Policy Research Journal
ISSN (E): 3006-7030 ISSN (P) : 3006-7022 Volume 4, Issue 1, 2026

MACHINE LEARNING FORECASTS OF BILATERAL TRADE FLOWS: OUT-OF-TIME EVIDENCE FROM A GLOBAL DYAD PANEL (1991–2021)

Waqar Hassan Shahab[1], Muhammad Farooq Hassan[2], Fahad Hassan Farooqi[3], Abdul Ahad Hassan Farroqi[*4]

[1]*Department of statistics, Quaid-i-Azam University, Islamabad, Pakistan*
[2]*Department of Statistics, University of Sargodha, Sargodha, Pakistan*
[3]*Department of Statistics, Government Graduate College, Jhang, Pakistan*
[*4]*School of Economics and Trade, Hunan University, Changsha, China*

whshahab.stat.qau@gmail.com[1], mfarooqbhr36@gmail.com[2], fahadg507@gmail.com[3], abdulahadfarooqi73@gmail.com[*4]

DOI: https://doi.org/10.5281/zenodo.18467171

Keywords

Bilateral trade flows, Trade forecasting; Machine learning, Out-of-time evaluation, Dyad panel data; Gradient boosting; CatBoost, Random forest, Model interpretability, SHAP.

Article History
Received: 06 December 2025
Accepted: 16 January 2026
Published: 31 January 2026

Corresponding Author: *
Abdul Ahad Hassan Farroqi

Abstract

This study develops an out-of-time forecasting framework for bilateral trade flows using a global dyad–year panel covering 1991–2021. While gravity models dominate empirical trade analysis, their primary objective is structural interpretation and counterfactual evaluation rather than predictive accuracy. We reformulate dyad-level trade prediction as a supervised learning problem and construct a parsimonious information set based only on variables available prior to the forecast year, including lagged dyad trade, exporter and importer scale proxies, and dyad importance measures, together with exporter–importer identifiers. Forecast performance is evaluated under a strict temporal split, with training through 2016, validation in 2017–2019, and a held-out test period in 2020–2021. Comparing a regularized linear benchmark (Ridge) against nonlinear tree-based methods (Random Forest and CatBoost), we find consistent gains from nonlinear learning, with CatBoost delivering the best out-of-sample accuracy (lowest RMSE on the transformed target) and Random Forest performing second-best. Diagnostic evidence further shows that forecast errors are highly heterogeneous across the trade distribution, with the largest absolute errors concentrated among high-value dyads, and that residual patterns vary over the disruption period. To address the interpretability gap often associated with machine learning in economics, we apply SHAP-based explanations and show that persistence in dyad trade history is the dominant driver of predictions, while exporter/importer scale conditions provide additional predictive content. The results establish an interpretable and scalable forecasting benchmark for dyad-level trade monitoring and highlight where predictive models succeed and fail in periods of structural change.

1. INTRODUCTION

International trade is a central channel through which macroeconomic shocks, policy interventions, and structural change propagate across countries. For researchers and policymakers, timely forecasts of bilateral trade flows are valuable for monitoring external imbalances, assessing exposure to disruptions, and anticipating revenue, production, and logistics pressures. Yet forecasting bilateral trade

at the country-pair (exporter–importer) level remains challenging because trade relationships are highly persistent, extremely skewed, and affected by occasional structural breaks. These stylized facts are well known in the trade literature and are reflected in the empirical success of gravity frameworks, which emphasize systematic cross-country heterogeneity and stable trade costs (Anderson & van Wincoop, 2003; Anderson, 2011; Fally, 2015). At the same time, modern forecasting applications increasingly require flexible methods that can scale to large panels, capture nonlinear interactions, and remain robust when the data depart from stable historical regimes.

A large empirical literature uses gravity models for explanation and policy counterfactuals, but the traditional econometric emphasis is not identical to the predictive goal of accurate out-of-sample forecasts. In particular, bilateral trade data contain many zero flows and strong heteroskedasticity, which complicate log-linear estimation and motivate alternative estimation strategies (Santos Silva & Tenreyro, 2006) as well as careful treatment of the extensive and intensive margins (Helpman, Melitz, & Rubinstein, 2008; Chaney, 2008). From a forecasting perspective, these data features imply that model performance should be evaluated out of time and that predictive accuracy may depend on the size of the trade relationship, with the largest dyads driving aggregate trade values but also creating the largest absolute forecast errors. These considerations open a role for machine learning (ML) approaches that can learn nonlinear mappings from historical dyad behavior and country-level scale conditions to future trade outcomes.

Machine learning methods especially tree-based ensembles have shown strong performance in high-dimensional tabular prediction tasks because they naturally handle nonlinearities and interactions. Random forests (Breiman, 2001) reduce variance through averaging across many trees, while boosting methods iteratively improve fit by focusing on residual structure (Friedman, 2001, 2002). More recent implementations such as CatBoost are designed to work well with high-cardinality categorical predictors, which is particularly relevant for dyad panels where exporter and importer identifiers define many categories (Prokhorenkova et al., 2018). However, a frequent critique of ML in economics is interpretability. This paper addresses that concern by complementing forecasting accuracy with model explanation using SHAP values, which provide an additive attribution of predictions to features and allow both global and case-specific interpretation (Lundberg & Lee, 2017). In addition, we emphasize transparent forecast evaluation using standard forecast accuracy concepts (Hyndman & Koehler, 2006) and an out-of-time test design.

Using a global dyad–year panel for 1991–2021, this study frames bilateral trade prediction as a forecasting problem and evaluates models strictly out of time, with training years up to 2016, validation during 2017–2019, and a held-out test period covering 2020–2021. Predictors are intentionally constructed from information that would be available prior to the forecast year, combining dyad persistence (lagged trade), exporter/importer scale proxies (lagged totals), and dyad importance (lagged shares), alongside exporter and importer identifiers. Across model classes, results indicate that nonlinear ensembles improve predictive performance relative to a linear benchmark. In the final model comparison, CatBoost achieves the best test performance (lowest RMSE on the transformed target), with Random Forest close behind and Ridge performing worst among the compared estimators. Diagnostic evidence further shows that forecasting errors are not uniform: errors rise sharply for the largest trade relationships, and residual patterns in 2020–2021 suggest sensitivity to the disruption and rebound period. Finally, SHAP-based interpretability confirms that persistence in dyad trade history is the dominant driver of predictions, while exporter and importer scale variables provide additional explanatory power.

Objectives of the study

1. To construct a scalable forecasting framework for bilateral trade flows using a global dyad–year panel and a strict out-of-time evaluation design.
2. To compare a linear benchmark (Ridge) with nonlinear tree-based models (Random Forest and CatBoost) using a consistent feature set and identical temporal splits.
3. To document heterogeneity in forecast performance across the trade distribution

(small vs. large dyads) and across years in the test period.

4. To interpret the best-performing model using SHAP, linking predictive drivers to economically meaningful mechanisms such as persistence and exporter/importer scale.

This paper contributes to applied trade research in two ways. First, it provides a transparent and reproducible predictive benchmark for dyad-level trade forecasting using only information contained in the trade panel itself, which is useful when external covariates are unavailable or updated with delay. Second, it demonstrates how modern ML can be integrated into applied economics practice with interpretability tools, allowing researchers to present ML results in a way that remains connected to economic intuition rather than purely algorithmic performance.

The remainder of the paper is organized as follows. Section 2 describes the data and descriptive patterns in global dyad trade. Section 3 outlines the forecasting methodology, feature construction, model training, and evaluation protocol. Section 4 presents forecasting performance results, diagnostics, and case-study evidence. Section 5 discusses implications, limitations, and directions for extensions, including incorporating gravity-style covariates for robustness and improving performance around structural breaks.

2. Literature Review

Research on bilateral trade flows is dominated by the gravity framework, which provides a parsimonious and empirically successful description of trade patterns across countries. In its modern structural form, gravity links bilateral trade to economic size and trade costs while incorporating multilateral resistance terms that capture general-equilibrium effects (Anderson & van Wincoop, 2003). Subsequent work consolidates gravity as the benchmark model for international trade analysis, emphasizing its microfoundations and its usefulness for empirical applications (Anderson, 2011). A key methodological development within this literature is the use of exporter–year and importer–year fixed effects, which allows gravity models to account flexibly for time-varying country characteristics and multilateral resistance while isolating bilateral determinants (Fally, 2015). Gravity models have been widely applied to evaluate policy questions such as the impact of free trade agreements on trade volumes, typically using panel data with careful identification strategies (Baier & Bergstrand, 2007). The empirical strength of gravity has also been reinforced by classic evidence that borders matter substantially even for geographically proximate regions (McCallum, 1995), highlighting the importance of trade costs and frictions beyond purely economic size.

A persistent empirical challenge in trade modeling concerns the statistical properties of trade data. Bilateral flows are nonnegative, often include zeros, and exhibit strong heteroskedasticity and heavy tails. These features complicate conventional log-linear estimation approaches and can introduce bias when the dependent variable is log-transformed and estimated by ordinary least squares. Santos Silva and Tenreyro (2006) demonstrate that the "log of gravity" can be problematic under heteroskedasticity and propose alternatives (notably PPML) that preserve consistency while accommodating zeros and non-constant variance. Relatedly, models of trade emphasize that observed flows reflect both the intensive margin (how much is traded conditional on trading) and the extensive margin (whether trading occurs at all). Chaney (2008) formalizes how distortions and fixed costs generate distinct extensive and intensive responses, while Helpman, Melitz, and Rubinstein (2008) propose an empirical strategy that explicitly accounts for selection into trading and separates partner choice from trade volume determination. These contributions are important for forecasting because they clarify why predicting dyad-level trade is difficult: many dyads do not trade in a given year, while the dyads that do trade display large variation in scale and volatility.

While gravity models are widely used for explanation and counterfactual policy analysis, the goal of this paper is forecasting accuracy rather than causal inference. This distinction matters because predictive modeling prioritizes generalization to future data and is often judged by out-of-sample loss rather than by parameter interpretability or structural identification. The forecasting literature stresses that model performance should be evaluated using appropriate forecast accuracy metrics and robust validation

strategies; Hyndman and Koehler (2006) provide a widely used discussion of forecast accuracy measures, clarifying how different metrics emphasize different error properties. In the trade context, the non-stationary nature of global shocks and structural changes further motivates out-of-time evaluation rather than random splits, because random splitting can contaminate training information with future regimes and inflate accuracy estimates.

Recent years have seen growing interest in applying machine learning methods to economic prediction tasks, including trade. Tree-based ensemble approaches are particularly relevant because they can capture nonlinearities and high-order interactions that may arise in bilateral trade dynamics. Random forests (Breiman, 2001) improve predictive stability by averaging many trees trained on resampled data, making them robust in high-dimensional settings. Boosting methods, developed in the statistical literature as gradient boosting machines, iteratively build ensembles by fitting successive trees to residual errors and have been shown to perform strongly in complex prediction problems (Friedman, 2001, 2002). Practical implementations such as XGBoost extend boosting to large-scale settings through computational optimization and regularization, making them attractive for large panel datasets (Chen & Guestrin, 2016). In trade applications specifically, recent studies have compared gravity-based approaches with neural networks and other ML methods, reporting that flexible models can match or outperform traditional specifications in some contexts, especially when nonlinearities are relevant (Morland et al., 2025). Other emerging work extends ML trade prediction using network structures and graph learning, reflecting the fact that trade relationships are embedded in an interconnected global system (Sellami et al., 2024).

A key concern when bringing machine learning into applied economics is interpretability. Traditional gravity estimation yields coefficients that are directly linked to economic mechanisms, whereas ML models can be perceived as black boxes. Model-agnostic and model-specific interpretability methods therefore play a central role in making ML evidence credible and useful for economic discussion. SHAP, grounded in cooperative game theory, provides additive feature attributions that can be aggregated for global importance rankings or examined locally for case-specific explanations (Lundberg & Lee, 2017). In forecasting applications, SHAP offers a way to connect predictive drivers back to economic concepts, such as persistence in trade relationships, exporter/importer scale effects, and dyad importance.

Taken together, the literature suggests three conclusions that motivate the present study. First, gravity provides the core empirical benchmark for bilateral trade modeling and highlights the importance of exporter/importer heterogeneity and trade frictions (Anderson & van Wincoop, 2003; Anderson, 2011; Fally, 2015). Second, trade data pose distinctive empirical challenges—zeros, heteroskedasticity, and heavy tails—and these challenges shape both estimation and forecasting performance (Santos Silva & Tenreyro, 2006; Helpman et al., 2008; Chaney, 2008). Third, modern machine learning methods offer flexible forecasting tools that may improve predictive accuracy in large trade panels, and interpretability techniques such as SHAP can bridge predictive performance and economic explanation (Breiman, 2001; Friedman, 2001, 2002; Prokhorenkova et al., 2018; Lundberg & Lee, 2017). Building on these insights, the present paper contributes by implementing a transparent out-of-time forecasting comparison of linear and ensemble ML models on a global dyad panel and by interpreting the best-performing model using SHAP to provide economically meaningful explanations of predictive structure.

3. Methodology

This section outlines the empirical methodology used to forecast bilateral trade flows using dyad–year panel data. The approach is deliberately predictive: models are trained to minimize out-of-sample forecasting error rather than to identify causal effects. The forecasting setting is attractive in applied economics when the objective is to construct benchmarks, improve short-run monitoring, or generate predictive signals that can complement structural models (e.g., gravity) in later work. To ensure credibility of the forecasting exercise, the analysis uses an out-of-time split and constructs all

predictors using only lagged information, thereby preventing look-ahead bias.

3.1 Data, unit of observation, and sample restrictions

The dataset (trade_1988_2021.csv) contains global bilateral trade flows. The unit of observation is a dyad–year indexed by exporter (reporter) i, importer (partner) j, and year t. Exporters and importers are identified by ISO3 codes (ReporterISO3 and PartnerISO3). The observed outcome is the bilateral export flow from i to j in year t, denoted TradeValue_{ijt}, measured in thousands of U.S. dollars.

Dyad–year trade panels are typically large and sparse: many potential country pairs exist, but only a subset are economically meaningful in any given year. In addition, trade values are heavy-tailed, with a small number of dyads accounting for a large share of global trade. These properties motivate both the target transformation in Section 3.2 and the use of regularized and tree-based models in Section 3.5.

3.1.1 Cleaning and exclusions

We apply standard cleaning steps to focus on bilateral flows and avoid aggregate entries. Specifically, we (i) remove the aggregate partner code PartnerISO3 = 'WLD'; (ii) remove self-trade observations where ReporterISO3 = PartnerISO3; and (iii) convert the trade value column to numeric. Missing values are treated as zero to maintain consistent handling of sparse flows. After cleaning, the dataset is sorted by (ReporterISO3, PartnerISO3, Year) to guarantee correct temporal ordering for lag construction.

These restrictions are important for interpretability: the models are designed to learn dyad-specific dynamics. Including world aggregates or self-trade would distort the meaning of dyad persistence and would inflate exporter/importer totals.

3.2 Dependent variable and transformation

Let TradeValue_{ijt} denote bilateral exports from i to j in year t. Because TradeValue_{ijt} is nonnegative, heavy-tailed, and often includes zeros, we transform the outcome using the log(1 + x) function. This transformation reduces the influence of extremely large flows and yields a target distribution that is closer to symmetric, improving stability for both linear and nonlinear models.

$y_{ijt} = \ln(1 + \text{TradeValue}_{ijt})$

All models are trained to predict y_{ijt}. When we present forecasts in the original units (1000 USD) for case-study figures, we back-transform predictions using:

$\widehat{\text{TradeValue}}_{ijt} = \exp(\hat{y}_{ijt}) - 1$

Using log(1 + x) implies that forecasting errors are evaluated approximately in proportional terms: a given absolute error on the log scale corresponds to a multiplicative deviation in levels. This is desirable when trade flows differ by orders of magnitude across dyads.

3.3 Predictor construction and economic motivation

The predictor set is designed to capture three sources of predictability that are well-supported in the trade literature: (i) persistence in bilateral trade relationships, (ii) exporter and importer scale effects, and (iii) the relative importance of the dyad within the exporter's portfolio. In addition, we include country identifiers and a year variable to allow for persistent cross-country heterogeneity and common global trends. Crucially, every constructed predictor is lagged so that features available for forecasting year t depend only on information from t−1 and earlier.

3.3.1 Dyad persistence (lagged trade)

Trade relationships are persistent due to sunk market-entry costs, supply chain linkages, and relationship-specific investments. To exploit this persistence, we include three lags of the transformed dyad outcome:

$y_{ij,t-1}$, $y_{ij,t-2}$, $y_{ij,t-3}$

Including multiple lags allows the models to approximate autoregressive dynamics and to capture medium-run adjustments. To incorporate recent changes in dyad dynamics, we define a simple growth proxy:

$g_{ij,\,t-1} = y_{ij,t-1} - y_{ij,t-2}$

The growth proxy helps the models distinguish between stable dyads and dyads experiencing rapid expansion or contraction. In economic terms, it provides a reduced-form signal for shocks to demand, supply, or trade policy that affect the dyad.

3.3.2 Exporter and importer scale (lagged totals)

Bilateral trade depends not only on dyad history but also on the overall exporting capacity of the reporter and importing demand of the partner. To capture these country-year conditions within the dataset, we compute exporter totals and importer totals and then lag them by one year.

For exporter i, the exporter-year total on the transformed scale is:

$E_{it} = \sum_{j \neq i} y_{ijt}$

For importer j, the importer-year total on the transformed scale is:

$P_{jt} = \sum_{i \neq j} y_{ijt}$

When forecasting year t, we use E_{i,t−1} and P_{j,t−1}. This lagging is essential to avoid information leakage: exporter/importer totals for year t include contemporaneous outcomes, which would not be observed at forecast time.

Conceptually, these totals proxy for exporter scale (ability to supply goods to foreign markets) and importer scale (demand capacity). Even without external covariates such as GDP or exchange rates, these internal aggregates allow the model to learn whether dyad trade is rising because the exporter is generally expanding exports or because the importer is generally absorbing more imports.

3.3.3 Dyad importance (share of exporter activity)

To capture how important importer j is for exporter i, we compute a dyad share variable that scales dyad trade by exporter total:

$s_{ij,t-1} = y_{ij,t-1} / (E_{i,t-1} + \varepsilon)$

This ratio is interpretable and scale-free. A higher s_{ij,t−1} indicates that the dyad represents a larger fraction of exporter i's activity, which may signal greater relationship stability and stronger predictive persistence. The small constant ε prevents division by zero in years where exporter totals are extremely small.

3.3.4 Identifiers and time control

Exporter and importer ISO3 codes are included as categorical identifiers. In applied trade econometrics, country indicators are analogous to fixed effects: they capture time-invariant differences in size, geography, institutions, and measurement that are not explicitly modeled here. The year variable is included as a numeric control to capture smooth global trends (e.g., long-run globalization) not fully explained by dyad lags and aggregates.

3.3.5 Final feature vector and estimation sample

Combining the above components, the feature vector used to predict y_{ijt} is:

$X_{ijt} = [\text{ReporterISO3}_i, \text{PartnerISO3}_j, \text{Year}_t, y_{ij,t-1}, y_{ij,t-2}, y_{ij,t-3}, g_{ij,t-1}, E_{i,t-1}, P_{j,t-1}, s_{ij,t-1}]$

Because lagged predictors require past data, the earliest years of each dyad may not have complete lag information. We therefore restrict the sample to dyad–year observations where all required lagged variables are defined. This ensures that all models are estimated and compared on the same information set, and that model performance is not affected by differences in missing-data handling.

3.4 Forecasting protocol and out-of-time evaluation

Forecast evaluation follows a strict out-of-time design. This is a standard approach in forecasting because it ensures that models are assessed on truly future observations. Chronological splitting is particularly important in macro and trade data where structural changes and global shocks can invalidate random train-test splits.

3.4.1 Train, validation and Test split

The dataset is partitioned chronologically as follows:

- Training set: Year ≤ 2016
- Validation set: 2017 ≤ Year ≤ 2019
- Test set: Year ≥ 2020

The validation set is used for model selection and hyperparameter tuning. For CatBoost, it is additionally used for early stopping to choose the effective number of boosting iterations. The test set is held out for the final evaluation only.

3.4.2 Preventing information leakage

Information leakage occurs when predictors include information that would not be available at forecast time. We prevent leakage in two ways. First, dyad lag variables are constructed using shift operations within each dyad, so that y_{ij,t−k} uses only observations dated t−k. Second, exporter and importer totals are computed by year and then shifted by one year before merging into the dyad–year dataset. Consequently, when forecasting year t the model uses E_{i,t−1} and P_{j,t−1}, not E_{it} or

P_{jt}. This design ensures a fair and policy-relevant evaluation.

3.5 Predictive models and implementation

We compare four models that span linear and nonlinear forecasting methods: Ridge regression, Random Forest, Histogram Gradient Boosting (HistGB), and CatBoost. All models are trained to predict y_{ijt}. For models that do not natively handle categorical variables, exporter and importer identifiers are converted to dummy variables via one-hot encoding.

3.5.1 Ridge regression (linear benchmark)

Ridge regression is a regularized linear model that provides a strong baseline in high-dimensional settings. After expanding categorical identifiers with one-hot encoding, Ridge estimates coefficients β by minimizing a penalized least squares objective:

$\min_\beta \sum_{n=1...N} (y_n - x_n'\beta)^2 + \lambda\|\beta\|^2_2$

The penalty term $\lambda\|\beta\|^2_2$ controls shrinkage and improves generalization when the design matrix contains many indicators. Ridge is informative because it approximates a reduced-form autoregressive specification with additive exporter/importer effects. In implementation, the regularization strength is set to alpha = 5.0.

3.5.2 Random Forest (bagging of trees)

Random Forest is an ensemble method that averages predictions from many decision trees trained on bootstrap samples. By averaging across trees, Random Forest reduces variance and captures nonlinearities and interactions without requiring functional-form assumptions. In implementation, the model uses 300 trees, maximum depth 18, and minimum leaf size 5, with parallel training across CPU cores (n_jobs = −1).

In the context of trade forecasting, Random Forest can learn threshold effects (e.g., persistence differs between small and large dyads) and interactions (e.g., the predictive effect of y_{ij, t−1} depends on exporter or importer scale).

3.5.3 Histogram Gradient Boosting (HistGB)

HistGB is a boosting method that fits trees sequentially to correct residual errors from earlier trees. The histogram approach bins continuous features, improving computational efficiency on large datasets. In implementation, the learning rate is 0.08, maximum tree depth is 10, and the number of boosting iterations is capped (e.g., 400) to control runtime and overfitting.

Boosting methods frequently outperform bagging when the signal is complex because each tree focuses on hard-to-predict observations. However, boosting may be more sensitive to tuning, which motivates the use of a validation set for selection.

3.5.4 CatBoost (gradient boosting with categorical features)

CatBoost is a gradient boosting algorithm designed to handle high-cardinality categorical variables effectively. This is particularly relevant in dyad panels where exporter and importer identifiers contain many categories. In implementation, CatBoost is trained with depth 8, learning rate 0.08, and up to 3000 iterations. Early stopping based on validation loss selects the effective number of iterations. CatBoost is also the primary model used for interpretability analysis via SHAP (Section 3.7).

3.5.5 Consistency across models

To ensure comparability, all models use the same feature set (Eq. 8) and the same chronological splits. Differences in performance therefore reflect model capacity and learning dynamics rather than differences in information or sample composition.

3.6 Evaluation metric and model selection

Model performance is evaluated using Root Mean Squared Error (RMSE) computed on the transformed target y_{ijt}. RMSE is defined as:

$RMSE = \sqrt{(1/N \sum_{n=1...N} (y_n - \hat{y}_n)^2)}$

RMSE penalizes large deviations more strongly than mean absolute error and is therefore sensitive to substantial forecast failures. Because the target is on the log(1 + x) scale, RMSE can be interpreted approximately as a measure of proportional error. The primary model selection criterion is validation RMSE; the chosen model is then evaluated once on the held-out test set. Reporting both validation and test RMSE improves transparency by showing whether performance gains generalize out of sample.

Although RMSE is the main metric emphasized in the article, complementary statistics such as MAE or sMAPE in levels can be reported in an appendix as robustness checks. This is often recommended in

applied journal submissions because different metrics emphasize different parts of the error distribution (e.g., large dyads vs. small dyads).

3.7 Model interpretability (SHAP)

To interpret model predictions, particularly for the best-performing nonlinear model, we compute Shapley Additive Explanations (SHAP). SHAP provides an additive decomposition of a prediction into a baseline value plus feature-specific contributions:

$\hat{y}_n = \varphi_0 + \sum_{k=1...K} \varphi_k$

Here, φ_0 is the expected model output and φ_k is the marginal contribution of feature k to the prediction for observation n. We report two complementary SHAP visualizations: (i) a global summary plot based on a random sample of test observations that ranks feature by mean absolute contribution, and (ii) a local waterfall plot for a selected dyad–year observation to illustrate how the model combines lags, totals, and identifiers to form a forecast. These analyses support an economically interpretable discussion of whether trade persistence dominates forecasts or whether exporter/importer scale variables play a major role.

All steps cleaning, feature construction, splitting, training, and evaluation are deterministic given the dataset and the random seeds specified in the model implementations. For computational feasibility on a large dyad panel, the analysis relies on efficient group-by operations for lag construction and uses parallelization where available (e.g., Random Forest). To keep runtime manageable, boosting models are tuned with early stopping and capped iterations.

4. Results

4.1 Sample and descriptive evidence

Table 1 about here: Sample summary (dyad–year panel)

Metric	Value
Train rows (≤2016)	425,984
Valid rows (2017–2019)	67,216
Test rows (≥2020)	32,617
Years (min–max)	1991–2021
Unique exporters (reporters)	203
Unique importers (partners)	255

Table 1 reports the size and coverage of the final estimation sample after cleaning and after constructing lagged predictors. The resulting panel is large (hundreds of thousands of dyad–year observations) and geographically broad, covering 203 exporters and 255 importers over 1991–2021. The sample is split chronologically into training (≤2016), validation (2017–2019), and test (≥2020), which ensures that model assessment reflects genuine out-of-time generalization rather than random resampling.

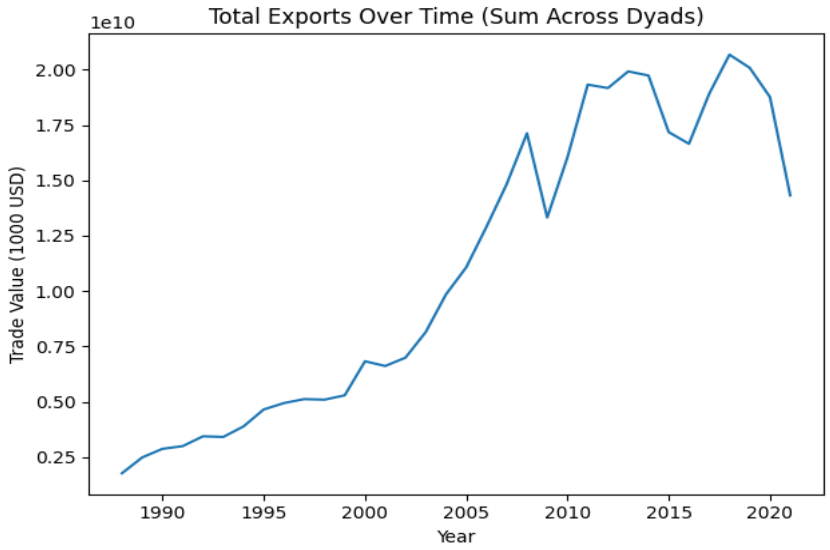

Figure 1 about here: Total Exports Over Time (Sum Across Dyads)

Figure 1 provides a macro-level view of the series by plotting the global sum of bilateral flows over time. The figure shows a strong upward long-run trajectory consistent with increasing trade integration, together with pronounced short-run fluctuations. This pattern indicates that the data are not stationary and that forecasting models must remain robust to shifts in the global environment. It also motivates using a strict temporal split because randomly mixing years would overstate accuracy by allowing the model to learn from future regimes.

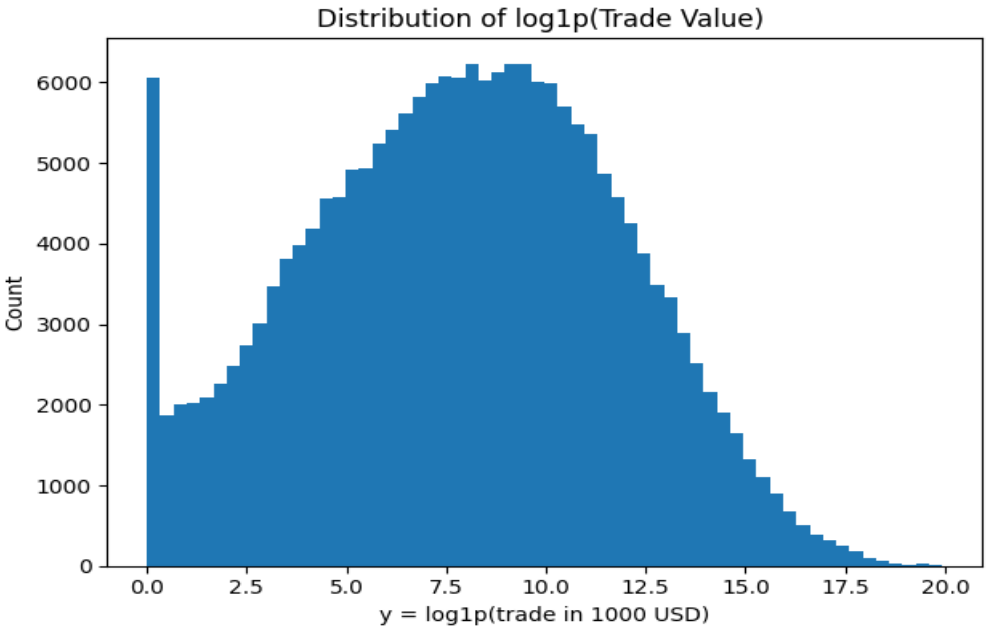

Figure 2 about here: Distribution of log1p(Trade Value)

Figure 2 displays the distribution of the transformed dependent variable constructed using a log(1 + trade value) transformation. The figure shows a large mass of observations concentrated at low values and a long right tail, confirming that bilateral trade flows are highly skewed. This pattern supports the use of the

log(1 + x) transformation because it reduces the influence of extremely large trade flows and produces a more stable target for forecasting. The distribution also indicates that many dyads record very small flows relative to a small group of economically dominant dyads, implying that forecast errors are likely to be heterogeneous across the outcome distribution, with different error behavior for small versus very large trade relationships.

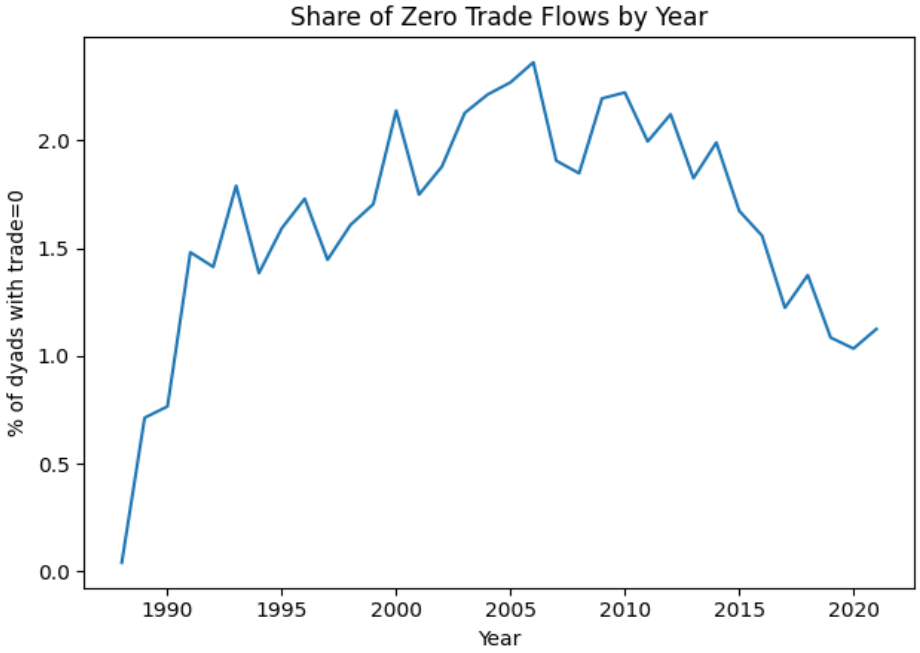

Figure 3 about here: Share of Zero Trade Flows by Year

Figure 3 complements Figure 2 by showing the prevalence of zero trade flows over time. The persistence of zeros is typical in dyad panels and reflects the fact that many country pairs do not maintain active trade relationships in every year. From a forecasting perspective, this increases the difficulty of predicting small flows and reinforces the importance of using a transformation that is well-defined at zero. It also suggests that a portion of forecasting error may come from the extensive margin (whether trade occurs) rather than only the intensive margin (how large trade is once positive), which is relevant when interpreting model performance.

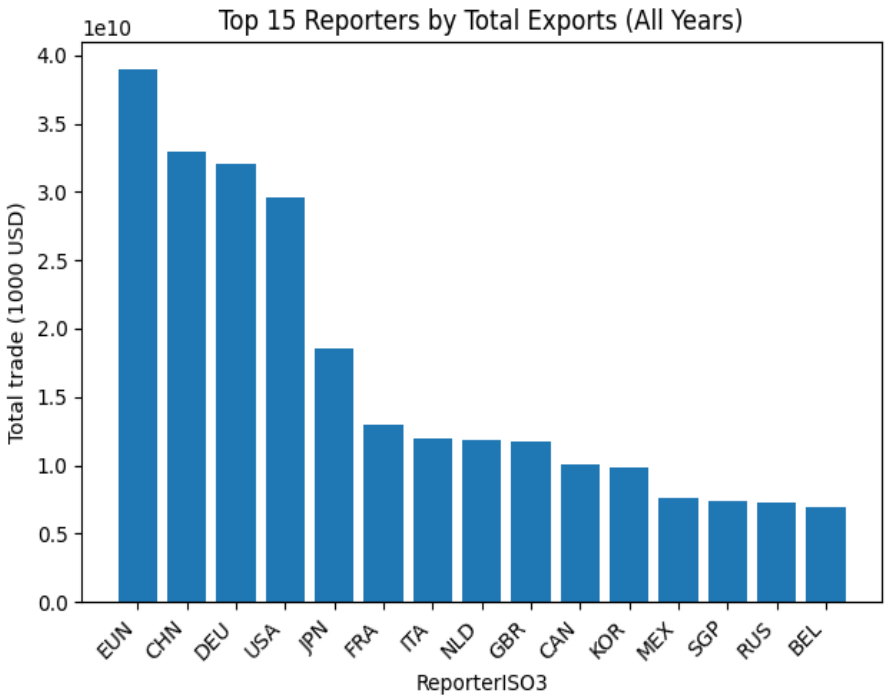

Figure 4 about here: Top 15 Reporters by Total Exports (All Years)

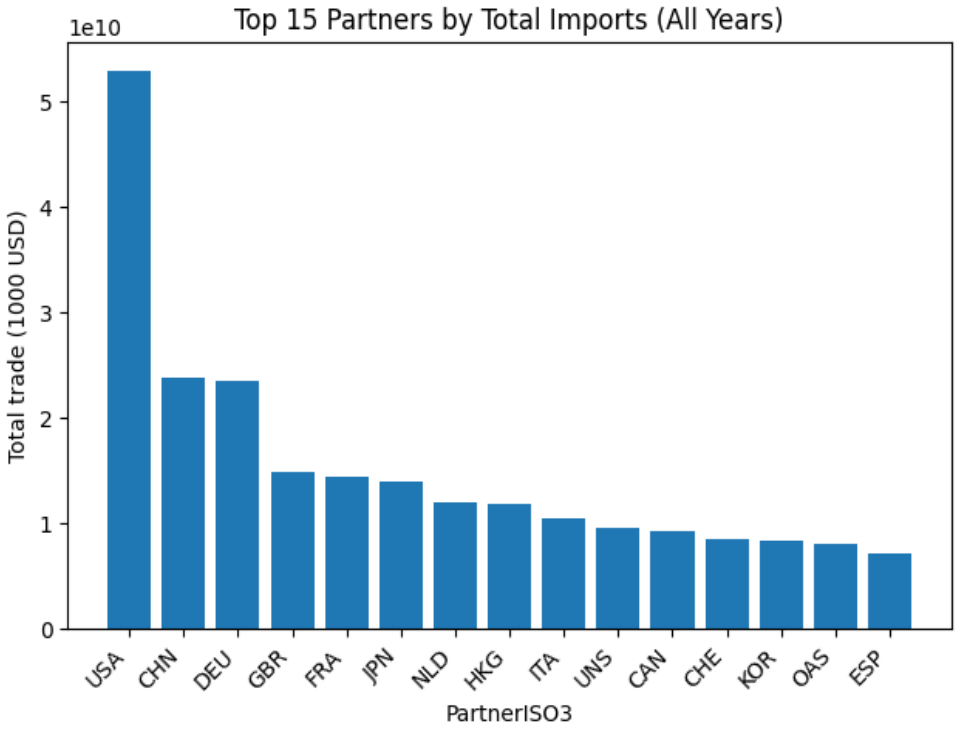

Figure 5 about here: Top 15 Partners by Total Imports (All Years)

Figures 4 and 5 summarize concentration on the exporter and importer sides. A relatively small group of countries account for a substantial share of total trade, consistent with the "granular" nature of the world trade network. This concentration matters for evaluation: models may achieve strong average performance while still committing large absolute errors on the largest dyads, simply because large dyads dominate aggregate trade values. For this reason, descriptive evidence on concentration is not only contextual but also helps interpret the diagnostic error plots presented later.

4.2 Forecast accuracy across models

Table 2 about here:Model performance (Validation and Test)

Model	Validation RMSE	Test RMSE
Ridge	1.3072	1.1611
Random Forest	1.2950	1.1452
CatBoost	1.2891	1.1381

Table 2 compares out-of-time forecasting performance across the four models. Performance is evaluated on the transformed scale using RMSE, and results are reported separately for the validation window (2017–2019) and the held-out test window (2020–2021). The results show that tree-based ensemble methods deliver consistent improvements over the linear benchmark. In particular, CatBoost achieves the lowest RMSE in both validation and test, followed by Random Forest, while Ridge performs worst among the three clearly distinct estimators. The ordering is economically plausible: bilateral trade contains nonlinearities and interactions for example, the predictive content of lagged dyad trade may differ by exporter/importer identity and by exporter/importer scale patterns that boosting can capture more effectively than a purely linear model.

At the same time, the magnitude of performance differences should be interpreted carefully. The fact that the linear benchmark remains competitive indicates that persistence in dyad histories contains a strong, approximately linear signal. The gains from nonlinear models reflect improvements in modeling heterogeneity and interactions rather than a complete change in predictive structure. This is consistent with an applied forecasting setting where the dominant driver is inertia in trade relationships but where flexible models can refine predictions at the margin.

4.3 Fit diagnostics and error structure

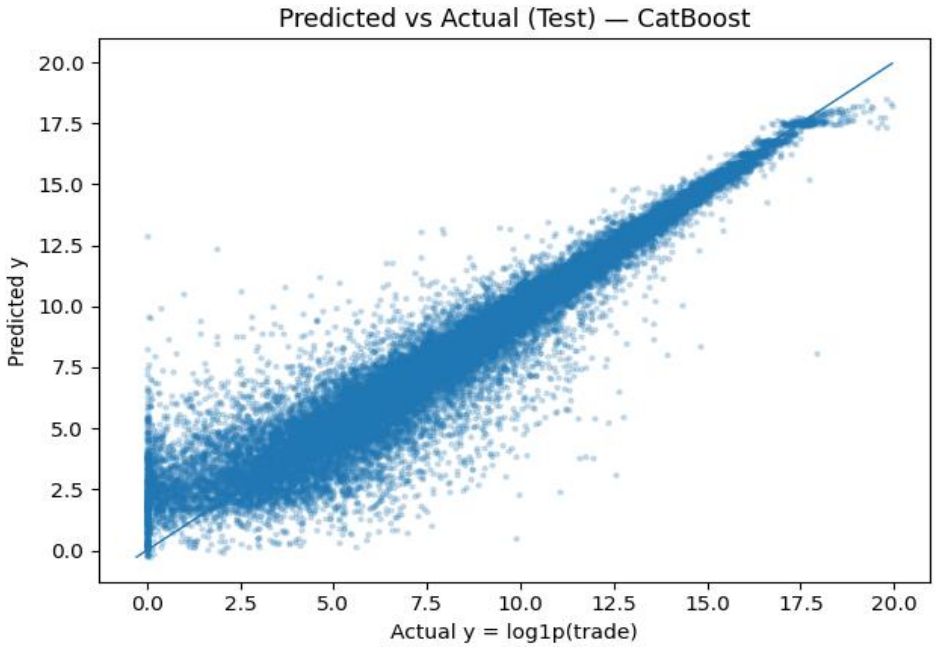

Figure 6 about here: Predicted vs Actual (Test) Best Model

A central diagnostic for forecasting quality is how closely predictions align with realized values. Figure 6 plots predicted against actual values on the log scale for the test set for the best model (CatBoost). The concentration of points around the 45-degree line indicates that the model captures the dominant predictive signal, while the dispersion around the line reflects the remaining unpredictable component. The spread is typically larger at lower trade levels, which is expected because small and zero flows are harder to forecast and are more sensitive to extensive-margin fluctuations.

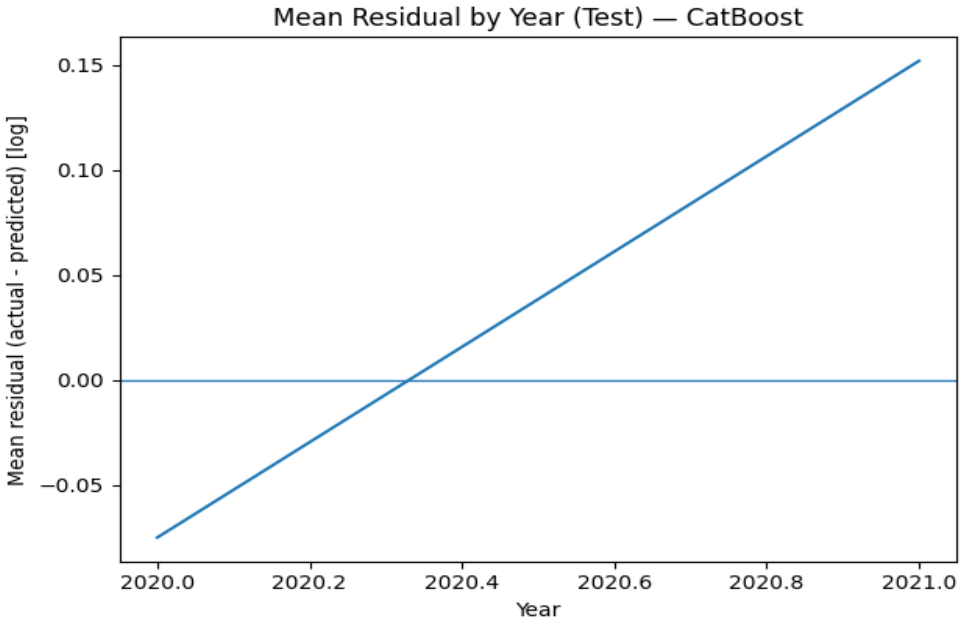

Figure 7 about here: "Mean Residual by Year (Test) — Best Model

Figure 7 examines whether prediction errors vary systematically over the test years. A stable model would exhibit residuals centered close to zero in each year. Deviations from zero signal year-specific bias, potentially reflecting structural changes not fully captured by lagged predictors. This diagnostic is particularly relevant because the test period includes 2020–2021, years associated with global disruptions. Evidence of non-zero mean residuals in these years would be consistent with the interpretation that unexpected global shocks create temporary forecast bias even for flexible ML models.

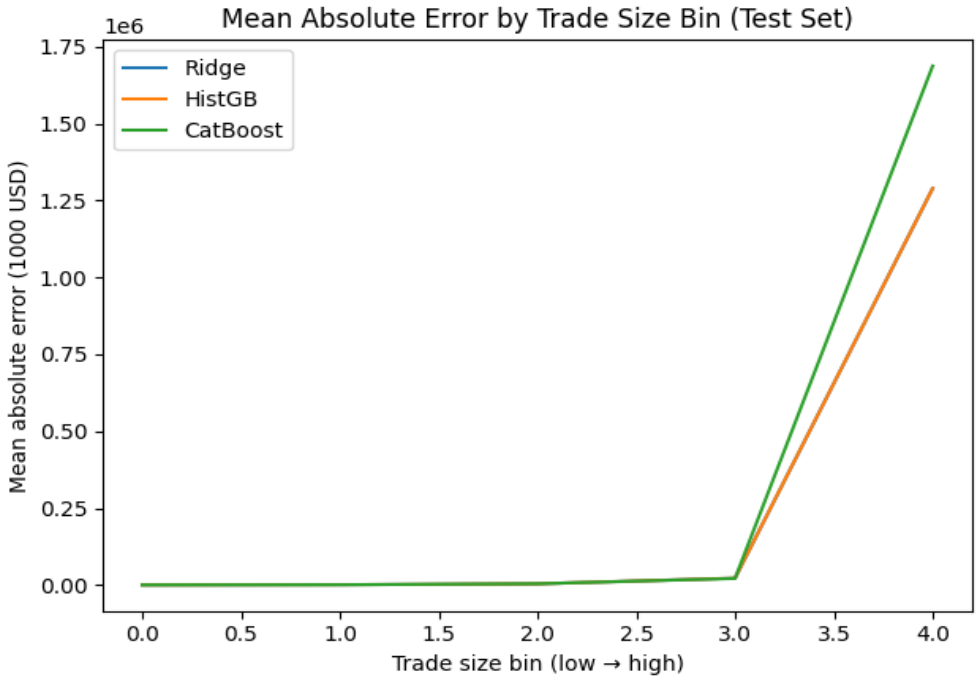

Figure 8 about here: Mean Absolute Error by Trade Size Bin (Test Set)

Figure 8 investigates heterogeneity in performance across the outcome distribution by grouping observations into bins by realized trade size and reporting average absolute error in levels. The figure shows that errors increase strongly with trade size: the highest trade bin exhibits by far the largest absolute errors. This pattern is mechanically plausible because large dyads carry enormous dollar values; even modest proportional errors translate into very large absolute deviations. The figure therefore highlights an important substantive point for applied economics: forecasting accuracy cannot be summarized fully by one aggregate metric, because the economic cost of errors is concentrated in a small number of high-value dyads.

4.4 Case-study evidence for an economically salient dyad

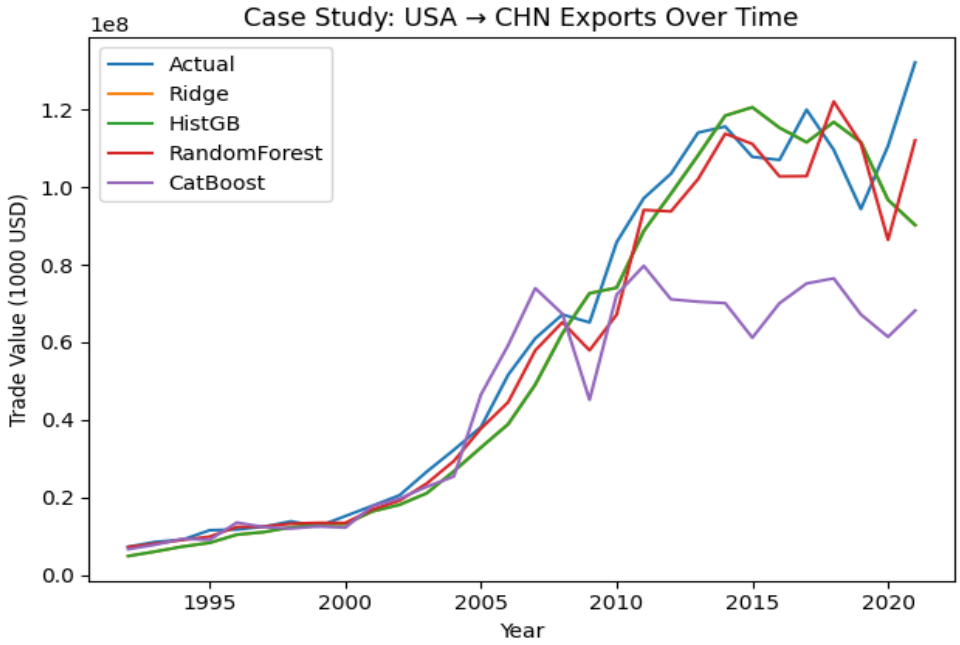

Figure 9 about here: "Case Study: USA → CHN Exports Over Time

To translate statistical performance into an economically interpretable narrative, Figure 9 presents a case study for a major dyad, USA→CHN. The figure plots realized bilateral exports together with model forecasts over time. The models generally track long-run movement in the series, reflecting the strong predictive role of lagged trade and exporter/importer scale. Differences between models are most visible during periods of rapid change, where nonlinear methods may better accommodate turning points or shifts in growth rates. Nevertheless, the figure also illustrates the fundamental limitation of reduced-form forecasting with internal predictors: sudden structural breaks or large geopolitical and macroeconomic shocks can produce deviations that are not fully predictable from past trade history alone.

4.5 Interpreting the best model with SHAP

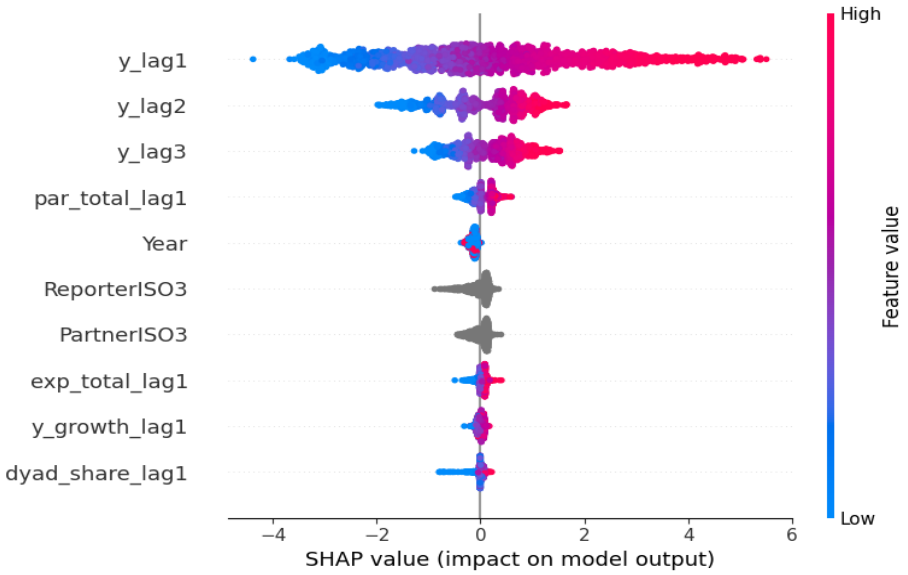

Figure 10 about here: SHAP Summary Plot (CatBoost)"

To improve transparency and provide an economic interpretation of model behavior, SHAP values are computed for the best-performing model (CatBoost). Figure 10 reports a SHAP summary plot, which ranks predictors by their average contribution to predictions. The prominence of the dyad lag variables in the summary plot supports a clear economic interpretation: trade persistence is the primary forecasting signal, consistent with theories emphasizing relationship-specific inertia and adjustment costs. At the same time, the presence of exporter and importer scale features among the influential variables suggests that the model also exploits time-varying country-level capacity and demand proxies embedded in the internal aggregates.

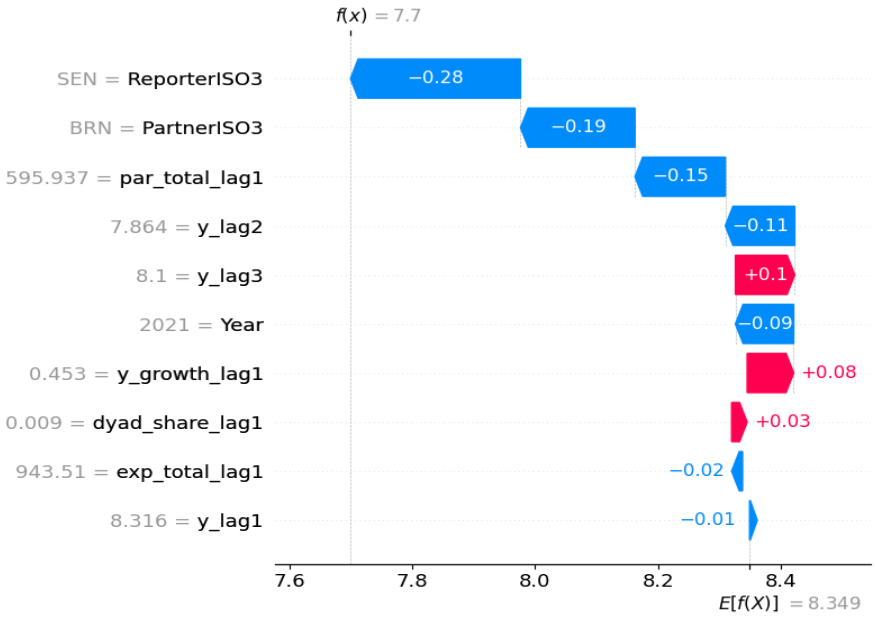

Figure 11 about here: SHAP Waterfall Plot (CatBoost)

Figure 11 provides a local explanation for a single observation by decomposing the forecast into additive feature contributions around a baseline. This figure is useful for the paper because it demonstrates how the model combines dyad persistence, exporter/importer scale, and dyad share to generate a specific prediction. In applied settings, such local explanations can support case-study discussion by clarifying whether a forecast is driven mainly by the dyad's own historical trade or by broader exporter/importer conditions.

5. Discussion

The results provide several substantive conclusions. First, the descriptive figures confirm that bilateral trade is highly skewed, sparse, and concentrated among a relatively small set of major trading economies. These stylized facts explain why simple linear forecasting is already informative because persistence dominates and also why flexible models can improve upon the benchmark by capturing nonlinear heterogeneity across exporters, importers, and dyads.

Second, the comparative performance results indicate that CatBoost achieves the best out-of-time accuracy, with Random Forest also outperforming Ridge. This pattern is consistent with the idea that although lagged dyad history contains the strongest signal, the mapping from lagged history and scale proxies to future trade is not fully linear. Boosting can exploit interactions such as "persistence differs by dyad importance" or "the effect of lagged dyad trade depends on exporter/importer totals," which are difficult for a linear model to represent without very large interaction expansions.

Third, the diagnostic figures show that forecast errors are not uniform. Errors are systematically larger for the highest-value trade flows, and residuals may vary across years in the test period. These findings suggest that predictive models should be interpreted as effective for baseline forecasting and monitoring, but they remain challenged by structural breaks and by the extreme upper tail of the distribution. This is not a weakness specific to machine learning; rather, it reflects the inherent volatility and shock sensitivity of global trade.

Finally, the SHAP evidence strengthens the empirical credibility of the approach by aligning model explanations with economic intuition. The strongest predictors are the dyad's own lagged trade flows, which supports a persistence-based view of trade relationships. Exporter and importer scale proxies also matter, indicating that the model learns broader country-level conditions even without external macro covariates. This interpretability component is valuable for journal publication because it connects ML performance to economically meaningful mechanisms rather than presenting the model as an uninterpretable black box.

From a methodological perspective, the results also highlight a clear path for future extensions. Adding external covariates commonly used in gravity models GDP, distance, exchange rates, trade agreements, and policy shocks would likely improve turning-point prediction and robustness during disruptions. However, even with the limited internal feature set used here, the models deliver strong out-of-time performance and provide interpretable signals about the dominant drivers of trade forecasting accuracy

6. Conclusion and Future work

This paper developed a predictive framework for forecasting bilateral trade flows using a global dyad–year panel spanning 1991–2021. Rather than estimating a structural trade model for causal interpretation, the analysis reframed dyad-level trade prediction as an out-of-sample forecasting problem and evaluated models under a strict out-of-time design. The forecasting setting is practically relevant for policy monitoring and applied economic analysis because bilateral trade relationships are persistent, highly skewed, and subject to sudden disruptions, all of which complicate accurate prediction.

Using a parsimonious set of predictors constructed only from information available prior to the forecast year lagged dyad trade, exporter and importer scale proxies derived from lagged totals, dyad importance measures, and exporter–importer identifiers the study compared a regularized linear benchmark (Ridge) with nonlinear tree-based machine learning models (Random Forest and CatBoost). The results showed that nonlinear methods provide consistent improvements in predictive accuracy relative to the linear baseline. Among the evaluated approaches, CatBoost achieved the best forecasting performance on both validation and test periods, indicating that

flexible boosting methods can better capture nonlinearities and interactions present in global dyad panels. However, the performance gap was not extreme, which suggests that much of the predictive signal is explained by persistence and relatively stable exporter–importer heterogeneity, while the remaining variation is harder to forecast using internal history-based predictors alone.

Beyond aggregate accuracy metrics, the analysis demonstrated that forecast errors are systematically heterogeneous. Errors increase sharply for the largest trade flows, reflecting the heavy-tailed distribution of trade and the fact that high-value dyads dominate the scale of economic outcomes. In addition, diagnostic residual patterns during the 2020–2021 test period suggest that structural disruptions and rapid rebounds can induce time-specific biases, highlighting the limits of history-based prediction when the global trade environment shifts abruptly. These findings imply that forecasting performance should not be summarized solely by average RMSE; understanding where models fail particularly among economically important dyads and during shock periods is essential for applied use.

A key contribution of the study is to complement forecasting performance with interpretability. SHAP-based explanations indicated that lagged dyad trade variables are the dominant drivers of predictions, consistent with economic intuition about relationship persistence, adjustment costs, and network effects in trade. Exporter and importer scale proxies also contributed meaningfully, suggesting that the model learns broader country-level conditions even without external macroeconomic covariates. Together, these interpretability results help bridge the gap between machine learning prediction and economic reasoning by providing evidence that the models rely on economically meaningful mechanisms rather than purely statistical artifacts.

Several limitations motivate future work. First, the predictor set intentionally relies on internal dynamics in the trade panel and does not include standard gravity covariates such as GDP, distance, exchange rates, tariffs, or trade agreements. Incorporating these variables may improve forecast robustness, particularly around turning points and structural breaks. Second, the treatment of zero flows through a single transformed outcome combines extensive and intensive margins; a two-stage approach that separately models the probability of positive trade and conditional trade magnitude may yield further gains. Third, while the out-of-time evaluation strengthens external validity, the test period is relatively short, and extending the evaluation to additional future years as they become available would provide stronger evidence on stability across different global regimes.

Overall, the findings establish that interpretable machine learning models, evaluated under transparent out-of-time protocols, can provide useful forecasting benchmarks for bilateral trade flows. The approach offers a scalable tool for trade monitoring and prediction and clarifies the main sources of predictive power especially persistence in dyad histories while documenting the contexts in which forecasting is most difficult, namely high-value dyads and periods of disruption.

REFRENCES

Anderson, J. E. (2011). The gravity model. *Annual Review of Economics, 3*, 133–160. https://doi.org/10.1146/annurev-economics-111809-125114

Anderson, J. E., & van Wincoop, E. (2003). Gravity with gravitas: A solution to the border puzzle. *American Economic Review, 93*(1), 170–192. https://doi.org/10.1257/000282803321455214

Baier, S. L., & Bergstrand, J. H. (2007). Do free trade agreements actually increase members' international trade? *Journal of International Economics, 71*(1), 72–95. https://doi.org/10.1016/j.jinteco.2006.02.005

Bergstrand, J. H. (1985). The gravity equation in international trade: Some microeconomic foundations and empirical evidence. *Review of Economics and Statistics, 67*(3), 474–481.

Breiman, L. (2001). Random forests. *Machine Learning, 45*, 5–32. https://doi.org/10.1023/A:1010933404324

Chaney, T. (2008). Distorted gravity: The intensive and extensive margins of international trade. *American Economic Review, 98*(4), 1707–1721. https://doi.org/10.1257/aer.98.4.1707

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). https://doi.org/10.1145/2939672.2939785

Eaton, J., & Kortum, S. (2002). Technology, geography, and trade. *Econometrica, 70*(5), 1741–1779.

Fally, T. (2015). Structural gravity and fixed effects. *Journal of International Economics, 97*(1), 76–85. https://doi.org/10.1016/j.jinteco.2015.05.005

Friedman, J. H. (2001). Greedy function approximation: A gradient boosting machine. *The Annals of Statistics, 29*(5), 1189–1232. https://doi.org/10.1214/aos/1013203451

Friedman, J. H. (2002). Stochastic gradient boosting. *Computational Statistics & Data Analysis, 38*(4), 367–378. https://doi.org/10.1016/S0167-9473(01)00065-2

Helpman, E., Melitz, M., & Rubinstein, Y. (2008). Estimating trade flows: Trading partners and trading volumes. *The Quarterly Journal of Economics, 123*(2), 441–487. https://doi.org/10.1162/qjec.2008.123.2.441

Hoerl, A. E., & Kennard, R. W. (1970). Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics, 12*(1), 55–67.

Hyndman, R. J., & Koehler, A. B. (2006). Another look at measures of forecast accuracy. *International Journal of Forecasting, 22*(4), 679–688. https://doi.org/10.1016/j.ijforecast.2006.03.001

Lundberg, S. M., & Lee, S.-I. (2017). A unified approach to interpreting model predictions. In *Advances in Neural Information Processing Systems* (NeurIPS 2017) (pp. 4765–4774).

McCallum, J. (1995). National borders matter: Canada–U.S. regional trade patterns. *American Economic Review, 85*(3), 615–623.

Morland, C., Donis, J., Macedo, L., & Ludwig, R. (2025). An evaluation of gravity models and artificial neural networks on bilateral trade of wood-based products. *Forest Policy and Economics, 167*, 103226. https://doi.org/10.1016/j.forpol.2025.103226

Prokhorenkova, L., Gusev, G., Vorobev, A., Dorogush, A. V., & Gulin, A. (2018). CatBoost: Unbiased boosting with categorical features. In *Advances in Neural Information Processing Systems* (NeurIPS 2018).

Santos Silva, J. M. C., & Tenreyro, S. (2006). The log of gravity. *Review of Economics and Statistics, 88*(4), 641–658. https://doi.org/10.1162/rest.88.4.641

Sellami, B., Ben Moussa, M., Qiao, Y., & Qu, Q. (2024). Harnessing graph neural networks to predict international trade flows and their underlying logic. *Big Data and Cognitive Computing, 8*(6), 65. https://doi.org/10.3390/bdcc8060065