Can Machines Beat Professional Economic Analysts in Forecasting Macroeconomic Indicator(s)? Master Thesis

Student: Anna Vlasova
Supervisor: Thierry Foucault

June 21, 2023

Abstract

This Thesis deals with the forecasts of inflation in the Euro Area made by experts (Survey of Professional Forecasters by the ECB), explores their properties such as stickiness and compares them to the machine-made forecasts on various horizons. Machine Learning predictions are made using Random Forest and Ridge regressions. We find that the machine-made predictions are generally more accurate than the Forecasters' predictions. This superior performance is mostly explained by the shift in inflation paradigm that happened in the post-Covid period that the machine learning models were able to foresee to a significantly higher degree than the Forecasters.

Contents

1 Introduction

Inflation is the key macroeconomic variable which is central to all economic agents behaviour. Inflation may impact consumers' finances, firms' decision making, the sentiment of the financial markets, etc. As it is crucial for the economy, inflation is also the main driver of the central banks' decisions. To keep the economy stable, they aim to keep the inflation at the long-term target rate, which is usually around 1.5-2% in the developed countries. To keep the inflation in check, the central banks use a variety of instruments such as policy rates which ultimately affect the real economy. It is then only natural that accurate forecasts of inflation are of utmost importance. To this date, inflation predictions made by experts are often used as a benchmark for inflation expectations. Not only that, Machine Learning is increasingly used to predict inflation as well.

In this work, we would like to explore the properties of inflation forecasts made by professionals and compare them to the Machine Learning predictions. We would like to see if humans are biased in their predictions, and if humans, as well as machines, can react to shifts in inflation paradigms as happened recently in the post-Covid period. In this work, we focus on the Euro Area.

A central piece in our work is the data from the ECB Survey of Professional Forecasts (SPF). It contains macroeconomic forecasts of experts (Forecasters) working in financial or non-financial institutions. This data provides a good benchmark for inflation expectations of professional analysts and allows us to look for patterns, and compare human performance to machine performance. The data spans from 1999 to 2022. The Forecasters' predictions are made on various horizons. We begin this work by describing this extensive data-set in the Section Forecasts Data Description.

The Survey of Professional Forecasters (SPF) in the Euro area has been explored in a number of papers. In particular, *Geoff, Genre, Bowles, Friz, Meyler, & Rautanen (2007)* wrote a review on the SPF after 8 years of experience for the ECB Occasional Paper Series. In particular, they found that, over the sample period, the Forecasters systematically underestimated the inflation. Using statistical tests, they find strong evidence that the SPF aggregate inflation forecasts were biased. However, *Geoff, Genre, Bowles, Friz,*

Meyler, & Rautanen (2007) treat their results with caution, as the sample period is not only relatively short, but also contained a number of hardly predictable inflationary shocks. We find the same pattern over our sample period, which is a considerably longer one. We connect it to the post-Covid inflation, which came as an unprecedented upward shock. When looking at the pre-Covid period, the bias is less pronounced.

Forecasters are likely are a heterogeneous group of experts. It is plausible that some Forecasters possess superior prediction capabilities and consistently outperform their peers. Conversely, some of the experts may consistently produce low-quality predictions due to a number of factors such as inattentiveness or poor understanding of inflation dynamics. We call these 2 Forecasters pools 'Star' and 'Underperformers' respectively. We find the 'Star' and the 'Underperformers' groups of Forecasters data in the Section Forecasters' Classification on the criterion of predictions' accuracy for each horizon. Top 10% Forecasters for each horizon are placed into the 'Star' group, bottom 10% for each horizon are placed into the 'Underperformers' group.

Stickiness is a well-known property of forecasts. Forecasters may 'stick' to their previous forecasts by being inattentive and not updating their information sets. In the Section Stickiness of Forecasters, we check if and how sticky the SPF forecasts are. We find that the forecasts are sticky on many of the horizons, and that the stickiness degree varies between horizons. Moreover, the stickiness degree varies among Forecasters, some of whom exhibit almost constant levels of stickiness for a given horizon. Furthermore, Forecasters incorporate new information better in recessions than in expansions.

Our central objective is to compare human and machine-made forecasts. To generate machine-made predictions, we use Machine Learning models, namely the Ridge Regression and the Random Forest Regression. We thoroughly describe our Machine Learning approach, including the choice of predictor variables, in the Section Machine Learning Models.

Our findings in the Subsection General Case of the Section Results indicate that the Ridge Regression outperforms both the Forecasters and the Random Forest Regressions in predicting inflation on most of the horizons.

We also compare the performance of 'Star' Forecasters and the Machine Learning models. 'Star' Forecasters performance is on par with the Ridge Regression performance on the short-term horizons, suggesting that 'Star' Forecasters may indeed possess superior forecasting capabilities and the ability to use more relevant information for forecasts than the general pool of Forecasters. We proceed to assuming that Ridge Regression represents the rational prediction and find the Forecasters' bias on the 1 rolling year and 2 rolling years horizons to be strongly negative. We partially link the negativity to the post-Covid upward inflation shock that the Forecasters may have failed to capture.

A key point to be addressed in our analysis is the shift in inflation regimes that happened in the post-Covid period. It might be possible that the Forecasters did not fully realise the shift in paradigms and / or that the machine is unable to fully capture the shift in frameworks. To analyse the pre-Covid and the post-Covid performance, we split the test data into 2 subsets encompassing low and high inflation regimes. Then, human and machine-made predictions are compared in the Subsection High vs Low Inflation Regimes of the Section Results. This subsection presents a slightly different picture from the Subsection General Case. In the pre-Covid period, the Ridge Regression does not outperform the Forecasters on most of the horizons. This may be connected to the nature of these horizons, as those on which Forecasters outperform the machine, have limited data, not allowing the model to reach its full potential during training. Still, the 'Star' Forecasters exhibit a better performance than the Machine Learning models on one of the 2 horizons where the sample size was big enough for training. As for the post-Covid period, Ridge Regression performance is significantly better than the Forecasters' performance. Forecasters have likely not adapted to the shift in paradigm and did not update their information sets, exhibiting stickiness. Adaptability of the Ridge Regression permits it to capture such change in trends significantly better than humans. As for the 'Star' Forecasters, they have exhibited the least stickiness and were able to foresee the regime shift better than the general group of Forecasters. We propose that the best experts in the field remain such thanks to their ability to capture and use new information. Overall, the performance of the 'Star' Forecasters does not lag as far behind from the performance of the Ridge regression as the performance of all Forecasters. 'Star' Forecasters' edge can be explained by many factors: su-

perior understanding of inflation-forming mechanisms and relationships between different macroeconomic variables, access to the extra information that the machine does not get in the form of features (i.e. news, people's sentiment, etc). However, machine's clear advantage is in the lack of stickiness and in its responsiveness to new information.

Bias of human analysts has already been explored by *van Binsbergen, Han, and Lopez-Lira (2022)* who introduced "a real-time measure of conditional biases to firms' earnings forecasts". The measure consists of the difference between analysts' expectations and statistically optimal unbiased Machine Learning benchmark. In their work, they use Random Forest regression as their primary unbiased machine-made forecast benchmark. This enables *van Binsbergen, Han, and Lopez-Lira (2022)* to compute the real-time estimated analysts' bias. Our approach is similar, as we treat the machine made forecasts as rational and compute biases of Forecasters for two of the horizons in the Section Results. Moreover, *van Binsbergen, Han, and Lopez-Lira (2022)* find that analysts are generally upwardly biased, thus overly optimistic in case of earnings' forecasts. We are unable to give a conclusive answer as to whether in our data Forecasters are upwardly or downwardly biased in relation to the rational predictions. In fact, we find that the bias depends on horizon.

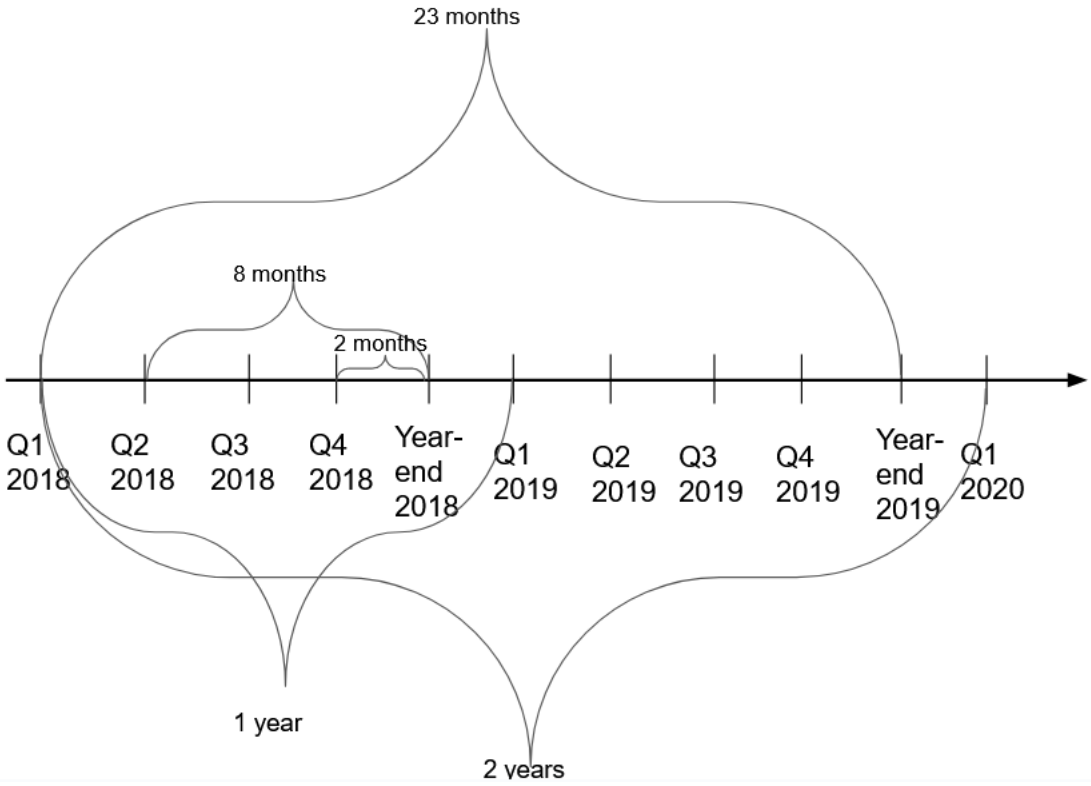
Our work is also closely related to *Coibion and Gorodnichenko (2015)* who introduced a new approach to test the degree of the stickiness of forecasts which we use in this work. *Coibion and Gorodnichenko (2015)*'s approach consists in testing the null hypothesis of full-information rational expectations by relating mean forecasts errors to subsequent forecasts revisions. In a full-information rational expectations framework, there should be no relation between these 2 variables. However, they find there usually is and this relationship is linked one-to-one to the degree of information rigidity (stickiness). *Coibion and Gorodnichenko (2015)* focus on the inflation expectations data from the Survey of Professional Forecasters in the U.S. with sample period from 1969 to 2010. They estimate the degree of information rigidity to be positive, and we get similar results using the European data in this work. *Coibion and Gorodnichenko (2015)* also study if stickiness varies with economic conditions, and find that it increases in times of lower volatility, which coincides with our results for expansions and recessions.

Genre, Kenny, Meyler, and Timmermann (2013) explored various techniques (principal components and trimmed means, performance-based weighting, etc) of combining SPF individual forecasts to obtain better performance than the simple average performance. Even though they treat the results with caution, they find the strongest evidence for improvement for the inflation forecasts. In the Section Conclusion of our work, we hypothesize that combining professional forecasts with machine forecasts using some of the sophisticated techniques from *Genre, Kenny, Meyler, and Timmermann (2013)* may produce superior inflation predictions.

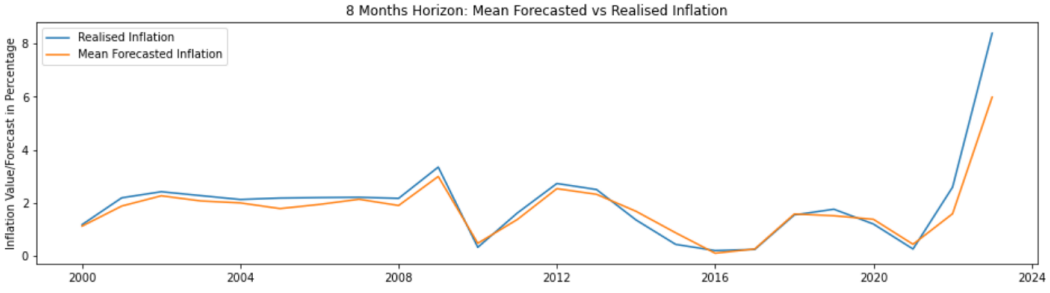
2 Forecasts Data Description

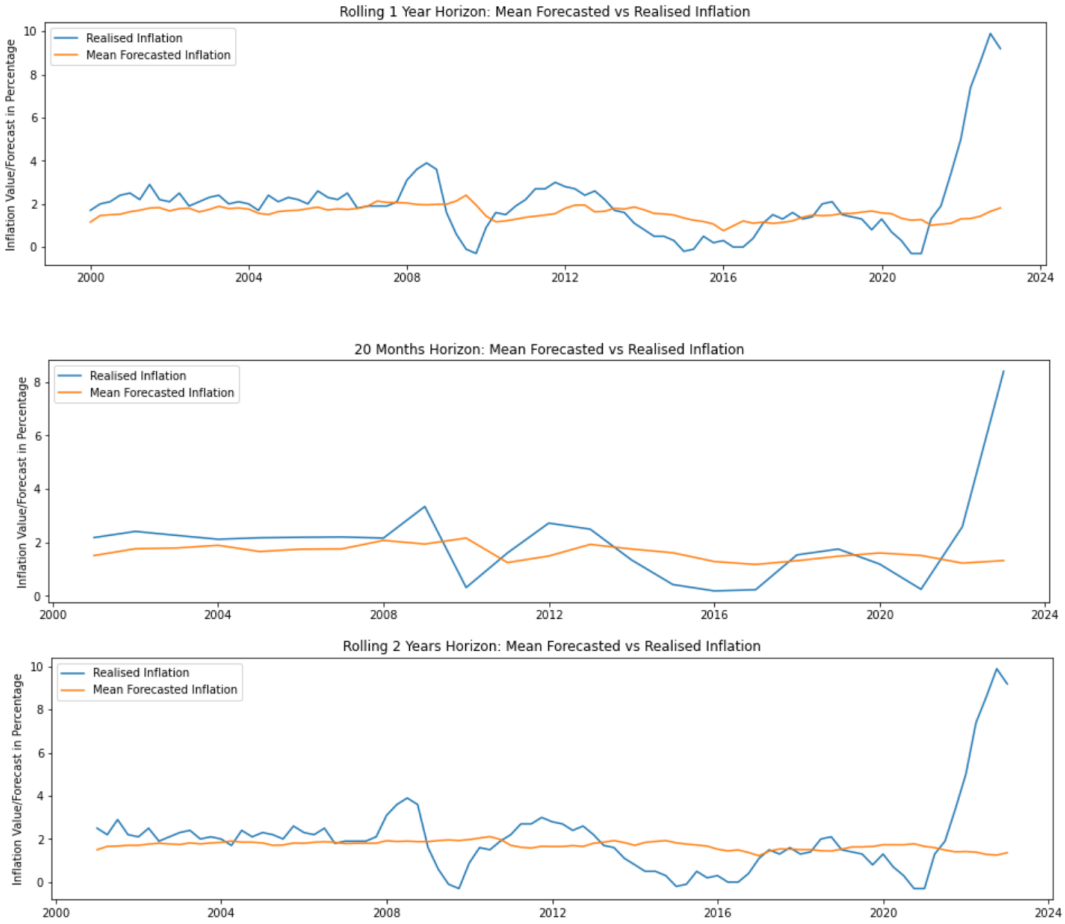
In this Section, we aim to describe the inflation forecasts data. We use the Survey of Professional Forecasters (SPF) from the ECB. This Survey is a comprehensive source of macroeconomic expectations for the Euro Area and has been conducted quarterly since 1999. The Survey provides not only forecasts for the Euro Area HICP inflation, but also expectations for Euro Area unemployment and GDP growth. However, for our analysis, we focus solely on the Euro Area inflation forecasts. According to the ECB, respondents of the Survey are experts of the financial or non-financial institutions. Each respondent (Forecaster) has a unique identification number, so that their answer can be traced across Surveys. Total number of Forecasters is 135, although it is unlikely that all of them participated in a given single Survey. The data we use spans from Quarter 1 1999 to Quarter 3 2022. The Survey collects HICP inflation expectations at various horizons, including the current calendar year, the next calendar year, the calendar year in 2 years, the calendar year in 4 years, the calendar year in 5 years, as well as 1 rolling year ahead, 2 rolling years ahead, and 5 rolling years ahead horizons.
For the realised inflation rates data, we use the Euro Area HICP inflation data from the ECB.

For our analysis, we create customised horizons based on the calendar year horizons in the Survey. For example, we form horizons of 2 months (available only in Quarter 4 as it represents the current calendar year forecast made in Q4 with 2 months remaining until the year-end), 5 months (available in Q3), 7 months (available in Q2), and so on. Below is a schematic depiction of selected horizons:



We compare the consensus inflation expectations with the realised inflation rates graphically at selected horizons:





As expected, quality of forecasts generally decreases with horizon. On the short-term horizons, Forecasters are quite reactive to shocks and change their estimates often. However, their long-term expectations appear to be anchored. The Forecasters consistently expect the long-term inflation to be between 1.5% and 2.0%. Forecasters appear to deviate very little from this target value on longer-term horizons. Due to this, we would like to explore the standard deviations of forecasts. There are 2 qualitatively different ways to do so on this type of data: a) find the average standard deviation between Forecasters by finding the standard deviation for each forecast date and then taking the average (Mean Standard Deviation) b) find the average standard deviation between forecasts made at different

dates by taking the standard deviation between the aggregate forecasts for each date (Standard Deviation of Mean). The former measure can tell us how much do Forecasters agree with each other on average, while the latter provides the variability of the consensus expectations over time. The degree to which Forecasters agree with each other on average can give us valuable insights on how heterogeneous the Forecasters are, and whether there are disagreements among experts. We report both measures for all horizons for which we have enough data. We also report the mean forecast for each horizon over the sample period to have a better understanding of the data.

Horizon	Mean	Mean Standard Deviation	Standard Deviation of Mean
1 rolling year	1.59	0.3	0.31
2 rolling years	1.71	0.26	0.19
2 months	1.95	0.11	1.61
5 months	1.91	0.16	1.43
8 months	1.77	0.21	1.16
11 months	1.58	0.23	0.6
14 months	1.63	0.22	0.39
17 months	1.63	0.24	0.36
20 months	1.63	0.26	0.28
23 months	1.62	0.26	0.23
26 months	1.7	0.23	0.2
29 months	1.74	0.24	0.19

Standard deviation of the mean decreases with horizon. Longer-term expectations change little over time, and the Forecasters always expect inflation to converge very close to the target level. The short-term expectations (less than 1 year horizon) follow the realised inflation more closely and vary with it, hence their higher standard deviation. The short-term expectations react to shocks, while the long-term expectations reveal Forecasters' confidence in the ability of the ECB to bring the inflation to the target

level.

As for the mean standard deviation, there is a general trend: as horizon goes up, uncertainty goes up. There are more disagreements and different opinions among Forecasters on longer-term horizons. It also appears that the standard deviation goes down for the longest horizons. This could be a coincidence or caused by attentiveness decreasing for the longest horizons, decreasing evident disagreements.

Geoff, Genre, Bowles, Friz, Meyler, & Rautanen (2007) study the errors of the aggregate SPF inflation forecasts and find that, over their sample period, the Forecasters tended to underestimate the inflation. They show that the mean error (ME), which they define to be equal to the average difference between inflation realisation and aggregate forecast, is positive for the 1 rolling year and 2 rolling years horizons. If the Forecasters are unbiased and the shocks to inflation are symmetric, then the ME is expected to be equal to 0 over a long time period. In the *Geoff, Genre, Bowles, Friz, Meyler, & Rautanen (2007)*'s work, the sample period is not long enough and they highlight that the inflation shocks seen during that period were upward and hardly predictable, which would explain the positive errors. We would like to repeat the exercise as we have a longer sample period covering several business cycles at our disposal. Special attention should be paid to the post-Covid period as it represents a period of an unprecedented upward inflation shock. Unprecedented inflation started in September 2021, which we mark as the start of the post-Covid inflation period. We measure the ME for the 1 rolling year and the 2 rolling years prediction horizons for the whole sample period, and for the pre-Covid / post-Covid periods separately. In our case, the mean error is defined as the average difference between mean forecast and inflation realisation, thus the signs of errors would be the opposite to the signs in *Geoff, Genre, Bowles, Friz, Meyler, & Rautanen (2007)*.

Horizon	Overall ME	Pre-Covid ME	Post-Covid ME
1 rolling year	-0.44	-0.07	-5.81
2 rolling years	-0.31	0.09	-5.9

Strongly negative post-Covid MEs are expected and can be fully linked

to the unprecedented upward inflation shock that happened in the 2021-2022 period. In the subsequent sections, we will explore if this shock was predictable to a degree. Overall, as the MEs are likely strongly affected by the post-Covid data points, looking at the pre-Covid MEs is preferred to get a more objective picture and remove the largest shock in the data. Forecasters appear to be biased in different directions when it comes to these two prediction horizons. Forecasters underestimate the inflation for the 1 rolling year horizon and overestimate the inflation for the 2 rolling years horizon. This may be partially explained by the properties of these horizons and by the Euro Area inflation dynamics. Mean forecast is higher for the 2 rolling years horizon than for the 1 rolling year horizon. Additionally, the standard deviation of mean is lower for the 2 rolling years prediction horizon. Before Covid, the Euro Area inflation seldom reached or exceeded the target rate of 2%. Moreover, several disinflationary periods happened (2014-2016, 2009-2010) in the sample period. For the longer-term horizon of 2 years, the Forecasters have not revised their expectations downwards enough and maintained predicting inflation convergence close to the target rate, causing slight positive bias. For the 1 rolling horizon, the Forecasters have been slightly more responsive to changes in inflation signals and deviated more from the target level, but have failed to capture some upward inflation shocks.
However, both MEs are close to 0 over this long-term period, so Forecasters' bias appears to be small, when excluding the post-Covid shock. The differences in the MEs for these 2 horizons highlight the complexity linked to inflation forecasting.

3 Forecasters' Classification

We suppose that the quality of forecasts varies from one Forecaster to another. We also suppose that the quality of forecasts is not random and is consistent for a given Forecaster over time. Thus, there should exist 'Star' Forecasters whose predictions are consistently better comparing to the aggregate predictions among Forecasters. These 'Star' Forecasters may have superior prediction capabilities due to their expertise, experience, access to information, and other relevant factors. However, it is important to note that even "Star" Forecasters are not infallible.

Identifying 'Star' Forecasters is not a straightforward task. The approach taken to identify them may vary depending on the goals of the identification process. We aim to describe and explain how we identify the 'Star' Forecasters in our case given our data properties and our objectives:

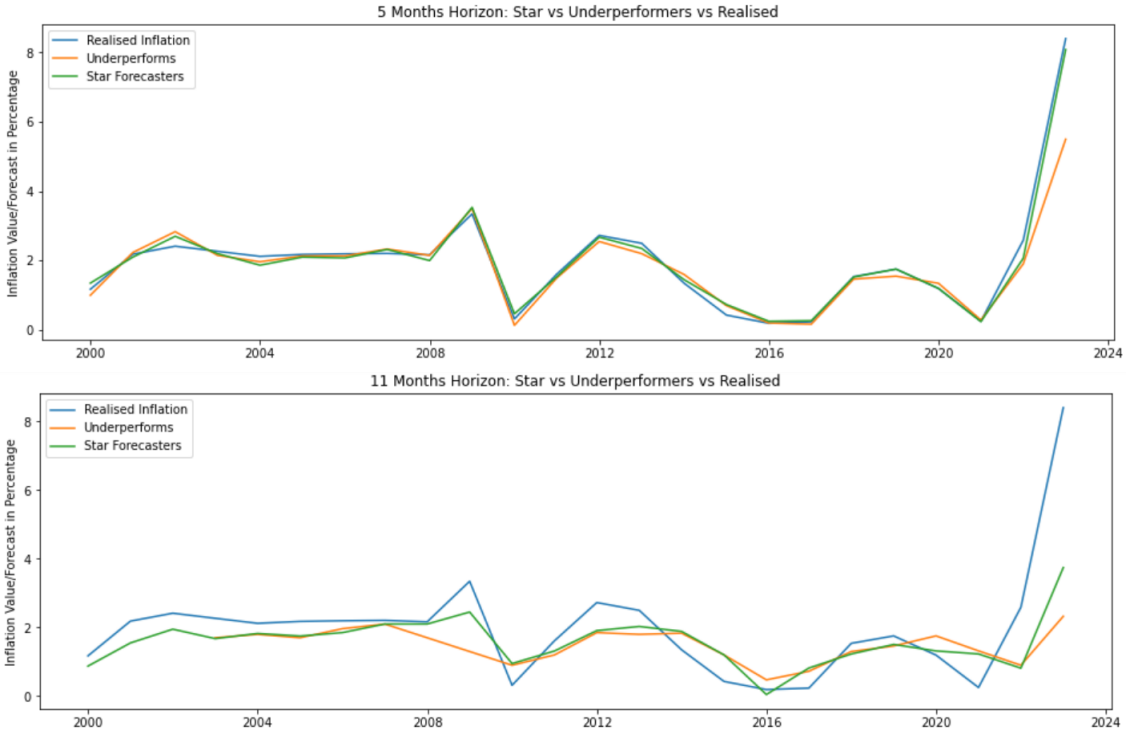
- First of all, we need to decide whether we evaluate 'Star' Forecasters across all prediction horizons simultaneously or separately for each prediction horizon. If we choose to search for the best performers across all horizons at once, we would assess their performance across different time-frames. This approach may provide a holistic view on Forecasters' prediction capabilities. However, as we know, performance is naturally better for the short-term forecasts. This would give the Forecasters with more short-term forecasts an advantage. In order to avoid this situation, we would have to come up with different weightings for different horizons, which we judge as unnecessary complexity for our task as it would make the process more opaque and could produce sub-optimal results. Moreover, by treating each horizon independently, we would recognize that Forecasters may have different strengths depending on the prediction horizon. Finally, by applying this approach, we would ensure that there are 'Star' Forecasters for each horizon, thus ensuring the comparison of the best performers with the consensus and machine learning predictions for all horizons. After considering all the advantages and disadvantages, we decide to treat each horizon independently.
- It is crucial to verify that the superior performance of the 'Star' Forecasters is consistent and is not merely attributed to chance. In order to do so, we only retain the Forecasters who have a minimum of 10 observations for a given horizon. Filtering out Forecasters with too few observations ensures that their performance is not simply attributed to randomness. Bigger sample size enables us to evaluate a Forecaster's 'true' prediction capabilities better, as the noise impact should be diminished.
- As will be discussed later in this work (specifically in the Subsection High vs Low Inflation Regimes of the Section Results), inflation paradigm has undergone a change in the post-Covid period with inflation rates soaring in the Euro Area and beyond. Particularly high inflation values were recorded in 2022. This has posed significant challenges to Forecasters. We propose that in order to be considered

a 'Star' Forecaster, the expert should have foreseen this paradigm shift to some degree and should have predicted the 2022 inflation more accurately than the majority of their counterparts. This condition would show a deeper understanding of inflation drivers and macroeconomic patterns. To ensure the condition is respected, we only retain Forecasters who have had made predictions targeting year 2022. If we do not do this, 'Star' Forecaster would not include Forecasters who have made predictions concerning 2022, as the errors associated to inflation in that year were unprecedented. Moreover, this condition helps us retain the Forecaster who are currently active and who have made predictions in the test period (year 2016 and after) that we are investigating.

- Finally, after having retained only the Forecasters that correspond to the listed conditions (at least 10 observations, forecasts targeting 2022) for a each horizon separately, the next step is to determine which among them should be designated as 'Star'. For this, an evaluation metric is necessary. Consistent with the methodology used in this work, we use the mean squared error normalised by the mean squared inflation. This metric takes into account both the magnitude of the forecast error and the corresponding inflation values. Subsequently, Forecasters are sorted according to their normalised mean squared error for each horizon. We decide to designate the top 10% of Forecasters as the 'Star' Forecasters as this percentage represents a balance between representation (large enough number of Forecasters being recognised as 'Star') and superior capabilities (meaningfully more accurate than average performance).

After the evaluation process, we have identified 34 'Star' Forecasters across all horizons. Considering that there are 111 Forecasters in total and 74 Forecasters after applying the baseline selection criteria, this number of 'Star's highlights that strengths and weaknesses of experts differ from one prediction horizon to another. Some Forecasters may exhibit a better understanding of short-term forecasting, allowing them to capture shocks. Conversely, other Forecasters may have a foresight to predict long-term trends. Furthermore, this result highlights the need of tailoring forecasting approach according to the prediction horizon.
Furthermore, Forecasters can be further divided into groups according to their competence. While it is true that we have so far focused on the top

performers, we can find the worst performers using the same methodology. We do so by simply selecting the bottom 10% Forecasters for each horizon from the pool of 74 Forecasters that meet the baseline criteria. Applying this approach leaves us with 25 Forecasters that we will call 'Underperformers'. The fact that there are more 'Star' Forecasters than 'Underperformers' implies that incompetence in forecasting is relatively consistent across horizons for the underperforming Forecasters. This incompetence can be attributed to Forecaster-specific factors such as lack of understanding of inflation dynamics and inability to take relevant information into account. These issues would prevent a Forecaster from making an accurate forecast for any time-frame.
Let us now visualise how predictions of the 'Star' Forecasters and 'Underperformers' evolved over time and compare them to the realised inflation on selected horizons:



These graphs have several implications:

- For some horizons, predictions do not exist for either 'Star' or 'Underperformers' group of Forecasts for some time periods. One explanation could be that the Forecasters that newly joined the Survey (or the ones that stopped responding to the Survey) exhibit prediction capabilities either superior or inferior to all their predecessors.
- 'Star' Forecasters and 'Underperformers' essentially capture and predict similar inflation trends, however 'Star' Forecasters are better at understanding the subtleties such as magnitude of a trend.
- As the time-frame increases, the normalised difference between 'Star's and 'Underperformers' decreases, and 'Star' Forecasters do not exhibit far superior capabilities of forecasting long-term trends. Moreover, we suggest that some cases of 'Star' Forecasters being closer to realised inflation rates may be partially attributable to chance. This highlights the inherent uncertainty associated with inflation forecasts.

4 Stickiness of Forecasters

Forecasters may not revise their expectations optimally as new information comes in, 'sticking' to their previous opinion. We would like to explore how sticky the forecasts in the Survey are. We do so by estimating Forecasters' information rigidity degree or the probability with which they acquire no information in a period.

4.1 Set Up

Here we reference the sticky information model for forecasts used by *Coibion and Gorodnichenko (2015)* and first introduced by *Mankiw and Reis (2002)*.
The model is written as follows:

$$F_t x_{t+h} = (1-\lambda) E_t x_{t+h} + \lambda F_{t-1} x_{t+h}$$

where $F_t x_{t+h}$ is the consensus forecast across agents at time t of a macroeconomic variable x at time $t+h$. $E_t x_{t+h}$ is the current rational expectation x at time $t+h$ in the full information framework. λ can be interpreted as the degree of stickiness of expectations, or information ridigity, or probability with which agents acquire no information in a period. In this framework, $\frac{1}{1-\lambda}$ is the average time duration between information sets updates.
From this set-up, Coibion and Gorodnichenko (2015) derive a relationship between the ex-post mean forecast error across agents and the ex-ante mean forecast revision:

$$x_{t+h} - F_t x_{t+h} = c + \beta(F_t x_{t+h} - F_{t-1} x_{t+h}) + error_t$$

where $\lambda = \frac{\beta}{1+\beta}$.
In particular, when $\lambda = 0$, forecasts represent rational full-information expectations, and the forecast error cannot be predicted using information at date t.
In our case, the variable of interest x is inflation. Since we have the forecasts data over several decades, we can estimate the βs, and therefore the λs across different horizons empirically. We now aim to do exactly this using our data.

4.2 Estimation of λ Across Horizons

Let us start with the 1 rolling year and 2 rolling years horizons.
Forecasts targeting inflation for the same period have a 1 year difference between them (i.e. 2-years ahead forecast in Q3 2018 and a 1-year ahead forecast in Q3 2019). We therefore can calculate the forecast revision after 1 year.
Define t as the date of the 1 rolling year forecast, $t-1$ as the date 1 year before (2 rolling years forecast date), and x_{t+1} as the targeted inflation at a period $t+1$, so in a year from the 1 rolling year forecast date. We then estimate the regression in Python:

$$x_{t+1} - F_t x_{t+1} = c + \beta(F_t x_{t+1} - F_{t-1} x_{t+1}) + error_t$$

We estimate $\hat{\beta} = 0.677$ and so $\hat{\lambda} = 0.67$ with the p-value equal to 0.002. This means that agents update their information every 8 months on average. $\hat{\lambda}$ is quite high and statistically significant. Forecasters update their information slowly and 'stick' to their previous opinions.
Let us now move forward to the calendar year derived horizons. There are significantly more horizons and forecasts revisions to explore, so we will be able to investigate some patterns in agents' stickiness.
Define h the horizon (in months) of the current forecast and Δ time difference (in months) between the current forecast and the previous forecast used to calculate forecast revision. This gives us the following regression:

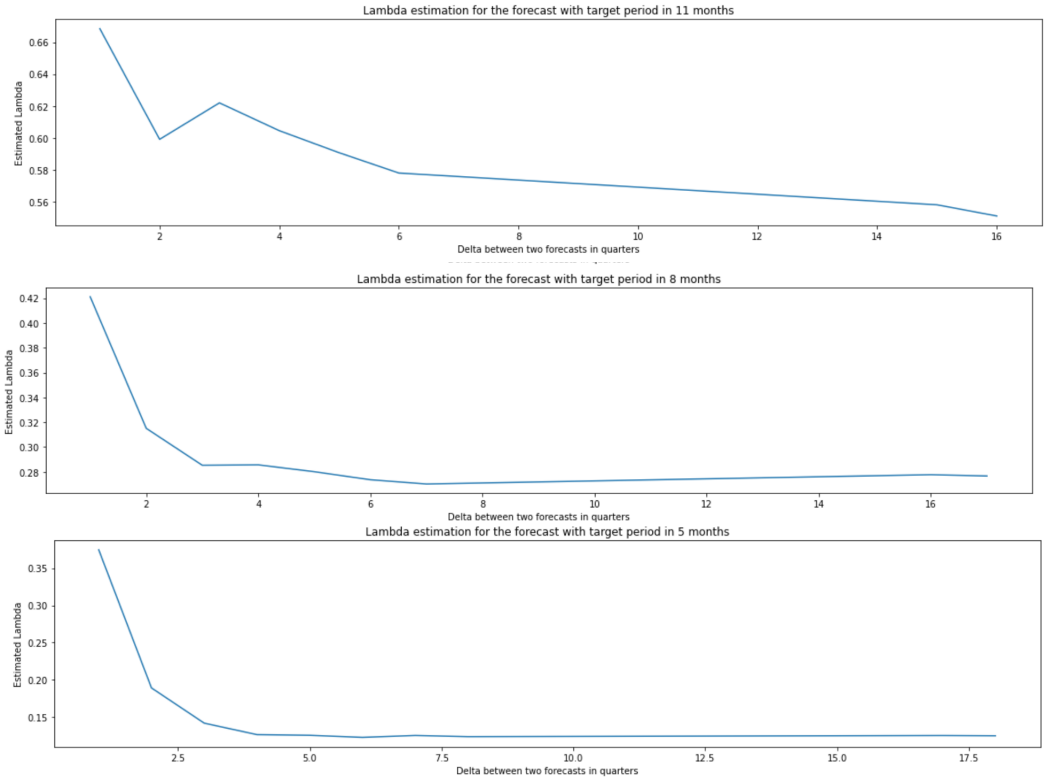
$$x_{t+h} - F_t x_{t+h} = c + \beta(F_t x_{t+h} - F_{t-\Delta} x_{t+h}) + error_t$$

We estimate this regression for different hs and Δs. Let us present findings in a table. We first report $\hat{\beta}$, then $\hat{\lambda}$, so that $0.14; 0.12$ means $\hat{\beta} = 0.14$ and $\hat{\lambda} = 0.12$. We report SI for statistically insignificant results (p-value above 0.05), NA for no data.

$\Delta \vert h$	$h=2$	$h=5$	$h=8$	$h=11$	$h=14$	$h=17$	$h=20$
$\Delta=3$	SI	0.14; 0.12	0.72; 0.42	2.01; 0.67	5.36; 0.84	SI	SI
$\Delta=6$	SI	0.14; 0.12	0.46; 0.32	1.49; 0.6	4.21; 0.81	SI	SI
$\Delta=9$	SI	0.17; 0.14	0.4; 0.29	1.65; 0.62	2.4; 0.71	SI	SI
$\Delta=12$	SI	0.14; 0.13	0.4; 0.29	1.53; 0.6	SI	SI	SI
$\Delta=15$	SI	0.14; 0.13	0.39; 0.28	1.44; 0.59	SI	SI	SI
$\Delta=18$	SI	0.14; 0.12	0.38; 0.27	1.37; 0.58	SI	SI	NA

Results start becoming statistically insignificant at a point of 14 months until the target period. For 14 months horizon, there are only 3 statistically significant results, for bigger horizons, there are none. This is due to both diminishing amount of data and properties of these forecasts.
The Forecasters always under-react to available information across all horizons, as estimated λ always has a positive sign.
The most interesting results relate to the forecasts with 5-11 months horizons. Other horizons have many / all statistically insignificant data points. For the 2 months horizon, as the target date is very close, it is likely that Forecasters are attentive and tend to use all available information. As on to why forecasts with longer horizons are not statistically significant, we propose that it is due to a big amount of noise and small sample sizes. We also notice that $\hat{\lambda}$ tends to converge as Δ increases.
Forecasts are rather sticky and get less sticky as horizon decreases. Therefore, Forecasters react better to information when the target date is close. As expected, forecasts are also the most sticky with relation to the closest previous forecast, so stickiness increases as Δ decreases. The Forecasters stick to their most recent forecast the most.

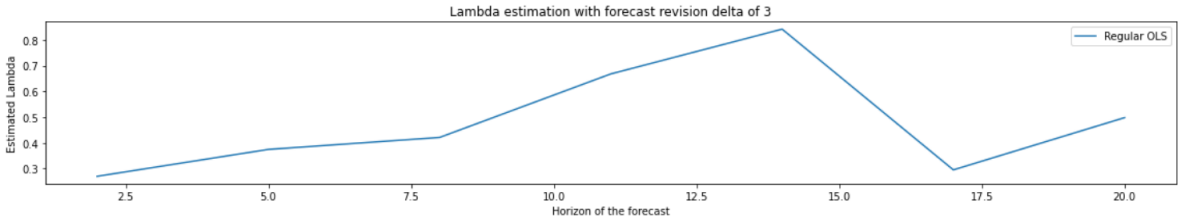
We will plot some graphs, fixing the horizon h of the reference forecast $F_t x_{t+h}$ and varying the time difference Δ between it and the previous forecast used to compute the forecast revision $F_{t-\Delta t} x_{t+h}$. Here we focus on the forecasts with 5 - 11 months horizon. Some statistically insignificant results may figure in the graphs too.

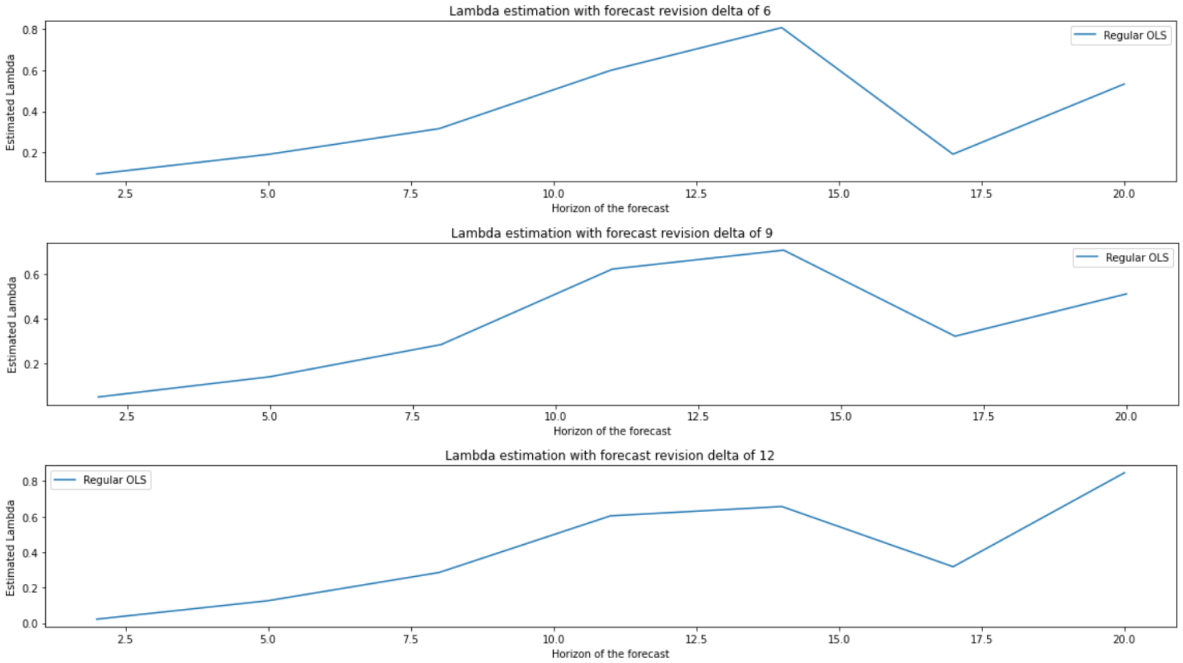


For a given horizon h, $\hat{\lambda}$ converges as Δ increases.

4.3 Dependence of $\hat{\lambda}$ on h

We plot graphs that of dependence of $\hat{\lambda}$ on h for a given Δ.





The relationship between h and $\hat{\lambda}$ is non-linear. This relationship has a similar pattern for different Δs. The overall tendency is that the Forecasters update information less for further horizons.

4.4 Comparision with the U.S. results

Coibion and Gorodnichenko (2015) also estimate λ using similar survey data in the U.S. The horizon in their data is current quarter + next three quarters (similar to our 11 months horizon). Their Δ is equal 1 quarter. They find $\hat{\lambda} = 0.55$ which is approximately the number our 11 months horizon $\hat{\lambda}$ converges as Δ increases. However, when we take $\Delta = 1$ quarter and $h = 11$, our $\hat{\lambda} = 0.67$, so this would be mean that the US experts update their information more frequently than their counterparts in Europe.

4.5 Business Cycle and Stickiness

Coibion and Gorodnichenko (2015) have also found that calm times are associated with stronger information rigidities. We proceed to check this finding in the European data. We would like to test if business cycle has any effect on the stickiness of forecasts. We do this by including a dummy variable $\delta_{recession}$ into the regression (equal to 0 in an expansion, and to 1 in a recession). The Euro Area business cycle data is taken from the FRED.

We set up the regression as follows:

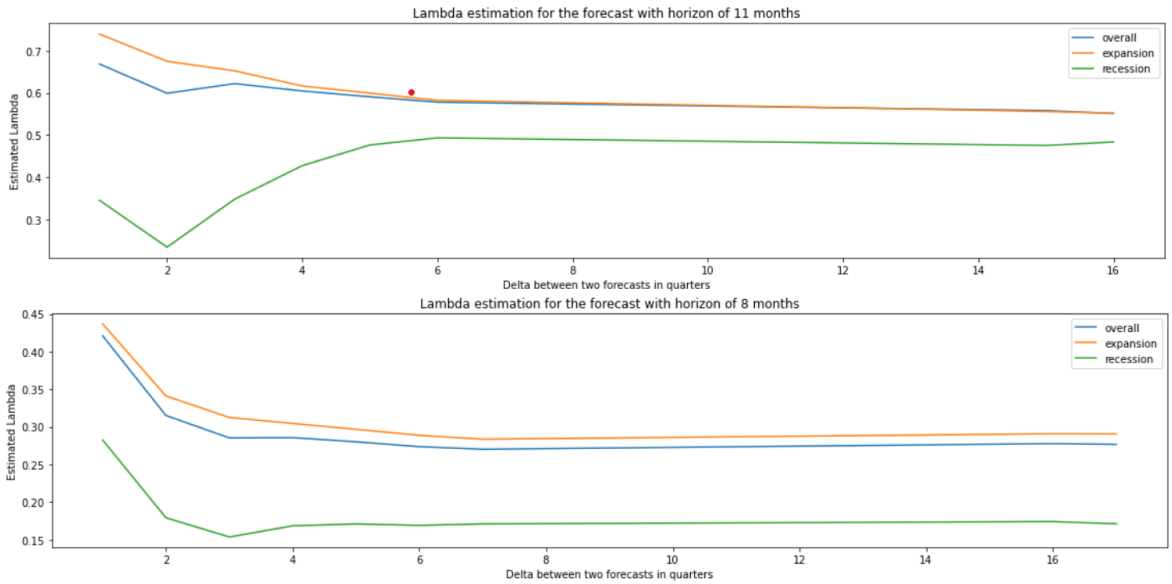
$$x_{t+1} - F_t x_{t+1} = c_1 + \beta_1(F_t x_{t+1} - F_{t-1}x_{t+1}) + \delta_{recession}c_2 + \beta_2(F_t x_{t+1} - F_{t-1}x_{t+1})_{recession}) + error_t$$

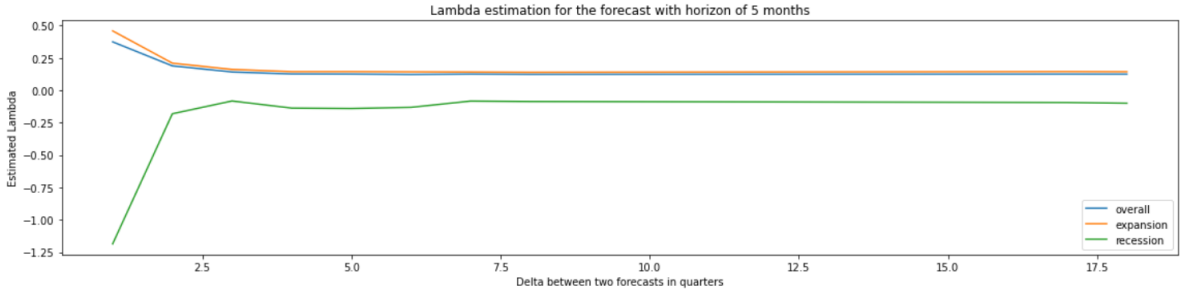
i.e. we allow β and c to vary according to the recession indicator. We treat $(F_t x_{t+1} - F_{t-1}x_{t+1})_{recession})$ as an independent variable that is equal to 0 in an expansion and to $F_t x_{t+1} - F_{t-1}x_{t+1}$ in a recession, so β_2 allows to capture additional stickiness in a recession. $\delta_{recession}$ allows us to add an intercept c_2 responsible for a recession.
We do not run regression on h and Δ pairs that were already statistically insignificant in the previous set-up.
We get from the results that c_2 is always statistically insignificant, so the intercept for recession and expansion should be statistically the same.
As for β_2, it is statistically significant for some h and Δ. We report results in the table below. NA stands for either no data or statistically insignificant in the previous general set-up. SI means β_2 is statistically insignificant, i.e. recession has no additional effect on stickiness for the given parameters.
We report values in the following order: $\hat{\beta_1}; \hat{\lambda_1}; \hat{\beta_2}; \hat{\lambda_2}$ ($\hat{\lambda_2} = \frac{\hat{\beta_1}+\hat{\beta_2}}{1+\hat{\beta_1}+\hat{\beta_2}}$)

$\Delta \vert h$	$h=5$	$h=8$	$h=11$	$h=14$
$\Delta=3$	$0.85; 0.45; -1.4; -1.19$	$0.78; 0.44; -0.38; 0.28$	$2.83; 0.74; -2.31; 0.35$	SI
$\Delta=6$	$0.27; 0.21; -0.41; -0.18$	$0.52; 0.34; -0.3; 0.18$	$2.08; 0.67; -1.77; 0.23$	SI
$\Delta=9$	$0.19; 0.16; -0.27; -0.08$	$0.45; 0.31; -0.27; 0.15$	SI	SI
$\Delta=12$	$0.17; 0.14; -0.29; -0.14$	SI	SI	SI
$\Delta=15$	$0.17; 0.14; -0.29; -0.14$	SI	SI	NA
$\Delta=18$	$0.17; 0.14; -0.28; -0.13$	SI	SI	NA

The estimated λ is always smaller in recessions than in expansions and even becomes negative when $h=5$, which means that, on average, Fore-

casters overreact to the available information.
Overall, forecasts in recession are less stickier than in expansions. This is consistent with what *Coibion and Gorodnichenko (2015)* conclude in their Section Information Rigidities over the Business Cycle. In recessions, agents react to information quicker and more importantly than otherwise.
There are less statistically significant results for recessions than in general, and that number of statistically significant results decreases as h increases. This may mean that agents update their information better in recessions only for the short-term horizons that are more worrying.
The trend of $\hat{\lambda}$ increasing with horizon h is still present. Agents incorporate less information into their forecasts for longer-term horizons in recessions as well as in expansions.
We now plot some graphs, fixing the horizon h of the forecast $F_t x_{t+h}$ and varying the time difference Δ used to compute the forecast revision. Here we focus on the forecasts with 5 - 11 months horizon. Some statistically insignificant results may figure in the graphs too for illustration purposes. We plot for recession $\hat{\lambda}_2$, expansion $\hat{\lambda}_1$, and the overall $\hat{\lambda}$.





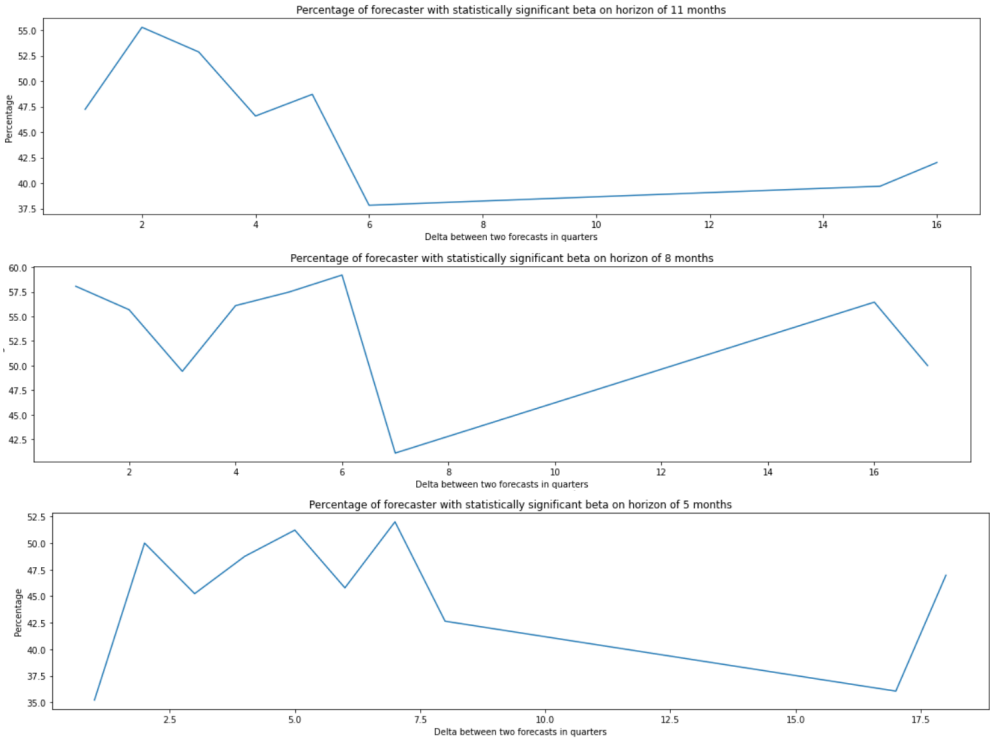
The convergence of estimated information degree is observed both for recessions and expansions. Forecasts made in recessions consistently exhibit less stickiness than those made in expansions. In recessions, Forecasters increase their attention levels and grasp the available information. They may even overreact to the information on the short-term horizons which are the most alarming.

4.6 Forecaster-specific λ

We would like to investigate whether individual Forecasters exhibit different level of stickiness. For this, we estimate the following regressions for each of the Forecasters:

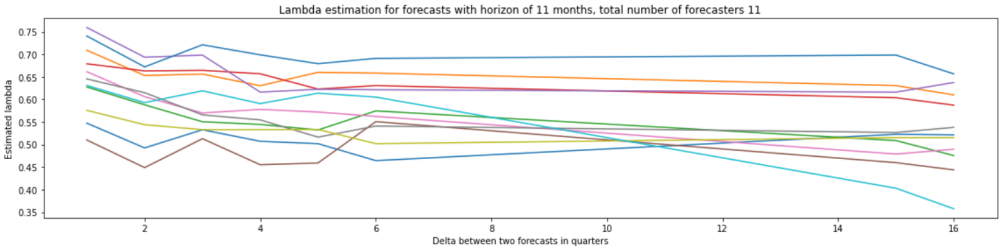
$$x_{t+h} - F_{i,t}x_{t+h} = c_i + \beta_i(F_{i,t}x_{t+h} - F_{i,t-1}x_{t+h}) + error_{ti}$$

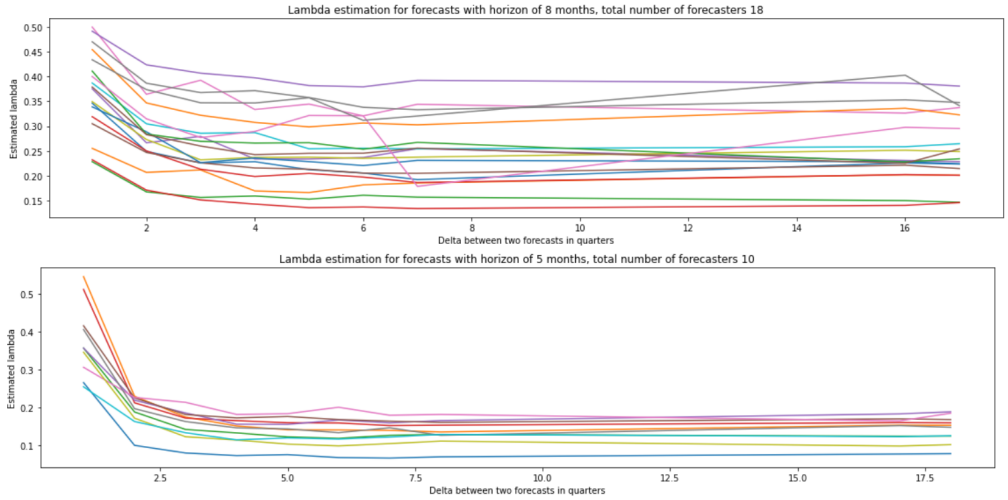
The estimation is made for the Forecasters for whom there are more than 2 observation available for a given h and Δ pair.
Some $\hat{\beta}$ are strongly statistically significant, while some are not at all. Therefore, some Forecasters are consistently showing the same level of stickiness while others either fully update their information sets or update them with varying frequency. In this case, the best way to illustrate the results is graphically.
We first investigate the issue of statistical significance. We estimate the percentage of Forecasters whose $\hat{\beta}$ is statistically significant at the 5% level, i.e. percentage of Forecasters who exhibit a consistent level of stickiness.



Statistical significance tends to be decreasing with Δ, but the effect is not strong. Moreover, the horizon effect on statistical significance is not strong either. 40% to 60% of Forecasters show statistically significant stickiness depending on the parameters.

we proceed to focusing on the Forecasters whose stickiness level is consistently statistically significant. We find the Forecasters whose $\hat{\beta}$s are statistically significant for all Δs for a given h. As a result, depending on the horizon, there are between 10 and 18 such Forecasters. We will plot their $\hat{\lambda}$ as a function of Δ for a given h.





For these Forecasters, $\hat{\lambda}$s converge very quickly. Moreover, their levels of stickiness are very correlated. Some Forecasters exhibit constant stickiness and consistently fail to incorporate new information to the same degree.

5 Machine Learning Models

5.1 Overview

We need to implement machine learning models to compare the Forecasters' performance with machine performance. Our models of choice are random forest regression and ridge regression.
Models are trained on the historical data corresponding to forecasts made from the beginning of 1999 until the end of 2015. Models are tested on the data corresponding to forecasts made from the beginning of 2016 to the end of 2022. The prediction models are implemented for some of the prediction horizons described in the Section Forecasts Data Description: 1 rolling year, 2 rolling years, 3-29 months. Only horizons for which there is enough data are used. To get a forecast, we give the models the latest information available at the forecast time. By construction, when testing the model, we get out-of-sample results.
For the rolling-year(s) horizons (i.e. 1 year ahead, 2 years ahead), the train data is taken at the monthly frequency to achieve a better accuracy of the model, and the test data is taken at the quarterly frequency as it corresponds to the forecasts' frequency. For the calendar year derived horizons

(i.e. 2 months ahead, 5 months ahead, 8 months ahead, etc) train and test data are taken at the yearly frequency as it is the only possibility.

5.2 Predictors

We need a set of predictors to be able to develop the two machine learning models. As inflation is a macroeconomic variable, we will use other macroeconomic variables to predict it. Predictor variables should either drive inflation or or be closely related to inflation and co-occur with it.
We aim to explain our predictors' choice and describe the predictor variables:

1. GDP growth is known to be closely connected to inflation. Growth in GDP leads to inflation increase as economic actors have more resources to spend, which leads to inflationary pressures. Conversely, when GDP growth is depressed, consumers tend to save more, which results in reduced demand and deflationary pressures. The Euro Area GDP growth rate same period, previous year data is taken from the FRED website as it provides the GDP data on a monthly basis, while the ECB only provides GDP growth data only on a quarterly basis.
2. Relationship between inflation and government debt may be complex. Some governments may attempt to increase inflation to dilute the government debt and its servicing costs. Moreover, a higher level of debt can lead to an increase in public spending, which may lead to inflationary pressures. On the other hand, a higher level of public debt may be associated to events such as recessions which depress the inflation in the long-run. Therefore, other variables may be necessary for a model to capture the relationship of debt with inflation. Our solution to this issue is using polynomial feature scaling in the ridge regression, allowing for non-linear relationships and for interactions between variables. The Euro area government debt as percent of GDP data is taken from the ECB website.
3. As defined by the ECB, M1 monetary aggregate is the "sum of currency in circulation and overnight deposits in the economy". When M1 increases, there is more money for economic agents to spend,

which may lead to a rise in inflation. Conversely, decrease in M1 may lead to saving behaviours, decreased demand, and lower prices. M1 has a direct influence on consumer spending. The Euro area M1 monetary aggregate is measured as an index and the data is taken from the ECB website.

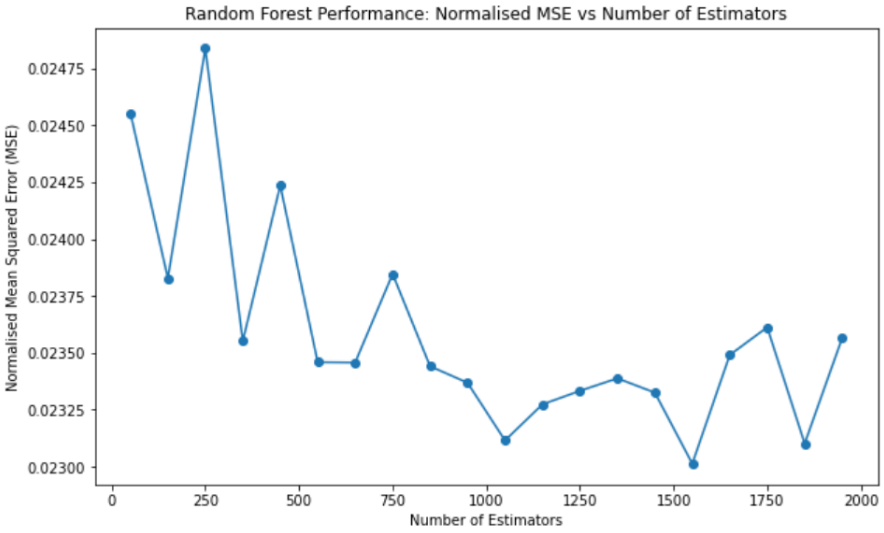
4. M3 is the broadest monetary aggregate. As defined by the ECB, It contains "financial instruments that closely resemble deposits, such as repurchase agreements, shares issued by money market funds and short-term debt securities issued by banks" and M2 which itself consists of M1 and "deposits with an agreed maturity of up to two years, or those redeemable at notice with a notice period of up to three month". It is related to inflation in a similar, even though less direct way than M1. Its impact on inflation may be more long-term through indirect channels such as investment. The Euro area M3 monetary aggregate is measured as an index and the data is taken from the ECB website.
5. Increases in production prices may be passed on to consumers, generating inflation. It is important to monitor producer prices to anticipate inflation. The effect of producer prices on inflation should be more pronounced in the short-term as it may be considered as a supply-side shock. We use the Euro area producer price index (PPI), domestic sales, year-on-year percent change data from the ECB website to represent this effect.

Some predictors, in particular, the PPI, are more important in the short-to-medium-term forecasting than in the long-term as they may represent shocks. Hence, we only use the PPI for predicting inflation on horizons smaller than 2 years.
Moreover, as will be discussed later, the sample size for the calendar year derived horizons (i.e. 2-29 months horizons) is small as it only contains yearly data-points, thus for these horizons model training is more challenging. To decrease the risk of overfit, for the 2-29 months horizons, M1 and M3 should not be used as predictors at the same time as these features account for similar trends. M1 is used for shorter-term horizons, M3 is used for longer-term horizons. For the 1 and 2 rolling years horizons, both M1 and M3 are used.

5.3 Random Forest Regression

Just as *van Binsbergen, Han, and Lopez-Lira (2022)* use the random forest regression to predict their variable of interest (corporate earnings), we use the random forest regression to predict inflation. Random forest regression is a non-linear supervised ensemble method. It consists of multiple decision tree regressors, which allows for improved accuracy and stability of predictions comparing to individual decision trees.
A decision tree regressor has a hierarchical structure. It recursively partitions the input data into subsets based on features. Each decision node in the tree represents a decision based on a feature, while each leaf node represents a predicted value of the target variable. The average value of variables on a leaf node represents the resulting prediction. The threshold used to partition the data is chosen so that it minimises the the expected mean squared error of the resulting prediction over all subsequent nodes. A new data point follows the path from the root node to a leaf node, making decisions based on its feature values. The leaf node reached provides the prediction equal to the average outcome for that leaf. In decision trees, the depth represents the maximum number of splits in the tree. Depth is the length of the longest path from the root node to a leaf node. A greater depth allows the tree to capture more complex relationships in the data, but it also increases the risk of overfit. The maximum number of features parameter helps to control the tree complexity by limiting the number of features considered at each decision node. By setting a maximum number of features, the decision tree algorithm selects the best feature subset among the available options at each node, and so reduces noise by selectings the most imporant features. Popular values for the maximum number of features parameter are 'sqrt' (square root of the total number of features) and 'log2' (base-2 logarithm of the total number of features). As decision trees may exhibit high variance and overfitting due to the risk of fitting too specific patterns or noise in the data, ensemble methods such as random forest are preferred to get more stable and accurate results.
Random forest regressors are an ensemble of decision tree regressors. For learning of each decision tree, a sub-sample is randomly drawn from the sample with replacement, thus noise impact is reduced and generalisation is improved. Moreover, for each decision tree, a random subset of features is selected. This helps limit correlation among the trees and prevents any single feature from dominating the predictions. The final outcome of a

random forest is the average outcome of decision trees it consists of. The number of decision trees in a random forest can be configured. Higher the number of decision trees, higher the performance is expected to be as the impact of biased or noisy trees is reduced. However, when the number of decision trees is high enough, a plateau in performance is expected and accuracy improvements from adding more decision trees are only marginal. Below is an example of the evolution of the normalised mean squared error on train data as a function of number of estimators keeping all other parameters constant:

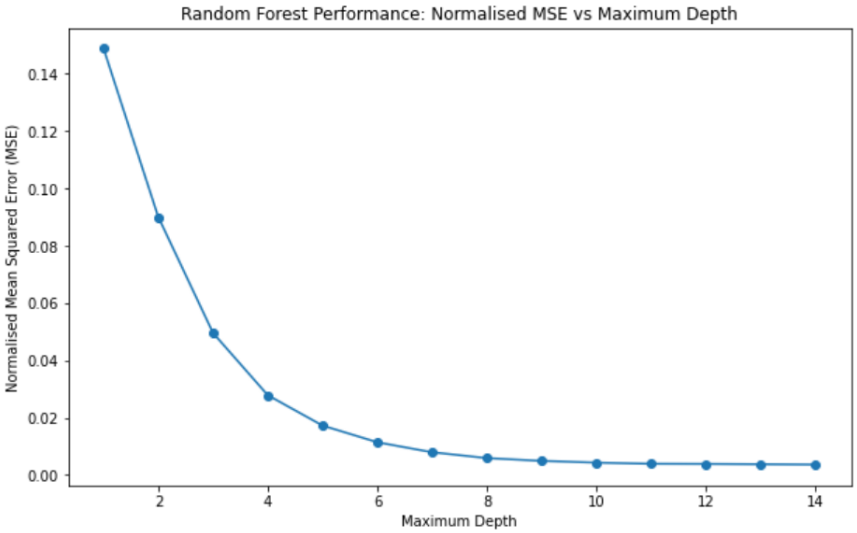


A plateau is reached after about $n = 1000$ estimators. We choose this number as optimal for our task from the performance and computational point of views.

When it comes to the choosing the hyper-parameters such as maximum number of features or their maximal depth, the optimal parameters depend on data and its complexity. In our case, we find the optimal hyper-parameters using cross-validation which is a re-sampling technique for hyper-parameter tuning. It involves dividing the data into subsets and iteratively evaluating the model's performance on different combinations of these subsets. Cross-validation provides an unbiased estimate of the model's generalization ability for various hyper-parameters. The hyper-parameter values associated with the best performance are chosen. The hyper-parameter tuning is performed on the train data.

As a result of hyper-parameter tuning, we get that the 'sqrt' maximum

number of features is optimal. We also look for the optimal tree depth. We include an example of evolution of the normalised mean squared error on train data number of estimators equal to 1000 and maximum number of features equal to 'sqrt', the horizon of prediction is 1 rolling year:



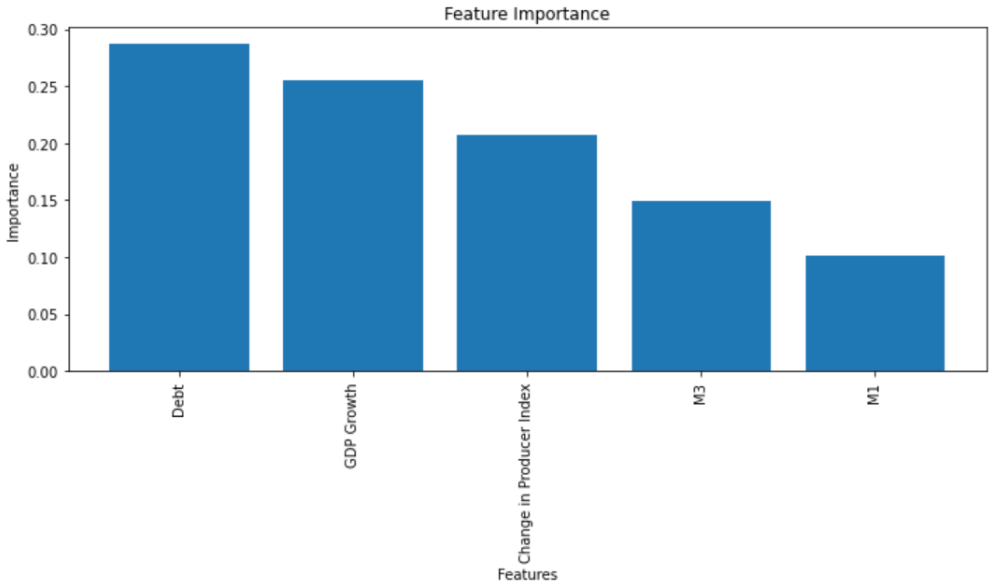
We noticed that the train MSE first decreases exponentially with maximum depth, then reaches a plateau after maximum depth equal to 6. From the graph, we suspect that the model we use is prone to overfitting. Moreover, as will be discussed later, there has been a change in inflationary regime post-Covid, so we prefer a model that did not fully fit to the previous inflation regime. Therefore, even though the optimal maximum depth suggested by the cross-validation is equal to 7-10 depending on the horizon used, we will use the maximum depth equal to 6 for all horizons to have a less complex model.

Mean squared error normalised by squared inflation on the train data (0.6) is significantly higher the normalised MSE on the train data (0.01). This indeed confirms that the model is prone to overfitting.

Finally, we would like to report the feature importance in the model. Feature importance is a metric of the relative significance of each feature. The importance is computed by measuring the decrease in model performance when a particular feature is randomly shuffled, causing its original relationship with the target variable to be disrupted. The greater the decrease in performance, the more important the feature is considered.

The feature importance for the 1 rolling year prediction using the chosen

parameters is reported below:



Feature importance is balanced and none of the features is dominating others, which should help to avoid overfitting. However, the model still exhibits an overfit which is evident from comparing the train and the test errors. The next method we use aims at tackling this issue.

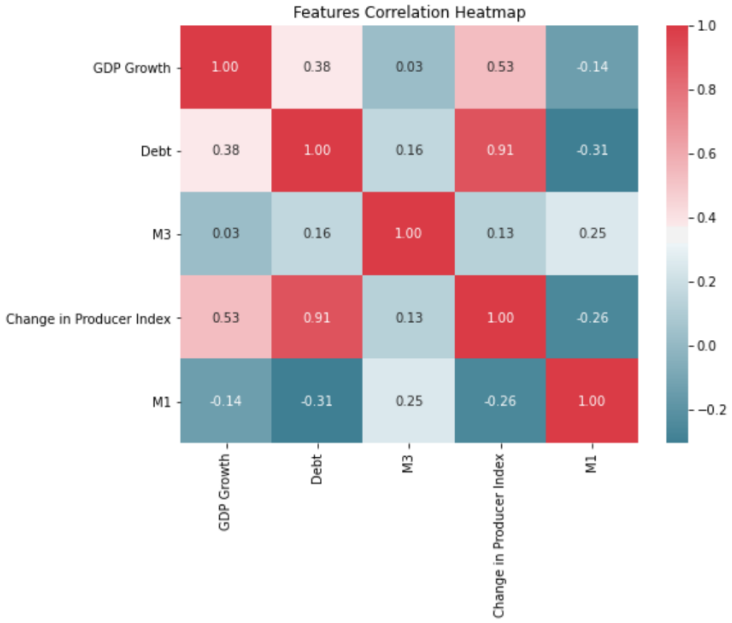
5.4 Ridge Regression

The second method we use for predicting inflation is the ridge regression with the polynomial scaling of features for some horizons.
Ridge regression is a regularised regression method used when there is multicollinearity among predictors. It extends ordinary least squares regression by introducing a regularization term to the objective function.
The method adds a penalty (regularisation) term, controlled by the hyperparameter λ, to the OLS objective function. Increasing λ shrinks the regression coefficients towards zero and so reduces their magnitude. Thus, the ridge regression model reduces the influence of highly correlated predictors and keeps coefficient values balanced without any feature dominating others.
Mathematically, ridge regression minimizes objective function which is the sum of squared errors (OLS) plus the L2 penalty term proportional to the square of the coefficients (regularisation). This penalty term balances

the fit and the complexity of the model. The ridge regression objective function is written as:

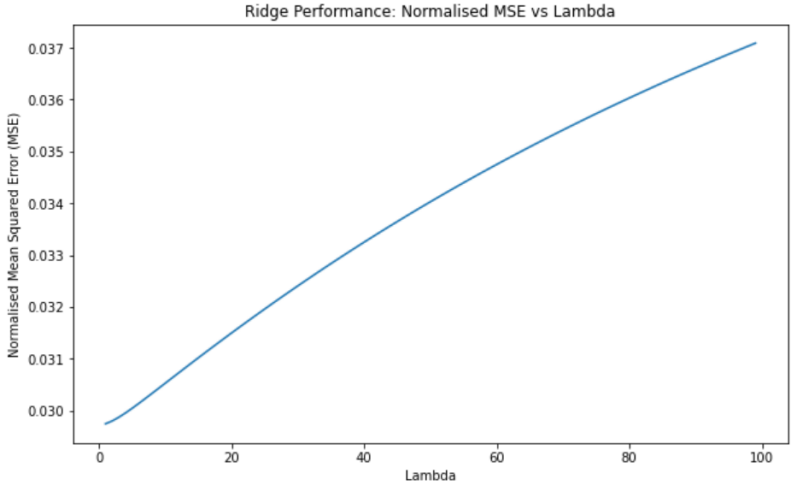
$$(y - X\beta)^T(y - X\beta) + \lambda\beta^T\beta$$

where y is the vector of target variables, X is the matrix of features, and β is the vector of coefficients.
As we have observed in the Subsection Random Forest Regression, the random forest model exhibited an overfit. This can be partially explained by the fact that our predictors, macroeconomic variables are correlated, which can be confirmed by this correlation heatmap:



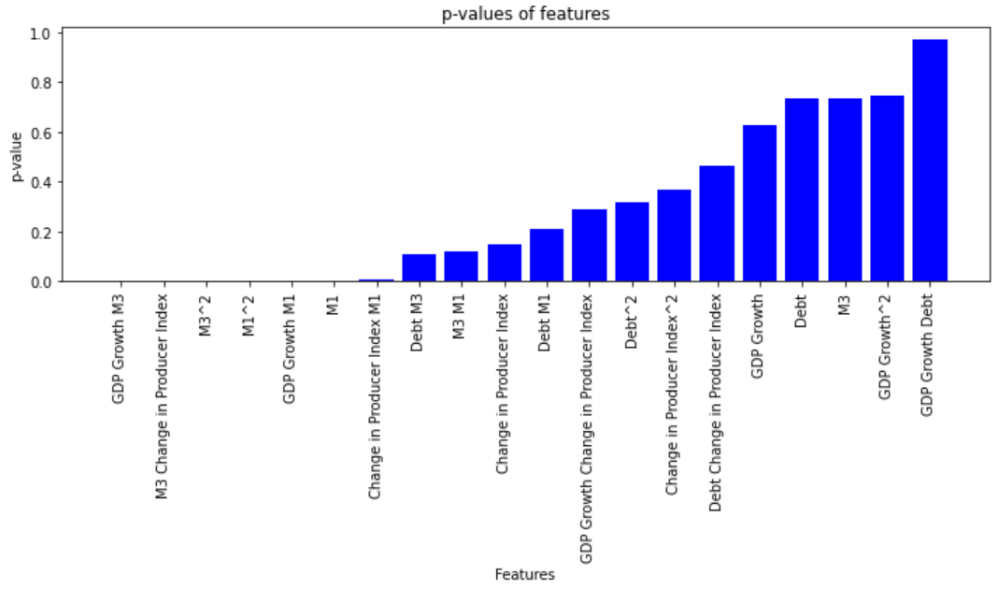
As our features are correlated, the ridge regression is a more suitable choice in our case, as it prevents overfit and is able to handle multicollinearity.
We use the polynomial feature scaling for some horizons. This choice is based on our understanding that the relationship between predictors and inflation is likely non-linear and is dependent on interaction between predictors. Polynomial feature transform the original predictors into higher-order polynomial terms. By incorporating polynomial terms, we allow for more complex interactions among predictors and inflation, enhancing the model's ability to capture the subtleties of inflation dynamics. We use the maximum polynomial degree of 2, as it represents the optimal trade-off

between model fit and complexity. With a degree higher than 2, the model will become too specific to the train data and it may fit noise in the data. The degree of 2 allows us to capture curvature and interactions in the data, making it the optimal choice.
We only use the polynomial features scaling on 2 horizons: 1 rolling year and 2 rolling years ahead predictions due to sample size issues. For these 2 horizons, train data is available monthly over a long period of time spanning predictions made from 1999 to 2015. Even though there are 20 features after polynomial scaling, for these 2 horizons the degree of freedom remains high, allowing for the optimal fit and capturing of patterns, not the noise. However, in case of the calendar year derived horizons (2-29 months), as the data can only be taken yearly, the sample size is from 15 to 17 data points, making the degree of freedom negative or close to 0. In general, a small degree of freedom is detrimental for regression quality for several reasons. It increases the risk of overfit. With a small amount of data, adding more coefficients can lead to a more complex model that is prone to fitting noise. Estimating a larger number of coefficients requires a larger sample size to achieve reliable and statistically significant results. With few data points, the estimates may have high variability. We thus decide against using polynomial scaling for calendar year derived horizons as it represents unnecessary complexity and the risk of overfit in the case of limited data.
The hyper-parameter λ allows for control over the trade-off between model fit and complexity. We present the mean squared error normalised by the squared inflation on the train data evolvution with lambda for the 1 rolling year horizon:



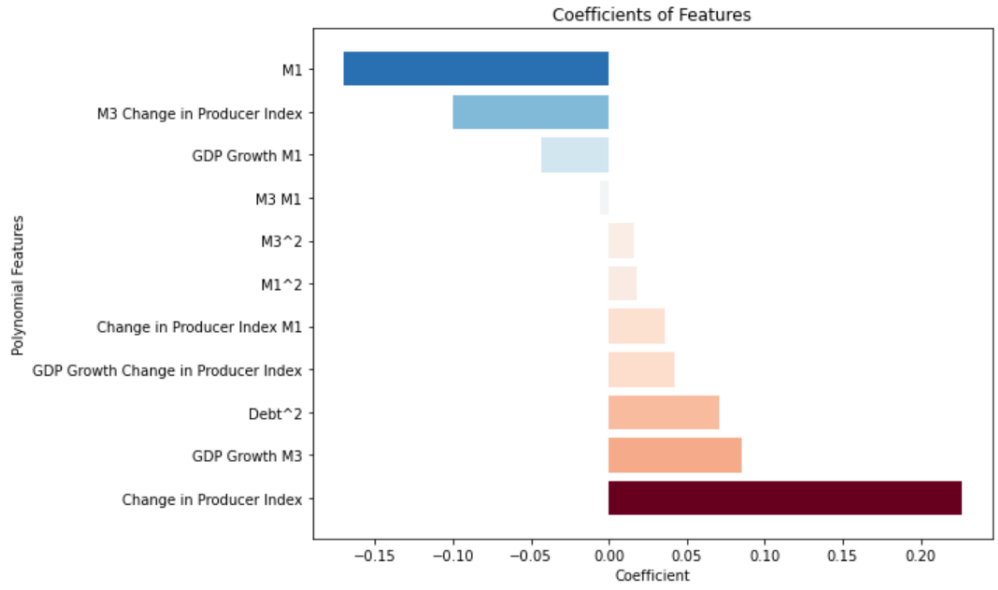
As expected, the train MSE is going up with λ. As before, the optimal λ can be chosen using cross-validation which tackles the overfit risk on the train data. The optimal λ varies from one horizon to another. It is expected to be higher in case of more complex features (polynomial) to deal with complexity and multicollinearity in the data.

It is important to consider statistical significance of the variables. For the 1 rolling year horizon, 65% of the features are statisically insignificant at the 5% level. We report the p-values of all the scaled features for the 1 rolling year horizon:



To improve model interpretability, model fit, and decrease the risk of overfit, we remove statistically insignificant variables one by one in descending order of p-value until all of the remaining features are statistically significant at the 5% level. As a result, we are left with 11 features from 20 initial features for the 1 rolling year horizon. This process is only executed for the 1 and 2 rolling year horizons, as these 2 horizons are the only ones with sample size big enough to ensure robust statistical significance estimates. For the 2-29 months prediction horizons, the variability is too high and we instead rely on economic intuition we have described in the Subsection Predictors to select features.

A clear advantage of the ridge regression is its interpretability. Each feature has a corresponding coefficient, which serves as a quantitative measure of the feature's influence on the final prediction. This analysis of the coefficients allows for a understanding of the relative impact of each feature. It can provide insights on drivers of inflation. We present the coefficients associated with the features used for the 1 rolling year horizon after the selection process:



*M*1 and *Change in Producer index* are the only two first-order individual features left. As we have hypothesized before, inflation is driven by complex non-linear factors and interactions. We note that the features left

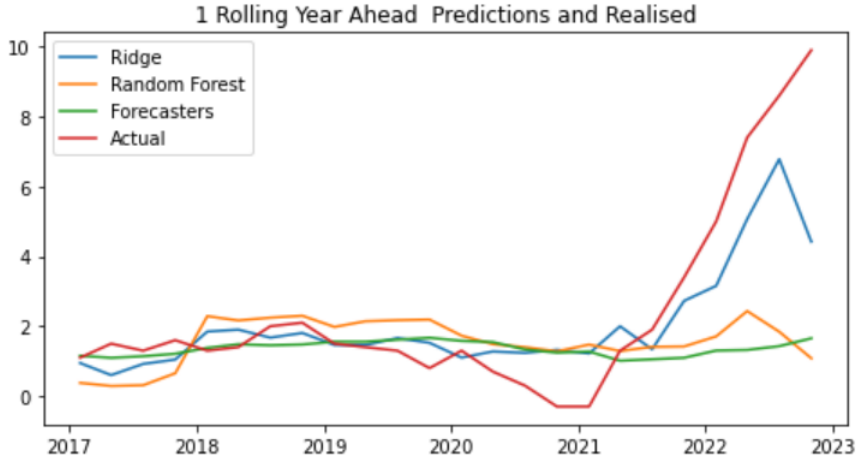
out would be different for the 2 rolling year horizon depending on the p-values of the coefficients. Indeed, some drivers of inflation are more important in the short-term, and some are more important in the long-term.

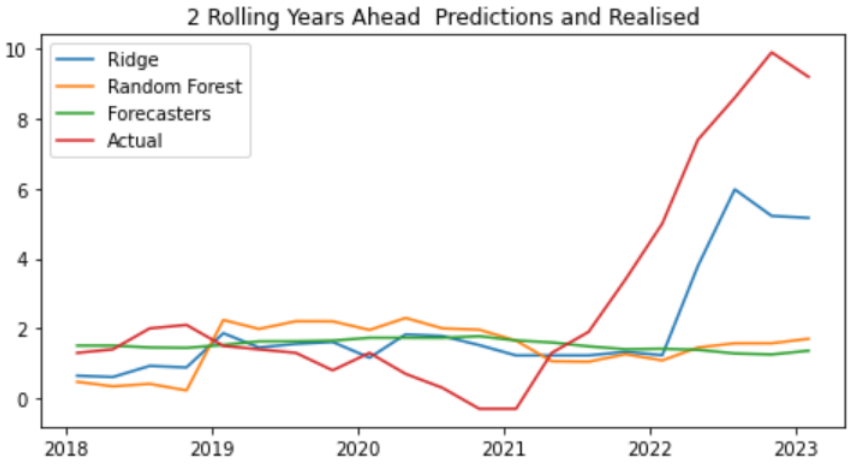
This model still exhibits an overfit with test $R^2 = 0.69$ and the train $R^2 = 0.84$. Increasing the λ would decrease the overfit and bring the R^2 values closer to each other. We avoid doing so as this would break the rules of testing a model out-of-sample. Nevertheless, the ridge regression provides us with a reasonable fit and, as will be discussed in the Section Results, the ridge regression performs significantly better than the random forest regression.

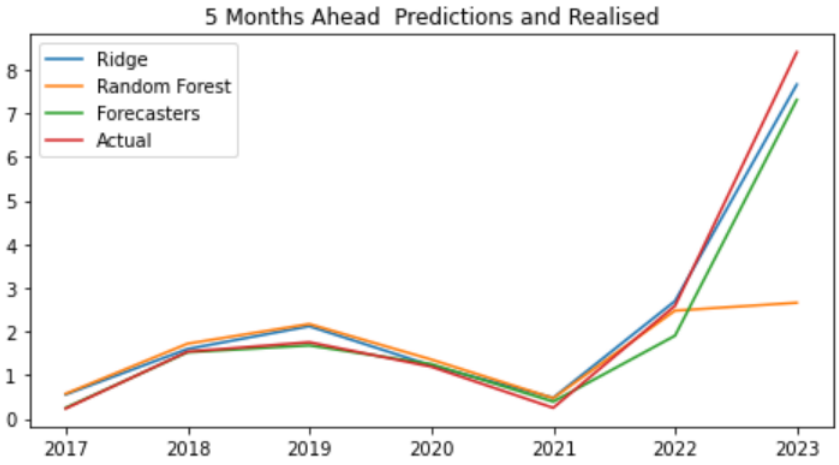
6 Results

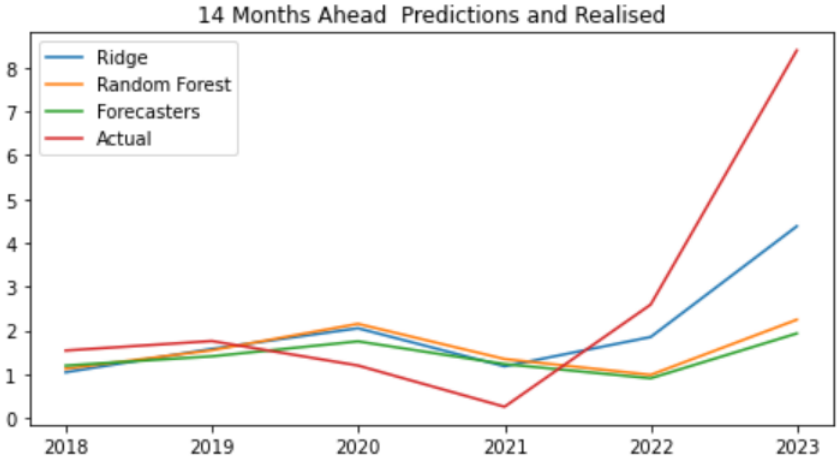
6.1 General Case

Firstly, we present some plots depicting Forecasters' consensus predictions, random forest predictions, ridge predictions, and realised inflation over time for selected horizons to illustrate our models' performance.




2 Rolling Years Ahead Predictions and Realised
Ridge
Random Forest
Forecasters
Actual
10
8
6
4
2
0
2018
2019
2020
2021
2022
2023


5 Months Ahead Predictions and Realised
Ridge
Random Forest
Forecasters
Actual
8
7
6
5
4
3
2
1
0
2017
2018
2019
2020
2021
2022
2023


14 Months Ahead Predictions and Realised
Ridge
Random Forest
Forecasters
Actual
8
7
6
5
4
3
2
1
0
2018
2019
2020
2021
2022
2023

Ridge regression generally produces better results than random forest regression. As expected, predictions' and forecasts' quality decreases with horizon. Moreover, from these graphs, it is not always clear whether humans or machines perform better, and more information is needed. However, it is obvious from the graphs, that during the post-Covid period, machines produced strikingly better predictions than humans did. Machines better captured the new trend of soaring inflation rates. It may be connected to the fact that humans do not update their information enough as discussed in the Section Stickiness of Forecasters or to their bias which will be explored later. In order to better understand the properties of machine and human forecasts, a study of the realised errors across all horizons is needed.

We now report the mean squared error normalised by the squared inflation for various horizons for Forecasters, ridge predictions, and random forest predictions over the test period (predictions made from 2016 to 2022). We also report these values for the 'Star' and 'Underperformers' ('Under') categories of Forecasters to compare the machine with the best and the worst performers among experts.

Horizon	Forecasters	'Star'	'Under'	Random Forest	Ridge
1 rolling year	0.63	0.53	0.69	0.6	0.18
2 rolling years	0.69	0.59	0.7	0.69	0.24
2 months	0.003	0.001	0.007	0.43	0.04
5 months	0.03	0.005	0.12	0.4	0.01
8 months	0.1	0.03	0.27	0.39	0.02
11 months	0.4	0.32	0.5	0.45	0.13
14 months	0.55	0.48	0.58	0.51	0.3
17 months	0.62	0.58	0.66	0.48	0.45
20 months	0.65	0.6	0.67	0.51	0.44
23 months	0.65	0.61	0.7	0.64	0.61
26 months	0.68	0.63	0.72	0.65	0.53
29 months	0.67	0.59	0.68	0.7	0.79

Forecasters generally perform worse than the Ridge regression. For the

short-term and medium-term horizons, Forecasters also lose precision at a faster rate than the machine. It is consistent with our hypothesis that Forecasters update their information sets less for forecasts with further horizons and that their stickiness grows with horizon.

However, the table reveals that the Forecasters perform significantly better than both machine learning models on the 2-month horizon. This discrepancy can be attributed to the nature of the forecast for this particular horizon. The forecast is constructed as a calendar year forecast, typically made Quarter 4 of each year. Human forecasters have access to approximately 10 out of 12 months' worth of data that will contribute to the inflation rate for that year. However, effectively communicating such data to a machine learning model poses difficulties. Thus, given the attentiveness of human forecasters for the 2 months horizon, that we observed in the Section Stickiness of Forecasters, it is expected that Forecasters outperform the machine learning models on the 2 months prediction horizon.

When analyzing the performance of the ridge model for different prediction horizons, it is observed that the model shows a significantly worse performance for the 17-29 months horizons comparing to the 2 rolling years horizon, even though the time-frames for these horizons are similar. Moreover, performance of the ridge model tends to converge closer to the performance of the random forest model on these horizons. We connect it to the fact that the data for the calendar year derived horizons is limited, and as the horizon increases, too much noise goes into the model, which fails to make fully relevant predictions. Therefore, the fact that human Forecasters outperform both models on the longest (29 months) horizon is also expected, as the data for the models to learn was limited and not enough for the models to distinguish noise from patterns on such a long horizon.

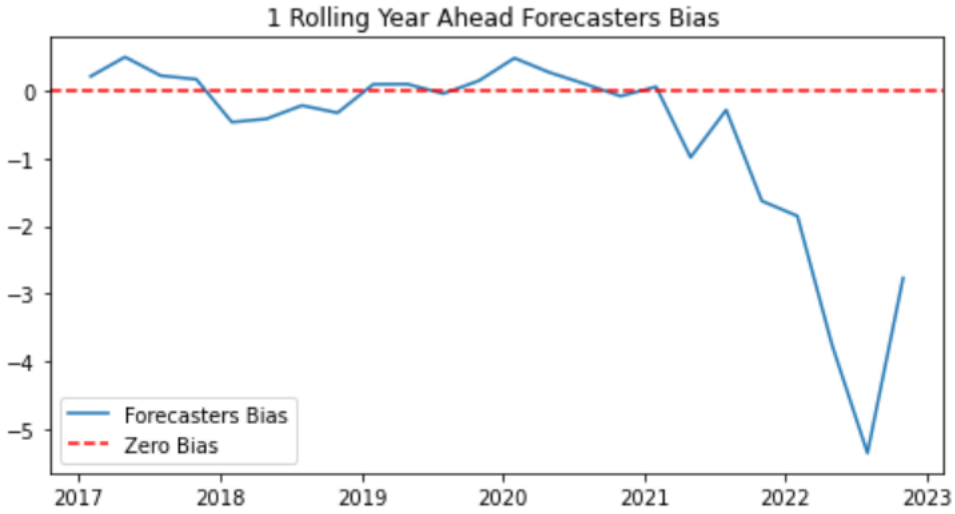
As we have expected, the ridge model beats the random forest on all but one horizon (29 months) by a significant margin. This is likely due to the ridge model being more adjustable and less prone to overfitting. In our test period, there has been a shift in inflation paradigm which we will discuss in the next Subsection. From the table, it is clear that the ridge model is more suitable for our task, and so we will use it as the benchmark machine model to compare with the Forecasters.

We consider the result obtained for the 'Star' and 'Underperformers' groups of Forecasters very valuable as they provide insights about differences between Forecasters. On the 5-8 months horizons and the 23-26 months horizons, normalised MSEs of the ridge model and the 'Star' Forecasters are of similar order. We hypothesize that the similar performance for the 23-26 months horizons may be explained not so much by the superior capabilities of the 'Star' Forecasters on these horizons, but rather by limitations of the ridge regression on these horizons that we have just discussed. Indeed, when comparing the normalised MSEs of the 'Star' Forecasters and the ridge model for the 2 rolling year horizon (similar time-frame to the 23-26 months horizons), the ridge model significantly outperforms the top experts among Forecasters. However, for the 5-8 months prediction horizons, such inference cannot be made. Thus, indeed, the 'Star' Forecasters' performance is on par with the ridge regression performance for the short-term horizons. Therefore, the 'Star' Forecasters make close to rational predictions on the short-term horizons. However, their performance still decreases on longer-term horizons, in line with our hypothesis that Forecasters exhibit more stickiness for further horizons. Therefore, even the best Forecasters exhibit stickiness even if to a smaller degree than the general pool of experts and do not update their information sets often enough. As for the 'Underperformers', quality of their forecasts decreases at a more rapid rate a horizon grows. We thus hypothesize that the 'Underperformers' do not update their information and exhibit a significant degree of stickiness for all, but the smallest horizon (2 months)

The results obtained for the 1 and 2 rolling years prediction horizons, for which the training data was the most extensive represent an essential aspect to emphasize. The ridge model has had the opportunity to learn from a more comprehensive dataset than for the calendar year derived horizons, resulting in a higher quality of predictions. Therefore, the performance of the model in these horizons may reflect the true potential of machine learning in forecasting macroeconomic variables.
Ridge regression produces significantly more accurate predictions than the Forecasters on these two horizons. As discussed earlier, one contributing factor in this accuracy difference is that the Forecasters update their information sets less frequently on longer-term horizons. We also attribute

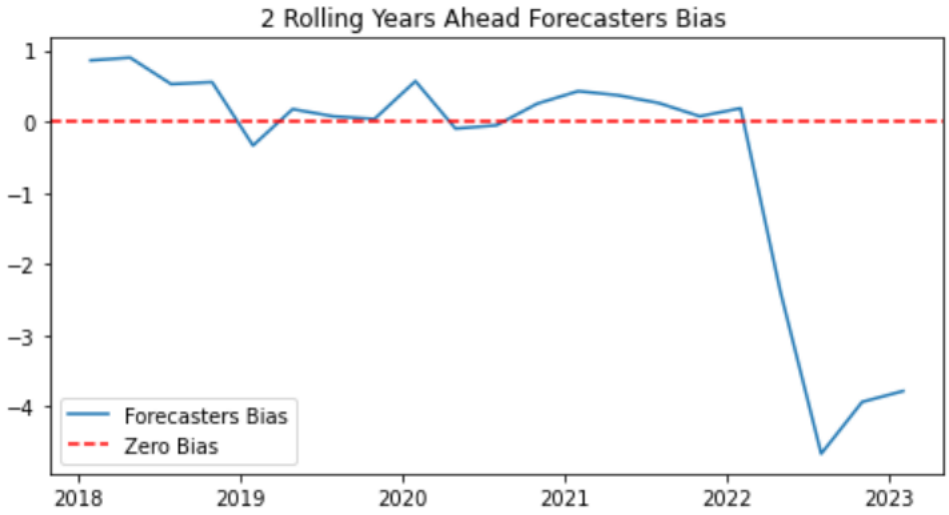
the ridge's success to it being the right model for this set-up when enough data is available due to it being able to balance between complexity and model fit.

Given the superior performance of the ridge regression model on the 1 and 2 year horizons, we will assume that its predictions are rational expectations of inflation at the time the forecasts are made. This assumption provides a basis for exploring any biases present in the human forecasters' predictions relative to this rational forecast. We will examine the differences between the human forecasts and the ridge regression predictions to find the potential biases of human forecasters.



The graph above represents the $F_t x_{t+1} - E_t x_{t+1}$, the difference between the Forecasters' average forecast and the ridge prediction for the 1 rolling year prediction horizon. This bias is calculated over the test period, including forecasts made between 2016 and 2021, with inflation realized in 2017-2022.

From this graph, the Forecasters have exhibited overly optimistic forecasts, predicting lower inflation than the rational benchmark provides during the post-Covid period. However, it may in fact be connected not to the inherent optimism of forecasters, but to the post-Covid shift of paradigm in inflation, which they failed to adapt to. Indeed, before 2021 the bias was more balanced.



The graph above represents the $F_t x_{t+2} - E_t x_{t+2}$, the difference between the Forecasters' average forecast and the ridge prediction for the 2 rolling years prediction horizon, calculated over the test period. This graph is similar to the 1 rolling year prediction bias graph, leading us to the same questions as those outlined above.

To conclude this Subsection, we present the mean error and the mean squared error of the Forecasters relative to the rational (ridge) predictions, both normalised (by the inflation and the squared inflation, respectively), for the 1 and 2 rolling years prediction horizons and over the test period.

Horizon	Normalised Mean Error	Normalised Mean Squared Error
1 rolling year	-0.28	0.2
2 rolling years	-0.17	0.16

In the next subsection we will explore whether the bias seen is caused by optimism, stickiness, or a combination of both.

6.2 High vs Low Inflation Regimes

There has been a significant shift in paradigm post-Covid, with inflation rising unprecedentedly. This sudden change in the inflation dynamics may have posed a challenge to both humans forecasts and machines predictions. Our test data includes the post-Covid period which we aim to analyse and compare with the pre-Covid period.

Ridge regression generally performs significantly better than the random forest regression. This can be attributed to the properties of the ridge regression and to the shift in inflation paradigm. Ridge regression, in contrast to the random forest regression, is more adaptable, and less prone to overfit. It can handle scenarios with extreme values as it remembers the magnitude of influence of each feature and will apply them to each value seen, even if extreme, thus generalising the influence of features. Moreover, by applying regularization, ridge regression prevents the model from relying too heavily on specific data points, leading to a more balanced representation of the underlying patterns. On the other hand, random forest lacks this extrapolation capability. This limitation arises because the individual decision trees in the random forest make predictions based on the average of the training samples on a leaf node. If there are no unusual values in train data, the model may not be able to accurately predict them, even if the pattern remains the same, just with more extreme values. Therefore, the better performance of ridge relative to the random forest may be partially explained by the shift in inflation paradigm it has better adapted to.

Moreover, as we have discussed that the Forecasters tend to be inattentive, we hypothesise that they may have failed to adapt to the changing paradigm. A large part of the ridge model's advantage over the Forecasters may be attributed to inattentiveness of the Forecasters. If this is the case, then we should see more balanced difference in performance between the machine and the Forecasters in the pre-Covid period. If it is not the case, then the difference in performance should remain as pronounced in the pre-Covid period and it may be attributed not only to information rigidity, but also to the Forecasters' inherent bias such as excessive optimism or excessive pessimism.

We define the post-Covid period to have started in September 2021, as it was when the inflation rate exceeded 3%. The corresponding 1 rolling year horizon forecasts started in September 2020, and the 2 rolling years horizon forecasts in September 2019, etc. We first report the normalised mean squared error in the **low inflation regime** for all the Forecasters, the 'Star' Forecasters, the 'Underperformers' ('Under') Forecasters, the random forest regression, and the ridge regression for the same horizons as before:

Horizon	Forecasters	'Star'	'Under'	Random Forest	Ridge
1 rolling year	0.34	0.33	0.58	0.49	0.28
2 rolling years	0.58	0.18	0.69	1.04	0.54
2 months	0.004	0.001	0.01	0.19	0.1
5 months	0.02	0.003	0.02	0.06	0.04
8 months	0.05	0.06	0.05	0.08	0.09
11 months	0.27	0.24	0.31	0.47	0.36
14 months	0.23	0.22	0.18	0.33	0.31
17 months	0.25	0.21	0.28	0.5	0.39
20 months	0.3	0.35	0.31	0.61	0.47
23 months	0.33	0.35	0.36	0.55	0.42
26 months	0.58	0.84	0.66	0.78	0.55
29 months	0.57	0.96	0.69	1.02	0.8

The results present a different picture than that presented by overall MSEs. During the pre-Covid period, ridge regression outperformed the Forecasters on the 1 rolling year and the 2 rolling years horizons. However, the results for the 1 rolling year horizon are more impressive, as the ridge model outperformed 'Star' Forecasters, which is not the case for the 2 rolling years horizon. This may be connected to luck of 'Star' Forecasters in this particular case, to their superior forecasting capabilities in the 'usual' inflation framework, or to the under-performance of the ridge regression in this case. A very surprising result is that the pre-Covid performance of the 'Star' Forecasters is significantly better for the 2 rolling years horizons than for the 1 rolling year horizon. Indeed, then, the su-

perior performance of the 'Star' Forecasters may be partially attributed to luck in this particular case. In general, the performance of the Forecasters has been much closer, and in many cases superior, to the performance of the ridge regression in the regular inflation paradigm. This plays into our hypothesis that the overall out-performance of the ridge model relative to the Forecasters is mostly though not fully, explained by the inattentiveness or stickiness of the experts who were unable to foresee the post-Covid change in framework. An important aspect to take into account is that the pre-Covid under-performance of ridge relative to the Forecasters for calendar year derived horizons may be explained by the limited train data size. With so few data, the ridge model may have been able to learn general inflation patterns, and not the subtleties differentiating 1.5% and 2.0% inflation rates, for example.

We now report the normalised MSEs in the **high inflation regime** for all the Forecasters, the 'Star' Forecasters, the 'Underperformers' ('Under') Forecasters, the random forest regression, and the ridge regression for the analysed horizons:

Horizon	Forecasters	'Star'	'Under'	Random Forest	Ridge
1 rolling year	0.67	0.56	0.7	0.62	0.16
2 rolling years	0.7	0.63	0.7	0.66	0.22
2 months	0.002	0.001	0.007	0.45	0.03
5 months	0.03	0.005	0.13	0.42	0.007
8 months	0.1	0.03	0.28	0.42	0.01
11 months	0.41	0.32	0.52	0.44	0.11
14 months	0.58	0.5	0.62	0.52	0.3
17 months	0.65	0.61	0.69	0.47	0.46
20 months	0.68	0.61	0.69	0.51	0.44
23 months	0.68	0.63	0.73	0.65	0.62
26 months	0.69	0.62	0.72	0.65	0.53
29 months	0.68	0.57	0.68	0.68	0.79

Ridge regression outperforms Forecasters ('Star' Forecasters included) on

the majority of horizons in the post-Covid period. This is especially evident on shorter-term horizons, where higher inflation rates were more foreseeable. The increase in inflation rates was significantly better captured by the ridge model, while the Forecasters have likely not updated their information sets and exhibited stickiness to the previous inflation framework. It should be noted that the 'Star' Forecasters have better adapted to the shift in inflation paradigm, especially on shorter-term horizons. Indeed, best experts should foresee a change in paradigm and not 'stick' to the previous framework. However, on the longer-term horizons, the 'Star' Forecasters did not capture the shift.

As for the random forest model, as expected, it generally performs better pre-Covid, in the usual inflation framework. This explained by inherent random forest properties. It is better to use random forest in cases where test data falls in the same range as train data.

Let us close this section by comparing the Forecasters' biases in the pre-Covid and post-Covid periods. Again, for the 1 and 2 year horizons, we will assume that ridge regression's predictions are rational expectations of inflation at the time the forecasts are made. We present the mean error and the mean squared error of Forecasters relative to the rational (ridge) predictions, both normalised (by inflation and squared inflation, respectively), for the 1 and 2 rolling years prediction horizons over the pre-Covid, **low inflation period**.

Horizon	Normalised Mean Error	Normalised Mean Squared Error
1 rolling year	-0.02	0.06
2 rolling years	0.28	0.13

We present the same table for the post-Covid, **high inflation period**.

Horizon	Normalised Mean Error	Normalised Mean Squared Error
1 rolling year	-0.44	0.21
2 rolling years	-0.32	0.16

As expected, in the post-Covid predictions, Forecasters were further away from the machine predictions than pre-Covid and exhibited a negative bias in relation to the ridge prediction. This may be entirely explained by the Forecasters not expecting the post-Covid shift in the inflation paradigm and sticking to their previous predictions. Thus, the bias exhibited post-Covid is not so pessimistic or optimistic, but rather explained by the information rididity of the Forecasters.

However, in the pre-Covid world, the Forecasters were very close to the ridge predictions for the 1 rolling year horizon. As seen in the Section Forecasts Data Description, the pre-Covid mean error (spanning the total sample period and not only the test period) of the Forecasters relative to the realised inflation values is equal to -0.07, thus negative as well for the 1 rolling year horizon. The bias shown for this horizon may be explained by a small degree of optimism or by mere randomness. Its magnitude is not high enough for us to conclude that the Forecasters have shown a significant negative bias on the 1 rolling year horizon in the pre-Covid period.
For the 2 rolling years horizon, the Forecasters exhibited a large positive bias, thus predicting higher inflation values than the machine. Pre-Covid test period contains low inflation values seen during Covid, which were unpredictable 2 years before, as the pandemic had not emerged yet. Thus, the positive bias may be partially explained by this downward shock. In the Section Forecasts Data Description, the ME of the Forecasters relative to the realised inflation values is positive as well. Based on this findings, we conclude that the Forecasters have shown a positive bias on the 2 rolling years horizon in the pre-Covid period.

7 Conclusion

We have started this work with the question 'Can Machines Beat Professional Economic Analysts in Forecasting Macroeconomic Indicator(s)?'. In our case, the indicator has been inflation. We cannot give a 100% conclusive answer that would work for all the situations as a result of this work, but we have gotten closer to giving it.

We have used the ECB Survey of Professional Forecasters data to represent the predictions of the professional economic analysts. We have hy-

pothesized that the Forecasters are not a homogeneous group and found the 'Star' and 'Underperformers' groups of Forecasters. We have run a stickiness analysis on Forecasters checking if they regularly update their information to make predictions. We have found that the Forecasters tend to exhibit stickiness, and more so in expansions than in recessions, when they may even overreact to the existing information. In our opinion, stickiness is the key property of the forecasts that we have explored and that explains many of their shortcomings.

We have used the random forest regression and the ridge regression machine learning models to represent the machine forecasts, prioritising the latter model due to its superior performance. The ridge model has outperformed the Forecasters on the majority of the horizons. When looking for reasons of this superior performance, we have looked at the machine and human performance in the pre-Covid and the post-Covid periods separately. There has been a major shift in inflation paradigm post-Covid with inflation going up to rates unprecedented since decades. The ridge model has been able to adapt to the inflation paradigm shift to a much greater degree than the Forecasters. Ridge regression performance pre-Covid is far superior than that of humans. This can be explained by the Forecasters' stickiness and underreaction to the new information that may indicate the inflation paradigm shift. Thus, the machine is far better at adjusting to the new conditions and can definitely beat the human analysts in this aspect.

However, when looking at the pre-Covid performance, the situation is more balanced. While on the horizons (1 rolling year and 2 rolling years) where there is the most train data for the models, the ridge regression has outperformed the Forecasters, it has done so with a small advantage. Moreover, the 'Star' Forecasters have shown superior performance relative to the ridge model on one of these 2 horizons, even though in our opinion luck has been a major factor. Furthermore, on the horizons where there was is train data (2-29 months), Forecasters have made more accurate predictions than the machine. Therefore, in the usual inflation paradigm context the winner is currently harder to determine. More studies with more data are needed. When there is limited data and when there is no shift in conditions happening, professional economic intuition and understanding of the usual patterns may be preferred. When train data sample is big enough, in the 'regular' framework, machines may outperform

the general pool of the Forecasters with a slight margin. When there is a shift in macroeconomic conditions, machine predictions are preferred as the Forecasters exhibit stickiness to the previous framework.

We have also searched for the evidence of persistence positive and negative bias that the Forecasters may exhibit on the 1 rolling and 2 rolling years prediction horizons. When taking the post-Covid inflation out of the picture, the Forecasters exhibit a slight negative bias which may not be significant on the 1 rolling year prediction horizon. However, on the 2 rolling years prediction horizon, the Forecasters exhibit a significant positive bias. In our opinion, it is due to Forecasters' long-term expectations being anchored to the ECB target inflation rate and due to the inflation in the Euro area consistently being lower than the target in the pre-Covid period.

In summary, while the potential for machines to outperform human analysts definitely exists, machine learning models require sufficient training. Machines may be preferred in uncertain and changing situations, as information rigidity is the main drawback of human forecasts we have identified. Combining machine and human predictions may represent a superior approach and could be a direction of further research.

References

[1] Jules H. van Binsbergen and others (2022). *Man versus Machine Learning: The Term Structure of Earnings Expectations and Conditional Biases.* The Review of Financial Studies, Volume 36, Issue 6, June 2023, Pages 2361–2396.

[2] Hastie, T., Tibshirani, R., & Friedman, J. H. (2009). *The elements of statistical learning: data mining, inference, and prediction.* 2nd ed. New York, Springer.

[3] Coibion, Olivier and Yuriy Gorodnichenko (2015). *Information Rigidity and the Expectations Formation Process: A Simple Framework and New Facts.* American Economic Review, 105 (8): 2644-78.

[4] Mankiw, N. Gregory, and Ricardo Reis (2002). *Sticky information versus sticky prices: A proposal to replace the new Keynesian Phillips curve.* Quarterly Journal of Economics 117(4): 1295-1328.

[5] Sandeep Mazumder (2018). *Inflation in Europe after the Great Recession.* Economic Modelling, Volume 71, Pages 202-213.

[6] Véronique Genre, Geoff Kenny, Aidan Meyler, Allan Timmermann (2013). *Combining expert forecasts: Can anything beat the simple average?* International Journal of Forecasting, Volume 29, Issue 1, 2013, Pages 108-121.

[7] Christiane Nickel, Gerrit Koester and Eliza Lis (2022). *Inflation Developments in the Euro Area Since the Onset of the Pandemic* Intereconomics, 2022, 57(2), 69-75.

[8] Kenny, Geoff, Genre, Véronique, Bowles, Carlos, Friz, Roberta, Meyler, Aidan, & Rautanen, Tuomas (2007). *The ECB survey of professional forecasters (SPF) - A review after eight years' experience.* Occasional Paper Series 59, European Central Bank.

[9] Philipp F. M. Baumann and Enzo Rossi and Alexander Volkmann. (2022). *What drives inflation and how? Evidence from additive mixed models selected by cAIC.* Swiss National Bank Working Papers.

[10] European Central Bank. Monetary Aggregates. Retrieved from https://www.ecb.europa.eu/stats/money_credit_banking/monetary_aggregates/html/index.en.html.

[11] European Central Bank. Survey of Professional Forecasters. Retrieved from https://www-ecb-europa-eu.translate.goog/stats/ecb_surveys/survey_of_professional_forecasters/html/index.en.html?_x_tr_sl=en&_x_tr_tl=ru&_x_tr_hl=ru&_x_tr_pto=sc.

[12] European Central Bank. Euro Area HICP Inflation. Retrieved from https://sdw.ecb.europa.eu/.

[13] Federal Reserve Bank of St. Louis. OECD based Recession Indicators for Euro Area from the Period following the Peak through the Trough https://fred.stlouisfed.org/series/EUROREC#0.

[14] Federal Reserve Bank of St. Louis. Leading Indicators OECD: Reference Series: Gross Domestic Product (GDP): Original Series for the Euro Area (19 Countries) https://fred.stlouisfed.org/series/EA19LORSGPORGYSAM.

[15] European Central Bank. Euro Area general government debt-to-GDP ratio. Retrieved from https://sdw.ecb.europa.eu/quickview.do?SERIES_KEY=325.GFS.A.N.I8.W0.S13.S1.C.L.LE.GD.T._Z.XDC_R_B1GQ._T.F.V.N._T.

[16] European Central Bank. Euro Area Monetary aggregate M1. Retrieved from https://sdw.ecb.europa.eu/quickview.do?SERIES_KEY=325.GFS.A.N.I8.W0.S13.S1.C.L.LE.GD.T._Z.XDC_R_B1GQ._T.F.V.N._T.

[17] European Central Bank. Euro Area Monetary aggregate M3. Retrieved from https://sdw.ecb.europa.eu/quickview.do;jsessionid=33BDB7935C26650ECC7CD679D3E77B02?SERIES_KEY=117.BSI.M.U2.Y.V.M30.X.I.U2.2300.Z01.A.

[18] European Central Bank. Euro Area Producer Price Index, domestic sales . Retrieved from https://sdw.ecb.europa.eu/quickview.do?org.apache.struts.taglib.html.TOKEN=3ec83e4477dee4159aaa887bfe813dee&SERIES_KEY=132.STS.M.

I8.N.PRIN.2C0000.4.000&start=&end=&submitOptions.x=0&
submitOptions.y=0&trans=YPC.