Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Benchmarking Predictive Machine Learning for Maritime Decision Support Across 91 Economies

Ariyono Setiawan

rmaryo4u@poltekpel-sby.ac.id

Politeknik Pelayaran Surabaya

Choo Wou Onn

INTI International University

Fahmi Umasangadji

Sekolah Tinggi Ilmu Pelayaran Jakarta

Sutoyo Sutoyo

Politeknik Pelayaran Surabaya

Iie Suwondo

Politeknik Pelayaran Surabaya

Abdullah Alghoson

University of Jeddah

Research Article

Keywords: predictive machine learning, intelligent decision support, external feature evaluation, gradient boosting, temporal generalization, maritime forecasting

Posted Date: November 23rd, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11709363/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

Abstract

Reliable intelligent decision support requires predictive models to demonstrate value beyond simple benchmarks and to justify the inclusion of external features. This study benchmarks regularized linear and nonlinear machine learning using maritime forecasting as the application domain. An audited cargo archive is combined with official container-throughput data for 91 economies during 2010–2023, yielding 1,274 annual observations. Seven local and pooled specifications are evaluated on 546 common one-step-ahead targets in 2018–2023. Ridge models use nested temporal tuning, while shallow gradient boosting uses fixed settings. Paired feature ablations test whether lagged cargo signals improve predictions beyond throughput history. Persistence has the lowest equal-economy mean absolute log error, at 7.018 log points. History-only boosting achieves the lowest volume-weighted absolute percentage error, at 4.040%, compared with 4.811% for persistence. Adding cargo increases equal-economy error by 0.029 log points for ridge and 0.146 for boosting; crossed economy–year resampling ranges include zero for both differences. Longer cargo lags, economy holdouts and an inception-cohort sensitivity do not establish an aggregate ridge benefit from the additional features. The contribution is a reproducible evaluation of model choice, feature value and loss-dependent performance for predictive decision-support components. The findings favor explicit benchmark and feature-admission checks over assuming that a more complex model or richer input set is preferable. Six test years, retrospective vintages and incomplete release calendars limit generalization and real-time deployment claims.

1 Introduction

Predictive machine learning can inform intelligent decision support by supplying forecasts for planning and monitoring. Its practical value depends on whether the learned predictions improve an explicitly defined objective and remain credible when new observations arrive. Maritime applications provide a relevant setting: learning methods are used for demand prediction and other port decisions [1], while the connection between predictive outputs and operational optimization remains an active research question [2]. Multisource container forecasting also illustrates the use of external economic information [3]. The central issue examined here is the evaluation of predictive learning components, with maritime throughput serving as the application domain.

Short annual histories challenge the evaluation of data-driven prediction. Pooling across economies increases the observations available to a shared learning function [4], but a larger row count does not remove temporal dependence or establish generalization. The experiment therefore evaluates a concrete information-design question: whether an external cargo feature block improves future-year predictions relative to the same learner using throughput history alone. This comparison separates the value of the learning method from the value of additional inputs.

The distinction between explanation and prediction is fundamental [5]. A high contemporaneous correlation does not establish forecast value, just as an increase in output does not measure technical efficiency. Leakage can further exaggerate apparent performance when targets or target-derived

quantities enter predictors [6]. Forecast evaluation requires an explicit information set, competitive benchmarks and error measures that correspond to the intended comparison [7]. PCA can summarize variation but does not independently establish an AI innovation or the validity of a performance construct [8].

The research questions concern predictive model evaluation, external feature value and transferability. Do lagged cargo growth and composition improve forecasts beyond throughput history within regularized linear and shallow nonlinear learners? Does the evaluation loss change the observed model ranking? Are feature-block comparisons stable under longer information lags, economy holdouts and alternative coverage rules? The United States provides a worked data-audit example, while the main empirical benchmark spans 91 economies.

The contribution lies in a reproducible evaluation of predictive machine learning for decision support. Source-level aggregation checks establish what the inputs represent. Nested temporal fitting and matched targets support comparisons between learning methods, while paired feature ablations isolate the observed value of the external cargo block. Loss-dependent rankings and transfer checks then qualify the model-selection evidence. The contribution is an empirical evaluation design and its documented findings; the algorithms themselves are established methods. Data, configurations, tuning selections and individual forecasts are supplied for replication.

2 Related research and analytical expectations

2.1 Predictive learning and external information

Maritime forecasting research offers a setting for evaluating statistical and machine-learning prediction [1, 3]. The distinction between a predictive component and a decision system is important: forecasts estimate future quantities, whereas choosing actions requires objectives, constraints and costs [2]. The present study evaluates the predictive component through measured forecast losses. It provides evidence relevant to the design of intelligent decision support, while operational effects such as congestion reduction require a subsequent decision-level evaluation.

The cargo block considered here has a plausible but imperfect connection to the target. Loaded and discharged mass describe trade direction and cargo composition, while container handling includes empty units and transshipment. Cargo may therefore contain useful leading information, redundant information already reflected in container history, or variation outside the target's reporting boundary. These alternatives are evaluated through restricted and augmented versions of the same learning family rather than by interpreting correlation as proof of usefulness.

Open-data transport prediction has also been investigated by Setiawan et al. [9], who applied LSTM to Indonesian transport-service export values. That study concerns export value rather than container lifts; its relevance here is the use of openly sourced temporal data for transport prediction. Setiawan et al. [10] examine container-ship capacity differences across five Indo-Pacific economies using PCA, ANOVA and

clustering. That cross-economy perspective motivates examining heterogeneity, while the present analysis tests incremental forecasting value on matched future-year targets rather than treating descriptive separation as predictive validation.

2.2 Pooling and evaluation objectives

Global forecasting methods learn across multiple series, whereas local methods fit each series separately [4]. Pooling is attractive when each economy has a short history, but its benefit is empirical, not guaranteed. Large benchmark exercises emphasize learning from comparative forecast evidence rather than relying on model labels [11]. Here, pooled ridge and boosting are compared with local persistence, drift and trend forecasts using identical targets.

Scale also changes the comparison. Absolute TEU errors emphasize high-volume economies, while proportional errors give smaller systems a meaningful role. Forecast accuracy measures answer different questions [12]. Accordingly, this study fixes an equal-economy log-error measure as the main summary and reports MASE and volume-weighted absolute percentage error as complementary views. No method is called universally best when the metrics disagree.

2.3 Temporal design and dependence

Expanding-origin evaluation and simple benchmarks are established forecasting tools [13, 14]. Cross-validation is not automatically invalid for every time-series model, but its validity depends on the dependence structure and the prediction setting [15]. Empirical work documents differences among performance-estimation procedures [16], and structured data can require blocking along more than one dimension [17]. The main experiment therefore tunes only within past years, and additional economy holdouts examine transfer to series excluded from pooled training.

The analytical expectation is deliberately conditional: adding cargo is useful only if it lowers the selected out-of-sample loss beyond a throughput-history model. A contrary result does not demonstrate that trade is economically irrelevant or that all external variables are unhelpful. It establishes the observed value of a specific feature block, horizon and model set under a documented dataset and information assumption.

3 Data and methods

3.1 Sources and population construction

Cargo records come from the recovered UNCTAD-format archive US_SeabornTrade.csv, which covers multiple economies despite its filename prefix. The measure is labeled metric tons in thousands, with economy codes, cargo-type codes and source footnotes. Official seaborne-trade concepts concern international shipments and exclude cabotage and transshipment [18, 19]. The archive's original download date and release identifier are unknown; its raw bytes and checksum are retained. It is not presented as a newly downloaded current UNCTAD release.

Container throughput comes from the World Bank World Development Indicators API, indicator IS.SHP.GOOD.TU, attributed to UNCTAD [20]. The extract was retrieved on 12 November 2026 and reports lastupdated = 2026-11-11. Metadata include coastal and international movements, empty units, and two lifts for transshipment at intermediate ports [21]. The official API query structure and country metadata support the documented extraction [22]. The two source vintages are retained separately rather than silently merged into a uniform historical release.

World Bank aggregate-region entries are removed using country metadata. A bundled ISO 3166 numeric–alpha-3 crosswalk connects economy codes without manual name matching. Among 158 economies with at least one positive matched throughput and cargo year during 2010–2023, 91 have complete positive values throughout all 14 years; 67 are excluded from the balanced panel for incomplete coverage. This yields 1,274 economy-year rows. Economies and territories are treated as reporting economies, not all as sovereign states. No missing observations are interpolated.

The balanced panel fixes the analytical population across six test years. Because full-period availability is a retrospective selection rule, it may favor better-reported systems. A separate inception sensitivity fixes membership using complete positive 2010–2017 coverage and evaluates available subsequent matched targets. It includes 132 initial economies, of which 131 contribute 650 test pairs; target coverage falls from 131 economies in 2018 to 91 in 2023. This sensitivity assesses, but cannot eliminate, reporting-related selection.

Table 1
Sample selection and temporal evaluation coverage

Stage	Count	Record definition
Positive matched candidate economies	158	At least one matched year in 2010–2023
Excluded candidate economies	67	Incomplete positive full-period coverage
Balanced analytical economies	91	Complete 2010–2023 levels
Annual analytical observations	1,274	91 economies × 14 years
First-origin pooled training rows	546	Targets 2012–2017 for the 2018 forecast
Last-origin pooled training rows	1,001	Targets 2012–2022 for the 2023 forecast
Common test targets per model	546	91 economies × 6 years, 2018–2023
Inception sensitivity test pairs	650	Available targets from the initial-coverage cohort

Counts describe distinct stages, not additive independent samples. The 546 common targets span only six calendar years; the inception sensitivity uses a separate reporting-coverage rule.

Table 2
Core quantities and construction rules

Quantity	Definition	Use
C_{i}	Container port traffic in TEU	Forecast target and lagged history
L_{i}	CargoType 10, total goods loaded, thousand metric tons	Cargo growth and composition
D_{i}	CargoType 20, total goods discharged, thousand metric tons	Cargo growth and composition
T_{i}	L_{i} + D_{i}	Audit and descriptive quantity
Source flags	Original footnotes retained in raw archive	Qualification of reporting quality

3.2 Aggregation audit and information boundaries

The cargo totals are constructed as:

$$T_{it} = L_{it} + D_{it}.$$

1

Loaded categories 11, 12 and 13 comprise total 10; discharged categories 21, 22 and 23 comprise total 20, up to rounding. Adding all eight categories therefore counts approximately twice T_{i}. A United States audit illustrates the overlapping aggregation alternative. Across the balanced panel, 2,149 direction-year records have complete component breakdowns and agree with their reported totals within 0.0011 thousand metric tons. Another 399 direction-year records lack a complete breakdown; no component values are imputed. Their reported totals remain available for analysis. Component rows are used for validation, never added to the totals in predictors.

The target is an activity measure. No labor, capital, berth capacity, digitalization index or recovery counterfactual is observed, so technical efficiency and resilience are not estimated. The international cargo and port-traffic definitions overlap incompletely [18, 21], which limits interpretation of their association. The wider disruption context in maritime transport [23] motivates temporal testing but is not used to attribute forecast errors to a particular policy or technology.

3.3 Target and feature blocks

The supervised target is annual container log growth:

$$g_{it} = \log C_{it} - \log C_{i,t-1}.$$

2

The restricted feature block contains previous-year container log growth and previous-year log throughput:

$$\mathbf{x}_{it}^{(0)} = \left(g_{i,t-1}, \log C_{i,t-1}\right)^{\mathsf{T}}.$$

3

The augmented block adds previous-year loaded growth, discharged growth and loaded share:

$$\mathbf{x}_{it}^{(1)} = \left(\mathbf{x}_{it}^{(0)\mathsf{T}}, \Delta \log L_{i,t-1}, \Delta \log D_{i,t-1}, \frac{L_{i,t-1}}{L_{i,t-1} + D_{i,t-1}}\right)^{\mathsf{T}}.$$

4

All features are indexed no later than t − 1. Training targets begin in 2012 because the lagged growth variables require two earlier levels. The cargo block contains no contemporaneous target or throughput-to-tonnage ratio. Pooling uses no economy identifier, country dummy or region label. Predictions are therefore learned from shared relationships among the observed features rather than economy-specific intercepts.

3.4 Models and tuning

Three local benchmarks are included. Persistence repeats the last observed level. Log drift extrapolates the average annual log change from 2010 to the current origin. Log trend regresses historical log throughput on a linear time index and extrapolates one step. Both trend methods are exponentiated without additional adjustment.

Ridge regression regularizes correlated predictors [24]. Within each training window, feature means and population standard deviations are estimated from training rows only. The unpenalized intercept and penalized slopes minimize:

$$\min_{a,\mathbf{b}} \sum_{(i,s)\in \mathcal{T}_t} \left(g_{is} - a - \mathbf{x}_{is}^{\mathsf{T}}\mathbf{b}\right)^2 + \lambda \left\| \mathbf{b} \right\|_2^2.$$

5

The penalty is selected from λ ∈ {0.1, 1, 10, 100}. At outer target year t, three inner validation targets, t − 3 through t − 1, are predicted using labels strictly before each inner year. The selection criterion is mean absolute log-growth error pooled across those validation targets. Scaling is refitted inside each inner training window. An exact tie selects the larger penalty. The chosen model is then refitted on all labels before t. The penalty grid and feature definitions are fixed for this experiment; the analysis is exploratory and was not preregistered.

Gradient boosting provides a nonlinear comparison using shallow trees [25]. History-only and augmented versions use identical settings: squared-error loss, 100 boosting iterations, learning rate 0.05, maximum depth 2, at least 20 samples per leaf, L2 regularization 1, no early stopping, and random seed 32. These settings are fixed rather than selected on the outer test errors. The scikit-learn implementation [26] and exact software versions are recorded in the replication environment.

Table 3
Model comparison and feature-block control

Model	Estimation unit	Information	Selection
Persistence	Each economy	Last throughput	None
Log drift	Each economy	Throughput history	None
Log trend	Each economy	Throughput history and time	OLS refit
Ridge history	Pooled economies	Restricted block	Nested temporal λ
Ridge + cargo	Pooled economies	Augmented block	Nested temporal λ
Boost history	Pooled economies	Restricted block	Fixed configuration
Boost + cargo	Pooled economies	Augmented block	Fixed configuration

Learned log-growth predictions are converted to positive throughput forecasts:

$$\widehat{C}_{it} = C_{i,t-1}\exp\left[\widehat{f}_{t}\left(\mathbf{x}_{it}\right)\right].$$

6

No retransformation bias correction is applied. The forecasts represent exponentiated log-location predictions rather than claimed conditional arithmetic means. Both learned families fit squared log-growth error, while evaluation includes absolute-error criteria; consequently, rankings are specific to this loss and estimation design, not an exhaustive comparison of loss-optimized algorithms.

3.5 Temporal testing and accuracy measures

Outer targets are 2018–2023. At the first origin, pooled models use targets 2012–2017, giving 546 training rows; at the last origin they use 2012–2022, giving 1,001 rows. The 91 economies contribute exactly 546 common test targets to each of the seven methods. Earlier test outcomes may enter later training windows only after their calendar year has passed, as in an expanding-origin design.

The principal summary is the mean absolute log error multiplied by 100:

$$\mathrm{MALE}_{100} = \frac{100}{NH}\sum_{i=1}^{N}\sum_{t\in\mathcal{H}}\left|\log C_{it} - \log\widehat{C}_{it}\right|.$$

7

Here N = 91, H = 6 and $\mathcal{H}$ = {2018,…,2023}. Equal series lengths give each economy equal weight. Log points approximate percentage errors for small deviations but are not labeled exact percentages. MASE divides each absolute TEU error by that economy's mean absolute first difference over the current training history and then averages the ratios [12]. This origin-specific denominator differs from a fixed six-year persistence error, so an observed MASE above one does not contradict a method outperforming persistence over those targets.

Volume-weighted absolute percentage error is:

$$\mathrm{WAPE} = 100 \frac{\sum_{i,t} \left| C_{it} - \widehat{C}_{it} \right|}{\sum_{i,t} C_{it}}.$$

8

It gives more influence to large throughput systems. Each metric is reported for all models, and the selected main summary is not changed after inspecting the ranking. No multiplicity-adjusted claim of a universally superior algorithm is made.

3.6 Robustness and uncertainty

Within each learning family, the cargo ablation is the paired difference between augmented and restricted absolute log errors. Positive differences favor history alone. Economy-level resampling preserves each economy's six-year error sequence. A second, crossed economy–year resampling independently samples 91 economy indices and six year indices with replacement, following the crossed-array logic of the pigeonhole bootstrap [27]. Ten thousand draws and seed 32 generate central 95% resampling ranges. Models are not refitted inside these draws.

With only six year clusters, these ranges are sensitivity summaries rather than well-calibrated confirmatory intervals. They do not capture every form of serial dependence, common shocks or model-selection uncertainty. Leave-one-test-year summaries, 2010 size quartiles and exclusion of the United States provide additional descriptive checks.

A two-year cargo-lag experiment keeps container history at t − 1 but shifts cargo features to t − 2; its training targets begin in 2013, and restricted ridge is refitted on the same smaller training sample. A five-group economy holdout assigns sorted ISO3 codes cyclically to deterministic groups. For each group and origin, both ridge variants are tuned temporally and fitted using the other four groups only, then evaluated on the excluded group. No held-out economy contributes fitting labels or scaling information. These holdouts assess transfer within the observed reporting population rather than prove representativeness of missing economies.

Evaluation workflow for predictive learning components. Audited inputs precede temporal model fitting; restricted and augmented feature blocks are compared on matched targets. Loss and transfer checks qualify the evidence for decision-support use.

4 Results

4.1 Audit of the motivating United States series

The recovered United States archive gives mean loaded-plus-discharged cargo of 1,367.307 million metric tons for 2010–2023. Summing component and total rows gives 2,734.614 million, demonstrating the consequence of a reproducible overlap in the aggregation rule. This is a data-construction check rather than an alternative estimate of national trade. Halving this incorrectly doubled predictor leaves Pearson correlation and simple-regression R^2 unchanged; it corrects its economic meaning and units, not its apparent association.

Using the current official throughput vintage, United States mean throughput is 49.360 million TEU, the level correlation with corrected cargo is 0.8212, and the correlation of 13 annual log changes is 0.4416. These descriptive relationships motivate, but do not validate, a forecasting model. Historical throughput values are subject to revision; the current vintage is retained consistently throughout this experiment.

4.2 Model ranking depends on the loss

Table 4
Forecast accuracy on 546 common economy-year targets during 2018–2023

Model	MALE × 100	MASE	WAPE %
Persistence	7.018	1.217	4.811
Log drift	7.524	1.293	4.376
Log trend	9.562	1.641	5.968
Ridge history	7.077	1.255	4.215
Ridge + cargo	7.106	1.262	4.159
Boost history	7.050	1.232	4.040
Boost + cargo	7.196	1.271	4.170

Persistence has the smallest equal-economy log error, 7.018, and the smallest MASE, 1.217. History-only boosting is close on log error, 7.050, but has the lowest WAPE, 4.040%, compared with 4.811% for persistence. Its WAPE is approximately 16.0% lower in relative terms. This volume-weighted improvement is an observed test-period result, not an established operational cost saving.

The contrast shows that pooled learning can be useful under one weighting objective without dominating under another. Log trend performs poorly across the panel, with log error 9.562 and WAPE 5.968%. A trend that is competitive in a single national series therefore cannot be assumed to generalize across all economies.

Panel (a) reports equal-economy mean absolute log error multiplied by 100; panel (b) reports volume-weighted absolute percentage error. Labels show three-decimal values. All methods use the same 546 targets, and smaller values indicate lower error.

4.3 Incremental value of cargo features

Adding the cargo block changes ridge log error from 7.077 to 7.106, a deterioration of 0.029 log points. The corresponding boosting change is from 7.050 to 7.196, a deterioration of 0.146 log points. Cargo improves the economy-level mean error in 39 of 91 economies for ridge and 33 of 91 for boosting. Ridge's WAPE declines slightly, from 4.215% to 4.159%, whereas boosting's WAPE increases from 4.040% to 4.170%. Thus, the selected cargo block does not show consistent incremental value across either model families or weighting criteria.

Economy-only resampling gives narrower ranges than crossed economy–year resampling. The crossed central 95% ranges for cargo-minus-history log error are − 0.042 to 0.103 for ridge and − 0.075 to 0.377 for boosting. Both include zero. The evidence supports a statement of no demonstrated aggregate improvement, rather than a confirmatory claim that cargo is always harmful.

Table 5
Paired cargo feature ablation and resampling uncertainty

Learning family	Change in MALE × 100	Economies improving	95% crossed resampling range
Ridge	0.029	39 / 91	[-0.042, 0.103]
Boosting	0.146	33 / 91	[-0.075, 0.377]

Change is augmented minus restricted loss on the same 546 targets; positive values favor history alone. An economy improves when its six-year mean loss falls. Ranges resample economies and years 10,000 times without refitting and are descriptive with six year clusters.

Cargo helps both model families in the 2020 target year, with paired differences of − 0.092 for ridge and − 0.111 for boosting. The direction changes in other years. Removing any one test year leaves the average cargo-minus-history difference positive for both families, but that sign stability does not remove the uncertainty associated with the short period.

Panel (a) gives annual paired differences; panel (b) shows pooled differences and central 95% crossed economy–year resampling ranges. Positive values favor history alone. Ranges describe fixed forecasts and are not confirmatory confidence intervals.

4.4 Lag availability and transfer checks

Table 6
Restricted versus augmented ridge under alternative designs

Design	Comparison targets	History MALE × 100	Cargo MALE × 100
Main balanced panel	546	7.077	7.106
Cargo lagged two years	546	7.020	7.096
Five-group economy holdout	546	7.095	7.134
Inception cohort with available targets	650	8.333	8.357

The two-year cargo lag does not reverse the aggregate comparison: history ridge has log error 7.020 and cargo ridge 7.096. Group holdouts yield 7.095 and 7.134, respectively. The inception sensitivity gives 8.333 and 8.357; persistence has error 8.065 on these same available targets. These designs address different limitations, but none establishes an aggregate ridge advantage from the selected cargo signals.

Size heterogeneity is consequential. Among the 23 economies in the largest 2010-throughput quartile, history-only boosting has log error 4.557 compared with 4.762 for persistence. In the smallest quartile, both errors are above 9.1. These are descriptive strata formed from initial-year data, not evidence of a causal scale effect or a validated model-allocation rule. Removing the United States leaves the principal equal-economy ranking unchanged.

For the six United States targets, local log trend has log error 5.586, compared with 5.927 for persistence and 6.249 for history ridge. Cargo ridge is 6.179, while cargo boosting is 7.088. These log-growth panel specifications differ from a contemporaneous national level regression. The focal economy illustrates heterogeneity rather than serving as independent external validation.

5 Discussion

5.1 External feature value in predictive learning

Loaded and discharged cargo are substantively connected to maritime activity, yet the experiment provides no consistent aggregate gain from adding their lagged growth and composition to throughput history. This does not contradict studies using richer hinterland or foreland information [3]: the input definitions, frequency and forecasting population differ. The finding instead narrows the claim that can be made for a specific low-frequency national cargo block.

The result also distinguishes model learning from the value of an external data source. History-only boosting reduces volume-weighted error relative to persistence, so the analysis does not support a blanket rejection of machine learning. Its advantage cannot be attributed to cargo, because the relevant model does not use that block. Separating those comparisons prevents a feature-rich model from receiving credit for gains that may be available from throughput history alone.

These results complement the open-data transport forecasting application of Setiawan et al. [9] and the cross-economy capacity analysis of Setiawan et al. [10]. Together they motivate attention to temporal prediction and heterogeneity. Their response variables, units and validation designs differ from the present throughput experiment, so published accuracy figures and component-variance shares are not treated as comparable performance benchmarks.

5.2 Implications for intelligent decision support

For an intelligent forecasting component, model selection should follow the intended evaluation objective. Equal-economy log error treats a proportional miss in a smaller economy as relevant alongside one in a larger economy, while WAPE emphasizes absolute volume. A monitoring application should therefore declare whose errors matter, which loss is reported and what information was available at the forecast origin. The observed difference between persistence and history-only boosting makes those design choices consequential.

The paired comparisons also inform feature admission: an external block should be assessed against the same learner and targets before being added to a predictive component. Here, the selected cargo block does not provide a consistent aggregate improvement. An operational extension would need release-time checks, monitoring of forecast errors and a decision model connecting throughput predictions to actions and costs [2]. These are requirements for subsequent system development; the present benchmark evaluates forecasts and feature blocks.

5.3 Auditability and interpretation of predictive evidence

Auditable data preparation supports reliable intelligent systems. The worked example shows that a duplicated tonnage measure can retain a convincing correlation while misrepresenting the quantity by a factor of two. Source-lineage checks and temporal evaluation therefore serve complementary purposes: they establish the meaning of a predictor and its observed contribution to forecasting. The supplied cargo-code rules, training splits and saved predictions make both stages inspectable.

Ridge coefficients and model feature blocks are transparent, but they are not causal parameters. The ablation measures the change in forecasting loss when a block is available to a specified learner; it does not estimate the economic effect of trade on ports. This interpretation follows the separation between explanatory and predictive modeling [5] and avoids treating a learning algorithm as a mechanism identification strategy.

5.4 Limitations

The panel adds economies, not additional historical years. All systems share six test periods, so 546 targets must not be treated as 546 independent temporal shocks. Crossed resampling and year deletion make this limitation visible, but six year clusters do not support precise long-run inferential claims. Resampling fixed forecasts also omits refitting uncertainty.

Both data vintages are retrospective, and the cargo archive's original release is unresolved. The main one-year-lag experiment assumes prior-year information is available; the two-year sensitivity is more conservative but still does not reconstruct actual release calendars. Consequently, these are pseudo-out-of-sample results, not a real-time deployment backtest. Source revision errors and publication delays may materially alter attainable performance.

The balanced sample favors complete reporting. The inception analysis expands coverage but uses available matched targets, leaving missingness potentially informative. Future work should examine the excluded economies and port-level or monthly records. Aggregated cargo mass and container lifts have different boundaries, and the selected features omit route connectivity, capacity and detailed economic indicators. Alternative signals may have value that is not captured here.

The model set is deliberately limited. Shallow boosting settings are fixed; the penalty grid is small; and loss functions are not optimized separately for every evaluation measure. No claim is made that these specifications dominate modern forecasting architectures. The study is an auditable empirical benchmark, not a new algorithm, a causal efficiency model or a systematic review.

6 Conclusion

This study benchmarks predictive machine learning as a component of intelligent maritime decision support across 91 economies. Persistence has the lowest equal-economy log error, while history-only boosting has the lowest volume-weighted error on the six common test years. The external cargo block does not consistently improve either learning family, and lag, transfer and coverage checks do not establish an aggregate ridge benefit. The resulting contribution is evidence for evaluating model choice, external feature value and temporal generalization under explicit data and loss assumptions. Source audits, simple benchmarks and paired comparisons should precede claims of predictive improvement. Real-time availability checks and decision-level validation remain necessary before operational use.

Declarations

Ethics and consent

The study uses publicly available aggregate economy-level statistics and involves no human participants or identifiable personal data. Human-participant ethical approval, consent to participate and consent for publication are not applicable. Clinical trial registration is not applicable.

Generative AI disclosure

ChatGPT assisted with language editing, interpretation and the submitted manuscript remains with the authors.

Funding

and competing interests

This research received no specific grant from funding agencies in the public, commercial or not-for-profit sectors. The authors declare no competing interests.

Author Contribution

Ariyono Setiawan: Conceptualization, Methodology, Software, Formal analysis, Visualization, Writing – original draft. Choo Wou Onn: Methodology, Validation, Supervision, Writing – review & editing. Fahmi Umasangadji: Investigation, Data curation, Writing – review & editing. Sutoyo: Investigation, Validation, Writing – review & editing. Iie Suwondo: Data curation, Validation, Writing – review & editing. Abdullah Alghoson: Methodology, Validation, Writing – review & editing.

Data Availability

All datasets are publicly available via UNCTADstat, [World Bank Open Data](https:/data.worldbank.org) , and [BTS.gov](https:/www.bts.gov) . No confidential or personal data were used. The dataset used in this study is archived at the institutional repository of Politeknik Pelayaran Surabaya and available upon request for reproducibility.

References

1. Filom S, Amiri AM, Razavi S. Applications of machine learning methods in port operations—A systematic literature review. Transp Res E. 2022;161:102722. https://doi.org/10.1016/j.tre.2022.102722.
2. Jahangard M, Xie Y, Feng Y. Leveraging machine learning and optimization models for enhanced seaport efficiency. Maritime Econ Logistics. 2025;27:710–51. https://doi.org/10.1057/s41278-024-00309-w.
3. Zeng F, Xu S. A hybrid container throughput forecasting approach using bi-directional hinterland data of port. Sci Rep. 2024;14:25502. https://doi.org/10.1038/s41598-024-77376-9.
4. Montero-Manso P, Hyndman RJ. Principles and algorithms for forecasting groups of time series: Locality and globality. Int J Forecast. 2021;37(4):1632–53. https://doi.org/10.1016/j.ijforecast.2021.03.004.
5. Shmueli G. To explain or to predict? Stat Sci. 2010;25(3):289–310. https://doi.org/10.1214/10-STS330.
6. Kapoor S, Narayanan A. Leakage and the reproducibility crisis in machine-learning-based science. Patterns. 2023;4(9):100804. https://doi.org/10.1016/j.patter.2023.100804.

7. Hewamalage H, Ackermann K, Bergmeir C. Forecast evaluation for data scientists: Common pitfalls and best practices. Data Min Knowl Disc. 2023;37:788–832. https://doi.org/10.1007/s10618-022-00894-5.
8. Jolliffe IT, Cadima J. (2016). Principal component analysis: A review and recent developments. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences, 374*(2065), 20150202. https://doi.org/10.1098/rsta.2015.0202
9. Setiawan A, Handoko W, Hadi ARBA, Onn CW. Prediksi ekspor jasa transportasi Indonesia menggunakan LSTM berbasis data perdagangan global terbuka. MIND (Multimedia Artif Intell Netw Database) J. 2025;10(2):130–44. https://doi.org/10.26760/mindjournal.v10i2.130-144.
10. Setiawan A, Otok BW, Handoko W, Hadi ARA, Onn CW, Arli D. Uneven transitions in container ship capacity across Indo-Pacific economies (2010–2022): Integrating PCA, ANOVA, and clustering evidence. Int J Adv Data Inform Syst. 2026;7(1):23–41. https://doi.org/10.59395/ijadis.v7i1.1442.
11. Makridakis S, Spiliotis E, Assimakopoulos V. The M4 Competition: 100,000 time series and 61 forecasting methods. Int J Forecast. 2020;36(1):54–74. https://doi.org/10.1016/j.ijforecast.2019.04.014.
12. Hyndman RJ, Koehler AB. Another look at measures of forecast accuracy. Int J Forecast. 2006;22(4):679–88. https://doi.org/10.1016/j.ijforecast.2006.03.001.
13. Hyndman RJ, Athanasopoulos G. (2021). *Forecasting: Principles and practice*(3rd ed.). OTexts. https://otexts.com/fpp3/
14. Tashman LJ. Out-of-sample tests of forecasting accuracy: An analysis and review. Int J Forecast. 2000;16(4):437–50. https://doi.org/10.1016/S0169-2070(00)00065-0.
15. Bergmeir C, Hyndman RJ, Koo B. A note on the validity of cross-validation for evaluating autoregressive time series prediction. Comput Stat Data Anal. 2018;120:70–83. https://doi.org/10.1016/j.csda.2017.11.003.
16. Cerqueira V, Torgo L, Mozetič I. Evaluating time series forecasting models: An empirical study on performance estimation methods. Mach Learn. 2020;109:1997–2028. https://doi.org/10.1007/s10994-020-05910-7.
17. Roberts DR, Bahn V, Ciuti S, Boyce MS, Elith J, Guillera-Arroita G, Hauenstein S, Lahoz-Monfort JJ, Schröder B, Thuiller W, Warton DI, Wintle BA, Hartig F, Dormann CF. Cross-validation strategies for data with temporal, spatial, hierarchical, or phylogenetic structure. Ecography. 2017;40(8):913–29. https://doi.org/10.1111/ecog.02881.
18. United Nations Conference on Trade and Development. (2023a). *Handbook of Statistics 2023: Maritime transport concepts and definitions*. https://unctad.org/system/files/official-document/tdstat48_FS013_en.pdf
19. United Nations Conference on Trade and Development. (n.d.). *World seaborne trade by types of cargo, annual*[Data set]. UNCTADstat. https://unctadstat.unctad.org/datacentre/dataviewer/US.SeaborneTrade

20. World Bank. (n.d.-a). Container port traffic (TEU: 20 foot equivalent units)[Data set]. World Development Indicators. Retrieved November 12. 2026, from https://data.worldbank.org/indicator/IS.SHP.GOOD.TU
21. World Bank. (n.d.-b). *Metadata glossary: IS.SHP.GOOD.TU*. https://databank.worldbank.org/metadataglossary/world-development-indicators/series/IS.SHP.GOOD.TU
22. World Bank. (n.d.-c). *API basic call structures*. Indic API Doc. https://datahelpdesk.worldbank.org/knowledgebase/articles/898581
23. United Nations Conference on Trade and Development. (2023b). *Review of Maritime Transport 2023*. https://unctad.org/publication/review-maritime-transport-2023
24. Hoerl AE, Kennard RW. Ridge regression: Biased estimation for nonorthogonal problems. Technometrics. 1970;12(1):55–67. https://doi.org/10.1080/00401706.1970.10488634.
25. Friedman JH. Greedy function approximation: A gradient boosting machine. Annals Stat. 2001;29(5):1189–232. https://doi.org/10.1214/aos/1013203451.
26. Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, Blondel M, Prettenhofer P, Weiss R, Dubourg V, Vanderplas J, Passos A, Cournapeau D, Brucher M, Perrot M, Duchesnay E. Scikit-learn: Machine learning in Python. J Mach Learn Res. 2011;12:2825–30. https://www.jmlr.org/papers/v12/pedregosa11a.html.
27. Owen AB. The pigeonhole bootstrap. Annals Appl Stat. 2007;1(2):386–411. https://doi.org/10.1214/07-AOAS122.

Figures

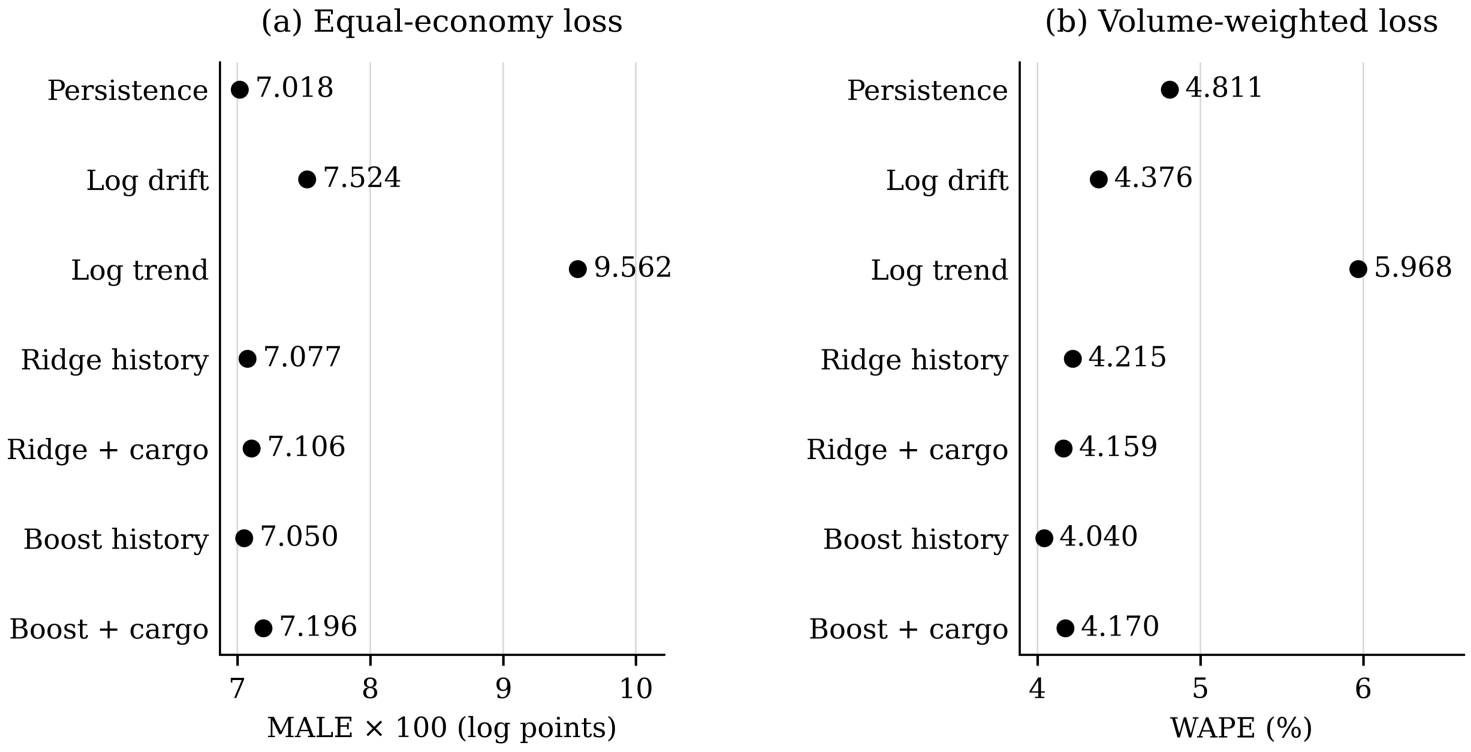

Figure 1

Forecast accuracy under two weighting objectives.

Panel (a) reports equal-economy mean absolute log error multiplied by 100; panel (b) reports volume-weighted absolute percentage error. Labels show three-decimal values. All methods use the same 546 targets, and smaller values indicate lower error.

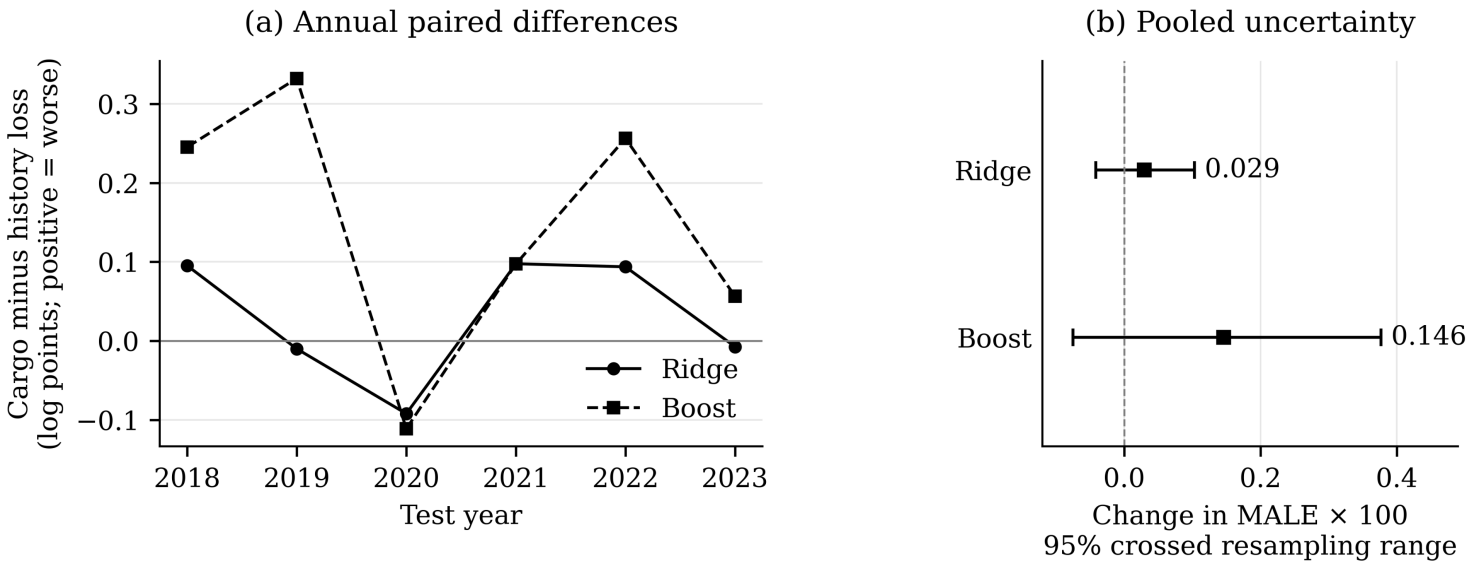

Figure 2

Cargo feature ablation by time and uncertainty.

Panel (a) gives annual paired differences; panel (b) shows pooled differences and central 95% crossed economy–year resampling ranges. Positive values favor history alone. Ranges describe fixed forecasts and are not confirmatory confidence intervals.

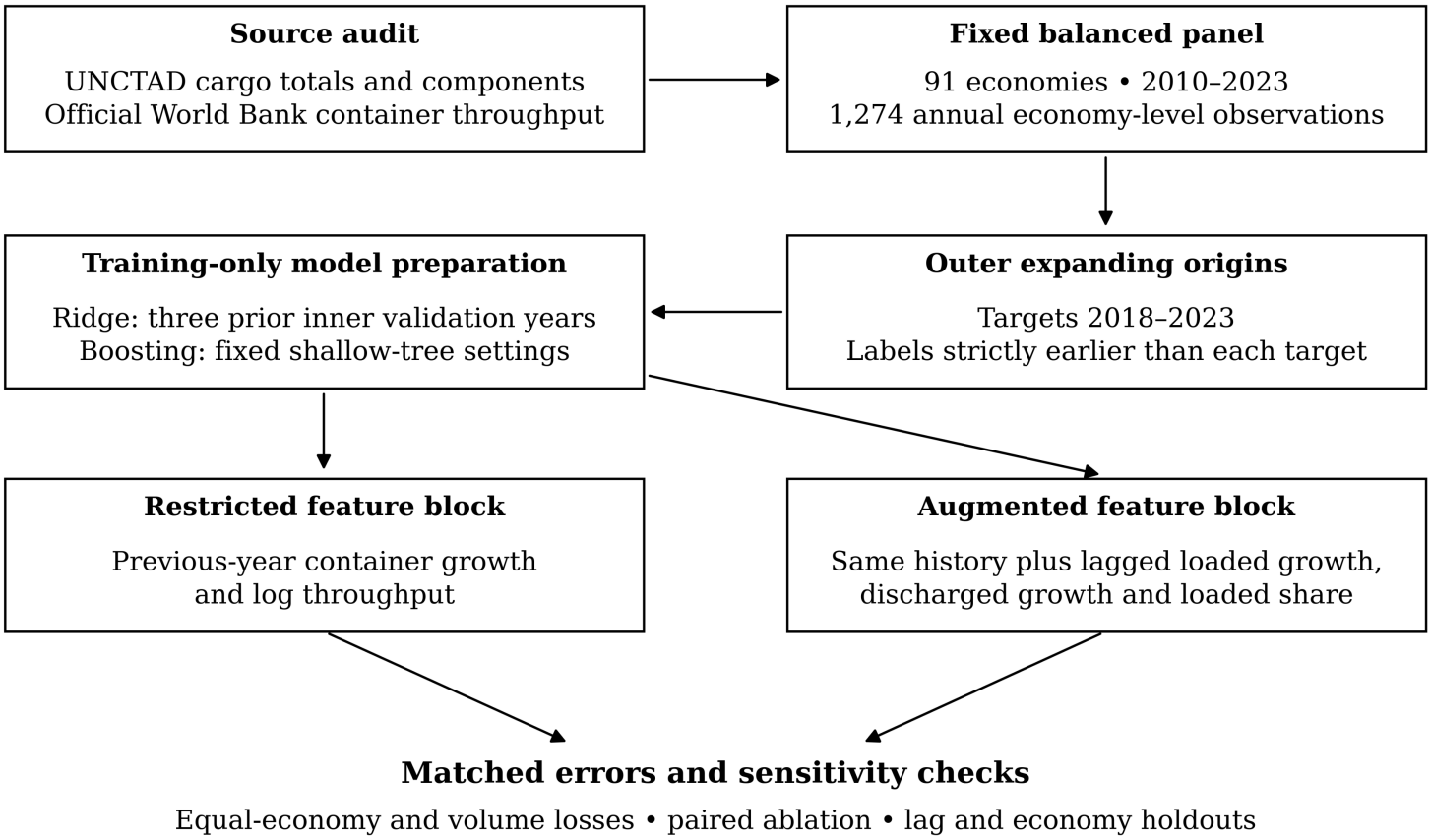

Figure 3

Unnumbered image in the Data and methods section.