Non-life insurance: The state of the art of determining the superior method for pricing automobile insurance premiums using archival technique

Sandeile Johannes Buthelezi(a)  Taurai Hungwe(b)*  Solly Matshonisa Seeletse(c)
Vimbai Mbirimi-Hungwe(d)

(a) PhD Student, Department of Statistical Sciences, Sefako Makgatho University of Health Sciences, South Africa
(b) Senior Lecturer, Department of Computer Science and Information Technology, Sefako Makgatho University of Health Sciences, South Africa
(c) Professor, Department of Statistical Sciences, Sefako Makgatho University of Health Sciences, South Africa
(d) Senior Lecturer, Department of Academic Literacy and Science Communication, Sefako Makgatho University of Health Sciences, South Africa

A B S T R A C T

The pricing of insurance premiums in the non-life insurance sector remains a challenging and complex task. It demands a delicate balance between accurately estimating risk exposure and ensuring profitability for insurers. Generalised Linear Regression Models (GLMs) have become the preferred methods for premium price modelling in the motor insurance sector. While the approach of using a single superior model on which predictions are based ignores the use of robust estimator models. This paper examines various methodologies and sheds light on superiority of twenty-two models compared to each other for pricing automobile insurance. These methods vary from traditional actuarial methods to the modern statistical models such as machine learning algorithms. By using archival technique, their inferiority and superiority are explored, considering the ever-changing landscape of risk factors and market dynamics. Furthermore, it highlights the potential benefits of leveraging these methods and the mechanism for pricing short-term insurance, particularly in motor vehicle insurance. It also develops a framework that can be used in pricing to cater to risk analysis constituents to mitigate uncertainties and provide good services to clients. Our findings show that ANN, NN, XG boost, Random forest (RF) are superior models, and we conclude that the modern statistical methods can accurately estimate the risk exposure as compared to traditional methods such as GLMs.

ARTICLE INFO

Article history:
Received 11 February 2024
Received in rev. form 12 March 2024
Accepted 25 March 2024

Keywords:
Archival technique, Automobile Insurance, Insurance Premium, Non-life insurance, Superior method

JEL Classification:
G52O, C450, C610

© 2024 by the authors. Licensee SSBFNET, Istanbul, Turkey. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (http://creativecommons.org/licenses/by/4.0/).

Introduction

Mathematical models are crucial in insurance topics for predicting costs, estimating uncertainty, and estimating claims and outstanding claims, ultimately affecting insurance premium pricing and claims to reserve (Satayatham et al., 2012). Non-life insurance pricing is the practice of determining the cost of an insurance policy while considering the various properties of the insured object and the policyholder. The primary source for use is historical data on policies and claims, sometimes augmented with external data. In a tariff analysis, this data is used to develop a model that depicts how the claim cost of an insurance policy is affected by the various explanatory variables. Generalised linear models (GLMs) are presented as a technique for tariff analysis and have become the standard approach in many countries (Ohlsson & Johansson, 2010). Kaaková and Kvěvánková (2014) for instance, used GLM to analyse a portfolio of automobile insurance data and the main reason for use was that GLMs are not constrained by inflexible preconditions.

* Corresponding author. ORCID ID: 0000-0003-2719-0839
© 2024 by the authors. Hosting by SSBFNET. Peer review under responsibility of Center for Strategic Studies in Business and Finance.
https://doi.org/10.20525/ijrbs.v13i2.3211
Claim severity analysis in automotive third-party liability insurance was conducted using GLM and it was demonstrated that this technique can be used in estimating claim severity, and frequency and determining net premiums for motor insurance (Soltes et al., 2019). GLMs are suggested as appropriate and superior predictive modelling techniques for reviewing auto insurance rate filings (Xie & Lawniczak, 2018). GLM was used to explore Industry-Level Fairness of Premiums across auto insurance and the results showed a good fit (Xie et al., 2022).

While GMLs have become the preferred method for claim modelling in the insurance sector, the approach of utilising a single superior model on which predictions are based ignores the use of robust estimator models. Most often, insurance claim data includes categorical variables, with a large number of levels. In such circumstances, some level subsets may be combined to reduce the overall number of levels for greater model simplicity and interpretability. This limits the full use of data in the disposal and may lead to biases and understating of risk assessment.

This study aims to investigate modern statistical methods against classical methods (GLMs) for superiority with the intent to give insight into pricing methodologies that can be utilised for premium pricing and further highlight the potential benefits of leveraging them. Moreover, it intends to transform the current mechanism and develop a framework that can be used to cater for risk analysis constituents to mitigate uncertainties and provide good services to clients.

**Literature Review**

Vehicle Insurance is a protection scheme that gives financial protection to cars, trucks, motorbikes, and other road vehicles against any car damages and bodily injury caused by traffic accidents. It also protects against liability in the event of a traffic collision. The specific terms and types of auto insurance differ according to legal constraints (Denuit et al., 2007). These protection schemes are derived from a process called vehicle insurance risk selection, where the insurers evaluate whether to insure an applicant and what premium to charge. The insurance premium is often calculated based on the annual frequency of claims, which is calculated using statistical data (Kaae, 2008). Traditional car insurance pricing models base the premium on self-reported rating variables (e.g., age and postal code), which capture characteristics of the policyholder and the covered vehicle, however, they are usually only indirectly related to the accident risk (Verbelen et al., 2017).

Recently, many insurance topics included the development of mathematical models that may be used to anticipate or predict insurance costs. As a result, modelling is a crucial aspect in estimating the degree of uncertainty of when claims will be made and how much will be paid. The modelling of claims and outstanding claims results in the pricing of insurance premiums and the estimation of claim reserving (Sattayatham et al., 2012). The study will review the relevant literature to methods used to determine pricing of motor insurance.

For instance, Bayesian model averaging was used as a hybrid model to combat such challenges (Hu, 2018). Osei and Yooku (2023) applied the Double Stochastic Poisson Process for modelling premiums for Pay-As-You-Drive data and suggested it as a best fit. The following machine learning techniques: logistic regression, decision tree, random forest, xBoost, and feed-forward network were investigated and used to predict claims severity and they were suggested to be superior techniques (Baran & Rola, 2022). A supervised driving risk scoring neural network (NN) model was also proposed and shown to predict accurate premiums with discounts (Meng et al., 2022). The combination of the Back Propagation Neural Network and credibility theory was investigated and suggested that they can conduct accurate claim amount estimation and pricing for vehicle insurance, which can effectively improve the current situation of the automotive insurance companies and encourage the development of the insurance industry (Yu et al., 2021). The use of multivariate mixed models to characterise the simultaneous dynamics of telematics data and claim frequencies was proposed. Furthermore, it was suggested that future premiums can then be calculated based on the anticipated distribution of claim characteristics given past data (Denuit et al., 2019).

**Research and Methodology**

The archival technique is a method used to draw inferences about a study (Ventresca & Mohr, 2017). In its most classic sense, the technique involves the study of historical documents, that is, documents created at some point in the relatively distant past, providing us access that we might not otherwise have to the individuals, and events of that earlier time. This technique is also employed by scholars engaged in non-historical investigations of documents and texts produced by and about contemporary organisations, often as tools to supplement other research strategies (field methods, survey methods, and so on.) (Ventresca & Mohr, 2017). Thus, in our study, it is used with the aim of analysing digital texts, including electronic databases and web pages on articles that compare classical statistical methods against modern methods such as RF and NN, among others. These articles range from the period of 2018 to 2023, and they were selected based on the criteria that they trailed and tested against other statistical methods (comparison) to a price premium for motor insurance as shown in the sampled table, Table 1. The dataset consists of twenty-two methods reviewed.
Table 1: Sample of data collected (articles) by the archival technique

<table>
<thead>
<tr>
<th>Year</th>
<th>Datatype</th>
<th>Model</th>
<th>Performance</th>
<th>Type of Model</th>
<th>Group Category</th>
</tr>
</thead>
<tbody>
<tr>
<td>2020</td>
<td>Claims Data</td>
<td>Neural Network</td>
<td>Superior</td>
<td>Machine Learning</td>
<td>Modern Statistics</td>
</tr>
<tr>
<td>2020</td>
<td>Claims Data</td>
<td>Autoregressive Integrated</td>
<td>Inferior</td>
<td>Traditional</td>
<td>Classical Statistics</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Moving Average</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2020</td>
<td>Claims Data</td>
<td>Generalized linear model</td>
<td>Inferior</td>
<td>Traditional</td>
<td>Classical Statistics</td>
</tr>
<tr>
<td></td>
<td></td>
<td>(GLM)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2019</td>
<td>Telematics Data</td>
<td>Random Forest</td>
<td>Moderate</td>
<td>Machine Learning</td>
<td>Modern Statistics</td>
</tr>
<tr>
<td>2019</td>
<td>Telematics Data</td>
<td>Logistic Regression</td>
<td>Moderate</td>
<td>Machine Learning</td>
<td>Modern Statistics</td>
</tr>
</tbody>
</table>

Details on modern statistical methods

Modern statistical methods play a crucial role in insurance pricing. They revolutionize the way premiums are determined and risks are assessed. These methods leverage data science, actuarial statistics, and financial models to enhance accuracy and efficiency in pricing strategies. Traditionally, insurers relied on actuarial tables and historical data for risk assessment, but the advent of data science has transformed this landscape. The integration of data science allows insurers to move beyond traditional methods (Verbelen et al., 2018). By harnessing the power of data analytics, insurers can now assess risk more accurately and calculate premiums based on a more comprehensive understanding of individual risk profiles. Actuarial statistics form the backbone of insurance pricing (Parodi, 2023). These statistics are essential for risk assessment, pricing, reserving, and claims analysis. They provide the framework for determining premiums by evaluating various risk factors and predicting future outcomes. Modern financial theory offers methodologies to mitigate risks related to interest rate fluctuations (Xie et al., 2022). This is particularly valuable for insurance firms managing portfolios of fixed assets, enabling them to optimise pricing strategies and reduce financial uncertainties. Applying modern statistical methods is key in defining premiums and enhancing risk assessment in general insurance (Grize, 2015). By utilising advanced statistical techniques, insurers can uncover new insights, refine pricing models, and improve decision-making processes. Therefore, the convergence of data science, actuarial statistics, and financial models has reshaped insurance pricing by introducing more sophisticated and accurate approaches to premium determination and risk evaluation. Insurers now have powerful tools at their disposal to navigate the complexities of the insurance landscape with greater precision and insight.

Scope of comparative analysis

To compare modern statistical methods and generalised linear models (GLMs), we need to consider the following criteria and parameters model assumptions, model complexity, model flexibility, model estimation, model diagnostics, model interpretation, model selection, computational efficiency, handling missing data, and handling complex structures (Biecek & Burzykowski, 2021; Henley et al., 2020). GLM are more explicit in assumptions, are suitable when there is complexity, are relevant in estimation, diagnostics, interpretation, efficiency, missing data, but is limited or lacks flexibility. That means the dynamism of modern data may become unsuitable for GLM.

Explicitly stating potential benefits

Insurance pricing can offer various benefits to both insurers and customers. Dynamic pricing, for example, is seen as a potential paradigm shift that can incentivise customers to adopt safer behaviours like defensive driving or proactive health management (Schlereth et al., 2018). Personalised approaches can lead to more tailored and fair pricing structures. Moreover, according to Hodula et al. (2021), insurance premiums, whether for life insurance or health insurance, allow individuals to obtain and maintain coverage by paying a monthly or annual amount. Understanding the factors that influence group employee benefits and general insurance pricing, such as upselling/cross-selling potential, taxes, and investment income, is crucial in determining pricing strategies (Parodi, 2023). Insurance companies have a strong incentive to ‘risk price’ effectively by using sophisticated rating factors to accurately assess risk and offer competitive pricing (Frees & Huang, 2023). Though this approach benefits both insurers and policyholders by ensuring fair premiums that reflect the level of risk involved, they may not be leading to ideal or optimal level. Hence, there are benefits of comparing models. When comparing insurance pricing, several potential benefits to consider include cost saving, better coverage, avoiding over insurance, efficiency and convenience, transparency, competition and cost-effective marketing (Haller et al., 2024). By actively and effectively comparing insurance options, individuals can make informed decisions that align with their needs and budget while potentially saving money and ensuring optimal coverage.

Clarity of framework development

The envisaged framework for risk analysis typically includes three main components: risk assessment, risk management, and risk communication (Kitsios et al., 2022). Risk assessment involves identifying and characterizing risks, while risk management focuses...
on controlling and mitigating risks. Risk communication, on the other hand, involves consulting with stakeholders regarding risks and ensuring that all parties are informed about the risks and their potential impacts. The framework also includes risk context, which defines the scope and boundaries within which risk is assessed, managed, and communicated. Risk criteria are used to evaluate the significance of risks, and risk estimates are determined by a blend of result and likelihood assessments. Risk identification involves postulating risk scenarios and determining those that warrant detailed risk characterisation. Risk treatment involves selecting and implementing measures to reduce risk, and stakeholders are those people and organisations that may affect, be affected by, or perceive themselves to be affected by a decision, action, or risk. The context of the framework is also important, as the final form has to be appropriate and fit its purpose.

Findings and Discussions

Findings

Case I: Data Structure

The study sampled articles published on electronic databases and web pages contributing to 22 statistical methods reviewed ranging from classical to the modern methods, divided into two datasets known as claims and telematics data. In Figure 1, we observed that claims data have been the most favoured dataset compared to the telematic dataset even though it has gained more attraction in recent years especially in the period of 2019 to 2023. Secondly, we also observe that more modern statistical methods have started gaining attraction, especially from the year 2019 to date as compared to the classical methods, even though the classical methods are still preferred.

![Figure 1](image)

Figure 1: The boxplot that shows the Data type used in the period of 2018 and 2023 in pricing modelling

Case II: Performance

The twenty-two methods ranging from traditional to deep learning were observed based on their performance (superior, inferior, and moderate) stated in Table 2. It is seen that traditional methods such as GLMs, Linear Regression, Multiple linear Regression, Autoregressive Integrated Moving Average, and Actuarial Methods perform inferior (poor) when compared to modern statistical methods, while machine learning (ML) such as decision trees, support vector machine, and RF contributed 50%, 33% and 10% of inferiority respective. It is also observed that most of the modern methods are moderate (better) compared to the classic methods, while DL methods have been shown to be superior when compared to both ML and traditional methods. We further noticed that methods developed under the NN methodology have more superiority with a 100% convergence. Moreover, it is also noted that GLMs perform superior when fused with other methods (Hybrid) with 100% convergence, while 17% of the time it has proven to be superior when tested sole with a failure rate of 83% (i.e., success of only mere 17%). Overall, modern techniques show superiority when compared to traditional methods of seating with a performance of 37% superiority and 69% combined with moderate performance.
### Table 2: The table that shows the performance of inferiority and superiority of different modelling techniques

<table>
<thead>
<tr>
<th>Performance</th>
<th>Inferior</th>
<th>Moderate</th>
<th>Superior</th>
</tr>
</thead>
<tbody>
<tr>
<td>Actuarial Method</td>
<td>100%</td>
<td>0%</td>
<td>0%</td>
</tr>
<tr>
<td>Autoregressive Integrated Moving Average</td>
<td>100%</td>
<td>0%</td>
<td>0%</td>
</tr>
<tr>
<td>Bayesian network</td>
<td>0%</td>
<td>100%</td>
<td>0%</td>
</tr>
<tr>
<td>Combined Actuarial Neural Networks</td>
<td>0%</td>
<td>0%</td>
<td>100%</td>
</tr>
<tr>
<td>Decision Trees</td>
<td>50%</td>
<td>0%</td>
<td>50%</td>
</tr>
<tr>
<td>Deep Neural Networks</td>
<td>0%</td>
<td>0%</td>
<td>100%</td>
</tr>
<tr>
<td>Gaussian Process regression</td>
<td>0%</td>
<td>0%</td>
<td>100%</td>
</tr>
<tr>
<td>Generalized Additive Models for Location, Scale and Shape (GAMLSS)</td>
<td>0%</td>
<td>0%</td>
<td>100%</td>
</tr>
<tr>
<td>Generalized Linear Model (GLM)-Generalized Pareto Distribution (GPD)</td>
<td>0%</td>
<td>0%</td>
<td>100%</td>
</tr>
<tr>
<td>Generalized linear model (GLM)</td>
<td>83%</td>
<td>0%</td>
<td>17%</td>
</tr>
<tr>
<td>Gradient Boosting Regressor (GBR)</td>
<td>0%</td>
<td>50%</td>
<td>50%</td>
</tr>
<tr>
<td>K-nearest neighbours</td>
<td>0%</td>
<td>100%</td>
<td>0%</td>
</tr>
<tr>
<td>Linear Regression</td>
<td>100%</td>
<td>0%</td>
<td>0%</td>
</tr>
<tr>
<td>Logistic Regression</td>
<td>0%</td>
<td>100%</td>
<td>0%</td>
</tr>
<tr>
<td>Multiple Linear Regression</td>
<td>100%</td>
<td>0%</td>
<td>0%</td>
</tr>
<tr>
<td>Neural Network</td>
<td>0%</td>
<td>20%</td>
<td>80%</td>
</tr>
<tr>
<td>Random Forest</td>
<td>10%</td>
<td>60%</td>
<td>30%</td>
</tr>
<tr>
<td>Support Vector Machine</td>
<td>33%</td>
<td>67%</td>
<td>0%</td>
</tr>
<tr>
<td>XG Boost</td>
<td>0%</td>
<td>54%</td>
<td>46%</td>
</tr>
<tr>
<td>artificial neural network (ANN)</td>
<td>0%</td>
<td>0%</td>
<td>100%</td>
</tr>
<tr>
<td>fuzzy logic</td>
<td>0%</td>
<td>100%</td>
<td>0%</td>
</tr>
<tr>
<td>Tab Net</td>
<td>0%</td>
<td>0%</td>
<td>100%</td>
</tr>
<tr>
<td><strong>All</strong></td>
<td><strong>31%</strong></td>
<td><strong>32%</strong></td>
<td><strong>37%</strong></td>
</tr>
</tbody>
</table>

### Case III: Superior Models

Figure 2 shares insights on the most favoured and most superior models within the twenty-two tested models. The results show that K nearest neighbours seem to be one of the most favoured algorithms, and it started drawing more attention from the year 2018 to date, followed by logistic regression. These models are labelled or adjudged moderate when compared to other statistical methods. XG Boost and RF models seem to be more likely to perform moderate or superior when compare, while ANN and NN models shown to be superior. Some of the 22 models show no conclusive evidence of their performance since there are no-enough evidence when trailed and tested against other methods. On the other hand, GLM seem to perform inferior most of the time. However, there are times that this model would outperform some of the modern methods given the fact that one of the weaknesses of ML models is to underperform when the dataset used to train the algorithms is small. This is when traditional methods would take superiority. Overall, ANN, NN, XGB, RF are labelled superior given the fact that they have been trailed and tested many times as compared to other methods.
Figure 2: The performance of superiority and inferiority of the period of time

Case IV: Hypothesis Test

This section investigates the claims that traditional methods such as GLMs are suggested to have appropriate and equal superior predictive modelling techniques for reviewing auto insurance rate filings similar to modern statistical techniques.

Test of hypothesis 1: The first hypothesis focuses on Traditional methods of understanding the conditions of the contract. The null and alternative hypotheses are:

\[ H_0: \text{Traditional methods have appropriate and equal superior predictive modelling techniques for reviewing auto insurance rate filings similar to modern statistical techniques.} \]

\[ H_a: \text{Traditional methods do not have appropriate and equal superior predictive modelling techniques for reviewing auto insurance rate filings similar to modern statistical techniques.} \]

Level of significance: \( \alpha = 0.05 \)

p-value = 0.000

Test verdict: Given the above results, the null hypotheses should be rejected. Under the 95% level of confidence modern statistical methods have a significant impact on reviewing auto insurance rate filings.

Furthermore, the results on the matrix, Figure 3 evidence a strong relationship between the Group Category (Modern Statistics and Classical Methods) and the performance of the model with a correlation of 52%, while a 62% correlation is between the model used
and the group category. This result implies that the model used has a direct influence on the outcome performance and the accuracy of pricing premiums for auto insurance.

![Figure 3: The table that shows the relationship between variables](image)

**Discussion**

Modern statistical methods have started gaining attraction, especially from the year 2019 to date as compared to the classical methods, even though the classical methods are still preferred. We also noted that the claims data is the most used historical data as compared to the telematic dataset even though recently more studies are being done on it. Furthermore, we observed that machine learning and deep learning techniques are superior methods when compared to the classical methods.

However, one traditional method known as the GLMs has shown to be superior when tested against other methods such as machine learning algorithms. This finding alluded to one of the weaknesses that machine learning algorithms tend to be inferior when tested against a small dataset especially methods such as Support Vector Machines, Decision Trees, and RF. Moreover, we also found that the GLMs method tends to be superior when fused with other statistical methods to create a hybrid method. Overall, ANN, NN, XGB, and RF are found to be superior methods given the fact that they have been trailed and tested many times as compared to other statistical methods.

We also noted that more research needs to be done on most of the models to stress their superiority, while gaps in creating hybrid traditional methods may be of interest especially when the modern statistical methods perfume poor due to small data provided for testing as seen in our founding. This framework has created a basis for researchers to further explore these methods with the aim of finding more superior of methods that can be used to price for non-life insurance. Our results have proven that traditional methods do not have appropriate and equal superior predictive modelling techniques for reviewing auto insurance rate filings similar to modern statistical techniques.

**Conclusions**

Modern statistical methods have gained popularity since 2019, with claims data being the most used historical data. Machine learning and deep learning techniques are superior to classical methods. Traditional methods like GLMs show superiority when tested against small datasets and when fused with other methods. ANN, NN, XGB, and RF are considered superior methods. Further research is needed to emphasize their superiority, and gaps in hybrid traditional methods may be of interest to small datasets where machine learning methods prove to be inferior. This framework encourages researchers to explore these methods in finding more superior predictive modelling techniques for reviewing auto insurance rate filings.

**Limitations of the study**

The data lacks detailed information explaining conditions that led to the methods being dropped and new ones being introduced to understand the historical developments and the contexts of each method. Such information could shed light on the historical dynamics of insurance methods in order to enable forecasting and development of robust methods in insurance pricing.
Suggestions for future research

On suggestions for future research, one may use data analytics in non-life insurance pricing to develop new algorithms or machine learning models to improve the accuracy and efficiency of pricing methods. Another area of further research may be to examine the potential benefits of incorporating artificial intelligence and machine learning techniques into non-life insurance pricing. This could involve developing new methods for setting prices based on the properties of the insured, as well as exploring the potential for automating pricing processes. Another way may be to explore the role of risk assessment in non-life insurance pricing, particularly in the context of automobile insurance. This could involve developing new risk assessment models or refining existing ones to improve the accuracy of premium calculation. We may also investigate the relationship between different pricing methods and customer behaviour in automobile insurance. This could involve examining how different pricing methods affect customer satisfaction, loyalty, and retention, as well as exploring the potential for personalised pricing based on individual customer characteristic.

References


Publisher’s Note: SSBFNET stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.

© 2024 by the authors. Licensee SSBFNET, Istanbul, Turkey. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (http://creativecommons.org/licenses/by/4.0/).

International Journal of Research in Business and Social Science (2147-4478) by SSBFNET is licensed under a Creative Commons Attribution 4.0 International License.