Individual Health Insurance Reserving Using Gradient Boosting Technique.

Ayyoub SAOUDI[1], Ghita HAJRAOUI[1], Jamal ZAHI[1]

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

Abstract. This article delves into the significance of individual health insurance reserving and its correlation with the solvency of insurance companies. It highlights the shift from traditional actuarial methods, such as the Chain Ladder method, to the utilization of Gradient Boosting in machine learning. This technique is applied to a Moroccan health insurance company. The results indicate a high accuracy of the model and robust evaluation metrics. Gradient Boosting proves its effectiveness in predicting reserving amounts, thereby enhancing risk management and insurers' solvency. The linearity observed in the graphical results confirms the consistent alignment between predictions and actual values, underscoring the merits of this emerging approach in health insurance.

Index Terms— health insurance, solvency, actuarial methods, Gradient Boosting, machine learning, risk management, insurers, predictive modeling, evaluation metrics, Moroccan insurance industry.

Introduction

Health insurance holds a pivotal role in society, serving as a crucial mechanism to address the escalating costs of healthcare and ensuring equitable access to essential medical services. The accurate estimation of reserves is paramount to an insurance company's ability to uphold its commitments to policyholders. Solvency, intricately linked to the quality of reserving, is now more than ever dependent on precise and sophisticated methods to guarantee the financial stability of insurance providers.

The advent of Machine Learning has ushered in a new era for insurers, providing access to vast sets of Big Data and advanced tools that significantly enhance the accuracy of individual reserving calculations. These innovative approaches empower insurers to conduct in-depth analyses of policyholder data, delving into intrinsic characteristics such as medical history. By leveraging these technologies, insurers can tailor their estimates to the nuances of each policyholder's profile, ultimately elevating the precision of provisions.

Risk management is another critical facet of health insurance. Insurers face the obligation of anticipating potential risks and appraising them accurately. Machine Learning methods, when applied to individual reserving, offer a responsive and dynamic solution for bolstering risk management. This enables insurance companies to safeguard their solvency and adapt swiftly to evolving trends and changes in policyholders' risk profiles. The central challenge in this context lies in optimizing individual reserving to ensure both insurer solvency and policyholder satisfaction.

Against the backdrop of these challenges, the central question emerges: How can insurance companies take advantage of emerging machine learning methods to calculate their individual reserving? This article delves into the practical application of machine learning algorithms, particularly focusing on the Gradient Boosting technique, in the context of individual reserving within a Moroccan health insurance company. The insights provided contribute to a broader understanding of how machine learning methods can fortify company solvency and elevate risk management practices in the realm of health insurance. As the industry grapples with evolving dynamics, embracing these innovative approaches becomes imperative for ensuring the continued efficacy and sustainability of health insurance provision.

1 Health insurance reserving: From actuarial methods to machine learning algorithms

1.1 Traditional Approaches and Limitations

Insurance companies have traditionally estimated their technical reserves using deterministic methods that have long Insurance companies traditionally employed deterministic methods, notably the Chain Ladder method [1], to estimate their technical reserves. Despite their historical reliability, these methods face limitations in capturing the intricacies of individual characteristics and the dynamic risk factors inherent in the insurance market. With the increasing volume of data, especially in healthcare, conventional models are increasingly proving inadequate, unable to fully harness this wealth of information.

Alternative methodologies, including the BornhuetterFerguson, Cape Cod, and Benktander-Hovinen methods, largely extend the core concepts of the Chain-Ladder technique. The application of Run-off triangles allows for a thorough tracking of claims costs, offering valuable insights into their distribution. This, in turn, facilitates the selection of suitable modeling techniques. In cases where claims demonstrate a Gaussian distribution, deterministic methods may be viable, while stochastic methods are more commonly utilized for modeling various other distribution types.[2]

1.2 Evolution to Machine Learning and Gradient Boosting

In response to these challenges, the advent of machine learning methods presents an opportunity to address the shortcomings of traditional techniques in estimating individual reserving for health insurance. Pioneering work by [3] introduced neural networks to individual reserving, marking a pivotal moment. However, new challenges have emerged, particularly in dealing with the nature of data, whether static or dynamic [4]. Accounting for the chronological evolution of certain variables becomes imperative for accurately modeling claims payments.

Recent contributions in the actuarial field have explored the potential of tree-based machine learning algorithms. Notably, [5] developed an approach utilizing Gradient Boosting, comparing its results with traditional aggregated data methods such as the generalized linear model and the Mack model. The distinct advantage of the Gradient Boosting algorithm lies in its exceptional performance on structured data, coupled with rapid computational capabilities.

The training process of Gradient Boosting involves iteratively adjusting weights of training examples, with a focus on poorly predicted examples from previous trees. The final model emerges as a weighted combination of predictions from each tree. Renowned for its efficacy in regression and classification, Gradient Boosting stands out for its high accuracy and adeptness in handling complex data, as emphasized by [6].

2 Methodology

The estimation of individual reserving in health insurance using the Gradient Boosting methodology requires a methodical and multifaceted process, beginning with scrupulous data collection and extending to a thorough evaluation of the resulting model's performance.

Our data source originates from a health insurance company that has chosen to remain anonymous for confidentiality reasons, the specifics of which are deliberately kept confidential to maintain data privacy and security. The dataset in question encompasses settlements of 2019 and comprises a spectrum of variables that encapsulate diverse aspects of insured individuals. These variables contain demographic details such as gender, socio-professional category, and age, as well as pertinent healthcare-related information, including medical procedure category, predisposition to long-term illness, frequency of care, average cost of medical procedures, and the anticipated reimbursement amount for the insured.

To enhance the quality of the dataset and optimize subsequent analyses, various preprocessing steps were meticulously applied. A pivotal aspect of this preprocessing involved the encoding of categorical variables. By transforming categorical data into a numerical format, this step ensures compatibility with the Gradient Boosting model, enabling it to effectively interpret and utilize this information during its developmental phase.

The entire analytical process, spanning from data acquisition to preprocessing and the actual implementation of the Gradient Boosting model, was executed using the Python programming language. Renowned for its flexibility and the availability of rich libraries dedicated to machine learning, Python facilitated seamless data manipulation, model development, and precise performance evaluation.

This systematic approach guarantees a reliable and robust estimation of individual reserving in the realm of health insurance. The resulting model not only captures the inherent complexity of individual characteristics but also stands as a testament to the advancements offered by machine learning techniques, specifically Gradient Boosting. The insights

derived from this process not only provide valuable understanding into the dynamics of health insurance reserving but also establish a sturdy foundation for informed decision-making and strategic actions within the broader landscape of health insurance management and risk assessment.

3 Results

3.1 Model Development and Training

The development of our model for individual reserving was grounded in the consideration of key variables outlined earlier: the insured individual's gender, socio-professional category, age, treatment category, predisposition to long-term illness, treatment frequency, average treatment cost, and the payment amount. These variables collectively form the basis for a comprehensive analysis of the factors influencing individual reserving in the realm of health insurance.

Our dataset covers a considerable volume of information, including a total of 55 245 observations. To ensure the robustness and generalizability of our model, we adopted a two-step process. Firstly, the dataset was divided into two sub-samples: 80% of the observations were allocated to our training set, and the remaining 20% were reserved for testing the model.

The training set, consisting of most observations, served as the foundation for building and adjusting the Gradient Boosting model. This process involved iteratively enhancing the model's predictive capabilities based on the patterns and relationships present in the data.

Subsequently, the testing set, which remained untouched during the model development phase, was employed to evaluate the model's performance. This step is crucial to determine how well the model generalizes to new, unseen data. The 20% of observations dedicated to testing allowed us to measure the model's accuracy, precision, and reliability in predicting individual reserving outcomes.

3.2 Model Evaluation and Performance Metrics

By adopting this meticulous approach of splitting the dataset into training and testing sets, we aimed to create a Gradient Boosting model that not only captures the shades of the provided data but also demonstrates robustness in predicting individual reserving across a broader spectrum. This methodology ensures that our model is well-equipped to handle real-world scenarios and provides a reliable tool for estimating individual reserving in the context of Moroccan health insurance companies.

Table 1. Evaluation metrics for the Gradient Boosting Regressor model

Metric	Value
R^2	90,15%
MSE	669696.8229927918
MAE	159.24306625601463
AIC	711528.037575469
BIC	711580.215777106

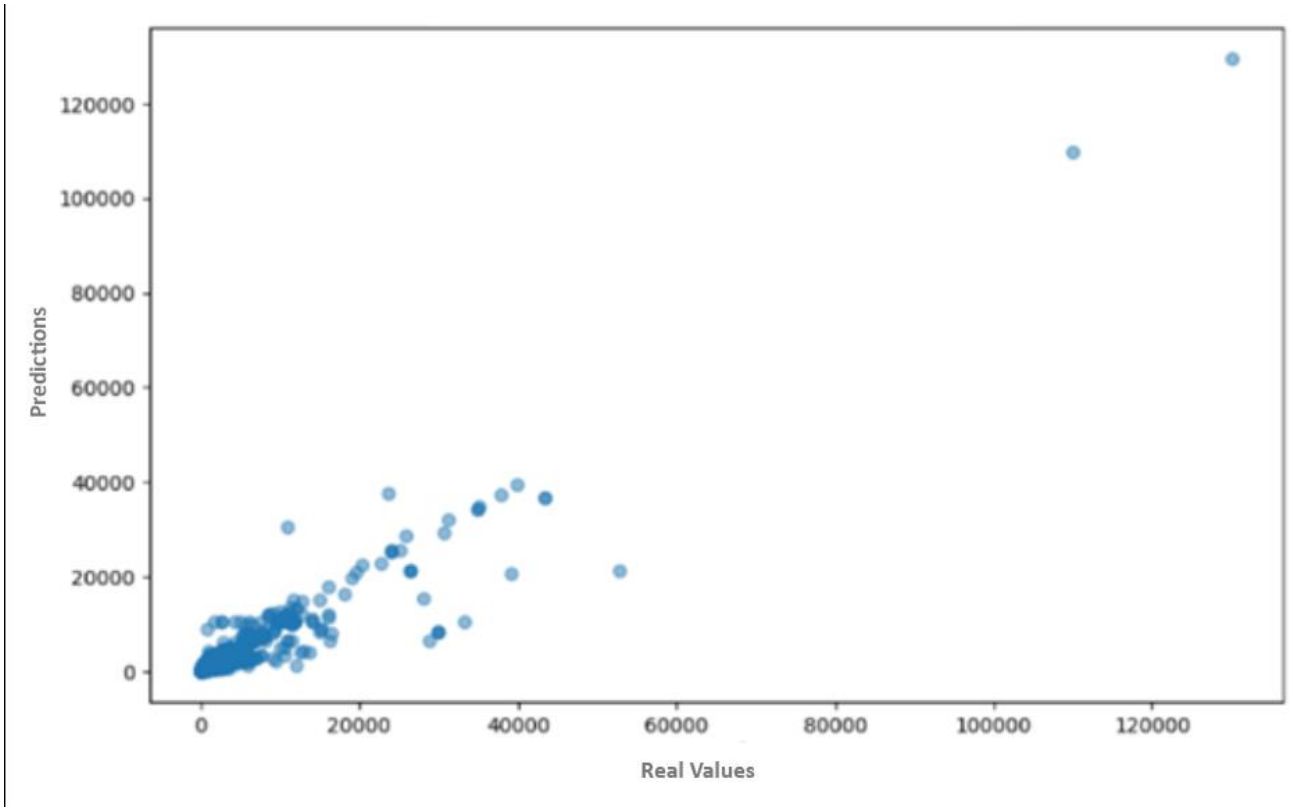
Source: Designed by the authors

The coefficient of determination R^2, standing at 90.15%, indicates the level to which the model's independent variables can predict the variance in the dependent variable. This high R^2 value suggests that the model shows an exceptional overall accuracy, showcasing a robust ability to forecast reserves.

The evaluation metrics for the Gradient Boosting Regressor model further affirm its solid overall performance. The mean squared error (MSE), computed at 669 696.82, represents the average of the squared differences between predicted and actual values. The relatively low MSE indicates acceptable model accuracy, emphasizing its capability to closely match predictions with real-world outcomes. Also, the mean absolute error (MAE) stands at 159.24, indicating a relatively low average of the absolute values of prediction errors.

Specifically, the Gradient Boosting model yields total predicted amounts of **36 895 230.80 MAD**. Information criteria such as the Akaike Information Criterion (AIC) and Bayesian Information Criterion (BIC) are calculated at 711 528.04 and 711 580.22, respectively. These information criteria serve as additional measures of model effectiveness, and the relatively low values indicate that the model performs well in balancing complexity and goodness of fit.

Collectively, these evaluation metrics, along with information criteria, provide a comprehensive view of the positive performance of the Gradient Boosting model in predicting reserve amounts.



Source: Designed by the authors

Fig. 1. Predictions compared to actual values.

Examining the graph illustrating the model's predictions reveals a distinctive pattern: the plotted points form a straight line, diverging from the expected curve. This linear alignment signifies a consistent and direct correlation between the model's predictions and the actual observed values. The absence of a curved trajectory in the graph underscores the model's precision and excellent alignment with real-world data, implying a high level of accuracy.

In statistical terms, this close alignment between predicted and actual values indicates that the model effectively captures the inherent patterns and relationships within the dataset. The lack of a curve signifies the model's ability to provide accurate predictions across a diverse range of values, establishing its reliability in estimating individual reserving within the nuanced domain of health insurance.

Beyond highlighting the model's accuracy, the straight-line pattern suggests a consistent and systematic performance in predictions, reinforcing the suitability of the Gradient Boosting model for the task at hand. This graphical representation not only visually verifies the model's ability in closely matching predictions with actual outcomes but also establishes a sense of trust in its ability to make accurate predictions for estimating reserve amounts in the dynamic context of health insurance in Morocco.

Conclusion

The integration of machine learning methodologies, mainly the application of Gradient Boosting, within the realm of individual health insurance reserving has displayed promising outcomes [6]. Our research indicates that the use of machine learning algorithms helps the insurance companies to ensure their solvency with a high precision.

Concerning the studied Moroccan health insurance company, The R^2 value of 90.15% signifies the model's strong predictive accuracy. The Gradient Boosting Regressor demonstrates solid performance indicating low prediction errors. The model predicts a total of 36,895,230.80 MAD. This application demonstrated heightened accuracy in forecasting reserve amounts, providing insurers with a more effective means to navigate and manage risks [7]. The evaluation process incorporated performance metrics such as R^2, Mean Absolute Error, Root Mean Squared Error, and Mean Absolute Percentage Error, showcasing superior overall performance.

These results not only underscore the potential of machine learning but also shed light on its capacity to enhance risk management, fortify solvency, and cultivate adaptability within the health insurance industry. As technological advancements persist, the incorporation of sophisticated machine learning approaches stands poised to revolutionize the landscape of health insurance, fostering an industry that is more resilient and responsive to dynamic challenges [7].

The incorporation of machine learning, specifically Gradient Boosting, in individual health insurance reserve prediction has proven effective. This research indicates that these algorithms empower insurance companies to maintain solvency. This application provides insurers with improved risk management tools, showcasing the potential of machine learning to enhance risk management and increase industry adaptability. The adoption of advanced machine learning approaches is expected to revolutionize the industry, making it more resilient and responsive to dynamic challenges.

References

1. E. Astesan, Les réserves techniques des sociétés d'assurances contre les accidents d'automobiles (Librairie générale de droit et de jurisprudence, 1938).
2. Saoudi, F. El Kassimi, J. Zahi, Technical reserving in non-life insurance: A literature review of aggregated and individual methods, J. Integr. Stud. Econ. Law Tech. Sci. Commun. **1**(2) (2023).
3. P. Mulquiney, Artificial Neural Networks in Insurance Loss Reserving, in Joint Conference on Information Sciences (2006).
4. G. C. Taylor, G. McGuire, J. Sullivan, Individual claim loss reserving conditioned by case estimates, Ann. Actuar. Sci. **3**, 215-256 (2008).
5. L. Ferraris, P. Liautaud, É. Borel, Méthode de Gradient Boosting (2022).
6. K. Kaushik et al., Machine Learning-Based Regression Framework to Predict Health Insurance Premiums, Int. J. Environ. Res. Public Health **19**(13), 7898 (2022).
7. U. Orji, E. Ukwandu, Machine Learning for an Explainable Cost Prediction of Medical Insurance, Mach. Learn. Appl. **15**, 100516 (2024).