ISSN: 2658-8455
Volume 4, Issue 2-1 (2023), pp. 307-319

IJAFAME
International Journal of Accounting, Finance, Auditing, Management & Economics

The use of Machine Learning in non-life insurance: Literature review

Chadia BEKKAYE, *(PhD student)*
Laboratory of Modeling Applied to Economics and Management (MAEGE)
Faculty of Juridical Sciences, Economic and Social Ain Sebaa
Hassan II University in Casablanca, Morocco

Tarek ZARI, *(Professor-Researcher)*
Laboratory of Modeling Applied to Economics and Management (MAEGE)
Faculty of Juridical Sciences, Economic and Social Ain Sebaa
Hassan II University in Casablanca, Morocco

Correspondence address :	Faculté des sciences Juridiques, Economiques et Sociales Ain Sebaa Université Hassan II de Casablanca- Maroc Téléphone : 0522766984 Maroc (Casablanca)
Disclosure Statement :	Authors are not aware of any findings that might be perceived as affecting the objectivity of this study
Conflict of Interest :	The authors report no conflicts of interest.
Cite this article :	BEKKAYE, C., & ZARI, T. (2023). The use of Machine Learning in non-life insurance: Literature review. International Journal of Accounting, Finance, Auditing, Management and Economics, 4(2-1), 307-319. https://doi.org/10.5281/zenodo.7827163
License	**This is an open access article under the CC BY-NC-ND license**

Received: February 23, 2023 *Accepted: April 19, 2023*

The use of Machine Learning in non-life insurance: Literature Review

Abstract

Insurance companies using risk modelling mainly focus on the mastery of Genelized linear models. Nevertheless, such models hinder constraints on the structure of risk and the interactions between the risk explanatory variables. Then, these limits can lead to a biased estimation of the insurance premium in certain populations of policyholders.
The traditional insurers have to face these existential challenges. Indeed, they need a focus on data strategy and implementation of statistical learning to achieve better pricing. In the last decades, computer performance has been continuously increasing, which has allowed a widespread application of the so-called statistical learning theory (Machine Learning) in many field. Non-life insurance pricing occupies as paradoxical place in actuarial science, hence the need for the application of different algorithms to evaluate the risks that insurance companies must face. Indeed, actuaries put forward the classical methods, linear algorithms mainly generalized linear model (GLM). Unfortunately, restrictions linked to this type of model, which can bias in the estimation of the insurance premium, have pushed actuaries to opt for efficient algorithms, referred to as statistical learning models. To do this, it is essential to look at the principals of classical GLM method, to identify their limitations and then to discuss the contributions of certain statistical learning methods in non-life insurance.

Keywords: Pricing, Non-life insurance, Generalized Linear Models GLM, Statistical Learning, Classification and Regression Trees CART, Random Forest, XGBoost, Neural Networks
Classification JEL: B23, C60
Paper Type: Theoretical research

ISSN: 2658-8455
Volume 4, Issue 2-1 (2023), pp. 307-319

IJAFAME
International Journal of Accounting, Finance, Auditing, Management & Economics

1. Introduction

Insurance has a crucial role to play in the indemnification of an uncertain and random event known as "risk". These named risks can be corporal (death, incapacity of work due to an accident, diseases, etc) as well as material and moral type, which affects the victim or its close relations. Other damages that lead to material destruction or loss of revenues (fire, theft, etc) the risk, by its random nature, is unknown in advance and the benefit reimbursed by the insurer is unknown.

This is the reason why actuaries have elaborated a large number of algorithms to predict these future costs. Indeed, methods such as Chain Ladder and Mack used for modeling risks.

The insurer must set an appropriate rate for each insured according to the risk, taking into consideration the rate variables that influence the loss experience. Indeed, pricing based on to the fact that pure premium must cover the risk guaranteed by the insurer, in other words, the pure premium must be proportional to the risk.

In the literature, the construction of the model is dependent on different regression methods and neural networks that have become competitive (Smith & Mason, 1997). Many authors have improved traditional statistical tools for prediction namely the Discriminant analysis (Flury & Riedwyl; Press & Wilson), Bayesian approach (Buntine & Weigend; Duda & Hart) also Multiple Regression (Menard, Myers).

In fact, these models permit studying the relationship between a response variable and a set of explanatory variables. Such models, invented by John Nelder and Robert Wedderburn (1972), are currently applied in the fields of statistics and actuarial science. In the 1980s, GLMs replaced the classical methods, which allow the modelling of non-linear behaviors and Gaussian residual distributions.

For years, the industry has experienced a change in classical statistical methods such as GLM. Recently, the CART method has increasingly become the most suitable tool for several disciplines. It belongs to the partitioning and segmentation methods developed by Breiman and col. under the acronym of CART: Classification and Regression Tree.

Buntine & Weigend proposes that many disciplines admit statistical methods such as regression analysis, Bayesian theory, least squares approximation models in a wide range of decisions.

According to Azzone & al, random forest represents a model that allows solving the policy lapse problem by retaining a benefit of this statistical learning algorithm compared to classical linear models.

Roel et al. and Guillen opt for premium personalization. In fact, their actors rely on telematics information to price car insurance premium.

The authors Sendhil Mullainathan and Jann Spiess advanced that machine learning algorithms have potential in econometric namely solving the outcome prediction problem and estimating treatment effects.

In a new study, Tober propose three machine leaning algorithms in the context of insurance frequency modelling represented by decision tree, random forest and gradient boosting machines.

In addition, the work performed by Gao & Wuthrich and Gao et al. consists in extracting the predictive capacity of the retained covariates of the telematics data on vehicule driving.

These sources provide both an overview of statistical learning methods and detailed references on their implementation.

Following these developments, researchers in computer science invented a large number of algorithms with the objective of predicting values, on the one hand, and classifying individuals on the other. This work gave birth to the theory of statistical learning. Indeed, decision trees, Classification and Regression Trees (CART), Random Forest, XGBoosting and neural networks, represent the best-known models.

This article aims to present statistical learning algorithms that improve pricing in the non-life insurance domain. The inadequacies and limitations of classical algorithms hinder the correct pricing that pushes the insurer to implement more developed machine learning models.
The objective of this paper is to present a recent literature on the benefits of statistical learning algorithms to underwriting in the non-life insurance sector. The first part is devoted to the study of the principles of the classical GLM model while identifying the limits of their application. A new section will discuss the implementation of Artificial Intelligence and Machine Learning in the Moroccan insurance domain and the beneficial role of these technologies on both sides: insured and insurance company.
The contributions of statistical learning models are then discussed (case of the CART algorithm in automobile insurance). Finally, the last part focused to the issues of comparison between GLM and Machine Learning.

2. Empirical literature review on Machine Learning application in non-life insurance

The study initiated by F.Kiema focuses on the calculation of file/folders provisions by applying Machine Learning methods and by declining to classical methods.
The calculation of the MSE seems useful in order to compare the algorithms developed with standard methods. The modeling of the ultimate charge by the CART, Random Forest and Neural Networks methods produces reserves near the ultimate charge on average, whereas the method applied by the claims managers overprovides in large majority.
The MSE defines by:

$$MSE = \frac{\sum(\hat{y} - y)^2}{n}$$

$\hat{y}$: Variable to predict
y : Real variable
n : Sample size
The study focuses on the analysis of CART, Random Forest and Neural Networks (NN model with 1cache layer and NN model with 2 cache layers). This comparison is made on performances measures: RMSE (Root Mean Square Error) and MAE (Mean Absolute Error).
According to the research, the synthesis revealed:

- The Neural Network NN to be the most prudent in terms of provisioning
- The Random Forest model more efficient when applied to RMSE and MAE.

Then, the project made by A.Yah Andrea & M.Fall includes provisioning methods using statistical learning concepts. Modeling the amount of charges from year to year and the duration of the payments attached seems significant to bridge the problem.
The results are satisfactory and prove the importance of applying individual model compared to the more classical methods of Chain Ladder. This non-life provisioning approach offers more valuable perspectives for an insurer, especially in terms of the evolution of charges that would be known by claim; even this approach is complex and costly.
There are various performance measures to evaluate the models adopted in order to achieve values near the real ones. This measures represented by AUC (Area Under the Curve), Accuracy, Sensibility, Specificity, ROC (Receiver Operator Characteristics), F-score, RMSE, etc.
We can mention several mathematical formulations of performance measures such as:

ISSN: 2658-8455
Volume 4, Issue 2-1 (2023), pp. 307-319

IJAFAME
International Journal of Accounting, Finance, Auditing, Management & Economics

- Accuracy

$$Accuracy = \frac{TP + TN}{(TP + FP + TN + FN)}$$

TP and TN define the positive and negative instance correctly ordered. While FP and FN denote positive and negative samples that are incorrectly classified.
In car insurance, the TP index would indicate the absence of a claim, while the TN corresponds to a claim.

- Sensibility

$$Sensibility = \frac{TP}{(TP + FN)}$$

- Specificity

$$Specificity = \frac{TN}{(TN + FP)}$$

- RMSE

$$RME = \sqrt{MSE}$$

Other papers have addressed the topic of prediction in the insurance industry using Machine Learning models.

Tableau 1: Various quantitative studies on Machine Learning implementation in non-life insurance

Authors	Topic	Algorithms tested	Evaluation models	Model adopted
Smith et al. 2000	An analysis of customer retention and insurance claim patterns using data mining	Decision Tree Neural Network	Accuracy ROC	Neural Network
Fang et al. 2016	Customer profitability forecasting using Big Data analysis: A case study of the insurance industry	Random Forest Logistic Regression Model Decision Trees Support Vector Machine Model Gradient Boost Model	R-squares RMSE	Random Forest
Subudhi & Panigrahi 2017	Use of optimized Fuzzy C-Means clustering and supervised classifiers for automobile insurance fraud detection	Decision Tree Multilayer Perceptron Model Super Vector Machine M model	Accuracy Sensitivity Specificity	Super Vector Machine Model

Mau et al. 2018	Forecasting the likely next purchase events of insurance customers: A case study on the value of data-rich multichannel environments	Random Forest	Accuracy F-score AUC ROC	Random Forest
Dewi et al. 2019	Analysis Accuracy of Random Forest Model for Big Data: A case Study of claim Severity Prediction in Car Insurance	Random Forest	Mean Square Error MSE	Random Forest
Abdelhadi et al. 2020	A proposed model to predict auto insurance claims using machine learning techniques	J48 Neural Networks XGBoost Naïve Bayes	Accuracy ROC	XGBoost

Source: Authors

The table represents the various authors who implemented Machine Learning algorithms precisely in the non-life insurance branch by focusing on specific performance measures.

3. Classical GLM method (Genelized Linear Models)

3.1. Presentation of GLM-Method

For a long time, the classical Gaussian linear model represents the basis of actuaries. However, methods that are more exhaustive developed due to the complexity of statistical problems: the generalized linear models GLM. For example, automobile pricing based on Poisson regression. These models are classified into three components (see, DENUIT and CHARPENTIER, 2004). The GLM founded on the same principle as the classical linear model; that is, both types allow the study of the correlation between a variable to predict and a predictor variable.
The GLM comprises three components:

3.1.1. Random component:

The variable to be predict denotes $Y = (Y_1 \dots Y_n)'$whose densities belong to the exponential family. The function f_{Y_i} belongs to the exponential family if and it is possible to find $\alpha_i \in \mathbb{R}$ (canonical, or mean, parameter), $\emptyset \in \mathbb{R}$ (dispersion parameter), a represents a non-zero function defined on $\mathbb{R}$, b constitutes a function on $\mathbb{R}$ that is twice derivable, c represents the function defined on $\mathbb{R}^2$, such that

$$f_{\alpha_i}(y_i) = \exp\left\{\frac{\alpha_i y_i - b(\alpha_i)}{a(\emptyset)} + c(y_i; \emptyset)\right\} \quad (1)$$

Where $\alpha_i \in \mathbb{R}, \emptyset \geq 0$ is a dispersion parameter and a, b and c are $\mathbb{R}$-value functions.

ISSN: 2658-8455
Volume 4, Issue 2-1 (2023), pp. 307-319

IJAFAME
International Journal of Accounting, Finance, Auditing, Management & Economics

3.1.2. Deterministic component:

The deterministic component stands for a parametric family of functions. A standard formulation is considered. Assume that it admits an expression based on a linear combination of predictors.

$$\eta(x_i) = \beta_1 x_{i1} + \cdots + \beta_p x_{ip} \quad (2)$$

The matrix product determines the linear predictor, the deterministic component of the model:

$$\eta = X\beta$$

3.1.3. Link function:

It consists on the strictly monotonic deterministic function g defined on $\mathbb{R}$. It specifies the link between the two components already mentioned, more precisely the link between the expectation of the random component and the deterministic component:
"g" represents the invertible function named link function

$$g_n(E[Y]) = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip} = x_i'\beta = \eta(x_i) \quad (3)$$

A specific link function is dedicated to each of the probability distributions of the exponential family, referred to as "canonical", and defined by $\theta = \eta$. The canonical link exists as $g(\mu_i) = \theta_i$, yet it is shown that $\mu_i = b'(\theta_i)$ so formally $g^{-1} = b'$.

3.2. Constraints of GLM

Fist, the GLM model should be parametric. Indeed, it requires fixing a law for the variable to predict $Y|X = x$. Otherwise, the model is moreover linear, which also influences the explanatory variables. This assumption is relaxed when we use generalized additive models (GAM).
For the same explanatory variable, another limitation is in place, which concerns the inability of the GLM algorithm to model different effects: the same coefficient applied for a continuous variable. There is a form of monotony. Adding also that the treatment of atypical or missing values considered delicate.
Finally, the fact of modelling the interactions between variables, although possible, is often the case according to expert opinion, as is their selection upstream. Indeed, eliminating a variable by an AIC- based method can be interesting if this variable is couple to another one. Another limitation takes into consideration. It revolves around the longevity of running sometimes these models; this leads to not testing all the possible interactions to keep the best model.
The GLM model relies on assumptions about the probabilistic distribution of data. (FOFANA, 2017)
These limitations do not allow determining a good pricing; which leads to develop new non-parametric models of machine learning, that is to say, search for statistical learning methods that have been able to extract dependency structures and features of the data not detected by generalized linear models.

4. Implementation of Statistical Learning Algorithms

In contrast to classical models using assumptions about the structure and distribution of the data, statistical learning theory assumes just one hypothesis. The data to predict, noted Y_i, are generated in identical and independent ways by a process P from the vector of explanatory variables X_i learning requires the construction of a function f and the parameter β, which characterizes the statistical learning model used.it, represents the number of neurons in a neural network or the number of nodes in a decision tree. The complexity of this function comes into

play as the algorithm allows for modeling of singularities in the data structure (interactions or nonlinear behaviors). The learning algorithm is in a break phase when we reach a stage where the complication of the model $\hat{f}(X, i\beta)$ leads to a decrease in its prediction performance on another database. This phenomenon is known as overfitting.
The learning problem relies on the formulation of a function f and the determination of the value of parameter β for which the predicted values of the model are as close as possible to the actual value (measurements). A distance must exist between the predictions and the measurements. It takes the form of the least squares cost function (denoted as quadratic cost):

$$R(Y_i, \hat{Y}_i) = (Y_i - \hat{Y}_i)^2 = (f(X_i, \beta) - \hat{f}(X_i, \beta))^2 \quad (4)$$

The problem is now in the form:

$$argmin\left(\frac{1}{N}\sum_{i=1}^{N}(Y_i - \hat{Y}_i)^2\right) = argmin\left(\frac{1}{N}\sum_{i=1}^{N}(f(X_i - \beta) - \hat{f}(X_i, \beta))^2\right) \quad (5)$$

The model which allows to have the smallest distance is retained, otherwise, the one which minimizes this distance.
Other cost functions are available. The value absolute can be applicable, for example:

$$R'(Y_i, \hat{Y}_i) = |Y_i - \hat{Y}_i| = |f(X_i, \beta) - \hat{f}(X_i, \beta)| \quad (6)$$

These two error measures are not complicated (easy to understand): they represent an overall indication of the forecast error. The quadratic error is to sanction large errors more severely than the absolute error.

4.1. Moroccan Insurance: A New Form of Intelligence

In Morocco as well as the rest of the world, the implementation of Artificial Intelligence “AI” and Machine Learning “ML” is creating new horizons.
In fact, it already exists in our current lives through our smartphones, GPS more and more in our cars. In companies through automatic translation or chatbot to contact customers online.
According to the study initiated by Mckliney & Company, AI offers a crucial opportunity for Morocco. Indeed, there exist currently eight sectors mature enough to achieve full profit by these technologies: Banking, Insurance, Telecoms, Automotive industry, Agriculture, Energy, Self-employment, and E-Gov.
In the insurance sector, Machine Learning algorithms provide predictions on the policy risk (claim frequency, claim cost, etc). The more the database is large, the more the algorithm will give conclusive results. The ML is based on the principle of prediction of products purchased by customer or that will attract his interest, using product history or other behaviors.
In sum, the application of Machine Learning is beneficial for both sides. For the insured, it allows for personalized and competitive pricing, a more specific estimate or quote. For the insurance company, it is a great opportunity to conquer the market and reduce costs by adopting automatic operations.
Today, such AI technologies facilitate our daily life in many arrears. Nevertheless, despite its rapid evolution, Artificial Intelligence and Machine Learning have not yet achieved several sectors of activity and this is the case of Morocco.

4.2. CART Algorithms in car insurance

Let us move to the domain of automobile insurance where the CART algorithm represents a powerful tool allowing (i) the improvement of segmentation as well as the classification of

ISSN: 2658-8455
Volume 4, Issue 2-1 (2023), pp. 307-319

IJAFAME
International Journal of Accounting, Finance, Auditing, Management & Economics

vehicles into different homogenous classes according to risk, and then (ii) the granting to each class average according to the cost or frequency of claim. (Paglia, 2011)
Indeed, the principle of this algorithm relies on the fact of selecting the explanatory variables based on a large number of modalities. These allow for more flexibility in the construction of two subgroups.
The CART method consists of two types of decision trees: classification trees and regression trees, which represent two non-parametric estimation methods, i.e., both types, used to solve classification and regression problems. Our case study focuses on a problem related to a regression on the amount and the frequency of claims in order to estimate the pure premium.
Classifying vehicles into homogenous risk groups is one of the segmentation issues that allows the insurer to differentiate itself from the competition and to differentiate the premium. This method also contributes to the improvement of the tariff model, and to obtaining information on the risk from the characteristics of the vehicle and on the behavior of the insured.

4.3. Comparison GLM & Statistical Learning Methods

The GLM method has been the basic tool in recent years for developing automobile insurance rates. It is a major element in the determination of the pure premium, which is traditionally estimated by actuaries based on a "$frequency \times average\ cost$" multiple model.
The GLM model is a parametric model, i.e., it relies on assumptions about the shape of the prediction function. It allows us to identify the linear dependencies between the variable to predict and the predictor variables.
On the other hand, the existence of limitations have allowed us to analyze it. First, the GLM model uses this condition, which in practice remains unverified when dealing with statistical tests. Moreover, this classical method does not allow providing a coefficient to model the interaction of two or more variables. Thus, it is important to know the variables a priori to facilitate the task of the user who must specify the interactions between these variables.
Moreover, this type of model is not reliable and capable of capturing the interaction between quantitative variables. As a result, we are obliged to perform a discretization of the continuous variables.
Unfortunately, the robustness of GLM capacity does not exist, especially in the estimation of the average reimbursements according to the representations of the histograms of the real and estimated data. (FOFANA, 2017)
These limitations have led to the innovation of other non-parametric methods that can compete with the classical GLM method.
First, we start with the CART algorithm, which does not require upstream assumptions about the distribution because actuaries have the possibility to use the model directly without going through the analysis of the data distribution. This is another reason to apply this method because it detects the structure of a problem by identifying the most complex interactions between the variable to predict and the explanatory variables, unlike the GLM, which is a model that relies on making assumptions about the shape of the prediction function.
The CART method also helps to segment a quantitative variable for each node. It allows the choice of the number of modalities to use for each variable, as using continuous or categorical data does not allow the tree to change. On the other hand, when applying the GLM method, the user must submit the definition of each modalities, as well as the set of endogenous variables to apply. The essential advantage of the CART method consists in integrating complex parameters in order to reduce or avoid the risk of over-learning, (see, L.BRIEMAN, 1984 and SIMON, 1960). Moreover, the results are easier to interpret.
CART performs an efficient and very elegant treatment for missing values following the use of the "surrogate split" method. This algorithm provides efficient treatment to outliers based on

the instability of CART used to measure the influence of individuals on the analysis (see, Bar-Hen, 2015).
In terms of performance, studies conducted to compare the CART algorithm to the GLM method. The results retained show better prediction when applying CART than GLM (see, Dissard, 2013).
Therefore, a main limitation of the CART method is its lack of robustness. Indeed, the construction of an optimal tree is likely to vary even for the smallest modification of a node;
this leads to the conclusion that this method is unstable and therefore the estimates vary enormously.
A method known as Random Forest used to overcome this limitation. It is a method allowing the improvement of the robustness.
We have already mentioned that tree regression represents a simple method to identify interactions between the variable to predict and the explanatory variables without having to set an upstream hypothesis on the law of the variable to predict, on the selection of the predictors and on the effectuation of the modalities within the predictors.
There is another non-parametric method represented by XGBoost, which represents one of the most powerful, and popular supervised algorithms.
This algorithm stands out for its simplicity and speed of application, as it relies on the selection of relevant variables automatically compared to generalized linear models, which are rather cumbersome to apply, as they must go through a prior step of variable selection.
The XGBoost algorithm allows the presentation of many effects, which are heterogeneous unlike the GLM, which relies on the principle of the non-existence of interaction between variables. The detection of these interactions is possible with the help of several tools developed before, namely H-statistics, ICE curves or PDP (or ALE) graphs allowing the combination of two variables (see, Delcaillau, 2019).
The neural network also represents a non-parametric method that comes to model the pure premium by having a node triggered with input data. This node will trigger other nodes to which it receives output data that represent the solution to a given problem.

4.4. Issues in Comparing GLM & Statistical Learning Algorithms

The insurer's first challenge is to achieve a good measurement of risk. Our goal is to compare the performance of different Machine Learning methods with respect to classical GLM models. This comparison step will allow us to present the functioning of the reference algorithms as well as the functioning of the CART method and how it is adapted to non-life insurance, noted CART-ANV
The second issue, which is the major economic and strategic issue, based on the improvement of the segmentation of the policyholder's portfolios. The insurer's objective is to develop market share in those segments that provide both a competitive advantage and a profit.
We can measure the quality of a risk group segmentation based on four major criteria: fairness, homogeneity, feasibility and incentive. (Feldblum, 2006)
The equity criterion represented by the absence of bias between the measured risk and the predicted risk; this stipulates that the premiums paid by the insured's should reflect the losses incurred by this group. The homogeneity criterion is that the risks within a group are homogenous, i.e. the subdivision of this group into several subgroups with significantly different premiums is impossible following this criterion.

5. Conclusion

The non-life insurance sector, and more specifically the car insurance sector, is a very competitive sector, which makes it necessary to develop an appropriate rate for each insured

ISSN: 2658-8455
Volume 4, Issue 2-1 (2023), pp. 307-319

IJAFAME
International Journal of Accounting, Finance, Auditing, Management & Economics

according to his risk. In other words, each risk guaranteed by the insurer covered by the pure premium.

Statistical learning plays a key role in non-life insurance, hence the need for more powerful methods such as CART, Random Forest, XGBoost and neural networks. A comparison of these algorithms with classical methods, especially GLM representing models that have limitation in determining the right pricing, will be useful.

The GLM method has been the basic tool for the last few years for the elaboration of automobile insurance rates. It plays a major role in determining the pure premium. It allows identifying the linear dependencies between the variable to predict and explicative variables.

On both insides, the existence of limits have allowed us to criticize it. First, the GLM model based on the principle of imposing a law on the target variable. Moreover, this classical method does not provide a coefficient to model the interaction of two or more variables.

Unfortunately, the robustness of the GLM capability does not exist especially in the estimation of the average reimbursements according to the representations of the histograms of the real and estimated data. (FOFANA)

These limitations have led to the innovation of other non-parametric methods (CART, Random Forest, XGBoost, neural networks) that can compete with the classical GLM method.

The CART algorithm does not impose assumptions on the distribution. It also detects the structure of a problem by identifying the most complex interactions between the variable to predict and explanatory variables. In particular: it identifies the most complex interactions between the modeling of the cost and frequency of claims (variable to predict) in automobile pricing and the various explanatory variables (age of driver, make of the vehicule, density, etc)

The key advantage of CART is to incorporate complex parameters to reduce or avoid the risk of overlearning and to interpret the results easily. In addition, it provides an effective treatment for outliers based on the CART instability used to measure the influence of individuals on the analysis (Bar-Hen, 2015). In terms of performance, the CART method helped to achieve better prediction results allowing for a comparison of these algorithms than GLM in auto insurance pricing (Dissard, 2013)

A method named Random Forest developed to overcome the limitation of the lack of robustness of the CART method.

There is another non-parametric method: XGBoost which is one of the most powerful and popular supervised algorithms. It allows the presentation of many effects that are heterogeneous. In contrast to the GLM, which depends on the principle of the non-existence of interaction between variables. The detection of these interactions is done using several tools developed before, namely H-statistics, ICE curves or PDP (or ALE) graphs allowing the combination of two variables.(Delcaillau, 2019)

There is another non-parametric method represented by neural networks that allows to model the pure premium (variable to be predicted) by triggering a node with input data in order to solve the given problem (the output data).

To conclude, we have been able to identify the limitations of generalized linear models GLM that are widely used in actuarial science. GLMs presuppose a particular form of risks and their interactions; this hinders the understanding of the subtleties of the portfolio. This is one more reason to implement statistical learning algorithms.

Currently, some studies suggest coupling the two approaches, namely GLM and Statistical learning. Moreover, the use of generalized additive models GAM is also essential to relax the linearity assumption by opting to smooth techniques.

References

(1). Weigend, Wray L. Buntine (1991). Bayesian Back-Propagation.
(2). Abdelhadi et al. (2020). A proposed model to predict auto insurance claims using machine learning techniques. Journal of Theoretical and Applied Information Technology 98: 3428–3437.
(3). Aman-Yah Andréa, Ehui & Mayoro Fall (2018). Application de méthodes de machine learning au provisionnement non-vie. EURIA Euro-Institut d'Actuariat .
(4). Johansson, E. Ohlsson (2010). Non-life insurace princing with Genelized Linear Models . Springer Verlag.
(5). Lebzar, D. Ait El Bour (2020). Artificial intelligence in the face of Moroccan companies, What challenges? International Journal of Digital Economy , vol 2, N°1.
(6). Bellina, R. (2014). Learning methods applied to non-life pricing. ISFA.
(7). Breiman L., Friedman, J. Olshen , R. Strone, C. (1984). Classification And Regressio Trees.
(8). Chapman & Hall/CRC.
(9). Christian Partrat, J.-L. B. (december 2004). Non-life insurance: Modelisation and simulation. Economica Edition.
(10). Dewi et al. (2019). Analysis Accuracy of Random forest Model for Big Data—A Case tudy of Claim Severity Prediction in Car Insurance. 5th International Conference on Science in Information Technology ICSITech, Indonesia oct 23-24, pp 60-65.
(11). Dutang, C. (2020). Actuarial services for non-life insurance. ENSAE Courses.
(12). Fabre-Rudelle, D. (2018). Contribution of statistical learning methods for individual provisioning in non-life insurance. Actuary thesis, ISUP.
(13). Fang et al. (2016). Customer profitability forecasting using Big Data analytics: A case study of the insurance industry. Computers & Industrial Engineering 101: 554–64.
(14). Fotia Santsa, R. P. (2018). Statistical learing methods applied in non-life insurance: automobile pricing. thesis.
(15). Gao et al. (2019) Claims frequency modeling using telematics car driving data. Scandinavian Actuarial Journal: 143–62.
(16). H. Riedwyl, B. Flury (1988). Mutivariate Statistics: A Practical Approach . Chapman and Hall, v + 296 pp.
(17). Henin P. (2016). A model for line-by-line liability insurance provisioning. Master's thesis, ISUP.
(18). J. Spiess, S. Mullainathan (2017). Machine Learning: An Applied Econometric Approach. Journal of Economic Perpectives-vol 31, Number 2-Spring 2017, pp 87-106
(19). James G. et al. (2013). Tree-based methods.In: An Introction to Statistical Learning:with applications in R,. Springer , New York, volume 103.
(20). LI, C. (2016). A Gentle Introduction to Gradient Boosting.
(21). M. Azzone, E. Barucci , G. Giuffra Moncayo & D. Marazzina. “A machine learning model for lapse prediction in life insurance contracts”. https://doi.org/10.1016/j.eswa.2021.116261
(22). Mathilde Clement (2020). Utilisation d'arbres de régression pour la prédiction des coû^ts automobiles. Institut des Actuaires, Dauphine. Université Paris.
(23). Mau et al. (2018). Forecasting the next likely purchase events of insurance customers: A case study on the value of data-rich multichannel environments. International Journal of Bank Marketing 36: 6.
(24). Michel Denuit, Arthur Charpentier (2004). Mathematics of non-life insurance. Volume I: Fundamental Principles of Risk Theory. Economica Edition.
(25). Michel Denuit, Arthur Charpentier (2009). Mathematics of non-life insurance. Volume II: Pricing and Provisioning. Economica Edition.

ISSN: 2658-8455
Volume 4, Issue 2-1 (2023), pp. 307-319

IJAFAME
International Journal of Accounting, Finance, Auditing, Management & Economics

(26). Mohamed Hanafy & Ruixing Ming. M. (2021). Machine Learning Approaches for Auto Insurance Big Data. Risks, 9, 42. MDPI.
(27). Oskar, s. (2019). Predicting the Customer Churn with Machine Learning Methods: Case: Private Insurance Customer Data. Master's thesis. LUT University, Finland.
(28). Ottou, P. (2017). Machine learning methods applied to line-by-line reserving in non-life insurance. Actuary thesis, Dauphine.
(29). Paglia, A. a.-G. (2011). .Risk pricing in non-life insurance, a statistical learning model approach. Bull. Fr. actuary 11 49-81.
(30). R.Wedderburn, J. N. (1972). Genelized Linear Models. Journal of the Royal Statistical Society,Series A 135, pp. 370-384.
(31). Richard O . Duda, Petter E. Hart & Nils J. Nilson (2014). Subjective Bayesian methods for rule-based inference systems. Reading in Arificial Intelligence, pp192-199.
(32). Robin G. Poggi2 J.M. (2017). CART trees and Random Forest: Importance and variable selection. arXiv :1610.08203V.
(33). Roel et al. (2017). Unraveling the predictive power of telematics data in car insurance pricing.. Journal of the Royal Statistical Society, SSRN, 2872112.
(34). Ruimy, M. (2016). Elaboration of a vehicule in car insurance. Master thesis, ISUP.
(35). S. James Press, Sandra Wilson. Choosing between logistic regression and disciminant analysis.
(36). S. Tober (2020). Tree-Based Machine Learning Models: with applications in Insurance Frequency Modelling. https://www.diva-portal.org/smash/get/diva2:1438321/FULLTEXT01.pdf
(37). Simon, R.B. (1960). Two studies in automobile insurance ratemaking. ASTIN Bulletin, 1(4), 192-217.
(38). Smith et al. (2000). An analysis of customer retention and insurance claim patterns using data mining . Journal of the Operational Research Society 51: 532–41.
(39). Subudhi & Panigrahi (2017). Use of optimized Fuzzy C-Means clustering and supervised classifiers for automobile insurance fraud detection. Journal of King Saud University-Computer and Information Sciences 32: 568–75.
(40). T M. Hastie, R. Tibshirani & J.Friedman (2008). The Elements of Statistical Learning. Springer Series in Statistics.
(41). W.Keener, R. (2006). Statistical Theory: Notes for a Course in Theorical Statistics. Springer, p 27-28; 32-33.
(42). Wedderburn, J. N.(1972). Introduction to Nelder and Wedderburn (1972) Genelized Linear Models.
(43). Wüthrich & Gao (2018). Feature extraction from telematics car driving heatmaps. European Actuarial Journal 8:383-406.
(44). Zhou, Z. H. (2012). Zhou, Zhi Hua. 2012. Ensemble Methods: Foundations and Algorithms. Boca Raton: CRC Press.