Data Science in a pricing process

Michaël Casalinuovo
Consultant, ADDACTIS® Software
michael.casalinuovo@addactis.com

addactis
Actuarial & Software Solutions

Contents

02 GLM

04 Machine learning methods

08 Mixture of methods

10 Conclusion

Nowadays, we live in a continuously changing market environment, Pricing has become a challenge. Non-life insurance companies started a price competition. Among other reasons, this competition is due to the emergence of price Comparison websites. The aim is to calculate the right tariff in order to keep customers satisfied while creating business. This implies that they need to:

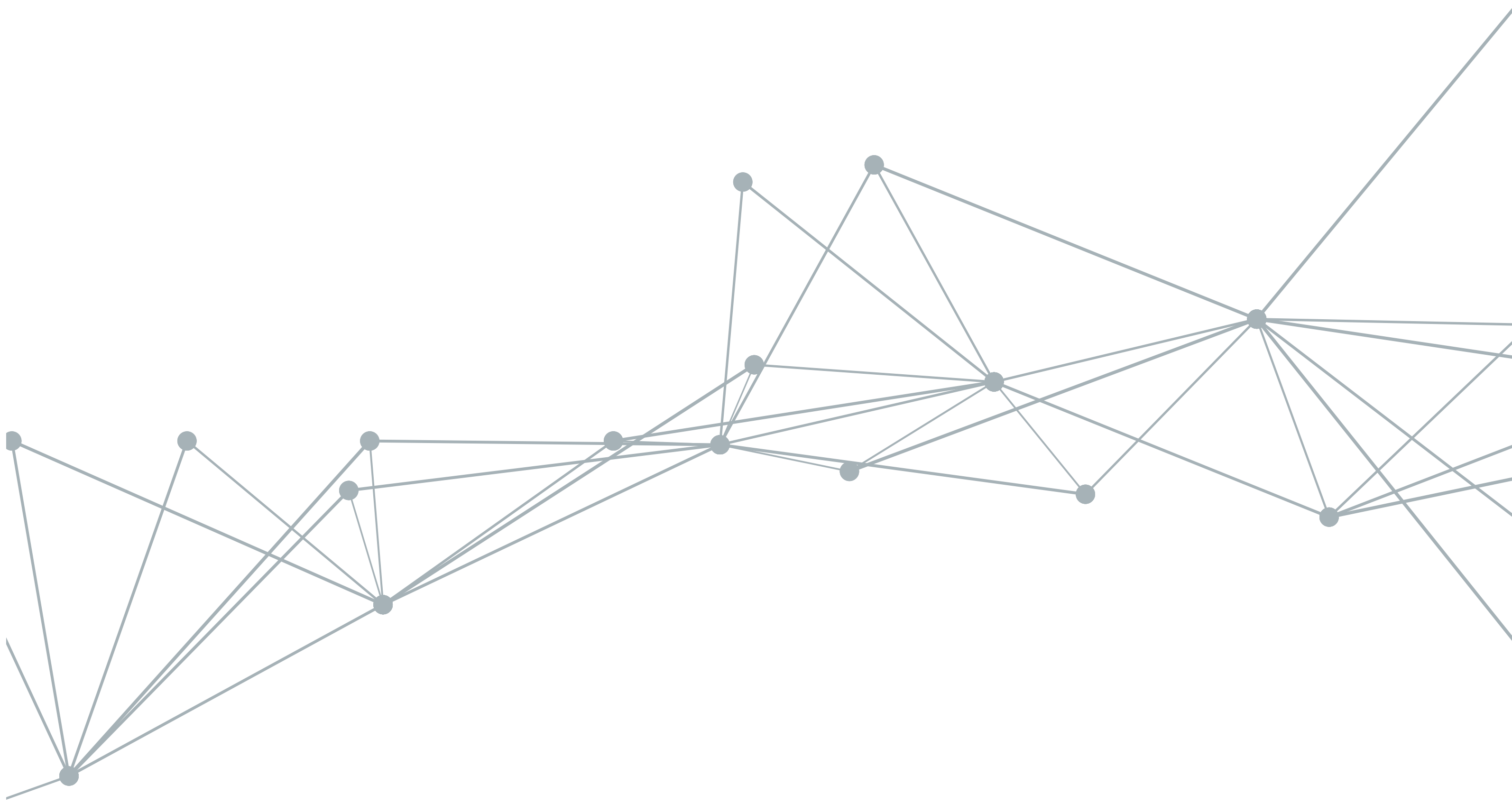
- compensate the risk they take;
- continue to be competitive;
- avoid anti selection.

In addition to that, insurers need to:

- operate very quickly;
- use appropriate data, sophisticated and best practice pricing models;
- have well documented and transparent pricing processes in place.

Currently, GLM models are commonly used by Non-life insurance companies for their pricing process. These models have a lot of advantages: transparency, understanding, etc. However, Data science methods and algorithms are growing and creating new niches not only for insurance companies but for many businesses. Therefore, some companies try to integrate machine learning methods in their pricing process.

As a result, we can wonder if Data science would replace GLM and become the new standard in the Pricing field.



GLM

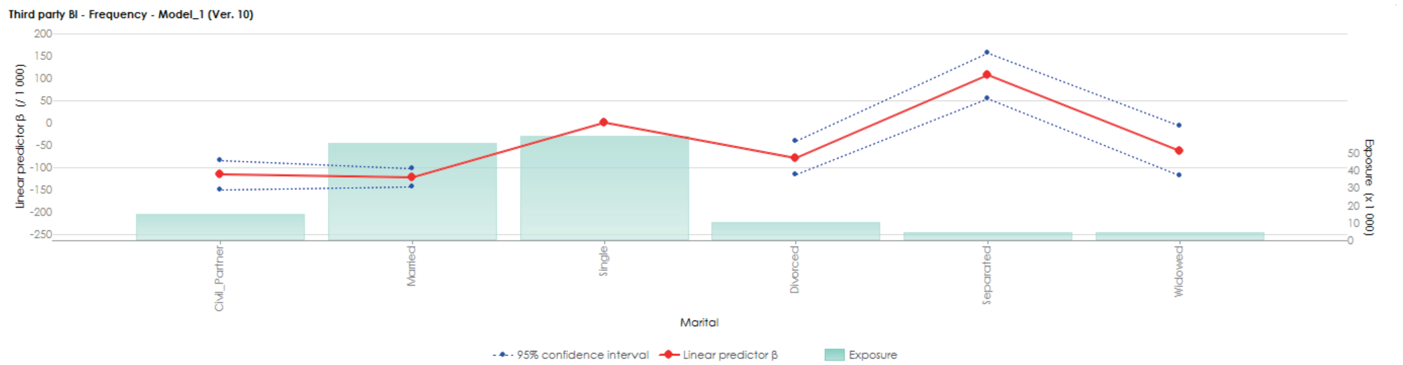
GLM is the current market standard for pricing. Indeed, these models are used to model, for a segment, pricing indicators like frequency, average cost or directly risk premium. Then, GLM will be used to model estimations made by previous models in order to obtain a tariff structure.

1- Advantages

First of all, model's behavior is easy to understand. Indeed, GLMs are an improvement of linear models. As a result, the user can easily configure the model he wants to run.

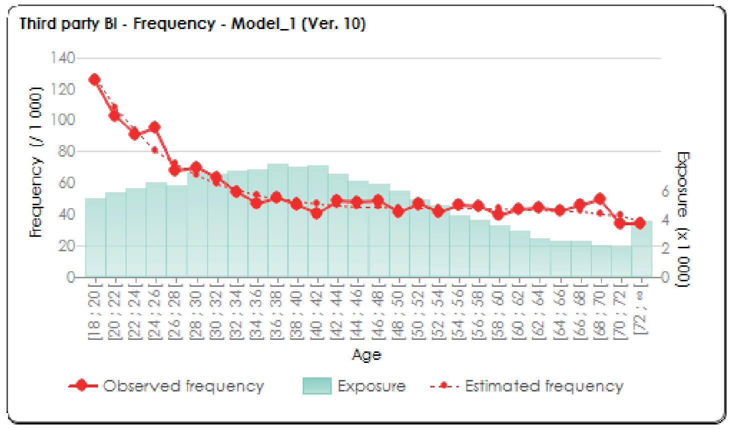
Moreover, many software packages offer to use GLM. On the one hand, open source software with several packages and on the other, actuarial software which embeds a powerful GLM engine like addactis® Pricing.

In addactis® Pricing, once a GLM has run, results are very easy to analyze.



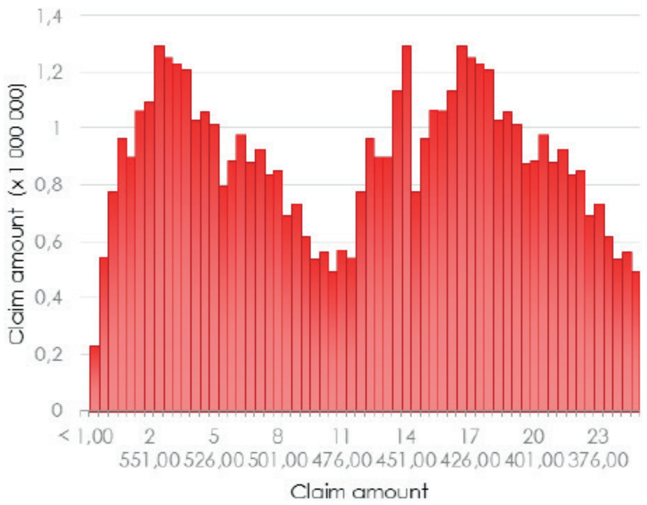
As it is a multiplicative model, the user is able to visualize the gap between modalities of a variable, all other things being equal. Besides, tariff structure is also commonly multiplicative.

Finally, GLM are flexible, the user is able to integrate constraints or smoothing if he has to review its model for legal or commercial reasons.



2- Disadvantages

However, GLM also have disadvantages.
First, GLMs are parametric methods. That means, the user has to specify a distribution when he sets up the model. And sometimes, it could be difficult to find a distribution which fits the data. As a result, this can cause prediction errors in the model.



Moreover, the user has to choose by himself variables to include in the model. Even if there are methods enabling to identify which variables are the more discriminant in the model, it does not take into account interaction between variables. In addition to that, if two variables are highly correlated, the user will not be able to include them together in the model.

Last, GLMs need a data preparation phase before they are run. Indeed, explicative factors have to be prepared for them to be added on a model. That means, the user has to check if there are no outliers or missing values and has to handle modalities with small exposures. Data preparation step can be time consuming and the time spent on this step is not spent on the GLM analysis itself.

2 MACHINE LEARNING METHODS

Indeed, Data Science is an interdisciplinary field. This is a data-driven science and it actually tries to extract insights from large amount of data. Many algorithms are based on statistical learning and machine learning methods offer new perspectives for data analyses and models conception.
These booming techniques have proved their effectiveness on the web and multimedia fields, for online advertising targeting for example.
In the pricing sector, with exogeneous data through the arrival of open data, connected objects and satellite data which are used as explanatory variables, machine learning methods could be adapted.

As a first approach, we have decided to study 3 algorithms which can be adapted to model pricing indicators:

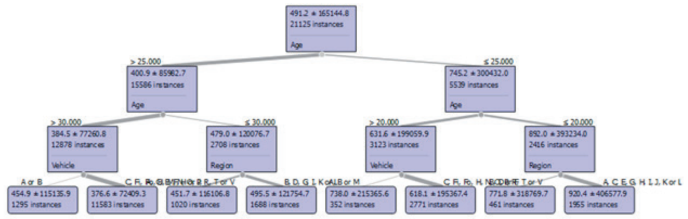
- CART
- Random forest
- Gradient Boosting

1- Algorithms

The first advantage of these methods is that they are non-parametric.
Also, these methods need to split the initial base into two bases: training set and testing set. The model is applied on the training set and the testing set enables to validate the model by calculating statistics.
There are 2 groups of algorithms: classification methods and regression methods. As pricing indicators are quantitative, regression methods will be used.
The first method studied is a regression by tree and the others are refinements of the previous one.

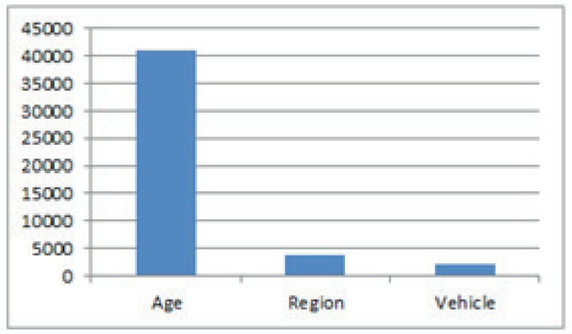
a. CART

The CART method means "Classification And Regression Trees". It is a recursive algorithm: based on the training set, at each step it searches, among all the available variables, which criteria reduce variance the most and splits the base according to this criterion. The algorithm stops when there are no criteria left or according to parameters specified by the user.



The tree is easy to read. The top of the tree is the training set, intermediate layer named "node" represents the splitting criteria. The last node is named "leaf" and contains the estimation for observations which fulfil all the criteria from the base to the leaf, this estimation is calculated by taking the observations values average for this group.

Also, this method calculates variable importance by summing all the variance reduction brought by the criterion based on this variable.



With this graph, the user can easily detect which variable is the most significant in the model.
Finally, the model is applied on the testing set and statistics are calculated by comparing observed values and estimated values:

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \widehat{\mu}_i)^2$$

$$RMSE = \sqrt{MSE}$$

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|y_i - \widehat{\mu}_i|$$

$$MAPE = \frac{1}{n}\sum_{i=1}^{n}\frac{|y_i - \widehat{\mu}_i|}{y_i}$$

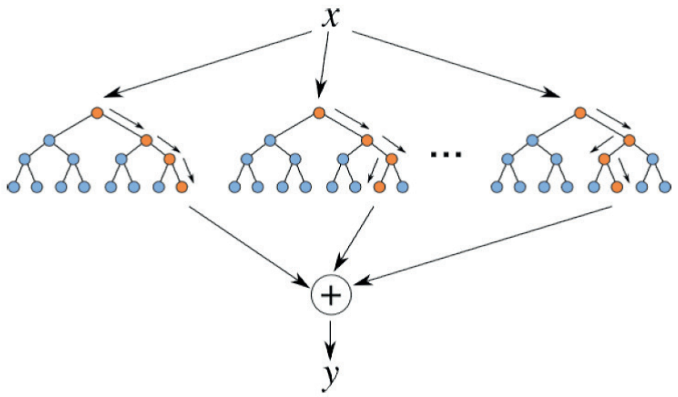
Where:

n : testing size
y_i : observed value for line i
μ_i : estimated value for line i

The user can compare several models with these statistics, the smallest the value, the better.

b. Random Forest

The previous method depends a lot on the training set, resulting in a risk of overlearning.
As a result, researches have been made in order to improve this algorithm. Random forest methods have been introduced as a generalization of CART. These methods have the advantage to reduce the prediction variance.
The Random Forest method consists in building a high number of independent trees.



It is an iterative method; the user has to choose:

- number of simulations to run
- the number of variables to use (m)
- CART parameters for all trees

At each iteration, a new base is created, randomly sampling the training set with replacement. Then a CART is fitted by randomly picking m variables (when m is equal to the initial number of variables, the method is also called "Bagging").
Model estimation is calculated by taking average estimations from all individual CART.

$$\hat{f}_B(x) = \frac{1}{B}\sum_{b=1}^{B}\hat{\phi}_b(x)$$

This method is more robust than CART but results interpretation are more difficult.

c. Boosting

Again, this last method is based on trees but it works differently than Random Forest. The aim of it is to reduce, at each step, prediction errors by fitting a tree on it.

This algorithm is iterative, it begins by fitting a weak model on the training set. Then, the next steps are the following:

- Calculate the loss function gradient for each prediction
- Fit a CART to explain them
- Sum the estimations of the previous CART with the estimations of the previous iteration in order to obtain new predictions

The user can choose the number of iterations and some other parameters in order to optimize the model.

The main advantage of this method is, besides reducing the variance, that it reduces the bias and its predictive power is better than the previous algorithms.
However, results are difficult to analyze and time processing can take longer than other methods.

2- Limits

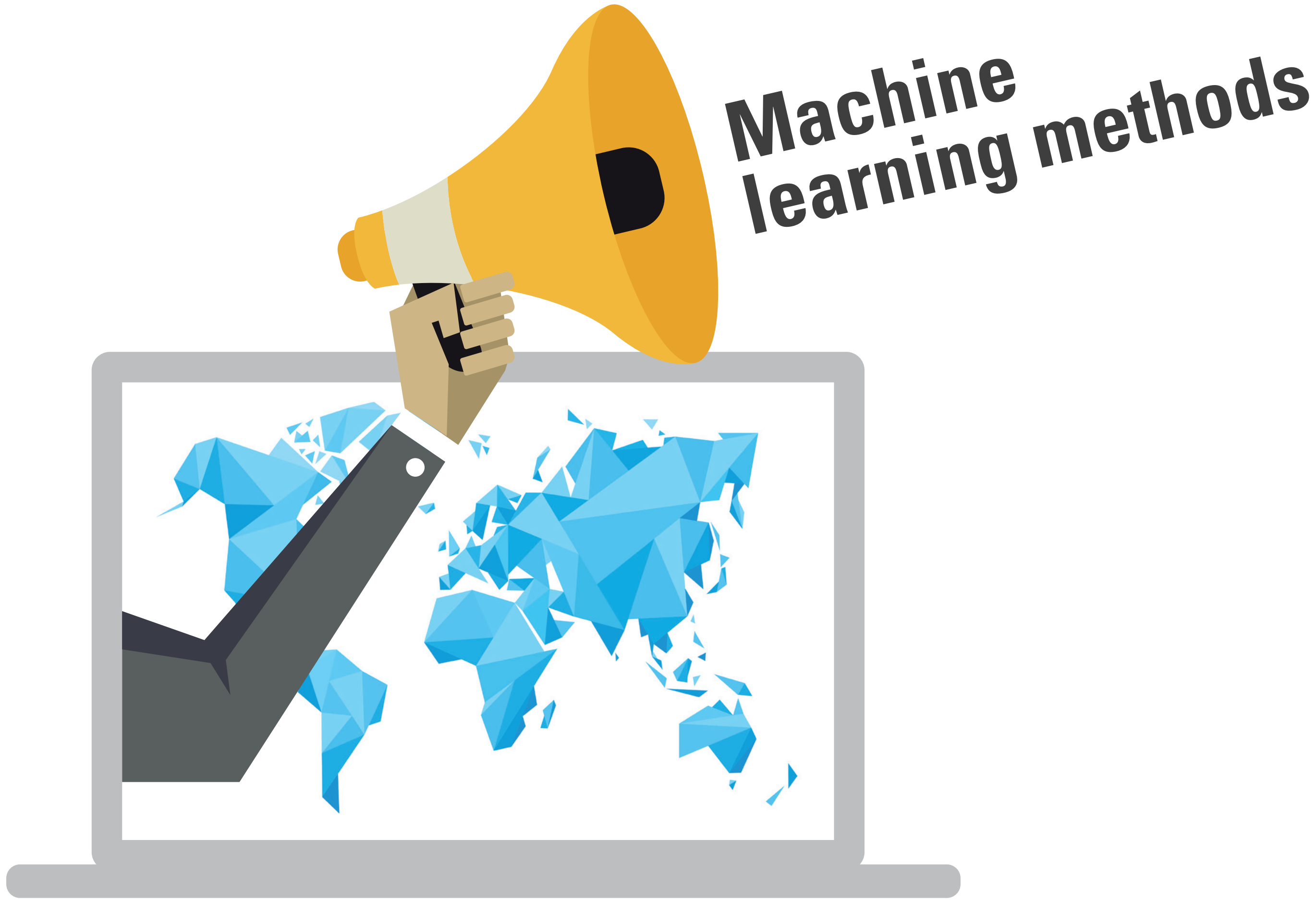
To conclude, machine learning methods can have better prediction quality, but they have also some limits.

First of all, unlike standard methods, data science is slightly more difficult to implement and computational time is higher. Moreover, many people consider these algorithms as a black box because they do not totally handle actions made during the process. Then, even if they are non-parametric algorithms, there are several parameters the user has to define and understand.

Furthermore, results are more difficult to analyze and the structure is not multiplicative as for GLM models.

Finally, there is no possibility for the user to fix constraints for certain variables or to smooth estimations.

As a result, in the pricing sector, machine learning enables to more precisely model pricing indicators but it is not totally adapted to define tariff structure.



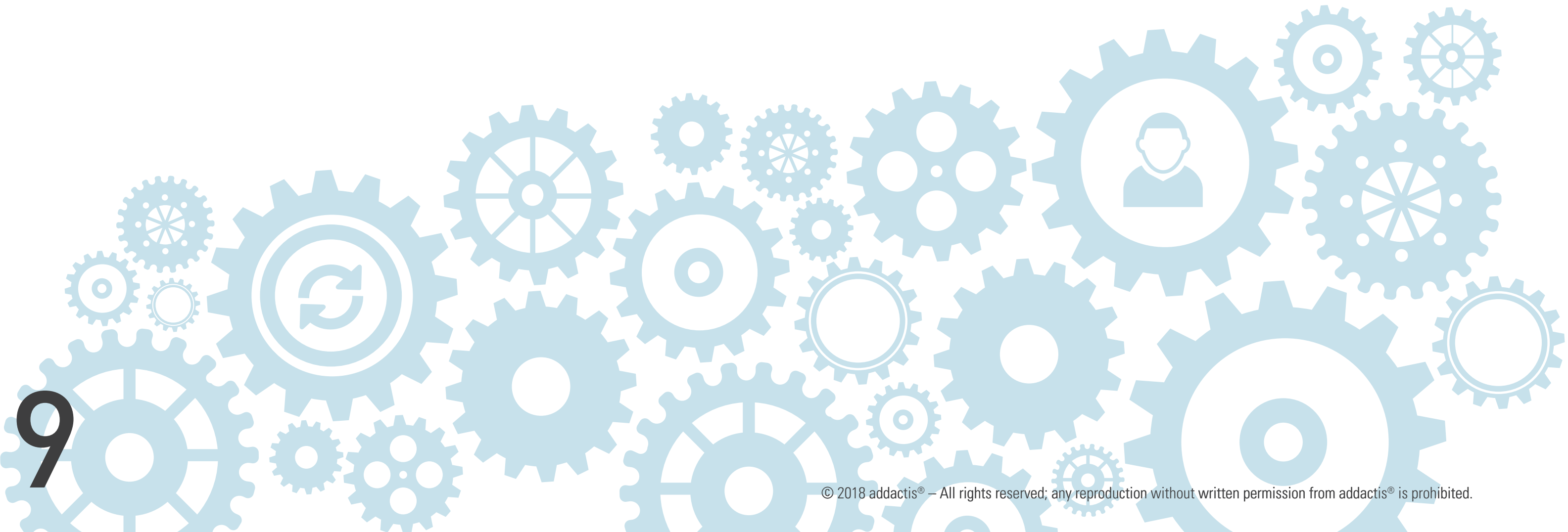
MIXTURE OF METHODS

Failing to totally replace GLM models, machine learning methods can be combined with GLM in order to improve them.

First of all, data science can bring another method to detect significant factors and modalities. Indeed, just like already widely used methods, machine learning algorithms can detect significant variables. But, in addition, it enables to define the best way to handle modalities; that means groups to create for nominal variables and interval for ordinal and continuous variables. Once this step is made, the user will apply these variables and modalities on a GLM model.

Besides, the user could use machine learning methods to detect interaction between variables and then be able to create and add these interactions in a GLM model.

Finally, as we have seen at the end of the previous part, machine learning algorithms results have not a multiplicative structure. Therefore, an alternative can be the following; use these algorithms to model the pricing indicators (frequency, amount, risk premium) and once these models have been consolidated the user can run GLM to explain the estimated premium in order to obtain a multiplicative tariff structure and to smooth estimators.



CONCLUSION

Recently, Machine learning methods emerge in a lot of fields, including insurance. More specifically, in the pricing domain, these methods can bring a lot in a continuously changing market environment.
However, even machine learning has advantages that current GLM methods haven't. Also, they have limits which can be blocking in a pricing process: lack of transparency, difficult to analyse, etc. As a result, from my point of view and even if people think so, the GLM age is not achieved.

The next step is to learn more about machine learning methods and to find the best way to combine machine learning methods with current GLM methods in order to optimize the pricing process.

GET IN TOUCH WITH

addactis
Actuarial & Software Solutions

contact@addactis.com
www.addactis.com

There is always an addactis® expert close to you! Closer to your needs, with a fine understanding of your market and requirements.

We are YOUR ALTERNATIVE!