risks

MDPI

Article

A Generalized Linear Model and Machine Learning Approach for Predicting the Frequency and Severity of Cargo Insurance in Thailand's Border Trade Context

Praiya Panjee and Sataporn Amornsawadwatana *

School of Engineering, University of the Thai Chamber of Commerce, Bangkok 10400, Thailand; 1810751101006@live4.utcc.ac.th
* Correspondence: sataporn_amo@utcc.ac.th

Abstract: The study compares model approaches in predictive modeling for claim frequency and severity within the cross-border cargo insurance domain. The aim is to identify the optimal model approach between generalized linear models (GLMs) and advanced machine learning techniques. Evaluations focus on mean absolute error (MAE) and root mean squared error (RMSE) metrics to comprehensively assess predictive performance. For frequency prediction, extreme gradient boosting (XGBoost) demonstrates the lowest MAE, indicating higher accuracy compared to gradient boosting machines (GBMs) and a generalized linear model (Poisson). Despite XGBoost's lower MAE, it shows higher RMSE values, suggesting a broader error spread and larger magnitudes compared to gradient boosting machines (GBMs) and a generalized linear model (Poisson). Conversely, the generalized linear model (Poisson) showcases the best RMSE values, indicating tighter clustering and smaller error magnitudes, despite a slightly higher MAE. For severity prediction, extreme gradient boosting (XGBoost) displays the lowest MAE, implying better accuracy. However, it exhibits a higher RMSE, indicating wider error dispersion compared to a generalized linear model (Gamma). In contrast, a generalized linear model (Gamma) demonstrates the lowest RMSE, portraying tighter clustering and smaller error magnitudes despite a higher MAE. In conclusion, extreme gradient boosting (XGBoost) stands out in mean absolute error (MAE) for both frequency and severity prediction, showcasing superior accuracy. However, a generalized linear model (Gamma) offers a balance between accuracy and error magnitude, and its performance outperforms extreme gradient boosting (XGBoost) and gradient boosting machines (GBMs) in terms of RMSE metrics, with a slightly higher MAE. These findings empower insurance companies to enhance risk assessment processes, set suitable premiums, manage reserves, and accurately forecast claim occurrences, contributing to competitive pricing for clients while ensuring profitability. For cross-border trade entities, such as trucking companies and cargo owners, these insights aid in improved risk management and potential cost savings by enabling more reasonable insurance premiums based on accurate predictive claims from insurance companies.

Keywords: machine learning; prediction model; cargo insurance; generalized linear model; gradient boosting; extreme gradient boosting

Citation: Panjee, Praiya, and Sataporn Amornsawadwatana. 2024. A Generalized Linear Model and Machine Learning Approach for Predicting the Frequency and Severity of Cargo Insurance in Thailand's Border Trade Context. *Risks* 12: 25. https://doi.org/10.3390/risks12020025

Academic Editor: Shengkun Xie

Received: 28 December 2023
Revised: 21 January 2024
Accepted: 25 January 2024
Published: 30 January 2024

1. Introduction

As of 2020, road transport constituted a significant 79.7% share of all goods movement within Thailand. This dominance in transportation modes results from successive Thai government policies that have emphasized prioritizing the development of the road system over alternative transport networks. Consequently, roads currently account for a staggering 91.6% of the total distance covered by various transportation methods within the country. The deliberate focus on road infrastructure has notably favored road haulage as the primary means of transport. Its distinct advantage lies in offering door-to-door transportation services, enabling shippers to seamlessly move goods from their source or origin directly to the recipient in a single stage. This convenience and efficiency in door-to-door transit have

significantly contributed to the prevalence of road transport for moving goods throughout Thailand and border trade (Krungsri Research 2022).

Thailand's road freight industry boasts significant advantages in the realm of cross-border transportation. Situated strategically in Southeast Asia, recent trade patterns have revealed a notable surge in connectivity, particularly concerning import and export border trade between Thailand and its key mainland Southeast Asian neighbors: Cambodia, Laos, Malaysia, and Myanmar (Chirathivat and Cheewatrakoolpong 2015). This trend underscores a significant evolution in cross-border trade relationships within the region. Thailand has experienced an upswing in trade activities, emphasizing enhanced connectivity and collaboration with these neighboring countries. Furthermore, Thailand's dedication to enhancing connectivity is exemplified by its ongoing infrastructure development endeavors. Initiatives like the enhancement of road networks play a pivotal role in fortifying cross-border trade and fostering increased opportunities for investment (Deputy Prime Minister and Minister of Foreign Affairs of Thailand 2023).

However, amid the success and extensive reliance on road transport, there is a significant concern: the vulnerability of cargo during transportation, particularly in cross-border operations. Despite the robustness of Thailand's road freight industry, the risk of cargo damage remains a pressing issue. Historical data spanning from 2016 to 2022 underscore this concern, showcasing consistent evidence of cargo damage risks, especially in cross-border transportation, as shown in the below table.

The data in Table 1 cover the period from 2016 to 2022, showcasing the claims data within cargo insurance for road freight pertaining to import/export activities between Thailand and Myanmar, Laos, Cambodia, and Malaysia. Sourced from the Insurance Premium Rating Bureau in Thailand, this comprehensive record provides historical insights into the diverse array of reasons causing cargo insurance claims in road freight.

Table 1. Causes of claims from 2016 to 2022.

Claim Cause Category	Loss Year						
	2016	**2017**	**2018**	**2019**	**2020**	**2021**	**2022**
Breakage	×	×	×	×	×	×	×
Bend/Dent/Scratch	×	×	×	×	×	×	×
Tear/Cut	×	×	×	×	×	×	×
Rainwater Damage		×	×		×	×	×
Overturning		×	×		×		
Shortage/Leakage			×		×		
Contamination			×	×	×	×	
Theft/Pilferage/Missing			×	×		×	
Stain			×			×	
Rust/Oxidation/Corrosion					×		
Others	×	×	×	×	×	×	×

Source: The Insurance Premium Rating Bureau (Thailand).

It encompasses various causes such as breakage, bending/denting/scratching, tearing/cutting, rainwater damage, overturning, shortage/leakage, contamination, theft/pilferage/missing, staining, rust/oxidation/corrosion, and an 'others' category. Each classification represents distinctive challenges and risks encountered during road freight transit. Spanning from 2016 to 2022, this expansive dataset shows the common risks and incidents impacting cargo shipments by road freight. It offers crucial insights to fortify strategies aimed at safeguarding shipments and mitigating financial liabilities from various threats.

This detailed understanding directly correlates with Thailand's notable surge in road freight cargo insurance participation during the same period. The increase in net premium

amounts paid from 2016 to 2022 signifies an industry-wide recognition of these risks, leading to a proactive approach to protecting cargo during transit. Furthermore, the alignment between the dataset's insights and the heightened engagement in cargo insurance underscores a collective effort within Thailand's road freight industry. This concerted action aims to mitigate risks and fortify protection measures for cargo in transit.

Thailand is experiencing a significant increase in road freight cargo insurance participation, as indicated by the net premium amounts paid from 2016 to 2022, as shown in Figures 1 and 2.

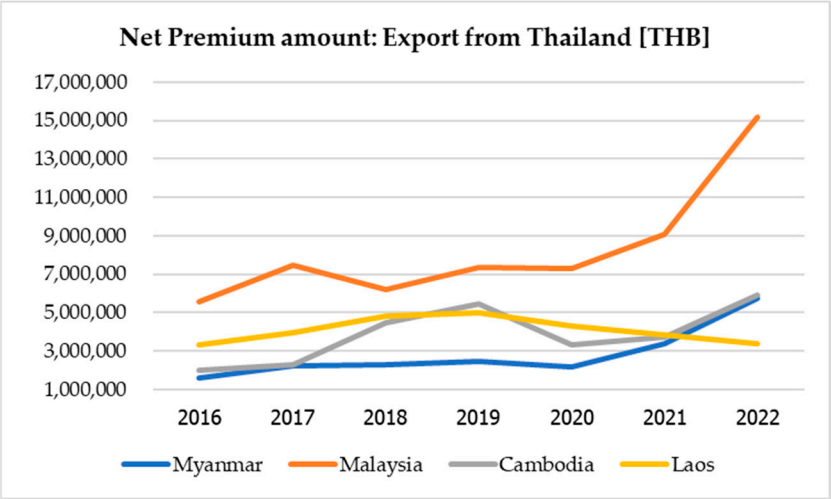

Figure 1. Net Premium Amount Data for Road Freight Exports from Thailand from 2016–2022. Source: The Insurance Premium Rating Bureau (Thailand).

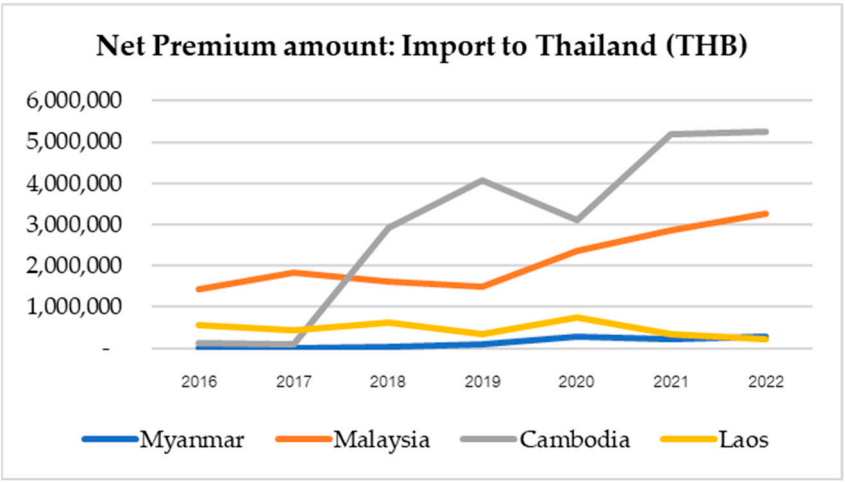

Figure 2. Net Premium Amount Data for Road Freight Imports to Thailand from 2016–2022. Source: The Insurance Premium Rating Bureau (Thailand).

Figure 1 depicts the trend of net premium amounts paid from 2016 to 2022. It is evident that the trend lines of destination countries such as Malaysia, Myanmar, and Cambodia show a clear increase over this period. However, in contrast, Laos exhibits a decreasing trend during the same timeframe.

Figure 2, which delineates the importation data, shows the trends of origin countries such as Malaysia, Myanmar, and Cambodia.

In Figure 2, the trends in net premium amounts paid between 2016 and 2022 are depicted. The trend lines from the origin countries of Malaysia, Myanmar, and Cambodia demonstrate a noticeable increase over this period. Conversely, Laos displays a declining trend during the same timeframe.

The trends in net premium amounts paid, as depicted in Figures 1 and 2, further emphasize this connection. The increasing trend lines for cargo exportation to countries like Malaysia, Myanmar, and Cambodia, as well as the rising trends in importation from these countries, demonstrate a growing awareness of the risks associated with road freight. Conversely, the declining trend in Laos indicates potential areas where further attention might be needed to address vulnerabilities in cargo transit.

The dataset covering cargo insurance claims from 2016 to 2022 for road freight between Thailand and neighboring countries like Myanmar, Laos, Cambodia, and Malaysia provides valuable insights into the risks faced during transit. It highlights various challenges, from breakage and theft to contamination and weather-related damages, offering a comprehensive view of vulnerabilities impacting cargo shipments.

The correlation between these data and the rise in net premium amounts paid for cargo insurance in Thailand is significant. The increase in premiums reflects the industry's acknowledgment of these risks. It signifies a proactive shift towards safeguarding cargo during transportation. Essentially, the insights drawn from the dataset directly influenced the surge in cargo insurance participation, indicating a collective effort within the industry to mitigate risks and protect shipments.

Cargo insurance plays a pivotal role in mitigating financial liabilities arising from diverse perils encountered during transit. The alignment between the insights gleaned from the dataset and the increased engagement in cargo insurance underlines the strategic response of the industry to fortify protection measures. Ultimately, this trend signifies a proactive approach to safeguarding shipments and minimizing the financial impact of the risks inherent in road freight transportation. Cargo insurance in cross-border transportation serves as a critical safeguard against a spectrum of uncontrollable risks that threaten goods during transit. Despite meticulous packaging and handling, unforeseen incidents like accidents, theft, breakage, and environmental damage can jeopardize cargo integrity. This insurance covers a myriad of potential risks, including theft, pilferage, breakage, rainwater damage, shortages, contamination, and more. And despite the abundance of research on supply chain risk management, there remains a notable gap in addressing the risk associated with cargo accumulation. This aspect has received limited attention in existing studies, despite past events demonstrating its potential for significant damage (Freichel et al. 2022). The core aim of insurance or coverage revolves around shifting risk and providing compensation in case of loss or damage (Ritonga et al. 2021). It acts as a financial shield, alleviating the burden of lost or damaged goods and ensuring stability for both shippers, consignees, and the trucking company. Lacking suitable cargo insurance coverage can lead to dire consequences for exporters/importers or businesses, resulting in devastating outcomes (Socorro and Karina 2019). Additionally, in many instances, cargo insurance is not just a best practice; it is a legal or contractual necessity, ensuring compliance with regulations and agreements.

The correlation between the insights derived from cargo insurance data and the increased engagement in fortifying protection measures through insurance aligns with the evolving landscape of predictive modeling in the insurance domain, particularly in Thailand. The traditional use of generalized linear models (GLMs) for actuarial purposes has been the foundation for risk assessment and premium calculations (Thai General Insurance Association 2016). In recent years, there has been a notable increase in research that compares generalized linear models (GLMs) with advanced machine learning (ML) approaches, including extreme gradient boosting (XGBoost) and gradient boosting. These studies have demonstrated the superior performance of ML models across many domains, as shown in Table 2.

In essence, the increased reliance on ML models reflects an industry-wide response to leverage more robust and versatile tools for risk evaluation and prediction. This parallels the proactive approach observed in the cargo insurance realm, where the insights gleaned from historical data drive a strategic response to fortify protection measures.

Table 2. Comprehensive reviews of machine learning in non-life insurance actuarial science.

Article	Year	Methodologies/Approaches	The Best Model
Modelling Motor Insurance Claim Frequency and Severity Using Gradient Boosting (Clemente et al. 2023)	2023	Gradient Boosting Machines, Generalized Linear Models	Gradient Boosting Machines [Frequency] and Generalized Linear Models [Severity]
Machine Learning in Forecasting Motor Insurance Claims (Poufinas et al. 2023)	2023	Support Vector Machines, Decision Trees, Random Forests, Extreme Gradient Boosting	Random Forests Limited Depth and Extreme Gradient Boosting
Predicting Motor Insurance Claims—XGBoost versus Logistic Regression (Murekatete 2022)	2022	Logistic Regression and Extreme Gradient Boosting	Extreme Gradient Boosting
The Impact of Machine Learning and Aggregated Data on Corporate Insurance Modelling (Hellestol and Eriksen 2022)	2022	CART, Random Forest, Extreme Gradient Boosting, Neural Network, Generalized Linear Model	Extreme Gradient Boosting
Boosting insights in insurance tariff plans with tree-based machine learning methods (Henckaerts et al. 2021)	2021	Generalized Linear Models, Regression Trees, Random Forests, Gradient Boosting Machines	Gradient Boosting Machines
A proposed model to predict auto insurance claims using machine learning techniques (Blier-Wong et al. 2020)	2020	Artificial Neural Network, Decision Tree, Naïve Bayes, Extreme Gradient Boosting	Extreme Gradient Boosting
Extreme Gradient Boosting Machine Learning Algorithm for Safe Auto Insurance Operations (Dhieb et al. 2019)	2019	Extreme Gradient Boosting, Naïve Bayes, Nearest Neighbor, Decision Tree	Extreme Gradient Boosting
Claims Reserving using Gradient Boosting and Generalized Linear Model (Ahlgren 2018)	2018	Generalized Linear Models, Gradient Boosting	Generalized Linear Models
The Accuracy of XGBoost for Insurance Claim Prediction (Fauzan and Murfi 2018)	2018	Extreme Gradient Boosting, Stochastic Gradient Boosting (Stochastic GB), AdaBoost, Random Forest, Neural Network	Extreme Gradient Boosting
P&C Reinsurance Modeling Pure Premium Estimation and Creation of a Reinsurance Program (Chasseray et al. 2017)	2017	Random Forests, Generalized Linear Models, Support Vector Machines, Gradient Boosting Machines, Extreme Gradient Boosting	Extreme Gradient Boosting

The predictive powers of machine learning (ML) models are often improved in various situations. However, when it comes to predicting claim frequency and severity in the complex domain of cargo insurance for cross-border transportation, it is crucial to carefully evaluate and select the most suitable model, whether it is an ML model or a generalized linear model (GLM). The selection of models for cross-border cargo insurance requires a nuanced approach due to the presence of distinctive variables, including complexities in transportation routes, diverse border rules, and distinct risk factors connected with international trade.

In conclusion, although there is a prevailing trend towards improved predictive accuracy of machine learning (ML) models, the decision on the most suitable model for predicting claim frequency and severity when it comes to cargo insurance for cross-

border transportation necessitates a thorough assessment of distinct variables, contextual complexities, and responsiveness to the dynamic nature of international trade risks.

The primary objective of this study is to determine the most optimal approach, whether generalized linear models (GLMs) or machine learning, for developing predictive models for claim frequency and severity within cargo insurance for the cross-border transportation domain. Through detailed performance metric evaluations, mean absolute error (MAE) and root mean squared error (RMSE) are used to gauge the performance of predictive models. MAE calculates the average magnitude of prediction errors, giving a straightforward measure of how far off predictions are from actual values. RMSE offers an interpretable measure by taking the square root of the average squared errors, providing insight into the spread of prediction errors in the same unit as the predicted output. Together, these metrics help evaluate model accuracy and guide improvements in predictive performance. The research aims to ascertain which method—generalized linear models (GLMs) or advanced machine learning—more accurately predicts claim frequency and severity within this specialized insurance domain.

2. Literature Review

From 2017 to 2023, a sequence of extensive studies and research initiatives were conducted to examine the comparative capabilities of machine learning models and the generalized linear models (GLMs) framework in the field of insurance. The primary objective of these studies was to thoroughly examine the effectiveness and efficiency of different machine learning methodologies, including gradient boosting machines, extreme gradient boosting, random forests, neural networks, and generalized linear models (GLMs). As shown in Table 2, the papers conducted thorough evaluations and comparisons of models to determine the superior predictive abilities and appropriateness of machine learning techniques and GLMs in different aspects of insurance. These aspects include modeling claim frequency and severity, estimating pure premiums, and developing reinsurance programs.

Machine Learning vs. Generalized Linear Models (GLMs)

Machine learning (ML) at its core represents a subset of artificial intelligence, empowering computers to autonomously think and learn. Its essence lies in enabling computers to adapt their actions to enhance accuracy (Alzubi et al. 2018). The discipline of machine learning (ML) continues to undergo rapid development, as it exists at the intersection of computer science and statistics, playing a fundamental role in the domains of artificial intelligence (AI) and data science. Machine learning is categorized into four main groups: supervised learning, unsupervised learning, semi-supervised learning, and reinforcement learning (Pugliese et al. 2021). ML techniques have demonstrated superior performance across various tasks compared to traditional methods. However, the heightened adaptability of ML models brings certain challenges. Their complexity and training algorithms often pose hurdles to ensuring performance reliability and impede model interpretability. Additionally, ML models typically demand substantial training data. Nonetheless, if high performance is a priority for a specific task and adequate training data are available, the advantages of ML might outweigh these challenges (Bianco et al. 2019).

Machine learning algorithms such as extreme gradient boosting (XGBoost) and gradient boosting machines (GBMs) leverage the potential of ensemble learning to effectively identify intricate patterns and nonlinear associations within datasets. Extreme gradient boosting (XGBoost) represents an enhanced iteration of gradient boosting machines (GBMs), incorporating parallel preprocessing at the node level, enhancing its speed compared to gradient boosting machines (GBMs). Additionally, extreme gradient boosting (XGBoost) introduces diverse regularization methods aimed at mitigating overfitting (Chen and Guestrin 2016). The versatility of these techniques has the potential to make them highly effective in predicting the frequency and severity of claims, especially in situations where there are complex risk environments. The Hellestol and Eriksen (2022) study involves a comparison of machine learning methods, including CART, random forest, XGBoost,

and neural networks, with benchmark GLMs. The results show that all machine learning models outperformed GLMs when classifying claim occurrences. In the paper by Blier-Wong et al. (2020), a comprehensive review of machine learning in property and casualty (P&C) insurance reveals that within pricing and reserving, extreme gradient boosting (XGBoost) and gradient boosting trees emerge prominently as the favored and widely-used frameworks. These models stand out as the most popular choices within the industry for enhancing pricing methodologies and reserve estimations.

Despite the surge in machine learning's prominence, ongoing comparisons between predictive models remain crucial. Recent research, exemplified by Tuininga's study in 2022, reveals that even when trained on vehicle insurance data, models like GBR, XGB, RF, and NNs (neural networks) could not surpass the performance of the generalized linear model (GLM). This highlights the enduring significance and relevance of established statistical methods like GLMs in certain domains, despite advancements in machine learning (Tuininga 2022). Furthermore, in certain scenarios where machine learning exhibits superior performance compared to GLMs, the improvement observed tends to be only marginally better. Jan Mikael Yousif conducted a study titled "A Comparative Analysis Between Various Machine Learning Models and Generalized Linear Models". The study aimed to assess improvements made by machine learning (ML) models compared to generalized linear models (GLMs). The improvements seen with machine learning models were not as significant as expected. This was largely because the generalized linear models (GLMs) had already demonstrated strong predictive abilities for the particular dataset (Yousif 2023).

The study from Clemente et al. (2023) also shows that when studying both machine learning and GLMs, the results from assessing performance outside the sample indicate that the gradient boosting model (GBM) demonstrates better predictive accuracy than the standard generalized linear models (GLMs) in the Poisson claim frequency model. However, in terms of claim severity, generalized linear models (GLMs) performed better than the gradient boosting model. This study shows that it remains valuable to compare prediction model performances before selecting the appropriate model for predicting specific data.

Cargo insurance falls under the category of non-life insurance and specifically caters to protecting goods during transit. For example, the study conducted by Monemar and Wallin (2015), titled "Premium Allocation for the Electrolux Cargo Insurance Program using Generalized Linear Models", delves into the application of generalized linear models (GLMs) within the realm of cargo insurance. Actuaries in insurance companies play a vital role in assessing premiums, reserves, and risk analysis. They use historical and current data along with mathematical and statistical methodologies to forecast future risk events. For cargo insurance, determining the actual premium involves considering claim frequency and severity, revealing low accident rates but substantial damages, resulting in high-value claims per occurrence.

3. Study Design

3.1. Machine Learning

3.1.1. Extreme Gradient Boosting (XGBoost)

XGBoost made its initial introduction through the collaborative efforts of Chen and Guestrin (2016). XGBoost, short for extreme gradient boosting, stands out as an immensely efficient machine learning algorithm centered around decision tree techniques as its fundamental building blocks. It constructs a robust model composed of a forest that houses numerous decision trees. Notably, XGBoost's effectiveness surpasses several other machine learning algorithms, such as artificial neural networks, gradient boosting machines, and random forests, as proven by Hellestol and Eriksen (2022), Abdelhadi et al. (2020), and Chasseray et al. (2017). XGBoost is a well-recognized supervised learning algorithm, distinguished by its composition of an objective function and base learners. The objective function incorporates a loss function that measures the difference between expected and actual values. Furthermore, it incorporates a regularization term that quantifies the discrep-

ancy between expected and observed values. In XGBoost's ensemble learning framework, a set of base learners, which are several models, is crucial for predicting a single result. Within this particular framework, a regressor undertakes the task of accurately modeling a given set of attributes and then predicting the value of an unknown output (Avanijaa 2021). XGBoost models have exhibited their proficiency in both classification and general regression tasks (Kankanamge et al. 2019). This innovation enhanced the algorithm's efficacy in training while significantly accelerating computation, as shown in Ge et al.'s (2022) study, which indicates that while conventional trees rely solely on first-order derivatives, XGBoost regression (XBR) incorporates second-order derivatives and regularization terms. The basic function of the XGBoost regression algorithm in mathematical equations, starting from the objective function to the ensemble model, can be represented as follows:

Given that X represents the matrix of predictor variables, y denotes the target variable (continuous for regression tasks). T signifies the number of boosting iterations. $h_t(x)$ represents the prediction of the tth model. γ denotes the learning rate. $L(y, \hat{y})$ represents the loss function, measuring the difference between predicted $\hat{y}$ and actual y. $\Omega(h)$ represents the regularization term for the weak learner h. The XGBoost regression objective function is a combination of the loss function and regularization, as shown below:

$$\text{Objective} = \sum_{i=1}^{n} L(y_i, \hat{y}_i) + \gamma \sum_{t=1}^{T} \Omega(h_t) \tag{1}$$

Step-by-Step Procedure: Start with initialization by first setting the initial predictions at zero, $\hat{y}_0 = 0$, and $t = 1$ to T. Then compute the pseudo-residuals (negative gradient) of the loss function:

$$g_i = -\left[\frac{\partial L(y_i, \hat{y}_i)}{\partial \hat{y}_i}\right]_{\hat{y}_i = \hat{y}_i^{(t-1)}} \tag{2}$$

After that, fit a weak learner (e.g., a decision tree) to the pseudo-residuals (negative gradient):

$$h_t(x) = \arg\min_{h} \sum_{i=1}^{n} (g_i - h(x_i))^2 + \Omega(h) \tag{3}$$

Update the ensemble predictions. In the context of XGBoost regression, the predicted output $(\hat{y})$ at each iteration t is derived from the ensemble model, incorporating the predictions of individual weak learners (often decision trees) into the overall model. The predicted output equation can be represented as follows:

At each boosting iteration t, the predicted output $(\hat{y}_t)$ is updated based on the ensemble model:

$$\hat{y}_t = \hat{y}_{t-1} + \gamma h_t(x) \tag{4}$$

where $\hat{y}_{t-1}$ represents the predicted output from the previous iteration, $h_t(x)$ denotes the prediction of the tth weak learner (e.g., decision tree) for the input x, and γ signifies the learning rate, controlling the step size during each update. This equation describes how the ensemble model aggregates the predictions of individual weak learners $(h_t(x))$ into the overall predicted output $(\hat{y}_t)$ at each boosting iteration. The predictions from each weak learner are scaled by the learning rate γ and added to the previous ensemble predictions to refine and improve the model's overall prediction.

This stepwise representation outlines the XGBoost regression algorithm in mathematical equation, detailing the objective function, computation of pseudo-residuals (negative gradient), fitting of weak learners, and the iterative update of the ensemble predictions.

3.1.2. Gradient Boosting Machines (GBMs)

A Gradient Boosting Machine (GBM) is a powerful ensemble machine learning technique used for regression and classification tasks. A GBM operates by sequentially combining multiple weak predictive models, often decision trees, to create a more robust and accurate final model. In the realm of regression, gradient boosting machines work by

iteratively minimizing the errors or residuals of the preceding models. A formulation of boosting methods based on gradient descent was developed (Friedman 2001). Each new model is built to correct the mistakes made by its predecessors, optimizing the overall prediction performance by focusing on the remaining errors. This sequential approach involves fitting the new model to the residuals of the combined ensemble, gradually reducing prediction errors (Ridgeway 2007). Gradient boosting stands out for its dual capacity to achieve superior predictive accuracy while enabling model interpretability. This distinctive trait holds immense value, particularly in business environments where models are assessed by decision-makers without statistical expertise. These stakeholders prioritize comprehensibility, requiring an understanding of the model's outputs. Its capability to balance accuracy with interpretability caters well to this need, empowering non-statistically trained individuals to grasp and trust the insights gleaned from the model's predictions (Cordeiro 2023). However, a GBM is not immune to challenges. It can be sensitive to overfitting, especially when the number of trees is too high or when the individual trees become overly complex. Regularization techniques are implemented to calibrate the model training process, aiming to strike a balance between the accuracy of the model on the training data and its predictive capability on new data (Elith et al. 2008). The success of gradient boosting machines has led to the development of optimized implementations like XGBoost (Chen and Guestrin 2016) and LightGBM (Ke et al. 2017), which enhance training efficiency and scalability, making them suitable for large-scale datasets and real-time applications. In essence, gradient boosting machines, especially in regression tasks, leverage the strengths of ensembling weak learners to create a powerful predictive model by iteratively minimizing errors and handling complex data relationships, albeit requiring careful tuning to avoid overfitting. The basic function of a gradient boosting regression algorithm in mathematical equation, starting from the objective function to the ensemble model, can be represented as follows:

Given that X represents the matrix of predictor variables, y denotes the target variable (continuous for regression tasks). T signifies the number of boosting iterations. $h_t(x)$ represents the prediction of the tth model. γ denotes the learning rate. $L(y, \hat{y})$ represents the loss function, measuring the difference between predicted $\hat{y}$ and actual y. $\Omega(h)$ represents the regularization term for the weak learner h. The gradient boosting regression objective function is defined as follows:

$$\text{Objective } = \sum_{i=1}^{n} L(y_i, \hat{y}_i) \tag{5}$$

$L(y_i, \hat{y}_i)$ denotes the loss function, typically squared error loss or another regression-specific loss function. However, the below step is very similar to XGBoost. Start with initialization by setting the initial predictions at zero, $\hat{y}_0 = 0$, and $t = 1$ to T, then compute the pseudo-residuals (negative gradient) of the loss function:

$$g_i = -\left[\frac{\partial L(y_i, \hat{y}_i)}{\partial \hat{y}_i}\right]_{\hat{y}_i = \hat{y}_i^{(t-1)}} \tag{6}$$

Fit a weak learner (e.g., a decision tree) to the pseudo-residuals (negative gradient):

$$h_t(x) = \underset{h}{\text{argmin}} \sum_{i=1}^{n} (g_i - h(x_i))^2 + \Omega(h) \tag{7}$$

Update the ensemble model predictions:

$$\hat{y}_t = \hat{y}_{t-1} + \gamma h_t(x) \tag{8}$$

where $\hat{y}_{t-1}$ represents the predicted output from the previous iteration, $h_t(x)$ denotes the prediction of the tth weak learner (e.g., decision tree) for the input x, and γ signifies the

learning rate, controlling the step size during each update. This equation describes how the ensemble model aggregates the predictions of individual weak learners $(h_t(x))$ into the overall predicted output $(\hat{y}_t)$ at each boosting iteration. This stepwise representation outlines the gradient boosting regression algorithm in mathematical terms. The algorithm sequentially fits weak learners to the negative gradient (pseudo-residuals) and updates the ensemble model to minimize the loss function and improve predictions iteratively.

In both gradient boosting machines (GBMs) and extreme gradient boosting (XGBoost), computing the negative gradient of the loss function is a fundamental step during each boosting iteration. However, while the concept of computing the negative gradient remains the same, the detailed implementation might differ between the two algorithms due to optimizations and additional features introduced in extreme gradient boosting (XGBoost). The concept of computing the negative gradient remains consistent between gradient boosting machines (GBMs) and extreme gradient boosting (XGBoost). However, the actual implementation might differ in extreme gradient boosting (XGBoost) due to optimizations like approximate algorithms, weighted quantile sketches, and other advanced techniques. These enhancements aim to make the computation more efficient without fundamentally changing the underlying principle of computing the negative gradients. In summary, while the fundamental concept of computing the negative gradient of the loss function is shared between gradient boosting machines (GBMs) and extreme gradient boosting (XGBoost), extreme gradient boosting (XGBoost) might employ optimizations and improvements in its implementation to compute these gradients more efficiently or accurately compared to traditional gradient boosting.

3.2. Generalized Linear Models (GLMs)

The generalized linear model (GLM) is a fundamental statistical framework that extends traditional linear regression to accommodate a broader range of data distributions and relationships between variables. It was introduced by Nelder and Wedderburn in 1972, revolutionizing statistical modeling (Nelder and Wedderburn 1972). At the core of the GLM is the linear predictor, which combines predictors linearly to model the relationship with the response variable. A GLM incorporates a link function that connects the linear predictor to the expected value of the response variable. Different link functions are utilized based on the nature of the response variable, including the logit, log-link, and identity functions. Unlike traditional linear regression, a GLM is not limited to the normal distribution. It can handle various distributions such as binomial, Poisson, gamma, and others, making it versatile for analyzing different types of data. A GLM's flexibility in handling diverse data distributions and accommodating non-linear relationships between variables has made it widely applicable across numerous fields, including epidemiology, ecology, finance, and social sciences. A generalized linear model (GLM) employs iterative algorithms like iteratively reweighted least squares or maximum likelihood estimation to estimate parameters efficiently. It allows for hypothesis testing on model coefficients, aiding in determining the significance of predictors in explaining the response variable's variance. The ability of generalized linear models (GLMs) to handle a wide range of data distributions and incorporate various link functions (Abhishek 2023) has significantly contributed to their popularity and utility in statistical modeling, offering researchers and practitioners a powerful and adaptable tool for analyzing complex datasets. A GLM's fundamental principles underpin more advanced modeling techniques and have spurred the development of extensions and variations, allowing for the creation of tailored models to suit specific research questions and datasets. Its versatility continues to be a cornerstone in statistical analysis and predictive modeling. In the framework of generalized linear models (GLMs), Poisson regression is utilized to model count data by assuming a Poisson distribution for the response variable. This model expresses the logarithm of the expected counts as a linear combination of predictor variables. The canonical link function for Poisson regression is the logarithm, linking the mean of the response variable to the linear combination of predictors. Its application is particularly effective when dealing with

count-based outcomes such as the number of events, occurrences, or frequencies in a fixed period or area. Gamma regression, another component of GLMs, is designed for continuous, positively skewed data, assuming a gamma distribution for the response variable. It models the logarithm of the expected value of the dependent variable as a linear combination of predictors. The structural equation for the Poisson and Gamma regressions within the framework of generalized linear models (GLMs) is indeed the same. Both regressions utilize a log-linear relationship between the expected value (μ) of the dependent variable and the predictors. The notation for the model equation for both Poisson and Gamma regressions is detailed below.

Y is the dependent variable, representing the counts of events (assumed to follow a Poisson distribution) for Poisson and continuous, positively skewed data for Gamma. $\beta_0, \beta_1, \beta_2, \ldots, \beta_p$ are coefficients corresponding to the intercept and predictor variables. $x_1, x_2, \ldots, x_p$ are predictor variables. μ is the expected value (mean) of the dependent variable Y, given the values of the predictors. e is the base of the natural logarithm. The Poisson and Gamma regression model assumes a log-linear relationship between the expected value of the dependent variable μ and the predictors. The model equation for the ith observation is:

$$\log(\mu_i) = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_p x_{ip} \tag{9}$$

$\log(\mu_i)$ is the natural logarithm of the expected value of Y for the ith observation. $x_{i1}, x_{i2}, \ldots, x_{ip}$ are the values of the predictor variables for the ith observation. $\beta_0, \beta_1, \ldots, \beta_p$ are coefficients corresponding to the intercept and predictor variables. The relationship between μ and Y is detailed below.

The expected value μ_i of the dependent variable Y is related to the model through the exponential function:

$$\mu_i = e^{\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_p x_{ip}} \tag{10}$$

For the model, $\log(\mu_i) = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_p x_{ip}$, and for the expected count and value, $\mu_i = e^{\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_p x_{ip}}$. This model helps to estimate the expected count of events and the expected value of the independent Y based on the given predictors $x_1, x_2, \ldots, x_p$ using a logarithmic link function.

Extreme gradient boosting (XGBoost), gradient boosting machines (GBMs), and generalized linear models (GLMs) stand out as models of choice for cargo insurance due to their unique strengths aligning with the intricacies of this domain. XGBoost's speed and ability to handle voluminous data and complex relationships suit the analysis of diverse cargo-related variables, aiding in predicting and mitigating risks associated with transportation. A GBM's iterative learning and adaptability to different loss functions make it adept at capturing patterns among heterogeneous cargo-related factors, allowing for nuanced risk assessment. Meanwhile, a GLM's flexibility in accommodating various data distributions is invaluable for modeling different types of cargo damage or loss occurrences, enhancing predictive capabilities within the context of specific damage scenarios. Together, these models offer a comprehensive toolkit to address the multifaceted challenges of cargo insurance, from complex risk patterns to nuanced damage predictions.

3.3. Mathematical Formulation for Claim Frequency and Severity Prediction Models

3.3.1. Claim Frequency Prediction Problem Formulation

Predicting the frequency of insurance claims based on various shipment and policy parameters.

Objective: To estimate the count of insurance claims based on various input features provided, including the nature of the shipment, countries involved, and insurance details.

In assessing the factors that contribute to claim frequency, various variables were considered. As detailed in Table 3, the variables include the status of goods, cargo type, packaging type, the countries of origin and destination, the sum insured amount, and the number of

claims. These variables are categorized accordingly, with some being categorical and others continuous numerical.

Table 3. Features for claim frequency prediction.

Variable	Sub-Variable	Features
Status of Goods	Import, Export	Categorical
Cargo Type	Cargo Type Group 1, Cargo Type Group 2, Cargo Type Group 3, Cargo Type Group 4, Cargo Type Group 5, Cargo Type Group 6	Categorical
Packaging Type	In bulk, Carton/Box, Case/Crate, Tin/Drum, Bag/Sack, Pallet/Skid, Bundle/Bale, Roll/Coil, Others	Categorical
Start Country	Thailand, Laos, Myanmar, Cambodia, Malaysia	Categorical
Destination Country	Thailand, Laos, Myanmar, Cambodia, Malaysia	Categorical
Sum Insured Amount	0,. . .,∞	Continuous numerical
Number of Claims	0,. . .,∞	Continuous numerical

Prediction Target: The number of insurance claims within a specific context or time frame [Claim Frequency].

Mathematical Equations

Input:

X represents the feature matrix, with each row x_i containing the mentioned features.
X_{i1}: Status of Goods (S_i)
X_{i2}: Cargo Type (C_i)
X_{i3}: Packaging Type (P_i)
X_{i4}: Start Country (SC_i)
X_{i5}: Destination Country (DC_i)
X_{i6}: Sum Insured Amount (SI_i)
X_{i7}: Number of Claims (NC_i)

Output:

Y represents the predicted claim frequency (CF).

The prediction problem can be represented as finding a function f that maps input features X to the predicted claim frequency CF:

$$CF = f(S, C, P, SC, DC, SI, NC) + \epsilon \tag{11}$$

where f is the regression function, S, C, P, SC, DC, SI, NC represent the input features, CF is the predicted count of claim frequency, and ϵ represents the residual error term.

This formulation defines the problem of predicting claim frequency based on specific features and aims to model the count of insurance claims by leveraging these features.

3.3.2. Claim Severity Prediction Problem Formulation

Predicting the severity of insurance claims based on various shipment and policy parameters.

Objective: To estimate the monetary value or cost associated with individual insurance claims based on various input features, including shipment details, countries involved, and insurance details.

In assessing the factors that contribute to claim severity, various variables were considered. As detailed in Table 4, the variables include the status of goods, cargo type, packaging

type, the countries of origin and destination, the sum insured amount, and the Incurred claims. These variables are categorized accordingly, with some being categorical and others continuous numerical.

Table 4. Features for claim severity prediction.

Variable	Sub-Variable	Features
Status of Goods	Import, Export	Categorical
Cargo Type	Cargo Type Group 1, Cargo Type Group 2, Cargo Type Group 3, Cargo Type Group 4, Cargo Type Group 5, Cargo Type Group 6	Categorical
Packaging Type	In bulk, Carton/Box, Case/Crate, Tin/Drum, Bag/Sack, Pallet/Skid, Bundle/Bale, Roll/Coil, Others	Categorical
Start Country	Thailand, Laos, Myanmar, Cambodia, Malaysia	Categorical
Destination Country	Thailand, Laos, Myanmar, Cambodia, Malaysia	Categorical
Sum Insured Amount	0,. . .,∞	Continuous numerical
Incurred Claims	0,. . .,∞	Continuous numerical

Prediction Target: The monetary value or cost associated with individual insurance claims [Claim Severity].

Mathematical Equations

Inputs:

X represents the feature matrix, with each row x_i containing the mentioned features.
X_{i1}: Status of Goods (S_i)
X_{i2}: Cargo Type (C_i)
X_{i3}: Packaging Type (P_i)
X_{i4}: Start Country (SC_i)
X_{i5}: Destination Country (DC_i)
X_{i6}: Sum Insured Amount (SI_i)
X_{i7}: Incurred Claims (IC_i)

Output:

Y represents the predicted claim severity (CS).

The prediction problem can be represented as finding a function g that maps input features X to the predicted claim severity CS:

$$CS = g(S, C, P, SC, DC, SI, IC) + \epsilon \tag{12}$$

where g is the regression function, S, C, P, SC, DC, SI, IC represent the input features, CS is the predicted severity of claims, and ϵ represents the residual error term.

This formulation defines the problem of predicting claim severity based on specific features and aims to model the monetary value associated with individual claims using these features.

3.4. Hyperparameter Tunning

Hyperparameter tuning is the process of optimizing the hyperparameters of a machine learning model to enhance its performance. These parameters are set prior to the training process and influence the learning process's behavior and complexity. Selecting the right hyperparameters is crucial, as they directly impact a model's ability to learn and generalize from the training data to new, unseen data (Yang and Abdallah 2020). Various methods exist for hyperparameter tuning, such as grid search, random search, Bayesian optimization, and evolutionary algorithms. In Python, libraries like Scikit-learn,

TensorFlow, and Keras offer built-in functionalities to perform hyperparameter tuning efficiently, enabling researchers and practitioners to automate and optimize this critical aspect of model development. The effectiveness of machine learning models significantly relies on hyperparameters, which control the learning process. For example, in the extreme gradient boosting (XGBoost) model, parameters like criteria, maximum depth, and the number of estimators is pivotal. These settings notably impact how easily a model can be trained. Hyperparameter optimization aims to uncover the best combination of these values, ensuring optimal model performance within a feasible time frame and enhancing its learning and predictive capabilities (Dalal et al. 2022).

In Table 5 provides descriptions of several parameters: the learning rate, the number of estimators, the maximum depth of the trees, and the alpha value.

Table 5. Description of the parameters.

Parameter	Description
learning_rate	Initial learning rate
n_estimators	Number of decision trees
max_depth	Maximum tree depth
alpha_value	Controls the shape of the distribution

Source: learning_rate, n_estimators, and max_depth descriptions from Zhao et al. (2022).

3.5. Outlier Detection

Outliers, in data analysis, are observations that significantly differ from the majority of the dataset. Detecting outliers is vital to ensuring data integrity, as they can distort statistical analyses and model performances. Various techniques exist for outlier detection, aiming to identify these anomalies and investigate their potential causes (Chandola et al. 2009). Outlier detection serves as a crucial task. This practice is contingent on the domain and has undergone comprehensive exploration, finding significant utility in pinpointing uncommon instances across various real-world applications. Its applications span diverse domains, encompassing network intrusion detection, medical diagnosis, fraud detection, and the identification of manufacturing defects (Alimohammadi and Chen 2022).

3.6. Z-Score Method

The Z-score method is a prevalent statistical technique used for outlier detection based on standard deviations from the mean. It involves calculating the Z-score for each data point, indicating how many standard deviations it is from the mean. Typically, a Z-score threshold of 3 or -3 is employed to identify outliers. This implies that any data point with a Z-score exceeding 3 or falling below -3 is considered an outlier (Yaro et al. 2023). Frequently utilized in diverse fields, basic statistical tools like the Z-score play a routine role in outlier identification within datasets. The Z-score calculates the distance between a data point and its mean, with values exceeding ± 3 commonly categorized as outliers (Jha et al. 2022).

3.7. Data Encoding

One-hot encoding is an approach utilized to transform categorical variables into a numerical format suitable for machine learning algorithms (Tuininga 2022). This encoding creates a binary column for each category present in the variable, setting the corresponding bit to 1 for the category and 0 for all others.

Figure 3 displays a sample representation of one-hot encoding, where each category within the variable is transformed into a unique binary column.

Country	Thailand	Malaysia	Cambodia
Thailand	1	0	0
Malaysia	0	1	0
Cambodia	0	0	1

Figure 3. Example of one-hot encoding.

3.8. K-Fold Cross-Validation

In machine learning studies, the dataset is typically divided into training and test sets. The training set is utilized by a machine learning model to establish a mathematical correlation between features and target variables. The training set typically outweighs the test set in size, leading to a potential issue with an unrepresentative test set structure affecting model performance—either excessively well or poorly. K-fold cross-validation addresses this by repetitively using the same dataset for both training and testing, mitigating biases and enhancing the model's robustness (Tuininga 2022). K-fold cross-validation is adept at enhancing the model's generalization, while an ensemble model can yield superior predictive accuracy compared to an individual model (Zhu et al. 2019).

Figure 4 presents a visual example of the K-fold cross-validation process with k set to 5, demonstrating how the data set is partitioned into five distinct subsets for validation and training.

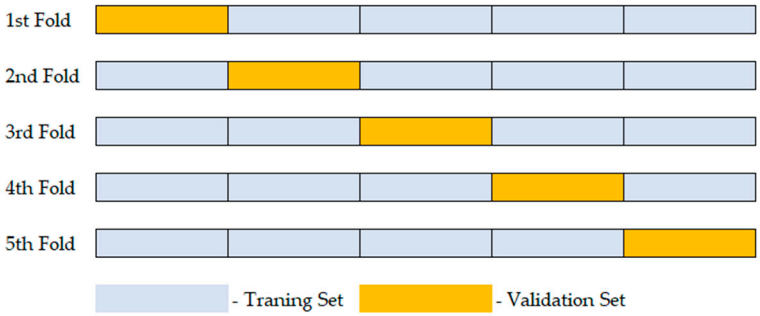

Figure 4. K-fold cross-validation (k = 5).

The figure above illustrates a standard K-fold cross-validation with k = 5.

3.9. Model Comparison

The Root Mean Squared Error (RMSE) and Mean Absolute Error (MAE) are commonly utilized measures for assessing the performance of models (Hodson 2022). RMSE highlights large errors by squared differences, while MAE measures absolute differences and is less sensitive to outliers. The choice of metric relies on data characteristics and desired sensitivity to diverse error types, offering distinct insights into model performance. A combination of metrics is often necessary for a comprehensive assessment of model performance (Chai and Draxler 2014).

Root mean squared error (RMSE):

$$\text{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2} \tag{13}$$

where n is the number of samples, y_i represents the actual value for the ith sample, and $\hat{y}_i$ denotes the predicted value for the ith sample.

Mean absolute error (MAE):

$$\text{MAE} = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i| \tag{14}$$

where n is the number of samples, y_i represents the actual value for the ith sample, and $\hat{y}_i$ denotes the predicted value for the ith sample.

These equations provide a mathematical representation of how RMSE and MAE are calculated based on differences between predicted $(\hat{y}_i)$ and actual (y_i) values across a dataset of n samples.

Comparing MAE and RMSE offers diverse insights into prediction accuracy. MAE focuses on average error magnitude, and RMSE balances sensitivity to larger errors with interpretability. These metrics aid a nuanced understanding of model behavior; for instance, lower RMSE implies better accuracy, while MAE is suitable for outlier sensitivity. Selecting the right metric hinges on specific objectives and the data context. This comparison guides informed decisions in model selection and fine-tuning, offering clarity on trade-offs between error types in various applications.

4. Research Methodology

4.1. Data Collection

The dataset gathered for this study was acquired from the Insurance Premium Rating Bureau in Thailand, covering cargo insurance data related to road transportation from 2016 to 2022. This dataset stands as an extensive repository of information specifically concerning cargo insurance within the specified timeframe and in the context of road transportation.

4.2. Dataset Size

The cargo insurance dataset from 2016 to 2022 is composed of 9803 insurance data points collected for this research.

4.3. Dataset Description

Table 6 meticulously delineates categorical variables pivotal in understanding the nuances of cross-border trade entities, encompassing elements like the status of goods (e.g., 'Import' or 'Export'), distinct cargo types (e.g., 'Group 1' through 'Group 6'), varied packaging types (ranging from 'In bulk' to 'Roll/Coil'), and the originating and destination countries (including 'Thailand', 'Laos', 'Myanmar', 'Cambodia', and 'Malaysia'). These categorical variables offer crucial insights into the diverse facets of trade operations and logistical intricacies.

Table 6. Description of the categorical variables.

Variable	Sub-Variable (Categorical)	Description
Status of Goods	Import, Export	Represents the classification of goods based on their intended importation or exportation.
Cargo Type	Cargo Type Group 1, Cargo Type Group 2, Cargo Type Group 3, Cargo Type Group 4, Cargo Type Group 5, Cargo Type Group 6	Represents the different groups categorizing the type of cargo being transported.
Packaging Type	In bulk, Carton/Box, Case/Crate, Tin/Drum, Bag/Sack, Pallet/Skid, Bundle/Bale, Roll/Coil, Others	Represents the various forms or methods of packaging used for the transported goods.
Start Country	Thailand, Laos, Myanmar, Cambodia, Malaysia	Represents the countries from which the cargo shipments originate.
Destination Country	Thailand, Laos, Myanmar, Cambodia, Malaysia	Represents the countries to which the cargo shipments are destined.

Simultaneously, Table 7 shows continuous numerical variables, such as the sum insured amount, the number of claims (utilized for predicting claim frequency), and the incurred claims (used for predicting claim severity). These continuous variables provide quantitative data on insurance-related figures, aiding in statistical analysis and predictive modeling to ascertain the risks and financial implications associated with insurance claims for entities engaged in cross-border trade activities.

Table 7. Description of the numerical variables.

Variable	Value (Continuous Numerical)	Description
Sum Insured Amount	0,. . .,∞	Represents the maximum amount an insurance company agrees to pay in the event of a covered loss or damage.
Number of Claims	0,. . .,∞	Represents the count or frequency of claims.
Incurred Claims	0,. . .,∞	Represents the total value or amount of all claims that an insurer anticipates paying or has already paid during a specific period.

4.4. Research Tools

Within this research methodology, this research leveraged Google Colab as the operational platform and utilized Python version 3.10.12 as the primary programming language within the toolset. Additionally, Microsoft Excel for Microsoft 365 MSO (Version 2312 Build 16.0.17126.20132) 64-bit served as a supplementary tool for specific data manipulation and analysis tasks within the workflow.

4.5. Research Methods

Figure 5 explains that our research methodology begins with the acquisition of historical insurance data sourced from the Insurance Premium Rating Bureau (Thailand). This dataset undergoes a division into two core subsets: the first contains claim frequency historical data, constituting the original frequency dataset, while the second contains historical information on claim value, shaping the original severity dataset. Following this partitioning, our methodology meticulously addresses data quality. Missing data were removed, and outliers were detected and rectified using the Z-score method within Google Colab, resulting in two refined and cleansed datasets primed for analysis. Given the presence of textual variables within the datasets, a crucial step involves encoding these textual elements into numeric formats, ensuring compatibility with analytical algorithms without compromising data integrity. Employing K-fold validation with k = 5, the datasets undergo a meticulous data train–test split, enabling robust model development and unbiased evaluation. Moving forward, our methodology entails algorithm refinement through hyperparameter tuning. Algorithms like extreme gradient boosting (XGBoost) and gradient boost models undergo this process, optimizing their performance. While XGBoost requires tuning various hyperparameters, it is essential to note that for the Poisson model, hyperparameter tuning in this study is not involved. For the Gamma distribution, the adjustment of parameters like 'alpha' becomes crucial for model refinement. Executing our research methodology in the Google Colab environment, harnessing Python version 3.10.12, facilitates comprehensive model evaluation. Key performance metrics such as mean absolute error (MAE) and root mean squared error (RMSE) are computed across all models, enabling an insightful comparison to determine their efficacy and suitability for our analysis.

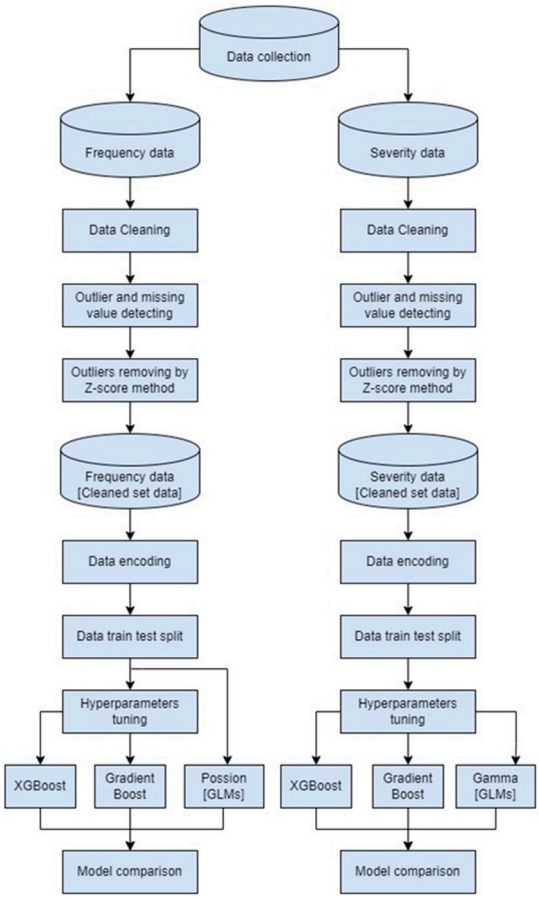

Figure 5. Research methods.

5. Result

Table 8 shows the original dataset that needs to be divided into two subsets based on specific criteria: the "original frequency dataset" denotes the initial dataset concentrating on the frequency of events or incidents observed within a specified period, and the "original severity dataset" represents the initial dataset detailing the intensity, impact, or seriousness of those recorded occurrences or events found in the frequency dataset. These datasets stand as raw, unprocessed datasets before any data cleaning steps, including the removal of missing values or outliers, have been applied. Subsequently, the cleaned dataset represents the modified dataset resulting from the data cleaning process after addressing missing values and removing outliers. Once the original dataset was divided into subsets—one focusing on claim frequency and the other on claim severity—our next step involved handling missing values and managing outliers. Missing values and outliers, which are extreme or uncommon data points, can distort statistical analyses or machine learning models were removed. To identify and address outliers, the Z-score method was employed. This technique detects outliers based on how many standard deviations a data point deviates from the mean. The code implemented removed outliers using Z-scores, setting a threshold of 3. A comparison between the dataset before and after data cleaning is presented in Table 8 to illustrate the impact of this process.

Table 8 exhibits a reduction in the total number of data entries after the removal of outliers and missing values in both the frequency and severity datasets. In the frequency dataset, the count decreased from an initial 9803 data entries to 9631 entries post-processing. Similarly, in the severity dataset, the count decreased from 150 entries initially to 147 entries after the data cleansing process. This decrease in data count signifies that some entries were identified as outliers or contained missing values, leading to their removal. While the reduction in data size might affect the overall sample size, it ensures a more refined

dataset free from extreme values or incomplete information. Consequently, the resulting datasets are more focused and potentially more suitable for subsequent analysis, modeling, or statistical inference by providing a cleaner and more reliable set of data.

Table 8. Frequency and severity datasets.

Variables		Frequency Dataset		Severity Dataset	
		Total Number of Data (Original Dataset)	**Total Number of Data (Cleaned Dataset)**	**Total Number of Data (Original Dataset)**	**Total Number of Data (Cleaned Dataset)**
Status of Goods	Import	2714	2670	56	55
	Export	7089	6961	94	92
Cargo Type Group	Group 1	3118	3013	66	65
	Group 2	26	26	2	2
	Group 3	1122	1109	21	20
	Group 4	90	90	-	-
	Group 5	399	395	12	12
	Group 6	5048	4998	49	48
Packaging Type	In bulk	33	33	-	-
	Carton/Box	2514	2477	49	48
	Case/Crate	667	666	5	5
	Tin/Drum	108	108	5	5
	Bag/Sack	369	365	12	12
	Pallet/Skid	836	811	28	28
	Bundle/Bale	695	694	-	-
	Roll/Coil	133	133	2	2
	Others	4448	4344	49	47
Start Country	Thailand	7080	6952	95	93
	Laos	144	140	1	1
	Myanmar	84	83	2	2
	Cambodia	105	83	5	4
	Malaysia	2390	2373	47	47
Destination Country	Thailand	2714	2670	55	54
	Laos	2008	1971	41	41
	Myanmar	2403	2398	11	11
	Cambodia	1045	1027	20	19
	Malaysia	1633	1565	23	22
Average Sum Insured Amount			9.88 million THB		10.98 million THB
Average Number of Claims			0.017		-
Average Incurred Claims			-		86,791.40 THB
Total Number of Data		9803 (100%)	9631 (100%)	150 (100%)	147 (100%)

Following the resolution of outliers and missing data, the subsequent phase entails encoding the datasets by code implementation in Google Colab. Transitioning into the data splitting stage, the data were split into five subsets for cross-validation, utilizing K-fold with a value of 5. Subsequently, leveraging a predefined table, hyperparameter tuning was performed, an essential step in optimizing the model's configuration for enhanced performance and robustness.

Table 9 shows the range hyperparameter and the best hyperparameter in each model approach. In complement to the hyperparameter tuning specified in the table, this study

initializes the code to retain the mean absolute error (MAE) of the superior models achieved through extreme gradient boosting (XGBoost), gradient boosting machines (GBMs), and generalized linear models (Poisson–Gamma). It acts as a container for storing the MAE values obtained from the best-performing models within each algorithm during the tuning process. Following the execution of Python package code in Google Colab for the frequency and severity datasets using extreme gradient boosting (XGBoost), gradient boosting machines (GBMs), and generalized linear models, a comprehensive assessment was obtained. The objective of this study is to conduct a thorough comparison of key performance metrics, specifically MAE and RMSE, among several models: extreme gradient boosting (XGBoost), gradient boosting machines (GBMs), and generalized linear models (Poisson–Gamma). This comparison aims to discern and evaluate the predictive accuracy and efficacy of each model variant, shedding light on their respective strengths and weaknesses within the context of the analysis.

Table 9. Hyperparameter tuning.

Model Approach	Frequency		Severity	
	Range Hyperparameters	**Best Hyperparameters**	**Range Hyperparameters**	**Best Hyperparameters**
Extreme Gradient Boosting	learning_rate = [0.01, 0.02, 0.03, 0.04, 0.05], n_estimators = [39, 40, 41, 42, 43], max_depth = [17, 18, 19, 20, 21]	learning_rate = [0.03], n_estimators = [41], max_depth = [19]	learning_rate = [0.05, 0.06, 0.07, 0.08, 0.09] n_estimators = [6, 7, 8, 9, 10] max_depth = [5, 6, 7, 8, 9]	learning_rate = [0.07], n_estimators = [8], max_depth = [7]
Gradient Boosting Machines	learning_rate = [0.12, 0.13, 0.14, 0.15, 0.16] n_estimators = [12, 13, 14, 15, 16] max_depth = [5, 6, 7, 8, 9]	learning_rate = [0.14], n_estimators = [12], max_depth = [7]	learning_rate = [0.14, 0.15, 0.16, 0.17, 0.18] n_estimators = [8, 9, 10, 11, 12] max_depth = [1, 2, 3, 4, 5]	learning_rate = [0.16], n_estimators = [10], max_depth = [3]
Generalized Linear Models [Poisson–Gamma]	-	-	alpha_values = [1000000000000000000, 1100000000000000000, 1200000000000000000]	alpha_values = [1100000000000000000]

Metrics Performance Comparison

Table 10 provides metrics that showcase the performance of different machine learning models and GLMs, highlighting their effectiveness in predictive tasks.

Table 10. Performance comparison.

Model Approach	Metrics Performance	Frequency	Severity
Extreme Gradient Boosting	MAE	0.0309	113,085.07
	RMSE	0.1534	234,877.46
Gradient Boosting Machines	MAE	0.0328	116,708.85
	RMSE	0.1445	237,625.14
Generalized Linear Models [Poisson–Gamma]	MAE	0.0338	121,727.67
	RMSE	0.1421	230,341.28

The mean absolute error (MAE) and root mean squared error (RMSE) metrics provide comprehensive insights into the predictive performance of different machine learning models (see Figures 6 and 7). Focusing on MAE as the crucial metric for evaluating accuracy, extreme gradient boosting (XGBoost) showcases the lowest MAE among the models, standing at 0.0309. This signifies that, on average, XGBoost's predictions deviate by

approximately 0.0309 units from the actual values, making it the most accurate in this aspect. Gradient boosting machines (GBMs) follow with a slightly higher MAE of 0.0328, indicating a slightly larger average absolute difference between their predictions and the ground truth compared to extreme gradient boosting (XGBoost). Despite the marginally higher MAE, gradient boosting machines (GBMs) still demonstrate commendable predictive accuracy. Generalized linear models (Poisson) exhibit the highest MAE among the models at 0.0338. While this value is slightly larger than extreme gradient boosting (XGBoost) and gradient boosting machines (GBMs), it is important to note that the difference in MAE is relatively small, indicating a GLM's overall ability to make predictions with a slightly larger average absolute deviation from the actual values. Moving on to RMSE, it complements the MAE by providing insights into the spread or dispersion of errors. Extreme gradient boosting (XGBoost) has the highest RMSE among the models, indicating a broader spread of errors despite having the lowest MAE. Conversely, a generalized linear model (Poisson) has the lowest RMSE, suggesting a tighter clustering of prediction errors compared to extreme gradient boosting (XGBoost) and gradient boosting machines (GBMs).

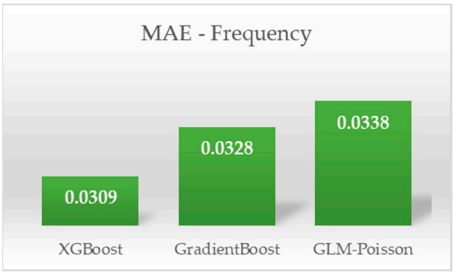

Figure 6. MAE frequency.

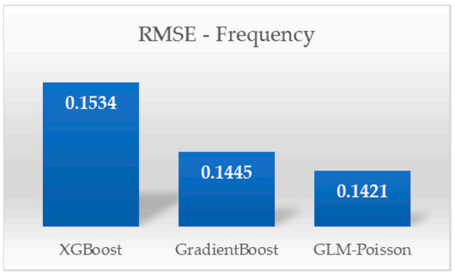

Figure 7. RMSE frequency.

In conclusion, while extreme gradient boosting (XGBoost) demonstrates the lowest MAE, implying better accuracy in predictions, it also shows higher RMSE values, suggesting a wider spread of errors. Gradient boosting machines (GBMs) closely follow extreme gradient boosting (XGBoost) in terms of MAE, and generalized linear models (Poisson) rank slightly higher in MAE but showcase the best RMSE values among the models, reflecting a smaller spread and magnitude of errors. Overall, the choice of the optimal model might depend on the specific emphasis on either absolute accuracy (MAE) or the distribution and magnitude of errors (RMSE) for the task at hand.

Figures 8 and 9 provide the MAE and RMSE severity for the different machine learning models. Focusing first on MAE, extreme gradient boosting emerges as the model with the lowest MAE at 113,085.07, showcasing the superior accuracy of this model compared to gradient boosting machines (MAE: 116,708.85) and generalized linear models (Gamma) (MAE: 121,727.67). Moving to RMSE, it provides insights into the spread of errors. Despite XGBoost's lower MAE, it presents a higher RMSE of 234,877.46, suggesting a wider dispersion of prediction errors compared to the other models. In contrast, generalized linear models (Gamma) exhibit the lowest RMSE (230,341.28), indicating a tighter clustering of prediction errors, and gradient boosting machines (GBMs) have the highest RMSE (237,625.14).

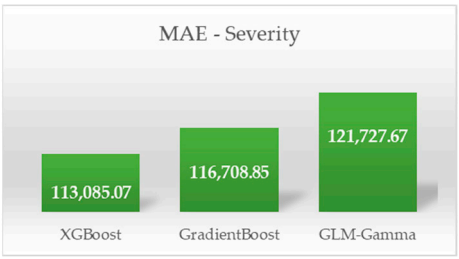

Figure 8. MAE severity.

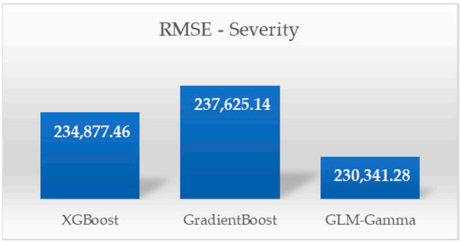

Figure 9. RMSE severity.

In summary, while extreme gradient boosting (XGBoost) demonstrates the lowest MAE, implying better accuracy in predictions, it also showcases higher RMSE values, suggesting a wider spread and larger magnitudes of errors. Conversely, generalized linear models (Gamma) exhibit the lowest RMSE, reflecting tighter clustering and smaller magnitudes of errors despite a slightly higher MAE. Gradient boosting machines fall between the two models in terms of MAE but have the highest RMSE.

6. Discussion

6.1. Frequency Predictive Performance

6.1.1. Extreme Gradient Boosting (XGBoost) in Predicting Claim Frequency

The MAE of 0.0309 for extreme gradient boosting (XGBoost) indicates a high degree of accuracy in its predictions for claim frequency, with the model's forecasts deviating minimally by approximately 0.0309 units from the actual observed values. Such precision underscores XGBoost's proficiency in closely mirroring actual claim frequency, a testament to its advanced ensemble learning capabilities. These capabilities enable extreme gradient boosting (XGBoost) to adeptly navigate and model the intricate nonlinear patterns within the data, a feature particularly advantageous in the realm of insurance. In this sector, the precision of claim frequency predictions is crucial, directly influencing risk assessment and the determination of premiums. XGBoost's ability to deliver predictions with such a minimal average deviation not only signifies its accuracy but also its significant potential in enhancing the reliability and effectiveness of insurance operations.

The RMSE value of 0.1534, when compared with the lower MAE, reveals a more nuanced aspect of XGBoost's predictive performance. Although the average deviation from the actual claim frequency is modest, as indicated by the MAE, the larger RMSE points to a broader range of prediction errors. This suggests that while extreme gradient boosting (XGBoost) generally provides predictions that are close to the actual values, there are instances where the model's forecasts exhibit more substantial discrepancies. This broader error distribution, captured by the RMSE, hints at the model's varying degrees of precision across different instances, possibly reflecting its sensitivity to outliers or anomalies within the frequency data. While extreme gradient boosting (XGBoost) demonstrates a robust ability to track the central trend of the data accurately, the RMSE underscores the presence of outliers or extreme cases where the model's predictions diverge more significantly from the observed values. This duality in XGBoost's performance—high accuracy on average coupled with a propensity for larger errors in specific cases—provides a comprehensive view of its predictive capabilities and limitations.

Application and Considerations:

XGBoost's strength is that it excels at providing accurate predictions for claim frequency on average (low MAE). This accuracy is valuable for insurance companies to estimate risk accurately. But XGBoost's wider spread of errors (higher RMSE) highlights the importance of considering potential outliers or instances where predictions significantly differ from actual claim frequency, despite the model's overall accuracy.

In summary, for claim frequency prediction, extreme gradient boosting (XGBoost) showcases relatively accurate predictions on average (low MAE), but the wider spread of errors (higher RMSE) implies occasional instances of larger deviations in predicting claim frequency from the observed values. This nuanced understanding is crucial for risk assessment in insurance applications, where precision in predicting claim frequency is vital.

6.1.2. Gradient Boosting Machines (GBMs) in Predicting Claim Frequency

The MAE of 0.0328 for gradient boosting machines (GBMs) demonstrates that the model's predictions for claim frequency typically diverge from the actual observed values by an average of 0.0328 units. This figure represents the mean absolute deviation and sheds light on the general magnitude of error in a GBM's frequency predictions.

This MAE indicates that gradient boosting machines (GBMs) deliver predictions with a proximity to actual values that is comparable to extreme gradient boosting (XGBoost), with a marginally higher deviation on average. This difference may stem from the inherent characteristics of a GBM's gradient-boosting framework, which, while robust, might not fully encapsulate the complexity of the data to the same extent as extreme gradient boosting (XGBoost), as reflected in the MAE.

The RMSE of 0.1445, reflecting the root mean squared error for gradient boosting machines (GBMs), quantifies the spread of prediction errors by taking the square root of the average squared discrepancies between the predicted and actual frequency values. Although the MAE portrays a relatively precise average prediction accuracy, the RMSE value points to a wider dispersion of errors, suggesting variability in the model's predictions that includes both minor and more substantial deviations from the actual observed frequencies.

Interestingly, while the RMSE value mirrors a similar pattern of error distribution as seen with extreme gradient boosting (XGBoost), it is marginally lower, implying a slightly tighter clustering of errors. This observation could indicate that gradient boosting machines (GBMs) possess a certain degree of robustness against outliers or extreme values, potentially offering more stable predictions across a broader spectrum of data. However, this comes with the trade-off of a slightly less precise capture of the central trend, as denoted by the marginally higher MAE in comparison to extreme gradient boosting (XGBoost).

Application and Considerations:

A GBM's predictions, on average, deviate by 0.0328 units. The higher RMSE implies a wider spread of errors, emphasizing the presence of larger deviations beyond the average error, which might have significant implications for risk assessments or financial estimations.

Gradient boosting machines (GBMs) demonstrate an average absolute deviation of approximately 0.0328 units in predicting claim frequency, reflecting the typical error made by the model. Despite the relatively low MAE, the higher RMSE of 0.1445 suggests a wider spread of errors, indicating occasional larger deviations from the observed frequency values beyond the average error. While gradient boosting machines (GBMs) offer reasonably accurate predictions on average, the presence of occasional larger deviations beyond the average error needs consideration for risk assessment or financial estimations related to claim frequency.

In essence, gradient boosting machines (GBMs) exhibit a relatively low average absolute deviation in claim frequency prediction but show a wider spread of errors beyond the average deviation.

6.1.3. Generalized Linear Model (Poisson) in Predicting Claim Frequency

The MAE of 0.0338 for the generalized linear model (Poisson) denotes an average deviation in its predictions for claim frequency, with forecasts typically straying by about 0.0338 units from the actual observed values. This figure, reflecting the mean absolute deviation, offers a look into the standard level of error associated with the model's frequency predictions.

When contrasted with the performance of extreme gradient boosting (XGBoost) and gradient boosting machines (GBMs), the marginally higher MAE for the generalized linear model (Poisson) subtly implies its slightly reduced precision in capturing the intricacies of the dataset. The inherent linear structure of the generalized linear model (Poisson) might contribute to this, potentially restraining its effectiveness in fully grappling with the complex dynamics embedded within the claim frequency data.

The RMSE of 0.1421 for the generalized linear model (Poisson) illuminates the degree of spread in the model's prediction errors, revealing the variability in how much the predicted values deviate from the actual frequency observations. This measure, indicating the square root of the average squared discrepancies, points to a range of error magnitudes, covering both minor and more pronounced deviations.

Interestingly, the Poisson model's RMSE, which is relatively lower compared to that of extreme gradient boosting (XGBoost) and gradient boosting machines (GBMs), hints at a more concentrated distribution of errors. This tighter clustering suggests enhanced model robustness, particularly in mitigating the impact of outliers or extreme data points. Such a characteristic is often inherent to linear models like Poisson, renowned for their stability and consistent error behavior across a spectrum of data scenarios.

Application and Considerations:

Generalized linear model (Poisson) predictions, on average, deviate by 0.0338 units. Understanding this average error magnitude is crucial for evaluating model performance in claim frequency prediction.

Generalized linear models (Poisson) demonstrate an average absolute deviation of approximately 0.0338 units in predicting claim frequency, reflecting the typical error made by the model. While a generalized linear model (Poisson) provides reasonably accurate predictions on average, the presence of occasional larger deviations beyond the average error needs consideration in risk assessment related to claim frequency.

In summary, generalized linear models (Poisson) show a relatively low average absolute deviation in claim frequency prediction, as indicated by their MAE of 0.0338. Additionally, their lower RMSE compared to XGBoost and GBMs suggests a tighter clustering of errors. This implies that while there are deviations in predictions, they are generally consistent and less variable, highlighting the Poisson model's reliability in capturing the central trend of claim frequency data.

6.2. Severity Predictive Performance

6.2.1. Extreme Gradient Boosting (XGBoost) in Predicting Claim Severity

The MAE of 113,085.07 for extreme gradient boosting (XGBoost) signifies that the model's severity predictions, on average, stray from the actual observed severity values by this magnitude. This metric offers a quantified look into the model's typical deviation when estimating claim severities.

While extreme gradient boosting (XGBoost) demonstrates a capacity to gauge the severity of claims with an average deviation of approximately 113,085.07, this figure also highlights the inherent challenges in severity prediction. The model's proficiency in modeling complex and nonlinear data patterns contributes to its predictive capability. Nonetheless, the substantial size of the average deviation underscores the intricate nature of severity prediction, a domain often characterized by its high variability and larger numeric scale. This divergence indicates that, although extreme gradient boosting (XGBoost) is

adept at capturing trends, the precision of its severity predictions may not mirror the tightness observed in its frequency predictions.

The RMSE of 234,877.46, compared with the lower MAE, paints a picture of considerable variability in XGBoost's severity predictions. While the MAE reflects a more focused deviation of around 113,085.07, the substantially higher RMSE reveals a broader spectrum of errors, indicating that the model's predictions are not uniformly close to the actual values. This disparity suggests that alongside the average deviations, there are scenarios where the model's estimates veer significantly farther from the actual severity figures, introducing a wider range of error magnitudes.

This RMSE underscores the model's fluctuating precision, particularly in the context of severity predictions where outliers or extreme values are more prevalent. Such high variability in prediction errors necessitates a prudent approach to interpreting the model's outputs, especially when dealing with claims of higher value where the financial stakes are substantial. The RMSE serves as a reminder of the inherent complexities and potential volatilities in modeling claim severities, advocating for a cautious and comprehensive understanding of the model's predictive behavior.

Application and Considerations:

This insight is essential for understanding the typical error magnitude. The higher RMSE suggests that there are larger deviations beyond the average error, emphasizing the presence of occasional predictions that significantly differ from the actual severity values.

In summary, extreme gradient boosting (XGBoost) demonstrates an average deviation of approximately 113,085.07 in predicting severity, but the wider RMSE of 234,877.46 indicates variability in errors, including instances of larger deviations beyond the average error.

6.2.2. Gradient Boosting Machines (GBMs) in Predicting Claim Severity

The MAE of 116,708.85 for gradient boosting machines (GBMs) reflects that the model's severity predictions, on average, diverge from the actual observed severity values by this amount. This metric elucidates the standard deviation from accuracy in a GBM's severity predictions, providing a clear indication of the model's average error magnitude.

This MAE suggests a level of prediction accuracy for claim severity that is comparably close to that of extreme gradient boosting (XGBoost), signifying that gradient boosting machines (GBMs) also offer a credible estimation of claim severity. The marginally higher MAE observed in gradient boosting machines (GBMs), relative to extreme gradient boosting (XGBoost), might be reflective of the model's intrinsic structure and its interaction with the complexity inherent in the severity data, potentially impacting its ability to fully encapsulate certain intricate data patterns.

The RMSE of 237,625.14 for gradient boosting machines (GBMs), slightly surpassing that of extreme gradient boosting (XGBoost), reveals an extended range of prediction errors. This metric underscores the variance in a GBM's predictions, indicating that, although the model's average deviation is approximately 116,708.85, there are instances where its predictions significantly overshoot or undershoot the actual severity values. The elevated RMSE underscores the existence of larger-than-average deviations, painting a picture of a broader error spectrum for a GBM's severity predictions.

This marginally higher RMSE compared to extreme gradient boosting (XGBoost) implies a wider dispersion of errors within a GBM's predictions. It denotes that while gradient boosting machines (GBMs) generally align closely with the actual values, mirroring the central trend of the data, they are not immune to considerable deviations, especially in cases involving higher-severity claims. This tendency suggests that gradient boosting machines (GBMs), similar to extreme gradient boosting (XGBoost), adeptly capture the overarching pattern in the data but may encounter challenges in accurately predicting the outliers or the more extreme values within the severity spectrum.

Application and Considerations:

Understanding this average error magnitude is essential for evaluating model performance. The higher RMSE indicates a wider spread of errors, underscoring the presence of occasional predictions that significantly differ from the actual severity values. Recognizing the wider spread of errors beyond the average deviation is crucial for risk assessments. Instances of larger deviations might have significant implications for financial risk estimations or insurance-related decisions. Insights from RMSE and MAE guide model evaluation and refinement efforts, highlighting areas where the model's performance deviates significantly and aiding in improving prediction accuracy.

In essence, gradient boosting machines (GBMs) exhibit an average deviation in predicting severity, but the higher RMSE indicates a wider range of errors, emphasizing the necessity of understanding the broader spectrum of errors beyond the average deviation for a comprehensive evaluation of the model's predictive performance.

6.2.3. Generalized Linear Model (Gamma) in Predicting Claim Severity

The mean absolute error (MAE) of 121,727.67 for the generalized linear model (Gamma) reveals that the model's predictions for severity typically differ from the actual observed values by about 121,727.67 units. This value reflects the usual scale of prediction error for severity made by the model. The MAE of the generalized linear model (Gamma) quantifies the mean absolute discrepancy, shedding light on the usual size of the prediction errors.

The indicated MAE points to a marginally greater average prediction discrepancy for the generalized linear model (Gamma) compared to machine learning models. This increased deviation may stem from the linear framework of the model, potentially restricting its capability to accurately represent the intricate variations found within claim severity data.

The root mean squared error (RMSE) of 230,341.28, despite the elevated MAE, indicates that prediction errors are relatively more compactly grouped. This RMSE value reflects how the errors are distributed, indicating that although the average discrepancy is about 121,727.67, the errors tend to be more closely bunched around the actual severity values compared to those from extreme gradient boosting (XGBoost) and gradient boosting machines (GBMs). Despite the slightly higher MAE, the RMSE signifies a denser aggregation of errors, hinting at a uniform pattern of deviations closely encircling the observed severity figures.

The RMSE for the generalized linear model (Gamma) is less than those for extreme gradient boosting (XGBoost) and gradient boosting machines (GBMs), denoting a more condensed clustering of prediction errors. This implies that while the generalized linear model (Gamma) may have constraints in precisely capturing the most extreme severity claims, it tends to offer more consistent predictions across different severity levels.

Application and Considerations:

Understanding this average error magnitude is crucial for evaluating model performance. The lower RMSE indicates a more clustered distribution of errors, emphasizing the model's consistent performance in predicting severity, despite a slightly higher average deviation.

Recognizing the more clustered distribution of errors is essential for risk assessments. The model's consistent performance around the observed severity values could be advantageous in certain risk estimation scenarios. Insights from RMSE and MAE facilitate model comparisons and guide improvements, highlighting areas where the model can be refined for better predictive accuracy.

In summary, generalized linear models (Gamma) exhibit an average deviation in predicting severity, but the lower RMSE suggests a more concentrated clustering of errors around the observed severity values. This concentrated error distribution, despite a slightly higher average deviation, indicates a consistent prediction pattern centered around the observed severity, offering insights crucial for decision-making and model enhancement.

6.3. Model Performance Evaluation

Assessing machine learning models through mean absolute error (MAE) and root mean squared error (RMSE) metrics offers vital insights into predictive accuracy and the distribution of prediction errors. Our analysis centered on three models—extreme gradient boosting (XGBoost), gradient boosting machines (GBMs), and generalized linear models (GLMs)—across two distinct datasets, each highlighting either frequency or severity as the focal point.

6.3.1. Frequency Predictive Performance Comparison

In analyzing the performance metrics of extreme gradient boosting (XGBoost), gradient boosting machines (GBMs), and generalized linear models (Poisson) for claim frequency prediction, extreme gradient boosting (XGBoost) emerges as the most favorable model based on a comprehensive assessment of MAE and RMSE metrics.

Extreme gradient boosting (XGBoost) exhibits a lower MAE, signifying closer predictions to actual frequency values on average. The model's accuracy (lower MAE) is crucial in insurance contexts for precise risk assessment and premium calculations. While extreme gradient boosting (XGBoost) showcases a wider spread of errors (higher RMSE), its accuracy (lower MAE) outperforms both gradient boosting machines (GBMs) and generalized linear models (Poisson). XGBoost's overall performance strikes a balance between providing accurate predictions (lower MAE) and acknowledging the potential for larger deviations (higher RMSE). XGBoost's prioritization of a lower MAE showcases its precision in predicting claim frequency, crucial for risk assessments in insurance. Despite the higher RMSE indicating occasional larger deviations, XGBoost's consistently accurate predictions (lower MAE) make it an optimal choice. Given its superior accuracy reflected in the lower MAE, extreme gradient boosting (XGBoost) is the preferred model for claim frequency prediction. Its ability to provide close predictions to actual frequency values on average outweighs occasional larger deviations indicated by the wider spread of errors (higher RMSE), making it a robust and dependable choice for insurance applications where precision is paramount.

6.3.2. Severity Predictive Performance Comparison

Analyzing the performance metrics of extreme gradient boosting (XGBoost), gradient boosting machines (GBMs), and the generalized linear model (Gamma) for claim severity prediction reveals compelling insights, suggesting that generalized linear models (Gamma) stand out as the preferred model based on a comprehensive evaluation of MAE and RMSE metrics.

In severity prediction, an RMSE holds more significance than an MAE as it accounts for the spread and variability of errors, which is crucial for assessing prediction reliability. Extreme gradient boosting (XGBoost) shows a lower MAE but a notably wider spread of errors (higher RMSE), indicating occasional larger deviations from actual severity values. Gradient boosting machines (GBMs) demonstrate a similar trend with a slightly higher MAE and a higher RMSE, suggesting a broader range of errors. Despite the generalized linear model (Gamma) showing a slightly higher MAE, it exhibits a notably lower RMSE, indicating a more clustered distribution of errors around observed severity values. The lower RMSE of the generalized linear model (Gamma) implies a more clustered error distribution around actual severity values, highlighting consistent predictions compared to extreme gradient boosting (XGBoost) and gradient boosting machines (GBMs). Despite a slightly higher average deviation (MAE), the generalized linear model (Gamma)'s concentrated error distribution ensures more reliable predictions closer to observed severity values.

The generalized linear model (Gamma) outperforms both extreme gradient boosting (XGBoost) and gradient boosting machines (GBMs) in a balanced evaluation considering multiple performance metrics. The generalized linear model (Gamma) showcases a slightly higher average deviation (MAE) compared to extreme gradient boosting (XGBoost) and gradient boosting machines (GBMs). However, its strength lies in maintaining a notably lower

spread of errors (RMSE). This balance between metrics is crucial in assessing predictive reliability. The generalized linear model (Gamma)'s ability to exhibit a more clustered error distribution around observed severity values, despite a slightly higher average deviation, underlines its consistency and reliability in predicting severity. In scenarios demanding precise estimations, this model's capacity to offer predictions with a more reliable and stable pattern proves pivotal for informed decision-making. This balanced performance makes the generalized linear model (Gamma) the optimal model among the three, showcasing reliability and stability in severity prediction tasks.

6.3.3. Choice of Model

In the realm of managing features like Status of Goods, Cargo Type, Packaging Type, Start Country, and Destination Country, extreme gradient boosting (XGBoost), generalized linear models (GLMs), and gradient boosting machines (GBMs) each exhibit distinct strengths. Extreme gradient boosting (XGBoost), celebrated for its prowess in capturing intricate interactions and nonlinear patterns, stands out for its robustness in navigating the multifaceted risk elements inherent in insurance claim predictions. Meanwhile, gradient boosting machines (GBMs), like extreme gradient boosting (XGBoost), excel at handling complex relationships due to their boosting of weak learners sequentially. In contrast, generalized linear models (GLMs), while proficient in managing categorical data, may lack extreme gradient boosting (XGBoost) and a GBM's finesse in capturing nuanced nonlinear relationships. However, a generalized linear model (GLM)'s linear nature presents a benefit in terms of simplicity, offering better interpretability and ease of implementation when complex nonlinear patterns are less critical for accurate predictions.

Extreme gradient boosting (XGBoost) offers superior accuracy (lower MAE) in predicting claim frequency, essential for insurance risk assessment and premium calculations. Despite a wider error spread (higher RMSE), its precision in average predictions makes it a robust choice.

The generalized linear model (Gamma) stands out as the optimal model despite a slightly higher average deviation (MAE) due to its notably lower spread of errors (lower RMSE). Its clustered error distribution around observed severity values signifies reliability and consistency in predictions. The selection of the optimal model is not solely based on a single metric but rather on a balance between multiple metrics, especially in contexts like severity prediction, where both the MAE and RMSE contribute to a comprehensive evaluation.

The selection of the optimal model is not solely based on a single metric but rather on a balance between multiple metrics, especially in contexts like severity prediction, where both the MAE and RMSE contribute to a comprehensive evaluation. While extreme gradient boosting (XGBoost) excels in frequency prediction with its accuracy, the generalized linear model (Gamma) shines in severity prediction due to its consistent and reliable performance, especially in maintaining a clustered error distribution around observed severity values. Therefore, considering the distinct requirements in frequency and severity predictions, extreme gradient boosting (XGBoost) and the generalized linear model (Gamma) emerge as the optimal choices, each excelling in different aspects crucial for their respective applications.

6.4. Potential Beneficiaries of Predictive Model Performance in Insurance: Leveraging Insights for Risk Assessment and Decision-Making

The comprehensive analysis of predictive model performance in insurance, covering frequency and severity predictions, offers more than just model selection. These insights refine insurers' risk assessment, inform regulatory policies, and bolster risk management strategies for cross-border entities, fostering stability and efficiency across multiple sectors.

6.4.1. Insurance Companies

Improved risk assessment by understanding which models perform better in predicting claim frequency and severity enables insurance companies to refine their risk

assessment processes. This insight aids in setting appropriate premiums, managing reserves, and mitigating potential losses by accurately forecasting claim occurrences and magnitudes. Accurate predictive models contribute to better pricing strategies, allowing insurance companies to set premiums that more accurately reflect the expected risk. This can lead to more competitive pricing for clients while maintaining profitability.

6.4.2. Government and Regulatory Bodies

Insights from predictive modeling can assist government bodies in understanding insurance trends, ensuring regulatory compliance, and making informed policy decisions. This understanding helps in creating a balanced regulatory environment that safeguards both insurers and insured parties.

6.4.3. Cross-Border Trade Entities (Trucking Companies, Shippers, and Consignees)

For entities involved in cross-border trade, knowing the predictive models for claim frequency and severity helps in understanding and mitigating potential risks associated with cargo transportation. This insight aids in better risk management, potentially leading to cost savings by identifying and addressing high-risk areas in logistics operations. Trucking companies, shippers, and consignees might benefit from more accurate predictive claims, resulting in more reasonable insurance premiums offered by insurance companies.

7. Conclusions

The pursuit of determining the optimal approach between generalized linear models (GLMs) and advanced machine learning (ML) in developing predictive models for claim frequency and severity within the domain of cargo insurance for cross-border transportation forms the core objective of this study. Utilizing mean absolute error (MAE) and root mean squared error (RMSE) as key evaluation metrics, this research aims to ascertain which method—GLMs or ML—more accurately predicts claim frequency and severity within this specialized insurance domain.

The metrics employed in the assessment, particularly MAE and RMSE, offer crucial insights into the accuracy and distribution of prediction errors. The evaluation focused on three models—extreme gradient boosting (XGBoost), gradient boosting machines (GBMs), and generalized linear models—across two distinct datasets, emphasizing either frequency or severity as focal points. The assessment revealed distinct performances for different predictive models in addressing these specific aspects.

In the realm of claim frequency prediction, extreme gradient boosting (XGBoost) emerged as the most favorable model, showcasing superior accuracy with a lower MAE. Despite a wider error spread (higher RMSE), its precision in average predictions proved it to be a robust choice for insurance applications that prioritize precision. However, in the prediction of claim severity, the generalized linear model (Gamma) showcased remarkable performance. Despite a slightly higher average deviation (MAE), it demonstrated notably lower spread errors (lower RMSE), ensuring a more clustered error distribution around observed severity values, signifying reliability and consistency in predictions.

The optimal model was not based solely on a single metric but rather on a balance between multiple metrics, notably MAE and RMSE, given their significant contributions to a comprehensive evaluation. While extreme gradient boosting (XGBoost) excelled in frequency prediction, showcasing accuracy, the generalized linear model (Gamma) stood out in severity prediction due to its consistent and reliable performance in maintaining a clustered error distribution around observed severity values. Therefore, considering the distinct requirements in frequency and severity predictions, extreme gradient boosting (XGBoost) and the generalized linear model (Gamma) emerged as the optimal choices, each excelling in different aspects crucial for their respective applications.

This study involves an in-depth analysis of predictive model performance in insurance, encompassing both frequency and severity predictions. This analysis not only aids insurers in refining risk assessment but also provides valuable insights for enhancing risk

management strategies for cross-border entities. These insights empower insurance companies to enhance risk assessment, set suitable premiums, manage reserves, and forecast claim occurrences accurately. This contributes to more competitive client pricing while ensuring profitability. For entities in cross-border trade, these insights aid in improved risk management, potentially leading to cost savings. Additionally, more accurate predictive claims may result in these entities receiving more reasonable insurance premiums from insurance companies.

7.1. Contribution

7.1.1. Comparison of Predictive Modeling Approaches

- Conducts a comprehensive comparison between generalized linear models (GLMs) and advanced machine learning techniques.
- Focuses specifically on claim frequency and severity in the cross-border cargo insurance sector.

7.1.2. Optimal Model Approach Identification

- Aims to identify the optimal modeling approach by evaluating model performance based on mean absolute error (MAE) and root mean squared error (RMSE) metrics.

7.1.3. Insights on Predictive Accuracy and Error Metrics

For Claim Frequency Prediction:

- Extreme gradient boosting (XGBoost) demonstrates higher predictive accuracy, as indicated by the lowest MAE.
- XGBoost shows higher RMSE values, suggesting a broader error spread compared to the generalized linear model (Poisson).
- The generalized linear model (Poisson) showcases the best RMSE values, indicating tighter error clustering and smaller error magnitudes.

For Severity Prediction:

- XGBoost exhibits the lowest MAE, implying superior accuracy.
- However, it also presents a higher RMSE, indicating wider error dispersion compared to the generalized linear model (Gamma).
- The generalized linear model (Gamma) demonstrates the lowest RMSE, portraying tighter error clustering and smaller error magnitudes, despite a slightly higher MAE.

7.1.4. Strategic Implications for Insurance Companies

- Findings enable insurers to refine risk assessment processes, set appropriate premiums, manage reserves, and accurately forecast claim occurrences.
- Contributes to competitive pricing strategies for clients while ensuring profitability for insurers.

7.1.5. Benefits for Cross-Border Trade Entities

- Insights aid trucking companies and cargo owners in improved risk management and potential cost savings.
- Enables more reasonable insurance premium settings based on accurate predictive claim models from insurance companies.

The contributions of this study are multifaceted, encompassing not only the comparative analysis of predictive models but also the practical implications of these findings in the domain of cargo insurance. The study's contribution lies in its systematic comparison and evaluation of well-established methodologies, helping stakeholders make informed decisions when selecting predictive models for claim frequency.

8. Limitations and Future Research Directions

8.1. Limitations

While the metrics used in this study, namely mean absolute error (MAE) and root mean squared error (RMSE), provided valuable insights into model performance, there are potential areas for further exploration. Future research might benefit from examining additional models or conducting more extensive hyperparameter tuning to potentially enhance metric performance. Exploring a broader spectrum of models or fine-tuning the parameters further could offer a deeper understanding of their comparative strengths and weaknesses. This avenue of study could potentially yield refined metrics, providing a more comprehensive assessment of predictive accuracy and reliability within the cargo insurance domain.

8.2. Future Research Directions

Future research endeavors in this domain could explore hybrid modeling techniques that integrate the strengths of generalized linear models (GLMs) and advanced machine learning (ML) methods. Hybrid models offer the promise of leveraging GLMs' interpretability while harnessing ML algorithms' predictive prowess, potentially leading to enhanced predictive accuracy in cross-border cargo insurance scenarios. Additionally, investigations could focus on tailored feature engineering and selection methods specific to the intricacies of the insurance domain. Incorporating domain knowledge and creating domain-specific features might substantially enhance model performance.

Further exploration into the impact of various features on claim frequency and severity could unveil hidden patterns crucial for more accurate predictions. Understanding the influence of temporal elements, such as trends and seasonality, through time-series analysis could refine predictive models by illuminating their impact on claim occurrences.

Moreover, investigating the influence of external factors, such as economic indicators or geopolitical events, on claim frequency and severity could deepen our understanding of insurance dynamics. Integrating external data sources into predictive models might lead to more adaptive and accurate predictions, further enhancing risk assessment and management in cross-border cargo insurance. These future research directions hold the potential to fortify predictive models, enabling a more comprehensive and precise evaluation of risks in the insurance landscape.

Author Contributions: Writing—Original Draft Preparation, P.P.; Supervision and validation, S.A. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Data Availability Statement: The dataset can be made accessible upon receiving a corresponding request.

Acknowledgments: This study owes its existence to the collaborative efforts of the School of Engineering at the University of the Thai Chamber of Commerce and the Insurance Premium Rating Bureau (IPRB), Thailand. Our sincere gratitude goes to all individuals whose support and contributions were instrumental in making this study possible.

Conflicts of Interest: The authors declare no conflicts of interest.

References

Abdelhadi, Shady, Khaled Elbahnasy, and Mohamed Abdelsalam. 2020. A Proposed Model to Predict Auto Insurance Claims using Machine Learning Techniques. *Journal of Theoretical and Applied Information Technology* 98: 3428–37.

Abhishek. 2023. Generalized Linear Models (GLMs). Medium. Available online: https://abhic159.medium.com/generalized-linear-models-glms-7b6e6c475d82 (accessed on 27 November 2023).

Ahlgren, Marcus. 2018. *Claims Reserving Using Gradient Boosting and Generalized Linear Models*. Stockholm: KTH Royal Institute of Technology.

Alimohammadi, Hamzeh, and Shengnan Nancy Chen. 2022. Performance Evaluation of Outlier Detection Techniques in Production Time Series: A Systematic Review and Meta-Analysis. *Expert Systems with Applications* 191: 116371. [CrossRef]

Alzubi, Jafar, Anand Nayyar, and Akshi Kumar. 2018. Machine Learning from Theory to Algorithms: An Overview. *Journal of Physics Conference Series* 1142: 012012. [CrossRef]

Avanijaa, Jangaraj. 2021. Prediction of House Price Using XGBoost Regression Algorithm. *Turkish Journal of Computer and Mathematics Education (TURCOMAT)* 12: 2151–55.

Bianco, Michael J., Peter Gerstoft, James Traer, Emma Ozanich, Marie A. Roch, Sharon Gannot, and Charles-Alban Deledalle. 2019. Machine Learning in Acoustics: Theory and Applications. *The Journal of the Acoustical Society of America* 146: 3590–628. [CrossRef] [PubMed]

Blier-Wong, Christopher, Hélène Cossette, Luc Lamontagne, and Etienne Marceau. 2020. Machine Learning in P&C Insurance: A Review for Pricing and Reserving. *Risks* 9: 4.

Chai, Tianfeng, and Roland R. Draxler. 2014. Root Mean Square Error (RMSE) or Mean Absolute Error (MAE)?—Arguments Against Avoiding RMSE in the Literature. *Geoscientific Model Development* 7: 1247–50. [CrossRef]

Chandola, Varun, Arindam Banerjee, and Vipin Kumar. 2009. Anomaly Detection: A Survey. *ACM Computing Surveys (CSUR)* 41: 1–58. [CrossRef]

Chasseray, Paul, Gauthier Eldin, and Aurégann Lefebvre. 2017. *P&C Reinsurance Modelling: Pure Premium Estimation and Creation of a Reinsurance Program*. Euro-Institut d'Actuariat AXA Global P&C. Brest: Université de Bretagne Occidentale, pp. 1–88.

Chen, Tianqi, and Carlos Guestrin. 2016. Xgboost: A Scalable Tree Boosting System. Paper presented at the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Francisco, CA, USA, August 13–17; pp. 785–94.

Chirathivat, Suthiphand, and Kornkarun Cheewatrakoolpong. 2015. *Thailand's Economic Integration with Neighboring Countries and Possible Connectivity with South Asia*. Tokyo: Asian Development Bank Institute.

Clemente, Carina, Gracinda R. Guerreiro, and Jorge M. Bravo. 2023. Modelling Motor Insurance Claim Frequency and Severity Using Gradient Boosting. *Risks* 11: 163. [CrossRef]

Cordeiro, Miguel Filipe Martins. 2023. *A Machine Learning Approach to Predict Health Insurance Claims*. Lisbon: Universidade Nova de Lisboa.

Dalal, Surjeet, Bijeta Seth, Magdalena Radulescu, Carmen Secara, and Claudia Tolea. 2022. Predicting Fraud in Financial Payment Services Through Optimized Hyper-Parameter-Tuned XGBoost Model. *Mathematics* 10: 4679. [CrossRef]

Deputy Prime Minister and Minister of Foreign Affairs of Thailand. 2023. *Intervention Delivered at the 8th Mekong-Lancang Co-operation Foreign Ministers' Meeting*; Bangkok: Ministry of Foreign Affairs, Kingdom of Thailand, December 8. Available online: https://www.mfa.go.th/en/content/mlcfmm2023-intervention-as-delivered-by-dpm-fm-2?page=5d5bd3dd15e39c306002ab20&menu=5f72d46f81ae194a461ef512 (accessed on 10 December 2023).

Dhieb, Najmeddine, Hakim Ghazzai, Hichem Besbes, and Yehia Massoud. 2019. Extreme Gradient Boosting Machine Learning Algorithm for Safe Auto Insurance Operations. Paper presented at the 2019 IEEE International Conference on Vehicular Electronics and Safety (ICVES), Cairo, Egypt, September 4–6; pp. 1–5.

Elith, Jane, John R. Leathwick, and Trevor Hastie. 2008. A working guide to boosted regression trees. *Journal of Animal Ecology* 77: 802–13. [CrossRef]

Fauzan, Muhammad Arief, and Hendri Murfi. 2018. The Accuracy of XGBoost for Insurance Claim Prediction. *International Journal of Advances in Soft Computing and Its Applications* 10: 159–71.

Freichel, Stephan L. K., Johannes K. Wö, Arthur Haas, and Lars ter Veer. 2022. Cargo Accumulation Risks in Maritime Supply Chains: A new perspective towards Risk Management for Theory, and Recommendations for the Insurance Industry and Cargo Shippers. *Logistics Research* 15: 4.

Friedman, Jerome H. 2001. Greedy Function Approximation: A Gradient Boosting Machine. *Annals of Statistics* 29: 1189–232. [CrossRef]

Ge, Jiankun, Linfeng Zhao, Zihui Yu, Huanhuan Liu, Lei Zhang, Xuewen Gong, and Huaiwei Sun. 2022. Prediction of Greenhouse Tomato Crop Evapotranspiration Using XGBoost Machine Learning Model. *Plants* 11: 1923. [CrossRef] [PubMed]

Hellestol, Tonje, and Petter Eriksen. 2022. The Impact of Machine Learning and Aggregated Data on Corporate Insurance Modelling: An Empirical Study on the Prospective Gains of Machine Learning Techniques Using New Data Sources in the Insurance Industry. Master's thesis, Norwegian School of Economics (NHH), Bergen, Norway.

Henckaerts, Roel, Marie-Pier Côté, Katrien Antonio, and Roel Verbelen. 2021. Boosting Insights in Insurance Tariff Plans with Tree-Based Machine Learning Methods. *North American Actuarial Journal* 25: 255–85. [CrossRef]

Hodson, Timothy O. 2022. Root-mean-square error (RMSE) or mean absolute error (MAE): When to use them or not. *Geoscientific Model Development* 15: 5481–7. [CrossRef]

Jha, H. S., A. Khanal, H. M. D. Seikh, and W. J. Lee. 2022. A Comparative Study on Outlier Detection Techniques for Noisy Production Data from Unconventional Shale Reservoirs. *Journal of Natural Gas Science and Engineering* 105: 104720. [CrossRef]

Kankanamge, Kusal D., Yasiru R. Witharanage, Chanaka S. Withanage, Malsha Hansini, Damindu Lakmal, and Uthayasanker Thayasivam. 2019. Taxi Trip Travel Time Prediction with Isolated XGBoost Regression. Paper presented at the 2019 Moratuwa Engineering Research Conference (MERCon), Moratuwa, Sri Lanka, July 3–5; pp. 54–59.

Ke, Guolin, Qi Meng, Thomas Finley, Taifeng Wang, Wei Chen, Weidong Ma, Qiwei Ye, and Tie-Yan Liu. 2017. LightGBM: A Highly Efficient Gradient Boosting Decision Tree. Paper presented at the Advances in Neural Information Processing Systems 30 (NIPS 2017), Long Beach, CA, USA, December 4–9.

Krungsri Research. 2022. Road Freight Transportation 2022–2024. Available online: https://www.krungsri.com/en/research/industry/industry-outlook/logistics/road-freight-transportation/io/road-freight-transportation-2022%E2%80%932024 (accessed on 2 December 2023).

Monemar, Magnus, and Erik Wallin. 2015. Premium Allocation for the Electrolux Cargo Insurance Program Using Generalized Linear Models. Available online: https://urn.kb.se/resolve?urn=urn:nbn:se:umu:diva-106904 (accessed on 18 October 2023).
Murekatete, Delphine. 2022. *Predicting Motor Insurance Claims—XGBoost versus Logistic Regression*. Kigali: African Institute for Mathematical Sciences (AIMS).
Nelder, John Ashworth, and Robert W. M. Wedderburn. 1972. Generalized Linear Models. *Journal of the Royal Statistical Society Series A: Statistics in Society* 135: 370–84. [CrossRef]
Poufinas, Thomas, Periklis Gogas, Theophilos Papadimitriou, and Emmanouil Zaganidis. 2023. Machine Learning in Forecasting Motor Insurance Claims. *Risks* 11: 164. [CrossRef]
Pugliese, Raffaele, Stefano Regondi, and Riccardo Marini. 2021. Machine Learning-based Approach: Global Trends, Research Directions, and Regulatory Standpoints. *Data Science and Management* 4: 19–29. [CrossRef]
Ridgeway, Greg. 2007. Generalized Boosted Models: A Guide to the GBM Package. *Update* 1: 2007.
Ritonga, Ali Imran, Kundori Kundori, Karolus G. Sengadji, and Hilda Emeraldo Ahmad. 2021. Optimizing the Process of Management of Marine Cargo Insurance Claims at PT. ABC. *Jurnal Logistik Indonesia* 5: 166–73. [CrossRef]
Socorro, Trujillo, and María Karina. 2019. International Marine Cargo Insurance: Building generic and thematic competences in commercial translation. *Journal of Specialised Translation* 32: 262–79.
Thai General Insurance Association. 2016. Manual for Practitioners in Actuarial Mathematics. Available online: https://www.tgia.org/upload/file_group/3/download_861.pdf (accessed on 18 July 2023).
Tuininga, Frits. 2022. A Machine Learning Approach for Modeling Frequency and Severity. Master's thesis, University of Twente, Enschede, UK.
Yang, Li, and Shami Abdallah. 2020. On hyperparameter optimization of machine learning algorithms: Theory and practice. *Neurocomputing* 415: 295–316. [CrossRef]
Yaro, Abdulmalik Shehu, Filip Maly, and Pavel Prazak. 2023. Outlier Detection in Time-Series Receive Signal Strength Observation Using Z-Score Method with S n Scale Estimator for Indoor Localization. *Applied Sciences* 13: 3900. [CrossRef]
Yousif, Jan Mikael. 2023. A Comparative Analysis between Various Machine Learning Models and Generalized Linear Models. Master's thesis, Stockholm University, Stockholm, Sweden.
Zhao, Xin, Qiushuang Li, Wanlei Xue, Yihang Zhao, Huiru Zhao, and Sen Guo. 2022. Research on ultra-short-term load forecasting based on real-time electricity price and window-based XGBoost model. *Energies* 15: 7367. [CrossRef]
Zhu, Ruijin, Weilin Guo, and Xuejiao Gong. 2019. Short-Term Photovoltaic Power Output Prediction Based on k-Fold Cross-Validation and an Ensemble Model. *Energies* 12: 1220. [CrossRef]