The Seybold Report
ISSN 1533-9211
DOI: 10.5281/zenodo.10053411

PREDICTING CARGO INSURANCE CLAIMS USING MACHINE LEARNING – A CASE STUDY OF THAILAND'S BORDER TRADE

PRAIYA PANJEE [1] and SATAPORN AMORNSAWADWATANA [2]*

[1, 2] School of Engineering, University of the Thai Chamber of Commerce, Thailand.
E-mail: [1]1810751101006@live4.utcc.ac.th, [2]sataporn_amo@utcc.ac.th (*Corresponding Author)

Abstract

The surge in cross-border trade in Thailand has led to a significant increase in shipments, subsequently raising the potential risks involved. As a result, both goods owners and carriers are compelled to explore effective measures to mitigate the impact in case of unforeseen incidents affecting their cargo. This has intensified their focus on obtaining comprehensive cargo insurance coverage. However, one ongoing issue that both policyholders and insurers have is precisely predicting whether a policyholder will submit a claim to determine a reasonable price for purchasing an insurance policy. The objective of this study is to evaluate and compare individual classifiers in order to determine which provides the most accurate predictions for cross-border freight insurance. This study looked at the relative performance of XGBoost, Logistic Regression, Light GBM, Gradient Boost, Catboost, and Random Forest approaches for predicting cargo insurance claims. The dataset comprises data sourced from The Insurance Premium Rating Bureau (IPRB) from 2016 to 2022 with a specific focus on road transportation in Thailand's border trade. The findings strongly indicate that GradientBoost is the superior model for handling cargo insurance claim predicting. It shows the best score in multiple metrics such as logloss, ROC AUC, precision, and accuracy.

Keywords: Cargo insurance, Risk prediction, Predictive model, Machine learning, XGBoost, GradientBoost

I. INTRODUCTION

Currently, road transport accounts for a major share of all logistics operations in the Association of Southeast Asian Nations (ASEAN)[1] The expansion of road transportation, particularly in the cross-border context within this region, is expected to be sustainable, aided by the implementation of new measures aimed at promoting interregional trade, economic recovery, and responding to the escalating demand for logistics solutions, which is being driven by the burgeoning e-commerce market. Notably, cross-border transportation is an important pillar in the European-South American logistics network, and routes connecting the United States to Africa and Asia are also important. Furthermore, the logistics solutions sector is constantly growing, and it plays a critical role in tackling complicated international long-distance transportation difficulties across Asia.

The dominance of road freight as the leading logistical option in the region is undeniable, indicating its rising popularity. This mode of transportation offers a cost-effective and sustainable option for international long-distance freight movements, made possible by the seamless deployment of cross-border trucking operations within the Asian landscape, enhancing operational efficiency and cost reduction in the logistics business. The absence of trade barriers in Asia, as well as the rise of the manufacturing industry, all contribute to road transportation's hopeful trajectory within the area.

The Seybold Report
ISSN 1533-9211
DOI: 10.5281/zenodo.10053411

Notably, the implementation of Road and Multimodal Solutions is likely to result in more enticing logistical choices in the ASEAN region. As East Asian economies are expected to drive trade growth in 2021, the ASEAN road freight industry is expected to increase at a rate greater than 8% per year from 2020 to 2025[1]

Despite the benefits of international goods transportation, it is critical to recognize the inherent hazards of moving goods across borders, as evidenced by accident statistics in Thailand, emphasizing the necessity for ongoing improvement in risk management methods. Cross-border traders must be proactive in limiting risks by obtaining product insurance, assuring adequate protection for the items or property being carried. Cargo insurance plays an important role in protecting products during transit, whether by sea, commercial airplane, or postal parcel, by providing coverage against a variety of potential mishaps, such as ship fires, sinkings, and damages sustained during the loading and unloading operations. The specific perils covered under cargo insurance depend on the coverage options selected by the insured. Considering the significant impact of cross-border transportation on businesses and trade, cargo insurance serves as an indispensable risk mitigation tool for effective origin-to-destination goods transportation.

Actuaries at insurance businesses are responsible for the assessment and determination of insurance premiums, insurance reserves, and risk analysis. These specialists use historical and current event analysis, as well as mathematical and statistical approaches, to model and anticipate future risk events. Accurate insurance premium computation, taking into account the related risks, is vital information that informs future premium judgments. The prediction of cargo claims is crucial to calculating insurance premiums in cargo insurance. However, as new artificial intelligence approaches emerge, the task of picking an acceptable model remains a continuous one. This article compares and contrasts the predictive capabilities of recently introduced approaches such as XGBoost, Logistic Regression, Light GradientBoost, GradientBoost, Catboost, and Random Forest. The historical insurance policy data from 2016 to 2022 were used in this study.

II. RESEARCH OBJECTIVES

The primary aim of this research is to conduct a comprehensive evaluation and comparison of various individual classifiers, seeking to ascertain which among them yields the most precise predictions in the domain of cross-border freight insurance. The study rigorously investigates the predictive capabilities of six distinct methodologies, namely XGBoost, Logistic Regression, Light GBM, GradientBoost, Catboost, and Random Forest. These models are subjected to rigorous scrutiny to determine their effectiveness in forecasting cargo insurance claims, a critical task within the realm of insurance underwriting and risk assessment.

In summary, this study endeavors to shed light on the most suitable modeling approaches for predicting cargo insurance claims, contributing to more effective risk management, and pricing strategies in the realm of cross-border freight insurance, thereby benefiting both insurers and insured parties alike.

The Seybold Report
ISSN 1533-9211
DOI: 10.5281/zenodo.10053411

III. THEORETICAL BACKGROUND AND LITERATURE REVIEW

Machine learning

Machine learning models serve as the backbone of data-driven decision-making processes in a wide array of domains. These models are computational algorithms that enable computers to autonomously learn from data and subsequently make predictions, classifications, or decisions without explicit programming. They are instrumental in identifying patterns, extracting meaningful insights, and delivering forecasts or classifications based on data patterns.

One of the pivotal distinctions in machine learning models lies in the dichotomy of supervised and unsupervised learning. In supervised learning, models are trained using labeled data, where input data is paired with corresponding correct outputs. This paradigm allows models to discern relationships between inputs and desired outcomes, rendering it suitable for tasks such as classification and regression. Common algorithms in supervised learning include decision trees, support vector machines, and neural networks[2]

In contrast, unsupervised learning deals with unlabeled data, concentrating on the identification of inherent patterns or structures within the data. Techniques such as clustering and dimensionality reduction are applied in unsupervised learning. For instance, K-means clustering and Principal Component Analysis (PCA) are key unsupervised learning methods[3]

Furthermore, machine learning models encompass various categories, including regression models for the prediction of continuous values, classification models for the categorization of data into predefined classes, and clustering models for grouping similar data points. Additionally, time series models are designed to handle sequential data, while reinforcement learning models excel at decision-making in dynamic environments[4]

In practice, the effectiveness of machine learning models hinges on comprehensive data preprocessing, feature engineering, and judicious model selection. Model evaluation is integral, involving the application of metrics such as accuracy, precision, recall, F1-score, and the ROC AUC curve, contingent on the nature of the problem and the model type.

Selecting the most suitable model is contingent upon the problem's characteristics and the available data. Researchers and practitioners frequently undertake a thorough exploration and validation of diverse models to identify the one that aligns most effectively with their objectives and data. Given the continuous evolution of the field, staying current with the latest techniques and best practices is essential for harnessing the full potential of this transformative technology[5]

Gradient Boosting

Gradient Boosting is a general ensemble learning method that builds a powerful predictive model by sequentially adding weak learners, typically decision trees, to the ensemble. It corrects the errors made by previous learners and minimizes a loss function through gradient descent. Gradient boosting is widely recognized for its ability to improve predictive accuracy and is a fundamental concept that forms the basis of algorithms like XGBoost and CatBoost[6]

The Seybold Report
ISSN 1533-9211
DOI: 10.5281/zenodo.10053411

the Gradient Boosting Classifier approach is employed for classification. It is an ensemble technique that can work with small data[7]

XGBoost (Extreme Gradient Boosting)

XGBoost is an ensemble learning method that extends the concept of gradient boosting. It combines the predictions of multiple weak learners, typically decision trees, to create a robust and high-performance predictive model. XGBoost focuses on optimizing the model's performance by minimizing a loss function through gradient descent. It incorporates regularization techniques to prevent overfitting and provides excellent speed and efficiency, making it a popular choice for various machine learning tasks[8] XGBoost employs a boosting technique to progressively build a sequence of decision trees. It starts by training the first tree to predict the outcome, and then, in subsequent iterations, it trains additional trees using the residuals obtained from the previous predictions. This iterative process allows XGBoost to enhance its predictive accuracy over time[9]

CatBoost

CatBoost, short for Categorical Boosting, is a gradient boosting algorithm designed to address the challenges posed by categorical features. It uses ordered boosting and oblivious trees to enhance predictive performance. CatBoost includes strategies for optimizing hyperparameters automatically, and it handles categorical data efficiently without the need for extensive preprocessing. This algorithm is user-friendly and effective for a wide range of machine learning applications[10]

LightGBM (Light Gradient Boosting Machine)

LightGBM is another gradient boosting framework that emphasizes efficiency and speed. It employs a histogram-based learning approach, which significantly reduces memory usage and accelerates the training process. LightGBM is well-suited for large datasets and high-dimensional feature spaces. It's designed to handle categorical features efficiently and has become a popular choice for machine learning competitions and applications[11]

Random Forest

Random Forest is an ensemble learning method that combines the predictions of multiple decision trees to create a robust and high-performing predictive model. It introduces randomness in the feature selection and bootstrapping processes, which helps reduce overfitting and enhance model accuracy. Each decision tree in the ensemble is trained on a different subset of the data, and their predictions are combined through voting (for classification) or averaging (for regression). This ensemble approach improves predictive performance and generalization[12] The Random Forest algorithm is an enhancement of the decision tree supervised machine learning approach[13]

Logistic Regression:

Logistic Regression is a fundamental statistical model used for binary and multi-class classification tasks. It models the probability of a binary outcome based on one or more

The Seybold Report
ISSN 1533-9211
DOI: 10.5281/zenodo.10053411

predictor variables. The logistic function, also known as the sigmoid function, is used to map the linear combination of input features to a value between 0 and 1, representing the probability of the positive class. Logistic Regression is interpretable, making it a valuable tool for understanding the relationships between predictor variables and the probability of an event occurring[14]

These machine learning algorithms play significant roles in insurance data analytics, offering various strengths and capabilities to address the diverse challenges and tasks within the insurance industry. Researchers and practitioners leverage these algorithms to enhance risk assessment and more, ultimately improving the efficiency and accuracy of insurance processes.

The Confusion Matrix

Confusion matrix

The confusion matrix is an important tool in the field of machine learning, largely used to evaluate the performance of classification models. It allows for the assessment of a model's ability to appropriately categorize situations, providing significant information into its strengths and faults. Although the confusion matrix has no single point of genesis, it is internationally recognized and widely used in the fields of machine learning and statistics[16]

A confusion matrix is employed in binary classification problems to effectively discern the accuracy of predicted class outputs. It serves as a valuable tool for distinguishing between correct and incorrect predictions for each class[15]

Table 1: Confusion matrix

	Predicted Positive	**Predicted Negative**
Actual Positive	True positive (TP)	False negative (FN)
Actual Negative	False positive (FP)	True negative (TN)

A confusion matrix is a 2x2 table that categorizes the model's predictions and actual class labels into four categories:

1. True Positives (TP): The number of instances correctly classified as the positive class.
2. True Negatives (TN): The number of instances correctly classified as the negative class.
3. False Positives (FP): The number of instances incorrectly classified as the positive class (actual negatives incorrectly predicted as positives).
4. False Negatives (FN): The number of instances incorrectly classified as the negative class (actual positives incorrectly predicted as negatives).

The common performance metrics that can be calculated using the values from the confusion matrix.

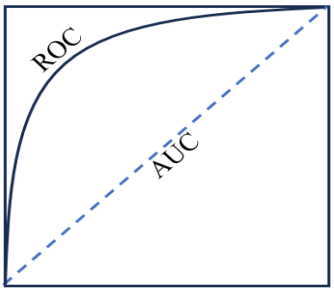
1. **Log Loss (Logarithmic Loss)** measures the performance of a classification model where the prediction output is a probability value between 0 and 1. The formula as below.

ISSN 1533-9211

DOI: 10.5281/zenodo.10053411

$$Logloss = -\frac{1}{n}\sum_{i=1}^{n}[y_i \log(\hat{y}_i) + (1 - y_i)\log(1 - \hat{y}_i)] \quad (19)$$

2. **ROC AUC (Receiver Operating Characteristic Area Under the Curve)** score can be probabilistically interpreted as follows: When you randomly select a positive case and a negative case, the AUC represents the probability that the positive case will have a higher predicted value than the negative case, based on the ranking of these cases according to the classifier's predictions.
3.



AUC Values	Test quality
0.9-1.0	Excellent
0.8-0.9	Very good
0.7-0.8	Good
0.6-0.7	Satisfactory
0.5-0.6	Unsatisfactory

(18)

4. **Precision** is the proportion of true positive predictions made by the model out of all positive predictions made by the model. The formula as below

$$\text{Precision} = \frac{TP}{TP+FP} \quad (17)$$

5. **Recall** (Sensitivity or True Positive Rate): Recall measures the proportion of true positive predictions among all actual positive instances. The formula as below

$$\text{Recall} = \frac{TP}{TP+FN} \quad (19)$$

6. **F1 Score** is a harmonic mean of precision and recall, balancing the trade-off between them. The formula as below

$$\text{F1 Score} = \frac{2 \cdot Precision \cdot Recall}{Precision + Recall} \quad (17)$$

7. **Accuracy** measures the ratio of accurate predictions in relation to the total number of instances assessed. The formula as below

$$\text{Accuracy} = \frac{TP+TN}{TP+FP+TN+FN} \quad (17)$$

ISSN 1533-9211

DOI: 10.5281/zenodo.10053411

Table 2: The description of the variables in the cargo insurance claims dataset.

Variables		Non-Occurance of Claims (Y = 0)	Occurance of Claims (Y = 1)	Total
Effective year	2016	1043	7	1050
	2017	982	13	995
	2018	1114	35	1149
	2019	1314	24	1338
	2020	1551	33	1584
	2021	1847	37	1884
	2022	1776	27	1803
Status of Goods	Import	2646	68	2714
	Export	6981	108	7089
Cargo type group	Group 1	3043	75	3118
	Group 2	24	2	26
	Group 3	1097	25	1122
	Group 4	90	0	90
	Group 5	386	13	399
	Group 6	4987	61	5048
Packaging Type	In bulk	33	0	33
	Carton/Box	2459	55	2514
	Case/Crate	658	9	667
	Tin/Drum	103	5	108
	Bag/Sack	357	12	369
	Pallet/Skid	803	33	836
	Bundle/Bale	695	0	695
	Roll/Coil	131	2	133
	Others	4388	60	4448
Start Country	Thailand	6971	109	7080
	Laos	142	2	144
	Myanmar	81	3	84
	Cambodia	100	5	105
	Malaysia	2333	57	2390
Destination Country	Thailand	2647	67	2714
	Laos	1961	47	2008
	Myanmar	2387	16	2403
	Cambodia	1023	22	1045
	Malaysia	1609	24	1633
Average of Suminsured Amount (THB)		19,731,401.60	22,134,837.36	21,984,622.62
Total		9627 (98%)	176 (2%)	9803 (100%)

The Seybold Report
ISSN 1533-9211
DOI: 10.5281/zenodo.10053411

Table 3: The description of cargo type group

Cargo type group	Description
Group 1	Electric, Electronic part/computer parts (as the component parts)
	Electrical Appliances/Computer (as the Finished Goods)
	Fertilizer
	Food & Seasoning
	Germ, Jewelry, Precious Stone, Precious Metal
	Hot / Cold Rolled Steel
	Leathers & Products
	Mineral & Ore
	Other Agriculture Product
	Other Steel/Metallic products
	Pharmaceutical product
	Plastic Resin
	Pulp & Paper
	Raw Cotton
	Rice
	Rubber (From Rubber Tree)
	Steel Rod
	Textile Product
	Timber / Wood
	Vehicles Parts
	Wooden products
Group 2	Beverage
Group 3	Machinery/Equipment & Parts
	Motorcycle in complete built-up
	Other vehicles in complete built-up
	Sedan Car in complete built-up
	Truck in complete built-up
Group 4	Frozen/Chilled/Refrigerated Cargo
Group 5	Chemical Product
	Other Petrochemical Product
	Petroleum
Group 6	Others
	Various

From table 3. It shows that the data categorizes the cargo types insured by a specific cargo insurance company in Thailand. Cargo insurance is a crucial aspect of the shipping and logistics industry, offering protection against the potential risks and uncertainties associated with transporting goods. By classifying and cataloging the cargo types covered by this insurance company, it provides a comprehensive view of the diverse range of products and commodities being safeguarded during transit.

By meticulously classifying these cargo types, the insurance company can better assess and mitigate risks, optimize its underwriting processes, and price policies more accurately. This dataset also aids in aligning insurance offerings with the unique needs and challenges associated with different cargo categories, thus enhancing its ability to provide comprehensive and tailored coverage to its clients.

ISSN 1533-9211

DOI: 10.5281/zenodo.10053411

IV. RESEARCH METHODS

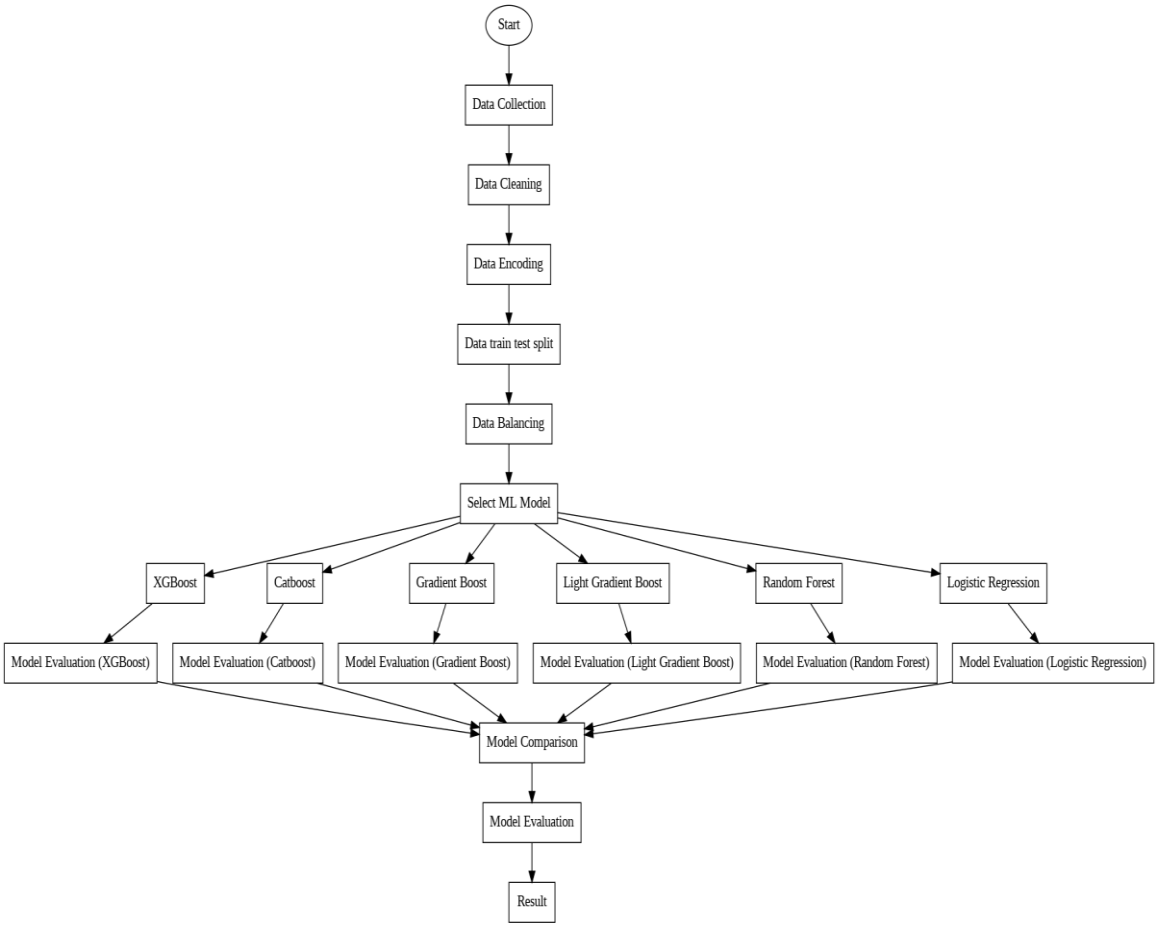


Figure 1: Research method

The study at hand revolves around an analysis of import/export land shipment data, generously provided by the Insurance Premium Rating Bureau [IPRB] This dataset covers the years from 2016 to 2022, offering a comprehensive glimpse into the dynamics of this industry over time. The dataset begins with the data cleaning process, conducted with the assistance of Microsoft Excel, to ensure data quality and reliability.

Since the primary objective is to classify whether an insurance claim exists, making this a binary classification problem. Six diverse machine learning models were employed to tackle this challenge: XGBoost Classifier, CatBoost Classifier, GradientBoost Classifier, LightGBM Classifier, Random Forest Classifier, and Logistic Regression. These models, each with its unique strengths and characteristics, were rigorously crafted to discern the presence or absence of insurance claims.

To assess the effectiveness and reliability of these models, a robust methodology is employed. For each model, a k-fold cross-validation process is executed, coupled with hyperparameter

ISSN 1533-9211

DOI: 10.5281/zenodo.10053411

tuning to optimize their performance. The dataset is subdivided into training and testing sets for each fold, with an 80% allocation to the training data and 20% to the testing data. The training data is then employed to fit the model, while the model's performance is rigorously evaluated on the independent test dataset.

In the context of claim count, which is intrinsically an imbalanced dataset (with only 2% of instances representing claims), a Synthetic Minority Over-sampling Technique (SMOTE) is thoughtfully employed. SMOTE is a valuable tool for augmenting the training dataset, ensuring that the model receives balanced and representative samples of both claim and no-claim instances. This method helps alleviate the class imbalance problem, thereby enhancing the model's capacity to discern claims accurately.

To assess the performance of these models, several key metrics are considered. For claim count, the metrics include Logloss, F1, ROC AUC, Precision, Recall, and Accuracy. These metrics collectively provide insights into the model's predictive capability, its ability to differentiate between claims and non-claims, and the trade-offs between precision and recall.

In summary, this comprehensive study is a testament to the power of machine learning and data-driven analytics in the insurance domain. The approach is not only robust but also carefully designed to account for data imbalances through techniques like SMOTE, ensuring that the models are both accurate and reliable in their assessments of claims. Through rigorous cross-validation and hyperparameter tuning, the study aims to produce models that can be of significant value to the insurance industry, enabling more accurate risk assessment, claims processing, and ultimately, better decision-making.

VI. RESEARCH RESULTS

Confusion matrix result

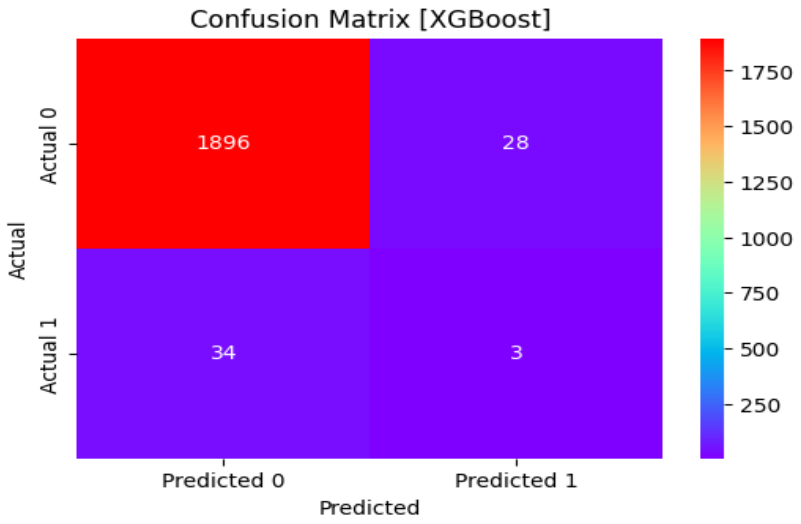


Figure 2: Confusion matrix – XGBoost

ISSN 1533-9211

DOI: 10.5281/zenodo.10053411

True Positives (TP): 3

These are the number of actual positive cases (Actual 1) correctly predicted as positive (Predicted 1). In this case, XGBoost correctly identified 3 positive cases.

True Negatives (TN): 1896

These are the number of actual negative cases (Actual 0) correctly predicted as negative (Predicted 0). XGBoost accurately identified 1896 negative cases.

False Positives (FP): 28

These are the number of actual negative cases (Actual 0) incorrectly predicted as positive (Predicted 1). XGBoost made 28 false positive predictions.

False Negatives (FN): 34

These are the number of actual positive cases (Actual 1) incorrectly predicted as negative (Predicted 0). XGBoost made 34 false negative predictions.

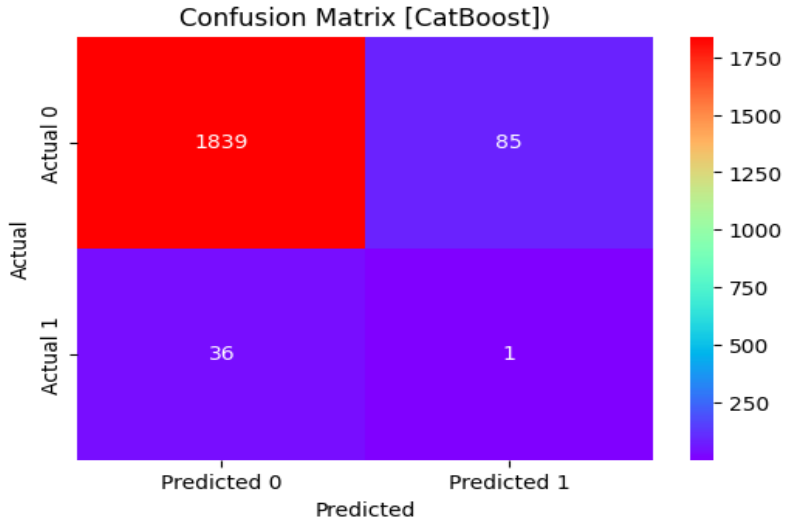


Figure 3: Confusion matrix – CatBoost

True Positives (TP): 1

These are the number of actual positive cases (Actual 1) correctly predicted as positive (Predicted 1). In this case, CatBoost correctly identified 1 positive case.

True Negatives (TN): 1839

These are the number of actual negative cases (Actual 0) correctly predicted as negative (Predicted 0). CatBoost accurately identified 1839 negative cases.

ISSN 1533-9211

DOI: 10.5281/zenodo.10053411

False Positives (FP): 85

These are the number of actual negative cases (Actual 0) incorrectly predicted as positive (Predicted 1). CatBoost made 85 false positive predictions.

False Negatives (FN): 36

These are the number of actual positive cases (Actual 1) incorrectly predicted as negative (Predicted 0). CatBoost made 36 false negative predictions.

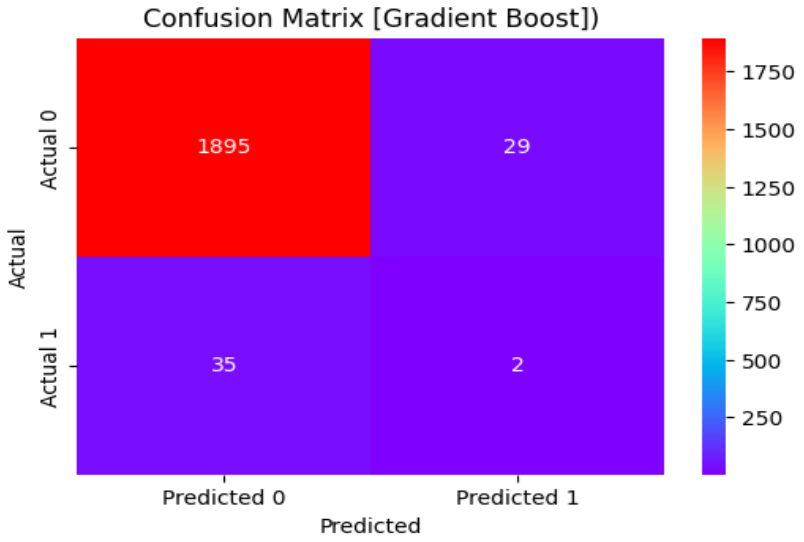


Figure 4: Confusion matrix – GradientBoost

True Positives (TP): 2

These are the number of actual positive cases (Actual 1) correctly predicted as positive (Predicted 1). In this case, Gradient Boosting correctly identified 2 positive cases.

True Negatives (TN): 1895

These are the number of actual negative cases (Actual 0) correctly predicted as negative (Predicted 0). Gradient Boosting accurately identified 1895 negative cases.

False Positives (FP): 29

These are the number of actual negative cases (Actual 0) incorrectly predicted as positive (Predicted 1). Gradient Boosting made 29 false positive predictions.

False Negatives (FN): 35

These are the number of actual positive cases (Actual 1) incorrectly predicted as negative (Predicted 0). Gradient Boosting made 35 false negative predictions.

ISSN 1533-9211

DOI: 10.5281/zenodo.10053411

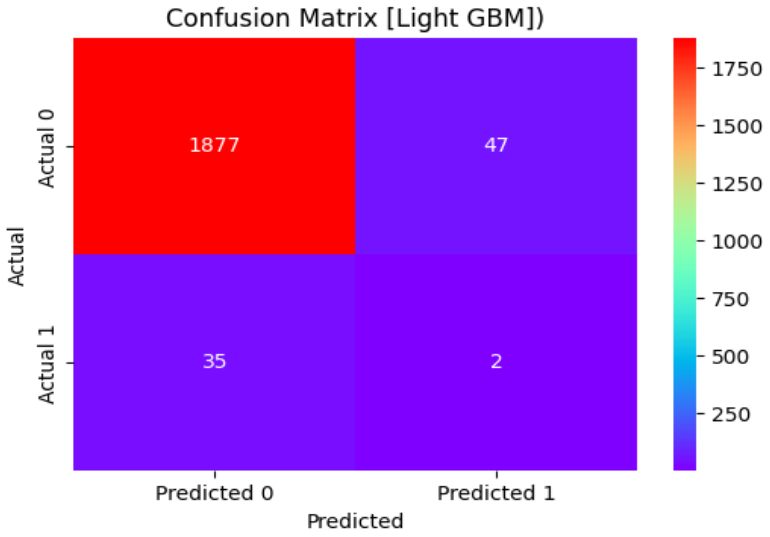


Figure 5: Confusion matrix – Light GBM

True Positives (TP): 2

These are the number of actual positive cases (Actual 1) correctly predicted as positive (Predicted 1). In this case, Light Gradient Boosting correctly identified 2 positive cases.

True Negatives (TN): 1877

These are the number of actual negative cases (Actual 0) correctly predicted as negative (Predicted 0). Light Gradient Boosting accurately identified 1877 negative cases.

False Positives (FP): 47

These are the number of actual negative cases (Actual 0) incorrectly predicted as positive (Predicted 1). Light Gradient Boosting made 47 false positive predictions.

False Negatives (FN): 35

These are the number of actual positive cases (Actual 1) incorrectly predicted as negative (Predicted 0). Light Gradient Boosting made 35 false negative predictions.

ISSN 1533-9211

DOI: 10.5281/zenodo.10053411

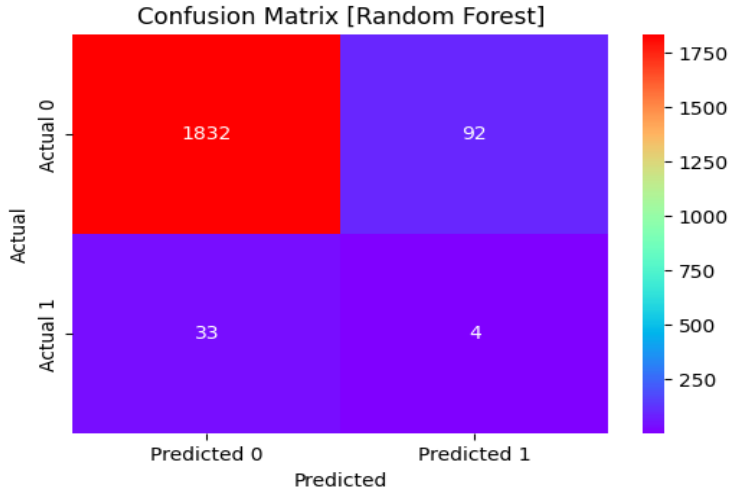


Figure 6: Confusion matrix – Random Forest

True Positives (TP): 4

These are the number of actual positive cases (Actual 1) correctly predicted as positive (Predicted 1). In this case, Random Forest correctly identified 4 positive cases.

True Negatives (TN): 1832

These are the number of actual negative cases (Actual 0) correctly predicted as negative (Predicted 0). Random Forest accurately identified 1832 negative cases.

False Positives (FP): 92

These are the number of actual negative cases (Actual 0) incorrectly predicted as positive (Predicted 1). Random Forest made 92 false positive predictions.

False Negatives (FN): 33

These are the number of actual positive cases (Actual 1) incorrectly predicted as negative (Predicted 0). Random Forest made 33 false negative predictions.

The Seybold Report
ISSN 1533-9211
DOI: 10.5281/zenodo.10053411

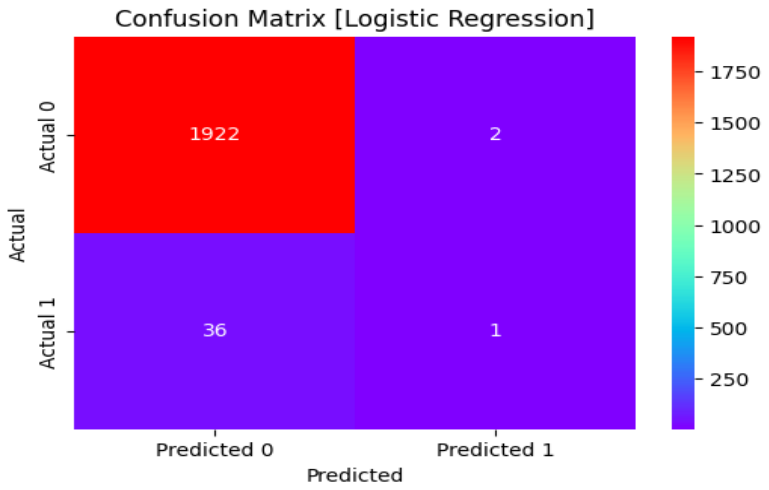


Figure 7: Confusion matrix – Logistics Regression

True Positives (TP): 1

These are the number of actual positive cases (Actual 1) correctly predicted as positive (Predicted 1). In this case, Logistic Regression correctly identified 1 positive case.

True Negatives (TN): 1922

These are the number of actual negative cases (Actual 0) correctly predicted as negative (Predicted 0). Logistic Regression accurately identified 1922 negative cases.

False Positives (FP): 2

These are the number of actual negative cases (Actual 0) incorrectly predicted as positive (Predicted 1). Logistic Regression made 2 false positive predictions.

False Negatives (FN): 36

These are the number of actual positive cases (Actual 1) incorrectly predicted as negative (Predicted 0). Logistic Regression made 36 false negative predictions.

Matrics Performance Comparison:

Table 4: Performance comparison

	XGBoost	Catboost	Gradient Boost	Light GBM	Random Forest	Logistic Regression
Logloss	0.1367	0.1367	0.1091	0.2041	0.2089	0.1962
ROC AUC	0.6953	0.6807	0.7143	0.6996	0.6971	0.6357
Precision	0.0892	0.0477	0.0876	0.0499	0.0491	0.0148
Recall	0.1083	0.0908	0.0682	0.0681	0.1249	0.0343
Accuracy	0.9664	0.9509	0.9713	0.9588	0.9411	0.9415
F1	0.0965	0.0622	0.0762	0.0570	0.0696	0.0206

The Seybold Report
ISSN 1533-9211
DOI: 10.5281/zenodo.10053411

From Table 4. It can explain and analyze as below.

Logloss: Lower logloss values indicate better model performance in classification problems. In this case, "Gradient Boost" (0.1091) has the lowest logloss, followed closely by "XGBoost" and "Catboost" (0.1367). "Gradient Boost" performs the best in this metric.

ROC AUC: ROC AUC measures a model's ability to distinguish between positive and negative classes. "Gradient Boost" (0.7143) has the highest ROC AUC, which indicates it has the best discriminatory power. "XGBoost" (0.6953) and "Light GBM" (0.6996) also perform well in this metric.

Precision: Precision measures the proportion of true positive predictions out of all positive predictions. In this case, "Gradient Boost" (0.0876) has the highest precision, followed by "XGBoost" (0.0892).

Recall: Recall measures the proportion of true positives out of all actual positives. "Random Forest" (0.1249) has the highest recall, followed by "XGBoost" (0.1083).

Accuracy: Accuracy measures the overall correct predictions. "Gradient Boost" (0.9713) has the highest accuracy, followed by "XGBoost" (0.9664).

F1 Score: F1 score is the harmonic mean of precision and recall. "Gradient Boost" (0.0762) has the highest F1 score, followed by "XGBoost" (0.0965).

Based on the metrics "Gradient Boost" performs well in terms of logloss, ROC AUC, precision, and accuracy. "Random Forest" has the highest recall, indicating its strength in capturing true positives. "XGBoost" also performs well across multiple metrics, including logloss, precision, and recall.

Table 5: Variable Importance

Level of importance	First	Second	Third
XGBoost	Cargo type group_Group 6	Destination Country_THA	Start Country_THA
Catboost	Start Country_THA	PackagingType_Others	Suminsured Amount
GradientBoosting	Suminsured Amount	Cargo type group_Group 6	Start Country_THA
LightGBM	Suminsured Amount	Effective year	Packaging Type_Others
Random Forest	Cargo type group_Group 6	Start Country_THA	Packaging Type_Others
Logistics Regression	Effective year	Cargo type group_Group 2	Start Country_MMR

Table 5. Shows the feature importance rankings for each model is important for understanding which variables or features are considered most influential in making predictions. The ranking of features can provide insights into the factors that significantly impact the model's decision-making process. These rankings are essential for feature selection, model interpretation, and understanding the relationship between features and the target variable.

The Seybold Report
ISSN 1533-9211
DOI: 10.5281/zenodo.10053411

DISCUSSION

In summary, Gradient Boost performs well in terms of ROC AUC, precision, logloss, and accuracy, making it a strong candidate for the best model overall. In summary, the "Gradient Boost" model excels in terms of ROC AUC, log loss, and accuracy. It demonstrates strong discrimination power and accurate probability estimation. However, it has lower recall, which means it might miss some positive cases. The choice of this model should consider the specific problem, objectives, and trade-offs between precision and recall. Overall, the selection of the ideal model should be based on a careful consideration of these factors and how it aligns with the specific goals of the insurance application, whether it's risk assessment, claims prediction, or premium calculation. Additionally, model fine-tuning and feature engineering could further enhance model performance.

CONCLUSION

Insurance claim prediction is an important procedure for both insurance firms and policyholders. Especially for the cargo insurance, as every minute, there are a lot of shipments transport by air sea road and rail. However, it can be a complex task that necessitates the use of proper methodologies and a methodical approach to assure excellent data quality, which eventually leads to the construction of an effective classifier.

The model's accuracy is critical in enabling insurers in accurately determining the fair cost of insurance for potential policyholders. An improved model improves the insurer's capacity to make precise decisions.

The primary objective of this study was to conduct a comprehensive evaluation and comparison of six distinct predictive models, with the aim of identifying the model that delivers the highest level of prediction accuracy across various performance metrics. The study successfully achieved this objective and revealed that the ensemble machine learning technique, Gradient Boost, outperforms the other models on multiple fronts, including ROC AUC, precision, logloss, and accuracy.

As a result of this analysis, Gradient Boost emerges as a robust candidate for the optimal choice as the overall model for predicting insurance claims in the context of cross-border transportation in Thailand. Its superior performance across key metrics underscores its efficacy in providing accurate predictions and highlights its potential to enhance decision-making and risk assessment in the domain of cargo insurance within this specific context.

Acknowledgement

The study would not have been possible without the collective contributions of School of Engineering, University of the Thai chamber of commerce and The Insurance Premium Rating Bureau (IPRB), Thailand. We would like to express my heartfelt thanks to all those who supported and contributed to this study.

The Seybold Report
ISSN 1533-9211
DOI: 10.5281/zenodo.10053411

References

1) Salika. (2021, August 18). Road Freight Rising in ASEAN. www.Salika.co. https://www.salika.co/2021/08/18/road-freight-rising-in-asean/

2) Mitchell, T. M. (1997). Machine learning. McGraw-Hill.

3) Hastie, T., Tibshirani, R., & Friedman, J. (2009). The elements of statistical learning: Data mining, inference, and prediction. Springer.

4) Sutton, R. S., & Barto, A. G. (2018). Reinforcement learning: An introduction. MIT Press.

5) Goodfellow, I., Bengio, Y., Courville, A., & Bengio, Y. (2016). Deep learning (Vol. 1). MIT press Cambridge.

6) Friedman, J. H. (2001). Greedy function approximation: a gradient boosting machine. Annals of Statistics, 1189-1232.

7) Khan, M. S. I., Islam, N., Uddin, J., Islam, S., & Nasir, M. K. (2022). Water quality prediction and classification based on principal component regression and gradient boosting classifier approach. Journal of King Saud University-Computer and Information Sciences, 34(8), 4773-4781.

8) Chen, T., & Guestrin, C. (2016). XGBoost: A Scalable Tree Boosting System. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (pp. 785-794).

9) Zhang, T., Zhu, W., Wu, Y., Wu, Z., Zhang, C., & Hu, X. (2023). An explainable financial risk early warning model based on the DS-XGBoost model. Finance Research Letters, 104045.

10) Prokhorenkova, L., Gusev, G., Vorobev, A., Dorogush, A. V., & Gulin, A. (2018). CatBoost: unbiased boosting with categorical features. In NeurIPS (pp. 6638-6648).

11) Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., ... & Yu, T. (2017). LightGBM: A highly efficient gradient boosting decision tree. In Advances in neural information processing systems (pp. 3146-3154).

12) Breiman, L. (2001). Random forests. Machine learning, 45(1), 5-32.

13) Ong, A. K. S., Prasetyo, Y. T., Velasco, K. E. C., Abad, E. D. R., Buencille, A. L. B., Estorninos, E. M., ... & Sittiwatethanasiri, T. (2022). Utilization of random forest classifier and artificial neural network for predicting the acceptance of reopening decommissioned nuclear power plant. Annals of Nuclear Energy, 175, 109188.

14) Hosmer Jr, D. W., Lemeshow, S., & Sturdivant, R. X. (2013). Applied logistic regression. John Wiley & Sons.

15) Baran, S., & Rola, P. (2022). Prediction of motor insurance claims occurrence as an imbalanced machine learning problem. arXiv preprint arXiv:2204.06109.

16) Analytics Vidhya. (2020). A Comprehensive Guide to Confusion Matrix in Machine Learning. Analytics Vidhya. https://www.analyticsvidhya.com/blog/2020/04/confusion-matrix-machine-learning/

17) Hossin, M., & Sulaiman, M. N. (2015). A review on evaluation metrics for data classification evaluations. International journal of data mining & knowledge management process, 5(2), 1.

18) Trifonova, O. P., Lokhov, P. G., & Archakov, A. I. (2014). Metabolic profiling of human blood. Biomeditsinskaia Khimiia, 60(3), 281-294.

19) Scikit-learn. (n.d.). sklearn.metrics.recall_score. Retrieved from https://scikit-learn.org/stable/modules/generated/sklearn.metrics.recall_score.html.

20) Kravchenko, Y., Dakhno, N., Leshchenko, O., & Tolstokorova, A. (2020). Machine Learning Algorithms for Predicting the Results of COVID-19 Coronavirus Infection. In IT&I Workshops (pp. 371-381).