# Forêt aléatoire pour la régression d'une variable censurée

Yohann Le Faou [1] & Arnaud Cohen [2] & Guillaume Gerber [3] &
Olivier Lopez [4] & Michael Trupin [5]

[1] *Forsides & Univ. Pierre et Marie Curie Paris VI, yohann.le_faou@etu.upmc.fr*
[2] *Forsides, 52 rue de la Victoire , 75009 Paris, France, arnaud.cohen@forsides.fr*
[3] *Forsides, 52 rue de la Victoire , 75009 Paris, France, guillaume.gerber@forsides.fr*
[4] *Univ. Pierre et Marie Curie Paris VI, 4 place Jussieu, 75005 Paris, France, olivier.lopez0@upmc.fr*
[5] *Santiane, 38 avenue des Champs-Elisées, 75008, Paris, France, m.trupin@santiane.fr*

**Résumé.** Sur le marché du courtage de produits d'assurance, les commissions perçues par les courtiers dépendent fortement de la résiliation observée sur les contrats. Dans l'optique d'optimiser un processus commercial, un scoring de prospects doit donc prendre en compte cette composante de résiliation. Nous proposons d'utiliser une forêt aléatoire pondérée pour prédire le facteur résiliation intervenant dans le score. Notre modèle est adapté à la censure des observations, omniprésente dans l'étude des mécanismes de résiliation. A travers des estimations sur données réelles et simulées, nous comparons notre approche à d'autres méthodes standards d'étude de variables censurées qui s'appliquent dans notre situation. Nous montrons que notre approche est compétitive en termes d'erreur quadratique pour répondre au problème posé.

**Mots-clés.** Forêt aléatoire, Données censurées, Régression, Assurance

**Abstract.** In the insurance broker market, commissions received by brokers are closely related to the surrender of the insurance contracts. In order to optimize a commercial process, a scoring of prospects should then take into account this surrender component. We propose a weighted Random Forest model to predict the surrender factor which is part of the scoring. Our model handles censoring of the observations, a classical issue when working on surrender mechanisms. Through careful studies of real and simulated data, we compare our approach with other standard methods which apply in our setting. We show that our approach is very competitive in terms quadratic error to address the given problem.

**Keywords.** Random Forest, Censored Data, Regression, Insurance

# 1 Introduction

Given a quantitative random variable $T$, a function $\phi$ and a vector of covariates $X$, a common problem in statistics, called regression, is to estimate $E[\phi(T)|X]$ as a function of $X$. A well known regression technique brought by L. Breiman in the early 2000s ([2]) is the Random Forest algorithm. We propose to adapt the Random Forest method to the case where $T$ is right-censored by a random variable $C$. Our method inspires from [10] which describes a CART algorithm for the study of a censored variable. We emphasize practical aspects of our work, as one of our purposes is to build a scoring system for the use of an insurance broker. Of particular interest is the computation of the observation weights we use in our method, which we carefully discuss. We also compare performances of our method to other state of the art models in real and simulated data studies.

Random Survival Forest have been proposed in [6] and [5] to extend Random Forests to the censored case. This algorithm aims to model the entire survival function of $T$, given $X$, and thus can be used to estimate $E[\phi(T)|X]$. Our approach is more direct than the latter since it does not rely on the estimation of the whole conditional distribution of $T$. Indeed, our algorithm relies on the weighting of the observations by the inverse-probability-of-censoring weighting principle. The same idea is studied in [12] and [4] but these articles restrict to single tree models in the applications and don't go into details about the practical computation of the weights, two subjects we believe we bring new contributions.

Before, first regression settings involving a censored variable of interest were developed in late 70s, beginning with extensions of the linear model : [11], [3]. Also, [7] is to our knowledge the first apparition of the inverse-probability-of-censoring weighting principle for regression in the survival field. Many tree-based algorithms, including with bagging extension, have been investigated in the past for the study of right-censored data and [1] gives an overview of the state of the art in the domain.

Our work is motivated by an application to insurance that we describe in part 3

# 2 Mathematical formulation

## 2.1 Definitions

Let $T$ a right-censored random variable. We call $C$ the censoring variable of $T$. This means each experiment does not lead to an observation of $T$. In fact, each experiment leads to an observation of :

$$Y = \min(T, C)$$
$$\delta = \mathbb{1}_{T \leq C}$$

Let $X \in \mathbb{R}^p$ a vector of covariates and $\phi$ a real valued function. In this context, we are interested in estimating the influence of $X$ on $\phi(T)$.

## 2.2 Presentation of the method

### 2.2.1 Weighted Random Forest

We have observations $(Y_i, \delta_i, X_i)_{i=1,..,n}$ and we look for estimations of $f(x) = E[\phi(T)|X = x]$. As we know, $f$ is the solution to the optimization problem :

$$f = \underset{g}{\operatorname{argmin}}\, E\left[(\phi(T) - g(X))^2\right] \tag{1}$$

We then choose the Random Forest algorithm with the mean squared error splitting criteria to estimate $f$. This leads us in looking for estimators of quantities of the form : $E[\psi(T, X)]$. Under some hypothesis it is possible to estimate the quantity $E[\psi(T, X)]$ asymptotically without bias, with $\psi$ a real valued function. We use the inverse-probability-of-censoring weighting principle to do so. It is a general principal that provide unbiased estimate of the law of a couple $(Z, X)$ when observation of $X$ is complete and observation of $Z$ is censored :

**Proposition 1** *Let* $\gamma = \begin{cases} 1 \text{ if } Z \text{ is censored} \\ 0 \text{ if } Z \text{ is censored} \end{cases}$ *and* $Z' = \gamma Z$. *Let* $p(X, Z) = P(\gamma = 1|X, Z)$

*Then for any function* $\psi$,

$$E\left[\frac{\gamma}{p(X, Z')} \cdot \psi(Z', X)\right] = E\left[\psi(Z, X)\right]$$

This result states that given a couple $(X_i, Z_i)$ if we only observe $(X_i, \gamma_i, Z_i')$ we can still get an unbiased estimator of the distribution of $(X, Z)$. This is done attributing weights $\frac{\gamma_i}{p(X_i, Z_i')}$ to the observations, with $p(x, z)$ the probability of $Z$ being non-censored, given $X = x$ and $Z = z$.

In our case, the probability of being non-censored given $X$ and $T$ is $P(\delta = 1|X, T) = P(T \leq C|X, T)$. In the survival censoring scheme, it is impossible to infer the latter since it is well known it is impossible to estimate the dependence between $T$ and $C$ (see section 4.1 in [8]). Therefore, we have to make assumptions about the dependence between $T$ and $C$. Let **H1** and **H2** denotes the following hypothesis :

$$\textbf{H1} : P(T \leq C|X, T) = P(T \leq C)$$
$$\textbf{H2} : P(T \leq C|X, T) = P(T \leq C|X)$$

Sufficient conditions for these hypothesis to be satisfied are, respectively, $T \perp\!\!\!\perp C$ (**H1**) and $T \perp\!\!\!\perp C$ conditionally on $X$ (**H2**).

Let $S_C$ the survival function of $C$, $S_C(\cdot|X)$ the survival function of $C$ given $X$, and denote by $\hat{S}_C$ and $\hat{S}_C(\cdot|X)$ estimators of these functions. Then, depending on the hypothesis we make, let $\hat{W}_i = \frac{\delta_i}{\hat{S}_C(Y_i)}$ or $\frac{\delta_i}{\hat{S}_C(Y_i|X_i)}$. We estimate $E[(\phi(T) - g(X))^2]$ by

$$\frac{1}{n}\sum_{i=1}^{n} \hat{W}_i \cdot (\phi(Y_i) - g(X_i))^2 \tag{2}$$

The Random Forest adaptation we propose is then a weighted Random Forest. Weights are taken into account in the bootstrap procedure. Indeed, during the sampling of a bootstrap set, we do a sample with replacement where each observation has probability $\hat{W}_i$ of being sampled.

This way, each observation accounts in the growing of the forest proportionally to its weight.

#### 2.2.2 Calculation of the weights

In practice, computation of the weights $\hat{W}_i$ requires to model $S_C$ and $S_C(\cdot|X)$.

To estimate $S_C$ under the **H1** hypothesis, we use a Kaplan-Meier estimator [13]. But for $S_C(\cdot|X)$ under **H2** it is necessary to use statistical models which take into accounts influences of covariates. We try and compare 2 techniques to estimate $S_C(\cdot|X)$ :

- RSF : Random Survival Forest (see [5])
- Cox model (see [9])

We test performances of our Random Forest model with these different weights and we analyse our results. We discuss which of these techniques achieves the best performance. In this step it is necessary to deal with a trade-off between bias and variance : indeed, some $S_C(\cdot|X)$ estimators like RSF may indicate very big weights $\hat{W}_i$ at some points. At the opposite Kaplan-Meier weights tend to be smoother. Big weights usually reduce bias but if a weight is too big the model in its $x$-neighbourhood is too much influenced by a single observation resulting in very high variance for the prediction.

# 3 Results

## 3.1 Application Scheme

Our work is motivated by an application in insurance where $T$ corresponds to termination time of a contract and $\phi$ gives the amounts of commissions received by an insurance broker per unity of premium. $\phi$ then represents the impact of the termination time of a contract on the turnover this contract brings.

## 3.2 Performances of the method

We compare the performances of the weighted random forest with 2 benchmarks we call *direct Cox* and *direct RSF*. As there names highlight, these models also rely respectively on the Cox model and the RSF model. But here we use them differently : in each case (*direct Cox* and

*direct RSF*), we use the model to estimate $S_T(\cdot|X)$ (let $\hat{S}_T(\cdot|X)$ the estimator) and then we get an estimate of $E[\phi(T)|X = x]$ integrating $\phi$ against $\hat{S}_T(\cdot|X)$ : $\hat{f}(x) = -\int_0^{+\infty} \phi(t) d\hat{S}_T(t|X = x)$.

Using a train-test approach, we compute the performances of the 5 models we compare : 3 weighted RF and 2 benchmarks. This is made on both real data and simulated data. We show that our method is competitive in terms of quadratic error for the prediction of $\phi$.

# Bibliographie

# References


[1] I. Bou-Hamad, D. Larocque, H. Ben-Ameur, et al. A review of survival trees. *Statistics Surveys*, 5:44–71, 2011.

[2] L. Breiman. Random forests. *Machine learning*, 45(1):5–32, 2001.

[3] J. Buckley and I. James. Linear regression with censored data. *Biometrika*, 66(3):429–436, 1979.

[4] T. Hothorn, P. Bühlmann, S. Dudoit, A. Molinaro, and M. J. Van Der Laan. Survival ensembles. *Biostatistics*, 7(3):355–373, 2006.

[5] H. Ishwaran and U. B. Kogalur. Random survival forests for r. *R news*, 7, 2007.

[6] H. Ishwaran, U. B. Kogalur, E. H. Blackstone, and M. S. Lauer. Random survival forests. *The annals of applied statistics*, pages 841–860, 2008.

[7] H. Koul, V. v. Susarla, and J. Van Ryzin. Regression analysis with randomly right-censored data. *The Annals of Statistics*, pages 1276–1288, 1981.

[8] S. Lagakos. General right censoring and its impact on the analysis of survival data. *Biometrics*, pages 139–156, 1979.

[9] D. Y. Lin and L.-J. Wei. The robust inference for the cox proportional hazards model. *Journal of the American statistical Association*, 84(408):1074–1078, 1989.

[10] O. Lopez, X. Milhaud, P.-E. Thérond, et al. Tree-based censored regression with applications in insurance. *Electronic journal of statistics*, 10(2):2685–2716, 2016.

[11] R. G. Miller. Least squares regression with censored data. *Biometrika*, 63(3):449–464, 1976.

[12] A. M. Molinaro, S. Dudoit, and M. J. Van der Laan. Tree-based multivariate regression and density estimation with right-censored data. *Journal of Multivariate Analysis*, 90(1):154–177, 2004.

[13] W. Stute and J.-L. Wang. The strong law under random censorship. *The Annals of Statistics*, pages 1591–1607, 1993.