiers using multiple biomarkers. Biometrics 78, 128–140.