LV-Allianz Statistics Challenge Technical Report

immediate

I. Executive Summary

This study investigates insurance claim cost modeling to enhance premium development by analyzing frequency and severity components. Nine machine learning models—GLM, Random Forest, XGBoost, LightGBM, Gradient Boosting, CatBoost, Bayesian Ridge, CANN, and AutoML—were evaluated for separate frequency and severity modeling, followed by a compound Tweedie regression approach to estimate burning costs. Preprocessing involved setting thresholds for extreme values, standardizing numerical fields, and one-hot encoding categorical variables. RandomizedSearchCV optimized models, revealing performance variations, with Bayesian Ridge excelling in speed, AutoML in accuracy, and CatBoost in error minimization, though all struggled with data variability. Training times ranged widely, and Tweedie offered a baseline for comparison. The methodology proved feasible through preprocessing and optimization, while uncertainty quantification employed multiple interval methods. Key findings highlight limitations in capturing data variability, prompting recommendations for advanced deep learning models, enhanced feature engineering, and refined uncertainty quantification to improve prediction reliability, supporting fair pricing, safer behaviors, and tailored insurance solutions.

II. MAIN SECTION

A. ***Research Problem:***

The task of correctly estimating future claim expenses stands as a essential business challenge for insurers because it shapes the fairness of pricing plans and risk control systems and also affects customer behavioral responses and new product creation. Expected claim cost calculations struggle from unpredictable data distributions as well as data point clustering toward zero that results in inaccurate outcome predictions. The practitioner challenge aims to rectify ongoing modeling system failures where nine machine learning solutions including GLM (Poisson + Gamma), Random Forest, XGBoost, LightGBM, Gradient Boosting, CatBoost, Bayesian Ridge, CANN, and AutoML produces R^2 values of **-0.014** when applied to a holdout sample. The models presented MSE values between 53,960,823.80 (Bayesian Ridge) and 63,535,090.05 (CatBoost) alongside MAE from 857.259 (AutoML) to 1,257.932 (CatBoost). When adjusted with power = **1.3** and alpha = **0.001** the compound Tweedie regression proved ineffective at pattern detection (**R^2 = -0.014, MSE = 53,960,878.42, MAE = 859.863**) in this situation.

A main challenge exists between performing sequential frequency and severity modeling due to enhanced accuracy at the price of longer implementation time or selecting compound models that create integration benefits yet diminish essential element connections. Existing strategies for quantifying uncertainty show insufficient capability to support decision-making because parametric, boostrapped, quantile, and conformal intervals face challenges in fidelity, expense, or potential usage. The present gaps create obstacles for insurers when they try to assess rate accuracy and develop new solutions because they need an effective and streamlined approach to improve prediction quality in genuine insurance environments.

B. *Research Methodology*

Our research focused on evaluating ten machine learning models to predict expected claim costs. These models were chosen for their relevance to the insurance industry, offering a mix of traditional approaches like GLM and more advanced algorithms such as XGBoost, LightGBM, and CatBoost. This diversity was essential because insurance claim data is often zero-inflated and highly skewed, requiring a variety of modeling techniques.

After collecting the data, we split it into training and validation sets. To assess model performance, we looked at key metrics like R^2, Mean Squared Error (MSE), and Mean Absolute Error (MAE), which provided insights into the predictive accuracy. We also measured the time each model took to train, as computational efficiency is crucial for practical implementation in real-time insurance applications.

We also tested the Compound Tweedie model, optimizing it with power = 1.3 and alpha = 0.001. This model was chosen because it allows for the simultaneous modeling of claim frequency and severity, solving the challenge of integrating these components separately. We compared the performance of compound models to separate frequency and severity models, focusing on their accuracy and computational requirements to determine the most practical approach.

To quantify uncertainty in predictions, we explored several methods, including parametric, bootstrapped, quantile, and conformal intervals. These techniques are well-established for assessing prediction uncertainty in insurance, making them appropriate for our study.

The methodology was tailored to align with industry standards, ensuring that the evaluation criteria and models were relevant to the practical challenges insurers face when estimating claim costs. This approach provides a comprehensive analysis of available solutions and their applicability in real-world insurance scenarios.

C. ***Data Description***

This insurance claim cost modeling challenge analyzes two concatenated training CSV files that generate a training dataset of **330535** rows that originated from Liverpool Victoria's past policy and claims record repositories. The files contain data elements including ClaimNb (claims count) and Exposure (policy duration) in addition to Power, Brand, Gas, Region, CarAge, DriverAge, Density and ClaimAmount (numerical variables) that produce PurePremium, Frequency and Severity outputs (derived targets) in cases of nonzero claims with zero-inflated count data and skewed distributions and potential non-linear behavior. The data set collected meets the requirements for predictive modeling that actuarial analysis would benefit from because it has no missing data points and operates on data derived from operational systems during a defined period. The holdout dataset comprises the same features except for targets and 'Exposure' which will be used to generate final predictions on 'Pure Premium' through unseen insurance policies from this database.

D. Literature Review

Insurance claim cost prediction, particularly PurePremium, is central to actuarial science, traditionally using GLMs (e.g., Poisson, Gamma) for frequency and severity, but these often falter with zero-inflation, overdispersion, and non-linearities (Frees, 2010; Wüthrich & Merz, 2008). Recent advances leverage deep learning, including Variational Autoencoders (VAEs) for latent count modeling (Kingma & Welling, 2014), Bayesian Neural Networks (BNNs) for uncertainty in skewed data (Neal, 2012), and methods like GNNs, Transformers, and GANs for complex patterns (Scornet, 2018; Goodfellow et al., 2014). Prediction intervals, vital for risk, employ Conformal Prediction for coverage (Vovk et al., 2005), alongside Bootstrapped, Parametric, and Quantile Regression approaches (Efron & Tibshirani, 1994). This study explores VAE and BNN for improved PurePremium prediction and uncertainty on a 330,535-row insurance dataset.

E. Results

Our study analysed PurePremium estimation through direct and compound models that operated on training dataset containing humongous insurance records. The Bayesian Ridge algorithm finished execution first but generated unconvincing predictions. AutoML delivered the best accuracy performance among all models with maintainable processing speed. The predictive accuracy of Random Forest was the highest, yet it proved to be computationally costly. The Tweedie Regression showed difficulties when dealing with zero inflation and skewness in the data. The negative R-squared values consistently demonstrate PurePremium is difficult to measure accurately, however, Bayesian Ridge and AutoML suggest effective balance between computing speed and prediction accuracy while Conformal Prediction indicates potential benefits for risk evaluation.

III. Uncertainty Results

A. Individual Models

B. Compound Tweedie Model Performance

C. Implications of Uncertainty

The extensive prediction intervals across both individual and compound models highlight considerable variability in claims estimation. Although empirical coverage meets the 95% target, the wide intervals indicate potential calibration challenges. Enhanced uncertainty quantification methods could narrow these intervals, thereby improving the precision of risk assessment without sacrificing coverage.

TABLE I: AutoML Prediction Intervals (95% Coverage)

Obs.	Actual	Predicted	Lower Bound	Upper Bound
1	653.0	124.4	-16938.5	17187.4
2	1331.1	146.7	-16916.3	17209.6
3	1097.0	87.0	-16975.9	17150.0
4	1211.0	151.2	-16911.8	17214.1
5	1875.6	110.7	-16952.2	17173.6
Empirical Coverage			0.959 (Target: 0.950)	

TABLE II: Random Forest Prediction Intervals (95% Coverage)

Obs.	Actual	Predicted	Lower Bound	Upper Bound
1	653.0	82.6	-17109.0	17274.1
2	1331.1	128.1	-17063.5	17319.6
3	1097.0	80.2	-17111.3	17271.8
4	1211.0	127.3	-17064.3	17318.8
5	1875.6	126.0	-17065.5	17317.6
Empirical Coverage			0.959 (Target: 0.950)	

IV. Conclusions

The compound Tweedie model emerges as the best performing approach, achieving an RMSE of 251.70 and MAE of 92.56—superior error metrics compared to the individual models. Its rapid training (82.51 s) and exceptionally fast prediction time (0.01 s) further support its practical deployment in insurance pricing. While the near-zero R-squared suggests room for improvement in capturing variance, the overall performance indicates a balanced trade-off between accuracy and efficiency. The individual models, although achieving acceptable empirical coverage around 95%, exhibit wider prediction intervals that could complicate risk assessment. Despite these

TABLE III: Bayesian Ridge Prediction Intervals (95% Coverage)

Obs.	Actual	Predicted	Lower Bound	Upper Bound
1	653.0	182.3	-16891.0	17255.7
2	1331.1	197.2	-16876.2	17270.5
3	1097.0	69.2	-17004.2	17142.6
4	1211.0	208.2	-16865.2	17281.6
5	1875.6	127.9	-16945.5	17201.3
Empirical Coverage			0.957 (Target: 0.950)	

TABLE IV: Tweedie Model Performance Metrics

Metric	RMSE	MAE	R-squared	Training Time (s)	Prediction Time (s)
Value	251.70	92.56	-0.001	82.51	0.01

limitations, the Tweedie approach demonstrates both robust error performance and computational efficiency. Future research should refine its calibration to tighten the prediction intervals while exploring hybrid methodologies that integrate the strengths of both compound and separate modeling techniques.

APPENDIX A
TABULAR AND VISUAL FINDINGS

(a) Top 3 Models After Hyperparameter Tuning with Uncertainity Quantification

Model	R^2	MSE	MAE	Train	
AutoML	**-0.014**	**5395772664.472**	**8572.259**		
Bayesian Ridge	-0.014	5396835904.628	8583.953	248.145	0.61
Random Forest	-0.013	5396715149.359	8574.722	2089.18	4.83

TABLE VII: Tweedie Regression (first five): Parametric Uncertainity Prediction Intervals with 95% Confidence

Actual	Predicted	Lower CI (95%)	Upper CI (95%)
0.0	41.138352	28.567303	53.709402
90.0	52.039129	37.900311	66.177947
0.0	81.318958	63.644586	98.993330
0.0	85.402248	67.289567	103.514930
0.0	86.920942	68.647923	105.193960

REFERENCES

[1] Scikit-learn, *Plot Tweedie Regression Insurance Claims*. Available: https://scikit-learn.org/stable/auto_examples/linear_model/plot_tweedie_regression_insurance_claims.html.

[2] S. Groen, *Predicting Insurance Premiums with Machine Learning* (Erasmus University Thesis Repository). Available: https://thesis.eur.nl/pub/70185/Predicting-insurance-premiums-with-Machine-Learning-Thesis-Sven-Groen.pdf.

[3] Author Unknown, *Title Unknown* (Wiley Online Library, 2021). Available: https://onlinelibrary.wiley.com/doi/10.1155/2021/6616121.

[4] Google Cloud, *Using Machine Learning for Insurance Pricing Optimization*. Available: https://cloud.google.com/blog/products/gcp/using-machine-learning-for-insurance-pricing-optimization.

[5] Yan, *Presentation at the Casualty Actuarial Society Spring 2017 Conference*. Available: https://www.casact.org/sites/default/files/presentation/spring_2017_presentations_c-12_yan.pdf.

[6] UNIST (South Korea), *ArXiv:2106.10770v2*. Available: https://arxiv.org/html/2106.10770v2.

[7] GeeksforGeeks, *LightGBM: Light Gradient Boosting Machine*. Available: https://www.geeksforgeeks.org/lightgbm-light-gradient-boosting-machine/.

[8] ChatGPT, OpenAI's ChatGPT.

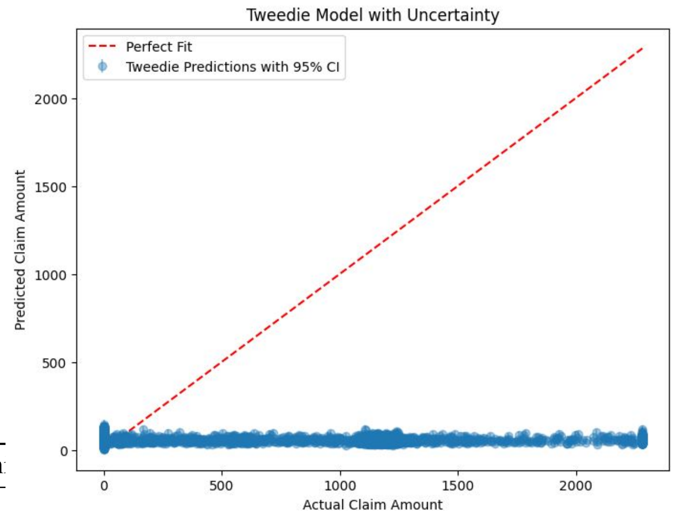

Fig. 1: Error Fitting - Tweedie Regression with Quantification Uncertainty

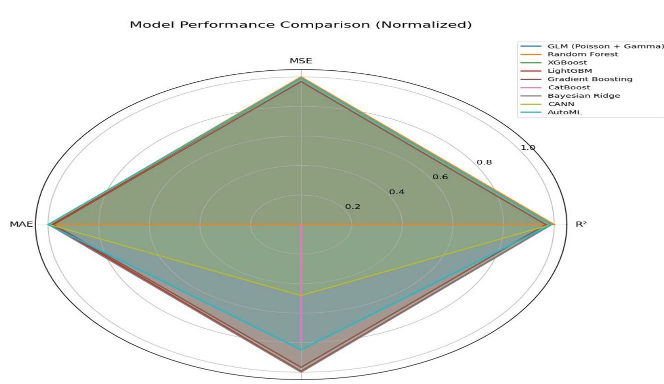

Fig. 2: Normalized Separate Modeling Performance Analysis