International Journal of Computer Engineering and Data Science (ISSN: 2737-8543)
Volume 2– Issue 3, July – September 2022

Health insurance pricing using CART decision trees algorithm

Fatima EL KASSIMI[1], Jamal ZAHI[2]

[1,2] University Hassan 1st, Faculty of Economics and Management, LM2CE, Settat, Morocco

E-mail : f.elkassimi@uhp.ac.ma , zahi71@hotmail.com

Article history

Received Sep 04, 2022
Revised Sep 26, 2022
Accepted Sep 27, 2022
Published Sep 28, 2022

ABSTRACT

Compulsory health insurance is intended to cover, in terms of medical care and expenditures, a heterogeneous set of insureds in terms of their health status; these insureds present different levels of risk and a wide range of health conditions. However, in Morocco, compulsory levies are collected independently of the health status, making low-risk people bear the cost of care instead of high-risked ones. Nevertheless, these levies must be based on the risk presented by the insured so that the rate of contribution is proportional to the risk that the insurance company bears. The purpose of this paper is to propose a different approach to pricing in health insurance, based on machine learning methods, namely CART algorithm.

Keywords: ***Pricing, Health insurance, Machine learning, Decision tree, CART.***

I. INTRODUCTION

Traditional pricing models such as GLM assume a functional form to explain severities and frequencies. However, data can be noisy and make it challenging to apply a single form [1] and [2]. In contrast, statistical learning models such as random forests and decision trees generally do not rely on assumptions. Furthermore, in traditional models, the explanatory variables are determined in advance [3], which is unnecessary for statistical learning techniques [4] and [5], where the algorithms select meaningful models and relevant variable [6] and [7]. For this reason, we have opted for a Machine Learning ML method to set tariff structures for our portfolio, allowing us to discharge any constraints for implementing the pricing model and select possible relevant pricing variables. Indeed, the present paper implements a supervised learning model based on trees, known by its simplicity as well as robustness in classification, and regression, namely the "classification and regression trees" CART algorithm.

The rest of the paper is organized as follows, we start section 2 by presenting the dataset, section 3 introduces CART, focusing on applying the algorithm to pricing a health insurance portfolio, and thus by estimating the severities. Section 4 provides the results of the frequency model using a similar process. Section 5 delivers a pricing matrix constructed using the results of the severity and frequency trees. Section 6 provides models final performance. Finally, we conclude with a short conclusion in section 7.

II. DATASET PRESENTATION

Our portfolio is managed by a mutual health insurance company whose identity is not disclosed for confidentiality reasons. The latter is part of the private sector governed by the National Social Security Fund. The database contains information on 98 000 health insurance claims observed during the year (2019). The characteristics present in the database cannot be used directly by a statistical model, and must be processed beforehand. In addition, the existence of missing values and inconsistent values poses many difficulties. In the same vein, some continuous variables require adequate segmentation, while discrete features with multiple modalities need to be reduced. The post-processed database has 96 540 rows and 20 variables, of which four features will be used as pricing features in our model.

- the modalities of the variable "Cat_Lib" are derived from a clustering of about 20 modalities into 12 groups.
- The quantitative variable "age_adh" has been segmented into 8 age groups, this choice of segmentation is essentially based on the opinion of an expert.

The table below shows the features that may explain the claims experience of the insureds, including those who do not pay. Each line refers to a single insured. Among its characteristics are:

TABLE I. FEATURES SELECTION

Feature	Modalities
Age range	T1 : [0,10[, T2 : [10,20[, T3 : [20,30[, . . ., T7 : [60,70[, T8 : 70 and plus,
Gender	M: Male, F: Female
Presence of Chronic disease	Y: Yes, N: No
Nature of the care consumed CAT_LIB	ACT_DENT, SURGICAL PROCEDURE, BIOLOGY, CARD_INTER, RAD_VAS, CARDIOVASCULAR, CONSULTATION, ..., EXPL_RAD, HOSPITALISATION, EXPLOR, ONCOLOGY, RAD_INTERV, RADIOLOGY, MED.

III. CART ALGORITHM FOR SEVERITIES MODEL

The literature review does not include enough work featuring the CART algorithm [8] for insurance pricing purposes. Nevertheless, we can mention the works of [9], [10], [11] and [12]. This algorithm's working principle consists of partitioning the predictors' space into various risk classes and assigning to each class the average number of claims for the frequency model and the average value taken by the claim amount in the class for the severity model. In order to build our model, we used the CART algorithm from the *rpart* package available in R.

In this paper, we will opt for the frequency/severity method i.e., we will estimate the average cost and frequency models separately (see [13], [14], and [15] for more details). Their product will represent the insurance tariff, and thus using the following risk factors: the insured's gender, his age, whether he suffers from a chronic illness, and the nature of the consumed care. Before proceeding with any modeling, it should be noted that the initial database will be split into two parts. One will be used as a training base (80% of the data). Moreover, the other 20% will be used as a test base. We also used the k=5 cross validation procedure to ensure the stability of the results.

A. *Saturated tree*

The saturated tree obtained using the *rpart* query (6275 leaves) is given below (cf. Figure 1.). This tree often tends to reproduce the values of the training data base, which leads, in some cases, to a very refined tree. However, this presents deficiencies in prediction, as a tree that is relatively dependent on the training base will not allow for the correct estimation of new data on the test base. Indeed, we aim to predict average claims costs, so the model should consider more generalities rather than exceptions. To overcome this problem, it is often customary to find a compromise between the adequacy of the model to the data and its degree of complexity (the so-called bias-variance trade-off). The Pruning technique introduced by [8] makes it possible to obtain a less complex model with exceptional predictive power.

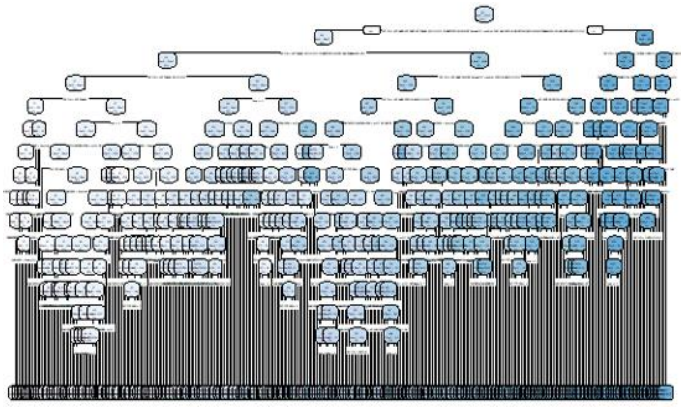

Figure 1. Maximal tree of severities

B. *Tree pruning*

This technique seeks to control the depth of the tree and improve its predictive ability. It consists of adjusting the tree complexity, which can be measured by the number of leaves in the tree. Many authors have studied this problem, [16] proposed to control the complexity by pruning. The purpose is to find less complex subtrees that predict well while avoiding overlearning. [8] present the method called 'Optimal pruning algorithm,' which consists of building a multitude of sub-trees of the maximal tree by successive pruning and then choosing, among this sequence, the optimal tree obtained *via* the complexity parameter noted '*cp*,' this penalty parameter allows to control the compromise between the length of the sub-tree and its adjustment to the training data.

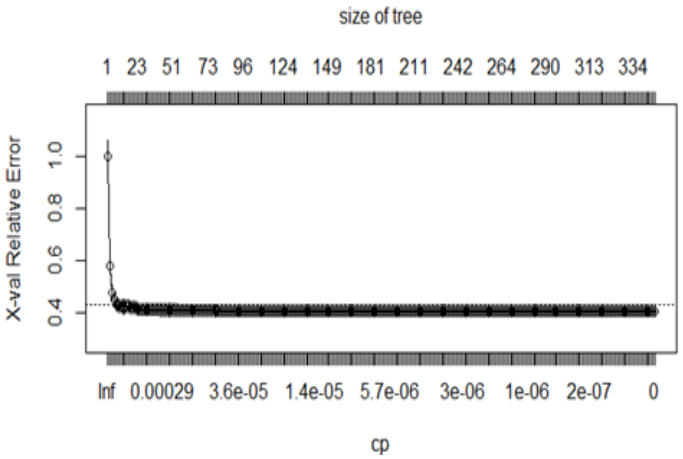

Figure 2. Optimal cp for severities decision tree

Therefore, we are interested in the tree that minimizes the standard deviation of the cross-validation error noted *xstd*. For our case study, we can clearly see on the Figure 2. that the curve reaches its minimum for the value of *cp*=2.2069e-03.

C. *Optimal tree*

The **Error! Reference source not found.**.illustrates the CART model. After pruning, we obtain a tree with 8 leaves.

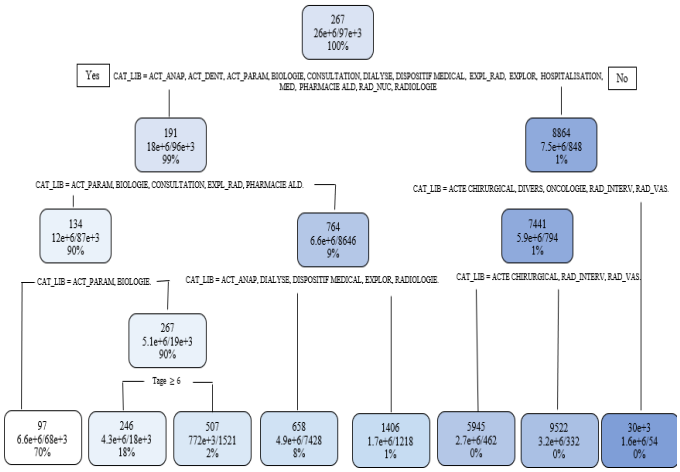

Figure 3. Optimal tree for severities model

D. *Discussion of results*

Insureds are successively separated into two subgroups according to their characteristics measured by our predictors (cf. Figure 3). Each end leaf corresponds to a group with an average cost of claims. Thus, a premium will be allocated depending on the individual's characteristics. We notice right away that the most discriminating variable is the CAT_LIB variable, which is almost present on all the separations; this variable refers to the type of care or treatment received by the insured. The table below (cf. TABLE II) summarizes the main assignment rules from the CART tree. It should be noted that some tariff classes have been merged due to their small size (< 1%) and the resemblance of the nature of the care consumed, which are quite similar.

TABLE II. SUMMARY OF SEVERITY ASSIGNMENT RULES FROM CART

Consumed care		**Severities in dhs**
- *Paramedical procedures* - *Biology procedures*		**97**
Age ≥60	- *Consultation* - *Radiological examination*	**246**
Age ≤ 60	- *Chronic disease pharmacy*	**507**
- *Dialysis procedures* - *Medical devices* - *Exploratory procedures* - *Radiology procedures*		(658 - 1 406) **1032**
- *Surgical intervention* - *Interventional Radiology* - *Vascular radiology*		(5 945 - 9 522) **7 734**
- *Oncology* - *Other procedures*		**30 000**

IV. CART MODEL FOR FREQUENCIES

The estimation of the frequency of a claim will be done following the same procedure as for the estimation of the average cost.

A. *Saturated tree*

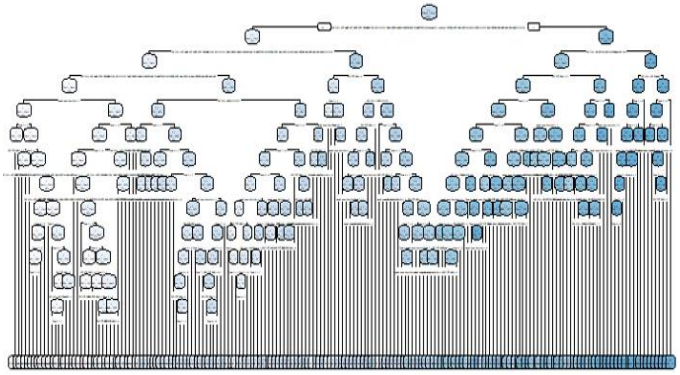

Figure 4. Maximum frequency tree

As expected, the regression tree on Figure 4 is very developed. It has perfectly replicated the training data, hence the interest of pruning to simplify it and avoid over-learning. This step remains essential in a prediction model since it allows the generalization of the model, thus increasing its performance. It should be noted that these saturated trees are, most of the time, quite unstable, insofar as the assignment rules can easily change following a change in the training data.

This overfitting situation must be avoided in order to give rise to more parsimonious and robust models. This can be achieved through the pruning procedure, mentioned before, consisting in the construction of a series of sub-trees from the saturated tree *via* a successive pruning.

B. *Tree pruning*

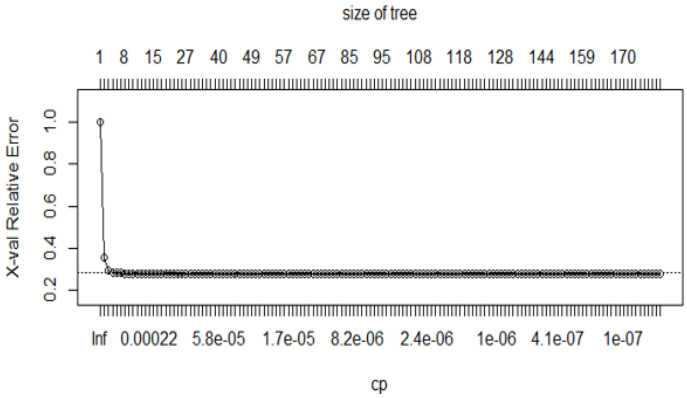

Figure 5. Optimal cp for frequencies decision tree

The Figure 5 shows that the optimal cp value, which minimizes the relative error, is cp = 1.9035e-03. We will use this value as a stopping rule to generate the pruned tree.

C. Optimal tree

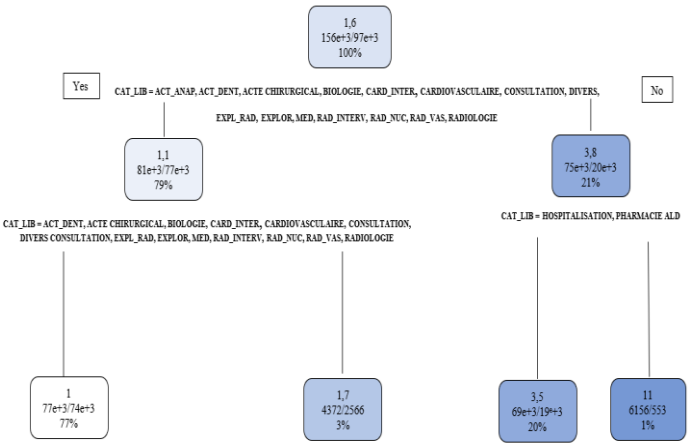

Figure 6. Optimal tree for frequencies model

The present regression tree predicts the frequency of care from predefined covariates. The average frequencies are given by the end leaves, each end leaf on (cf. Figure 6) correspond to a group of members with an associated average frequency.

D. Discussion of results

We immediately note the dominance of a single predictor, as shown in the figure 6 above. The "CAT_LIB" variable continues to be the most relevant. Indeed, the first division of the tree separates the insureds according to the care item. They are divided into two sub-groups:

The first subgroup (79% of members), contains insureds whose consumption in terms of care consists of dental procedures, surgical procedures, biology, cardiovascular care, consultations, exploratory radiology procedures, drug purchases, interventional radiology procedures and nuclear radiology, vascular radiology or radiology procedures. Among the members of this subgroup, we distinguish two groups of insureds: Those who actually consume these care items incurring an average of only one claim in the year (77% of insureds), and those who do not consume them incurring an average of two claims in the year (around 3%).

The second subgroup (21% of insureds) includes policyholders whose consumption of health care is limited to hospitalization and chronic diseases, among these policyholders, we distinguish two groups of members: those who actually consume these care items incurring an average of four claims per year (20% of insureds), and those who do not consume them, incurring an average of seven claims per year (1% of insureds).

TABLE III. SUMMARY OF FREQUENCY ASSIGNMENT RULES FROM CART

Consumed care	Frequency
- *Dental procedures* - *Surgical procedures* - *Biology procedures* - *Cardiovascular care* - *Consultations* - *Exploratory radiology* - *Drug purchases* - *Interventional Radiology* - *Nuclear Radiology* - *Vascular Radiology* - *Radiology procedures*	1
- *Unable to read the profile*	1 , 7
- *Hospitalization procedures* - *Chronic disease medication*	3 , 5
- *Oncology* - *Medical device* - *Other procedures*	11

V. FORECASTING THE PURE PREMIUM

Using the two pruned trees (claims frequency tree and claims severity tree), we constructed the matrix on TABLE IV below. This matrix will be used as a pricing matrix, i.e., the values in the matrix are pure premiums to which administrative and other expenses must be added in order to establish the commercial premiums.

TABLE IV. PRICING MATRIX DERIVED FROM THE CART SEVERITY AND FREQUENCY MODELS

			Frequency			
			1	***1,7***	***3,5***	***11***
			<=1	***]1,3[***	***[3,11[***	***>=11***
Severity	***97***	***<246***	97	164,9	339,5	1 067
	246	***]97,246]***	246	418,2	861	2 706
	507	***]246,507]***	507	861,9	1 774,5	5 577
	1032	***]507, 1406]***	1032	1 754	3 612	11 352
	7 734	***]1406, 9522]***	7 734	13 148	27 069	85 074
	30 000	***]9522,30000]***	30 000	51 000	105 000	330 000

On TABLE IV, the intersection of the first row "1" with the first column of the matrix, the amount "97" corresponds to the annual pure premium associated with insureds consuming a single "Biology" procedure per year.

The amount "3 612" corresponds to the annual pure premium associated with insureds consuming on average an annual cost ranging between "507" and "1 406" and whose claims frequency is between "3" and "11".

VI. PRICING MODEL PERFORMANCE

Let us first examine the residual errors using the root mean square error (RMSE) measure:

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2} \quad (1)$$

This index (1) provides an indication of the dispersion or variability of the prediction quality. The RMSE can be related to the variance of the model. Often, RMSE values are difficult to interpret because one is not able to tell whether a variance value is small or large. To overcome this effect, it is more interesting to normalize the RMSE so that this indicator is expressed as a percentage of the mean value of the observations. This can be used to make the indicator more meaningful.

TABLE V. PRICING MODEL PERFORMANCE

Severities model		*Frequencies model*	
CART-RMSE	*Mean*	*CART-RMSE*	*Mean*
2 429,53	121 476.345	1,1158	22,1957

The RMSE from the CART algorithm for the severity model is 2 429.53 which is relatively low because the average of the observations is 121 476.345. The same observation is valid for the frequency model. Indeed, in the first case, the variance of the model corresponds to only 2% of the average of the observations, whereas in the second case, the variance reaches less than 5% of the average of the observations.

We thus see from (cf. TABLE V) that both CART models clearly perform well. However, we must remember that these outputs are largely conditioned by the data used, as well as the configuration of the parameters. For the two models of severity and frequency, the CART algorithm seems to perform very well insofar as it reduces the variance for both models. These results therefore tend to favor the simple tree models. However, we must insist on the fact that the order of magnitude of the RMSE is not really interpretable operationally in the concrete framework of health pricing, which further restricts the formulated conclusions.

VII. CONCLUSION

In this paper, we have applied CART decision trees algorithm to the problem of health insurance pricing, taking pricing out of its traditional GLM framework by applying the algorithm to frequency and severity models. This paper makes multiple contributions to the existing literature. First, we develop comprehensive tariff structures through ML techniques for a health insurance portfolio. We have used the Poisson distribution method in this process, which is more adapted to the actuarial context. Secondly, the use of cross-validation at (k=5 folds) provides a relevant tuning procedure, which will evaluate the performance of the said methods and ensure the stability of the results over several folds of data. Finally, we have paid great attention to the interpretability of the models obtained.

REFERENCES

[1] E. Frees and E. Valdez, "Hierarchical Insurance Claims Modeling," Journal of the American Statistical Association, 103(484), 2008, pp. 1457-1469.

[2] Frees, E. W., Derrig, R. A., & Meyers, G. (Eds.). Predictive modeling applications in actuarial science (Vol. 1). Cambridge University Press, 2014.

[3] J. Paefgen, T. Staake and F. Thiesse, "Evaluation and aggregation of pay-as-you-drive insurance rate factors: A classification analysis approach," Decision Support Systems, 56(1), 2013, pp. 192-201.

[4] M. Kuhn and K. Johnson, "Applied Predictive Modeling," Springer, 2013.

[5] T. Hastie, R. Tibshirani and J. Friedman, "The Elements of Statistical Learning: Data Mining, Inference, and Prediction.," Springer, 2nd edition, 2009.

[6] Narwani, B., Muchhala, Y., Nawani, J., & Pawar, R. Categorizing driving patterns based on telematics data using supervised and unsupervised learning. In 2020 4th International Conference on Intelligent Computing and Control Systems (ICICCS). Piscataway: IEEE,2020, 302–6.

[7] Kuo, K., & Lupton, D. Towards Explainability of Machine Learning Models in Insurance Pricing .

[8] L. Breiman, J. Friedman, C. J Stone and R. A. Olshen, "Classification and Regression Trees," CRC press, 1984.

[9] Paglia, A., Phélippé-Guinvarc'h, M., & Lenca, P. Adaptation de l'algorithme CART pour la tarification des risques en assurance non-vie. EURO Institut d'actuariat EURIA,2011, 1-12.

[10] Henckaerts, Roel. "Insurance Pricing in the Era of Machine Learning and Telematics Technology." PhD diss., KU Leuven, 2021.

[11] Diao, L., & Weng, C. Regression tree credibility model. North American Actuarial Journal, 2019, 169–96.

[12] Henckaerts, R., Côté, M.-P., Antonio, K., & Verbelen, R. Boosting insights in insurance tariff plans with tree-based machine learning methods. North American Actuarial Journal, 2020, 1–31.

[13] Lotsi, A., Mettle, F., & Adjorlolo, P. K. Application of Bühlmanns-Straub Credibility Theory in Determining the Effect of Frequency-Severity on Credibility Premium Estimation. ADRRI Journal of Physical and Natural Sciences, 2019, 1-24.

[14] Sakthivel, K. M., & Rajitha, C. S. . Artificial intelligence for estimation of future claim frequency in non-life insurance. Global Journal of Pure and Applied Mathematics, 2017, 13, 10.

[15] Gao, G., Meng, S., & Wuthrich, M. Claims Frequency Modeling Using Telematics Car Driving Data. Scandinavian Actuarial Journal, 2018.

[16] Biau, Gérard, and Luc Devroye. "On the layered nearest neighbour estimate, the bagged nearest neighbour estimate and the random forest method in regression and classification." Journal of Multivariate Analysis 101, no. 10, 201, 2499-2518.