Survival model assessment via polarization-based discrimination indexes: an application to churn risk

Chiara Gigliarano[1], Silvia Figini[1], Pietro Muliere[1]

[a]*Department of Economics, Università dell'Insubria, via Monte Generoso 71, 21100 Varese, Italy.Email: chiara.gigliarano@uninsubria.it*
[b]*Department of Political and Social Sciences, Università di Pavia, Strada Nuova 65, 27100 Pavia, Italy. Email: silvia.figini@unipv.it*
[c]*Department of Decision Sciences, Università L. Bocconi, via Roentgen 1, 20136, Milano, Italy. Email: pietro.muliere@unibocconi.it*

Abstract

In this paper we propose a novel class of indices derived from polarization measures useful to compare survival predictive models. More precisely, the main aim of this paper is to review the most relevant performance measures aimed to assess temporal dependent models and to underline how polarization measures can be useful for model comparison. Our idea could be of interest for a wide range of applications where temporal dependent models can be employed. Finally, after a theoretical discussion we test our proposal on real data related to churn risk.

Keywords: ROC curve, survival models, polarisation, model selection, churn risk

1. Introduction

A large number of applications fall into the framework of predictive classification. In such problems the aim is to construct the best predictive model using a set of data. One of the important issues in performance evaluation is that of which criterion to choose to measure classifier performance. In the literature a wide range of performance indicators are available, such as the misclassification (or error) rate, the Kolmogorov-Smirnov (KS) statistic, likelihood ratios, the area under the ROC curve (or, equivalently, the Gini coefficient), pairs of measures such as specificity and sensitivity or precision and recall, measures of accuracy of probability estimates such as Brier or log score, and many others (see, for example, Flach, 2003; Hand, 1997; Dodd and Pepe, 2003).

The most popular index recognized in the literature is the AUC (Area Under the ROC Curve). The AUC is a single number derived from a classification rule, so that comparisons of classification rules can be made in a straightforward way. It is objective, requiring no choices of parameter values to be made by the user, so that different researchers would obtain the same results from the same data. However, it also has some well-known weaknesses (see Gigliarano et al. 2014). For example, if ROC curves cross then it is possible that one curve has a larger AUC (and so is apparently better) even though the alternative may show superior performance over almost the entire range of values of the classification threshold (defined below) for which the curve will be used. In many practical applications, it is likely that the ROC curves being compared will cross. One reason for this is that comparisons are likely to be between classifiers with similar performance. As discussed in Hand (2009), in many situations, an empirical process of classifier improvement is undertaken, adjusting the classifier a small step at a time so as to gradually improve the KS, AUC, or whatever performance measure is being used. The result is a series of comparisons between similar classifiers, which are therefore likely to have similar ROC curves. When curves are similar, it is unlikely that one will dominate another - unlikely that one will have a superior sensitivity for all choices of specificity.

The risks of comparing classifiers on the basis of simple summary measures which fail to take account of the potential for ROC curves to cross are well-known. Having noted this weakness of the AUC as a measure of classifier performance, in Gigliarano et al. (2014) are presented alternative measures to compare predictive models characterized by a binary dependent variable. Starting from the contribution of Gigliarano et al. (2014), this paper extend model comparison for temporal dependent models proposing a novel class of polarization measures. The paper is structured as follows: Section 2.1 describes in general the problem of model assessment and performance indicators for survival models; Section 2.2 briefly reviews the polarization measures; Section 3 shows the novel class of polarization based discrimination indexes useful to compare predictive models for survival data; Section 4 discusses the results from an illustrative application to churn risk and, finally, Section 5 underlines the conclusions and further ideas of research.

2. Background

2.1. Measuring the predictive accuracy of survival models

Let us consider a binary response setting, where D is the status indicator, with $D = 1$ if the subject developed an event of interest within a given period of time (such as, e.g., having a default) and or $D = 0$ otherwise .
Let the dependence between a covariates vector X and the condition of being bad be specified through a given binary model (i.e. logistic regression); $z(X)$ is the linear predictor, where greater values of $z(X)$ indicate grater predicted probability of being bad (i.e. having a default).
The model predictor can be thought as a test, any value v of $z(X)$ defines a prediction rule to classify a subject as bad if $z(X) > v$, or as good if $z(X) \leq v$. The probabilities of correct classification conditional on the status, i.e. sensitivity $P(z(X) > v|D = 1)$ and specificity $P(z(X) \leq v|D = 0)$ at v, represent the predictive accuracy of the rule.
The ROC curve is the plot of sensitivity against the complement to one of the specificity (varying v), and represents the accuracy of the whole set of rules; see Figure **??**. The area under the ROC curve (AUC) is a summary accuracy index, which is equal to:

$$AUC = P(z(X_i) > z(X_j)|D_i = 1 \wedge D_j = 0) \tag{1}$$

where (i, j) denote any pair of randomly chosen subjects, and the symbol $\wedge$ represents the logical conjunction "and". See, among others, Krzanowski and Hand (2009), Pencina and D'Agostino (2004, 2008).
Let us now consider a survival time setting, where (T, D) denote the random variables of concern. Let $T = min(T_S, T_C)$ be the observed time, where T_S is the failure time and T_C is the right censoring time. Typically, in practical applications a discrete time scale is considered (days, weeks, months etc.). Therefore T can be considered as a discrete random variable, that assume values $t_{(1)} < \ldots < t_{(K)}$, where $t_{(K)}$ is the end of the study period. D is the status indicator, $D = 0$ if $T = T_C$ or $D = 1$ if $T = T_S$.
Let the dependence between a vector of fixed covariates X and T_S be specified through a given survival model. In the absence of censoring, for any pair of subjects (i, j), labelled such as $T_{Si} < T_{Sj}$, a desirable property of the model is to have the same ranking between the corresponding predicted times. In such a case, the pair is said to be concordant and the probability of concordance is a natural indicator of model discrimination.

In the light of this idea, the popular Harrell C index of discrimination evaluates the probability of concordance addressing for the presence of right censored times (see Harrell et al.,1982). The Harrell C concordance index is related to the area under the ROC curve (see e.g. Heagerty and Zheng 2005) and the relative interpretation as a misclassification probability appears as particularly attractive.
The starting point to define whether the subjects (i, j) are concordant is that they are comparable, meaning that their survival times T_{Si} and T_{Sj} can be ranked. The ranking between T_{Si} and T_{Sj} can be determined from the observed data $(T_i; D_i)$ and $(T_j; D_j)$ if and only if the minimum between T_i and T_j is an event time, i.e. if $T_i < T_j$ and $D_i = 1$ or $T_i > T_j$ and $D_j = 1$. Thus, assuming the subjects are labelled such as $T_i < T_j$, they are comparable if and only if $D_i = 1$, arguing that $T_{Si} < T_{Sj}$. The probability that (i, j) are comparable is $\pi_{comp} = P(T_i < T_j \wedge D_i = 1)$ where the presence of censored times is considered as a population characteristic and is included in the definition of π_{comp}.
Considering a proportional hazard model, i.e. $log(h(t_{(k)}|X) = z(X)+log(h_0(t_{(k)})$ for $k = 1, \dots, K$, where $h(t_{(k)}|X)$ is the hazard of an individual having covariate equal to X, $h_0(t_{(k)})$ is the baseline hazard and $z(X)$ the model linear predictor, C can be defined without the need of predicting individual survival times. The ranking between the predicted times is obtained resorting to the linear predictors, since the inequality $z(X_i) > z(X_j)$ is equivalent to $S(t_{(k)|X_i}) < S(t_{(k)|X_j})$ for any $t_{(k)}$, and this implies a smaller predicted time for the subject i.
Thus, in the absence of censoring, any pair (i, j), labelled such that $T_{si} < T_{sj}$, is concordant if and only if $z(X_i) > z(X_j)$. In the presence of censoring, the C index is the probability that (i, j) are concordant given that they are comparable:

$$C = Pr\{z(\mathbf{X}_i) > z(\mathbf{X}_j)|T_i < T_j \wedge D_i = 1\} = \frac{\pi_{conc}}{\pi_{comp}} \tag{2}$$

where the numerator is the probability of concordance $\pi_{conc} = P((z(X_i) > z(X_j)) \wedge (T_i < T_j) \wedge (D_i = 1))$. In analogy with AUC, Harrell C index ranges from 0.5 to 1 and has the same interpretation.
An alternative definition of C index was given (Klawonn et al., 2011) for a survival model which specifies the relation between a vector of possibly time dependent covariates $X(t)$ and T_s, originating a one to one correspondence between predicted times and predicted survival probabilities at any point t.

This is guaranteed for proportional hazard models, and for a more general class of transformation models. It has to be pointed out, as in the absence of a specification about how to predict individual survival times, the ranking between $S(t|X_i(t))$ and $S(t|X_j(t))$ is possible if and only if the survival functions are separated over the whole time. Thus, i has lower predicted time than j if and only if $S(t|X_i(t)) < S(t|X_j(t))$ for any t. Therefore, the probability of concordance becomes $\pi_{conc} = P(S(t|X_i(t)) < S(t|X_j(t)) \wedge T_i < T_j|D_i = 1)$ and the C index becomes

$$C = \frac{\pi_{conc}}{\pi_{comp}} = P(S(t|X_i(t)) < S(t|X_j(t))|T_i < T_j \wedge D_i = 1) \tag{3}$$

Assuming, instead, of dealing with survival models in the more general framework where the one to one correspondence does not necessarily hold (see, e.g., Antolini et al., 2005) . A basic notation is introduced referring to discrete times. Let $D(t_{(k)})$ be the status at time $t_{(k)}$; more precisely, $D(t_{(k)}) = 1$ if the subject experiences event at $t_{(k)}$, $D(t_{(k)}) = 0$ if does not experience event until $t_{(k)}$ and $D(t_{(k)})$ is not defined if subject is not at risk at $t_{(k)}$. To evaluate the model ability to discriminate among subjects at risk at t_k, between churned and non churned till t_k, the predicted survival probability $S(t_{(k}|X(t))$ is the natural quantity to consider. In this case, the probabilities of correct classification conditional on the status at $t_{(k)}$, i.e. sensitivity $Pr\{S(t_{(k)}|X(t)) \leq v|D_j(t_{(k)}) = 1\}$ and dynamic specificity: $Pr\{S(t_{(k)}|X(t)) > v|D_j(t_{(k)}) = 0\}$ represent the accuracy of the rule, and the area $AUC(t_{(k)})$ under the corresponding ROC curve:

$$AUC(t_{(k)}) = Pr\{S(t_{(k)}|\mathbf{X}_i(\mathbf{t})) < S(t_{(k)}|\mathbf{X}_j(\mathbf{t}))|D_i(t_{(k)}) = 1 \wedge D_j(t_{(k)}) = 0\} \tag{4}$$

represents the ability of discrimination of $S(t_{(k)}|\mathbf{X}(\mathbf{t}))$. To summarize the ability of $S(t_{(k)}|\mathbf{X}(\mathbf{t}))$ to discriminate between $D(t_{(k)}) = 1$ and $D(t_{(k)}) = 0$ over the whole follow-up a weighted average of $AUC(t_{(k)})$ over time can be written as follows:

$$\begin{aligned} C^{td} &= \frac{\sum_{k=0}^{K} AUC(t_{(k)}) \cdot w(t_{(k)})}{\sum_{k=0}^{K} w(t_{(k)})} \\ &= P(S(T_i|X_i(t)) < S(T_i|X_j(t))|T_i < T_j \& D_i = 1) \end{aligned}$$

where $w(t_{(k)}) = Pr\{D_i(t_{(k)}) = 1 \,\&\, D_j(t_{(k)}) = 0\}$ is the probability of comparable pairs. It is worth of note as $w(t_{(k)})$ is the same weighting system

adopted in the derivation of the Harrell's C as weighted average over time of time dependent accuracy.
Let us observe as the difference between C and C^{td} is that in the latter the predicted survival is evaluated in T_i, which indeed depends on (i, j), instead on a fixed t. C^{td} is the probability that (i, j) are concordant given they are comparable. More specifically, given a comparable pair (i, j), labelled such as $T_i < T_j$ and $D_i = 1$, a desirable property of the model is that the predicted survival probability, at the time where the subject i developed the event, is greater for the subject j who actually is still free from the event. We refer this condition as 'td concordance'. If the one to one correspondence holds, C^{td} reduces to the Harrell's C.

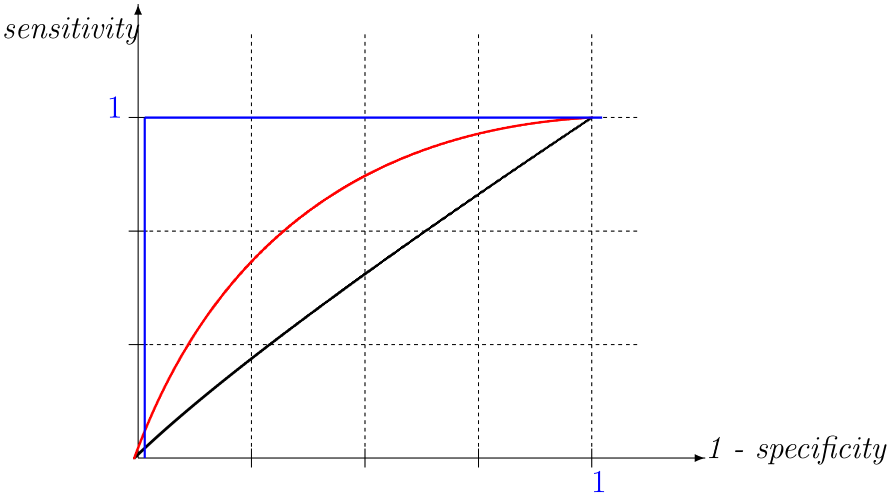


Figure 1: Example of ROC curve

Summing up, the well-known Harrell's C measure for assessing the predictive accuracy of survival models is a measure of concordance between model predictions and status indicator. This measure is based only on rankings and not on the differences in survival times. In the next sections we rather propose to assess the discriminative ability by looking also at differences in survival times, recurring, in particular, to indices of polarization.

2.2. Measuring polarization

Polarization is commonly connected with the division of a society into groups as possible cause of social conflicts. It is measured by quantifying and comparing socio-economic disparity, not only in terms of differences among individuals (as inequality measurement does) but also in terms of differences among population groups.
Based on the seminal papers by Esteban and Ray (1994) and Wolfson (1994), two different strands of the literature have emerged, underlining different aspects of the phenomenon of income polarization. The first string, originating from Esteban and Ray (1994), focuses on the rise of separated income groups: polarization increases if the groups become more homogeneous internally, more separated from each other and more equal in size. This approach is followed, among others, by Gradin (2000), D'Ambrosio (2001), Duclos et al. (2004).

The second strand, going back to Wolfson (1994), describes the decline of the middle class, measuring how the center of the income distribution is emptied. This approach is often referred to as "bi-polarization" and assumes the presence of only two groups which are divided by the median income.

In general most of the existing contributions to polarization measurement mainly focus on income, being income the characteristic both that forms groups and that distinguishes individuals and groups from each other. In contrast, Zhang and Kanbur (2001) and Gradin (2000) derive polarization measures that allow for other characteristics than income to form groups and for income to measure polarization among those exogenously formed groups.

Let us first introduce some notation. Let $\mathbf{x} = (x_1, \ldots, x_n)$ be a vector of incomes for a given population of size n and with mean $\bar{x}$. Now suppose that the population can somehow be separated into k groups and let $\mathbf{x}_j = \left(x_{j1}, \ldots, x_{jn_j}\right)$ be the income vector of individuals belonging to group j, $\bar{x}_j$ be the corresponding mean income and $\frac{n_j}{n}$ the population share of group j, $j = 1, \ldots, g$.

Esteban and Ray (1994) set up a system of four axioms and derive the following polarization measure

$$P^{ER}(\mathbf{x}) = K \sum_{j=1}^{g} \sum_{l=1}^{g} \left(\frac{n_j}{n}\right)^{1+\alpha} \cdot \frac{n_l}{n} \cdot \left|\bar{x}_j - \bar{x}_l\right|, \tag{5}$$

where $K > 0$ is a normalizing constant and $\alpha \in (1; \alpha^*], \alpha^* \approx 1.6$ is the so called polarization sensitivity. Note that the inclusion of α makes the difference between polarization and inequality in this approach, since for $\alpha = 0$ and $K = \frac{1}{2 \cdot \bar{x} \cdot n^2}$ expression (**??**) gets back to the Gini index.[1]

Since $P^{ER}(\mathbf{x})$ does not incorporate any within-group heterogeneity, Esteban et al. (2007) proposed an extension of the original index that is function also of within-group Gini index. Esteban et al (2007) recommend that their measure should be applied after the original vector of incomes has been grouped by a statistical approach that minimizes within-group dispersion. Gradin (2000) proposed the same correction term to the original index, though in his approach groups can be defined according to variables different from income, being – at least to our knowledge – the first measure of "socioeconomic" polarization.

The second strand of the literature on income polarization is based on the polarization measure of Wolfson (1994). Additional to the existing notation, let m denote the median of the income vector $\mathbf{x}$ and $L(z)$ be the value of the Lorenz curve at the z-quantile of $\mathbf{x}$. Wolfson (1994) proposes a polarization curve analogous to the Lorenz curve for inequality measurement and defines his polarization measure to be

$$P^W(\mathbf{x}) = \frac{2\bar{x}}{m} \cdot (1 - 2 \cdot L(0.5) - G) = \frac{2\bar{x}}{m} \cdot (G_B - G_W), \tag{6}$$

where G, G_W and G_B are, respectively, the total, the within- and the between-group Gini indices.

Therefore, the Wolfson polarization measure is a normalized function of the difference between the inequality between groups G_B and the inequality within groups G_W.

Summing up, the Wolfson measure can be used in case of two non-overlapping groups, while the Esteban and Ray index is suitable for a generic number of groups, also overlapping.

3. Methodology: a polarization-based discrimination index

As discussed above, predictive accuracy of survival models are usually assessed using Harrell's C measure, which is a measure of concordance be-

[1] Being precisely, for $\alpha = 0$ we would obtain the Gini index for classified data. This is because Esteban and Ray (1994) argue that an individual feels perfect identification with each member of his or her own subgroup, regardless of possible income differences. Therefore, each income of a group can be replaced by the respective group mean.

tween model predictions and status indicator. This measure is based only on rankings and not on the differences in survival times.

Here we propose to assess the discriminative ability of survival models by looking also at differences in survival times, in particular at the polarization in the predicted survival times, based on the overlap of the groups. In this way we are able to capture a different set of information.

Following the Esteban and Ray (1994) approach, discussed above, the distribution of a random variable, such as survival time, is said to be polarized in case of rise of groups well separated and with low disparity inside. Polarization is, therefore, characterized by three features: high homogeneity within each group, high heterogeneity across groups, small number of significantly sized groups.

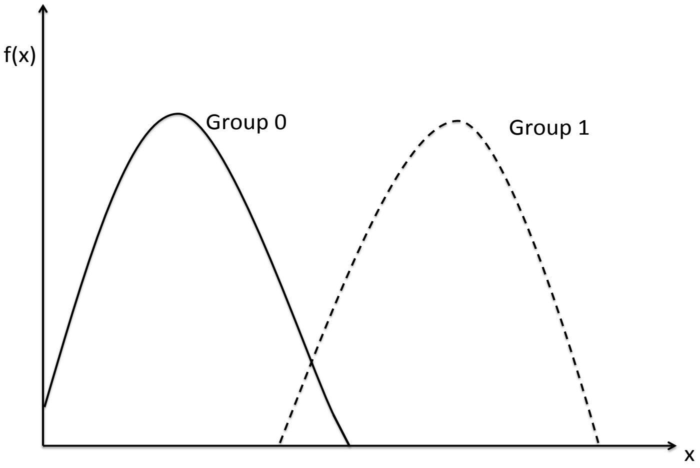
Absence of polarization in the predicted failure times reveal a weak discriminatory power, while strong polarization suggest that the survival model has high predictive accuracy.

If the distributions of the predicted survival times (or predicted probabilities of survival) do not overlap, it means that the model perfectly discriminate the individuals (see, for example, Figure **??**). The higher the degree of overlap, the worse is the survival model. The case of perfect discrimination would correspond to the case of maximum bi-polarization (therefore, zero overlap).

We propose to apply polarization indices in survival analysis to measure concentration in the predicted probabilities of survival within groups of subjects and to detect differences in heterogeneity between the predicted survival probabilities of two groups of subjects. Differences in concentration between groups may suggest the presence of a differential covariates effect, thus providing information of the discriminative power of a survival model.

Analogously to Section 2, let T denote the time to some failure of interest, with possible right-censoring: $T = min(T_S, T_C)$. We follow individuals for a given period of time, until $t_{(K)}$. We consider a statistical model that provides the predicted survival time for each individual. At each given point in time $t_{(1)} \leq ... \leq t_{(k)} \leq ... \leq t_{(K)}$, we have two categories of subjects: those who developed the event before $t_{(k)}$ (*events*) and those who did not (*non-events*). For any $t_{(k)}$, $D(t_{(k)})$ denotes the status at (or until) $t_{(k)}$, where $D(t_{(k)}) = 1$ if the individual experiences the event at $t_{(k)}$ and $D(t_{(k)}) = 0$ has not experienced the event until $t_{(k)}$. $D(t_{(k)})$ is not defined if the individual is not at risk at $t_{(k)}$. Therefore, for each time $t_{(k)}$, we have the group $G_1(t_{(k)})$ of the subject who experience the event at time $t_{(k)}$ (or the group of bad items) and the group $G_0(t_{(k)})$ of the subject who have not experienced the event

Figure 2: Example of distribution with a high level of polarization



until $t_{(k)}$ (or the group of good items).
Let $T_1, T_2, ..., T_n$ be the actual survival times and $Z_1(t), Z_2(t), ..., Z_n(t)$ the corresponding predicted probabilities of survival at t for a sample of size n. Note that in the most common survival models we can interchange the predicted survival times and the predicted probabilities of survival.

Following the Esteban and Ray's approach as defined in expression (**??**), we define a *polarization-based measure of the discriminatory power of the survival model* at time $t(k)$ as

$$P^{ER}(t_{(k)}) = K\left(\left(\frac{n_0}{n}\right)^{1+\alpha}\frac{n_1}{n}+\left(\frac{n_1}{n}\right)^{1+\alpha}\frac{n_0}{n}\right)\left(\left|\bar{Z}_0(t_{(k)})-\bar{Z}_1(t_{(k)})\right|\right), \quad (7)$$

where n_j and $\bar{Z}_0(t_{(k)})$ are, respectively, the size and the average predicted probability of group $G_j(t_{(k)})$ for $j = 0, 1$. Moreover, $K > 0$ is a normalization constant and $\alpha \in (1; 1.6]$ measures polarization sensitivity.
A continuous diagnostic test has a good performance (in terms of discrimination) if the inequality of the predicted probabilities for the the group of bad items G_1 is smaller than the inequality of the predicted probabilities for the the group of good items G_0, meaning that the model attaches on average smaller predicted probabilities of survival to the bad and higher predicted

probabilities of survival to the bad.

Higher values of $P^{ER}(t_{(k)})$ reveal high discriminatory power of the survival model at time $t_{(k)}$.

We propose to synthesize the overall predictive accuracy in a single time-independent measure of the discrimination performance of a predictor for the whole follow-up, by considering:

$$T = \int_0^\infty P^{ER}(t)w(t)dt.$$

This time-independent measure of discrimination is a time-dependent weighted mean. We follow the proposal by Lambert and Chevret (2013), by considering as time-dependent weights $w(t)$ the marginal density of failure times, that is $w(t) = f(t)$. Since practical analyses typically consider a restricted time range (t_{min}, t_{max}), we restrict also the overall measure T as follows:

$$T_r = \int_{t_{min}}^{t_{max}} P^{ER}(t)w_r(t)dt,$$

where $w_r(t) = \frac{f(t)}{\int_{t_{min}}^{t_{max}} f(t)dt}$. Since $f(t)dt = -dS(t)$, where $S(t)$ is the survival function, hence the previous expression becomes

$$T_r = -\frac{1}{S(t_{min}) - S(t_{max})} \int_{t_{min}}^{t_{max}} P^{ER}(t)dS(t), \tag{8}$$

A nonparametric estimator for T_r could be

$$\hat{T}_r = \frac{1}{\hat{S}(t_{min}) - \hat{S}(t_{max})} \sum_{k=0}^{K} \hat{P}^{ER}(t_{(k)})(\hat{S}(t_{(k)-1} - \hat{S}(t_{(k)}),$$

where $\hat{S}(t)$ is the Kaplan-Meier estimator of the survival function and $\hat{T}(t)$ is a nonparametric estimator of the overall index of discrimination.

The survival model discriminates well between the two groups, if the overall index (**??**) is significantly greater than zero.

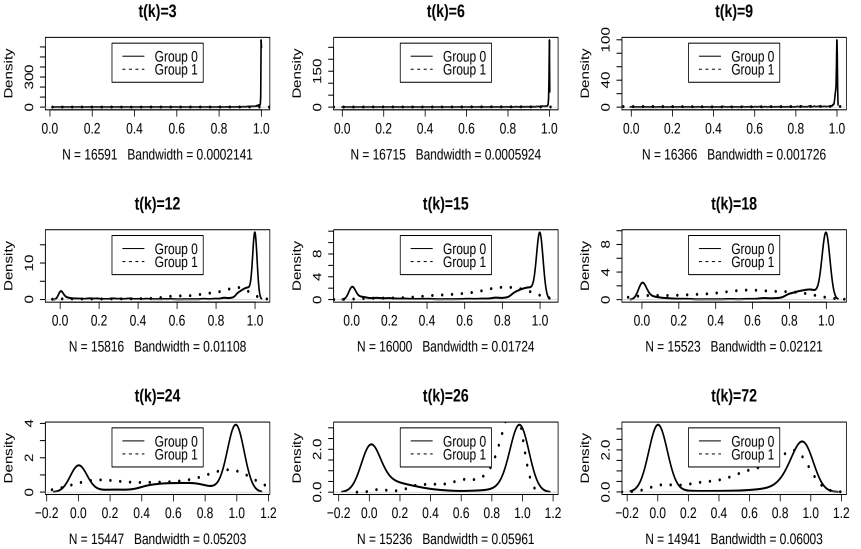
4. An application to churn risk

Our case study concerns a media service company. The main objectives of such a company are to maintain its customers, in an increasingly competitive market, and to reduce the churn risk of such customers, by carefully designing appropriate marketing actions. A variety of statistical techniques arising from medical survival analysis can be applied to tenure modeling and churn risk analysis (see e.g. Andreeva and Crook 2005; Backiel et al. 2016; Im et al.2012). In this section we look at tenure prediction using classical survival analysis based on Cox regression models and compare it with non parametric techniques based on Random Survival Forest. The tenure prediction models we have developed generate, for a given customer, a hazard curve or a hazard function, that indicates the probability of cancellation at a given time t in the future. A hazard curve can be converted to a survival curve or to a survival function which plots the probability of survival (non-cancellation) at any time, given that the customer was alive (active) at time t-1.

In our application at hand, in order to build a survival analysis model, we have constructed two variables: one variable of status (distinguishing between active and non active customers) and one of duration (indicator of customer seniority) . The data set is composed of 17,000 observations. The explanatory variables employed to run the Cox Model and the Random Survival Forest are: socio- demographic information about the customers; information about their contractual situation and about its changes in time; information about contacting the customers (through the call centre, promotion campaigns, etc).

The variables regarding customers contain demographic information (age, gender, marital status, location, number of children, job and degree) and other information about customer descriptive characteristics: hobbies, pc possession at home, address changes. The variables regarding the contract contain information about its chronology (signing date and starting date, time left before expiration date), its value (fees and options) at the beginning and at the end of the survey period, about equipments needed to use services (if they are rented, leased or purchased by the customer) and binary variables which indicate if the customer has already had an active, cancelled or suspended contract. There are also information about invoicing (invoice amount compared to different period of time 2, 4, 8, 12 months). The variables regarding payment conditions include information about the type of payment of the monthly subscription (postal bulletin, account charge, credit card), as well as other info about the changes of the type of payment. The data set used for the analysis also includes variables that provide information about the type of the services bought, about the purchased options, and about specific ad-hoc purchases, such as number and total amount of spe-

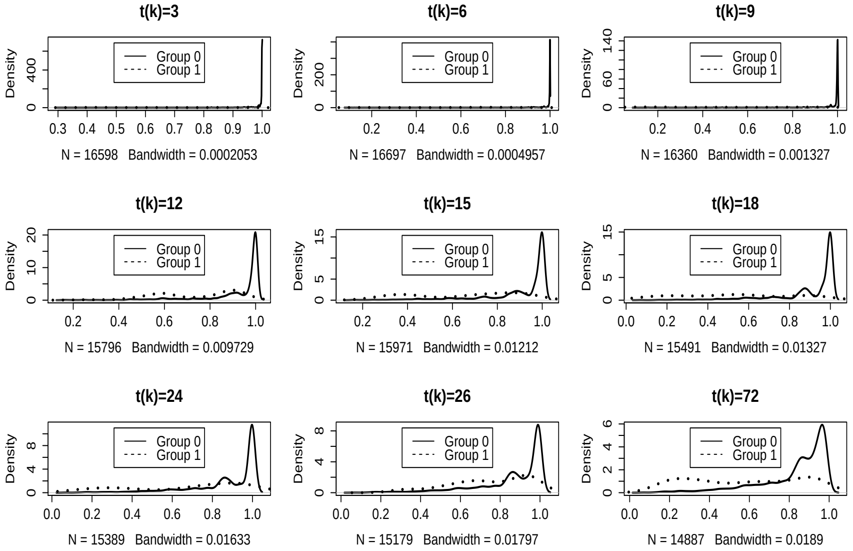
Figure 3: Cox model: Kernel density estimation of the predicted probabilities of survival at time $t_{(k)}$, by groups



cific purchases during the last month and the last 2 months. The variables regarding contacts with the customer contain information about any type of contact between the customer and the company (mostly through calls to the call centre). They include many types of calling categories (and relatives sub-categories). They also include information about the number of questions made by every customer and temporal information, such as the number of calls made during the last month, the last two months and so on.
In order to show how our proposal works, it is necessary to evaluate the models, besides the statistical aspects, also by comparing the potential advantages using one model rather than another.
We have discretized the survival times in $t = (3, 6, 9, 12, 15, 18, 24, 36, 72)$ months. Figures **??** and **??** illustrate the group-specific kernel density estimates of the predicted probabilities of survival at each fixed time for Cox model (Figure **??**) and for the Random Survival Forest (Figure **??**). We note that both models differentiate more the groups as time $t_{(k)}$ increases.

The temporal dependent models implemented (Cox regression models and Random Survival Forest) have been evaluated in terms of performance indicators able to reflect the time dependent nature of data. For each time

Figure 4: RSF model: Kernel density estimation of the predicted probabilities of survival at time $t_{(k)}$, by groups



$t = (3, 6, 9, 12, 15, 18, 24, 36, 72)$, we have computed the Harrel C concordance index defined in expression (**??**), the time-dependent $AUC(t)$ defined in (**??**), and our proposal of polarization-based discrimination index $P^{ER}(t)$ as in (**??**). Results are shown in Table **??** and Figure **??**. We note that the AUC measure prefers the Random Survival Forest for each period of time, while the opposite result is obtained if looking at the Harrell C measure. Differently, the measure $P^{ER}(t)$ that we have proposed, which is based on the values of the estimated predicted probabilities rather than their rankings, does not seem to detect differences in the two survival models.

Figure 5: Time dependent model comparison, using different discrimination indices

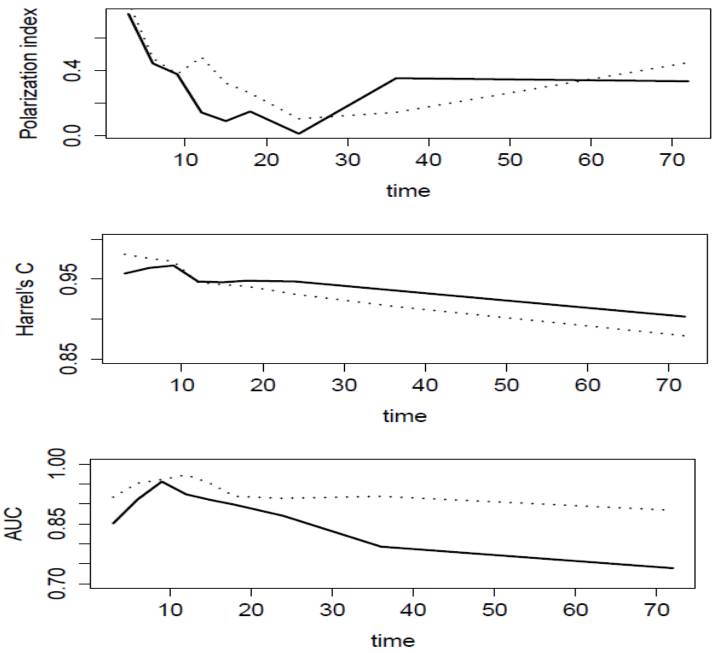


Table 1: Comparing discrimination indices

	t_3	t_6	t_9	t_{12}	t_{15}	t_{18}	t_{24}	t_{36}	t_{72}
Cox									
$P^{ER}(t)$	0.750	0.445	0.3788	0.143	0.090	0.149	0.012	0.354	0.335
$AUC(t)$	0.851	0.911	0.956	0.924	0.910	0.898	0.870	0.793	0.739
Harrell C	0.957	0.964	0.967	0.947	0.946	0.948	0.947	0.936	0.903
RSF									
$P^{ER}(t)$	0.825	0.477	0.379	0.484	0.324	0.262	0.104	0.144	0.450
$AUC(t)$	0.917	0.952	0.961	0.973	0.951	0.919	0.914	0.919	0.884
Harrell	0.981	0.976	0.972	0.946	0.943	0.941	0.931	0.916	0.879

5. Conclusion

In this paper we have revised statistical indexes to compare the accuracy of temporal models and we have proposed a novel class of discrimination measures particularly interesting to assess predictive models based for survival data. More precisely using polarization indexes we are able to measure concentration in the predicted probabilities of survival within groups of subjects and to detect differences in heterogeneity between the predicted survival probabilities of two groups of subjects.

Furthermore, differences in the concentration between groups may suggest the presence of a differential covariate effects, thus providing information of the discriminative power of a survival model.

The empirical results at hand underline that the novel class of indexes is more informative with respect to the indexes of performance proposed in the literature to compare survival models. Future works might consider alternative measures of polarization, such as overlap measures or extensions of the Wolfson index in case of two overlapping groups.

References

Andreeva, G. and Crook J., 2005. Modelling the purchase propensity: analysis of a revolving store card, Journal of the Operational Research Society, Volume 56, Issue 9, pp 1041-1050.

Antolini L., Boracchi P., Biganzoli E., 2005. A time-dependent discrimination index for survival data, Statistics in Medicine, 24, 3927-3944.

Backiel, A. Baesens B. and Claeskens G., 2016. Predicting time-to-churn of prepaid mobile telephone customers using social network analysis. Journal of the Operational Research Society (under publication).

D'Ambrosio, C., 2001. Household characteristics and the distribution of income in Italy: An application of a social distance measures. Review of Income and Wealth 47, 43-64.

Dodd, L. E., and Pepe, M. S., 2003. Partial AUC estimation and regression. Biometrics, 59, 614–623.

Duclos, J., Esteban, J. and Ray, D., 2004. Polarization: Concepts, measurement, estimation. Econometrica 72, 1737-1772.

Esteban, J. and Ray, D., 1994. On the measurement of polarization. Econometrica 62, 819-851.

Esteban, J., Gradin, C. and Ray, D., 2007. An extension of a measure of polarization, with an application to the income distribution of five OECD countries. Journal of Economic Inequality 5, 1-19.

Hand D.J. 1997. Construction and Assessment of Classification Rules. Chichester Wiley.

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

Harrell Jr. Califf R.M., Pryor D.B., F.E., Lee K.L., Rosati R.A. 1982, Evaluating the yield of medical tests, Journal of the American Medical Association, 247, 2543-2546.

Heagerty P. J. and Zheng Y., 2005. Survival model predictive accuracy and ROC curves, Biometrics, 61, 92-105.

Klawonn, F., Hoeppner, F., and May, S., 2011. An alternative to ROC and AUC analysis of classifiers. In Gama, J., Bradley, E., and Hollmen, J., eds. Advances in Intelligent Data Analysis, 210–221, Berlin: Springer.

Krzanowski W.J. and Hand D.J. 2009. ROC curves for continuous data. Chapman and Hall, London.

Flach P.A. 2003. The geometry of ROC space: understanding machine learning metrics through isometrics. Proc. 20 th International Conference on Machine Learning (ICML03), pp. 194-201.

Gigliarano C., Figini S. and Muliere P., 2014. Making Classifier Performance Comparisons when ROC Curves intersect. Computational Statistics and Data Analysis, vol. 77, pp. 300-312.

Gradin C., 2000. Polarization by sub-populations in Spain, 1973-91. Review of Income and Wealth 46, 457-474.

Im J., Apley, D.W., Qi C. and Shan X., 2012. A time-dependent proportional hazards survival model for credit risk analysis. Journal of the Operational Research Society, Volume 63, Issue 3, pp 306-321.

Lambert J. and Chevret S., 2014. Summary measure of discrimination in survival models based on cumulative/dynamic time-dependent ROC curves, Statistical Methods in Medical Research, DOI: 10.1177/0962280213515571

Lee, W., 1999. Probabilistic analysis of global performances of diagnostic tests: interpreting the Lorenz curve-based summary measures. Statistics in Medicine, 18, 455–471.

Lloyd, C.J., 1998. Using smoothed receiver operating characteristic curves to summarize and compare diagnostic systems. Journal of the American Statistical Association, 93, 1356-1364.

Lusted, L. B., 1971. Signal detectability and medical decision-making. Science, 171, 1217–1219.

McClish, D. K., 1989. Analyzing a portion of the ROC curve. Medical Decision Making, 9, 190–195.

Moon T. H. , Sohn S. Y. , 2011. Survival analysis for technology credit scoring adjusting total perception. Journal of the Operational Research Society, Volume 62, Issue 6, pp 1159-1168.

Pencina, M.J., and D'Agostino, R. B., 2004. Overall C as a measure of discrimination in survival analysis: model specific population value and confidence interval estimation. Statistics in Medicine, 23, 2109–2123.

Pencina, M. J., D'Agostino, R. B. Sr, D'Agostino R. B. Jr, Vasan, R. S., 2008. Evaluating the added predictive ability of a new marker: from area under the ROC curve to reclassification and beyond, Statistics in Medicine, 27, 157-72.

Walter, S. D., 2005. The partial area under the summary ROC curve. Statistics in Medicine, 24, 2025-2040.

Wang, Z., and Chang, Y. I., 2011. Marker selection via maximizing the partial area under the ROC curve of linear risk scores. Biostatistics, 12, 369–385.

Wolfson, M. C., 1994. When inequalities diverge. The American Economic Review 48, 353-358.

Yousef, W. A., 2013. Assessing classifiers in terms of the partial area under the ROC curve, Computational Statistics and Data Analysis, 64, 51-70.

Zhang, B., 2006. A semiparametric hypothesis testing procedure for the ROC curve area under a density ratio model, Computational Statistics and Data Analysis, 50, 1855 - 1876.

Zhang, X. and Kanbur, R., 2001. What difference do polarisation measures make? An application to China. The Journal of Development Studies 37, 85-98.