Longevity 16
13/08/2021

Model-based Recursive Partitioning for Mortality

Lou SEMI
SCOR
BRM & Knowledge Team

SCOR The Art & Science of Risk | Life

Executive summary

- [Introduction-Objective of the study]
- [Description of the algorithm used: GLM Tree]
- [Presentation of the data]
- [Results and discussion]
- [Conclusion-Perspective]

Introduction-Objective of the study

Company business may face adverse claims experience including higher than expected claims on key pockets of business. We propose to use a **model-based recursive partitioning approach** to:

- monitor these emerging areas of focus,
- understand underlying mortality drivers and,
- potential management actions can be used.

Main Objective: Identify the key segments that significantly deviate from the assumptions.

Generalized linear model combined with decision tree method is used for clustering pockets of business in terms of A/E.

The **application** focus on US industry data where deviations are identified relative to the standard 2015 VBT table.

Description of the algorithm used : GLM Tree

MOB algorithm

- MOB is a generic algorithm for model-based recursive partitioning (Zeileis, Hothorn, and Hornik 2008).
- Considering a parametric model M(Y, θ). This model could be a **normal** distribution for Y , a **psychometric** model for a matrix of responses Y , or some kind of **regression** model when Y = (y, x) can be split up into a dependent variable y and regressors x.(Zeileis, Hothorn 2014).
- Model-based recursive partitioning is used to partition data into groups that differ in terms of the parameters in the model.
- Rather than fitting one global model to a dataset, it estimates local models on subsets of data that are “learned” by recursively partitioning.
- The basic idea is to grow a tree in which every node is associated with a model of type M
- New “mobsters” dedicated to specific models, lmtree() and glmtree() for MOBs of (generalized) linear models exist.

Description of the algorithm used : GLM Tree

GLM Tree function

- GLM Tree is an extension of the MOB algorithm to obtain a more interpretable tree.
- The approach combines parametric models such as Generalized Linear Models with decision tree models.
- We choose the GLM method to facilitate the applications and in particular the logistic regression to facilitate the computations.

Main steps in the algorithm

1. Model and parameters estimation
2. Instability tests
3. Partitioning
4. Pruning

Description of the algorithm used : GLM Tree

MAIN STEPS IN THE ALGORITHM

1 Model and parameters estimation

- Fit the model once to all observations in the current node by minimizing some objective function.

$$\sum_{i=1}^{n} \Psi(Y_i, \theta)$$

- The estimation of the vector of parameters θ can be computed by solving the first order conditions:

$$\sum_{i=1}^{n} \psi(Y_i, \hat{\theta}) = 0, \quad \psi(Y, \theta) = \frac{\partial \Psi(Y, \theta)}{\partial \theta}$$

- The score function evaluated at the estimated parameters $\hat{\psi}_i = \psi(Y_i, \hat{\theta})$ is then inspected for systematic deviations.

2 Instability Tests

- To assess whether splitting of the node is necessary the general class of score-based fluctuation tests is employed.
- The test implemented differs depending on whether the partitioning variable is categorical or numerical.

.

- Zj with the minimal p-value is chosen for splitting the node.

Description of the algorithm used : GLM Tree

MAIN STEPS IN THE ALGORITHM

3 Partitioning

- For each conceivable split, the model is estimated on the two resulting subsets and the resulting objective functions are summed.
- The split that optimizes the segmented objective function is then selected as the optimal.

4 Pruning

- For determine the optimal size of the tree, one can either use a pre-pruning or post-pruning strategy.
- For the former, the algorithm stops when no significant parameter instablities are detected in the current node.
- For the latter, one would first grow a large tree and then prune back splits that did not improved the model judging by information criteria such as AIC or BIC.

Presentation of the data

Sources

- SOA aggregate data from 2003-2013.
- Data are from the ILEC members (18) and the MIB's Actuarial and Statistical Research Group.

Detailed data description

- From the original database with 26 millions rows and 33 variables, we worked with an 8 millions rows and 12 variables base.
- We choose the following variables:
  - Insurance Plan,
  - Duration,
  - Face amount band,
  - Attained age,
  - Risk class,
  - Smoker status,
  - Number of deaths,
  - Death claim amount,
  - Policies exposed,
  - Amount exposed,
  - Expected death, QX2015VBT by amount,
  - Expected Death,QX2015VBT by policy.

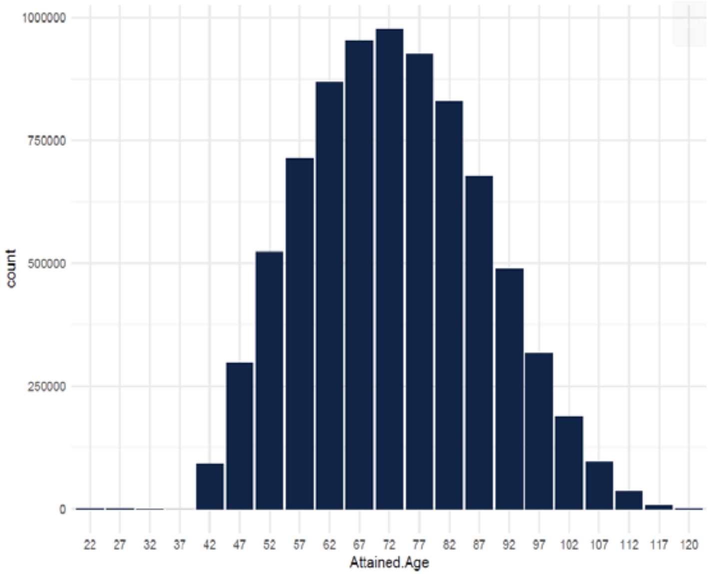
Presentation of the data

Detailed data treatment

- Delete rows with null expositions and missing values.
- Combine 3 variables : Preferred class, number of preferred classes and smoker status to construct the variable risk class smoker. Risk class smoker follows the order : "SS","PS","SNS","S+NS","PNS","SPNS“.
- Group attained age by quinquennial age.
- From 14 modalities of the variable Face amount to 6.
- From 25 modalities of the variable Duration to 6.

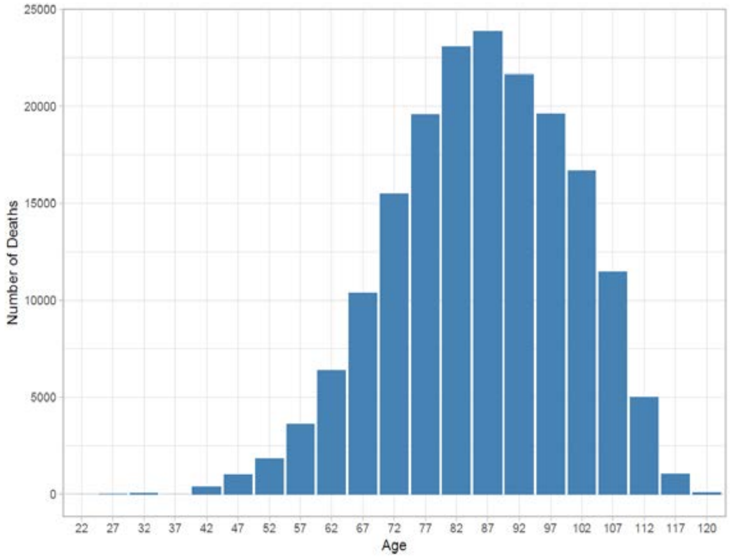
Presentation of the data

Descriptive Statistics : Age



Distribution of the overall data according to the age.

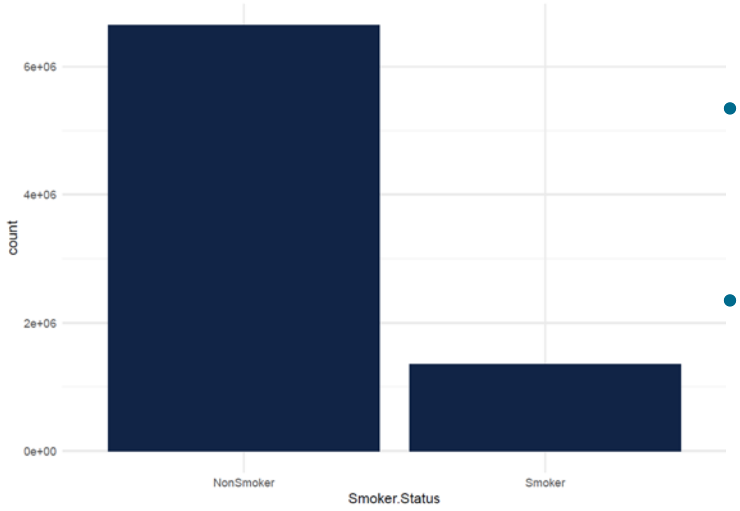
- Population quite old.
- Same distribution but different scales : <150000 deaths.
- Decreasing at old ages.



Distribution of the number of deaths according to the age.

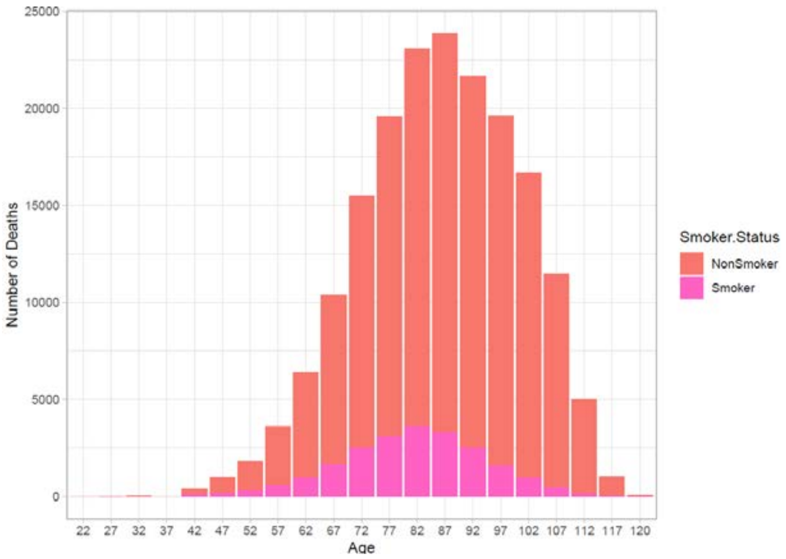
Presentation of the data

Descriptive Statistics : Smoker or not



Distribution of the overall data according to the smoker status.

- 78% of non-smoker in the overall population.
- => More deaths of non-smoker.

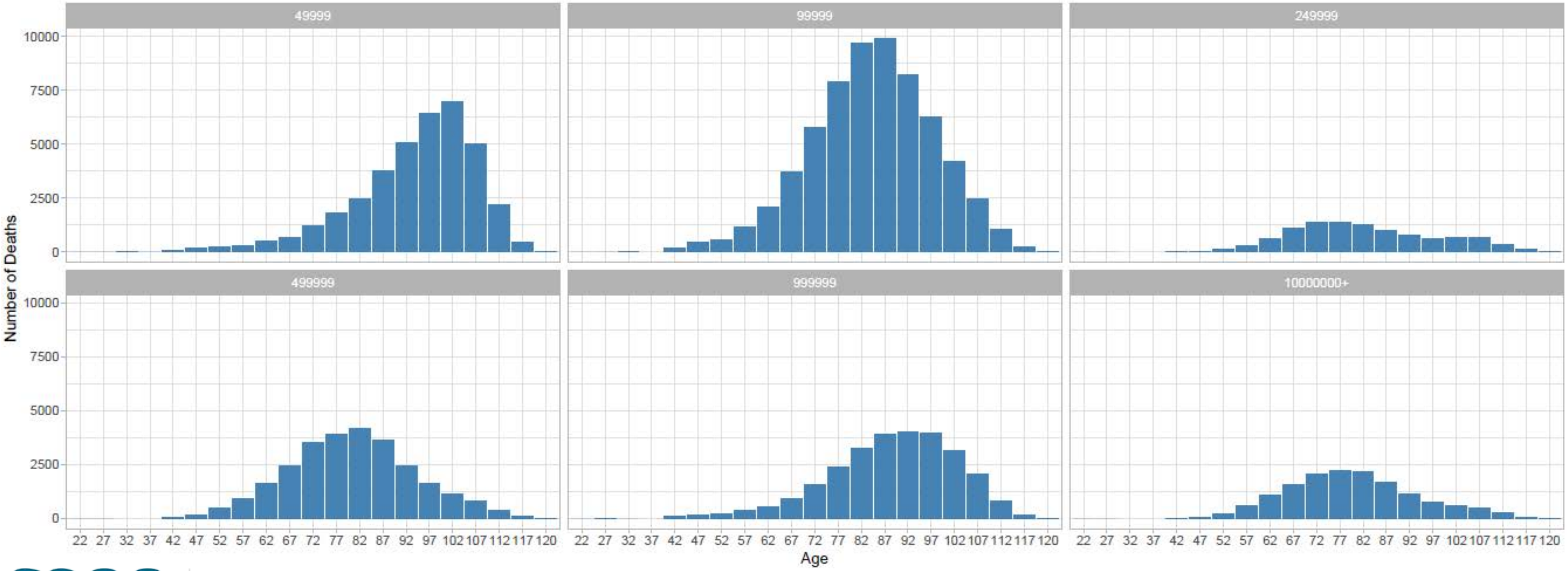


Distribution of the number of deaths according to the age and smoker status.

Presentation of the data

Descriptive Statistics : Face Amount

- Same distribution as previously.
- More deaths for face amount less than 99999.



Distribution of the number of deaths according to the age and face amount.

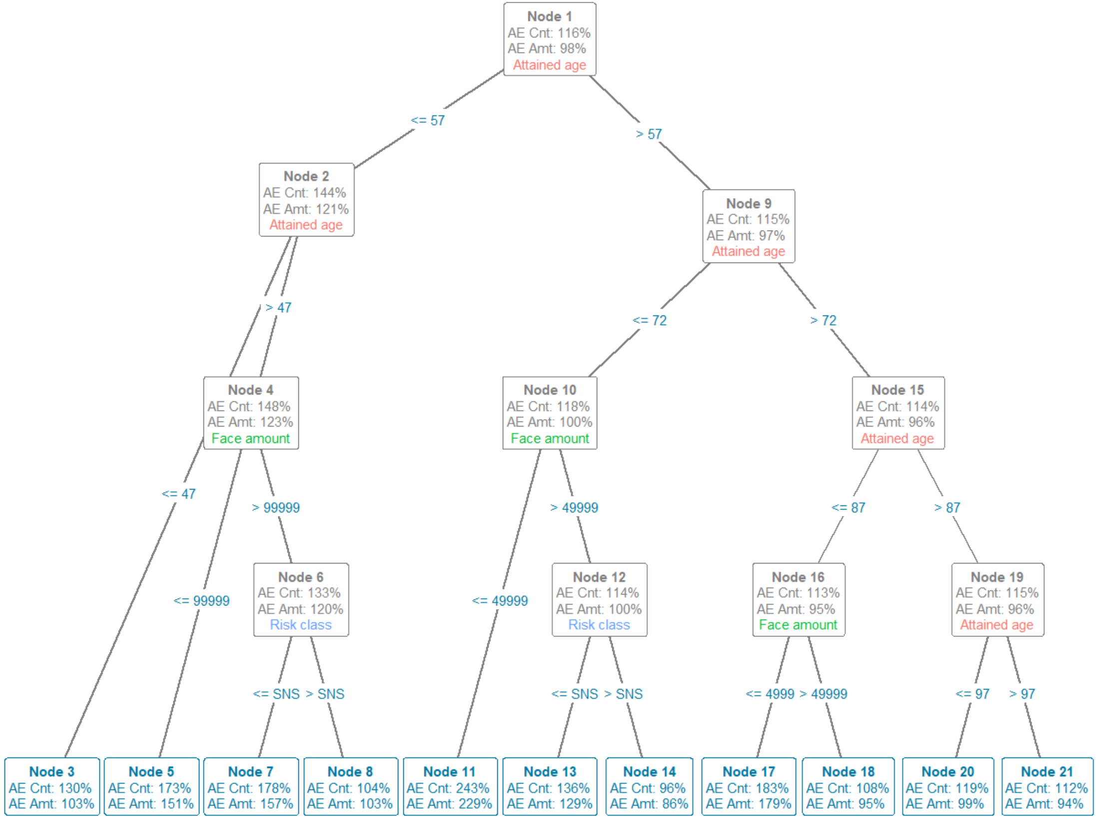
Results and discussion

The order of the splits shows which mortality differential are the most important.
Among, attained age, duration, risk class, face amount, insurance plan, we observe:

1. Attained age (x2)
2. Face amount
3. Risk class

Attained age appears first, indicating that this variable has the highest instability. We observe the splits for age 18-47, 48-57, 58-72 and 72+.

This highlights a different shape with age between industry's data and the standard 2015 VBT table.



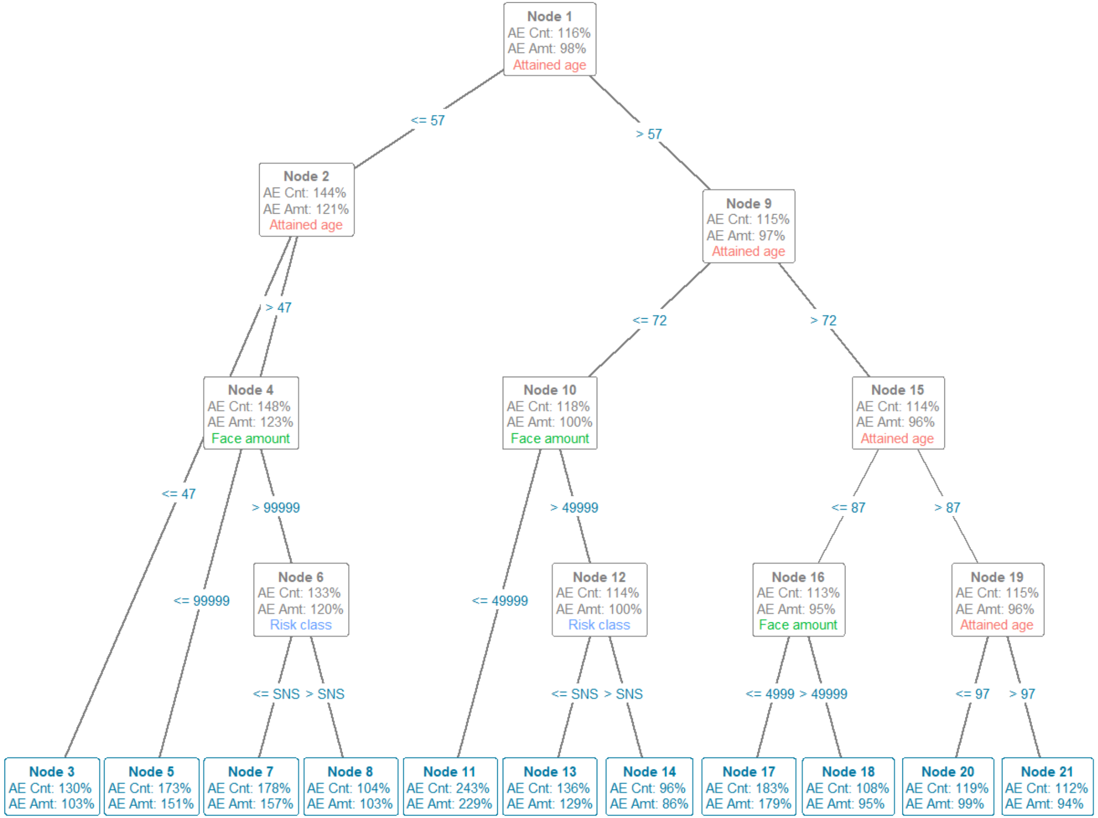
Results and discussion

- Face amount is also important leading to split between small and larger. amounts
- Risk class appears third and leads to a split between SM + standard NS and healthier classes (S+NS, PNS and SPNS)

Large deviations can be seen for:

1. The small amounts (nodes 5,11,17)
2. SM + standard NS and large amount (nodes 7,13)

- For extreme ages (<48 and >87), we observe deviations in count basis, but the table captures the amount effect as A/E in amount are relatively close to 100%



Results and discussion

A/E for males

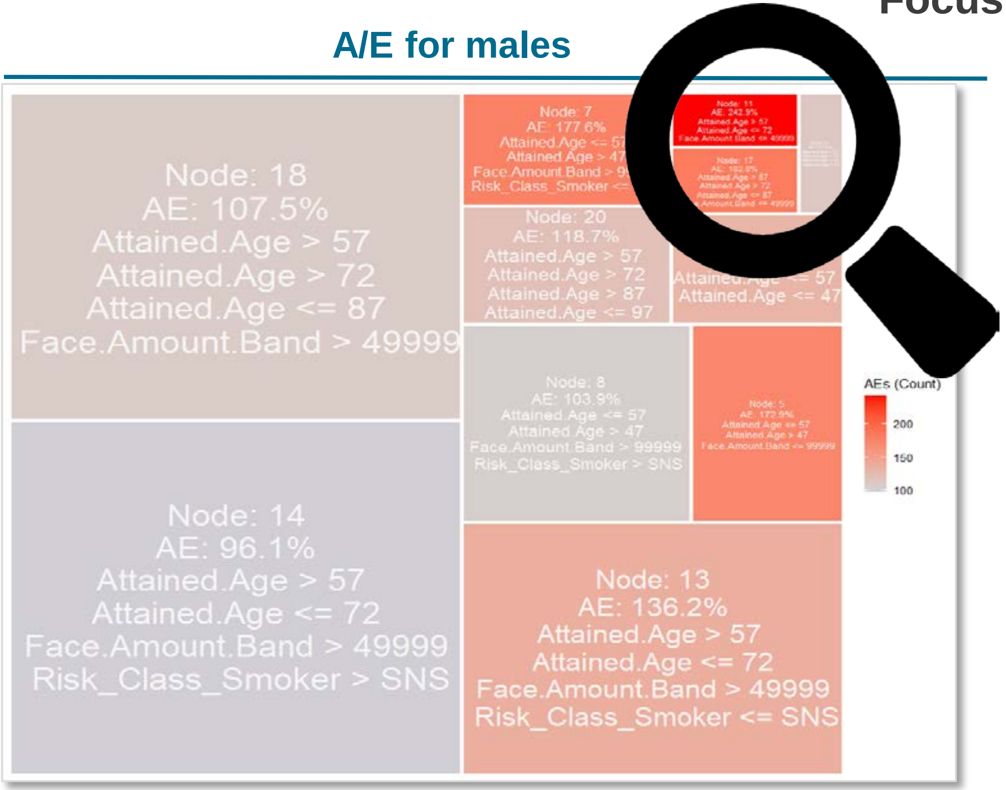
Discussion

- The size of the cell represents the size of the node.
- The color of the cell represents how good or bad we are compared to the standard 2015 VBT table: greyer is the cell, more we overestimate the mortality !
- AE < 100% => prudent ! ☺
- AE > 100% => high than expected claims ! ☹
- Nodes 14,18 and 8 can be considered as prudent. Theses nodes gather most of the observations. The standard table captures the mortality pattern adequately for these pockets of business.

Results and discussion

Focus on this side !

A/E for males

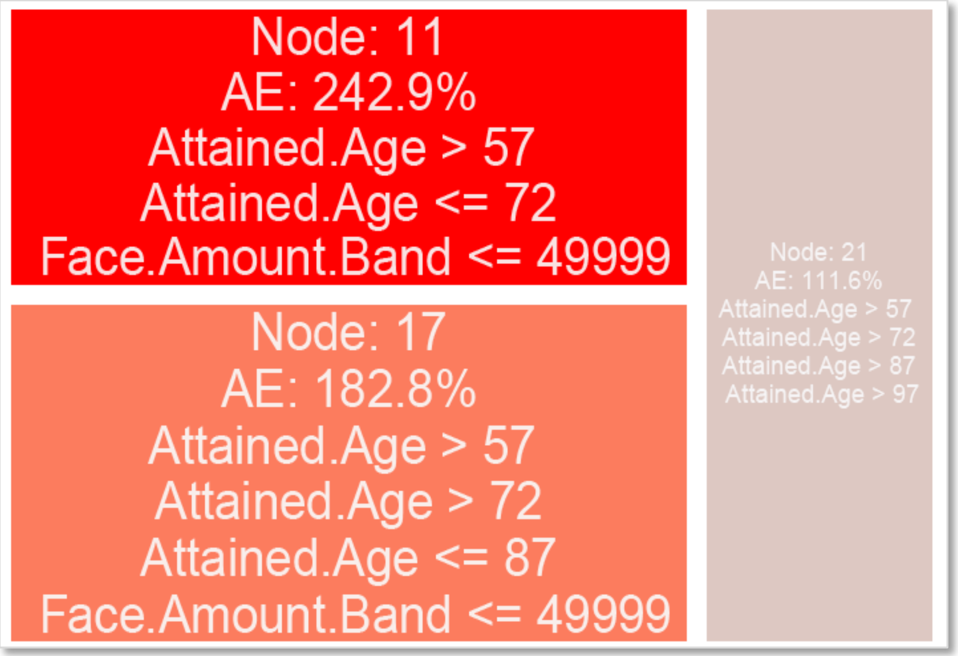


Discussion

- The size of the cell represents the size of the node.
- The color of the cell represents how good or bad we are compared to the standard 2015 VBT table: greyer is the cell, more we overestimate the mortality !
- AE < 100% => prudent ! ☺
- AE > 100% => high than expected claims ! ☹
- Nodes 14,18 and 8 can be considered as prudent. Theses nodes gather most of the observations. The standard table captures the mortality pattern adequately for these pockets of business.

Results and discussion

A/E for males

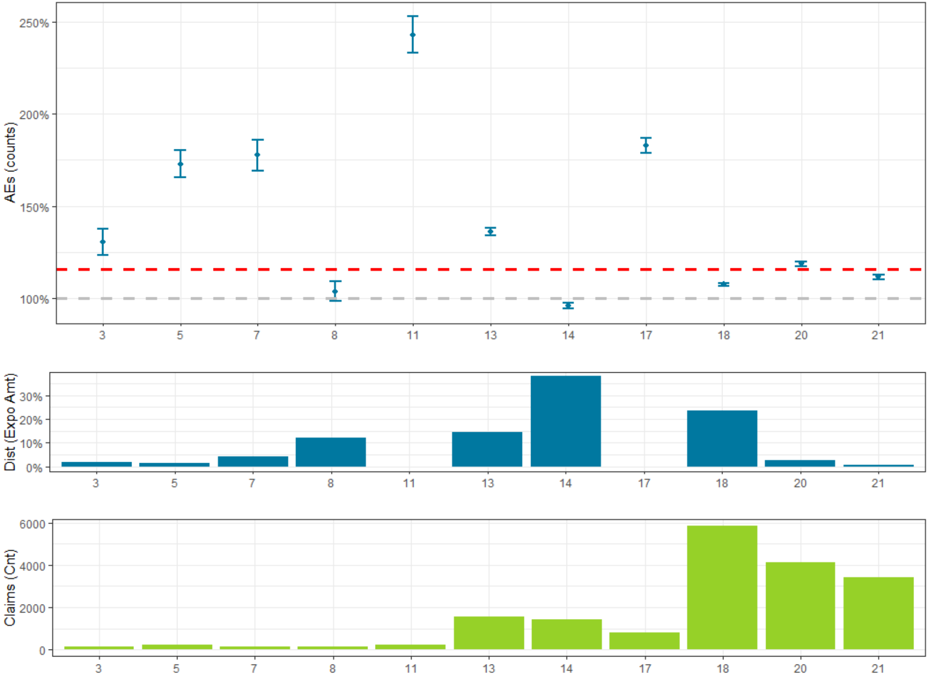


Discussion

- For small face amount, we observe large deviations. The count and amount affects are not well captured by the 2015 VBT table.
- For the node 11, we observe the highest deviation.
- The standard table captures the mortality for this segment of the population not adequately.
- For the oldest ages, we have fewer observations and relatively higher than expected claims.

Results and discussion

A/E Count

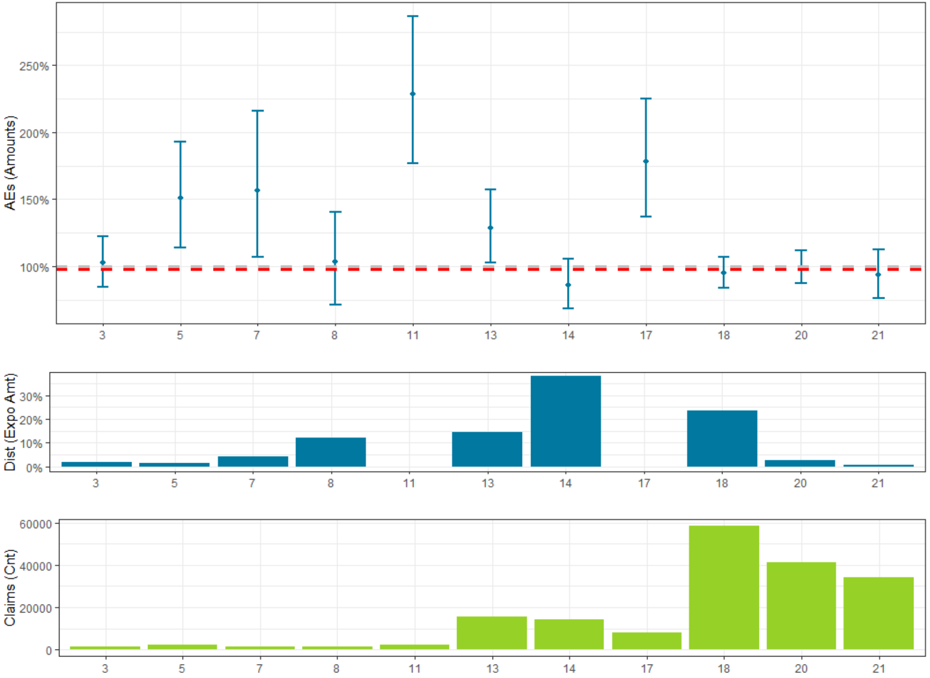


Discussion

- The grey dashed line illustrate the 100% A/E.
- The red dashed line is the 115% overall A/E in count.
- Node 11 and 17 (>57 and small amount) have the highest deviations however, exposed face amount and number of claims are small.
- Node 13 (Smoker with high amount) has a relatively high adverse claims experience and represents more than 10% of the face amount insured.

Results and discussion

A/E Amount



Discussion

- In an actuarial perspective, the studying A/E in Amount is important to refine the conclusions.
- We focus on nodes where there is no intersection between the CI and the 100% line
- Nodes 5 and 7 are above the 100% A/E but the uncertainties are as large as illustrated by the size of the confidence intervals and they represent very few claims and low face small amount exposed.
- Nodes 11 and 17, large deviations but very low face amount exposed.
- Node 13 with the CI just above the grey line, but with large amount and claims.
- For most of the nodes a deviation in count is observed, the table in amount captures the mortality pattern.

Conclusion-Perspective

- Identify the problematics segments which deviate from the assumptions : For any ages, high deviations for small face amount.

- We propose to use a model-based recursive partitioning approach to:
  monitor these emerging areas of focus,
  and potential management actions can be used.

- Same study done for women: same conclusion with the face amount.

- Perspective: Study the mortality trend instead of level.

Bibliography

- Zeileis A, Hothorn T, Hornik K (2008). “Model-Based Recursive Partitioning.” Journal of Computational and Graphical Statistics, 17(2), 492–514. doi:10.1198/106186008X319331.
- Zeileis A, Hothorn T, (2014).”Parties, Models, Mobsters: A new implementation of Model-based recursive partitioning in R”. Working Paper 2014-10, Working Papers in Economics and Statistics, Research Platform Empirical and Experimental Economics, Universität Innsbruck.