# Insurance Premium Calculation Using Machine Learning Methodologies

by

Foad Haji Mohammad

A thesis submitted to the Faculty of Graduate and Postdoctoral Affairs in partial fulfillment of the requirements for the degree of

Master of Science

in

Management

Carleton University
Ottawa, Ontario

## Abstract

This research introduces a novel, data-driven optimization methodology for determining insurance premiums and illustrates it using an Iranian health insurance company as a case study. Using the optimal classification machine learning model, the insureds are divided into groups based on their potential to cause disruptions. Next, the indemnity of each group is estimated. Then a linear programming model is used to determine the premium for each group based on their estimated indemnification. Multi-criteria decision-making is utilized to identify the best machine learning algorithm after evaluating the performance of many algorithms using five metrics: accuracy, precision, recall, F1 score, and Matthews correlation coefficient. This study addresses the two common and significant disruptions in the insurance sector: errors in predicting possible losses and a decrease in the insurer's market share. The methodology discussed in this work can be applied to other insurance domains, expanding its practical applications.

*Keywords*: Insurance premiums, Actuarial models, Risk assessment, Data-driven optimization, Machine learning

## Acknowledgments

I want to express my most profound appreciation to the following individuals who have contributed to the successful completion of my Master's thesis.

Firstly, I would like to thank my thesis supervisor, Professor Aaron Nsakanda, for his guidance, support, and valuable feedback throughout my research. His help was invaluable in shaping the direction and scope of this thesis.

I am also grateful to my thesis committee members, Professor Ahmed Doha and Professor Mohamed Al Guindy, for their insightful comments and suggestions that helped me improve my work.

Moreover, I would like to thank my family and friends for their unconditional love and support throughout my academic journey. Their encouragement and motivation kept me going during challenging times.

Lastly, I express my gratitude to the University, which provided me with the opportunity to pursue my academic goals and allowed me to develop as a scholar.

# Table of Contents

## List of Tables

## List of Figures

## List of Annexes

# Chapter 1: Introduction

This chapter will begin with the background and research motivation, then proceed to discuss the research objectives and contributions. Furthermore, the last section will present the thesis outline to understand its structure clearly. By the end of this chapter, readers will have gained insight into the research's context, goals, and organization.

## 1.1 Background and Research Motivation

One of the most common contracts is the insurance contract which determines the claims the insurer is legally required to cover and the premium insureds have to pay (Torraca & Fanzeres, 2021). The insurance industry is one of the world's most essential industries generating billions of dollars annually (Ranjbari et al., 2019). For instance, health insurance is the second most popular type of insurance contract in Iran after vehicle insurance, generating 456.6 million dollars in 2017 only (FinancialTribune, 2017).

The nature of insurance decisions, like any other financial industry, is to make a balance between possible growth and stability. The insurance industry has been structured based on the concept of risk-sharing. Therefore, the insurer tries to anticipate the prospective indemnity to reduce the risk of bankruptcy (Ayuso & Santolino, 2012).

In the insurance industry, as in every other industry, there are disruptions that, if they occur, pose a significant threat to the industry's continued existence. In general, a disruption is an unforeseen problem that disrupts an activity or process and, if severe enough, causes it to deviate from its optimal state (Gans, 2016).

Overall, there are four primary disruptions in the insurance industry (Ceballos & Kramer, 2019; Douven et al., 2020; Liu et al., 2017; Settipalli & Gangadharan, 2021):

- **Insurance frauds and the increase of false losses:** This disruption occurs when an insured falsifies accidents to receive damages from the insurer. Unreal medical documents for health insurance (for example, seeking compensation for an individual illness other than the insured person) and manipulated accidents for accident insurance (for example, intentionally causing accidents or aggravating their effects) are such insurance disorders (Derrig, 2002).
- **Decrease in the insurer's market share:** This disruption occurs when the insurer's share in the relevant market decreases because of the decrease in the insurer's desirability. When an insurer is not desirable, the insured will not renew their contracts for the next year, or even in some cases, they will cancel the contracts of the current year. One of the most common reasons for the decrease in the insured's favorability is that the insurer declares a high premium for its services. Among other reasons for reducing the value of the insurance provider are the low level of risk coverage, inappropriate advertising, low service speed, etc.; however, none are as comprehensive and essential as the impact of the insurance premium price (Halbersma et al., 2011).
- **Error in predicting possible future losses:** This disruption occurs when the insurer mispredicts its possible future losses. As mentioned earlier, insurance is based on the concept of risk sharing. Therefore, naturally, the insurer predicts possible future losses to reduce the risk of bankruptcy. If these losses are predicted to be less than the actual amount, it is evident that the insurer will suffer. Also, if

this prediction is more than the actual damage to the insured, the investment benefit for the insured will be lost. In cases where the insurer only uses classical methods, such as statistical methods, this disruption becomes more apparent. Even today, even though data science is used in advanced countries to predict possible future losses, this is still considered one of the most common disruptions for insurers (Ceballos & Kramer, 2019).

- **Occurrence of unforeseen obligations:** This disruption occurs when the insurer must cover risks that did not exist at the time of signing the contract. For example, we can mention the global epidemic of the coronavirus. When the insurer estimated the severity of the risks covered, there was no risk of such a disease. Although this rarely happens, it is considered one of the most critical risks for the insurer.

The most prevalent disruptions are errors in predicting possible future losses and a decrease in the insurer's market share (Ceballos & Kramer, 2019; Douven et al., 2020; Liu et al., 2017). To better understand the errors in predicting possible future losses, it can be helpful to state this simple example. Suppose an insurer has $\alpha$ insureds, and the insurance premium received for each insured is $\beta$. In this case, the insurer's income ($\pi$) will be equal to $\pi = \alpha \times \beta$. Suppose during a financial period, a number of insureds ($\gamma$) have an accident (disruption) and claim damages ($\theta$) according to their contract with the insurer. In that case, the insurer will suffer a loss. The larger the quantity $\gamma \; or \; \theta$ is, the greater the damage inflicted on the insurer, i.e., $\delta = \gamma \times \theta$, will be more significant. With these interpretations, it is evident that as the value of $\delta$ increases, the insurer gets closer and closer to bankruptcy. If the value of $\delta$ is greater or equal to the value of $\pi$, the insurer has reached bankruptcy.

As demonstrated in the preceding paragraph, the insurer will be disrupted in the event of an increase in the number of insureds who have experienced a disruption or in the amount of claimed damages (generally, in the event of an increase in the amount of monetary damage caused to the insurer). Therefore, the more capable and accurate an insurer estimates the number of disruptions and potential indemnity, the more profitable it will be and the lower its bankruptcy risk.

To better understand a decrease in the insurer's market share, it is essential to closely examine how insurance premiums are calculated. Historically, there were two primary approaches to determining insurance premiums (David, 2015). One is from the insured's perspective, which seeks to reduce the amount of insurance premiums. The other is from the insurer's perspective, which naturally strives to increase the amount of insurance premiums to lower the risk of disruption. These two perspectives are entirely irrelevant in today's market. On the one hand, non-profitable premiums for the insurer will eventually lead to bankruptcy. On the other hand, if the announced insurance premium is not competitive, the insured will immediately sever ties with the present insurer for the upcoming year and select a new insurer to finalize the contract. The increase in insurance premiums will lead to the lost loyalty of even the most devoted consumers, decreasing the insurer's market share.

## 1.2 Research Objectives and Contributions

Based on the two critical disruptions mentioned in the previous section, a fundamental question is posed which this study seeks to answer: "How can the insurance premium

determined by the insurer for a specific type of insurance be defined so that the insurer is at the lowest level of bankruptcy risk while maintaining an acceptable level of attractiveness and benefit for the insured?"

Correctly determining insurance premiums is vital for enhancing profitability and the insurer's existence and avoiding insolvency. If the determined insurance premium cannot compete with the premiums of the insurer's competitors, the insured's tendency goes toward the competitors. In addition, as indicated previously, the insurer's bankruptcy is inevitable if the insurance premium is assessed to be too low to meet the risks associated with the insurance liability.

Therefore, the primary objective of this study is to determine how to find an insurance premium that will cover as much of the insurer's losses as possible. In addition, to prevent the insurer's insolvency, the determined insurance premium should not result in the insured's investment loss.

This study also seeks to answer the following sub-questions:

- How can the insurer predict the indemnity of its prospective insureds?
- What information is needed for this prediction?
- How can the predicted indemnity be used for calculating the optimal insurance premium?

In order to attain this objective (considered a significant objective in the financial markets), we found it necessary in this study to classify the insureds based on their behaviour and characteristics and to determine the insurance premium for each group separately. The logic behind doing this is that in non-life insurance, indemnities will be claimed by people struggling with a disruption (for instance, a disease for health insurance). Therefore, it is

essential to pinpoint the people more likely to be disrupted, and this can be reached through assessing each insured (in the case of health insurance, it would be beneficial to study the insured's health information). Ultimately, this can raise the insured's satisfaction, stabilize the insurer's position in the competitive financial markets, distribute the risk level more appropriately, and boost the insurer's profits.

Overall, in order to fulfill our research objective, we need to analyze customer behaviour and identify patterns in the data collected from them using machine learning and data mining techniques, determine the level of potential future damage caused to the insurer by each insured through their grouping, and determine the insurance premium for each group based on the damage caused by the same group.

Our work contributes to the theory and practice by:

- Presenting a new and optimal method for determining insurance premiums based on the behaviours and characteristics of the insureds
- Providing different insurance premiums for different classified groups
- Considering the two fundamental subjects of profitability and competitiveness in developing a mathematical model to determine the appropriate insurance premium for each insured in a classified group

This research can boost the insurance companies' profitability and their level of preparedness to confront disruptions for the following reasons:

- Finding the right insurance premium for each group of insureds improves the utility for consumers who are typically less concerned and, as a result, feel less need to purchase insurance (raising the benefit for the insured and the insurer's market share).

- Classifying the insureds and determining the specific insurance premium can consolidate the insurer’s position in a competitive market.
- Grouping the insureds can share the risks between the parties more appropriately and fairly.

This thesis's applicability extends beyond insurance contracts and the insurance industry. As a result, we anticipate that our developed methodology can affect substantial shifts across the financial sector.

## 1.3 Thesis Outline

The rest of this thesis will be structured into five chapters. Chapter two will examine and review the relevant literature and identify the existing research gaps. The research problem will be defined and formulated in detail in chapter three. In chapter four, the solution methodology will be introduced inclusively. The proposed method will be validated using a case study, and the results will be assessed in chapter five. Finally, the research reaches its end with a conclusion in chapter six, also providing managerial implications and limitations of the study.

# Chapter 2: Literature Review

This chapter will begin by presenting a series of studies conducted to calculate insurance premiums. Additionally, we will discuss recent research studies that highlight the application of machine learning in the insurance industry. Finally, we will identify the research gaps and discuss the ones our work aims to fill.

## 2.1 Overview of Significant Studies in Insurance Premium Calculation

Different methodologies have been proposed in the literature for calculating insurance premiums in various contexts, such as auto, livestock, fire, and cyber insurance. While most works from actuarial science and research use traditional statistical methods, others have used new principles or models for premium calculation, such as continuous cumulative prospect theory, simulation methods and the analytic hierarchy process (AHP) technique. Overall, these articles demonstrate the diverse approaches that can be taken in actuarial science to calculate insurance premiums and manage risk.

David (2015) proposed a generalized linear model to calculate auto insurance premiums using the frequency of insurance claims. By analyzing the frequency of insurance claims, insurers can better understand the risk associated with insuring a particular driver and adjust premiums accordingly to maintain profitability while remaining competitive. Hence, this approach could reduce the level of risk associated with insuring high-risk drivers, such as young or inexperienced drivers, by more accurately pricing their insurance premiums based on their individual risk profiles. Overall, this work demonstrates the potential for

statistical modelling techniques to be used in the actuarial science field to develop more accurate and effective insurance pricing models.

Pai et al. (2015) used credibility analysis to calculate the insurance premium of livestock mortality insurance. Credibility analysis is a statistical method that combines an individual's experience (e.g., claims history) with the experience of a larger group to improve the accuracy of risk estimates. By using this approach, Pai et al. estimated the probability of livestock death and developed a pricing model that reflected the true risk of insuring against this event. The authors demonstrated that their approach resulted in more accurate premiums and helped to reduce the volatility of insurers' profits. This study illustrates how statistical techniques can refine insurance pricing, resulting in fairer premiums for policyholders while maintaining profitability for insurers.

Nardon and Pianca (2019) defined a new principle for calculating the insurance premium using the continuous cumulative prospect theory. This theory suggests that individuals evaluate potential outcomes based on subjective value rather than objective probability, which can lead to inconsistencies in decision-making. By incorporating this principle into insurance pricing models, the authors sought to develop more accurate pricing schemes that reflect the subjective nature of risk perception. The authors demonstrated the effectiveness of their approach by applying it to the case of crop insurance, which can be subject to significant variations in weather patterns and other factors that affect crop yield. Their results showed that the continuous cumulative prospect theory could be used to develop more accurate and fair insurance premiums for crop insurance, which can help reduce farmers' financial risk and improve insurance market stability. This study highlights

the potential benefits of incorporating behavioural economics principles into insurance pricing models, leading to more equitable and efficient insurance markets.

Gumus and Uzekmek (2019) calculated the insurance premium for fire insurance by determining the risk score using the AHP technique. The AHP technique involves breaking down complex decision-making processes into smaller, more manageable components to facilitate decision-making. The authors used this approach to determine the risk score for each policyholder by evaluating various factors, such as the property's location, building materials, and fire safety measures in place. They then used these risk scores to determine the appropriate insurance premium for each policyholder, which reflected the actual risk of a fire occurring and the potential cost of damages. Their study demonstrated that the AHP technique could be a valuable tool for insurance companies to develop more accurate and transparent insurance pricing models, which can lead to fairer premiums for policyholders and improve the overall stability of the insurance market.

Finally, Yang et al. (2020) developed a new framework for calculating insurance premiums in the context of cyber insurance, which protects against losses arising from electronic intrusions or cyberattacks. The authors utilized steady-state simulation results to model the potential electronic intrusions and their hypothesized direct impacts, such as damage to data and systems, business interruption, and loss of reputation. They then integrated these simulation results with traditional actuarial techniques, such as risk assessment and loss modelling, to determine the appropriate insurance premium for each policyholder. The framework also incorporated an analysis of risk factors, including the type of business, size of the organization, and the security measures in place, to ensure that premiums were tailored to each policyholder's specific needs and risks. The authors

demonstrated that their framework could help insurance companies to more accurately and reasonably price cyber-insurance policies, ultimately leading to more effective risk management and improved cybersecurity practices for businesses.

It should be noted that none of the above studies use machine learning methodologies for insurance premium calculation. In the uses of non-machine learning methods to predict indemnity and calculate insurance premiums, two issues may arise: the calculation time may become too lengthy (Zhu et al., 2021), and the results may be unreliable in some cases (Ij, 2018).

## 2.2 Overview of Noteworthy Studies in the Application of Machine Learning in the Insurance Industry

The papers included in this section all relate to the application of data science and machine learning (ML) in the insurance industry. They address different aspects of the industry, such as fraud detection, risk assessment, profitability, claims management, and customer behaviour analysis. However, they all share a common theme of using advanced analytical techniques to extract insights from large and complex insurance data sets. Overall, these papers illustrate the potential for data science and machine learning to improve the insurance industry's efficiency, profitability, and customer experience. Several studies conducted in the relevant fields have proved the speed, reliability, and accuracy of ML models (Mahesh, 2019). Here it should be mentioned that ML models consider not only different aspects of a problem to reach a final decision but also, in some cases, they can benefit from experts' knowledge (Alipour-Vaezi et al., 2022).

Yeo and Smith (2006) used a clustering algorithm and a neural network classifier to determine the price of automobile insurance premiums based on data from an Australian insurance company. They employed a hierarchical clustering algorithm to partition the drivers into different risk groups. They applied a neural network, a meta-heuristic method, to estimate the expected claim cost for each group. The neural network model was trained using a backpropagation algorithm, and the model's performance was evaluated using mean squared error and mean absolute error. The study demonstrated the effectiveness of clustering algorithms and neural networks for predicting automobile insurance premiums based on driver characteristics and driving behaviour.

Dora and Sekharan (2015) proposed a method for detecting and predicting possible fraud in insurance premiums using big data, a Hadoop environment, and analytical methods. The authors used clustering algorithms to group policies based on similar characteristics and identify potential fraud cases. They also used outlier detection techniques to identify policyholders with unusual patterns in their behaviour. The proposed method involved data mining and machine learning techniques such as decision trees, logistic regression, and neural networks to detect fraudulent activities. The type of insurance is not specified in the paper, but the authors suggest that the proposed method can be applied to different types of insurance, including health, life, and property insurance.

Fang et al. (2016) introduced a novel profitability method for the insurance industry by incorporating the insurer's debt reserve factor. Fang's method effectively measures the insurer's actual share by factoring in purchase behaviour (purchase history) and future cash flow. In addition, regression and random forest methods, as well as methods for big data analysis, such as data mining, were used to predict the insurance company's profitability.

Their empirical research revealed that geographic region, age, insurance status, gender, and customer origin are the most significant predictors of customer profitability.

Lin et al. (2017) developed a large-scale data mining approach for the insurance industry using an innovative sampling technique combined with ensemble learning, a meta-heuristic approach, and random forest algorithms. The authors also employed parallel computing and a memory-sharing mechanism to address the challenge of analyzing large-scale data sets. The proposed approach was tested on a real-world auto insurance claim data set, demonstrating its effectiveness in improving prediction accuracy and reducing model complexity.

Wang and Xu (2018) introduced a novel deep-learning technique for insurance fraud detection in non-life insurance. This paper developed a text analysis model based on Latent Dirichlet Allocation to extract accident text description features. The suggested model uses textual and standard numerical features to identify fraudulent claims. The authors suggest that their approach can be a useful tool for insurance companies to detect fraudulent claims in a timely and accurate manner, which can help to reduce losses and improve the overall efficiency of the insurance industry.

Huang and Meng (2019) presented a comprehensive framework based on rapid classification techniques for the insurance industry. Their article examined the effects of the most complex driving behaviour variables on automobile insurance premiums. Their model improved pricing precision and interpretation by combining data collection and machine learning techniques.

Krashinakova et al. (2019) proposed a new approach to model the insurance renewal price adjustment problem as a sequential decision-making problem using the Markov decision-

making process. The article aimed to determine the optimal pricing strategy for the renewal of insurance contracts over multiple periods, considering the uncertainty in future claim occurrences and policyholder behaviour. The proposed model integrated reinforcement learning algorithms to determine the optimal pricing strategy in a dynamic environment (a form of meta-heuristic optimization). The article does not use game theory explicitly, but it uses mathematical modelling extensively, including the use of stochastic processes, probability distributions, and optimization techniques.

Dutta et al. (2021) used a data-driven forecasting method to predict the demand for health and treatment insurance premiums. They used a combination of time-series analysis and machine learning algorithms to analyze the historical data and forecast future demand. They compared the performance of different models and found that the Gradient Boosting Regression Tree (GBRT) algorithm performed the best in terms of accuracy and efficiency. Also, their studies made valuable comparisons between actual and predicted values, proving this research's accuracy.

Petneházi (2021) conducted a study using machine learning (ML) algorithms to predict the value at risk in the health insurance industry. Specifically, the study used an artificial neural network (ANN) trained by price history to make predictions. Data mining was used in this study to analyze and process large volumes of insurance data. The study found that ensemble-based ML algorithms were more precise than other methods for mining insurance data and could be trusted for this purpose. Furthermore, the research suggests that this approach could be used to help insurers identify risk factors and make more accurate predictions about future claims. Overall, the study highlights the potential of ML in the health insurance industry to improve risk management and increase efficiency.

Lastly, Zhang et al. (2021) aimed to determine the insurance indemnity for health insurance using a linear regression algorithm and optimal control techniques. The study focused on developing a mathematical model to estimate the optimal insurance indemnity for a given patient based on their health information and history. The authors used linear regression to estimate the health expenditure of patients and optimize the insurance indemnity accordingly. The optimal control techniques were then applied to find the optimal insurance indemnity that maximizes the expected utility of the patient. The study used real-world data from a large health insurance provider in China to validate the proposed model's accuracy.

Table 1 summarizes the main aspects of the insightful papers presented in this section.

**Table 1 Summary of the main aspects of the recent insightful papers**

| Papers | Insurance Type | | Data mining | MADM | Solution Methodology | | Game theory | Mathematical Modeling |
|---|---|---|---|---|---|---|---|---|
| | Non-life | Life | | | Metaheuristic | Exact | | |
| Yeo and Smith (2006) | • | | • | | • | | | • |
| Dora and Sekharan (2015) | • | • | • | | | | | |
| Fang et al. (2016) | • | | • | | | | | • |
| Lin et al. (2017) | • | | • | | • | | | • |
| Wang and Xu (2018) | • | | • | | | | | |
| Huang and Meng (2019) | • | | • | | | | | |
| Krasheninnikova et al. (2019) | • | | • | | • | | | • |
| Dutta et al. (2021) | • | | • | | | | | • |
| Petneházi (2021) | • | | • | | | | | |
| Zhang et al. (2021) | • | | • | | | • | | • |

## 2.3 Research Gap

This section summarizes research gaps identified using the information presented in this chapter and table 1 and discusses the research gap our work aims to fill.

- Compared to non-life insurance, life insurance has a lower priority on the research agenda of academics and requires more attention.

- Little to no research was identified that uses data analysis in the insurance industry to eliminate the insurance provider's readiness for disruption.
- Although ML has been used widely in the insurance industry, most studies have used traditional methods, including expert-based techniques and game theory, to determine the insurance premium. Therefore, it can be concluded that researchers have not discovered the use of data-mining techniques with the readiness analysis approach to determine the insurance premium.
- Based on the research findings, no studies were found using multi-criteria decision-making techniques to select the optimal data mining algorithm.

As shown in table 2, from the gaps identified, our work can cover a substantial part to a fair extent by synthesizing data mining methods, mathematical modelling based on actuarial concepts, and multi-indicator decision-making techniques. Moreover, the proposed methodology can even aid decision-makers worldwide in addressing the previously mentioned pertinent insurance-related issues.

**Table 2 Summary of the main aspects of the recent insightful papers and current research**

| Papers | Insurance Type | | Data mining | MADM | Solution Methodology | | Game theory | Mathematical Modeling |
|---|---|---|---|---|---|---|---|---|
| | Non-life | Life | | | Metaheuristic | Exact | | |
| Yeo and Smith (2006) | • | | • | | • | | | • |
| Dora and Sekharan (2015) | • | • | • | | | | | |
| Fang et al. (2016) | • | | • | | | | | • |
| Lin et al. (2017) | • | | • | | • | | | • |
| Wang and Xu (2018) | • | | • | | | | | |
| Huang and Meng (2019) | • | | • | | | | | |
| Krasheninnikova et al. (2019) | • | | • | | • | | | • |
| Dutta et al. (2021) | • | | • | | | | | • |
| Petneházi (2021) | • | | • | | | | | |
| Zhang et al. (2021) | • | | • | | | • | | • |
| Current Research | • | | • | • | | • | | • |

# Chapter 3: Problem Definition and Formulation

This chapter provides an introduction to the fundamentals of calculating insurance premiums. The concept of risk evaluation and financial stability in the insurance industry will be presented first, followed by an overview of actuarial models, which play a crucial role in assessing risk and determining appropriate premiums.

It should be noted that the term "actuary" is a unique concept within the insurance industry that does not have an equivalent in other fields. Therefore, throughout this thesis, the term "actuary" will be used to refer to this important topic.

## 3.1 Evaluating Risk and Financial Stability in Insurance

Rotar (2014)'s book "Actuarial Models: Insurance Calculations" clearly shows actuarial and financial industry risks. It is assumed that two investors, "A" and "B", expect random incomes corresponding to the variables $X_1$ $and$ $X_2$, respectively. For simplicity's sake, assume that the distributions of these two random variables are independent. As a result, their mean and variance will be $m = E\{X_i\}$ and $\sigma^2 = Var\{X_i\}$, respectively. Moreover, Rotar's study supposes that "A" and "B" measure their level of investment risk based on the income variance criterion. Since each is risk averse, they attempt to minimize the risk imposed on their income. In order to achieve this objective, both investors divide their total income equally. In this instance, the resultant random income is computed using the formula (3-1).

$$Y = \frac{1}{2}(X_1 + X_2) \tag{3-1}$$

In addition, their new average income is determined using the equation (3-2), which is equal to its value before risk sharing.

$$E\{Y\} = \frac{1}{2}(E\{X_1\} + E\{X_2\}) = \frac{1}{2}(m + m) = m \quad (3\text{-}2)$$

However, the essential point is about the new variance. This value is represented by the equation (3-3).

$$Var\{Y\} = \frac{1}{4}(Var\{X_1\} + Var\{X_2\}) = \frac{1}{4}(\sigma^2 + \sigma^2) = \frac{1}{2}\sigma^2 \quad (3\text{-}3)$$

Considering this critical, let us assume that our investors' number extends to $n$ (n>0). In this case, the income variables in question are equal to $X_1.X_2.\ldots.X_n$. Accordingly, the income variable will be:

$$Y_n = \frac{1}{n}(X_1 + X_2 + \cdots + X_n) \quad (3\text{-}4)$$

Moreover, the new mean and variance will be calculated based on the following equations, respectively, in (3-5) and (3-6):

$$E\{Y\} = \frac{1}{n}(E\{X_1\} + E\{X_2\} + \cdots + E\{X_n\}) = m \quad (3\text{-}5)$$

$$Var\{Y\} = \frac{1}{n^2}(Var\{X_1\} + Var\{X_2\} + \cdots + Var\{X_n\}) = \frac{n\sigma^2}{n^2} = \frac{\sigma^2}{n} \quad (3\text{-}6)$$

We know that the above equation tends to be zero for large values of $n$. Similarly, it can be inferred that the random variable $Y_n$ will tend toward the value m. Therefore, it can be concluded that if there are many partners, the risk of each will be reduced to zero.

This phenomenon is known as risk redistribution or sharing, and it is present in most financial stabilization mechanisms, particularly in the insurance industry. In other words, the existence of the insurance industry can be attributed to the insurer’s utilization of the concept of risk redistribution.

It is worth mentioning that variance is not the only risky feature. This feature is far from the best risky feature. The following criteria can be mentioned for comparison and selection among risky options:

- The mean value criterion
- Value at risk
- The mean-variance criterion

## 3.2 Insurance Premiums and Portfolio of Insurance Projects

Suppose the random variable $X_i$ represents the payment made to customer i. When $X$ is revealed, $n$ insureds are organized as a group. This group is also referred to as the risk portfolio. In this instance, the total payment random variable is denoted by the expression (3-7):

$$S_n = X_1 + X_2 + \cdots + X_n \tag{3-7}$$

For a relatively small size n, the above equation is generally accurately calculated. However, we must employ various approximations for a large size n, the most well-known of which is the Central Limit Theorem in probability theory.

Insurance companies typically receive insurance premiums to cover their clients' risks. $c_n$ represents the total insurance premium for the risk portfolio of size n. The insurance premium $c_n$ should be greater than the average total payment $\mathrm{E}\{\mathrm{S_n}\}$ if a company wishes to maintain its financial stability. The typical term for the average $\mathrm{E}\{\mathrm{S_n}\}$ is the net premium. Security loading, an additional payment for the risk borne by the insurer, is another factor that must be included. The equation (3-8) can be used to calculate security loading:

$$\Delta_n = c_n - E\{S_n\} \tag{3-8}$$

Determining the lowest acceptable value of $\Delta_n$ is the most high-priority task and function of actuarial modelling. One consideration in this regard is that the larger the portfolio, the less security is required for each client. This issue is similar to the earlier-discussed issue of risk redistribution.

In reality, the total payment random variable ($S_n$) is such that the number of variables is uncertain. In other words, the quantity of the number of variables is a random variable. In this case, the total payment random variable is the relation (3-9), where N is a random variable representing the total number of claims in a given time frame.

$$S_N = X_1 + X_2 + \cdots + X_N \quad (3\text{-}9)$$

The above model can be used in two modes:

The first mode is related to the future risk portfolio. That is when the number of prospective customers is uncertain, resulting in their number being a random variable.

The second mode, more prevalent and widely used in the insurance industry, is employed when calculating the total losses for a specific current portfolio instead of each insured's losses. The company must pay by check. In this instance, $N$ is the number of future claims the company will receive within a given time frame, and $X$ is the payments associated with these damage claims. Examining the possible distributions of the variable $N$ is the first step in this model's analysis. Other statistical distributions should be considered despite the prevalence of the Poisson distribution.

The two models briefly described above are both static. If time is also factored into the preceding relationship, the model becomes dynamic and can be verified using the following equation:

$$S_{N_t} = X_1 + X_2 + \cdots + X_{N_t} \quad (3\text{-}10)$$

The model above represents the total cumulative damage claim up to time $t$. The company's cash flow of insurance premiums should be considered along with the claim flow. Let $c_t$ be the total premium collected up to time $t$. Moreover, at time $t = 0$, the company has an initial surplus of $u$. In this case, the company's surplus at time $t$ is a random variable that is defined as shown in relation (3-11):

$$R_t = u + c_t - S_{N_t} \tag{3-11}$$

It should be added that this model is one of the well-known surplus models, and others can be implemented too.

### 3.3 Actuarial Equations for Determining Insurance Premiums

In actuarial equations, equation (3-12) is used to determine the premium classically. $c$, $\theta$, and $E\{S_n\}$ represent the insurance premium, loading coefficient, and mathematical expectation of the random variable payment, respectively.

$$c = (1 + \theta) \cdot E\{S_n\} \tag{3-12}$$

To calculate the equation (3-12), it is necessary first to determine the parameter of the loading factor. Equation (3-13) is presented to determine this parameter. It is worth noting that $Var\{S_n\}$ indicates the variance of the random variable payment $S_n$. Moreover, $q_{\beta_s}$ represents the (1-$\beta_s$)th percentile of the standard normal distribution. In other words, it is the value such that the area under the standard normal distribution to the left of $q_{\beta_s}$ is equal to 1-$\beta_s$. The value of $q_{\beta_s}$ is often used in actuarial science to determine the amount of risk associated with a particular event.

$$\theta = \frac{q_{\beta_s}\sqrt{Var\{S_n\}}}{\sqrt{E\{S_n\}}} \tag{3-13}$$

The purpose of providing these relations is that the insurer's bankruptcy risk becomes lower than a specific limit. This issue is formulated in relation (3-14). In this regard, $\beta$ limit is mentioned as the limit of damage claim coverage in many texts.

$$P(S_n \geq c) \leq \beta \tag{3-14}$$

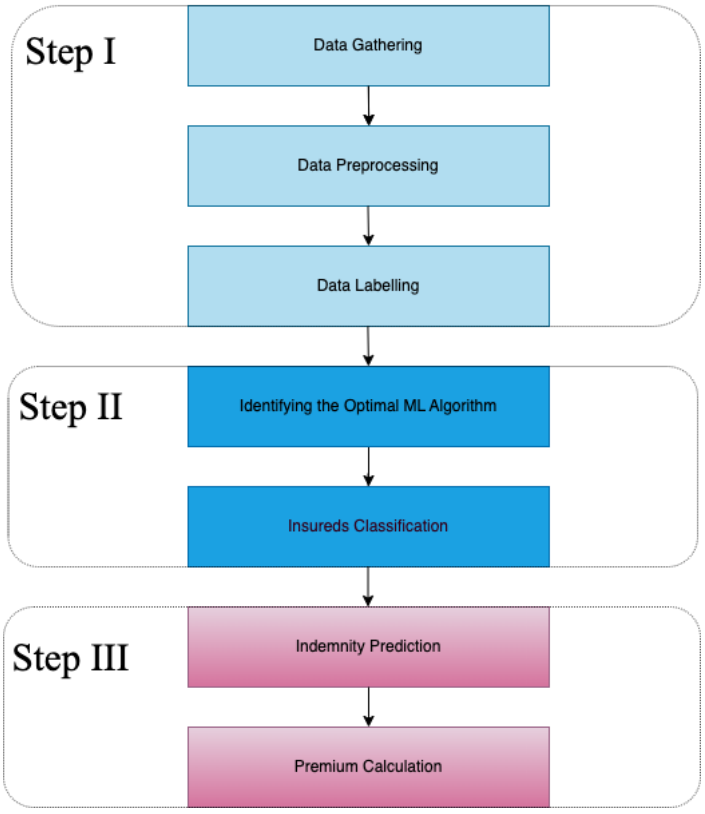
In conclusion, this chapter provided a good foundation for understanding the fundamentals of calculating insurance premiums. The average total payment, net premium, and security loading are all important components in determining the insurance premium and should be considered in developing our own model for calculating the insurance premium in the next chapter. However, it is essential to note that insurance companies today may have additional factors and considerations when determining premiums, which should also be considered in developing our model.

# Chapter 4: Solution Methodology

This thesis chapter provides a novel solution methodology for managing two prevalent and fundamental insurance disruptions defined in the first chapter (a decline in the insurer's market share and an error in predicting potential losses). The essential aspect of the offered methodology is using data mining and machine learning techniques to analyze the insured's probability of causing damage in the insurance sector. In this methodology, it is considered that the insurer strives to set a premium that ensures competitiveness in today's competitive market while avoiding bankruptcy, as discussed in the previous chapter.

We propose a three-step methodology that follows the analytical process summarized in figure 1. Our methodology is not restricted to a particular form of insurance (such as car insurance, health and medical insurance, housing, accidents, etc.).

**Figure 1 A summary of the solution methodology steps**



In the subsequent sections of this chapter, each step of the proposed methodology is discussed in detail.

## 4.1 Data Gathering

Collecting relevant data to investigate and comprehend the research problem is crucial in this study. Moreover, appropriate data collection is required for training and evaluating the machine learning models. As depicted in figure 1, data collection is the first step in the proposed process.

## 4.2 Data Preprocessing

This is the second and most time-consuming stage. Occasionally, the data acquired from primary information sources are accompanied by flaws that analysts must detect and attempt to rectify or correct. Therefore, prior to analysis and result acquisition, acquired data is edited and cleaned.

In the preprocessing step, the following should be generally carried out to prepare the data for further analysis:

- Data cleansing
- Data editing
- Data reduction
- Data wrangling
- Removing missing values
- Data normalization

## 4.3 Data Labelling

In our research methodology, it is proposed to implement data labelling steps by industry experts to allow their knowledge to be utilized efficiently. Furthermore, a correct understanding of the data can be attained by incorporating the views of experts and professionals.

Health and medical insurance are examples in which the mentioned step can be employed extensively because many dangerous diseases, such as the global coronavirus epidemic, lack a history that can be used for data analysis or decision-making. This is occurring at a time when many insurers provide insurance for medical services associated with these disorders (even in the case of the new coronavirus disease, we saw the introduction of insurance for medical services related to this disease). Consequently, this strategy is required for making defensible conclusions. At the end of this stage, the experts classify insured individuals into extremely high-risk, high-risk, moderate-risk, low-risk, and extremely low-risk groups.

## 4.4 Identifying the Optimal Machine Learning Algorithm

After data collection, preprocessing, and labelling, the first step of the research is complete, and the second begins. The first and critical stage of this step is to extract knowledge from the labelled data using data mining methods and train our machine learning algorithms. Then, we test our classification algorithms to forecast the category and class of each insured at this point. As previously discussed, classification algorithms are one sort of supervised algorithm. Because insurance industry data are uneven, employing hybrid learning algorithms at this stage is advisable.

After employing multiple machine learning algorithms in our labelled dataset, one of the most fundamental questions that may arise is which method produces the best results with the given

data. This stage focuses on five criteria used to assess machine learning algorithms. The following criteria will be used to evaluate the performance of different algorithms, considering the scope of the application and the nature of the insurance data:

- Accuracy
- Precision
- Recall
- F1-score
- Matthews Correlation Coefficient (MCC)

We calculate the relative weight of each criterion in the subsequent step, which will identify the significance of each of the enumerated evaluation criteria. After calculating the required weights, there are two main approaches for obtaining the best algorithm:

- Only the criterion with the highest weight is considered, and the algorithm that performed best based on it is selected as the best-chosen algorithm.
- Selecting the optimal algorithm based on a relationship of utility maximization, considering all criteria and the relation (4-1). The $w_s$ in equation (4-1) indicate the weights derived for each criterion. Moreover, the index "a" means different algorithms that have been reviewed.

$$\text{Max}\{(w_{\text{Accuracy}}\text{Accuracy}_a) + (w_{\text{Precision}}\text{Precision}_a) + (w_{\text{Recall}}\text{Recall}_a) + (w_{\text{F1 score}}\text{F1 score}_a) + (w_{\text{MCC}}\text{MCC}_a)\} \quad \forall a \quad (4\text{-}1)$$

Many researchers consider one of the criteria to be the primary criterion (without employing multi-criteria decision-making methods) and select the optimal algorithm based on this criterion. However, it is recommended to employ multi-criteria decision-making to evaluate all criteria as it evaluates many parts of an algorithm according to their significance, efficacy, and participation in decision-making.

### 4.5 Insureds Classification

After the optimal machine learning algorithm is identified, it could be used to develop the ability to estimate the risk level each new insured will impose on the insurer. As previously mentioned, in the initial phase of data labelling for a given data set, the experts classified the insured individuals into extremely high-risk, high-risk, moderate-risk, low-risk, and extremely low-risk groups. Using the labelled data set for training, the optimal machine learning algorithm can now forecast the level of risk for each new insured.

### 4.6 Indemnity Prediction

To compute the insurance premium at the final stage, it is initially essential to forecast the insured's expected indemnity ($E\{S\}$). Since we have categorized the insureds according to their risk to the insurer, the mathematical expectation of their indemnity is derived independently and solely for each class ($E\{S_i\}$). This results in an insurance premium that is particular to each class. After selecting the appropriate ML model and classifying insureds based on training and testing data, this objective can be achieved.

## 4.7 Premium Calculation

Having estimated the mathematical expectation of a potential loss for each group of insureds in the previous step, the appropriate insurance premium for each insured group is now calculated using a linear mathematical model considering the insurer's diverse demands. As indicated previously, the insurer must consider two factors when calculating the insurance premium:

A) **Profitability**: Although government interventions and regulations are highly in place to ensure that the industry operates fairly and responsibly and to protect the interests of policyholders, the insurance industry, like every other industry, works in its specialized economic activity to maximize profit. Profit is the difference between expenses and revenues. When it comes to insurance firms, expenses are the damages paid to policyholders. On the other hand, insurance income is the total premiums insurers receive—the more significant the gap between these two parameters, the greater the insurer's profitability. Insurer's profitability, the objective function in our study, is carefully modelled and constrained to consider government regulations, ensuring that profitability is achieved while still meeting regulation requirements.

B) **Competitiveness**: Even though the insurance industry is an oligopoly with a few dominant players, the market has become more competitive with the emergence of new insurers. If a competitor offers a lower insurance premium than an insured individual's current insurer, the individual can terminate their contract and switch to a new insurer. This can ultimately affect the number of insured individuals and the insurer's profitability. Additionally, it could harm the insurer's reputation and lead to contract cancellations. To ensure the insurer remains competitive, we address this issue as a constraint in our mathematical model.

In the provided model, we ensured the profitability objective of the insurer by using the fundamentals introduced in chapter 3. This is represented in relation (4-2). In the following functions, $c_i$ indicates the insurance premium for the $ith$ group of insureds. Also, in these relationships, $x_i$, $E\{S_i\}$, and $E\{N_i\}$ represent the number of insureds in the ith category, the mathematical expectation of the possible loss of the insureds in the ith category, and the mathematical expectation of the possible number of insureds in the ith category who suffer indemnities.

$$\text{Max } Z = \sum_{i=1}^{I} (c_i x_i - E\{S_i\} \cdot E\{N_i\}) \tag{4-2}$$

In order to guarantee the insurer's competitiveness, relation (4-3) stipulates that each category's premium must be lower than the average premium disclosed by the insurer's rivals ($c_R$). This is one of the most fundamental requirements for an insurer to thrive in today's market.

$$c_i \leq c_R \qquad \forall i \tag{4-3}$$

In (4-4), it is noted that the raised insurance premium must not exceed the amount of the return rate. This principle is based on the idea that insurance premiums should not increase at a rate higher than the expected return on investment. If the premiums were to increase disproportionately to the return rate, it could lead to excessive premium inflation, making insurance unaffordable or undesirable for policyholders. Therefore, in order to guarantee the insurer's competitiveness, the premium increase should be limited to the return rate. In this relationship, $c_i'$ denotes the insurance premium announced for the $ith$ category in the previous year, whereas $\varphi$ represents the return rate.

$$c_i \leq (1+\varphi) \cdot c_i' \qquad \forall i \qquad (4\text{-}4)$$

As stated in (4-5), in order to guarantee the insurer's profitability, the insurer's income is more or at least equal to the sum of the overhead costs (α) and the cost of the insurer's loss.

$$\sum_{i=1}^{I} c_i x_i \geq \alpha + \sum_{i=1}^{I} E\{S_i\} \cdot E\{N_i\} \qquad (4\text{-}5)$$

Logically, the insurance premium received for higher-risk groups is higher than those for other groups. The relation (4-6) ensures this importance will be satisfied.

$$c_{i+1} \geq c_i \qquad \forall i \qquad (4\text{-}6)$$

Finally, the following constraints have been added to the model to ensure that the final prices and service costs are within government regulations and non-negative.

$$L \leq c_i \leq U \qquad \forall i \qquad (4\text{-}7)$$

$$c_i \geq 0 \qquad \forall i \qquad (4\text{-}8)$$

# Chapter 5: Case Study

In this chapter, an Iranian health insurance firm was studied as a case study to illustrate and verify the methodology described in the preceding chapter. The insurance company we looked into is one of Iran's most significant private insurance companies, with more than 1000 agencies nationwide to provide people with insurance services. For privacy reasons, we will not disclose the name of the company. In the subsequent sections, each step of the solution methodology is implemented in our case study.

## 5.1 Data Gathering, Preprocessing, and Labelling

The first dataset used for this study (dataset 1) was gathered from one of the hospitals in Tehran, Iran. This dataset included the information of about 2000 patients and was used for training the machine learning algorithms and selecting the optimal one in the following steps. Further details regarding its features are provided in table 3.

**Table 3 Data description of dataset 1 and dataset 2**

| Feature | Type |
|---|---|
| ID | Numerical |
| Age | Numerical |
| Sex | Boolean |
| Pregnancy status | Boolean |
| Heart disease | Boolean |
| Blood pressure disease | Boolean |

| | |
|---|---|
| Kidney disease | Boolean |
| Diabetes | Boolean |
| Liver disease | Boolean |
| Lung disease | Boolean |
| Immune Deficiency disease | Boolean |

Dataset 1 was edited and cleaned before analysis and result acquisition. Microsoft Power BI and Microsoft Excel were used extensively in this study for this purpose. A total of 1094 records from dataset 1 were deemed acceptable and labelled by a team of experts in Iran, including two general practitioners and two insurance managers. This team of experienced experts was chosen for their respective fields of expertise. Their opinions were gathered after they all arrived at the same conclusion. At the end of this stage, the experts classified insured individuals into extremely high-risk, high-risk, moderate-risk, low-risk, and extremely low-risk groups.

It is important to note that even though dataset 1 was cross-sectional, meaning it was gathered from a diverse group of subjects at a specific moment, it still provided enough information for the experts to determine the labels accurately.

Another dataset (dataset 2) was collected by the health insurer under study at this stage. Dataset 2 included the medical records of 103 potential new insureds and was used for implementing our solution methodology in the next steps. The features of dataset 2 were the same as dataset 1 (refer to table 3).

Finally, the team of experts' opinions about the weights of performance measures (necessary for selecting the optimal ML algorithm), the possible indemnity of each insured in each

classification group (necessary for premium calculation), and the possible percentage of disrupted insureds in each classification group (necessary for premium calculation) were collected. This information is summarized in tables 4, 5 and 6, respectively. Their opinions were gathered after they arrived at the same conclusion.

**Table 4 Experts' opinion about the weights of performance measures**

| Performance measure | Weight |
|---|---|
| Accuracy | $w_{\text{Accuracy}} = 0.1968$ |
| Precision | $w_{\text{Precision}} = 0.1591$ |
| Recall | $w_{\text{Recall}} = 0.2032$ |
| F1-score | $w_{\text{F1 score}} = 0.2334$ |
| MCC | $w_{\text{MCC}} = 0.2075$ |

**Table 5 Experts' opinion about the possible indemnity of each insured in each classification group**

| Classification group | Possible indemnity of each insured |
|---|---|
| 1 | 30,000,000 IRR |
| 2 | 65,000,000 IRR |
| 3 | 200,000,000 IRR |
| 4 | 550,000,000 IRR |
| 5 | 780,000,000 IRR |

**Table 6 Experts' opinion about the percentage of disrupted insureds in each classification group**

| Classification group | Percentage of disrupted insureds |
|---|---|
| 1 | 33% |
| 2 | 20% |
| 3 | 20% |
| 4 | 40% |
| 5 | 50% |

## 5.2 Identifying the Optimal Machine Learning Algorithm and Insureds Classification

Following the collection, preprocessing, and labelling of dataset 1, five machine learning algorithms (random forest, extra trees, gradient boosting, cat boost, and decision tree) were chosen to be trained with it. We attempted to use more hybrid algorithms, including proven and novel ones, to increase the method's reliability. These algorithms were selected based on their ability to classify data using different techniques, their effectiveness in handling various data types, and their widespread usage in machine-learning libraries. For instance, the decision tree algorithm, as described by Breiman et al. (1984), is a self-sufficient model that generates a tree-like structure to enable accurate predictions or classifications. On the other hand, the rest of the algorithms we opted for are ensemble methods that leverage multiple models to enhance accuracy. Their effectiveness, robustness, and widespread adoption in academia and industry make them trusted and proven choices for various machine-learning tasks, including risk group classification (Fernández-Delgado et al., 2014).

At this stage, 80% of labelled dataset 1 was utilized for training purposes, while 20% was reserved for testing. Our selected machine learning algorithms were trained with carefully selected hyperparameters. The choice of these hyperparameters was informed by an extensive review and analysis of previous literature, where they have been established to yield optimal performance in similar classification tasks. For this purpose, we utilized works such as "Sequential model-based optimization for general algorithm configuration" (Hutter et al., 2011), "Random search for hyper-parameter optimization" (Bergstra & Bengio, 2012), "Algorithms for hyper-parameter optimization" (Bergstra et al., 2011), "An Introduction to Statistical Learning" (James et al., 2013), and "Scikit-learn: Machine learning in Python" (Pedregosa et al., 2011). By leveraging the collective knowledge and expertise present in the

field, we aimed to build upon the established best practices and ensure the robustness and effectiveness of our models.

After training was completed, the labelling predictions of the five algorithms were tested and compared with experts' opinions, resulting in performance measures that are outlined in table 7.

**Table 7 The performance of each investigated model**

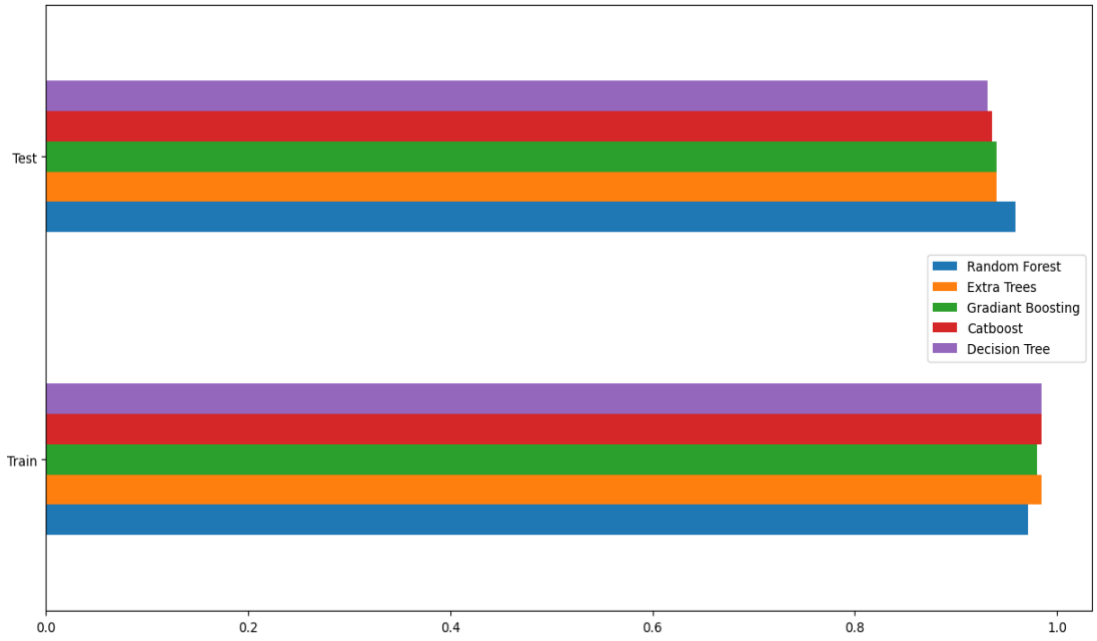
| Algorithms | Performance Measure | | | | |
|---|---|---|---|---|---|
| | **Accuracy** | **Precision** | **Recall** | **F1-score** | **MCC** |
| Random Forest | 0.9589 | 0.9633 | 0.9589 | 0.9603 | 0.9169 |
| Extra Trees | 0.9406 | 0.9577 | 0.9406 | 0.9460 | 0.8864 |
| Gradient Boosting | 0.9406 | 0.9577 | 0.9406 | 0.9460 | 0.8864 |
| Cat Boost | 0.9361 | 0.9529 | 0.9361 | 0.9414 | 0.8766 |
| Decision Tree | 0.9315 | 0.9548 | 0.9315 | 0.9390 | 0.8697 |

Before proceeding forward, it was necessary to compare each algorithm's performance in the training and testing processes based on the accuracy criterion to ensure that these algorithms are not over-fitted or under-fitted.

According to Hastie et al. (2009), overfitting occurs when a machine learning model performs well on the training data but does not perform well on new, unseen data. This means that an overfit model may have a high accuracy on the training data but a significantly lower accuracy on the testing or validation data. On the other hand, underfitting happens when a machine-learning model is too simple to capture the underlying patterns in the training data. As a result, an underfit model may show low accuracy both on the training data and the testing or validation data.

Figure 2 demonstrates that none of the machine learning algorithms used in the training and testing phases are over-fitted or under-fitted since their accuracy is acceptable and almost the same throughout both phases.

**Figure 2 Comparing the accuracy of used algorithms by considering both the training and testing data**




Now, for selecting the optimal ML algorithm, we used the experts' opinions about the weights of performance measures (refer to table 4) and determined the ideal algorithm based on the relationship between the parameters (refer to relation 4-1).

As stated in relation 4-1, the values of $W_a$ for each algorithm are determined based on the evaluation criteria for each algorithm provided in Table 7 and the weights for the performance measures specified by experts in Table 5. The $W_a$ values are shown in Table 8. This table demonstrates that the random forest was the optimal machine-learning algorithm in terms of performance.

**Table 8 Machine learning algorithms ranking in terms of performance**

| Algorithms | $W_a$ | Rank |
|---|---|---|
| Random Forest | 0.9512 | 1 |
| Extra Trees | 0.9369 | 2 |
| Gradient Boosting | 0.9369 | 2 |
| Cat Boost | 0.9276 | 4 |
| Decision Tree | 0.9242 | 5 |

Now that the optimal machine learning algorithm was identified (Random Forest), the new potential insureds in dataset 2 were classified using the Random Forest algorithm and each new insured's risk factor was determined. A summary of the results is shown in Table 9.

**Table 9 Results of potential insureds classification**

| Classification group (risk factor) | Insureds number |
|---|---|
| 1 | 9 |
| 2 | 50 |
| 3 | 34 |
| 4 | 8 |
| 5 | 2 |

## 5.3 Indemnity Prediction and Premium Calculation

Finally, using the experts' opinions about the possible indemnity of each insured in each classification group and the possible percentage of disrupted insureds in each classification group (refer to tables 5 and 6), we could calculate the appropriate insurance premiums for each group using the mathematical model in relations 4-2 to 4-8.

All the input data used in the mathematical model is shown in Table 10.

Here, it is necessary to note that because our datasets were cross-sectional in nature due to challenges in accessing healthcare data, we relied on experts' opinions to calculate $E\{S_i\}$ and $E\{N_i\}$ rather than computing them directly using the predictive power of ML algorithms.

**Table 10 Input data for the mathematical model**

| Parameter | Notations | Value | | | | |
|---|---|---|---|---|---|---|
| | | $i = 1$ | $i = 2$ | $i = 3$ | $i = 4$ | $i = 5$ |
| Number of insureds in the ith category | $x_i$ | 9 | 50 | 34 | 8 | 2 |
| The mathematical expectation of the possible indemnity of each insured in the ith category | $E\{S_i\}$ | 30,000,000 IRR | 65,000,000 IRR | 200,000,000 IRR | 550,000,000 IRR | 780,000,000 IRR |
| The mathematical expectation of the possible number of disrupted insureds in the ith category | $E\{N_i\}$ | (0.33) (9) = 3 | (0.2) (50) = 10 | (0.2) (34) = 7 | (0.4) (8) = 3 | (0.5) (2) = 1 |
| Insurance premium announced for the $ith$ category in the previous year | $c_i'$ | 20,000,000 IRR | 45,000,000 IRR | 120,000,000 IRR | 120,000,000 IRR | 850,000,000 IRR |

| | | |
|---|---|---|
| Average premium disclosed by the insurer's rivals | $c_R$ | 480,000,000 IRR |
| Return rate | $\varphi$ | 0.3 |
| Overhead costs | $\alpha$ | 350,000,000 IRR |
| Acceptable premium range | $(L, U)$ | (20,000,000 , 5,000,000,000) IRR |

As shown in Table 11, using GAMS Software and the introduced model in relations 4-2 to 4-8., $c_i$s indicating the insurance premium for the $ith$ group of insureds were obtained.

**Table 11 Calculated results**

| Parameter | Optimal value |
|---|---|
| $c_1$ | 26,000,000 IRR |
| $c_2$ | 58,500,000 IRR |
| $c_3$ | 156,000,000 IRR |
| $c_4$ | 156,000,000 IRR |
| $c_5$ | 480,000,000 IRR |
| $Z$ | 6,101,000,000 IRR |

# Chapter 6: Conclusion, Managerial Implications, and Limitations

This research presented a novel, data-driven optimization methodology for determining insurance premiums and validated it using an Iranian health insurance company as a case study. Using the optimal classification machine learning model, the insureds were divided into five groups based on their readiness to cause disruptions. Next, the indemnity of each class was estimated. Lastly, in order to determine the premium for each class depending on its indemnification, a novel mathematical model was used. The optimal machine learning algorithm was obtained by considering five performance metrics as criteria, including accuracy, precision, recall, F1 score, and MCC.

The suggested technique helps health insurance managers handle the two frequently occurring and significant disruptions in the insurance sector: errors in predicting possible future losses and a decrease in the insurer's market share. Furthermore, this research aimed to calculate the optimal premium for each prospective insured class as a sensible and possible initial point.

There are certain limits on the case study used for validation: First, there were only 1094 acceptable records in the initial dataset; a more extensive dataset would undoubtedly improve the performance of the algorithms. Additionally, as previously indicated, our datasets were limited to cross-sectional data due to difficulties in obtaining healthcare data. If longitudinal data were available, $E\{S_i\}$ and $E\{N_i\}$ could have been directly calculated instead of relying on our team of experts' opinions. Lastly, we had access to a team of four experts; the greater the number of specialists, the more precise the findings.

It is strongly suggested that future research address the following directions: Initially, it should exploit the results of this study for any other non-life insurance to validate it further. Secondly, the behaviour of insureds may be analyzed using big data analysis and longitudinal data, and the findings could be compared to this paper. Finally, in future studies, it may be helpful to utilize more advanced MADM approaches like AHP or Bayesian Best-Worst Method (BBWM) in determining the weight of criteria to use when selecting the best ML algorithm.

In conclusion, we believe that this new data-driven methodology will eliminate errors in predicting indemnities for any non-life insurance. Also, this method can fortify any insurance company against unprecedented and unforeseen obligations.

## References


Alipour-Vaezi, M., Aghsami, A., & Rabbani, M. (2022). Introducing a novel revenue-sharing contract in media supply chain management using data mining and multi-criteria decision-making methods. Soft Computing, 26(6), 2883-2900.

Ayuso, M. and Santolino, M., (2012). Forecasting the maximum compensation offer in the automobile BI claims negotiation process. Group Decision and Negotiation, 21(5), 663-676.

Bergstra, J., Bardenet, R., Bengio, Y., & Kégl, B. (2011). Algorithms for hyper-parameter optimization. Advances in Neural Information Processing Systems, 2546-2554.

Bergstra, J., & Bengio, Y. (2012). Random search for hyper-parameter optimization. Journal of Machine Learning Research, 13(Feb), 281-305.

Breiman, L., Friedman, J., Olshen, R., & Stone, C. (1984). Classification and Regression Trees. CRC Press.

Ceballos, F. and Kramer, B., (2019). From index to indemnity insurance using digital technology: Demand for picture-based crop insurance.

David, M. (2015). Auto insurance premium calculation using generalized linear models. Procedia Economics and Finance, 20, 147-156.

Derrig, R.A., (2002). Insurance fraud. Journal of Risk and Insurance, 69(3), 271-287.

Dora, P., & Sekharan, G. H. (2015). Healthcare insurance fraud detection leveraging big data analytics. IJSR, 4(4), 2073-2076.

Douven, R., Van der Heijden, R., McGuire, T. and Schut, F., (2020). Premium levels and demand response in health insurance: relative thinking and zero-price effects. Journal of Economic Behavior & Organization, 180, 903-923.

Dutta, K., Chandra, S., Gourisaria, M. K., & Harshvardhan, G. (2021). A Data Mining based Target Regression-Oriented Approach to Modelling of Health Insurance Claims. Paper presented at the 2021 5th International Conference on Computing Methodologies and Communication (ICCMC)

Fang, K., Jiang, Y., & Song, M. (2016). Customer profitability forecasting using Big Data analytics: A case study of the insurance industry. Computers & Industrial Engineering, 101, 554-564.

Fernández-Delgado, M., Cernadas, E., Barro, S., & Amorim, D. (2014). Do we need hundreds of classifiers to solve real-world classification problems? Journal of Machine Learning Research, 15(1), 3133-3181.

FinancialTribune. (2017). Iran's Insurance Premium Income Tops $2b. FinancialTribune.com, Vol. 2022.

Gans, J. (2016). The disruption dilemma. MIT press.

Gumus, F., & Uzekmek, F. (2019). An application on risk and premium calculation of fire insurance.

Halbersma, R.S., Mikkers, M.C., Motchenkova, E. and Seinen, I., (2011). Market structure and hospital–insurer bargaining in the Netherlands. The European Journal of Health Economics, 12(6), 589-603.

Hastie, T., Tibshirani, R., & Friedman, J. (2009). The Elements of Statistical Learning: Data Mining, Inference, and Prediction. New York, NY: Springer.

Huang, Y., & Meng, S. (2019). Automobile insurance classification ratemaking based on telematics driving data. Decision support systems, *127*, 113156.

Hutter, F., Hoos, H. H., & Leyton-Brown, K. (2011). Sequential model-based optimization for general algorithm configuration. Proceedings of the 5th International Conference on Learning and Intelligent Optimization, Springer, 507-523.

Ij, H. (2018). Statistics versus machine learning. Nat Methods*,* 15(4), 233.

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). An Introduction to Statistical Learning. Springer.

Krasheninnikova, E., García, J., Maestre, R., & Fernández, F. (2019). Reinforcement learning for pricing strategy optimization in the insurance industry. Engineering applications of artificial intelligence, 80, 8-19.

Lin, W., Wu, Z., Lin, L., Wen, A., & Li, J. (2017). An ensemble random forest algorithm for insurance big data analysis. Ieee Access*, 5*, 16568-16575.

Liu, Y., Li, X., Wang, D. and Cui, L., (2017). The bounds of premium and a fuzzy insurance model under risk aversion utility preference. 13th International Conference on Natural Computation, Fuzzy Systems and Knowledge Discovery (ICNC-FSKD), 1357-1362.

Mahesh, B. (2019). Machine Learning Algorithms - A Review. doi:10.21275/ART20203995.

Nardon, M., & Pianca, P. (2019). Insurance premium calculation under continuous cumulative prospect theory. University Ca'Foscari of Venice, Dept. of Economics Research Paper Series No, 3.

Pai, J., Boyd, M., & Porth, L. (2015). Insurance premium calculation using credibility analysis: an example from livestock mortality insurance. Journal of Risk and Insurance, 82(2), 341-357.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., ... & Vanderplas, J. (2011). Scikit-learn: Machine learning in Python. Journal of Machine Learning Research*, 12*(Oct), 2825-2830.

Petneházi, G., (2021). Quantile convolutional neural networks for Value at Risk forecasting. Machine Learning with Applications, 6, 100096.

Ranjbari, M., Shams Esfandabadi, Z. & Scagnelli, S.D. (2019). Sharing economy risks: Opportunities or Threats for insurance companies? A Case study on the Iranian insurance industry. The Future of Risk Management, Volume II, 343-360.

Rotar, V. I. (2014). Actuarial models: the mathematics of insurance: CRC Press.

Settipalli, L. and Gangadharan, G.R., (2021). Provider profiling and labeling of fraudulent health insurance claims using Weighted MultiTree. Journal of Ambient Intelligence and Humanized Computing, 1-22.

Torraca, A. P., & Fanzeres, B. (2021). Optimal insurance contract specification in the upstream sector of the oil and gas industry. European journal of operational research, 295(2), 718-732.

Wang, Y., & Xu, W. (2018). Leveraging deep learning with LDA-based text analytics to detect automobile insurance fraud. Decision support systems, 105, 87-95.

Yang, Z., Liu, Y., Campbell, M., Ten, C. W., Rho, Y., Wang, L., & Wei, W. (2020). Premium calculation for insurance businesses based on cyber risks in IP-based power substations. IEEE Access, 8, 78890-78900.

Yeo, A. C., & Smith, K. A. (2006). Implementing a data mining solution for an automobile insurance company: Reconciling theoretical benefits with practical considerations. Cases on Database Technologies and Applications, 189-201.

Zhang, Y., Wu, Y. and Yao, H. (2021). Optimal Insurance Indemnity and Reinsurance Strategy for Health Insurance. Mathematical Problems in Engineering.

Zhu, M., Anwar, A.H., Wan, Z., Cho, J.-H., Kamhoua, C.A. & Singh, M.P. (2021). A survey of defensive deception: Approaches using game theory and machine learning. IEEE Communications Surveys & Tutorials, 23(4), 2460-2493.

# Annexes

## Annex A: Code Development

### A.1 Python Codes:

**Importing the libraries**

```
import numpy as np
import pandas as pd
import xlsxwriter
from sklearn import tree
from sklearn.ensemble import RandomForestClassifier
GradientBoostingClassifier, ExtraTreesClassifier
from sklearn.metrics import accuracy_score,
classification_report, matthews_corrcoef, recall_score
from sklearn.model_selection import train_test_split,
GridSearchCV
from sklearn.tree import DecisionTreeClassifier
from catboost import CatBoostClassifier
```

**Loading dataset 1**

```
df=pd.read_csv("Dataset.csv")
df=df.drop('ID',axis=1)
df.head(3)

X=df.iloc[:,:-1]
y=df.iloc[:,-1]
```

**Train/Test split**

```
X_train, X_test, y_train, y_test = train_test_split(X,y,
test_size=0.2)
```

**Defining evaluation function**

```
def evaluate(model, X_train, X_test, y_train, y_test):
    y_test_pred = model.predict(X_test)
    y_train_pred = model.predict(X_train)

    print("TRAINING RESULTS:
\n==============================")
```

```
    clf_report =
pd.DataFrame(classification_report(y_train, y_train_pred,
output_dict=True))
    print(f"ACCURACY SCORE:\n{accuracy_score(y_train,
y_train_pred):.4f}")
    print(f"MCC SCORE:\n{matthews_corrcoef(y_train,
y_train_pred):.4f}")
    print(f"CLASSIFICATION REPORT:\n{clf_report}")

    print("TESTING RESULTS:
\n============================")
    clf_report = pd.DataFrame(classification_report(y_test,
y_test_pred, output_dict=True))
    print(f"ACCURACY SCORE:\n{accuracy_score(y_test,
y_test_pred):.4f}")
    print(f"MCC SCORE:\n{matthews_corrcoef(y_test,
y_test_pred):.4f}")
    print(f"CLASSIFICATION REPORT:\n{clf_report}")
```

**Random Forest**

**# Defining parameters range**

```
param_grid = {'n_estimators': [int(x) for x in
np.linspace(start = 10, stop = 100, num = 10)],
              'max_features': ['auto', 'sqrt'],
              'max_depth': [2, 3, 4, 5, 6, 7, 8, 9, 10],
              'min_samples_split': [2, 3, 4, 5, 6, 7, 8,
9, 10],
              'min_samples_leaf': [1, 2, 3, 4, 5, 6, 7, 8,
9, 10],
              'bootstrap': [True, False]}
```

**# Training**

```
rf = RandomForestClassifier()
rf_Grid = GridSearchCV(estimator = rf, param_grid =
param_grid, cv = 5, verbose = 2, n_jobs = 4)
rf_Grid.fit(X_train, y_train)
```

**# Random Forest best parameters with gridCV**

```
rf_Grid.best_params_
```

**# Evaluation**

```
evaluate(rf_Grid , X_train, X_test, y_train, y_test)
```

**Extra Trees**

**# Training**

```
ex_tree_clf = ExtraTreesClassifier(n_estimators=1000,
max_features=7, random_state=42)
ex_tree_clf.fit(X_train,y_train)
```

**# Evaluation**

```
evaluate(ex_tree_clf, X_train, X_test, y_train, y_test)
```

**Stochastic Gradient Boosting**

**# Training**

```
grad_boost_clf =
GradientBoostingClassifier(n_estimators=100,
random_state=42)
grad_boost_clf.fit(X_train, y_train)
```

**# Evaluation**

```
evaluate(grad_boost_clf, X_train, X_test, y_train, y_test)
```

**Cat Boost**

**# Training**

```
catboost = CatBoostClassifier()
catboost.fit(X_train, y_train)
```

**# Evaluation**

```
evaluate(catboost, X_train, X_test, y_train, y_test)
```

**Decision Tree Classifier**

**# Training**

```
dtc=DecisionTreeClassifier(criterion='entropy' ,
random_state=0)
dtc.fit(X_train,y_train)
```

**# Evaluation**

```
evaluate(dtc, X_train, X_test, y_train, y_test)
```

**Comparing the accuracy of algorithms by considering both the training and testing data**

```
scores = {
    'Random Forest': {
        'Train': accuracy_score(y_train,
rf_Grid.predict(X_train)),
        'Test': accuracy_score(y_test,
rf_Grid.predict(X_test)),
    }
```

```
}
scores['Extra Trees'] = {
        'Train': accuracy_score(y_train,
ex_tree_clf.predict(X_train)),
        'Test': accuracy_score(y_test,
ex_tree_clf.predict(X_test)),
    }
scores['Gradiant Boosting'] = {
        'Train': accuracy_score(y_train,
grad_boost_clf.predict(X_train)),
        'Test': accuracy_score(y_test,
grad_boost_clf.predict(X_test)),
    }
scores['Catboost'] = {
        'Train': accuracy_score(y_train,
catboost.predict(X_train)),
        'Test': accuracy_score(y_test,
catboost.predict(X_test)),
    }
scores['Decision Tree'] = {
        'Train': accuracy_score(y_train,
dtc.predict(X_train)),
        'Test': accuracy_score(y_test,
dtc.predict(X_test)),
    }

scores_df = pd.DataFrame(scores)

scores_df.plot(kind='barh', figsize=(15, 8))
```

**Loading dataset 2**

```
df2 = pd.read_csv("Case Study.csv")
df2.head(3)

label = rf_Grid.predict(df2)
```

**Adding label to dataset 2**

```
d2 = df2.insert(loc=0, column ='Label', value = label)
d2 = df2
df2

df_csv = df2.to_csv('df_csv', index=True)

with open('df_csv', 'r') as file:
    csv_string = file.read()
```

```
print('\nCSV String:\n', csv_string)
writer = pd.ExcelWriter('Case Study Results.xlsx',
engine='xlsxwriter')

df2.to_excel(writer, sheet_name='Sheet1')

writer.save()
```

## A.2 GAMS Codes

```
sets
i index for insureds' categories /1*5/

parameters
x(i) Number of insureds in the ith category /1 9, 2 50, 3
34, 4 8, 5 2/
ES(i) Mathematical expectation of the possible indemnity of
each insured in the ith category /1 3000000, 2 6500000, 3
20000000, 4 55000000, 5 78000000/
EN(i) Mathematical expectation of the possible number of
disrupted insureds in the ith category /1 3, 2 10, 3 7, 4
3, 5 1/
c_prime(i) Insurance premium announced for the ith category
in the previous year /1 20000000, 2 45000000, 3 120000000,
4 120000000, 5 850000000/;

scalar
c_R Average premium disclosed by the insurer's rivals
/480000000/
phi Return rate /0.3/
alpha Total overhead costs /350000000/
L Lower bound of insurance premiums /20000000/
U Upper bound of insurance premiums /5000000000/;

variables
Z

positive variable c(i);

Equations
obj

co1(i)
co2(i)
co3
co4(i)
```

```
co5(i);

obj .. Z =e= sum(i,(c(i)*x(i))-(ES(i)*EN(i)));

co1(i) .. c(i) =l= c_R;
co2(i) .. c(i) =l= (1 + phi) * c_prime(i);
co3 .. sum(i,c(i)*x(i)) =g= alpha + sum(i,ES(i)*EN(i));
co4(i) .. c(i) =g= L;
co5(i) .. c(i) =l= U;


model Thesis /all/
option optca = 0, optcr = 0, lp = cplex, reslim = 3600;

solve thesis using Lp max Z;

display Z.l, c.l;

execute_unload 'Final Results.gdx';
```