Bachelor Thesis

“Machine Learning Models and Inflation in USA: Accuracy of Models and relevance of Variables "

Álvaro Ortiz de Pazos

Student number: 11131

Academic Supervisor: Javier Cerezo

May 2023

Final paper submitted in partial fulfillment of the requirements

for the Degree of Bachelor of Economics

"I hereby declare that I understand and abide by IE University's policy concerning plagiarism; and that this thesis contains no material previously published, or written by another person or AI software, except where properly cited."

ABSTRACT

Inflation has become a relevant issue since the Covid-19 pandemic, as prices of goods and services have increased across the board. To address this issue, I explore the ability of several machine learning (ML) models including linear, regularized (Lasso, Ridge and Elastic Net), principal components and non-linear models including Random Forest and Ada Boost, to forecast inflation in the United States. Compared to previous works, I explore how these models work and which variables are relevant during the recent high inflation episode. The results show that Machine Learning models are useful to forecast inflation in the USA. The Non-Linear models (Random Forest & Ada Boost) are more accurate and better to capture the recent high inflation period. Moreover, I found that traditional variables and hypothesis such as price dynamics and the Phillips curve remains a solid explanation of inflation. However, the high inflation period looks driven more by inflation´s own dynamics and a weakening role of the traditional Phillips curve indicators. This could be the result of the non-linearity of this period.

KEYWORDS: Machine Learning, Inflation, Forecasting, Random Forest, Regularized models

Contents

1 Introduction

After years of relatively low prices, inflation has been a major issue since the pandemic. Since then, inflation has risen sharply affecting many advanced and emerging economies globally. The issue of inflation has come to the forefront of the political agenda, particularly with respect to monetary policy. In this sense, forecasting price inflation accurately in the short, medium and long term can be of great help for the implementation of monetary policies.

High inflation is normally bad for the economy including households and firms. Price increases are normally translated into a loss of purchasing power as nominal income does not normally match goods and services price increases. Thus, real income (a proxy for our standard of living), falls. High inflation also creates uncertainty as firms and families find it more difficult to make plans for the future which may lead to lower consumption, investment and finally employment. However, low inflation can also have negative consequences. When prices are falling, consumers delay purchases if they anticipate lower prices in the future. Moreover, firms can also delay investment resulting in low growth and employment creation. Japan is one country with a long period of nearly no economic growth, largely due to deflation. [1]

The US Inflation rate has experienced different episode during the last decade, from periods of abnormally low inflation after the Financial Crisis (an average of 0.15% monthly growth rate or 1.5% annualized) to accelerating inflation since the Covid-19 (when inflation reached an average of near 0.6% monthly or 6.6% annualized).

Inflation can change from low to high regimes and vice versa. Furthermore, accurately forecasting inflation is essential for central banks as they need to respond properly to changing inflation conditions and maintain optimal conditions for the economy.

In this thesis, I analyze the relevance of alternative Machine Learning Models (ML) and a wide range variables in forecasting the inflation in the USA over the past few years. To do this, I use a large database including different groups of variables which could potentially be used to forecast inflation. Among them, I include variables such the own price components (goods, services, transport...), producer and commodity prices, economic activity indicators,

[1] See Oner(2019) for a good description of the effects of inflation on the economy

labor market conditions, money and credit variables, financial indicators and housing prices data.. Once the database has been compiled, I test the forecasting accuracy of Machine learning models to forecast inflation.

Specifically, I estimate several machine learning models to evaluate the forecasting accuracy, including a linear model, regularized linear models such as Lasso, Ridge, and Elastic Net, a factor model such as principal component analysis, and non-linear models like Random Forest and Ada Boost. The estimation or training sample spans almost 44 years, from February 1968 to December 2011.

After estimating the models, I assess their forecasting accuracy at different horizons and compare their performance to that of a naive model during the testing sample, which spans almost ten years from January 2012 to July 2022. To evaluate the forecasting accuracy during different inflation regimes, I differentiate the testing sample into three episodes: the total test set (January 2012 to July 2022) which include the Covid-19 months, the low inflation test set (from January 2012 to January 2020) and the high inflation test set (from July 2020 to July 2022)."

Once the forecasting has been tested I use the "Shapley Values"[2] technique to check the relevance of the different variables in the best models. Furthermore, I analyze the importance of different models and variables in forecasting inflation at different inflation regimes.

Finally, after evaluating the results on the models and variables I will present the limitations and potential for further research, as well as the main conclusions.

[2]The Shapley Values technique was developed by Shapley(1951) and is a Machine Learning based on Game Theory to calculate the contribution of the independent variables to the forecast of model

2 Theoretical Framework: Recent Literature on Inflation and Machine Learning models

After years of relatively low prices we have been experiencing a increase of inflation rates worldwide. The surge of high inflation during the post-pandemic recovery was not anticipated for most of the analysts nor the international organizations. In general terms, there is no agreement on one specific reason behind this episode but some of them operating at the same time. Koch and Noureldin (2023) summarize some of more relevant factors behind the recent inflation episode:

- Demand Factors: Some analysts and authors have suggested that a more rapid economic recovery than expected has been a key driver of inflation. This may be due to the high uncertainty during the pandemic, leading to the implementation of a large-scale policy support. However, although the Covid-19 shock was very severe, it was short lived. As a result, the demand recovery was boosted by fiscal policies, fueling extra inflation.

- Supply factors and Global Value chains: These factors also played a role in the inflationary pressures. The rapid recovery in demand was accompanied by problems in global value chains. These included scarcities of some parts of goods, as a result of the necessary lock-downs implemented to fight the COVID-19 pandemic.

- Sectorial Reallocation: The lock-downs triggered important changes on the sectorial composition of the economy. Some goods and services were prohibited while other were boosted (i.e Teleworking, Food...) affecting relative prices.

- Very tight labor market (low unemployment): This continues to persist in some advanced economies and particularly in the US Economy. Low unemployment would be amplifying the demand pressures. This is consistent with the traditional Philips curve hypothesis suggesting an inverse relationship between inflation and unemployment: as labor market tightens and unemployment falls employers need to offer higher rates to attract or retain workers. The increase in wages will increase production costs which will be finally translated on to consumers in the form of high prices.[3]

[3]The original Phillips curve was proposed by A.Phillips which analyzed the wage and unmployment data in the UK Phillips (1958)

- Changing role of price expectations: Some authors (Reis, 2022) are suggesting that we are experiencing changes in the way the consumers and firms form their expectations on prices. For some of them expectations were high but were neglected due to overconfidence by central banks after several years achieving targets

- Extremely Loose Policies: The inflation revival could have been supported by loose monetary policies during a long period of time before the Covid reflected in low interest rates and high credit. Moreover, as mentioned large fiscal packages approved to fight the Covid had also expansionary stimulus.

Forecasting inflation is not an easy task and while some authors as Stock & Watson (1999) showed that many activity demand indicators worked as potential predictors of US inflation, Atkeson & Ohanian (2001) showed that in many cases, the Phillips curve (which relates excess demand from activity indicators or labor market) failed to beat simple naive models. These contrasting results inspired researchers to look for alternative models and variables.

While some authors look for alternative variables to explain inflation others began to explore the ability of models in forecasting inflation. The first group of researchers look for alternative variables. According to Madeiros et al(2019), some authors tested expectation variables (Groen et al.,2013), commodity prices (Chen et al., 2014), financial indicators (Forni et al., 2003) and, more recently, supply side or bottlenecks (Jesee and Santacreu,2022).

The literature on inflation forecasting models has been growing fast during the last years due to the improvement of technology and rapid development of machine learning techniques. Thanks to these advances and the development of the Big Data much larger databases have been collected, the processing capacity have increased and better algorithms are developed every year . In the case of Machine Learning Models to forecast inflation we have several examples.

For example, Garcia et al.(2017) applies different ML methods to forecast Brazilian inflation using ML models as LASSO, Ada-LASSO, or Random Forest. Medeiros et al (2019) focuses on the United States which allows them to work with a larger database, that the authors divided into eight different groups. This database is very similar to what I use in this paper. However, in this work I have updated the database until 2022, which allow me to test the models in this

recent high inflationary environment . Given that inflation started to accelerate after Covid-19, I have the opportunity to evaluate the forecasting accuracy and the relevance of variables in a higher inflation period, from half of 2020 to 2022.

Finally, a most recent paper (Joseph et al., 2021) focuses on the inflation in the UK using a similar method to my work to evaluate the influence of different variables ("Shapley Values"). Moreover, I will follow him to evaluate the contribution of the variables to explain inflation with the most accurate models of this work and compare the results of my work on the US economy with the United Kingdom.

In sum, with this work, I join the recent literature on using Machine Learning Models to forecast inflation. As some authors before, I evaluate the forecasting accuracy of some Machine Learning linear and non-linear models. I extend these works by including the most recent inflationary period and evaluate the relevance of different variables to explain inflation in both normal and high inflation episodes.

3 The data, the models and the methodology

In this section I outline the database, describe the different models I will use as candidates and explain the strategy to assess the forecasting accuracy of the models and the relevance of the variables under different regimes.

The main source of data is the FRED database developed and maintained by the US Federal Reserve of St. Louis as in Medeiros et al (2019). Once the data are cleaned and transformed, we estimate the Machine Learning Model to forecast inflation at different monthly horizons withing the Training Sample in Python using the scikit-learn package. Once the Models are estimated in the training sample, I evaluate their forecasting a accuracy with several metrics for errors (RMSE and MAE) and compare their performance against the Naive model. As I describe later, I will evaluate the results in different test sets, including a total inflation test set, a low inflation test set and high inflation test set.

Finally, I evaluate the relevance of the variables for the best models through the "Shapley Val-

ues" for the best model, checking the relevance of variables the different testing samples.

The dataset and the computer code can be found in github. The documentation is accessible in a notebook and is presented equivalent to common Python documentation. The link contains the entire code for the benchmark along with the other alternative models, as well as the code for the Shapley values and several additional tables, matrices, and graphs used in the thesis. Furthermore, some of the graphs are created in Excel using data imported from Python.
I describe deeper the data process, the estimation of machine learning models and the relevance of the variables in the following lines.

3.1 The Data

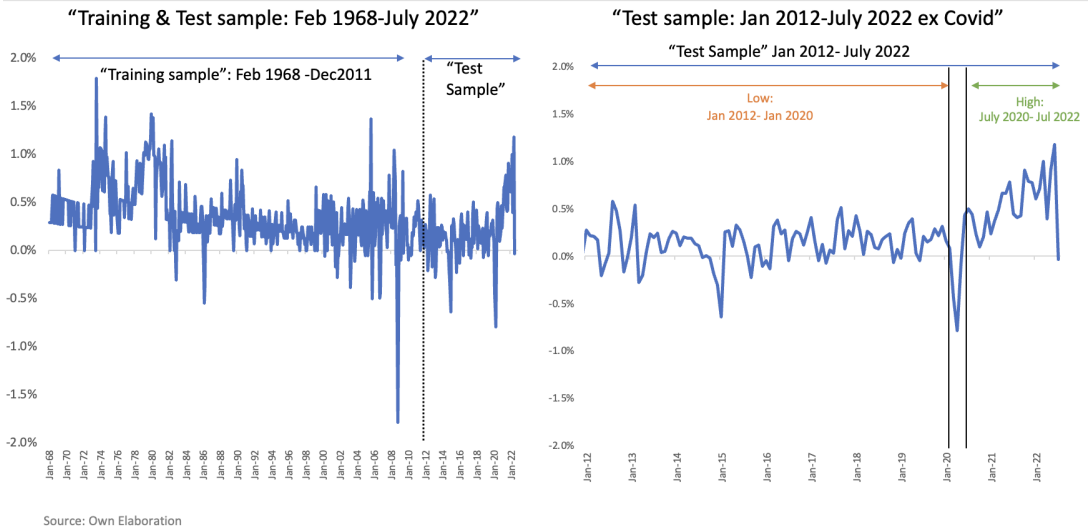
The main source of the data used in this thesis correspond to monthly indicators for the United States obtained from the Federal Reserve Economic Data base or FRED Database [4]. The database is an online database maintained by the Research division of the Federal Reserve Bank of St. Louis which includes hundreds of thousands of economic data time series from national, international, public, and private sources.

3.1.1 The Dependent Variable: Inflation under different regimes

The key variable ion this thesis is the US inflation rate. In particularly I focus in the the Consumer Price Index for All Urban Consumers (All items) which is produced and seasonally adjusted by the Bureau of Labor Statistics. In this paper I focus on the monthly rate of the seasonally adjusted series to avoid the yearly statistical base effects.

[4]The FRED database can be accessed at this link https://fred.stlouisfed.org/

Figure 1: US Inflation during Training & Testing Sample



Source: Own Elaboration

Table 1: Data information & Inflation

Data	Dates	Monthly Inflation	Year on Year Inflation
Total Sample	Jan 1968 - July 2022	0.33%	4.10%
Training Sample	Jan 1968 - Dec 2011	0.36%	4.50%
Test Sample	Jan 2012 - July 202	0.20%	2.50%
Low Inflation Test Sample	Jan 2012 - Jan 2020	0.14%	1.60%
High Inflation Test Sample	July 2020 - July 2022	0.55%	5.40%

Source:Own elaboration

The total sample included in this exercise extends from February 1968 to July 2022 (654 observations) and the average monthly rate of this period is equal to 0.33% (4.1% year or year). As described in Figure 1 and Table 1, the total sample is divided in various samples and periods:

- The Training Sample (1968-2011): The training sample will be used to estimate the models and runs from February 1968 until December 2011. It includes around 81% of the data, with an average inflation slightly higher than the total sample (0.36% monthly and 4.5% yearly)

- The Testing Sample (1968-2011): The testing sample include monthly data from January 2012 to the end of the total sample, July 2012. . The average inflation is lower (0.2% monthly and 2.5% yearly) and close to the actual US Federal Reserve inflation target of 2.0%. This period include the aftermath of the financial crisis and the large shock of Covid-19 which triggered important changes in the behavior of inflation. In order to deal with these different regimes I have divided this sample in two periods :

- The Low Inflation Test Sample (2012-2019): This sub-sample corresponds to the first period of the test sample including January 2012 to January 2020 (eight years). The US Economy was under the consequences of the Financial Crisis (2009) and the inflation rate undershoot the actual target as the inflation rate was significantly lower (0.14% monthly and 1.60% yearly) than the historical average. during some periods of 2014 and 2015 the economy was experiencing some deflationary episodes.
- The High Inflation Test sample Period (July 2020-2022): The last testing sample is the “high inflation test set" running from July 2020 to July 2022. During this period the economy has been living with the consequences of the Covid-19 which generated important changes on inflation. During this sample the inflation rate has been significantly higher (0.55% monthly and 5.40% yearly) clearly overshooting the target of the Federal Reserve.

The fact that inflation has been experiencing different regimes in the sample will allow to test the performance of the Machine Learning models and the explanatory variables under very different circumstances. The dependent variable (i.e inflation) will enter in the model forwarded

at six different horizons: 1, 2, 3, 4 , 5 and 6 months. The basic idea is that the current values of the explanatory variables should help to forecast inflation in future at different horizons.This strategy is in line with Garcia et al. (2017) and Medeiros et al (2019)[5]. have limited the horizons to six, as one of my primary contributions to the literature is to evaluate the accuracy of models and variables during the most recent high-inflation period, which is constrained by limited data availability.

3.1.2 The Explanatory Variables: Different Groups

The FRED data base contains a large numbers of variables for a long sample. The database included in this paper is the one suggested by Medeiros et al(2019) and include 99 variables differentiated in groups including the dependent variable and eight different groups of indepen-

[5]Medeiros et al. (2019) use additionally lagged variables of explanatory variables. Since an important part of my strategy is the evaluation of the recent inflationary period, I use only the independent variables at time t. As we will see later the results of both strategies for the common samples are similar

dent or explanatory variables. Moreover,all the variables are transformed[6] in order to have all the data in stationary terms. (See Tables 7 to Table 14 of the Appendix for a detailed description and transformation of the variables included in each of the groups). The different groups of the variables are the following:

- Price components: This group includes some of the key consumer price indexes of goods and services sub-indexes. It will help us to evaluate the relevance of some special groups and test also the role of the own dynamics of prices or expectations.

- Price costs: This include producer prices and key input prices of goods and materials as oil prices. This can be useful as a proxy for how the companies translate their cost to the consumer prices.

- Unemployment or Labor conditions: This is a key variable for the Central Banks and in particular the Federal Reserve. The conditions of the labor market are a good indicator on how likely workers demand wage increases when the labor market conditions are "tight" following a traditional Phillips Curve. This group also include average hourly earnings by sector.

- Activity Indicators: As the previous labor conditions group, the activity indicators are also a proxy of the "Phillips Curve" but rather than focusing on unemployment they express in terms of the "output-gap" (or activity demand gap over a level of potential output) which is also widely used in Central Bank Models to forecast inflation. When demand conditions are very buoyant the risks that wages and other input prices pass on to consumer prices is higher. Here we include several indicators such as industrial production index, or different industrial productions of goods and materials.

- Labor Force: Play a similar role than the activity indicators reflecting the state of the business cycle This group includes the number of employees by sector and average hourly earnings by sector.

[6]The different transformations used in this exercise are the ones suggested by Medeiros et al(2019) : (1) no transformation; (2) Δxt; (3) $\Delta 2xt$; (4) $\log(xt)$; (5) $\Delta \log(xt)$; (6) $\Delta 2 \log(xt)$; and (7) $\Delta(xt/xt-1 - 1)$. Therefore each table represents a group of variables and it contains the different variables, with each specific transformation and a small description of the variable.

- Money and credit: According to some economic streams of literature, in the medium run inflation is a monetary phenomena. In this group we include some variables to test the role of monetary variables in inflation.

- Interest rates: This group includes different treasury, corporate rates and the US stock market as a measure of how financial conditions can affect inflation. Moreover, interest rates are the main tool of the Central Banks to control inflation

- Housing market: These variables have been associated with inflation since the last financial crisis due to the potential of housing market to affect inflation or even generate deflation during the big housing crisis.

Given that this work include a wide set of variables and indicators, and with different economic meaning, this will facilitate the analysis of the relevance of several hypothesis of the drivers of inflation.

3.2 The Models

3.2.1 General expression of Machine Learning models of inflation

Consider the following general Model for monthly inflation at different h=1, 2, 3, 4, 5, 6 monthly horizons. Where π_{t+h} is the inflation in month t + h and $x_t = (x_{1t}, x_{2t}, \ldots, x_{nt})$ is a n-vector of explanatory variables and/or common factors; G_h represents different models between the x_t covariates at time t[7] and future inflation π_{t+h}; and u_{t+h} a zero mean random error.The function G_h can be a single model or an ensemble of different specifications: [8]

$$\pi_{t+h} = G_h(x_t) + u_{t+h} \quad h = 1, .., H, t = 1, 2, \ldots T. \ (1)$$

A general expression embedding the structure of most of the algorithms can be defined as the below one (see Mullainathanet and Spiess(2017). Basically, it includes a function class $f(x)$

[7]Medeiros et al also include the possibility of lagged variables as explanatory variables

[8]Rather than working in a rolling-window scheme as Medeiros et al. (2019) I opt to estimate on the training model and apply the estimated model to the complete sample. The reason for this is that as I will describe later, I will divide the sample in two sub-samples to deal with the structural break post covid.

which can include a linear function, a factor, a tree, neural network ... and a “Regularizer" or $R(f)$ which refers to how “complexity" (or having a large number of coefficients and variables in the model) will be treated. A diagram explaining the strategy can be observed below:

$$\underbrace{minimize \sum_{i=1}^{n} L(y\, y_{i,t+h}, f(X_{i,t})}_{\text{in-sqmple loss function}} over \overbrace{f \in F}^{\text{function class}} subject\ to \underbrace{R(f) \leq c}_{\text{complexity restriction}}$$

According to Mullainathan and Spiess (2017)[9] “picking the prediction function involves two steps: The first step is, conditional on a level of complexity, to pick the best in-sample loss-minimizing function and in a second step is to estimate the optimal level of complexity using empirical tuning". In the general form below, I select the deviance between the data and the fit of our model $(y_i - \beta X_i)$ as Loss function and we assume a a general function the regularizer $\lambda||f||_{l_q}$.

The “Regularizer" can be expressed as the product of the strength parameter of the regularization (λ) and its functional form (l_q).. Note that as we are adding a penalty on the size of each β_k, this strategy penalizes complexity as the β_k coefficients control the influence of input variables X_i to predict y_i.

We can express the problem in the general form of (5) and (6) which can be solved by numerical optimization methods (i.e Newton, Gradient Descent..).

$$Minimize \sum_{i=1}^{n} L(y_{i,t+h}, f(X_{i,t})) + \lambda||f||_{l_q} \tag{5}$$

$$\hat{\beta} = argmin \left\{ \sum_{i=1}^{n} (y_{i,t+h} - \beta X_{i,t})^2 + \lambda \sum_k c(\beta_k) \right\} \tag{6}$$

3.2.2 Multinomial Linear Regression model

The simplest regression algorithm is the Multinomial Linear regression model. This model is well suited to capture linear problems, but it is not prepared to deal with very complex models

[9]Mullainathan, S., & Spiess, J. (2017). Machine learning: An applied econometric approach. Journal of Economic Perspectives, 31(2), 87–106. https://doi.org/10.1257/jep.31.2.87

or nonlinear data.

In this model we assume that there a linear relationship between the dependent variable y_i, and the independent variables X_i or $m(\beta X_i)$. The optimization problem consists in finding the β values which minimize the loss function or the deviance of the true data and the fit of our model ($y_{i,t+h} - \beta X_{i,t}$). In this case, we assume no treatment (or zero strength) of regularization of the complexity ($\lambda = 0$) leading to the following simple solution of the estimators:

$$Minimize \sum_{i=1}^{n} L(y_{i,t+h}, m(\beta X_{i,t})) + \lambda ||f||_{l_q} \qquad (7)$$

$$\hat{\beta} = argmin \left\{ \sum_{i=1}^{n} (y_{i,t+h} - \beta X_{i,t})^2 \right\} \qquad (8)$$

3.2.3 Regularized Models: Lasso, Ridge, and Elastic Net

The Multinomial regression model assumes that there will not be treatment for the complexity ($\lambda = 0$) but if the number of the variables is large this can lead to “overfitting". This is a general problem in Machine Learning models including a large set of variables which leads to low "in sample" error but higher "out of the sample" in the testing sample.

As the number of covariates or explanatory variables in our exercise is large we can use some of the regularization strategies to deal with. To solve this problem we can add a cost penalty to the loss function (i.e $\lambda > 0$) or “Regularization“which can take different patterns of functional forms $c(\beta_k)$. Given the functional pattern of the regularization the estimation of the λ parameter is common. Depending on the form of the function c one can define different alternative models described below: The Lasso , Ridge and Elastic Net Models.

- **The Regularized Lasso Model**

  The LASSO (Least Absolute Shrinkage and Selection Operator) is a regression model used for feature selection and regularization in machine learning introduced by (Tibshirani,1996). The penalty introduced in the Lasso model encourages the model to select a subset of the most important features and shrink the coefficients of the less important

features towards zero.The lasso's absolute value penalty, $|\beta|$, places a constant penalty on deviations from zero. Taddy (2019)[10] signal some of the advantages of using Lasso as "some of the solved k values will be exactly equal to zero—not close to zero, but zero as in "they are not in the model, so you don't need to store or think about them." The penalty and the estimations of the LASSO model are the following:

$$Minimize \sum_{i=1}^{n} L(y_{i,t+h}, m(\beta X_{i,t})) + \lambda \sum_{j=1}^{p} \left|\beta_j\right| \quad (9)$$

$$\hat{\beta}^{lasso} = argmin \left\{ \sum_{i=1}^{n} (y_{i,t+h} - \beta X_{i,t})^2 + \lambda \left|\beta_j\right| \right\} \quad (10)$$

- **The Ridge Model**

The shape of the ridge penalty $\lambda \left|\beta_j\right|^2$ is different and places little penalty on small values of β but a rapidly increasing penalty on large values. "This will be appropriate for scenarios where you believe each covariate has a small effect, with no big coefficients dominating the model" (Taddy M. , 2019) .

$$Minimize \sum_{i=1}^{n} L(y_{i,t+h}, m(\beta X_{i,t})) + \lambda \sum_{j=1}^{p} \left|\beta_j\right|^2 \quad (11)$$

$$\hat{\beta}^{Ridge} = argmin \left\{ \sum_{i=1}^{n} (y_{i,t+h} - \beta X_{i,t})^2 + \lambda \left|\beta_j\right|^2 \right\} \quad (12)$$

- **The Elastic Net Model**

The Elastic Net model was introduced by Zou, H. and Hastie, T. (2005) and is the last of our three shrinkage models. This model introduces a penalization which is a combination or weighted average of the previous regularization methods Lasso and Ridge.

$$Minimize \sum_{i=1}^{n} L(y_{i,t+h}, m(\beta X_{i,t})) + \alpha\lambda \sum_{j=1}^{p} \beta^2 + \alpha\lambda(1-\alpha) \sum_{j=1}^{p} \left|\beta_j\right| \quad (13)$$

$$\hat{\beta}^{E.Net} = argmin \left\{ \sum_{i=1}^{n} (y_{i,t+h} - \beta X_{i,t})^2 + \alpha\lambda\beta^2 + \lambda(1-\alpha) + |\beta| \right\} \quad (14)$$

[10]Taddy, M. (2019). Business Data Science: Combining Machine Learning and economics to optimize, automate, and accelerate business decisions. (page 155) McGraw-Hill Education.

3.2.4 Reduction based methods PCA (Principal Component Analysis) model

Given the high number of explanatory variables I have explored some of the regularization regressions in the paragraphs above. Another possibility is to use reduction methods. The Principal Component Analysis model is a type of dimensionality reduction technique used for large datasets based on Factor Models. Basically, it summarizes a high number of variables in a reduced number of variables or components which are a linear combination of the original variables. The first principal component explains the greatest variance in the data, and the

principal components are ranked according to how much variance they account for in the data. Factor Models are helpful for displaying high-dimensional data and can do this by building a linear model for x as a function of these unknown factors. Following (Taddy M. , 2019) we can express the principal components as

$$E\lceil x_i \rceil = \varphi_1 v_{i1} + \varphi_2 v_{i2} + \ldots + \varphi_k v_{iK} \quad (15)$$

where x_i and φ_k are all length-p vectors, while the v_{iK} are uni-variate scores indicating how observation i loads on factor k. Each observation x_i is mapped to K factors v_{iK} representing a low dimensional summary of x_i. But v_{iK} factors are latent ("unknown") variables and need to be estimated. A way to estimate them is through Principal Component Analysis where the vector of variables v_{iK}:

$$v_{ki} = x_i \varphi_k = \sum_{j=1}^{p} \varphi_{kj} x_{ij} \quad (16)$$

Where vectors of variables can be expressed as a function of components $\Phi = \{\varphi_1, \varphi_2, \ldots, \varphi_k\}$. Once the principal components are estimated and decided the optimum number of principal components to be used, we can reformulate the problem as a Linear Regression Model but replacing the X set of variables by the Φ_k components as in the following representation:

$$Minimize \sum_{i=1}^{n} L(y_{i,t+h}, m(\beta \Phi_k(X_{i,t})) \quad (17)$$

$$\hat{\beta} = argmin \left\{ \sum_{i=1}^{n} (y_{i,t+h} - \beta \Phi_k(X_{i,t}))^2 \right\} \quad (18)$$

3.2.5 Non-Linear Tree Based Models: Random Forest and ADA Boost

Tree-based models are a non-parametric method which can be used for regression and for classification purposes. According to Joseph (2022) “The idea behind them is to consecutively split the training dataset until an assignment or stopping criterion with respect to the target variable into a “data bucket" or leaf is reached"[11]. “Splitting the vector of predictors z_t (predictors and lags of dependent variable) into N leaf, $Z = \left\{Z_1, Z_2, \ldots, Z_{Nleaf}\right\}$, the optimal estimates of the β “coefficients" is just the average of the training target values y within each leaf of a tree". The optimization problem and its estimator forms would be as the following

$$Minimize \sum_{m=1}^{Nleaf} L(y_{i,t+h}, I(z_t)) \qquad (19)$$

$$\hat{\beta} = argmin\{1/|Z_m| \sum_{i=1}^{n} (y_{i,t+h} - I(z_m))\} \qquad (20)$$

A disadvantage of regression trees is that they are not identically distributed: they are built adaptively to reduce the bias. This may lead to severe over-fitting. To deal with this problem some ensemble approaches are used for different models:

-**Random Forest** (Breiman, 2001)[12]: Random Forest are frequently used to overcome this problem. They use bagging as a pick with replacement (i.e bagging or bootstrap sampling) and obtain multiple observations or examples of the training sample. The assemble predictor is obtained from the expectation or the average of the different examples so in bagging, each training example is equally likely to be picked.

-**AdaBoost** (Freund and Schapire, 1995)[13]: While in bagging each training example is equally likely, in boosting, the probability of a particular example being in the training set of a particular example depends on the performance of the prior on that example. At each iteration of the process a weight w_{it} is assigned to each sample in the training set which is equal to current error on that sample. These weights can be used in the training of the weak learner (Trees). For instance, decision trees can be grown which favor the splitting of sets of samples with large weights.

[11]

[12]Breiman, L. (2001). Machine Learning, 45(1), 5–32. https://doi.org/10.1023/a:1010933404324

[13]Freund, Y., & Schapire, R. E. (1995). A desicion-theoretic generalization of on-line learning and an application to boosting. Lecture Notes in Computer Science, 23–37. https://doi.org/10.1007/3-540-59119-2_166

3.3 Methodology: Strategy on Models and Variables Evaluation

Once the models have been estimated, I proceed to evaluate the relevance of the models and the importance of the variables. The strategy for each is different. In the case of the models, we will evaluate their relevance based on the errors in the testing sample compared to the actual data. To assess the significance of the variables, I will rely on the "Shapley values", which measure the average contribution of each variable in the model to the final forecast.

3.3.1 Evaluation of the Models: Comparing Errors Out of Sample

The comparison of the forecasted values of the models with the real data in the testing samples determines the error term of any of the models at any of the horizons as defined in (21) and (22):

$$\hat{\pi}_{t+h/t} = \hat{G}_{h,t}(x_t) \qquad (21)$$

$$\hat{e}_{t,m,h} = \pi_t - \hat{\pi}_{t,m,h} \qquad (22)$$

Where $\hat{\pi}_{t,m,h}$ is the inflation forecast for month t made by the model "m "with information up to time t. The models are compared according to two different statistics, namely, the root mean squared error (RMSE) and the mean absolute error (MAE) as follows:

$$\mathrm{RMSE}_{m,h} = \sqrt{\frac{1}{T - T_0 + 1} \sum_{t=T_0}^{T} \widehat{e}_{t,m,h}^2}, \quad \mathrm{MAE}_{m,h} = \frac{1}{T - T_0 + 1} \sum_{t=T_0}^{T} \left|\widehat{e}_{t,m,h}\right|,$$

The first two measures above are the usual ones in the forecasting literature. Reporting both MAE in addition to RMSE is important for confirming that the results are not due to a few large forecasting errors.

Besides contrasting the error of the different models of inflation at different horizons I also compare all the alternatives with a benchmark model. In this case, as described in equation (23), I use a naive model. In this model, I forecast of inflation at the different horizons based

exclusively on the latest inflation monthly rate. I will compare the RMSE and MAE of the algorithms with the RMSE and MAE of the naive model:

$$\hat{\pi}_{t+h/t} = \pi_t \quad (23)$$

$$RMSE_{Naive,h} = \sqrt{\frac{1}{T-T_0-1}} \sum_{t=T_o}^{T} e^2_{t,Naive,h} \text{ and } MAE_{Naive,h} = \frac{1}{T-T_0-1} \sum_{t=T_o}^{T} \left|\hat{e}^2_{t,naive,h}\right|$$

The naive model basically says that the most likely result of future inflation (monthly) is the latest released monthly inflation rate. The meaning of using this model as a benchmark, is that any of the alternative models should be at least better than this model at any horizon. To specifically account for this, and standard in the literature, I compute the ratios of any specific models $\hat{G}_{h,t}$ relative to the Naive Model as:

$$RatioRMSE = \frac{RMSE_{m,h}}{RMSE_{Naive,h}} \text{ and } RatioMAE = \frac{MAE_{m,h}}{MAE_{Naive,h}}$$

Any values of the ratios of RMSE or MAE equal or above one stand for models which are not competitive with a simple naive model. In contrast, models with ratios below one reflect that the alternative models add some extra value in terms of forecasting accuracy as the errors are lower than the naive model. The lower the ratio the higher the accuracy of the alternative models.

3.3.2 Evaluation of Variables: Analysis of Shapley Values

Evaluating the relevance of variables is a critical aspect when working with machine learning models. According to Joseph (2019), this involves addressing questions such as, "How important is a variable for a model's performance?" or "How crucial is a variable in generating a forecast?".

While determining the importance of explanatory variables in a linear model might be straightforward, disentangling their influence in non-linear models, such as Random Forests or Neural Networks, can be challenging. A recent tool called "SHAP" (Lundberg and Lee, 2017) has gained popularity as an effective method for identifying the significance of variables in a model (see Joseph, 2019).

Shapley values, on which SHAP is based, stem from a mathematical theory that ensures a fair distribution of variable contributions. Decomposing a given prediction into Shapley values enables local interpretability, providing insights into the roles of individual variables in complex models. The method decomposes the predictions made by any algorithm into variable contributions (their "Shapley values"), thus decomposing individual predictions f(x) of observation t into the contributions of the individual features or variables:

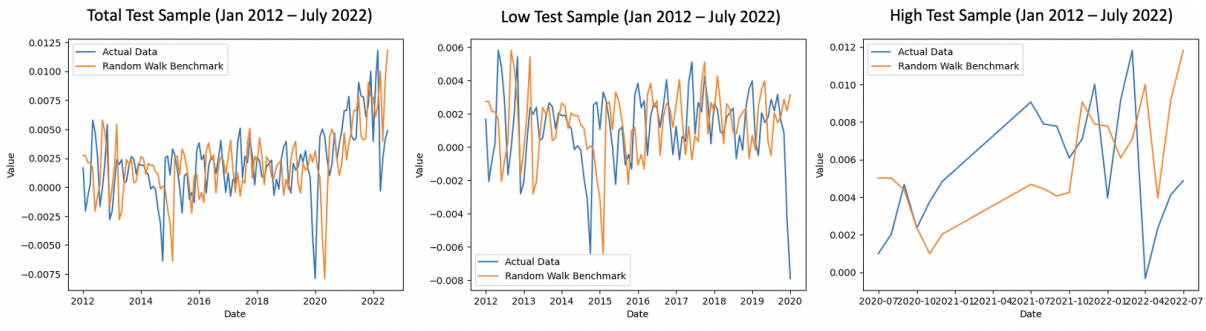
$$f(x_t) = \sum_{k=0}^{N} \phi_k(x_t) \quad (24)$$

4 Results: Accuracy of Inflation Forecasting of different Machine Learning Models

In this section, I describe the main results of the different algorithms and compare the results with the benchmark model. Starting with the simplest one, the results of the Naive model for one, three and six months can be observed in figure 2 and Table 1. As explained, the rest of the algorithms will be linked to this one as I also evaluate them in relative terms to this naive model.

Figure 2: Naive Benchmark model for the 3 different Tests samples

(Total, Low and High Inflation) in 3 months Horizon



Source: Own Elaboration

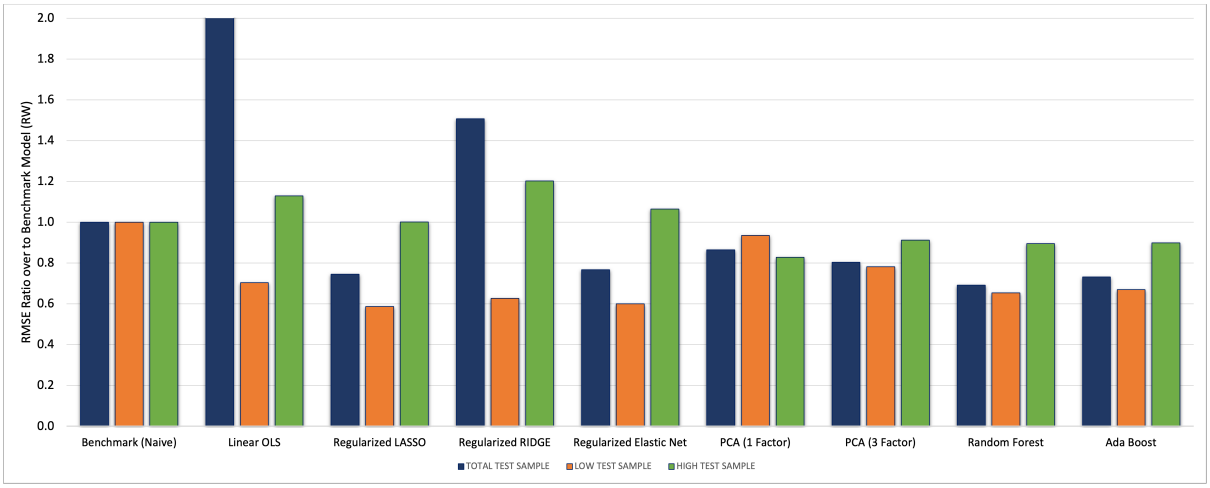
Table 2: Naive Benchmark model RMSE statistics in 1, 3 & 6 months

BENCHMARK RMSE	1 MONTH	3 MONTHS	6 MONTHS
TOTAL	0.34%	0.35%	0.35%
LOW	0.28%	0.28%	0.31%
HIGH	0.43%	0.41%	0.45%

Source:Own Elaboration

Table 2 shows that the RMSE error for the total inflation test set is relatively high and stable across the short (1 month) and medium term horizons (6 months), with a minimal increase from 0.34% to 0.35%. Remember that the average monthly inflation rate of the test sample is 0.20% so the error is relatively high. When checking the error term of the sub-samples (low and high inflation) some of the differences arise. In both regimes the error is stable but at different levels. During the low inflation period, the RMSE error fluctuates around 0.28%-0.31% but increases significantly to 0.43%-0.45% during the high inflation years. In general terms, the errors of the naive model tell us that inflation is not easy to be forecast with simple Naive models and that the high inflation period is even more difficult to forecast.This should open the door for the machine learning models for some improvement.

Figure 3: RMSE Ratio to Benchmark Model in 1 month horizon



Source: Own Elaboration

Table 3: RMSE Ratio to Benchmark Model in 1 month horizon

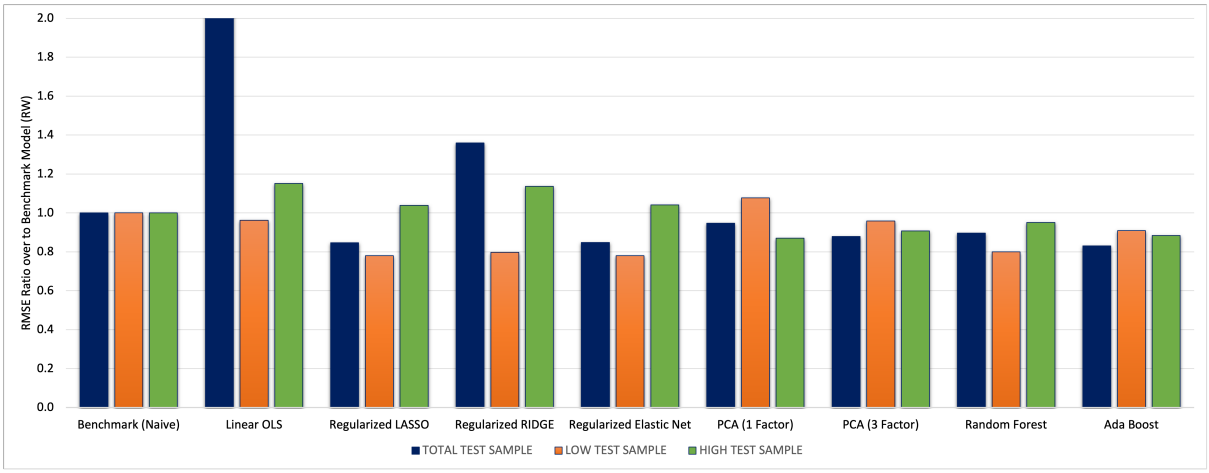
RMSE 1 MONTH	Benchmark (Naive)	Linear OLS	Regularized LASSO	Regularized RIDGE	Regularized Elastic Net	PCA (1 Factor)	PCA (3 Factor)	Random Forest	Ada Boost
TOTAL TEST SAMPLE	1	8.53	0.75	1.51	0.77	0.86	0.80	0.69	0.73
LOW TEST SAMPLE	1	0.70	0.59	0.63	0.60	0.93	0.78	0.65	0.67
HIGH TEST SAMPLE	1	1.13	1.00	1.20	1.06	0.83	0.91	0.90	0.90

Source: Own Elaboration

The results of the alternative machine learning models at the short term horizon (1 month) can be observed in Figure 3 and Table 3. The graph and the table show the ratios of the RMSE of all of the models relative to the benchmark model (the naive one). The first thing to note is that the RMSE Ratio of the total sample is significantly higher in the case of the linear model. This is mainly the result of the inclusion in the total inflation test set of the Covid-19 severe months (January 2020 to June 2020) and shows how difficult is for the traditional linear models to capture or explain the large price fluctuations (otherwise the RMSE ratio would be higher for all the models). This is even true for some of the regularized models such as Ridge, which is also affected by this Covid-19 effect. In fact, they are clearly under-performing the Naive model. Some of the Regularized modes (Lasso and Elastic Net) and specially the non-linear models as Random Forest and Ada Boots outperform the rest as the present lower RMSE errors.

The Covid-19 effect disappears when I analyze the low and high inflation periods, as the Covid-19 effect has been eliminated in this sub-samples. The results corresponding to the "low inflation period" show now that the linear models is relatively good and the regularize models are slightly better. In fact, their performance is similar or even better than the the nonlinear models. The results change when we evaluate the RMSE during the "high inflation period". During this sample, the RMSE errors of the non-linear models such as PCA, Random Forest and AdaBoost are significantly lower, showing a better performance than the benchmark as well as the linear and regularize models in the very short run (1 month).

Figure 4: RMSE Ratio to Benchmark Model in 3 months horizon



Source: Own Elaboration

Table 4: RMSE Ratio to Benchmark Model in 3 months horizon

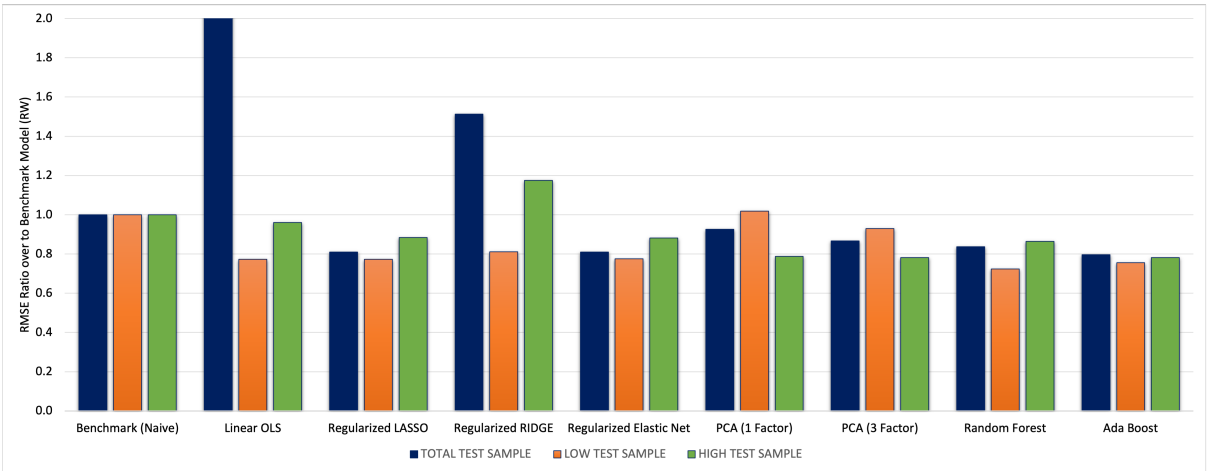
RMSE 3 MONTH	Benchmark (Naive)	Linear OLS	Regularized LASSO	Regularized RIDGE	Regularized Elastic Net	PCA (1 Factor)	PCA (3 Factor)	Random Forest	Ada Boost
TOTAL TEST SAMPLE	1	11.59	0.85	1.36	0.85	0.95	0.88	0.90	0.83
LOW TEST SAMPLE	1	0.96	0.78	0.80	0.78	1.08	0.96	0.80	0.91
HIGH TEST SAMPLE	1	1.15	1.04	1.14	1.04	0.87	0.91	0.95	0.88

Source:Own Elaboration

Figure 4 and Table 4 show the RMSE ratio to benchmark model at short-medium horizon (3 months). In general the RMSE ratios to the benchmark model are in line with the short-term horizon ones (1 month) . The linear models have many difficulties, even failing to outperform the benchmark in explaining the large variations in prices, the similar happens to the Ridge model . In contrast, these two models perform relatively well in explaining periods of "low inflation". In addition, we can observe again how the regularization models such as LASSO or Elastic Net outperform the linear model and their performance is in line with the nonlinear models during this period.

Again, the nonlinear models have a better predictive performance in the "high inflation periods", with PCA (1 factor) and AdaBoost being the best models. In sum, the results are similar than those from the very short term horizon, but the differences between the non-linear and linear models in the high inflation periods are now smaller. Although there still seems to be some advantage of using the non-linear models in the total and high inflation test samples.

Figure 5: RMSE Ratio to Benchmark Model in 6 months horizon



Source: Own Elaboration

Table 5: RMSE Ratio to Benchmark Model in 6 months horizon

RMSE 6 MONTH	Benchmark (Naive)	Linear OLS	Regularized LASSO	Regularized RIDGE	Regularized Elastic Net	PCA (1 Factor)	PCA (3 Factor)	Random Forest	Ada Boost
TOTAL TEST SAMPLE	1	37.56	0.81	1.51	0.81	0.93	0.87	0.84	0.80
LOW TEST SAMPLE	1	0.77	0.77	0.81	0.78	1.02	0.93	0.72	0.76
HIGH TEST SAMPLE	1	0.96	0.88	1.18	0.88	0.79	0.78	0.86	0.78

Source:Own Elaboration

Finally, Figure 5 and Table 5 show the comparative results for the medium term horizon (6-month horizon). As in the previous horizons, the linear and the regularized ridge model do not outperform the benchmark model in the full sample for the same reasons linked to the sharp swings in inflation due to Covid-19. However, during the "low inflation" the results show that the regularized and nonlinear models outperform the benchmark model easily being the best models in this case. However, in the "high inflation" test set, neither the linear nor the regularized models manage to outperform the benchmark model, being the nonlinear models again the outperforming ones over both the naive ann linear models.

In summary, Table 6 presents the overall results of the comparative analysis of the RMSE ratios of the models for all the horizons. The table show the number of times that any of the models outperform the Naive Model including the number of times and the % of the total. The results show that most of the models outperform the naive model in the full test sample, except for the traditional linear model. This should be taken with care as it must be taken into account

that this sample include a very complicated environment with large variations due to the Covid shock. The rest of the ML models are useful during the full test sample, as they outperform the naive model and, therefore, they are better predictors of inflation. Most of the models in the low inflation years outperform the naive (except the PCA) and behave in a similar way, revealing the gains of using ML models.

Table 6: Nº of times each model beats the Benchmark in each of the 6 horizons

Nº Times it beats the Benchmark	Linear OLS	Regularized LASSO	Regularized RIDGE	Regularized Elastic Net	PCA (1 Factor)	PCA (3 Factor)	Random Forest	Ada Boost
TOTAL TEST SAMPLE	0	6	0	6	6	6	6	6
	0%	100%	0%	100%	100%	100%	100%	100%
LOW TEST SAMPLE	6	6	6	6	2	5	6	6
	100%	100%	100%	100%	33%	83%	100%	100%
HIGH TEST SAMPLE	1	3	1	3	5	5	5	5
	17%	50%	17%	50%	83%	83%	83%	83%

Source: Own Elaboration

An important result is the superiority of the nonlinear models during the high inflation sample years. On average of the three samples, the best models as AdaBoost and Random Forest, out beat 94% (17/18) of the times the naive model. This performance is better at the total and low inflation test set when they outperform 100%. Besides, the non linear models are the only ones out beating the naive most of the times during the high inflation period. This is consistent with the findings of Medeiros et al. (2019), who find that the Random Forest model is the one with the lower forecasting errors for the US inflation.

Another important result is that there are some gains to introduce regularization on the linear model (particularly Lasso and Elastic Net). This should not be strange given the high number of variables included in our exercise.

5 Relevance of variables: The Shapley values

An important aspect of this work is the analyze the relevance of the different groups of variables in the best machine learning model for forecasting inflation. In this particular case, I examine and analyze the most important variables and groups of variables within the Random Forest model (the best model for forecasting inflation in this work). Given the Non-Linearity of the

Random Forest Model, I use the Shapley values technique to evaluate the contribution of the individula and different group of variables in the machine learning models. As described in the theoretical section the Shapley values decompose the forecast of the model in the contribution of all the variables to the forecast.

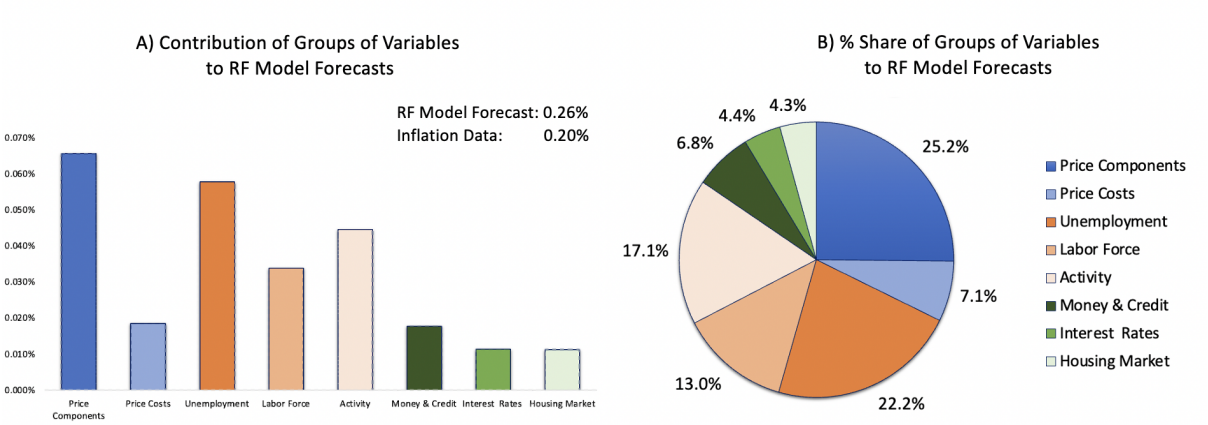
One special interest in my work is to check whether the relevance of the variables remains consistent across different time samples and, in particular, during the different inflation regime defined earlier . Therefore, I will focus in analyzing and identifying the most relevant variables and their groups during the total, the low and high inflation test sets. By doing so, I can obtain a more comprehensive understanding of the key factors driving inflation forecasting in different inflation regimes.

When calculating the importance of each group of variables in the model, the Shapley values are computed for the different test samples: the total inflation test set (January 2012 - July 2022), the low inflation test set (January 2012 - January 2020), and the high inflation test set (July 2020 - July 2022). Given that the Shapley values are additive (Joseph, 2019), after computing the individual variable contribution, I also compute the aggregated contribution of the eight different groups of variables (Graph 6A) and the share of each group to the forecast of the model (Graph 6B).[14]

The results for latest decade, the total inflation test set (2012-2022), are somehow consistent with some of the discussions in both the empirical and theoretical literature. Figure 6A and 6B shows the relevance and the contribution by group of variables to the Random forest forecast of this sample (0.26% average inflation), the forecast of this period was not bad but still overstimate the average inflation of this period (0.20%) . The group contributions show that price components or price dynamics (in blue) and the conditions of the labor market or unemployment (orange) are the key variables for inflation in this sample. The greater importance of price components (own dynamic inflation) could be revealing the relevance of inflation dynamics and expectations in the evolution of prices. Moreover, activity and labor force indicators are also relevant with the group of unemployment showing a high contribution to the forecast.

[14]In the appendix I also show the average individual contribution of the variables in Tables 16, 17 and 18 as the numbers of variables in the groups are not uniform. However, the relative importance remains mainly unchanged

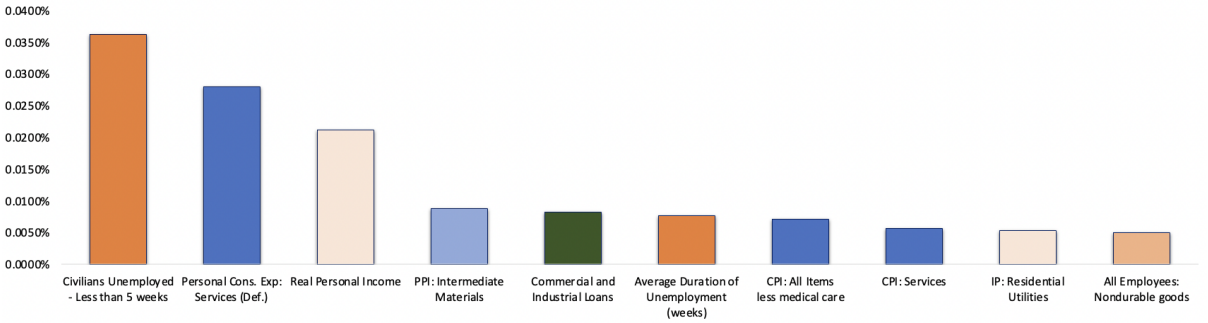
Figure 6: Contribution of each group of variables in Total Inflation Test set (January 2012 - July 2022) in Random Forest model



Source:Own Elaboration

This looks to be consistent with the traditional view that during the last decade the Phillips curve is a valid hypothesis. In the long sample, the price cost group or cost inflation look less relevant. Finally, the groups of financial variables such as money and credit, interest rates, and the housing market seem to have little effect.[15]. After the analysis of the contribution to the total explanation in 6B, the share in % of the groups it is clear that the price components (25.2%), the unenploymnet (22.2%) and activity (13.0%) and labor force (17.1%) explain more than two thirds of the model.

Figure 7: Top 10 variables with more influence in Total Inflation Test Set (January 2012 - July 2022) in RF model

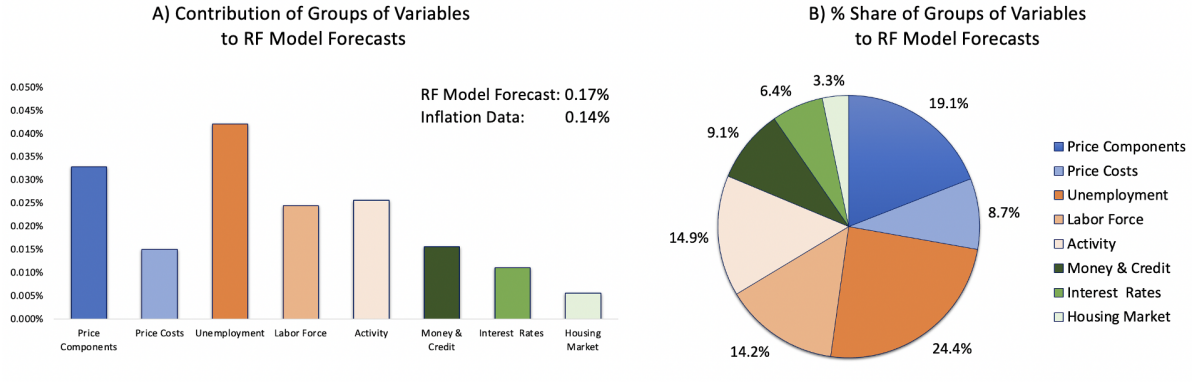


Source:Own Elaboration

[15]In general terms it is assumed that inflation and money are linked in the long run while we are evaluating this relationship in the very short run. The same could be true for interest rates with takes some times to affect inflation

Figure 7 shows the top 10 of the most individual relevant variables in the model during the total the total inflation test set (2012-2022). The graphs shows that the two most explanatory variables are short-term unemployment and the services consumption deflator. The latest reflecting that more stable component of prices, the services, are more sticky and more difficult to experience large fluctuations. This could be relevant for monetary policy because these prices are more stable one but also the ones reacting less to interest rates.

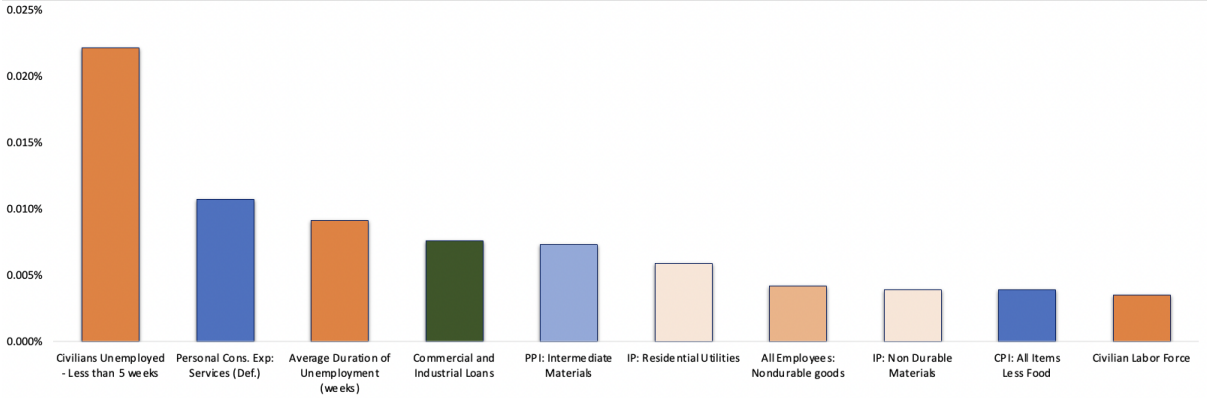
Figure 8: Contribution of each group of variables in Low Inflation Test set (January 2012 - January 2020) in Random Forest model



Source:Own Elaboration

Figure 8 shows the relevance of the variables during the "low inflation" test set (2012-2019). It shows some differences. First, the importance of the price components or own price dynamics, were not as important as the total sample . This could be the result of lower influence of expectations due to the relative stability of this period of low prices. However, unemployment (a proxy of labor market tightness and the Philips curve) become the most relevant group (third column in orange). As in the entire test sample, financial variables did not contribute much to the prediction of inflation. In sum, it looks that during this "stable inflation" period the traditional Phillip curve worked relatively well, while the role of price dynamics were less relevant. The figure 8B shows how unemployment group is now the most relevant group with a relevance of near 25% (24.4%), higher than the price components variables (19.1%).

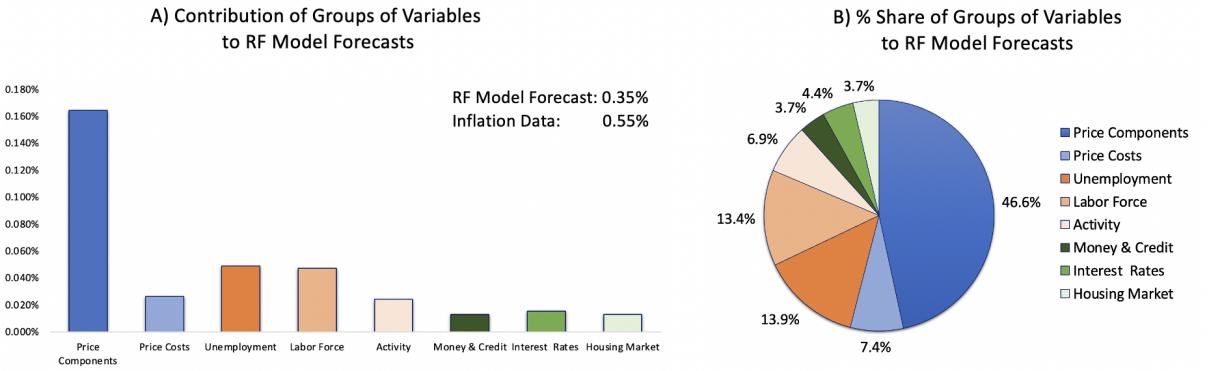
Figure 9: Top 10 variables with more influence in Low Inflation Test Set (2012 - January 2020) in RF model



Source:Own Elaboration

During this stable period, among the top 10 variables with the greatest influence were the unemployment, labor force or activity group confirming the idea of the Phillips curve and demand being specially relevant. The variable with the greatest importance is the number of unemployed people less than 5 weeks while the personal consumption expenditure (price component) diminishes

Figure 10: Contribution of each group of variables in High Inflation Test Set (July 2020 - July 2022) in RF model

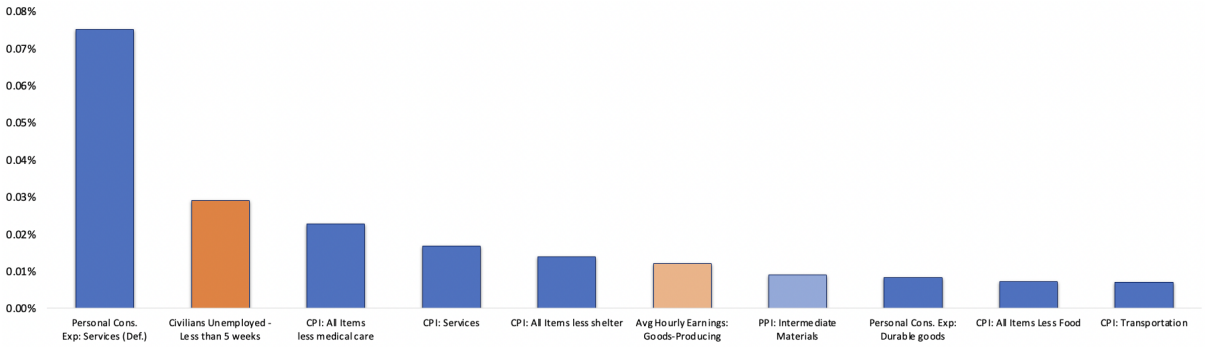


Source:Own Elaboration

Finally, Figure 10 shows the analysis of the variables during the recent "High Inflation" test set (2020-2022). The graphs show significant changes relative to rest of the periods. We can

clearly observe that the group of price components has the highest relevance in this period (most of the contribution in figure 10.A and reaching 46.6% of the forecast in graph B), and maintain a large difference with respect to the rest of the groups. This could be the result of an increasing importance of "price dynamics" after the Covid-19 which could led people and firms to internalize higher inflation leading to an effective increase. Expectations could be playing a higher role and this could lead to a higher difficulty for the monetary policy to control inflation[16]. The unemployment group, although with a much smaller contribution than previous test sets, is still very relevant in these times of high inflation. The lower relevance of unemployment is in line with some authors suggesting recently diminishing but still important role of the Phillips curve (Del Negro et al, 2020) [17]. Finally, the groups of financial variables confirm previous results showing small relevance in explaining inflation.

Figure 11: Top 10 variables with more influence in High Inflation Test Set (July 2020 - July 2022) in RF model



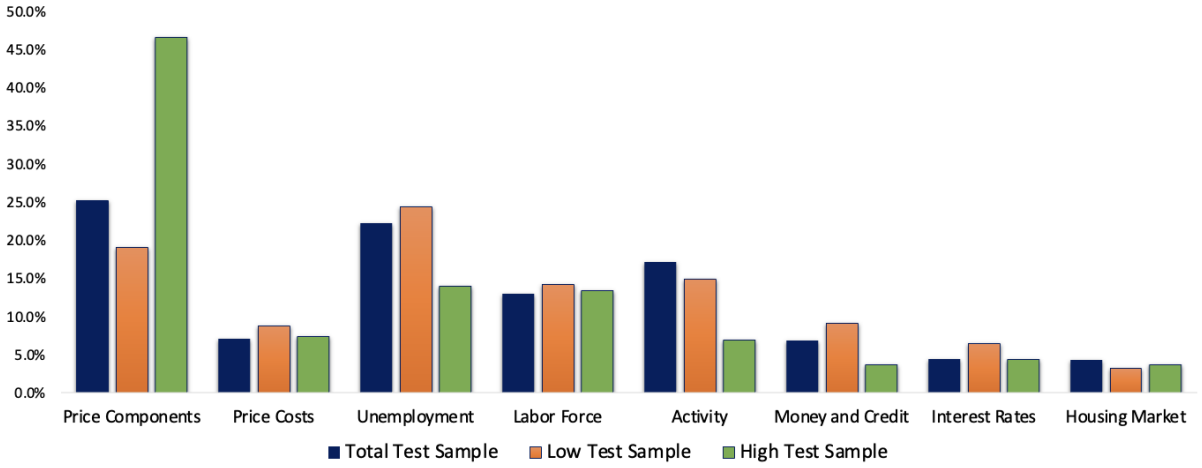
Source:Own Elaboration

Relative to individual variables, and since the group of price component variables accounts for almost 47% of the total contribution, most of the top 10 variables come from this group. As in the entire test sample, personal consumption expenditures deflator on services is the most influential variable, but this time it is the one that contributes the most relative to the rest by a large margin. This is consistent with Medeiros et al. (2019), who show this variable as one of the top four variables. Unemployment of less than 5 weeks is still relevant, but this time much lower than services inflation.

[16]See Federal Reserve Governor speech "Monetary Policy and Price Stability" Powell(2022)

[17]These authors argue that the Phillips curve has been flattening recently

Figure 12: Comparison of the "total "contribution of groups of variables to Inflation in Random Forest model between the 3 test sets



Source:Own Elaboration

The Figure 12 summarizes the relevance of the variables in our best model across the samples. The results show the importance of the price components and the labor market conditions (unemployment) followed by the activity indicators and labor force growth. In general, this results are consistent with traditional explanations of price dynamics and expectations as well as excess demand (or "Phillips Curve" models) playing a prominent role. However the analysis of "Low" and High" inflation show important differences. In the low inflation period, unemployment and the other groups somehow associated with the Phillips curve hypothesis, such as labor force and activity, have greater relevance in contributing to the prediction of inflation in the Random Forest model. In contrast, during the recent high inflation periods the price components increase relevance, reinforcing the role of price dynamics and expectations which coincides with a lower influence of the labor market conditions and activities(i.e the Phillips Curve model).

6 Limitations and further research

While the empirical approach I use in this work have many advantages, it has also limitations. First of all, while the sample is large, balanced and standard with the machine learning literature (81% of training sample vs. 19% of test sample), I use only 2 years of data to evaluate or test

the "high inflation" sample. However, I use more than 8 years to test the low inflation one. This is a problem of lack of data for the most recent period. Further research could include the analysis of previous high inflation periods in the past to confirm my results. While the database

of explanatory variables is large, some of the variables are general and specific components of the same variables as the price components. This could be affecting the relevance in the Shapley values, as there are some very general variables jointly with other very specific, thus some of them are more likely to contribute to a model than others. There are some strategies to cope with this in the future. This could include techniques to select the more relevant previously to the estimation, or using reduction theories to make synthetic variables of the groups as PCA.

To improve the thesis, a second benchmark model could be added as a naive forecast with a rolling mean or median, which is used in papers such as Medeiros et al (2019). In addition, some other models could have been included, such as XGBoost, which is an improved version of Random Forest, or Neural Networks. These models would have been good to include since nonlinear models have proved to be good models of inflation specially when inflation is high

To check the robustness of the results, it would be possible to include other countries like Spain. This should require efforts to build the Dataset, as normally the data for countries other than the USA are poorer.

7 Conclusion

This empirical work present important results. First, I show that using machine learning models can help to forecast important variables as inflation. In general terms, most of the algorithms can improve the forecasts of a naive model at different short-medium term horizons. One can therefore conclude that, it is possible to improve the forecasts of inflation in the United States thanks to machine learning. Furthermore, such methods can also be useful to forecast inflation in other countries, or for other economic variables. As a result of the advances in machine learning and big data, I believe that this models could become increasingly important in the coming years.

Secondly, the results show a superiority of the non-linear models (Random Forest and ADA Boost) in forecasting inflation in the USA. These are the two machine learning models that offer the lowest RMSE and MAE errors, and manage to beat the Naive model at all the different horizons and samples test sets. This superiority increase during the recent "High Inflation" years, maybe signaling the non-linearity of the recent inflation episode.

Third, the analysis of the relevance of the variables in the best model reveal interesting results which are part of the recent debate on the drivers of the recent inflation period. The results underscore the relevance of price dynamics, unemployment and activity indicators ("Philips Curve") as key determinants of for inflation forecasting. However, they show differences in the importance of the variables during the recent "high inflation" period. It is worth highlighting the higher importance of the price components group (price dynamics) reflecting some inflation dynamics relevance and/or the importance of expectations. In contrast, indicators associated with the Phillips Curve hypothesis (Unemployment, activity ..) appeared to be weakened in relevance.

Last, the monetary and financial variables and the housing market ones do not show an special relevance for forecast short term inflation forecast in the USA during the recent years.

References

[1] Ang, A. (2017, February). Ridge regression, a.k.a ridge = Argmin X. Ridge Regression. https://angms.science/doc/Regression/Regression$_{5R}idgeRegression.pdf$

[2] Araujo, G. S., Gaglianone, W. P. (2023). Machine learning methods for inflation forecasting in Brazil: New contenders versus Classical Models. Latin American Journal of Central Banking, 4(2), 100087. https://doi.org/10.1016/j.latcb.2023.100087

[3] Breiman, L. (2001). Machine Learning, 45(1), 5–32. https://doi.org/10.1023/a:1010933404324

[4] Chen, Y.-C., Turnovsky, S. Zivot, E. (2014). 'Forecasting inflation using commodity price aggregates', Journal of Econometrics 183, 117–134.

[5] Cybenko, G. (1989). Approximation by superpositions of a sigmoidal function. Mathematics of Control, Signals, and Systems, 2(4), 303–314. https://doi.org/10.1007/bf02551274

[6] Del Negro, Marco, Michele Lenza, Giorgio E. Primiceri, and Andrea Tambalotti. 2020. "What's up with the Phillips Curve?" Brookings Papers on Economic Activity, Spring, 301-373.

[7] Forni, M., Hallin, M., Lippi, M. Reichlin, L. (2003). 'Do financial variables help forecasting inflation and real activity in the euro area?', Journal of Monetary Economics 50, 1243–1255.

[8] Labelle and Santacreu, A., (2022). Global Supply Chain Disruptions and Inflation During the Covid-19 Pandemic (2022). Available at SSRN: https://ssrn.com/abstract=4029211

[9] Freund, Y., Schapire, R. E. (1995). A desicion-theoretical generalization of on-line learning and an application to boosting. Lecture Notes in Computer Science, 23–37. https://doi.org/10.1007/3-540-59119-2$_1$66

[10] Garcia, M. G. P., Medeiros, M. C., Vasconcelos, G. F. R. (2017). Real-time inflation forecasting with high-dimensional models: The case of Brazil. International Journal of Forecasting, 33(3), 679–693. https://doi.org/10.1016/j.ijforecast.2017.02.002

[11] Hoerl, A. E., Kennard, R. W. (1970). Ridge regression: Biased estimation for nonorthogonal problems. Technometrics, 12(1), 55–67. https://doi.org/10.1080/00401706.1970.10488634

[12] Joseph, A., Kalamara, E., Kapetanios, G., Potjagailo, G. (2021). Forecasting UK inflation bottom up. SSRN Electronic Journal. https://doi.org/10.2139/ssrn.3819286

[13] Koch, C., Noureldin, D. (2023, March 1). How we missed the recent inflation surge. IMF. https://www.imf.org/en/Publications/fandd/issues/2023/03/how-we-missed-the-recent-inflation-surge-koch-noureldin .

[14] Medeiros, M. C., Vasconcelos, G. F., Veiga, Á., Zilberman, E. (2019). Forecasting inflation in a data-rich environment: The benefits of Machine Learning Methods. Journal of Business Economic Statistics, 39(1), 98–119. https://doi.org/10.1080/07350015.2019.1637745

[15] Mullainathan, S., Spiess, J. (2017). Machine learning: An applied econometric approach. Journal of Economic Perspectives, 31(2), 87–106. https://doi.org/10.1257/jep.31.2.87

[16] Powell, J. (2022) "Monetary Policy and Price Stability. Speech at At “Reassessing Constraints on the Economy and Policy,” an economic policy symposium sponsored by the Federal Reserve Bank of Kansas City, Jackson Hole, Wyoming. https://www.federalreserve.gov/newsevents/speech/powell20220826a.htm

[17] Reis, R. (2022). The Burst of High Inflation in 2021-22: How and Why Did We Get Here?”, CEPR Press Discussion Paper No. 17514

[18] Oner, C. (2019). Inflation: Prices on the rise. IMF. https://www.imf.org/en/Publications/fandd/issues/Series/Back-to-Basics/Inflation

[19] Phillips, A. W. (1958). The Relation between Unemployment and the Rate of Change of Money Wage Rates in the United Kingdom, 1861-1957. Economica, 25(100), 283-299.

[20] Taddy, M. (2019). Business Data Science: Combining Machine Learning and economics to optimize, automate, and accelerate business decisions. McGraw-Hill Education.

[21] Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. Journal of the Royal Statistical Society: Series B (Methodological), 58(1), 267–288. https://doi.org/10.1111/j.2517-6161.1996.tb02080.x

[22] What is Fred?. Getting To Know FRED. https://fredhelp.stlouisfed.org/fred/about/about-fred/what-is-fred/

[23] Shapley, L, (1951). Notes on the n-Person Game II: The Value of an n-Person Game" (PDF). Santa Monica, Calif.: RAND Corporation.

[24] Zou, H., Hastie, T. (2005). Regularization and variable selection via the elastic net. Journal of the Royal Statistical Society Series B: Statistical Methodology, 67(2), 301–320. https://doi.org/10.1111/j.1467-9868.2005.00503.x

8 Appendixes

The dataset and the computes conde are available from: `https://github.com/AlvaroOrtiz2001/THESIS-FORECASTING-INFLATION-US`

The column *"tcode"* denotes the following data transformation for a series x:

- (1) no transformation;
- (2) Δxt;
- (3) Δ2xt;
- (4) log(xt);
- (5) Δ log(xt);
- (6) Δ2 log(xt);
- (7) Δ(xt/xt-1 $-$ 1).

The Name Variable (FRED) column gives the name of the variable according to the FRED data name. Next to it we include a short description of the variable

Table 7: Data description group 1: Price components

GROUP 1. PRICE COMPONENTS			
Nº	**Name Variable (FRED)**	**Description**	**Tcode**
1	G1_CPIAPPSL	CPI: Apparel	5
2	G1_CPITRNSL	CPI: Transportation	5
3	G1_CPIULFSL	CPI: All Items Less Food	5
4	G1_CUSR0000SA0L5	CPI: All Items less medical care	5
5	G1_CUUR0000SA0L2	CPI: All Items less shelter	5
6	G1_CUSR0000SAC	CPI: Commodities	5
7	G1_CUUR0000SAD	CPI: Durables	5
8	G1_CPIMEDSL	CPI: Medical Care	5
9	G1_DDURRG3M086SBEA	Personal Cons. Exp: Durable goods	5
10	G1_DNDGRG3M086SBEA	Personal Cons. Exp: Nondurable goods	5
11	G1_DSERRG3M086SBEA	Personal Cons. Exp: Services	5
12	G1_CUSR0000SAS	CPI: Services	5

Table 8: Data description group 2: Price components

GROUP 2. PRICE COSTS			
Nº	**Fred Name**	**Description**	**Tcode**
1	G2_OILPRICE	Oil Price	5
2	G2_WPSFD49207	PPI: Finished Goods	5
3	G2_WPSFD49502	PPI: Finished Consumer Goods	5
4	G2_WPSID61	PPI: Intermediate Materials	5
5	G2_WPSID62	PPI: Crude Materials	5
6	G2_PPICMM	PPI: Metals and metal products	5

Table 9: Data description group 3: Unemployment

GROUP 3. UNEMPLOYMENT			
Nº	**Fred Name**	**Description**	**Tcode**
1	G3_CLF16OV	Civilian Labor Force	5
2	G3_CE16OV	Civilian Unemployment	5
3	G3_UEMPLT5	Civilians Unemployed - Less than 5 weeks	5
4	G3_UEMP5TO14	Civilians Unemployed for 5-14 weeks	5
5	G3_UEMP15T26	Civilians Unemployed for 15-26 weeks	5
6	G3_UEMP15OV	Civilians Unemployed 15 weeks and over	5
7	G3_UEMP27OV	Civilians Unemployed 27 weeks and over	5
8	G3_UEMPMEAN	Average Duration of Unemployment (weeks)	2
9	G3_UNRATE	Civilians Unemployed Rate	2

Table 10: Data description group 4: Interest Rates

GROUP 4. INTEREST RATES			
Nº	**Fred Name**	**Description**	**Tcode**
1	G4_TB3SMFFM	3-Month Treasury C minus FEDFUNDS	1
2	G4_TB3MS	3-Month Treasury Bill:	2
3	G4_TB6MS	6-Month Treasury Bill:	2
4	G4_GS10	10-Year Treasury Rate	2
5	G4_BAA	Moody's Seasoned Aaa Corporate Bond Yield	2
6	G4_T1YFFM	1-Year Treasury C minus FEDFUNDS	1
7	G4_T5YFFM	5-Year Treasury C minus FEDFUNDS	1
8	G4_TB6SMFFM	6-Month Treasury C minus FEDFUNDS	1
9	G4_T10YFFM	10-Year Treasury C minus FEDFUNDS	1
10	G4_FEDFUNDS	Effective Federal Funds Rate	2
11	G4_GS1	1-Year Treasury Rate	2
12	G4_AAA	Moody's Seasoned Baa Corporate Bond Yield	2
13	G4_BAAFFM	Moody's Baa Corporate Bond Minus FEDFUNDS	1
14	G4_AAAFFM	Moody's Aaa Corporate Bond Minus FEDFUNDS	1
15	G4_STOCK	U.S. Stock Market	5

Table 11: Data description group 5: Money & Credit

GROUP 5. MONEY & CREDIT			
Nº	**Fred Name**	**Description**	**Tcode**
1	G5_BUSLOANS	Commercial and Industrial Loans	6
2	G5_M1SL	M1 Money Stock	6
3	G5_M2SL	M2 Money Stock	6
4	G5_M2REAL	Real M2 Money Stock	5
5	G5_REALLN	Real State Loans at All Commercial Banks	6
6	G5_TOTRESNS	Total Reserves of Depository Institutions	6
7	G5_NONBORRES	Reserves of Depository Institutions	7

Table 12: Data description group 6: Activity

GROUP 6. ACTIVITY			
Nº	Fred Name	Description	Tcode
1	G6_IPCONGD	IP: Consumer Goods	5
2	G6_IPDCONGD	IP: Durable Consumer Goods	5
3	G6_IPNCONGD	IP: Nondurable Consumer Goods	5
4	G6_CUMFNS	Capacity Utilization: Manufacturing	2
5	G6_INDPRO	IP Index	5
6	G6_IPDMAT	IP: Durable Materials	5
7	G6_IPBUSEQ	IP: Business Equipment	5
8	G6_IPFPNSS	IP: Final Products and Nonindustrial Supplies	5
9	G6_IPFINAL	IP: Final Products (Market group)	5
10	G6_IPMANSICS	IP: Manufacturing (SIC)	5
11	G6_IPMAT	IP: Materials	5
12	G6_IPB51222S	IP: Residential Utilities	5
13	G6_IPNMAT	IP: Non Durable Materials	5
14	G6_W875RX1	Real personal income ex transfer receipts	5
15	G6_RPI	Real Personal Income	5

Table 13: Data description group 7: Labor Force

GROUP 7. LABOR FORCE			
Nº	**Fred Name**	**Description**	**Tcode**
1	G7_USCONS	All Employees: Construction	5
2	G7_DMANEMP	All Employees: Manu	5
3	G7_USGOOD	All Employees: Goods-Producing Industries	5
4	G7_CES1021000001	All Employees: Mining and Logging: Mining	5
5	G7_PAYEMS	All Employees: Total nonfarm	5
6	G7_NDMANEMP	All Employees: Nondurable goods	5
7	G7_SRVPRD	All Employees: Service-Providing Industries	5
8	G7_USWTRADE	All Employees: Wholesale Trade	5
9	G7_USGOVT	All Employees: Government	5
10	G7_USFIRE	All Employees: Financial Activities	5
11	G7_MANEMP	All Employees: Manufacturing	5
12	G7_USTPU	All Employees: Trade, Transportation & Utilities	5
13	G7_CES2000000008	Avg Hourly Earnings: Construction	6
14	G7_CES0600000008	Avg Hourly Earnings: Goods-Producing	6
15	G7_CES3000000008	Avg Hourly Earnings: Manufacturing	6
16	G7_CES0600000007	Avg Weekly Hours: Goods-Producing	1
17	G7_AWHMAN	Avg Weekly Hours: Manufacturing	1
18	G7_AWOTMAN	Avg Weekly Overtime Hours: Manufacturing	2

Table 14: Data description group 8: Housing Market

GROUP 8. HOUSING MARKET			
Nº	**Fred Name**	**Description**	**Tcode**
1	G8_HOUSTMW	Housing Starts, Midwest	4
2	G8_HOUSTNE	Housing Starts, Northeast	4
3	G8_HOUSTS	Housing Starts, South	4
4	G8_HOUSTW	Housing Starts, West	4
5	G8_HOUST	Housing Starts: Total New Privately Owned	4
6	G8_PERMITMW	New Private Housing Permits, Midwest (SAAR)	4
7	G8_PERMITNE	New Private Housing Permits, Northeast (SAAR)	4
8	G8_PERMITS	New Private Housing Permits, South (SAAR)	4
9	G8_PERMITW	New Private Housing Permits, West (SAAR)	4
10	G8_PERMIT	New Private Housing Permits (SAAR)	4

Table 15: RMSE from Total Test (January 2012 to July 2022) of all variables

RMSE - TOTAL	1 MONTH	2 MONTHS	3 MONTHS	4 MONTHS	5 MONTHS	6 MONTHS
Benchmark (Naive)	1	1	1	1	1	1
Linear OLS	8.53	6.34	11.59	29.22	31.77	37.56
Regularized LASSO	0.75	0.80	0.85	0.88	0.82	0.81
Regularized RIDGE	1.51	1.38	1.36	1.35	1.37	1.51
Elastic Net	0.77	0.80	0.85	0.88	0.82	0.81
PCA (1)	0.86	0.92	0.95	0.99	0.94	0.93
PCA (3)	0.80	0.85	0.88	0.91	0.88	0.87
Random Forest	0.69	0.84	0.90	0.94	0.85	0.84
Ada Boost	0.73	0.85	0.83	0.90	0.81	0.80

Source: Own Elaboration

Table 16: RMSE from Low Test (January 2012 to January 2020) of all variables

RMSE- LOW	1 MONTH	2 MONTHS	3 MONTHS	4 MONTHS	5 MONTHS	6 MONTHS
Benchmark (Naive)	1	1	1	1	1	1
Linear OLS	0.70	0.77	0.96	0.95	0.85	0.77
Regularized LASSO	0.59	0.68	0.78	0.84	0.78	0.77
Regularized RIDGE	0.63	0.68	0.80	0.87	0.81	0.81
Elastic Net	0.60	0.67	0.78	0.84	0.78	0.78
PCA (1)	0.93	0.99	1.08	1.15	1.06	1.02
PCA (3)	0.78	0.85	0.96	1.03	0.96	0.93
Random Forest	0.65	0.75	0.80	0.87	0.79	0.72
Ada Boost	0.67	0.86	0.91	0.91	0.83	0.76

Source: Own Elaboration

Table 17: RMSE from High Test (July 2020 to July 2022) of all variables

RMSE -HIGH	1 MONTH	2 MONTHS	3 MONTHS	4 MONTHS	5 MONTHS	6 MONTHS
Random Walk	1	1	1	1	1	1
Linear OLS	1.13	1.38	1.15	1.04	1.11	0.96
Regularized LASSO	1.00	1.20	1.04	0.92	0.83	0.88
Regularized RIDGE	1.20	1.38	1.14	1.04	0.99	1.18
Elastic Net	1.06	1.21	1.04	0.92	0.84	0.88
PCA (1)	0.83	1.03	0.87	0.78	0.72	0.79
PCA (3)	0.91	1.09	0.91	0.81	0.73	0.78
Random Forest	0.90	1.05	0.95	0.84	0.79	0.86
Ada Boost	0.90	1.08	0.88	0.90	0.85	0.78

Source: Own Elaboration

Table 18: MAE from Total Test (January 2012 to July 2022) of all variables

MAE - TOTAL	1 MONTH	2 MONTHS	3 MONTHS	4 MONTHS	5 MONTHS	6 MONTHS
Benchmark (Naive)	1	1	1	1	1	1
Linear OLS	2.06	1.76	2.41	4.65	4.67	5.21
Regularized LASSO	0.71	0.74	0.83	0.88	0.78	0.73
Regularized RIDGE	1.56	1.42	1.52	1.43	1.41	1.60
Elastic Net	0.74	0.75	0.83	0.87	0.78	0.73
PCA (1)	0.95	0.95	1.00	1.06	0.97	0.91
PCA (3)	0.83	0.84	0.88	0.93	0.85	0.81
Random Forest	0.69	0.77	0.86	0.89	0.80	0.77
Ada Boost	0.74	0.85	0.84	0.90	0.76	0.76

Source: Own Elaboration

Table 19: MAE from Low Test (January 2012 to January 2020) of all variables

MAE - LOW	1 MONTH	2 MONTHS	3 MONTHS	4 MONTHS	5 MONTHS	6 MONTHS
Benchmark (Naive)	1	1	1	1	1	1
Linear OLS	0.72	0.79	0.98	0.98	0.87	0.71
Regularized LASSO	0.59	0.65	0.74	0.81	0.73	0.67
Regularized RIDGE	0.67	0.69	0.78	0.84	0.82	0.79
Elastic Net	0.61	0.65	0.75	0.81	0.74	0.67
PCA (1)	1.03	1.03	1.14	1.22	1.09	0.99
PCA (3)	0.82	0.86	0.95	1.04	0.92	0.84
Random Forest	0.65	0.71	0.78	0.82	0.72	0.65
Ada Boost	0.68	0.87	0.91	0.93	0.77	0.68

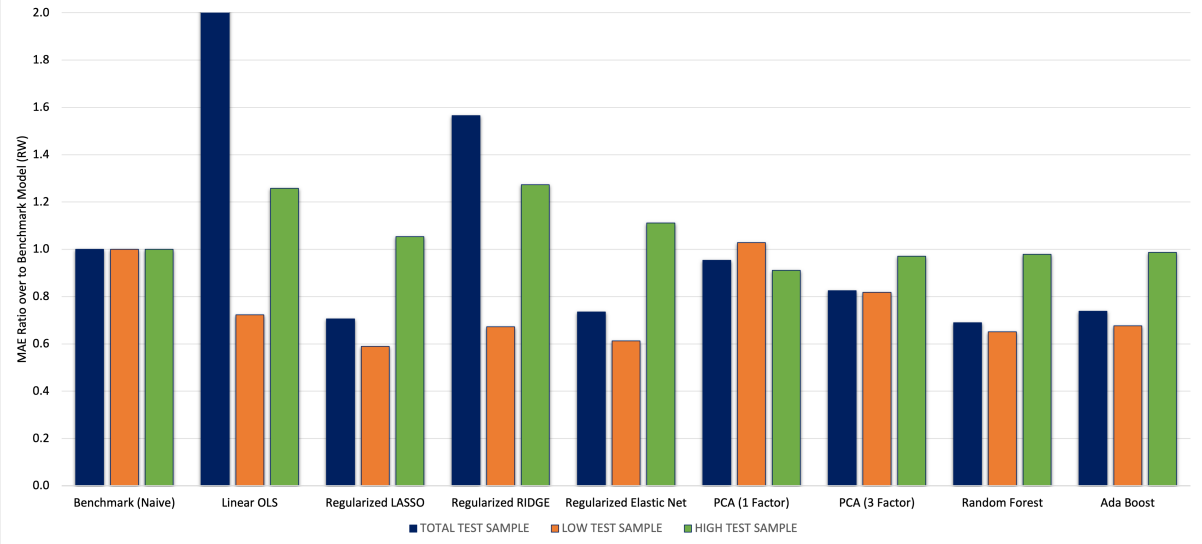
Source: Own Elaboration

Table 20: MAE from High Test (July 2020 to July 2022) of all variables

MAE -HIGH	1 MONTH	2 MONTHS	3 MONTHS	4 MONTHS	5 MONTHS	6 MONTHS
Benchmark (Naive)	1	1	1	1	1	1
Linear OLS	1.26	1.39	1.15	1.04	1.14	0.97
Regularized LASSO	1.05	1.09	1.01	0.90	0.78	0.84
Regularized RIDGE	1.27	1.28	1.14	1.00	0.94	1.19
Elastic Net	1.11	1.09	1.01	0.90	0.79	0.84
PCA (1)	0.91	0.97	0.83	0.74	0.74	0.74
PCA (3)	0.97	1.00	0.85	0.77	0.68	0.73
Random Forest	0.98	0.98	0.96	0.84	0.77	0.87
Ada Boost	0.99	0.99	0.84	0.89	0.80	0.75

Source: Own Elaboration

Figure 13: MAE Ratio to Benchmark Model in 1 month horizon



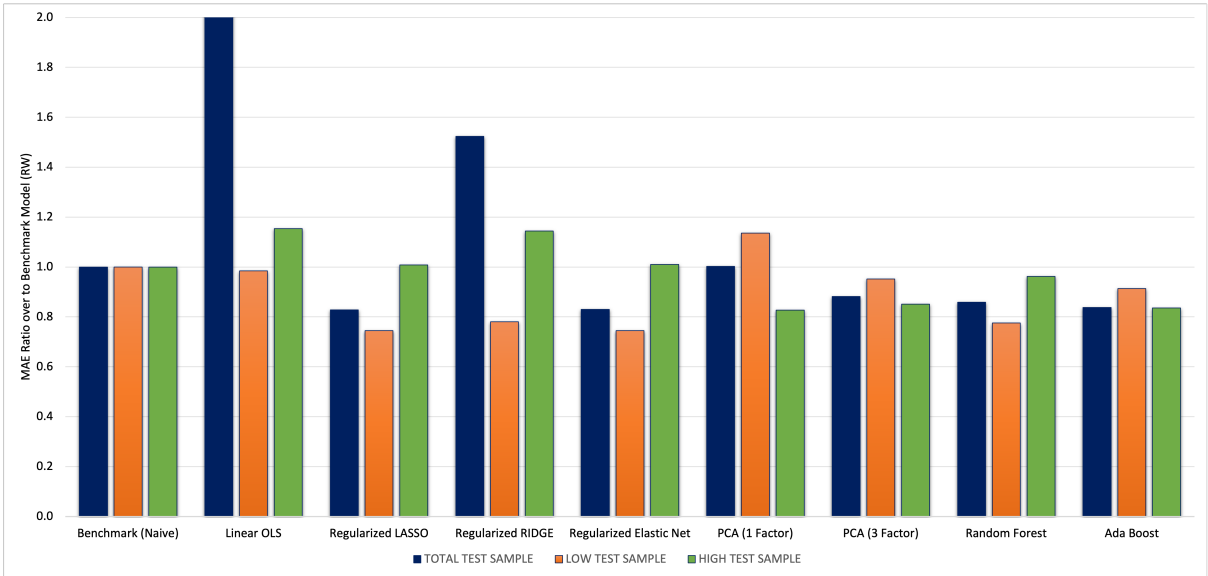
Source: Own Elaboration

Table 21: MAE Ratio to Benchmark Model in 1 month horizon

MAE 1 MONTH	Benchmark (Naive)	Linear OLS	Regularized LASSO	Regularized RIDGE	Regularized Elastic Net	PCA (1 Factor)	PCA (3 Factor)	Random Forest	Ada Boost
TOTAL TEST SAMPLE	1	2.06	0.71	1.56	0.74	0.95	0.83	0.69	0.74
LOW TEST SAMPLE	1	0.72	0.59	0.67	0.61	1.03	0.82	0.65	0.68
HIGH TEST SAMPLE	1	1.26	1.05	1.27	1.11	0.91	0.97	0.98	0.99

Source: Own Elaboration

Figure 14: MAE Ratio to Benchmark Model in 3 month horizon



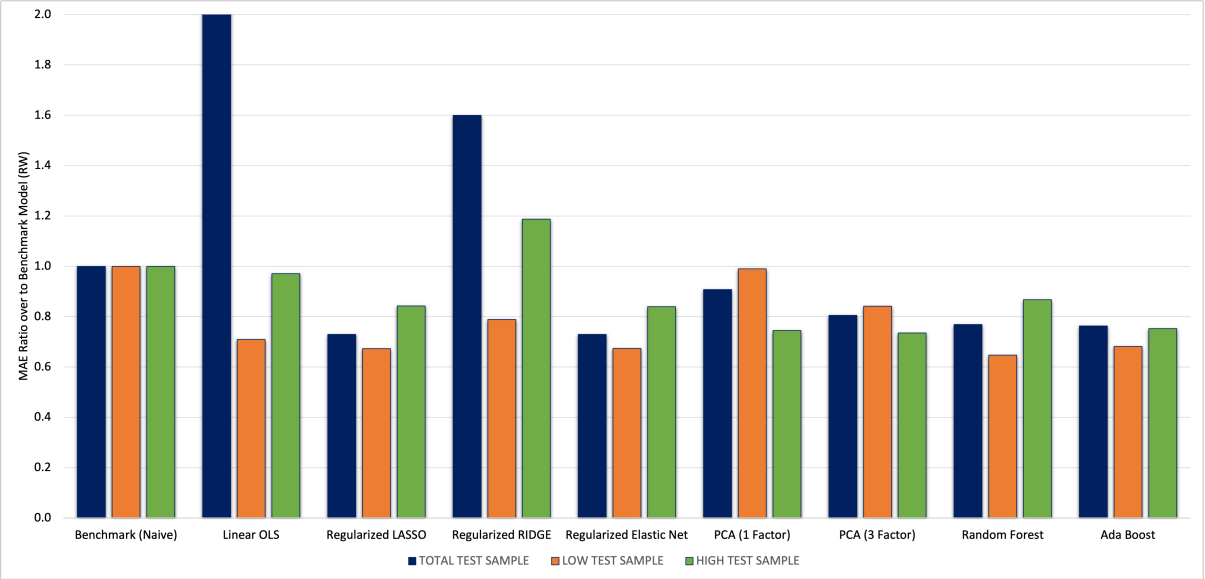
Source: Own Elaboration

Table 22: MAE Ratio to Benchmark Model in 3 month horizon

MAE 3 MONTH	Benchmark (Naive)	Linear OLS	Regularized LASSO	Regularized RIDGE	Regularized Elastic Net	PCA (1 Factor)	PCA (3 Factor)	Random Forest	Ada Boost
TOTAL TEST SAMPLE	1	2.41	0.83	1.52	0.83	1.00	0.88	0.86	0.84
LOW TEST SAMPLE	1	0.98	0.74	0.78	0.75	1.14	0.95	0.78	0.91
HIGH TEST SAMPLE	1	1.15	1.01	1.14	1.01	0.83	0.85	0.96	0.84

Source: Own Elaboration

Figure 15: MAE Ratio to Benchmark Model in 6 month horizon



Source: Own Elaboration

Table 23: MAE Ratio to Benchmark Model in 6 month horizon

MAE 6 MONTH	Benchmark (Naive)	Linear OLS	Regularized LASSO	Regularized RIDGE	Regularized Elastic Net	PCA (1 Factor)	PCA (3 Factor)	Random Forest	Ada Boost
TOTAL TEST SAMPLE	1	5.21	0.73	1.60	0.73	0.91	0.81	0.77	0.76
LOW TEST SAMPLE	1	0.71	0.67	0.79	0.67	0.99	0.84	0.65	0.68
HIGH TEST SAMPLE	1	0.97	0.84	1.19	0.84	0.74	0.73	0.87	0.75

Source: Own Elaboration

Table 24: Top 10 Variables with more contribution to Random Forest Model in Total Test (January 2012 - July 2022

TOTAL TEST SAMPLE	TOP 10 variables with more contribution	Not normalized	Normalized
G3_UEMPLT5	Civilians Unemployed - Less than 5 weeks	0.0363%	13.9%
G1_DSERRG3M086SBEA	Personal Cons. Exp: Services	0.0280%	10.7%
G6_RPI	Real Personal Income	0.0212%	8.1%
G2_WPSID61	PPI: Intermediate Materials	0.0088%	3.4%
G5_BUSLOANS	Commercial and Industrial Loans	0.0083%	3.2%
G3_UEMPMEAN	Average Duration of Unemployment (wee	0.0077%	3.0%
G1_CUSR0000SA0L5	CPI: All Items less medical care	0.0071%	2.7%
G1_CUSR0000SAS	CPI: Services	0.0057%	2.2%
G6_IPB51222S	IP: Residential Utilities	0.0053%	2.0%
G7_NDMANEMP	All Employees: Nondurable goods	0.0050%	1.9%

Source: Own Elaboration

Table 25: Top 10 Variables with more contribution to Random Forest Model in Low Test (July 2020 - January 2020)

LOW TEST SAMPLE	TOP 10 variables with more contribution	Not normalized	Normalized
G3_UEMPLT5	Civilians Unemployed - Less than 5 weeks	0.0221%	12.8%
G1_DSERRG3M086SBEA	Personal Cons. Exp: Services	0.0107%	6.2%
G3_UEMPMEAN	Average Duration of Unemployment (weeks)	0.0091%	5.3%
G5_BUSLOANS	Commercial and Industrial Loans	0.0076%	4.4%
G2_WPSID61	PPI: Intermediate Materials	0.0073%	4.2%
G6_IPB51222S	IP: Residential Utilities	0.0059%	3.4%
G7_NDMANEMP	All Employees: Nondurable goods	0.0042%	2.4%
G6_IPNMAT	IP: Non Durable Materials	0.0039%	2.3%
G1_CPIULFSL	CPI: All Items Less Food	0.0039%	2.3%
G3_CLF16OV	Civilian Labor Force	0.0035%	2.0%

Source: Own Elaboration

Table 26: Top 10 Variables with more contribution to Random Forest Model in High Test (July 2020 - July 2022)

HIGH TEST SAMPLE	TOP 10 variables with more contribution	Not normalized	Normalized
G1_DSERRG3M086SBEA	Personal Cons. Exp: Services	0.0751%	21.3%
G3_UEMPLT5	Civilians Unemployed - Less than 5 weeks	0.0290%	8.2%
G1_CUSR0000SA0L5	CPI: All Items less medical care	0.0228%	6.5%
G1_CUSR0000SAS	CPI: Services	0.0168%	4.8%
G1_CUUR0000SA0L2	CPI: All Items less shelter	0.0139%	3.9%
G7_CES0600000008	Avg Hourly Earnings: Goods-Producing	0.0121%	3.4%
G2_WPSID61	PPI: Intermediate Materials	0.0090%	2.6%
G1_DDURRG3M086SBEA	Personal Cons. Exp: Durable goods	0.0083%	2.4%
G1_CPIULFSL	CPI: All Items Less Food	0.0073%	2.1%
G1_CPITRNSL	CPI: Transportation	0.0070%	2.0%

Source: Own Elaboration

Figure 16: "Average" individual contribution of groups of variables to Inflation in RF model: Total Inflation Test set (January 2012 - July 2022)

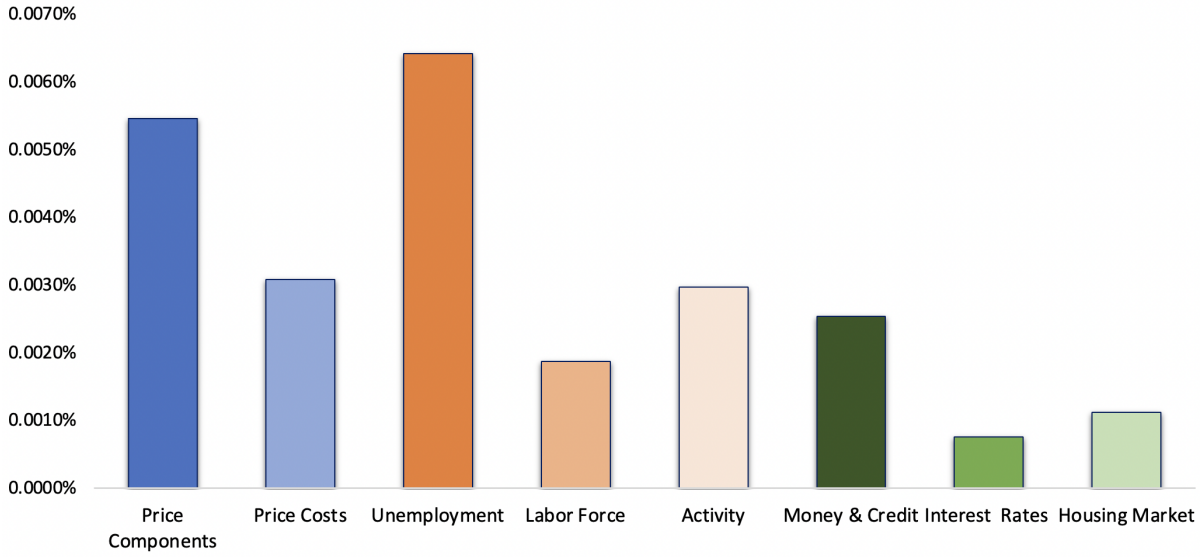


Figure 17: "Average" individual contribution of group of variables to Inflation in RF model: Low Inflation Test set (January 2012 - January 2020)

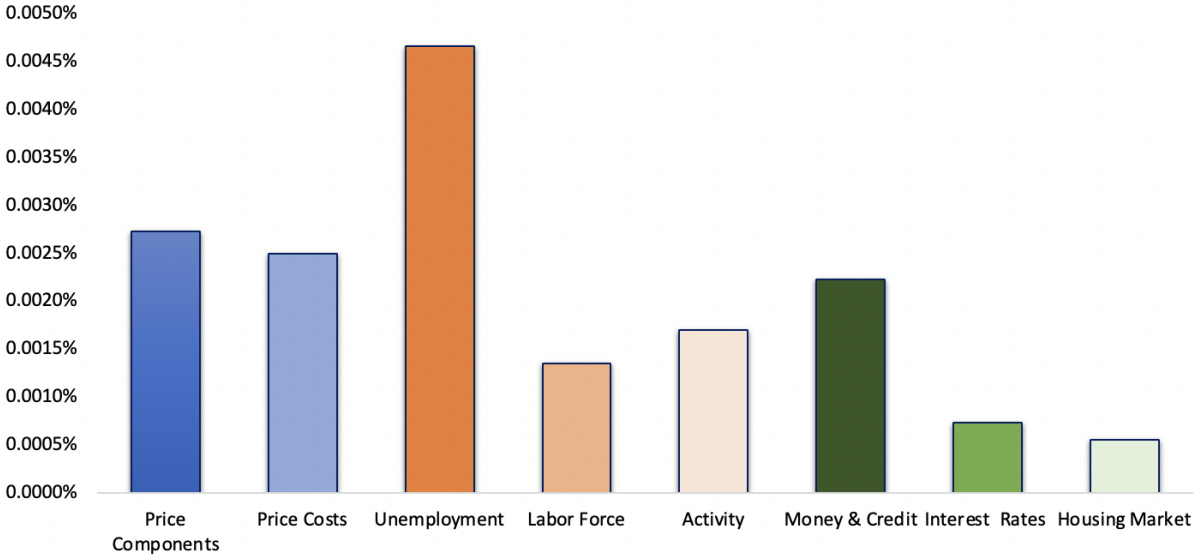


Figure 18: "Average" individual contribution of group of variables to Inflation in RF model: High Inflation Test set (July 2020 - July 2022)