Random Forest for Regression of a Censored Variable

Yohann Le Faou

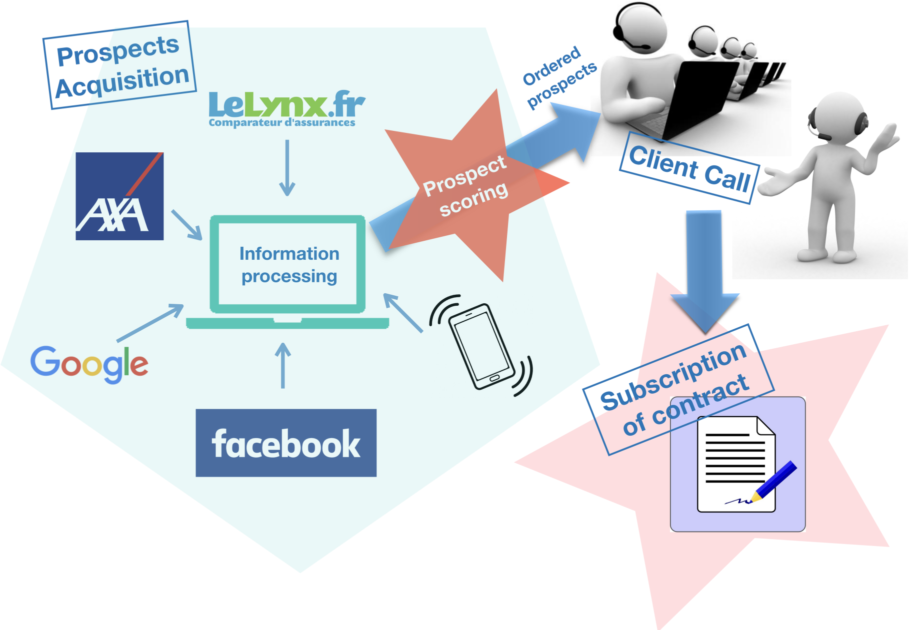
July 4, 2017

Insurance, Mathematics and Economics 2017

Table of contents

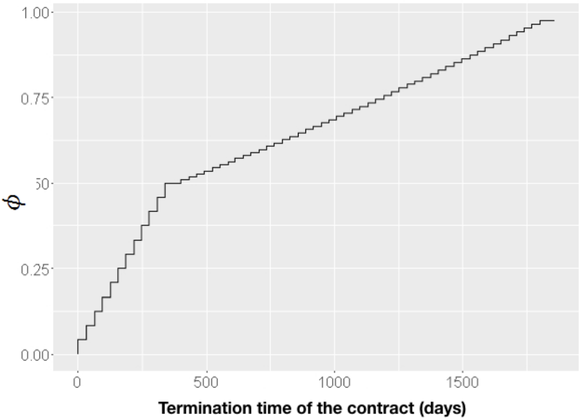
Introduction

Imagine if you were an insurance broker



Commissioning

ϕ : Commissioning function of the insurance broker (per unit of annual premium)



Mathematical Formulation

- T : Termination time of the contract (may be censored)
- 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 :

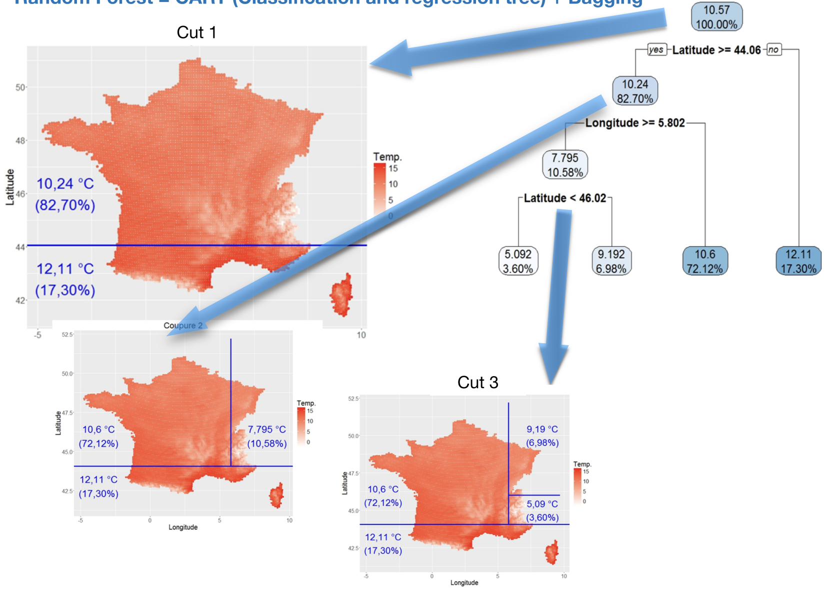
- $Y = min(T, C)$
- $\delta = \mathbb{1}_{T \leq C}$

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

Weighted Random Forest and IPCW method

Random Forest

Random Forest = CART (Classification and regression tree) + Bagging



To solve our problem

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

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

$\Longrightarrow$ Need an estimate of $E\left[(\phi(T) - g(X))^2\right]$ 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

IPCW principle

- IPCW : Inverse Probability of Censoring Weighting

IPCW principle]

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

Then for any bounded function ψ,

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

Reminder

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

IPCW principle

Proof

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

Estimation of the weights

Hypothesis

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

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

Weighted Random Forest

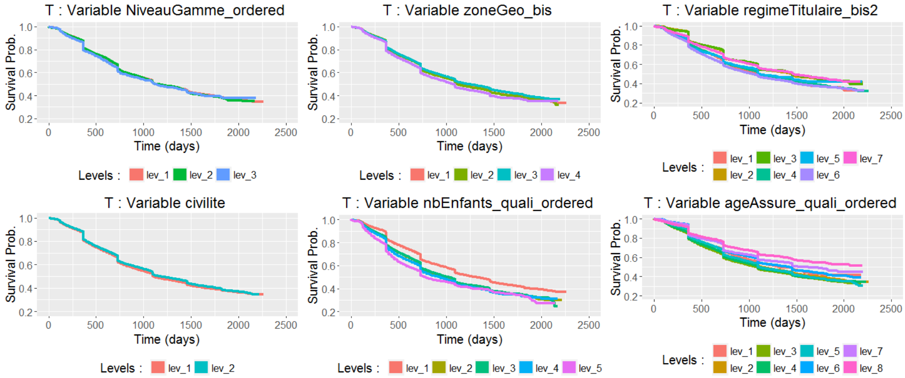
- Let $\hat{W}_i = \frac{\delta_i}{\hat{S}_C(Y_i)}$ or $\frac{\delta_i}{\hat{S}_C(Y_i|X_i)}$
- We estimate $E[(\phi(T) - g(X))^2]$ by

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

- **Weights are taken into account in the bootstrap of the Random Forest**

Experiments

Survival curves by subgroup of individuals



Setting of the Experiments

Weighted RF

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

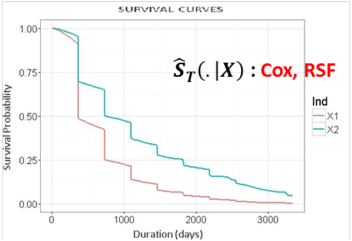
$$\hat{S}_C(.\,|X) \begin{cases} \color{red}{KM} \\ \color{red}{RSF} \\ \color{red}{Cox} \end{cases}$$

$\hat{\phi} = \hat{f}(x)$: prediction of the random forest

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

Alternative methods

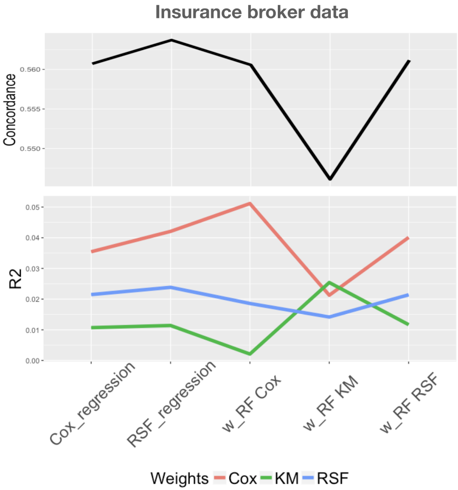
Train data
.....

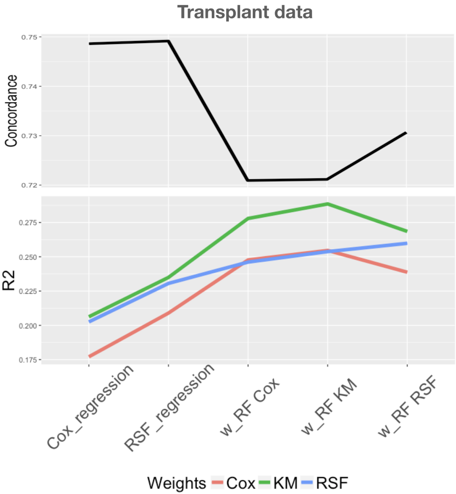


$\hat{S}_T(.\,|X)$: Cox, RSF

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

Results


Insurance broker data
Concordance
0.560
0.555
0.550
R2
0.05
0.04
0.03
0.02
0.01
0.00
Cox_regression
RSF_regression
w_RF Cox
w_RF KM
w_RF RSF
Weights Cox KM RSF


Transplant data
Concordance
0.75
0.74
0.73
0.72
R2
0.275
0.250
0.225
0.200
0.175
Cox_regression
RSF_regression
w_RF Cox
w_RF KM
w_RF RSF
Weights Cox KM RSF

Summary

- We can adapt the Random Forest algorithm to the case where the target Y is censored using IPCW principle.
- Weighted Random Forest is competitive with other standard methods

Outlook

- Implementation of the method in a R package
- Theoretical study of the consistency of the method

Acknowledgments

Thanks to Olivier Lopez (UPMC, Paris, France), Guillaume Gerber (Forsides, Paris, France) and Michael Trupin (Groupe Santiane, Paris, France)

Thank you for listening

mail : yohannlefa@gmail.com