Deep Science Publishing
https://doi.org/10.70593/978-93-49307-44-5_6

Chapter 6: Utilizing machine learning and predictive analytics to enhance risk assessment and claims processing in insurance

6.1. Introduction

Insurance companies gather a growing variety of data for use in the insurance process. Many emerging data sources, such as the ones collected from the Internet of Things (IoT) and social media, may complement traditional data to provide better insights to predict future losses in an insurance contract. However, the variety of data sources implies a diversity of data formats. Some of these data sources come in a non-vectorial format, which makes them difficult to incorporate in the existing pre-established ratemaking models. This paper will present some of these emerging data sources and a unified framework for actuaries to incorporate them in existing ratemaking models.

The goal of this framework is to bring these novel data to actuaries through a new representation of the variables in the modelling process. The proposed approach stems from representation learning, which is a toy machine learning framework aiming to create representations of raw data. A useful representation will transform the original data into a dense vector space where the ultimate predictive task is simpler to model. To achieve this goal, two steps are needed: (1) learning representations from the emerging data with deep neural networks, and (2) adding the representation vectors to the data of the existing ratemaking models, which are then retrained with these new variables.

Historically, the field of actuarial science has a well-established methodology for the prediction of future losses in property and casualty (P&C) insurance contracts. Due to the computational expense of the models, until recently, these models were often structured based on relatively few variables, relying on the expert knowledge of actuaries to decide the important terms and how to represent them in a model. In addition, most modelling practices in actuarial science consider data in a tabular format.Therefore, to apply better pricing risk and claim processing, vehicle insurance companies need to understand how risk assessment features impact insurance claims, choosing the proper techniques. Insurance claims size prediction is one of the best ways of understanding how various risk assessment features affect vehicle insurance claims. A key challenge for the insurance industry is for each customer to charge an appropriate price for the risk that customer represents. Risk varies widely, and a deep understanding of the underlying processes generating this risk is crucial for predicting the likelihood and cost of insurance claims. Focusing on a specific nonlife market, namely vehicle insurance.

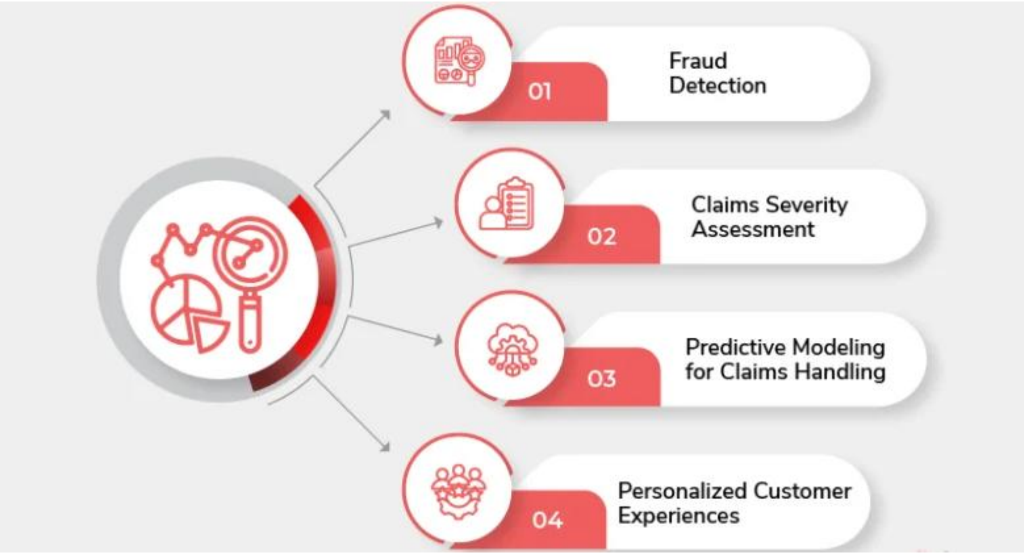
01
Fraud Detection
02
Claims Severity Assessment
03
Predictive Modeling for Claims Handling
04
Personalized Customer Experiences

Fig 6.1: Predictive Analytics in Claims Processing

6.1.1. Background and Significance

Insurance is a risk management tool that controls the possible financial consequences of uncertain events. It is a contract in which one party transfers the risk of loss to the other in exchange for a consideration called a premium. An insurance policy is an agreement between an insurer and an insured or policyholder. A claim is a request made by the policyholder for insurance benefits. For example, a claimant who has a car accident submits relevant documents and requests compensation for the repair of the accident-damaged vehicle; this is called motor insurance claims. The core premise of insurance involves pooling resources, sharing risks, and providing compensation to each other in the event of a loss.

In vehicle insurance, the insurer transfers the risk of loss of the vehicle to the insured. Insurance companies charge a premium based on different risk assessment features; based on the claims reporting to the insurance company, the risk premium is increased, and for better risk transfer, the benefits to the policyholder remain in a better situation. An insurance claim is a request for payment that a policyholder submits to an insurer for covered losses. The insured or a claimant can claim for covering losses owned due to accident occurrence in each insurance company. Insurance companies charge for the probability of the total claim amount and management of the loss as premium payment. Understanding the primary amount of each possible insurance claim enables the insurance company to set a fair price for its products. Therefore, the claim number and size prediction are essential for the better preparation of insurance companies to meet the client's compensation requests.

6.2. Background of Insurance Industry

The insurance industry is changing due to technology, data, climate change, and public opinion. Organizations use machine learning and advanced analytics for better performance and growth, recognizing the value of collaboration among relevant domains. The insurance industry is under pressure to keep pricing competitive, while achieving a long-term sustainable operating profit and maintaining financial resilience and capital efficiency. The complexity of products, and ever-increasing advancement in computer technology, has posed a daunting task of handling risk management, product pricing, reserving and claims assessment. Underwriting as the first touch point of client interaction in the insurance value chain serves a crucial role in identifying potentially risky clients, evaluating correct premiums, and developing long-term resilience of the organization.

Manual risk assessment is time-consuming, requiring human judgement and leading to potential bias. Firstly, to keep pace with the change in risk theory, analyzing risk data and properly designing rules that are relevant to risk theories is inherently very difficult. Over the years, a set of risk assessment logic and rules are designed mainly by actuaries using manual rules.

6.2.1. Research design

This section presents the research design for the study using a flow chart. The different components of the research are listed along with the techniques to be used in each step. The research design contains the research methods, data source, feature selection, evaluation metrics, and investigation methods.

The study used various machine learning techniques to build predictive models, inputting the features to different algorithms and setting the parameters as required. The models are executed on four different tree-based models. The model's performance is measured with the metrics, including accuracy, precision, recall & F1 score for classification models and RMSE, MAE, and R-squared for regression models. Data cleaning is developed, and feature selections. Hyperparameters tuning is done, and the modeling is developed. The best model will be selected. The best model is developed, and the implementation is done to automate it by creating a web app. For the analysis, some of the most renowned and widely used machine learning methodologies are discussed. So, statements can be derived that the conversation and coding are quite open-ended. Usage of a variety of no. of clustering analysis methods, and applying each of them.

The process of enhancing or increasing the period of technological objects, technical processes, or facilities which have not achieved their designated lifetime has a wide range of definitions. Claims size is the amount of money that will ultimately be paid out by an insurer once a loss has occurred at their insured assets. It is possible that neither the claim nor the claim size will come to the attention of an insurer until a long time after the coverage inception date. Thus, the claim size prediction analysis is today essentially irrelevant for risk selection, in vivid contrast to vehicle insurance pricing.

6.3. Understanding Risk Assessment

In regulatory contexts, insurance companies are required to monitor and control the risks linked to underwriting and asset management activities (Dhawan, 2024; Cyriac et al., 2025; Klein, 2025). Managing the premium sufficiency risk regarding future claims is required for underwriting risks, while assessing the adequacy of current reserves to pay outstanding claims is needed for reserving risks. In both events, estimates of the future claims payment (cash-flow) are modeled to quantify a confidence level on whether the expected cash flow will be sufficient to pay for its liabilities. Countries usually demand that estimates of the expected future claims, along with a measure of its uncertainty, be provided, leading to the need for stochastic models for this task. The estimate of the future claims the insurer will have to pay (reserves) is calculated by summing the predictions of a stochastic model that is able to accurately estimate the future claims. Estimating the uncertainty regarding these reserves is usually performed via residual processes of traditional stochastic models, such as autocorrelation and/or residual distributions adjusted to a pre-defined pseudo-stochastic process.

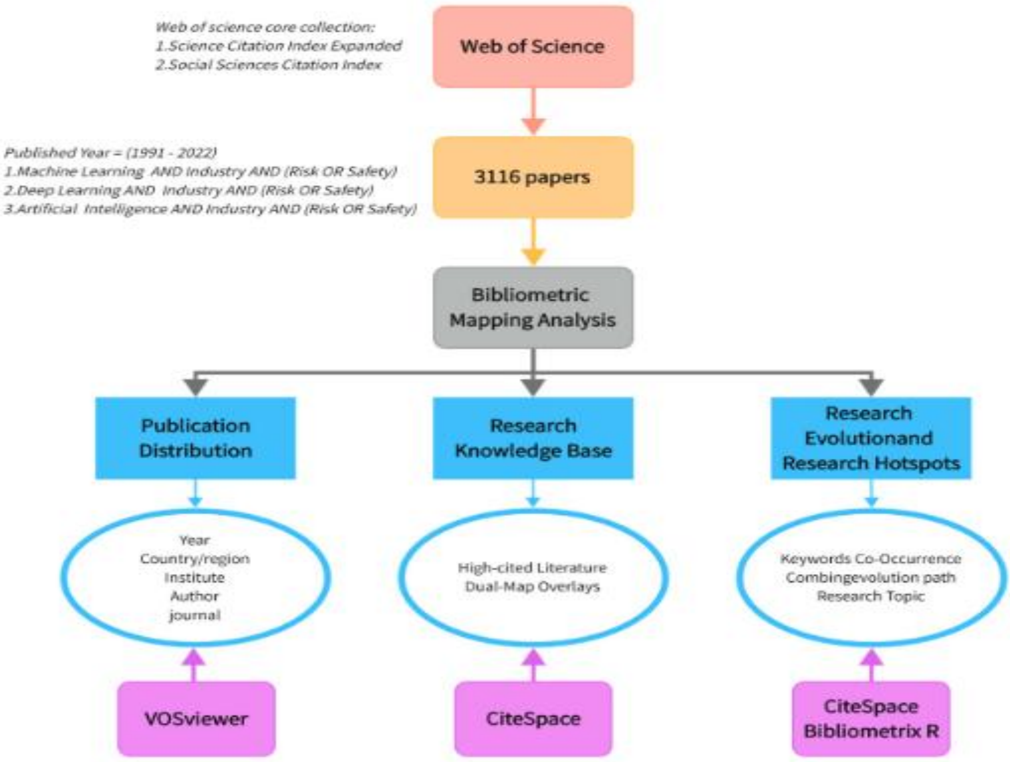
Web of science core collection:
1.Science Citation Index Expanded
2.Social Sciences Citation Index
Web of Science
Published Year = (1991 - 2022)
1.Machine Learning AND Industry AND (Risk OR Safety)
2.Deep Learning AND Industry AND (Risk OR Safety)
3.Artificial Intelligence AND Industry AND (Risk OR Safety)
3116 papers
Bibliometric Mapping Analysis
Publication Distribution
Research Knowledge Base
Research Evolutionand Research Hotspots
Year
Country/region
Institute
Author
journal
High-cited Literature
Dual-Map Overlays
Keywords Co-Occurrence
Combingevolution path
Research Topic
VOSviewer
CiteSpace
CiteSpace
Bibliometrix R

Fig 6.2: Machine Learning in Industrial Risk Assessment

In particular, claims reserves are usually calculated by the iterative run-off triangle approach or a chained run-off method. However, traditional stochastic models usually applied to calibrate claims reserves have an important drawback: their residual processes are not the most appropriate for their integration with predictive non-probabilistic machine learning models. Development year typed cumulative claims reserves are considered to cascade stochastic modes and predictions in a hierarchy since they are the only data available to the actuary. More sophisticated stochastic models such as Generalized Lindstad, Mack and its variations, Hierarchical Poisson.

6.3.1. Definition and Importance

Predictive analytics is a crucial tool for businesses because it allows them to assess risk, evaluate customers, and enhance marketing efforts in order to increase profitability. It relies on statistical and machine learning techniques to analyze data so that predictions can be made regarding the possibility of losing money on a customer, with fraud being a constant concern in these efforts. Additionally, predictive analytics lends itself well to behavioral segmentation strategies in conjunction with other data sources such as next-best-offer predictions or propensity to respond predictions. However, implementing these strategies can be difficult in practice. Addressing these limitations requires new techniques and innovations, especially in terms of address matching and geography data disambiguation in processing and enrichment.

Predictive analytics presents an opportunity for insurance companies to compete on price, something they currently cannot do due to regulatory restrictions. Calculating the optimal price for insurance depends on good predictions of the expected future claims cost and the insurer's administrative costs, as well as a fair margin. Competition between insurance providers can leave many firms fighting over smaller slices of the profit pie, leaving the winner with only small margins. Due to this competition, claims that may otherwise have gone uncovered have an increased chance of being paid out, further threatening the profit margins of these firms in a high-risk scenario.

6.3.2. Traditional Risk Assessment Methods

The key data used is the history of the insured contracts (Lloyds Banking Group, 2024; Rowe, 2024). This includes events affecting the insured contracts, as well as prior claim amounts and accident periods. Ideally, the starting point of the analysis is to create a homogeneous dataset of these insured contracts. Then, within the design of a method to perform the analysis, actuaries need to fix several important points. These include the number of time periods per step on which the prediction is performed, whether to consider only the raw insured contract features or to transform some features, the way of selecting an exhaustive set of features to maximize the alignment of the regression output with the predictive task, and whether to create a machine learning method to analyze the data and learn the correlations and complex relationships within a set of features.

The main prediction target is the expected value of the primary payout of the insured contracts. This provides information on the natural logarithm of the predicted claims amount. This variable is transformed into the modeling gum using the uniform approximation in the box [0, 1], which can enhance the prediction performance of the model. Note that the gradient boosting models are hence suitable since they rely on thresholds on the distribution of the data. These models provide information across the full range of the prediction cost. The aim is to provide the design and an associated pre-implementation study of predictive methods of machine learning to assist the actuary in the risk assessment step of property and casualty insurance contracts under variable long- and short-tailed errors.

6.4. Overview of Claims Processing

Claims occur when the policyholder creates a formal request to an insurer so that it can provide the coverage or compensation for accidents. Insurance claims are payments for compensations made by claims assistants within the specified period for covered loss, which could be property claims, hail claims, accident claims, liability claims, earthquake claims, health claims, theft claims, fire claims, flood claims, etc. If the claim is accepted, the insurance company compensates the risk taken where most of its earnings derive. For example, in a vehicle insurance company, claims can occur when the property, its passengers, or third parties get damaged in a vehicle accident. The injuries can be covered in different categories: indemnity payment, expense coverage, and loss of income coverage. These claims cannot be a loss, but the insurance coverage could minimize the costs involved with the risk related to the property and liability of the insured vehicle. Insurance claims are payouts done by insurers to compensate their clients' losses due to accidents or other covered events. In this case, the insurer's earnings are claimed only if they occur in favor of the insured consumers. Loss occurred in this transaction comes through payment done to the clients. Thus, it needs to predict the loss or possible claim count for any vehicle insurance company.

Insurance companies have a huge chance of using ML and PL for claim severity prediction due to the availability of innovative data sources and learning algorithms. ML is being widely used in the insurance industry, specifically in claim severity prediction. Claim count prediction in non-life insurance is a fundamental but very challenging task. Claim prediction approaches in non-life insurance range from traditional econometric models to recent ML models. ML techniques are finding a wide variety of applications in non-life insurance. Since its introduction in vehicle insurance, it plays a significant role in providing smooth premium adjustment in the company. Claim prediction in the vehicle insurance sector is a roof task for the estimation of premium. Claims happen where the insurer compensates the loss of insured objects such as vehicles, health, life and property etc to the policyholders. Premium is charged for a risk taken in favor of the client. Premium for coverage is either de-principles based on risk factors or pooled among the insureds.

6.4.1. Claims Lifecycle

The use of descriptive statistics and preliminary analysis to gain insights about the claim size was then examined. Then the data was examined further by looking at the distribution of the target claim size variable and some of the other predictor variables. The first step in modeling the data was to check for the best predictors of the target claim size using Tree-Based regression techniques. Using Random Forest as a base estimator, Bagging was implemented using Neighborhood Components Analysis which is an algorithm that while originally meant for classification, can also be used for regression. Using the same principle, Gradient Boosting was then applied as another ensemble regression technique. Finally, all of the results were evaluated and compared using plots from the mean absolute percentage error metric. The most effective model on the dataset was then taken to be the final model and its predicted results were plotted against the original test data to visualize the results. Insurance is a financial transaction between two parties wherein the party taking the insurance (the policyholder) pays a certain sum of money to the other party (the insurer or insurance company). In return, the insurance company promises to cover the financial losses that would incur to the policyholder in case of damage to either the policyholder's property or loss due to third-party liability. For companies to remain profitable, they need to charge each customer an appropriate price for the risk they represent. Claims might be pure risk meaning that there is no possible gain or profit, but losses are to be covered. Also, claims might come due to very expensive accidents. Hence, predicting customer claims is a core part of a company's worth. Vehicles represent a significant share of the economy, and their insurance contributes to the reduction of risk. As a result, it is no wonder that many insurance companies use the vehicle insurance case as study objects for better predicting and charging regarding risk.

6.4.2. Challenges in Current Practices

The below challenges and recommendations concern the utilization of machine learning models for risk assessment and claims processing, which was the focus of the project worked on during the internship. The application of machine learning (ML) algorithms can have a plethora of applications in various sectors. In the insurance sector the algorithms have become ever more efficient in performing tasks from risk assessment, claims fraud detection, and claims processing amongst a variety of others. The unprecedented increase in data due to technological advancement as well as the increasing variety in data prompts for further investigation. Especially the accessibility of data sources outside the traditional data sources such as social media and even the dark web offers new insight into assessing the risk for insurance contracts. However, traditional ratemaking models are usually not designed to accommodate such heterogeneous data sources.

Nonetheless, various innovations have already been made in integrating machine learning in insurance processes. A systematic overview of which predictive analytics and machine learning methods are commonly used in which insurance processes are given. Next to this, the most frequently used data sources in each process are disclosed as well as the most common way of data pre-processing. This makes the current practices transparent to the actors that are innovating in this field. A number of challenges that current practices face are revealed and additional recommendations to tackle these challenges are provided. Lastly, further directions regarding extensions to this project are discussed. Companies are more and more incorporating machine learning models in their insurance processes. The main aim of these projects is to provide higher predictive power, with respect to more traditional models.

6.5. Machine Learning in Insurance

The insurance industry deals with uncertainties by offering coverage for uncertain events to customers and collecting insurance premiums to pre-finance potential claims. As a tradeoff for the risk it takes with customers, an insurance company applies a premium to customers. The initial prices for a premium are based on risk factors. A key challenge for the insurance industry is to charge each customer an appropriate price for the risk they represent. Risk varies widely from customer to customer, and a deep understanding of different risk factors helps to predict the likelihood and cost of insurance claims. Insurance companies have to predict how many claims are going to occur, and the severity of these claims to set a fair price for their insurance products. Claim prediction for the vehicle insurance sector is the cornerstone of premium estimates. Given numerous risk factors an insurance company holds through the insurance policy data, it is imperative to analyze which of these risk factors are important to accurately predict claims. Several studies have been done to personalize the premium estimate. By applying different heuristic techniques to public datasets of health insurance, it has been demonstrated that analysis of risk factors for insurance prediction has huge possible benefits. In order to set a fair premium for a risk, a key requirement is knowing how much of the risk is going to be claimed. The accuracy of premium estimates on claim prediction can be significantly improved by analyzing information from telematics. Currently, many insurance companies are transitioning to ML techniques to predict claims size. Flexible ML techniques have to be investigated to make accurate predictions for claims size by analyzing a large vehicle dataset. It has been designed to apply the tree-based ML methods to the dataset, and these are bagging, random forest, and gradient boosting.

Claims prediction is the cornerstone of setting a fair price for insurance products. Prepared historical claims can give insights into selling new insurance policies. The majority of insurance companies either partially or fully in-house their premium aspirants, but they commonly estimate the premium ranges using regression models based on risk factors. Based on limited initial information, only the mileage-based risk factor is considered to build a rule-based classification model. More advanced modelling techniques for risk classification are proposed, and they address the telematics-based classification with ML and rule-based classification methods, which are comparative.

However, these methods focus on premium classification. With scarce public datasets of insurance claims and privacy concerns, they can only train a black-box prediction model and make a limited number of queries to predict claim amounts since insurance companies require comprehensive knowledge about risk factors to explore new business opportunities. Systematic investigation of feature learning methods to derive importable features for the black-box model's prediction curves is required.

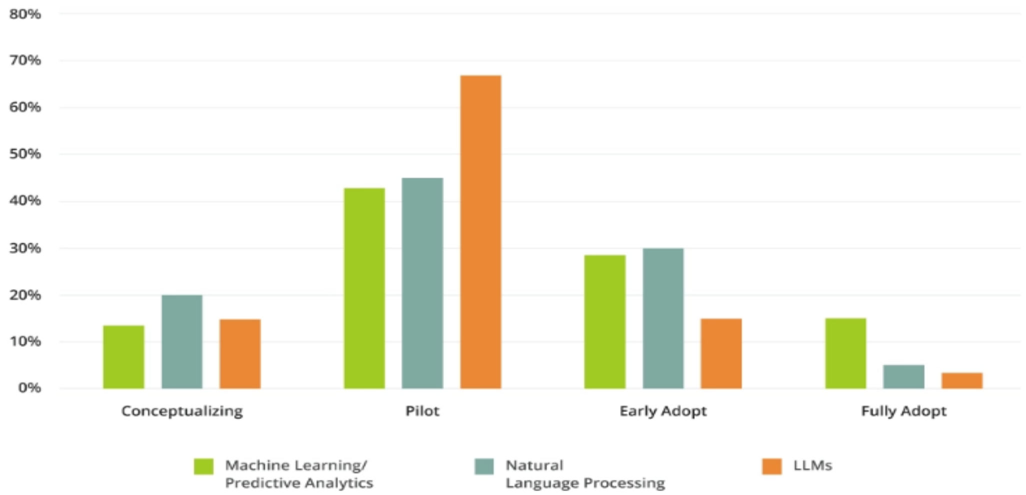

Fig : Risk Assessment for Insurance

6.5.1. Introduction to Machine Learning

In recent years, there has been a proliferation of data and information generated from many sources, and the incorporation of Data Science, Data Analytics, Predictive Analytics, Machine Learning, Artificial Intelligence in businesses has skyrocketed. This progress is attributable to a variety of circumstances, including the rapid advancement of computing resources and the growing understanding of the importance of data to business and society. However, despite all the opportunities that accompany this, many firms lack the capacity to comprehend the enormous volume of data they collect. Although many firms collect data, the majority of them continue to use outdated reporting metrics and summary statistics to brief different stakeholders and make decisions based on observation rather than on rigorous analysis. The insurance sector is no exception to this.

With the advent of e-commerce and social media, the insurance business faces opportunities in its operations, risk management, and the marketplace. E-commerce and social media platforms have provided innovative chances for the creation of new forms of insurance products. However, there is also more competition than from outside the sector and from within. More and more competitors are entering the market with simpler products and more competitive pricing, assisted by the evolution of platforms and sales organizations. The insurance industry has only limited experience with big data. A key challenge for the insurance industry is to charge each customer an appropriate price for the risk they represent. Risk varies widely from customer to customer, and a deep understanding of different risk factors helps to predict the likelihood and cost of insurance claims. This study studies non-life insurance, with a specific focus on vehicle insurance. An insurance claim arises when a policyholder creates a formal request to an insurer for coverage or compensation for an unfortunate event because of some kind of accident. Insurance companies must predict on average how many claims are going to occur and the severity of these claims in order to enable the insurer to set a fair price for their insurance products accordingly. Thus, the cornerstone of premium estimates in the vehicle insurance sector is claim prediction. For motor insurance, several studies have been done to personalize the premium estimate.

6.5.2. Machine Learning Algorithms Used in Insurance

Tree-based machine learning methods are used to solve a challenging problem for vehicle insurance companies, predicting claims size. Different kinds of predictive models are applied, including linear regressions, regression tree and tree-based ensemble learning methods. Empirical evidence indicates that tree-based ensemble learning algorithms are very effective machine learning methods. They have a good predictive performance on unseen data, provide important information about the modeling, can easily handle a mixture of continuous and categorical features and are robust to outliers. Despite their advantages, the use of tree-based ensemble learning methods in the vehicle insurance domain is not yet well researched.

In this study, different tree-based machine learning methods are used to predict the claim size of a vehicle insurance company in Ethiopia. The performances of these models are evaluated and compared and the drivers of prediction are explored. The results indicate that tree-based ensemble methods outperform the classical least square method in predicting claims size. Their predictive power is high as compared to the Least Square method, and the random forest model is the best performing method. The tree-based ensemble models also provide a lot of information about feature importance in predicting claims size.

Insurance companies build an extensive portfolio of diverse risk types. They collect personal data of varying attributes of all the risk types relevant to their business. Data preprocessing and cleansing are carried out to enhance data quality and accuracy for predictive modeling. Data size reduction enhances the performance of all machine learning applied models. Transformation functionalities of the SAS digital analytics environment are useful for numeric size reductions. SAS data exploration is preferred over R and Python because of its efficient data linking, sampling, grouping and aggregating features. Needed variables are selected within the business context to start the exploratory modeling with IV-based statistical variables.

6.6. Conclusion

This paper proposed a hybrid ML approach to tackle the widely studied vehicle insurance claims prediction. Although this insurance data science problem is studied by various methods, due to the fact that it is essentially different from most well performed benchmarks, traditional ML methods are also investigated in detail. The work experimented on a wide range of ML algorithms using different paradigms and an extensive set of features from current and new data sources. The work pointed out a new challenge in large scale claim size modeling in the insurance industry; by attacking the design of features, tuning of models, selections of arbitrary ensembles, and construction of other maturity ambitiously. The work provided a relatively fine-established benchmark for the study of claim prediction for insurance data science researchers. The very narrow sweet zone of performance is a challenge to insurance data science practitioners. It is hoped that the deep learning ensemble from a more intelligent construction and design of features according to the actual situation of the insurance industry can push the performance that batch-wise rule of thumb-translation ML fails.

The claim size prediction from raw data appears to be a more embroidered yet fruitful prospect, and intermediate data preprocessing is meaningful. To begin with, to model claim sizes in vehicle insurance and respond with an organization of honor. Random forests, decision trees, and NN outperform first-to-try and scalable tree-based machine learning methods with enhanced stacks. Secondly, to refine prediction, a debiasing ensemble is necessary but challenging due to training sample selection. Lastly, future work on deep encoders for hidden representations, integration with holistic loss and revenue-driven profit simulation, and the exploration of data-driven intelligence for effective product design and pricing is foreseen.

6.6.1. Future Trends

In light of the rapid transformation of the insurance industry, it is essential to take stock of the emerging trends affecting the property and casualty (P&C) insurance market. Technology is gaining more traction in companies' search for much-needed efficiency. Proliferation of insurance technology (insurtech) firms. Newer entrants are particularly strong in the fields of risk selection through improved underwriting tools and service, which includes transparency in pricing and claims processing. Fee-for-service models, alternatives to the commission-based agency model, are gaining traction. Artificial intelligence-based predictive analytics methods are rapidly deployed by firms to improve risk assessment and claims processing.

Despite the skepticism around the fitness of some of these new methods, the P&C insurance market is constantly being reshaped by these changes. ‘Big data’ is increasingly sought-after for use in prediction. Vast amounts of new data and sources are being monitored and assessed for modeling purposes. While word-of-mouth and prior experience had long been the only ‘social’ data vehicles, social media in its various forms is becoming commonplace in the assessment of moral hazard (or indemnity). More qualitative data sources, such as reports based on imagery, are also becoming regular fare for firms seeking to assess risk.

The development of self-driving cars and informations systems in commercial aggregate transport, such as Telematics for commercial transport professionally monitored by fleet managers and on-board cameras; in ride-sharing, such as dash cam videos from taxi service vehicles; and in micro-mobility, such as data from dockless electric scooters, are gathering much attention from research and firms alike on claims processing needs. Access to these vast amounts and newer types of data sources, however, is raising some deep-rooted concerns. Discussants fear that established policy considerations and frameworks may go by the wayside amidst the heated competition to seize opportunities in the fast-evolving marketplace.

References

Cyriac, T., Regenstein, J., & McConnell, S. (2025). Agentic AI in Financial Services and Insurance. Snowflake. Retrieved from snowflake.comSnowflake AI Data Cloud

Rowe, T. (2024). How Agentic AI is Transforming the Banking Industry. Intelligent Core™. Retrieved from intelligentcore.ioIntelligent Core™

Lloyds Banking Group. (2024). Lloyds hires an Amazon Web Services executive as its new AI chief. Financial Times. Retrieved from ft.comFinancial Times

Klein, C. (2025). Agentic AI is a 'big next step' in AI's evolution. Axios. Retrieved from axios.comAxios

Dhawan, R. (2024). AWS is helping financial giants like JPMorgan and Bridgewater with their AI ambitions. Business Insider. Retrieved from businessinsider