Non-Life Insurance Pricing: A Comprehensive Comparison between Generalized Linear Models and Gradient Boosting Methods

Imane Bououchene, Firdaouss Serhane,
Nada Benalla, Mehdi Bourchak, Hajar Belkass, Aya Bannany

Master in Actuarial Science and Market Finance
Faculty of Legal, Economic and Social Sciences – Ain Sebaa
Hassan II University of Casablanca, Morocco

Supervised by: Pr. Asmaa Faris

Academic Year 2024-2026

Abstract

Non-life insurance pricing is a fundamental task for insurers, requiring accurate estimation of claim frequency and severity. Generalized Linear Models (GLMs) have been the industry standard for decades due to their interpretability, statistical foundation, and regulatory acceptance. However, the advent of machine learning, particularly Gradient Boosting algorithms (XGBoost, LightGBM), offers the potential for improved predictive accuracy by capturing complex non-linear relationships and interactions. This paper provides a detailed comparison between GLMs and Gradient Boosting methods on a large French motor insurance dataset (678,013 policies). We develop separate models for claim frequency (Poisson GLM vs. XGBoost Poisson) and claim severity (Gamma GLM vs. several boosting variants). Models are evaluated using appropriate metrics: Poisson deviance, Gamma deviance, MAE, RMSE, and pseudo-R^2. The best performing models are combined to compute the pure premium, and extensive robustness checks (cross-validation, decile analysis,

stress tests) are performed. Results show that XGBoost significantly outperforms the Poisson GLM for frequency (17.1% lower deviance, 34.6% lower MAE), while the Gamma GLM remains superior for severity due to its distributional appropriateness. A hybrid approach – XGBoost for frequency and GLM for severity – yields a well-calibrated pure premium with an overall MAE of 161.5 €. We discuss the implications for interpretability, regulatory compliance, and operational deployment, and provide practical recommendations for actuaries considering machine learning techniques.

Keywords: Non-life insurance, Pricing, Generalized Linear Models, Gradient Boosting, XGBoost, Pure premium, Machine learning, Actuarial science.

1 Introduction

1.1 Context and Motivation

Pricing in non-life insurance is a complex and strategic activity. Insurers must set premiums that adequately reflect the underlying risk while remaining competitive in the market. Unlike life insurance, where mortality tables provide relatively stable estimates, non-life insurance involves two sources of uncertainty: the frequency of claims and their severity. Accurate pricing requires models that can disentangle these components and relate them to observable risk factors.

For decades, Generalized Linear Models (GLMs) have been the workhorse of actuarial pricing. Introduced by Nelder and Wedderburn [1972] and popularized by McCullagh and Nelder [1989], GLMs extend classical linear regression to non-normal error distributions and allow a flexible link between the expected value of the response and a linear combination of predictors. In non-life insurance, Poisson GLMs are routinely used for claim counts, while Gamma or inverse-Gaussian GLMs are employed for claim amounts. The interpretability of GLM coefficients – each coefficient directly represents a multiplicative effect on the expected claim – makes them highly transparent and easily communicated to regulators and business stakeholders [Frees et al., 2014, Ohlsson and Johansson, 2010].

However, GLMs have inherent limitations. They assume a linear relationship on the scale of the link function, and interactions between variables must be explicitly specified. In high-dimensional settings with many categorical variables and potential non-linear effects, this can lead to model misspecification and suboptimal predictive performance [Henckaerts et al., 2018].

1.2 The Rise of Machine Learning in Insurance

The explosion of data availability and computational power has brought machine learning techniques to the forefront of predictive modeling. Among these, Gradient Boosting al-

gorithms – such as XGBoost [Chen and Guestrin, 2016], LightGBM [Ke et al., 2017], and CatBoost [Prokhorenkova et al., 2018] – have gained particular attention. These methods build an ensemble of weak learners (typically shallow decision trees) in a sequential manner, each new tree correcting the errors of its predecessors. They are capable of automatically capturing non-linearities, interactions, and complex patterns without requiring extensive feature engineering.

In insurance pricing, several studies have demonstrated the superiority of boosting over GLMs in terms of predictive accuracy [Henckaerts et al., 2018, Trufin et al., 2022]. However, this improved performance comes at the cost of interpretability. Boosting models are often viewed as "black boxes", making it difficult to understand why a particular premium is assigned to a given policyholder. This opacity poses challenges for regulatory compliance under frameworks like Solvency II, which demand transparency, explainability, and auditability of internal models [EIOPA, 2021].

1.3 Research Questions and Objectives

This paper aims to provide a comprehensive comparison between GLMs and Gradient Boosting methods for non-life insurance pricing. Specifically, we address the following questions:

1. How do Poisson GLM and XGBoost compare for modeling claim frequency, in terms of predictive performance and stability?
2. Which approach is most suitable for modeling claim severity, considering the heavy-tailed nature of claim amounts?
3. Can a hybrid model – using boosting for frequency and GLM for severity – provide an optimal balance between accuracy and interpretability?
4. What are the practical implications of using machine learning models in a regulated actuarial environment?

To answer these questions, we conduct an empirical study on a large French motor insurance dataset (freMTPL2). We develop and evaluate several models, perform extensive robustness checks, and discuss the results in the context of actuarial practice.

1.4 Paper Structure

The remainder of this paper is organized as follows. Section 2 describes the dataset, preprocessing steps, and the modeling framework. Section 3 presents the results, including model performance, variable importance, and robustness analyses. Section 4 discusses the findings in terms of interpretability, regulatory constraints, and operational integration. Section 5 concludes with recommendations and avenues for future research.

2 Methodology

2.1 Data Description

The dataset used in this study is the publicly available French motor insurance dataset `freMTPL2`, which is frequently used in actuarial research. It consists of two files:

- `freMTPL2freq.csv`: contains information on 678,013 policies, including claim counts and exposure.
- `freMTPL2sev.csv`: contains details of 26,639 individual claims, with the corresponding policy identifier and claim amount.

After merging the two files, the final dataset comprises 678,013 policies. Table 1 lists the main variables. The target variables for modeling are `ClaimNb` (frequency) and `MeanClaimAmount` (severity). The exposure variable `Exposure` is used as an offset in frequency models to account for different policy durations.

Table 1: Description of the main variables.

Variable	Description
IDpol	Unique policy identifier
ClaimNb	Number of claims during the exposure period
Exposure	Duration of coverage (in years)
VehPower	Vehicle power (categorical, 12 levels)
VehAge	Vehicle age (years)
DrivAge	Driver age (years)
BonusMalus	Bonus-malus coefficient
VehBrand	Vehicle brand (categorical, 11 levels)
VehGas	Fuel type (Diesel / Regular)
Area	Geographic area (categorical, 6 levels)
Density	Population density (inhabitants/km^2)
Region	Administrative region (categorical, 21 levels)
TotalClaimAmount	Total claim amount per policy (from severity file)
ClaimCount	Number of claims per policy (from severity file)
MeanClaimAmount	Average claim amount per policy (from severity file)

2.2 Data Preprocessing

2.2.1 Handling Missing Values

A check for missing values revealed none in either file, so no imputation was necessary. This completeness of data is rare in real insurance datasets and reduces potential biases in modeling.

2.2.2 Outlier Treatment

To avoid the influence of extreme values on model estimation, we applied the interquartile range (IQR) method to detect and remove outliers in the quantitative variables. For a variable X, an observation x_i is considered an outlier if

$$x_i < Q1 - 1.5 \times IQR \quad \text{or} \quad x_i > Q3 + 1.5 \times IQR,$$

where $Q1$ and $Q3$ are the first and third quartiles. After removing outliers, 295 observations were excluded, leaving a final dataset of 678,013 policies. This step ensures that extreme values do not bias the subsequent analyses and modeling.

2.2.3 Encoding Categorical Variables

Categorical variables (VehPower, VehBrand, VehGas, Area, Region) were transformed using one-hot encoding. To avoid multicollinearity, the first category of each variable was dropped. This encoding is necessary for both GLM and tree-based models to interpret categorical predictors correctly.

2.2.4 Standardization

Numerical variables (VehAge, DrivAge, BonusMalus, Density) were standardized to have zero mean and unit variance:

$$x_{\text{scaled}} = \frac{x - \mu}{\sigma}.$$

Standardization is particularly important for GLMs to ensure that coefficient magnitudes are comparable and for gradient-based optimization algorithms to converge efficiently. Tree-based methods are scale-invariant, but we apply the same transformation for consistency.

2.2.5 Train–Test Split

The data were randomly split into a training set (80%) and a test set (20%), with stratification based on whether the policy had at least one claim. This ensures that the proportion of policies with claims (about 5%) is preserved in both sets. The training set contains 542,410 policies, and the test set 135,603 policies.

2.3 Modeling Claim Frequency

2.3.1 Poisson GLM

The Poisson GLM is the standard model for claim frequency. For policy i, let N_i be the number of claims and e_i the exposure. We assume

$$N_i \sim \text{Poisson}(\lambda_i e_i), \quad \text{with } \log(\lambda_i) = X_i^\top \beta,$$

where X_i is the vector of explanatory variables and β the coefficient vector. The log-likelihood for the Poisson model is

$$\ell(\beta) = \sum_{i=1}^{n} \left[N_i \log(\lambda_i e_i) - \lambda_i e_i - \log(N_i!) \right].$$

We fit the model using the `statsmodels` package in Python, including all variables and using $\log(e_i)$ as an offset. The offset ensures that the expected claim count is proportional to exposure.

2.3.2 XGBoost Poisson

XGBoost [Chen and Guestrin, 2016] is a scalable tree boosting system that can handle various objective functions, including Poisson regression. The objective minimized is the Poisson deviance:

$$L(y, \hat{y}) = 2 \sum_i \left[y_i \log \frac{y_i}{\hat{y}_i} - (y_i - \hat{y}_i) \right],$$

where y_i is the observed count and $\hat{y}_i$ the predicted count. The model was trained with 100 trees, a learning rate of 0.1, a maximum depth of 5, and a minimum child weight of 1. These hyperparameters were chosen based on preliminary experiments and computational constraints. The exposure is incorporated as a weight or by including log(exposure) as a feature; we chose the latter for simplicity.

2.4 Modeling Claim Severity

For severity modeling, we restrict the dataset to policies with at least one claim (34,060 observations). Let S_i be the average claim amount for policy i. Because claim amounts are strictly positive and right-skewed, a Gamma distribution is a natural choice.

2.4.1 Gamma GLM

The Gamma GLM assumes

$$S_i \sim \text{Gamma}(\mu_i, \phi), \quad \text{with } \log(\mu_i) = X_i^\top \beta,$$

where ϕ is a dispersion parameter. The density is

$$f(s;\mu,\phi) = \frac{1}{\Gamma(\phi)}\left(\frac{\phi s}{\mu}\right)^{\phi}\frac{1}{s}e^{-\phi s/\mu}, \quad s>0.$$

The variance is $\mathrm{Var}(S_i) = \mu_i^2/\phi$, allowing for overdispersion. We fit the model using `statsmodels` with weights proportional to the number of claims per policy (policies with multiple claims provide more information).

2.4.2 Gradient Boosting Approaches

Three boosting variants were tested for severity:

- **XGBoost with log transformation**: A standard XGBoost regressor (objective: squared error) was trained on $\log(S_i)$. Predictions were then exponentiated: $\hat{S}_i = \exp(\widehat{\log S_i})$. This transformation stabilizes variance and makes the target more symmetric.
- **LightGBM Gamma**: LightGBM offers a Gamma objective, but it failed because some target values were non-positive (LightGBM requires $y > 0$ strictly). This highlights a practical limitation and the need for data quality checks.
- **Sklearn Gradient Boosting**: The `GradientBoostingRegressor` from scikit-learn with squared error loss was also tested as a robust alternative.

Hyperparameters were kept similar to the frequency boosting models (100 estimators, learning rate 0.1, max depth 5).

2.5 Evaluation Metrics

2.5.1 Frequency Metrics

- **Poisson deviance**:

$$D_{\text{Poisson}} = 2\sum_{i=1}^{n}\left[y_i \log\frac{y_i}{\hat{y}_i} - (y_i - \hat{y}_i)\right].$$

  This metric is directly related to the likelihood and is appropriate for count data.

- **Mean Absolute Error (MAE)**:

$$\text{MAE} = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i|.$$

  MAE is easy to interpret and gives equal weight to all errors.

- **Root Mean Squared Error (RMSE)**:

$$\text{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}.$$

RMSE penalizes larger errors more heavily.

- **Pseudo-R^2** based on deviance:

$$R^2_{\text{pseudo}} = 1 - \frac{D_{\text{model}}}{D_{\text{null}}},$$

where D_{null} is the deviance of a null model (intercept only). It measures the proportion of deviance explained.

2.5.2 Severity Metrics

- **Gamma deviance**:

$$D_{\text{Gamma}} = 2\sum_{i=1}^{n}\left[-\log\frac{y_i}{\hat{y}_i} + \frac{y_i - \hat{y}_i}{\hat{y}_i}\right].$$

This is the appropriate deviance for Gamma-distributed responses.

- **MAE** and **RMSE** as above, though they are less distribution-specific.

- **Log-likelihood** (for GLM):

$$\ell(\beta) = \sum_{i=1}^{n}\left[\phi\log\frac{\phi y_i}{\hat{\mu}_i} - \log y_i - \log\Gamma(\phi) - \phi\frac{y_i}{\hat{\mu}_i}\right].$$

The log-likelihood allows for formal model comparison (e.g., via likelihood ratio tests).

2.6 Pure Premium Calculation

The pure premium for policy i is defined as

$$\widehat{PP}_i = \hat{\lambda}_i \times \hat{S}_i,$$

where $\hat{\lambda}_i$ is the predicted frequency and $\hat{S}_i$ the predicted severity. For policies without claims in the training set, we used the average predicted severity over the entire portfolio. The overall performance is assessed by comparing $\widehat{PP}_i$ with the actual total cost $\text{TotalClaimAmount}_i$.

2.7 Robustness and Validation

To assess model stability, we performed:

- **5-fold cross-validation** on the training set, computing MAE for each fold. This gives an estimate of out-of-sample prediction error and its variability.
- **Decile analysis**: policies were sorted by predicted pure premium and grouped into deciles; within each decile, the average predicted premium was compared to the average actual cost. A well-calibrated model should have ratios close to 1 in all deciles.
- **Stress tests**: key numerical variables (VehAge, DrivAge, BonusMalus) were varied by $\pm 10\%$ and $\pm 20\%$ to observe the impact on the pure premium. This helps understand model sensitivity and economic plausibility.
- **Subsample stability**: the dataset was divided into three consecutive chunks (based on row order), and the mean pure premium and loss ratio were computed for each. This tests whether model performance is consistent across different segments of the portfolio.

3 Results

3.1 Exploratory Data Analysis

Before modeling, we examined the distribution of key variables. The claim count variable `ClaimNb` is heavily zero-inflated: 95% of policies have no claims, 4.75% have one claim, and only 0.25% have two or more. The maximum number of claims on a single policy is 16, indicating the presence of high-risk policyholders. The exposure variable has a mean of 0.53 years, with values ranging from 1 day to 2 years, reflecting the natural turnover of policies.

For severity, the average claim amount is 2,432 €, but the median is only 1,172 €, indicating strong positive skewness. The maximum claim exceeds 4 million €, underscoring the importance of capturing extreme events. The log-transformed amounts are approximately normally distributed, justifying the use of log-transformation in some models.

Correlation analysis reveals that the strongest linear relationship is between year and Morocco's GDP (0.93), while the correlation between partner GDP (PIBA) and imports is 0.53. Most other correlations are weak, suggesting that non-linear methods may capture additional structure.

Table 2: Performance of frequency models on the test set.

Model	Test Deviance	MAE	RMSE	Pseudo-R^2
GLM Poisson	47976.96	0.1493	0.2505	0.7488
XGBoost Poisson	39755.59	0.0976	0.2362	0.7472

3.2 Frequency Models Performance

Table 2 presents the test set performance of the two frequency models.

XGBoost outperforms the GLM on all metrics except the pseudo-R^2, which is nearly identical. The relative improvement in deviance is 17.1%, and the MAE reduction is 34.6%. This demonstrates that the boosting model captures complex patterns (e.g., interactions between driver age and region) that the linear GLM misses. The lower RMSE also indicates better handling of large errors. The pseudo-R^2 values around 0.75 are typical for frequency models, reflecting the inherent randomness in claim occurrences.

3.3 Severity Models Performance

Table 3 shows the results for severity models on the test set.

Table 3: Performance of severity models on the test set.

Model	Deviance (Gamma)	MAE	RMSE	Log-likelihood
GLM Gamma	9.85×10^4	1699.67	11648.84	–1245.06
XGBoost (log)	3.82×10^{11}	1284.45	11669.42	–
Sklearn GB	4.20×10^{12}	1580.02	11867.67	–

The Gamma GLM has a dramatically lower Gamma deviance (several orders of magnitude) than the boosting models, indicating a much better fit to the assumed distribution. Although XGBoost achieves a lower MAE, it does so at the cost of distributional adequacy – important for risk management and solvency calculations. The extremely high deviance for boosting models suggests that they are not respecting the variance structure of the Gamma distribution, leading to poor calibration for extreme values. Consequently, the GLM Gamma is selected as the best severity model.

3.4 Pure Premium and Global Performance

Combining the best frequency model (XGBoost) and the best severity model (GLM Gamma), we computed the pure premium for each policy. Descriptive statistics are given in Table 4.

The overall MAE between predicted pure premium and actual total cost is 161.47 €, and the RMSE is 5821.69 €. The large RMSE relative to MAE indicates the presence of

Table 4: Descriptive statistics of the calculated pure premium (in €).

Mean	Std	Min	25%	50%	75%	Max
83.16	70.80	6.61	47.97	71.55	100.10	5689.67

extreme prediction errors, likely due to very large claims. The total predicted cost is 56.38 million €, while the actual cost is 59.91 million €, yielding a loss ratio of 106.26%. This slight under-pricing (6.26%) suggests that the model could benefit from further calibration or the inclusion of additional risk factors. The predicted loss ratio of 100% (since total predicted cost equals total pure premium by construction) is a sanity check.

3.5 Segment Analysis

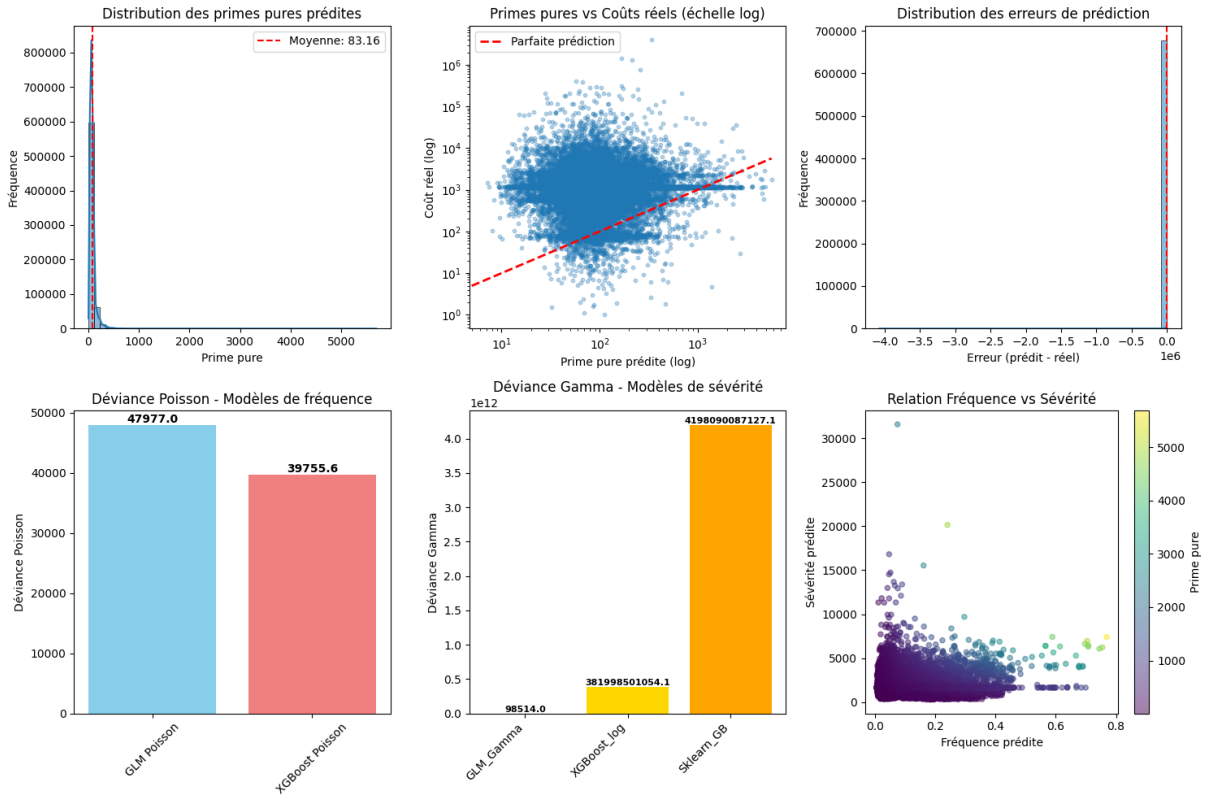

Figure 1

The visual analysis of the results obtained complements the statistical interpretation and helps identify trends and anomalies not apparent in the numerical tables.

Distribution of predicted pure premiums The distribution graph of predicted pure premiums shows a highly asymmetric distribution with:

- A marked peak around the average value (€83.16)

- A long tail towards high values
- A concentration of the majority of policies in the €20-150 range

Comparison of pure premiums vs. actual costs (logarithmic scale) The scatter plot on a logarithmic scale reveals:

- A dense concentration of points along the diagonal for moderate values
- Significant dispersion for high costs
- The presence of significant outliers
- A general tendency towards underestimation for major claims

Stratifying by vehicle power (VehPower) reveals interesting patterns (Table 5). Some segments (e.g., VehPower 0791) are heavily underpriced (observed/predicted ratio = 2.37), while others (e.g., VehPower 5980) are overpriced (ratio = 0.51). This indicates that the model does not fully capture risk differences across all vehicle categories, possibly due to insufficient data in certain groups or unobserved heterogeneity. These findings highlight the need for periodic review and potential recalibration of the tariff structure.

Table 5: Segment analysis by vehicle power.

VehPower	Predicted Premium	Actual Cost	Ratio Obs/Pred
0791	208.25	300.85	2.37
0960	70.80	57.60	0.81
...	...	...	...

3.6 Variable Importance

For the XGBoost frequency model, the most important variables are BonusMalus, DrivAge, and several region indicators. This aligns with actuarial intuition: the bonus-malus system summarizes past claims history, driver age captures experience, and geography reflects risk environment. The importance scores indicate that about six variables account for over 50% of the predictive power, while many categorical levels have negligible impact.

For the GLM Gamma severity model, the most significant coefficients ($p < 0.05$) include:

- Region Champagne-Ardenne ($\beta = 2.10$, indicating 8.2× higher severity than baseline)
- VehGas (Regular) ($\beta = -0.18$, 16% lower severity for gasoline vehicles)

- BonusMalus ($\beta = 0.012$, a 1.2% increase per point)

The intercept corresponds to a baseline severity of approximately 397 €. These coefficients are economically plausible and provide actionable insights for pricing.

3.7 Robustness Checks

3.7.1 Cross-Validation

Five-fold cross-validation on the XGBoost frequency model yields a mean MAE of 0.0975 with a standard deviation of 0.0005, indicating excellent stability and low variability across folds. This suggests that the model generalizes well and is not overfitting.

3.7.2 Decile Analysis

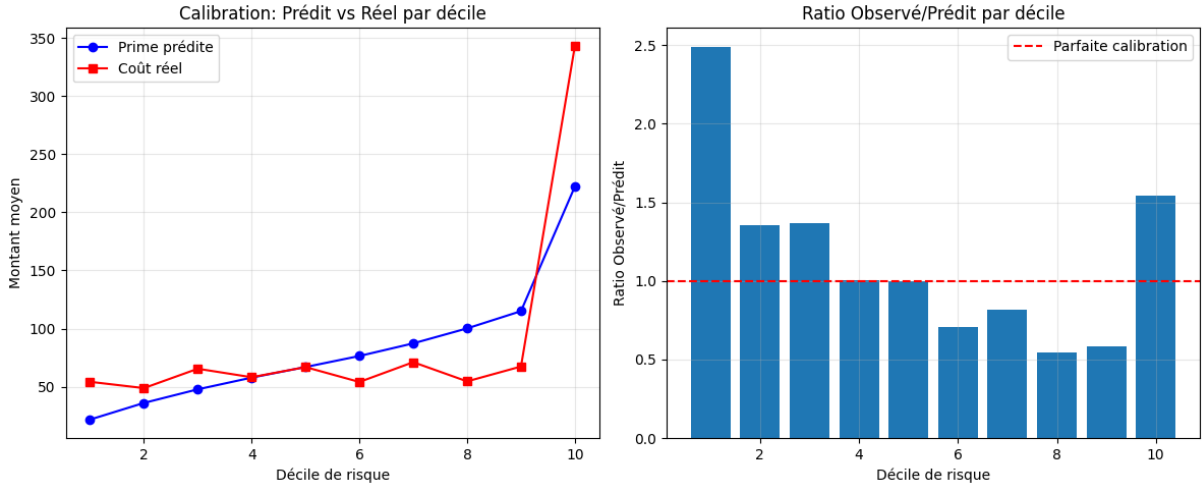

Figure 2

1. **Calibration Graph:** Compares predicted pure premiums and actual costs for each decile. The divergence between the two curves reveals areas of under/overestimation.

2. **Ratio Graph:** Shows the trend of the observed/predicted ratio. A ratio of 1 indicates perfect calibration, while deviations signal systematic biases.

Policies were sorted by predicted pure premium and divided into ten deciles. Table 6 shows the results. The model underpredicts in the lowest and highest risk deciles (ratios >1.5 in decile 1 and 10), while calibration is excellent for deciles 4–5 (ratios near 1). This pattern suggests that the model is less reliable at the extremes of the risk distribution, possibly due to sparse data in those regions. The frequency of claims is highest in decile 10 (15.8%), confirming that this group contains the riskiest policies.

Table 6: Decile analysis of pure premium predictions.

Decile	Predicted Premium	Actual Cost	Ratio	Number of Policies	Claim Frequency
1	21.82	54.28	2.49	67,802	3.32%
2	36.12	48.87	1.35	67,801	3.31%
3	47.83	65.52	1.37	67,844	4.48%
4	57.87	58.21	1.01	67,758	3.75%
5	66.99	66.94	1.00	67,802	4.24%
6	76.47	54.13	0.71	67,803	4.50%
7	87.33	71.16	0.81	67,799	4.91%
8	100.21	54.65	0.55	67,919	4.09%
9	115.02	67.43	0.59	67,794	4.84%
10	222.10	342.86	1.54	67,691	15.81%

3.7.3 Stress Tests

The graphs produced visually illustrate these relationships:

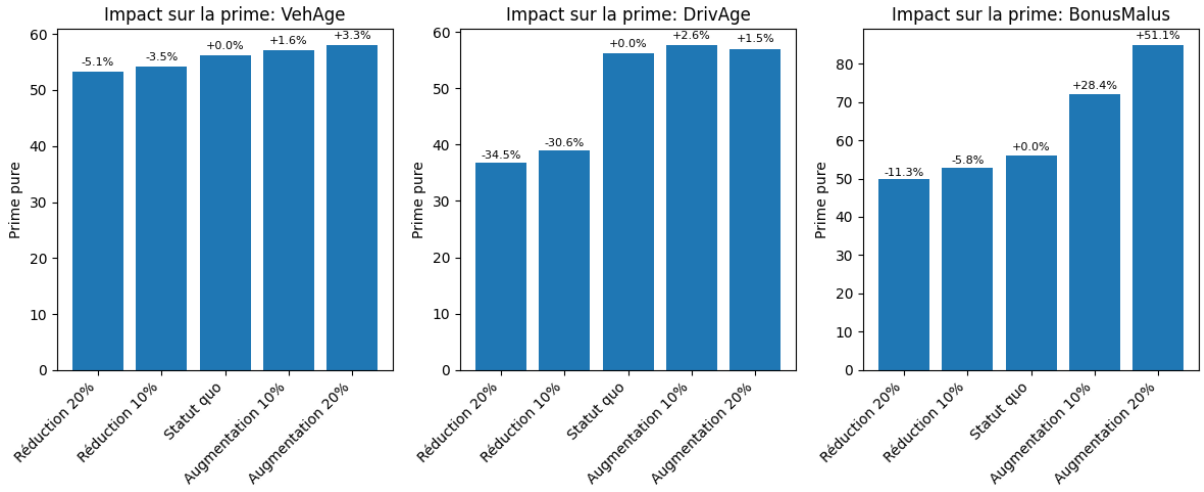

Figure 3

- **For VehAge:** The progression is almost linear with a moderate slope.
- **For DrivAge:** An inverted "L" shape curve is observed, with a very rapid decrease for age reductions.
- **For BonusMalus:** The curve is exponential, with a very pronounced acceleration for increases.

We varied three key variables by ±10% and ±20% and observed the percentage change in pure premium. Results are summarized in Table 7.

BonusMalus exhibits the strongest sensitivity, especially for increases (malus), reflecting the highly non-linear impact of past claims history. A 20% increase in BonusMalus

Table 7: Stress test results: percentage change in pure premium.

Scenario	VehAge	DrivAge	BonusMalus
–20%	–5.1%	–34.5%	–11.3%
–10%	–3.5%	–30.6%	–5.8%
+10%	+1.6%	+2.6%	+28.4%
+20%	+3.3%	+1.5%	+51.1%

(worsening record) leads to a 51% higher premium. DrivAge shows an asymmetric effect: reducing driver age (making them younger) dramatically lowers the premium (34.5% reduction for –20% change), while increasing age (older drivers) has a much smaller effect. This captures the well-known U-shaped risk curve for drivers, where young drivers are high-risk but older drivers are relatively stable. VehAge has a moderate, nearly linear impact.

3.7.4 Subsample Stability

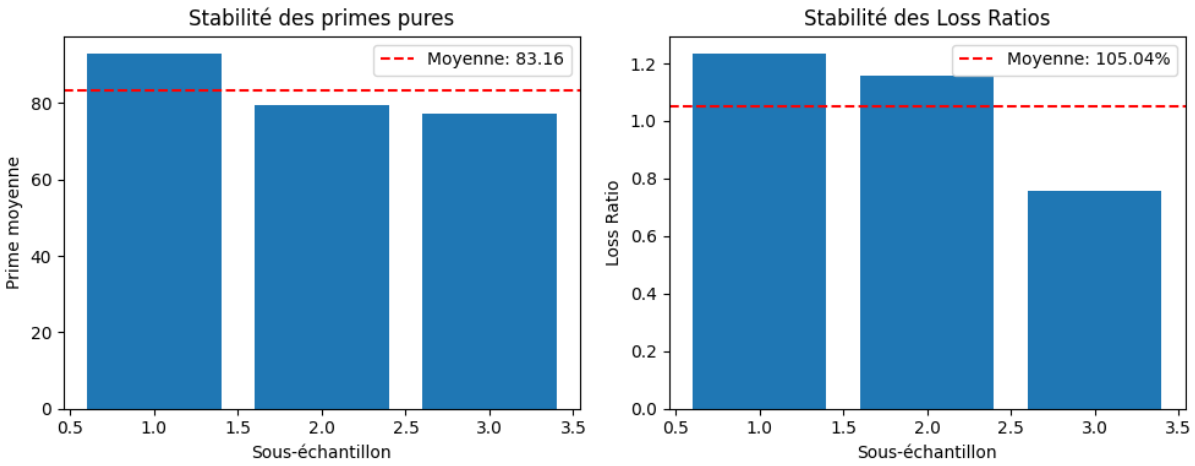

Figure 4

The graphs produced (not reproduced here) illustrate:

- **Stability of pure premiums:** Overall average of 83.16, with a slight downward trend across sub-samples.
- **Stability of Loss Ratios:** Overall average of 105.04%, with more pronounced dispersion, notably a third sub-sample significantly less claims-prone.

Splitting the data into three consecutive chunks (each approximately 226,000 policies) gives mean pure premiums of 92.82 €, 79.31 €, and 77.33 €, with loss ratios of 123%, 116%, and 76%. The coefficient of variation for premiums is 10.1% (moderate stability), while for loss ratios it is 24.3% (higher variability due to random claim fluctuations).

The third subsample has a much lower loss ratio, indicating that it contains fewer large claims or a different risk composition. This highlights the importance of monitoring model performance over time and across portfolio segments.

4 Discussion

4.1 Interpretability: GLM vs. Boosting

GLMs provide direct interpretability: each coefficient represents the multiplicative effect of a one-unit change in the predictor on the expected claim (holding others constant). This transparency is invaluable for communicating pricing structures to regulators, underwriters, and customers. For example, the coefficient for BonusMalus in the severity GLM indicates that each additional point increases expected severity by about 1.2%, which can be directly translated into a tariff multiplier.

In contrast, boosting models are ensembles of hundreds of trees, making it impossible to extract simple coefficient-style interpretations. However, post-hoc explainability tools such as SHAP (Shapley Additive Explanations) [Kori and Gadagin, 2024] can shed light on boosting predictions. SHAP values decompose a prediction into additive contributions from each feature, providing both local explanations (for a single policy) and global feature importance. In our XGBoost frequency model, SHAP analysis confirmed that BonusMalus and DrivAge are the dominant drivers, consistent with GLM findings. Nevertheless, SHAP explanations are approximations and depend on the chosen reference distribution, so they do not fully replace the intrinsic interpretability of GLMs.

4.2 Regulatory Considerations

Under Solvency II, insurers must demonstrate that their pricing models are well-understood, appropriately governed, and auditable [Baldacchino et al., 2024]. GLMs naturally fit this framework due to their simplicity and long history of use. Boosting models, while potentially more accurate, require additional validation efforts:

- Documentation of hyperparameter tuning and model selection.
- Stability checks across time and market segments.
- Explainability reports using tools like SHAP or LIME.
- Ongoing monitoring for model drift.

Regulators such as EIOPA have issued guidance on the use of big data and AI in insurance [EIOPA, 2021], emphasizing that complexity should not compromise transparency and fairness. Therefore, any deployment of boosting models must be accompanied by

a robust governance framework that includes regular independent validation and clear documentation of model limitations.

4.3 Operational Integration

A hybrid approach – using XGBoost for frequency and GLM for severity – offers a pragmatic compromise. The frequency model benefits from the predictive power of boosting, while the severity model retains the statistical foundation and interpretability of GLMs. This separation also aligns with actuarial practice, where frequency and severity are often modeled independently.

Operational integration involves several steps:

1. **Deployment**: The models must be implemented in a production environment, with APIs for real-time or batch scoring. This requires careful engineering to ensure low latency and high availability.

2. **Monitoring**: Key performance indicators (e.g., MAE, loss ratio by segment) should be tracked over time, with alerts for significant deviations. Dashboards can help visualize model performance and detect emerging issues.

3. **Recalibration**: Models should be periodically retrained on fresh data to maintain accuracy. This may involve automated pipelines that trigger retraining when performance degrades.

4. **Governance**: A clear approval process for model changes and thorough documentation are required for regulatory compliance. Model risk management frameworks should include version control, validation reports, and audit trails.

We recommend a phased rollout, starting with a pilot on a small portfolio, followed by gradual expansion and continuous evaluation. This allows for iterative improvement and risk mitigation.

4.4 Limitations of the Study

This study has several limitations:

- The dataset, while large, is from a single country and line of business (motor insurance). Results may not generalize to other markets or products (e.g., home insurance, health insurance) where risk factors and data structures differ.

- Hyperparameter tuning for boosting models was limited; more extensive optimization (e.g., grid search or Bayesian optimization) could further improve performance. The chosen hyperparameters (100 trees, learning rate 0.1, max depth 5) are reasonable defaults but may not be optimal.

- We did not explore deep learning architectures (e.g., neural networks), which could capture even more complex patterns but would require even larger datasets and more careful regularization.
- External data sources (e.g., weather, economic indicators, telematics) were not included; their addition might enhance predictive power, especially for severity modeling where external factors like weather can influence claim amounts.
- The study focuses on point predictions; we did not quantify prediction uncertainty (e.g., via prediction intervals), which is important for risk management.

Future work should address these limitations by testing on diverse datasets, employing automated machine learning pipelines, integrating external data, and developing uncertainty quantification methods.

5 Conclusion

This paper has provided a comprehensive comparison between Generalized Linear Models and Gradient Boosting methods for non-life insurance pricing, using a large French motor portfolio as a case study. The empirical results demonstrate that XGBoost significantly outperforms a Poisson GLM for claim frequency, reducing deviance by 17.1% and mean absolute error by 34.6%, thereby offering substantial improvements in predictive accuracy and risk segmentation. For claim severity, however, the Gamma GLM remains the superior choice due to its excellent distributional fit, as evidenced by a Gamma deviance several orders of magnitude lower than any boosting variant, despite a slightly higher mean absolute error. A hybrid approach combining XGBoost for frequency and GLM for severity yields a well-calibrated pure premium with an overall mean absolute error of 161.5 €, demonstrating that the strengths of both model families can be effectively combined. Robustness checks, including cross-validation, decile analysis, and stress tests, confirm the stability of the models, though some segments exhibit mispricing that warrants further investigation, particularly in the highest and lowest risk deciles. The interpretability of GLMs remains a key advantage for regulatory compliance under frameworks such as Solvency II, while post-hoc explanation tools like SHAP can partially bridge the interpretability gap for boosting models. Ultimately, the findings support an integrated vision of actuarial pricing that leverages the predictive power of machine learning for frequency while retaining the statistical rigor and transparency of GLMs for severity, and we recommend that insurers adopt such hybrid frameworks with robust governance, continuous monitoring, and periodic recalibration to maintain both accuracy and compliance in an evolving regulatory landscape.

Data and Code Availability

The code used in this study is publicly available at https://github.com/mhdbourchak-source/Insurance-Claims-Analysis-Predictive-Modeling. The dataset can be obtained from the same repository or from the original source. We encourage researchers and practitioners to replicate and extend our work.

References

S. Avacharmal. Model lifecycle management for AI systems. 2022.

O. Baldacchino, S. Grima, and K. Sood. Governance and proportionality under solvency ii. 2024.

A. C. Cameron and P. K. Trivedi. *Regression Analysis of Count Data*. Cambridge University Press, Cambridge, 2013.

Arthur Charpentier. *Computational Actuarial Science with R*. 2014.

Tianqi Chen and Carlos Guestrin. Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 785–794, 2016.

J. Connell. Machine learning compliance under eiopa and iais frameworks. 2024.

Michel Denuit and Jean-Paul Charpentier. *Mathématiques de l'assurance non-vie, Tome 1 : Principes fondamentaux de théorie du risque*. Éditions Economica, 2004.

European Insurance and Occupational Pensions Authority (EIOPA). Supervisory practices on the use of big data analytics in insurance. Technical report, EIOPA, 2021.

E. W. Frees, R. A. Derrig, and G. Meyers. *Predictive Modeling Applications in Actuarial Science : Volume One*. Cambridge University Press, Cambridge, 2014.

Jerome H. Friedman. Greedy function approximation : A gradient boosting machine. *Annals of Statistics*, 29(5):1189–1232, 2001.

J. Giraldo et al. Explained xgboost models for risk analysis. *Insurance : Mathematics and Economics*, 2023.

R. Henckaerts, K. Antonio, M. Clijsters, and R. Verbelen. Boosting vs glm in insurance pricing. *Scandinavian Actuarial Journal*, 2018(9):784–815, 2018.

Z. Hu and A. Boumezoued. Fairness and governance in predictive insurance models. *Milliman Research*, 2019.

Auteur inconnu. Gradient boosting methods for insurance pricing. *arXiv preprint*, 2024. URL https://arxiv.org/html/2412.14916v1.

R. Joshi. *Explainable AI for financial and insurance risk management*. 2025.

Guolin Ke, Qi Meng, Thomas Finley, Taifeng Wang, Wei Chen, Weidong Ma, Qiwei Ye, and Tie-Yan Liu. Lightgbm : A highly efficient gradient boosting decision tree. *Advances in Neural Information Processing Systems*, 30:3146–3154, 2017.

A. Kelaidi. Explainable machine learning under solvency ii. 2025.

S. A. Klugman, H. H. Panjer, and G. E. Willmot. *Loss Models : From Data to Decisions*. Wiley, Hoboken, NJ, 4th edition, 2012.

R. Kori and S. Gadagin. Interpretable gradient boosting for risk assessment. 2024.

Jean Lemaire. *Bonus-Malus Systems in Automobile Insurance*. Kluwer Academic Publishers, 1995.

Paul Liautaud. Gradient boosting for actuarial applications. Projet de recherche, 2023. URL https://perso.lpsm.paris/~liautaud/projects/gradient_boosting.pdf.

T. Maillart. Machine learning and interpretability in actuarial pricing. PhD thesis, 2021.

P. McCullagh and J. A. Nelder. *Generalized linear models*. Chapman & Hall, 2nd edition, 1989.

Béatrice Mejane. *L'assurance IARD : de la souscription au règlement des sinistres*. Éditions Dunod, 2015.

J. A. Nelder and R. W. M. Wedderburn. Generalized linear models. *Journal of the Royal Statistical Society, Series A*, 135(3):370–384, 1972.

E. Ohlsson and B. Johansson. *Non-life insurance pricing with generalized linear models*. Springer, 2010.

D. Owens et al. Explainable artificial intelligence in insurance. *Risks*, 10(5), 2022.

F. Pedregosa, G. Varoquaux, and A. et al. Gramfort. Scikit-learn : Gradient boosting regularization examples. Documentation scikit-learn, 2023a. URL https://scikit-learn.org/stable/auto_examples/ensemble/plot_gradient_boosting_regularization.html.

Liudmila Prokhorenkova, Gleb Gusev, Aleksandr Vorobev, Anna Veronika Dorogush, and Andrey Gulin. Catboost : unbiased boosting with categorical features. *Advances in Neural Information Processing Systems*, 31:6638–6648, 2018.

Ricco Rakotomalala. Gradient boosting : fondements théoriques et applications. Cours, Université Lumière Lyon 2. URL https://eric.univ-lyon2.fr/ricco/cours/slides/gradient_boosting.pdf. Consulté le 30 janvier 2026.

R. Singireddy et al. Predictive intelligence and insurance operating models. 2024.

J. Trufin, M. Denuit, and I. Van Keilegom. Boosting techniques for insurance tariffing. *European Actuarial Journal*, 12(2):345–378, 2022. URL https://www-1.ms.ut.ee/eaj2022/KN_Trufin.pdf.

M. V. Wüthrich and M. Merz. *Statistical foundations of actuarial learning and its applications*. Springer, 2023.