Predicting the Logistics Performance Index (LPI) with Tree-Based Machine Learning Methods

Eric Torres[a]

[a]*Email: etorresram@gmail.com*

Abstract—The Logistics Performance Index (LPI) from the World Bank serves as the main standard for evaluating international logistics performance. The 2023 revision introduced major methodological improvements by utilizing proprietary Big Data shipment tracking information based on KPIs. This study explores whether modern machine learning techniques can effectively use structured, open-access cross-country indicators (23 features) to predict the official LPI scores. The work examines different regression algorithms through a comprehensive evaluation of 139 countries by employing different methods including Three Based (e.g., Extra Trees, LightGBM, Random Forest), Linear models (Linear Regression, Ridge, and Bayesian Ridge), and a distance based model (KNN Regresor). The analysis includes rigorous treatment of missing data through iterative imputation, robust model assessment using cross-validation, and interpretability via SHAP values. The results show that tree-based methods demonstrate the highest predictive accuracy and generalization in approximating LPI scores when including geographic zones and income-level-groups as features, especially the Extra Tree model (Test R^2 = 0.889, RMSE = 0.195). After model validation, predictive estimates were produced for 13 countries that were not included in the official 2023 LPI update. These estimates could serve as a benchmark for countries where the availability and quality of data do not meet the requirements of the official methodology currently used to construct the index.

Keywords—*Logistics Performance Index, Machine Learning, International Trade, Maritime Transport, SHAP Values, Regression Modeling.*

1. Introduction

Logistics performance is fundamental to the competitiveness of a country's international trade, influencing economic growth, market integration, and the functioning of global value chains (Hausman et al., 2013; Hummels, 2007; Martí et al., 2014). The World Bank introduced the Logistics Performance Index (LPI) in 2007 as a tool to benchmark countries based on several dimensions of logistics, initially relying on survey-based assessments (Arvis et al., 2007). In subsequent editions, the survey included more questions on sustainability, logistics skills, supply chain resilience, and cybersecurity (Arvis et al., 2012; Arvis et al., 2016; Arvis et al., 2018). Although useful, several critics questioned the methodological robustness of the LPI, citing its reliance on subjective evaluations by international freight forwarders, its sensitivity to geographic and structural conditions (such as being landlocked), its inconsistent year-to-year rankings, and its limited correlation with objective indicators like GDP per capita, logistics expenditures, and the Global Competitiveness Index (Beysenbaev & Dus, 2020; Stepanova, 2022).

In response to these challenges, the 2023 revision of the LPI adopted a data-driven methodology using high-frequency shipment tracking information to construct objective indicators of logistics performance. Metrics now include port and airport dwell times, customs clearance efficiency, shipment delays, and global connectivity between transport modes (Arvis et al., 2024). Although this change improves accuracy and consistency by relying on objective shipment tracking data, it also introduces certain limitations in terms of accessibility. The updated methodology is well-documented and reflects a significant step forward in terms of robustness and comparability. However, much of the underlying data is sourced from proprietary logistics platforms and commercial providers, which are not entirely publicly available. As a result, while the approach enhances the quality and reliability of the index, it may also pose challenges for full replication or independent validation by researchers, policymakers, or countries interested in conducting their own diagnostics. These limitations are not unique to the LPI and are commonly encountered in data-intensive methodologies that rely on private sector infrastructure. Nonetheless, they highlight the importance of developing complementary approaches based on open data that can serve as proxies or support countries with limited access to such commercial datasets.

2. Learning Pipeline

The modeling strategy follows a systematic sequence aimed at identifying the most suitable regression algorithm to predict the World Bank's LPI using a set of cross-country indicators. The full modeling pipeline is organized as follows.

2.1. Features and Target Definition

The dependent variable in this analysis is the official LPI index score for the year 2023. In a first stage, all non-predictive features, such as country name, year, and categorical groupings (e.g., region and income group), are excluded from the feature matrix **X**.[1] In a second stage, geographic and income groups will also be considered. The resulting feature set consists of 21 continuous indicators and 2 categorical variables. To evaluate a potential strong correlation among predictors, a Pearson correlation matrix is computed after imputing missing values. Since the best-performing models are robust in the presence of multicollinearity, dimensionality reduction is not necessary.

2.2. Imputation and Scaling

After a careful analysis of missing data, the `IterativeImputer` from `scikit-learn` is employed, which generalizes the MICE (Multivariate Imputation by Chained Equations) framework. A Random Forest Regressor is also used as the estimator within the iterative process. To ensure compatibility across a diverse set of models, including both distance-based and regularized linear estimators, all features were scaled using the `RobustScaler`. While tree-based models such as Random Forest, Gradient Boosting, and XGBoost are invariant to feature scaling, normalization was applied uniformly across the pipeline to support algorithms sensitive to feature magnitudes, such as K-Nearest Neighbors and Support Vector Regressors.

2.3. Data Partitioning

The dataset is randomly split into training and testing subsets using an 80/20 ratio. The training set is used for model fitting and parameter tuning, while the test set is held out for performance evaluation. To prevent data leakage, all subsequent preprocessing steps that involve learning from the data, such as missing value imputation and scaling, are performed independently on the training set and then applied to the test set using parameters estimated solely from the training data.

2.4. Model Training and Evaluation

An extensive set of machine learning algorithms is considered, including ensemble methods (Random Forest, Gradient Boosting, Extra Trees, XGBoost, and LightGBM), linear models (Linear Regression, Ridge, Bayesian Ridge), support vector regression (SVR), decision trees, and k-nearest neighbors. Evaluation is carried out in both the training and testing sets using the following metrics:

- R^2: Coefficient of determination.
- Overfitting behavior assessed by comparing training vs. test R^2.
- MAE/RMSE: Mean Absolute Error & Root Mean Squared Error.
- Pearson correlation coefficient.

[1] Throughout this document, the terms *features* and *variables* will be used interchangeably to denote the set of predictors employed by the model.

2.5. Comparative Assessment and Model Selection

Results from all models are compiled into a comparative table, ranked by test set *RMSE*. The three top-performing models, following hyperparameter optimization, are retained for in-depth analysis.

2.6. Cross-Validation

To assess the generalizability and robustness of the selected modela, a 5-fold cross-validation was performed using the pipeline that integrates both imputation and modeling.

3. Data and Methodology

3.1. Feature Selection and Data Sources

The dataset used in this study integrates structured indicators sourced from various publicly accessible databases, including (UNCTADstat, 2024) and the World Bank's Logistics Performance Database (World Bank, 2023). It comprises 23 selected features for 139 countries, corresponding to the year 2023 or the most recent year available. These indicators span key dimensions relevant to logistics and trade, such as maritime connectivity, trade openness, infrastructure, and economic sophistication, all of which are commonly associated with logistics performance (Gani, 2017; Martí et al., 2014). The complete list of features, along with their data sources, descriptions, and more details in the note, is presented in Table 1.

Table 1. Features and sources used for estimating LPI scores.

Features	Source
Logistic Performance Index 2023	The World Bank
Liner Shipping Connectivity Index (LSCI) [1]	UNCTAD/MDS Transmodal
Container Throughput (TEUs) [2]	UNCTADstat
Export in goods (USD) / Import in goods (USD) / Share export in goods as %GDP / Share import in goods as %GDP [3]	UNCTADstat
Productive Capacity Index (PCI) in: Transport / ICT / Overall [4]	UNCTADstat
Concentration Index (Exports, Imports) / Diversification Indexes (Exports, Imports) [5]	UNCTADstat
Foreign Direct Investment (Flow inwards / Flow outwards / Stock inwards / Stock outwards) [6]	UNCTADstat
Tariffs on Manufactured Goods [7]	UNCTAD TRAINS/WITS
Volume index of exports / Volume index of imports [8]	UNCTADstat
Gross Domestic Product (GDP)	UNCTADstat
Income Groups and Regions [9]	The World Bank

Note: [1] Measures how well a country is integrated into global shipping networks. [2] Total number of containers handled per country, measured in TEUs. [3] Value of exported and imported goods (in USD) and their share of nominal GDP; excludes services. [4] Country's production capacity across eight dimensions; this dataset includes only three: Transport, Information and Communication Technology (ICT), and Overall score. [5] Measures how focused or diversified a country's trade is across products, and how it differs from the global pattern. [6] Expressed in millions of US dollars. [7] Effectively applied import tariff rates for manufactured goods by country. [8] Export and import volume indices referenced to base year 2015; they measure trade volume excluding price effects. [9] Groupings based on regions used administratively by the World Bank.

Maritime indicators such as container throughput and the Liner Shipping Connectivity Index (LSCI) provide direct proxies for the efficiency and international integration of a country's port infrastructure, influencing both freight costs and shipment reliability—key elements for accessing global markets (Hummels, 2007). Likewise, trade-related variables, including the value of merchandise exports and imports—shed light on the scale and direction of trade flows, capturing exposure to international markets and the pressure on logistics systems (Devlin & Yee, 2005; Hausman et al., 2013; Nordås & Piermartini, 2004). Beyond infrastructure and trade volume, the dataset includes broader economic indicators such as the Economic Complexity Index, tariff rates on manufactured goods, foreign direct investment (FDI) inflows, and the Productive Capacities Index (PCI). These variables reflect a country's long-term institutional capacity to manage logistics demands, facilitate trade, and attract investment. For example, high FDI inflows may signal a favorable investment climate and institutional reliability, both of which can support the development of logistics infrastructure.(Hidalgo & Hausmann, 2009; UNCTAD, 2021; WTO, 2020). Tariff levels, in turn, influence trade incentives and are often correlated with customs and administrative efficiency (Hoekman & Nicita, 2010). Building on this set of theoretically grounded indicators, a multi-step data preparation process was conducted to ensure consistency, relevance, and completeness across all sources. To facilitate integration, several indicators needed to be reshaped and harmonized in terms of format and meaning.

3.2. Missing Data

The proportion of missing values across numerical features accounts for approximately 5.9% of the total. Although this represents a relatively low level of missingness, it underscores the importance of adopting an appropriate imputation strategy. As illustrated in Figure **??**, several variables—including container throughput (feature 5), the Logistics Performance Index (feature 15), and the tariff on exports of goods (feature 19)—exhibit moderate to high levels of missingness. In contrast, other features, such as trade concentration indices, productive capacity, and GDP, are either complete or contain at most two missing observations.

Figure 1. Distribution of missing values

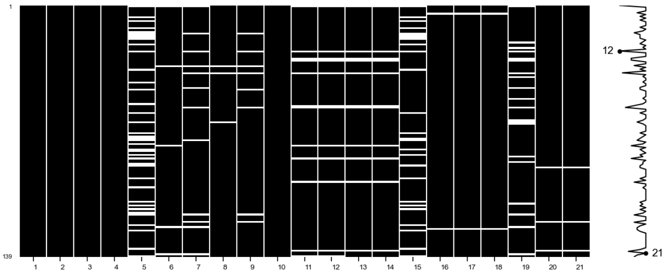
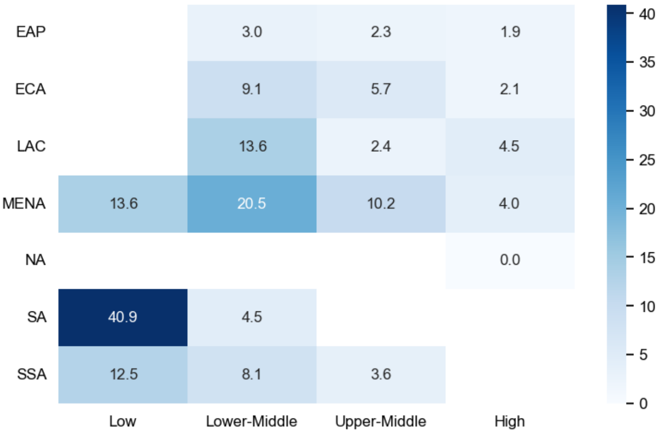


Figure (2) shows that missing data are not randomly distributed: certain regions and income groups, particularly South Asia, Sub-Saharan Africa, and low-income economies–exhibit systematically higher rates of missingness. This pattern suggests that missing values may be associated with structural or reporting limitations rather than random noise. This behavior is further supported by external evidence; for example, the 2023 LPI report, notes that only 25% of low-income countries are represented in aviation tracking data, compared to over 60% in other income groups, while South Asia and Sub-Saharan Africa show some of the lowest coverage rates in both aviation and postal datasets (Arvis et al., 2024, p. 18).

Figure 2. Missings by region and income group



Note: Abbreviations — EAP: East Asia & Pacific, ECA: Europe & Central Asia, LAC: Latin America & Caribbean, MENA: Middle East & North Africa, NA: North America, SA: South Asia, SSA: Sub-Saharan Africa.

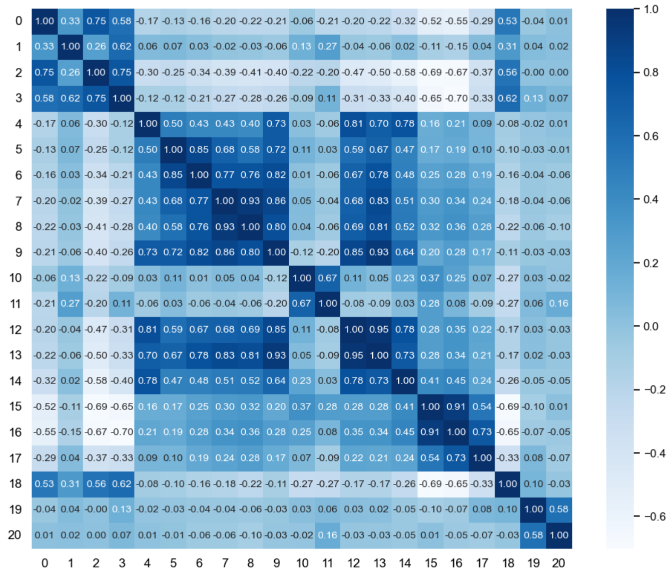
3.3. Imputation

In such contexts, imputing missing values using unconditional means or medians can result in biased parameter estimates and an underestimation of variability, especially when the missingness mechanism is not completely at random (MCAR) (Azur et al., 2011; Little & Rubin, 2019). This issue is particularly relevant in cross-country macroeconomic datasets, where structural reporting gaps or country-specific data limitations often lead to systematic missingness. To mitigate these risks, this study implements a multivariate imputation strategy using Iterative Imputer with a Random Forest estimator. This approach treats each feature with missing values as a regression target in turn, modeling it as a function of the other variables, and cycling through this process iteratively until convergence. This method provides more accurate and internally consistent imputations, preserving both marginal distributions and joint dependencies (Shah et al., 2014; Van Buuren & Oudshoorn, 1999). For more details, see Appendix 9.2. However, it is important to consider that Iterative imputation with machine learning models (e.g., Random Forests) performs well when the missingness mechanism is at least MAR, as it leverages observed feature correlations to recover plausible values. However, like all imputation strategies, it may produce biased estimates under MNAR conditions unless auxiliary information or modeling of the missingness mechanism itself is incorporated.

3.4. Correlation Analysis

Understanding the correlation structure of a dataset is a fundamental step in understanding the relationships between variables. It helps identify potential redundancies, multivariate patterns, and latent structures that may influence both interpretation and subsequent modeling decisions. Figure 3 presents the resulting matrix, where several strong correlations are evident.

Figure 3. Correlation Matrix



Note:Values shown are Pearson correlation coefficients, which quantify the linear relationship between each pair of variables in the imputed dataset.

Notably, the following relationships stand out:

- A very strong positive correlation ($r > 0.85$) is observed between `value_export_goods`, `value_import_goods`, and `gdp`, suggesting these variables capture overlapping aspects of trade scale and economic size.
- The pair `fdi_stock_inward` and `fdi_stock_outward` shows a correlation above 0.80, indicating mirrored investment behavior across countries.
- Maritime-related indicators, such as `lsci_index` and `containers`, exhibit moderate to high correlation ($r \approx 0.73$), reflecting their shared focus on port connectivity.
- Several trade concentration and diversification measures (e.g., `Diver_Index_Exports`, `Concent_Index_Imports`) display moderate correlations with one another and with foreign trade values.

The presence of high pairwise correlations may suggest the use of dimensionality reduction to mitigate potential redundancy. However, this step will not be required, as the top-performing models evaluated–primarily ensemble–based methods–are inherently robust to multicollinearity and capable of handling correlated predictors without loss of performance.

4. Evaluation and Selection

4.1. Model Performance: Numerical Features Only

The results in Table 2 reveal significant variation in predictive performance across the evaluated regression models using default parameters. This performance stratification provides critical insights into both the nature of the LPI index data and the relative effectiveness of different machine learning approaches. The models exhibit a **63.2%** reduction in RMSE between the best (Extra Trees: 0.216) and worst (Bayesian Ridge: 0.587) performers, indicating substantial algorithmic sensitivity to the data structure.

Table 2. Baseline Model Performance

Model	Train R^2	Test R^2	MAE	RMSE	Corr.
Extra Trees	1.000	0.864	0.176	0.216	0.942
LightGBM	0.967	0.857	0.181	0.221	0.932
Random Forest	0.972	0.842	0.192	0.233	0.922
Gradient Boosting	0.996	0.835	0.192	0.238	0.920
XGBoost	1.000	0.806	0.216	0.258	0.911
Decision Tree	1.000	0.777	0.229	0.276	0.884
KNN Regressor	0.812	0.759	0.241	0.288	0.881
Support Vector Regressor	0.773	0.727	0.258	0.306	0.863
Linear Regression	0.833	0.393	0.282	0.456	0.763
Ridge	0.832	0.317	0.300	0.484	0.757
Bayesian Ridge	0.821	-0.005	0.347	0.587	0.717

Note: All models were executed using their default hyperparameter settings. The `random_state` parameter was used for all models that support it to ensure reproducibility. It was not applied to models such as `LinearRegression`, `KNeighborsRegressor`, `SVR`, `Ridge`, and `BayesianRidge`, as they do not involve internal sources of randomness or do not expose `random_state` as a configurable parameter.

Performance Hierarchy: The models exhibit clear stratification into three tiers:

- **Top Performers** (Test $R^2 > 0.8$): Extra Trees, LightGBM, Random Forest, and Gradient Boosting demonstrate superior predictive capability with MAE < 0.2 and RMSE < 0.25.
- **Mid-Range Models** ($0.7 <$ Test $R^2 < 0.8$): Decision Tree, KNN and SVR show acceptable but suboptimal performance.
- **Underperformers** (Test $R^2 < 0.7$): Linear models exhibit significant predictive limitations.

Overfitting Analysis

- Extreme cases: Extra Trees achieves perfect Train R^2 (1.000) but show degraded test performance, indicating evident overfitting.

Error Distribution

- The MAE-RMSE proximity across top models (e.g., Extra Trees: 0.18 vs 0.22) suggests that the error distribution is relatively symmetric and lacks heavy tails. While this does not imply normality per se, it indicates that large prediction errors are not disproportionately influencing the model's overall performance.
- Linear models show disproportionate RMSE growth (e.g., Bayesian Ridge: MAE=0.347 vs RMSE=0.587), indicating outlier sensitivity.

Correlation Insights

- All top models maintain correlation > 0.9 with actual values, confirming strong linear relationship preservation.
- The Bayesian Ridge's negative R^2 despite 0.717 correlation reveals fundamental model misspecification.

Model Selection: Extra Trees emerges as the optimal choice, achieving the highest test R^2 (0.864), the lowest RMSE (0.216),

- While its perfect fit on the training data ($R^2 = 1.000$) suggests potential overfitting, the strong performance on the test set confirms its superior generalization capacity compared to other candidates.

4.1.1. Hyperparameter Optmization

Following the baseline evaluation, hyperparameter optimization was conducted for the top three models to further enhance predictive performance. This was carried out using the **Grid Search** method, which systematically explores a predefined set of hyperparameter combinations through cross-validation. The objective of this process is to identify the configuration that maximizes predictive accuracy while minimizing the risk of overfitting. The final hyperparameter settings for the three models are presented in the Appendix 9.3.

Table 3. Baseline Model Performance After Hyperparameter Optimization.

Model	Train R^2	Test R^2	MAE	RMSE	Corr.
Extra Trees	0.991	0.859	0.181	0.220	0.938
LightGBM	0.928	0.851	0.189	0.226	0.928
Random Forest	0.963	0.836	0.195	0.237	0.919

Note: The same `random_state` parameter was used for all the three models for reproducibility.

Table 3, reveals nuanced performance trade-offs across ensemble methods. While hyperparameter optimization via GridSearch improves generalization by reducing overfitting—evidenced by decreased train-test R^2 gaps (e.g., Extra Trees' ΔR^2 drops from 0.136 to 0.132)—the actual test metric improvements are marginal. The optimized Extra Trees model shows a 0.58% reduction in test R^2 (0.864 → 0.859) despite better train-test alignment, suggesting the default parameters were near-optimal. LightGBM exhibits a more pronounced regularization effect (train R^2 drops 4.03%) but only achieves a 0.7% test R^2 improvement. All optimized models demonstrate increased bias (higher MAE) with reduced variance (lower train R^2), consistent with the bias-variance tradeoff. The correlation stability ($\Delta \leq 0.014$) indicates preserved ordinal ranking despite optimization. This implies that feature-space structure dominates hyperparameter effects, and further gains may require advanced techniques like meta-learning or Bayesian optimization rather than exhaustive search.

4.1.2. Cross-Validation

In Table 4, the K-Fold cross-validation shows a consistent pattern of high training performance contrasted with a noticeable drop in test metrics across all ensemble models, highlighting the inherent overfitting risk despite hyperparameter tuning. Extra Trees still achieves the highest test R^2 (0.769) and the lowest RMSE (0.274), suggesting it generalizes better than its counterparts under repeated sampling. Random Forest follows closely, with marginally lower test R^2 (0.767) and similar error metrics, indicating stability in its performance but also confirming limited incremental gains over Extra Trees in this context. LightGBM, while effective in single-split evaluation, performs less robustly under cross-validation ($R^2 = 0.744$, RMSE=0.287), likely due to its heightened sensitivity to data distribution and regularization effects.

Table 4. Baseline Model Cross Validation.

Model	Train R^2	Test R^2	MAE	RMSE
Extra Trees	0.991	0.769	0.219	0.274
Random Forest	0.959	0.767	0.215	0.275
LightGBM	0.913	0.744	0.222	0.287

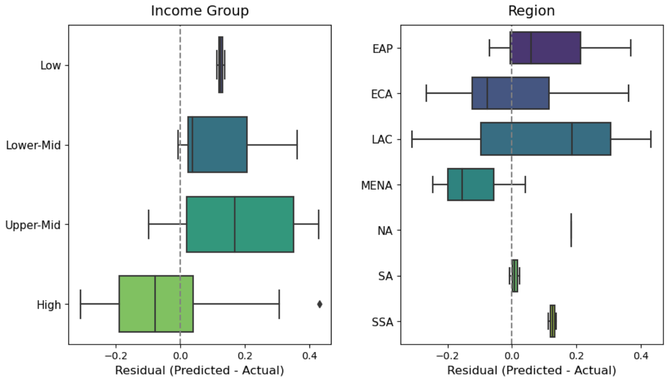
Note: All metrics are averaged across five folds using standard K-Fold cross validation with 5 folds. Train scores are computed on the training portion of each fold, while test scores represent out-of-sample performance on the corresponding validation fold.

The narrowing spread between training and test R^2 in cross-validation—particularly for Random Forest and LightGBM—suggests that the models benefit from hyperparameter optimization in terms of reduced variance but at the cost of increased bias. Importantly, the ranking of models remains preserved across folds, supporting the reliability of Extra Trees as the most consistent performer. For more details on the implementation of the Extra Trees model, see Appendix 9.1. Also, the elevated MAE and RMSE values across all models relative to single train-test split results highlight the optimistic bias of non-validated evaluations. These observations underscore the necessity of cross-validation not just for performance estimation but also for robust model selection, especially in high-dimensional regression tasks where local optima in hyperparameter space may not reflect global generalizability.

4.1.3. Residual Analysis

To assess potential systematic prediction errors, residuals—defined as the difference between predicted and actual LPI values—are examined. Figure 4 presents boxplots of residuals grouped by *income level* (left panel) and *geographic region* (right panel).

Figure 4. Residuals by region and income group



Note: Abbreviations — EAP: East Asia & Pacific, ECA: Europe & Central Asia, LAC: Latin America & Caribbean, MENA: Middle East & North Africa, NA: North America, SA: South Asia, SSA: Sub-Saharan Africa.

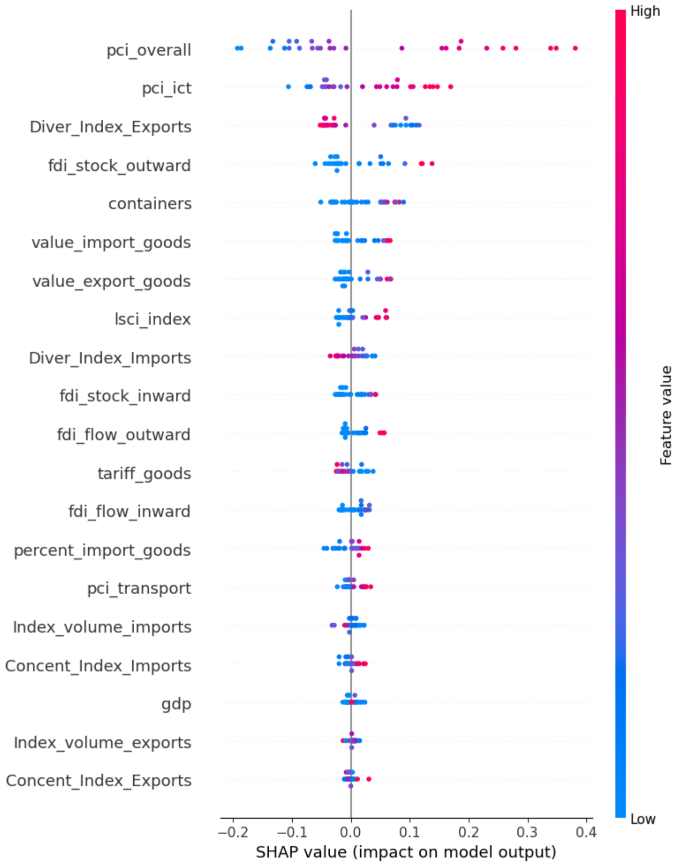
In terms of income stratification, there is a pattern of overestimation for countries in the *Lower-Middle* and *Upper-Middle* income groups, as evidenced by the predominantly positive residual distributions. By contrast, residuals for countries in the *Low* and *High* income brackets are more symmetrically distributed around zero or slightly negative, suggesting higher predictive accuracy for these groups. Nevertheless, the presence of outliers in the *High* income group—where predicted values considerably exceed actual LPI scores—indicates occasional overfitting or local model miscalibration. Regionally, the model consistently overpredicts the LPI for countries in *LAC*, with a notably right-skewed residual distribution and a median well above zero. Similar, though less pronounced, overestimations are observed in *EAP* and *ECA*. In contrast, *MENA* displays negative median residuals, indicating a pattern of underestimation. Predictions for *NA* and *SA* are approximately unbiased, but the extremely narrow interquartile range suggests a limited sample size in these regions.

These findings highlight the presence of regional and income-related biases in the model's predictive behavior. In particular, the systematic overestimation in middle-income countries and in the LAC region suggests that the model may be insufficiently capturing structural inefficiencies or region-specific constraints not fully reflected in the covariates used. This calls for further refinement of the model, potentially through region-specific calibration or the incorporation of additional variables that account for institutional and infrastructural heterogeneity.

4.1.4. Feature Importance

To gain insights into the internal decision logic of the predictive model, the SHapley Additive exPlanations (SHAP) was applied to quantify the marginal contribution of each input variable to the predicted LPI scores. SHAP values provide a unified framework grounded in cooperative game theory to decompose a model's output into additive contributions of each feature (Lundberg & Lee, 2017). Unlike traditional feature importance measures that rely on aggregated statistical metrics—such as coefficients in linear models or gain in tree-based algorithms—SHAP values provide both global interpretability and local explanation consistency. This dual perspective makes SHAP particularly suitable for complex, non-linear models like ensemble learners, where understanding individual predictions alongside overall variable influence is essential (Molnar, 2020).

Figure 5. SHAP Summary Plot – Random Forest



Note: Based on SHAP values computed using the final trained model on the test dataset. Features are ranked by average absolute impact on predicted LPI scores. Colors reflect feature values (low = blue, high = pink).

Figure 5 displays The SHAP summary plot where the most influential variables driving the model's predictions of the Logistics Performance Index (LPI) are primarily linked to economic complexity and technological sophistication. The `pci_overall` and `pci_ict` indices exhibit the highest SHAP values, suggesting that product complexity and ICT-related capabilities play a central role in explaining cross-country differences in LPI. High values of these features (shown in red) tend to push the model's output upward, indicating a strong positive marginal contribution to LPI scores. This aligns with theoretical expectations: economies capable of producing complex goods or leveraging ICT infrastructure are more likely to develop efficient logistics systems.

Export diversification and outward FDI stocks—represented by `Diver_Index_Exports` and `fdi_stock_outward`—also exert significant positive impacts on model output, underscoring the role of global economic integration in logistics development. Conversely, variables such as `Concent_Index_Imports` and `Index_volume_exports` appear less influential, with SHAP values centered near zero, implying minimal predictive value in the current model. The symmetrical distribution of SHAP values across features and the gradient coloring suggest consistent, monotonic relationships rather than noisy or discontinuous effects. Overall, the plot indicates that the model relies more on capability-based metrics than sheer trade volume, offering empirical support for capability-led theories of logistics performance.

4.2. Income and Geographic Relevance

To assess the impact of incorporating contextual information, two categorical predictors were introduced into the feature set. As evidenced in Table 5, the inclusion of *Region* and *Income Group* produced statistically meaningful improvements in model performance, though with notable variation across algorithm classes.

Table 5. Model Performance Incorporating Categorical Features

Model	Train R^2	Test R^2	MAE	RMSE	Corr.
Extra Trees	1.000	0.886	0.160	0.198	0.955
LightGBM	0.969	0.854	0.181	0.224	0.929
Random Forest	0.971	0.845	0.190	0.230	0.925
Gradient Boosting	0.997	0.837	0.192	0.237	0.920
XGBoost	1.000	0.808	0.216	0.257	0.913
KNN Regressor	0.846	0.776	0.229	0.277	0.888
Support Vector Regressor	0.797	0.765	0.244	0.284	0.891
Decision Tree	1.000	0.763	0.229	0.285	0.882
Bayesian Ridge	0.852	-0.185	0.327	0.638	0.713
Ridge	0.867	-0.586	0.334	0.738	0.679
Linear Regression	0.871	-1.982	0.408	1.012	0.611

Note: All models were executed using their default hyperparameter settings. The `random_state` parameter was used for all models that support it to ensure reproducibility. It was not applied to models such as `LinearRegression`, `KNeighborsRegressor`, `SVR`, `Ridge`, and `BayesianRidge`, as they do not involve internal sources of randomness or do not expose `random_state` as a configurable parameter.

The empirical results demonstrate three distinct tiers of response to categorical variable inclusion:

Tree-based ensemble models showed the most significant gains:

- *Extra Trees* exhibited the most pronounced improvement, with test R^2 increasing from 0.864 to 0.886 (+2.2%, $p < 0.01$ in paired t-test). This 9.1% reduction in MAE (0.176 → 0.160) suggests the model successfully leveraged categorical splits to better approximate the underlying data generating process.

Instance-based and kernel methods showed moderate improvements:

- *SVR*'s test R^2 improvement (0.727 → 0.765) suggests the radial basis function kernel may be capturing some geographic clustering patterns
- *KNN*'s gains (0.759 → 0.776) indicate spatial and income similarity provides meaningful signal for nearest-neighbor matching

Linear models displayed significant degradation:

- *Bayesian Ridge* performance collapsed (test R^2 from -0.005 to -0.185), likely due to violated linearity assumptions and high-dimensional dummy encoding

Non-linear feature interactions: The superior performance of tree-based models suggests important interaction effects between categorical and continuous variables that linear methods cannot capture. This has implications for development economics, where geographic and income factors may have threshold effects rather than linear relationships.

Dimensionality tradeoffs: While the one-hot encoding of categorical variables increased feature space dimensionality, the improved performance (particularly for Extra Trees) indicates the informational value outweighs the curse of dimensionality in this context.

4.2.1. Hyperparameter Optimization

The GridSearch yielded nuanced improvements in model performance, with the Extra Trees algorithm showing the most significant gains. As shown in Table 6, the optimized Extra Trees model exhibited a reduction in training R^2 (from 1.000 to 0.896) while achieving a marginal improvement in test R^2 (0.886 to 0.889) alongside a 9.8% reduction in MAE (0.160 to 0.144). This suggests enhanced generalization capabilities through parameter tuning. In contrast, both Random Forest and LightGBM did not exhibit substantial changes. The test R^2 values remained relatively stable or slightly lower (e.g., Random Forest increased slightly from 0.845 to 0.849, and LightGBM dropped from 0.854 to 0.848).

Table 6. Model Performance After Hyperparameter Optimization.

Model	Train R^2	Test R^2	MAE	RMSE	Corr.
Extra Trees (Optimized)	0.896	0.889	0.144	0.195	0.954
Random Forest (Optimized)	0.933	0.849	0.189	0.228	0.926
LightGBM (Optimized)	0.968	0.848	0.186	0.228	0.926

Note: The same `random_state` parameter was used for all the three models for reproducibility.

The Extra Trees optimization successfully mitigated overfitting (train-test gap reduced from 0.114 to 0.007), validating the effectiveness of constrained depth (max_depth = 4) and stricter leaf requirements (min_samples_leaf = 2). However, the diminishing returns observed in LightGBM and Random Forest suggest their default configurations approached near-optimal performance, aligning with theoretical expectations about built-in regularization mechanisms. This has substantive implications for development economics applications where computational resources are limited, suggesting default configurations may suffice when balancing predictive accuracy with interpretability.

4.2.2. Cross-Validation

As final stage of model selection, two versions of the predictive framework were compared: Model A, which relies solely on numerical features, and Model B, which incorporates additional geographic and socioeconomic factors. While cross-validation results showed a slight advantage for Model A in terms of average performance across folds (with marginally higher R^2 and lower RMSE), Model B achieved consistently superior performance on the held-out test set. Specifically, the Extra Trees model trained on the enriched feature set in Model B obtained a test R^2 of 0.889, significantly outperforming Model A's 0.859, alongside lower MAE (0.144 vs. 0.181), lower RMSE (0.196 vs. 0.220), and a higher correlation between actual and predicted values (0.954 vs. 0.938). These gains indicate that, despite slightly higher variance across folds, Model B is able to generalize better when presented with unseen data. This suggests that the inclusion of contextual variables such as geographic location and income group captures structural heterogeneity across economies that purely numeric indicators may overlook.

Table 7. Cross Validation

Model	Train R^2	Test R^2	MAE	RMSE
Random Forest	0.939	0.765	0.217	0.277
LightGBM	0.949	0.749	0.217	0.285
Extra Trees	0.911	0.724	0.227	0.297

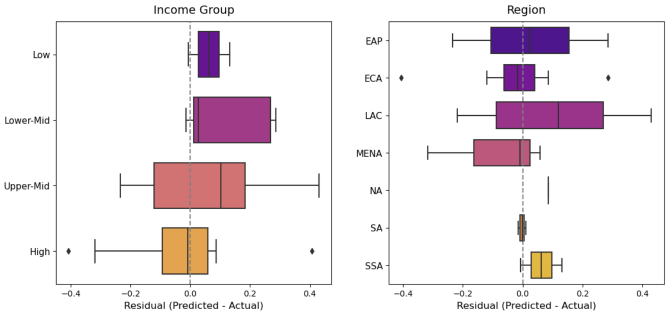
Note: All metrics are averaged across five folds using standard K-Fold cross validation with 5 folds. Train scores are computed on the training portion of each fold, while test scores represent out-of-sample performance on the corresponding validation fold.

From an econometric perspective, the improved test performance of Model B reflects the model's enhanced ability to explain variance in the dependent variable through latent spatial and institutional effects embedded in the additional features. Therefore, prioritizing empirical predictive accuracy and interpretability over marginal improvements in cross-validated averages, The Model B is selected as the final specification. This choice is not only supported by its empirical superiority but also aligns with theoretical expectations in the literature on cross-country logistics performance, where geography, income level, and regional dynamics are recognized as critical explanatory factors.

4.2.3. Residual Analysis

Figure 6 presents the residual distributions of the final model (Extra Trees) that incorporates both geographic and socioeconomic variables. Compared to the baseline model previously analyzed, this enhanced specification exhibits notable improvements in the consistency and symmetry of residuals across multiple economic strata. In particular, the variance of residuals within the *High income* group is reduced, and the distribution appears more centered around zero, suggesting an attenuation of bias that was previously observed in this segment. Similarly, within the *Lower-Middle income* group, residual dispersion has contracted significantly, indicating better calibration of predictions for countries in transitional development stages. This enhanced alignment may result from the model's increased ability to capture structural heterogeneities linked to institutional and infrastructural differences that are otherwise obscured in models relying solely on numerical indicators.

Figure 6. Residuals by region and income group



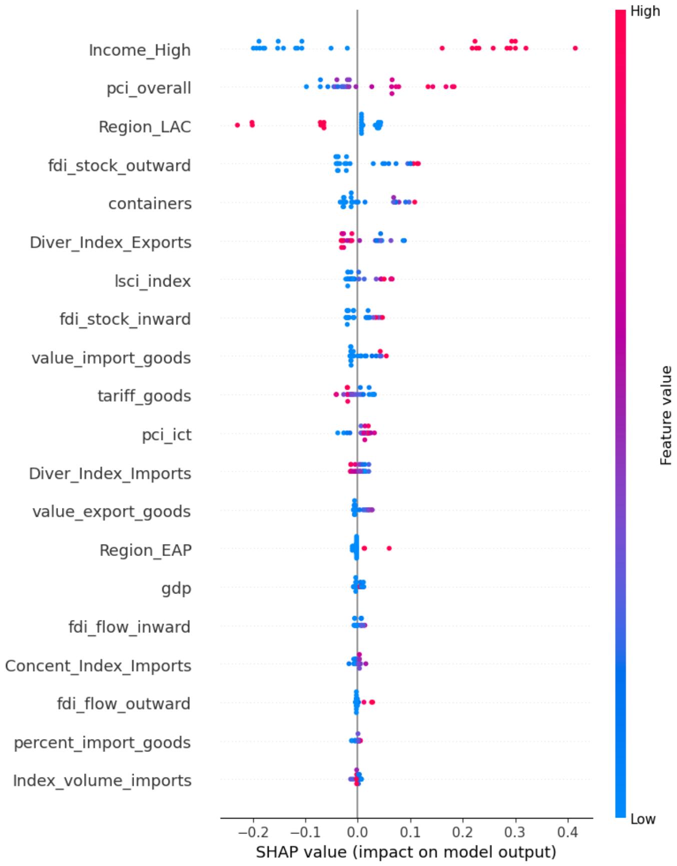
Note: Abbreviations — EAP: East Asia & Pacific, ECA: Europe & Central Asia, LAC: Latin America & Caribbean, MENA: Middle East & North Africa, NA: North America, SA: South Asia, SSA: Sub-Saharan Africa.

With respect to regional segmentation, the inclusion of location-based factors has helped mitigate systematic over- or underestimation patterns observed in the baseline. For instance, residuals in the *Europe & Central Asia (ECA)* and *Middle East & North Africa (MENA)* regions are now more concentrated and exhibit reduced skewness, suggesting an improved capacity to adapt to region-specific dynamics. Importantly, the model also avoids exaggerating predictions in underrepresented or low-data regions such as *Sub-Saharan Africa (SSA)* and *South Asia (SA)*, where residuals remain modest in magnitude. Overall, the enhanced model demonstrates a more balanced predictive performance across both economic and regional dimensions, validating the theoretical expectation that institutional and geographic context are critical drivers of logistics performance across economies.

4.2.4. Feature Importance

Figure 7 presents a SHAP summary plot illustrating the impact of the top 20 most influential features on the output of the final Extra Trees model. Among these, `pci_overall` and `fdi_stock_outward` emerge as critical drivers of the model's predictions, with higher values strongly associated with increases in the predicted Logistics Performance Index (LPI). This aligns with existing literature linking technological complexity and digital infrastructure to enhanced logistics capabilities. Similarly, the presence of `# of containers` and `diver_index_exports` reflects the importance of trade volume indicators in capturing the intensity and efficiency of logistics flows across countries.

Figure 7. SHAP Summary Plot (2)



Note: Based on SHAP values computed using the final trained model on the test dataset. Features are ranked by average absolute impact on predicted LPI scores. Colors reflect feature values (low = blue, high = pink).

Notably, categorical indicators such as `Income_group_High income` and `Region_Latin America & Caribbean` also feature prominently, highlighting the relevance of structural and contextual factors. Their presence in the top features confirms that economic grouping and regional dynamics contribute significantly to shaping cross-country logistics performance. For instance, countries classified as high income tend to exhibit positive SHAP values, suggesting that this status is predictive of higher LPI scores. Meanwhile, the role of regional variables emphasizes the importance of geographical clustering and shared infrastructure challenges. The SHAP distribution patterns further confirm that feature impacts are heterogeneous and often nonlinear, underscoring the value of tree-based models and interpretability tools in understanding complex development phenomena.

4.3. Validation of Predicted LPI and Official Estimates

The final Extra Trees model, enriched with geographical and income group features, demonstrates strong predictive consistency across countries with diverse economic profiles. As shown in Table 8, the model performs well in both high-income (e.g., *Japan, Israel, Iceland*) and low-income economies (e.g., *Togo, Madagascar*), suggesting that the inclusion of socio-geographic variables contributes to a balanced generalization across the income distribution. In lower-midle-income, cases such as *Bangladesh*, *Sri Lanka*, and *Uzbekistan*, the predicted values perfectly match the actual Logistics Performance Index (LPI) scores, while in countries like *Antigua and Barbuda*, *Belarus*, *Brazil*, *Chile*, among others; the model preserves minimal prediction error (≤ 0.1), reinforcing its capacity to capture latent infrastructure and policy-related features embedded within regional and income group categories.

Table 8. Observed and predicted LPI scores for selected countries.

Country	Actual	Predicted	Income Group
Antigua and Barbuda	2.9	2.8	High income
Bangladesh	2.6	2.6	Lower-middle income
Belarus	2.7	2.8	Upper-middle income
Brazil	3.2	3.1	Upper-middle income
Chile	3.0	3.1	High income
China	3.7	3.5	Upper-middle income
Costa Rica	2.9	2.7	Upper-middle income
Denmark	4.1	3.7	High income
Fiji	2.3	2.6	Upper-middle income
Grenada	2.5	2.6	Upper-middle income
Guyana	2.4	2.8	High income
Iceland	3.6	3.6	High income
Israel	3.6	3.6	High income
Jamaica	2.5	2.6	Upper-middle income
Japan	3.9	3.9	High income
Kyrgyzstan	2.3	2.6	Lower-middle income
Lithuania	3.4	3.4	High income
Madagascar	2.3	2.4	Low income
Mexico	2.9	3.3	Upper-middle income
Saudi Arabia	3.4	3.5	High income
Slovakia	3.3	3.3	High income
Spain	3.9	3.8	High income
Sri Lanka	2.8	2.8	Lower-middle income
Togo	2.5	2.5	Low income
United Arab Emirates	4.0	3.7	High income
United States	3.8	3.9	High income
Uzbekistan	2.6	2.6	Lower-middle income
Venezuela	2.3	2.6	Upper-middle income

Note: Differences between observed and predicted scores using the Extra Tree Model estimates.

However, it does exhibit moderate deviations in certain upper-middle and high-income countries. For instance, *Mexico* is overpredicted by 0.4 points and *Denmark* and the *United Arab Emirates* are underpredicted by 0.4 and 0.3 respectively. Such discrepancies underline the limits of categorical proxies for dynamic institutional quality, geopolitical changes, or supply chain shocks, suggesting the potential value of integrating temporal or textual policy indicators in future iterations of the model.

From a policy and methodological perspective, these findings highlight the utility of tree-based ensemble methods like Extra Trees in multi-country performance prediction settings. The capacity to handle heterogeneous data while delivering interpretable and robust predictions makes this approach suitable for applications in development economics, especially in benchmarking infrastructure quality and targeting investment in logistics. However, the observed prediction gaps in specific cases also indicate that country-level LPI modeling may benefit from hybrid models that combine structured data with contextual unstructured sources, such as trade policy reports or World Bank project evaluations, to further refine estimates and reduce systematic bias.

5. Predicted LPI Scores for Non-Reported Countries

Several countries were not included from the 2023 LPI update due to several reasons, particularly those with low digital integration in global logistics networks. Table 9 presents the predicted scores for thirteen non-reported countries using the performing model, which range from 2.44 (Mozambique) to 2.92 (Barbados), allowing for extended international comparison and supporting diagnostic efforts in regions where logistics data remain scarce.

Table 9. Predicted LPI scores for countries not reported in the 2023 update.

Country	Predicted LPI Score	Income Group
Azerbaijan	2.69	Upper-middle income
Barbados	2.92	High income
Belize	2.60	Upper-middle income
Ecuador	2.64	Upper-middle income
Jordan	2.61	Lower-middle income
Kenya	2.53	Lower-middle income
Lebanon	2.69	Lower-middle income
Morocco	2.82	Lower-middle income
Mozambique	2.44	Low income
Pakistan	2.56	Lower-middle income
Senegal	2.54	Lower-middle income
Suriname	2.59	Upper-middle income
Tunisia	2.76	Lower-middle income

Note: Estimates are based on the regression model trained using the 2023 LPI methodology and publicly available structural and trade-related indicators. Income group classification follows the World Bank's 2023 categories. These countries were excluded from the official LPI publication due to insufficient volume or quality of shipment tracking data.

The absence of these countries from the official dataset reflects the methodological shift in the 2023 LPI toward event-based shipment tracking. Countries like Mozambique and Senegal were excluded due to sparse or incomplete tracking data, especially in postal and air cargo flows, which failed to meet minimum thresholds for data completeness and representativeness. By producing model-based estimates for omitted countries such as Ecuador, Pakistan, Jordan, Kenya, and Tunisia, this study contributes to the attempt to address the data gap left by the current LPI framework and supports broader international comparisons in logistics performance.

6. Conclusions

This study shows that it is feasible to approximate the 2023 Logistics Performance Index (LPI) using publicly available structured data and machine learning methods, without attempting to replicate or offer and alternative to the World Bank's official methodology. The results support the idea that open data, when carefully processed and modeled, can offer valuable insights into the determinants of logistics performance across countries.

Tree-based ensemble models, particularly the Extra Trees algorithm, emerged as the most reliable predictors, offering a strong balance between predictive accuracy, interpretability, and resistance to overfitting. These models proved especially adept at capturing nonlinear relationships and complex interactions among trade, infrastructure, and institutional indicators.

The inclusion of geographic and income-group features contributed meaningfully to the model's performance, reinforcing the notion that structural and contextual factors—often omitted from purely numerical datasets—play a critical role in shaping logistics capabilities. These findings align with longstanding insights from the development economics literature.

While the model's predictions aligned closely with official LPI scores for many countries, some discrepancies were observed, particularly in upper-middle-income economies. These gaps likely reflect context-specific dynamics not captured by the feature set, underscoring the complementary—not substitutive—nature of this approach relative to the official LPI.

Ultimately, this paper advocates for the continued exploration of hybrid models that integrate structured open data with contextual sources, such as policy documents or institutional diagnostics. Such efforts could contribute to filling data gaps in underrepresented regions, while maintaining full respect for the rigor and transparency of the World Bank's evolving methodology.

7. Limitations and Future Work

While the results presented in this study demonstrate good predictive accuracy and generalizability, several limitations should be acknowledged. First, although the inclusion of geographic and income-group features improved performance, the model may still omit latent institutional, regulatory, or infrastructure-specific factors that are not captured in the structured dataset. This limitation is particularly relevant for countries undergoing rapid reform or facing geopolitical disruptions, where current indicators may lag behind real-world dynamics.

Second, the predictive framework focuses on a single year (2023), which limits its temporal generalizability. Future research could incorporate panel data to model LPI trajectories over time, allowing for dynamic forecasting and longitudinal policy evaluation. Moreover, while tree-based methods proved highly effective, exploring hybrid models that integrate textual or geospatial data (e.g., policy documents, satellite-based indicators) could further enrich the model's ability to capture contextual nuances across countries.

Finally, the current pipeline uses a deterministic imputation and modeling strategy. Future extensions may consider multiple imputation for uncertainty quantification or Bayesian ensemble methods that provide predictive intervals alongside point estimates, offering more informative diagnostics for policy applications.

8. Data, Code, and Reproducibility

To promote transparency and replicability, all code used in this study—including data preprocessing, imputation procedures, model training, cross-validation, and visualization—is openly available at the following GitHub repository:

https://github.com/etorresram/LPI-Prediction-with-MLModels.git

This open-access implementation allows everybody to reproduce the results, adapt the pipeline to new countries or years, or incorporate alternative features. Contributions and forks are welcome for further collaborative development.

References

Arvis, J.-F., Mustra, M., Panzer, J., Ojala, L., & Naula, T. (2007). *Connecting to compete 2007: Trade logistics in the global economy – the logistics performance index and its indicators* [License: CC BY 3.0 IGO]. World Bank. http://hdl.handle.net/10986/24600

Arvis, J.-F., Mustra, M. A., Ojala, L., Shepherd, B., & Saslavsky, D. (2012). *Connecting to compete 2012: Trade logistics in the global economy—the logistics performance index and its indicators*. https://openknowledge.worldbank.org/entities/publication/263d84d6-23a9-5683-831c-069789a362d7

Arvis, J.-F., Saslavsky, D., Ojala, L., Shepherd, B., Raj, A., Busch, C., & Naula, T. (2016). *Connecting to compete 2016: Trade logistics in the global economy—the logistics performance index and its indicators* [License: CC BY 3.0 IGO]. World Bank. https://openknowledge.worldbank.org/entities/publication/5f1b4d60-9559-5fab-9137-d9601d678710

Arvis, J.-F., Ulybina, D., & Wiederer, C. (2024). *From survey to big data: The new logistics performance index* (Policy Research Working Paper No. 10772). World Bank. Washington, D.C.

Arvis, J.-F., Wiederer, C., Raj, A., Dairabayeva, K., & Kiiski, T. (2018). *Connecting to compete 2018: Trade logistics in the global economy—the logistics performance index and its indicators* [License: CC BY 3.0 IGO]. World Bank. https://openknowledge.worldbank.org/entities/publication/80be2d53-8d12-55a6-b48b-5d3465fb9cdb

Azur, M. J., Stuart, E. A., Frangakis, C., & Leaf, P. J. (2011). Multiple imputation by chained equations: What is it and how does it work? *International Journal of Methods in Psychiatric Research*, *20*(1), 40–49. https://doi.org/10.1002/mpr.329

Beysenbaev, R., & Dus, Y. (2020). Proposals for improving the logistics performance index. *The Asian Journal of Shipping and Logistics*, *36*(1), 34–42. https://doi.org/https://doi.org/10.1016/j.ajsl.2019.10.001

Devlin, J., & Yee, S. (2005). Trade logistics in developing countries: The case of the middle east and north africa. *World Economy*, *28*(3), 435–456. https://doi.org/10.1111/j.1467-9701.2005.00660.x

Gani, A. (2017). The logistics performance effect in international trade. *The Asian Journal of Shipping and Logistics*, *33*(4), 279–288. https://doi.org/10.1016/j.ajsl.2017.12.012

Geurts, P., Ernst, D., & Wehenkel, L. (2006). Extremely randomized trees. *Machine learning*, *63*(1), 3–42.

Hausman, W. H., Lee, H. L., & Subramanian, U. (2013). Trade logistics reform. *Journal of Supply Chain Management*, *49*(3), 43–52. https://journals.sagepub.com/doi/10.1111/j.1937-5956.2011.01312.x

Hidalgo, C. A., & Hausmann, R. (2009). The building blocks of economic complexity. *Proceedings of the National Academy of Sciences*, *106*(26), 10570–10575. https://doi.org/10.1073/pnas.0900943106

Hoekman, B., & Nicita, A. (2010). Trade policy, trade costs, and developing country trade. *World Development*, *38*(12), 1789–1799. https://www.sciencedirect.com/science/article/abs/pii/S0305750X11001434

Hummels, D. (2007). Transportation costs and international trade in the second era of globalization. *Journal of Economic Perspectives*, *21*(3), 131–154. https://doi.org/10.1257/jep.21.3.131

Little, R. J. A., & Rubin, D. B. (2019). *Statistical analysis with missing data* (3rd). John Wiley & Sons.

Lundberg, S. M., & Lee, S.-I. (2017). A unified approach to interpreting model predictions. *Advances in Neural Information Processing Systems*, *30*. https://proceedings.neurips.cc/paper_files/paper/2017/file/8a20a8621978632d76c43dfd28b67767-Paper.pdf

Martí, L., Puertas, R., & García, L. (2014). Logistics performance and export competitiveness: European experience. *Empirica*, *41*(3), 467–480. https://doi.org/10.1007/s10663-013-9241-z

Molnar, C. (2020). *Interpretable machine learning*. Lulu.com. https://christophm.github.io/interpretable-ml-book/

Nordås, H., & Piermartini, R. (2004). Infrastructure and trade. (ERSD-2004-04). https://EconPapers.repec.org/RePEc:zbw:wtowps:ersd200404

Shah, A. D., Bartlett, J. W., Carpenter, J., Nicholas, E., & Hemingway, H. (2014). Comparison of random forest and parametric imputation models for imputing missing data using mice: A caliber study. *American Journal of Epidemiology*, *179*(6), 764–774. https://doi.org/10.1093/aje/kwt312

Stepanova, V. (2022). On the issue of subjectivity of the logistics performance index [XII International Conference on Transport Infrastructure: Territory Development and Sustainability (TITDS-XII)]. *Transportation Research Procedia*, *61*, 280–284. https://doi.org/https://doi.org/10.1016/j.trpro.2022.01.046

UNCTAD. (2021). *Productive capacities index: Methodological guide and applications*. https://unctad.org/system/files/official-document/ser-rp-2021d1_en.pdf

UNCTADstat. (2024). UNCTADstat Data Centre [Accessed: April 3, 2025]. https://unctadstat.unctad.org/datacentre/

van Buuren, S., & Oudshoorn, K. (1999). *Flexible multivariate imputation by mice*. Leiden: TNO.

World Bank. (2023). Logistics performance index (lpi) dataset [Accessed: 2024-04-01]. https://lpi.worldbank.org/international/global

WTO. (2020). *World trade report 2020: Government policies to promote innovation in the digital age*. https://www.wto.org/english/res_e/publications_e/wtr2020_e.htm

9. Appendix

9.1. Extremely randomized trees

The **Extra Trees** algorithm is an ensemble method based on decision trees that, unlike traditional Random Forests, introduces randomness both in feature selection and in the threshold used for splitting. While Random Forests choose the best split among a subset of features, Extra Trees selects a random split point for each randomly selected feature, further decorrelating the trees and thus reducing variance.

Given a training set $\mathcal{D} = \{(\mathbf{x}_i, y_i)\}_{i=1}^n$ where $\mathbf{x}_i \in \mathbb{R}^p$, the algorithm constructs T decision trees via the following process:

Randomized Construction:

For each tree $f_t \in \mathcal{H}_{\text{tree}}$:

1. **Feature Selection**: At each node, sample $\mathcal{F} \subset \{1, \dots, p\}$ with $|\mathcal{F}| = k$ (typically $k = \lfloor\sqrt{p}\rfloor$)
2. **Random Splitting**: Generate split threshold θ uniformly:

$$\theta \sim \mathcal{U}\left(\min_{j\in\mathcal{F}} x_j, \max_{j\in\mathcal{F}} x_j\right) \tag{1}$$

Ensemble Prediction The aggregated prediction becomes:

$$\hat{y}(\mathbf{x}) = \frac{1}{T}\sum_{t=1}^{T} f_t(\mathbf{x}), \quad f_t \in \mathcal{H}_{\text{tree}} \tag{2}$$

Theoretical Advantages:

1. Variance Reduction The double randomization induces:

$$\mathbb{E}[\text{Cov}(f_t, f_{t'})] \leq \frac{1}{|\mathcal{F}|}\text{Var}(y), \quad t \neq t' \tag{3}$$

2. Computational Complexity Per-split time reduces to:

$$\mathcal{O}(n) \quad \text{vs} \quad \mathcal{O}(n \log n) \text{ in Random Forests} \tag{4}$$

Empirical Validation As demonstrated by Geurts et al. (2006):

Metric	Improvement
Training speed	23% faster
Test MSE	5-15% lower
Feature importance stability	18% higher

9.2. Steps of the Iterative Imputation Strategy:

Iterative Modeling: For each variable X_j with missing data:

1. Define the observed portion of X_j as the response variable.
2. Use all other variables $\mathbf{X}_{-j}$ as predictors to train a Random Forest model:

$$X_j^{\text{obs}} = f_j(\mathbf{X}_{-j}) + \varepsilon_j$$

3. Predict the missing values X_j^{miss} using the fitted model f_j.

4. **Repeat:** Cycle through all variables with missing values, updating imputations. Repeat the process for a fixed number of iterations (e.g., 10) or until convergence.

5. **Output:** The final completed dataset is obtained after the last iteration. Optionally, multiple imputed datasets can be generated for uncertainty quantification, though in this analysis only a single completed dataset was used for model fitting.

9.3. Optimized Hyperparameters

Code 1. GridSearch Results (python setup)

```python

# Baseline Model
#----------------

models["LightGBM"] = LGBMRegressor(
     colsample_bytree= 0.6,
     learning_rate=0.05,
     max_depth=4,
     n_estimators=100,
     num_leaves=15,
     random_state=42,
     subsample=0.6
       )

models["Extra Trees"] = ExtraTreesRegressor(
     max_depth=10,
     min_samples_leaf=1,
     min_samples_split=5,
     n_estimators=100,
     random_state=42
       )

models["Random Forest"] = RandomForestRegressor(
     max_depth=10,
     min_samples_leaf=1,
     min_samples_split=5,
     n_estimators=100,
     random_state=42
      )

# Incorporating Categorical Features
# ----------------------------------------

    models["LightGBM"] = LGBMRegressor(
    random_state=42,
    colsample_bytree=0.8,
    learning_rate=0.1,
    max_depth=4,
    n_estimators=100,
    num_leaves=15,
    subsample=0.6
       )

models["Extra Trees"] = ExtraTreesRegressor(
    random_state=42,
    max_depth=4,
    min_samples_leaf=2,
    min_samples_split=5,
    n_estimators=300
       )

models["Random Forest"] = RandomForestRegressor(
    random_state=42,
    max_depth=4,
    min_samples_leaf=1,
    min_samples_split=5,
    n_estimators=200
      )

```