Economics School of Louvain - ESL

Inflation Forecasting in Spain

Combining Machine Learning with Alternative Data Sources

Author : Jauder Uribarri Sánchez

Thesis Director : Francesca Monti

Thesis Reader : Sébastien Van Bellegem

Academic Year 2024 2025

Master in Economics – 120 credits – Econometrics Focus

Université catholique de Louvain - Economics School of Louvain - Place Montesquieu 3, 1348 Louvain-La-Neuve

Abstract

This thesis presents a machine learning framework that forecasts Spanish inflation using multiple measures, including monthly, quarterly, annual, and core indices. This approach integrates traditional macroeconomic indicators with alternative data sources, such as real estate transactions, commodity prices, internet search intensity, and sentiment indices. A systematic feature engineering pipeline generates lagged predictors, and model interpretability is ensured through SHAP values, providing a transparent understanding of inflation drivers.

LightGBM consistently emerges as the strongest performer among the competing algorithms, with results highlighting the central importance of lagged values in capturing inflation dynamics. Benchmarking exercises are carried out against the Bank of Spain's macroeconomic projections and random walk models. While the proposed framework outperforms institutional forecasts at certain time horizons, the random walk benchmark remains difficult to surpass, highlighting the challenges of inflation forecasting.

Overall, the results show that combining conventional and unconventional information enhances the predictive ability of inflation models. In addition to improving predictive accuracy, using interpretable machine learning methods provides valuable insights into the role of various economic and financial drivers. This dual focus on accuracy and interpretability renders the framework relevant for both academic research and policymakers responsible for anticipating and responding to changes in price dynamics.

1 Introduction

Understanding and forecasting inflation dynamics has long been a central concern for policymakers, researchers, and financial institutions. Inflation influences critical economic decisions, including monetary policy, wage negotiations, investment strategies, and broader macroeconomic stability. In Spain, inflation has fluctuated significantly in recent decades due to both domestic conditions and external shocks. Periods of stability have alternated with periods of heightened volatility, often driven by fluctuations in energy prices, food market prices, and international crises. These dynamics highlight the challenge of identifying reliable predictors of price growth and the limitations of traditional models in fully capturing the underlying drivers.

Traditional approaches to inflation forecasting have relied heavily on macroeconomic indicators such as output, wages, and monetary aggregates. Yet the growing availability of non-traditional data — including internet search intensity, sentiment indicators, and high-frequency market information — offers new opportunities to enhance predictive performance. Machine learning methods, with their capacity to handle high-dimensional and heterogeneous datasets, provide a natural complement to standard econometric techniques by exploiting these richer inputs.

This thesis develops a comprehensive dataset for Spain that integrates both conventional macroeconomic information and alternative sources. Modern predictive models are then applied to forecast inflation across multiple horizons. The framework incorporates lagged variables, feature relevance through SHAP[1] values, recursive forecasting procedures, and systematic benchmarking against established references such as the *Banco de España* projections and random walk models. By considering monthly, quarterly, annual, and core CPI, the analysis provides a broad assessment of how different components of inflation evolve and how machine learning models perform in anticipating them.

2 Literature Review

Traditional euro-area inflation forecasting has long relied on reduced-form Phillips curve models based on macroeconomic theory. For example, Bobeica et al. (2020) examines a variety of Phillips curve vector autoregressions (VAR) and autoregressive distributed lag (ADL) specifications. They demonstrate that, although model performance is unstable over time, certain specifications outperform a univariate benchmark. Averaging these specifications provides more robust forecasts, especially during volatile periods. Building on this, Consolo, Cristadoro, and Forni (2022) proposes a bottom-up approach that estimates augmented Phillips curves for subcomponents of the Harmonized Index of the Consumer Prices (HICP). This approach shows that the disaggregated framework improves forecasting accuracy, particularly during the shocks caused by the pandemic and rising energy prices. It also highlights the role of international energy and food prices as key drivers of recent inflation dynamics. Similarly, Capolongo and Pacella (2021) develop a Bayesian panel vector autoregression (VAR) model that integrates key inflation drivers, cross-country dynamics, and country-specific variables. Their results show that incorporating multi-country and sectoral information enhances the accuracy of core inflation forecasts, while aggregate models tend to perform better for headline inflation. Earlier work by Gerlach and Svensson (2008) assesses a wide range of monetary and economic indicators. It shows that in the early years of the Economic and Monetary Union (EMU), money-based forecasts —particularly those relying on the broad monetary aggregate $M3$ (which includes currency in circulation, bank deposits, and other liquid assets, adjusted for structural breaks)— showed strong predictive power.. However, they lost accuracy thereafter. Nonetheless, combining monetary with broader economic indicators continues to yield improved forecasting performance

[1]SHAP (SHapley Additive exPlanations) is a widely used interpretability method that attributes each predictor's contribution to model outputs, helping to identify the main drivers behind the forecasts.

over simple benchmarks.

In the context of Spain, traditional forecasting frameworks have also emphasized a blend of econometric models and judgment. Álvarez and Sánchez (2017) describe a suite of models used by the *Banco de España*—including Phillips curves, transfer-function models, and time-varying mean models—supplemented with expert judgment to account for heterogeneity in price dynamics. Complementing this, Bańbura and Bobeica (2021) shows that incorporating survey-based inflation expectations from professional forecasters into Phillips curve and factor models improves forecast accuracy, particularly for core inflation. Expectations from markets, firms, or households, however, do not yield systematic gains.

An important strand of the literature emphasizes that inflation is inherently difficult to forecast, especially at short horizons. Atkeson and Ohanian (2001) show that, for U.S. inflation, a simple naive model that sets future inflation equal to last year's average consistently outperforms Phillips curve specifications. The authors argue that the weak empirical basis of the Phillips curve and the instability of slack–inflation relationships imply that renewed attempts to improve inflation forecasts with Phillips curve models are unlikely to succeed. This reinforces the value of naive forecasts as a benchmark. Relatedly, Stock and Watson (2007) found that the decline in U.S. inflation volatility since the mid-1980s has been accompanied by a reduction in forecastability. Although unobserved-components models with stochastic volatility fit the data well, simple benchmarks, such as random walks or autoregressive models, are difficult to outperform in real-time applications. The authors also suggest that combining forecasts, such as staff or survey forecasts that pool model-based and judgmental components, can lead to improvements. They imply that robustness often comes less from individual model choice than from the systematic use of forecast combinations.

Moving from traditional econometric specifications to data-driven approaches, a first benchmark is Medeiros et al. (2021), who evaluates a broad range of machine learning methods within a data-rich environment for U.S. inflation. The authors demonstrate that tree-based models, especially random forests, consistently outperform standard benchmarks, with the most significant improvements observed during recessions and periods of high uncertainty. These results highlight the importance of capturing nonlinearities and stable variable selection mechanisms, especially when there are many predictors available from the price, financial, housing, and labour markets. A complementary cross-country perspective is offered by Kohlscheen (2021), who argues that random forests substantially improve inflation forecasting in advanced economies. Their analysis also highlights the ongoing, albeit diminished, role of inflation expectations and emphasizes the need for future research to integrate the flexibility of machine learning (ML) tools with stronger economic interpretability and incorporation into macroeconomic models. Gradient-boosted tree models, such as *XGBoost* and *LightGBM*, are gaining traction in economic forecasting, including in some European contexts. Although there is still limited evidence of inflation in Spain, Ivaşcu (2023) use these methods to forecast Romanian inflation. Their results are mixed: strong short-term performance, but no clear improvement over AR benchmarks. This highlights the challenges of forecasting in data-scarce environments. In a related study, Michalakopoulos et al. (2025) investigated electricity day-ahead market prices in Greece, Belgium, and Ireland. They found that *LightGBM* consistently achieved the highest forecasting accuracy and robustness, especially when trained on short historical windows. Their results suggest that *LightGBM* could be effective in capturing seasonal dynamics and extreme events under volatile, data-limited conditions, and that similar gains could be achieved when applying boosted tree models to inflation forecasting in European economies such as Spain.

A recent contribution is Lenza, Moutachaker, and Paredes (2025), who apply quantile regression forests (QRF) to forecast euro-area headline and core inflation at 3–12 month horizons. Their framework combines the broad scope of the Phillips curve with the flexibility of machine learning to deliver competitive density forecasts against Bayesian VAR combinations and linear quantile regressions. The results demonstrate the effectiveness of QRF for core inflation, where nonlinear relationships between predictors and inflation are more prevalent, such as during periods of demand-supply imbalances or sudden changes in energy pass-through. This demonstrates that linear models may fail to capture state-dependent dynamics while the QRF approach can adapt flexibly to different regimes. Forecast performance also varies across economic states. Interpretability tools reveal the key drivers and demonstrate the policy relevance of the model during the 2023–2024 disinflation episode.

A critical challenge in modern macroeconomic forecasting is reconciling the high predictive accuracy of machine learning with the need for economic interpretability. Buckmann, Joseph, and Robertson (2023) address this issue by developing a workflow that combines comparative model evaluation with Shapley-based feature decompositions and inference. When applied to U.S. unemployment forecasting, their framework demonstrates that ML models, such as boosting and neural networks, outperform linear benchmarks. Additionally, Shapley value regressions enable the identification of economically significant nonlinearities and variable importance, comparable to traditional econometrics. The study shows that the optimal balance between accuracy and interpretability is achieved by using a small set of theory-driven predictors. In a Spanish context, Díaz Sobrino et al. (2020) offers a compelling narrative-based interpretability approach, treating the qualitative written reports from the *Banco de España* as "shadow forecasts." Using machine learning, they extract predictive signals, such as economic sentiment and textual cues, from quarterly central-bank narratives. This demonstrates that information beyond standard macroeconomic variables, such as central bank narratives and text-based signals, can improve inflation forecasts. Combining these qualitative insights with traditional quantitative predictors provides a clearer, more intuitive understanding of macroeconomic developments in real time.

The literature on inflation forecasting in the euro area and Spain shows a clear evolution. Early studies relied on Phillips curve models and monetary indicators, later complemented by forecast combinations and survey expectations. These approaches improved robustness but often struggled to outperform naive benchmarks, particularly in volatile environments. More recent research highlights the promise of machine learning methods, especially tree-based models, which capture nonlinearities and perform well in high-dimensional settings. However, most applications remain either focused on the U.S. or at the euro-area aggregate, while evidence for Spain is scarce. Moreover, although boosting algorithms such as *XGBoost* and *LightGBM* are increasingly popular in related forecasting tasks, their use in inflation forecasting—particularly in Spain—remains underexplored.

This gap is where the present work contributes. The study provides new evidence on the performance of modern machine learning methods, particularly boosted tree models, relative to traditional benchmarks by applying them to forecast Spanish inflation. This approach's distinctive feature is its systematic use of non-economic data, such as high-frequency indicators and alternative information sources, alongside conventional macroeconomic variables. This allows the forecasts to capture dimensions of inflation dynamics that standard models sometimes miss.

3 Methodology

This section outlines the empirical strategy adopted to forecast Spanish inflation using machine learning methods. The approach combines traditional econometric benchmarks with modern algorithms, while explicitly incorporating a wide range of non-economic predictors to enrich the forecasting environment. The methodology proceeds in several stages: data collection and preprocessing, feature engineering, model estimation, forecasting, and benchmark evaluation.

3.1 Data Collection and Preprocessing

A broad dataset is compiled to cover both macroeconomic and non-macroeconomic dimensions of Spanish inflation dynamics. Sources include official price indices, financial and labour market indicators, as well as unconventional inputs such as internet search trends, commodity prices, and textual indicators. All series are aligned to a monthly frequency and preprocessed to handle missing values, ensure temporal consistency, and apply necessary transformations.

Table 1 summarizes the datasets employed in the study, detailing their frequency, origin, and classification into macroeconomic and non-macroeconomic categories.

Table 1: Summary of Datasets Used for Inflation Forecasting in Spain

Dataset	**Frequency**	**Origin**	**Type**
		Macroeconomic Data	
National CPI (IPC)	Quarterly	INE	Price Index
Interest Rates	Monthly	Banco de España	Financial Indicator
Unemployment Rate	Monthly	INE	Labour Market
GDP (Demand Side)	Quarterly	INE	National Accounts
Household Consumption	Quarterly	INE	National Accounts
R&D Investment	Annual	INE / Cotec	Investment Indicator
Government Spending	Annual	INE	Fiscal Indicator
IBEX 35	Monthly	Investing.com	Stock Market Index
Imports and Exports	Monthly	Datacomex	Trade Indicator
GDP Deflator	Annual	World Bank	Price Index
Average Salary	Annual	datosmacro.com	Labour Market
Europe Brent Spot Price	Monthly	EIA	Commodity Price
		Non-Macroeconomic Data	
Google Trends Data	Daily	Google	Behavioural Data
Housing Prices	Monthly	Idealista	Real Estate
Real Estate Transactions	Quarterly	Spanish ministry	Real Estate
Gasoline & Diesel Prices	Monthly	Spanish ministry	Energy Prices
Drought Indicators (SPEI)	Monthly	AEMET	Environmental
Sentiment Analysis	Daily	El País	Textual/Sentiment
Sentiment Analysis	Daily	El Economista	Textual/Sentiment
Olive Oil Prices	Monthly	Spanish ministry	Commodity Price
Car Matriculations	Monthly	DGT	Automotive Indicator
Commodity Prices	Monthly	IMF	Global Prices
Food Prices	Monthly	FAO	Global Prices
HICP data	Monthly	Eurostat	Price Indexes

The Harmonized Index of Consumer Prices (HICP, or IPCA in Spanish) published by the *Instituto Nacional de Estadística* (INE) serves as the target variable. From this index, four dependent variables are constructed to capture different aspects of inflation dynamics. First, monthly changes in headline inflation provide the baseline reference. Second, the monthly core inflation rate (*inflación subyacente*) is considered, excluding energy and unprocessed food, as reported by the INE. In addition, quarterly and annual changes in the headline index are computed to extend the analysis to medium- and longer-term horizons. Specifically, quarterly changes are obtained as the difference in the index between a given month and the same index three months earlier, while annual changes are computed as the difference relative to twelve months earlier. Missing values in the initial periods are replaced with zeros to ensure continuity of the series.

The macroeconomic dataset incorporates a variety of indicators from official Spanish and international sources. Core variables include interest rates, unemployment rates, and GDP, all of which are converted from a quarterly to a monthly frequency to ensure comparability. Additional measures of domestic demand and structural dynamics include household consumption, R&D investment, government spending and stock market performance through *IBEX 35*, while imports and exports capture trade flows. The GDP deflator and average salaries offer complementary perspectives on price formation and labour market costs. Finally, the Europe Brent spot price for oil represents global factors. All series are processed to harmonize different reporting formats, standardize units, and convert lower-frequency data into consistent monthly series. Quarterly observations, in particular, are expanded by replicating each value across the three months of the quarter to which it corresponds, thereby ensuring temporal alignment with monthly indicators.

The non-macroeconomic dataset introduces complementary information that captures behavioural, environmental, and textual dimensions of Spanish inflation dynamics. Google Trends data was obtained by selecting relevant search topics (e.g., *paro*, *gasolina*, *oro*)[*unemployment*, *gasoline* and *gold* in Spanish] and retrieving their interest over time in Spain. These series were resampled to a monthly frequency to ensure comparability with other indicators. Environmental conditions were accounted for using the Standardized Precipitation Evapotranspiration Index (SPEI), which was processed both nationally and by province. For each region, monthly averages were computed by aggregating gridded data across spatial coordinates, yielding a comprehensive measure of drought intensity.

A textual dimension was introduced via sentiment indicators, constructed from archival news articles. Headlines from major Spanish newspapers such as *El País* and *El Economista* were scraped across multiple years. The first one being a general newspaper, while the second focuses on the economy. A sentiment analysis pipeline was then applied to classify each daily headline into positive, neutral, or negative categories. Specifically, the implementation relied on a widely used pretrained model, `pysentimiento/robertuito-sentiment-analysis` Pérez et al. (2021). Then, sentiment indices were derived as the monthly proportion of headlines falling into each category. This provides a measure of media tone and public discourse by capturing sentiment fluctuations that may affect consumer expectations and market behaviour.

In addition, a diverse set of non-macroeconomic series was incorporated. These include housing prices from *Idealista*, real estate transactions from official registries, fuel prices from the Spanish Ministry for Ecological Transition, olive oil prices from the Ministry of Agriculture, and car matriculations from the General Directorate of Traffic (DGT). To capture global influences, commodity price series from the International Monetary Fund (IMF) and international food price indices from the Food and Agriculture Organization (FAO) were included. Moreover, item-level Harmonised Index of Consumer Prices (HICP) data from Eurostat were added, filtered specifically for Spain, which allows a more detailed representation of consumption categories.

3.2 Feature Engineering

All sources are merged into a single, time–aligned monthly panel. After converting the date field to a proper *datetime* index, every series is coerced to numeric and three rate-of-change transformations are created for each variable: the *monthly* change ($t - t_{-1}$), the *quarterly* change ($t - t_{-3}$), and the *annual* change ($t - t_{-12}$). The unified panel is then combined with the IPC targets and the macroeconomic data, keeping only numeric columns to avoid leakage from identifiers or text fields. Two calendar dummies—*month* and *year*—are appended to capture seasonal and low-frequency effects.

To represent delayed transmission mechanisms, a structured set of lagged predictors is generated by shifting the entire feature matrix by $\{1, 2, 3, 6, 12\}$ months and appending these lagged copies to the contemporaneous features. The final dataset combines the original features with their lagged versions and removes the initial rows where lags generate missing values. The result is a complete, expanded panel ready for model training and forecasting.

3.3 Modelling Approach

The modelling strategy is implemented through a custom function that combines forecast evaluation, feature selection using SHAP values, and recursive out-of-sample prediction. A key element is the definition of forecast horizons: for each decision date t, the model is trained to predict the realized value of inflation at $t+h$. Formally, the target is constructed as $y_t^{(h)} = y_{t+h}$ but indexed at the decision date t, meaning that the model uses only information available at t to produce a forecast h months ahead. This alignment procedure guarantees consistency across horizons while avoiding the use of future information. Predictors are limited to numerical variables, excluding contemporaneous inflation to prevent leakage. Including inflation at time t as a predictor would give the model direct information about the target variable, resulting in overly optimistic, unreplicable forecasts.

A baseline model is then trained on the evaluation window using median imputation to replace missing values. To ensure interpretability, SHAP values are computed, decomposing each prediction into the contribution of its features. Features with an average absolute SHAP importance above a pre-specified threshold are retained, while the remainder are discarded. This step reduces dimensionality and yields a more interpretable feature set that is consistently used across folds and out-of-sample forecasts. Beyond its technical role, this procedure also highlights which predictors—whether traditional macroeconomic indicators or alternative data sources—carry the greatest influence on inflation forecasts, thereby facilitating an economic interpretation of the models.

For model evaluation, a time-series cross-validation (TSCV) procedure is employed. At each *fold*, the model is trained on an expanding training set and evaluated on the next block of observations. A *fold* refers to one iteration of this process, where the training set is extended and the model is tested on the following unseen period. Using expanding windows is particularly suitable for inflation forecasting, as it allows the model to incorporate the most recent structural changes and shocks while preserving the full history of past dynamics. This ensures that forecasts reflect both long-term patterns and newly emerging trends. In this work, a total of 10 *folds* are used for all experiments to ensure consistency in model evaluation. Predictions from each fold are collected, alongside performance metrics including MSE[2], MAE[3], R^2[4], and CRPS[5]. SHAP values are also recomputed within each fold, enabling assessment of the stability and evolution of feature relevance across time.

Once cross-validation is completed, a final model is trained on the entire evaluable sample, and predictions are generated for the future period if requested. Results are provided in two aligned time indices: at decision dates t (reflecting when forecasts are made) and at target dates $t+h$

[2]**MSE** (Mean Squared Error): penalizes large mistakes more heavily. Lower is better.

$$\text{MSE} = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$$

[3]**MAE** (Mean Absolute Error): scale-preserving, more robust to outliers. Lower is better.

$$\text{MAE} = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i|$$

[4]$\mathbf{R^2}$ (Coefficient of Determination): variance explained relative to the mean model. Higher is better.

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}$$

[5]**CRPS** (Continuous Ranked Probability Score): measures the accuracy of the entire forecast distribution compared to the observed value. Lower is better.

(reflecting the horizon of interest). Diagnostic plots are produced to compare predicted versus realized inflation, highlight validation folds, and summarize feature importance through both SHAP barplots and fold-level heatmaps.

3.4 Forecasting and Prediction

The analysis is structured by examining each of the four inflation indicators in turn: monthly CPI, core CPI, quarterly CPI, and annual CPI. For every indicator, the performance of *LightGBM*, *XGBoost*, and *Random Forest* is compared using standard evaluation metrics. Once the most accurate model is identified, two additional diagnostics are presented: a comparison between predicted and realized values at decision dates, and a SHAP-based heatmap that highlights the most influential predictors. This structure provides both a benchmark of model accuracy and insights into the key drivers of inflation at different horizons.

3.4.1 Monthly CPI

Table 2: Model performance for Monthly CPI across horizons

Model	Horizon	RMSE	MAE	R^2	CRPS	N Feat.
LightGBM	$h=1$	0.188	0.308	0.553	0.308	20
XGBoost	$h=1$	0.193	0.313	0.511	0.313	30
Random Forest	$h=1$	0.296	0.388	0.230	0.388	20
LightGBM	$h=3$	0.202	0.311	0.511	0.311	19
XGBoost	$h=3$	0.194	0.324	0.499	0.324	21
Random Forest	$h=3$	0.287	0.383	0.242	0.383	27
LightGBM	$h=6$	0.202	0.321	0.490	0.321	30
XGBoost	$h=6$	0.174	0.306	0.537	0.306	19
Random Forest	$h=6$	0.293	0.387	0.219	0.387	25

Best model and diagnostics. *Best performing model for $h=1$:* **LightGBM**. Figure 1 compares predicted and realized monthly inflation at decision dates; Figure 2 shows the fold-aggregated SHAP importance heatmap.

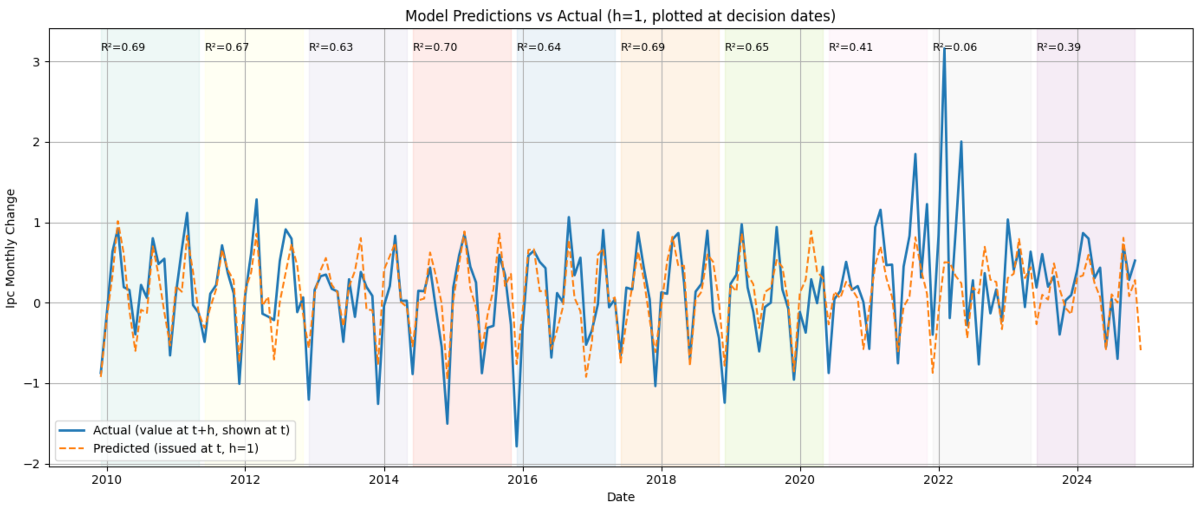

Figure 1: Monthly CPI: predictions vs. realized values at decision dates ($h=1$)

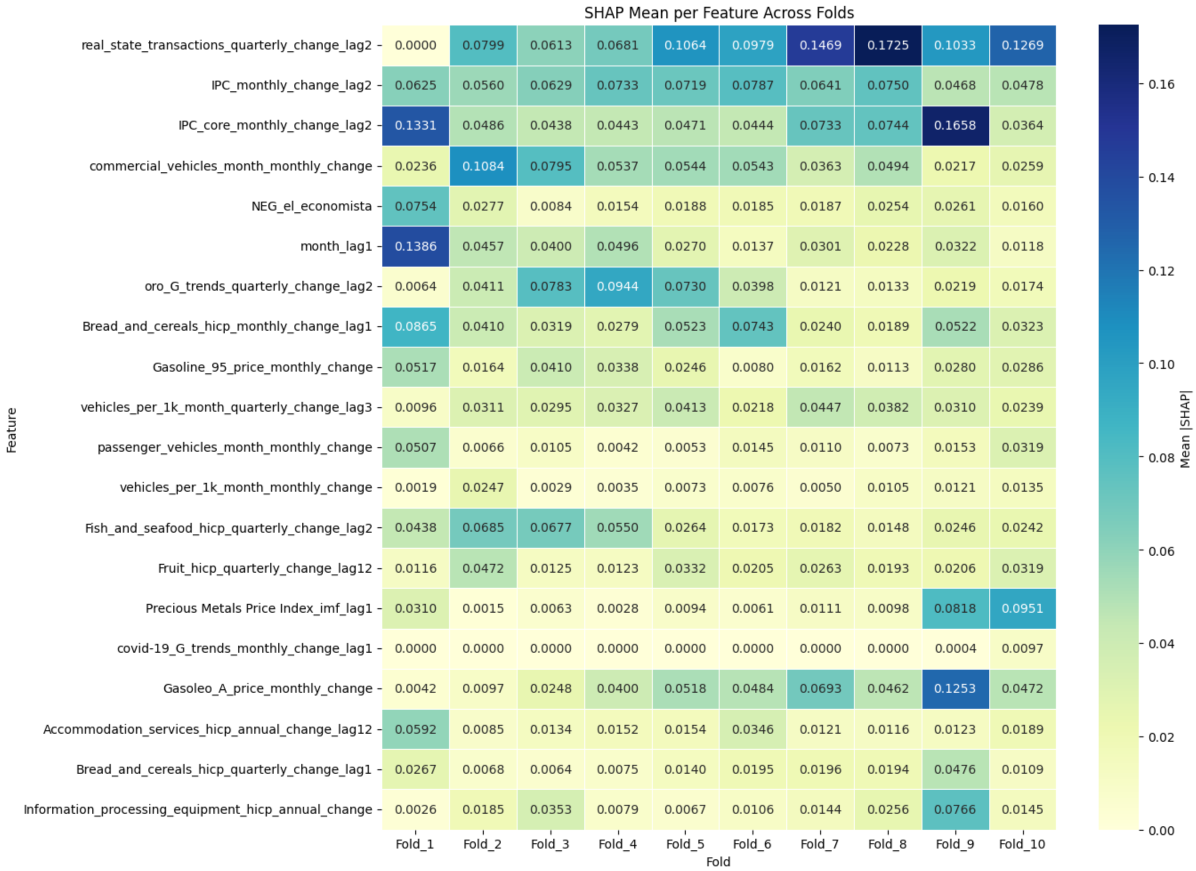

Figure 2: Monthly CPI: SHAP mean per feature across folds ($h = 1$)

Economic interpretation. The SHAP analysis shows that core inflation dynamics are primarily driven by their own past values, underlining the strong persistence typical of underlying price pressures. Indicators related to non-energy industrial goods and services, such as clothing and restaurants, appear as key contributors, consistent with their structural role in shaping the more stable component of inflation. The relative absence of volatile food and energy categories aligns with the construction of the core index, which excludes these elements.

Beyond standard categories, variables such as real estate transactions and vehicle registrations exhibit predictive power. Their relevance suggests that demand-side activity in durable goods markets carries information about medium-term inflation trends. Google Trends indicators, such as those related to geopolitical risks, also appear, likely capturing shifts in uncertainty that influence expectations and service-sector pricing behaviour.

The prediction charts indicate that the models track the core inflation path with greater stability compared to the headline index, though some deviations occur during episodes of macroeconomic shocks. This reflects the slower-moving, persistent nature of core inflation, where both sectoral activity indicators and sentiment-related measures play a complementary role in explaining dynamics.

3.4.2 Core CPI

Table 3: Model performance for Core CPI across horizons

Model	Horizon	RMSE	MAE	R^2	CRPS	N Feat.
LightGBM	$h = 1$	0.084	0.206	0.710	0.206	31
XGBoost	$h = 1$	0.175	0.241	0.460	0.241	33
Random Forest	$h = 1$	0.256	0.352	0.309	0.352	11
LightGBM	$h = 3$	0.069	0.195	0.782	0.195	23
XGBoost	$h = 3$	0.063	0.178	0.753	0.178	23
Random Forest	$h = 3$	0.239	0.334	0.412	0.334	17
LightGBM	$h = 6$	0.087	0.211	0.759	0.211	20
XGBoost	$h = 6$	0.094	0.201	0.720	0.201	23
Random Forest	$h = 6$	0.225	0.322	0.450	0.322	11

Best model and diagnostics. *Best performing model for $h = 1$:* **LightGBM**. Figure 3 compares predicted and realized monthly inflation at decision dates; Figure 4 shows the fold-aggregated SHAP importance heatmap.

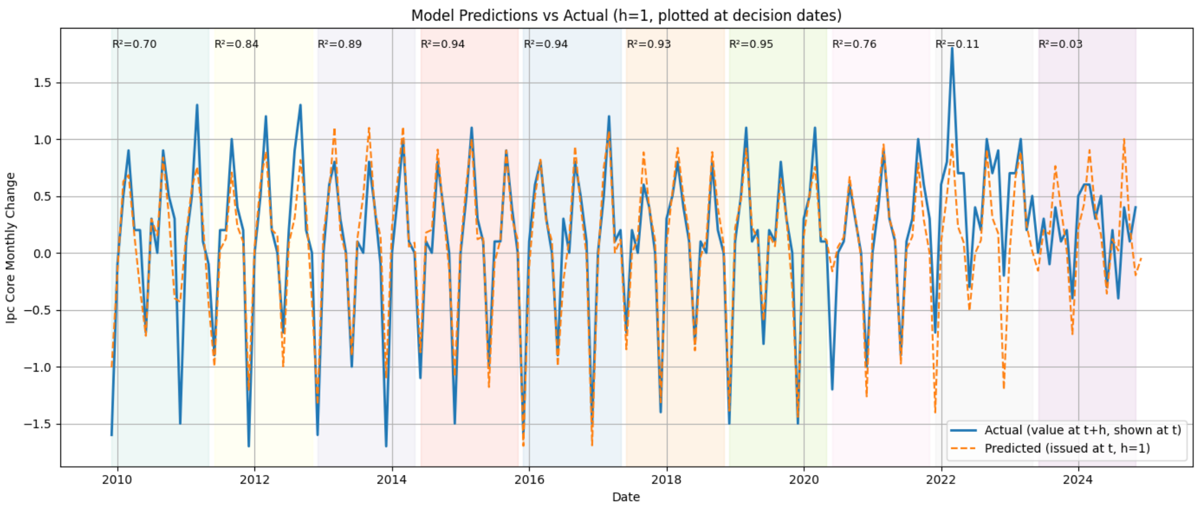

Figure 3: Core CPI: predictions vs. realized values at decision dates ($h = 3$)

Figure 4: Core CPI: SHAP mean |value| per feature across folds ($h = 3$)

Economic interpretation. The prediction charts indicate that the models are able to track short-term changes in monthly inflation with reasonable accuracy, although performance varies across time and events. Peaks and troughs in inflation are generally captured, but large shocks remain more difficult to anticipate, as is typical in high-volatility episodes. Nevertheless, the models provide a useful signal of inflation dynamics at short horizons.

The SHAP analysis highlights that beyond core indicators such as the lagged monthly CPI and core CPI, other variables play a significant role. Real estate transactions and car registrations appear as relevant predictors, likely because they capture household consumption dynamics and durable goods demand, which are closely linked to inflationary pressures. The presence of food-related variables, such as fish and bread prices, reflects the strong pass-through of food markets into overall inflation.

Interestingly, non-traditional predictors such as Google Trends for *'war in Ukraine'* and *'gold'* also emerge as influential. These may proxy uncertainty and risk sentiment, which indirectly shape inflation expectations and consumption behaviour. For example, spikes in *'Ukraine'* searches coincide with geopolitical shocks that triggered energy and food price surges. Similarly, gold-related searches might reflect shifts in investor sentiment that correlate with inflationary environments. Together, these findings suggest that unconventional indicators can provide complementary information to standard economic series in forecasting Spanish inflation.

3.4.3 Quarterly CPI

Table 4: Model performance for Quarterly CPI across horizons

Model	Horizon	RMSE	MAE	R^2	CRPS	N Feat.
LightGBM	$h = 3$	0.589	0.580	0.506	0.580	28
XGBoost	$h = 3$	0.787	0.639	0.322	0.639	23
Random Forest	$h = 3$	0.937	0.746	0.229	0.746	22
LightGBM	$h = 6$	0.571	0.590	0.499	0.590	28
XGBoost	$h = 6$	0.751	0.664	0.343	0.664	22
Random Forest	$h = 6$	0.941	0.730	0.221	0.730	20
LightGBM	$h = 12$	0.677	0.634	0.423	0.634	31
XGBoost	$h = 12$	0.722	0.638	0.393	0.638	22
Random Forest	$h = 12$	0.992	0.756	0.205	0.756	24

Best model and diagnostics. *Best performing model for h = 3:* **LightGBM**. Figure 5 compares predicted and realized monthly inflation at decision dates; Figure 6 shows the fold-aggregated SHAP importance heatmap.

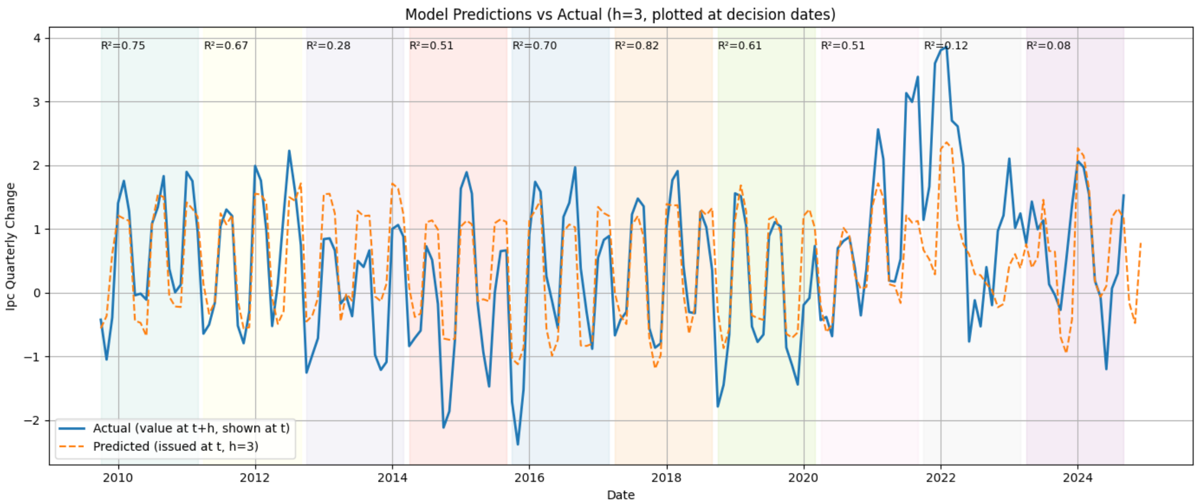

Figure 5: Quarterly CPI: predictions vs. realized values at decision dates ($h = 3$)

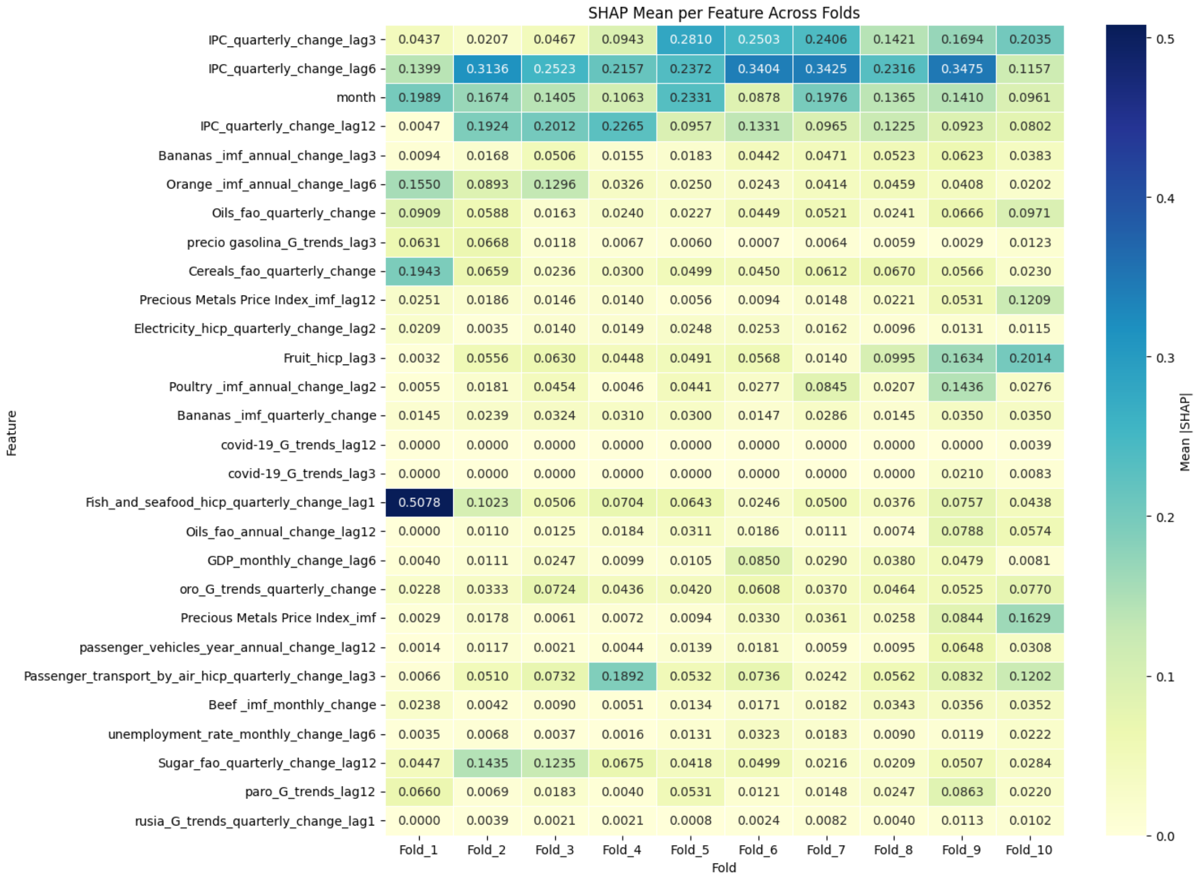

Figure 6: Quarterly CPI: SHAP mean |value| per feature across folds ($h = 3$)

Economic interpretation. The SHAP analysis reveals that quarterly CPI dynamics are heavily influenced by their own lags, which highlights the strong autocorrelation typical of broader price indices. Energy-related variables, such as oil and fuel prices, also emerge as relevant, reflecting the role of energy costs in driving quarterly shifts in consumer prices. Seasonal patterns, proxied by the *month* variable, further contribute to explaining inflation fluctuations across time.

Interestingly, some seemingly less direct variables show predictive power. Real estate–related indicators, such as housing and transaction measures, may capture aggregate demand and broader macroeconomic momentum feeding into price pressures. Likewise, agricultural commodities like bananas, poultry, and fish appear relevant, possibly because they represent volatile components that, when aggregated quarterly, leave discernible signals for headline inflation. The presence of search-based indicators, such as Google Trends on unemployment or geopolitical themes, as well as Covid-19 related searches, suggests that shifts in sentiment and uncertainty are also informative for price-setting behaviour.

The prediction charts confirm that the models broadly capture the trajectory of quarterly inflation, though deviations arise during high-volatility episodes. This performance reflects the interplay of persistent dynamics and sector-specific drivers, where a combination of traditional price components, housing activity, and expectation-based proxies jointly enrich the forecasting framework.

3.4.4 Annual CPI

Table 5: Model performance for Annual CPI across horizons

Model	Horizon	RMSE	MAE	R^2	CRPS	N Feat.
LightGBM	$h = 6$	2.902	1.177	-2.619	1.177	22
XGBoost	$h = 6$	3.489	1.313	-4.560	1.313	22
Random Forest	$h = 6$	4.854	1.395	-2.030	1.395	22
LightGBM	$h = 12$	3.067	1.220	-2.740	1.220	26
XGBoost	$h = 12$	2.834	1.169	-2.882	1.169	21
Random Forest	$h = 12$	6.128	1.672	-4.947	1.672	15
LightGBM	$h = 24$	3.068	1.158	-0.942	1.158	26
XGBoost	$h = 24$	2.339	1.102	-1.675	1.102	24
Random Forest	$h = 24$	4.716	1.475	-3.258	1.475	15

Best model and diagnostics. *Best performing model for $h = 12$:* **LightGBM**. Figure 7 compares predicted and realized monthly inflation at decision dates; Figure 8 shows the fold-aggregated SHAP importance heatmap.

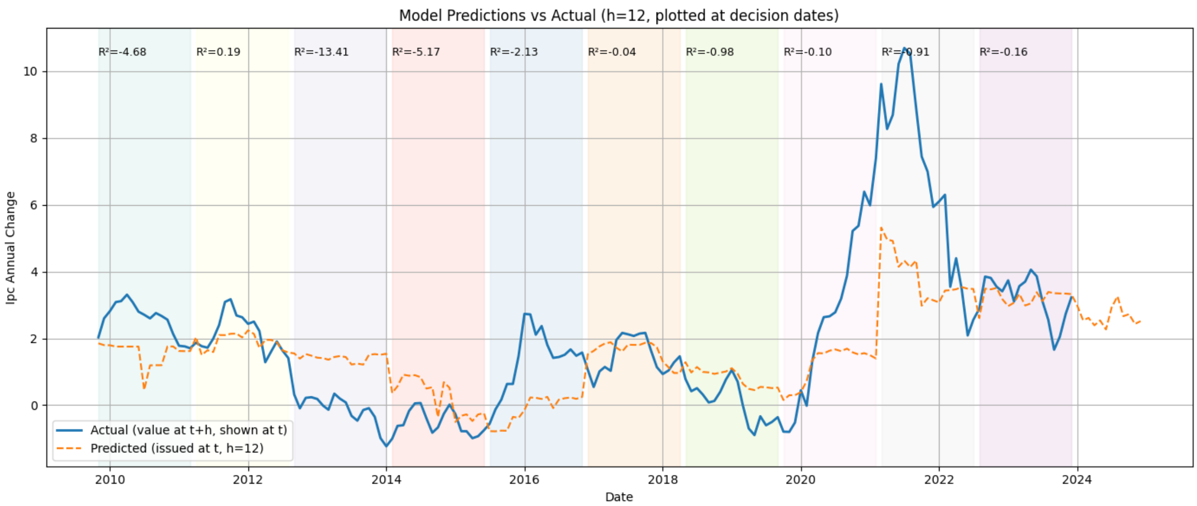

Figure 7: Annual CPI: predictions vs. realized values at decision dates ($h = 12$)

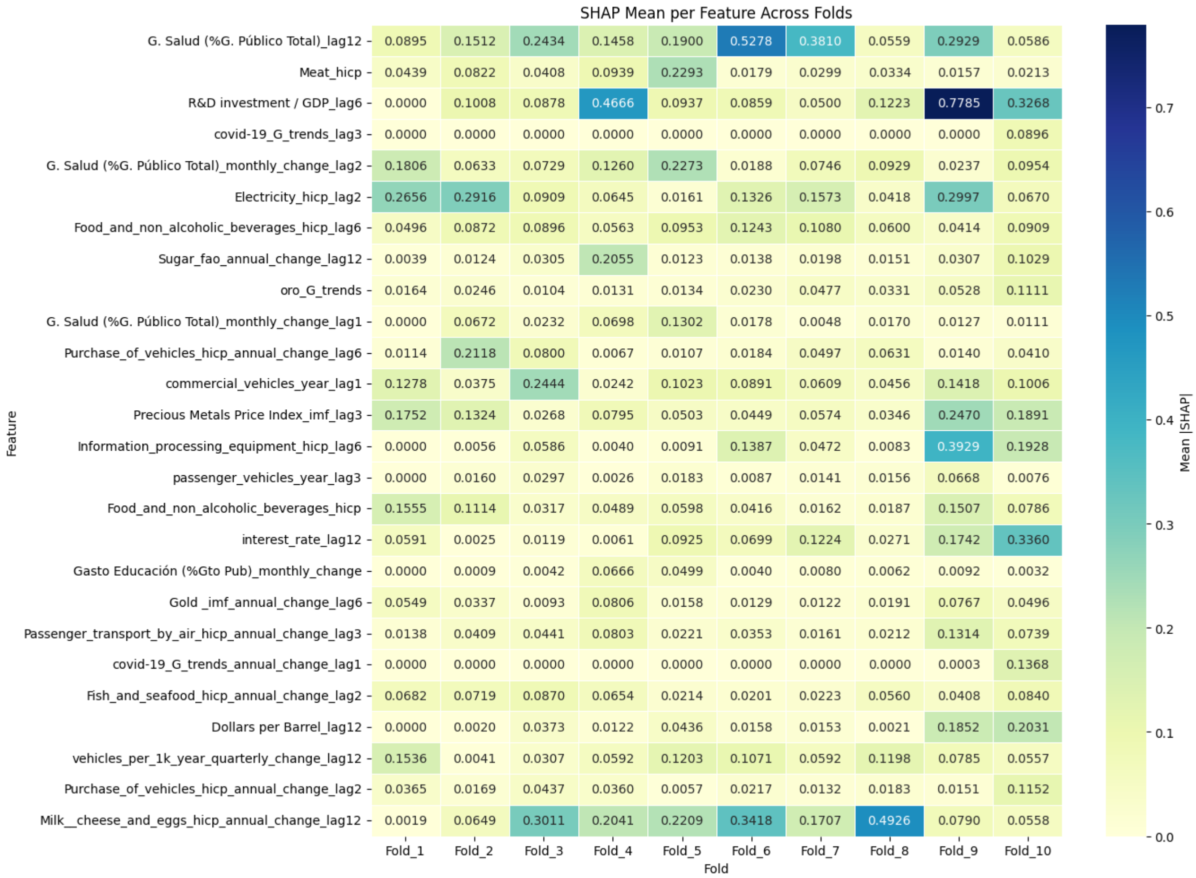

Figure 8: Annual CPI: SHAP mean |value| per feature across folds ($h = 12$)

Economic interpretation. The SHAP analysis for annual CPI indicates that health expenditure, energy-related components such as electricity, and certain food categories are among the main predictors. Interestingly, variables like R&D investment, gold prices, and Google Trends indicators also emerge as relevant, hinting that broader macro-financial and sentiment factors may carry information about longer-run price pressures. These results suggest that the model captures not only standard consumption categories but also more unconventional drivers that might proxy for uncertainty or investment-related demand.

Despite these insights, the overall predictive performance of the models for annual inflation remains weak, particularly at longer horizons. The negative R^2 values and relatively high error metrics underscore the difficulty of anticipating annual inflation dynamics, which are often shaped by structural shocks, policy interventions, and global developments that are not easily captured by historical regularities in the data.

The prediction charts confirm that while the models occasionally align with broad turning points, they fail to consistently track the level of annual CPI. This reinforces the interpretation that, unlike shorter-term horizons, annual inflation is influenced by a wide set of complex and less predictable factors. As a result, the usefulness of these forecasts may lie more in identifying potential risk factors and explanatory channels rather than in delivering accurate point predictions.

4 Benchmarking and Evaluation

To evaluate the relative performance of the machine learning forecasts, the results are benchmarked against two reference standards. First, a comparison is made with the official macroeconomic projections published by the *Banco de España*, which serve as an institutional baseline for inflation expectations. Second, following the approach of Stock and Watson (2007), forecasts are also contrasted with a simple random walk, a common benchmark in empirical macroeconomic forecasting.

4.1 Benchmarking against *Banco de España* forecasts

Every March, the *Banco de España* (BdE) publishes a macroeconomic forecast bulletin that includes projections for annual inflation. These forecasts provide end-of-year inflation changes for the current year (a 9-month horizon) and the following year (a 21-month horizon). To benchmark the performance of the machine learning approach, the *LightGBM* model was adjusted to match these horizons. Predictions were computed on a monthly basis, averaged across the year to yield an annual inflation forecast comparable to the BdE projections.

Among the models considered, *LightGBM* consistently outperforms alternatives in terms of error metrics and stability across horizons. Its ability to handle non-linearities, interactions, and large sets of lagged predictors makes it particularly well-suited for capturing the dynamics of inflation. For this reason, *LightGBM* is selected as the benchmark model when comparing results against the *Banco de España* forecasts.

Figures 9 and 10 show the comparative evolution of actual inflation, BdE forecasts, and model-based projections at both horizons. At the 9-month horizon, the BdE tends to provide more accurate short-term assessments, particularly during periods of heightened volatility, while the model captures the overall dynamics but exhibits higher forecast errors. At the 21-month horizon, performance is closer, with both approaches facing challenges in anticipating inflationary shocks, though the model delivers a more balanced error distribution.

Table 6 reports forecast evaluation metrics against realized inflation. For the 9-month horizon, the BdE achieves a lower MAE (0.56 vs. 1.04) and smaller cumulative errors, while the model underperforms. However, at the 21-month horizon, the *LightGBM* approach exhibits superior accuracy (MAE 1.04 vs. 1.54) and a comparable bias. Diebold–Mariano tests[6] indicate no statistically significant difference between the two forecast sources, underscoring that the machine learning model provides performance on par with institutional forecasts over longer horizons.

One possible explanation for this pattern is that, in the short term, institutional forecasts may benefit from judgmental adjustments and access to qualitative information, such as policy measures, negotiations, and expert assessments, that are not captured by purely data-driven models. However, over longer timeframes, the systematic use of lagged variables and nonlinear relationships enables *LightGBM* to more accurately capture persistent dynamics, thereby reducing forecast errors compared to the benchmark. This suggests that machine learning models are particularly valuable for forecasting beyond the immediate horizon, where institutional judgment alone becomes less reliable.

[6]The Diebold–Mariano test *(Diebold and Mariano 1995)* evaluates whether the predictive accuracy of two competing forecasts differs significantly, based on the loss differential series (e.g., squared or absolute forecast errors). It is widely applied in forecast comparison exercises.

Table 6: Forecast errors vs. actual annual inflation (IAPC).

Model	MAE	Mean Error (Bias)	Cumulative Error
Horizon: 9 months			
BdE	0.560	-0.240	-3.600
Model	1.044	-0.607	-9.112
DM statistic = 1.495, p-value = 0.135			*(not statistically significant)*
Horizon: 21 months			
BdE	1.540	-0.620	-9.300
Model	1.044	-0.607	-9.112
DM statistic = -1.096, p-value = 0.273			*(not statistically significant)*

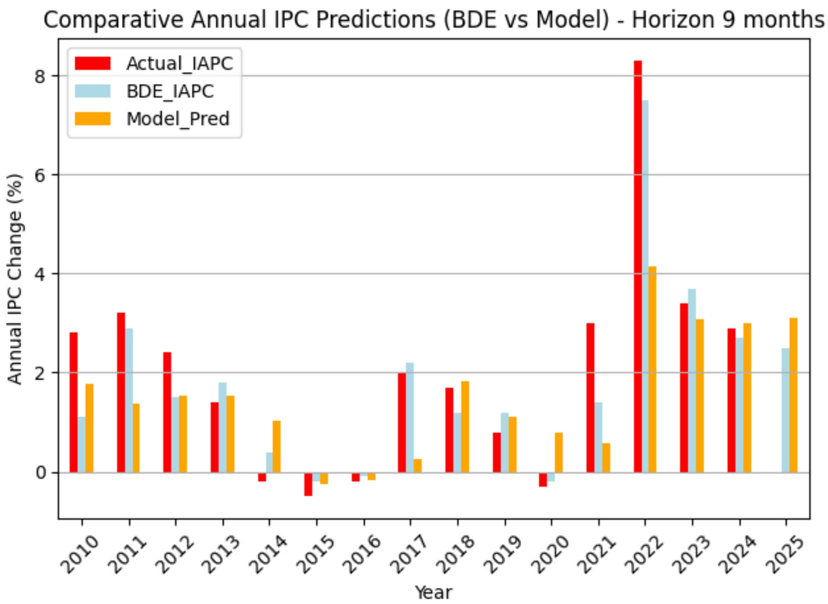

Figure 9: Comparative Annual IPC Predictions (BdE vs. Model) – Horizon 9 months

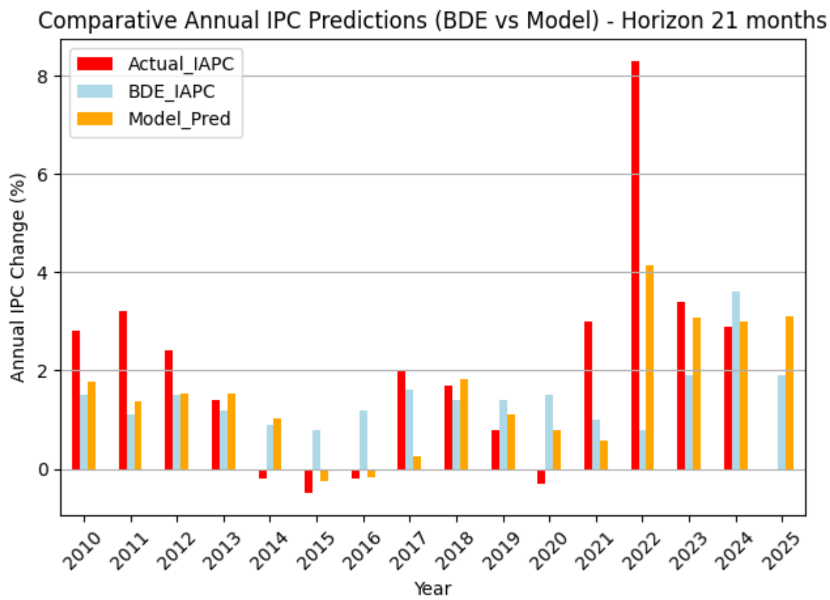

Figure 10: Comparative Annual IPC Predictions (BdE vs. Model) – Horizon 21 months

4.2 Benchmarking against a Random Walk

To assess whether the machine learning models extract predictive structure beyond simple persistence, forecasts are compared against a driftless random walk (RW) benchmark. In this setup, the forecast for horizon h is simply the most recently observed value of inflation at time t, projected forward to $t+h$.

RW forecasts are generated using an expanding-window evaluation. At each origin of the forecast, the most recent observation is used for the prediction. There is also the option to include a constant drift term, which is equal to the average historical change in the series. For each horizon, the forecast path is aligned with actual realizations, and performance is evaluated using the same error metrics as in the machine learning models (RMSE, MAE, CRPS, R^2).

This rolling-origin procedure ensures that the RW benchmark is computed under the same conditions as the machine learning forecasts, allowing for a fair comparison. Statistical differences in predictive accuracy are formally tested using Diebold–Mariano (DM) tests.

4.2.1 Monthly CPI

The monthly index exhibits strong short-run inertia, so the RW can be a demanding benchmark. In this case, the *LightGBM* model delivers lower RMSE and MAE compared to the random walk, showing that high-frequency predictors add value beyond pure persistence. The RW, however, still captures broad trends, reflecting the importance of short-term carry-over. The DM test confirms this.

Table 7: Monthly CPI: model vs. random walk ($h = 1$)

Model	R^2	RMSE	MAE
LightGBM	0.524	0.430	0.311
Random Walk	-0.668	0.789	0.625
DM statistic $= -6.307$, p-value $= 0.000$ *(significant)*			

Reading guide. Lower RMSE/MAE and higher R^2 indicate better accuracy. A significant DM test in favor of the model implies predictive gains beyond a persistence benchmark; lack of significance indicates performance comparable to a random walk.

4.2.2 Core CPI

For the core index, the *LightGBM* model shows a clear advantage over the random walk, with much lower RMSE and MAE values. This highlights the model's ability to capture underlying inflation dynamics beyond persistence. The DM test confirms this.

Table 8: Core CPI: model vs. random walk ($h = 1$)

Model	R^2	RMSE	MAE
LightGBM	0.744	0.330	0.234
Random Walk	-0.667	0.837	0.693
DM statistic = -9.744, p-value = 0.000 *(significant)*			

4.2.3 Quarterly CPI

For the quarterly index, the *LightGBM* model also outperforms the random walk, delivering lower RMSE and MAE. This suggests that the model is able to capture signals beyond pure persistence, even at a lower frequency. The random walk still follows broad dynamics, but with less precision. The DM test confirms this.

Table 9: Quarterly CPI: model vs. random walk ($h = 3$)

Model	R^2	RMSE	MAE
LightGBM	0.534	0.805	0.611
Random Walk	-1.506	1.855	1.652
DM statistic = -12.273, p-value = 0.000 *(significant)*			

4.2.4 Annual CPI

For the annual horizon, the model delivers clear improvements over the random walk, with substantially lower RMSE and MAE. This suggests that predictors add meaningful information for longer-run inflation dynamics. The DM test confirms this.

Table 10: Annual CPI: model vs. random walk ($h = 12$)

Model	R^2	RMSE	MAE
LightGBM	0.480	1.730	1.220
Random Walk	-0.390	2.747	1.962
DM statistic = -4.940, p-value = 0.000 *(significant)*			

5 Conclusions

This study developed an extensive forecasting framework for Spanish inflation that combines traditional macroeconomic indicators with alternative data sources, including real estate activity, commodity markets, internet search intensity, and sentiment indicators. A systematic feature engineering pipeline was designed to produce lagged variables and enriched predictors. Then, forecasting models were trained using machine learning methods, with feature selection guided by SHAP values. This methodology was applied to four key measures of inflation: monthly, core, quarterly, and annual indices.

Empirical results demonstrate that lagged values of inflation play a central role in explaining dynamics across all horizons, highlighting the persistence inherent in price formation. Beyond these, SHAP analyses reveal the relevance of both conventional predictors and less traditional variables, such as Google Trends indicators and real estate transactions. Among the forecasting methods, *LightGBM* consistently emerges as the best-performing model, producing relatively stable and accurate forecasts across the different measures of inflation.

Benchmarking exercises against forecasts from the Bank of Spain and random walk models demonstrate the strengths and limitations of the proposed framework. While the models perform competitively in some settings, random walk benchmarks are difficult to outperform, particularly at annual horizons where predictive accuracy is weaker. Overall, this approach shows the potential of incorporating lagged inflation, macroeconomic, and non-macroeconomic data into modern predictive models. This approach provides valuable insights for researchers and policymakers interested in monitoring and anticipating price dynamics in Spain.

References

Álvarez, Luis J. and Isabel Sánchez (2017). *A Suite of Inflation Forecasting Models.* Occasional Paper 1703. Banco de España.

Atkeson, Andrew and Lee E. Ohanian (2001). "Are Phillips Curves Useful for Forecasting Inflation?" In: *Federal Reserve Bank of Minneapolis Quarterly Review* 25.1, pp. 2–11.

Bańbura, Marta and Elena Bobeica (2021). *Do inflation expectations improve model-based inflation forecasts?* Working Paper 2138. Banco de España.

Bobeica, Elena et al. (2020). *Does the Phillips curve help to forecast euro area inflation?* ECB Working Paper 2471. European Central Bank.

Buckmann, Marcus, Andreas Joseph, and Helena Robertson (2023). "An interpretable machine learning workflow with an economic prior". In: *International Journal of Central Banking.* Interpretable ML framework with SHAP and economic priors for macro forecasting.

Capolongo, Antonio and Caterina Pacella (2021). "Forecasting inflation in the euro area: countries matter!" In: *Empirical Economics* 61.5, pp. 2477–2499.

Consolo, Agostino, Riccardo Cristadoro, and Mario Forni (2022). *A Bottom-Up Reduced-Form Phillips Curve for the Euro Area.* IMF Working Paper 2022/260. International Monetary Fund.

Díaz Sobrino, Natalia et al. (2020). *The narrative about the economy as a shadow forecast: an analysis using Banco de España quarterly reports.* Working Paper 2020/XX. Banco de España.

Diebold, Francis X. and Roberto S. Mariano (1995). "Comparing Predictive Accuracy". In: *Journal of Business & Economic Statistics* 13.3, pp. 253–263.

Gerlach, Stefan and Lars E. O. Svensson (2008). *Do monetary indicators lead euro area inflation?* ECB Working Paper 867. European Central Bank.

Ivaşcu, Codruţ (2023). "Can Machine Learning Models Predict Inflation?" In: *Proceedings of the International Conference on Business Excellence* 17.1. Includes XGBoost and LightGBM for CPI forecasting in Romania, pp. 1748–1756.

Kohlscheen, Emanuel (2021). *Using Machine Learning to Forecast Inflation: New Evidence from Advanced Economies.* BIS Working Paper 980. Bank for International Settlements.

Lenza, Michele, Inès Moutachaker, and Joan Paredes (2025). "Density Forecasts of Inflation: A Quantile Regression Forest Approach". In: *European Economic Review* 178, p. 105079.

Medeiros, Marcelo C. et al. (2021). "Forecasting Inflation in a Data-Rich Environment: The Benefits of Machine Learning Methods". In: *Journal of Business & Economic Statistics* 39.1, pp. 98–119.

Michalakopoulos, Vasilis et al. (2025). "Data-driven Day Ahead Market Prices Forecasting: A Focus on Short Training Set Windows". In: *arXiv preprint arXiv:2506.10536.*

Pérez, Juan Manuel et al. (2021). “pysentimiento: A Python Toolkit for Opinion Mining and Social NLP tasks”. In: *arXiv preprint arXiv:2106.09462*.

Stock, James H. and Mark W. Watson (2007). “Why Has U.S. Inflation Become Harder to Forecast?” In: *Journal of Money, Credit and Banking* 39.1. Supplement to Vol. 39, No. 1, February, pp. 3–33.