



MASTER'S DEGREE THESIS

Can Machine Learning Outperform Traditional Econometric Models?

Evidence from Euro Area Inflation Forecasting

Author:

Andrea Češková

Supervisor:

Prof. Dr. Guido Ascari

Co-Supervisor:

Prof. Dr. Bernhard Schipp

A thesis submitted in fulfillment of the requirements for the double degree in

Master in Economics, Development and Innovation

University of Pavia

and

Master in Public and International Economics

TU Dresden

in the

Department of Economics and Management, University of Pavia

September, 2026

Acknowledgements

I would like to express my sincere gratitude to my thesis supervisor, Prof. Guido Ascari, for his guidance, valuable feedback, and continuous support throughout the research and writing of this thesis. His expertise and encouragement were instrumental in shaping this work. I would also like to express my deepest gratitude to my partner, Denis, for his unwavering support throughout the years of my university studies. His patience, encouragement, and belief in me carried me through the most challenging moments of this journey, and I could not have completed this thesis without him by my side.

List of Figures

2.1	Euro Area HICP Inflation and Component Contributions	3
3.1	Euro Area HICP Inflation: Target Variable by Forecast Horizon	21
5.1	Variable importance in respective ML models	57
5.2	SHAP importance across forecast horizons, Random Forest and XGBoost	58
5.3	Total SHAP attribution magnitude by horizon, Random Forest vs. XGBoost	59
5.4	Local SHAP decomposition, Random Forest, September 2022, $h = 12$. . .	59
5.5	Local LIME explanations, $h = 1$, Random Forest and XGBoost	61
5.6	Local LIME explanations, $h = 12$, Random Forest and XGBoost	62
5.7	Giacomini-Rossi fluctuation test: rolling DM statistic vs RW(3m)	63
5.8	Giacomini-Rossi fluctuation test: $\sup t $ relative to the 5% critical value . .	64
6.1	Accuracy gains of lagged predictors in LASSO and RIDGE	66
6.2	RMSFE by model and horizon, expanding vs. rolling (120-month) window	69
6.3	RMSFE rank stability across window regimes, by horizon	70
7.1	Pearson correlation of every predictor with core HICP inflation	90

List of Tables

3.1	Model Overview	20
3.2	Mean selected penalty and model size by horizon	33
3.3	Hyperparameters used for ML models	33
4.1	Target Variable properties	47
4.2	Most correlated predictors with core euro area inflation	48
5.1	Relative forecast accuracy: RMSFE, MAE, and MAD vs. RW(3m)	52
5.2	Forecasting accuracy with period split (Pre-2020, Post-2020)	53
6.1	RMSFE by model, horizon, and estimation window regime	68
7.1	Predictor panel: source, frequency, and transformation	85
7.2	Relative forecast accuracy: RMSFE, MAE and MAD after introducing lagged predictors	91

Abstract (English)

This thesis asks whether machine learning can outperform traditional econometric models in forecasting euro area core HICP inflation, and whether any improvement stems from a richer information set or from nonlinearity. A hierarchy of benchmark (random walk), econometric (UCSV, a dynamic factor model), and machine-learning (Ridge, Lasso, Elastic Net, Random Forest, XGBoost) models are estimated recursively under an expanding window (Jan. 2000–Feb. 2026), with pseudo-out-of-sample evaluation from Jan. 2010 at horizons $h = 1, 3, 6, 12$ months. The predictor panel combines EA-MD-QD with supplementary ECB, Eurostat and FRED series, yielding 62 real-time-consistent monthly predictors, with every pipeline step designed so no model ever uses information unavailable at forecast time.

In point-estimate terms, the dynamic factor model and the shrinkage estimators (Ridge, Lasso, Elastic Net) deliver the most consistent gains over the random-walk benchmarks, particularly at short-to-medium horizons and post-2020, while Random Forest and XGBoost deteriorate with the horizon on RMSFE despite Random Forest attaining the lowest median absolute error of any model at several horizons: its typical forecast error is small, but a handful of large errors, concentrated around the 2020–2023 disruption, drive its RMSFE up disproportionately. Formal inference is more cautious than these point estimates: Diebold-Mariano tests reject equal predictive ability against the random walk in only a handful of candidate-horizon combinations, mostly pre-2020, and the Model Confidence Set fails to distinguish almost any candidate at any horizon. Robustness checks reinforce this caution: a Giacomini-Rossi test traces the short-horizon edge to a 2020–2021 artefact, Ridge improves substantially once lagged predictors are added while Lasso does not, and Lasso's long-horizon advantage disappears under a rolling window. The evidence therefore favours a richer information set over nonlinear flexibility as the source of machine learning's value here. This advantage is real but modest, and rarely confirmed with strong statistical significance.

Keywords: inflation forecasting, machine learning, dynamic factor model, LASSO, Random Forest, gradient boosting, real-time forecasting, euro area

Abstract (Italian)

Questa tesi indaga se il machine learning sia in grado di sovraperformare i modelli econometrici tradizionali nella previsione dell'inflazione core HICP dell'area euro, e se un eventuale miglioramento derivi da un insieme informativo più ricco o dalla non linearità. Una gerarchia di modelli benchmark (random walk), econometrici (UCSV, un modello a fattori dinamici) e di machine learning (Ridge, Lasso, Elastic Net, Random Forest, XGBoost) viene stimata in modo ricorsivo con finestra espansiva (gen. 2000–feb. 2026), con valutazione pseudo-fuori-campione da gen. 2010 agli orizzonti $h = 1, 3, 6, 12$ mesi. Il pannello combina EA-MD-QD con serie supplementari di BCE, Eurostat e FRED, per 62 predittori mensili coerenti con un contesto di previsione in tempo reale, con ogni fase della pipeline costruita in modo che nessun modello utilizzi informazioni non ancora disponibili al momento della previsione.

In termini di stime puntuali, il modello a fattori dinamici e gli stimatori di shrinkage (Ridge, Lasso, Elastic Net) offrono i guadagni più consistenti rispetto ai benchmark random walk, in particolare agli orizzonti brevi-medi e nel periodo successivo al 2020, mentre Random Forest e XGBoost peggiorano al crescere dell'orizzonte in termini di RMSFE, nonostante Random Forest raggiunga l'errore assoluto mediano più basso tra tutti i modelli a diversi orizzonti: il suo errore di previsione tipico è contenuto, ma un ristretto numero di errori di grande entità, concentrati attorno alla discontinuità del 2020–2023, ne gonfia l'RMSFE in modo sproporzionato. L'inferenza formale è più cauta di quanto suggeriscano queste stime puntuali: i test di Diebold-Mariano rifiutano l'ipotesi di uguale capacità predittiva rispetto al random walk solo in una minoranza delle combinazioni modello-orizzonte considerate, per lo più nel periodo precedente al 2020, e il Model Confidence Set non riesce a distinguere quasi nessun candidato a nessun orizzonte. I controlli di robustezza rafforzano questa cautela: un test di Giacomini-Rossi riconduce il vantaggio di breve periodo a un artefatto legato alla finestra 2020–2021, Ridge migliora sensibilmente con l'aggiunta di predittori ritardati mentre Lasso no, e il vantaggio di Lasso ai lunghi orizzonti scompare passando a una finestra mobile. L'evidenza favorisce dunque un insieme informativo più ricco, piuttosto che la flessibilità non lineare, come fonte del valore aggiunto del machine learning in questo contesto – un vantaggio reale ma modesto, e raramente confermato con una significatività statistica robusta.

Keywords: previsione dell'inflazione, machine learning, modello a fattori dinamici, LASSO, Random Forest, gradient boosting, previsione in tempo reale, area euro

Contents

Acknowledgements	i
List of Figures	ii
List of Tables	iii
1 Introduction	1
2 Literature review	3
2.1 Inflation forecasting and the limits of traditional models	3
2.1.1 Univariate benchmarks: Random Walk and UCSV	4
2.1.2 Factor Models	5
2.2 Machine learning in macroeconomic forecasting	9
2.2.1 Shrinkage Methods (Regularization)	10
2.2.2 Tree Models: Random Forests and Gradient Boosting	12
2.2.3 Interpretability	14
2.2.4 Empirical evidence	16
2.3 Synthesis and implications for the research design	17
3 Methodology	19
3.1 Target Variable and Forecast Horizons	20
3.2 Econometric models	23
3.2.1 Random walk	23
3.2.2 UCSV	24
3.2.3 Factor Model	25
3.3 ML models	29
3.3.1 Shrinkage methods	29
3.3.2 Tree methods	31
3.3.3 Hyperparameters	31
3.3.4 Variable Selection and Interpretability: SHAP and LIME	34
3.4 Evaluation	37
3.4.1 Metrics and Forecast Combination	37
3.4.2 Estimation Windows and sub-periods	40
4 Data	42
4.1 Core Dataset: EA-MD-QD	42
4.2 Supplementary Series	43
4.3 Preprocessing	43

4.4	Missing Data	44
4.5	Outlier Treatment	45
4.6	Data Properties	46
5	Results	50
5.1	Forecast evaluation	50
5.2	Variable Selection and Interpretability: SHAP and LIME	55
5.3	Model Stability	63
6	Robustness Check	65
6.1	Predictor Lags: Extending the Feature Set for Shrinkage Methods	65
6.2	Rolling Window Estimation	67
7	Discussion	72
	Bibliography	77
	Appendix	85

1 Introduction

The 2021-2023 inflation surge was difficult for most forecasters to anticipate, and it renewed a longstanding question about whether the models commonly used for inflation forecasting, largely developed and validated during the calmer years since the Global Financial Crisis, remain well suited to more turbulent periods. Simple benchmarks such as the random walk and the Unobserved Components Stochastic Volatility (UCSV) model have historically proven difficult to beat (Atkeson and Ohanian, 2001; James H. Stock and Mark W. Watson, 2007), and dynamic factor models tend to improve on them only modestly, by pooling information from large macroeconomic panels while remaining strictly linear (James H. Stock and Mark W. Watson, 2002b; Eickmeier and Ziegler, 2008). Machine learning methods are often proposed as a way to do more: by relaxing the linearity restriction, tree-based ensembles such as Random Forest and gradient boosting can in principle capture threshold effects, asymmetries, and interactions that a linear model cannot, and some early evidence reported RMSE gains of around 30% over standard benchmarks (Medeiros, Vasconcelos, et al., 2021). More recent evidence is less conclusive: Naghi, O'Neill, and Danielova Zaharieva (2024) find that this Random Forest advantage weakens considerably once the 2020-2022 period enters the sample, and Goulet Coulombe, Leroux, et al. (2022) suggest that ML's apparent advantage may come from two different sources, a larger information set or genuine nonlinearity, which a simple accuracy comparison alone may not be able to distinguish.

This thesis takes up that distinction for euro area core HICP inflation, a target for which the underlying question, and the evidence, remain comparatively scarce. It asks:

- (Q1) Can machine learning models outperform the random walk, UCSV, and dynamic factor model benchmarks in forecasting euro area core inflation?
- (Q2) Holding the predictor panel fixed, does relaxing the linearity restriction, moving from the penalised linear models (Ridge, LASSO, Elastic Net) to tree-based ensembles (Random Forest, gradient boosting), improve forecast accuracy?
- (Q3) Are these patterns stable across forecast horizons and across the pre- and post-2020 regimes, or do they depend on the specific episode being forecast?

To answer these questions, eight models, spanning the random walk, UCSV, a dynamic factor model, three regularised linear models (Ridge, LASSO, Elastic Net), and two tree-based ensembles (Random Forest, gradient boosting), are estimated recursively on a panel of 62 real-time-consistent monthly predictors and evaluated pseudo-out-of-sample at

horizons $h = 1, 3, 6, 12$ months, from January 2010 through February 2026. Because each model in this ladder changes exactly one methodological choice relative to the one below it (information set, regularisation, or linearity), the pattern of relative performance across the ladder, backed by Diebold-Mariano tests, model confidence sets, and SHAP/LIME decompositions of the tree-based models, is designed to answer Q1–Q3 directly.

The remainder of the thesis is organised as follows: Chapter 2 reviews the literature on univariate benchmarks, factor models, and machine-learning methods in macroeconomic forecasting, and draws out the implications of that literature for the empirical design adopted here. Chapter 3 sets out the forecasting framework in detail: the target variable and horizons, the three econometric benchmarks, the five machine-learning models, and the forecast evaluation methodology. Chapter 4 describes the data, including the construction of the predictor panel and the preprocessing steps applied to it. Chapter 5 presents the main forecasting results, including the sub-period accuracy comparison, the formal statistical tests, and the SHAP/LIME interpretability analysis. Chapter 6 subjects these results to the robustness checks described above, and Chapter 7 discusses their implications, situates them within the broader literature on machine learning in macroeconomic forecasting, and concludes.

2 Literature review

2.1 Inflation forecasting and the limits of traditional models

This chapter reviews the literature that underpins the empirical analysis of euro area core inflation forecasting in this thesis. It focuses on three broad strands of the forecasting literature: simple univariate benchmarks, factor models, and newer machine-learning approaches. The purpose of this review is not only to summarize the main findings in each strand, but also to clarify why these approaches are relevant for the comparison carried out in the empirical part of the thesis.

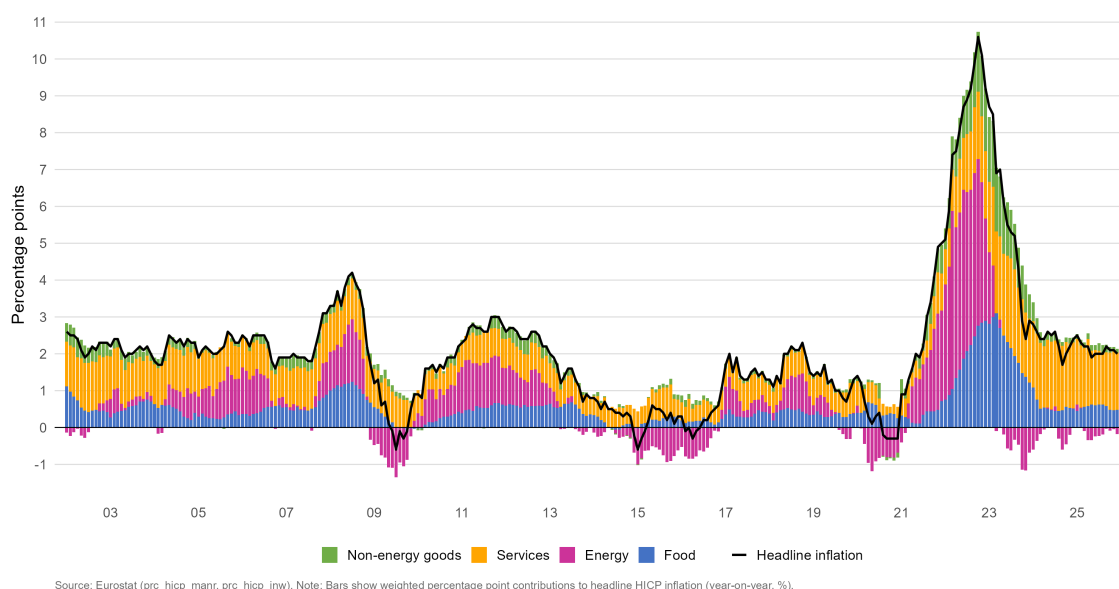


Figure 2.1: Euro Area HICP Inflation and Component Contributions

Figure 2.1 presents the evolution of euro area headline HICP inflation and the contributions of its main components. It illustrates the changing nature of inflation dynamics and the different roles played by energy, food, non-energy industrial goods, and services. While inflation remained relatively stable during much of the pre-2020 period, it increased sharply after the COVID-19 shock, particularly during 2021–2023. The inflation surge was initially driven mainly by energy prices and was later reinforced by food and other components. These developments highlight the importance of distinguishing between headline and underlying inflation. Since volatile energy and food prices can strongly influence headline HICP, this thesis focuses on core HICP as a measure more closely related to persistent domestic inflationary pressures.

2.1.1 Univariate benchmarks: Random Walk and UCSV

The univariate benchmark literature establishes a surprisingly high standard. Atkeson and Ohanian (2001) propose a four-quarter-ahead random-walk forecast: the average inflation over the next four quarters is set equal to the average over the previous four. They evaluate three Phillips-curve specifications using the unemployment gap, Chicago Fed National Activity Index, and a composite leading indicator. They find that the ratio of Phillips-curve RMSE to this naive RW benchmark exceeds unity in every case over 1984-1999, with reported RMSE ratios in the range 1.05-1.35. As a result, the Atkeson-Ohanian (AO) random walk has since become a standard reference benchmark for the post-1984 "Great Moderation" period. The model's strength lies in its agnosticism: it imposes no assumptions about the determinants of inflation and therefore avoids misspecification. For the same reason, I adopt this approach as my first benchmark in the forecasting exercise.

James H. Stock and Mark W. Watson (2007) extend the models comparison by Atkeson and Ohanian (2001) and document that, for US four-quarter GDP-deflator inflation, the absolute RMSFE of an AR(AIC) forecast fall from 1.89 % in 1970 – 1983 to 0.62 % in 1984 – 2004. This corresponds to a two-thirds reduction. Phillips-curve forecasts collapse over the same period: at $h = 8$ quarters the capacity-utilisation Phillips curve lies 45 % below the AR in 1970 – 1983 but 52% above in 1984 – 2004. Within this shifting environment, the AO random walk is hard to beat at four-quarter horizons. Building on this evidence, James H. Stock and Mark W. Watson (2007) introduce the Unobserved Components Stochastic Volatility (UCSV) model, which decomposes inflation into a random-walk trend with stochastic-volatility innovations and a transitory component with separate stochastic volatility. They capture the changing dynamics of U.S. inflation and document that the mean squared error of four-quarter-ahead inflation forecasts from a univariate trend-cycle model falls from 1.99 in the 1970–1983 period to 0.35 after 1984, broadly tracking the absolute decline in inflation volatility. At the same time, they argue that inflation has become less predictable in a structural sense, because the predictable (trend) component of inflation has shrunk relative to the unpredictable (transitory) component. This pattern implies that more sophisticated multivariate models struggle to outperform simple univariate specifications in the post-1984 period.

In the literature on macroeconomic forecasting, these univariate models often serve as the foundational benchmarks against which more complex structural or multivariate models are measured. In the euro area, researchers such as Bańbura and Bobeica (2023) have used RW and UCSV to evaluate the performance of Phillips curve models. They found that while newer Phillips curve specifications (such as those incorporating inflation detrending) can sometimes beat these benchmarks, the gains are often episodic. Over a 25-year sample in the euro area, the UCSV and RW models yield similar Root Mean Squared Forecast Errors (RMSFE), though UCSV is particularly useful for density forecasting because it explicitly models time-varying volatility (Bańbura and Bobeica, 2023).

This robustness of atheoretical benchmarks extends beyond advanced economies: Duncan and Martínez-García (2019) find that the RW-AO model consistently outperforms

open-economy Phillips curves, factor-augmented models, and time-varying parameter specifications across 14 emerging market economies (1980-2016), attributing its success to an inherent ability to downweight past data that guards against the structural breaks and misspecification common in those economies. Faust and Wright (2013) similarly identify RW-AO and UCSV as exceptionally resilient non-stationary benchmarks, performing "reasonably well" and often statistically comparable to a "fixed ρ " AR(1)-in-gap-form benchmark.

Despite their predictive accuracy, these benchmark models also have several limitations. Because they are purely data-driven, they lack structural economic insights and cannot explain why inflation is changing. They are unable to simulate policy interventions, such as a change in a central bank's inflation target or the impact of a specific economic shock. Both RW and UCSV share another common deficit: they cannot anticipate turning points driven by forces that have not yet appeared in the inflation series itself. During a regime break driven by exogenous forces, be it a commodity price spike, a global supply chain disruption, or a sudden fiscal expansion, the UCSV can only begin updating its trend estimate once the shock has passed through into realised inflation. By that point, a model with access to energy prices, supply chain pressure indices, or import cost data would already have incorporated the signal. It is precisely at those turning points, the 2013-2019 disinflationary drift and, more severely, the 2021-2023 surge (both shown in Figure 2.1), that the case for data-rich and nonlinear alternatives is expected to be the strongest.

In this thesis, the RW and UCSV models therefore serve as core univariate benchmarks: RW represents the most agnostic forecast that exploits inflation's near-unit-root behavior, while UCSV adds a flexible time-varying trend and volatility structure, still relying only on inflation's own history.

2.1.2 Factor Models

Building up on simple univariate models discussed in the previous section, this thesis aims to implement also a data-intensive benchmark, as the data might contain signals relevant for this forecasting exercise. Factor models fill the "limited information" constraint inherent in univariate benchmarks like the Random Walk and UCSV by operating in a "data-rich environment" that allows them to draw on dozens or even hundreds of macroeconomic predictors. However, including high-dimensional data in a single equation model is problematic for many reasons: overfitting, multicollinearity, collapse of degrees of freedom or parameter uncertainty. As Rossi (2026) notes, forecasting with Big Data typically requires one of three strategies: aggregating the data and then forecasting, forecasting first and then aggregating, or jointly forecasting and aggregating.

Dynamic factor models (DFMs) are models specifically designed to exploit the temporal dimension of a data-rich environment (Darracq Paries et al., 2021). They are a form of dimensionality reduction and rest on single premise: the co-movements of a large number of macroeconomic time series can be summarised by a small number of unobserved (or latent) common factors. Fluctuations in any individual timeseries are thus decomposed

into a common component, driven by these shared factors, and an idiosyncratic component capturing series-specific dynamics or measurement error. As James H. Stock and Mark W. Watson (2016) observe in their work, this is a remarkably simple idea that is also strongly supported in empirical studies across a wide range of macroeconomic applications.

In principle, “dynamism” can enter a factor model at several distinct stages, and the literature offers estimators for each. Dynamics can be built into the extraction step itself, as in the original dynamic factor model of Geweke (1977) and Sargent and Sims (1977) and its frequency-domain generalisation, the GDFM of Forni et al. (2005); cast in state-space form and estimated via the Kalman filter combined with quasi-maximum-likelihood (EM), as in Doz, Giannone, and Reichlin (2011), an approach whose natural handling of unbalanced, ragged-edge panels explains its popularity in real-time nowcasting at the Federal Reserve and the ECB (Giannone, Reichlin, and Small, 2008; Bańbura, Giannone, and Reichlin, 2011); or introduced downstream by embedding the factors in a VAR alongside observable variables, as in the factor-augmented VAR (FAVAR) of Bernanke, Boivin, and Elias (2005).

James H. Stock and Mark W. Watson (2002b), by contrast, locate dynamism not in the extraction step but in the forecasting equation: factors are estimated by static principal components, and any dynamic dependence is introduced afterward by including lagged factors and lagged dependent variables in a factor-augmented regression. Stock and Watson show formally that a dynamic factor model with q dynamic factors and s relevant lags admits an equivalent “static” representation with $r = q(s + 1)$ static factors, meaning that static PCA applied to a sufficiently rich panel consistently recovers the factor space spanned by the current and lagged dynamic factors, even without estimating their law of motion explicitly. This result constitutes a foundational contribution to the architecture of factor models in macroeconomic forecasting. More generally, James H. Stock and Mark W. Watson (2002b) establish the consistency of PCA-based factor estimates as both the number of series N and the number of time periods T grow large, under the weak assumptions of the approximate factor model formalised by Chamberlain and Rothschild (1982). This result removes the need to impose parametric restrictions on the distribution of the idiosyncratic disturbances, and makes it possible to estimate factor models using datasets containing hundreds of macroeconomic series.

The DFM will serve as the key traditional multivariate benchmark in this thesis: it will exploit the same large information set as the ML models but through a strictly linear mechanism. If ML models outperform the DFM, the excess forecasting gain will be directly attributable to the ability to capture nonlinearities rather than simply exploiting a larger information set. Goulet Coulombe, Leroux, et al. (2022) confirm that this distinction is empirically meaningful, showing that both data richness and nonlinearity contribute independently to ML forecasting gains in macroeconomic applications.

Factor Extraction: Dynamic or Static (Stacked)?

A cornerstone contribution to the architecture of factor modeling is the formalization by James H. Stock and Mark W. Watson (2002b) of the relationship between static and dynamic factor extraction. In its dynamic form, the factor model expresses an $N \times 1$ vector of observed series X_t as loading on q latent dynamic factors f_t through a distributed-lag factor structure, plus a mean-zero idiosyncratic component uncorrelated with the factor innovations at all leads and lags. If the idiosyncratic disturbances are additionally uncorrelated across series, the model is *exact*: the common component is then the sole source of co-movement between any two series. In practice, series drawn from common surveys or sharing sector-specific dynamics may exhibit residual cross-sectional correlation, which motivates the *approximate* DFM of Chamberlain and Rothschild (1982), permitting limited correlation among the idiosyncratic terms.

James H. Stock and Mark W. Watson (2002b) show that this dynamic representation admits an observationally equivalent *static*, or stacked, form (James H. Stock and Mark W. Watson, 2016). Letting p be the degree of the dynamic factor loadings' lag polynomial and stacking the current and p lagged dynamic factors into the $r \times 1$ vector $F_t = (f'_t, f'_{t-1}, \dots, f'_{t-p})'$, $r \geq q$, the DFM can be written as:

$$X_t = \Lambda F_t + e_t, \quad (2.1)$$

$$F_t = \Phi(L) F_{t-1} + G \eta_t, \quad (2.2)$$

where Λ is the $N \times r$ matrix of static factor loadings and $\Phi(L)$ absorbs the dynamics of the original factor VAR. The static form is preferred for forecasting because it renders the model amenable to principal-components estimation and makes the factor-augmented forecasting regression directly tractable (James H. Stock and Mark W. Watson, 2002b; Smeeke and Wijler, 2018), without requiring the dynamic factors' own law of motion to be estimated explicitly. This is the representation adopted throughout Chapter 3.

Determining the Number of Factors

A closely related methodological question is how to determine the number of factors r to retain, and how to conduct valid inference once those factors are used in a subsequent forecasting regression. Both issues are addressed by Bai and Ng (2002) and Bai and Ng (2006), respectively, and both papers directly inform the estimation strategy adopted in this thesis.

Bai and Ng (2002) propose a class of information criteria to choose the number of static factor r of the form:

$$IC(r) = \ln V_r(\hat{\Lambda}, \hat{F}) + r \cdot g(N, T), \quad (2.3)$$

where $V_r(\hat{\Lambda}, \hat{F})$ is the least-squares objective function evaluated at the principal components (PC) estimators and $g(N, T)$ is a penalty satisfying $g(N, T) \rightarrow 0$ and $\min(N, T) \cdot g(N, T) \rightarrow \infty$ as $N, T \rightarrow \infty$. Among the proposed variants, the IC_{p_2} criterion employs $g(N, T) = [(N + T)/NT] \ln[\min(N, T)]$, which simplifies to twice the BIC penalty when

$N = T$ and has been shown to perform well in simulation settings calibrated to macroeconomic data James H. Stock and Mark W. Watson (2016). Unlike traditional criteria like AIC or BIC, which were developed for fixed N , Bai and Ng's criteria are consistent as both N and T tend to infinity (Bai and Ng, 2002). This is essential for macroeconomic datasets where the cross-sectional dimension is often as large as the temporal one (Forni et al., 2005).

A second, related question arises once the estimated factors $\hat{F}_t$ (rather than the true, unobserved factors F_t) are used as regressors in a forecasting equation of the diffusion-index type (James H Stock and Mark W Watson, 2002a). Bai and Ng (2006) address this "generated regressors" problem for such factor-augmented regressions. They show that least-squares estimates and forecasts constructed from $\hat{F}_t$ in place of F_t remain consistent and asymptotically normal, but that unless the ratio of the cross-sectional to the time dimension, N/T , tends to zero, the sampling uncertainty introduced by estimating F_t is not asymptotically negligible and must be explicitly incorporated into valid inference. They derive analytical formulas for confidence and prediction intervals that remain valid regardless of the magnitude of N/T , including in settings where the factors are nonstationary. Together, Bai and Ng (2002) and Bai and Ng (2006) thus provide the two pieces of asymptotic theory underpinning the factor-augmented regression framework: a consistent, data-driven procedure for selecting r , and a formal basis for treating estimated factors as valid regressors in a second-stage forecasting equation.

Bai and Ng (2008) propose targeting the predictor set to the series being forecast, via hard- or soft-thresholding pretests, before extracting factors from the reduced panel, and report a non-trivial reduction in forecast error for inflation specifically. This targeted-predictor approach lies beyond the scope of this thesis and is not implemented: the factors used throughout are extracted from the full, quality-filtered predictor panel (Section 3), without a target-relevance screening step. The omission is deliberate: Lasso and Ridge, both already included in the model comparison, perform their own target-aware dimensionality reduction directly on the panel (sparse selection and continuous shrinkage, respectively), so a targeted-predictors DFM variant would largely duplicate a mechanism already represented elsewhere rather than add a genuinely distinct benchmark. The DFM is instead retained as a *target-agnostic* method, extracting factors purely from the panel's common comovement structure, providing a cleaner contrast with Lasso and Ridge's target-aware approaches to the same dimensionality-reduction problem.

Review of empirical applications

Following the theoretical developments in dimensionality reduction, the empirical literature has extensively tested the James H Stock and Mark W Watson (2002a) "diffusion index" specification. In the original application, James H Stock and Mark W Watson (2002a) used 215 monthly U.S. predictors to forecast eight series at 6-, 12-, and 24-month horizons, outperforming univariate autoregressions, small VARs, and leading-indicator models, especially for real-activity variables. Real-time testing introduced more caution: Bernanke and Boivin (2003) found factor models beat small-model benchmarks even with

vintage data but were still outperformed by the Federal Reserve’s judgmental Greenbook forecast, a “judgment gap” that Faust and Wright (2009) and Faust and Wright (2013) attribute to purely statistical models’ inability to capture the current-state nowcast and the long-run inflation trend.

For the euro area, results have been more cautious still. Angelini, Henry, and Mestre (2003) find that euro area price variables are driven by a single nominal factor that is a useful leading indicator for inflation, though its advantage over simple univariate benchmarks is less clear-cut than in the U.S.; Marcellino, James H Stock, and Mark W Watson (2003) further show that pooling country-specific univariate forecasts often outperforms models built directly on area-wide aggregates. Against this, a meta-analysis of 52 studies by Eickmeier and Ziegler (2008) finds that factor models do better for euro area inflation than for U.S. inflation, and Marra (2017) report that dynamic extensions of the static Stock-Watson specification can improve on it for inflation in large EU panels. Duncan and Martínez-García (2019) find factor models are a “reasonable second-best” in emerging markets, more exposed to structural breaks than simpler adaptive strategies.

Factor models also face two recurring criticisms directly relevant to this thesis’s model ladder. Standard PC extraction maximises variation across the whole panel, so it can miss a signal that is important for inflation specifically but explains little total variance (Faust and Wright, 2009; Faust and Wright, 2013); and the resulting factors are an “overly aggregated” representation that can suppress granular predictive information (Medeiros, Vasconcelos, et al., 2021). Recent ML research addresses exactly these limitations: Medeiros, Vasconcelos, et al. (2021) show that Random Forest can outperform factor specifications by up to 30% in MSE by exploiting dense, high-dimensional signal that PCA-based factors compress away, though Naghi, O’Neill, and Danielova Zaharieva (2024) find this advantage is sensitive to extreme regimes such as COVID-19. These approaches are discussed in the next section.

2.2 Machine learning in macroeconomic forecasting

While traditional econometric models remain the foundational workhorses of macroeconomic analysis due to their theoretical grounding and interpretability, they frequently struggle to capture the complex nonlinear dynamics and evolving structural relationships inherent in inflation (Bolívar, 2026). A primary limitation of these classical approaches is the “degree of freedom” problem, which restricts the number of predictors to a small subset of the available data, often leading to information loss in an increasingly data-rich environment (Aras and Lisboa, 2022). Classical OLS in high dimensions produces low-bias but high-variance estimates that generalise poorly to new data, whereas regularised or nonparametric estimators explicitly sacrifice bias to reduce variance, often yielding better out-of-sample performance.

In this sense, machine learning (ML) methodologies represent a “philosophical shift” by prioritizing out-of-sample predictive accuracy $\hat{y}$ over consistent parameter estimation $\hat{\beta}$, effectively bridging these gaps by discovering intricate patterns directly from data

with fewer a priori assumptions (J. C. Chen et al., 2022). By leveraging regularization and shrinkage to handle high-dimensional datasets and manage the bias-variance trade-off, ML models like Random Forests and Neural Networks can identify subtle interactions among hundreds of variables that would overwhelm traditional models (Abadie and Kasy, 2019). This flexibility makes ML models particularly resilient during turbulent regimes and structural breaks, such as the Global Financial Crisis or the COVID-19 pandemic—where traditional linear benchmarks often fail to adapt to rapidly shifting economic conditions (Benecká, 2026).

2.2.1 Shrinkage Methods (Regularization)

The most direct route from classical regression to machine learning in the forecasting context runs through penalised linear models. Shrinkage methods, also known as regularization techniques, represent a family of estimation methods specifically engineered to handle high-dimensional problems where the number of parameters or predictive covariates is large relative to the number of observations (Abadie and Kasy, 2019). Their primary utility in empirical macroeconomics is to mitigate overfitting by introducing a penalty term to the objective function, which constrains the magnitude of the regression coefficients (Abadie and Kasy, 2019). The model family is characterized by a capacity to filter informative signals from noise, with different penalty specifications, determining whether the model performs continuous shrinkage or automatic variable selection. Unlike ordinary least squares (OLS), which often results in highly volatile and uninterpretable forecasts when faced with many correlated variables, shrinkage methods provide a stable and regularized framework for discovering structural patterns directly from data-rich environments (Zou and Hastie, 2005). This thesis will use 3 different methods: Ridge, LASSO, and Elastic Net.

Ridge regression, proposed by Hoerl and Kennard (1970), is a foundational penalized regression technique that stabilizes estimation by applying an ℓ_2 penalty (the sum of squared coefficients) to the objective function. This model minimizes the sum of squared residuals plus a penalty term governed by a tuning parameter λ , which controls the strength of the regularization. As λ increases, the penalty for large coefficients becomes more severe, leading to smaller estimated values. A defining characteristic of ridge regression is its continuous shrinkage mechanism; unlike methods like the LASSO, it shrinks coefficients toward zero without ever setting them to exactly zero, thereby retaining all predictors in the final model (Aras and Lisboa, 2022). Because it does not perform variable selection, it is often described as a "dense" model, making it particularly effective in macroeconomic environments where many correlated variables collectively contribute to the forecast (Joseph et al., 2021). Its primary advantage in empirical research is its ability to handle multicollinearity by managing the bias-variance trade-off and reducing the high variance associated with unconstrained estimators (Aras and Lisboa, 2022; Abadie and Kasy, 2019). In practical forecasting "horse races," ridge regression has frequently proven robust, consistently emerging as a top performer for its ability to capture stable linear relationships and accurately identify turning points in inflation dynamics.

The Least Absolute Shrinkage and Selection Operator (LASSO), introduced by Tibshirani (1996), serves as a foundational regularization technique that addresses the limitations of ordinary least squares in high-dimensional settings by replacing the ℓ_2 penalty with an ℓ_1 norm (the sum of the absolute values of the coefficients) to the objective function. The defining characteristic of the LASSO is its ability to perform automatic variable selection; unlike ridge regression, the LASSO's geometric properties allow it to shrink the coefficients of irrelevant or redundant predictors exactly to zero (Aras and Lisboa, 2022). This produces a sparse model that is highly interpretable, as it isolates a parsimonious subset of the most influential drivers of inflation from an otherwise overwhelming pool of candidate variables. This makes the method most useful, when only a small subset of the pool of potential predictors is truly relevant. In empirical macroeconomic research, LASSO-based models have frequently outperformed traditional ARIMA, VAR, and factor model benchmarks, particularly in environments where the number of predictors exceeds the number of observations ($p > n$) (Benecká, 2026). Despite these strengths, the LASSO faces a significant challenge when dealing with highly correlated predictors where it tends to arbitrarily select only one variable from a group of collinear features and ignore the others. Furthermore, its selection process can be unstable over time, meaning that small additions to the dataset can lead to a radically different set of selected variables, a drawback that has led to the adoption of more robust extensions such as the Adaptive LASSO and the Elastic Net (Li and W. Chen, 2014).

The Elastic Net, introduced by Zou and Hastie (2005), functions as a hybrid regularization technique designed to capitalize on the strengths of both ridge regression and the LASSO while mitigating their respective weaknesses. This model applies a weighted combination of ℓ_1 and ℓ_2 penalties, governed by a mixing hyperparameter, which allows it to perform automatic variable selection and continuous shrinkage simultaneously. Its primary contribution to macroeconomic forecasting is the grouping effect, which encourages highly correlated predictors to be included or excluded from the model as a collective unit (Aras and Lisboa, 2022). This mechanism specifically addresses the LASSO's tendency to arbitrarily select a single predictor from a cluster of correlated variables and ignore the rest of the group. Additionally, the Elastic Net provides a more robust framework for high-dimensional environments where the number of parameters exceeds the sample size ($p > n$); unlike the LASSO, which can select at most n variables before it saturates, the Elastic Net can potentially select all p predictors if necessary (Araujo and Gaglianone, 2023). To avoid excessive bias from double shrinkage, the definitive version of the estimator often incorporates a rescaling factor to undo the extra bias introduced by applying both penalties, thereby enhancing predictive accuracy (Joseph et al., 2021). In empirical research, the Elastic Net is frequently favored for its ability to balance the sparsity of the LASSO with the stability of ridge regression, making it highly effective in settings where predictors are numerous, interdependent, and characterized by complex co-movements (Zou and Hastie, 2005).

However, regularized linear models remain subject to an important limitation: despite their ability to handle many predictors, they generally impose an additive and linear relationship between the predictors and future inflation, unless nonlinear transformations

or interaction terms are specified in advance. This restriction may be problematic in macroeconomic environments characterized by threshold effects, asymmetric responses, state dependence, and interactions among economic variables. For example, the predictive relevance of an activity indicator may depend on whether inflation is already elevated, whether energy-price growth exceeds a particular threshold, or whether the economy is operating in a high- or low-volatility regime.

This limitation motivates the consideration of tree-based machine-learning methods as a second model family, which will be further discussed in the following chapter.

2.2.2 Tree Models: Random Forests and Gradient Boosting

Tree-based models constitute a prominent family of nonparametric machine learning techniques that have gained significant traction in macroeconomic forecasting due to their unique ability to model complex nonlinear relationships and high-order interactions among predictors (Chakraborty and Joseph, 2017). Unlike the penalized linear regressions previously discussed, tree models do not assume a specific functional form; instead, they rely on the recursive binary partitioning of the covariate space. The algorithm operates by consecutively branching the training data based on specific feature thresholds to minimize a predefined loss function, typically the mean squared error (MSE), within the resulting "data buckets" or leaves. This hierarchical structure allows for a flexible, data-driven mapping between predictors and inflation, where each path down the branches effectively captures unique interactions between various economic indicators that traditional linear models often fail to represent. While single decision trees provide a highly interpretable and visually intuitive framework, they are frequently characterized by high prediction variance and a significant risk of overfitting noisy training data (Maehashi and Shintani, 2020). The empirical literature on inflation forecasting predominantly leverages ensemble methods which aggregate the predictions of multiple trees to enhance model stability and maximize out-of-sample predictive accuracy.

Random Forest (RF), introduced by Breiman (2001), is a powerful ensemble learning method that aggregates a collection of independent regression trees to enhance predictive stability and minimize the large variance often associated with single decision trees. The model operates through a dual mechanism of randomness: first, it employs bootstrap aggregation (bagging), as proposed in Breiman (1996), where each individual tree in the ensemble is trained on a random subsample of the data drawn with replacement (Aras and Lisboa, 2022). Second, to further decorrelate the trees, RF utilizes a "random subspace projection" by selecting only a random subset of predictors for consideration at each candidate split node rather than the full set of variables (Araujo and Gaglianone, 2023). The final forecast is obtained by taking the average of the predictions from all individual trees, a process that leverages the Law of Large Numbers to ensure the model does not overfit as more trees are added (Aras and Lisboa, 2022). Its predictive advantage is most pronounced during turbulent economic regimes and structural breaks, where it can identify subtle shifting patterns among hundreds of indicators. Furthermore, RF addresses the "black box" critique by providing internal diagnostics through out-

of-bag (OOB) estimation, which yields unbiased measures of generalization error and variable importance (Breiman, 2001). These importance scores, typically derived from the permutation-based mean decrease in accuracy, allow researchers to identify which economic drivers are most influential in the model's forecasting process.

While Random Forests leverage bagging to reduce prediction variance through the aggregation of independent trees, another powerful branch of tree-based ensembles is Gradient Boosting. This family of models represents a fundamental shift in ensemble logic: rather than building trees in parallel, boosting techniques operate through sequential learning, where each new model is iteratively added to the ensemble to specifically correct the errors—or residuals—of its predecessors (Aras and Lisboa, 2022).

By consecutively updating the estimator in the direction that most reduces a differentiable loss function, this process effectively "boosts" a collection of weak learners (often shallow decision trees) into a single, high-performance predictive model. In the empirical inflation literature, this sequential refinement is particularly valued for its ability to adapt to complex, shifting dynamics and high-volatility regimes, often yielding accuracy gains that rival or surpass those of bagging-based methods (Angelopoulos et al., 2025). This framework has laid the groundwork for advanced implementations like XGBoost and CatBoost, which incorporate sophisticated regularization and optimization techniques to handle high-dimensional macroeconomic datasets with greater speed and precision.

eXtreme Gradient Boosting (XGBoost), introduced by T. Chen and Guestrin (2016), is a scalable implementation of gradient boosting that has gained wide prominence for its predictive accuracy and computational speed (Aras and Lisboa, 2022). Like other boosting methods, it builds an ensemble of regression trees sequentially, each correcting the residuals of its predecessors, but adds a regularised learning objective, penalising both the number of terminal leaves and the magnitude of leaf weights, together with algorithmic optimisations for speed and scalability (T. Chen and Guestrin, 2016; Aras and Lisboa, 2022).

The economic interpretation of tree-based forecasting gains is explained by Goulet Coulombe, Leroux, et al. (2022), who find that the nonlinearity channel is the single largest contributor to forecast improvement at medium and long horizons. Trees capture threshold effects, asymmetric responses, and interaction terms without the investigator specifying them in advance. These features are particularly relevant for inflation dynamics, where the relationship between activity and prices can be nonlinear, state-dependent, or subject to regime shifts. Conversely, at very short horizons, nonlinearity contributes little and the performance gap between tree-based and linear shrinkage methods narrows considerably (Medeiros, Vasconcelos, et al., 2021; Aras and Lisboa, 2022).

Including both shrinkage and tree-based methods enables the empirical analysis to distinguish the contribution of regularization in a linear framework from the additional forecasting value of nonlinear and interaction effects. The following section discusses interpretability tools applied to these models, before the empirical evidence on their forecasting performance is reviewed in Section 2.2.4.

2.2.3 Interpretability

The growth of ML applications has been accompanied by growing concern over their interpretability. The "black-box" critique has historically been among the primary obstacles to their adoption in central bank and policy settings, where model transparency is a practical requirement for communicating policy decisions. The core difficulty is that the mapping from inputs to outputs in a deep tree ensemble or a high-dimensional kernel machine does not admit a compact analytical representation, so standard regression-table communication is unavailable.

A natural starting point for interpreting tree-based forecasts is the out-of-bag (OOB) permutation importance mechanism intrinsic to Random Forests. Because each tree is trained on a bootstrap subsample, the remaining observations form an out-of-bag validation set that requires no separate hold-out sample. Variable importance is then measured by randomly permuting the values of each predictor in the OOB set and recording the resulting increase in mean squared error: a large deterioration signals that the variable carries genuine predictive information, while a small one indicates the model is largely indifferent to it (Breiman, 2001). This permutation-based importance score is computationally cheap and available at no additional cost as a by-product of model training, making it the standard diagnostic in applied RF inflation forecasting (Medeiros, Vasconcelos, et al., 2021; Joseph et al., 2021). Its key limitation, however, is that importance scores are unsigned: they reveal *which* variables matter but not *how* they affect the forecast, nor do they decompose contributions at the level of individual predictions. These limitations motivate the use of model-agnostic post-hoc tools discussed below.

To address this, several model-agnostic interpretability techniques have emerged, making it possible to quantify additive contribution of each feature to the predictive performance of a model:

- Shapley Values (SHAP) by Scott M Lundberg and Lee (2017): Drawing from cooperative game theory, SHAP assigns an importance value to each feature for a specific prediction, ensuring consistency and local accuracy (Aras and Lisboa, 2022). Consistency is the key property for economic applications: if a contribution of a feature to the model's predictions increases, its Shapley value cannot decrease, preserving the ordering of feature importance under model changes. This technique is possible to use also on complex nonlinear models, but their effectiveness depends on data structure (sparsity) and the specific economic context (Goulet Coulombe, Göbel, and Klieber, 2024).
- Local Interpretable Model-Agnostic Explanations (LIME) by Ribeiro, Singh, and Guestrin (2016): In contrast to the axiomatic approach of SHAP, LIME focuses on approximating the behavior of a machine learning model locally around a specific observation. The core mechanism involves training a simpler, interpretable surrogate model (such as a sparse linear regression or decision tree) on a dataset of perturbed samples generated around the instance of interest. These perturbed

samples are weighted by their proximity to the original instance using a similarity kernel, ensuring the explanation achieves high local fidelity to the complex model's local decision boundary.

While LIME is computationally flexible and provides a qualitative understanding of feature contributions, researchers Aras and Lisboa (2022) note that its reliance on heuristically chosen parameters can lead to results that violate the consistency and local accuracy properties inherently guaranteed by the SHAP framework. Despite these theoretical limitations, LIME remains a practical tool for forecasters to determine whether an individual prediction should be trusted by translating complex model outputs into human-intelligible artifacts.

However, the choice between these local approximation methods is further complicated by the inherent properties of macroeconomic data. As already mentioned, predictors are often highly correlated, which is a significant obstacle for standard interpretability tools. Aras and Lisboa (2022) argue that standard SHAP can mistakenly assign non-zero importance to uninformative variables in such environments. To address this, they propose a modification that integrates feature selection with SHAP framework to eliminate uninformative features from the start, aiming to provide a sparser and more reliable explanation.

Beyond achieving sparsity, SHAP values remain fundamentally descriptive rather than inferential: Aras and Lisboa (2022) notes that they cannot support formal statistical inference on model forecasts, and Zhao and Hastie (2021) formalise the conditions under which post-hoc tools such as SHAP or partial dependence yield causally meaningful conclusions, warning that they typically do not. Buckmann and Joseph (2022) take a step toward closing this gap by incorporating a Shapley-regression approach developed by Joseph (2019), which enables formal hypothesis tests on individual predictor contributions. They note that Shapley values, unlike permutation importance, reveal the direction and magnitude of a predictor's contribution, including interaction effects. This is valuable when the aim is to understand *how* a model uses a variable, not only *which* variables it uses.

Buckmann and Joseph (2022) provide the first systematic application of the Shapley interpretability workflow to macroeconomic forecasting, using unemployment forecasting as their running example. Their framework makes an important conceptual distinction between *global* and *local* attributions. Global attributions summarise the average contribution of each predictor across the full sample, identifying the most consistently important variables for the model's predictions. Local attributions, by contrast, decompose the prediction at each individual time point, revealing which variables drove the forecast at a specific observation. This local-global decomposition is particularly valuable for inflation forecasting: the predictors relevant during the high-inflation episode of 2021–2023 may differ substantially from those important in the pre-2020 low-inflation regime, and aggregating across the full sample would obscure this heterogeneity.

The "Dual Interpretation" framework of Goulet Coulombe, Göbel, and Klieber (2024)

similarly reframes ML forecasts as combinations of past historical episodes, or "historical analogies," letting a forecast's credibility be judged by whether it leans on plausible historical precedents.

2.2.4 Empirical evidence

This section synthesizes the empirical evidence on ML performance in macroeconomic forecasting, focusing on inflation. Since 2018, a robust consensus has emerged that ML models systematically improve forecast accuracy over the traditional benchmarks discussed above (Araujo and Gaglianone, 2023; Pijush Kanti Das and Prabir Kumar Das, 2024; Medeiros, Vasconcelos, et al., 2021), challenging the earlier view that the RW and UCSV were nearly impossible to beat (Capolongo and Pacella, 2021; Medeiros, Vasconcelos, et al., 2021). XGBoost has increasingly joined Random Forest as a primary model in this literature (Naghi, O'Neill, and Danielova Zaharieva, 2024; Vancsura, Tatay, and Bareith, 2025; Aras and Lisboa, 2022), with their relative ranking depending on the dataset, inflation volatility, and training-window length.

The benchmark study is Medeiros, Vasconcelos, et al. (2021), who forecast US CPI inflation (1990-2015) using 508 predictors built from the FRED-MD panel and four lags. Random Forest is the most effective model, cutting RMSE by roughly 30% against the RW benchmark across most horizons, with Random Forest, Ridge, and a hybrid linear-RF model the only specifications retained in the 50% Model Confidence Set at every horizon; gains are statistically larger during recessions and high-uncertainty periods, and feature importance rotates from price and autoregressive terms at short horizons to labour and activity variables at longer ones. Goulet Coulombe, Leroux, et al. (2022) confirm this using the same database, showing nonlinearity contributes more than the sparse/dense regularisation choice, with ML's marginal value rising post-1995 and in high-volatility regimes.

Three mechanisms explain this advantage. First, ML handles high-dimensional predictor sets without the information loss of traditional approaches (Liu, Pan, and Xu, 2024; Pijush Kanti Das and Prabir Kumar Das, 2024): Smeeke and Wijler (2018) show lasso-type regression is robust to misspecification and competitive with or superior to factor models for price indices specifically, and Joseph et al. (2021) find gains from forecasting CPI sub-components before aggregating. Second, nonlinearity is the central channel: Goulet Coulombe, Leroux, et al. (2022) call nonparametric nonlinearity the "true game changer," most valuable when uncertainty is high or financial conditions are tight, since linear models tend to fail at turning points (Rossi, 2021). Third, macroeconomic prediction is generally dense rather than sparse: Giannone, Lenza, and Primiceri (2021) find negligible posterior probability of sparsity across six datasets including US inflation, favouring Ridge-type shrinkage over LASSO-type selection and supporting the inclusion of both in empirical comparisons.

These findings generalise across countries: Brazil (Araujo and Gaglianone, 2023), India (Pijush Kanti Das and Prabir Kumar Das, 2024), Japan (Liu, Pan, and Xu, 2024), Turkey (Aras and Lisboa, 2022). Gains are typically largest in volatile episodes: Random Forest

improves on ARIMA by nearly 43% for Indian inflation during COVID-19 but loses its edge in the calmer pre-pandemic sample (Pijush Kanti Das and Prabir Kumar Das, 2024), a pattern that anticipates the pre-/post-2020 split adopted in this thesis.

The euro area literature is more limited, reflecting the shorter sample available since the currency's inception in 1999. Capolongo and Pacella (2021) find a multi-country Bayesian VAR outperforms UCSV for HICP at short horizons but not longer ones, and that core HICP has a stronger domestic component than headline HICP, which is more exposed to global commodity shocks. Vancsura, Tatay, and Bareith (2025) apply a broad suite of models, including XGBoost, GRU, LSTM, and wavelet neural networks, to euro area HICP and its twelve subcomponents (2000-2023), finding RNN/LSTM architectures consistently outperform SVR and Random Forest in volatile environments.

The 2020-2023 period is a critical stress test for ML inflation forecasting. Naghi, O'Neill, and Danielova Zaharieva (2024) extend the Medeiros, Vasconcelos, et al. (2021) replication to Canadian and UK data through October 2022 and find Random Forest performance deteriorates substantially post-pandemic, with a stochastic-volatility model and some gradient-boosting variants becoming more accurate. Pijush Kanti Das and Prabir Kumar Das (2024) document the same asymmetry for Indian inflation, and Liu, Pan, and Xu (2024) find penalised regressions with external predictors beat the random walk during 2022–2023, though gains fade as the surge unwinds. The post-2020 sample thus offers a natural test of the stability of ML forecasting gains.

2.3 Synthesis and implications for the research design

The literature reviewed above carries five implications for the empirical study undertaken in this thesis.

First, the choice of univariate benchmark matters. Atkeson and Ohanian (2001), James H. Stock and Mark W. Watson (2007), Bańbura and Bobeica (2023), and Rossi (2025) all establish that AO random walk and the UCSV trend-cycle decomposition set a high bar that many multivariate models fail to clear. Reporting RMSEs against a UCSV benchmark (and not just AR or RW) is therefore the disciplined comparison.

Second, in the data-rich strand, the central finding of James H. Stock and Mark W. Watson (2002b) and James H Stock and Mark W Watson (2002a), Bernanke and Boivin (2003), Marcellino, James H Stock, and Mark W Watson (2003) and the Eickmeier and Ziegler (2008) meta-analysis is that factor-augmented forecasts deliver moderate but consistent gains, of the order of 10-25% in relative RMSE. These gains shrink during the Great Moderation and re-expand in volatile regimes. A DFM benchmark implemented along the lines of James H Stock and Mark W Watson (2002a) with Bai and Ng (2002) factor selection is therefore a natural intermediate rung between the univariate benchmarks and ML.

Third, the ML literature converges on a robust empirical regularity, discussed above: random forests and gradient boosting dominate in volatile and non-stationary regimes

(Medeiros, Vasconcelos, et al., 2021; Goulet Coulombe, Leroux, et al., 2022), while Vancsura, Tatay, and Bareith (2025) provide complementary deep-learning evidence for the euro area. Penalised regressions remain competitive in stable regimes, and the gains from nonlinearity exceed those from variable selection (Goulet Coulombe, Leroux, et al., 2022; Giannone, Lenza, and Primiceri, 2021), which justifies a model ladder organised by data richness (RW $\rightarrow$ UCSV $\rightarrow$ DFM $\rightarrow$ Ridge/LASSO) and then by nonlinearity (RF, XGBoost).

Fourth, the post-pandemic ML inflation literature Vancsura, Tatay, and Bareith (2025), Naghi, O'Neill, and Danielova Zaharieva (2024), Liu, Pan, and Xu (2024), and Simionescu (2025) delivers a more nuanced verdict than the pre-pandemic optimism: random forest dominance is not always robust to the 2020-2022 episode, and the comparative advantage of ML depends on country, feature set and choice of benchmark. Disaggregated bottom-up architectures Joseph et al. (2021), Latypov et al. (2024), and Berben, Rasiawan, and Winter (2025) have emerged as a state-of-the-art benchmark; while this thesis focuses on aggregate core HICP rather than such a disaggregated architecture, this is a natural direction for a possible future extension.

Fifth, formal forecast comparison must distinguish equal predictive ability Diebold and Mariano (1995) from time-varying relative performance Giacomini and Rossi (2010) and from set-valued model selection Hansen, Lunde, and Nason (2011). Sub-period evaluation across a pre-2020 / post-2020 split, combined with Diebold-Mariano and Giacomini-Rossi fluctuation tests, follows directly from the literature. SHAP-based interpretability for tree ensembles complements the comparison.

3 Methodology

This chapter sets out the forecasting framework evaluated in the remainder of the thesis. Every candidate model is cast as an approximation to the same reduced-form problem: given information available at the forecast origin t , denoted $\mathcal{I}_t$, produce a point forecast of the h -period-ahead target $\pi_{t+h}^{(h)}$ defined in Chapter 3.1,

$$\pi_{t+h}^{(h)} = f(\mathcal{I}_t; \theta) + \varepsilon_{t+h}^{(h)}, \quad \mathcal{I}_t \subseteq \{\pi_t, \pi_{t-1}, \dots\} \cup \{X_t, X_{t-1}, \dots\}, \quad (3.1)$$

where X_t is the panel of predictor series described in Chapter 4. The eight models compared in this thesis are distinguished by exactly three choices within this common structure: (i) the information set $\mathcal{I}_t$ they are permitted to draw on: either inflation history alone, or inflation history plus the full predictor panel; (ii) whether f is estimated freely or shrunk toward a simpler alternative; and (iii) whether f is restricted to be linear in its inputs. Table 3.1 summarises where each model sits along these three dimensions.

Organising the models this way is what allows the empirical results in Chapter 5 to be read as a decomposition rather than a simple horse race. The step from the random walk to the UCSV to the DFM isolates the value of *data richness*: each uses progressively more information, but all remain linear (or, for UCSV, univariate) and unregularised in the sense of imposing no variable-selection penalty. The step from the DFM to Ridge and LASSO isolates the value of *regularisation*: both use the same predictor panel as the DFM, but replace unsupervised factor extraction with a penalty estimated directly on the forecasting objective. The step from the linear ML models to Random Forest and XGBoost isolates the value of *nonlinearity*: both retain the full predictor panel and a comparable estimation philosophy to Ridge/LASSO, but relax the linearity restriction on f .

The remainder of this chapter is organised as follows. Chapter 3.1 defines the target variable and the four forecast horizons evaluated throughout. Chapter 3.2 sets out the three econometric benchmarks (Random Walk, UCSV, DFM). Chapter 3.3 sets out the five ML models (Ridge, LASSO, Elastic Net, Random Forest, XGBoost), with separate subsections focusing on interpretability of the models and hyperparameter and variable selection. Chapter 3.4 sets out the forecast evaluation framework: loss metrics, forecast combination, the Diebold-Mariano and Model Confidence Set procedures, the Giacomini-Rossi fluctuation test, and the choice of estimation window and sub-periods.

All data processing, model estimation and forecast evaluation described in this and the following chapters are conducted in R, with package versions pinned via `renv` to ensure

Table 3.1: Model Overview

Category	Model	Data	Estimation
Naive benchmark	Random Walk	Inflation only	None
Univariate benchmark	UCSV	Inflation only	Bayesian MCMC
Traditional multivariate	DFM	Large dataset	OLS + PCA
Linear ML	Ridge	Large dataset	GCV
Linear ML	LASSO	Large dataset	BIC
Linear ML	Elastic Net	Large dataset	BIC + GCV
Nonlinear ML	Random Forest	Large dataset	Fixed
Nonlinear ML	Gradient Boosting	Large dataset	Fixed + early stopping

reproducibility.

3.1 Target Variable and Forecast Horizons

The variable to be forecast is euro area core HICP inflation, defined as the Harmonised Index of Consumer Prices excluding energy and unprocessed food. Core inflation is the natural target for this study for two reasons. First, it is the measure most closely watched by the ECB in the context of its medium-term price stability mandate, as it filters out the volatile commodity-driven components and better captures the underlying trend in consumer prices (Bańbura and Bobeica, 2023).

Second, core inflation is the more informative test bed for the models compared in this thesis specifically because it is demand-driven rather than supply-driven: Capolongo and Pacella (2021) show that euro area core HICP carries a stronger domestic, country-specific component, while headline HICP is more exposed to global commodity-price shocks. A forecasting exercise built around a large panel of domestic macroeconomic indicators: labour-market slack, credit aggregates, wages, survey expectations, has little to say about the exogenous, globally-determined energy and food shocks that dominate headline inflation; whatever predictive signal such a panel contains should show up, if anywhere, in the component of inflation that domestic conditions and monetary policy actually influence. Core HICP strips supply shocks that are outside the central bank's control, thus making it a cleaner test of model forecasting ability for the policy-relevant underlying trend.

Following James H Stock and Mark W Watson (2002a), the target variable is constructed

as the h -period average of month-on-month annualised inflation:

$$\pi_{t+h}^{(h)} = \frac{1200}{h} \sum_{j=1}^h \ln\left(\frac{P_{t+j}}{P_{t+j-1}}\right), \quad (3.2)$$

where P_t denotes the core HICP index level in month t . The scalar 1,200 converts the log-difference from a monthly rate to an annualised percentage. For $h = 1$ this reduces to the standard month-on-month annualised rate; for $h \in \{3, 6, 12\}$ it yields the average annualised inflation rate over the next quarter, half-year, and year, respectively. This multi-step direct-forecast formulation follows James H. Stock and Mark W. Watson (2007) and is used, alongside an iterated alternative, in related work on euro area inflation by Bańbura and Bobeica (2023). It avoids error accumulation from chaining one-step-ahead forecasts and allows each model to be estimated directly on the target horizon.

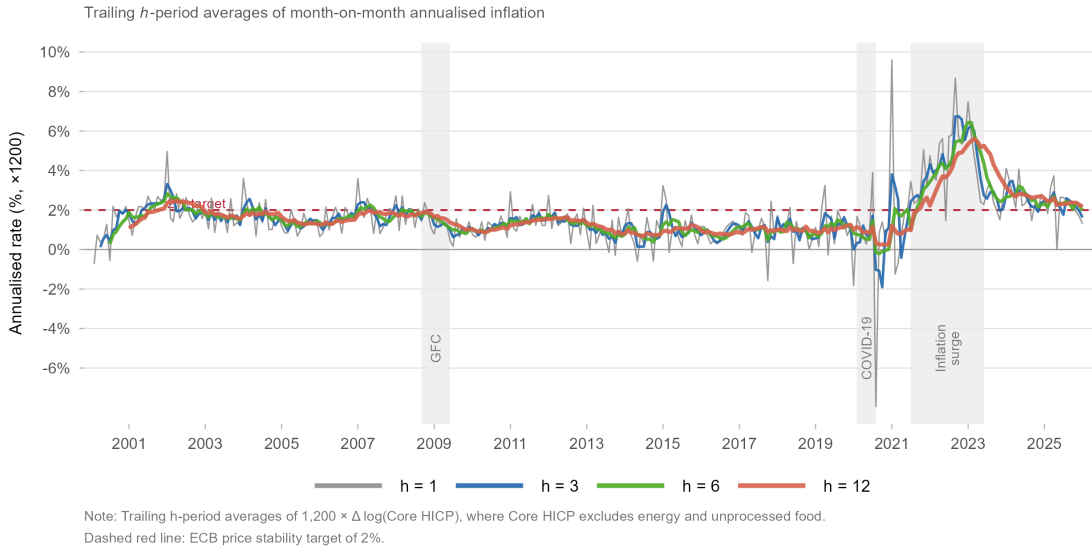


Figure 3.1: Euro Area HICP Inflation: Target Variable by Forecast Horizon

The use of month-on-month changes rather than year-on-year differences as the base rate is also motivated by the strong serial correlation inherent in twelve-month inflation: because eleven of the twelve monthly changes that compose the annual rate are already known at the forecast origin, year-on-year inflation is mechanically predictable and therefore an undemanding benchmark (Hall, Tavlas, and Wang, 2023). The present approach resolves this by building the multi-horizon target from month-on-month log-differences directly, aggregated via the h -period average in equation (3.2).

Forecasts are generated at four horizons: $h = 1, 3, 6$, and 12 months ahead. This choice

reflects both the monthly frequency of the data and the policy cycle of the ECB, which operates primarily at the one-to-four quarter horizon. Evaluating across multiple horizons is central to the thesis's contribution, as different model families are expected to exhibit comparative advantages at different horizons.

- $h = 1$ month ahead (short-run): The horizon at which high-frequency ML features (such as energy prices, financial conditions indices, and nowcasting-type indicators) are most informative. ML models are expected to hold their strongest advantage here.
- $h = 3$ months ahead (short-to-medium run): A key policy-relevant horizon. The Dynamic Factor Model (DFM), which pools information across many macroeconomic series, is expected to be competitive at this horizon.
- $h = 6$ months ahead (medium run): Added to provide finer resolution between the short- and long-run results, and to capture the horizon at which the UCSV's persistence properties are expected to dominate.
- $h = 12$ months ahead (long-run): The horizon at which the Stock-Watson UCSV and inflation expectations-based models typically perform best, reflecting the slow-moving trend component of inflation. ML performance is expected to deteriorate relative to shorter horizons.

Figure 3.1 illustrates the target variable across the four forecast horizons over the full sample period. Several features of the data are worth noting.

First, the degree of smoothing increases mechanically with the horizon: the one-month series ($h = 1$) exhibits the greatest volatility, while the twelve-month average ($h = 12$) traces the lower-frequency trend in underlying inflation. This property is a direct consequence of the averaging in equation (3.2) and implies that short-horizon forecasting is inherently a noisier task.

Second, from approximately 2003 to 2019 all four series remain close to the ECB's 2% price stability reference value, with subdued volatility interrupted only briefly around the Global Financial Crisis. This prolonged period of anchored inflation at low variance motivates the binary pre/post-2020 sub-period split adopted in the evaluation framework (Section 3): the post-2020 period constitutes a genuinely distinct regime in terms of both the level and volatility of inflation. Third, the COVID-19 shock in early 2020 produces a sharp transitory deflation at $h = 1$ that is largely absorbed at longer horizons, highlighting the sensitivity of short-horizon evaluation metrics to outliers during crisis episodes.

Finally, the inflation surge of 2021-2023 is visible across all horizons, though its peak is attenuated and delayed at longer averaging windows. The return of all series to approximately 2% by 2024-2025 closes the sample on a regime resembling the pre-2020 anchor, providing a natural bracket for the out-of-sample evaluation period.

3.2 Econometric models

3.2.1 Random walk

The naive benchmarks in this thesis follow the random walk tradition of Atkeson and Ohanian (2001), who show that a simple average of recent inflation outperforms Phillips-curve-based forecasts of US inflation at the one-year horizon. This model predicts that next period's inflation equals current inflation:

$$\hat{\pi}_{t+h}^{(h)} = \pi_t \quad (3.3)$$

This is a standard approach of macroeconomic forecasting exercises identified in the literature. Their result is frequently misquoted as single-period persistence ($\hat{\pi}_{t+h} = \pi_t$); in fact, their benchmark forecasts future inflation using the *average* inflation rate over the preceding four quarters, precisely because single-period persistence is an unreasonably weak benchmark once the underlying series is as volatile as month-on-month inflation.

This distinction matters for the present application. Month-on-month core HICP inflation, even after excluding energy and unprocessed food, is subject to substantial transitory noise from timing effects, seasonal residuals, and idiosyncratic price-setting behaviour in narrow expenditure categories. A forecast that simply carries forward the most recent monthly observation, $\hat{\pi}_{t+h}^{(h)} = \pi_t$, would therefore propagate this noise directly into the point forecast, producing an artificially weak benchmark that any model exploiting a modest amount of information could improve upon.

Instead, I construct two trailing-average random walk benchmarks, matching the logic of Atkeson and Ohanian (2001) to the multi-horizon direct-forecast design set out in later sections:

$$\hat{\pi}_{t+h, \text{RW}(3)}^{(h)} = \frac{1}{3} \sum_{j=0}^2 \pi_{t-j}, \quad (3.4)$$

$$\hat{\pi}_{t+h, \text{RW}(12)}^{(h)} = \frac{1}{12} \sum_{j=0}^{11} \pi_{t-j}. \quad (3.5)$$

Both benchmarks are horizon-invariant, which means the same trailing average is used as the point forecast for every $h \in \{1, 3, 6, 12\}$. This is because a pure random walk in the level of trend inflation implies no expected change over any horizon; only the window used to estimate that trend level differs between the two specifications.

The two windows are chosen to bracket the persistence properties relevant to the horizons under study. RW(3) uses a short trailing window and therefore tracks recent momentum in inflation, making it the more demanding benchmark at short horizons ($h = 1, 3$), where transitory dynamics still carry information. RW(12) uses a full trailing year and is close to a backward-looking proxy for year-on-year inflation, making it the more demanding benchmark at longer horizons ($h = 12$), where the h -period-average forecast target itself becomes closer to trend inflation and any information advantage from the current month's reading has largely averaged out. Reporting both benchmarks, rather than a single naive reference, avoids the risk of a model appearing to add value only because it was compared against an artificially weak, single-window random walk, and is consistent with the practice of setting multiple performance floors before evaluating more complex models.

3.2.2 UCSV

The UCSV model of James H. Stock and Mark W. Watson (2007) is another standard univariate benchmark in the inflation forecasting literature, as discussed in the Section 2.1.1. It decomposes inflation into a permanent trend component τ_t and a transitory component, where both feature independent stochastic volatility:

$$\pi_t = \tau_t + e^{\frac{1}{2}(h_0 + \omega_h \tilde{h}_t)} \varepsilon_t^\pi, \quad \varepsilon_t^\pi \sim \mathcal{N}(0, 1) \quad (3.6)$$

$$\tau_t = \tau_{t-1} + e^{\frac{1}{2}(g_0 + \omega_g \tilde{g}_t)} \varepsilon_t^\tau, \quad \varepsilon_t^\tau \sim \mathcal{N}(0, 1) \quad (3.7)$$

where the log-volatility states $\tilde{h}_t$ and $\tilde{g}_t$ each follow a random walk:

$$\tilde{h}_t = \tilde{h}_{t-1} + \varepsilon_t^h, \quad \varepsilon_t^h \sim \mathcal{N}(0, 1) \quad (3.8)$$

$$\tilde{g}_t = \tilde{g}_{t-1} + \varepsilon_t^g, \quad \varepsilon_t^g \sim \mathcal{N}(0, 1) \quad (3.9)$$

The parameters h_0 and g_0 are intercepts controlling the baseline level of volatility, while

ω_h and ω_g govern how much the volatility states $\tilde{h}_t$ and $\tilde{g}_t$ contribute to the overall variance of each component.

Following Chan (2018), this thesis implements the **non-centred reparameterisation** of the model. In the original centred parameterisation of James H. Stock and Mark W. Watson (2007), the log-volatilities enter the observation equations multiplicatively, which causes the MCMC sampler to mix poorly when the signal-to-noise ratio is low. This is precisely the situation that prevails during periods of stable, low inflation such as 2013–2019 in the euro area. The non-centred form separates the baseline volatility level (h_0, g_0) from the time-varying deviation $(\tilde{h}_t, \tilde{g}_t)$, which yields substantially better mixing and more reliable posterior inference at little computational cost.

The model is estimated via Bayesian MCMC using a Gibbs sampler. The point forecast is the median of simulated forward draws of $\hat{\pi}_{t+h|t}^{(h)} = \frac{1}{h} \sum_{i=1}^h \pi_{t+i|t}$. The stochastic volatility component allows the model to adapt automatically to changing inflation regimes, making it a particularly demanding benchmark that has proven difficult to beat in euro area applications (Bańbura and Bobeica, 2023). It sets the bar for what can be achieved using inflation data alone, and any multivariate model must clear this bar to justify its additional complexity.

3.2.3 Factor Model

This section sets out the factor model in four steps: the choice of a static, principal-components specification over a state-space alternative; the two-step estimation strategy and its asymptotic justification; the formal estimator and the selection of the number of factors; and the direct multi-step forecasting equation that uses the estimated factors as predictors.

This thesis estimates the DFM in its static form, with factors extracted nonparametrically by principal components, rather than via the parametric state-space/Kalman-filter alternative. The choice is motivated by Smeekes and Wijler (2018), who show that the Kalman-based maximum-likelihood estimator of Doz, Giannone, and Reichlin (2011) deteriorates relative to standard PC once the panel’s idiosyncratic component exhibits substantial cross-sectional correlation, a realistic feature of a macroeconomic dataset drawn from related institutional and sectoral aggregates. In their calibrated simulation, non-sphericity cuts the variance share explained by the first principal components from roughly 72% to 41%, degrading the parametric estimator while PC remains robust.

The two-step estimation strategy used throughout this analysis first extracts the common

factors $\hat{F}_t$ from X_t by principal components, and then treats $\hat{F}_t$ as observed data in the direct multi-step OLS forecasting regression below (Equation (3.11)), rather than estimating the factors and the forecasting equation jointly. The decisive theoretical justification for this two-step approach is the result of Bai and Ng (2006): as both N and T grow, the estimation error in $\hat{F}_t$ vanishes at a rate fast enough that the estimated factors can be treated as observed in subsequent regressions, without requiring a generated-regressor correction. This licenses the procedure of extracting factors by PC and then running OLS in the forecasting regression, while retaining the consistency and asymptotic normality of the OLS estimator.

More formally, the estimator solves the least-squares problem:

$$\min_{F_1, \dots, F_T, \Lambda} \frac{1}{NT} \sum_{t=1}^T (X_t - \Lambda F_t)' (X_t - \Lambda F_t) \quad (3.10)$$

subject to the normalisation $N^{-1}\Lambda'\Lambda = I_r$ and $F'F$ diagonal. When there are no missing observations the solution is the matrix of eigenvectors of the sample covariance matrix $\hat{\Sigma}_X = T^{-1} \sum_{t=1}^T X_t X_t'$ associated with its r largest eigenvalues, with factor estimates $\hat{F}_t = N^{-1} \hat{\Lambda}' X_t$.

The number of static factors r was assessed using the IC_{p2} criterion of Bai and Ng (2002), the standard choice across the forecasting literature consulted in this thesis, including Medeiros, Vasconcelos, et al. (2021) and Naghi, O'Neill, and Danielova Zaharieva (2024), and reported to perform well in macroeconomic simulation designs (James H. Stock and Mark W. Watson, 2016). For this panel, however, IC_{p1} , IC_{p2} , and IC_{p3} all decrease monotonically through $r = 40$ with no interior minimum: the penalty term never comes to dominate the reduction in residual variance, reflecting a panel whose cumulative variance share rises gradually (19% at $r = 1$, 71% at $r = 8$, 90% at $r = 20$) rather than plateauing, so the criterion fails to select r for this panel. r is therefore selected instead via the eigenvalue-ratio (ER) and growth-ratio (GR) estimators of Ahn and Horenstein (2013), which choose r as the value that maximises the ratio of consecutive eigenvalues (ER) or of consecutive remaining-variance terms (GR), rather than requiring an information-criterion penalty to dominate, and remain well defined under exactly this condition. The ER and GR estimators agree on $r = 4$ for the full-sample panel, which is the value used throughout.

One caveat applies to this choice: r is selected once from the full 2000–2026 panel and then held fixed at every recursive origin, rather than being re-selected from each origin's own

expanding window. This is a minor departure from the real-time-consistency principle enforced elsewhere in the recursive design (Section 4.4), since an origin in, say, 2011 could not have observed the post-2020 comovements that inform this selection. The likely impact is small, since ER and GR agree on the same r over the full sample and DFM accuracy is itself stable across the independent rolling-window check (Section 6.2), but re-selecting r recursively at each origin was not implemented here and is left as a direction for future robustness checking.

It is also important to note, that since the PC estimator is *nonparametric*, no distributional assumption is placed on e_t or η_t , and consistency holds under the weak conditions of the approximate factor model of Chamberlain and Rothschild (1982). This requires only that the cross-sectional correlation of the idiosyncratic terms remains bounded as $N \rightarrow \infty$. While the DFM *model* imposes a parametric factor structure on X_t , the principal components *estimation method* is nonparametric in that it makes no distributional claims about the data-generating process for the disturbances.

Multi-Step Forecasting

Given estimated factors $\hat{F}_t$, the h -step-ahead inflation forecast is obtained via a direct OLS regression:

$$\pi_{t+h}^{(h)} = \alpha^{(h)} + \beta^{(h)}(L) \hat{F}_t + \gamma^{(h)}(L) \pi_t^{(h)} + \varepsilon_{t+h}, \quad (3.11)$$

where $\pi_{t+h}^{(h)}$ denotes the h -period average of month-on-month annualised inflation scaled by 1,200, as defined in Section 3.1. A separate regression is estimated for each horizon $h \in \{1, 3, 6, 12\}$, and all parameters are re-estimated at each recursive origin t using only data available up to and including period t . Dynamics are introduced not through a state-space factor VAR, but explicitly, by including lags of the estimated factors and of the target variable in the forecasting equation itself.

$\alpha^{(h)}$ is a horizon-specific intercept. $\beta^{(h)}(L) = \beta_0^{(h)} + \beta_1^{(h)}L + \dots + \beta_{p_f-1}^{(h)}L^{p_f-1}$ is a lag polynomial of order p_f applied to the r -dimensional factor vector, so $\beta^{(h)}(L)\hat{F}_t$ contributes the contemporaneous factors $\hat{F}_t$ together with $p_f - 1$ further lags, $r \times p_f$ regressors in total. $\gamma^{(h)}(L)$ is an analogous scalar lag polynomial of order p_y applied to the target's own lags, $\pi_t^{(h)}, \pi_{t-1}^{(h)}, \dots, \pi_{t-p_y+1}^{(h)}$, giving the regression a direct-forecast autoregressive component alongside the factor information.

The lag orders $p_f = p_y$ are horizon-dependent rather than fixed across h : 4 for $h \leq 3$, 2 for $h \leq 6$, and 1 for $h = 12$. Shorter lag structures at longer horizons reflect the shrinking effective estimation sample as h grows (the usable regression sample runs only from $t = \max(p_f, p_y)$ to $t = T_{\text{est}} - h$ within a given origin's window) and the diminishing

marginal information content of additional lags once the target is already averaged h months forward. The four outlier-month dummies of Section 4.4 also enter the regression as additional regressors, one 0/1 indicator per flagged month once it falls inside the training range; they are suppressed from Eq. (3.11) for notational compactness but follow the same inclusion rule as elsewhere in the thesis. Each horizon-origin regression is solved by ordinary least squares (via a QR decomposition of the design matrix, for numerical stability), and the h -step forecast is produced by evaluating the fitted equation at the most recent p_f factor lags and p_y target lags available at origin t .

The direct approach is preferred over iterated forecasting, where the factor VAR (2.2) is stepped forward h periods. Marcellino, James H. Stock, and Mark W. Watson (2006) show that this choice involves a trade-off: iterated forecasts are more efficient when the underlying one-step model is correctly specified, since they extrapolate a single compact model forward, whereas direct forecasts sacrifice some of that efficiency in exchange for not depending on a correctly-specified, stable one-step dynamic model compounded across every intermediate step. Three considerations favour the direct approach in this setting:

1. **Robustness to structural breaks.** The recursive evaluation window spans clear structural breaks (the COVID-19 shock and the 2021–2023 energy-price surge), during which a fixed one-step factor VAR would be especially prone to misspecification if iterated forward.
2. **Computational tractability.** The recursive out-of-sample design re-estimates the factor structure at every forecast origin, across both expanding- and rolling-window specifications, which makes the closed-form PCA estimator considerably more tractable than the iterative EM/Kalman-filter estimation a full state-space DFM would require.
3. **Ragged-edge handling is not needed here.** The principal advantage of the state-space approach is its ability to handle ragged-edge data, with series arriving at different real-time release dates, as is typical in nowcasting applications. This feature is not directly relevant here, since the predictor panel is aligned to a common monthly frequency with missing values imputed prior to estimation.

Under the classical large- N , large- T conditions of Bai and Ng (2002) and James H. Stock and Mark W. Watson (2002b), static PCA yields consistent factor estimates without requiring the additional distributional assumptions or convergence properties of EM-based estimation, making it a suitable and computationally tractable choice for the

recursive forecasting exercise conducted in this thesis.

The DFM occupies an intermediate position in the model hierarchy. It exploits the same large information set as the ML models but imposes a strictly linear structure: predictive content is channelled entirely through the extracted factors, with no interactions or threshold effects permitted. The DFM extracts structure that is common across all variables, whereas the following ML models extract structure that is predictive of inflation specifically. If the ML models outperform the DFM, the excess gain is attributable to the ability to capture nonlinearities rather than simply exploiting a richer predictor panel. Conversely, if the DFM matches or surpasses the ML models, the evidence would suggest that the structure of euro area inflation is well approximated by linear factor dynamics, and the additional flexibility of ML methods does not compensate for their higher estimation variance.

3.3 ML models

The four ML models are estimated on the full monthly predictor panel described in Section 4, applying a direct multi-step approach at each horizon $h \in \{1, 3, 6, 12\}$.

For Ridge and LASSO, predictors are standardised to zero mean and unit variance within each recursive estimation window prior to fitting, following the approach from Naghi, O'Neill, and Danielova Zaharieva (2024). Standardisation is not applied for Random Forest and Gradient Boosting, since tree-based splits are invariant to monotonic rescaling of individual predictors.

The Factor model and ML models share the same predictor universe; the key distinction is that the Factor model summarises it through a small number of linear factors, while the ML models interact directly with the raw predictors and can capture nonlinear relationships.

3.3.1 Shrinkage methods

Section 2.2.1 set out the general case for shrinkage over unpenalised OLS in high dimensions, and the qualitative behaviour of the ℓ_1 and ℓ_2 penalties. This section gives the formal specification of Ridge, LASSO, and Elastic Net as implemented in this thesis, and works through one consequence of the ℓ_1/ℓ_2 distinction that is specific to the predictor panel assembled in this thesis: how each penalty handles its internal collinearity.

Unlike principal components, which extracts directions of maximum common variance

from the full predictor panel so that every factor absorbs the joint variation of *all* N series, LASSO operates directly on the predictor space and selects variables individually. Under high collinearity between two series, LASSO can arbitrarily discard one of the pair while retaining the other, even when both carry genuinely independent predictive content (Smeekes and Wijler, 2018). This is particularly relevant here, since the panel contains several conceptually overlapping series, discussed further in Chapter 4: ECB BSI loans to non-financial corporations, M3 outstanding, and SPF inflation expectations, among others, where pairwise correlations may be substantial. The practical consequence is that LASSO variable selection should be interpreted with caution: an excluded series does not imply irrelevance, only that a correlated substitute was retained. Ridge does not share this pathology, since its ℓ_2 penalty shrinks correlated predictors symmetrically rather than eliminating one arbitrarily.

Formally, Ridge solves the ℓ_2 -penalised least squares problem:

$$\hat{\beta}_h^{\text{Ridge}} = \arg \min_{\beta} \sum_t \left(\pi_{t+h}^{(h)} - x_t' \beta \right)^2 + \lambda \|\beta\|_2^2 \quad (3.12)$$

where x_t is the $p \times 1$ vector of standardised predictors at time t and $\lambda \geq 0$ is the shrinkage penalty. Ridge shrinks all coefficients toward zero while retaining every predictor, making it the natural linear ML benchmark: it uses the same raw information set as Random Forest and XGBoost but is constrained to a linear prediction surface.

LASSO replaces the ℓ_2 penalty with an ℓ_1 penalty:

$$\hat{\beta}_h^{\text{LASSO}} = \arg \min_{\beta} \sum_t \left(\pi_{t+h}^{(h)} - x_t' \beta \right)^2 + \lambda \|\beta\|_1 \quad (3.13)$$

The ℓ_1 penalty induces sparsity by shrinking irrelevant coefficients to exactly zero, performing automatic variable selection. Comparing LASSO to Ridge isolates the value of sparsity within the class of linear models.

Elastic Net combines both penalties via a mixing parameter $\alpha \in [0, 1]$, interpolating between LASSO's sparsity and Ridge's more graceful handling of highly correlated predictors:

$$\hat{\beta}_h^{\text{ENet}} = \arg \min_{\beta} \sum_t \left(\pi_{t+h}^{(h)} - x_t' \beta \right)^2 + \lambda \left[\alpha \|\beta\|_1 + (1 - \alpha) \|\beta\|_2^2 \right] \quad (3.14)$$

Setting $\alpha = 1$ recovers the LASSO penalty in (3.13), and $\alpha = 0$ recovers the Ridge penalty in (3.12).

3.3.2 Tree methods

Random Forest constructs an ensemble of B regression trees, each estimated on a bootstrap subsample of the training data and a random subset of predictors at each split. The forecast is the average across all trees:

$$\hat{\pi}_{t+h}^{(h)} = \frac{1}{B} \sum_{b=1}^B T_b(x_t; \Theta_b) \quad (3.15)$$

where $T_b(\cdot; \Theta_b)$ denotes the b -th tree with randomisation parameters Θ_b . By averaging over many trees estimated on different subsamples, Random Forest reduces variance without requiring any assumption about the functional form of the relationship between predictors and inflation. It can capture interactions and threshold effects that are invisible to linear models.

Rather than averaging independent trees, Gradient Boosting (XGBoost) builds an additive ensemble sequentially, with each tree fitted to the residuals of the current ensemble:

$$\hat{\pi}_{t+h}^{(h)} = \sum_{b=1}^B \eta \cdot T_b(x_t) \quad (3.16)$$

where η is the learning rate controlling the contribution of each successive tree. The objective combines a differentiable loss function with ℓ_1 and ℓ_2 regularisation on the tree weights, providing protection against overfitting. XGBoost is included as the second nonlinear model to test whether the sequential, residual-focused learning mechanism offers advantages over the parallel averaging of Random Forest.

3.3.3 Hyperparameters

Rather than cross-validation, tuning follows the approach of Medeiros, Vasconcelos, et al. (2021). For LASSO, λ is chosen to minimise the Bayesian Information Criterion (BIC),

$$\text{BIC}(\lambda) = n \log \left(\frac{\text{RSS}(\lambda)}{n} \right) + \text{df}(\lambda) \log n, \quad (3.17)$$

where $\text{df}(\lambda)$ is the number of nonzero coefficients at that λ (Kock and Callot, 2015; Medeiros and Mendes, 2016). This requires a single path fit rather than repeated refitting across folds. Ridge never produces exact zeros, so a nonzero-coefficient BIC is uninformative for it ($\text{df}(\lambda) = p$ at every λ); λ is instead chosen by generalised cross-validation (GCV) using the classical SVD-based effective degrees of freedom,

$$\text{df}(\lambda) = \sum_{i=1}^p \frac{d_i^2}{d_i^2 + n\lambda}, \quad (3.18)$$

where d_i are the singular values of the standardised design matrix, and λ is chosen to minimise

$$\text{GCV}(\lambda) = \frac{\text{RSS}(\lambda)/n}{(1 - \text{df}(\lambda)/n)^2}. \quad (3.19)$$

This is the natural analogue of the BIC rule above for a penalty that does not induce sparsity. Medeiros, Vasconcelos, et al. (2021) also do not extend their BIC rule to ridge, since it is scoped to "the LASSO family".

For Elastic Net, α and λ are selected jointly rather than α being fixed *a priori*: for each α on a grid $\alpha \in \{0.25, 0.5, 0.75\}$, λ is chosen by the same BIC rule as LASSO above: `glmnet`'s nonzero-coefficient degrees of freedom carries over unchanged to any $\alpha > 0$ and the (α, λ) pair attaining the lowest BIC across the entire grid is used. This keeps Elastic Net's selection in the same single-path-fit, no-cross-validation spirit as LASSO and Ridge, just extended across the mixing parameter.

Table 3.2 summarises what the BIC/GCV selection procedures above actually chose, averaged over the 182 recursive origins (2010-01 to 2025-02) at each core horizon.

LASSO's selected model grows steadily less sparse with the horizon: from a mean of 4.8 nonzero coefficients at $h = 1$ (as low as 0 at some origins) to 19.7 at $h = 12$, alongside a roughly fifteen-fold decline in the mean selected λ (0.155 to 0.010). Ridge, in contrast, retains most of the panel's effective dimensionality throughout: mean effective degrees of freedom rise from 28.4 at $h = 1$ to 43.8 at $h = 12$, so its selected λ mainly rescales coefficient magnitudes rather than meaningfully reducing model complexity.

Elastic Net's selected α is heavily concentrated at 0.75 across origins (95.1% of origins at $h = 1$, 99.5% at $h = 3$, 97.3% at $h = 6$, 98.4% at $h = 12$; the remainder select $\alpha = 0.5$, and only at $h = 1$ occasionally $\alpha = 0.25$). Its resulting mean model size (5.1 to 21.0 nonzero coefficients across horizons) tracks LASSO's own closely and sits far below Ridge's effective dimensionality. The data consistently prefers the sparse end of the mixing grid, not the dense one.

For Random Forest, hyperparameters are fixed rather than tuned, for the same reason cross-validation is avoided for the shrinkage methods above: the recursive expanding

	$h = 1$	$h = 3$	$h = 6$	$h = 12$
LASSO: mean nonzero coefs.	4.8	12.8	13.6	19.7
LASSO: mean λ	0.155	0.040	0.028	0.010
Ridge: mean effective df	28.4	40.7	43.9	43.8
Ridge: mean λ	0.542	0.120	0.043	0.036
Elastic Net: modal α	0.75	0.75	0.75	0.75
Elastic Net: % origins at modal α	95.1	99.5	97.3	98.4
Elastic Net: mean nonzero coefs.	5.1	14.4	14.6	21.0

Table 3.2: Mean selected penalty and model size by horizon

Averaged over 182 recursive origins (2010-01 to 2025-02). LASSO/Elastic Net degrees of freedom are counts of nonzero coefficients; Ridge reports SVD-based effective degrees of freedom, since no Ridge coefficient is ever set to exactly zero.

window starts thin, with a rows-to-predictors ratio as low as 1.6 : 1 in the earliest (2010) estimation window, leaving little scope to split the sample into folds without estimation becoming unreliable. Settings follow Medeiros, Vasconcelos, et al. (2021): $mtry = \lfloor K/3 \rfloor$ (the `randomForest` R package's own regression default), a minimum leaf size of five observations, and 500 trees. Medeiros, Vasconcelos, et al. (2021) report identical settings and find results "reasonably stable" across the alternative combinations they tried, so no further tuning is attempted here.

Table 3.3: Hyperparameters used for ML models

Model	Hyperparameter	Value	Selection method
LASSO	penalty, α	1	Fixed
	penalty strength, λ	—	BIC, df = nonzero coefficient count
	target lags, p_y	4 / 2 / 1	Fixed, by horizon ($h \leq 3 / 4 \leq h \leq 6 / h > 6$)
	standardisation	yes	Fixed
	outlier dummies	4 dates	Fixed, added once observed with training support
Ridge	penalty, α	0	Fixed
	penalty strength, λ	—	GCV, SVD-based effective degrees of freedom
	target lags, p_y	4 / 2 / 1	Fixed, by horizon ($h \leq 3 / 4 \leq h \leq 6 / h > 6$)
	standardisation	yes	Fixed

(continued on next page)

(continued from previous page)

Model	Hyperparameter	Value	Selection method
	outlier dummies	4 dates	Fixed, added once observed with training support
Elastic Net			
	mixing parameter, α	{0.25, 0.5, 0.75}	Grid, selected jointly with λ by global BIC
	penalty strength, λ	—	BIC (per α), df = nonzero coefficient count
	target lags, p_y	4 / 2 / 1	Fixed, by horizon ($h \leq 3$ / $4 \leq h \leq 6$ / $h > 6$)
	standardisation	yes	Fixed
	outlier dummies	4 dates	Fixed, added once observed with training support
Random Forest			
	number of trees	500	Fixed (Medeiros, Vasconcelos, et al., 2021)
	mtry	$\lfloor K/3 \rfloor$	Fixed, package default (Medeiros, Vasconcelos, et al., 2021)
	minimum node size	5	Fixed (Medeiros, Vasconcelos, et al., 2021)
	maximum tree depth	unconstrained	Fixed; a cap of 10 or 20 would not bind
	target lags, p_y	4 / 2 / 1	Fixed, by horizon ($h \leq 3$ / $4 \leq h \leq 6$ / $h > 6$)
	outlier dummies	4 dates	Fixed, added once observed with training support
XGBoost			
	maximum tree depth	3	Fixed
	learning rate, η	0.10	Fixed
	subsample ratio	0.8	Fixed
	column subsample ratio	0.8	Fixed
	boosting rounds	≤ 500	Early stopping (20 rounds patience) on a trailing 15% time-ordered validation slice; refit on full sample
	target lags, p_y	4 / 2 / 1	Fixed, by horizon ($h \leq 3$ / $4 \leq h \leq 6$ / $h > 6$)
	outlier dummies	4 dates	Fixed, added once observed with training support

3.3.4 Variable Selection and Interpretability: SHAP and LIME

Variable importance is computed for the five models with a per-predictor structure: Ridge, Lasso, Elastic Net, Random Forest, XGBoost and not RW, UCSV, or the DFM, whose forecasts do not decompose onto individual raw predictors in the same way. Random Forest's and XGBoost's importance measures inherit the same two limitations noted for out-of-bag permutation importance in Section 2.2.3: they are unsigned, and they summarise the whole estimation sample at once, with no per-observation decomposition.

Ridge, Lasso and Elastic Net’s coefficient-based measure (below) already carries a sign, and, since the model is linear, a local effect identical to its global one, so SHAP and LIME are applied only to Random Forest and XGBoost. Hyperparameters and target lags p_y throughout this section are unchanged from Section 3.3.3 (Table 3.3). The one departure from the rest of this thesis is that every model here, for the importance measure and for the SHAP/LIME decomposition alike, is fit *once* on the full available sample rather than recursively at each expanding-window origin: these fits exist purely to be read, and are distinct from the models that produced the forecasts evaluated in Chapter 5 and Chapter 6.

For Ridge, Lasso and Elastic Net, predictor j ’s importance at horizon h is its fitted coefficient’s absolute value, rescaled by the predictor’s own sample standard deviation:

$$\text{Imp}_{h,j} = \left| \hat{\beta}_{h,j} \right| \times \text{sd}(x_j), \quad (3.20)$$

where $\hat{\beta}_{h,j}$ is the j -th element of whichever coefficient vector ($\hat{\beta}_h^{\text{Ridge}}$, $\hat{\beta}_h^{\text{LASSO}}$, or $\hat{\beta}_h^{\text{ENet}}$) was selected at horizon h . For Random Forest, importance is impurity-corrected importance (Nembrini et al., 2018), a bias-corrected variant of the usual variance-reduction impurity measure that can take small negative values for a pure-noise predictor rather than the spuriously positive value uncorrected impurity importance would assign it. For XGBoost, importance is *gain*: the total reduction in training loss attributable to splits on that predictor, summed across all trees. These are aggregated into nine categories (own lags, employment, consumption, interest and exchange rates, output/income/housing, money and credit, prices and wages, expectations, and stocks), clipped at zero (impurity-corrected importance can otherwise contribute a small negative share) and rescaled to sum to one within each panel, stacked as a share across the full $h = 1, \dots, 12$ profile (Section 5.2).

SHAP Scott M Lundberg and Lee (2017) explains a prediction as an additive attribution over predictors,

$$\hat{\pi}_{t+h}^{(h)} = \phi_0 + \sum_{i=1}^M \phi_i(t), \quad (3.21)$$

where ϕ_0 is the model’s expected output and $\phi_i(t)$ is predictor i ’s attributed contribution at origin t . Local accuracy (Equation (3.21) holding exactly), missingness, and consistency together pin down a unique solution for $\phi_i(t)$: the Shapley value from cooperative game theory,

$$\phi_i(t) = \sum_{S \subseteq F \setminus \{i\}} \frac{|S|! (|F| - |S| - 1)!}{|F|!} \left[f_{x_t}(S \cup \{i\}) - f_{x_t}(S) \right], \quad (3.22)$$

where F is the full predictor set and $f_{x_t}(S)$ is the model’s expected output at x_t given

only the predictors in coalition S are known. Exact enumeration of (3.22) is $O(2^{|F|})$ and infeasible for $M \approx 60$ predictors; this thesis instead uses TreeSHAP (Scott M. Lundberg et al., 2020), the exact, polynomial-time variant of (3.22) for tree ensembles, rather than the slower, sampling-based Kernel SHAP the same paper proposes for arbitrary models. The R `treeshap` package's `ranger.unify()/xgboost.unify()` convert a fitted Random Forest or XGBoost object into a common tree representation, from which $\phi_i(t)$ is obtained exactly for every predictor i and origin t ; ϕ_0 is taken as the fitted model's own mean prediction over the full sample, and (3.21) holding exactly for every t is a direct check on the computation itself. Global importance at each horizon is

$$|\overline{\phi_i}| = \frac{1}{T} \sum_{t=1}^T |\phi_i(t)|, \quad (3.23)$$

magnitude only, since a predictor whose effect flips sign with its own level can average toward zero under a signed measure.

LIME Ribeiro, Singh, and Guestrin (2016) instead fits a local surrogate model around a single instance x_t :

$$\xi(x_t) = \arg \min_{g \in G} \mathcal{L}(f, g, \pi_{x_t}) + \Omega(g), \quad (3.24)$$

where G is a class of interpretable models (here, sparse linear regression), $\pi_{x_t}(z)$ weights a perturbed sample z by its proximity to x_t , $\mathcal{L}$ is the resulting locally-weighted squared loss, and $\Omega(g)$ penalises complexity. In practice, $g(z) = \beta_0 + \sum_{k \in K} \beta_k z_k$ is fit on N perturbed samples drawn around x_t , and β_k is reported as predictor k 's "LIME weight". This is implemented via the R `lime` package, which departs from Ribeiro, Singh, and Guestrin (2016)'s original algorithm in three respects, each made necessary by a panel of continuous macroeconomic predictors with no natural discretisation: perturbation is continuous rather than binary (`bin_continuous = FALSE`); $\pi_{x_t}(z)$ is a Gower-distance kernel rather than the paper's exponential kernel; and feature selection (10 features requested) uses `lime`'s ridge-based *highest-weights* rule (a single ridge fit on all predictors, keeping only the K largest absolute coefficients), rather than the paper's K -LASSO, which instead uses a Lasso regularisation path to select the K surviving predictors before refitting their weights by unpenalised least squares. Predictors are z-scored before this fit, via a wrapper that un-scales back to the model's native units before calling `predict()`, leaving the underlying Random Forest/XGBoost fit untouched: left unstandardised, the β_k in (3.24) are mechanically biased toward whichever predictor has the smallest native scale. Level/rate predictors here have $\text{sd} \approx 1\text{--}3$, month-on-month growth-rate predictors have $\text{sd} \approx 0.002\text{--}0.004$ independent of relevance, so standardisation is necessary for

the resulting weights and the highest-weights selection step, to be comparable across predictors. Whether LIME's selected predictors actually vary across the same dates and horizons as the local SHAP decompositions, or stay similar regardless of which month is explained (the specific failure mode Aras and Lisboa (2022) attribute to LIME's non-axiomatic parameter choices) is examined in Section 5.

3.4 Evaluation

3.4.1 Metrics and Forecast Combination

To quantify the predictive performance of the various models, this thesis employs three standard loss functions: **Root Mean Squared Forecast Error (RMSFE)**, **Mean Absolute Error (MAE)**, and **Median Absolute Deviation (MAD)**. RMSFE is the most common metric in the forecasting literature. It is calculated as the square root of the average squared difference between the forecasted value ($\hat{\pi}_{t+h|t}$) and the realized observation (π_{t+h}):

$$RMSFE_h = \sqrt{\frac{1}{T - T_0 + 1} \sum_{t=T_0-h}^{T-h} (\hat{\pi}_{t+h|t} - \pi_{t+h})^2} \quad (3.25)$$

The quadratic nature of the loss function means that the RMSFE penalizes large forecast errors more heavily. This makes it a critical metric for evaluating performance during high-volatility periods, such as the 2021-2022 inflation surge, where "outlier" errors can have significant policy implications making it sensitive to the type of outlier errors that characterised the 2021-2022 Eurozone inflation surge.

The MAE provides a linear measure of accuracy by averaging the absolute differences between forecasts and realizations:

$$MAE_h = \frac{1}{T - T_0 + 1} \sum_{t=T_0-h}^{T-h} |\hat{\pi}_{t+h|t} - \pi_{t+h}| \quad (3.26)$$

Unlike the RMSFE, the MAE is more robust to outliers and inflation shocks, providing a more balanced view of typical model performance. While RMSFE and MAE summarise average performance, both remain sensitive to a small number of unusually large errors. Following Medeiros, Vasconcelos, et al. (2021), I also report the **median absolute deviation (MAD)** of the forecast errors from their own median:

$$MAD_h = \text{median}(|e_{t+h,h} - \text{median}(e_{t+h,h})|), \quad e_{t+h,h} = \pi_{t+h} - \hat{\pi}_{t+h|t} \quad (3.27)$$

where the median is taken over $t = T_0 - h, \dots, T - h$. Unlike RMSFE and MAE, MAD is robust to outliers and asymmetries in the error distribution. Reporting RMSFE, MAE, and MAD together helps confirm that the ranking of models reflects broadly consistent performance, rather than being driven by a handful of large forecasting errors.

While accuracy metrics provide a ranking, the **Diebold-Mariano (DM) test** (Diebold and Mariano, 1995) is required to determine if the differences in performance between two models are statistically significant. The test evaluates the null hypothesis of equal predictive ability between a challenger model and a benchmark (such as the Random Walk or UCSV). The DM statistic is a t -test calculated using the mean of the loss differentials between two models, utilizing Newey-West HAC standard errors to account for serial correlation in multi-step forecasts. This test is primarily valid for non-nested models and is most reliable when applied to out-of-sample forecasts obtained via a rolling scheme.

Additionally, I also implement the **Model Confidence Set (MCS)** procedure of Hansen, Lunde, and Nason (2011), which tests the full set of models jointly rather than pair by pair. The MCS is a sequential testing procedure that starts with the full set of models and, at each step, tests the null hypothesis of equal predictive ability across the surviving set; if rejected, the model contributing most to that rejection is eliminated and the test is repeated on the reduced set. This continues until the null fails to be rejected for all remaining models. I use the $t_{\max}$ statistic for this equivalence test, chosen over the range statistic T_R that Hansen, Lunde, and Nason (2011) propose for its higher power when the candidate set is small, with a significance level of $\alpha = 0.10$ and $B = 5000$ bootstrap replications. The result is a set of statistically equivalent models that provides a more honest representation of model uncertainty than a simple point-estimate ranking.

Given the structural breaks introduced by the pandemic and the subsequent inflation regime shift, a constant-parameter evaluation may be misleading. I utilize the **Giacomini-Rossi fluctuation test** (Giacomini and Rossi, 2010) to assess how the relative performance of models evolves over time. This test is computed over a rolling window of m observations. Formally, the test statistic is a rolling, Newey-West-standardised Diebold-Mariano statistic,

$$\hat{t}(\tau) = \frac{1}{\sqrt{m} \hat{\sigma}} \sum_{s=\tau-m+1}^{\tau} (L_{1,s} - L_{2,s}), \quad L_{i,s} = e_{i,s}^2 \quad (3.28)$$

computed over a rolling window of m observations, where $L_{i,s}$ is the squared forecast error of model i at origin s and $\hat{\sigma}$ is a Newey-West long-run variance estimate of the loss differential. Rolling the window across all admissible origins τ and taking the maximum

absolute value yields the **sup-statistic**,

$$S = \sup_{\tau} |\hat{t}(\tau)|, \quad (3.29)$$

and the null hypothesis of stable relative performance is rejected if S exceeds the critical value tabulated in Giacomini and Rossi (2010) for the relevant $\pi = m/T$ ratio, where T is the number of common out-of-sample origins. It allows for the detection of episodes where machine learning models might significantly outperform statistical benchmarks, even if their average performance over the full sample is similar. This is particularly valuable for identifying "puzzles" like the "missing inflation" or "missing disinflation" periods, where traditional models often fail.

Following Giacomini and Rossi (2010), the window length is set as a fixed fraction of the evaluation sample, $m = \max(12, \text{round}(0.4 \times T))$, where T is the number of common out-of-sample origins across all candidates. For the $T = 182$ -origin evaluation sample used throughout this thesis, this gives $m = 73$ months (just over six years) and $\pi = m/T \approx 0.40$. The corresponding critical value is obtained by linear interpolation, over the π grid tabulated in Giacomini and Rossi (2010), of the critical values at the 10%, 5%, and 1% significance levels; at $\pi \approx 0.40$ this yields a 5% critical value of approximately 2.67. All rejection decisions reported in Section 5.3 are evaluated at the 5% level. To ease comparison across candidates and horizons, results there are reported as the ratio

$$\text{ratio} = \frac{\sup_{\tau} |\hat{t}(\tau)|}{\text{CV}_{5\%}}, \quad (3.30)$$

where a ratio above 1 indicates that the sup-statistic exceeds the 5% critical value, i.e. rejection of the null of stable relative performance, and a ratio below 1 indicates non-rejection.

In addition I also explore **Forecast Combination** as a strategy to improve robustness and hedge against model uncertainty, since diversifying across multiple models often yields more stable forecasts than relying on a single specification (Araujo and Gaglianone, 2023; Doojav and Narmandakh, 2025). I implement the following:

- **Simple Average:** the unweighted mean of the M candidate forecasts at each origin t and horizon h ,

$$\hat{\pi}_{t+h}^{(h),\text{Avg}} = \frac{1}{M} \sum_{m=1}^M \hat{\pi}_{t+h}^{(h),m}. \quad (3.31)$$

- **Trimmed Average:** the mean after discarding the single highest and lowest

forecast at each origin, so that the combination is not distorted by whichever individual model happens to be furthest from the pack at that point in time,

$$\hat{\pi}_{t+h}^{(h),\text{Trim}} = \frac{1}{M-2} \sum_{m \in S_t} \hat{\pi}_{t+h}^{(h),m}, \quad (3.32)$$

where S_t excludes the maximum and minimum of $\{\hat{\pi}_{t+h}^{(h),1}, \dots, \hat{\pi}_{t+h}^{(h),M}\}$ at that origin.

- **Median:** the median across the M candidate forecasts,

$$\hat{\pi}_{t+h}^{(h),\text{Med}} = \text{median} \left(\hat{\pi}_{t+h}^{(h),1}, \dots, \hat{\pi}_{t+h}^{(h),M} \right), \quad (3.33)$$

which, like the trimmed average, is robust to a single divergent forecast, but without discarding as much information from the remaining models.

Following the approach of Medeiros, Vasconcelos, et al. (2021), all three schemes are recomputed at every recursive origin over the $M = 5$ candidate models (DFM, LASSO, Ridge, Elastic Net, Random Forest, and XGBoost). The univariate benchmarks (UCSV, RW(3m), RW(12m)) are combined *against* in the evaluation framework rather than combined *into* the ensemble, since their role is to set a performance floor rather than contribute forecasting information.

3.4.2 Estimation Windows and sub-periods

To evaluate the predictive power of machine learning and traditional econometric models, this thesis simulates a real-time forecasting environment using specific estimation schemes. The choice of estimation scheme is a fundamental component of any real-time forecasting exercise, as it dictates the volume and relevance of historical information available to models at any given point in time. In this thesis I adopt an **expanding window (recursive) scheme** as the **primary estimation framework** and a **rolling window scheme** for **robustness checks**.

In an expanding window design, the estimation sample begins in January 2000 and expands forward with each new observation. Out-of-sample evaluation begins from January 2010, after an initial ten-year burn-in period, providing approximately 15 years of forecast evaluation through end-2024. For each month in the out-of-sample period, only the information available up to that specific forecast origin is used, thereby respecting real-time availability and avoiding look-ahead bias. This is the standard design in the academic forecasting literature and ensures direct comparability with Medeiros, Vasconcelos, et al. (2021) and Naghi, O'Neill, and Danielova Zaharieva (2024). The expanding window makes

full use of all available history, which is particularly valuable for the tree-based models whose splits become more reliable with larger training samples (Buckmann and Joseph, 2022). Its main limitation is that it is slow to adapt to structural breaks: observations from the extended low-inflation period of 2013-2019 continue to influence parameter estimates even after the 2021 regime shift.

As a secondary check, I implement a rolling window scheme with a fixed length of 120 months. This length is chosen to match exactly the size of the very first expanding-window estimation sample (January 2000 – December 2009): both schemes therefore begin the out-of-sample evaluation period from an *identical* estimation sample and diverge only in how that sample evolves thereafter, with the expanding window continuing to accumulate history while the rolling window discards the oldest observation each time a new one is added. This design isolates the effect of window *growth* from the effect of window *length*, rather than confounding the two by starting the rolling scheme from a shorter initial sample than the expanding scheme.

The debate between expanding and rolling windows is not about which is universally "correct," but rather which set of assumptions better fits the data-generating process at a given time. The expanding window prioritizes statistical efficiency, using all data to filter out transitory noise in stationary environments (Angelopoulos et al., 2025). Alternatively, the rolling window prioritizes adaptability and providing protection against regime shifts where "long memory" become a liability (Angelopoulos et al., 2025). If ML models perform substantially better under the rolling window than under the expanding window, it would suggest that their advantage stems from adaptability to structural change rather than from modelling power per se. Conversely, if the relative ranking of models is stable across the two schemes, the results are more likely to reflect genuine differences in forecasting ability that are robust to the choice of estimation window.

To examine whether model performance varies across inflation regimes, the out-of-sample evaluation period is also split into two sub-periods: a **pre-2020 period** (January 2010 - December 2019), covering the low and stable inflation environment following the global financial crisis, and a **post-2020 period** (January 2020 - end of sample), covering the pandemic disruption and the subsequent inflation surge and disinflation.

4 Data

This chapter introduces the dataset and variables used in the empirical analysis and also provides basic summary statistics. Section 4.1 introduces the primary macroeconomic dataset. Section 4.2 presents the supplementary series appended to the core dataset. Section 4.3 describes the preprocessing applied to merge and transform the two data blocks into a single monthly panel. Section 4.4 discusses the extent of missing data and the real-time-consistent procedure used to handle it within the recursive estimation loop. Section 4.5 describes the treatment of outlying observations in the target series.

This thesis uses aggregate data for euro area. It is therefore expected, that this modeling approach doesn't account for some country-specific drivers which are important for explaining inflation. As Menz (2024) points out, the aggregate specification neglects the country dimension.

4.1 Core Dataset: EA-MD-QD

The core of the predictor panel is the EA-MD-QD database (February 2026 vintage, Zenodo record 18804061), a FRED-MD-style monthly macroeconomic panel for the euro area maintained in the tradition of Barigozzi and Lissona (2026). The raw database is processed through the toolkit's own MATLAB preprocessing routine (`routine_data.m`) with settings `country = EA`, `frequency = M`, `transformation = light`, and `imputation = none`. Under the light-transformation setting, every series is differenced according to standard unit-root tests, with the exception of interest rates and the unemployment rate, which are retained in levels; any series found to be $I(2)$ is differenced twice. With imputation switched off, no outlier correction or missing-value filling is performed at this stage. Ragged edges and any gaps introduced by differencing are left as missing values and carried forward into the panel untouched. This yields a pre-transformed, already-stationary block of macroeconomic and price series (industrial production and turnover by category, producer and consumer price sub-indices, survey confidence indicators, monetary aggregates, interest rates, the unemployment rate, and the exchange rate, among others).

4.2 Supplementary Series

To broaden the information set beyond the EA-MD-QD coverage, a number of supplementary monthly series are appended. These are collected separately from four sources: the **ECB Statistical Data Warehouse** (credit aggregates to non-financial corporations, households and the private sector; M3; the Composite Indicator of Systemic Stress, CISS; the ECB’s non-energy commodity price index; Survey of Professional Forecasters inflation expectations at the 1-year, 2-year and long-term horizons; negotiated wage growth; government expenditure; total industrial production), **Eurostat** (capacity utilisation, short-term consumer inflation expectations, and import prices in industry), **FRED** (Brent crude oil, a European natural gas price proxy, and the German 10-year government bond yield), and two standalone sources: the **Global Supply Chain Pressure Index** constructed by Federal Reserve Bank of New York (2026) and the **Holston-Laubach-Williams natural-rate/output-gap estimates** (2023 vintage) by Holston, Laubach, and Williams (2023). A shadow policy rate (Wu-Xia/Krippner style) by Wu and Xia (2017) is also included; this series’ underlying source stops in April 2021 and was not extended with a post-2021 proxy, a limitation discussed in Section 4.4.

4.3 Preprocessing

The EA-MD-QD block and the supplementary block are merged into a single monthly panel, `panel_X`, spanning January 2000 to February 2026 (314 monthly observations) and containing 62 predictor series (two raw HICP level series, used only to construct the target above, are explicitly excluded from the predictor set to avoid using the target’s own level as a regressor).

The supplementary block mixes monthly and quarterly source series. Quarterly series: the three SPF expectations horizons, negotiated wage growth, capacity utilisation, government expenditure, and the HLW output gap from Holston, Laubach, and Williams (2023), are forward-filled (last observation carried forward) onto the monthly date grid before merging, so that within a given quarter all three months carry that quarter’s most recently released value. This is because a forecaster in month 2 of a quarter genuinely only has the previous quarter’s released figure available.

Unlike the EA-MD-QD block, the supplementary series arrive in their natural units and are not pre-transformed. A lightweight transformation is applied at merge time, following the same logic as the EA-MD-QD light-transformation setting described in Section 4.1: series that are level-nonstationary stocks or flows (loans to non-financial corporations, loans

to households, total private loans, M3, the non-energy commodity price index, import prices, Brent crude, the gas price proxy, and total industrial production) are converted to monthly percentage changes, $100 \times \Delta \log(x_t)$. The remaining series (the CISS, capacity utilisation, the shadow rate, a short-term inflation expectations indicator, the Global Supply Chain Pressure Index, negotiated wage growth, government expenditure as a share of GDP, the HLW output gap, and the German government bond yield) are already rates, ratios, gaps, indices, or standardised survey balances and are kept in levels. As with the EA-MD-QD block, no imputation is performed at this stage: gaps in the monthly supplementary series that are not addressed by the quarterly forward-fill above are left as missing values and carried into `panel_X` untouched, to be handled by the column-quality filter and imputation described in Section 4.4.

4.4 Missing Data

Forty-six of the 62 predictor series (70%) contain at least one missing observation. The large majority are short ragged edges of two to four months at the start of a series, arising from heterogeneous series start dates combined with the EA-MD-QD toolkit's no-imputation setting. Three series carry more substantial gaps: the shadow policy rate (114 missing months, 36% of the sample, since its underlying source ends in April 2021 and the panel extends to 2026), import prices in industry (87 missing months, 28%), and government expenditure (33 missing months, 11%, a consequence of its quarterly source frequency combined with a later start date than the rest of the panel).

These gaps are deliberately *not* imputed at the panel construction stage. Instead, missingness is handled inside the recursive estimation loop itself. At every forecast origin, a column-quality filter drops any predictor with more than 30% missing observations *within that origin's own estimation window*, drops any near-constant predictor (standard deviation below 10^{-8}), and then mean-imputes any remaining missing values using only the data available up to that origin. In practice this means a series such as the shadow rate remains part of the predictor set for the earlier part of the recursive evaluation and is automatically dropped once the growing estimation window pushes its within-window missing share past 30%. This is a data-driven, real-time-consistent alternative to hand-curating an inclusion list, applied identically to the factor model and to every machine-learning model.

4.5 Outlier Treatment

Four calendar months are flagged as extreme observations in the target series π_t (standardised residual $|z| > 3$): August 2020, January 2021, September 2022, and January 2023, corresponding to pandemic-era base effects and measurement disruptions and to the most acute prints of the 2022-2023 energy-price shock. Rather than winsorising or removing these observations, each is handled with a dedicated dummy regressor added to the direct-forecast design matrix, following the outlier-dummy treatment in Medeiros, Vasconcelos, et al. (2021): a single 0/1 indicator per flagged month, enters alongside the predictors and target lags in every model's regression at every origin and horizon. A given dummy is only included once its associated month actually falls inside that regression's training range; introducing it earlier would produce an identically-zero column and a rank-deficient design matrix, since the "outlier" would not yet have been observed.

A similar caveat applies here as to the DFM's factor count (Section 2.1.2): the four flagged months are identified from the standardised residuals of the full 2000–2026 series, not from a residual computed recursively at each origin using only the data available up to that point. The timing of each dummy's inclusion in the design matrix is unaffected, since a dummy is only ever added once its associated month has actually occurred, as described above; what is not real-time-consistent is the ex-ante identification of which months qualify as extreme in the first place. A direct check confirms that all four flagged months would still cross the $|z| > 3$ threshold under a genuinely recursive, expanding-window residual computed using only data through that month. The same recursive rule, however, would also flag eight further months not treated as outliers here, one earlier in the sample (October 2017) and seven clustered in the run-up to the 2021–2023 surge (January 2020, November 2021, and five months across 2022). This occurs because an expanding-window standard deviation, estimated before the full magnitude of the 2020–2023 volatility regime has entered the sample, is smaller than the eventual full-sample standard deviation used here, so moderate deviations earlier in the sample cross the same fixed threshold more easily in real time than they do once judged against the calmer, hindsight-informed full-sample benchmark. The consequence of the simplification adopted here is the following: a genuinely real-time outlier-dummy procedure would have treated several additional months during the 2021–2022 build-up as extreme observations, which the present design does not.

4.6 Data Properties

A natural question raised by working with an inflation series is whether its persistence properties matter for the validity of the forecasting methods used in this thesis. They do not, at least not formally. The random-walk benchmarks involve no estimation at all; UCSV explicitly models a non-stationary, random-walk trend component rather than assuming stationarity away; the dynamic factor model's factors are extracted from a predictor panel that is stationary by construction for the large majority of series (Table 7.1): the EA-MD-QD block is differenced according to unit-root tests, and non-stationary supplementary stocks and flows are converted to growth rates. A minority of series, interest rates, the unemployment rate, and a handful of ratios and expectations series, is instead retained in levels on the working assumption that these are stationary or sufficiently bounded not to distort the extracted factors. π_t 's own lags enter as regressors alongside the factors in the DFM's direct-forecast design, and alongside the predictor panel in every other model's design (Section 3), rather than π_t appearing only as the dependent variable. The Lasso, Ridge, Elastic Net, Random Forest and gradient-boosted-tree candidates carry no formal stationarity assumption either. None of the eight models compared in Chapter 3 requires π_t to satisfy a formal stationarity condition to be validly estimated: an AR-augmented direct-forecast regression with a persistent dependent variable and persistent own-lag regressors is standard practice in this literature and remains well defined regardless of whether π_t is formally stationary.

This does not make the question uninteresting, however, since a reader encountering the descriptive statistics below might reasonably wonder whether the target series behaves in a way that undermines the exercise. The remainder of this section addresses that concern directly, and identifies the one place where a related risk, a permanent shift in the level of inflation, is genuinely relevant to one family of candidate models.

Table 4.1 reports descriptive statistics for the target variable, core HICP inflation (π_t , month-on-month, annualised), over the sample February 2000 to January 2026 (312 monthly observations), along with headline HICP inflation for comparison. Both series show substantial excess kurtosis (10.33 for core, 9.62 for headline) relative to the Gaussian benchmark of zero, reflecting the extreme observations of the COVID-19 period and the subsequent inflation surge; headline inflation is additionally far more volatile (standard deviation 2.88 vs. 1.48) and skewed (1.91 vs. 0.61) than core inflation, consistent with the exclusion of energy and food prices being the relevant distinction between the two series.

An Augmented Dickey-Fuller (ADF) test applied to π_t over the full sample fails to reject

Descriptive statistics for the target variable (core HICP, m/m annualised) and, for context, headline HICP over the same sample. ADF: Augmented Dickey-Fuller test statistic and p -value (H_0 : unit root).

Variable	N	Mean	SD	Min	Max	Skewness	Kurtosis	ADF stat	ADF p
Core HICP (target)	312	1.67	1.48	-7.96	9.60	0.61	10.33	-2.69	0.29
Headline HICP	312	2.09	2.88	-5.44	21.50	1.91	9.62	-3.37	0.06

Note: Sample: 2000-01-01 to 2026-02-01 ($T = 314$ monthly observations). Core/headline HICP m/m rates are $1200 \times \Delta \log$ of the seasonally adjusted index. Skewness/kurtosis: bias-corrected (type 2); kurtosis is excess kurtosis (0 under normality).

Table 4.1: Target Variable properties

the null hypothesis of a unit root at conventional significance levels (test statistic -2.69 , $p = 0.29$); the same test applied to headline inflation narrowly rejects the null at the 10% level ($p = 0.06$). Taken at face value, this would suggest core inflation is non-stationary over this sample.

This result does not establish that core inflation is genuinely non-stationary. The standard ADF test assumes a stable mean under its alternative hypothesis, and a single large, temporary shift in that mean. This could be for example the 2021–2023 energy-price shock and its subsequent reversion, which is well known to bias the test toward non-rejection even when the underlying process is stationary around two distinct regimes rather than integrated (Perron, 1989). Since the sample used here visibly contains exactly this kind of episode, this is a more plausible explanation for the observed result than a literal unit root in the transformed series.

To test this explanation formally, a Zivot-Andrews unit-root test (Zivot and Andrews, 1992) is applied to π_t . Unlike the standard ADF test, this test allows for one endogenously-selected structural break in the series' intercept and tests the unit-root null against a trend-stationary alternative with a break. The test rejects the unit root decisively (test statistic -16.11 , against critical values of -5.34 , -4.80 and -4.58 at the 1%, 5% and 10% levels respectively), and the estimated break date in December 2020 falls immediately before the onset of the inflation surge rather than at an arbitrary point in the sample. Because the Zivot-Andrews alternative models a one-time *permanent* shift in the mean, its rejection establishes that π_t is not a pure random walk with unbounded memory, but it does not by itself establish that the break was temporary rather than permanent; that distinction has to be read from the series itself rather than from the test.

The one substantive implication of this analysis is narrower than stationarity itself. The Zivot-Andrews result rules out a literal unit root, but whether the December 2020 break

is temporary or permanent is better read from the series itself than from the test: core inflation returns to approximately 2% by 2024–2025 (Section 3.1, Figure 3.1), consistent with a temporary rather than permanent shift. Read this way, the identified episode does not threaten the validity of any candidate model directly; the risk it points to is that Random Forest and gradient boosting cannot extrapolate beyond the range of target values observed during estimation. This means a genuinely *permanent* shift in the level of inflation, rather than the temporary one identified here, could degrade their out-of-sample accuracy in a way it would not for the other candidates. This is precisely the failure mode the sub-period robustness check in Chapter 6 is designed to detect empirically, and it is addressed there rather than requiring any adjustment to the target variable itself.

Table 4.2 reports the ten predictors most correlated (in absolute value) with core inflation, out of the full 62-series panel. The same correlations are shown for every predictor in Figure 7.1, as part of the Appendix.

The 10 predictors most correlated (in absolute value) with the target, out of the 62-series panel (excluding HICP sub-components).

Variable	Corr. w/ target	Mean	SD	Min	Max
SPF inflation expectations (long-term)	0.42	1.90	0.11	1.65	2.18
Unemployment rate (<25, youth)	-0.42	18.77	3.28	14.00	25.60
Unemployment rate (total)	-0.40	8.92	1.65	6.20	12.30
HLW output gap	0.40	0.51	2.32	-4.96	7.01
Unemployment rate (25+)	-0.40	7.84	1.48	5.30	10.90
Producer prices: non-durable consumer goods	0.38	0.00	0.00	-0.01	0.02
Negotiated wage growth	0.38	2.28	0.83	1.12	5.38
Producer prices: consumer goods	0.37	0.00	0.00	-0.01	0.02
Producer prices: durable consumer goods	0.31	0.00	0.00	-0.01	0.01
Consumer inflation expectations (short-term)	0.29	20.87	10.37	-6.80	64.50

Note: Correlation computed pairwise-complete over the full sample. Units are mixed across predictors (rates/ratios in levels, price and volume series in 100 x monthly log-differences).

Table 4.2: Most correlated predictors with core euro area inflation

The ten most correlated predictors cluster into four economically recognisable channels, which is itself a useful sanity check on the panel: long-term SPF inflation expectations ($\rho = 0.42$) and short-term consumer inflation expectations ($\rho = 0.29$) represent the expectations channel of an expectations-augmented Phillips curve; three measures of unemployment (total, 25-and-over, and under-25) together with the Holston-Laubach-Williams output gap (all $|\rho| \approx 0.40$) represent labour-market and real-activity slack; three producer-price sub-indices (non-durable, general, and durable consumer goods, $\rho = 0.31$ – 0.38) represent upstream cost-push pressure working its way into consumer prices with

a lead; and negotiated wage growth ($\rho = 0.38$) represents the wage-price channel. That a panel spanning labour markets, prices, expectations, credit, sentiment and financial conditions produces exactly these four channels at the top, without any channel-specific series having been hand-selected for inclusion, is consistent with the panel measuring the macroeconomic relationships the forecasting exercise is intended to exploit.

Equally informative is the ceiling on these correlations: no predictor in the 62-series panel exceeds $|\rho| = 0.42$ with the target (Figure 7.1). Inflation dynamics over this sample are thus not well summarised by any single indicator, which is a first piece of evidence for pooling many weak predictors via a factor model or shrinkage/ensemble method rather than relying on one strong one.

5 Results

5.1 Forecast evaluation

Table 5.1 reports RMSE, MAE, and MAD ratios for each model relative to RW(3m), computed at every horizon from $h = 1$ to $h = 12$. At short and medium horizons ($h = 1$ to $h = 6$), the DFM achieves the lowest RMSE ratio in every column, ranging from 0.87 at $h = 1$ to 0.84 at $h = 5$ – $h = 6$. Beyond $h = 6$, the picture is more mixed: Ridge is lowest at $h = 7$, $h = 8$, $h = 9$, and $h = 11$ (RMSE ratios of 0.87, 0.88, 0.89, and 0.90), Elastic Net is lowest at $h = 10$ (0.89, narrowly ahead of Lasso and Ridge, which both round to 0.90 without being exactly tied: 0.899 and 0.902, respectively), and Lasso is lowest at $h = 12$ (0.925), narrowly ahead of Ridge and Elastic Net, which round to the same displayed value (0.93) but are not tied with it (0.929 and 0.930, respectively). Elastic Net otherwise tracks Lasso and Ridge closely at every horizon without ever posting the single best RMSE ratio elsewhere, consistent with it interpolating between the two penalties rather than dominating either. The MAE ratios tell a cleaner story than the RMSE ratios: the DFM is lowest at every horizon from $h = 1$ to $h = 12$ except $h = 10$, where the Median combination (0.87) edges it out by a hundredth (DFM: 0.88); neither Ridge, Lasso, nor Elastic Net ever overtakes the DFM on this metric at any horizon. The DFM’s factor structure appears best suited to capturing near-term co-movements in the data-rich predictor set on a squared-error basis, while the RMSE handoff to Ridge/Lasso/Elastic Net beyond $h = 6$ reflects the shrinkage estimators’ advantage on the (rarer) large-error observations that RMSE weights heavily, more than any general superiority over the DFM in typical conditions.

Random Forest and XGBoost tell a less favourable story under RMSE and MAE. Both models are competitive with RW(3m) at $h = 1$ (ratios of 0.97 and 1.03, respectively) but deteriorate steadily as the horizon lengthens: the Random Forest RMSFE ratio rises from 0.97 at $h = 1$ to a peak of 1.19 at $h = 9$, settling around 1.16–1.18 for $h = 10$ – $h = 12$, and the XGBoost ratio rises from 1.03 to a peak of 1.28 at $h = 7$. Read in isolation, this would suggest that the nonlinear, tree-based models offer no advantage over a naive benchmark at longer horizons and may actively hurt accuracy.

This conclusion, however, does not survive a closer look at the MAD ratios. Here, Random Forest is in fact the best-performing model at several horizons, achieving MAD ratios as low as 0.58 at $h = 10$ and 0.63–0.66 at $h = 8$ and $h = 9$, which is far below its own RMSFE ratio at the same horizons. This is exactly the divergence the RMSFE-MAE-

MAD distinction predicts: Random Forest’s forecast errors are *typically* small and tightly clustered around their own median (hence the strong MAD) but occasionally very large, plausibly clustered around the 2020–2022 disruption, and it is these rare extreme misses that RMSFE’s quadratic penalty picks up so heavily, with MAE registering them only mildly and MAD, anchored to the errors’ own median rather than to zero, not registering them at all. Put differently, Random Forest is not simply a worse forecaster than its RMSFE ranking suggests; it is a forecaster whose typical-case accuracy is good but whose tail risk is large, and RMSFE alone conflates the two. This suggests that tree-based methods do not fail to learn useful nonlinear structure in typical conditions, but rather struggle to generalise that structure during regime shifts, when the handful of large misses that separate its RMSFE from its MAD are concentrated.

Forecast combinations (simple average, trimmed mean, median: pooled across all six individual model candidates) perform respectably at short-to-medium horizons under all three metrics, and the trimmed mean is in fact the best-performing model of any kind, individual or combined, on the MAD metric at $h = 3$ and $h = 4$ (0.86 at both). This pairing is not a coincidence: the trimmed mean and median combinations are themselves robust statistics, constructed specifically to down-weight or discard whichever constituent forecast is most extreme in a given month, and MAD is a robust statistic built on the same principle, applied to the resulting errors rather than to the forecasts. A combination scheme that suppresses the influence of one wildly-off constituent model naturally produces an error series with few genuinely extreme values, which is exactly the property MAD is designed to reward and RMSFE is largely blind to. Beyond that narrow window, however, the combinations do not dominate the best individual model at any horizon, and their relative advantage erodes noticeably beyond $h \approx 8$. This is broadly consistent with the combination literature’s expectation that pooling helps most when individual model errors are large and imperfectly correlated, and helps less once a single model (here, Ridge, Lasso, or the DFM) is already well-specified for the horizon in question.

Table 5.2 repeats the RMSFE-ratio comparison separately for the pre- and post-2020 sub-samples, each with its own within-panel Diebold-Mariano test and Model Confidence Set.

Point estimates aside, formal tests of predictive superiority find little full-sample evidence to discriminate between the candidate models and the random-walk benchmark. The Diebold-Mariano test against RW(3m) rejects equal predictive ability at conventional levels in only two of the 36 candidate $\times$ horizon combinations tested at $h \in \{1, 3, 6, 12\}$: the

RMSFE, MAE, and MAD ratios with respect to RW(3m), by forecasting horizon ($h = 1, \dots, 12$), full sample. Values below 1 indicate the model is more accurate than RW(3m) on that metric/horizon; bold marks the lowest ratio in each column. RW(3m) is the reference benchmark.

Model	Forecasting horizon											
	1	2	3	4	5	6	7	8	9	10	11	12
RMSFE ratio												
DFM	0.87	0.90	0.86	0.87	0.84	0.84	0.91	0.94	0.97	0.99	0.99	1.01
Lasso	1.02	0.97	0.90	0.90	0.90	0.90	0.89	0.91	0.97	0.90	0.94	0.93
Ridge	1.04	0.99	0.90	0.88	0.86	0.87	0.87	0.88	0.89	0.90	0.90	0.93
Elastic Net	1.03	0.97	0.91	0.90	0.90	0.91	0.89	0.88	0.96	0.89	0.93	0.93
Random Forest	0.97	0.96	0.93	0.99	1.06	1.15	1.17	1.18	1.19	1.18	1.16	1.16
XGBoost	1.03	1.08	1.04	1.12	1.24	1.27	1.28	1.25	1.24	1.22	1.21	1.22
UCSV	0.99	0.99	0.95	0.94	0.95	0.94	0.94	0.94	0.95	0.95	0.94	0.95
RW (12m)	0.97	0.97	0.96	0.98	0.99	1.01	1.01	1.02	1.02	1.02	1.02	1.02
<i>Forecast combinations</i>												
Average	0.96	0.95	0.89	0.91	0.93	0.96	0.97	0.97	1.00	0.98	0.99	1.00
Trimmed Mean	0.97	0.95	0.88	0.91	0.91	0.94	0.95	0.95	0.99	0.96	0.98	0.99
Median	0.97	0.94	0.88	0.88	0.88	0.89	0.91	0.91	0.96	0.93	0.96	0.97
MAE ratio												
DFM	0.88	0.88	0.87	0.86	0.83	0.80	0.84	0.83	0.87	0.88	0.87	0.89
Lasso	1.00	0.92	0.92	0.93	0.92	0.88	0.92	0.93	0.96	0.92	0.95	0.95
Ridge	1.01	0.92	0.91	0.93	0.90	0.89	0.90	0.89	0.91	0.90	0.90	0.92
Elastic Net	1.00	0.93	0.92	0.92	0.92	0.89	0.91	0.91	0.97	0.90	0.93	0.94
Random Forest	0.94	0.91	0.93	0.96	0.97	0.98	0.98	0.97	0.99	0.98	0.99	1.02
XGBoost	1.02	1.01	1.02	1.06	1.10	1.07	1.10	1.09	1.11	1.14	1.15	1.19
UCSV	0.97	0.97	0.99	0.98	0.97	0.97	0.97	0.97	0.99	0.98	0.97	0.99
RW (12m)	0.95	0.93	0.99	0.99	0.98	0.97	0.98	0.97	1.00	0.99	0.98	0.99
<i>Forecast combinations</i>												
Average	0.94	0.90	0.90	0.91	0.92	0.89	0.91	0.89	0.92	0.90	0.91	0.94
Trimmed Mean	0.95	0.90	0.90	0.91	0.91	0.89	0.91	0.89	0.92	0.89	0.91	0.93
Median	0.95	0.90	0.90	0.90	0.89	0.86	0.89	0.88	0.90	0.87	0.90	0.92
MAD ratio												
DFM	0.86	0.86	0.92	0.88	0.75	0.70	0.75	0.67	0.69	0.71	0.70	0.68
Lasso	0.83	0.79	0.95	0.89	0.82	0.73	0.81	0.87	0.81	0.79	0.71	0.75
Ridge	0.97	0.83	0.92	0.95	0.83	0.76	0.84	0.86	1.01	0.84	0.81	0.80
Elastic Net	0.84	0.83	0.94	0.90	0.80	0.76	0.80	0.90	0.86	0.80	0.73	0.73
Random Forest	0.95	0.86	0.91	0.86	0.80	0.75	0.71	0.63	0.66	0.58	0.64	0.72
XGBoost	0.95	0.89	0.93	0.98	0.89	0.82	0.83	0.89	0.99	0.95	1.02	1.03
UCSV	0.97	1.00	1.14	1.10	0.93	0.92	0.99	1.04	1.02	0.95	0.99	0.94
RW (12m)	0.91	0.80	1.01	0.91	0.82	0.81	0.83	0.86	0.94	0.80	0.87	0.83
<i>Forecast combinations</i>												
Average	0.86	0.81	0.87	0.87	0.80	0.73	0.82	0.76	0.78	0.74	0.73	0.77
Trimmed Mean	0.88	0.80	0.86	0.86	0.80	0.75	0.79	0.76	0.79	0.70	0.69	0.73
Median	0.88	0.80	0.87	0.88	0.79	0.73	0.78	0.77	0.81	0.73	0.73	0.77

Note: Full sample; computed from a separate recursive re-estimation at every horizon 1-12

Table 5.1: Relative forecast accuracy: RMSFE, MAE, and MAD vs. RW(3m)

RMSFE ratio to RW(3m), by forecasting model and horizon, Pre-2020 vs. Post-2020. Values below 1 indicate the model is more accurate than RW(3m) within that panel; bold marks the lowest ratio in each column-panel. RW(3m) is the reference benchmark.

Model	h = 1	h = 3	h = 6	h = 12
Pre-2020				
DFM	0.897***†	0.804***†	0.713***†	0.740**†
Lasso	0.932**†	0.787***†	0.760**†	1.023†
Ridge	0.929*†	0.848***†	0.840*†	0.983†
Elastic Net	0.940**†	0.782***†	0.744**†	1.017†
Random Forest	0.929***†	0.791***†	0.726**†	0.803*†
XGBoost	0.939**†	0.852***†	0.785**†	0.924†
UCSV	1.030	0.998	0.967	0.954†
RW (12m)	0.911†	0.778†	0.707†	0.715†
<i>Forecast combinations</i>				
Average	0.906***†	0.785***†	0.719***†	0.880
Trimmed Mean	0.908***†	0.786***†	0.722**†	0.889
Median	0.910***†	0.786***†	0.721**†	0.896
Post-2020				
DFM	0.866*†	0.873†	0.867†	1.053†
Lasso	1.049†	0.928†	0.929†	0.908†
Ridge	1.070†	0.917†	0.880†	0.919†
Elastic Net	1.051†	0.936†	0.938†	0.915†
Random Forest	0.984†	0.956†	1.223†	1.210†
XGBoost	1.078†	1.130†	1.372†	1.251†
UCSV	0.983†	0.938†	0.937†	0.949†
RW (12m)	0.985†	1.000†	1.066†	1.063†
<i>Forecast combinations</i>				
Average	0.973†	0.922†	1.006†	1.021†
Trimmed Mean	0.986†	0.905†	0.986†	1.006†
Median	0.984†	0.903†	0.926†	0.989†

Note: */**/** denote Diebold-Mariano significance (10%/5%/1%) vs RW(3m); a dagger superscript marks inclusion in the 90% Model Confidence Set (both computed separately within each panel). Pre-2020 = evaluation start to Dec-2019; Post-2020 = Jan-2020 to end of sample. Ratios, DM, and MCS are all re-run within each panel's own sub-sample; a model's Pre-2020 and Post-2020 figures are not directly comparable to each other, since the RW(3m) denominator differs across panels.

Table 5.2: Forecasting accuracy with period split (Pre-2020, Post-2020)

DFM's improvement at $h = 1$ (5% level) and Ridge's improvement at $h = 3$ (10% level). No other model, including the DFM at $h = 3$ and $h = 6$ despite its lowest point estimate there, reaches significance at any horizon. The Model Confidence Set procedure tells the same story more starkly: at the 90% confidence level, every candidate and every benchmark, including RW(3m) and RW(12m) themselves, remains inside the confidence set at every horizon except $h = 1$, where the simple average alone is excluded. This inability of the MCS to statistically discriminate among almost all candidate models echoes Hall, Tavlas, and Wang (2023)'s finding that tests of forecast superiority for euro area inflation tend to have limited power over the sample lengths typically available, and raises a natural question: is a candidate's full-sample edge over RW(3m) genuinely negligible, or is it being averaged away across a fifteen-year window spanning two very different inflation regimes? The sub-period split addresses this directly.

In the pre-2020 sub-sample, RW(12m) achieves the lowest RMSFE ratio at $h = 3$ (0.778), $h = 6$ (0.707), and $h = 12$ (0.715); at $h = 1$, however, it is the DFM (0.897) that is lowest, narrowly ahead of RW(12m) itself (0.911). Ridge and Random Forest are tied a step further back at $h = 1$ (0.929 each), essentially matching Lasso (0.932) and XGBoost (0.939) in a tight middle cluster. This is a marked change from the full sample, where Ridge, Lasso, and Elastic Net all sit *above* 1 at $h = 1$ (Table 5.1) and so do not beat RW(3m) at all once the post-2020 surge is blended in. This result is still consistent with the broader finding that in a low, stable, and well-anchored inflation environment, sophisticated conditioning information adds little over a simple average of recent outcomes (Hall, Tavlas, and Wang, 2023): the DFM's edge at $h = 1$ is a narrow, single-horizon improvement, not a broad one, and RW(12m) otherwise dominates every other horizon outright, which is exactly where the Diebold-Mariano stars in Table 5.2 also concentrate most heavily.

The post-2020 sub-sample overturns this picture, and mirrors the full-sample ranking (Table 5.1) closely. The DFM is again the lowest-ratio model at $h = 1$, $h = 3$, and $h = 6$ (0.866, 0.873, and 0.867), and Lasso is again lowest at $h = 12$ (0.908), narrowly ahead of Elastic Net (0.915) and Ridge (0.919). Elastic Net again sits between the two without ever leading at any horizon. The *relative* position against RW(12m) is where the more interesting result lies: dividing the DFM's ratio to RW(3m) by RW(12m)'s own ratio to RW(3m) recovers the DFM's position relative to RW(12m) directly, and that position improves substantially between panels, from essentially tied pre-2020 ($0.713/0.707 \approx 1.01$ at $h = 6$) to a clear lead post-2020 ($0.867/1.066 \approx 0.81$ at $h = 6$). Part of this narrowing gap, however, is not solely attributable to the models themselves: the naive benchmarks reorder the same way. RW(3m)/RW(12m), recovered the same way from the RW(12m) row, falls from 1.29–1.41 at $h = 3$ – $h = 12$ pre-2020 to 0.94–1.00 over the same horizons

post-2020. A trailing 12-month average is, by construction, slow to register a genuine regime change, so RW(12m) itself becomes a weaker benchmark once inflation stops being stable. The apparent gain from data-rich or factor-based methods post-2020 is therefore partly a story about the naive benchmark deteriorating, not only about the models improving. Random Forest and XGBoost (the two models already flagged above as more fragile to outlier-driven errors) deteriorate sharply from pre- to post-2020, most visibly at $h = 6$ and $h = 12$: Random Forest's ratio rises from 0.726/0.803 to 1.223/1.210, and XGBoost's from 0.785/0.924 to 1.372/1.251.

Taken together, the sub-period results suggest that whatever edge data-rich, regularised, or factor-based forecasting methods hold over the random-walk benchmarks is concentrated in the post-2020 period of structural instability, and negligible or confined to a single short horizon during the low, stable, anchored pre-2020 period (at least when measured against RW(12m) in point-estimate terms). This should not be read as unqualified evidence of model skill: part of it reflects RW(12m) itself becoming a weaker benchmark once inflation stops being stable, rather than the candidate models becoming more accurate in an absolute sense. The Diebold-Mariano and MCS annotations in Table 5.2 make this distinction explicit, and in fact point the other way: statistical significance against RW(3m) is common and often strong (up to the 1% level) throughout the pre-2020 panel, and all but disappears post-2020, where only the DFM at $h = 1$ retains a single asterisk. The two sub-periods therefore tell different stories depending on which benchmark and which question is asked: against RW(12m) in point-estimate terms, the models' apparent edge is a post-2020 phenomenon; against RW(3m) in statistical terms, the only sub-period where that edge is confirmed with any confidence is pre-2020. Read together with the full-sample result above, this is consistent with a candidate edge that is real but modest in the calm pre-2020 regime, but too small, or too swamped by 2021-2023 volatility, to survive a formal test once the random walk itself becomes the more competitive benchmark.

5.2 Variable Selection and Interpretability: SHAP and LIME

Figure 5.1 decomposes variable importance by predictor category (own lags, employment, consumption, interest and exchange rates, output/income/housing, money and credit, prices and wages, expectations, and stocks) for the five models for which importance measures are available: Ridge, Lasso, Elastic Net, Random Forest, and XGBoost.

Three patterns are common across the linear and Random Forest models. First, the importance of own lags (AR) declines monotonically with the forecast horizon, from a non-trivial share at $h = 1$ (particularly pronounced for Lasso, where own lags account

for over a third of importance at $t+1$) to a negligible share beyond roughly $h = 6$. This is the expected signature of direct multi-step forecasting: near-term inflation dynamics retain autoregressive persistence, but this persistence has largely dissipated by the time the target is six to twelve months ahead.

Second, *Prices and wages* and *Expectations* play different roles across models rather than moving together. For Ridge, Lasso, and Elastic Net, *Expectations* is the larger of the two at short horizons and gives ground gradually to *Prices and wages*, which is roughly level with or modestly ahead of it by $t+9-t+12$; combined, the pair remains the single largest or second-largest contributor to importance at almost every horizon for these three models, consistent with the view that inflation persistence and expectations-anchoring are the dominant forces shaping medium-term core inflation dynamics. Random Forest leans almost entirely on *Prices and wages* instead: its share widens steadily with the horizon while *Expectations* stays flat and comparatively small throughout, indicating that Random Forest draws on realised cost-push pricing rather than on the forward-looking expectations channel the linear models rely on more evenly.

Third, *Output, income and housing* is a large and generally increasing contributor across horizons for all five models, plausibly reflecting the role of slower-moving real-activity indicators in signalling demand-driven inflationary pressure that materialises with a lag.

XGBoost importance is markedly less stable across horizons than that of the other four models. Short-horizon importance is split fairly evenly across Output, income and housing, Employment, and Interest and exchange rates; from around $t+7$ onward, Output, income and housing abruptly comes to dominate, accounting for more than half of total importance at every longer horizon, with Interest and exchange rates a distant second and Employment, Prices and wages, and Expectations all comparatively small throughout. This abrupt, discontinuous reallocation around $t+6-t+7$ contrasts with the smooth, gradual category shifts visible for Ridge, Lasso, and Random Forest.

This instability is unlikely to reflect a genuinely different economic mechanism operating within XGBoost forecasts, and more likely reflects a known limitation of gain-based importance measures for gradient-boosted trees: gain importance is sensitive to the correlation structure among predictors and can allocate importance somewhat arbitrarily among a set of collinear candidates, particularly in a data-rich environment where many macroeconomic series are highly correlated (Smeeke and Wijler, 2018).

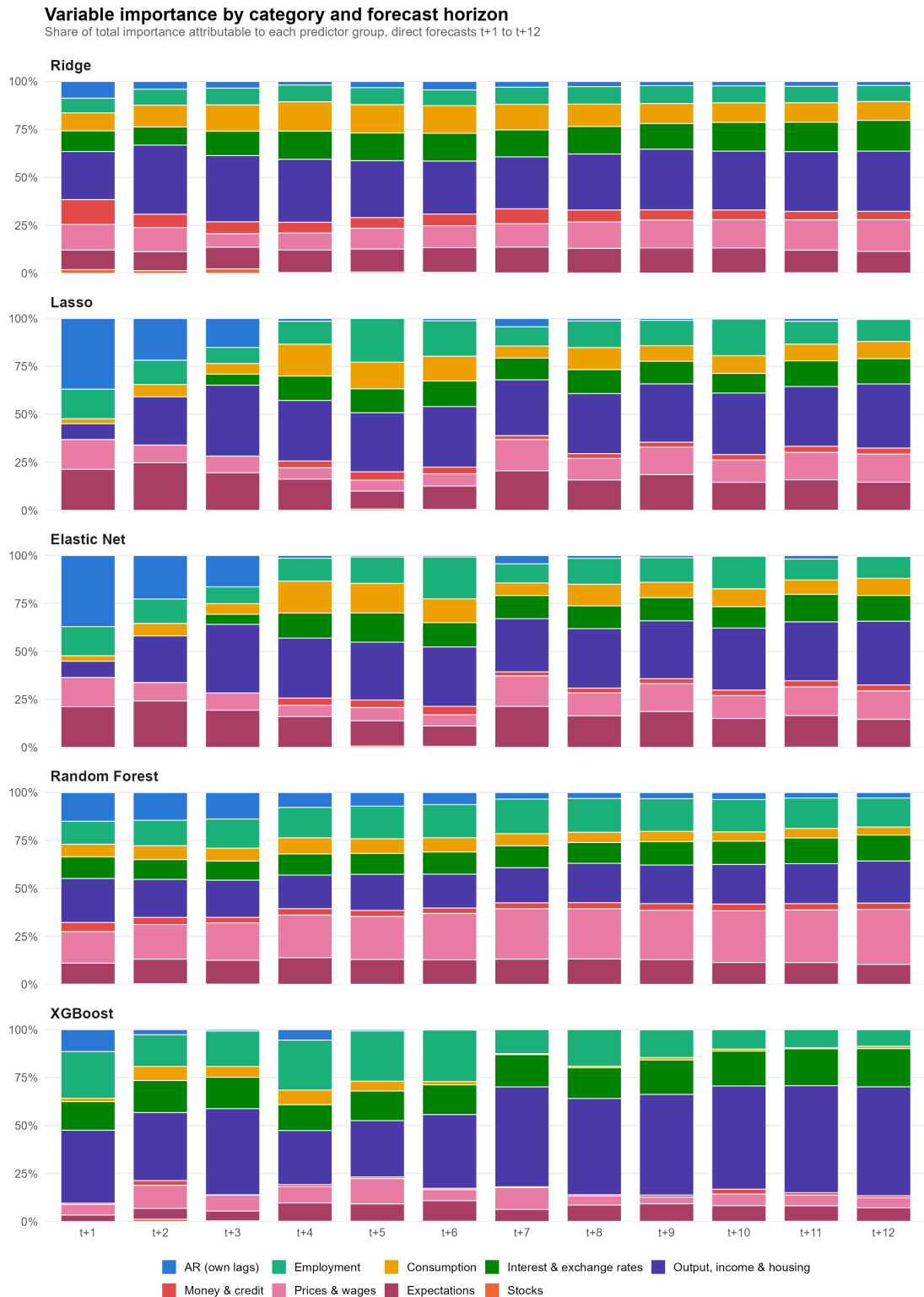


Figure 5.1: Variable importance in respective ML models

SHAP (Section 3.3.4) offers an independent check on this reading, since TreeSHAP’s exact per-observation decomposition does not share gain importance’s sensitivity to collinear splitting. Figure 5.2 shows each model’s top 15 predictors by relative SHAP importance across $h = 1, \dots, 12$. For Random Forest, the picture is a gradual rotation rather than an abrupt switch: youth unemployment is the dominant predictor at short and medium horizons, and the HLW output gap climbs steadily to overtake it by $h \approx 10\text{--}12$, with total and 25-and-over unemployment and the two money-market rates filling a broad, fairly stable mid-tier throughout. XGBoost’s profile is markedly more concentrated: importance is carried almost entirely by the HLW output gap and youth unemployment across the whole horizon range, with every other predictor, including the price and expectations series that feature prominently for Ridge, Lasso and Random Forest in Figure 5.1, contributing comparatively little at any horizon. A model relying this heavily on two predictors is, mechanically, more exposed to any instability in those two predictors’ own relationship with inflation than one that spreads its attribution across six or seven, which is consistent with, and gives a second, SHAP-based reason for, the instability already identified through gain importance above.

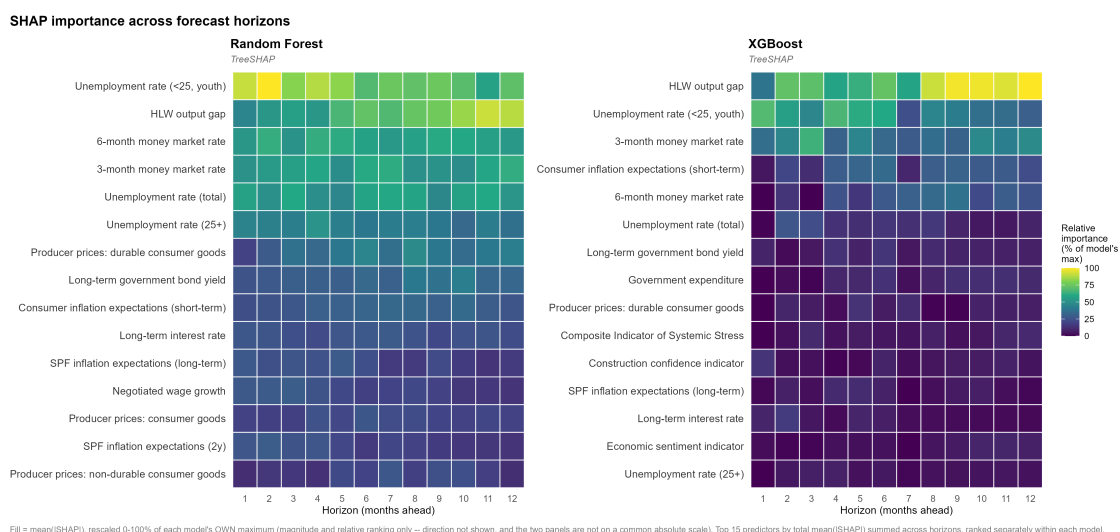


Figure 5.2: SHAP importance across forecast horizons, Random Forest and XGBoost

This concentration is compounded by instability in the *total* magnitude of XGBoost’s attributions across horizons (Figure 5.3). Summing $\text{mean}(|\phi_i|)$ across all predictors, Random Forest’s total declines smoothly from 1.17pp at $h = 1$ to 0.89pp at $h = 12$, while XGBoost’s is markedly non-monotonic: it rises to a peak of 1.57pp at $h = 3$, falls back to 1.21–1.24pp at $h = 5\text{--}h = 6$, drops sharply to a trough of 0.64pp at $h = 7$, and then recovers to 1.03–1.14pp for $h = 8\text{--}h = 12$. Since XGBoost is refit independently at every horizon with its boosting-round count chosen by early stopping (Section 3.3.4), this

jaggedness is consistent with the selected number of rounds varying substantially from one horizon to the next, obtained here via a completely different attribution mechanism from gain importance, and so corroborating rather than merely repeating the instability already noted above.

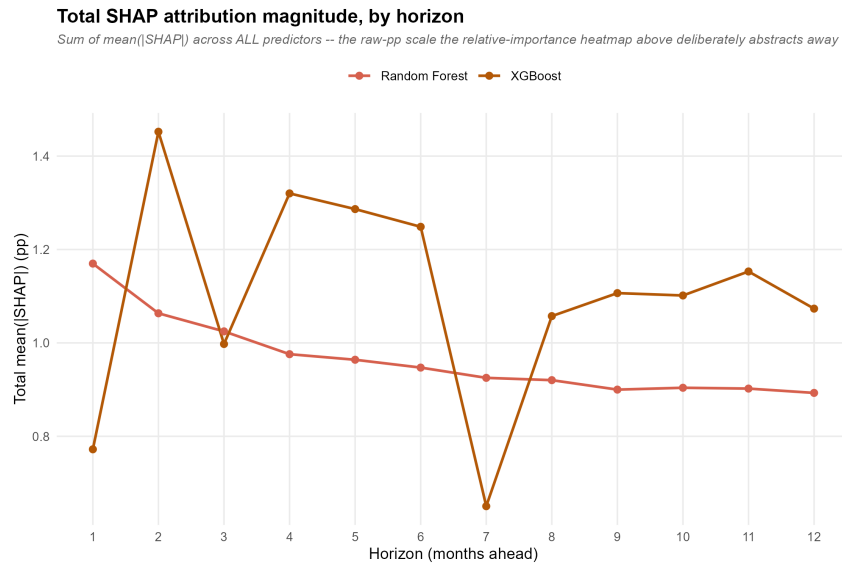


Figure 5.3: Total SHAP attribution magnitude by horizon, Random Forest vs. XGBoost

Random Forest — 2022-09 ($h = 12$)

SHAP decomposition of the fitted forecast (pp, core HICP m/m annualised)

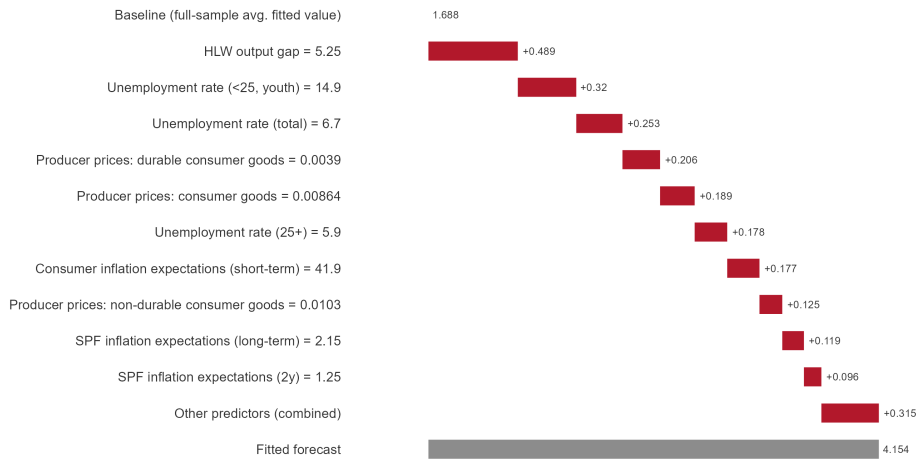
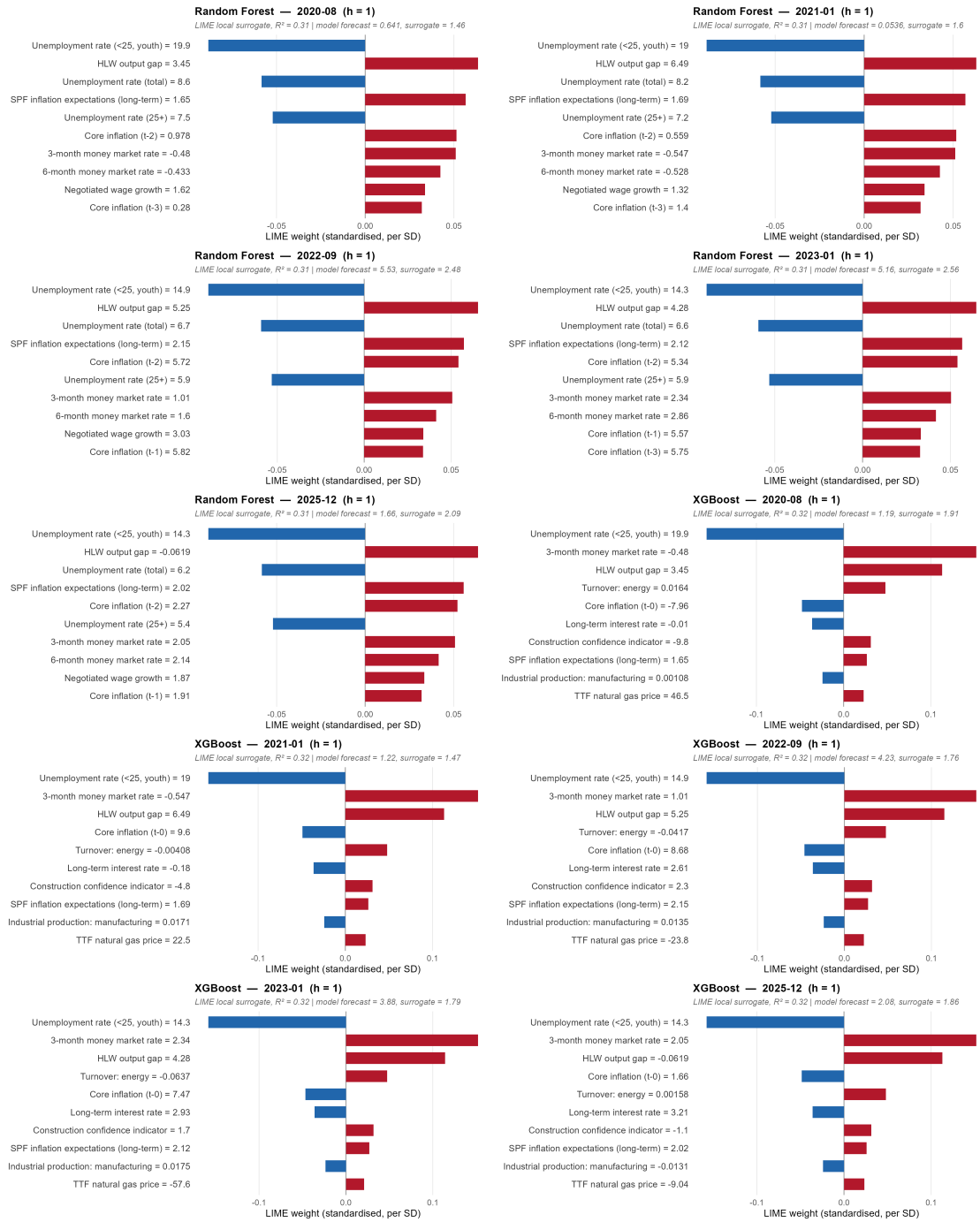


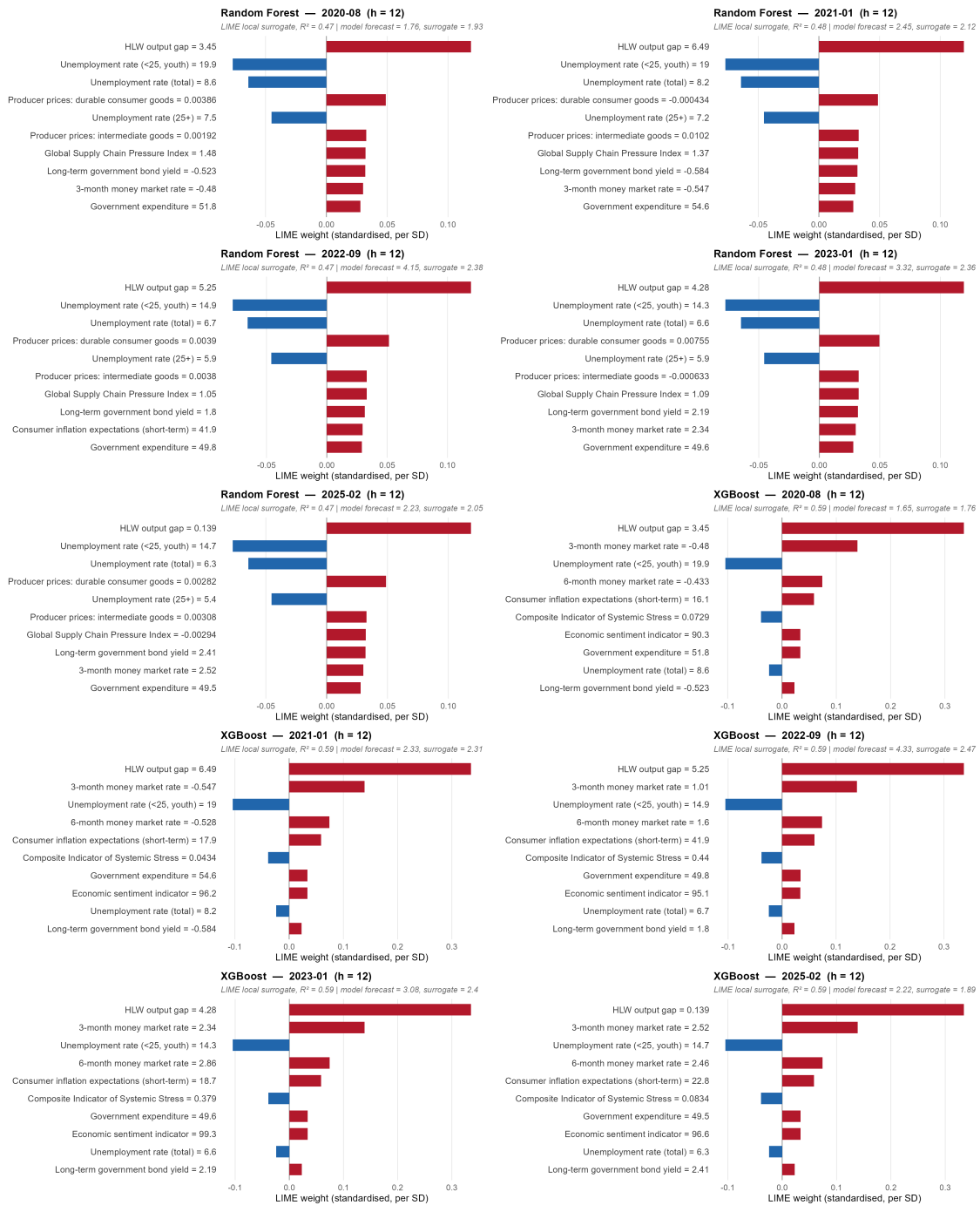
Figure 5.4: Local SHAP decomposition, Random Forest, September 2022, $h = 12$

The local decomposition makes this concrete for a single origin. Figure 5.4 decomposes Random Forest's $h = 12$ fitted forecast for September 2022: the peak of the 2022–2023 energy-price shock (Section 4.5). The full-sample baseline of 1.69pp is built up to a fitted value of 4.15pp almost entirely through positive contributions: the HLW output gap

(+0.49pp) and youth unemployment (+0.32pp) contribute the most, followed by total unemployment, a cluster of producer-price sub-components, and inflation expectations (all individually below +0.26pp). Every one of the top ten predictors pushes the forecast in the same direction for this origin, which is not the case for the other three flagged outlier months at $h = 12$, where positive and negative contributions are more evenly mixed: the model's mechanism, not a single dominant predictor, is what generates the unusually elevated medium-term path around the energy shock.

LIME (Section 3.3.4) closes the question left open in the methodology: whether its selected predictors actually vary across the same origins and horizons examined via local SHAP decomposition above, or stay similar regardless of which month is explained. Figures 5.5 and 5.6 report the local surrogate at $h = 1$ and $h = 12$ for the same five origins (the four outlier months plus the latest available date) and both models. The local R^2 is essentially constant within each model-horizon combination rather than varying with the date being explained: Random Forest: $R^2 = 0.31$ at every $h = 1$ origin, $R^2 = 0.47$ – 0.48 at every $h = 12$ origin; XGBoost: $R^2 = 0.37$ at $h = 1$, $R^2 = 0.56$ at $h = 12$. The top-10 selected predictors are, for a given model and horizon, nearly identical in both composition and ranking across five origins spanning the COVID trough, the 2021 base-effect episode, the energy-price peak, the post-crisis disinflation, and the most recent available month. This is precisely the failure mode flagged in Section 3.3.4 (Aras and Lisboa, 2022): rather than adapting to what is locally relevant at each origin, LIME's continuous, Gower-kernel-weighted perturbation recovers close to the same linear approximation of the model regardless of which month is explained. The predictors it does select overlap substantially with SHAP's own top global predictors (the unemployment measures, the HLW output gap, and the money-market rates) so LIME and SHAP broadly agree on *which* predictors matter in aggregate; the disagreement is that LIME cannot distinguish, locally, *why* a specific month's forecast is what it is, the one property TreeSHAP's exact per-observation decomposition is designed to guarantee.

Local LIME explanations, $h = 1$ Figure 5.5: Local LIME explanations, $h = 1$, Random Forest and XGBoost

Local LIME explanations, $h = 12$ Figure 5.6: Local LIME explanations, $h = 12$, Random Forest and XGBoost

5.3 Model Stability

The sub-period comparison in Section 5.1 showed that model rankings against random-walk benchmarks differ sharply between the low-inflation pre-2020 period and the high-inflation post-2020 period. This raises a natural question: is a candidate's average edge over RW(3m) actually stable across the full evaluation window, or does it come from a narrow sub-period, with the model doing no better (or even worse) elsewhere? The Giacomini and Rossi (2010) fluctuation test addresses this directly, testing the null hypothesis that a candidate's relative out-of-sample performance against a reference benchmark is constant over time, against the alternative that it fluctuates.

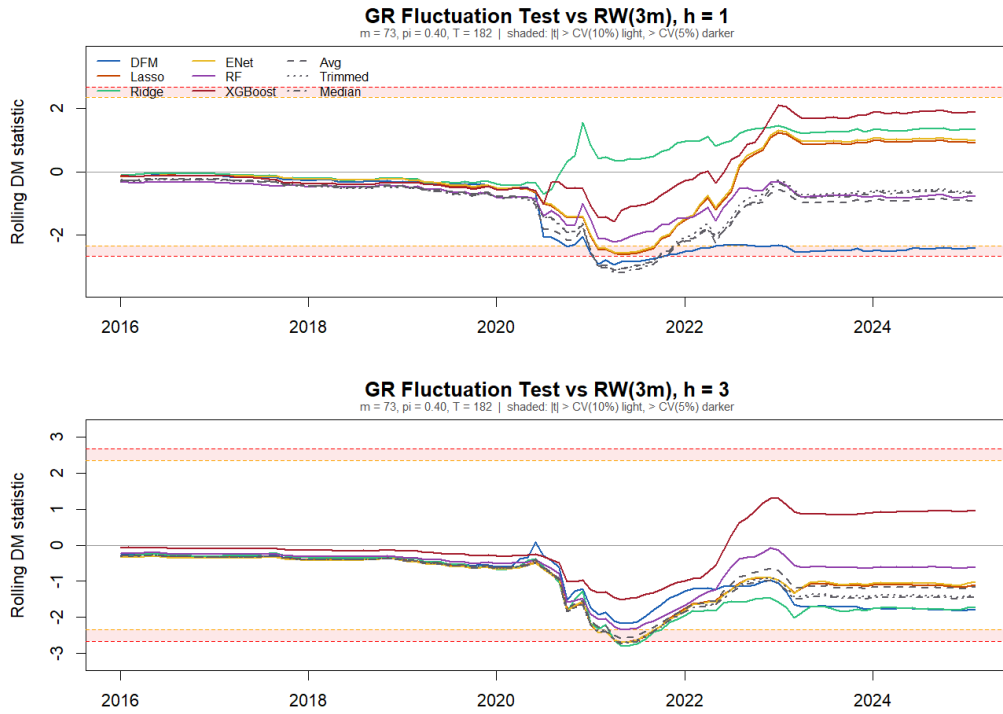


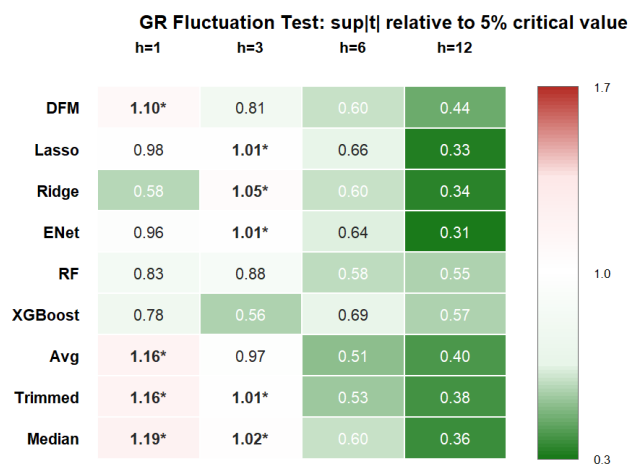
Figure 5.7: Giacomini-Rossi fluctuation test: rolling DM statistic vs RW(3m)

Figure 5.7 plots the rolling statistic for all nine candidates (the six individual models: DFM, Lasso, Ridge, Elastic Net, Random Forest, XGBoost and the three forecast combinations) at $h = 1$ and $h = 3$. This is the two horizons where the sup-statistic breaches the 5% critical value for at least one candidate. (Figure 5.8 summarises $\sup|t|$ relative to the 5% critical value for all four horizons). Both panels show the same pattern: every candidate's rolling statistic falls sharply and stays deeply negative from around 2020 through 2021, which is exactly the period in which core inflation departed most from its recent history, before drifting back toward zero, and in several cases crossing into positive territory, over

2022-2023. A negative statistic indicates the candidate's squared error is below RW(3m)'s over that window; the swing from strongly negative to flat or positive therefore does not mean these models forecast badly in absolute terms during 2022-2023, but that their advantage over the naive benchmark narrows or disappears as the sample re-normalises.

At $h = 1$, four of the nine candidates: the DFM and all three forecast combinations (Average, Trimmed Mean, Median), have $\sup|t|$ ratios above 1 relative to the 5% critical value (1.10, 1.16, 1.16, and 1.19, respectively), formally rejecting stability of their relative performance against RW(3m). Lasso, Ridge, Elastic Net, Random Forest, and XGBoost do not reject at $h = 1$. At $h = 3$, the pattern shifts: Lasso, Ridge, Elastic Net, Trimmed Mean, and Median now reject (ratios of 1.01, 1.05, 1.01, 1.01, and 1.02), while the DFM, Random Forest, XGBoost, and the simple average do not. At $h = 6$ and $h = 12$, no candidate rejects stability at any conventional level (the highest ratio is 0.69, for XGBoost at $h = 6$); relative performance against RW(3m) is uniformly stable at these longer horizons.

These results confirm with a formal test, that whatever edge the short-horizon models and combinations hold over RW(3m) is not constant across the sample. It is concentrated in, and driven by, the 2020-2021 window, exactly where the split-sample results already pointed. At the longer horizons, by contrast, the absence of any rejection indicates that the relative ranking established in Table 5.1 is not an artefact of a single turbulent sub-period, but holds up consistently across the fifteen-year evaluation window.



Values > 1 (marked *, red) reject stability of the candidate's relative performance vs RW(3m) at the 5% level

Figure 5.8: Giacomini-Rossi fluctuation test: $\sup|t|$ relative to the 5% critical value

6 Robustness Check

6.1 Predictor Lags: Extending the Feature Set for Shrinkage Methods

Macroeconomic determinants of inflation in the panel (Section 7.1) typically transmit to prices with a delay rather than instantaneously: the effect of a change in the money supply or in labour costs on core inflation is generally not fully reflected within the same month. A model that only ever observes the contemporaneous vector X_t may therefore be discarding information that is already available to the forecaster at the time the forecast is made, simply because that information entered the predictor panel one or more periods earlier. Allowing the estimator to see X_{t-1} alongside X_t is a direct, data-driven way to let it exploit this transmission-lag structure without imposing a specific lag length for each of the 62 series individually.

The direct h -step design matrix was extended with an additional parameter p_x , denoting the number of extra predictor lags included beyond the contemporaneous vector. For $p_x = 0$ (the default, used by DFM, Random Forest, and XGBoost) the design is unchanged from the baseline specification in Section 3. For $p_x = 1$, the K -column predictor block X_t is replaced by the $2K$ -column block $[X_t, X_{t-1}]$, stacking one additional lag of every series, while the dependent variable (the h -step-ahead average of π_t^{core}), the autoregressive lags of the target, and the outlier dummies (Section 4.5) are left unchanged. The prediction row used to generate the out-of-sample forecast is extended identically, so estimation and prediction always use the same feature space.

Medeiros, Vasconcelos, et al. (2021) include four lags of every predictor, but their panel has a substantially larger ratio of observations to predictors than the one used here. In this thesis, the earliest recursive estimation windows (starting 2010-01) contain on the order of 120 monthly observations against a 62-series panel, which is already a thin observations-to-predictors ratio before any lag is added. Extending to four lags would imply $K \times 5 = 330$ feature columns against as few as 120-180 training rows in the early part of the recursive sample, i.e. more predictors than observations even before the target lags and outlier dummies are added. While `glmnet` remains numerically well-defined in this regime, the information-criterion-based λ selection (BIC for Lasso, GCV for Ridge) loses much of its power to separate genuine signal from noise as dimensionality grows relative to sample size. One additional lag ($p_x = 1$) was therefore chosen as the largest lag depth this sample could plausibly support, doubling rather than quintupling the feature

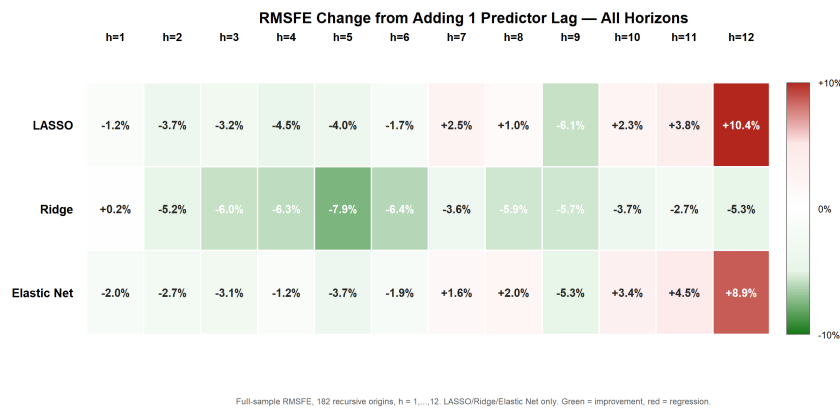


Figure 6.1: Accuracy gains of lagged predictors in LASSO and RIDGE

space. This same sample-size argument, that added dimensionality, is only useful if the estimator has a way to filter it.

The lag extension was therefore applied only to the two shrinkage estimators, not to Random Forest, XGBoost, or the DFM. Lasso and Ridge are built to handle wide, collinear feature sets: their ℓ_1/ℓ_2 penalties are precisely the mechanism that allows the model to down-weight or eliminate lagged columns that add no information, while still exploiting those that do. Random Forest and XGBoost instead select splits greedily from whichever column performs best in a given bootstrap sample or boosting round, and the DFM extracts a small number of unsupervised common factors from the panel as a whole; neither has an analogous built-in safety valve against added dimensionality. This expectation is consistent with earlier findings in this thesis (Section 5.2), where Random Forest and XGBoost’s forecast accuracy proved considerably more sensitive to the width and noise content of the predictor set than Lasso or Ridge’s. The lag experiment was therefore targeted at the two models best positioned to benefit from the additional features, and its results are reported in Table 7.2 and summarized in Figure 6.1.

Figure 6.1 disaggregates the predictor-lag comparison by individual horizon. The two rows reveal markedly different responses to the added lag, which in turn helps explain why Ridge, and not Lasso, emerges as the main beneficiary of this extension in the results Table 7.2.

Ridge improves at every horizon except $h = 1$, where it is the only case in which the extra lag makes accuracy marginally worse (+0.2%). The gains are largest in the $h = 3$ – $h = 6$ band, peaking at -7.9% at $h = 5$, and stay broadly in the -4% to -8% range through $h = 9$. They then fade at the two longest horizons (-3.7% at $h = 10$, -2.7% at $h = 11$) before ticking back up to -5.3% at $h = 12$. This is consistent with how Ridge regularises: its ℓ_2

penalty shrinks coefficients smoothly rather than selecting a subset of columns, letting it spread weight across X_t and X_{t-1} for a given series instead of choosing between them. When the true transmission lag of a predictor is ambiguous or split across both periods, this is exactly the setting in which a dense shrinkage estimator has an advantage over a sparse one — though the fade at $h = 10-h = 11$ suggests this advantage weakens rather than strengthens as the horizon grows very long.

Lasso's response is far more uneven. It improves at every one of $h = 1-h = 6$, though only modestly (-1.2% to -4.5%), turns unfavourable at $h = 7$ and $h = 8$ ($+2.5\%$, $+1.0\%$), swings back to a sizeable improvement at $h = 9$ (-6.1%), and deteriorates again through $h = 10-h = 12$, most sharply at $h = 12$ ($+10.4\%$). Because the ℓ_1 penalty forces Lasso to select between X_t and X_{t-1} for each series rather than blend them, its BIC-based lambda selection must correctly identify which of the two nearly collinear copies carries the relevant signal; the reversal at $h = 9$, sitting between two deteriorating horizons, suggests this selection becomes unstable rather than smoothly worse at long horizons, and $h = 12$ is a clear case of the model locking onto the less informative lag.

At $h = 1$, in contrast to the rest of the range, neither model reacts strongly to the added lag: Lasso still improves marginally (-1.2%) and Ridge is only marginally worse ($+0.2\%$). This is the one horizon where the extra lag is least likely to add genuinely new information, since the contemporaneous predictor vector is already the most directly relevant input, so widening the feature space mainly adds estimation noise rather than transmission-lag signal.

6.2 Rolling Window Estimation

This section repeats the full recursive out-of-sample exercise under the rolling-window scheme described in Section 3.4.2: the same 182 forecast origins (2010-01 to 2025-02), the same four horizons, the same models estimated with an identical specification, but with the estimation sample fixed at 120 months and sliding forward one month at a time rather than expanding. The two random walk benchmarks, RW(3m) and RW(12m), are mechanically invariant to this choice, since both are trailing averages computed directly from the realised inflation series rather than estimated over a sample; they are included in Table 6.1 purely as a fixed point against which to read the other models' sensitivity.

The clearest result to survive both estimation schemes is the DFM's performance at short and medium horizons. It is the single most accurate model at $h = 1$, $h = 3$, and $h = 6$ under both the expanding and the rolling window (Figure 6.3), and at the two intermediate

Model	$h = 1$		$h = 3$		$h = 6$		$h = 12$	
	Exp.	Roll.	Exp.	Roll.	Exp.	Roll.	Exp.	Roll.
DFM	1.359	1.412	0.924	0.856	0.802	0.769	1.046	1.074
LASSO	1.592	1.677	0.969	1.038	0.858	0.964	0.954	1.662
Ridge	1.618	1.605	0.971	1.012	0.832	0.949	0.958	1.142
Random Forest	1.512	1.507	0.995	0.999	1.096	1.089	1.198	1.256
XGBoost	1.630	1.614	1.161	1.166	1.227	1.255	1.248	1.355
Average	1.492	1.496	0.963	0.968	0.915	0.945	1.034	1.156
Trimmed	1.507	1.526	0.948	0.966	0.900	0.949	1.021	1.147
Median	1.506	1.519	0.946	0.966	0.851	0.941	1.007	1.135
UCSV	1.548	1.577	1.020	1.045	0.897	0.910	0.980	0.972
RW(3m)	1.557	1.557	1.074	1.074	0.952	0.952	1.032	1.032
RW(12m)	1.507	1.507	1.032	1.032	0.964	0.964	1.053	1.053

Table 6.1: RMSFE by model, horizon, and estimation window regime

Expanding (Exp.) vs. rolling 120-month window (Roll.), both re-estimated over the same 182 recursive origins (2010-01 to 2025-02). Bold marks the lowest RMSFE within each horizon-regime column. RW(3m)/RW(12m) are identical across regimes by construction.

horizons it is not merely stable but *more* accurate under the rolling scheme, with RMSFE falling by 7.3% at $h = 3$ and 4.1% at $h = 6$. Restricting the factor extraction to the most recent ten years of data therefore does not just fail to hurt the DFM: it modestly helps, consistent with a factor structure that shifts slowly enough for the discarded pre-2010 observations to have become a source of estimation noise rather than useful signal by the time these origins are reached.

The result that does *not* survive the robustness check is LASSO's apparent advantage at the longest horizon. Under the expanding window, LASSO is the best-performing model at $h = 12$ (RMSFE 0.954, ahead of even UCSV and both random walk benchmarks). Under the rolling window, it becomes the *worst*-performing model at the same horizon (RMSFE 1.662, a 74.1% deterioration, last of eleven). This is the single largest ranking flip in Table 6.1 and Figure 6.3, and it is the sharpest illustration in this thesis of exactly the failure mode the fluctuation and window-robustness checks are designed to catch: a model whose long-horizon accuracy depends on being permitted to anchor its BIC-selected coefficient path on fifteen-plus years of history, rather than on a stable underlying relationship between the panel and future inflation. Section 6.1 already documented that



Figure 6.2: RMSFE by model and horizon, expanding vs. rolling (120-month) window

LASSO's variable selection at $h = 12$ is unstable in the sense of reacting unpredictably to an additional predictor lag; the present result shows the same instability along a second, independent dimension, the length of the estimation history, which reinforces rather than duplicates that earlier finding. Ridge, which uses the same predictor panel but never sets a coefficient exactly to zero, degrades far more gently over the same comparison (RMSFE rising from 0.958 to 1.142 at $h = 12$, and more gradually at the shorter horizons), consistent with the general expectation from Section 3.2 that a dense shrinkage estimator is less prone to abrupt selection-driven instability than a sparse one.

UCSV moves in the opposite direction from LASSO at the longest horizon: it is the best-performing model of all eleven at $h = 12$ under the rolling window (RMSFE 0.972), having ranked third under the expanding window, and its RMSFE is the most stable of any model across the full horizon range (a maximum change of 2.5% in either direction). This is consistent with Hall, Tavlas, and Wang (2023)'s account of *why* rolling windows should add little for a univariate, stochastic-volatility model applied to already-stationary post-2000 inflation: UCSV's trend-cycle decomposition adapts to changing volatility within the model itself, via the state-space stochastic-volatility structure described in Section 3.2, so truncating its estimation history mechanically removes information without offsetting this against a comparable gain in adaptability, and the model's RMSFE reflects exactly

that near-indifference.

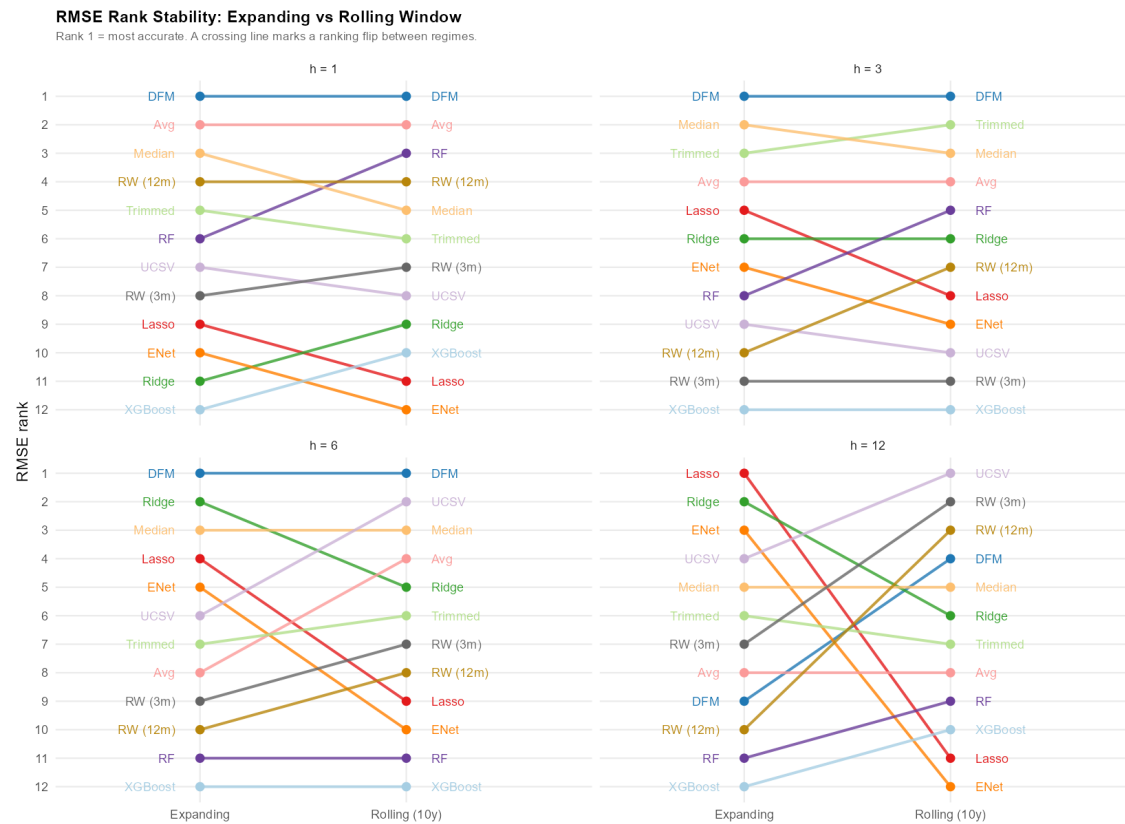


Figure 6.3: RMSFE rank stability across window regimes, by horizon

Random Forest and XGBoost are the two worst-performing candidate models at every horizon under both schemes, which is one respect in which the window choice makes no difference to the substantive conclusion: whatever limits these two models' accuracy in this application, it is not the length of the estimation window. This dovetails with the sensitivity to predictor width and noise already documented for both models in Section 6.1 and Section 5.2; a 120-month rolling window leaves the same 62-series panel to split on with fewer training rows than the expanding scheme provides in its later origins, which is, if anything, the more demanding setting for exactly the reason identified there.

Hall, Tavlás, and Wang (2023) points out, that rolling window forecasts for euro, US and UK inflation in years 1999-2022 did not yield much improvements on forecasting performance. The authors also point out that rolling window technique was essentially developed to handle non-stationary data, which was the case between 1970s-1990s. Starting from 2000s and 2010s, the inflation data becomes stationary, therefore the benefit of including rolling window disappears.

Taken together, these results qualify rather than simply confirm Hall, Tavlás, and Wang

(2023)’s argument that rolling windows add little once inflation is stationary. That claim holds essentially by construction for the two random walk benchmarks, and holds empirically for UCSV and, at the horizons where it matters most, for the DFM. It does not hold for LASSO, whose $h = 12$ ranking is entirely an artefact of the expanding window’s growing estimation sample rather than a stable property of the underlying model-panel relationship. Whether the choice of estimation window matters is therefore not a property of the inflation series alone, as a purely time-series argument would suggest, but depends on how a given model turns additional history into additional parameters. This is also why the forecast-combination schemes introduced in Section 3.4.1 are a useful hedge in practice, and not only against the possibility of having selected the wrong individual candidate model, as motivated there, but against the possibility of having selected the wrong *estimation window* for that candidate: the combination forecasts (Average, Trimmed, Median) move far less than LASSO does between the two schemes at $h = 12$ (a 12-13% change against LASSO’s 74%), precisely because no single unstable model can dominate a five-way average.

7 Discussion

In the final chapter, I would like to return to the three questions posed in the introduction of this thesis and draw out broader implications and limitations of this forecasting exercise.

Can machine learning outperform the benchmarks?

The answer to Q1 is yes, but narrowly, unevenly across horizons, and fragile to how the comparison is framed. Point estimates favour the data-rich models: the DFM posts the lowest RMSFE ratio against RW(3m) at every horizon from $h = 1$ to $h = 6$ (0.87 down to 0.84), and Ridge, Lasso, or Elastic Net take over beyond $h = 6$, settling around 0.87–0.93. Formal inference is far less generous. Across the full sample, the Diebold-Mariano test rejects equal predictive ability against RW(3m) in only two of 36 candidate-horizon combinations (the DFM at $h = 1$, Ridge at $h = 3$), and the Model Confidence Set retains essentially every candidate, including both random-walk benchmarks themselves, at every horizon. The sub-period split resolves part of this puzzle but complicates another part of it: pre-2020, RW(12m) is the best point forecast at three of four horizons and DM rejections against RW(3m) are common, up to the 1% level; post-2020, the DFM and the shrinkage estimators pull well ahead of RW(12m) in point-estimate terms, but DM significance against RW(3m) nearly disappears, surviving only for the DFM at $h = 1$. Whatever edge the data-rich models hold is therefore real in the sense of being detectable pre-2020, when the sample is calm enough for a test to have power, but too small, or too swamped by 2021-2023 volatility, to be confirmed statistically once that period is included. The Giacomini-Rossi fluctuation test adds a further qualification: at $h = 1$ and $h = 3$, where several candidates and combinations reject stability against RW(3m), the rejection is driven almost entirely by the 2020–2021 window; at $h = 6$ and $h = 12$, no candidate rejects stability at all. The short-horizon "edge" is therefore itself an artefact of a two-year window, not a feature of the full fifteen-year sample, while the longer-horizon rankings, modest as they are, hold up consistently. Taken together, the honest answer to Q1 is that traditional econometric benchmarks remain very hard to beat with any statistical confidence over the full evaluation window, and that whatever advantage the data-rich models do have is concentrated in specific horizons and periods rather than uniform across the exercise.

Does the gain come from information or from nonlinearity?

Q2 was deliberately framed as the sharpest test the design supports: holding the predictor panel fixed, does relaxing linearity, moving from Ridge, Lasso, and Elastic Net to Random Forest and XGBoost, improve accuracy? On the metric that matters most for the model comparison in this thesis, RMSFE, the answer is no. Random Forest and XGBoost are competitive with the linear models only at $h = 1$ (RMSFE ratios 0.97 and 1.03) and deteriorate steadily thereafter, peaking at 1.19 and 1.28 respectively, while Ridge, Lasso, and Elastic Net stay in the 0.87–0.93 range across the same horizons. On MAD, however, the picture reverses: Random Forest is the single best-performing model in the entire comparison at several horizons (0.58 at $h = 10$, 0.63–0.66 at $h = 8$ – $h = 9$), far below its own RMSFE ratio at those same horizons. This divergence is itself the answer to Q2, not a complication of it. Nonlinearity does appear to buy something: Random Forest’s *typical* forecast error, the one MAD is built to measure, is genuinely small. What it does not buy is protection against the rare, large errors that RMSFE’s quadratic loss penalises heavily, and it is exactly these rare errors, concentrated around the 2020–2023 disruption, that dominate the RMSFE-based ranking on which Q1 was answered. The SHAP and gain-importance evidence in Section 5.2 is consistent with a mechanism for this fragility rather than merely restating it: XGBoost’s variable attribution is markedly less stable across horizons than Ridge, Lasso, or Random Forest’s, and becomes heavily concentrated on two predictors (the HLW output gap and youth unemployment) precisely in the regime where accuracy deteriorates most. A model that leans this narrowly on so few series is mechanically more exposed to instability in those series’ own relationship with inflation, which is one plausible reason nonlinear flexibility translates into typical-case accuracy without translating into tail-robust accuracy. This is consistent with the reading, echoed in Naghi, O’Neill, and Danielova Zaharieva (2024), that Random-Forest-style dominance documented in earlier, calmer samples (Medeiros, Vasconcelos, et al., 2021) is not a robust feature of tree-based nonlinearity as such, but a feature of the regime in which it is estimated: in the roughly 120–190 monthly observations available at any point in this recursive sample, a flexible nonparametric learner appears to fit typical dynamics well but has too little data to also learn a functional form that degrades gracefully once the regime shifts. The information-richness side of Q2 is answered only indirectly by this design, and more cautiously: the gap between the univariate benchmarks and the data-rich linear models (DFM, Ridge, Lasso) is consistent with a richer panel adding real value, but because RW and UCSV differ from the DFM in estimation method and functional form as well as in information set, this comparison cannot isolate data richness from the change in model class in the way the linear-versus-nonlinear comparison isolates

nonlinearity. Read together, the evidence in this thesis is more supportive of the "richer information set" channel than of the "nonlinearity" channel, but this asymmetry partly reflects which of the two channels the design can test cleanly, and should be read with that caveat rather than as a fully symmetric horse race.

How stable are the results?

Q3 has two different answers depending on which result is being asked about. Some findings are robust to both the sub-period split and the choice of estimation window. The DFM's short-to-medium-horizon strength is one: it is the best-performing model at $h = 1$, $h = 3$, and $h = 6$ under both the expanding and the rolling window, and its accuracy at $h = 3$ and $h = 6$ actually improves under the rolling scheme (by 7.3% and 4.1% respectively), suggesting its factor structure is not merely surviving the robustness check but genuinely benefiting from a shorter, more homogeneous estimation history. Random Forest and XGBoost's underperformance is equally robust in the opposite direction: both are the two worst-performing candidates at every horizon under both window regimes, so whatever limits their accuracy in this application is not an artefact of window length. Other findings are not robust at all. LASSO's apparent win at $h = 12$ under the expanding window (RMSFE 0.954, best of eleven candidates) is entirely reversed under the rolling window, where it becomes the single worst-performing model at the same horizon (RMSFE 1.662, a 74% deterioration and the largest ranking flip in the exercise). Because this collapse coincides with the predictor-lag robustness check already showing LASSO's variable selection at $h = 12$ to be unusually sensitive to the addition of a single extra lag (Section 6.1), the two independent checks reinforce the same conclusion: LASSO's long-horizon accuracy in the main results reflects the BIC selector's access to a long, slowly-growing estimation history rather than a stable relationship between the panel and future inflation. Ridge, which never sets a coefficient to exactly zero, degrades far more gently under the same comparison, which is the pattern the dense-versus-sparse shrinkage distinction in Section 2.2.1 would predict. The overall lesson for Q3 is that stability itself needs to be checked independently of accuracy: a model's headline ranking in Table 5.1 is informative only once it survives a change in window and a change in feature width, and, for one of the models compared here, it did not.

Read together, these results narrow rather than resolve the question in the thesis title. Machine learning, in the specific form of tree-based nonlinearity examined here, does not outperform a well-specified linear, data-rich benchmark on the metric, RMSFE, that this and most of the comparable literature treats as decisive; where it wins is on typical-case error, a real but different property that a policymaker interested in worst-case forecast risk

around turning points would not prioritise. The more consistent performer throughout, the DFM, owes its advantage to exactly the channel the design can test least ambiguously in its favour by elimination: it uses the same rich panel as the ML models but processes it linearly, and it is competitive with or ahead of the penalised linear models at short and medium horizons without ever requiring a variable-selection step that a later robustness check could unravel. This is broadly consistent with Goulet Coulombe, Leroux, et al. (2022) and Giannone, Lenza, and Primiceri (2021)’s point that macroeconomic prediction problems are typically dense rather than sparse, favouring estimators like the DFM and Ridge that spread weight across many correlated predictors, over LASSO’s sparse selection or, more surprisingly given the pre-pandemic literature, over Random Forest and XGBoost’s nonparametric flexibility. The variable-importance results reinforce this reading economically rather than merely statistically: own-lag persistence, price and expectations indicators, and slow-moving real-activity series (Section 5.2) carry most of the predictive weight across models and horizons, which is exactly the kind of dense, broadly economically interpretable signal a linear model is well suited to exploit and a tree ensemble, in a panel and sample of this size, does not obviously exploit any better.

Three limitations qualify these conclusions. First, the sample available for estimation is short relative to the 62-series predictor panel: the earliest recursive windows contain on the order of 120 monthly observations, a regime in which flexible nonparametric learners are known to need more data than linear, regularised alternatives to generalise reliably; the underperformance of Random Forest and XGBoost documented here may reflect this specific data-to-predictor ratio rather than a general property of tree-based nonlinearity in inflation forecasting. Second, the evaluation window contains a single, large structural break (the pandemic and the 2021–2023 energy shock), and the pre-/post-2020 split, while a natural one given the data, cannot speak to how these models would compare under a different kind of regime change, such as the more gradual disinflation of the 1970s and 1980s or a demand-driven rather than supply-driven shock. Third, the robustness checks in Chapter 6 were themselves selectively targeted: the predictor-lag extension was applied only to Ridge and Lasso, the two models best equipped by their penalty structure to exploit added, partially redundant features, so it remains an open question whether Random Forest or XGBoost’s accuracy would respond differently to the same lag structure rather than to the contemporaneous panel used throughout the main results.

Three extensions follow directly from the limitations above and from choices flagged but not pursued earlier in the thesis. A targeted-predictor variant of the DFM, screening the panel for relevance to core inflation before extracting factors, as proposed by Bai and Ng (2008) and explicitly deferred in Section 2.1.2, would test whether the DFM’s

advantage over the ML models widens or narrows once it is given the same target-aware selection mechanism Ridge and Lasso already have built in. A bottom-up architecture forecasting core HICP's sub-components separately before aggregating (Joseph et al., 2021; Latypov et al., 2024) would test whether Random Forest and XGBoost's tail-risk problem is a property of forecasting an aggregate that itself averages over heterogeneous component dynamics, or persists at the disaggregated level. Finally, extending the lag-augmented feature set used for Ridge and Lasso in Section 6.1 to Random Forest and XGBoost, and evaluating the headline HICP series alongside core HICP, would complete the two robustness dimensions this thesis leaves partially open, and would let future work confirm whether the conclusions reached here for core inflation and a contemporaneous panel generalise along either of those two directions.

Machine learning, as implemented here, does not deliver an unambiguous improvement over well-specified traditional models for euro area core inflation forecasting. The most consistent forecasting gains over the random-walk benchmarks come from the dynamic factor model, a strictly linear method whose only departure from the univariate benchmarks is access to a large predictor panel; the two nonlinear, tree-based models add genuine accuracy in typical conditions but pay for it in fragility precisely when accuracy matters most, around the 2020–2023 regime shift that motivated this thesis in the first place. If there is a single lesson to draw from the evidence gathered here, it is that in this data and sample regime, the size and organisation of the information set forecasters condition on appears to matter more than the flexibility of the function used to process it, a conclusion that qualifies rather than overturns the more optimistic strand of the machine-learning forecasting literature, and one that a further decade of post-2026 data, and the extensions outlined above, will be well placed to test again.

Bibliography

- Abadie, Alberto and Maximilian Kasy (Dec. 1, 2019). “Choosing among Regularized Estimators in Empirical Economics: The Risk of Machine Learning”. In: *The Review of Economics and Statistics* 101.5, pp. 743–762. ISSN: 0034-6535. DOI: 10.1162/rest_a_00812.
- Ahn, Seung C. and Alex R. Horenstein (2013). “Eigenvalue Ratio Test for the Number of Factors”. In: *Econometrica* 81.3. _eprint: <https://onlinelibrary.wiley.com/doi/pdf/10.3982/ECTA8968>, pp. 1203–1227. ISSN: 1468-0262. DOI: 10.3982/ECTA8968.
- Angelini, Elena, Jerome Henry, and Ricardo Mestre (2003). *Diffusion Index-Based Inflation Forecasts for the Euro Area*. DOI: 10.2139/ssrn.356081.
- Angelopoulos, Georgios et al. (Oct. 14, 2025). *A new proposal for forecasting inflation in the eurozone: a global model*. Vol. 350. Working paper. Bank of Greece. DOI: 10.52903/wp2025350.
- Aras, Serkan and Paulo J. G. Lisboa (Nov. 30, 2022). “Explainable inflation forecasts by machine learning models”. In: *Expert Systems with Applications* 207, p. 117982. ISSN: 0957-4174. DOI: 10.1016/j.eswa.2022.117982.
- Araujo, Gustavo Silva and Wagner Piazza Gaglianone (June 1, 2023). “Machine learning methods for inflation forecasting in Brazil: New contenders versus classical models”. In: *Latin American Journal of Central Banking* 4.2, p. 100087. ISSN: 2666-1438. DOI: 10.1016/j.latcb.2023.100087.
- Atkeson, Andrew and Lee E. Ohanian (2001). *Are Phillips Curves Useful for Forecasting Inflation?* URL: <https://doi.org/10.21034/qr.2511> (visited on 03/30/2026).
- Bai, Jushan and Serena Ng (2002). “Determining the Number of Factors in Approximate Factor Models”. In: *Econometrica* 70.1, pp. 191–221. ISSN: 1468-0262. DOI: 10.1111/1468-0262.00273.
- (2006). “Confidence Intervals for Diffusion Index Forecasts and Inference for Factor-Augmented Regressions”. In: *Econometrica* 74.4, pp. 1133–1150. ISSN: 1468-0262. DOI: 10.1111/j.1468-0262.2006.00696.x.
- (Oct. 1, 2008). “Forecasting economic time series using targeted predictors”. In: *Journal of Econometrics*. Honoring the research contributions of Charles R. Nelson 146.2, pp. 304–317. ISSN: 0304-4076. DOI: 10.1016/j.jeconom.2008.08.010.
- Bañbura, Marta and Elena Bobeica (Jan. 1, 2023). “Does the Phillips curve help to forecast euro area inflation?” In: *International Journal of Forecasting* 39.1, pp. 364–390. ISSN: 0169-2070. DOI: 10.1016/j.ijforecast.2021.12.001.

- Bañbura, Marta, Domenico Giannone, and Lucrezia Reichlin (July 8, 2011). “Nowcasting”. In: *The Oxford Handbook of Economic Forecasting*. Ed. by Michael P. Clements and David F. Hendry. Oxford University Press, p. 0. ISBN: 978-0-19-539864-9. DOI: 10.1093/oxfordhb/9780195398649.013.0008.
- Barigozzi, Matteo and Claudio Lissona (Feb. 1, 2026). *EA-MD-QD: Large Euro Area and Euro Member Countries Datasets for Macroeconomic Research*. DOI: 10.5281/zenodo.18804061.
- Benecká, Soňa (Aug. 1, 2026). “Forecasting Disaggregated Producer Prices: A Fusion of Machine Learning and Econometric Techniques”. In: *Journal of Forecasting* 45.5, pp. 2458–2501. ISSN: 1099-131X. DOI: 10.1002/for.70149.
- Berben, Robert-Paul, Rajni Rasiawan, and Jasper de Winter (Feb. 5, 2025). *Forecasting Dutch Inflation Using Machine Learning Methods*. Rochester, NY. DOI: 10.2139/ssrn.5126947.
- Bernanke, Ben S. and Jean Boivin (Apr. 1, 2003). “Monetary policy in a data-rich environment”. In: *Journal of Monetary Economics*. Swiss National Bank/Study Center Gerzensee Conference on Monetary Policy under Incomplete Information 50.3, pp. 525–546. ISSN: 0304-3932. DOI: 10.1016/S0304-3932(03)00024-2.
- Bernanke, Ben S., Jean Boivin, and Piotr Elias (Feb. 1, 2005). “Measuring the Effects of Monetary Policy: A Factor-Augmented Vector Autoregressive (FAVAR) Approach”. In: *The Quarterly Journal of Economics* 120.1, pp. 387–422. ISSN: 0033-5533. DOI: 10.1162/0033553053327452.
- Bolivar, Osmar (Mar. 1, 2026). “High-frequency inflation forecasting: A two-step machine learning methodology”. In: *Latin American Journal of Central Banking* 7.1, p. 100172. ISSN: 2666-1438. DOI: 10.1016/j.latcb.2025.100172.
- Breiman, Leo (Aug. 1, 1996). “Bagging predictors”. In: *Machine Learning* 24.2, pp. 123–140. ISSN: 1573-0565. DOI: 10.1007/BF00058655.
- (Oct. 1, 2001). “Random Forests”. In: *Machine Learning* 45.1, pp. 5–32. ISSN: 1573-0565. DOI: 10.1023/A:1010933404324.
- Buckmann, Marcus and Andreas Joseph (2022). “An Interpretable Machine Learning Workflow with An Application to Economic Forecasting”. In: *SSRN Electronic Journal*. ISSN: 1556-5068. DOI: 10.2139/ssrn.4130517.
- Capolongo, Angela and Claudia Pacella (Nov. 1, 2021). “Forecasting inflation in the euro area: countries matter!” In: *Empirical Economics* 61.5, pp. 2477–2499. ISSN: 1435-8921. DOI: 10.1007/s00181-020-01959-4.
- Chakraborty, Chiranjit and Andreas Joseph (Sept. 1, 2017). *Machine Learning at Central Banks*. Rochester, NY. DOI: 10.2139/ssrn.3031796.

- Chamberlain, Gary and Michael Rothschild (Oct. 1982). *Arbitrage, Factor Structure, and Mean-Variance Analysis on Large Asset Markets*. DOI: 10.3386/w0996.
- Chan, Joshua C. C. (Sept. 14, 2018). "Specification tests for time-varying parameter models with stochastic volatility". In: *Econometric Reviews* 37.8, pp. 807–823. ISSN: 0747-4938. DOI: 10.1080/07474938.2016.1167948.
- Chen, Jeffrey C. et al. (2022). "Off to the Races: A Comparison of Machine Learning and Alternative Data for Predicting Economic Indicators". In: Abraham, Katharine G. et al. *Big Data for Twenty-First-Century Economic Statistics*. University of Chicago Press. ISBN: 978-0-226-80125-4 978-0-226-80139-1. DOI: 10.7208/chicago/9780226801391.001.0001.
- Chen, Tianqi and Carlos Guestrin (Aug. 13, 2016). "XGBoost: A Scalable Tree Boosting System". In: *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. KDD '16. New York, NY, USA: Association for Computing Machinery, pp. 785–794. ISBN: 978-1-4503-4232-2. DOI: 10.1145/2939672.2939785.
- Darracq Paries, Matthieu et al. (2021). "Review of Macroeconomic Modelling in the Eurosystem: Current Practices and Scope for Improvement". In: *SSRN Electronic Journal*. ISSN: 1556-5068. DOI: 10.2139/ssrn.3928288.
- Das, Pijush Kanti and Prabir Kumar Das (June 1, 2024). "Forecasting and Analyzing Predictors of Inflation Rate: Using Machine Learning Approach". In: *Journal of Quantitative Economics* 22.2, pp. 493–517. ISSN: 2364-1045. DOI: 10.1007/s40953-024-00384-z.
- Diebold, Francis X. and Roberto S. Mariano (1995). "Comparing Predictive Accuracy". In: *Journal of Business & Economic Statistics* 13.3, pp. 253–263. ISSN: 0735-0015. DOI: 10.2307/1392185.
- Doojay, Gan-Ochir and Batbold Narmandakh (July 3, 2025). "Forecasting Inflation in Mongolia Using Machine Learning". In: *International Economic Journal* 39.3, pp. 568–590. ISSN: 1016-8737. DOI: 10.1080/10168737.2025.2491386.
- Doz, Catherine, Domenico Giannone, and Lucrezia Reichlin (Sept. 1, 2011). "A two-step estimator for large approximate dynamic factor models based on Kalman filtering". In: *Journal of Econometrics*. Annals Issue on Forecasting 164.1, pp. 188–205. ISSN: 0304-4076. DOI: 10.1016/j.jeconom.2011.02.012.
- Duncan, Roberto and Enrique Martínez-García (July 1, 2019). "New perspectives on forecasting inflation in emerging market economies: An empirical assessment". In: *International Journal of Forecasting*. Forecasting issues in developing economies 35.3, pp. 1008–1031. ISSN: 0169-2070. DOI: 10.1016/j.ijforecast.2019.04.004.
- Eickmeier, Sandra and Christina Ziegler (2008). "How successful are dynamic factor models at forecasting output and inflation? A meta-analytic approach". In: *Journal*

- of Forecasting* 27.3. _eprint: <https://onlinelibrary.wiley.com/doi/pdf/10.1002/for.1056>, pp. 237–265. ISSN: 1099-131X. DOI: 10.1002/for.1056.
- Faust, Jon and Jonathan H. Wright (Oct. 1, 2009). “Comparing Greenbook and Reduced Form Forecasts Using a Large Realtime Dataset”. In: *Journal of Business & Economic Statistics* 27.4, pp. 468–479. ISSN: 0735-0015. DOI: 10.1198/jbes.2009.07214.
- (Jan. 1, 2013). “Chapter 1 - Forecasting Inflation”. In: *Handbook of Economic Forecasting*. Ed. by Graham Elliott and Allan Timmermann. Vol. 2. Handbook of Economic Forecasting. Elsevier, pp. 2–56. DOI: 10.1016/B978-0-444-53683-9.00001-3.
- Federal Reserve Bank of New York (2026). *Global Supply Chain Pressure Index*.
- Forni, Mario et al. (2005). “The Generalized Dynamic Factor Model: One-Sided Estimation and Forecasting”. In: *Journal of the American Statistical Association* 100.471, pp. 830–840. ISSN: 0162-1459.
- Geweke, John (1977). “The dynamic factor analysis of economic timeseries models”. In: *Latent variables in socio-economic models*. Latent variables in socio-economic models. - Amsterdam : North-Holland Publishing Company, ISBN 0-7204-0526-2. - 1977, p. 365–383, p. 365.
- Giacomini, Raffaella and Barbara Rossi (2010). “Forecast comparisons in unstable environments”. In: *Journal of Applied Econometrics* 25.4, pp. 595–620. ISSN: 1099-1255. DOI: 10.1002/jae.1177.
- Giannone, Domenico, Michele Lenza, and Giorgio E. Primiceri (2021). “Economic Predictions With Big Data: The Illusion of Sparsity”. In: *Econometrica* 89.5, pp. 2409–2437. ISSN: 1468-0262. DOI: 10.3982/ECTA17842.
- Giannone, Domenico, Lucrezia Reichlin, and David Small (May 1, 2008). “Nowcasting: The real-time informational content of macroeconomic data”. In: *Journal of Monetary Economics* 55.4, pp. 665–676. ISSN: 0304-3932. DOI: 10.1016/j.jmoneco.2008.05.010.
- Goulet Coulombe, Philippe, Maximilian Göbel, and Karin Klieber (Nov. 21, 2024). *Dual Interpretation of Machine Learning Forecasts*. Rochester, NY. DOI: 10.2139/ssrn.5029492.
- Goulet Coulombe, Philippe, Maxime Leroux, et al. (2022). “How is machine learning useful for macroeconomic forecasting?” In: *Journal of Applied Econometrics* 37.5, pp. 920–964. ISSN: 1099-1255. DOI: 10.1002/jae.2910.
- Hall, Stephen G., George S. Tavas, and Yongli Wang (2023). “Forecasting inflation: The use of dynamic factor analysis and nonlinear combinations”. In: *Journal of Forecasting* 42.3, pp. 514–529. ISSN: 1099-131X. DOI: 10.1002/for.2948.
- Hansen, Peter R., Asger Lunde, and James M. Nason (2011). “The Model Confidence Set”. In: *Econometrica* 79.2, pp. 453–497. ISSN: 1468-0262. DOI: 10.3982/ECTA5771.

- Hoerl, Arthur E. and Robert W. Kennard (Feb. 1, 1970). "Ridge Regression: Biased Estimation for Nonorthogonal Problems". In: *Technometrics* 12.1, pp. 55–67. ISSN: 0040-1706. DOI: 10.1080/00401706.1970.10488634.
- Holston, Kathryn, Thomas Laubach, and John C. Williams (2023). "Measuring the Natural Rate of Interest after COVID-19". In: *SSRN Electronic Journal*. ISSN: 1556-5068. DOI: 10.2139/ssrn.4482053.
- Joseph, Andreas (Mar. 8, 2019). "Parametric inference with universal function approximators". In: *Bank of England working papers*. Number: 784.
- Joseph, Andreas et al. (Mar. 26, 2021). *Forecasting UK Inflation Bottom Up*. Rochester, NY. DOI: 10.2139/ssrn.3819286.
- Kock, Anders Bredahl and Laurent Callot (June 1, 2015). "Oracle inequalities for high dimensional vector autoregressions". In: *Journal of Econometrics*. High Dimensional Problems in Econometrics 186.2, pp. 325–344. ISSN: 0304-4076. DOI: 10.1016/j.jeconom.2015.02.013.
- Latypov, Rodion et al. (2024). "Bottom-up Inflation Forecasting Using Machine Learning Methods". In: *Russian Journal of Money and Finance* 83.3, pp. 23–44.
- Li, Jiahua and Weiye Chen (Oct. 1, 2014). "Forecasting macroeconomic time series: LASSO-based approaches and their forecast combinations with dynamic factor models". In: *International Journal of Forecasting* 30.4, pp. 996–1015. ISSN: 0169-2070. DOI: 10.1016/j.ijforecast.2014.03.016.
- Liu, Yang, Ran Pan, and Rui Xu (Sept. 2024). "Mending the Crystal Ball: Enhanced Inflation Forecasts with Machine Learning". In: *IMF Working Papers* 2024.206, p. 1. ISSN: 1018-5941. DOI: 10.5089/9798400285387.001.
- Lundberg, Scott M and Su-In Lee (2017). "A Unified Approach to Interpreting Model Predictions". In: *Advances in Neural Information Processing Systems*. Vol. 30. Curran Associates, Inc.
- Lundberg, Scott M. et al. (Jan. 2020). "From local explanations to global understanding with explainable AI for trees". In: *Nature Machine Intelligence* 2.1, pp. 56–67. ISSN: 2522-5839. DOI: 10.1038/s42256-019-0138-9.
- Maehashi, Kohei and Mototsugu Shintani (Dec. 1, 2020). "Macroeconomic forecasting using factor models and machine learning: an application to Japan". In: *Journal of the Japanese and International Economies* 58, p. 101104. ISSN: 0889-1583. DOI: 10.1016/j.jjie.2020.101104.
- Marcellino, Massimiliano, James H Stock, and Mark W Watson (Feb. 1, 2003). "Macroeconomic forecasting in the Euro area: Country specific versus area-wide information". In: *European Economic Review* 47.1, pp. 1–18. ISSN: 0014-2921. DOI: 10.1016/S0014-2921(02)00206-4.

- Marcellino, Massimiliano, James H. Stock, and Mark W. Watson (Nov. 1, 2006). “A comparison of direct and iterated multistep AR methods for forecasting macroeconomic time series”. In: *Journal of Econometrics* 135.1, pp. 499–526. ISSN: 0304-4076. DOI: 10.1016/j.jeconom.2005.07.020.
- Marra, Fabio Della (Mar. 28, 2017). “A forecasting performance comparison of dynamic factor models based on static and dynamic methods”. In: *Communications in Applied and Industrial Mathematics* 8.1, pp. 43–66. ISSN: 2038-0909. DOI: 10.1515/caim-2017-0003.
- Medeiros, Marcelo C. and Eduardo F. Mendes (Mar. 1, 2016). “Regularization of high-dimensional time-series models with non-Gaussian and heteroskedastic errors”. In: *Journal of Econometrics* 191.1, pp. 255–271. ISSN: 0304-4076. DOI: 10.1016/j.jeconom.2015.10.011.
- Medeiros, Marcelo C., Gabriel F. R. Vasconcelos, et al. (Jan. 2, 2021). “Forecasting Inflation in a Data-Rich Environment: The Benefits of Machine Learning Methods”. In: *Journal of Business & Economic Statistics* 39.1, pp. 98–119. ISSN: 0735-0015. DOI: 10.1080/07350015.2019.1637745.
- Menz, Jan-Oliver (2024). *Sources of post-pandemic inflation in Germany and the euro area: An application of Bernanke and Blanchard (2023)*. Working Paper 02/2024. Technical Paper.
- Naghi, Andrea A., Eoghan O’Neill, and Martina Danielova Zaharieva (2024). “The benefits of forecasting inflation with machine learning: New evidence”. In: *Journal of Applied Econometrics* 39.7, pp. 1321–1331. ISSN: 1099-1255. DOI: 10.1002/jae.3088.
- Perron, Pierre (1989). “The Great Crash, the Oil Price Shock, and the Unit Root Hypothesis”. In: *Econometrica* 57.6, pp. 1361–1401. ISSN: 0012-9682. DOI: 10.2307/1913712.
- Ribeiro, Marco Tulio, Sameer Singh, and Carlos Guestrin (Aug. 13, 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*. KDD ’16. New York, NY, USA: Association for Computing Machinery, pp. 1135–1144. ISBN: 978-1-4503-4232-2. DOI: 10.1145/2939672.2939778.
- Rossi, Barbara (Dec. 2021). “Forecasting in the Presence of Instabilities: How We Know Whether Models Predict Well and How to Improve Them”. In: *Journal of Economic Literature* 59.4, pp. 1135–1190. ISSN: 0022-0515. DOI: 10.1257/jel.20201479.
- (June 3, 2025). “Inflation forecasting”. In: *Research Handbook on Inflation*. Red. by Guido Ascari and Riccardo Trezzi. Section: Research Handbook on Inflation. Edward Elgar Publishing, pp. 440–474. ISBN: 978-1-0353-2776-8.

- (2026). “Keynote Speech: Forecasting in the Presence of Instabilities and Big Data”. In: ECB Conference on Forecasting Techniques: Artificial intelligence in the analysis of economic narratives, forecasting, and risk assessment. Vol. 13. European Central Bank.
- Sargent, Thomas and Christopher Sims (1977). *Business cycle modeling without pretending to have too much a priori economic theory*. Working paper 55. Federal Reserve Bank of Minneapolis. DOI: 10.21034/wp.55.
- Simionescu, Mihaela (Jan. 2025). “Machine Learning vs. Econometric Models to Forecast Inflation Rate in Romania? The Role of Sentiment Analysis”. In: *Mathematics* 13.1, p. 168. ISSN: 2227-7390. DOI: 10.3390/math13010168.
- Smeeke, Stephan and Etienne Wijler (July 1, 2018). “Macroeconomic forecasting using penalized regression methods”. In: *International Journal of Forecasting* 34.3, pp. 408–430. ISSN: 0169-2070. DOI: 10.1016/j.ijforecast.2018.01.001.
- Stock, James H and Mark W Watson (Apr. 1, 2002a). “Macroeconomic Forecasting Using Diffusion Indexes”. In: *Journal of Business & Economic Statistics* 20.2, pp. 147–162. ISSN: 0735-0015. DOI: 10.1198/073500102317351921.
- (2002b). “Forecasting Using Principal Components from a Large Number of Predictors”. In: *Journal of the American Statistical Association* 97.460, pp. 1167–1179. ISSN: 0162-1459.
- (2007). “Why Has U.S. Inflation Become Harder to Forecast?” In: *Journal of Money, Credit and Banking* 39 (s1), pp. 3–33. ISSN: 1538-4616. DOI: 10.1111/j.1538-4616.2007.00014.x.
- (Jan. 1, 2016). “Chapter 8 - Dynamic Factor Models, Factor-Augmented Vector Autoregressions, and Structural Vector Autoregressions in Macroeconomics”. In: *Handbook of Macroeconomics*. Red. by John B. Taylor and Harald Uhlig. Vol. 2A. Elsevier, pp. 415–525. ISBN: 978-0-444-59469-3. DOI: 10.1016/bs.hesmac.2016.04.002.
- Tibshirani, Robert (Jan. 1, 1996). “Regression Shrinkage and Selection Via the Lasso”. In: *Journal of the Royal Statistical Society: Series B (Methodological)* 58.1, pp. 267–288. ISSN: 0035-9246. DOI: 10.1111/j.2517-6161.1996.tb02080.x.
- Vancsura, László, Tibor Tatay, and Tibor Bareith (Dec. 2025). “Enhancing Policy Insights: Machine Learning-Based Forecasting of Euro Area Inflation HICP and Subcomponents”. In: *Forecasting* 7.4, p. 63. ISSN: 2571-9394. DOI: 10.3390/forecast7040063.
- Wu, J. C. and F. D. Xia (2017). “Time-varying lower bound of interest rates in Europe (Chicago Booth Research Paper no. 17-06)”. In: *University of Chicago Booth School of Business*.
- Zhao, Qingyuan and Trevor Hastie (Jan. 2, 2021). “Causal Interpretations of Black-Box Models”. In: *Journal of Business & Economic Statistics* 39.1, pp. 272–281. ISSN: 0735-0015. DOI: 10.1080/07350015.2019.1624293.

- Zivot, Eric and Donald W. K. Andrews (1992). “Further Evidence on the Great Crash, the Oil-Price Shock, and the Unit-Root Hypothesis”. In: *Journal of Business & Economic Statistics* 10.3, pp. 251–270. ISSN: 0735-0015. DOI: 10.2307/1391541.
- Zou, Hui and Trevor Hastie (Apr. 1, 2005). “Regularization and Variable Selection Via the Elastic Net”. In: *Journal of the Royal Statistical Society Series B: Statistical Methodology* 67.2, pp. 301–320. ISSN: 1369-7412. DOI: 10.1111/j.1467-9868.2005.00503.x.

Appendix

Table 7.1: Predictor panel: source, frequency, and transformation

Variable	Description	Source	Freq.	Transformation
Employment				
UNETOT_EA	Unemployment rate, total	EA-MD-QD	M	Level
UNE025_EA	Unemployment rate, 25+	EA-MD-QD	M	Level
UNEU25_EA	Unemployment rate, under 25	EA-MD-QD	M	Level
Consumption				
CCONFIX_EA	Consumer confidence index (EC)	EA-MD-QD	M	Level
RTCONFIX_EA	Retail confidence indicator	EA-MD-QD	M	Level
SCONFIX_EA	Services confidence indicator	EA-MD-QD	M	Level
CCI_EA	Consumer Confidence Index	EA-MD-QD/OECD	M	Level
Interest and exchange rates				
IRT3M_EACC	3-month interest rate	EA-MD-QD	M	Level
IRT6M_EACC	6-month interest rate	EA-MD-QD	M	Level
LTIRT_EACC	Long-term (10y, EMU criterion) rate	EA-MD-QD	M	Level
long_gbond	Germany 10y government bond yield	FRED (IRLTLT01DEM156N)	M	Level
shadow_rate	Wu-Xia euro area shadow short rate	Wu and Xia (2016)	M	Level
REER42_EA	Real effective exchange rate (42 partners)	EA-MD-QD	M	Δx_t
ERUS_EA	EUR/USD exchange rate	EA-MD-QD	M	Δx_t
ciss	Composite Indicator of Systemic Stress (sovereign)	ECB CISS	M	Level
Output, income and housing				

(continued on next page)

(continued from previous page)

Variable	Description	Source	Freq.	Transformation
IPMN_EA	Industrial Production Index: Manufacturing	EA-MD-QD	M	$100 \times \Delta \log$
IPCAG_EA	IP Index: Capital Goods	EA-MD-QD	M	$100 \times \Delta \log$
IPCOG_EA	IP Index: Consumer Goods	EA-MD-QD	M	$100 \times \Delta \log$
IPDCOG_EA	IP Index: Durable Consumer Goods	EA-MD-QD	M	$100 \times \Delta \log$
IPNDCOG_EA	IP Index: Non-Durable Consumer Goods	EA-MD-QD	M	$100 \times \Delta \log$
IPING_EA	IP Index: Intermediate Goods	EA-MD-QD	M	$100 \times \Delta \log$
IPNRG_EA	IP Index: Energy	EA-MD-QD	M	$100 \times \Delta \log$
TRNMN_EA	Turnover Index: Manufacturing	EA-MD-QD	M	$100 \times \Delta \log$
TRNCAG_EA	Turnover Index: Capital Goods	EA-MD-QD	M	$100 \times \Delta \log$
TRNCOG_EA	Turnover Index: Consumer Goods	EA-MD-QD	M	$100 \times \Delta \log$
TRNDCOG_EA	Turnover Index: Durable Consumer Goods	EA-MD-QD	M	$100 \times \Delta \log$
TRNNDCOG_EA	Turnover Index: Non-Durable Consumer Goods	EA-MD-QD	M	$100 \times \Delta \log$
TRNING_EA	Turnover Index: Intermediate Goods	EA-MD-QD	M	$100 \times \Delta \log$
TRNNRG_EA	Turnover Index: Energy	EA-MD-QD	M	$100 \times \Delta \log$
CAREG_EA	Passenger Car Registrations	EA-MD-QD	M	$100 \times \Delta \log$
tot_indprod	Total industrial production (excl. construction)	ECB STBS	M	$100 \times \Delta \log$
ICONFIX_EA	Industrial confidence indicator	EA-MD-QD	M	Level

(continued on next page)

(continued from previous page)

Variable	Description	Source	Freq.	Transformation
ESENTIX_EA	Economic Sentiment Indicator	EA-MD-QD	M	Level
KCONFIG_EA	Construction sentiment indicator	EA-MD-QD	M	Level
BCI_EA	Business Confidence Index	EA-MD-QD/OECD	M	Level
cap_util	Capacity utilisation, manufacturing	Eurostat (ei_bsin_q_r2)	Q→M (ff)	Level
gov_expn	Government total expenditure, % of GDP	ECB GFS	Q→M (ff)	Level
hlw_outgap	Holston-Laubach-Williams euro area output gap	Holston, Laubach, and Williams (2023)	Q→M (ff)	Level
Money and credit				
CURR_EACC	Money stock: currency in circulation	EA-MD-QD	M	$100 \times \Delta \log$
M1_EACC	Money stock: M1	EA-MD-QD	M	$100 \times \Delta \log$
M2_EACC	Money stock: M2	EA-MD-QD	M	$100 \times \Delta \log$
m3	Money stock: M3, outstanding amount	ECB BSI	M	$100 \times \Delta \log$
loans_nfc	Loans to non-financial corporations	ECB BSI	M	$100 \times \Delta \log$
loans_hh	Loans to households	ECB BSI	M	$100 \times \Delta \log$
loans_total	Loans, private sector total	ECB BSI	M	$100 \times \Delta \log$
Prices and wages				
PPICAG_EA	Producer Price Index: Capital Goods	EA-MD-QD	M	$100 \times \Delta \log$
PPICOG_EA	Producer Price Index: Consumer Goods	EA-MD-QD	M	$100 \times \Delta \log$

(continued on next page)

(continued from previous page)

Variable	Description	Source	Freq.	Transformation
PPIDCOG_EA	Producer Price Index: Durable Consumer Goods	EA-MD-QD	M	$100 \times \Delta \log$
PPINDCOG_EA	Producer Price Index: Non-Durable Consumer Goods	EA-MD-QD	M	$100 \times \Delta \log$
PPIING_EA	Producer Price Index: Intermediate Goods	EA-MD-QD	M	$100 \times \Delta \log$
PPINRG_EA	Producer Price Index: Energy	EA-MD-QD	M	$100 \times \Delta \log$
brent	Brent crude oil price	FRED (DCOILBRETEU)	M	$100 \times \Delta \log$
ttf_gas	EU natural gas price	FRED (PNGASEUUSDM)	M	$100 \times \Delta \log$
gscpi	Global Supply Chain Pressure Index	FED New York	M	Level
import_prices	Import prices, industry excl. energy	Eurostat (sts_inpi_m)	M	$100 \times \Delta \log$
nec	Nominal effective exchange rate / competitiveness indicator	ECB NEC	M	$100 \times \Delta \log$
neg_wages	Negotiated wages, growth rate	ECB STS	Q→M (ff)	Level
Expectations				
spf_1y	1-year-ahead HICP inflation expectation, point forecast avg.	ECB SPF	Q→M (ff)	Level
spf_2y	2-year-ahead HICP inflation expectation	ECB SPF	Q→M (ff)	Level
spf_1t	Long-term HICP inflation expectation	ECB SPF	Q→M (ff)	Level

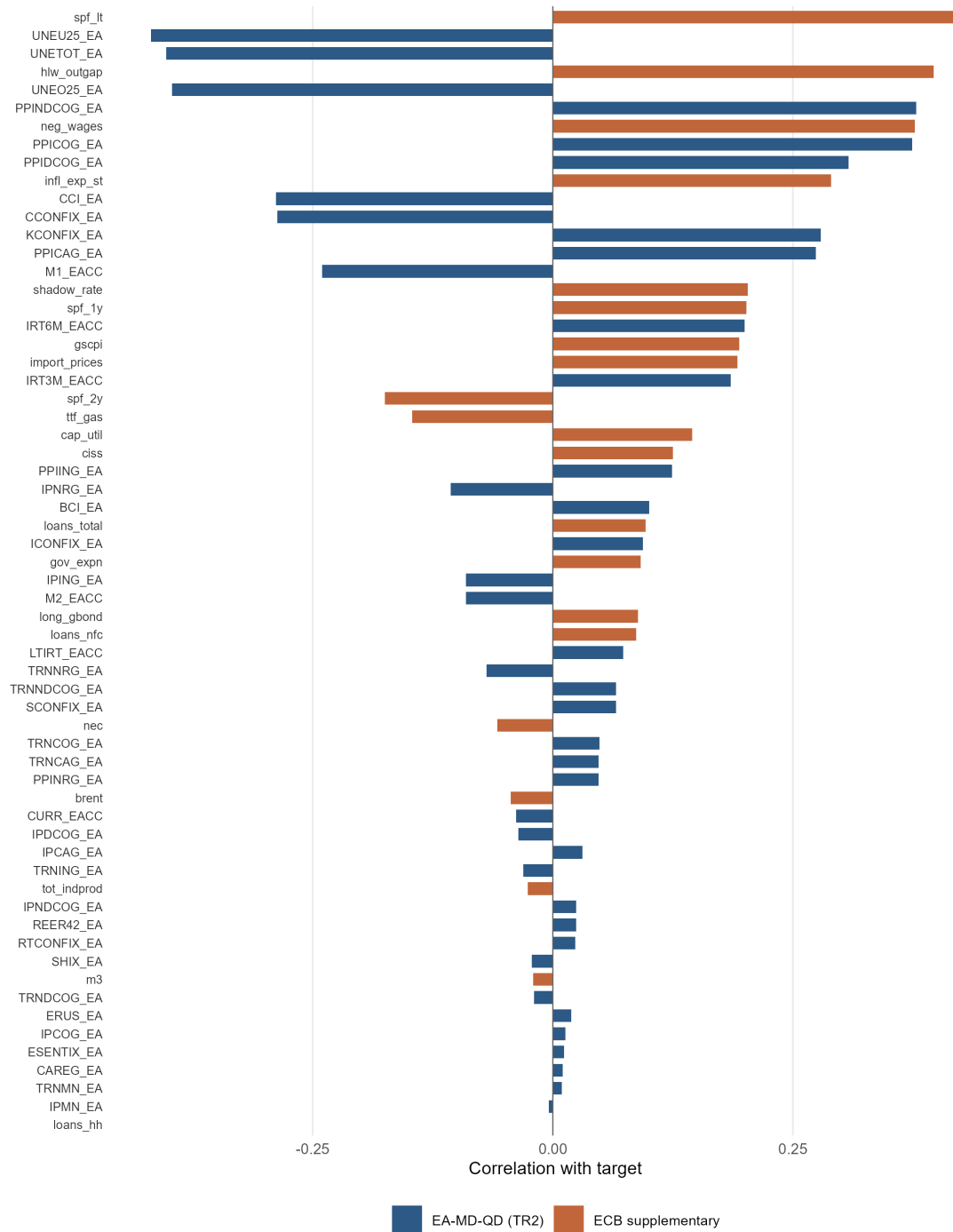
(continued on next page)

(continued from previous page)

Variable	Description	Source	Freq.	Transformation
infl_exp_st	Short-term consumer inflation expectations	Eurostat (ei_bsco_m)	M	Level
Stocks				
SHIX_EA	Stock price index	EA-MD-QD	M	$100 \times \Delta \log$

Predictor correlation with the target

Pearson correlation with core HICP (m/m annualised), full sample



Note: pairwise-complete correlations over the full estimation sample.

Figure 7.1: Pearson correlation of every predictor with core HICP inflation (month-on-month, annualised), full sample, pairwise-complete. Bars are coloured by source block (EA-MD-QD TR2 vs. ECB supplementary series, Section 4)

RMSFE, MAE, and MAD ratios to RW(3m), by horizon ($h = 1, \dots, 12$), full sample. Bold marks the lowest ratio per column. Lasso/Ridge/Elastic Net rows are shaded by the % change from adding one predictor lag (green = improved, red = worsened; baseline and lag-variant share the same colour).

Model	Forecasting horizon											
	1	2	3	4	5	6	7	8	9	10	11	12
RMSFE ratio												
<i>Models</i>												
DFM	0.87	0.90	0.86	0.87	0.84	0.84	0.91	0.94	0.97	0.99	0.99	1.01
Lasso	1.02	0.97	0.90	0.90	0.90	0.90	0.89	0.91	0.97	0.90	0.94	0.93
Lasso (+1 lag)	1.01	0.94	0.87	0.86	0.86	0.89	0.91	0.92	0.91	0.92	0.97	1.02
Ridge	1.04	0.99	0.90	0.88	0.86	0.87	0.87	0.88	0.89	0.90	0.90	0.93
Ridge (+1 lag)	1.04	0.94	0.85	0.83	0.79	0.82	0.84	0.83	0.84	0.87	0.88	0.88
Elastic Net	1.03	0.97	0.91	0.90	0.90	0.91	0.89	0.88	0.96	0.89	0.93	0.93
Elastic Net (+1 lag)	1.01	0.95	0.88	0.89	0.86	0.89	0.90	0.90	0.91	0.92	0.97	1.01
Random Forest	0.97	0.96	0.93	0.99	1.06	1.15	1.17	1.18	1.19	1.18	1.16	1.16
XGBoost	1.03	1.08	1.04	1.12	1.24	1.27	1.28	1.25	1.24	1.22	1.21	1.22
UCSV	0.99	0.99	0.95	0.94	0.95	0.94	0.94	0.94	0.95	0.95	0.94	0.95
RW (12m)	0.97	0.97	0.96	0.98	0.99	1.01	1.01	1.02	1.02	1.02	1.02	1.02
<i>Forecast combinations</i>												
Average	0.96	0.95	0.89	0.91	0.93	0.96	0.97	0.97	1.00	0.98	0.99	1.00
Trimmed Mean	0.97	0.95	0.88	0.91	0.91	0.94	0.95	0.95	0.99	0.96	0.98	0.99
Median	0.97	0.94	0.88	0.88	0.88	0.89	0.91	0.91	0.96	0.93	0.96	0.97
MAE ratio												
<i>Models</i>												
DFM	0.88	0.88	0.87	0.86	0.83	0.80	0.84	0.83	0.87	0.88	0.87	0.89
Lasso	1.00	0.92	0.92	0.93	0.92	0.88	0.92	0.93	0.96	0.92	0.95	0.95
Lasso (+1 lag)	0.99	0.90	0.91	0.89	0.87	0.87	0.91	0.89	0.88	0.88	0.92	0.93
Ridge	1.01	0.92	0.91	0.93	0.90	0.89	0.90	0.89	0.91	0.90	0.90	0.92
Ridge (+1 lag)	0.98	0.86	0.87	0.86	0.82	0.84	0.87	0.86	0.88	0.88	0.88	0.88
Elastic Net	1.00	0.93	0.92	0.92	0.92	0.89	0.91	0.91	0.97	0.90	0.93	0.94
Elastic Net (+1 lag)	0.99	0.91	0.91	0.91	0.88	0.87	0.90	0.88	0.88	0.88	0.90	0.92
Random Forest	0.94	0.91	0.93	0.96	0.97	0.98	0.98	0.97	0.99	0.98	0.99	1.02
XGBoost	1.02	1.01	1.02	1.06	1.10	1.07	1.10	1.09	1.11	1.14	1.15	1.19
UCSV	0.97	0.97	0.99	0.98	0.97	0.97	0.97	0.97	0.99	0.98	0.97	0.99
RW (12m)	0.95	0.93	0.99	0.99	0.98	0.97	0.98	0.97	1.00	0.99	0.98	0.99
<i>Forecast combinations</i>												
Average	0.94	0.90	0.90	0.91	0.92	0.89	0.91	0.89	0.92	0.90	0.91	0.94
Trimmed Mean	0.95	0.90	0.90	0.91	0.91	0.89	0.91	0.89	0.92	0.89	0.91	0.93
Median	0.95	0.90	0.90	0.90	0.89	0.86	0.89	0.88	0.90	0.87	0.90	0.92
MAD ratio												
<i>Models</i>												
DFM	0.86	0.86	0.92	0.88	0.75	0.70	0.75	0.67	0.69	0.71	0.70	0.68
Lasso	0.83	0.79	0.95	0.89	0.82	0.73	0.81	0.87	0.81	0.79	0.71	0.75
Lasso (+1 lag)	0.89	0.79	0.88	0.88	0.75	0.74	0.83	0.80	0.83	0.74	0.62	0.61
Ridge	0.97	0.83	0.92	0.95	0.83	0.76	0.84	0.86	1.01	0.84	0.81	0.80
Ridge (+1 lag)	0.84	0.70	0.86	0.90	0.78	0.79	0.92	0.79	0.85	0.84	0.81	0.77
Elastic Net	0.84	0.83	0.94	0.90	0.80	0.76	0.80	0.90	0.86	0.80	0.73	0.73
Elastic Net (+1 lag)	0.88	0.83	0.87	0.86	0.78	0.79	0.81	0.81	0.85	0.70	0.67	0.65
Random Forest	0.95	0.86	0.91	0.86	0.80	0.75	0.71	0.63	0.66	0.58	0.64	0.72
XGBoost	0.95	0.89	0.93	0.98	0.89	0.82	0.83	0.89	0.99	0.95	1.02	1.03
UCSV	0.97	1.00	1.14	1.10	0.93	0.92	0.99	1.04	1.02	0.95	0.99	0.94
RW (12m)	0.91	0.80	1.01	0.91	0.82	0.81	0.83	0.86	0.94	0.80	0.87	0.83
<i>Forecast combinations</i>												
Average	0.86	0.81	0.87	0.87	0.80	0.73	0.82	0.76	0.78	0.74	0.73	0.77
Trimmed Mean	0.88	0.80	0.86	0.86	0.80	0.75	0.79	0.76	0.79	0.70	0.69	0.73
Median	0.88	0.80	0.87	0.88	0.79	0.73	0.78	0.77	0.81	0.73	0.73	0.77

Note: Values below 1 indicate the model is more accurate than RW(3m) on that metric/horizon. Full sample; computed from a separate recursive re-estimation at every horizon 1-12

Table 7.2: Relative forecast accuracy: RMSFE, MAE and MAD after introducing lagged predictors