Apuntes 99, 5-42
ISSN: 0252-1865
eISSN: 2223-1757
doi: 10.21678/apuntes.99.2189

Artículo recibido el 29 de septiembre de 2023
Artículo aceptado para publicación el 3 de enero de 2025

GDP nowcasting with machine learning and unstructured data[1]

Juan Tenorio
Universidad Peruana de Ciencias Aplicadas
pcefjten@upc.edu

Wilder Perez
Universidad Científica del Sur, Lima; Ministerio de Economía y Finanzas del Perú
wperezc@cientifica.edu.pe

Abstract. Nowcasting models based on machine learning (ML) algorithms deliver a noteworthy advantage for decision-making in the public and private sectors due to their flexibility and ability to handle large amounts of data. This article introduces real-time forecasting models for the monthly Peruvian GDP growth rate. These models merge structured macroeconomic indicators with high-frequency unstructured sentiment variables. The analysis spans January 2007 to May 2023, encompassing a set of 91 leading economic indicators. Six ML algorithms were evaluated to identify the most effective predictors for each model. The findings underscore the remarkable capability of ML models to yield more precise and foresighted predictions compared to conventional time series models. Notably, the *gradient boosting machine*, *LASSO*, and *elastic net* models emerged as standout performers, achieving a reduction in prediction errors of 20% to 25% compared to autoregression and various specifications of dynamic factor model. These results could be influenced by the analysis period, which includes crisis events featuring high

1 This article is a revised and expanded version of a working paper published by the Peruvian Economics Association (Tenorio & Perez, 2023),. The opinions expressed in this article and any errors and omissions are solely the authors' responsibility and do not necessarily reflect the views of the Ministry of Economy and Finance. This work would not have been possible without the support of Adrian Colonna, Arian Segil, and Shadia Muñoz-Najar, the valuable comments of Carlos Montoro, Ricardo Quineche, Marco Vega, Youel Rojas, Juan Carlos Aquino, and Fernando Pérez Forero of Central Reserve Bank of Peru, Jose Luis Bustamante, Wilder Ramirez, Ricardo Najarro, the staff of the General Directorate of Macroeconomic Policy of the Ministry of Economy and Finance, and comments from the attendees of the 41st Meeting of Economists of Central Reserve Bank of Peru 2023.

uncertainty, where ML models with unstructured data improve significance.
Clasification JEL: C32, C53, E37, C52, E32.

Key Words: nowcasting, machine learning, GDP growth.

Resumen. Los modelos de *nowcasting* basados en algoritmos de Machine Learning (ML) ofrecen una ventaja notable para la toma de decisiones en los sectores público y privado debido a su flexibilidad y capacidad para manejar grandes cantidades de datos. Este documento presenta modelos de pronóstico en tiempo real para la tasa de crecimiento mensual del PIB peruano. Estos modelos combinan indicadores macroeconómicos estructurados con variables de sentimiento no estructurados de alta frecuencia. El análisis comprende desde enero de 2007 hasta mayo de 2023, abarcando un conjunto de 91 indicadores económicos principales. Se evaluaron seis algoritmos de ML para identificar los predictores más eficaces de cada modelo. Los resultados subrayan la notable capacidad de los modelos de ML para producir predicciones más precisas y previsoras que los modelos convencionales de series temporales. En particular, Gradient Boosting Machine, LASSO y Elastic Net destacaron por sus resultados, logrando una reducción de los errores de predicción de entre el 20% y el 25% en comparación con los modelos AR y varias especificaciones de DFM. Estos resultados podrían estar influenciados por el periodo de análisis, que incluye acontecimientos de crisis con un alto grado de incertidumbre, en los que los modelos ML con datos no estructurados mejoran la significación.

Clasificación JEL: C32, C53, E37, C52, E32.

Palabras clave: *nowcasting*, machine learning, crecimiento del PBI.

1. Introduction

Real-time decision-making is a significant challenge for policymakers who often face delays in obtaining updated information about macroeconomic indicators. In most cases, the economic variables present a delay of 30–45 days on average, including the time for revisions and retrospectives. Nevertheless, the continuous strides toward the new generation of high-frequency data have changed how prediction models address the inherent uncertainty in this information. As a result, in the past few years, central banks and international institutions have adopted methodological focuses that incorporate machine learning and take advantage of the abundant quantities of data that come from search engines and social media, as shown in Araujo et al. (2023); Chakraborty and Joseph (2017); Richardson and Mulder (2018).

These automated learning techniques have gained greatly in popularity compared to the conventional focus of traditional time series models to project macroeconomic variables. An often highlighted characteristic of these algorithms is their capacity to formulate parametric selections in large amounts of data sets, based on training a specific percentage of the model's information. The objective of this paper is to explore the benefits of utilizing several machine learning methodologies. We do so by combining the use of conventional leading indicators (structured data) and sentiment data indexes (non-structured or unstructured data) to forecast in real-time (nowcast) Peru's monthly real GDP growth rate.

The data set consists of both local and international variables, which can be broken down into 53 structured variables and 38 nonstructured variables, giving a total of 91 predictors. We examine these predictive variables based on the model, between September 2014 and May 2023, to evaluate the optimum performance of each. Furthermore, following Romer and Romer (2008) we performed a predictive accuracy analysis using two models as reference, the traditional autoregressive time series and a dynamic factor model, based on the leading indicator of electricity production used in the economic literature and by Peruvian political and economic consulting firms. This facilitates an exhaustive evaluation of the performance of machine learning algorithms.

The results show that the immediate predictions of machine learning models are more robust than the benchmark auto-regressive model and perform better than dynamic factor model (DFM). Specifically, the *random forest*, *gradient boosting machine*, and *adaptive LASSO* models demonstrate a superior ability to reduce the average projection error in a range of 20-25%. In addition, following the methodology proposed by Armstrong (2001), we corroborate that using the average projection value all the machine learning

algorithms adds significant value to the root mean square error (RMSE), which contributes positively to a more precise prediction of GDP. Even though other models—*ridge*, the least absolute shrinkage and selection operator (*LASSO*), and *elastic vet*—do not reach the same level of predictive ability as the aforementioned machine learning (ML) methodologies, they still outperform the control model.

Further, the proof of forecasting evaluation and consistency assessment confirms that most of the machine learning models improve the prediction significantly, in line with previous literature applied in other contexts (Richardson & Mulder, 2018; Varian, 2014; Zhang et al., 2023).

This article contributes to the literature that highlights the success of machine learning applications in contrast to more traditional methodologies.[2] However, given the lack of evidence in Latin America,[3] and in particular Peru,[4] regarding the use of these algorithms in conjunction with non-structured data, this research project also highlights the need foreground a discussion about what these models entail. Barrios et al. (2021), Richardson and Mulder (2018), and Döpke et al. (2017) have shown through the implementation of diverse machine learning algorithms that the results of these methods are better suited to carrying out forecasts in real-time when a large amount of information is available to the forecaster. For example, Longo et al. (2022) carried out a forecast of quarterly GDP in the US for the combination of a neuronal recurrent network and a dynamic factor model with a temporal variation of the median. This combination of models demonstrated a substantial decrease in the forecast error, as well as a capability to capture the period of recession associated with the COVID-19 pandemic and the subsequent economic recovery. Similarly, in the case of El Salvador and Belize, Barrios et al. (2021) implemented a large array of machine learning methods and predictive variables to forecast the quarterly growth of GDP. The results concluded that the application of these tools represents a robust alternative to prediction, and its benefits led the authors to recommend their use in other countries in the region. Other researchers have extended the application of machine learning models beyond GDP to the likes of forecasting, inflation, yield curve, and active prices. These efforts have yielded notable results in precise forecasting (Giglio et al., 2022; Medeiros et al., 2021).

---

2 It is important to mention the pre-publication of our paper, Tenorio & Perez (2023), by the Central Bank of Peru and the Peruvian Economics Association as well as at meetings of economists organized by the Central Bank of Peru, where we received valuable feedback from other experts.

3 See Barrios et al. (2021).

4 See Escobal D'Angelo and Torres (2002); Perez Forero (2018).

It is still important to highlight that the implementation of these methods still presents challenges, sparking major debates. For instance, Green and Armstrong (2015) and Makridakis et al. (2018), when comparing multiple models of machine learning, found that the results of the forecasting were less significant in comparison with statistical smoothing approaches and ARIMA models. These authors warned that the computational complexity inherent to variable selection and use in the machine learning model makes immediate forecasting difficult and less practical for policymakers.

The remainder of this article is structured as follows. The next section presents a literature review that explores the relevance of the nowcasting methodology in the context of machine learning and big data, both at the national and international levels. Thereafter, a section is devoted to the methodology, models, and data sets used. The results are then displayed in a specific section, followed by the robustness tests and the conclusion.

2. Literature review

Economists aim to provide the most accurate GDP forecasts using the most efficient approaches. Stock and Watson (1989) were the first to propose an economic cycle index using factor models. However, a critical challenge is the increase in uncertainty in the estimates, an area in which traditional models, which use a limited set of variables, often fall short. The literature has therefore been exploring new models using machine learning techniques to balance the trade-off between bias and variance.

Nowcasting methods seeks to address the issue of extended delays in the publication of key economic aggregates, as well as aims to predict the present, the very near future, and the very recent past (Bánbura et al., 2013). A traditional reference nowcasting model is the DFM, widely used in central banks to predict GDP (Bánbura & Rünstler, 2011; Bok et al., 2018; Giannone et al., 2008; González-Astudillo & Baquero, 2019; Rusnák, 2016). Two seminal studies have formalized this process into statistical models. On one hand, Giannone et al. (2008) proposed a methodology to assess the marginal impact of the publication of monthly-updated data on forecasts of quarterly-published real GDP growth. The authors proposed a method to track the real-time flow of information that central banks monitor through large datasets with staggered publication dates. The proposed method works by updating primary forecasts (forecasts for the current quarter) each time new higher-frequency data is published. This is done using progressively larger datasets that reflect the unsynchronized data publication dates. On the other hand, Evans (2005) performed real-time estimations of the current state of the US economy using an approach that included data

complexity and provided useful information about the relationship between macroeconomics and asset prices. The author modeled monthly time series using a DFM in a state-space system. Once the state-space representation was settled, Kalman filter techniques were estimated for GDP forecasting, as they automatically adapt to changes according to the data available. In the present study, we followed the proposals of Evans (2005) to perform our DFM specifications in addition to the implementation suggestions of Doz et al. (2012).

An additional advantage of nowcasting models is the constant improvement in wider information availability and data frequency heterogeneity (González-Astudillo & Baquero, 2019; Zhang et al., 2023). Thus, ML methods are now being incorporated to enhance the nowcasting approach. ML algorithms deliver better performance in handling large amounts of data, capturing non-linear relationships, and adapting to changing economic conditions.

ML methods provide more accurate predictions by incorporating various variables and sources of unstructured data. As noted by Athey (2018), these techniques can be divided into two main categories: unsupervised and supervised ML. The former seek groups of observations that are similar in terms of their covariance. Thus, a "dimensionality reduction" can be performed. Unsupervised MLs commonly use videos, images, and text as sources of information, in techniques such as K-means clustering. For instance, Blei et al. (2003) applied pooling models to find topics in textual data. In turn, Woloszko (2020) presented a weekly indicator of economic activity for 46 OCDE countries and the G20 using search data from Google Trends. The author illustrated the power of prediction of specific topics, including "bankruptcies," "economic crises," "investment," "baggage," and "mortgages." Calibration was performed using a neural network that captured nonlinear patterns, which were shown to be consistent with economic intuition using ML Shapley values interpretation tools.

On the other hand, supervised ML algorithms entail the use of a group of variables or features to predict a specific indicator result (Varian, 2014). The variety of supervised ML regression methods in circulation include *LASSO, ridge, elastic net, random forest, regression trees, support vector machines, neural nets, matrix factorization, and model averaging,* among others.

Several studies highlight the advantages of supervised ML models over traditional methods in forecasting macroeconomic series. Ghosh and Ranjan (2023) presented a compilation of ML techniques and conventional time series methods to predict the Indian GDP, estimating ML in the DFM context with financial and economic uncertainty data. They

employed *random forest* and *prophet* models along with conventional time series models such as ARIMA to nowcast Indian GDP, finding that hybrid models stand out. Similarly, Richardson and Mulder (2018) detected that a *ridge* regression model outperformed a DFM for a GDP nowcast GDP of New Zealand. Muchisha et al. (2021) built and compared ML models to forecast the GDP of Indonesia. They evaluated six ML algorithms, *random forest, LASSO, ridge, elastic net, neural networks,* and *support vector machines,* using 18 variables between 3Q2013 and 4Q2019. Their results illustrate the outstanding performance of ML versus auto-regressive models, especially the *random forest* model.

For their part, Zhang et al. (2023) tested ML, DFM, and static factor and MIDAS regression models to nowcast the GDP rate growth of China, observing the superior accuracy of ML compared to DFM. *Ridge regression* surpassed all other ML models in prediction and early anticipation of crises such as the global financial crisis and COVID-19. Kant et al. (2022) compared models applied to the Dutch economy between 1992 and 2018, with *random forest* algorithms standing out. Using novel variables such as Google Search and air quality, Suphaphiphat et al. (2022) ran standard DFM and ML on European economies during normal times and crises. They showed that most MLs significantly outperformed the AR (1) reference model; DFM tended to perform better in normal times, while many of the ML methods excelled in identifying turning points. Moreover, ML proved able to predict adequately in very disparate economies. Meanwhile, Barrios et al. (2021) assessed adjusted ML models on the Belizian and Salvadoran economies and found that they delivered robust predictions, adding to the evidence that ML algorithms are effective in very different country contexts.

Another relevant consideration is Big Data due to its benefits in broadening the range and use of available data to provide valid information on the behavior of the economy and anticipate certain economic indicators (Einav & Levin, 2014). As mentioned in Eberendu et al. (2016), the digital era has seen the emergence of digital news platforms, social media technologies, smartphones, and online advertising. Nevertheless, many of the new data types—text, XML, email, images, videos, and so on—lack a pre-fixed format, raising new challenges and attracting new research. Eberendu et al. (2016) proposed a general description of this type of data. Some studies show relevant results on the use of these techniques. For instance, Varian (2014) proposed that a search for "initial claims for unemployment" in Google Trends offered good basis on which to forecast unemployment, CPI, and consumer confidence in countries such as the US, UK, Canada, Germany, and Japan. The author focused on immediate out-of-sample forecasting and

extended the Bayesian structural time series model using the Hamiltonian sampler for variable selection, obtaining good results for unemployment but less so for CPI or consumer confidence.

In the Latin America context, Barrios et al. (2021), Richardson and Mulder (2018), and Döpke et al. (2017) have shown through the implementation of diverse ML algorithms that the results are more promising for carrying out forecasts in real-time when a large amount of information is at the researchers' disposal. Caruso (2018) noted the benefits of using external indicators in short-term GDP forecasting in Mexico, assessing a DFM model that deals with the mixed frequency of macroeconomic indicators. Gálvez-Soriano (2020) showed that the *bridge equation* model did better than DFM and principal components analysis in predicting monthly Mexican GDP. Corona et al. (2022) illustrated the gains on DFM models of including nontraditional variables such as Google Trends with regard to the Mexican Global Economic Activity Indicator (IGAE). Bolivar (2024) nowcasted monthly economic growth by using machine learning algorithms and integrating data from both traditional and remote-sensing sources, for the case of Bolivia. The results indicated that these tools (ML and Big Data) represented a solid alternative to prediction, and their benefits lend to usage in other countries in the region. Other researchers have extended the application of ML models to GDP, inflation, yield curve, and active prices. These efforts have yielded notable results in precise forecasting (Giglio et al., 2022; Medeiros et al., 2021).

In the case of the Peruvian economy, previous works have focused on the anticipated estimation of monthly GDP growth based on a set of leading indicators (structured data). However, the limited application of machine learning models and the inclusion of unstructured data in GDP forecasting is evident. For instance, Escobal D'Angelo and Torres (2002) built a joint leading indicator that allows the tracking of Peruvian GDP with only 14 variables. Kapsoli Salinas and Bencich Aguilar (2002) performed a forward GDP estimation with a nonlinear neural network model. In turn, Etter et al. (2011) proposed a leading indicator using an expectations survey conducted by the Central Bank of Peru (BCRP). Martınez and Quineche (2014) forecast the GDP growth rate based only on the electricity production indicator. Following Aruoba et al. (2009), Forero et al. (2016) proposed a leading indicator of Peruvian economic activity, obtained as a common unobservable component that explains the co-movement among six variables: electricity production, domestic cement consumption, adjusted domestic sales tax, chicken sales, metal mining production, and real GDP. Finally, Pérez Forero (2018) attempted to solve the difficulties about best

leading indicators selection under the approach of Varian (2014). Perez Forero estimated a steady state system through the Bayesian Gibbs-Sampling methods and a spike-and-slab to perform stochastic search variable selection (SSVS), calculating the probability of inclusion of a large set of variables in the best model to predict GDP.

Finally, the studies applied to Peru and focused on implementing machine learning techniques include those by Tenorio and Perez (2024) and Tenorio and Perez (2023), which are working papers that were updated and reviewed. These pre-publications have provided valuable feedback and insights from experts on the subject, allowing us to refine our contribution.

3. Methodology

This section briefly describes the different regularization methods and decision trees used to select the best predictors for the monthly nowcasting model and calibrate the hyperparameters, in a series from January 2007 to May 2023. The six methods that are used are *random forest* (RF), *gradient boosting machine* (GBM), *LASSO* regression, *ridge*, *elastic net*, and, as a benchmark, an autoregressive (AR) and dynamic factor model (DFM).

3.1 Autoregressive model (AR)

As a starting point, we established an autoregressive AR model for monthly GDP growth (y_t), which reflects the value of a variable in terms of its previous values. A model of order 1, following these characteristics, exhibits the following structure:

$$y = \beta_0 + \beta_1 y_{t-1} + e_t \tag{1}$$
$$e_t \sim N(0,\sigma^2)$$

where β_0 is a constant term, β_1 is a parameter, and e_t is a term that follows a normal distribution with a mean of zero and a constant variance σ^2 and captures the randomness of the model.

3.2 Stepwise least squares

Stepwise regression is a method that sequentially fits a model by adding or removing variables iteratively based on different statistical criteria, with the aim of minimizing the mean squared error. This model combines simplicity and robustness with which to improve the model's projection capability. The variable selection process can be carried out through either forward selection, backward selection, or a combination of both known as bidirectional stepwise regression. This study aims to find the best choice within the universe of 91 leading indicators.

3.3 Dynamic factor model (DFM)

DFMs are estimated in the form of state-space systems using the Kalman filter and various types of algorithms. Following the proposal by Doz et al. (2011), one of the most popular in the economic literature is the expectation maximization algorithm due to its robust numerical properties, which make it an efficient estimator for bigger datasets.

The canonical reference DFM can be described as follows:

$$x_t = Bf_t + e_t \qquad e_t \backsim N(0, R) \tag{2}$$

$$f_t = \sum_{j=1}^{p} A_j f_{t-j} + u_t \qquad u_t \backsim N(0, D) \tag{3}$$

Where equation (2) is identified as the measurement equation, and equation (3) as the transition equation, allowing the unobservable factor f_t to evolve as in a vector autoregressive model. These equations do not include trends or intercepts, as the included data must be stationary and standardized before estimation.

The matrix system is as follows:

x_t: a vector of n x 1 observable time series at time t: $(x_t, \ldots, x_{nt})'$, which allows for missing data.

f_t: a vector of r x 1 factors at time t: $(f_t, \ldots, f_{rt})'$.

B: a matrix of n x r observable time series with lag j.

D: a matrix of r x r state covariances.

R: a matrix of r x r measurement covariances. This matrix is diagonal under the assumption that all covariances between the series are explained by the factors $E\,[x_{it}\,|x_{-i,t}, f_t] = b_i f_t, \forall i$, where b_i is the $i-th$ row of B.

This model can be estimated using a classical form of the Kalman filter and the maximum likelihood estimation algorithm, after transformation into a state–space model. In a VAR expression, it would be as follows:

$$x_t = CF_t + e_t \qquad e_t \backsim N(0, R) \tag{4}$$

$$F_t = AF_{t-1} + u_t \qquad e_t \backsim N(0, Q) \tag{5}$$

As a benchmark model, we use the efficient estimation of a DFM via the EM algorithm on stationary and seasonally adjusted data with time-invariant system matrices and classical assumptions, while permitting missing data (Bánbura & Modugno, 2014).

3.4 Penalized regression models

These methodologies are employed to optimize the selection of predictor variables and control the model complexity, which is crucial to preventing

overfitting in high-dimensional settings. The literature suggests different forms of penalization to estimate the parameters β_j accurately. We will briefly explore the characteristics of the *ridge*, *LASSO*, *elastic net*, and *adaptive LASSO* models, emphasizing how these techniques allow for proper weighting of coefficients and how their application affects the inclusion and relevance of variables in the final model.

3.4.1 Ridge regression

The *ridge* model is defined by adding a penalty based on the sum of squares of the coefficients of the predictor variables. This penalty compels the coefficients to be very small and prevents them from taking extremely high values, thus reducing the influence of less relevant variables. To estimate the coefficients $\hat{\beta}^{Ridge}$, the equation must be expressed as:

$$min_{\beta}\left(\sum_{i=1}^{n}(y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j)^2 + \lambda\sum_{j=1}^{p}\beta_j^2\right) \tag{6}$$

Where y_i is the observed value of the dependent variable for observation i, x_{ij} is the value of predictor variable j in observation i, β_j is the coefficient associated with predictor variable j, p is the number of predictor variables, and λ is the regularization hyperparameter that controls the magnitude of the penalty. The sum of the terms β_j^2in the penalty prevents the coefficients from reaching large values, thereby contributing to stability and reducing the risk of overfitting.

3.4.2 LASSO regression

The *LASSO* model, introduced by Tibshirani (1996), employs a penalty based on the sum of the absolute values of the coefficients of the predictor variables. This penalty forces some coefficients to reach exactly zero, automatically selecting a subset of more relevant predictor variables and eliminating less significant ones. The LASSO coefficients $\hat{\beta}^{Lasso}$ are estimated as:

$$min_{\beta}\left(\sum_{i=1}^{n}(y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j)^2 + \lambda\sum_{j=1}^{p}|\beta_j|\right) \tag{7}$$

The change lies in the hyperparameter λ which, by summing the absolute values of the coefficients $|\beta_j|$ in the penalty, leads to model selection and simplification by allowing some coefficients to be zero. This provides a more precise variable selection approach regarding the degree of importance of all variables.

3.4.3 Elastic net regression

The *elastic net* model combines appropriately the constraints of both the *LASSO* and *ridge* models. Zou and Hastie (2005) noted that its advantage lies in correcting the model when the number of regressors exceeds the number of observations ($p > n$), which improves variable grouping. The penalty includes both the sum of the absolute coefficient values and the sum of squares of the predictor variable coefficients. The equation for estimating the coefficients $\hat{\beta}^{Enet}$ is expressed as:

$$min_{\beta}\left(\sum_{i=1}^{n}(y_i - \beta_0 - \sum_{j=1}^{p}x_{ij}\beta_j)^2 + \lambda\sum_{j=1}^{p}(\alpha\,|\beta_j| + (1-\alpha)\beta_j^2)\right) \qquad (8)$$

where λ is the global regularization hyperparameter and α is the hyperparameter that controls the mix between *LASSO* ($\alpha = 1$) and *ridge* ($\alpha = 0$) penalties. The combination of both penalties in the *elastic net* model allows for a higher degree of flexibility in variable selection and coefficient alignment.

3.4.4 Adaptive LASSO regression

Following Zou (2006) the *a*daptive *LASSO* model is a variant of the *LASSO* model that introduces a penalty approach to adaptively adjust the magnitude of the penalties for each coefficient of the predictor variables. This adaptation allows for penalties to be different for different coefficients, potentially resulting in a more precise selection of relevant variables. Liu (2014) argued that this process can be efficiently performed using the LARS algorithm. The equation for the *adaptive LASSO* model ($\hat{\beta}^{AdL}$) is expressed as:

$$min_{\beta}\left(\sum_{i=1}^{n}(y_i - \beta_0 - \sum_{j=1}^{p}x_{ij}\beta_j)^2 + \lambda\sum_{j=1}^{p}w_j|\beta_j|\right) \qquad (9)$$

Where λ is the regularization hyperparameter, and w_j is the adaptation factor for the coefficient β_j. It is important to note that the exact form of the adaptation factors w_j depends on the specific implementation and may vary. In general, these factors are calculated based on the absolute values of the coefficients in previous iterations of the algorithm.

3.5 Decision tree models

Decision Tree models are machine learning algorithms that represent decisions and actions in the form of a tree. In the present case, we will present two algorithms where each internal node of the tree represents a feature or attribute, and each branch represents a decision or rule based on that attribute. The training data is divided based on these decisions until it reaches leaf nodes,

which correspond to the predictions, in our case, related to monthly GDP growth. The use of these trees also allows for an improvement in variable selection by handling non-linear relationships in the model.

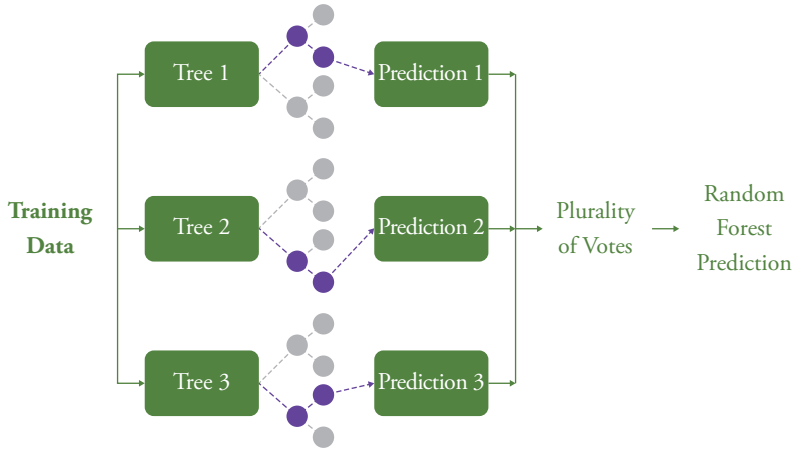
3.5.1 Random forest

This method is based on constructing decision trees using variables from a matrix X and a random selection of features. In addition, it involves randomly selecting subsets of data from X with replacement to train each tree in the ensemble, distinguishing it from other tree-based techniques. Each tree generates a prediction of the target variable (in this case, monthly GDP), and the final model selects the most voted prediction in the ensemble of trees (Breiman, 2001). According to Tiffin (2016) *random forest* has the advantage of combining predictions from multiple trees and selecting those with lower error, thereby reducing the influence of potential individual errors (if the correlation between trees is low). In sum, this method recursively divides the data in χ_i into optimized regions and uses variable-based criteria to forecast the target variable, then calculates the dependent variable as the average (avg) of these regions.

$$\hat{f}(\chi) = \sum_{m} \hat{c}_m I(\chi \in \mathrm{X}_r) : \hat{c}_m = avg(y_i \,|x_i \in \mathrm{X}_r) \tag{10}$$

The algorithm has certain advantages, such as efficiency in handling large datasets with many variables, providing an estimation of variable importance, and offering an unbiased estimation of generalization error during its construction (Breiman, 2001). However, it also has disadvantages, such as difficulty in interpreting results beyond predictions and computationally intensive demand for training and hyperparameter tuning. Therefore, it was necessary to fine-tune this model through cross-validation, thus achieving better performance on unseen data.

Figure 1
Simple representation of the random forest algorithm



Source: Compiled by authors.

3.5.2 Gradient Boosting Machine

This algorithm builds a sequence of decision trees in which each tree is fitted to the residual errors of the previous tree. Therefore, each iteration obtains a new tree that minimizes the remaining error. These prediction models are trained using the errors from the accumulated set of weak predictions[5] in a way that provides a progressive improvement in regression performance compared to the initial model (Natekin & Knoll, 2013).

In essence, each tree in this algorithm contributes its prediction, which is added to the sequence of predictions from previous trees to enhance the final prediction of the model. According to Boehmke and Greenwell (2020), this method can be summarized by the following equation:

$$F(\chi) = \sum_{z=1}^{Z} F_z(x) \tag{11}$$

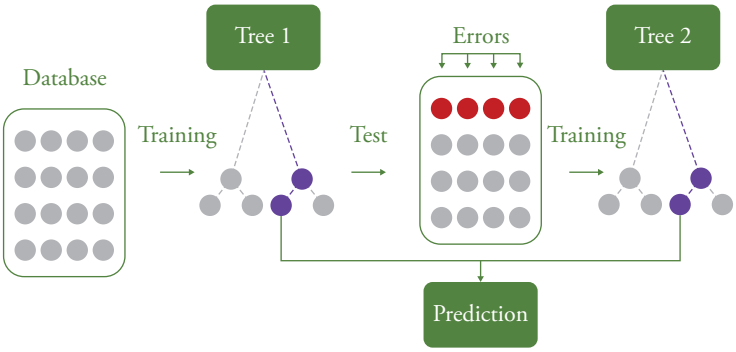
Where z is the number of trees that cumulatively sum the errors from all preceding trees. That is, if the first tree $y = F_1(x)$, then the second tree will be $F_2(x) = F_2(x) + e_1$ and so on, successively, to minimize $F(x)$ as the following expression:

5 Brownlee (2016) indicated that weak models are not necessarily weaker than accurate models, as they have the advantage of being able to correct the overfitting problem.

$$L = Min \sum_{Z} L(y_z, F_z(x)) \tag{12}$$

Therefore, as new decision trees are incorporated, the accuracy of the final projection gradually, improves resulting in more precise forecasts for monthly GDP.

Figure 2
Simple representation of the gradient boosting machine algorithm



Source: Boehmke & Greenwell (2020)

3.6 Data

The model database comprises a variety of variables, ranging from macroeconomic and financial data to unstructured information related to sentiment or "trend" (See Tables 6, 7, and 8). This information set encompasses consumption indicators, such as credits, deposits, consumer surveys, and local activity indicators, including electricity production, hydrocarbons, economic expectations, and others. Investment indicators are also incorporated, including internal cement consumption, capital goods imports, and so forth. A set of monetary indicators covering consumer and producer price indices, among others, is included. It is important to highlight the inclusion of economic sector variables related to fishing and agricultural production, which constitutes a unique feature compared to other nowcasting models. Furthermore, the database covers information on foreign trade, the labor market, and climate data.

In addition to conventional variables, we have incorporated unstructured data related to perception in various areas, such as the economy, consumption, labor market, politics, tourism, government support, and natural phenomena. These variables can capture the general sentiment of the population and its potential influence on economic indicators. In particular, the use of vast search engines, such as Google, stands out as a powerful tool

for providing real-time information. Scott and Varian (2013) have pointed out that the inclusion of online searches as variables provides substantial benefits to short-term forecasting models, especially in detecting periods of high volatility. This is demonstrated in the ability to anticipate both the recession caused by the COVID-19 pandemic and the subsequent period of economic recovery. Consequently, the effectiveness of this approach has been widely investigated and adopted by central banks and international institutions. Thus, we estimated ten groups (See Table 6) of variables with the aim of tracking Google search queries, which are updated daily and can be downloaded from Google Trends. The selection of words (variables) was intended to convey different aspects of the economy; for instance, the consumption-related group is constructed based on searches for words like "Kia," "restaurants," "Toyota," "credits," "loans," "deals," "mortgages," and "cinema."

Once this textual data was converted into numerical data, we evaluated the inclusion of these series in the estimations of an optimal model using Gibbs sampling, following Garcia-Donato and Martinez-Beneito (2013). For this we used 50,000 iterations, an initial burning of 1,000 iterations, and constant beta priors (see Figure 10). This indicates the high relevance of the group of unstructured variables, such as the search frequency for the likes of "flights," "peruflight us," "visa," or "El Niño", which reflect the dynamics of tourism and climatic conditions, among others. Furthermore, we compared the results of this estimation with another by confining the sample to 2019 (see Figure 11); unstructured data becomes more important when incorporating the pandemic period into the sample, which is in line with the findings of Richardson and Mulder (2018) and Woloszko (2020). Moreover, we performed a contemporaneous correlation analysis of these variables against monthly GDP, finding that more than half of the unstructured sample has a correlation greater than 30%.

The data frequency for constructing the model ranges from daily to monthly records. We assessed each variable in terms of its predictive ability regarding monthly GDP growth. Then, to facilitate comparison and analysis, we transformed these variables into annualized monthly percentage changes and standardized them. This standardization process allowed us to maintain a common reference framework and ensure that different variables contributed equitably to the model.

After obtaining a total set of 91 predictors spanning January 2008 to May 2023, we conducted the evaluation and selection of optimal predictors independently for each machine learning algorithm employed. We specify how we handled the data for the forecast update process in Section 4.1, and

how we tested the model accuracy comparison in Section 4.2. This approach enabled us to refine the process of choosing the most efficient prediction model, thereby achieving enhanced performance.

3.7 Forecast evaluation strategy

To assess the accuracy in the projection of each model we used the root mean square error (RMSE), following the equation:

$$RMSE = \sqrt{\frac{1}{\mathrm{T}}\sum_{t=1}^{T}(y_t - \hat{y}_t)^2} \tag{13}$$

Where y_t represents the observed value of monthly GDP growth, $\hat{y}_t$ is the forecasted value, and T is the total number of projections made. Following this initial assessment of prediction fit, we employed the method proposed by Diebold and Mariano (1995) to determine if the projections generated by each machine learning model significantly differed from the *benchmark* model.

4. Results

This section begins with a brief description of the database training period and hyperparameter optimization estimation, and finishes with a thorough analysis of the results.

4.1 Estimation and hyperparameter calibration

To estimate machine learning models, the selection of hyperparameters plays a crucial role in terms of efficiency and accuracy. Furthermore, calibrating the hyperparameters of each model with a smoothing range provides flexibility, reduces noise, and enhances forecast stability and accuracy.

The optimal determination of these values requires the division of the sample data into three parts: i) a training set, ii) a validation set, and iii) a testing set. First, we estimated the model using the training set (in-sample), which comprises the first set of hyperparameters. Then, the cross-validation method is used to calculate the best hyperparameters with the validation set. This process involves training and five-fold validation of the ML model in which every partition or fold is used as the validation set and the others as the training set on each iteration. Hence, we obtained five performance metrics, one for each fold, which we then averaged. Moreover, to identify the optimal hyperparameters we ran the cross-validation Bayesian optimization algorithm, closely following Snoek et al. (2012).

Table 1
Estimation testing strategy

Training dataset 2008m01-2014m08	**Validation set** 2014m09-2022m05	**Testing set** 2022m06-2023m05
⟷	⟷	⟷
Fold 1 Fold 2 Fold 3 Fold 4 Fold 5		

Source: Compiled by authors.

We then used cross-validation techniques to carry out the search process for the optimal values that minimize the mean quadratic error of projections (MSE[6]). The cross-validation entailed forecasting the growth (y_{t+h}) with the available data at time $t(y_{t+h}|I_t)$[7], with the hyperparameters obtained for each fold.[8] Once we identified the optimal values, we assessed the accuracy of the model in the testing set (out-sample) by evaluating the MSE between the projection growth with the available data available at time $t(y_{t+h}|I_t)$ and the available data at time $t + h(y_{t+h}|I_t)$. We repeated these steps to attain the minimization of the MSE value as shown in Figures 6 to 9 for each type of ML model.[9]

To prevent overfitting in the ML models, we bounded the hyperparameters within ranges recommended in the literature reviewed. (See Zou & Hastie, 2005.) This approach contributed significantly to the model's ability to make robust predictions, allowing for more effective exploration in estimating monthly GDP rate growth without the risk of overfitting.

6 Indicator that measures the average of the squared errors between the predictions of a model and the real values, without applying the square root, used for validation of parameters in ML models.

7 **I** is the available information set where we obtained the full available data of 91 predictors variables.

8 The h can be interpreted as the horizon to forecast, which in a nowcasting context s usually $h = 1$.

9 In case of partial availability of the information set or of the data pertaining to the 91 predictor variables, estimation could be performed using other techniques such as DFM with the modified EM algorithm of Bánbura and Modugno (2014), which also accounts for missing data in the EM iterations.

Table 2
Priors and hyperparameter ranges

Model	Hyperparameter	Range	Optimised Value
Lasso	Lambda	0.001 to 0.009	0.007
Ridge	Lambda	0.01 to 0.09	0.310
Elastic Net	Alpha Lambda	0.1 to 0.9 0.01 to 0.09	0.500 0.040
Adaptive Lasso	Lambda	0.01 to 0.09	0.670
Random Forest	Omega #Tress	0.1 to 0.9 1 to 400	0.340 281
Gradient Boosting Machine	#Tress Distribution Shrinkage	1 to 5000 Normal 0.001 to 0.009	19 Bernoulli 0.300

Source: Compiled by authors.

4.2 Model comparison

Table 3 presents a comparison of the prediction performance of the ML and benchmark models for the validation and test set, from September 2014 to May 2023. As to the forecast evaluation using the RMSE, the ML models succeeded in significantly minimizing the projection error in comparison with the benchmark AR model and the three different specifications of dynamic factor models.[10] Every projection model compared the forecast with the full available data set at time $t + h$ with the actual GDP rate growth $t + h$. The models that stand out over the others were the *gradient boosting machine*, *LASSO* and *elastic net* , each of which reduced the forecast error by around 20% to 25%.

We also estimated the Diebold–Mariano statistic, which is used to compare the accuracy of two forecast models. According to this statistic,[11] most of the ML models are statistically significant, in line with previous research (Richardson & Mulder, 2018; Varian, 2014; Zhang et al., 2023); most showed p-values below 0.05, suggesting that their forecasts are significantly different from the actual GDP values. This indicates that the predictions for these models are statistically distinguishable from the real outcomes. Adaptive *LASSO* (p=0.126) and *random forest* (p=0.089) presented higher p-values, indicating that their forecasts are not significantly different from the actual GDP values at the conventional significance levels. This could suggest these models provide more accurate predictions of the real GDP outcomes.

10 Bánbura & Modugno (2014).
11 Diebold & Mariano (1995).

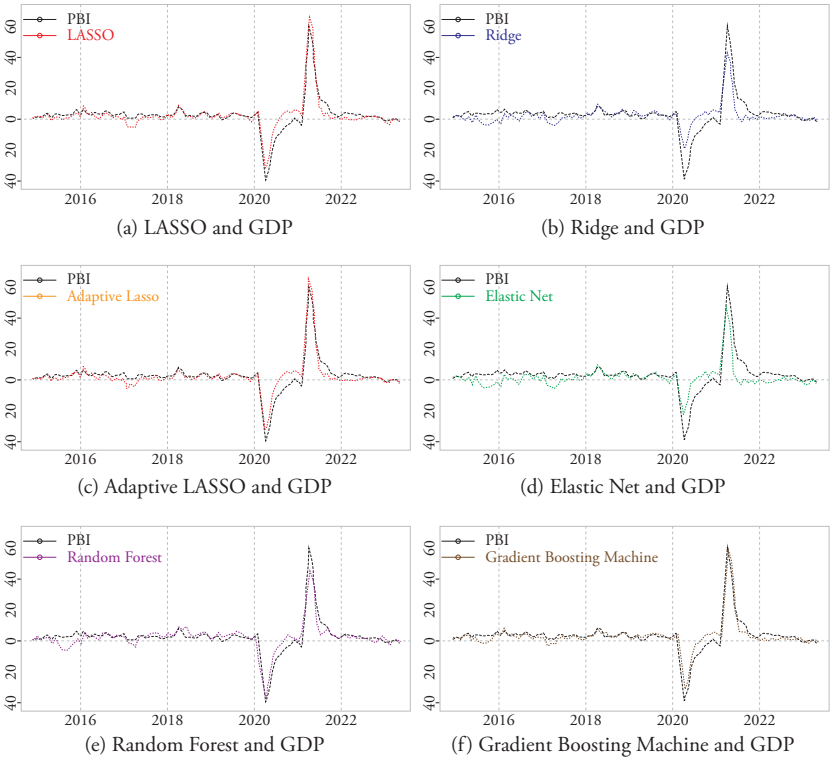
On the other hand, it is important to highlight that the real-time forecasts presented in this paper successfully anticipated the economic contraction in the Peruvian context associated with by the COVID-19 pandemic in March 2020, and also accurately captured the subsequent economic recovery period in March of the following year. This illustrates the usefulness and effectiveness of using penalty models and/or decision trees to forecast high-frequency economic variables.

Table 3
Evaluation of model and benchmark forecasts
2014m09–2023m05

Model	MAE	RMSE	RMSE (Rel. to AR)[1]	p-value (DM)
Lasso	0,29	0,26	0,10	0,014
Ridge	0,38	0,34	0,13	0,043
Elastic Net	0,33	0,28	0,11	0,039
Adaptive Lasso	0,51	0,68	0,27	0,126
Random Forest	0,4	0,45	0,18	0,089
Gradient Boosting Machine	0,11	0,17	0,07	0,016
Stepwise[2]	1,66	1,63	0,59	0,001
DFM full[3]	0,67	0,93	0,36	0,005
DFM best[4]	0,55	0,72	0,28	0,004
DFM structured[5]	0,86	1,05	0,41	0,003
Autoregressive model (AR)	2,14	2,55	0,00	

1/ $RMSE(Model)_i/RMSE(AR)$. 2/ Uses variables within unstructured data and structured selected by iteratively adding or removing variables based on statistical criteria. 3/ DFM full uses the 91 variables within unstructured as well as structured data. 4/ DFM best uses variables within unstructured data and structured selected by the Gibbs sampling as best estimators to predict GDP. 5/ DFM structured uses only 56 structured variables.
Source: compiled by authors.

Figure 3
Machine learning model projection and GDP



(a) LASSO and GDP (b) Ridge and GDP

(c) Adaptive LASSO and GDP (d) Elastic Net and GDP

(e) Random Forest and GDP (f) Gradient Boosting Machine and GDP

Source: Compiled by authors.

4.3 Consistency

To test the consistency of the results and determine whether the ML model projections contribute positively to the accuracy of predictions of monthly GDP versus the benchmark models, we used the Romer and Romer (2008) approach, except we replaced an official's prediction with a DFM estimation that incorporates electricity production as the main leading indicator—a popular approach among economic studies departments in Peru. We estimated the following regression model:

$$y_t = \beta_1 DFME_t + \beta_2 ML_{it} + e_t \tag{14}$$

Where y_t represents the real monthly GDP growth, $DFME_t$ is the dynamic factor model estimated using electricity production, and ML_i is the out-sample prediction for each machine learning model. The results obtained indicate

that all the projections of machine learning contribute significantly to the GDP projection, with the best model being the *gradient boosting machine* in line with the Akaike criterion. Likewise, analyzing the estimation errors of the models generated by equation (14), we applied the test proposed by Harvey et al. (1997) with the long-run variance autocorrelation estimator proposed by Diebold and Mariano (1995) to evaluate the accuracy gains in the estimates from the results of the ML models. The p–value is shown in the last column of Table 4, where the alternative hypothesis is that the models in equation (14), which include the ML model projection, are more accurate than the predictions under the dynamic factor model alone. These values indicate the superior accuracy of the models that incorporate ML at a 10% confidence level in the case of the *LASSO* and *ridge* models, but at 5% in the others.

Table 4
β_2^e value and validation criteria

Model	Estimated value	AIC	p-value	p-value (DM)
Lasso	0,714	520,32	0,00	0,079
Ridge	0,936	554,73	0,00	0,057
Elastic Net	0,839	549,80	0,00	0,055
Adaptive Lasso	0,703	517,49	0,00	0,046
Random Forest	0,783	534,20	0,00	0,049
Gradient Boosting Machine	0,810	492,09	0,00	0,041

Source: Compiled by authors.

5. Conclusions

In this study we evaluated the prediction accuracy of the most popular ML algorithms to nowcast—tracking in real time—the monthly growth rate of Peruvian GDP. The analysis window was between 2008 and 2023 and worked with several leading indicators to assess the dynamic of GDP components measured by way of the expenditure and productive sector approach. Furthermore, we enriched our approach by incorporating a sentiment data index built through Google Trends, which have proven effective in estimating advanced economic activity. The ML approach allowed the use of 91 variables simultaneously, incorporating structured data non-structured data, including a larger dataset for the Peruvian GDP prediction case. The evaluation results and consistency exercise provide evidence that the positive contribution of ML models and sentiment data significantly improve the model accuracy and allow the early detection of periods of high volatility—an aspect that conventional models often fail to capture.

Our results shed light on how ML can outperform AR, stepwise and DFM models in prediction accuracy, which opens up a new agenda for emerging economies to improve the forecasting of relevant macroeconomic variables such as consumption, employment, and investment, among others.

These models have been implemented by the Department of Macroeconomic Projections in the Ministry of Economics and Finance of Peru, perform successfully, and are incorporated into monthly activities; therefore, we would like to suggest three specific outstanding agendas based on our application expertise. First, there is a need to analyze the marginal prediction gains from the inclusion of unstructured data in reducing forecast error, since our results have shown improvements in accuracy. However, a key question arises: Would the period analyzed influence the results? Between 2004 and 2023, which includes high volatility events such as the pandemic, the global financial crisis, and climate shocks in 2017 and 2023, ML models with unstructured data gained in predictive capacity by track daily frequency data from Google Trend searches. This question could be tested by performing a variance analysis of the projection errors by comparing ML models with other more traditional ones during a period of relative normality and other periods of crisis. Second, in the estimates we observed the unsynchronized availability of approximately 45% of the dataset variables (91), which raises the question of whether consistent results would be equally obtained with a smaller number of variables. This proportion could be evaluated in subsequent studies by reducing the software requirements. Third, the treatment of the unstructured data could be improved; in this study we used a simple and didactic management of non-structured data, but monthly weighting of searched words in Google Trends could be considered to smooth the high variability related to this type of data.

References

Araujo, D., Bruno, G., Marcucci, J., Schmidt, R., & Tissot, B. (2023). Machine learning: applications in central banking. *Journal of AI, Robotics & Workplace Automation, 2*(3), 271–293.

Armstrong. (2001). *Principles of forecasting: A handbook for researchers and practitioners* (Vol. 30). Springer.

Aruoba, S. B., Diebold, F. X., & Scotti, C. (2009). Real-time measurement of business conditions. *Journal of Business & Economic Statistics, 27*(4), 417-427.

Athey, S. (2018). The impact of machine learning on economics intelligence: An agenda. In *The economics of artificial* (pp. 507-547). University of Chicago Press.

Bánbura, M., & Modugno, M. (2014). Maximum likelihood estimation of factor models on datasets with arbitrary pattern of missing data. (W. O. Library, Ed.) *Journal of applied econometrics, 29*(1), 133-160.

Bánbura, M., & Rünstler, G. (2011). A look into the factor model black box: Publication lags and the role of hard and soft data in forecasting gdp. *International Journal of Forecasting, 27*(2), 333-346.

Bánbura, M., Giannone, D., Modugno, M., & Reichlin, L. (2013). Now-casting and the real-time data flow. In Elsevier (Ed.), *Handbook of economic forecasting* (Vol. 2, pp. 195-237).

Barrios, J. J., Escobar, J., Leslie, J., Martin, L., & Peña, W. (2021). *Nowcasting para predecir actividad económica en tiempo real: Los casos de Belice y El Salvador.* Inter-American Development Bank.

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003, Jan). Latent dirichlet allocation. *Journal of machine Learning research, 3*, 993-1022.

Boehmke, B., & Greenwell, B. (2020). Chapter 12: Gradient boosting. In *Hands-on machine learning with R.* Chapman & Hall.

Bok, B., Caratelli, D., Giannone, D., Sbordone, A. M., & Tambalotti, A. (2018). Macroeconomic nowcasting and forecasting with big data. *Annual Review of Economics, 10*, 615-643.

Bolivar, O. (2024). Gdp nowcasting: A machine learning and remote sensing data-based approach for Bolivia. *Latin American Journal of Central Banking, 5*(3).

Breiman, L. (2001). Random forests. In *Machine learning* (Vol. 45, pp. 5-32). Springer.

Brownlee, J. (2016). Bagging and random forest ensemble algorithms for machine learning. In Master *Machine learning algorithms* (pp. 4-22). Machine Learning Mastery.

Caruso, A. (2018). Nowcasting with the help of foreign indicators: The case of Mexico. In *Economic Modelling* (Vol. 69, pp. 160-168). Elsevier.

Chakraborty, C., & Joseph, A. (2017). *Machine learning at central banks.* Bank of England working paper.

Corona, F., González-Farías, G., & López-Pérez, J. (2022). Timely estimates of the monthly Mexican economic activity. *Journal of Official Statistics, 38*(3), 733-765.

Diebold, F. X., & Mariano, R. S. (1995). Comparing predictive accuracy. *Journal of Business and Economic Statistics, 13*(3), 253-263.

Döpke, J., Fritsche, U., & Pierdzioch, C. (2017). Predicting recessions with boosted regression trees. *International Journal of Forecasting, 33*(4), 745-759.

Doz, C. G. (2011). A quasi–maximum likelihood approach for large, approximate dynamic factor models. *Review of economics and statistics, 94*(4), 188-205.

Doz, C. G. (2011). A two-step estimator for large approximate dynamic factor models based on kalman filtering. *Journal of Econometrics, 164*(1), 188-205.

Eberendu, A. C., & al., e. (2016). Unstructured data: An overview of the data of big data. *International Journal of Computer Trends and Technology, 38*(1), 46-50.

Einav, L., & Levin, J. (2014). The data revolution and economic analysis. *Innovation Policy and the Economy, 14*(4), 1-24.

Escobal D'Angelo, J., & Torres, J. (2002). *Un sistema de indicadores lideres del nivel de actividad para la economía peruana.*

Etter, R. G., & al., e. (2011). *A composite leading indicator for the peruvian economy based on the bcrp's monthly business tendency surveys (tech. rep.).* Banco Central de Reserva del Perú.

Evans, M. (2005). *Where are we now? real-time estimates of the macro economy.*

Forero, F. J., Aguilar, O. J., & Vargas, R. F. (2016). *Un indicador lider de actividad real para Perú.*

Gálvez-Soriano, O. d. (2020). Nowcasting Mexico's quarterly GDP using factor models and bridge equations. *Estudios Económicos (México, DF), 35*(2), 213-265.

Garcia-Donato, G., & Martinez-Beneito, M. A. (2013). On sampling strategies in bayesian variable selection problems with large model spaces. *Journal of the American Statistical Association, 108*(501), 340-352.

Ghosh, S., & Ranjan, A. (2023). A machine learning approach to GDP nowcasting: An emerging market experience. *Buletin Ekonomi Moneter dan Perbankan, 26*, 33-54.

Giannone, D., Reichlin, L., & Small, D. (2008). Nowcasting: The real-time informational content of macroeconomic data. *Journal of monetary economics, 55*(4), 665-676.

Giglio, S., Kelly, B., & Xiu, D. (2022). Factor models, machine learning, and asset pricing. *Annual Review of Financial Economics, 14*, 337-368.

González-Astudillo, M., & Baquero, D. (2019). A nowcasting model for Ecuador: Implementing a time-varying mean output growth. *Economic Modelling, 82*, 250-263.

Green, K. C., & Armstrong, S. (2015). Simple versus complex forecasting: The evidence. *Journal of Business Research, 68*(8), 1678-1685.

Harvey, D., Leybourne, S., & Newbold, P. (1997). Testing the equality of prediction mean squared errors. *International Journal of Forecasting, 13*(2), 281-291.

Kant, D., Pick, A., & deWinter, J. (2022). *Nowcasting GDP using machine learning methods.* Nederlandsche Bank Working Paper.

Kapsoli Salinas, J., & Bencich Aguilar, B. (2002). *Indicadores lideres, redes neuronales y predicción de corto plazo.* Pontificia Universidad Católica del Perú. Departamento de Economía.

Liu, Z. Z. (2014). *The doubly adaptive lasso methods for time series analysis.* The University of Western Ontario (Canada).

Longo, L., Riccaboni, M., & Rungi, A. (2022). A neural network ensemble approach for gdp forecasting. *Journal of Economic Dynamics and Control*, 134.

Makridakis, S., Spiliotis, E., & Assimakopoulos, V. (2018). Statistical and machine learning forecasting methods: Concerns and ways forward. *PloS one, 13*(3), e0194889.

Martınez, M., & Quineche, R. (2014). *Un indicador lider para el nowcasting de la actividad económica del perú (tech. rep.).* Mimeo.

Medeiros, M. C., Vasconcelos, G. F., Veiga, Á., & Zilberman, E. (2021). Forecasting inflation in a data-rich environment: The benefits of machine learning methods. *Journal of Business & Economic Statistics, 39*(1), 98-119.

Muchisha, N. D., Tamara, N., Andriansyah, A., & Soleh, A. M. (2021). Nowcasting Indonesia's GDP growth using machine learning algorithms. *Indonesian Journal of Statistics and Its Applications, 5*(2), 355-368.

Natekin, A., & Knoll, A. (2013). Gradient boosting machines, a tutorial. *Frontiers in neurorobotics, 7*(21).

Pérez Forero, F. (2018). *Nowcasting Peruvian GDP using leading indicators and bayesian variable selection (tech. rep.).* Banco Central de Reserva del Perú.

Richardson, A., & Mulder, T. (2018). *Nowcasting New Zealand GDP using machine learning algorithms.* CAMA Working Paper.

Romer, C., & Romer, D. (2008). The fomc versus the staff: Where can monetary policymakers add value? *American Economic Review, 98*(2), 230-235.

Rusnák, M. (2016). Nowcasting Czech GDP in real time. *Economic Modelling, 54*, 26-39.

Scott, S. L., & Varian. (2013). *Bayesian variable selection for nowcasting economic time series (tech. rep.).* National Bureau of Economic Research.

Snoek, J., Larochelle, H., & Adams, R. P. (2012). Practical Bayesian optimization of machine learning algorithms. *Advances in neural information processing systems, 25*.

Stock, J. H., & Watson, M. W. (1989). New indexes of coincident and leading economic indicators. *NBER macroeconomics annual, 4*, 351-394.

Suphaphiphat, N., Wang, Y., & Zhang, H. (2022). *A scalable approach using DFM, machine learning and novel data, applied to european economies.*

Tenorio, J., & Pérez, W. . (2023). GDP nowcasting with machine learning and unstructured data to Peru. *Perueconomics*, (No. 197).

Tenorio, J., & Perez, W. (2024). GDP nowcasting with machine learning and unstructured data. (No. 2024-003).

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society Series B: Statistical Methodology, 58*(1), 267-288.

Tiffin, M. A. (2016). *Seeing in the dark: A machine-learning approach to nowcasting in Lebanon.* International Monetary Fund.

Varian, H. (2014). Machine learning and econometrics. *Slides package from talk at University of Washington*.

Woloszko, N. (2020). A weekly tracker of activity based on machine learning and google trends.

Zhang, Q., Ni, H., & Xu, H. (2023). Nowcasting Chinese GDP in a data-rich environment: Lessons from machine learning algorithms. *Economic Modelling, 122*, 106204.

Zou, H. (2006). The adaptive lasso and its oracle properties. *Journal of the American statistical association, 101*(476), 1418-1429.

Zou, H., & Hastie, T. (2005). Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society Series B: Statistical Methodology, 67*(2), 301-320.

6. Appendix

Table 5
Literature on Nowcasting

Author	Year	Methodology	Country
Literature International			
Banbura and others	2013	DFM	Europe
Evans	2005	DFM	US
Giannone and others	2008	DFM	US
Nowcasting with machine learning			
Richardson and others	2018	Various models ML	New Zealand
Giannone and others	2008	DFM	US
Ghosh and Ranjan	2023	Various ML	India
Muchisha and others	2020	Various ML vs DFM	Indonesia
Zhang, Ni and Xu	2023	Various ML	China
Kant and others	2022	Various ML	Netherlands
Suphaphiphat and others	2022	Various ML	Europe
Nowcasting with big data			
Blei, Ng and Jordan	2003	LDA	US
Athey, Mobius and Pal	2017	Google News	Spain
Woloszko	2020	Google Trends	USA
Niesert and otros	2020	Google Trends	Advanced Economies
Peruvian main references			
Escobal and Torres	2002	DFM	Peru
Pérez Forero	2016	DFM	Peru
Kapsoli and Bencich	2002	Neuronal Networks	Peru
Pérez Forero	2018	Bayesian VAR	Peru
Etter and Graff	2011	Surveys	Peru
Martinez and Quineche	2014	Neuronal Networks	Peru

Source: Own elaboration.

Table 6
List of no structured variables included in the model

Unstructured variable details		
Units of Measure	Frequency	Source
Search Index (0 to 100)	Daily	Google Trends
	Variables	
1.- Searched Words on Economic		
Inflation	Recession	
2.- Searched Words on Consumption		
Kia	Toyota	Movies
Restaurants	Credits	Loans
Mortgages	Deals	
3.- Searched Words on Labor Market		
Employment	Unemployment	Labor
4.- Searched Words on Sectorial Industry		
Mining	Investment	
5.- Searched Words on Current Situation		
Peruvian Crisis	Bankruptcy	Economy
Economic Crisis		
6.- Searched Words on Real Estate Market		
Land	Real Estate	
7.- Searched Words on Politics		
Elections		
8.- Searched Words on Tourism		
Travel	Machu Picchu	Flights
Visa	Flights to the US	Accommodations
Hotels	Vacations	
9.- Searched Words on Bonds and Pensions		
Bonds	CTS	AFP
10.- Searched Words on Weather and Natural Phenomena		
Rains	ENSO	Droughts
Frosts	Huaico	

Source: compiled by authors.

Table 7
List of structured variables included in the model (a)

No.	Variable	Units of Measure	Frequency	Source
		Main Indicator		
1	GDP	Index 2007 = 100	Monthly	INEI
		Consumption Indicators		
2	Credit	S/ Millions	Monthly	BCRP
3	Credit	US$ Millions	Monthly	BCRP
4	Credit (constant exchange rate)	S/ Millions	Monthly	BCRP
5	Consumer credits	S/ Millions	Monthly	BCRP
6	Mortgage Loans	S/ Millions	Monthly	BCRP
7	Deposits	S/ Millions	Monthly	BCRP
8	Deposits	S/ Millions	Monthly	BCRP
9	Sales of chickens	Metric Tons	Dayly	MIDAGRI
10	Consumer Confidence Index	Points	Monthly	Apoyo Consultoria
		Activity Indicators		
11	Electricity Production		Monthly	INEI
12	Hydrocarbon Production		Dayly	MINEM
13	3-Month Economic Expectations	Points	Monthly	BCRP
14	Oil	B/D	Dayly	MINEM
15	Natural Gas	MCF	Dayly	MINEM
		Investment Indicators		
16	Domestic Cement Consumption	Index	Weekly	INEI
17	Import of Intermediate Inputs	Index	Weekly	INEI
18	Import of Capital Goods	Index	Weekly	INEI
		Labor Market Indicators		
19	Employed Labor Force	Thousands	Monthly	INEI
20	Properly Employed Population[1]	Thousands	Monthly	INEI
		Public Investment Indicators		
21	Non-Financial Gov. Expenditures	S/ Millions	Monthly	BCRP
22	IAFO	Index	Monthly	INEI
		Foreign Trade Indicators		
23	Volume of Imported Inputs	Index	Monthly	INEI
24	Terms of Trade	Index	Monthly	BCRP
25	IPX	Index	Monthly	BCRP
26	IPM	Index	Monthly	BCRP

		Financial Indicators		
27	General Stock Market Index[2]	Percentages	Dayly	Bloomberg
28	Liquidity	Millions of Soles	Monthly	BCRP
		Monetary Indicators		
29	CPI	Index	Monthly	INEI
30	Non Food and Energy Price Index	Index	Monthly	BCRP
31	Wholsale Price Index	Index	Monthly	BCRP
32	Core CPI	Index	Monthly	BCRP

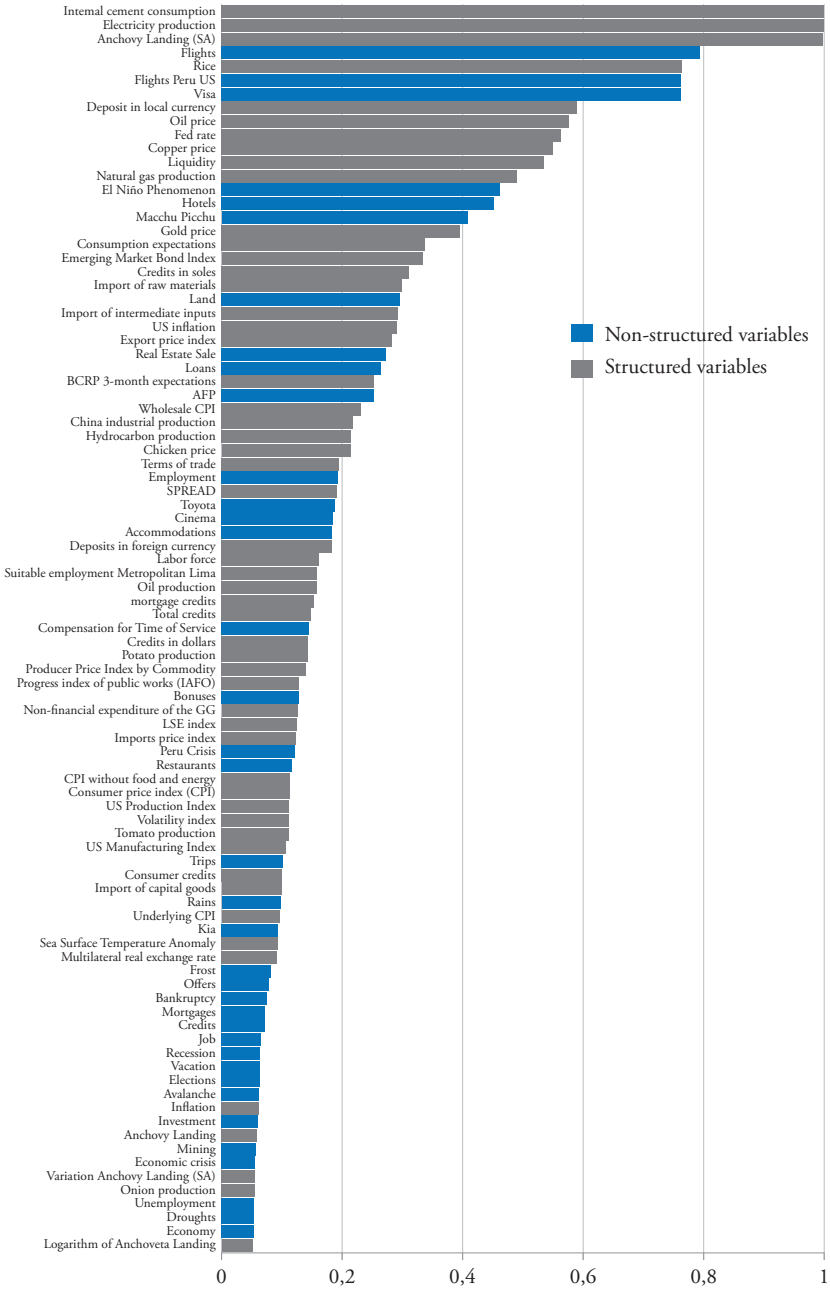
1/ Metropolitan Lima. 2/ Lima
Source: compiled by authors.

Table 8
List of structured variables included in the model (b)

No.	Variable	Units of Measure	Frequency	Source
		International Indicators		
33	Multilateral Real Exchange Rate	(2009=100)	Monthly	BCRP
34	EMBIG Perú	Pbs	Dayly	BCRP
35	Oil WTI	Dollars per Barrel	Dayly	Bloomberg
36	USIPC	Index	Monthly	FRED
37	Industrial Production Index	YoY	Quarterly	Bloomberg
38	Copper	cUS$/lb.	Dayly	Bloomberg
39	Gold	US$/oz. tr.	Dayly	Bloomberg
40	US Manufacturing PMI	Points	Monthly	Bloomberg
41	FED Interest Rate (Upper Limit)	Percentages	Monthly	Bloomberg
42	VIX Index	Percentages	Dayly	Bloomberg
43	Spread 2Y-5Y		Monthly	Bloomberg
44	China Industrial Production	YoY	Monthly	Bloomberg
45	PPI by All Commodities	(1982=100)	Monthly	FRED
		Climate Indicators		
46	ATSM	Degrees Celsius	Monthly	IMARPE
		Fisehry Indicators		
47	Anchoveta Landing	Metric Tons	Dayly	IMARPE
48	Logarithm of Anchoveta Landing		Dayly	Own elaboration
49	Anchoveta Landing[1]		Dayly	Own elaboration
50	Variation Anchoveta Landing[2]		Dayly	Own elaboration
		Agricultural Indicators		
51	Paddy Rice production	Tons	Monthly	MIDAGRI
52	Potato production	Tons	Monthly	MIDAGRI
53	Onion production	Tons	Monthly	MIDAGRI
54	Tomato production	Tons	Monthly	MIDAGRI

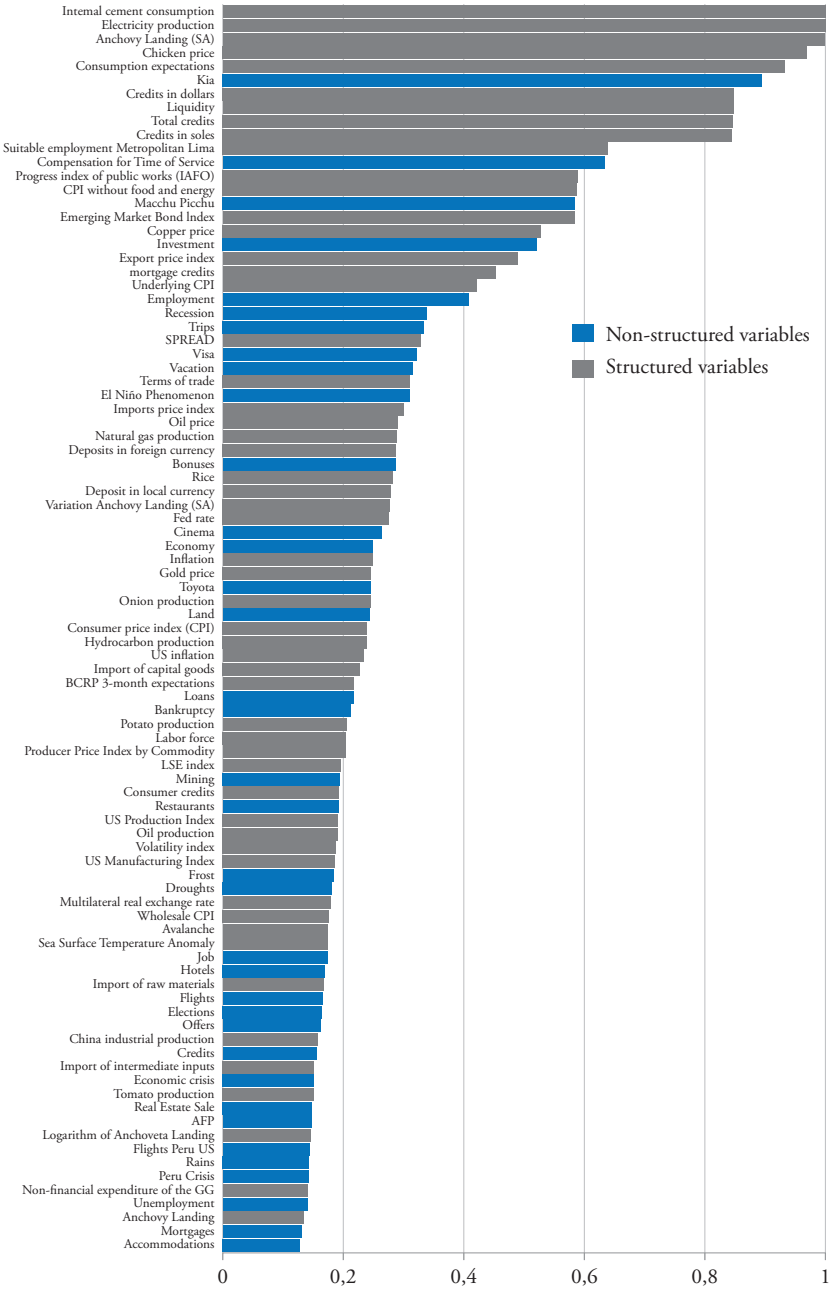
1/ Seasonally Adjusted. 2/ Seasonally Adjusted.
Source: compiled by authors.

Figure 4
Gibb sampling (2004-2023) - probability of inclusion in optimal model



Source: compiled by authors.

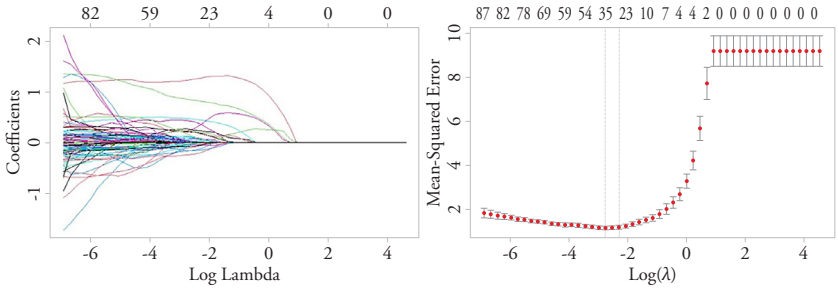
Figure 5
Gibb sampling (2004-2019) - probability of inclusion in optimal model



Source: compiled by authors.

Figure 6
LASSO Optimal Parameters

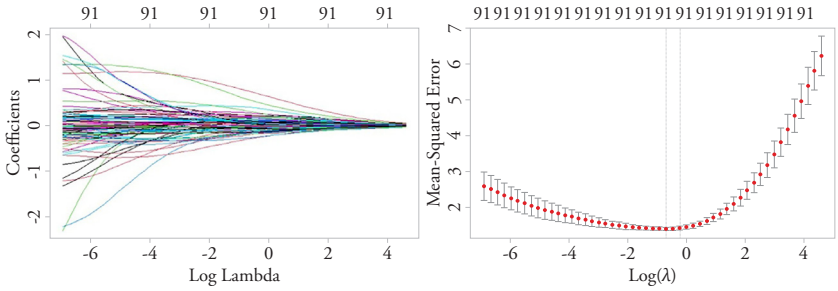
(a) Coef. (e) by hyperparameter range (b) Optimal hyperparameter vs. MSE



Source: compiled by authors.

Figure 7
Ridge Optimal Parameters

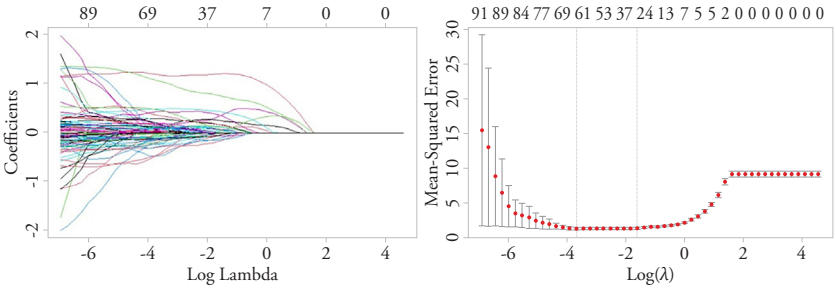
(a) Coef. (e) by hyperparameter range (b) Optimal hyperparameter vs. MSE



Source: Own elaboration

Figure 8
Elastic Net Optimal Parameters

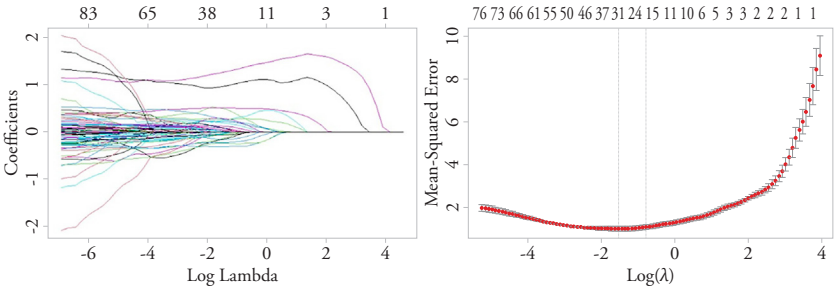
(a) Coef. (e) by hyperparameter range (b) Optimal hyperparameter vs. MSE



Source: compiled by authors.

Figure 9
Adaptive LASSO Optimal Parameters

(a) Coef. (e) by hyperparameter range (b) Optimal hyperparameter vs. MSE



Source: compiled by authors.

Figure 10
Dynamic correlations of the structured variables

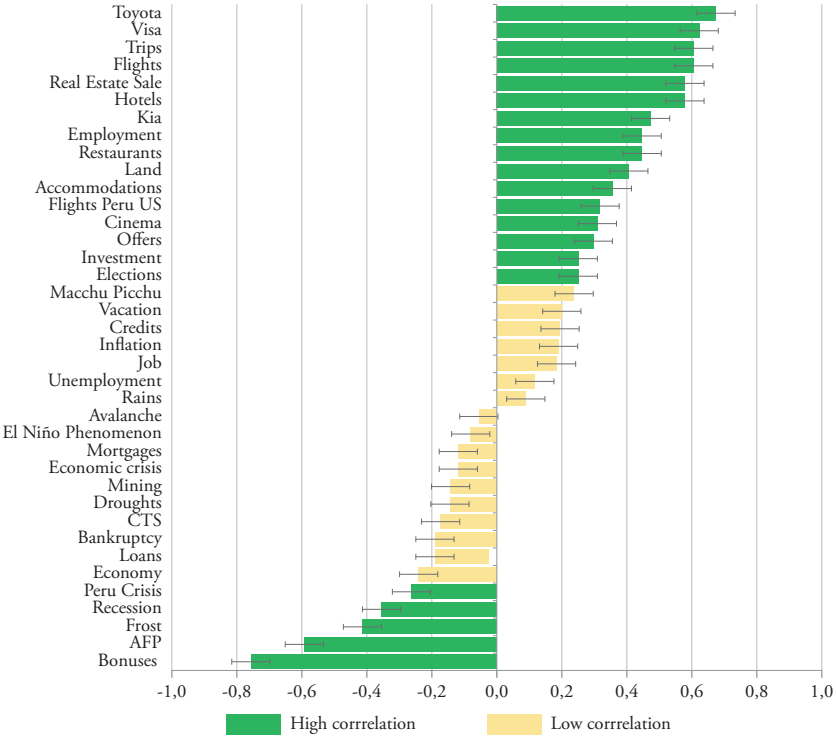
N°	Variables	Time						
		t-3	t-2	t-1	t	t+1	t+2	t+3
1	Electricity production	0.5	0.6	0.6	0.8	0.6	0.5	0.4
2	Imports of capital goods	0.5	0.6	0.7	0.7	0.7	0.8	0.7
3	Deposit in local currency	0.6	0.6	0.6	0.7	0.7	0.7	0.6
4	Internal cement consumption	0.6	0.6	0.6	0.7	0.5	0.5	0.4
5	Import of raw materials	0.6	0.6	0.6	0.7	0.5	0.5	0.4
6	Consumer credits	0.4	0.4	0.5	0.6	0.6	0.7	0.7
7	Producer Price Index by Commodity	0.5	0.5	0.6	0.6	0.6	0.5	0.5
8	Price index of imports	0.5	0.5	0.6	0.6	0.5	0.5	0.4
9	Total credits	0.3	0.4	0.5	0.6	0.6	0.6	0.6
10	Wholesale CPI	0.2	0.3	0.4	0.5	0.6	0.6	0.6
11	Natural gas production	0.3	0.4	0.4	0.5	0.5	0.5	0.5
12	China industrial production	0.6	0.6	0.6	0.5	0.4	0.4	0.3
13	Credits in foreing currency	0.4	0.4	0.5	0.5	0.5	0.5	0.5
14	Consumption expectations	0.6	0.6	0.5	0.5	0.4	0.3	0.2
15	US inflation	0.4	0.4	0.5	0.5	0.5	0.4	0.4
16	Import of intermediate inputs	0.6	0.5	0.5	0.5	0.4	0.3	0.3
17	Liquidity	0.2	0.2	0.3	0.5	0.5	0.6	0.6
18	Oil price	0.5	0.5	0.5	0.4	0.4	0.3	0.3
19	Gold price	0.5	0.4	0.4	0.4	0.3	0.3	0.2
20	Export price index	0.5	0.5	0.4	0.4	0.3	0.2	0.1
21	Suitable employment Metropolitan Lima	0.4	0.4	0.4	0.4	0.3	0.3	0.2
22	US Production Index	0.4	0.4	0.4	0.4	0.3	0.2	0.2
23	US Manufacturing Index	0.4	0.4	0.4	0.4	0.3	0.2	0.2
24	Mortgage credits	0.3	0.3	0.3	0.4	0.4	0.4	0.5
25	Non-financial expenditure of the GG	0.4	0.4	0.3	0.3	0.2	0.1	0.1
26	Labor force	0.4	0.4	0.4	0.3	0.3	0.2	0.1
27	Credits in soles	0.1	0.2	0.3	0.3	0.4	0.4	0.4
28	Progress index of public works (IAFO)	0.1	0.2	0.1	0.3	0.2	0.2	0.2
29	Anchovy Landing (SA)	0.1	0.0	0.1	0.3	0.1	0.1	0.0
30	Copper price	0.4	0.4	0.3	0.3	0.2	0.1	0.0
31	Consumer price index (CPI)	-0.1	0.0	0.1	0.2	0.3	0.4	0.4
32	Hydrocarbon production	-0.1	0.0	0.0	0.2	0.1	0.2	0.2
33	Volatility index	0.0	0.1	0.1	0.2	0.2	0.3	0.3

34	Anchovy Landing	0.0	0.0	0.0	0.2	0.1	0.1	0.0
35	Tomato production	0.1	0.1	0.2	0.1	0.2	0.2	0.1
36	Logarithm of Anchoveta Landing	0.1	0.1	0.1	0.1	0.1	0.1	0.1
37	Terms of trade	0.3	0.3	0.2	0.1	0.0	-0.1	-0.2
38	Rice	-0.1	-0.1	0.0	0.1	0.0	0.0	0.1
39	Fed rate	0.0	0.0	0.1	0.1	0.0	0.0	0.0
40	Onion production	0.2	0.2	0.1	0.0	0.1	0.0	0.0
41	Oil production	-0.2	-0.2	-0.1	0.0	-0.1	0.1	0.1
42	Chicken price	0.1	0.1	0.1	0.0	0.0	-0.1	-0.1
43	Potato production	0.0	0.0	0.0	0.0	0.0	-0.1	0.0
44	LSE index	0.2	0.1	0.1	0.0	-0.1	-0.1	-0.2
45	Emerging Market Bond Index	0.2	0.1	0.0	-0.1	-0.1	-0.2	-0.2
46	Consumption expectations	0.1	0.0	-0.1	-0.1	-0.2	-0.2	-0.2
47	Underlying CPI	-0.3	-0.2	-0.2	-0.1	0.0	0.1	0.2
48	Multilateral real exchange rate	-0.2	-0.2	-0.1	-0.1	-0.1	-0.1	-0.1
49	Variation Anchovy Landing (SA)	-0.1	-0.1	-0.2	-0.1	-0.2	-0.1	-0.2
50	SPREAD	-0.1	-0.2	-0.1	-0.1	-0.1	-0.1	-0.1
51	Deposits in foreign currency	-0.3	-0.3	-0.3	-0.2	-0.1	0.0	0.0
52	Sea Surface Temperature Anomaly	-0.2	-0.2	-0.2	-0.2	-0.2	-0.2	-0.2
53	CPI without food and energy	-0.6	-0.6	-0.5	-0.5	-0.4	-0.3	-0.2

High correlation Low correlation

Source: compiled by authors.

Figure 11
Correlations of the main nonstructured variables



Source: compiled by authors.