AI-Enhanced Forecasting Models for Insurance Claims

By Dr. Imad Hout

Associate Professor of Computer Science, American University of Beirut (AUB)

1. Introduction to AI-Enhanced Forecasting Models

Forecasting is widely recognized as a valuable tool. It gives business experts the ability to guide decision-makers by providing estimates of future outcomes using historical and current data. Businesses have always tried to apply the knowledge and expertise of specialist data scientists in forecasting to gain insights from vast amounts of data. These insights help judge the expected future outcomes and use these probabilities in making better business decisions. Business abstracts, like most business transactions, entail a degree of danger, and forecasting aids in mitigating that risk.

The development of more reliable forecasting models is becoming increasingly important. From an insurance perspective, the aim of forecasting is to predict potential losses that businesses are likely to encounter and reserve for such losses by building up provisions beforehand. As technology has improved over the years, access to quality historical data has increased, and advances in computing power have made it viable to use that data in estimating future claims. As a result, predictive analytics is being used widely in the insurance industry to develop models that can forecast future insurance claims. Improved forecasting accuracy has many beneficial effects in the context of insurance. Insurance companies can get a better overview of potential claims, put in place adequate reinsurance treaties which can help optimize their capital allocations. In other words, using forecasting to estimate insurance claims can result in a substantial competitive advantage. The integration of AI in insurance claims forecasting could revolutionize the market.

Given the size of the industry and the importance of making accurate predictions, many applications have been identified where AI could revolutionize insurance claims forecasting. In any business, customer satisfaction is vital. The ability of an insurer to pay claims to the satisfaction of the person who has lost can mean the difference between successful retention of that customer or losing the customer to the competition. Providing a more accurate and therefore more speedy claims service is another potential application of automation

technology and AI in insurance claims. In light of the events in current times, fraud detection is another possible application in the industry. In this paper, we will discuss the current processes and available techniques being used, where necessary.

1.1. Significance of Predictive Analytics in Insurance Industry

The insurance industry in Europe is among the most developed and can be extremely influential across a variety of sectors, such as pensions and savings management systems for individuals and families and capital markets. In today's fast-moving world of digital transformation, predictive analytics represents a game changer in the insurance space, heralding the beginning of an era of 'Personalized Protection.' Insurers have embarked on a quest to supply customers with personalized products and solutions and innovative tools to enhance and optimize customer experiences while increasing operational efficiency. When resolving a claim, predictive models can support risk assessment, underwriting, pricing policy, fraud detection, and claims management while maintaining an insurer's marketplace competitiveness and presence. The data-driven strategy is likely to pave the way for the future of the insurance market in different segments by leveraging real-time predictive analytics, considering customers' needs and preferences and immediate actions upon those insights.

The application of predictive analytics has the potential to harness variously sourced and complex datasets and unleash new business opportunities and create valuable insights via a governing body for internal business evaluative measurements. However, these have limitations due to solid business strategies, data quality, the predictive analytics lifecycle, and data integration difficulties. As a pioneer in the insurance reserving field, numerous success stories have been published by insurers in recent years. Predictive analytics has transformed insurance by not only predicting claims but also enhancing operational efficiencies and reshaping and encouraging higher customer retention and satisfaction levels. Predictive analytics can drive innovations and digital transformation, leading to turning risk into business management and growth for insurers.

2. Fundamentals of Machine Learning in Insurance

Machine learning is a subset of artificial intelligence and represents a wide range of techniques that can be used to leverage information hidden in the vast amounts of data available

nowadays. Solving insurance-related problems on the basis of automated learning of associations between inputs and desired outputs, predictive modeling, and other methods have been attracting growing attention recently. The machine learning domain brings many new concepts relating to algorithms, which represent the basic processing units that perform learning. These algorithms can build various models that represent the internal representational state of the system during learning, such as neural networks or probabilistic graphical models, and they also differ fundamentally from the statistical models, which are typically used in insurance analytics. Furthermore, these concepts are complex, and there is no standard method on how to use them in a given problem.

The success of predictive modeling, massive amounts of data, and the increase in computing power is mainly dependent on large training data sets of good quality and with a broad known variability. Traditionally, both insurance and other financial institutions have been collecting large amounts of information, such as policy data for residential or commercial buildings, personal or motor insurance. In the majority of cases, this valid data is stored for long periods of time and is available within company databases and could be used to train machine learning models. Insurance areas in which machine learning is used for forecasting mainly concern claim frequency, frequency of customer contact, lapse rates, and claims severity; yet, in property and casualty insurance, a growing number of applications focus on assessing risk or fraud detection. The type of machine learning technique strongly depends on the distribution of data, and the success in using it for optimal prediction is strongly dependent on the method chosen, to which there is no one easy answer. Furthermore, a trend to automate more complex analysis, the so-called advanced analytics, is increasing, and with it also the research into new machine learning methods, particularly in order to adapt to market changes.

2.1. Supervised Learning Techniques

One of the most commonly used techniques in predictive models for insurance claims is supervised learning. Supervised modeling, as opposed to unsupervised methods in clustering and segmentation, refers to the use of labeled training observations. This bifurcates supervised learning into two categories: regression models (for continuous outcomes) and categorical classification models (for categorical outcomes). In insurance claims settings, for

example, the use of labeled training data provides a 'yes' or 'no' to the algorithms. Then, claims managers assign outcomes to data points for which the results are known and act as the label. The labeled training set is used to associate features with inputs and inputs with outputs (hence, supervised). Model outcomes are then used, based on the algorithms and data used, to predict the unknown outcomes.

Supervised modeling algorithms have the advantage of constructing models that are accurate and have a definite output. They are more closely associated with real-world problems, with many examples in both general theoretical literature and the claims management literature demonstrating their applicability. In applications of insurance claim forecasting and other actuarial areas, the key considered elements are tangible risk reduction and profitability of the different enterprise activities. These models face challenges such as overfitting, where a model learns the details of the training samples and lacks generalization on new information. Also, these models require a well-curated dataset that accurately represents the real world. Simple real-world applications in actuarial science, insurance claims, and related areas illustrate the benefits of using labeled data in training the various machine learning models.

One of the world's leading reinsurers found in a practical case study that with the application of predictive modeling for claim frequencies and severities, two years in advance, companies are able to quantify the risk and price more accurately. The result seen was greater and more optimal capital backing, overall contributing to economic value added. Another research firm effectively employs supervised learning techniques across various industrial insurance lines of business to forecast reserve requirements and claim savings for its claims handling operations. The findings have shown improved allocation of reserves and the financial stability of the company. With the examples above, it is evident in the market that these techniques are relevant and useful given the availability of data and specialized expertise to curate a dataset.

3. Data Preprocessing and Feature Engineering

Welcome to the third chapter of the guide! In the following sections, we will extensively discuss some very important parts of the entire process of developing forecasting models, enabling future claims prediction. The first important step and aspect of those analyses, as well as their main input, is data preprocessing. Once we have raw, original information, we

need to prepare it for further analysis. Data preprocessing is a step-by-step process of altering, cleaning, and aggregating that data. Without careful data preprocessing, predictive modeling is likely to deliver nonsense and possibly misleading results. Then, we go into feature engineering, a set of procedures for choosing the variables to include in our model. These are often core to various types of modeling, related to a concept called variable selection, used in many analytical fields, including time-series forecasting and machine learning.

Our focus here is to transform raw data into a version that helps the model make simple and accurate predictions. To that end, we standardize our variables, which makes their averages equal to zero, their standard deviations equal to one, and scales their probability distributions to meet the Gaussian distribution. We also categorize, allowing a model to compute target variable interactions in much smaller geographical spaces. Data preprocessing can be particularly tricky since many real-world datasets are quite large, and momentum builds to relax or even ignore protocols to accommodate size. Automated methods can be quite useful, but such procedures require substantial attention to block overfitting and too eagerly dismissing variables without appropriate care. Measurements of feature relevance can be misleading, especially requiring transformations that aren't made in advance. To illustrate more: A model of loss, in some cases, may seem directly responsive to a feature, like the following.

3.1. Handling Missing Data

Introducing any form of artificial intelligence into the domain of real-world claims forecasting is a compelling proposition. Nonetheless, the method will only be premium if it possesses principled methods for dealing with confounding practical limitations. They exist. For instance, missing data is one of the serious issues in the development of a robust forecasting model for insurance claims; that is, using machine learning or statistical methods to predict future claims volumes or costs based on historical claims data. There are broadly three types of 'missingness' in the data: missing completely at random, missing at random, and not missing at random. The presence of missing data in the claims dataset could affect the forecast accuracy. In the most basic case, where data missing is not random, this can further lead to biased forecasts or a model analysis that is more problematic.

There are several ways to address issues concerning the handling of missing data. The simplest approach to treat missing data is to ignore the entire case altogether if there is only ever one data point missing. The strategy is called complete case analysis. Another approach to missing data is data imputation, in which missing values are supplanted with estimated ones to allow for mathematical computation involved in the analysis. In addition, three common data imputation methods include mean imputation, single imputation, and multiple imputation. While imputation can be a successful mechanism to alleviate the negative impacts of missing data, there are debates in the literature on the robustness of imputation methods. It should also be highlighted that multiple imputations require the assumption of data missing completely at random or data missing at random conditions and mainly apply to larger datasets. To assess how much missing data can be said to impact model reliability, three questions from the guidelines for dealing with missing data are used: whether missing item carries a major decision impact, whether the complete-case analysis is feasible, and techniques to quantify the impact of missing data will be assessed.

Referring to similar challenges in practice, there was a case of failing to handle missing data in claims development forecasting, which led to suboptimal claims provisions. In some cases, misspecification of the regression equation caused substantial underestimation of historical claim development factors with systematic under- or overestimation of overall claims amounts.

4. Model Selection and Evaluation

Model selection is a key step in building any forecasting system and considers several criteria. Besides good forecasting quality, measured with standard scoring rules, further significant points include that the forecast model should match business objectives, use appropriate data sources, contain as few parameters as possible, remain comprehensive and easy to interpret, and maintain computational efficiency. One of the most crucial aspects of model selection is thus the choice of the model family and the model structure. Often, the above-mentioned aspects require trade-offs, and it is important to take into account the goals of the forecast and its intended use. In many application domains, more accurate and complex modeling techniques have been developed and used over the last decades, such as advanced neural network architectures, state-of-the-art modeling of time and space with RNNs, fully

automated machine learning-based architecture search, unsupervised learning techniques, reinforcement learning, or deep clustering methods, among others. A key point is the iterative nature of the process. The recurring character of desirable characteristics of forecasting methods is emphasized.

Selecting an appropriate model also requires a sufficiency of evaluation procedures. Cross-validation is the best tool for evaluating forecast model quality, where different splitting schemes and cross-validation techniques need to be examined. For time series data analysis, predictability can be evaluated using the Box-Jenkins approach. If it is a forecasting task in which different models should be considered, A/B testing can be useful. Relevant performance indicators include the mean absolute deviation (MAD), mean squared error (MSE), mean absolute percentage error (MAPE), or the Theil-U statistic in supply chain management. Since health insurance claims forecasting differs from other types of forecasting, certain other indicators are relevant. Regulatory institutions also play an important role since these institutions have complex duties in terms of solvency, risk management, and supervision. The features of forecasting claims could be different; thus, a technique used in one country, one insurance company, or one portfolio could not be applicable directly to other insurance companies. However, the main performance indicator in health insurance is data science. Health claims cost forecasting has become a part of this data science work. The work continues iteratively, aiming to improve the initial models. Model A could be used in the first nth month for forecasting, whereas model B could be used for comparison purposes. Model B, if found to be superior, will replace model A in the nth+1 month, and the process goes on.

4.1. Performance Metrics for Insurance Claim Models

In order to evaluate the insurance claim forecasting models, we need a specific predictive performance measure. The accuracy, precision, recall, F1 score, area under the receiver operating characteristic curve, and the area under the precision-recall curve are often used to evaluate the performance of binary classification models. Accuracy is defined as $(\sum (TP + TN))/(\sum (TP + TN + FP + FN))$, where TP is the number of true positives, TN the number of true negatives, and FP and FN the numbers of false positives and false negatives, respectively. Although accuracy is easily interpretable, it may not be suitable for all problems, particularly when the classes have an imbalanced distribution. Precision and recall can be calculated as

TP/(TP + FP) and TP/(TP + FN), respectively, whereas the F1 score is defined as the harmonic mean of precision and recall.

Nevertheless, it is worth noting that different objectives may require the use of different performance measures, which can lead to inconsistent results regarding model comparisons. It is also important to emphasize that most of the previously mentioned classification performance measures are more suitable for addressing the operational aspects of the insurance business than the statistical characteristics of individual loss distributions, which may limit their use. Reported changes in forecasting accuracy when precision and recall metrics were used to evaluate the model performance prompted useful model revision. However, it was pointed out that one cannot infer from these results whether suitable classifications were maximized in a real-world context, emphasizing the considerable practical challenges in interpreting and applying advanced metrics such as precision and recall. All of their results suggest that suitable metrics are more helpful in model performance enhancement and benchmarking within the same area or function rather than in an objective and standardized comparison among researchers.

5. Applications of AI-Enhanced Forecasting Models in Insurance

An important application for time series forecasting models in insurance is the forecasting of claims properties in the near future. To quantify and budget the risks related to their products, insurers and reinsurers employ millions of time series models to forecast future claims payments or future premiums. Claims forecasting uncovers hidden information in the data and enhances the efficiency of risk management. The properties of forecasts, such as coverage, accuracy, and underlying uncertainty, are used by insurers for different pricing and hedging strategies. In this paper, we list several specific applications and case studies related to forecasting in the insurance industry. First, the efficiencies of fraud detection can be significantly improved with the combined strength of forecasting and AI. AI can be used to identify and prioritize claims that need investigation. Second, other studies have used forecasting for claims reserving purposes. By leveraging distributional forecasts, one might be able to use a snapshot approach based on estimates. Third, forecasting has been used to imply monetary targets for performance management. The need for accuracy and planning is, for instance, particularly important in the area of case management, wherein initial reserves

and final payouts need to be justified to both internal and external stakeholders. AI and forecast-based claim reserving can also be applied to develop more accurate loss ratio predictions. This can be used by an insurer to improve the underwriting quality of their business. In summary, forecast-based insurance models offer a tool for strategic underwriting decisions by an insurer and underpin a transformation strategy for the insurer that moves from handling general schemes to a more tailored client offering, wherein value-based discussions can be held at the level of corporates. This extension of our paper also offers ethical and regulatory considerations surrounding the use of these approaches.

5.1. Fraud Detection and Prevention

Fraud detection is a critical part of client engagement and is ranked as the highest operational priority by insurers. Insurance firms widely apply predictive modeling. Many algorithms, such as decision trees, support vector machines, neural networks, and random forests, have been used. Other technologies that are being experimented with in insurance include text mining, which screens correspondence with specific keywords, and the use of AI in conjunction with data analytics for amenability prediction and automation, as well as for spotting severe events. In the insurance field, artificial intelligence technologies such as regular expressions, statistical examination algorithms, and algorithms with linguistic interfaces are utilized for routine claims. Natural language interfaces are now utilized for property and casualty insurance claims reporting so that responses can be given 24 hours a day.

A properly developed artificial intelligence technique will quickly review claims information, search through the client's account data, and render a decision. To search for unaccountable activity, trade irregularity, or money laundering warning scenarios, insurance firms use claims data, operational data (including the types of treatments offered), and publicly available information, and they normally go back around ten years. The sophisticated software locates using a method called pattern recognition that has been highly modified and is reassessed daily from earlier detections and from millions of data variables. Insurance firms utilize predictive models in enforced regulations for business partner contracts in order to avert breaches of terms and conditions. An experience model is usually used to compare data from new pooled claims in real time with insurance firms' historical data. Data is provided

from various sources, including actual claim details, claims activity warnings, as well as relationships with various particular points and patterns. The effect of a new indicator on the model changes every day and reports results in real time. Mixing real-time analysis of data with indicators of fraud is a significant focus of additional insurance investments.

There are numerous examples of case studies in the insurance industry that explain success in reducing fraud. In this area, there are many challenges to making more progress. Insurance underwriters and regulators are concerned with false positives. An issue that often arises is that numerous people think their claims are not valid. The majority of insurance firms are now concentrating on the development of machine learning strategies. Following the analysis, it is important to recognize alterations in trends and strategies. It is frequently referred to as model control to ensure that the fraud analysis is carried out in a thorough and regular manner. Machine learning strategies are improved continuously, but their development is more dependent on the abilities of individuals than on improvements in technology. Policymakers must pay attention to this trend.

6. Future Direction

AI has the potential to reshape the forecasting landscape in the insurance industry. The future direction for forecasting in insurance will likely be characterized by an increase in AI-enhanced models as insurance companies are now able to gather and store an increasing amount of data. There are a number of identified upcoming trends in the field of machine learning for forecasting. The first is an increase in the degree of automation, from simply fitting a time series model to automating the whole data processes. There might also be further integration of complex algorithms in automated ML platforms that journey through the forecasting process. Other trends include more seamless integration of feature learning into forecasting models and increasing consumer-grade anomaly detection. The final but perhaps most relevant trend is the ability of machine learning models to continually learn. It is now possible with new machine learning models to retrain and continue learning in the face of changing data, and so the continued relevance of these models seems evident.

The challenge for the future development of AI-enhanced insurance forecasting is not application but rather how to collaborate together to incrementally create value for the industry. It is worth getting excited about the applications, but we should try to develop them

responsibly. In order to manage the rapid evolution of the field, there are regulations and commercial considerations that AI applications should follow. In particular, there are ethical considerations around data security, which are extremely important when faced with vast amounts of data. Third-party data breaches are not uncommon, and they carry significant regulatory penalties. Furthermore, it is also important to ensure that AI is used responsibly and that it is free from biases that temper users' expectations. Discrimination of any kind is unethical and can lead to expensive lawsuits. Furthermore, we have a fiduciary responsibility to our clients. Lastly, the trend of integrating AI with blockchain and the Internet of Things has not had a massive impact on forecasting yet but has important potential, considering the robustness and reliability of data collected. This may influence who gets into the future of forecasting in insurance, since companies at the frontier of IoT could have a significantly competitive edge.

The use of this data and the trends listed above could drastically change the future of forecasting for insurance, allowing for a more granular, continuously learning, and adaptive approach. However, when developing forecasting models, it is important to keep in mind changing legislation, fiduciary duties, and ethical implications. This landscape is gradually evolving, and therefore the future direction is not certain. We believe, however, that forecasting in insurance has a bright future, but it is important that we continue to work together, share our experiences, and develop our methods in a transparent and comprehensible manner. This may enable the formation of a consensus on what the key factors in forecasting are and standardize best practices or benchmarks. This consensus could come in the way of leading the development in line with the three principles of fairness, privacy, and transparency, and in collaboration to minimize the potential limitations.

7. Conclusion

This thematic essay is directed towards quantitative professionals working either in the insurance industry or on insurance-related problems. We review works on AI-enhanced forecasting models for insurance claims. Using more advanced predictive analytics to allow better and more detailed, risk-informed decision-making in (re-)insurance quotes can help balance the effects of new technologies such as telematics for cars and wearables for health insurance. In this essay we have tried to show the potential of transformative changes in

current methods of insurance pricing and risk evaluation in the era of predictive analytics, Machine Learning, and AI technologies. Traditional insurance pricing relies on the use of 'experience forecasting' — i.e., the statistical projection of annual aggregated claim cost distributions. We reviewed a growing body of literature on advanced predictive modeling for 'portmanteau (or portfolio) forecasting'. Instead of dealing with aggregated claim outcomes, forecasting is performed on an individual 'unit' of risk, e.g. a health or a car insurance policyholder. We also assessed the relative performance between model types, sizes, and techniques on diverse insurance data sets. Using more granular data combined with predictive analytics will enable placing insurance products where they are most needed both because they will satisfy potential buyers of such products and because they will bring in enough premiums to stabilize the insurance division or the company. In this thematic essay, we have reviewed and offered a critique of the state of research in predictive analytics, regressing machine learning, AI, DLT, and blockchain techniques for insurance forecasting.

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.