Data Mining 1 (2025) 1-10 Submitted 05/12; Published 12/12

Predictive risk modelling in motor vehicle insurance: Comparative study of two-stage and Tweedie models

Onkar Basu ONKAR.BASU@MIF.STUD.VU.LT
Mantas Janušas MANTAS.JANUSAS@MIF.STUD.VU.LT
Katarina Koiro KATARINA.KOIRO@MIF.STUD.VU.LT
Edvin Macel EDVIN.MACEL@MIF.STUD.VU.LT
Data Science study programme
Faculty of Mathematics and Informatics

Advisor: Jurgita Markevičiūtė

Abstract

The paper explores two key approaches for modelling motor insurance claim outcomes: the two-stage frequency-severity framework and the Tweedie approach. Drawing on a real-world dataset, the aim is to compare the predictive performance of these methods in estimating claim frequency and claim severity, which together inform expected loss. By identifying the most effective modelling strategy , this research contributes to the development of robust, data-driven pricing frameworks that support competitive advantage in motor insurance.

Keywords: Motor insurance, frequency-severity framework, Tweedie model, two-stage risk modelling

1. Introduction

According to the Allianz Global Insurance Report (2025), the insurance industry has remained one of the largest business sectors globally in recent decades, with premiums as a percentage of global GDP fluctuating between 7.1% and 7.4% during 2022–2024. Within this domain, the non-life personal property and casualty (P&C) segment represents the most substantial portion of the total insurance market, based on McKinsey & Company's Global Insurance Report (2024). A key component of this segment is motor vehicle insurance, which constitutes a significant share—particularly in developed economies—driven by rising vehicle ownership and the prevalence of mandatory third-party liability coverage. Considering the scale of the motor insurance sector and its growing relevance in data-driven decision-making, this study constructs a predictive risk model of motor insurance claims to inform risk-based pricing, emphasizing claim frequency and financial severity—the cost of a claim. Specifically, the objective of this research is to compare the performance of two-stage modelling approaches with the Tweedie model in predicting motor insurance claim outcomes. By evaluating these frameworks, the study aims to identify the modelling strategy that provides the most accurate and robust predictions for motor insurance data.

2. Literature Review

Given the sector's scale and potential for competitive research insights, predictive modelling in motor insurance is increasingly salient as insurers seek to improve risk assessment, pricing, and portfolio management. Most actuarial studies use the frequency–severity approach, estimating claim frequency (expected number of claims) and claim severity (average cost per claim), with total cost calculated as their product, representing the insurer's expected payout; the corresponding mathematical formulation is deferred to Appendix A. Two principal modelling strategies arise for estimating total loss: two-stage models, which predict frequency and severity separately before combining them to obtain the expected cost, and a single-stage approach, such as the Tweedie model, which jointly models the distribution of total claim amounts without explicitly decomposing frequency and severity components.

2.1 Two-stage models

A substantial body of research has examined the risk assessment of automotive insurance data by developing separate models for claim frequency and claim severity. For instance, Kafková and Křivánková (2014) employed Generalized Linear Models (GLMs) with a Poisson distribution and log-link to predict annual claim frequency, identifying policyholder's age, the age of the insured vehicle, and the region of residence as the most significant predictors. Putra et al. (2021) extended this work by similarly modelling claim frequency with a Poisson distribution, but additionally developing a separate model for claim severity based on a Normal distribution. The selected predictors were largely consistent with the previous, excluding the age variable, while also incorporating additional factors such as occupation type, claim reason, and marital status.

However, GLM-based frequency–severity models have limitations. Primarily, Henckaerts et al. (2021) and Clemente et al. (2023) note that they assume linear relationships between predictors and response variables and independence between claim frequency and severity. The authors employ a Gradient Boosting Machine (GBM) to capture the underlying non-linear relationships more effectively, significantly outperforming GLMs in predicting claim frequency, however, the study shows that GLMs prediction for severity remains competitive.

If GLMs produce comparable results for severity relative to more advanced models, recent studies further question the reliance on Poisson-based count regression for claim frequency due to the high prevalence of zero claims (Alomair, 2024; Kollongei and Onyango, 2024; Zhang et al., 2025). Alomair (2024) proposes Support Vector Machines as a more accurate alternative to count regression models, such as the Poisson and Negative Binomial, while Kollongei and Onyango (2024) employ XGBoost to address zero inflation, overdispersion, and independence assumptions in large motor insurance datasets. Similarly, Zhang et al. (2025) apply multivariate zero-inflated and zero-modified models to handle comparable distributional challenges in insurance data. Taken together, while GLMs remain competitive for modelling severity, these studies indicate that traditional count regression models, as used by Kafková and Křivánková (2014) and Putra et al. (2021), lack predictive advantage for claim frequency, supporting the adoption of methods capable of accommodating the substantial proportion of zero observations.

2.2 Tweedie models

A common approach for insurance prediction is Tweedie modelling and its variants. One-stage Tweedie models are particularly suited for datasets with zeros and positive continuous outcomes, enabling direct prediction of total claim cost per policy in a single GLM without explicitly separating frequency and severity; the corresponding mathematical formulation is deferred to Appendix B. The power parameter determines the Tweedie response distribution, ranging from Normal to Inverse Gaussian. Specifically for insurance claims, $p \in (1, 2)$, yielding the Compound Poisson–Gamma distribution with a point mass at zero and a continuous positive component.

Classical implementations of this framework include Tweedie GLMs (Jørgensen and Paes De Souza, 1994; Smyth and Jørgensen, 2002) and the Tweedie GAM (Wood, 2001), which provide competitive predictive accuracy but often serve primarily as benchmarks. Since the Tweedie approach is extendable, Yang et al. (2016) applying gradient tree boosting to the Tweedie compound Poisson model, capturing complex predictor interactions and relaxing the linearity constraint of the logarithmic mean, resulting in more accurate premium predictions than traditional GLM and GAM models. According to Wilson et al. (2024), Tweedie models remain competitive in automotive insurance, offering interpretable results, though they are limited in handling real-time streaming data and non-linear trends, and are slightly outperformed by machine learning methods such as GBM, ANN, and hybrid GBM–ANN models. On the other hand, Shi et al. (2015) note that the Tweedie compound Poisson–Gamma model can introduce bias by assuming independence between claim frequency and severity, reflecting the conceptual separation of the Poisson and Gamma distributions. They emphasize the inherent interrelation between the number and size of claims, providing a rationale for two-stage modelling the expected loss as the product of frequency and severity. Shi et al. (2015) highlight that, in some contexts, the inherent dependence between claim frequency and claim size may be more pronounced than in others, which aligns with the reasoning put forward by above-mentioned Clemente et al. (2023). In the dataset used in their study, a noticeable correlation between these two factors is observed, which aligns with the empirical results, demonstrating that explicitly accounting for the correlation between frequency and severity yields superior predictive performance compared to approaches that treat these components as independent.

2.3 A perspective

Both approaches to modelling risk in motor insurance are comparable in analytical strength, each presenting distinct advantages and limitations. The Tweedie GLM is straightforward, requiring fewer parameters and offering ease of implementation, however, it may lose precision in contexts where claim frequency and severity exhibit strong dependence. In contrast, two-stage models provide greater interpretability by isolating the drivers of frequency and severity, allowing for more flexible modelling when these components follow different patterns. Nevertheless, aggregating the models can introduce additional complexity. The methodological framework of this study therefore seeks to examine the interdependence between claim likelihood and claim size and to challenge a Tweedie model with various implementations of two-stage models in order to identify the most effective predictive approach.

3. Data

3.1 Data set

This project utilizes the dataset "Dataset of an actual motor vehicle insurance portfolio" by Lledó and Pavía (2024), comprising 105,555 annual insurance policies from a Spanish insurer collected over three years (November 2015 to December 2018). The dataset contains 30 variables covering policyholder demographics, vehicle characteristics and claims history. Prior to public release, the authors of this dataset performed thorough quality control, addressing missing values, duplicates and formatting issues, resulting in a dataset ready for analysis (Lledó and Pavía, 2024). Initial check proved that data contains no duplicates and minimal missingness, with only two vehicle features showing notable gaps: `Length` (10%, 10,329 observations) and `Type_fuel` (2%, 1,764 observations). The missingness seems to be structurally related, i.e. whenever `Type_fuel` is missing, `Length` is also missing, indicating a MAR (missing at random) pattern consistent with specific vehicle types. A full variable description is provided in Appendix C.

3.2 Exploratory data analysis

A binary indicator `missing_Length` was derived to assess the relationship between missingness and other observed variables. Numerical features show that missing records are associated with low-specification vehicles: median `Power` (12 HP vs. 98 HP), `Weight` (149 kg vs. 1,239 kg), and `Value_vehicle` (€3,499 vs. €18,331) are substantially lower for the missing group. Similarly, categorical features confirm a specific profile: 94% fall into `Payment = 0` (half-yearly) and 82% into `Type_risk = 1` (motorbikes). These patterns suggest the missing `Length` cases likely correspond to small, low-specification vehicles such as mopeds or light motorcycles, for which `Type_fuel` may also be difficult to classify. Detailed comparisons are provided in Appendix D.

From the correlation heatmap shown in Figure 1, the physical parameters of the vehicle, i.e. weight, power, cylinder capacity and value exhibit strong mutual correlations. This aligns with real-world expectations. In terms of insurance risk, `N_claims_year` shows notable correlations with `R_Claims_history` and `N_claims_history`, suggesting that past claims frequency is one of the most informative indicators of future risk.

The outlier summary was generated using the $1.5 \times$ IQR rule. Both *N_claims_year* and *Cost_claims_year* contain the highest proportion of outliers, each accounting for approximately 18.61% of all observations, followed by *Weight* (10.95%) and *Cylinder_capacity* (10.15%). To address outliers, a logarithmic transformation was applied to the affected variables. This transformation helped stabilize variance and reduce skewness. The quantitative summary of variables and their outliers with relevant outlier percentages is provided in Table 9 in the Appendix F.

Figure 2 visualizes the relationship between seniority and number of claims. On the x axis we have divided the number of claims into 3 segments, i.e 1,2,2+ claims and on the y axis we represent the seniority of a person, which measures the number of years the customer has been with the insurance company. We observe that people who have filed 0 claims are customers who have been a client of the insurance company for a longer period of time. In contrast, customers who file claims are relatively new clients of the insurance

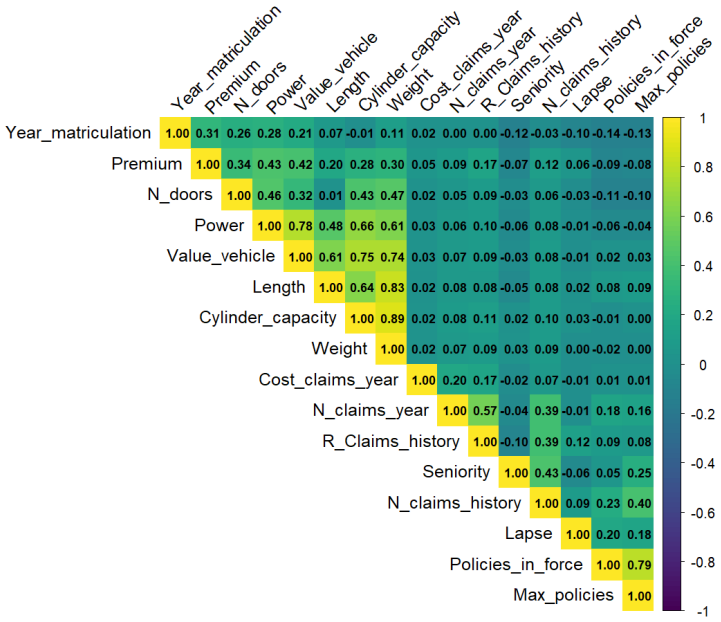

Figure 1: Correlation heatmap of numerical variables.

company. This suggests that customer loyalty (seniority) is a good predictor of lower risk. Figure 3 visualizes the proportion of number of claims by the policies in force. The x-axis shows different numbers of policies in force, ranging from 1 to 17, while the y-axis represents the proportion of claims. It can be observed that lower numbers of policies in force (1–4), the proportion of 0 claims is relatively higher compared to other policy counts. In the range of 5-8 policies, the share of 1 claim and 2+ claims grows significantly. This might imply that as the number of policies increases, the likelihood of at least 1 claim rises, possibly due to greater exposure or higher total insured value.

3.3 Feature engineering

For frequency of claims estimation, exposure (the time at risk used to scale the observed number of claims) was first approximated as the number of days between `Date_last_renewal` and `Date_next_renewal`, resulting in a distribution tightly centered around one year with only minor deviations. However, several policies experienced mid-year lapses, reducing their (actual) exposure to an average of 0.26 years compared to one full year for active policies, implying that treating all contracts as fully exposed would underestimate their claim frequency by roughly a factor of 3.8. To correct this, actual exposure was defined as the time from `Date_last_renewal` to either `Date_lapse` (for lapsed policies) or `Date_next_renewal` (for active policies), with a minimum threshold of 0.1 years (approximately 36 days) to avoid extreme annualized rates. An annualized measure, `claim_rate_annual`, was computed as `N_claims_year` divided by this exposure. Additionally, a missingness indicator was created for `Length` (also `Type_fuel`, though the latter was subsequently removed due to near-zero variance). The variable `Length`, which exhibited structured missingness, was imputed us-

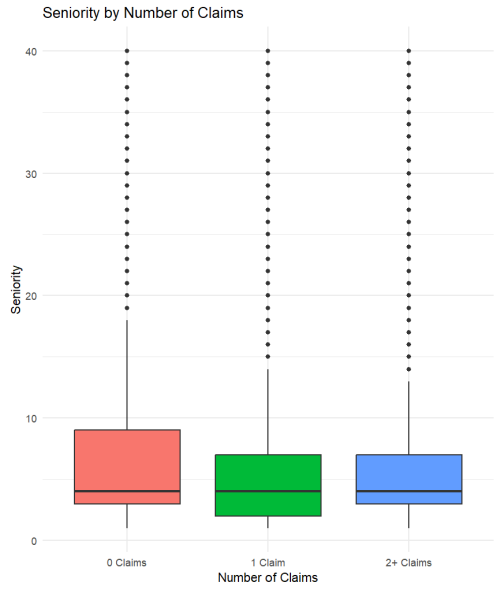

Figure 2: Relationship of seniority with number of claims

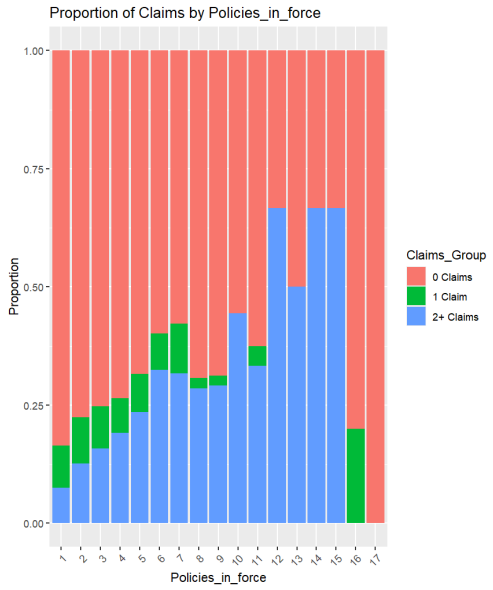

Figure 3: Relationship of policies in force with number of claims

ing the `missRanger` algorithm based on a set of correlated technical attributes, ensuring consistency between observed and imputed distributions. Categorical fuel type missingness was encoded explicitly by assigning the level "Unknown."

To address extreme skewness in claims cost and vehicle characteristics, logarithmic transformations were applied to `Cost_claims_year`, `Weight`, `Cylinder_capacity`, `Power`, and `N_claims_history` using the $\log(x+1)$ transformation to handle zero values. For severity modelling, `Cost_claims_year` was back-transformed from its log scale and `Claim_Severity` was defined as total cost divided by the number of claims for policies with at least one claim. Other engineered features include age-related features (`Driver_Age`, `Driving_Experience`), vehicle-related features (`Vehicle_Age`, `Power_to_Weight_Ratio`), and contract-related features (`Product_Holding_Ratio`). A unique row identifier (`unique_row_id`) was created to enable correct merging of predictions across different modeling components, as policyholder IDs (`ID`) are not unique across policy-years. A complete list of engineered features, along with their descriptions and calculations, is provided in Table 8 in the Appendix.

4. Modelling

The dataset was split 75/25 into training (79,167 rows) and test (26,388 rows) sets using stratified sampling on a zero-indicator for `N_claims_year` to preserve the 81% zero-claim proportion in both sets.

4.1 Frequency modelling of claim counts

Three models were evaluated for annual claim frequency prediction: Poisson GLM, Zero-Inflated Poisson (ZIP), and Random Forest (RF). Both Poisson and ZIP employed exposure as an offset term to model claim rates. The ZIP model used only statistically significant predictors ($p < 0.05$) from the Poisson model in both the count and zero-inflation components. The Random Forest was trained on annualized rates (`claim_rate_annual` = `N_claims_year` / `exposure_years_actual`) with 500 trees, $m_{try} = \sqrt{p}$ and minimum node size of 10. Full specifications and rationale are provided in Appendix G. Model performance is summarized in Table 1.

Table 1: Frequency model comparison

Model	RMSE	MAE	AIC	R^2 (OOB)
Poisson GLM	1.20	0.41	88,451	-
ZIP (sig. pred.)	1.20	0.44	89,604	-
Random Forest	1.06	0.34	-	50.35%

Random Forest achieved the best predictive accuracy (MAE = 0.34 claims/year, so 18% improvement over Poisson), explaining 50% of variance in out-of-bag samples. Between parametric models, Poisson is preferred over ZIP (ΔAIC = 1,153), as ZIP offered no improvement despite additional complexity. Figure 4 shows predictions clustering near the diagonal, while Figure 5 confirms unbiased residuals for typical policies.

The model systematically underpredicts extreme high-frequency policies (rate >10, N=42, 0.16% of portfolio), with mean prediction 5.2 versus actual 17.1. Performance across claim frequency buckets is detailed in Table 10 (Appendix I). A two-stage approach with specialized RF for high-rate policies improved the tail (MAE: 12.1→10.3), however, degraded overall performance (MAE: 0.34→0.35). Given marginal gains on a negligible segment, the single-stage RF was selected as the final model for frequency of claims prediction (more details are in Appendix H).

4.2 Claim severity modelling

Severity modeling focused on predicting average cost per claim for policies with at least one claim (N=4,900 training observations). Four approaches were compared: Linear Regression on log-transformed severity, Random Forest, XGBoost with Tweedie objective and GAMLSS (Generalized Additive Models for Location, Scale and Shape).

Table 2: Claim severity model comparison

Model	**MAE (€)**	**RMSE (€)**	**MAPE**
Linear Regression	410.68	1792.29	148.03%
Random Forest	418.75	1791.66	169.70%
XGBoost (Tweedie)	457.00	1761.39	323.85%
GAMLSS	410.61	1791.96	146.89%

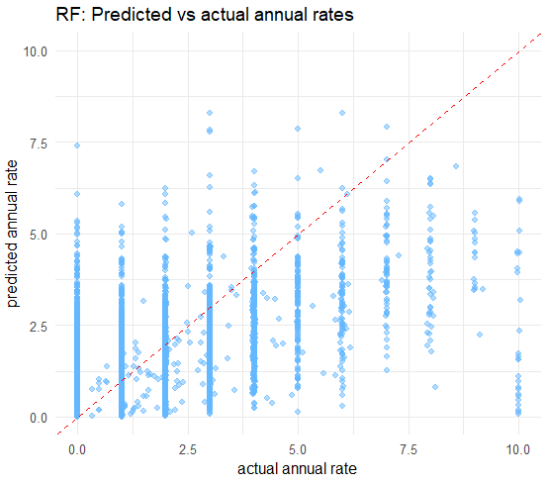

Figure 4: RF: Predicted vs actual rates

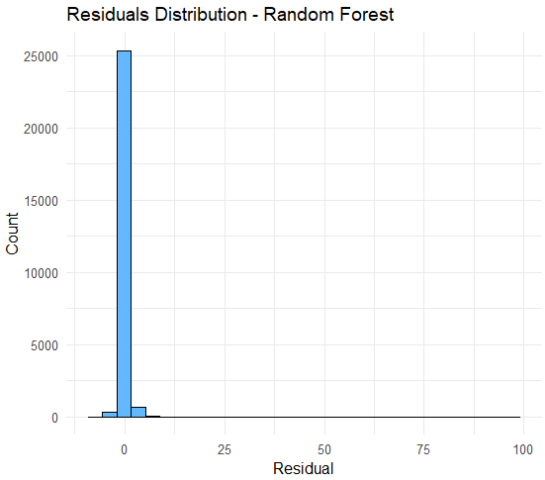

Figure 5: RF: Residual distribution

GAMLSS provides the best overall performance, achieving the lowest MAE (€410.61) and MAPE (146.89%). GAMLSS explicitly models the distribution of claim severity, capturing skewness and heavy tails typical in insurance data. Unlike Linear Regression, Random Forest, or XGBoost (which primarily minimize mean-squared error without accounting for distribution shape) GAMLSS fits a flexible, heavy-tailed distribution (Inverse Gaussian), leading to better prediction of both small and large claims. XGBoost with Tweedie objective performed poorly despite tuning. A two-stage approach with separate models for claims below/above €1,000 showed marginal changes (MAE +€2, RMSE -€1) but no meaningful improvement. The lack of significant gains is attributable to insufficient data for claims exceeding €1,000. XGBoost typically requires several thousand observations to perform efficiently.

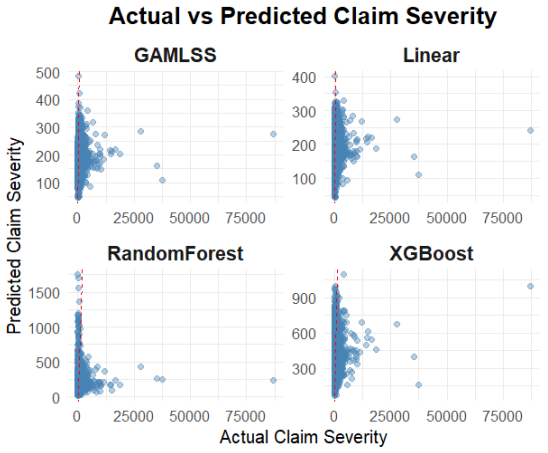

Figure 6: Actual vs predicted severity across models

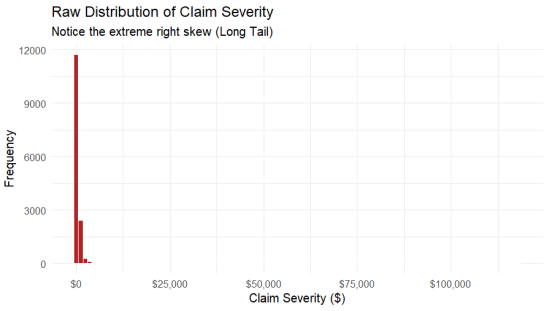

Figure 7: Distribution of claim severity

Figure 6 exhibits a distinct clustering pattern, with predictions heavily concentrated in the €0-500 range across all models. This phenomenon is directly attributable to the extreme right-skew of the training data (Figure 7): the vast majority of claim severity observations fall within €0-500, while large claims (> €5,000) are sparse. Consequently, models—which optimize for overall error minimization—prioritize fitting the dense cluster of small claims. Lacking sufficient examples of high-severity events to learn robust patterns, models struggle to generalize to the tail, resulting in systematic underprediction of large losses. Essentially, the models have become highly specialized at predicting minor accidents but remain "blind" to catastrophic risk due to data scarcity in the upper tail.

4.3 Tweedie model for expected loss

As an alternative to the two-stage modelling framework, the expected annual loss per policy was modelled directly using a compound Poisson–Gamma Tweedie GLM with log-transformed response (`Cost_claims_year_log`) to handle extreme skewness. Variables were excluded based on multicollinearity (VIF diagnostics) and relevance to loss prediction. Notably, actual exposure (`exposure_years_actual`) was included as a standard predictor rather than as an offset term, allowing flexible exposure-cost relationships. The exposure variable proved highly significant (coefficient = 1.574, $p < 0.001$).

The final specification, obtained through iterative variable selection minimizing AIC and out-of-sample error, retained predictors significant at the 1% level related to policyholder characteristics, vehicle specifications, policy details and claims history. Full model coefficients and significance levels are provided in Table 12 (Appendix J).

Table 3: Tweedie model results

Metric	**Value**
Power parameter (p)	1.0454
Dispersion (ϕ)	5.6107
Exposure coef.***	1.574
MAE (log scale)	1.12
RMSE (log scale)	2.37
Spearman ρ	0.5545
Pred./Actual ratio	0.975

*** $p < 0.001$

As per Table 3, estimated power parameter $p = 1.0454$ lies within the compound Poisson range ($1 < p < 2$), confirming the suitability of the Tweedie family for these data. The model achieved moderate rank correlation ($\rho = 0.5545$), indicating reasonable policy-level risk ordering despite substantial individual variability. Most importantly, the prediction-to-actual ratio of 0.975 demonstrates great aggregate calibration with only 2.5% underestimation at the portfolio level.

4.4 Model comparison and final selection

The two-stage approach (frequency × severity) and single-stage Tweedie model were compared on portfolio-level loss prediction. Predictions were merged via `unique_row_id` to account for duplicate policyholder IDs across policy-years. For the two-stage model, expected loss per policy was calculated as:

$$\hat{L}_i = \hat{N}_i \times \hat{S}_i$$

where $\hat{N}_i$ is the predicted claim count (RF frequency × exposure) and $\hat{S}_i$ is the predicted average severity. Policies with zero predicted claims were assigned zero loss.

Table 4: Portfolio-level loss comparison

Approach	**Total Predicted**	**Total Actual**	**Ratio**
Two-Stage (Freq × Sev)	1,313,209	4,144,883	0.317
Tweedie (single-stage)	27,544*	28,264*	0.975

*Sum of log-predictions; not directly comparable in euro terms

The accuracy metrics for two-stage model were calulated: MSLE is 2.4974, MAE is 724.61, and RMSE is 4,183. Even though the MAE and RMSE values are quite high, suggesting that predictions may not be very precise, these results are common in the insurance market, since insurance companies often work with variables of varying magnitudes. Overall, the Tweedie model clearly outperforms the Two-Stage model, when comparing RMSE and MAE results with Table 3. The ratio also shows that the Tweedie model produces smaller loss. Complementary plots visualizing the prediction metrics are presented in Appendix K.

5. Conclusions

This project developed predictive models for motor insurance claims using 26,388 policies, comparing two-stage (frequency × severity) and single-stage (Tweedie GLM) approaches.

The Random Forest frequency model achieved strong performance (MAE = 0.34 claims/year, R^2 = 50%), with accurate predictions for typical policies and systematic underprediction only in extreme cases (0.16% of portfolio). For severity, GAMLSS performed best (MAE = €410.61) but all models exhibited conservative predictions for high-cost claims due to data sparsity in the tail. On the other hand, Tweedie GLM achieved near-perfect aggregate calibration (ratio = 0.975) by directly modeling log-transformed loss, effectively handling extreme skewness without explicit decomposition. In contrast, the two-stage framework produced conservative estimates (ratio ≈ 0.3-0.4), primarily due to severity model limitations in predicting rare, high-cost events.

Key methodological contributions include proper exposure adjustment for mid-year lapses (preventing 3.8× frequency underestimation), structured missing data analysis, and empirical comparison of modeling frameworks. Future work should prioritize enhanced tail modeling, especially for severity modelling.

References

Allianz Research. Allianz Global Insurance Report 2025: Rising demand for protection. Technical report, Allianz Group, May 2025. URL https://www.allianz.com/content/dam/onemarketing/azcom/Allianz_com/economic-research/publications/specials/en/2025/may/2025-05-27-global-insurance-report.pdf.

Gadir Alomair. Predictive performance of count regression models versus machine learning techniques: A comparative analysis using an automobile insurance claims frequency dataset. *PLOS ONE*, 19(12):e0314975, December 2024. ISSN 1932-6203. doi: 10.1371/journal.pone.0314975. URL https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0314975. Publisher: Public Library of Science.

Carina Clemente, Gracinda R. Guerreiro, and Jorge M. Bravo. Modelling Motor Insurance Claim Frequency and Severity Using Gradient Boosting. *Risks*, 11(9):163, September 2023. ISSN 2227-9091. doi: 10.3390/risks11090163. URL https://www.mdpi.com/2227-9091/11/9/163. Publisher: Multidisciplinary Digital Publishing Institute.

Roel Henckaerts, Marie-Pier Côté, Katrien Antonio, and Roel Verbelen. Boosting Insights in Insurance Tariff Plans with Tree-Based Machine Learning Methods. *North American Actuarial Journal*, 25(2):255–285, April 2021. ISSN 1092-0277. doi: 10.1080/10920277.2020.1745656. URL https://doi.org/10.1080/10920277.2020.1745656. Publisher: Routledge _eprint: https://doi.org/10.1080/10920277.2020.1745656.

Bent Jørgensen and Marta C. Paes De Souza. Fitting Tweedie's compound poisson model to insurance claims data. *Scandinavian Actuarial Journal*, 1994(1):69–93, January 1994. ISSN 0346-1238. doi: 10.1080/03461238.1994.10413930. URL https://doi.org/10.1080/03461238.1994.10413930. Publisher: Taylor & Francis _eprint: https://doi.org/10.1080/03461238.1994.10413930.

Silvie Kafková and Lenka Křivánková. Generalized Linear Models in Vehicle Insurance. *Acta Universitatis Agriculturae et Silviculturae Mendelianae Brunensis*, 62(2):383–388, February 2014. ISSN 12118516, 24648310. doi: 10.11118/actaun201462020383. URL https://doi.org/10.11118/actaun201462020383. Publisher: Acta Universitatis Agriculturae et Silviculturae Mendelianae Brunensis.

Naomi Kollongei and Fredrick Onyango. Motor Insurance Claim Frequency Prediction Using XGBoost. *Asian Journal of Probability and Statistics*, 26(10):155–170, October 2024. ISSN 2582-0230. doi: 10.9734/ajpas/2024/v26i10665. URL https://journalajpas.com/index.php/AJPAS/article/view/665.

Josep Lledó and Jose M. Pavía. Dataset of an actual motor vehicle insurance portfolio. *European Actuarial Journal*, 2, July 2024. doi: 10.17632/5cxyb5fp4f.2. URL https://data.mendeley.com/datasets/5cxyb5fp4f/2. Publisher: Mendeley Data.

McKinsey & Company. Global Insurance Report 2025: The pursuit of growth. Technical report, McKinsey & Company, November 2024. URL https://www.mckinsey.com/industries/financial-services/our-insights/global-insurance-report#/.

Tri Andika Julia Putra, Donny Citra Lesmana, and I. Gusti Putu Purnaba. Prediction of Future Insurance Premiums When the Model is Uncertain. *Atlantis Press*, pages 128–135, May 2021. doi: 10.2991/assehr.k.210508.054. URL https://www.atlantis-press.com/proceedings/icmmed-20/125956477. ISSN: 2352-5398.

Peng Shi, Xiaoping Feng, and Anastasia Ivantsova. Dependent frequency–severity modeling of insurance claims. *Insurance: Mathematics and Economics*, 64:417–428, September 2015. ISSN 0167-6687. doi: 10.1016/j.insmatheco.2015.07.006. URL https://www.sciencedirect.com/science/article/pii/S0167668715001183.

Gordon K. Smyth and Bent Jørgensen. Fitting Tweedie's Compound Poisson Model to Insurance Claims Data: Dispersion Modelling. *ASTIN Bulletin: The Journal of the IAA*, 32(1):143–157, May 2002. ISSN 0515-0361, 1783-1350. doi: 10.2143/AST.32.1.1020. URL https://www.cambridge.org/core/journals/astin-bulletin-journal-of-the-iaa/article/fitting-tweedies-compound-poisson-model-to-insurance-claims-data-dispersion-modelling/DEF0B49F96FC015C7FBE076BC0A5C3AC.

Alinta Ann Wilson, Antonio Nehme, Alisha Dhyani, and Khaled Mahbub. A Comparison of Generalised Linear Modelling with Machine Learning Approaches for Predicting Loss Cost in Motor Insurance. *Risks*, 12(4):62, April 2024. ISSN 2227-9091. doi: 10.3390/risks12040062. URL https://www.mdpi.com/2227-9091/12/4/62. Publisher: Multidisciplinary Digital Publishing Institute.

Simon N Wood. mgcv: GAMs and Generalized Ridge Regression for R. *R news*, 1, 2001.

Yi Yang, Wei Qian, and Hui Zou. Insurance Premium Prediction via Gradient Tree-Boosted Tweedie Compound Poisson Models. *Journal of Business & Economic Statistics*, April 2016. doi: 10.48550/arXiv.1508.06378. URL http://arxiv.org/abs/1508.06378. arXiv:1508.06378 [stat].

Pengcheng Zhang, David Pitt, and Xueyuan Wu. A comparative analysis of several multivariate zero-inflated and zero-modified models with applications in insurance. *Communications in Statistics - Theory and Methods*, 54(7):2130–2157, April 2025. ISSN 0361-0926. doi: 10.1080/03610926.2024.2360079. URL https://doi.org/10.1080/03610926.2024.2360079. Publisher: Taylor & Francis _eprint: https://doi.org/10.1080/03610926.2024.2360079.

Appendix A. Frequency-Severity Method

The frequency–severity method is an actuarial approach used to estimate claim frequency (expected number of claims) and claim severity (average cost per claim), with total cost calculated as their product, representing the insurer's expected payout:

$$\mathbb{E}[L] = \widehat{f}_{\text{pred}} \times \widehat{s}_{\text{pred}} \tag{1}$$

Definition of terms.
$\mathbb{E}[L]$: total expected cost.
$\widehat{f}_{\text{pred}}$: predicted frequency.
$\widehat{s}_{\text{pred}}$: predicted severity.

Appendix B. Tweedie Model Method

A one-stage Tweedie model uses a compound Poisson–Gamma distribution to simultaneously predict the total claim cost per policy in a single GLM without explicitly separating frequency and severity:

$$Y_i \sim \text{Tweedie}_p(\mu_i, \phi), \quad \mathbb{E}[Y_i] = \mu_i, \quad \text{Var}(Y_i) = \phi\, \mu_i^p \tag{2}$$

Definition of terms.
Y_i: total claim amount for policyholder.
μ_i: the expected value of Y_i.
ϕ: dispersion parameter.
p: power parameter. Determines the variance–mean relationship and defines the distribution family. For insurance claims, $1 < p < 2$ corresponds to a compound Poisson–Gamma model.

Appendix C. Description of Variables

Table 5: Description of variables for motor vehicle insurance data source (Lledó and Pavía, 2024)

Variable Name	Description
ID	Internal identification number assigned to each annual contract. Each policyholder may appear in multiple rows, representing different annuities.
Date_start_contract	Start date of the policyholder's contract (DD/MM/YYYY).
Date_last_renewal	Date of last contract renewal (DD/MM/YYYY).
Date_next_renewal	Date of the next contract renewal (DD/MM/YYYY).
Distribution_channel	Channel through which the policy was contracted: 0 for Agent, 1 for Insurance broker.
Date_birth	Date of birth of the insured (DD/MM/YYYY).
Date_driving_licence	Date of issuance of the insured's driver's license (DD/MM/YYYY).
Seniority	Number of years the insured has been associated with the insurance entity.
Policies_in_force	Total number of policies held by the insured during the reference period.
Max_policies	Maximum number of policies ever held by the insured.
Max_products	Maximum number of products simultaneously held by the insured.
Lapse	Number of policies cancelled or terminated for nonpayment, excluding those replaced by another policy.
Date_lapse	Date of contract termination (DD/MM/YYYY).
Payment	Last payment method: 1 for half-yearly, 0 for annual.
Premium	Net premium amount for the current year.
Cost_claims_year	Total cost of claims during the current year.
N_claims_year	Number of claims incurred during the current year.
N_claims_history	Total number of claims filed over the policy's lifetime.
R_Claims_history	Ratio of claims to policy duration in years.
Type_risk	Type of risk: 1 for motorbikes, 2 for vans, 3 for passenger cars, 4 for agricultural vehicles.
Area	Area classification: 0 for rural, 1 for urban (over 30,000 inhabitants).
Second_driver	1 if multiple regular drivers declared, 0 otherwise.
Year_matriculation	Year of vehicle registration (YYYY).
Power	Vehicle power in horsepower.
Cylinder_capacity	Cylinder capacity of the vehicle.
Value_vehicle	Market value of the vehicle on 31/12/2019.
N_doors	Number of vehicle doors.
Type_fuel	Type of fuel: Petrol (P) or Diesel (D).
Length	Vehicle length in meters.
Weight	Vehicle weight in kilograms.

Appendix D. Summary of missingness

Table 6: Numerical feature comparisons by missing Length indicator

Variable	Test Used	p-value	Median (Missing=0)	Median (Missing=1)
Year_matriculation	Wilcoxon	8.20e-186	2005	2004
Value_vehicle	Wilcoxon	0.0	18330.87	3499
Power	Wilcoxon	0.0	98	12
Weight	Wilcoxon	0.0	1239	149

Note: All reported p-values are substantially below the 0.001 threshold due to the large sample size. Practical differences should be interpreted using the median values reported.

Table 7: Categorical feature distribution by missing Length indicator

Variable	Value	Missing=0 (n)	%	Missing=1 (n)	%	Test Used	p-value
Area	0	69502	73.0	7142	69.1	Chi-squared	1.01e-16
Area	1	25724	27.0	3187	30.9	Chi-squared	1.01e-16
Distribution_channel	0	51546	54.1	6371	61.7	Chi-squared	1.64e-48
Distribution_channel	1	43680	45.9	3958	38.3	Chi-squared	1.64e-48
Payment	0	62146	65.3	9718	94.1	Chi-squared	0.0
Payment	1	33080	34.7	611	5.9	Chi-squared	0.0
Second_driver	0	82250	86.4	10247	99.2	Chi-squared	1.60e-309
Second_driver	1	12976	13.6	82	0.8	Chi-squared	1.60e-309
Type_risk	1	2	0.0	8500	82.3	Chi-squared	0.0
Type_risk	2	12954	13.6	258	2.5	Chi-squared	0.0
Type_risk	3	82247	86.4	743	7.2	Chi-squared	0.0
Type_risk	4	23	0.0	828	8.0	Chi-squared	0.0

Note: All reported p-values are substantially below the 0.001 threshold, reflecting statistical significance primarily due to the large sample size. Practical differences should be interpreted using the proportions and counts reported.

Appendix E. Engineered Features

Table 8: Summary of engineered features with descriptions and calculations

Feature Name	**Description**	**Calculation / Source**
Driver_Age	Age of policyholder at contract start	Floor(Date_start_contract - Date_birth) in years
Driving_Experience	Years of driving experience	Floor(Date_start_contract - Date_driving_licence) in years
Vehicle_Age	Age of vehicle at contract start	Year(Date_start_contract) - Year_matriculation
Power_to_Weight_Ratio	Relative vehicle performance	Power / Weight (0 if Weight = 0)
Contract_Duration	Length of current policy period	Date_next_renewal - Date_start_contract (days)
Product_Holding_Ratio	Ratio of products to policies	Max_products / Max_policies (0 if Max_policies = 0)
Has_Claimed_In_History	Indicates past claims	1 if N_claims_history > 0, else 0
exposure_years_actual	Actual policy exposure with lapse adjustment	Days active / 365.25, minimum 0.1 years
claim_rate_annual	Annualized claim frequency	N_claims_year / exposure_years_actual
missing_Length	Missing vehicle length indicator	1 if Length is missing, else 0
unique_row_id	Unique policy-year identifier	Sequential row number
Cost_claims_year_log	Log-transformed annual claim cost	$\log(1 + \text{Cost\_claims\_year})$
Weight_log	Log-transformed vehicle weight	$\log(1 + \text{Weight})$
Cylinder_capacity_log	Log-transformed cylinder capacity	$\log(1 + \text{Cylinder\_capacity})$
Power_log	Log-transformed vehicle power	$\log(1 + \text{Power})$
N_claims_history_log	Log-transformed historical claims	$\log(1 + \text{N\_claims\_history})$
Claim_Severity	Average cost per claim	Cost (in euros) / N_claims_year

Appendix F. Outlier Analysis

Table 9: Summary of the number of outliers present in each variable of our dataset

Feature	Number of Outliers	% of All Observations
Seniority	7,603	7.20
Policies_in_force	3,882	3.68
Max_policies	7,898	7.48
Premium	7,381	6.99
Cost_claims_year	19,646	18.61
N_claims_year	19,646	18.61
N_claims_history	5,013	4.75
R_Claims_history	8,541	8.09
Year_matriculation	2,906	2.75
Power	9,691	9.18
Cylinder_capacity	10,748	10.18
Value_vehicle	3,344	3.17
Length	2,480	2.35
Weight	11,558	10.95

Appendix G. Random Forest specifications

The Random Forest model for claim frequency was configured with the following hyperparameters:

- **Number of trees**: 500. This ensures stable predictions with diminishing marginal returns beyond this threshold.
- **Variables per split (m_{try})**: $\sqrt{p}$, where p is the number of predictors. This helps balancing decorrelation between trees and predictive power.
- **Minimum node size**: 10 observations. This prevents overfitting while maintaining sufficient data for terminal node predictions.
- **Sampling**: Bootstrap sampling with replacement (standard bagging procedure).
- **Response variable**: Annualized claim rate accounting for actual exposure, allowing the model to capture non-linear relationships without explicit exposure modeling.

This specification balances model complexity with computational efficiency while enabling automatic detection of interactions and non-linearities without manual feature engineering.

Appendix H. Two-Stage Random Forest approach

To address systematic underprediction of high-frequency policies, a two-stage approach was explored:

Stage 1: Initial RF predicts rates for all policies (parameters as in Appendix G).
Stage 2: For policies with predicted rate ≥ 3 claims/year, predictions are replaced by an auxiliary RF trained exclusively on high-frequency training observations (rate ≥ 3). The auxiliary model uses 500 trees and $m_{try} = \sqrt{p}$ but reduces minimum node size to 5 to capture patterns in the limited subsample.
Results: The approach improved predictions for extreme cases (rate > 10) from MAE 12.1 to 10.3, but overall MAE increased from 0.34 to 0.35 due to higher variance in the 3-10 claims/year segment. As this segment represents only 42 policies (0.16% of portfolio) and likely contains data anomalies, the added complexity was deemed unjustified.

Appendix I. Random Forest Performance by Claim Frequency Bucket

Table 10 presents Random Forest performance stratified by actual claim frequency. The model performs well for typical policies (0-5 claims/year, 99.8% of portfolio) but systematically under-predicts extreme high-frequency cases.

Table 10: RF performance by claim frequency bucket (initial single-stage model)

Bucket	**N**	**Mean Actual**	**Mean Predicted**	**RMSE**	**MAE**
0	21,493	0.00	0.42	0.53	0.42
0-1	2,264	0.55	0.66	0.43	0.35
1-2	2,153	1.42	1.36	0.60	0.46
2-5	409	2.84	2.47	1.28	1.01
5-10	27	6.30	3.96	3.71	2.74
10+	42	17.08	5.23	17.54	12.14
Overall	**26,388**	**0.43**	**0.52**	**1.06**	**0.34**

The 10+ bucket contains only 42 policies (0.16% of portfolio) with annual rates up to 100 claims/year, suggesting potential data errors or misclassified fleet/commercial policies. While these cases substantially inflate bucket-level MAE (12.14), their impact on overall portfolio metrics remains minimal due to their rarity.
Table 11 compares the initial single-stage RF with the two-stage approach for high-frequency policies.

Table 11: Comparison of single-stage vs two-stage RF by bucket

	Single-Stage RF			**Two-Stage RF**		
Bucket	**N**	**Mean Pred**	**MAE**	**N**	**Mean Pred**	**MAE**
0	21,493	0.42	0.42	21,477	0.18	0.18
0-1	2,264	0.66	0.35	319	0.64	0.55
1-2	2,153	1.36	0.46	2,202	1.09	0.59
2-5	409	2.47	1.01	2,006	2.10	1.21
5-10	27	3.96	2.74	342	4.49	2.64
10+	42	5.23	12.14	42	7.24	10.34
Overall	**26,388**	**0.52**	**0.34**	**26,388**	**0.53**	**0.35**

The two-stage approach improves extreme tail predictions (10+: MAE 12.14→10.34, mean prediction 5.23→7.24) but increases error in the 2-5 range (MAE 1.01→1.21), resulting in marginally worse overall performance (0.34→0.35). This trade-off, combined with the negligible portfolio weight of extreme cases (0.16%), led to selection of the simpler single-stage model.

Appendix J. Tweedie Model Specification

Table 12 presents the full coefficient estimates for the final Tweedie GLM.

Table 12: Tweedie model coefficient estimates

Variable	Estimate	Std. Error	t-value	p-value	
(Intercept)	-20.66	54.76	-0.377	0.7060	
Seniority	-0.0399	0.00143	-27.862	<2e-16	***
Policies_in_force	0.0569	0.00902	6.307	2.86e-10	***
Max_policies	-0.0763	0.00991	-7.703	1.34e-14	***
R_Claims_history	0.2027	0.00225	90.164	<2e-16	***
Type_risk	-0.0559	0.01437	-3.892	9.95e-05	***
Area	-0.0179	0.01185	-1.508	0.1315	
Second_driver	0.1102	0.01547	7.122	1.08e-12	***
Value_vehicle	3.13e-06	7.55e-07	4.141	3.46e-05	***
missing_Length	-0.3906	0.04032	-9.688	<2e-16	***
Driver_Age	0.00320	0.000882	3.626	0.0003	***
Driving_Experience	-0.00696	0.000920	-7.572	3.72e-14	***
Vehicle_Age	-0.00696	0.000989	-7.036	2.00e-12	***
Contract_Duration	-3.46e-04	7.28e-06	-47.503	<2e-16	***
Product_Holding_Ratio	0.6440	0.03073	20.956	<2e-16	***
Has_Claimed_In_History	18.99	54.76	0.347	0.7287	
exposure_years_actual	1.574	0.02030	77.513	<2e-16	***
Power_log	0.02975	0.01565	1.901	0.0573	.
N_claims_history_log	0.6030	0.01020	59.125	<2e-16	***

Signif. codes: *** 0.001, ** 0.01, * 0.05, . 0.1
Dispersion: 5.6107, Index parameter: 1.0454
Residual deviance: 151,043 on 79,148 df, AIC: 103,140

The model was estimated on 79,167 training observations. Exposure (`exposure_years_actual`) exhibits the strongest effect (t = 77.51), followed by historical claims variables. Three variables show weak or non-significant effects: `Area` (p = 0.13), `Has_Claimed_In_History` (p = 0.73, likely collinear with `N_claims_history_log`), and `Power_log` (p = 0.057). These were retained based on domain knowledge and overall model performance.

Appendix K. Claim and loss prediction plots

The claim and loss prediction plots are presented below.

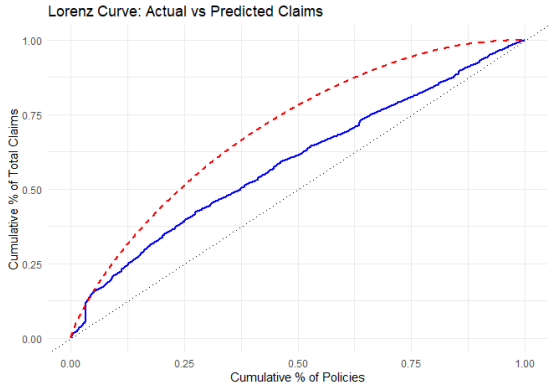

Figure 8: Lorenz Curve: Actual vs Predicted Claims

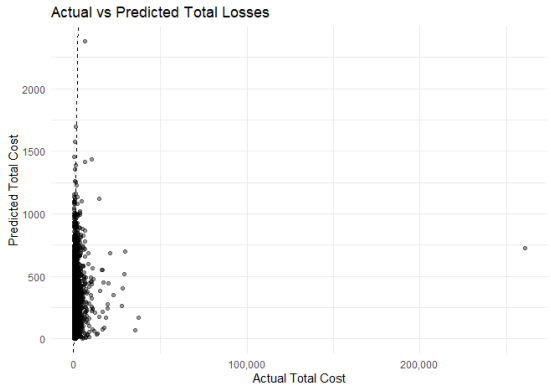

Figure 9: Actual vs Predicted Total Losses

Disclosure of Artificial Intelligence

This section outlines the various artificial intelligence (AI) tools utilized throughout this paper.

- AI was used to supplement literature finding process. Engines: scite.ai
- AI was used to assist with refining academic language and grammar; generating synonyms, rephrasing and shortening overly complex sentences to improve clarity. Engines: Copilot, ChatGPT, Overleaf Writefull.
- AI was used to get technical help with generating LaTeX code snippets for equations, tables and figures; ensuring the consistent format of the paper. Engines: Copilot, ChatGPT.