FEDERAL RESERVE BANK *of* ATLANTA WORKING PAPER SERIES

The Anatomy of Out-of-Sample Forecasting Accuracy

Daniel Borup, Philippe Goulet Coulombe, David E. Rapach,
Erik Christian Montes Schütte, and Sander Schwenk-Nebbe

Working Paper 2022-16c
November 2022 (Revised September 2024)

Abstract: We introduce the performance-based Shapley value (PBSV) to measure the contributions made by each of the individual predictors in fitted time-series forecasting models to the out-of-sample loss. The PBSVs for the individual predictors sum to the out-of-sample loss, so our new metric produces an exact decomposition of out-of-sample performance. In essence, the PBSV anatomizes out-of-sample forecasting accuracy, thereby providing valuable information to decision makers for interpreting fitted time-series forecasting models. The PBSV is model agnostic–so it can be applied to any fitted prediction model, including "black box" models in machine learning–and it can be used for any loss function. We also develop the TS-Shapley VI, a version of the conventional Shapley value that gauges the importance of predictors for explaining the in-sample predictions in the entire sequence of fitted prediction models that generates the time series of out-of-sample forecasts. We then propose the model accordance score to compare predictor ranks based on the TS-Shapley-VI and PBSV, thereby linking predictors' in-sample importance to their contributions to out-of-sample forecasting accuracy. We illustrate our new metrics in an application forecasting US inflation using a variety of machine-learning models and a large number of predictors.

JEL classification: C22, C45, C52, C53, E31, E37

Key words: model interpretation, machine learning, time-series data, Shapley value, loss function, inflation

https://doi.org/10.29338/wp2022-16

The authors thank seminar and conference participants at the European Commission Joint Research Center: Online Seminar, 2022 International Symposium on Forecasting, the Workshop on Advances in Alternative Data and Machine Learning for Macroeconomics and Finance, the Federal Reserve Bank of Atlanta, the 12th European Central Bank Conference on Forecasting Techniques, and the International Association for Applied Econometrics 2023 Conference, as well as Daniele Bianchi, Giulio Caperna, Todd Clark, Marco Colagrossi, Jonas N. Eriksen, Claudia Foroni (Workshop on Advances in Alternative Data and Machine Learning discussant), Nikolay Gospodinov, Andreas Joseph, Juri Marcucci, Michael McCracken, Marcelo Medeiros, Stig Møller, Mirco Rubin, and Michel van der Wel (ECB Conference on Forecasting Techniques discussant), for insightful comments. The views expressed here are those of the authors and not necessarily those of the Federal Reserve Bank of Atlanta or the Federal Reserve System. Any remaining errors are the authors' responsibility.

Daniel Borup (dborup@econ.au.dk) is with Aarhus University. Philippe Goulet Coulombe (p.gouletcoulombe@gmail.com) is with the Université du Québec à Montréal. David E. Rapach (dave.rapach@gmail.com) is with the Federal Reserve Bank of Atlanta. Erik Christian Montes Schütte (christianms@econ.au.dk) is with Aarhus University and DFI. Sander Schwenk-Nebbe (sandersn@econ.au.dk) is with Aarhus University. Please address questions regarding content to David Rapach (corresponding author), Research Department, Federal Reserve Bank of Atlanta, 1000 Peachtree Street NE, Atlanta, GA 30309.

The Anatomy of Out-of-Sample Forecasting Accuracy

Abstract

We introduce the *performance-based Shapley value* (PBSV) to measure the contributions made by each of the individual predictors in fitted time-series forecasting models to the out-of-sample loss. The PBSVs for the individual predictors sum to the out-of-sample loss, so our new metric produces an exact decomposition of out-of-sample performance. In essence, the PBSV anatomizes out-of-sample forecasting accuracy, thereby providing valuable information to decision makers for interpreting fitted time-series forecasting models. The PBSV is model agnostic—so it can be applied to any fitted prediction model, including "black box" models in machine learning—and it can be used for any loss function. We also develop the TS-Shapley-VI, a version of the conventional Shapley value that gauges the importance of predictors for explaining the in-sample predictions in the entire sequence of fitted prediction models that generates the time series of out-of-sample forecasts. We then propose the model accordance score to compare predictor ranks based on the TS-Shapley-VI and PBSV, thereby linking predictors' in-sample importance to their contributions to out-of-sample forecasting accuracy. We illustrate our new metrics in an application forecasting US inflation using a variety of machine-learning models and a large number of predictors.

Keywords: Model interpretation, Machine learning, Time-series data, Shapley value, Loss function, Inflation

JEL classifications: C22, C45, C52, C53, E31, E37

1. Introduction

Time-series forecasting models play a fundamental role in decision making for many economic agents, such as managers, financial market participants, and policy makers. Forecasting models in macroeconomics and finance are among the most important, as they provide decision makers with insight into future general economic and financial market conditions. With the advent of "big data," the use of machine learning for out-of-sample time-series forecasting in macroeconomics and finance is burgeoning. Macroeconomic applications forecast a host of variables, such as inflation, output and employment growth, the unemployment rate, unemployment insurance initial claims, and housing starts[1]; applications in finance often involve forecasting stock returns.[2] The growing literature provides evidence that machine-learning models improve forecasting accuracy in these domains. Of course, forecasting accuracy is central to a model's usefulness. However, the ability to *interpret* fitted time-series forecasting models is also crucial for informing decision making. This is especially relevant for machine-learning models, as many are "black boxes." In particular, it is vital to understand how the predictors in fitted machine-learning models contribute to forecasting

[1] See, for example, Medeiros and Mendes (2016), Medeiros et al. (2021), Borup and Schütte (2022), Goulet Coulombe et al. (2022), Borup et al. (2023), and Hauzenberger et al. (2023).

[2] See, for example, Chinco et al. (2019), Freyberger et al. (2020), Gu et al. (2020), Dong et al. (2022), and Avramov et al. (2023).

accuracy, thereby making the black boxes more transparent by revealing the roles of the model inputs in determining time-series forecasting success. In this paper, we develop the first metric—the *performance-based Shapley value* (PBSV_p)—that provides such an understanding for fitted time-series forecasting models.

Specifically, the PBSV_p estimates the contribution of a predictor p in a sequence of fitted time-series forecasting models to a loss measure over the out-of-sample forecast evaluation period.[3] As its name suggests, we employ the logic of Shapley (1953) values to fairly allocate the marginal contributions of a model's predictors to the out-of-sample loss. By a property of Shapley values, the sum of the PBSV_p values across all of the predictors equals the out-of-sample loss measure. Thus, by computing the PBSV_p for each of the predictors, we can exactly decompose the out-of-sample loss into the components attributable to the individual predictors.

In essence, the PBSV_p allows us to anatomize forecasting accuracy in a time-series context, identifying the predictors that enhance out-of-sample performance, as well as those that detract from it. By understanding the roles of predictors in determining out-of-sample performance, economic agents can make better-informed decisions. We emphasize that the PBSV_p is very flexible: it is model agnostic—so it can be used for any fitted prediction model (parametric or nonparametric, linear or nonlinear)—and it can be applied to any loss function, including the popular mean squared error (MSE), mean absolute error (MAE), and root mean squared error (RMSE) criteria.

As machine learning has grown in popularity over the past few decades, a variety of tools have been developed for interpreting fitted prediction models, including a number that are model agnostic. One set of tools analyzes how the in-sample predictions generated by fitted models vary with the individual predictors. Such methods include partial dependence plots (Friedman, 2001), Shapley values (Shapley, 1953; Štrumbelj and Kononenko, 2010, 2014; Lundberg and Lee, 2017), individual conditional expectation curves (Goldstein et al., 2015), locally interpretable model-agnostic explanations (Ribeiro et al., 2016), and accumulated local effects (Apley and Zhu, 2020). A related set of tools measures variable importance, namely, how important individual predictors are in accounting for the predictions produced by fitted models. Variable-importance metrics include those based on partial dependence plots (Greenwell et al., 2018), permutations (Fisher et al., 2019), and Shapley values (Lundberg and Lee, 2017; Casalicchio et al., 2018).

Existing tools for interpreting fitted prediction models are typically applied in a manner appropriate for cross-sectional data. Specifically, a researcher divides the total sample of observations into training and test samples. The researcher then fits a prediction model using data from the

[3] It can also be computed for any subsample of the forecast evaluation period, including a single observation.

training sample and uses the fitted model to generate predictions for the test-sample observations. To interpret the model that generates the forecasts, the researcher computes, for example, the variable importance for each predictor based on the fitted model and training data used to estimate the model. This conventional approach is eminently reasonable, especially in a cross-sectional context.[4] However, it is not necessarily appropriate in a time-series setting. In such a setting, a researcher usually re-estimates the prediction model each period using an expanding or rolling window of data, as they generate a sequence of out-of-sample forecasts. Thus, instead of a single model, there is a sequence of estimated models to interpret. Our new PBSV_p metric explicitly accounts for a time-series setting by recognizing that the prediction model is re-estimated regularly as new data become available when generating the sequence of out-of-sample forecasts. The conventional approach also focuses on the predictors' contributions to the in-sample predictions, while the PBSV_p estimates their contributions to the out-of-sample loss—the ultimate object of interest for assessing forecasting accuracy.

We develop two additional metrics that, in conjunction with the PBSV_p, link the predictors' in-sample importance in fitted models to their contributions to out-of-sample forecasting accuracy. First, we introduce the TS-Shapley-VI_p, an extension of the conventional in-sample Shapley-based variable-importance measure that aggregates predictor p's in-sample variable importance across the entire set of fitted models that generates the sequence of out-of-sample time-series forecasts.

Second, we define the *model accordance score* (MAS) to assess the extent to which the in-sample importance of predictors in a sequence of fitted forecasting models aligns with the predictors' contributions to out-of-sample forecasting accuracy. Specifically, in the spirit of the Spearman rank correlation, we compare the ranks of the predictors in terms of their in-sample importance based on the TS-Shapley-VI_p and their contributions to out-of-sample forecasting accuracy based on the PBSV_p. A relatively high MAS indicates that the predictors that are the most important for generating the in-sample fitted values in a sequence of time-series forecasting models are also the most responsible for improving out-of-sample forecasting accuracy. As the MAS declines, there are greater discrepancies between the in-sample importance of predictors and their contributions to out-of-sample accuracy. While a performance metric like the RMSE focuses solely on out-of-sample performance, the MAS evaluates whether a model's out-of-sample success mirrors what it has learned from the in-sample data. In this sense, the MAS paired with a performance metric such

[4] For example, this approach is used on numerous occasions for the applications in the insightful textbook by Molnar (2023).

as the RMSE provides insight into the model's intentional success. We also develop a procedure for testing the statistical significance of the MAS.

Philosophically, model interpretation tools can be either *true to the model* or *true to the data* (Chen et al., 2020). The former means that we are interested in interpreting the particular fitted prediction model (or sequence of fitted models) that generates the out-of-sample forecasts. This is precisely what interests us in the present paper, so we are true to the model in constructing our metrics. We discuss remaining true to the model versus true to the data in more detail in Section 2.

We illustrate the use of our new metrics in an empirical application forecasting US inflation. A spate of recent studies finds that large datasets in conjunction with nonlinear machine-learning models, including random forests and neural networks, significantly improve inflation forecasts (e.g., Medeiros et al., 2021; Goulet Coulombe, 2022; Goulet Coulombe et al., 2022; Hauzenberger et al., 2023). We generate inflation forecasts using a set of approximately 120 predictors, primarily from the FRED-MD database (McCracken and Ng, 2016), and a variety of leading machine-learning methods, including principal component regression (Stock and Watson, 2002a,b), elastic net (Zou and Hastie, 2005) estimation of a linear model, random forests (Breiman, 2001), XGBoost (Chen and Guestrin, 2016), and neural networks. We also consider ensembles of individual forecasts generated by different models. The forecasting models consistently outperform a standard autoregressive (AR) benchmark model in terms of RMSE at horizons ranging from one to twelve months, in line with the recent literature.

We employ our new PBSV_p to measure how the predictors contribute to the accuracy of the out-of-sample inflation forecasts. At shorter horizons, the PBSV_p identifies the price of oil as a leading predictor for improving forecasting accuracy across different models, in line with the relevance of commodity price fluctuations for short-term inflation. Across all reported horizons and a variety of models, the PBSV_p points to the durables component of the CPI, the medical services component of the CPI, and the spread between the Baa-rated corporate bond yield and the federal funds rate as important predictors for improving the accuracy of out-of-sample inflation forecasts.

The MAS values reflect the degree of agreement in terms of predictor ranks between the TS-Shapley-VI_p and PBSV_p for the different forecasting models. For some models, we find a relatively low RMSE combined with a relatively low MAS, suggesting that luck played a significant role in the model's out-of-sample success. For other models, a low RMSE coincides with a high MAS, indicating that the sequence of fitted models learned from the in-sample data in a manner that reliably improves out-of-sample forecasting accuracy. In this regard, the random forest and ensemble forecasts generally perform the best.

The rest of the paper is organized as follows. Section 2 derives the PBSV_p, TS-Shapley-VI_p, and MAS metrics for analyzing predictor relevance in a time-series context. Section 3 presents the empirical application forecasting US inflation. Section 4 concludes. We created the Python package anatomy to implement the algorithms for computing our new metrics.

2. Methodology

We use the following notation in our time-series context. We index individual predictors by p and collect the predictors in the index set $S = \{1, \ldots, P\}$. The period-t P-dimensional vector of predictor observations is denoted by $\boldsymbol{x}_t = [\begin{array}{ccc} x_{1,t} & \cdots & x_{P,t} \end{array}]'$. The prediction model is given by

$$y_{t+1:t+h} = f(\boldsymbol{x}_t) + \varepsilon_{t+1:t+h}, \tag{1}$$

where $y_{t+1:t+h} = (1/h)\sum_{k=1}^{h} y_{t+k}$ is the target, h is the horizon for the forecast, f is the conditional mean (i.e., prediction) function, and $\varepsilon_{t+1:t+h}$ is an additive, zero-mean disturbance term. We denote the fitted prediction model by $\hat{f}$, while $W_i = \{t_{i,\text{start}}, \ldots, t_{i,\text{end}} - (h-1) - 1\}$ denotes the set of observations used to train the model in Equation (1) based on window W_i. The fitted prediction model evaluated at instance $\boldsymbol{x}_t$ and trained using W_i for horizon h is denoted by $\hat{f}(\boldsymbol{x}_t\,; W_i, h)$. As in Štrumbelj and Kononenko (2014), the only assumption we need is that the fitted model maps the predictors from a known input space to a known codomain.

2.1. Shapley Values in a Time-Series Context

Shapley values draw on coalitional game theory to utilize the analogy between the predictors in a model and players in a cooperative game earning payoffs, where an individual predictor's payoff corresponds to its contribution to the model's prediction. In a time-series setting, the aim of a Shapley value is to quantify the marginal contribution of predictor $x_{p,t}$ to the prediction $\hat{f}(\boldsymbol{x}_t\,; W_i, h)$, given the presence of all of the other predictors $(S \setminus \{p\})$. For now, we assume that the predictors in S are independent; we subsequently explain how we can relax this assumption. Allocating the contributions of the predictors to the prediction is far from trivial, especially when the predictors interact and there are complex nonlinearities in the fitted model. Viewed through the lens of coalitional game theory, Shapley values provide a means for fairly allocating the contributions of the predictors to a prediction for any fitted model.

Adapting Štrumbelj and Kononenko (2014) to our time-series context, the Shapley value for predictor p and instance $\boldsymbol{x}_t$ for a model trained using window W_i for horizon h is given by

$$\phi_p(\boldsymbol{x}_t\,; W_i, h) = \sum_{Q \subseteq S \setminus \{p\}} \frac{|Q|!(P-|Q|-1)!}{P!} \left[\xi_{Q \cup \{p\}}(\boldsymbol{x}_t\,; W_i, h) - \xi_Q(\boldsymbol{x}_t\,; W_i, h)\right] \tag{2}$$

for $p \in S$ and $t \in W_i$, where Q is a subset of predictors (i.e., a coalition), $Q \subseteq S \setminus \{p\}$ is the set of all possible coalitions of $P-1$ predictors in S that exclude predictor p, $|Q|$ is the cardinality of Q, $|Q|!(P-|Q|-1)!/P!$ is a combinatorial weight,

$$\xi_Q(\boldsymbol{x}_t\,; W_i, h) = \mathbb{E}\Big[\hat{f} \,\Big|\, X_{j,t} = x_{j,t} \,\forall\, j \in Q\,; W_i, h\Big] \tag{3}$$

is the value function, and $\mathbb{E}$ is the expectation operator. Equation (3) is the prediction of the fitted model conditional on the predictors in coalition Q, so $\xi_{Q \cup \{p\}}(\boldsymbol{x}_t\,; W_i, h) - \xi_Q(\boldsymbol{x}_t\,; W_i, h)$ in Equation (2) measures the change in the prediction, conditional on the predictors in coalition Q, when the predictor p is included in the conditioning information set. The difference $\xi_{Q \cup \{p\}}(\boldsymbol{x}_t\,; W_i, h) - \xi_Q(\boldsymbol{x}_t\,; W_i, h)$ is computed for all possible coalitions of $P-1$ predictors that exclude predictor p, with each quantity receiving the weight $|Q|!(P-|Q|-1)!/P!$ in the summation in Equation (2) (the weights sum to one). In essence, the Shapley value uses coalitions to control for the other predictors when measuring the contribution of predictor p to the prediction corresponding to instance $\boldsymbol{x}_t$.

The Shapley value in Equation (2) has a number of attractive properties, including the following (which we express in terms of our time-series setting).

- *Efficiency* (also known as *local accuracy*):

$$\sum_{p \in S} \phi_p(\boldsymbol{x}_t\,; W_i, h) = \hat{f}(\boldsymbol{x}_t\,; W_i, h) - \mathbb{E}\Big[\hat{f}\,; W_i, h\Big], \tag{4}$$

  where $\mathbb{E}[\hat{f}\,; W_i, h]$ is the baseline prediction, which corresponds to the unconditional expectation of $\hat{f}$ (i.e., the prediction based on the empty coalition set).

- *Missingness*:

$$\forall\, R \subseteq S \setminus \{p\} : \xi_{R \cup \{p\}}(\boldsymbol{x}_t\,; W_i, h) = \xi_R(\boldsymbol{x}_t\,; W_i, h) \Rightarrow \phi_p(\boldsymbol{x}_t\,; W_i, h) = 0. \tag{5}$$

- *Symmetry*:

$$\forall\, R \subseteq S \setminus \{p,q\} : \xi_{R\,\cup\,\{p\}}(\boldsymbol{x}_t\,; W_i, h) = \xi_{R\,\cup\,\{q\}}(\boldsymbol{x}_t\,; W_i, h) \Rightarrow \\ \phi_p(\boldsymbol{x}_t\,; W_i, h) = \phi_q(\boldsymbol{x}_t\,; W_i, h). \tag{6}$$

- *Linearity*: For any real numbers c_1 and c_2 and models $\hat{f}(\boldsymbol{x}_t\,; W_i, h)$ and $\hat{f}'(\boldsymbol{x}_t\,; W_i, h)$,

$$\phi_p\Big(c_1\Big[\hat{f}(\boldsymbol{x}_t\,; W_i, h) + c_2\hat{f}'(\boldsymbol{x}_t\,; W_i, h)\Big]\Big) = \\ c_1\phi_p\Big(\hat{f}(\boldsymbol{x}_t\,; W_i, h)\Big) + c_1c_2\phi_p\Big(\hat{f}'(\boldsymbol{x}_t\,; W_i, h)\Big). \tag{7}$$

Efficiency says that we can exactly decompose the fitted model prediction corresponding to instance $\boldsymbol{x}_t$ (in terms of the deviation from the baseline prediction) into the sum of the Shapley values for the individual predictors for that instance. Missingness and symmetry are intuitively appealing, while linearity is useful for computing Shapley values for ensembles of prediction models.[5]

In general, it is practically infeasible to compute the exact Shapley value in Equation (2) for even a moderate number of predictors, as the prediction function has to be evaluated for all possible coalitions both with and without predictor p. We use a modified version of the algorithm in Štrumbelj and Kononenko (2014) to estimate the Shapley value. To derive the algorithm, we first express Equation (2) in the equivalent form:

$$\phi_p(\boldsymbol{x}_t\,; W_i, h) = \frac{1}{P!}\sum_{\mathcal{O}\,\in\,\pi(P)}\left[\xi_{\mathrm{Pre}_p(\mathcal{O})\,\cup\,\{p\}}(\boldsymbol{x}_t\,; W_i, h) - \xi_{\mathrm{Pre}_p(\mathcal{O})}(\boldsymbol{x}_t\,; W_i, h)\right] \tag{8}$$

for $p \in S$ and $t \in W_i$, where $\mathcal{O}$ is an ordered permutation for the predictor indices in S, $\pi(P)$ is the set of all ordered permutations for S, and $\mathrm{Pre}_p(\mathcal{O})$ is the set of indices that precede p in $\mathcal{O}$. The algorithm is based on making a random draw m with replacement for an ordered permutation from $\pi(P)$, which we denote by $\mathcal{O}_m$. Using $\mathcal{O}_m$, we compute

$$\theta_{p,m}(\boldsymbol{x}_t\,; W_i, h) = \frac{1}{|W_i|}\sum_{s\,\in\,W_i}\Big[\hat{f}(\boldsymbol{x}_{j,t} : j \in \mathrm{Pre}_p(\mathcal{O}_m) \cup \{p\},\, \boldsymbol{x}_{k,s} : k \in \mathrm{Post}_p(\mathcal{O}_m)\,; W_i, h) - \\ \hat{f}(\boldsymbol{x}_{j,t} : j \in \mathrm{Pre}_p(\mathcal{O}_m),\, \boldsymbol{x}_{k,s} : k \in \mathrm{Post}_p(\mathcal{O}_m) \cup \{p\}\,; W_i, h)\Big] \tag{9}$$

for $p \in S$ and $t \in W_i$, where $\mathrm{Post}_p(\mathcal{O})$ is the set of indices that follow p in $\mathcal{O}$. Equation (9) approximates the effect of removing predictors not in the coalition by replacing them with background data

[5] As subsequently explained, for missingness in Equation (5) to hold, the Shapley value needs to be computed in a manner that is true to the data.

from the training sample. Background data refer to the data used to integrate out the predictors not in the coalition when estimating the conditional expectation in Equation (3).

Using $\theta_{p,m}(\boldsymbol{x}_t\,; W_i, h)$ in Equation (9), the estimate of $\phi_p(\boldsymbol{x}_t\,; W_i, h)$ in Equation (8) is given by

$$\hat{\phi}_p(\boldsymbol{x}_t\,; W_i, h) = \frac{1}{2M}\sum_{m=1}^{2M} \theta_{p,m}(\boldsymbol{x}_t\,; W_i, h) \tag{10}$$

for $p \in S$ and $t \in W_i$, where M is the number of random draws. To increase computational efficiency, we follow Castro et al. (2009) and compute Shapley values for each predictor $p \in S$ for a randomly drawn ordered permutation from $\pi(P)$. In addition, we implement antithetic sampling as a variance-reduction technique by computing $\theta_{p,m}(\boldsymbol{x}_t\,; W_i, h)$ in Equation (9) for the original order of a randomly drawn ordered permutation, as well as when the order is reversed (Mitchell et al., 2022). Based on arguments in Štrumbelj and Kononenko (2014), $\hat{\phi}_p(\boldsymbol{x}_t\,; W_i, h)$ in Equation (10) provides an unbiased and consistent estimate of $\phi_p(\boldsymbol{x}_t\,; W_i, h)$ in Equation (8). Equation (10) retains the attractive properties in Equations (4) to (7), including efficiency:

$$\sum_{p \in S} \hat{\phi}_p(\boldsymbol{x}_t\,; W_i, h) = \hat{f}(\boldsymbol{x}_t\,; W_i, h) - \underbrace{\bar{\hat{f}}(W_i, h)}_{\hat{\phi}_\emptyset(W_i,h)}, \tag{11}$$

where $\bar{\hat{f}}(W_i, h) = (1/|W_i|)\sum_{t \in W_i} \hat{f}(\boldsymbol{x}_t\,; W_i, h)$ is the average in-sample prediction for the model trained using sample W_i, which corresponds to the baseline or unconditional forecast (i.e., the forecast based on the empty coalition set, which we denote by $\hat{\phi}_\emptyset(W_i, h)$).

We now relax the assumption that the predictors in S are independent. Equation (9) effectively samples from the marginal distribution based on the training sample for the predictors not in the coalition. This corresponds to the *interventional* Shapley value, which coincides with remaining true to the model in Chen et al. (2020). Alternatively, we could sample from the conditional distribution for the predictors not in the coalition. This corresponds to the *observational* Shapley value, which equates with remaining true to the data in Chen et al. (2020).

At first glance, it may seem inappropriate to sample from the marginal instead of the conditional distribution when the predictors are dependent. However, when the predictors are dependent, Janzing et al. (2020) use insights from Pearl (2009) to argue that, to fairly allocate the contributions across the individual predictors, it is more appropriate to use the interventional in lieu of the observational Shapley value via the marginal distribution and thus remain true to the model.[6]

[6] Janzing et al. (2020) point out that sampling from the marginal distribution effectively implements the do-operator in Pearl (2009).

Along this line, Janzing et al. (2020) and Sundararajan and Najmi (2020) observe that, unlike the interventional Shapley value, the observational Shapley value can attribute importance to irrelevant predictors, so the missingness property in Equation (5) does not hold. Whether to remain true to the model or true to the data is ultimately a philosophical question that depends on the context of the problem being analyzed (Chen et al., 2020). In light of the above considerations and in our context—where we are interested in interpreting the sequence of fitted models that generates the out-of-sample forecasts—all of our Shapley-based metrics are interventional and thus true to the model.

The Shapley value $\hat{\phi}_p(\boldsymbol{x}_t\,; W_i, h)$ provides a local measure of the contribution of predictor p to the prediction corresponding to instance $\boldsymbol{x}_t$ in the training sample. A global measure of the importance of predictor p for the training sample can be computed by taking the average of the absolute values of the Shapley values for predictor p across the training-sample observations:

$$\text{Shapley-VI}_p(W_i, h) = \frac{1}{|W_i|} \sum_{t \in W_i} \left| \hat{\phi}_p(\boldsymbol{x}_t\,; W_i, h) \right| \tag{12}$$

for $p \in S$. The variable-importance measure in Equation (12) is a popular metric for assessing predictor importance in machine-learning applications (e.g., Molnar, 2023, Chapter 9.6). Equation (12) is based on a single training sample. Tools for interpreting fitted models are usually applied in this manner, which is appropriate for cross-sectional data (or time-series data if the prediction model is only estimated once). In a time-series context, however, researchers often re-estimate the prediction model on a regular basis over time as additional data become available, so there are multiple training samples. Next, we develop a variable-importance metric more suited to this practice.

Suppose that we are forecasting a monthly variable at horizon h and that we re-estimate the prediction model each month as additional data become available. This is typically done using either an expanding or rolling window, where the estimation sample becomes longer (remains the same size) for the former (latter). Assume that there are $t = 1, \ldots, T$ total observations available. The initial in-sample period ends in $t = T_{\text{in}}$, while the remaining $T - T_{\text{in}} = D$ observations constitute the out-of-sample period.

Mimicking the situation of a forecaster in real time, we proceed as follows. We first use observations from $t = 1$ through $t = T_{\text{in}} - (h-1) - 1$ to fit the prediction model in Equation (1) and generate an out-of-sample forecast of $y_{T_{\text{in}}+1:T_{\text{in}}+h}$. For an expanding (rolling) window, we then use observations from $t = 1$ ($t = 2$) through $T_{\text{in}} - (h-1)$ to fit Equation (1) and generate a forecast of $y_{T_{\text{in}}+2:T_{\text{in}}+h+1}$. Continuing in this manner, we generate a sequence of $D-(h-1)$ out-of-sample forecasts, where, for the final forecast, we use observations from the first period (period $T-D-(h-1)$) through $T-2h$ for an expanding (rolling) window to fit Equation (1) and generate a forecast of $y_{T-(h-1):T}$. Note that we only use data available at the time of forecast formation to train the model so that there is no "look-ahead" bias in the out-of-sample forecasts. We denote the sequence of time-series forecasts by $\hat{y}_{T_{\text{in}}+1:T_{\text{in}}+h}$, $\hat{y}_{T_{\text{in}}+2:T_{\text{in}}+h+1}, \ldots, \hat{y}_{T-(h-1):T}$.

The Shapley-based variable importance in Equation (12) corresponds to a prediction model trained once using the observations in W_i. To accommodate the sequence of $D-(h-1)$ time-series forecasts for models regularly retrained with an expanding or rolling window, we denote the set of training samples by $W = \{W_1, \ldots, W_{D-(h-1)}\}$. In this context, we define the *time-series Shapley-based variable importance* as

$$\text{TS-Shapley-VI}_p(W, h) = \frac{1}{|W|} \sum_{i \in W} \text{Shapley-VI}_p(W_i, h) \tag{13}$$

for $p \in S$, which is the average of the variable-importance measures for predictor p across all of the training samples used to generate the sequence of time-series forecasts.

2.2. Performance-Based Shapley Values

Out-of-sample forecasts are typically assessed using a loss function. Accordingly, we propose the PBSV_p to decompose the loss over the out-of-sample period into the components attributable to the individual predictors $p \in S$. We begin by defining the Shapley value for the fitted model and the vector of predictors used to generate an out-of-sample forecast, which corresponds to an out-of-sample version of Equation (8):

$$\begin{aligned}&\phi_p^{\text{out}}\big(\boldsymbol{x}_{T_{\text{in}}+(i-1)}; W_i, h\big) = \\ &\quad \frac{1}{P!} \sum_{\mathcal{O} \in \pi(P)} \big[\xi_{\text{Pre}_p(\mathcal{O}) \cup \{p\}}\big(\boldsymbol{x}_{T_{\text{in}}+(i-1)}; W_i, h\big) - \xi_{\text{Pre}_p(\mathcal{O})}\big(\boldsymbol{x}_{T_{\text{in}}+(i-1)}; W_i, h\big)\big]\end{aligned} \tag{14}$$

for $p \in S$ and $i = 1, \ldots, D-(h-1)$, where $\boldsymbol{x}_{T_{\text{in}}+(i-1)}$ is the vector of predictors plugged into the fitted prediction model that is trained with W_i and used to generate the ith out-of-sample forecast, which is given by $\hat{y}_{T_{\text{in}}+i:T_{\text{in}}+h+(i-1)} = \hat{f}\big(\boldsymbol{x}_{T_{\text{in}}+(i-1)}; W_i, h\big)$.

To estimate Equation (14), we use a suitably modified version of the algorithm in Section 2.1. For a random draw m of an ordered permutation, we modify Equation (9) to

$$\begin{aligned} &\theta^{\text{out}}_{p,m}\big(\boldsymbol{x}_{T_{\text{in}}+(i-1)}\,;W_i,h\big) = \\ &\quad \frac{1}{|W_i|}\sum_{s\,\in\,W_i}\Big[\hat{f}\big(\boldsymbol{x}_{j,T_{\text{in}}+(i-1)} : j\in \text{Pre}_p(\mathcal{O}_m)\cup\{p\},\, \boldsymbol{x}_{k,s} : k\in\text{Post}_p(\mathcal{O}_m)\,;W_i,h\big) - \\ &\qquad\qquad \hat{f}\big(\boldsymbol{x}_{j,T_{\text{in}}+(i-1)} : j\in \text{Pre}_p(\mathcal{O}_m),\, \boldsymbol{x}_{k,s} : k\in\text{Post}_p(\mathcal{O}_m)\cup\{p\}\,;W_i,h\big)\Big], \end{aligned} \tag{15}$$

while Equation (10) becomes

$$\hat{\phi}^{\text{out}}_p\big(\boldsymbol{x}_{T_{\text{in}}+(i-1)}\,;W_i,h\big) = \frac{1}{2M}\sum_{m=1}^{2M}\theta^{\text{out}}_{p,m}\big(\boldsymbol{x}_{T_{\text{in}}+(i-1)}\,;W_i,h\big) \tag{16}$$

for $p\in S$ and $i=1,\ldots,D-(h-1)$. Equation (15) continues to approximate the effect of removing predictors not in the coalition by replacing them with background data from the training sample W_i. The $\hat{\phi}^{\text{out}}_p\big(\boldsymbol{x}_{T_{\text{in}}+(i-1)}\,;W_i,h\big)$ estimate in Equation (16) is again characterized by the properties in Equations (4) to (7). Based on efficiency, we can decompose the out-of-sample forecast corresponding to $\boldsymbol{x}_{T_{\text{in}}+(i-1)}$:

$$\sum_{p\,\in\,S}\hat{\phi}^{\text{out}}_p\big(\boldsymbol{x}_{T_{\text{in}}+(i-1)}\,;W_i,h\big) = \hat{f}\big(\boldsymbol{x}_{T_{\text{in}}+(i-1)}\,;W_i,h\big) - \hat{\phi}_{\emptyset}(W_i,h) \tag{17}$$

for $i=1,\ldots,D-(h-1)$.

The key insight for computing the PBSV$_p$ is to wrap a loss function around the predictions in Equation (15). We denote a generic loss function by

$$L\Big(y_{T_{\text{in}}+i:T_{\text{in}}+h+(i-1)}, \hat{f}\big(\boldsymbol{x}_{T_{\text{in}}+(i-1)}; W_i,h\big)\Big) \tag{18}$$

for $i=1,\ldots,D-(h-1)$. To incorporate the loss function, we further modify the algorithm. For a random draw m of an ordered permutation, we adjust Equation (15) as follows:

$$\begin{aligned} &\theta^{\text{out}}_{p,m}\big(\boldsymbol{x}_{T_{\text{in}}+(i-1)}\,;W_i,h,L\big) = \\ &\quad L\Bigg(y_{T_{\text{in}}+i:T_{\text{in}}+h+(i-1)}, \frac{1}{|W_i|}\sum_{s\,\in\,W_i}\hat{f}\big(\boldsymbol{x}_{j,T_{\text{in}}+(i-1)} : j\in \text{Pre}_p(\mathcal{O}_m)\cup\{p\},\, \boldsymbol{x}_{k,s} : k\in\text{Post}_p(\mathcal{O}_m)\,;W_i,h\big)\Bigg) - \\ &\quad L\Bigg(y_{T_{\text{in}}+i:T_{\text{in}}+h+(i-1)}, \frac{1}{|W_i|}\sum_{s\,\in\,W_i}\hat{f}\big(\boldsymbol{x}_{j,T_{\text{in}}+(i-1)} : j\in \text{Pre}_p(\mathcal{O}_m),\, \boldsymbol{x}_{k,s} : k\in\text{Post}_p(\mathcal{O}_m)\cup\{p\}\,;W_i,h\big)\Bigg) \end{aligned} \tag{19}$$

for $p \in S$ and $i = 1, \ldots, D - (h-1)$. Equation (16) becomes

$$\hat{\phi}_p^{\text{out}}\big(\boldsymbol{x}_{T_{\text{in}}+(i-1)}\,;W_i,h,L\big) = \frac{1}{2M}\sum_{m=1}^{2M}\theta_{p,m}^{\text{out}}\big(\boldsymbol{x}_{T_{\text{in}}+(i-1)}\,;W_i,h,L\big) \tag{20}$$

for $p \in S$ and $i = 1, \ldots, D - (h-1)$.

The local PBSV$_p$ in Equation (20) measures the contribution of predictor p to the loss incurred by the ith out-of-sample forecast. Like Equation (15), Equation (19) approximates the effect of removing predictors not in the coalition by replacing them with background data from the training sample W_i. Based on the logic of Shapley values, the local PBSV$_p$ in Equation (20) fairly allocates the loss among the predictors for the ith out-of-sample forecast.[7] Along with the properties in Equations (5) to (7), Equation (20) is characterized by efficiency in Equation (4):

$$\begin{aligned}\sum_{p\,\in\,S}\hat{\phi}_p^{\text{out}}\big(\boldsymbol{x}_{T_{\text{in}}+(i-1)}\,;W_i,h,L\big) &= \\ L\Big(y_{T_{\text{in}}+i:T_{\text{in}}+h+(i-1)},\hat{f}\big(\boldsymbol{x}_{T_{\text{in}}+(i-1)}\,;W_i,h\big)\Big) &- \hat{\phi}_{\emptyset}^{\text{out}}(W_i,h,L)\end{aligned} \tag{21}$$

for $i = 1, \ldots, D - (h-1)$, where $\hat{\phi}_{\emptyset}^{\text{out}}(W_i,h,L)$ corresponds to the loss for the baseline or unconditional prediction based on the empty coalition set.

We are primarily interested in the performance of the entire sequence of out-of-sample forecasts, so we define the global PBSV$_p$. To obtain the global PBSV$_p$, we again modify the algorithm. Specifically, we expand Equation (19) to reflect the average loss for the out-of-sample period:

$$\begin{aligned}&\theta_{p,m}^{\text{out}}(W,h,L) = \\ &\frac{1}{|W|}\sum_{i\,\in\,W} L\left(y_{T_{\text{in}}+i:T_{\text{in}}+h+(i-1)},\frac{1}{|W_i|}\sum_{s\,\in\,W_i}\hat{f}\big(\boldsymbol{x}_{j,T_{\text{in}}+(i-1)} : j \in \text{Pre}_p(\mathcal{O}_m)\cup\{p\},\,\boldsymbol{x}_{k,s} : k \in \text{Post}_p(\mathcal{O}_m)\,;W_i,h\big)\right) - \\ &\frac{1}{|W|}\sum_{i\,\in\,W} L\left(y_{T_{\text{in}}+i:T_{\text{in}}+h+(i-1)},\frac{1}{|W_i|}\sum_{s\,\in\,W_i}\hat{f}\big(\boldsymbol{x}_{j,T_{\text{in}}+(i-1)} : j \in \text{Pre}_p(\mathcal{O}_m),\,\boldsymbol{x}_{k,s} : k \in \text{Post}_p(\mathcal{O}_m)\cup\{p\}\,;W_i,h\big)\right)\end{aligned} \tag{22}$$

for $p \in S$. Equation (22) again approximates the effect of removing predictors not in the coalition by replacing them with background data from the training sample. Equation (20) is now given by

$$\hat{\phi}_p^{\text{out}}(W,h,L) = \frac{1}{2M}\sum_{m=1}^{2M}\theta_{p,m}^{\text{out}}(W,h,L) \tag{23}$$

[7] Section A.1 of the Online Appendix provides the local PBSV$_p$ for the special case of a linear model (with no interactions) and squared error loss, for which we can derive an analytical expression. More generally, we need to rely on the algorithm to compute the PBSV$_p$.

for $p \in S$.

The global PBSV_p in Equation (23) allows us to decompose the average loss for a sequence of out-of-sample forecasts into the contributions of each of the P predictors. In this way, we anatomize out-of-sample performance by fairly assessing how the individual predictors contribute to out-of-sample forecasting accuracy. In addition to the properties in Equations (5) to (7), Equation (23) is again characterized by efficiency in Equation (4):

$$\sum_{p \in S} \hat{\phi}_p^{\text{out}}(W, h, L) = \frac{1}{|W|} \sum_{i \in W} L\Big(y_{T_{\text{in}}+i:T_{\text{in}}+h+(i-1)}, \hat{f}\big(\boldsymbol{x}_{T_{\text{in}}+(i-1)}; W_i, h\big)\Big) - \hat{\phi}_{\emptyset}^{\text{out}}(W, h, L), \tag{24}$$

where $\hat{\phi}_{\emptyset}^{\text{out}}(W, h, L)$ corresponds to the average loss for the sequence of baseline forecasts based on the empty coalition set.[8]

The PBSV_p bears some resemblance to the Shapley feature importance (SFIMP) in Casalicchio et al. (2018), as both are computed using a loss function for the test sample. However, there are important differences between the PBSV_p and SFIMP. The SFIMP assumes that the prediction model is estimated only once, which is more appropriate for cross-sectional data, while the PBSV_p is explicitly designed for time-series data when the out-of-sample forecasts are generated by a sequence of fitted models based on an expanding or rolling window. Furthermore, there are substantive differences in the algorithms used to compute the PBSV_p and SFIMP (beyond the fact that the former is based on a sequence of fitted models, while the latter is not). For example, the SFIMP uses background data from the test sample to control for predictors not in the coalition when computing Shapley values; in contrast, Equation (22) always uses background data from the training sample, so we remain true to the fitted models that generate the out-of-sample forecasts.[9] In sum, the PBSV_p furnishes a means for fairly allocating the out-of-sample loss for a sequence of time-series forecasts across the individual predictors, thereby showing exactly how each predictor contributes to out-of-sample performance. In this way, the PBSV_p provides an anatomy of out-of-sample forecasting accuracy.[10]

As an example of computing the PBSV_p for a specific loss function, consider the RMSE criterion:

$$\text{RMSE} = \left\{ \frac{1}{|W|} \sum_{i \in W} \Big[y_{T_{\text{in}}+i:T_{\text{in}}+h+(i-1)} - \hat{f}\big(\boldsymbol{x}_{T_{\text{in}}+(i-1)}\,; W_i, h\big)\Big]^2 \right\}^{0.5}. \tag{25}$$

[8] In addition to the entire out-of-sample period, the PBSV_p in Equation (23) can be computed for any subsample of the forecast evaluation period; for an example, see Figure 2 for the empirical application in Section 3.

[9] The PBSV_p has a different focus from the Shapley regressions proposed by Joseph (2021). Shapley regressions relate the realized target values to Shapley values for the out-of-sample observations in a linear regression framework.

[10] We use $M = 500$ for the algorithms when computing the TS-Shapley-VI_p in Equation (13) and the PBSV_p in Equation (23) for the empirical application in Section 3.

To obtain the global PBSV_p for the RMSE, we use the following version of Equation (22):

$$
\begin{aligned}
&\theta_{p,m}^{\text{out}}(W, h, \text{RMSE}) = \\
&\left\{ \frac{1}{|W|} \sum_{i \in W} \left[y_{T_{\text{in}}+i:T_{\text{in}}+h+(i-1)} - \frac{1}{|W_i|} \sum_{s \in W_i} \hat{f}\big(\boldsymbol{x}_{j,T_{\text{in}}+(i-1)} : j \in \text{Pre}_p(\mathcal{O}_m) \cup \{p\},\, \boldsymbol{x}_{k,s} : k \in \text{Post}_p(\mathcal{O}_m)\,; W_i, h\big) \right]^2 \right\}^{0.5} - \\
&\left\{ \frac{1}{|W|} \sum_{i \in W} \left[y_{T_{\text{in}}+i:T_{\text{in}}+h+(i-1)} - \frac{1}{|W_i|} \sum_{s \in W_i} \hat{f}\big(\boldsymbol{x}_{j,T_{\text{in}}+(i-1)} : j \in \text{Pre}_p(\mathcal{O}_m),\, \boldsymbol{x}_{k,s} : k \in \text{Post}_p(\mathcal{O}_m) \cup \{p\}\,; W_i, h\big) \right]^2 \right\}^{0.5}
\end{aligned}
\tag{26}
$$

for $p \in S$. The global PBSV_p in Equation (23) is then given by

$$
\hat{\phi}_p^{\text{out}}(W, h, \text{RMSE}) = \frac{1}{2M} \sum_{m=1}^{2M} \theta_{p,m}^{\text{out}}(W, h, \text{RMSE}) \tag{27}
$$

for $p \in S$. According to the efficiency property,

$$
\sum_{p \in S} \hat{\phi}_p^{\text{out}}(W, h, \text{RMSE}) = \text{RMSE} - \hat{\phi}_{\emptyset}^{\text{out}}(W, h, \text{RMSE}), \tag{28}
$$

so we exactly decompose the out-of-sample RMSE into the contributions made by each of the predictors.

2.3. Model Accordance Score

We use the MAS to compare predictor ranks according to the TS-Shapley-VI_p in Equation (13) and the PBSV_p in Equation (23). The aim is to gauge how well the in-sample importance of the predictors in the sequence of fitted forecasting models aligns with the predictors' roles in determining out-of-sample forecasting accuracy. The MAS is a type of Spearman rank correlation between a list of P strictly positive ranks $A \in \{1, \ldots, P\}$ (corresponding to the TS-Shapley-VI_p) and a list of P both negative and positive ranks $B \in \{-P, \ldots, -1, 1, \ldots, P\}$ (corresponding to the PBSV_p). A is derived from the TS-Shapley-VI_p by ranking the predictors in ascending order (with the highest variable importance receiving the highest rank). B is derived from the PBSV_p by ranking "good" predictors (i.e., those that reduce the out-of-sample loss) separately from "bad" predictors (i.e., those that increase the out-of-sample loss). Good predictors are ranked in ascending order from one to the number of good predictors, where the best predictor receives the highest rank (which is at most P, if all of the predictors are good); bad predictors are ranked from -1 to the negative of the number of bad predictors, where the worst predictor receives the most negative rank (which is $-P$ in the limit if all of the predictors are bad). In the case where all of the predictors contribute to lowering out-of-sample loss and the relative importance is the same according to the TS-Shapley-VI$_p$ and PBSV$_p$, then $A = B$. At the other extreme, the most important predictors according to the TS-Shapley-VI$_p$ are the worst according to the PBSV$_p$ and contribute to increasing out-of-sample loss, so $B = -A$.

We define the MAS as

$$\text{MAS} = 1 - \frac{\text{MSDR}}{\mathbb{E}[\text{MSDR}]}, \tag{29}$$

where MSDR is the weighted mean squared deviation in ranks:

$$\text{MSDR} = \frac{1}{P}\sum_{p=1}^{P} w_p \Big[\text{rank}\big(\text{TS-Shapley-VI}_p(W,h)\big) - \text{signed-rank}\Big(\hat{\phi}_p^{\text{out}}(W,h,L)\Big)\Big]^2, \tag{30}$$

w_p is the weight for predictor p (the weights are scaled to sum to P), and we standardize the MSDR in Equation (29) by dividing by the expectation of Equation (30) under the assumption that good predictors are as likely as bad predictors in terms of the out-of-sample loss.[11] The greater the accord in ranks between the TS-Shapley-VI$_p$ and PBSV$_p$, the lower (higher) the MSDR (MAS) will be; when there is exact agreement between the ranks ($A = B$), the MSDR (MAS) reaches its minimum (maximum) value of zero (one).

In empirical applications, certain predictors often receive substantially higher in-sample variable importance, while others have variable importance close to zero. To account for such differences in in-sample variable importance, we weight the differences in ranks in Equation (30) proportionally to the TS-Shapley-VI$_p$ by setting $w_p = \frac{\text{TS-Shapley-VI}_p}{\frac{1}{P}\sum_{p=1}^{P}\text{TS-Shapley-VI}_p}$. Note that the scaling of the weights implies that $\sum_{p=1}^{P} w_p = P$, so the average value of the weights is one. The equal-weighted case corresponds to $w_p = 1$ for $p = 1, \ldots, P$.

We test for a significant relationship between the ranks for the TS-Shapley-VI$_p$ and PBSV$_p$ (A and B, respectively). We do so by generating a distribution for the MSDR under the null hypothesis of no relationship between the ranks and computing an empirical p-value for the MSDR corresponding to the original data. We generate random (i.e., unrelated) ranks under the null hypothesis as follows. To simulate a random rank of predictors for B (PBSV$_p$), we first draw $P_+ \sim$

[11] When good predictors are as likely as bad predictors in terms of the out-of-sample loss and the weights sum to P, it can be shown that

$$\mathbb{E}[\text{MSDR}] = \frac{1}{P}\left(\sum_{p=1}^{P}\Big[w_p\,\text{rank}\big(\text{TS-Shapley-VI}_p(W,h)\big)^2\Big] + \sum_{a=0}^{P}\left\{\binom{P}{a}0.5^P[S(a)+S(P-a)]\right\}\right),$$

where $S(n) = [n(n+1)(2n+1)]/6$.

Binomial(P, α), where α is a hyperparameter corresponding to the proportion of good predictors anticipated by the researcher under the null hypothesis, and set $P_- = P - P_+$.[12] Then, we randomly draw a sequence of P elements from $\{-P_-, .., -1, 1, .., P_+\}$ without replacement. Based on the original weights and ranks for the TS-Shapley-VI$_p$ and the randomly drawn predictor ranks for B, we compute the MSDR in Equation (30). Repeating this many times, we generate an empirical distribution for the MSDR under the null hypothesis and compute the empirical p-value as the proportion of generated MSDR values that are less than or equal to the MSDR for the original data.

We created the Python package anatomy to implement the algorithms for calculating the TS-Shapley-VI$_p$, PBSV$_p$, and MAS. Section A.2 of the Online Appendix provides computational details for the algorithms in the package.[13]

3. Forecasting Inflation

In this section, we use the metrics developed in Section 2 to analyze predictor relevance in out-of-sample forecasts of US inflation. Recent evidence shows that traditional inflation benchmark forecasts can be outperformed by the use of big data in conjunction with machine-learning methods and that the outperformance is largely attributable to nonlinearities, especially at longer horizons (e.g., Medeiros et al., 2021; Goulet Coulombe, 2022; Goulet Coulombe et al., 2022; Hauzenberger et al., 2023). We forecast inflation using a large dataset and a variety of machine-learning models.

3.1. Forecasting Models

Consider the following prediction model for inflation:

$$\pi_{t+1:t+h} = f\left(\boldsymbol{\pi}_{t-L:t}^{\text{AR}}, \boldsymbol{w}_t, \boldsymbol{w}_t^{\text{MA}(q)}\right) + \varepsilon_{t+1:t+h}, \tag{31}$$

where $\pi_{t+1:t+h} = (1/h)\sum_{k=1}^{h} \pi_{t+k}$, $\pi_t = \log(\text{CPI}_t) - \log(\text{CPI}_{t-1})$, CPI$_t$ is the month-t US consumer price index (CPI), $\boldsymbol{\pi}_{t-L:t}^{\text{AR}} = [\ \pi_t \quad \cdots \quad \pi_{t-L}\ ]'$ collects the AR components in inflation, $\boldsymbol{w}_t$ is a vector of predictors, and $\boldsymbol{w}_t^{\text{MA}(q)} = (1/q)\sum_{k=1}^{q} \boldsymbol{w}_{t-(k-1)}$ is a vector of moving averages (MAs) of order q for the predictors in $\boldsymbol{w}_t$. We gather the entire set of predictors in the P-dimensional vector $\boldsymbol{x}_t = [\ \boldsymbol{\pi}_{t-L:t}^{\text{AR}}{}' \quad \boldsymbol{w}_t' \quad \boldsymbol{w}_t^{\text{MA}(q)\,\prime}\ ]'$. The inclusion of MAs of the predictors is motivated by Goulet

[12] Setting the hyperparameter α depends on the forecasting environment. Specifically, it should be set to the proportion of predictors expected to contribute to reducing the loss against the model with an empty set of predictors (commonly the unconditional mean forecast).

[13] While we present the TS-Shapley-VI$_p$, PBSV$_p$, and MAS metrics in terms of a regression problem, it is straightforward to adapt the metrics for a classification problem.

Coulombe et al. (2021), who find that MAs of predictors provide substantive out-of-sample gains for forecasting macroeconomic variables. We set $q = 3$, which allows predictors up to a quarter in the past to affect the prediction. In terms of the AR components, we set $L = 11$, corresponding to twelve lags of inflation in Equation (31). Based on Equation (31), the forecast of $\pi_{t+1:t+h}$ is given by

$$\hat{\pi}_{t+1:t+h} = \hat{f}(\boldsymbol{x}_t), \tag{32}$$

where $\hat{f}$ is the fitted prediction function based on data through t.

We consider a variety of machine-learning models for forecasting inflation based on Equation (31).

- Principal component regression (PCR, Stock and Watson, 2002a,b)
- Elastic net (ENet, Zou and Hastie, 2005) estimation of a linear model
- Random forest (Breiman, 2001)
- XGBoost (Chen and Guestrin, 2016)
- Neural network

The first two models are linear, while the last three allow for nonlinearities in the prediction function. We also consider ensembles of individual forecasting models, which are popular in machine learning. An ensemble forecast can be straightforwardly computed as a simple average of the forecasts generated by the models in the ensemble.[14]

- Ensemble-linear: average of the PCR and ENet forecasts
- Ensemble-nonlinear: average of the random forest, XGBoost, and neural network forecasts
- Ensemble-all: average of PCR, ENet, random forest, XGBoost, and neural network forecasts

Section A.3 of the Online Appendix provides details for the construction of the different forecasting models.

[14] The algorithm for computing the PBSV$_p$ accommodates ensemble forecasts (as shown in Section A.2 of the Online Appendix).

3.2. Data

We measure inflation based on the US CPI. CPI data are from the FRED database at the Federal Reserve Bank of St. Louis (ticker CPIAUCSL). The predictors are from two data sources. We use a set of 118 predictors from the FRED-MD database (McCracken and Ng, 2016), which is employed by a number of recent macroeconomic forecasting studies (e.g., Kotchoni et al., 2019; Medeiros et al., 2021; Borup and Schütte, 2022; Goulet Coulombe et al., 2022; Hauzenberger et al., 2023). We also include three predictors from the University of Michigan Survey of Consumers.[15] The sample period covers 1960:01 to 2022:12. We specify 1960:01 to 1989:12 as the initial in-sample period and compute out-of-sample forecasts for 1990:01 to 2022:12. As in Medeiros et al. (2021), among others, we generate out-of-sample inflation forecasts using a rolling estimation window.

3.3. Results

An AR model of order k serves as the benchmark, where we determine k using the Bayesian information criterion (BIC, Schwarz, 1978), considering a maximum value of twelve. We also estimate the AR benchmark model via a rolling window. The AR model is a standard benchmark in the macroeconomic forecasting literature, including for inflation (e.g., Kotchoni et al., 2019; Medeiros et al., 2021).

We evaluate the forecasts using the RMSE criterion. Table 1 reports results for the accuracy of the inflation forecasts for horizons of one, three, six, and twelve months. The table provides the RMSE for the AR benchmark forecast, as well as the RMSE ratio for each of the competing models vis-à-vis the AR benchmark. We use the Diebold and Mariano (1995) and West (1996) statistic to test the null hypothesis that the MSE (in population) for the AR benchmark forecast is less than or equal to that for the competing forecast against the (one-sided, upper-tail) alternative that the AR forecast MSE is greater than the competing forecast MSE.[16]

The RMSE for the AR benchmark forecast decreases monotonically with the horizon from 0.26% ($h = 1$) to 0.16% ($h = 12$) in Table 1. At the one-month horizon in the second column, six of the eight competing forecasts deliver a lower RMSE than the AR benchmark (the exceptions are PCR and XGBoost), and the improvement in forecasting accuracy is statistically significant for the ENet, neural network, ensemble-nonlinear, and ensemble-all forecasts. The ENet, ensemble-nonlinear, and ensemble-all forecasts provide the largest improvements in accuracy, each with an RMSE ratio of 0.93. Seven of the eight competing forecasts outperform the AR benchmark at the three-month

[15] Section A.4 of the Online Appendix provides a complete list of the inflation predictors.

[16] We use a robust standard error (Newey and West, 1987) to compute the Diebold and Mariano (1995) and West (1996) statistic, which accounts for the autocorrelation induced by overlapping observations when $h > 1$.

Table 1. Out-of-sample forecasting results

The table reports the root mean squared error (RMSE) for the autoregressive benchmark forecast and the RSME ratio for the competing forecast in the first column vis-à-vis the autoregressive benchmark forecast for inflation for the 1990:01 to 2022:12 out-of-sample period and the horizon (h) in the column heading. The Diebold and Mariano (1995) and West (1996) statistic is used to test the null hypothesis that the benchmark forecast MSE is less than or equal to the competing forecast MSE against the (one-sided, upper-tail) alternative hypothesis that the benchmark forecast MSE is greater than the competing forecast MSE; *, **, and *** indicate significance at the 10%, 5%, and 1% level, respectively.

(1) Forecast	(2) $h = 1$	(3) $h = 3$	(4) $h = 6$	(5) $h = 12$
Autoregressive benchmark RMSE	0.26%	0.23%	0.20%	0.16%
Principal component regression	1.08	1.01	0.96	0.92^{**}
Elastic net	0.93^{**}	0.95^{*}	0.96	0.94
Random forest	0.96	0.97	0.92^{*}	0.82^{***}
XGBoost	1.00	0.98	0.91^{**}	0.85^{***}
Neural network	0.94^{**}	0.93^{**}	0.94	0.83^{***}
Ensemble-linear	0.96	0.96	0.93^{*}	0.90^{**}
Ensemble-nonlinear	0.93^{**}	0.93^{**}	0.90^{**}	0.81^{***}
Ensemble-all	0.93^{**}	0.93^{**}	0.90^{**}	0.84^{***}

horizon in the third column (the exception is PCR). The improvements are again significant for the ENet, neural network, ensemble-nonlinear, and ensemble-all forecasts. The biggest gain in accuracy is for the neural network, ensemble-nonlinear, and ensemble-all forecasts (RMSE ratio of 0.93 for each). The results are fairly similar at the six-month horizon in the fourth column, although now all of the competing forecasts outperform the AR benchmark, and the improvement is significant in five cases (random forest, XGBoost, ensemble-linear, ensemble-nonlinear, and ensemble-all). The largest gain in accuracy is for the ensemble-nonlinear and ensemble-all forecasts (RMSE ratio of 0.90 for each).

The best overall results are at the twelve-month horizon in the last column of Table 1. All eight of the competing forecasts outperform the AR benchmark, and seven of the improvements are significant (the exception is the ENet). The nonlinear forecasts perform very well for $h = 12$, with RMSE reductions of 18%, 15%, and 17% vis-à-vis the AR benchmark for the random forest, XGBoost, and neural network, respectively. This pattern is consistent with the recent literature that finds that nonlinear machine-learning models are particularly useful for forecasting inflation at longer horizons. The ensemble forecasts also perform well at the twelve-month horizon, as each delivers a significant improvement in forecasting accuracy. Reiterating the strong performance of the nonlinear models, the ensemble-nonlinear forecast performs the best at the 12-month horizon, reducing the RMSE by 19% relative to the AR benchmark.

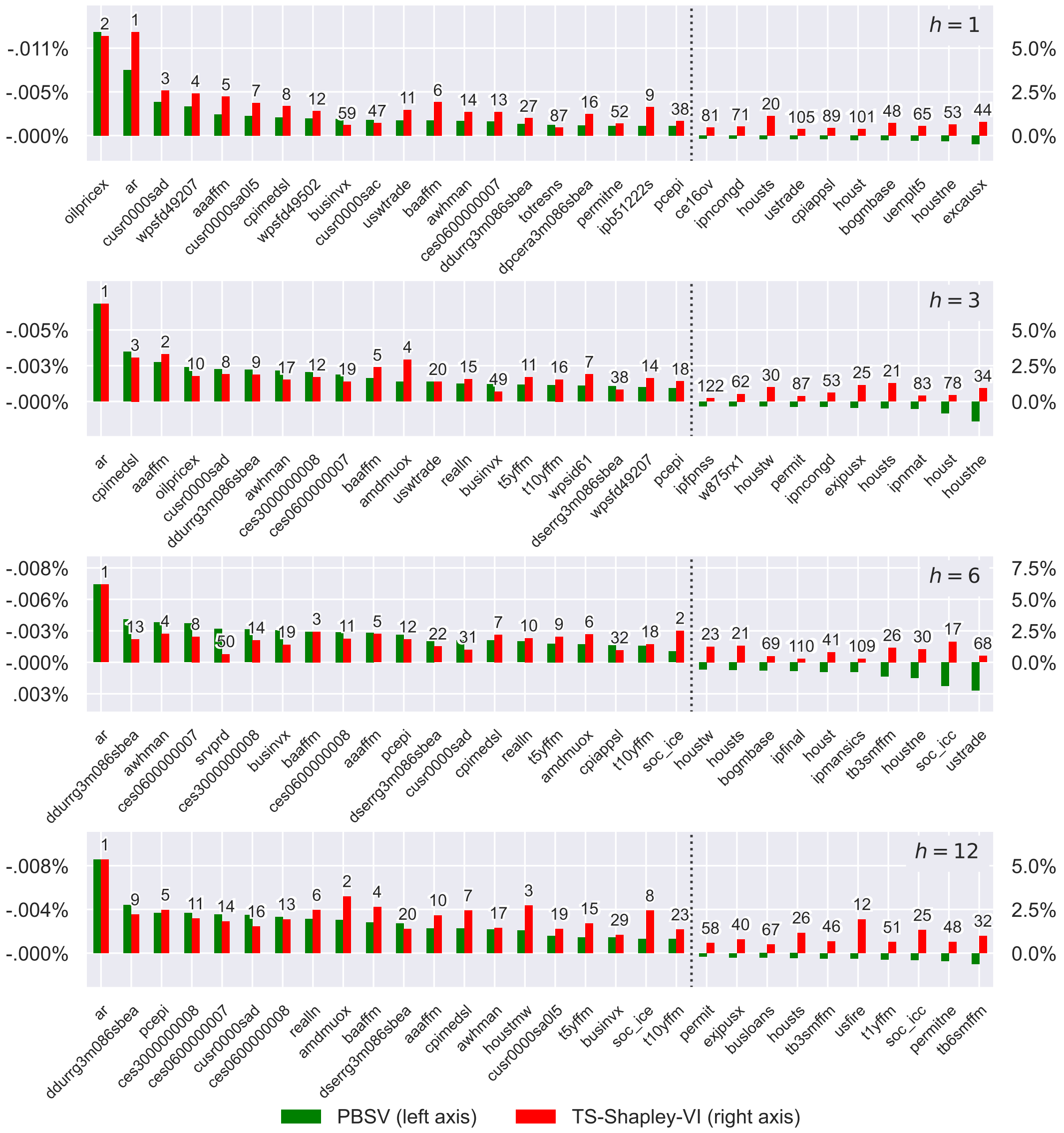

Figure 1. PBSV and TS-Shapley-VI: ensemble-nonlinear. The figure shows the PBSV$_p$ (left axis) and TS-Shapley-VI$_p$ (right axis) for the ensemble-nonlinear inflation forecast for the 1990:01 to 2022:12 out-of-sample period. The predictors on the horizontal axis are the top 20 and the bottom ten ordered according to the PBSV$_p$ in terms of improving out-of-sample forecasting accuracy. The numbers associated with the red bars are the predictor ranks according to the TS-Shapley-VI$_p$.

Next, we demonstrate how our new metrics in Section 2 can be used to interpret fitted prediction models by anatomizing out-of-sample forecasting performance. Figure 1 depicts the PBSV$_p$ based on the RMSE and the TS-Shapley-VI$_p$ for the ensemble-nonlinear forecast. We focus on the ensemble-nonlinear forecast to conserve space and because it performs well overall in Table 1.

Figures A.1 to A.7 in the Online Appendix provide analogous versions of Figure 1 for the other forecasts.

The different panels in Figure 1 display results for the different horizons. The predictors on the horizontal axis in each panel are ordered according to the PBSV_p in terms of their contributions to improving out-of-sample forecasting accuracy. We refer to the predictors on the horizontal axis by their FRED-MD abbreviations; see Table A.1 in the Online Appendix. The green (red) bars correspond to the PBSV_p (TS-Shapley-VI_p).[17] To conserve space, the horizontal axis shows the top 20 and bottom ten predictors based on the PBSV_p. The numbers associated with the red bars are the predictor ranks based on the TS-Shapley-VI_p.

The green bars to the left of the dotted vertical line in each panel of Figure 1 identify the 20 predictors that contribute the most to lowering the RMSE (i.e., improving forecasting accuracy) for the ensemble-nonlinear forecast. At the one-month horizon, the price of oil (`oilpricex`) is the top contributor, highlighting the importance of oil price fluctuations in affecting short-run inflation. The price of oil also ranks fourth at the three-month horizon. The AR components (`ar`) make major contributions at all reported horizons: they rank second at the one-month horizon and first at the other reported horizons. Other predictors that consistently rank highly across all reported horizons in Figure 1 include the durables component of the CPI (`cusr0000sad`), the medical services component of the CPI (`cpimedsl`), the durable goods component of the personal consumption expenditures price index (`ddurrg3m086sbea`), average weekly hours for the goods producing sector (`ces0600000007`), average weekly hours in manufacturing (`awhman`), the personal consumption expenditures price index (`pcepi`), and the spreads between Aaa- and Baa-rated corporate bond yields and the federal funds rate (`aaaffm` and `baaffm`, respectively).

According to the green bars to the right of the dotted vertical lines in Figure 1, there are a number of predictors that substantively detract from out-of-sample forecasting accuracy, including a number relating to housing, such as total housing starts (`houst`) and housing starts in the South (`housts`) at all reported horizons; housing starts in the Northeast (`houstne`) at the one-, three-, and six-month horizons; housing starts in the West at the three- and six-month horizons; total new housing permits (`permit`) at the six- and twelve-month horizons; and new housing permits in the Northeast (`permitne`) at the twelve-month horizon.

Comparing the red and green bars in Figure 1, many of the predictors listed above that are leading contributors to out-of-sample forecasting accuracy based on the PBSV_p are also among the

[17] In Figure 1, we sum the Shapley values for each predictor and its corresponding $\text{MA}(q)$ term. We also sum the Shapley values for the twelve lags of inflation.

most important predictors on an in-sample basis according to the TS-Shapley-VI$_p$. Nevertheless, there are a few predictors that evince marked differences across the PBSV$_p$ and TS-Shapley-VI$_p$ to the right of the vertical dashed lines in Figure 1. For example, housing starts in the South, which contributes adversely to out-of-sample performance, ranks among the most important variables on an in-sample basis at all reported horizons. Other predictors exhibiting a similar pattern include the index of current economic conditions (`soc_icc`) at the six- and twelve-month horizons and employment in the financial activities sector (`usfire`) at the twelve-month horizon. The MAS values subsequently reported in Table 2 quantify the degree of accordance between predictor ranks based on the PBSV$_p$ and TS-Shapley-VI$_p$.

In sum, the PBSV$_p$ quantifies the contributions of predictors to the accuracy of CPI inflation forecasts for the 1990:01 to 2022:12 out-of-sample period. It allows us to pinpoint predictors that play leading roles in accounting for the out-of-sample gains in forecasting accuracy, as well as to identify predictors that detract from out-of-sample forecasting accuracy. Based on Figure 1 and Figures A.1 to A.7 in the Online Appendix, in terms of the most important predictors for improving the accuracy of inflation forecasts across the different models, the PBSV$_p$ identifies the price of oil at shorter horizons, as well as the AR components, the durables component of the CPI, the medical services component of the CPI, and the spread between the Aaa-rated corporate bond yield and the federal funds rate at all reported horizons.

We also illustrate how the PBSV$_p$ can shed light on the most important contributors to forecasting accuracy for subsamples of the entire sequence of time-series forecasts. This provides a sense of the predictor contributions to forecasting accuracy over time. Figure 2 plots the cumulative difference in squared errors (CDSE, Goyal and Welch, 2003, 2008) between a naïve forecast that ignores the information in the predictors and the ensemble-nonlinear forecast. We again focus on the ensemble-nonlinear forecast to conserve space. Figures A.8 to A.14 in the Online Appendix provide analogous versions of Figure 2 for the other forecasts. To further conserve space, we report results for horizons of one, six, and twelve months in Figure 2.

The CDSE is a convenient and informative graphical device for ascertaining whether a competing forecast is more accurate than the naïve forecast for any subsample of the out-of-sample period. In terms of Figure 2, we compare the CDSE at the beginning and end of the interval corresponding to a subsample. If the curve lies to the right (left) at the end of the interval relative to the beginning, then the ensemble-nonlinear (naïve) forecast is more accurate in terms of MSE for the subsample. In addition, we compute the PBSV$_p$ for the predictors for the ensemble-nonlinear forecast for non-overlapping twelve-month rolling subsamples. The abbreviation to the right (left)

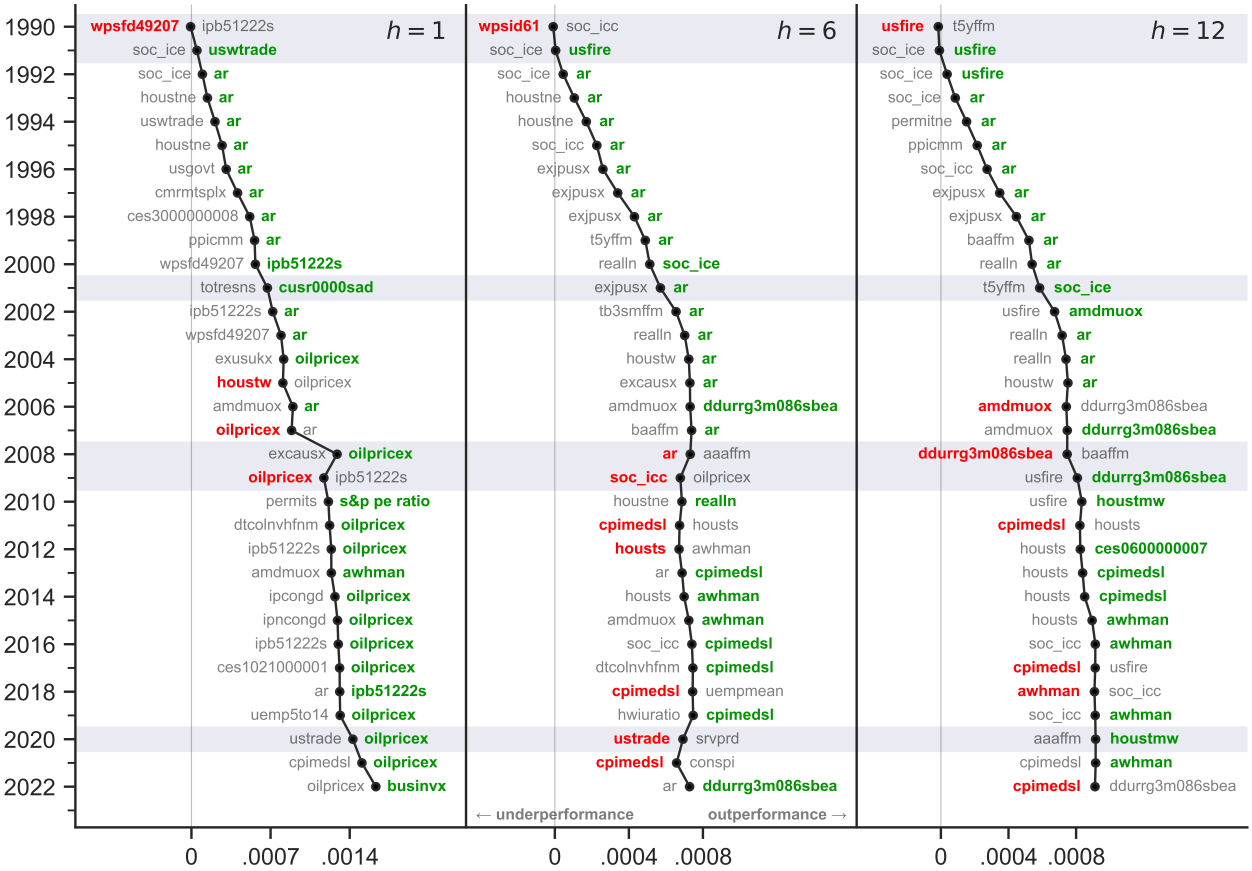

Figure 2. Cumulative difference in squared errors: ensemble-nonlinear. The figure shows the cumulative difference in squared errors for a naïve forecast that ignores the information in the predictors vis-à-vis the ensemble-nonlinear forecast for the 1990:01 to 2022:12 out-of-sample period. Shifts to the right (left) imply an improvement (deterioration) in forecasting accuracy relative to the naïve forecast. The figure also shows the top (bottom) contributor to the improvement (deterioration) in forecasting performance as identified by the PBSV$_p$ for non-overlapping twelve-month subsamples; a green (red) color for the predictor indicates that the subsample is associated with an improvement (deterioration) in performance. Horizontal gray bars delineate twelve-month subsamples that contain an NBER-dated recession.

of the curve in Figure 2 indicates the predictor that contributes the most to improving (detracting from) performance during a subsample. A predictor in green (red) to the right (left) of the curve indicates that the ensemble-nonlinear forecast delivers a lower (higher) MSE than the naïve forecast for the twelve-month subsample. The horizontal gray bars delineate twelve-month subsamples that contain an NBER-dated recession.

The CDSE plots in Figure 2 are consistently positively sloped (when viewed from top to bottom), so the ensemble-nonlinear forecast outperforms the naïve forecast on a consistent basis over time. For numerous twelve-month periods before the Great Recession in 2008, the AR components are most responsible for the outperformance of the ensemble-nonlinear forecast, consistent with the top and bottom two panels of Figure 1. In line with the top panel of Figure 1, at the one-month horizon in the left panel of Figure 2, there are eleven twelve-month periods when the price of oil is the predictor most responsible for the outperformance of the ensemble-nonlinear forecast, including during the Great Recession and the recent recession corresponding to the advent of COVID-19, as well as the inflation surge starting in mid 2021. However, there are two twelve-month periods when the price of oil detracts the most from forecasting accuracy, pointing to noteworthy time variation in the predictor's contribution to forecasting accuracy.

The medical services component of the CPI is the leading predictor in terms of the outperformance of the ensemble-nonlinear forecast for four of the twelve-month subsamples at the six-month horizon in the middle panel of Figure 2, consistent with the third panel of Figure 1. Economically, it accords with Bils and Klenow (2004), who rank medical care among the stickiest components of the CPI (in terms of its low frequency of price adjustment), and it is an important component in the Federal Reserve Bank of Atlanta's Sticky-Price CPI. However, there are a few twelve-month subsamples in the middle panel of Figure 2 when the medical services component of the CPI detracts the most from forecasting accuracy. Thus, like the price of oil at the one-month horizon, the medical services component of the CPI evinces important time variation in its contribution to forecasting accuracy at the six-month horizon. A similar situation holds for the medical services component of the CPI at the twelve-month horizon in the right panel of Figure 2.

Table 2. Model accordance scores

The table reports the model accordance score (MAS) for the inflation forecast in the first column for the 1990:01 to 2022:12 out-of-sample period and the horizon (h) in the column heading. The MAS compares the predictor ranks in terms of the in-sample TS-Shapley-VI$_p$ and out-of-sample PBSV$_p$, where a higher score indicates greater agreement in predictor ranks; *, **, and *** indicate significance at the 10%, 5%, and 1% level, respectively, with the hyperparameter for the proportion of good predictors under the null hypothesis set to $\alpha = 2/3$.

(1)	(2)	(3)	(4)	(5)
Forecast	$h = 1$	$h = 3$	$h = 6$	$h = 12$
Principal component regression	0.60***	0.58***	0.49**	0.45*
Elastic net	0.55**	0.37	0.59**	0.40
Random forest	0.82***	0.62***	0.75***	0.88***
XGBoost	0.53**	0.34	0.46*	0.57***
Neural network	0.56***	0.57***	0.27	0.46*
Ensemble-linear	0.64***	0.52**	0.54**	0.53**
Ensemble-nonlinear	0.65***	0.61***	0.43	0.59***
Ensemble-all	0.68***	0.64***	0.48**	0.65***

Next, we analyze the MAS in Equation (29). Table 2 reports the MAS for the different forecasts and horizons, where we set the hyperparameter α equal to $2/3$. Recall that the MAS measures the agreement (in terms of predictor ranks) between the in-sample TS-Shapley-VI$_p$ and the out-of-sample PBSV$_p$. As the MAS increases, there is greater agreement between the predictors that are deemed important in the sequence of fitted models that generate the forecasts and those that contribute to improvements in out-of-sample forecasting accuracy. A higher MAS indicates that the training of the sequence of prediction models identifies the most relevant predictors for improving out-of-sample performance, thereby inspiring more confidence in the reliability of the model and implying less reliance on good luck in accounting for a model's out-of-sample success.

All of the MAS measures are positive in Table 2, ranging from 0.27 (neural network, $h = 6$) to 0.88 (random forest, $h = 12$). Many are statistically significant at conventional levels: 27, 24, and 16 of the 32 MAS metrics are significant at the 10%, 5%, and 1% level, respectively. Thus, there is generally considerable agreement between the in-sample importance of predictors in the fitted models and their contributions to out-of-sample forecasting accuracy.

The MAS measures for the ensemble-nonlinear forecast in Table 2 align with the impressions of the results in Figure 1. The MAS is relatively large and statistically significant for the ensemble-nonlinear forecast at the one-, three-, and twelve-month horizons. It is smaller and insignificant at conventional levels at the six-month horizon, in line with the more sizable discrepancies in rankings to the right of the dashed line in the third panel of Figure 1. Recall from Table 1 that the ensemble-nonlinear forecast outperforms the AR benchmark forecast at the six-month horizon (RMSE ratio of 0.90, significant at the 10% level). The results for the ensemble-nonlinear forecast at the six-month horizon in Figure 1 and Table 2 suggest that the ability of the forecast to outperform the AR benchmark involves some luck. Also, recall from Table 1 that the ensemble-nonlinear forecast significantly outperforms the AR benchmark at horizons of one, three, and twelve months. The MAS results in Table 2 indicate that luck plays a more limited role in the out-of-sample success of the ensemble-nonlinear forecast at those horizons.

Finally, Figure 3 provides additional insight into links between the MAS and out-of-sample forecasting accuracy as measured by the RMSE. Each panel in the figure depicts a quadrant plot with RMSE (MAS) on the horizontal (vertical) axis, where both measures are standardized in the form of Z-scores. The top-right *intentional success* quadrant is the most desirable, as forecasts located there have above-average MAS and below-average RMSE.[18] In contrast, the bottom-right *unintentional success* quadrant is less desirable, in the sense that forecasts located there have

[18] Note that the RMSE on the horizontal axis is decreasing from left to right.

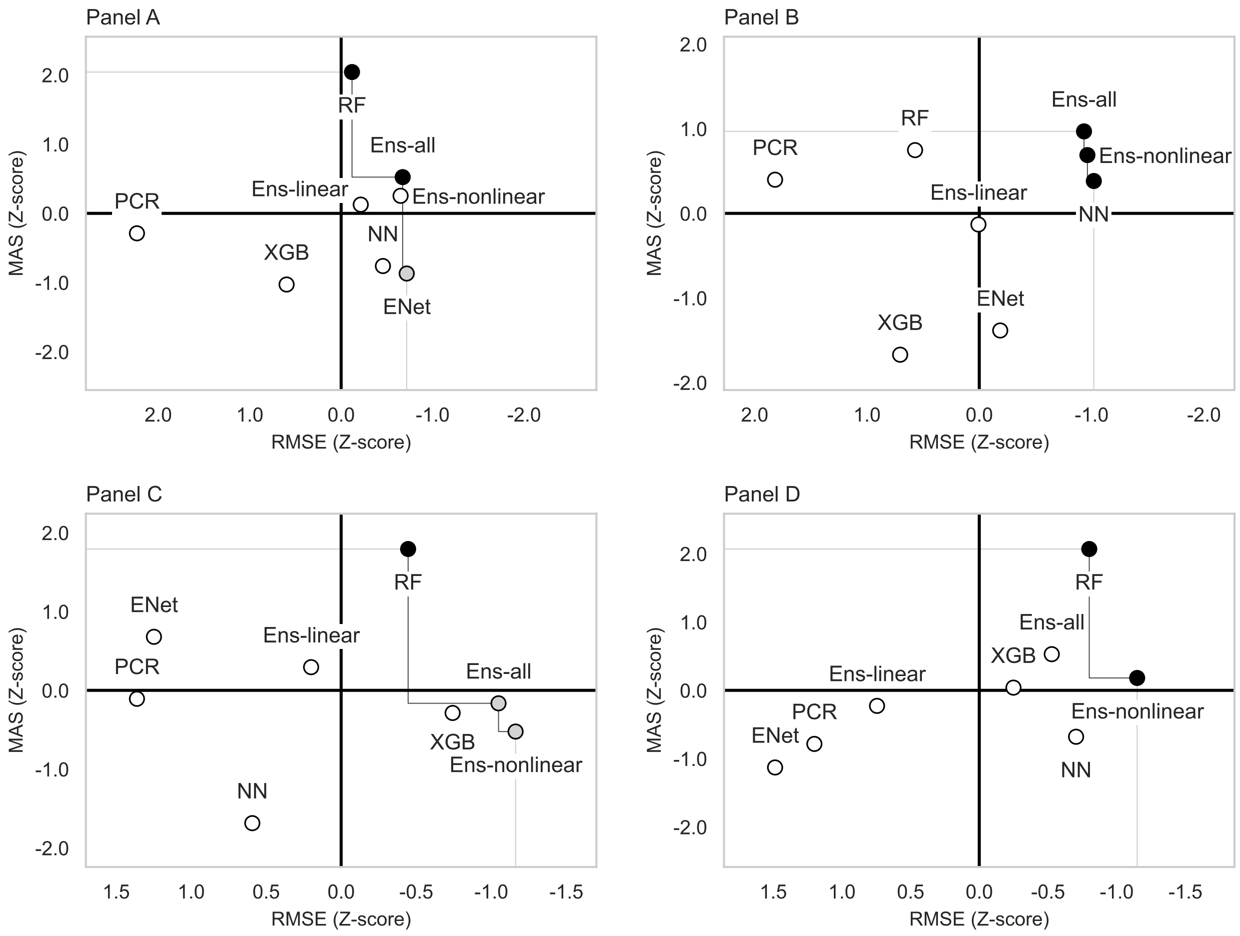

Figure 3. Quadrant plots. The figure shows the RMSE and MAS Z-scores in quadrant plots for forecast horizons of one, three, six, and twelve months in Panels A through D, respectively. The frontier of non-dominated forecasts is highlighted (a forecast is non-dominated if no other model has a higher MAS and a lower RMSE). Black dots are non-dominated forecasts (i.e., are on the frontier) that have above-average MAS and below-average RMSE. Gray dots are non-dominated forecasts that have below-average MAS and RMSE.

lower-than-average RMSE but also below-average MAS, suggesting that luck plays a larger role in accounting for their out-of-sample success. Each quadrant plot includes a frontier of non-dominated forecasts. A forecast is non-dominated if no other forecast has both a higher MAS and a lower RMSE. Forecasts with black dots in Figure 3 are non-dominated forecasts with above-average MAS and below-average RMSE (i.e., are in the intentional success quadrant); non-dominated forecasts with gray dots have below-average MAS and RMSE (i.e., are in the unintentional success quadrant).

Panel A indicates that the random forest and ensemble-all forecasts perform well at the one-month horizon in terms of the MAS-RMSE combination. Both forecasts are in the intentional success quadrant and lie on the frontier. The ensemble-linear and ensemble-nonlinear forecasts also reside in the intentional success quadrant, and the latter is very close to the frontier. At the three-month horizon in Panel B, the neural network, ensemble-nonlinear, and ensemble-all forecasts stand out, as they all lie on the frontier and are in the intentional success quadrant.

Turning to the six-month horizon in Panel C, the random forest lies on the frontier and is the only forecast in the intentional success quadrant. The ensemble-nonlinear and ensemble-all forecasts are also on the frontier and in the unintentional success quadrant. However, the MAS metrics for these forecasts lie close to the boundary for the intentional success quadrant, so it is unlikely that luck plays an outsize role in explaining their out-of-sample success. At the twelve-month horizon in Panel D, the random forest and ensemble-nonlinear forecasts lie on the frontier and are in the intentional success quadrant, while the ensemble-all and XGBoost forecasts are in the intentional success quadrant but not on the frontier. Overall, the quadrant plots in Figure 3 identify the random forest, ensemble-nonlinear, and ensemble-all forecasts as among the best forecasts with regard to intentional success.

4. Conclusion

Many economic agents rely extensively on time-series forecasts when making decisions, including forecasts of macroeconomic and financial variables. As large datasets and machine learning grow in popularity in macroeconomics and finance, the interpretation of forecasting models fitted with time-series data is becoming increasingly important. We develop the PBSV$_p$ to measure the contributions of individual predictors in fitted machine-learning models to out-of-sample forecasting accuracy, thereby furnishing a powerful new model-interpretation tool that fosters a deeper understanding of the sources of a model's out-of-sample performance. The PBSV$_p$ is model agnostic—so it can be applied to any machine-learning model—and can be used for any loss function, making it a very flexible tool.

We develop two additional metrics to complement the PBSV$_p$. The first is the TS-Shapley-VI$_p$, which extends the conventional Shapley-based variable-importance metric by measuring a predictor's importance across the entire set of fitted prediction models that generates the sequence of out-of-sample forecasts. The second is the MAS, which compares predictor ranks based on the TS-Shapley-VI$_p$ and PBSV$_p$. As the MAS increases, there is greater accord between the predictors' importance in the sequence of fitted models used to generate the out-of-sample forecasts and their importance with respect to out-of-sample performance. A relatively high MAS together with a low average loss indicates that the model learned from the in-sample data in a manner that leads to out-of-sample success, while a relatively low MAS and average loss suggest that luck plays a more substantive role in the model's out-of-sample success. In the former (latter) case, we can be more (less) confident that a model will continue to perform well on an out-of-sample basis going forward.

To demonstrate the use of the PBSV$_p$, TS-Shapley-VI$_p$, and MAS metrics, we undertake an empirical application forecasting monthly US inflation based on a large number of predictors and a variety of machine-learning methods. In line with recent studies, machine-learning forecasts generally outperform a standard AR benchmark at horizons ranging from one to twelve months. The outperformance is the greatest at the twelve-month horizon for machine-learning methods that allow for nonlinearities.

According to the PBSV$_p$, predictors that play leading roles in improving forecasting accuracy across the different models include the price of oil at shorter horizons, as well as the durables component of the CPI, the medical services component of the CPI, and the spread between the Baa-rated corporate bond yield and the federal funds rate at all reported horizons. Using the MAS to compare predictor ranks based on the TS-Shapley-VI$_p$ and PBSV$_p$, we find considerable agreement between the in-sample importance of predictors in fitted models and their contributions to out-of-sample forecasting accuracy, although the link is relatively weak for some models. We use quadrant plots to identify models that deliver a relatively low RMSE combined with a high MAS; such models successfully learn from the in-sample data to reliably improve out-of-sample forecasting accuracy. The random forest, ensemble-nonlinear, and ensemble-all forecasts generally perform the best in terms of the quadrant plots. Overall, our new metrics provide keen insight into the sources of the out-of-sample forecasting accuracy of machine-learning forecasts of US inflation.

We created the Python package anatomy to facilitate the implementation of the new metrics developed in this paper to better understand the sources of the out-of-sample forecasting accuracy of fitted machine-learning models. In ongoing research, we are exploring strategies for using our new metrics to refine forecasting models over time to potentially improve future out-of-sample performance, so the metrics can serve as both interpretation and development tools for time-series forecasting models.

References

Apley, D. W. and J. Zhu (2020). Visualizing the effects of predictor variables in black box supervised learning models. *Journal of the Royal Statistical Society. Series B (Statistical Methodology) 82*(4), 1059–1086.

Avramov, D., S. Cheng, and L. Metzker (2023). Machine learning versus economic restrictions: Evidence from stock return predictability. *Management Science 69*(5), 2587–2619.

Bils, M. and P. J. Klenow (2004). Some evidence on the importance of sticky prices. *Journal of Political Economy 112*(5), 947–985.

Borup, D., D. E. Rapach, and E. C. M. Schütte (2023). Mixed-frequency machine learning: Nowcasting and backcasting weekly initial claims with daily internet search volume data. *International Journal of Forecasting 39*(3), 1122–1144.

Borup, D. and E. C. M. Schütte (2022). In search of a job: Forecasting employment growth using Google Trends. *Journal of Business & Economic Statistics 40*(1), 186–200.

Breiman, L. (2001). Random forests. *Machine Learning 45*(1), 5–32.

Casalicchio, G., C. Molnar, and B. Bischl (2018). Visualizing the feature importance for black box models. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pp. 655–670.

Castro, J., D. Gómez, and J. Tejada (2009). Polynomial calculation of the Shapley value based on sampling. *Computer and Operations Research 36*(5), 1726–1730.

Chen, H., J. D. Janizek, S. Lundberg, and S.-I. Lee (2020). True to the model or true to the data? Working Paper arXiv:2006.16234v1.

Chen, T. and C. Guestrin (2016). XGBoost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 785–794.

Chinco, A., A. D. Clark-Joseph, and M. Ye (2019). Sparse signals in the cross-section of returns. *Journal of Finance 74*(1), 449–492.

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

Dong, X., Y. Li, D. E. Rapach, and G. Zhou (2022). Anomalies and the expected market return. *Journal of Finance 77*(1), 639–681.

Fisher, A., C. Rudin, and F. Dominici (2019). All models are wrong, but many are useful: Learning a variable's importance by studying an entire class of prediction models simultaneously. *Journal of Machine Learning Research 20*(177), 1–81.

Freyberger, J., A. Neuhierl, and M. Weber (2020). Dissecting characteristics nonparametrically. *Review of Financial Studies 33*(5), 2326–2377.

Friedman, J. H. (2001). Greedy function approximation: A gradient boosting machine. *Annals of Statistics 29*(5), 1189–1232.

Goldstein, A., A. Kapelner, J. Bleich, and E. Pitkin (2015). Peeking inside the black box: Visualizing statistical learning with plots of individual conditional expectation. *Journal of Computational and Graphical Statistics 24*(1), 44–65.

Goulet Coulombe, P. (2022). A neural Phillips curve and a deep output gap. Working Paper arXiv:2202.04146v1.

Goulet Coulombe, P., M. Leroux, D. Stevanovic, and S. Surprenant (2021). Macroeconomic data transformations matter. *International Journal of Forecasting 37*(4), 1338–1354.

Goulet Coulombe, P., M. Leroux, D. Stevanovic, and S. Surprenant (2022). How is machine learning useful for macroeconomic forecasting? *Journal of Applied Econometrics 37*(5), 920–964.

Goyal, A. and I. Welch (2003). Predicting the equity premium with dividend ratios. *Management Science 49*(5), 639–654.

Goyal, A. and I. Welch (2008). A comprehensive look at the empirical performance of equity premium prediction. *Review of Financial Studies 21*(4), 1455–1508.

Greenwell, B. M., B. C. Boehmke, and A. J. McCarthy (2018). A simple and effective model-based variable importance measure. Working Paper arXiv:1805.04755v1.

Gu, S., B. Kelly, and D. Xiu (2020). Empirical asset pricing via machine learning. *Review of Financial Studies 33*(5), 2223–2273.

Hauzenberger, N., F. Huber, and K. Klieber (2023). Real-time inflation forecasting using non-linear dimension reduction techniques. *International Journal of Forecasting 39*(2), 901–921.

Janzing, D., L. Minorics, and P. Blöbaum (2020). Feature relevance quantification in explainable AI: A causal problem. In *23rd International Conference on Artificial Intelligence and Statistics*, pp. 2907–2916.

Joseph, A. (2021). Shapley regressions: A framework for statistical inference on machine learning models. Working Paper arXiv:1903.04209v1.

Kotchoni, R., M. Leroux, and D. Stevanovic (2019). Macroeconomic forecast accuracy in a data-rich environment. *Journal of Applied Econometrics 34*(7), 1050–1072.

Lundberg, S. M. and S.-I. Lee (2017). A unified approach to interpreting model predictions. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, pp. 4768–4777.

McCracken, M. W. and S. Ng (2016). FRED-MD: A monthly database for macroeconomic research. *Journal of Business & Economic Statistics 34*(4), 574–589.

Medeiros, M. C. and E. F. Mendes (2016). ℓ_1-regularization of high-dimensional time-series models with non-Gaussian and heteroskedastic errors. *Journal of Econometrics 191*(1), 255–271.

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

Mitchell, R., J. Cooper, E. Frank, and G. Holmes (2022). Sampling permutations for Shapley value estimation. *Journal of Machine Learning Research 23*(43), 1–46.

Molnar, C. (2023). *Interpretable Machine Learning: A Guide for Making Black Box Models Explainable* (Second ed.). Independently published.

Newey, W. K. and K. D. West (1987). A simple, positive semi-definite, heteroskedasticity and autocorrelation consistent covariance matrix. *Econometrica 55*(3), 703–708.

Pearl, J. (2009). *Causality* (Second ed.). Cambridge: Cambridge University Press.

Ribeiro, M. T., S. Singh, and C. Guestrin (2016). "Why Should I Trust You?" Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 1135–1144.

Schwarz, G. (1978). Estimating the dimension of a model. *Annals of Statistics 6*(2), 461–464.

Shapley, L. S. (1953). A value for n-person games. *Contributions to the Theory of Games 2*(28), 307–317.

Stock, J. H. and M. W. Watson (2002a). Forecasting using principal components from a large number of predictors. *Journal of the American Statistical Association 97*(460), 1167–1179.

Stock, J. H. and M. W. Watson (2002b). Macroeconomic forecasting using diffusion indexes. *Journal of Business & Economic Statistics 20*(2), 147–162.

Štrumbelj, E. and I. Kononenko (2010). An efficient explanation of individual classifications using game theory. *Journal of Machine Learning Research 11*(1), 1–18.

Štrumbelj, E. and I. Kononenko (2014). Explaining prediction models and individual predictions with feature contributions. *Knowledge and Information Systems 41*(1), 647–665.

Sundararajan, M. and A. Najmi (2020). The many Shapley values for model explanation. In *Proceedings of the 37th International Conference on Machine Learning*, pp. 9269–9278.

West, K. D. (1996). Asymptotic inference about predictive ability. *Econometrica 64*(5), 1067–1084.

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