Modeling Pure Premium in Auto Insurance Using GLMs and Python

Douglas Fulcher

Bellevue University

Applied Data Science DSC680 T301 - Fall 2025

Instructor: Amirfarrokh Iranitalab

Modeling Pure Premium in Auto Insurance Using GLMs and Python

Business Problem Statement

Auto insurers must match premiums to each policyholder's expected claim cost. The challenge is to convert noisy signal data into calibrated risk estimates that yield fair, defensible prices. This study evaluates generalized linear models (GLMs) as a statistically principled framework that answers the questions:

- Which driver and vehicle characteristics most impact pure premium?
- How can segmentation analysis highlight risk factors while remaining fair and explainable?
- Does Tweedie GLM outperform a two-part Frequency x Severity model?
- What biases or limitations in the dataset could affect fairness?
- How much improvement over the null model can be achieved?

Background and History of GLMs in Auto Insurance Pricing

For years, auto insurers have priced policies using broad categories and manual adjustments—useful, but often too coarse to accurately reflect each driver's actual risk. As data volume and complexity grew, more rigorous, data-driven methods became necessary (Frees & Valdez, 2008). Generalized linear models (GLMs) answered that call. By allowing non-normal outcomes and flexible link functions, GLMs fit insurance realities—claim counts and amounts rarely behave like simple Gaussians (McCullagh & Nelder, 1989). With multiple risk factors in play (driver age, vehicle attributes, geography, exposure), GLMs produce calibrated estimates of claim frequency and severity that support fairer, more competitive pricing and more apparent justification to stakeholders (Ohlsson & Johansson, 2010).

Regulatory scrutiny and market competition accelerated the adoption of GLM, prompting carriers to justify the fairness of their prices—not just their value. In many personal lines markets, GLMs are now standard practice, providing transparency, stability, and a disciplined approach to moving beyond crude averages.

Data Explanation

The dataset, sourced from the CASDatasets R package (also available on OpenML and Kaggle), contains detailed records of French automobile insurance policies, including driver demographics, vehicle characteristics, and claim information. Frequency and severity data share a unique policy identifier (IDpol), enabling seamless joining. A complete data dictionary is provided in Table 1 of the Appendix.

Data retrieval via OpenML was straightforward, with no missing values. We visualized distributions and frequency (see Figures 1, 4, and 5), which revealed typical insurance patterns: long-tailed distributions and claim sparsity. We addressed these extremes through capping and transformation strategies (see Figures in the Appendix).

Additionally, we checked for multicollinearity (see Figure 3) among the predictors and found no significant issues, which supports the inclusion of all major features in our modeling. Our EDA also informed feature engineering decisions, such as binning driver ages and capping outlier values, to improve model interpretability and performance. To evaluate different capping and binning strategies, we added additional columns to the data frame to store the values of these strategies. No source data was updated in place.

Methods

This section provides a concise overview of the modeling approach, tools, and evaluation metrics used in this effort.

Modeling Approach

We model insurance risk with generalized linear models (GLMs) using insurance-friendly families (McCullagh & Nelder, 1989; Ohlsson & Johansson, 2010). We estimate three target types aligned to standard pricing workflows:

- **Frequency**: expected claim count per exposure (Poisson as baseline; Negative Binomial to evaluate overdispersion).
- **Severity**: Claim Amount divided by Claim Number (Gamma).
- **Pure premium**: expected claim cost per unit exposure (Tweedie; or Frequency * Severity).

All models use exposure-appropriate weighting (exposure for frequency/pure premium; claim counts for severity) and a log link, so coefficients act multiplicatively on the mean.

Data Selection

We used Scikit-Learn's ShuffleSplit with a fixed seed value to segment the data into an 80/20 test/train split. This was done to ensure the same train/test split was used across multiple models, where we evaluated different features and options related to capping/binning strategies. After selecting the train/test set, we needed to evaluate our weighting requirements. For frequency modeling, Exposure, which represents the length of time a policy has been in effect, was used as the sample weight. For severity and pure premium models, claim number (the number of claims during a policy term) was used. Severity was modeled only on records with a Claim Amount greater than zero. After determining our filters for each model, we needed to evaluate which predictors were likely

to influence frequency and severity. Predictors were selected from the available outputs of the EDA phase to evaluate the impacts of binning/capping strategies.

Estimation and Software

For this project, we use glum's *GeneralizedLinearRegressor*. This regressor handles the appropriate family of models (Poisson, Negative Binomial, Gamma, and Tweedie) used in Frequency and Severity modeling. Glum was chosen as an academic learning opportunity in order to expand the available packages/libraries at my disposal. During modeling, we employed a log link and penalty search with L1 (Lasso) regularization to minimize noise and improve reproducibility. The code for modeling was primarily pulled from the glum package's documentation (Quantco, 2024). However, we added a significant amount of code to allow for multiple model simulations to be run without overwriting the results of prior models. In its current state, this code can be configured using an input dictionary to define: which model configurations should run for frequency, which model configurations should run for severity, which combinations of frequency and severity model runs should be evaluated in the Two-Part method for pure premium, and which model configurations should run for Tweedie combined pure premium methodology.

Model Evaluation Metrics

To compare variations of models, we attempt to report the following for the models:

- **Mean deviance (family):** average deviance under the model's own distribution (e.g., Poisson, Gamma, Tweedie).
- **Mean Poisson Deviance (cross-family-yardstick):** computed for every model to enable apples-to-apples comparison regardless of family.
- **Improvement over null (%):** A representation of the percentage of reduction in test deviance vs. a null model that predicts a single global mean rate/level.
- **Overfitting gap:** test minus train deviance (smaller absolute gaps indicate better generalization.
- **MAE (severity & pure premium):** mean absolute error to provide scale-aware performance alongside deviance.

These metrics focus on calibration and generalization rather than on in-sample fit. These metrics are more in line with GLMs, providing ranked candidate specifications before business or regulatory review.

Analysis

This section outlines the modeling experiments conducted and the key findings from our baseline and feature engineering approaches. Detailed results and figures will be added after further review and analysis.

Baseline Models

We began by running Frequency, Severity, and Tweedie GLMs using input transformations recommended in the glum documentation. This established a baseline and confirmed that our code and data pipeline were functioning correctly. We ran a comparison model, which binned age and unbinned most of the features from the baseline model to determine

whether these changes would hurt the model. Contrarily, these manual bins outperformed the tutorial features. All models showed improvement over the null model, with test set deviance reduced by approximately 3–5% compared to the null (see Table 2 and Table 4 in the Appendix for summary metrics).

Model Comparison and Evaluation

All models were evaluated against the null model, which predicts the average outcome for all policies. Across all approaches, the best models showed only modest improvements—typically a 3–5% reduction in test set deviance over the null. This suggests that, while feature engineering and binning strategies can influence model performance, the overall predictive power of the available features is limited for this dataset.

Conclusion

For our Frequency Models, we can see that the Manual Bins strengthened stability and calibration when using the Poisson distribution. With a 0.4575 deviance and P/O RMSE of 0.101 (see Figure 7), we demonstrate more reliable frequency distributions from this model compared to the Glum baseline, which has 0.4579 and 0.114, respectively (see Figure 7).

Evaluating Severity, we see that Manual Binning again demonstrates a slight edge over the Glum baseline. Both model variants show a relatively stable ~6.5% underprediction bias, suggesting that further refinement may be needed. ElasticNet regularization benefitted both models, resulting in a small 0.096 overfit gap for the Glum baseline and only 0.091 on the manual binned model (see Figures 8 and 9).

In the Two-Part Pure Premium evaluation (Frequency * Severity), we observe that both models have a ~6.5% underprediction of the observed values in the test set. Manually binning provided a marginal edge over the baseline but did not significantly alter the results (see Figures 10 and 11).

The glum baseline performed well in our Tweedie Models. Here, we see that the baseline had a lower deviance of 76.63, with p = 1.5, compared to the manual bin models, which had a deviance of 77.28, with the lowest deviance at p = 1.7. The best Tweedie model had a ~5.1% underprediction, compared to approximately 6.0% for the manual bin model (see Figures 12 and 13).

From these results, we can see that feature engineering matters. Identifying the optimal capping and binning strategy can have a significant impact on Pure Premium. With Two-Part models, we can break down which factors drive the likelihood of loss separately from those that impact the severity of loss, which could aid in providing transparency to consumers and regulators. Tweedie models simplify model implementation, but they do come at the cost of transparency by providing factors that attempt to balance both needs. However, it does this reasonably well on this dataset by reducing the underprediction. See the Coefficient plots in the appendix for a better understanding of the top 20 features affecting each model's results.

Research Questions

Through this effort, we were able to answer the research questions we posed. We found that driver and vehicle age have a meaningful impact on our models. We also demonstrated how segmentation (capping/binning) strategies can alter the fit of our models to achieve business outcomes. In our case, we observed that Tweedie models were able to better predict our target, albeit at the cost of some explainability. The dataset lacked sufficient characteristics and was outdated, which may be masking important risk factors. In a more robust dataset, we would need to be careful not to use features such as age or gender unfairly. Most meaningfully, we found that this limited set of policy and severity characteristics yielded only a 3-5% improvement over random guessing. Additional feature engineering, such as investigating interactions or splines, may improve results.

Limitations

A primary limitation of this project was the scope and depth of the dataset. While the French motor insurance data provided a solid foundation for exploring GLMs, it lacked the breadth of features typically available in real-world insurance pricing—such as household composition, conviction history, detailed vehicle specifications, and coverage-level payment data. This limited feature set likely contributed to the modest improvement of the models over the null model. Additionally, the age of the data (approximately 20 years old) may not reflect current risk patterns or market conditions, further constraining the generalizability of the findings.

Challenges

As someone without an actuarial or statistics background, the steepest learning curve was understanding the core concepts of frequency and severity modeling. Interpreting statistical methods for handling long-tailed distributions and outliers was particularly challenging, leading to delays in model development. Evaluating which features to transform, cap, or bin required significant research and trial-and-error, highlighting the importance of domain expertise in actuarial modeling.

Future Uses/Applications

While this project will not result in a production model, the experience gained is directly applicable to leading and supporting future projects involving GLMs in insurance pricing. In future work, exploring more advanced modeling techniques—such as random forests or gradient boosting machines—could yield better predictive performance, especially with higher quality datasets. However, these methods must be balanced against the need for transparency and interpretability in regulated environments.

Recommendations

For similar projects, it is recommended to invest time early in understanding the data and the business context. Collaborating with domain experts can accelerate the learning process and improve feature engineering decisions. When possible, seek out or request datasets with a broader range of characteristics to maximize model performance. Additionally, prioritize transparency in modeling choices, mainly when results will be used for business or regulatory purposes.

Implementation Plan

No implementation into a production system is planned for this project. The concepts and modeling code developed here provide a strong, reusable foundation for future work. Acquiring more comprehensive data, refining feature engineering, piloting more advanced modeling techniques, and implementing robust error handling and logging would be required before considering production deployment.

Ethical Assessment

The dataset used in this project presented minimal ethical concerns. It was highly aggregated, contained no personally identifiable information, and was sufficiently dated to avoid privacy risks. In future applications involving more granular or current data, it will be important to maintain strict data protection practices and consider the fairness and transparency of any models developed.

Appendix

Potential Audience Questions

1. How did you decide on the specific binning strategies for driver and vehicle age, and did you test alternative approaches like monotonic binning?
2. Did you consider using interaction terms or polynomial features in your GLMs, and if so, what impact did they have on model performance?
3. How did you handle exposure weighting in the presence of policies with very short or very long durations?
4. What diagnostics did you use to check for overdispersion in the Poisson models, and how did you decide when to switch to Negative Binomial?
5. Can you elaborate on why OptBinning performed worse than manual binning in your case? Was it due to a lack of monotonicity, or was there something else at play?
6. What is a Generalized Linear Model, and how is it different from regular linear regression?
7. Why do insurance companies need to use models like GLMs instead of just averaging past claims?
8. What does 'binning' mean, and why is it important for modeling?
9. How do you know if your model is good or not? What does 'deviance' measure?
10. Why is it important to keep models transparent for regulators and consumers?

Tables

Table 1 – Data Dictionary

Dataset	Column Name	Column Description	Data Type	Sample Values
Frequency	IDpol	Unique policy identifier	int32	1, 3, 5
Frequency	ClaimNb	Number of claims on the policy (count)	int64	0, 1, 2

Dataset	Column Name	Column Description	Data Type	Sample Values
Frequency	Exposure	Policy exposure (fraction of year)	float64	0.10000, 0.77000, 0.00274
Frequency	Area	Territorial code (ordinal: A→F)	category	D, B, E
Frequency	VehPower	Vehicle power band (ordinal category)	int64	4, 5, 6
Frequency	VehAge	Vehicle age in years	int64	0, 2, 6
Frequency	DrivAge	Driver age in years	int64	29, 41, 55
Frequency	BonusMalus	Bonus–malus score (experience-based rating)	int64	50, 54, 95
Frequency	VehBrand	Vehicle make / brand code (nominal)	category	B12, …
Frequency	VehGas	Fuel type (nominal)	category	Regular, Diesel
Frequency	Density	Population density (numeric)	int64	54, 76, 3317
Frequency	Region	Geographic region code	category	R82, R22, R72
Severity	IDpol	Unique policy identifier (joins to frequency data)	int32	139, 190, 414

Dataset	Column Name	Column Description	Data Type	Sample Values
Severity	ClaimAmount	Total claim payout for the policy (sum)	float64	303.00, 1981.84, 10834.00

Figures

Figure 1: Severity Scale

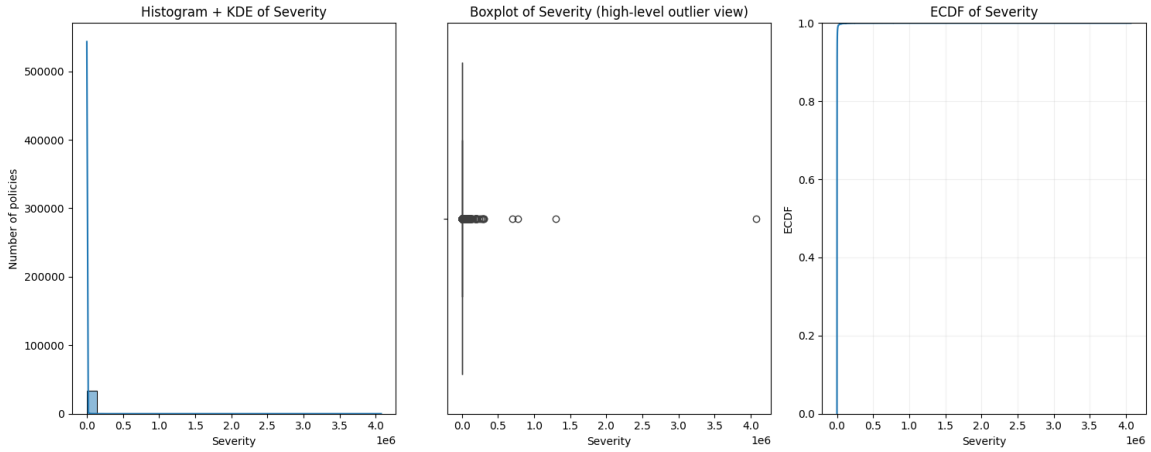


Figure 2: Log Scale Severity

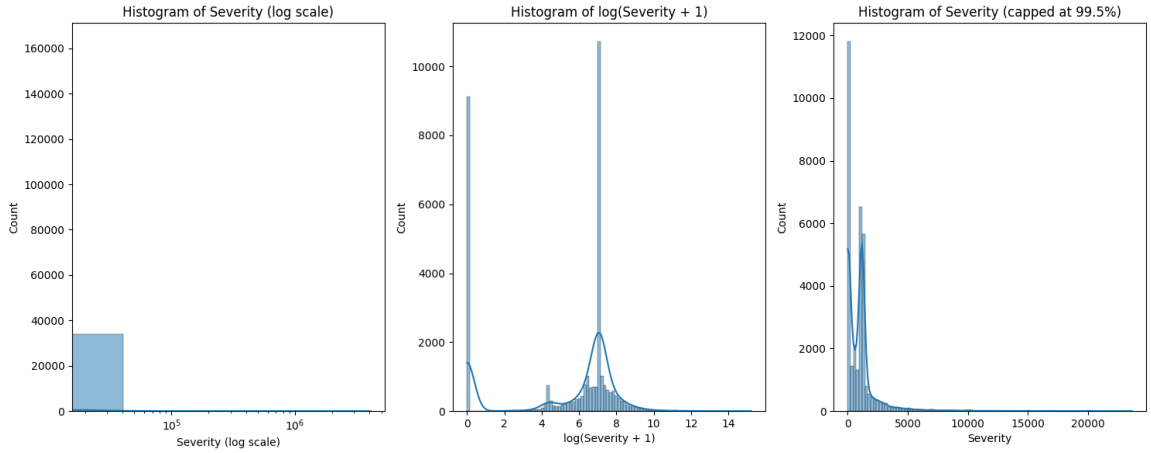


Figure 3: Feature Correlation

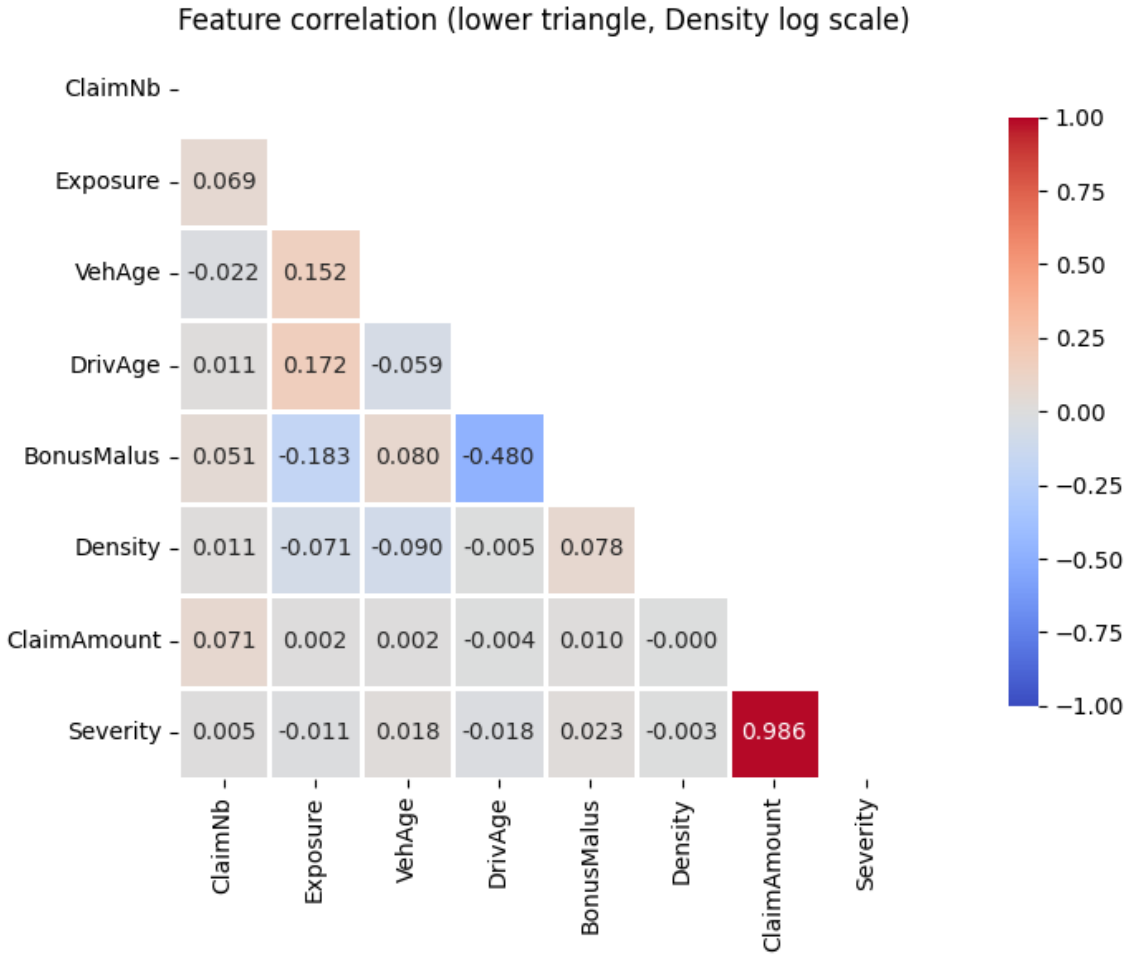


Figure 4: Driver Age Count and Frequency

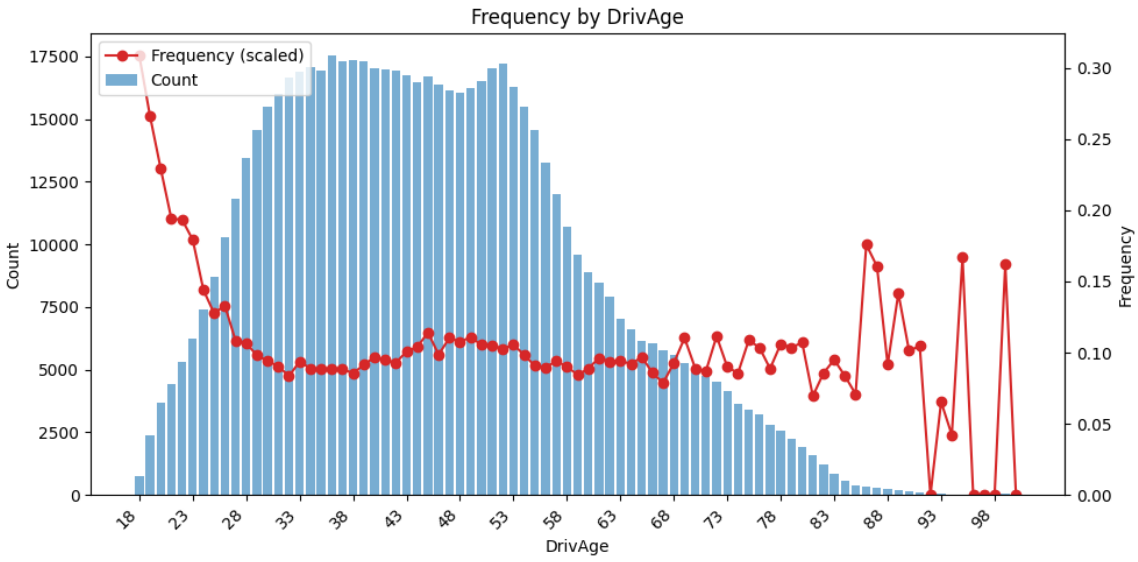
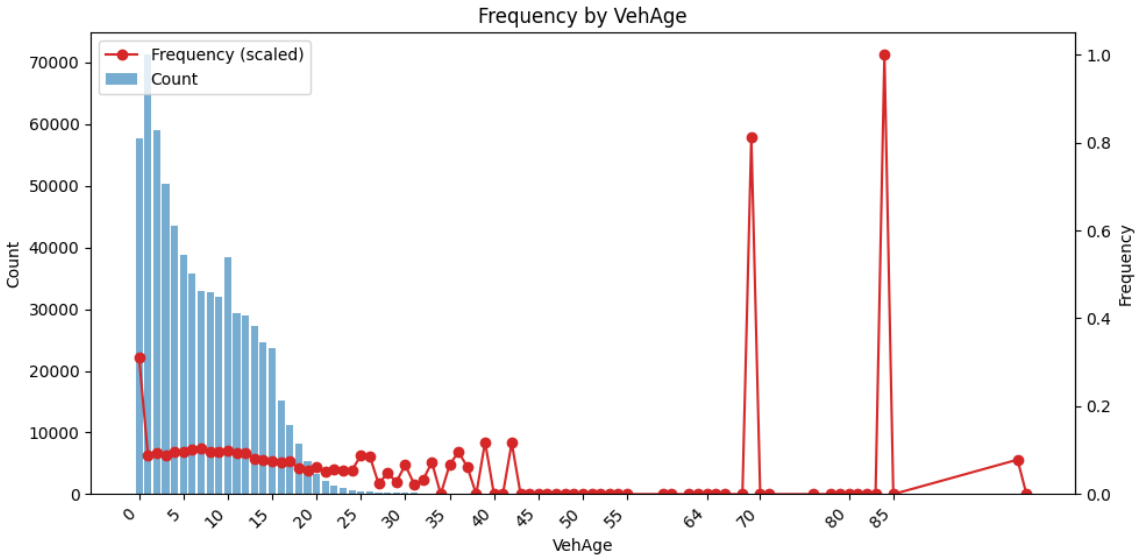


Figure 5: Vehicle Age Count and Frequency



Frequency Model Results

Glum Baseline

Figure 6: Glum Baseline Frequency Metrics

```
============================================================
FREQUENCY MODEL COMPARISON — [GLUM_BASELINE]
============================================================

Model                          Test Deviance    Improve %    Overfit
------------------------------------------------------------------------
Poisson GLM                         0.465663          4.2   0.010639
Tweedie (freq)                      0.467030          3.9   0.010629
Negative Binomial                   0.469966          3.3   0.010558

Model                 Test Deviance   Improvement %   Overfit Gap     P/O RMSE        #Features
Poisson GLM           0.465663        4.153809        0.010639        0.081325        63
Tweedie (freq)        0.467030        3.872612        0.010629        0.163027        63
Negative Binomial     0.469966        3.268150        0.010558        0.275617        56
```

Manual Binned

Figure 7: Manual Binned Frequency Metrics

```
============================================================
FREQUENCY MODEL COMPARISON — [MANUAL_BIN_GLUM_TGT]
============================================================

Model                          Test Deviance    Improve %    Overfit
--------------------------------------------------------------------
Poisson GLM                         0.465573          4.2   0.010853
Tweedie (freq)                      0.467052          3.9   0.010619
Negative Binomial                   0.470085          3.2   0.010545

Model                 Test Deviance   Improvement %   Overfit Gap     P/O RMSE        #Features
Poisson GLM           0.465573        4.172332        0.010853        0.067387        72
Tweedie (freq)        0.467052        3.867915        0.010619        0.174239        72
Negative Binomial     0.470085        3.243612        0.010545        0.249604        65
```

Severity Model Results

Glum Baseline

Figure 8: Glum Baseline Severity Metrics

```
============================================================
SEVERITY MODEL COMPARISON — [GLUM_BASELINE]
============================================================

Model                          Test Deviance    Improve %    Overfit
--------------------------------------------------------------------
Gamma + ElasticNet                  1.375032         -0.2   0.095847
Tweedie p=1.9                       1.388009         -1.2   0.122702
Gamma GLM                           1.392754         -1.5   0.128373

Model                 Test Deviance   Improvement %   Overfit Gap     P/O RMSE        #Features
Gamma + ElasticNet    1.375032        -0.248398       0.095847        0.104285        63
Tweedie p=1.9         1.388009        -1.194545       0.122702        0.146330        63
Gamma GLM             1.392754        -1.540479       0.128373        0.140495        63
```

Manual Binned

Figure 9: Manual Binned Severity Metrics

```
============================================================
SEVERITY MODEL COMPARISON — [MANUAL_BIN_GLUM_TGT]
============================================================

Model                          Test Deviance    Improve %    Overfit
--------------------------------------------------------------------
Gamma + ElasticNet                  1.367991          0.3   0.090574
Tweedie p=1.9                       1.385951         -1.0   0.123770
Gamma GLM                           1.391562         -1.5   0.130433

Model                 Test Deviance   Improvement %   Overfit Gap     P/O RMSE        #Features
Gamma + ElasticNet    1.367991        0.264934        0.090574        0.093394        78
Tweedie p=1.9         1.385951        -1.044466       0.123770        0.141548        78
Gamma GLM             1.391562        -1.453516       0.130433        0.174199        78
```

Two-Part Results

Glum Baseline

Figure 10: Glum Baseline Two-Part Metrics

```
glum_baseline × glum_baseline | F=Poisson GLM, S=Gamma + ElasticNet — pointwise
deviance (train/test): 1.279185 / 1.375032
wMAE     (train/test): 1534.4 / 1597.6
null dev (test):       1.371625
two-part aggregates (Exposure × Freq × Sev):
train: pred=38,851,608.24  obs=38,869,709.33  P/O=1.000
test : pred=9,708,999.34  obs=10,387,183.21  P/O=0.935
```

Manual Binned

Figure 11: Manual Binned Two-Part Metrics

```
manual_bin_glum_tgt × manual_bin_glum_tgt | F=Poisson GLM, S=Gamma + ElasticNet — pointwise
deviance (train/test): 1.277417 / 1.367991
wMAE     (train/test): 1529.1 / 1592.5
null dev (test):       1.371625
two-part aggregates (Exposure × Freq × Sev):
train: pred=38,846,092.57  obs=38,869,709.33  P/O=0.999
test : pred=9,704,108.50  obs=10,387,183.21  P/O=0.934
```

Tweedie Summary Outputs

Glum Baseline

Figure 12: Glum Baseline Tweedie Outputs

```
============================================================
PURE PREMIUM — glum_baseline_pp
============================================================

Tweedie (p=1.9) (p=1.899999976158142)
training loss (deviance): 33.546899
testing  loss (deviance): 33.885491
train totals  observed = 38,869,709.33, predicted = 40,735,162.56
test  totals  observed = 10,387,183.21, predicted = 10,173,501.52

Tweedie (p=1.7) (p=1.7000000476837158)
training loss (deviance): 35.900331
testing  loss (deviance): 37.002309
train totals  observed = 38,869,709.33, predicted = 39,835,183.03
test  totals  observed = 10,387,183.21, predicted = 9,954,784.11

Tweedie (p=1.5) (p=1.5)
training loss (deviance): 73.021887
testing  loss (deviance): 76.633913
train totals  observed = 38,869,709.33, predicted = 39,295,679.83
test  totals  observed = 10,387,183.21, predicted = 9,822,788.68

Tweedie (p=1.3) (p=1.2999999523162842)
training loss (deviance): 189.789315
testing  loss (deviance): 203.048486
train totals  observed = 38,869,709.33, predicted = 38,998,966.52
test  totals  observed = 10,387,183.21, predicted = 9,749,653.84
```

Manual Binned

Figure 13: Manual Binned Tweedie Results

```
============================================================
PURE PREMIUM — manual_bin_glum_tgt_pp
============================================================

Tweedie (p=1.9) (p=1.899999976158142)
training loss (deviance): 33.681231
testing  loss (deviance): 33.973876
train totals  observed = 38,869,709.33, predicted = 39,193,237.41
test  totals  observed = 10,387,183.21, predicted = 9,790,386.10

Tweedie (p=1.7) (p=1.7000000476837158)
training loss (deviance): 36.236011
testing  loss (deviance): 37.242931
train totals  observed = 38,869,709.33, predicted = 39,040,618.34
test  totals  observed = 10,387,183.21, predicted = 9,751,197.77

Tweedie (p=1.5) (p=1.5)
training loss (deviance): 73.871545
testing  loss (deviance): 77.283188
train totals  observed = 38,869,709.33, predicted = 38,944,585.98
test  totals  observed = 10,387,183.21, predicted = 9,726,065.03

Tweedie (p=1.3) (p=1.2999999523162842)
training loss (deviance): 191.964528
testing  loss (deviance): 204.781273
train totals  observed = 38,869,709.33, predicted = 38,892,041.61
test  totals  observed = 10,387,183.21, predicted = 9,711,592.26
```

Coefficient Plots

Glum Baseline

Figure 14: Top 20 Glum Baseline Frequency Features

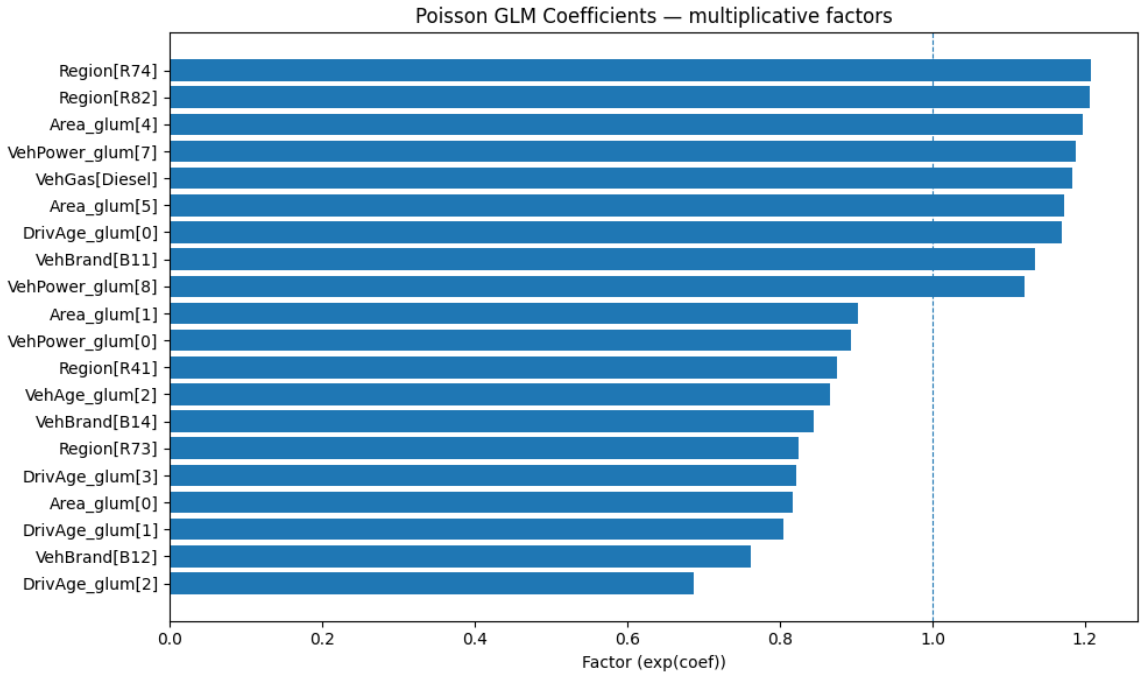
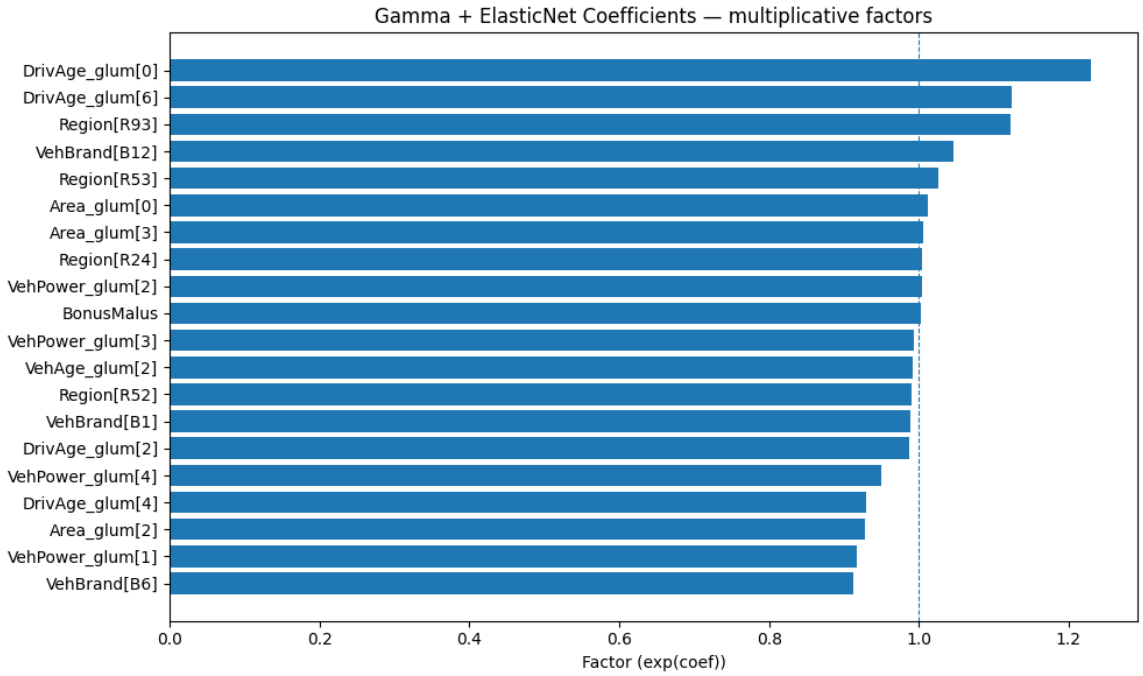


Figure 15: Top 20 Glum Baseline Severity Features



Manual Binned

Figure 16: Top 20 Manual Binned Frequency Features

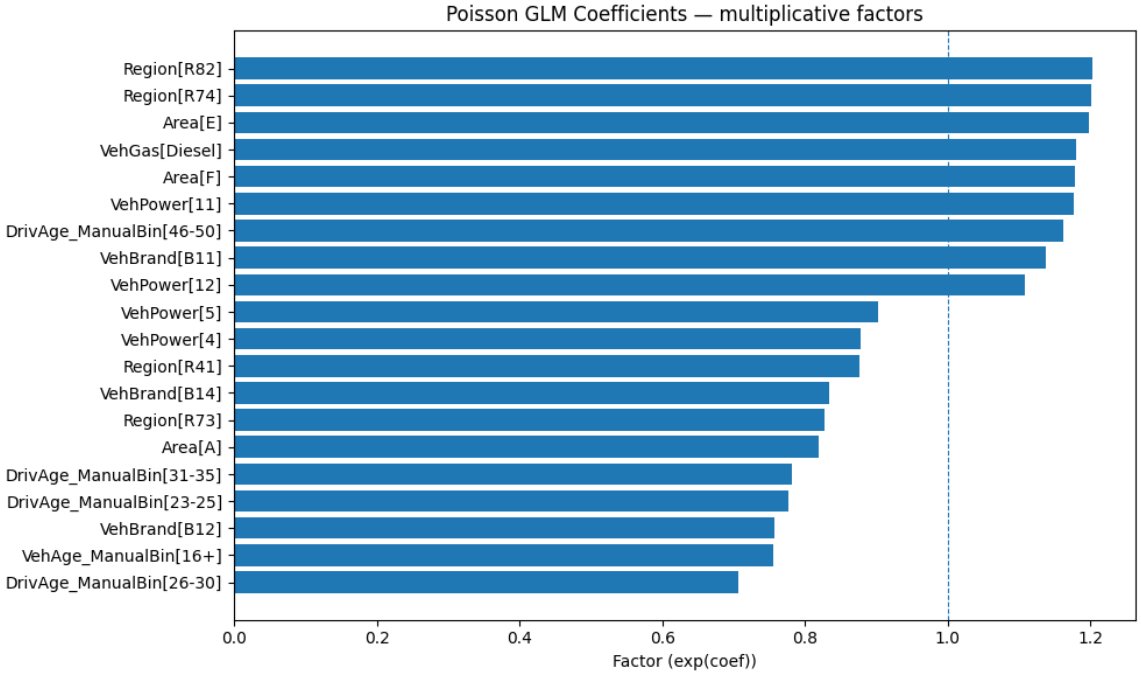
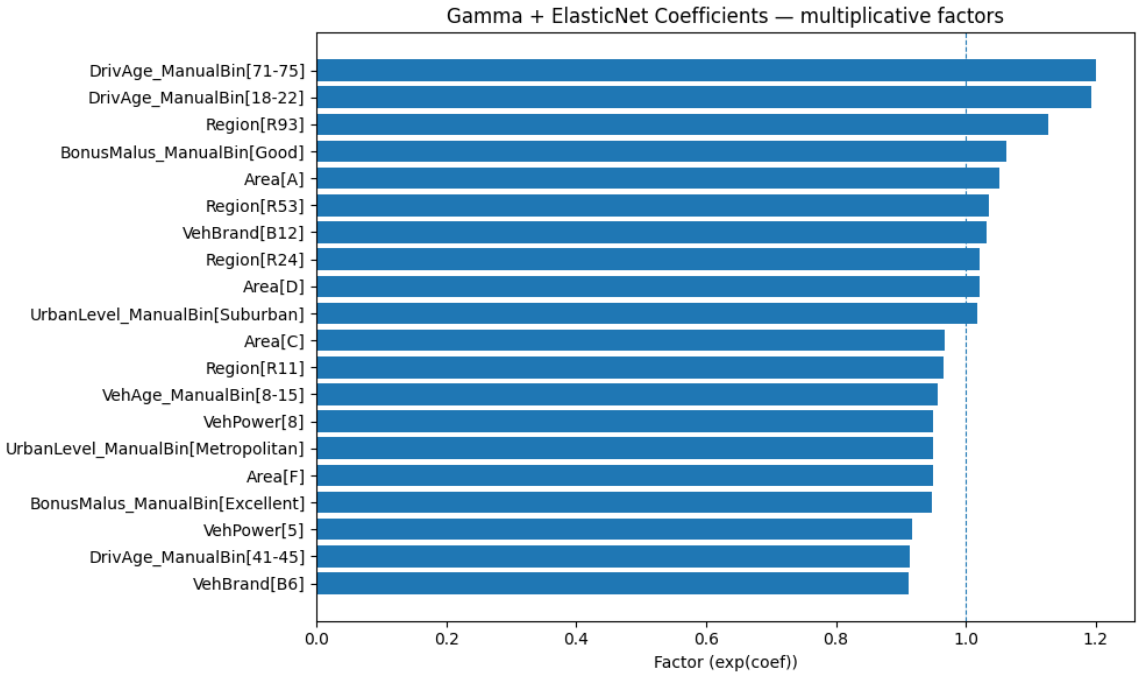


Figure 17: Top 20 Manual Binned Severity Features



Tweedie Frequency Features

Figure 18: Top 20 Tweedie Glum Baseline Features (Winner only)

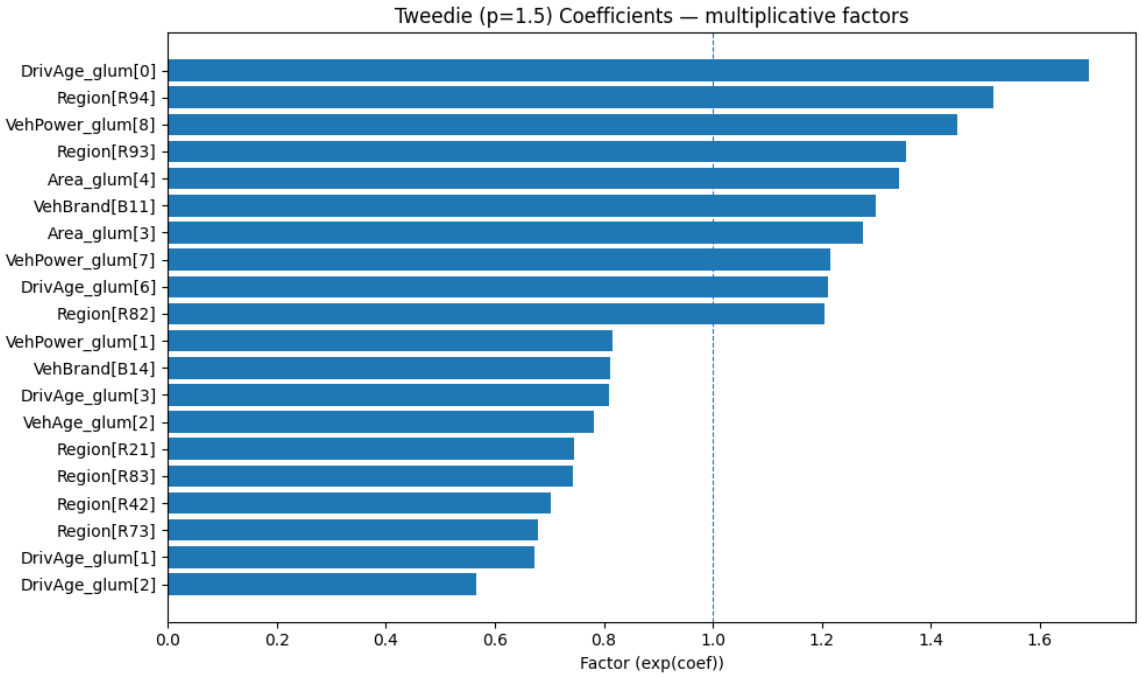
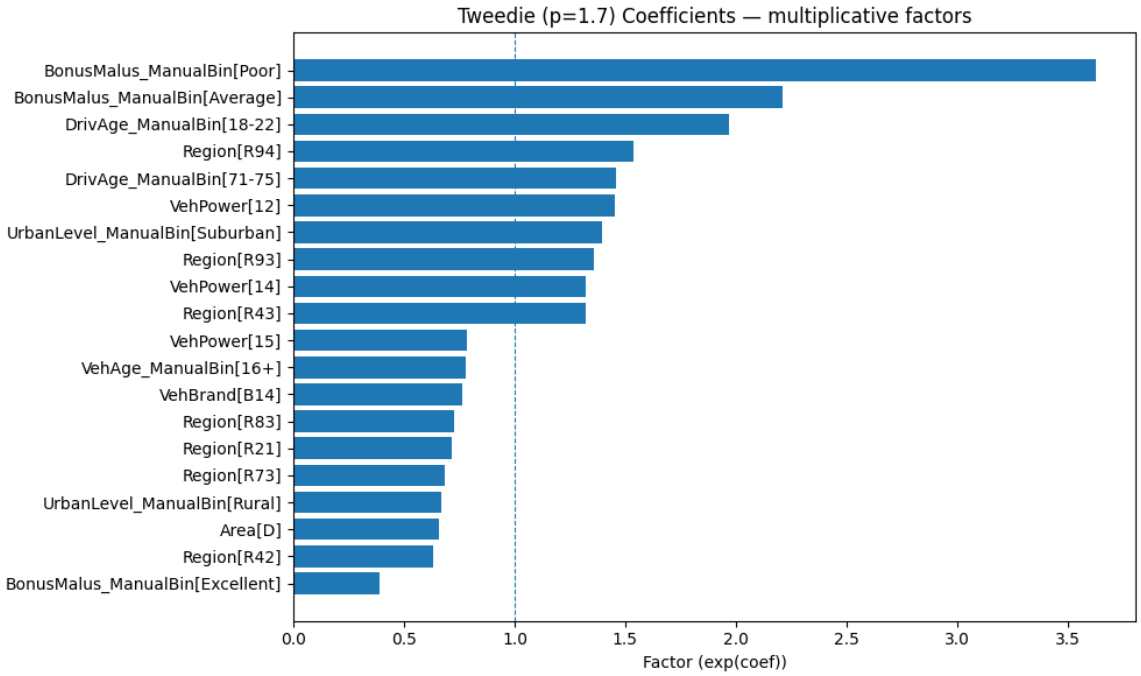


Figure 19: Top 20 Tweedie Manual Binned Features (Winner only)



References

- CAS Monograph. (2025). *Generalized Linear Models for Insurance Rating Second Edition (2025 Revision)*. Casualty Actuarial Society. Retrieved from https://www.casact.org/sites/default/files/2021-01/05-Goldburd-Khare-Tevet.pdf. Last Accessed September 2025.
- Frees, E. W., Derrig, R. A., & Meyers, G. G. (2021). *Predictive Modeling Applications in Actuarial Science: Volume 1*. Cambridge University Press.
- Dutang, C. (2025). *CASdatasets: Insurance Data for Actuarial Science*. Retrieved from https://github.com/dutangc/CASdatasets and last accessed September 2025.
- Dutang, C. (2024). CASdatasets: Insurance datasets. Retrieved from https://cas.uqam.ca/ and last accessed October 2025.
- Quantco. (2024). *glum documentation*. Retrieved from https://glum.readthedocs.io. Last Accessed September 2025.
- Sarpal, K. (2020, November 7). *FREMTPL - French motor TPL insurance claims data*. Kaggle. https://www.kaggle.com/datasets/karansarpal/fremtpl-french-motor-tpl-insurance-claims. Last Accessed September 2025.