MODELING CLAIMS FREQUENCY IN THE ALGERIAN AUTOMOBILE INSURANCE MARKET USING MACHINE LEARNING

Walid **OUCHERIF**[*]
Nassim **TOUCHE**[**]

Received: 21/03/2023/ **Accepted**: 19/09/2023 / **Published**: 08/01/2024
Corresponding author: walid.oucherif@univ-bejaia.dz

ABSTRACT

The Algerian automobile insurance market faces significant challenges in pricing insurance policies due to the lack of reliable predictions for insurance losses. In this paper, we introduce a new ratemaking system that leverages advanced data analysis techniques, including Generalized Linear Models and machine learning algorithms like Neural Networks, boosting, and stacking algorithms, to model claims frequency. By analyzing data and statistics of drivers in the Algerian market, this system offers a data-driven solution that helps insurers to better understand their risk exposure and make informed pricing decisions. The proposed system has implications for both insurers and policyholders in terms of fairer and more accurate pricing, which will ultimately benefit the Algerian economy.

KEY WORDS: Auto insurance, statistical learning, Neural Nets, GBM, XGBoost.

JEL CLASSIFICATION : G22, G52, C15, C44.

[*] Research Unit LaMOS, University of Bejaia (Algeria), walid.oucherif@univ-bejaia.dz, https://orcid.org/0000-0003-4169-3336

[**] Research Unit LaMOS, University of Bejaia (Algeria), nassim.touche@univ-bejaia.dz, https://orcid.org/0000-0002-9185-3433

نماذج الخسائر المتكررة في السوق الجزائري لتأمين السيارات باستخدام تعلم الآلة

ملخص

يواجه سوق التأمين على السيارات الجزائري تحديات كبيرة في تسعير عقود التأمين بسبب عدم وجود تنبؤات موثوقة لخسائر التأمين. في هذا البحث، قدمنا نظاماً جديداً للتسعير يستفيد من تقنيات حديثة لتحليل البيانات، بما في ذلك النماذج الخطية المعممة وخوارزميات التعلم الآلي مثل الشبكات العصبونية الاصطناعية وخوارزميات التعزيز والتكديس لنمذجة تكرار الخسائر. من خلال تحليل بيانات وإحصاءات السائقين في السوق الجزائرية، يقدم هذا النظام حلاً يعتمد على البيانات التي تساعد شركات التأمين على فهم تعرضهم للمخاطر بشكل أفضل واتخاذ قرارات تسعير صحيحة. النظام المقترح له آثار على كل من شركات التأمين والمأمن عليه من حيث التسعير الأكثر عدلاً ودقة، الأمر الذي يستفيد منه الاقتصاد الجزائري في نهاية المطاف.

كلمات المفتاحية: التأمين على السيارات، التعلم الإحصائي، الشبكات العصبونية الاصطناعية، XGBoost، GBM

MODÉLISATION DE LA FRÉQUENCE DE SINISTRALITÉ SUR LE MARCHÉ ALGÉRIEN DE L'ASSURANCE AUTO À L'AIDE DE L'APPRENTISSAGE AUTOMATIQUE

RÉSUMÉ

Le marché algérien de l'assurance automobile est confronté à des défis importants dans la tarification des polices d'assurance en raison du manque de prévisions fiables pour les pertes en assurance. Dans cet article, nous introduisons un nouveau système de tarification qui s'appuie sur des techniques avancées d'analyse de données, notamment des modèles linéaires généralisés et des algorithmes d'apprentissage automatique tels que les réseaux de neurones, les algorithmes de boosting et de stacking, pour modéliser la fréquence des sinistres. En analysant les données et les statistiques des conducteurs sur le marché algérien, ce système offre une solution basée sur les données qui aide les assureurs à mieux comprendre leur exposition au risque et à prendre des décisions de tarification éclairées. Le système proposé a des implications tant pour les assureurs que pour les assurés en termes de tarification plus juste et plus précise, ce qui profitera en fin de compte à l'économie algérienne.

MOTS CLÉS : Assurance auto, Apprentissage statistique, Réseaux de neurones, GBM, XGBoost.

INTRODUCTION

Car insurance is one of the most important segments of the insurance industry in any country. Automobile insurance is mandatory for all registered vehicles and covers third-party liability and physical damage from collisions with other vehicles or objects. To provide adequate financial coverage, Insurers use various techniques to create personalized premium rates for individual customers. For many decades, one of the most popular methods for determining premiums based on different risk factors has been statistical tools such as Generalized Linear Models (GLMs). By using GLMs, insurers can predict the likelihood of specific events and adjust premiums accordingly, ensuring customers receive the right level of coverage for their individual circumstances.

GLMs are very useful in this context (Renshaw, 1994), especially because predictions can be explained easily and be interpretable according to the rating variables. The primary basis for calculating automobile insurance premiums is the statistical concept of "frequency and severity" for which GLM regressions are adequate tools. For instance, the use of Poisson regression has been widely adopted for the estimation of the frequency in actuarial literature. In practice, there is often a dependence between frequency and severity, which makes it important to model them jointly (Garrido, 2016). Furthermore, in the process of modeling automobile insurance data to forecast claims frequency, it is common to observe dispersion in the data, which can result in inaccuracies in the predictions. The negative binomial and generalized Poisson regression models are commonly used to handle overdispersion (Noriszur, 2007).

However, what about using data science and Artificial Intelligence (AI) for insurance? With the development of new technologies and AI, the application of machine learning algorithms to make predictions for insurance is becoming increasingly appealing, despite the potential trade-off between interpretability and accuracy when compared to traditional statistical models. Numerous investigations (Wuthrich, 2021, Ferrario, 2019) have extensively explored the pricing techniques and data analysis methods for non-life insurance. For instance, some studies (Noll, 2020, Ferrario, 2020) evaluated various machine learning algorithms against GLMs, utilizing the *CAS datasets* package in *R* (Charpentier, 2015) and reported the superiority of AI based methods. Deep learning algorithms have been utilized to explore alternative approaches. One such approach is the Combined Actuarial Neural Networks (Schelldorfer, 2019), which processes categorical features using neural network models. The XGBoost framework has gained popularity as an interesting approach for handling tabular data. Its usage has led to a more accurate process for predicting accident claims in car insurance (Pesantez-Narvaez, 2019).

The issue with the current approach to auto insurance pricing in Algeria is that it lacks fairness and accuracy. By only considering the value of the vehicle in the premium calculation, insurers are not taking into account the individual risk factors associated with each driver and their driving history. This leads to situations where some drivers are unfairly charged higher premiums than they should, while others are given lower premiums than they should, based on their actual risk.

A solution to this issue requires the development of a new ratemaking process that takes into account drivers' statistics. By including other characteristics of the driver, such as age, gender, and driving history, insurers can more accurately assess the risk associated with each driver and customize premiums accordingly. This approach is not only more ethical, as it ensures that each driver pays a fair price for their insurance coverage, but it also promotes stability in the insurance market. Moreover, accurate loss prediction can not only improve the solvency of insurers by mitigating unexpected losses and enhancing their financial stability, but it can also significantly improve the stability of the Algerian insurance market at a large scale.

We aimed to develop a ratemaking process that takes into account drivers' statistics and, to that end, employed a range of analytical tools—including GLMs, Neural Networks (NN), boosting, and stacking algorithms (LightGBM and XGBoost)—to estimate policyholders' claim frequency.

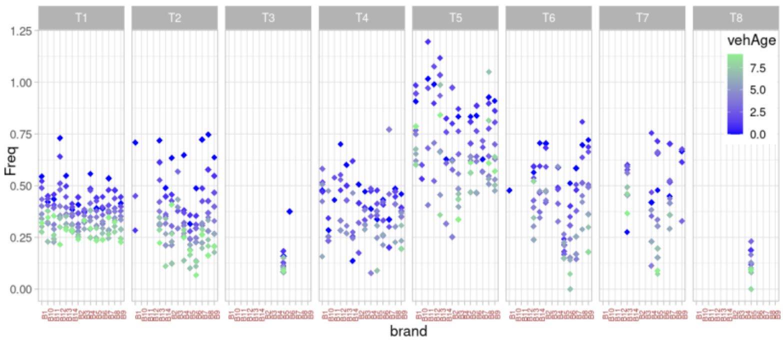
1- DATA AND EVALUATION METRIC

The goal of this paper is to use actuarial techniques to model insurance losses of the portfolio from the largest car insurer in Algeria *Société Nationale d'Assurance*. The data set contains 398 830 insurance policies of the All-Risks guarantee.

To model insurance losses, we have nine features: driver's age and his driving license seniority, vehicle's power, age and value, the exposure, vehicle's car brand, region of driving and lastly, the type of the vehicle. The latter three features are categorical variables, whereas the first six are continuous.

Performing an exploratory data analysis is a crucial step prior to constructing any model. Its purpose is to identify patterns, relationships, and anomalies in the data through the use of various statistical graphs and visualization techniques.

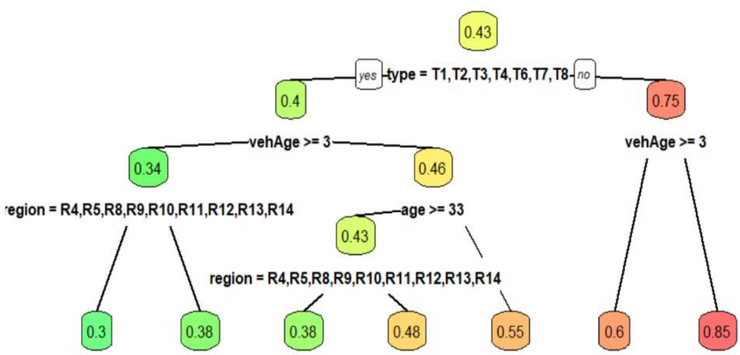
Figure 1. Claims frequency as a function of the vehicle's brand, age and type.



Source : the authors

Our primary objectives were to extract significant variables from the dataset and identify outliers and anomalies. To achieve this, we built a small binary decision tree to gain initial insight into the most relevant variables that should be incorporated into our models:

Figure 2. Small depth decision tree that shows important features.



Source: the authors

In order to evaluate the performance of the count models, we will use the Poisson deviance, which is a measure that quantifies variations in the data that are not explained by the model under consideration. For a given estimator $\hat{\mu}$ of y, the Poisson deviance loss of a data set with n rows is equal to

$$D = 2\sum_{i=1}^{n}\left[y_i \log\left(\frac{y_i}{\hat{\mu}_i}\right) - (y_i - \hat{\mu}_i)\right]$$

2- ESTIMATION OF CLAIMS FREQUENCY

2.1- Generalized linear models

The Generalized Linear Model (GLM) extends the ordinary linear regression model to accommodate response variables that have distribution models other than a normal distribution, thus providing a more flexible framework for modeling. It is a powerful statistical tool, especially in the context of insurance losses, where the data is typically skewed, not necessarily normal.

Given a response variable y, the GLM is:

$$f(y) = c(y,\varphi)\exp\left(\frac{y\theta - a(\theta)}{\varphi}\right), \qquad g(\mathbb{E}(y)) = x'\beta.$$

The equation $f(y)$ states that the distribution of y is in the exponential family. The second one specifies that a transformation of the mean, $g(\mathbb{E}(y))$, is linearly related to explanatory variables contained in x.

GLMs consist of three primary components:

1. **The random component:** it refers to the probability distribution of the response variable$y$ (e.g., normal, binomial, Poisson);
2. **The systematic component:** it relates to the explanatory variables, similar to the predictors in a traditional linear regression model. In general, we have: $\eta(x) = x'\beta$;
3. **The link function$g$:** it serves to connect the systematic component to the random component: $g(\mathbb{E}(y)) = x'\beta$. It describes the relationship between the linear predictor and the mean of the distribution function.

The flexibility of GLMs arises from the wide variety of distributions and link functions that can be used. For example, in our case, a Poisson distribution and a logarithmic link function are a common choice for modeling claims frequency. For more details about GLMs, one can refer to the relevant literature.

The initial model we will assess is the intercept-only model, which employs the training set's y mean to predict the claims frequency. It is labeled GLM1, and the in-sample and out-of-sample actual claim frequencies for the learning and test sets are 42.84% and 42.37%, respectively. Note that in-sample refers to the learning set and out-of-sample refers to the test set.

Table 1. Performance comparison of various GLM models. Losses are in 10^{-2}.

	In-Sample Loss	**Out-of-Sample Loss**	**Predicted Average Frequency**	**Improvement GLM2 baseline**
GLM1 (mean)	97.27	96.58	42.84%	/
GLM2	92.79	92.25	42.88%	100%
GLM3	92.37	91.84	42.87%	109.4%
GLM4	92.01	91.57	42.86%	115.7%

Source: the authors

Following that, we construct a Poisson GLM model that solely considers pertinent features (cited previously) while disregarding any plausible correlations between them. This model is identified as GLM2.

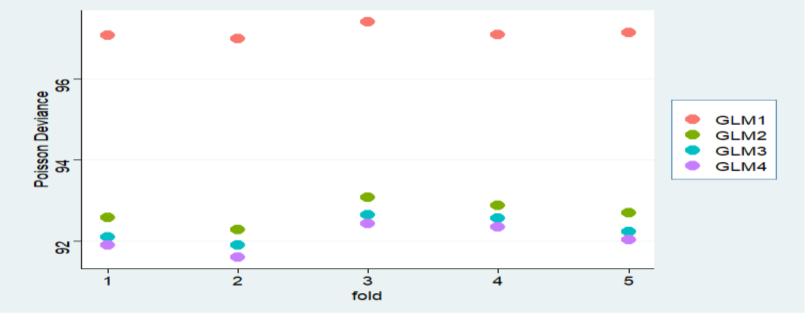
It is apparent that the GLM2 model has led to a significant enhancement in our forecasts compared to the GLM1 model (see table 1). From now on, we will be using the deviance loss of the GLM2 as the benchmark for our forthcoming model comparisons.

Next, we build the GLM3 model where we will adopt an appropriate continuous functional form for some of the continuous variables such as the driver's age, vehicle's value… (Schelldorfer, 2019).

$$AgeDriver \longmapsto \beta_j AgeDriver + \beta_{j+1}\log(AgeDriver) + \sum_{k=2}^{m} \beta_{j+k}(AgeDriver)^k$$

The parameter $m$is derived after training several models, with the aim of capturing the polynomial effect of the variable in question while avoiding overfitting the training data. This highlights the importance of splitting the data set prior to modeling. The final GLM model, in addition to considering the polynomial effects of variables, also takes into account potential correlations among features. This model is referred to as GLM4. Table 1 shows that GLM4 performs better in predicting in-sample and out-of-sample data.

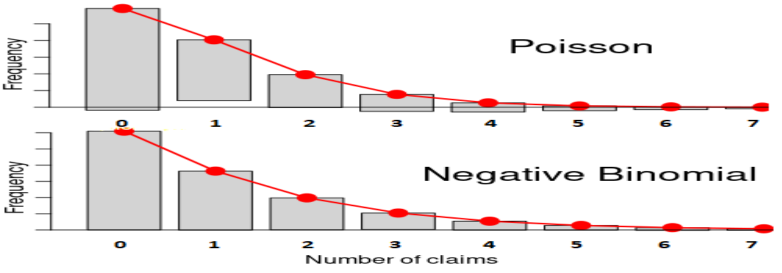
Figure 3. K-Cross Validation GLM models.



Source: the authors

To prevent overfitting, we conducted a cross-validation procedure to verify the generalization performance of our models. The evaluation metrics' high variance between blocks indicates poor prediction generalization of our models. The results of this procedure are presented in Figure 3, indicating that our models generally perform well.

Figure 4. Negative Binomial type 2 (NB2) vs Poisson on theoretical vs data set distribution of response *y*.



Source: the authors

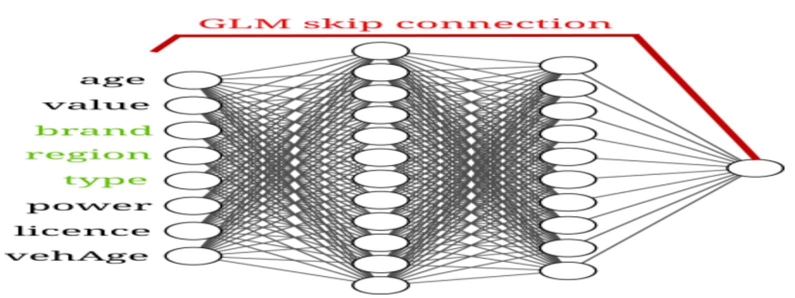
However, the actual frequency of the response variable y is considerably high, and our data manifests over-dispersion, suggesting that the variance of y is greater than the mean. To verify this fact, we used the *vcd* library in R and plotted the theoretical distribution (in red line) against the distribution of the count variable in our data set (in gray histograms), as shown in Figure 4. Employing a Negative Binomial type 2 (NB2) model resulted in equivalent predictions to the Poisson models, but the variances varied and scaled with predictions, which aided in identifying over-dispersion

2.1- Combined Actuarial Neural Network

This part is a reproduction of the deep learning tutorial by (Schelldorfer, 2019). The focus on the latter article was not on tuning hyperparameters but rather on embedding layers, because it seems to provide a better use of categorical variables within neural networks than the classical dummy transformations.

The goal of the Combined Actuarial Neural Network is to improve the classical GLM regression models using neural networks. The idea is to nest the GLM into a network architecture by injecting it in a commonly named *skip connection* that directly links the input layer to the output layer, see figure 5.

Figure 5. CANN architecture with *GLM skip connection* (Schelldorfer, 2019).



Source : the authors

Note that categorical features in green have been embedded but not represented as is in figure 5. For the adaptation to our data set, we went through the same methodology and used the same preprocessing procedure on our features, see (Ferrario and Noll 2020, Schelldorfer, 2019) for processing features tutorial.

Regarding the performance of the CANN model, according to table 2 below, it appears to surpass the best GLM model that we have built, even though the GLM incorporates some complex interactions between features. It is worth noting that fine-tuning the hyperparameters of the CANN approach might lead to even better results. However, we have decided not to proceed with fine-tuning our model since the marginal improvement that could be achieved after several hours of optimization is not worth the effort.

2.1- LightGBM

LightGBM (Guolin, 2017) is an open-source gradient boosting framework developed by Microsoft that uses tree-based learning algorithms designed to process data faster and provide better accuracy. LightGBM's uniqueness lies in its "Gradient-based One-Side Sampling" (GOSS) and "Exclusive Feature Bundling" (EFB) algorithms. Unlike traditional boosting models that grow trees horizontally, LightGBM grows trees vertically, choosing the leaf with maximum loss to grow. Major boosting algorithms are based on gradient tree boosting. For a more comprehensive understanding of decision trees, one can consult relevant literature.

A regression tree can be represented by:

$$f(x_i) = \omega_{\tau_i} \text{where} \tau_i = q(x_i)$$

- $x_i \in \mathbb{R}^m$ is the vector of the *i-th* observation's features, and m is the number of features;
- τ_i is an integer between 1 and T and is the leaf index in the tree. T is the total number of leaves in the tree;
- $\omega_{\tau_i} \in \mathbb{R}$ is the score of leaf τ_i. It is also called *leaf weight*;
- q is a function mapping from $\mathbb{R}^m$ to the integer space $\{1,2,\dots,T\}$

Upon the principles of boosting, gradient tree boosting employs a sequential approach where each subsequent predictor learns from the errors of its predecessor. The ultimate predicted outcome is calculated as a weighted sum of the ensemble of decision trees. We have:

$$\hat{y} = \phi(x_i) = \sum_{k=1}^{K} f_k(x_i), \qquad f_k \in F,$$

Where K is the number of trees, f_k is the *k-th* tree model. $f_k(x_i)$ is the score of the *i-th* observation obtained from the *k-th* tree. $F$is the tree space.

LightGBM can be used for regression, classification, ranking and many other machine learning applications. What we are interested in here is the Poisson regression.

After a few hours of hyperparameter tuning, we achieved better results than the CANN approach, as demonstrated in the table 2 below. Even though LightGBM performs better than neural networks in terms of out-of-sample data prediction, the difference is not substantial. Therefore, we have decided not to continue optimizing hyperparameters, as we believe that the gains obtained are not worth the time and effort given the results of the following boosting algorithm.

2.1- XGBoost (eXtreme Gradient Boosting)

XGBoost (Tianqi, 2016) is a model well-known for its high predictive performance. It is an optimized gradient boosting library designed to be highly efficient, flexible and portable. It also provides a parallel tree boosting that help train models more efficiently by utilizing all cores of a CPU for example. XGBoost also uses stacking to make predictions. Stacking is an ensemble machine learning technique designed to train multiple models to solve similar problems and then combine them into one that makes more accurate predictions.

As mentioned above, LightGBM uses a leaf-wise growth strategy whereas XGBoost employs a level-wise growth strategy for decision trees, which can lead to much deeper trees. XGBoost optimizes its model using the following regularized objective function:

$$L(\phi) = \sum_{i=1}^{n} l(\hat{y}_i, y_i) + \sum_{k=1}^{K} \Omega(f_k)$$

$$s.t. \;\; \Omega(f_k) = \gamma T_k + {}^{1}/_{2}\, \lambda \|\omega^k\|^2,$$

- $l(\widehat{y_i}, y_i)$ is a differentiable convex loss function (Mean Squared Error MSE for example);
- $\Omega(f_k)$is a regularization term that penalizes the complexity of the model to avoid overfitting. Ω is defined by the number of leavesT_k and the weights of leaves $\omega^k = \left(\omega_1^k, \omega_2^k, \dots, \omega_{T_k}^k\right)$.

Hyperparameters are the configuration settings used to optimize machine learning algorithms. In XGBoost, several hyperparameters can be tuned to optimize the performance of the model. Here are some of the key ones:

- N estimators: The number of gradient boosted trees to use. More trees can lead to better performance but also risk of overfitting.
- Max depth: The maximum depth of a tree. Deeper trees can capture more complex patterns but also risk overfitting.
- Learning rate: This is the step size shrinkage used in each boosting step to prevent overfitting.
- Gamma: A node is split only when the resulting split gives a positive reduction in the loss function. Gamma specifies the minimum loss reduction required to make a split.
- Lambda: L2 regularization term on weights, used to handle the regularization part of XGBoost.

Tuning these hyperparameters can be done manually, through trial and error, or more systematically with approaches like grid search or random search.

Table 2. Performance comparison of all models. Losses are in 10^{-2}.

	In-Sample Loss	**Out-of-Sample Loss**	**Predicted Average Frequency**	**Improvement GLM2 baseline**
GLM1 (mean)	97.27	96.58	42.84%	/
GLM2	92.79	92.25	42.88%	100%
GLM4	92.01	91.57	42.86%	115.7%
CANN	91.62	91.35	42.86%	120.8%
LightGBM	90.62	91.25	41.19%	123.2%
XGBoost	**89.33**	**90.81**	**42.81%**	**134.1%**

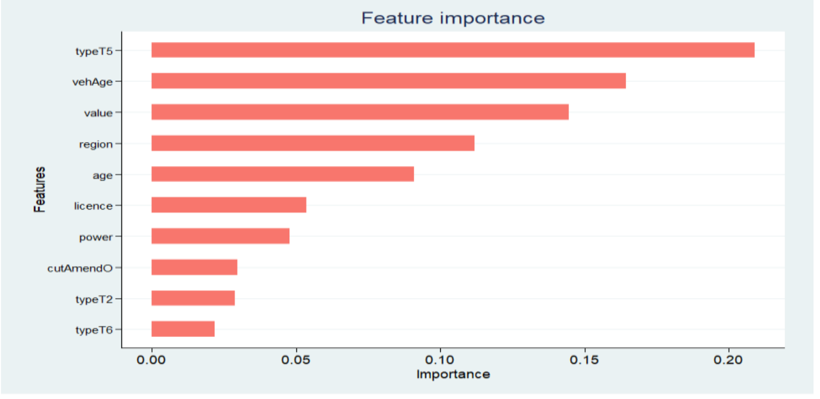
Source: the authors

We can see that XGBoost outperforms all previous models on predicting both in-sample and out-of-sample data set.

However, there is an important downside to using XGBoost which is limited interpretability to how predictions are made. This is the reason why it is commonly characterized as a "black box" model. There still exists some tools that can give some insight on the features that are most correlated to the response variable y. We can also build SHAP plots (SHapley Additive exPlanations), (Lundberg, 2017). The goal of SHAP is to explain the prediction of an input vector x by computing the contribution of each feature to the prediction.

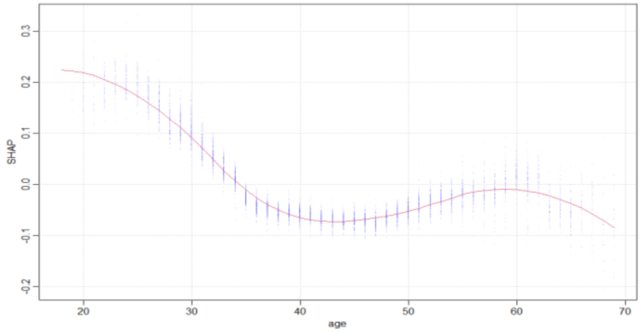
We can now visualize the most important features in our data set, as well as some SHAP plots that can show us how features correlate to the response variable y.

Figure 6. Feature importance.



Source: the authors

Figure 7. SHAP with respect to age of driver.



Source: the authors

Finally, we perform the K-cross validation procedure once more (Figure 8) to ensure that all the machine learning models we have developed so far generalize well to data that were not used during the training step. The results obtained are quite definitive. The deviation in deviance losses across different folds is minimal, indicating that our models can indeed accurately predict out-of-sample data. Among all the models considered, XGBoost continues to exhibit the best performance.

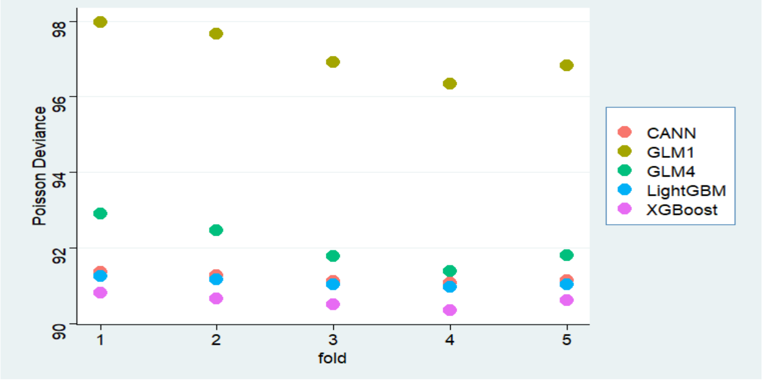


Figure 8. K-cross validation all models.

Source: the authors

3- INTERPRETATIONS

Our research findings suggest that there are numerous ways to efficiently estimate the claims frequency of policyholders using a variety of explanatory variables. The application of such estimations can be beneficial for both insurers and policyholders within the Algerian car insurance market, which currently relies on a less dynamic, more simplified premium calculation model.

From an insurer's perspective, the present market is marked by instability. This instability stems from a lack of consideration for individual risk factors, such as driving history, when setting premiums, with insurers primarily focusing on the value of the vehicle. This approach leaves insurers vulnerable to significant solvency risk and market volatility. Our study reveals that adopting statistical and machine learning methods can lead to considerable improvements in risk assessment and quantification. The adoption of these techniques will enhance insurer solvency, improve return on investment, and lower the probability of ruin.

For policyholders, the implications are twofold. On one hand, the current pricing model creates unfair discrepancies in premium costs. Low-risk drivers often find themselves paying more than their risk profile justifies, while high-risk drivers are, in contrast, undercharged based on their actual risk. This unfairness could lead low-risk drivers to opt out of insurance or seek alternative providers. On the other hand, a more risk-reflective pricing model, like the one we have proposed, can alleviate this issue. Low-risk drivers will appreciate lower premiums corresponding to their risk level, whereas high-risk drivers may be prompted to adapt safer driving habits in response to higher premiums. Thus, an improved pricing model may encourage overall safer driving behavior, benefiting the Algerian society as a whole.

CONCLUSION

Our research has mentioned critical inadequacies in the current ratemaking practices in the Algerian automobile insurance market. The absence of robust loss estimations and over-reliance on the value of the vehicle as the primary determinant of premiums is neither supportive of market stability nor equity. It not only exposes insurers to unnecessary risks and potential insolvency but also perpetuates an unfair system for policyholders, who often end up paying premiums disproportionate to their risk profiles.

We explored multiple statistical and machine learning tools to model claims frequency. We started with the conventional GLM approach, which yielded promising results with straightforward interpretation. Then, we integrated neural networks with GLM to enhance prediction accuracy, at the cost of some interpretability.

Finally, we turned to gradient boosting models, specifically LightGBM and XGBoost. Here, XGBoost outperformed all other models, indicating its potential as a powerful tool in risk assessment and insurance pricing. It is our belief that leveraging such advanced methods can significantly advance the Algerian insurance market. These proposed solutions can lead to a fairer and more efficient system, reducing financial risk for insurance providers while ensuring more equitable premiums for policyholders.

References

Charpentier A., (2015). *"Computational Actuarial Science with R"*. CRC Press.

Dunn P. K., & **Gordon K. S., (2018)** *"Generalized Linear Models With Examples in R"*. Springer Texts in Statistics.

Ferrario A., & **Hammerli R., (2019).** *"On boosting: theory and applications"*. Actuarial Data Science of the Swiss Association of Actuaries SAV.

Ferrario A., Noll, A., & **Wuthrich M. V., (2020).** *"Insights from Inside Neural Networks".* Actuarial Data Science of the Swiss Association of Actuaries SAV.

Garrido J., Genest C., & **Schulz J., (2016).** "Generalized linear models for dependent frequency and severity of insurance claims". *Insurance: Mathematics and Economics* (70). pp.205-215.

Guolin K., Qi M., Thomas F., Taifeng W., Wei C., Weidong M., Qiwei Y., &Tie-Yan L., (2017). "Light GBM A Highly Efficient Gradient Boosting Decision Tree". Advances in N*eural Information Processing Systems* 30 NIP 2017.

Lundberg S., & **Lee, S., (2017).** *"A Unified Approach to Interpreting Model Predictions"*. NIPS'17: Proceedings of the 31st International Conference on Neural Information Processing Systems. Curran Associates.

Noll A., Salzmann R., & **Wuthrich M. V., (2017).** *"Case Study: French Motor Third-Party Liability Claims".* Actuarial Data Science of the Swiss Association of Actuaries SAV.

Noriszura I., (2007). *"Handling Overdispersion with Negative Binomial and Generalized Poisson Regression Models".* Casualty Actuarial Society Forum, Winter 2007, 103-158.

Pesantez-Narvaez J., Guillen M., & **Alcañiz M., (2019).** "Predicting Motor Insurance Claims Using Telematics Data—XGBoost versus Logistic Regression". R*isks*, 7, 70.

Renshaw A.E., (1994). *"Modelling the claims process in the presence of covariates".* ASTIN Bull., 24 (2).

Schelldorfer J., & **Wuthrich M. V., (2019).** *"Nesting Classical Actuarial Models into Neural Networks".* Actuarial Data Science of the Swiss Association of Actuaries SAV.

Tianqi C., & **Carlos G., (2016).** *"XGBoost: A Scalable Tree Boosting System".* CoRR, abs/1603.02754.

Wuthrich M. V., & **Buser C., (2021).** *"Data Analytics for Non-Life Insurance Pricing".* Swiss Finance Institute Research Paper No. 16-68.