Pontifícia Universidade Católica
do Rio de Janeiro

Tito Guedes Bruni

Machine Learning performance under different predictors' transformations: forecasting 12-month-ahead Brazilian inflation

Monografia de Final de Curso

Orientador: Gilberto Boaretto

Declaro que o presente trabalho é de minha autoria e que não recorri, para realizá-lo, a nenhuma forma de ajuda externa, exceto quando autorizado pelo professor tutor.

Rio de Janeiro, Dezembro de 2022

Abstract

In this essay, we use three well known machine learning (ML) models, namely Random Forest (RF), least absolute shrinkage and selection operator (LASSO), and complete subset regression (CSR) to forecast 12-month-ahead Brazilian inflation. Since ML methods started to be used in the context of forecasting economic variables, academic research has been focused on (i) comparing the performance of different models and (ii) analysing which explanatory variables are more relevant. In this paper, we contribute to the literature by analysing if accumulating explanatory variables enhances the forecasts of accumulated inflation.

Keywords: machine learning, time series, forecast, inflation

Contents

1 Introduction

Accurately forecasting inflation is important for several reasons. First of all, modern central banks calibrate their economic policies based on expected inflation (Iversen et al, 2016). Bad forecasts would result in ineffective policies with high social costs. Secondly, many long-term contracts are set in nominal terms and therefore bad inflation forecasts would generate undesired uncertainty. Finally, expectations about future prices are a key factor for households' decisions concerning future consumption and investments.

In this paper we use Machine Learning (ML) methods to forecast 12-month-ahead Brazilian inflation. Namely, the models used were: random forest (Breiman, 2001), LASSO (Tibshirani, 1996), Ridge (McDonald, 2009) and CSR (Elliotti; Gargano; Timmermann, 2013). Different authors have already approached this subject. Forni et al. (2003) showed that multivariate models beat univariate models for 12-month forecasts in the main countries of Europe. More recently, Medeiros et al. (2021) showed ML models outperform univariate models specially when inflation is more volatile. In particular, they showed that random forest (RF) provides improvements of almost 25% in terms of the root mean squared error (RMSE) when compared to the random walk (RW) for 12-month forecasts.

This study is innovative because we accumulate explanatory variables in two to twelve months to investigate whether it would enhance our 12-month inflation's forecasts. Recently, Coulombe et al. (2021) showed that transformations in macroeconomic data can enhance forecasts' accuracy. However, the transformations used by the authors were different from the one we are using. Namely, the transformations they used were moving average factors (MAF) and moving average rotation (MARX).

Unlike other papers, our benchmark is not an univariate model. Our benchmark is FOCUS' expected inflation. FOCUS is a report which consists of the projections of many economic variables (such as inflation) and its projections take into consideration the forecasts of more than a hundred professional forecasters.

2 Data and Method

2.1 Data

Our data consists of 84 variables extracted from the Time Series Management System from the Brazilian Central Bank. This plataform gathers data from various sources such as the the Brazilian Institute of Geography and Statistics (IBGE) and the Brazilian Central Bank. We selected variables related to prices, commodities, economic activity, employment, electricity, confidence, finance, credit, government and international trade. The period of analysis goes from January 2006 until May 2022. Instead of extracting each series individually from the Central Bank's website, we used the *R-package* **GetBCBData**.

2.2 Method

We aim to analyse how 12-month ahead forecasts of yearly inflation rate change when we accumulate monthly independent variables in $h = 1, 2, ..., 12$ months. We applied three types of tranformations depending on the properties of each variable:

Transformation 1	Transformation 2	Transformation 3
$\boldsymbol{X}_t^h = \boldsymbol{X}_t - \boldsymbol{X}_{t-h}$	$\boldsymbol{X}_t^h = \left(\frac{\boldsymbol{X}_t}{\boldsymbol{X}_{t-h}} - 1\right)100$	$\boldsymbol{X}_t^h = \left(\left(\Pi_{t=1}^h 1 + \frac{\boldsymbol{X}_t}{100}\right) - 1\right)100$

Given we accumulate the inflation rate in 12 months, we lose the first eleven observations of our data sets. Also, given we are computing direct 12 months ahead forecasts, we lose the last eleven observations. After these operations, each of our data sets has 173 observations. We compute rolling window forecasts for the period of January 2019 to March 2022 which means 38 forecasts. Our windows have 134 observations.

Notice that to predict inflation at time $t + 12$ we use data that was available at time t. Therefore, if the value x_t of the variable X is only known at $t + 3$, our models are going to use x_{t-3} as if it was x_t.

2.3 Statistical Methods

To compute the predictions, we use six statistical methods with specific properties which allow us to deal with: variables' selection, non-linearity and dimensionality reduction without loss of information.

2.3.1 Shrinkage Methods

When fitting data by using least squares, the coefficients $\beta_0, \beta_1, ..., \beta_p$ are the ones that minimize the following function:

$$RSS = \sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{ij} \right)^2 \tag{2.1}$$

The Least Absolute Shrinkage and Selection Operator (or simply LASSO) is a statistical model similar to OLS but unlike OLS, it's able to exclude independent variables from the model when these variables are considered irrelevant. The LASSO is able to do this because of the presence of a *shrinkage penalty.* In particular, it penalizes the sum of the absolute values of the coefficients. The parameter that penalizes the number of independent variables is λ:

$$\hat{\boldsymbol{\beta}}_{\boldsymbol{LASSO}}(\lambda) = \arg\min_{\boldsymbol{\beta}} \left\{ RSS + \lambda \sum_{j=1}^{p} |\beta_j| \right\} \tag{2.2}$$

The Ridge regression is another method that implements shrinkage of the coefficients' values. However, unlike the LASSO, the Ridge regression is not able to implement variable selection. This happens because of the format of its *shrinkage penalty.*

$$\hat{\boldsymbol{\beta}}_{\boldsymbol{Ridge}}(\lambda) = \arg\min_{\boldsymbol{\beta}} \left\{ RSS + \lambda \sum_{j=1}^{p} \beta_j^2 \right\} \tag{2.3}$$

2.3.2 Random Forest

The Random Forest method allows to deal with non-linearity. It applies bootstrap to regression trees. It generates many regression trees by randomly selecting some of the variables and some of the observations. The goal is to generate regression trees which are considerably different from each other. In the end, it takes the average of the regression trees. Let B be the total number of trees. The Random Forest prediction is given by:

$$\frac{1}{B}\sum_{b=1}^{B} T_b^*(\mathbf{X})$$

where T_1^* is the first bootstrap regression tree, T_2^* is the second, and so on and so forth.

2.3.3 Complete Subset Regression

The Complete Subset Regression (CSR) consists of estimating a large number of linear regressions with a fixed number of explanatory variables and computing the mean of the predictions generated by all the models. Our data sets have 83 predictors and even if we estimated models with only four variables, there would be almost 2 million possible combinations of models to be computed. And given we are using rolling window predictions, we would have to compute all these models for each one of our 38 forecasts.

To deal with the complexity of the models we used the R-package *HDeconometrics* which has a function that previously selects 20 predictors (based on the *t-statistic* of the coefficients of each predictor obtained by estimating regressions) and compute all possible models with four explanatory variables.

2.4 Forecasts' performance

To measure the performance of our forecasts we compare the models based on three statistics: the root mean squared error (RMSE), the mean absolute error (MAE) and the median absolute deviation from the median (MAD). Notice that each error depends on the time t of the observation and it also depends on the model m used to estimate the prediction:

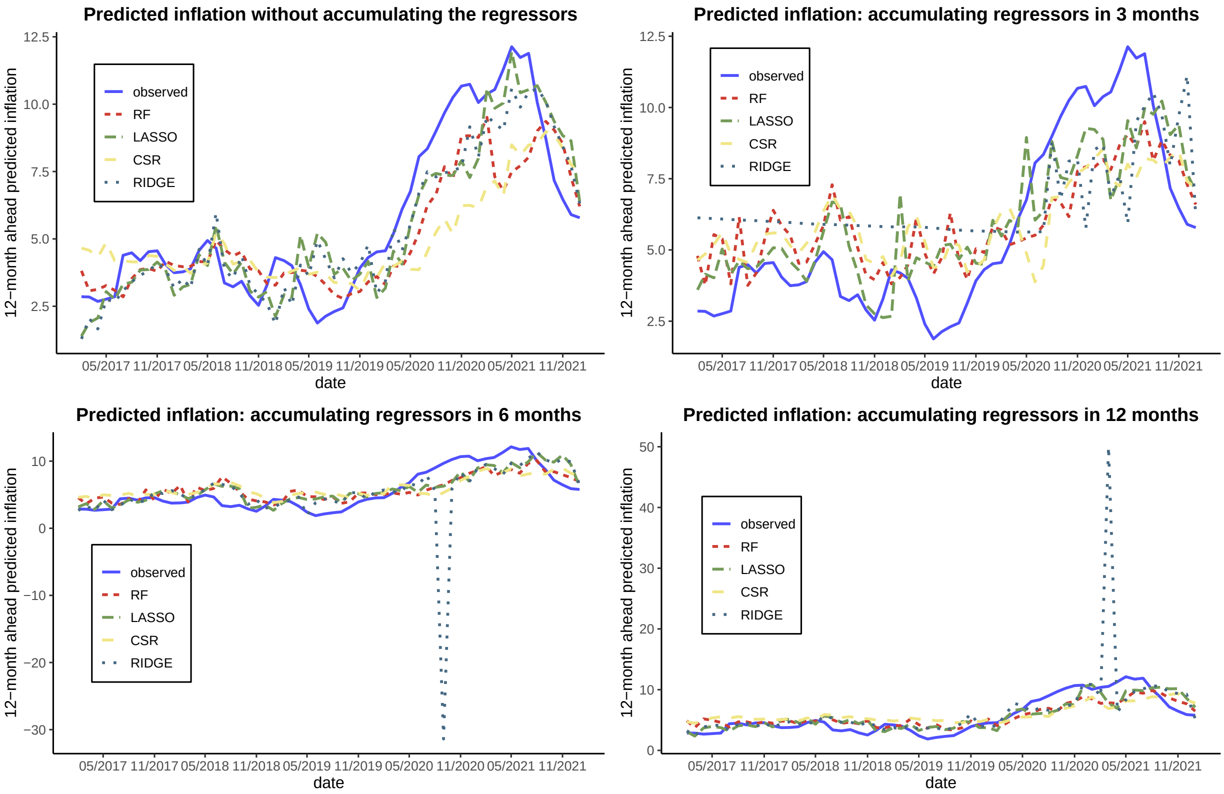
$$RMSE_m = \sqrt{\frac{1}{T - T_0 + 1} \sum_{t=T_0}^{T} \hat{e}_{t,m}^2}$$

$$MAE_m = \frac{1}{T - T_0 + 1} \sum_{t=T_0}^{T} |\hat{e}_{t,m}|$$

$$MAD_m = \text{median}\Big|\hat{e}_{t,m} - \text{median}(\hat{e}_{t,m})\Big|$$

3 Results

3.1 Trajectory



3.2 RMSE

We show that Random Forest (RF) and LASSO forecasts' are usually better and their accuracy does not seem to be impacted by the transformations (accumulations) implemented in the regressors. Therefore, accumulating explanatory variables does not enhance the quality of the forecasts of the 12-month-ahead Brazilian inflation. In addition, even though the *Ridge regression* performed poorly compared to the other models in the majority of the datasets, the best predictions were obtained when combining the *Ridge regression* with the dataset of monthly variables.

We chose the FOCUS's forecasts to be our benchmark. Therefore, to evaluate the quality of our models, we compare their RMSEs with the RMSE obtained from FOCUS's forecasts. We show that the machine learning models used usually beat FOCUS' forecast and this result does not depend on whether or not the explanatory variables were accumulated.

Table 1 – RMSE

	RF	LASSO	CSR	Ridge
1	1.695	1.459	2.202	1.378
2	2.031	1.700	2.236	3.046
3	2.006	1.786	2.248	2.590
4	1.969	1.907	2.286	2.240
5	1.982	1.801	2.286	6.632
6	1.949	1.875	2.277	5.599
7	1.953	7.544	2.263	9.218
8	1.868	6.605	2.268	13.450
9	1.851	5.862	2.234	12.705
10	1.750	1.648	2.231	5.903
11	1.735	1.627	2.250	3.963
12	1.620	1.670	2.182	5.352

Each line corresponds to a transformation in the regressors. For instance, line 3 has the RMSE of the models where regressors were accumulated in 3 months.

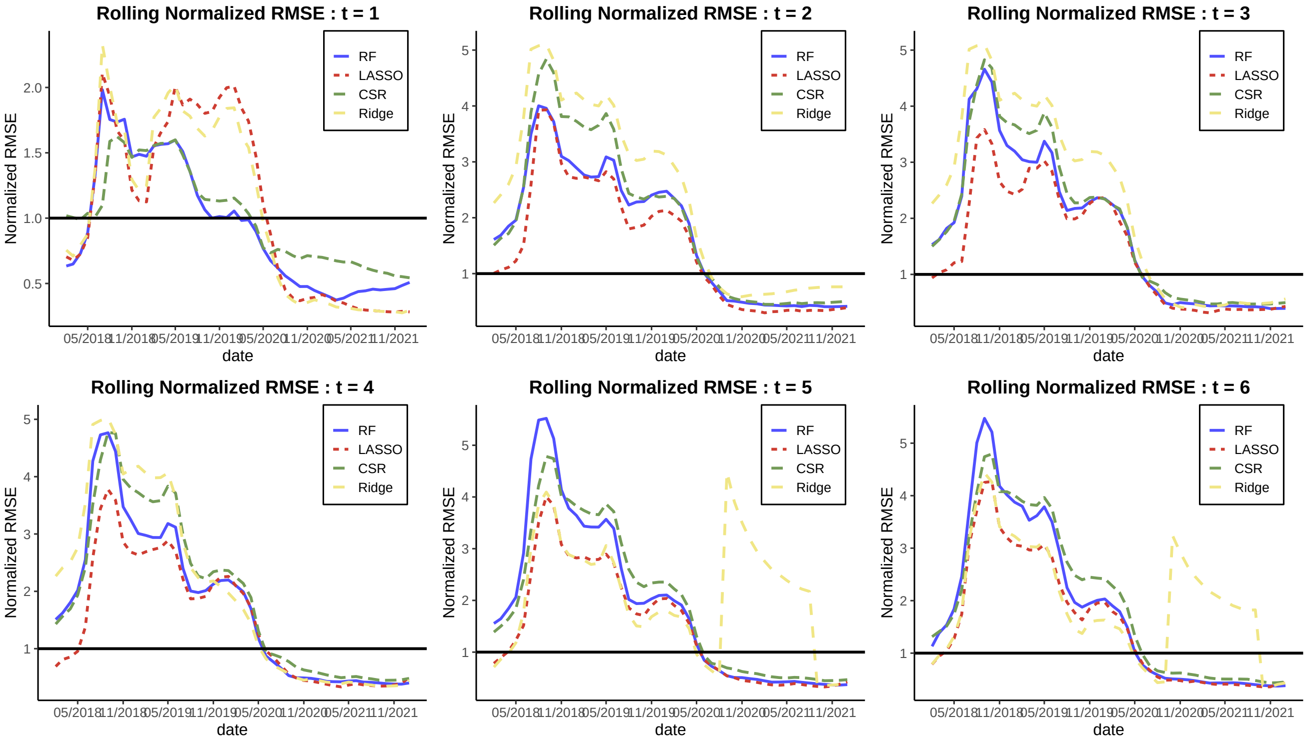
Table 2 – Normalized RMSE

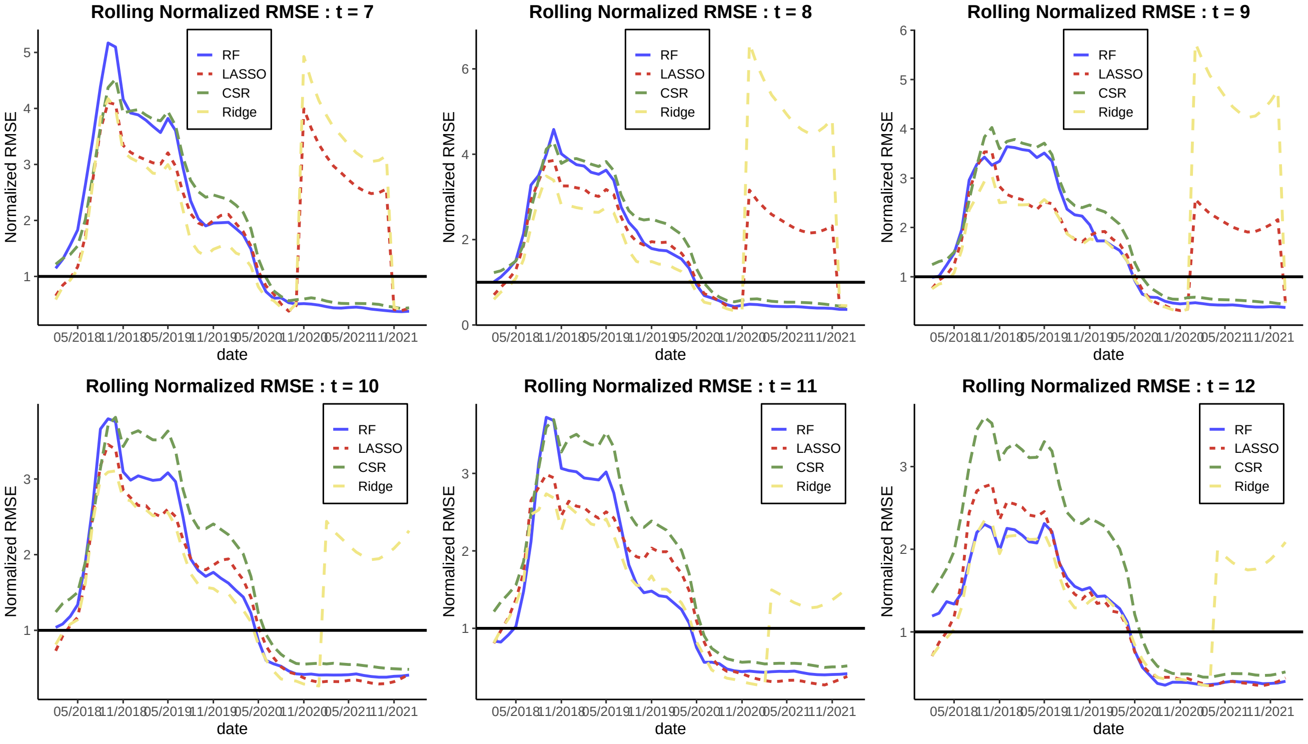
	RF	LASSO	CSR	RIDGE
1	0.499	0.429	0.648	0.406
2	0.598	0.500	0.658	0.896
3	0.590	0.526	0.662	0.762
4	0.579	0.561	0.673	0.659
5	0.583	0.530	0.673	1.952
6	0.574	0.552	0.670	1.648
7	0.575	2.220	0.666	2.713
8	0.550	1.944	0.667	3.958
9	0.545	1.725	0.657	3.739
10	0.515	0.485	0.657	1.737
11	0.511	0.479	0.662	1.166
12	0.477	0.491	0.642	1.575

We divide the RMSE of each model by the RMSE from FOCUS.

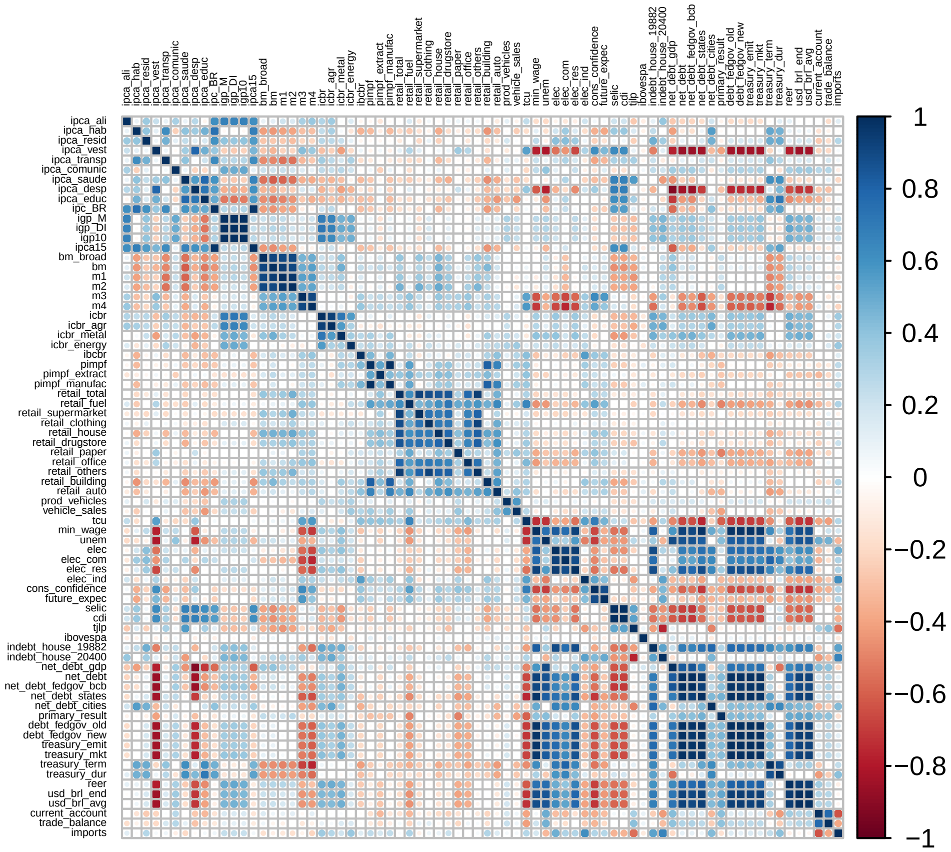
3.3 Rolling RMSE

To have a more dynamic view of the error measures, I estimated the values of RMSE, MAD and MAE in windows of 12 months:





Finally, I also generated a correlation matrix of the predictors:



Bibliography

BREIMAN, L. Random forests. *Machine learning*, Springer, v. 45, p. 5–32, 2001.

COULOMBE, P. G. et al. Macroeconomic data transformations matter. *International Journal of Forecasting*, Elsevier, v. 37, n. 4, p. 1338–1354, 2021.

ELLIOTT, G.; GARGANO, A.; TIMMERMANN, A. Complete subset regressions. *Journal of Econometrics*, Elsevier, v. 177, n. 2, p. 357–373, 2013.

FORNI, M. et al. Do financial variables help forecasting inflation and real activity in the euro area? *Journal of Monetary Economics*, Elsevier, v. 50, n. 6, p. 1243–1255, 2003.

IVERSEN, J. et al. Real-time forecasting for monetary policy analysis: The case of sveriges riksbank. *Riksbank Research Paper Series*, n. 142, 2016.

MCDONALD, G. C. Ridge regression. *Wiley Interdisciplinary Reviews: Computational Statistics*, Wiley Online Library, v. 1, n. 1, p. 93–100, 2009.

MEDEIROS, M. C. et al. Forecasting inflation in a data-rich environment: The benefits of machine learning methods. *Journal of Business & Economic Statistics*, Taylor Francis, v. 39, n. 1, p. 98–119, 2021. Disponível em: ⟨https://doi.org/10.1080/07350015.2019.1637745⟩.

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.