

Extending and Understanding:
An Application of Machine Learning to the World Bank's Logistics
Performance Index
Ben Shepherd and Tanaporn Sriklay

March 9th, 2023.

Abstract

Purpose

We extend the World Bank's Logistics Performance Index (LPI) for 30 additional countries and 13 additional years. We develop an inexpensive method for extending survey data when frequent, universal surveys are unavailable. We identify groups of country characteristics that influence LPI scores.

Design/Methodology/Approach

Using data from the World Development Indicators—the broadest global dataset of country socio-economic features—we test machine learning algorithms for their ability to predict the LPI. We examine importance scores to identify factors that influence LPI scores.

Findings

The best performing algorithm produces predictions on unseen data that account for nearly 90% of observed variation, and are accurate to within 6%. It performs twice as well as an OLS model with per capita income as the only predictor. Explanatory factors are: business environment; economic structure; finance; environment; human development; and institutional quality.

Practical Implications

Machine learning offers a simple, inexpensive way of extending the coverage of survey data. Our dataset provides a richer picture of logistics performance around the world. The factors we identify as predicting higher LPI scores can help policymakers and practitioners target interventions.

Originality/Value

This paper is one of the first applications of machine learning to extend coverage of an index based on an international survey. We use the new data to provide the most wide-ranging analysis of logistics performance across countries and over time. The output is an important resource for policymakers tracking performance, and researchers particularly in smaller and lower income countries. We also examine a wider range of explanatory factors for LPI scores than previous work.

1 INTRODUCTION

Logistics performance is an important determinant of firm-level outcomes, and the ability to engage with international markets, including through Global Value Chains (Memedovic et al., 2008; Ojala et al., 2008). However, obtaining estimates of logistics costs is data intensive, and requires sophisticated methodologies (e.g., Rantasila and Ojala, 2015; Engblom et al., 2012; Solakivi et al., 2018).

Against this background, the World Bank's Logistics Performance Index (LPI) was launched in 2007, with the objective of providing a multi-dimensional summary of performance in the logistics sector (Arvis et al., 2007, 2010, 2012, 2014, 2016, 2018). The LPI rapidly become an indicator used by policymakers and analysts alike. In 2018, it covered 160 countries. The LPI dataset is based on a survey administered online to around 1,000 respondents, who grade performance on a Likert scale in six core areas, which are aggregated into an overall index using Principal Component Analysis.¹ The six dimensions are: competence and quality of logistics service providers; ease of arranging competitively priced shipments; timeliness of delivery; ability to track and trace shipments; quality of trade-related infrastructure; and performance of border agencies. These components provide a view of logistics performance that is of particular importance from the point of view of international trade integration (Arvis et al., 2007). The objective is to capture the key dimensions of country performance that determine the ability of shippers and consignees to arrange timely, reliable, and competitively priced movements of goods across borders.

While the LPI represents an important step forward in the ability to understand and track performance in logistics, it suffers from two limitations. First, it has not been produced since 2018, and before then was produced on a two or three year schedule. There are gaps in the data, which means that it is difficult to accurately understand performance trends. Second, country coverage varies across editions. This selection effect introduces an additional difficulty in terms of tracking trends through time: some countries experience long gaps between observations if they are not included in every edition. Similarly, while the survey approach ensures that major traders are covered in each edition, the same is not true of smaller, lower income countries. While logistics is becoming a more important trade policy issue around the world, the information base remains incomplete in significant ways.

The incompleteness of the existing LPI information base is non-random. The 2018 edition of the index covers 64% of high income countries, 77% of upper middle income countries, 85% of lower middle income countries, and 89% of low income countries. While these figures show that the World Bank has been diligent in including developing countries in the LPI, the data behind reports such as Arvis et al. (2007, 2010, 2012, 2014, 2016, 2018) show that survey sample sizes for smaller and lower income countries are often small, and coverage of countries varies substantially from one survey round to another. The exclusion of smaller countries due to the difficulty of collecting data there is an issue for economic analysis of the impacts of improvements in logistics performance, as it necessarily skews results towards what is seen in larger countries.

¹ The 2007 edition of the LPI also included a seventh sub-index. However, it was not included in the calculation of the overall LPI, and was dropped from subsequent editions.

Coverage is one issue that arises for users of the LPI, both in a temporal and cross-sectional (country) sense. A related concern is resource intensity. Running a survey, even online, entails major direct costs. As such, it is largely beyond the reach of researchers outside international organizations like the World Bank. Indeed, the lack of an LPI edition since 2018 suggests that even within well-resourced international organizations, it may be difficult to mobilize the financial and technical support required to put in place such a large project on a regular basis.

Against this background, the present paper asks a simple question: is it possible to use existing information from a standard global dataset to fill in the blanks in the existing LPI dataset, and thereby potentially provide a basis for extending coverage in the country and time series dimensions? The answer is a qualified “yes”. It is possible to extract meaningful, though potentially complex, relationships with variables that are observed more directly and regularly in standard international statistics, without the need for a survey. Concretely, we apply a suite of machine learning (ML) techniques with the aim of using the World Development Indicators dataset to predict the LPI. We start with that dataset because it is the broadest-based international dataset covering economic, social, governance, geography, infrastructure, and other country-level variables that could plausibly be related to logistics performance. Our objective is to use ML to choose a subset of data series and a prediction algorithm that combine to produce a synthetic measure that closely mirrors the observed LPI within-sample, but also enables us to extend the series to new countries and years with no observed LPI score. We will make the full, filled-in dataset available to other researchers upon request in standard (CSV) format, along with input data and replication code.

A secondary output of our approach is that it supplies data-driven insights regarding the country-level factors that influence LPI scores. Hillberry and Zhang (2018) also use machine learning to analyze the LPI, but with a much narrower range of explanatory factors than we use. They too start from the World Development Indicators as their primary data source, but they apply analyst judgment to choose a subset of possible explanatory factors. Our approach allows a greater role for the data to speak independently of analyst judgment, and we therefore start with a broader information base in areas like economic performance, trade openness, infrastructure quality, and social characteristics of countries. We therefore contribute to a broader understanding of the factors that influence logistics performance by grouping important predictors into six groups that correspond to a mix of economic and other factors. Our primary objective is to extend the LPI dataset, but a secondary objective is to better understand the potentially wide range of factors that affect a country’s ability to provide logistics services in a competitive way.

While ML techniques are increasingly familiar in the context of “big data” applications, they have recently found a range of applications in economics and related fields (e.g., Athey and Imbens, 2019). There are other studies which apply ML techniques to fill data gaps in various fields. By using stimulated data related to securities, Cagala (2017) shows the random forest model has higher performance in the dataset with complex-nonlinear relationships compared to the Logit model. Moreover, Rees and Haig (2018) highlight an ability of ML to extend coverage of a stock risk assessment where analyst assessment is unavailable. The random forest is also often applied to close a gap of continuous time-series in environment datasets (Kang et al., 2019; Zhu et al., 2022). Other ML techniques such as Support Vector Machines are effectively applied to address energy consumption data gaps (Zhang et al., 2018). An important feature of ML techniques is that they are theoretically agnostic: they do not impose a given framework a priori,

but instead allow the data to speak to the existing correlations and predictive power of different elements. While unconventional, our approach makes it possible, in principle, both to predict LPI values where they are not currently observed using only commonly available country-level data, and to understand a broader range of factors that may lie behind performance as it varies across country income groups and geographical regions.

Given the relatively small amount of data available to “train” ML models, our approach is remarkably successful. Our preferred model produces predicted LPI values that account for nearly 90% of observed variation in the original LPI series. On average, the error in our predictions is around 0.2 points on the one to five LPI scale, or about 6%. We conclude that ML therefore represents a promising set of techniques for expanding the coverage of indicators like the LPI.

The great benefit of our approach is that it greatly expands the global knowledge base on logistics performance. The original LPI covers between 146 and 155 countries for six periods between 2007 and 2018. Our ML-generated predictions, on the other hand, systematically cover 199 countries for all years between 2001 and 2019. We therefore provide a longer-term framework within which to understand evolutions in performance, as well as one that is more representative in terms of countries covered.

While many papers in a supply chain management, logistics, engineering, and economics use the LPI as an explanatory variable or predictor, there is only one other we are aware of that uses the LPI as a dependent variable. The closest paper to ours is therefore Hillberry and Zhang (2018), which uses one ML technique—Least Absolute Shrinkage and Selection Operator (LASSO) regression—to identify the most important determinants of logistics performance. Their approach shows the power of ML in terms of producing accurate prediction, as well as what is termed “feature selection”, namely narrowing down the set of explanatory variables in the models to those with the most power in terms of producing accurate predictions. We build on and extend their work by considering other ML algorithms. Although our focus is more on prediction than inference, we follow Hillberry and Zhang (2018) in using the output of our feature selection algorithms applied as part of an overall ML technique to identify groups of economic and other factors that positively or negatively impact LPI scores. Our approach differs from theirs in that we start with the same dataset, but do not rely on analyst judgment to select possible correlates of logistics performance: we therefore have a wider range of data available, with greater disaggregation in areas like infrastructure, economic performance, and openness to trade.

From a methodological standpoint, our paper takes a different approach from most. It is not theory driven, in the sense of choosing explanatory variables using researcher judgment. Instead, it is data driven: it is patterns in the data themselves that determine which variables play an important predictive role. Our standpoint is therefore an empirical one, based on prediction, rather than a theoretical one, based on understanding or explanation. Our approach does not seek to supplant theoretical work, but rather to complement it. As mentioned above, the key contribution of our paper in a concrete sense is a freely available, expanded dataset of logistics performance. We hope other researchers will use these outputs to better understand logistics and supply chain processes, including through the use of a range of theoretical perspectives. A key benefit of favoring this kind of data-driven research, in addition to theory-driven contributions, is that it can open up a field to new insights and changing views. In economics,

for instance, empirical work that started as a pure measurement exercise—i.e., with very little theoretical input—has radically altered the profession’s views on the impacts of minimum wages (see initial work by Card and Krueger, 1995; and a consolidation of subsequent work by Cengiz et al., 2019).

A key concern for any ML approach, including ours, relates to the quality of the input data. We have chosen the WDI database as our data source because it is the standard reference source for economists and other social scientists who study economic and social structure of countries. It contains detailed information on trade, which is particularly important for the LPI, such as the total value of trade flows, trade relative to GDP, and the composition of trade by broad groups (manufactures, food, fuels, high technology goods, etc.). The fact that the WDI is the standard data source for analysis of development-related issues around the world means that data quality is as high as possible, even though there are reliability issues in the source data for some countries. We deal with the issue of missing data in the WDI dataset by using multiple imputation, a standard statistical technique. Finally, an indirect check on the quality of the input data is to see how well they predict the output data, in this case the LPI. If data are overly noisy or inaccurate, they will not produce systematically accurate predictions of an output variable. We find the opposite here: ML techniques applied to the WDI data make it possible to produce highly accurate predictions of the LPI. So we are confident that data quality is as high as possible, given our objective of global coverage, and the need to focus in particular on smaller, lower-income countries where data collection is very challenging.

Against this background, the paper proceeds as follows. The next section discusses our dataset, including transformation and pre-treatment that is necessary prior to applying ML. Section 3 then looks in detail at our methodology, starting from a general description of the problems of prediction and feature selection, and moving into a presentation of the different ML techniques we employ. The following section then presents results, and discusses their implications. The final section concludes.

2 DATA

2.1 Dependent Variable

Our primary objective in this paper is to use ML to produce predictions of observed LPI scores that are as accurate as possible. Our dependent variable in all cases is therefore a set of LPI scores. The untransformed index ranges between one and five. We retain all six editions of the LPI, from 2007 to 2018. In line with the way the LPI is reported by the World Bank, we round scores to two decimal places.

We split the LPI sample into training and testing sub-samples by random assignment. The training sub-sample is an estimation sample in econometrics, while the testing sub-sample is used for out-of-sample predictions. The dataset is small by ML standards, with 683 observations covering 162 countries in the training sub-sample, and 227 observations covering 133 countries in the testing sub-sample.

We perform two transformations on the LPI. First, we use the max-min normalization—subtraction of the minimum and division by the range—to produce a variable that ranges between zero and unity. We then apply a logit transformation to give a series with infinite range. Using an infinitely ranged series broadens the range of ML techniques we have access to, and

requires weaker assumptions. Back-transformation of predictions means that our predicted LPI values lie on the same scale as the original observations.

2.2 Explanatory Variables

Unlike Hillberry and Zhang (2018), we do not have any theoretical priors about the number or type of variables that could be useful in predicting LPI scores. We therefore start with the broadest possible dataset. Conceptually, a wide range of variables could be associated with logistics performance, ranging from infrastructure quantity and quality, to per capita income, to international linkages such as trade flows and air or maritime connections.

We therefore use the full World Development Indicators (WDI) dataset as the starting point for our analysis. Hillberry and Zhang (2018) use a subset of this data. The rationale for using the WDI dataset is that it is the broadest dataset available with a significant time series of observations on country features for essentially all countries in the world. Those features cover a broad range of characteristics, including economic, social, geographical, and policy factors. We retain 2001 through 2019. The dataset includes 1421 variables, however there are many missing observations. We therefore drop all series that have more than 25% missing observations. We also drop series that have a correlation coefficient of 0.75 or greater: individual series do not add much information above what is contained in one of the set. This process leads to 140 variables, to which we add four dummy variables for World Bank income groups. We complete remaining missing observations in the data using multiple imputation (Honaker et al., 2012).

Logistics performance is a many-faceted concept. Its determinants potentially cover a wide range of areas, from economic and social variables, to infrastructure quality, governance, and geography. The LPI is an index that covers six independent dimensions. Starting from the broadest possible set of variables therefore has the advantage of allowing this full range of factors to play a role in predicting unobserved LPI values to extend it. At the same time, feature selection can play a descriptive role by identifying country characteristics that are most predictive of performance—information that is important for policymakers interested in improving performance over time.

The WDI database contains numerous variables that are of direct relevance to a country's level of logistics performance as captured by the LPI. For instance, the ratio of trade to GDP is an indicator of a country's level of trade dependence: a higher value creates an incentive to improve logistics performance, so as to maintain export income. Another example is the proportion of paved roads as an indicator of infrastructure quality (one of the LPI dimensions), or the amount of traffic moving through ports both as an indicator of infrastructure quality and the level of service provision (both of which are core dimensions of the LPI). Similarly, the proportion of the population living in urban centers proxies for the extent to which logistics at international gateways is determinative of a country's overall level of performance: given that the LPI survey is administered to international professionals (i.e., not resident in the country being ranked) they likely focus on the performance of international gateways, rather than logistics in the hinterland (Arvis et al., 2007). Finally, the version of the WDI database used here also includes information from the Doing Business project on time delays and costs associated with moving goods across borders, which are related to the efficiency of border clearance processes and the timeliness with which goods reach their destination (also core dimensions of the LPI).

In economic growth, a wide variety of country-level variables found in the WDI dataset have been shown drive increases in per capita income. Because of conditional or unconditional

convergence over time—a phenomenon also noted in relation to the LPI (Arvis et al., 2007, 2010, 2012, 2014, 2016, 2018)—these factors are of particular importance to low income countries, which is where coverage of the LPI is sparsest. Examples of important growth predictors include years of education (a proxy for human capital; Mankiw et al., 1992), and control of corruption, (Grundler and Potrafke, 2019). In addition, it is primarily low and middle income countries where the most work remains to be done in terms of fostering trade integration (e.g., Hoekman and Nicita, 2011).

In terms of pre-treatment, we apply the Box and Cox (1964) transformation to all 140 variables. This transforms them to normal distributions, which will perform better with ML. Finally, we center and scale the variables by subtracting the mean and dividing by the standard deviation.

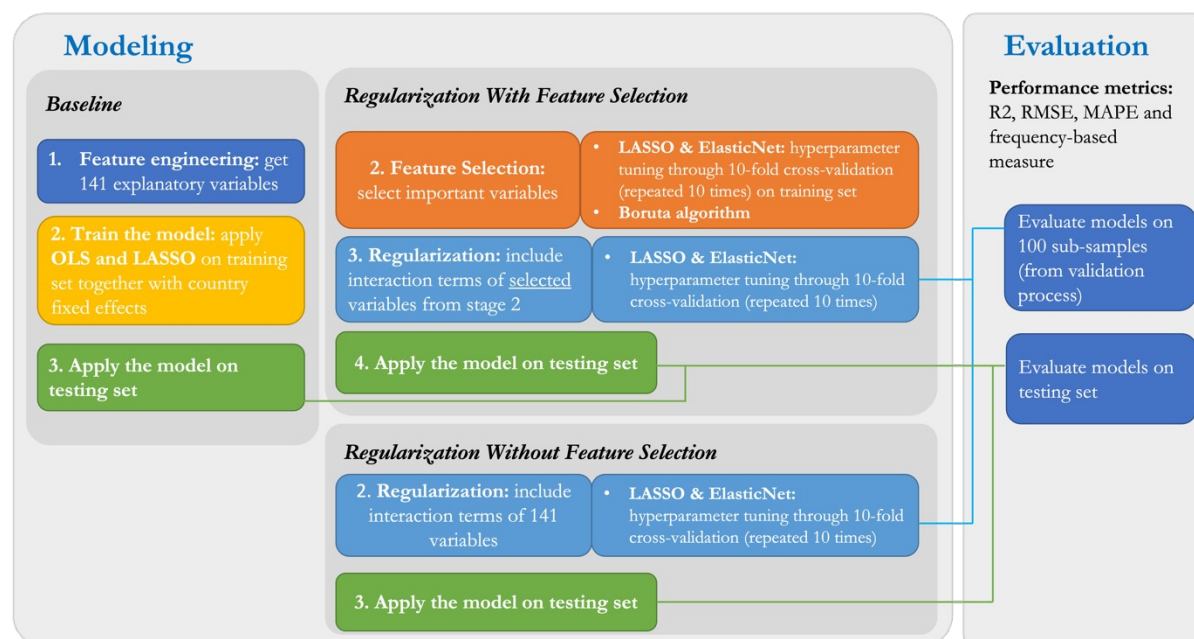
Our full database of explanatory variables consists of all interactions and squared terms, as well as the original variables. This procedure—common in ML—maximizes variation in the data, and helps capture nonlinearities. It results in a dataset of 19,740 variables. The final models use a smaller subset of those variables selected by the ML algorithms, but this approach allows us to start with the broadest possible information base.

3 METHODOLOGY

A wide range of ML techniques are available to assist with the prediction problem we have developed. We are agnostic as to which of these techniques will perform best, so we move from simple and well-established methodologies to more recent developments. A particular concern is reconciling the selection of important explanatory variables (feature selection) with predictions that are as accurate as possible.

Our approach applies a variety of different models. Figure 1 sets out our general methodology.

Figure 1: General modeling approach and methodology flowchart.



(Figure by authors.)

In assessing model performance, it is important to have a benchmark. We therefore employ a simple statistical model that nonetheless captures an important aspect of how logistics performance is conceptualized. We run an OLS regression using per capita GDP as the only predictor. This baseline model is important, as it represents a “naïve” model that could be used to predict logistics performance, since it is widely believed to correlate with country income levels (e.g., Arvis et al., 2007). Comparing the per capita income model with the ML models will make clear whether or not our approach has any substantial value added.

3.1 Feature Selection

Feature selection is one part of an ML algorithm, and means using a methodology and a benchmark to narrow the set of explanatory variables so as to retain only the most important from a prediction standpoint. Applying feature selection before including interaction terms is common practice for two reasons (Kuhn and Johnson, 2019). First, it is more efficient in term of computation cost. Second, the larger number of irrelevant interactions lead to a stronger penalty on interaction variables and lower the chance of getting truly important interactions, which usually are small. Nevertheless, feature selection should lower the overfitting problems created by large numbers of crossed features with a large set of predictors.

The first set of feature selection algorithms we apply focus on prediction accuracy using the standard 10-fold cross-validation method, combined with different approaches for eliminating features with little predictive power. The LASSO and ElasticNet approaches fit within this category. Both are regression models that are closely related to OLS. LASSO (Tibshirani, 1996) applies a penalty in the objective function based on the absolute value of the coefficients, while ElasticNet (Zou and Hastie, 2005) applies a penalty that is a combination of the LASSO penalty and a second term based on the square of the coefficient. In both cases, the penalty “shrinks” coefficient estimates towards zero relative to OLS. Features that have coefficients equal to zero can be dropped from the model.

Formally, the ElasticNet, which nests LASSO as a special case, can be written as follows, where $\hat{B}$ is an estimated coefficient. Each algorithm solves a problem like the following (Hastie et al., 2009):

$$\hat{B} = \underset{B}{\operatorname{argmin}} \left\{ \sum_{i=1}^N \left(y_i - B_0 - \sum_{j=1}^p x_{ij} B_j \right)^2 + \sum_{j=1}^p a |B_j| + (1-a) B_j^2 \right\}$$

In this case, y is the variable we want to predict, while the set of $p \times x$ terms represents the available predictors (features). The first term in parentheses is simply the OLS objective function. The final terms are the shrinkage penalties referred to above. a is a parameter that determines the weights of the two penalty factors, with the special case of $a=1$ producing Lasso estimates. The parameter a is a model hyperparameter that is chosen by grid search with cross-validation.

Standard practice for using the LASSO and ElasticNet algorithms to choose explanatory variables is to combine the penalty approach relative to OLS with 10-fold cross-validation (repeated 10 times). In this approach, the data is randomly split into ten equally sized sub-samples. The regularization algorithm is then run 10 times, on each occasion retaining one sub-sample for out-of-sample predictions. To minimize sampling bias, the process is repeated 10 times for each value of the penalty parameter. The selected model is then the one with the lowest error (RMSE) or highest goodness-of-fit out of all the models estimated, averaging over the 100 held-out-

samples. The rationale for combining the penalty with cross-validation is that it focuses on out-of-sample prediction, which is ultimately the objective of applying ML in this context. By introducing randomization into the construction of sub-samples, cross-validation helps ensure that models are not over-fitting the data used for regularization.

The second feature selection approach we use is the Boruta algorithm (Kursa and Rudnicki, 2010). This algorithm is a first stage that can feed into regularization using other approaches, such as Lasso. The essence of this approach is to create a “shadow” dataset by randomly resampling from the explanatory variables. The next stage is to apply an ML technique to produce predictions of the variable of interest, using both the actual and shadow datasets. By comparing statistical significance between observed and shadow variables based on standard metrics, the algorithm can assess whether or not a given variable is performing better than would be expected by chance, since the shadow dataset is based on random resampling. The algorithm then retains those features that perform relatively well by iterating a set number of times.

3.2 Regularization

Regularization refers to the algorithm used to fit a model to the data, similar to the choice of estimator in econometrics. Following feature selection, we apply two regularization methods. Using the set of explanatory variables identified in the first stage, the LASSO estimates coefficients by penalizing OLS by a factor depending on the absolute value of the coefficients, thereby shrinking them towards zero. The ElasticNet takes the same approach, but applies a mix of penalties, based on the absolute values of the coefficients, and their squares. In both cases, the penalty parameters are selected through the 10-fold cross-validation with 10 resamples.

3.3 Advanced ML Techniques

Advanced ML techniques refer to those algorithms that go substantially beyond what can be contained in a regression-like framework. Basic regularization algorithms are, like OLS, linear models. By creating a set of explanatory variables that consists of all interactions and squared terms, we allow for a degree of nonlinearity. However, the approach is somewhat mechanical, and does not capture nonlinearities outside of these strict confines.

Advanced ML techniques incorporate more elaborate nonlinearities directly. We consider two such techniques here: Random Forests, and Support Vector Machines. In both algorithms, the optimal hyperparameters are selected based on the best average RMSE over the 100 held-out-samples.

The Random Forests algorithm (Breiman, 2001) aggregates performance across a large number of decision tree regressions. The basic approach is to average a potentially large number of noisy but unbiased estimates. Each decision tree is constructed by splitting the dataset on particular attributes to obtain the maximum reduction in variance in the prediction variable, i.e. from unconditional to conditional measures. Aggregation takes place by choosing a set of data points at random from the dataset and constructing a decision tree regression, then repeating across a pre-set number of trees. The model does not impose linearity.

More formally, Hastie et al. (2009) set out the Random Forests algorithm as follows:

1. For $b = 1$ to B :
 - a. Draw a bootstrap sample Z^* of size N from the training data.

- b. Grow a random forest tree T_b to the bootstrapped data, by recursively repeating the following steps for each terminal node of the tree, until the minimum node size n_{min} is reached.
 - i. Select m variables at random from the p variables.
 - ii. Pick the best variable/split-point among the m .
 - iii. Split the node into two daughter nodes.
 2. Output the ensemble of trees $\{T_b\}_1^B$.
 3. To make a prediction at a new point x :
 - a. $\hat{f}^B(x) = \frac{1}{B} \sum_{b=1}^B T_b(x)$.

The key ingredient in the Random Forests algorithm is the regression tree. To see how this technique works, we follow Hastie et al. (2009) in considering a dataset of p inputs and a response variable for N observations. We partition the model into M regions R , and model the response as a constant in each region:

$$f(x) = \sum_{m=1}^M c_m I(x \in R_m)$$

Based on a minimization of the sum of squares criterion, the optimal choice $\hat{c}_m$ is simply the average of y in each region.

Starting with the full dataset, we now consider a splitting variable j and a split point s and define the pair of half-planes:

$$R_1(j, s) = \{X | X_j \leq s\}; R_2(j, s) = \{X | X_j > s\}$$

Using the same optimality criterion, the optimal value is again simply the average in the region:

$$\hat{c}_1 = \text{ave}(y_i | x_i \in R_1(j, s)); \hat{c}_2 = \text{ave}(y_i | x_i \in R_2(j, s))$$

The (j, s) pairs are determined by scanning through all of the inputs. Once the best split is available, the data can be partitioned into two regions, and the process can be repeated until some stopping criterion is reached. It is this algorithm that is at the center of the Random Forests approach, presented above.

Finally, a Support Vector Machine (Cortes and Vapnick, 1995) is a regression model that uses a hyperplane rather than a line, as in linear regression, to make predictions on the dependent variable. Data points close to the hyperplane are called support vectors, and influence its shape. SVMs allow for nonlinearities, and use a kernel function to identify the hyperplane that produces the most accurate predictions of the variable of interest. In this study, the Radial Basis Function Kernel is employed.

The simplest representation of a Support Vector Machine is in terms of a regression problem (Hastie et al., 2009). We start with the standard linear regression model:

$$f(x) = x' B + B_o$$

To estimate the parameters B , we consider minimization of the following loss function:

$$H(B, B_0) = \sum_{i=1}^N V(y_i - f(x_i)) + \frac{\lambda}{2} \|B\|^2$$

Where:

$$V_c(r) = \begin{cases} 0 & \text{if } |r| < c \\ |r| - c & \text{otherwise} \end{cases}$$

This error measure effectively ignores very small errors, below some threshold c that is chosen as a model hyperparameter.

Hastie et al. (2009) show that the solutions to minimizing the above expression for H take the following form:

$$\begin{aligned} \hat{B} &= \sum_{i=1}^N (\hat{a}_i^* - \hat{a}_i) x_i \\ \hat{f}(x) &= \sum_{i=1}^N (\hat{a}_i^* - \hat{a}_i) \langle x, x_i \rangle + B_0 \\ (\hat{a}_i^*, \hat{a}_i) &= \min_c \sum_{i=1}^N (a_i^* + a_i) - \sum_{i=1}^N y_i (a_i^* - a_i) + \frac{1}{2} \sum_{i,i'=1}^N (a_i^* - a_i) (a_{i'}^* - a_{i'}) \langle x_i, x_{i'} \rangle \end{aligned}$$

Subject to:

$$\begin{aligned} 0 &\leq a_i, a_i^* \leq \frac{1}{\lambda} \\ \sum_{i=1}^N (a_i^* - a_i) &= 0 \\ a_i a_i^* &= 0 \end{aligned}$$

3.4 Assessing Model Performance

Each of the ML models produces predicted values of the LPI. To assess model performance, we focus on out-of-sample fit: we use the training data to select model parameters and, in some cases, a set of explanatory variables, and then we use the model to produce predictions of the LPI for data not used in training, that is the 25% of the total sample reserved for out-of-sample testing.

To assess performance, we do not assume that any one metric is superior. As a baseline, we use R^2 , as in a standard regression context. It summarizes the proportion of observed variation in out-of-sample LPI that is accounted for by variation in predicted LPI. So it is a simple measure of goodness-of-fit. Kuhn and Johnson (2019) argued that R^2 can have very encouraging performance in models where the predicted variables have large variance.

In addition, we calculate the square root of the average squared error (difference between observed and predicted LPI values; RMSE). This measure is one way of summarizing how large the prediction error is. A complementary measure is mean absolute percentage error (MAPE), which translates prediction errors into absolute valued percentage terms, then takes the average.

Finally, we use the fact that the LPI contains an estimated 80% confidence interval to construct a frequency-based measure of goodness-of-fit. We consider a prediction “accurate” if it is within the 80% confidence interval, and “inaccurate” if it is not. So the measure summarizes the proportion of out-of-sample data that is accurately predicted, in this sense.

In an ML context, each indicator can be calculated using up to three sets of data. All methodologies have scores for the training and validation sub-samples, i.e. the 75%-25% split with the former used to fit the model, and the latter used for out-of-sample prediction (an average, in the case of cross-validation).

4 DISCUSSION AND IMPLICATIONS

4.1 Model Performance and Selection

Table 1 contains results for the above performance metrics. In the best performing models, RMSE is around 0.2 in the validation data. Results for the validation and testing sub-samples are very close, as hypothesized above. Moreover, the difference between RMSE for the training data and the validation data is not too large, so there is little evidence of overfitting. An RMSE of 0.2 suggests that an “average” error is around 0.2 points on the one to five scale used by the LPI. In quantitative terms, this difference is not very large, given that overall performance is spread across a substantial band within the overall space. However, a well-known issue with the LPI data is that countries tend to cluster in the middle range (Arvis et al., 2007, 2010, 2012, 2014, 2016, 2018); so for many countries, a difference in score of 0.2 is not negligible in terms of the distribution of scores within a particular range, but this sensitivity is more a function of the original data, which display more variation in country scores in a year-on-year sense than could be explained by substantive changes alone.

A similar picture emerges when considering errors in percentage terms, through the MAPE metric. The best performing models have scores of around 0.06, which means that an “average” error is approximately 6% of the observed score. Again, this deviation is not large in relative terms, as we argued above for the case of absolute deviations. But the issue of clustering of values in the mid-ranges of the original data remains. However, considering only scores, the predicted values out-of-sample are remarkably close to the observed scores given that only a relatively small amount of data is available for training the models.

With these findings in mind, the remaining metrics show that the out-of-sample predictions closely match observed variation in the data. The R^2 of the best performing models is just under 0.9, so even out of sample, predicted values account for around 90% of observed variation in the true data. This level of fit indicates a very close relationship between actual and fitted values. Again, there is only a relatively small difference in performance between the training and validation scores, which means that overfitting is not too much of a problem. Finally, the frequency measure shows that the leading measures predict a value of the LPI that lies within the estimated 80% confidence interval from the original data around 64% of the time. This is a relatively strong performance, but suggests that sampling variation in the underlying data—which is based on a survey, as discussed above—means that some movement in scores is idiosyncratic and difficult to replicate using other observed data.

A key comparison is between the ML models and the baseline model, namely the OLS regression with per capita income as the only predictor. Table 1 shows that, in approximate terms, the ML models all produce predictions that are around twice as accurate as this simple model. In other

words, our approach adds a large amount of value in terms of predicting logistics performance relative to what could be understood using per capita income only. It is important both to use different regularization algorithms, and also different input datasets, to properly understand the determinants of logistics performance, and use them to predict the LPI out of sample.

Using these metrics, a key choice relates to the best performing model. Overall, that is a variation on the LASSO using the Boruta algorithm to choose features in a first step. There is no ambiguity in this finding, as the chosen model is the best (or equal best) performer on all metrics. We therefore use results from this version of the predicted data in presenting further results below.

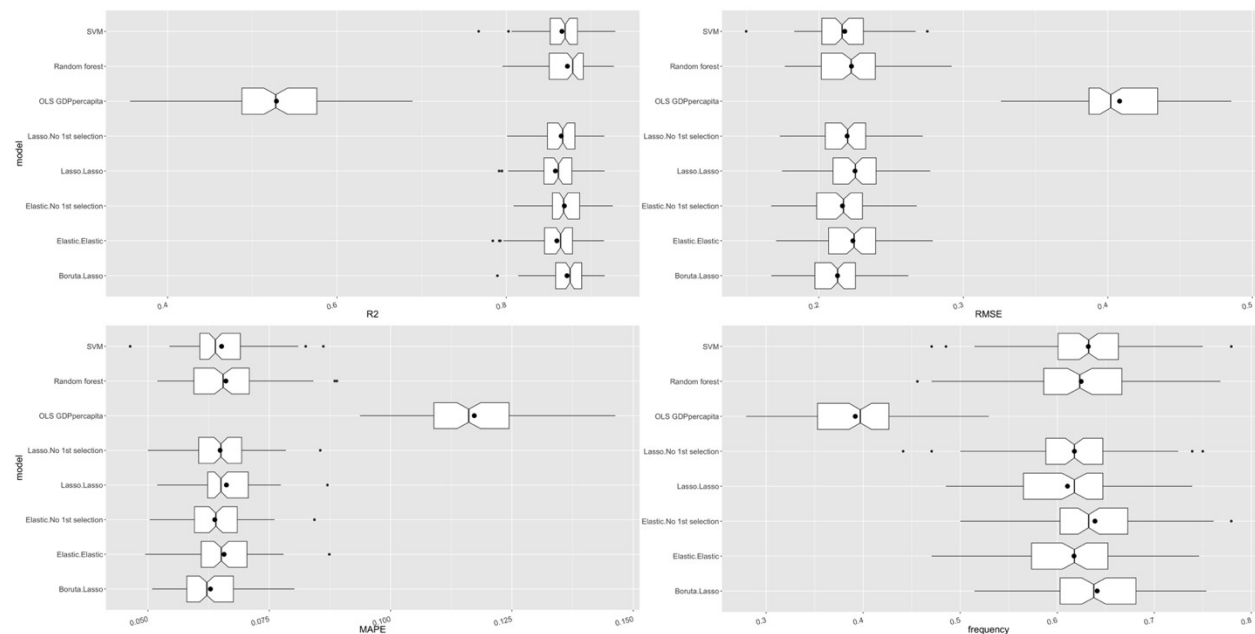
Table 1: Model performance metrics on test sample

Model	R2			RMSE			MAPE			Frequency		
	Test	Train	Valid.	Test	Train	Valid.	Test	Train	Valid.	Test	Train	Valid.
<i>Baseline: GDP per capita only</i>												
OLS	0.485	0.522	0.529	0.414	0.409	0.408	0.117	0.113	0.117	0.388	0.392	0.392
<i>Regularization Without Feature Selection</i>												
LASSO	0.866	0.917	0.865	0.211	0.173	0.220	0.063	0.050	0.065	0.599	0.703	0.618
ElasticNet	0.862	0.928	0.869	0.214	0.162	0.216	0.064	0.047	0.064	0.612	0.736	0.639
<i>Regularization With Feature Selection</i>												
Boruta-LASSO	0.872	0.935	0.872	0.206	0.152	0.213	0.062	0.045	0.063	0.617	0.769	0.641
LASSO	0.851	0.904	0.858	0.222	0.186	0.225	0.065	0.053	0.066	0.599	0.690	0.611
ElasticNet	0.843	0.915	0.860	0.228	0.176	0.224	0.066	0.050	0.066	0.626	0.700	0.617
<i>Advanced Models</i>												
RF	0.841	0.941	0.872	0.236	0.158	0.223	0.068	0.045	0.066	0.581	0.760	0.625
SVM	0.829	0.968	0.866	0.238	0.106	0.218	0.068	0.029	0.065	0.634	0.908	0.632

(Table by authors.)

To look in more detail at model results, Figure 2 contains boxplots of our four metrics for each model, focusing on testing sub-samples. There are multiple testing samples because we use 10-fold cross-validation (repeated 10 times), so the figures in Table 1 represent averages over those different samples. Figure 2 presents the same information more completely. The first key finding is that all of the ML models add a substantial amount of predictive accuracy and understanding relative to a simple model based only on country per capita income. Second, the overall result is unchanged, as the LASSO model with first stage Boruta algorithm performs best in median terms (the central line) across all metrics excepting R2. However, the interquartile range makes clear that there is in fact relatively little difference in performance from one ML model to another. The overall conclusion stands, but with the nuance that the chosen model's performance is not clearly superior to the set of candidate models as a whole in an overall sense as opposed to an average or median sense. Nonetheless, an additional reason for preferring for the LASSO model with Boruta feature selection is that it is relatively sample compared with some of the other options, and therefore has the benefit of parsimony. The same is true of the variable set, as the two approaches to feature selection—Boruta algorithm, followed by LASSO with cross-validation for penalty weight selection—results in a model that uses fewer variables than some of the alternatives, in particular the Random Forests and Support Vector Machine models.

Figure 2: Model results, testing sub-samples.



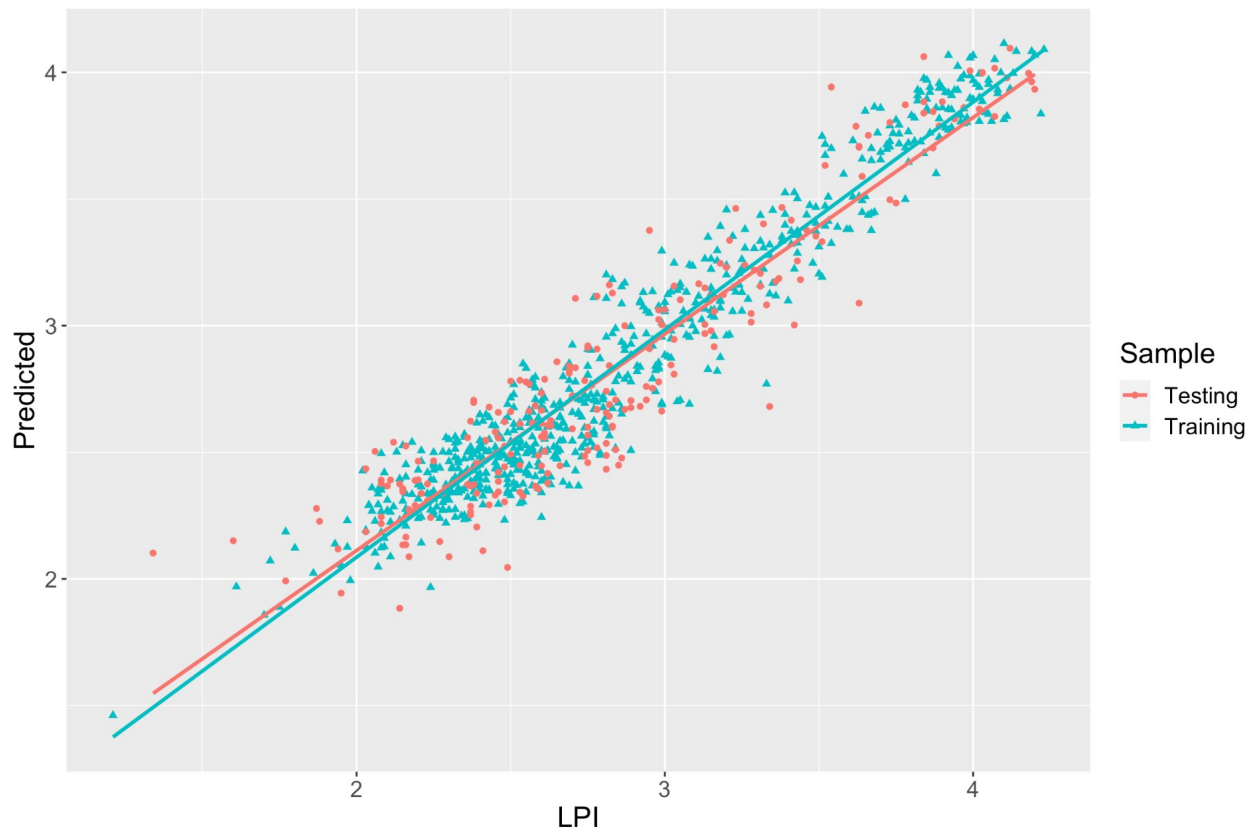
(Figure by authors.)

Figure 3 summarizes results from the preferred model—LASSO with first stage Boruta feature selection—by plotting actual against predicted values. The association is very close, and the lines of best fit are almost identical for the training and testing sets. Visually, there is no evidence that predictions in the training data are substantially better than predictions in the testing data, which is an indication that the model has been appropriately specified and selected.

Of course, there is some amount of variability in the predicted values around the observed values. This outcome is to be expected. But the LPI is a survey methodology and so contains a measure

of sampling variability in addition to information on key dimensions of performance. It is in line with expectations that seeking to reproduce the index should also produce results that vary somewhat. The important point is that much of that variation appears random, as it is not significantly different depending on whether high, low, or mid-performance countries are considered. Indeed, a remarkable feature of Figure 3 is that a simple model based only on readily available data can nonetheless yield very accurate predictions in an overall sense, subject to the issue noted above about bunching of observations in the mid-ranges of the LPI, which emerges clearly from Figure 3.

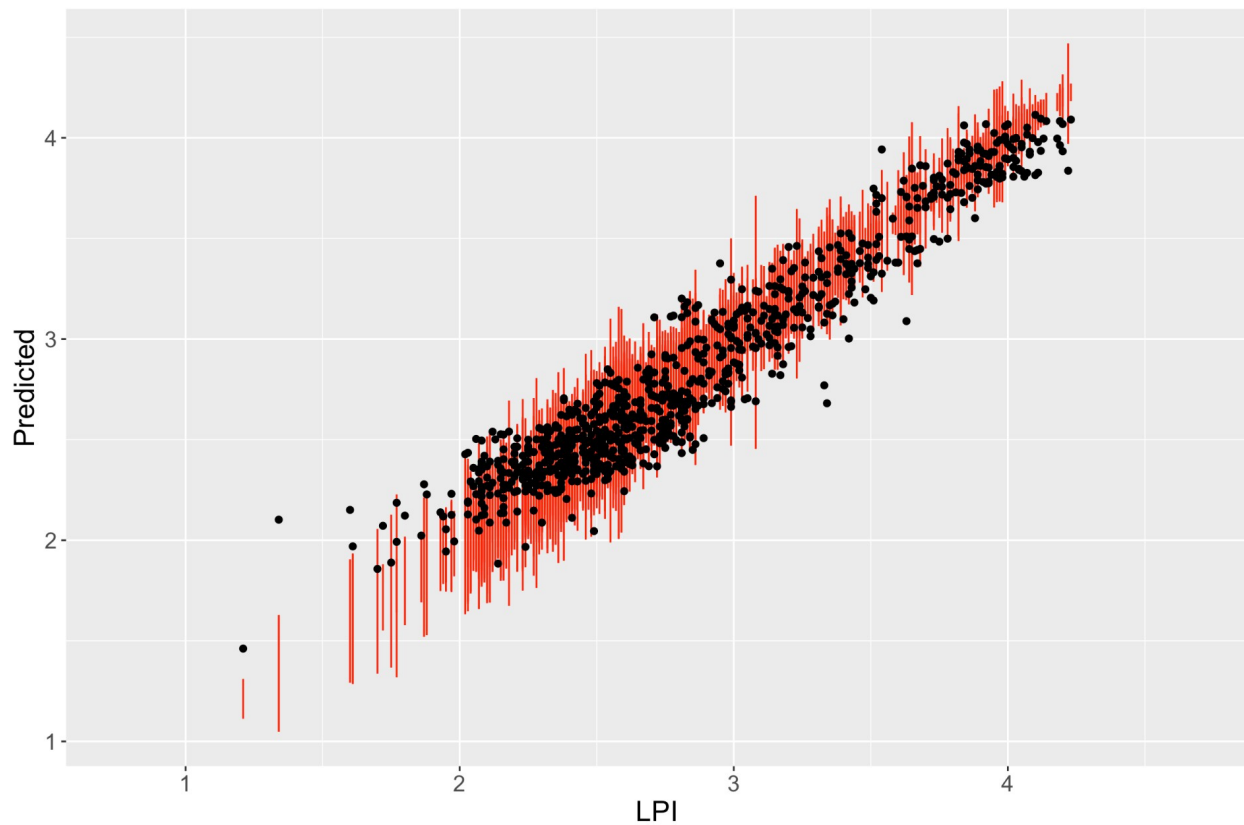
Figure 3: Actual vs. predicted LPI values.



(Figure by authors.)

Figure 4 addresses the issue of sampling variability more directly. It plots predicted LPI scores from the ML model against vertical lines representing the estimated 80% confidence interval around the reported LPI scores from the World Bank. If the predicted value of the LPI falls within the vertical band, it is within the confidence interval, and so the hypothesis that the observed and predicted values of the LPI are identical cannot be rejected. As the figure makes clear, only a relatively small number of observations falls outside the confidence interval. In most cases, the predicted value is well within the confidence interval, which we interpret as indicating that systematic—as opposed to random—differences between the model's predicted values and observed LPI values are limited. While the problem is not absent, the overall result again suggests that the model has performed well, in particular given the limited quantity of data available for training.

Figure 4: Actual values with 80% confidence intervals vs. predicted values.



(Figure by authors.)

Table 2 shows that while using an ML model to predict the LPI results in some degree of variability beyond what is contained in the original data, the benefits are compelling. The first column shows that the LPI covers 146-155 countries every two to three years from 2007 through 2018. By contrast, the ML model extends this coverage to 199 countries, every year from 2001 through 2019. So country coverage is wider, including in particular smaller, lower income countries. In addition, the availability of a longer time series makes it possible to say more about the ways in which performance changes throughout the sample period.

Table 2: Country coverage by year.

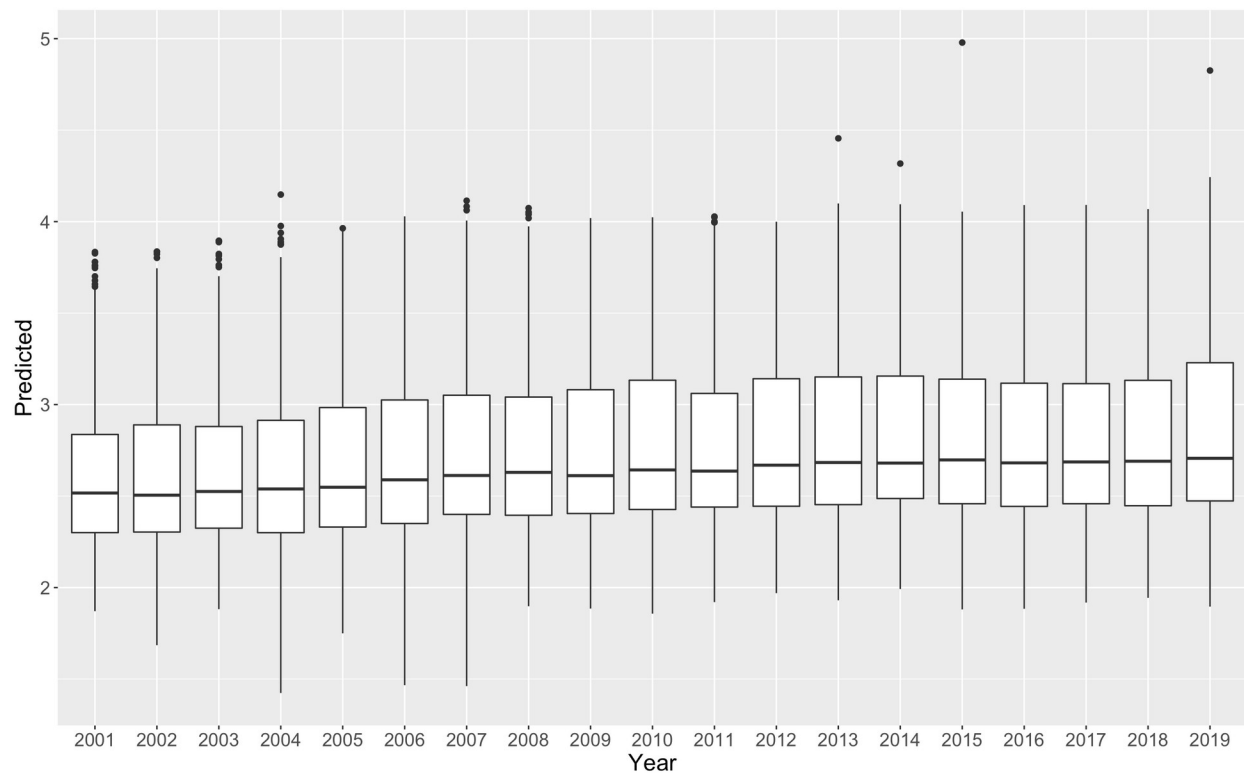
	LPI	Model
2001	0	199
2002	0	199
2003	0	199
2004	0	199
2005	0	199
2006	0	199
2007	146	199
2008	0	199
2009	0	199
2010	150	199

2011	0	199
2012	149	199
2013	0	199
2014	155	199
2015	0	199
2016	155	199
2017	0	199
2018	155	199
2019	0	199

(Table by authors.)

Figure 5 turns to this issue of the time series distribution of scores. Using model estimates, it contains a boxplot summarizing the distribution of scores by year. The first dynamic that emerges is that median scores only see very minor changes over time. By contrast, there is more movement at the upper and lower ends of the distribution, with notable improvements in performance. However, there are also outliers, in particular in later years, with very high scores. Examining the boxplots suggests that it would be prudent in policy analysis to discard these scores, as they are very much outliers in terms of the underlying pattern of performance observed across countries and years. Looking in more detail at outlier countries—the thirty largest prediction errors—shows that they are robustly associated with states that could be considered disrupted or fragile. For instance, correlation between the prediction error and indices of refugees and internally displaced persons, external intervention, security concerns, and group grievances range between 0.3 and 0.4. So the algorithm has greater difficulty predicting scores in environments where state performance is weak, which in turn would likely be associated with poorer quality of the underlying WDI data, as well as, greater uncertainty affecting logistics. The reason for this conclusion is that the WDI dataset is effectively a data repository: for the most part, the World Bank does not collect WDI data directly, but instead relies on official statistics furnished by governments and other international agencies. So in environments of weak state capacity, the information supplied by governments may be of poorer quality. Our results reflect this dynamic.

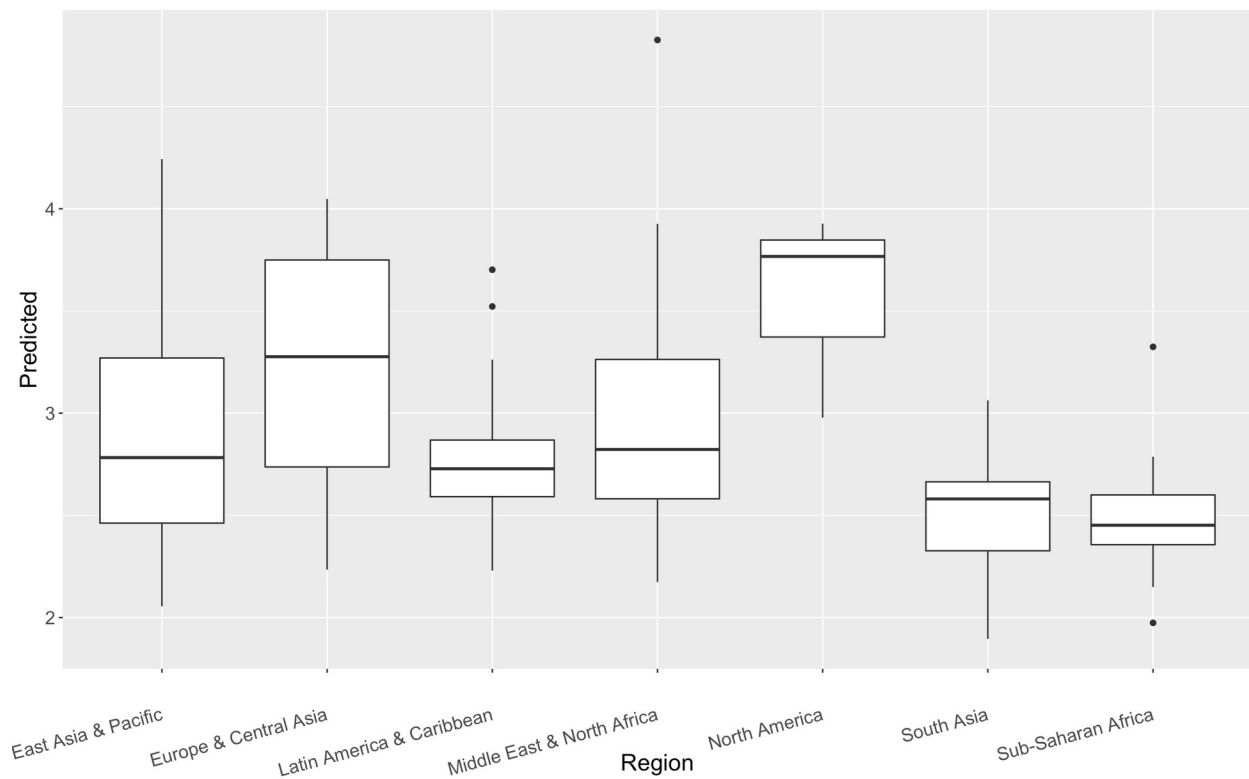
Figure 5: Distribution of LPI scores by year.



(Figure by authors.)

Figure 6 breaks out the predicted LPI scores by World Bank region. In line with Arvis et al. (2007, 2010, 2012, 2014, 2016, 2018), we find that performance in most regions is overlapping—that is, all regions contain a substantial range of performance at the country level. However, North America and Europe and Central Asia stand out as two regions with substantially higher performance, while South Asia and Sub-Saharan Africa stand out as the lowest performing regions. While the gap between the lowest and highest performing regions is clear, the new data suggest—in line with Arvis et al. (2007, 2010, 2012, 2014, 2016, 2018)—that the dominant feature for other regions is dispersion of scores across countries, from relatively low to relatively high performance. This dispersion is linked to the phenomenon of outliers, discussed above. The geographical split of results clearly suggests that environments with relatively weak or disrupted states are prone to having outlier LPI predictions, likely for the reasons discussed above.

Figure 6: Distribution of LPI scores by region, 2019.



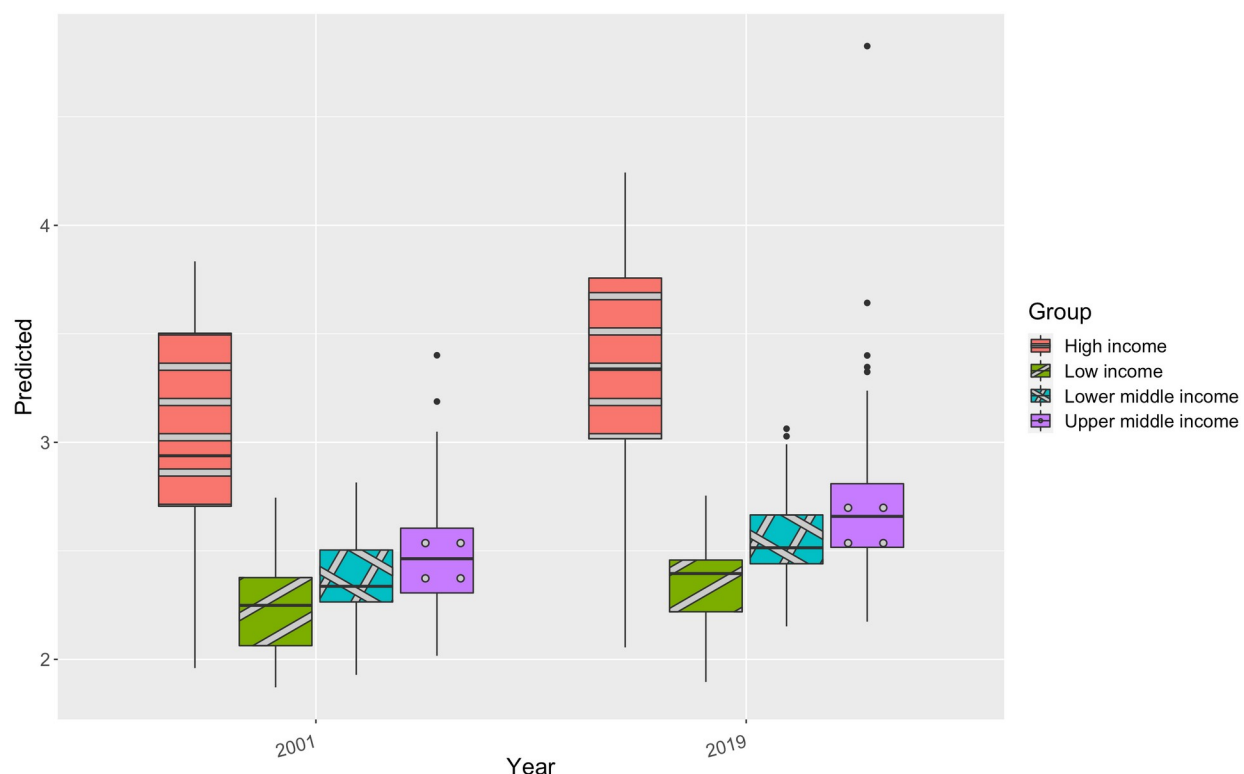
(Figure by authors.)

Finally, Figure 7 shows boxplots for scores by income group at the endpoints of the sample. Overall findings are in line with Arvis et al. (2007, 2010, 2012, 2014, 2016, 2018), but are based on a much broader range of countries. The high income group stands out in both periods as having substantially higher performance than other groups. Moreover, there is a noticeable improvement in performance between 2001 and 2019. Performance changes are more muted in other groups, but the trend towards improvement is general. However, there is a persistent correlation between per capita income and country LPI performance. As in the previous figures, however, outliers also play an important role in some groups, particularly middle income countries. Analyzing these anomalous scores suggests that they are linked to relatively low state capacity, which could in turn translate into poor quality of underlying data, as well as less certain logistics performance. For policy purposes, the most serious outliers would likely be dropped from the dataset, as previously noted; they are retained in these figures for the purposes of demonstrating the overall performance of our models, as well as for highlighting the issues that make it difficult both to measure and predict logistics performance in particular country environments.

While there is an overall correlation between per capita income and LPI scores, which is consistent with the observation that higher income countries tend to exhibit stronger logistics performance, there is also considerable dispersion within income groups. For instance, the best performing low income countries exhibit a level of performance that is seen in some of the weaker high income countries. The clear implication is that per capita income is one factor that

influences logistics performance, but many others combine to produce a complex relationship between country income level and the LPI.

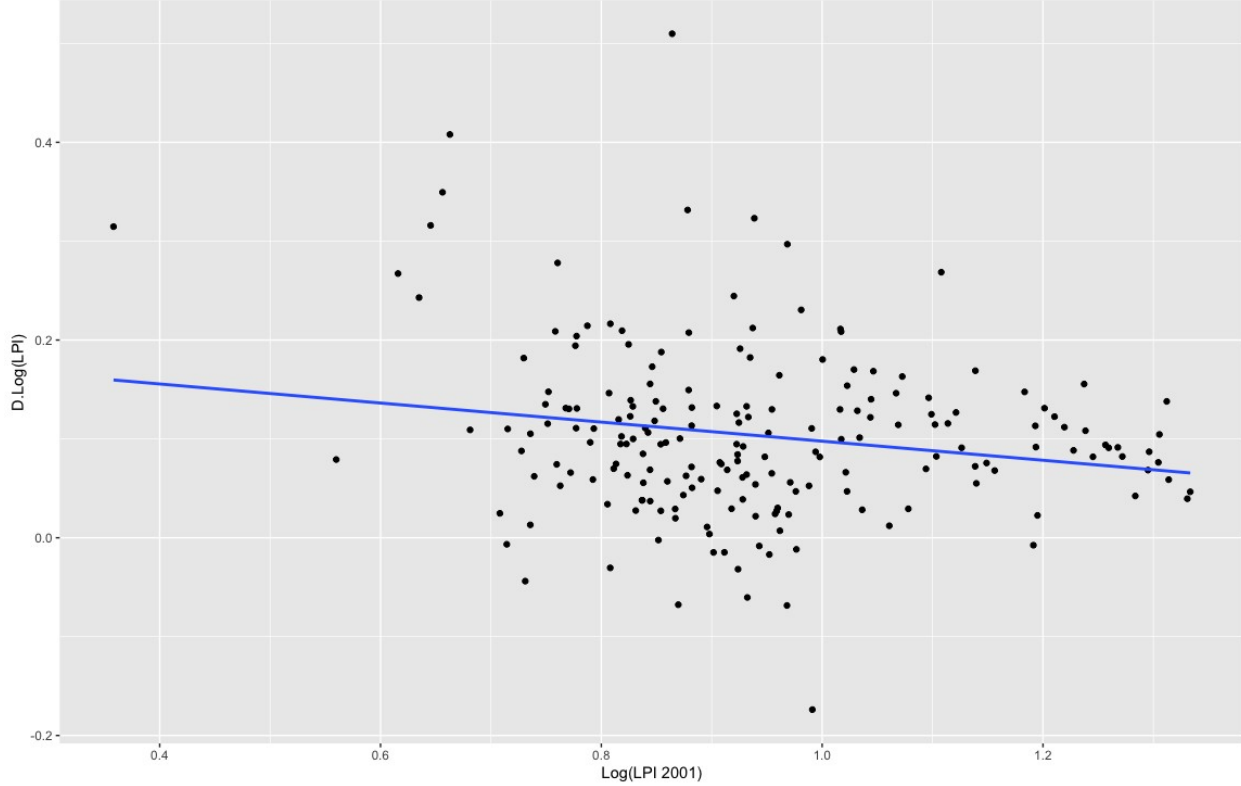
Figure 7: Distribution of predicted scores by income group, 2001 and 2019.



(Figure by authors.)

A key question in assessing changes in logistics performance over time is whether or not there is a “catch up” effect at play, i.e. that lower scoring countries improve at a faster relative rate than higher scoring countries. Given that we have more countries and time periods available than the various World Bank reports, our dataset is well-suited to answering this question. Following the economic growth literature, which this kind of catch up is referred to as “beta convergence”, Figure 8 presents the association between our predicted 2001 LPI score and the difference between the predicted scores in 2019 and 2001, all in logarithms. By plotting log first differences, we are capturing the growth rate in a country’s LPI score in percentage terms. As the figure shows, there is a clear negative association: countries with a higher initial score tend to see a slower rate of increase than countries with a low initial score. So the data support the catch up hypothesis, although there is of course some level of dispersion around the regression line. To be clear, although the catch up hypothesis is reminiscent of empirical work on economic growth, our argument is not that per capita GDP is determinative of the process at work for LPI scores. Rather, we simply note that a variety of processes—global diffusion of norms, capacity building by international organizations, and policy reforms by governments—could contribute to faster score growth in countries with initially lower scores relative to countries with initially higher scores. Economic growth could also be part of that process, but we do not single it out.

Figure 8: Change predicted LPI 2001-2019 vs. LPI score 2001 (both in logarithms).



(Figure by authors.)

5 FACTORS THAT INFLUENCE LOGISTICS PERFORMANCE

As noted above, our primary objective is to increase coverage of the World Bank's LPI dataset, and to show that ML techniques are a viable and inexpensive option for extending survey-based datasets. But a secondary objective is to understand the factors that influence logistics performance at a country level. We therefore seek to extend findings in the literature that go in this direction, by using a broader set of initial data series, and relying exclusively on a data-driven approach through the feature selection algorithm described above, namely LASSO regression with a Boruta algorithm first stage.

Previous studies have showed there are relationships between logistics performance and a wide variety of socioeconomic factors. For example, Qazi et al. (2022) studied an association between all six components of the LPI and six categories of risks including business environment, corruption, economic, environmental, financial, health and safety, and political risks. They identified these features based on risk assessments. They found logistics competence, as well as economic and financial risks are the most important drivers for all six LPI sub-indices. By using the Global Competitiveness Index, Sergi et al. (2021) found that human factors and infrastructure are key determinants of LPI scores in Europe, Asia, and Oceania. Meanwhile, those factors and infrastructure are important for improving Africa's logistic development. Magazzino et al. (2021) showed that crucial factors for improving logistics performance include technological innovation, Human Development Index (HDI), urbanization, and trade openness.

Against this background, our ML-based analysis offers potentially the broadest variety of individual indicators, given that we have started with the WDI as a feature database. But in interpreting results, it is important to distinguish correlation from causation. Our ML model is based on prediction, which is a function of correlations in the data that may or may not be causal in nature. Moreover, there is clear potential for measured variables to act as proxies for unobserved variables. There is therefore a substantial role for interpretation in grouping factors together that appear to be related in a structural sense. We therefore focus more on identifying the types of factors that are important from a predictive point of view, rather than individual variables that have that characteristic. Given the nature of the WDI dataset, many individual variables are not structural in nature, in the sense of describing deep country characteristics. Rather, they are outward expressions of structural characteristics, or measured variables that are correlated with deep determinants of country-specific factors. It is therefore important not to place emphasis on possible interpretations behind individual variables, but rather to identify broad classes of factors that the model shows play an important role in predicting LPI scores.

Our analytical tool in undertaking this assessment is an importance score calculated as an index between zero and 100, based on a rescaling of the absolute values of regression coefficients. Given that all data are standardized before entering the ML model, simple regression coefficients are exactly equivalent to standardized regression coefficients. The sign of the coefficient indicates the direction of correlation between the variable and a country's LPI score. The absolute value of the magnitude of the coefficient is an indicator of effect size.

As suggested by Figure 7, and noted by the World Bank in successive reports (Arvis et al., 2007, 2010, 2012, 2014, 2016, and 2018), a country's level of per capita income is one of the most important determinants of the LPI. However, results from our baseline model show that per capita income alone does a relatively poor job of predicting LPI scores. So it is one factor among many. Indeed, some of the variables identified as important in our analysis are correlated with per capita income, thereby highlighting the importance of economic development as a determinant of performance. But in analyzing the potential role of other factors, we aim to go beyond this finding to look at a wider range of factors that combine with income.

Our preferred model (Boruta Lasso) has 170 features including 154 interaction features and 16 original features. However, many features play only a marginal role in prediction. While the model uses the full set of features, our analysis of its results focuses on the 50 most important features, as identified using a standard metric. This approach makes it possible to see that the interaction between high/upper middle income dummy variables and other socioeconomic factors strongly influences LPI scores. Half of the interaction terms (12 out of 24 features) are present in the 20 most important features, again as identified using our standard metric. This finding highlights that apart from economic aspects, factors related to the business environment, sustainable development, and institutional vulnerability are important predictors. Some of the variables identified may at first glance seem unrelated to logistics performance. They likely proxy unobserved country characteristics, such as environmental management or human capital development, which bears a closer relationship. Our selection of the 20 or 50 most important features is based on a standard metric, but the numbers are chosen arbitrarily and purely in order to facilitate text-based exploration of the determinants of performance.

For all countries, the predicted LPI is increased by good infrastructure and large trade of services or manufactured goods, while it is pulled down by a combination of poor institutional

quality, migration, resource depletion, and an underdeveloped health system. These factors are important across the board in determining logistics performance. But interactions with income group dummies also make it possible to highlight particular factors that matter more or less for particular groups.

For middle income countries, the role of trade openness stands out: variables capturing trade with other regions, in particular imports, are important determinants of LPI scores. For high income countries, economies with better environmental performance, human development, institutional quality, and business environment tend to have greater predicted LPI. However, the positive sign of the interaction terms of high income and the share of carbon dioxide emissions from solid fuel consumption could be explained by the relationship of industrialization and environmental development. Verbeke and de Clercq (2002) found that when the income of a country reaches a threshold point, the share of carbon dioxide emissions from solid fuels decreases thanks to better technologies and lower demand for dirty manufacturing.

The most important factors for low income countries remain those noted above: infrastructure, trade engagement, institutional quality, migration, resource use, and health. There is a clear progression across income groups in terms of the factors that matter most for performance. Policy priorities correspondingly differ from one setting to another. For instance, environmental protection is more important in high income countries, and, to a lesser extent, in the upper middle-income group. It is less of a priority in low and lower middle-income countries. This result sits well with the findings of Arvis et al. (2018), who show that shippers most often demand environmentally friendly options in high and upper middle-income countries. Similarly, business environment is more of a priority in high income countries than in low income countries, perhaps because it only determines performance once a certain number of basic components—such as infrastructure quality—are already in place.

In Table 3, we present a list of the most important features, namely the fifty features identified as having the highest relevance scores. They represent correlations within the data, and so cannot always be given causal interpretations. But we have sought to use them to identify the structural determinants of logistics performance by focusing on socioeconomic factors they are correlated with, and grouping them together into sets of variables with similar interpretations.

The Table shows that 21 of the fifty most important features include at least one component that relates directly to international economic integration. The most common variables relate to international trade, such as the proportion of different types of trade in total merchandise trade (agricultural raw materials, commodities, and information and communication technology products), the trade balance, or the ratio of trade to GDP. All of these measures are of direct relevance to the LPI, as it captures logistics performance primarily relating to international trade, and countries that trade more, particularly in manufactured goods, have an incentive to upgrade performance. In addition, there are measures of international financial integration, such as portfolio flows and remittances, both of which are correlated with a country's level of financial and investment capacity, which could influence its ability to build costly infrastructure.

Other variables among the fifty most important features capture the importance of the services sector, which is directly relevant to the LPI because one of its pillars is the quality and

competence of service providers. Indicators include the proportion of particular types of services trade in total services trade, and some measures of the proportion of employment in services. Also common among the list of most important features are measures of economic stability, such as the level of international reserves (relevant to avoiding balance of payment crises), broad money as a percentage of GDP and the GDP deflator (both relevant to the inflation rate), as well as domestic credit to the private sector, the savings rate, and the labor force participation rate. All of these variables are general macroeconomic indicators that proxy for both for a country's ability to weather shocks and crises, but also its quality of macroeconomic management, and its capacity to mobilize resources for major investment projects, such as infrastructure improvement (one of the LPI components).

A final set of indicators relates to a country's human resources. The LPI is partly based on perceptions of the quality of service providers, which is in turn partly determined by human capacity. Variables like the level of basic education, population makeup, and health spending all correlate with labor output and productivity, and so could influence logistics performance through that channel.

Table 3: Fifty most important features in the preferred model.

Rank	Feature Importance Score	Sign of effects	Description of variable1	Description of variable2	Category of variable1	Category of variable2
1	100.00	Positive	International tourism, number of arrivals		Economics: Trade of services	
2	82.12	Negative	Consumer price index (2010 = 100).ZG	High income country	Economics: Stability	Economics: Income & Debt (Dummy variables)
3	54.85	Positive	Mobile cellular subscriptions.P2		Business environment: Infrastructure	
4	50.88	Negative	Prevalence of anemia among non-pregnant women (% of women ages 15-49)	High income country	Human Development: Health	Economics: Income & Debt (Dummy variables)
5	49.85	Negative	Adjusted savings: net forest depletion (% of GNI)	High income country	Environment: Resources	Economics: Income & Debt (Dummy variables)
6	49.76	Negative	Travel services (% of commercial service exports)	High income country	Economics: Trade of services	Economics: Income & Debt (Dummy variables)
7	48.79	Positive	Employment in services, female (% of female employment) (modeled ILO estimate)	High income country	Institutional quality: Gender equality	Economics: Income & Debt (Dummy variables)
8	40.69	Positive	CO2 emissions from solid fuel consumption (% of	Upper middle income country	Environment: Emissions	Economics: Income & Debt

Rank	Feature Importance Score	Sign of effects	Description of variable1	Description of variable2	Category of variable1	Category of variable2
			total)			(Dummy variables)
9	40.69	Negative	Food imports (% of merchandise imports)		Environment: Agricultural production	
10	36.61	Positive	Domestic credit to private sector (% of GDP)		Finance: Assets	
11	34.73	Negative	Food exports (% of merchandise exports)	High income country	Environment: Agricultural production	Economics: Income & Debt (Dummy variables)
12	32.87	Negative	Food imports (% of merchandise imports)	High income country	Environment: Agricultural production	Economics: Income & Debt (Dummy variables)
13	31.43	Positive	ICT goods exports (% of total goods exports)		Business environment: Infrastructure	
14	31.06	Positive	PPP conversion factor, GDP (LCU per international \$)C.RF		Economics: Income & Debt	
15	30.30	Positive	Portfolio investment, net (BoP, current US\$)	Adjusted savings: net forest depletion (% of GNI)	Finance: Investment	Environment: Resources
16	29.86	Positive	Labor force participation rate, total (% of total population ages 15-64) (modeled ILO estimate)	High income country	Human Development: Labor	Economics: Income & Debt (Dummy variables)
17	29.73	Positive	Ores and metals imports (% of merchandise imports)	High income country	Environment: Resources	Economics: Income & Debt

Rank	Feature Importance Score	Sign of effects	Description of variable1	Description of variable2	Category of variable1	Category of variable2
						(Dummy variables)
18	28.26	Positive	Agricultural raw materials imports (% of merchandise imports)	Upper middle income country	Environment: Resources	Economics: Income & Debt (Dummy variables)
19	27.82	Negative	Domestic private health expenditure (% of current health expenditure)	High income country	Human Development: Health	Economics: Income & Debt (Dummy variables)
20	27.71	Negative	Proportion of seats held by women in national parliaments (%)	Prevalence of anemia among non-pregnant women (% of women ages 15-49)	Institutional quality: Gender equality	Human Development: Health
21	26.55	Negative	Permanent cropland (% of land area)	Insurance and financial services (% of service exports, BoP)	Environment: Resources	Economics: Trade of services
22	24.85	Positive	Refugee population by country or territory of asylum		Human Development: Migration	
23	24.21	Negative	Refugee population by country or territory of asylum.OR		Human Development: Migration	
24	24.08	Negative	Armed forces personnel (% of total labor force)	Refugee population by country or territory of asylum	Institutional quality: Political risk	Human Development: Migration
25	23.76	Positive	Cereal yield (kg per hectare)		Environment: Agricultural production	
26	23.32	Positive	Proportion of seats held by	High income country	Institutional	Economics:

Rank	Feature Importance Score	Sign of effects	Description of variable1	Description of variable2	Category of variable1	Category of variable2
			women in national parliaments (%)		quality: Gender equality	Income & Debt (Dummy variables)
27	21.44	Negative	Claims on central government, etc. (% GDP)	Food exports (% of merchandise exports)	Institutional quality: Government spending	Environment: Agricultural production
28	19.86	Negative	Reserves and related items (BoP, current US\$)	School enrollment, primary (gross), gender parity index (GPI)	Finance: Account	Human Development: Education
29	19.00	Positive	Employment in services, female (% of female employment) (modeled ILO estimate)		Institutional quality: Gender equality	
30	18.54	Positive	Adjusted savings: carbon dioxide damage (% of GNI)	Incidence of tuberculosis (per 100,000 people)	Environment: Emissions	Human Development: Health
31	18.18	Positive	Current account balance (% of GDP)	High income country	Finance: Account	Economics: Income & Debt (Dummy variables)
32	17.95	Negative	External balance on goods and services (current US\$)	Merchandise exports to low- and middle-income economies in Sub-Saharan Africa (% of total merchandise exports)	Economics: Trade	Economics: Trade of goods
33	17.95	Negative	Portfolio investment, net (BoP, current US\$)	School enrollment, primary (gross), gender parity index (GPI)	Finance: Investment	Human Development: Education
34	17.64	Positive	Compulsory education,	Current health	Human	Human

Rank	Feature Importance Score	Sign of effects	Description of variable1	Description of variable2	Category of variable1	Category of variable2
			duration (years)	expenditure (% of GDP)	Development: Education	Development: Health
35	17.61	Negative	Time required to enforce a contract (days)	High income country	Business environment: Ease of doing business	Economics: Income & Debt (Dummy variables)
36	17.60	Positive	Statistical performance indicators (SPI): Pillar 1 data use score (scale 0-100)	Merchandise imports from low- and middle-income economies in South Asia (% of total merchandise imports)	Institutional quality: Transparency	Economics: Trade of goods
37	17.51	Negative	Rural population.ZG	High income country	Human Development: Demography	Economics: Income & Debt (Dummy variables)
38	17.15	Negative	Death rate, crude (per 1,000 people)	Export value index (2000 = 100)	Human Development: Health	Economics: Trade
39	17.11	Negative	Foreign direct investment, net outflows (% of GDP)	Statistical performance indicators (SPI): Pillar 3 data products score (scale 0-100)	Finance: Investment	Institutional quality: Transparency
40	16.07	Positive	Ores and metals exports (% of merchandise exports)	Merchandise exports by the reporting economy, residual (% of total merchandise exports)	Environment: Resources	Economics: Trade of goods
41	15.08	Positive	CO2 emissions from solid fuel consumption (% of total)	Merchandise imports from low- and middle-income economies in Latin America & the	Environment: Emissions	Economics: Trade of goods

Rank	Feature Importance Score	Sign of effects	Description of variable1	Description of variable2	Category of variable1	Category of variable2
				Caribbean (% of total merchandise imports)		
42	15.02	Positive	Domestic general government health expenditure (% of general government expenditure)	High income country	Human Development: Health	Economics: Income & Debt (Dummy variables)
43	14.52	Positive	Manufactures exports (% of merchandise exports)		Economics: Trade	
44	13.26	Negative	PPP conversion factor, GDP (LCU per international \$)C.RF	Merchandise exports to low- and middle-income economies in Sub-Saharan Africa (% of total merchandise exports)	Economics: Income & Debt	Economics: Trade of goods
45	13.15	Positive	Prevalence of anemia among non-pregnant women (% of women ages 15-49)	Mortality caused by road traffic injury (per 100,000 population)	Human Development: Health	Business environment: Infrastructure
46	12.95	Negative	Statistical performance indicators (SPI): Pillar 1 data use score (scale 0-100)	GDP deflator (base year varies by country).AD	Institutional quality: Transparency	Economics: Stability
47	12.82	Negative	Total reserves in months of imports	Proportion of seats held by women in national parliaments (%)	Economics: Trade	Institutional quality: Gender equality
48	12.60	Negative	Total reserves in months of imports	PPP conversion factor, GDP (LCU per international \$)C.RF	Economics: Trade	Economics: Income & Debt
49	12.58	Positive	Portfolio investment, net (BoP, current US\$)	Merchandise exports to low- and middle-income economies in South Asia (% of total merchandise	Finance: Investment	Economics: Trade of goods

Rank	Feature Importance Score	Sign of effects	Description of variable1	Description of variable2	Category of variable1	Category of variable2
				exports)		
50	12.46	Positive	Broad money (% of GDP)	Merchandise exports to low- and middle-income economies in East Asia & Pacific (% of total merchandise exports)	Finance: Money supply	Economics: Trade of goods

(Table by authors.)

With these elements in mind, the features that exhibit predictive power in our model are:

1. Business environment: Infrastructure and ease of doing business.
2. Economic: Income and debt, trade in goods and services, and stability.
3. Finance: Banking, investment, assets, money supply, and accounts.
4. Environment: Agricultural production, resources, and emissions.
5. Human development: Health, education, demography, migration, and labor.
6. Institutional quality: Government spending, political risk, transparency, and gender equality

As noted above, the interplay among these factors differs depending on country income status. Because income status is not equally distributed geographically, countries in different locations are also more or less sensitive to these factors compared with other regions. Where high income economies are dominant, as in Europe or North America, factors such as environmental protection, human development, institutional quality, and business environment are the most important for enhancing logistics performance. In low income countries, primarily found in Africa and some parts of Asia, the priority is on infrastructure, trade engagement, institutional quality, migration, resource use, and health (perhaps as a proxy for human capital development). For middle-income countries, which are common in Asia and Latin America, the most important area is trade openness.

6 CONCLUSION

This paper has used ML techniques to extend the LPI, and examine the country-level determinants of performance. After accurately reproducing the original series, we extend its temporal coverage from six years to nineteen, while we expand country coverage by around one-third. The result is a more complete picture of logistics performance around the world and across time periods. Qualitatively, our results are comparable to those in the various World Bank LPI reports, when it comes to the pattern of variation we see across income groups and geographical regions. Where our work adds real substantive value is in showing that improvements in logistics performance are relatively slow over time, and tend to occur at the upper and lower ends of the distribution rather than in the middle.

From the perspective of policymakers, there is value in extending and expanding the LPI dataset in this way. Some countries, such as Indonesia and Morocco (Arvis et al., 2007, 2010, 2012, 2014, 2016, 2018) have used the LPI as the starting point for a large scale rethinking of logistics regulation and service provision. Similarly, the Asia Pacific Economic Cooperation (APEC) group of 21 economies has used the LPI as a key performance indicator for tracking progress on supply chain connectivity. Our approach opens up these options to a wider range of countries. In addition, we enable policymakers to take a longer term perspective both on performance and on the effects of policy changes by extending temporal coverage. Together, these innovations improve the information base available to policymakers looking to improve logistics performance.

Beyond our substantive results, our paper is also proof of concept for the idea that ML is a relatively simple and cost-effective way of extending the coverage of policy-relevant indicators that are costly to collect. There is great scope for future research to apply this finding to other policy indicators in international trade, including indices of policy restrictions that typically

have similarly incomplete country and temporal coverage. There are also many applications outside international trade, including in areas where data collection efforts have been discontinued at an international level, as is the case with the World Bank's Doing Business project. While there were many difficulties associated with that dataset, it was also extensively used by governments, as well as by analysts in empirical and applied work. Our approach shows that independent researchers can, given the data already collected, successfully extend coverage even when international organizations elect not to collect additional data.

From the perspective of logistics researchers, our work opens up the perspective of more detailed investigation of the country-level determinants of performance. ML has made it possible to narrow down the set of variables that most strongly influence LPI scores in particular groups of countries to six categories, despite starting from hundreds of individual series. Future research could examine these factors in greater detail, and seek to identify where the biggest impact lies in individual countries. On the one hand, country case studies and examinations of business processes could help demonstrate the salience of more detailed performance determinants in particular contexts. But at the same time, future statistical analysis could build on our results to more aggressively narrow down the range of retained features, and to focus on testing theories of logistics performance and supply chain operations by examining the importance of theory-based measures for predicting logistics performance using models similar to the ones we have used here. There is also scope to use ML techniques with additional or alternative data sources that provide additional details on aspects not included in the WDI. Such research would be more focused on explanation than prediction, and could perhaps focus on areas like productivity analysis (e.g., using the Penn World Tables).

The categories we have identified as important, and the additional data we have produced and made freely available, may assist future researchers in developing theory-based approaches that help understand logistics and supply chain processes. A primary objective of our work is to enable future research. Theoretical perspectives are an important part of that future development. There is also scope for additional empirics. First, researchers can work with the set of most important features we have identified, to look for additional or unexpected connections between economic and social factors, and logistics performance. Second, our replication code is freely available, so other researchers can use alternative ML techniques, or theory-driven models, to compare prediction accuracy, and refine understanding of global logistics performance.

REFERENCES

- Arvis, J.-F., A. Mustra, J. Panzer, L. Ojala, and T. Naula. 2007. *Connecting to Compete 2007: Trade Logistics in the Global Economy—The Logistics Performance Index and its Indicators*. Washington, D.C.: World Bank.
- Arvis, J.-F., A. Mustra, L. Ojala, B. Shepherd, and D. Saslavsky. 2010. *Connecting to Compete 2010: Trade Logistics in the Global Economy—The Logistics Performance Index and its Indicators*. Washington, D.C.: World Bank.
- Arvis, J.-F., A. Mustra, L. Ojala, B. Shepherd, and D. Saslavsky. 2012. *Connecting to Compete 2012: Trade Logistics in the Global Economy—The Logistics Performance Index and its Indicators*. Washington, D.C.: World Bank.

- Arvis, J.-F., D. Saslavsky, L. Ojala, B. Shepherd, C. Busch, and A. Raj. 2014. *Connecting to Compete 2014: Trade Logistics in the Global Economy—The Logistics Performance Index and its Indicators*. Washington, D.C.: World Bank.
- Arvis, J.-F., D. Saslavsky, L. Ojala, B. Shepherd, C. Busch, A. Raj, and T. Naula. 2016. *Connecting to Compete 2016: Trade Logistics in the Global Economy—The Logistics Performance Index and its Indicators*. Washington, D.C.: World Bank.
- Arvis, J.-F., L. Ojala, C. Wiederer, B. Shepherd, A. Raj, K. Dairabayeva, and T. Kiiski. 2018. *Connecting to Compete 2018: Trade Logistics in the Global Economy—The Logistics Performance Index and its Indicators*. Washington, D.C.: World Bank.
- Box, G., and D. Cox. 1964. “An Analysis of Transformations.” *Journal of the Royal Statistical Society, Series B*.26, 211-252.
- Breiman, L. 2010. “Random Forests.” *Machine Learning*, 45: 5-32.
- Card, D., and A. Krueger. 1995. “Minimum Wages and Employment: A Case Study on the Fast Food Industry in New Jersey and Pennsylvania.” *American Economic Review*, 84(4): 772-793.
- Cortes, C., and V. Vapnick. 1995. “Support Vector Networks.” *Machine Learning*, 20(3): 273-297.
- Cagala, T. 2017. “Improving Data Quality and Closing Data Gaps with Machine Learning”, in “Data needs and Statistics compilation for macroprudential analysis”, IFC Bulletin, 46.
- Cengiz, D., A. Dube, A. Lindner, and B. Zipperer. 2019. “The Effect of Minimum Wages on Low-Wage Jobs.” *Quarterly Journal of Economics*, 134(3): 1405-1454.
- Engblom, J., T. Solakivi, J. Toyli, and L. Ojala. 2012. “Multiple-Method Analysis of Logistics Costs.” *International Journal of Production Economics*, 137(1): 29-35.
- Grundler, K., and N. Potrafke. 2019. “Corruption and Economic Growth: New Empirical Evidence.” *European Journal of Political Economy*, 60: 101 810.
- Hastie, T., R. Tibshirani, and J. Friedman. 2009. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. New York: Springer.
- Hillberry, R., and X. Zhang. 2018. “Policy and Performance in Customs: Evaluating the Trade Facilitation Agreement.” *Review of International Economics*, 26(2): 438-480.
- Hoekman, B., and A. Nicita. 2011. “Trade Policy, Trade Costs, and Developing Country Trade.” *World Development*, 39(12): 2069-2079.
- Honaker, J., G. King, and M. Blackwell. 2012. “Amelia II: A Program for Missing Data.” Working Paper, <https://r.iq.harvard.edu/docs/amelia/amelia.pdf>.
- Kang, M., K. Ichii, J. Kim, Y. Indrawati, J. Park, M. Moon, J. Lim and J. Chun. 2019. “New Gap-Filling Strategies for Long-Period Flux Data Gaps Using a Data-Driven Approach”. *Atmosphere*, 10(10): 568.
- Kuhn, M., and K. Johnson. 2019. *Feature Engineering and Selection: A Practical Approach for Predictive Models*. London: CRC Press

Kursa, M., and W. Rudnicki. 2010. "Feature Selection with the Boruta Package." *Journal of Statistical Software*, 36(11): 1-13.

Magazzino, C., A. Alola, and N. Schneider. 2021. "The trilemma of innovation, logistics performance, and environmental quality in 25 topmost logistics countries: A quantile regression evidence." *Journal of Cleaner Production*: 322, 129050.

Mankiw, G., D. Romer, and D. Weil. 1992. "A Contribution to the Empirics of Economic Growth." *Quarterly Journal of Economics*, 107(2): 407-437.

Memedovic, O., L. Ojala, J.-P. Rodrigue, and T. Naula. 2008. "Fueling the Global Value Chains: What Role for Logistics Capabilities?" *International Journal of Technological Learning, Innovation, and Development*, 1(3): 353-374.

Ojala, L., D. Andersson, and T. Naula. 2008. "Linking to Global Logistics Value Chains: An Imperative for Developing Countries." *International Journal of Technological Learning, Innovation, and Development*, 1(3): 427-450.

Qazi, A. 2022. "Adoption of a probabilistic network model investigating country risk drivers that influence logistics performance indicators." *Environmental Impact Assessment Review*, 94: 106760.

Rantasila, K., and L. Ojala. 2015. "National-Level Logistics Costs: An Overview of Extant Research." *International Journal of Logistics Research and Applications*, 18(4): 313-324.

Rees, B. and A. Haig. 2018. "The Potential for Machine Learning in Investment Research: Evidence from Analyst and Machine-Based Risk Assessments". *Capital Markets: Asset Pricing & Valuation eJournal*.

Sergi, B., V. D'Aleo, S. Konecka, K. Szopik-Depczyńska, I. Dembińska, and G. Ioppolo. 2021. "Competitiveness and the Logistics Performance Index: The ANOVA method application for Africa, Asia, and the EU regions." *Sustainable Cities and Society*, 69: 102845.

Solakivi, T., L. Ojala, H. Lorrentz, J. Toyli, and S. Laari. 2018. "Estimating the Size of the National Logistics Market: A Method to Include both Market-Based Demand and In-House Services." *International Journal of Physical Distribution and Logistics Management*, 48(5): 488-503.

Tibshirani, R. 1996. "Regression Shrinkage and Selection via the LASSO." *Journal of the Royal Statistical Society B*, 58(1): 267-288.

Verbeke, T., and M. De Clercq. 2002. "Environmental quality and economic growth." *SSRN Electronic Journal*. <https://doi.org/10.2139/ssrn.298318>.

Zhang, W., C. Robinson, S. Guhathakurta, V. Garikapati, B. Dilkina, M. Brown, and R. Pendyala. 2018. "Estimating Residential Energy Consumption in Metropolitan Areas: A Microsimulation Approach". *Energy*, 155: 162-173.

Zhu, S., R. Clement, J. McCalmont, C. Davies, and T. Hill. 2022. "Stable Gap-Filling for Longer Eddy Covariance Data Gaps: A Globally Validated Machine-Learning Approach for Carbon Dioxide, Water, and Energy Fluxes". *Agricultural and Forest Meteorology*, 314, 108777.

Zou, H., and T. Hastie. 2005. "Regularization and Variable Selection via the Elastic Net." *Journal of the Royal Statistical Society B*, 67: 301-320.