Pricing Models in Individual Health Insurance

Ângelo Cunha[1] and A. Manuela Gonçalves[1,2](✉)

[1] Department of Mathematics, University of Minho, Guimarães, Portugal
mneves@math.uminho.pt
[2] University of Minho, Centre of Mathematics, Guimarães, Portugal

Abstract. Accurate premium estimation is fundamental in insurance pricing, ensuring both fairness for policyholders and financial stability for insurers. Generalized Linear Models (GLMs) have been widely adopted due to their interpretability and compliance with regulatory standards. However, recent advancements in Machine Learning (ML) offer improved predictive accuracy by capturing non-linear relationships and complex risk structures. This study evaluates the performance of GLMs against ML models such as Decision Trees, Random Forest, Gradient Boosting (GBM), and XGBoost in predicting claim frequency and severity. Results demonstrate that GBM and XGBoost outperform GLMs in predictive accuracy, yet challenges remain in ensuring model interpretability and regulatory transparency. While GLMs provide an established actuarial framework, ML techniques offer optimization potential, leading to enhanced risk assessment and fairer pricing. The findings highlight the trade-off between predictive accuracy and interpretability in insurance pricing, emphasizing the need for hybrid approaches that integrate the strengths of both GLMs and ML models. This study contributes to the ongoing evolution of actuarial science by demonstrating how data-driven techniques can complement traditional methodologies while maintaining compliance with industry regulations.

Keywords: Actuarial Science · Generalized Linear Models · Machine Learning · Predictive Modeling · Pricing

1 Introduction

Insurance pricing is a critical process that balances fairness for policyholders with financial stability for insurers. Traditionally, Generalized Linear Models (GLMs) [11] have been the industry standard due to their interpretability and widespread acceptance by regulators [?]. GLMs provide a structured framework for modeling claim frequency and severity while maintaining transparency in pricing methodologies [4]. However, recent advancements in Machine Learning (ML) techniques, such as tree-based models and deep learning, offer promising alternatives that enhance predictive accuracy by capturing non-linear relationships and complex risk structures [15].

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2026
O. Gervasi et al. (Eds.): ICCSA 2025 Workshops, LNCS 15887, pp. 347–364, 2026.
https://doi.org/10.1007/978-3-031-97589-9_24

This study focuses on individual health insurance pricing, comparing GLMs with ML techniques, specifically Decision Trees, Random Forest, Gradient Boosting (GBM), and XGBoost. The primary goal is to evaluate whether ML models can improve claim frequency and severity predictions, leading to more precise pricing strategies. Prior research highlights the benefits of ML in actuarial science, including applications in frequency and severity modeling [8]. Despite their strong predictive capabilities, ML models pose challenges related to interpretability, regulatory compliance, and potential bias [14].

Regulatory bodies require transparent pricing methodologies, which traditional GLMs provide but ML models often lack [14]. The use of black-box models in a highly regulated industry raises concerns about fairness, bias, and accountability [13]. As Olhede and Wolfe (2018) [14] point out, predictive models must adhere to strict guidelines ensuring that decision-making processes remain explainable and justifiable.

Given these challenges, this research aims to identify the best-performing models while ensuring they remain practical for actuarial and regulatory applications. This study contributes to the evolving field of actuarial science and data-driven risk assessment by exploring the trade-off between accuracy and explainability in insurance pricing. Moreover, it highlights the potential of hybrid approaches that integrate ML techniques with traditional actuarial methods to optimize predictive performance while preserving interpretability.

2 Generalized Linear Models: Application to Pricing

Insurance pricing relies on a tariff system that evaluates historical claims data and policyholder characteristics to determine fair premiums. The cost associated with a policy i can be divided into two primary components:

- **Frequency** (N_i): Number of claims per exposure unit.
- **Severity** (X_{ij}): Total loss amount per claim.

The total loss for policy i is given by:

$$S_i = \sum_{j=1}^{N_i} X_{ij}. \tag{1}$$

The expected loss or pure premium is computed as:

$$E[S_i] = E[N_i] \cdot E[X_i]. \tag{2}$$

For an insurance portfolio of n policies, the total expected loss and variance are:

$$\mu_S = \sum_{i=1}^{n} E[S_i], \tag{3}$$

$$\sigma_S^2 = \sum_{i=1}^{n} \mathrm{Var}[S_i]. \tag{4}$$

Frequency Modeling

Claim frequency is often modeled using the **Poisson distribution**, given by:

$$P(N = x) = \frac{e^{-\lambda}\lambda^x}{x!}, \quad E[N] = \lambda, \quad \mathrm{Var}[N] = \lambda. \tag{5}$$

When overdispersion is present ($\mathrm{Var}[N] > E[N]$), the **Negative Binomial distribution** is preferred:

$$P(N = x) = \frac{\Gamma(x+\kappa)}{\Gamma(x+1)\Gamma(\kappa)} \left(\frac{\kappa}{\kappa+\mu}\right)^{\kappa} \left(\frac{\mu}{\kappa+\mu}\right)^{x}. \tag{6}$$

Severity Modeling

Claim severity is commonly modeled using the **Gamma distribution**:

$$f(y) = \frac{y^{\alpha-1}e^{-y/\theta}}{\theta^{\alpha}\Gamma(\alpha)}, \quad E[Y] = \theta\alpha, \quad \mathrm{Var}[Y] = \theta^2\alpha. \tag{7}$$

An alternative is the **Inverse Gaussian distribution**, useful for heavy-tailed severity data:

$$f(y) = \sqrt{\frac{\lambda}{2\pi y^3}} \exp\left(-\frac{\lambda(y-\mu)^2}{2\mu^2 y}\right), \quad E[Y] = \mu, \quad \mathrm{Var}[Y] = \frac{\mu^3}{\lambda}. \tag{8}$$

GLM Framework for Pricing

Generalized Linear Models (GLMs) are used to estimate claim frequency and severity. The **log-link function** is commonly applied:

$$\ln(\mu_i) = \beta_0 + \sum_{j=1}^{p} \beta_j x_{ij}. \tag{9}$$

This results in a **multiplicative model**:

$$\mu_i = e^{\beta_0} \times e^{\beta_1 x_{1i}} \times \ldots \times e^{\beta_p x_{pi}}. \tag{10}$$

Exposure, Offsets, and Weights

Exposure (Exposure_i) accounts for varying risk durations. In GLMs, exposure can be incorporated as an **offset**:

$$g(\mu_i) = \beta_0 + \sum_{j=1}^{p} \beta_j x_{ij} + \log(\text{Exposure}_i). \tag{11}$$

Alternatively, exposure can be included as a **weight variable** to adjust for variance:

$$\text{Var}(y_i) = V(\mu_i) \cdot a(\phi) \cdot \omega_i. \tag{12}$$

Where:

- $\text{Var}(y_i)$ is the variance of the ith observation.
- $V(\mu_i)$ is the variance function.
- $a(\phi)$ is a function of the dispersion parameter ϕ.
- ω_i is the weight assigned to observation i, commonly related to exposure in insurance data.

Premium Components

The total premium charged consists of:

- **Pure Premium**: Expected loss.
- **Security Margin**: Covers uncertainty.
- **Other Charges**: Administrative costs and taxes.

Policyholders with similar risk factors are grouped into **tariff classes** and charged the same premium.
Using GLMs, insurers can improve risk segmentation, optimize pricing strategies, and enhance financial stability.

3 Methodologies

Preprocessing

The following preprocessing steps were applied using R: Data anonymization: Removed personally identifiable information and regional identifiers for privacy compliance. Handling missing values and unrealistic data points: Ensured data integrity by removing incomplete or implausible records. Although alternative imputation strategies—such as mean substitution or model-based imputation—were considered, they were ultimately deemed inappropriate for this specific dataset. Imputing values without a strong underlying justification could have

introduced artificial patterns or noise, particularly in an actuarial context where reliability and explainability are critical. Normalization: Applied to numerical features for ML models to enhance training stability. Encoding categorical variables: Converted categorical data into numerical formats. Outlier handling. These preprocessing techniques ensured that the dataset was optimized for both GLMs and ML models, allowing for a robust comparison of pricing methodologies.

Generalized Linear Models (GLMs)

GLMs were used as the baseline models for health insurance pricing due to their interpretability and regulatory acceptance.

For claim frequency, the Poisson and Negative Binomial distributions were evaluated. The Negative Binomial GLM was selected as the best model, as it effectively handled overdispersion (i.e., when the variance exceeds the mean), which was a key issue in the dataset.

For claim severity, models suited for positive, right-skewed continuous distributions were considered, including Gamma GLM, Inverse Gaussian GLM, and Gaussian GLM. Given the skewed nature of incurred claims, the Gaussian model was ruled out, and the Inverse Gaussian GLM was selected as the best-performing model.

The primary goal in GLM modeling is not to achieve a perfect fit to the data's distribution but to identify the distribution that best approximates the observed data while ensuring interpretability and business applicability.

To determine the best GLM models, the following techniques were applied:

- Stepwise Akaike Information Criterion (AIC) – sequentially removed non-significant predictors to balance model goodness-of-fit and complexity;
- Bayesian Information Criterion (BIC) – used to prioritize simpler models when differences in fit were marginal;
- Deviance analysis – assessed how well each model explained variability in the data;
- Residual analysis – identified systematic patterns that could indicate model misspecification;
- 10-fold cross-validation - ensured robustness of model predictions on unseen data;
- Performance metrics (MSE and MAE) – used to quantitatively compare model accuracy.

To refine the rating structure, multiple models were tested, some including interaction effects between categorical variables (e.g., Age Band * Gender). Only interactions that significantly improved predictive accuracy were retained in the final models.

Machine Learning Models (MLM)

To explore improvements in predictive accuracy, four machine learning (ML) models were tested:

Decision Tree Regression, Random Forest, Gradient Boosting Machine (GBM), XGBoost.

To ensure ML models could effectively process the dataset, the following preprocessing steps were applied:

- Feature Engineering – created Claim Frequency (Number of Claims / Exposure) and Claim Severity (Incurred Claims / Exposure), as ML models do not incorporate exposure as an offset;
- Normalization – applied to improve convergence of algorithms sensitive to data scaling, such as GBM and XGBoost;
- Categorical Encoding – converted categorical features (Age Band, Gender, Region, Benefit Type) into a numerical format;
- Handling Imbalanced Data – removed INS. D (Benefit Type) due to having only one observation, which could bias the model and disrupt cross-validation;
- Train-Test Split – data was split into 80% training and 20% testing to ensure robust model evaluation;
- Hyperparameter Tuning – conducted manually.

Regarding the hyperparameter tuning, it is important to notice that, unlike GLMs, ML models require this to optimize performance. Hyperparameters were adjusted manually by iterating through different configurations and selecting those with the best performance metrics on validation data. The key parameters adjusted for each model were:

- **Decision Tree** – complexity parameter (to prevent overfitting);
- **Random Forest** – number of trees, maximum depth, and minimum samples per split;
- **GBM** – learning rate, number of boosting iterations, maximum depth, and minimum samples per leaf;
- **XGBoost** – learning rate, number of boosting rounds, maximum tree depth, and subsampling ratio.

Model Evaluation and Cross-Validation

To ensure model generalizability and reduce the risk of overfitting, **10-fold cross-validation** was employed. In this approach, the dataset was randomly partitioned into ten approximately equal-sized subsets (folds). Each model was trained on nine of these folds and validated on the remaining one, rotating the validation fold through all ten combinations. This process ensured that every observation was used for both training and validation exactly once, providing a more reliable estimate of model performance. The final performance metrics

were calculated by averaging the results across all ten iterations, offering a robust assessment of how the model would perform on unseen data. The evaluation included: The final performance was assessed using: Mean Squared Error (MSE) Mean Absolute Error (MAE) Additionally, cross-validation MAE boxplots and line plots were analyzed to assess the stability and consistency of each model across folds.

The test dataset was used to compare model performance at the end.

4 Case Study

4.1 Data Collection and Methods

In this section, the data sources and methodologies employed are described, providing a comprehensive account of the research approach undertaken in the project.

The dataset used in this study was obtained from Grupo Actuarial and consists of 233,745 individual health insurance policies. To ensure confidentiality, all regional identifiers and specific insurance details were anonymized. Additionally, records containing missing, erroneous, or nonsensical values were excluded to maintain data integrity and ensure the dataset was suitable for analysis.

The dataset includes nine variables: Age Band, which categorizes individuals into different age groups, Gender, which specifies the policyholder's gender, Benefit Type, which defines the type of insurance coverage, and Region, which provides geographical context. Additionally, Exposure represents the time an individual has been at risk, Commercial Premium denotes the amount paid for the insurance policy, Sum Insured reflects the total coverage amount under the policy, Incurred Claims measures the total amount claimed by the policyholder, and Number of Claims indicates the count of claims filed.

To prepare the dataset for modeling, several preprocessing steps were applied. First, data cleaning was performed by removing records with missing or unrealistic values. Feature engineering was then carried out by creating two key variables: Claim Frequency, calculated as the number of claims divided by exposure, and Claim Severity, obtained by dividing incurred claims by exposure. Categorical variables were converted into numerical representations to facilitate machine learning applications, and normalization was applied to improve model convergence and performance, particularly for gradient boosting algorithms.

By implementing these preprocessing steps, the final dataset was optimized for developing and evaluating Generalized Linear Models (GLMs) and Machine Learning (ML) models in insurance pricing.

4.2 Exploratory Data Analysis Response Variables

In insurance pricing, response variables represent the key financial risks associated with policyholders. This study models claim frequency using the **Number of Claims** and claim severity using **Incurred Claims**. Both variables exhibit highly skewed distributions, which require specialized statistical models for accurate prediction.

Number of Claims (Frequency Model). is a discrete variable representing the count of claims reported per policyholder. The distribution is highly right-skewed, with most policyholders not submitting any claims during the observed period. The mean number of claims per policy is 0.0419, with a standard deviation of 0.4583, indicating a large disparity between policyholders. The minimum and first three quartiles (Q1, Q2, and Q3) are all zero, confirming that the vast majority of policyholders do not file claims. However, the maximum recorded number of claims is 53, highlighting the presence of a small subset of individuals who submit significantly more claims than the average.
The skewness of this variable is 32.61, indicating extreme asymmetry, while the kurtosis is 2238.62, suggesting a heavy-tailed distribution with significant outliers. This distribution pattern is typical for insurance claim data, where a small proportion of policyholders generate most of the claims. Given the high variance relative to the mean, a Poisson regression model may not be appropriate, as it assumes equality between mean and variance. Instead, the Negative Binomial model, which accommodates over-dispersion, provides a more suitable alternative.

Incurred Claims (Severity Model). represents the total monetary amount paid in claims per policyholder. Like the Number of Claims, this variable is also highly skewed, with a mean of 184.7 monetary units (m.u.) and a standard deviation of 787.9 m.u. The minimum, first quartile, and median values are all zero, reinforcing the observation that most policyholders do not incur claims. However, the maximum recorded claim reaches nearly 10,000 m.u., demonstrating the presence of extreme cases.
The analysis of percentiles reveals that 80% of policies have incurred claims equal to zero, while the 90th percentile reaches 296.1 m.u., and the top 1% of claims exceed 4,292.3 m.u. This confirms that a small fraction of policyholders accounts for a disproportionately large share of total claim costs. The skewness value of 6.67 and kurtosis of 56.61 further confirm that the distribution is heavily right-skewed with extreme values. Given these properties, standard Gaussian models are unsuitable, and severity is better modeled using the Gamma or Inverse Gaussian distributions, both of which are designed for strictly positive and right-skewed data.

Exposure (Offset Variable). represents the total duration of risk coverage per policyholder and serves as an offset variable in Generalized Linear Models (GLMs). This ensures that claim frequency estimates account for differences in policy duration, enabling fair comparisons across policyholders with varying coverage periods. The dataset reveals that 76.6% of policies have full-year exposure, meaning they were active for an entire 12-month period. A significant proportion, 20.2%, has an exposure of less than 0.2 years, reflecting policyholders who had coverage for a much shorter duration.
Adjusting for exposure is crucial, as policies with longer coverage periods are naturally more likely to generate claims than those with shorter durations. Incor-

porating exposure as an offset variable in GLMs ensures that claim frequencies are evaluated on a per-unit-time basis, preventing bias in the estimation of risk factors (Table 1).

Table 1. Summary of Response Variables

Variable	Type	Mean	Std. Dev.	Max	Distribution Considerations
Number of Claims	Discrete	0.0419	0.4583	53	Highly skewed, requires Poisson or Negative Binomial models.
Incurred Claims	Continuous	184.7 m.u.	787.9 m.u.	9999.6 m.u.	Heavy-tailed, modeled with Gamma or Inverse Gaussian.
Exposure	Offset	-	-	1 (full year)	Adjusts claim frequency for varying policy durations.

4.3 Exploratory Data Analysis Feature Variables and Bivariate Analysis

The feature variables in this dataset include Age Band, Gender, Benefit Type, and Region, each providing key insights into the characteristics of policyholders.

Age Band: This variable categorizes individuals into six age groups. The dataset is skewed towards younger individuals, with the "25 to 35" years category being the most populous, followed by the "36 to 45" years group. Older age bands have fewer individuals, especially those above 65 years.

Gender: There is a significant gender imbalance, with male policyholders almost doubling the number of female policyholders. This disparity should be considered when analyzing gender-specific trends.

Benefit Type: The majority of individuals have "INS. A" (69.44%) and "INS. B" (30.38%) benefits. Other benefit types have very few individuals, indicating a strong preference for these two insurance types.

Region: The dataset is predominantly concentrated in Region A, which has 53.66% of the individuals. Other regions, such as Region B and Region C, follow but with significantly fewer individuals. "Region F" has an almost negligible number of individuals.

In terms of **bivariate analysis**, it was conducted in order to identify relationships between variables, revealing trends that may have impact in claim frequency and severity and looking for possible variable interactions. For pricing models, understanding these dependencies is crucial for setting fair premiums, detecting risk patterns, and improving model accuracy. For example, analyzing the correlation between exposure and incurred claims can refine risk assessment, while relationships between sum insured and claims frequency help adjust pricing strategies effectively and also help avoiding multicollinearity that can complicate the interpretation of coefficients and destabilize the model.

The correlation matrix showed weak or negligible correlations between most variables, except for a strong positive correlation (0.40) between Incurred Claims and Number of Claims. This indicates that as the number of claims increases, the total amount incurred also rises, which aligns with expectations.

The analysis also highlighted that Commercial Premium had little correlation with both Incurred Claims and Number of Claims, suggesting that policyholders who pay higher premiums do not necessarily file more claims. Exposure exhibited a weak negative correlation with the number of claims, implying that individuals with longer coverage periods do not always report more claims (Table 2).

Table 2. Correlation Matrix of the Variables

	Exposure	Commercial Premium	Sum Insured	Incurred Claims	Number Claims
Exposure	1.0000	0.0110	0.0677	0.0108	−0.1103
Commercial Premium	-	1.0000	0.0709	0.0034	0.0016
Sum Insured	-	-	1.0000	0.0779	0.0591
Incurred Claims	-	-	-	1.0000	0.4025
Number Claims	-	-	-	-	1.0000

To confirm the significance of these relationships, a Pearson's Correlation Test was conducted. Most correlations were statistically significant, except for those involving Commercial Premium, reinforcing that some variables may contribute little to predictive modeling.

4.4 Modeling The Claim Frequency and Severity Finding the Most Parcimonious GLM Model

Claim Frequency: Link Function: Logarithm

$$\begin{aligned}\eta(\text{estimated}) = &-5.3200 \\ &+ 0.3598 \cdot \text{Age_Band25 to 35} + 0.7135 \cdot \text{Age_Band36 to 45} \\ &+ 0.9851 \cdot \text{Age_Band46 to 55} + 3.9480 \cdot \text{Age_Band56 to 65} \\ &+ 5.3880 \cdot \text{Age_BandAbove 65} - 0.0433 \cdot \text{GenderMale} \\ &+ 2.1580 \cdot \text{Benefit_TypeINS. B} + 2.1300 \cdot \text{Benefit_TypeINS. C} \\ &- 29.8100 \cdot \text{Benefit_TypeINS. E} - 0.0716 \cdot \text{Region B} + 1.2230 \cdot \text{Region C} \\ &+ 0.1318 \cdot \text{Region D} + 0.6810 \cdot \text{Region E} - 28.8800 \cdot \text{Region F}\end{aligned}$$

Offset: Log of Exposure

The Negative Binomial GLM was identified as the optimal model for predicting claim frequency, selected through a comprehensive evaluation using AIC, BIC, deviance, and residual analysis. While the variable "Gender" was not statistically significant, as indicated by the Wald and ANOVA tests (p-value = 0.3425), it was retained in the model to ensure interpretability and relevance for business applications. Its inclusion did not substantially affect the model's parsimony, as the AIC and BIC values remained nearly identical with or without it.

The model analysis revealed a strong positive relationship between age and claim frequency, with individuals aged 56 to 65 filing approximately 52.93 times more claims than those in the 0 to 25 age group. This effect became even more

pronounced for policyholders aged above 65, who exhibited a 216.55-fold increase in claims compared to the baseline. Gender differences were also examined, with females demonstrating a 5.8% higher claim frequency than males, although this effect was not statistically significant.

The type of insurance coverage was another crucial factor influencing claim frequency. Policies under INS. B and INS. C showed a substantial increase in claim frequency, estimated at 8.65 and 8.41 times higher, respectively, compared to INS. A. In contrast, INS. E exhibited a drastic decrease in claims, with a claim frequency nearly 29.81 times lower than the baseline. This extreme reduction suggests potential data anomalies or unique characteristics associated with this benefit type.

Regional variations were also evident in the model. Region C exhibited the most significant increase in claim frequency, approximately 3.40 times higher than Region A. Meanwhile, Region B and Region D displayed only moderate deviations. A particularly striking result was observed for Region F, where claim frequency decreased drastically, likely due to data sparsity or insufficient policy representation in this category.

These findings reinforce the significance of age, benefit type, and region as key rating factors in insurance pricing. The insights gained from this model support better risk segmentation and more accurate premium calculations, ensuring that insurers can refine their pricing strategies based on data-driven decision-making. Additionally, even though the variable "Gender" was not statistically significant, it was retained in the frequency model due to business interpretability and regulatory considerations. An ANOVA test confirmed that including "Gender" did not significantly improve model fit (p-value = 0.3425). Despite its lack of predictive power, gender was retained to align with business applications and industry standards.

Claim Severity: Link Function: Logarithm

Estimated Model:

$$\begin{aligned}\eta(\text{estimated}) = {} & 6.76541 \\ & + 0.08226 \cdot \text{Region1 B} - 0.10544 \cdot \text{Region1 C} \\ & + 0.05213 \cdot \text{Region1 D} + 0.11703 \cdot \text{Region1 E} \\ & + 0.01820 \cdot \text{Age1 25 to 35} + 0.25587 \cdot \text{Age1 36 to 45} \\ & + 0.44167 \cdot \text{Age1 46 to 55} + 1.06005 \cdot \text{Age1 56 to 65} \\ & + 1.49170 \cdot \text{Age1 Above 65} + 0.49103 \cdot \text{BenType1 INS. B} \\ & + 0.63113 \cdot \text{BenType1 INS. C} - 0.47847 \cdot \text{BenType1 INS. E} \\ & - 0.08005 \cdot \text{Gender1 Male}\end{aligned}$$

Offset: Log of Exposure

To model claim severity, an Inverse Gaussian GLM with a log-link function was applied. The model incorporated key predictors such as age, region, benefit

type, and gender, with log exposure included as an offset to adjust for varying policy durations.

The results showed that region, age, and benefit type significantly influenced claim severity. Policyholders from Region B had an 8.6% increase in expected claim amounts, whereas those in Region C had a 10% decrease, both statistically significant. Regions D and E showed modest increases of 5.3% and 12.4%, respectively.

Age had a strong effect on severity, with claim costs rising as age increased. Compared to the baseline group (ages 0–25), individuals aged 36–45 had 29.2% higher claim amounts, while those aged 46–55 and 56–65 had increases of 55.5% and 106%, respectively. The oldest policyholders (65+) exhibited the highest severity, with a 149% increase in incurred claims.

Benefit type was also a key driver of claim severity. INS. B and INS. C increased expected claim amounts by 63.4% and 88%, respectively, relative to the baseline (INS. A). Conversely, INS. E showed a 38.1% reduction, though this effect was not statistically significant.

These findings reinforce the importance of demographic and policy-related factors in determining claim severity, highlighting the value of GLMs in risk assessment and premium calculation.

4.5 Pure Premium Model Using GLM

As stated previously, the Pure Premium is calculated as the product of the Frequency and Severity models. Given that both models are in the form of a log-link function, the linear predictors can be added together. The final model for Pure Premium is obtained by taking the exponential of the combined linear predictors. for this project, using GLM models, the combined linear predictor for the Pure Premium is:

$$\begin{aligned}\eta_{\text{Pure Premium}} = {} & 1.44541 \\ & + 0.44206 \cdot \text{Age_Band25 to 35} + 0.96937 \cdot \text{Age_Band36 to 45} \\ & + 1.42677 \cdot \text{Age_Band46 to 55} + 5.00805 \cdot \text{Age_Band56 to 65} \\ & + 6.87970 \cdot \text{Age_BandAbove 65} + 0.44773 \cdot \text{GenderMale} \\ & + 2.64903 \cdot \text{Benefit_TypeINS. B} + 2.76113 \cdot \text{Benefit_TypeINS. C} \\ & - 30.28847 \cdot \text{Benefit_TypeINS. E} + 0.04543 \cdot \text{Region B} \\ & + 1.30526 \cdot \text{Region C} + 0.18393 \cdot \text{Region D} \\ & + 0.79803 \cdot \text{Region E} - 29.35847 \cdot \text{Region F}\end{aligned}$$

Finally, the Pure Premium model is given by:

$$\text{Pure Premium} = \exp(\eta_{\text{Pure Premium}})$$

4.6 Machine Learning Models for Frequency and Severity

The study explores the use of Machine Learning (ML) techniques alongside Generalized Linear Models (GLMs) to improve predictive accuracy in claim

frequency modeling. Data preprocessing included renaming columns, converting categorical variables into factors, and creating "Frequency" and "Severity" variables to accommodate the absence of an offset in ML models. Normalization was applied to enhance model performance, and an 80-20 train-test split ensured robust evaluation. Due to insufficient observations, the "INS. D" category was excluded to prevent bias in cross-validation.

Four ML models—Decision Tree Regression, Random Forest, Gradient Boosting Machine (GBM), and XGBoost—were tested. Hyperparameter tuning played a crucial role in optimizing model performance. For Decision Trees, complexity was controlled to prevent overfitting. Random Forest tuning focused on the number of trees, depth, and minimum samples per split. GBM optimization involved tuning boosting iterations, learning rate, and tree depth to balance underfitting and overfitting. Similarly, XGBoost fine-tuned boosting rounds, tree depth, learning rate, and subsampling ratio to improve stability and generalization.

Overall, careful data preprocessing, feature engineering, and model-specific hyperparameter tuning contributed to robust model performance and better risk estimation in health insurance pricing.

Comparison of Model Performance During "training": Cross-Validation MAE for Different Frequency and Severity Models.

To deepen the model evaluation, analyzing graphics with, for example, the MAE values across folds recorded during cross-validation. These plots enable to examine not only overall performance but also variability across folds, giving a fuller view of model consistency. This analysis is essential because a model may perform well on average yet exhibit high variability across folds, which could suggest potential instability when applied to new data. By doing this, it is possible to assess the stability and consistency of the model.

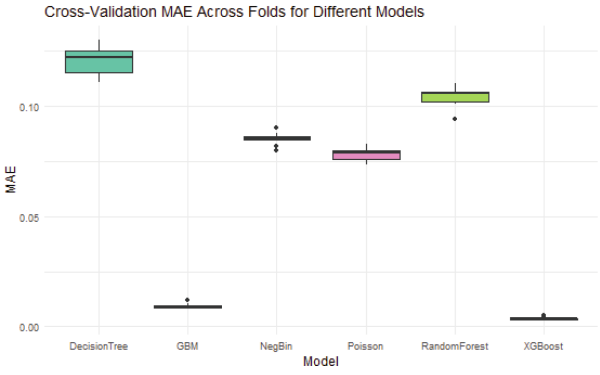

Fig. 1. Cross-Validation MAE Boxplot for the different Frequency Models.

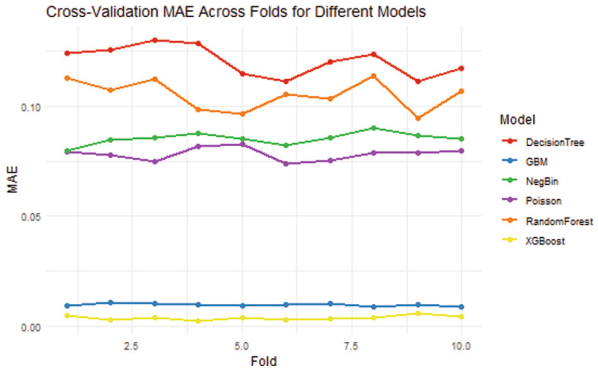

Fig. 2. Cross-Validation MAE Line plot for the different Frequency Models.

Figure 1 and Fig. 2 show that the Frequency Models using GBM and XGBoost have the lowest MAE, and are very consistent and stable across folds, indicating they perform better in terms of predictive accuracy compared to the other models. Decision Tree and Random Forest has the highest MAE, meaning it performs the worst among these models. The Negative Binomial and Poisson Models perform somewhere in between but not as well as GBM and XGBoost.

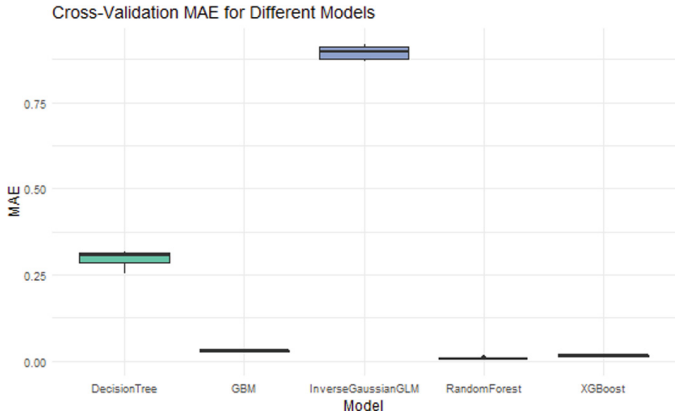

Fig. 3. Cross-Validation MAE Boxplot for the different Severity Models.

Figure 3 and Fig. 4 show that the severity Models using Random Forest, GBM and XGBoost have the lowest MAE, and are very consistent and stable across folds, indicating they perform better in terms of predictive accuracy compared to the other models. Decision Tree has the highest MAE, among the machine learning models, meaning it performs the worst among these models. The Inverse Gaussian model performs poorly when compared with the machine learning models.

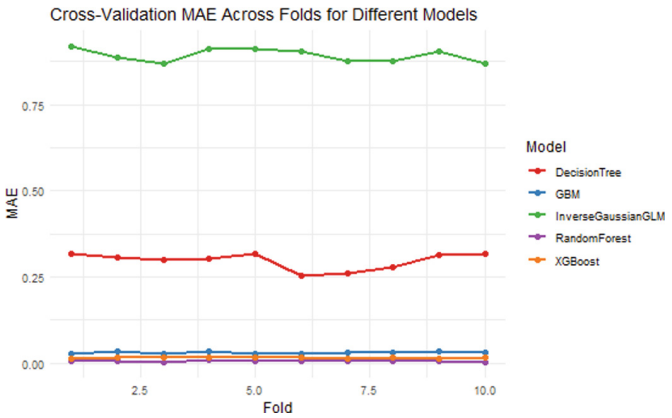

Fig. 4. Cross-Validation MAE Line plot for the different Severity Models.

Assessment of Model Accuracy in Test Phase

The performance of each model was evaluated using the Test Data. The Table 3 summarizes the results for each model, indicating the Mean Squared Error (MSE) and Mean Absolute Error (MAE) for the Frequency Models.

Table 3. Performance Metrics for Different Frequency Models

Model	MSE	MAE
Poisson GLM	0.2152	0.0798
Negative Binomial GLM	0.4716	0.0865
Decision Tree	0.2111	0.0717
Random Forest	0.2055	0.0700
GBM	0.1976	0.0626
XGBoost	0.1620	0.0493

Based on the metrics, is easy to notice that XGBoost has the lowest MSE (0.1620), and MAE (0.0493), making it the best performing model according to the metrics. GBM follows as the next best model with the second lowest MSE (0.1976), and MAE (0.0626). Random Forest and Decision Tree also perform relatively well but are outperformed by GBM and XGBoost. The Poisson GLM and Negative Binomial GLM models show higher error rates across all metrics, indicating lower performance compared to the other models. Among the models evaluated, XGBoost emerges as the best model due to its superior performance across all the considered metrics: MSE, RMSE, and MAE. Therefore, XGBoost is recommended as the optimal choice.

As for the Severity, the Table 4 summarizes the results for each model, indicating the Mean Squared Error (MSE) and Mean Absolute Error (MAE) for the Frequency Models.

Table 4. Performance Metrics for Different Severity Models

Model	MSE	MAE
Inverse Gaussian GLM	2.2340	1.3711
Decision Tree	0.3394	0.3143
Random Forest	0.0025	0.0044
GBM	0.0118	0.0275
XGBoost	0.0024	0.0150

Based on the metrics, Random Forest and XGBoost stand out as the best-performing models. The Random Forest Model has the second lowest MSE (0.0025) and the lowest MAE (0.0044). XGBoost follows closely with a slightly lower MSE (0.0024) and an MAE of 0.0150. GBM also performs well with an MSE of 0.0118 and an MAE of 0.0275. The Decision Tree shows moderate performance, while the Inverse Gaussian GLM has a substantially higher MSE (2.2340) and MAE (1.3711), indicating lower predictive accuracy compared to the machine learning models. Overall, Random Forest and XGBoost are the recommended models due to their superior performance in both MSE and MAE.

5 Conclusions

This study analyzed different pricing models for individual health insurance, comparing Generalized Linear Models (GLMs) with Machine Learning (ML) techniques to improve premium estimation accuracy. While GLMs remain widely used due to their interpretability, results show that ML models, particularly Random Forest, GBM, and XGBoost, provide higher predictive accuracy for claim frequency and severity modeling.

By integrating ML into insurance pricing, insurers can better align premiums with underlying risks, ensuring a fairer pricing structure that benefits both insurance companies and policyholders. The improved predictive performance of ML models also enhances financial stability and competitive positioning in an evolving market.

A major challenge was the availability of high-quality data, highlighting the need for stronger collaborations with companies to obtain richer datasets while reducing noise and outliers.

This research emphasizes the importance of continuous innovation in actuarial science. As insurance risks and regulatory requirements evolve, pricing methodologies must adapt to ensure accuracy and fairness. The findings lay the groundwork for future advancements in the use of ML for insurance pricing, reinforcing its practical applications and industry adoption.

6 Future Work

This study highlights several directions for future research in insurance pricing models. While Generalized Linear Models (GLMs) and Machine Learning (ML) techniques (Decision Trees, Random Forest, GBM, and XGBoost) were analyzed, further advancements can be made by incorporating neural networks and deep learning methods. These models may improve predictive accuracy but pose challenges in terms of interpretability, which is critical for regulatory compliance in insurance.

Future research should focus on enhancing model explainability while maintaining high predictive power. Methods such as permutation feature importance and partial dependence plots (PDPs) can provide deeper insights into model behavior, improving transparency in decision-making. These techniques can help identify key risk factors, making ML models more interpretable and applicable in real-world insurance pricing.

Another promising avenue is hybrid modeling approaches [10], which integrate GLMs with ML models to balance accuracy and interpretability. The combination of actuarial expertise with modern ML tools could refine risk assessment methodologies, ensuring robust, fair, and transparent pricing structures.

Additionally, the methodologies developed in this study can be extended beyond individual health insurance to other sectors, such as life, automobile, and property insurance. Each insurance line presents unique risk factors, creating opportunities for further optimization and adaptation of ML-based pricing models.

By addressing these challenges, future research can drive the evolution of insurance pricing strategies, enhancing both predictive performance and regulatory transparency while expanding ML applications in actuarial science.

Acknowledgements. I would like to express my sincere gratitude to Doutora A. Manuela Gonçalves for the invaluable support and mentorship throughout the course of this research. Lastly, I would like to express my heartfelt gratitude to my family, friends and colleagues for all the support and encouragement shown throughout my academic journey. Their belief in my abilities has been a constant source of motivation, and I am truly grateful for their enduring support. A. Manuela Gonçalves was partially financed by Portuguese Funds through FCT (Fundação para a Ciência e a Tecnologia) within the Project UID/00013: Centro de Matemática da Universidade do Minho (CMAT/UM).

References

1. Breiman, L.: Random forests. Mach. Learn. **45**(1), 5–32 (2001)
2. Chen, T., Guestrin, C.: XGBoost: a scalable tree boosting system. In: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 785–794 (2016)
3. Ferrario, A., Candelieri, A., Archetti, F.: Integrating neural networks and generalized linear models to predict health insurance claims. Artif. Intell. Med. **107**, 101877 (2020)

4. Frees, E.W., Valdez, E.A.: Hierarchical insurance claims modeling. J. Am. Stat. Assoc. **103**(484), 1457–1469 (2008)
5. Friedman, J.H.: Greedy function approximation: a gradient boosting machine. Ann. Stat. **29**(5), 1189–1232 (2001)
6. Haberman, S., Renshaw, A. E.: Generalized Linear Models and Actuarial Science. J. Royal Stat. Soc. Ser. D Stat. **45**(4), 407 (1996)
7. Hastie, T., Tibshirani, R., Friedman, J.: The Elements of Statistical Learning: Data Mining, Inference, and Prediction. 2nd edn. Springer (2009). https://doi.org/10.1007/978-0-387-84858-7
8. Henckaerts, R., Antonio, K., Clijsters, M., Verbelen, R.: A data science perspective on predicting insurance claims using generalized linear models. Insurance Math. Econ. **100**, 402–418 (2021)
9. Kohavi, R.: A study of cross-validation and bootstrap for accuracy estimation and model selection. In: International Joint Conference on Artificial Intelligence (IJCAI), pp. 1137–1145 (1995)
10. Kuo, Y.-C., Yao, C.-C.: Hybrid modeling approach in insurance pricing: combining GLMs with machine learning. J. Risk Finan. Manage. **15**(2), 95 (2022). https://doi.org/10.3390/jrfm15020095
11. Nelder, J.A., Wedderburn, R.: Generalized linear models. J. Royal Stat. Soc. Ser. A (General) **135**(3), 370–384 (1972)
12. De Jong, P., Heller, G. Z.: Generalized Linear Models for Insurance Data. Cambridge University Press (2008)
13. Spedicato, G., Dutang, C., Petrini, L.: Machine learning methods to perform pricing optimization: a comparison with standard generalized linear models. Casualty Actuarial Soc. (2023)
14. Olhede, S.C., Wolfe, P.J.: The growing ubiquity of algorithms in society: implications, impacts, and innovations. Proc. Natl. Acad. Sci. **115**(50), 12516–12523 (2018)
15. Wüthrich, M. V., Buser, C.: Tree-based machine learning methods in actuarial science. Scandinavian Actuarial J., 159–179 (2021)
16. Yang, J., Zhang, Y., Zhou, X.: Direct modeling of total loss using gradient boosting and Tweedie distribution. Ann. Actuarial Sci. **12**(1), 73–91 (2018)