Fabio Augusto Fujita

Projeção da inflação de bens industriais brasileira usando métodos de machine learning

Brasil

2022

Fabio Augusto Fujita

Projeção da inflação de bens industriais brasileira usando métodos de machine learning

Dissertação apresentada ao Curso de Mestrado Acadêmico em Economia, Universidade de Brasília, como requisito parcial para a obtenção do título de Mestre em Economia

Universidade de Brasília - UnB

Faculdade de Administração Contabilidade e Economia - FACE

Departamento de Economia - ECO

Programa de Pós-Graduação

Orientador: Daniel Oliveira Cajueiro, Dr.

Brasil

2022

Fabio Augusto Fujita
Projeção da inflação de bens industriais brasileira usando métodos de machine learning/ Fabio Augusto Fujita. – Brasil, 2022-
53p. : il. (algumas color.) ; 30 cm.

Orientador: Daniel Oliveira Cajueiro, Dr.

Dissertação (Mestrado) – Universidade de Brasília - UnB
Faculdade de Administração Contabilidade e Economia - FACE
Departamento de Economia - ECO
Programa de Pós-Graduação, 2022.

1. Inflação de bens industriais. 2. Previsão. 3. Machine learning. II. Universidade de Brasília. III. Faculdade de Administração, Contabilidade e Economia - FACE. IV. Departamento de Economia IV. Projeção da inflação de bens industriais brasileira usando métodos de machine learning

Fabio Augusto Fujita

Projeção da inflação de bens industriais brasileira usando métodos de machine learning

Dissertação apresentada ao Curso de Mestrado Acadêmico em Economia, Universidade de Brasília, como requisito parcial para a obtenção do título de Mestre em Economia

Trabalho aprovado. Brasil, 10 de agosto de 2022:

Daniel Oliveira Cajueiro, Dr.
Orientador

Marina Delmondes de Carvalho Rossi, Dra.
Convidado 1

Fabio Augusto Reis Gomes, Dr.
Convidado 2

Brasil
2022

Resumo

Há um grande interesse em melhorar as projeções de inflação para o planejamento e a tomada de decisão pelas famílias, setor privado e formuladores de políticas. No entanto, superar até mesmo modelos univariados pode ser uma tarefa difícil. Usamos métodos de machine learning e um grande conjunto de dados para prever a inflação de bens industriais no IPCA brasileiro para horizontes até $t + 12$, considerando dados entre janeiro de 2007 e agosto de 2021. Avaliamos as previsões de quatro métodos lineares regularizados e dois métodos não lineares baseados em árvores, considerando random walk e modelos autorregressivos como benchmarks, utilizando uma metodologia pseudo out-of-sample. Também avaliamos os resultados sem dados de desemprego como regressores, levando em consideração as discussões em torno da relevância dos dados de desemprego na previsão de inflação. Os métodos não lineares superam os métodos lineares regularizados e os benchmarks. Também encontramos evidências de que os mecanismos de seleção de variáveis dos métodos random forest e gradient tree boosting têm um desempenho melhor do que os de modelos lineares regularizados para prever a inflação de bens industriais. As random forests se destacam em termos de erro de previsão e como o método que melhor controla o trade-off viés-variância. O método também exibe um desempenho mais uniforme do que o gradient tree boosting ao longo dos horizontes de previsão.

Palavras-chave: Previsão, machine learning, inflação de bens industriais.

Abstract

There is great interest in improving inflation forecasts for better planning and decision making by households, the private sector, and policy makers. However, even outperforming univariate models can be a difficult task. We use machine learning methods and a large data set to forecast industrial goods inflation on Brazilian IPCA for horizons up to $t + 12$, considering the time span between January 2007 and August 2021. We assess the forecasts of four regularized linear methods and two nonlinear tree based methods, with random walk and AR models as benchmarks, in a pseudo out-of-sample framework. We also assess the results without unemployment data as regressors, considering the discussions around the relevance of unemployment data on inflation forecasting. The nonlinear methods outperform the regularized linear methods and the benchmarks. We also find evidence that the variable selection mechanisms of random forest and gradient tree boosting perform better than on linear regularized models to forecast industrial goods inflation. Random forest stands out in terms of forecasting error and as the method that better controls the bias-variance trade-off. It also displays a more uniform performance than gradient tree boosting across the forecasting horizons.

Keywords: Forecasting, machine learning, industrial goods inflation.

List of Figures

List of Tables

List of abbreviations and acronyms

adaLASSO	Adaptative Least Absolute Shrinkage and Selection Operator
ANP	National Agency of Petroleum, Natural Gas and Biofuels
AR	Autoregressive
BCB	Banco Central do Brasil
CPI	Consumer Price Index
ECB	European Central Bank
FGV	Fundação Getúlio Vargas
FRED	Federal Reserve Economic Data
IBGE	Brazilian Institute of Geography and Statistics
IGP-M	General Market Price Index
IPCA	Extended National Consumer Price Index
IPEA	Institute for Applied Economic Research
LASSO	Least Absolute Shrinkage and Selection Operator
MAE	Mean Absolute Error
ML	Machine Learning
MSE	Mean Squared Error
OLS	Ordinary Least Squares
POF	Consumer Expenditure Survey
RF	Random Forest
RW	Random Walk
SGS	Time Series Management System
SVM	Support Vector Machines

Contents

1 Introduction

Inflation forecasts are essential to guide the decisions of the agents. However, forecasting inflation is not an easy task, as it seems to be among the least stable macroeconomic variables (Elliott; Timmermann, 2008). This is due to its dependency on the relationship between several variables that are subject to shifts and shocks. Machine learning (ML) (Bishop, 2006; Izenman, 2008) techniques used to manipulate and analyze data is one of the results of the very productive collaborations between computer scientists and statisticians in the last decades (Varian, 2014). Designed primarily for prediction problems (Mullainathan; Spiess, 2017), its impacts on economic literature and inflation forecasting are emerging (Athey, 2019).

This work uses machine learning methods to forecast industrial goods inflation in Brazil measured by IPCA[1]. Inflation forecasts are particularly important for Brazil, an emerging country which adopted inflation targeting in 1999, after controlling hyperinflation in the mid 90's with the implementation of the Plano Real. Inflation targeting regimes require credibility and transparency on the conduction of the monetary policy to effectively anchor the expectations (Svensson, 2010). Thus, reliable inflation forecasts are crucial for the success of inflation-targeting strategies on emerging countries (Mishkin, 2000), which usually exhibit higher inflation rates and inflationary uncertainty than developed countries (Bansal; Dahlquist, 2000).

We contribute to the literature by expanding the list of machine learning methods tested for inflation forecasting and by not focusing on headline prices variation, but on sectoral inflation. Breaking down inflation can be useful for a better comprehension of the inflationary dynamics and of the impacts of monetary policy on different sectors (Altissimo et al., 2009; Clark, 2006). Additionally, the evolution of relative prices can be useful to analyse the effects of sectorial shocks (Aoki, 2001) and relates to an important part of headline inflation (Reis; Watson, 2010). There are also attempts to improve the accuracy of inflation forecasts by using disaggregated data to model each component by its intrinsic characteristics, even though there is no consensus in the literature about the effectiveness (Hubrich, 2005).

We consider one of the breakdowns adopted by Banco Central do Brasil (BCB), which disaggregates IPCA into four segments[2] with different dynamics: food at home, industrial goods, services and administered prices (BCB, 2019). We opt to work with

[1] IPCA is the Extended National Consumer Price Index. It is the official reference for the Brazilian inflation targeting system. The Brazilian Institute of Geography and Statistics (IBGE) calculates and publishes IPCA in a monthly basis.

[2] This is also the breakdown that BCB considers to collect disaggregated market expectations for IPCA, after the reformulation of Focus Survey in September/2021.

industrial goods inflation due to the inclusion of tradable goods and the influence of imported inflation and cost based variations to the sector (ECB, 2019), which may potentiate the benefits of using a rich data set, including international data like commodity prices and industrial costs.

We assess the forecasts of four regularized linear models (ridge regression, LASSO, adaLASSO and elastic net) and two nonlinear tree based methods (random forests and gradient tree boosting). We use autoregressive models and random walk as benchmarks. The forecasts follows a pseudo out-of-sample framework, meaning that we use only data available by the cut-off date to estimate the models, which we consider as five business days before the publication of IPCA_{t+1}. We estimate the models using a 60 period rolling window and report mean squared errors (MSE) and mean absolute errors (MAE) for five horizons: $t+1$, $t+3$, $t+6$, $t+9$ and $t+12$.

Our data covers the period between January 2007 and August 2021 (176 observations) and includes monthly time series for inflation indexes, exchange rates, commodity prices, producer prices, industrial manufacturing and transportation costs, industrial wages, exports, local and international interest rates, monetary variables, public debt, energy consumption and prices, economic activity indicators and unemployment. We also include market expectations and professional forecasts for IPCA and exchange rate from BCB's Focus Survey. We consider eleven lags of all the variables, with the exception of the market expectations from Focus.

The literature on inflation forecasting is vast and includes several traditional approaches and its variations, such as Phillips curves, unobserved component stochastic volatility models, structural models, vector autoregressions, bayesian models, factor models, DSGE models and many others (Faust; Wright, 2013). The broader availability of data and advancements in statistical methods and computational power also favored the use of less traditional methods to forecast inflation, such as the use of big data (Cavallo; Rigobon, 2016) and machine learning.

Our work connects particularly to works that use ML methods to forecast Brazilian inflation. Recent publications compare the forecasting accuracy of ML methods like random forests and several instances of regularized models (such as ridge regression, LASSO, adaLASSO and elastic net) to traditional econometric approaches, factor models and reduced-form structural models. The results of some of the first publications are not entirely favorable to ML methods. For instance, Medeiros et al. (2016) find that LASSO based methods only outperform AR models when forecasting Brazilian headline inflation in the short-term[3] and without statistical difference. However, examples in which ML methods

[3] IPCA up to $t+4$ and IGP-M at $t+1$. IGP-M is the General Market Price Index. Fundação Getúlio Vargas (FGV) calculates and publishes IGP-M in a monthly basis. It includes consumer, wholesale and construction prices.

outperform traditional approaches are emerging in the literature. Some examples are the works of Garcia et al. (2017) and Araujo and Gaglianone (2020) forecasting Brazilian headline IPCA and Chakraborty and Joseph (2017) and Medeiros et al. (2021), predicting UK CPI and USA CPI, respectively. Random forest stands out among the ML methods in these works, specially in periods of greater uncertainty (Chakraborty; Joseph, 2017; Medeiros et al., 2021).

Furthermore, it is worth mentioning that our work is an instance of the so-called *prediction policy problem*. It is crucial to distinguish prediction policy problems from those that deal with causal inference. We may find a rich discussion of the differences in the work of Kleinberg et al. (2015). In a broad view, these authors argue that a wide range of relevant problems of public policy formulation does not necessarily require causal inference but rather predictive inference.

We find that the nonlinear methods have the best performance, ranking first and second in terms of smallest MSE and MAE for all forecasting horizons, with the exception of $t+1$, in which ridge regression has the second best performance. Random forest outperforms all the other methods, except for $t+12$, when gradient tree boosting reports smaller errors. Among the regularized linear models, ridge regression stands out, even though the methods did not beat the AR benchmark in $t+3$ and $t+6$ horizons.

In order to assess the relevance of unemployment data on industrial goods inflation forecasting, we repeat the tests for all models after removing it from the list of regressors. Even though the negative correlation between inflation and unemployment is a historically important basis for inflation forecasting (Stock; Watson, 1999), many aspects of its dynamics are still unexplained and its relevance is even questioned by some authors (Mankiw, 2001). The removal of unemployment data has minor effects on the results of the nonlinear methods, which continue to have better overall performance, and conflicting contributions among the forecasting horizons on LASSO based models. However, there are improvements in the MSE and MAE for elastic net and ridge regression in all forecasting horizons, with the latter beating both benchmarks.

The results give further evidence that not only the nonlinear methods are able to better represent the complex dynamics of inflation, but also that the variable selection for random forest and gradient tree boosting perform better than for the shrinkage methods to forecast industrial goods inflation. Moreover, even though we do not intend to assess the causal inference between inflation and unemployment, we have an indication that either (i) unemployment data may not be relevant to forecast industrial goods inflation when using a large data set, even though ML methods frequently select it; or (ii) the influence of unemployment on industrial goods inflation is episodic; or (iii) IBGE's open unemployment is not a good proxy to forecast inflation.

This work is organized as follows. The next section introduces the machine learning

methods we use to forecast the industrial goods segment of IPCA, as well as defines the benchmarks. Section 3 describes the procedures we adopt to tune and estimate the models. We detail the data set we use in Section 4 and present the results in Section 5. Section 6 summarizes and concludes the work.

2 Models

2.1 Benchmarks: Random Walk and AR

We use random walk and autoregressive models as benchmarks. Despite being basic univariate models, the literature has several evidence that beating these naive models in forecasting inflation is not an easy task. For example, Atkeson and Ohanian (2001) show that even state of the art (by the time the article was written) NAIRU based Phillips curves cannot forecast US inflation four quarters ahead better than a simple random walk, over the period between 1984 and 1999. Considering inflation in Brazil, Medeiros et al. (2016) find that AR models outperform LASSO and adaLASSO models for forecasting IPCA horizons longer than four months ahead and IGP-M horizons other than one month ahead.

The random walk model forecasts all horizons as the last value observed for the variable as in $\hat{x}_{t+h} = x_t$, where h is the forecast horizon.

The autoregressive models are linear models represented as:

$$\hat{x}_{t+h} = c + \sum_{i=1}^{p} \Phi_i x_{t-i} \,, \tag{2.1}$$

where c is a constant, Φ_i are the parameters of the model estimated using regression on past observations of the variable and p is the order of the model, determined by some model selection criteria.

2.2 Regularized linear models

Linear models for time series forecasting have the form:

$$y_{t+h} = \boldsymbol{\beta}_0 + X_t \hat{\boldsymbol{\beta}} + \epsilon_{t+h} \,, \tag{2.2}$$

where X_t is a matrix of dimensions Txn with the regressor variables as columns, T is the number of observations, n is the number of regressors, $\hat{\boldsymbol{\beta}}$ is the model coefficients column vector, $\boldsymbol{\beta}_0$ is the intercept, y_{t+h} is the variable we want to forecast and ϵ_{t+h} is the error term.

Traditional linear regression models use OLS to estimate $\hat{\boldsymbol{\beta}}$ by minimizing the sum of squared residuals, as in:

$$\hat{\boldsymbol{\beta}} = \arg\min_{\boldsymbol{\beta}} \left(\sum_{t=1}^{T} \left(y_{t+h} - \boldsymbol{\beta}_0 - X_t \boldsymbol{\beta} \right)^2 \right) . \tag{2.3}$$

One of the problems of OLS is that, when $n > T$, the OLS estimator is not unique, as the minimization problem is rank deficient. Considering the broad availability of data, it is possible that $n \gg T$, specially in time series forecasting problems using machine learning methods. In these cases, it is possible to completely overfit (Masini et al., 2021). Another characteristic of OLS is that the forecasts often have low bias, but large variance, which might lead to bad prediction accuracy. Sometimes it is useful to sacrifice bias to reduce the variance of the predicted values and improve the overall prediction accuracy (Hastie et al., 2009).

To address these problems, regularized linear methods introduce penalties for overfitting the data, controlling model complexity by shrinking parameters values towards zero, unless supported by the data (Bishop, 2006). The coefficients of the linear model are the result of the following minimization problem:

$$\hat{\boldsymbol{\beta}} = \arg\min_{\boldsymbol{\beta}} \left(\sum_{t=1}^{T} \left(y_{t+h} - \boldsymbol{\beta}_0 - X_t \boldsymbol{\beta} \right)^2 + p(\boldsymbol{\beta}, \lambda, \omega) \right) , \tag{2.4}$$

where the regularization term p is a function of the coefficients $\boldsymbol{\beta}$, a tuning parameter $\lambda \geq 0$ that controls how heavily the method penalizes overfitting and possibly other vector or scalar parameters ω. The intercept is not part of the regularization function, as the results would depend on the starting value of the forecasted variable. In the following sections we present the regularization functions and the characteristics of ridge regression, LASSO, adaLASSO, and elastic net methods.

2.2.1 Ridge Regression

Hoerl and Kennard (1970) introduced Ridge Regression as a method to reduce the mean error by adding bias and reducing the variance of OLS estimators to fight multicolinearity. The regularization term of Equation (2.4) uses the ℓ_2 norm of the coefficients, as in:

$$p(\boldsymbol{\beta}, \lambda) = \lambda ||\boldsymbol{\beta}||_2^2 = \lambda \sum_{i=1}^{n} \beta_i^2 . \tag{2.5}$$

By using the ℓ_2 norm to penalize complexity, the solution to the minimization problem is not sparse, as the coefficients are rarely set to zero.

2.2.2 LASSO

LASSO is the acronym for Least Absolute Shrinkage and Selection Operator, a method proposed by Tibshirani (1996) that employs the ℓ_1 norm to regularize the coefficients in Equation (2.4):

$$p(\boldsymbol{\beta}, \lambda) = \lambda ||\boldsymbol{\beta}||_1 = \lambda \sum_{i=1}^{n} |\beta_i| \,. \tag{2.6}$$

By using the ℓ_1 norm in the regularization term, the solution of LASSO is sparse, meaning that the model performs shrinkage and variable selection at the same time. Ridge regression does a proportional shrinkage, while LASSO performs a "soft thresholding", translating each coefficient by the constant factor λ, truncating at zero (Hastie et al., 2009). Another difference between ridge regression and LASSO is that, because of the nature of the convex optimization problem, when the number of variables n is greater than the number of observations T, LASSO selects at most n variables before it saturates (Zou; Hastie, 2005).

2.2.3 AdaLASSO

Zhao and Yu (2006) show that the variable selection performed by LASSO may be inconsistent under certain conditions and that, in particular, if an irrelevant predictor is highly correlated with the predictors in the true model, LASSO may not be able to distinguish it from the true predictors with any amount of data and any amount of regularization.

To address this issue, Zou (2006) proposed the Adaptative LASSO (AdaLASSO), a two step estimation that uses the coefficients of a first model to generate weights to penalize different coefficients in LASSO ℓ_1 penalty. The regularization function in Equation (2.4) is:

$$p(\boldsymbol{\beta}, \lambda, \boldsymbol{\omega}) = \lambda \sum_{i=1}^{n} \omega_i |\beta_i| \,, \tag{2.7}$$

where $\omega_i = |\hat{\beta}_j|^{-\tau}$ are the adaptative weights based on the coefficients $\hat{\beta}_j$ of the first step estimation, and $\tau > 0$ is an additional tuning parameter. The first step estimation can be any consistent estimator of β, in which case adaLASSO is said to have what Fan and Li (2001) define as the Oracle properties. The author suggests OLS unless collinearity is a concern, in which case Ridge Regression may be a better option.

2.2.4 Elastic Net

Zou and Hastie (2005), who first proposed elastic net, describe it as a method that simultaneously performs automatic variable selection, continuous shrinkage and, unlike LASSO, encourages a grouping effect of strongly correlated predictors.

As a compromise between Ridge Regression and LASSO, the elastic net uses a linear combination of $\ell 1$ and $\ell 2$ norms, and the regularization function in Equation (2.4) becomes:

$$p(\boldsymbol{\beta}, \lambda, \alpha) = \lambda \left(\alpha \sum_{i=1}^{n} \beta_i^2 + (1-\alpha) \sum_{i=1}^{n} |\beta_i| \right) = \lambda \left(\alpha ||\boldsymbol{\beta}||_2^2 + (1-\alpha) ||\boldsymbol{\beta}||_1 \right) , \tag{2.8}$$

where $\alpha \in [0, 1]$.

2.3 Nonlinear models

For nonlinear models we focus on two regression tree based methods: Random Forests and Gradient Tree Boosting.

Regression trees approximate an unknown nonlinear function with local predictions using recursive partitioning of the space of covariates. The starting point is the root node. A node is a subset of the set of variables and a non-terminal node splits into two child nodes. The algorithm searches for the best split and applies this boolean condition on the value of the variables. A node that does not split is called a terminal node or leaf. The value on the terminal node determines the output variable (Izenman, 2008).

The tree size is a tuning parameter governing the model's complexity. Large trees may overfit the data, while small trees may not capture the data's structure. Hence the importance of choosing the optimal size adaptively, based on the data (Alpaydin, 2014).

2.3.1 Random Forest

Random Forest is a method originally proposed by Breiman (2001) used for classification and regression. When used for regression, the algorithm grows a collection of regression trees specifying each tree in a bootstrapped sub-sample of the original data. When growing the trees, the method considers a limited random subset of variables to determine the best split at each node. The prediction for each tree is at the terminal nodes (or leaves), and the final prediction is formed by taking the average of all the trees. In other terms:

$$\hat{y}_{t+h} = \frac{1}{B} \sum_{b=1}^{B} f_b(\mathbf{x}_t) \, , \tag{2.9}$$

where y_{t+h} is the variable we want to forecast, B is the number of regression trees, and $f_b(\mathbf{x}_t)$ is the result of each regression tree built considering a bootstrapped sub-sample of the training data and the randomization of the subset of variables considered to determine the best split at each node.

Bagging involves averaging across models estimated with several different bootstrap samples in order to improve the performance of an estimator (Varian, 2014). Considerable evidence has been accumulated since the introduction of bagging that demonstrates its effectiveness on improving the accuracy of some estimators, specially nonlinear ones like trees and neural networks (Friedman; Hall, 2007). Random forests goes beyond and introduce randomness to the tree growing process to reduce the correlation between individual trees (Efron; Hastie, 2016; James et al., 2013).

Random forests are very popular and can perform remarkably well in representing complex relations in the data, with very little tuning required when compared to methods like deep neural networks (Athey; Imbens, 2019).

2.3.2 Gradient Tree Boosting

Gradient Boosting is a greedy method originally proposed by Friedman (2001) based on constructing additive models by sequentially fitting a base learner to current pseudo-residuals at each iteration. The pseudo-residuals are the gradient of the loss function being minimized, with respect to the model values at each training data point evaluated at the current step. In other terms, if we want to estimate the function $f(\mathbf{x})$, we iteratively use regression trees and compute:

$$f_m(\mathbf{x}) = f_{m-1}(\mathbf{x}) - \rho_m \frac{\partial \ell(f(\mathbf{x}))}{\partial f(\mathbf{x})}\bigg|_{f(\mathbf{x})=f_{m-1}(\mathbf{x})}, \tag{2.10}$$

where $\ell(f(\mathbf{x}))$ is the loss fuction and ρ_m is the learning rate (a multiplicative factor to the contribution of each tree that controls the convergence speed). The method repeats the process for the number of iterations specified (Izenman, 2008).

Gradient tree boosting is a special case of gradient boosting, in which the base learners are short regression trees. The idea is to iteratively add regression trees to the model that are fit to the negative gradient of the loss function, reducing the estimation error. The researcher should use some model selection criteria to tune the parameters of gradient boosting, as well as the base parameters for the regression trees.

We used XGBoost, a variant of gradient tree boosting introduced by Chen and Guestrin (2016) that implements algorithmic optimizations that improve scalability and reduces running time. According to the authors, the impact of the system has been widely

recognized in machine learning and data mining competitions like Kaggle's and KDD Cup, providing state of the art results on a wide range of problems.

3 Methodology

We forecast industrial goods inflation on IPCA and report mean squared errors and mean absolute errors for five horizons: $t+1$, $t+3$, $t+6$, $t+9$ and $t+12$. The pseudo out-of-sample methodology simulates real time forecasts, using only data available by the cut-off date, which we consider as five business days before the publication of IPCA_{t+1}.

We estimate the models using a 60 period rolling window and tune the hyper-parameters using grid-search and out-of-sample evaluation[1] on twelve forecasts. In other words, at the beginning of the sample, we estimate the models with the 60 period rolling window and forecast $t+h$ for every combination of hyper-parameters. We store the forecasts and compute the mean squared errors after twelve steps. In cases of higher computational cost or higher number of hyper-parameters to tune, the combination with the smallest MSE is used to forecast the complete test sample. This is the procedure we adopt for gradient tree boosting. In other cases (ridge regression, LASSO, elastic net and random forest), we perform the hyper-parameter validation at every step, by adding to the records another set of forecasts for every combination of hyper-parameters on the grid and removing the first set (thus keeping twelve forecasts to choose the smallest MSE).

For adaLASSO, we use ridge regression coefficients as the adaptative weights, as Zou (2006) suggests[2]. We calculate the ridge regression coefficients at each step and validate the shrinkage factor λ only once. The parameters selected for the benchmark AR models vary between AR(3) and AR(4) models, with very small differences in MSE. We adopt an AR(3) for parcimony.

Section 4 describes the data we use for the estimation of the models. To take into account possible longer pass-through effects, we include the last observation and eleven lags of each variable, totalling one year of data. The exceptions are the market expectations for headline IPCA and exchange rate from the Focus survey. For the regularized models, we standardize the predictors to prevent the magnitude of the variables from affecting the regularization.

We test the predictive performance[3] of the machine learning methods against the

[1] Tashman (2000) emphasizes the importance of out-of-sample evaluation tests of forecasting accuracy and describes this rolling window implementation. The main drawbacks of this methodology are the computational cost, the fact that we reduce the period of valid forecasts and do not benefit from some properties of the traditional k-fold cross-validation framework. However, we keep on simulating a true out-of-sample evaluation and also avoid theoretical problems with respect to temporal evolutionary effects and dependencies within the data (Bergmeir; Benitez, 2012).

[2] Zou (2006) suggests $\hat{\boldsymbol{\beta}}(ols)$ for general cases and $\hat{\boldsymbol{\beta}}(ridge)$ when collinearity is a concern. Besides the concerns with collinearity, we use $\hat{\boldsymbol{\beta}}(ridge)$ as $n > T$.

[3] Diebold (2015) published a paper with his personal perspective about the use of Diebold-Mariano test, twenty years after the publication of the 1995 original paper. He emphasizes that the Diebold-Mariano

benchmarks using Diebold-Mariano test, in which the null hypothesis is of equal forecast accuracy (Diebold; Mariano, 1995). Diebold-Mariano test properties are likely to differ under the null hypothesis with nested models, as forecasts errors are asymptotically the same and therefore perfectly correlated (Clark; McCracken, 2001). However, Giacomini and White (2006) prove that the test is valid if we estimate the models using a rolling window framework.

We estimate the models using standard Python packages: Statsmodels for ARMA models; Scikit-Learn for LASSO, Ridge Regression, Elastic Net and Random Forest; ASGL[4] package for AdaLASSO and XGBoost package for XGBoost.

test is intended for comparing forecasts and not for comparing models. Despite Diebold's criticism about the use of Diebold-Mariano test for pseudo out-of-sample model comparisons, we are interested in the comparative historical predictive performance with a limited sample, one of the cases the author claims that may justify its use.

[4] ASGL is a package for regularized linear and quantile regression related to the research of Mendez-Civieta et al. (2021)

4 Data

The data covers the period between January 2007 and August 2021 (176 observations). We opt to use data starting from 2007 as it is the first full year after the weighting structure revision of IPCA, implemented on July 2006. The previous weighting structure considered the Consumer Expenditure Survey (POF) IBGE conducted during the biennium of 1995/1996, a period of major changes[1] in the Brazilian economy and stabilization of the country inflation after the implementation of the Plano Real, in 1994. IBGE carried out two additional weighting structure reviews of IPCA after July 2006, considering the Consumer Expenditure Surveys of 2008/2009 and 2017/2018. The inflation rate in Brazil, however, was significantly more stable.

We obtained the time series from IBGE, Fundação Getúlio Vargas (FGV), the Time Series Management System (SGS) maintained by BCB, the Brazilian National Agency of Petroleum, Natural Gas and Biofuels (ANP), the IPEADATA maintained by the Institute for Applied Economic Research (IPEA) and the Federal Reserve Economic Data (FRED), maintained by the Federal Reserve Bank of St. Louis. The data set includes monthly time series for inflation indexes, exchange rates, commodity prices, producer prices, industrial manufacturing and transportation costs, industrial wages, exports, local and international interest rates, monetary variables, public debt, energy consumption and prices (electric energy and fuels), economic activity indicators and unemployment. The full list of variables and original sources are available in Appendix A.

We also use data from BCB's Focus Market Readout to include market expectations and professional forecasts. The report collects market expectations for several economic variables from professional forecasters, banks, asset managers, consultancies and other institutions that operate in the financial market. The agents inform their forecasts on business days and BCB publishes the compiled statistics of the expectations collected until Fridays on the next Monday. BCB ranks the forecasters and the five best for each forecast horizon (short-term, mid-term and long-term) are part of Focus TOP5. We use the market expectations for headline[2] IPCA and monthly average exchange rates (USD/BRL) for the next twelve months, considering the median of the forecasts of the TOP5 mid-term ranking.

[1] To illustrate this fact, the twelve month accumulated inflation shifted from 631,54% in January 1995 to 9,56% in December 1996.

[2] In September/2021, BCB announced the inclusion of forecasts for the IPCA Industrial Goods segment to the Focus Survey, among other additions. As our focus is on forecasting IPCA industrial goods inflation, it would be interesting to use the specific market expectations for the segment (besides the headline inflation) among the independent variables. The data, however, was not available for the time span we used. When this data is available for a reasonable time span, it should be considered as one of the regressors or even as a benchmark.

In the cases in which only daily data is available, we consider the average value for the month. Additionally, for regularized models, we standardize the predictors to prevent the magnitude of the variables from affecting the regularization. Other than that, the only transformation on the data are the ones necessary to achieve stationarity.

To take into account possible longer pass-through effects, we include the last observation and eleven lags of each variable, totalling one year of data. The exception are the market expectations from the Focus survey. Following the methodology described in Section 3, the pseudo out-of-sample window goes from January/2014 to August/2021.

5 Results

5.1 Complete sample

We forecast industrial goods segment inflation on IPCA using the methods described in Section 2 for five horizons: $t+1$, $t+3$, $t+6$, $t+9$ and $t+12$. In this subsection we report the results using the complete data described in Section 4 and Appendix A as regressors.

Table 1 reports the mean absolute error and mean squared error for each of the forecast horizons. Overall, the nonlinear methods have the best performance, with random forest and gradient tree boosting ranking first and second in terms of smallest MSE and MAE for all forecasting horizons, with the exception of t+1, in which ridge regression has the second best performance. Random forest dominates all the other models, linear and non linear, both in terms of MSE and MAE for all forecasting horizons, except for t+12, when gradient tree boosting is the best performing method.

Among the ML regularized linear models, ridge regression has the smallest MSE and MAE for all forecasting horizons, except for t+6, in which adaLASSO has the best overall performance, and t+12, in which LASSO has the smallest MAE. However, these ML methods do not beat the AR(3) benchmark in t+3 and t+6 horizons. For some applications in the literature, LASSO based models prove to be more accurate than other shrinkage methods, including ridge regression. We attribute the superiority of ridge regression in forecasting industrial goods inflation to the high correlation between the predictors, since the data set includes several industrial costs, commodity prices, inflation indexes and some of its breakdowns. In this situation, specially when the number of regressors is larger than the number of observations, Zou and Hastie (2005) and Tibshirani (1996) observe empirically that ridge regression dominates LASSO.

Garcia et al. (2017) report that LASSO based models have the best performance forecasting IPCA in the short-term, when compared to other traditional and ML methods, including random forest. They attribute the success of the methods to the use of expert forecasts in the list of regressors, which tend to be very precise in the short-term. One possible explanation for the bad performance of LASSO based models in our assessment is that, even though we are considering expert forecasts as regressors, these expectations are for headline inflation and not specifically for the for the industrial sector inflation.

Table 2 reports the results of the Diebold-Mariano test of the ML methods forecasts against the two benchmarks (AR and RW). Even though the test has its limitations relative to nested models, Giacomini and White (2006) prove it to be valid when estimating the models in a rolling-window framework, as discussed in Section 3.

Random forest forecasts are the only ones with statistical significance of at least 10% in Diebold-Mariano test for all horizons against the forecasts of the two benchmarks. No ML linear regularized model forecast reaches statistical significance of at least 10% against the AR benchmark for horizons of t+6 or longer and for horizons up to t+9 against RW.

Figure 1 shows the boxplots for the forecasting errors, considering each method and forecast horizon. Table 3 summarizes the interquantile range of the errors. The information can be useful to analyze the error distribution and variance.

Random forest seems to be the method with the best bias-variance trade-off balance. It reports the smallest error variance for $t+3$, $t+6$ and $t+9$ horizons, as well as on average, as evidenced by the lower error interquantile range. Among the linear regularized models, ridge regression has the lowest error interquantile range for all forecasting horizons (for $t+1$, it has the lowest variance among all the methods). The error interquantile range values for random forest, gradient tree boosting and ridge regression are close, but the median and average errors in the boxplots suggest that the linear model has more biased errors than the nonlinear ones in our sample.

Other authors also highlight the good performance of random forest in forecasting headline inflation when compared to a variety of methods using data from different countries. Some examples are the works of Araujo and Gaglianone (2020) forecasting Brazilian IPCA, Medeiros et al. (2021) addressing USA CPI prediction and Chakraborty and Joseph (2017) considering CPI in the United Kingdom.

Considering that random forest achieves the best overall performance, we assess the relevance of the variables for this model. The feature importance algorithm we use calculates the weighted average decrease in node impurity using Gini importance, in which the weights relate to the probability of reaching that node. The higher the decrease in node impurity when splitting a node into two child nodes based on a given feature, the more important is the feature. The feature importance for random forest is the average feature importance considering all the trees. We compute the frequency at which each variable is among the 5% most important for random forest at each forecast (the eighteen most important features for each window). We categorize the variables into six groups: (i) AR components and inflation indexes; (ii) money and exchange rates; (iii) interests, bonds and debt; (iv) costs, economic activity and trade; (v) commodities and energy and (vi) unemployment.

Figure 2 shows that the AR components and inflation indexes category is by far the most important for shorter horizons and gradually loses importance for longer horizons. The decreasing importance of AR components towards longer horizons is somewhat expected, as the AR models with smaller MSE tends to be AR(3) or AR(4). It is important to point out that this category also includes the market expectations for future headline

IPCA, which tends to be more accurate for shorter horizons. Oppositely, commodities and energy and money and exchange groups gain importance towards longer horizons, at least partially due to the pass-through lag. Unemployment is the least important category, particularly for shorter forecasting horizons.

MSE x 1000	Forecast Horizon				
(MAE x 1000)	t+1	t+3	t+6	t+9	t+12
AR	0.808 (2.174)	0.976 (2.503)	1.225 (2.681)	1.361 (2.84)	1.444 (2.927)
Random Walk	1.097 (2.572)	1.165 (2.703)	1.615 (3.078)	1.857 (3.286)	1.977 (3.384)
LASSO	1.338 (2.671)	1.402 (2.823)	2.064 (3.325)	1.516 (3.070)	1.352 (2.776)
AdaLASSO	1.266 (2.745)	1.386 (2.817)	1.291 (2.835)	1.713 (3.099)	1.427 (3.015)
Ridge Regression	**0.762** **(2.039)**	0.996 (2.288)	1.425 (2.985)	1.305 (2.753)	1.268 (2.815)
Elastic Net	1.171 (2.578)	1.435 (2.925)	1.466 (3.065)	1.350 (2.889)	1.551 (3.151)
Random Forest	0.663 (1.997)	0.763 (2.137)	0.814 (2.223)	0.870 (2.276)	**0.955** **(2.397)**
Gradient Tree Boosting	0.788 (2.179)	**0.804** **(2.234)**	**0.860** **(2.325)**	**1.033** **(2.424)**	0.733 (2.043)

The table shows the mean squared errors and mean absolute errors (in parentheses) for each of the forecast horizons. The values are multiplied by 1000. The shaded cells indicate the models with the lowest MSE and MAE for each horizon. The values in bold represent the models with the second smallest MSE and MAE for each horizon.

Table 1 – Mean Squared Error and Mean Absolute Error - Complete sample (with unemployment data)

Against AR	Diebold-Mariano test: statistical significance				
	t+1	t+3	t+6	t+9	t+12
LASSO	**	**			
AdaLASSO	***	**			
Ridge Regression					
Elastic Net		***			
Random Forest	**	***	***	***	***
Gradient Tree Boosting			**		***

Against RW	Diebold-Mariano test: statistical significance				
	t+1	t+3	t+6	t+9	t+12
LASSO					**
AdaLASSO					
Ridge Regression					**
Elastic Net					
Random Forest	**	*	**	***	***
Gradient Tree Boosting			**	**	***

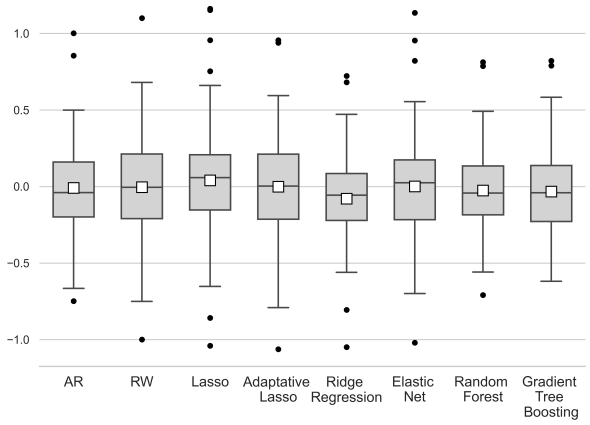
* 10% statistical significance
** 5% statistical significance
*** 1% statistical significance

Table 2 – Diebold-Mariano test statistical significance against the benchmarks - Complete sample (with unemployment data)

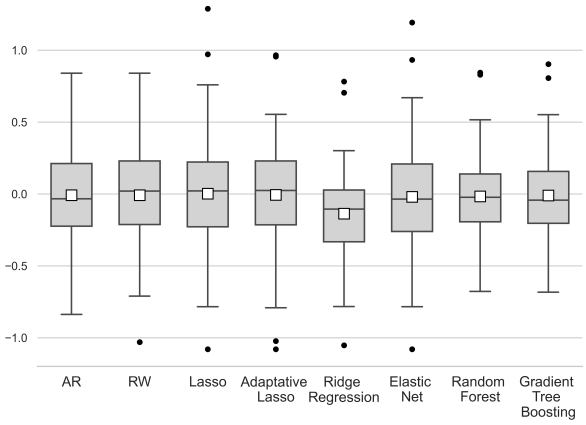
	t+1	t+3	t+6	t+9	t+12	average
AR	0.358	0.436	0.414	0.420	0.446	0.415
RW	0.423	0.443	0.485	0.520	0.480	0.470
LASSO	0.360	0.450	0.545	0.513	0.410	0.456
AdaLASSO	0.425	0.444	0.463	0.510	0.503	0.469
Ridge Regression	0.306	0.359	0.410	0.399	0.377	0.370
Elastic Net	0.391	0.469	0.525	0.489	0.531	0.481
Random Forest	0.320	0.333	0.362	0.370	0.375	0.352
Gradient Tree Boosting	0.366	0.361	0.376	0.371	0.314	0.358

The table shows the error interquantile range for each of the forecast horizons. The values are multiplied by 100. The shaded cells indicate the models with the lowest interquantile range for each horizon.

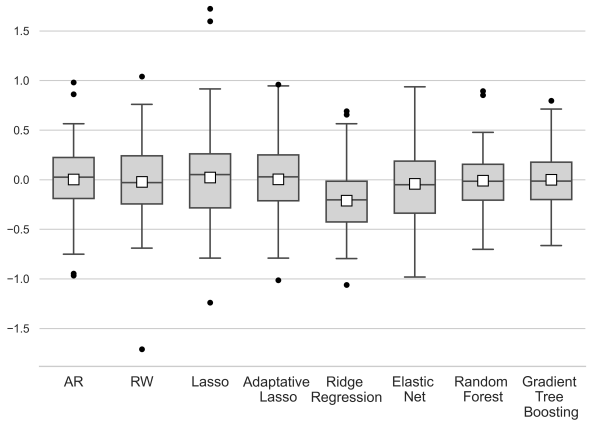
Table 3 – Error interquantile range - Complete sample (with unemployment data)



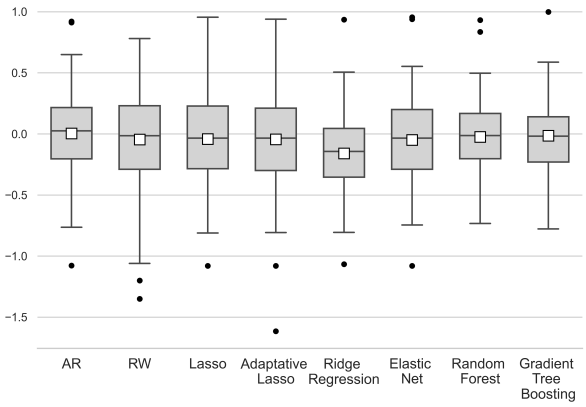
(a) t+1



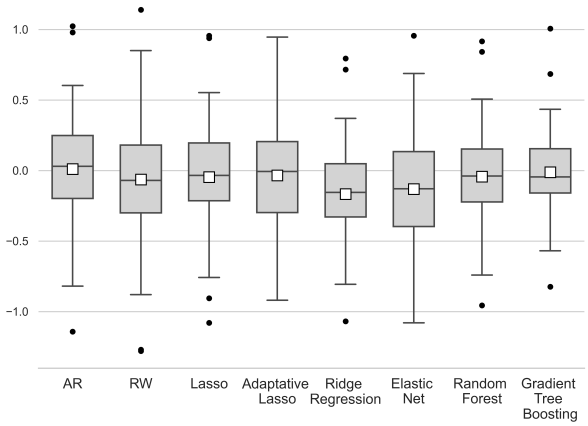
(b) t+3



(c) t+6



(d) t+9



(e) t+12

Boxplots for the errors, considering each forecast horizon. The white square represents the mean value. The whiskers are 1.5 times the interquantile range.

Figure 1 – Error distribution - Complete sample (with unemployment data)

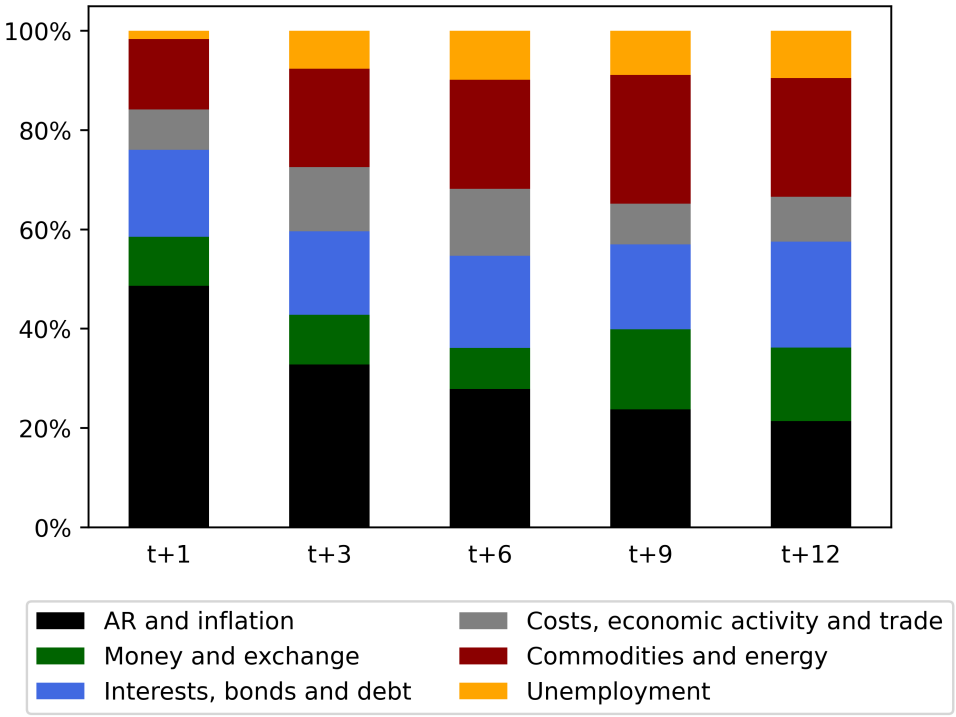


Figure 2 – Variable importance for random forest - Complete sample (with unemployment data)

5.2 Sample without unemployment data

To assess the relevance of unemployment data on industrial goods inflation forecasting, we repeat the tests for all models after removing it from the list of regressors. Although machine learning methods seek to automate as many decisions as possible when looking for patterns in the data, it is misleading to think that including every possible data is the best alternative, as economic theory and content expertise are important to guide where the algorithm looks for structure (Mullainathan; Spiess, 2017). The unemployment-inflation relationship is the basis for the traditional Phillips curve and has played an important role in empirical macroeconomics (Stock; Watson, 1999). However, many aspects of its dynamics are still unexplained, and some authors question its relevance (Mankiw, 2001). Atkeson and Ohanian (2001) show that even state of the art (by the time the article was written) NAIRU based Phillips curves could not forecast US inflation four quarters ahead better than a random walk. Also working with US inflation, Stock and Watson (2008) consider the performance of unemployment based Phillips curves to be episodic.

There are also discussions about the relevance of unemployment data to forecast Brazilian inflation (Sachsida, 2013). Mendonça et al. (2012) find that the effect of unemployment over inflation is close to zero, even though in some cases it is possible to observe the expected negative correlation in the short-term, depending on the proxies. Using high dimensional models to forecast IPCA and IGP-M, Medeiros et al. (2016) also find that unemployment is not among the the most relevant variables and consider it to be evidence that inflation mechanisms in Brazil are not those stated by the Phillips curve, especially when it comes to the unemployment-inflation relationship.

Table 4 shows the MSE and MAE after the removal of unemployment data. Random forest continues to rank first or second in terms of smallest MSE and MAE in all forecasting horizons, with negligible increase on the error (increase in MAEx1000 and MSEx1000 ≤ 0.005). Gradient tree boosting has the best performance for t+9 and t+12. Table 5 reports that, once again, random forest is the only model that beat the forecasts of the two benchmarks with statistical significance of at least 10%.

The removal of unemployment data has minor effects on the results of the nonlinear methods (which continue to have better overall performance) and conflicting contributions among the forecasting horizons on LASSO based models. However, there are improvements in the MSE and MAE for ridge regression and elastic net in all forecasting horizons.

Ridge regression is the best performing method in terms of MSE and MAE for t+1 horizon, the method with smallest MAE for t+3 horizon and the second best model for t+3 and t+6 in terms of MSE. It also reports the smallest MSE and MAE for all forecasting horizons among the ML regularized linear models. The method beats both benchmarks in all forecasting horizons in terms of MAE and MSE. Table 5 shows that

the forecast without unemployment data reaches at least 10% statistical significance in Diebold-Mariano test in three horizons against AR models and four horizons against RW[1].

Table 6 and Figure 3 show that there are no major changes in the error interquantile range relative to the sample with unemployment data. Futhermore, the boxplots confirm that there are only subtle changes to the sample error distributions for random forest and gradient tree boosting, whereas it is possible to observe the mean and median error reduction for ridge regression.

Figure 4 shows the variable groups importance for random forest using the methodology described in subsection 5.1. As in the complete sample, the AR and inflation related features continue as the most important for shorter horizons, gradually decreasing in importance towards longer horizons, when commodities and energy gain relevance. Compared to the variable importances of the sample with unemployment data, the category of interests, bonds and debt has the bigger increase in importance, followed by the categories of AR and inflation features and commodities and energy.

We do not intend to assess the causal inference between inflation and unemployment in this work. There are methodologies and instruments much more suitable for this. However, we have further indication that either (i) unemployment data may not be relevant to forecast industrial goods inflation when using a large data set, even though ML methods frequently select it; or (ii) the influence of unemployment on industrial inflation is episodic; or (iii) IBGE's open unemployment is not a good proxy to forecast inflation. We also have evidence that the variable selection for random forest and gradient tree boosting perform better than for the shrinkage methods to forecast industrial goods inflation.

[1] With the complete sample, ridge regression forecasts has statistical significance of at least 10% on Diebold-Mariano test only for $t + 12$ horizon against RW (and none against the AR model).

MSE x 1000	Forecast Horizon				
(MAE x 1000)	t+1	t+3	t+6	t+9	t+12
AR	0.808 (2.174)	0.976 (2.503)	1.225 (2.681)	1.361 (2.840)	1.444 (2.927)
Random Walk	1.097 (2.572)	1.165 (2.703)	1.615 (3.078)	1.857 (3.286)	1.977 (3.384)
LASSO	1.459 (2.806)	1.393 (2.816)	1.716 (3.052)	1.495 (3.062)	1.606 (3.074)
AdaLASSO	1.024 (2.513)	1.282 (2.725)	1.766 (3.008)	1.355 (2.838)	1.585 (3.016)
Ridge Regression	0.616 (1.880)	**0.793** (2.098)	**0.855** (2.276)	0.957 (2.438)	1.147 (2.736)
Elastic Net	0.925 (2.332)	1.316 (2.760)	1.336 (2.858)	1.284 (2.847)	1.492 (3.027)
Random Forest	**0.662** **(1.999)**	0.774 **(2.182)**	0.845 (2.269)	**0.910** **(2.331)**	**0.997** **(2.443)**
Gradient Tree Boosting	0.738 (2.176)	0.810 (2.200)	0.876 **(2.274)**	0.904 (2.276)	0.828 (2.207)

The table shows the mean squared errors and mean absolute errors (in parentheses) for each of the forecast horizons. The values are multiplied by 1000. The shaded cells indicate the models with the lowest MSE and MAE for each horizon. The values in bold represent the models with the second smallest MSE and MAE for each horizon.

Table 4 – Mean Squared Error and Mean Absolute Error - Sample without unemployment data

Against AR	Diebold-Mariano test: statistical significance				
	t+1	t+3	t+6	t+9	t+12
LASSO	**	***			
AdaLASSO	*	**			
Ridge Regression	**		*	**	
Elastic Net		*			
Random Forest	**	***	***	***	***
Gradient Tree Boosting			*	**	***

Against RW	Diebold-Mariano test: statistical significance				
	t+1	t+3	t+6	t+9	t+12
LASSO					
AdaLASSO					
Ridge Regression	***		*	**	***
Elastic Net					
Random Forest	**	*	**	***	***
Gradient Tree Boosting	**		**	***	***

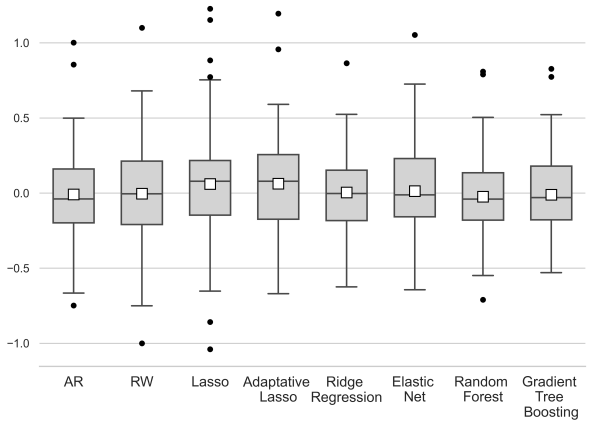
* 10% statistical significance
** 5% statistical significance
*** 1% statistical significance

Table 5 – Diebold-Mariano test statistical significance against the benchmarks - Sample without unemployment data

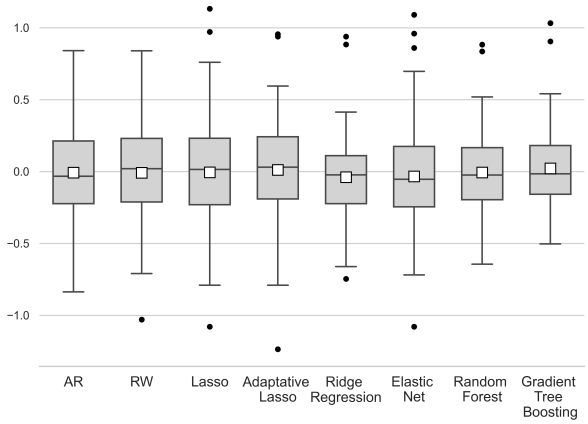
	t+1	t+3	t+6	t+9	t+12	average
AR	0.358	0.436	0.414	0.420	0.446	0.415
RW	0.423	0.443	0.485	0.520	0.480	0.470
Lasso	0.364	0.461	0.479	0.541	0.462	0.461
AdaLasso	0.430	0.433	0.414	0.437	0.497	0.442
Ridge Regression	0.336	0.334	0.368	0.380	0.399	0.363
Elastic Net	0.387	0.420	0.458	0.462	0.457	0.437
Random Forest	0.316	0.363	0.354	0.388	0.357	0.355
Gradient Tree Boosting	0.357	0.338	0.382	0.365	0.378	0.364

The table shows the error interquantile range for each of the forecast horizons. The values are multiplied by 100. The shaded cells indicate the models with the lowest interquantile range for each horizon.

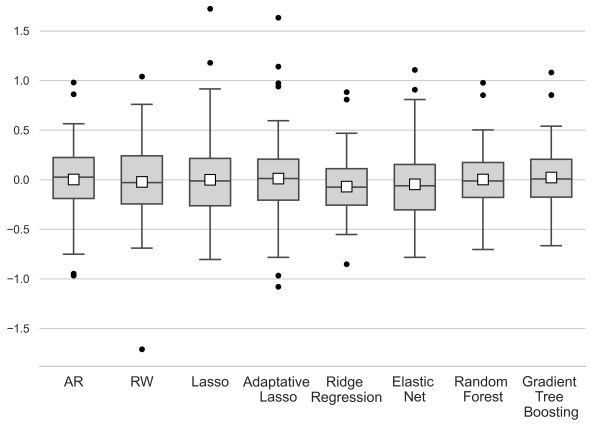
Table 6 – Error interquantile range - Sample without unemployment data



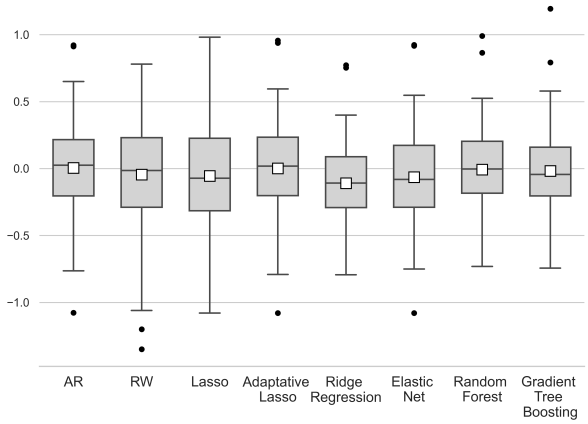
(a) t+1



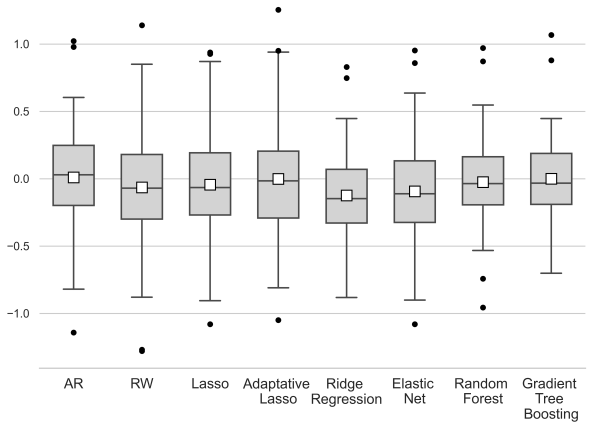
(b) t+3



(c) t+6



(d) t+9



(e) t+12

Boxplots for the errors, considering each forecast horizon. The white square represents the mean value. The whiskers are 1.5 times the interquantile range.

Figure 3 – Error distribution - Sample without unemployment data

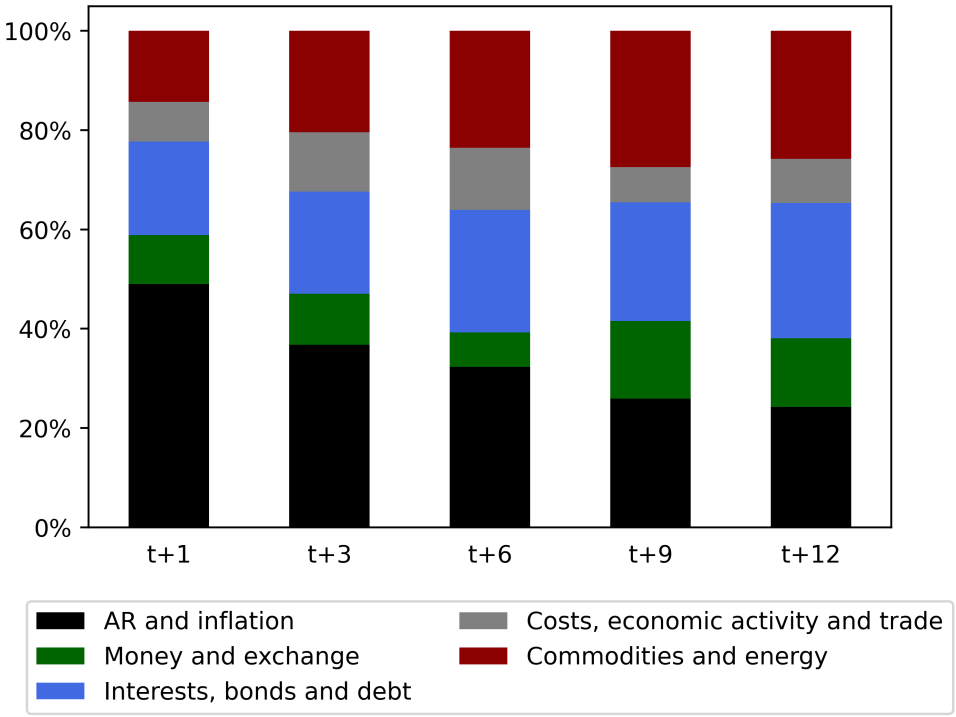


Figure 4 – Variable importance for random forest - Sample without unemployment data

6 Summary and conclusion

We forecast industrial goods inflation on Brazilian IPCA using machine learning methods and find further evidence that ML can be a valuable tool. Nonlinear tree based methods perform better than linear shrinkage methods and outperform the benchmarks RW and AR models. Ridge regression performs better than LASSO based models, probably due to the high correlation among the regressors, which include several industrial costs, commodity prices, inflation indexes and some of its breakdowns. Random Forest stands out with lower errors and a more uniform performance across the forecasting horizons, being the method that better controls the bias-variance trade-off. The results give further evidence that the nonlinear methods are able to better represent the complex dynamics of inflation, in line with the literature on Brazilian headline inflation forecasting. Regarding the variable importance for random forest, we find that autoregressive components and inflation related features are the most important for shorter horizons, gradually decreasing in importance towards longer horizons, when commodities and energy gain relevance.

Considering the forecasts without unemployment data as regressors, we find that the variable selection mechanisms of random forest and gradient tree boosting perform better than on linear regularized models to forecast industrial goods inflation. We also find further indication that either: (i) unemployment data may not be relevant to forecast industrial goods inflation when using a large data set, even though ML methods frequently select it; or (ii) the influence of unemployment on inflation is episodic; or (iii) IBGE's open unemployment is not a good proxy to forecast inflation.

One possible extension to this work is to use machine learning methods to forecast the other inflation breakdown segments (administered prices, services and food at home) and assess whether there are improvements with disaggregated inflation forecasting using combinations of the best methods for each segment. Another possibility is to explore other labor market proxies, such as labor market flow, unemployment duration, underemployment rate and discouraged workers rate, for a more in depth assessment of the effects of employment data on inflation forecasting using machine learning methods.

Bibliography

Alpaydin, E. *Introduction to Machine Learning*. 3. ed. : MIT Press, 2014. Cited on page 24.

Altissimo, F.; Mojon, B.; Zaffaroni, P. Can aggregation explain the persistence of inflation? *Journal of Monetary Economics*, n. 56, p. 231–241, 2009. Cited on page 17.

Aoki, K. Optimal monetary policy responses to relative-price changes. *Journal of Monetary Economics*, v. 48, p. 55–80, 2001. Cited on page 17.

Araujo, G. S.; Gaglianone, W. P. Machine learning methods for inflation forecasting in Brazil: new contenders versus classical models. *CEMLA - XXV Meeting of The Central Bank Researchers Network, October 2020*, 2020. Cited 2 times on pages 19 and 32.

Athey, S. The impact of machine learning on economics. In: Agrawal, A.; Gans, J.; Goldfarb, A. (Ed.). *The Economics of Artificial Intelligence: An Agenda*. : University of Chicago Press, 2019. p. 507–547. Cited on page 17.

Athey, S.; Imbens, G. W. Machine learning methods that economists should know about. *Annual Review of Economics*, v. 11, p. 685–725, 2019. Cited on page 25.

Atkeson, A.; Ohanian, L. E. Are phillips curves useful for forecasting inflation? *Federal Reserve Bank of Minneapolis Quarterly Review*, v. 25, n. 1, p. 2–11, 2001. Cited 2 times on pages 21 and 37.

Bansal, R.; Dahlquist, M. The forward premium puzzle: different tales from developed and emerging economies. *Journal of International Economics*, v. 51, p. 115–144, 2000. Cited on page 17.

BCB. Evolução dos preços relativos no IPCA. *Estudos Especiais do Banco Central*, Banco Central do Brasil, n. 36, 2019. Cited on page 17.

Bergmeir, C.; Benitez, J. M. On the use of cross-validation for time series predictor evaluation. *Information Sciences*, v. 191, p. 192–213, 2012. Cited on page 27.

Bishop, C. M. *Pattern recognition and machine learning*. 1. ed. : Springer, 2006. Cited 2 times on pages 17 and 22.

Breiman, L. Random forests. *Machine Learning*, Springer, n. 45, p. 5–32, 2001. Cited on page 24.

Cavallo, A.; Rigobon, R. The Billion Prices Project: Using online prices for measurement and research. *Journal of Economic Perspectives*, v. 30, n. 2, p. 151–178, 2016. Cited on page 18.

Chakraborty, C.; Joseph, A. Machine learning at central banks. *Staff Working Paper*, Bank of England, n. 674, 2017. Cited 2 times on pages 19 and 32.

Chen, T.; Guestrin, C. XGBoost: A scalable tree boosting system. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, Association for Computing Machinery, p. 785–794, August 2016 2016. Cited on page 25.

Clark, T. E. Disaggregate evidence on the persistence of consumer price inflation. *Journal of Applied Econometrics*, n. 21, p. 563–587, 2006. Cited on page 17.

Clark, T. E.; McCracken, M. W. Tests of equal forecast accuracy and encompassing for nested models. *Journal of Econometrics*, n. 105, p. 85–110, 2001. Cited on page 28.

Diebold, F. X. Comparing predictive accuracy, twenty years later: A personal perspective on the use and abuse of Diebold–Mariano tests. *Journal of Business and Economic Statistics*, v. 1, n. 33, p. 1–9, 2015. Cited on page 27.

Diebold, F. X.; Mariano, R. Comparing predictive accuracy. *Journal of Business and Economic Statistics*, n. 13, p. 253–265, 1995. Cited on page 28.

ECB. What is behind the change in the gap between services price inflation and goods price inflation? *ECB Economic Bulletin*, European Central Bank, n. 5, 2019. Cited on page 18.

Efron, B.; Hastie, T. *Computer Age Statistical Inference*. 1. ed. : Cambridge University Press, 2016. Cited on page 25.

Elliott, G.; Timmermann, A. Economic forecasting. *Journal of Economic Literature*, v. 1, n. 46, p. 3–56, 2008. Cited on page 17.

Fan, J.; Li, R. Variable selection via nonconcave penalized likelihood and is Oracle properties. *Journal of the American Statistical Association*, v. 96, n. 456, p. 1348–1360, 2001. Cited on page 23.

Faust, J.; Wright, J. H. Forecasting inflation. In: Elliot, G.; Timmermann, A. (Ed.). *Handbook of Economic Forecasting*. : Elsevier, 2013. v. 2A, p. 2–56. Cited on page 18.

Friedman, J. H. Greedy function approximation: A gradient boosting machine. *The Annals of Statistics*, v. 29, n. 5, p. 1189–1232, 2001. Cited on page 25.

Friedman, J. H.; Hall, P. On bagging and nonlinear estimation. *Journal of Statistical Planning and Inference*, Elsevier, v. 137, p. 669–683, 2007. Cited on page 25.

Garcia, M. G. P.; Medeiros, M. C.; Vasconcelos, G. F. R. Real-time inflation forecasting with high-dimensional models: The case of Brazil. *International Journal of Forecasting*, Elsevier, v. 33, n. 3, p. 679–693, 2017. Cited 2 times on pages 19 and 31.

Giacomini, R.; White, H. Tests of conditional predictive ability. *Econometrica*, v. 74, n. 6, p. 1545–1578, 2006. Cited 2 times on pages 28 and 31.

Hastie, T.; Tibshirani, R.; Friedman, J. *The elements of statistical learning*. 2. ed. : Springer, 2009. Cited 2 times on pages 22 and 23.

Hoerl, A.; Kennard, R. Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics*, v. 12, n. 1, p. 55–67, 1970. Cited on page 22.

Hubrich, K. Forecasting Euro area inflation: Does aggregating forecasts by HICP component improve forecast accuracy? *International Journal of Forecasting*, n. 21, p. 119–136, 2005. Cited on page 17.

Izenman, A. J. *Modern multivariate statistical techniques: Regression, classification and manifold learning*. : Springer, 2008. Cited 3 times on pages 17, 24, and 25.

James, G. et al. *An Introduction to Statistical Learning*. 1. ed. : Springer, 2013. Cited on page 25.

Kleinberg, J. et al. Prediction policy problems. *American Economic Review*, v. 105, n. 5, p. 491–95, 2015. Cited on page 19.

Mankiw, N. G. The inexorable and mysterious tradeoff between inflation and unemployment. *The Economic Journal*, Royal Economic Society, v. 111, p. C45–C61, 2001. Cited 2 times on pages 19 and 37.

Masini, R. P.; Medeiros, M. C.; Mendes, E. F. Machine learning advances for time series forecasting. *Jounal of Economic Surveys*, Wiley, p. 1–36, 2021. Cited on page 22.

Medeiros, M. C.; Vasconcelos, G. F. R.; Freitas, E. H. Forecasting brazilian inflation with high dimensional models. *Brazilian Review of Econometrics*, v. 36, n. 2, p. 223–254, 2016. Cited 3 times on pages 18, 21, and 37.

Medeiros, M. C. et al. Forecasting inflation in a data-rich environment: The benefits of machine learning methods. *Journal of Business and Economic Statistics*, v. 39, n. 1, p. 98–119, 2021. Cited 2 times on pages 19 and 32.

Mendez-Civieta, A.; Aguilera-Morillo, M.; Lillo, R. E. Adaptive sparse group LASSO in quantile regression. *Advances in Data Analysis and Classification*, Springer, n. 15, p. 547–573, 2021. Cited on page 28.

Mendonça, M. J. C.; Sachsida, A.; Medrano, L. A. T. Inflação versus desemprego: novas evidências para o Brasil. *Economia Aplicada*, v. 16, n. 3, p. 475–500, 2012. Cited on page 37.

Mishkin, F. Inflation targeting in emerging-market countries. *American Economic Review*, v. 90, p. 105–109, 2000. Cited on page 17.

Mullainathan, S.; Spiess, J. Machine learning: An applied econometric approach. *Journal of Economic Perspectives*, v. 31, n. 2, p. 87–106, 2017. Cited 2 times on pages 17 and 37.

Reis, R.; Watson, M. W. Relative goods' prices, pure inflation and the Phillips correlation. *American Economic Journal*, v. 2, p. 128–157, 2010. Cited on page 17.

Sachsida, A. Inflação, desemprego e choques cambiais: uma revisão da literatura sobre a curva de Phillips no Brasil. *Revista Brasileira de Economia*, v. 67, n. 4, p. 549–559, 2013. Cited on page 37.

Stock, J. H.; Watson, M. W. Forecasting inflation. *Journal of Monetary Economics*, Elsevier, n. 44, p. 293–335, 1999. Cited 2 times on pages 19 and 37.

Stock, J. H.; Watson, M. W. Phillips curve inflation forecasts. *NBER Working Papers, September 2008*, n. 14322, 2008. Cited on page 37.

Svensson, L. Inflation targeting. In: Friedman, B.; Woodford, M. (Ed.). *Handbook of Monetary Economics*. : Elsevier, 2010. v. 3, p. 1237–1302. Cited on page 17.

Tashman, L. J. Out-of-sample tests of forecasting accuracy: an analysis and review. *International Journal of Forecasting*, v. 16, p. 437–450, 2000. Cited on page 27.

Tibshirani, R. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological)*, Wiley Online Library, v. 58, n. 1, p. 267–288, 1996. Cited 2 times on pages 23 and 31.

Varian, H. R. Big data: new tricks for econometrics. *Journal Economic Perspectives*, v. 28, n. 2, p. 3–28, 2014. Cited 2 times on pages 17 and 25.

Zhao, P.; Yu, B. On model selection consistency of Lasso. *Journal of Machine Learning Research*, v. 7, p. 2541–2563, 2006. Cited on page 23.

Zou, H. The adaptative lasso and its oracle properties. *Journal of the American Statistical Association*, American Statistical Association, v. 101, n. 476, p. 1418–1429, 2006. Cited 2 times on pages 23 and 27.

Zou, H.; Hastie, T. Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: Series B*, v. 67, n. 2, p. 301–320, 2005. Cited 3 times on pages 23, 24, and 31.

Appendix

APPENDIX A – List of variables

List of variables	Original source
IPCA	IBGE
IPCA - Industrial Goods Segment	BCB
IPCA - Industrial goods core	BCB
IPCA - Industrial goods - Volatile items	BCB
Commodity Index - Brazil (IC-Br)	BCB
Commodity Index - Brazil Agriculture (IC-Br - Agriculture)	BCB
Commodity Index - Brazil Energy (IC-Br - Energy)	BCB
Commodity Index - Brazil Metal (IC-Br - Metal)	BCB
General Price Index - Market (IGP-M)	FGV
Wholesale Price Index-Market (IPA-M)	FGV
Wholesale Price Index-Market (IPA-M) - Agriculture	FGV
Wholesale Price Index-Market (IPA-M) - Industrial goods	FGV
Wholesale Price Index-Market (IPA-M) - Extractive industry	FGV
Wholesale Price Index-Market (IPA-M) - Transformation industry	FGV
Wholesale Price Index-Market (IPA-M) - Ethyl ethanol	FGV
Wholesale Price Index-Market (IPA-M) - Anhydrous ethyl ethanol	FGV
Wholesale Price Index-Market (IPA-M) - Computer hardware, electronic and optical products	FGV
CPI USA All items	U.S. Bureau of Labor Statistics
CPI USA All items 12M	OECD
M0	BCB
M1	BCB
M2	BCB
M3	BCB
M4	BCB

List of variables	Original source
Total exports	MDIC
EMBI+Brazil (Emerging Markets Bond Index Plus Brazil)	JP Morgan
Selic	BCB
Federal Funds Effective Rate - FEDFUNDS	Federal Reserve
Libor rate - 3M	ICE Benchmark Administration
Average exchange rate (USD/BRL)	BCB
Power purchasing parity (USD/BRL)	Ipea
Total electricity consumption	Eletrobras
Electricity average price	Eletrobras
U.S. Gulf Coast Conventional Gasoline Regular Spot Price FOB	U.S. EIA
Total fuel sales - 12 months	ANP
Gasoline sales - 1 month	ANP
Ethanol sales - 1 month	ANP
Diesel fuel sales - 1 month	ANP
Deep Sea Freight Transportation Services	U.S. Bureau of Labor Statistics
Producer Price Index by Industry: Total Manufacturing Industries	U.S. Bureau of Labor Statistics
Producer Price Index by Commodity: All Commodities	U.S. Bureau of Labor Statistics
Central Bank Economic Activity Index (IBC-Br)	BCB
Current economic conditions index	Fecomercio - SP
Net public debt - total (% GDP)	BCB
Real wages - industry	CNI
Unemployment rate - open	IBGE

Focus survey - Market expectations	**Original source**
IPCA t+1 - TOP5 Mid-term median	BCB
IPCA t+2 - TOP5 Mid-term median	BCB
IPCA t+3 - TOP5 Mid-term median	BCB
IPCA t+4 - TOP5 Mid-term median	BCB
IPCA t+5 - TOP5 Mid-term median	BCB
IPCA t+6 - TOP5 Mid-term median	BCB
IPCA t+7 - TOP5 Mid-term median	BCB
IPCA t+8 - TOP5 Mid-term median	BCB
IPCA t+9 - TOP5 Mid-term median	BCB
IPCA t+10 - TOP5 Mid-term median	BCB
IPCA t+11 - TOP5 Mid-term median	BCB
IPCA t+12 - TOP5 Mid-term median	BCB
Exchange rate (USD/BRL) t+1 - TOP5 Mid-term median	BCB
Exchange rate (USD/BRL) t+2 - TOP5 Mid-term median	BCB
Exchange rate (USD/BRL) t+3 - TOP5 Mid-term median	BCB
Exchange rate (USD/BRL) t+4 - TOP5 Mid-term median	BCB
Exchange rate (USD/BRL) t+5 - TOP5 Mid-term median	BCB
Exchange rate (USD/BRL) t+6 - TOP5 Mid-term median	BCB
Exchange rate (USD/BRL) t+7 - TOP5 Mid-term median	BCB
Exchange rate (USD/BRL) t+8 - TOP5 Mid-term median	BCB
Exchange rate (USD/BRL) t+9 - TOP5 Mid-term median	BCB
Exchange rate (USD/BRL) t+10 - TOP5 Mid-term median	BCB
Exchange rate (USD/BRL) t+11 - TOP5 Mid-term median	BCB
Exchange rate (USD/BRL) t+12 - TOP5 Mid-term median	BCB