Forecasting Dutch inflation using machine learning methods*

Robert-Paul Berben†[1], Rajni N. Rasiawan[1,2], and Jasper M. de Winter[1]

[1] De Nederlandsche Bank
[2] Vrije Universiteit Amsterdam

November 11, 2025

Abstract

This paper evaluates the forecasting performance of machine learning models for Dutch inflation over the period 2010-2023, leveraging a large dataset. We find that certain machine learning models outperform simple benchmark models, particularly in forecasting core inflation and services inflation. However, these models face challenges in consistently outperforming the inflation forecasts of De Nederlandsche Bank. Notably, they demonstrate potential in enhancing forecasts for non-energy industrial goods inflation. path-average forecasts are more accurate than direct forecasts, while the inclusion of non-linearities, factors, or targeted predictors yields limited improvements in forecasting performance. Overall, Ridge regression achieves the highest forecast accuracy among the models considered.

Keywords: Inflation forecasting, Big data, Machine learning, Random Forest, Ridge regression
JEL Classification: C22, C53, C55, E17, E31

* We gratefully acknowledge valuable comments from Maurice Bun, Peter van Els, Kostas Mavromatis, participants of the DNB and Ortec Finance research seminars, and attendees of the ECB Workshop "Harnessing Artificial Intelligence for Inflation Assessment". The views expressed in this paper are solely those of the authors and do not necessarily reflect those of De Nederlandsche Bank or the Eurosystem.
† Corresponding author: r.p.berben@dnb.nl.

1 Introduction

Inflation forecasts serve as critical inputs for monetary policy formulation. However, accurately forecasting inflation remains a considerable challenge, as simple time series models frequently produce the most accurate forecasts. Quoting Faust and Wright (2013) "We find that [...], extremely simple inflation forecasts—that however take account of nowcasting and secular changes in the local mean inflation rate—are just about the best that are available." In recent years, however, various papers have been published showing that the combination of "big data" and machine learning (ML) models can lead to more accurate inflation forecasts. For example, Medeiros et al. (2021) apply several ML models using the FRED-MD database (McCracken and Ng, 2016) to forecast U.S. consumer price index (CPI) inflation, demonstrating that these models outperform conventional benchmarks, with the random forest (RF) model yielding the smallest forecast errors. This paper examines the broader applicability of ML models by applying them to inflation in the Netherlands, one of the larger economies in the euro area.

We contribute to the existing literature by evaluating the effectiveness of ML models in forecasting Dutch HICP inflation across both the pre- and post-COVID-19 pandemic periods. Model performance is benchmarked against both standard time series approaches and the official inflation projections published by De Nederlandsche Bank (DNB), the central bank of the Netherlands.

Building on Medeiros et al. (2021), we utilize a large-scale dataset and evaluate multiple ML models for forecasting the year-on-year inflation rate. The dataset comprises 129 monthly Dutch and international time series, spanning January 1990 to January 2024. Target variables include the Harmonized Index of Consumer Prices (HICP) and three of its sub-components: services inflation (HICPS), non-energy industrial goods (NEIG) inflation, and core inflation (HICP excluding energy and food). To generate year-on-year inflation forecasts, we implement two distinct approaches. The first is a direct approach, which forecasts the year-on-year inflation rate using a single model—a conventional method in the literature. The second, introduced by Goulet Coulombe et al. (2021) and referred to as the path-average method, involves forecasting month-on-month changes in the price index and aggregating them to derive the year-on-year inflation rate.

Our main results can be summarized as follows. First, ML models outperform a simple benchmark random walk model in forecasting Dutch inflation, especially for core inflation and services inflation. Second, non-linear models, such as random forests, while effective in certain contexts,

do not consistently outperform linear models. Third, ML models generally yield more accurate forecasts at longer horizons (6 to 12 months) than at shorter ones. Additionaly, path-average forecasts frequently outperform direct methods in forecasting year-on-year inflation. Fourth, ML models struggle to consistently outperform DNB's official inflation forecasts. However, for NEIG inflation, certain ML models outperform DNB's forecast accuracy over specific horizons. Finally, based on the comparative forecasting performance accros ML models and benchmarks, the preferred ML model is path-average Ridge regression.

The remainder of the paper is structured as follows. Section 2 provides a review of the literature. Section 3 outlines the dataset, the forecast framework and the forecast evaluation metrics. Section 4 introduces the ML models employed in the analysis. Section 5 presents the main empirical findings. Section 6 offers concluding remarks. Appendix A provides detailed information on the dataset. A comprehensive set of results and supplementary figures is available in the online appendix (here).

2 Literature

Since the seminal article by Medeiros et al. (2021), a growing body of research has demonstrated the usefulness of ML models in forecasting inflation. Naghi et al. (2024) build upon Medeiros et al. (2021) by extending the sample period to 2022, significantly broadening the set of ML models considered, and applying the methodology to U.S., Canadian and U.K. datasets. Their findings suggest that the reported superiority of the RF model for forecasting U.S. inflation extends to U.K. inflation, though less convincingly to Canadian data. Moreover, the RF model exhibits diminished performance on U.S. data from 2020 onward. Similarly, Huang et al. (2025) use a large dataset of macroeconomic and financial predictors to forecast Chinese CPI inflation and producer price index (PPI) inflation. They find that penalized linear regression models outperform standard benchmark models. Maehashi and Shintani (2020) analyze Japanese CPI and wholesale price index (WPI) inflation, among other things. Their results indicate that beyond the very short-term horizon, both the RF model and boosted trees outperform the benchmark autoregressive (AR) model. Das and Das (2024) conclude that for Indian CPI inflation, the RF model yields substantially more accurate forecasts than the ARIMA model, but only when the COVID-19 pandemic period is included. Using pre-pandemic data, however, the ARIMA model remains difficult to outperform.

Lenza et al. (2025) apply the quantile random forest (QRF) model to euro area inflation data. Their QRF-based density forecasts demonstrate competitive performance relative to those produced by linear models and the European Central Bank's Survey of Professional Forecasters.

While inflation dynamics in major economic blocks have been extensively studied, empirical evidence for smaller economies, such as the Netherlands, remains relatively limited. Kohlscheen (2022) employs a RF model to examine the drivers of CPI inflation across a panel of high-income countries, including the Netherlands; however, country-specific results for the Netherlands are not reported. Vedder and van de Winkel (2024), replicate the methodology of Goulet Coulombe et al. (2022) using Dutch data and demonstrate that several ML models yield significantly more accurate CPI inflation forecasts than an AR model.

Our analysis also relates to the literature on multi-period-ahead forecasting, which encompasses a range of methodological approaches. Marcellino et al. (2006) use a large dataset of U.S. macroeconomic time series and find that iterated AR(p) forecasts generally outperform direct AR(p) forecasts, with relative performance improving as the forecast horizon increases. Quaedvlieg (2021) reaches similar conclusions using tests for multi-horizon superior predictive ability. Kock and Teräsvirta (2016) examine multi-period-ahead forecasting using nonlinear neural network-based prediction techniques and observed that iterated and direct forecasts often yield comparable performance, with their relative ranking depending on the characteristics of the dataset. Goulet Coulombe et al. (2021) offer a novel perspective by systemically comparing direct forecasts with path-average forecasts across several target variables. Path-average forecasts can enhance accuracy by decomposing the forecasting problem into simpler, horizon-specific sub-problems. Their findings suggest that the choice between direct versus path-average forecasting is often variable specific, although path-average methods tend to perform better for variables that exhibit strong co-movement with the business cycle. Beck and Wolf (2025) confirm these findings in the context of Swiss inflation forecasting.

Finally, our analysis also relates to the literature examining the properties of institutional forecasts. In academic research, simple time series models are frequently employed as benchmark models. A more policy-relevant question is whether newly developed models can outperform existing institutional forecasts. In other words, it is crucial to assess whether incorporating ML models into the existing suite of forecasting tools yields tangible benefits. Lenza et al. (2025) show that QRF forecasts of euro area inflation perform comparably to the published Eurosystem

inflation projections. Yoon (2021) shows that boosting and RF models can generate forecasts of Japanese GDP growth that surpass those produced by the IMF and the Bank of Japan in terms of accuracy. Araujo and Gaglianone (2023) find that ML models frequently outperform traditional econometric approaches in forecasting Brazilian inflation.

3 Data, forecast design and forecast evaluation

3.1 Data

We compile a dateset comprising monthly observations of 129 Dutch and international macroeconomic time series. All series were downloaded on March 4^{th}, 2024, and span the period from January 1990 to January 2024. The target variables include four inflation series: headline HICP inflation (HICP), services inflation (HICPS), non-energy industrial goods inflation (HICPNEIG), and core inflation (HICPMEF), defined as headline HICP excluding energy and food. Following McCracken and Ng (2016) and Medeiros et al. (2021), the time series are categorized into eleven groups: (1) output & income, (2) labor market, (3) consumption, (4) orders & inventories,(5) money & credit, (6) interest & exchange rates, (7) commodity prices, (8) producer prices, (9) domestic prices, (10) price expectations, and (11) stock market. In addition to the raw monthly series, we include four principal components (factors) derived from the full set of time series as additional predictors. Further methodological details are provided in Section 4.5. We incorporate four lags of each variable and four autoregressive terms of the dependent variable, resulting in a total of 532 potential predictors. Table A.1 in Appendix A provides additional details on the time series, including seasonal adjustments and transformations of all series. We consider two approaches for constructing year-on-year inflation forecasts: the direct method and the path-average method. Section 3.3 outlines both forecasting approaches. The datasets used for forecasting differ in the treatment of trending predictors to ensure stationarity. For path-average forecasts, trending time series are rendered stationarity by applying (log) first differences. For direct forecasts, year-on-year changes in the predictors variables are employed.

Figure 1 illustrates the development of the four target inflation series. Following a prolonged period of inflation near 2 percent, headline HICP inflation rose markedly during the COVID-19 pandemic. After the first confirmed cases of COVID-19 in the Netherlands in February 2020,

inflation began a steep ascent in the mid-2021, peaking at 17.1% in September 2022 before subsequently declining. This surge was mainly driven by substantial increases in energy and food prices. As shown in Figure 1, the rise in core inflation during the pandemic was noticeable more moderate, though still apparent. Non-energy industrial goods inflation displays significantly greater volatile than core inflation. The more subdued increase in core inflation during the pandemic likely reflects the relatively stable path of services inflation.

- insert Figure 1 about here -

The transmission of elevated energy and food prices is considerably more pronounced in the industrial sector than in the services sector. This pattern is further evident when examining distributional characteristics in Figures 1. Among the inflation series, non-energy industrial goods inflation has the highest coefficient of variation (1.8), followed by headline HICP inflation (0.9), core inflation (0.7) and services inflation (0.5).

3.2 Forecast design

The first forecast is made January 2010. The forecasts are based on a rolling window with a fixed length of 216 months (18 years). This means that the number of forecasts depends on the forecast horizon. One advantage of a rolling-window framework is that it provides some protection against structural breaks or trends in the target variable. We re-calculate the seasonal inflation factors for each estimation window to avoid look-ahead bias in these factors.

We employ a quasi real-time design, taking into account the data publication delays as of our download date (March 4^{th}, 2024). However, we ignore the possibility of data revisions for the predictors, such as industrial production. The latter implies that we might overestimate the forecast accuracy of the ML models. Unfortunately, a large real-time dataset for the Netherlands does not yet exist. Furthermore, Bernanke and Boivin (2003) have shown that the scope of the dataset appears to matter more for forecast accuracy than the use of real-time (unrevised) data. Moreover, the target inflation series are not revised after the initial publication, which further mitigates the drawback of our quasi real-time design. Overall, it is very unlikely that the relative ranking of the ML models in terms of forecast accuracy will change in a meaningful way when conducting a full-fledged real-time analysis. We evaluate the out-of-sample forecast performance of the ML models over the 1- to 12-month forecast horizon ($h = 1, \ldots, 12$). Since the maximum

horizon of the institutional forecast is only 10 months, the comparison between these forecasts and the ML models is limited to the 1- to 10-month horizon.

3.3 Direct and path-average forecasts

We consider two approaches to forecast year-on-year inflation: direct forecasting and path average forecasting, cf. Goulet Coulombe et al. (2021). Let Y_t be a target inflation series (HICP, HICPMEF, HICPNEIG or HICPS), and let x_t be a large macroeconomic dataset comprising of N predictors, for $t = 1, \ldots, T$. When needed, the predictors are transformed to stationarity by taking annual (log) differences. Y_t is not seasonally adjusted. Our target variable is y_{t+h}, the year-on-year percentage change in Y_t, h periods into the future: $y_{t+h} = 100 * (Y_{t+h} - Y_{t+h-12})/Y_{t+h-12}$. Direct forecasts of y_{t+h} can be obtained using the following prediction model:

$$y_{t+h}^{dir} = G_h(x_t) + \epsilon_{t+h} \tag{1}$$

where $G_h(\cdot)$ is a (potentially non-linear) mapping between the predictor variables x_t and future inflation. Denote $\hat{G}_h(x_t)$ and $\hat{y}_{t+h}^{dir}$ the fitted model and its forecast. Then the forecast error is defined as $e_{t+h} = y_{t+h} - \hat{y}_{t+h} = y_{t+h} - \hat{G}_h(x_t)$. Following Goulet Coulombe et al. (2021), another method to compute forecasts of year-on-year inflation h periods into the future is by 'averaging' 1 to h periods ahead forecasts of month-on-month changes in Y_t. Consider the following alternative prediction model:

$$y'_{t+h} = G_h(x'_t) + \epsilon'_{t+h} \tag{2}$$

where $y'_{t+h} = \ln(Y'_{t+h}/Y'_{t+h-1})$ and Y'_t is equal to Y_t corrected for seasonal factors, ζ_t, i.e. $Y_t = Y'_t + \zeta_t$. x'_t is a second macroeconomic dataset, in which predictors are transformed to stationarity by taking first (log) differences, see Appendix A. A path-average forecast of year-on-year inflation, denoted by $\hat{y}_{t+h}^{pa}$, is then computed as:

$$\hat{Y}_{t+h}^{pa} = \exp(\ln(Y'_t) + \sum_{i=1}^{h} \hat{y}'_{t+i}) + \zeta_{t+h} \tag{3}$$

$$\hat{y}_{t+h}^{pa} = 100 * (\hat{Y}_{t+h}^{pa} - Y_{t+h-12})/Y_{t+h-12} \tag{4}$$

In this case, the forecast error is calculated as $e_{t+h} = y_{t+h} - \hat{y}_{t+h}^{pa}$.

3.4 Forecast evaluation

Following the literature, we measure the accuracy of the forecasts using the root mean squared forecast error (RMSFE). We compute the gain in RMSFE relative to a benchmark model for each horizon separately. The standard Diebold and Mariano (1995) (DM) test procedure is utilized to test the significance of the gains in relative predictive accuracy[1].

To gain insight into the evolution of the forecast errors over time, we examine the forecast accuracy of the ML models over both the full sample and over the pre-pandemic sample. The full sample runs from January 2010 (2010M1) to December 2023 (2023M12), meaning that the first forecast is produced in January 2010 and the last (1-month ahead) forecast in December 2023. The pre-pandemic sample ends in January 2019, ensuring that even the 12-month ahead forecast refers to a date before the outbreak of the pandemic. Furthermore, we calculate the cumulated sum of squared forecast errors difference (CSSFED), which is defined as:

$$CSSFED_{M,h} = -\sum_{t=t_0}^{t_1}(e^2_{M,t,h} - e^2_{BM,t,h}) \tag{5}$$

where M represents a ML model and BM represents a benchmark model. A CSSFED *above* zero indicates that the forecasts of the ML model have a *lower* CSSFE up until that point in time, and are therefore more accurate than the benchmark model's forecasts. Conversely, a CSSFED *below* zero indicates that the benchmark model has higher forecast accuracy at that point in time. Additionally, a *decrease* in the CSSFED indicates that the model performance of the ML model is decreasing relative to the benchmark model, while an *increase* indicates the opposite.

Finally, we analyze the influence of several model specifications in the spirit of Goulet Coulombe et al. (2022). First, we compare the forecast accuracy of direct forecasts against path-average forecasts. Second, we compare forecasts derived from a dataset that includes factors among the predictors to those that exclude them. Third, we analyze the impact of using targeted predictors. We follow the approach of Bai and Ng (2008) to target the variables and select approximately 30 relevant predictors using soft thresholding combined with the LASSO regularizer. For each

[1] Results for the test for average Superior Predictive Ability (Quaedvlieg, 2021) and the Model Confidence Set (Hansen et al., 2011) are available in the online appendix (here).

model M, we compute the $OOS.R^2$ as follow:

$$OOS.R^2_{M,t,h} = 1 - \frac{\epsilon^2_{M,t,h}}{\frac{1}{T}\sum_{t=1}^{T}(y_{t+h} - \hat{y}_{t+h})} \tag{6}$$

where $\epsilon^2_{M,t,h}$ is the squared forecast error of model M for horizon h at time t. To assess the influence of feature f, we run Diebold-Mariano-style regressions for model M_f, which includes the feature, and model M_{-f}, which does not include the feature:

$$OOS.R^2_{M_f,t,h} - OOS.R^2_{M_{-f},t,h} = \alpha_{M_f} + \nu_{t,h} \tag{7}$$

The main advantage of equation (7) is that the coefficient α_{M_f} can be interpreted as the gain in the $OOS.R^2$, and is not unit- or series-dependent. We then count the number of models for which the version with feature f makes significantly more accurate forecasts than the version without feature f, and vice versa. We also calculate the median difference between $OOS.R^2$ for models with and without feature f, separately for the various model types.

4 Models

We construct h-period ahead forecasts of the target variable y_t using various models, distinguishing five 'types' of models: benchmark models, shrinkage models, tree models, ensemble models, and factor models. Model names are abbreviated according to the following convention: `[method].[model].[factor].[modelsel]`. `[method]` can take on two values, depending of the method used for constructing the forecasts. `[method]` equal to `Y` ('year-on-year') denotes direct forecasts. Path-average forecasts are represented by `M`, as they are computed using forecasts of month-on-month changes in the price index. `factor` indicates whether factors are included in the set of predictors (`F`) or not (`NF`). For some models, tuning of the parameters is done in multiple ways. In those cases, `[modelsel]` can take on the values `BIC` (Bayesian Information Criterion) and `CV` (Cross Validation). If only a single method is used to tune the model, `[modelsel]` is left blank. In the following, we present a brief overview of the models we use in our analysis.

4.1 Benchmark model

The first model in the class of benchmark models is the random walk model. Direct forecast using the random walk model (`Y.RW`) matches the no-change forecast in Atkeson and Ohanian (2001), i.e. past year's inflation is used as a forecast: $y_{t+h} = y_t$. In path-average forecasts, on the other hand, the no-change forecast (`M.RW`) is past month's month-on-month inflation, similar to Stock and Watson (1999): $y'_{t+h} = y'_t$. The second benchmark extends the two random walk models with a 'drift' parameter: $y_{t+h} = \beta_0 + y_t$ (`Y.RWD`) and $y'_{t+h} = \beta_0 + y'_t$ (`M.RWD`), respectively. The third benchmark is the autoregressive model of order p, where p is determined by the BIC, with a maximum of 4 lags. The AR(p) is used both to produce direct forecasts $y_{t+h} = \beta_0 + \beta_1 y_t + \cdots \beta_p y_{t+p-1}$ (`Y.AR`) and path-average forecasts $y'_{t+h} = \beta_0 + \beta_1 y'_t + \cdots \beta_p y'_{t+p-1}$ (`M.AR`), respectively. The parameters are estimated by OLS. Considering all forecast horizons (12) and target variables (4), `M.RWD` has the lowest RMSFE in 23 out of the 48 cases, and is hence selected as the main benchmark model and referred to as 'the benchmark model' in the remainder of the paper.

4.2 Shrinkage models

We estimate several shrinkage estimators where $G_h(x_t) = \beta_h x_t$. All methods minimize the objective function:

$$\sum_{t=1}^{T} \left\{ (y_{t+h} - \beta_h x_t)^2 + \lambda J(\beta_h) \right\} \tag{8}$$

where λ is the hyperparameter determining the degree of regularization. The methods differ in terms of the specification of the penalty term $J(\beta_h)$. We choose λ either by `BIC` or `CV`.

LASSO (`LAS`) The least absolute shrinkage and selection operator (LASSO) was introduced by Tibshirani (1996) and corresponds to the penalty term given by $J(\beta_h) = \sum_{i=1}^{N} |\beta_{h,i}|$. The penalty term of LASSO is the L_1 norm, which shrinks parameters of irrelevant predictors to zero. To achieve consistent model selection, Zou (2006) proposed adaLASSO (`ALAS`). AdaLASSO is similar to LASSO but includes weighting parameters ω_i obtained from a first-step estimation. In this paper, we use LASSO for this purpose, with the penalty given by $J(\beta_h) = \sum_{i=1}^{N} \omega_i |\beta_{h,i}|$.

Ridge regression (`RR`) was proposed by Hoerl and Kennard (1970) and assumes an L_2 norm penalty $J(\beta_h) = \sum_{i=1}^{N} \beta_{h,i}^2$. The parameters of less-relevant predictors can become very small,

but unlike LASSO, will rarely be exactly zero.

Elastic Net (`EN`) was designed to make the most out of LASSO and Ridge regression, and includes these models as special cases. The penalty term of Elastic Net is given by $J(\beta_h) = \omega \sum_{i=1}^{N} |\beta_{h,i}| + (1-\omega) \sum_{i=1}^{N} \beta_{h,i}^2$, where ω is an additional tuning parameter setting the relative importance of the L_1 and L_2 penalty, respectively. We follow Medeiros et al. (2021) and fix ω at 0.5. For estimation of shrinkage models we use R-package `glmnet`.

4.3 Tree models

A regression tree with S terminal nodes can be written as:

$$y_{t+h} = \sum_{i=1}^{S} \beta_{h,i} 1_{\{x_t \in R_i\}} + \epsilon_{t+h} \tag{9}$$

where $1_{\{\cdot\}}$ is an indicator function, R_i is a partition of the space of x_t and $\beta_{h,i}$ is the sample average of y_{t+h} given $x_t \in R_i$. Estimation of (9) entails finding the best tree structure to minimize $\sum_{t=1}^{T} \epsilon_{t+h}^2$. A strength of regression trees is its capability to deal with non-linearity and interaction terms among predictors. A weakness, though, is that due to over-fitting the out-of-sample forecasting properties can be very poor. To deal with the issue of over-fitting, we consider three types of ensemble methods.

Random Forest The random forest (`RF`) model was introduced by Breiman (2001). The method is based on bootstrap aggregation (bagging) of randomly constructed regression trees. While the forecast of a regression tree in each bootstrap sample may suffer from over-fitting, averaging forecasts of bootstrap samples diminishes the variation and yields a more stable forecast. Ideally, regression trees of different bootstrap samples should not be highly correlated, since otherwise averaging may not be effective in lowering the variance of the forecast. In the random forest model, a dropout procedure is used to de-correlate the regression trees of the bootstrap samples. For estimation we use R-package `ranger`, with default settings and 500 trees.

Boosted Tree In Boosted Trees (`BTREE`) multiple regression trees are constructed similarly to bootstrap aggregation to overcome over-fitting. The algorithm starts with an initial regression tree:

$$f_0(x_t) = \sum_{i=1}^{S} \beta_{h,i} 1_{\{x_t \in R_i\}} \tag{10}$$

Then the model is updated in an iterative fashion from $k-1$ to k according to the following rule:

$$f_k(x_t) = f_{k-1}(x_t) + \eta \sum_{i=1}^{S_k} \beta_{h,k,i} 1_{\{x_t \in R_{k,i}\}} \tag{11}$$

where η is a learning rate (that we set at 0.05) and the regression tree in step k is estimated from the residual from step $k-1$, $y_{t+h} - f_{k-1}(x_t)$. For estimation we use R-package `xgboost`, with the maximum depth of a tree equal to 4 and the maximum number of boosting iterations set to 1000.

BART The `BART` model is a sum-of-trees model introduced by Chipman et al. (2010). The Bayesian approach addresses the problem of over-fitting by using prior distributions to regularize the fit of each individual tree. Consequently, each tree explains only a small fraction of the variation in the target variable. To compute `BART` forecasts, we follow recommendations of Prüser (2019). We set the number of trees to 200, and α and β, which jointly control the depth of the trees, to 0.1 and 1, respectively. The hyperparameters k and q, which together determine the tightness of the prior of the values of the terminal nodes, are set at 2 and 0.9.

Borup et al. (2023) argue that "RF applied in high dimensions without an initial weeding out of irrelevant predictors may fail to reach its full potential". We follow their advise and add `TRF` and `TBART` to the list of tree models. In these models, we first select around 30 relevant targeting predictors, applying soft thresholding in combination with the LASSO regularizer, as suggested by Bai and Ng (2008).

4.4 Ensemble models

Random Subset Regression Unlike traditional regression methods that select a single best subset of predictors, the idea of Complete Subset Regression (Elliott et al., 2013) is to generate a large number (K) of forecasts based on different subsets of x_t, denoted x_t^i. The final forecast $\hat{y}_{t+h}$ is then computed by taking a simple average of the individual forecasts $\hat{y}_{t+h}^i$:

$$y_{t+h}^i = \beta_h x_t^i + \epsilon_{t+h}, \qquad i = 1 \dots K \tag{12}$$

$$\hat{y}_{t+h} = \sum_{i=1}^{K} \hat{y}_{t+h}^i / K \tag{13}$$

In this paper, we average over 1000 Random Subset Regressions (RSR), cf. Boot and Nibbering (2019). In each regression, the number of predictors, in addition to the four lags of the target variable, is randomly selected, with a maximum of four. We also investigate the performance of a targeted version of Random Subset Regression (TRSR), following Bai and Ng (2008) and Kotchoni et al. (2019).

Bagging In bagging (BAGG), bootstrap samples of the original predictor variables and the target variable are repeatedly generated. We construct $K = 100$ samples using the block bootstrap method, with a fixed block length of five months. For each bootstrap sample, we select around 10 relevant predictors using soft thresholding. Then, a regression is applied to compute a forecast y^i_{t+h}. The final forecast from bagging is constructed as the simple average of the forecasts from the individual bootstrap samples.

4.5 Factor models

Following Stock and Watson (1999), we compute k common factors F^k_t as the first k principle components of all predictor variables x_t. The h-period ahead forecast is constructed by running a principal components regression of the form:

$$y_{t+h} = \beta_h F^k_t + \epsilon_{t+h} \tag{14}$$

where β_h is a $(1 \times k)$ vector of coefficients. To select the number of factors, we consider the information criteria of Bai and Ng (2002). In the full sample, their IC_1, IC_2 and IC_3 criteria suggests that 2, 2, and 9 factors, respectively, are appropriate. We select four factors, consistent with Medeiros et al. (2021). In addition to this plain vanilla factor model (FACT), we also include a factor model with targeted predictors (TFACT) and a boosted factor model (BFACT). In the targeted factor model, we follow Bai and Ng (2008) and select around 30 relevant predictors using soft thresholding in combination with the LASSO regularizer. In the boosted factor, we adopt the boosting algorithm as in Bai and Ng (2009) to the select the factors and the number of lags in the model. The maximum number of factors and lags is set at 10 and 4, respectively.

5 Results

This section presents the results of the out-of-sample forecasting experiments. Subsection 5.1 provides graphical representations of relative RMSFEs across all target variables, models, and forecast horizons. Subsection 5.2 identifies the best-performing models for each target variable and forecast horizon. Subsection 5.3 examines which model specification choices enhance forecast accuracy, following Goulet Coulombe et al. (2021). Finally, Subsection 5.4 evaluates whether ML models can outperform DNB's official inflation forecast. Detailed results, including RMSFEs, $OOS.R^2$ statistics, p-values from the Diebold and Mariano-test as well as results for the test for average Superior Predictive Ability (Quaedvlieg, 2021) and the Model Confidence Set (Hansen et al., 2011), are available in the online appendix (here).

5.1 Forecast accuracy of machine-learning models: a bird's eye view

Figures 2 and 3 present the RMSFE of all 61 models, compared to `M.RWD`—the path-average random walk with drift model—across the full sample and the pre-pandemic sample, respectively. Individual model performances are represented by blue dots, indicating their relative RMSFE. To illustrate the central tendency of the RMSFE distribution, the bold vertical line indicates the median across all models, while the red dot marks the mean. The boxplot boundaries correspond to the 75th and 25th percentiles, summarizing the interquartile range (IQR) of the distribution. The right whisker is positioned at the smaller of the maximum RMSFE value and the 75th percentile plus 1.5$\times$ IQR. The left whisker is positioned at the larger of the minimum RMSFE value and the 25th percentile minus 1.5$\times$IQR.

Figures 2 and 3 show that outperforming a simple benchmark model (`M.RWD`) at short forecast horizons is quite challenging. However, ML models demonstrate substantial gains in forecast accuracy at longer horizons, particularly when evaluated over the full sample. The magnitude of improvement over the benchmark varies considerably across models, ranging from negligible to as much as 48%, depending on the target variable and forecast horizon.

Second, ML models show a steadily increasing gain in the forecast accuracy for services inflation. These gains are markedly smaller in the pre-pandemic sample, likely reflecting the relative stability of services inflation before the onset of the COVID-19 pandemic.

Third, for NEIG inflation, most models fail to outperform the simple benchmark model in the

pre-pandemic sample. However, in the full sample, several models outperform the benchmark, particularly at medium-term forecast horizons. The findings for NEIG and services inflation are consistent with those for core inflation, where the median forecast across all ML model clearly outperforms the benchmark in the full sample.

- insert Figure 2 - 3 about here -

Fourth, the contrast between the full sample and the pre-pandemic sample is particularly notable for headline HICP inflation. While most models are outperformed by the benchmark model (`M.RWD`) in the full sample, their relative performance improves with the forecast horizon in the pre-pandemic sample. This discrepancy is likely related to the heightened volatility in food and energy prices during the post-pandemic period. These fluctuations had a direct impact on headline HICP inflation, whereas the other inflation series were either unaffected or indirectly influenced through round-effects. The ML models appear ill- equipped to capture these dynamics[2].

Fifth, both the full-sample and the pre-pandemic sample figures show that the gap between the median and mean forecast errors widens as the forecast horizon shortens. Specifically, the mean relative RMSFE rises above the median, suggesting that a small subset of models perform substantially worse at shorter horizons. This implies that the distribution of forecast errors becomes increasingly skewed at shorter horizons, driven by the under-performance of a limited number of models.

Finally, it is noteworthy that forecast dispersion has increased since the onset of the pandemic, as indicated by the larger boxplots in Figure 2 relative to Figure 3. Specifically, the interquartile range of the RMSFE distribution in the full sample was, on average across all horizons and target variables, 3.3 times larger than in the pre-pandemic sample[3]. This evidence suggests that the post-pandemic surge in inflation has amplified differences in model accuracy, thereby improving the ability to discriminate between more and less accurate forecasting models. The next section examines these shifts in model performance in greater detail.

[2] This may also reflect limitations in the predictor set used, particularly the lack of long monthly time series on energy components and food prices for the Netherlands.

[3] The range between the 90th and 10th percentiles of the RMSFE distribution in the full sample was, on average, 2.6 times greater than in the pre-pandemic sample across all horizons and target variables.

5.2 The best performing machine learning models

To gain deeper insight into the best-performing models, we conduct a more detailed analysis of their forecasting performance. Figure 4 and 5 show the *relative* forecasting performance of all 62 models compared to the benchmark model ($\text{RMSFE}_{\text{ML}}/\text{RMSFE}_{\text{M.RWD}}$) offering a more garnular view of which model specifications perform best within each model class. White cells indicate the forecast accuracy of the ML model is not statistically different from that of the benchmark model, based on the Diebold and Mariano test at the 10% significance level. Green cells indicate a relative improvement in forecast accuracy of 5% or more, while red cells indicate a deterioration of at least 5%. The color intensity ranges from light to dark green (or red), corresponding to improvements (or declines) in forecast accuracy ranging from 5% to 50% or more.

The first conclusion from Figure 4 it that, in the full sample, none of the models consistently outperforms the benchmark model in forecasting headline HICP inflation. Moreover, using direct forecasts (denoted by `Y.`) often results in substantially lower forecast accuracy compared to path-average forecasts (denoted by `M.`). A possible explanation is that the volatility in headline HICP inflation is primarily driven by fluctuations in energy and food prices, while the predictor set lacks sufficient coverage of energy- and food prices related series. Additionally, (announced) tax changes may also contribute to this volatility and are not adequately captured by the models.

Second, for core inflation, NEIG inflation and services inflation, ML models significantly outperform the benchmark across all horizons. Specifically, three path-average Ridge regression specifications—`M.RR.NF.BIC`, `M.RR.F.CV`, and `M.RR.F.BIC`—consistently rank among the top-performing models for HICPMEF, HICPNEIG and HICPS. These models outperform the benchmark for core and services inflation across all forecast horizons, often by a substantial margin. For NEIG inflation, these Ridge regressions outperform the benchmark in at least seven forecast horizons.

Third, some factor model specifications yield lower forecast errors than Ridge regressions, though their performance is less consistent across horizons and inflation measures. For example, the `Y.BFACT.F` model outperforms Ridge regressions in forecasting core inflation 4 to 8 months ahead, as indicated by darker green cells. However, at certain horizons, this factor model fails to outperform the benchmark model, whereas Ridge regressions frequently do—often by a considerable margin.

Fourth, the forecast accuracy of the tree-based models is relatively poor. They do not consistently outperform linear shrinkage and factor models, as indicated by the darker green cells associated with the latter. Moreover, tree-based models show less consistent performance across horizons and inflation measures. The only model that consistently outperforms the benchmark in the full sample is the path-average random forest model with factors (`M.RF.F`). This is a notable finding, given that tree-based methods are frequently among the top-performing models in the existing literature. For instance, Medeiros et al. (2021) calculate the average p-values of the Model Confidence Set (MCS) introduced by Hansen et al. (2011) across all forecast horizons, finding that this average is higher for the Random Forest (0.95) than for the Ridge Regression (0.77), where the best model corresponds to the highest p-value. We find that this p-value is higher for the Ridge Regression than for the Random Forest, except in the case of NEIG inflation, as detailed in the online appendix.

Fifth, the results indicate that path-average inflation forecasts generally outperform direct forecast specification. Red cells—indicating poor performance—are almost exclusively associated with direct forecast specifications. Strikingly, Ridge regression is the worst-performing ML model when applied using the direct method, yet becomes the best-performing model when implemented using the path-average approach.

- insert Figure 4 - 5 about here -

In the pre-pandemic sample (Figure 5), the results are somewhat different from those observed over the full sample. Several ML model now now outperform the benchmark for headline HICP inflation. However, for HICPMEF, HICPNEIG and HICPS, the pattern is reversed: there are significantly fewer dark green cells and more red cells, indicating weaker performance. Despite this, the path-average Ridge regression models again emerge as the best-performing models, consistently excelling in forecasting HICP inflation, core inflation, and services inflation. With the exception of one Ridge regression specification (`M.RR.F.CV`) for headline HICP inflation, no model outperforms the benchmark model across all forecast horizons. The forecasting performance of the factor models and tree-based models deteriorates significantly compared to their performance in the full sample. Nonetheless, the relatively high forecast accuracy of the path-average Ridge regression models remains evident in the pre-pandemic sample.

5.3 A closer look at model specification

The previous two subsections examined the forecasting performance of different model *types*. This section shifts the focus to differences in model *specification*. As outlined in Section 4, models can differ along several dimensions. We examine three specification choices: (1) the transformation of the target variable—comparing path-average forecasts with direct forecasts, (2) the inclusion versus exclusion of statistical factors in the predictor set; and (3) the use of the full predictor set versus a subset of targeted predictors. While additional modeling choices could be explored, many are specific to individual model types. For example, some models require hypermeter tuning based on criteria such as the `BIC` (Bayesian Information Criterion) or `CV` (Cross Validation), which are beyond the scope of the current analysis.

To assess the impact of these specification choices, we estimate regressions inspired by Goulet Coulombe et al. (2021) and Goulet Coulombe et al. (2022); see Equation (7) in Section 3. Figure 6 presents the results of the comparison between path-average and direct forecasts for the full sample. Blue bars indicate the percentage of models for which the $OOS.R^2$ of path-average forecasts (`M`) is *not* significantly different from that of the direct forecasts (`Y`). Red bars represent the share of models where path-average forecasts significantly outperform direct forecasts at the 10% level. Green bars indicate cases where direct forecasts significantly outperform path-average forecasts, also at the 10% level. Figure 7 evaluates whether including statistical factors in the predictor set improves forecast accuracy. Finally, Figure 8 investigates whether using targeted predictors improves forecast accuracy or not[4].

The main message from Figure 6 is that using path-average forecasts for forecasting headline HICP inflation is a 'no-regret' policy. The combined height of the blue (no significant difference) and red bars (path-average forecast superior) sum to 100%, indicating that direct forecasts never outperform path-average forecasts. This pattern also holds for nearly all forecast horizons when forecasting NEIG inflation. The results are more mixed for forecasting core inflation and services inflation. Path-average forecasts are advantageous in capturing short-term volatility, which contributes to more stable medium-term forecasts. On the other hand, direct forecasts may avoid the accumulation of errors inherent in the iterative nature of path-average forecasts. The relatively low volatility of core and services inflation reduces the need to rely on path-average forecast as a

[4] Factor models are excluded from Figure 7. Shrinkage models, which always use the full predictor set, are excluded from Figure 8.

form of 'insurance' against short-term fluctuations. For the full sample, there is strong empirical support for both including statistical factors (Figure 7) and using targeted predictors (Figure 8). Across nearly all forecast horizons and inflation measures, the inclusion of factors and the use of targeted predictors consistently enhance or do not impair forecast accuracy.

- insert Figure 6 - 8 about here -

To gain more insight into the quantitative differences between direct and path-average forecasts, Figure 9 shows the gain in $OOS.R^2$ from path-average forecasts relative to direct forecasts, disaggregated by model type. For headline HICP inflation and NEIG inflation, the gains are more pronounced at longer horizons, especially for shrinkage models. Ensemble models tend to perform worse when using path-average forecasts. However, this result should be interpreted with caution, as ensemble models rarely ranked among the best-performing models in the previous analysis. Depending on the forecast horizon, factor models occasionally perform better when using the direct forecasting approach.

Improvements in $OOS.R^2$ resulting from the inclusion of factors and targeted predictors are relatively limited, with factor inclusion showing particularly modest effects.

Qualitatively, the results for the pre-pandemic sample (available in the online appendix) closely resemble those observed in the full sample. For example, path-average forecasts generally yield higher $OOS.R^2$ values. A notable exception arises with factor models, where direct forecasts outperform path-average forecasts across all inflation measures except headline HICP inflation. This pattern extends to ensemble models, which consistently demonstrate weak performance across specifications.

The main takeaway from the $OOS.R^2$ analysis is that, overall, path-average forecasts outperform other methods for inflation forecasting. An exception arises with factor models, where the results are more mixed—particularly in the pre-pandemic sample. While incorporating factors and using targeted predictors yields only modest improvements, the benefits of path averaging are considerably greater. Our findings align with Goulet Coulombe et al. (2021), who attribute the strong performance of path averaging to its ability to leverage sparsity by aggregating separate horizon forecasts, simplifying learning for models like Random Forest and LASSO. Direct forecasts are denser and harder to estimate. Consistent with Beck and Wolf (2025), we also find that

aggregating short-horizon forecasts into longer-horizon projections outperforms direct forecasting, reinforcing path forecasting as a robust strategy in predictive modeling.

- insert Figure 9, 10, 11 about here -

5.4 ML models versus institutional forecasts

This section investigates whether ML models can not only outperform simple benchmark models in forecasting Dutch inflation, but also exceed the accuracy of an established institutional forecast—namely DNB's official inflation forecast.

Since the euro's introduction in 1999, DNB has been part of the Eurosystem. The Eurosystem publishes macroeconomic projections quarterly (see here). Darracq Pariès et al. (2021) assess the current modeling framework used within the Eurosystem, while Conrad and Enders (2024) evaluate the accuracy of its inflation projections. DNB contributes to these quarterly exercises by producing forecasts of Dutch annual inflation, covering horizons from 1 to 10 month ahead. These forecasts, referred to as the Narrow Inflation Projection Exercise (NIPE), will be denoted as NIPE throughout the remainder of the paper.

DNB's NIPE forecasts are generated using a suite-of-models approach. This suite includes linear models for key HICP inflation sub-components—such as NEIG, food and services inflation—building on Den Reijer and Vlaar (2006). Additionally, SARIMA models are employed to forecast over 200 individual COICOP price sub-components of the HICP. Final forecasts are derived through informal model averaging. The forecasts incorporate announced government measures, including changes to VAT rates and energy taxes. The model-based forecasts are sometimes adjusted using expert-judgment to account for relevant off-model information.

DNB produces inflation forecasts in February, May, August, and November. This differs from the ML forecasts analyzed earlier, which are produced every month. To ensure a fair comparison, we restrict ML forecasts to the same four months in which NIPE forecasts are available. For example, 1-month ahead forecasts correspond to March, June, September, and December, 2-month ahead forecasts to April, July, October, and January; and so forth.

Figures 12 and 13 illustrate the forecasting performance of all 62 models compared to the NIPE forecasts, offering deeper insights into which models within each model type perform best. The figures display heatmaps of the forecasting performance. The heatmap colors range from

a gain in forecast accuracy against the benchmark model ($RMSFE_{ML}$-$RMSFE_{NIPE}$) of 5% (light blue) to 50% (dark blue). Relative gains are displayed only when the improvement over NIPE is at least 5% and statistically significant at the 10% level, based on the Diebold and Mariano test. Figure 12 depicts forecasting performance over the full sample. The key insights are as follows:

First, none of the models is capable of consistently outperforming the NIPE headline HICP inflation forecast. This is not surprising given the weak performance of ML models in forecasting headline HICP compared to the naive benchmark model in Figure 4. Surprisingly, however, NIPE's headline HICP forecasts are often significantly more accurate those of the ML models, particularly at very short horizons (1 and 2 months), suggesting that relying solely on ML models may reduce forecast accuracy.

Second, the weak forecasting performance of the ML models becomes even more evident when comparing the NIPE forecasts for core inflation and services inflation. Once again, none of the ML models is able to outperform the NIPE forecast in a meaningful and statistically significant way. Given the previously relatively good performance of the ML models against the benchmark model, these outcomes are noteworthy (see Figure 4), as the usefulness of ML models was most widespread for these inflation measures.

Third, the outcomes for the NEIG inflation forecast present a somewhat different picture compared to the other inflation measures. Although almost none of the models can outperform the NIPE model, the differences are much smaller for most models, as indicated by the white cells. Only two models can outperform the NIPE forecast on one or more horizons: `M.RF.F` and `M.BTREE.F`.

A possible explanation for the relatively weak performance of the ML models in the full sample is that inflation during the observed period was strongly influenced by announced tax changes and levies. These announcements could be considered by researchers when making the NIPE forecast, but are not incorporated into the mechanical forecasts of the machine-learning models.

Turning to the pre-pandemic sample, the main conclusions from Figure 13 are:

- insert Figure 12 , 13 about here -

First, the forecasting performance for headline HICP inflation during the pre-pandemic sample echoes the lackluster forecasting performance of the ML models over the full sample. Using a ML model instead of the NIPE forecast hurts forecasting performance for almost all horizons and

models. The same holds for the forecasts of ML models for core inflation and services inflation, although to a somewhat lesser extent. The exception is `Y.BFACT.T`, which outperforms the NIPE core inflation forecast over an 8-month horizon.

Second, some ML models can outperform the NIPE forecast for NEIG inflation at specific forecast horizons. The relatively strong performance of several Ridge regression specifications is noteworthy. Three of the path-average Ridge regressions (`M.RR.NF.BIC`, `M.RR.F.CV`, `M.RR.F.BIC`) are either better or equally good according to our measures of economic and statistical significance across all horizons[5]. Comparing the forecasting performance of the NIPE and the naive benchmark models (`M.RWD`) suggests that the NIPE model has relatively limited forecasting power for NEIG inflation in the pre-pandemic period. The simple benchmark model has equal forecast accuracy to the NIPE forecast for NEIG inflation across all horizons and even outperforms the NIPE forecast over a 5-month horizon.

To better understand the performance of the Ridge regression models for NEIG inflation over time, Figure 14 illustrates the evolution of realized NEIG inflation (left-hand axis) and the evolution of the CSSFED of `M.RR.F.BIC` compared to the NIPE model for forecast horizons of 1 to 10 months ahead (right-hand axis).

- insert Figure 14 about here -

Between 2010 and 2019, `M.RR.F.BIC` shows some marginal forecasting gains over the NIPE forecasts for most forecast horizons. From 2020 onward, as NEIG inflation edged up slightly, there was an initial slight deterioration in the CSSFED. However, as NEIG inflation surged later on, `M.RR.F.BIC` clearly outperformed the NIPE forecasts, especially at longer forecast horizons.

Figure 15 highlights the predictors that most significantly contributed to the enhanced performance of `M.RR.F.BIC` in forecasting NEIG inflation during the pandemic and post-pandemic periods. The figure displays the contributions of various predictor sets. The contribution of each individual predictor is calculated as the product of its regression coefficient and its value (in deviation from the mean). These contributions are closely related to the concept of Shapley values, which are commonly used to quantify variable importance in ML models. In linear regression

[5] We tested for average Superior Predictive Ability, as described by Quaedvlieg (2021). The Ridge regression models have - on average - statistically significant lower mean Squared Forecast Error compared to the NIPE over short horizons (1- to 5-months), long horizons (6- to 10-months) and all horizons (1- to 10-months).

models, such as the Ridge regression model, predictor contributions are equivalent to Shapley values if the predictors are orthogonal—which is not the case in our dataset.

During the pandemic, real activity predictors, expectations, and financial predictors were the primary contributors to the `M.RR.F.BIC` forecasts. From the end of 2021 onward, producer prices (PPI) and domestic price pressures, including headline HICP inflation and its main sub-components (excluding NEIG inflation), became the dominant driving forces. This shift reflects the pass-through of energy prices, ongoing supply chain disruptions, and some overheating of the Dutch economy.

- insert Figure 15 about here -

6 Concluding remarks

This paper investigates whether machine learning models can produce accurate short- and medium-term inflation forecasts for the Netherlands.

In the first 'horse race', we evaluate the forecasting performance of ML models against a simple benchmark model. The relative improvements in RMSFE range from negligible at short horizons to over 40% at longer horizons, depending on the inflation measure. ML models exhibit limited effectiveness in forecasting headline HICP inflation. Across the full sample period, none of the ML models consistently outperform the benchmark model. However, when excluding the pandemic and post-pandemic periods, ML model performance improves notably. In contrast, for core inflation, NEIG inflation, and services inflation, the performance of ML models improve notably when pandemic and post-pandemic data are included.

In the second evaluation exercise, we assessed the performance of ML models against DNB's official NIPE forecast. We found that the performance of the ML models was weak in both the full and pre-pandemic samples. The main exception is NEIG inflation, where some of the Ridge regression models are able to outperform the NIPE forecast.

Combining the outcomes of both forecasting 'horse races', the preferred ML model is the path-average Ridge regression. The `M.RR.F.BIC` model either outperformed or matched the forecast accuracy of the benchmark and NIPE models. In contrast to other research, we find that the non-linear random forest is not very helpful in forecasting inflation, whether in tranquil or volatile times. The forecast accuracy of the best linear models was comparable to or significantly better

than that of non-linear models. Furthermore, we show that using path-average forecasts is superior to using direct forecasts, while the use of statistical factors or targeting predictors adds little value.

Future research could investigate whether these outcomes apply to other countries and time periods as well. Another promising avenue for future research would be to incorporate textual or other big-data sources into the dataset to assess whether these data sources can enhance forecasting performance.

References

Araujo, G. S. and W. P. Gaglianone (2023). Machine learning methods for inflation forecasting in Brazil: new contenders versus classical models. *Latin American Journal of Central Banking 4*(2), 1–29. link.

Atkeson, A. and L. E. Ohanian (2001). Are Phillips curves useful for forecasting inflation? Quarterly Review 2511, Federal Reserve Bank of Minneapolis. link.

Bai, J. and S. Ng (2002). Determining the number of factors in approximate factor models. *Econometrica 70*(1), 191–221. link.

Bai, J. and S. Ng (2008). Forecasting economic time series using targeted predictors. *Journal of Econometrics 146*(2), 304–317. link.

Bai, J. and S. Ng (2009). Boosting diffusion indices. *Journal of Applied Econometrics 24*, 607–629. link.

Beck, E. and M. Wolf (2025). Forecasting inflation with the hedged random forest. Working Paper 2025-07, Swiss National Bank. link.

Bernanke, B. S. and J. Boivin (2003). Monetary policy in a data-rich environment. *Journal of Monetary Economics 50*(3), 525–546. link.

Boot, T. and D. Nibbering (2019). Forecasting using random subspace methods. *Journal of Econometrics 209*(2), 391–406. link.

Borup, D., B. J. Christensen, N. S. Mühlbach, and M. Slot Nielsen (2023). Targeting predictors in random forest regression. *International Journal of Forecasting 39*(2), 841–868. link.

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

Chipman, H. A., E. I. George, and R. E. McCulloch (2010). BART: Bayesian additive regression trees. *Annals of Applied Statistics 4*(1), 266–298. link.

Conrad, C. and Z. Enders (2024). The limits to the ECB's inflation projections. Policy Brief 945, SUERF. link.

Darracq Pariès, M., A. Notarpietro, J. Kilponen, N. Papadopoulou, S. Zimic, P. Aldama, G. Langenus, L. J. Alvarez, M. Lemoine, and E. Angelini (2021). Review of macroeconomic modelling in the eurosystem: current practices and scope for improvement. Occasional Paper Series 267, European Central Bank. link.

Das, P. K. and P. K. Das (2024). Forecasting and analyzing predictors of inflation rate: Using machine learning approach. *Journal of Quantitative Economics 22*, 493–517. link.

Den Reijer, A. and P. Vlaar (2006). Forecasting inflation: An art as well as a science! *De Economist 154*, 19–40. link.

Diebold, F. X. and R. S. Mariano (1995). Comparing predictive accuracy. *Journal of Business & Economic Statistics 13*(3), 134–144. link.

Elliott, G., A. Gargano, and A. Timmermann (2013). Complete subset regressions. *Journal of Econometrics 177*(2), 357–373. link.

Faust, J. and J. H. Wright (2013). Forecasting inflation. *Handbook of Forecasting 2*(A), 2–56. link.

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

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

Hansen, P. R., A. Lunde, and J. M. Nason (2011). The model confidence set. *Econometrica 69*(2), 453–497. link.

Hoerl, A. E. and R. W. Kennard (1970). Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics 12*(1), 55–67. here.

Huang, N., Y. Qi, and J. Xia (2025). China's inflation forecasting in a data-rich environment: based on machine learning algorithms. *Applied Economics 57*(17), 1995–2020. link.

Kock, A. B. and T. Teräsvirta (2016). Forecasting macroeconomic variables using neural network models and three automated model selection techniques. *Econometric Reviews 35*(8), 1753–1779. link.

Kohlscheen, E. (2022). What does machine learning say about the drivers of inflation? Working Papers 980, BIS. link.

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

Lenza, M., I. Moutachaker, and J. Paredes (2025). Density forecasts of inflation: a quantile regression forest approach. *European Economic Review 178*, 105079. link.

Maehashi, K. and M. Shintani (2020). Macroeconomic forecasting using factor models and machine learning: an application to Japan. *Journal of The Japanese and International Economies 58*, 101104. link.

Marcellino, M., J. H. Stock, and M. W. Watson (2006). A comparison of direct and iterated multistep AR methods for forecasting macroeconomic time series. *Journal of Econometrics 135*(1), 499–526. link.

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

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

Naghi, A. A., E. O. E., and M. D. Zaharieva (2024). The benefits of forecasting inflation with machine learning: New evidence. *Journal of Applied Econometrics forthcoming*, 1321—1331. link.

Prüser, J. (2019). Forecasting with many predictors using Bayesian additive regression trees. *Journal of Forecasting 38*(1), 40–53. link.

Quaedvlieg, R. (2021). Multi-horizon forecast comparison. *Journal of Business & Economic Statistics 39*, 621–631. link.

Stock, J. H. and M. W. Watson (1999). Forecasting inflation. *Journal of Monetary Economics 44*, 293–335. link.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B 58*, 267–288. link.

Vedder, C. and E. van de Winkel (2024). Is machine learning beneficial for macroeconomic forecasting with limited observations? Discussion paper, CPB. link.

Yoon, J. (2021). Forecasting of real GDP growth using machine learning models: Gradient boosting and random forest approach. *Computational Economics 57*, 247–265. link.

Zou, H. (2006). The adaptive lasso and its oracle properties. *Journal of the American Statistical Association 101*(476), 1418–1429. link.

A Data set

Table A.1 provides the 129 monthly series that have been used for the estimation of the models in the main text. As mentioned in the main text, the data set can be split-up into eleven groups: 1) output & income: 27 series, 2) labor market: 10 series, 3) consumption: 14 series, 4) orders & inventories: 6 series, 5) money & credit: 9 series, 6) interest & exchange rates: 14 series, 7) commodity prices: 7 series, 8) producer prices: 20 series, 9) domestic prices: 13 series, 10) price expectations: 3 series, and 11) stock market: 6 series. A complete list of the series is provided in Table A.1.

We constructed two databases: one for computing path-average forecasts and another for calculating direct forecasts. First, we collected all monthly series in the economics and finance sections from data warehouses (see column 'Source') that started in January 1990 or earlier, and hand-picked series within each of the eleven groups we defined. Next, for the path-average database, we performed seasonal adjustment on all non-financial series that were not seasonally adjusted at the source using the US Census X13 ARIMA-SEATS method. The financial series, i.e. all series in the groups stock market, money and credit, interest and exchange rates are not seasonally adjusted. Then, we transformed all variables to stationarity. For the path-average forecast database we mostly applied the first difference or log first difference operator to the series, whereas for the direct forecast database we employed annual differences or annual log differences[6]. Details on the transformations applied in both databases are given in the 'Transf.' column in Table A.1. We use the monthly current and one year ahead consumer price inflation forecast for the Netherlands from Consensus Forecasts to calculate a monthly series of consumer price inflation expectations 12 months ahead.

- insert Table A.1 about here -

Table A.1: Description monthly database

Nr.	Description	Source	Transf.	Last
1. Output & Income				
1.	Manufacturing (total)	EUR	2,5	Dec-23
2.	Manufacturing of food products, beverages & tobacco	EUR	2,5	Dec-23
3.	Manufacturing of textiles	EUR	2,5	Dec-23
4.	Manufacturing of wearing apparel	EUR	2,5	Dec-23
5.	Manufacturing of leather & related products	EUR	2,5	Dec-23
6.	Manufacturing of wood & of products of wood & cork	EUR	2,5	Dec-23
7.	Manufacturing of paper & paper products	EUR	2,5	Dec-23
8.	Printing & reproduction of recorded media	EUR	2,5	Dec-23
9.	Manufacturing of coke & refined petroleum products	EUR	2,5	Dec-23
10.	Manufacturing of pharmaceutical products	EUR	2,5	Dec-23
11.	Manufacturing of rubber & plastic products	EUR	2,5	Dec-23
12.	Manufacturing of other non-metallic mineral products	EUR	2,5	Dec-23
13.	Manufacturing of basic metals	EUR	2,5	Dec-23
14.	Manufacturing of fabricated metal products	EUR	2,5	Dec-23
15.	Manufacturing of computer, electronic & optical products	EUR	2,5	Dec-23
16.	Manufacturing of machinery & equipment	EUR	2,5	Dec-23

Continued on next page ...

[6] There are two exceptions to this rule: the economic sentiment indicator and business climate indicator are included in levels as proxies for the output gap.

Table A.1 —Continued from previous page

Nr.	Description	Source	Transf.	Last
17.	Manufacturing of motor vehicles & trailers	EUR	2,5	Dec-23
18.	Manufacturing of furniture	EUR	2,5	Dec-23
19.	Other manufacturing	EUR	2,5	Dec-23
20.	Supply of natural gas	CBS	2,5	Jan-24
21.	Use of natural gas	CBS	2,5	Jan-24
22.	Hotels & similar accommodation, arrivals, foreigners	EUR	2,5	Dec-23
23.	Hotels & similar accommodation, arrivals, total	EUR	2,5	Dec-23
24.	Hotels & similar accommodation, nights spend, foreigners	EUR	2,5	Dec-23
25.	Hotels & similar accommodation, nights spend, total	EUR	2,5	Dec-23
26.	Economic sentiment indicator	EUR	0,0	Feb-24
27.	Business climate indicator	EUR	0,0	Feb-24
2. Labor market				
28.	Unemployment rate, total	ECB	1,4	Jan-24
29.	Unemployment rate, female	ECB	1,4	Jan-24
30.	Unemployment rate, total, < 25 years	ECB	1,4	Jan-24
31.	Unemployment rate, female, < 25 years	ECB	1,4	Jan-24
32.	Hourly wages	CBS	2,5	Oct-23
33.	Consumer confidence, unemployment > 12 months	EUR	1,4	Feb-24
34.	Construction confidence, employment > next 3 months	EUR	1,4	Feb-24
35.	Construction confidence, limiting factors, shortage of labour	EUR	1,4	Feb-24
36.	Industrial confidence, employment > 3 months	EUR	1,4	Feb-24
37.	Retail confidence, employment expectations > 3 months	EUR	1,4	Feb-24
3. Consumption				
38.	Consumer confidence, headline	EUR	1,4	Feb-24
39.	Consumer confidence, financial situation < 12 months	EUR	1,4	Feb-24
40.	Consumer confidence, financial situation > 12 months	EUR	1,4	Feb-24
41.	Consumer confidence, general economic situation < 12 months	EUR	1,4	Feb-24
42.	Consumer confidence, general economic situation > 12 months	EUR	1,4	Feb-24
43.	Consumer confidence, major purchase > 12 months	EUR	1,4	Feb-24
44.	Consumer confidence, major purchases, current	EUR	1,4	Feb-24
45.	Consumer confidence, statement on financial situation of household	EUR	1,4	Feb-24
46.	Consumer confidence, savings > 12 months	EUR	1,4	Feb-24
47.	Consumer confidence, current ec. situation is adequate for savings	EUR	1,4	Feb-24
48.	Retail confidence, headline	EUR	1,4	Feb-24
49.	Retail confidence, volume of stocks currently held	EUR	1,4	Feb-24
50.	New passenger car	ECB	2,5	Jan-24
51.	Retail trade	ECB	2,5	Jan-24
4. Orders & Inventories				
52.	Industrial confidence, assessment of the current level of stocks	EUR	1,4	Feb-24
53.	Construction confidence, limiting factors, none	EUR	1,4	Feb-24
54.	Construction confidence, limiting factors, insufficient demand	EUR	1,4	Feb-24
55.	Construction confidence, limiting factors, weather conditions	EUR	1,4	Feb-24
56.	Construction confidence, limiting factors, material/equipment	EUR	1,4	Feb-24
57.	Construction confidence, limiting factors, other	EUR	1,4	Feb-24
5. Money & Credit				
58.	Loans, excl. government (EUR)	ECB	3,6	Jan-24
59.	External assets (EUR)	ECB	3,6	Jan-24
60.	External liabilities (EUR)	ECB	3,6	Jan-24
61.	Overnight deposits (EUR)	ECB	3,6	Jan-24
62.	Deposits <2 years, redeemable at notice < 3 months (EUR)	ECB	3,6	Jan-24
63.	Repo's, debt securities, shares < 2 years (EUR)	ECB	3,6	Jan-24
64.	M1 (EUR)	ECB	3,6	Jan-24
65.	M2 (EUR)	ECB	3,6	Jan-24

Continued on next page . . .

Table A.1 —Continued from previous page

Nr.	Description	Source	Transf.	Last
66.	M3 (EUR)	ECB	3,6	Jan-24
6. Interest & Exchange Rates				
67.	Real effective exchange rate, deflator consumer price index	ECB	2,5	Feb-24
68.	Real effective exchange rate, deflator producer price index	ECB	2,5	Feb-24
69.	Nominal effective exchange rate euro	ECB	2,5	Feb-24
70.	10-year government bond interest rate (%)	ECB	1,4	Feb-24
71.	3-month interbank interest rate (%)	ECB	1,4	Feb-24
72.	3-month deposits interest rate (%)	ECB	1,4	Feb-24
73.	Loans non-financial corporations, new, total (%)	ECB	1,4	Jan-24
74.	Loans non-financial corporations, new, $\leq$ EUR 1 million (%)	ECB	1,4	Jan-24
75.	Loans non-financial corporations, new, > EUR 1 million (%)	ECB	1,4	Jan-24
76.	Loans consumption, new, total (%)	ECB	1,4	Jan-24
77.	Loans house purchases, new, total (%)	ECB	1,4	Jan-24
78.	UK pound sterling/EUR exchange rate (%)	EUR	2,5	Feb-24
79.	Japenese yen/EUR exchange rate	EUR	2,5	Feb-24
80.	US dollar/EUR exchange rate	EUR	2,5	Feb-24
7. Commodity Prices				
81.	Harmonized index of consumer prices, energy	ECB	2,7	Jan-24
82.	Europe brent spot price, USD (barrel)	ECB	2,5	Feb-24
83.	Food commodities	ECB	2,5	Feb-24
84.	Non-energy commodities	ECB	2,5	Feb-24
85.	Terms of trade	CBS	2,5	Dec-23
86.	Import prices	CBS	2,5	Dec-23
87.	Export prices	CBS	2,5	Dec-23
8. Producer Prices				
88.	Producer price index, manufacturing	EUR	2,5	Dec-23
89.	Producer price index, mining & quarrying	EUR	2,5	Dec-23
90.	Producer price index, food products, beverages & tobacco products	EUR	2,5	Dec-23
91.	Producer price index, textiles	EUR	2,5	Dec-23
92.	Producer price index, wearing apparel	EUR	2,5	Dec-23
93.	Producer price index, leather & related products	EUR	2,5	Dec-23
94.	Producer price index, wood & of products of wood & cork	EUR	2,5	Dec-23
95.	Producer price index, paper & paper products	EUR	2,5	Dec-23
96.	Producer price index, printing & reproduction of recorded media	EUR	2,5	Dec-23
97.	Producer price index, coke & refined petroleum products	EUR	2,5	Dec-23
98.	Producer price index, chemicals & chemical products	EUR	2,5	Dec-23
99.	Producer price index, rubber & plastic products	EUR	2,5	Dec-23
100.	Producer price index, other non-metallic mineral products	EUR	2,5	Dec-23
101.	Producer price index, basic metals	EUR	2,5	Dec-23
102.	Producer price index, fabricated metal products	EUR	2,5	Dec-23
103.	Producer price index, computer, electronic & optical products	EUR	2,5	Dec-23
104.	Producer price index, machinery & equipment	EUR	2,5	Dec-23
105.	Producer price index, motor vehicles & trailers	EUR	2,5	Dec-23
106.	Producer price index, electricity, gas, steam & air conditioning supply	EUR	2,5	Dec-23
107.	Producer price index, water collection, treatment & supply	EUR	2,5	Dec-23
9. Domestic Prices				
108.	Harmonized index of consumer prices, headline	ECB	2,5	Jan-24
109.	Harmonized index of consumer prices, headline excl. energy & food	ECB	2,5	Jan-24
110.	Harmonized index of consumer prices, headline excl. energy	ECB	2,5	Jan-24
111.	Harmonized index of consumer prices, unprocessed food	ECB	2,5	Jan-24
112.	Harmonized index of consumer prices, food	ECB	2,5	Jan-24
113.	Harmonized index of consumer prices, services	ECB	2,5	Jan-24
114.	Harmonized index of consumer prices, industrial goods excl. energy	ECB	2,5	Jan-24

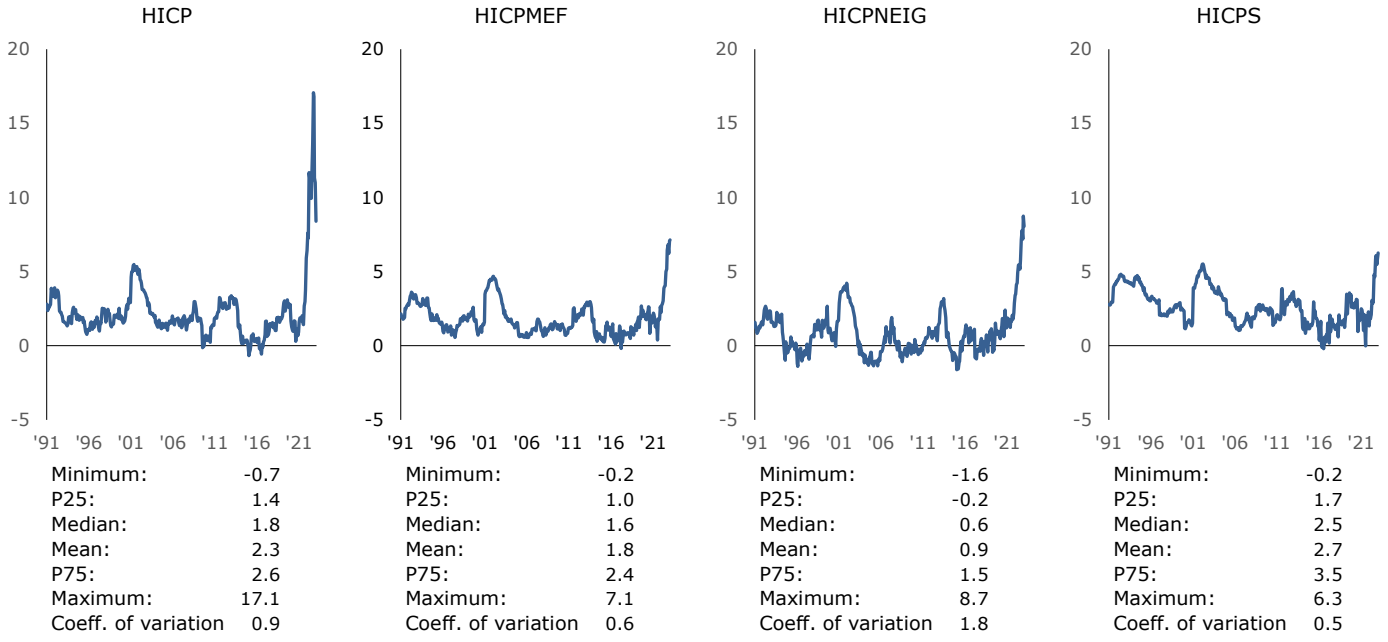
Continued on next page . . .

Table A.1 —Continued from previous page

Nr.	Description	Source	Transf.	Last
115.	Harmonized index of consumer prices, processed food	ECB	2,5	Jan-24
116.	Consumer price index, all items	ECB	2,5	Jan-24
117.	Construction costs, material prices, residential	CBS	2,5	Jan-24
118.	Housing prices	CBS	2,5	Dec-23
119.	Price materials new construction home	CBS	2,5	Jan-24
120.	Consumer confidence, price trend < 12 months	EUR	1,4	Feb-24
10. Price Expectations				
121.	Consumer confidence, price trend > 12 months	EUR	1,4	Feb-24
122.	Construction confidence, price expectations > 3 months	EUR	1,4	Feb-24
123.	Consensus forecast consumer price index > 12 months	CF	1,4	Feb-24
11. Stock Market				
124.	Amsterdam exchange index (AEX)	ECB	2,5	Feb-24
125.	Amsterdam midkap index	DS	2,5	Feb-24
126.	Dow Jones euro stoxx 50 index	ECB	2,5	Feb-24
127.	Financial stability index, Germany	ECB	1,4	Jan-24
128.	Financial stability index, United Kingdom	ECB	1,4	Jan-24
129.	Financial stability index, Netherlands	ECB	1,4	Jan-24

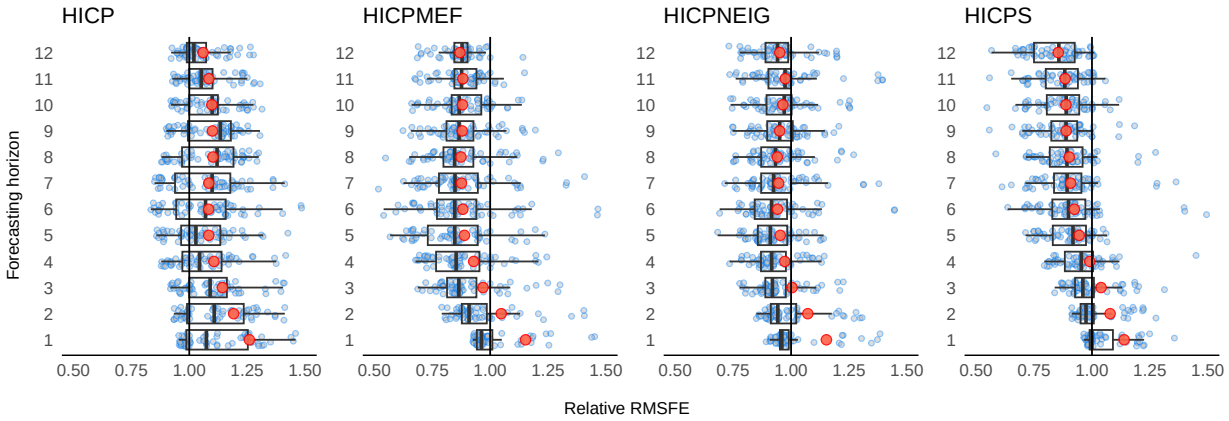
Notes: Nr.: Number of indicator; Description: Indicator description; Source: CBS: Statistics Netherlands, CF: Consensus Forecasts, DS: Refinitiv datastream, ECB: European Central Bank, EUR: Eurostat; Transf. = x,y: Transformation of variable in path-average (x) and direct (y) forecast database, respectively, 0 = level, 1 = first difference, 2 = log difference, 3 = difference of log difference, 4 = annual difference, 5 = annual log difference, 6 = difference of annual log difference; Last: Last monthly observation.

Figure 1: HICP inflation and its main sub-components†



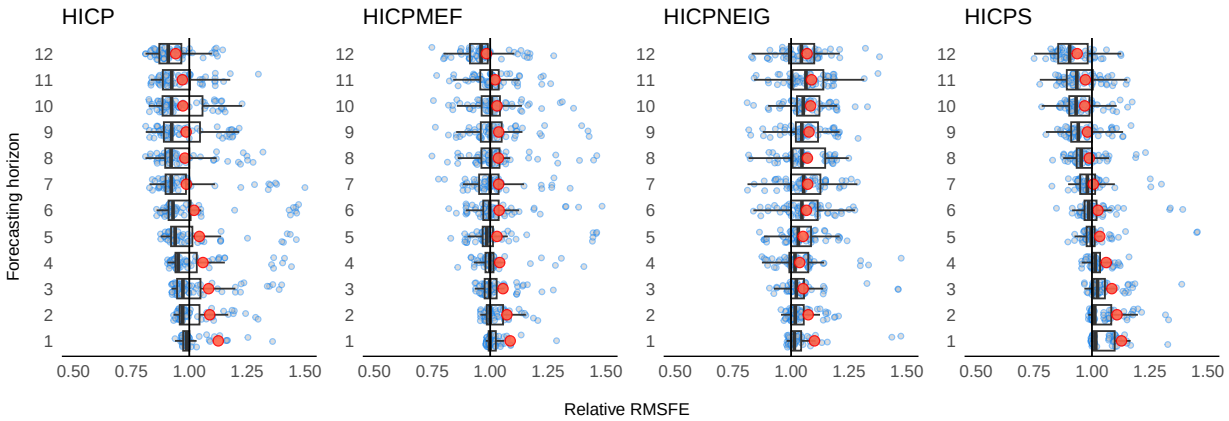
† HICP = headline HICP inflation, HICPMEF = HICP inflation excluding food & energy, HICPNEIG = non-energy industrial goods inflation, HICPS = services inflation.

Figure 2: Relative RMSFE, full sample (2010M1-2023M12)†



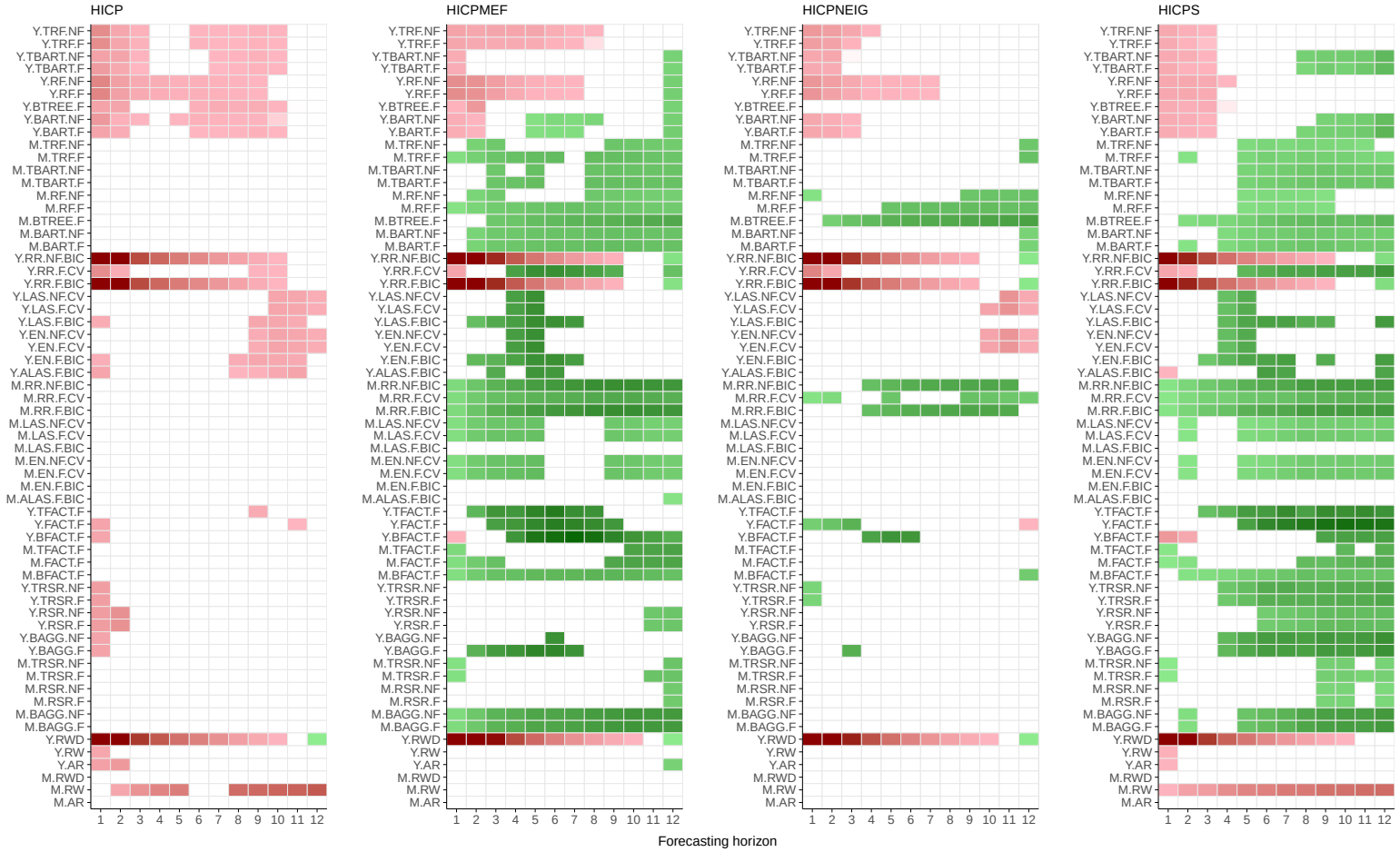
† (RMSFE indicator model)/(RMSFE M.RWD); HICP = headline HICP inflation, HICPMEF = HICP inflation excluding food & energy, HICPNEIG = non-energy industrial goods inflation, HICPS = services inflation.

Figure 3: Relative RMSFE, pre-pandemic sample (2010M1-2019M12)†



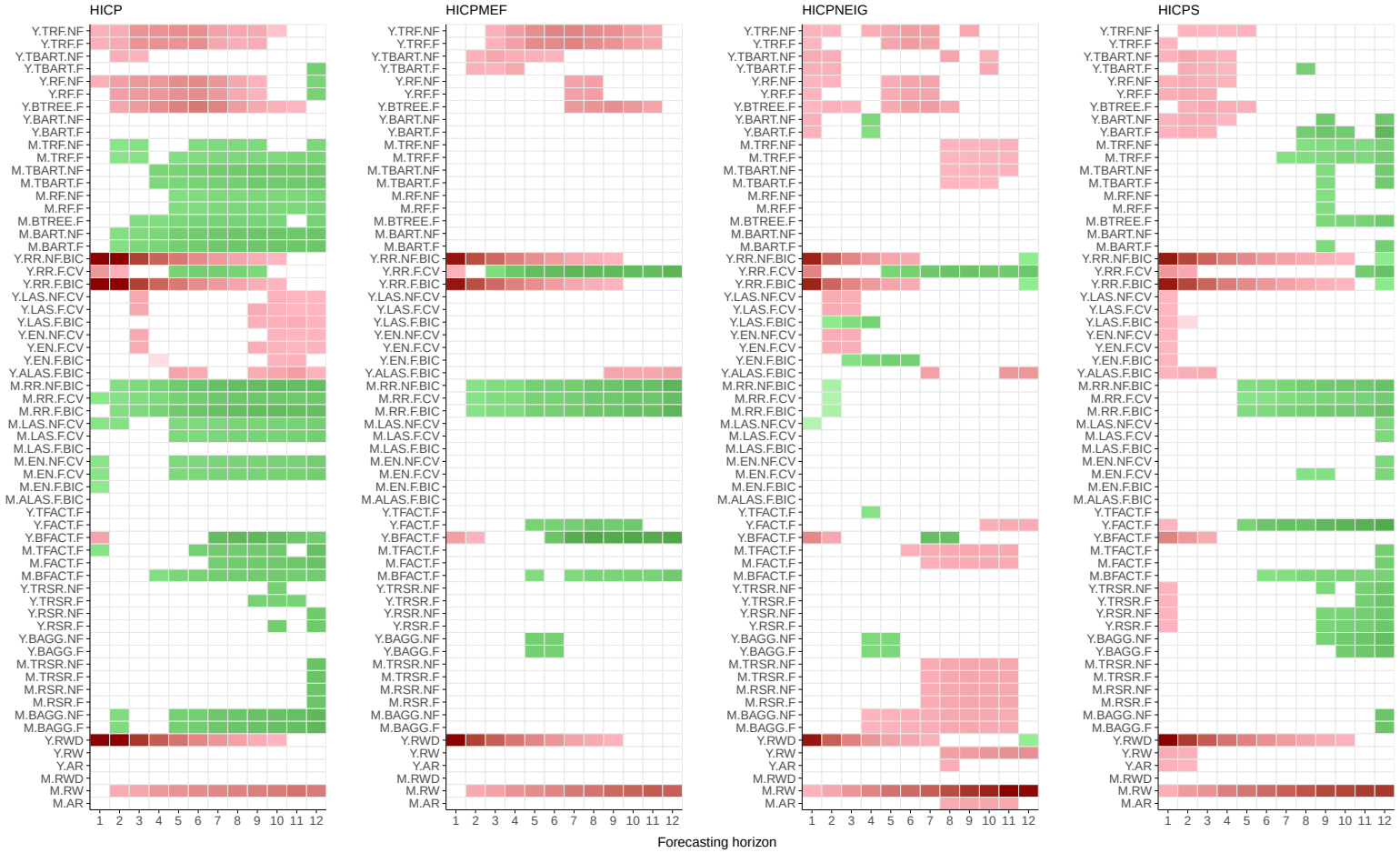
† (RMSFE indicator model)/(RMSFE M.RWD); HICP = headline HICP inflation, HICPMEF = HICP inflation excluding food & energy, HICPNEIG = non-energy industrial goods inflation, HICPS = services inflation.

Figure 4: Heatmap relative RMSFE, full sample (2010M1-2023M12) †



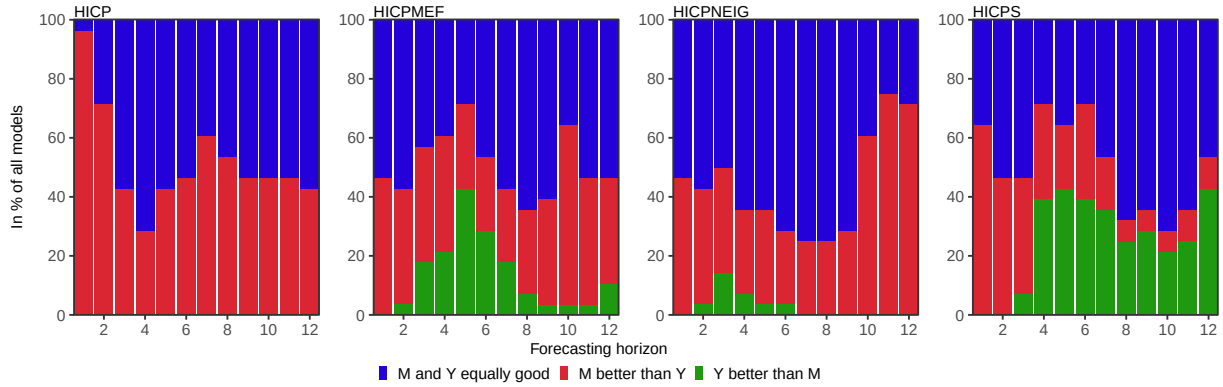
† White cells indicate the forecast accuracy of the machine-learning model is not statistically different from the benchmark model, i.e. the Diebold and Mariano-test is not statistically significant at the 10% level. A green cell indicates the relative gain in forecast accuracy of the ML model is 5% or larger, red cells indicate there is 5% or larger decline in forecast accuracy from using the ML model. The colors range from light green(red) to dark green(red), corresponding to a gain(loss) in forecast accuracy ranging from 5% tot 50% or more.

Figure 5: Heatmap relative RMSFE, pre-pandemic sample (2010M1-2019M12) †



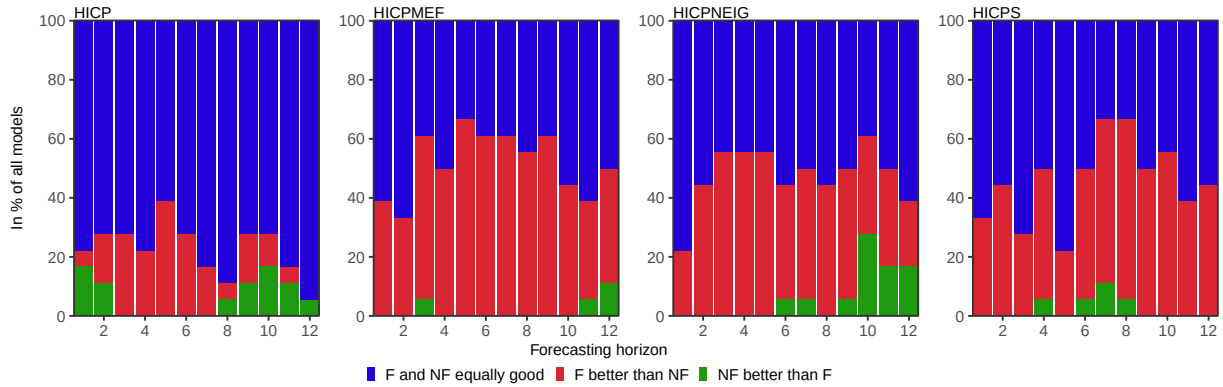
† White cells indicate the forecast accuracy of the machine-learning model is not statistically different from the benchmark model, i.e. the Diebold and Mariano-test is not statistically significant at the 10% level. A green cell indicates the relative gain in forecast accuracy of the ML model is 5% or larger, red cells indicate there is 5% or larger decline in forecast accuracy from using the ML model. The colors range from light green(red) to dark green(red), corresponding to a gain(loss) in forecast accuracy ranging from 5% tot 50% or more.

Figure 6: Test of direct versus path-average forecast, 10% significance level, full sample (2010M1-2023M12)†



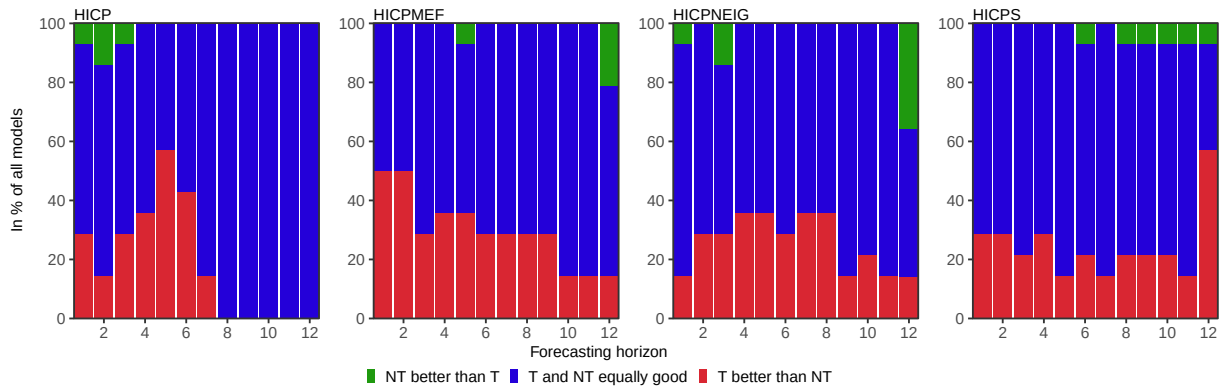
† M: path-average forecast, Y: direct forecast, HICP = headline HICP inflation, HICPMEF = HICP inflation excluding food & energy, HICPNEIG = non-energy industrial goods inflation, HICPS = services inflation.

Figure 7: Test of factor versus no factor forecast, 10% significance level, full sample (2010M1-2023M12)†



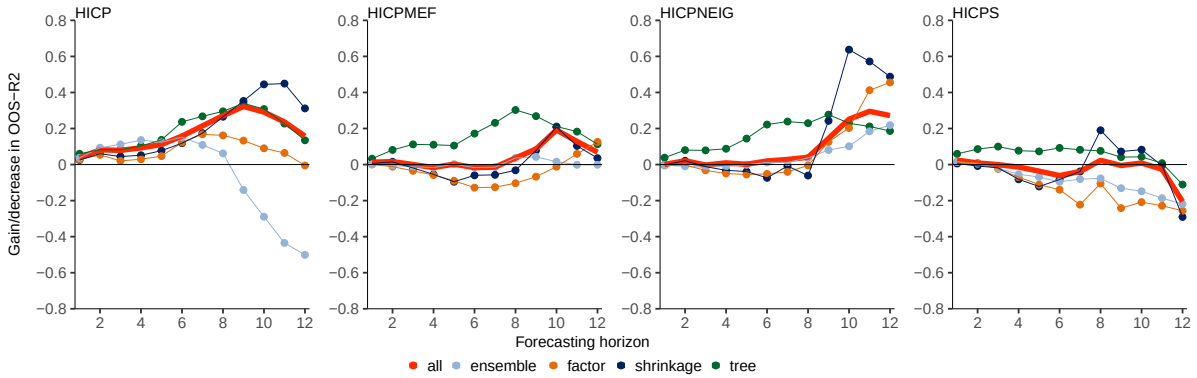
† F: factor augmented forecast (excluding factor models), NF: forecast without factors (excluding factor models), HICP = headline HICP inflation, HICPMEF = HICP inflation excluding food & energy, HICPNEIG = non-energy industrial goods inflation, HICPS = services inflation.

Figure 8: Test of targeting versus no targeting of predictors, 10% significance level, full sample (2010M1-2023M12)†



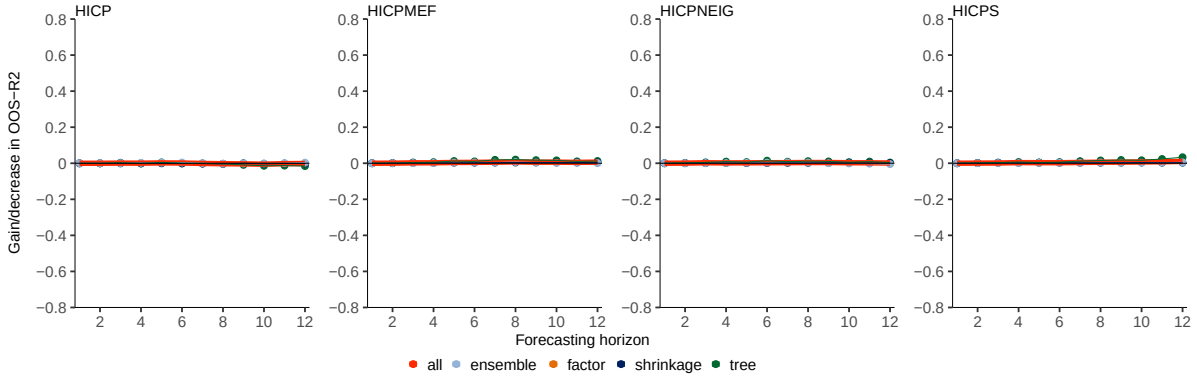
† T: targeted predictors (excluding shrinkage models), NT: predictors not targeted (excluding shrinkage models), HICP = headline HICP inflation, HICPMEF = HICP inflation excluding food & energy, HICPNEIG = non-energy industrial goods inflation, HICPS = services inflation.

Figure 9: Gain in $OOS.R^2$ of path-average forecast compared to direct forecast, full sample (2010M1-2023M12)†



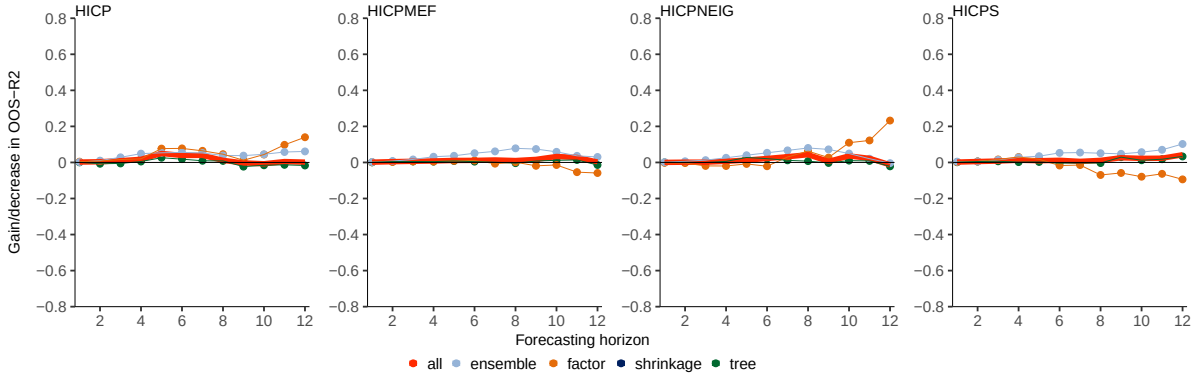
† Gain/decrease in OOS-R2 of using path-average forecasts, HICP = headline HICP inflation, HICPMEF = HICP inflation excluding food & energy, HICPNEIG = non-energy industrial goods inflation, HICPS = services inflation.

Figure 10: Gain in $OOS.R^2$ of factor versus no factor forecast, full sample (2010M1-2023M12)†



† Gain/decrease in OOS-R2 of using factors (excluding factor models), HICP = headline HICP inflation, HICPMEF = HICP inflation excluding food & energy, HICPNEIG = non-energy industrial goods inflation, HICPS = services inflation.

Figure 11: Gain in $OOS.R^2$ of targeting versus no targeting of indicators, full sample (2010M1-2023M12)†



† Gain/decrease using targeted factors or indicators (excluding shrinkage models), HICP = headline HICP inflation, HICPMEF = HICP inflation excluding food & energy, HICPNEIG = non-energy industrial goods inflation, HICPS = services inflation.

Figure 12: Heatmap relative RMSFE, full sample (2010M1-2023M1) †

† White cells indicate the forecast accuracy of the machine-learning model is not statistically different from the NIPE forecast, i.e. the Diebold and Mariano-test is not statistically significant at the 10% level. A green cell indicates the relative gain in forecast accuracy of the ML model is 5% or larger, red cells indicate there is 5% or larger decline in forecast accuracy from using the ML model. The colors range from light green(red) to dark green(red), corresponding to a gain(loss) in forecast accuracy ranging from 5% tot 50% or more.

Figure 13: Heatmap relative RMSFE, pre-pandemicsample (2010M1-2019M12) †

† White cells indicate the forecast accuracy of the machine-learning model is not statistically different from the NIPE forecast, i.e. the Diebold and Mariano-test is not statistically significant at the 10% level. A green cell indicates the relative gain in forecast accuracy of the ML model is 5% or larger, red cells indicate there is 5% or larger decline in forecast accuracy from using the ML model. The colors range from light green(red) to dark green(red), corresponding to a gain(loss) in forecast accuracy ranging from 5% tot 50% or more.

Figure 14: NEIG inflation: CSSFED `M.RR.F.BIC` versus NIPE benchmark

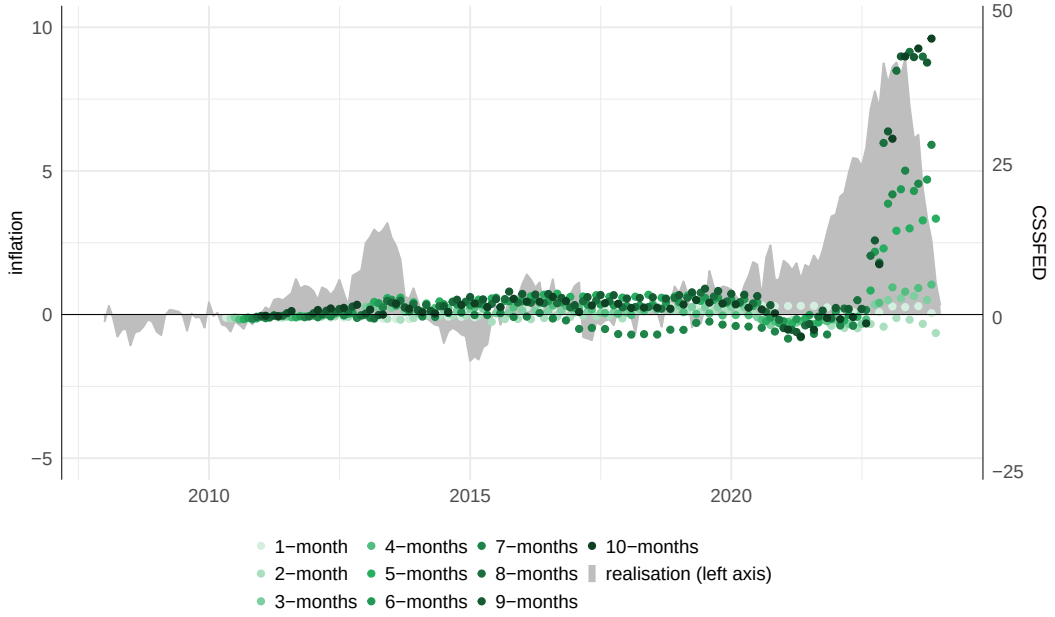


Figure 15: Decomposition of the `M.RR.F.BIC` forecast