DNB Working Paper
No 828/February 2025

Forecasting Dutch inflation using machine learning methods

Robert-Paul Berben, Rajni Rasiawan and Jasper de Winter

DeNederlandscheBank
EUROSYSTEEM

Forecasting Dutch inflation using machine learning methods

Robert-Paul Berben, Rajni Rasiawan and Jasper de Winter *

* Views expressed are those of the author and do not necessarily reflect official positions of De Nederlandsche Bank.

Working Paper No. 828

February 2025

De Nederlandsche Bank NV
P.O. Box 98
1000 AB AMSTERDAM
The Netherlands

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

February 4, 2025

Abstract

This paper examines the performance of machine learning models in forecasting Dutch inflation over the period 2010 to 2023, leveraging a large dataset and a range of machine learning techniques. The findings indicate that certain machine learning models outperform simple benchmarks, particularly in forecasting core inflation and services inflation. However, these models face challenges in consistently outperforming the primary inflation forecast of De Nederlandsche Bank for headline inflation, though they show promise in improving the forecast for non-energy industrial goods inflation. Models employing path averages rather than direct forecasting achieve greater accuracy, while the inclusion of non-linearities, factors, or targeted predictors provides minimal or no improvement in forecasting performance. Overall, Ridge regression has the best forecasting performance in our study.

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

* We gratefully acknowledge comments from Maurice Bun, Peter van Els, Kostas Mavromatis and participants of the ECB Workshop "Harnessing Artificial Intelligence for inflation assessment".

† Corresponding author: r.p.berben@dnb.nl.

1 Introduction

Inflation forecasts are important inputs for monetary policy decision making. Unfortunately, forecasting inflation turns out to be a rather challenging task, with simple time-series models often yielding 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) consider several ML models use the FRED-MD database McCracken and Ng (2016) to forecast U.S. consumer price index (CPI) inflation, and are able to outperform standard benchmarks, with the random forest (RF) model delivering the smallest errors. This paper examines the broader applicability of ML models for forecasting inflation by applying these methods to inflation in the Netherlands, one of the larger countries in the euro area.

We contribute to the existing literature by examining the usefulness of ML models for forecasting (sub-components of) HICP inflation in the Netherlands, both before, during and after the COVID-19 pandemic. The analysis benchmarks the performance of these models against both a simple time-series model and the inflation projections provided by De Nederlandsche Bank (DNB), the central bank of the Netherlands.

We build on Medeiros et al. (2021) by using a large dataset and evaluating multiple ML models to forecast the year-on-year inflation rate. Our dataset consists of 129 monthly Dutch and international series spanning from January 1990 to January 2024. The target variables in our analysis are the harmonized index of consumer price (HICP) and three of its sub-components: services inflation, non-energy industrial goods (NEIG) inflation, and core inflation (HICP excluding energy and food). To generate year-on-year inflation forecasts we usetwo different approaches. The first is the direct approach, which involves forecasting the year-on-year inflation rate directly using a single model. This is the more conventional method. The second approach, known as the path average method, introduced by Goulet Coulombe et al. (2021), involves forecasting month-on-month changes in the price index and then cumulating these forecasts to derive the year-on-year inflation rate.

Our main results can be summarized as follows. First, ML models can outperform simple benchmark models in forecasting Dutch inflation, particularly when forecasting core inflation and services inflation. Second, non-linear models, such as the random forest, although effective in some contexts, do not consistently outperform linear models. Third, ML models typically provide more accurate forecasts for longer horizons (6 to 12 months) compared to shorter periods. Moreover, path average forecasts often outperform direct forecasting methods for year-on-year inflation. Fourth, ML models face challenges to consistently outperform DNB's official inflation forecasts, particularly for headline HICP inflation. However, for NEIG inflation, some ML models outperform DNB's forecasts over certain horizons. Finally, combining the results of the forecasting 'horse races' between the machine learning models and the two benchmarks, the preferred machine learning model is the path average Ridge regression.

The rest of the paper is organized as follows. Section 2 reviews the literature. In Section 3 we describe the data, the forecast design and the forecast evaluation metrics. In Section 4 we introduce the ML models. The main results are in Section 5. Concluding remarks are in Section 6. Details on the dataset and additional charts are shown in Appendices A and B, respectively. A complete set of all results is in the online Appendix (here).

2 Literature

Following the seminal article by Medeiros et al. (2021), numerous studies have been published demonstrating the usefulness of ML models for forecasting inflation. Naghi et al. (2024) elaborate on the work by Medeiros et al. (2021) by extending the sample from 2016 to 2022, considerably expanding the set of machine learning models, and also applying the methods to Canadian and U.K. datasets. They find that the claimed superiority of the RF model for forecasting U.S. inflation carries over to U.K. inflation, but much less so to Canadian data. Furthermore, from 2020 onward the RF model performs poorly on U.S. data. Similarly, Huang et al. (2024) 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 benchmarks. Maehashi and Shintani (2020) analyze, among other things, Japanese CPI and wholesale price index (WPI) inflation. They show that beyond the very short-term horizon, the RF model and boosted trees outperform the benchmark AR model. Das and Das (2024) conclude that for Indian CPI the RF model provides much more accurate forecasts than an ARIMA model, but only when the COVID-19 pandemic period is included. Using pre-pandemic data the ARIMA model turns out to be difficult to beat. Lenza et al. (2023) apply the quantile random forest (QRF) model to euro area data. QRF density forecasts appear competitive with those of a linear model and the ECB survey of professional forecasters.

Although inflation in many large economic blocks has been studied extensively, evidence for smaller countries, such as the the Netherlands, remains limited. Kohlscheen (2022) employs the RF model to examine the drivers of CPI inflation in a panel of high-income countries, including the Netherlands. Results for the Netherlands separately are not provided however. Vedder and van de Winkel (2024), replicating Goulet Coulombe et al. (2022) on Dutch data, demonstrate that many ML models can generate significantly better forecasts of CPI inflation than an AR model.

Our analysis also relates to studies on multiperiod-ahead forecasting. Multiperiod-ahead times series forecasts can be computed in various ways. Marcellino et al. (2006) use a large dataset of U.S. macroeconomic time-series and demonstrate that iterated AR(p) forecasts tend to outperform direct AR(p) forecasts, with the relative performance improving as the forecast horizon extends. Quaedvlieg (2021) reaches similar conclusions using tests for multi-horizon superior predictive ability. Kock and Teräsvirta (2016) examine multiperiod-ahead forecasting using nonlinear neural network prediction methods. They find that iterated and direct forecasts often exhibit similar performance, with their ranking depending on the dataset. Goulet Coulombe et al. (2021) provide a new perspective on this literature. They systemically compare direct forecasts to path average forecasts for several target variables. In a linear context, the two approaches should yield fairly similar results. However, for nonlinear models, path average forecasts offer greater flexibility, potentially leading to more accurate forecasts. They demonstrate that the preference for direct versus path average forecasts is often variable specific, although the latter is generally preferred for variables that strongly co-move with the business cycle. They find that for U.S. CPI inflation, direct forecasts are usually more accurate than path average forecasts.

Finally, our analysis is also related to the literature that analyses the properties of institutional forecasts. In the academic literature, simple time-series models are often used as benchmarks. A more pertinent issue for policymakers and other users of inflation forecasts is whether newly developed models can outperform existing institutional forecasts. In other words, it is important to determine whether adding one or more ML models to the existing suite of models is beneficial. Lenza et al. (2023) show that QRF forecasts of euro area inflation perform on par with the published Eurosystem inflation projections. Yoon (2021) demonstrates that boosting and RF models can produce forecasts of Japanese GDP growth that are more accurate than those made by the IMF and the Bank of Japan. Araujo and Gaglianone (2023) show that ML models can, in numerous cases, outperform traditional econometric models for Brazilian inflation.

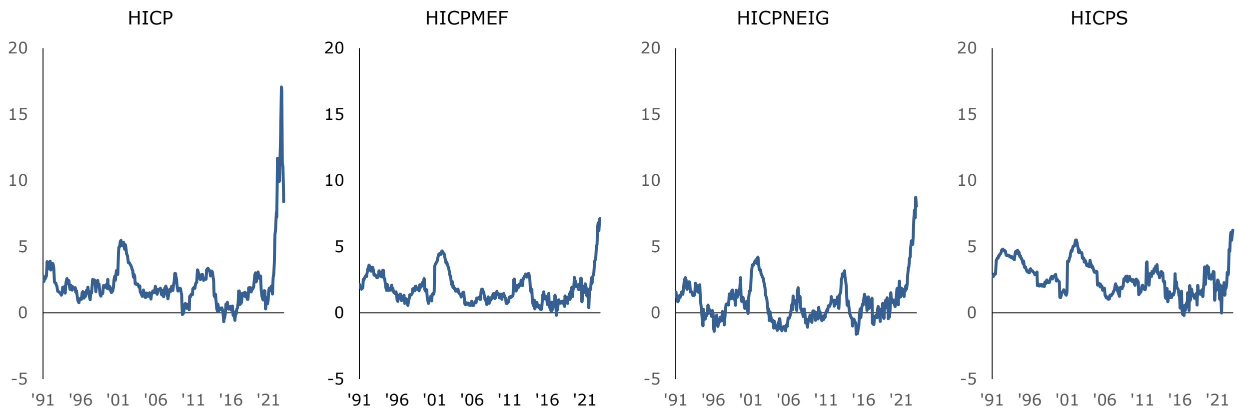
3 Data, forecast design and forecast evaluation

3.1 Data

We have compiled a dateset consisting of monthly observations of 129 Dutch and international macroeconomic time-series. All series were downloaded on March 4^{th}, 2024, and cover the period from January 1990 to January 2024. The target variables in our analysis are four inflation series: headline HICP (HICP) inflation, services inflation (HICPS), non-energy industrial goods inflation (HICPNEIG), and core inflation (headline HICP excluding energy and food, HICPMEF). Our list of time-series can be subdivided into eleven groups, following McCracken and Ng (2016) and Medeiros et al. (2021), namely: (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 monthly series in our database, we include as potential predictors the four principal components (factors) computed from this set of variables. Further details are provided in Section 4.5. We consider four lags of all variables, as well as four autoregressive terms of the dependent variable. Hence, the analysis contemplates 532 potential predictors. Table A.1 in Appendix A presents more details on the data series used, seasonal adjustments, and transformations of all series. In this paper, we consider two methods for constructing forecasts of year-on-year inflation: the direct method and the path average method. Details on both methods are presented in Section 3.4. The datasets used to compute these forecasts differ in how trending predictors are transformed to achieve stationarity. For path average forecasts, trending time-series are transformed to stationarity by taking (log) first differences. For direct forecasts, year-on-year changes of the predictors are used.

Figure 1 illustrates the development of headline HICP inflation and its three sub-components. After a period of inflation around 2 percent, headline HICP inflation increased during the pandemic. Following the first case of COVID-19 in the Netherlands in February 2020, inflation began a steep ascent in the summer of 2021, peaking at 17.1% in September 2022 before starting to decline. The main driver behind this increase was the significant rise in energy and food prices. This is evident from Figure 1, which shows that the increase in core inflation (HICPMEF) during the pandemic was much less steep, although still noticeable. Non-energy industrial goods inflation (HICPNEIG) is much more volatile than core inflation. The smaller increase in core inflation during the pandemic appears to be due to the relatively stable path of services inflation.

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



	HICP	HICPMEF	HICPNEIG	HICPS
Minimum:	-0.7	-0.2	-1.6	-0.2
P25:	1.4	1.0	-0.2	1.7
Median:	1.8	1.6	0.6	2.5
Mean:	2.3	1.8	0.9	2.7
P75:	2.6	2.4	1.5	3.5
Maximum:	17.1	7.1	8.7	6.3
Coeff. of variation	0.9	0.6	1.8	0.5

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

The pass-through of higher energy and food prices in the industrial sector is much stronger than in the services sector. This is also clearly seen by examining some distributional characteristics in Figures 1. 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 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 hindsight 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 forecasting 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 forecasting accuracy than the use of real-time (unrevised) data. Moreover, HICP inflation and its sub-components, our target variables, 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 forecasting 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 forecasting 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 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 benchmark models 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.

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:

$$CSSFED_{M,h} = -\sum_{t=t_0}^{t_1}(e^2_{t,M,h} - e^2_{t,BM,h}) \tag{1}$$

where M represents a machine learning model and BM represents a benchmark model. A CSS-

FED *above* zero indicates that the forecasts of the machine learning 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 a lower forecast accuracy at that point in time. Additionally, a *decrease* in the CSSFED indicates that the model performance of the machine learning 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 model M, we compute the $OOS.R^2$ as follow:

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

where $\epsilon^2_{t,h,M}$ 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:

$$OOS.R^2_{t,h,M_f} - OOS.R^2_{t,h,M_{-f}} = \alpha_{M,f} + \nu_{t,h} \tag{3}$$

The main advantage of equation (3) is that the coefficient $\alpha_{M,f}$ 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.

3.4 Direct and path average forecasts

Let Y_t be (a sub-component of) the monthly HICP index, and let x_t be a large macroeconomic dataset comprising of N predictors, for $t = 1, \ldots, T$. 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}, denoted by $\hat{y}^{dir}_{t+h}$, can be obtained using the following prediction model:

$$y^{dir}_{t+h} = G_h(x_t) + \epsilon_{t+h} \tag{4}$$

where $G_h(\cdot)$ is a (potentially non-linear) mapping between the predictor variables x_t and future inflation. 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:

$$y'_{t+h} = G_h(x'_t) + \epsilon'_{t+h} \tag{5}$$

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. $Y_t = Y'_t + \zeta_t$. A path average forecast of year-on-year inflation, denoted by $\hat{y}^{pa}_{t+h}$, is then computed as:

$$\hat{Y}^{pa}_{t+h} = \exp(\ln(Y'_t) + \sum_{i=1}^{h} \hat{y}'_{t+i}) + \zeta_{t+h} \tag{6}$$

$$\hat{y}^{pa}_{t+h} = 100 * (\hat{Y}^{pa}_{t+h} - Y_{t+h-12})/Y_{t+h-12} \tag{7}$$

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 benchmark 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 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 M terminal nodes can be written as:

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

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:

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

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

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

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 $\hat{y}_{t+h}^i$:

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

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

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 settle for four factors. 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. In Subsection 5.1, we graphically display the relative RMSFEs for all target variables, models, and forecast horizons. In Subsection 5.2, we identify the best-performing models for each inflation measure and forecast horizon. Subsection 5.3 investigates which specification choices contribute to forecasting accuracy (following Goulet Coulombe et al., 2021). Finally, Subsection 5.4 examines whether ML models can outperform DNB's official inflation forecast. Detailed results, including RMSFEs, $OOS.R^2$ statistics and p-values of the Diebold and Mariano-test, are provided in the online Appendix (here).

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

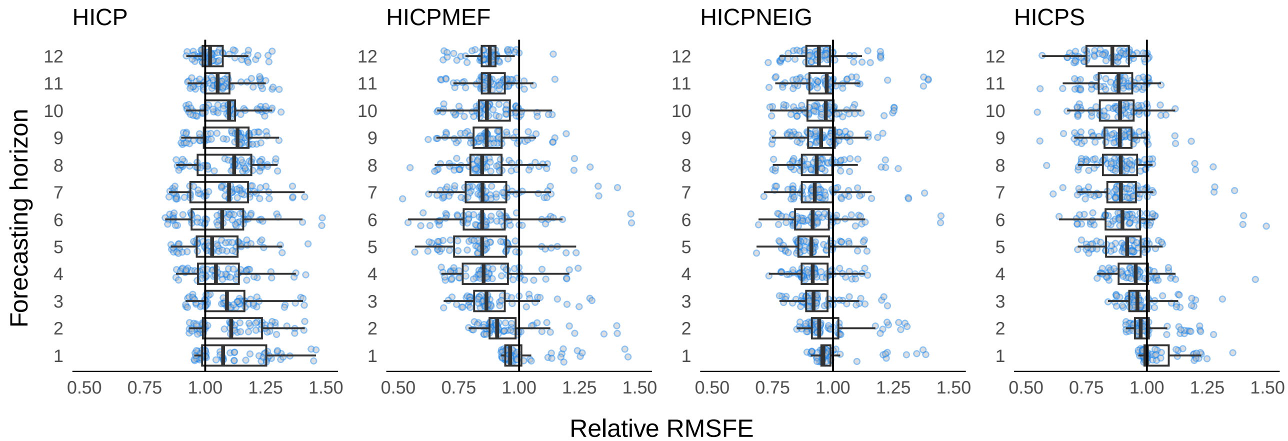
Figures 2 and 3 display the RMSFE relative to `M.RWD` –the path average random walk with drift model–for all 61 models, over the full sample and the pre-pandemic sample, respectively. The dots represent the relative RMSFEs of the individual models. To visualize the central tendency of the RMSFE distribution, the bold vertical line indicates the median across all models. The right and left boundaries of the boxplot represent the 75th and 25th percentiles, respectively, summarizing the spread of the distribution. The rightmost whisker is positioned at the smaller of the maximum RMSFE value and the 75th percentile plus 1.5 times the interquartile range (IQR). The leftmost whisker is positioned at the larger of the minimum RMSFE value and the 25th percentile minus 1.5 times the IQR.

The first conclusion that can be drawn from Figures 2 and 3 is that outperforming a simple benchmark model (`M.RWD`) for short forecasting horizons is quite challenging. However, ML models show substantial improvements over the benchmark for longer forecasting horizons, particularly in the full sample. The improvements in forecast accuracy over the benchmark model vary widely between models, ranging from minimal to as much as 48%, depending on the target variable and the forecasting horizon.

Second, there is a steadily increasing gain in the forecasting accuracy of the ML models for services inflation. However, the gains are markedly smaller in the pre-pandemic sample, which may be related to the stability of services inflation prior to the pandemic.

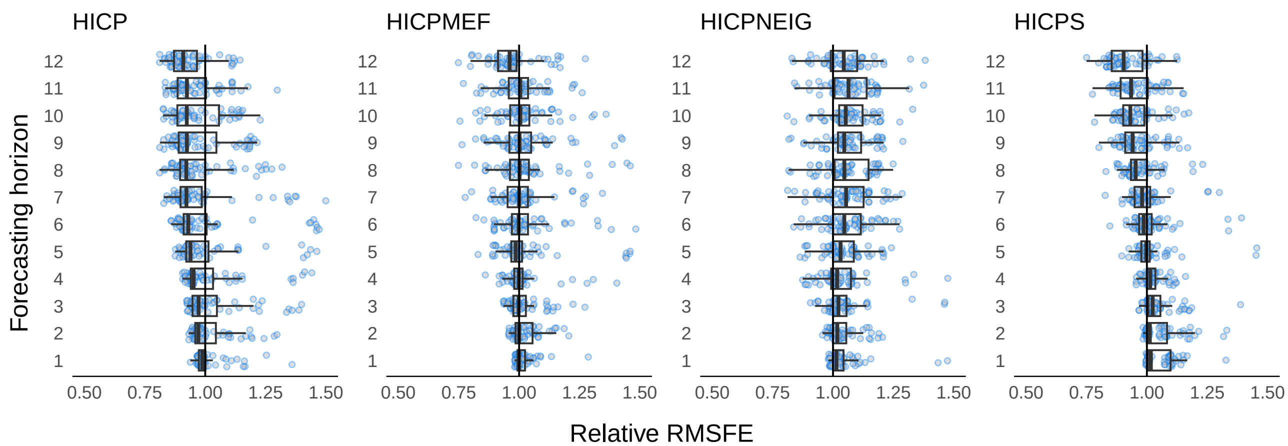
Third, for NEIG inflation, the median forecast across all ML model is unable to outperform

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.

the simple benchmark model in the pre-pandemic sample. However, the benchmark model can be outperformed in the full sample, particularly for medium term forecast horizons. The results for NEIG inflation 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.

Fourth, the differences between the full sample and the pre-pandemic sample are particularly notable for headline HICP. While most models are outperformed by the benchmark model (`M.RWD`) in the full sample, in the pre-pandemic sample, the median gain increases with the forecasting horizon. This is most likely related to the high volatility in food and energy prices in the post-pandemic period. The fluctuations in food and energy prices directly impacted HICP inflation, whilst the other inflation series were not or only indirectly affected via second round-effects by the high food en energy inflation. The machine learning models seem ill equipped to pick up these movements[1].

Finally, it is interesting to note that the spread of the forecasts appears to have widened since the pandemic, as shown by the larger boxplots in Figure 2 compared to Figure 3. This suggests that the surge in inflation since the pandemic has helped to distinguish between more and less accurate forecasting models.

5.2 The best performing machine learning models

To gain more insight into the best-performing model we analyze the model performance in more detail. Figure 4 and 5 show the *relative* forecasting performance of all 62 models analyzed against the benchmark model ($\mathrm{RMSFE}_{\mathrm{ML}}/\mathrm{RMSFE}_{\mathrm{M.RWD}}$) to provide a deeper insight into which models within the model types perform best. White cells indicate the forecasting 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 forecasting accuracy of the machine learning model is 5% or larger, red cells indicate there is 5% or larger decline in forecasting accuracy from using the machine learning model. The colors range from light green(red) to dark green(red), corresponding to a gain(loss) in forecasting accuracy ranging from 5% tot 50% or more.

The first conclusion from Figure 4 it that in the full sample, none of the models is capable of systemically outperforming the headline HICP inflation forecast of the benchmark model. Moreover, using direct forecasts (denoted by `Y.`) can even lead to a substantial lower forecast accuracy then using path average forecasts (denoted by `M.`). A possible explanation is that, since the volatility of HICP inflation is largely due to volatility in the energy and food components, our set of predictors is not sufficiently rich in terms of energy and food prices related series. Also, (announced) tax changes may play a role.

Second, for core inflation, NEIG inflation and services inflation, machine learning models can significantly outperform the benchmark for all horizons. Specifically, three specifications of the path average Ridge regression forecasts (`M.RR.NF.BIC`, `M.RR.F.CV`, `M.RR.F.BIC`) are among the

[1] This might also be caused by the selection of indicators in our database. There is a lack of long monthly time-series on energy components and food prices for the Netherlands.

best performing models across the three sub-components of inflation. These models outperform the benchmark model for core inflation and services inflation on all forecasting horizons, often by a substantial margin. For NEIG inflation, these specific Ridge regressions outperform the benchmark for at least seven forecasting horizons.

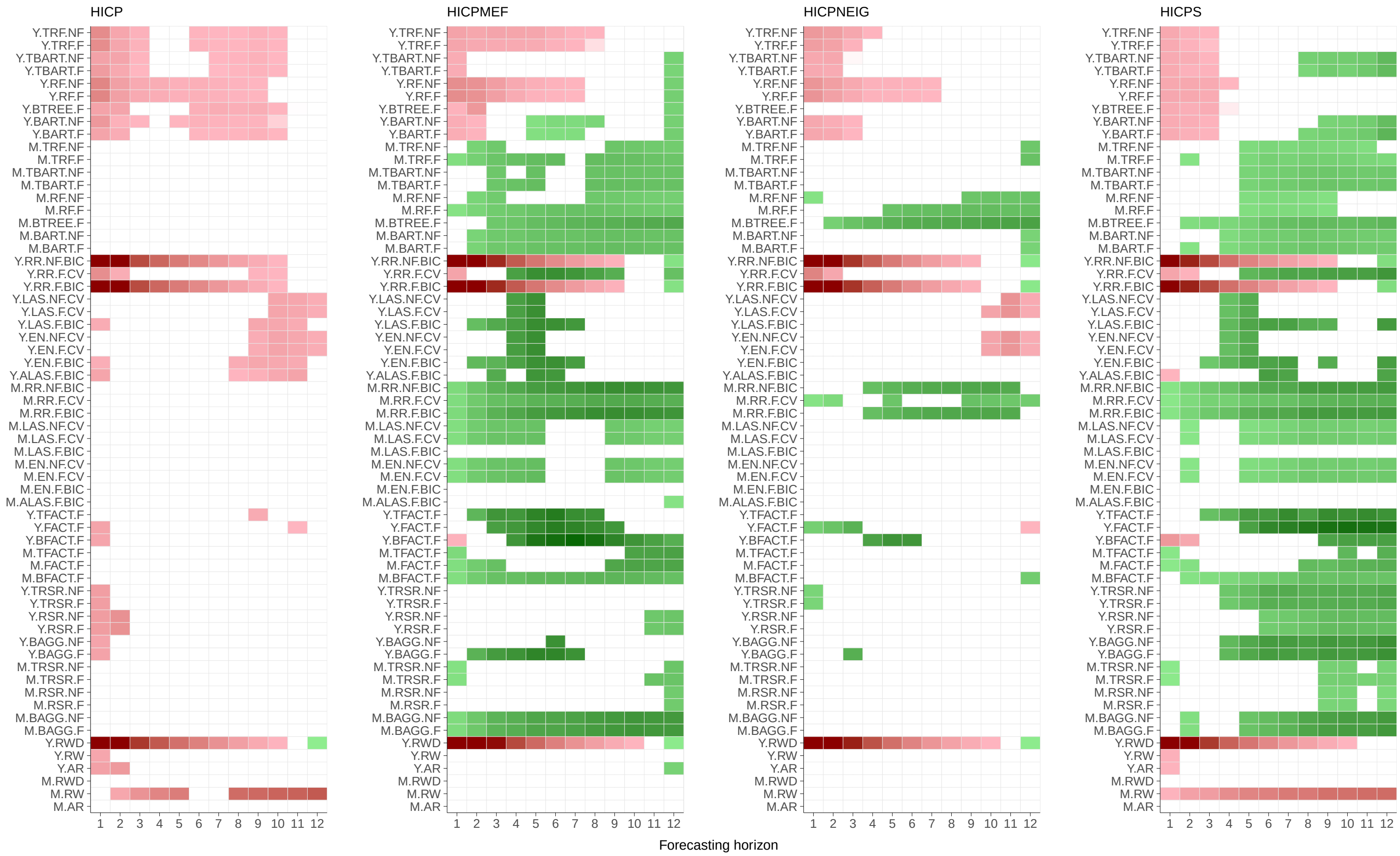
Third, some specifications of the factor model have smaller forecast errors than the Ridge regression model, but their performance is less consistent across horizons and target variables. For instance, the `Y.BFACT.F` model performs better than the Ridge regressions for forecasting core inflation 4 to 8 months ahead (the cells are darker green). However, on some forecasting horizons, this factor model does not outperform the benchmark model, whereas the Ridge regressions often do, and by a substantial margin.

Fourth, the forecasting accuracy of the tree-based models is quite poor. They do not systematically perform better than the linear shrinkage and factor models, as indicated by the darker green colors for the latter models. Moreover, the tree-based models show less consistent performance. The only model consistently outperforming the benchmark in the full sample is the path average random forecast model including factors (`M.RF.F`). This is a striking outcome, as in the literature tree-based methods are often found to be among the best-performing models.

Fifth, the outcomes indicate that path average inflation forecasts perform better than direct forecasts. The red cells are almost exclusively reserved for some of the direct forecasts. Strikingly, the Ridge regression, when using the direct forecast method, is the worst-performing machine learning model, while using the path average method, it is the best-performing machine learning model.

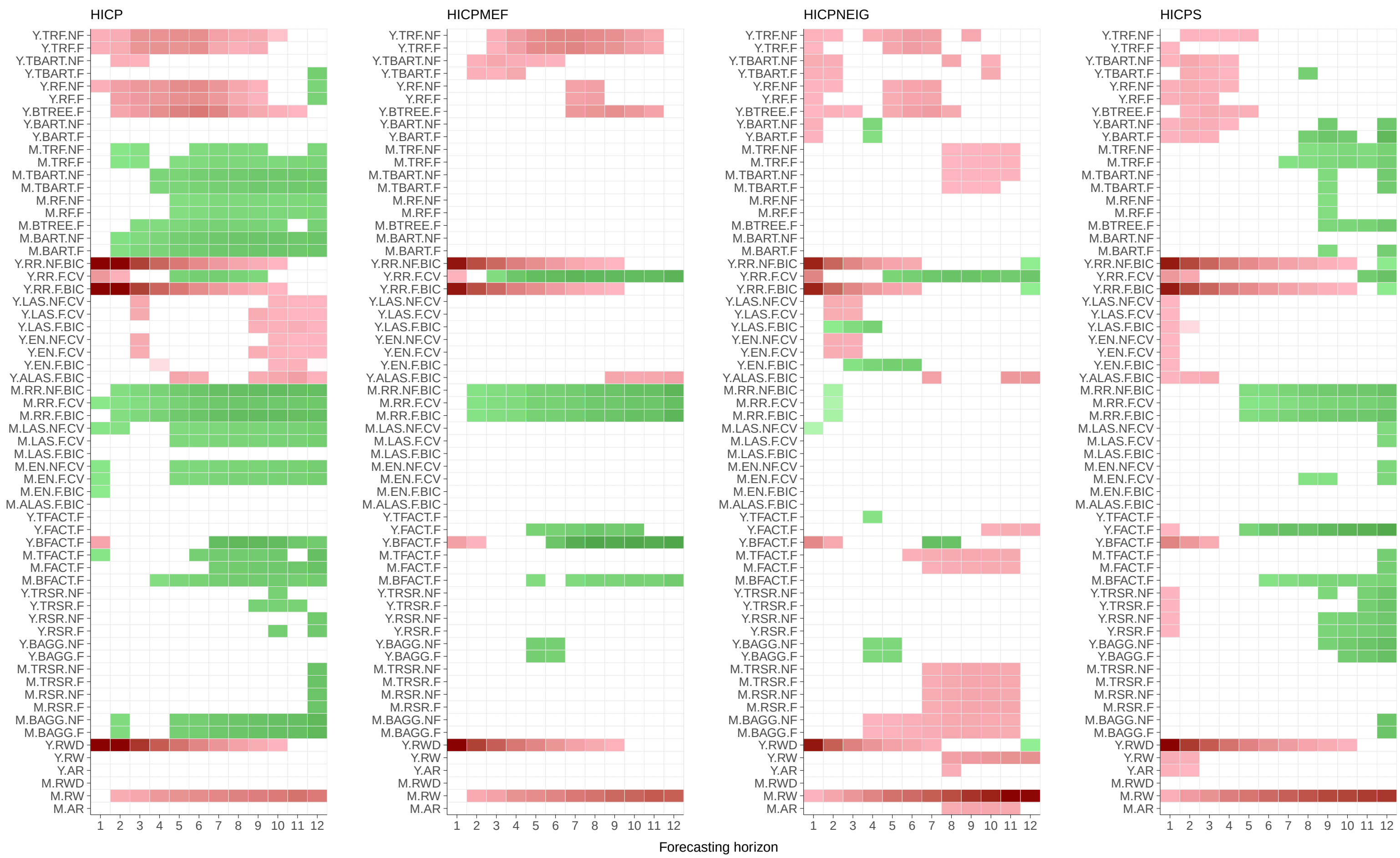
In the pre-pandemic sample (Figure 5), the results are somewhat different. For headline HICP inflation, various machine learning models now perform better than the benchmark. For the sub-components of inflation, the outcomes are actually reversed: there are significantly fewer (dark) green cells and more red cells. The path average Ridge regression models once again emerge as the best-performing models, consistently excelling in forecasting HICP inflation, core inflation, and services inflation. Except for one Ridge regression model (`M.RR.F.CV`) for headline HICP, none of the models surpass the benchmark model across all forecasting horizons. The forecasting performance of the factor models and tree-based models is significantly worse than when evaluated over the full sample. The relatively high forecasting accuracy of the path average Ridge regression is also evident in the pre-pandemic sample.

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



† White cells indicate the forecasting 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 forecasting accuracy of the machine learning model is 5% or larger, red cells indicate there is 5% or larger decline in forecasting accuracy from using the machine learning model. The colors range from light green(red) to dark green(red), corresponding to a gain(loss) in forecasting accuracy ranging from 5% tot 50% or more.

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



† White cells indicate the forecasting 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 forecasting accuracy of the machine learning model is 5% or larger, red cells indicate there is 5% or larger decline in forecasting accuracy from using the machine learning model. The colors range from light green(red) to dark green(red), corresponding to a gain(loss) in forecasting accuracy ranging from 5% tot 50% or more.

5.3 A closer look at model specification

The previous two subsections presented our findings focusing on the forecasting performance of model *types*. This section examines our outcomes through the lens of differences in model *specification*. As described in Section 4, models can differ across several dimensions. In this section, we look at three specification choices: (1) the transformation of the target variable, where we compare the relative performance of path average forecasts and direct forecasts, (2) the inclusion or exclusion of factors in the set of predictors, and (3) the use of all predictors versus using only targeted predictors. The list of modeling choices could be extended further, but some of them are specific to the model type. For example, for some models, the tuning of the parameters is done in several ways, such as via the `BIC` (Bayesian Information Criterion) or `CV` (Cross Validation). We do not consider these differences here.

We run regressions inspired by Goulet Coulombe et al. (2021) and Goulet Coulombe et al. (2022); see equation (3) in Section 3. Figure 6 shows the outcome of our comparison of path average versus direct forecasts for the full sample. The blue bars show the percentage of models for which the $OOS.R^2$ of the path average forecasts (`M`) is *not* significantly different from the direct forecasts (`Y`). On the other hand, the red bars show the percentage of models for which the path average forecasts are significantly better than the direct forecasts at the 10% level. Lastly, the green bars refer to cases for which the direct forecasts are significantly better than the path average forecasts, again at the 10% level. Figure 7 tests whether it is beneficial to add statistical factors to the set of predictors. Finally, Figure 8 examines whether using targeted predictors improves forecasting accuracy or not[2].

The main message from Figure 6 is that it is a 'no-regret' policy to forecast headline HICP inflation using path average forecasts. The blue (no difference between path average and direct forecast) and red bars (path average forecast better than direct forecast) sum to 100%, indicating that direct forecasts are never better. The same holds true for almost all forecasting horizons when forecasting NEIG inflation. The results are more ambiguous when forecasting core inflation and services inflation. The advantage of path average forecasts is their ability to model short-term volatility, leading to more stable long-term forecasts. On the other hand, direct forecasts may avoid cumulating errors, which are inherent in path average forecasts. The lower volatility of

[2] Factor models are excluded from the comparison in Figure 7. Shrinkage models always use the complete set of predictors and are excluded from Figure 8.

core inflation and services inflation lessen the need for using path average forecast as ‘insurance’ against short-term volatility. For the full sample, both the case for including factors in the set of predictors (Figure 7) and for using targeted predictors (Figure 8) is strong. For almost all horizons and measures of inflation, including factors doesn’t hurt and often improves forecasting accuracy. The same conclusion holds for the use of targeted predictors.

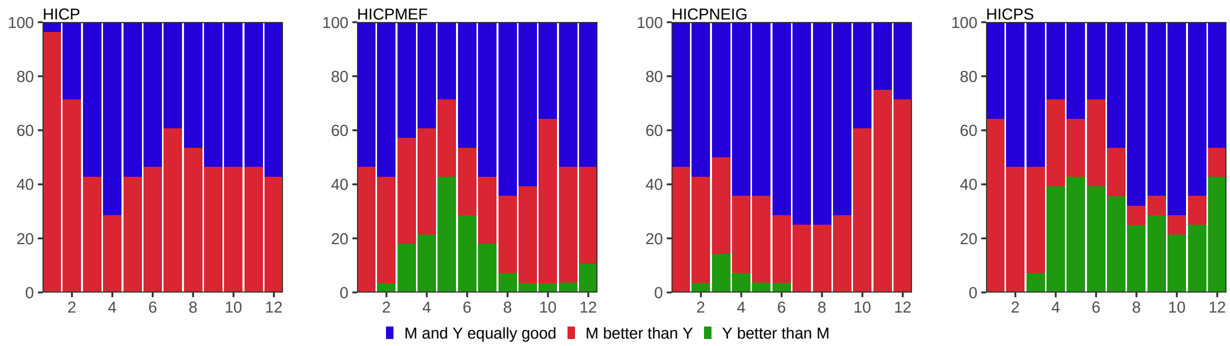
To gain more insight into the quantitative differences between direct and path average forecasts, Figure 9 shows the gain in $OOS.R^2$ of path average forecasts over direct forecasts, broken down by model type. For headline HICP inflation and NEIG inflation, the gains are larger at longer horizons, particularly for shrinkage models. Ensemble models perform worse when using path average forecasts. However, the weight given to the latter result should be low, because in the previous two sections, ensemble models hardly featured among the best-performing models. Depending on the forecasting horizon, factor models are sometimes better when using the direct forecasting approach.

The gains in the $OOS.R^2$ from adding factors and targeted predictors are much smaller, especially from adding factors.

Section B in the Appendix presents the results for the pre-pandemic sample. Qualitatively, the results are broadly similar to those for the full sample. For example, the use of path average forecasts also tends to raise the $OOS.R^2$. A key exception is that for factor models, it is clearly better to use *direct* forecasts for all inflation measures, apart from headline HICP inflation. This also holds for ensemble models, but given their weak forecasting performance it is inadvisable to use this model type anyway.

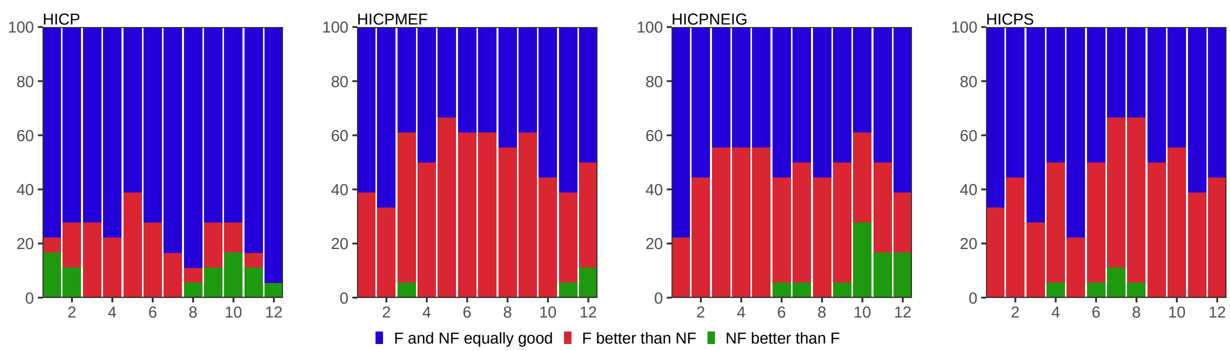
The main message from the $OOS.R^2$ analysis is that, in general, path average forecasts are the preferred option for inflation forecasting in our dataset. Exceptions are factor models, where the picture is more mixed, especially in the pre-pandemic sample. The gains from following this strategy are sizable. Including factors and using targeted predictors is generally a ‘no-regret‘ policy, although the gains are relatively small compared to the much larger gains from using the path average forecasts.

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



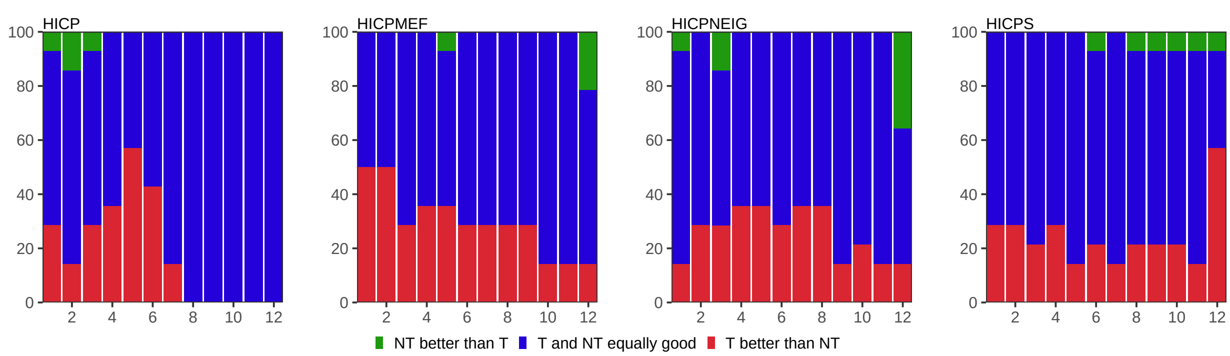
† M: path average forecast, Y: direct forecast, HICP = headline HICP inflation, HICPMEF = HICP inflation excluding food & energy, HICPNEIG = non-energy industrial goods inflation, HICPS = services inflation.

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



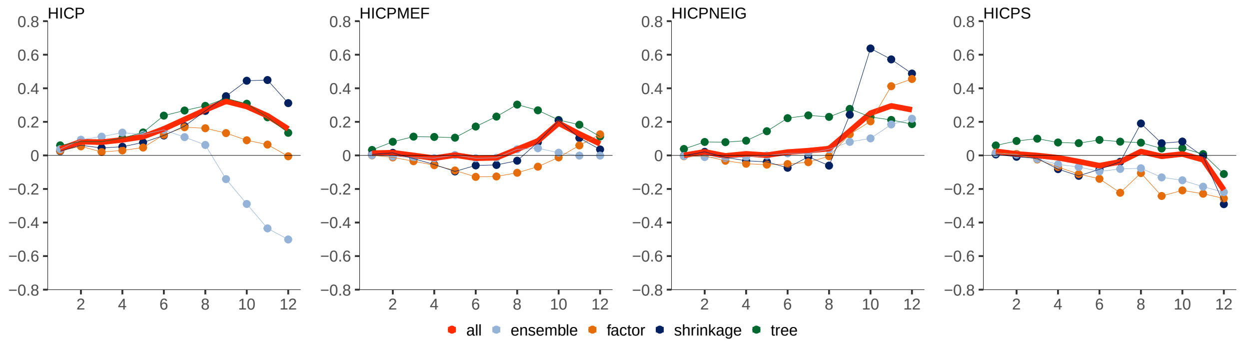
† F: factor augmented forecast (excluding factor models), NF: forecast without factors (excluding factor models), HICP = headline HICP inflation, HICPMEF = HICP inflation excluding food & energy, HICPNEIG = non-energy industrial goods inflation, HICPS = services inflation.

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



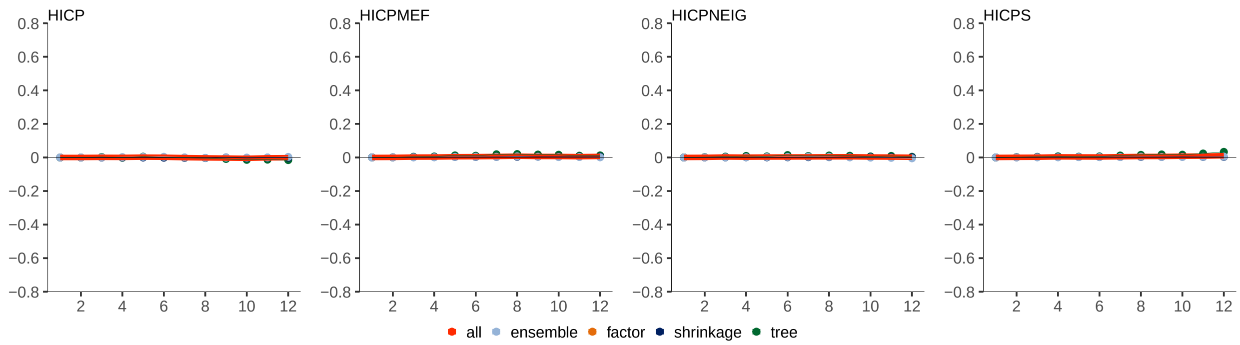
† T: targeted predictors (excluding shrinkage models), NT: predictors not targeted (excluding shrinkage models), HICP = headline HICP inflation, HICPMEF = HICP inflation excluding food & energy, HICPNEIG = non-energy industrial goods inflation, HICPS = services inflation.

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



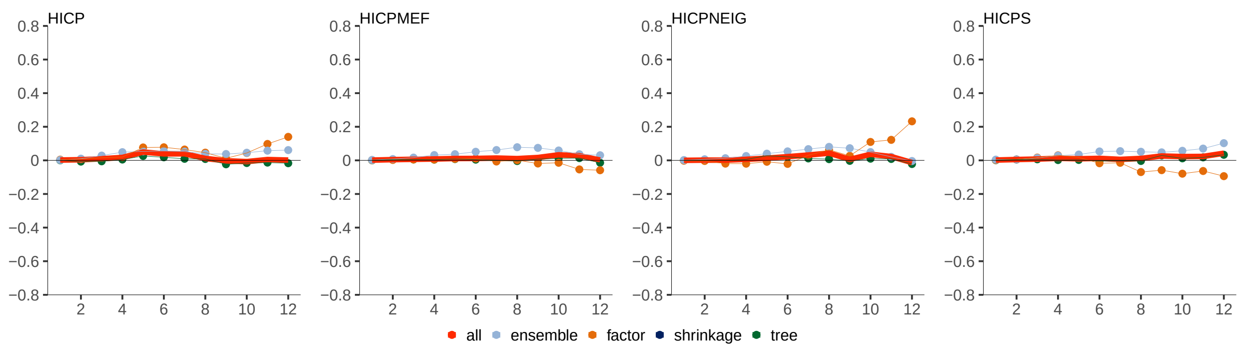
† Gain/decrease in OOS-R2 of using path average forecasts, HICP = headline HICP inflation, HICPMEF = HICP inflation excluding food & energy, HICPNEIG = non-energy industrial goods inflation, HICPS = services inflation.

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.

5.4 ML models versus institutional forecasts

In this section, we examine whether ML models can not only produce more accurate forecasts of Dutch inflation than simple benchmark models, but also outperform an established institutional forecast, DNB's official inflation forecast.

Since the introduction of the euro in 1999, DNB has been a member of the Eurosystem. The Eurosystem publishes macroeconomic projections four times a year[3]. Darracq Pariès et al. (2021) presents a recent assessment of the modeling toolbox currently in use within the Eurosystem, while Conrad and Enders (2024) investigate the accuracy of the inflation projections. DNB contributes to the quarterly projection exercises by providing forecasts of the Dutch annual inflation rate, ranging from 1- to 10-months into the future. These forecasts, known as the Narrow Inflation Projection Exercise (NIPE), will be referred to as NIPE throughout the remainder of the paper.

DNB's NIPE forecasts are based on a suite-of-models approach. This suite includes linear models for forecasting key components of HICP inflation, such as NEIG, food and services inflation, building on Den Reijer and Vlaar (2006). Additionally, SARIMA models are used to separately forecast more than 200 individual (COICOP) price sub-components of the HICP. Informal model averaging is employed to arrive at the final forecasts. The forecasts also consider announced government measures, such as changes in the VAT rate or energy taxes. Finally, the model-based forecasts are sometimes adjusted using expert-judgment to reflect relevant off-model information.

DNB produces inflation forecasts in four specific months of the year: February, May, August, and November. This contrasts with the ML model forecasts analyzed before, which are made every month. To fairly compare the forecast quality between NIPE forecasts and the ML forecasts, we will only use the forecasts from the four months when DNB forecasts are available. This means, for example, that we will only have 1-month ahead forecasts for March, June, September, and December, and 2-months ahead forecasts for April, July, October, and January, and so on.

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 forecasting accuracy against the benchmark model (RMSFE_{ML}-$\text{RMSFE}_{\text{NIPE}}$) of 5% (light

[3] See https://www.ecb.europa.eu/press/projections/html/index.en.html and https://www.ecb.europa.eu/pub/pdf/other/staffprojectionsguide201607.en.pdf

blue) to 50% (dark blue). The relative gains are only shown if the forecast gain against the NIPE is *at least* 5% *and* the gain is statistically significant at the 10% level according to the Diebold and Mariano test.

Figure 12 depicts the forecasting performance over the full sample. The key insights for this sample 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 machine learning in forecasting headline HICP compared to the naive benchmark model in Figure 4. What is quite surprising is that the NIPE model's headline HICP forecast is often significantly better than the machine learning models, especially for very short horizons (1 and 2 months). This suggests that relying solely on a machine learning model can negatively impact forecasting performance.

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

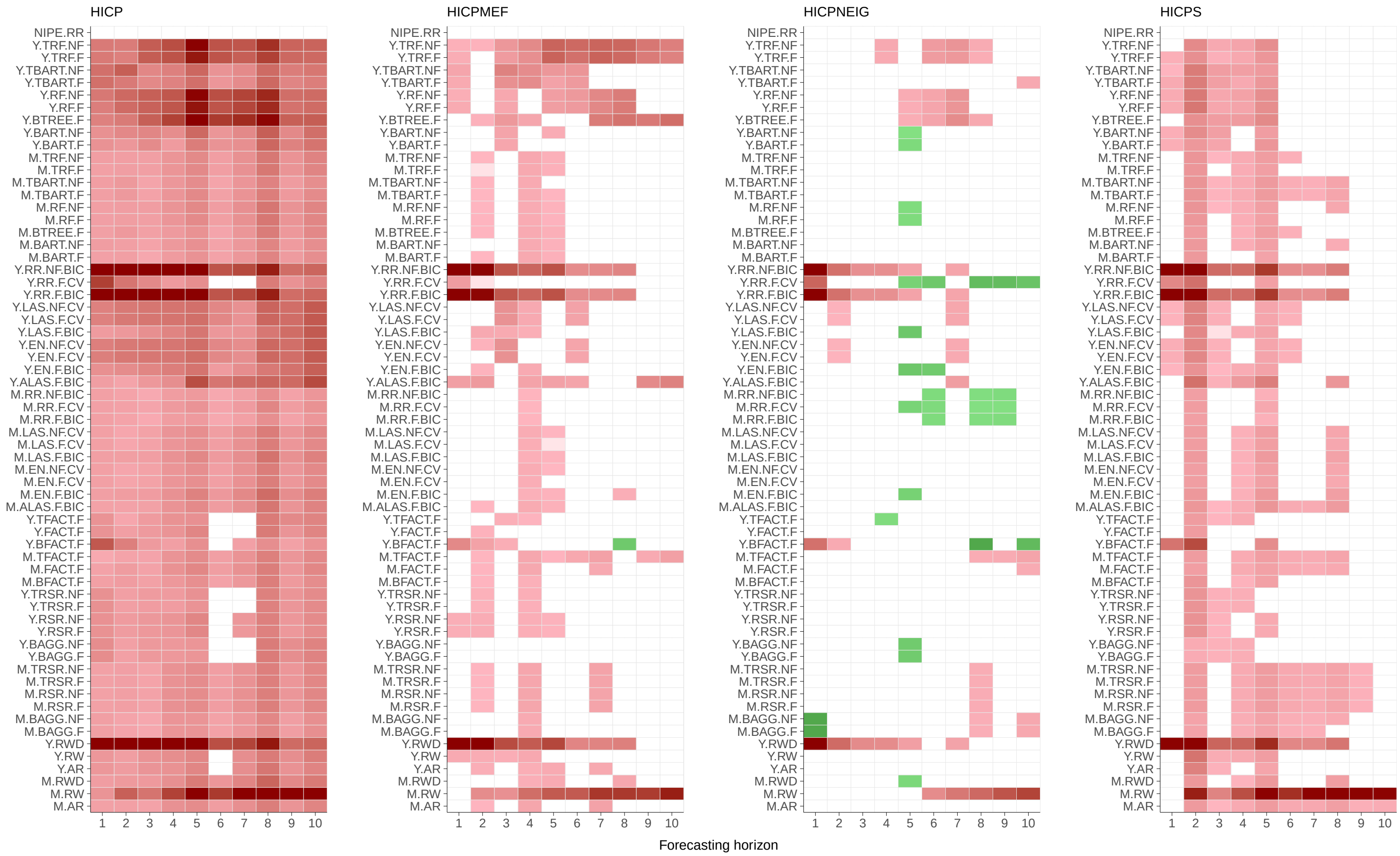
Figure 12: Heatmap relative RMSFE, full sample (2010M1-2023M1) †

† White cells indicate the forecasting 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 forecasting accuracy of the machine learning model is 5% or larger, red cells indicate there is 5% or larger decline in forecasting accuracy from using the machine learning model. The colors range from light green(red) to dark green(red), corresponding to a gain(loss) in forecasting accuracy ranging from 5% tot 50% or more.

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

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



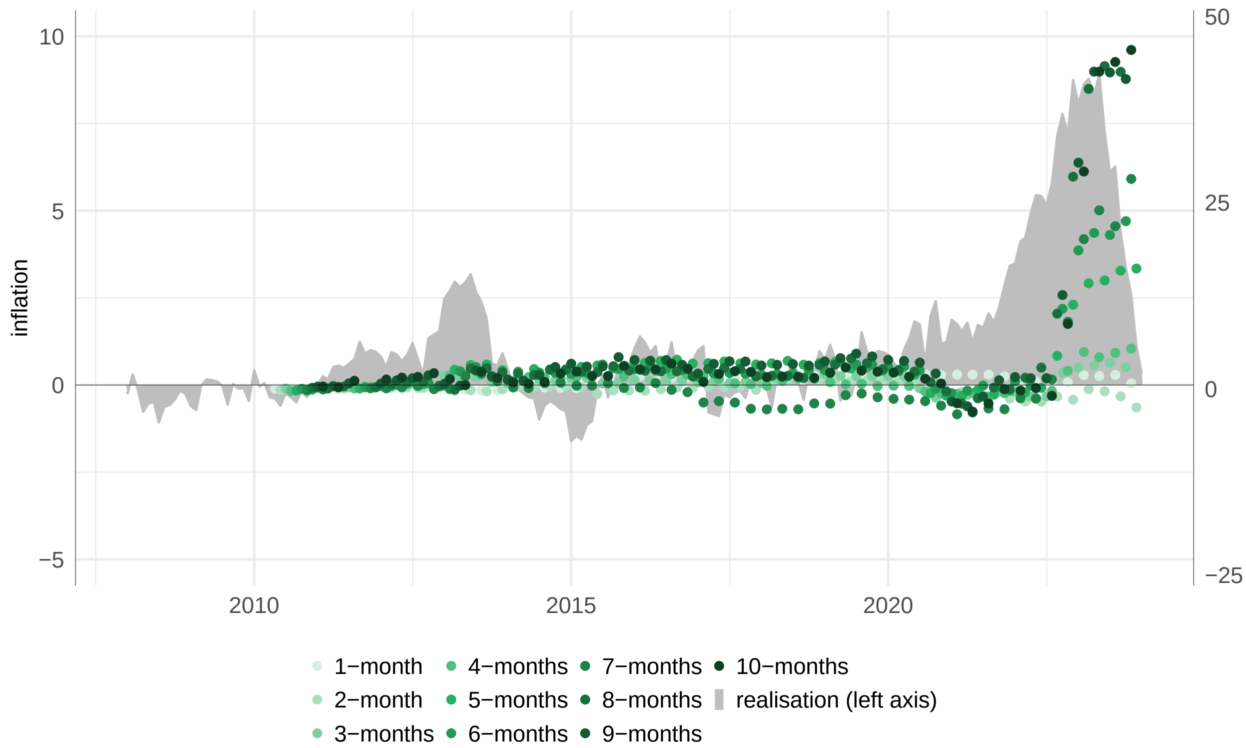
† White cells indicate the forecasting accuracy of the machine-learning model is not statistically different from the NIPE forecast, i.e. the Diebold and Mariano-test is not statistically significant at the 10% level. A green cell indicates the relative gain in forecasting accuracy of the machine learning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

tical significance across all horizons[4]. 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 forecasting 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 forecasting horizons of 1 to 10 months ahead (right-hand axis).

[4] We tested for average superior predictive ability, as described by Quaedvlieg (2021). The ridge models have - on average - statistically significant lower mean Square 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).

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

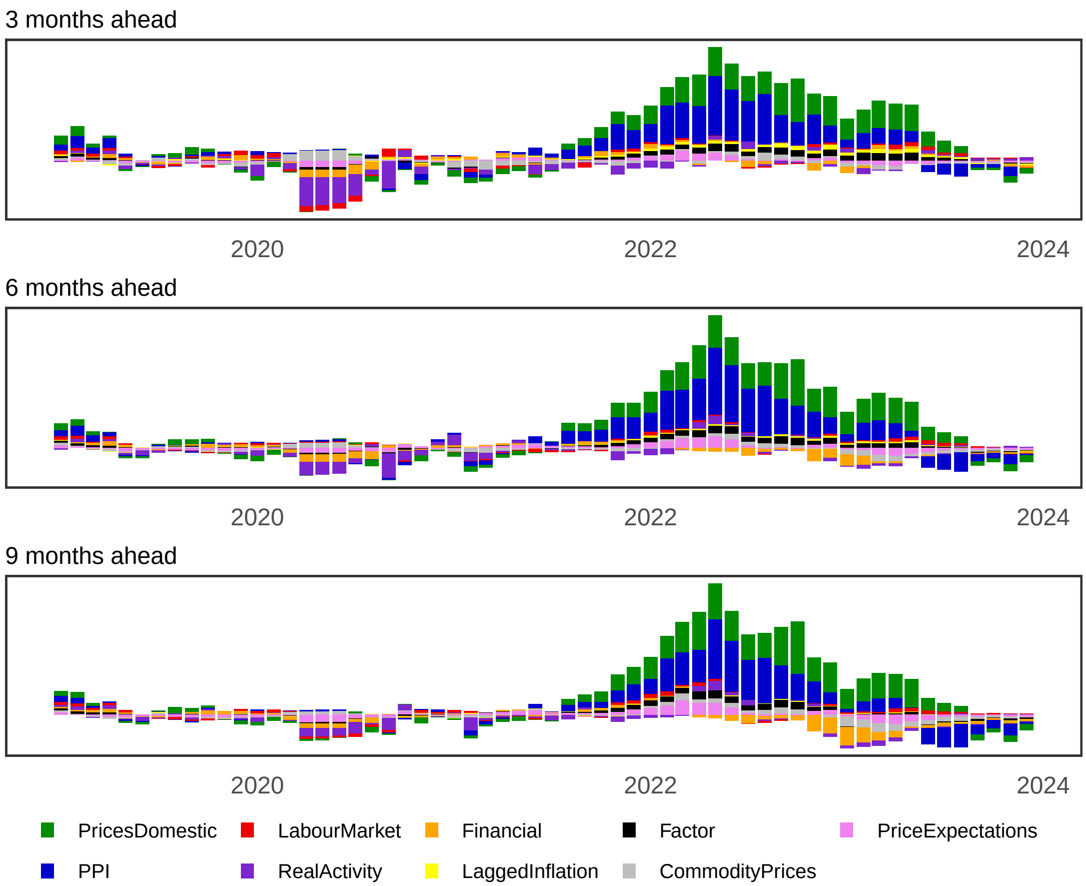


Between 2010 and 2019, `M.RR.F.BIC` shows some marginal forecasting gains over the NIPE forecasts for most forecasting horizons. From 2020 onwards, as NEIG inflation edged up sligthly, there was an initial slight deterioration in the CSSFED. However, as NEIG inflation surged later on, `M.RR.F.BIC` clearly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

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



6 Concluding remarks

In this paper, we examine whether machine learning methods can generate accurate short- and long-term inflation forecasts for the Netherlands.

In the first 'horse race', we compared forecasts of machine learning models to those of a simple benchmark model. We found that the improvements in relative RMSFE ranged from negligible at short horizons to more than 40% at longer horizons, depending on the inflation measure. Machine learning models are not very good at forecasting headline HICP inflation. Across the full sample, none of the machine learning models outperformed the simple benchmark model. However, when excluding the pandemic and post-pandemic periods, the results for the machine learning models are more favorable. Bycontrast, for core inflation, NEIG inflation, and services inflation, the relative forecast performance of machine learning models improves when the pandemic and post-pandemic data are included in the evaluation sample.

In the second evaluation exercise, we assessed the performance of machine learning models against DNB's official NIPE forecast. We found that the performance of the machine learning 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 outperformed the NIPE forecast.

Combining the outcomes of both forecasting 'horse races', the preferred machine learning model is the path average Ridge regression. The `M.RR.F.BIC` model either outperformed or matched the forecasting 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.

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.

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 (2024). China's inflation forecasting in a data-rich environment: based on machine learning algorithms. *Applied Economics forthcoming*, 1–26. 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 (2023). Density forecasts of inflation: a quantile regression forest approach. Working Paper 2830, European Central Bank. 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, A. 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 series that were not seasonally adjusted at the source using the US Census X13 ARIMA-SEATS method. 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[5]. 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.

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

Continued on next page ...

[5] 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
10.	Manufacturing of chemicals & chemical 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
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 dem&	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

Continued on next page ...

Table A.1 – Continued from previous page

Nr.	Description	Source	Transf.	Last
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
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, headline	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				

Continued on next page ...

Table A.1 – Continued from previous page

Nr.	Description	Source	Transf.	Last
108.	Harmonized index of consumer prices, headline	ECB	2,7	Jan-24
109.	Harmonized index of consumer prices, headline excl. energy & food	ECB	2,7	Jan-24
110.	Harmonized index of consumer prices, headline excl. energy	ECB	2,7	Jan-24
111.	Harmonized index of consumer prices, unprocessed food	ECB	2,7	Jan-24
112.	Harmonized index of consumer prices, food	ECB	2,7	Jan-24
113.	Harmonized index of consumer prices, services	ECB	2,7	Jan-24
114.	Harmonized index of consumer prices, industrial goods excl. energy	ECB	2,7	Jan-24
115.	Harmonized index of consumer prices, processed food	ECB	2,7	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, 7 = annual percentage change; Last: Last monthly observation.

B Additional OOS.R2 results

Figure B.1: Test of direct versus path average forecast, pre-pandemic sample (2010M1-2019M12)

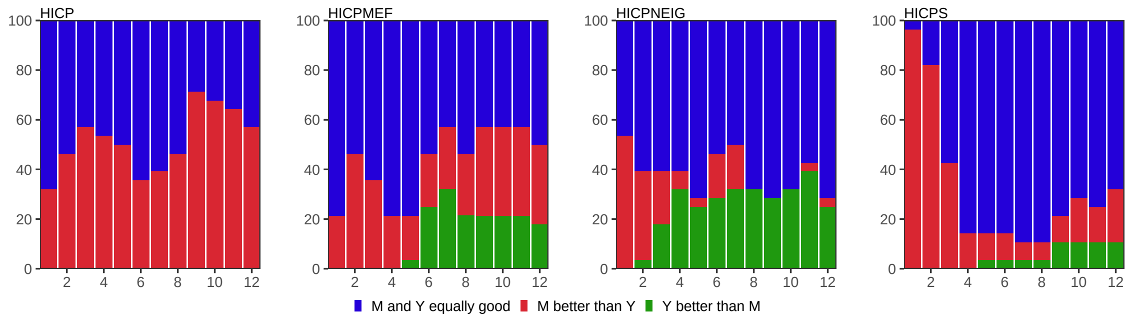


Figure B.2: Test of factor versus no factor forecast, pre-pandemic sample (2010M1-2019M12)

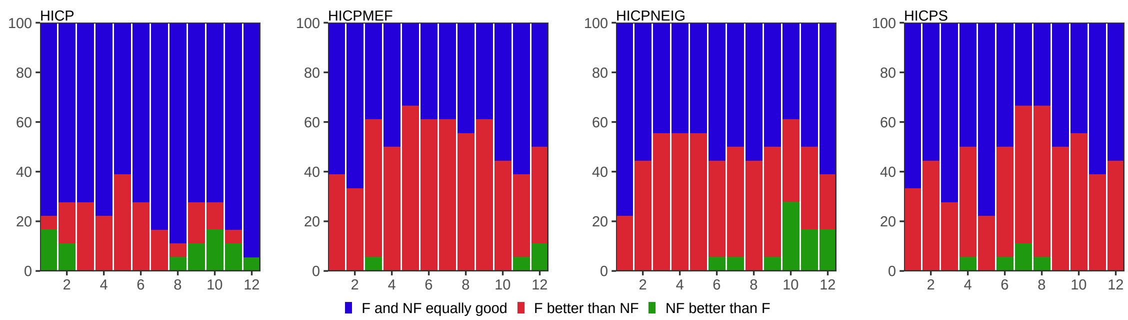


Figure B.3: Test of of targeting versus no targeting of predictors, pre-pandemic sample (2010M1-2019M12)

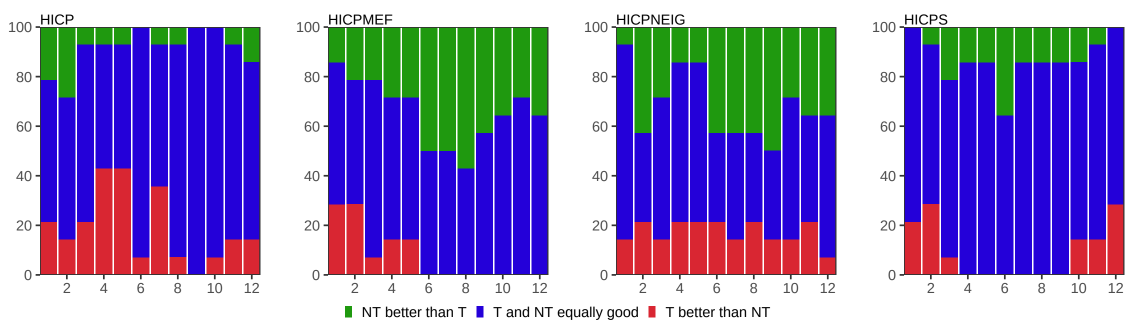


Figure B.4: Gain in $OOS.R^2$ of path average forecast over direct forecast, pre-pandemic sample (2010M1-2019M12)

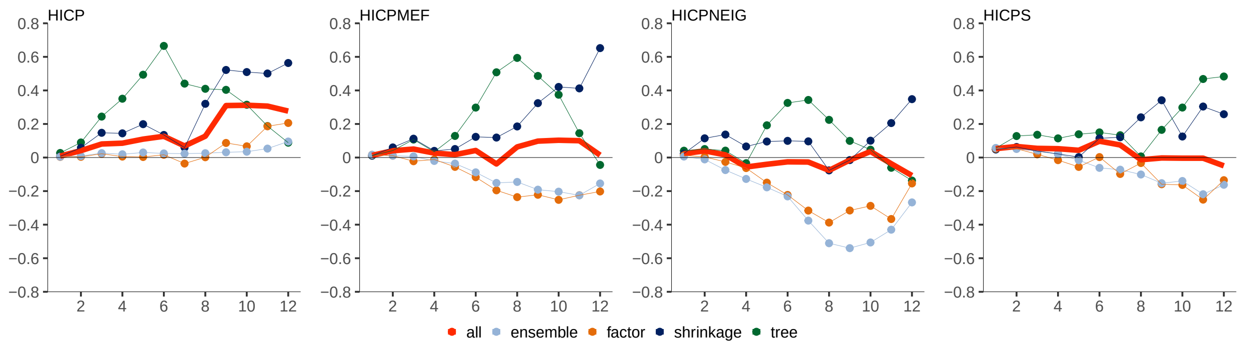


Figure B.5: Gain in $OOS.R^2$ of factor versus no factor forecast, pre-pandemic sample (2010M1-2019M12)

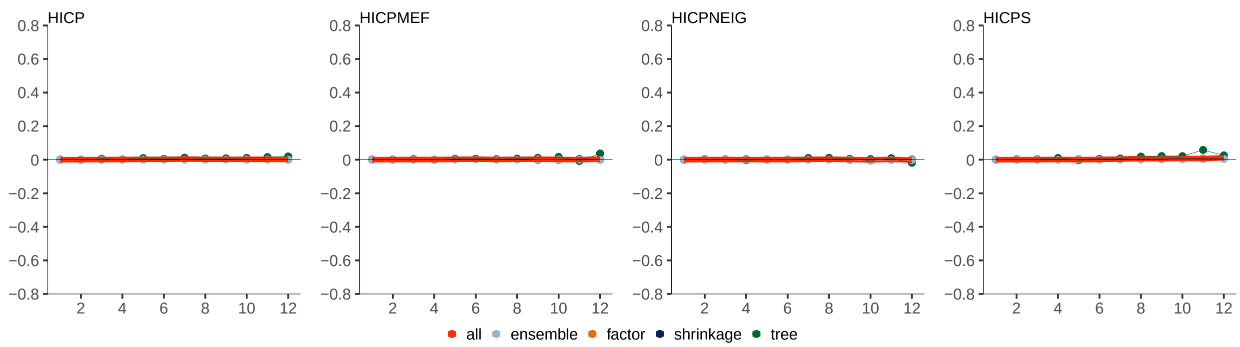
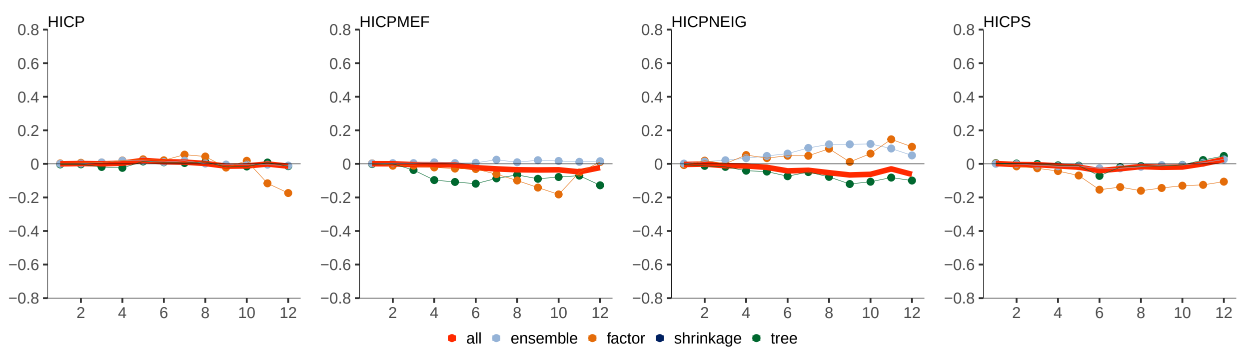


Figure B.6: Gain in $OOS.R^2$ of targeting versus no targeting of predictors, pre-pandemic sample (2010M1-2019M12)



DeNederlandscheBank

EUROSYSTEEM

De Nederlandsche Bank N.V.
Postbus 98, 1000 AB Amsterdam
020 524 91 11
dnb.nl