

Latent Structures in Zero-Inflated Risk Domains: An Elastic–Tweedie Synergy for Claim Forecasting

P. B. W. S. R. Kumarasinghe^{1*}, N. A. D. N. Napagoda²

¹*Department of Information Technology, SIBA Campus, Pallekele, Sri Lanka*

²*Department of Mathematical Sciences, Faculty of Applied Sciences, Wayamba University of Sri Lanka, Kuliyaipitiya, Sri Lanka.*

Corresponding author*: rasadari.k@siba.edu.lk

Abstract

The frequency of insurance claims presents a unique modeling challenge due to high-dimensional inputs, strong feature correlations, and the dominance of zero-inflated outcomes. Conventional statistical models often fall short under these conditions, failing to capture the underlying structure of complex data sets. This study proposes an advanced predictive framework integrating Elastic Net regularization and a Tweedie-distribution-based XGBoost algorithm to address these issues in the context of motor insurance. Those methodologies were applied to the French Motor Claims data set, which contains over 678,000 policies, to distill influential variables while suppressing redundancy and noise. Lasso Regression, Elastic Net and the Boruta algorithm were employed to select relevant features. Elastic Net, in particular proved effective in identifying critical predictors including Exposure, Vehicle Age, Driver Age, BonusMalus, Area, and Fuel Type by balancing sparsity and multicollinearity. These features were used to train both standard and Tweedie-distribution-based XGBoost models. Performance was evaluated using RMSE, MAE, and R^2 , where the Tweedie XGBoost model guided by Elastic Net-selected features achieved the highest accuracy and explanatory power. The proposed architecture not only offers superior generalization and interpretability but also exhibits robustness in modeling skewed, zero-dominated distributions inherent to claim data. Beyond predictive enhancement, this framework has practical implications for actuarial science, particularly in dynamic pricing strategies, refined segmentation, and adaptive underwriting. This approach marks a shift toward more nuanced and scalable machine learning paradigms in insurance analytics by integrating statistically grounded feature selection with distribution-aware boosting.

Keywords: Claim frequency prediction, Feature selection, Tweedie-distribution-based XGBoost, Elastic Net, Lasso Regression

Introduction

Accurately predicting insurance claim frequency is essential for effective underwriting, pricing, and risk management in the insurance industry. While traditional models like Poisson and Negative Binomial regression have been widely used, they often fall short with modern, complex data sets that are high-dimensional and zero-inflated (Quan *et al.*, 2020). Building on this understanding, Quan *et al.* (2020) proposed hybrid tree-based models for insurance claims, demonstrating greater accuracy compared to traditional statistical approaches. This study introduces a machine learning approach combining advanced feature selection and gradient boosting to enhance predictive performance of insurance claim frequency models.

Therefore, methodology employs Lasso Regression, Elastic Net and the Boruta algorithm to select relevant features from the comprehensive French Motor Claims data set. Lasso and Elastic Net encourage sparsity by removing irrelevant variables, while Boruta identifies statistically significant features using random forests. These complementary methods capture both linear and non-linear relationships (Friedman, Hastie, & Tibshirani, 2010 and Kursa & Rudnicki, 2010).

Using features selected by Lasso, Elastic Net or Boruta, six models were created as three for standard XGBoost approach and another three for Tweedie XGBoost approach. Gradient boosting models namely XGBoost have been shown to effectively handle complex and structured data sets (Chen & Guestrin, 2016). Tweedie XGBoost is well-suited to claim data due to its ability to handle zero-inflated and skewed distributions. Models were evaluated using RMSE and R², and feature importance was assessed through gain-based scores. This research identifies the most effective feature selection and model combinations for predicting claim frequency, contributing practical insights for actuaries and data scientists and promoting the integration of machine learning in insurance analytics (Biagini, 2022).

The novelty of this study lies in combining advanced feature selection methods with a Tweedie-based XGBoost framework for claim frequency prediction, which has not been extensively explored in prior insurance analytics research.

Materials and Methods

This study is based on the French Motor Claims data set, which contains detailed records of 678,013 motor insurance policies. Each policy record includes a variety of information particularly policyholder demographics, vehicle characteristics, regional data, and historical claims history. This massive data set provides a comprehensive basis for modeling and predicting insurance claim frequencies.

Data preprocessing

Before building any models, the data set underwent several preprocessing steps to ensure data quality and suitability for analysis. Handling missing data, duplicate data and outliers are not used here since the data set doesn't have such conditions. Categories were converted into numerical representations with categorical encoding, and the data set was split into two subsets one was used for training the models (80%), while the other was held out as a test set (20%).

Feature selection

Feature selection is a crucial step in building reliable and interpretable predictive models, especially when working with large data sets containing numerous variables. The primary aim is to identify the most relevant features that significantly influence the target variable—in this case, the frequency of motor insurance claims—while reducing noise and complexity that could degrade model performance. This study employed three complementary feature selection techniques.

Lasso regression

Lasso applies an L1 regularization penalty to the regression coefficients, encouraging sparsity by shrinking some coefficients exactly to zero. This enables automatic feature selection by eliminating less important predictors. The tuning parameter λ was selected through cross-validation to control the strength of regularization and balance the trade-off between model simplicity and predictive accuracy.

$$\hat{\beta}_{LASSO} = \arg \min \left\{ \sum_{i=1}^n (y_i - x_i \beta)^2 + \lambda \sum_{j=1}^p |\beta_j| \right\} \text{----- (1)}$$

Where;

- y_i – actual response variable
- x_i – feature vector for observation
- β – vector of coefficients
- λ – regularization parameter controlling sparsity

Elastic net

Elastic Net combines the L1 penalty of Lasso and the L2 penalty of Ridge regression. This approach helps in scenarios where predictors are highly correlated, as it tends to select groups of related features together rather than arbitrarily choosing one. Parameters λ and α were selected through cross-validation to govern the balance between the two penalties, allowing for flexible feature selection that can handle multicollinearity effectively.

$$\hat{\beta}_{EN} = \arg \min \left\{ \sum_{i=1}^n (y_i - x_i \beta)^2 + \lambda \sum_{j=1}^p |\beta_j| + \frac{1-\alpha}{2} \sum_{j=1}^p (\beta_j)^2 \right\} \text{----- (2)}$$

where;

- y_i – actual response variable
- x_i – feature vector for observation
- β – vector of coefficients
- λ – controls the overall regularization strength
- α – mixing parameter

Boruta algorithm

Boruta is a wrapper method based on random forests that identifies all features carrying meaningful information by comparing their importance with randomized “shadow” features. It uses a p-value threshold of 0.01 to determine which features are significantly relevant. This rigorous process ensures that no important variables are overlooked, even those with weaker but still significant effects.

These methods were applied to the data set to systematically reduce the feature space to a core subset that includes variables related to policy exposure, geographic location, vehicle attributes, driver demographics, and historical claim discounts. This strategic variable selection facilitates model optimization by emphasizing the most influential predictors, thereby enhancing both generalizability and cross-context interoperability. Feature selection (Lasso, Elastic Net, and Boruta) was carried out within training folds of cross-validation to avoid information leakage and ensure unbiased model evaluation.

Model Development

The predictive modeling approach involved training gradient boosting models, known for their effectiveness with structured tabular data and their ability to capture complex, nonlinear relationships. Two types of models were developed.

Standard XGBoost

This model utilized the traditional squared loss function to minimize the difference between predicted and actual claim counts. XGBoost builds an ensemble of decision trees sequentially, where each new tree attempts to correct the errors of the previous ensemble. The algorithm incorporates regularization to control overfitting and can handle missing values and mixed feature types efficiently.

$$0 = \sum_{i=1}^n \Gamma(y_i, \hat{y}_i^{(t)}) + \sum_{k=1}^t \Omega(f_k) \text{-----} (3)$$

Where;

- y_i – actual value
- $\hat{y}_i^{(t)}$ – predicted value at iteration t
- f_k – kth decision tree
- $\Omega(f)$ – regulation for tree complexity
- Γ – loss function

Tweedie-distribution-based XGBoost

A customized Tweedie XGBoost model was employed to better accommodate the distributional characteristics of insurance claim data, which typically include many zeros and positive skewness was employed. The Tweedie distribution is appropriate for modeling non-negative data with a mass at zero and continuous positive values. Integrating this loss function into XGBoost allows the model to better fit the specific statistical nature of claim frequency data.

For both models, the data set was split into training and testing sets to ensure unbiased model development. Hyperparameters, primarily learning rate, tree depth, and the number of trees were tuned using cross-validation techniques to optimize predictive performance while preventing overfitting.

$$L_{TWEEDIE}(y_i, \hat{y}_i) = \frac{1}{\phi} \left[\frac{(y_i^{2-p})_i}{(1-p)(2-p)} - \frac{y_i \exp((1-p)\hat{y}_i)}{1-p} + \frac{\exp((2-p)\hat{y}_i)}{2-p} \right] \text{-----} (4)$$

Where;

- y_i – observed target
- \hat{y}_i – predicted log-link value
- p – Tweedie power parameter
- ϕ – dispersion parameter

Model evaluation

Model performance was assessed using Root Mean Squared Error (RMSE), R-squared value (R²), and Mean Absolute Error (MAE). These metrics provided a comprehensive view of performance by balancing sensitivity to outliers with overall predictive accuracy.

Results

Evaluation of feature selection techniques: Lasso, Elastic Net, and Boruta

Table 1 presents the outcome of the feature selection process using three different techniques: Lasso Regression, Elastic Net, and Boruta. A check mark (✓) indicates that the feature was selected by the corresponding method. Each method identified a subset of features from the original data set deemed most relevant for predicting claim frequency. Common features chiefly Exposure, Vehicle Age (VehAge), and BonusMalus were consistently selected across all three methods, indicating their strong predictive potential. Lasso and Elastic Net, both regularization-based methods, additionally selected features like Area, Driver Age (DrivAge), and Vehicle Fuel Type (VehGas), while Boruta, a wrapper method based on random forests, was more conservative in its selection. The presence of both overlapping and unique variables highlights the complementary nature of these feature selection strategies.

Table 1: Features Selected by Lasso, Elastic Net, and Boruta Algorithms

Variables	Lasso	Elastic Net	Boruta
	Optimal lambda: 0.01	Optimal lambda: 0.03 Alpha: 0.2	p-Value: 0.01
Intercept			
IDpol (policy ID)			
Exposure (exposure period)	✓	✓	✓
Area (area code)	✓	✓	
VehPower (power of the car)			
VehAge (vehicle age, in years)	✓	✓	✓
DrivAge (driver age)	✓	✓	
BonusMalus (Bonus/malus)	✓	✓	✓
VehBrand (car brand)			
VehGas (car gas, Diesel or regular)	✓		
Density (density of inhabitants)			
Region (policy regions in France)			

Model performance

Table 2 summarizes the predictive performance of the models built using XGBoost and XGBoost Tweedie algorithms, with features selected by Lasso, Elastic Net, and Boruta. Performance was assessed using three evaluation metrics: Root Mean Squared Error (RMSE), R-squared (R^2), and Mean Absolute Error (MAE). Among all models, the XGBoost Tweedie model with Elastic Net-selected features demonstrated the best performance, achieving the lowest RMSE (0.1093) and MAE (0.0875), along with the highest R^2 score (0.7921). These results indicate that the combination of the Tweedie distribution (which is well-suited for zero-inflated count data) and Elastic Net feature selection yields a highly accurate and robust model for predicting insurance claim frequency. Other models also performed reasonably well, with XGBoost using Boruta features coming close in terms of R^2 and MAE, but not outperforming the Tweedie variant.

Table 2: Performance Metrics of XGBoost and XGBoost Tweedie Models with Different Feature Selection Methods

Model	XGBoost (Lasso)	XGBoost (Elastic Net)	XGBoost (Boruta)	XGBoost Tweedie (Lasso)	XGBoost Tweedie (Elastic Net) *	XGBoost Tweedie (Boruta)
RMSE	0.1311	0.1158	0.1220	0.1266	0.1093	0.1163
R^2	0.6998	0.7645	0.7421	0.7195	0.7921	0.7654
MAE	0.1049	0.0927	0.0976	0.1013	0.0875	0.0930

Discussion

Despite claim frequency being count data, the zero-inflated and skewed nature of the dataset justified using a Tweedie-based boosting framework. Though typically applied to claim severity, Tweedie boosting can effectively handle zero-dominated counts and delivered superior performance here, making it a pragmatic predictive choice. This study aimed to predict how frequently insurance policyholders file claims, a task complicated by a data set with many zero claims and a skewed distribution. In this study, three feature selection methods (Lasso, Elastic Net and Boruta) were employed to retain meaningful variables and reduce redundancy. Key predictors like **Exposure**,

VehAge, BonusMalus, and DrivAge were consistently identified as important. The best-performing model was XGBoost with a Tweedie model, using features selected by Elastic Net. This combination achieved the highest predictive accuracy (lowest error and highest R²), effectively handling the zero-inflated data and ensuring model robustness through relevant input features.

Overall, the Tweedie XGBoost Model with Elastic Net feature selection demonstrated strong performance, balancing accuracy, interpretability, and efficiency, and highlighting the value of carefully tailored machine learning techniques for real-world insurance analytics and risk assessment.

Conclusions

This integration of Elastic Net feature selection with a Tweedie XGBoost model represents a novel contribution to claim frequency modeling, offering both predictive strength and practical value. This study demonstrates the effectiveness of combining Elastic Net feature selection with a Tweedie-based XGBoost model to improve the prediction of insurance claim frequency. The challenge stemmed from a highly zero inflated data set, where most policyholders reported zero claims conditions under which traditional models often underperform. Elastic Net efficiently selected key correlated features notably Exposure, Vehicle Age, Bonus-Malus, and Driver Age while Tweedie XGBoost handled the zero-inflated, positively skewed distribution of claim counts. This Tweedie XGBoost Model with Elastic Net feature selection achieved superior accuracy and interpretability compared to other models. The results provide practical value for actuarial decision-making, aiding in pricing, risk assessment, and portfolio management. The method is scalable and adaptable, with potential applications in broader insurance analytics and risk modeling. These findings extend beyond predictive accuracy to inform actionable decision-making for insurers. The key features identified through Elastic Net; Exposure, Vehicle Age, Driver Age, BonusMalus, Area, and Vehicle Fuel Type represent core risk factors that can guide pricing structures, underwriting rules, and policyholder segmentation. Incorporating these variables into operational models allows insurers to refine risk assessment processes and develop adaptive policy strategies. The Tweedie XGBoost model further supports scalability and long-term monitoring of claim behavior across diverse customer profiles. The use of Tweedie XGBoost, although unconventional for count data, was motivated by the zero-inflated and skewed nature of the claims frequency dataset and proved effective in predictive accuracy.

References

- Biagini, L. (2022). Applications of machine learning for the ratemaking in agricultural insurances. *arXiv*.
- Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 785–794.
- Friedman, J., Hastie, T., & Tibshirani, R. (2010). Regularization paths for generalized linear models via coordinate descent. *Journal of Statistical Software*, 33(1), 1–22.
- Kursa, M. B., & Rudnicki, W. R. (2010). Feature selection with the Boruta package. *Journal of Statistical Software*, 36(11), 1–13.
- Quan, Z., Wang, Z., Gan, G., & Valdez, E. A. (2020). Hybrid Tree-based Models for Insurance Claims. *arXiv preprint arXiv:2006.05617*.