

Executive Summary

This report details the development of a predictive model to calculate the pure premium for a portfolio of private vehicle insurance policies. The model successfully identifies key risk factors from the data and provides a robust, interpretable framework for pricing individual policyholders, which has been implemented in an interactive dashboard.

The final analysis reveals that the primary drivers of risk are a combination of driver history, vehicle characteristics, and policyholder demographics. Notably, factors such as **driving experience, vehicle ownership, vehicle year, gender, marital status, and past accidents** amongst a host of others are highly significant in predicting the likelihood of a claim.

Interestingly

- male married drivers were more likely to file a claim
- Older drivers were less likely to file a claim
- Higher annual mileage was associated with a higher claim probability, which aligns with the increased exposure to risk
- Although vehicles being older were associated with higher claim probability, this effect wore down as the experience of drivers increase, suggesting most of the claims from older vehicles are from younger drivers

For claim severity, **driving experience, vehicle age, and credit score** were found to be key indicators of potential cost. The model confirms that less experienced drivers, non-vehicle owners, and those with a history of accidents represent a higher risk and therefore warrant a higher premium.

The pricing tool developed from this model allows the pricing team to instantly calculate a pure premium by inputting a policyholder's specific profile, ensuring a consistent, transparent, and data-driven approach to risk assessment.

Technical Details

1. Modelling Approach

A frequency-severity approach was adopted to calculate the pure premium, based on the principle that

$$\begin{aligned}\text{Pure Premium} &= E[\text{CLAIMS}] \\ &= P(\text{OUTCOME} = 1) \times E(\text{CLAIMS} \mid \text{OUTCOME} = 1)\end{aligned}$$

This methodology involves developing two separate Generalised Linear Models (GLMs)

- **A Frequency Model:** To predict the probability of a claim occurring.
- **A Severity Model:** To predict the expected cost of a claim, given that one has occurred.

2. Data Preparation and Validation Framework

To ensure a robust and unbiased evaluation, the dataset was first split into a training set (80%) and a test set (20%). This was performed **before** any pre-processing to prevent data leakage, ensuring the test set remained "unseen" for a fair final evaluation.

There were missing values in the `ANNUAL_MILEAGE` and `CREDIT_SCORE` predictors, affecting $\approx 10\%$ of the data. An investigation was conducted to determine if this missingness was systematic. A series of independence tests (t-tests for continuous variables, Chi-squared tests for categorical) revealed no strong evidence that the missingness was dependent on other predictors, suggesting the data was Missing at Random (MAR).

To avoid the significant loss of statistical power and potential bias from listwise deletion, an imputation strategy was employed. After encountering technical challenges in applying a PCA-based imputation model consistently across both train and test sets, a **k-Nearest Neighbors (kNN) imputation** strategy was chosen. The kNN model was trained *only* on the training data's numeric columns and then applied to transform both the train and test sets.

3. Frequency Model: Claim Occurrence

A logistic regression model—a Binomial GLM with a logit link—was selected to model the probability of a claim.

Justification: Several models were considered, including GLMs with probit links and machine learning models like Random Forest and SVM. Machine Learning models were ultimately not selected due to their 'black box' nature. For a pricing task, model transparency and the ability to explain results to stakeholders are paramount. Therefore, the **logit link GLM (logistic regression) was selected for the final model due to its superior interpretability**. The ability to convert logit coefficients into odds ratios allows for clear business insights, such as stating that a one-unit increase in a predictor changes the odds of a claim by a specific percentage. This level of transparency was deemed more valuable than the minor uplift in predictive power offered by the less interpretable alternatives.

Model Selection & Diagnostics: The full model, including all two-way interaction terms, was simplified using a combination of forward and backward stepwise selection based on AIC to prioritize parsimony. A Likelihood Ratio Test confirmed that the simpler, stepwise-selected model was preferred. Diagnostic checks using randomized quantile residuals showed that the residuals were approximately normally distributed, indicating the logistic model was a good fit for the data.

4. Severity Model: Claim Size

An initial Extreme Value Theory (EVT) analysis was conducted on the claim sizes. A Mean Excess plot showed a clear upward trend, and a GEV fit to block maxima yielded a significantly positive shape parameter, confirming that the claim size distribution is heavy-tailed. Based on this, two models were developed: a Gamma GLM and a Lognormal model. A Gamma GLM with a log link was selected as the final model.

Justification: This model is the industry standard for severity modeling for several key reasons:

1. **Data Characteristics:** The Gamma distribution is well-suited for positive, right-skewed, heavy-tailed data.
2. **Log Link Function:** The log link ensures positive predicted severities and models predictor effects as multiplicative, which aligns with actuarial pricing principles.
3. **Heteroskedasticity:** The Gamma GLM's variance structure ($Var(Y) \propto \mu^2$) naturally accounts for the heteroskedasticity inherent in claims data, where larger predicted claims are associated with greater uncertainty.
4. **Interpretability:** The log link allows for straightforward interpretation of coefficients in terms of percentage changes, which is valuable for business insights and decision-making.
5. **Empirical Performance:** The Gamma GLM demonstrated a slightly lower RMSE and MAE compared to the Lognormal model, and was ultimately chosen for its improved AIC. Further tail behaviour were meticulously analysed, which demonstrated little to no systematic differences between the lognormal and gamma models.

Model Selection & Diagnostics: Similarly to the Severity Model, a full model, including all two-way interaction terms, was simplified using the same technique of stepwise selection based on AIC to prioritize parsimony. LR tests were used to justify the preference of certain models over another, and model diagnostic plots such as standardised residual plots were used to confirm goodness-of-fit and lack of identifiable patterns in the residuals. The final model's dispersion parameter was approximately 1.20, indicating mild overdispersion, which is common in insurance claims data.

5. Model Diagnostics and Challenges

Influential Points: An analysis of the Gamma model using Cook's distance identified 97 highly influential observations. Rather than removing these points, which would bias the model and ignore legitimate high-risk scenarios, they were investigated. The analysis confirmed they were not data errors but represented valid, high-risk profiles. To assess their impact, the model's dispersion parameter was compared

before and after their removal. The dispersion remained unchanged (approximately 1.20), indicating that these points were not the root cause of the model’s overdispersion. Given that they represent legitimate risks and their removal did not improve the overall model fit, the final decision was made to **retain them** to ensure an unbiased and robust pricing model.

Multicollinearity: VIF values were monitored during model selection. High VIFs for main effects were deemed acceptable when their corresponding interaction term was also present in the model, in adherence with the principle of hierarchy.

6. Final Model Validation on Test Set

Both the frequency and severity models were validated on the unseen test data.

Frequency Model: The final logistic model and a probit alternative were compared. On the test set, their performance was nearly identical (AUC of 0.881 vs 0.882), and a DeLong test confirmed no statistically significant difference. Therefore, the more interpretable logistic model was confirmed as the final choice.

Severity Model: The final Gamma GLM was compared against a Lognormal model (with a Duan smearing correction for back-transformation). The results showed a clear trade-off:

- The Gamma model had a slightly lower Root Mean Squared Error (RMSE), indicating it was better at avoiding very large prediction errors.
- Both models had very similar Mean Absolute Error (MAE).
- Crucially, the Lognormal model exhibited a significant systematic bias. The Gamma GLM, by contrast, was virtually unbiased, with a mean error of only -\$2.48.

Due to its lack of bias and superior performance on large claims (lower RMSE), the **Gamma GLM was confirmed as the final severity model**. The combination of the selected logistic and Gamma models provides a robust and transparent foundation for the pricing dashboard.

Improvements and Future Work

While the current model performs well, several avenues for future enhancement exist:

- The use of Machine Learning techniques (e.g., Gradient Boosting Machines or Neural Networks) could be explored to potentially capture complex non-linear relationships and interactions that GLMs might miss. However, this would need to be balanced against the requirement for model interpretability.
- Advanced imputation methods could be used instead of KNN, accounting for the relationships between **all** numerical AND categorical variables. The current imputation system overlooks relationships with categorical variables and only imputes numerical values
- Use of compound distributions to model the claim counts could be explored, such as a Negative Binomial distribution to account for overdispersion in claim counts. However we did not receive any counts data.
- Further exploration of interaction terms and non-linear effects using splines or polynomial terms could enhance model flexibility.