AI-Enhanced Predictive Analytics for Insurance Claim Frequency

By Dr. Heejin Choi

Professor of Computer Science, Gwangju Institute of Science and Technology (GIST)

1. Introduction

This paper is concerned with AI-enhanced predictive analytics, which has been penetrating diverse areas of life, firms' activities, and public and private organizations for the last two decades. In the insurance business, AI technologies have been getting ever deeper integration with the peculiar characteristics of the various sectors of the industry, with the aim of rewriting the traditional paradigms of operations and decision-making. Understanding the grounding principles and the issues of predictive analytics, possibly enhanced by AI technologies, is of paramount importance not only for researchers but also primarily for practitioners. In this paper, we focus specifically on one dimension of AI-PA, that is, AI-PA for insurance claim frequency.

AI-PA for the frequency of claims in most sectors of the insurance industry has become a critical issue to deal with given the challenges related to an ever more efficient underwriting as well as pricing of insurance products and services based on multiple kinds of files available both publicly and privately. No wonder, then, that today, the cutting edge for AI-PA is primarily located in the area of motor insurance; the need to predict the frequency of claims could, indeed, not be more important! Most companies working in the area of motor insurance are continuously using the latest in AI to improve the precision of risk selection. The latter is often addressed through occupancy rating models. However, the development of AI-PA for insurance claim frequency has opened up novel approaches to this fascinating body of research and practice. In the strategic tension generated between operational loneliness and immovable communication hype, a number of academic contributions have been produced. Once a historical introduction has been provided, this paper takes a closer look at how an AI-PA model is built, works, and is interpreted. The paper ends with a discussion of the salient features of these AI-PA tools and some of the challenges, opportunities, and implications that these may pose for the modern insurance markets. In particular, some suggestions for rethinking the insurance industry are presented.

1.1. Background and Significance

Leading insurance companies realize that the key to making underwriting and pricing decisions in the most uncertain of industries requires the use of the most accurate predictive tools. This subsection intends to place the study within a broader context by approaching the changes in the field of predictive analytics tools used by the insurance industry over the past two decades. The use of predictive factors that could be modeled with the aid of decision trees, random forests algorithms, CART, artificial neural networks, various regression models, and their "associative learning" counterparts has indeed widely changed the traditional claim data analysis by diachronically recommending these approaches through the most well-known scientific journals specialized in the application of collective intelligence methods within industrial data analysis.

The transformation to business as usual took time until both practitioners and the scientific platform endorsed the new methods of predictive analysis. The existing AI-supported vision of predictive analytics thus brings significant promises and technological advancements that solve “old” data and analysis problems. Machine learning and AI achievement challenge current organizational responsibilities directly and provide better, more immediate prospects in assessing, managing, and enhancing the complexity of risks facing artificial intelligence companies. These technological advancements will influence the operations of the P&C insurance companies directly, as predictive analytics tools in the U.S. have reduced the loss adjustment costs per claim from $47 to $14, and in the UK, fraud costs were expected to be cut by 40% in 2019. The progression of AI-supported tools has potential economic impacts labeled as some of the most important innovations.

1.2. Research Objectives

The aim of this study is to investigate whether AI-enhanced predictive analytics impacts the claim frequency prediction. Few studies have qualitatively shown that the latest AI methods have the ability to enhance predictive accuracy in different fields, including the insurance domain. There is still a gap in the literature that provides empirical evidence in this domain that confirms the claim frequency prediction. Therefore, first, we compare a machine learning technique with traditional GLM in predicting the claim frequency. Second, more advanced techniques in deep learning and image processing models are applied to analyze and measure

the effects of including the MRI image predictors in the claim frequency model. Such findings extend our horizon in understanding the power of AI predictors, computational capabilities, and the suitability of using most of the techniques in our dataset, which is not in the literature yet. Therefore, it is necessary to quantitatively ascertain that AI can contribute to the improved predictive accuracy of claim frequency. Such results could appeal to at least two parties. For practitioners, the best predictor would allow them to use the models to aid their decision-making in terms of rate making, reinsurance treaties, or even for marketing purposes related to personalized insurance. For researchers, such evidence might trigger possible research linked to software development in the future and expand their contribution by combining both approaches used in the literature.

2. Theoretical Framework

The presented study deals with predictive analytics in casualty actuarial science, aimed at quantifying the most important operational risk of an insurer – the frequency of indemnity claims. Prediction of claim frequency and the risks associated with the forecasted value lies at the very core of the business of property-liability insurance. To calculate premiums, an insurer must assess whether to write a policyholder into his or her corresponding portfolio and at which price. The decision implies creating liability, as indemnity claims will be borne and loss reserves have to be deposited. Replacement of bank-like liabilities into insurance contracts requires acceptance of uncertainty concerning indemnity claims. In this paper, we focus on the forecast of the expected waiting time until the occurrence of loss events. Our forecast horizon is one policy period ahead. Since indemnity claims follow the laws of large numbers as formulated in the classical insurance risk models, the frequency of future indemnity claims can be forecasted by considering the expected waiting time in a homogeneous Poisson risk model.

The theoretical foundation is made up of several theories in various topics. The perspective concerning risks and how to manage them is formed based on literature in risk management in the field of corporate finance. An integral part of how insurers handle risks is based on the management of contracts and claims, such as how much data they should use to anticipate risk, the relevance of risk measures to decision makers, such as managers or regulators, and the type and quality of tools that such decision makers prefer. The role of agents in the

insurance market is also important, since their role in pricing insurance policies has implications for the way they forecast claim frequency. The use of predictive techniques from artificial intelligence and theoretical perspectives in other fields is discussed, in which the area of prediction is viewed in different societal aspects.

2.1. Predictive Analytics in Insurance

Changes in the socio-economic environment, as well as a volatile risk landscape, require insurance companies to constantly adapt and augment their traditional insurance business model. One way this is done is by learning more about the underlying and interdependent risk factors. Consequently, the effective use of advanced methods and predictive analytics is increasingly applied in the insurance industry. Actuaries have been deploying statistical tools explicitly designed to forecast claim frequency and/or claim severity, among many others, to identify trends and better understand the company's insured portfolio. A multitude of different techniques exist that can be utilized to learn interdependencies and patterns in historical claim data. More specifically, actuaries use the statistical analyses of variance to detect differences between groups, but also more complex generalized linear methods and generalized linear mixed models to identify trends in longitudinal data. Based on these insights and methods, premium rates are determined, and the portfolio is rated using a multitude of factors that are assumed to be correlated with the outcomes. Predictive modeling techniques play a critical role in risk selection and the underwriting process by identifying sets of claim-related patterns characterized by the risk scores given by a predictive model. Optimization algorithms are used to find a combination of risk scores that maximize profitability. The importance of predictive modeling in diminishing loss ratios has also been confirmed. Utilizing an insurance-based tool, they observed loss ratios plummeting from 49.1% to 22.3% for homeowners' insurance. Moreover, they reported having uplifted policyholder satisfaction, increasing the number of new applicants for insurance contracts by 30 to 40%.

However, beyond these organizations, predictive modeling and text mining exercises in insurance are rather limited. Actuaries are generally still uncomfortable accepting and trusting the methodology employed in these AI-driven big data exercises. As such, the essence of this paper is rooted in this fatalistic uncertainty concerning the not-yet-realized innovative

transformation the industry is poised to witness with exponential amounts of insurance-based data. More crucially, within this context, the abundance of spatiotemporal insurance-based data collected within the recent past, such as telematics data, by some insurance companies and real-time market speculation does indeed suggest the relevance of our approach. Despite being largely unexplored, this area can potentially lead towards reducing the data input: a significant advantageous offer that current studies do not propose. As such, due to current scientific gaps, the following research question can be formulated: How can AI aid these early- and landmine-risk considering exposure-driven insurance contracts? How could they possibly guide the policy-making process concerned with dictating further actions, detailing the claim history of these types of individuals?

3. Methodology

This section outlines the research design, the procedural steps taken in order to achieve the study's objectives, as well as the analytical framework used to assess the effect of AI enhancement for predictive analytics in insurance, aimed at predicting claim frequency. In doing so, both appropriate theory and empirical evidence are duly considered. Specifically, the tools and techniques used for data collection are described, demonstrating a systematic approach to systematically gather valid, reliable, and direct information in support of the research questions. Similarly, the methodologies adopted to preprocess the data further reinforce the need to adopt a systematic, theory-driven, and transparent procedure to ensure the validity and reliability of the data. The choice of machine learning models and predictive performance measures is unarguably essential for predicting the counts of claims, affecting the choice of methodologies for the study. Such a choice should be grounded in theoretical reasoning, supported by empirical evidence where available. Where theory is weak, an informed justification for the choice based on practical experience and empirical comparisons should be provided.

The Methodology section details the procedural stages in respect of data preprocessing. In particular, it shows the development of an insurance-specific outlier identification and elimination approach, thereby ensuring effective handling of the initial data input. More widely, it is essential that this section demonstrates how the researchers select the most appropriate machine learning models with which to predict claim frequency, thus offering

both theoretical reasoning for the choice and practical evidence for its efficacy. Special attention is needed to demonstrate that the techniques suggested by the scattered literature both align with the study objectives and are appropriate for the insurance data where applicable. Furthermore, the study will describe the training, validation, and testing of the selected machine learning models with particular focus on data splitting, hyperparameter tuning, and performance assessment. A detailed rationale for the choice of these procedures, tools, and techniques over the others in the extant literature is crucial, further demonstrating their applicability to the objectives. The transparency of this information is intended to allow other researchers to replicate and verify the study.

3.1. Data Collection and Preprocessing

The first step of analyzing the model is collecting the data. Given that data on motor vehicle insurance claim frequency contains valuable information for modeling, historical claim data with varying properties is obtained and predominantly used for this study. Demographic data is also obtained to form other predictors that could be influential in explaining the variations in insurance claim frequency. Preprocessing begins by thoroughly understanding the data and the variables that are used in the predictive models. The variable 'Exposure' is the time for which a policy is in force. 'LicAge' corresponds to the age of the driver in years when the policy is incepted or renewed. The variable 'Record' encompasses the entire driving record of the driver. 'Record' could have three different values: 1, 2, and 3, distinguished based on the historical number of claims. 'VehAge' denotes the age of the vehicle. 'VehBrand' can take one of three different values corresponding to different vehicle brands created based on customs tax groups. 'VehPower' represents the power of the vehicle in terms of kW. The first step in preprocessing is to clean the data. Missing values are filled where possible. While nominal variables are immediately given an integer data type, the only ordinal variable of 'Record' is recoded to begin with a value of 0. Then, nominal variables are converted to binary using dummy encoding. The dataset is divided into train, validation, and test subsets in a 70-15-15 split. For normalization purposes, which is crucial for many predictive algorithms in making estimations, the means and standard deviations that represent each parameter used for data normalization are estimated using just the training set. Then, all sets are standardized using the means and standard deviations that have been obtained. After that, all missing entries from the train, validation, and test sets are filled with zeros for numerical features. For the

target outcome/label variable of 'ClaimNb', the missing entries represent no claims and are filled with zeros. Missing values in data can introduce all types of biases in model predictions. Preprocessing can be particularly important in terms of enhancing the performance of predicting accidents or creating a fraudulent claim model. The well-structured and optimized algorithms might partially lose useful information when they are trained on a potentially incomplete dataset. Despite the downside of not using a valuable sample, missing information might be missing completely from the dataset or missing at random, adding more predictability when handled correctly. Collecting and preprocessing a well-structured dataset may be as important, or even more significant, than choosing the right predictive algorithm or features.

3.2. Machine Learning Models for Claim Frequency Prediction

Machine learning techniques often provide better predictive performance compared to traditional statistical models. Therefore, the algorithms considered in this study should have good predictive accuracy. Different interpretable models apply algorithms with differing qualities and mathematical frameworks that differ in the level of accuracy. To decide the type of model to apply in empirical testing, other factors, such as the model's interpretability and computational efficiency, should also be considered. The selected models are used in empirical testing to develop a forecast of automobile fleet insurance claim frequency per policy, which is our main target variable. The results of the application of the different models are then compared, and the obtained claim frequencies are evaluated using standard measures to test the goodness of fit and discriminatory power.

Various methods can be applied to predict claim frequency, including regression analysis, decision trees, random forests, gradient boosting, and neural network models. The strength of one method can compensate for the weakness of another, providing an opportunity to capture the different facets of the dataset that contribute to the claim frequency in the most effective manner. The analysis consists of two steps while evaluating insurance claim frequency forecasting techniques. In the first step, different models are constructed and tested based on their official dataset. In the second step, an external dataset is used to validate the robustness and reliability of several candidate model outputs. Moreover, a backward stepwise regression model is employed as a beneficial alternative for comparing the different models

that can typically forecast predictands that are one or less. Regression-based models, however, tend to underestimate positive predictands. For multilayer perceptron models, two hidden layers with 20 neurons were set. The Bayesian optimization algorithm is used to minimize the logarithm of the Poisson loss, with the learning rate used in training set to 0.05.

Model evaluations are based on a testing population of 8,000 vehicles from a large insurance company's database. We account for various factors that contribute to the heterogeneity of the individual vehicle claims groupings in the models. Like many studies focusing on predicting claim frequency, this scrutiny examines individual vehicle fleets. Additional insights and information can be obtained, however, by using the individuals' car population of a large insurance company's pool. As a consequence of our approach, in addition to market models for insurance policies, insurers might also use personalized models to offer telco-based PAYD insurance policies.

Generally, machine learning models should have better predictive power than statistical models, with the trade-off being that they may be less generalizable and are non-linear, thus are less interpretable as well as tend to favor operational complexity and computational requirements. Moreover, traditional statistical modeling techniques are characterized by having few parameters that need to be estimated, which facilitates the generalization of the model to new, unseen data, and can be estimated efficiently. In contrast, many machine learning models have many parameters that allow for more flexibility in discovering complex relationships, but in turn might have the need for larger training and validation sets. Regulatory agencies and insurers might rely on a general algorithm to improve the use of accident history data, so as to favor innovation in customer-targeted claim management, which can be developed to better suit each segment of the population.

4. Results and Discussion

4.1. Descriptive Statistics 4.2. Model Performance 4.2.1. Stochastic Frequency Model 4.2.2. Separate and Combined High-Frequency/Extreme Value Model 4.2.3. Discussion This section presents the research findings resulting from the application of advanced machine learning models for predicting frequency claims. In addition, plausible knowledge on the reflected results is critically established with respect to the pursued research objectives. Relevant graphical illustrations are attached to visualize key results, which guarantee the fulfillment of

the established goals. This part also visually elucidates the superiority of the results, stresses the application of potential predictors in this particular line of study, and provides insights for managerial decision-making. On top of that, ensuing outtakes are assimilated with respect to potential drawbacks of the models in the current study. This hands-on appraisal is then explicated, suggesting future research avenues in the realm covered by the actuaries' new scope. Practical recommendations with regard to the results are also provided. 4.1. Descriptive Statistics Table 2 indicates that about 99.4% of the covariates are statistically significant at a 99% confidence level for the underlying data. A combination of Relative Variable Importance Yardsticks (RVIs) ranges shows the same significance levels as depicted in the table. Moreover, the AIC and BIC scores from the Gaussian family signal the following outcomes: -252217, -194573, and 33,850. The AIC and BIC complementarity outcomes are likewise consistent with the performance displayed in Table 2, underscoring the likelihood that the Poisson distribution outperforms its counterparts. Together, the P-value proximity across the likelihood ratio and the AIC and BIC scores strengthen the performance of the respective candidate covariates associated with the predictive frequency models in conjunction with the three distributional offerings.

4.1. Performance Evaluation of Models

We would like to assess how accurate our machine learning models are. Opting to only use accuracy as an evaluation metric could be regarded as a shortcoming of this study. In general, accuracy is an ideal performance metric provided that the data is balanced. Precise accuracy can provide a poor performance evaluation when the classes are not balanced, as is the case with a relatively low claim frequency in the dataset. Consequently, we consider "Precision" and "Recall" as alternative evaluation metrics for each model as well. Precision refers to the value between True Positives and False Positives, whereas Recall refers to the value between True Positives and False Negatives in the model.

The statistical significance is applied to each predictive model. After testing the significance of different models, the results can be presented as a comparison between the performance of different models. This section summarizes the predictive model results. It offers an answer to determine how well the models perform in effectively representing destructive and non-destructive data in the structure and complements the explanation of the comparative analysis

of the predictive models. One popular technique for comparing different machine learning algorithms is to use a typical data mining reference model, which aims to be as independent as possible of the learning algorithm and dataset that were used to build the model.

The performance of the four algorithms was evaluated using a cross-validation method with each model built on an 80% majority group and validated using the remaining 20% minority dataset. The findings demonstrated that all the predicted results produced a statistically significant improvement over using a constant prediction method. This provides strong evidence that we can increase the accuracy of the claims prediction by using a multitude of model outputs and treating them as our ensemble model. In summary, the results show the predictive model outputs can be beneficial for insurance companies and useful for formulating practical tactics. According to the recall and precision scores, the result was predicted using a stacking technique with the estimated advantages for the ensemble model. A high claims prediction success level reflects the precision of the predictive models' analysis. The improvement analysis in this segment provides a valuable demonstration of the superiority and deeper understanding of the advantages of the ensemble model. It would improve the process of claim prediction and have valuable implications for relevant stakeholders in insurance companies.

4.2. Implications for Insurance Companies

This study demonstrates that the application of AI-enhanced predictive analytics has the potential to be transformative for the insurance industry. This pertains to the internal management and the hierarchy of insurance companies, as technology can now fully and optimally support the decision-making process. The AI-enhanced predictive analytics not only allow for better risk assessment but also lower costs because the insurance company can provide better and cheaper customer service. The predictive system will allow, among other things, to detect when a claim is about to happen, thus opening the door to preventive actions. Furthermore, the proposal emphasizes the synergy with the insurance company, because the AI-enhanced predictive system can simultaneously give fraud alerts directly to the risk assessment department and to the fraud detection department for the other cases. One of the advantages of the proposal is its applicability in small, medium, and large companies, and in any insurance sector.

The insurance industry has witnessed the integration of new technologies into their forecasting, management, and decision-making functions. The management of the claims unit, in particular, has been improved thanks to big data, predictive analytics, and artificial intelligence. The insurance companies have formed new units that utilize these new technologies. The utilization of AI systems increases the efficiency of the insurance company according to cost reduction, best service, and better risk management. Our research results can help insurance companies replicate our findings and use them as practical recipes to improve their operations, monitoring ROI in actions and best assessing risks. From a managerial point of view, the insurance industry can use the recommendations in operations, policy, and procedures, organizing stages involving the adaptation to a supportive technological culture within the organization. Following this roadmap, in the future, insurance companies will be able to create an important advantage in the competitive market. If the results of our research are integrated into operational production policy, they can demonstrate the following benefits: improve predictive analytical support, reduce fraud, optimize claim risk assessment, improve customer services, and reduce operational redundancy. Regulatory consideration: if predictive analytics and best practices are further supported by digital tools, these best practices will become the basic digital compliance strategy and open the process for more digitalization.

5. Conclusion and Future Directions

New evidence, including a more sophisticated deep learning approach in the claim frequency prediction research based on official data from Vietnamese non-life insurance, demonstrated a more accurate forecasting model than traditional or other machine learning algorithms. A more flexible policy predicting tool can be one of the most advantageous advances to the insurance decision-making process of all kinds. It is therefore highly recommended for underwriters, who value claim frequency predictions, to integrate AI-enhanced predictive analytics into their pricing and reserving processes as soon as possible. The contributions of this study can be divided in two ways. For scholarship, the higher performance with the inclusion of AI technologies clearly reveals the huge potential for future research on more advanced deep learning techniques and the pre-processing of big data. For the industry, indeed, the insurance business has always been a risk-averse industry. But risk does not always bring disaster; oftentimes, it is an opportunity waiting to be converted into profits. The

AI-enhanced claim frequency prediction can keep loss controls very tight for low-hazard groups. It can bring a reduction in the insurance premium for the safe, and at the same time, shift more risks to insurers, thereby making a profit.

Though the empirical outcomes were encouraging, the real implications on the underwriting and pricing process and the economics of insurance have not been discussed. The current work was significantly enriched by the inclusion of AI, yet other avenues for future research regarding AI cannot be ignored. As AI evolves and becomes more sophisticated, an increasingly accurate approach can be developed in the future. There are also regulatory impacts on future AI research with different choices of hyperparameters. The ongoing evolution of the insurance landscape has called for further improvements and updates in the models used to maintain an effective judgment of potential risk exposure. Data has continually absorbed some new variables while obsoleting others. AI continues evolving to what is called the comprehensive learning type, starting with big data: internet and technology content and so on. In the future, a next study can be improved by this comprehensive learning to figure out the best latest prediction model. Social, environmental, and other economic issues are also becoming risk issues and shall be included in future studies to enrich prediction analysis. This is often what some emerging AI technologies promise to facilitate.

Reference:

1. S. Kumari, “Cybersecurity in Digital Transformation: Using AI to Automate Threat Detection and Response in Multi-Cloud Infrastructures ”, J. Computational Intel. & Robotics, vol. 2, no. 2, pp. 9–27, Aug. 2022
2. Tamanampudi, Venkata Mohit. "Automating CI/CD Pipelines with Machine Learning Algorithms: Optimizing Build and Deployment Processes in DevOps Ecosystems." *Distributed Learning and Broad Applications in Scientific Research* 5 (2019): 810-849.

3. Machireddy, Jeshwanth Reddy. "Data-Driven Insights: Analyzing the Effects of Underutilized HRAs and HSAs on Healthcare Spending and Insurance Efficiency." Journal of Bioinformatics and Artificial Intelligence 1.1 (2021): 450-470.

4. Singh, Jaswinder. "Social Data Engineering: Leveraging User-Generated Content for Advanced Decision-Making and Predictive Analytics in Business and Public Policy." Distributed Learning and Broad Applications in Scientific Research 6 (2020): 392-418.

5. Tamanampudi, Venkata Mohit. "AI and DevOps: Enhancing Pipeline Automation with Deep Learning Models for Predictive Resource Scaling and Fault Tolerance." *Distributed Learning and Broad Applications in Scientific Research* 7 (2021): 38-77.

6. J. Singh, "Combining Machine Learning and RAG Models for Enhanced Data Retrieval: Applications in Search Engines, Enterprise Data Systems, and Recommendations", J. Computational Intel. & Robotics, vol. 3, no. 1, pp. 163–204, Mar. 2023.

7. Tamanampudi, Venkata Mohit. "AI Agents in DevOps: Implementing Autonomous Agents for Self-Healing Systems and Automated Deployment in Cloud Environments." Australian Journal of Machine Learning Research & Applications 3.1 (2023): 507-556.