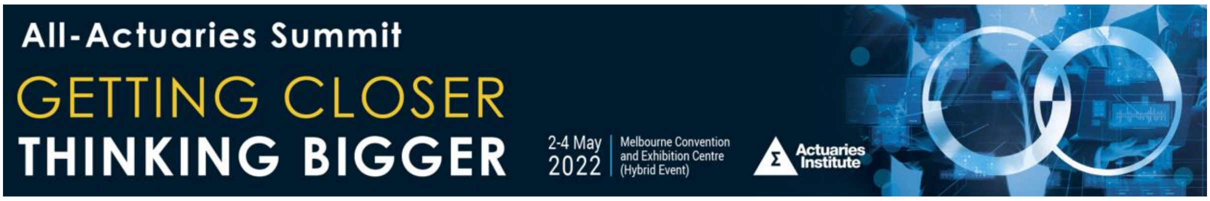

Approaches to Better Utilising Machine Learning Models for Efficient Modelling and Pricing Delivery

Prepared by Zhijing Xu

Presented to the Actuaries Institute
2022 All-Actuaries Summit
2-4 May 2022

Institute of Actuaries of Australia
ABN 69 000 423 656
Level 2, 50 Carrington Street, Sydney NSW Australia 2000
t +61 (0) 2 9239 6100
e actuaries@actuaries.asn.au **w** www.actuaries.asn.au

Approaches to Better Utilising Machine Learning Models for Efficient Modelling and Pricing Delivery

Zhijing Xu

Abstract

Retail insurance pricing can be a lengthy and complex process involving many technical and practical considerations. To respond to rapid market changes, insurers require not only an established framework to conduct pricing reviews, but also the capabilities to translate data into pricing responses in an accurate and efficient manner. One challenge faced by insurers is that output from tree-like models may not be directly implementable, either due to governance concerns or the inability of rating engines to accommodate machine learning models. Another challenge is how to make use of recent data, which may provide valuable insights into emerging trends, yet is underdeveloped.

In this paper, we propose approaches to better utilising machine learning models to improve insurers' pricing capabilities, which could be well integrated into insurers' existing pricing algorithms. The approaches aim to overcome the previously highlighted two challenges and to enable an efficient risk modelling and pricing delivery process, by directly leveraging the machine learning results. A case study is presented to demonstrate the viability and to highlight the advantages of the approaches using actual insurance claim data. The pricing solutions are expected to significantly boost insurers' pricing capabilities under market conditions that change rapidly.

1 Introduction

Retail insurance pricing can be a lengthy and complex process involving many technical and practical considerations. A pricing review may be considered a strategic pricing initiative or a tactical pricing review. A strategic pricing initiative covers all aspects of an end-to-end pricing process (as described in [1]), and is usually carried out when an insurer wants to develop and launch a new product or revamp an existing product to meet observed or perceived market demands. A tactical pricing review addresses divergence between technical premiums (the insurer's expected cost to supply insurance plus a margin) and market prices, or adjusts the insurer's desired exposures to a particular segment. Due to the broader impact of pricing, insurers generally would establish a governance framework for pricing reviews. With increased market competition, regulatory requirements, and more sophisticated customers, it is essential that insurers not only establish an effective governance framework, but also improve the pricing capabilities to translate data into pricing responses in an efficient manner.

The insurance industry has widely adopted general linear models (GLMs) for risk modelling and pricing [2]). The interpretability of GLM and its easy implementation have made it an acceptable framework to build insurance pricing models to ensure compliance in a heavily regulated environment [3, 4]. However, machine learning methods have gained growing attention and become increasingly popular in recent years. Research has demonstrated their capabilities and advantages in reserving [5, 6], fraud detection [7], risk modelling [3, 5, 8-9], pricing optimisation [10], and marketing [11]. Despite the popularity, there have been a few challenges highlighted in the risk modelling and pricing delivery processes when applying machine learning methods.

One challenge faced by the insurers is that results from the tree-like models may not be directly implementable, due to either governance concerns (e.g., transparency and explainability of the resulting insurance premium) or the incapacity of rating engines to accommodate machine learning models. In practice, insurers use different approaches to address this problem, for example, by building GLMs on top of the machine learning model, or manually adjusting existing rating tables along several most significant rating factors. Another challenge is how to make use of recent data, which may provide valuable insights into emerging trends yet is underdeveloped. In short-tail retail insurance pricing, insurers generally avoid the most recent 6 months data when building the models. Although the data might be subsequently used for validation, the exclusion from modelling stage means the latest experience is not reflected in the obtained risk relativities.

In this paper, we propose approaches to overcome the risk modelling and pricing delivery challenges highlighted above. The solutions aim to better utilise machine learning models and have the flexibility to be easily integrated into insurers' existing pricing algorithms without undesired modifications. For the implementation challenge arising with machine learning models, the proposed approach leverages the partial dependence plot to extract risk relativities as substitutes of GLM coefficients. For the underdeveloped data challenge, the proposed approach suggests modelling each risk component using different data periods. The viability and advantages of the proposed approaches are discussed in detail with a case study using actual insurance claim data. The solutions are expected to significantly boost insurers' pricing capabilities and therefore offers a competitive advantage under market conditions which are changing rapidly.

2 Data and Models

2.1 Data

The case study was based on a dataset of approx. 678,000 French motor third-party liability policies observed mostly over one year [12]. For demonstration purposes, we only built a claim frequency model, and ignored claim size, although claim amount was available in the data. The data contains 12 columns with 10 risk features collected on the policies, as shown in Table 1. Three models were built to predict claim frequency.

Table 1. Risk features of the dataset [12]

Column	**Description**
IDpol	The policy ID
ClaimNb	Number of claims during the exposure period
Exposure	The period of exposure for a policy, in years
VehPower	The power of the car (ordered values)
VehAge	The vehicle age, in years
DrivAge	The driver age (≥ 18), in years
BonusMalus	Bonus/malus, between 50 and 350: <100 means bonus, >100 means malus
VehBrand	The car brand (unknown categories)
VehGas	The car fuel, diesel or regular
Area	The density value of the city community where the car driver lives in: from "A" for rural area to "F" for urban centre
Density	The density of inhabitants (number of inhabitants per square-kilometre) of the city where the car driver lives in
Region	The policy region in France (based on the 1970-2015 classification)

2.2 Machine Learning Models

For machine learning methods, we built a regression model using gradient boosting machines (GBMs) [13, 14]. GBMs develop an ensemble of weak-performing decision trees, by training each of the trees on different labels, that collectively formulates strong predictions. They can be used in regression and classification tasks. In this paper, we used an open-source package Catboost [15], given its capability of handling categorical features without one-hot encoding pre-processing. Furthermore, it allows monotonic constraints to be applied to certain features in the model specification. This is of particular importance in pricing practice, which will be discussed in the following section. Figure 1 gives an excerpt of one of the decision trees in the trained GBM and illustrates how the decision tree works. At each node of the decision-making process, the model determines which branch to follow by answering a "yes-no" question using the criteria in the node, and a prediction value will be arrived at the leaf nodes of the tree. The final prediction value is calculated by summing up all prediction values from each of the trees in the ensemble.

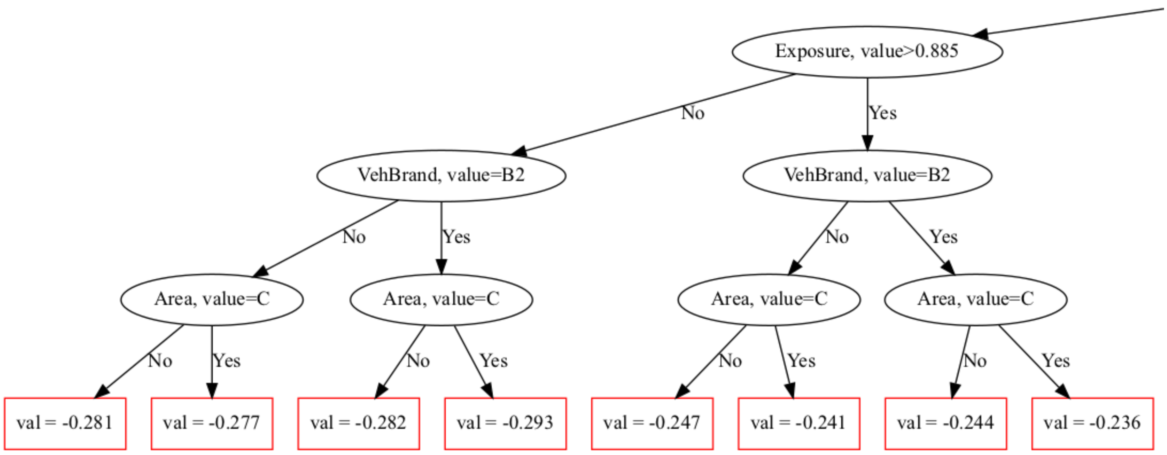

Figure 1. Decision process

2.3 Generalized Additive Models

For GLMs, two models were built using generalized additive models (GAMs) [16]. A GAM is a special type of GLM in which the linear response variable depends on smooth functions of the predictor variables: $g\big(E(Y)\big) = \alpha + \sum_{i=1}^{k} f_i(\,x_i) + \varepsilon$, where the error term ε is independent of the x_i's and $f_i(\,x_i)$ is a smooth function of x_i. The primary advantage of GAMs, compare to GLMs, is that they allow the relationship between the response variable and the predictor variables to be additive, but not necessarily in a monomial form. In the context of insurance pricing, this means the rate of risk change can be different along the range of a predictor variable, which is particularly useful. For example, for motor insurance, the change of risk when the vehicle age turns from 0 (brand new) to 1 year, compare to from 15 to 16 years, could be significantly different. In this paper, a GAM was developed by directly fitting to the raw claim data. Furthermore, a second GAM was built by fitting to the predicted values of the GBM, to acknowledge that insurers often develop an unconstrained machine learning model and refer to as the source of truth regarding the technical premiums.

3 Results

3.1 Model Performance

A GBM and a GAM ("GAM-Raw") were first trained on the raw dataset and a second GAM ("GAM-GBM") was then trained using the predicted values of the GBM. As in Figure 2a, the top four features that explain the highest variations of the risk in the data are: exposure, bonus malus, vehicle age, and driver age, while area only contributes marginally to predicting the risk. To compare the goodness of fit, the log-likelihood values of the three models were calculated. Of the three models, the GBM was found to outperform both the GAMs, and the two GAMs didn't differ significantly from each other. Figure 2b compares the discrimination power of the GBM and GAM-GBM. Compared to the GAM-GBM, the GBM has a higher Gini gain, indicating an overall better fitting, which is consistent with results from log-likelihood values. On individual policy basis, the log-likelihood values suggest the GBM has a better performance for approximately 60% of the 223,500 out-of-sample test samples.

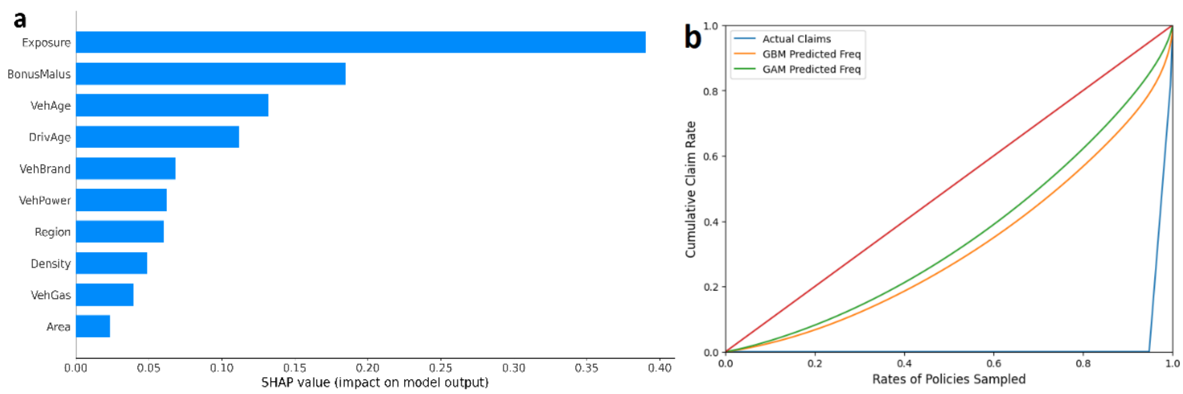

Figure 2. (a) Feature importance from GBM; (b) discrimination power of GBM vs GAM-GBM.

Figure 3 compares the observed and predicted claim frequency of the three models along the top four features. As in the chart, the three predicted curves are generally close to each other, however the GBM does adapt better to the observations by allowing for higher risk variabilities. A particular example is the claim risk by vehicle age. The data indicates a brand-new vehicle (VehAge = 0) incurs a much higher claim risk compared to a vehicle of 1 year old. This feature was highlighted in the GBM by acknowledging that the claim risk would drop significantly when VehAge changes from 0 to 1. However, the risk variation was smoothed out in the GAMs since GAM could only produce a smoothed curve for a numerical predictor variable.

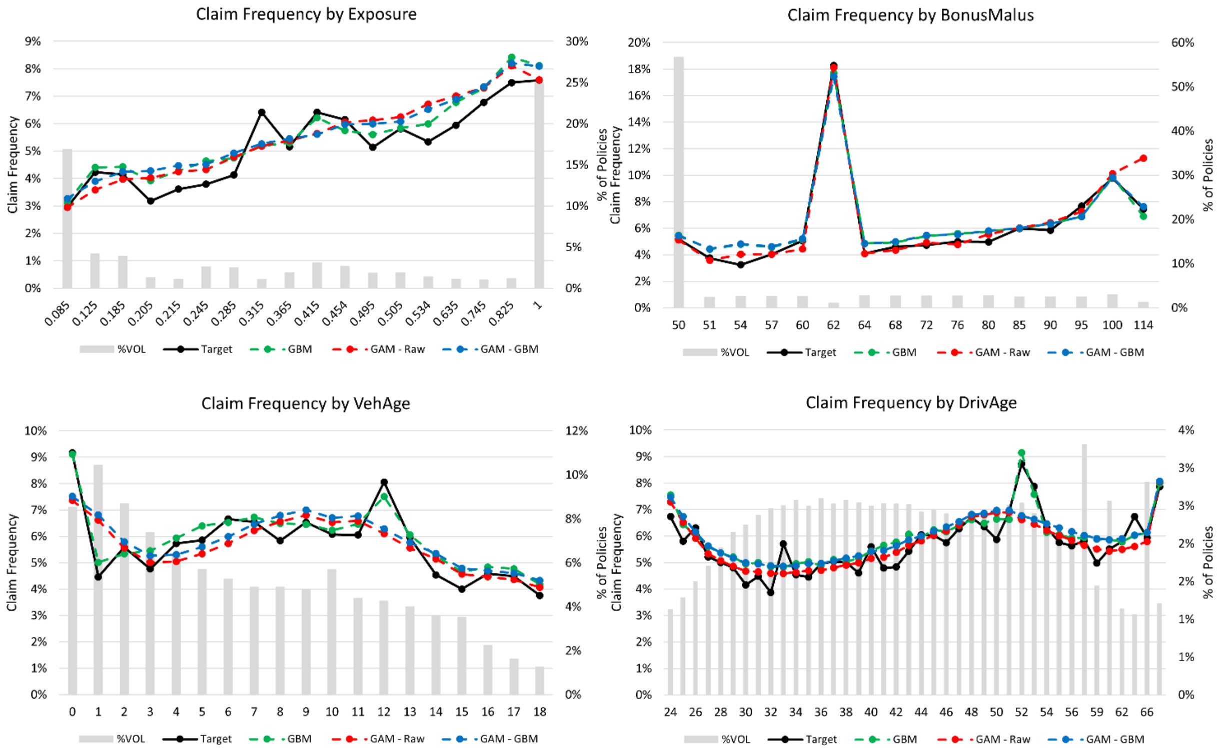

Figure 3. Observed (target) and predicted (GBM; GAM-Raw; GAM-GBM) claim frequency

3.2 Feature Relativity and Market Prices

One of the reasons that GLMs are widely adopted in the insurance industry is their interpretability and ease of implementation. However, the results of machine learning models can also be easily interpreted using Shapley values [17, 18]. We will demonstrate that the partial dependence plot has provided a solution to the direct implementation of machine learning results.

Looking at GLMs, the relationship between the response variable and the predictor variables is given by: $Y \sim g^{-1}(\alpha + \sum_i \beta_i x_i)$. In insurance pricing, a log link function is the most popular choice. In this case, $Y \sim \exp(\alpha + \sum_i \beta_i x_i) = \alpha' \times (\beta_1')^{x_1} \times (\beta_2')^{x_2} \times \cdots \times (\beta_n')^{x_n}$, where $\alpha' = e^{\alpha}$ and $\beta_i' = e^{\beta_i}$. This means essentially, that each predictor variable x_i will contribute to a multiplicative factor $r_i = (\beta_i')^{x_i}$ in the final pricing. In a score card rating engine, variable x_i may only take the value of either 0 (if not belong to a group) or 1 (if belonging to a group). This implies x_i will simply contribute a factor of $r_i = \beta_i'$ in the final pricing. In this case, β_i' can be interpreted as the risk relativity of variable x_i, and α' can be interpreted as the baseline premium. In Shapley values, the partial dependence plot can be interpreted in a similar way. This implies insurers could use the partial dependence function to derive the market prices to be implemented into the rating engine.

Figure 4 gives the risk relativities of the top four features in the three models. Despite the difference, in general, the risk relativities of the three models follow the same pattern, except that the GBM allows for non-smooth risk variations along the predictor variables. In contrast, GAMs can only produce smoothed risk relativity curves for numerical predictor variables. The results imply risk relativities from GBMs can be directly implemented into rating engines for pricing purposes.

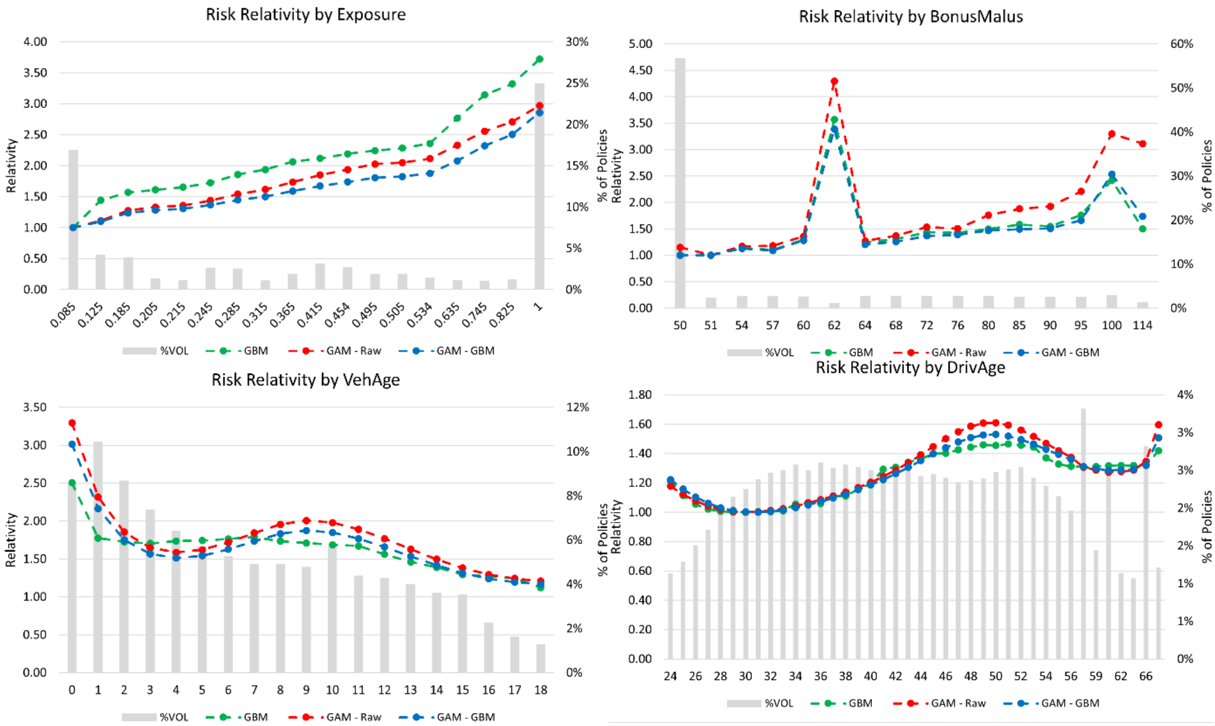

Figure 4. Risk relativity of top four features: exposure, bonus malus, vehicle age and driver age.

The difference in the market prices of the GBM and GAM-GBM was investigated to assess the impact of large premium movements. As in Figure 5a, for 53% of the policies, the difference in the market prices of the two models is less than 10%; and for 94% of the policies, the difference is less than 30%. This indicates that the suggested modelling approach, while more accurately aligning to expected outcomes, is not too radical to be implemented. Figure 5b compares the performance of the derived GBM rating and the GAM rating. It indicates the proposed approach still outperforms the GAM approach despite some loss of the discrimination power during the implementation process.

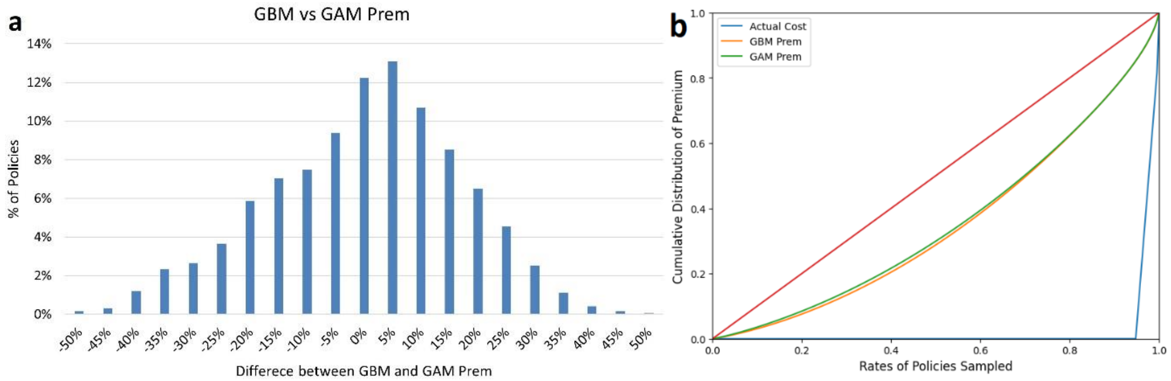

Figure 5. (a) Distribution of difference in the pricing between GAM and GAM-GBM; (b) discrimination power of GBM vs GAM-GBM rating.

3.3 Drifted Experience and Implications

One challenge in insurance pricing is determining how to leverage the most recent experience. In short-tail retail pricing, one common practice is to exclude the most recent 6 months' data when building a model, because it is underdeveloped due to lags in reporting. Although the data might be used to adjust the model outputs in a later validation stage, this is not ideal as the obtained risk relativities may not reflect latest experience.

One observation is that claim frequency and claim severity typically have different development patterns. In practice, for short-tail business, claim frequency tends to be fully developed or close to fully developed much faster than claim severity. Furthermore, the unexpected emerging trends are often characterised by changes in claim frequency. As to the claim severity, the trend is often driven by changes in the inflation and supply chains, which can be anticipated to some extent. In this case, we propose to build two separate models (i.e., claim frequency and claim severity models) to account for the different components of the technical premiums, using different periods of claim data. For example, for short-tail retail pricing, insurers could use one year's claim experience up to one month ago to build a claim frequency model. Insurers could then use one to two years' claim experience, adjusted for anticipated changes in the inflation and supply chains, to build a claim severity model. The longer period of the data used is to account for the higher volatility embedded in claim severity experience. The claim frequency and claim severity model together determine the technical premium for a particular policy.

To demonstrate how the proposed approach could address the emerging trends in claim frequency, we randomly selected 25% policies from the raw data to proximate the latest three months' experience. A claim risk drift was applied to the selected policies, by assuming that claim frequency increased by 2.0%, 4.0% and 6.0% in a particular area (Area = B) during the three months. The same drift assumption was also applied to one of the vehicle brands (VehBrand = B5) to account for potential impact of feature interactions. Two GBMs were then trained separately on the raw data and the drifted data, and a comparative analysis was done based on the two models.

Figure 6a and b compare the risk relativities for area and vehicle brand using the raw data and the drifted data. The GBM trained on the drifted data clearly identified the changes in the claim risk in area "B" and for vehicle brand "B5". More importantly, Figure 6c indicates that due to the drifted claim risk, the explanatory power of the predicator variable area has increased compared to the original model (as in Figure 2a).

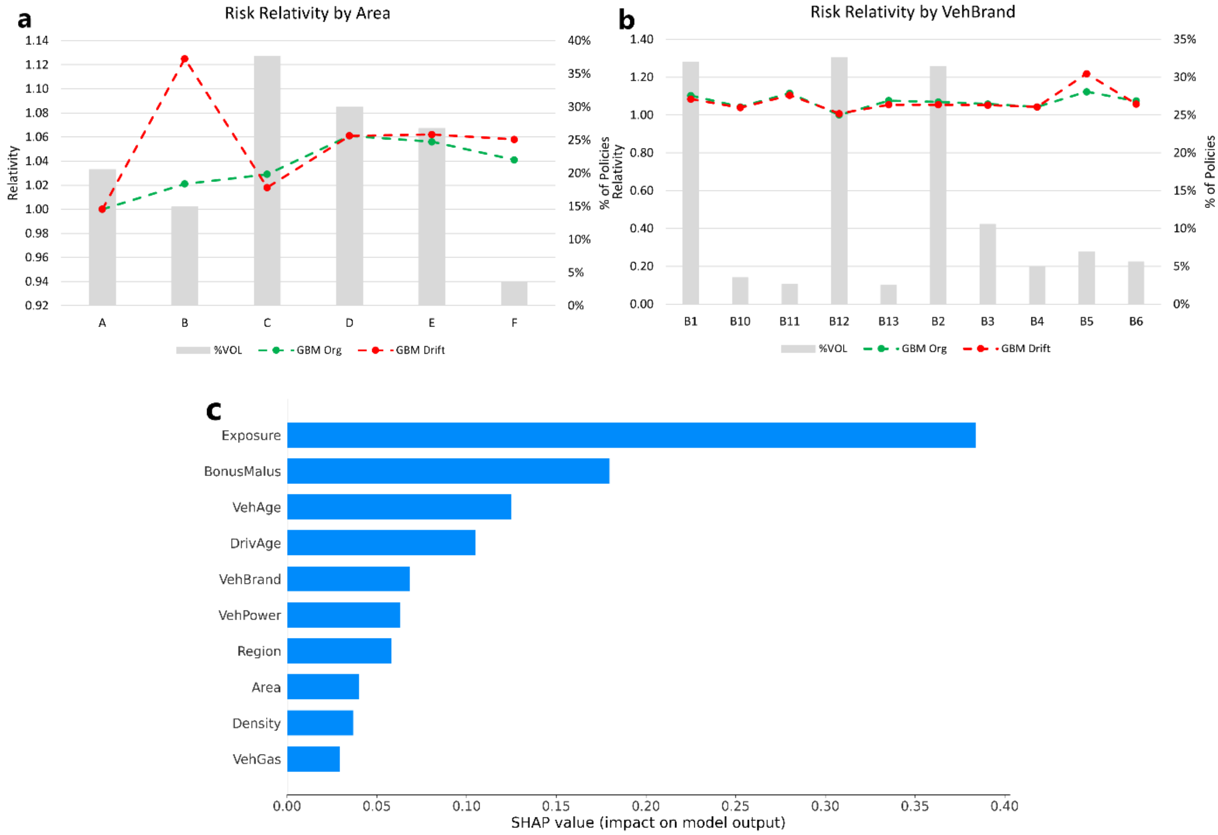

Figure 6. (a)(b) Risk relativity of GBMs based on the raw data and the drifted data; (c) feature importance based on the drifted data.

4 Discussion

In this paper, we propose approaches to overcome two particular challenges faced by the insurers in the risk modelling and pricing delivery processes, i.e., how to implement machine learning results, and how to better utilise more recent data in risk modelling. A case study was performed based on a motor third-party liabilities dataset to demonstrate the approaches in detail.

In section 3.1, the comparison between model performance indicated GBMs can offer a better risk differentiation compared to GAMs. This suggests potential commercial advantages by shifting towards machine learning models in the pricing process. Due to inherent restrictions, for numerical feature, GAMs can only produce a smoothed curve based on a smooth function of the underlying predictor, as shown in the risk relativity curve. One possible solution to allow for higher variations in risk relativities of numerical features in GAMs is to treat them as categorical features. However, additional data pre-processing steps are generally required to group the numerical values into a limited number of groups, which could be a subjective and onerous process.

Section 3.2 further highlighted a solution based on risk relativities from GBMs is practically equivalent to using GLM coefficients. The risk relativities can be easily obtained with the partial dependence plot and implemented into rating engines. This saves the time needed to build a GLM on top of the GBM. Since GBMs offer a better risk differentiation, the approach will help target more accurately on the profitable segments and support sustainable growth for the insurers. An important point needs to be made is that this solution doesn't require a change to the structure of insurers' existing pricing algorithms. A change to the structure of pricing algorithms is often a significant process given the broad impact and potential unexpected consequences due to algorithm flaws and other matters. Our proposed solutions provide better risk modelling and more efficient pricing delivery process that easily fits into the insurers' current business model. However, the proposed solution does support the removal of existing rating factors from, and addition of new rating factors to, current algorithms. This is generally a lower risk compared to a comprehensive structural change to the algorithms. Section 3.2 also demonstrated that our proposed solution will generate a similar position of market pricing compared to the conventional GLM approach, which further substantiates the feasibility of the approach to pricing decision makers. The concept of productionising machine learning models in a reliable and efficient way is similar to MLOps [19], which seeks to increase automation and improve the quality of production models while maintaining compliance with business and regulatory requirements.

Section 3.3 proposed a solution of modelling risk components using different data to address the emerging trends. The practice of building a separate model for each component is quite common in the industry. For example, insurers may build claim frequency and claim severity models; credit institutions may develop probability of default, exposure at default and loss given default models. That said, these models are generally trained on the same dataset. Our solution highlighted the potential benefits of using different data when building the models in consideration of the development nature of the underlying risk. In the example of motor claims, the claim frequency generally develops much faster compared to the claim severity. Besides, the claim cost tends to be highly correlated with the inflation rate and changes in supply chains. Compared to the claim frequency, the trend in the claim cost could be anticipated to some extent based on economic and market research. In this case, insurers could build the claim severity model using an earlier data period subject to adjustments for forward-looking expectations, while building the claim frequency model based on more recent experience. The case study demonstrated that simply ignoring recent experience may not be ideal, as emerging trends will not be reflected in the model. This may have also provided a solution to utilising the experience from the period of the COVID-19 pandemic for many industries. However, we do acknowledge that each industry is different and special considerations need to be made before ingestion of atypical data.

5 Acknowledgements

The author would like to thank Rick Shaw, Mudit Gupta, Elliot Dawson, Nate Xie, Ningfei Zhu and Ally Luo for their valuable comments and feedback during drafting of this paper.

6 References

[1] M. Gupta, "Framework for determining the marketable price of retail insurance," in *2021 All-Actuaries Virtual Summit*, 2021.

[2] A. Diana, J. E. Griffin, J. S. Oberoi and J. Yao, "Machine-Learning Methods for Insurance Applications-A Survey," Society of Actuaries, 2019.

[3] L. Guelman, "Gradient boosting trees for auto insurance loss cost modeling and prediction," *Expert Systems with Applications,* vol. 39, pp. 3659-3667, 2012.

[4] R. Henckaerts, M.-P. Cote, K. Antonio and R. Verbelen, "Boosting insights in insurance tariff plans with tree-based machine learning methods," *North American Actuarial Journal,* vol. 25, pp. 255-285, 2021.

[5] C. Blier-Wong, H. Cossette, L. Lamontagne and E. Marceau, "Machine learning in P&C insurance: A review for pricing and reserving," *Risks,* vol. 9, p. 4, 2021.

[6] K. Ding, B. Lev, X. Peng, T. Sun and M. A. Vasarhelyi, "Machine learning improves accounting estimates: Evidence from insurance payments," *Review of accounting studies,* vol. 25, pp. 1098-1134, 2020.

[7] J. Ai, P. L. Brockett and L. L. Golden, "Assessing consumer fraud risk in insurance claims: An unsupervised learning technique using discrete and continuous predictor variables," *North American Actuarial Journal,* vol. 13, pp. 438-458, 2009.

[8] A. Ghahari, N. K. Newlands, V. Lyubchich and Y. R. Gel, "Deep learning at the interface of agricultural insurance risk and spatio-temporal uncertainty in weather extremes," *North American Actuarial Journal,* vol. 23, pp. 535-550, 2019.

[9] B. Hartman, R. Owen and Z. Gibbs, "Predicting high-cost health insurance members through boosted trees and oversampling: An application using the HCCI database," *North American Actuarial Journal,* vol. 25, pp. 53-61, 2020.

[10] G. A. Spedicato, C. Dutang and L. Petrini, "Machine learning methods to perform pricing optimization. A comparison with standard GLMs," *Variance,* vol. 12, pp. 69-89, 2018.

[11] R. Gupta and C. Pathak, "A machine learning framework for predicting purchase by online customers based on dynamic pricing," *Procedia Computer Science,* vol. 36, pp. 599-605, 2014.

[12] C. Dutang and A. Charpentier, "CASdatasets: Insurance datasets," 2020-12-11. [Online]. Available: http://cas.uqam.ca/.

[13] J. H. Friedman, "Greedy function approximation: a gradient boosting machine," *Annals of statistics,* pp. 1189-1232, 2001.

[14] M. Bowles, Machine learning in Python: essential techniques for predictive analysis, John Wiley & Sons, 2015.

[15] L. Prokhorenkova, G. Gusev, A. Vorobev, A. V. Dorogush and A. Gulin, "CatBoost: unbiased boosting with categorical features," *Advances in neural information processing systems,* vol. 31, 2018.

[16] E. W. Frees, R. A. Derrig and G. Meyers, Predictive Modeling Applications in Actuarial Science - Volume I: Predictive Modeling Techniques, Cambridge University Press, 2014.

[17] S. M. Lundberg, G. Erion, H. Chen, A. DeGrave, J. M. Prutkin, B. Nair, R. Katz, J. Himmelfarb, N. Bansal and S.-I. Lee, "From local explanations to global understanding with explainable AI for trees," *Nature Machine Intelligence,* vol. 2, pp. 2522-5839, 2020.

[18] M. Proksch, H2O, [Online]. Available: https://www.h2o.ai/blog/from-glm-to-gbm-part-2/. [Accessed 07 03 2022].

[19] C. Breuel, "ML Ops: Machine Learning as an Engineering Discipline," [Online]. Available: https://towardsdatascience.com/ml-ops-machine-learning-as-an-engineering-discipline-b86ca4874a3f. [Accessed 14 03 2022].