David and Goliath in Inflation Forecasting:

Competing with Institutional Forecasts using a Machine Learning Slingshot

Dominik Hecker*

November 4, 2025

Abstract

This paper introduces a machine learning model as a univariate benchmark for real-time U.S. inflation forecasting. The model uses a Random Forest regression framework and derives all predictive inputs from engineered features. Using signal processing techniques, I extract the features from past inflation dynamics. Evaluated over the volatile inflation period from 2020Q1 to 2024Q4, the model reduces average forecast errors by about 14% relative to the Survey of Professional Forecasters and performs competitively with standard univariate benchmarks, particularly at short- and medium-term horizons. In addition, this paper shows that the Random Forest model provides economically meaningful measures of forecast uncertainty and that the engineered features contribute substantially to predictive accuracy. The results support including univariate machine learning models in forecasting model suites, especially in periods of heightened inflation volatility.

JEL-Codes: C53, E37, C45, E31, C52

Keywords: Real-time Inflation forecasting; Random Forest regression; machine learning; feature engineering; signal processing

*Kiel University, E-mail: hecker@economics.uni-kiel.de. I thank Jan Ewald, Fabio Verona, and Maik Wolters for helpful comments and suggestions. All errors are my own.

1 Introduction

In the realm of inflation forecasting, institutional forecasts have historically played the role of Goliath.[1] Professional forecasters and policy institutions rely on extensive macroeconomic datasets, large-scale structural models, and expert judgment to predict inflation. However, simple univariate time-series models-those that rely solely on historical inflation data-have often been difficult to outperform (Stock and Watson, 2007), as U.S. inflation exhibited strong persistence and relatively stable dynamics during the Great Moderation. The pandemic period, however, introduced large supply and demand shocks that generated persistent and unexpected inflationary pressures, challenging both complex institutional models and traditionally robust univariate benchmarks. Against this backdrop, this paper evaluates a new univariate benchmark that applies statistical learning techniques to extract predictive structure from inflation's own history.

This paper develops a simple machine learning model for real-time inflation forecasting. By design, it restricts the information set to inflation's own history, providing a disciplined test of how much predictive power can be extracted from inflation itself, without relying on large panels of macroeconomic predictors or expert judgment. The model employs a Random Forest regression framework and incorporates features engineered from historical inflation data using signal-processing techniques. The analysis evaluates the model's real-time forecasting performance in comparison to institutional and established univariate benchmarks, explores the model's forecast uncertainty, and examines the relevance of the engineered features for determining the forecasts.

This paper's contributions are threefold:

(1) Real-time forecast performance under fair conditions. The first contribution assesses the real-time forecast performance of the Random Forest (RF) model relative to a set of standard univariate time-series models and an institutional benchmark, namely the Survey of Professional Forecasters (SPF). Under *fair* conditions, all models are evaluated using identical information sets, and for the out-of-sample period from 2020Q1 to 2024Q4. *Empirically*, the RF reduces forecast errors by on average 14% compared to the SPF across horizons. At horizons up to two quarters ahead, it achieves substantial reductions in forecast errors relative to both benchmarks. The longer the horizons, the less pronounced is the forecasting superiority of the RF, while it is still performing competitively

[1] This paper uses the David and Goliath legend purely as a metaphorical framework and without any religious sentiment.

or on par with the benchmarks.

(2) Interpretable forecast uncertainty. The second contribution evaluates the quality and interpretability of forecast uncertainty produces by the Random Forest relative to that of the SPF. Uncertainty for the RF model is obtained directly from the distribution of individual tree forecasts, while SPF uncertainty is proxied by the cross-sectional dispersion among professional forecasters. I compute statistical measures that assess whether forecast uncertainty co-moves with actual inflation volatility. *Empirically*, the RF provides economically meaningful and informative measures of forecast uncertainty. Across forecast horizons, the RF forecast uncertainty correlates positively with realized and forecast error volatility indicating that higher forecast dispersion coincides with greater fundamental and predictive uncertainty, respectively. In terms of all statistical measures, the SPF shows less connection with uncertainty in the economic environment.

(3) Feature importance and interpretability. Finally, the paper examines the drivers behind the Random Forest model's predictive performance. Beyond producing accurate forecasts, Random Forest regressions offer interpretability by quantifying how much each predictor contributes to the model's output. I assess feature relevance using Shapley values, which decompose individual forecasts into additive feature-level contributions and summarize each variable's average marginal impact. This analysis also serves as a robustness check of the feature-engineering approach—if the engineered transformations are meaningful, they should yield consistently positive Shapley contributions across horizons.

Empirically, the engineered features—capturing cyclical, spectral, and energy-based information from inflation's past—consistently enhance predictive performance. Short-term inflation persistence dominates up to three quarters ahead, while medium-frequency cyclical and energy-related components gain importance at longer horizons. This temporal shift in relevance shows that the model relies on economically intuitive signals that evolve with the forecast horizon, confirming that the selected transformations are both statistically useful and economically interpretable.

Taken together, these findings demonstrate that a minimalist, data-driven machine learning model relying solely on inflation's own history can match or even exceed the real-time forecasting performance of institutional benchmarks and standard univariate models. The Random Forest framework provides economically meaningful measures of uncertainty, while the feature-importance analysis confirms the economic validity of the feature-engineering design and its link to the model's predictive power.

Literature. This paper relates to the rapidly growing body of research that applies machine learning (ML) methods to forecast inflation in real time, with the seminal study by Medeiros, Vasconcelos, Veiga, and Zilberman (2021) marking the foundation of this literature. Existing studies differ primarily along two dimensions: (i) the choice of ML algorithm and (ii) the construction and richness of the underlying feature set.

Regarding the forecasting algorithms, most studies compare the predictive capabilities of a wide range of ML approaches. Medeiros et al. (2021) examine an extensive suite of models encompassing regularized linear estimators (such as variants of LASSO and ridge regression), factor-based and ensemble methods (including factor models, bagging, boosting, and model averaging), as well as nonlinear learners like tree-based algorithms and neural networks. Naghi, O'Neill, and Danielova Zaharieva (2024) extend this framework by incorporating advanced tree-based and Bayesian ensemble methods, together with support vector regressions (SVR) and neural networks of varying architectures. Similarly, Araujo and Gaglianone (2023) evaluate a broad set of 50 forecasting methods across 501 macroeconomic time series. Across these comparisons, the RF consistently emerges as one of the most accurate and robust models in terms of mean squared forecast errors. Building on this evidence, the present study focuses exclusively on the RF as a benchmark ML forecaster.

A second major source of heterogeneity in the literature lies in the construction of predictor variables. Most studies employ large macroeconomic datasets that undergo regularization, factor extraction, or shrinkage to reduce dimensionality and multicollinearity. Notably, Medeiros et al. (2021) expand a dataset of 122 macroeconomic indicators by including principal component factors and up to four lags for each variable, resulting in 508 potential predictors. Naghi et al. (2024) further enlarge this framework by incorporating additional variables and extending the analysis to the United Kingdom and Canada.

While recent research increasingly incorporates novel information sources—such as text-based indicators or high-frequency financial variables—this paper adopts a deliberately minimalist approach: The present study relies exclusively on inflation vintages, intentionally excluding external predictors or auxiliary macroeconomic series. Instead, I extract predictive information directly from the inflation process through systematic transformations. This approach represents, to my knowledge, the first explicit application of *feature engineering* in a macroeconomic forecasting context.[2] The combination

[2] A related exception is Verona (2025), who employ wavelet-based decompositions to extract frequency-domain information from macroeconomic time series—an approach conceptually akin to feature engineering, although not explicitly

of feature engineering with a simple machine learning model reflects the idea that forecasting accuracy depends not only on algorithmic sophistication but also on the construction and representation of the underlying data (Verdonck, Baesens, Óskarsdóttir, and vanden Broucke, 2024).

The remainder of the paper is organized as follows. Section 2 outlines the econometric framework, including the Random Forest and feature-engineering approach, while Section 3 details the data and training design. Sections 4–6 present results on forecasting performance, uncertainty, and feature relevance, respectively. Section 7 concludes.

2 Econometric Framework

The baseline forecasting framework relates the h-step-ahead inflation rate to a model-based mapping of the information available at time t,

$$\pi_{t+h} \;=\; \mathbf{M}^{(h)}(\mathbf{X}_t) \;+\; \varepsilon_{t+h}, \qquad t = 1, \ldots, T-h, \quad h = 1, \ldots, H, \tag{1}$$

where inflation is defined as the quarter-over-quarter percentage change in the annualized price level, $\pi_t \equiv 100\,[\ln P_t - \ln P_{t-1}]$. The function $\mathbf{M}^{(h)}(\cdot)$ denotes a horizon-specific forecasting model that generates a prediction of π_{t+h} using the predictor set $\mathbf{X}_t$ observed at time t. The term ε_{t+h} is the corresponding h-step-ahead forecast error.

Forecast accuracy is evaluated using the root mean squared error (RMSE) and mean absolute error (MAE), both computed using strictly out-of-sample forecast errors $\hat{\varepsilon}_{s+h}$,

$$\mathrm{RMSE}^{(h)} = \sqrt{\frac{1}{N_h}\sum_{s} \hat{\varepsilon}^2_{s+h}} \qquad \mathrm{MAE}^{(h)} = \frac{1}{N_h}\sum_{s} \left|\hat{\varepsilon}_{s+h}\right|$$

where the summation is taken over all forecast origins s such that the realized value π_{s+h} is available. The RMSE is the appropriate loss function when the model targets the conditional mean of π_{t+h}, whereas the MAE is the proper choice when the conditional median is the object of prediction. Following Fulton and Hubrich (2021), I report both metrics to provide a robust assessment of model performance.

labeled as such.

2.1 Random Forests

In this application, the baseline $\mathbf{M}^{(h)}$ takes the form of a random forest regression, denoted by $\mathbf{RF}^{(h)}$. Random forests are an ensemble learning method that combines decision trees with bagging, originally introduced by Breiman (2001). Decision trees use a tree structure to identify a variable's future value by imposing sensible splits on its observation values. The tree structure leads to a predicted value at its final leaf nodes. Bagging (Bootstrap Aggregating) involves fitting multiple decision trees on bootstrapped subsets of the original training data. The resulting ensemble of bagged decision trees forms the random forest, which simply averages the individual trees' predictions. While stand-alone decision trees fit the training data well, the bagging step enhances the forecasting performance and prevents overfitting by not only training on randomized data subsets but also by inducing randomness in the choice of variables and splits for each individual tree.

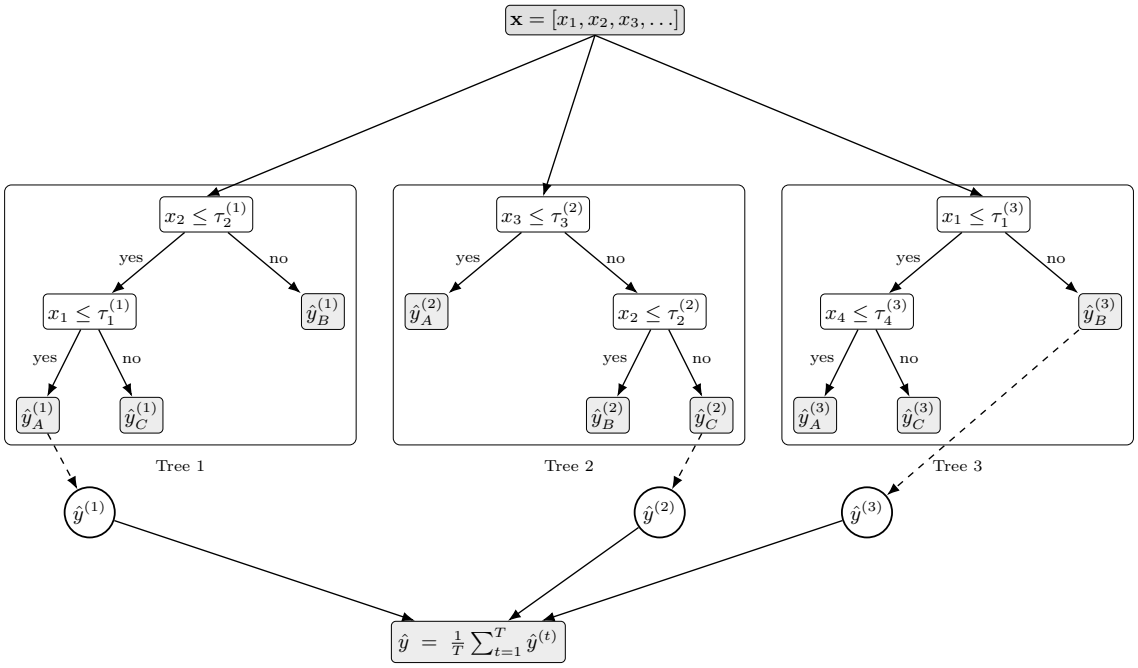

Figure 1: Schematic representation of a Random Forest.

Notes: The feature vector $\mathbf{x} = [x_1, x_2, x_3, \ldots]$ serves as the common input to an ensemble of regression trees (Tree 1–3). Each tree recursively partitions the feature space through binary splits of the form $x_j \leq \tau_j^{(t)}$ until reaching terminal nodes that yield local predictions, e.g., $\hat{y}^{(t)}A$ in Tree 1. The overall forecast is the average across all T trees, $\hat{y} = \frac{1}{T}\sum t = 1^T \hat{y}^{(t)}$.

Figure 1 provides a schematic illustration of how a random forest produces a forecast. The input vector $\mathbf{x}$ enters multiple decision trees, each of which operates on a randomly selected subset of predictors. In practice, hundreds of trees are typically grown, though the figure displays only three for clarity. Within each tree, the data are recursively split according to learned threshold criteria τ_i, producing

branches corresponding to binary decisions (e.g., "yes" or "no"). This hierarchical splitting continues until a terminal leaf node is reached, where the tree assigns a forecast value $\hat{y}^{(i)}$. During training, the algorithm determines both the optimal split criteria and the resulting leaf predictions based on a loss-minimization principle. Once all trees are trained, the random forest aggregates the individual predictions by simple averaging, yielding the final forecast $\hat{y} = \frac{1}{T}\sum_{i=1}^{T} \hat{y}^{(i)}$.

Random forests are the only machine learning method used in this paper for four main reasons. First, they have consistently demonstrated competitive performance in macroeconomic forecasting compared to more complex models, such as neural networks and boosting variants (Medeiros et al., 2021; Naghi et al., 2024). Second, they can be efficiently trained, cross-validated, and deployed on standard computing hardware, eliminating the need for high-performance clusters or GPU acceleration. This enhances both accessibility and reproducibility.

Third, random forests provide a relatively high degree of interpretability. Unlike deep learning architectures, they allow researchers to trace predictions back to their underlying structure and assess the relative importance of different predictors (see Section 6). This transparency is particularly valuable in economic forecasting, where understanding the factors driving predictive performance can be just as important as achieving accuracy.

Finally, random forests have inspired dedicated extensions designed for economic forecasting applications, such as hedged random forests (Beck, Kozbur, and Wolf, 2024). These extensions optimize forecast combination weights rather than relying on uniform averaging. They have proven effective in recent inflation forecasting applications (Beck and Wolf, 2025), highlighting the adaptability and ongoing relevance of random forests in this domain.

2.2 Benchmark models

To benchmark the forecasting performance of the Random Forest models, I estimate a set of standard univariate time-series models. For each forecast horizon h, a separate model is estimated using only information available at the forecast origin t, following a direct forecasting approach. The benchmark set includes a random walk (RW), an autoregressive (AR) model, an autoregressive moving average (ARMA) model, and an autoregressive integrated moving average (ARIMA) model.

The choice of benchmarks follows the canonical forecasting setup in Medeiros et al. (2021) and the

univariate tradition in Stock and Watson (2007). The ARMA-type models require selecting hyperparameters (p, q, D), where p denotes the autoregressive lag order, q the moving-average lag order, and D the degree of differencing. These are chosen by minimizing the Bayesian Information Criterion (BIC).

Hyperparameter selection is initially conducted using the final inflation series introduced in Section 3. For comparability with the Random Forest setup, all benchmark models are subsequently re-estimated in a fully real-time, expanding-window fashion: at each forecast origin t, estimation is based exclusively on inflation observations that would have been available at that time.

RW (random walk): as a naïve and deliberately "light-learning" but very well established benchmark, the forecast is given by the rolling four-quarter average of past inflation for any $h \geq 1$,

$$\widehat{\pi}^{\mathrm{RW}}_{t+h|t} \;=\; \frac{1}{4}\sum_{j=0}^{3} \pi_{t-j}. \tag{2}$$

AR(p): as a standard benchmark capturing autoregressive persistence, I estimate for each horizon h a direct autoregressive model,

$$\pi_{s+h} \;=\; \alpha^{(h)} + \sum_{j=1}^{p} \phi_j^{(h)}\, \pi_{s-j+1} + u^{(h)}_{s+h}, \qquad s = p, \ldots, t-h. \tag{3}$$

This specification defines a direct autoregressive forecasting model of order p for horizon h. The parameter $\alpha^{(h)}$ is a horizon-specific intercept, and $\phi_j^{(h)}$ are the corresponding autoregressive coefficients. The term $u^{(h)}_{s+h}$ represents the forecast error at horizon h, assumed to be a mean-zero innovation.

The h-step-ahead forecast then is

$$\widehat{\pi}^{\mathrm{AR}}_{t+h|t} \;=\; \hat{\alpha}^{(h)} + \sum_{j=1}^{\hat{p}^{(h)}} \hat{\phi}_j^{(h)}\, \pi_{t-j+1}. \tag{4}$$

ARMA(p, q): allowing for moving-average dynamics under stationarity, I estimate a direct autoregressive moving average model of order (p, q) for forecast horizon h

$$\pi_{s+h} \;=\; \alpha^{(h)} + \sum_{j=1}^{p} \phi_j^{(h)}\, \pi_{s-j+1} + u^{(h)}_{s+h} + \sum_{k=1}^{q} \theta_k^{(h)}\, u^{(h)}_{s-k+h}, \qquad s = \max(p, q), \ldots, t-h, \tag{5}$$

The moving average component is captured by coefficients $\theta_k^{(h)}$ applied to past forecast errors $u_{s-k+h}^{(h)}$ for $k = 1, \ldots, q$. The forecast then is

$$\widehat{\pi}_{t+h|t}^{\text{ARMA}} \;=\; \hat{\alpha}^{(h)} + \sum_{j=1}^{\hat{p}^{(h)}} \hat{\phi}_j^{(h)}\, \pi_{t-j+1} + \sum_{k=1}^{\hat{q}^{(h)}} \hat{\theta}_k^{(h)}\, \hat{u}_{t-k+h}^{(h)}. \tag{6}$$

ARIMA(p, d, q): to allow for differenced dynamics, I extend the ARMA specification by integration through differencing and estimate

$$(1-L)^{d^{(h)}} \pi_{s+h} \;=\; \alpha^{(h)} + \sum_{j=1}^{p} \phi_j^{(h)} (1-L)^{d^{(h)}} \pi_{s-j+1} + u_{s+h}^{(h)} + \sum_{k=1}^{q} \theta_k^{(h)} u_{s-k+h}^{(h)}, \tag{7}$$

where the operator $(1-L)^{d^{(h)}}$ applies $d^{(h)}$-order differencing to the inflation series, where L denotes the lag operator ($L\pi_t = \pi_{t-1}$). The dependent variable $(1-L)^{d^{(h)}} \pi_{s+h}$ is therefore the $d^{(h)}$-times differenced h-step-ahead inflation rate. The conditional mean forecast $\widehat{\pi}_{t+h|t}^{\text{ARIMA}}$ is obtained from the estimated model and, if $d^{(h)} = 1$, reintegrated back to the level of π.

2.3 Feature Engineering

The machine learning literature distinguishes between the *data set* and the *feature set*. While the data set comprises raw observations, the feature set consists of transformed variables-called *features*-that serve as inputs to the forecasting model. The process of transforming the raw data set into a more informative feature set is called *feature engineering*. In principle, one could provide a machine learning model with only raw data and rely on it to extract relevant relationships internally. However, empirical studies suggest that omitting feature engineering often leads to inferior forecasting performance (see e.g., Guyon and Elisseeff (2006); Oliveira and Torgo (2015)).

Although the terminology *feature engineering* may seem unfamiliar in the context of economic forecasting, the practice itself is not new. Economists have long engineered features—often implicitly—by constructing transformations such as rolling means, variances, growth rates, or detrended components to summarize dynamics and reveal latent structure. Such transformations enrich the information set by highlighting persistent trends, cyclical movements, or volatility episodes. In practice, feature engineering is often ad hoc and driven by domain expertise. In this paper, feature engineering is made explicit and systematic, serving as a structured bridge between economic time series behavior and the

learning capabilities of machine learning models.

The random forest models $\mathbf{RF}^{(h)}$ developed in this paper rely exclusively on features engineered from past inflation dynamics. This approach is deliberately orthogonal to the prevailing practice in the machine learning and economic forecasting literature, where models are typically trained on large-scale datasets with dozens or even hundreds of macroeconomic and financial predictors, often embedded within increasingly complex architectures such as deep neural networks or hybrid ensembles. Instead, the goal here is to extract the maximum predictive content from the univariate inflation series alone, using systematic feature engineering to compensate for the absence of additional covariates.

Accordingly, this paper does not seek to demonstrate incremental gains from ever-growing pools of predictors or from the adoption of more computationally intensive learning architectures. Rather, it asks a more fundamental question: *how much forecasting performance can be recovered by combining classical time-series intuition with modern machine learning techniques in a purely univariate setting?* In this sense, the proposed model serves as a machine-learning-based univariate benchmark, designed to test the upper bound of what inflation's own past can reveal when processed through a carefully constructed feature space and a flexible, nonlinear prediction algorithm.

Since the relationship between these engineered features and future inflation is unknown a priori, machine learning provides a natural framework to detect and exploit complex, possibly non-linear dependencies. Random forests, in particular, offer an additional interpretive advantage: after training, they assign relative importance weights to each feature, quantifying its contribution to predictive accuracy. Unlike automated feature engineering pipelines that generate vast numbers of candidate variables—often leading to high-dimensional feature spaces and overfitting risks (Cerqueira, Moniz, and Soares, 2024)—the proposed approach keeps the feature set compact and interpretable. Positive feature importance weights indicate that a variable materially improves forecast precision, while near-zero or negative weights suggest redundancy or noise. Section 6 discusses these importance weights in detail and validates the relevance of the engineered features for capturing the dynamics of U.S. inflation.

The choice of transformations in this paper is grounded in signal processing, which has gained increasing prominence as a powerful framework for time series forecasting. As noted by Praveen, Dekka, Sai, Chennamsetty, and Chinta (2025), "integrating signal processing techniques into forecasting models can improve their accuracy, adaptability, and robustness to various market conditions and data

characteristics." A growing body of empirical work supports this view, showing that decompositional and spectral representations can extract latent structure that is otherwise difficult to detect in raw time-domain data. For example, De Brabandere, Robberechts, Op De Beéck, and Davis (2019) employ energy measures, energy ratios, Fast Fourier Transforms (FFT), power spectral density, and related spectral indicators to classify time series dynamics, while Shu and Gao (2020) and others demonstrate that Empirical Mode Decomposition (EMD) serves as an effective preprocessing step for enhancing predictive performance in neural network-based stock price forecasting. Consistent with this literature, I assume that inflation dynamics can be decomposed into distinct components—such as trend, cycles, and volatility patterns—which provide complementary predictive information. By systematically engineering features that capture these dimensions, the forecasting model can access a richer signal representation than is available from raw inflation levels alone. Inflation is a process that combines persistent components, cyclical fluctuations, and occasional shock-driven surges—making it particularly well-suited to decomposition-based feature engineering.

Figure 2 illustrates the feature engineering pipeline applied to past inflation data. Starting from a rolling window of recent inflation observations (shaded in gray), the signal is decomposed into intrinsic mode functions (IMFs) using Empirical Mode Decomposition (EMD), transformed using the Hilbert–Huang Transform (HHT) to extract instantaneous amplitudes and phases, analyzed via the Short-Time Fourier Transform (STFT) to quantify frequency-specific band powers, and evaluated using the Teager–Kaiser Energy Operator (TKEO) to capture localized energy or volatility. These complementary representations summarize distinct aspects of the inflation signal's dynamics. Statistical summaries derived from each transformation form the engineered feature vector $\mathbf{X}_\omega$, which serves as part of the input to the machine learning forecasting model.

I split the full predictor vector into three components:

$$\mathbf{X_t} = (\pi_t, \pi_{t-1}, \mathbf{X}_\omega)\,,$$

where (π_t, π_{t-1}) denote the last two inflation lags. Since inflation exhibits strong persistence, autoregressive dynamics contain useful predictive content, justifying their inclusion. The engineered feature vector $\mathbf{X}_\omega$ is constructed from a rolling window of $\omega = 20$ past observations $\pi_t, \ldots, \pi_{t-\omega}$ and consists

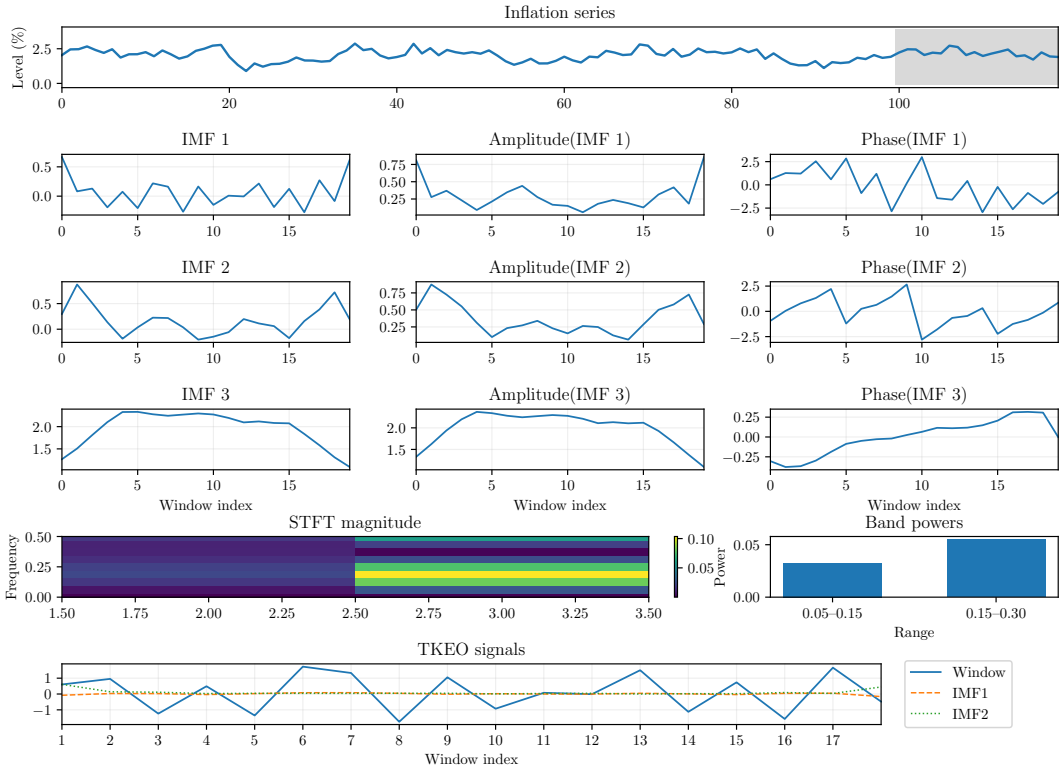

Figure 2: Schematic representation of feature engineering.

Notes: A rolling window of the inflation series (gray-shaded area) is decomposed into intrinsic mode functions (IMF 1–3) using Empirical Mode Decomposition (EMD). From these components, the Hilbert–Huang Transform (HHT) extracts instantaneous amplitudes and phases, while the Short-Time Fourier Transform (STFT) computes average band powers over selected frequency ranges (0.05–0.15 and 0.15–0.30 cycles per quarter). The Teager–Kaiser Energy Operator (TKEO) captures localized energy and volatility dynamics. Together, these transformations generate the engineered feature vector $\mathbf{X}_\omega$ used as input to the Random Forest forecasting model.

of three sets of transformations derived from signal processing techniques:

$$\mathbf{X}_\omega = \begin{bmatrix} \mathcal{F}_{\text{HHT}}(\pi_w) = [A(\text{IMF}_1), \Phi(\text{IMF}_1), A(\text{IMF}_2), \Phi(\text{IMF}_2), A(\text{IMF}_3), \Phi(\text{IMF}_3)], \\ \mathcal{F}_{\text{STFT}}(\pi_w) = [P_{0.05-0.15}, P_{0.15-0.30}], \\ \mathcal{F}_{\text{TKEO}}(\pi_w) = [E(\pi_w), E(\text{IMF}_1), E(\text{IMF}_2)] \end{bmatrix}. \tag{8}$$

The first feature set, $\mathcal{F}_{\text{HHT}}(\pi_w)$, is derived from the first three IMFs obtained via EMD, which decompose the window into oscillatory components ordered from high to low frequency. The Hilbert–Huang Transform then computes instantaneous amplitudes A and phases Φ, which capture the strength and position of each cycle in time.

The second feature set, $\mathcal{F}_{\text{STFT}}(\pi_w)$, captures how short- and medium-run cycles contribute to the

windowed signal. Applying the STFT produces a spectrogram, from which average band powers are extracted over economically relevant frequency intervals. I retain only the medium- and high-frequency ranges, (0.05–0.15) and (0.15–0.30) cycles per quarter, as the low-frequency bandpower does not inform short- to medium-horizon inflation forecasts.

The final feature set, $\mathcal{F}_{\text{TKEO}}(\pi_w)$, evaluates the signal's instantaneous energy using the TKEO operator, which is sensitive to local bursts and shocks. I compute mean energy values for the original window and its first two IMFs to detect volatility or sudden changes in inflation dynamics.

With the feature vector $\mathbf{X}_\omega$ defined, I now turn to how these inputs are used within the Random Forest regression framework for inflation forecasting.

3 Training the Random Forest Model

3.1 Inflation Data and Forecast Benchmarks

Final inflation is measured as the quarter-over-quarter percentage change in the personal consumption expenditure (PCE) price index, published by the Bureau of Economic Analysis under the code DPCERG.[3] This chain-type index is the Federal Open Market Committee's targeted inflation measure and serves as the policy-relevant benchmark (Fulton and Hubrich, 2021). The series is expressed at an annualized rate and the June 2025 release provides the reference for forecast evaluation.

For real-time forecasting, I use vintage versions of PCE inflation from the ArchivaL Federal Reserve Economic Data (ALFRED) as of June 2025.[4] Multiple releases within a quarter are averaged to create a single observation, and growth rates are computed as for the final series. The vintage sample period ranges from 2000Q3 to 2025Q1, yielding 98 quarterly observations. Given that the feature engineering described in Section 2.3 relies on a rolling-window of inflation, with a window size of five years, i.e. 20 quarters, the forecast evaluation sample starts in 2005Q1 to allow for lagged predictors. This reduces the sample to 80 quarterly observations.

To compare the forecasting performance of the RF relative to other models and institutional benchmarks, the natural choices to do so are the Survey of Professional Forecasters and the Tealbook pro-

[3]The data is available through the Federal Reserve Economic Data under the following link: https://fred.stlouisfed.org/series/PCEPI.

[4]The data is obtained from ALFRED under the following link: https://alfred.stlouisfed.org/series?seid=PCEPI.

jections. The Survey of Professional Forecasters is hosted by the Federal Reserve Bank of Philadelphia und provides quarterly PCE inflation forecasts at horizons of one to six quarters ahead.[5] Forecasts are expressed in annualized growth rates and correspond to the end of each quarter. I align SPF release dates with the real-time targets and restrict the sample to overlapping periods. This produces a benchmark series that is directly comparable to the Random Forest forecasts.

The Tealbook projections are published with a lag of five years. As the focus of this paper is the volatility inflation period after 2020Q1, Tealbook projections cannot serve as an additional institutional benchmark. After having introduced data sources and forecast benchmarks, I now specify how I train the random forest model.

3.2 Training and Testing Sets

I adopt an expanding-window training approach. For a given quarter t, the model is trained on all observations up to t, using the engineered features derived from the corresponding rolling window. The model is then used to predict inflation at horizon $t + h$. In the next period, once new inflation data become available, the training set expands to include this information and the model is retrained. This recursive re-estimation process reflects how a real-time forecasting system would be updated with each new data release.

A key objective of this paper is to evaluate forecasting performance under conditions that closely replicate real-time forecasting. This is particularly important in the period following 2020, when inflation dynamics exhibited extreme uncertainty. After the sharp contraction in 2020Q1, it was unclear whether inflation would remain subdued or reaccelerate. Similarly, during the post-2021 inflation surge, policymakers and economists debated whether the spike was temporary or persistent. Any learning algorithm that is trained on future observations belonging to these shock episodes would benefit from knowledge unavailable to real-time forecasters, thereby introducing a so-called "look-ahead bias" or "data leakage". To ensure that the Random Forest faces the same informational constraints as professional forecasters, the model is trained only on data available up to each forecast origin.

Unlike traditional machine learning workflows that rely on distinct training, validation, and testing

[5]The data for the SPF forecasts can be obtained from the Federal Reserve Bank of Philadelphia's website under https://www.philadelphiafed.org/surveys-and-data/real-time-data-research/survey-of-professional-forecasters.

splits, Random Forests internally perform validation through bootstrapped sampling, which makes an explicit validation set unnecessary. However, a genuine testing set is preserved by evaluating forecasts only on future observations not included in the training set. This setup ensures that forecast performance is assessed truly out-of-sample and that comparisons with institutional benchmarks, such as the SPF, are conducted under equivalent information conditions.

Each forecast horizon $h = 1, \ldots, H$ is modeled using a separate Random Forest regression, following a direct forecasting strategy as in Beck et al. (2024). This specification allows the model to learn horizon-specific relationships between engineered features and future inflation, recognizing that the predictive relevance of short-run fluctuations, medium-frequency dynamics, or local volatility may differ depending on how far ahead the forecast is made.

To ensure that each Random Forest model learns a generalizable relationship between the engineered features and future inflation, the training data are constructed from randomly drawn periods rather than sequential time blocks. This prevents the model from simply memorizing the temporal structure of the data. In addition, a small amount of noise is added to each feature to avoid overfitting and to reinforce robustness. Both steps ensure that the RF model receives only information that would have been available in real time and cannot exploit unintended time-series regularities. For each cut-off date T_c, the training dataset of size $N = 100$ is generated as follows:

1. Randomly draw a period index t_i from the real-time sample, excluding the most recent $\omega = 20$ observations prior to the cut-off date T_c.

2. Construct a window of real-time PCE inflation data ending at t_i and engineer features using the procedure described in Section 2.3.

3. Augment the engineered feature vector with the last two available lags of inflation, π_{t_i} and π_{t_i-1}.

4. Add noise from a normal distribution with variance 1×10^{-4} to each feature to enhance robustness.

5. Pair the resulting feature vector X_i with the corresponding final revised value of PCE inflation h periods ahead, denoted by y_i, to form a training observation (X_i, y_i).

With the training samples (X_i, y_i) constructed for each forecast horizon and cut-off date, I now turn to the estimation of the Random Forest models and their diagnostic evaluation.

3.3 Training, Tuning, and Diagnostics of Random Forest Models

The Random Forest models are estimated on the training data set. Another advantage of RF models is that they do not require the user to manually specify hyperparameters. Instead, hyperparameters are automatically optimized during training, in contrast to neural networks, where tuning can be an iterative complex and computationally intensive process. In this application, two hyperparameters are subject to optimization: the minimum leaf size and the number of learners. The minimum leaf size controls the minimum number of observations contained in each terminal node of a decision tree, influencing the granularity of the splitting structure. The number of learners determines the size of the ensemble, with larger forests typically improving predictive accuracy at the cost of increased computational time.

Hyperparameter optimization is conducted during estimation using 5-fold cross-validation. For each candidate combination of hyperparameters, the training sample is partitioned into five folds. The model is then estimated five times, each time using four folds for training and one fold for validation, rotating the validation fold in each iteration. The cross-validation root mean squared error (RMSE) is computed for each round and averaged across folds. The hyperparameter combination that minimizes the average RMSE is selected.

Figure 3 reports the Out-of-Bag (OOB) mean squared forecast error as a function of the number of decision trees for the most recent information vintage. Across all horizons, the OOB error declines sharply as more trees are added, before stabilizing at a plateau. The vertical red line indicates the number of trees minimizing the OOB error, beyond which adding further learners yields negligible gains in predictive performance. While most horizons exhibit smooth convergence, the model for $h = 5$ displays some initial volatility in OOB error before settling, suggesting greater variability in feature relevance at this horizon. Overall, the convergence patterns confirm that the RF models reach stable prediction accuracy after a sufficiently large number of learners.

To verify that the trained models rely on economically meaningful inputs rather than noise, I later analyze the stability and relevance of feature importance rankings across horizons in Section 6. Before turning to this interpretability analysis, however, it is essential to first establish whether the Random Forest models deliver competitive forecasting accuracy. The next section therefore evaluates their predictive performance relative to traditional univariate benchmarks and institutional forecasts.

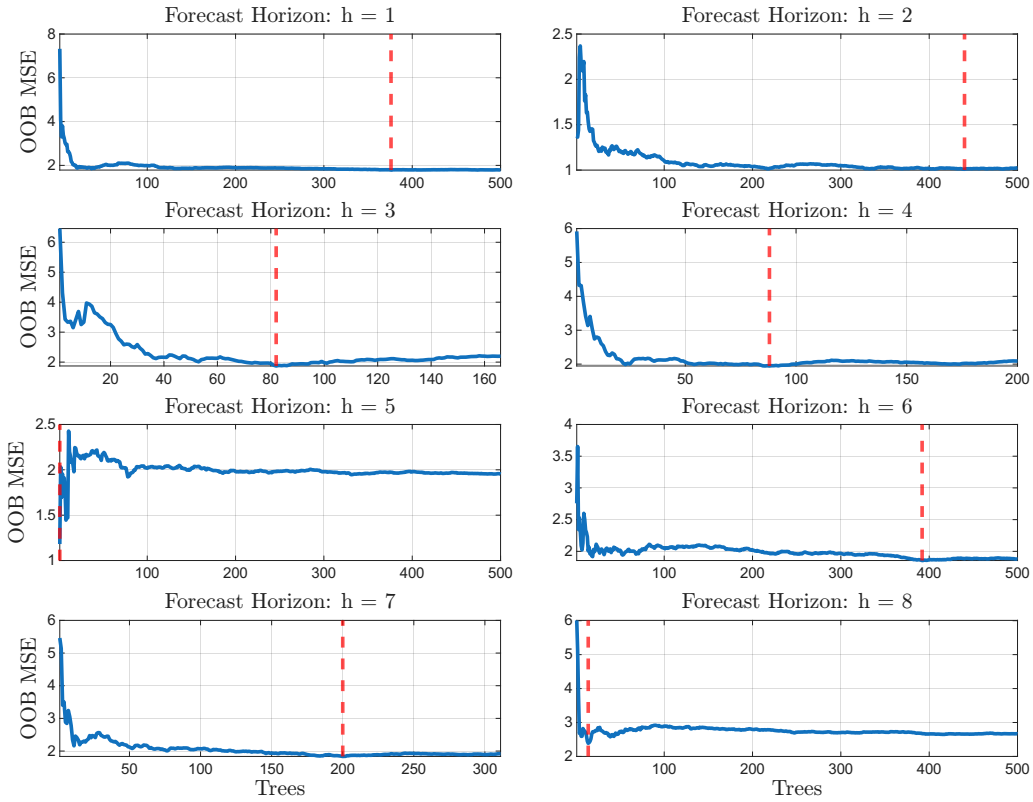

Figure 3: Out-of-Bag (OOB) mean squared forecast error as a function of the number of decision trees for each forecast horizon ($h = 1, \ldots, 8$)

Notes: Each panel shows how prediction error evolves as more trees are added to the Random Forest ensemble. The OOB error declines rapidly with additional trees before stabilizing at a plateau, indicating convergence of forecast accuracy. Vertical red lines mark the number of trees minimizing the OOB error, beyond which further learners yield negligible improvements.

4 Forecasting Performance

This section evaluates the forecasting performance of the RF model relative to both univariate time-series benchmarks and institutional forecasts. The out-of-sample period begins in 2020Q1 and ends in 2024Q4.

Table 1 summarizes the model performance by computing summary statistics that give a first impression of how the RF model's forecasting abilities compare to those of univariate and institutional benchmarks. The average RMSE and MAE values indicate that the AR model performs best overall, achieving the lowest average forecast errors across horizons (RMSE = 1.95, MAE = 1.36). It also records the highest number of horizon-specific wins (3 for RMSE and 4 for MAE). The Random Forest model ranks closely behind (avg. RMSE = 2.14, MAE = 1.52), outperforming both institutional (SPF) and ARIMA-type models in terms of average accuracy, and achieving three horizon wins in both RMSE and MAE. While the SPF exhibits relatively stable performance with the lowest disper-

sion (std. RMSE = 0.11), it consistently remains outperformed by RF and AR in terms of absolute accuracy. The Random Walk model provides moderate results, with one win in each metric, while ARMA and especially ARIMA tend to show weaker average performance, with ARIMA ranking last in both average RMSE and MAE and exhibiting higher variability. Overall, the results confirm the strong and robust performance of RF relative to institutional and ARIMA-type benchmarks, though the parsimonious AR model remains a competitive and consistent univariate baseline.

Table 1: Summary statistics of model performance across horizons (OOS period)

	Average		Maximum		Minimum		# Horizons Best		Std. Dev.	
Model	RMSE	MAE	RMSE	MAE	RMSE	MAE	RMSE	MAE	RMSE	MAE
RF	2.14	1.52	2.99	2.58	1.21	1.05	3	3	0.70	0.49
SPF	2.48	1.74	2.56	1.79	2.22	1.66	0	0	0.11	0.04
RW	2.15	1.55	2.51	1.81	1.67	1.27	1	1	0.32	0.20
AR	1.95	1.36	2.14	1.47	1.69	1.23	3	4	0.18	0.08
ARMA	2.10	1.49	2.34	1.60	1.88	1.40	0	0	0.17	0.07
ARIMA	2.35	1.66	2.68	1.84	1.83	1.29	1	0	0.30	0.20

Note: "# Horizons Best" counts the number of horizons (out of 8) for which the model achieves the lowest RMSE or MAE (ties allowed). All statistics are computed over the out-of-sample period.

Building on this first impression, forecast accuracy is compared across horizons using pairwise RMSE and MAE ratios, where each entry reports the relative error of the row model to that of the column model. Ratios below unity therefore indicate that the row model is more accurate. In addition, I formally assess statistical differences in predictive accuracy using the Diebold–Mariano test (Diebold and Mariano, 2002), incorporating the Harvey–Leybourne–Newbold (HLN) small-sample adjustment (Harvey, Leybourne, and Newbold, 1997) and Newey–West autocorrelation-robust variance estimation (Newey and West, 1987). The RMSE ratios are displayed in Table 2.

Overall, the results indicate strong predictive performance of the RF model, particularly at short- and medium-term horizons. Across most horizons up to six quarters ahead, the RF model typically outperforms the institutional benchmark provided by the Survey of Professional Forecasters, often by substantial margins. In several cases, the RF cuts the forecasting error by more than 50% compared with the SPF. In comparison with univariate benchmarks such as the random walk, AR, ARMA, and ARIMA models, the RF generally performs competitively or better, especially in the short run. While performance differences become smaller or more mixed at longer horizons (seven to eight quarters ahead), the RF remains broadly comparable to or better than the majority of univariate alternatives. Nonetheless, given the limited size of the out-of-sample evaluation window, statistical significance is

difficult to establish consistently.

Turning to specific horizons, the results are especially favorable in the short run. At $h = 1$, the RF nearly halves the error of the SPF benchmark and performs on par with the random walk and AR models, while reducing the RMSE by up to 13% relative to ARIMA. Averaged across all benchmarks, the reduction is approximately 15%. At $h = 2$, the RF again markedly outperforms the SPF with a reduction of around 65% in RMSE, although it is around 7% worse than the random walk. Still, it improves on AR by roughly 4%, ARMA by 25%, and ARIMA by 56%, corresponding to an average reduction of around 29% across all benchmarks. These results highlight the particular strength of the RF at short forecast horizons.

At medium horizons, performance remains favorable. At $h = 3$, the RF reduces the error by 30% compared to the SPF, though it slightly underperforms the AR model by about 6%. On average, the RMSE declines by about 9%. At $h = 4$, the RF still outperforms all benchmarks, cutting the SPF error by roughly 34% and the AR by around 2%. The reduction relative to ARMA reaches 25% and is statistically significant.

At $h = 5$, the RF shows clear gains relative to SPF, ARMA, and ARIMA (with ratios around 0.5), resulting in an average reduction of around 33%. At $h = 6$, the reductions are even more pronounced: the RF lowers the RMSE by approximately two thirds when compared with SPF and ARIMA, and remains superior to RW, AR, and ARMA. The average reduction at this horizon is around 38%. At longer horizons, a more nuanced picture emerges.

At $h = 7$, performance becomes mixed: the RF outperforms SPF by around 32% but exhibits substantially larger forecast errors than ARIMA (by nearly 88%). On average, the RF produces a 17% increase in RMSE at this horizon. At the longest horizon of $h = 8$, performance is again mixed. The average RMSE reduction across benchmarks is around 5%, with the RF being roughly on par with SPF, slightly worse than AR, but outperforming ARIMA by more than 50%.

Overall, the results indicate that while forecasting accuracy deteriorates uniformly across models at longer horizons, the RF remains competitive and continues to perform well in relative terms against more complex univariate benchmarks such as ARIMA.

Although the Random Forest achieves economically large reductions in forecast errors—sometimes exceeding 50% relative to the SPF—these improvements are not always statistically significant according

Table 2: RMSE Ratios of Random Forest (RF) Relative to Benchmark Models

Benchmark Model	$h=1$	$h=2$	$h=3$	$h=4$	$h=5$	$h=6$	$h=7$	$h=8$
SPF	0.47	0.35	0.70	0.66	0.50	0.32	0.68	0.99
RW	1.00	1.07	0.94	0.88	0.93	0.81	1.05	1.05
AR	0.98	0.96	1.06	0.98	1.01	0.93	1.17	1.14
ARMA	0.94	0.75	0.98	**0.75**	0.47	0.73	1.05	1.02
ARIMA	0.87	0.44	0.88	0.59	0.42	0.32	1.88	0.47

Notes: Each entry reports the ratio of the root mean squared error (RMSE) of the Random Forest (RF) model to that of the benchmark model, i.e. $\text{Ratio}_{\text{RF},j} = \frac{\text{RMSE}_{\text{RF}}}{\text{RMSE}_j}$. Values below one indicate that the RF outperforms the benchmark (lower forecast error), while values above one indicate inferior performance. Bold numbers denote statistically significant differences in predictive accuracy according to the Diebold–Mariano test at the 5% level. Horizon h refers to the number of quarters ahead being forecasted.

to the Diebold–Mariano test. This reflects the limited length and volatility of the out-of-sample evaluation period, which reduces the power of formal significance tests even in the presence of substantial economic gains.

Table 3: MAE Ratios of Random Forest (RF) Relative to Benchmark Models

Benchmark Model	$h=1$	$h=2$	$h=3$	$h=4$	$h=5$	$h=6$	$h=7$	$h=8$
SPF	0.54	0.47	0.59	0.61	0.51	0.43	0.79	0.81
RW	1.01	1.08	0.89	0.86	0.82	0.90	1.31	0.87
AR	1.00	0.90	0.97	0.97	0.83	0.93	1.54	0.99
ARMA	0.90	0.69	0.92	0.72	0.41	0.71	1.42	0.89
ARIMA	1.02	0.52	0.80	0.49	0.51	0.35	2.17	0.60

Notes: Each entry reports the ratio of the mean absolute error (MAE) of the Random Forest (RF) model to that of the benchmark model, i.e. $\text{Ratio}_{\text{RF},j} = \frac{\text{MAE}_{\text{RF}}}{\text{MAE}_j}$. Values below one indicate that the RF achieves lower forecast errors than the benchmark (better performance), while values above one indicate higher forecast errors. Horizon h refers to the number of quarters ahead being forecasted. Bold numbers can denote statistically significant differences in predictive accuracy according to the Diebold–Mariano test at the 5% level.

The MAE ratios reported in Table 3 confirm the same qualitative patterns observed under the RMSE metric. In particular, the RF model delivers sizeable performance gains at short horizons, cutting MAE relative to the SPF by roughly 45–55% at $h = 1$ and $h = 2$, while remaining broadly competitive with the univariate benchmarks, especially AR and RW. At medium-term horizons ($h = 3$ to $h = 6$), the RF continues to outperform SPF and maintains lower or comparable MAE levels than the univariate models, often achieving 20–40% reductions relative to ARMA and ARIMA. At longer horizons ($h = 7$ and $h = 8$), performance becomes more mixed: although the RF still tends to outperform SPF, it occasionally underperforms relative to ARIMA and, to a lesser extent, RW and AR. Nonetheless, even at these horizons the RF frequently remains within a competitive range of the strongest univariate benchmarks. Overall, the MAE-based comparisons reinforce the conclusion from the RMSE analysis that the RF model exhibits robust short- and medium-horizon advantages, with only limited and

horizon-specific deterioration in relative accuracy at longer forecast horizons.

5 Forecast Uncertainty

A key criterion for evaluating a forecasting model is not only how accurately it predicts future outcomes, but also whether it provides meaningful insights into the uncertainty surrounding those predictions. In the context of inflation, such uncertainty has important economic implications. In his Nobel lecture, Friedman (1977) argued that higher inflation tends to increase inflation uncertainty, which in turn impairs the price system's ability to allocate resources efficiently, thereby generating economic inefficiency and lower output growth. This theoretical link suggests that reliable inflation forecasts should reflect underlying economic uncertainty—models whose forecast dispersion responds to shifts in inflation volatility are likely capturing economically relevant information.

In this section, I assess whether the Random Forest model produces economically meaningful uncertainty signals by relating its internal forecast dispersion to observable forms of inflation volatility. Unlike traditional univariate benchmarks such as the Random Walk or ARIMA models, which provide only conditional mean forecasts and assume homoskedastic residuals, the Random Forest model yields a full predictive distribution. The dispersion of tree-level forecasts offers an intrinsic measure of forecast uncertainty.

Specifically, I compare the spread of tree-level forecasts to (i) realized inflation volatility and (ii) the volatility of the model's own forecast errors, thereby quantifying both the calibration and responsiveness of the RF's uncertainty estimates in the spirit of Naghi et al. (2024). For each forecast horizon, I compute the interquartile range (IQR) across the ensemble of tree forecasts, defined as $IQR_t^{RF} = \hat{y}_{t,0.75}^{RF} - \hat{y}_{t,0.25}^{RF}$. The IQR summarizes the middle 50% of the forecast distribution and serves as a standard measure of forecast uncertainty in both professional surveys and empirical forecasting studies (e.g., Abel, Rich, Song, and Tracy, 2016; Clements, Rich, and Tracy, 2025).

To evaluate these relationships, I compute a set of complementary statistics that capture three aspects of forecast uncertainty: its level, calibration, and responsiveness. First, *uncertainty levels* are summarized by the coverage rate of the IQR, i.e., the share of realized inflation outcomes falling within the middle 50% of the RF's forecast distribution—values near 50% imply well-calibrated uncertainty bands. Second, *calibration* is assessed through the average ratios of the IQR to realized inflation

volatility and to forecast error volatility, indicating whether the model's ex-ante uncertainty scales appropriately with both fundamental and predictive variability. Finally, *responsiveness* is measured by the correlations between the IQR and the two volatility metrics, which reveal whether periods of higher inflation volatility are associated with wider forecast dispersion.

Table 4: Forecast Uncertainty Metrics for the Random Forest Model

Measure	Definition	h=1	h=2	h=3	h=4	h=5	h=6	h=7	h=8
Uncertainty levels									
Coverage	$\frac{1}{T}\sum_{t=1}^{T}\mathbb{I}\left[\hat{\pi}_{t,0.25}^{RF} \leq \pi_t \leq \hat{\pi}_{t,0.75}^{RF}\right]$	0.70	0.78	0.79	0.66	0.82	0.79	0.78	0.84
IQR_t	$\hat{\pi}_{t,0.75}^{RF} - \hat{\pi}_{t,0.25}^{RF}$	1.18	1.42	1.16	0.97	1.38	1.20	1.28	1.66
Calibration									
R_1	$\frac{1}{T}\sum_{t=1}^{T}\frac{IQR_t^{RF}}{\sigma_t^{\text{real}}}$	0.40	0.43	0.62	0.48	0.52	0.55	0.55	0.53
R_2	$\frac{1}{T}\sum_{t=1}^{T}\frac{IQR_t^{RF}}{\sigma_t^{\text{FE},RF}}$	0.54	0.64	0.86	0.65	0.89	0.77	0.80	0.80
Responsiveness									
ρ_1	$\text{corr}(IQR_t^{RF}, \sigma_t^{real})$	0.32	0.18	0.44	0.27	0.27	0.39	0.25	0.26
ρ_2	$\text{corr}(IQR_t^{RF}, \sigma_t^{FE,RF})$	0.37	0.22	0.47	0.22	0.27	0.51	0.37	0.39

Notes: The table reports forecast uncertainty measures for the Random Forest (RF) model across horizons h = 1–8. IQR_t^{RF} denotes the interquartile range of tree-level forecasts. R_1 and R_2 compare average IQR width to realized and forecast error volatility, while ρ_1 and ρ_2 capture the correlation of forecast uncertainty with these volatility measures. Values near one indicate well-calibrated uncertainty levels; positive correlations imply that forecast dispersion increases when inflation becomes more volatile.

Table 4 summarizes the resulting measures across forecast horizons. The coverage rates of the RF's interquartile ranges lie between 66% and 84% across horizons h = 1–8, indicating that the model's uncertainty estimates are somewhat conservative—its forecast bands tend to be slightly wider than necessary, yet still provide reasonable probabilistic coverage of realized inflation outcomes. While Clark, Ganics, and Mertens (2025) document systematic undercoverage in SPF-based fan charts, the RF model achieves more balanced coverage, suggesting that data-driven ensembles can better approximate true predictive uncertainty. The median IQR width fluctuates between roughly one and one and a half percentage points, with slightly larger bands at longer horizons (h = 6–8), reflecting the natural increase in forecast uncertainty over time.

The calibration ratios further confirm that the RF's forecast dispersion scales proportionately with both realized inflation volatility and the variability of its own forecast errors. Across horizons, R_1 ranges from about 0.4 to 0.6 and R_2 from roughly 0.5 to 0.9, suggesting that the model's internal uncertainty tracks the magnitude of true and predictive volatility reasonably well. In line with Rossi, Sekhposyan, and Soupre (2016), who distinguish between realized and expectational uncertainty, the RF's dispersion measures align with both realized and forecast error volatility, indicating that its

uncertainty estimates capture both ex-ante and ex-post components of inflation uncertainty. This implies that the model's ex-ante forecast bands neither systematically understate nor exaggerate actual variability, even at medium and long horizons.

The positive correlations between forecast uncertainty and both realized and forecast error volatility—ranging from roughly 0.2 to 0.5 across horizons—confirm that the RF's forecast dispersion is economically meaningful. Consistent with Abel et al. (2016), who show that forecast dispersion among professional forecasters rises in periods of heightened volatility, the RF's IQR-based uncertainty measures expand when inflation becomes more volatile, reinforcing their interpretation as meaningful indicators of uncertainty. Moreover, whereas Clements et al. (2025) find that professional forecasters tend to underreact to rising uncertainty, the RF's uncertainty measures respond contemporaneously to changes in realized and forecast error volatility, indicating a more adaptive and data-responsive updating mechanism.

Finally, following Lahiri and Sheng (2010), who argue that forecast disagreement is informative only when it comoves with objective measures of uncertainty, the RF model's IQRs exhibit significant correlations with realized inflation volatility. This supports the view that the model's dispersion reflects genuine uncertainty rather than mechanical variation in forecasts. Taken together, these results demonstrate that the Random Forest model produces forecast uncertainty estimates that are both well-calibrated and responsive to underlying inflation dynamics, offering a credible and interpretable measure of predictive confidence rather than arbitrary statistical noise.

6 Relative Feature Importance

To gain insight into how individual features contribute to the Random Forest's inflation forecasts, I compute Shapley values, a widely used interpretability measure from cooperative game theory (see Štrumbelj and Kononenko (2014) and Lundberg and Lee (2017)). Shapley values decompose each prediction into additive feature-level contributions relative to a baseline prediction, typically the mean of the training data. Formally, each forecast $\hat{y}_i$ can be expressed as

$$\hat{y}_i = \phi_0 + \sum_j \phi_{ij},$$

where ϕ_{ij} denotes the marginal contribution of feature j to observation i. In this application, I compute Shapley values for every tree in the ensemble and average them across all trees and sample observations. The resulting mean absolute Shapley values summarize each variable's average influence on the model's forecasts, providing an interpretable measure of relative importance that complements traditional split-based importance scores (see e.g. Buckmann, Potjagailo, and Schnattinger (2025) for a related application).

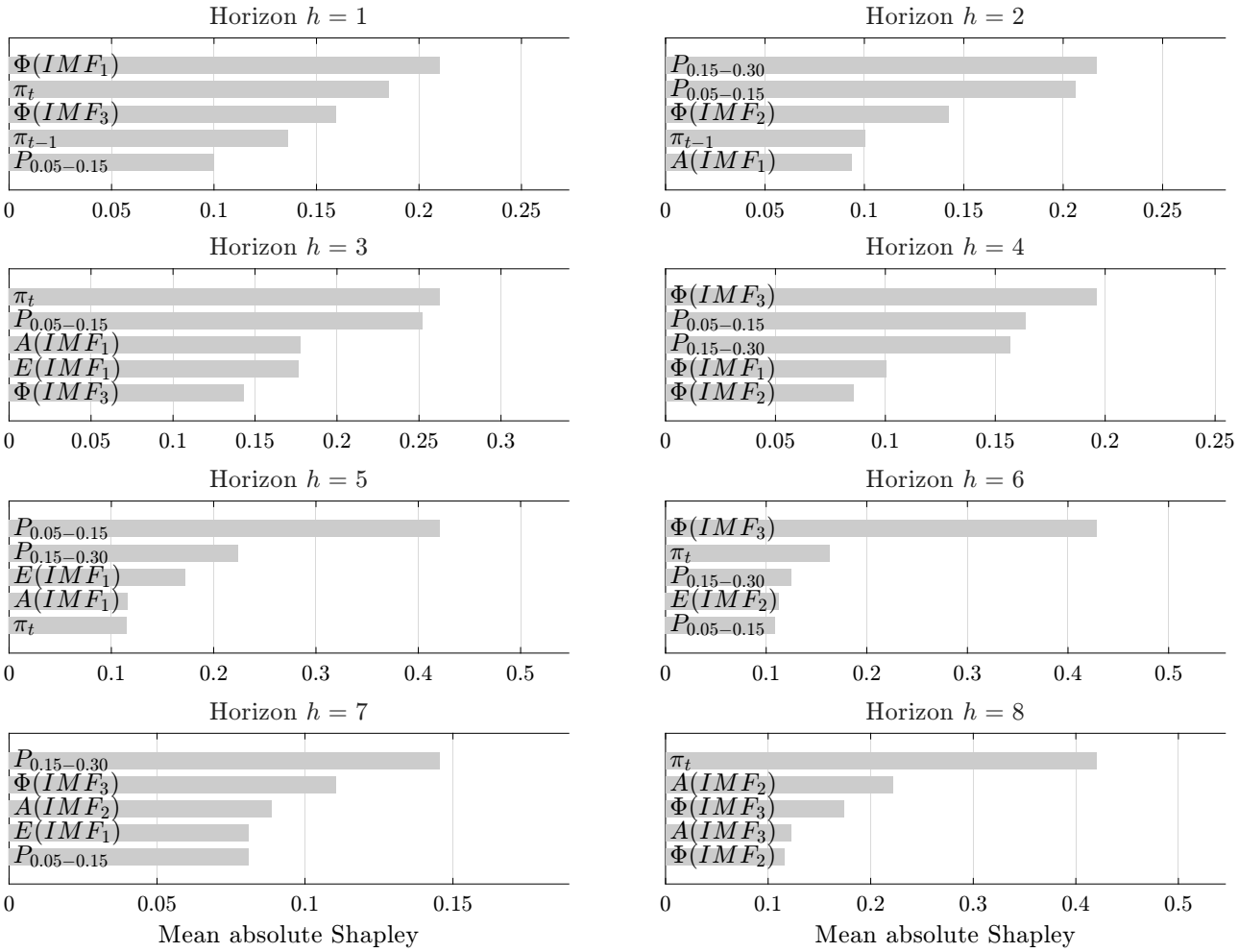

Figure 4: Five highest mean absolute Shapley values across features for each forecast horizon ($h = 1, \ldots, 8$)

Notes: Each panel displays the five predictors with the highest mean absolute Shapley values for the Random Forest inflation forecasting model across forecast horizons. Shapley values quantify each variable's average contribution to the model's predictions with higher values indicating stronger influence on forecast outcomes.

Figure 4 displays the five most influential predictors at each forecast horizon based on these mean absolute Shapley values. The Shapley framework thus provides a transparent decomposition of the Random Forest's predictions, directly quantifying how changes in individual predictors affect the model's output and revealing how the relevance of different features evolves across forecast horizons.

Across horizons, past inflation (π_t, π_{t-1}) remains among the dominant drivers at short horizons,

consistent with the strong persistence of inflation dynamics. At medium-term horizons ($h = 3$–5), cyclical and frequency-domain features—particularly those associated with the bandpass components ($P_{0.05-0.15}$, $P_{0.15-0.30}$) and IMF-based amplitudes—gain prominence, indicating that cyclical patterns become more predictive further ahead. At longer horizons ($h = 6$–8), energy-related features ($E(\pi_w)$, $E(\text{IMF}_1)$) and low-frequency components contribute more substantially, suggesting that slow-moving or structural factors increasingly shape inflation expectations.

Overall, the Shapley analysis confirms that the Random Forest model's predictive structure is economically interpretable: short-run forecasts rely primarily on inflation inertia, while medium- and long-run forecasts draw on cyclical and energy-based information that captures the evolving nature of inflation dynamics. The investigation of relative feature importance also provides an indirect robustness check of the feature-engineering approach—if the engineered transformations are informative, they should exhibit consistently positive Shapley contributions across horizons. This expectation is largely met. Only the mean energy of the past inflation window, $E(\pi_w)$, never appears among the top five predictors at any horizon, suggesting that the model attributes less marginal importance to this particular transformation. Nonetheless, its Shapley values remain positive, confirming that it contributes modestly to the forecasts. The same pattern emerges when importance is computed directly from the Random Forest via impurity-based measures, underscoring the stability of the model's feature relevance across evaluation methods.

7 Concluding Remarks

This paper shows that a minimalist, data-driven framework—relying exclusively on inflation's own historical dynamics—can generate forecasts that not only rival but frequently exceed the accuracy of professional forecasters and well established univariate benchmark models. Rather than introducing a new high-dimensional dataset, innovative transformation technique, or novel learning algorithm, the contribution lies in establishing a robust, reproducible, and interpretable univariate benchmark.

The purpose of this paper is not to demonstrate incremental gains across many economies or to tailor the model to country-specific institutional environments. The United States serves as a natural testing ground, given its central role in the global macroeconomic and monetary policy landscape and the extensive use of U.S. inflation forecasts for model evaluation. Starting from this benchmark case

establishes a clear upper bound on what can be achieved from inflation's own past dynamics alone. Should a researcher's objective be to further enhance predictive performance, the next step would be to augment the feature set with external macroeconomic or financial indicators, rather than to alter the core modeling architecture. In this sense, the contribution of the paper is to provide a disciplined baseline: it isolates the forecasting value of inflation's intrinsic temporal structure before additional sources of information are layered on.

The proposed model is not intended to replace more complex, theory-rich, or institutionally curated forecasting systems; instead, it complements them by offering a transparent and low-cost tool that responds flexibly to rapid changes in inflation dynamics. By embedding real-time sampling, systematic feature design, and uncertainty quantification into a unified and operationally simple structure, the paper provides policymakers and researchers with a practical forecasting benchmark that can serve as a responsive early indicator, a robustness check against overfitting in richer models, or a baseline for model comparison in real-time policy environments.

References

Abel, J., R. Rich, J. Song, and J. Tracy (2016). The measurement and behavior of uncertainty: evidence from the ecb survey of professional forecasters. *Journal of Applied Econometrics 31*(3), 533–550.

Araujo, G. S. and W. P. Gaglianone (2023). Machine learning methods for inflation forecasting in brazil: New contenders versus classical models. *Latin American Journal of Central Banking 4*(2), 100087.

Beck, E., D. Kozbur, and M. Wolf (2024). The hedged random forest. *Available at SSRN 5032102*.

Beck, E. and M. Wolf (2025). Forecasting inflation with the hedged random forest. *Swiss National Bank Working Paper*.

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

Buckmann, M., G. Potjagailo, and P. Schnattinger (2025). Blockwise boosted inflation: Non-linear determinants of inflation using machine learning. *Bank of England Staff Working Paper Series (1143)*.

Cerqueira, V., N. Moniz, and C. Soares (2024). Vest: Automatic feature engineering for forecasting. *Machine Learning 113*(7), 4523–4545.

Clark, T. E., G. Ganics, and E. Mertens (2025). Constructing fan charts from the ragged edge of spf forecasts. *Review of Economics and Statistics*, 1–45.

Clements, M. P., R. W. Rich, and J. Tracy (2025). An investigation into the uncertainty revision process of professional forecasters. *Journal of Economic Dynamics and Control 173*, 105060.

De Brabandere, A., P. Robberechts, T. Op De Beéck, and J. Davis (2019). Automating feature construction for multi-view time series data. In *ECMLPKDD Workshop on Automating Data Science*, pp. 1–16. N/A.

Diebold, F. X. and R. S. Mariano (2002). Comparing predictive accuracy. *Journal of Business & economic statistics 20*(1), 134–144.

Friedman, M. (1977). Nobel lecture: inflation and unemployment. *Journal of political economy 85*(3), 451–472.

Fulton, C. and K. Hubrich (2021). Forecasting us inflation in real time. *Econometrics 9*(4), 36.

Guyon, I. and A. Elisseeff (2006). An introduction to feature extraction. In *Feature extraction: foundations and applications*, pp. 1–25. Springer.

Harvey, D., S. Leybourne, and P. Newbold (1997). Testing the equality of prediction mean squared errors. *International Journal of forecasting 13*(2), 281–291.

Lahiri, K. and X. Sheng (2010). Measuring forecast uncertainty by disagreement: The missing link. *Journal of Applied Econometrics 25*(4), 514–538.

Lundberg, S. M. and S.-I. Lee (2017). A unified approach to interpreting model predictions. *Advances in neural information processing systems 30*.

Medeiros, M. C., G. F. Vasconcelos, Á. 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.

Naghi, A. A., E. O'Neill, and M. Danielova Zaharieva (2024). The benefits of forecasting inflation with machine learning: New evidence. *Journal of Applied Econometrics 39*(7), 1321–1331.

Newey, W. K. and K. D. West (1987). Hypothesis testing with efficient method of moments estimation. *International Economic Review*, 777–787.

Oliveira, M. and L. Torgo (2015). Ensembles for time series forecasting. In *Asian Conference on Machine Learning*, pp. 360–370. PMLR.

Praveen, M., S. Dekka, D. M. Sai, D. P. Chennamsetty, and D. P. Chinta (2025). Financial time series forecasting: A comprehensive review of signal processing and optimization-driven intelligent models. *Computational Economics*, 1–27.

Rossi, B., T. Sekhposyan, and N. Soupre (2016). Understanding the sources of macroeconomic uncertainty. *CEPR Discussion Papers* (11415), 52–69.

Shu, W. and Q. Gao (2020). Forecasting stock price based on frequency components by emd and neural networks. *Ieee Access 8*, 206388–206395.

Stock, J. H. and M. W. Watson (2007). Why has us inflation become harder to forecast? *Journal of Money, Credit and banking 39*, 3–33.

Štrumbelj, E. and I. Kononenko (2014). Explaining prediction models and individual predictions with feature contributions. *Knowledge and information systems 41*(3), 647–665.

Verdonck, T., B. Baesens, M. Óskarsdóttir, and S. vanden Broucke (2024). Special issue on feature engineering editorial. *Machine learning 113*(7), 3917–3928.

Verona, F. (2025). From waves to rates: Enhancing inflation forecasts through combinations of frequency-domain models. *Bank of Finland Research Discussion Papers 1*.