Random Forest for Regression of a Censored Variable

Yohann Le Faou *(yohann.lefaou@forsides.fr)*
Forsides & Sorbonne Universités - Université Pierre et Marie Curie (Paris 6, LSTA)

Contribution

- We propose a method to estimate $E[\phi(T)|X]$, where :
  - T is a **censored** duration
  - X is a vector of covariates
  - ϕ is a given function
- We adapt the well known **Random Forest** method to handle such case
- We study the performance of our algorithm through computations on real data and simulated data, and we compare it to alternative methods that may be used
- Our work is motivated by an **application to insurance**

Introduction

Practical case :

- Insurance broker business : An insurance broker takes a commission when it subscribes a contract for an insurance company
- Given a prospect, we aim to build a model which predicts the amount of commissioning (per unit of premium) it will meet
  - In our case, **the amount of commissioning (per unit of annual premium) is a function of T : the termination time of the contract** (we note ϕ this function : Figure 1)
  - If the contract didn't terminate, information about $\phi(T)$ is **censored**
  - The model should take into account the influence of characteristics of the prospect : age, gender, number of people insured, social security regime, range of insurance, geographical zone (Figure 2)

Mathematical Formulation :

- T : Termination time of the contract
- C : Censoring time
- $X \in R^d$: Covariates about the prospect : 6 covariates
- Goal : Build a model to estimate $f(x) = E[\phi(T)|X = x]$

Observations

We observe $(Y_i, \delta_i, X_i)_{1 \le i \le n}$ i.i.d with :
- $Y_i = \min(T_i, C_i)$
- $\delta_i = 1_{T_i \le C_i}$

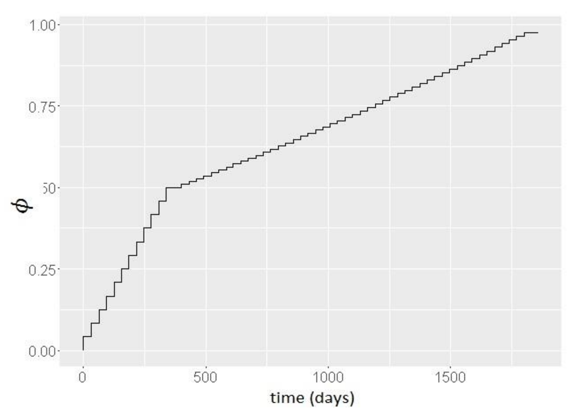


Figure 1. ϕ : Commissioning function of the insurance broker

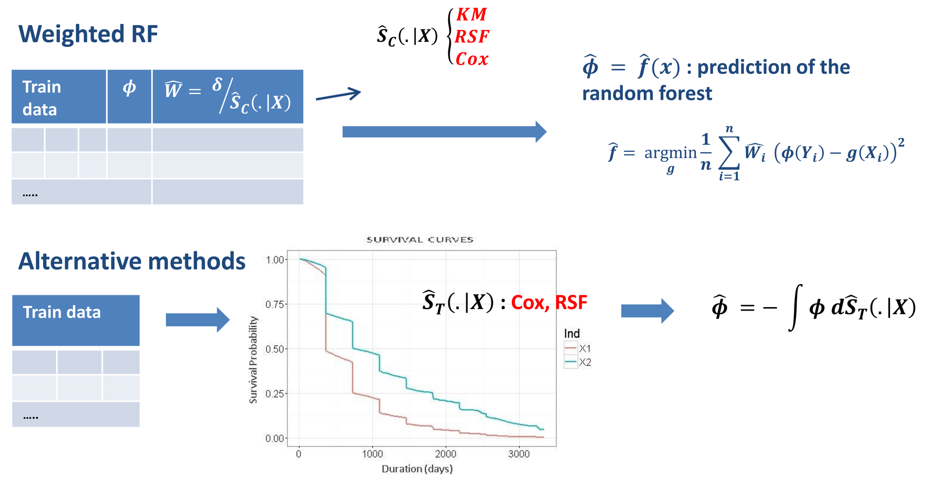


Figure 3. Different compared models

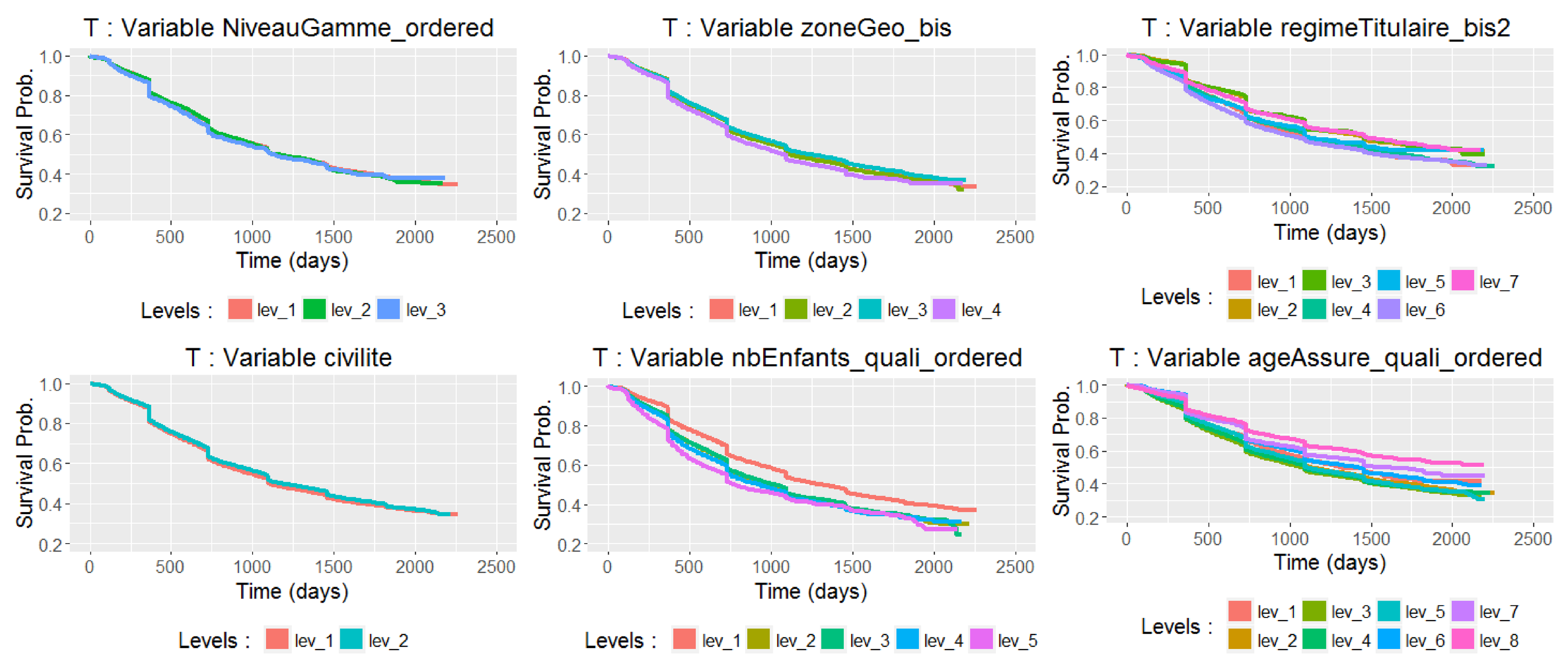


Figure 2. Descriptive statistics (data was blurred for confidentiality)

Weighted Random Forest

- We know that $f = \underset{g}{\operatorname{argmin}} E[(\phi(T) - g(X))^2]$ and we address this optimization problem using Random Forest
- We use **IPCW** principle to estimate $E[(\phi(T) - g(X))^2]$ (where T is censored) :

IPCW principle (Inverse Probability of Censoring Weighting)

Let $p(t, x) = P(\delta = 1 | T = t, X = x)$
Then, for any bounded function ψ :

$$E[W \cdot \psi(Y, X)] = E[\psi(T, X)] \quad (\text{with } W = \frac{\delta}{p(Y,X)})$$

Hypothesis

- **H1** : $P(T \le C | T, X) = P(T \le C)$ (true if C and (T, X) are independent)
- **H2** : $P(T \le C | T, X) = P(T \le C | X)$ (true if C and T are independent conditionaly on X)

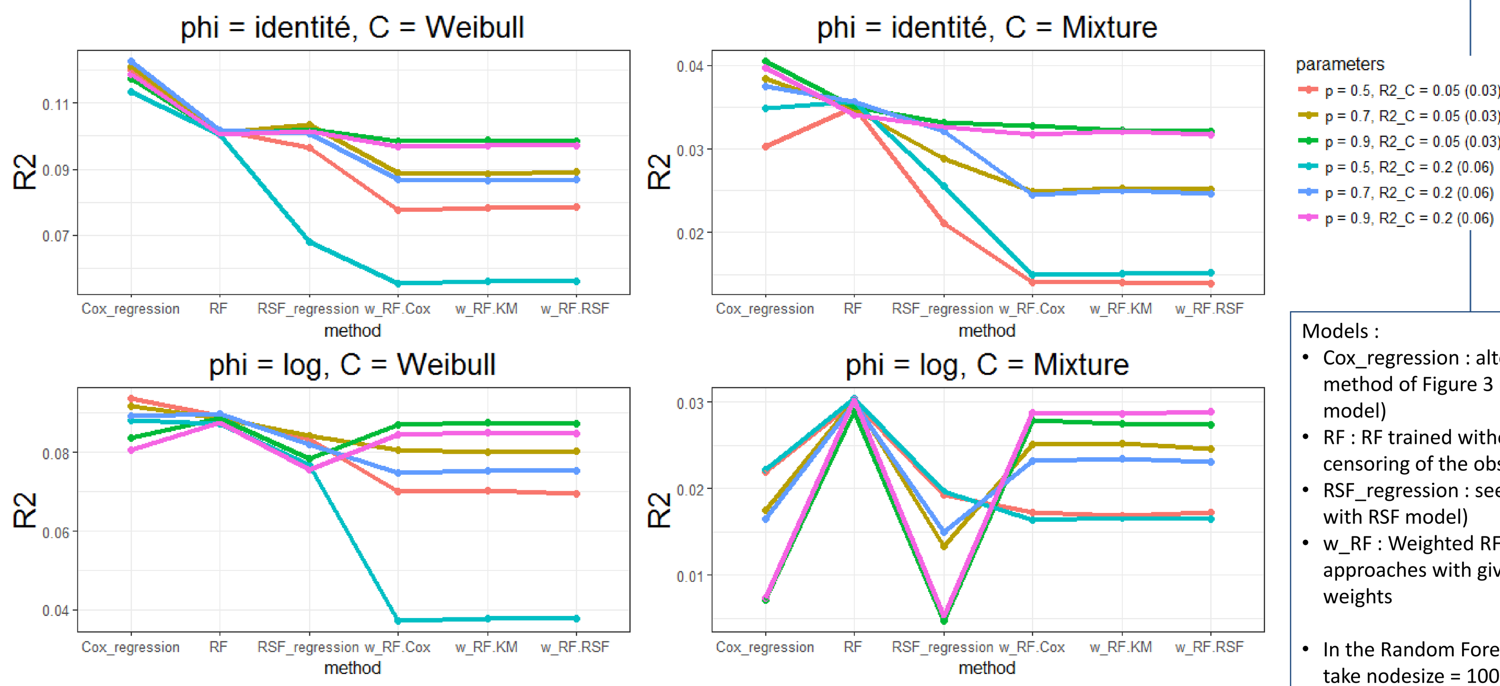
- Expression of the weights :
  - Under **H1** : $p(t, x) = P(T \le C | T = t, X = x) = P(t \le C) = S_C(t)$
  - Under **H2** : $p(t, x) = P(t \le C | X = x) = S_C(t | X = x)$
- Weighted Random Forest
  - Let $\hat{S}_C$ (resp. $\hat{S}_C(\cdot|X)$) an estimate of S_C (resp. $S_C(\cdot|X)$)
  - Depending on the hypothesis we make, let $\widehat{W}_i = \frac{\delta_i}{\hat{S}_C(Y_i)}$ or $\frac{\delta_i}{\hat{S}_C(Y_i|X_i)}$.
  - We estimate $E[(\phi(T) - g(X))^2]$ by $\frac{1}{n}\sum_{i=1}^n W_i \cdot (\phi(T_i) - g(X_i))^2$
  - **Weights are taken into account in the bootstrap of the Random Forest** : during the sampling of a bootstrap set, we do a sample with replacement where each observation has probability W_i of being sampled.

Setting of the experiments

- We compare the performances of 5 models (Figure 3):
  - 3 Weighted RF :
    - weights estimated under **H1** : Kaplan Meier (KM)
    - weights estimated under **H2** : Cox, or RSF (Random Survival Forest)
  - 2 Alternative methods : "Cox regression" and "RSF regression"
- Data
  - Simulated data
    - $X \sim Unif([-1;1]^6)$
    - 2 settings for the law of T and C:
      - Setting 1 : Weibull - $T|X \sim Weibull(\lambda_1 . e^{-t\beta_1 X}, k_1)$, $C|X \sim Weibull(\lambda_2 . e^{-t\beta_2 X}, k_2)$
      - Setting 2 : Mixture of Weibull - $T|X, G \sim Weibull(\lambda_1 . e^{-t\beta_{1,G} X}, k_1)$, $C|X, G \sim Weibull(\lambda_2 . e^{-t\beta_{2,G} X}, k_2)$ with $G \sim unif(\{1,2,3,4\})$
  - Real data
    - Data from a health insurance broker ≈ 70 000 observations
    - 47,8% is non censored
    - 6 qualitative covariates with some of them ordered (like age brackets) : age, gender, number of people insured, social security regime, range of insurance, geographical zone
- Methodologies :
  - Simulated data : train = 1000, test = 1000 (we can compute exact criteria)
  - Real data : train = 10000, test = 50000 (we can't compute exact criteria)
  - Means and standard deviations of the studied models are calculated using 100 bootstrap samples of data
- Practical issues :
  - In practice we replace T by $\min(T, const)$ so that we can calculate quantities like $\int \phi \, d\hat{S}_T$
  - We threshold the weights (in the w_RF method) so that the ratio between the smallest and the biggest weight doesn't exceed "max ratio weights"

Results

- **On Simulated data :**



Models :
- Cox_regression : alternative method of Figure 3 (with Cox model)
- RF : RF trained without censoring of the observations
- RSF_regression : see Figure 3 with RSF model)
- w_RF : Weighted RF approaches with given weights

- In the Random Forests, we take nodesize = 100 (resp. 200) for C = Weibull (resp. Mixture)
- max_ratio_weights = 20 in w_RF

  - R2_C : percentage of explained variance in the model $C|X$. (values of 0.03 and 0.06 in the mixture model)
  - p : proportion of non-censored observation

- **On Real data :**

Performance criteria :
- Notation : mean criteria (mean rank) over 100 bootstraps
- R2_20 and Kendall_20 are the classical R2 and Kendall statistics, but computed from groups of observations (of size 20)
- To design the groups, test observations are ranked in increasing order of predicted value, and then we slice each 20 observations
- Each group is then associated to an empirical Y value given by the Kaplan Meier estimator of the group
- This technique allows us working with a least square criteria in the context of censoring
- The concordance index is given as a comparison to Kendall_20

			ϕ = Commissioning function (Fig.1)			ϕ = log		
method	nodesize	max_ratio_weights	R2_20	Kendall_20	Concordance	R2_20	Kendall_20	Concordance
w_RF.KM	200	10	0.097 (25.74)	0.624 (25.9)	0.549 (25.48)	0.084 (25.9)	0.617 (26.08)	0.548 (25.52)
w_RF.KM	200	50	0.097 (25.91)	0.624 (25.87)	0.549 (25.16)	0.085 (25.85)	0.617 (26.04)	0.548 (25.22)
w_RF.KM	500	10	0.157 (14.36)	0.634 (18.72)	0.552 (16.47)	0.142 (14.12)	0.627 (18.7)	0.551 (17.04)
w_RF.KM	500	50	0.157 (14.26)	0.635 (18.29)	0.552 (15.63)	0.143 (14.44)	0.628 (18.3)	0.551 (16.47)
w_RF.KM	1000	10	0.162 (12.09)	0.637 (16.42)	0.552 (18.14)	0.146 (12.53)	0.63 (16.85)	0.551 (18.35)
w_RF.KM	1000	50	0.163 (11.61)	0.638 (16.07)	0.552 (17.55)	0.147 (12.24)	0.63 (16.66)	0.551 (17.78)
w_RF.KM	2000	10	0.149 (17.71)	0.65 (8.16)	0.555 (9.48)	0.135 (17.71)	0.644 (8.53)	0.555 (9.57)
w_RF.KM	2000	50	0.149 (18.18)	0.649 (8.95)	0.554 (10.92)	0.134 (18.09)	0.643 (9.09)	0.554 (10.1)
w_RF.RSF	200	10	0.097 (25.9)	0.623 (26.22)	0.549 (25.58)	0.084 (25.87)	0.617 (26.04)	0.548 (25.45)
w_RF.RSF	200	50	0.097 (25.72)	0.624 (25.99)	0.549 (25.54)	0.085 (25.57)	0.617 (25.95)	0.548 (25.49)
w_RF.RSF	500	10	0.158 (13.92)	0.635 (18.29)	0.552 (16.24)	0.143 (14.04)	0.628 (18.47)	0.551 (16.45)
w_RF.RSF	500	50	0.157 (14.45)	0.635 (18.63)	0.552 (16.33)	0.142 (14.4)	0.627 (18.41)	0.551 (16.65)
w_RF.RSF	1000	10	0.162 (12.4)	0.637 (16.81)	0.552 (18.25)	0.146 (12.19)	0.63 (16.53)	0.551 (17.55)
w_RF.RSF	1000	50	0.162 (12.46)	0.637 (16.29)	0.552 (17.79)	0.146 (12.33)	0.63 (16.5)	0.551 (18.18)
w_RF.RSF	2000	10	0.15 (17.31)	0.65 (8.16)	0.555 (10.19)	0.136 (17.05)	0.644 (8.12)	0.555 (9.58)
w_RF.RSF	2000	50	0.15 (17.54)	0.65 (8.24)	0.555 (9.84)	0.135 (17.18)	0.644 (8.32)	0.555 (9.25)
w_RF.Cox	200	10	0.095 (26.26)	0.623 (26.17)	0.549 (25.66)	0.084 (25.98)	0.617 (26.25)	0.548 (25.95)
w_RF.Cox	200	50	0.095 (26.42)	0.623 (26.63)	0.549 (25.49)	0.084 (26.06)	0.617 (26.12)	0.548 (25.74)
w_RF.Cox	500	10	0.157 (14.5)	0.634 (18.73)	0.552 (16.16)	0.142 (14.19)	0.627 (18.72)	0.551 (16.76)
w_RF.Cox	500	50	0.159 (13.42)	0.635 (17.93)	0.552 (15.64)	0.142 (14.29)	0.627 (18.24)	0.551 (16.58)
w_RF.Cox	1000	10	0.162 (12.05)	0.637 (16.85)	0.552 (18.36)	0.147 (12)	0.63 (15.98)	0.551 (17.61)
w_RF.Cox	1000	50	0.162 (12.35)	0.637 (16.35)	0.552 (18.49)	0.146 (12.83)	0.63 (16.75)	0.551 (18.2)
w_RF.Cox	2000	10	0.15 (17.27)	0.65 (8.41)	0.555 (9.81)	0.136 (17.21)	0.645 (7.94)	0.555 (9.25)
w_RF.Cox	2000	50	0.149 (17.93)	0.65 (8.24)	0.555 (9.66)	0.135 (17.73)	0.644 (8.26)	0.555 (9.13)
RSF regression	200	NA	0.214 (4.48)	0.657 (6.03)	0.559 (5.36)	0.2 (4.32)	0.652 (5.85)	0.558 (5.58)
RSF regression	500	NA	0.238 (1.78)	0.664 (2.59)	0.561 (2.08)	0.225 (1.67)	0.659 (2.63)	0.561 (2.13)
RSF regression	1000	NA	0.24 (1.43)	0.666 (1.85)	0.562 (1.27)	0.227 (1.49)	0.661 (1.64)	0.562 (1.25)
RSF regression	2000	NA	0.218 (4.02)	0.663 (3.26)	0.56 (3.68)	0.202 (4.13)	0.658 (2.95)	0.56 (3.65)
Cox regression	NA	NA	0.222 (3.53)	0.659 (4.95)	0.559 (4.75)	0.206 (3.59)	0.654 (5.08)	0.559 (4.52)

Analysis of results

- **Simulated data** :
  - 2 important parameters which impact the performances
    - function ϕ : w_RF perform better with $\phi = log$ than with $\phi = identity$
    - Censoring rate : performances of w_RF decrease rapidly with the censoring rate
  - Curves organize by group of 2 since $R2\_C$ has low importance. Moreover, there is no significant differences in the results given by w_KM, w_Cox and w_RSF, hence we don't need to estimate the X-conditional weights, KM weights are sufficient.
- **Real data** :
  - The censoring rate is high and our results confirm that in this case w_RF doesn't work as good as RSF regression or Cox regression.
  - The nodesize is the crucial parameter to calibrate in the Random Forest. Here 1000 is the best value in terms of both R2 and Kendall

Conclusion

- We show through a quantitative study that the weighted Random Forest method is competitive compared to other algorithm, and achieves the best performances in some settings
- We provide a least square criteria to do model selection in the context of right censoring
- **Soon : a R package and an article**