Statistical and Machine Learning Methods for Auto Insurance Claims Prediction: A Survey

Yara S. Kenupp[1], **Paulo Andre L. Castro**[1]

[1]Aeronautics Institute of Technology
Brazil

yara@ita.br, pauloac@ita.br

Abstract. *This paper reviews statistical and machine-learning methods for predicting claim frequency, claim severity, and loss cost in auto insurance. We cover generalized regression models, tree ensembles, distributional methods, neural and hybrid models, and attention-based architectures. The review compares datasets, targets, validation strategies, and empirical findings, with emphasis on exposure treatment, calibration, uncertainty, interpretability, and temporal validation. We also discuss open challenges related to benchmark scarcity, portfolio drift, high-cardinality variables, joint frequency-severity modeling, and the responsible use of telematics, text, and multimodal data.*

1. Introduction

Accurately estimating individual risk is essential for technical balance, financial sustainability, and competitiveness [de Jong and Heller 2008]. In individualized pricing, this estimation also supports actuarial fairness: lower-risk policyholders should not systematically subsidize higher-risk policyholders, while higher-risk profiles should be priced according to the expected losses they generate [Verbelen et al. 2018]. In pricing, the gross premium adds expense loadings, taxes, profit, and risk margins to the pure premium, the expected loss per unit of exposure, traditionally decomposed as frequency $\times$ severity [Frees et al. 2014].

The foundation of pricing lies in modeling frequency and severity; miscalibration compromises sustainability and can induce adverse selection. The growing availability of granular data (vehicle sensors, public datasets, telematics) has expanded the ML solution space with operational impact, though portfolio-level EBIT effects remain nascent [McKinsey Company 2025].

Historically, Generalized Linear Models (GLMs) have been the mainstay because of their transparency, statistical coherence, and regulatory compliance [Frees et al. 2014]. Their limitations motivate distributional extensions for overdispersion, excess zeros, heavy-tailed severities, and pure premium modeling [Hastie and Tibshirani 1990, Rigby and Stasinopoulos 2005, Lambert 1992, Mullahy 1986, Smyth and Jørgensen 2002, Jørgensen 1997]. When relationships are nonlinear or high-dimensional, gradient boosting and neural networks can improve out-of-sample performance under careful validation [Henckaerts et al. 2021, Guelman 2012, Wilson et al. 2024]. Calibration should also be assessed with proper scoring rules and distributional diagnostics [Gneiting and Raftery 2007]. More recently, self-attention models

have shown competitive results on tabular benchmarks and early insurance applications [Huang et al. 2020, Gorishniy et al. 2021, Sun et al. 2024].

Related work has examined machine learning in insurance at different levels of scope: Blier-Wong et al. [Blier-Wong et al. 2021] review applications in property and casualty insurance pricing and reserving, whereas Hanafy and Ming [Hanafy and Ming 2021] provide a focused empirical comparison of classifiers for binary claim-occurrence prediction using the Porto Seguro dataset. Against this background, this article systematizes the model landscape for frequency, severity, and pure premium in auto insurance, covering GLM-family distributions and modern machine-learning methods, including attention-based architectures; synthesizes evaluation criteria, such as out-of-sample performance and calibration, and practical aspects, such as interpretability, cost, and compliance; and offers scenario-driven guidelines for class imbalance, high-cardinality categorical variables, telematics, and excess zeros.

The remainder is organized as follows. Section 2 formalizes GLMs for frequency, severity, and pure premium (exposure offsets, families, and extensions such as Tweedie and zero-inflated models). Section 3 reviews **twenty-one** representative studies covering GBMs (including probabilistic variants), neural/hybrid models, and Transformers, as well as telematics and frequency–severity dependence. Section 4 concludes.

2. Foundations of Actuarial Modeling

This section formalizes the GLM structure used in insurance. We close by noting practical limitations that motivate more flexible variants.

2.1. The Classical Structure of GLMs

Generalized Linear Models (GLMs), introduced by [Nelder and Wedderburn 1972], became the backbone of frequency and severity modeling in non-life insurance pricing beginning in the 1990s [de Jong and Heller 2008]. The GLM framework models a response variable Y, such as the number of claims or the cost per event, as a function of covariates $X_1, X_2, \ldots, X_p$, assuming the conditional expectation $\mathbb{E}[Y \mid X]$ belongs to the exponential family and relates to the predictors through a link function $g(\cdot)$.

The basic specification is

$$g(\mu_i) = \eta_i = \beta_0 + \sum_{j=1}^{p} \beta_j x_{ij} + \text{offset}_i, \tag{1}$$

where $\mu_i = \mathbb{E}[Y_i]$, g is the link function (e.g., *log*), and the parameters β are estimated by maximum likelihood. For frequency models, the usual offset_i is $\log(\text{exposure}_i)$.

In actuarial applications, the choice of family and link depends on the target: for frequency (counts), Poisson with a log link and, in the presence of overdispersion, the Negative Binomial; for severity (positive values), distributions such as Gamma and Inverse Gaussian (log link) or lognormal regression (by modeling the log of severity); and for the pure premium (a mixture of point mass at zero and a positive continuous part), the compound Tweedie (Poisson–Gamma) with power parameter $1 < p < 2$ [Frees et al. 2014, Smyth and Jørgensen 2002].

In settings with excess zeros in frequency, extensions such as *zero-inflated* or hurdle models are common [Lambert 1992, Mullahy 1986]. GLMs are attractive for their interpretability and ease of implementation, as well as sound statistical properties under appropriate assumptions. For example, one can quantify the marginal effect of a covariate on the expected response, which is crucial when justifying rate variations to regulators [Frees et al. 2014].

On the other hand, GLMs have known limitations: they require explicit specification of nonlinearities and interactions (via splines and cross-terms), can be sensitive to outliers and to the choice of family/link, and may not accommodate excess zeros or heavy tails without appropriate extensions. These constraints motivate variants such as GAM/GAMLSS and, in high-dimensional settings with complex relationships, the use of machine-learning techniques [Hastie and Tibshirani 1990, Rigby and Stasinopoulos 2005, Dugas et al. 2003].

2.2. Key Challenges in Insurance Data

GLMs are a solid starting point, but insurance data often violate equidispersion, linearity, and light-tail assumptions, which hurts performance and motivates extensions and ML [de Jong and Heller 2008, Wüthrich and Merz 2023]. For frequency, excess zeros and overdispersion weaken the plain Poisson; the Negative Binomial adds variance [Cameron and Trivedi 2013], while zero-inflated and hurdle models address surplus zeros [Lambert 1992, Mullahy 1986]. Recent setups pair a binary occurrence model with a conditional count model and ML variants for classification/count [Poufinas et al. 2023, Wilson et al. 2024].

For severity, costs are positive and skewed; Gamma/Inverse Gaussian GLMs (log link) and lognormal regression are standard [de Jong and Heller 2008, Frees et al. 2014]. With heavy tails, tail components (e.g., GPD/Pareto) and high-quantile validation are used [McNeil et al. 2015, Hastie et al. 2009]. Tree-based and deep models, including quantile/distributional variants (quantile boosting, Tweedie boosting, QRF/NGBoost), deliver out-of-sample (OOS) gains and enable quantiles/intervals [Henckaerts et al. 2021, Friedman 2001, Meinshausen 2006, Yang et al. 2018, Duan et al. 2020].

High-cardinality categoricals, rare claims, and latent risk factors further complicate insurance modeling. High-cardinality variables can be handled through supervised binning, leakage-controlled target encoding, CatBoost's ordered statistics, or learned embeddings [Henckaerts et al. 2021, Micci-Barreca 2001, Prokhorenkova et al. 2018, Huang et al. 2020, Gorishniy et al. 2021]. Rare-claim settings require threshold tuning, probability calibration, and metrics such as precision–recall, while out-of-time validation is important under portfolio drift [Saito and Rehmsmeier 2015, He and Garcia 2009]. Richer data sources such as telematics, claim text, and attention-based models can reduce heterogeneity and improve explanations, but increase the need for stability and governance [Verbelen et al. 2018].

Beyond predictive accuracy, individualized pricing also raises fairness and governance concerns. More granular segmentation can reduce cross-subsidies among risk groups, but it may also amplify undesirable disparities when rating variables act as proxies for sensitive or operationally restricted attributes. Therefore, models based on richer data sources should be assessed not only for discrimination and calibration, but also for

stability, interpretability, and fairness-oriented governance.

2.3. Expanding with Machine Learning

Beyond GLMs, ML is most useful under high dimensionality, high-cardinality categoricals, complex interactions, or unstructured inputs (telematics, text). In auto insurance, modern ensembles often improve out-of-sample (OOS) and out-of-time (OOT) loss-cost performance over GLMs when validation is carefully designed [Wilson et al. 2024, Guelman 2012]. Gradient boosting remains a strong baseline and is auditable via SHAP, PD, and ICE; monotonicity aids governance [Henckaerts et al. 2021, Yang et al. 2018]. Distributional variants model full conditional laws (not just means) through probabilistic boosting [Sprangers et al. 2021, März and Kneib 2022, Chevalier and Côté 2025], while hurdle/zero-inflated extensions address surplus zeros [So and Valdez 2024].

For actuarial tabular data, categorical embeddings and attention (feature tokenization) show competitive pricing results [Brauer 2024], and transformer hybrids deliver accuracy with SHAP-based explanations [Sun et al. 2024]. In frequency–severity benchmarks, NNs are competitive end-to-end [Holvoet et al. 2025], related designs and prediction-combination strategies remain useful references [McDonnell et al. 2023, Ye et al. 2022], and AGLM preserves a GLM backbone while adding ML-learned terms [Fujita et al. 2020]. Recent integrations link credibility/random effects with attention (Credibility Transformer) [Richman et al. 2025].

3. Literature Review

This survey adopts a narrative, non-systematic approach. The twenty-one studies were purposively selected to represent the main methodological directions in auto-insurance claims prediction. Priority was given to works addressing claim frequency, claim severity, or pure premium, using public or widely adopted datasets when available, reporting out-of-sample evaluation, and providing distinct methodological or practical insights. Adjacent studies on text, multimodal claims, and time-to-event modeling were retained when their methods were transferable to pricing. Works outside the insurance domain or without sufficient methodological and empirical detail for comparison were not retained. The resulting corpus, summarized in Tables 1–3, is representative rather than exhaustive.

3.1. Explicit actuarial losses and extensions

Methods that directly optimize an actuarial distribution (Poisson, Gamma, Tweedie), preserving exposure semantics via an offset and enabling target-consistent evaluation (frequency, severity, or pure premium). This category includes boosting with a Tweedie loss, distributional/probabilistic variants, and zero-inflated Tweedie.

In [Yang et al. 2018], TDboost trains gradient boosting directly under the Tweedie loss, aligning the loss with the actuarial target of pure premium; this allows modeling, in a single target, both the mass at zero and the positive part, avoiding the coupling of separate frequency and severity models. In applications (including motor), TDboost outperforms GLM–Tweedie on out-of-sample metrics based on the ordered Lorenz curve and the Gini index. The central contributions are: coherence between target and loss; the ability to capture nonlinearities and interactions without extensive engineering; and gains with minimal changes to the actuarial workflow (the Tweedie semantics are preserved). The authors estimate the power index and dispersion (ρ, ϕ) by profile likelihood.

In [Chevalier and Côté 2025], point-estimate and probabilistic GBMs are compared for claim frequency and severity across five public datasets. The results indicate that LightGBM and XGBoostLSS are computationally efficient, CatBoost performs well with high-cardinality categorical variables, and probabilistic variants provide uncertainty estimates without systematically sacrificing mean predictive quality. The evaluation combines exposure-aware frequency modeling with distributional scoring and calibration (CRPS, quantile loss, and DPIT residuals) and Gini-based tariff comparisons.

Focusing on datasets with many zeros, [So and Valdez 2024] proposes zero-inflated Tweedie trained via gradient boosting (CatBoost), preserving actuarial semantics (exposure as an offset) and evaluating on the aggregate loss target. The authors compare two formulations: (i) separate models for the mean μ and the zero probability q; and (ii) a linked model with $q = f(\mu)$ (for example, $q = 1/(1 + \mu^{\gamma})$), using a custom loss. In synthetic telematics data (100k policies, strong zero inflation), ZITw-boosted models outperform Tweedie–GLM and Tweedie–boosting on MAD and deviance; the linked version also leads on Gini and is more robust on ordered Gini (min–max). Tests with compositional variables (CLR/ALR/ILR/PPCA) indicate that boosting handles them well without additional transformations.

A complementary direction is to improve *estimation* of Tweedie models themselves. In [Gao 2024], the Tweedie compound Poisson model for pure premium is fitted via an EM algorithm that is equivalent to an iteratively re-weighted Poisson–Gamma on augmented data. This avoids saddlepoint approximations used by DGLMs when zeros are abundant, reduces the computational burden of profiling the power parameter (ρ), and preserves exposure semantics through the offset. Empirically, the EM fit improves numerical stability and likelihood relative to profile-likelihood baselines in high–zero regimes.

Zero inflation can also be handled by explicitly modeling the extra mass at zero within a boosting framework. [Zhou et al. 2022] propose *EMTboost*, a boosting-assisted zero-inflated Tweedie that estimates the Tweedie mean and the zero probability jointly via EM, while the Tweedie component is learned non-parametrically with gradient tree boosting. On synthetic auto-claims with extreme zero proportion, EMTboost outperforms Tweedie-GLM and standard Tweedie boosting on deviance and risk-ordering metrics, providing a principled alternative when zero inflation is severe.

Table 1. Explicit actuarial losses and distributional extensions.

Ref.	Dataset / target	Method and main finding
Yang et al.	P&C, incl. MTPL / Pure premium	TDboost with Tweedie loss improved risk ordering over GLM/GAM.
Chevalier and Côté	P&C / Frequency + severity	Probabilistic GBMs provided calibrated quantiles and intervals; CatBoost helped with high-cardinality categoricals.
So and Valdez	Auto/telematics / Pure premium	Zero-inflated Tweedie with CatBoost improved deviance and Gini over Tweedie baselines.
Gao	P&C, incl. MTPL / Pure premium	EM estimation for Tweedie compound Poisson improved stability in high-zero regimes.
Zhou et al.	Auto, synthetic / Pure premium	EMTboost combined zero-inflated Tweedie and boosting, improving deviance and risk ordering.

3.2. Tree Ensembles and Hybrids

This strand covers tree-based boosting, GLM–ANN hybrids (stacking/CANN), forecast combination, joint frequency–severity modeling, and telematics (UBI) applications, with out-of-sample evaluation using target-consistent deviances, Lorenz/Gini ordering, and scale metrics (MAE, RMSE, and relative sum error, SUM).

In [Guelman 2012], a temporal holdout shows that GBM outperformed GLM on loss cost, using bucket orderings (GBM/GLM ratio) and loss-ratio behavior. Using a Belgian portfolio, [Henckaerts et al. 2021] build frequency/severity tariffs with a single tree, random forest, and gradient boosting, trained with Poisson deviance with exposure offset (frequency) and Gamma (severity); GBM reports gains over GLM on business indicators (e.g., lift, concentration) while retaining interpretability (VI/PDP/ICE), and a GLM surrogate is provided for deployment.

On `FreMTPL2`, [Wilson et al. 2024] compare GLM, GBM, neural networks, and a GLM+ANN hybrid for loss cost; the hybrid attains the lowest MAE on the test protocol (approximately 2012 vs. 2154 for NN, 2390 for GBM, and 2871 for GLM), with Actual–vs–Expected plots corroborating alignment. Across four datasets, [Holvoet et al. 2025] compare GLM, GBM, NNs, and CANN (GBM$\to$NN): for frequency, the flexible CANN typically achieves the lowest deviances; for severity, there is no universal winner, though embeddings help. They also compare tariffs via Lorenz/Gini and offer a GLM surrogate.

On LFHS/Kangaroo auto data ($\sim$94% zeros), [Ye et al. 2022] show that forecast combinations (ARM, ARM–Tweedie, LR–C, etc.) improve over single models when no Gini-dominant model exists (typical gains of $\sim$10% in MAE/RMSE, $\sim$25% in Re–RMSE, and $\sim$30% in SUM, relative to the best single candidate). When a strong Gini-dominant model is present (e.g., $G \approx 0.955$), simple averaging degrades performance.

To relax linearity/additivity and independence between frequency and severity, [Su and Bai 2020] propose D-FSBoost (Poisson for $\log \lambda$ and conditional Gamma for $\log \mu_N$), which attains smaller out-of-sample log-likelihood losses on `FreMTPL2` and lower Lorenz/Gini vulnerability under a minimax analysis. For UBI, [Henckaerts and Antonio 2022] build a baseline tariff with GBMs (Poisson with exposure offset; Gamma) and update prices multiplicatively via GLMs using telematics; no material selection bias is reported, LASSO highlights four dominant frequency predictors (mileage, harsh braking, lateral movements, night driving), deviance improves (about 2.6% train; 1.5% test) and Gini rises (train $0.275 \to 0.329$, test $0.136 \to 0.207$), with $\sim$60% of young drivers receiving average discounts of $\approx$5%.

Beyond accuracy, hybrids can also improve deployability. [Lindholm and Palmquist 2024] propose a *black-box–guided* workflow that distills a GBM (or another predictor) into a tariff GLM: they use partial-dependence (PD) functions from the black box to create partitions/groupings of covariates (including interactions), and then fit a categorical GLM on those partitions. In non-life case studies, the guided GLM preserves interpretability and approximates the ensemble's risk ordering, serving as a production-ready *surrogate*.

For MTPL pricing under stronger governance, [Ibrahim et al. 2025] show that depth-limited GBMs with monotonicity exhibit GAM-like partial effects, retain OOS gains vs. GLM under temporal holdout, and yield stable partial curves for audit. The

study discusses trade-offs among depth, smoothing, and tariff stability, suggesting shallow GBMs as a practical compromise between accuracy and transparency.

Table 2. Tree ensembles, hybrid models, telematics, and frequency–severity dependence.

Ref.	Dataset / target	Method and main finding
Wilson et al.	FreMTPL2 / Loss cost	GLM+ANN hybrid achieved the lowest MAE under OOS holdout.
Guelman	Real auto portfolio / Loss cost	GBM improved OOS loss-cost prediction over GLM with temporal split.
Henckaerts et al.	P&C / Frequency, severity, tariffs	Tree ensembles improved OOS performance and business lift, with interpretable diagnostics.
Holvoet et al.	MTPL / Frequency + severity	CANN improved frequency deviance; embeddings helped severity in some datasets.
Ye et al.	Kangaroo / Loss cost	Forecast combinations reduced errors when no single model dominated.
Lindholm and Palmquist	P&C / Tariff pricing	Guided GLM approximated ensemble risk ordering while preserving interpretability.
Ibrahim et al.	MTPL / Pricing	Depth-limited monotone GBM improved OOS performance and tariff stability.
Su and Bai	France / Frequency–severity	D-FSBoost improved OOS likelihood and tariff stability versus independent models.
Henckaerts and Antonio	MTPL telematics / UBI tariff	Dynamic UBI updates improved deviance/Gini and supported driver discounts.

3.3. Transformers and Attention in Tabular Data

Attention-based architectures have recently been adapted to actuarial tabular data, especially through hybrid designs that combine feature tokenization, learned embeddings, and classical actuarial structure. In this setting, the main objective is not simply to replace GLMs or tree ensembles, but to capture nonlinear interactions and high-cardinality effects while preserving exposure treatment, target-consistent losses, and interpretability requirements.

In [Brauer 2024], tabular transformers with feature tokenization are applied to non-life pricing (FreMTPL2freq), comparing GLM, FFNN, CANN, and FT-Transformer. The central result is that the proposed hybrids, CAFTT and LocalGLMftt, which combine GLM and transformer, achieve lower frequency deviance than the benchmarks, showing how to embed GLM structure (tokenization/embeddings, regularization) to preserve actuarial controls while capturing complex interactions. Embeddings help with high-cardinality categoricals; the trade-off is higher compute and tuning cost.

In [Sun et al. 2024], the Actuarial Transformer combines tree-based residual modeling with self-attention for auto-insurance risk evaluation. Using an auto-insurance dataset, the model is evaluated mainly through Poisson deviance and an Improvement Index, and it outperforms classical and machine-learning baselines in the reported experiments. SHAP analyses are used to identify domain-consistent risk drivers. This supports hybrid attention mechanisms that retain actuarial alignment while scaling to richer interactions.

Beyond purely tabular hybrids, claim text can add predictive signal. [Troxler and Schelldorfer 2024] present case studies (incl. multilingual car-accident nar-

ratives) showing how transformer-based NLP features from long, domain-specific descriptions are combined with tabular covariates in actuarial workflows; [Xu et al. 2023] use BERT embeddings from claim narratives to feed frequency and severity models, improving out-of-sample accuracy and stability when text is informative.

Although adjacent to pricing, multimodal auto-claims work further supports attention in claims pipelines: [Yang et al. 2023] combine text (transformer encoders), images, and structured data for fraud detection, reporting gains over structured-only baselines. This suggests that, when auxiliary unstructured inputs exist, attention representations can complement tree/GLM models while preserving exposure-weighted losses and calibration.

Complementing exposure-weighted hybrids, [Richman et al. 2025] introduce the Credibility Transformer, adding a credibility-inspired token that blends portfolio-level priors with observation-level evidence. The mechanism stabilizes training on tabular insurance data and improves out-of-sample deviance/log-loss versus deep baselines, and is complementary to [Brauer 2024, Sun et al. 2024].

A complementary direction concerns the timing of insurance events. Although claim frequency and loss cost remain the main focus of this survey, time-to-event modeling is relevant because exposure, claim occurrence, and claim development are intrinsically temporal. In this context, [Li et al. 2025] proposes a decoder-style transformer for survival analysis in insurance underwriting, showing how attention-based architectures can model time-to-claim outcomes. This line of work suggests a possible bridge between frequency, severity, and timing models in future actuarial architectures.

Table 3. Transformer and attention-based studies for actuarial and claims applications.

Ref.	Dataset / target	Method and main finding
Brauer	FreMTPL2 / Frequency	CAFTT and LocalGLMftt combine GLM and Transformer structures; lower deviance than GLM/NN, with higher tuning cost.
Sun et al.	Auto insurance / Risk evaluation	Actuarial Transformer combines tree residuals and attention; improves Poisson deviance and uses SHAP for explanations.
Troxler and Schelldorfer	Auto + property / Text applications	Transformer-based NLP features from claim descriptions are combined with tabular covariates in actuarial workflows.
Xu et al.	Claim narratives / Frequency, severity, loss	BERT-based text features improve OOS accuracy and stability when claim text is informative.
Yang et al.	Auto claims / Fraud	Multimodal model combines text transformers, images, and tabular data; improves AUC/F1 over structured-only baselines.
Richman et al.	P&C tabular / Pricing and risk	Credibility Transformer uses prior and evidence tokens; improves OOS deviance/log-loss over deep baselines.
Li et al.	Underwriting / Time-to-event	SurvFormer applies a decoder-style Transformer to survival analysis; improves concordance and calibration.

4. Conclusions

The reviewed studies suggest that empirical improvements in auto-insurance prediction tend to be more consistent in three situations: when the loss function is aligned with the actuarial target, when the model is adapted to the data regime, and when validation reflects realistic deployment conditions. For pure premium, Tweedie and zero-inflated Tweedie models preserve the compound structure of aggregate losses. For frequency and severity, boosting, neural networks, distributional methods, and hybrid architectures are often reported to improve risk ranking, deviance, calibration, or predictive accuracy relative to classical baselines. However, these gains are not automatic: they depend on careful exposure treatment, leakage prevention, temporal validation, calibration assessment, and model governance.

Several open challenges remain. First, the literature still relies heavily on a small number of public or widely used datasets, especially FreMTPL2, which limits the external validity of empirical comparisons. A promising research direction is the construction of broader benchmark protocols with standardized train/test splits, OOT validation, exposure-weighted losses, and comparable metrics for frequency, severity, and pure premium. Second, calibration and uncertainty quantification remain unevenly reported. Future work should evaluate not only point predictions, but also predictive distributions, quantiles, prediction intervals, and calibration under portfolio drift.

Third, joint modeling of frequency and severity remains an important research opportunity. Many practical pricing workflows still treat these components separately, although claim count and claim size may be dependent. Models that combine frequency-severity dependence, distributional learning, and actuarial constraints may contribute to both pricing adequacy and risk understanding. Fourth, high-cardinality categorical variables, telematics, claim text, and multimodal data open new possibilities for representation learning, especially with embeddings and attention-based models. These approaches are promising, but require stronger evidence on stability, interpretability, fairness, and operational feasibility.

Finally, the most useful future models are likely to be neither purely classical nor fully black-box. A promising path is the development of actuarially coherent hybrid models that combine the transparency of GLMs, the predictive flexibility of machine learning, and the uncertainty awareness of probabilistic methods. In this direction, SHAP/PD-ICE analyses, monotonicity constraints, GLM surrogates, calibration monitoring, and periodic recalibration should increasingly be treated as important components of responsible machine-learning deployment in auto insurance.

References

Blier-Wong, C., Cossette, H., Lamontagne, L., and Marceau, E. (2021). Machine learning in p&c insurance: A review for pricing and reserving. *Risks*, 9(1):4.

Brauer, A. (2024). Enhancing actuarial non-life pricing models via transformers. *European Actuarial Journal*, 14:991–1012.

Cameron, A. C. and Trivedi, P. K. (2013). *Regression Analysis of Count Data*. Number 53 in Econometric Society Monographs. Cambridge University Press, Cambridge, 2 edition.

Chevalier, D. and Côté, M.-P. (2025). From point to probabilistic gradient boosting for claim frequency and severity prediction. *European Actuarial Journal*, 15:707–752.

de Jong, P. and Heller, G. Z. (2008). *Generalized Linear Models for Insurance Data*. International Series on Actuarial Science. Cambridge University Press, Cambridge.

Duan, T., Avati, A., Ding, D. Y., Thai, K. K., Basu, S., Ng, A., and Schuler, A. (2020). NGBoost: Natural gradient boosting for probabilistic prediction. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 2690–2700. PMLR.

Dugas, C., Bengio, Y., Chapados, N., Vincent, P., Denoncourt, G., and Fournier, C. (2003). Statistical learning algorithms applied to automobile insurance ratemaking. In Shapiro, A. F. and Jain, L. C., editors, *Intelligent and Other Computational Techniques in Insurance: Theory and Applications*, pages 137–197. World Scientific.

Frees, E. W., Derrig, R. A., and Meyers, G., editors (2014). *Predictive Modeling Applications in Actuarial Science: Volume I, Predictive Modeling Techniques*. International Series on Actuarial Science. Cambridge University Press, Cambridge.

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

Fujita, S., Tanaka, T., Kondo, K., and Iwasawa, H. (2020). AGLM: A hybrid modeling method of GLM and data science techniques. Technical report, Actuarial Colloquium Paris.

Gao, G. (2024). Fitting Tweedie's compound poisson model to pure premium with the EM algorithm. *Insurance: Mathematics and Economics*, 114:29–42.

Gneiting, T. and Raftery, A. E. (2007). Strictly proper scoring rules, prediction, and estimation. *Journal of the American Statistical Association*, 102(477):359–378.

Gorishniy, Y., Rubachev, I., Khrulkov, V., and Babenko, A. (2021). Revisiting deep learning models for tabular data. In *Advances in Neural Information Processing Systems*, volume 34 of *NeurIPS*. arXiv:2106.11959.

Guelman, L. (2012). Gradient boosting trees for auto insurance loss cost modeling and prediction. *Expert Systems with Applications*, 39(3):3659–3667.

Hanafy, M. and Ming, R. (2021). Machine learning approaches for auto insurance big data. *Risks*, 9(2):42.

Hastie, T., Tibshirani, R., and Friedman, J. (2009). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, volume 2. Springer.

Hastie, T. J. and Tibshirani, R. J. (1990). *Generalized Additive Models*. Monographs on Statistics and Applied Probability. Chapman and Hall, London.

He, H. and Garcia, E. A. (2009). Learning from imbalanced data. *IEEE Transactions on Knowledge and Data Engineering*, 21(9):1263–1284.

Henckaerts, R. and Antonio, K. (2022). The added value of dynamically updating motor insurance prices with telematics collected driving behavior data. *Insurance: Mathematics and Economics*, 105:79–95.

Henckaerts, R., Côté, M.-P., Antonio, K., and Verbelen, R. (2021). Boosting insights in insurance tariff plans with tree-based machine learning methods. *North American Actuarial Journal*, 25(2):255–285.

Holvoet, F., Antonio, K., and Henckaerts, R. (2025). Neural networks for insurance pricing with frequency and severity data: A benchmark study from data preprocessing to technical tariff. *North American Actuarial Journal*, pages 1–44.

Huang, X., Khetan, A., Cvitkovic, M., and Karnin, Z. S. (2020). TabTransformer: Tabular data modeling using contextual embeddings. arXiv:2012.06678.

Ibrahim, J., Murfi, H., Novkaniza, F., and Pramono, A. (2025). Rethinking the interpretability of XGBoost for insurance pricing: A case study on motor third-party liability insurance. SSRN preprint.

Jørgensen, B. (1997). *The Theory of Dispersion Models*. Chapman & Hall/CRC Monographs on Statistics & Applied Probability. Chapman & Hall/CRC, London.

Lambert, D. (1992). Zero-inflated poisson regression, with an application to defects in manufacturing. *Technometrics*, 34(1):1–14.

Li, Y. et al. (2025). SurvFormer: A transformer based framework for survival analysis in insurance underwriting. In *Companion Proceedings of the Web Conference 2025*. ACM.

Lindholm, M. and Palmquist, J. (2024). Black-box guided generalised linear model building with non-life pricing applications. *Annals of Actuarial Science*, 18(3):675–691.

März, A. and Kneib, T. (2022). Distributional gradient boosting machines. arXiv:2204.00778.

McDonnell, K., Murphy, F., Sheehan, B., Masello, L., and Castignani, G. (2023). Deep learning in insurance: Accuracy and model interpretability using TabNet. *Expert Systems with Applications*, 217:119543.

McKinsey Company (2025). The state of AI. Online report. Accessed 10 July 2025.

McNeil, A. J., Frey, R., and Embrechts, P. (2015). *Quantitative Risk Management: Concepts, Techniques and Tools*. Princeton University Press, Princeton, NJ.

Meinshausen, N. (2006). Quantile regression forests. *Journal of Machine Learning Research*, 7:983–999.

Micci-Barreca, D. (2001). A preprocessing scheme for high-cardinality categorical attributes in classification and prediction problems. *SIGKDD Explorations*, 3(1):27–32.

Mullahy, J. (1986). Specification and testing of some modified count data models. *Journal of Econometrics*, 33(3):341–365.

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

Poufinas, T., Gogas, P., Papadimitriou, T., and Zaganidis, E. (2023). Machine learning in forecasting motor insurance claims. *Risks*, 11(9):164.

Prokhorenkova, L., Gusev, G., Vorobev, A., Dorogush, A. V., and Gulin, A. (2018). CatBoost: Unbiased boosting with categorical features. In *Advances in Neural Information Processing Systems*, volume 31. arXiv:1706.09516.

Richman, R., Scognamiglio, S., and Wüthrich, M. V. (2025). The credibility transformer. *European Actuarial Journal*, 15:345–379.

Rigby, R. A. and Stasinopoulos, D. M. (2005). Generalized additive models for location, scale and shape. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 54(3):507–554.

Saito, T. and Rehmsmeier, M. (2015). The precision-recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets. *PLOS ONE*, 10(3):e0118432.

Smyth, G. K. and Jørgensen, B. (2002). Fitting Tweedie's compound poisson model to insurance claims data: Dispersion modelling. *ASTIN Bulletin*, 32(1):143–157.

So, B. and Valdez, E. A. (2024). Zero-inflated Tweedie boosted trees with CatBoost for insurance loss analytics. arXiv:2406.16206.

Sprangers, O., Schelter, S., and de Rijke, M. (2021). Probabilistic gradient boosting machines for large-scale probabilistic regression. In *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*.

Su, X. and Bai, M. (2020). Stochastic gradient boosting frequency-severity model of insurance claims. *PLOS ONE*, 15(8):e0238000.

Sun, T., Yang, J., Li, J., Chen, J. Y., Liu, M., Fan, L., and Wang, X. (2024). Enhancing auto insurance risk evaluation with transformer and SHAP. *IEEE Access*.

Troxler, A. and Schelldorfer, J. (2024). Actuarial applications of natural language processing using transformers: Case studies for using text features in an actuarial context. *British Actuarial Journal*, 29:e4.

Verbelen, R., Antonio, K., and Claeskens, G. (2018). Unravelling the predictive power of telematics in car insurance pricing. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 67(5):1275–1304.

Wilson, A. A., Nehme, A., Dhyani, A., and Mahbub, K. (2024). A comparison of generalised linear modelling with machine learning approaches for predicting loss cost in motor insurance. *Risks*, 12(4):62.

Wüthrich, M. V. and Merz, M. (2023). *Statistical Foundations of Actuarial Learning and its Applications*. Springer Actuarial. Springer Nature Switzerland AG, Cham.

Xu, S., Manathunga, V., and Hong, D. (2023). Framework of bert-based nlp models for frequency and severity in insurance claims. *Variance*, 16(2).

Yang, J., Chen, K., Ding, K., Na, C., Wang, M., et al. (2023). Auto insurance fraud detection with multimodal learning. *Data Intelligence*, 5(2):388–412.

Yang, Y., Qian, W., and Zou, H. (2018). Insurance premium prediction via gradient tree-boosted Tweedie compound poisson models. *Journal of Business & Economic Statistics*, 36(3):456–470.

Ye, C., Zhang, L., Han, M., Yu, Y., Zhao, B., and Yang, Y. (2022). Combining predictions of auto insurance claims. *Econometrics*, 10(2):19.

Zhou, H., Yang, Y., and Qian, W. (2022). Tweedie gradient boosting for extremely unbalanced zero-inflated data. *Communications in Statistics – Simulation and Computation*.