Insurance Pricing Adequacy & Risk Segmentation Analysis

Dustin Corbett

Abstract

This actuarial pricing analysis evaluates whether insurance premiums accurately reflect modeled risk using synthetic policyholder data. We employ a frequency-severity framework with Poisson and Gamma Generalized Linear Models (GLMs) to estimate expected losses. By comparing actual premiums to estimated pure premiums, we identify systemic underpricing and simulate repricing scenarios constrained by practical caps. The study delivers both technical validation and actionable recommendations for risk-adjusted pricing improvements.

1 Introduction

Actuarial pricing strategies are critical in maintaining both profitability and fairness in insurance portfolios. Leveraging machine learning and statistical modeling, this project assesses whether current premiums align with the underlying risk profile of policyholders. Using open-source synthetic data from Kaggle's Insurance Claims and Policy Data, the analysis spans:

- Modeling claim frequency with Poisson regression
- Modeling claim severity with Gamma regression
- Estimating pure premium as frequency $\times$ severity
- Detecting pricing gaps by policy product
- Simulating repricing under customer retention constraints

2 Exploratory Data Analysis

The premium distribution is skewed right, with multiple peaks, suggesting a mix of product tiers or pricing strategies. Boxplots reveal minimal average premium difference between customers who filed claims and those who did not.

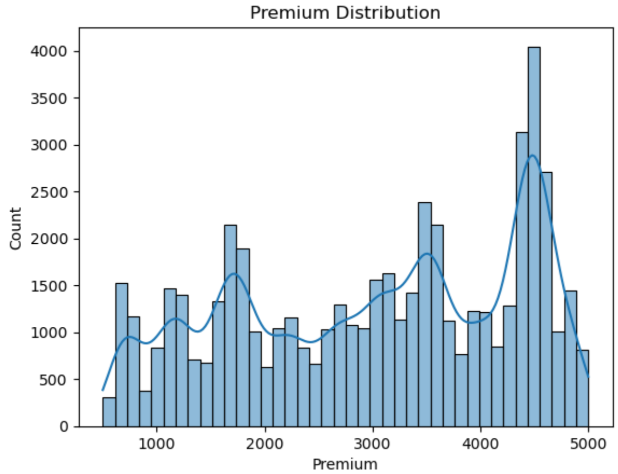

Figure 1: Premium Distribution

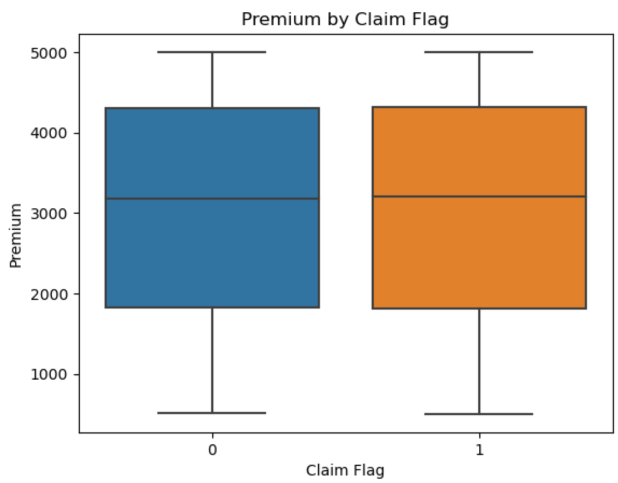

Figure 2: Premium by Claim Flag

3 Claim Modeling

3.1 Frequency (Poisson GLM)

Poisson regression reveals statistically significant predictors such as gender, occupation, and education, although the pseudo-R^2 remains low (0.004), as typical for sparse claim count data. Cross-validation yields a normalized RMSE of 35%, indicating moderate model error relative to the scale of claim counts.

3.2 Severity (Gamma GLM)

Gamma regression is fitted on observed premiums among claimants. The model achieves a high pseudo-R^2 (0.796) and good deviance residual distribution, supported by a normalized cross-validation RMSE of approximately 7.4%, indicating high predictive accuracy on the premium values. A p-value heatmap highlights statistically significant and non-significant predictors, aiding interpretation and variable selection.

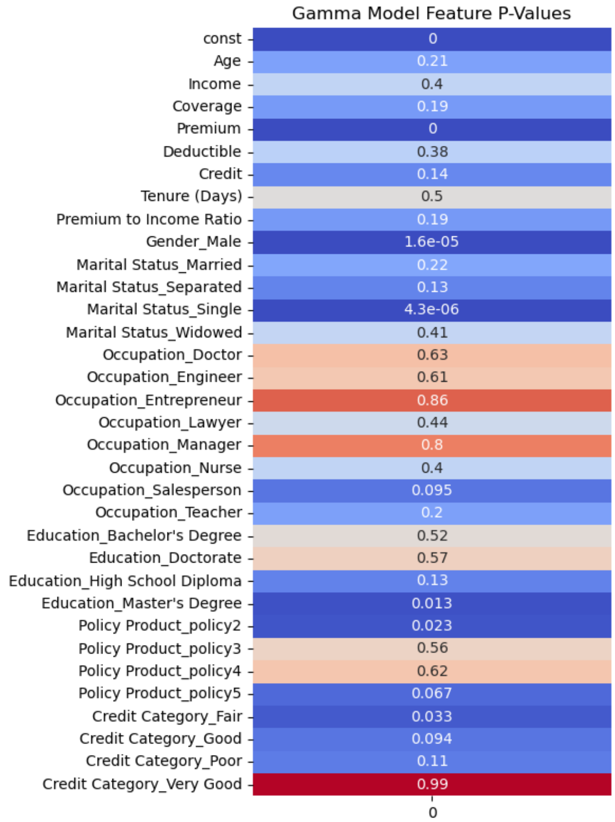

Figure 3: Gamma Model Feature P-Values

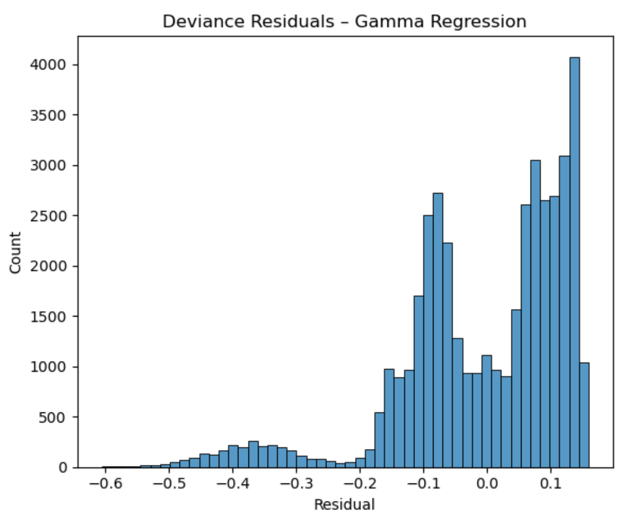

Figure 4: Deviance Residuals – Gamma Regression

4 Premium Adequacy

The estimated pure premium is computed as the product of predicted frequency and severity. When plotted against actual premiums, a consistent pattern of undercharging is observed across policyholders.

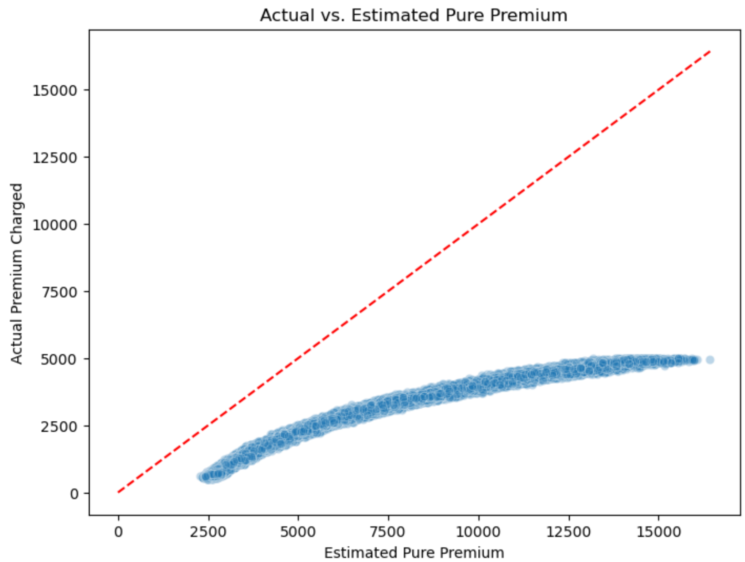

Figure 5: Actual vs. Estimated Pure Premium

5 Underpricing Analysis

We calculate pricing gaps and visualize them by policy segment. All major policy types exhibit underpricing, with a portfolio-level loss of $248 million in expected revenue.

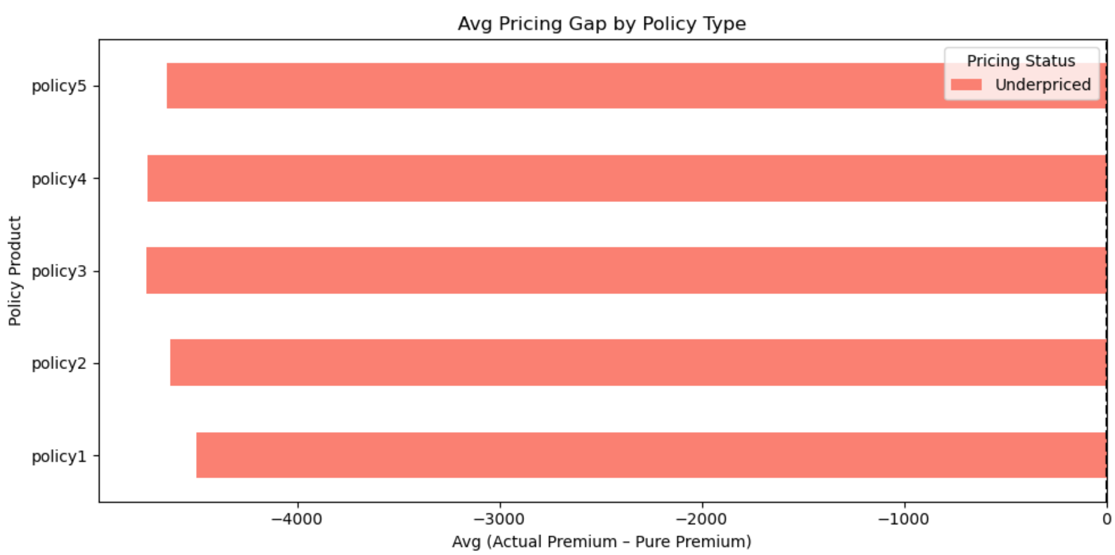

Figure 6: Avg Pricing Gap by Policy Type

6 Repricing Simulation

To mitigate underpricing, we simulate premium increases capped at 20% per policy. The resulting shift in premium distribution and segment averages are displayed below.

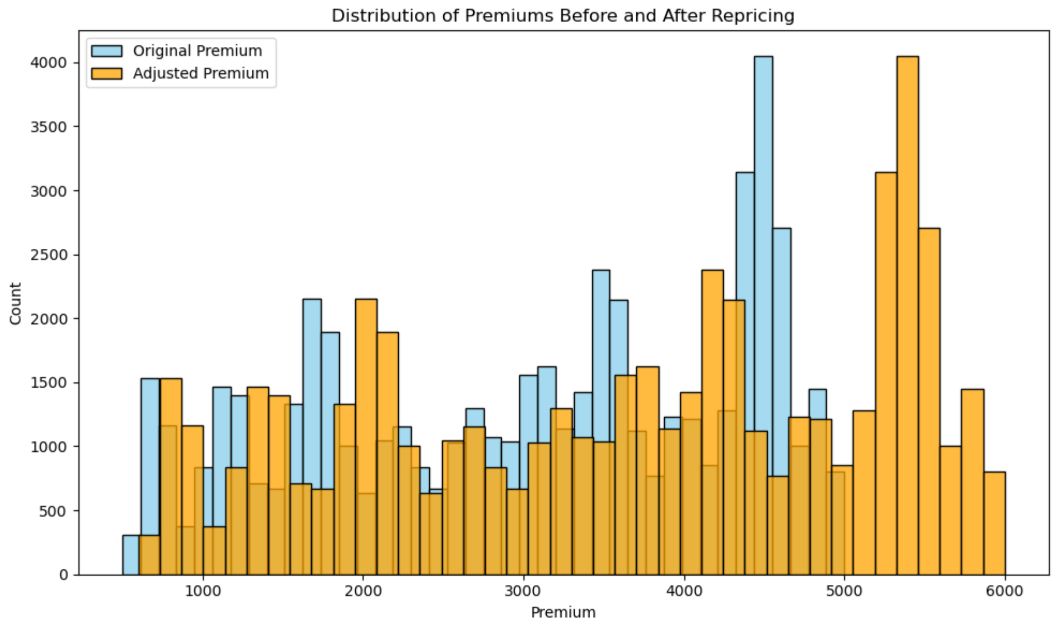

Figure 7: Distribution of Premiums Before and After Repricing

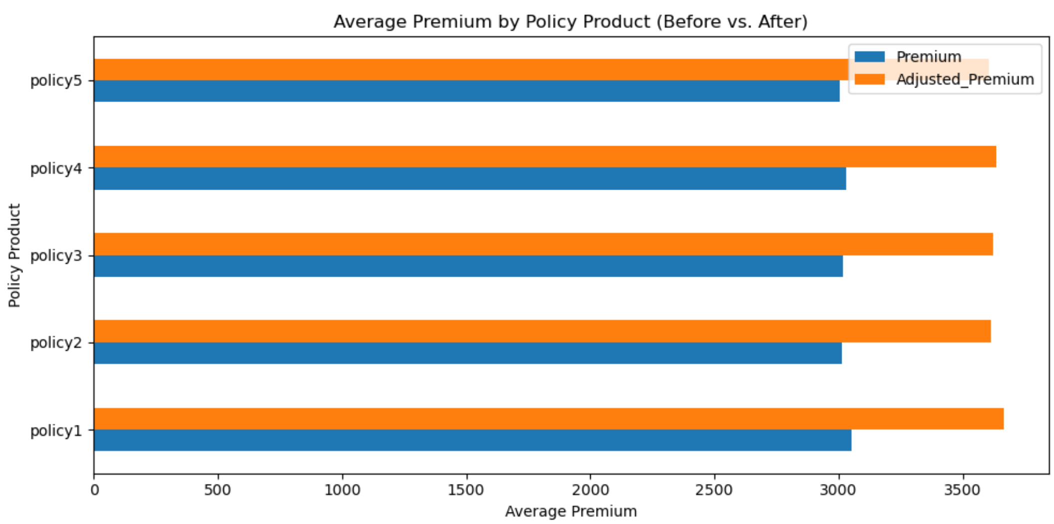

Figure 8: Average Premium by Policy Product (Before vs. After)

7 Business Recommendations

- **Revenue Opportunity:** The current portfolio is underpriced by an estimated \$32.4 million.

- **Corrective Action:** Premiums should be adjusted upward for underpriced policies, constrained by a 20% cap to reduce churn.
- **Priority Targets:** All policy product types exhibit varying degrees of underpricing. While products 3, 4, and 5 show the largest gaps, repricing should be considered across the entire portfolio for fairness and accuracy.
- **Portfolio Stability:** This repricing approach improves fairness and profitability while maintaining retention.

8 Conclusion

This project presents a robust actuarial pricing framework combining predictive modeling, validation, and repricing simulation. Key contributions include:

- Accurate modeling of frequency and severity using GLMs with cross-validation
- Identification of systemic underpricing across multiple policy segments
- Quantified impact of correcting pricing inefficiencies under retention constraints

The methodology highlights how statistical modeling can support better insurance pricing decisions, aligning premiums with expected loss and driving sustainable profitability.