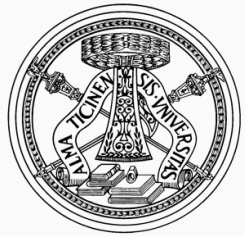


A bootstrap method to model small datasets with survival outcome

Elena Ballante

Department of Political and Social Sciences, University of Pavia

INFN Pavia

Next_AIM general meeting, 13 - 15 February 2023

Table of contents

Introduction and purposes

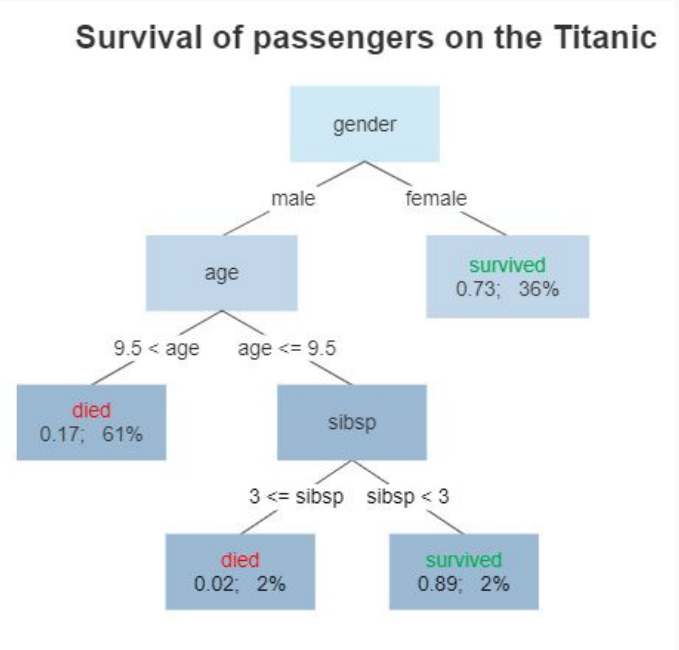
Purposes

The aim of this work is to propose an improved version of survival bagging trees that could be applied in the context of small datasets to obtain more reliable and stable results.
The difference with respect to classical survival bagging trees is the introduction of an extension of Efron's bootstrap procedure.

Survival tree models

Let be U the true survival time and C the true censoring time. The observed data is then composed of $\tau = \min(U, C)$, the time until either the event occurs or the subject is censored; $\delta = I(U \leq C)$, an indicator that takes a value of 1 if the true time–to–event is observed and 0 if the subject is censored; and $X = (X_1, ..., X_p)$, a vector of p covariates. Data is available for N independent subjects (τ_i, δ_i, X_i), $i = 1, ..., N$. The basic setup assumes that the covariate values are available at time 0 for each subject.

Decision tree models are non-parametric predictive tools used to make inference about an unknown function f that relates the time-to-event t with a p dimension vector of covariates **x**.
These models are easily interpretable and competitive in terms of predictions, although they are recognized to be an **unstable procedure**.



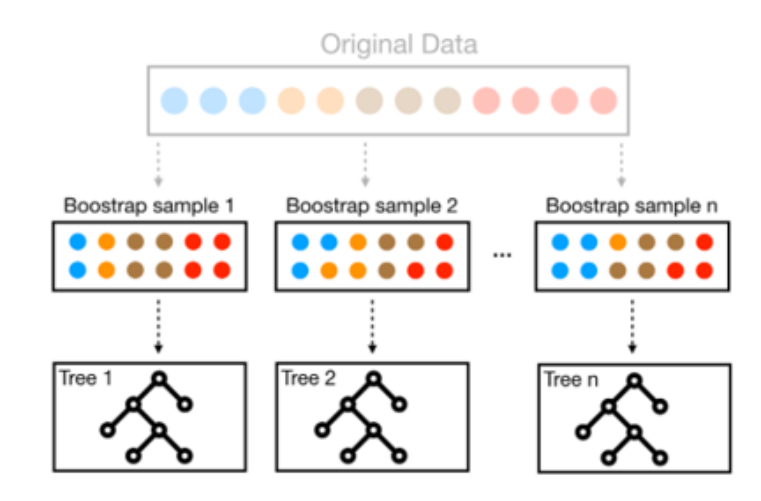
Bagging procedure

Bagging algorithm**

Bagging algorithms combine B different weak predictors to improve the stability of the model and reduce the model error.

Bagging procedure are proved to work well with unstable models, for this reason decision trees are good candidates to be employed in bootstrap procedures.

Breiman's Bagging procedure is based on Efron's bootstrap.



** Ref: Breiman, L. (1996). *Bagging Predictors*, Machine Learning 24(2), pp. 123–140.

*** Ref: Hothorn, T., Lausen, B., Benner, A. and Radespiel-Tröger, M.(2004). *Bagging survival trees*, Statist. Med., 23, pp. 77-91.

Efron's Bootstrap

Define a random sample of size n drawing with replacement from the original dataset. The new sample is called bootstrap sample. The bootstrap samples is the one used to train ensemble models.
This is equivalent to associate to data points a vector of weights $(\pi_1, ...\pi_n)$ where $\pi_i = \frac{c_i}{n}$ and $(c_1, ..., c_n) \sim Multinom(n, (\frac{1}{n}, ..., \frac{1}{n}))$.

Rubin's Bootstrap

The Rubin's bootstrap, also called Bayesian bootstrap, modifies the Efron's bootstrap defining the vector of weights as $(\pi_1, ...\pi_n) \sim \textit{Dirichlet}(1, 1, ..., 1)$.
In Taddy et al (2015), the authors introduce the idea of replacing Efron's bootstrap with Rubin's bootstrap in bagging algorithm, defining the Empirical Bayesian Forests
Two main drawbacks of Efron's and Rubin's bootstrap:

- No prior opinions are taken into account
- Inference and prediction are based only on observed values

Proper Bayesian Bootstrap

Proper Bayesian Bootstrap*

The prior of F is assumed to be $Dir(kF_0)$ where F_0 is a proper distribution function and k is the level of confidence in the initial choice F_0. Thus the posterior of F results to be $Dir(kF_0 + nF_n)$.
The distribution F is then approximated using:

$$F^*(x) = \sum_{i=1}^{r} w_i \mathbb{I}_{[x_i, \text{inf}]}(x)$$

where $(w_1, ..., w_m) \sim Dir((k+n)p_i)$.

From a computational point of view a Bootstrap resample X_m^* is generated from $(n+k)^{-1}(kF_0 + nF_n)$ and the distribution F is approximated using:

$$F^*(x) = \sum_{i=1}^{m} w_i \mathbb{I}_{[X_i^* \leq x]}$$

where $(w_1, ..., w_m) \sim Dir(\frac{n+k}{m})$.

* Ref: Muliere, P., and Secchi, P. (1996). *Bayesian Nonparametric predictive inference and bootstrap techniques*, Ann. Inst. Statist. Math., 48(4), pp. 663–673.

Generalized Bayesian Ensemble Trees for Survival Analysis

Generalized Bayesian Ensemble Trees algorithm

Generalized Bayesian Ensemble Trees algorithm

Input: Training set T

for *b in 1:B* **do**

Sample $(\mathbf{x}_1^*, y_1^*), ..., (\mathbf{x}_m^*, y_m^*)$ from $(k+n)^{-1}(kF_0 + nF_n)$;

Draw $\mathbf{w}^b$ from $Dir(\frac{n+k}{m}, ..., \frac{n+k}{m})$;

Get $\tau^b = \tau(\mathbf{w}^b)$ running weighted tree on the new sample $\mathbf{x}_1^*, ..., \mathbf{x}_m^*$

end

Where $F_0(y, \mathbf{x}) = \prod_{k=1}^{P} F_0(x_k) F_0(y|x_1, x_2, ..., x_P)$ where the distribution function models the relations between y and the covariates chosen using prior knowledge on the data.

* Ref: Galvani, M., Bardelli, C., Figini, S., Muliere, P. (2021). *A Bayesian Nonparamtric Learning Approach to Ensamble Models Using the Proper Bayesian Bootstrap*, Algorithms, 14(1), 11

Generalized Bayesian Ensemble Trees for Survival Analysis

The main differences introduced for the application to the survival analysis are the two:

- the response variable associated to the vector of covariates generated from the prior is obtained with a suitable survival model;
- the aggregation method of the predictions should be performed taking into account the nature of the time-to-event data.

The prior relation between y and $\mathbf{x}$ is evaluated with an exponential regression model.
About the second point, the output of the model is a bootstrap aggregated version of the estimated conditional survival function S for a new observation $\mathbf{x}_{new}$ computed by averaging the cumulative hazard functions (obtained by Nelson-Aalen estimator) of each leaf where the new observation falls.

* Ref: Hothorn, T., Lausen, B., Benner, A. and Radespiel-Tröger, M. .(2004). *Bagging survival trees*, Statist. Med., 23, pp. 77-91.
** Ref: Ishwaran, H., Kogalur, U.B., Blackstone,E.H., Laure, M. (2008). *Random Survival Forest*, The Annals of Applied Statistics, 2(3), pp. 841–860

Empirical Results on simulated data

Experimental setting

The simulated datasets are composed of 5 numerical covariates and a time-to-event target variable with a 20% of censored observations. The simulation of the data is performed by the flexible-hazard method as described in Harden and Kropko (2018). The sample sizes are set as $N = 30$, $N = 50$ and $N = 100$.

The priors are uniform distributions on the range of each covariate in the original dataset and the parameter k is set such that the weight $w = \frac{k}{k+n}$ assigned to the prior F_0 is equal to 0.25.

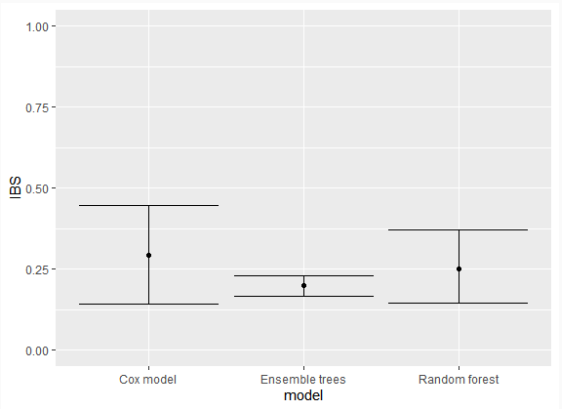
The proposed model is compared with the most common models in the survival analysis: the **Survival Random Forest** and the **Cox model**.

Prediction performance was evaluated in terms of Integrated Brier Score (IBS) in a 5-fold cross validation exercise. For each setting, 100 datasets are generated. Mean values and nonparametric confidence intervals of the resulting IBSs are presented.

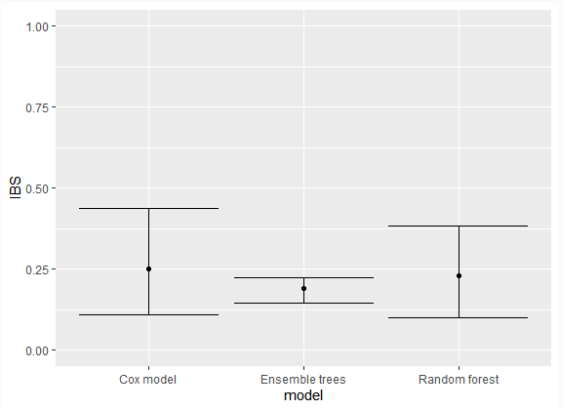
* Ref: Harden, J. J. and Kropko, J. (2018). *Simulating Duration Data for the Cox Model*, Political Science Research and Methods

Empirical Results on simulated data

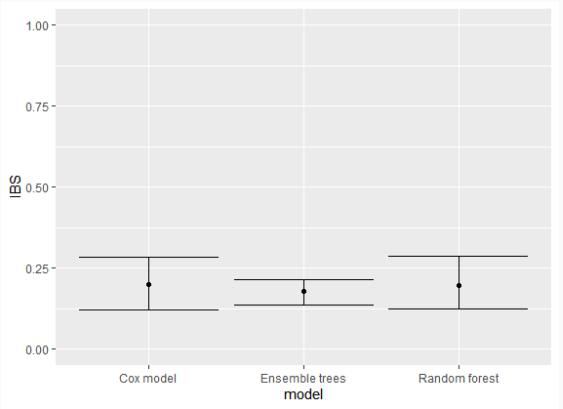
Figure 1: *Comparison of mean and nonparametric confidence intervals for IBS obtained in cross validation for the 100 simulated datasets.*



(a) *N=30*



(b) *N=50*



(c) *N=100*

Application on Blue Sky Radiomics study

The BlueSky study

The "Blue Sky Radiomics" study (NCT04364776) aims to investigate the prognostic role of radiomic features in predicting progression-free survival (PFS) in a series of stage III, unresectable, PD-L1 positive non-small-cell lung cancer(NSCLC) patients undergoing chemoradiotherapy (CRT) and maintenance durvalumab.
We consider CT images of n=57 patients have been acquired with intravenous contrast medium, and different scanners at the diagnosis time (T0). The ROI (primary lung tumor) was segmented using Oncentra Masterplan software; radiomic features have been extracted with LIFEx and harmonized with the ComBat tool*.

* Ref: Cabini,R.F. et al (2018). *Preliminary report on harmonization of features extraction process using the ComBat tool in the multi-ceneter "Blue Sky Radiomics" study on stage III unresectable NSCLC*, Insights Imaging (2022)

Empirical Results on real data

We considered for the analysis a dataset of 51 subjects and 47 covariates. Variables involved are 42 radiomic variables and 5 variables related to clinical and histological information.

Performance	BBBtrees $w = 0$	BBBtrees $w = 0.25$	BBBtrees $w = 0.5$	RF	Cox
IBS	0.1276194	0.2396914	0.2396914	0.2292683	-
CV IBS	0.09757971	0.2680563	0.2999776	0.2029304	-

Table 1: *Results on BlueSky radiomic dataset in terms of integrated Brier score*

In this specific application we can observe that the use of Rubin bootstrap as a particular case of the proper Bayesian bootstrap lead to an important improvement with respect to the other methods.

Conclusions and further ideas of research

Conclusions

- Generalized Bayesian Ensemble Tree model introduces a proper Bayesian framework in the original bagging procedure.
- Generalized Bayesian Ensemble Trees model allows the introduction of prior knowledge thus considering also observations not included in the data at hand.
- Model performance improves with respect to the other classical models especially for low sample sizes.
- On the basis of the analysed simulated data, obtained prediction models using the proper Bayesian bootstrap are more stable.
- An important room of improvement could be the deployment of non parametric methods for the generation of survival time in synthetic generated data.

References i

[1] Galvani, M., Bardelli, C., Figini, S., Muliere, P.: A Bayesian Nonparamtric Learning Approach to EnsambleModels Using the Proper Bayesian Bootstrap. Algorithms **14**(1), 11 (2021) doi: 10.3390/a14010011

[2] Muliere, P., Secchi, P.: Bayesian Nonparametric Predictive Inference and Bootstrap Techniques.Annals of the Institute of Statistical Mathematics. **48**(4), 663–673 (1996)

[3] Efron, B.: Bootstrap methods: another look at the Jackknife, Annals of Statistics, **7**(1), 1–26 (1979).

[4] Rubin, D. B.:The Bayesian Bootstrap, Annals of Statistics, **9**(1), 130–134 (1981).

[5] Gordon, L. and Olshen, R.A.: Tree-structured survival analysis. Cancer treatment reports **69** (10) 1065–1069 (1985)

[6] Ishwaran, H., Kogalur, U.B., Blackstone,E.H., Laure, M.: Random Survival Forest. The Annals of Applied Statistics **2**(3), 841–860 (2008) doi: 10.3390/a14010011

References ii

[7] Hothorn, T., Lausen, B., Benner, A. and Radespiel-Tröger, M.: Bagging survival trees. Statist. Med., **23**, 77-91 (2004). https://doi.org/10.1002/sim.1593

[8] Harden, J. J. and Kropko, J.: Simulating Duration Data for the Cox Model. Political Science Research and Methods (2018) https://doi.org/10.1017/psrm.2018.19

[9] Andersen, P. and Gill, R.: Cox's regression model for counting processes, a large sample study. Annals of Statistics **10**, 1100-1120 (1982).

Thank you for the attention!