Random Forest for Regression of a Censored Variable

Guillaume Gerber[1] Yohann Le Faou[*,1,2] Olivier Lopez[2] Michael Trupin[3]

[1]Forsides Innovation, Paris, France

[2]Univ. Pierre et Marie Curie, Paris, France

[3]Groupe Santiane, Paris, France

Perspectives on Actuarial Risks in Talks of Young Researchers, january 2017

Outline

Outline

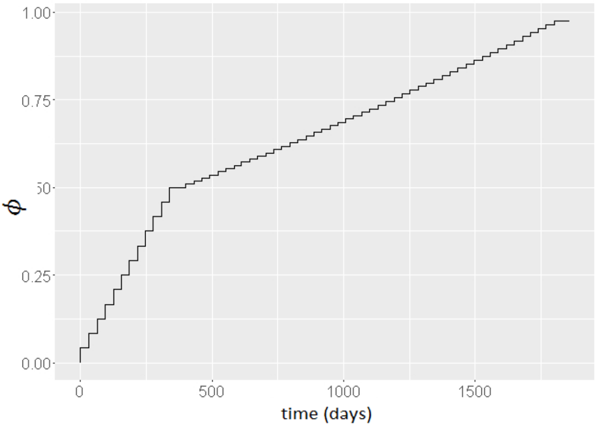
1. Introduction
   - Practical case
   - Mathematical formulation
2. Weighted Random Forest and IPCW method
   - Random Forest
   - IPCW principle
   - Estimation of the weights
3. Results of the simulations
   - Real data application

Insurance broker market

- An insurance broker takes a commission when it subscribes a contract for an insurance company
- This commission depends on the behavior of the policy holder
  - For instance, the commission is low if the contract terminate rapidly.
  - The broker may refund part of the commission
- 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)

Commissioning

- In our case, **the amount of commissioning (per unit of premium) is a function of T : the termination time of the contract** (we note ϕ this function)



Goal

- Given a prospect, we aim to build a model which predicts the amount of commissioning (per unit of premium) it will meet
  - 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.

Outline

Mathematical Formulation

- T : Termination time of the contract
- C : Censoring time
- $X \in \mathbb{R}^d$: Covariates about the prospect : 6 covariates

Observations

We observe $(Y_i, \delta_i, X_i)_{1 \leq i \leq n}$ i.i.d. with :

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

- Goal : Build a model for $f(x) = E[\phi(T)|X = x]$

Outline

Random Forest

- We want to estimate $f(x) = E[\phi(T)|X = x]$
- We know that :

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

and we address this optimization problem using Random Forest

$\Longrightarrow$ We need an estimate of the quantity $E\left[(\phi(T) - g(X))^2\right]$ with T censored.
$\Longrightarrow$ More generally, for any bounded ψ, we can estimate $E\left[\psi(T, X)\right]$ with T censored using IPCW principle

Outline

IPCW principle

- IPCW : Inverse Probability of Censoring Weighting

Proposition (IPCW principle)

Let $p(T,X) = P(\delta = 1 | T, X)$

Then for any bounded function ψ,

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

Reminder

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

IPCW principle

Proof.

$$
\begin{aligned}
E\left[\frac{\delta}{p(Y,X)}\cdot\psi(Y,X)\right] &= E\left[\frac{\delta}{p(T,X)}\cdot\psi(T,X)\right] \\
&= E\left[\frac{\psi(T,X)}{p(T,X)}\cdot\underbrace{E[\,\delta\,|T,X]}_{p(T,X)}\right] \\
&= E\left[\psi(T,X)\right]
\end{aligned}
$$

$\square$

Outline

1. Introduction
   - Practical case
   - Mathematical formulation
2. Weighted Random Forest and IPCW method
   - Random Forest
   - IPCW principle
   - Estimation of the weights
3. Results of the simulations
   - Real data application

Estimation of the weights

Hypothesis

H1 : $P(T \leq C|X,T) = P(T \leq C)$ (true if $C \perp\!\!\!\perp (T,X)$)
H2 : $P(T \leq C|X,T) = P(T \leq C|X)$ (true if $C \perp\!\!\!\perp T$ conditionally on X)

- Under **H1** : $p(T,X) = P(\delta = 1|T,X) = P(T \leq C|X,T) = P(T \leq C) = S_C(T)$
- Under **H2** : $p(T,X) = S_C(T|X)$

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 $\hat{W}_i = \frac{\delta_i}{\hat{S}_C(Y_i)}$ or $\frac{\delta_i}{\hat{S}_C(Y_i|X_i)}$. We estimate $E[(\phi(T) - g(X))^2]$ by

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

- **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 $\hat{W}_i$ of being sampled.

Outline

Data

- Data from a Health insurance broker
- 70000 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
- 29 levels

Methodologies

- Measure of performances :
  - train data : 20000 obs. / test data : 50000 obs.
  - mean and standard deviation of the performances of studied models are computed using 100 bootstrap samples of data.
- Models :
  - 3 Weighted Random Forest : weights estimated with **H1** : Kaplan Meier and **H2** : Cox model, RSF (Random Survival Forest)
  - 2 Benchmark models : Cox, RSF (Random Survival Forest)

Methodologies (to sum up)

Weighted RF

Train data	ϕ	$\widehat{W} = \delta / \widehat{S}_C(.\|X)$
.....		

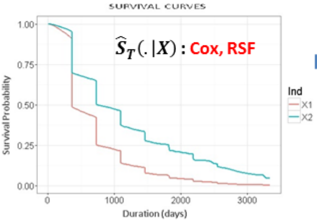
$\widehat{S}_C(.|X)$ $\begin{cases} KM \\ RSF \\ Cox \end{cases}$

$\widehat{\phi} = \widehat{f}(x)$: **prediction of random forest**

$$\widehat{f} = \underset{g}{\operatorname{argmin}} \frac{1}{n} \sum_{i=1}^{n} \widehat{W}_i \left(\phi(Y_i) - g(X_i)\right)^2$$

Benchmark

Train data		
.....		

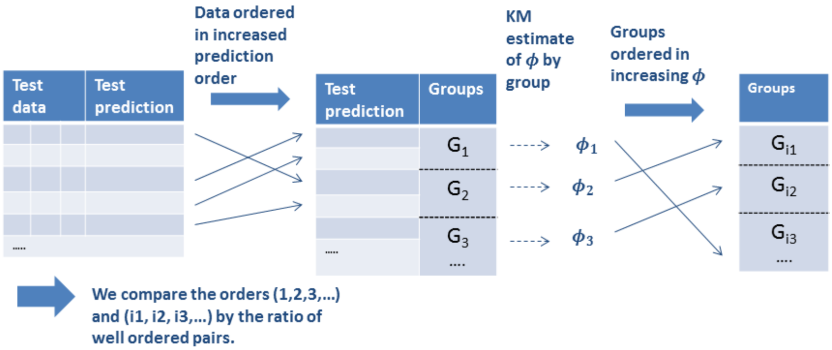


$\widehat{S}_T(.|X)$: Cox, RSF

$\widehat{\phi} = -\int \phi \, d\widehat{S}_T(.|X)$:

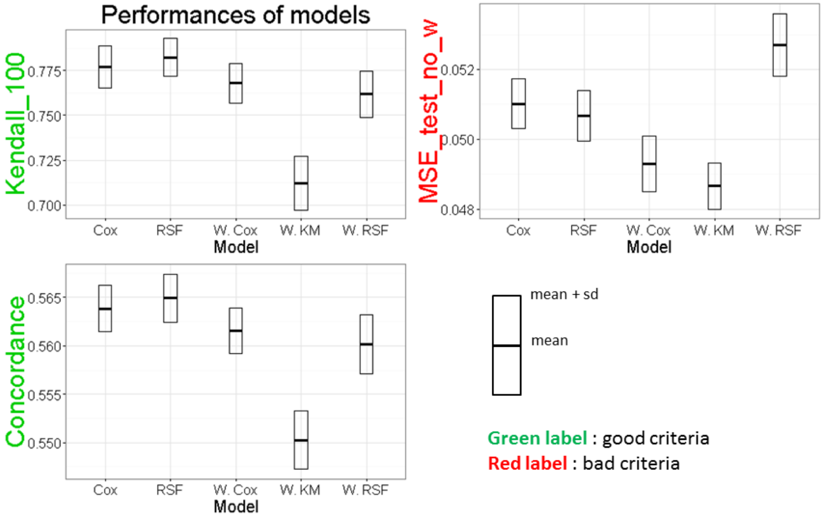
prediction of random forest

New criteria to compare model



- We cut the test set by groups of 100 obs. We call "Kendall_100" the ratio of well ordered pairs among all pairs of groups.

Results



Green label : good criteria

Red label : bad criteria

Summary

- We can adapt the Random Forest algorithm to the case where the target Y is censored using IPCW principle.
- We get better results using conditional weights $\hat{W}_i = \frac{\delta_i}{\hat{S}_C(Y_i|X_i)}$ (Cox, RSF) rather than non-conditional $\frac{\delta_i}{\hat{S}_C(Y_i)}$ (Kaplan Meier)
- Our weighted RF method didn't achieve as good performances as our benchmarks on this data.
- Outlook
  - Study the method in a high dimension setting
  - Theoretical study of the consistency of the method

Thank you for listening

For Further Reading I

Lopez, Olivier and Milhaud, Xavier and Thérond, Pierre-Emmanuel
Tree-based censored regression with applications to insurance
2015

Molinaro, Annette M and Dudoit, Sandrine and Van der Laan, Mark J
Tree-based multivariate regression and density estimation with right-censored data
2004, Journal of Multivariate Analysis, p. 154–177

Ishwaran, Hemant and Kogalur, Udaya B and Blackstone, Eugene H and Lauer, Michael S
Random survival forests
2008, The annals of applied statistics

For Further Reading II

Hothorn, Torsten and Bühlmann, Peter and Dudoit, Sandrine and Molinaro, Annette and Van Der Laan, Mark J
Survival ensembles
2006, Biostatistics

Biau, Gérard and Scornet, Erwan
A random forest guided tour
2016, Test