An application of machine learning regression to feature selection: a study of logistics performance and economic attribute

Suriyan Jomthananachai · Wai Peng Wong · Khai Wah Khaw

Received: 18 November 2021 / Accepted: 29 March 2022 / Published online: 28 April 2022
© The Author(s) 2022

Abstract

This study demonstrates how to profit from up-to-date dynamic economic big data, which contributes to selecting economic attributes that indicate logistics performance as reflected by the Logistics Performance Index (LPI). The analytical technique employs a high degree of productivity in machine learning (ML) for prediction or regression using adequate economic features. The goal of this research is to determine the ideal collection of economic attributes that best characterize a particular anticipated variable for predicting a country’s logistics performance. In addition, several potential ML regression algorithms may be used to optimize prediction accuracy. The feature selection of filter techniques of correlation and principal component analysis (PCA), as well as the embedded technique of LASSO and Elastic-net regression, is utilized. Then, based on the selected features, the ML regression approaches artificial neural network (ANN), multi-layer perceptron (MLP), support vector regression (SVR), random forest regression (RFR), and Ridge regression are used to train and validate the data set. The findings demonstrate that the PCA and Elastic-net feature sets give the closest to adequate performance based on the error measurement criteria. A feature union and intersection procedure of an acceptable feature set are used to make a more precise decision. Finally, the union of feature sets yields the best results. The findings suggest that ML algorithms are capable of assisting in the selection of a proper set of economic factors that indicate a country’s logistics performance. Furthermore, the ANN was shown to be the best effective prediction model in this investigation.

Keywords Feature selection · Machine learning regression · Artificial neural network · Logistics performance index

1 Introduction

The World Bank’s Logistics Performance Index (LPI) is well-known practical information available to policymakers for judging a country’s logistics performance [1]. The LPI, which has been assessing nations’ logistics performance on a biannual basis by analyzing survey data, is perhaps the most important tool to emerge from trade facilitation. It delivers a macroeconomic overview on how policymakers may favorably affect global supply chain capabilities and the performance of relevant businesses representing the efficiency of the clearance process, trade quality, and transportation-related infrastructure [2]. Perhaps it would be advantageous if there was constant and up-to-date information reflecting the country’s logistical efficiency. It may monitor changes in variable data and continuously evaluate trends or predict logistics efficiency to provide policymakers with rapid access to projected logistics performance to improve the country’s logistics and supply chain capabilities.

Nonetheless, prior research has shown that variables such as institutional reforms and resource enhancements significantly accelerate logistics performance. Countries with a low level of corruption and a stable political environment, according to [3], are more likely to have a high level of logistics performance, and improvements in resource supply such as infrastructure, technology, labor, and education related to the country’s competitiveness have a significant positive effect on performance.
Similarly, [1] emphasizes that governance weaknesses and societal instability might reduce performance. However, the aforementioned study variables are static data, and certain components were also collected through the survey.

Since then, the components, particularly the economic component, have demonstrated a significant link between a nation’s logistics performance (expressed by LPI scores) and an economic element, such as the country’s economic development indicator [1, 4]. For example, GDP per capita [1], export and import volume show LPI components have a substantial beneficial influence on expanding international commerce for both import and export [5]. There are, nevertheless, unstudied substantial economic aspects that connect with logistics efficiency.

This study demonstrates how to profit from up-to-date dynamic economic big data, which contributes to the selection of economic attributes that indicate logistics performance as reflected by the LPI. The analytical technique employs a high degree of productivity in the field of machine learning (ML) for prediction or regression using an adequate set of economic feature subsets. Because the accuracy of ML prediction outcomes is reliant not only on the model structure and associated training algorithm but also on the feature space constructed using the initial feature set and feature selection algorithm [6]. In ML applications, feature selection is often employed as a portion of the pre-processing phase to get a subset of features by reducing elements with minimal predictive information [7].

There are two aims in this study: (1) to determine the ideal subset of economic features that best represents a particular anticipated variable for predicting a country's logistics performance and (2) to improve prediction accuracy by employing a set of alternative ML regression algorithms. This research looks into two major research questions: to begin, can ML algorithms assist in selecting the proper subset of economic features that reflect the country's logistics performance? Second, what is the appropriate ML regression approach for predicting logistics performance based on certain economic attributes? This paper’s structure is as follows. Section 2 includes a review of the literature on feature selection and regression machine learning. Section 3 contains methodology, which includes the ML feature selection procedure, data sources and data preparation, data analysis, and parameter setup. Sections 4 and 5 provide the results and discussion, as well as the concluding remark and future work.

## 2 Literature review

### 2.1 Feature selection

The process of selecting a subset of important features, especially variables, for model building is known as feature selection. It is well recognized that a subset of relevant features may be beneficial in improving model performance. With minimum information loss, the feature selection process aims to eliminate duplicate or superfluous features and other features that are closely connected in the data. It is often used to make the model more comprehensible and to improve generality by decreasing variance [8]. The three forms of feature selection strategies are filter, wrapper, and embedded.

To rank all features, filter techniques rely on statistical properties. The challenge of feature selection is viewed as a ranking problem by the majority of filter techniques. These techniques are independent of the ML learning algorithm that will be used with the selected subset [9]. This category comprises mutual information-based, correlation-based, Chi-square test-based and principal component analysis-based techniques. Filter techniques are often used in high-dimensional datasets because of their processing efficiency [8].

Wrapper techniques select feature subsets based on how useful they are to a certain predictor or classifier. Selection is viewed as a search problem in these techniques, with various feature combinations created, evaluated, and compared to other combinations. The search is driven by heuristic intelligent optimization techniques. Simplified methods, such as sequential search, or evolutionary algorithms, such as particle swarm optimization (PSO) or genetic algorithm (GA), that generate local optimal results and are computationally viable, are used to achieve good results [6]. As a selection criterion, these approaches use the performance of the inductive algorithm. They wrap the learning algorithm with feature selection and estimate the advantages of adding or deleting a feature using cross-validation [9]. Wrapper techniques outperform filter methods because each cycle, a new prediction model or learning algorithm evaluates a different feature subset [10].

By embedding feature selection within the model learning, embedded methods provide a trade-off solution between filter techniques and wrapper methods. They return both the learned model and the selected features at the same time [11]. The learning and feature selection components of embedded techniques cannot be separated [12]. Embedded techniques include feature selection into the model training process; for example, starting regularization processes while the model is being trained is a frequent example [6]. Regularization models such as sparse
linear discriminant analysis regularized support vector machine (SVM) and LASSO are the most commonly used embedded approaches [11]. For the instance of an embedded method, the LASSO technique normalizes the parameters of a linear model using an L1-norm penalty, reducing the less correlated coefficients to zero [13]. Many new sparse learning approaches for multi-class classification have been suggested, including L2, 1-norm regularized regression models [11].

Since there are several methods for selecting features, in our work, we used filter techniques to rank the predictors under consideration. Then, using the well-known ML regression, we evaluated several potential subsets of selected features using a particular learning algorithm and ranked those that performed the best. Furthermore, this work compares the embedded techniques of penalized linear regression, which technique is helpful since it allows for simultaneous feature selection and prediction [14].

2.2 ML of regression

ML regression approaches are rapidly being used. The following fundamental techniques are utilized in the literature: artificial neural network (ANN), SVM, and random forest (RF) [15]. These are effective data-driven techniques. These models can also give regression results in a variety of fields, including energy, environmental, waste and pollution, medical, information technology, finance, and business and economics.

One of the most significant artificial intelligence approaches is ANN [16] and extensions of MLP-ANN models are possibly the most popular and frequently utilized in the field of machine learning prediction. To make precise predictions, these models serve as a reliable predictive tool [17], [18] use nonlinear regression techniques to anticipate ozone levels based on main pollutants and meteorological variables. The findings produced from the nonlinear regression techniques ANN were satisfactory, and it had shown its robustness as a helpful tool for evaluating and forecasting air quality situations. To improve simulation for medical operations, a versatile and reliable predictor of body size, shape, and ligament thickness is required. Using clinical data, the ANN can predict patient conditions, and it produced more accurate results than traditional regression analysis methods [19]. It is used to validate the prediction of yearly generation rates of household, commercial, and building and demolition wastes for MLP-ANN. MLP-ANN models demonstrated high prediction accuracy, making them useful for forecasting trash generation rates from various sources and potentially a cost-effective strategy for developing integrated municipal solid waste management systems [20]. In business and economics, researchers suggested ANN regression models cope with the challenge of predicting GDP growth. It is demonstrated that the ANN model can predict GDP growth rates significantly more accurately than a corresponding linear model [21]. Moreover, ANN models are appropriate tools for analyzing economic data such as GDP and GDP per capita since they allow for a trade-off between the capacity to predict these features and model size [21, 22]. MLP-ANN was used by [23] to predict customer quality in e-commerce social networks. By employing word-of-mouth marketing tactics, MLP-ANN produces a strong model for predicting which referrers will attract high-quality referrals (in terms of transaction volume). And the MLP-ANN technique performed better in terms of estimating building costs [16].

The use of SVM for classification and regression issues has grown significantly. Support vector regression (SVR) is a subset of SVM [24]. SVR techniques are typically used for predicting, and the results are usually satisfactory. As an example, consider a large office building’s energy consumption prediction, in which the summer hourly cooling load statistics are utilized as energy consumption data. In terms of accuracy, robustness, and generalization ability, the findings show that the suggested SVR-based technique outperforms generally used methods [25]. [26] utilized the ε-SVR and ν-SVR using linear, polynomial, radial basis function, and sigmoid kernels for predicting software enhancement effort in the information technology area. When prediction accuracies for both types of SVR were compared to those of statistical regressions, they were statistically better than statistical regression. In banking, SVR and enhanced versions of SVR techniques are used to forecast corporate bond losses in the event of default [27]. Overall, their empirical findings indicate that SVR techniques are a promising method for banks to utilize to anticipate loss given failure.

Random forest RF regression (RFR) is a widely used method for analyzing high-dimensional data. Due to poor predictors, its advantages may be reduced in sparse environments, necessitating a pre-estimation dimension reduction (targeting) phase. Nonetheless, this approach is usually useful for predicting and frequently produces an adequate outcome. For example, [28] suggested RFR-based techniques for battery capacity estimate, with experimental findings demonstrating that the proposed technique is capable of evaluating the health statuses of various batteries and promising for online battery capacity estimation. Furthermore, when the RFR performance was compared to the multiple linear regression (MLR) techniques, RFR has a significantly better predictive potential than a typical linear regression model. RFR is regarded as a particularly promising approach for large-scale modeling of groundwater nitrate contamination [29]. In terms of economics, [30] used RFR to estimate GDP at the town scale versus
MLR, with the RFR model achieving considerably greater accuracy.

3 Methodology

3.1 ML feature selection process

Figure 1 depicts the merging of two major approaches in the feature selection process. To begin with, we use correlation and principal component analysis (PCA) approaches to identify the possible feature set. Second, the embedded technique employs the ML of linear regression analysis approaches, namely LASSO and Elastic-net (E-net) regression. In the case of LASSO and E-net, the selected feature can be validated continuously after the model has been trained. Remembering the probable feature set of filter methods, the data associated with those feature sets will be trained using ANN, MLP-ANN, SVR, RFR, and Ridge regression. LASSO or E-net have a high accurate subset selection but lack optimal prediction rates. The dataset of features selected by both regression algorithms is then used to supervise using the suggested ML regression method. The models based on specified sets of features will be validated continuously using the test dataset. Finally, the model’s performance will be evaluated.

3.2 Data sources and data preparation

3.2.1 LPI of World Bank

The World Bank’s Logistics Performance Index (LPI) is a biennially announced indicator for judging a country’s logistics performance. For this research period from 2010 to 2018 (5 periods), there are 134 countries with LPI information available (see Table 11 of the ’Appendix’).

3.2.2 The economic statistics data of S&P global market intelligence

The economic features are derived from the macroeconomic data of S&P Global Market Intelligence’s economic statistics. The economic and demographic statistics data provide the macroeconomic attribute, which includes 52 features divided into five categories: (1) market size and growth (total of 10 features), (2) macroeconomic stability (total of 17 features), (3) personal income and labor (total of 9 features), (4) external sector (total of 13 features), and (5) tax rates (total of 3 features).

3.2.3 Data preparation

According to data preparation, in the study of 5 periods from 2009 to 2018 (biannual average to matching to LPI data), initial 26 features were selected, which are accessible for 100 countries when mapping to the LPI data (see Table 11 of the ’Appendix’). Table 1 displays the 26 possible economic features. The first feature selection is

---

Fig. 1 The ML feature selection process

- All features
  - Feature selection approach
    - Correlation
    - PCA
  - Machine learning regression approach
    - ANN/MLP-ANN of regression
    - Support vector regression
    - Random forest regression
    - Ridge regression
    - LASSO/E-net regression
  - Validation of selected features
motivated by the trade-off between the number of instances and the missing value of country economic data. When the instance must be maximized due to the typically high accuracy even with big datasets [31], the missing value must be minimized. Because missing values might cause bias and reduce analytical efficiency, they should be avoided [32].

For the dataset of 500 instances (5 periods of 100 countries), 70% of the data is utilized as a training set for ML methods, while the remainder is used to test or validate model performance. Furthermore, just one missing item in the dataset is replaced with a mean value based on the country’s economic attribute.

### Table 1 The available 26 economic features

| Category                        | Symbolic and abbreviation of feature | Feature*          |
|---------------------------------|--------------------------------------|-------------------|
| Market Size and Growth (7 of 10)| X1 R_GDP_Gr                           | Real GDP Growth (%) |
|                                 | X2 N_GDP                              | Nominal GDP (SB)  |
|                                 | X3 Pop_Gr                             | Population Growth (%) |
|                                 | X4 GDP_C                              | GDP per Capita ($) |
|                                 | X5 Pri_C_Gr                           | Private Consumption Growth (%) |
|                                 | X6 Gov_C_Gr                           | Government Consumption Growth (%) |
|                                 | X7 GFI_Gr                             | Gross Fixed Investment Growth (%) |
| Macroeconomic Stability (8 of 17)| X8 CAB                               | Current Account Balance (SB) |
|                                 | X9 CP_G                               | Consumer Prices Growth (%) |
|                                 | X10 CP_Idx                            | Consumer Price Index, Average |
|                                 | X11 BB/GDP                            | Budget Balance/ GDP (%) |
|                                 | X12 GDP_D                             | GDP Deflator (2005 = 100) |
|                                 | X13 PD/GDP                            | Public Debt/ GDP (%) |
|                                 | X14 BE/GDP                            | Budget Expenditure/ GDP (%) |
|                                 | X15 GNS_Rt                            | Gross National Savings Rate (%) |
| Personal Income and Labor (1 of 9)| X16 LF                               | Labor Force (M) |
| External Sector (10 of 13)      | X17 CAB/GDP                           | Current Account Balance/ GDP (%) |
|                                 | X18 Exp                               | Goods: Exports (FOB) (SB) |
|                                 | X19 Imp                               | Total Imports (CIF) (SB) |
|                                 | X20 NDIF                              | Net Direct Investment Flows (SB) |
|                                 | X21 FER                               | Foreign Exchange Reserves (SB) |
|                                 | X22 Iwd_DI                            | Inward Direct Investment (SB) |
|                                 | X23 Owd_DI                            | Outward Direct Investment (SB) |
|                                 | X24 TB                                | Trade Balance (SB) |
|                                 | X25 MMI_Rt                            | Money Market Interest Rate (%) |
|                                 | X26 DC_Gr                             | Domestic Credit Growth (%) |

*The feature description is shown in Table 12 of ‘Appendix’*

Source: S&P Global Market Intelligence’s macroeconomic data

### 3.3 Data analysis and parameter setting

#### 3.3.1 Feature selection

##### 3.3.1.1 Correlation

The Pearson correlation coefficient is a measure of the linear correlation between two variables in a collection of variables in statistics. It has a value between $-1$ and $+1$, where $+1$ represents completely positive linear correlation, 0 represents nonlinear correlation, and $-1$ represents completely negative linear correlation [33]. Pearson correlation coefficient $r$ is defined as follows:

$$
r = \frac{1}{n} \sum_{i=1}^{n} (x_i - \overline{x})(y_i - \overline{y}) \div \left( \sqrt{\frac{1}{n} \sum_{i=1}^{n} (x_i - \overline{x})^2} \sqrt{\frac{1}{n} \sum_{i=1}^{n} (y_i - \overline{y})^2} \right),
$$

where $n$ is the sample size; $x_i$ and $y_i$ are the individual sample points indexed with $i$; $\overline{x}$ is the sample mean representing as $\frac{1}{n} \sum_{i=1}^{n} x_i$, and analogously for $\overline{y}$.
The correlation matrix is used to select the subset of relevant features. In this study, the high correlated input with output attributes will be considered [34] stated that it will be weak positive correlation if $r$ value = 0 to 0.25, fair positive correlation if $r$ value = 0.25 to 0.5, good correlation if $r$ value = 0.5 to 0.75, excellent correlation if $r$ value is more than 0.75. It’s also a similar rank to the negative correlation if $r = [0, -1]$. A tool for scientific data analysis of correlation package in Microsoft Excel is utilized to obtain the result. The suggested Excel tool is a cost-effective alternative to costly software, and it is simple to use for basic data analysis.

3.3.1.2 PCA PCA is arguably the most common multivariate statistical technique for reducing data with multiple dimensions, and it is frequently used to reduce well-being indicators to a single index of well-being. PCA was performed in this study using RStudio (version 4.1.1), which was utilized to extract the major components input and output variables after the input and output datasets were centered and scaled. On the one hand, while using PCA, the factor loading criterion is set to 0.3, which means that only variables with absolute factor loading equal to or greater than 0.3 are considered. The varimax factor rotation approach was used to minimize variables with excessive loading on a factor to assure factor interpretability because a loading less than 0.3 is regarded as insignificant [35]. PCA-biplots, on the other hand, were created for feature selection [36]. PC1 is designated as dimension one, whereas PC2 is designated as dimension two. To display and validate, a biplot-based PCA technique is employed, which is a form of statistics graph that may be used to depict the relationships between multiple parameters [37]. Vectors are the projector variables in a PCA-based biplot. Furthermore, the attributes were centered and scaled during the PCA preprocessing.

3.3.1.3 Penalized linear regression Penalized regression models (for example, least absolute shrinkage and selection operator (LASSO) or Elastic-net (E-net)). The high performance of LASSO and E-net is since these models avoid overfitting and minimize model complexity by penalizing the size of coefficients [38]. These models use the simultaneous feature selection–prediction process. [39] introduced LASSO, a famous penalized technique for choosing individual variables, which is based on the following stated model:

$$E(y|X = x) = \beta_0 + \beta'x.$$  (2)

Finding the elements of $\beta$ that equal zero is how the feature selection issue is expressed. Estimates are selected by

$$\min_{\beta} \sum (y_i - \beta_0 - \beta'x_i)^2,$$

subject to $\sum |\beta_j| < t$.

where $y_i$ is the dependent variable, $x_i$ indicates the predictor variables, $\beta_0$ intercept and $\beta'$ show the unknown parameters of the regression equation. Furthermore, $t > 0$ is a tuning parameter that governs the degree of shrinkage applied to the estimations. It is the same as minimizing,

$$\min_{\beta} \sum (y_i - \beta_0 - \beta'x_i)^2 + \lambda \sum |\beta_j|.$$  (4)

For high coefficient estimations, this is the common least squares with a penalty term set by $\lambda$. $\sum |\beta_j|$ is a coefficient vector constraint that yields a sparse solution vector $\beta_j$; as increases $\lambda$, more members of $\beta_j$ become zero. LASSO is able to choose a limit of $N - 1$ features, where $N$ is the sample size [40]. This might be an issue when performing a regression with a limited number of samples but a large number of features [41].

The elastic-net approach attempts to overcome the constraints of the LASSO technique; it is especially effective when there are numerous correlated features [42]. This model is equivalent to minimizing,

$$\min_{\beta} \sum (y_i - \beta_0 - \beta'x_i)^2 + \lambda \sum \left(\alpha |\beta_j| + \frac{1}{2} (1 - \alpha) \|\beta_j\|^2\right).$$  (5)

The values for $\alpha$ of Elastic-net lie between $\alpha = [0, 1]$, if $\alpha = 1$ it is the formulation of the LASSO algorithm.

In this paper, the penalized linear regression analysis was carried out using RStudio in which the glmnet package is employed. In addition, the datasets were centered and scaled, and tenfold cross-validation was performed to produce internally valid performance metrics.

3.3.2 Regression and validation

3.3.2.1 ANN Three critical aspects influence ANN: the unit’s input and activation functions, network architecture, and the weight of each input connection [43]. It is composed of three levels of nodes (neurons), namely the input, hidden, and output layers (Fig. 2a). The data sample is accepted by the input layer, and the target category is returned by the output layer [44]. The neuron, the fundamental unit of these networks, mimics the human counterpart, having dendrites for taking input variables and emitting an output value that may be used as input for other neurons [45]. The neural network’s layers of fundamental processing units are interconnected, with weights assigned to each connection [46], which are changed during the network’s learning process. This step improves not only the interconnections between the layers of neurons, but also the
parameters of the transfer functions between one layer and another, reducing mistakes. Finally, the neural network’s final layer is in charge of combining all of the signals from the preceding layer into a single output signal—the network’s reaction to specific input data [15].

A basic ANN structure is depicted in Fig. 2b, which includes neuron connections, biases assigned to neurons, and weights assigned to connections. Two equations can be used to identify a neuron $k$ [47]:

$$y_k = f(u_k + b_k)$$

and

$$u_k = \sum_{i=1}^{N} w_{ki} x_i,$$

where $x_1, x_2, \ldots, x_n$ are the inputs, $w_{k1}, w_{k2}, \ldots, w_{kn}$ are the neuron weights, $u_k$ is the result of weighted input calculation, $b_k$ is the bias term, $f(\cdot)$ is the activation function, and $y_k$ is the output. There are numerous algorithms that may be used to train a network [43].

The MATLAB 2020b Neural network toolbox was used in this investigation. The ANN was built using the default network and parameters of hidden layer sizes of $1 \times 10$ (one hidden layer with ten nodes). For prediction tasks, a feed-forward ANN with backpropagation learning has been built as a default. Since then, backpropagation has been the most often used supervised algorithm [48]. TRAINLM is a network training function that uses the Levenberg–Marquardt optimization technique to alter the weight and bias variables. TRAINLM is a fast algorithm, although it takes up more memory than other algorithms. LEARNGDM (Gradient descent with momentum weight and bias learning function) is used for error minimization. This function computes the weight change regarding a specific neuron while accounting for the input and error terms, weight and bias, learning rate, and momentum term of the neuron, and is equal to gradient descent with momentum backpropagation. The tangent sigmoid function (TANSIG) is used as a transfer function in the following equation for the input variable $x$ [49]:

$$TANSIG(x) = \frac{2}{1 + e^{-2x}} - 1.$$  \hspace{1cm} (8)

TANSIG is employed in both the hidden and output layers. They calculate the output based on the net input. The values returned by this activation function range from 1 to +1.

3.3.2.2 MLP-ANN The network architecture refers to the structure of connectivity between distinct neurons in ANN. One of the most frequent and useful ANN architectures is the multi-layer perceptron (MLP) network. Each neuron in MLP-ANN is linked to many of its neighbors, with variable weights indicating the relative importance of the individual neuron inputs to the other neurons. MLP is a type of network that belongs to the feed-forward ANN family, and its learning method is backpropagation [50].

In this study, ANN multilayer perceptron is designed based on hidden layer sizes of $10 \times 10$ (ten hidden layers with ten nodes each). The MATLAB 2020b Neural network toolbox which parameter setting similar to the abovementioned ANN was used in this investigation.

3.3.2.3 SVR SVR is an analytical technique used to explore the connection between one or more predictor variables and a real-valued (continuous) dependent variable [51]. When addressing nonlinear problems, SVR uses a kernel function to transfer the nonlinear regression problem to a higher latitude space, allowing it to determine the best hyperplane to separate the sample points [24].
The RF technique is a tree-based ensemble approach that was created to overcome the limitations of the classic classification and regression tree (CART) method \[52\]. RFR is an ensemble learning technique that employs regression algorithms and decision trees \[53\]. The RF regression technique employs regression trees as base learners. \[52\] RF regression technique employs regression trees as base learners. \[53\] The RF regression technique employs regression trees as base learners. \[52\]

\[
\begin{align*}
\text{max} \left[-\frac{1}{2} \sum_{i=1}^{k} \sum_{j=1}^{k} (a_i - a_i^*) (a_j - a_j^*) K(X_i, X_j) \\
- \sum_{i=1}^{k} (a_i - a_i^*) \varepsilon + \sum_{i=1}^{k} (a_i - a_i^*) Y_i \right],
\end{align*}
\]

subject to \( \sum_{i=1}^{k} (a_i - a_i^*) = 0 \), \( 0 \leq (a_i - a_i^*) \leq \xi \) and \( i = 1, 2, \ldots, l \), where \( X_i \) is the sample data; \( l \) is the sample size; \( C \) is the penalty coefficient; \( \varepsilon \) surpasses the penalty size of the error sample; \( K(X_i, X_j) \) is the kernel function to the optimal solution of \( a \).

In RStudio SVR setting as library caret for classification and regression training, library e1071 for a regression machine in which epsilon-regression type is applied, the radial basis kernel is used in predicting method, and cost of constraints violation is set as default (= 1). In the case of a probabilistic regression model, the fitted model of the sigma parameter is the scale parameter of the hypothesized (zero-mean) Laplace distribution calculated by maximum likelihood.

### 3.3.2.4 RFR

The RF technique is a tree-based ensemble approach that was created to overcome the limitations of the classic classification and regression tree (CART) method \[52\]. RFR is an ensemble learning technique that employs regression algorithms and decision trees \[53\]. The RF regression technique employs regression trees as base learners. \[52\] RF regression technique employs regression trees as base learners. \[53\] The RF regression technique employs regression trees as base learners. \[52\]

\[
f_{RF}^C(x) = \frac{1}{C} \sum_{i=1}^{C} T_i(x),
\]

where \( x \) is the vectored input variable, \( C \) is the number of trees, and \( T_i(x) \) is a single regression tree created from a subset of input parameters and the bootstrapped samples.

In RStudio RFR, we use the libraries caret for classification and regression training, and randomForest to implement Breiman’s random forest method for classification and regression. Number of trees to grow, or \( \text{ntree} = 500 \), this should not be set too low to guarantee that every input row is forecasted at least a few times \[54\]. The regression problem’s default parameters are \( \text{mtry} \) or the number of variables randomly selected as candidates at each split, and importance = TRUE.

### 3.3.2.5 Penalized linear regression

One kind of penalized linear regression is ridge regression. This approach has the potential to reduce the magnitude of the regression coefficients, resulting in improved generalizability for predicting unseen data \[53\]. The ridge coefficients are calculated using the following equation:

\[
\min_{\beta} \sum_{i=1}^{N} (y_i - \beta_0 - \beta_j x_j)^2 + \lambda \sum_j \beta_j^2.
\]

Ridge regression has one apparent drawback: it includes all predictors in the final model. It will reduce all of the coefficients toward zero, but not precisely \[55\]. The LASSO and E-net regression are newer alternatives to Ridge regression that help to address this limitation. As previously stated, this study used LASSO and E-net regression for feature selection and prediction, as Ridge regression is only employed to carry out the regression technique.

Similar to LASSO and E-net regression, RStudio of the glmnet package is also employed in Ridge regression with centered and scaled datasets pre-processing and tenfold validation are utilized.

### 3.3.3 Performance evaluation

The mean absolute errors (MAE):

\[
\text{MAE} = \frac{1}{N} \sum_{k=1}^{N} |y(k) - \hat{y}(k)|,
\]

mean absolute percentage errors (MAPE):

\[
\text{MAPE} = \frac{1}{N} \sum_{k=1}^{N} \left| \frac{y(k) - \hat{y}(k)}{y(k)} \right| \times 100\%,
\]

Root-mean-square error (RMSE):

\[
\text{RMSE} = \sqrt{\frac{1}{N} \sum_{k=1}^{N} (y(k) - \hat{y}(k))^2},
\]

Nash – Sutcliffe efficiency coefficient (NSE):

\[
\text{NSE} = 1 - \frac{\sum_{k=1}^{N} (y(k) - \hat{y}(k))^2}{\sum_{k=1}^{N} (y(k) - \overline{y}(k))^2},
\]

and determination coefficient \( (R^2) \):

\[
R^2 = \left[ \frac{\sum_{k=1}^{N} (y(k) - \overline{y}(k)) (\hat{y}(k) - \overline{\hat{y}}(k))^2}{\sum_{k=1}^{N} (y(k) - \overline{y}(k))^2} \right] / \left[ \sum_{k=1}^{N} (y(k) - \overline{y}(k))^2 \right].
\]
will be used to evaluate the performance of the models for the prediction validation method [56–59], where \( N \) is the amount of validation data, \( y \) is the real output of LPI score which \( \bar{y} \) is its average value and \( \hat{y} \) is the prediction of the output of LPI score which \( \bar{\hat{y}} \) is the average of the predicted value.

### 3.4 Analysis time

Using the big O-notation analysis, we determine the theoretical computational time complexity of ML models. The O-notation is used to present any asymptotic computing characteristics by estimating the worst-case computational time. Table 2 displays the time complexity of the study ML models.

### 4 Result and discussion

#### 4.1 Feature selection result

##### 4.1.1 The result of correlation method

Figure 3 depicts the result of the correlation study performed using the Microsoft Excel data analysis tool. When a regression type prediction is used, the input of a correlation model spanning both the dependent and predictor variables is used. We begin by constructing a feature set of

| Prediction model          | Time complexity | Remark                                      |
|---------------------------|-----------------|---------------------------------------------|
| ANN/MLP-ANN               | \( O(1) \)     | Constant [60]                              |
| SVR                       | \( O(n \times S) \approx O(n^2) \) | \( n \) indicates the number of samples, and \( S \) symbolizes the number of support vectors [60] |
| RFR                       | \( O(n \log n \times p \times C) \) | \( n \) indicates the number of samples, \( p \) indicates the number of features, and \( C \) indicates the number of trees in random forest [61] |
| Ridge/ LASSO/ E-net       | \( O(n \times p^2) \) | \( n \) indicates the number of samples, and \( p \) indicates the number of features [62–64] |

![Table 2 Time complexity of models](image)

![Fig. 3 The result of correlation method](image)
predictor factors that have a direct good or outstanding correlation to the dependent variable of LPI \((r \geq 0.5\), as shown in the red border in Fig. 3) (namely set A). A set A’s predictor variables are \(X_4\) (GDP_C; \(r = 0.75\)), \(X_18\) (Exp; \(r = 0.52\)), and \(X_19\) (Imp; \(r = 0.5\)), for a total of three features. Furthermore, the predictor variables that have a strong or outstanding correlation with a member of set A \((r \geq 0.5\), as shown in the yellow border in Fig. 3) are taken into account and subsequently extended to a member of set B. The additional predictor variables of a set A into set B include \(X_2\) (N_GDP; \(r = 0.83\) (with Exp) and \(r = 0.93\) (with Imp)), \(X_16\) (LF; \(r = 0.62\) (with Exp) and \(r = 0.57\) (with Imp)), \(X_21\) (FER; \(r = 0.6\) (with Exp) and \(r = 0.52\) (with Imp)), \(X_{22}\) (Iwd_DI; \(r = 0.54\) (with Exp) and \(r = 0.59\) (with Imp)), and \(X_{23}\) (Owd_DI; \(r = -0.56\) (with Exp) and \(r = -0.6\) (with Imp)) The total number of features in set A and B.

![Fig. 4 The percentage of variation based on principle component](image_url)

| Economic feature | LASSO | E-net_0.1 | E-net_0.25 | E-net_0.5 | E-net_0.75 | E-net_0.9 |
|------------------|-------|-----------|------------|-----------|------------|-----------|
|                  |       |           |            |           |            |           |
| \(X_1\) R_GDP_Gr | –     | ~ 0 0.64  | –          | –         | –          | –         |
| \(X_2\) N_GDP   | – 0.09| 26.49     | ~ 0 11.2   | ~ 0 13.9  | ~ 0 19.6   | ~ 0 27.2  |
| \(X_3\) Pop_Gr  | – 0.08| 23.66     | ~ 0 27.9   | ~ 0 25.7  | ~ 0 24.6   | ~ 0 24.1  |
| \(X_4\) GDP_C   | 0.33  | 100       | 0.27 100   | 0.3 100   | 0.32 100   | 0.33 100  |
| \(X_5\) Pri_C_Gr| –     | –         | –          | –         | –          | –         |
| \(X_6\) Gov_C_Gr| –     | –         | –          | –         | –          | –         |
| \(X_7\) GFI_Gr  | –     | –         | –          | –         | –          | –         |
| \(X_8\) CAB     | –     | –         | –          | –         | –          | –         |
| \(X_9\) CP_G    | – 0.03| 7.7       | ~ 0 11.5   | ~ 0 9.3   | ~ 0 8.49   | ~ 0 8.18  |
| \(X_{10}\) CP_Idx| –     | –         | –          | –         | –          | –         |
| \(X_{11}\) BB/GDP| –     | –         | –          | –         | –          | –         |
| \(X_{12}\) GDP_D| –     | ~ 0 0.01  | –          | –         | –          | –         |
| \(X_{13}\) PD/GDP| 0.05  | 13.91     | 0.06 22    | 0.05 17.9 | 0.05 15.6  | 0.05 14.7 |
| \(X_{14}\) BE/GDP| –     | ~ 0 0.04  | –          | –         | –          | –         |
| \(X_{15}\) GNS_Rt| –     | 0.02 5.56 | 0.01 3.55  | ~ 0 1.1   | ~ 0 0.29   | ~ 0       |
| \(X_{16}\) LF   | –     | –         | –          | –         | –          | –         |
| \(X_{17}\) CAB/GDP| –     | 0.01 4.1  | –          | –         | –          | –         |
| \(X_{18}\) Exp  | – 0.05| 14.69     | 0.08 29.7  | 0.08 27.1 | 0.08 24.1  | 0.06 18   |
| \(X_{19}\) Imp  | 0.18  | 53.65     | 0.09 33.1  | 0.1 32.9  | 0.12 39    | 0.17 51.5 |
| \(X_{20}\) NDIF | –     | ~ 0 0.2   | –          | –         | –          | –         |
| \(X_{21}\) FER  | –     | ~ 0 0.88  | –          | –         | –          | –         |
| \(X_{22}\) Iwd_DI| –     | –         | –          | –         | –          | –         |
| \(X_{23}\) Owd_DI| –     | –         | –          | –         | –          | –         |
| \(X_{24}\) TB   | –     | –         | –          | –         | –          | –         |
| \(X_{25}\) MMI_Rt| – 0.04| 11.18     | ~ 0 19.4   | ~ 0 15.4  | ~ 0 12.9   | ~ 0 11.8  |
| \(X_{26}\) DC_Gr| – 0.01| 1.53      | ~ 0 1.99   | ~ 0 1.52  | ~ 0 1.46   | ~ 0 1.76  |
| **Total selected**| **9** | **15**    | **10**     | **10**    | **10**     | **9**     |

Table 3 The results of penalized linear regression method
B is $3 + 5 = 8$. Table 4 shows a summary of the subset of features selected using the correlation approach.

4.1.2 The result of PCA method

The PCA result is generated by RStudio in which the dependent variable and predictor variables are used as PCA model input to select the feature for the regression purpose [56]. Figure 4 depicts the proportion of variance of each principal component based on the overall result (only PC1 to PC10 out of a total of 27 PCs). The first and second principal components (i.e., PC1 and PC2) exhibited 34.9 percent variance, whereas PC1 through PC10 may encompass roughly 80 percent of the variation (81.41 percent). Furthermore, when PC1 to PC3 were evaluated, the variation was 46.33 percent, which is more than half of the range of PC1 to PC10. When PC1 to PC5 is considered half of the 10 PCs from PC1 to PC10, the variance is 62.12 percent. To construct a collection of selected features, we examined the attribute that provides a high loading on a factor (equal to or greater than 0.3). The detected attributes in PC1 to PC3 (46.33 percent variance), PC1 to PC5 (62.12 percent variation), and PC1 to PC10 (81.41 percent variation) are allocated to feature sets C, D, and E, respectively. Set C such as $X_1$ (R_GDP_Gr), $X_2$ (R_GDP_Gr), $X_4$ (GDP_C), $X_5$ (Pri_C_Gr), $X_{11}$ (BB/GDP), $X_{14}$ (BE/GDP), $X_{15}$ (GNS_Rt), $X_{18}$ (Exp), $X_{19}$ (Imp), $X_{22}$ (Iwd_DI), and $X_{23}$ (Owd_DI), 11 features by total. Set D of 16 features is set C plus 5 features which are $X_8$ (CAB), $X_9$ (CP_G), $X_{12}$ (GDP_D), $X_{21}$ (FER), and $X_{24}$ (TB). And set E has a total of 24 from the overall 26 features that exclude $X_{16}$ (LF) and $X_{17}$ (CAB/GDP).

Furthermore, the feature selection while constructing a PCA-biplot is illustrated in Fig. 5, with the selected features represented by blue vectors. The selection is motivated by the interrelationships of each feature to LPI. The direction of the feature vector reflects the positive or negative correlations [65]. When a feature has a comparable direction that is the smallest in the angle of the vector relative to the LPI vector, it indicates the strongest positive correlations, while the opposite direction indicates negative correlations. Vectors close to perpendicular to the LPI vector, on the other hand, are weakly correlated (orange vectors in Fig. 5.) Based on the PCA-biplot, the selected
features of set F, i.e., $X_2$ (R_GDP_Gr), $X_4$ (GDP_C), $X_{16}$ (LF), $X_{18}$ (Exp), $X_{19}$ (Imp), $X_{21}$ (FER), $X_{22}$ (Iwd_DI), $X_{23}$ (Owd_DI), and $X_{25}$ (MMI_Rt), which are 9 features in total. The summary of the subset of features selected using the PCA method is shown in Table 4.

### 4.1.3 The result of penalized linear regression method

Table 3 displays the results of the LASSO and E-net penalized linear regression methods. Using RStudio, the model for LASSO regression of Eq. (4) has been reducing the predictor parameters from 26 to 10, which offer various interception values and parameter significance.

The 9 features selected by LASSO (set G) include $X_2$ (R_GDP_Gr), $X_3$ (Pop_Gr), $X_4$ (GDP_C), $X_9$ (CP_G), $X_{13}$ (PD/GDP), $X_{18}$ (Exp), $X_{19}$ (Imp), $X_{25}$ (MMI_Rt), and $X_{26}$ (DC_Gr). For Elastic-net related to Eq. (5), in this study, we vary the $\alpha$ as 0.1, 0.25, 0.5, 0.75, and 0.9. The results of feature selection from RStudio which provides the preferred parameters that the model does not shrink are displayed in Table 3. It was found that when $\alpha = 0.9$ the set of selected features is similar to the results of LASSO. When $\alpha$ is assigned with the value of 0.25, 0.5, and 0.75, they provide the likely set of 10 selected features (set H). Finally, for $\alpha = 0.1$, we found that 15 parameters were non-shrink (set I). Set H contains all attributes of set G which $X_{15}$ (GNS_Rt) is added. And set I comprised all elements of set G with $X_1$ (R_GDP_Gr), $X_{14}$ (BE/GDP), $X_{17}$ (CAB/GDP), $X_{20}$ (NDIF), and $X_{21}$ (FER) combined. The summary of the subset of features selected using penalized linear regression method is shown in Table 4.
### Table 5: Summary of ML regression model performance evaluation

| Criteria | Prediction model | All feature | Selected feature | Principal Component Analysis | Penalized Linear Regression | Average |
|----------|------------------|-------------|------------------|-------------------------------|----------------------------|---------|
|          |                  |             |                  | Correlation                   |                            |         |
|          |                  |             |                  | Cor (set A)                  | Cor_related (set B)        |         |
| MAE      | ANN              | 0.1595      | 0.1725           | 0.1488                        | 0.1413                     | 0.1494  |
|          | MLP-ANN          | 0.1727      | 0.1901           | 0.1411                        | 0.1483                     | 0.1543  |
|          | SVR              | 0.1433      | 0.1739           | 0.1712                        | 0.1506                     | 0.1555  |
|          | RFR              | 0.1693      | 0.1781           | 0.1609                        | 0.1678                     | 0.1695  |
|          | Ridge            | 0.2746      | 0.2535           | 0.2379                        | 0.2341                     | 0.2375  |
|          | LASSO            | –           | –                | –                             | –                          | 0.2463  |
|          | E-net_0.5        | –           | –                | –                             | –                          | 0.2488  |
|          | E-net_0.1        | –           | –                | –                             | –                          | 0.2536  |
| MAPE     | ANN              | 5.4742      | 5.8787           | 5.0874                        | 4.7489                     | 5.1049  |
|          | MLP-ANN          | 5.806       | 6.4713           | 4.9371                        | 5.0301                     | 5.0214  |
|          | SVR              | 4.919       | 6.0464           | 6.0184                        | 5.2344                     | 5.3963  |
|          | RFR              | 5.8853      | 6.1349           | 5.5814                        | 5.7829                     | 5.8821  |
|          | Ridge            | 9.3567      | 8.5306           | 8.0882                        | 8.0178                     | 8.1623  |
|          | LASSO            | –           | –                | –                             | –                          | 8.5189  |
|          | E-net_0.5        | –           | –                | –                             | –                          | 8.6205  |
|          | E-net_0.1        | –           | –                | –                             | –                          | 8.8439  |
| RMSE     | ANN              | 0.1963      | 0.2127           | 0.19                           | 0.1868                     | 0.2005  |
|          | MLP-ANN          | 0.2332      | 0.2325           | 0.1937                        | 0.1986                     | 0.215   |
|          | SVR              | 0.1925      | 0.217            | 0.2188                        | 0.1978                     | 0.2027  |
|          | RFR              | 0.211       | 0.2194           | 0.2033                        | 0.2069                     | 0.2117  |
|          | Ridge            | 0.3353      | 0.3067           | 0.2902                        | 0.2906                     | 0.2935  |
|          | LASSO            | –           | –                | –                             | –                          | 0.3094  |
|          | E-net_0.5        | –           | –                | –                             | –                          | 0.3112  |
|          | E-net_0.1        | –           | –                | –                             | –                          | 0.3162  |
| NSE      | ANN              | 0.8724      | 0.8502           | 0.8804                        | 0.8844                     | 0.8668  |
|          | MLP-ANN          | 0.8199      | 0.821            | 0.8758                        | 0.8693                     | 0.8469  |
|          | SVR              | 0.8812      | 0.8491           | 0.8467                        | 0.8746                     | 0.8683  |
|          | RFR              | 0.8574      | 0.8457           | 0.8675                        | 0.8629                     | 0.8564  |
|          | Ridge            | 0.6208      | 0.6827           | 0.7159                        | 0.7152                     | 0.7095  |
|          | LASSO            | –           | –                | –                             | –                          | 0.6985  |
|          | E-net_0.5        | –           | –                | –                             | –                          | 0.6948  |
|          | E-net_0.1        | –           | –                | –                             | –                          | 0.685   |
4.2 Regression and validation result

According to the subset of selected features (set A to set I), 70% of datasets are trained utilizing identified ML methods such as ANN, MLP-ANN, SVR, RFR, and Ridge. Furthermore, the LASSO and E-net models constantly train their datasets using only the selected feature set that they have been trained on. Furthermore, the entire collection of all features is compared. The test sets are then utilized to validate the model. The validation findings are represented by a performance evaluator or criterion such as MAE, MAPE, RMSE, NSE, and $R^2$.

The summary of performance evaluation findings is given in Table 5 and Fig. 6 for easier comparison. SVR of a feature set I has the greatest MAE performance (*0.1349) (minimum value), while this model with a comparable set has the best MAPE performance as the lowest value (*4.6387). For RMSE, ANN with a comparable feature set I achieves the best performance (*0.1808) (minimum value). NSE values range between $-\infty$ and 1 to indicate the prediction performance, whereas NSE values close to 1 indicate best prediction performance [57] that the SVR with a comparable feature set I achieves the most excellent performance (*0.8938). Moreover, SVR of a feature set I has the highest $R^2$ performance (*0.8964) (maximum value). And if $R^2 > 0.8$, there is a strong correlation between actual values and model estimations [16].

When ML is examined using the average value based on each model for all feature sets, the best performing model is ANN, with average MAE, MAPE, RMSE, and NSE values of **0.1525, **5.1775, **0.1985, and **0.8691, respectively. When compared to ANN, MLP-ANN, SVR, and RFR show satisfactory performance for all criteria. When compared to each admissible model, the performance of all penalized linear regression methods (Ridge, LASSO, and E-net) is a lesser amount of performance. We determined the average performance of admissible ML models that omit penalized linear regression approaches when we focused on the set of selected features. Set H has the greatest performance for MAE, MAPE, and $R^2$ (**0.1497, **5.1775, and **0.8803, respectively), whereas set C has the best performance for RMSE and NSE (**0.1975 and **0.8728, respectively).

Because of a different perspective or set of criteria produces a different set of optimal results, hence, in this study, we reprocessed using the feature union and intersection operations. To reorganize the acquired feature subsets, a feature union and intersection procedure are presented [66]. Table 6 displays the new feature set based on sets C, H, and I. However, certain feature sets are close enough that we can merge them while reprocessing, such as C ∪ I & C ∪ H ∪ I and C ∩ H & C ∩ H ∩ I. Or like...
the previous set that we do not reprocess such as H & H ∩ I and I & H ∪ I. Table 7 shows a summary of the ML regression model performance evaluation of the reprocess of a new feature set that has been merged with the parent sets.

For MAE, the best performance is obtained by ANN of a feature set C ∪ H (*0.1318), and this model with a comparable set performs best in MAPE (*4.4512), RMSE (*0.1723), NSE (*0.9017), and R² (*0.9033). When ML is examined using the average value based on each model for all new and parent feature sets, the top-performing model is

![Fig. 6 Summary of ML regression model performance evaluation](image-url)
ANN, with average MAE, MAPE, RMSE, NSE, and $R^2$ values of **0.1412, **4.7946, **0.1874, **0.8834 and **0.887, respectively, while SVR is second. [67] noted that when the connections between parameters become noninvertible (due to a large number of predictor variables), the input and output configurations used in ANN have a major influence on the accuracy. Furthermore, ANN outperforms linear models in terms of accuracy (where the number of important predictor variables is restricted) [68].

Moreover, concentrating on the set of selected features, we determined the average performance of the four suitable ML models described above. The best performance is exhibited in set C, which offered the best for all performance metrics, namely MAE, MAPE, RMSE, NSE, and $R^2$ as **0.1463, **5.0066, **0.1908, **0.8811, and **0.888, respectively. The members of set C affect the accuracy of LPI when predicting the following features: X1 (R_GDP_Gr), X2 (R_GDP_Gr), X3 (Pop_Gr), X4 (GDP_C), X5 (Pri_C_Gr), X6 (BB/GDP), X7 (PD/GDP), X8 (BE/GDP), X9 (GNS_Rt), X10 (Exp), X11 (Imp), X12 (Iwd_DI), X13 (Owd_DI), X14 (MMI_Rt), and X15 (DC_Gr), total 16 features. As previously stated, the instance must be maximized due to the normally high accuracy even with large datasets, and the number of missing values must be minimized because it can be involved in reducing bias and improving the efficiency of the analysis; furthermore, a limitation on the number of features may support this. Taking into account the parent features of C and H, this limits the number of features to 11 and 10, respectively. Those features may be used as an alternative since they give an adequate performance (closest to the best). The other explanation, set C, is supplied by the PCA method, which is one of the algorithms with higher performance than the other algorithms, resulting in many studies. While the penalized linear of E-net regression provides set H, this regression technique

| Economic feature | C | H | I | Union | Intersection |
|------------------|---|---|---|-------|-------------|
| X1 R_GDP_Gr      | o | o | o | o     | o           |
| X2 N_GDP         | o | o | o | o     | o           |
| X3 Pop_Gr        | o | o | o | o     | o           |
| X4 GDP_C         | o | o | o | o     | o           |
| X5 Pri_C_Gr      | o | o | o | o     | o           |
| X6 Gov_C_Gr      | o | o | o | o     | o           |
| X7 GFI_Gr        | o | o | o | o     | o           |
| X8 CAB           | o | o | o | o     | o           |
| X9 CP_G          | o | o | o | o     | o           |
| X10 CP_Idx       | o | o | o | o     | o           |
| X11 BB/GDP       | o | o | o | o     | o           |
| X12 GDP_D        | o | o | o | o     | o           |
| X13 PD/GDP       | o | o | o | o     | o           |
| X14 BE/GDP       | o | o | o | o     | o           |
| X15 GNS_Rt       | o | o | o | o     | o           |
| X16 LF           | o | o | o | o     | o           |
| X17 CAB/GDP      | o | o | o | o     | o           |
| X18 Exp          | o | o | o | o     | o           |
| X19 Imp          | o | o | o | o     | o           |
| X20 NDIF         | o | o | o | o     | o           |
| X21 FER          | o | o | o | o     | o           |
| X22 Iwd_DI       | o | o | o | o     | o           |
| X23 Owd_DI       | o | o | o | o     | o           |
| X24 TB           | o | o | o | o     | o           |
| X25 MMI_Rt       | o | o | o | o     | o           |
| X26 DC_Gr        | o | o | o | o     | o           |

Total selected feature: 11 10 15 16 19 15 19 5 7 10 5
has accurate subset selection but lacks optimum prediction rates.

The four acceptable ML models do not differentiate from each other for the best performance set of C ∪ H and the parent set of C and H based on the errors shown by the boxplots (Fig. 7), and the RFR is circled as the biggest error values. However, it is discovered that the extreme error levels of all models are nearly the same. Furthermore, Taylor diagrams were created for the evaluation of the acquired results, and they allow for the determination of the correctness of the developed models in many areas [57, 58]. Figure 8 clearly shows that the prediction results of the set of C ∪ H and the parent set of C and H based on the four acceptable ML models are close to the observations. What is interesting about the findings shown in Fig. 8c is that the ANN model outperformed the other models for the set of C ∪ H that gives the shortest distance to the observation. The statistical significance of the acquired data was examined using the Kruskal–Wallis test in this study, as well as an analysis of whether the predicted and observed or logistics performance index distributions, were consistent [57, 69]. H0 denotes a hypothesis based on the statistically significant difference between mean predicted and observed LPI values. Table 8 reveals that the H0 hypothesis was rejected (P value > 0.05) in all C, H, and C ∪ H set predictions; in other words, there is no significant difference between predicted and observed averages. H0 hypotheses were rejected, similarly, this indicates all of the ANN, MLP-ANN, SVR, and RFR models produce more accurate results. This suggests that the pre-processing of data preparation and feature selection had a statistically significant beneficial influence on ML predictions.

Table 7 Summary of ML regression model performance evaluation of novel feature set

| Criteria | Prediction model | C | H & H ∩ I | I & H ∪ I | C ∪ H | C ∪ I & C ∪ H ∩ I | C ∩ H & C ∩ H ∩ I | C ∩ I | Average |
|----------|------------------|---|-----------|-----------|------|-------------------|-------------------|------|---------|
| MAE      | ANN              | 0.1413 | 0.1511 | 0.1395 | *0.1318 | 0.1398 | 0.1379 | 0.1468 | **0.1412 |
|          | MLP-ANN          | 0.1483 | 0.1446 | 0.1741 | 0.148 | 0.1519 | 0.1573 | 0.1769 | 0.1573 |
|          | SVR              | 0.1506 | 0.1434 | 0.1349 | 0.1442 | 0.1384 | 0.156 | 0.1493 | 0.1450 |
|          | RFR              | 0.1678 | 0.1595 | 0.1586 | 0.163 | 0.1613 | 0.1681 | 0.1663 | 0.1635 |
| Average  |                  | 0.152 | 0.1497 | 0.1518 | ***0.1463 | 0.1479 | 0.1548 | 0.1598 |
| MAPE     | ANN              | 4.7489 | 5.0782 | 4.7913 | *4.4512 | 4.8051 | 4.6502 | 5.037 | **4.7946 |
|          | MLP-ANN          | 5.0301 | 4.9631 | 5.9207 | 4.9673 | 5.0641 | 5.3018 | 6.1291 | 5.3395 |
|          | SVR              | 5.2344 | 5.0204 | 4.6387 | 4.9568 | 4.736 | 5.4587 | 5.1901 | 5.0336 |
|          | RFR              | 5.7829 | 5.5189 | 5.4903 | 5.6511 | 5.5798 | 5.8296 | 5.7612 | 5.6591 |
| Average  |                  | 5.1991 | 5.1452 | 5.2103 | ***5.0066 | 5.0463 | 5.3101 | 5.5294 |
| RMSE     | ANN              | 0.1868 | 0.2027 | 0.1808 | *0.1723 | 0.1972 | 0.1833 | 0.1886 | **0.1874 |
|          | MLP-ANN          | 0.1986 | 0.2011 | 0.2382 | 0.1938 | 0.2063 | 0.227 | 0.2346 | 0.2142 |
|          | SVR              | 0.1978 | 0.1917 | 0.1821 | 0.1948 | 0.1886 | 0.2014 | 0.1924 | 0.1927 |
|          | RFR              | 0.2069 | 0.1972 | 0.1967 | 0.2024 | 0.2003 | 0.2086 | 0.2034 | 0.2022 |
| Average  |                  | 0.1975 | 0.1982 | 0.1995 | ***0.1908 | 0.1981 | 0.2051 | 0.2048 |
| NSE      | ANN              | 0.8844 | 0.8639 | 0.8917 | *0.9017 | 0.8712 | 0.8887 | 0.8822 | **0.8834 |
|          | MLP-ANN          | 0.8693 | 0.866 | 0.812 | 0.8756 | 0.859 | 0.8293 | 0.8177 | 0.847 |
|          | SVR              | 0.8746 | 0.8822 | 0.8938 | 0.8784 | 0.886 | 0.87 | 0.8814 | 0.8809 |
|          | RFR              | 0.8629 | 0.8755 | 0.876 | 0.8687 | 0.8714 | 0.8606 | 0.8674 | 0.8689 |
| Average  |                  | 0.8728 | 0.8719 | 0.8684 | ***0.8811 | 0.8719 | 0.8622 | 0.8622 |
| R²       | ANN              | 0.8862 | 0.8765 | 0.8935 | *0.9033 | 0.8741 | 0.8906 | 0.8825 | **0.887 |
|          | MLP-ANN          | 0.8695 | 0.873 | 0.8141 | 0.8854 | 0.862 | 0.8396 | 0.8181 | 0.8591 |
|          | SVR              | 0.8793 | 0.8868 | 0.8964 | 0.881 | 0.888 | 0.8749 | 0.8846 | 0.8809 |
|          | RFR              | 0.8724 | 0.885 | 0.8881 | 0.8823 | 0.8838 | 0.8668 | 0.8772 | 0.8753 |
| Average  |                  | 0.8769 | 0.8803 | 0.8730 | ***0.888 | 0.877 | 0.868 | 0.8656 |
4.3 Discussion

Finally, as shown in Table 10, we discussed the finding outcomes based on both feature selection techniques of filter and embedding method which is focused on the suggested statistical property and ML algorithm. The discussion describes the advantages and disadvantages of models that influence the findings of this study.

To get good results, effective wrapper strategies, such as sequential search, or evolutionary algorithms, such as Particle Swarm Optimization (PSO) or Genetic Algorithm (GA), provide local optimum solutions and are computationally viable, are utilized. Because of the potential of overfitting and computationally costly [72], wrappers have a significant disadvantage, particularly in terms of computational inefficiency, which becomes more obvious as the feature space develops. The wrapper technique is thus eliminated from this analysis, although it will be significant in future studies.

5 Concluding remark and future work

In conclusion, the current study illustrates an application of machine learning regression to feature selection. In this study, we looked at the impact of logistics performance utilizing the World Bank’s LPI and the economic attributes of S&P Global Market Intelligence’s macroeconomic data source. The 500 case samples ranged from 2009 to 2018, with an initial set of 26 economic features accessible. Furthermore, the number of instances (maximize) and the missing value of nation economic data have been traded-off in the first feature selection (minimize). The filter methods of correlation and PCA are employed in the suggested feature selection procedure. The ML regression algorithms ANN, MLP-ANN, SVR, RFR, and Ridge are then utilized to train and verify the data set depending on the selected feature. To select the feature, the embedded technique of penalized linear regression of LASSO and E-net is also used, followed by continuous training and validation of the dataset. In feature selection, the proposed ML regression uses a subset of penalized linear regression features to train and validate the dataset. According to the results of the model’s performance based on the MAE, MAPE, RMSE, NSE, and R² criteria, the feature set of PCA (set C), and E-net (set H and I) offer the most closely acceptable performance.

Then, using parent sets (C, H, and I), a feature union and intersection operation are performed. Finally, the set of C U H (a total of 16 features) performs the best across all criteria. Furthermore, the findings may address the study issue that ML algorithms can select the appropriate set of economic features that reflect the country’s logistics performance. In response to the question: what is the best ML regression technique for predicting logistics performance based on selected economic attributes? The findings indicate that ANN is the most effective model for prediction in this study. Furthermore, we note that features C and H limit the number of features to 11 and 10, respectively. Those features may be used as an alternative since they give an adequate performance (near to the best) when it is necessary to maximize the instance and reduce the missing data of the dataset.

Furthermore, in a future study, the focus may be on utilizing more diverse feature dimensions integrated with economic attributes. The unique elements connected to the megatrend, such as the carbon emissions rate, the cost and consumption rate of fuel and renewable energy, the e-commerce market size, and growth, may reflect on logistics performance in the new era of a global supply
Fig. 8 Taylor diagrams of the best performance set of C ∪ H and the parent set of C and H (a) set C (b) set H (c) set C ∪ H.

Table 8 Result of Kruskal–Wallis test (95% significance level)

| Prediction model | C P-value | *H0 | H P-value | *H0 | C ∪ H P-value | *H0 |
|------------------|----------|-----|-----------|-----|---------------|-----|
| ANN              | 0.9474   | Reject | 0.8489   | Reject | 0.9766       | Reject |
| MLP-ANN          | 0.9532   | Reject | 0.8719   | Reject | 0.8374       | Reject |
| SVR              | 0.7922   | Reject | 0.6452   | Reject | 0.7434       | Reject |
| RFR              | 0.5556   | Reject | 0.4809   | Reject | 0.451        | Reject |

*H0 denotes a hypothesis based on the statistical significance discrepancy between mean predicted and observed LPI values.
Table 9 | Analysis time of ML algorithms (second)

| Prediction model | C | Train | Validate | C | H | Train | Validate | C | H | Train | Validate |
|------------------|---|-------|----------|---|---|-------|----------|---|---|-------|----------|
| ANN              | < 1 | < 1   | < 1      | < 1 | < 1 | < 1   | < 1       | < 1 | < 1 | < 1   | < 1       |
| MLP-ANN          | 8.3 | < 1   | 7.9      | < 1 | 8.2 | < 1   | < 1       | 0.089| 0.014| 0.517 | 0.018     |
| SVR              | 0.087| 0.016 | 0.091    | 0.012| 0.089| 0.014 |           |     |     |       |           |
| RFR              | 0.475| 0.016 | 0.453    | 0.014| 0.517| 0.018 |           |     |     |       |           |

Table 10 | Finding discussion based on the study model

| Feature selection strategy | Approach | Advantage | Disadvantage | Finding discussion |
|---------------------------|----------|-----------|--------------|--------------------|
| Filter                    | Correlation | For a variety of features, effective and quick [70] | Some potentially beneficial features may be included in the weak correlation subset and hence cannot be exploited effectively [70] | The less potentially beneficial feature set of A and B results in a strong correlation of the dependent and predictor variables. Because of the model disadvantage for the feature selection approach, this model may be less successful for this research data set |
| PCA                       | Low complexity and the absence of features [71] | Lower computational cost and time [72] | Good generalization ability [72] | The PCA approach provides the potentially important feature set of Set C. That it has a low absence of valuable features is one of its advantages. However, the difference in criterion, i.e., the factor loading and percentage of variation, is a major change that will modify the result of the feature set based on its drawback |
| ANN                       | Better at identifying very complex patterns and making accurate predictions [74] | Model networks are used to approximate or estimate functions that generally need a considerable quantity of training data [74] | Because the pattern of the LPI and economic attributes of various countries is extremely complex because the research data contains different economic backgrounds based on country, i.e., underdeveloped, developing, and developed countries. According to its advantages, ANN is the most effective model for prediction in these characteristics of the data pattern |
| MLP-ANN                   | When the amount of input features and data complexity is substantially bigger, utilize non-linear correlations among variables to produce more accurate predictions [75] | Slow convergence velocity, resulting in a preference for the local optimum [76] | Neural networks require a large number of parameters and are unable to see the learning process among them [76] | The result is difficult to comprehend, and the learning time is excessive [77] | MLP-ANN may benefit from data complexity as well, but its performance is lower than that of ANN because of the limitation of modification network design in this investigation (layer sizes of 1 x 10 fixed), which may prevent it from observing the optimum effective network in this study |
| Feature selection strategy | Approach | Advantage | Disadvantage | Finding discussion |
|----------------------------|----------|-----------|--------------|--------------------|
| **SVR**                    | Statistical property | Performs effectively with a linear Kernel that can be linear or nonlinear depending on the Kernel [77] | Nonlinear problems are difficult to solve, and it might be challenging to identify an appropriate kernel function [77] | SVR achieves an acceptable result, demonstrating adequate performance across all criteria. However, this study may have had a disadvantage due to the difficulty in solving nonlinear problems |
|                            | ML algorithm | High generalization ability with a limited training sample set [78] | When the number of observation samples is enormous, the effectiveness of SVR may be low [78] | |
| **RFR**                    | Statistical property | Can handle high-dimensional data, has a quick training speed, and can recognize the mutual impact of features [77] | In some noisy classification or regression issues, there will be overfitting, and attributes with more values will have a higher influence on the results [77] | RFR receives an acceptable result, demonstrating adequate performance across all categories. However, this work may have suffered from overfitting in some noisy regression issues with higher values of economic attributes |
|                            | ML algorithm | It can balance mistakes in uneven data sets, and even if a major portion of the features are destroyed, accuracy can still be preserved [77] | The estimations are biased [78] | Ridge regression results are restricted in this study since the data set contains poorly correlated environmental variables |
| **Ridge**                  | Statistical property | Capable of dealing with highly correlated environmental variables (multicollinearity) [78] | | |
|                            | ML algorithm | Is useful when the amount of data is small in comparison to the number of variables [78] | As an advantage, LASSO may be utilized for feature selection. However, when compared to other ML algorithms of the filter approach, it has limitations when used in the regression process |
| **Embedded LASSO**         | Statistical property | Provides an interpretable model and selects a subset of predictors having the greatest influence on the response variable [79] | Selects one covariate at random from a set of highly collinear variables to incorporate in the model and discards the others [79] | This strategy provides a potentially important feature set of set H and set I, which may be used as the parent set for feature union and intersection operations. However, when compared to other ML algorithms of the filter approach, it has limitations when used in the regression process |
|                            | ML algorithm | When less data is available, it might be utilized for feature selection [79] | | |
| **E-net**                  | Statistical property | Performs well when the number of parameters is greater than the number of samples [79] | Cannot be utilized for feature selection when there is a limited amount of data available because it overwhelms the data with too many model variables [79] | |
Table 11: List of study countries

134 countries with LPI information available from year 2010 to 2018

| Afghanistan* | Colombia | Germany | Kuwait | Nigeria | Spain |
|-------------|----------|---------|--------|---------|-------|
| Angola*     | Congo, Dem. Rep.* | Greece | Lao PDR* | Oman* | Sweden |
| Argentina   | Congo, Rep. | Guatemala* | Latvia | Pakistan | Switzerland |
| Armenia     | Costa Rica | Guinea* | Lebanon | Panama | Syrian Arab Rep.* |
| Australia   | Côte d’Ivoire | Guinea-Bissau* | Liberia* | Papua New Guinea | Taiwan, China |
| Austria     | Croatia | Guyana* | Libya | Paraguay | Tajikistan* |
| Bahamas*    | Cuba* | Haiti* | Lithuania | Peru | Thailand |
| Bahrain     | Cyprus | Honduras | Luxembourg | Philippines | Togo* |
| Belgium     | Czech Rep | Hong Kong SAR, China | Macedonia, FYR* | Poland | Tunisia |
| Benin*      | Denmark | Hungary | Madagascar | Portugal | Turkey |
| Bhutan*     | Djibouti* | Iceland | Malaysia | Qatar | Ukraine |
| Bolivia     | Dominican Rep.* | India | Maldives* | Romania | United Arab Emirates |
| Bosnia and Herzegovina* | Ecuador | Indonesia | Malta | Russian Federation | United Kingdom |
| Brazil      | Egypt, Arab Rep | Iraq* | Mexico | Rwanda | United States |
| Bulgaria    | El Salvador | Ireland | Moldova | Saudi Arabia | Uruguay* |
| Burkina Faso | Eritrea* | Italy | Mongolia* | Senegal | Uzbekistan |
| Cambodia*   | Estonia | Jamaica | Montenegro* | Serbia | Venezuela, RB |
| Cameroon    | Fiji | Japan | Myanmar* | Singapore | Vietnam |
| Canada      | Finland | Jordan | Nepal* | Slovak Rep | |
| Chad        | France | Kazakhstan | Netherlands | Slovenia | |
| Chile       | Gabon | Kenya | New Zealand | Solomon Islands* | |
| China       | Georgia | Korea, Rep | Niger | South Africa | |

*Excluded in 100 countries for feature selection study

Source: World bank LPI data

Table 12: Economic attribute description

| Economic attribute | Description |
|--------------------|-------------|
| Real GDP growth (%) | Percent change in real GDP, over previous period |
| Nominal GDP (SB) | Nominal GDP |
| Population growth (%) | Percent change in population, over previous period |
| GDP per capita ($) | Nominal GDP divided by population |
| Private consumption growth (%) | Percent change in private consumption, over previous period |
| Government consumption growth (%) | Percent change in real government consumption, over previous period |
| Gross fixed investment growth (%) | Percent change in real gross fixed investment, over previous period |
| Current account balance (SB) | Trade balance, plus net services, plus net income, plus net current transfers |
| Consumer prices growth (%) | Percent change in consumer price index (12-month average), over previous period |
| Consumer price index, average | Consumer price index, period average. Base years vary |
| Budget balance/ GDP (%) | Budget balance as a percent of GDP |
| GDP deflator (2005 = 100) | GDP deflator index in local currency, period average (2005 = 100) |
| Public debt/ GDP (%) | Public debt as a percent of GDP |

Central government receipts minus central government outlays, as a percent of GDP
Table 12 (continued)

| Economic attribute                                               | Description                                                                                                                                 |
|------------------------------------------------------------------|--------------------------------------------------------------------------------------------------------------------------------------------|
| Budget expenditure/ GDP (%).                                     | Aggregate national savings by the public and private sector as a percent of nominal GDP. National savings equals gross domestic investment plus the current-account balance |
| Gross national savings rate (%).                                 | Economically active population                                                                                                                     |
| Labor force (M)                                                  | Current account balance as a percent of GDP                                                                                                            |
| Current account balance/ GDP (%)                                 | Total exports of goods on a free on board (FOB) basis                                                                                                  |
| Goods: Exports (FOB) ($B)                                        | Total imports of goods on a cost, insurance and freight (CIF) basis                                                                                    |
| Total Imports (CIF) ($B)                                         | Net flows of direct investment                                                                                                                         |
| Net direct investment flows (SB)                                 | Total reserves (excluding gold), including foreign exchange, reserve position with the IMF and SDRs at end period                                       |
| Forex reserves (SB)                                              | Net flows of direct investment capital by non-residents into the country                                                                            |
| Inward direct investment (SB)                                    | Net flows of direct investment capital by domestic residents out of the country                                                                     |
| Outward direct investment (SB)                                   | Exports of goods less imports of goods                                                                                                               |
| Trade balance (SB)                                               | Interest rate on money market lending                                                                                                                  |
| Money market interest rate (%)                                   | Percentage change in bank lending to public and private sectors, plus bank lending in domestic currency                                                       |

Source: S&P Global Market Intelligence’s macroeconomic data

**Funding** Open Access funding enabled and organized by CAUL and its Member Institutions.

**Declarations**

**Conflict of interest** The authors declare that they have no conflict of interest.

**Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article’s Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article’s Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visithttp://creativecommons.org/licenses/by/4.0/.

**References**

1. World Bank (2018) Connecting to Compete 2018 Trade Logistics in the Global Economy The Logistics Performance Index and Its Indicators. http://hdl.handle.net/10986/29971. Accessed 31 August 2021

2. Gerschberger M, Manuj I, Freinberger PP (2017) Investigating supplier-induced complexity in supply chains. Int J of Phys Distr Logist Manag 47(8):688–711

3. Wong WP, Tang CF (2018) The major determinants of logistic performance in a global perspective: evidence from panel data analysis. Int J of Logist Res Appl 21(4):431–443

4. D’Aleo V, Sergi BS (2017) Does logistics influence economic growth? European Exp Manag Decis 55(8):1613–1628

5. Takele TB (2019) The relevance of coordinated regional trade logistics for the implementation of regional free trade area of Africa. JTSCM 13(1):1–11

6. Chandraeshekar G, Sahin F (2014) A survey on feature selection methods. Comput Electr Eng 40(1):16–28

7. Vieira SM, Sousa JM, Runkler TA (2010) Two cooperative ant colonies for feature selection using fuzzy models. Expert Syst Appl 37(4):2714–2723

8. Muthukrishnan R, Rohini R (2016) LASSO: A feature selection technique in predictive modeling for machine learning. In: Proceeding of the 2016 IEEE international conference on advances in computer applications (ICACA), pp. 18–20

9. Khmaissia F et al (2018) Accelerating band gap prediction for solar materials using feature selection and regression techniques. Comput Mater Sci 147:304–315

10. Sikora R, Piramuthu S (2007) Framework for efficient feature selection in genetic algorithm based data mining. Eur J Oper Res 180(2):723–737

11. Lu M (2019) Embedded feature selection accounting for unknown data heterogeneity. Expert Syst Appl 119:350–361

12. Lal TN et al (2006) Embedded methods, in Feature extraction. Springer, pp 137–165.
55. García-Nieto PJ, García-Gonzalo E, Paredes-Sánchez JP (2021) Prediction of the critical temperature of a superconductor by using the WOA/MARS, ridge, lasso and elastic-net machine learning techniques. Neural Comput Appl 33:17131–17145
56. Kong X et al (2015) Wind speed prediction using reduced support vector machines with feature selection. Neurocomputing 169:449–456
57. Başakın EE et al (2021) A new insight to the wind speed forecasting: robust multi-stage ensemble soft computing approach based on pre-processing uncertainty assessment. Neural Comput Appl 34:783–812
58. Ullah QZ et al (2021) A Cartesian genetic programming based parallel neuroevolutionary model for cloud server’s CPU usage prediction. Electronics 10:67
59. Guo J et al (2019) An XGBoost-based physical fitness evaluation model using advanced feature selection and Bayesian hyper-parameter optimization for wearable running monitoring. Comput Netw 151:166–180
60. Citakoglu H (2021) Comparison of multiple learning artificial intelligence models for estimation of long-term monthly temperatures in Turkey. Arab J Geosci 14:2441
61. Huang R et al (2021) Machine learning in natural and engineered water systems. Water Res 205:117666
62. Ahmadi-Nedushan B et al (2006) A review of statistical methods for the evaluation of aquatic habitat suitability for instream flow assessment. River Res Appl 22(5):503–523
63. Boucher TF et al (2015) A study of machine learning regression methods for major elemental analysis of rocks using laser-induced breakdown spectroscopy. Spectrochim Acta B 107:1–10

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.