Data Science and Machine Learning

Example of application in Non Life Insurance

November 2016

Written by Julie, translated by Chloë,
Périclès Actuarial team members

Machine Learning in a P&C context : practical applications

How to use it despite internal constraints

The more the market **is transparent and driven by prices,** the more insurers have to adapt their strategy.

In P&C pricing, models have to be redesigned to better understand the risk and to improve the related segmentation : **innovation is a strategic challenge.**

If Machine Learning algorithms now offer a lot of possibilities and already proved in several studies their supremacy against conventional GLM models…

… **They remain little used in replacement of conventional GLM models**. It is especially due to IT constraints (unsuitable production system).

A way to use Machine Learning algorithms potential under the constraint of a fixed pricing model is to **exploit them to improve pricing variables segmentations and to identify new variables.**

Machine Learning in a P&C context : practical applications

Application in the case of a car insurance (1/2)

- In the case of a car insurance, **Machine Learning models are an innovating and powerful tool to improve segmentation and construct, for example, a vehicles classification** by homogeneous risk.

- The vehicles classification is a central issue for segmentation. Indeed, it allows the insurer to make the difference with the market and can, depending on the **construction method chosen**, have these **3 advantages** :
  - Improve the pricing model, by reducing, for example, the unexplained variance ;
  - Collecting information on the risk from all the technical caracteristics of the vehicle ;
  - Collecting information on the policyholder behaviour from the car he/she drives.

- One of the **key issues of a vehicle classification** is also to be able to integrate **new cars** or even cars that have been outside of the study scope during the construction.

Machine Learning in a P&C context : practical applications

Application in the case of a car insurance (2/2)

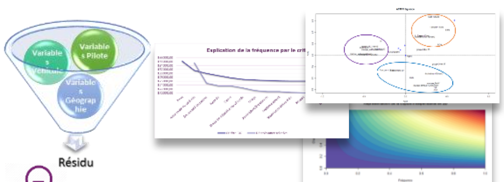
- Example of a method that has been developped to construct a vehicles classification. This classification reduces the unexplained variance in the case of a small database with <u>a strong heterogeneity</u> thanks to a Machine Learning algorithm : **the CART decision tree.**

PART 1 : Residual approach	PART 2 : Use of credibility to manage data heterogenity	PART 3 : Machine Learning to classify vehicles in homogeneous risks
Isolation of the risk part due to other factors than the vehicle risk part	Non conventional use of a credibility method to define cars with trustable information that will be used as a learning base for the Machine Learning algorithm.	Explication of residues of « credible » vehicles with car variables thanks to Machine Learning algorithms.

PART 1

GLM for severity and frequency

$$g(E[Y|X_1,\ldots,X_p]) = \beta_0 + \sum_{k=1}^{I} \beta_k X_k$$



Feedback

- Explanatory variables have to be selected considering the link between the driver and the car
- The choice of the parametric distribution and of the link function has to take into account the future use of residues
- A cost x frequency classification allows the obtention of a vehicle classification for each dimension and the comparison to the SRA classification

PART 2

Bühlmann-Straub model

$$\widehat{\mu(\theta_i)} = Z_i \bar{X}_i + (1 - Z_i)\mu_0$$

$$\mu_0 = \sum_{i=1}^{I} \frac{Z_i}{Z_\bullet} X_i$$

$$Z_i = \frac{w_{i\bullet}}{w_{i\bullet} + \frac{\sigma^2}{\eta^2}}$$

$$w_\bullet = \sum_{i}^{I} Z_i$$

Determination of the Bühlmann-Straub factor's limit from which vehicles are considered credible.

Feedback

- The credibility step improved the trees learning and the building of a more relevant classification
- The credibility factor limit and the method to define it have to be cautiously chosen by taking into account the information loss on the learning base

PART 3

CART regression

$$\gamma_n(u) = \frac{1}{n} \sum_{i=1}^{n} (Y_i - u(X_i))^2$$

$$\overline{Y}_{t_f} = \frac{1}{\#\{(X_i, Y_i) \in \mathcal{L}\ ;\ X_i \in t_f\}} \sum_{\{X_i\ ;\ X_i \in t_f\}} Y_i$$

Feedback

- The CART regression is the mean to isolate the noise signal and to directly create vehicles categories
- Decision trees have the advantage to create clear rules that will be used to classify future vehicles
- The classes number is at stake, the segmentation degree has to be kept in order to avoid the over-learning issue of Machine Learning algorithms