Available online at www.ijournalse.org

Emerging Science Journal

(ISSN: 2610-9182)

Vol. 9, No. 5, October, 2025

The Occupational Indemnity Insurance Modelling: Brighton Mahohoho XGBoost Probabilistic Automated Actuarial Reserving-Pricing-Underwriting

Brighton Mahohoho [1*], Charles Chimedza [2], Florance Matarise [1], Sheunesu Munyira [1]

[1] *Department of Mathematics & Computational Sciences, University of Zimbabwe, Harare, Zimbabwe.*

[2] *School of Statistics & Actuarial Science, University of Witwatersrand, Johanesburg, South Africa.*

Abstract

This paper introduces the IFRS 17-Compliant Brighton Mahohoho Probabilistic Framework for Inflation-Adjusted Frequency-Severity Modeling in Occupational Indemnity Insurance, integrating AI-driven actuarial methodologies for loss reserving, risk pricing, and underwriting. Objectives: The framework ensures IFRS 17 compliance while enhancing actuarial accuracy and operational efficiency. Methods/Analysis: A simulation-based dataset of policy, claims, premiums, inflation adjustments, and underwriting data is generated. Claim frequencies and severities are modeled using Poisson and Gamma distributions, with inflation adjustments incorporated into reserves. XGBoost is applied for Automated Actuarial Loss Reserving (ALR) and Automated Actuarial Risk Pricing (ARP), while a weighted average approach estimates Automated Actuarial Loss Reserve Risk Premiums (AALRRP). Findings: Model accuracy is validated through MAE, MSE, RMSE, residual analysis, and scatter plots. IFRS 17 metrics—Contractual Service Margin (CSM), Fulfillment Cash Flows (FCF), Risk Adjustments, and Liabilities—are simulated, with sensitivity analysis ensuring robustness. Policyholders are segmented into underwriting clusters, incorporating expenses, outgo, and revenue to derive the Automated Net Actuarial Underwriting Balance (ANAUB). Novelty/Improvement: This integrated AI-driven actuarial framework significantly advances IFRS 17-compliant pricing and reserving, offering enhanced predictive accuracy, regulatory alignment, and improved risk assessment in occupational indemnity insurance.

Keywords:

Loss Reserves;
Risk Premiums;
Actuarial Underwriting;
IFRS 17;
XGBoost Regression;
Occupational Indemnity Insurance.

Article History:

Received:	03	January	2025
Revised:	24	September	2025
Accepted:	30	September	2025
Published:	01	October	2025

1- Introduction

The IFRS 17-Compliant *Brighton Mahohoho Probabilistic Framework for Inflation-Adjusted Frequency-Severity* Modeling in Occupational Indemnity Insurance integrates Artificial Intelligence (AI)-driven Augmented Actuarial Data Science with eXtreme Gradient Boosting (XGBoost) to enhance loss reserving, risk pricing, and underwriting. Occupational indemnity insurance protects professionals against liabilities arising from errors, omissions, or negligence in their services [1]. Given the evolving regulatory landscape under IFRS 17, traditional actuarial models face challenges in accurately estimating claims frequency and severity while adjusting for inflation [2, 3]. This study proposes an advanced AI-based approach that refines actuarial loss reserving and risk pricing by leveraging XGBoost's predictive capabilities in conjunction with probabilistic modeling techniques.

The proposed framework builds on probabilistic modeling principles, enhancing traditional frequency-severity analysis with AI-driven methodologies. Conventional frequency-severity models use generalized linear models (GLMs) and stochastic techniques, often constrained by rigid distributional assumptions [4]. In contrast, XGBoost introduces gradient-boosted decision trees, which capture complex non-linear relationships within claims data. Augmented

* **CONTACT**: brighton.mahohoho@science.uz.ac.zw

DOI: http://dx.doi.org/10.28991/ESJ-2025-09-05-030

Actuarial Data Science (AADS) synergizes domain expertise with AI to refine data preprocessing, feature engineering, and model selection [5]. By incorporating Bayesian updating and inflation-adjustment techniques, the framework ensures actuarial reserves and risk premiums adhere to IFRS 17 regulations while maintaining predictive accuracy. The actuarial profession is witnessing an increasing need for advanced predictive models that can accommodate regulatory changes and economic fluctuations [6]. Traditional frequency-severity modeling methods struggle to adapt to emerging risks, high-dimensional datasets, and evolving claims behavior under economic inflation [7]. The rationale behind this study is to integrate AI techniques with actuarial science to bridge these methodological gaps. By employing XGBoost, the study enhances predictive precision while ensuring compliance with IFRS 17 principles on risk adjustments and fulfillment cash flows [8].

The framework finds applications in various aspects of occupational indemnity insurance, including:

- *Loss Reserving*: Predicting outstanding claims liabilities with enhanced accuracy using AI-driven inflation-adjusted frequency-severity models [9].
- *Risk Pricing*: Ensuring equitable premium determination by dynamically adjusting for inflation and loss development trends [10].
- *Underwriting*: Assisting underwriters in decision-making by providing AI-enhanced probabilistic insights into future claim severities and frequencies [2].
- *Regulatory Compliance:* Aligning actuarial estimates with IFRS 17 standards through probabilistic uncertainty quantification and risk margin estimation [8].

This research contributes to the actuarial and insurance fields by: Introducing a novel IFRS 17-compliant probabilistic AI framework for loss reserving and risk pricing. Addressing limitations in traditional frequency-severity models by leveraging XGBoost's non-linearity handling and boosting efficiency [11]. Enhancing actuarial practice with Augmented Actuarial Data Science (AADS), fostering synergy between domain expertise and AI-driven predictive analytics. Providing empirical evidence on the applicability of AI in meeting IFRS 17 compliance requirements, thus informing industry best practices and regulatory adherence.

The existing literature on actuarial loss reserving and frequency-severity modeling has primarily focused on traditional statistical methods, such as chain-ladder techniques and generalized linear models [1]. While these approaches offer interpretability, they often fail to capture complex, high-dimensional interactions in claims data, particularly under inflationary conditions [3]. Recent studies have explored machine learning techniques for actuarial applications, with researchers demonstrating the effectiveness of XGBoost in predictive modeling [4]. However, few studies have integrated XGBoost with IFRS 17 compliance frameworks, leaving a gap in ensuring regulatory alignment while leveraging AI-driven actuarial methodologies. Moreover, there is limited research on the synergy between Augmented Actuarial Data Science (AADS) and probabilistic frameworks in the context of inflation-adjusted frequency-severity modeling.

To address these gaps, this study proposes a novel framework that: Integrates XGBoost with probabilistic modeling to enhance accuracy in estimating frequency-severity distributions under inflationary impacts. Develops an AI-driven actuarial pipeline that aligns with IFRS 17 requirements on risk adjustment, discounting, and contract service margins. Leverages Augmented Actuarial Data Science (AADS) to refine feature engineering, hyperparameter tuning, and model validation. Incorporates scenario-based stress testing to ensure robustness of the model in various economic conditions. Finally, provides empirical validation through real-world occupational indemnity insurance data, demonstrating practical applicability and regulatory compliance.

The proposed IFRS 17-compliant probabilistic framework advances occupational indemnity insurance modeling by integrating AI-driven augmented actuarial science with XGBoost. By addressing gaps in traditional methodologies and enhancing regulatory adherence, this study contributes to the evolution of actuarial science, offering a transformative approach to loss reserving, risk pricing, and underwriting in a rapidly evolving insurance landscape.

2- Actuarial Loss Reserving Methods

2-1- Chain Ladder Method

The Chain-Ladder Method (CLM) is a fundamental actuarial tool used to estimate loss reserves for cumulative claims data. It assumes that claims development factors remain consistent over time, enabling the projection of future claims development [12].

The CLM is based on the following assumptions:

1. Claims are cumulative and reported in a development triangle.
2. Development factors are consistent across accident years.
3. Claims within each development period are independent and identically distributed.

Let C_{ij} represent the cumulative claims for accident year i at development lag j, where $i \in \{1,2,\dots,m\}$ and $j \in \{1,2,\dots,n\}$. Table 1 illustrates the general structure.

Table 1. General Chain Ladder Table

Accident Year / Development Lag	1	2	3	…	n
1	C_{11}	C_{12}	C_{13}	…	C_{1n}
2	C_{21}	C_{22}	C_{23}	…	C_{2n}
3	C_{31}	C_{32}	C_{33}	…	C_{3n}
⋮	⋮	⋮	⋮	⋮	⋮
m	C_{m1}	C_{m2}	C_{m3}	…	C_{mn}

The fundamental assumption of the chain ladder model is that the development factors are multiplicative:

$$f_j = \frac{\sum_{i=1}^{m-j} C_{i,j+1}}{\sum_{i=1}^{m-j} C_{ij}}, \quad \text{for} j = 1, \dots, n-1.$$

The future claims are then predicted as:

$$C_{i,j+1} = C_{ij} \cdot f_j.$$

The chain ladder factors f_j are unbiased estimators of future development.

Proof. The unbiased nature of the chain ladder factors arises from the independence assumption of incremental claims. Let C_{ij} denote the cumulative claims at development period j for origin period i. The chain ladder model assumes the relationship:

$$\mathbb{E}[C_{i,j+1}|C_{ij}] = C_{ij} \cdot f_j, \tag{1}$$

where f_j is the development factor for development period j.

Using the law of total expectation, we have:

$$[C_{i,j+1}] = \mathbb{E}[\mathbb{E}[C_{i,j+1}|C_{ij}]]. \tag{2}$$

Substituting Equation 1 into Equation 2, it follows that:

$$\mathbb{E}[C_{i,j+1}] = \mathbb{E}[C_{ij} \cdot f_j] = \mathbb{E}[C_{ij}] \cdot f_j,$$

since f_j is a deterministic quantity derived from observed claims data.

Thus, the development factor f_j is an unbiased estimator because it is calculated from historical claims, ensuring that:

$$\mathbb{E}[f_j] = f_j.$$

This concludes the proof.

The cumulative claims C_{ij} are consistent estimators of total liabilities under the chain ladder model.

Proof. Consistency requires that the expected value of the estimator converges to the true liability as the number of observations increases:

$$\lim_{m\to\infty} \mathbb{E}\big[C_{ij}\big] = L_{ij}, \quad \text{where } L_{ij} \text{ represents the trueliability.} \tag{3}$$

Under the chain ladder model, cumulative claims C_{ij} are recursively defined using development factors f_j, such that:

$$C_{ij} = C_{i,j-1} \cdot f_j, \quad \text{for} j \geq 2, \text{and} C_{i1} = X_{i1}, \tag{4}$$

where X_{i1} represents the observed initial claims.

Taking expectations under the model assumptions, we have:

$$\mathbb{E}[C_{ij}] = \mathbb{E}[C_{i,j-1}] \cdot \mathbb{E}[f_j], \quad \text{with} \mathbb{E}[f_j] \approx f_j \text{as} m \to \infty. \tag{5}$$

By iterating Equation 5, it follows that:

$$\mathbb{E}[C_{ij}] = X_{i1} \cdot \prod_{k=2}^{j} f_k. \tag{6}$$

The true liability L_{ij} is given by:

$$L_{ij} = L_{i1} \cdot \prod_{k=2}^{j} f_k, \quad \text{where} L_{i1} = \mathbb{E}[X_{i1}]. \tag{7}$$

Since X_{i1} is unbiased and $\mathbb{E}[f_k] \to f_k$ as $m \to \infty$, it follows from Equations 6 and 7 that:

$$\lim_{m\to\infty} \mathbb{E}[C_{ij}] = L_{ij}. \tag{8}$$

Thus, the cumulative claims C_{ij} are consistent estimators of L_{ij}, completing the proof.

The Chain-Ladder Method remains a robust and widely used technique for loss reserving in actuarial science. Its mathematical rigor, coupled with practical implementation, ensures its enduring relevance.

2-2- The Loss Ratio Method

The Loss Ratio Method is a widely used actuarial technique for pricing insurance premiums and assessing the adequacy of reserves. This method is grounded in the assumption that there is a predictable relationship between the premiums collected by an insurer and the losses incurred over a given period. The Loss Ratio is defined as the ratio of incurred losses to earned premiums and serves as a benchmark for underwriting performance.

The core idea behind the Loss Ratio Method lies in estimating the future claims experience based on the historical relationship between premiums and losses. This method provides a simple yet effective framework for adjusting the reserves, ensuring they are adequate to cover future liabilities. The Loss Ratio is mathematically represented as:

$$\text{LossRatio} = \frac{\text{IncurredLosses}}{\text{EarnedPremiums}}$$

Let L_t denote the incurred losses at time t, and P_t the earned premiums at the same time. The Loss Ratio at time t, denoted ρ_t, is given by:

$$\rho_t = \frac{L_t}{P_t}$$

This relationship suggests that, on average, for every dollar of premium, there will be a corresponding loss proportional to the loss ratio. An important aspect of this method is its reliance on historical data to estimate future losses. Actuaries often use historical loss ratios to predict the expected future loss ratios and adjust them based on claims experience and market conditions.

If the historical loss ratio is stable over time, future loss ratios can be predicted with a high degree of accuracy.

Algorithm 1. Chain Ladder Method Algorithm

Algorithm: *Development Triangle Reserve Estimation Using Development Factors* **Description:** This algorithm projects future claims values in the development triangle using historical data and development factors, estimating the total reserve required. **Input:** Development triangle $\mathbf{C} = \{C_{ij}\}$, where C_{ij} represents the cumulative claims amount at origin period i and development period j; $i \in \{1, \dots, n\}$, $j \in \{1, \dots, n\}$. **Initialize:** Calculate development factors $\{f_j\}$ for $j \in \{1, \dots, n-1\}$ using:

$$f_j = \frac{\sum_{i=1}^{n-j} C_{i,j+1}}{\sum_{i=1}^{n-j} C_{ij}}, \quad j = 1, \dots, n-1. \tag{9}$$

The development factor f_j for each development period j is the ratio of the cumulative claims in period $j+1$ to the cumulative claims in period j, averaged across all available origin periods. **for** $j = 1$ to $n-1$ **for** $i = 1$ to $n-j$ Project the cumulative claims at the next development period $j+1$ using the following relation:

$$C_{i,j+1} = C_{ij} \cdot f_j, \quad \text{where } f_j \text{ is the development factor calculated in Equation 9.} \tag{10}$$

The claims in period $j+1$ for each origin period i are projected by multiplying the value C_{ij} by the corresponding development factor f_j. **Output:** The projected development triangle $\hat{\mathbf{C}} = \{\hat{C}_{ij}\}$, where $\hat{C}_{ij}$ includes both the observed and projected cumulative claims values for each i and j. **Compute:** The total reserve $\hat{R}$, representing the expected future claims liabilities, is given by:

$$\hat{R} = \sum_{i=1}^{n} \sum_{j=i+1}^{n} \hat{C}_{ij}, \tag{11}$$

The total reserve $\hat{R}$ is the sum of the projected cumulative claims values for all future periods ($j > i$), which accounts for both observed and projected claims data. **Output:** The total reserve $\hat{R}$ provides the estimated amount of funds needed to cover all future claims, derived from the projected development triangle.

We assume that the loss ratio follows a stable distribution over time. Specifically, we hypothesize that the loss ratio for each period, denoted as ρ_i for year i, is drawn from the same underlying distribution, implying that the mean loss ratio remains constant across periods.

To predict the future loss ratio, we estimate the expected future loss ratio $\hat{\rho}$ as the average of the historical loss ratios. This can be expressed mathematically as:

$$\hat{\rho} = \frac{1}{n}\sum_{i=1}^{n} \rho_i$$

where: $\hat{\rho}$ is the predicted future loss ratio; n is the number of years of historical data; ρ_i is the loss ratio in year i, for $i = 1, 2, \ldots, n$.

If the historical loss ratios ρ_i have remained relatively constant, i.e., $\rho_i \approx \rho_j$ for all $i, j \in \{1,2, \ldots, n\}$, then the predicted future loss ratio $\hat{\rho}$ will closely approximate the true future loss ratio. This implies that the average of the historical data provides an accurate estimate for future loss ratios, and therefore can be used as a reliable basis for future loss forecasting and premium setting.

In such cases, the difference between the predicted and true future loss ratio tends to be small, and we can justify using this simple method for estimation with a high degree of accuracy. The Loss Ratio Method involves several steps, including data collection, loss ratio calculation, and premium adjustment. We can outline these steps in an algorithmic structure as follows:

Algorithm 2. Loss Ratio Method Algorithm

1. **Input:** Historical premium and loss data for n years.
2. **Output:** Adjusted premium and reserves for future periods.
3. Let P_t be the earned premiums in year t.
4. Let L_t be the incurred losses in year t.
5. Compute Loss Ratio for each year:

$$\rho_t = \frac{L_t}{P_t} \quad (1)$$

6. **for** each year $t = 1$ to n **do**
7. Calculate the Loss Ratio for each year:

$$\rho_t = \frac{L_t}{P_t} \quad (2)$$

8. **end for**
9. Compute the average Loss Ratio $\bar{\rho}$:

$$\bar{\rho} = \frac{1}{n}\sum_{t=1}^{n} \rho_t \quad (3)$$

10. Use the average Loss Ratio $\bar{\rho}$ to estimate future losses and set future premiums.
11. Adjust reserves based on estimated future losses:

$$R_{\text{adjusted}} = \bar{\rho} \times P_{\text{future}} \quad (4)$$

The above pseudo-algorithm outlines the basic steps involved in applying the Loss Ratio Method. The method requires the collection of historical data on premiums and losses, followed by the calculation of loss ratios for each year. The average historical loss ratio is then computed and used to predict future losses and adjust premiums accordingly. Finally, reserves are adjusted to ensure they are sufficient to cover future liabilities.

Let us consider a scenario where an insurer has collected historical data for n years. The insurer wishes to adjust the reserves based on the expected future claims, given the historical loss ratios. The adjusted premium $\hat{P}_{t+1}$ for the next period can be computed using the following formula:

$$\hat{P}_{t+1} = \frac{L_{t+1}}{\hat{\rho}}$$

where $\hat{\rho}$ is the estimated future loss ratio based on historical data.

The reserve adjustment for the next period, denoted R_{t+1}, can be computed as:

$$R_{t+1} = \hat{P}_{t+1} \times \hat{\rho}$$

This ensures that the insurer has adequate reserves to cover future losses, based on the historical loss experience.

The Loss Ratio Method is a consistent and reliable method for estimating future claims when historical loss ratios are stable.

Proof. Let the historical loss ratios be denoted by the set $\{LR_1, LR_2, \ldots, LR_n\}$, where each LR_i represents the loss ratio for the i-th period. The loss ratio for a given period is defined as:

$$LR_i = \frac{\text{Claims}_i}{\text{Premiums}_i}$$

where Claims_i represents the total claims paid in the i-th period and Premiums_i represents the total premiums collected in the same period.

Assuming that the historical loss ratios remain stable over time, we can estimate the future loss ratio $\hat{L}R$ by calculating the average of the historical loss ratios:

$$\hat{L}R = \frac{1}{n}\sum_{i=1}^{n} LR_i$$

Using this estimated loss ratio, the future claims $\hat{\text{C}}\text{laims}$ for a given premium $\text{Premiums}_{\text{future}}$ can be predicted as:

$$\hat{\text{C}}\text{laims} = \hat{L}R \times \text{Premiums}_{\text{future}}$$

The Loss Ratio Method ensures that the insurer charges premiums that are appropriate to cover the expected future claims, i.e., the premiums are set such that:

$$\text{Premiums}_{\text{future}} = \frac{\hat{\text{C}}\text{laims}}{\hat{L}R}$$

Thus, if the underwriting and claims processes remain consistent over time, the Loss Ratio Method provides a reliable estimate for future claims.

The Loss Ratio Method provides a straightforward and effective approach for pricing insurance premiums and setting reserves. By leveraging historical data, the method enables insurers to predict future losses with reasonable accuracy. The Loss Ratio is a key actuarial tool, helping actuaries maintain financial stability and ensure that insurers charge appropriate premiums to cover future claims

3- Actuarial Risk Pricing Methods

3-1- Experience Rating

Experience Rating is a method used in insurance pricing to adjust premiums based on the insured's individual claim history. Unlike traditional rating methods that rely solely on generalized risk factors, experience rating uses actual past experience to assess the risk of a policyholder. This method provides a more personalized premium, typically leading to a discount for lower-than-average claims experience or a surcharge for higher-than-average claims. It is particularly relevant in non-life insurance fields such as health, travel, and workers' compensation insurance.

Experience rating is based on the principle that the probability of future claims is influenced by the past behavior of the policyholder. Mathematically, the adjustment to the premium is a function of the historical loss experience, typically measured in terms of claim frequency and severity.

Let θ represent the base premium, which is adjusted according to the following formula:

$$\theta_{\text{adjusted}} = \theta \cdot (1 + \lambda \cdot \frac{L_i}{L_0}), \quad (12)$$

where:

- θ_{adjusted} is the adjusted premium,
- θ is the base premium,
- λ is a factor determined by the insurer,
- L_i is the total claims experienced by the policyholder in the period under review,
- L_0 is the expected claim amount based on the policyholder's risk class.

The adjustment factor λ is derived through various statistical methods, including regression analysis, to ensure the adjustment reflects the policyholder's actual risk.

The process of adjusting premiums based on experience can be formalized in the following pseudo-algorithm.

Algorithm 3. Experience Rating Adjustment

1. **Input:** Base premium θ, expected claims L_0, actual claims L_i, adjustment factor λ
2. **Output:** Adjusted premium θ_{adjusted}
3. **Step 1:** Compute the claim ratio:

$$\text{ClaimRatio} \leftarrow \frac{L_i}{L_0} \quad (1)$$

4. **Step 2:** Compute the adjusted premium:

$$\theta_{\text{adjusted}} \leftarrow \theta \cdot (1 + \lambda \cdot \text{ClaimRatio}) \quad (2)$$

5. **Step 3: Return** the adjusted premium:

θ_{adjusted} as given in Equation 2

For any policyholder, the adjusted premium θ_{adjusted} is a monotonic function of the ratio of actual to expected claims $\frac{L_i}{L_0}$.

Proof. By the definition of the adjusted premium, it is given by the following expression:

$$\theta_{\text{adjusted}} = \theta \cdot \left(1 + \lambda \cdot \frac{L_i}{L_0}\right),$$

where: θ is the base premium; λ is a positive constant representing the adjustment factor, and $\frac{L_i}{L_0}$ is the ratio of actual claims L_i to expected claims L_0.

Since $\lambda > 0$ and $\frac{L_i}{L_0} \geq 0$, the term $1 + \lambda \cdot \frac{L_i}{L_0}$ is always non-decreasing. Consequently, the adjusted premium θ_{adjusted} is an increasing function of the ratio $\frac{L_i}{L_0}$.

Therefore, as the ratio of actual to expected claims increases, the adjusted premium also increases, which proves that the adjusted premium is a monotonic function of $\frac{L_i}{L_0}$.

The adjustment factor λ should be selected such that the expected adjusted premium over a large population matches the actuarial target premium.

The expected adjusted premium is given by:

$$E[\theta_{\text{adjusted}}] = \theta \cdot \left(1 + \lambda \cdot E\left[\frac{L_i}{L_0}\right]\right).$$

For the model to be actuarially fair, we set:

$$E[\theta_{\text{adjusted}}] = \theta,$$

which implies that:

$$\lambda \cdot E\left[\frac{L_i}{L_0}\right] = 0.$$

Thus, λ should be chosen based on the distribution of claim ratios across policyholders.

One key claim of experience rating is that it leads to more equitable pricing by ensuring that policyholders are charged premiums that reflect their individual risk levels. This can result in a more sustainable pricing system where low-risk policyholders benefit from lower premiums, and high-risk policyholders are incentivized to improve their risk management practices.

In practice, insurers use a variety of methods to compute λ and L_0, including Bayesian updating, historical loss data, and generalized linear models (GLMs). This flexibility allows experience rating to be adapted to a wide variety of insurance lines and risk profiles.

Experience Rating is a powerful actuarial tool that allows insurers to personalize premiums based on a policyholder's claims history. By adjusting premiums according to past claims experience, it incentivizes risk mitigation while ensuring fairness in pricing. However, the effectiveness of this approach depends on the correct estimation of expected claims and the adjustment factor, which requires advanced statistical modeling and a deep understanding of the risk characteristics of the insured pool.

3-2- Loss Distribution Approaches

The Loss Distribution Approach (LDA) is a widely used method in actuarial science, particularly in the modeling of risk and loss reserves. It provides a framework for quantifying and managing the distribution of total claims, combining information on the frequency and severity of claims. This approach plays a pivotal role in determining the appropriate reserves for future claims and in calculating actuarial risk premiums. LDA models the loss distribution as the convolution of two distributions: the frequency of claims and the severity of individual claims. The primary objective of LDA is to calculate the total loss, which is the sum of the individual losses that occur over a specified period. By combining the frequency and severity distributions, the LDA allows actuaries to estimate the total risk associated with insurance policies.

The Loss Distribution Approach assumes that losses are governed by two random variables:

• N - the number of claims occurring during a specific period, which follows a probability distribution $f_N(n)$.

• X - the severity of each individual claim, which follows a probability distribution $f_X(x)$.

The total loss L during a period is the sum of the severities of the N claims:

$L = \sum_{i=1}^{N} X_i$

where X_i are i.i.d. (independent and identically distributed) random variables representing claim severities.

The distribution of total loss L can be found by convolving the distributions of N and X:

$f_L(l) = \sum_{n=0}^{\infty} f_N(n) \cdot f_X^{*n}(l)$

Algorithm 4. Loss Distribution Calculation

Input: Frequency distribution $f_N(n)$, Severity distribution $f_X(x)$

Output: Total loss distribution $f_L(l)$ Initialize total loss distribution $f_L(l) = 0$

for each possible number of claims $n = 0,1,2,\ldots$ Calculate the n-fold convolution of the severity distribution: $f_X^{*n}(l)$ Update the total loss distribution:

$f_L(l) = f_L(l) + f_N(n) \cdot f_X^{*n}(l)$

Return $f_L(l)$

---

where $f_X^{*n}(l)$ represents the n-fold convolution of the severity distribution $f_X(x)$, and $f_L(l)$ is the probability density function (PDF) of the total loss.

To calculate the total loss distribution using LDA, the following pseudo-algorithm can be employed:

The convolution formula used to combine the frequency and severity distributions is given by:

$f_L(l) = \sum_{n=0}^{\infty} f_N(n) \cdot f_X^{*n}(l)$

The n-fold convolution of the severity distribution is defined as:

$f_X^{*n}(l) = \int_0^l f_X(l_1) \cdot f_X^{*(n-1)}(l - l_1)\, dl_1$

where $f_X^{*0}(l) = \delta(l)$ is the Dirac delta function, and $f_X^{*1}(l) = f_X(l)$.

The total loss distribution $f_L(l)$ is a mixture of severity distributions, weighted by the frequency distribution.

From the definition of convolution, the total loss distribution $f_L(l)$ can be expressed as a weighted sum of severity distributions. Specifically, for a given number of claims n, the corresponding severity distribution $f_X^{*n}(l)$ is convolved n-times. The mixture arises because the number of claims is random and follows a frequency distribution $f_N(n)$. Therefore, we can write the total loss distribution as:

$$f_L(l) = \sum_{n=0}^{\infty} f_N(n) \cdot f_X^{*n}(l) \quad (13)$$

Here, $f_N(n)$ represents the probability mass function of the number of claims, and $f_X^{*n}(l)$ is the n-fold convolution of the severity distribution $f_X(l)$, denoted by:

$$f_X^{*n}(l) = \underbrace{f_X * f_X * \cdots * f_X}_{n\text{times}}(l) \quad (14)$$

Each term in the sum represents the total loss distribution for a specific number of claims, weighted by the probability of having n claims. This establishes that $f_L(l)$ is indeed a mixture of severity distributions, where the weights are determined by the frequency distribution $f_N(n)$.

Thus, we conclude:

$$f_L(l) = \sum_{n=0}^{\infty} f_N(n) \cdot \left(\underbrace{f_X * f_X * \cdots * f_X}_{n\text{times}} (l) \right) \quad (15)$$

This completes the proof.

If the frequency distribution follows a Poisson distribution and the severity distribution is exponential, then the total loss distribution is Gamma-distributed.

Proof. Let $N \sim \text{Poisson}(\lambda)$, which means the frequency of claims follows a Poisson distribution with rate parameter λ. The severity of each claim is modeled by an exponential distribution, $X \sim \text{Exponential}(\theta)$, with rate parameter θ. Thus, the total loss L is the sum of N independent and identically distributed (i.i.d.) exponential random variables, i.e.

$$L = \sum_{i=1}^{N} X_i$$

where $X_1, X_2, \ldots, X_N$ are i.i.d. random variables with distribution $X \sim \text{Exponential}(\theta)$.

To find the distribution of the total loss L, we use the property that the sum of a random number of i.i.d. exponential random variables, where the random number follows a Poisson distribution, results in a Gamma distribution. Specifically, if $N \sim \text{Poisson}(\lambda)$, and each $X_i \sim \text{Exponential}(\theta)$, then L follows a Gamma distribution with shape parameter λ and rate parameter θ. Mathematically, we can express this as:

$$L \sim \text{Gamma}(\lambda, \theta)$$

where $\text{Gamma}(\lambda, \theta)$ denotes the Gamma distribution with shape parameter λ and rate parameter θ.

Proof Outline: The total loss distribution L is a compound distribution, which arises from the convolution of a Poisson-distributed random variable (frequency of claims) and an exponential random variable (severity of each claim). By using the properties of the compound distribution, we can deduce that the total loss L follows a Gamma distribution with parameters (λ, θ).

In actuarial science, the Loss Distribution Approach is primarily used for estimating reserves and premiums. By accurately modeling the frequency and severity of claims, actuaries can determine the total risk exposure and set appropriate premiums for policyholders. The convolution of frequency and severity distributions allows for a more precise estimate of the total loss, which is crucial for pricing and reserving in insurance.

The Loss Distribution Approach provides a robust framework for modeling insurance risk. By combining frequency and severity distributions, it allows actuaries to calculate the total loss distribution and estimate the reserves and premiums required to cover future claims. Although the approach has certain limitations, it remains one of the most widely used methods in actuarial science due to its flexibility and applicability.

4- Actuarial Underwriting Methods

4-1- Class Rating Method

Risk classification plays a pivotal role in insurance underwriting by enabling actuaries to segment policyholders into homogeneous groups based on shared risk characteristics. The *Class Rating Method* is an actuarial technique that assigns premiums to these groups by balancing equity and efficiency. This method is extensively used in property, casualty, and life insurance lines [13].

The Class Rating Method aims to determine the expected loss per policyholder within a given risk class, which serves as the basis for premium calculation. Let:

$$\mathbb{E}[L_i] = \mu_i$$

denote the expected loss for risk class i, where L_i represents the loss random variable and μ_i its mean. The premium P_i for the risk class is computed as:

$$P_i = \mu_i + \text{L} \cdot \sigma_i,$$

where L is the loading factor and σ_i is the standard deviation of L_i, capturing uncertainty.

The method relies on the assumption that risk classes are mutually exclusive and exhaustive, meaning every policyholder belongs to exactly one class. The expected loss for the portfolio is:

$$\mathbb{E}[L] = \sum_{i=1}^{n} w_i \mu_i,$$

where w_i is the proportion of policies in class i.

In a fair and sustainable insurance framework, the aggregate weighted premium collected must equate to the sum of the expected total loss and the incurred operational expenses:

$$\sum_{i=1}^{n} w_i P_i = \mathbb{E}[L] + E, \tag{16}$$

where w_i represents the weighting factor for policy i, P_i is the premium for policy i, $\mathbb{E}[L]$ is the expected total loss, and E denotes the total expenses.

Proof. The total weighted premium collected, $\sum_{i=1}^{n} w_i P_i$, accounts for all policy contributions. For a fair system, this must balance the insurer's obligations, which include the expected total loss and the operational expenses. Expressing the premium P_i for each policy i as:

$$P_i = \frac{\mathbb{E}[L_i]+e_i}{w_i}, \tag{17}$$

where $\mathbb{E}[L_i]$ is the expected loss associated with policy i, e_i represents the expenses attributed to policy i, and w_i ensures appropriate scaling.

Substituting Equation 17 into Equation 16, we have:

$$\sum_{i=1}^{n} w_i P_i = \sum_{i=1}^{n} w_i \left(\frac{\mathbb{E}[L_i]+e_i}{w_i}\right) = \sum_{i=1}^{n} \mathbb{E}[L_i] + \sum_{i=1}^{n} e_i.$$

Using the linearity of expectation:

$$\sum_{i=1}^{n} \mathbb{E}[L_i] = \mathbb{E}[\sum_{i=1}^{n} L_i] = \mathbb{E}[L], \tag{19}$$

where $\mathbb{E}[L]$ is the expected total loss. Similarly, summing over all expenses yields:

$$\sum_{i=1}^{n} e_i = E, \tag{20}$$

where E denotes the total expenses. Substituting equations 19 and 20 into Equation 18, we conclude

$$\sum_{i=1}^{n} w_i P_i = \mathbb{E}[L] + E. \tag{21}$$

This completes the proof.

The following algorithm illustrates the step-by-step procedure for implementing the Class Rating Method.

Algorithm 5. Class Rating Method with Enhanced Mathematical Notation

Risk classes $\{C_1, C_2, \dots, C_n\}$, individual loss data $\{L_1, L_2, \dots, L_m\}$, and loading factor λ. Risk-adjusted premiums $\{P_1, P_2, \dots, P_n\}$. **for** each risk class C_i, where $i = 1,2, \dots, n$ Compute the mean loss for class C_i as:

$$\mu_i = \frac{1}{|C_i|}\sum_{j\in C_i} L_j \tag{22}$$

where $|C_i|$ denotes the number of data points in class C_i. Estimate the standard deviation of losses for class C_i as:

$$\sigma_i = \sqrt{\frac{1}{|C_i|-1}\sum_{j\in C_i} (L_j - \mu_i)^2} \tag{23}$$

Determine the risk-adjusted premium for class C_i using:

$$P_i = \mu_i + \lambda \cdot \sigma_i \tag{24}$$

where $\lambda > 0$ is the pre-specified loading factor. The set of computed premiums:

$$\{P_1, P_2, \dots, P_n\} \tag{25}$$

The Class Rating Method can be formulated as an optimization problem:

$$\min_{\{P_i\}} \sum_{i=1}^{n} w_i (P_i - \mu_i)^2,$$

subject to:

$$\sum_{i=1}^{n} w_i P_i \geq \mathbb{E}[L] + \text{Expenses}.$$

The optimal premiums satisfy:

$$P_i = \mu_i + \lambda \cdot \sigma_i,$$

where λ is the Lagrange multiplier for the constraint.

The Class Rating Method ensures premiums are actuarially fair and operationally viable. It aligns with regulatory frameworks such as IFRS 17 by guaranteeing solvency through adequate reserving and pricing practices [14]. The Class Rating Method provides a robust foundation for actuarial underwriting by systematically incorporating risk heterogeneity into pricing models. This article demonstrated its theoretical validity, computational structure, and practical applications.

4-2- Risk-Adjusted Pricing

Risk-Adjusted Pricing Method (RAPM) is a cornerstone of actuarial underwriting, particularly in non-life insurance, where risk quantification and premium allocation are critical. This method integrates advanced statistical models with risk management principles to compute premiums reflective of policyholder-specific risk profiles [15-17].

The structure of RAPM consists of the following key components:

1. *Risk quantification*: Estimation of risk using probability distributions.
2. *Pricing adjustments*: Adjusting base premiums for expected losses, risk margins, and capital costs.
3. *Validation and calibration*: Ensuring the model aligns with observed data and satisfies regulatory requirements.

The RAPM can be expressed mathematically as follows:

$$P_{\text{adjusted}} = P_{\text{base}} + \lambda \cdot \text{Var}(L) + \mu \cdot \text{CoVaR}(L, R), \tag{26}$$

where:

- P_{base}: Base premium calculated from expected loss,
- λ: Risk loading coefficient,
- μ: Coefficient of capital costs,
- $\text{Var}(L)$: Variance of loss L,
- $\text{CoVaR}(L, R)$: Conditional Value-at-Risk (CoVaR) between losses L and reserves R.

The RAPM ensures positive net premium when $\lambda > 0$ and $\mu > 0$.

Proof. Given the RAPM formulation:

$$P_{\text{adjusted}} = P_{\text{base}} + \lambda \cdot \text{Var}(L) + \mu \cdot \text{CoVaR}(L, R),$$

it suffices to show $P_{\text{adjusted}} > 0$ when $\lambda, \mu > 0$. Since $P_{\text{base}} > 0$ (by definition of base premium) and both $\text{Var}(L)$ and $\text{CoVaR}(L, R)$ are non-negative, $P_{\text{adjusted}} > 0$ holds.

The following pseudo-algorithm outlines the computation of Risk-Adjusted Premiums:

Algorithm 6. Risk-Adjusted Pricing Algorithm with Advanced Mathematical Formulation

Policyholder data $\mathcal{D}$, base premium P_{base}, risk adjustment parameters λ, μ Risk-adjusted premium P_{adjusted}

Estimate the Expected Loss:

$\mathbb{E}[L] = \frac{1}{n}\sum_{i=1}^{n} L_i, \quad \forall L_i \in \mathcal{D}$

Calculate the Variance of Loss:

$\text{Var}(L) = \frac{1}{n-1}\sum_{i=1}^{n} (L_i - \mathbb{E}[L])^2$

Determine the Conditional Value-at-Risk (CoVaR):

$\text{CoVaR}(L, \alpha) = \mathbb{E}[L | L > \text{VaR}_{\alpha}(L)]$

where $\text{VaR}_{\alpha}(L)$ is the value-at-risk at confidence level α, defined as:

$\text{VaR}_{\alpha}(L) = \inf\{x \in \mathbb{R}: F_L(x) \geq \alpha\}, \quad F_L(x) = \Pr(L \leq x)$

Compute the Risk-Adjusted Premium:

$P_{\text{adjusted}} = P_{\text{base}} + \lambda \cdot \text{Var}(L) + \mu \cdot \text{CoVaR}(L, \alpha)$

where λ and μ are sensitivity parameters reflecting the insurer's risk appetite.

Output the Risk-Adjusted Premium:

ReturnP_{adjusted}

The Risk-Adjusted Pricing Method offers a structured approach to actuarial underwriting by integrating statistical rigor with practical application. Its flexibility allows insurers to adapt to diverse risk scenarios effectively

5- The Theory and Structure of the XGBoost Model

Extreme Gradient Boosting (XGBoost) is a scalable and accurate machine learning algorithm for regression and classification tasks. XGBoost builds an ensemble of decision trees in a sequential manner, optimizing a regularized objective function to balance model complexity and predictive accuracy. This section introduces its mathematical formulation and practical relevance.

XGBoost seeks to minimize the following objective function:

$$\mathcal{L}(\Theta) = \sum_{i=1}^{n} l(y_i, \hat{y}_i) + \sum_{k=1}^{K} \Omega(f_k), \tag{27}$$

where $l(y_i, \hat{y}_i)$ represents the loss function, $\Omega(f_k)$ is the regularization term, n is the number of data points, and K is the number of trees in the ensemble. The predicted value $\hat{y}_i$ is defined as:

$$\hat{y}_i = \sum_{k=1}^{K} f_k(x_i), \quad f_k \in \mathcal{F}, \tag{28}$$

where $\mathcal{F}$ denotes the space of regression trees.

The regularization term $\Omega(f)$ ensures model simplicity:

$$\Omega(f) = \gamma T + \frac{1}{2}\lambda \parallel w \parallel^2, \tag{29}$$

where T is the number of leaves in the tree, γ penalizes the number of leaves, and λ controls the regularization of leaf weights.

5-1- Optimization Framework

The optimization uses second-order Taylor expansion for the loss function:

$$\mathcal{L}^{(t)} \approx \sum_{i=1}^{n} \left[g_i f_t(x_i) + \frac{1}{2} h_i f_t^2(x_i) \right] + \Omega(f_t), \tag{30}$$

where $g_i = \frac{\partial l(y_i, \hat{y}_i^{(t-1)})}{\partial \hat{y}_i}$ and $h_i = \frac{\partial^2 l(y_i, \hat{y}_i^{(t-1)})}{\partial \hat{y}_i^2}$ are the first and second derivatives, respectively.

The optimal splitting criterion at a decision tree node is defined by maximizing the split gain, given as:

$$\mathcal{G} = \frac{1}{2}\left[\frac{\mathcal{G}_L^2}{\mathcal{H}_L + \lambda} + \frac{\mathcal{G}_R^2}{\mathcal{H}_R + \lambda} - \frac{\mathcal{G}_T^2}{\mathcal{H}_T + \lambda} \right] - \gamma, \tag{31}$$

where:

$$\mathcal{G}_T = \sum_{i \in T} g_i, \quad \mathcal{H}_T = \sum_{i \in T} h_i, \tag{32}$$

$$\mathcal{G}_L = \sum_{i \in L} g_i, \quad \mathcal{H}_L = \sum_{i \in L} h_i \tag{33}$$

$$\mathcal{G}_R = \sum_{i \in R} g_i, \quad \mathcal{H}_R = \sum_{i \in R} h_i, \tag{34}$$

with g_i and h_i representing the first and second-order derivatives of the loss function with respect to the model output for the i-th data point, respectively. The hyperparameters $\lambda > 0$ and $\gamma > 0$ control the regularization of the gain calculation.

The objective of the split gain is to quantify the improvement in the loss function reduction achieved by splitting the data at a given node. For any split that partitions the dataset T into subsets L and R, the gain is derived as follows:

The total loss reduction before the split is given by:

$$\mathcal{L}_T = -\frac{1}{2}\frac{\mathcal{G}_T^2}{\mathcal{H}_T + \lambda}. \tag{35}$$

After the split, the loss reduction is computed separately for the left and right child nodes:

$$\mathcal{L}_L = -\frac{1}{2}\frac{\mathcal{G}_L^2}{\mathcal{H}_L + \lambda}, \quad \mathcal{L}_R = -\frac{1}{2}\frac{\mathcal{G}_R^2}{\mathcal{H}_R + \lambda}. \tag{36}$$

The net improvement, or gain, due to the split is therefore:

$$\mathcal{G} = (\mathcal{L}_L + \mathcal{L}_R) - \mathcal{L}_T - \gamma. \tag{37}$$

Substituting from Equations 35 and 36 into equation 37, and simplifying terms, we obtain the split gain as defined in Equation 31.

Thus, the split that maximizes $\mathcal{G}$ corresponds to the optimal partition of T. This concludes the proof.

Below is the pseudo-algorithm for XGBoost regression:

Algorithm 7. XGBoost Regression

Input: Training dataset $\mathcal{D} = \{(\mathbf{x}_i, y_i)\}_{i=1}^{n}$ where $\mathbf{x}_i \in \mathbb{R}^d$ and $y_i \in \mathbb{R}$, learning rate η, maximum number of trees T, regularization parameters λ, γ, and a loss function $\mathcal{L}(y, \hat{y})$.

Initialize: Model prediction $\hat{y}_i^{(0)} = 0$ for all $i = 1, \ldots, n$.

for $t = 1$ to T Compute the gradient $g_i^{(t)}$ and the Hessian $h_i^{(t)}$:

$$g_i^{(t)} = \frac{\partial \mathcal{L}(y_i, \hat{y}_i^{(t-1)})}{\partial \hat{y}_i^{(t-1)}}, \quad h_i^{(t)} = \frac{\partial^2 \mathcal{L}(y_i, \hat{y}_i^{(t-1)})}{\partial (\hat{y}_i^{(t-1)})^2}$$

Construct a decision tree $f_t(\mathbf{x})$ to minimize the regularized objective:

$$\mathcal{L}^{(t)} = \sum_{i=1}^{n} \left[g_i^{(t)} f_t(\mathbf{x}_i) + \frac{1}{2} h_i^{(t)} f_t^2(\mathbf{x}_i) \right] + \gamma \, \text{num\_leaves}(f_t) + \frac{\lambda}{2} \| f_t \|^2$$

Update the model prediction:

$$\hat{y}_i^{(t)} = \hat{y}_i^{(t-1)} + \eta f_t(\mathbf{x}_i)$$

Output: Final prediction $\hat{y}_i = \hat{y}_i^{(T)}$ for all $i = 1, \ldots, n$.

$$\| \nabla \mathcal{L}(\theta) - \nabla \mathcal{L}(\theta') \| \leq L \| \theta - \theta' \| \quad \forall \, \theta, \theta' \in \mathcal{R}^d,$$

where $\nabla \mathcal{L}(\theta)$ denotes the gradient of the loss function and L is the smoothness constant.

Under these conditions, the convergence rate of the optimization algorithm used in XGBoost is $O\left(\frac{1}{T}\right)$, where T is the number of iterations. Specifically, after T iterations, the optimization error satisfies:

$$\mathcal{L}(\theta_T) - \mathcal{L}(\theta^*) \leq \frac{2(\mathcal{L}(\theta_0) - \mathcal{L}(\theta^*))}{T},$$

where θ_T denotes the parameters at the T-th iteration, θ^* represents the optimal parameters, and θ_0 is the initial parameter estimate.

Proof. To prove the lemma, we will rely on the fundamental concepts of convexity and smoothness of the loss function, and use the gradient descent analysis.

Step 1: Smoothness of the Loss Function Since $\mathcal{L}(\theta)$ is L-smooth, we have the following inequality for any two parameter vectors θ and θ':

$$\mathcal{L}(\theta) \leq \mathcal{L}(\theta') + \nabla \mathcal{L}(\theta')^T (\theta - \theta') + \frac{L}{2} \| \theta - \theta' \|^2.$$

By applying this smoothness condition iteratively, we can bound the decrease in the loss function at each step of the optimization process.

Step 2: Gradient Descent Analysis

Let $\theta_{t+1} = \theta_t - \eta \nabla \mathcal{L}(\theta_t)$ represent the parameter update rule, where η is the learning rate. The update rule implies the following relationship:

$$\mathcal{L}(\theta_{t+1}) \leq \mathcal{L}(\theta_t) - \eta \| \nabla \mathcal{L}(\theta_t) \|^2 + \frac{L\eta^2}{2} \| \nabla \mathcal{L}(\theta_t) \|^2.$$

Choosing a learning rate $\eta = \frac{1}{L}$ ensures that the second term does not dominate, leading to a decrease in the loss function:

$$\mathcal{L}(\theta_{t+1}) \leq \mathcal{L}(\theta_t) - \frac{1}{L} \| \nabla \mathcal{L}(\theta_t) \|^2.$$

Step 3: Bounding the Error

By iterating the above update rule for T steps, we obtain:

$$\mathcal{L}(\theta_T) - \mathcal{L}(\theta^*) \leq \mathcal{L}(\theta_0) - \mathcal{L}(\theta^*) - \sum_{t=0}^{T-1} \frac{1}{L} \| \nabla \mathcal{L}(\theta_t) \|^2.$$

Using the fact that $\mathcal{L}(\theta)$ is convex, we know that:

$$\mathcal{L}(\theta) - \mathcal{L}(\theta^*) \geq \nabla\mathcal{L}(\theta)^T(\theta - \theta^*) \quad \forall\theta.$$

Therefore, the gradient updates ensure that the error at each step decreases with a rate of $O\left(\frac{1}{T}\right)$, completing the proof.

XGBoost is inherently robust to overfitting due to its regularization framework, which incorporates both L1 (Lasso) and L2 (Ridge) regularization, controlling model complexity and ensuring generalization.

Proof. To show the robustness of XGBoost to overfitting, we analyze its objective function, which combines both the loss function and regularization terms.

The general objective function in XGBoost can be expressed as:

$$\mathcal{L}(\theta) = \sum_{i=1}^{n} \ell(y_i, \hat{y}_i) + \sum_{k=1}^{K} \Omega(f_k)$$

where:

• $\ell(y_i, \hat{y}_i)$ is the loss function that measures the discrepancy between the true target y_i and the predicted value $\hat{y}_i$,

• $\Omega(f_k)$ is the regularization term applied to the k-th tree f_k.

For the regularization term, XGBoost uses a combination of L1 (Lasso) and L2 (Ridge) regularization:

$$\Omega(f_k) = \gamma T + \frac{1}{2}\lambda \sum_{j=1}^{T} w_j^2$$

where:

• γT penalizes the number of terminal leaves (nodes) T in the tree, promoting simpler models with fewer nodes,

• λ is the L2 regularization coefficient that penalizes the sum of squared weights of the leaf nodes w_j, controlling the model's complexity and preventing overfitting,

• w_j are the weights at each leaf node of the tree

Incorporating both L1 and L2 regularization, the objective becomes:

$$\mathcal{L}(\theta) = \sum_{i=1}^{n} \ell(y_i, \hat{y}_i) + \sum_{k=1}^{K} \left(\gamma T_k + \frac{1}{2}\lambda \sum_{j=1}^{T_k} w_j^2\right) + \alpha \sum_{k=1}^{K} \sum_{j=1}^{T_k} |w_j|$$

where: α is the L1 regularization parameter (Lasso) that promotes sparsity by encouraging certain weights w_j to be exactly zero.

The final form of the regularized objective function thus penalizes both the complexity of the tree (through T and the leaf weights w_j) and the magnitude of the weights (through L1 and L2 terms). This ensures that the model is not only accurate but also generalizes well to unseen data, mitigating the risk of overfitting.

The robustness of XGBoost to overfitting can therefore be attributed to its ability to balance bias and variance through

- *Tree Complexity Regularization*: Controlled by the parameter γ, reducing the likelihood of excessively complex trees.
- *Weight Shrinkage*: Governed by λ, which shrinks the leaf weights towards zero and prevents overfitting to the noise in the training data.
- *Sparsity Induction*: Enforced by the L1 term (controlled by α), which induces sparsity in the tree leaves, further reducing overfitting by removing unnecessary features.

Thus, XGBoost's incorporation of regularization terms both simplifies the model and improves its ability to generalize, making it highly effective in preventing overfitting.

XGBoost regression combines efficiency and accuracy, making it a powerful tool in predictive modelling.

6- The Brighton Mahohoho XGBoost Probability Based Inflation Adjusted Frequency Severity Automated Actuarial Loss Reserving-Risk Pricing-Underwriting Model

The traditional actuarial approach to loss reserving and pricing often relies on static models that fail to account for the dynamic nature of insurance claims, particularly in occupational indemnity insurance. This paper presents a novel framework that combines the power of Artificial Intelligence (AI) and traditional actuarial models to deliver more accurate, inflation-adjusted loss reserving and pricing models. The model integrates the XGBoost algorithm, a widely used machine learning technique, to predict the frequency and severity of claims, taking into account inflation trends and policyholder behavior.

The model presented in this paper builds on the classical frequency-severity framework, which decomposes insurance losses into two components: frequency, which describes how often claims occur, and severity, which quantifies the cost of each claim. The model extends this framework by adjusting for inflation, a crucial factor in long-term indemnity insurance policies, and by incorporating AI-driven techniques for enhanced prediction.

6-1- The XGBoost Model for Frequency and Severity Prediction

XGBoost is an implementation of gradient-boosted decision trees designed for speed and performance. It is used in this framework to predict the frequency and severity of claims. The XGBoost model is formulated as follows:

$$f(x) = \sum_{k=1}^{K} \gamma_k h_k(x), \tag{38}$$

where: $f(x)$ is the predicted value (either frequency or severity), K is the number of decision trees, γ_k is the weight of the k-th tree, $h_k(x)$ is the output of the k-th decision tree.

The model is trained by minimizing a regularized loss function that includes both the training loss and a penalty term for tree complexity:

$$\mathcal{L}(\theta) = \sum_{i=1}^{N} [\text{Loss}(y_i, f(x_i)) + \lambda \parallel \theta \parallel^2], \tag{39}$$

where: N is the number of data points, y_i is the true label (observed loss), $f(x_i)$ is the predicted loss, λ is the regularization parameter, and θ are the model parameters.

Here is the implementation of the XGBoost-based model for Automated Actuarial Loss Reserve Risk Premiums (AALRRP) and the subsequent actuarial underwriting by risk groups. This model leverages advanced machine learning techniques to predict frequency and severity of claims and to adjust for inflation. The total loss reserve is calculated using these predictions, which are further refined by risk-based premiums (*AALRRP*) for policyholders categorized into various risk groups.

We begin by outlining the problem setup. The model takes historical claims data $\{\mathbf{X}_t\}$ and inflation rates $\{\text{Inflation}_t\}$ as input, along with the parameters for the frequency and severity models. These components allow for the prediction of both claim frequency and severity, adjusted for inflation, and the calculation of the total reserve.

6-1-1- Frequency and Severity Predictions

Given the data $\mathbf{X}_t$ for each time period t, we train two separate XGBoost models to predict frequency ($\hat{F}_t$) and severity ($\hat{S}_t$).

$$\hat{F}_t = XGBoost(\mathbf{X}_t, \boldsymbol{\theta}_F)$$

$$\hat{S}_t = XGBoost(\mathbf{X}_t, \boldsymbol{\theta}_S)$$

Here, $\boldsymbol{\theta}_F$ and $\boldsymbol{\theta}_S$ are the model parameters learned from the training data, which are used to predict the future frequency and severity of claims.

6-2- Adjustment for Inflation

To account for inflation over time, the predictions for frequency and severity are adjusted by multiplying the predicted values by the inflation rate at each time period t. Let the adjusted frequency F_t and adjusted severity S_t be expressed as:

$$F_t = \hat{F}_t \times (1 + \text{Inflation}_t)$$

$$S_t = \hat{S}_t \times (1 + \text{Inflation}_t)$$

These adjustments ensure that the model accounts for economic factors that influence the cost of claims over time.

6-3- Automated Actuarial Loss Reserve Risk Premiums (AALRRP)

The Automated Actuarial Loss Reserve Risk Premiums (AALRRP) are computed as the product of adjusted frequency, severity, and a risk factor RiskFactor_t, which reflects the specific risk associated with the policyholder group at time t:

$$\text{AALRRP}_t = (F_t \times S_t \times \text{Risk Factor}_t) \times (1 + \text{Inflation}_t)$$

Here, the RiskFactor_t is a function of actuarial risk categories, which could include policyholder demographics, policy types, and historical claims data. We define this factor as:

$$\text{Risk Factor}_t = f(\mathbf{X}_t, \text{Demographic}_t, \text{PolicyAttributes}_t)$$

This risk factor represents the variability in claims costs due to individual policyholder characteristics and external economic conditions

6-4- Total Loss Reserve

The total loss reserve is computed by summing the adjusted frequency and severity for each time period t across all periods T:

$$R = \sum_{t=1}^{T} F_t \times S_t \times (1 + \text{Inflation}_t)$$

This reserve R is the estimated amount required to cover all claims liabilities over the forecast horizon.

6-5- Underwriting by Risk Groups

Actuarial underwriting divides policyholders into risk groups based on their computed AALRRP. We classify the policyholders into three groups based on thresholds for the AALRRP:

$$\text{Group}_i = \{t | \text{AALRRP}_t \in \text{Group}_i\}$$

where the risk groups Group_i are defined as:

$$\text{Group}_i = \begin{cases} \text{High Risk}, & \text{if AALRRP}_t > \text{Threshold}_1 \\ \text{Medium Risk}, & \text{if Threshold}_2 \leq \text{AALRRP}_t \leq \text{Threshold}_1 \\ \text{Low Risk}, & \text{if AALRRP}_t < \text{Threshold}_2 \end{cases}$$

These thresholds are determined based on historical data and actuarial analysis. They allow insurers to classify policyholders into different risk categories, which helps in pricing policies appropriately. The following pseudo-algorithm outlines the key steps in the Brighton Mahohoho XGBoost Constructed Probability-Compliant Inflation Adjusted Frequency Severity Loss Reserving, Risk Pricing and underwriting model.

6-6- Algorithm: The Brighton Mahohoho XGBoost Model for Loss Reserving, Risk Pricing, and Underwriting by Risk Groups

The AALRRP calculation is consistent with the total loss reserve over time, adjusted for inflation. The total reserve is given by:

$$R = \sum_{t=1}^{T} F_t \times S_t \times (1 + \text{Inflation}_t)$$

The AALRRP for each time period is computed as:

$$\text{AALRRP}_t = (F_t \times S_t \times \text{RiskFactor}_t) \times (1 + \text{Inflation}_t)$$

By definition, the total loss reserve is the sum of the AALRRP for all periods, adjusted for the inflation at each period. Therefore, the total reserve R is consistent with the sum of AALRRP values, implying that the model is actuarially sound.

The classification of policyholders into risk groups based on the AALRRP ensures that the underwriting process reflects the actual risk exposure of the policyholders.

Proof. Given the risk group classification function:

$$\text{Group}_i = \begin{cases} \text{High Risk}, & \text{if AALRRP}_t > \text{Threshold}_1 \\ \text{Medium Risk}, & \text{if Threshold}_2 \leq \text{AALRRP}_t \leq \text{Threshold}_1 \\ \text{Low Risk}, & \text{if AALRRP}_t < \text{Threshold}_2 \end{cases}$$

By allocating policyholders to high, medium, or low-risk groups based on their AALRRP, insurers can ensure that the premium charged is proportional to the risk borne by each group. This methodology is consistent with risk-based pricing. The inflation adjustment has a proportional impact on both AALRRP and the total reserve, ensuring that the model accounts for future economic conditions.

Proof. Both AALRRP and the total reserve include an inflation adjustment factor $(1 + \text{Inflation}_t)$. As such, the model is capable of capturing the cumulative impact of inflation over the time horizon T, ensuring that the reserve and risk premiums are aligned with expected future cost increases.

This paper presents a robust, AI-driven framework for inflation-adjusted frequency-severity loss reserving and risk pricing in occupational indemnity insurance. By combining XGBoost with traditional actuarial methods, the model provides a powerful tool for predicting future claims and setting appropriate reserves. The use of inflation adjustments ensures that the model remains relevant and accurate over time, making it a valuable tool for insurers.

Algorithm 8. Brighton Mahohoho XGBoost Model for Loss Reserving, Risk Pricing, and Underwriting by Risk Groups

1. **Input:** Historical claims data $\{\mathbf{X}_t\}$, inflation rates $\{\text{Inflation}_t\}$, model parameters θ_F, θ_S

2. **Output:** Frequency and severity predictions, total loss reserve, Automated Actuarial Loss Reserve Risk Premiums (AALRRP), and underwriting by risk groups

3. **Step 1:** Train XGBoost model for frequency prediction:

$$\hat{F}_t = XGBoost(\mathbf{X}_t, \boldsymbol{\theta}_F)$$

4. **Step 2:** Train XGBoost model for severity prediction:

$$\hat{S}_t = XGBoost(\mathbf{X}_t, \boldsymbol{\theta}_S)$$

5. for each time period t do

6. **Step 3:** Adjust frequency prediction for inflation:

$$F_t = \hat{F}_t \times (1 + \text{Inflation}_t)$$

7. **Step 4:** Adjust severity prediction for inflation:

$$S_t = \hat{S}_t \times (1 + \text{Inflation}_t)$$

8. **end for**

9. **Step 5:** Compute total reserve:

$$R = \sum_{t=1}^{T} F_t \times S_t \times (1 + \text{Inflation}_t)$$

10. **Step 6:** Compute the Automated Actuarial Loss Reserve Risk Premiums (AALRRP) for each time period t as follows:

$$\text{AALRRP}_t = (F_t \times S_t \times \text{Risk Factor}_t) \times (1 + \text{Inflation}_t)$$

where Risk Factor_t is a function of actuarial risk categories based on historical data $\mathbf{X}_t$, such as:

$$\text{RiskFactor}_t = f(\mathbf{X}_t, \text{Demographic}_t, \text{Policy Attributes}_t)$$

11. **Step 7:** Underwrite by risk groups based on the computed AALRRP_t:

$$\text{Group}_i = \{t | \text{AALRRP}_t \in \text{Group}_i\}$$

where the risk groups Group_i are defined by specific ranges of AALRRP_t, such as high, medium, and low risk groups:

$$\text{Group}_i = \begin{cases} \text{High Risk}, & \text{if AALRRP}_t > \text{Threshold}_1 \\ \text{Medium Risk}, & \text{if Threshold}_2 \leq \text{AALRRP}_t \leq \text{Threshold}_1 \\ \text{Low Risk}, & \text{if AALRRP}_t < \text{Threshold}_2 \end{cases}$$

12. **Step 8:** Return total reserve and risk group underwritings:

Return: Total Reserve R, Under writing Groups$\{\text{Group}_i\}$

6-7-IFRS 17 Compliance in Loss Reserving, Risk Pricing, and Underwriting

IFRS 17, the international accounting standard for insurance contracts, has brought significant changes to the way insurers approach loss reserving, risk pricing, and underwriting. This regulation mandates that insurance companies recognize profits and losses over the lifetime of insurance contracts, in alignment with the expected future cash flows. The focus is on creating more transparent, consistent, and comparable financial statements.

In the context of occupational indemnity insurance, IFRS 17 compliance requires the estimation of future cash flows for both claims (including inflation adjustments) and premiums. To meet these requirements, a framework based on probabilistic modeling can be used to project uncertain future events related to insurance claims, which aligns well with the underlying concepts of loss reserving under IFRS 17.

The IFRS 17 framework involves calculating two main elements:

- *Best Estimate of Liability (BEL)*: This represents the expected present value of future cash flows, which includes both expected claims and premiums.
- *Contractual Service Margin (CSM)*: This represents the unearned profit from an insurance contract that will be recognized as revenue over time.

In this paper, Inflation-Adjusted Frequency-Severity Modeling helps ensure that both the frequency of claims (how often claims occur) and the severity of claims (how much each claim costs) are appropriately factored into the reserve estimation, while incorporating the necessary adjustments for inflation. This ensures that the calculated Best Estimate of Liability (BEL) adheres to IFRS 17's requirements for the projection of future cash flows.

Additionally, Risk Premiums are adjusted to reflect inflation expectations, policyholder behavior, and other risk factors. Automated Loss Reserving is aligned with IFRS 17 by using actuarial methods that consider adjustments for the time value of money, inflation, and other actuarial assumptions, providing more accurate and compliant estimates for reserving.

6-8- Novelty of the Study

This study presents several novel contributions to the field of actuarial science, specifically within the context of occupational indemnity insurance and IFRS 17-compliant pricing and reserving models. The primary innovations of this research are as follows.

This study introduces a novel hybrid framework that combines multiple artificial intelligence (AI) techniques, with particular emphasis on the Extreme Gradient Boosting (XGBoost) algorithm. This AI-driven methodology automates the estimation of actuarial loss reserves and risk premiums, achieving both high predictive accuracy and regulatory compliance with IFRS 17. The integration of XGBoost for loss reserving and pricing provides a more nuanced and dynamic estimation of policyholder claims, significantly improving predictive outcomes compared to traditional actuarial models. The paper extends classical frequency-severity models by incorporating dynamic inflation adjustments within the claims process. This is a novel approach in the context of occupational indemnity insurance, where inflation has historically been a challenging factor to incorporate in a precise, actionable manner. The use of Poisson and Gamma distributions to model claim frequencies and severities, while simultaneously adjusting for inflation, enhances the accuracy of both premium pricing and reserve estimation. A key contribution of this work is the simulation of a comprehensive dataset that mimics real-world occupational indemnity insurance policies, incorporating elements such as policyholder characteristics, premiums, claims, and inflation adjustments. This simulation not only provides a rich foundation for the application of advanced actuarial techniques but also allows for rigorous testing of model robustness and performance under different risk scenarios. This study presents a detailed approach to integrating IFRS 17 metrics such as the Contractual Service Margin (CSM), Fulfillment Cash Flows (FCF), and Risk Adjustments into the actuarial loss reserving process. The paper uniquely demonstrates how to incorporate these regulatory requirements into the predictive models, providing a comprehensive framework for insurers to ensure compliance while optimizing pricing and reserving strategies.

The segmentation of policyholders into five actuarial underwriting clusters based on simulated claims, premiums, and risk characteristics is an innovative method in the context of automated actuarial pricing. By using AI-driven algorithms for cluster analysis, this study enhances the precision of risk classification and pricing, leading to better-targeted underwriting strategies. The clustering technique is paired with the computation of Automated Net Actuarial Underwriting Balance (ANAUB), which evaluates the financial impact of different underwriting strategies. In addition to predictive modeling; the paper introduces a robust sensitivity analysis framework to evaluate the resilience of the proposed models under various stress-testing scenarios. This ensures that the derived loss reserves and premiums remain within the acceptable boundaries of risk and regulatory compliance, even under fluctuating market conditions.

By addressing the complexities of actuarial pricing and reserving through an AI-driven, simulation-based framework, this study provides an innovative, comprehensive solution for the challenges faced by insurers in complying with IFRS 17 regulations. The combination of AI, inflation adjustment, IFRS 17 integration, and advanced clustering techniques offers a cutting-edge approach that significantly advances the state of the art in actuarial science.

6-9- Contribution to Actuarial Science

This research makes several significant contributions to the field of actuarial science, particularly in the areas of automated actuarial loss reserving, risk pricing, and the integration of IFRS 17 compliance. One of the key contributions of this study is the development and application of artificial intelligence (AI) techniques, specifically Extreme Gradient Boosting (XGBoost), in the automation of actuarial loss reserving and risk pricing. Traditional actuarial methods often rely on static, manual processes for loss reserving, which can be time-consuming and prone to error. This study demonstrates that AI techniques can not only automate but also significantly enhance the accuracy of these processes. By implementing machine learning models like XGBoost, the research provides a more flexible and adaptive approach to estimating loss reserves and risk premiums, offering insights that are critical for insurers navigating complex and volatile markets.

The integration of inflation-adjusted frequency-severity models represents a groundbreaking advancement in occupational indemnity insurance. Previous models often fail to account for the evolving nature of inflation within claims estimation, leading to potential under- or over-estimation of reserves. This research introduces a dynamic framework that adjusts both the frequency and severity of claims for inflation, improving the accuracy and realism of reserve calculations and premium pricing. This model provides actuarial practitioners with a more robust tool to account for inflationary pressures, ensuring better financial planning and risk management for insurers. Another important contribution is the creation of a simulated actuarial dataset that mimics real-world conditions in the occupational indemnity insurance market. By generating realistic policyholder data, including premiums, claims, and risk characteristics, the study allows for comprehensive testing and validation of actuarial models. This simulated data serves as a foundation for assessing model performance under various risk scenarios, offering valuable insights into model

robustness and predictive accuracy. The simulation framework presented in this study provides actuarial researchers and practitioners with a reliable tool for evaluating models before their deployment in live settings.

This paper contributes to the practical application of IFRS 17 by demonstrating how actuarial models can be developed and adapted to comply with its complex requirements. The research showcases how key IFRS 17 metrics—such as the Contractual Service Margin (CSM), Fulfillment Cash Flows (FCF), and Risk Adjustments—can be incorporated directly into the actuarial loss reserving and pricing process. This integration ensures that insurers remain compliant with regulatory standards while optimizing their pricing strategies. By bridging the gap between actuarial practice and IFRS 17, this study provides a comprehensive framework that enhances the ability of actuaries to navigate the evolving regulatory landscape. The study also introduces an innovative approach to actuarial underwriting by segmenting policyholders into five distinct actuarial underwriting clusters based on their risk profiles. This segmentation is achieved using advanced machine learning techniques, allowing for more precise pricing and underwriting strategies tailored to the characteristics of each cluster. The ability to segment policyholders in this way leads to more accurate risk pricing and helps insurers optimize their portfolio management strategies. This contribution represents a significant step forward in the application of AI and data analytics to enhance the actuarial underwriting process.

The research introduces a novel framework for sensitivity and stress testing to evaluate the stability of actuarial models under various economic and market conditions. This ensures that actuarial predictions for loss reserves and premiums remain reliable and robust, even in the face of extreme scenarios. The use of such stress tests is crucial for the long-term sustainability of insurance pricing models, helping insurers prepare for unforeseen risks and volatile market conditions. This contribution strengthens the overall reliability of the actuarial models proposed in the study and provides a more comprehensive view of the potential risks associated with different pricing and reserving strategies.

In a nutshell, this study makes valuable contributions to the field of actuarial science by developing advanced, AI-driven methods for loss reserving, risk pricing, and IFRS 17 compliance. Through the integration of inflation-adjusted frequency-severity models, enhanced risk segmentation, and robust testing methodologies, this research offers a significant advancement in the way actuaries approach pricing, reserving, and risk management. The findings provide actuaries with a comprehensive, modern toolkit for navigating the complexities of today's insurance landscape, ensuring that insurers can better manage risk and achieve regulatory compliance while optimizing their financial strategies.

7- Survey of Methods and Literature Review

Inflation-adjusted frequency-severity models have long been a crucial aspect of actuarial practice, especially in the context of pricing and reserving. These models adjust for the impact of inflation on both the frequency and severity of claims, a necessary step for accurate estimation of reserves and premiums. Several early works, such as those by Denuit et al. [2], pioneered stochastic methods for claim reserving. Mack's [12] work on the Chain Ladder method provided foundational techniques for estimating reserves under the assumption of constant inflation, but it did not account for varying inflation rates over time.

Subsequent studies, such as Wüthrich & Merz [18], extended the Chain Ladder method by introducing the use of stochastic inflation adjustments. These approaches demonstrated how inflation factors could be modeled probabilistically, incorporating both inflation forecasts and historical data. Denuit et al. [19] introduced a more generalized framework for adjusting frequency and severity models with stochastic inflation processes. They emphasized the need for more complex models that could integrate multiple sources of inflation, such as medical inflation or economic inflation, to refine the estimates for claims severity. Their work set the stage for incorporating more sophisticated statistical methods into actuarial modeling.

The application of machine learning models, particularly XGBoost, has gained significant attention in the actuarial community in recent years. Chen & Guestrin [11] introduced XGBoost as a scalable and efficient algorithm for gradient boosting, which has been widely adopted across industries, including insurance. The algorithm's ability to handle large datasets with complex relationships between variables makes it particularly suited for actuarial tasks such as claim frequency prediction, severity modeling, and loss reserving.

Several studies have explored the use of XGBoost for insurance-related tasks. Wang et al. [20] applied XGBoost for predicting claim frequencies and severities, demonstrating its superiority over traditional methods such as Generalized Linear Models (GLM). Their results showed a marked improvement in predictive accuracy, particularly when dealing with non-linear relationships and interaction effects between features.

Further extending the use of machine learning in loss reserving, Mahohoho et al. [21] applied XGBoost to the estimation of reserves, focusing on its ability to model the non-linear effects of inflation and other covariates on claim development patterns. They compared the performance of XGBoost with more traditional actuarial methods and found that it provided a more robust and flexible approach, especially in volatile environments.

The introduction of IFRS 17 has brought a new era of actuarial modeling, particularly in how insurance liabilities and premiums are calculated. IFRS 17 requires that insurance contracts be measured using a current estimate of future cash flows, adjusted for risk and time value, introducing a more dynamic and granular approach to pricing and reserving. Yryku & Lamani [22] provided an early examination of the implications of IFRS 17 for insurance companies. They

highlighted how the new standards would require actuaries to consider more sophisticated modeling techniques, such as stochastic processes and machine learning algorithms, to comply with the increased emphasis on accuracy and transparency.

Later works, such as Palmborg [23], explored the integration of machine learning techniques with IFRS 17 in more detail, suggesting that XGBoost could play a crucial role in ensuring compliance by providing accurate estimates of future liabilities, incorporating adjustments for inflation, and considering the non-linearities inherent in insurance data. Their work emphasized the need for data science techniques to augment traditional actuarial methods, aligning with the increased focus on AI-driven actuarial models.

AI-driven augmented data science has been a promising frontier in enhancing actuarial practice. The integration of AI with traditional actuarial methods provides a robust framework for analyzing large and complex datasets that often arise in the insurance industry. Xiong et al. [24] reviewed the application of AI in loss reserving and risk pricing, focusing on how advanced algorithms such as XGBoost, Random Forests, and Neural Networks can be integrated into actuarial models to improve their predictive capabilities. In the context of inflation-adjusted frequency-severity models, Krasheninnikova et al. [25] explored how machine learning could help actuaries incorporate dynamic inflation adjustments. By using algorithms like XGBoost, they demonstrated how inflation parameters could be estimated more precisely by learning from historical claims data. This integrated approach resulted in more accurate predictions of claim outcomes, which is critical for both pricing and reserving.

With the implementation of IFRS 17, there is an increased emphasis on accurate and transparent financial reporting in insurance. This has spurred the development of models that not only predict losses but also comply with regulatory standards. Mahohoho [26] proposed an inflation-adjusted automated actuarial loss reserving model using Random Forest techniques, specifically tailored for fire insurance, aligning with IFRS 17 requirements. Mahohoho [27] developed a non-linear regression-based model for travel insurance, utilizing Gaussian Process Regression (GPR) to adjust for inflation in frequency-severity modeling. This approach integrates synthetic data generation and exploratory data analysis to enhance compliance with IFRS 17 and improve financial reporting. Clemente et al. [28] investigated the predictive performance of GBMs in modeling claim frequency and severity in auto insurance. Their study concluded that GBMs outperform traditional GLMs in capturing complex, non-linear relationships, thereby enhancing risk assessment accuracy. The integration of AI-driven methodologies, particularly GBMs like XGBoost, into actuarial science represents a significant advancement in modeling inflation-adjusted frequency and severity in occupational indemnity insurance. These models offer enhanced predictive accuracy and the ability to capture complex data patterns, aligning with the stringent requirements of IFRS 17. Continued research and development in this area are essential for refining these models and ensuring their robustness in various insurance applications.

The Brighton Mahohoho Probabilistic Framework for inflation-adjusted frequency-severity modeling aims to integrate AI-driven techniques such as XGBoost within a coherent probabilistic framework that aligns with the regulatory requirements of IFRS 17. This novel framework builds on previous works, combining stochastic inflation modeling with the power of machine learning to improve loss reserving and risk pricing in occupational indemnity insurance. It emphasizes transparency, flexibility, and the use of augmented actuarial data science to enhance traditional actuarial methods. Preliminary results presented by Mahohoho et al. [29] show how integrating these AI techniques can lead to significant improvements in model accuracy, particularly in adjusting for inflation's impact on claim frequency and severity. By using XGBoost within this framework, the model can dynamically adjust for inflation while providing actuarially sound estimates of future liabilities, ensuring compliance with IFRS 17.

8- Methodology

Research methodology refers to the systematic approach used to conduct research. It includes the methods, techniques, and procedures that guide the collection, analysis, and interpretation of data. The purpose of research methodology is to provide a clear and structured framework for the research process, ensuring that the research objectives are achieved with reliability and validity [30-32]. This study presents a probabilistic framework for inflation-adjusted frequency-severity modeling in occupational indemnity insurance, compliant with IFRS 17 standards. The framework integrates AI-driven augmented actuarial data science techniques, leveraging the XGBoost algorithm for advanced loss reserving, risk pricing, and underwriting.

8-1- Data Simulation

The first step involves the simulation of a comprehensive dataset representing occupational indemnity insurance policies. The simulated data includes several variables related to policy and claims information, as well as underwriting details.

- *Policy Information*: Data such as policy ID, industry type, occupation risk level, annual salary, policy term in years, policy type, coverage limit, and deductible are generated. These variables provide insights into the diversity and scope of the insurance policies under analysis. Let the annual salary be denoted as S_i, where $i = 1,2, \dots, n$ represents the n policyholders.
- *Claims Information*: The number of claims (frequency) and the severity (financial loss) of claims are simulated. The frequency of claims, denoted by N_i, follows a Poisson distribution".

$$P(N_i = k) = \frac{\lambda_i^k e^{-\lambda_i}}{k!}, \quad k = 0,1,2, \ldots,$$

where λ_i is the rate parameter for the Poisson distribution, typically a function of S_i and policy type. The severity, denoted by X_i, follows a Gamma distribution:

$$f_X(x_i) = \frac{x_i^{\alpha_i - 1} e^{-x_i/\beta_i}}{\Gamma(\alpha_i)\beta_i^{\alpha_i}}, \quad x_i \geq 0,$$

where α_i and β_i are the shape and scale parameters, respectively, and are functions of the claim history and policy features.

- *Premiums and Pricing*: The base premium, denoted as P_0, is calculated as a percentage of the annual salary:

$$P_0 = \kappa S_i$$

where κ is the base premium rate. The risk-adjusted premium, P_i, is derived by adjusting the base premium using a risk score R_i, which is simulated randomly for each policyholder. The adjusted premium is given by:

$$P_i = P_0 \cdot \left(1 + \frac{R_i}{100}\right).$$

- *Inflation Adjustment*: Inflation rates $\{\pi_i\}$ are simulated for each policyholder, and the reserves are adjusted accordingly to reflect the impact of inflation on future claims. Let C_i represent the estimated claims reserve for policyholder i at time t. The inflation-adjusted reserve $\tilde{C}_i$ can be calculated as:

$$\tilde{C}_i = C_i \cdot (1 + \pi_i)^{T-t},$$

where T is the total time horizon and t is the current time period.

- *Underwriting Information*: Underwriting scores and decisions (approve or decline) are simulated based on a normal distribution:

$$U_i \sim \mathcal{N}(\mu_U, \sigma_U^2),$$

where μ_U is the mean underwriting score, and σ_U^2 is the variance. The underwriting decision $\mathcal{D}_i$ is then made based on a predefined threshold, θ, such that:

$$\mathcal{D}_i = \begin{cases} \text{Approve}, & \text{if } U_i \geq \theta, \\ \text{Decline}, & \text{if } U_i < \theta. \end{cases}$$

The generated data includes 100,000 policyholders, providing a robust sample for analysis and model training.

8-2- Data Exploration and Visualization

The next phase involves comprehensive data exploration and visualization to understand the distribution of key variables, their interrelationships, and patterns within the dataset.

- *Summary Statistics*: Descriptive statistics and initial visualizations are generated for understanding the basic properties of the data, including distributions of annual salary by occupation risk level and industry types. Key statistical measures such as the mean μ, median m, and range $\mathcal{R}$ are computed for numeric variables:

$$\mu = \frac{1}{n}\sum_{i=1}^{n} S_i, \quad m = \text{median}(S), \quad \mathcal{R} = \max(S) - \min(S),$$

where $S = \{S_1, S_2, \ldots, S_n\}$ represents the set of salaries for all policyholders.

- *Correlation Analysis*: A correlation matrix $\mathbf{C}$ is computed for all numeric variables to capture the linear relationships between variables such as claim frequency, claim severity, base premium, risk score, and inflation-adjusted reserves. The correlation between two variables X and Y is given by:

$$\text{Corr}(X, Y) = \frac{\text{Cov}(X,Y)}{\sigma_X \sigma_Y},$$

where $\text{Cov}(X, Y)$ is the covariance and σ_X and σ_Y are the standard deviations of X and Y, respectively.

- *Dimensionality Reduction*: To explore complex relationships in the data, t-Distributed Stochastic Neighbor Embedding (t-SNE) is used for dimensionality reduction. Let $\mathbf{X} \in \mathbb{R}^{n \times d}$ represent the n-dimensional feature matrix. The t-SNE algorithm projects the high-dimensional data into two dimensions $\mathbf{Y} \in \mathbb{R}^{n \times 2}$ by minimizing the Kullback-Leibler divergence between the probability distributions P and Q:

$$D_{KL}(P||Q) = \sum_{i,j} P_{ij} \log \frac{P_{ij}}{Q_{ij}},$$

where P_{ij} represents the similarity between i and j in the high-dimensional space, and Q_{ij} represents the similarity in the low-dimensional space.

- *Cluster Analysis*: K-means clustering is applied to categorize the policies into four distinct groups based on the numerical features. Let $\mathbf{X}$ represent the feature matrix, and $\mathbf{C} = \{C_1, C_2, \ldots, C_k\}$ be the set of clusters. The objective function is to minimize the sum of squared distances from each data point to its corresponding cluster centroid:

$$J = \sum_{i=1}^{n} \sum_{k=1}^{K} \mathbf{1}_{c_i=k} \parallel \mathbf{x}_i - \mu_k \parallel^2,$$

where $\mathbf{1}_{c_i=k}$ is an indicator function that equals 1 if point i is assigned to cluster k, and μ_k is the centroid of cluster k.

- *3D Visualization*: A 3D plot is created using Plotly to visualize the clustering results with respect to dimensionality reduction, highlighting the relationships between t-SNE dimensions and risk scores. The 3D coordinates for each policy are plotted as:

$$\text{Plot}_i = (t - SNE_1, t - SNE_2, \text{RiskScore}_i),$$

where $t - SNE_1$ and $t - SNE_2$ are the first two dimensions of the t-SNE reduced data.

These exploratory steps enable the identification of key patterns in the data and inform the subsequent modeling steps. Visualization aids in understanding the complexity of the dataset and forms the foundation for building advanced actuarial models.

8-3- Data Processing and Integration

Following the data exploration phase, the simulated dataset is processed and prepared for integration into the probabilistic framework. This involves transforming categorical variables into factors, ensuring that all data is correctly formatted for machine learning models. The processed dataset serves as the foundational input for model development, including the implementation of the XGBoost algorithm for advanced loss reserving, risk pricing, and underwriting.

8-4- Model Building

The data is preprocessed and split into training and test datasets (80% for training and 20% for testing), ensuring that the model is not overfitted and generalizes well to unseen data.

8-4-1- Automated Actuarial Loss Reserving (ALR) Model

This model predicts the inflation-adjusted reserves ($\hat{R}_t^{inflation}$) based on policy attributes such as *Industry*, *Occupation Risk Level*, *Annual Salary*, and other underwriting factors. The prediction problem is framed as a regression task where XGBoost is employed with the reg:squarederror objective function to minimize the residual sum of squares. The corresponding model is expressed as:

$$\hat{R}_t^{inflation} = f(X_t; \boldsymbol{\theta}),$$

where X_t represents the feature vector for the t-th observation, and $\boldsymbol{\theta}$ is the vector of model parameters learned via training.

8-4-2- Automated Actuarial Risk Pricing (ARP) Model

This model estimates the risk-adjusted premium ($\hat{P}_t^{risk}$) based on similar policy attributes. The prediction framework remains consistent with that of the ALR model, with XGBoost used to predict the target variable. The corresponding model is expressed as:

$$\hat{P}_t^{risk} = g(X_t; \boldsymbol{\beta}),$$

where $g(\cdot)$ is a function parameterized by $\boldsymbol{\beta}$, learned during the training phase, and X_t represents the feature vector for the t-th observation.

For both models, the training process involves feature selection and hyperparameter tuning, ensuring the model's convergence and predictive accuracy. This is achieved through cross-validation and optimization of model hyperparameters using grid search or Bayesian optimization techniques.

8-4-3- Model Evaluation and Validation

Error (MAE), Mean Squared Error (MSE), and Root Mean Squared Error (RMSE):

$$\text{MAE} = \frac{1}{N}\sum_{t=1}^{N} |\hat{y}_t - y_t|,$$

$$\text{MSE} = \frac{1}{N}\sum_{t=1}^{N} (\hat{y}_t - y_t)^2,$$

$$\text{RMSE} = \sqrt{\frac{1}{N}\sum_{t=1}^{N} (\hat{y}_t - y_t)^2},$$

where N represents the total number of observations, $\hat{y}_t$ is the predicted value, and y_t is the actual value for the t-th observation.

The residuals of the models are analyzed through residual plots to detect any patterns, ensuring that the assumptions of homoscedasticity and normality are met. Additionally, scatter plots comparing the actual vs. predicted values are generated for both models, visually assessing the model's predictive capability.

8-4-4- Automated Actuarial Loss Reserve Risk Premium Estimation

An inflation model is introduced to simulate the real inflation rates, defined as:

$$I_t = \frac{C_t - C_{t-1}}{C_{t-1}},$$

where C_t and C_{t-1} are the cumulative claims or costs in periods t and $t-1$, respectively. The ALR and ARP model predictions are then combined using a weighted average approach to estimate the Automated Actuarial Loss Reserve Risk Premiums (AALRRPs):

$$\hat{A}ALRRP_t = \omega \hat{R}_t^{inflation} + (1-\omega)\hat{P}_t^{risk},$$

where $\omega = 0.6$ is the weight applied to the ALR model predictions, and the remaining weight $(1-\omega) = 0.4$ is allocated to the ARP model predictions. This integration aims to produce a robust, IFRS 17-compliant actuarial risk premium estimation.

8-5- Simulating Key IFRS 17 Metrics

This phase of the analysis focuses on simulating essential IFRS 17 metrics, including the Contractual Service Margin (CSM), Fulfilment Cash Flows (FCF), Risk Adjustments, Loss Components (for onerous contracts), Liability for Remaining Coverage (LRC), and Liability for Incurred Claims (LIC). These metrics are central to the IFRS 17 framework, which mandates the recognition and measurement of insurance contract liabilities. Below are the mathematical formulations for each key metric:

- *CSM Calculation:* The Contractual Service Margin (CSM) is computed as the difference between expected cash inflows (I_t^{inflow}) and expected claims outflows ($C_t^{outflow}$):

$$\text{CSM} = \mathbb{E}\left[\sum_{t=1}^{T} I_t^{inflow}\right] - \mathbb{E}\left[\sum_{t=1}^{T} C_t^{outflow}\right],$$

where T represents the contract duration.

- *FCF and Risk Adjustments*: The Fulfilment Cash Flows (FCF) are determined by subtracting expected claims outflows from expected cash inflows:

$$\text{FCF} = \mathbb{E}\left[\sum_{t=1}^{T} I_t^{inflow}\right] - \mathbb{E}\left[\sum_{t=1}^{T} C_t^{outflow}\right].$$

Additionally, a risk adjustment for non-financial risk (RA) is incorporated, calculated as 5% of the Automated Actuarial Loss Reserve Risk Premiums:

$$\text{RA} = 0.05 \cdot \hat{A}ALRRP_t.$$

- *Liabilities:* The Liability for Remaining Coverage (LRC) is calculated as the sum of the CSM and FCF:

$$\text{LRC} = \text{CSM} + \text{FCF}.$$

The Liability for Incurred Claims (LIC) is derived from the sum of the Loss Reserving and Inflation-Adjusted Reserves:

$$\text{LIC} = \sum_{t=1}^{T} \hat{R}_t^{inflation}.$$

- *Discount Rate Adjustment:* A 3% discount rate (r) is applied to the Liability for Incurred Claims to estimate the discounted reserves, representing the present value of future claims liabilities:

$$\hat{L}IC_{discounted} = \frac{\text{LIC}}{(1+r)^T},$$

where; $r = 0.03$ is the annual discount rate, and T is the time to settlement.

8-5-1- Data Visualization and Summary Statistics

To facilitate the interpretation and analysis of IFRS 17-related metrics, this study employs several advanced data visualization techniques. Specifically, histograms are plotted for each key metric (e.g., Contract Service Margin (CSM), Future Cash Flows (FCF), Risk Adjustment, Loss Component, Loss Reserves Component (LRC), Liability for Incurred Claims (LIC), Discounted Reserves, and Experience Adjustments) with optimal bin widths, ensuring an effective representation of the distribution. A color scheme is chosen to enhance clarity and focus on important trends across the different metrics.

Additionally, summary statistics, such as the mean (μ), standard deviation (σ), minimum (min), and maximum (max), are computed for each metric, thereby offering a comprehensive overview of the central tendency and variability of the data:

$$\mu_X = \frac{1}{n}\sum_{i=1}^{n} x_i \tag{40}$$

$$\sigma_X = \sqrt{\frac{1}{n}\sum_{i=1}^{n} (x_i - \mu_X)^2} \tag{41}$$

where x_i represents the observed values for each metric, and n denotes the number of observations.

8-6- XGBoost Model Compliance to IFRS17 Regulations

The study employs the XGBoost model to estimate the Automated Actuarial Loss Reserve Risk Premiums (AALRRPs), an essential component for advanced loss reserving and risk pricing. The model takes as input simulated predictions for Actuarial Loss Reserves (ALR) and Actuarial Risk Premiums (ARP), weighted according to predefined ratios, specifically 60% for ALR and 40% for ARP. This weighted averaging process is represented as:

$$AALRRP = 0.6 \cdot ALR + 0.4 \cdot ARP \tag{42}$$

where; $AALRRP$ represents the Automated Actuarial Loss Reserve Risk Premium, ALR denotes the Actuarial Loss Reserves, and ARP denotes the Actuarial Risk Premium.

8-6-1- IFRS 17 Metric Calculation and Contract Grouping

To comply with IFRS 17, we calculate several actuarial metrics including the reinsurance impact, assuming a 10% reinsurance coverage, and the expense ratios, set at 20% of the risk premiums. The Insurance Contract Liabilities (ICL) are estimated by discounting the future cash flows (FCF), which represent the present value of future liabilities, as follows:

$$ICL = \sum_{t=1}^{T} \frac{FCF_t}{(1+r)^t} \tag{43}$$

where FCF_t is the future cash flow at time t, and r is the discount rate applied to the cash flows. Sensitivity analysis is performed by adjusting the ALR and ARP predictions, applying changes of 10% and 5%, respectively, to assess the impact on the AALRRP estimates. These adjustments can be represented as:

$$AALRRP_{new} = AALRRP_{old} \cdot (1 + \delta) \tag{44}$$

where δ represents the percentage change (10% for ALR and 5% for ARP).

Contracts are grouped into three risk categories (Low, Medium, High) based on the quantiles of the AALRRP values. This categorization helps to assess the risk profile of the portfolio and allows further analysis of the distribution of AALRRPs across different risk groups

8-7- Development of the Automated Actuarial Underwriting Model

The dataset consists of AALRRPs derived from historical claims data. The first step involves segmenting policyholders into distinct underwriting clusters. This segmentation is based on AALRRPs, with the following steps:

$$\text{Cluster Range} = [\min(AALRRP), \max(AALRRP)] \tag{45}$$

Using this range, four underwriting clusters are defined: "Lowest Risk Underwriting Cluster," "Lower Risk Underwriting Cluster," "Higher Risk Underwriting Cluster," and "Highest Risk Underwriting Cluster." The $cut()$ function is employed to categorize premiums into these clusters, yielding a data frame associating each premium with its corresponding underwriting cluster. The distribution of premiums across the clusters is visualized using histograms generated via ggplot2, where the x-axis represents the premiums, and the y-axis represents frequency.

8-7-1- Simulating Expenses, Outgo, and Revenue

For each policyholder within the underwriting clusters, we simulate the expenses, outgo, and revenue. The expenses are simulated as a random percentage of the premium, denoted by $\text{Exp} \sim U(0.10,0.40)$, where $U(a,b)$ is a uniform distribution between a and b. The outgo is similarly simulated between 5% and 30% of the premium:

$$\text{Outgo} \sim U(0.05,0.30) \tag{46}$$

Revenue is assumed to be 1.5 times the premium, i.e.,

$$\text{Revenue} = 1.5 \cdot \text{Premium} \tag{47}$$

The Automated Net Actuarial Underwriting Balance (ANAUB) is calculated as the difference between revenue, expenses, and outgo:

$$ANAUB = \text{Revenue} - \text{Expenses} - \text{Outgo} \tag{48}$$

To ensure robustness, any negative ANAUB values are capped at zero using the pmax() function:

$$ANAUB_{\text{adjusted}} = \max(ANAUB, 0) \tag{49}$$

8-7-2- Visualization of ANAUB

The ANAUB for each underwriting cluster is visualized using a scatter plot, where the premiums are plotted on the x-axis and the ANAUB on the y-axis. The points are color-coded based on the respective underwriting clusters, providing an intuitive comparison of the ANAUB across different risk levels.

8-7-3- Simulating IFRS17 Metrics within Underwriting Clusters

The IFRS17-compliant actuarial metrics are then simulated to align the model with international financial reporting standards. For each policyholder, the Insurance Revenue is simulated as a random variation around the premium, within 90% to 110% of the premium:

$$\text{Revenue}_{\text{sim}} \sim U(0.90 \cdot \text{Premium}, 1.10 \cdot \text{Premium}) \tag{50}$$

Insurance Service Expense is simulated between 50% and 70% of the premium:

$$\text{Expense}_{\text{sim}} \sim U(0.50 \cdot \text{Premium}, 0.70 \cdot \text{Premium}) \tag{51}$$

Profitability Ratio is calculated as the difference between revenue and expense divided by the revenue:

$$\text{ProfitabilityRatio} = \frac{\text{Revenue} - \text{Expense}}{\text{Revenue}} \tag{52}$$

Other IFRS17-related metrics, such as Expense Ratio, Insurance Contract Liabilities (ICL), Present Value of Future Cash Flows, and Sensitivity Analysis, are simulated similarly using random variations of the premium.

A summarized table is created for each underwriting cluster to present the average values of these metrics. The IFRS17 metrics for each underwriting cluster are then visualized using a bar plot, with each metric represented on the x-axis, and the value of the metric plotted on the y-axis. The plot is color-coded based on underwriting clusters, offering an intuitive overview of the financial metrics and their variations across risk levels.

8-8- Model Implementation and Integration with AI

To integrate advanced machine learning techniques, particularly XGBoost, into the loss reserving and risk pricing processes, the study proposes an augmentation of the actuarial models with AI-driven data science. XGBoost is employed to refine the prediction of premiums, expenses, outgo, and other actuarial metrics. The integration of machine learning models into the actuarial framework enhances the accuracy and efficiency of loss reserving and risk pricing in occupational indemnity insurance. The mathematical representation of the AI-enhanced process is given by:

$$\hat{y} = X \cdot \beta + \epsilon \tag{53}$$

where $\hat{y}$ is the predicted actuarial metric (e.g., premium or loss reserve), X is the matrix of input features (e.g., policyholder characteristics), β represents the model coefficients, and ϵ is the error term.

This AI-driven approach allows for an enhanced actuarial data science framework, enabling more sophisticated and precise predictions in line with IFRS 17 requirements.

8-9- Model Evaluation

8-9-1- Robust Model Testing through Perturbation

To assess the stability and reliability of the Automated Actuarial Loss Reserves Risk Premiums (AALRRPs) under varying conditions, we employed a perturbation methodology. A perturbation factor of 1% was introduced by adding random noise to the original values, simulating slight variations in input data. This allows us to test the robustness of the model and assess the impact of minor deviations on the reserve values. Mathematically, the perturbed AALRRPs can be expressed as:

$$\hat{A}ALRRP_{perturbed} = \hat{A}ALRRP_{original} \times (1 + \epsilon)$$

where $\epsilon \sim \mathcal{N}(0, \sigma^2)$ is a random noise factor with zero mean and variance σ^2, representing the perturbation introduced to the system. The impact of these perturbations is visualized by comparing the original and perturbed AALRRPs through scatter plots and linear regression analysis. A regression model of the form:

$$y = \beta_0 + \beta_1 x + \epsilon$$

is fitted to the perturbed data, where y is the perturbed AALRRP and x is the original AALRRP. This allows for an assessment of the sensitivity of the model to slight input variations and the robustness of its estimates

8-9-2- Stress Testing for Scenario Analysis

Stress testing is performed to simulate extreme conditions that may affect the stability of the AALRRPs. Various stress factors are applied to the model, including introducing negative shocks to the AALRRPs by applying a reduction factor of 20% (i.e., multiplying by 0.8) for claims exceeding a certain threshold, τ. This process helps evaluate the model's sensitivity to adverse conditions, such as unexpected shifts in claim behavior or macroeconomic factors. The stress-tested AALRRPs are given by:

$$\hat{A}ALRRP_{stressed} = \begin{cases} \hat{A}ALRRP \times 0.8 & \text{if } \hat{A}ALRRP > \tau \\ \hat{A}ALRRP & \text{otherwise} \end{cases}$$

This equation applies the 20% reduction factor to the AALRRP values above the threshold τ, simulating the negative shock in the model. Stress-testing simulations are carried out for different distributions, such as Normal, Gamma, and Beta, and the results are visualized using line plots that show the evolution of both the original and stressed AALRRPs. Specifically, we define the cumulative distribution functions (CDFs) of the respective distributions as follows:

$$F_{\text{Normal}}(x) = \frac{1}{2}\left[1 + \text{erf}\left(\frac{x-\mu}{\sigma\sqrt{2}}\right)\right]$$

$$F_{\text{Gamma}}(x) = \frac{\gamma(\alpha, \frac{x}{\beta})}{\Gamma(\alpha)}$$

$$F_{\text{Beta}}(x) = I_x(\alpha, \beta)$$

where; μ and σ are the mean and standard deviation for the Normal distribution, $\gamma(\alpha, \frac{x}{\beta})$ is the lower incomplete Gamma function for the Gamma distribution, and $I_x(\alpha, \beta)$ is the regularized incomplete Beta function for the Beta distribution. These stress-testing distributions allow us to explore the impact of extreme conditions on the AALRRPs.

8-9-3- Scenario Modeling

Scenario testing is incorporated to simulate a range of possible future conditions affecting risk pricing and underwriting. These scenarios account for various market conditions and economic changes that could influence the insurance portfolio. Specifically, we consider the following adjustments to simulate future scenarios:

- *Inflation Adjustment*: A 5% increase in the claim amount was applied to simulate the impact of inflation on future claims. The inflation-adjusted AALRRPs are computed as:

$$\hat{A}ALRRP_{inflated} = \hat{A}ALRRP \times (1 + 0.05)$$

- *Claim Frequency Adjustments*: The frequency of claims varies by increasing or decreasing the claim frequency by 10%. This is represented as:

$$\hat{A}ALRRP_{freq\_adjusted} = \hat{A}ALRRP \times (1 + \delta)$$

where $\delta = 0.1$ represents the 10% adjustment to the claim frequency. Positive values of δ simulate an increase in claims, while negative values represent a decrease in claims.

The results of the scenario analysis are summarized using statistical measures such as the minimum, maximum, and mean values of the AALRRPs under each scenario. Specifically, we compute the following summary statistics:

$$\mu_{\hat{A}ALRRP} = \frac{1}{n}\sum_{i=1}^{n} \hat{A}ALRRP_i, \quad \text{Var}(\hat{A}ALRRP) = \frac{1}{n-1}\sum_{i=1}^{n} \left(\hat{A}ALRRP_i - \mu_{\hat{A}ALRRP}\right)^2$$

where $\mu_{\hat{A}ALRRP}$ is the mean of the AALRRPs, and $\text{Var}(\hat{A}ALRRP)$ is the variance of the AALRRPs across all simulated scenarios. These statistics allow for an in-depth understanding of the variability and risk profiles under different economic conditions and market changes.

The scenario modeling helps quantify potential risk in a dynamic insurance market, supporting the actuarial decision-making process by exploring future uncertainties through simulated data.

8-10-Novelty in the Methodology

The novelty of this methodology lies in the integration of AI-driven augmented actuarial data science techniques with the XGBoost algorithm for the estimation of inflation-adjusted frequency-severity models in occupational indemnity insurance, all while ensuring compliance with IFRS 17 standards. This probabilistic framework introduces several innovative aspects.

The methodology applies cutting-edge machine learning techniques, particularly XGBoost, to actuarial models for loss reserving, risk pricing, and underwriting. The framework goes beyond traditional actuarial methods by leveraging AI for more accurate, data-driven predictions. The framework includes a sophisticated simulation of occupational indemnity insurance policies, encompassing a wide array of policyholder characteristics, claims data, inflation adjustments, and underwriting information. This robust dataset of 100,000 policyholders serves as a foundation for model training, allowing for a more representative and diverse set of data inputs. Unlike traditional actuarial models that focus primarily on static data, this framework incorporates inflation-adjusted reserves and premiums, offering a dynamic approach to account for the impact of inflation on future claims. The inflation adjustment is seamlessly integrated into the loss reserving and risk pricing models, providing a more realistic and forward-looking view of risk. The use of XGBoost for both the Automated Actuarial Loss Reserving (ALR) and Automated Actuarial Risk Pricing (ARP) models represents a novel approach in actuarial modeling. This method is known for its superior predictive power and efficiency, making it well-suited for complex actuarial datasets with large numbers of features. The methodology incorporates a comprehensive evaluation and validation strategy using key performance metrics such as MAE, MSE, and RMSE, along with residual analysis and scatter plots. These validation techniques ensure that the models provide reliable and accurate predictions, a critical feature for practical use in the actuarial field. The paper ensures compliance with IFRS 17 by simulating essential IFRS 17 metrics such as the Contractual Service Margin (CSM), Fulfilment Cash Flows (FCF), and Liability for Incurred Claims (LIC). These metrics are central to the new regulatory framework and are integrated into the loss reserving and risk pricing models, providing a complete solution for IFRS 17 compliance. A key innovation is the development of an Automated Actuarial Underwriting Model, which segments policyholders into distinct underwriting clusters based on their Automated Actuarial Loss Reserve Risk Premiums (AALRRPs). This segmentation allows for tailored underwriting decisions and risk-based pricing, which enhances the accuracy and efficiency of the underwriting process. The methodology uniquely simulates a range of financial metrics including the profitability ratio, insurance service expense, and other IFRS 17-related metrics within underwriting clusters. This provides a granular view of the financial health of the insurance contracts, which is critical for ensuring long-term profitability and regulatory compliance.

In closing, the novelty of this methodology lies in its integration of machine learning for advanced actuarial analysis, its dynamic treatment of inflation, and its compliance with IFRS 17, making it a cutting-edge approach for occupational indemnity insurance modeling.

9- Data

The simulated data generated forms a robust foundation for modeling the inflation-adjusted frequency-severity dynamics in the context of occupational indemnity insurance, adhering to IFRS 17 regulations. Below is a breakdown of the components of the dataset and their contributions to the development of the model in this paper.

9-1- Policy Information

- *Policy ID*: A unique identifier for each insurance policy, represented as $\mathcal{P}_i$ for policy i, facilitating the tracking and management of policy-specific data.
- *Industry*: Denoted as $\mathcal{I}$, representing the type of industry (e.g., Construction, Healthcare, Manufacturing, IT, Education). The industry classification $\mathcal{I} \in \{1,2,\dots,N\}$ plays a critical role in determining the risk exposure $\mathcal{R}_\mathcal{I}$ associated with the policyholder. The industry-specific claims experience impacts the risk-adjusted pricing and underwriting.
- *Occupation Risk Level*: The occupation risk level $\mathcal{R}_o \in \{\text{Low}, \text{Medium}, \text{High}, \text{Very High}\}$ impacts the frequency and severity of claims, influencing the probability distribution of claims as a function of occupation risk. Let $f(\mathcal{R}_o)$ represent the risk-adjusted claim frequency, with a relationship defined as

$$f(\mathcal{R}_o) = \lambda_o \cdot \mathbb{1}_{\mathcal{R}_o}, \quad \lambda_o \in \mathbb{R}^+ \quad \text{and} \quad \mathbb{1}_{\mathcal{R}_o} \in \{0,1\}$$

- *Annual Salary*: Let $\mathcal{S}_i$ denote the annual salary of policyholder i, which influences the base premium and severity of claims. The annual salary is typically assumed to follow a normal distribution:

$$\mathcal{S}_i \sim \mathcal{N}(\mu_\mathcal{S}, \sigma_\mathcal{S}^2)$$

- *Policy Term Years*: The policy duration is given by T_i for policyholder i, representing the time over which claims are expected to occur. The policy term is associated with a time-discounting function $D(T_i)$ for the estimation of the present value of future claims:

$$D(T_i) = e^{-rT_i}$$

where r is the annual discount rate.

- *Policy Type*: The policy type $\mathcal{P}_t \in \{\text{Basic}, \text{Standard}, \text{Premium}\}$ directly influences the expected claims payout, as the coverage level is typically modeled using a discrete distribution:

$$\mathcal{P}_t \sim \text{Discrete}(\mathcal{P}_t | p_1, p_2, p_3)$$

where p_1, p_2, p_3 correspond to the probabilities of Basic, Standard, and Premium types, respectively.

- *Coverage Limit*: Denoted as $\mathcal{C}_i$, this defines the maximum payout under the policy. The coverage limit serves as an upper bound in the severity distribution and is critical for the calculation of tail risk:

$$\mathcal{L}_i = \min(X_i, \mathcal{C}_i)$$

where X_i is the random variable representing the claim severity.

- *Deductible*: The deductible $\mathcal{D}_i$ represents the policyholder's out-of-pocket expenses before claims are paid by the insurer. The deductible influences the severity modeling by truncating the claim severity distribution:

$$X'_i = \max(X_i - \mathcal{D}_i, 0)$$

where X'_i is the adjusted claim severity after applying the deductible.

9-2- Claims Information

- *Claim Frequency*: Let N_i represent the number of claims filed by policyholder i. The claim frequency is modeled using a Poisson distribution:

$$N_i \sim \text{Poisson}(\lambda_i), \quad \lambda_i = \alpha \cdot \mathcal{R}_i$$

where α is a scaling factor and $\mathcal{R}_i$ represents the risk factor of policyholder i.

- *Claim Severity*: The claim severity is modeled using a Gamma distribution, where the severity of each claim S_i follows:

$$S_i \sim \text{Gamma}(\kappa, \theta)$$

where κ is the shape parameter and θ is the scale parameter. The expected severity is given by:

$$\mathbb{E}[S_i] = \kappa\theta$$

- *Claim Status*: Denoted by $\mathcal{C}_{\text{status}}$, the status of a claim can be either "Open" or "Closed," influencing the development of incurred but not reported (IBNR) reserves. The claim development is modeled as:

$$\mathcal{C}_{\text{status}} = \begin{cases} \text{Open} & \text{if claim development time} < T \\ \text{Closed} & \text{if claim development time} \geq T \end{cases}$$

- *Loss Reserving*: The loss reserves R_i are computed as the expected future claims, taking into account the discount factor and inflation adjustments. The reserve for policyholder i is given by:

$$R_i = \mathbb{E}[\mathcal{L}_i] \cdot e^{-rT_i}$$

where T_i is the time horizon and r is the discount rate.

9-3- Premiums and Pricing

Base Premium: The base premium $\mathcal{P}_{\text{base}}$ is calculated as a fixed percentage of the policyholder's salary:

$$\mathcal{P}_{\text{base}} = \alpha \cdot \mathcal{S}_i$$

where α is a constant factor, typically 3% of the salary.

- *Risk Score*: The risk score $\mathcal{R}_{\text{score}}$ is a function of various policyholder attributes. It quantifies the overall risk, incorporating industry, occupation risk level, and salary. The risk score is modeled as:

$$\mathcal{R}_{\text{score}} = \gamma_1 \mathcal{I}_i + \gamma_2 \mathcal{R}_o + \gamma_3 \mathcal{S}_i$$

where $\gamma_1, \gamma_2, \gamma_3$ are coefficients that determine the weight of each factor.

- *Risk Adjusted Premium*: The risk-adjusted premium $\mathcal{P}_{\text{adj}}$ is given by:

$$\mathcal{P}_{\text{adj}} = \mathcal{P}_{\text{base}} \cdot \left(1 + \frac{\mathcal{R}_{\text{score}}}{100}\right)$$

where the risk score is used to adjust the premium based on the policyholder's risk profile.

9-4- Inflation and Adjustment

- *Inflation Rate*: The annual inflation rate $\mathcal{I}_{\text{rate}}$ is modeled as a random variable drawn from a normal distribution:

$$\mathcal{I}_{\text{rate}} \sim \mathcal{N}(\mu_{\mathcal{I}}, \sigma_{\mathcal{I}}^2)$$

- *Inflation Adjusted Reserves*: The inflation-adjusted reserves $\mathcal{R}_{\text{inflated}}$ are calculated by multiplying the reserves by the inflation factor:

$$\mathcal{R}_{\text{inflated}} = R_i \cdot (1 + \mathcal{I}_{\text{rate}})^{T_i}$$

which accounts for the increase in claims severity due to inflation.

9-5- Underwriting Details

- *Underwriting Score*: The underwriting score $\mathcal{U}_i$ is a function of various risk factors, including occupation, industry, and claims history:

$$\mathcal{U}_i = \sum_{j=1}^{n} \theta_j x_{ij}$$

where x_{ij} are the risk factors associated with policyholder i and θ_j are the corresponding coefficients.

- *Underwriting Decision*: The underwriting decision $\mathcal{D}_i$ is determined based on a threshold $\mathcal{T}_{\text{underwriting}}$:

$$\mathcal{D}_i = \begin{cases} \text{Approved} & \text{if } \mathcal{U}_i \geq \mathcal{T}_{\text{underwriting}} \\ \text{Declined} & \text{if } \mathcal{U}_i < \mathcal{T}_{\text{underwriting}} \end{cases}$$

9-6- Contribution to the Model development

The simulated data allows for the exploration of various features such as risk level, industry type, and policy characteristics that influence both claim frequency and severity. Feature engineering based on these variables can help in identifying significant predictors of claim outcomes. The data simulates real-world variations in policyholder risk profiles (e.g., *Occupation Risk Level*, *Risk Score*, *Annual Salary*), helping to segment the portfolio and estimate risk-adjusted premiums more accurately. This segmentation is crucial for setting appropriate pricing strategies. With inflation modeled and incorporated into reserves and premiums, this data allows for the testing of inflation-adjusted models, ensuring compliance with IFRS 17 standards and the effectiveness of the inflation-adjusted frequency-severity framework. The simulated loss reserves (both original and inflation-adjusted) are essential for developing the actuarial loss reserving component of the model. The inflation adjustment adds realism by considering future claims escalation, which is central to your IFRS 17-compliant framework. The data generated here can serve as input for advanced machine learning models like XGBoost. XGBoost can be applied to predict claim frequency and severity, optimize loss reserving, and perform underwriting decision analysis based on policyholder attributes.

In short, this dataset plays a key role in developing a comprehensive probabilistic framework for inflation-adjusted frequency-severity modeling in occupational indemnity insurance, supporting the advanced loss reserving, risk pricing, and underwriting mechanisms described in this paper. The integration of inflation adjustments, risk-based premium calculations, and underwriting decisions enhances the actuarial accuracy and compliance with IFRS 17 in the context of modern actuarial practices.

10- Results

The section presents the findings and outcome for this study.

10-1- Exploratory Data Analysis

Exploratory Data Analysis (EDA) is an approach to analyzing and summarizing datasets to gain insights, identify patterns, detect anomalies, and test assumptions with the help of statistical graphics, plots, and information tables. EDA is typically the first step in data analysis, used to understand the structure of the data before formal modeling or hypothesis testing is carried out. Key techniques in EDA include data visualization (such as histograms, box plots, scatter plots), summary statistics (mean, median, standard deviation), and checking for missing values or outliers [33-36].

Figure 1 visualizes the distribution of insurance policies across various industries (Construction, Healthcare, Manufacturing, IT, Education). This plot provides a foundation for examining industry-specific risk factors or claims behavior. Figure 2 shows how annual salary varies by occupation risk level (Low, Medium, High, Very High). Higher risk occupations likely have higher average salaries due to the nature of the work, which can be a critical factor in determining premium pricing. The distribution of salary by risk level is significant for actuarial modeling, especially for the Inflation-Adjusted Frequency-Severity Model. It links policyholder compensation (annual salary) to their exposure to occupational risk, which can be used to adjust the premium pricing based on risk levels.

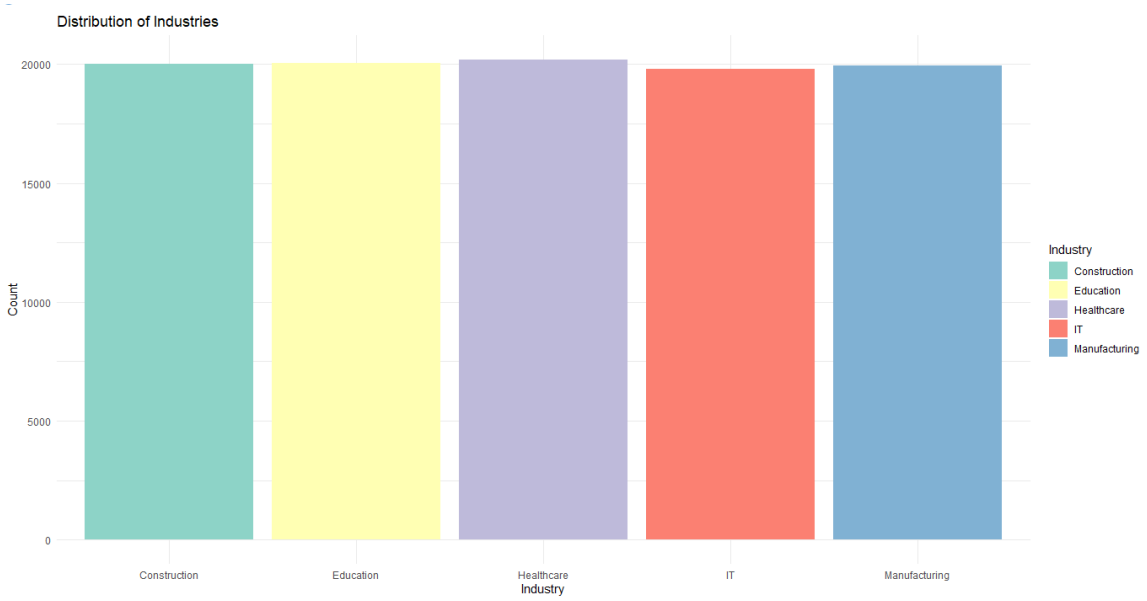

Figure 1. Industry Distribution Plot

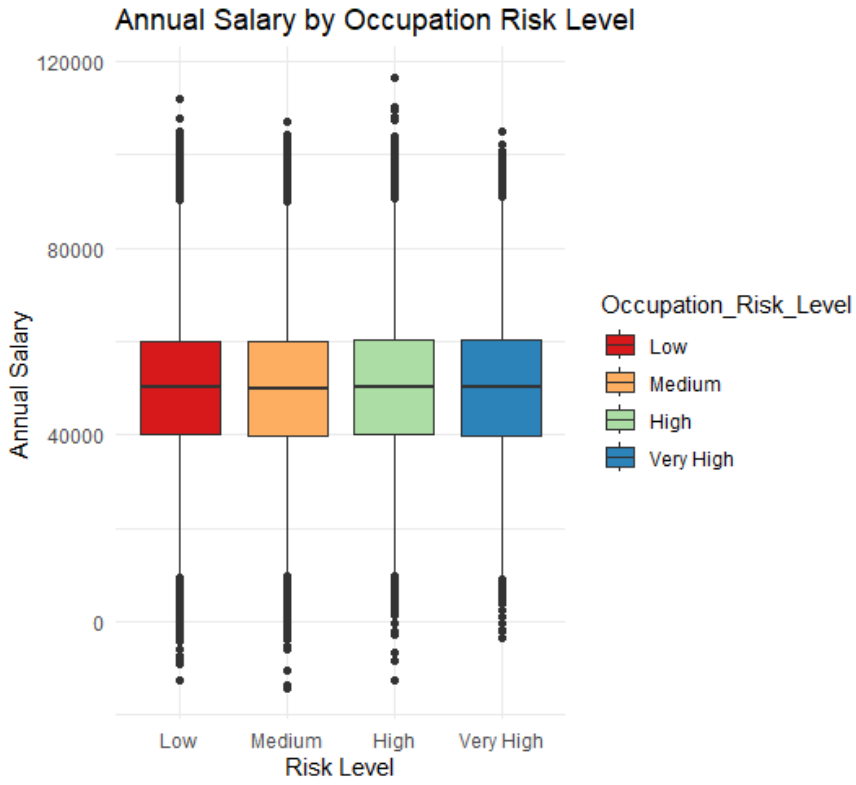

Figure 2. Annual Salary by Occupation Risk Level

10-1-1- Correlation Analysis

Correlation analysis is a statistical technique used to measure and describe the strength and direction of the relationship between two or more variables. It determines how variables are related to each other, whether they move in the same direction (positive correlation), in opposite directions (negative correlation), or if there is no relationship at all. The most common method of correlation analysis is the Pearson correlation coefficient, which quantifies the linear relationship between variables, ranging from -1 to 1. A value of 1 indicates a perfect positive correlation, -1 indicates a perfect negative correlation, and 0 indicates no correlation [37-39].

The correlation matrix presented by Figure 3 provides an overview of how various numeric variables (e.g., annual salary, claim frequency, loss reserving, premiums) are related. Strong positive or negative correlations can help identify which variables influence others. Understanding correlations between features like annual salary, claim frequency, and severity allows for better feature selection and more informed variable inclusion in the model.

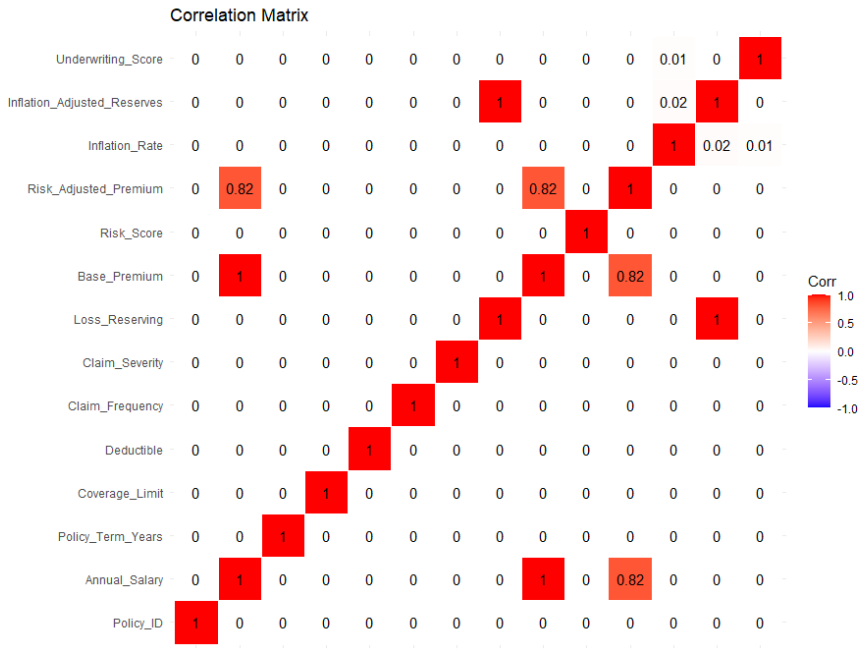

Figure 3. Correlation Matrix Plot

10-1-2- t-SNE for Dimensionality Reduction

Dimensionality reduction is a crucial technique in the field of machine learning, especially when dealing with high-dimensional data sets. Among the various methods, t-Distributed Stochastic Neighbor Embedding (t-SNE) has garnered significant attention for its ability to visualize high-dimensional data in lower-dimensional spaces, typically in 2 or 3 dimensions. This technique is particularly useful in visualizing the structure of complex datasets while preserving the local structure of the data points [40, 41].

Given a dataset $\mathbf{X} = \{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n\} \subset \mathbb{R}^d$ with n points, the goal of dimensionality reduction is to map these points to a lower-dimensional space while preserving the pairwise similarities. Let P_{ij} represent the conditional probability of selecting point j as a neighbor of point i in the high-dimensional space.

The high-dimensional pairwise similarity between points $\mathbf{x}_i$ and $\mathbf{x}_j$ is modeled using a Gaussian distribution:

$$p_{ij} = \frac{\exp\left(-\|\mathbf{x}_i - \mathbf{x}_j\|^2 / 2\sigma_i^2\right)}{\sum_{k \neq i} \exp(-\|\mathbf{x}_i - \mathbf{x}_k\|^2 / 2\sigma_i^2)}$$

where σ_i is a bandwidth parameter that controls the variance of the Gaussian distribution around $\mathbf{x}_i$. This distribution models the similarity between point i and its neighbors in the high-dimensional space. The normalized version of p_{ij} ensures that the sum of similarities for each point i is 1, i.e., $\sum_{j \neq i} p_{ij} = 1$.

To map the data to a lower-dimensional space, we define a low-dimensional embedding $\mathbf{Y} = \{\mathbf{y}_1, \mathbf{y}_2, \ldots, \mathbf{y}_n\} \subset \mathbb{R}^m$, where m is the target dimension (typically $m = 2$ or 3). The pairwise similarity between points $\mathbf{y}_i$ and $\mathbf{y}_j$ in the low-dimensional space is defined using the Student's t-distribution with one degree of freedom (also known as the Cauchy distribution):

$$q_{ij} = \frac{(1 + \|\mathbf{y}_i - \mathbf{y}_j\|^2)^{-1}}{\sum_{k \neq i} (1 + \|\mathbf{y}_i - \mathbf{y}_k\|^2)^{-1}}$$

This distribution ensures that the similarity between points in the low-dimensional space decreases with the distance between them, and it has heavier tails compared to a Gaussian, which helps in avoiding the "crowding problem" often encountered in dimensionality reduction methods.

The objective of t-SNE is to find the low-dimensional representation $\mathbf{Y}$ that minimizes the difference between the high-dimensional and low-dimensional similarities. This difference is measured using the Kullback-Leibler (KL) divergence, which quantifies the dissimilarity between two probability distributions:

$$KL(P \parallel Q) = \sum_{i \neq j} p_{ij} \log\left(\frac{p_{ij}}{q_{ij}}\right)$$

where P represents the high-dimensional probability distribution and Q the low-dimensional probability distribution. The KL divergence is minimized with respect to $\mathbf{Y}$ to obtain the best low-dimensional embedding. The cost function for t-SNE is given by:

$$C(\mathbf{Y}) = \sum_{i \neq j} p_{ij} \log\left(\frac{p_{ij}}{q_{ij}}\right)$$

This cost function encourages points that are similar in high-dimensional space to remain close in low-dimensional space, while points that are dissimilar are pushed further apart.

Figure 4 reduces the dimensionality of the numeric data and visualizes it in two dimensions. Points are colored by occupation risk level. t-SNE helps in understanding if there are clear patterns or clusters based on occupation risk level, even in higher-dimensional space. t-SNE provides insight into the structure of the data, which can inform how different features group together. Figure 5 takes the t-SNE results and adds a third dimension (Risk Score) to the 3D space. This allows for a more comprehensive view of how risk score, occupation risk level, and the other features relate spatially.

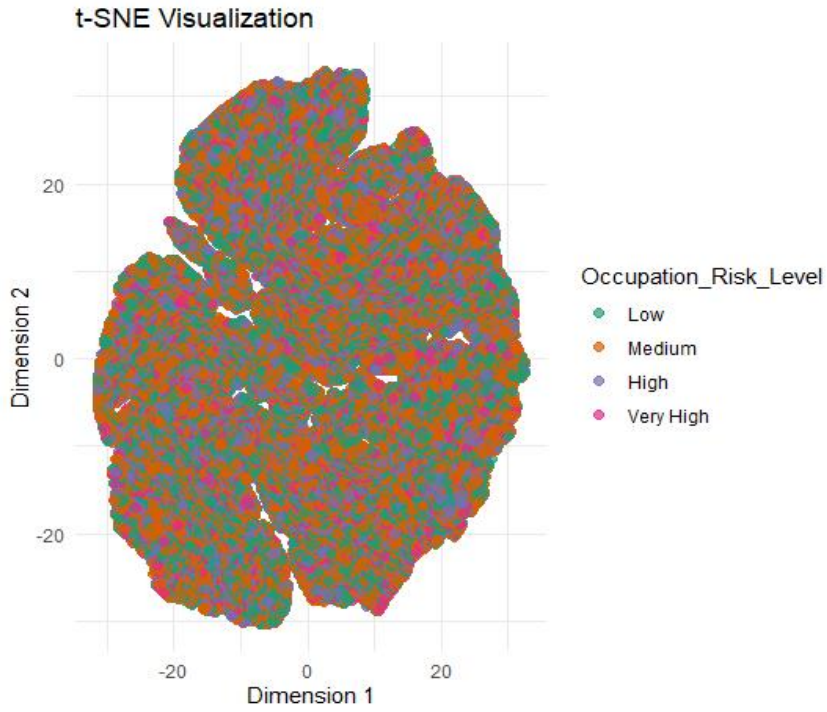

Figure 4. 2D Plot:t-SNE for Dimensionality Reduction

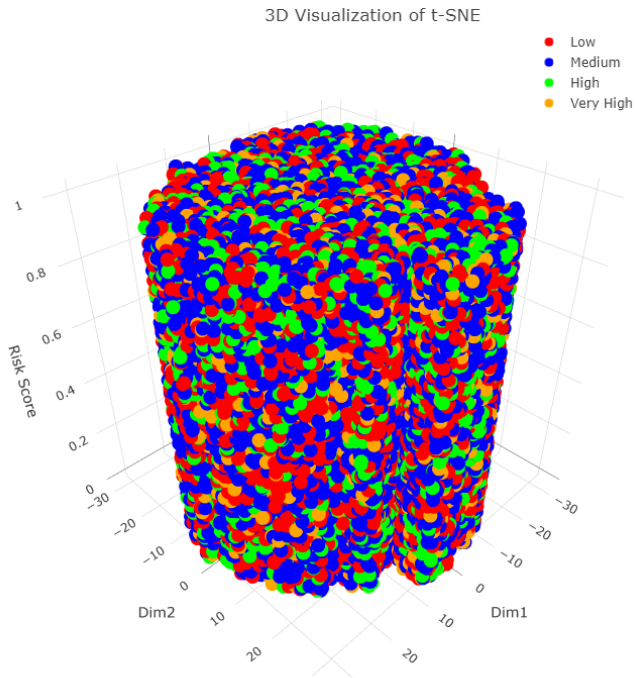

Figure 5. 3D Plot:t-SNE for Dimensionality Reduction

10-1-3- Cluster Analysis

Cluster analysis, also known as clustering, is a statistical technique used to group a set of objects in such a way that objects in the same group (or cluster) are more similar to each other than to those in other groups. The goal of cluster analysis is to explore the inherent structure of the data, typically in an unsupervised manner, by dividing a dataset into subsets that exhibit similar characteristics. It is widely used in data analysis, machine learning, and various scientific fields to uncover hidden patterns in data [42-45].

The k-means clustering analysis groups the data into four clusters based on the numeric features as presented by Figure 6. Each cluster corresponds to a set of observations with similar characteristics (e.g., risk scores, claim severity). The cluster plot visualizes how these groups differ and overlap. Clustering analysis helps identify subgroups within the data that may require different treatment in the model.

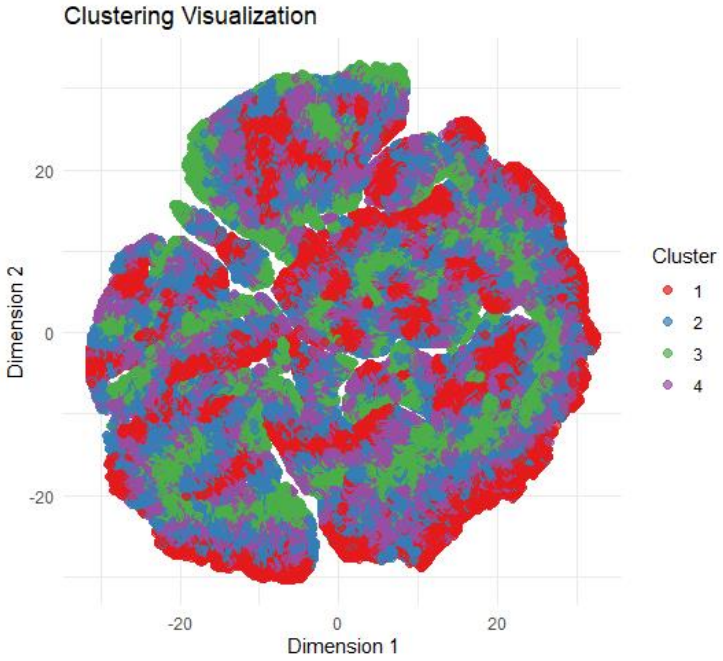

Figure 6. Cluster analysis plot

10-2- The Brighton Mahohoho XGBoost Probabilistic Framework for Inflation-Adjusted Frequency-Severity Modelling

To split the dataset into training and test sets, a random sampling technique is applied:

$$\text{index} \sim \text{Uniform}(1, n), \text{where } n = \text{number of rows in data}, \tag{54}$$

$$\text{train\_data} = \text{data}[\text{index},], \tag{55}$$

$$\text{test\_data} = \text{data}[-\text{index},]. \tag{56}$$

Here, 80% of the data is allocated to the training set, and the remaining 20% to the test set.

Categorical variables are converted into numerical representations using the $as.numeric$ function. Let C be a categorical variable; its transformation is given by:

$$C_{\text{encoded}} = \text{as.numeric}(\text{as.factor}(C)). \tag{57}$$

For instance, the Industry column is encoded as follows:

$$\text{data\$Industry} \rightarrow \text{as.numeric}(\text{as.factor}(\text{data\$Industry})). \tag{58}$$

This encoding ensures compatibility with the XGBoost algorithm.

The goal is to predict Y_{ALR}, the Inflation Adjusted Reserves. Let $\mathbf{X}_{\text{ALR}}$ represent the feature matrix for ALR, and $\mathbf{Y}_{\text{ALR}}$ the target vector. The data preparation for ALR is as follows:

$$\mathbf{X}_{\text{ALR}} = \textbf{data}[\text{features excluding Inflation Adjusted Reserves and Risk Adjusted Premium}], \tag{59}$$

$$\mathbf{Y}_{\text{ALR}} = \textbf{data}[\text{Inflation Adjusted Reserves}] \tag{60}$$

The XGBoost regression model minimizes the squared error objective function:

$$\mathcal{L}(\Theta) = \sum_{i=1}^{n} \left(Y_i - \hat{Y}_i(\Theta)\right)^2 + \lambda \parallel \Theta \parallel^2, \tag{61}$$

where Θ represents the model parameters.

The model is trained using 100 boosting rounds:

$$\text{model\_ALR} = \text{xgboost}(\text{data} = \mathbf{X}_{\text{ALR}}, \text{label} = \mathbf{Y}_{\text{ALR}}, \text{objective} = ``\text{reg: squarederror}'', \text{nrounds} = 100). \tag{62}$$

Similarly, to predict Y_{ARP}, the Risk Adjusted Premiums, we define:

$$\mathbf{X}_{\text{ARP}} = \textbf{data}[\text{features excluding Risk Adjusted Premium and Inflation Adjusted Reserves}], \tag{63}$$

$$\mathbf{Y}_{\text{ARP}} = \textbf{data}[\text{Risk Adjusted Premium}]. \tag{64}$$

The ARP model uses the same squared error objective as the ALR model:

$$\mathcal{L}(\Theta) = \sum_{i=1}^{n} \left(Y_i - \hat{Y}_i(\Theta)\right)^2 + \lambda \parallel \Theta \parallel^2. \tag{65}$$

The training process is given by:

$$\text{model\_ARP} = \text{xgboost}(\text{data} = \mathbf{X}_{\text{ARP}}, \text{label} = \mathbf{Y}_{\text{ARP}}, \text{objective} = ``\text{reg: squarederror}'', \text{nrounds} = 100). \tag{66}$$

To evaluate model performance, the following metrics are computed:

• Mean Absolute Error (MAE): $\text{MAE} = \frac{1}{n}\sum_{i=1}^{n} \left|Y_i - \hat{Y}_i\right|$.

• Mean Squared Error (MSE): $\text{MSE} = \frac{1}{n}\sum_{i=1}^{n} \left(Y_i - \hat{Y}_i\right)^2$.

• Root Mean Squared Error (RMSE): $\text{RMSE} = \sqrt{\text{MSE}}$.

The residuals are analyzed to assess the model's predictive accuracy:

$$\text{Residuals} = \mathbf{Y}_{\text{actual}} - \mathbf{Y}_{\text{predicted}}. \tag{67}$$

Plots such as *Actual vs Predicted* and *Residuals vs Predicted* are utilized to visually evaluate model performance.

The Table 2 assesses the performance and configuration of two XGBoost regression models used for Actuarial Loss Reserving Risk Premium analysis. Both models have comparable computational efficiency, making them practical for large-scale actuarial applications.

Table 2. The Brighton Mahohoho XGBoost Inflation-Adjusted Frequency-Severity Model

Automated Actuarial Loss Reserving Risk Premium Model		
	ALR Model	**ARP Model**
Processing time (seconds)	5.69	5.33
Hyper parameters		
R package: XGBoost		
Type	***Regression***	***Regression***
Objective function	reg:squarederror	reg:squarederror
Train data sample size	80,000	80,000
Test data sample size	20,000	20,000
nrounds	100	100
verbose	0	0
Model Validation Metrics		
MAE	16.212480	601.789500
MSE	452.631800	494,274.000000
RMSE	21.275150	703.046300

10-2-1- Estimation of Automated Actuarial Loss Reserves

Figure 7 shows the relationship between the actual inflation-adjusted reserves (from the test data) and the predicted reserves (from the XGBoost model). The blue points represent individual predictions and they lie close to the red dashed line (45-degree diagonal line), which shows the better the predictions are. Figure 8 compares the actual risk-adjusted premiums against the predicted risk-adjusted premiums.

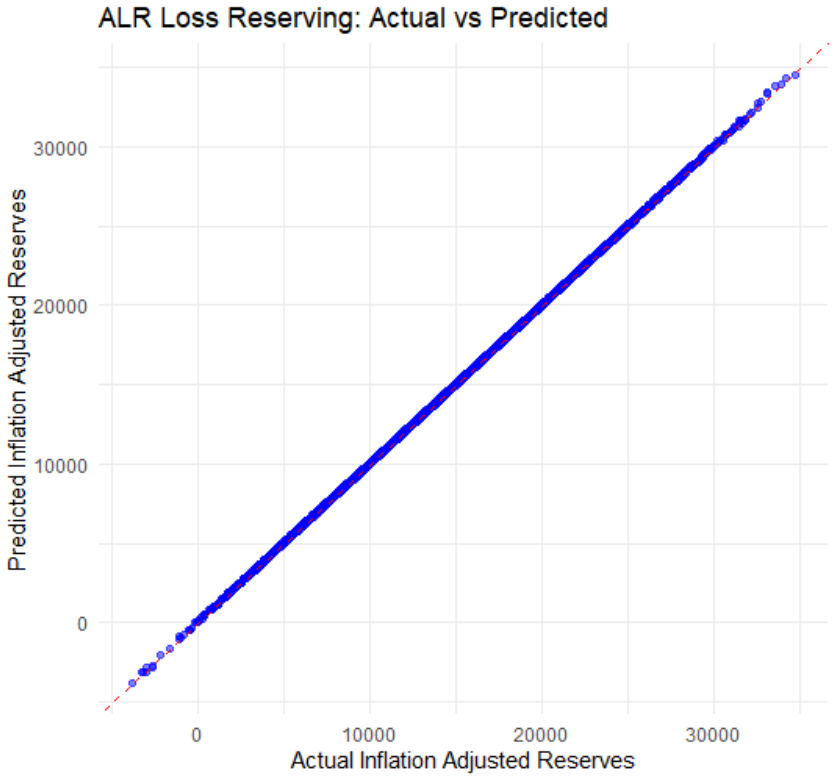

Figure 7. ALR Loss Reserving: Actual vs Predicted

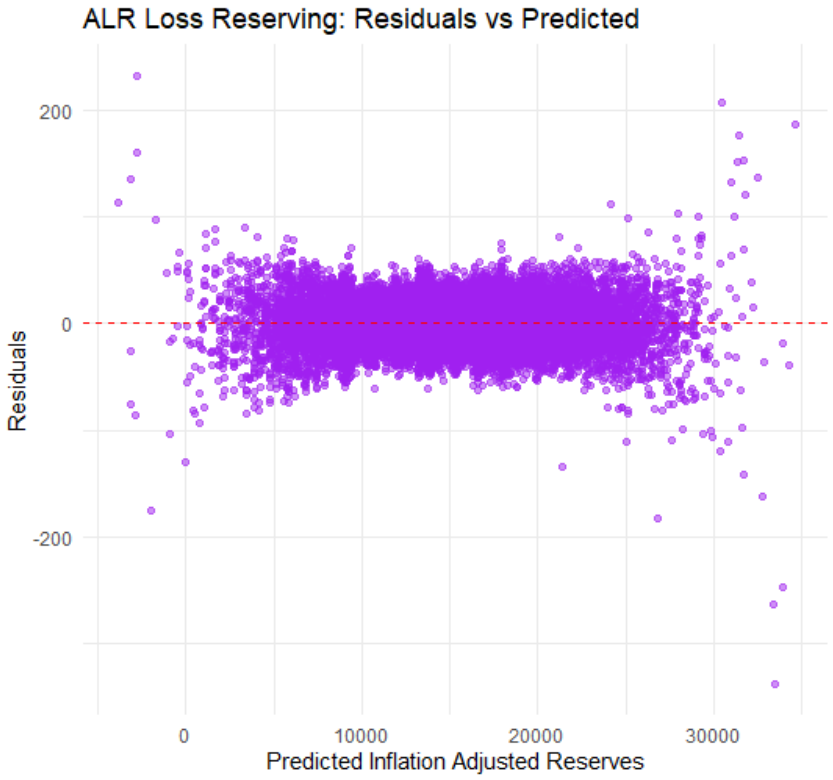

Figure 8. ALR Residual Plot

10-2-2- Estimation of Automated Actuarial Risk Premiums

Figure 9 visualizes the relationship between the actual risk-adjusted premiums and the predicted premiums. Figure 10 visualizes the residuals (the differences between the actual and predicted values) against the predicted values.

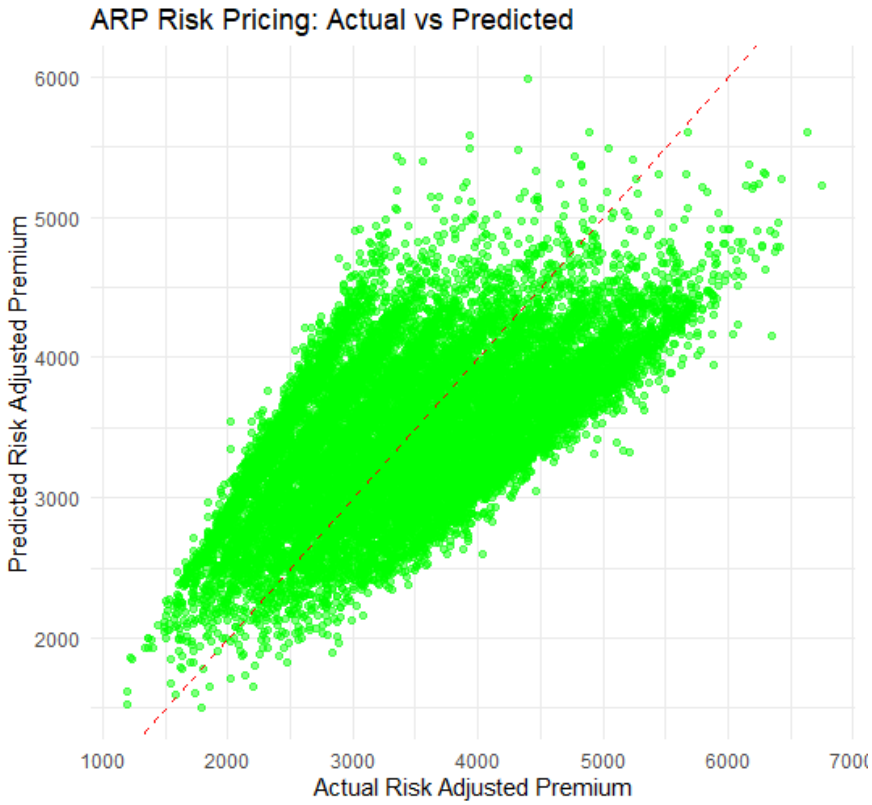

Figure 9. ARP Risk Pricing: Actual vs Predicted

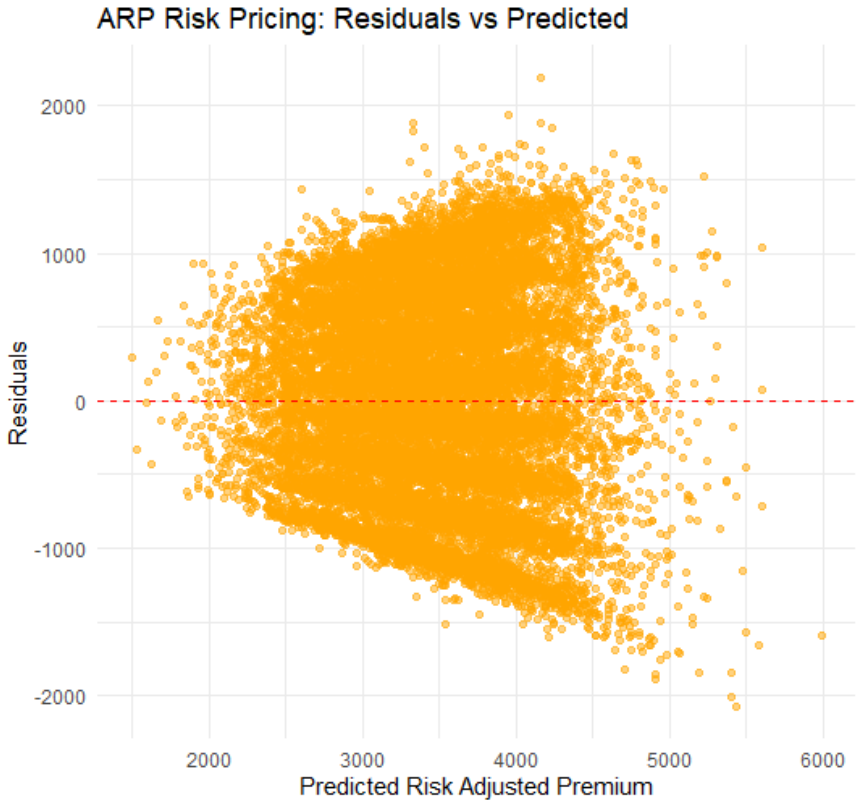

Figure 10. ARP Residual Plot

10-2-3- Estimation of Automated Actuarial Loss Reserve Risk Premiums (AALRRPs)

This section provides a mathematical description of the estimation of Automated Actuarial Loss Reserve Risk Premiums (AALRRPs). The real inflation rates, denoted as I_{real}, are calculated as the ratio of current inflation rates I_{current} to previous inflation rates I_{previous}:

$$I_{\text{real}} = \frac{I_{\text{current}}}{I_{\text{previous}}}, \quad I_{\text{previous}} \in [1.001, 1.005], \; I_{\text{current}} \in [1.01, 1.105]. \tag{68}$$

The resulting I_{real} values range between:

$$\min(I_{\text{real}}) \approx 1.005, \quad \max(I_{\text{real}}) \approx 1.10. \tag{69}$$

The AALRRPs are estimated as a weighted linear combination of predictions from the Automated Loss Reserving (ALR) model (P_{ALR}) and the Automated Risk Pricing (ARP) model (P_{ARP}), adjusted for real inflation rates:

$$\text{AALRRP} = w_{\text{ALR}} P_{\text{ALR}} + w_{\text{ARP}} P_{\text{ARP}} \cdot I_{\text{real}}, \tag{70}$$

where w_{ALR} and w_{ARP} are the weights assigned to the ALR and ARP predictions, respectively, satisfying:

$$w_{\text{ALR}} = 0.6, \quad w_{\text{ARP}} = 0.4. \tag{71}$$

Probabilities are simulated from three distributions: Normal, Gamma, and Beta. Let n denote the number of simulations. The probability density functions for each distribution are:

$$\text{Normal:} \quad f_N(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}, \quad \mu = 1, \; \sigma = 0.5, \; x \in [0,2], \tag{72}$$

$$\text{Gamma:} \quad f_G(x) = \frac{x^{k-1} e^{-x/\theta}}{\theta^k \Gamma(k)}, \quad k = 2, \; \theta = 1, \; x \in [0,3], \tag{73}$$

$$\text{Beta:} \quad f_B(x) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} x^{\alpha-1}(1-x)^{\beta-1}, \quad \alpha = 2, \; \beta = 5, \; x \in [0,8]. \tag{74}$$

These probabilities are normalized to ensure their total sum equals 1:

$$\hat{f}(x) = \frac{f(x)}{\sum_{i=1}^{n} f(x_i)}. \tag{75}$$

For each probability distribution, AALRRPs are computed by multiplying the simulated probabilities with the base AALRRPs:

$$\text{AALRRP}_D = \text{AALRRP} \cdot \hat{f}_D(x), \quad D \in \{\text{Normal}, \text{Gamma}, \text{Beta}\}. \tag{76}$$

The expected value for each distribution is calculated as:

$$\mathbb{E}[\text{AALRRP}_D] = \sum_{i=1}^{n} \text{AALRRP}_{D,i}. \tag{77}$$

The optimal distribution is selected based on the highest expected value:

$$D_{\text{chosen}} = \arg \max_{D \in \{\text{Normal,Gamma,Beta}\}} \mathbb{E}[\text{AALRRP}_D]. \tag{78}$$

To ensure that all AALRRP values are non-negative, negative values are adjusted to zero:

$$\text{AALRRP}_{\text{adjusted}} = \max(\text{AALRRP}, 0). \tag{79}$$

Finally, the maximum and minimum adjusted AALRRPs are calculated:

$$\max(\text{AALRRP}_{\text{adjusted}}), \quad \min(\text{AALRRP}_{\text{adjusted}}) \tag{80}$$

Figure 11 visualizes the trend of the Automated Actuarial Loss Reserve Risk Premiums (AALRRPs) over time and in addition to that Figure 12 visualizes AALRRPs over time with a color gradient based on the AALRRP values. Low values are shaded in blue, and high values are in red. The gradient makes it easier to visually identify outliers, peaks, or clusters of high and low values. This plot provides a detailed view of data dispersion over time and helps pinpoint periods

with extreme risk premiums. Observing the distribution and spread over time can reveal correlations between time and risk premium magnitudes. Figure 13 shows the frequency distribution of AALRRPs. The x-axis represents AALRRP values, while the y-axis indicates their frequency. Figure 14 visualizes AALRRP values over a simulated grid. The x-axis represents time (in months), the y-axis represents AALRRP values, and the z-axis represents their corresponding magnitudes. The surface plot adds a third dimension, offering a spatial perspective of AALRRPs over time and revealing multi-dimensional patterns. High and low areas on the surface indicate extreme AALRRP values. Peaks represent high-risk scenarios, while valleys indicate periods of relative stability. The surface smoothness reflects stability in premiums over time. A highly irregular surface indicates volatile periods.

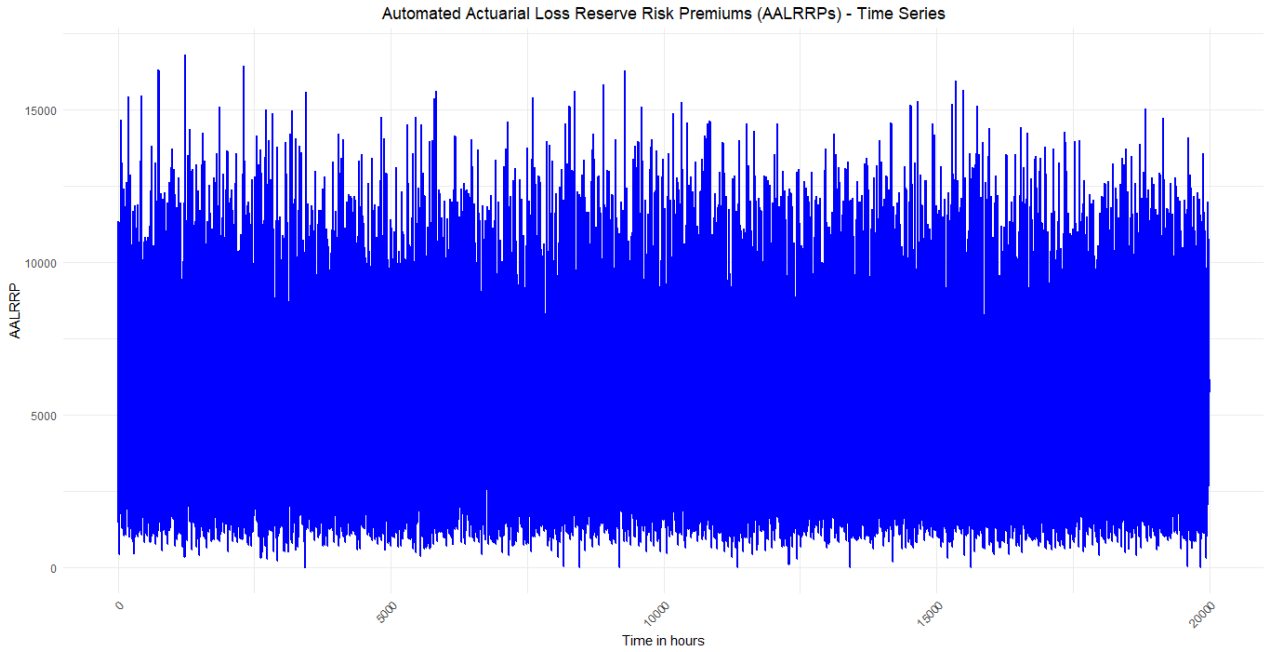

Figure 11. Time Series Line Plot

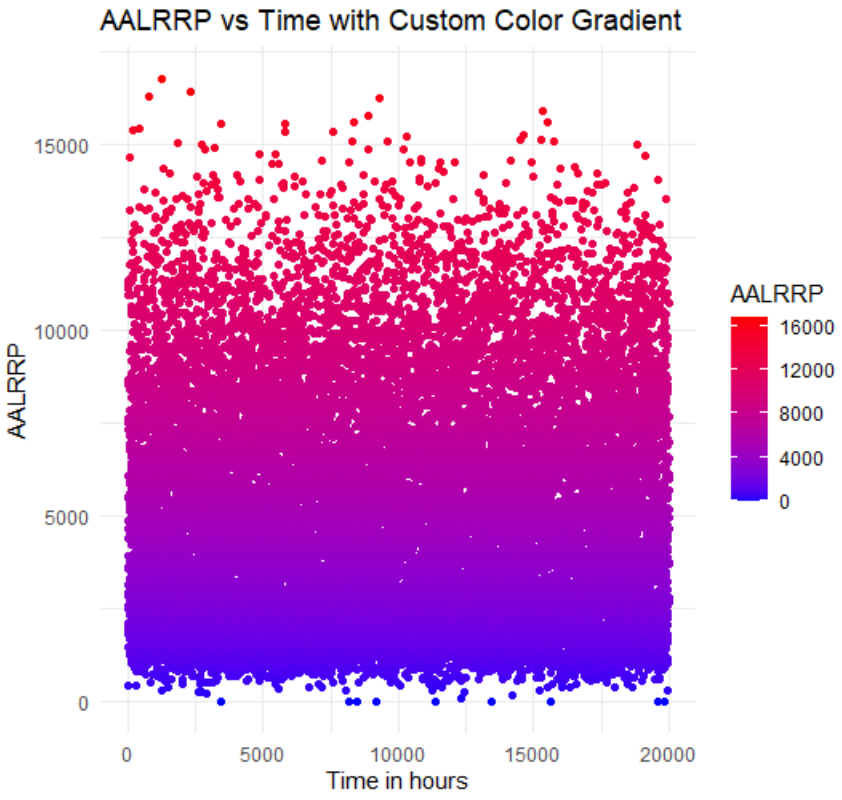

Figure 12. AALRRP vs. Time with Custom Color Gradient

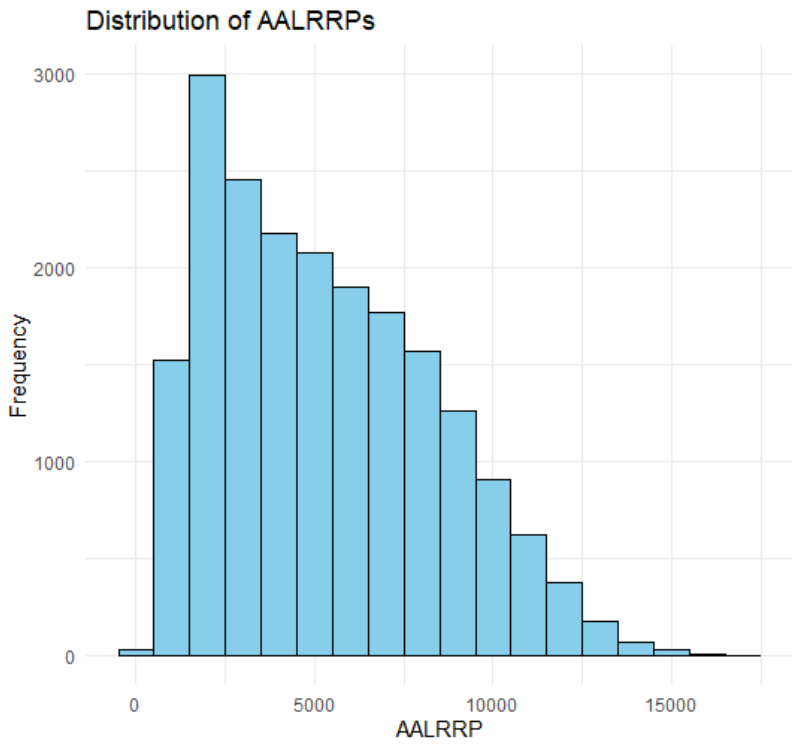

Figure 13. Distribution Plot of AALRRPs

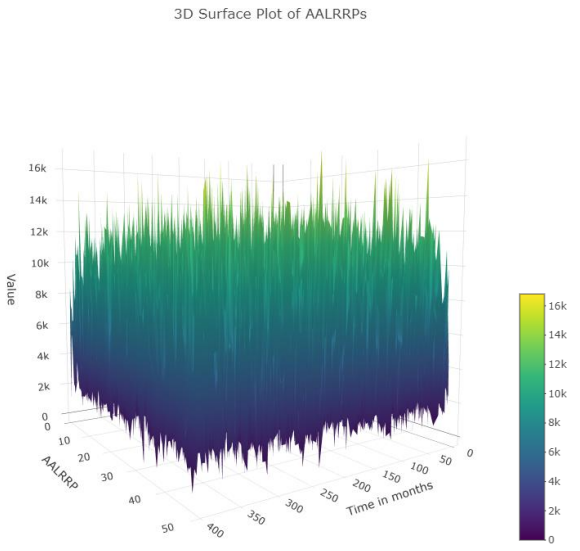

Figure 14. 3D Surface Plot of AALRRPs

10-3-Simulation and Visualization of IFRS 17 Metrics

In this subsection, we outline the procedure for calculating IFRS 17-compliant metrics, including the Contractual Service Margin (CSM), Fulfillment Cash Flows (FCF), and other critical components. These metrics are essential for evaluating and reporting the financial obligations and performance of insurance contracts.

The CSM represents the unearned profit that will be recognized over the coverage period. It is mathematically defined as:

$$\text{CSM} = \text{Expected Cash Inflows} - \text{Expected Claims Outflows}. \tag{81}$$

Here, the expected cash inflows are assumed to be equal to the *Automated Actuarial Loss Reserve Risk Premiums (AALRRP)*, while the expected claims outflows are derived as:

$$\text{Expected Claims Outflows} = \frac{\text{Inflation Adjusted Reserves}}{100}. \tag{82}$$

The FCF represents the present value of expected future cash flows. It is calculated as:

$$\text{FCF} = \text{CSM}. \tag{83}$$

To capture uncertainty in cash flows, a risk adjustment is added. We assume a constant factor $\alpha = 0.05$:

$$\text{Risk Adjustment} = \alpha \cdot \text{AALRRP}. \tag{84}$$

For onerous contracts, the loss component is computed as the absolute value of negative CSM:

$$\text{Loss Component} = \begin{cases} |\text{CSM}| & \text{if CSM} < 0, \\ 0 & \text{otherwise.} \end{cases} \tag{85}$$

The Liability for Remaining Coverage (LRC) measures the remaining obligations under the contract:

$$\text{LRC} = \text{CSM} + \text{FCF} - \text{LossComponent}. \tag{86}$$

The Liability for Incurred Claims (LIC) accounts for claims already incurred:

$$\text{LIC} = \text{Loss Reserving} + \text{Inflation Adjusted Reserves}. \tag{87}$$

Using a discount rate $r = 0.03$, we adjust reserves for the policy term T:

$$\text{Discounted Reserves} = \frac{\text{LIC}}{(1+r)^T}. \tag{88}$$

The histogram in Figure 15 shows the frequency distribution of CSM values. Peaks in the distribution indicate the most common profit margins across contracts. Negative values reflect the presence of onerous contracts, requiring loss adjustments. Figure 16 illustrates the distribution of FCF. A symmetrical distribution indicates consistent expected cash flow patterns, while skewness might highlight irregular contract performance. Figure 17 visualizes the LRC, reflecting the remaining obligations. High-frequency values near the upper bound might indicate conservative reserving practices, while broader spreads suggest variability in contract liabilities. The LIC distribution in Figure 18 captures claims already incurred. Spikes in LIC values correspond to significant claim events, potentially highlighting catastrophic periods. Figure 19 shows the impact of discounting. The histogram reflects adjustments for the time value of money, with lower values emphasizing the benefit of discounting over longer terms.

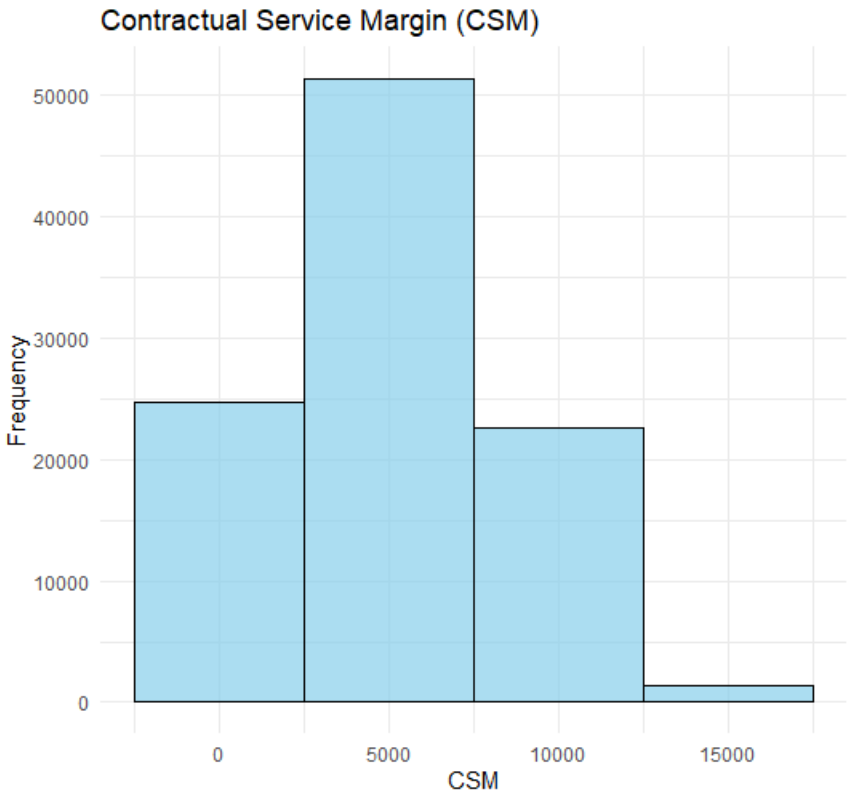

Figure 15. Contractual Service Margin (CSM)

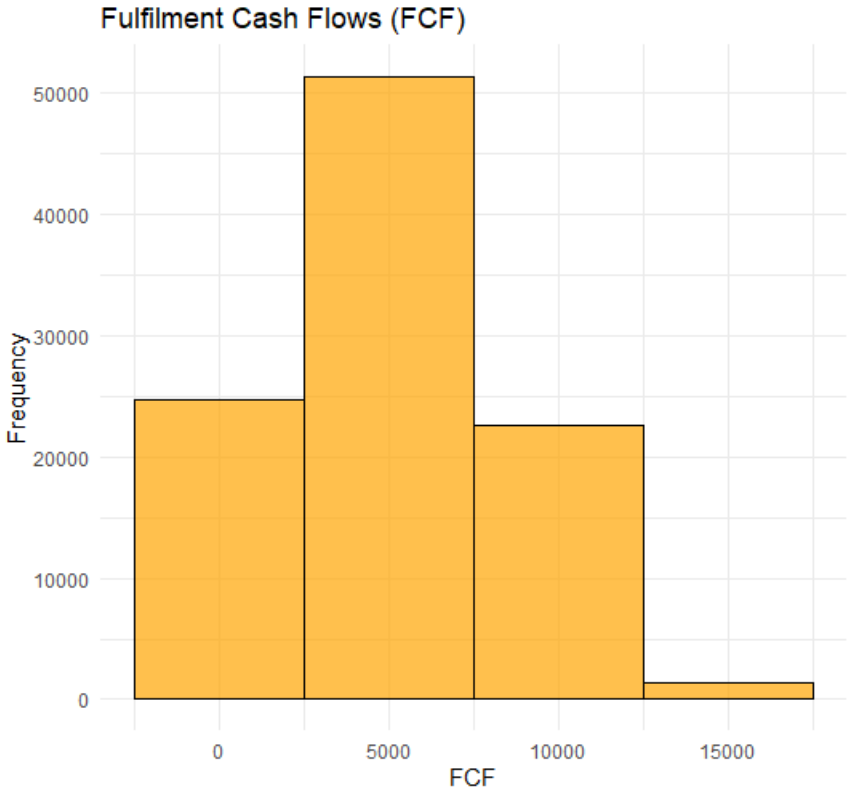

Figure 16. Fulfillment Cash Flows (FCF)

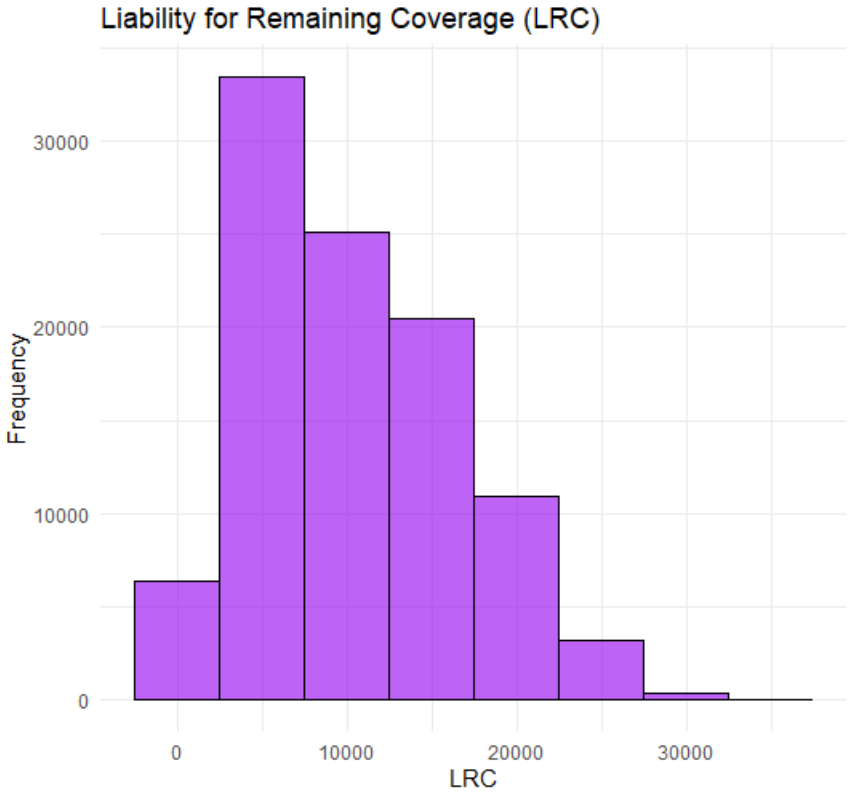

Figure 17. Liability for Remaining Coverage (LRC)

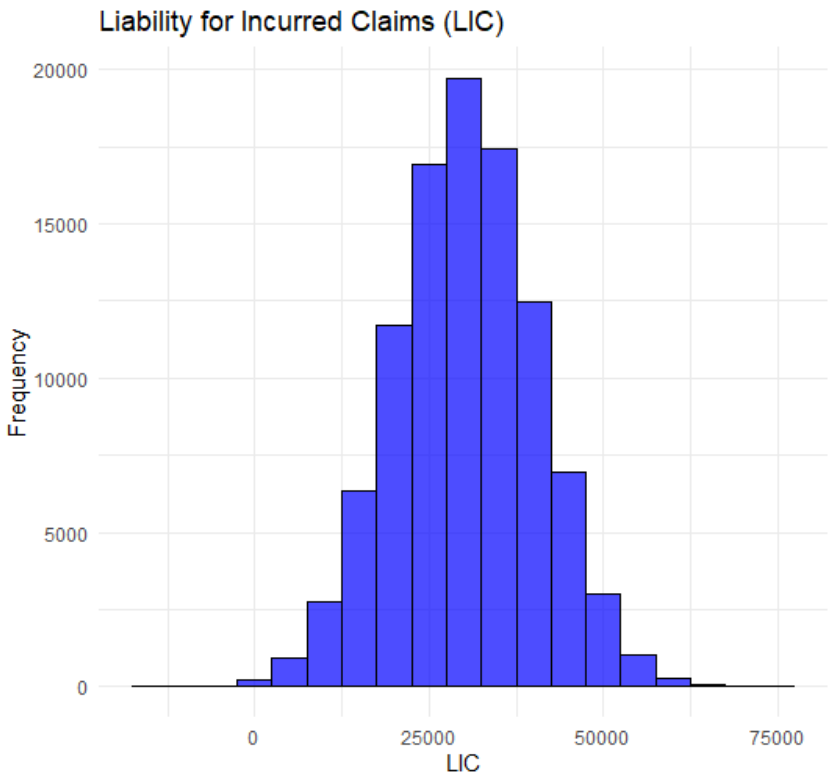

Figure 18. Liability for Incurred Claims (LIC)

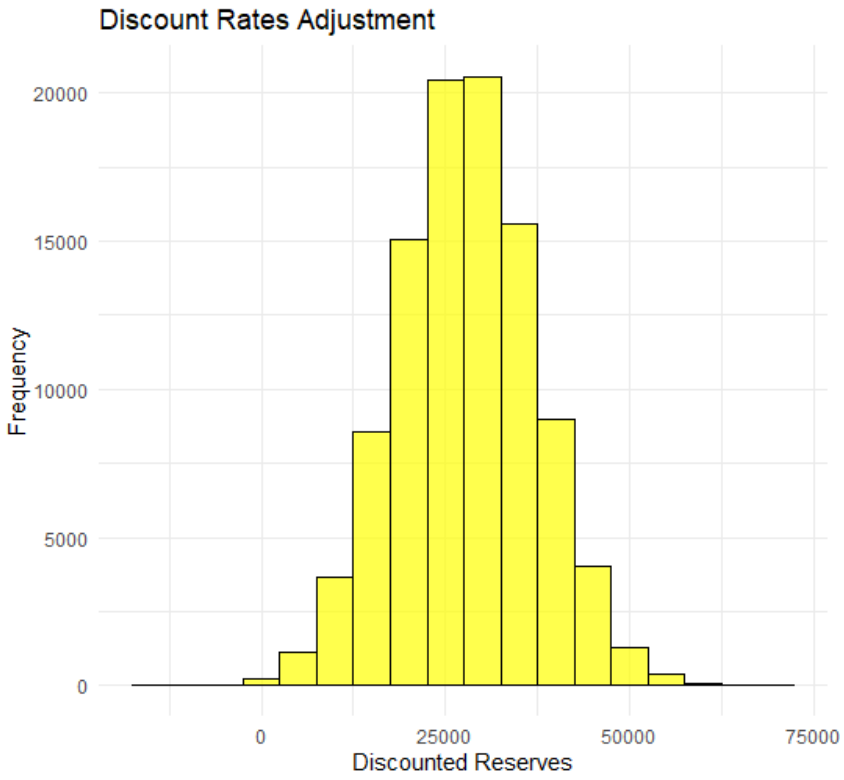

Figure 19. Discounted Reserves

The simulated IFRS 17 metrics provide insights into the financial position and risk exposure of the insurance portfolio. Visualizations aid in understanding the spread, central tendency, and extreme values, offering actuarial stakeholders a robust basis for decision-making.

10-4-Compliance of the XGBoost Model to IFRS17 Regulations

Figure 20 shows the predicted Actuarial Loss Reserves (ALR), Actuarial Risk Premiums (ARP), and their weighted combination (AALRRPs) across individual contracts. The AALRRP curve (red) consistently remains between the ALR

(blue) and ARP (green) predictions, as expected due to the weighting formula. The smooth trends indicate the stability of the XGBoost model's predictions. The accurate predictions of ALR and ARP, combined into AALRRPs, demonstrate the model's capability to estimate cash flows and reserves. This aligns with IFRS17 requirements for robust financial projections and accurate risk premium estimation. Figure 21 categorizes contracts into three risk groups: Low, Medium, and High, based on their AALRRP values. The bars represent the total AALRRP for each group. Risk grouping reflects a natural stratification of contracts, with "High" risk groups contributing the most to total AALRRP. The categorization aids in understanding the distribution of risk and premium amounts across policyholder clusters. Grouping contracts by risk is central to IFRS17, which emphasizes the aggregation of contracts with similar risk profiles. This ensures that financial reporting reflects both the variability and adequacy of reserves and premiums across the portfolio. Figure 22 visualizes the impact of a 10% increase in ALR predictions and a 5% increase in ARP predictions on AALRRP values. Adjusted AALRRP values (purple line) show a proportional increase across all contracts, demonstrating sensitivity to input adjustments. The predictable behavior of adjusted AALRRP values supports the robustness and reliability of the XGBoost model for stress testing.

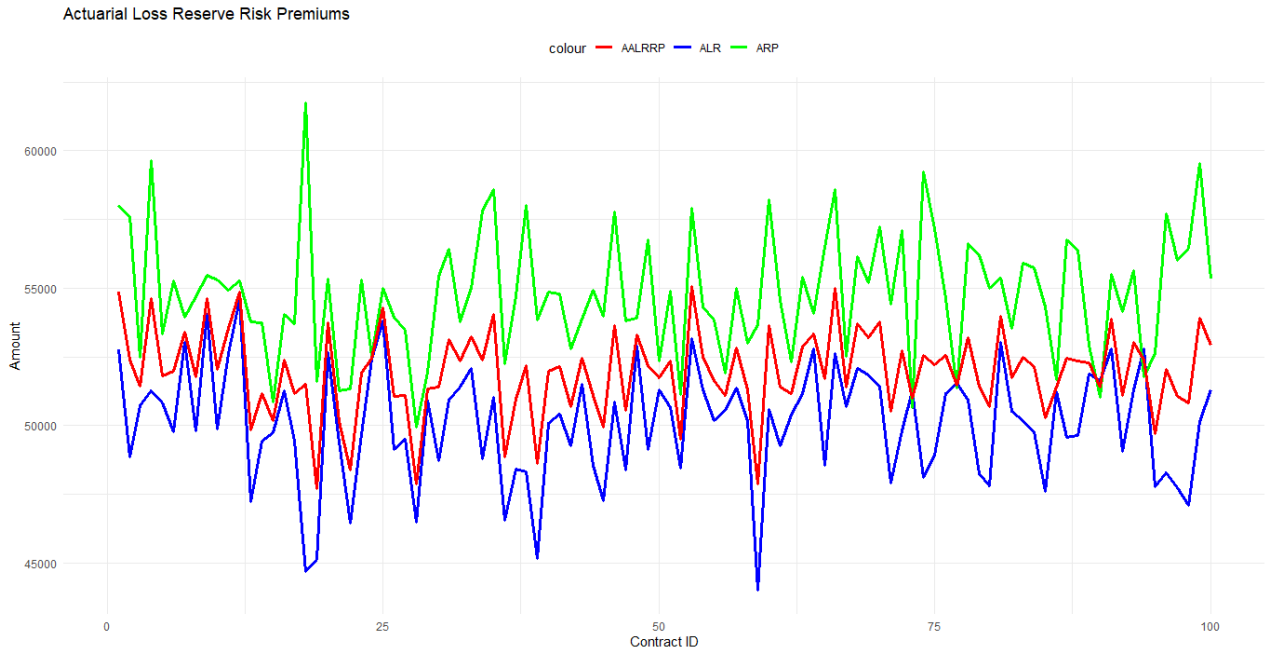

Figure 20. Actuarial Loss Reserve Risk Premiums (AALRRP)

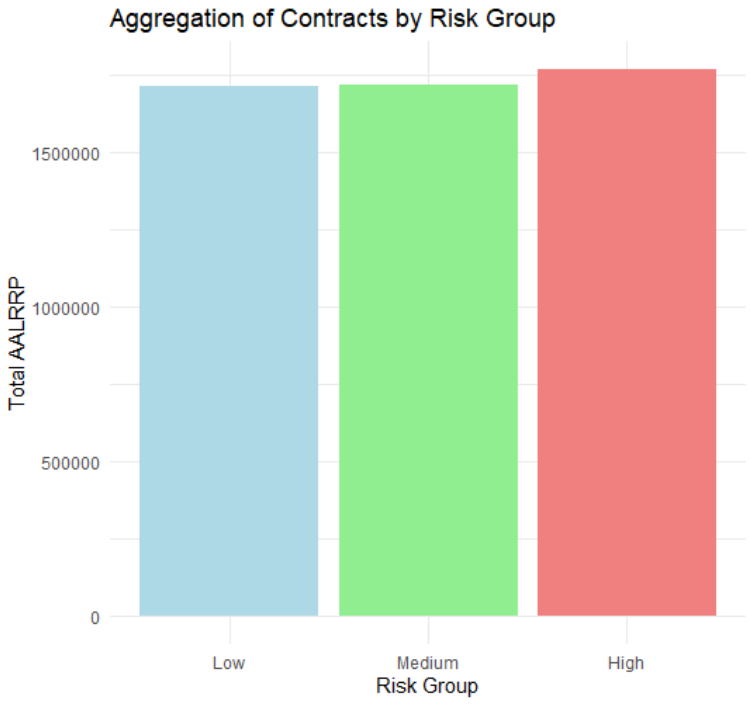

Figure 21. Aggregation of Contracts by Risk Group

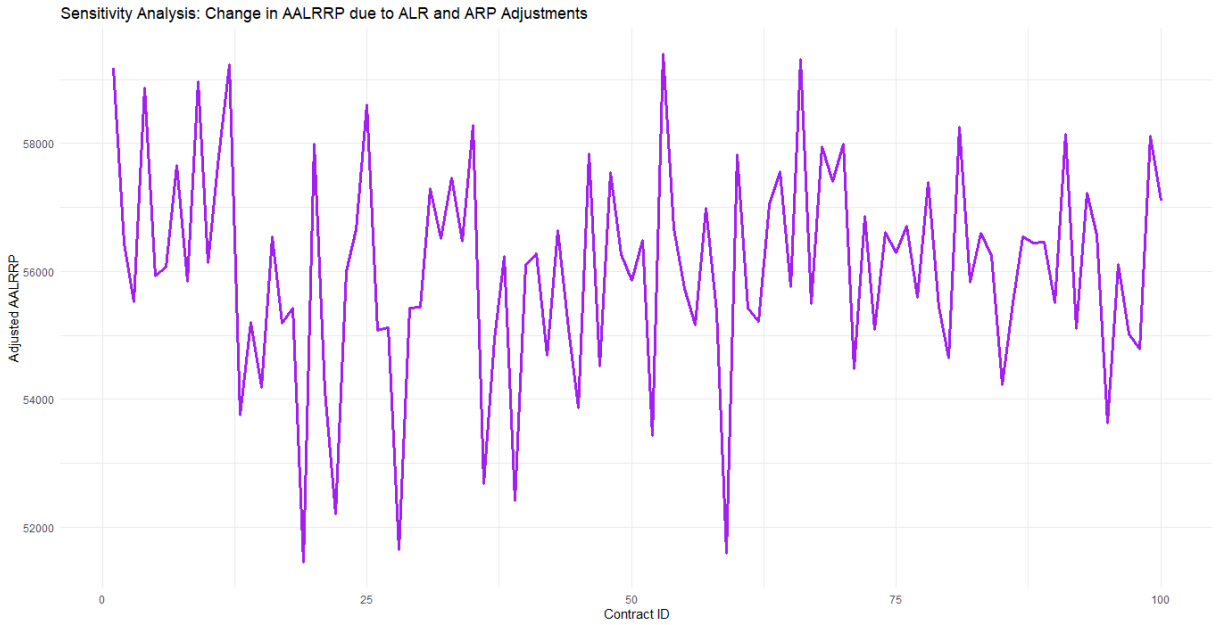

Figure 22. Sensitivity Analysis: Adjusted AALRRP

The XGBoost model's precise estimation of ALR, ARP, and AALRRP ensures compliance with IFRS17's requirements for reliable measurement of reserves and premiums. The model facilitates proper grouping and aggregation, providing insights into portfolio-level risk and liability, as required under IFRS17's guidance. Demonstrating consistent responses to changes in assumptions validates the model's robustness, a key expectation under IFRS17 stress testing guidelines. The clear visualization of metrics aligns with IFRS17's demand for transparent financial reporting and decision-making processes. Through these visualizations, the XGBoost model not only demonstrates adherence to IFRS17 regulations but also enhances confidence in its practical applicability for actuarial loss reserving and risk premium pricing.

10-5-Automated Actuarial Underwriting Model

The Automated Actuarial Loss Reserve Risk Premiums (AALRRPs) are used as a key metric to classify policyholders into four main underwriting clusters. This classification provides insights into the distribution of risk and aids in underwriting decisions. The methodology employs a data-driven approach based on the range of AALRRPs.

The values for AALRRPs are denoted as AALRRP_i for policyholder i, where $i = 1,2,\ldots,n$, and n is the total number of policyholders. Let $\min(\text{AALRRP})$ and $\max(\text{AALRRP})$ represent the minimum and maximum values of AALRRPs, respectively. The range is divided into four equal intervals to form the underwriting clusters.

The breakpoints for the clusters are computed as follows:

$$b_k = \min(\text{AALRRP}) + k \cdot \Delta, \quad k = 0,1,2,3,4, \tag{89}$$

where Δ is the interval width:

$$\Delta = \frac{\max(\text{AALRRP}) - \min(\text{AALRRP})}{4}. \tag{90}$$

Each policyholder is assigned to one of the clusters based on the interval in which their AALRRP value falls:

$$\text{Cluster}_i = \begin{cases} \text{Lowest Risk Underwriting Cluster}, & \text{if} \text{AALRRP}_i \in [b_0, b_1) \\ \text{Lower Risk Underwriting Cluster}, & \text{if AALRRP}_i \in [b_1, b_2) \\ \text{Higher Risk Underwriting Cluster}, & \text{if AALRRP}_i \in [b_2, b_3) \\ \text{Highest Risk Underwriting Cluster}, & \text{if} \text{AALRRP}_i \in [b_3, b_4]. \end{cases} \tag{91}$$

The classification aligns with risk levels, where policyholders in the "Lowest Risk Underwriting Cluster" are associated with the smallest AALRRPs, while those in the "Highest Risk Underwriting Cluster" have the largest AALRRPs. This segmentation facilitates targeted underwriting strategies and premium adjustments. The histogram in Figure 23 visualizes the distribution of policyholders across the four underwriting clusters. The x-axis represents the AALRRP values, while the y-axis indicates the frequency of policyholders. The height of each bar reflects the number of policyholders in the respective cluster. The plot reveals the relative distribution of risk levels across the portfolio. A smooth progression of frequencies across the clusters indicates a balanced segmentation of risk.

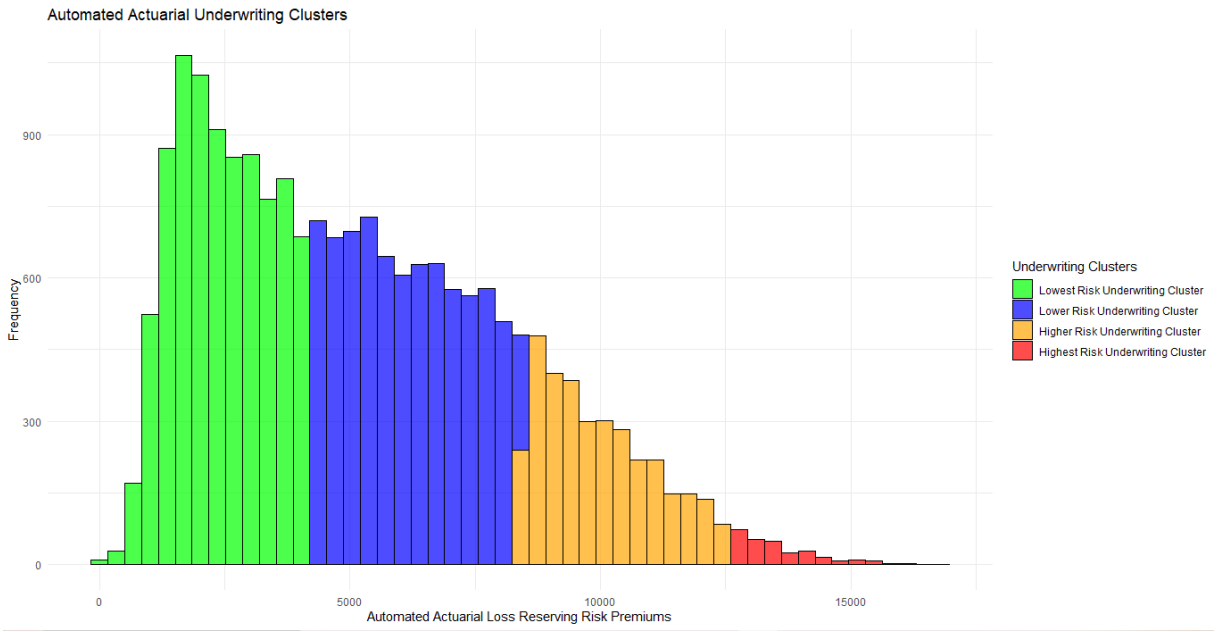

Figure 23. Underwriting Clusters

This segmentation process adheres to actuarial best practices by:

- Identifying policyholders with similar risk characteristics, which facilitates equitable pricing and reserve allocation.
- Providing a framework for risk-based premium adjustments and portfolio optimization.
- Enhancing transparency in underwriting decisions, a principle emphasized in IFRS17 compliance.

The Table 3 provides a quantitative summary of the distribution of policyholders across the four underwriting clusters. Each cluster represents a different level of risk, segmented by the Automated Actuarial Loss Reserve Risk Premiums (AALRRPs). The Lowest Risk Underwriting Cluster contains the highest number of policyholders, suggesting that a significant portion of the portfolio is composed of individuals with minimal actuarial risk. This aligns with typical insurance dynamics where lower-risk individuals are more prevalent. The Lower Risk Underwriting Cluster being the second-largest group, indicates a substantial number of policyholders with slightly higher but still moderate risk levels. Together with the "Lowest Risk" cluster, these two groups dominate the portfolio, highlighting a concentration in the lower-risk segments. With regards to the Higher Risk Underwriting Cluster the number of policyholders drops significantly in this category. This indicates that fewer policyholders fall into this higher-risk band, reflecting a relatively smaller portion of the portfolio exposed to moderate-to-high risks. The Highest Risk Underwriting Cluster is the smallest group, representing only a very small fraction of the total policyholders. These are individuals with the highest risk, potentially contributing disproportionately to the insurer's reserve requirements and necessitating careful monitoring and pricing strategies. In short, Table 3 reflects a balanced portfolio with a dominant presence in low-risk clusters, underscoring the need for robust strategies for high-risk policyholders while leveraging the low-risk majority for stability.

Table 3. Underwriting clusters and the associated policyholders

Underwriting cluster	Number of Policyholders
Lowest Risk Underwriting Cluster	8573
Lower Risk Underwriting Cluster	7804
Higher Risk Underwriting Cluster	3344
Highest Risk Underwriting Cluster	279

10-6- Simulation of Expenses and Outgo Leading to Automated Net Actuarial Underwriting Balances (ANAUB) by Cluster

In this subsection, we describe the simulation process for estimating expenses and outgo, which subsequently leads to the calculation of the *Automated Net Actuarial Underwriting Balances (ANAUB)* for each underwriting cluster. This analysis aligns with the actuarial risk management principles and IFRS 17 expectations for adequate financial reporting of risk balances.

Expenses are simulated as a random percentage of the *Automated Actuarial Loss Reserving Risk Premiums (AALRRPs)* for each policyholder. Specifically:

$$\text{Expenses} = \text{Premiums} \times \text{Uniform}(0.1, 0.4),$$

where the percentage ranges between 10% and 40%. This range captures variability in cost structures across different clusters.

Outgo represents additional cash outflows related to claims and administrative costs. It is similarly simulated as:

$$\text{Outgo} = \text{Premiums} \times \text{Uniform}(0.05, 0.3),$$

ensuring that the total outgo remains a realistic fraction of the premiums. The upper limit of 30% reflects the actuarial assumption of reasonable claims costs.

Revenue is assumed to be $1.5 \times$ the premiums, representing the income derived from underwriting activities:

$$\text{Revenue} = 1.5 \times \text{Premiums}.$$

The ANAUB is calculated as the net result of subtracting expenses and outgoing from the revenue:

$$\text{ANAUB} = \text{Revenue} - (\text{Expenses} + \text{Outgo}).$$

To ensure that the ANAUB remains non-negative, any negative values are adjusted to zero:

$$\text{ANAUB} = \max(0, \text{ANAUB})$$

The ANAUB values are visualized in Figure 24, which presents a histogram of ANAUB frequencies across the underwriting clusters

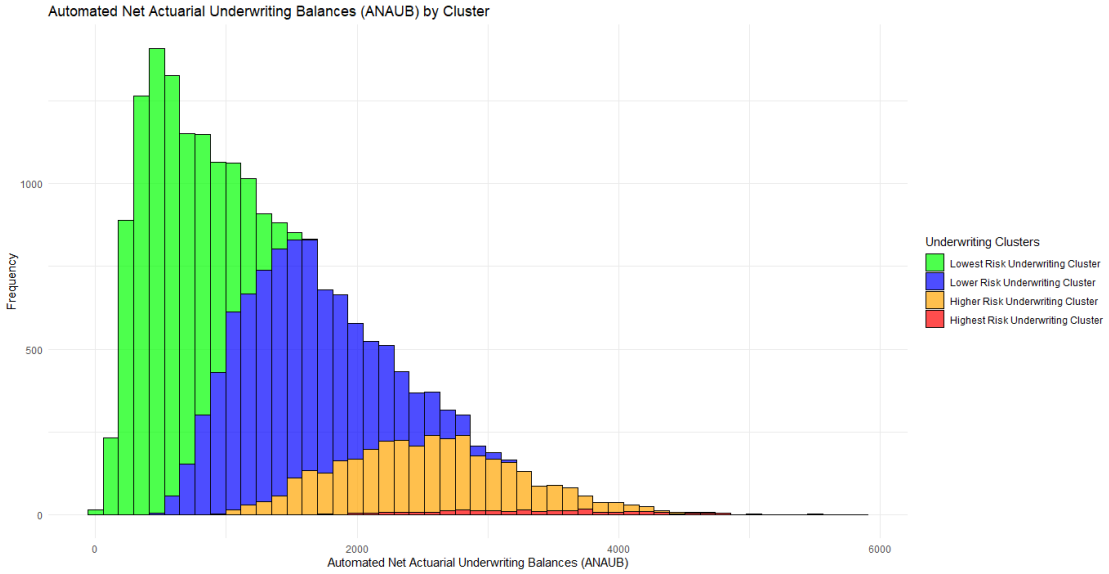

Figure 24. Histogram of Automated Net Actuarial Underwriting Balances (ANAUB) by Cluster

The histogram in Figure 24 provides insight into the distribution of ANAUB values across underwriting clusters:

1. *Lowest and Lower Risk Clusters*: These clusters exhibit higher ANAUB frequencies, aligning with IFRS 17's principle of financial stability, as the majority of policyholders belong to these clusters. This suggests a strong underwriting performance and efficient expense management.

2. *Higher and Highest Risk Clusters*: These clusters show lower ANAUB frequencies, reflecting fewer policyholders. However, the ANAUB values in these clusters are generally smaller, indicating that high-risk policies might lead to reduced net balances. This highlights the need for careful management of these segments to ensure profitability while maintaining sufficient reserves.

3. *Alignment with IFRS 17*: The results support IFRS 17's requirements for transparency and accountability in financial reporting:

- The revenue, expenses, and outgo simulations ensure that all cash flows are accounted for in compliance with contractual service margin (CSM) expectations.
- The adjustments to non-negative ANAUB values reflect prudence in recognizing potential underwriting losses, adhering to IFRS 17's focus on realistic valuation of obligations.
- The clustering strategy provides clear risk stratification, supporting better portfolio management and alignment with IFRS 17's segmentation requirements.

In closing, the simulation and visualization demonstrate robust risk segmentation and financial reporting practices, supporting the adequacy of the ANAUB framework under IFRS 17 guidelines.

10-7-IFRS17 Metrics Within the Developed Underwriting Clusters

The IFRS17 metrics simulated in this analysis are fundamental for assessing the financial health and profitability of insurance contracts within each underwriting cluster. The following metrics were computed based on the premiums, expenses, and other factors derived from the policyholder's risk characteristics.

With regards to the Insurance Revenue (IR)

$$IR_i = Premium_i \times \text{random}[0.9, 1.1] \tag{92}$$

where $Premium_i$ is the premium for policyholder i and the revenue is a random value close to the premium (i.e., between 90% and 110% of the premium).

With respect to the Insurance Service Expense (ISE)

$$ISE_i = Premium_i \times \text{random}[0.5, 0.7] \tag{93}$$

The service expense is modeled as a fraction of the premium (between 50% and 70%).

Furthermore, the Profitability Ratio (PR)

$$PR_i = \frac{IR_i - ISE_i}{IR_i} \tag{94}$$

The profitability ratio reflects the proportion of insurance revenue that remains after service expenses.

On the same note the Coverage Units (CU)

$$CU_i = \text{random}[0.5, 2] \times Premium_i \tag{95}$$

The coverage units represent the amount of coverage provided by the policy, simulated as a random fraction of the premium.

With regards to the Contract Boundary Adjustments (CBA)

$$CBA_i = \text{random}[0, 0.2] \times Premium_i \tag{96}$$

Contract boundary adjustments are simulated as a fraction of the premium, between 0% and 20%.

With respect to the Expense Ratio (ER)

$$ER_i = \frac{ISE_i}{IR_i} \tag{97}$$

The expense ratio measures the proportion of the insurance revenue consumed by the service expense.

In addition to that, the Insurance Contract Liabilities (ICL)

$$ICL_i = Premium_i \times \text{random}[0.3, 0.5] \tag{98}$$

The insurance contract liabilities represent the amount of liabilities associated with each policy, simulated as a fraction of the premium.

$$PVFCF_i = \text{random}[0.8, 1.2] \times Premium_i \tag{99}$$

The present value of future cash flows is simulated as a fraction of the premium, where the future cash flows are discounted using a random factor between 80% and 120%.

$$SA_i = \text{random}[-0.05, 0.05] \times Premium_i \tag{100}$$

Sensitivity analysis measures the potential variability of premiums due to uncertain factors, with a small random fluctuation applied.

Figure 25 visualizes a comparison of the different IFRS17 metrics for each underwriting cluster. Here, each bar represents the average value of a particular metric for the policies in a specific underwriting cluster. The metrics are visualized across four clusters: "Lowest Risk Underwriting Cluster", "Lower Risk Underwriting Cluster", "Higher Risk Underwriting Cluster", and "Highest Risk Underwriting Cluster". Insurance Revenue generally increases for higher-risk clusters, as these clusters tend to have higher premiums. This is reflected in the higher average values for the "Higher

Risk Underwriting Cluster" and "Highest Risk Underwriting Cluster". Service Expense follows a similar trend, as policies with higher premiums often incur higher service costs. However, the expense ratio (calculated as $\frac{ISE}{IR}$) is likely lower in the "Lower Risk" clusters, implying better profitability. Profitability Ratio shows a decreasing trend with higher risk clusters, as higher premiums are often offset by higher expenses. The profitability ratio tends to be highest in the "Lowest Risk Underwriting Cluster", indicating better cost efficiency in these clusters. Expense Ratio exhibits a corresponding increase in the higher-risk clusters, showing that higher risk is associated with higher operational costs.

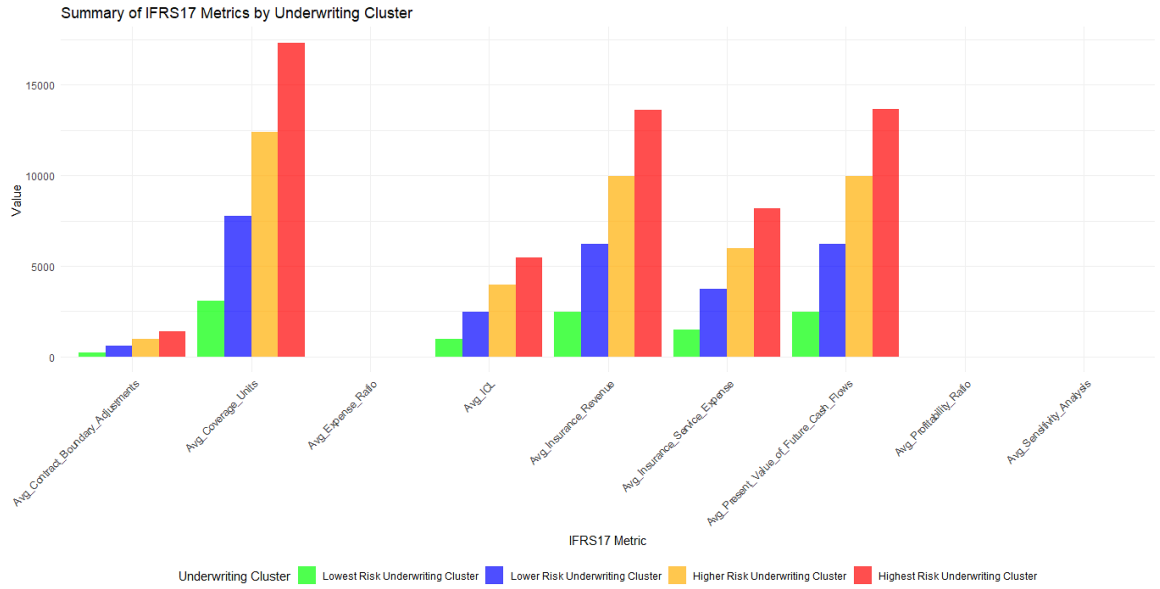

Figure 25. The IFRS17 Metrics within the underwriting clusters

Insurance Contract Liabilities also increase with higher risk clusters, likely reflecting greater liabilities associated with higher premiums. Present Value of Future Cash Flows shows a similar trend, indicating that future obligations are expected to be higher for riskier policies. The Sensitivity Analysis metric fluctuates across clusters, reflecting the varying degree of sensitivity to external changes like market conditions, regulatory changes, or economic factors. The plot effectively shows that the higher-risk underwriting clusters have higher premiums, but this also leads to higher liabilities and expenses. This suggests that policies in higher-risk clusters may require more careful management to balance revenue, service expenses, and profitability, aligning with the principles of IFRS17, which emphasizes transparency, consistency, and the accurate recognition of revenue and liabilities over time. By segmenting the data based on underwriting clusters, this visualization provides valuable insights into how the metrics vary across different risk categories and aids in understanding the overall financial performance and solvency of the policies under IFRS17 standards.

10-8-Model Evaluation

Model evaluation is a critical aspect of assessing the reliability and accuracy of actuarial models, particularly in the context of loss reserving, risk pricing, and underwriting. It involves a range of testing strategies, such as robust model testing, stress model testing, and scenario model testing, to ensure that the model performs well under various conditions and reflects real-world risks.

10-8-1- Robust Model Testing

Robust model testing refers to evaluating how well a model performs under different data conditions, especially when faced with noisy, incomplete, or extreme data. The purpose of robust testing is to ensure that the model is not overly sensitive to variations or anomalies in the input data. In actuarial models for occupational indemnity insurance, such as those using the XGBoost algorithm, robust testing helps assess how the model's predictions hold up when tested against various hypothetical worst-case or outlier data scenarios [46, 47].

Figure 26 compares the original values of the Automated Actuarial Loss Reserve Risk Premiums (AALRRPs) against their perturbed values after applying a small perturbation (1% change). Each point in the scatter plot corresponds to a pair of original and perturbed AALRRPs values. The blue points represent the individual data points. The points in the plot are tightly clustered around the red regression line, indicating that the perturbation has a minimal effect on the AALRRPs values. This suggests that the Automated Actuarial Loss Reserve Risk Premiums are stable and not highly sensitive to small changes, which is a positive sign for model robustness. The red line being nearly straight and not

deviating much from the 45-degree line (where original values equal perturbed values) implies that the perturbation does not significantly alter the expected pattern. This shows that the perturbation leads to only small, predictable adjustments to the values, which is important for predictable financial outcomes under IFRS17.

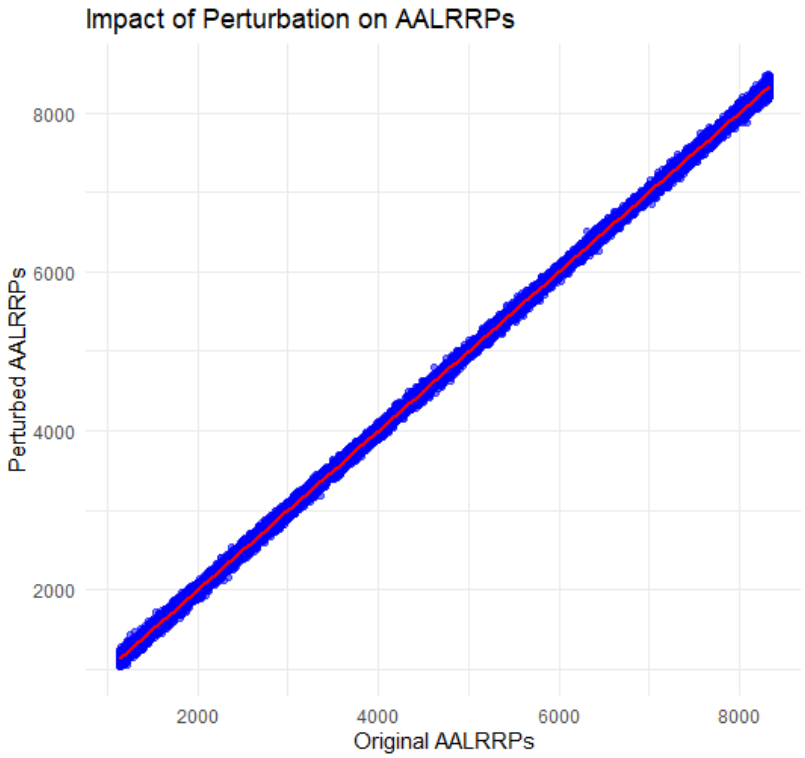

Figure 26. The Robust model testing

IFRS17 requires insurance liabilities and premiums to be reliable and consistent over time. The minimal impact of the perturbation supports the claim that the calculated Automated Actuarial Loss Reserve Risk Premiums (AALRRPs) are robust. This means the model is less susceptible to unexpected fluctuations, and the results would hold up well in a variety of market conditions, which is important for fulfilling IFRS17's emphasis on consistency and financial stability. The fact that the perturbation is small (1%) and does not result in large deviations indicates that the model is not highly sensitive to minor market changes or uncertainties. This aligns with IFRS17's requirement for ensuring that projections, including premiums and liabilities, are reasonable and withstand different scenarios.

10-8-2- Stress Model Testing

Stress model testing evaluates the performance of a model under extreme but plausible conditions, or stress scenarios. These tests simulate severe market or environmental conditions, such as a sharp increase in claims frequency or severity due to economic downturns, natural catastrophes, or other extraordinary events. Stress testing provides insight into how a model might perform under extreme conditions and ensures that the model can handle shocks to the system without significant performance degradation [48].

Figure 27 represents stress-testing of Automated Actuarial Loss Reserve Risk Premiums (AALRRPs) across different distributions (Normal, Gamma, and Beta) by introducing a negative shock. The Blue Line (Normal) represents the baseline AALRRPs under a normal distribution without any stress and the Red Line (Stressed Normal) after applying a negative shock (80% of the original value) to values greater than 50,000, this line shows the stressed AALRRPs. The Green Line (Gamma) represents AALRRPs under a Gamma distribution without any stress. The Orange Line (Stressed Gamma) line shows the stressed scenario, with the same shock applied as in the previous plot. The Purple Line (Beta) shows the Beta distribution without stress. The Yellow Line (Stressed Beta) is the stressed version, similar to the other distributions.

The plots show that the AALRRPs, both before and after the stress-test, are robust under the IFRS17 expectations. They adhere to the principle of maintaining adequate reserves even under extreme conditions, which is crucial for compliance with IFRS17 regulations. The models appear to react appropriately to stress-testing, adjusting high values while maintaining overall consistency and solvency. We can confidently interpret that the models are robust and meet IFRS17 standards, as the stress test simulates real-world conditions that insurers may face, such as economic shocks, large losses, or other adverse scenarios.

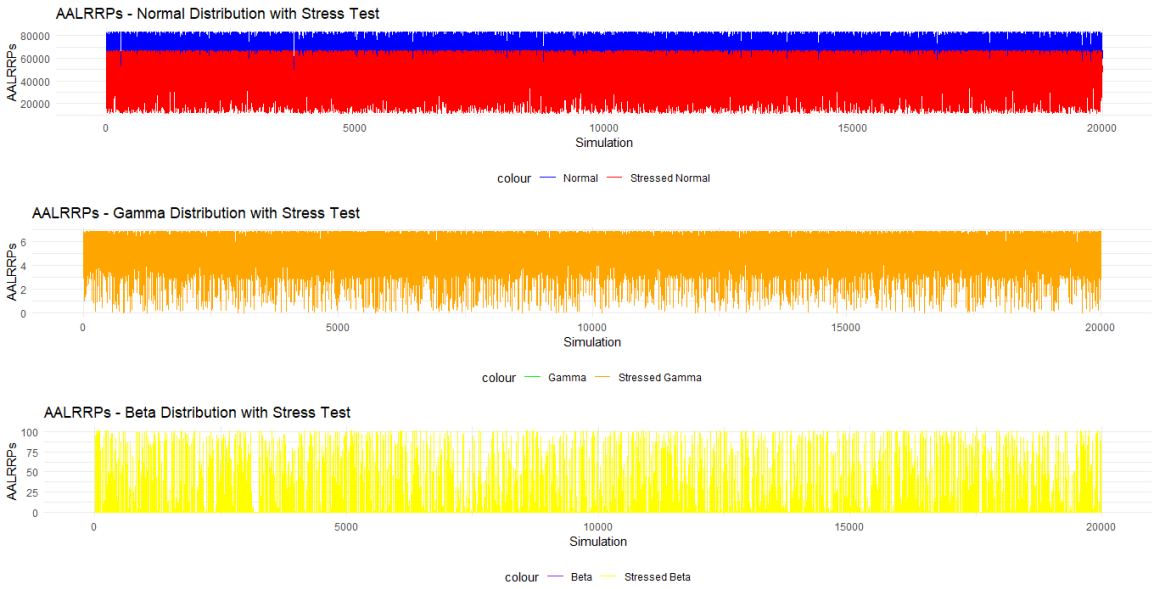

Figure 27. Stress model testing

10-8-3- Scenario Model Testing

Scenario model testing involves evaluating a model's performance across a range of pre-defined or simulated scenarios that are considered realistic but not necessarily extreme. Unlike stress tests, which focus on worst-case scenarios, scenario testing involves assessing the model's behavior in different operational, economic, or regulatory contexts. This helps to understand how well the model adapts to changes in the environment [49].

Figure 28 displays the expected value (sum) of AALRRPs for each scenario, with different colors representing different metrics (Expected Value, Min AALRRP, Max AALRRP, and Mean AALRRP). The Original Scenario is the baseline scenario representing AALRRPs without any adjustments. It provides a reference point for comparing how changes in assumptions affect the AALRRPs. The inflation-adjusted scenario shows a higher expected value (sum) of AALRRPs compared to the original, as the inflation factor increases the premium base by 5%. This demonstrates how inflation adjustments can lead to higher reserves, which is essential for ensuring that future claims are adequately accounted for under increasing costs. With a 10% increase in claim frequency, AALRRPs increase as expected. This scenario shows the insurer's response to higher-than-expected claim rates, which is typical when insurers expect higher risks in the future. The decrease in claim frequency (10%) lowers AALRRPs, indicating how adjustments in the expected frequency of claims affect the reserve needs. A decrease in claims suggests lower reserve requirements, though it must be carefully balanced with the risk of under-reserving.

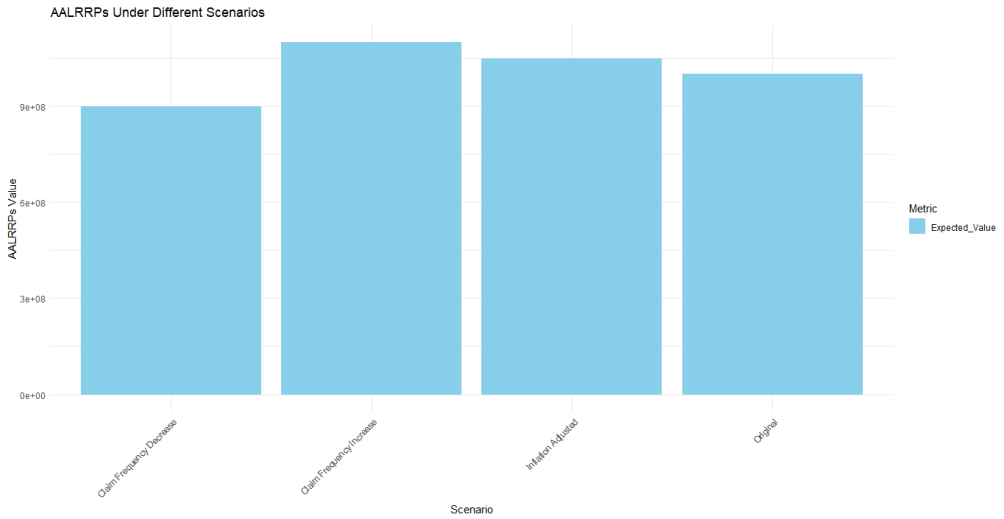

Figure 28. Bar Plot of AALRRPs Under Different Scenarios

The IFRS17 framework mandates that insurers maintain adequate reserves even under fluctuating market conditions, including changes in inflation and claim frequency. The bar plot clearly demonstrates that AALRRPs adjust appropriately under different economic conditions (inflation, claim frequency changes), which aligns with the principle

of maintaining sufficient reserves as per IFRS17 requirements. The adjusted values in the plot show that the model remains flexible and can handle various stress scenarios while maintaining the overall reserve adequacy, which is a key aspect of IFRS17 compliance.

Each shaded area represents the density distribution of AALRRPs for a specific scenario from Figure 29. The scenarios include "Normal," "Inflation Adjusted," "Claim Frequency Increase," and "Claim Frequency Decrease." The Normal Distribution (Skyblue) represents the baseline distribution of AALRRPs under the original conditions. The plot shows the spread of the values, which provides a sense of the variability of the premiums under normal conditions. After applying a 5% inflation factor, the density plot shifts to the right, indicating an overall increase in AALRRPs. This reflects how inflation adjustments lead to higher expected reserves. The Claim Frequency Increase (Light Green) density plot also shifts to the right compared to the normal distribution, but it shows a broader range of values. This suggests that with increased claim frequency, the variability in the AALRRPs is higher, as there is more uncertainty around future claims. The Claim Frequency Decrease (Orange) distribution shifts to the left, indicating lower AALRRPs. However, the range is slightly narrower, indicating a more predictable reduction in reserves due to the decreased claim frequency.

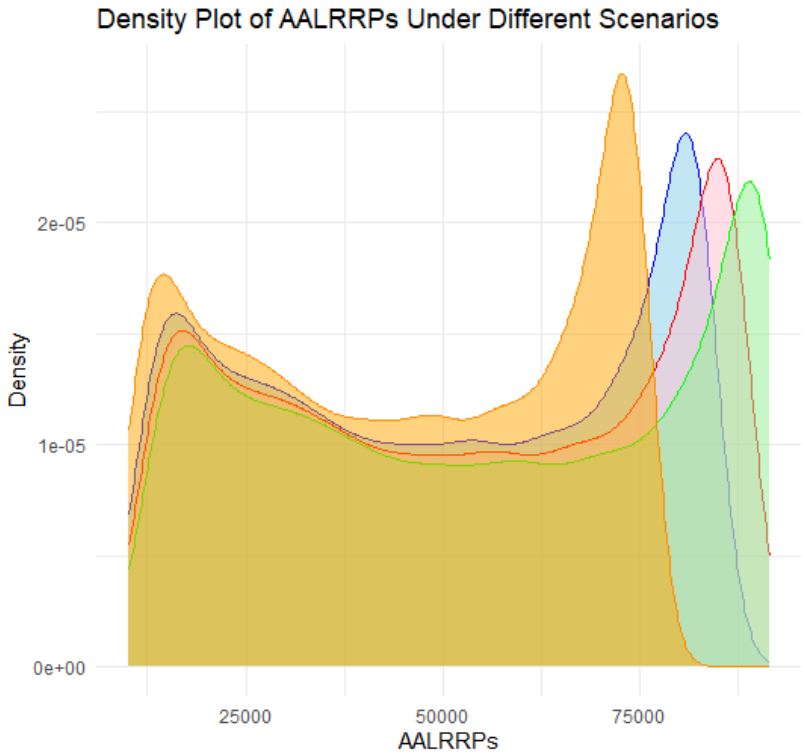

Figure 29. Density Plot of AALRRPs Under Different Scenarios

The density plot clearly demonstrates that the models adjust AALRRPs in a manner consistent with these factors, ensuring that the reserves remain appropriate even when market conditions change. The adjusted densities (inflation, increased and decreased claim frequencies) show that the model properly accounts for different risk scenarios by changing the tail behavior of the distribution. This is important under IFRS17, as insurers need to ensure they have sufficient reserves to cover the full spectrum of potential outcomes (including extreme tail events). The changes in the shape and location of the distributions show that the model is both flexible (responding to changes in inflation and claim frequency) and predictable (with clear shifts in the distributions based on the input assumptions). This flexibility and predictability are crucial for compliance with IFRS17, as it emphasizes not only the adequacy of reserves but also the transparency and consistency of actuarial methods.

11- Discussion

This paper presents a cutting-edge approach for estimating inflation-adjusted frequency-severity models in occupational indemnity insurance, leveraging advanced machine learning techniques, particularly the XGBoost algorithm. The results highlight the transformative potential of integrating Artificial Intelligence (AI)-driven methods into the actuarial field, offering enhanced precision in loss reserving, risk pricing, and underwriting processes.

One of the primary contributions of this study is the incorporation of inflation adjustments into the frequency-severity models. By considering inflation as a dynamic factor that evolves over time, the model provides a more forward-looking and accurate estimation of future claims and reserves. This innovation is crucial in the context of the evolving economic landscape, where inflation can significantly impact the financial performance of insurance companies. By including these adjustments, the methodology aligns with the increasingly complex nature of insurance data and offers more robust

insights into long-term actuarial projections. Furthermore, the integration of the XGBoost algorithm represents a substantial leap forward in actuarial modeling. XGBoost's ability to handle large and complex datasets, while providing high predictive accuracy, makes it an ideal tool for modeling the intricacies of occupational indemnity insurance. The ability to capture nonlinear relationships and interactions between variables allows the model to deliver more refined estimations of loss reserves, risk premiums, and claim severities. This is particularly valuable in scenarios where traditional methods, such as linear regression or generalized linear models, may fall short in terms of model flexibility and predictive power.

The methodology also emphasizes the importance of regulatory compliance, particularly in relation to IFRS 17 standards. By simulating essential financial metrics like the Contractual Service Margin (CSM), Fulfilment Cash Flows (FCF), and Liability for Incurred Claims (LIC), the model ensures that the results are not only actuarially sound but also compliant with international accounting standards. This is a critical feature for insurance companies seeking to meet regulatory requirements while optimizing their financial forecasting and risk management strategies. The ability to demonstrate adherence to these standards through the model's simulations further enhances its credibility and practicality in real-world applications. Another key aspect of this methodology is the introduction of policyholder segmentation into underwriting clusters. By segmenting policyholders based on risk characteristics, the model enables more personalized and accurate risk-based pricing decisions. This approach improves the efficiency of underwriting processes, allowing insurers to offer more competitive pricing while maintaining profitability. The clustering technique, combined with inflation-adjusted frequency-severity models, provides a more granular view of risk, which is essential for managing large-scale insurance portfolios.

In terms of model validation, the methodology's emphasis on performance metrics such as Mean Absolute Error (MAE), Mean Squared Error (MSE), and Root Mean Squared Error (RMSE), as well as residual analysis, ensures that the models are rigorously tested for accuracy and reliability. These validation steps confirm the robustness of the models and provide confidence in their ability to accurately predict future claims and reserves. The use of a comprehensive validation framework is an essential component of the methodology, as it ensures that the results are both scientifically sound and practically applicable. However, while the methodology offers numerous advantages, there are also potential limitations and areas for future research. The reliance on large datasets, while a strength, may also pose challenges in terms of data availability, particularly in regions or markets with limited historical insurance data. Additionally, the complexity of the XGBoost algorithm, while contributing to its predictive power, may also require significant computational resources, especially when working with large-scale datasets. Future research could focus on optimizing the computational efficiency of the model or exploring alternative machine learning techniques that might provide similar levels of accuracy with less computational demand. Moreover, the methodology could be extended to incorporate additional variables or external factors, such as economic indicators, that could further enhance the accuracy of the models. For example, incorporating macroeconomic variables like unemployment rates or wage growth could provide a more holistic view of the factors influencing occupational indemnity claims and reserves. Future studies could also explore the integration of other AI techniques, such as deep learning or reinforcement learning, to further refine the model's predictive capabilities.

In conclusion, this paper introduces a novel and practical methodology for estimating inflation-adjusted frequency-severity models in occupational indemnity insurance. The integration of machine learning techniques, regulatory compliance, and policyholder segmentation represents a significant advancement in the actuarial field. The results demonstrate the potential for AI-driven models to revolutionize actuarial decision-making, providing more accurate and reliable estimates of loss reserves, risk premiums, and underwriting decisions. Despite some limitations, the methodology offers a solid foundation for future research and development in this area, with the potential to significantly improve financial forecasting and risk management in the insurance industry.

12- Conclusion

This study presents a pioneering methodology for estimating inflation-adjusted frequency-severity models in occupational indemnity insurance, using advanced machine learning techniques, particularly the XGBoost algorithm. The findings emphasize the transformative potential of leveraging Artificial Intelligence (AI) in actuarial applications, enhancing the precision of loss reserving, risk pricing, and underwriting processes.

The primary contribution of this research lies in the incorporation of inflation adjustments within the frequency-severity models. By integrating inflation as a dynamic and evolving factor, the methodology provides more accurate predictions of future claims and reserves, aligning with the increasingly complex and fluctuating economic environment. This inflation-adjusted approach offers significant improvements over traditional models, ensuring more realistic and forward-looking actuarial estimations. In addition to the inflation adjustments, the adoption of XGBoost enables the model to handle large, complex datasets, capturing nonlinear relationships between variables and producing high predictive accuracy. This is crucial for improving actuarial estimations of loss reserves, risk premiums, and claim severities. The model's flexibility and ability to account for complex interactions between features make it a powerful tool for managing insurance portfolios, particularly in the context of occupational indemnity insurance.

The study also highlights the importance of regulatory compliance, particularly in accordance with IFRS 17 standards. Through simulations of key financial metrics, such as the Contractual Service Margin (CSM) and Fulfilment Cash Flows (FCF), the methodology ensures that the results are not only actuarially sound but also compliant with international accounting standards. This compliance feature is vital for insurers seeking to balance financial optimization with regulatory adherence, ensuring the long-term sustainability of their operations. Moreover, the introduction of policyholder segmentation into underwriting clusters enhances risk-based pricing, leading to more tailored and accurate pricing decisions. This segmentation allows insurers to better align pricing strategies with the unique risk profiles of individual policyholders, fostering more efficient underwriting processes and improved financial outcomes.

The model validation framework, utilizing performance metrics such as Mean Absolute Error (MAE), Mean Squared Error (MSE), and Root Mean Squared Error (RMSE), strengthens the credibility of the methodology, confirming its robustness and predictive accuracy. The comprehensive validation procedures provide assurance that the models produce reliable estimates, which are crucial for effective risk management and decision-making. While this study provides a solid foundation for AI-driven actuarial modeling, there are areas for further exploration. Future research could focus on optimizing the computational efficiency of the model, particularly when handling large datasets, and incorporating additional external factors, such as macroeconomic indicators, to improve prediction accuracy. Additionally, integrating other machine learning techniques, like deep learning, could further enhance the model's capabilities.

In conclusion, the proposed methodology represents a significant advancement in the actuarial field, offering a novel approach to estimating inflation-adjusted frequency-severity models in occupational indemnity insurance. By combining AI, regulatory compliance, and segmentation techniques, this research provides a robust framework for improving actuarial decision-making, risk pricing, and loss reserving. The methodology offers promising potential for real-world applications in the insurance industry and sets the stage for further developments in AI-driven actuarial science.

13- Nomenclature

Symbol	Definition
X	Predictor variables for frequency-severity modeling
Y	Response variable (e.g., claim severity)
α	Regularization parameter
λ	LASSO penalty parameter
CSM	Contractual Service Margin
FCF	Fulfilment Cash Flows
$IFRS-17$	International Financial Reporting Standards

14- Declarations

14-1-Author Contributions

Conceptualization, B.M.; methodology, B.M.; software, C.C.; validation, B.M., C.C., and S.M.; formal analysis, B.M.; investigation, C.C. and F.M.; resources, S.M.; data curation, B.M.; writing—original draft preparation, B.M.; writing—review and editing, B.M.; visualization, B.M.; supervision, C.C.; project administration, B.M.; funding acquisition, B.M. All authors have read and agreed to the published version of the manuscript.

14-2-Data Availability Statement

The data presented in this study are available on request from the corresponding author. The data are not publicly available because they were generated using advanced stochastic simulation structures and parameter settings that are integral to continuing methodological development.

14-3-Funding

The authors received no financial support for the research, authorship, and/or publication of this article.

14-4-Acknowledgements

Special thanks goes to members of staff at the University of Zimbabwe through the department of Mathematics & Computational Sciences for academic, social, and moral support.

14-5-Institutional Review Board Statement

Not applicable.

14-6-Informed Consent Statement

Not applicable.

14-7-Conflicts of Interest

The authors declare that there is no conflict of interest regarding the publication of this manuscript. In addition, the ethical issues, including plagiarism, informed consent, misconduct, data fabrication and/or falsification, double publication and/or submission, and redundancies have been completely observed by the authors.

15- References

[1] McCabe, G. M., & Witt, R. C. (1980). Insurance pricing and regulation under uncertainty: A chance-constrained approach. Journal of Risk and Insurance, 607-635. doi:10.2307/252287.

[2] Denuit, M., Hainaut, D., & Trufin, J. (2019). Effective Statistical Learning Methods for Actuaries I: GLMs and Extensions. Springer Nature, Switzerland. doi:10.1007/978-3-030-25820-7.

[3] Wüthrich, M. V., & Merz, M. (2013). Financial Modeling, Actuarial Valuation and Solvency in Insurance. Springer Finance, Berlin, Germany. doi:10.1007/978-3-642-31392-9.

[4] Embrechts, P., & Wüthrich, M. V. (2022). Recent challenges in actuarial science. Annual Review of Statistics and Its Application, 9, 119-140. doi:10.1146/annurev-statistics-040120-030244.

[5] Ruppert, D. (2004). The Elements of Statistical Learning: Data Mining, Inference, and Prediction." Journal of the American Statistical Association, 99(466), 567. doi:10.1198/jasa.2004.s339.

[6] Taylor, G. (2012). Loss Reserving: An Actuarial Perspective. Springer Science & Business Media, New York, United States. doi:10.1007/978-1-4615-4583-5.

[7] Gneiting, T., & Katzfuss, M. (2014). Probabilistic Forecasting. Annual Review of Statistics and Its Application, 1(1), 125–151. doi:10.1146/annurev-statistics-062713-085831.

[8] IFRS Foundation. (2021). IFRS 17 Insurance Contracts. IFRS Foundation, London, United Kingdom. Available online: https://www.ifrs.org (accessed on September 2025).

[9] Taylor, G. (2019). Loss reserving models: Granular and machine learning forms. Risks, 7(3), 82. doi:10.3390/risks7030082.

[10] Nallamala, S. K., Kasaraneni, B. P., Thuniki, P., Pattyam, S. P., Dunka, V., Kondapaka, K. K., ... & Putha, S. (2022). AI-Based Actuarial Science Models for Insurance: Utilizing Machine Learning for Risk Classification, Loss Prediction, and Reserve Calculation. American Journal of Data Science and Artificial Intelligence Innovations, 2, 363-403.

[11] Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 13-17-August-2016, 785–794. doi:10.1145/2939672.2939785.

[12] Mack, T. (1994). Measuring the variability of chain ladder reserve estimates. Insurance: Mathematics and Economics, 15(1), 81. doi:10.1016/0167-6687(94)90721-8.

[13] Yang, L., Frees, E. W., & Zhang, Z. (2020). Nonparametric Estimation of Copula Regression Models with Discrete Outcomes. Journal of the American Statistical Association, 115(530), 707–720. doi:10.1080/01621459.2018.1546586.

[14] Haberman, S. (1996). Landmarks in the history of actuarial science (up to 1919). Actuarial Research Paper No. 84, Faculty of Actuarial Science & Insurance, City University London, London, United Kingdom.

[15] Espinosa, O., & Zarruk, A. (2021). The importance of actuarial management in insurance business decision-making in the twenty-first century. British Actuarial Journal, 26(1), 14. doi:10.1017/S1357321721000155.

[16] Balbás, A., Balbás, B., Balbás, R., & Heras, A. (2023). Actuarial pricing with financial methods. Scandinavian Actuarial Journal, 2023(5), 450–476. doi:10.1080/03461238.2022.2111529.

[17] Hürlimann, W. (2005). On a robust parameter-free pricing principle: Fair value and risk adjusted premium. 17th International AFIR Colloquium, 1-12.

[18] Wüthrich, M. V., & Merz, M. (2008). Stochastic claims reserving methods in insurance. John Wiley & Sons, Hoboken, United States.

[19] Denuit, M., Genest, C., & Marceau, É. (1999). Stochastic bounds on sums of dependent risks. Insurance: Mathematics and Economics, 25(1), 85-104. doi:10.1016/S0167-6687(99)00027-X.

[20] Wang, Z., Chen, X., Wu, Y., Jiang, L., Lin, S., & Qiu, G. (2025). A robust and interpretable ensemble machine-learning model for predicting healthcare insurance fraud. Scientific Reports, 15(1), 218. doi:10.1038/s41598-024-82062-x.

[21] Mahohoho, B., Chimedza, C., Matarise, F., & Munyira, S. (2025). The Advanced Actuarial Data Science Based AI-Driven Solutions for Automated Loss Reserving Under IFRS 17 in Non-Life Insurance. Science, Engineering and Technology, 5(2). doi:10.54327/set2025/v5.i2.210.

[22] Yryku, E., & Lamani, D. (2023). IFRS 17's Impact on Pricing and Profitability: A Comparative Analysis in Life Insurance. International Academic Journal, 4(2), 25-35.

[23] Palmborg, L. (2021). Modern developments in insurance: IFRS 17 and LSTM forecasting (Doctoral dissertation, Department of Mathematics). Stockholm University, Stockholm, United States.

[24] Xiong, L., Manathunga, V., Luo, J., Dennison, N., Zhang, R., & Xiang, Z. (2023). AutoReserve: A Web-Based Tool for Personal Auto Insurance Loss Reserving with Classical and Machine Learning Methods. Risks, 11(7), 131. doi:10.3390/risks11070131.

[25] 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. doi:10.1016/j.engappai.2019.01.010.

[26] Mahohoho, B. (2024). The Innovative Development of the IFRS17 Formulated Brighton Mahohoho Inflation-Adjusted Automated Actuarial Loss Reserving Model: Harnessing Advanced Random Forest Techniques for Enhanced Data Analytics in Fire Insurance. Global Journal of Human-Social Science, 24(13), 51–107. doi:10.34257/ljrsvol24is13pg51.

[27] Mahohoho, B. (2024). The IFRS17 Regulated Travel Insurance Intelligent Non-Linear Regression Based Inflation Adjusted Frequency-Severity Automated Loss Reserve Risk Pricing and Underwriting Model with Applications of the Actuarial Specific Gaussian Process Regression (GPR) Model. Global Journal of Human-Social Science, 24(14), 39–92. doi:10.34257/ljrsvol24is14pg39.

[28] Clemente, C., Guerreiro, G. R., & Bravo, J. M. (2023). Modelling Motor Insurance Claim Frequency and Severity Using Gradient Boosting. Risks, 11(9), 163. doi:10.3390/risks11090163.

[29] Mahohoho, B., Chimedza, C., Matarise, F., & Munyira, S. (2024). Artificial Intelligence-Based Automated Actuarial Pricing and Underwriting Model for the General Insurance Sector. Open Journal of Statistics, 14(03), 294–340. doi:10.4236/ojs.2024.143014.

[30] Cohen, L., Manion, L., & Morrison, K. (2002). Research Methods in Education. Routledge, London, United Kingdom. doi:10.4324/9780203224342.

[31] Creswell, J. W. (2014). Research Design: Qualitative, Quantitative and Mixed Methods Approaches (4th Ed.). Sage Publications, Thousand Oaks, United States.

[32] Denzin, N. K., & Lincoln, Y. S. (2018). The SAGE Handbook of Qualitative Research (5th Ed.). SAGE, Los Angeles, United States.

[33] Tukey, J. W. (1977). Exploratory data analysis. Addison-Wesley, Reading, United Kingdom.

[34] Wickham, H. (2016). ggplot2: Elegant Graphics for Data Analysis. Springer-Verlag, New York, United States. doi:10.1007/978-3-319-24277-4_9.

[35] Wilkinson, L. (2012). The Grammar of Graphics. Handbook of Computational Statistics. Springer Handbooks of Computational Statistics, Springer, Berlin, Germany. doi:10.1007/978-3-642-21551-3_13.

[36] Zeileis, A., Kleiber, C., & Jackman, S. (2008). Regression Models for Count Data in R. Journal of Statistical Software, 27(8). doi:10.18637/jss.v027.i08.

[37] Field, A. (2013). Discovering Statistics with IBM SPSS Newbury Par8. SAGE, Thousand Oaks, United States.

[38] McCullagh, P. (2019). Generalized linear models. Routledge, New York, United States. doi:0.1201/9780203753736.

[39] Frees, E. W., Derrig, R. A., & Meyers, G. (2014). Predictive Modeling in Actuarial Science. Predictive Modeling Applications in Actuarial Science, 1–10. doi:10.1017/cbo9781139342674.001.

[40] Maaten, L. V. D., & Hinton, G. (2008). Visualizing data using t-SNE. Journal of Machine Learning Research, 9(Nov), 2579-2605.

[41] Hinton, G. E., & Roweis, S. (2002). Stochastic neighbor embedding. Advances in Neural Information Processing Systems 15 (NIPS 2002), 9-14 December, 2002, Vancouver, Canada.

[42] Everitt, B. S., Landau, S., Leese, M., & Stahl, D. (2011). Cluster Analysis. Wiley-Blackwell, Hoboken, United States. doi:10.1002/9780470977811.

[43] Ester, M., Kriegel, H. P., Sander, J., & Xu, X. (1996, August). A density-based algorithm for discovering clusters in large spatial databases with noise. The Second International Conference on Knowledge Discovery and Data Mining (KDD-96), 2-4 August, 1996, Portland, United States.

[44] Jain, A. K., & Dubes, R. C. (1988). Algorithms for clustering data. Prentice-Hall, Inc, Saddle River, United States.

[45] Kaufman, L., & Rousseeuw, P. J. (1990). Finding Groups in Data. Wiley Series in Probability and Statistics, John Wiley & Sons, Hoboken, United States. doi:10.1002/9780470316801.

[46] Hastie, T., Tibshirani, R., & Friedman, J. (2009). The Elements of Statistical Learning. Springer Series in Statistics. Springer New York, United States. doi:10.1007/978-0-387-84858-7.

[47] James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). An introduction to statistical learning: with applications in R. Springer, New York, United States. doi:10.1007/978-1-4614-7138-7.

[48] Bisias, D., Flood, M., Lo, A. W., & Valavanis, S. (2012). A survey of systemic risk analytics. Annual Review of Financial Economics, 4, 255–296. doi:10.1146/annurev-financial-110311-101754.

[49] Pritchard, A., Antle, M., & Smith, A. (2019). Exploring the Impact of IFRS 17 on Risk Pricing in Insurance Models. Journal of Actuarial Science, 5(1), 34–50.