Jurnal Matematika UNAND
Vol. **15** No. **1** Hal. 78 – 94
Edisi Januari 2026
ISSN : 2303–291X
e-ISSN : 2721–9410

https://doi.org/10.25077/jmua.15.1.78-94.2026

TARIFF ANALYSIS OF MOTOR INSURANCE USING GENERALIZED LINEAR MODEL (GLM) AND GRADIENT BOOSTING MACHINE (GBM)

YUNIKE JEMIS FIFNELAVINDY ALSITANINGTYAS[a,*], HUBBI MUHAMMAD[b], ADHITYA RONNIE EFFENDIE[c]

[a,c] *Dept. of Mathematics, Universitas Gadjah Mada, Yogyakarta, Indonesia,*
[b] *Dept. of Mathematics, Universitas Pamulang, Serang, Indonesia,*
email : yalsitaningtyas@mail.ugm.ac.id, hubbi.muhammad@unpam.ac.id, adhityaronnie@ugm.ac.id

Received October 27, 2024 Received in revised form January 11, 2026
Accepted January 15, 2026 Available online January 31, 2026

Abstract. *The insurance sector operates by managing the transfer of risk from policyholders to insurance providers, where premiums are charged as compensation for the assumed risk. Traditionally, premium determination in motor vehicle insurance relies on the Generalized Linear Model (GLM), which requires the response variable to follow a distribution from the exponential family and may have limitations in capturing non-linear relationships and complex interactions among rating factors. To address these limitations, this study compares the performance of the Generalized Linear Model (GLM) and the Gradient Boosting Machine (GBM) in modeling claim frequency and claim severity for motor vehicle insurance premiums. The analysis is conducted using an insurance dataset obtained from a public data repository, and both models are evaluated using K-Fold Cross Validation. Model performance is assessed based on the Root Mean Square Error (RMSE), which measures the average magnitude of prediction errors and is commonly used to evaluate predictive accuracy. The results indicate that the GBM consistently produces lower RMSE values than the GLM for both claim frequency and claim severity modeling, indicating superior predictive performance. However, despite its higher accuracy, the GBM model lacks the interpretability inherent in the GLM framework, which remains crucial for transparency and regulatory considerations in insurance premium determination. These findings suggest that while GBM is effective for improving prediction accuracy, GLM remains valuable for interpretability, and a complementary use of both approaches may provide optimal results in actuarial pricing applications.*

Keywords: Gradient Boosting Machine, Generalized Linear Model, Insurance Premium.

1. Introduction

The insurance industry is one of the financial sectors directly affected by the Covid-19 pandemic. Motor vehicle insurance premiums experienced a 19.9% decline in the

*Corresponding Author

first quarter of 2021 compared to the first quarter of 2020. This decline is attributed to the Covid-19 pandemic, which resulted in policyholders traveling less and consequently a 30.4% decrease in motor vehicle insurance claims in the first quarter of 2021 compared to the same period in 2020. Despite policyholders traveling less frequently, the premiums paid did not differ significantly from the pre-Covid-19 pandemic period because the premium rate determination system for motor vehicle insurance products still relies solely on the insured vehicle's coverage amount. Insurance companies are competing to offer their products, and one strategy they employ is setting lower premium rates. According to Regulation No. Per-07/BL/2009 issued by the Chairman of the Capital Market and Financial Institutions Supervisory Agency, it stipulates that the premiums imposed on insurance participants consist of pure premiums plus administrative costs, acquisition costs, and the company's profit. Premiums must be calculated based on several aspects possessed by or characteristics of the insured, commonly known as rating factors. Premiums calculated considering these rating factors are referred to as premium rates.

The calculation of premiums can be performed using the Generalized Linear Models (GLMs) framework. In [1], as well as in [2], GLMs are utilized to calculate premiums in general insurance. Additionally, [3] researched a Compound Poisson model integrated with GLM. In [4], it is demonstrated that the assumption of independence between frequency and severity claims can be relaxed by employing Generalized Linear Models and incorporating rating factors into the model as a modification. In [5], it is stated that for large-sized data, Generalized Linear Models remain consistent in variable selection, suggesting their suitability for large datasets.

On the other hand, tree-based machine learning is a method that leverages a tree structure to make decisions without requiring any assumptions. This method is suitable for datasets with a substantial amount of data, and the distribution of the data is unknown. Despite the often-perceived "black box" nature of machine learning, numerous studies, such as those by [6], have explored its interpretability.

Based on the background outlined above, the researcher will develop a model capable of calculating motor vehicle premiums based on rating factors using both the Generalized Linear Model (GLM) and the Gradient Boosting Machine (GBM) methods. The criterion used to determine the best method will be the Root Mean Square Error (RMSE), employing K-fold cross-validation.

2. Theoretical Framework

2.1. *Generalized Linear Model*

Linear regression analysis is a statistical method aimed at identifying linear relationships between variables, involving a response variable (the affected variable) and predictor variables (the influencing factors). This method typically performs well when the response variable follows a normal distribution and data variability remains stable. Generalized Linear Models (GLMs), however, are designed for cases where the response variable does not conform to a normal distribution and data variability is constant [7]. GLMs extend linear regression by utilizing distributions from the exponential family, with the primary goal of estimating the response

variable based on insights provided by the predictor variables.

The observation variable Y, which follows an exponential family distribution, has the following probability density function [8]:

$$f(y|\theta,\phi) = \exp(\frac{y\theta - b(\theta)}{\phi} + c(y,\phi)), y \in S, \tag{2.1}$$

where y being a response variable, θ as the canonical parameter, ϕ as the scale parameter, and S being a subset of the set of natural numbers $\mathbb{N}$ or real numbers $\mathbb{R}$, where $b(\theta)$ and $c(y,\phi)$ are known functions. In the exponential family distribution, the following holds: $E(y) = b'(\theta)$ and $Var(y) = \phi b''(\theta)$.

GLM has the same objective as linear regression in general, which is to determine the conditional expectation of the response variable using existing observed data. In this case, parameters $\beta_1, \beta_2, \cdots, \beta_n$ will be determined through the link function of the mean values of explanatory variables (μ_i), which can be expressed as a linear combination of explanatory variables x_i as follows:

$$g(\mu_i) = \eta_i = \beta_0 + \sum_{j=1}^{n} \beta_j x_{ij} = x_i^T = \eta_i, i = 1, 2, \cdots, m.$$

The function that links the linear predictor η_i to the mean response μ_i is denoted by g, where g is a monotonic and differentiable function. In the framework of Generalized Linear Models (GLM), this function g is referred to as the link function. If the link function satisfies:

$$g(\mu_i) = \sum_{j=1}^{n} \beta_j x_{ij} = \theta_i,$$

then g is called the canonical link function.

2.2. *Pure Premium*

Pure premium is defined as the expected value of annual claim costs and is obtained by multiplying the expected claim frequency by the expected claim severity. Under the assumption that claim frequency and claim severity are independent, the expected aggregate claim amount, denoted by $\mathbb{E}(\sum_{i=1}^{N} C_i)$ can be expressed as follows:

$$\mathbb{E}(\sum_{i=1}^{N} C_i) = \mathbb{E}(N) \cdot \mathbb{E}(C_i), \tag{2.2}$$

the severity claim $(C_1, C_2, \cdots, C_N)$ are independent of the frequency claim (N).

2.3. *Gradient Boosting Machine*

The GBM algorithm focuses on the iterative use of weak learners to correct errors. In other words, GBM examines the errors produced by the previous weak learner and then constructs a new weak learner based on those error values. The weak learners to be used in this final project are decision trees. Please observe the GBM algorithm below.

(1) Input: Training data $\{(x_i, y_i)\}_{i=1}^{n}$ and *loss function* $\ell(y, \hat{f}(x))$.

(2) Define an initial model $\hat{f}_0(x) = \arg\min_{\rho} \sum_{i=1}^{n} \ell(y_i, \rho)$.

(3) For each iteration $m = 1, 2, \cdots, M$, the model is updated through the following steps:

   (a) the pseudo-residuals are computed as $r_{im} = -\left[\frac{\partial \ell(y_i, \hat{f}(x_i))}{\partial \hat{f}(x_i)}\right]_{\hat{f}(x)=\hat{f}_{m-1}(x)}, i = 1, 2, \cdots, n$.

   (b) *decision tree* for r_{im}, give a name to each *external node* as R_{jm} with an output of ρ_{jm} for $j = 1, 2, \cdots, J_m$.

   (c) For each terminal node R_{jm} an optimal update value ρ_{jm} is obtained by solving $j = 1, 2, \cdots, J_m$, calculate $\rho_{jm} = \arg\min_{\rho} \sum_{x_i \in R_{jm}} \ell(y_i, \hat{f}_{m-1}(x_i) + \rho)$.

   (d) Finally, the model is updated as $\hat{f}_m(x) = \hat{f}_{m-1}(x) + v \sum_{j=1}^{J_m} \rho_m I(x \in R_{jm})$.

(4) Output. $\hat{f}_M(x)$.

A decision tree is a weak learner that is prone to overfitting. Therefore, in GBM, there are hyperparameters that need to be adjusted to achieve the best results. The selection of hyperparameters is commonly done using the k-fold cross-validation method. Examples of some hyperparameters are Shrinkage (referred to as learning_rate in R), interaction.depth (referred to as max_depth in R), n.minobsinnode (referred to as min_samples_leaf in R), and n.trees.

2.4. *Machine Learning Interpretation*

A single decision tree is straightforward to interpret since it can be entirely represented and visualized in a two-dimensional graph. In contrast, ensemble models like GBM do not offer this transparency, making them appear as black-box models. To demystify and better understand the inner workings of models like GBM, various tools and methods are available. This section introduces the tools employed in this project to interpret the functioning of the GBM model.

(1) Variable Importance.
Variable importance was introduced by [9] and is a measure of how important explanatory variables are in predicting the response. For a specific explanatory variable $x_l, l \in \{1, 2, \cdots, p\}$, in the decision tree m, the variable importance is given by

$$I_l(m) = \sum_{j=1}^{J-1} I(v(j) = l)(\Delta L)_j. \tag{2.3}$$

To determine the significance of each variable, we calculate the total enhancement in the loss function ℓ across all internal nodes $J-1$ where variable x_l serves as the splitting criterion. Variables with higher relevance yield greater cumulative improvements in the loss function at these splits, distinguishing them from less impactful variables. To assess each variable's relative impact, we normalize

the importance scores so that their sum equals 100%. This approach can be extended to ensemble methods like GBM by averaging the importance score of variable x_l across all trees within the ensemble, as shown by:

$$I_l = \frac{1}{M} \sum_{m-1}^{M} I_l(m). \tag{2.4}$$

(2) Partial Dependence Plot
After identifying the most important variables, it is important to understand their influence on the response. Partial Dependence Plot (PDP) shows the marginal effect of a variable on predictions obtained from a model [10]. This means that we compute predictions for a particular variable x_l while averaging the values of other variables x_C according to:

$$\bar{f}_l(x_l) = \frac{1}{n} \sum_{i=1}^{n} f_{\text{mod } el}(x_l, x_{i,C}). \tag{2.5}$$

Here, C represents the complement set of l, ensuring that $l \cup C = 1, 2, \cdots, p$; $x_{i,C}$ denotes the values of other variables for observation i; and n indicates the total observations in the training dataset.

It is essential to recognize that PDP evaluates the influence of x_l on $f(x)$ after averaging the effects of the remaining variables x_c on $f(x)$. Consequently, potential interaction effects between x_l and other variables in x_c could obscure the isolated impact of x_l.

3. Result and Discussion

In this study, the data used are from one of the insurance companies in the United States. The data were obtained from the [11]. The data consists of information related to customers, with a total of 10,302 records. However, there are 1,663 censored data points. Censored data are observations obtained from incomplete observations, thus the analysis is only conducted on 8,639 data points. The variables analyzed in this study include the number of children of the customer, customer age, year of becoming a customer, monthly income, marital status (2 categories), gender (2 categories), education (5 categories), occupation (9 categories), duration of vehicle usage, vehicle usage (2 categories), legal bluebook levy, vehicle age, red car (2 categories), and vehicle type (6 categories).

3.1. *Generalized Linear Model*

3.1.1. *Frequency Claim*

The frequency claim data is suspected to follow one of the discrete distributions, however, after fitting the discrete distribution using the Anderson-Darling (A-D) test, the data does not follow any discrete distribution. Based on this, fitting was done to the Tweedie mixture distribution, using the "tweedie.profile" function in the R software. After testing the distribution of the claim frequency data, a p-value of 1.132653 was obtained. This indicates that the p-value is between 1 and 2, or it can

be said that the response variable of claim frequency follows a Tweedie distribution. Next, the determination of the best link function was conducted, where the best link function is the one with the smallest AIC value. Based on the test conducted, it was found that the log link function is the best link function with an AIC value of 16312.66. By using the log link function, the claim frequency model is obtained as follows:

$$g(\mu_i) = x_i^T\beta \leftrightarrow \ln\ \mu_i = x_i^T\beta \leftrightarrow \mu_i = \exp(x_i^T\beta). \tag{3.1}$$

The goodness of fit of the claim frequency model in Eq. (3.1) will be tested. The tests used to assess the goodness of fit of the model are the partial likelihood ratio test and the Wald test.

Table 1. The estimation of parameters for the claim frequency model

Source	β	*Std. Error*	*p-value*
Intercept	0.0519524	0.1494263	0.728092
KidSdriv	0.1090512	0.0323352	0.000749
Age	-0.0047564	0.0021432	0.026498
Income	-0.0010793	0.0004247	0.011482
MSTATUS = single	0.2029693	0.0355616	1.19e-08
Education = Magister	-0.3007894	0.1057419	0.004460
Occupation = Doctor	0.3232653	0.1496905	0.030841
Occupation = Lawyer	0.2241284	0.1102276	0.042058
Occupation = *Professional*	0.1913213	0.0738336	0.009583
Occupation = etc	0.4264444	0.1143232	0.000193
car use = personal	-0.2350002	0.0563800	3.11e-05
Bluebook	-0.0077682	0.0030025	0.009695
Car Type = *Sports car*	0.4132317	0.0753173	4.24e-08
Car Type = SUV	0.2765348	0.0650489	2.15e-05

From the results of the partial tests presented in Table 1, the claim frequency model is obtained as follows:

$$\begin{aligned} E[N_i] = \mu_f = \exp(&\beta_0 + \beta_1 x_1^i + \beta_2 x_2^i + \beta_4 x_4^i + \beta_5 x_5^i + \beta_7 x_7^i \\ &+ \sum_{j=1}^{9} \beta_{8j} x_{8j}^i + \beta_{10} x_{10}^i + \beta_{11} x_{11}^i + \sum_{j=1}^{6} \beta_{12j} x_{12j}^i). \end{aligned} \tag{3.2}$$

3.1.2. *Severity Claim*

Similar to claim frequency, the claim severity data is suspected to follow one of the continuous distributions. However, after fitting the continuous distribution using the chi-square test, the data does not follow any continuous distribution. Therefore, fitting was done to the Tweedie mixture distribution using the "tweedie.profile" function in the R software. After testing the distribution of the claim severity data,

a p-value of 1.757143 was obtained. This indicates that the response variable of claim severity follows a Tweedie distribution.

Next, the determination of the best link function was conducted, where the best link function is the one with the smallest AIC value. Based on the test conducted, it was found that the log link function is the best link function with an AIC value of 40390.79. By using the log link function, the claim severity model is obtained as follows:

$$g(\mu_i) = x_i^T\gamma \leftrightarrow \ln\ \mu_i = x_i^T\gamma \leftrightarrow \mu_i = \exp(x_i^T\gamma). \tag{3.3}$$

The fit quality of the claim severity model, as represented in equation 3.3, will be evaluated. To determine the model's adequacy, both the partial likelihood ratio test and the Wald test will be applied.

Table 2. The estimation of parameters for the claim severity model

Source	γ	*Std. Error*	*p-value*
Intercept	3.7648081	0.5313250	1.52e-12
Kidsdriv	0.2842356	0.1163967	0.014633
MStatus = Single	0.3163777	0.1273632	0.013013
Bluebook	-0.0387866	0.0107378	0.000306
Car Type = *Panel Truck*	0.6940844	0.3443632	0.043885
Car Type = *Pickup*	0.5198037	0.2071392	0.012115
Car Type = *Sports car*	0.5746832	0.2688504	0.032588
Car Type = Van	0.7248252	0.2655623	0.006361
Car Type = SUV	0.5283112	0.2260659	0.019469

From the results of the partial tests presented in Table 2, the claim severity model is obtained as follows.

$$E[C_i] = \mu_c = \exp(\gamma_0 + \gamma_1 x_1^i + \gamma_5 x_5^i + \gamma_{11} x_{11}^i + \sum_{j=1}^{6} \gamma_{12j} x_{12j}^i). \tag{3.4}$$

3.2. *Gradient Boosting Machine*

The package gbm() will be used to apply the GBM method to the data. First, hyperparameters such as shrinkage, interaction depth, and minobsinode will be determined using cross-validation. After determining these three hyperparameters, the package gbm() will determine the hyperparameter n.trees by reviewing the train error and test error.

3.2.1. *Frequency Claim*

We will use 10-fold cross-validation to determine the best shrinkage value. Several candidate values for shrinkage will be considered, including $0.01, 0.05, 0.1, 0.3, 0.5$, and 1.

Table 3. *Cross Validation* for *Hyperparameter Shrinkage*

Shrinkage	RMSE	*n.trees*
0.01	1.151192	1164
0.05	1.150657	202
0.10	1.151096	222
0.30	1.150924	45
0.50	1.152055	13
1.00	1.155289	20

Based on Table 3, it can be seen that the shrinkage value that provides the smallest RMSE is 0.05. Therefore, this value will be used in the algorithm.

The next step is to determine the hyperparameters interaction depth and minobsinnode. These two hyperparameters are often determined together. They will act as an early stopping rule on the decision tree built to prevent overfitting. Below are the sets of values to be used as candidate values for interaction depth and minobsinnode.

interaction.depth: $\{1, 2, 3, 4, 5, 6, 7, 8, 9, 10\}$

minobsinnode: $\{16, 17, 18, 19, 20, 21, 22, 23, 24, 25\}$

interaction.depth \ minobsinnode	16	17	18	19	20	21	22	23	24	25
1	1,149815	1,150325	1,148846	1,149666	1,149903	1,150087	1,150023	1,150401	1,149694	1,151145
2	1,149310	1,149727	1,151557	1,150263	1,151253	1,149746	1,150086	1,150714	1,150298	1,150163
3	1,150804	1,152129	1,150952	1,149026	1,150611	1,151442	1,149895	1,149318	1,150045	1,150317
4	1,151692	1,152184	1,151760	1,151166	1,150325	1,150086	1,150632	1,151340	1,150372	1,151196
5	1,152256	1,152067	1,150469	1,150208	1,150852	1,150943	1,150167	1,150141	1,150100	1,152450
6	1,151404	1,150444	1,150466	1,150233	1,150680	1,150657	1,151055	1,149606	1,151680	1,151951
7	1,150968	1,151535	1,149985	1,151455	1,151729	1,150795	1,151195	1,151364	1,152890	1,151200
8	1,151388	1,151450	1,152769	1,151675	1,151171	1,150537	1,151926	1,149775	1,151544	1,150966
9	1,151566	1,152373	1,150427	1,151212	1,152785	1,151348	1,152273	1,151751	1,151528	1,151929
10	1,151460	1,152238	1,151802	1,151470	1,153751	1,151453	1,150882	1,149987	1,151878	1,151320

Figure 1. *Cross Validation* for *hyperparameter interaction.depth* dan *minobsinnode*

We will use the combination that gives the smallest RMSE value, which is when the interaction depth is 1 and minobsinnode is 18. This combination yields an RMSE value of 1.148846.

The hyperparameters shrinkage, interaction depth, and minobsinnode have been determined. The final step is to determine the hyperparameter n.trees. Note that in Figure 2, the train error starts to increase after entering the 201st iteration. Therefore, it is decided to use 201 iterations in the final model built, which is the number of iterations just before overfitting occurs.

The hyperparameters to be used for the GBM algorithm are summarized in Table 4.

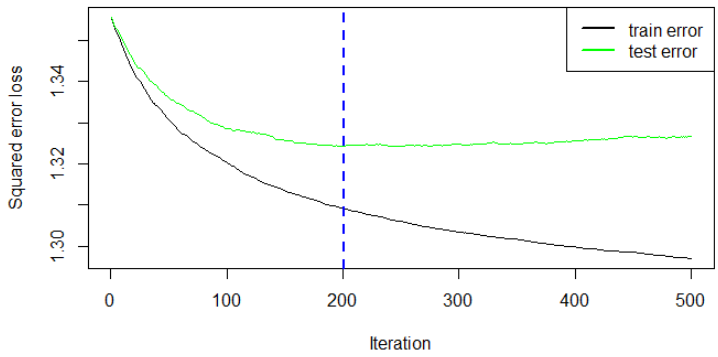

Figure 2. *Train Error* and *Test Error* for GBM

Table 4. *Hyperparameter* for GBM

shrinkage	*interaction.depth*	*minobsinnode*	*n.trees*
0.05	1	18	201

3.2.2. *Severity Claim*

We will use 10-fold cross-validation to determine the best shrinkage value. Several candidate values for shrinkage will be considered, including $0.01, 0.05, 0.1, 0.3, 0.5$, and 1.

Table 5. *Cross Validation* for *Hyperparameter Shrinkage*

Shrinkage	RMSE	*n.trees*
0.01	119.0160	246
0.05	119.0646	55
0.10	119.0799	29
0.30	119.0706	3
0.50	119.1124	2
1.00	119.2548	3

Based on Table 5, it can be seen that the shrinkage value that provides the smallest RMSE is 0.01. Therefore, this value will be used in the algorithm.

The next step is to determine the hyperparameters interaction depth and minobsinnode. These two hyperparameters are often determined together. They will act as an early stopping rule on the decision tree built to prevent overfitting. Below are the sets of values to be used as candidate values for interaction depth and minobsinnode.

interaction.depth: $\{1, 2, 3, 4, 5, 6, 7, 8, 9, 10\}$
minobsinnode: $\{16, 17, 18, 19, 20, 21, 22, 23, 24, 25\}$

interaction.depth \ minobsinnode	16	17	18	19	20	21	22	23	24	25
1	119,0149	118,9800	119,0687	119,0460	119,0966	119,0487	119,0477	119,0838	119,0673	119,0492
2	119,1053	119,1315	119,0992	119,0947	119,0526	119,1328	119,1432	119,0864	119,1004	119,1426
3	119,1880	119,1798	119,1675	119,1544	119,1564	119,1874	119,1652	119,1350	119,1624	119,1542
4	119,1621	119,1777	119,1027	119,1945	119,1982	119,1380	119,1671	119,1218	119,2238	119,1505
5	119,1457	119,1570	119,1981	119,1828	119,1192	119,1440	119,2023	119,1349	119,1189	119,1538
6	119,1370	119,1700	119,1468	119,1274	119,1887	119,1828	119,1590	119,1494	119,1992	119,1519
7	119,1727	119,2009	119,1879	119,1832	119,1834	119,1693	119,1774	119,1359	119,1865	119,1274
8	119,1028	119,1996	119,1844	119,2072	119,1684	119,1356	119,1212	119,1433	119,1492	119,1882
9	119,1398	119,1836	119,1861	119,1825	119,1654	119,1791	119,1434	119,1723	119,1523	119,2074
10	119,2227	119,1868	119,1562	119,2410	119,2117	119,1583	119,1886	119,1687	119,1486	119,2232

Figure 3. *Cross Validation* for *hyperparameter interaction.depth* and *minobsinnode*

We will use the combination that gives the smallest RMSE value, which is when the interaction depth is 1 and minobsinnode is 17. This combination yields an RMSE value of 118.9800.

The hyperparameters shrinkage, interaction depth, and minobsinnode have been determined. The final step is to determine the hyperparameter n.trees. Note that in Figure 4, the train error starts to increase after entering the 256th iteration. Therefore, it is decided to use 256 iterations in the final model built, which is the number of iterations just before overfitting occurs.

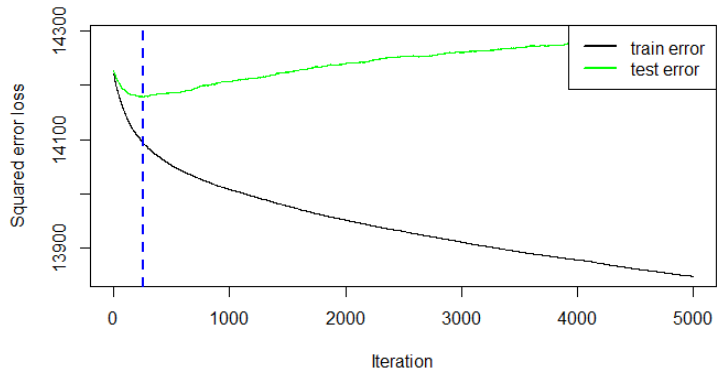

Figure 4. *Train Error* and *Test Error* untuk GBM

The hyperparameters to be used for the GBM algorithm are summarized in Table 6.

Table 6. *Hyperparameter* for GBM

shrinkage	*interaction.depth*	*minobsinnode*	*n.trees*
0.01	1	17	256

3.3. *Model Evaluation*

Please refer to Table 7 to see the comparison of RMSE generated from the GLM and GBM methods. The RMSE values will be used to compare the accuracy of the two methods.

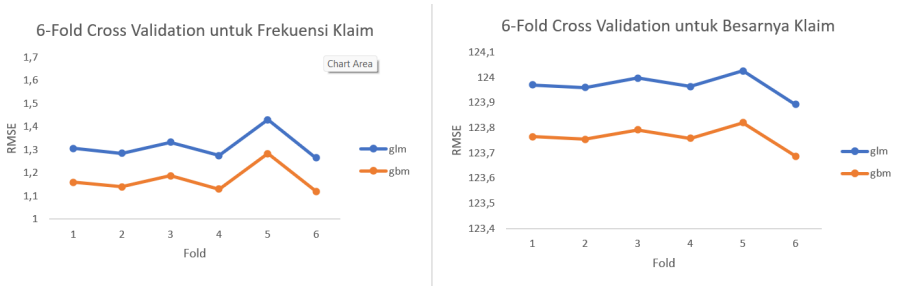

Figure 5. Comparison of RMSE using 6-Fold Cross Validation for Claim Frequency and Claim Severity

In Figure 5 the RMSE generated by both methods in claim frequency and claim severity modeling for each fold can be observed. It can be seen that the RMSE values for the GLM method are larger than the RMSE values for the GBM method. This indicates that, for the data under consideration, the GBM method is better suited than the GLM method for predicting claim frequency and claim severity.

Next, we will compare the average RMSE generated using 6-Fold Cross Validation for each claim frequency and claim severity modeling.

Table 7. Comparison of Average RMSE

Modeling	GLM	GBM
Frequency Claim	1,26578	1,11990
Severity Claim	123,8225	123,6169

From Table 7 it can be seen that the Gradient Boosting Machine (GBM) algorithm is able to produce models with the lowest RMSE, both for claim frequency and claim severity modeling. Therefore, it can be said that the Gradient Boosting Machine (GBM) algorithm is more suitable for modeling claim frequency and claim severity compared to the Generalized Linear Model (GLM) method.

3.4. *Interpretation Best Model*

Gradient Boosting Machine (GBM) has a feature that can be used to identify which variables contribute most significantly. This feature is known as variable importance. Variable importance quantifies the influence of a variable by calculating how often the variable is used as a splitting rule. Figure 6 displays the variable importance for Claim Frequency and claim severity Modeling with GBM. Based on the figure on the right-hand side, it can be seen that the customer's age (AGE), the type of vehicle (CAR-TYPE), and the customer's monthly income (INCOME) are the variables that have the most significant influence. Based on the figure on the left-hand side, it can be seen that the bluebook variable, customer's occupation (OCCUPATION), customer's age (AGE), and customer's vehicle type (CAR-TYPE) are the variables that have the most significant influence.

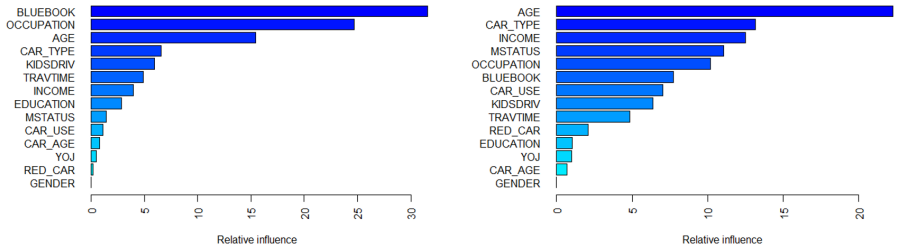

Figure 6. *Variable Importance* for Frequency Claim (right) and Severity Claim (Left) with GBM

We can see the marginal effect of customer age on claim frequency from the machine learning model in Figure 7 on the right-hand side. The risk of filing claims is high for policyholders in early adulthood and gradually decreases with age until stabilizing around the age of 35. The risk starts to decrease again at around the age of 40 and increases in elderly policyholders, around the age of 55. The early adulthood age group has a higher accident rate. This factor is often associated with a lack of driving experience and more risky driving behavior. Meanwhile, elderly policyholders are often more cautious in driving, but health factors and decreased physical response can affect the risk of accidents.

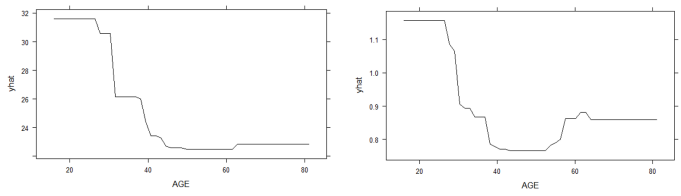

Figure 7. *Partial Dependence Plots* for Frequency Claim and Severity Claim against age customer

Based on the figure 7 on the left-hand side, we can see the marginal effect of

customer age on claim severity from the machine learning model. Early adulthood policyholders tend to have higher claim severity. This may be due to early adulthood policyholders often being involved in accidents with significant losses including repair or replacement costs for vehicles. The comparison of claim severity and customer age can be seen that the relationship between the two variables is somewhat linear.

In Figure 8 we can see the marginal effect of customer vehicle type on claim frequency (right) and claim severity (left) from the machine learning model. Pickup trucks and minivans have high claim frequencies, while sports cars and pickup trucks rank high in claim severity. This may be because some types of vehicles have higher safety levels than others. Sports cars and other luxury cars usually have higher premiums due to their high claim frequency and the expensive cost of replacement parts. SUVs have the lowest claim frequency and claim severity because this type of vehicle has performance and safety features with varying prices so that the insurance rates for SUVs can vary depending on various factors such as safety features, size, and reputation of certain models.

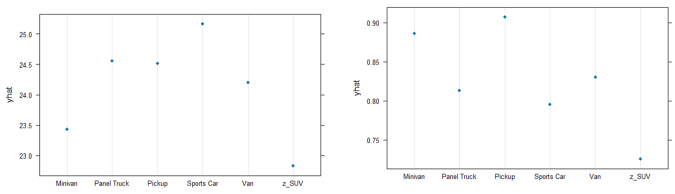

Figure 8. *Partial Dependence Plots* for frequency claim and severity claim against car type

In Figure 9 we can see the marginal effect of customer income on claim frequency (left) and claim severity (right) from the machine learning model. The higher the customer's income, the lower the claim frequency tends to be. This could indicate that people with higher incomes may be more cautious in obtaining insurance claims or may have access to better services that prevent claims. On the other hand, in terms of claim severity, we can see a pattern where claim severity tends to decrease and stabilize with the increase in customer income. This may suggest that people with higher incomes tend to have lower claim amounts.

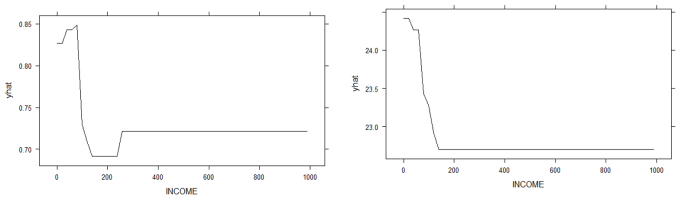

Figure 9. *Partial Dependence Plots* for frequency claim and severity claim against customer income

In Figure 10 we can see the marginal effect of customer age and customer vehicle type on claim frequency (right) and claim severity (left) from the machine learning model. It can be observed that for all customer vehicle types, the higher the customer's age, the lower the claim frequency and claim severity. However, in comparing claim severity, it can be seen that for all customer vehicle types, the relationship between customer age and claim severity tends to be more stable. On the other hand, interactions between customer age and customer vehicle type can also be observed. In claim frequency modeling, the interaction between customer age and customer vehicle type is not very significant. It can be seen that the customer vehicle types with the highest to lowest claim frequencies for all age points are pickup trucks, minivans, vans, panel trucks, sports cars, and SUVs. The patterns for each vehicle type are also very similar. Claim severity for each vehicle type varies in order for each customer age. Therefore, it can be said that there is an interaction between customer age and customer vehicle type.

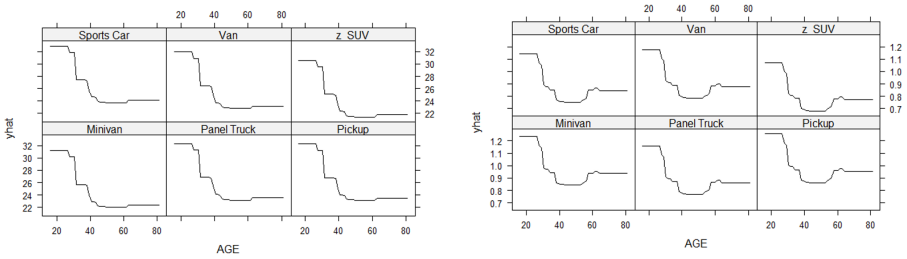

Figure 10. *Partial Dependence Plots* for frequency claim and severity claim against age and car type

Furthermore, the pure premiums of motor vehicle insurance can be determined using the Generalized Linear Model (GLM) and Gradient Boosting Machine (GBM). The policies whose premiums are to be determined originate from the test set. By employing GLM, the expected claim frequency and claim severity can be determined. Table 8 – Table 10 give randomly selected profile from the test set.

Table 8. Single profile Pure Premium

ID	Number of Children	Age	YOJ	Income	Marital Status	Gender	Education
493243717	0	49	8	$39,49	Not Married	Female	High School

Table 9. Single profile Pure Premium

ID	Occupation	Travtime	Car User	Bluebook	Car Type	Cary Age
493243717	Professional	61	Private	$5,450	SUV	1

Table 10. Single profile Pure Premium

ID	Frequency Pred.	Severity Pred.	Pure Premium
493243717	1.06331	27.98908	29.76114

In the same way, the pure premium values for motor vehicles can be obtained using combinations of the aforementioned categories.

Table 11. The Premium Rates for GLM

ID	Frequency Pred.	Severity Pred.	Pure Premium
196357678	1.28846	42.82746	55.18150
803846001	0,82219	31,40194	25,81835
798306776	0,67525	23,11107	15,60565
493274156	0,60405	11,65175	7,03830
421896035	0,77836	22,87542	17,80532

Table 11 presents the motor vehicle premium values for several profiles using the GLM model. The determination of motor vehicle premiums using GBM follows the same steps as in the GLM scenario. For example, let us consider Table 8 as the randomly selected profile from the test set. The profile just described was observed for only 3 months, so we can use the GLM predictions to determine the expected frequency and severity for that exposure period. However, in practice, premiums are updated or subscribed to annually, making it inappropriate to calculate premiums assuming an exposure value of 1. This is why the Annual Pure Premium is also calculated and included in the last column.

Determining motor vehicle premiums using GBM follows the same steps as in the GLM scenario. For example, consider a random profile selected from the test set as shown in Table 8, Table 9, and Table 10.

Table 12. Single profile Pure Premium for GLM & GBM

ID	Method	Frequency Pred.	Severity Pred.	Pure Premium	Annual Pure Premium
493243717	GLM	1,06331	27,98908	29,76114	119,04460
493243717	GBM	0,82162	32,39270	26,61443	106,45770

In the example from Table 12 it can be seen that the GLM and GBM models produce distinctly different predictions. In terms of frequency, the GLM yields higher predicted values, but for claim severity, the GBM model produces the highest predictions. However, the pure premium calculated is higher for the GLM model. In the same way, the pure premium values for motor vehicles can be obtained using combinations of the aforementioned categories.

Table 13 presents the motor vehicle premium values for several profiles using the GBM model. The profile just described was observed for only 3 months, so we

Table 13. The Premium Rates for GLM

ID	Frequency Pred.	Severity Pred.	Pure Premium	Annual Pure Premium
196357678	1,19318	32,29437	38,53307	154,13226
803846001	0,93637	24,790094	23,21270	92,85079
798306776	0,78555	20,16925	15,84399	63,37598
493274156	0,80321	20,59998	16,54614	66,18457
421896035	0,78937	23,78631	18,77618	75,10471

can use the GBM predictions to determine the expected frequency and severity for that exposure period. However, in practice, premiums are updated or subscribed to annually, making it inappropriate to calculate premiums assuming an exposure value of 1. This is why the Annual Pure Premium is also calculated and included in the last column. When a particular risk class or group of policies with very similar characteristics consistently shows higher (or lower) claim frequency and/or higher (or lower) claim severity, the model will provide higher (or lower) predictions in terms of frequency and/or severity. The final result will yield premium values that accurately represent and adjust to the true nature of the risk for each profile.

4. Conclusion

In our study we have shown that by using GBM models by tuning of the model hyperparameters, we can create highly competitive models of claim frequency and claim severity in a non-life insurance pricing setting. Given that we do not know about the interactions of our true frequency and severity function, we have also shown that the GBM outperforms the GLM in terms of both prediction accuracy and ranking of the claim frequency and claim severity risk. The results of this simulation study confirms other studies in that GBM models have the advantage compared to GLMs that they can automatically detect and model non-linear effects and interaction effects between explanatory variables. For GLMs, such non-linear effects or interaction effects need to be manually investigated and included in the modeling which often is a time consuming effort. On the other hand, we saw that GLMs are relatively stable to fit whereas the GBMs are more prone to overfitting the data. Another obvious advantage of GLMs are their immediate transparency, since the parameter estimates of a GLM directly shows the effect of a variable on the response. The application of variable importance and partial dependence plots offers an effective way to interpret ML algorithms like GBM, providing transparency and explaining the model's operations. Furthermore, GBM guided us in the right direction to develop GLMs with enhanced model performance.

5. Acknowledgment

Special thanks are extended to all those who took part in this research, as this work would not have been possible without their support.

Bibliography

[1] Nelder, J. A., Verrall, R. J., 1997, Credibility Theory and Generalized Linear Models, *ASTIN Bulletin* Vol. **27**(1): 71 – 82

[2] Ohlsson, E., Johanson, B., 2010, *Non-Life Insurance Pricing with Generalized Linear Models*, Springer, Berlin

[3] Rosenlund, S., 2013, *Integrating Ordinary GLM with Credibility in a Compound Poisson Model*, Stockholm, Swedia

[4] Garrido J, Genest C, Schulz J, 2016, Generalized Linear Models for Dependent Frequency and Severity of Insurance Claims, *Insur. Math. Econ.* Vol. **70**: 205 – 215

[5] Li N, Peng X, Kawaguchi E, Suchard MA, Li G, 2020, A Scalable Surrogate L0 Sparse Regression Method for Generalized Linear Models with Applications to Large Scale Data, *J. Stat. Plan. Inference* Vol. **213**: 262 – 281

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

[7] Agresti A., 2015, *Foundations Linear Generalized Linear Models*, Cambridge: John Wiley & Sons, Inc.

[8] Jong, P.D., Heller G.Z., 2008, *Generalized Linear Models for Insurance Data*, Cambridge: Cambridge University Press

[9] Breiman, L., Friedman, J., Stone, C. J., Olshen, R. A, 1984, *Classification and regression trees*, CRC Press

[10] Hastie, T., Tibshirani, R., Friedman, J., 2009, The elements of statistical learning: Data mining, inference, and prediction, Springer Science & Business Media

[11] "Car Insurance Claim (Cleaned)," Data.World, 2025, [Online], Available: https://data.world/saleem786/car-insurance-claims-analysis/workspace/file?filename=Car+Insurance+Claim+(Cleaned).xlsx. [Accessed: 08-02-2023].

[12] Tweedie, M.C.K., 1984, An Index Which Distinguishes between Some Important Exponential Families, in: Ghosh, J.K. and Roy, J., Eds., Statistics: Applications and New Directions, Proceedings of the Indian Statistical Institute Golden Jubilee International Conference, Indian Statistical Institute, Calcutta: 579 – 604