



Forecasting Euro-Area Sovereign Bond Spreads: Evidence on Accuracy and Economic Value

Jhon Dalton Cevallos Delgado

Dissertation written under the supervision of professor Dan Tran

Dissertation submitted in partial fulfilment of requirements for the MSc in
Economics, at the Universidade Católica Portuguesa, January 04th, 2026.

Abstract

This thesis studies the forecasting and economic performance of euro-area sovereign bond spread models, with a particular focus on the relationship between statistical accuracy and cross-sectional ranking ability. I develop an error-correction modeling framework that combines a theory-consistent econometric specification with a machine-learning extension, allowing long-run equilibrium relationships to be preserved while capturing nonlinear short-run dynamics. Model performance is evaluated not only using conventional forecasting metrics, but also through long–short portfolio strategies designed to test economic value.

Using monthly data for ten euro-area sovereign issuers over the period 2010–2024, I show that incorporating machine-learning techniques into the short-run dynamics substantially reduces forecast errors relative to a linear error-correction benchmark, even though out-of-sample R^2 values remain negative. However, improvements in statistical accuracy do not translate mechanically into superior cross-sectional ranking or uniformly higher risk-adjusted portfolio returns. Instead, economic performance depends on the interaction between forecast signals, persistence in spread dynamics, and the way predictions are implemented within portfolios under different interest-rate hedging configurations.

The results highlight a central accuracy–ranking puzzle: models that improve point-forecast precision do not necessarily generate better investment outcomes. For long–short strategies, the ability to correctly rank relative tightenings and widenings is more important than minimizing forecast errors. These findings underscore the importance of aligning forecasting objectives, evaluation metrics, and portfolio implementation when assessing predictive models in sovereign bond markets.

Title: Forecasting Euro-Area Sovereign Bond Spreads: Evidence on Accuracy and Economic Value

Author: Jhon Dalton Cevallos Delgado

Keywords: Sovereign Bond Spreads; Error-Correction Models; Machine Learning; Cross-Sectional Ranking; Long–Short Portfolios

Resumo

Esta tese analisa o desempenho preditivo e económico de modelos de spreads soberanos da área do euro, com foco na relação entre a precisão estatística e a capacidade de ordenação cross-section. Desenvolvo um modelo de correção do erro que combina uma especificação econométrica consistente com a teoria económica com uma extensão de *machine learning*, preservando as relações de equilíbrio de longo prazo e captando dinâmicas não lineares de curto prazo. O desempenho é avaliado tanto com métricas tradicionais de previsão como através de estratégias de portefólios *long–short*.

Com dados mensais de dez emissores soberanos da área do euro entre 2010 e 2024, mostro que a incorporação de *machine learning* reduz os erros de previsão face a um modelo linear, apesar de os valores de R^2 fora da amostra permanecerem negativos. No entanto, melhorias na precisão estatística não se traduzem necessariamente numa melhor ordenação cross-section nem em retornos ajustados ao risco mais elevados.

Os resultados evidenciam um *puzzle* entre precisão e ordenação: para estratégias *long–short*, a capacidade de ordenar corretamente apertos e alargamentos relativos é mais importante do que minimizar erros de previsão.

Título: Previsão dos Spreads da Dívida Soberana da Área do Euro: Evidência sobre Precisão e Valor Económico

Autor: Jhon Dalton Cevallos Delgado

Palavras-chave: Spreads da Dívida Soberana; Modelos de Correção de Erro; Aprendizagem Automática; Classificação Transversal; Portefólios Long–Short

Preface

The completion of this thesis represents the conclusion of a demanding but rewarding academic process, carried out alongside professional commitments. The experience required sustained effort and discipline, while also providing valuable opportunities for learning and intellectual development.

I would like to thank God for granting me health, patience, and clarity of mind throughout this period. I am especially grateful to my Mother for her constant support and guidance, and to my brother Kevin for the trust and encouragement they placed in me when I decided to pursue my studies abroad. Their support was fundamental to this decision and to the completion of this work.

I am also thankful to my partner, Angie, for her understanding, patience, and support during periods of intensive work and long research hours.

I would like to express my sincere gratitude to my supervisor, Professor Dan Tran. From the first course I attended, his way of communicating ideas and motivating students strongly resonated with me, encouraging me to pursue further coursework under his guidance and to remain engaged throughout my studies. His advice, availability, and support during the development of this thesis are greatly appreciated.

Finally, I acknowledge the support of the university for providing a rigorous academic environment and the resources that made this research possible.

Contents

1	Introduction	1
2	Literature Review and Theoretical Framework	3
3	Data	5
3.1	Overview and Coverage	5
3.2	Data Sources, Frequency, and Pre-processing	6
3.3	Global and Regional Variables	7
3.4	Country-Specific Variables	8
3.5	Bond-Level Structural Characteristics	8
3.6	Variable Construction	9
3.6.1	Dependent Variable: Sovereign Spread	9
3.6.2	Constructed Independent Variables	9
3.7	Descriptive Statistics	11
4	Methodology	12
4.1	Overview of Empirical Strategy	12
4.2	Error-Correction Model (ECM)	15
4.2.1	Long-Run Cointegration	15
4.2.2	Short-Run Dynamics and One-Step-Ahead Forecasting	16
4.2.3	XGBoost Model	16
4.2.4	Hyperparameter Tuning and Time-Series Cross-Validation	17
4.2.5	Forecasting Procedure	18
4.3	Portfolio Construction	19
4.3.1	Long–Short Portfolio Strategy	19
4.3.2	Interest-Rate Hedging Strategies	20
4.4	Performance Evaluation	21
5	Results	22
5.1	Overview of Results	22
5.2	Econometric Performance	23
5.2.1	Error–Correction Validity	23
5.2.2	Magnitude of Feature Contributions	24
5.2.3	Out-of-Sample R^2	26

5.2.4	Ranking Performance and Tercile Agreement	27
5.3	Portfolio Performance.	30
5.4	Alternative classification-based approach.	34
6	Conclusion	35
A	Technical Implementation Details.	37
A.1	Hyperparameter Tuning and Cross-Validation Design	37
B	Model Diagnostics and Robustness.	38
B.1	Feature Contributions by Country	39
B.2	Cross-Validation Design Comparison	42
C	Kitchen Sink Specifications	42
C.1	Kitchen Sink OLS Model	43
C.2	Kitchen Sink Machine Learning Model	44
D	Panel Specification with Country Fixed Effects	45
E	Ordered Logit Classification Model	46

List of Tables

1	Global and Regional Factors.	7
2	Country-Specific Factors.	8
3	Bond-Level Structural Characteristics.	9
4	Descriptive Statistics and Unit-Root Properties: Sovereign Yield Spread (10-Year).	11
5	Out-of-sample R^2 Statistics (ECM vs ML-ECM).	26
6	Hit Ratio by Country (ECM vs ML-ECM).	29
7	Portfolio performance (ECM and ML-ECM).	31
8	Average Active Weights by Country (ECM and ML-ECM).	33
9	Trading Frequency by Country (ECM vs ML-ECM).	34
B.1	Unit-Root Tests for Long-Run Fundamental Variables.	39
B.2	Portfolio Performance Across CV Specifications (ML-ECM).	42
C.1	Out-of-Sample R^2 (ECM vs Kitchen Sink).	43
C.2	Hit Ratios by Country (ECM vs Kitchen Sink).	43
C.3	Portfolio Performance (Kitchen Sink).	44
C.4	Out-of-sample R^2 Statistics (ML-ECM vs ML-Kitchen Sink).	44
C.5	Hit Ratios by Country (ML-ECM vs ML-Kitchen Sink).	44
C.6	Portfolio Performance (ML-Kitchen Sink).	45
D.1	Hit Ratios by Country (Panel Model).	45
D.2	Portfolio Performance (Panel Model).	46
E.1	Portfolio Performance (Ordered Logit).	46

List of Figures

1	Modeling and Evaluation Pipeline.	14
2	Expanding-window cross-validation scheme with a single hold-out validation split used for hyperparameter tuning (CV=1).	18
3	Rolling ADF stationarity flags for ECM residuals. Purple cells indicate I(0) residuals (stationary / cointegrated), while grey cells indicate I(1) residuals (non-stationary).	24
4	Average absolute feature contributions in the ECM and ML-ECM models. . . .	25
5	Long and Short Positions (ECM).	27
6	Long and Short Positions (ML-ECM).	28
7	Tercile Agreement (ECM).	28
8	Tercile Agreement (ML-ECM).	28
B.1	Feature Contributions by Country (ECM).	40
B.2	Feature Contributions by Country (ML-ECM).	41

1. Introduction

Sovereign bond spreads—the yield difference between a country’s government bonds and a risk-free benchmark—are a key indicator of fiscal conditions, market confidence, and systemic risk. For investors, policymakers, and central banks, understanding what drives these spreads and predicting their evolution is important for portfolio allocation, risk management, and crisis prevention. A large empirical literature documents the main determinants of sovereign spreads, including global risk appetite, monetary conditions, and country-specific fiscal and macroeconomic fundamentals. More recently, machine-learning methods have been applied to sovereign risk forecasting because of their ability to capture nonlinearities, interactions, and regime-dependent behavior that standard econometric models may miss.

Despite this progress, an important question remains underexplored: does improved forecast accuracy lead to better cross-sectional ranking of sovereign issuers? Long–short portfolio strategies do not rely on precise point forecasts. Instead, they require correctly identifying which countries are expected to experience spread tightening and which are expected to experience spread widening. A model that substantially reduces mean squared error generates economic value only if it improves the ranking of expected tighteners and wideners. This distinction between statistical accuracy and ranking performance—hereafter referred to as the accuracy–ranking puzzle—motivates this study.

This thesis addresses two related questions. First, can machine-learning methods improve the accuracy of sovereign spread forecasts relative to a theory-consistent error-correction model? Second, do improvements in statistical accuracy translate into better cross-sectional ranking and higher risk-adjusted portfolio performance?

To address these questions, I develop an empirical framework that combines econometric structure with modern predictive methods. I begin with a country-specific error-correction model (ECM) that decomposes sovereign spread dynamics into a long-run relationship with macro-financial fundamentals and a short-run adjustment component. This specification ensures that forecasts remain consistent with the economic structure documented in the sovereign-spread literature. I then extend this framework by applying XGBoost, a gradient-boosted tree algorithm, to the short-run forecasting equation while retaining the error-correction term. This hybrid ML–ECM approach preserves the long-run discipline imposed by cointegration while allowing for nonlinearities and interactions in short-run dynamics.

Forecasts from both models are evaluated using two complementary criteria: statistical accuracy and economic performance. From an econometric perspective, forecasting performance is assessed using out-of-sample R^2 , mean squared error (MSE), and directional hit ratios. These metrics capture whether the models deliver statistically superior predictions of monthly sovereign spread changes and correctly anticipate their direction.

In addition, I evaluate cross-sectional ranking accuracy by measuring how often each model correctly classifies countries into the top and bottom terciles of realized spread changes, corresponding to the strongest wideners and tighteners. This ranking dimension is central because the profitability of long–short strategies depends on the relative ordering of countries rather than

on point-forecast precision.

From an economic perspective, model forecasts are translated into long–short sovereign bond portfolios. Each month, countries are ranked according to their predicted spread changes and assigned to an equal-weighted portfolio that goes long sovereigns expected to tighten and short sovereigns expected to widen. Portfolio performance is evaluated relative to the iBoxx Euro Sovereign Index under several interest-rate hedging configurations. These include an unhedged benchmark as well as portfolios that neutralize duration, convexity, and basis-point value, both individually and jointly. Risk-adjusted performance is assessed using Sharpe ratios, Sortino ratios, and higher-moment statistics, allowing for a direct evaluation of downside risk and tail behavior.

All models are estimated using a strict expanding-window procedure to ensure temporal consistency and to avoid look-ahead bias. The forecasting exercise begins in January 2015 after a 60-month initialization period and extends through December 2024. This sample covers several distinct macroeconomic regimes, including the post-crisis recovery, the negative interest-rate environment, the pandemic shock, and the subsequent monetary tightening cycle. Portfolio performance is evaluated relative to the iBoxx Euro Sovereign Index.

The empirical analysis focuses on ten Eurozone sovereign issuers—Austria, Belgium, Finland, France, Greece, Ireland, Italy, the Netherlands, Portugal, and Spain—observed at monthly frequency from January 2010 to December 2024. The analysis is restricted to euro-denominated sovereign bonds issued by Eurozone member states, which removes exchange-rate movements as an adjustment mechanism. In a monetary union, a common monetary authority sets policy for economies with heterogeneous fiscal positions, economic structures, and sensitivities to macro-financial shocks. As a result, sovereign issuers are exposed to the same global and regional conditions but lack an independent exchange-rate channel to absorb imbalances.

In this setting, relative sovereign spread movements are shaped primarily by differences in how countries respond to common shocks rather than by idiosyncratic developments alone. The absence of exchange-rate adjustment implies that variation in exposure to changes in global risk appetite, financial stress, and geopolitical events plays a central role in determining which countries experience spread widening or tightening at a given point in time. From an investment perspective, country-specific fundamentals often reflect the outcomes of these shocks rather than their initial source.

This perspective informs the empirical design of the study. Global and regional risk variables are treated as the primary drivers of short-run sovereign spread changes, while country-specific variables are included as conditioning variables that govern the heterogeneous transmission of common shocks across issuers. This approach is consistent with the empirical literature, which finds that global risk measures and regional financial conditions explain a substantial share of sovereign spread variation at short horizons.

The empirical results show a clear disconnect between statistical forecasting accuracy and economic performance. While the ML–ECM delivers sizeable improvements in out-of-sample predictive accuracy relative to the linear ECM, these gains do not consistently translate into higher risk-adjusted portfolio returns. In several cases, the simpler econometric model produces

comparable or higher Sharpe and Sortino ratios. This pattern indicates that lower forecast errors do not necessarily imply better identification of relative tighteners and wideners. Additional evidence suggests that models identify spread-widening episodes more reliably than spread-tightening episodes, reflecting asymmetries in the distribution of spread changes and the prominence of stress-driven dynamics.

This thesis contributes to the literature in three ways. Methodologically, it integrates error-correction models with machine-learning techniques in the context of Eurozone sovereign spreads. Empirically, it evaluates forecasting performance over a challenging period characterized by regime shifts and structural breaks, using both statistical metrics and interest-rate-hedged portfolio strategies. Conceptually, it highlights the importance of aligning forecasting objectives with economic decision-making, showing that for long–short strategies, correctly ranking relative tightening and widening pressures is more important than minimizing point-forecast errors.

The remainder of the thesis is organized as follows. Section 2 reviews the related literature. Section 3 describes the data. Section 4 presents the empirical methodology. Section 5 reports the results. Section 6 concludes with limitations and directions for future research. The appendices provide technical details, robustness checks, and alternative model specifications.

2. Literature Review and Theoretical Framework

The theoretical foundation for understanding sovereign spread dynamics in the Eurozone is rooted in Mundell’s theory of optimal currency areas (Mundell, 1961). Mundell argues that a monetary union can function smoothly only if member countries possess effective mechanisms to absorb asymmetric shocks. These mechanisms include labor mobility, fiscal transfers, or flexible exchange rates that allow relative price adjustment. When countries adopt a common currency, they give up the exchange-rate channel, making alternative adjustment mechanisms essential for macroeconomic stability.

The Eurozone lacks several of these mechanisms. Labor mobility remains limited, and fiscal integration is weak, with no centralized system of automatic transfers. Most importantly, member states relinquished independent monetary policy and exchange-rate flexibility upon adopting the euro. In this context, Stiglitz (2016) argues that the institutional design of the Eurozone creates a structural vulnerability. When common shocks—such as financial crises or shifts in global risk appetite—hit economies with heterogeneous structures and fiscal capacities, the absence of exchange-rate adjustment forces the burden of adjustment onto internal channels, including unemployment, wage compression, and fiscal austerity.

This institutional setting changes how sovereign spreads behave in a monetary union. Rather than reflecting ex-ante differences in fiscal discipline alone, spreads increasingly capture market perceptions of each country’s ability to absorb common shocks without the stabilizing mechanism of currency adjustment. When adverse shocks occur, countries with weaker banking systems or less flexible labor markets experience prolonged adjustment periods. Fiscal deterioration and rising debt ratios often emerge endogenously as a result of these constraints rather than as the initial cause of spread widening. From this perspective, country-specific fundamentals act

primarily as conditioning variables, while global and regional shocks serve as the main drivers of spread variation.

A large empirical literature supports the dominant role of global variables in sovereign spread dynamics. Longstaff et al. (2011) analyze sovereign credit default swap (CDS) spreads for 26 countries and show that a common global variables explains more than 60% of their variation. This global component is closely linked to U.S. equity market volatility and corporate credit spreads, while traditional fiscal indicators play a secondary role. The authors conclude that sovereign credit risk behaves largely as a global asset class driven by international risk conditions.

Within the Eurozone, De Grauwe and Ji (2013) document that during the 2010–2012 sovereign debt crisis, spreads became disconnected from fiscal fundamentals and instead displayed self-fulfilling dynamics driven by panic and contagion. Countries with stable fiscal positions experienced sharp spread increases, and the correlation between spreads and debt ratios weakened during periods of stress. These findings are consistent with the institutional critique of the Eurozone, highlighting the vulnerability of sovereign bond markets to liquidity-driven crises in the absence of a lender of last resort.

Further evidence of regime dependence is provided by Arghyrou and Kontonikas (2012), who show that the sensitivity of Eurozone spreads to global and domestic variables varies over time. Before the global financial crisis, spreads were largely insensitive to fiscal indicators. After 2008, markets began pricing fiscal risks more aggressively as policy constraints became more salient. Similarly, Izadi and Hassan (2018) find that global risk measures, such as the VIX and U.S. high-yield spreads, consistently explain sovereign yield spreads across developed economies, while domestic inflation variability plays a more limited role.

Afonso et al. (2012, 2019) further document structural breaks in the determinants of Eurozone sovereign spreads. Prior to 2007, fiscal variables had little explanatory power. After the crisis, fiscal imbalances and liquidity risk became statistically significant, reflecting changes in market functioning and the emergence of liquidity premia. Taken together, this literature shows that sovereign spread dynamics are highly regime-dependent, with global and regional shocks playing a central role and country-specific fundamentals acting mainly as conditioning variables.

Against this backdrop, the error-correction framework provides a natural econometric structure for modeling sovereign spreads. Engle and Granger (1987) show that when nonstationary variables are cointegrated, their short-run dynamics can be represented through an error-correction model in which deviations from long-run equilibrium influence subsequent adjustments. This framework allows researchers to separate long-run relationships from short-run fluctuations while imposing economically meaningful structure.

In the context of sovereign spreads, cointegration captures the idea that spreads are anchored to macro-financial fundamentals in the long run, even if short-run deviations are large and persistent. The error-correction term measures disequilibrium or stress, while the speed-of-adjustment parameter governs how quickly spreads revert toward equilibrium. In a monetary union, where adjustment through exchange rates is unavailable, this process may be slow and uneven across countries.

While cointegration methods are often implemented in multivariate systems (Johansen, 1991), this study adopts country-specific error-correction models to allow for heterogeneity in long-run relationships and adjustment dynamics across Eurozone issuers. This choice is consistent with both the institutional constraints emphasized by Stiglitz and the empirical evidence of heterogeneous responses to common shocks within the monetary union.

Alongside econometric approaches, machine-learning methods have gained prominence in asset pricing due to their ability to model nonlinearities, interactions, and state-dependent relationships. Gu et al. (2020) provide a comprehensive comparison of machine-learning techniques for equity return prediction and show that nonlinear methods, particularly tree-based models, outperform linear benchmarks in environments with complex predictive structures. However, they also emphasize that the benefits of machine learning depend on data richness and signal strength.

In sovereign bond markets, Belly et al. (2023) apply machine-learning methods to Eurozone spreads and find that ensemble tree models, especially XGBoost, significantly reduce forecast errors relative to traditional econometric models. Their analysis focuses on point-forecast accuracy and does not examine whether these statistical gains translate into improved portfolio performance or cross-sectional ranking. Gomez-Gonzalez et al. (2024) similarly show that global financial stress indicators dominate predictive importance when using machine-learning techniques, reinforcing the central role of common shocks.

More broadly, the asset-pricing literature highlights a persistent disconnect between statistical forecast accuracy and economic profitability. Welch and Goyal (2008) document that many predictors fail to improve out-of-sample equity premium forecasts, while Campbell and Thompson (2008) show that statistical improvements generate economic value only when forecasts are used effectively in portfolio construction. These findings underscore that accuracy alone is not sufficient for economic relevance.

Despite this evidence, existing studies typically evaluate either forecasting accuracy or portfolio performance, but rarely both in a unified framework for Eurozone sovereign spreads. The present study fills this gap by examining whether improvements in sovereign spread forecasting accuracy translate into better cross-sectional ranking of spread tightenings and widenings and into superior risk-adjusted portfolio performance across alternative hedging configurations.

3. Data

3.1. Overview and Coverage

The dataset covers ten Eurozone sovereign issuers—Austria, Belgium, Finland, France, Greece, Ireland, Italy, the Netherlands, Portugal, and Spain—which together represent the main government bond markets in the monetary union. I focus on 10-year euro-denominated benchmark bonds to ensure comparability across issuers and a consistent basis for measuring sovereign yield spreads.

Following the literature, countries are classified into core and periphery groups to reflect

differences in fiscal positions, macroeconomic resilience, and sovereign risk perceptions. The core group comprises Austria, Belgium, Finland, France, and the Netherlands, while the periphery includes Greece, Ireland, Italy, Portugal, and Spain. This classification is used to interpret cross-country heterogeneity in spread dynamics.

The sample spans January 2010 to December 2024, yielding 179 monthly observations per country and a total of 1,790 country–month observations. This period covers several distinct macro-financial regimes, including the European sovereign debt crisis, the era of unconventional monetary policy and negative interest rates, the pandemic shock, and the subsequent monetary tightening cycle. As a result, the data capture sovereign spread behavior across markedly different economic environments.

Explanatory variables are grouped into three categories: (i) global and regional variables, capturing shocks common to all issuers; (ii) country-specific variables, reflecting domestic macroeconomic and fiscal conditions; and (iii) bond-level characteristics, summarizing pricing and interest-rate risk features. These categories align with the theoretical framework discussed in Section 2 and are detailed in the following subsections.

Portfolio returns are computed from monthly mid-price changes obtained from Refinitiv. Although dirty prices include accrued interest, mid-prices are consistently available across all sovereigns and therefore provide a uniform and comparable return measure.

3.2. Data Sources, Frequency, and Pre-processing

The majority of variables are obtained from the Refinitiv Workspace API, which provides harmonized definitions and consistent coverage across all ten countries. Two global financial variables—the U.S. High Yield Index option-adjusted spread and the VIX—are sourced from the Federal Reserve Economic Data (FRED) database. Together, these sources yield a coherent dataset spanning global risk variables, country-specific fundamentals, and bond characteristics.

Most variables are available at monthly frequency, while fiscal and macroeconomic indicators are released annually or quarterly. To maintain a realistic information set and avoid look-ahead bias, annual variables—specifically GDP growth and the debt-to-GDP ratio—are lagged by one year. Monthly inflation and reserve measures are lagged by one month. As a result, all regressors reflect information available to an investor at time t .

The econometric treatment of the data depends on the purpose of each estimation. When sovereign spreads are modeled in levels, regressors are also expressed in levels to preserve the long-run interpretation of coefficients. In contrast, when forecasting monthly changes in spreads, the dependent variable is expressed in first differences and all explanatory variables are transformed accordingly. This alignment allows the model to capture short-run dynamics, which are directly relevant for one-month-ahead forecasting and for the long–short trading strategies developed later.

Across all specifications, regressors are standardized prior to estimation. Standardization places variables on a comparable scale, improves numerical stability, and prevents differences in units from mechanically affecting coefficient magnitudes—an important consideration in

forecasting environments with heterogeneous predictors.

3.3. Global and Regional Variables

The first group comprises variables that capture global and Euro-area-wide financial conditions affecting all sovereign issuers simultaneously. These variables summarize common shocks arising from changes in global risk appetite, regional credit conditions, equity-market performance, and monetary policy.

Global credit conditions are proxied by the Bank of America U.S. High Yield Index option-adjusted spread, which reflects compensation for bearing non-investment-grade credit risk at the global level. Regional credit risk is captured by the iTraxx Europe 5-year CDS index, while conditions in the cash corporate bond market are summarized by the iBoxx EUR Corporates Index. Together, these variables reflect shifts in funding conditions and risk premia relevant for sovereign bond pricing.

Risk sentiment is measured using implied-volatility indices. The VIX captures global equity-market uncertainty, while the VSTOXX reflects risk conditions specific to the Eurozone. To complement these indicators, I include the STOXX 50 return, which tracks aggregate Eurozone equity performance and proxies changes in growth expectations.

Interest-rate conditions are represented by both global and regional benchmarks. The U.S. 10-year Treasury yield serves as a global risk-free reference, while the German 10-year Bund yield represents the natural risk-free benchmark within the euro area. Short-term liquidity conditions and monetary-policy expectations are captured by the 3-month Euribor. Eurozone CPI inflation is also included, as it directly affects ECB policy expectations and the term structure of interest rates.

Collectively, these global and regional variables capture the systematic component of sovereign spread variation driven by external and Euro-area-wide financial conditions.

Table 1: Global and Regional Factors.

Variable	Description	Frequency	Source
VIX	U.S. equity volatility index	Monthly	FRED
VSTOXX	European equity volatility index	Monthly	Refinitiv
STOXX50	European equity index return	Monthly	Refinitiv
U.S High Yield Spread	Global credit spread	Monthly	FRED
iTraxx Europe 5y CDS	European corporate credit risk	Monthly	Refinitiv
Iboxx EUR Corp	Euro corporate bond index	Monthly	Refinitiv
Euribor 3M	Eurozone money market rate	Monthly	Refinitiv
Eurozone CPI	Eurozone inflation rate	Monthly	Refinitiv
U.S Yield 10Y	U.S. risk-free rate	Monthly	Refinitiv
Germany Yield 10Y	German risk-free rate	Monthly	Refinitiv

Notes: Variables capturing systematic risk factors that affect all euro-area sovereign bonds simultaneously.

3.4. Country-Specific Variables

The second group consists of country-level variables capturing domestic macroeconomic conditions, fiscal positions, and financial resilience. These variables vary across sovereign issuers and condition how individual countries respond to common global and regional shocks.

Domestic equity-market returns are used to proxy country-specific financial conditions and investor sentiment, as equity prices respond rapidly to changes in growth prospects and perceived risk. Domestic price stability is captured by the Harmonized Index of Consumer Prices (HICP), which influences real yields and short-run monetary-policy expectations.

Economic activity is measured using official GDP growth rates, while fiscal sustainability is proxied by the debt-to-GDP ratio. Both variables are available at annual frequency and are lagged by one year to ensure consistency with the information set available to investors at time t . Although slow-moving, these indicators anchor long-run solvency perceptions and shape market assessments of adjustment capacity.

External liquidity is measured using foreign-exchange reserves and total reserves, which proxy a sovereign's ability to absorb external shocks and meet short-term obligations. These variables are particularly relevant for countries historically more exposed to external refinancing pressures.

Together, these country-specific variables summarize domestic conditions that differentiate sovereign risk across Eurozone member states.

Table 2: Country-Specific Factors.

Variable	Description	Frequency	Source
Stock Index	National equity market return	Monthly	Refinitiv
HCPI	Domestic inflation rate	Monthly	Refinitiv
Forex Reserves	Foreign-exchange reserve assets	Monthly	Refinitiv
Total Reserves	Total reserve assets	Monthly	Refinitiv
GDP Growth Rate	Economic growth rate	Yearly	Refinitiv
Debt-to-GDP Ratio	Public debt ratio	Yearly	Refinitiv

Notes: Variables capturing macroeconomic and fiscal conditions specific to each sovereign issuer.

3.5. Bond-Level Structural Characteristics

The third group comprises bond-level characteristics that capture pricing dynamics and interest-rate sensitivity. The analysis focuses on 10-year sovereign benchmark bonds, which represent each country's long-term financing cost and form the basis for computing spreads relative to Germany. Monthly returns are constructed using Refinitiv mid-prices, ensuring comparability across issuers.

Interest-rate exposure is summarized using duration, convexity, and basis point value (BPV). Duration measures first-order sensitivity to yield changes, convexity captures nonlinear price

responses to large yield movements, and BPV measures the absolute price change associated with a one-basis-point shift in yields. Together, these measures characterize the key dimensions of interest-rate risk embedded in sovereign bond prices and play a central role in the construction of hedged portfolio strategies.

Table 3: Bond-Level Structural Characteristics.

Variable	Description	Frequency	Source
Yield	10Y sovereign bond yield	Monthly	Refinitiv
Mid Price	Bond mid-market price	Monthly	Refinitiv
Duration	Interest rate sensitivity (1st order)	Monthly	Refinitiv
Convexity	Interest rate sensitivity (2nd order)	Monthly	Refinitiv
BPV	Price change per basis point	Monthly	Refinitiv

Notes: Variables measuring interest rate sensitivity and structural features of sovereign bonds.

3.6. Variable Construction

3.6.1 Dependent Variable: Sovereign Spread

Following standard practice in the sovereign bond literature, the sovereign spread is defined as the difference between each country’s 10-year benchmark yield and the corresponding German Bund yield:

$$\text{Spread}_{i,t} = \text{Yield}_{i,t} - \text{Yield}_{\text{Germany},t}. \quad (1)$$

Sovereign spreads exhibit strong persistence and are predominantly nonstationary. Accordingly, the empirical analysis focuses on monthly changes in spreads:

$$\Delta \text{Spread}_{i,t} = \text{Spread}_{i,t} - \text{Spread}_{i,t-1}. \quad (2)$$

Working with first differences serves both econometric and economic purposes. From a statistical perspective, differencing avoids spurious regression concerns. From an economic perspective, modeling changes in spreads rather than levels prevents the forecasting framework from mechanically favoring structurally high-spread sovereigns on the long side and low-spread sovereigns on the short side. Instead, the specification isolates short-run widening and tightening dynamics, aligning the dependent variable with the one-month horizon and cross-sectional ranking objective of the long–short trading strategy.

3.6.2 Constructed Independent Variables

Several explanatory variables are constructed to summarize relative risk conditions, policy differentials, and liquidity characteristics not directly observable from raw series.

Implied Volatility Premium (IVP). The implied volatility premium captures relative regional risk aversion by comparing European and global equity-market uncertainty:

$$IVP_t = VSTOXX_t - VIX_t. \quad (3)$$

Positive values indicate periods in which implied volatility in the Eurozone exceeds that of the United States, signaling heightened regional risk sentiment.

Bund–Treasury Premium (BTP). To capture differences in long-term risk-free benchmarks and monetary-policy stances between Europe and the United States, I define:

$$BTP_t = Y_{GER,t} - Y_{USA,t}. \quad (4)$$

This measure reflects relative safe-asset demand and divergence in expected policy paths between the ECB and the Federal Reserve.

Variance Risk Premium (VRP). Following Longstaff et al. (2011), adapted to a European context, the variance risk premium is constructed as the difference between implied variance derived from the VSTOXX and realized variance computed from STOXX 50 returns:

$$VRP_t = IV_t - RV_t. \quad (5)$$

Realized variance is estimated using the Garman and Klass (1980) open–high–low–close estimator over a rolling 20-day window. The VRP captures compensation for bearing variance risk and proxies shifts in market-wide risk premia.

Reserves Ratio. To proxy a sovereign’s short-term external liquidity position, I construct the ratio of foreign-exchange reserves to total reserves:

$$\text{Reserves Ratio}_{i,t} = \frac{\text{FX Reserves}_{i,t}}{\text{Total Reserves}_{i,t}}. \quad (6)$$

Higher values indicate a larger buffer of liquid foreign assets and greater resilience to external funding shocks.

Bid–Ask Spread. Market liquidity conditions are captured by the bid–ask spread on sovereign yields:

$$\text{Bid–Ask Spread}_{i,t} = \text{Ask Yield}_{i,t} - \text{Bid Yield}_{i,t}. \quad (7)$$

Wider bid–ask spreads indicate reduced secondary-market liquidity and heightened uncertainty among market participants.

3.7. Descriptive Statistics

Table 4 reports descriptive statistics for sovereign spreads and their first differences. Spread levels exhibit strong persistence across all countries, with an average first-order serial correlation of approximately 0.94. Augmented Dickey–Fuller (ADF) tests indicate that most spread series fail to reject the unit-root null, while complementary KPSS tests do not support stationarity in levels. Taken together, the evidence supports treating sovereign spreads as $I(1)$ processes.

To ensure a consistent treatment across countries and to avoid spurious inference, the forecasting analysis therefore focuses on monthly changes in spreads, $\Delta\text{Spread}_{i,t}$.

Persistence varies substantially across issuers. Half-life estimates implied by AR(1) coefficients range from 5.18 months for the Netherlands to 39.70 months for Ireland, with an average of 16.84 months and a median of 14.57 months. The difference between the mean and median reflects the presence of highly persistent cases, particularly among periphery countries.

Table 4: Descriptive Statistics and Unit-Root Properties: Sovereign Yield Spread (10-Year).

Panel A: Sovereign Yield Spread (10-Year).

Country	Mean	SD	Min	Median	Max	Serial Corr.	Half-life	ADF Stat.
Austria (AT)	0.378	0.218	0.060	0.308	1.276	0.937	10.604	-2.746*
Belgium (BE)	0.607	0.445	0.147	0.494	2.749	0.954	14.569	-2.268
Finland (FI)	0.277	0.145	-0.008	0.247	0.662	0.917	8.164	-2.694*
France (FR)	0.468	0.219	0.202	0.392	1.311	0.925	9.455	-2.351
Greece (GR)	6.363	6.146	0.843	4.803	31.442	0.959	17.262	-1.855
Ireland (IE)	1.536	2.004	0.207	0.570	8.846	0.982	39.700	-3.270**
Italy (IT)	1.918	0.871	0.775	1.639	5.215	0.936	10.231	-2.476
Netherlands (NL)	0.222	0.109	0.039	0.201	0.572	0.875	5.182	-2.412
Portugal (PT)	2.655	2.623	0.408	1.627	14.054	0.977	31.282	-1.513
Spain (ES)	1.528	1.047	0.617	1.130	5.432	0.968	22.033	-1.919

Panel B: Sovereign ΔSpread (10-Year).

Country	Mean	SD	Min	Median	Max	Serial Corr.	Half-life	ADF Stat.
Austria (AT)	-0.000	0.078	-0.221	-0.006	0.279	-0.223	∞	-6.392***
Belgium (BE)	0.002	0.137	-0.519	-0.010	0.766	-0.102	∞	-3.926***
Finland (FI)	0.001	0.059	-0.196	-0.002	0.209	-0.145	∞	-11.557***
France (FR)	0.003	0.084	-0.329	-0.001	0.379	-0.185	∞	-4.676***
Greece (GR)	-0.008	1.761	-12.648	-0.046	10.558	-0.030	∞	-4.354***
Ireland (IE)	-0.007	0.384	-2.015	-0.014	2.428	0.018	0.173	-6.199***
Italy (IT)	0.002	0.311	-1.037	-0.024	1.507	-0.032	∞	-13.707***
Netherlands (NL)	0.000	0.055	-0.173	-0.006	0.163	-0.193	∞	-9.747***
Portugal (PT)	-0.001	0.563	-2.547	-0.021	2.903	-0.144	∞	-4.120***
Spain (ES)	0.001	0.265	-0.947	-0.020	1.194	-0.128	∞	-3.201**

Notes: Panel A reports descriptive statistics for 10-year sovereign yield spreads relative to Germany. Panel B reports descriptive statistics for monthly changes (Δspread). The sample covers January 2010–December 2024, and all series contain 179 monthly observations. Serial Corr. is the first-order autocorrelation coefficient. Half-life is computed as $\ln(0.5)/\ln(\theta)$ for $0 < \theta < 1$ and set to ∞ when $\theta \leq 0$ or $\theta \geq 1$. Significance: *** $p < 0.01$, ** $p < 0.05$, * $p < 0.10$.

Volatility also differs markedly across countries, consistent with the core–periphery distinction. Greece exhibits the highest spread volatility, while the Netherlands shows the lowest. The distribution of ΔSpread is asymmetric: the median monthly change is negative, indicating that modest spread tightening is more frequent, whereas the mean is slightly positive, reflecting that widening episodes, while less frequent, tend to be larger in magnitude.

Differencing substantially reduces persistence. Serial correlation in ΔSpread becomes negative for most countries and close to zero for Ireland, consistent with a stationary process driven by short-run fluctuations. Large spread movements are concentrated in the 2010–2014 sovereign debt crisis, while the post-2015 period exhibits fewer extreme observations. These properties indicate that ΔSpread provides a suitable dependent variable for short-horizon forecasting and cross-sectional ranking exercises.

Overall, the data display strong persistence in spread levels, asymmetry between tightening and widening episodes, and substantial cross-country heterogeneity. These features call for a modeling framework that accommodates both long-run relationships and short-run dynamics, while allowing for heterogeneous responses to common shocks. The next section develops the econometric and machine-learning methodology used to address these challenges.

4. Methodology

4.1. Overview of Empirical Strategy

The empirical objective of this thesis is to assess whether monthly changes in Euro-area sovereign bond spreads can be forecasted and whether such forecasts translate into economically meaningful trading strategies. I develop an integrated framework that combines econometric modeling, machine-learning techniques, and fixed-income portfolio construction. Forecasts are generated country-by-country using a one-step-ahead expanding-window procedure, with the first out-of-sample forecast produced in January 2015 after a minimum estimation window of 60 months.

The forecasting target is the monthly change in the 10-year sovereign spread relative to Germany, $\Delta\text{Spread}_{i,t}$, chosen for its stationarity and its direct relevance for the long–short portfolio strategy. All explanatory variables follow the lagging conventions established in the Data section, ensuring consistency with the investor information set.

The empirical strategy proceeds in two stages. First, I estimate a country-specific Error-Correction Model (ECM) that decomposes spread dynamics into a long-run cointegrating relationship and a short-run adjustment mechanism. The ECM is re-estimated each month using the expanding window and serves as a linear, theory-consistent benchmark for forecasting $\Delta\text{Spread}_{i,t+1}$.

Second, I extend the ECM framework using XGBoost, a gradient-boosted tree model designed to capture nonlinearities, interaction effects, and regime-dependent behavior in macro-financial data. The resulting ML–ECM preserves the long-run and short-run structure of the ECM while allowing the short-run relationship between predictors and $\Delta\text{Spread}_{i,t+1}$ to vary flexibly.

Hyperparameters are selected using an expanding-window time-series cross-validation procedure.

Forecasts from the ECM and ML–ECM are translated into long–short sovereign bond portfolios. Each month, countries are ranked by their predicted spread changes, with equal-weighted long positions in the strongest expected tightenings and short positions in the strongest expected widenings. To account for interest-rate risk, I construct a set of hedged portfolios that neutralize duration (D), convexity (C), and basis-point value (B) exposures. Seven hedging configurations are considered: D , C , B , DC , DB , CB , and DCB . Portfolio weights are obtained by imposing linear constraints and solving the resulting optimization problems using the CVXPY framework (Diamond and Boyd, 2016).

Portfolio performance is evaluated relative to the iBoxx Euro-area Sovereign Index using annualized mean returns, annualized volatility, Sharpe ratios, Sortino ratios, and higher-moment statistics, which together summarize risk-adjusted performance, downside risk, and distributional properties of returns.

Figure 1 provides a schematic overview of the empirical workflow. The next section introduces the econometric Error-Correction Model (ECM), which constitutes the first stage of the analysis.

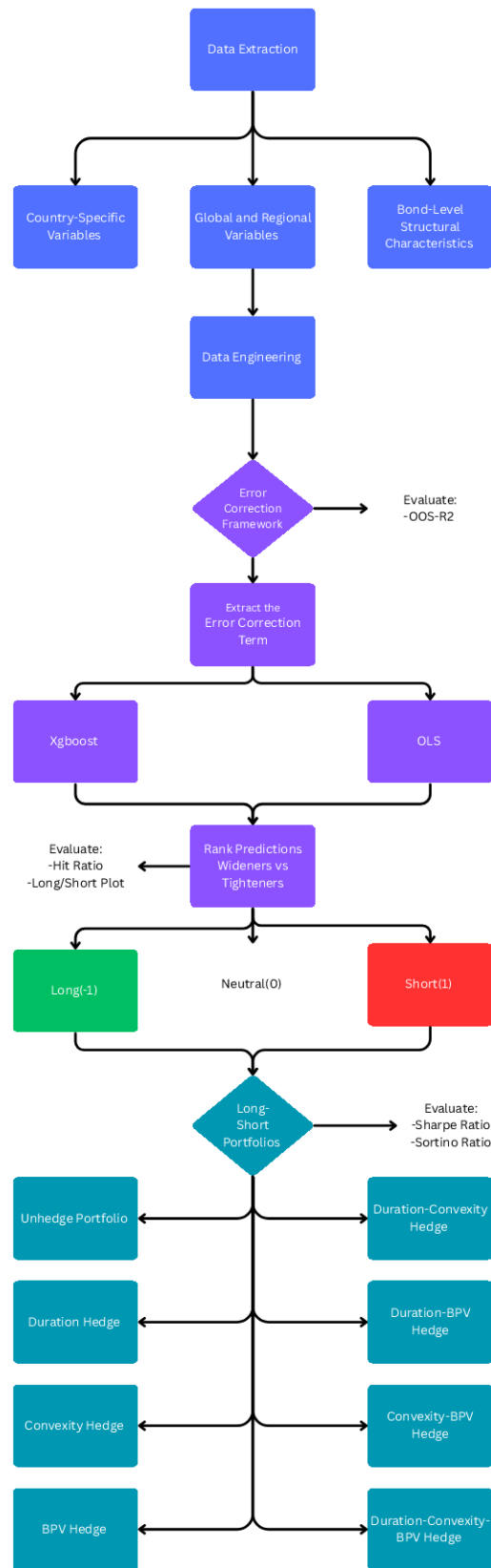


Figure 1: Modeling and Evaluation Pipeline.

4.2. Error-Correction Model (ECM)

4.2.1 Long-Run Cointegration

The first step of the empirical analysis models the long-run relationship between sovereign bond spreads and their macro–financial determinants. When nonstationary variables share a stable equilibrium relationship such that deviations from this relationship are stationary, the variables are said to be cointegrated (Engle and Granger, 1987). In the context of sovereign spreads, cointegration reflects the idea that spreads may drift with macro–financial conditions in the long run, while deviations from fundamentals tend to correct over time.

For each country i , the long-run equilibrium relationship is estimated using the static OLS regression:

$$\text{Spread}_t^{(i)} = \beta' X_t^{(i)} + u_t^{(i)},$$

where $\text{Spread}_t^{(i)}$ denotes the 10-year sovereign spread relative to Germany, $X_t^{(i)}$ is a vector of long-run global, regional, and country-specific fundamentals, and $u_t^{(i)}$ is the deviation from long-run equilibrium.

Consistent with the evidence presented in the Data section, sovereign spreads and their long-run determinants are predominantly $I(1)$, supporting the use of a cointegration-based specification.

The vector $X_t^{(i)}$ includes the following long-run determinants of sovereign spreads, selected based on the empirical literature and the institutional features of the Eurozone:

- **Implied Volatility Premium (VSTOXX–VIX)**, capturing relative risk aversion and regional uncertainty;
- **iTraxx Europe 5-year CDS index**, reflecting market-implied default risk for European investment-grade corporates;
- **iBoxx Euro corporate yield**, proxying regional corporate credit conditions;
- **Bund–Treasury yield spread**, measuring relative safe-asset valuations between Europe and the United States;
- **3-month Euribor rate**, capturing short-term monetary and liquidity conditions in the euro area;
- **Euro-area CPI inflation**, reflecting macroeconomic fundamentals and ECB policy expectations;
- **German 10-year yield**, representing the euro-area risk-free benchmark and flight-to-quality dynamics;
- **Domestic inflation (HICP)**, capturing country-specific price dynamics;
- **Debt-to-GDP ratio**, proxying long-run fiscal sustainability;

- **Reserves ratio**, measuring external liquidity buffers and financial resilience.

Together, these variables summarize the global, regional, and domestic forces that anchor sovereign spreads in the long run. The estimated residual $u_t^{(i)}$ constitutes the Error-Correction Term (ECT) and measures deviations from long-run equilibrium.

4.2.2 Short-Run Dynamics and One-Step-Ahead Forecasting

Given the estimated long-run relationship, short-run dynamics are modeled using a country-specific ECM. For each country i , monthly changes in sovereign spreads are described by:

$$\Delta \text{Spread}_{t+1}^{(i)} = \gamma' \Delta X_t^{(i)} + \lambda' \Delta F_t^{(i)} + \phi^{(i)} u_t^{(i)} + \varepsilon_{t+1}^{(i)},$$

where $u_t^{(i)}$ denotes the lagged deviation from long-run equilibrium.

The vector $\Delta X_t^{(i)}$ contains first differences of the long-run fundamentals included in the cointegration relationship, ensuring consistency between the long-run and short-run components of the model. Since these variables are predominantly nonstationary in levels, differencing is required for valid short-run inference.

The vector $\Delta F_t^{(i)}$ includes additional predictors that are stationary or weakly persistent and therefore unsuitable for inclusion in the long-run equilibrium. These variables capture short-horizon fluctuations relevant for forecasting and include:

- **Bid–ask spread** – capturing bond market liquidity;
- **Domestic stock index return** – reflecting country-specific financial conditions;
- **GDP growth** – proxying short-term macroeconomic momentum;
- **Variance risk premium (VRP)** – capturing compensation for variance risk;
- **U.S. high-yield spread** – reflecting global credit risk.

The coefficient $\phi^{(i)}$ governs the speed of adjustment toward the long-run equilibrium. Negative values indicate mean reversion following shocks, while values close to zero imply slow or incomplete adjustment.

The ECM is estimated using an expanding-window procedure and produces one-step-ahead forecasts for $\Delta \text{Spread}_{t+1}^{(i)}$, which serve as the benchmark inputs for the portfolio construction described in Section 4.3.

4.2.3 XGBoost Model

The ECM offers a theory-consistent linear benchmark, but its structure assumes constant marginal effects and cannot accommodate the nonlinear and state-dependent behaviour that characterises sovereign spreads. In practice, the sensitivity of spreads to macro–financial

conditions varies substantially across episodes of market stress, shifts in monetary policy, and changes in liquidity conditions. These features motivate the use of a more flexible forecasting model that can adapt across regimes while retaining the economic structure of the ECM.

To address this, I employ Extreme Gradient Boosting (XGBoost) (Chen, 2016), a regularised ensemble of decision trees widely used in empirical finance. XGBoost is well suited for this context for three reasons. First, tree-based models naturally capture nonlinearities and interactions: the effect of a predictor can change depending on the values of other variables or the prevailing market environment. Second, boosted trees are robust to structural breaks and shifts in relationships—features that have defined the euro-area sovereign bond market over the last decade, including the sovereign debt crisis, unconventional monetary policy, the negative-rate period, the pandemic shock, and the recent tightening cycle. Third, recent evidence shows that bond risk premia and yield changes often depend on interactions, thresholds, and tail behaviour that linear models cannot represent (Gu et al., 2020; Belly et al., 2023).

A *regression tree* partitions the predictor space into a set of regions defined by threshold rules on the input variables and assigns a constant prediction within each region. This structure allows the model to approximate nonlinear relationships and discrete regime shifts in a transparent way. *Boosting* builds a sequence of such trees iteratively, where each new tree is fitted to the residuals of the previous ones. By aggregating many weak learners, boosting produces a flexible function that improves forecast accuracy while controlling overfitting through regularisation.

Formally, XGBoost approximates the conditional expectation of the next-month change in spreads using an additive ensemble of regression trees:

$$\widehat{\Delta\text{Spread}}_{t+1}^{(i)} = \sum_{k=1}^K f_k(x_t^{(i)}), \quad f_k \in \mathcal{F},$$

where $x_t^{(i)}$ collects the ECM predictors, $x_t^{(i)} = (\Delta X_t^{(i)}, \Delta F_t^{(i)}, u_t^{(i)})$, namely the first differences of long-run fundamentals, the first differences of short-run variables, and the Error-Correction Term. The set $\mathcal{F}$ denotes the space of regression trees; each tree captures local nonlinear threshold effects, and boosting aggregates them into a smooth and flexible forecasting rule.

This hybrid design preserves the economic intuition embedded in the ECM while allowing the model to adjust its response across macro-financial states. As a result, XGBoost can capture nonlinear adjustment speeds, varying sensitivities to fundamentals, and interaction effects that are empirically relevant for sovereign spread dynamics. The next sections describe the hyperparameter tuning and real-time forecasting procedures used to implement the model.

4.2.4 Hyperparameter Tuning and Time-Series Cross-Validation

The performance of XGBoost depends on a set of hyperparameters controlling tree depth, learning rate, the number of boosting iterations, and subsampling rates. Hyperparameters are tuned using a time-series cross-validation scheme based on an expanding estimation window.

A minimum estimation window of 60 months is imposed before the first out-of-sample

forecast. Within this window, hyperparameters are selected using a single hold-out validation split ($CV = 1$), where the sample is divided into a training set comprising the first two-thirds of the observations and a validation set containing the remaining one-third.

As the estimation window expands over time, the 2/3–1/3 training–validation proportion is kept fixed. New observations are therefore not systematically added to the training sample only; instead, observations may enter either the training or the validation set depending on the timing, ensuring that the relative size of the validation block remains constant across windows.

The hyperparameter configuration that minimises the validation loss is retained. The model is then re-estimated on the full 60-month window and used to generate the first one-step-ahead forecast (month 61). Thereafter, the estimation sample expands recursively, and forecasts are produced strictly out of sample following the same expanding-window procedure.

Figure 2 illustrates the cross-validation and forecasting design. The expanding training sample is shown in blue, the validation block in green, and the red diagonal represents the sequence of out-of-sample forecasts.

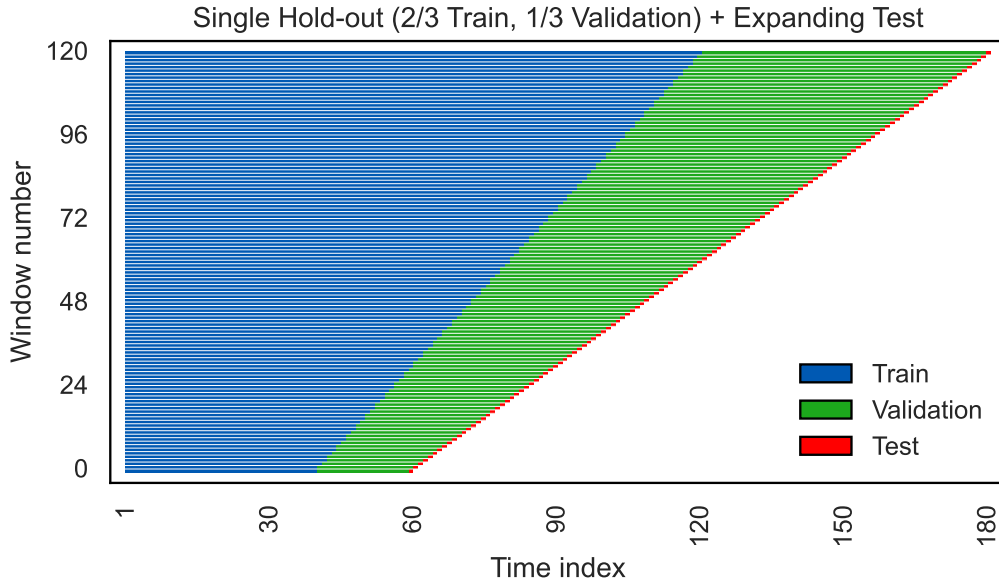


Figure 2: Expanding-window cross-validation scheme with a single hold-out validation split used for hyperparameter tuning ($CV=1$).

Further implementation details, including the hyperparameter grid and the `RandomizedSearchCV` setup, are provided in Appendix A.1. The use of a single validation split reflects a trade-off between preserving estimation sample size and maintaining temporal ordering in a relatively short and persistent time series.

4.2.5 Forecasting Procedure

With hyperparameters fixed, the XGBoost model is implemented within the same expanding-window forecasting structure used for the ECM. Let t_0 denote the first month in which the

minimum estimation window is satisfied. For each $t \geq t_0$, the model is re-estimated using all observations available up to period t and produces a one-step-ahead forecast for month $t + 1$.

For each country i , the predictor vector is defined as

$$z_t^{(i)} = \left(\Delta X_t^{(i)}, \Delta F_t^{(i)}, u_t^{(i)} \right),$$

where $\Delta X_t^{(i)}$ denotes first differences of the long-run fundamentals, $\Delta F_t^{(i)}$ contains additional country-specific short-run predictors, and $u_t^{(i)}$ is the Error-Correction Term capturing deviations from long-run equilibrium.

XGBoost maps this information set into a nonlinear forecast of the next-month change in the sovereign spread:

$$\widehat{\Delta \text{Spread}}_{t+1}^{(i)} = \widehat{f}\left(z_t^{(i)}\right),$$

where $\widehat{f}(\cdot)$ represents the fitted boosted-tree ensemble estimated over the expanding window.

Repeating this procedure over time yields a sequence of out-of-sample forecasts $\{\widehat{\Delta \text{Spread}}_{t+1}^{(i)}\}$ for each country. These forecasts serve as direct inputs to the portfolio construction framework described in Section 4.3, where sovereigns expected to experience spread tightening are assigned long positions and those expected to widen are assigned short positions.

4.3. Portfolio Construction

Forecasts generated by the ECM and ML-ECM models are translated into sovereign bond portfolio strategies to evaluate whether predictive signals carry economic value. The baseline specification is an equal-weighted long-short strategy based on predicted cross-sectional movements in sovereign spreads. Section 4.3.2 extends this baseline by introducing interest-rate hedging constraints that isolate the contribution of relative spread predictability from broader yield-curve movements.

4.3.1 Long-Short Portfolio Strategy

One-step-ahead forecasts of spread changes are converted into trading signals using the inverse relationship between bond prices and yields. A forecasted spread tightening implies a relative decline in yields and therefore a positive bond return, while a forecasted spread widening implies a negative return.

At each month t , predicted spread changes $\widehat{\Delta \text{Spread}}_{i,t+1}$ are sorted cross-sectionally into terciles:

- **Bottom tercile:** countries expected to tighten the most (long positions),
- **Top tercile:** countries expected to widen the most (short positions).

With $N = 10$ countries, each tercile contains approximately three sovereigns. In the equal-weighted implementation, long positions receive weight $+1/L$ and short positions $-1/S$, where

L and S denote the number of countries in the long and short legs, respectively. When $L = S = 3$, each position carries weight $\pm 1/6$. This construction yields a market-neutral portfolio satisfying

$$\sum_{i=1}^N w_{i,t} = 0, \quad \sum_{i=1}^N |w_{i,t}| = 1,$$

so that returns reflect purely cross-country relative performance.

Monthly portfolio returns are computed as

$$R_{p,t+1} = \sum_{i=1}^N w_{i,t} R_{i,t+1},$$

where $R_{i,t+1}$ denotes the realized return on sovereign bond i between t and $t + 1$. This long–short strategy serves as the unhedged benchmark.

4.3.2 Interest-Rate Hedging Strategies

To determine whether forecasting performance reflects genuine spread predictability rather than coincidental exposure to interest-rate movements, the baseline portfolio is augmented with hedged variants that neutralize different dimensions of interest-rate risk. These hedges remove return components driven by common yield-curve movements and ensure that performance reflects relative spread dynamics.

Hedging is implemented by imposing linear neutrality constraints on duration, convexity, and basis-point value (BPV). Duration captures sensitivity to parallel shifts in yields, convexity captures curvature effects, and BPV reflects the monetary impact of small yield changes. Constraints are imposed individually and jointly, yielding seven configurations: D , C , B , DC , DB , CB , and DCB .

Let $w_t = (w_{1,t}, \dots, w_{N,t})'$ denote the vector of portfolio weights, L_t the set of long positions, and S_t the set of short positions. Weights satisfy

$$\|w_t\|_1 = 1, \quad w_{i,t} \geq 0 \text{ for } i \in L_t, \quad w_{i,t} \leq 0 \text{ for } i \in S_t, \quad w_{i,t} = 0 \text{ otherwise.}$$

Neutrality constraints take the form

$$d'w_t = 0, \quad c'w_t = 0, \quad b'w_t = 0,$$

where d , c , and b denote vectors of durations, convexities, and BPVs, respectively. Joint hedges impose multiple constraints simultaneously. Importantly, the optimization preserves the sign of each position, ensuring that hedging alters exposure magnitudes but not the directional content of the forecasting signal.

All hedged portfolios are implemented using the CVXPY convex optimization framework (Diamond and Boyd, 2016). Comparing performance across increasingly restrictive hedging

schemes allows for a clear assessment of whether predictive signals generate economic value independent of interest-rate risk.

4.4. Performance Evaluation

To assess the economic value of the forecasting signals, portfolio performance is benchmarked against the iBoxx Euro Sovereign Index. The iBoxx Euro Sovereign Index is a widely used passive benchmark that tracks the market-value-weighted performance of euro-denominated sovereign bonds issued by Eurozone central governments. The index is constructed and maintained by IHS Markit and is designed to represent the broad investment-grade sovereign bond market in the euro area across countries and maturities. In this study, the iBoxx index is used exclusively as a passive reference point, providing a natural baseline for evaluating whether active long–short strategies based on spread forecasts generate economic value beyond aggregate market exposure.

I evaluate the full family of portfolio strategies introduced in Section 4.3. In total, this yields sixteen portfolios: eight based on forecasts from the ECM and eight based on forecasts from the ML–ECM. For each forecasting model, I consider an unhedged equal-weighted long–short strategy and seven interest-rate-hedged variants that neutralize duration (D), convexity (C), basis-point value (B), and their combinations (DC, DB, CB, DCB). Performance is summarized using a set of complementary metrics that capture returns, risk, downside exposure, and distributional characteristics.

First, I report annualized mean returns and annualized volatility, obtained by rescaling average monthly returns and monthly standard deviations to a yearly frequency following standard practice in the fixed-income literature.

Risk-adjusted performance is evaluated using the Sharpe ratio, defined as annualized mean return divided by annualized volatility. Because all portfolios are long–short and self-financing, the Sharpe ratio directly reflects the quality of the forecasting signal without requiring an explicit risk-free adjustment.

To account explicitly for downside risk, I also report the Sortino ratio, which replaces total volatility with downside volatility computed from negative returns only. The Sortino ratio is particularly informative in sovereign bond markets, where spread widening episodes tend to be abrupt and severe, while tightening episodes are typically more gradual. By focusing on downside variation, the Sortino ratio provides a more economically relevant assessment of performance during stress periods.

To characterize the shape of the return distribution, I report skewness and kurtosis. Skewness measures asymmetry in returns, while kurtosis captures tail risk—an important feature given the presence of crisis episodes, flight-to-quality dynamics, and non-normal return behavior in sovereign bond markets.

In addition to return-based metrics, I report trading frequency by country to provide descriptive information on how often each sovereign enters the long or short leg of the portfolio under each forecasting model. Table 9 summarizes the number of months in which a country is assigned to the tightening (long) or widening (short) tercile under the ECM and ML–ECM specifications.

These statistics offer a transparent view of the cross-sectional allocation patterns generated by each model and serve as a complementary diagnostic to the portfolio-level performance results.

Taken together, these metrics provide a comprehensive evaluation of the economic performance of the forecasting strategies. The next section presents the empirical results, focusing on whether the ML–ECM improves the identification of relative spread tighteners and wideners and whether interest-rate hedging enhances the robustness of portfolio returns.

5. Results

5.1. Overview of Results

This section evaluates whether the ECM and ML–ECM models generate meaningful forecasting improvements and whether these gains translate into superior economic performance. The analysis proceeds along three tightly linked dimensions. First, I assess traditional out-of-sample forecasting accuracy, measured through the OOS R^2 , to evaluate whether accounting for long-run cointegration and nonlinearities improves short-horizon spread predictions. Second, I examine cross-sectional ranking ability, which is central for identifying the countries expected to experience the strongest spread tightening or widening and therefore underpins the construction of relative-value strategies. Third, I map these forecasts into sixteen long–short sovereign bond portfolios—both unhedged and hedged against duration, convexity, and BPV risk—and evaluate their risk-adjusted performance.

Several broad conclusions emerge from this analysis. First, the two-stage cointegration framework is strongly supported by the data. Rolling ADF diagnostics indicate that first-stage residuals are stationary for the vast majority of the sample, validating the use of an error-correction mechanism as a state variable in spread forecasting. Consistent with this structure, feature contribution results show that one-month spread changes are driven primarily by global and regional risk conditions—captured by the variance risk premium, corporate credit spreads, and the euro-area interest-rate environment—while the error-correction term ranks among the most informative predictors in both the linear and machine-learning specifications. Country-specific fundamentals play a secondary role at the monthly horizon, indicating that short-run spread dynamics are dominated by common risk pricing and long-run adjustment forces.

Second, the machine-learning extension delivers clear gains in traditional forecasting accuracy relative to the linear ECM. While country-level OOS R^2 values remain negative for both models—highlighting the low signal-to-noise ratio of monthly spread changes—the ML–ECM consistently reduces forecast errors. The aggregate reduction in mean squared error relative to the ECM is economically large (around 59.5%), with the strongest improvements occurring precisely where the linear benchmark performs worst. At the same time, neither model outperforms a simple expanding historical average in R^2 terms, underscoring that short-horizon level forecasting remains intrinsically challenging even when nonlinearities and interactions are exploited.

Third, improvements in point-forecast accuracy do not mechanically translate into superior cross-sectional ranking performance. Evidence from heatmaps, tercile-agreement matrices,

hit ratios, and Spearman rank correlations indicates that both models struggle to recover the exact cross-country ordering of monthly spread changes, with average rank correlations close to zero and overall hit ratios around one-third. Nonetheless, the ranking results are not devoid of structure. Both models generate persistent signals—partly driven by the error-correction term—and correct short signals occur more frequently than correct long signals, implying that the models are relatively more informative about spread widening than tightening episodes. This asymmetry is economically relevant, as identifying prospective wideners strengthens the short leg of long–short portfolios and helps protect performance in adverse market conditions. More broadly, the results highlight a central theme of the analysis: profitable cross-sectional allocation does not require precise point forecasts, but rather a balance between accuracy and generalisation in an environment where realised spread changes are small and tightly clustered.

Finally, the portfolio evidence confirms that both the ECM and ML–ECM embed economically useful information, but that realised performance depends critically on how forecasting signals are translated into exposures. Hedging interest-rate risk typically reduces both mean returns and volatility substantially; in some cases this improves risk-adjusted performance, while in many others unhedged or lightly hedged strategies remain competitive. The ECM delivers relatively stable Sharpe ratios across hedging designs, whereas ML–ECM performance is more sensitive to the hedging choice, highlighting the need to understand which risk components drive the country selections implied by each model. Across all strategies, return distributions exhibit positive skewness alongside fat tails, with heavier tails more pronounced under the ECM. Taken together, these findings show that while the ML–ECM improves statistical forecasting accuracy, economic value ultimately arises from the interaction between forecasting signals, cross-sectional ranking, and portfolio construction rather than from any single performance metric in isolation.

5.2. Econometric Performance

5.2.1 Error–Correction Validity

Before assessing forecasting performance, it is important to verify that the error–correction term (ECT) in the ECM specification is stationary over the sample. Figure 3 reports rolling augmented Dickey–Fuller (ADF) test outcomes for the ECM residuals using an expanding-window procedure. Purple cells indicate rejection of the unit-root null ($I(0)$ residuals), while grey cells indicate windows in which the residuals are classified as non-stationary ($I(1)$).

The figure shows that for nearly all countries and estimation windows, the ECM residuals are classified as stationary. Only a small number of isolated episodes exhibit non-rejection of the unit-root null, such as brief windows for Ireland around 2015–2016 and Belgium around 2022. These cases appear to reflect temporary instabilities or limited effective sample sizes in short expanding windows rather than a persistent breakdown of the long-run relationship.

Overall, the rolling ADF evidence supports the presence of a stable cointegrating relationship between sovereign spreads and the set of macro–financial fundamentals used in the ECM. This validates the inclusion of the error–correction term as a stationary driver of short-run spread dynamics throughout the forecasting sample.

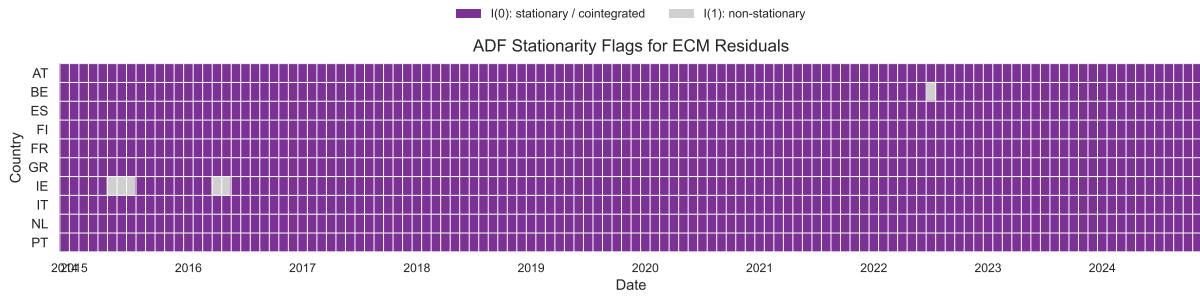


Figure 3: Rolling ADF stationarity flags for ECM residuals. Purple cells indicate I(0) residuals (stationary / cointegrated), while grey cells indicate I(1) residuals (non-stationary).

Appendix B reports detailed country-level ADF statistics for the variables entering the long-run equation. Most series are classified as I(1). The iTraxx Europe 5-year CDS index occasionally rejects the unit-root null at the 10% level; however, complementary KPSS tests do not provide strong evidence in favor of stationarity. Given its central economic role as a regional credit-risk indicator, iTraxx is retained in the long-run specification.

5.2.2 Magnitude of Feature Contributions

To complement the forecasting and ranking results presented above, Figure 4 reports the average absolute contribution of each predictor in the ECM and ML–ECM models. For the ECM, feature importance is measured as the average absolute value of the estimated coefficients across all out-of-sample forecasting windows and countries, rescaled so that the weights sum to one. For the ML–ECM, the same averaging and rescaling procedure is applied to the XGBoost feature-importance measure. These magnitudes should therefore be interpreted as indicators of relative predictive relevance for one-month-ahead spread changes rather than as structural elasticities.

Across both models, a clear pattern emerges: global and regional variables account for the largest share of predictive importance. The variance risk premium (VRP), U.S. high-yield spreads, German sovereign yields, and corporate credit indicators —particularly the iBoxx and iTraxx indices—consistently rank among the most important variables. These variables capture broad movements in global risk appetite, credit-market stress, and the euro-area interest-rate environment, which are central drivers of short-horizon sovereign spread dynamics.

A second key result is the prominent role of the error-correction term (ECT). In both the ECM and the ML–ECM, the ECT—reported as *Country Residual* in Figure 4—ranks among the top predictors. Its importance indicates that deviations from long-run equilibrium conditions play a significant role in shaping near-term spread adjustments, consistent with mean-reversion dynamics emphasized in the theoretical framework.

While the set of leading predictors is broadly similar across models, the distribution of importance differs. The linear ECM assigns relatively concentrated weight to a small set of persistent global credit-risk indicators, such as iTraxx, iBoxx, and U.S. high-yield spreads. In contrast, the ML–ECM allocates importance more evenly across predictors, including interest-rate

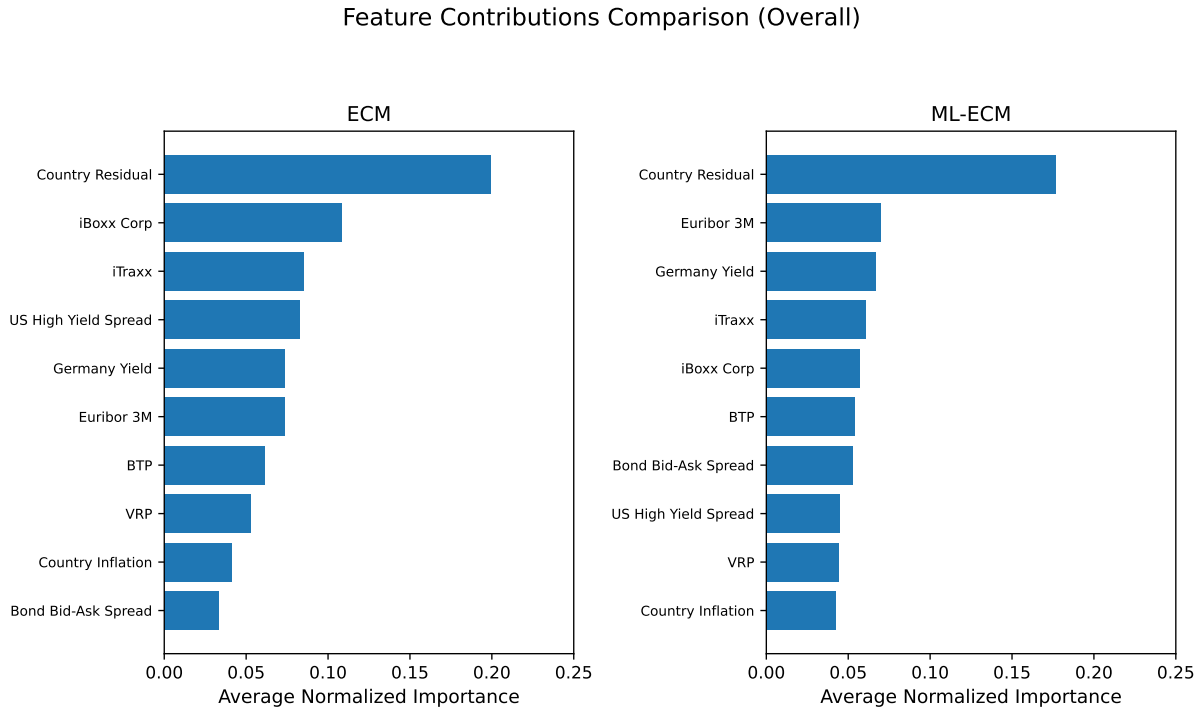


Figure 4: Average absolute feature contributions in the ECM and ML–ECM models.

level variables such as German yields and Euribor 3M. This reflects the nonlinear model’s ability to capture interactions between monetary conditions, global risk pricing, and country-level spread responses.

Country-specific variables contribute less to overall predictive importance in both models. Among these variables, domestic inflation appears most consistently, suggesting that short-run price dynamics retain some predictive content at the monthly frequency. At the bond level, the bid–ask spread emerges as the most relevant liquidity proxy, indicating that market liquidity conditions affect short-term spread movements, albeit to a lesser extent than global and long-run forces.

A country-by-country decomposition of feature contributions, reported in Appendix B.1, reveals additional heterogeneity. Across countries, the ordering of key predictors remains broadly stable, with the ECT ranking among the top three variables in both models for nearly all sovereigns. An exception arises in the ML–ECM for Ireland, where euro-area inflation, Euribor 3M, and the country residual jointly dominate predictive importance. For the ECM, the ECT accounts for more than 20% of total importance in Austria, Belgium, Ireland, Italy, Portugal, and the Netherlands. In the ML–ECM, only Austria and France allocate a comparable share of importance to a single variable, reflecting a more even distribution of weights across predictors.

Overall, these results reinforce the central empirical finding of the study: short-horizon sovereign spread changes are driven primarily by global and regional risk conditions, reinforced by long-run adjustment pressures, while country-specific fundamentals play a secondary and conditioning role. The ML–ECM captures this structure through a more balanced allocation of importance across interacting drivers, whereas the ECM concentrates predictive power in a

smaller set of linear variables.

5.2.3 Out-of-Sample R^2

Table 5 reports country-level out-of-sample R^2 statistics for the linear ECM benchmark (OLS) and the ML–ECM (XGBoost), together with a relative performance measure based on mean squared forecast errors. The first two rows present the out-of-sample R^2 values obtained by each model for the ten countries, while the third row reports the proportional reduction in mean squared error achieved by the ML–ECM relative to the ECM.

Table 5: Out-of-sample R^2 Statistics (ECM vs ML–ECM).

	AT	BE	ES	FI	FR	GR	IE	IT	NL	PT	Overall
ECM R^2	-0.805	-2.906	-0.845	-0.298	-1.414	-3.935	-16.396	-0.843	-0.723	-5.899	-3.775
ML–ECM R^2	-0.197	-0.426	-0.189	-0.181	-0.149	-0.948	-3.877	-0.477	-0.075	-1.732	-0.933
ML–ECM vs ECM R^2	0.337	0.635	0.355	0.090	0.524	0.605	0.720	0.199	0.376	0.604	0.595

Notes: Country-by-country out-of-sample R^2 statistics for the ECM benchmark (OLS) and the ML–ECM (XGBoost). Forecasts are generated from January 2015 to November 2024 using an expanding-window scheme. The relative out-of-sample statistic ML–ECM vs ECM R^2 is computed as $1 - \text{MSE}_{\text{XGB}}/\text{MSE}_{\text{OLS}}$.

Across all countries, the ML–ECM delivers higher (i.e. less negative) out-of-sample R^2 values than the linear ECM, indicating systematic improvements in forecast accuracy despite the inherent difficulty of predicting one-month changes in sovereign spreads. The smallest improvement is observed for Finland, where the reduction in mean squared error amounts to 0.090, reflecting the limited scope for gains given the already weak predictive performance of the linear benchmark. In contrast, the largest improvement is achieved for Ireland, with a reduction of 0.720. This result is consistent with the particularly poor performance of the ECM for this country ($R^2 = -16.396$) relative to the ML–ECM ($R^2 = -3.877$), leaving greater room for error reduction. At the aggregate level, the ML–ECM improves overall forecasting accuracy by 59.5% relative to the linear ECM.

The overall out-of-sample R^2 reported in the final column differs from the country-level statistics in that it is computed by pooling all country–month observations prior to evaluating forecast fit. As such, it reflects model performance at the euro-area level rather than assigning equal weight to each country. Using this pooled metric, the ECM attains an overall R^2 of -3.775 , while the ML–ECM improves this figure to -0.933 . The corresponding relative statistic (0.595) indicates a substantial reduction in aggregate forecast errors when nonlinearities are introduced through the machine-learning specification.

Both models perform particularly poorly for Ireland and Portugal, suggesting that spread dynamics in these countries remain difficult to capture with the available information set. More generally, there is no sharp distinction between core and periphery countries in terms of relative forecast improvements. However, average performance is somewhat stronger for core countries, consistent with the notion that global and regional variables—which dominate the information set—provide greater explanatory power for core sovereigns, while periphery spreads may depend more heavily on country-specific dynamics that are harder to model at a monthly frequency.

As an additional robustness check, Appendix C reports results from unrestricted “Kitchen Sink” specifications that exclude the error-correction term and rely exclusively on short-run dynamics. Dropping the ECT improves out-of-sample R^2 in both linear and machine-learning settings, although performance remains negative relative to the historical-mean benchmark. In particular, the Kitchen Sink OLS improves the ECM by 30.3% (from -3.775 to -2.327), while the ML–KS improves upon the ML–ECM by 35.0% (from -0.933 to -0.254). These results suggest that imposing long-run restrictions does not necessarily maximize point-forecast accuracy at short horizons, even though such restrictions may remain valuable for economic interpretation and ranking-based applications.

Finally, it is important to note that neither the ECM nor the ML–ECM outperforms a simple expanding historical average in terms of out-of-sample R^2 . This finding highlights the low signal-to-noise ratio inherent in short-horizon sovereign spread changes. Nevertheless, the ML–ECM delivers a marked improvement over the linear ECM benchmark, consistent with the ability of gradient-boosted trees to exploit nonlinearities and interaction effects, even when overall predictive power remains limited.

5.2.4 Ranking Performance and Tercile Agreement

To evaluate the cross-sectional signals generated by each model, Figures 5 and 6 display the long and short positions assigned to each country over time. Both models exhibit substantial persistence in their signals: once a country is classified as a tightener or widener, the corresponding position is typically maintained for several consecutive months rather than switching frequently.

Persistence is particularly pronounced for periphery sovereigns such as Greece, Portugal, and Italy, where spread dynamics tend to evolve in sustained phases. Core countries also exhibit persistence, though to a lesser extent and with more frequent neutral positioning. Comparing the two heatmaps, the linear ECM in Figure 5 displays stronger overall persistence, with positions often remaining unchanged across months. By contrast, the ML–ECM signals in Figure 6 appear more adaptive, with persistent exposures concentrated in episodes of sustained spread tightening or widening rather than mechanically carried forward.

This behaviour is consistent with the prominent role of the error-correction term (ECT) in both models. When spreads remain away from their long-run equilibrium, the models infer that adjustment is ongoing, generating naturally persistent signals until convergence progresses.

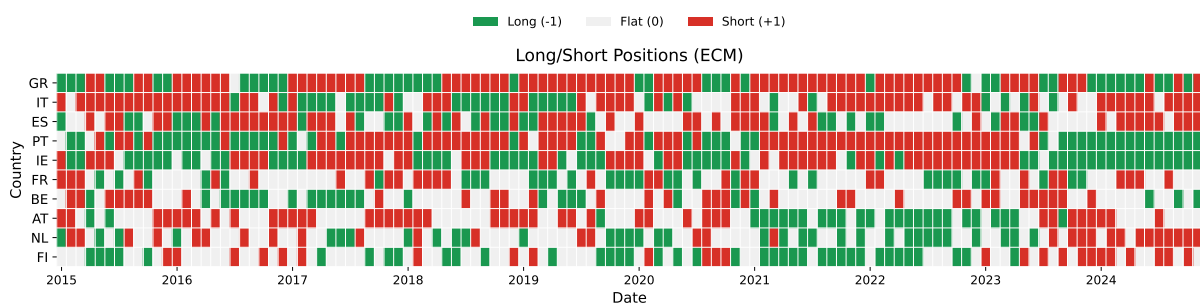


Figure 5: Long and Short Positions (ECM).

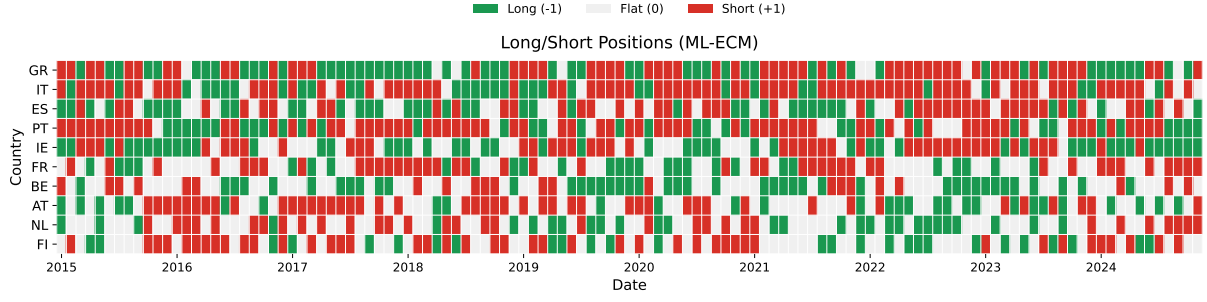


Figure 6: Long and Short Positions (ML-ECM).

Figures 7 and 8 assess tercile agreement between predicted and realised spread changes. Blue cells indicate correct long signals (countries in the bottom realised tercile, corresponding to spread tightening), while orange cells indicate correct short signals (countries in the top realised tercile, corresponding to spread widening). Grey cells denote misclassifications or months in which the country does not fall into an extreme tercile.

A notable asymmetry emerges: orange cells are more prevalent than blue cells in both models. This indicates that ranking performance is stronger for identifying spread wideners than spread tighteners. From an allocation perspective, this property is valuable, as correctly flagging prospective wideners improves the short leg of the portfolio and helps limit downside exposure during adverse market conditions.

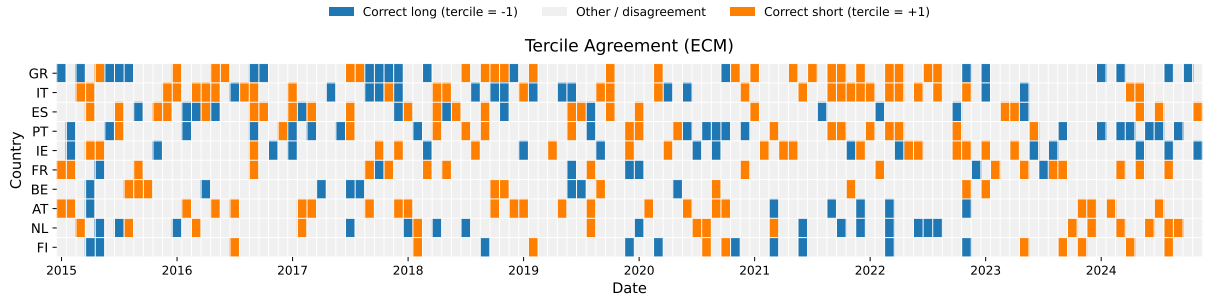


Figure 7: Tercile Agreement (ECM).

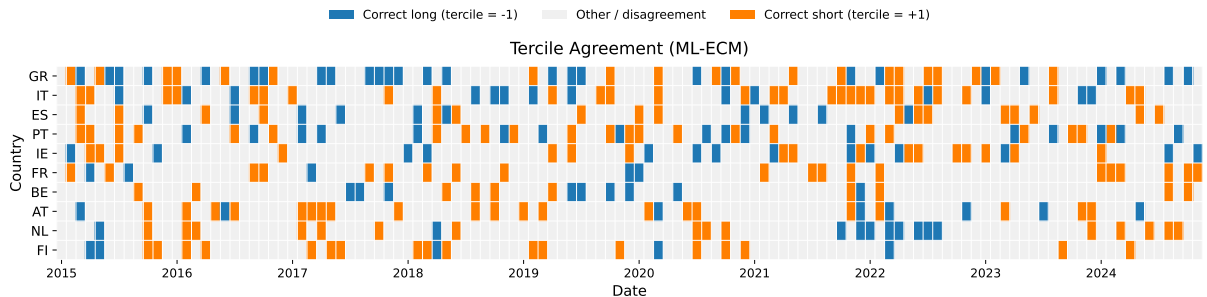


Figure 8: Tercile Agreement (ML-ECM).

Table 6 reports tercile hit ratios by country. Despite the substantial improvements in out-of-sample R^2 delivered by the ML-ECM, directional accuracy gains are mixed. The ML-ECM

outperforms the ECM for Belgium, Finland, Greece, and Portugal, matches it for Austria and France, and underperforms for Spain, Ireland, Italy, and the Netherlands. At the aggregate level, the ECM achieves a slightly higher overall hit ratio (0.385) than the ML–ECM (0.370), implying that both models correctly classify roughly one-third of active country–month observations.

Table 6: Hit Ratio by Country (ECM vs ML–ECM).

	AT	BE	ES	FI	FR	GR	IE	IT	NL	PT	Overall
ECM Hit Ratio	0.372	0.305	0.440	0.282	0.305	0.415	0.473	0.434	0.375	0.390	0.385
ML–ECM Hit Ratio	0.372	0.322	0.333	0.324	0.305	0.453	0.419	0.406	0.312	0.400	0.370

Notes: The hit ratio for each country is the number of correctly matched active terciles (actual = −1 or +1) divided by the number of periods in which the actual tercile is active (actual ≠ 0). The Overall value reports the pooled performance of the model using all countries and dates together.

Cross-sectional ranking performance is further assessed using the Spearman rank correlation between predicted and realised Δ Spread. The average Spearman correlation is -0.078 for the ECM and -0.027 for the ML–ECM. Both values are close to zero, indicating that precise cross-sectional ranking of monthly spread changes remains challenging, although the ML–ECM exhibits a modest improvement.

Results for the Kitchen Sink specifications reported in Appendix C.1 and Appendix C.2 are broadly consistent with these findings. In the linear case, removing the ECT slightly worsens average hit ratios, while the nonlinear Kitchen Sink specification improves directional accuracy modestly, particularly for Finland, Greece, and Portugal. However, absolute differences remain small.

Taken together, the heatmaps, tercile-agreement plots, hit ratios, and rank correlations indicate that improvements in point-forecast accuracy do not mechanically translate into superior ranking performance. While accurate level forecasts are not theoretically required for effective ranking, the tight clustering of monthly sovereign spread changes around zero makes rankings highly sensitive to noise.

To illustrate this point, consider a simple example. Suppose that Spain, France, and Italy realise monthly spread changes of -0.03 , 0.00 , and 0.03 , respectively. A model may generate noisy level forecasts such as 1.0 , 1.4 , and 3.4 , resulting in a relatively large mean-squared error. Nevertheless, the model would still rank the cross-section correctly by identifying Italy as the largest expected widener and Spain as the strongest expected tightener. In this setting, weak point-forecast accuracy does not preclude economically meaningful ranking signals for long–short portfolio allocation.

The ranking results suggest that, although the ML–ECM substantially reduces out-of-sample forecast errors, these gains do not translate mechanically into better tercile allocation. One potential limitation is that forecasts are generated country by country, without explicitly exploiting cross-sectional information. To address this concern, Appendix D presents a panel-regression specification that jointly models the cross-section, reported in Table D.1. The results are very similar: even after accounting for cross-sectional dependence, the hit ratio remains close to 0.376 .

This indicates that the ranking performance is not materially affected by whether the model is estimated individually or jointly across countries.

Taken together, these findings reinforce the idea that high R^2 accuracy is not a sufficient condition for effective ranking. While a minimum degree of forecast discipline is clearly required, the key difficulty arises from the nature of the signal itself. Monthly sovereign spread changes are often very small, measured in only a few basis points. In this setting, even minor deviations from the true value can easily alter the cross-sectional ordering, leading to unstable tercile assignments. As a result, improvements in point forecast accuracy may compress predicted values without necessarily improving the model's ability to separate countries cleanly into tighteners and wideners.

These results therefore point to a broader limitation of cross-sectional forecasting in low-volatility environments. When the signal-to-noise ratio is weak, ranking performance depends not only on forecast accuracy but also on how models translate small numerical differences into discrete allocation decisions. Specifications that emphasise flexibility may fit local movements more closely, while more regularised models may yield more stable ordering across countries. In this application, where spread changes are tightly clustered, this tension between precision and stability becomes a defining feature of the ranking problem.

The relevant question is therefore not whether a model achieves superior statistical accuracy in isolation, but whether its signals translate into economically meaningful portfolio outcomes. The next section evaluates this dimension directly by mapping the ECM and ML–ECM forecasts into long–short sovereign bond strategies and assessing their risk-adjusted performance.

5.3. Portfolio Performance.

Table 7 reports the performance of the 16 long–short asset-allocation strategies implied by the ECM and ML–ECM signals, together with the iBoxx Euro-area Sovereign Index, which serves as a passive benchmark. Over the sample period, the benchmark delivers low average returns and a Sharpe ratio of 0.065, reflecting the challenging environment faced by buy-and-hold sovereign investors. Volatility is largely downside-driven, indicating that return fluctuations are dominated by negative realizations rather than symmetric risk, while its weak skewness highlights the limited upside potential of passive exposure to euro-area sovereign debt.

The table also reports the performance of an expanding-average benchmark, which is included to mirror the naive forecasting rule commonly used to evaluate out-of-sample R^2 . While ECM and ML–ECM forecasts yield negative OOS R^2 relative to this benchmark—indicating no improvement in mean squared forecast error—the economic implications are markedly different. The expanding-average signal produces poor cross-sectional rankings, which translate into weak long–short portfolios and result in the only strategy in the analysis exhibiting both a negative Sharpe ratio and a negative Sortino ratio. This contrast highlights that statistical forecast accuracy relative to a naive benchmark is not a sufficient criterion for economic value; what matters is how forecast signals rank assets and translate into portfolio returns.

A first-order result for the active strategies is that hedging interest-rate exposures materially

affects both returns and risk. Once duration, convexity, and/or BPV are neutralised, average returns and volatility typically decline by roughly half relative to the unhedged specifications. In some cases, this reduction in volatility more than compensates for the loss in mean returns, leading to improved risk-adjusted performance; in many others, however, hedging reduces Sharpe ratios, indicating that unhedged portfolios already deliver competitive risk–return trade-offs in this setting.

Table 7: Portfolio performance (ECM and ML–ECM).

Panel A: Benchmarks

Benchmark	Mean (%)	Std (%)	Sharpe Ratio	Downside Std (%)	Sortino Ratio	Skewness	Kurtosis
iBoxx EUR Sovereign Debt	0.360	5.575	0.065	4.033	0.089	-0.363	0.681
Expanding Average	-3.143	9.943	-0.316***	7.909	-0.397	-0.770	4.523

Panel B: ECM Portfolios

Portfolio	Mean (%)	Std (%)	Sharpe Ratio	Downside Std (%)	Sortino Ratio	Skewness	Kurtosis
Unhedged	4.701	10.143	0.463***	6.005	0.783	0.135	4.707
Duration Hedged	2.177	4.956	0.439***	2.866	0.760	0.870	10.906
Convexity Hedged	1.926	4.959	0.389***	2.964	0.650	0.597	9.326
BPV Hedged	2.309	4.690	0.492***	2.428	0.951	1.391	6.796
Duration–Convexity Hedged	2.484	5.353	0.464***	2.152	1.154	3.865	25.695
Duration–BPV Hedged	2.522	5.282	0.478***	2.815	0.896	1.257	8.525
Convexity–BPV Hedged	2.451	5.254	0.467***	2.850	0.860	1.036	7.375
Duration–Convexity–BPV Hedged	2.637	5.639	0.468***	2.950	0.894	1.982	15.582

Panel C: ML–ECM Portfolios

Portfolio	Mean (%)	Std (%)	Sharpe Ratio	Downside Std (%)	Sortino Ratio	Skewness	Kurtosis
Unhedged	5.570*	9.696	0.574***	4.978	1.119	0.857	2.105
Duration Hedged	1.828	4.461	0.410***	2.391	0.765	0.986	2.829
Convexity Hedged	1.746	4.463	0.391***	2.433	0.718	0.894	2.582
BPV Hedged	1.403	4.949	0.283***	2.633	0.533	1.680	7.101
Duration–Convexity Hedged	1.644	4.424	0.372***	2.506	0.656	0.811	3.496
Duration–BPV Hedged	1.007	4.669	0.216**	2.547	0.395	1.559	5.935
Convexity–BPV Hedged	1.473	4.683	0.314***	2.513	0.586	1.391	5.186
Duration–Convexity–BPV Hedged	2.478	5.431	0.456***	2.756	0.899	1.548	6.568

Notes: Mean and standard deviation (Std) are annualised and reported in percentage terms. Downside Std is computed using only negative returns. The Sortino ratio is computed as Mean divided by Downside Std. Sharpe ratios are tested against the null hypothesis that the Sharpe ratio equals zero; significance levels are denoted by * $p < 0.10$, ** $p < 0.05$, and *** $p < 0.01$.

Looking more closely at the ECM portfolios, performance is relatively stable across hedging strategies. Sharpe ratios range from 0.391 for the convexity-hedged portfolio to 0.497 for the BPV-hedged specification, suggesting that the linear ECM delivers consistent risk-adjusted returns regardless of the precise hedging design. Hedging interest-rate exposures generally improves Sortino ratios even when Sharpe ratios remain broadly unchanged, indicating that hedging primarily reduces downside risk rather than overall volatility. Hedged ECM portfolios also exhibit higher kurtosis and more pronounced positive skewness than the unhedged specification, pointing to distributions with occasional large positive outcomes rather than increased downside tail risk. Among the hedged ECM strategies, the duration–convexity neutral portfolio achieves the highest Sharpe and Sortino ratios, together with the highest kurtosis, reflecting infrequent but sizeable positive payoffs. By contrast, the unhedged ECM portfolio displays lower Sharpe and

Sortino ratios, making it less clear-cut as the dominant specification within the linear framework.

By contrast, the ML–ECM portfolios exhibit substantially greater dispersion in performance across hedging strategies. Sharpe ratios range from a low of 0.216 for the duration–BPV hedged portfolio to a high of 0.583 for the unhedged strategy, highlighting the sensitivity of nonlinear signals to how interest-rate exposures are treated. The unhedged ML–ECM portfolio also delivers the highest Sortino ratio, together with positive skewness and comparatively lower kurtosis, pointing to a smoother payoff profile with fewer extreme realisations. Hedged ML–ECM portfolios, in contrast, tend to display weaker downside-adjusted performance, suggesting that aggressive hedging removes economically relevant variation exploited by the nonlinear model. Overall, the Sortino-based evidence clearly identifies the unhedged ML–ECM specification as the preferred portfolio within the nonlinear framework.

Across all active strategies, return distributions display positive skewness, a desirable feature from an investor’s perspective, as performance is supported by occasional large gains rather than frequent extreme losses. Excess kurtosis is nevertheless non-negligible, indicating fat-tailed behaviour, which is more pronounced for ECM-based portfolios than for ML–ECM portfolios. Relative to the passive benchmark, downside volatility is approximately halved across all ECM and ML–ECM specifications, implying that positive and negative return fluctuations are of more comparable magnitude. This asymmetric improvement translates into substantially higher Sortino ratios, which never fall below 0.395, highlighting that active sovereign spread strategies not only outperform passive exposure in risk-adjusted terms, but do so by materially reducing downside risk.

To complement the performance results, Table 8 reports the average active portfolio weights by country across hedging specifications. The ECM strategies tend to be short more frequently in four of the five periphery countries (Spain, Greece, Italy, and Portugal) and also take short exposure in core issuers such as Austria and Belgium. By contrast, the ML–ECM is short in only one periphery country on average (Italy), while its short exposure is more concentrated in core countries—most notably Austria, Finland, and France. Two clarifications are important. First, the weights in the table do not sum to one because they are time averages of active exposures; portfolio weights are fully invested each month and always satisfy the normalisation constraint in the optimisation step. Second, average weights mask substantial time variation: different hedging constraints and changing cross-sectional signals can reallocate exposure materially from month to month. A noteworthy implication is that the ECM appears to implement, on average, a “short periphery / long core” tilt, whereas the ML–ECM exhibits the opposite pattern more often, yet both frameworks still produce economically meaningful portfolio results.

Finally, Table 9 reports trading frequencies by country, distinguishing how often each model assigns a country to the long or short tercile. Both models take short positions more often than long positions on average, consistent with earlier evidence that the ranking mechanism is relatively more effective at identifying widening episodes. Trading intensity is concentrated in a subset of countries: the ECM trades Greece, Ireland, and Portugal most actively, while the ML–ECM trades Greece, Italy, and Portugal most frequently. Although Greece and Portugal are heavily traded under both models, their long–short composition differs. For Greece, the ECM is

Table 8: Average Active Weights by Country (ECM and ML–ECM).**Panel A: ECM Weights (%)**

Portfolio	AT	BE	ES	FI	FR	GR	IE	IT	NL	PT
Unhedged	-3.197	-1.977	-4.137	3.153	0.470	-3.922	0.435	-4.956	1.079	-2.327
Duration Hedged	-2.046	0.129	-3.110	5.481	1.311	-1.003	1.823	-2.193	0.627	-0.378
Convexity Hedged	-2.489	-0.059	-3.546	5.884	0.938	-0.935	2.510	-2.863	1.556	-0.666
BPV Hedged	-1.230	-0.662	-3.408	5.176	2.202	-0.317	1.121	-1.129	0.975	-1.739
Duration–Convexity Hedged	-2.210	-0.740	-2.947	7.557	1.530	-0.401	1.969	-3.209	0.328	-3.459
Duration–BPV Hedged	-3.502	-0.688	-2.645	4.232	0.645	-1.602	0.828	-0.855	0.178	-0.910
Duration–Convexity–BPV Hedged	-3.463	-2.378	-2.682	4.473	0.483	-2.195	1.354	-4.103	1.487	-2.310

Panel B: ML–ECM Weights (%)

Portfolio	AT	BE	ES	FI	FR	GR	IE	IT	NL	PT
Unhedged	-3.206	4.451	-0.622	-3.078	-3.102	-0.198	0.851	-5.469	-0.712	-4.416
Duration Hedged	-2.584	6.971	1.294	-1.091	-0.616	1.163	2.354	-4.531	0.919	-3.317
Convexity Hedged	-2.255	6.606	0.864	-0.652	-1.116	1.472	2.171	-4.320	0.715	-2.773
BPV Hedged	-1.996	6.286	1.208	-1.949	0.289	1.108	2.298	-3.895	0.875	-3.788
Duration–Convexity Hedged	-1.678	6.649	0.719	-1.616	-0.228	1.063	1.655	-5.120	0.011	-3.389
Duration–BPV Hedged	-2.683	6.710	1.482	-2.541	-2.600	0.503	2.065	-4.544	-0.045	-2.811
Duration–Convexity–BPV Hedged	-3.906	6.769	-0.319	-3.336	-1.682	0.356	1.197	-5.021	-0.052	-2.952

Notes: Entries report average *active* portfolio weights by country for the ECM and ML–ECM long–short specifications. All weights are expressed in percentage terms. For country i in portfolio p , the reported value corresponds to $\bar{w}_{i,p} = \frac{1}{T_{i,p}} \sum_{t: w_{i,p,t} \neq 0} w_{i,p,t}$, where $w_{i,p,t}$ denotes the long or short position at time t and $T_{i,p}$ counts only periods with active trades ($w_{i,p,t} \neq 0$). At each date t , portfolio weights satisfy $\sum_i |w_{i,p,t}| = 100$.

more skewed toward short positions, whereas the ML–ECM is more balanced between long and short calls. Ireland stands out as the only country for which both models generate more long than short positions, even though the overall balance remains relatively close. Cross-model differences are also informative: the ML–ECM assigns more short signals to Finland than the ECM, while the ECM assigns more short signals to Greece than the ML–ECM, and the ML–ECM assigns more long signals to Belgium. These trading-frequency patterns help explain how different model structures translate into distinct country exposures while still delivering positive portfolio performance.

Table 9: Trading Frequency by Country (ECM vs ML–ECM).

Country	ECM			ML–ECM (XGB)			Δ (ECM – ML–ECM)		
	Long	Short	Total	Long	Short	Total	Δ Long	Δ Short	Δ Total
AT	27	43	70	29	46	75	-2	-3	-5
BE	25	34	59	43	22	65	-18	12	-6
ES	29	52	81	42	46	88	-13	6	-7
FI	39	25	64	27	43	70	12	-18	-6
FR	32	30	62	28	43	71	4	-13	-9
GR	43	74	117	55	56	111	-12	18	6
IE	56	52	108	49	44	93	7	8	15
IT	31	64	95	33	73	106	-2	-9	-11
NL	34	30	64	29	32	61	5	-2	3
PT	46	64	110	36	69	105	10	-5	5
Mean	36.2	45.8	83.0	37.1	47.4	84.5	-0.9	-1.6	-1.5

Notes: Trading frequencies report the number of periods in which a country is assigned to the long or short tercile by each model. “Total” equals the sum of long and short signals. Differences are defined as ECM – ML–ECM. Positive Δ values indicate that the ECM model generates more trading signals than the ML–ECM model, while negative values indicate the opposite.

Taken together, the portfolio-performance, weighting, and trading-frequency results illustrate how differences in statistical modelling translate into distinct but economically meaningful allocation patterns. While the ECM and ML–ECM generate markedly different country exposures and respond differently to hedging constraints, both frameworks deliver positive risk-adjusted performance when embedded in long–short sovereign bond portfolios. These findings reinforce a central theme of the empirical analysis: economic value depends not on any single modelling dimension, but on the interaction between forecasting signals, cross-sectional ranking, and portfolio construction choices.

5.4. Alternative classification-based approach.

In addition to the regression-based framework discussed above, Appendix Section E explores an alternative modelling strategy in which the problem is treated as a classification task. Rather than forecasting spread changes and ranking countries ex post, this approach directly predicts the tercile to which each country belongs using an ordered logit model. The resulting portfolio performance is broadly comparable to that obtained under the regression-based ECM and ML–ECM frameworks, suggesting that the main economic conclusions are not driven by the choice between regression and classification.

One notable difference is that, under the classification setting, a larger share of total volatility—around 60%—is attributable to downside movements. This comparison should be interpreted with caution, however, as predicting discrete labels mechanically pushes forecasts toward extreme positions. Effectively, the strategy behaves more like a binomial model over the

full sample, despite allowing for three outcome categories, which limits the comparability of risk characteristics across approaches.

Appendix Tables D.2, C.6, and C.3 further report the portfolio performance of panel-based and kitchen-sink specifications. The results are qualitatively similar across these alternatives. Among them, the kitchen-sink OLS specification stands out as the most noteworthy benchmark, as it delivers the most stable and highest Sharpe ratios across models while preserving key properties of the ECM portfolios, including stable performance and favourable downside-risk characteristics as reflected in Sortino ratios. By contrast, the ML–KS specification underperforms relative to the ML–ECM, and the panel-based portfolios perform worse than all four baseline models.

Taken together, these findings indicate that a simple linear regression model can generate economic performance comparable to that of more complex alternatives, consistent with the evidence presented in the previous section. Nevertheless, the ECM framework remains the preferred specification in this study, as it combines competitive economic performance with a strong theoretical foundation and a more comprehensive representation of both long- and short-run dynamics in sovereign spread behaviour.

6. Conclusion

This thesis develops an error-correction modeling framework to examine whether jointly accounting for long-run equilibrium relationships and short-run dynamics improves both forecasting performance and economic outcomes in euro-area sovereign bond spreads. By combining a traditional econometric ECM with a machine-learning extension, the analysis bridges economic theory and modern predictive methods. Crucially, model performance is evaluated not only through conventional statistical metrics, but also through cross-sectional ranking ability and portfolio-level outcomes. This integrated perspective allows predictive models to be assessed in terms of their economic usefulness rather than forecasting accuracy alone.

The empirical results show that introducing nonlinearities through the ML–ECM leads to substantial reductions in forecast error relative to the linear benchmark, even though out-of-sample R^2 values remain negative. These findings are consistent with the broader sovereign spread literature and reflect the inherently low signal-to-noise ratio that characterizes short-horizon spread changes. At the same time, improvements in statistical accuracy do not translate mechanically into superior cross-sectional ranking or uniformly stronger portfolio performance. Instead, economic value depends on how forecasting signals interact with persistence, asymmetries in spread movements, and the manner in which predictions are mapped into long–short portfolios under realistic risk constraints. This disconnect between statistical accuracy and economic performance constitutes the central accuracy–ranking puzzle highlighted in this study.

These findings are also consistent with the institutional features of the euro-area sovereign bond market emphasized in the theoretical and empirical literature. In a monetary union, countries are exposed to common global and regional shocks but lack independent monetary policy and exchange-rate adjustment as stabilizing mechanisms. As a result, short-run spread dynamics are

shaped less by gradual changes in fundamentals and more by heterogeneous responses to shared shocks, liquidity conditions, and shifts in risk appetite. In this environment, forecasting models may capture meaningful nonlinearities and reduce average forecast errors without necessarily improving the relative ordering of outcomes that drives long–short investment strategies. This institutional setting therefore provides a natural context in which the disconnect between statistical accuracy and economic value becomes particularly transparent.

The portfolio analysis further demonstrates that both the ECM and the ML–ECM contain economically meaningful information, but that realized performance is highly sensitive to implementation choices. Hedging duration, convexity, and basis-point value exposures materially alters risk–return trade-offs, and overly mechanical hedging can remove economically relevant sources of return. These results underscore a practical implication for fixed-income investors: predictive models cannot be evaluated in isolation from portfolio construction, and forecasting accuracy alone is insufficient for assessing investment value. Understanding which risk variables drive country selection, and how those exposures are managed, is essential for designing robust sovereign bond strategies.

Several limitations of the present framework point to natural extensions for future research. First, implementing short positions in sovereign bonds is more complex than in equity markets and typically requires the use of repo transactions, credit default swaps, or futures contracts. Each of these instruments involves liquidity constraints, margin requirements, and transaction costs that are not explicitly modeled here and may affect real-world performance. Second, the machine-learning design is intentionally conservative and largely regression-based, reflecting the limited cross-sectional dimension of the dataset and the dominance of global and regional risk variables. Richer panel datasets with greater heterogeneity across issuers could support unified cross-country machine-learning architectures and reduce reliance on duplicated country-level signals. Finally, future research could frame spread prediction explicitly as a classification problem or employ long-horizon machine-learning architectures better suited to capturing persistence and regime shifts.

Overall, the findings of this thesis highlight the importance of aligning forecasting objectives, evaluation metrics, and portfolio implementation. In long–short investment strategies, economic value depends primarily on the ability to correctly rank relative winners and losers rather than on minimizing point-forecast errors. Forecasting accuracy is therefore neither a sufficient nor a reliable indicator of economic performance, even when statistically significant improvements are achieved. This distinction reflects a general principle of active portfolio management: predictive models should be evaluated according to the decision problem they are designed to solve. In the context of euro-area sovereign bonds, where common shocks dominate short-run dynamics and adjustment mechanisms are constrained, this principle becomes particularly clear. More broadly, the results underscore that predictive models must ultimately be judged by their contribution to relative asset selection and risk management, rather than by statistical accuracy alone.

A. Technical Implementation Details.

A.1. Hyperparameter Tuning and Cross-Validation Design

This appendix documents the hyperparameter grid, the definition of each parameter, the randomized search procedure, and the expanding-window cross-validation scheme used for tuning the XGBoost model.

Hyperparameter Definitions

- **n_estimators**: number of boosted trees.
- **max_depth**: maximum depth of each tree.
- **learning_rate**: shrinkage factor applied to each tree.
- **subsample**: fraction of observations sampled per tree.
- **colsample_bytree**: fraction of predictors sampled per tree.
- **min_child_weight**: minimum sum of instance weights required to form a leaf.
- **gamma**: minimum loss reduction required to make a split.

Search Grid Specification

The model is tuned over the following hyperparameter ranges:

$$\begin{aligned} \text{n_estimators} &\in \{200, 400, 800\}, \\ \text{max_depth} &\in \{2, 3, 4, 5\}, \\ \text{learning_rate} &\in \{0.01, 0.05, 0.1, 0.2\}, \\ \text{subsample} &\in \{0.6, 0.8, 1.0\}, \\ \text{colsample_bytree} &\in \{0.6, 0.8, 1.0\}, \\ \text{min_child_weight} &\in \{1, 3, 5\}, \\ \text{gamma} &\in \{0.0, 0.1, 0.5\}. \end{aligned}$$

RandomizedSearchCV Implementation

XGBoost hyperparameters are tuned country-by-country using RandomizedSearchCV with `n_iter_search = 100` and `random_state = 42`:

```
xgb_base = XGBRegressor(  
    objective='reg:squarederror',
```

```

        tree_method='hist',
        random_state=random_state,
        n_jobs=-1
    )
search = RandomizedSearchCV(
    estimator=xgb_base,
    param_distributions=param_dist,
    n_iter=n_iter_search,
    cv=tscv,
    scoring='neg_mean_squared_error',
    n_jobs=-1,
    verbose=0,
    random_state=random_state
)

```

The hyperparameter grid spans $3,888 = 3 \times 4 \times 4 \times 3 \times 3 \times 3 \times 3$ possible combinations per country. An exhaustive search would require $3,888 \times 10 \times 119 = 4,626,720$ model evaluations. `RandomizedSearchCV` samples 100 configurations to ensure computational feasibility.

Expanding-Window Validation Scheme

Hyperparameters are selected using an expanding-window cross-validation scheme. At each forecast origin t , the available data $[t_0, t - 1]$ is split into a training block (first two-thirds) and a validation block (final one-third). Each hyperparameter configuration θ is estimated on the training block and evaluated on the validation block using mean-squared error:

$$\text{MSE}(\theta) = \frac{1}{|V_t|} \sum_{\tau \in V_t} (y_\tau - \hat{y}_{\tau|\theta})^2.$$

The configuration achieving the lowest validation error is selected:

$$\theta_t^* = \arg \min_{\theta} \text{MSE}(\theta).$$

After selecting θ_t^* , the model is retrained on the full window $[t_0, t - 1]$ using these hyperparameters to produce the forecast for period t . The window then expands by one period and the procedure is repeated.

B. Model Diagnostics and Robustness.

Unit-Root Tests for Long-Run Variables

Table B.1 reports augmented Dickey–Fuller (ADF) unit-root test statistics for the long-run variables included in the error-correction framework. For most series, the null hypothesis of a

unit root cannot be rejected, consistent with I(1) behavior.

Table B.1: Unit-Root Tests for Long-Run Fundamental Variables.

	IVP	iTraxx	iBoxx Corp	BTP	Euribor 3M	Eurozone Inflation	Germany Yield	Country Inflation	Country Debt/GDP	Country Reserves Ratio
AT	-2.198	-2.577*	-1.586	-1.557	-2.261	-2.034	-1.812	-0.173	-1.933	0.129
BE	-2.198	-2.577*	-1.586	-1.557	-2.261	-2.034	-1.812	1.308	-2.496	1.174
ES	-2.198	-2.577*	-1.586	-1.557	-2.261	-2.034	-1.812	-0.313	-2.238	-1.640
FI	-2.198	-2.577*	-1.586	-1.557	-2.261	-2.034	-1.812	0.441	-1.872	-1.536
FR	-2.198	-2.577*	-1.586	-1.557	-2.261	-2.034	-1.812	-0.822	-1.249	-0.823
GR	-2.198	-2.577*	-1.586	-1.557	-2.261	-2.034	-1.812	-0.390	-2.535	-1.342
IE	-2.198	-2.577*	-1.586	-1.557	-2.261	-2.034	-1.812	-1.758	-1.604	-1.046
IT	-2.198	-2.577*	-1.586	-1.557	-2.261	-2.034	-1.812	-0.423	-1.879	-1.515
NL	-2.198	-2.577*	-1.586	-1.557	-2.261	-2.034	-1.812	0.468	-0.230	-0.926
PT	-2.198	-2.577*	-1.586	-1.557	-2.261	-2.034	-1.812	0.048	-1.676	-1.293

Notes: Entries report Augmented Dickey–Fuller (ADF) test statistics for the null of a unit root in each series. More negative values indicate stronger evidence against the unit-root null. Significance stars are based on p -values from `adfuller` in `statsmodels`, which implement MacKinnon (1994, 2010) one-sided p -values for the null of a unit root against the stationary alternative: *** $p < 0.01$, ** $p < 0.05$, * $p < 0.10$.

B.1. Feature Contributions by Country

Figures B.1 and B.2 report average absolute feature contributions by country for the ECM and ML–ECM specifications, respectively. The figures summarize the relative importance of predictors across sovereign issuers.

Feature Contributions by Country (ECM)

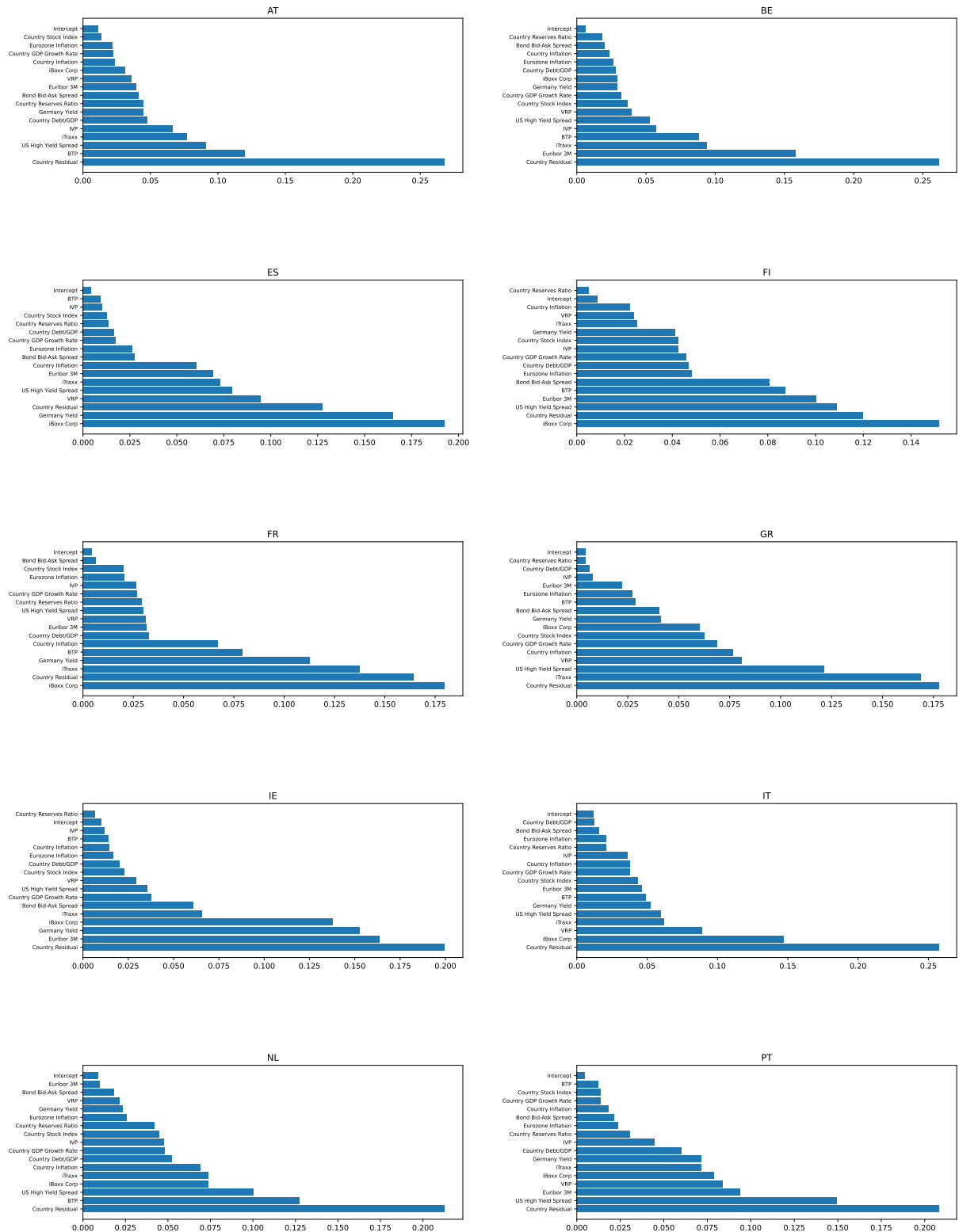


Figure B.1: Feature Contributions by Country (ECM).

Feature Contributions by Country (ML-ECM)

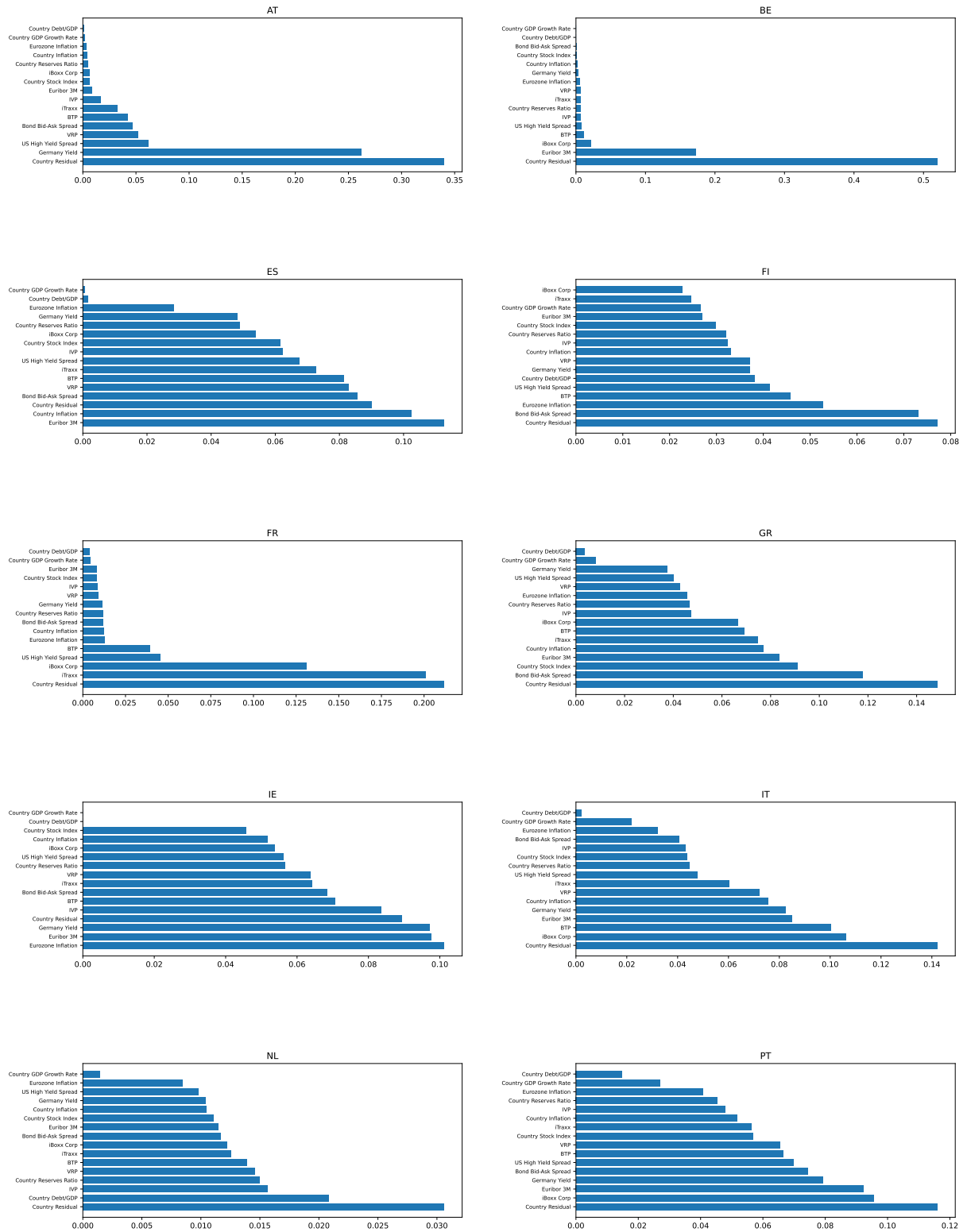


Figure B.2: Feature Contributions by Country (ML-ECM).

B.2. Cross-Validation Design Comparison

Table B.2 reports portfolio performance statistics for the ML–ECM model under four alternative cross-validation designs (CV1–CV4). The table summarizes results across the full sample and selected subsamples.

Table B.2: Portfolio Performance Across CV Specifications (ML-ECM).

Panel A: Period 2010–2022

Model	Mean (%)	Std (%)	Sharpe Ratio	Skewness	Kurtosis
CV4	5.827	10.270	0.567***	-0.164	3.554
CV3	4.368	9.547	0.458***	0.915	2.748
CV2	5.134	11.106	0.462***	0.360	4.154
CV1	5.515	9.526	0.579***	0.613	2.166

Panel B: Period 2022–2024

Model	Mean (%)	Std (%)	Sharpe Ratio	Skewness	Kurtosis
CV4	6.583	9.390	0.701***	0.490	1.379
CV3	3.450	9.496	0.363**	0.998	3.552
CV2	4.758	10.353	0.460***	0.862	2.235
CV1	6.367	9.710	0.656***	1.428	2.333

Panel C: Period 2010–2024

Model	Mean (%)	Std (%)	Sharpe Ratio	Skewness	Kurtosis
CV4	6.035*	10.025	0.602***	-0.015	2.964
CV3	4.071	9.533	0.427***	0.925	2.747
CV2	5.002	10.894	0.459***	0.477	3.536
CV1	5.570*	9.560	0.583***	0.857	2.105

Notes: Mean and standard deviation are annualised and expressed in percentage terms. Sharpe ratios are tested against the null of no predictability with significance levels denoted by * $p < 0.10$, ** $p < 0.05$, and *** $p < 0.01$.

C. Kitchen Sink Specifications

This appendix reports results from “Kitchen Sink” specifications that exclude the error-correction term. These unrestricted models rely exclusively on first-differenced predictors and are estimated as auxiliary specifications to assess the role of long-run restrictions in the forecasting and portfolio construction framework. Two variants are considered: a linear OLS specification and a machine-learning specification based on XGBoost.

C.1. Kitchen Sink OLS Model

The Kitchen Sink OLS specification is estimated using only first-differenced predictors:

$$\Delta \text{Spread}_{t+1}^{(i)} = \gamma' \Delta X_t^{(i)} + \lambda' \Delta F_t^{(i)} + \varepsilon_{t+1}^{(i)}, \quad (\text{C.1})$$

where $\Delta X_t^{(i)}$ denotes changes in country-specific fundamentals and $\Delta F_t^{(i)}$ denotes changes in global and regional financial variables.

Table C.1 reports out-of-sample R^2 statistics for the Kitchen Sink OLS specification.

Table C.1: Out-of-Sample R^2 (ECM vs Kitchen Sink).

	AT	BE	ES	FI	FR	GR	IE	IT	NL	PT	Overall
ECM R^2	-0.805	-2.906	-0.845	-0.298	-1.414	-3.935	-16.396	-0.843	-0.723	-5.899	-3.775
KS R^2	-0.735	-2.266	-0.659	-0.298	-0.713	-2.437	-7.399	-0.667	-0.715	-3.710	-2.327
KS vs ECM R^2	0.038	0.164	0.101	0.000	0.291	0.303	0.517	0.095	0.005	0.317	0.303

Notes: Country-by-country out-of-sample R^2 statistics for the ECM benchmark (OLS) and the Kitchen Sink (OLS). Forecasts are generated from January 2015 to November 2024 using an expanding-window scheme. The out-of-sample statistic R^2 is computed as $1 - \text{MSE}_{\text{model}} / \text{Var}(\Delta \text{Spread})$, and the relative measure KS vs ECM R^2 is computed as $1 - \text{MSE}_{\text{KS}} / \text{MSE}_{\text{ECM}}$. The Overall values report the pooled performance using all countries and dates together.

Table C.2 reports tercile classification accuracy by country.

Table C.2: Hit Ratios by Country (ECM vs Kitchen Sink).

	AT	BE	ES	FI	FR	GR	IE	IT	NL	PT	Overall
ECM Hit Ratio	0.372	0.305	0.440	0.282	0.305	0.415	0.473	0.434	0.375	0.390	0.385
KS Hit Ratio	0.308	0.254	0.429	0.366	0.305	0.453	0.392	0.462	0.338	0.410	0.381

Notes: The hit ratio for each country is the number of correctly matched active terciles (actual = -1 or +1) divided by the number of periods in which the actual tercile is active (actual $\neq$ 0). The Overall value reports the pooled performance of the model using all countries and dates together.

Table C.3 reports long–short portfolio performance constructed from Kitchen Sink OLS forecasts under alternative interest-rate hedging configurations. Reported statistics include mean returns, volatility, Sharpe ratios, Sortino ratios, skewness, and kurtosis.

Table C.3: Portfolio Performance (Kitchen Sink).

Portfolio	Mean (%)	Std (%)	Sharpe Ratio	Downside Std (%)	Sortino Ratio	Skewness	Kurtosis
Unhedged	7.006**	10.615	0.660***	5.522	1.269	0.777	5.492
Duration Hedged	3.177*	5.314	0.598***	2.763	1.150	1.028	7.020
Convexity Hedged	2.653	5.089	0.521***	2.788	0.952	0.763	6.477
BPV Hedged	2.864*	4.559	0.628***	2.117	1.353	1.618	6.122
Duration–Convexity Hedged	3.423	6.774	0.505***	2.853	1.200	3.267	18.570
Duration–BPV Hedged	3.308**	5.239	0.631***	2.636	1.255	1.184	6.056
Convexity–BPV Hedged	2.958*	5.415	0.546***	2.699	1.096	1.630	9.130
Duration–Convexity–BPV Hedged	2.784*	4.973	0.560***	3.096	0.899	-0.236	2.967

Notes: Mean and standard deviation (Std) are annualised and reported in percentage terms. Downside Std is computed using only negative returns. The Sortino ratio is computed as Mean divided by Downside Std. Sharpe ratios are tested against the null hypothesis that the Sharpe ratio equals zero; significance levels are denoted by * $p < 0.10$, ** $p < 0.05$, and *** $p < 0.01$.

C.2. Kitchen Sink Machine Learning Model

The machine-learning Kitchen Sink specification applies XGBoost to first-differenced predictors:

$$\widehat{\Delta\text{Spread}}_{t+1}^{(i)} = \sum_{k=1}^K f_k(\Delta X_t^{(i)}, \Delta F_t^{(i)}), \quad f_k \in \mathcal{F}, \quad (\text{C.2})$$

where f_k are regression trees capturing nonlinear interactions among predictors.

Table C.4 reports out-of-sample R^2 statistics for the machine-learning Kitchen Sink specification.

Table C.4: Out-of-sample R^2 Statistics (ML-ECM vs ML-Kitchen Sink).

	AT	BE	ES	FI	FR	GR	IE	IT	NL	PT	Overall
ML–ECM R^2	-0.197	-0.426	-0.189	-0.181	-0.149	-0.948	-3.877	-0.477	-0.075	-1.732	-0.933
ML–KS R^2	-0.135	-0.228	-0.269	-0.143	-0.089	-0.212	-3.381	-0.200	-0.163	-0.661	-0.254
ML–KS vs ML–ECM R^2	0.052	0.139	-0.067	0.032	0.052	0.378	0.102	0.187	-0.081	0.392	0.351

Notes: Country-by-country out-of-sample R^2 statistics for the ML–ECM (XGBoost) and ML–KS forecasting models, computed under an expanding-window scheme from January 2015 to November 2024. The out-of-sample statistic R^2 is defined as $1 - \text{MSE}_{\text{model}} / \text{Var}(\Delta\text{Spread})$, and the relative measure ML–KS vs ML–ECM R^2 is computed as $1 - \text{MSE}_{\text{ML-KS}} / \text{MSE}_{\text{ML-ECM}}$. The Overall values report the pooled performance using all countries and dates together.

Table C.5 reports tercile classification accuracy by country.

Table C.5: Hit Ratios by Country (ML-ECM vs ML-Kitchen Sink).

	AT	BE	ES	FI	FR	GR	IE	IT	NL	PT	Overall
ML–ECM Hit Ratio	0.372	0.322	0.333	0.324	0.305	0.453	0.419	0.406	0.312	0.400	0.370
ML–KS Hit Ratio	0.321	0.169	0.333	0.479	0.268	0.585	0.405	0.406	0.400	0.470	0.396

Notes: The hit ratio for each country is the number of correctly matched active terciles (actual = -1 or $+1$) divided by the number of periods in which the actual tercile is active (actual $\neq 0$). The Overall value reports the pooled performance of the model using all countries and dates together.

Table C.6 reports long–short portfolio performance based on machine-learning Kitchen Sink forecasts under alternative interest-rate hedging configurations. Reported statistics include mean

returns, volatility, Sharpe ratios, Sortino ratios, skewness, and kurtosis.

Table C.6: Portfolio Performance (ML–Kitchen Sink).

Portfolio	Mean (%)	Std (%)	Sharpe Ratio	Downside Std (%)	Sortino Ratio	Skewness	Kurtosis
Unhedged	4.053	10.731	0.378***	6.257	0.648	0.561	1.902
Duration Hedged	1.232	4.705	0.262***	2.802	0.440	0.752	2.767
Convexity Hedged	1.838	4.676	0.393***	2.563	0.717	1.014	2.944
BPV Hedged	1.029	5.542	0.186**	3.528	0.292	0.561	8.139
Duration–Convexity Hedged	0.999	5.305	0.188**	3.507	0.285	0.067	4.009
Duration–BPV Hedged	2.421	5.236	0.462***	2.572	0.941	1.493	4.522
Convexity–BPV Hedged	2.713	5.250	0.517***	2.569	1.056	1.417	4.276
Duration–Convexity–BPV Hedged	2.383	5.552	0.429***	2.949	0.808	1.276	4.466

Notes: Mean and standard deviation (Std) are annualised and reported in percentage terms. Downside Std is computed using only negative returns. The Sortino ratio is computed as Mean divided by Downside Std. Sharpe ratios are tested against the null hypothesis that the Sharpe ratio equals zero; significance levels are denoted by * $p < 0.10$, ** $p < 0.05$, and *** $p < 0.01$.

D. Panel Specification with Country Fixed Effects

This appendix reports results from a panel regression with country fixed effects, estimated as an auxiliary specification to account explicitly for cross-sectional dependence and common shocks across euro-area sovereign issuers.

The panel regression is specified as:

$$\Delta \text{Spread}_{i,t} = \alpha_i + \lambda' \Delta X_{i,t} + \theta' \Delta F_t + \gamma' (\Delta F_t \times \text{Periphery}_i) + u_{i,t}, \quad (\text{D.1})$$

where α_i denotes country fixed effects, $\Delta X_{i,t}$ are country-specific fundamentals, ΔF_t are global and regional variables, and Periphery_i is a dummy variable equal to one for Greece, Italy, Ireland, Portugal, and Spain. The interaction term allows for heterogeneous exposure to common variables across core and periphery countries.

Panel unit-root tests based on the cross-sectionally augmented Dickey–Fuller (CADF) procedure indicate that spreads are stationary once common components are accounted for. Accordingly, the specification focuses on short-run monthly spread variation.

Table D.1 reports the tercile agreement (hit ratios) of the panel model across countries.

Table D.1: Hit Ratios by Country (Panel Model).

	AT	BE	ES	FI	FR	GR	IE	IT	NL	PT	Overall
Panel Model	0.397	0.424	0.464	0.423	0.305	0.377	0.257	0.387	0.388	0.350	0.376

Notes: The hit ratio for each country is the number of correctly matched active terciles (actual = -1 or $+1$) divided by the number of periods in which the actual tercile is active (actual $\neq 0$). The Overall value reports the pooled performance of the model using all countries and dates together.

Table D.2 reports the performance of long–short portfolios constructed from panel-based forecasts under alternative interest-rate hedging configurations. Reported statistics include mean returns, volatility, Sharpe ratios, skewness, and kurtosis.

Table D.2: Portfolio Performance (Panel Model).

Portfolio	Mean (%)	Std (%)	Sharpe Ratio	Downside Std (%)	Sortino Ratio	Skewness	Kurtosis
Unhedged	3.266	8.536	0.383***	4.938	0.661	0.952	5.563
Duration–Hedged	1.099	4.900	0.224**	3.204	0.343	0.146	10.607
Convexity–Hedged	1.237	4.810	0.257***	3.022	0.409	0.550	9.170
BPV–Hedged	0.755	4.791	0.158*	2.662	0.283	2.511	16.667
Duration–Convexity–Hedged	2.210	4.574	0.483***	2.335	0.947	1.672	7.372
Duration–BPV–Hedged	1.032	5.439	0.190**	3.229	0.319	1.469	12.751
Convexity–BPV–Hedged	0.923	5.419	0.170*	3.161	0.292	1.686	12.135
Duration–Convexity–BPV–Hedged	1.702	4.224	0.403***	2.395	0.710	1.072	5.651

Notes: Mean and standard deviation (Std) are annualised and reported in percentage terms. Downside Std is computed using only negative returns. The Sortino ratio is computed as Mean divided by Downside Std. Sharpe ratios are tested against the null hypothesis that the Sharpe ratio equals zero; significance levels are denoted by * $p < 0.10$, ** $p < 0.05$, and *** $p < 0.01$.

E. Ordered Logit Classification Model

This appendix reports results from an ordered logit model that directly predicts tercile labels ($y_{t+1}^{(i)} \in \{-1, 0, 1\}$) rather than continuous spread changes. The model is estimated as an auxiliary specification that targets cross-sectional ranking outcomes instead of point forecasts.

The ordered logit model is based on a latent index:

$$y_{t+1}^{*(i)} = \alpha^{(i)} + \gamma' \Delta \mathbf{X}_t^{(i)} + \lambda' \Delta \mathbf{F}_t^{(i)} + \varepsilon_{t+1}^{(i)}, \quad (\text{E.1})$$

where $\varepsilon_{t+1}^{(i)}$ follows a logistic distribution. The observed signal is determined by threshold parameters $\kappa_1 < \kappa_2$:

$$y_{t+1}^{(i)} = \begin{cases} -1, & \text{if } y_{t+1}^{*(i)} \leq \kappa_1, \\ 0, & \text{if } \kappa_1 < y_{t+1}^{*(i)} \leq \kappa_2, \\ 1, & \text{if } y_{t+1}^{*(i)} > \kappa_2. \end{cases} \quad (\text{E.2})$$

Table E.1 reports the performance of long–short portfolios constructed from ordered logit signals under alternative interest-rate hedging configurations. Reported statistics include mean returns, volatility, Sharpe ratios, skewness, and kurtosis.

Table E.1: Portfolio Performance (Ordered Logit).

Portfolio	Mean (%)	Std (%)	Sharpe Ratio	Downside Std (%)	Sortino Ratio	Skewness	Kurtosis
Unhedged	3.883	11.728	0.331***	7.859	0.494	-0.245	4.476
Duration–Hedged	2.666	5.476	0.487***	3.009	0.886	0.980	4.353
Convexity–Hedged	1.935	5.216	0.371***	3.046	0.635	0.887	4.699
BPV–Hedged	3.381*	5.451	0.620***	2.900	1.166	0.923	4.483
Duration–Convexity–Hedged	2.436	5.818	0.419***	2.990	0.815	1.646	6.891
Duration–BPV–Hedged	3.430**	5.432	0.631***	2.771	1.238	1.109	4.695
Convexity–BPV–Hedged	3.145*	5.326	0.591***	2.787	1.129	1.057	4.761
Duration–Convexity–BPV–Hedged	1.986	5.654	0.351***	3.247	0.612	1.013	4.505

Notes: Mean and standard deviation (Std) are annualised and reported in percentage terms. Downside Std is computed using only negative returns. The Sortino ratio is computed as Mean divided by Downside Std. Sharpe ratios are tested against the null hypothesis that the Sharpe ratio equals zero; significance levels are denoted by * $p < 0.10$, ** $p < 0.05$, and *** $p < 0.01$.

References

- Afonso, A., Arghyrou, M., and Kontonikas, A. (2012). The determinants of sovereign bond yield spreads in the emu.
- Afonso, A., Jalles, J. T., and Kazemi, M. (2019). The effects of macroeconomic, fiscal and monetary policy announcements on sovereign bond spreads: an event study from the emu. Technical report, EconPol Working Paper.
- Arghyrou, M. G. and Kontonikas, A. (2012). The emu sovereign-debt crisis: Fundamentals, expectations and contagion. *Journal of International Financial Markets, Institutions and Money*, 22(4):658–677.
- Belly, G., Boeckelmann, L., Caicedo Graciano, C. M., Di Iorio, A., Istrefi, K., Siakoulis, V., and Stalla-Bourdillon, A. (2023). Forecasting sovereign risk in the euro area via machine learning. *Journal of Forecasting*, 42(3):657–684.
- Campbell, J. Y. and Thompson, S. B. (2008). Predicting excess stock returns out of sample: Can anything beat the historical average? *The Review of Financial Studies*, 21(4):1509–1531.
- Chen, T. (2016). Xgboost: A scalable tree boosting system. *Cornell University*.
- De Grauwe, P. and Ji, Y. (2013). Self-fulfilling crises in the eurozone: An empirical test. *Journal of International Money and finance*, 34:15–36.
- Diamond, S. and Boyd, S. (2016). Cvxpy: A python-embedded modeling language for convex optimization. *Journal of Machine Learning Research*, 17(83):1–5.
- Engle, R. F. and Granger, C. W. (1987). Co-integration and error correction: representation, estimation, and testing. *Econometrica: journal of the Econometric Society*, pages 251–276.
- Garman, M. B. and Klass, M. J. (1980). On the estimation of security price volatilities from historical data. *Journal of business*, pages 67–78.
- Gomez-Gonzalez, J. E., Uribe, J. M., and Valencia, O. (2024). Sovereign risk and economic complexity.
- Gu, S., Kelly, B., and Xiu, D. (2020). Empirical asset pricing via machine learning. *The Review of Financial Studies*, 33(5):2223–2273.
- Izadi, S. and Hassan, M. K. (2018). Impact of international and local conditions on sovereign bond spreads: International evidence. *Borsa Istanbul Review*, 18(1):41–51.
- Johansen, S. (1991). Estimation and hypothesis testing of cointegration vectors in gaussian vector autoregressive models. *Econometrica: journal of the Econometric Society*, pages 1551–1580.
- Longstaff, F. A., Pan, J., Pedersen, L. H., and Singleton, K. J. (2011). How sovereign is sovereign credit risk? *American Economic Journal: Macroeconomics*, 3(2):75–103.

- Mundell, R. A. (1961). A theory of optimum currency areas. *The American economic review*, 51(4):657–665.
- Stiglitz, J. E. (2016). *The euro: How a common currency threatens the future of Europe*. WW Norton & Company.
- Welch, I. and Goyal, A. (2008). A comprehensive look at the empirical performance of equity premium prediction. *The Review of Financial Studies*, 21(4):1455–1508.