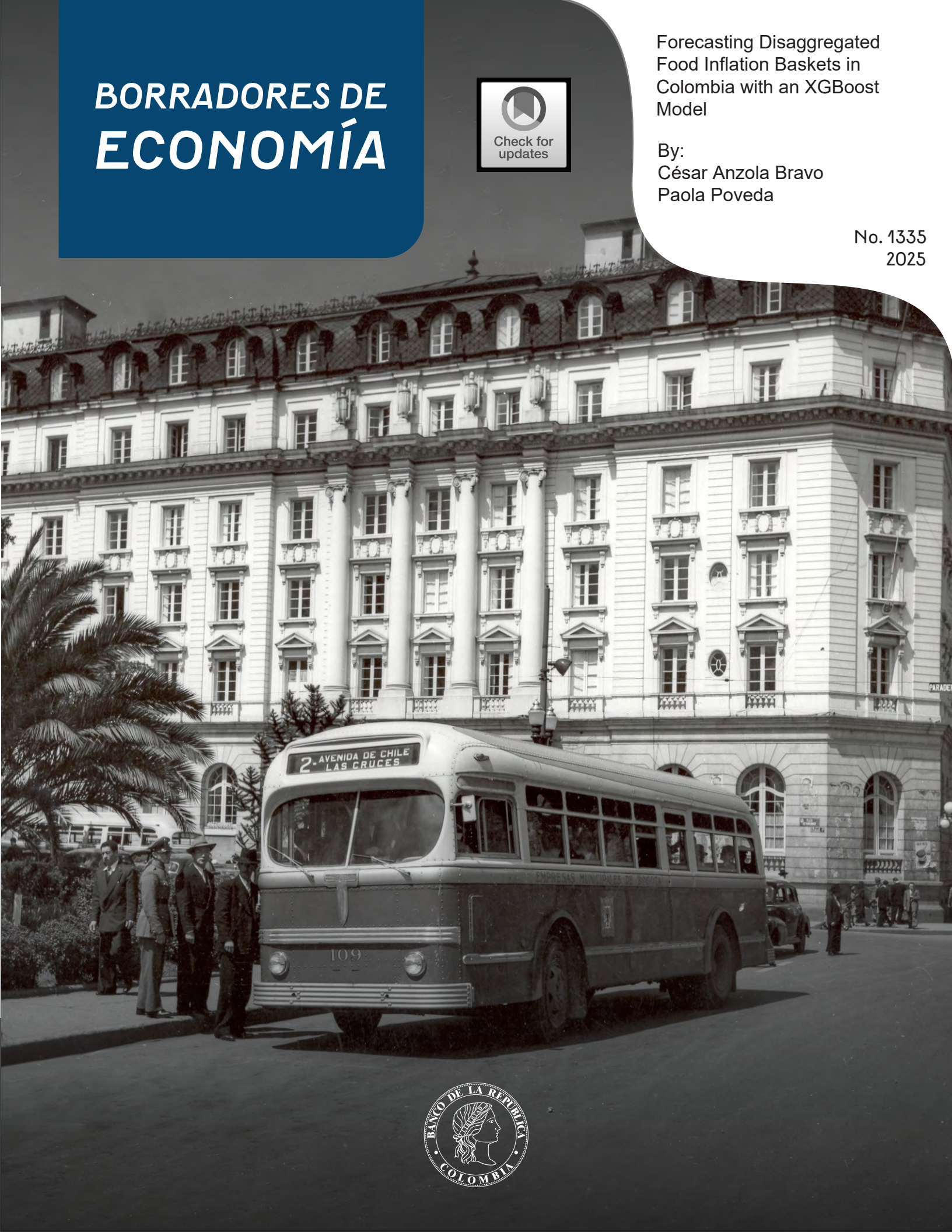
BORRADORES DE ECONOMÍA

Forecasting Disaggregated Food Inflation Baskets in Colombia with an XGBoost Model

By:
César Anzola Bravo
Paola Poveda

No. 1335
2025



BANCO DE LA REPÚBLICA
COLOMBIA

Forecasting Disaggregated Food Inflation Baskets in Colombia with an XGBoost Model

César Anzola Bravo*
canzolbr@banrep.gov.co

Paola Poveda †
apovedol@banrep.gov.co

Abstract

Food prices have consistently been one of the leading contributors to Colombia's inflation rate. They are particularly sensitive to exogenous factors such as extreme weather events, supply chain disruptions, and global commodity price shocks, often resulting in sharp and unpredictable price fluctuations. This document pursues two main objectives. First, it aims to estimate and evaluate methods for forecasting 33 homogeneous food inflation baskets, which together constitute the total food Consumer Price Index (Food CPI), offering tools that can assist policymakers in anticipating the drivers of future inflation. This includes both traditional time series models and modern machine learning approaches. Second, it seeks to enhance the interpretability of model predictions through explainable AI techniques. To achieve this, we propose a variable lag selection algorithm to identify optimal feature-lag pairs, and employ SHAP (SHapley Additive exPlanations) values to quantify the contribution of each feature to the model's forecast. Our findings indicate that machine learning models outperform traditional approaches in forecasting food inflation, delivering improved accuracy across most individual baskets as well as for aggregated food inflation.

Keywords: Macroeconomic Forecasts, Food Prices, Machine learning.

JEL codes: C53, E31, E37.

*Both authors are members of the Center of Studies on Production and Sectoral Trade at Banco de la Republica

†We would like to thank Margarita Gáfaro, Juan Esteban Carranza, Norberto Rodriguez, Juan Jose Ospina and Banco de la Republica seminar participants for their helpful comments.

Pronosticando inflaciones de canastas de alimentos desagregadas en Colombia usando un modelo XGBoost

César Anzola Bravo
canzolbr@banrep.gov.co

Paola Poveda
apovedol@banrep.gov.co

Resumen

Los precios de los alimentos han sido uno de los principales factores que contribuyen a la inflación en Colombia. Estos son particularmente sensibles a factores externos como choques climáticos, interrupciones en las cadenas globales de valor y choques en los precios de los productos básicos a nivel global, lo que resulta en fluctuaciones impredecibles de precios. Este documento tiene dos objetivos. En primer lugar, busca estimar y evaluar métodos para pronosticar 33 canastas homogéneas de inflación de alimentos, ofreciendo herramientas que puedan ayudar a los hacedores de política a anticipar los factores que afectan la inflación de alimentos futura. Esto incluye tanto modelos tradicionales de series de tiempo como enfoques modernos de machine learning. En segundo lugar, se propone mejorar la interpretabilidad de las predicciones de los modelos mediante técnicas de explainableAI. Para ello, proponemos un algoritmo de selección de variables que identifique las variables explicativas más relevantes, y utilizamos valores SHAP (SHapley Additive exPlanations) para cuantificar la contribución de cada variable explicativa en las predicciones del modelo. Nuestros hallazgos indican que los modelos de machine learning superan a los enfoques tradicionales en el pronóstico de la inflación de alimentos, logrando una mayor precisión tanto en la mayoría de las canastas individuales como en la inflación de alimentos agregada.

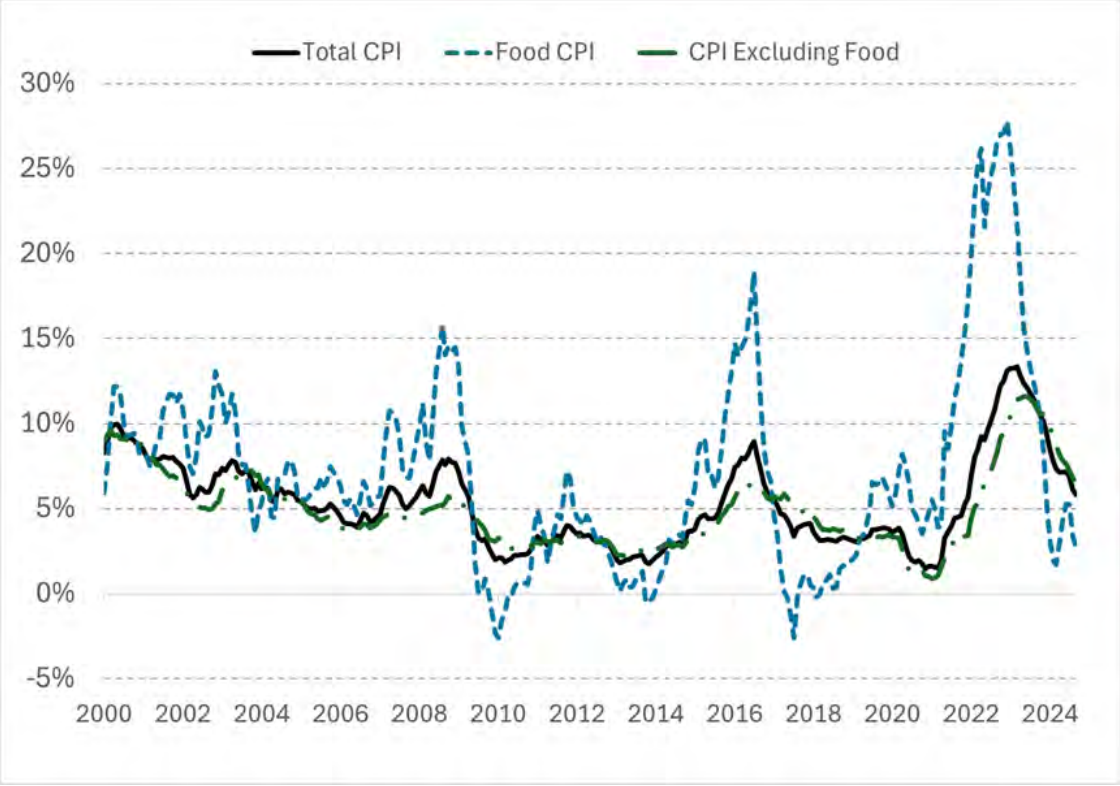
Keywords: Pronóstico Macroeconómico, Inflación de alimentos, Machine learning.

Códigos JEL: C53, E31, E37.

1 Introduction

In Colombia, the Consumer Price Index (CPI) is divided into two baskets: food and total CPI excluding food. Figure 1 shows the annual inflation rates for total CPI and these two baskets. Although food prices represent only 15% of the total CPI, they have been a major source of inflation volatility and have often remained above the inflation of other goods. High inflation levels and strong fluctuations, create challenges for low-income households, food producers, and monetary authorities for the following reasons. First, increasing food prices affect the affordability and accessibility of food, especially for low-income households. Second, food inflation shapes agricultural investments and food producers plating decisions. Third, food inflation increments have being transmitted to total CPI inflation rates, especially in the following periods: 2007-2009, 2014-2016, and 2021-2025. Thus, food inflation has caused total inflation rate to fall outside of the inflation target ranges announced by the central bank, directly affecting macroeconomic stability.

Figure 1: Inflation in Colombia - XXI Century



Notes: This figure shows Colombia's total CPI, food and non-food annual inflation rates. source: DANE, own calculations

Considering the above, forecasting food inflation and understanding its drivers is valuable for both policymakers and society in general. However, this task is challenging for two main reasons. First, food inflation CPI is comprised of a basket of highly heterogeneous products (including agricultural products, livestock, and processed foods) whose prices are influenced by distinct market

dynamics. Second, the relationship between food prices and their explanatory variables is often non-linear, which means that traditional linear time series methods may fail to adequately capture these dynamics. For instance, studies on agricultural commodities affected by droughts have shown that prolonged drought periods tend to exert increasingly severe impacts on crop yields and overall production, highlighting the complexity of modeling such effects (Birgani et al., 2022). Over time, these challenges have driven the development and refinement of sophisticated frameworks for monitoring inflation, incorporating advanced techniques in data collection, modeling, and forecasting.

This article pursues two objectives. First, it aims to estimate and evaluate methods for forecast up to 12 months ahead month to month inflation rates for 33 homogeneous food inflation baskets, which together constitute the total food CPI. This includes both traditional time series models and modern machine learning approaches. Second, it seeks to enhance the interpretability of model predictions through explainable AI techniques. To achieve this, we propose a lag structure optimization algorithm to identify the most relevant lags for each explanatory variable, and employ SHAP (SHapley Additive exPlanations) values to quantify the contribution of each lag feature to the model's predictions for each individual basket.

By decomposing the prediction into its main contributors, this article contributes to the literature on the dynamics of food inflation. Food prices are generally associated with temporary supply shocks, whether they come from a local or external source. From a domestic perspective, supply shocks are related to the nature of agricultural production and distribution. Those include things like extreme weather events, pests, and supply chain disruptions. In this regard, Abril-Salcedo et al. (2020) found that weather variables have a transitory and asymmetric effect on food inflation in Colombia. Their results suggest that a drought caused by El Niño increases food inflation up to nine months after the shock.

Similarly, in small open economies, the global trade system represents another potential source of inflationary pressure on food products. Among the most frequently cited factors are oil prices, exchange rates, and international demand for biofuels. However, the extent to which domestic food prices are affected by international trade depends on the pass-through which has been found to be higher in low income countries (Kohlscheen, 2022). Ultimately, many studies on the drivers of food prices usually find evidence for a combination of these two types of factors (Balogh & Sárvári, 2024; Ismaya & Anugrah, 2018; Köse & Ünal, 2024).

This study also adds to the forecasting literature that explores the use of ML methods to forecast macroeconomic indicators. This type of data is characterized by limited data from low-frequency observations, strong temporal dependencies, complex interdependencies between variables, and exposure to shocks. In the case of inflation, this complexity means that forecasters will often rely on subjective inflation expectations and structural models (Faust & Wright, 2013). ML methods such as tree-based models and neural networks have developed different tools to better accommodate these singularities (Petropoulos et al., 2022). With this exercise, we join the studies that have tested ML methods against other more traditional approaches and have found that ML can do a better job at forecasting macroeconomic indicators (Araujo & Gaglianone, 2023; Smalter Hall & Cook, 2017; Milunovich, 2020).

We combined the use of ML methods with the technique of aggregating forecasts of disaggregated variables as a way to exploit the heterogeneity they entangle.[1] In Colombia, several studies have tried a similar strategy with some differences. Martínez-Rivera et al. (2023) forecast 183-time series

[1] Martínez-Rivera et al. (2023) discuss this strategy in great detail.

items of the entire CPI using lags and a set of macroeconomic indicators. In addition, Zárate-Solano & Rodríguez-Niño (2024) highlight the importance of decomposing aggregate CPI trend variations into trend baskets and sub-sectors. Similarly, González-Molano et al. (2006) takes the food section of the CPI and compute forecasts for three groups: food away from home, processed foods, and unprocessed foods. Finally, Melo-Velandia et al. (2022) partitioned food products into perishable and processed goods. Except for González-Molano et al. (2006), these studies find evidence that supports aggregating forecasts to improve the overall forecast.

The rest of the article is divided as follows. Section 2 discusses the forecast objective, the data and the forecast methods used. Section 3 summarizes the forecast results, discusses performance and presents the contribution of the explaining variables to the prediction. Section 4 concludes.

2 Methodology

2.1 Data

2.1.1 Food CPI baskets

Colombia's Food CPI, published by DANE (National Administrative Department of Statistics), includes 59 items that may exhibit different price dynamics and determinants. However, some of these items share similar behaviors and underlying drivers, making it unnecessary to model each item individually. To simplify the analysis, we grouped the 59 items into 33 homogeneous baskets. The target variables from the study are the month-to-month inflation rates for these baskets from January 1999 to September 2024. Each basket combines items with comparable inflation patterns, both in mean and volatility, and similar responses to common influencing factors. These baskets belong to 3 groups: agricultural crops, animal products and industrial foods. The first group of products, such as vegetables and fruits, respond more to weather variables and input costs such as fertilizers. The second group, like pork or chicken, respond more to feed costs for livestock or substitute prices. The last group of products, such as bread or beverages, contains manufactured products with more complex production processes and input costs.The complete list of baskets is shown in Appendix A1

2.1.2 Features

We gathered a set of explanatory variables that fall into these four categories: nominal exchange rate,[2] commodity prices, energy costs, and weather variables. They were chosen due to their plausible economic links to production and price dynamics within each inflation basket. The selection process combined domain expertise with correlation analysis to ensure that the variables were both theoretically grounded and empirically relevant to the respective inflation rates. Moreover, each feature is carefully selected to be as exogenous as possible by checking low cross-correlations with the other explanatory variables, reducing the risk of misleading attributions and making computation faster.

Commodity prices influence food inflation through two main channels: the cost of raw materials used in food production and the role of global prices as determinants of local prices. Energy costs

[2]Nominal exchange rate is highly relevant for globally traded goods. This includes raw materials, final product exported prices and other second round macroeconomic effects.

affect food prices by influencing both transportation expenses and the cost of processing food in manufacturing plants. Thus, we included the price of diesel and the spot energy price at the national level. Weather impacts agricultural production by disrupting crop growth, soil conditions, and livestock health, which depend on stable climate patterns. These weather variables were calculated following a standard methodology used in weather shocks related literature (See Dell et al. (2014); Tirivarombo et al. (2018)). From a set of potential weather indicators, we retained those with the highest predictive ability for each basket's inflation rate (See Appendix A3). Finally, many of these inflation baskets have a seasonal component. To capture the cyclical nature, we apply cyclical encoding by transforming monthly dummy variables into their sine and cosine components.[3]

In Colombia, the production of vegetables, tubers, and fruits benefits from the country's diverse geography and tropical climate, which allows year-round cultivation without the seasonal constraints seen in temperate regions. This natural advantage supports a wide variety of crops and reduces dependency on imports. However, production costs remain a critical challenge, with fertilizers and transportation representing the largest share of expenses. Fertilizer costs shares vary significantly depending on the good, but in general it is a crucial input cost. Local fertilizers are closely tight to Urea prices worldwide which is why we used it as a driving factor for agricultural baskets.

We included futures contract prices for major global trading partners as driving factors for coffee, sugar and rice. Although coffee and sugar are locally produced in Colombia, prices for both goods are closely tight to international prices because of government regulations design to protect local producers. Internal coffee prices are determined with a formula that explicitly incorporate a global coffee price index. Similarly, local sugar prices respond to supply and demand factors but are managed through a band system that sets upper and lower limits based on international reference prices. Moreover, rice foreign prices are introduced to capture the effect of imported rice on local markets.

We also included futures contract prices for maize, soybean meals and pork prices. Maize and soybean meal are essential components of feed for poultry and pork production in Colombia, yet the country relies heavily on imports primarily from the United States and Brazil. This dependency leaves farmers highly exposed to international price fluctuations, as there are virtually no local substitutes when global prices rise. The vulnerability is compounded by the fact that feed costs represent a significant portion of total production expenses. Additionally, about one-quarter of Colombia's pork consumption is imported from the United States, meaning that U.S. pork prices exert a strong influence on local market dynamics of pork. The rest of the explanatory variables are other raw material prices that serve as inputs for the industrial foods group. We selected the raw material prices from the countries that represent Colombia's main sources of imports (See the full list of features included for each basket in Appendix A2)[4].

2.2 Model selection

Each food inflation basket is modeled individually using a combination of basket-specific predictors and shared predictors common across baskets. For each basket, we estimated two types of machine learning models: XGBoost and Elastic-net. The XGBoost model is a tree-based boosting algorithm

[3]This technique maps each month onto a unit circle, preserving the inherent periodicity of the calendar. This approach is particularly useful in models like XGBoost, which do not inherently recognize temporal relationships. For more information, see scikit-learn-doc "Time-related feature engineering"

[4]All features are included as monthly variations to match the transformation for inflation rates.

that sequentially reduces prediction errors by combining multiple weak learners, making it highly effective for structured data. On the other hand, the Elastic-net model is a regularized regression method that linearly combines the penalties of Lasso and Ridge regression. Both models were used to forecast up to 12 months ahead month to month inflation rate. We used a direct forecast approach were we estimated separate models for each forecast step h for a particular basket j. We compared predictions for both machine learning models with a traditional recursive forecast Sarimax.

The forecasting objective is formalized in Equations 1 and 2. In this framework, $X_t, \ldots, X_{t-L}$ represents a matrix of current and lagged explanatory variables that influence the target variable Y_{t+h} with a delay[5] and Y_t^j captures the inflation persistence. The model M^j is used to generate n forecasts for each basket j, based on a set of estimated parameters $\hat{\theta}$. The goal is to minimize the loss function $L(Y_i, \hat{Y}_i)$ by selecting the optimal combination of features X and parameters θ for each food inflation basket.

$$\hat{Y}^j_{t+h|t} = M^j(\hat{\theta}, Y^j_t, X^j_t, X^j_{t-1}, \ldots, X^j_{t-L}) \tag{1}$$

$$L(Y^j_i, \hat{Y}^j_i) = \frac{1}{n}\sum_{i=1}^{n}(Y^j_{i,t+h} - \hat{Y}^j_{i,t+h})^2 \tag{2}$$

Building a forecasting model also involves selecting the appropriate data and choosing an algorithm for identifying the relevant variables and tuning parameters. Regarding the data selection process, we partitioned the time series dataset into 3 standard subsets as shown in Table 1. This is a mandatory process in machine learning to ensure reliable model development and evaluation. The training set is used to fit the model, the validation set is used for tuning hyperparameters and preventing over-fitting, and the test set provides an unbiased estimate of the model's performance on unseen data. The validation period between January of 2014 and December of 2019 was selected to include both stable and volatile phases of food inflation, ensuring that the model is exposed to a diverse range of conditions during development. Furthermore, the test set between January 2020 and September 2024 includes predominantly volatile periods, providing a rigorous assessment of the model's robustness under challenging, high-uncertainty scenarios.

Table 1: Time series data partitioning

Dataset	**Time Period**
Training Set	January-1999 to December-2013
Validation Set	January-2014 to December-2019
Test Set	January-2020 to September-2024

In both the validation and test phase, we apply an expanding rolling window method for data selection which means that, at each step, a new model is trained using the most recent information. This enables the use of larger windows, which is more appropriate for stationary data such as inflation rates. (Rossi, 2013). As for the features and parameter selection, the chosen algorithm aims to account for the delayed impact of explanatory variables on inflation, driven by pass-through dynamics. A description of the lag selection algorithm is better describe in Algorithm 1.

[5]For example, global commodity price changes may take time to affect domestic food inflation due to transportation delays, inventory turnover, and pricing contracts.

Algorithm 1 Lag Selection Process

1. For a particular basket j we define a set of q candidate explanatory variables that have a lagged impact on y^j, $X^j = (x^j_{1,t}, x^j_{1,t-1}, ..., x^j_{1,t-12}, ..., x^j_{q,t}, x^j_{q,t-1}, ..., x^j_{q,t-12})$
2. To forecast y_{t+1} we select the lag $l1$ for X_1 that minimizes $L(y_{t+1}, \hat{y}_{t+1})$ on the validation set.
3. We fix $X_{1,t-l1}$ and combine it with lags for X_2 and keep the lag $l2$ that minimizes the loss function in the validation set.
4. We keep adding the rest of the q candidate features iteratively until we have an optimal lagged value for each candidate feature.
5. We do this process for each forecast step h $(y_{t+1}, ..., y_{t+h})$. Notice that each step can have different optimal lags for each feature.
6. After selecting each combination of lagged-features we tune hyper-parameters θ^j_{t+h} using an standard grid search that minimizes $L(y_{t+h}, \hat{y}_{t+h})$ on the validation sample for each step h
7. Finally we have a combination of X^j_{t+h} and $\hat{\theta}^j_{t+h}$ for each forecast step h $(y_{t+1}, ..., y_{t+h})$

We developed a "greedy", stepwise algorithm to select the optimal lag structure for each time step h. Every lagged information is known at time t and would be used to forecast up to 12 steps ahead. The algorithm begins by selecting the first variable and choosing the lag that minimizes the loss function on the validation set. The second variable is then added, and its optimal lag is selected while keeping the first variable and its lag fixed. This process continues until all candidate variables have been incorporated, resulting in a combination of optimal lags for each candidate variable with the best predictive power.[6] Hyperparameters are also tuned for optimal performance.[7]. Notice that with this approach each forecast horizon has a particular model with specific optimal lagged features and hyperparameters[8]

2.3 Models

2.3.1 XGBoost

XGBoost is a widely used gradient-boosting algorithm designed for supervised learning tasks. Gradient boosting builds models iteratively by combining multiple weak learners, typically decision trees, with each new learner focusing on correcting the prediction errors of its predecessors. While traditional gradient boosting can be computationally intensive and may struggle to scale with large datasets, XGBoost addresses these limitations by optimizing for computational efficiency without compromising predictive performance.

[6] As a greedy algorithm, the final configuration is locally optimal and depends on the order in which variables are introduced. However, this limitation is mitigated by the low correlations among carefully selected explanatory variables within each inflation basket, which reduces the risk of suboptimal selections due to variable ordering.

[7] The grid search for the XGBoost include the following parameter ranges: max depth [1,4], learning rate [0.1,0.4], number of trees [50,150], gamma [0.5,4]. The grid search for elastic-net includes regularization parameters. The Sarimax is tuned for optimal MA and AR components.

[8] While this approach enhances predictive performance, the selected lags vary across horizons and can be volatile. This variability reflects the changing relevance of past information depending on the forecast horizon. Appendix A4 shows an example that shows the relationship between forecast horizons and selected lags.

Equations 3 to 8 show the most relevant structure for XGBoost models. The first equation describes the objective function $\mathcal{L}$ which contains two parts: the loss function $l(y, \hat{y})$ and a regularization term $\Omega(f)$ which helps to avoid overfitting by penalizing overcomplicated trees. That is, penalizing the number of leafs T and larger leaf weights w. The optimal parameters enable the construction of a sequence of trees that balance predictive accuracy with simplicity, ensuring they remain computationally efficient and interpretable. Finally, each prediction at iteration t $\hat{y}_{i,t}$ is expressed by Equation 8. At each step, the model is trained to minimize the errors of the previous trees, f_k by iteratively adding new trees that correct the residuals from the prior iteration (See Chen & Guestrin (2016)). Figure 2 illustrates several ways of splitting the data across variable t in a decision tree fashion. Panel b) shows a high value for the regularization $\Omega(f)$ because the tree has too many splits (or leafs). Conversely, the panel c) has a wrong split for $t1$ which yields in a high value for the loss function $\mathcal{L}(f)$. Finally, panel d) shows an appropriate number of leafs that minimizes the loss function at the same time by choosing the right threshold.

$$\mathcal{L}_{\sqcup} = \sum_{i=1}^{n} l(y_i, \hat{y}_{i,t-1} + f_t(\mathbf{x}_i)) + \sum_{k=1}^{K} \Omega(f_t) \tag{3}$$

$$\Omega(f_t) = \gamma T + \frac{1}{2}\lambda \sum_{j=1}^{T} w_j^2 \tag{4}$$

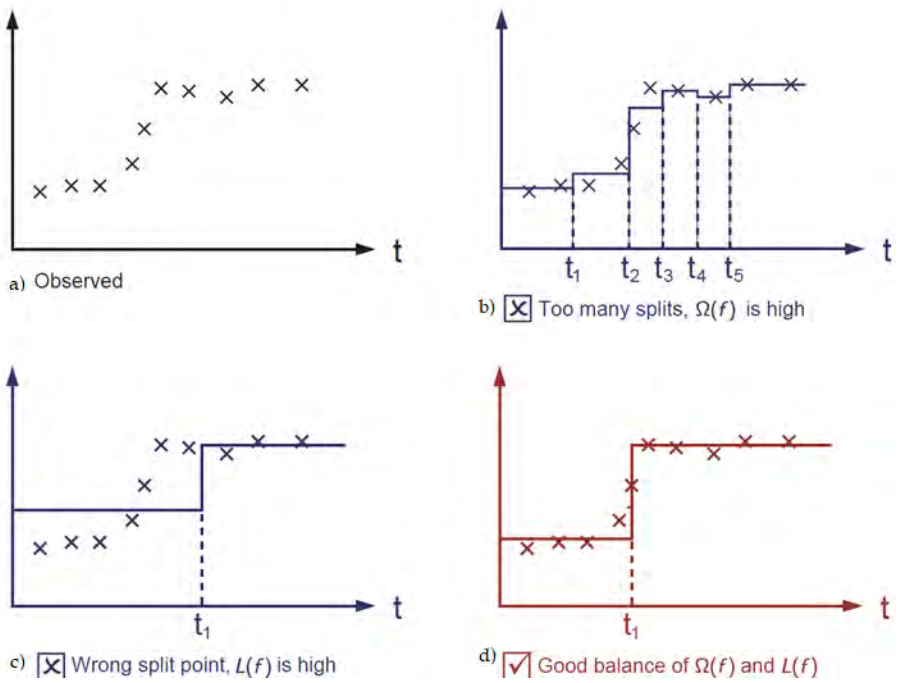
$$w_j^* = -\frac{\sum_{i \in I_j} g_i}{\sum_{i \in I_j} h_i + \lambda} \tag{5}$$

$$g_i = \frac{\partial l(y_i, \hat{y}_{i,t-1})}{\partial \hat{y}_{i,t-1}} \tag{6}$$

$$h_i = \frac{\partial^2 l(y_i, \hat{y}_{i,t-1})}{\partial (\hat{y}_{i,t-1})^2} \tag{7}$$

$$\hat{y}_{i,t} = \sum_{k=1}^{t} f_k(\mathbf{x}_i) \tag{8}$$

Figure 2: XGBoost



Notes: This figure illustrates how regularization parameters help to identify the appropriate cuts to prevent overfitting (Created by XGBoost developers).

2.3.2 Elastic-net

The Elastic Net is a regularized regression model that combines the penalties of the Lasso (Tibshirani, 1996) and Ridge (Hoerl & Kennard, 1970) methods to address the limitations associated with each approach when applied independently. Specifically, it introduces a convex penalty term that includes both the absolute value and the squared magnitude of the linear regression coefficients, thereby encouraging sparsity and group selection simultaneously. This dual penalty structure makes the Elastic Net particularly effective in high-dimensional settings where the number of predictors exceeds the number of observations or when predictors exhibit high multicollinearity. By tuning the mixing parameter, the model can interpolate between the Lasso and Ridge extremes, achieving a balance between variable selection and coefficient shrinkage. The optimization problem is shown in Equation 9. The goal is to estimate β and chose optimal regularization parameters λ and α (See Zou & Hastie (2005)).[9]

$$\hat{\boldsymbol{\beta}} = \arg\min_{\boldsymbol{\beta}} \left\{ \frac{1}{2n} \|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|_2^2 + \lambda \left[\alpha \|\boldsymbol{\beta}\|_1 + \frac{1}{2}(1-\alpha) \|\boldsymbol{\beta}\|_2^2 \right] \right\} \tag{9}$$

[9]We tuned regularization parameters α and λ using an standard grid search.

2.3.3 Sarimax

Sarimax (AutoRegressive Integrated Moving Average with exogenous variables) model extends the Sarima model by incorporating exogenous variables to improve forecasting accuracy. Like Sarima, it models a time series by combining three components: autoregression (AR), integration (I), and moving average (MA). The AR component captures the relationship between the current value of the series and its lagged values, the I component addresses non-stationarity by differencing the series, and the MA component models the relationship between the current value and past forecast errors. The Sarimax model includes these elements while also incorporating the effect of external regressors X_t (See (Jenkins & Box, 1976)).[10] For this document, we chose the Sarimax that minimizes the AIC criteria using an automatic arima.[11]

$$\Delta^d Y_t = c + \sum_{s=1}^{p} \phi_s \Delta^d Y_{t-s} + \sum_{s=1}^{q} \theta_s \epsilon_{t-s} + \sum_{s=0}^{r} \beta_s X_{t-s} + \epsilon_t, \tag{10}$$

3 Results

3.1 Forecast Accuracy

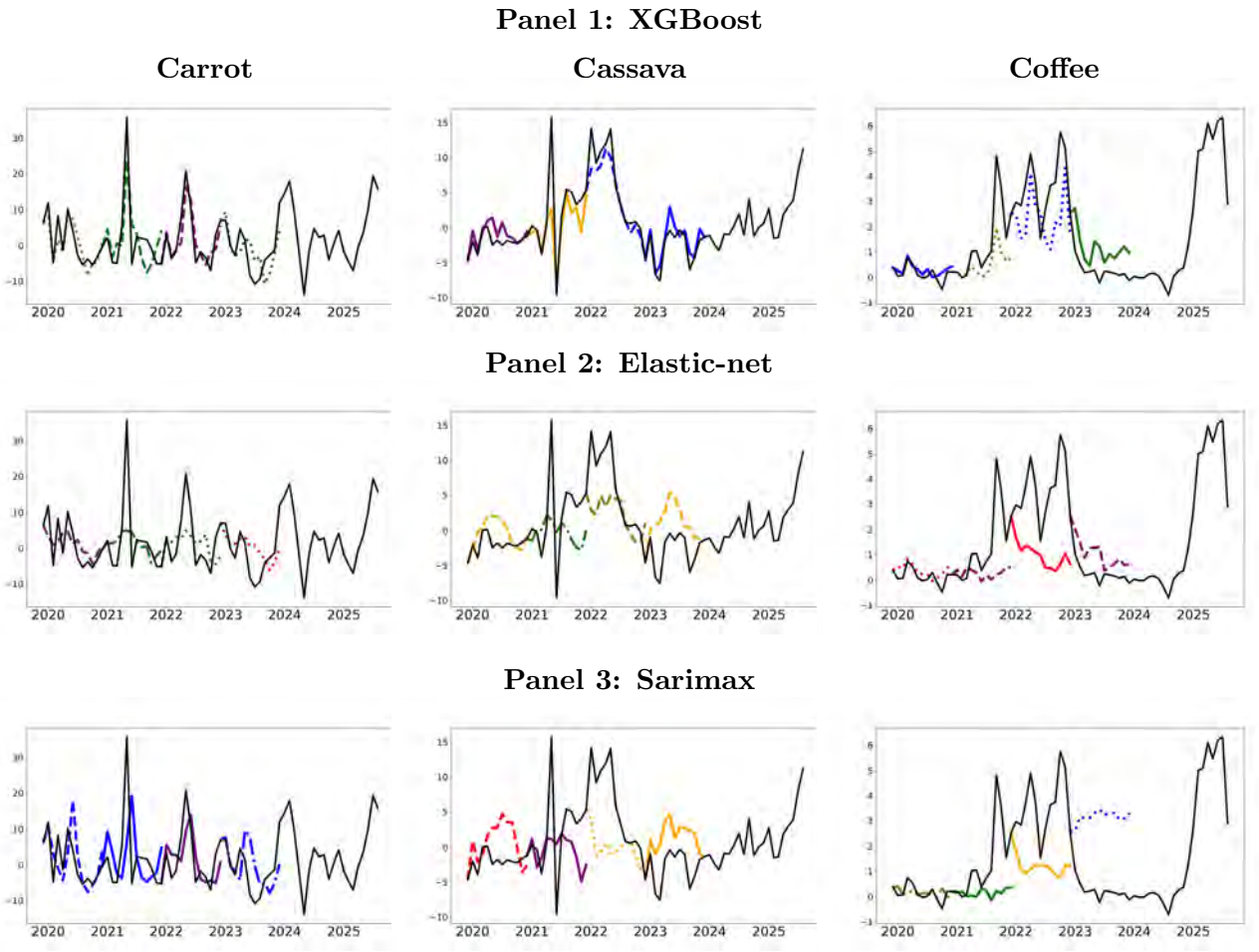
In this section, we discuss the models' forecast accuracy. First, we evaluate the models' forecast performance using an expanding window forecast on the test set from January 2020 to May 2024. Figure 3 presents the forecast across the three models (XGBoost, Elastic-net, and Sarimax) for three baskets: Carrot, Cassava, and Coffee. Visually, XGBoost appears to capture most short-term fluctuations and volatility, particularly for crops like cassava and coffee. This suggests that the model's non-linear structure and its ability to capture complex interactions may offer a distinct advantage in dynamic pricing environments. Elastic-net produces smoother forecasts, which may be beneficial for factors with more stable trends, but could miss abrupt changes. Sarimax, while generally more stable and consistent, seems less responsive to sudden shifts which may limit its effectiveness in highly volatile series.[12]

[10] To compare the forecasting performance of the three models, we included the same set of features as explanatory variables. The main difference is that for Sarimax the external regressor effects are linear.

[11] We used the pmdarima python package to select the best Sarimax model based on the standard AIC criteria.

[12] See Appendix A6 to check visually the inflation forecast expanding windows for other baskets with an important weight on food CPI

Figure 3: Expanding window forecast



Notes: shows an expanding window forecast use to evaluate the models predictive performance betweem 2020 and 2024. Black solid line represents the observed month to month inflation rate. Each line style - color combination represents a different 1 to 12 steps ahead forecast made with information known at the moment of forecasting.

Next, we compute the Mean Absolute Forecast Errors (MAFEs) from the test set and apply the Giacomini and White test (Giacomini & White, 2006) to assess wether the predictive performance of the XGBoost model surpasses that of Elastic-net and SARIMAX.[13] Tables 2, 3 and 4 show this analysis for 3 food groups (agricultural, animal and industrial foods products) and 5 forecast horizons (2, 4, 7, 10 and 12 months ahead). Overall, the tree-based model outperforms the other two models, at the 5% significance level, for most inflation baskets. The accuracy gains are especially pronounced for agricultural and animal food products, while improvements are less substantial for industrial food items.

Table 2 reports the MAFEs for agricultural food products, which tend to have the most volatile

[13]The Giacomini and White (2006) test is a statistical method used to compare the predictive ability of two forecasting models. The test is designed to work in out-of-sample settings and allows for conditional predictive ability, meaning it can account for time-varying forecast performance.

inflation rates and, consequently, the highest forecast errors. Given that the magnitude of these errors varies significantly depending on each product's volatility, we focus on how the non-linear model (XGBoost) enhances forecast accuracy across different baskets and time horizons. To assess this, we calculate the percentage change in forecast errors by dividing the errors from the XGBoost model by those from the second-best performing model. Forecast accuracy gains are larger for fruits, plantain, tomato and cassava. For the XGBoost model forecast errors are between 40% and 80% lower across time horizons compare to the second best performing model. On the other hand, for rice, sugar, coffee and onions forecast errors are between 20% and 50% lower than the other models. Potatoes show mixed results: XGBoost performs well in the first month but loses accuracy in subsequent periods. Lastly, for carrots and other vegetables, the gains are smaller and statistically significant in only 2 out of 5 forecast steps.

Table 2: Forecast errors for agricultural food products

Basket	Model	t+2	t+4	t+7	t+10	t+12
Carrot	Xgb	1.52*	3.57*	3.16*	5.80*	6.12
	Elast	5.91	6.11	6.10	7.06	6.04
	Sarimax	6.36	6.19	6.58	7.16	7.53
Cassava	Xgb	0.90*	0.46*	1.47*	0.69*	1.36*
	Elast	4.00	4.00	3.95	4.28	4.00
	Sarimax	4.79	5.16	5.06	4.98	4.88
Coffee	Xgb	0.87*	0.60*	0.85*	1.08*	1.21*
	Elast	1.10	1.21	1.33	1.45	1.37
	Sarimax	1.28	1.45	1.78	2.09	2.20
Fruits	Xgb	0.66*	0.70*	1.10*	1.50	1.40*
	Elast	2.91	2.82	2.94	3.19	2.91
	Sarimax	2.21	2.19	2.23	2.19	2.22
Onion	Xgb	5.58*	7.79	5.81*	5.05*	6.17*
	Elast	8.45	7.77	8.78	8.18	8.63
	Sarimax	9.70	7.86	8.23	7.99	8.38
Other Vegetables	Xgb	2.56	1.39*	2.33	2.16	1.03*
	Elast	2.78	2.35	2.16	1.98	2.01
	Sarimax	3.03	2.54	2.48	2.41	2.33
Plantain	Xgb	4.09*	3.17*	1.26*	4.74	0.97*
	Elast	5.61	5.25	5.23	5.27	5.33
	Sarimax	6.02	5.33	5.20	5.37	5.33
Potatoe	Xgb	8.56	8.04	4.08*	6.96*	6.66*
	Elast	8.83	8.73	9.04	8.56	10.17
	Sarimax	8.91	8.83	9.85	10.45	10.74
Rice	Xgb	1.01*	1.09*	1.81	0.91*	0.86*
	Elast	1.79	1.71	1.76	1.54	1.57
	Sarimax	1.93	1.92	1.95	1.68	1.64
Sugar	Xgb	0.94*	0.56*	0.63*	0.63*	0.74*
	Elast	1.07	1.19	0.90	1.01	1.00
	Sarimax	1.08	1.13	1.05	1.05	1.04
Tomato	Xgb	3.01*	7.63*	1.66*	7.26*	3.21*
	Elast	10.99	10.98	10.93	10.31	10.67
	Sarimax	10.39	10.64	10.74	10.49	10.73
Other Tubers	Xgb	1.23*	2.18*	1.72*	1.21*	2.19*
	Elast	3.00	2.88	2.94	3.18	2.96
	Sarimax	3.23	3.22	3.20	3.29	3.08

Notes: This table shows the MAFEs for the 3 models and for 5 forecast horizons: 2, 4, 7, 10 and 12 months ahead. The * symbol indicates that the XGBoost model has a lower forecast error than the other two models at the 10% significance, based on the Giacomini and White test.

Table 3 shows MAFEs for animal food products. These product prices are less volatile than agricultural food products, but are also significantly subject to weather and raw material shocks. Chicken, eggs and dairy are the baskets with the largest accuracy gains for this group. Forecast errors for the XGBoost model are between 30% to 50% lower across time horizon in contrast with the elastic-net model. Pork basket follows, with accuracy gains that range between 20% to 40%. Finally, for beef and Seafood their is no difference in predictive ability between the models.

Table 3: Forecast errors for animal food products

Basket	Model	t+2	t+4	t+7	t+10	t+12
Beef	Xgb	0.71	0.50	0.66	0.53	0.53
	Elast	0.88	0.98	0.99	1.00	1.02
	Sarimax	0.95	1.03	1.05	1.06	1.05
Chicken	Xgb	0.62*	0.56*	0.60*	0.54*	0.63*
	Elast	0.99	1.02	0.93	0.94	0.93
	Sarimax	0.86	0.95	0.89	0.86	0.87
Dairy	Xgb	0.36*	0.69*	0.62*	0.62*	0.59*
	Elast	0.78	0.82	0.89	0.96	0.98
	Sarimax	0.82	1.08	1.26	1.50	1.60
Egg	Xgb	1.14*	1.57*	0.77*	1.02*	0.79*
	Elast	2.24	2.10	2.07	2.04	1.99
	Sarimax	2.10	2.12	2.11	2.01	1.96
Pork	Xgb	0.54*	1.05	0.73*	0.67*	0.95*
	Elast	1.05	1.05	1.14	1.06	1.08
	Sarimax	1.13	1.21	1.22	1.18	1.13
Seafood	Xgb	0.56	0.59	0.55	0.56*	0.58
	Elast	0.58	0.60	0.56	0.61	0.61
	Sarimax	0.63	0.70	0.71	0.74	0.73

Notes: This table shows the MAE for the 3 models and for 5 forecast horizons: 2, 4, 7, 10 and 12 months ahead. The * symbol indicates that the Xgboost model has a lower forecast error than the other two models at the 10% significance, based on the Giacomini and White test.

Table 4 reports the MAFEs for industrial food baskets. This group exhibits relatively stable inflation rates and shows the smallest accuracy improvements when using the XGBoost. Two main reasons can explain this lower performance: (1) limited nonlinear relationships between features and inflation rates, and (2) missing variables that could be more influential for this group. Despite this, forecast improvements can be observed for most of this group basket. Beverages, snacks, corn flour, bread, brownies and sauces are the baskets with the largest accuracy gains for this group. Forecast errors for the XGBoost model are between 20% to 35% lower across time horizon in contrast with the elastic-net model. Moreover, chocolate, corn flour, vegetable oils and spices have significant accuracy gains for 3 out of 5 forecast steps with ranges between 5% and 20%. Finally, for the rest of the products their is no difference in predictive ability between the models.

Table 4: Forecast errors for industrial food products

Basket	Model	t+2	t+4	t+7	t+10	t+12
Beverages	Xgb	0.39*	0.47*	0.46*	0.42*	0.47*
	Elast	0.53	0.56	0.56	0.54	0.60
	Sarimax	0.53	0.74	0.87	0.90	0.93
Bread	Xgb	0.43*	0.79	0.70*	0.62*	0.68*
	Elast	0.85	0.88	0.90	0.92	0.92
	Sarimax	0.74	0.91	1.09	1.29	1.38
Brownies and Cookies	Xgb	0.45*	0.57*	0.60*	0.60*	0.65*
	Elast	0.58	0.67	0.72	0.74	0.76
	Sarimax	0.58	0.69	0.79	0.83	0.88
Chocolate Drink	Xgb	0.58*	0.74	0.75*	0.90	1.31*
	Elast	0.77	0.76	0.84	1.01	1.36
	Sarimax	0.84	1.04	1.13	1.34	1.73
Cold Meats	Xgb	0.69	0.83	0.87*	0.77	0.83
	Elast	0.73	0.89	0.92	0.94	0.94
	Sarimax	0.93	1.08	1.13	1.11	1.12
Corn Flower	Xgb	0.58*	0.68*	0.89*	0.88*	0.96*
	Elast	0.95	1.02	1.11	1.19	1.26
	Sarimax	0.82	1.05	1.29	1.56	1.62
Instant Soups	Xgb	0.64	0.53*	0.58*	0.56*	0.60
	Elast	0.65	0.68	0.66	0.65	0.68
	Sarimax	0.65	0.69	0.81	0.84	0.89
Other Flours	Xgb	0.52*	0.57*	0.52*	0.73	0.59*
	Elast	0.60	0.66	0.70	0.71	0.74
	Sarimax	0.64	0.73	0.82	0.90	0.95
Pasta	Xgb	0.56	0.62	0.67	0.67	0.70
	Elast	0.58	0.70	0.70	0.71	0.72
	Sarimax	0.68	0.82	0.86	0.82	0.84
Pulses	Xgb	2.02	1.92	1.69	1.60	1.71
	Elast	2.24	2.00	1.74	1.67	1.98
	Sarimax	2.61	2.20	1.75	1.85	1.78
Salt and Nuts	Xgb	0.50	0.57	0.58*	0.66	0.65*
	Elast	0.61	0.64	0.66	0.67	0.68
	Sarimax	0.53	0.60	0.67	0.73	0.78
Sauces and pastes	Xgb	0.54*	0.54*	0.70*	0.72*	0.66*
	Elast	0.66	0.72	0.85	0.83	0.82
	Sarimax	0.64	0.78	0.89	0.89	0.89
Candy and Snacks	Xgb	0.44	0.63	0.56*	0.58*	0.53*
	Elast	0.65	0.68	0.77	0.82	0.78
	Sarimax	0.48	0.59	0.68	0.76	0.81
Culinary Herbs	Xgb	2.00*	1.84	1.78*	1.85*	1.17*
	Elast	2.19	1.95	1.91	2.14	2.20
	Sarimax	2.41	2.23	2.15	2.30	2.23
Vegetable Oil	Xgb	0.66	1.04	1.36	0.91*	1.00*
	Elast	0.88	1.09	1.39	1.37	1.33
	Sarimax	0.83	1.08	1.24	1.37	1.40

Notes: This table shows the MAE for the 3 models and for 5 forecast horizons: 2, 4, 7, 10 and 12 months ahead. The * symbol indicates that the XGBoost model has a lower forecast error than the other two models at the 10% significance, based on the Giacomini and White test.

3.2 Forecast Interpretation

As discussed previously, most of the literature on food inflation forecasting has concentrated on developing models that improve predictive accuracy. In recent years, there has been a growing emphasis on deep learning models with complex architectures, which are capable of capturing intricate non-linear relationships between predictors and targets. However, this increase in complexity often comes at the cost of interpretability. While these advanced models may enhance forecast performance, their black-box structure makes it difficult to understand the underlying mechanisms driving the results. From a policymakers's perspective, interpretability is essential, as it allows them to trust and act upon model outputs. For this reason, in this study, we aim to develop a model that strikes a balance between predictive accuracy and interpretability—delivering reliable forecasts while remaining transparent.

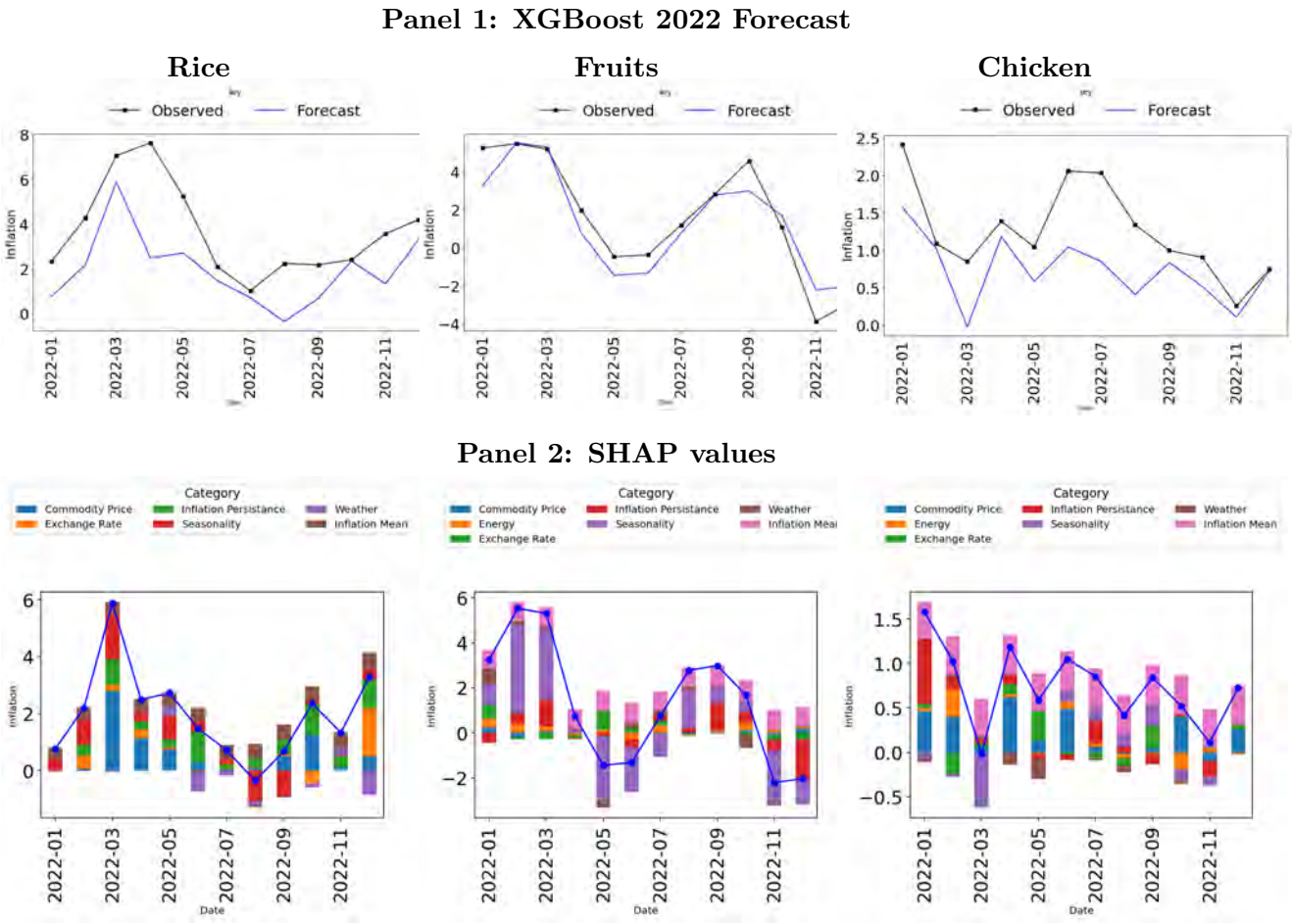
We use SHAP values (SHapley Additive exPlanations) to interpret the XGBoost model predictions following (Lundberg & Lee, 2017).[14] This approach has gained significant traction in recent years, with a growing body of literature applying SHAP to enhance the interpretability of complex machine learning models. Even though SHAP values are widely used for models interpretation, an important limitation is that the estimation is uncertain when features exhibit multicollinearity. In such cases, SHAP may distribute importance across correlated variables in a way that obscures the true drivers of model predictions. However, this limitation is mitigated in this document because model predictions rely on a carefully selected, limited set of features. Each feature is chosen to be as exogenous as possible and exhibits low correlation with the others, reducing the risk of misleading attributions and enhancing the reliability of the SHAP-based interpretation.[15]

For illustration, the first panel of Figure 4 shows one particular forecasts for three baskets: rice, fruits, and chicken, from January 2022 to December 2022 (blue line) and compare them with the actual values (black line). The second panel decomposes these predictions into SHAP values, revealing the influence of individual features on the model outcome. For instance, rice inflation forecast peak during the second quarter, which follows a similar pattern than the observed inflation, is mainly attributed to urea price increases (yellow color). In contrast, the inflation forecast for fruits during this period is largely influenced by the cyclical component of seasonal patterns. Lastly, the forecast for chicken inflation reflects the impact of maize prices, with additional, but less significant, contributions from other factors such as pork prices and the nominal exchange rate.

[14] SHAP (SHapley Additive exPlanations) values are a powerful tool for interpreting machine learning models by quantifying the contribution of each feature to a specific prediction. Based on cooperative game theory, SHAP assigns each feature an importance value that reflects how much it influenced the model's output.

[15] See Appendix A5 to check feature cross correlations used for individual baskets.

Figure 4: December 2021 forecast - SHAP values



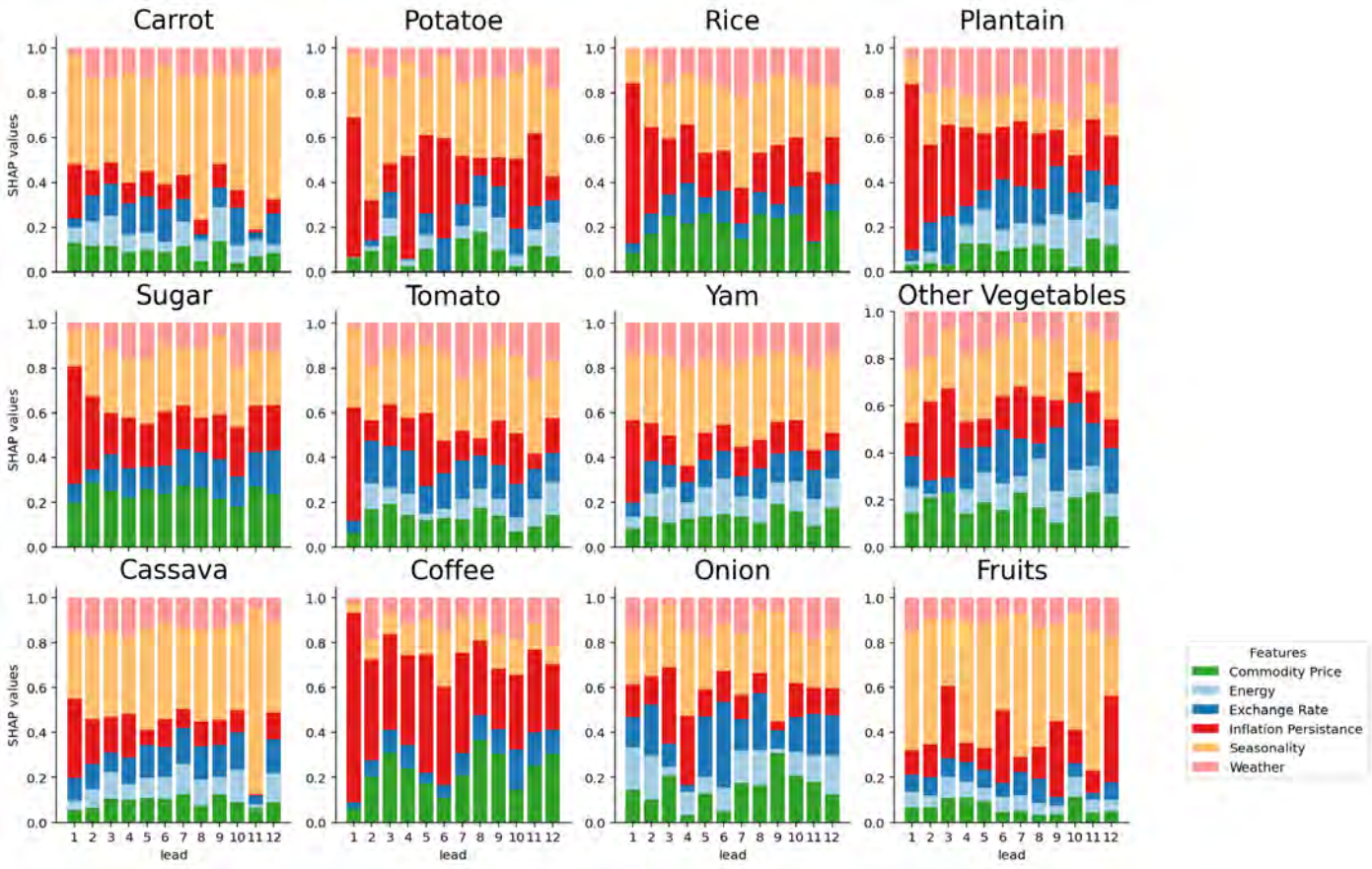
Notes: Panel 1 compares the 12 months ahead forecast made with information up to December 2021 (blue line) with the actual inflation rate (black line) for three baskets: rice, fruits and chicken. Panel 2 shows the SHAP values of each feature on the model forecast. The sum of the SHAP values is equivalent to the model prediction.

Figure 5 illustrates the average contribution of each feature to the forecast for agricultural food products between 2020 and 2024.[16] Since a distinct model is estimated for each forecast horizon, the influence of these features varies across the forecast steps (shown on the x-axis). Seasonal patterns play a significant role in these baskets, as many crops are harvested during specific weather conditions, such as dry seasons. The relevance of other features differs depending on the crop basket. For example, the last observed inflation value (highlighted in red), which reflects the persistence of inflation, is particularly important for coffee, plantain, and sugar products. In contrast, weather has a greater impact on plantain, rice, tomatoes, other vegetables, and yam, but are less influential for other baskets. Foreign prices and fertilizer costs are key factors for sugar, coffee and rice products, while their contribution is lower for other baskets. Finally, diesel prices have a notably higher

[16] We took the mean absolute value for the SHAP values of each feature by forecast horizon.

impact on onions and plantain forecasts compared to the other products.

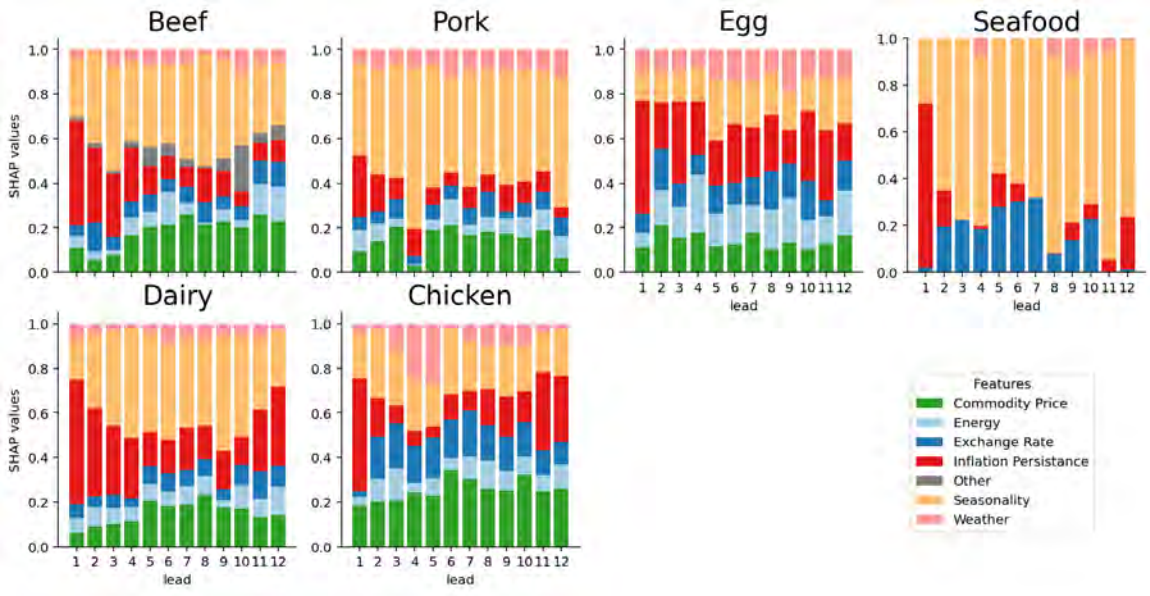
Figure 5: XGBoost SHAP Values - Agricultural Food Products



Notes: This figure shows the mean absolute SHAP values of each feature for all the forecasts samples between 2020 and 2024. X axis is the forecast horizon: 1 to 12 periods ahead, and Y axis are the absolute mean SHAP values. Because each forecast horizon is estimated with different models and parameters, features contributions change between forecast horizons.

Figure 6 illustrates the average contribution of each feature to the forecast of animal food products. For this group the seasonal component (in yellow) is important for pork, dairy and seafood, but it is less relevant for the rest of the products. Foreign prices (in green) are more relevant for chicken, dairy and pork than for the rest of the baskets. Weather (in red) is a contributing factor for chicken and eggs, but are not as relevant for the rest of the baskets. Inflation persistence (in yellow) is highly relevant for dairy products, especially during the first forecast horizons and in less degree for beef, chicken and eggs. In addition, diesel variations are more relevant for eggs and exchange rate contribution is especially important for seafood.

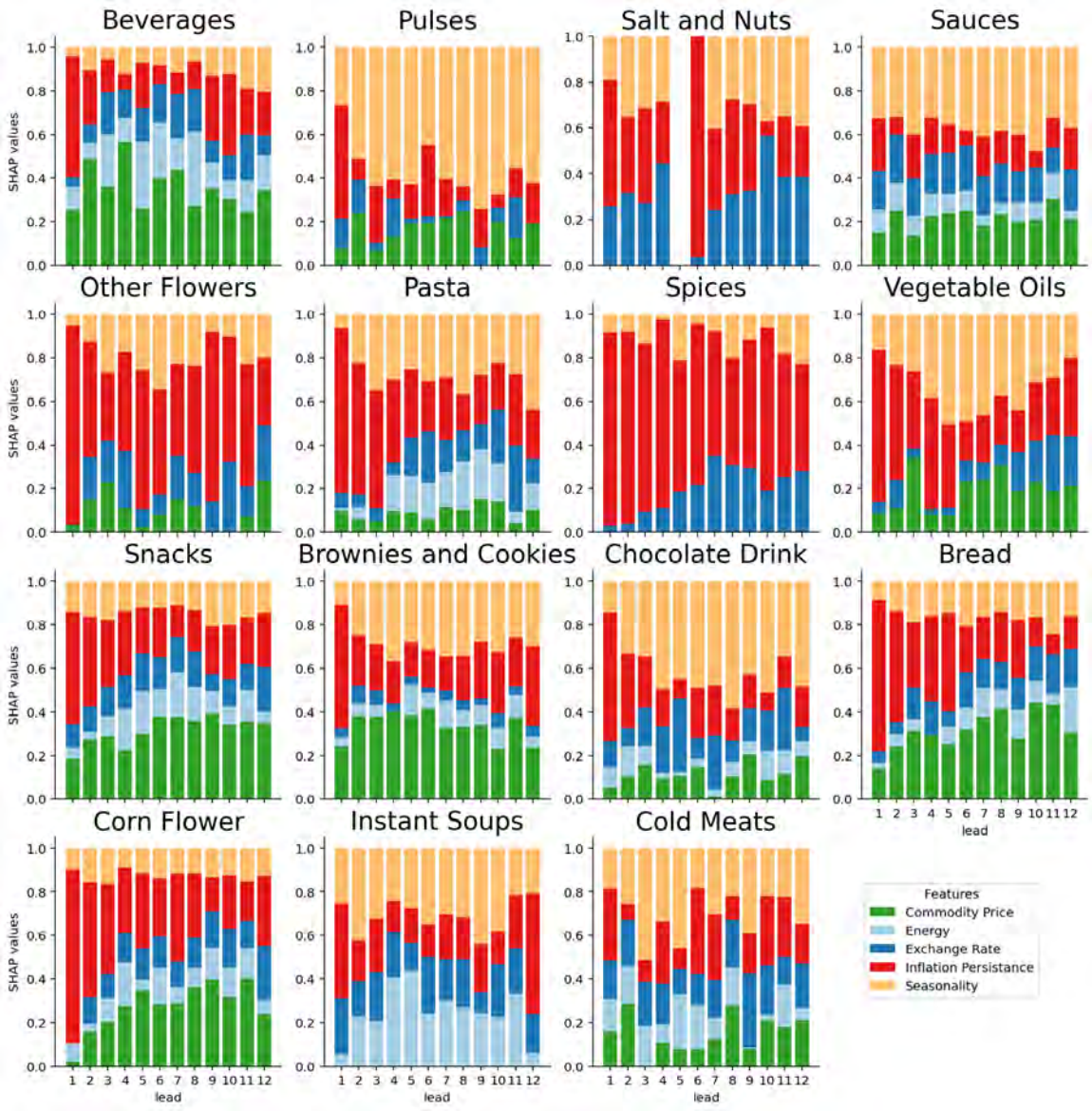
Figure 6: XGBoost SHAP values - Animal Food Products



Notes: This figure shows the mean absolute SHAP values of each feature for all the forecasts samples between 2020 and 2024. X axis is the forecast horizon: 1 to 12 periods ahead, and Y axis are the absolute mean SHAP values. Because each forecast horizon is estimated with different models and parameters, features contributions change between forecast horizons.

Figure 7 illustrates the average contribution of each feature to the forecast of industrial foods. For baskets like, beverages, brownies, sauces, snacks, bread, corn flower and vegetable oils, the main factor for model predictions were international prices. On the other hand, for baskets such as Other flowers, spices, pasta and salt the main factor is the inflation persistence. Seasonal component is also a relevant factor, especially for pulses, sauces and chocolate.

Figure 7: XGBoost SHAP values - Industrial Foods

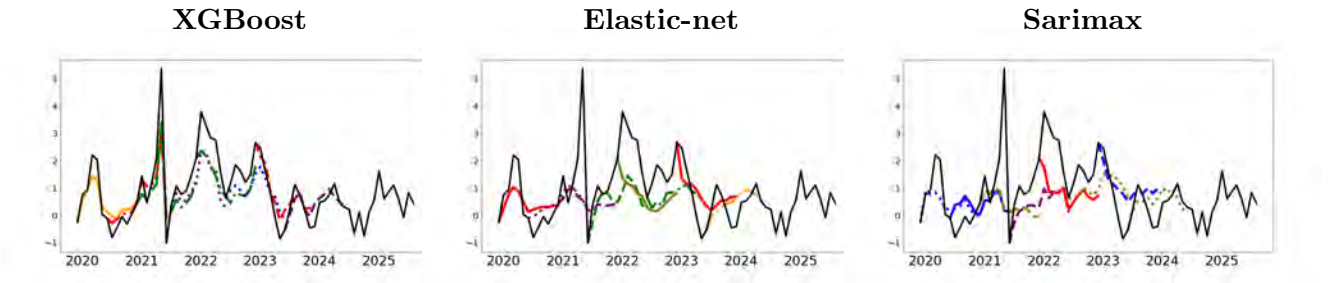


Notes: This figure shows the mean absolute SHAP values of each feature for all the forecasts samples between 2020 and 2024. X axis is the forecast horizon: 1 to 12 periods ahead, and Y axis are the absolute mean SHAP values. Because each forecast horizon is estimated with different models and parameters, features contributions change between forecast horizons.

3.3 Aggregated Food Inflation Basket

As a final forecasting exercise, we aggregate individual forecasts from the 33 food baskets to obtained an aggregated food CPI forecast. Figure 8 shows a visual representation of forecast performance across the three models: XGBoost, Elastic-net, and Sarimax. In addition, Table 5 present the aggregated food inflation forecast errors for the same horizons used to study disaggregated baskets. As in the earlier case, the most accurate model is the XGBoost with forecast errors statistically lower for all forecast steps, except for the first month. On average, the MAFEs from the tree-based model are approximately 25% lower than those from the other forecasting approaches.

Figure 8: Aggregated Food Inflation Forecast



Notes: shows the aggregated inflation forecasts from the expanding window between 2020 and 2024. This is calculated as the weighted average of individual basket forecasts. Black solid line represents the observed month to month food inflation rate. Each line style - color combination represents a different forecast made with information known at the moment of forecasting.

Table 5: Forecast Errors: Aggregated Food Inflation

Model	t+2	t+4	t+7	t+10	t+12
Xgb	0.57	0.76*	0.73*	0.76*	0.75*
Elast	0.68	1.03	0.97	0.99	1.03
Sarimax	0.71	1.02	0.95	0.98	0.93

Notes: This table shows the MAE for the 3 models and for 5 forecast horizons: 2, 4, 7, 10 and 12 months ahead. The * symbol indicates that the XGBoost model has a lower forecast error than the other two models at the 10% significance, based on the Giacomini and White test.

4 Conclusions

Although food prices account for 15% of total CPI index, they have consistently been one of the leading contributors to CPI inflation volatility in Colombia. In addition, food inflation rate has being considerably higher than other goods during a considerable part of the century. This is a crucial problem that affect low income households, farmers and monetary policy authorities. Therefore, forecasting food inflation is a highly relevant topic for policymakers and society in general.

Forecasting food inflation basket is challenging for two reasons. First, food inflation CPI is comprised of a basket of highly heterogeneous prices (including agricultural products, livestock, and processed foods), each influenced by distinct market dynamics. Second, the relationship between food prices and their explanatory variables is often non-linear.

To address the first challenge, this article estimates models for 33 homogeneous food inflation baskets, which together constitute the total food CPI. Studying inflation at this level of granularity helps identify which specific food baskets are driving inflation, leading to more accurate forecasts and enabling targeted policy responses. To tackle the second challenge, this article exploites the flexible structure of tree-based models to improve forecast accuracy for volatile baskets. This type of models prove to be more accurate than linear models for most of the disaggregated baskets, especially for longer forecast horizons. Forecast errors from XGBoost were between 5% to 60% lower than the linear models depending on the basket and the forecast horizon, and for the aggregate inflation basket the errors where 25% lower.

Beyond improving accuracy, the article enhances the interpretability of XGBoost forecasts in three ways. First, it identifies homogenous food baskets and corresponding explanatory variables. They were chosen due to their plausible economic links to production and price dynamics within each inflation basket. Second, it introduces a lag structure optimization algorithm to select lagged features with the highest predictive power, thereby reducing model complexity by reducing the number of explanatory variables. Third, it employs SHAP values to quantify the contribution of each factor to the model's predictions, offering deeper insights into the drivers of food price movements which vary significantly depending on the product.

These results highlight the importance of examining the dynamics of food inflation, given its volatility and its significant impact on overall inflation. We demonstrate how modern machine learning methodologies not only improve predictive accuracy but also enhance our understanding of the underlying patterns driving food prices. The observed patterns provide further evidence supporting the disaggregation of aggregate variables, offering deeper insights beyond improved forecasting performance. Future research should expand the analysis by incorporating other machine learning approaches, such as deep learning models. These models are widely recognized in forecasting literature for their ability to capture complex patterns and may enhance prediction accuracy. Additionally, further work should emphasize the use of explainable AI techniques to better interpret model outputs and ensure that the factors driving predictions are both transparent and robust.

References

Abril-Salcedo, D. S., Melo-Velandia, L. F., & Parra-Amado, D. (2020). Nonlinear relationship between the weather phenomenon el niño and colombian food prices. *Australian Journal of Agricultural and Resource Economics*, *64*(4), 1059–1086.

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

Balogh, J. M., & Sárvári, B. (2024). Evolution of food price inflation in the european union 27 member states. *Journal of Foodservice Business Research*, 1–27.

Birgani, R. A., Kianirad, A., Shab-Bidar, S., Djazayeri, A., Pouraram, H., & Takian, A. (2022). Climate change and food price: a systematic review and meta-analysis of observational studies, 1990-2021. *American Journal of Climate Change*, *11*(2), 103–132.

Chen, T., & Guestrin, C. (2016). Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining* (pp. 785–794).

Dell, M., Jones, B. F., & Olken, B. A. (2014). What do we learn from the weather? the new climate-economy literature. *Journal of Economic literature*, *52*(3), 740–798.

Faust, J., & Wright, J. H. (2013). Forecasting inflation. In *Handbook of economic forecasting* (Vol. 2, pp. 2–56). Elsevier.

Giacomini, R., & White, H. (2006). Tests of conditional predictive ability. *Econometrica*, *74*(6), 1545–1578.

González-Molano, E. R., Gómez, M. I., Melo-Velandia, L. F., Torres, J. L., et al. (2006). Forecasting food price inflation in developing countries with inflation targeting regimes: the colombian case. *Borradores de Economía; No. 409*.

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

Ismaya, B. I., & Anugrah, D. F. (2018). Determinant of food inflation. *Bulletin of Monetary Economics and Banking*, *21*(1), 81–94.

Jenkins, G. M., & Box, G. E. (1976). Time series analysis: forecasting and control. *(No Title)*.

Kohlscheen, E. (2022). Understanding the food component of inflation. *arXiv preprint arXiv:2212.09380*.

Köse, N., & Ünal, E. (2024). The effects of the oil price and temperature on food inflation in latin america. *Environment, Development and Sustainability*, *26*(2), 3269–3295.

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

Martínez-Rivera, W., González-Molano, E., & Caicedo-García, E. (2023). Forecasting inflation from disaggregated data: The colombian case. *Borradores de Economía; No. 1251*.

Melo-Velandia, L. F., Orozco-Vanegas, C. A., & Parra-Amado, D. (2022). Extreme weather events and high colombian food prices: A non-stationary extreme value approach 1. *Agricultural Economics*, *53*(S1), 21–40.

Milunovich, G. (2020). Forecasting australia's real house price index: A comparison of time series and machine learning methods. *Journal of Forecasting*, *39*(7), 1098–1118.

Petropoulos, F., Apiletti, D., Assimakopoulos, V., Babai, M. Z., Barrow, D. K., Taieb, S. B., ... others (2022). Forecasting: theory and practice. *International Journal of forecasting*, *38*(3), 705–871.

Rossi, B. (2013). Advances in forecasting under instability. In *Handbook of economic forecasting* (Vol. 2, pp. 1203–1324). Elsevier.

Smalter Hall, A., & Cook, T. R. (2017). Macroeconomic indicator forecasting with deep neural networks. *Federal Reserve Bank of Kansas City Working Paper*(17-11).

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, *58*(1), 267–288.

Tirivarombo, S., Osupile, D., & Eliasson, P. (2018). Drought monitoring and analysis: standardised precipitation evapotranspiration index (spei) and standardised precipitation index (spi). *Physics and Chemistry of the Earth, Parts a/b/c*, *106*, 1–10.

Zárate-Solano, H. M., & Rodríguez-Niño, N. (2024). Consumer prices trends in colombia: Detecting breaks and forecasting inflation.

Zou, H., & Hastie, T. (2005). Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, *67*(2), 301–320.

A1 Basket elements and weights

Table A.1: Baskets Elements and Weights

Basket	**Items**	**Weight (%)**
Beef	Beef and derivatives	12.5
Beverages	Sodas and malt beverages for home consumption; tea and infusions; mineral water; packaged liquid refreshments; concentrates for refreshments; energy drinks	5.1
Bread	Bread	4.5
Brownies and cookies	Brownies and soda cookies	1.8
Carrot	Carrot	0.3
Cassava	Cassava for home consumption	0.7
Chicken	Poultry	8.2
Chocolate Drink	Chocolate and chocolate-based products	1.2
Coffee	Coffee and coffee-based products	1.4
Cold Meats	Prepared meats, cold cuts and other meat-containing products	2.7
Corn Flour	Oats, wheat, corn and derivatives	0.8
Onion	Onion	1.4
Culinary Herbs	Condiments and culinary herbs	0.4
Dairy	Milk, cheese and other dairy products	11.9
Eggs	Eggs	4.50
Fruits	Fruits	6.6
Instant Soups	Soups, creams, broths and consommés	0.4
Other flours and cereals	Other flours, cereals and starches	2.0
Other Vegetables	Fresh vegetables and legumes	2.3
Pasta	Pasta	0.8
Plantain	Plantain	2.1
Potato	Potato	2.0
Pork	Pork and derivatives	3.2
Pulses	Dried legumes; canned and dehydrated legumes and vegetables	2
Rice	Rice	6.2
Salt and Nuts	Salt; other precooked and prepared foods; dried fruits and nuts, fruit peels and edible seeds	0.4
Sauces and Pastes	Sauces, pastes and dressings	1.1
Seafood	Freshwater fish, sea fish, and related products	4.2
Snacks and Candy	Snacks for home consumption; ice cream, sweets, jams, jellies and similar	2.0
Sugar	Sugar and other sweeteners; raw panela for home consumption	2.7
Tomato	Tomato	1.2
Yam	Products derived from tubers, roots and plantains; arracacha, yam and other tubers	0.2
Vegetable Oil	Oils and fats	3.5

Notes: This table presents the items included in the 33 inflation baskets, along with the corresponding food CPI weights used for aggregation.

A2 Features used for each inflation basket

Table A.2: Food Inflation Baskets Features

Basket Group	Basket	Commodity Price	Energy Costs	Weather
Agricultural Crops	Carrot	European urea	Diesel	SPEI 6
	Cassava	European urea	Diesel	SPEI 6
	Fruits	European urea	Diesel	SPEI 6
	Onion	European urea	Diesel	Atypical temperature
	Other vegetables	European urea	Diesel	SPEI 2
	Plantain	European urea	Diesel	SPEI 12
	Potatoes	European urea	Diesel	Atypical temperature
	Tomato	European urea	Diesel	Precipitation excess
	Other Tubers	European Urea	Diesel	Atypical temperature
	Coffee	Mild arabicas coffee price; European urea		SPEI 12
	Rice	Thailand white rice 5%		Dry days excess
	Sugar	US raw sugar import price		Dry days excess
Animal Products	Beef	US Pork swine leanhogs; US yellow maize No2; European urea	Diesel	Atypical precipitation
	Chicken	US Pork swine leanhogs; US yellow maize No2	Diesel	Atypical temperature
	Dairy	US Whole milk powder; CME soybean meal	Diesel	Atypical precipitation
	Eggs	US yellow maize No2	Diesel	Atypical temperature
	Pork	CME soybean meal; US Pork swine leanhogs	Diesel	SPEI 6
	Seafood	Exchange rate		ENSO
Industrial Food	Bread	US wheat No1 HRW; Malaysian palm oil; Dry whole milk powder	Energy price	
	Brownies and cookies	US raw sugar import price; US wheat No1 HRW; Malaysian palm oil; US Dry whole milk powder	Energy price	
	Candy and snacks	US raw sugar import price; ICCO cocoa beans; Malaysian palm oil	Energy Price	
	Chocolate drink	ICCO cocoa beans		
	Cold meats	US Pork swine leanhogs	Energy price	
	Pasta	US wheat HRW No1	Energy Price	
	Pulses	Canadian grains and specialty crops PPI		
	Maize, wheat and oats flours	US yellow maize No2; US wheat HRW No1	Energy Price	
	Non-alcoholic beverages	Canadian Barley No1; US raw sugar import price	Energy price	
	Other flours and cereals	US oats PPI		
	Sauces and pastes	Malaysian palm oil; US raw sugar import price	Energy Price	
	Culinary herbs			
	Instant Soups			
	Salt and Nuts			
	Vegetable oil	Soybean oil		

Notes: All baskets include the nominal exchange rate COP/USD to capture the general pass-through on goods. Beef inflation includes cow sacrifices to capture short run supply dynamics. Seafood includes ENSO index used to determined El niño and La niña weather events. CME stands for Chicago Mercantile Exchange. Check Appendix A3 to see the definition of weather variables.

A3 Weather Variables

This section outlines the procedure used to identify the most suitable weather predictor for each inflation basket. We begin by processing temperature and precipitation gridded data to get the daily mean at the Municipality-year-month level. Temperature data comes from the Earth Reanalysis fifth generation (ERA5) by the Copernicus Climate Change Service. They provide a reanalysis-type measure of air temperature at 2 meters above land surface, available in gridded format at a 0.25° × 0.25° resolution with hourly frequency since 1979. Precipitation data comes from The Climate Hazards Center Infrared Precipitation with Stations (CHIRPS) by the Climate Hazards Group. They produce a 0.05° gridded rainfall time series over land with daily frequency since 1981.

We expect deviations from the historical weather patterns of these variables to influence food inflation. For temperature, we rely on percentiles of its distribution to identify atypical events. We first compute percentiles 20th and 80th of daily temperature in each Municipality-quarter from 1979-01-01 to 2025-09-30. Then, we compute our measure of atypical temperature as the total number of days the temperature was above the 80th percentile or below the 20th percentile for each year-month.

To identify days with atypical precipitation shortages, we adopt an approach based on consecutive dry days, recognizing that-unlike temperature—precipitation is bounded at zero. We begin by calculating the number of consecutive dry days for each day throughout the sample period. For each month, we then determine the 80th percentile of this distribution to serve as a benchmark for extreme precipitation shortages. Using this threshold, we compute our indicator as the average number of dry days exceeding the monthly 80th percentile. For excess rainfall, we apply the same percentile-based method used for temperature. Also, to capture atypical precipitation events—whether due to excess or shortage—we use the maximum value between the two indicators.

Next, we compute the SPEI for each municipality-year-month using the monthly temperature and precipitation, and the average altitude in the municipality. The SPEI indicates deviations from the usual water balance for some time period in the municipality (Tirivarombo et al., 2018). *Water balance* refers to the interaction between precipitation and temperature that determines the lack or excess of water in the location. We consider as reference periods for these water balance 1,2,3,6 and 12 months. SPEI values range from -3 to 3, with positive values indicating water excess and negative values indicating water deficit.

In summary, we build 5 weather indicators of disturbances to the usual weather patterns at the Municipality-year-month level:

1. Atypical temperature: Number of days with atypical temperature where *atypical* refers to values above the 80th of below the 20th percentile of the historic distribution.

2. Precipitation excess: Number of days with precipitation above the 80th of the historic distribution.

3. Precipitation shortage: Average number of days exceeding the the 80th percentile of consecutive dry days.

4. Atypical precipitation: Maximum between precipitation excess and precipitation shortage.

5. SPEI: Number of standard deviations from the mean climate water balance values assessed for the reference period.

Subsequently, we expect atypical weather conditions to have a greater impact on food prices in regions where the corresponding food is more intensively produced. To account for this, we aggregate the monthly weather variables using Equation A.1, where W_{tj} represents the weather variable in year-month t for municipality j, and α_{mjb} denotes the production share of basket b in municipality j relative to national production in month m. The resulting variable, W_{tb}, captures the weather indicator for each basket at the national level.

$$W_{t,b} = \sum_{j}^{J} \alpha_{mjb} W_{tj} \tag{A.1}$$

These production shares are derived from four surveys that report provision or production of food: SIPSA, ESAG, EVA, and USPLECHE. The first two are administered by DANE, while the latter two are managed by the Ministry of Agriculture (MADR). The SIPSA and EVA surveys are conducted at the municipal level, allowing for the direct use of both municipal production shares and municipal-level weather variables. In contrast, ESAG and USPLECHE are conducted at the *departamento* level, meaning that production shares for beef, pork, and dairy are calculated at that level of aggregation. To match this level of aggregation, the monthly weather variables—originally available at the municipal level—are averaged across all municipalities within each *departamento* to obtain *departmento*-level variables.

As a final step, we run a series of linear regressions to identify the most relevant weather indicator for each inflation basket. Each basket's inflation series is regressed on individual weather indicators and their squared terms, following the specification in Equation A.2, where π denotes the inflation basket and W the weather variable. To account for the potential lagged effects of weather on production and to align with the SPEI format, we include lags of 1, 2, 3, 6, and 12 months for both temperature and precipitation variables. We select the weather variable that maximizes model fit, as measured by the R-squared.

$$\pi_t = \beta_1 W_t + \beta_2 W_t^2 + \gamma_{year(t)} + \delta_{month(t)} + \varepsilon_t \tag{A.2}$$

Table A.3 presents the final selection of weather variables used for predicting food inflation.:

Table A.3: Best weather indicator selected for each crop

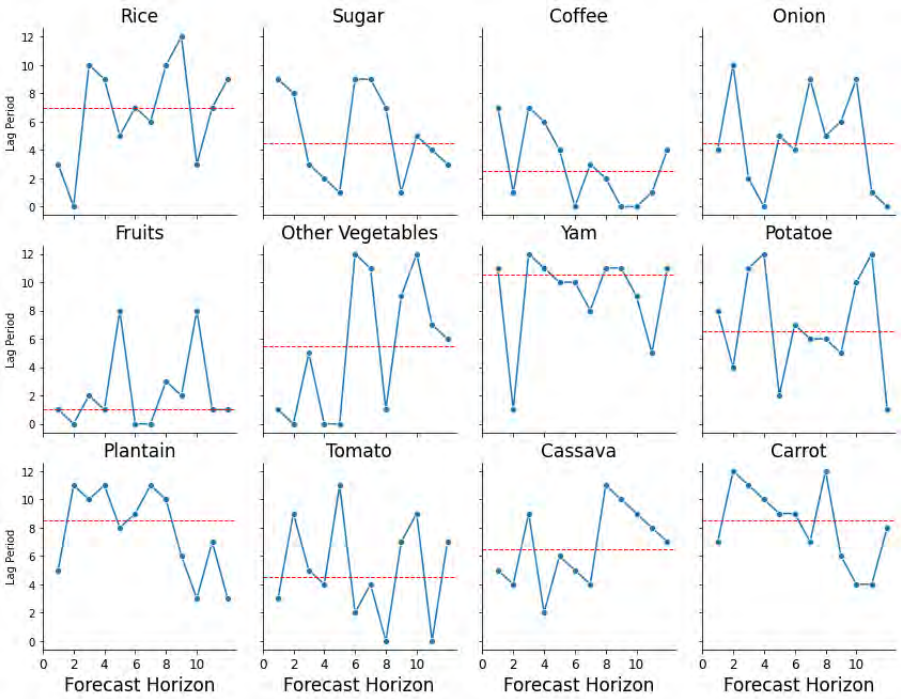
Crop	**Best Weather Indicator**
Sugar	Dry days excess (t-2)
Rice	Dry days excess (t-6)
Tomato	Precipitation excess (t-2)
Beef	Atypical precipitation
Dairy	Atypical precipitation (t-3)
Coffee	SPEI 12
Plantain	SPEI 12
Other vegetables	SPEI 2
Cassava	SPEI 6
Fruits	SPEI 6
Carrot	SPEI 6
Pork	SPEI 6
Eggs	Atypical temperature (t)
Onion	Atypical temperature (t-2)
Potato	Atypical temperature (t-2)
Other tubers	Atypical temperature (t-3)
Chicken	Atypical temperature (t-6)

Notes: *Dry days excess* refers to precipitation shortage; *atypical precipitation* refers to days with excess or shortage of precipitation; *atypical temperature* refers to days with atypical high or low temperature. *(t-i)* refers to the lag i of the variable.

A4 Lag selection

Following Algorithm 1 we selected the optimal lag for each feature based on forecast accuracy on the validation set previously mentioned. While this approach enhances predictive performance, the selected lags vary across horizons and can be volatile. This variability reflects the changing relevance of past information depending on the forecast horizon, but it also introduces heterogeneity and complexity in interpreting the model's behavior across time. Figure A.1 shows the selected lags for each forecast time step for the agricultural product weather variables (blue line) and the median across horizons (orange line). X axis correspond to the forecast horizon and Y axis correspond to the selected lag used in each of the 12 models to forecast the 12 steps ahead. Fruits, cassava, potatoes and onions exhibit an upward trend. Recent lags are selected for short run forecast, while older lags are selected for distant forecast. On the other hand, sugar, coffee and plantain exhibit a downward trend. Older lags are most important for short run forecast and then more recent lags are selected for the last 3 periods. Moreover, rice and tomatoes exhibit a U-shape relationship between lags and forecast horizons. Medium run forecast steps used older lags, while short run and last steps use recent lags. The selected lags for the rest of the features also show volatility across the forecast horizon.[17]

Figure A.1: XGBoost optimal lags of weather variables in agricultural food products



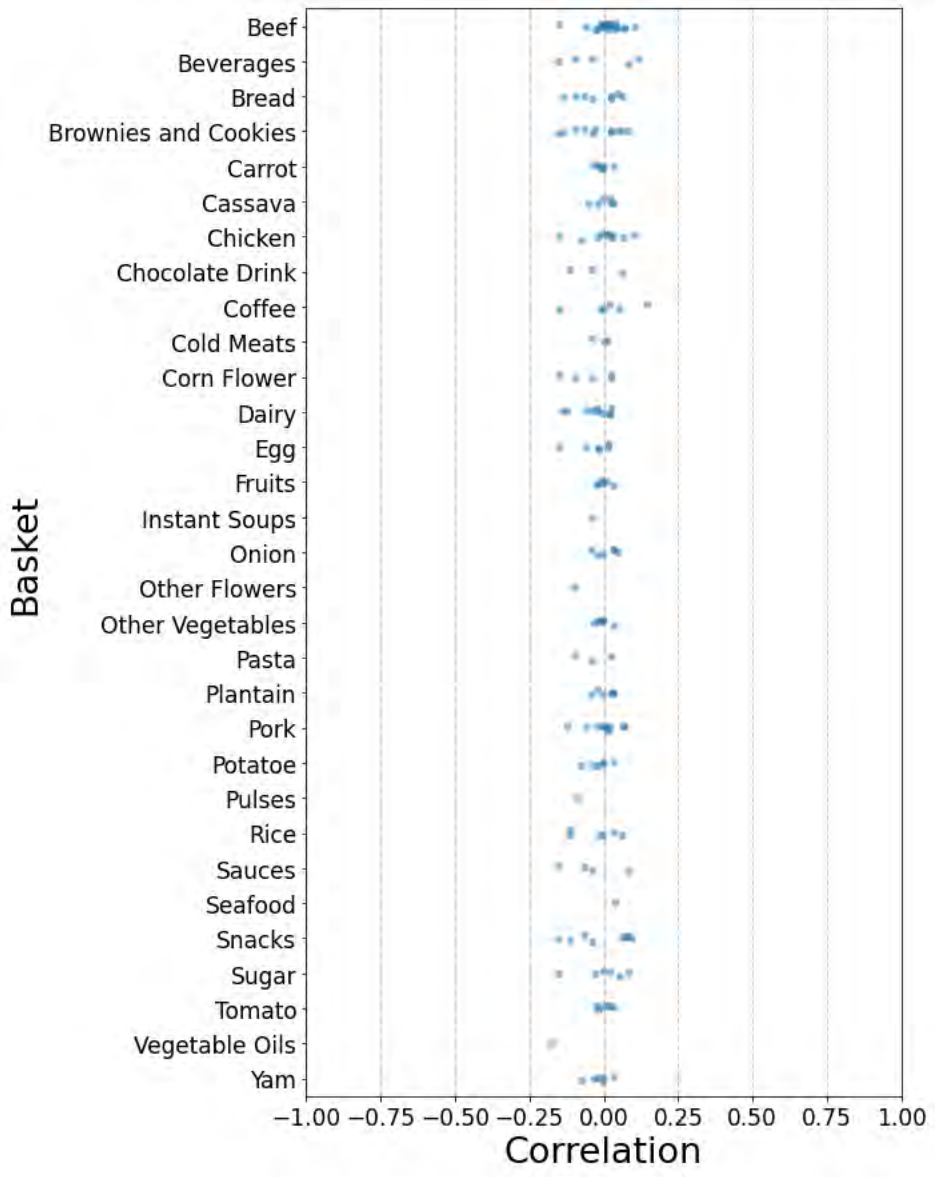
Notes: This figure correspond to the selected lags for the weather variable in the agricultural group. X axis correspond to the forecast horizon: 1 to 12 months ahead and Y axis corresponds to optimal lags selected for each forecast horizon: 1 to 12 months behind. Each foreast horizon uses a different model to forecast inflation rates with different parameters and lag-features combinations.

[17]The rest of the selected lags are not included in the document because they present a similar behavior.

A5 Features cross correlations

Figure A.2 presents the pairwise correlations between features within each inflation basket. All correlation values fall between -0.20 and 0.20, with most clustering near zero. This low level of cross-correlation indicates that the features are largely independent, which enhances the precision of SHAP value estimation. Each feature is selected for its exogeneity and minimal overlap with others, reducing the risk of misleading attributions and strengthening the reliability of the SHAP-based interpretation.

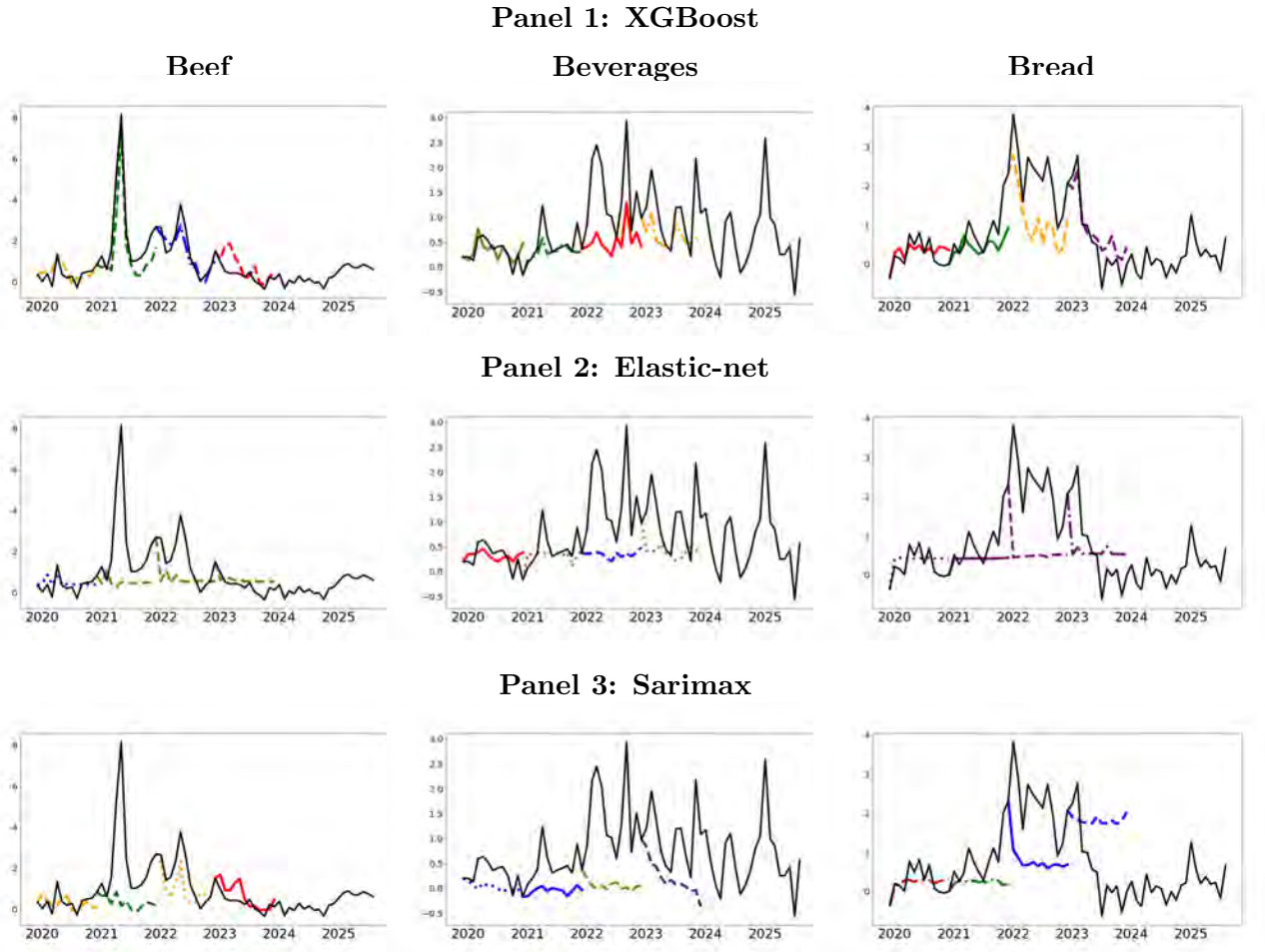
Figure A.2: Features cross correlations



Notess: This figure shows the cross correlations between features for each inflation basket. Each point represents the correlation value between 2 features.

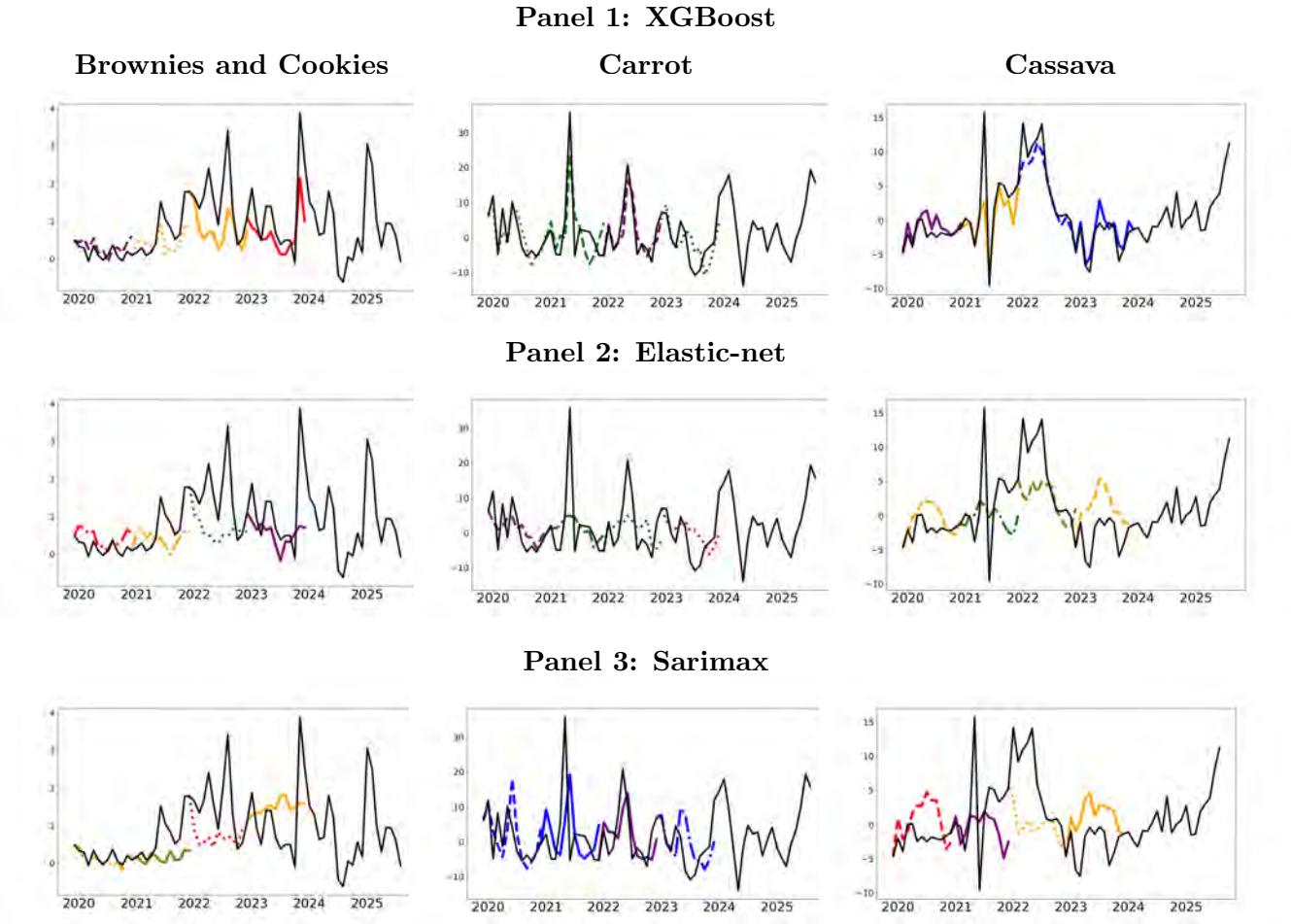
A6 Forecast Rolling Window

Figure A.3: Expanding window forecast



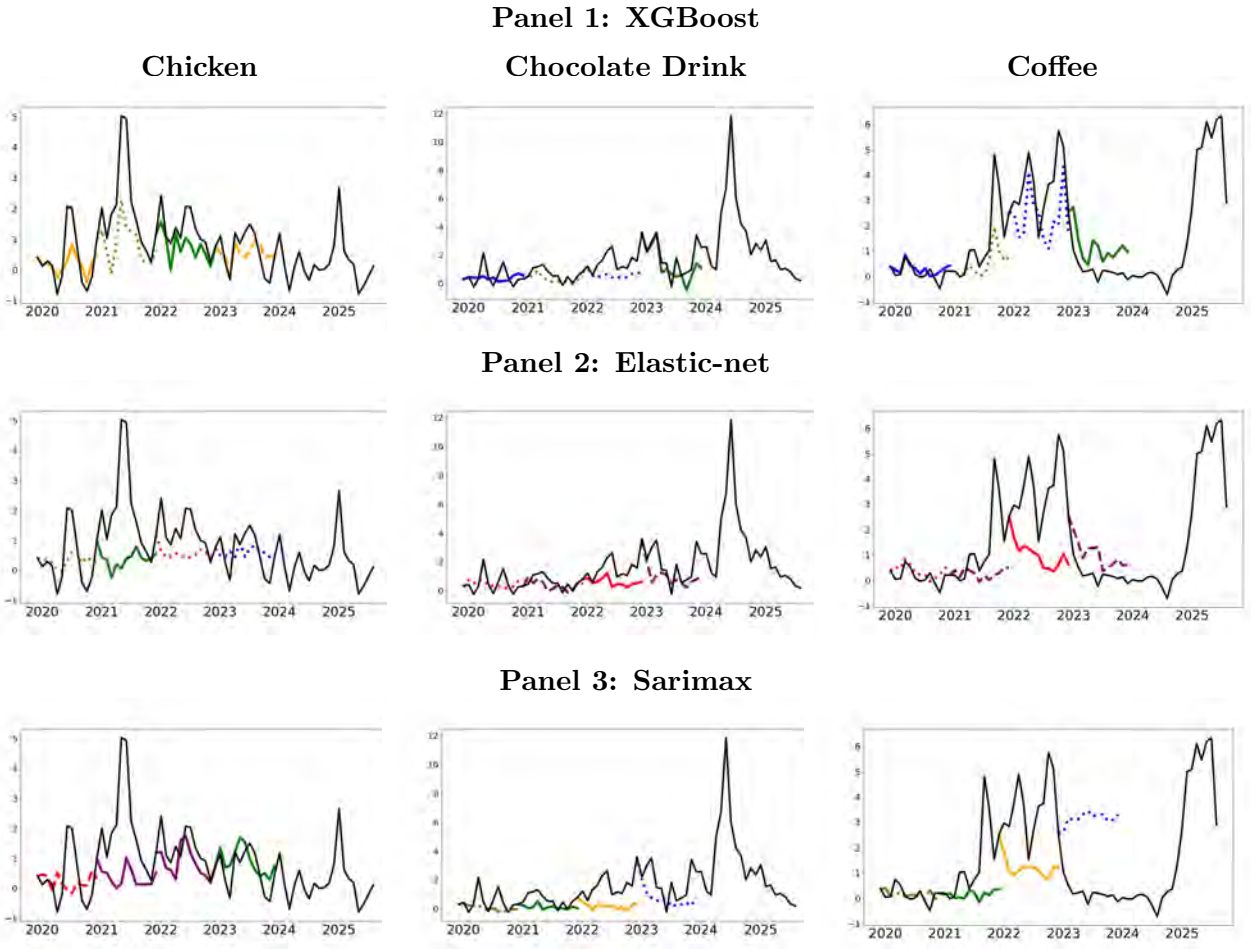
Notes: shows an expanding window forecast use to evaluate the models predictive performance between 2020 and 2024. Black solid line represents the observed month to month inflation rate. Each line style - color combination represents a different forecast made with information known at the moment of forecasting.

Figure A.4: Expanding window forecast



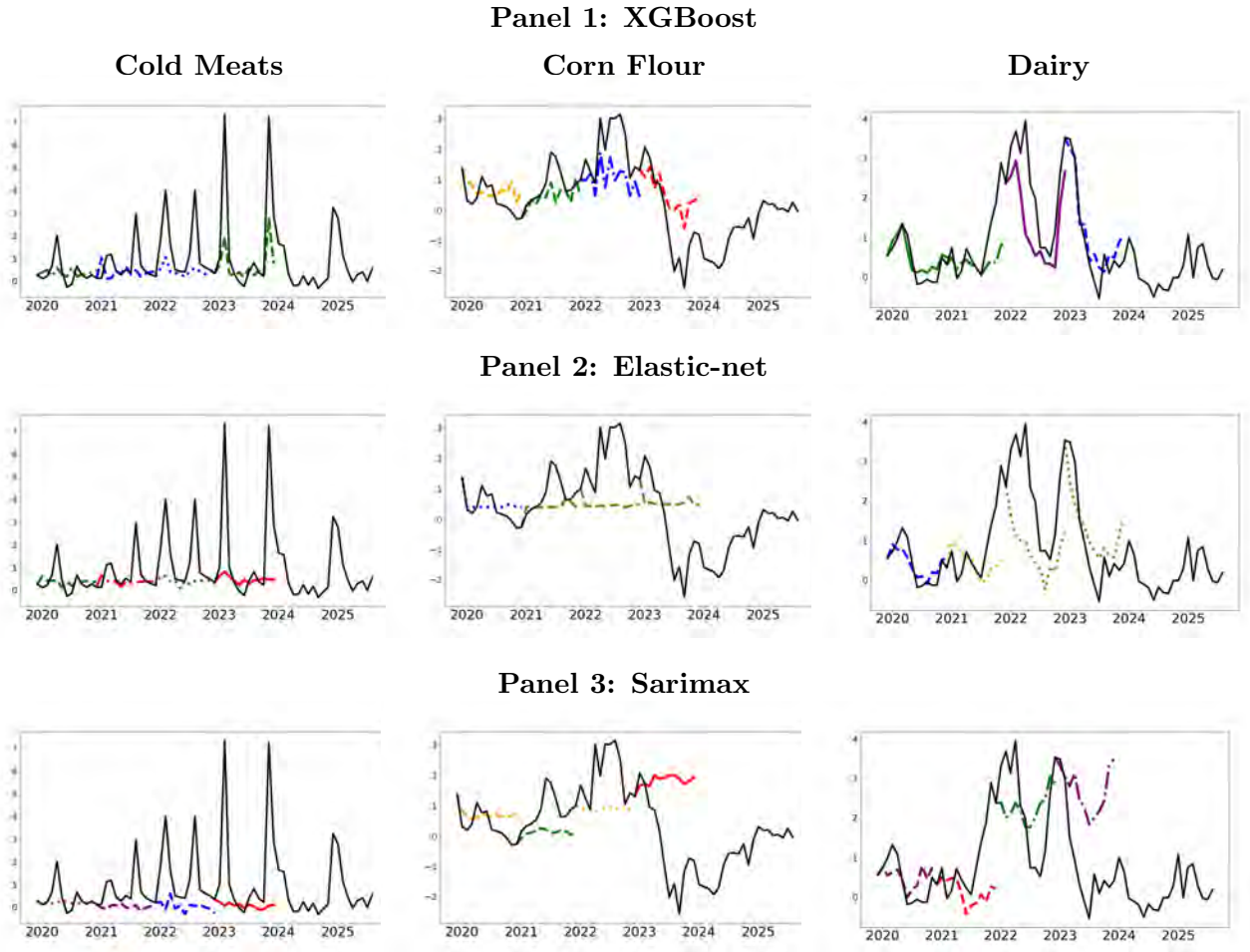
Notes: shows an expanding window forecast use to evaluate the models predictive performance between 2020 and 2024. Black solid line represents the observed month to month inflation rate. Each line style - color combination represents a different forecast made with information known at the moment of forecasting.

Figure A.5: Expanding window forecast



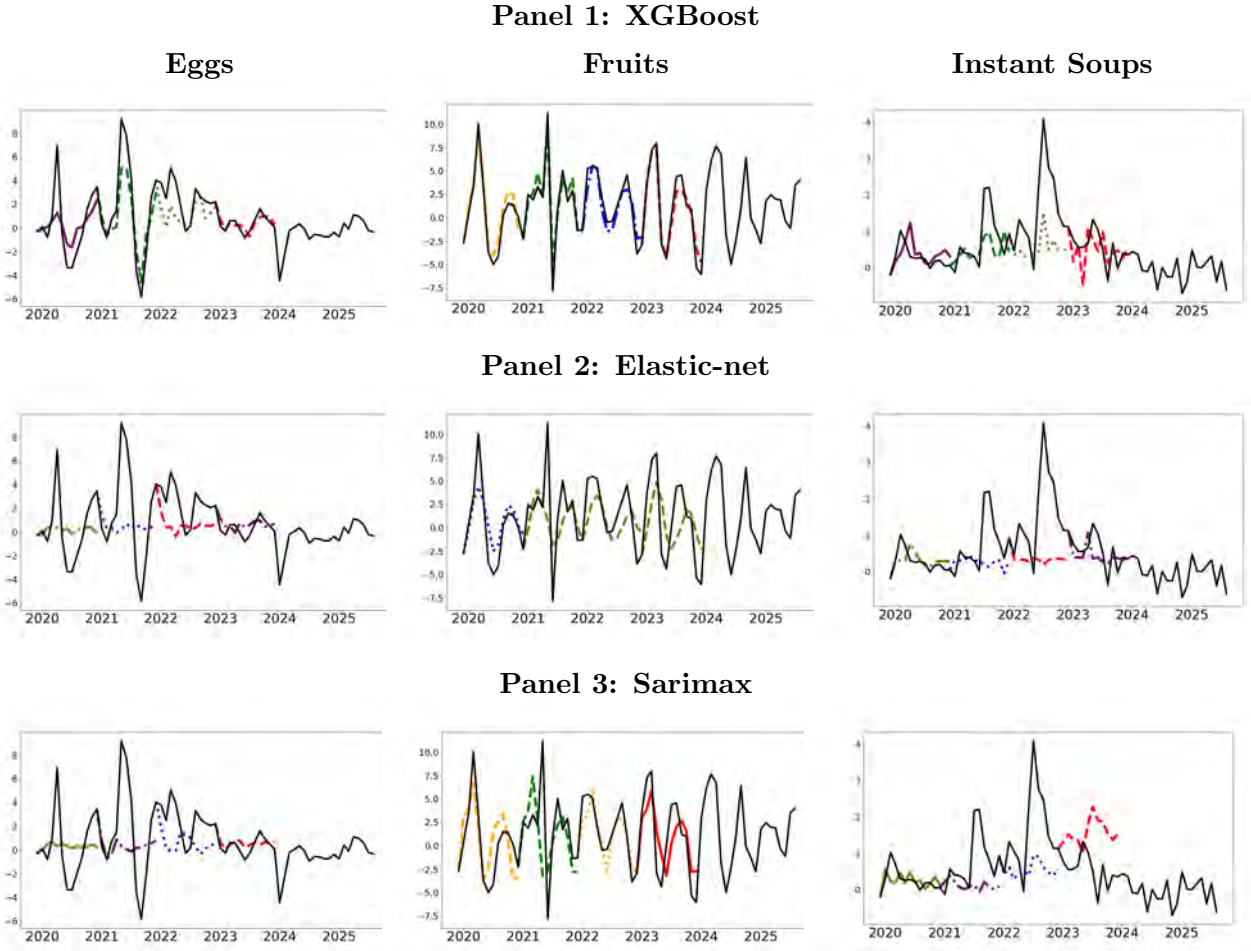
Notes: shows an expanding window forecast use to evaluate the models predictive performance between 2020 and 2024. Black solid line represents the observed month to month inflation rate. Each line style - color combination represents a different forecast made with information known at the moment of forecasting.

Figure A.6: Expanding window forecast



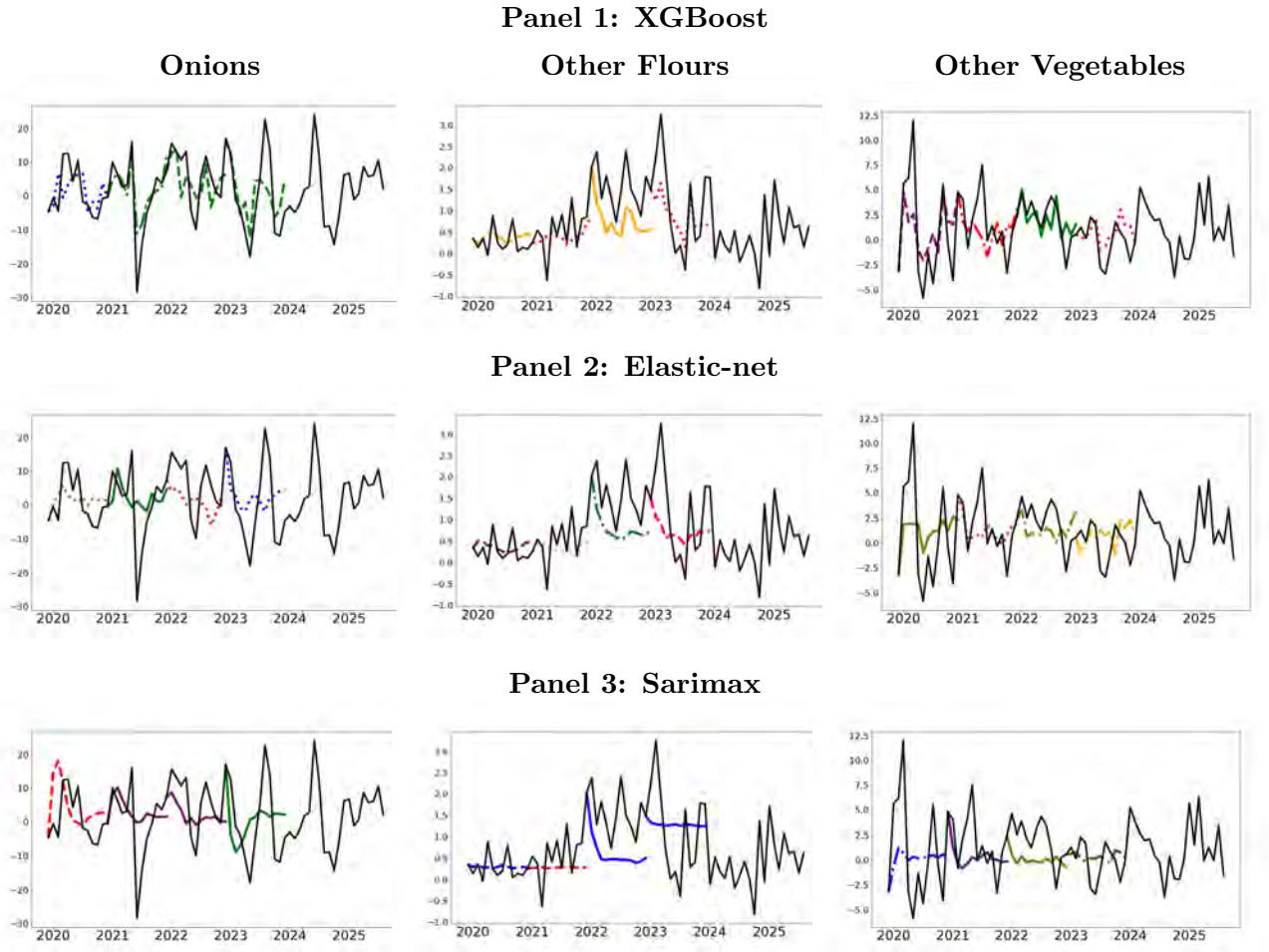
Notes: shows an expanding window forecast use to evaluate the models predictive performance between 2020 and 2024. Black solid line represents the observed month to month inflation rate. Each line style - color combination represents a different forecast made with information known at the moment of forecasting.

Figure A.7: Expanding window forecast



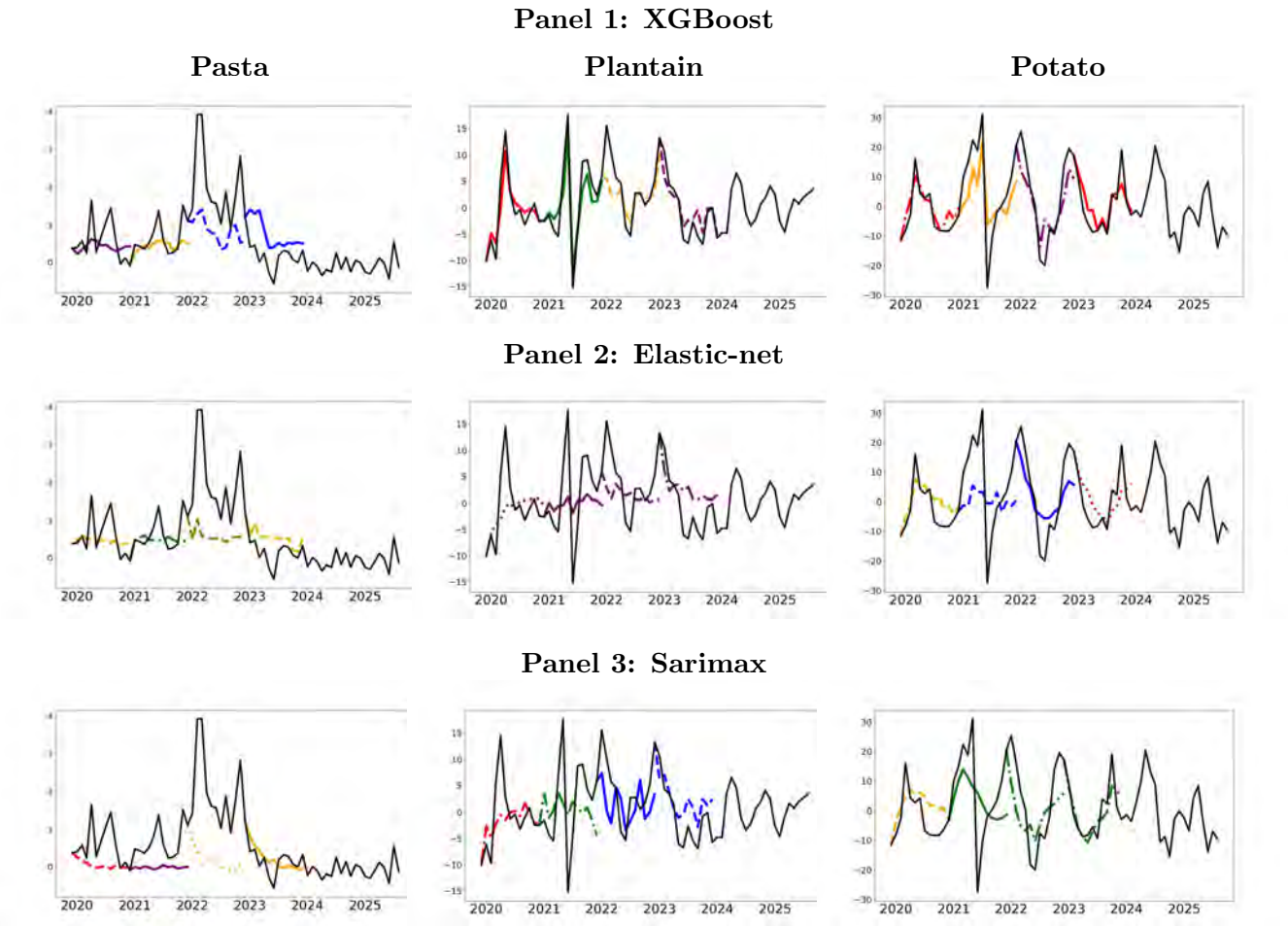
Notes: shows an expanding window forecast use to evaluate the models predictive performance between 2020 and 2024. Black solid line represents the observed month to month inflation rate. Each line style - color combination represents a different forecast made with information known at the moment of forecasting.

Figure A.8: Expanding window forecast



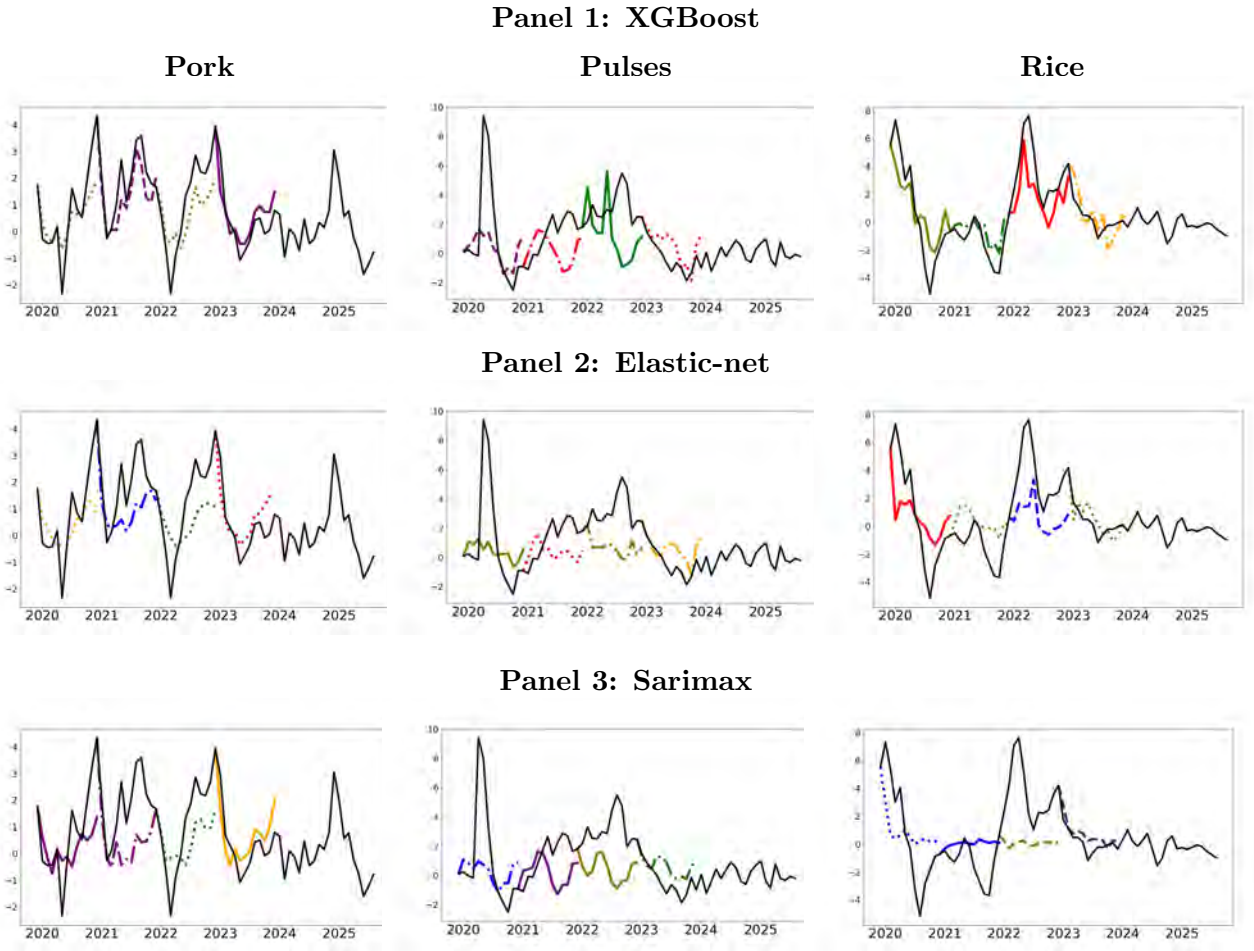
Notes: shows an expanding window forecast use to evaluate the models predictive performance between 2020 and 2024. Black solid line represents the observed month to month inflation rate. Each line style - color combination represents a different forecast made with information known at the moment of forecasting.

Figure A.9: Expanding window forecast



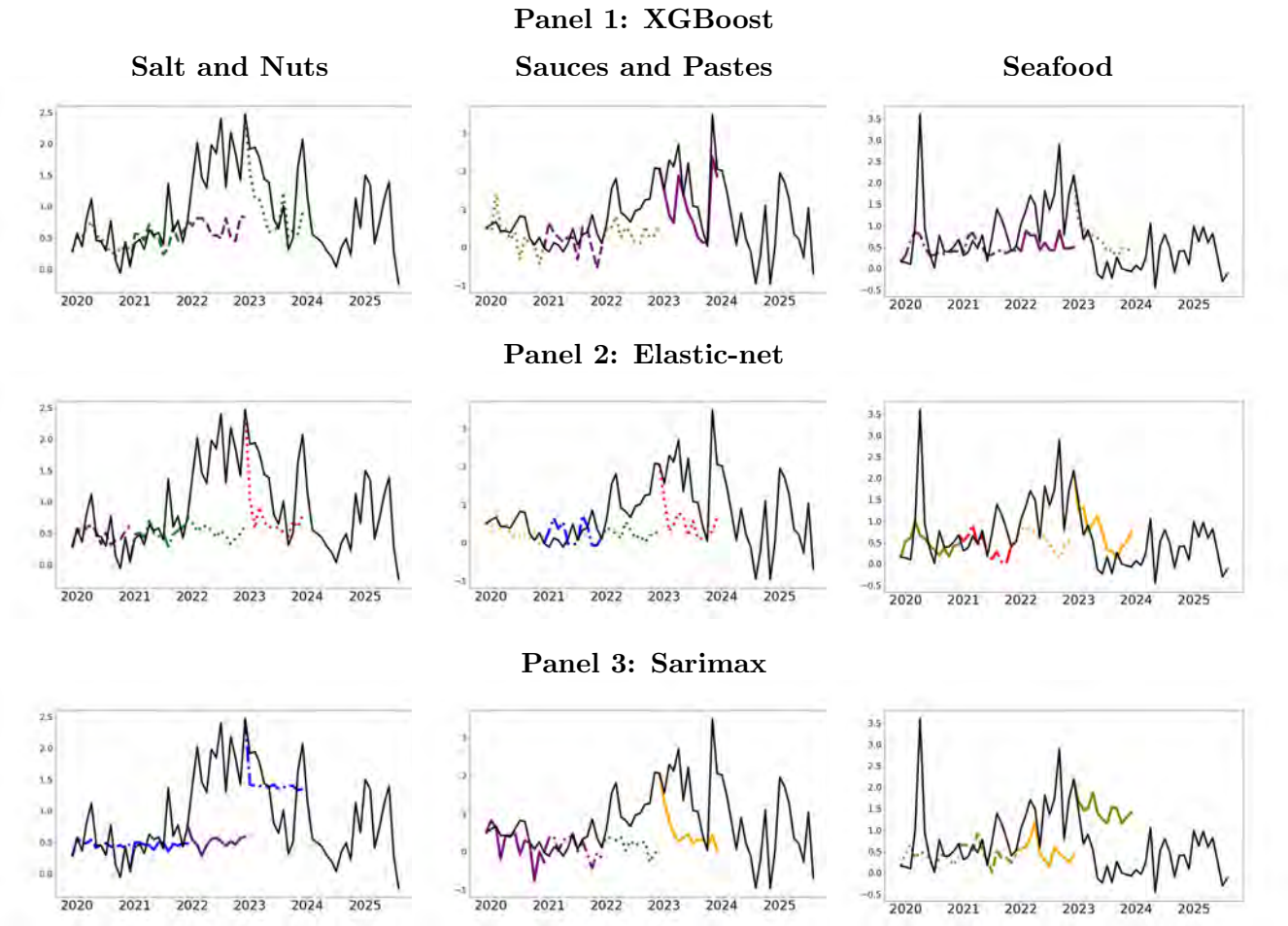
Notes: shows an expanding window forecast use to evaluate the models predictive performance between 2020 and 2024. Black solid line represents the observed month to month inflation rate. Each line style - color combination represents a different forecast made with information known at the moment of forecasting.

Figure A.10: Expanding window forecast



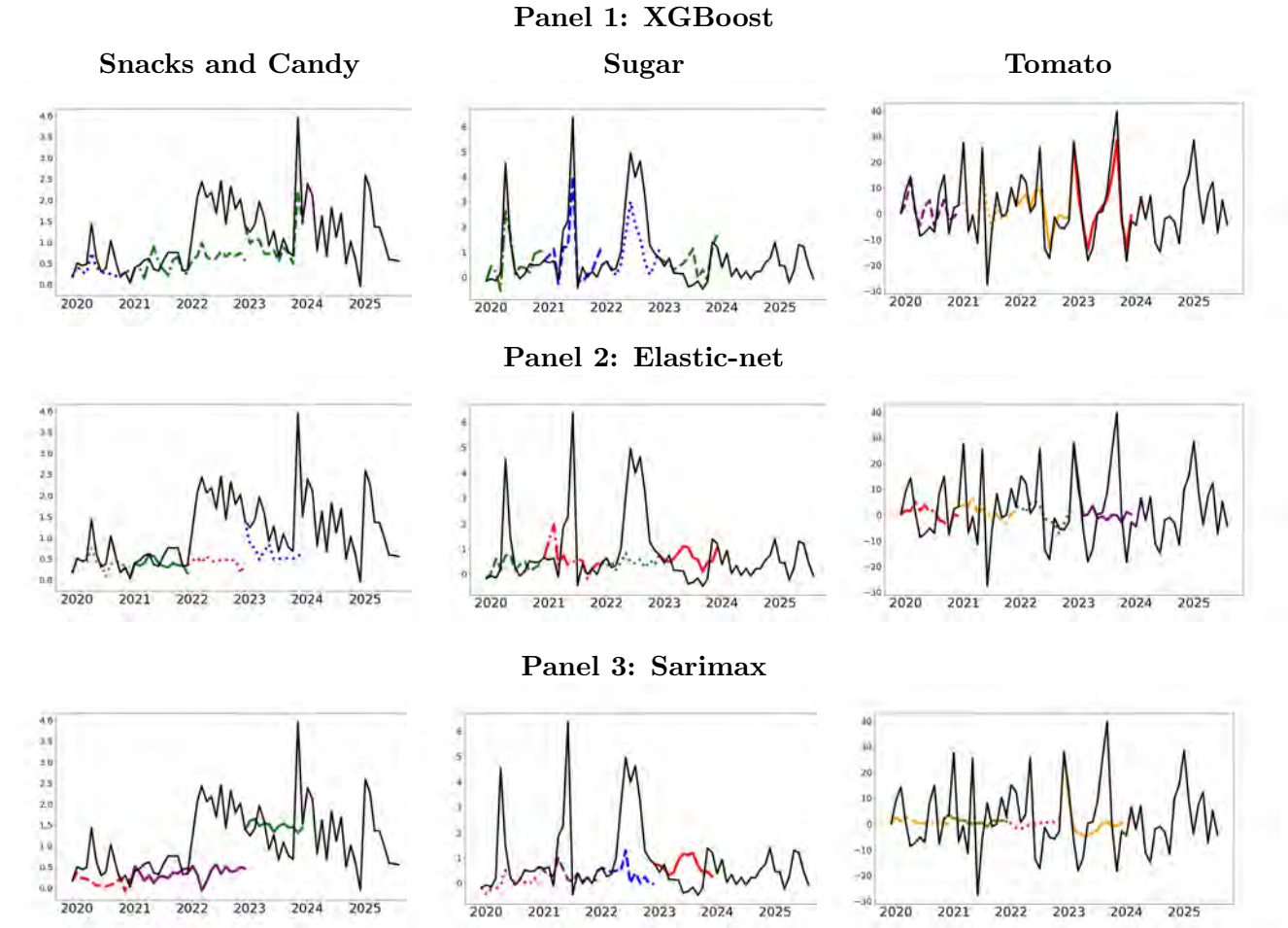
Notes: shows an expanding window forecast use to evaluate the models predictive performance between 2020 and 2024. Black solid line represents the observed month to month inflation rate. Each line style - color combination represents a different forecast made with information known at the moment of forecasting.

Figure A.11: Expanding window forecast



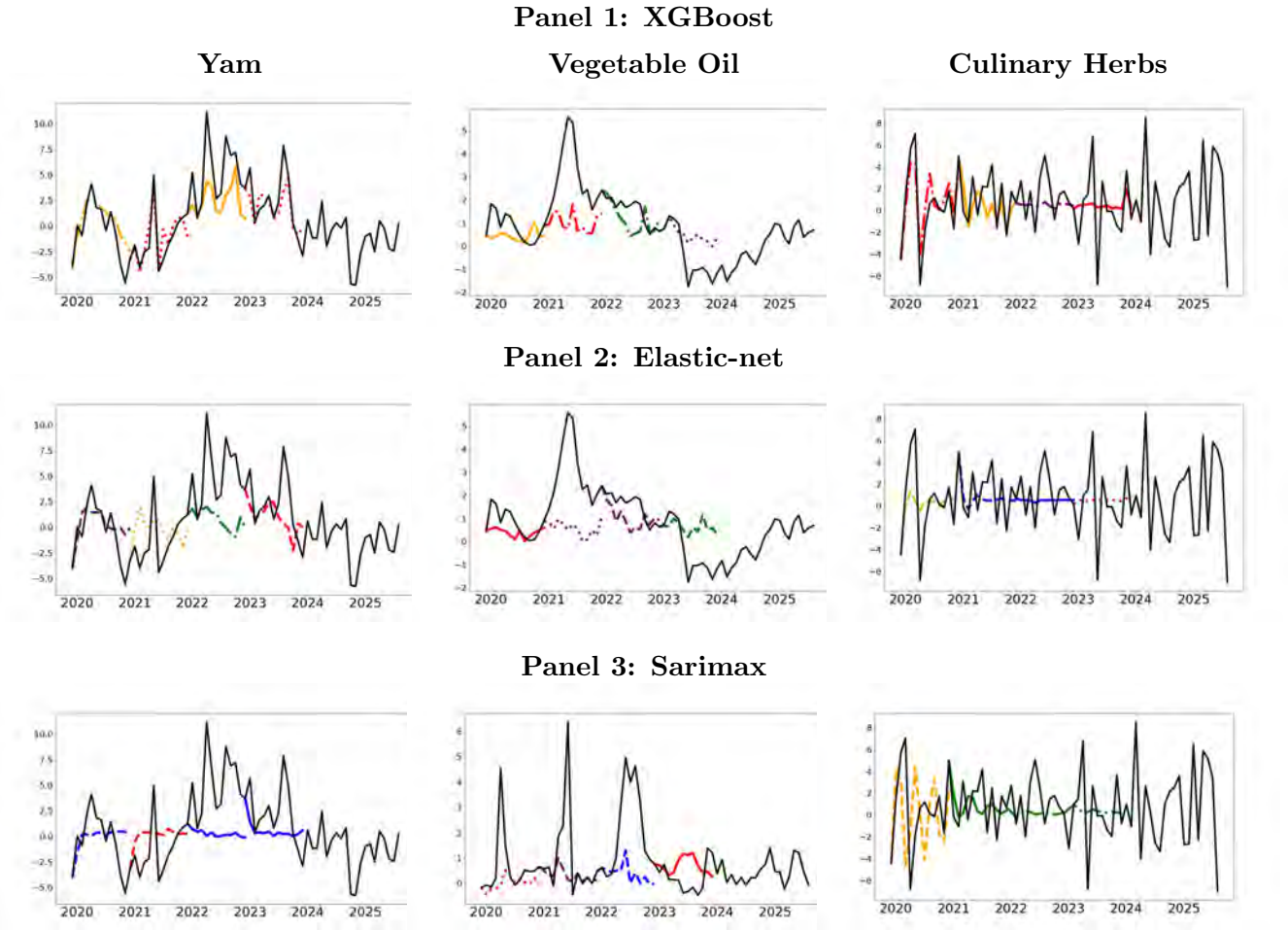
Notes: shows an expanding window forecast use to evaluate the models predictive performance between 2020 and 2024. Black solid line represents the observed month to month inflation rate. Each line style - color combination represents a different forecast made with information known at the moment of forecasting.

Figure A.12: Expanding window forecast



Notes: shows an expanding window forecast use to evaluate the models predictive performance between 2020 and 2024. Black solid line represents the observed month to month inflation rate. Each line style - color combination represents a different forecast made with information known at the moment of forecasting.

Figure A.13: Expanding window forecast



Notes: shows an expanding window forecast use to evaluate the models predictive performance between 2020 and 2024. Black solid line represents the observed month to month inflation rate. Each line style - color combination represents a different forecast made with information known at the moment of forecasting.