GDP nowcasting with Machine Learning and Unstructured Data [*]

Juan Tenorio[†] Wilder Perez[‡]

January 2024

Abstract

In a context of ongoing change, "nowcasting" models based on Machine Learning (ML) algorithms deliver a noteworthy advantage for decision-making in both the public and private sectors due to their flexibility and ability to drive large amounts of data. This document introduces projection models designed for real-time forecasting of the monthly Peruvian GDP growth rate. These models integrate structured macroeconomic indicators with high-frequency unstructured sentiment variables. The analysis spans from January 2007 to May 2023, encompassing a comprehensive set of 91 leading economic indicators. Six ML algorithms were rigorously 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, Gradient Boosting Machine, LASSO, and Elastic Net emerged as standout performers, demonstrating a prediction error reduction of 20% to 25% when contrasted with AR and various specifications of DFM. These results could be influenced by the analysis period, which includes crisis events featuring high uncertainty, where ML models with unstructured data improve significance.

Clasification JEL: C32, C53, E37, C52, E32.
Key Words: nowcasting, machine learning, GDP growth.

[*] The opinions expressed in this document, as well as any errors and omissions, are the sole responsibility of the authors 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, staff of the General Directorate of Macroeconomic Policy of the Ministry of Economy and Finance and comments from the attendees of the XLI Meeting of Economists of Central Reserve Bank of Peru 2023.

[†] Ministry of Economy and Finance, and UPC; jtenorio@mef.gob.pe and pcefjten@upc.edu.pe

[‡] Ministry of Economy and Finance, and UCS; wperez@mef.gob.pe and wperezc@cientifica.edu.pe

1 Introduction

Making decisions in real-time is a true challenge for policymakers, given that the primary barrier they face is the usual delay in the availability of updated information about macroeconomic aggregates. In most cases, the economic variables show a delay of between 30-45 days on average, including the time for revisions and retrospectives. Nevertheless, the continuous stride forward in the new generation of high-frequency data has changed how prediction models face the uncertainty inherent in this information. As a result, in the past few years, both central banks and international institutions have adopted methodological focuses that incorporate machine learning, and take advantage of the abundant quantity of data that come from search engines and social media as is shown in Richardson and Mulder (2018); Chakraborty and Joseph (2017); Araujo et al. (2023).

These automated learning techniques have gained great popularity in comparison with the conventional focus of traditional time series models that project macroeconomic variables. A characteristic of these algorithms that is often highlighted resides in their capacity to formulate parametric selections in large amounts of data sets, which find their base in training a specific percentage of the model's information. Hence, the objective of this document is to explore the benefits of utilizing several machine learning methodologies. This will be done 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) the monthly Peru's real GDP (Gross Domestic Product) rate growth. The data set consists of both local and international variables, which can be broken down into 53 structured variables in 38 nonstructured variables, giving a total of 91 predictors. These predictive variables are examined according to the model, to evaluate the optimum performance of each variable between September 2014 and May 2023. Furthermore, following C. Romer and D. Romer (2008), we perform an analysis of the evaluation of predictive accuracy using two models as reference: the traditional autoregressive time series, and a dynamic factor model, based on the leading indicator of production of electricity used by the economic literature, which is commonly used by authorities in the political and economic consulting Peruvian firms. This will facilitate an exhaustive evaluation of the performance of machine learning algorithms.

The results indicate that immediate predictions of the machine learning models are more solid in comparison with the benchmark auto-regressive model and display a better performance compared to DFM. Specifically, the Random Forest, Gradient Boosting Machine, and Adaptive Lasso show performance with a superior ability to reduce the average error of projection in a range from 20%-25%. Additionally, it is corroborated that following the methodology proposed by Armstrong (2001), the use of the average value of projection of all the machine learning algorithms, adds a significant value to the RMSE, which positively contributes to a more precise prediction of GDP. Even though other methodologies, such as *Ridge*, *LASSO* and *Elastic Net* do not reach the same level of predictive ability as the previously mentioned ML methodologies, they still outperform the control model in terms of performance. Further, the proof of forecasting evaluation and consistency assessment confirms that most of the machine learning models improve the prediction significantly which is in line with previous literature applied in other contexts

(Richardson and Mulder, 2018; H. Varian, 2014; Q. Zhang, Ni, and Xu, 2023).

This research document adds itself to the existing literature that highlights the success of machine learning applications in contrast to more traditional methodologies. However, given the lack of evidence in Latin America[1], and in particular in Peru[2], surrounding the utilization of these algorithms in conjunction with non-structured data, this research project also highlights the need to bring to the forefront of the discussion on what these models entail. Barrios et al. (2021), Richardson and Mulder (2018) and Döpke, Fritsche, and Pierdzioch (2017) have shown through the implementation of diverse machine learning algorithms that these method's results are more adequate in carrying out forecasts in real-time when a large amount of information is available to the forecaster. For example, Longo, Riccaboni, and Rungi (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 has demonstrated a substantial decrease in the forecast error, also showing a notable capability of capturing the period of recession caused by the COVID-19 pandemic and the economic recuperation that came thereafter. Similarly, in the case of El Salvador and Belize, Barrios et al. (2021) implemented a large array of machine, learning methods to forecast the quarterly growth of GDP, using a large amount of predictive variables. The results of this research study concluded that the application of these tools represents a robust alternative to prediction, and its benefits suggest a recommendation for its use in other countries in the region. Additionally, other researchers have extended the application of the machine, learning models further from GDP, including forecasting, inflation, yield curve, and active prices. These efforts have yielded notable results in precise forecasting (Medeiros et al., 2021; Giglio, Kelly, and Xiu, 2022).

It is still important to highlight that these methods present challenges in their implementation, which have led to some major debates surrounding the topic. In fact, Green and S. Armstrong (2015), as well as Makridakis, Spiliotis, and Assimakopoulos (2018), when comparing multiple models of machine learning, find that the deposition of the forecasting is less significant in comparison with the statistical smoothing approaches, and the ARIMA models. These authors warn that the computational complexity that is inherent to the selection and use of variables in the machine learning model makes immediate forecasting difficult and less practical for policymakers.

Finally, the rest of this document has the following sections. Initially, a literature review is carried out that explores the relevance of the nowcasting methodology in the context of machine learning and big data, both at the national and international levels. After this, a section dedicated to the methodology is presented in which details are given about the models that were utilised and the data sets. Afterwards, the results are displayed in a specific section, followed by the robustness tests and the conclusion.

[1] See Barrios et al. (2021).

[2] See Escobal D'Angelo and Torres (2002); Pérez Forero (2018).

2 Literature review

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

To address the issue of extended delays in the publication of key economic aggregates, the concept of nowcasting was proposed. This approach aims to predict the present, the very near future and the very recent past (Bańbura, Giannone, et al., 2013). A traditional reference nowcasting model is the Dynamic Factor Model, widely used in central banks to predict GDP (Giannone, Reichlin, and Small, 2008; Bańbura and Rünstler, 2011; Bok et al., 2018; Rusnák, 2016; González-Astudillo and Baquero, 2019). Two seminal articles have formalized this process into statistical models. On one side, Giannone, Reichlin, and Small (2008) proposed a methodology to assess the marginal impact of the publication of monthly-updated data on forecasts of quarterly-published real Gross Domestic Product (GDP) growth. The method presented by these authors was able to track the real-time flow of information that central banks monitor by handling 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) do real-time estimations of the current state of the US economy. This approach included data complexity and provided useful information about the relationship between macroeconomics and asset prices. The author models monthly time series with a dynamic factor model (Dynamic Factor Model - DFM) in a state space system. Once the state space representation is settled, Kalman filter techniques are estimated to make GDP forecasting, as they automatically adapt to changes according to the data available. To this document concern, we perform DFM specifications as benchmark models following Evans (2005) proposal and the implementation suggestions of Doz, Giannone, and Reichlin (2012).

An additional bright side of the nowcasting models is the constant improvement experienced from wider information availability and data frequency heterogeneity (Q. Zhang, Ni, and Xu, 2023; González-Astudillo and Baquero, 2019). Thus recently, machine learning methods are been incorporated enhances to the nowcasting approach. The algorithms of machine learning (ML) deliver better performance in handling large amounts of data, capturing non-linear relationships and adapting to changing economic conditions.

Those methods provide more accurate predictions by incorporating various variables and sources of unstructured data. As described Athey (2018), these techniques are divided into two main brands, supervised and unsupervised ML. Athey (2018) explains that unsupervised MLs are looking for 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 a source of information, in techniques such as grouping *k-medias*. For instance, Blei, Ng, and Jordan (2003) applied pooling models to find "topics" in textual data. Another example is the paper written by Woloszko (2020). Here, the author shows a weekly indicator of the economic activity for 46 OCDE countries and the G20 using search data from Google Trends. This document showcases the power of prediction of specific "topics", including "bankruptcies", "economic crises", "investment", "baggage" and "mortgages". Calibration is performed using a neural network that captures nonlinear patterns, which are shown to be consistent with economic intuition using ML Shapley values interpretation tools. On the other side, the supervised ML algorithms as pointed out by H. Varian (2014) implies the use of a group of variables features or variables to predict a specific indicator result. There is a variety of supervised ML methods regressions such as *LASSO, Ridge, Elastic Net, Random Forest, Regression Trees, Support Vector Machines, Neural Nets, and Matrix Factorisation*, among others as *Model Averaging*.

Several studies highlight the advantages of supervised ML models to forecast macroeconomic series that overcome traditional methods. An application is the research of Ghosh and Ranjan (2023), who present a compilation of Machine Learning techniques and conventional time series methods to predict the Indian GDP. They estimate the ML in the DFM context with financial and economic uncertainty data. They estimate Random Forest and Prophet models along with conventional time series models such as ARIMA to nowcast Indian GDP, where hybrid models stand out. Likewise, the results from Richardson and Mulder (2018) showed better performance of the Ridge regression model to the nowcast GDP of New Zealand over a Dynamic Factor Model. Muchisha et al. (2021) built and compared ML models to forecast the GDP of Indonesia. They evaluate six ML algorithms: Random Forest, LASSO, Ridge, Elastic Net, Neural Networks and Support Vector Machines. They used 18 variables between 3Q2013 and 4Q2019. Their results make clear the outstanding performance of ML than auto-regressive models, especially the Random Forest model. Also, Q. Zhang, Ni, and Xu (2023) test ML, DFM and static factor and MIDAS regression models to nowcast the GDP rate growth of China. They find superior accuracy of ML compared to DFM. The ML model that deserved more attention was Ridge Regression, which overcame the others based on prediction and early anticipation of crises such as the global financial crisis and COVID-19. Kant, Pick, and Winter (2022) compares models to the Netherlands economy between 1992 and 2018, where Random Forest algorithms stood out. Suphaphiphat, Wang, and H. Zhang (2022) uses novel variables such as Google Search and air quality. They run standard DFM and ML to European economies during normal times and crises. They show that most MLs significantly outperform the AR(1) reference model. They highlight that DFM tends to perform better in normal times, while many of the ML methods have excellent performance in identifying turning points. Moreover, ML can predict adequately in very disparate economies. Moreover, Barrios et al. (2021) assesses adjusted machine learning models to Belize and El Salvador economies where ML delivers good predictions, proving the effectiveness of ML algorithms in very different country contexts.

Another relevant aspect is Big Data due to its benefits of broadening the range and use of available data that can provide some valid information on the behaviour of the economy to anticipate certain economic indicators (Einav and Levin, 2014). As mentioned in Eberendu et al. (2016), the digital era has allowed the emergence of news channels and social network technologies, mobile phones and online advertising. Nevertheless, the source of new types of data without a pre-fixed format raises new challenges. This data is available in formats like text, XML, email, images, videos, etc. Eberendu et al. (2016) gives and general description of this type of data. Some studies show relevant results on the use of these techniques. For instance, H. Varian (2014) indicate how the search related to the "initial claims for unemployment" in Google Trends are good candidates to forecast unemployment, CPI and consumer confidence in countries such as the US, UK, Canada, Germany and Japan. They focus on immediate out-of-sample forecasting and extend the Bayesian structural time series model using the Hamiltonian sampler for variable selection. These authors obtain good results for unemployment, while for CPI or consumer confidence not so good.

Previous works applied to the Peruvian economy are focused on the anticipated estimation of monthly GDP growth based on a set of leading indicators (structured data). However, a scarce application of machine learning models and the inclusion of unstructured data in GDP forecasting is evident. In particular, Escobal D'Angelo and Torres (2002) built a joint leading indicator that allows the tracking of Peruvian GDP with only 14 variables. On the other hand, Kapsoli Salinas and Bencich Aguilar (2002) perform forward GDP estimation with a nonlinear neural network model. Additionally, Etter, Graff, et al. (2011) propose a leading indicator with the expectations survey conducted by the Central Bank of Peru (BCRP). Also, Martınez and Quineche (2014) forecast the growth rate of GDP based only on the electric production indicator. Following to Aruoba, Diebold, and Scotti (2009), Forero, Aguilar, and Vargas (2016) propose a leading indicator of Peruvian economic activity. This indicator is obtained as a common unobservable component that explains the co-movement among six variables: electricity production, domestic cement consumption, adjusted domestic IGV, chicken sales, mining metal production and real GDP. Finally, Pérez Forero (2018) try to solve the difficulties about best leading indicators selection under the approach of H. Varian (2014). Perez Forero estimated a stade state system through the Bayesian Gibbs-Sampling methods and the spike-and-slab to the stochastic selection variables (SSVS) and calculated the probability of the inclusion of a large set of variables in the best model to predict GDP.

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) are utilized.

3.1 Autoregressive Model (AR)

As a starting point for our reference, we establish an autoregressive AR model for the 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:

$$y_t = \beta_0 + \beta_1 y_{t-1} + e_t \tag{1}$$

where β_0 is a constant term, β_1 is a parameter, and e_t is a term that represents the error and captures the randomness of the model.

3.2 Dynamic Factor Model (DFM)

DFMs are estimated in the form of state-space systems and can be estimated using the Kalman filter and various types of algorithms. One of the most popular in the economic literature is the Expectation Maximization algorithm, due to its robust numerical properties following the proposal by Doz, Giannone, and Reichlin (2011), which is an efficient estimation to bigger datasets.

The canonical reference DFM can be described as follows:

$$x_t = C_0 f_t + e_t \qquad e_t \sim N(0, R) \tag{2}$$

$$f_t = \sum_{j=1}^{p} A_j f_{t-j} + u_t \qquad u_t \sim N(0, Q_0) \tag{3}$$

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

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

C_0: a matrix of $n \times r$ observable time series with lag j.

Q_0: a matrix of $r \times r$ state covariances.

R: a matrix of $r \times r$ measurement covariances. This matrix is diagonal under the assumption that all covariances between the series are explained by the factors $E[x_{it} \mid x_{-i,t}, f_t] = c_{0i} f_t, \forall i$, where c_{0i} is the $i - th$ row of C_0.

This model can be estimated using a classical form of the Kalman Filter and the Maximum Likelihood estimation algorithm, after transforming it into a State Space model. In a VAR expression, it would be as follows:

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

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

As a benchmark model, we use the efficient estimation of a Dynamic Factor Model via the EM Algorithm - on stationary data with time-invariant system matrices and classical assumptions while permitting missing data following the approach of Bańbura and Modugno (2014).

3.3 Penalized Regression Models

These methodologies are employed to optimize the selection of predictor variables and control the model's complexity, which is crucial in preventing overfitting in high-dimensional settings. The literature suggests various 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 impacts the inclusion and relevance of variables in the final model.

3.3.1 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, preventing 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^2 in the penalty prevents the coefficients from reaching large values, thereby contributing to stability and reducing the risk of overfitting.

3.3.2 LASSO Regression

The LASSO (Least Absolute Shrinkage and Selection Operator) 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 has the property of forcing some coefficients to exactly reach zero, resulting in the automatic selection of a subset of more relevant predictor variables and the elimination of less significant ones. The Lasso coefficients $\hat{\beta}^{Lasso}$ are estimated:

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

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.

3.3.3 Elastic Net Regression

The Elastic Net model appropriately combines the constraints of both the LASSO and Ridge models. In particular, Zou and Hastie (2005) mention 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 values of the coefficients and the sum of the squares of the coefficients of the predictor variables. The equation for estimating the coefficients $\hat{\beta}^{Enet}$ is expressed as:

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

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.

3.3.4 Adaptive Lasso Regression

Following Zou (2006), the Adaptive LASSO model is a variant of the LASSO model that introduces a penalty approach which adaptively adjusts 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), indicate that this process can be efficiently performed using the LARS algorithm. The equation for the Adaptive LASSO model ($\hat{\beta}^{AdL}$) is expressed as:

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

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.

3.4 Decision Tree Models

Decision Tree models are machine learning algorithms that represent decisions and actions in the form of a tree. In this 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. Additionally, the use of these trees allows for an improvement in variable selection by handling non-linear relationships in the model.

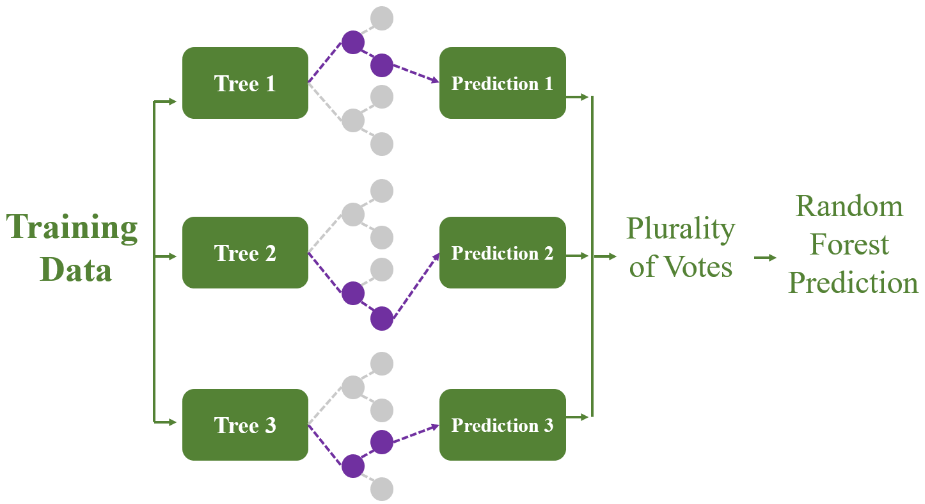
3.4.1 Random forest

This method is based on constructing decision trees using variables from a matrix X and a random selection of features. Additionally, 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 summary, this method recursively divides the data in X into optimized regions and uses variable-based criteria to forecast the target variable, then calculates the dependent variable as the average of these regions.

$$\hat{f}(x) = \sum_{m} \hat{c}_m I(x \in X_r); \hat{c}_m = avg(y_i | x_i \in X_r) \tag{10}$$

The algorithm has certain advantages, such as being efficient 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 has disadvantages like difficulty in interpreting results beyond predictions and a computationally intensive demand for training and hyperparameter tuning. Therefore, for this model, it was necessary to fine-tune it through cross-validation, achieving better performance on unseen data.

Figure 1: Simple Representation of the Random Forest Algorithm



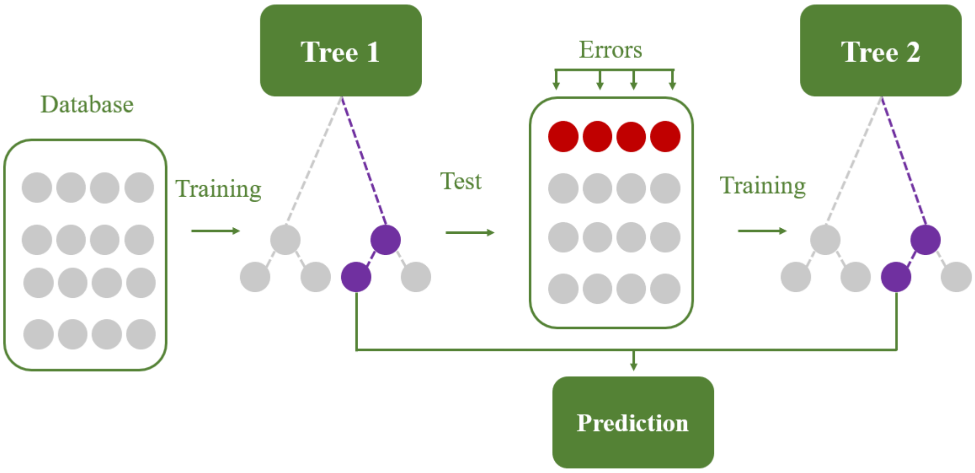
Source: Own elaboration

3.4.2 Gradient Boosting Machine

This algorithm builds a sequence of decision trees, where 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[3] in a way that provides a progressive improvement in regression performance compared to the initial model (Natekin and Knoll, 2013).

[3] Brownlee (2016), indicates that weak models do not necessarily mean they are better than accurate models, as they have the advantage of being able to correct the overfitting problem.

Figure 2: Simple Representation of the Gradient Boosting Machine Algorithm



Source: According to Boehmke and Greenwell (2020)

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. Boehmke and Greenwell (2020), mentions that this method can be summarized by the following equation.

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

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

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

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

3.5 Data

The model's 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, chicken sales, consumer surveys, and local activity indicators, including electricity production, hydrocarbons, economic expectations, and others. Investment indicators are also incorporated, such as 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 labour 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, labour 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 massive search engines, such as Google, stands out as a powerful tool for providing real-time information. Scott and Varian (2013) has 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 estimate 10 groups (See Table 6) of variables that aim to track Google search queries, which are updated daily and can be downloaded from Google Trends. The selection of these words (variables) aims to convey different aspects of the economy, such as the consumption-related group, which is constructed based on searches for words like "Kia", "Restaurants", "Toyota", "Credits", "Loans", "Deals", "Mortgages", and "Cinema". Once this textual data is converted into numerical data, the inclusion of these series is evaluated in the estimations of an optimal model using Gibbs sampling following the findings of Garcia-Donato and Martinez-Beneito (2013) and using 50,000 iterations, an initial burning of 1,000 iterations, and constant beta priors (see Figure 10). This indicates that there is a high relevance of the group of unstructured variables such as the search frequency for "flights", "peruflight_us", "visa", or "El Niño", which would reflect the dynamics of tourism and climatic conditions, among others. Furthermore, we compare the results of this estimation with another one by reducing the sample to 2019 (see Figure 11), where 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). Additionally, a contemporaneous correlation analysis of these variables against the monthly GDP is also performed, obtaining that more than 50% of the unstructured sample correlates greater than 30%.

The data frequency ranges from daily to monthly records in constructing the model. Each variable was assessed 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 allows us to maintain a common reference framework and ensure that different variables contribute equitably to the model.

Ultimately, we have a total set of 91 predictors spanning from January 2008 to May 2023. The evaluation and selection of optimal predictors will be conducted independently for each machine learning algorithm employed. We will specify how we handle the data for the forecast update process in section 4.1 and how we test the model accuracy comparison in section 4.2. This approach will enable us to refine the process of choosing the most efficient prediction model, thereby achieving enhanced performance.

3.6 Strategy of the forecast evaluation

The method that will assess the accuracy in the projection of each model will be done through the root mean square error (RMSE), following the equation:

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

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 will employ the method proposed by Diebold and Mariano (1995) to determine if the projections generated by each machine learning model significantly differ from the *benchmark* model.

4 Results

This section begins by providing 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 hyperparameters calibration

To estimate machine learning models, the selection of hyperparameters plays a crucial role in terms of efficiency and accuracy.

The optimal determination of these values requires the split of the sample data into three parts: i) a training set, ii) a validation set, and iii) a testing set. Initially, the model is estimated with the training set (in-sample) which turns out 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 validation of the ML model in 5 folds, by using every partition or fold as the validation set and the others as the training set on each iteration. Hence, we obtain 5 performance metrics, one by each fold, which are averaged. Also, to identify the optimal hyperparameters, we will run the cross-validation Bayesian optimization algorithm, following closely Snoek, Larochelle, and Adams (2012).

Then, the search process of the optimal values that minimize the mean quadratic error of projections (*MSE*[4]), through cross-validation techniques. The cross-validation consists of forecasting the growth (y_{t+h}) to time with the available data at time t ($y_{t+h}|I_t$)[5], with the hyperparameters obtained by each fold [6]. Once it is identified the optimal values, the accuracy of the model is assessed in the testing set (*out-sample*), evaluating the MSE between the projection growth with the available at time t ($y_{t+h}|I_t$) and with the available data at time $t+h$ ($y_{t+h}|I_{t+h}$). This is repeated to reach the minimization of the value of the

[4] 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.

[5] **I** is the available information set where we get the full available data of 91 predictors variables.

[6] The h can be interpreted as the horizon to forecast, where in a nowcasting context it usually is $h = 1$.

MSE as shown in figures 6 to 9 for each type of ML model[7].

In addition, to prevent overfitting in the ML models the hyperparameters are bounded within ranges recommended by the reviewed literature (See Zou and Hastie (2005)). This approach contributes 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.

Table 1: The strategy of testing estimations

Training dataset					**Testing set**
2008m1-2014m08					2014m09-2023m5
		⟷			↔
Fold 1	Fold 2	Fold 3	Fold 4	Fold 5	

Source: Own elaboration

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	0.1 to 0.9	0.500
	Lambda	0.01 to 0.09	0.040
Adaptive Lasso	Lambda	0.01 to 0.09	0.670
	Omega	0.1 to 0.9	0.340
Random Forest	# Trees	1 to 400	281
Gradient	# Trees	1 to 5000	19
Boosting	Distribution	Normal	Bernoulli
Machine	Shrinkage	0.001 to 0.009	0.300

Source: Own elaboration

4.2 Model comparison

A comparison of the prediction performance of the ML and benchmark models for the test set from September 2009 to May 2023 is presented in Table 3. In terms of the forecast evaluation using the RMSE, the ML models manage to significantly minimize the projection error in comparison with the benchmark AR model and the three different specifications of dynamic factor models [8]. Every projection model compares the forecast with the full available data set at time $t+h$ with the actual GDP rate growth at time $t+h$.

[7] In case of the partial availability of the information set or not full available data of 91 predictors variables, the lack variables could be estimated by others techniques such as DFM with the modified EM algorithm of Bańbura and Modugno (2014) which also accounts for missing data in the EM iterations.

[8] Bańbura and Modugno (2014).

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

Model	**RMSE**	**RMSE (Rel. to AR)**[9]	***p*-value**
Lasso	0.26	0.10	0.014
Ridge	0.34	0.13	0.043
Elastic Net	0.28	0.11	0.039
Adaptive Lasso	0.68	0.27	0.126
Random Forest	0.45	0.18	0.089
Gradient Boosting Machine	0.17	0.07	0.016
DFM full [10]	0.93	0.36	0.005
DFM best [11]	0.72	0.28	0.004
DFM structured [12]	1.05	0.41	0.003
AR	2.55	0.00	

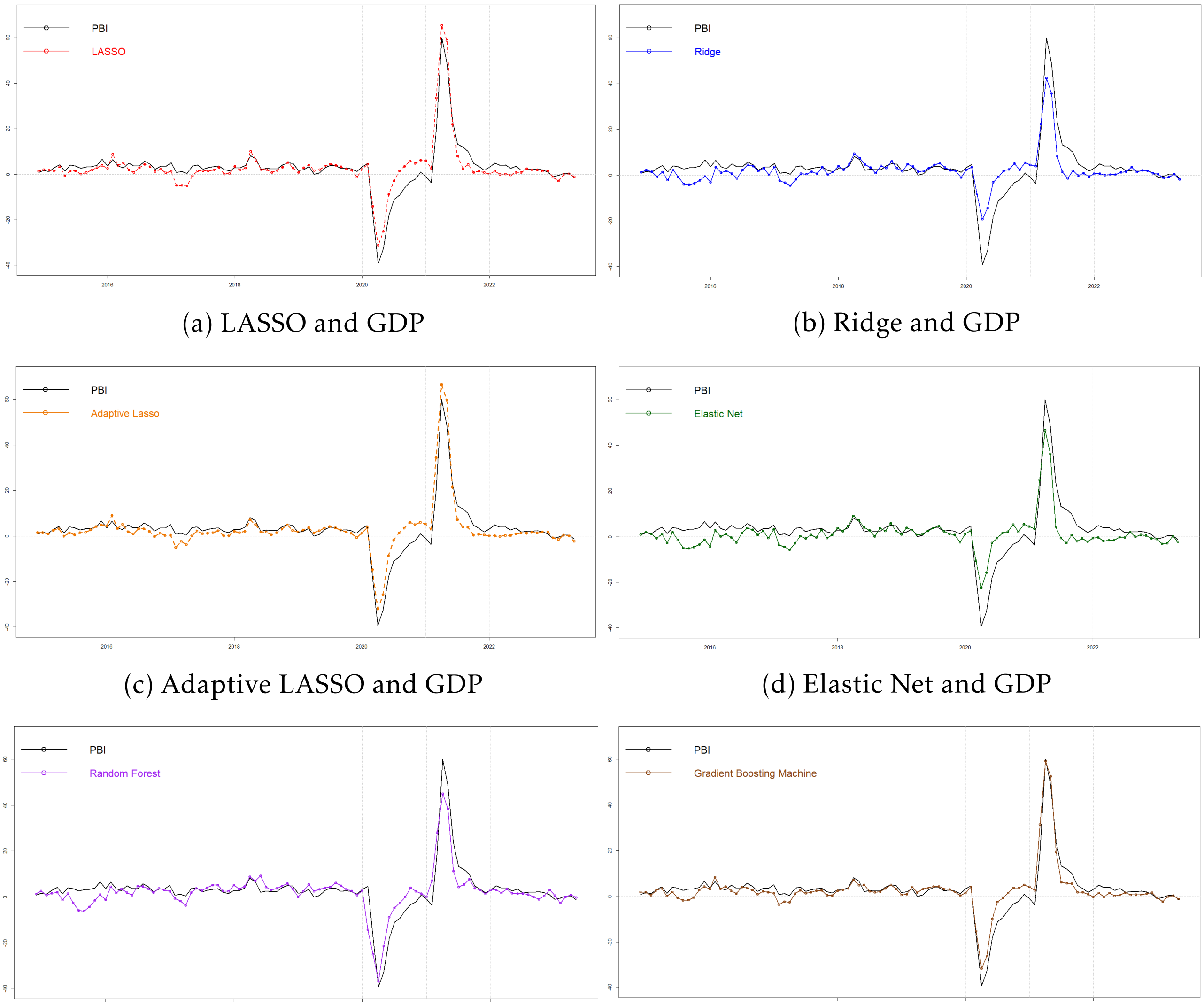
Source: Own elaboration. 9/ $RMSE(Model)_i/RMSE(AR)$. 10/ DFM full, use the 91 variables within unstructured as well as structured data. 11/ DFM best, uses variables within unstructured data and structured selected by the Gibbs sampling as best estimators to predict GDP. 12/ DFM structured, use only 56 structured variables.

Between the models that stand over the others, we get the *Gradient Boosting Machine*, *LASSO* and *Elastic Net* that archive to reduce the forecast error by around 20% to 25%. Also, Diebold-Mariano statistic[13] concludes that most of the ML models are statistically significant, in line with previous research. (Richardson and Mulder, 2018; H. Varian, 2014; Q. Zhang, Ni, and Xu, 2023).

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

[13] Diebold and Mariano (1995).

Figure 3: ML 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

4.3 Consistency

To test the consistency of the results and determine if the ML model projections contribute positively to the accuracy predictions of monthly GDP over the benchmark models, we use the C. Romer and D. Romer (2008) approach, instead of using an officials prediction, we replace to a DFM estimation that incorporates the electricity production as main leading indicator, which popular among Economic Studies Department in Peru. We estimate 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 with 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* according to the Akaike criterion. Likewise, analyzing the estimation errors of the models generated by equation 14, we applied the test proposed by Harvey, Leybourne, and Newbold (1997) with a long run variance autocorrelation estimator from Diebold and Mariano (1995), to evaluate the accuracy gains in the estimates by including 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 a superior accuracy of the models incorporating Machine Learning at a 10% confidence level in the case of the Lasso and Ridge model but at 5% in the others.

Table 4: β_2^e value and validation criteria

Models	Estimated value	AIC	p-value	p-value (DM)
Lasso	0.714	520.32	0.000	0.079
Ridge	0.936	554.73	0.000	0.057
Elastic Net	0.839	549.80	0.000	0.055
Adaptive Lasso	0.703	517.49	0.000	0.046
Random Forest	0.783	534.20	0.000	0.049
Gradient Boosting Machine	0.810	492.09	0.000	0.041

Source: Own elaboration

5 Conclusions

In this article, we evaluated the prediction accuracy of the most popular Machine Learning algorithms to do the nowcasting (tracking in real-time) of 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 the GDP's components measured by the expenditure and productive sector approach. Furthermore, it is worth mentioning that we have enriched our approach by incorporating a sentiment data index built through Google Trends, that shown to be helpful to predict in advance economic activity. The Machine Learning approach allows the use of 91 variables simultaneously between structured data and non-structured data, one of the documents that use a larger dataset used for the Peruvian GDP prediction case. The evaluation results and consistency exercise show evidence of 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 outperforming machine learning over the AR and DFM models in prediction accuracy, which opens a new agenda for emerging economies to improve the forecast of relevant macroeconomic variables such as consumption, employment, and investment, among others.

As these models have been implemented in the Department of Macroeconomic Projections of the Ministry of Economics and Finance of Peru with successful performance and incorporated into the monthly duties, we point out three specific pending agendas regards 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 the accuracy. However, one question arises. Would the analyzed period influence those results? given that between 2004 and 2023 which includes high volatility events such as the pandemic, the global financial crisis and various climate shocks in 2017 and 2023, where ML models with data that do not structured gain greater predictive capacity by being able to track daily frequency data from Google Trend searches. This could be achieved by performing a variance analysis of the projection errors comparing ML models with other more traditional ones during a period of relative normality and other periods of crisis. Second, a fact we observed in the estimates of the unsynchronized availability of the variables (91), which represented challenges and difficulties, which raises the question of whether consistent results are equally obtained with a smaller number of variables, we estimate this in roughly 45% of the 91 variables of the dataset. This proportion could be evaluated in subsequent studies reducing the software requirements. Third, the treatment of the unstructured data could be improved. In this document, we use a simple and didactic management of no-structured data, but it might be considered monthly weighting of searched words in GoogleTrend to smooth the high variability related to this type of data.

References

Araujo, Douglas et al. (2023). “Machine learning applications in central banking”. In: *Journal of AI, Robotics & Workplace Automation* 2(3), pp. 271–293.

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

Aruoba, S Boraǧan, Diebold, Francis X, and Scotti, Chiara (2009). “Real-time measurement of business conditions”. In: *Journal of Business & Economic Statistics* 27(4), pp. 417–427.

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

Bańbura, Marta, Giannone, Domenico, et al. (2013). “Now-casting and the real-time data flow”. In: *Handbook of economic forecasting*. Vol. 2. Elsevier, pp. 195–237.

Bańbura, Marta and Modugno, Michele (2014). “Maximum likelihood estimation of factor models on datasets with arbitrary pattern of missing data”. In: *Journal of applied econometrics* 29(1), pp. 133–160.

Bańbura, Marta and Rünstler, Gerhard (2011). “A look into the factor model black box: publication lags and the role of hard and soft data in forecasting GDP”. In: *International Journal of Forecasting* 27(2), pp. 333–346.

Barrios, Juan José et al. (2021). “Nowcasting para predecir actividad económica en tiempo real: los casos de Belice y El Salvador”. In.

Blei, David M, Ng, Andrew Y, and Jordan, Michael I (2003). “Latent dirichlet allocation”. In: *Journal of machine Learning research* 3(Jan), pp. 993–1022.

Boehmke, Bradly and Greenwell, BM (2020). “Chapter 12: Gradient Boosting”. In: *Hands-On Machine Learning with R*.

Bok, Brandyn et al. (2018). “Macroeconomic nowcasting and forecasting with big data”. In: *Annual Review of Economics* 10, pp. 615–643.

Breiman, Leo (2001). “Random forests”. In: *Machine learning* 45, pp. 5–32.

Brownlee, Jason (2016). “Bagging and random forest ensemble algorithms for machine learning”. In: *Machine Learning Algorithms*, pp. 4–22.

Chakraborty, Chiranjit and Joseph, Andreas (2017). “Machine learning at central banks”. In.

Diebold, Francis X and Mariano, Roberto S (1995). “Comparing predictive accuracy”. In: *Journal of Business and Economic Statistics* 13(3), pp. 253–263.

Döpke, Jörg, Fritsche, Ulrich, and Pierdzioch, Christian (2017). “Predicting recessions with boosted regression trees”. In: *International Journal of Forecasting* 33(4), pp. 745–759.

Doz, Catherine, Giannone, Domenico, and Reichlin, Lucrezia (2011). “A two-step estimator for large approximate dynamic factor models based on Kalman filtering”. In: *Journal of Econometrics* 164(1), pp. 188–205.

Doz, Catherine, Giannone, Domenico, and Reichlin, Lucrezia (2012). “A quasi–maximum likelihood approach for large, approximate dynamic factor models”. In: *Review of economics and statistics* 94(4), pp. 1014–1024.

Eberendu, Adanma Cecilia et al. (2016). “Unstructured Data: an overview of the data of Big Data”. In: *International Journal of Computer Trends and Technology* 38(1), pp. 46–50.

Einav, Liran and Levin, Jonathan (2014). “The data revolution and economic analysis”. In: *Innovation Policy and the Economy* 14(1), pp. 1–24.

Escobal D'Angelo, Javier and Torres, Javier (2002). "Un sistema de indicadores lideres del nivel de actividad para la economía peruana". In.

Etter, Richard, Graff, Michael, et al. (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, Martin (2005). *Where are we now? real-time estimates of the macro economy*.

Forero, Fernando J Perez, Aguilar, Omar J Ghurra, and Vargas, Rodrigo F Grandez (2016). "Un Indicador Lider de Actividad Real para el Perú". In.

Garcia-Donato, Gonzalo and Martinez-Beneito, Miguel A (2013). "On sampling strategies in Bayesian variable selection problems with large model spaces". In: *Journal of the American Statistical Association* 108(501), pp. 340–352.

Ghosh, Saurabh and Ranjan, Abhishek (2023). "A Machine Learning Approach To Gdp Nowcasting: An Emerging Market Experience". In: *Buletin Ekonomi Moneter dan Perbankan* 26, pp. 33–54.

Giannone, Domenico, Reichlin, Lucrezia, and Small, David (2008). "Nowcasting: The real-time informational content of macroeconomic data". In: *Journal of monetary economics* 55(4), pp. 665–676.

Giglio, Stefano, Kelly, Bryan, and Xiu, Dacheng (2022). "Factor models, machine learning, and asset pricing". In: *Annual Review of Financial Economics* 14, pp. 337–368.

González-Astudillo, Manuel and Baquero, Daniel (2019). "A nowcasting model for Ecuador: Implementing a time-varying mean output growth". In: *Economic Modelling* 82, pp. 250–263.

Green, Kesten C and Armstrong, Scott (2015). "Simple versus complex forecasting: The evidence". In: *Journal of Business Research* 68(8), pp. 1678–1685.

Harvey, David, Leybourne, Stephen, and Newbold, Paul (1997). "Testing the equality of prediction mean squared errors". In: *International Journal of Forecasting* 13(2), pp. 281–291.

Kant, Dennis, Pick, Andreas, and Winter, Jasper de (2022). "Nowcasting GDP using machine learning methods". In.

Kapsoli Salinas, Javier and Bencich Aguilar, Brigitt (2002). "Indicadores lideres, redes neuronales y predicción de corto plazo". In.

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

Longo, Luigi, Riccaboni, Massimo, and Rungi, Armando (2022). "A neural network ensemble approach for GDP forecasting". In: *Journal of Economic Dynamics and Control* 134, p. 104278.

Makridakis, Spyros, Spiliotis, Evangelos, and Assimakopoulos, Vassilios (2018). "Statistical and Machine Learning forecasting methods: Concerns and ways forward". In: *PloS one* 13(3), e0194889.

Martınez, M and Quineche, R (2014). *Un indicador lıder para el nowcasting de la actividad económica del perú*. Tech. rep. Mimeo.

Medeiros, Marcelo C et al. (2021). "Forecasting inflation in a data-rich environment: the benefits of machine learning methods". In: *Journal of Business & Economic Statistics* 39(1), pp. 98–119.

Muchisha, Nadya Dwi et al. (2021). "Nowcasting Indonesia's GDP Growth Using Machine Learning Algorithms". In: *Indonesian Journal of Statistics and Its Applications* 5(2), pp. 355–368.

Natekin, Alexey and Knoll, Alois (2013). "Gradient boosting machines, a tutorial". In: *Frontiers in neurorobotics* 7, p. 21.

Pérez Forero, Fernando (2018). *Nowcasting peruvian gdp using leading indicators and bayesian variable selection*. Tech. rep. Banco Central de Reserva del Perú.

Richardson, Adam and Mulder, Thomas (2018). "Nowcasting New Zealand GDP using machine learning algorithms". In.

Romer, Christina and Romer, David (2008). "The FOMC versus the staff: where can monetary policymakers add value?" In: *American Economic Review* 98(2), pp. 230–235.

Rusnák, Marek (2016). "Nowcasting Czech GDP in real time". In: *Economic Modelling* 54, pp. 26–39.

Scott, Steven L and Varian (2013). *Bayesian Variable Selection for Nowcasting Economic Time Series*. Tech. rep. National Bureau of Economic Research.

Snoek, Jasper, Larochelle, Hugo, and Adams, Ryan P (2012). "Practical bayesian optimization of machine learning algorithms". In: *Advances in neural information processing systems* 25.

Stock, James H and Watson, Mark W (1989). "New indexes of coincident and leading economic indicators". In: *NBER macroeconomics annual* 4, pp. 351–394.

Suphaphiphat, Nujin, Wang, Yifei, and Zhang, Hanqi (2022). "A Scalable Approach Using DFM, Machine Learning and Novel Data, Applied to European Economies". In.

Tibshirani, Robert (1996). "Regression shrinkage and selection via the lasso". In: *Journal of the Royal Statistical Society Series B: Statistical Methodology* 58(1), pp. 267–288.

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

Varian, Hal (2014). "Machine Learning and Econometrics". In: *Slides package from talk at University of Washington*.

Woloszko, Nicolas (2020). *A Weekly Tracker of activity based on machine learning and Google Trends*.

Zhang, Qin, Ni, He, and Xu, Hao (2023). "Nowcasting Chinese GDP in a data-rich environment: Lessons from machine learning algorithms". In: *Economic Modelling* 122, p. 106204.

Zou, Hui (2006). "The adaptive lasso and its oracle properties". In: *Journal of the American statistical association* 101(476), pp. 1418–1429.

Zou, Hui and Hastie, Trevor (2005). "Regularization and variable selection via the elastic net". In: *Journal of the Royal Statistical Society Series B: Statistical Methodology* 67(2), pp. 301–320.

6 Appendix

Table 5: Literature on Nowcasting

International			
Nowcasting			
Author	**Year**	**Methodology**	**Country**
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: Own elaboration

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	Daily	MIDAGRI
10	Consumer Confidence Index	Points	Monthly	Apoyo Consultoria
	Activity Indicators			
11	Electricity Production		Monthly	INEI
12	Hydrocarbon Production		Daily	MINEM
13	3-Month Economic Expectations	Points	Monthly	BCRP
14	Oil	B/D	Daily	MINEM
15	Natural Gas	MCF	Daily	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 [14]	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
	Finacial Indicators			
27	General Stock Market Index[15]	Percentages	Daily	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	Wholesale Price Index	Index	Monthly	BCRP
32	Core CPI	Index	Monthly	BCRP

Source: Own elaboration

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

	Structured variables			
	International Indicators			
33	Multilateral Real Exchange Rate	(2009=100)	Monthly	BCRP
34	EMBIG Perú	Pbs	Daily	BCRP
35	Oil WTI	Dollars per Barrel	Daily	Bloomberg
36	USIPC	Index	Monthly	FRED
37	Industrial Production Index	YoY	Quarterly	Bloomberg
38	Copper	cUS$/lb.	Daily	Bloomberg
39	Gold	US$/oz.tr.	Daily	Bloomberg
40	US Manufacturing PMI	Points	Monthly	Bloomberg
41	FED Interest Rate (Upper Limit)	Percentages	Monthly	Bloomberg
42	VIX Index	Percentages	Daily	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
	Fishery Indicators			
47	Anchoveta Landing	Metric Tons	Daily	IMARPE
48	Logarithm of Anchoveta Landing		Daily	Own elaboration
49	Anchoveta Landing[16]		Daily	Own elaboration
50	Variation Anchoveta Landing[17]		Daily	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

Source: Own elaboration

[14] Metropolitan Lima.

[15] Lima.

[16] Seasonally Adjusted.

[17] Seasonally Adjusted.

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

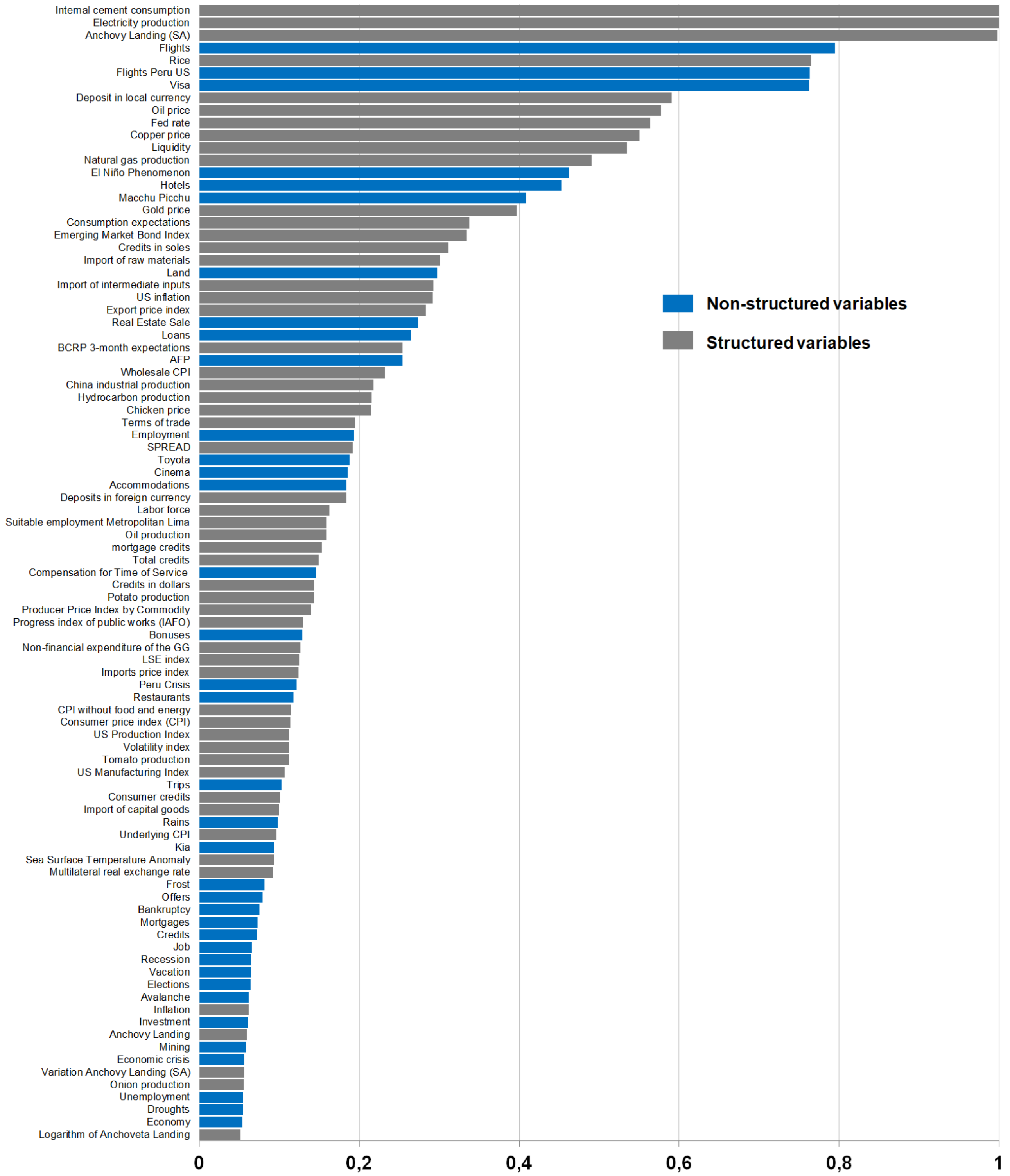


Figure 5: Gibb sampling (2004-2019) - probability of inclusion in optimal model

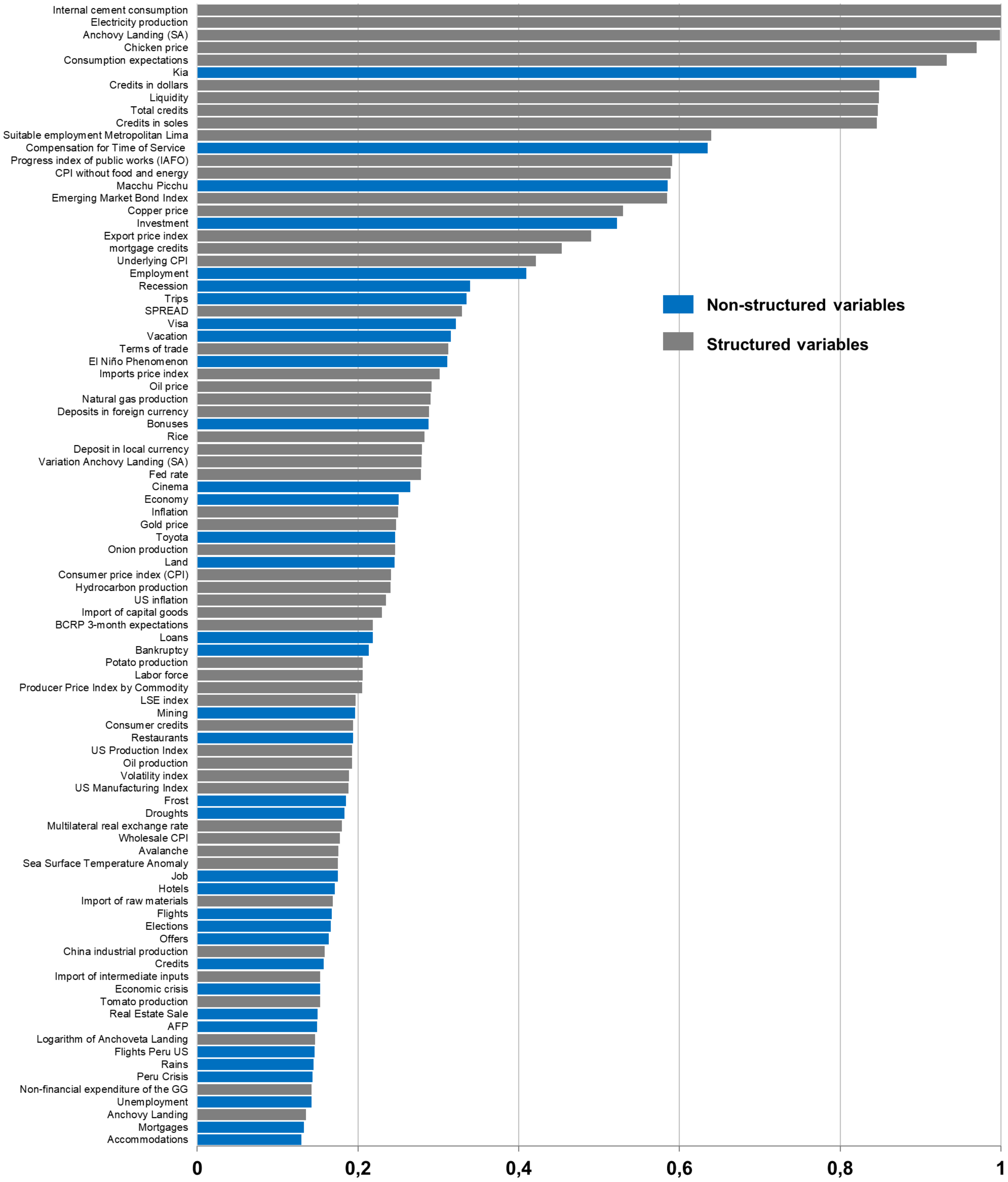
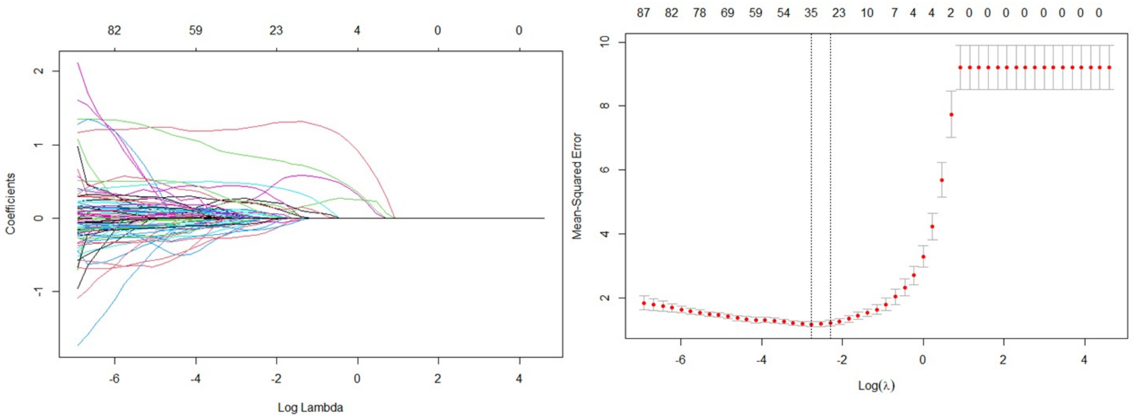


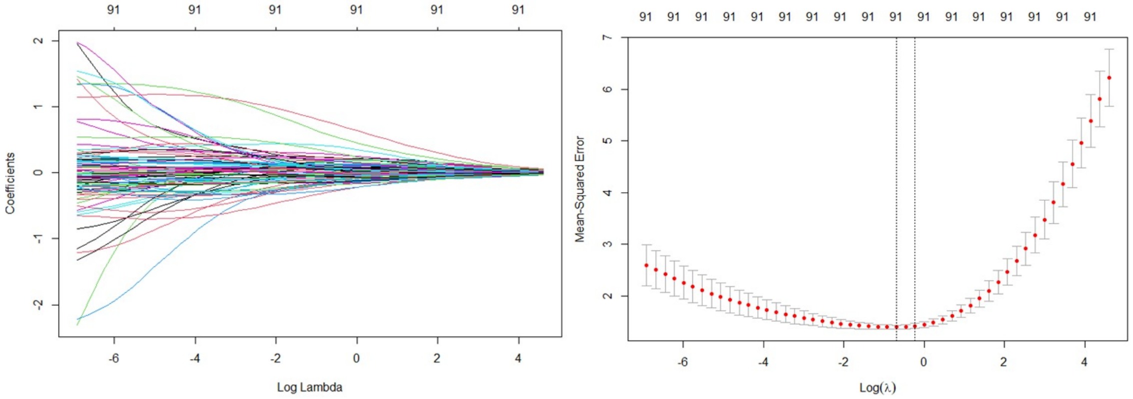
Figure 6: LASSO Optimal Parameters



(a) Coef. (e) by hyperparameter range

(b) Optimal hyperparameter vs. MSE

Figure 7: Ridge Optimal Parameters



(a) CCoef. (e) by hyperparameter range

(b) Optimal hyperparameter vs. MSE

Figure 8: Elastic Net Optimal Parameters

(a) Coef. (e) by hyperparameter range

(b) Optimal hyperparameter vs. MSE

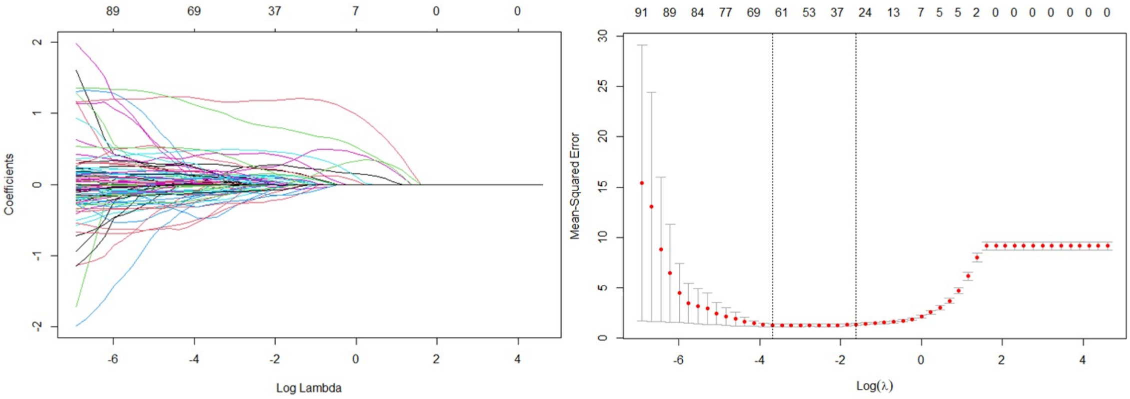


Figure 9: Adaptive LASSO Optimal Parameters

(a) Coef. (e) by hyperparameter range

(b) Optimal hyperparameter vs. MSE

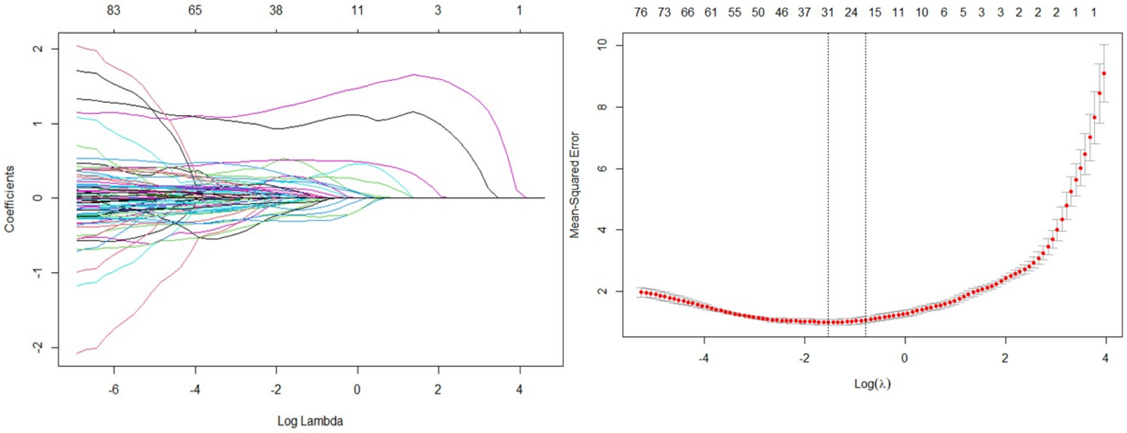
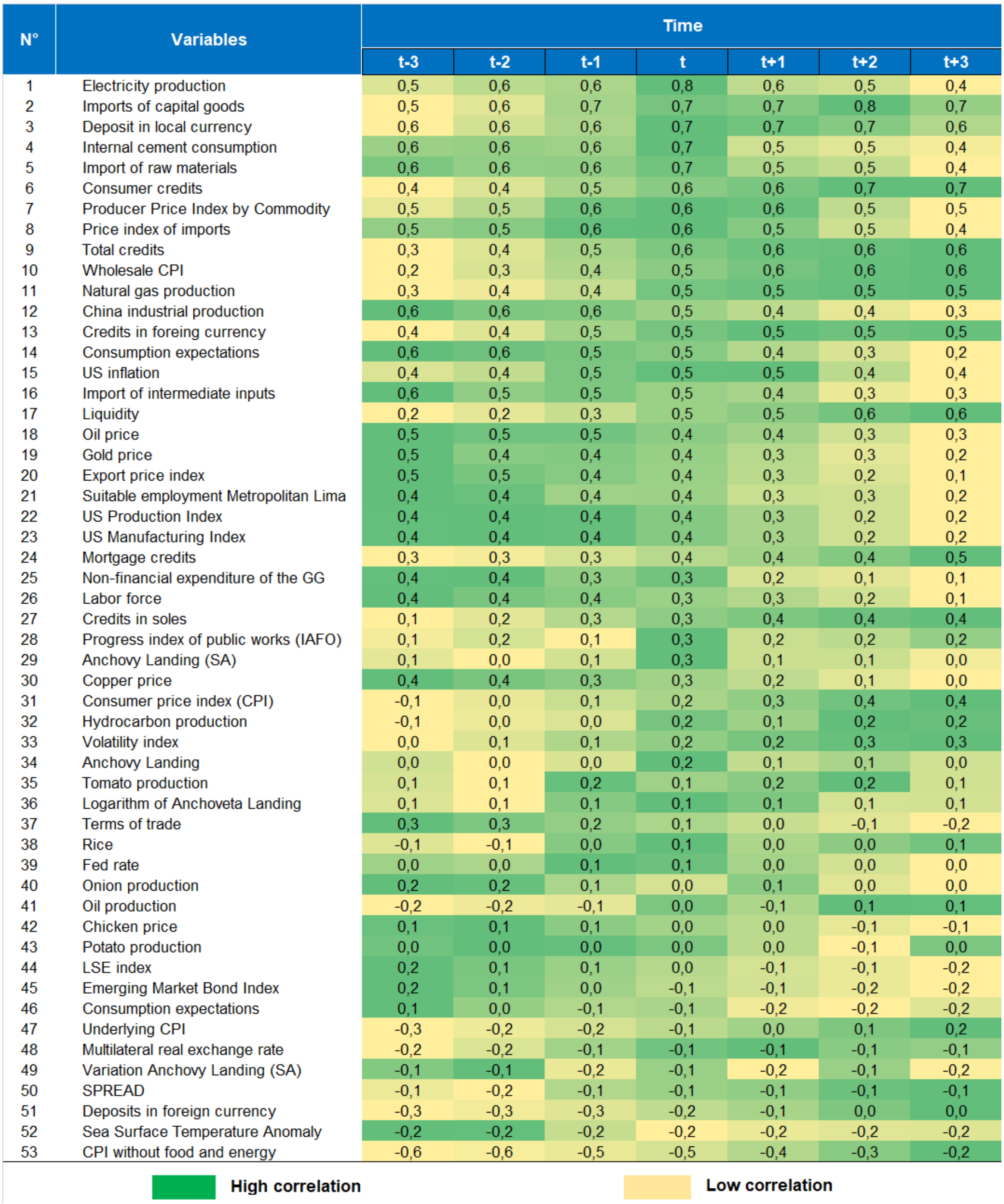


Figure 10: Dynamic correlations of the structured variables

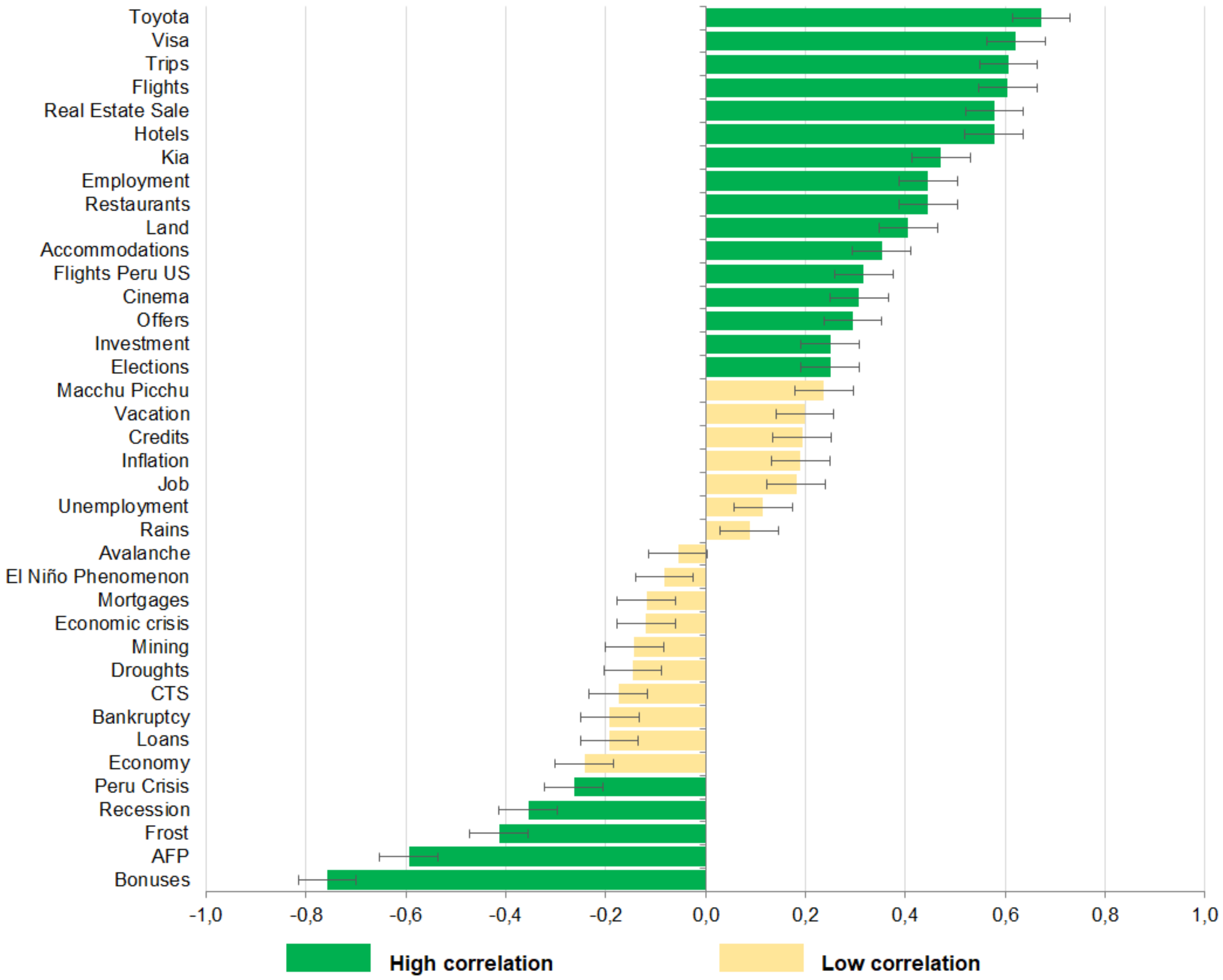


N°	Variables	Time						
		t-3	t-2	t-1	t	t+1	t+2	t+3
1	Electricity production	0,5	0,6	0,6	0,8	0,6	0,5	0,4
2	Imports of capital goods	0,5	0,6	0,7	0,7	0,7	0,8	0,7
3	Deposit in local currency	0,6	0,6	0,6	0,7	0,7	0,7	0,6
4	Internal cement consumption	0,6	0,6	0,6	0,7	0,5	0,5	0,4
5	Import of raw materials	0,6	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: Own elaboration

Figure 11: Correlations of the main nonstructured variables



Source: Own elaboration