Machine Learning : From GLM models to the CART tree passing by the Random Forest

The advent of Machine Learning models is transforming the P&C actuarial modelling, once dominated by Generalised Linear Models (GLM). These algorithms, combined to the appearance of more abundant and better structured data, multiply the possibilities of risk modelling and understanding. Moreover, to exploit at is finest the algorithms' result, it is required to perfectly master their functioning. In particular, to avoid widely spread overlearning issues.

1- Why Machine Learning algorithms are the key success factor in tomorrow's actuarial modelling?

A P&C market in transformation

On this market, well known for its strong competition, technological and regulatory changes (digitalisation, web pricing simulators, web comparators, etc…) tend to modify the policyholders behaviour: they are more than ever driven by the price. In this context, **players have to adapt their strategy to this new environment by doing things in a different way**.

Data revolution: a resource to be tapped

The **advent of data impacts the P&C insurance business**. It's the same thing on the actuarial one, on subjects like pricing or risks management.

The challenge is twofold:

- **Collect and craft data to make them workable;**
- **Choose the algorithm that will be able to understand data and to make them meaningful.**

Machine Learning: algorithms that redesign actuarial models with new data

Even if GLM models, commonly used in P&C insurance, allow the use of statistical tests to determine the quality of a model, they require strong hypotheses as inputs. That is why **common statistical modellings are restrictive and not adapted to data exploration**. But **Machine Learning algorithms are**.

2- What are Machine Learning algorithms?

The Machine Learning concept

Principle: The principle of these algorithms is to carry out a task based on the data experience.

Example: In non-life insurance, this task could be the scoring of policyholders' termination along with the data used to define the policyholders' characteristics.

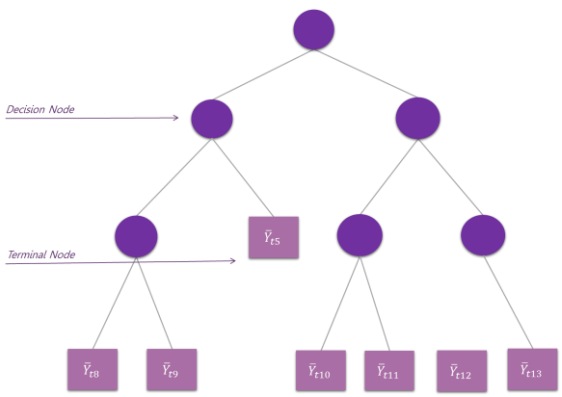
Advantages: These methods **don't use any strong assumption on the data distribution**. Thanks to Machine Learning algorithms, the interaction between data can be **collected and transcribed in order to improve the risk understanding**.

Several models: In insurance, the most famous and widely used Machine Learning models are **Decision Trees**, Neural Networks and Random Forest models.

Focus on Decision Trees

The Decision Tree became a popular method thanks to its **computing time speed**, to its capacity in managing every kind of variable and in selecting the most relevant ones. The **readability and easiness in understanding results** are also good points for this model.

The CART tree regression:
On the created partitions the CART tree regression builds constant by fragment estimators, from data, with a binary recursive cutting of explanatory variables.



Example of a Decision Tree

Limits: The principal drawback of Decision Trees is that the classification heavily depends on the order of the chosen variables. This order can harm the predictive

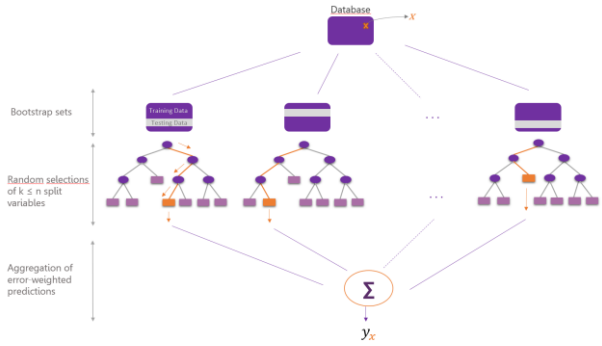
power of the model. This limit can be corrected by boosting or bagging methods.

Random Forest models

Random Forest models are specific cases of bagging for Decision Trees.

Principle:

The principle of this algorithm consists in the construction of a Decision Trees' family on bootstrap samples, and then in the aggregation of the models' forecasts. The algorithm is built to look for each split the best split for p explanatory variables randomly chosen in n variables (and not the best one of all the n variables).



Example of a Random Forest

3- Machine Learning models in practice

The use of Machine Learning algorithms requires the mastering of both the mathematical/technical methods and the software (R, python) used.

The Machine Learning approach is different from ordinary approaches and leads to some questions:

- The data quality and the meaning of these data have to be carefully checked**. The work on databases is a crucial step** during Data Science projects.

- Another point of questioning is the **data overlearning issue**. A lot of methods exist to reduce it. For example, the cross validation and the tuning.

  Example: The stopping criterion is one of the parameters to optimised for the Decision Tree.

- These models, more complex than statistical methods, are often considered as « black boxes » because of the subtlety of the algorithms and their settings. **The challenge of the Machine Learning models' understanding is strengthened by the results interpretation and sharing**.

From an operational point of view, the Machine Learning approach is not always that easy to implement (IT constraints, model management over time, etc...). In this context, to meet operational requirements, Machine Learning models can be combined with traditional approaches in order to improve hypotheses and the *a priori* choice of historical methods

Written by Julie, translated by Chloë,

Périclès Actuarial team members