# Tree-based estimators and actuarial applications

Lyon-Columbia Workshop (Lyon), 06/27/2016

Xavier Milhaud
Joint work with O. Lopez and P. Thérond

## Two problems with censoring - Lifetime / Claim amount

- Estimate some individual lifetime $T$ given features $X \in \mathbb{R}^d$,
- Only observe the follow-up time $Y$ : censored observation.

- The claim is still opened and has been under payment for a time $Y$ (the claim is not closed).
- The total claim amount $M$ is still unknown : just paid $N \leq M$.
- $M$ to predict (or total claim lifetime $T$) from $X \in \mathbb{R}^d$.

## Clustering by trees : key components

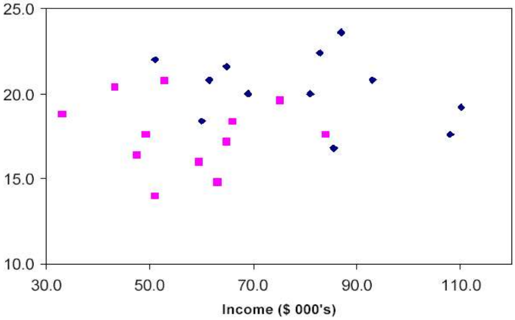
To estimate our quantity of interest, use a tree approach where :

1. the root : whole population to segment $\Rightarrow$ starting point ;
2. the branches : correspond to splitting rules ;
3. the leaves : homogeneous disjoint subsamples of the initial population, give the estimation of the quantity of interest.

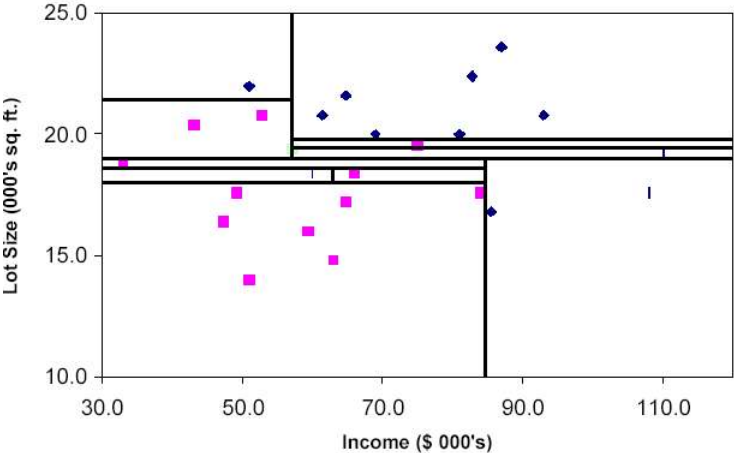
A reference in actuarial sciences $\rightarrow$ [Olb12] : builds experimental mortality tables of a reinsurance portfolio by predicting death rates.

# Example : predicting owner status | income and size

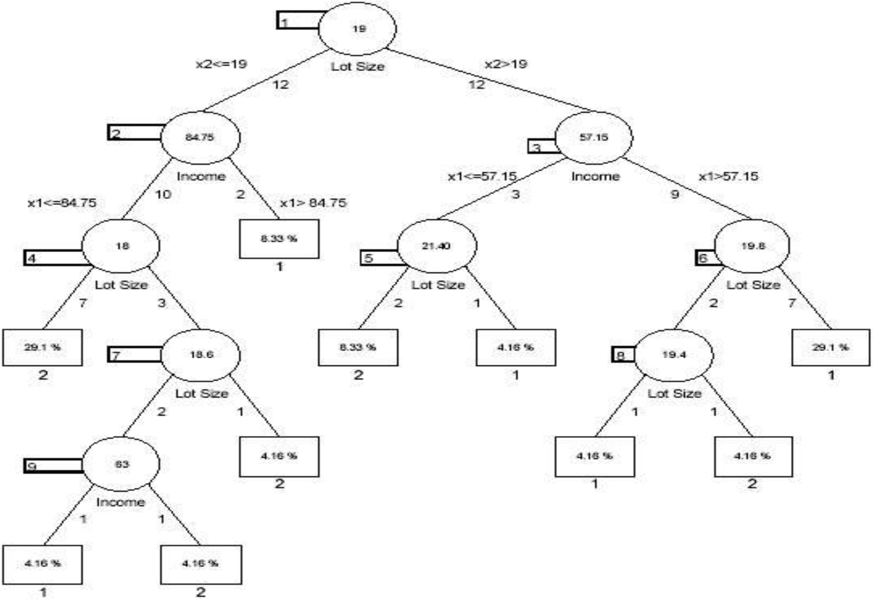
| Income ($ 000's) | Lot Size (000's sq. ft.) | Owners=1, Non-owners=2 |
|---|---|---|
| 60 | 18.4 | 1 |
| 85.5 | 16.8 | 1 |
| 64.8 | 21.6 | 1 |
| 61.5 | 20.8 | 1 |
| 87 | 23.6 | 1 |
| 110.1 | 19.2 | 1 |
| 108 | 17.6 | 1 |
| 82.8 | 22.4 | 1 |
| 69 | 20 | 1 |
| 93 | 20.8 | 1 |
| 51 | 22 | 1 |
| 81 | 20 | 1 |
| 75 | 19.6 | 2 |
| 52.8 | 20.8 | 2 |
| 64.8 | 17.2 | 2 |
| 43.2 | 20.4 | 2 |
| 84 | 17.6 | 2 |
| 49.2 | 17.6 | 2 |
| 59.4 | 16 | 2 |
| 66 | 18.4 | 2 |
| 47.4 | 16.4 | 2 |
| 33 | 18.8 | 2 |
| 51 | 14 | 2 |
| 63 | 14.8 | 2 |



## Partition and tree : maximal global homogeneity




Create subspaces maximizing homogeneity within each partitions.



1
19
Lot Size
x2<=19
x2>19
12
12
2
84.75
Income
3
57.15
Income
x1<=84.75
10
2
x1> 84.75
8.33 %
1
x1<=57.15
3
x1>57.15
9
4
18
Lot Size
5
21.40
Lot Size
6
19.8
Lot Size
7
3
2
1
2
7
29.1 %
2
7
18.6
Lot Size
8.33 %
2
4.16 %
1
8
19.4
Lot Size
29.1 %
1
2
1
1
1
9
83
Income
4.16 %
2
4.16 %
1
4.16 %
2
1
1
4.16 %
1
4.16 %
2

2. Building the tree - steps
   - Building steps to estimate the expectation
   - Stopping rules
   - Pruning criterion

## Regression trees : $Y$ continuous and fully observed

Regression problem :

$$\pi_0(\mathbf{x}) = E_0[Y \,|\, \mathbf{X} = \mathbf{x}] \tag{1}$$

$\rightarrow$ Most famous option : linear relationship b/w $Y$ and $X$ (limit ourselves to a given class of estimator) $\Rightarrow$ mean squared error.
$\rightarrow$ In full generality, we cannot consider all potential estimators of $\pi_0(\mathbf{x})$ $\Rightarrow$ trees are *another class* : piecewise constant functions.

**Building a tree provides a sieve of estimators**, obtained from successive splits of covariate space $\mathcal{X}$.

## CART estimator : a piecewise constant estimator

$$\hat{\pi}(\mathbf{x}) := \hat{\pi}^L(\mathbf{x}) = \sum_{l=1}^{L} \hat{\gamma}_l \, R_l(\mathbf{x}) \tag{2}$$

- $L$ is the number of leaves for the tree, $l$ its index,
- $R_l(\mathbf{x}) = \mathbb{1}(\mathbf{x} \in \mathcal{X}_l)$ : splitting rule,
- $\hat{\gamma}_l = E_n[Y \,|\, \mathbf{x} \in \mathcal{X}_l]$ : empirical mean of $Y$ in leaf $l$,
- The partitions $\mathcal{X}_l \subseteq \mathcal{X}$ are
  - disjoints ($\mathcal{X}_l \cap \mathcal{X}_{l'} = \emptyset$, $l \neq l'$),
  - exhaustive ($\mathcal{X} = \cup_l \mathcal{X}_l$).

This (piecewise constant) form can be generalized whatever the quantity of interest (expectation, median, ...).

## Building the tree : splitting criterion

→ Must be suitable to our task.
→ To solve (1), *OLS* are used since the solution is given by

$$\pi_0(\mathbf{x}) = \arg\min_{\pi(\mathbf{x})} E_0\Big[\phi(T, \pi(\mathbf{x})) \,|\, \mathbf{X} = \mathbf{x}\Big] \tag{3}$$

where $\phi(T, \pi(\mathbf{x})) = (T - \pi(\mathbf{x}))^2$ ($\phi$ loss function)

→ Here, results in minimizing the intra-node variance at each step.

→ If $T$ is fully observed, building the regression tree with this criterion is consistent ([BFOS84]).

## Pruning : penalize by tree complexity

CART principle : do not stop the splitting process, and buid the *"maximal"* tree (size $K(n)$), then prune it.
$\rightarrow$ We get a sieve of estimators $(\hat{\pi}^K(\mathbf{x}))_{K=1,\ldots,K(n)}$.

Avoid overfitting $\Rightarrow$ find the best subtree of the maximal tree, with a trade-off betwwen good fit and complexity :

$$R_\alpha(\hat{\pi}^K(\mathbf{x})) = E_n[\,\Phi(Y, \hat{\pi}^K(\mathbf{x}))\,] + \alpha\,(K/n).$$

If $\alpha$ fixed, the final estimator (pruned tree) yields

$$\boxed{\hat{\pi}^K_\alpha(\mathbf{x}) = \underset{(\hat{\pi}^K)_{K=1,\ldots,K(n)}}{\arg\min} \; R_\alpha(\hat{\pi}^K(\mathbf{x})).} \tag{4}$$

3. Extend to (potentially) censored data

## Back to our data

We observe a sample of i.i.d. random variables $(Y_i, N_i, \delta_i, X_i)_{1 \leq i \leq n}$ with same distribution $(Y, N, \delta, X)$, where

$$\left\{ \begin{array}{lcl} Y & = & \inf(T, C), \\ N & = & \inf(M, D), \end{array} \right.$$

and

$$\delta = \mathbf{1}_{T \leq C} = \mathbf{1}_{M \leq D}.$$

- $C$ et $D$ are the censoring variables, for instance :
  - $C =$ time b/w the declaration date and the extraction date ;
  - $D =$ current amount paid for this claim.

## Focus on lifetime $T$ : what we would like to do

In practice, we only observe i.i.d. replications $(Y_i, \delta_i, \mathbf{X}_i)_{1 \le i \le n}$ where

$$\begin{cases} Y & = & \inf(T, C) \\ \delta & = & \mathbf{1}_{T \le C} \end{cases}$$

- Current lifetime $Y$, not closed : $\delta = 0$.
- We seek
$$T^* = E\left[T \mid \delta = 0, Y, \mathbf{X}\right].$$
- Goal : find an estimator of $T^*$ from observations.

Pitfalls : we do not observe i.i.d. replications of $M \Rightarrow$ standard methods do not apply (LLN).

## Ingredients : Kaplan-Meier estimator and IPCW

- Assume that $T$ is independent from $C$.
- Define :
$$\hat{F}(t) = 1 - \prod_{Y_i \le t} \left( 1 - \frac{\delta_i}{\sum_{j=1}^n \mathbf{1}_{Y_j \ge Y_i}} \right).$$
- This estimator tends to $F(t) = \mathbb{P}(T \le t)$.

- Additive version : $\hat{F}(t) = \sum_{i=1}^n W_{i,n} \mathbf{1}_{Y_i \le t}$,
where
$$W_{i,n} = \frac{\delta_i}{n[1 - \hat{G}(Y_i -)]},$$
with $\hat{G}(t)$ the Kaplan-Meier estimator of $G(t) = \mathbb{P}(C \le t)$.

## Why does it work ?

1. Recall that $W_{i,n} = \frac{1}{n}\frac{\delta_i}{1-\hat{G}(Y_i-)}$ is "close" to $W^*_{i,n} = \frac{1}{n}\frac{\delta_i}{1-G(Y_i-)}$.
2. Moreover (LLN),

$$\sum_{i=1}^{n} W^*_{i,n}\phi(Y_i) = \frac{1}{n}\sum_{i=1}^{n}\frac{\delta_i\phi(Y_i)}{1-G(Y_i-)} \to_{p.s.} E\left[\frac{\delta\phi(Y)}{1-G(Y-)}\right].$$

### Proposition

For all function $\phi$ such that $E[\phi(T)] < \infty$,

$$E\left[\frac{\delta\phi(Y)}{1-G(Y-)}\right] = E[\phi(T)].$$

## Application to our context

Would like to estimate quantities like $E[\phi(T, X)]$ (see eq. (3)).

### Proposition

Assume that :

- $C$ is independent from $(T, X)$;

Then

$$E\left[\frac{\delta\phi(Y, X)}{n(1 - G(Y-))}\right] = E[\phi(T, X)],$$

and

$$E\left[\frac{\delta\phi(Y, X)}{n(1 - G(Y-))} \mid X\right] = E[\phi(T, X) \mid X].$$

- Thus to estimate $E[\phi(T, X)]$, we use

$$\frac{1}{n}\sum_{i=1}^{n}\frac{\delta_i\phi(Y_i, X_i)}{1-\hat{G}(Y_i-)} = \sum_{i=1}^{n} W_{i,n}\phi(Y_i, X_i).$$

- Therefore, to estimate quantities like

$$E\left[(\phi(T_i) - a)^2 \mathbf{1}_{X_i\in\mathcal{X}}\right],$$

where $\mathcal{X}$ is a subspace, we compute

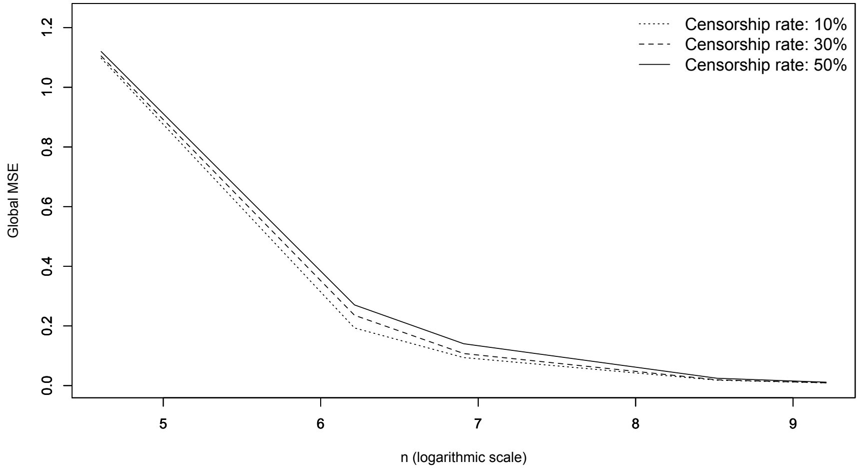
$$\boxed{\sum_{i=1}^{n} W_{i,n}(\phi(Y_i) - a)^2 \mathbf{1}_{X_i\in\mathcal{X}}.}$$

## Quality of our CART estimator : simulation study

Consider the following simulation scheme :

1. draw $n+v$ *iid* replications $(\mathbf{X}_1, ..., \mathbf{X}_n)$ of the covariate, with $\mathbf{X}_i \sim \mathcal{U}(0,1)$ ;
2. draw $n+v$ *iid* lifetimes $(T_1, ..., T_n)$ following an exponential distribution such that $T_i \sim \mathcal{E}(\beta = \alpha_1 \mathbb{1}_{\mathbf{X}_i \in [a,b[} + \alpha_2 \mathbb{1}_{\mathbf{X}_i \in [b,c[} + \alpha_3 \mathbb{1}_{\mathbf{X}_i \in [c,d[} + \alpha_4 \mathbb{1}_{\mathbf{X}_i \in [d,e]})$. (notice that there thus exist four subgroups in the whole population)
3. draw $n+v$ *iid* censoring times, Pareto-distributed : $C_i \sim \mathcal{P}areto(\lambda, \mu)$ ;
4. from the simulated lifetimes and censoring times, get for all $i$ the actual observed lifetime $Y_i = \inf(T_i, C_i)$ and the indicator $\delta_i = \mathbf{1}_{T_i \leq C_i}$ ;
5. compute the estimator $\hat{G}$ from the whole generated sample $(Y_i, \delta_i)_{1 \leq i \leq n+v}$.

| % of censored observations | Sample size $n$ | Group-specific MWSE Group 1 MWSE | Group 2 MWSE | Group 3 MWSE | Group 4 MWSE | Global MWSE |
|---|---|---|---|---|---|---|
| 10% | 100 | 0.19516 | 0.42008 | 0.17937 | 0.30992 | ***1.10454*** |
| | 500 | 0.03058 | 0.07523 | 0.03183 | 0.06029 | ***0.19796*** |
| | 1 000 | 0.01509 | 0.03650 | 0.01517 | 0.02619 | ***0.09306*** |
| | 5 000 | 0.00295 | 0.00714 | 0.00289 | 0.00530 | ***0.01804*** |
| | 10 000 | 0.00105 | 0.00378 | 0.00117 | 0.00292 | ***0.00910*** |
| 30% | 100 | 0.20060 | 0.43664 | 0.17448 | 0.29022 | ***1.10765*** |
| | 500 | 0.03736 | 0.07604 | 0.04301 | 0.06584 | ***0.22217*** |
| | 1 000 | 0.01748 | 0.04095 | 0.01535 | 0.02674 | ***0.10043*** |
| | 5 000 | 0.00319 | 0.00758 | 0.00291 | 0.00547 | ***0.01904*** |
| | 10 000 | 0.00117 | 0.00372 | 0.00125 | 0.00292 | ***0.00930*** |
| 50% | 100 | 0.19784 | 0.45945 | 0.17387 | 0.28363 | ***1.11476*** |
| | 500 | 0.04906 | 0.08993 | 0.05301 | 0.06466 | ***0.25668*** |
| | 1 000 | 0.02481 | 0.05115 | 0.01788 | 0.03004 | ***0.12387*** |
| | 5 000 | 0.00520 | 0.00867 | 0.00389 | 0.00516 | ***0.02299*** |
| | 10 000 | 0.00153 | 0.00407 | 0.00162 | 0.00308 | ***0.01057*** |


Censorship rate: 10%
Censorship rate: 30%
Censorship rate: 50%
Global MSE
n (logarithmic scale)
0.0
0.2
0.4
0.6
0.8
1.0
1.2
5
6
7
8
9

4 Applications

## Application 1 : income protection

We refer to short-term disability contracts over 6 years with the following information :

- 83 547 claims ;
- PH ID, cause (sickness or accident), gender, SPC, age, duration in disability state (censored or not), distribution channel ;
- the censoring rate equals 7.2% ;
- mean lifetime in disability state : 100 days.

Goal : find a segmentation to predict how much time the disability state lasts.

## Tree estimator : the age at claim seems to be key

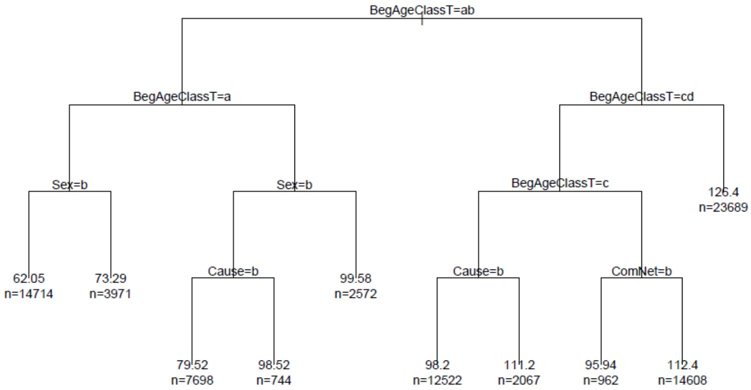



FIGURE: Disability duration explained by sex, SPC, network, age, cause.

Usually, the recovery rates used to compute technical provisions for this guarantee depends on the age at the claim date due to local prudential regulation ⇒ we fit a Cox PH with this covariate :

- leads to consider the high predictive power of this variable ;
- PH assumption rejected by all tests (LR, Wald and log-rank) ;
- obtained results will be considered as benchmarks to enable a comparison with those resulting from the tree approach.

| Classes | Mean Age | Tree | Cox |
|---|---|---|---|
| a | 26.83 | 64.44 | 80.01 |
| b | 34.19 | 85.48 | 96.35 |
| c | 39.57 | 100.04 | 110.19 |
| d | 45.05 | 111.38 | 126.03 |
| e | 51.29 | 126.40 | 146.28 |

TABLE: Expected disability time (days) depending on age at disability time.

→ We observe significant differences between Tree / Cox estimates.

→ These differences can be explained by two phenomena resulting from using the Cox proportional-hazards model :

- the estimation of the baseline hazard is very sensitive to highest disability durations (mainly concentrated in class *e*).
  → affect the estimates of all other classes ;

- our approach directly target the duration expectation while Cox partial-likelihood is focused on estimating the hazard rate.

## Application 2 : reserving

Seek $$E\left[M \mid \delta = 0, X, Y, N\right]$$

Get back to quantities only conditioned by covariates **X** :

$$\begin{aligned} E\left[M \mid \delta = 0, X = x, Y = y, N = n\right] &= E\left[M \mid M \geq n, T \geq y, X = x\right] \\ &= \frac{E\left[M\,\mathbf{1}_{M \geq n, T \geq y} \mid X = x\right]}{\mathbb{P}(T \geq y, M \geq n \mid X = x)}. \end{aligned}$$

Define $\phi_1(t, m) = m\mathbf{1}_{m \geq n, T \geq y}$, $\phi_2(t, m) = \mathbf{1}_{t \geq y, m \geq n}$.
Estimate the ratio of

(1) $E[\phi_1(T, M) \mid X = x]$ over (2) $E[\phi_2(T, M) \mid X = x]$.

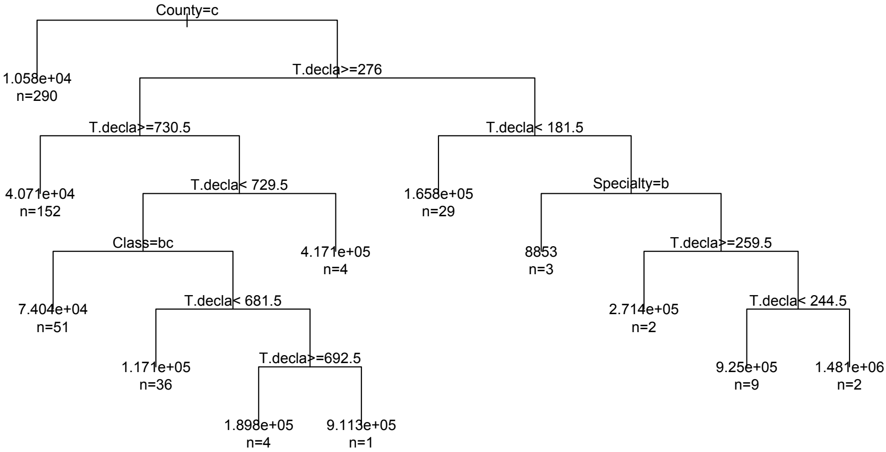
## Our data

Third-party insurance in medical field in US, with 648 claims and various individual characteristics (specialty, class, county, reopen status, ...) with large heterogeneity.

```
   Claim.entry Indemn.res  ALAE.res  (..) Cens. Already.paid Reserved
47  2000-07-14   0         0.00             1         3456        0
48  2000-07-24   5000     13880.25          0       138435    18880
49  2000-07-31   5000     11304.60          0         7300    16305
50  2000-07-31   5000    103471.31          0       118136   108471
51  2000-08-04   0         0.00             1        46587        0
52  2000-08-14   0         0.00             1         3083        0
53  2000-08-15   0         0.00             1            0        0
54  2000-08-28   0         0.00             1          980        0

> summary(myData$Observed.total)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
      0       0    2644   41760   18500 1557000
> (tx.censure.learning) ;  (tx.censure.validation)
[1] 32.19178                 [1] 34.375
```

## Predictions of quantity (1) : $E[M1_{(M>n,T>y)} \,|\, X = x]$ Pruned survival tree



## Predictions of quantity (2) : P(M > n, T > y | X = x) Pruned survival tree, numerical results

Error of the tree:

```
> (1.0 - (confusion.matrix[1,1]+confusion.matrix[2,2]) / sum(confusion.matrix))*
> cat("The test sample estimate of the prediction error in the pruned tree is",
The test sample estimate of the prediction error in the pruned tree is 18.6 %
```

Predicted probabilities for the denominator:

| (..) Censure | Already.paid | Reserved | Observed.total | KM.weight | Proba.censorship |
|---|---|---|---|---|---|
| 1 | 24 | 0 | 24 | 0.0017 | 0.1496063 |
| 1 | 1844 | 0 | 1844 | 0.0017 | 0.1496063 |
| 1 | 444 | 0 | 444 | 0.0017 | 0.1935484 |
| 1 | 0 | 0 | 0 | 0.0017 | 0.1496063 |
| 1 | 3907 | 0 | 3907 | 0.00176 | 0.2307692 |
| 0 | 0 | 81000 | 0 | 0 | 0.7500000 |
| 0 | 1061 | 42139 | 1061 | 0 | 0.7400000 |
| 0 | 1061 | 79939 | 1061 | 0 | 0.2307692 |
| 0 | 1061 | 12439 | 1061 | 0 | 0.7400000 |

# Final ratio (1)/(2) and comparison to experts' opinions

```
> ##########################################################################
> ## Final prediction of total claim amount for censored claims.
> ##########################################################################
> ## Comparison b/w predictions from the tree and the one from the expert.

 Censure Already.paid Reserved Obs.total Adj.predicted.claims Expert.prediction
       0            0    81000         0             70752.37             81000
       0            0    71600         0             10585.00             71600
       0            0        0         0             10585.00                 0
       0            0    13500         0             10585.00             13500
       0            0    52700         0             55008.11             52700
       0            0     2500         0             10585.00              2500
       0            0    55500         0             70752.37             55500
       0            0    62100         0             55008.11             62100
       0            0    81000         0             54274.67             81000
       0         1061    42139      1061             55008.11             43200
       0         4266    57834      4266             70752.37             62100

> ## Difference in % (due also to absent expert' opinion leading to no reserve)
> (Reserve.gap <- round((abs(Tree.totalLumpSum.toReserve - Expert.totalLumpSum.
[1] 14.47   => It seems that experts have tendency to overestimate the reserve
```

## Final remarks

+ Can reveal to be a useful method for many applications, e.g. experimental mortality databases, ...
+ Simple and easy-to-understand final estimator.
+ Consistent procedure and theoretical guarantees.
+ Discriminating power of covariates.
+ Extensions by working on the loss function.

- Instability : need to gain robustness (random forests, ...).

# References


L. Breiman, J. Friedman, R. A. Olshen, and C. J. Stone.
*Classification and Regression Trees*.
Chapman and Hall, 1984.

Walter Olbricht.
Tree-based methods : a useful tool for life insurance.
*European Actuarial Journal*, 2(1) :129–147, 2012.


And our working paper :

`https://hal.archives-ouvertes.fr/hal-01141228/file/TreeCensoredRegression-LopezMilhaudTherond.pdf`