REM WORKING PAPER SERIES

Sovereign Ratings and Risk Pricing, Agency Divergences in the European Union

António Afonso, José Alves, Periklis Gogas, Theophilos Papadimitriou

REM Working Paper 0424-2026

July 2026

REM – Research in Economics and Mathematics
Rua Miguel Lúpi 20,
1249-078 Lisboa,
Portugal

ISSN 2184-108X

Fundação para a Ciência e a Tecnologia
U LISBOA | UNIVERSIDADE DE LISBOA
ISEG Lisbon School of Economics & Management Universidade de Lisboa
REM RESEARCH IN ECONOMICS AND MATHEMATICS

REM – Research in Economics and Mathematics

Rua Miguel Lupi, 20
1249-078 LISBOA
Portugal

Telephone: +351 - 213 925 912
E-mail: rem@iseg.ulisboa.pt

https://rem.rc.iseg.ulisboa.pt/

https://twitter.com/ResearchRem

https://www.linkedin.com/company/researchrem/

https://www.facebook.com/researchrem/

Sovereign Ratings and Risk Pricing, Agency Divergences in the European Union[1]

António Afonso[2] José Alves[3] Periklis Gogas[4] Theophilos Papadimitriou[5]

July 2026

Abstract

Using annual EU-27 data for 1995-2024, we examine whether sovereign ratings mainly reflect common macro-fiscal fundamentals or whether agency-specific departures from that benchmark are also priced into sovereign funding conditions. Pooled ordered probits and a machine-learning diagnostic layer for nonlinearities and thresholds for Fitch, Moody's, and S&P identify a stable set of core rating determinants centred on inflation, debt-to-GDP, current account balance, budget balance rule indicator, output gap, old-age dependency, and revenue capacity, while within-country variation is narrower and concentrated mainly in inflation, debt, and unemployment. The machine-learning analysis confirms that flexible models absorb nearly all systematic variation between fundamentals and ratings, validating the shadow-rating decomposition used in the market-pricing test. ECB-based bond-yield regressions show that both the fundamentals-implied shadow rating and the agency-specific deviation are priced in euro-area Bund spreads: a one-notch more favourable value of either component is associated with about 50 basis points lower spreads. Evidence indicates that this pricing effect strengthens as debt rises and intensifies further once debt exceeds 100% of GDP, while a shorter crisis-period interaction is directionally similar but less precise. Sovereign ratings therefore appear to combine a common fundamentals core with discretionary overlays that markets treat as economically relevant signals.

Keywords: sovereign ratings; sovereign risk pricing; sovereign bond spreads; rating agencies; ordered probit; panel data; machine-learning; macro-financial transmission

JEL: C23; C25; E44; F34; G15; H63

---

[1] This work was supported by national funds through FCT – Fundação para a Ciência e a Tecnologia, I.P., in the framework of the unit UID/06522/2025 – https://doi.org/10.54499/UID/06522/2025. The opinions expressed herein are those of the authors and do not necessarily reflect those of the authors' employers. Any remaining errors are the authors' sole responsibility.

[2] ISEG Research, ISEG - Lisbon School of Economics and Management, Universidade de Lisboa, Lisbon, Portugal; UECE - Research Unit on Complexity and Economics; CESifo (Center for Economic Studies and Ifo Institute). E-mail: aafonso@iseg.ulisboa.pt

[3]ISEG Research, ISEG - Lisbon School of Economics and Management, Universidade de Lisboa, Lisbon, Portugal; UECE - Research Unit on Complexity and Economics; CESifo (Center for Economic Studies and Ifo Institute). E-mail: jalves@iseg.ulisboa.pt

[4] Department of Economics, Democritus University of Thrace. email: pgkogkas@econ.duth.gr

[5] Department of Economics, Democritus University of Thrace. email: papadimi@econ.duth.gr

1. Introduction

Sovereign credit ratings sit at the heart of the macro-financial transmission mechanism. They matter not only because they summarize default risk, but because they influence sovereign funding conditions, collateral frameworks, portfolio mandates, and the risk management of financial intermediaries. In that sense, ratings are not merely descriptive labels; they are state variables that shape how macroeconomic and fiscal information is translated into borrowing conditions. The literature has long shown that ratings are tightly linked to sovereign yields and spreads, while still retaining explanatory power beyond observable fundamentals (Cantor and Packer, 1996). That combination keeps a central question open: which macro-fiscal fundamentals anchor sovereign ratings, and where does agency-specific judgment begin?

This question is especially important in the European Union. Cross-border holdings, common financial shocks, and the sovereign-bank nexus make rating actions economically consequential beyond the rated sovereign itself. Downgrades can propagate through bank balance sheets, portfolio rebalancing, and cross-border funding conditions, while disagreement across agencies can alter the information set facing investors and regulators. In this environment, understanding ratings is not simply an exercise in classification; it is part of understanding sovereign-risk transmission inside an integrated macro-financial area.

Despite an extensive determinants related literature, two limitations remain. The first is methodological but not merely technical: most studies impose a relatively rigid functional form (often linear in covariates) and then interpret coefficients as "determinants," even though ratings are ordinal, threshold-based outcomes and agencies explicitly employ conditional logic ("debt matters more when growth is weak," "external balances matter more when financing conditions tighten"). Linear-additive models may be adequate approximations in tranquil regimes but become fragile exactly where ratings matter most - around regime breaks and threshold crossings. Consistent with this concern, evidence suggests that rating standards and determinants differ before and after the European debt crisis (Reusens and Croux, 2017) and that ratings and spreads co-move more tightly in stress states (Aizenman et al., 2013).

The second limitation is conceptual: many determinants' papers treat deviations from fundamentals as residual noise, rather than as an economically meaningful object that can reveal systematic agency behaviour, judgment, and potential bias. Yet research documents nontrivial subjectivity in sovereign ratings (De Moor et al., 2018), patterns consistent with home bias (Fuchs and Gehring, 2017), and incentives that can distort rating quality in equilibrium (Bolton et al., 2012; Mathis et al., 2009), with competition sometimes associated with lower quality (Becker and Milbourn, 2011) and with cyclicality in quality (Bar-Isaac and Shapiro, 2013). If

these mechanisms operate, then the central empirical question is not only which fundamentals correlate with ratings, but also when, how, and for whom ratings depart from fundamentals-implied benchmarks.

We study whether EU sovereign ratings mainly summarize a common macro-fiscal core or whether agency-specific deviations from that core also matter for sovereign risk pricing. The analysis covers long-term issuer ratings for the 27 EU member states over 1995-2024, harmonized onto a common ordinal and quantitative scale. The empirical backbone is a panel ordered probit in the tradition of Afonso (2003), Bissoondoyal-Bheenick (2005), and Afonso et al. (2011), which respects the ordinal nature of ratings and provides a transparent fundamentals-based benchmark from which shadow ratings and agency deviations can be constructed.

To complement that benchmark, we add three elements: a fixed-effects panel for the cross-agency average, ECB-based yield regressions that decompose ratings into shadow values and agency deviations, and a deliberately limited machine-learning diagnostic for nonlinearities and thresholds.

The nonlinear layer is not intended to replace the econometric benchmark. Its role is diagnostic: random forests, support vector machines, neural networks, and XGBoost (Chen & Guestrin, 2016) are used to test whether the ordered-probit index misses threshold effects or interactions that are difficult to pre-specify. Because sovereign ratings affect funding conditions, regulation, and policy narratives, the flexible models are paired with SHAP-based attributions, which apply Shapley (1953) values to model predictions (Lundberg & Lee, 2017) and partial-dependence diagnostics so that prediction remains auditable rather than merely black-box, in line with calls for interpretability in high-stakes settings (Rudin, 2019; Molnar, 2022).

Throughout, flexible models are evaluated under a time-consistent validation design, performance is assessed with ordinally meaningful criteria rather than exact-match accuracy alone, and fundamentals-implied shadow ratings are constructed from both ordered-probit and machine-learning probabilities so that observed ratings can be decomposed into a common fundamentals component and an agency deviation.

The credibility of the market-pricing test depends directly on the quality of the shadow rating used to construct the agency deviation. A linear model leaves systematic structure in the residual, meaning the deviation would confound genuine agency judgment with functional-form misspecification. The ML checks establish that flexible models recover essentially all systematic variation between fundamentals and ratings, confirming that the ordered-probit

residual deviation reflects genuine agency discretion rather than functional-form misspecification.

The central empirical result is that euro-area sovereign spreads price both the fundamentals-implied component of ratings and the agency-specific deviation around it. In the baseline decomposition, a one-notch more favourable shadow rating and a one-notch more favourable agency deviation are each associated with roughly 50 basis points lower Bund spreads. Appendix evidence shows that this pricing effect becomes stronger as debt rises and intensifies further once debt exceeds 100% of GDP. This pricing result rests on a stable ratings core: across the three agencies, inflation, debt-to-GDP, current account balance, the budget balance rule indicator, output gap, old-age dependency, and revenue capacity are the most consistent correlates, while within-country evidence is narrower and concentrated in fewer variables.

The remainder of the article is organized as follows. Section 2 reviews the literature on sovereign-rating determinants, ratings as macro-financial state variables, and agency behaviour. Section 3 presents the empirical framework, including the panel and market-pricing extensions. Section 4 describes the data, rating harmonization, market-yield construction, and stylized facts. Section 5 reports the main empirical results. Section 6 concludes.

2. Literature Review

The literature on sovereign ratings revolves around three connected questions: which fundamentals agencies appear to price, how rating actions affect markets and macro-financial outcomes, and how agency incentives shape the content and timing of rating decisions. Early determinants work shows that a relatively small set of macro-fiscal variables explains a large share of cross-country variation in ratings, while also showing that ratings retain explanatory power for sovereign yields beyond those observables (Cantor and Packer, 1996; Afonso, 2003; Bissoondoyal-Bheenick, 2005; Afonso et al., 2011, 2012). Related work documents that macroeconomic fundamentals also help price sovereign debt directly (Hilscher and Nosbusch, 2010), reinforcing the dual role of fundamentals as both rating inputs and pricing factors. That dual result matters for the present study because it implies that ratings are simultaneously outcomes of macro-fiscal conditions and inputs into financial conditions.

The euro area debt crisis intensified interest in whether the mapping from fundamentals to ratings is stable through time and whether agencies adjust standards in crisis states. Aizenman et al. (2013) examine the crisis period and show that rating actions and sovereign spreads are tightly linked when stress regimes prevail, which suggests that the marginal effect of fundamentals may become state-dependent. This is one reason a purely linear parametric model

can be too restrictive: what matters for ratings in tranquil times may be different from what matters during stress, and the same variable may have nonlinear effects depending on the fiscal or external backdrop. Reusens and Croux (2017) make that instability explicit by comparing determinants before and after the European debt crisis, which underscores a basic methodological warning: the "determinants" one estimates are partly a function of the regime being sampled.

A second strand studies the consequences of ratings and rating news. Here the key question is whether ratings merely reflect information already in markets or whether they move prices and quantities through information, coordination, and constraint channels. Kaminsky and Schmukler (2002) show that sovereign ratings affect country risk and stock returns in emerging markets, supporting the view that rating events can convey new information or coordinate beliefs. Gande and Parsley (2005) document spillovers in the sovereign debt market, implying that rating news is not confined to the rated country but propagates cross-sectionally, which is consistent with portfolio rebalancing, benchmark effects, or common-information updating. The coordination channel is central in Boot et al. (2006), who argue that ratings can function as focal points in contracting and regulation; importantly, this implies that discrete thresholds can generate discontinuities in market responses even when fundamentals move smoothly. That discontinuity logic matters directly for a determinants paper because it rationalizes why agencies might place special weight on variables that are predictive of crossing salient rating boundaries (for example, the cusp between investment grade and speculative grade), rather than treating all rating-notch movements as symmetric. The empirical implication is that determinants may be heterogeneous across the distribution of ratings, a feature that is awkward for standard parametric specifications but natural for tree-based methods and other nonlinear learners.

Another strand of the literature focuses on agency behaviour: whether ratings embed distortions due to conflicts of interest, competitive dynamics, or regulation. This literature is relevant not only because it interprets deviations from fundamentals, but also because it clarifies why two agencies might weight the same fundamental differently even in the same period. Evidence of herding among sovereign rating agencies (Chen et al., 2019) further suggests that agency-specific deviations need not be independent of one another. Bolton et al. (2012) formalize how issuer shopping and the strategic environment can inflate ratings, while Skreta and Veldkamp (2009) show that complexity can interact with ratings shopping to generate systematic inflation, an insight that generalizes beyond structured finance to any setting where the mapping from fundamentals to risk is sufficiently opaque that multiple plausible

assessments can be produced. Mathis et al. (2009) highlight limits to reputational discipline, which is crucial because it weakens the common assumption that "reputation fixes everything." If reputational incentives do not fully discipline agencies, then persistent deviations from fundamentals are not anomalies; they are equilibrium outcomes.

The same perspective also helps explain why competition and regulation can reshape rating content. Becker and Milbourn (2011) provide evidence that increased competition can reduce rating quality, while Bar-Isaac and Shapiro (2013) show that ratings quality can vary over the business cycle, complementing earlier evidence that agencies can amplify rather than smooth crisis dynamics (Ferri et al., 1999). Opp et al. (2013) emphasize that regulation tied to ratings changes incentives for agencies and market participants, which matters for sovereigns because sovereign exposures sit at the nexus of regulation, banking, and collateral eligibility. Kisgen and Strahan (2010) show that ratings-based regulation affects a firm's cost of capital; even though their setting is corporate, the mechanism is portable: when regulations embed ratings, ratings become state variables with real financing consequences. Taken together, these contributions imply that any empirical determinants exercise that treats ratings as a purely statistical function of fundamentals is missing a live part of the story.

Within Europe, the channels through which sovereign ratings affect the real economy are particularly salient because of the sovereign-bank nexus and cross-border holdings. Adelino and Ferreira (2016) show that sovereign downgrades transmit to bank ratings and constrain lending supply, which provides a concrete mechanism for how rating events can become macro-relevant shocks rather than passive reflections of deterioration. de Haan and Vermeulen (2021) document that sovereign debt ratings influence the country composition of cross-border holdings in the euro area, which is consistent with ratings-based investment constraints or internal risk management practices that mimic such constraints. Those two findings together create a feedback logic that determinants papers must confront: ratings respond to fundamentals, yet rating actions can themselves shift financing conditions in ways that feed back into fundamentals. If you ignore that endogeneity possibility, you can easily over-interpret coefficients as "determinants" when they partly capture the consequences of prior rating dynamics. In practice, this pushes the empirical strategy toward carefully timed covariates (to limit look-ahead) and toward comparing predictive mappings (which can still be useful for monitoring) with more structural interpretations (which require stronger assumptions).

A related debate is whether ratings contain systematic subjectivity or bias beyond what can be justified by observable fundamentals. De Moor et al. (2018) quantify subjectivity in sovereign ratings, which supports the view that discretionary judgment is not a residual

nuisance but a meaningful component of the rating production function. Abad et al. (2018) examine rating convergence and show that disagreement itself shapes spillovers, implying that the information environment is multi-dimensional: it is not just the level of ratings but also the dispersion across agencies that matters for markets. Fuchs and Gehring (2017) provide evidence consistent with home bias in sovereign ratings, pointing to political economy channels that would not be captured by standard macro-fiscal covariates. These contributions collectively motivate an approach that can separate what is common across agencies (plausibly fundamentals) from what is agency-specific (potentially discretion, incentives, or bias).

Methodologically, the natural benchmark for ratings is the ordered response model, because ratings are ordinal categories with unequal economic distances. Ordered probit is attractive precisely because it is transparent and makes the threshold structure explicit. Its limitation is that the researcher must pre-specify functional form and interactions, which can be restrictive if the true mapping is nonlinear, regime-dependent, or heterogeneous across rating levels.

For that reason, and because the regime-dependence documented in the crisis literature suggests interactions that are difficult to pre-specify, recent econometrics work treats machine learning as a useful complement for prediction and pattern discovery, as long as the models remain interpretable and properly validated (Varian, 2014; Kleinberg et al., 2015; Athey and Imbens, 2017, 2019; Mullainathan and Spiess, 2017). In ratings applications, flexible classifiers often outperform simpler linear benchmarks, but their value here is narrower: they help audit whether the ordered-probit benchmark is missing systematic threshold structure or interactions (Huang et al., 2004; Kim, 2005; Gu et al., 2020; Golbayani et al., 2020; Hashimoto et al., 2023). Machine learning is therefore used here as a controlled diagnostic layer rather than as a substitute for the econometric design.

Taken together, this literature suggests three things. Rating determinants are likely to be regime-dependent and partly endogenous because ratings interact with market prices, regulation, and macro-financial conditions. Agency behaviour and institutional incentives can generate persistent departures from a purely fundamentals-based benchmark. Flexible predictive models are useful only insofar as they help identify nonlinear structure without obscuring interpretation. Those points motivate an empirical strategy built around an ordered-response benchmark, a decomposition into shadow ratings and agency deviations, and a tightly controlled use of machine learning.

3. Empirical Framework

Sovereign ratings are modelled as ordered outcomes generated by an underlying, continuous assessment of creditworthiness. Let $R_{a,i,t}$ denote the long-term issuer rating assigned by agency $a \in \{S\&P, Moody's, Fitch\}$ to country i in year t, coded as an ordered category $k \in \{0,1,\dots,K\}$ where higher values indicate better credit quality (Appendix Table A1). The observed rating is assumed to be a discretized realization of a latent index $R^*_{a,i,t}$ that aggregates macroeconomic, fiscal, external, and structural information into a single creditworthiness measure, consistent with the ordered-response approach commonly applied in the sovereign ratings literature (Afonso, 2003; Bissoondoyal-Bheenick, 2005; Afonso et al., 2011). The rating assignment rule is

$$R_{a,i,t} = k, \text{if } \mu_{a,k-1} < R^*_{a,i,t} \leq \mu_{a,k}, k = 0, \dots, K, \quad (1)$$

with $\mu_{a,-1} = -\infty$ and $\mu_{a,K} = +\infty$, and where the thresholds $\{\mu_{a,k}\}$ are agency-specific thresholds. Thresholds formalize the idea that ratings are categorical and that rating changes can occur in discrete steps, an important feature given the coordination and institutional roles of ratings (Boot et al., 2006). The latent index is specified as

$$R^*_{a,i,t} = X'_{i,t}\beta_a + \alpha_{a,i} + \gamma_{a,t} + \varepsilon_{a,i,t}, \quad (2)$$

where $X_{i,t}$ is the vector of fundamentals observed for country i at time t, β_a is an agency-specific slope vector, $\alpha_{a,i}$ captures time-invariant unobserved heterogeneity, $\gamma_{a,t}$ captures common time effects (including global conditions and agency-wide shifts in standards), and $\varepsilon_{a,i,t}$ is an idiosyncratic disturbance assumed standard normal under ordered probit. Allowing β_a and $\{\mu_{a,k}\}$ to differ across agencies accommodates persistent heterogeneity in criteria and rating scales, which is consistent with evidence of discretion and subjectivity in sovereign ratings (De Moor et al., 2018), and the model is estimated by maximum likelihood.

The ordered probit structure implies that the probability of observing category k is

$$\Pr(R_{a,i,t} = k | X_{i,t}) = \Phi(\mu_{a,k} - \eta_{a,i,t} - X'_{i,t}\beta_a) - \Phi(\mu_{a,k-1} - \eta_{a,i,t} - X'_{i,t}\beta_a), \quad (3)$$

where $\Phi(\cdot)$ denotes the standard normal cumulative distribution function. Coefficients β_a summarize how fundamentals shift the latent index, but economically meaningful effects are

obtained through marginal effects on rating probabilities. For covariate x_m, the marginal effect on the probability of category k is

$$\frac{\partial \Pr(R=k|X)}{\partial x_m} = \beta_{a,m}\left[\phi\left(\mu_{a,k-1} - \eta - X'\beta_a\right) - \phi\left(\mu_{a,k} - \eta - X'\beta_a\right)\right], \quad (4)$$

with $\phi(\cdot)$ the standard normal density and $\eta_{a,i,t} \equiv \alpha_{a,i} + \gamma_{a,t}$. Because marginal effects depend on the position of $R^*_{a,i,t}$ relative to the thresholds, the influence of a given fundamental is inherently state-dependent: changes in covariates have larger probability impacts when an observation is close to a cut-point. This property is central when attention focuses on transitions around salient boundaries (e.g., investment-grade versus speculative-grade) and when ratings are interpreted as categorical signals with coordination effects (Boot et al., 2006). Accordingly, the analysis emphasizes probability objects that align with economic relevance, including the probability of being investment grade and the probability of crossing key thresholds, rather than relying only on coefficient signs.

The panel structure raises the issue of unobserved heterogeneity. Sovereign creditworthiness is shaped by slowly moving characteristics such as long-run institutional capacity and historical credibility that may not be fully captured by the available covariates. In practice, the main analysis therefore relies on clustered pooled ordered probits and on a country-and-year-dummy specification treated as a stress test rather than as a clean fixed-effects ordered-probit estimator. This keeps the interpretation aligned with the reported specifications while still confronting the possibility that slow-moving country characteristics matter for ratings.

Accordingly, the coefficient estimates should be read as shifts in latent creditworthiness, while the substantive interpretation rests on predicted probabilities and marginal effects. This distinction is central when comparing agencies: different thresholds and slopes allow Fitch, Moody's, and S&P to share a common fundamentals core while still assigning different weights to the same observable information. The ordered-probit benchmark therefore disciplines the decomposition by separating a transparent fundamentals-implied rating from the residual agency-specific judgment around it.

A strict timing convention is adopted to align the information set in $X_{i,t}$ with what could plausibly be available at the time ratings are assigned. To reduce look-ahead bias, fundamentals are measured using lagged or otherwise pre-determined values relative to the rating year. This timing discipline is particularly important because rating actions can influence financing conditions and macro outcomes, implying potential feedback from ratings to fundamentals over

time (Adelino and Ferreira, 2016). The empirical strategy therefore treats the ordered probit primarily as a transparent benchmark mapping from observables to ratings, rather than as a structural causal model.

To capture nonlinearities and high-order interactions that may characterize the rating function, a complementary machine learning framework is implemented. Let $g_a(\cdot)$ denote an agency-specific predictive mapping from fundamentals to rating outcomes. In multi-class form, the machine learning models estimate a vector of predicted class probabilities

$$\hat{p}_{a,i,t}(k) = \Pr\big(\hat{R}_{a,i,t} = k \mid X_{i,t}\big), k = 0, \dots, K. \qquad (5)$$

Four flexible model families are used as benchmarks: random forests (Breiman, 2001), support vector machines (Cortes and Vapnik, 1995), neural networks (Rumelhart et al., 1986), and gradient boosting via XGBoost (Friedman, 2001; Chen and Guestrin, 2016). The point is not to win a forecasting tournament for its own sake, but to see whether a richer nonlinear mapping materially changes the broad macro-fiscal structure recovered by the ordered-response benchmark.

Model evaluation follows a time-aware design. Training uses earlier periods and evaluation uses later periods, with tuning restricted to the training window to avoid information leakage. Performance is assessed with ordinally meaningful measures based on notch distance as well as exact and near-exact prediction rates, so the comparison remains tied to the economic meaning of rating errors rather than to headline accuracy alone.

A central objective is comparability between the econometric and machine learning approaches. To place both paradigms on a common footing, fundamentals-implied "shadow ratings" are constructed from each model's probability forecasts. Define a numerical rating score $\hat{S}^{obs}_{a,i,t}$ as the observed rating mapped to the quantitative scale in Appendix Table A1. For machine learning, an expected score can be computed as

$$\hat{S}^{ML}_{a,i,t} = \sum_{k=0}^{K} k \; \hat{p}_{a,i,t}(k). \qquad (6)$$

For an ordered probit, an analogous expected score is obtained from the model-implied probabilities $\hat{p}^{OP}_{a,i,t}(k)$ computed from the estimated thresholds and linear index:

$$\hat{S}^{OP}_{a,i,t} = \sum_{k=0}^{K} k \; \hat{p}^{OP}_{a,i,t}(k). \qquad (7)$$

These expected scores are then mapped back to letter-grade categories using Appendix Table A1, yielding shadow ratings that are directly comparable across models and agencies. The deviation between observed ratings and fundamentals-implied benchmarks is defined as

$$D^{m}_{a,i,t} = S^{obs}_{a,i,t} - \hat{S}^{m}_{a,i,t}, m \in \{OP, ML\}. (8)$$

A positive deviation indicates a rating that is more favourable than the benchmark implied by observed fundamentals, while a negative deviation indicates a more conservative rating. Operationally, on the common numerical scale, the observed rating is decomposed into a fundamentals-implied shadow score and an agency-specific deviation. Treating deviations as an explicit outcome provides an operational transparency measure aligned with evidence that sovereign ratings contain a discretionary component beyond observable fundamentals (De Moor et al., 2018).

The next step is to ask whether bond markets price only the common fundamentals component of ratings or also the agency-specific deviation. Using annual 10-year sovereign yields, the benchmark market specification regresses the sovereign spread versus the German Bund on the observed average rating and, more importantly, on the decomposition into the shadow rating and the agency deviation under country fixed effects and year effects. A lagged version with lagged macro-fiscal controls is preferred because ratings and funding conditions co-evolve within the year, and year-end spread, first-difference, and broader EU yield-level models are used as robustness checks. These regressions are interpreted as pricing associations rather than structural causal estimates.

Interpretability matters because the flexible models are used here as diagnostics rather than as standalone forecasting devices. SHAP values summarize feature contributions, and partial dependence plots trace how predicted ratings move with individual covariates.

The machine-learning evidence therefore plays a deliberately confirmatory role. It is used to check whether flexible benchmarks recover the same broad determinants as the ordered-probit benchmark, whether they reveal clear threshold behaviour in variables such as debt, current accounts, demographics, and revenue capacity, and whether their near-perfect fit validates the ordered-probit deviation used in the market-pricing test that follows.

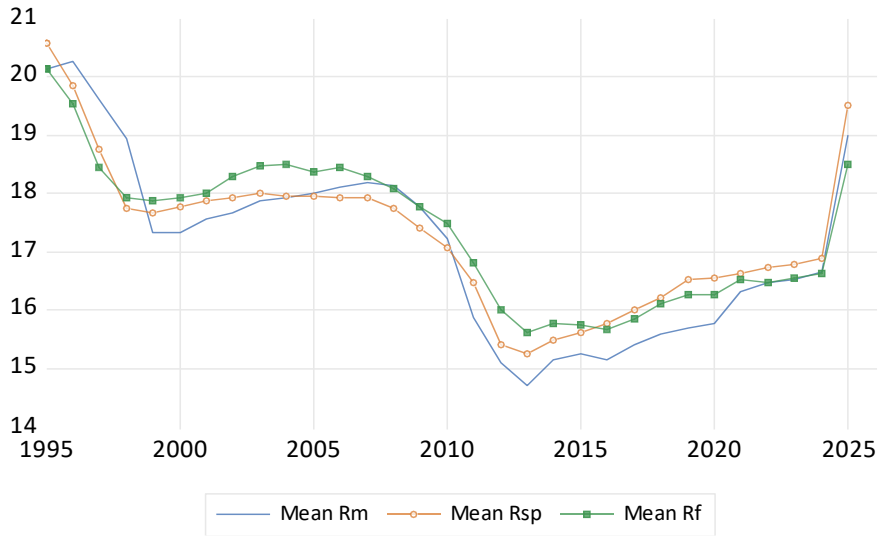
Overall, the empirical strategy combines an ordered-probit benchmark, a shadow-rating decomposition, time-consistent validation, and a limited set of nonlinear diagnostics that keep the exercise transparent and auditable.

4. Data and Stylized Facts

The empirical analysis is based on an annual panel for the countries that constitute the European Union (EU-27) at the end of the sample. The time dimension spans 1995-2024, covering the pre-euro convergence period, the global financial crisis, the euro-area sovereign debt crisis and the post-crisis policy environment, and the COVID-19 disruption. Annual frequency is a practical compromise between the institutional rhythm of sovereign rating decisions and the availability of comparable macro-fiscal data. Because ratings and fundamentals are persistent and some series are not available for every country-year, the working panel is unbalanced, and most within-country variation is concentrated around stress episodes and turning points.

Sovereign ratings are collected for the three major agencies - S&P, Moody's, and Fitch - and focus on long-term issuer assessments so that the underlying concept is comparable across agencies and over time. The raw letter grades are converted into the common 0-21 scale reported in Appendix Table A1, where 21 corresponds to AAA/Aaa and 0 to default. This harmonization preserves the ordinal nature of ratings while allowing numerically consistent cross-agency comparison. The analysis uses the three agency-specific series and their simple cross-agency average. That average is not treated as a literal agency rating; it serves as a compact summary of the common component in sovereign risk assessments.

Working with three agencies also makes disagreement observable within the same sovereign-year. We summarize that disagreement with the within-year range and the within-year standard deviation across agency scores. These measures help distinguish a strong common fundamentals signal from agency-specific judgment and motivate the later decomposition into a fundamentals-implied shadow rating and an agency deviation. Figure 1 shows the broad comovement of the three average rating series, including the common deterioration around the global financial crisis and the euro-area sovereign debt crisis.



Source: authors' calculations. Rm – Moody's; Rsp – S&P; Rf – Fitch. Numerical ratings on the vertical axis.

Figure 1. Quantitative average ratings by agency.

The covariate set combines macroeconomic performance indicators, fiscal sustainability variables, external position measures and demographic pressure indicators that are standard in the sovereign ratings determinants tradition (Cantor and Packer, 1996; Afonso, 2003; Afonso et al., 2011). The main variables are inflation, unemployment, the output gap, government debt, the current account balance, the old-age dependency ratio, a budget balance rule indicator, and general government revenue as a share of GDP, which is interpreted as a proxy for revenue capacity. These choices follow the standard sovereign-ratings determinants literature and cover the main macro-fiscal channels through which solvency, cyclicality, external vulnerability, and fiscal capacity shape ratings.

All variables are constructed so that they are comparable across countries and over time and are aligned with the rating-year information set. In the baseline design, fundamentals are treated as predetermined relative to the rating year whenever feasible, because rating actions can feed back into financing conditions and macro outcomes. The empirical models are therefore interpreted as mappings from observed fundamentals to ratings, not as structural causal systems.

The budget balance rule indicator is taken from the IMF Fiscal Affairs Department Fiscal Rules Dataset, 1985-2024 update (International Monetary Fund, 2025), using the budget balance rule in-place field.

For the market-pricing extension, monthly 10-year sovereign bond yields are taken from the ECB's long-term interest rate statistics and aggregated to annual averages and year-end observations (European Central Bank, 2026; Eurostat, 2026). The main pricing outcome is the

annual average spread against the German Bund, measured in basis points. Because the ECB reports euro-area long-term rates in euro and non-euro member rates in national currencies, the baseline spread regressions are estimated on euro-area sovereign-years only, while a broader EU yield-level panel is used as a robustness check. This design preserves a clean funding-cost interpretation for Bund spreads while retaining the wider public coverage of the ECB dataset.

Because the dataset is assembled from multiple sources and some series are not fully available for all country-years, the estimation sample is unbalanced in practice. To preserve comparability across modelling approaches, each empirical specification is estimated on a common set of country-year observations across the ordered-response and machine-learning models. This ensures that differences in performance or inferred determinants are not mechanically driven by differences in sample composition.

Table 1 summarizes the main variables, sources, transformations, and expected associations used in the analysis.

Table 1. Summary of variables

Variable	*Definition*	*Source*	*Expected sign*
Rf, Rm, Rsp	Long-term issuer ratings mapped to the common 0-21 ordinal scale	Fitch, Moody's, S&P	n.a.
Rravg	Cross-agency average of harmonized ratings	Authors' calculations	n.a.
Inflation	Consumer price inflation rate	AMECO	-
Debt-to-GDP	General government gross debt as a share of GDP	AMECO	-
Current account	Current account balance as a share of GDP	AMECO	+
Unemployment	Unemployment rate	AMECO	-
Budget balance rule	In-place budget balance rule indicator	IMF	+
Output gap	Output gap as a share of potential GDP	AMECO	+
Old-age dependency	Population aged 65+ relative to working-age population	Eurostat	-
Revenue capacity	General government revenue as a share of GDP	AMECO	+
10-year yield / Bund spread	Annual average and year-end 10-year yield; spread against Germany in the euro-area sample	ECB / Eurostat	-

Appendix Table A2 reports the correlation structure among the main variables, including the harmonized rating series. Two stylized facts are especially relevant for the modelling strategy.

First, agency ratings display an extremely strong common component. The pairwise correlations among Moody's, S&P, and Fitch ratings range from 0.954 to 0.971, and the cross-agency average is almost perfectly correlated with each agency series. For the EU sample, most of the variation in ratings is therefore shared across agencies, while still leaving economically meaningful room for systematic agency-specific deviations.

Second, ratings and fundamentals line up in economically intuitive ways, while the fundamentals themselves are sufficiently correlated to make interactions and threshold effects plausible. Ratings are lower when inflation, debt, unemployment, and demographic pressure are higher, and higher when external balances, cyclical conditions, the fiscal-rule indicator, and revenue capacity are stronger. At the same time, output gap and unemployment move closely

in opposite directions, debt co-moves with demographics and revenue capacity, and current-account balances are positively related to both ratings and fiscal capacity. These patterns support the ordered-response benchmark but also justify the limited nonlinear diagnostics reported below.

5. Empirical Analysis

For ease of interpretation, four specifications are central in what follows. The pooled ordered probits in Table 2 provide the baseline mapping from fundamentals to agency ratings. The fixed-effects average-rating model in Table 3 is the preferred within-country robustness specification. Section 5.3 describes the application of four flexible machine learning models (plus a stacking ensemble that combines their predictions) to the same eight fundamentals. In section 5.4, the preferred specification is the lagged Bund-spread regression with country and year fixed effects and lagged controls, because ratings and funding conditions co-evolve within the year. The country-and-year-dummy ordered probits, the year-end and first-difference spread regressions, the broader EU yield-level regressions, and the debt-interaction model are reported as robustness or stress tests.

5.1. Ordered Probit Results

We estimate agency-specific ordered probits because sovereign ratings are ordinal outcomes generated by an underlying continuous assessment of creditworthiness.

For each agency, country, and year, the latent rating index depends on the macro-fiscal covariates and country-specific effects, as shown in equation (9).

$$R_{it}^{*} = \beta X_{it} + \lambda Z_i + a_i + \mu_{it}. \qquad (9)$$

In equation (9), the explanatory vector includes inflation, debt-to-GDP, revenue capacity, the budget balance rule indicator, unemployment, the current account balance, old-age dependency, and the output gap. Observed ratings correspond to intervals of the latent index separated by estimated cut-off points, as shown in equation (10).

$$R_{it} = \begin{cases} AAA\ (Aaa) & if \quad {R_{it}}^{*} > c_{20} \\ AA+\ (Aa1) & if \quad c_{20} > {R_{it}}^{*} > c_{19} \\ AA\ (Aa2) & if \quad c_{19} > {R_{it}}^{*} > c_{18} \\ & \vdots \\ < CCC + (Caa1) & if \quad c_1 > {R_{it}}^{*} \end{cases}. \qquad (10)$$

The slope coefficients and cut-off points are estimated by maximum likelihood. Table 2 reports the baseline pooled ordered probit estimates for sovereign ratings assigned by Fitch, Moody's, and S&P, with standard errors clustered at the country level. A clear and economically coherent pattern emerges across the three agencies: inflation, public debt, and the old-age dependency ratio enter with negative signs, while the current account balance, the budget balance rule indicator, the output gap, and the revenue capacity enter positively. These seven covariates are precisely estimated in the pooled benchmark across agencies, while unemployment remains negative but is not estimated as precisely in the contemporaneous agency-level ordered probits.

The main takeaway is therefore slightly narrower than a simple all-variables story. The agency-level ordered probits still reveal a strong common fundamentals core, but the most stable elements are inflation, debt, current account balance, the budget balance rule indicator, output gap, old-age dependency, and revenue capacity. Labor-market conditions still matter economically, yet their strongest evidence appears in the average-rating panel and in the more demanding specifications rather than uniformly in every pooled ordered probit.

The economic interpretation should nevertheless focus on sign patterns, significance, and implied probability shifts rather than on a mechanical comparison of raw coefficient magnitudes across agencies. Because each ordered probit is separately scale-normalized on its latent index, coefficient sizes are not directly comparable one-for-one across Fitch, Moody's, and S&P. The more credible conclusion is therefore that the agencies share a stable determinants core, not that one agency reacts more strongly than another simply because a reported coefficient is larger in absolute value.

Read in probability terms, these coefficients primarily shift mass away from the upper part of the rating distribution and toward the middle rather than implying a uniform notch-for-notch change across all categories. For a highly rated sovereign, an adverse movement in debt or inflation is therefore more likely to raise the probability of slipping from the AAA/AA range into nearby lower investment-grade categories than to imply an immediate jump into speculative-grade territory. The ordered-probit estimates are therefore most informative about the direction and concentration of rating risk around economically relevant thresholds.

Table 2. Ordered probit results for sovereign ratings by agency.

	Fitch	*Moody's*	*S&P*
Inflation	-0.030***	-0.057***	-0.029***
	(0.007)	(0.018)	(0.007)
Debt-to-GDP	-0.010***	-0.013***	-0.009**
	(0.004)	(0.004)	(0.004)
Current account	0.059***	0.046**	0.063***
	(0.018)	(0.021)	(0.020)
Unemployment	-0.052	-0.068	-0.071
	(0.045)	(0.042)	(0.047)
Budget balance rule	0.783**	0.777***	0.668**
	(0.307)	(0.272)	(0.335)
Output gap	0.068**	0.080***	0.065**
	(0.027)	(0.027)	(0.027)
Old-age dependency	-0.104***	-0.100***	-0.111***
	(0.032)	(0.036)	(0.034)
Revenue capacity	0.123***	0.147***	0.132***
	(0.022)	(0.026)	(0.024)
cut point 1	-3.168*	-3.154**	-3.084**
	(1.634)	(1.470)	(1.493)
cut point 2	-2.948**	-2.774*	-2.531*
	(1.489)	(1.478)	(1.377)
cut point 3	-2.288*	-2.083	-2.218*
	(1.370)	(1.542)	(1.339)
cut point 4	-2.154	-1.982	-2.078
	(1.355)	(1.530)	(1.350)
cut point 5	-1.991	-1.710	-1.891
	(1.366)	(1.493)	(1.342)
cut point 6	-1.784	-1.503	-1.528
	(1.374)	(1.548)	(1.280)
cut point 7	-1.657	-1.400	-1.222
	(1.361)	(1.560)	(1.262)
cut point 8	-1.351	-1.073	-0.759
	(1.290)	(1.506)	(1.293)
cut point 9	-0.861	-0.912	-0.448
	(1.313)	(1.498)	(1.287)
cut point 10	-0.419	-0.681	-0.141
	(1.300)	(1.479)	(1.244)
cut point 11	-0.036	-0.328	0.267
	(1.269)	(1.415)	(1.216)
cut point 12	0.366	0.250	0.664
	(1.241)	(1.402)	(1.211)
cut point 13	0.678	0.528	0.966
	(1.220)	(1.405)	(1.195)
cut point 14	1.023	0.768	1.151
	(1.196)	(1.371)	(1.199)
cut point 15	1.378	1.058	1.662
	(1.210)	(1.361)	(1.235)
cut point 16	1.819	1.543	2.031
	(1.242)	(1.376)	(1.255)
cut point 17	2.084*	1.910	
	(1.254)	(1.373)	
cut point 18		2.120	
		(1.393)	
cut point 19		2.507*	
		(1.414)	
cut point 20		2.834**	
		(1.440)	
Obs.	724	710	746
Pseudo R^2	0.186	0.203	0.196

Notes: Ordered-probit estimates. Standard errors are reported in parentheses. Cut points are estimated thresholds on the latent rating index. Higher rating scores indicate stronger credit quality. ***, **, and * denote statistical significance at the 1%, 5%, and 10% levels, respectively.

The stronger robustness exercise is to add country and year dummies to the ordered probit while clustering at the country level. Once this is accounted for, several coefficients become weaker or unstable, and the models display substantial complete determination, which is a

warning sign rather than a result to celebrate. Therefore, these saturated specifications should be read as demanding stress tests, not as clean fixed-effects ordered probit estimators. Under this harsh specification, debt and the budget balance rule indicator remain comparatively resilient, unemployment becomes more important in several cases, while current account, output gap, old-age dependency, and revenue capacity lose much of their pooled cross-sectional strength. This is consistent with the broader interpretation that a large part of the sovereign-ratings signal is cross-sectional and slow-moving, whereas within-country movements are more selectively priced.

5.2. Cross-Agency Panel Evidence

To isolate the common component of agency assessments, Table 3 estimates panel regressions for the yearly cross-agency average rating using pooled OLS and a country fixed-effects specification with year dummies, both with country-clustered standard errors, under contemporaneous and one-year-lagged regressors. This exercise is best viewed as a robustness check rather than as a substitute for the ordered-response benchmark, because the averaged rating is treated as a quantitative score and therefore imposes a cardinal approximation on an inherently ordinal outcome. Its value is that it reveals which determinants survive once agency-specific noise is averaged out and the specification shifts from cross-sectional differences to within-country variation over time.

Table 3. Panel results for the average sovereign rating.

	(1)	(2)	(3)	(4)
Explanatory variables	***Contemporaneous***		***One-year lag***	
Inflation	-0.088***	-0.040***	-0.010**	-0.002
	(0.014)	(0.011)	(0.004)	(0.002)
Debt-to-GDP	-0.021*	-0.062***	-0.020**	-0.059***
	(0.010)	(0.010)	(0.009)	(0.008)
Current account	0.096**	-0.040	0.110**	-0.025
	(0.044)	(0.030)	(0.043)	(0.031)
Unemployment	-0.184*	-0.180**	-0.148	-0.124*
	(0.103)	(0.073)	(0.099)	(0.062)
Budget balance rule	1.466*	0.246	1.636*	0.615
	(0.829)	(0.604)	(0.872)	(0.614)
Output gap	0.188**	0.047	0.193**	0.068
	(0.077)	(0.044)	(0.086)	(0.056)
Old-age dependency	-0.211***	0.046	-0.197***	0.098
	(0.066)	(0.091)	(0.063)	(0.101)
Revenue capacity	0.257***	-0.048	0.259***	-0.021
	(0.040)	(0.081)	(0.044)	(0.070)
Constant	15.367***	24.816***	14.051***	20.559***
	(2.481)	(4.781)	(2.661)	(4.553)
Country FE & Year FE	No	Yes	No	Yes
Obs.	748	748	755	755
R^2	0.574	0.676	0.572	0.609

Notes: Country-clustered standard errors are reported in parentheses. Columns (1)-(2) use contemporaneous regressors; columns (3)-(4) use one-year lags. Higher rating scores indicate stronger credit quality. ***, **, and * denote statistical significance at the 1%, 5%, and 10% levels, respectively.

5.3. Machine Learning Evidence

The Machine Learning evidence is reported in full so that the cross-agency diagnostics can be evaluated alongside the pricing results. The purpose is not to turn the paper into a forecasting tournament, but to audit whether flexible learners recover the same macro-fiscal structure as the ordered-response benchmark, to identify where nonlinearities and interactions are economically material, and to assess whether agency-specific models add information beyond the cross-agency average. Four flexible model families are benchmarked against a standard OLS baseline. Random forests (Breiman, 2001) aggregate predictions from large ensembles of decision trees; they are famously robust to irrelevant features and multicollinearity. Support vector machines (Cortes and Vapnik, 1995) find the maximum-margin boundary in a transformed feature space; they ML method that best handles small datasets. Neural networks (Rumelhart et al., 1986) approximate arbitrary nonlinear functions through layered representations; here a feedforward architecture with dropout regularization to limit overfitting on the relatively small sovereign panel. XGBoost (Chen and Guestrin, 2016) builds an additive ensemble of shallow trees through gradient boosting and is particularly effective at capturing threshold effects and variable interactions. In addition to the four individual learners, a stacked ensemble combines their predictions through a meta-learner, allowing the final forecast to adaptively weight each model's contribution depending on where in the feature space it performs best. All models are trained on the same set of macro-fiscal fundamentals used in the ordered-probit benchmark, with performance assessed on a held-out validation window to ensure time consistency. Table 4 reports the results for the cross-agency average rating.

Table 4. Performance of the evaluated models: average rating.

Model	RMSE	MAE	MAPE	R2	Exact	Within 1
OLS baseline	2.405	1.863	0.131	0.574	0.092	0.433
SVM	1.370	0.799	0.053	0.862	0.369	0.789
XGBoost	0.118	0.079	0.005	**0.999**	**0.563**	**1.000**
Random Forest	0.573	0.361	0.026	0.976	0.439	0.941
Decision Tree	**0.117**	**0.022**	**0.001**	**0.999**	**0.563**	0.997
Neural Network	0.680	0.449	0.029	0.966	0.406	0.940
Stacking	0.900	0.569	0.041	0.940	0.393	0.849

Note: RMSE = root mean squared error; MAE = mean absolute error; MAPE = mean absolute percentage error; R2 = coefficient of determination; Exact = share of exact rating predictions; Within 1 = share of predictions within one notch.

Figure 2 reports the random-forest variable-importance ranking for the cross-agency average rating. The result is useful because it shows that the nonlinear model is not finding a diffuse, atheoretical prediction rule. Revenue capacity, measured in the underlying data as the tax-burden ratio, is the dominant predictor, followed by public debt and the current-account balance, implying an important fiscal policy implication. The ranking therefore mirrors the

ordered-response evidence: rating agencies appear to reward fiscal capacity and external resilience while penalizing debt overhang, with cyclical and institutional controls playing a secondary but still informative role.

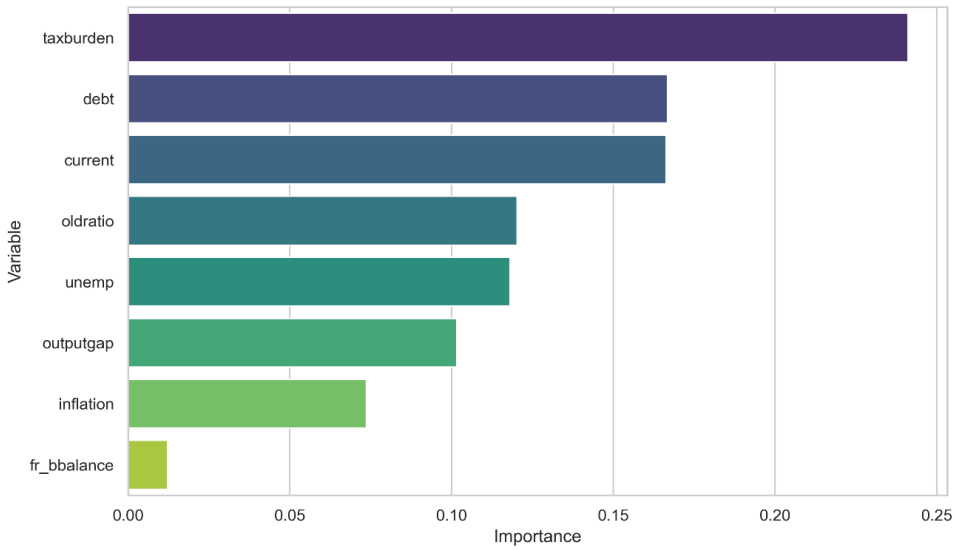


Figure 2. Variable-importance feature rankings for the cross-agency average rating.

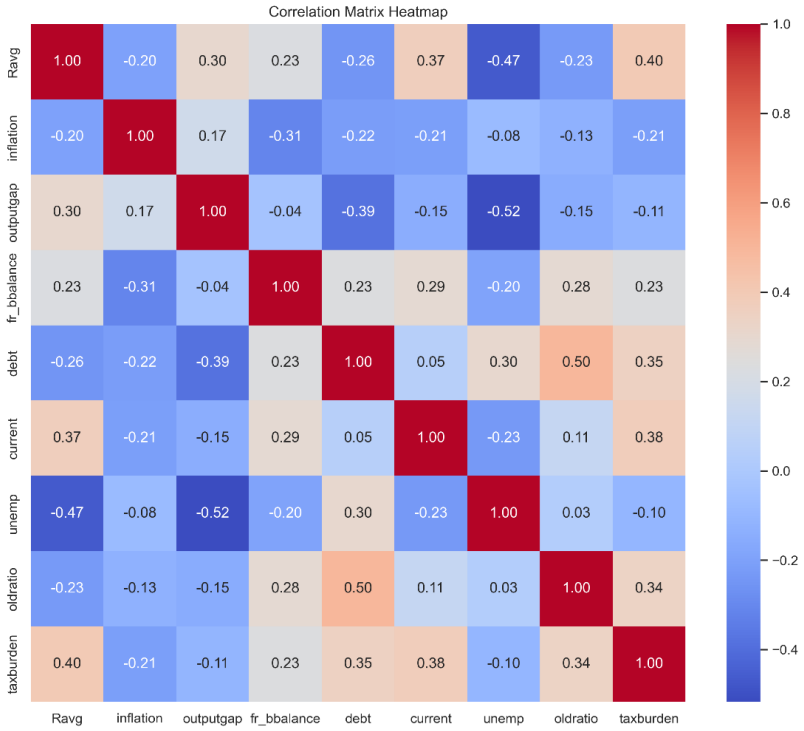


Figure 3. Correlation heatmap for the cross-agency average rating.

Figure 3 provides the bivariate discipline behind the multivariate results. The average rating is positively associated with output gaps, current-account balances, fiscal-rule coverage, and revenue capacity, and negatively associated with inflation, debt, unemployment, and old-age dependency. This is not meant to substitute for the ordered-probit and panel specifications, because many of these fundamentals are mutually correlated. Rather, the heatmap shows that the signs recovered by the econometric and machine-learning models are anchored in economically plausible raw covariation.

Figure 4 uses mean absolute SHAP values to ask whether the ranking in Figure 2 survives a contribution-based decomposition of the fitted model. It does. Revenue capacity remains the leading feature, while old-age dependency, the current account, debt, unemployment, output gap, inflation, and the budget-balance rule follow in an economically coherent order. The agreement between permutation-style importance and SHAP importance matters because it reduces the risk that the result is an artefact of a single importance metric.

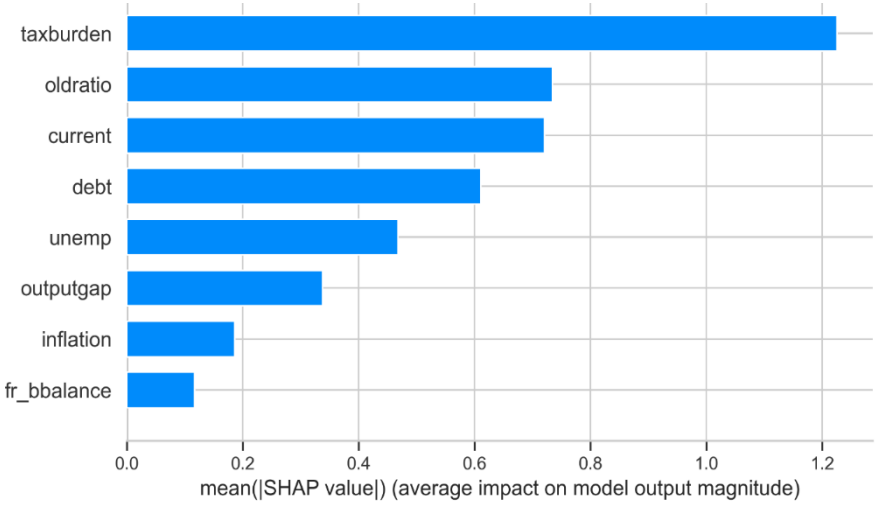


Figure 4. SHAP feature importance for the cross-agency average rating.

Figure 5 makes the revenue-capacity channel economically visible. The relationship is convex rather than linear: low revenue capacity is associated with weak predicted ratings, but the rating premium rises disproportionately once the revenue ratio moves into the upper part of the EU distribution. This pattern is consistent with the interpretation that fiscal capacity is valuable not only as a flow variable, but also as a signal of administrative capacity, repayment capacity, and policy space under stress.

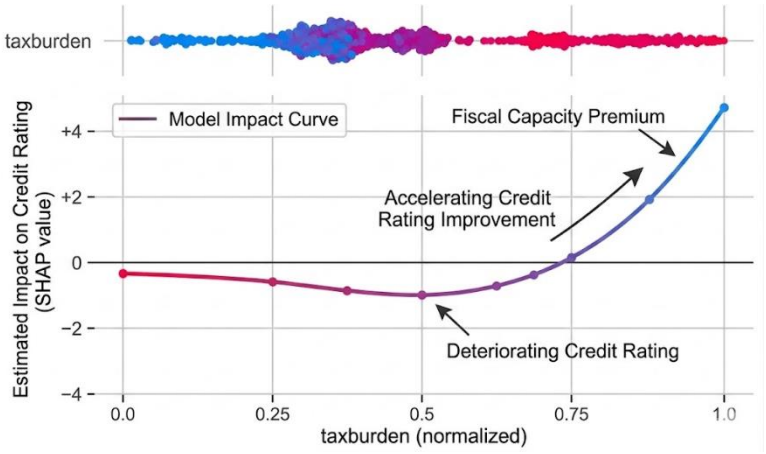


Figure 5. Convex relationship between revenue capacity and the predicted average rating.

Figure 6 reports the full SHAP summary plot for the cross-agency average rating. Unlike the bar ranking, the beeswarm display shows both magnitude and direction. High revenue capacity pushes predictions upward, while high debt, unemployment, inflation, and old-age dependency tend to lower predicted ratings. The wider horizontal spread for revenue capacity, debt, and the current account indicates that these variables matter most when they interact with the broader fiscal and external state of the sovereign.

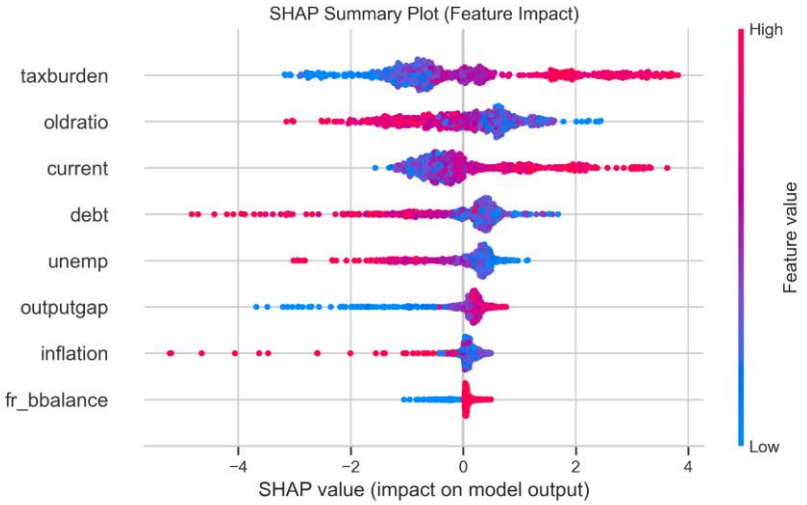


Figure 6. SHAP summary plot for the cross-agency average rating.

There is a very important note to add here: the SHAP analysis corresponds exactly to the sign probit coefficients. The directional signal is consistent across all eight variables, every variable with a positive probit coefficient contributes positively to the predicted ratings in the SHAP decompositions, and every variable with a negative coefficient contributes negatively. This agreement holds despite the two frameworks optimising different objectives: marginal linear effects in the probit versus nonlinear feature contributions in the ML.

Figure 7 complements the SHAP evidence with univariate Partial-Dependence Plots (PDPs). The shape of the PDPs is largely based on economic intuition. Debt seems to be irrelevant to ratings for debt to GDP ratios from 0 to 60%. After that threshold, it starts to negatively and significantly impact ratings as it is depicted by the steep negative slope of the PDP, while its effect softens after the 100% ratio of debt to GDP, where the slope becomes less negative. Revenue capacity has a positive rating reward region starting at 30% and up to the high 40s of tax to GDP ratio. After that it exhibits a significant jump and finally flattens at 50% as probably reasonable fears of over taxation and Laffer curve effects may rise. The current account ratio exhibits an almost flat (zero slope) region for deficits, it starts to reward the sovereign with higher ratings after it gets balanced, but these rewards evaporate after 5%. Demographic pressure seems to not affect ratings up to 35%, but after that the slope becomes negative, leading to lower ratings. Rising unemployment at low rates has a small negative effect, nonetheless, at approximately 10% we observe a small negative threshold effect.

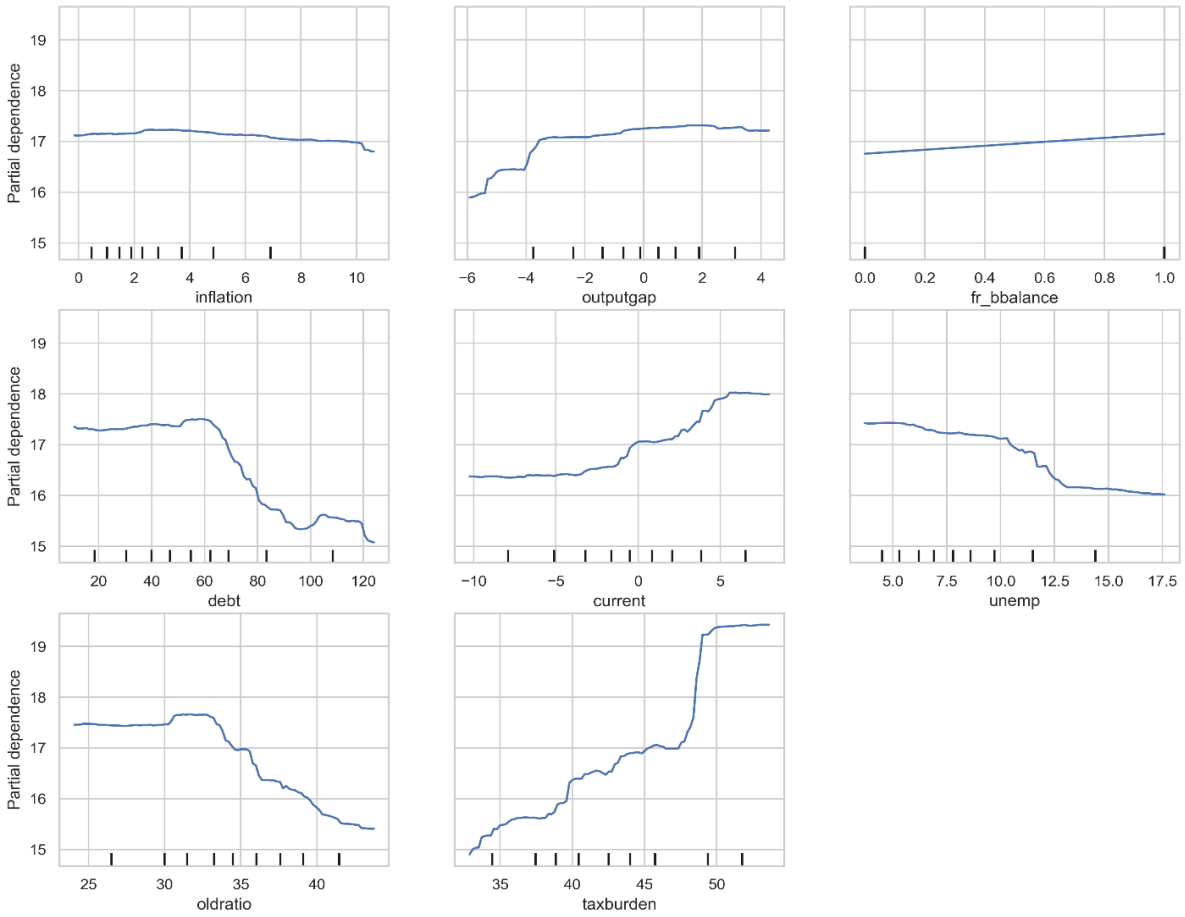


Figure 7. Partial-dependence plots for the cross-agency average rating.

Figure 8 shifts from one-dimensional nonlinearities to interactions among the leading fundamentals. The highest predicted ratings occur when strong revenue capacity is paired with low or moderate debt, and when revenue capacity is accompanied by a current-account surplus. Conversely, high debt is most damaging when it coincides with weaker fiscal capacity or an external deficit. This is precisely the type of conditional structure that an additive ordered-probit index may compress: agencies do not appear to price debt, fiscal capacity, and external balance as isolated signals, but as mutually reinforcing elements of sovereign resilience.

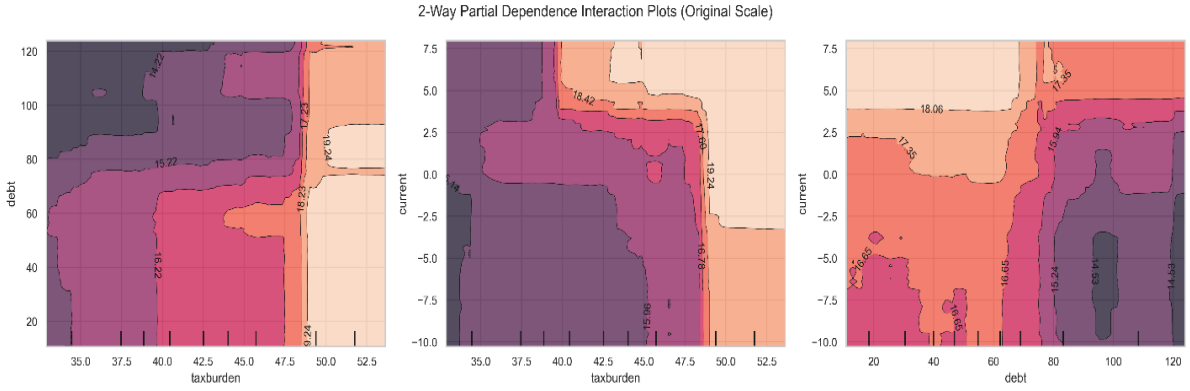


Figure 8. Two-way partial-dependence interaction plots for the cross-agency average rating.

Table 5 extends the performance comparison of our best models to agency-specific ratings for Fitch, Moody's, and S&P individually. The hierarchy across models is consistent with Table 4: XGBoost and Random Forest substantially outperform the OLS baseline across all three agencies on every metric. The OLS baseline R^2 ranges narrowly from 0.556 to 0.589 across agencies, while XGBoost achieves $R^2 = 0.999$ for all three. These differences provide evidence in support of the non-linear nature of the process of assigning sovereign ratings. Cross-agency differences in model performance are modest relative to the gap between the linear benchmark and the non-linear machine learning models, which is the main diagnostic message: the fundamentals-rating mapping contains structure that a linear specification systematically misses, and this holds equally for each of the three major agencies.

Table 5. Performance of the evaluated models: agency-specific ratings.

Agency	Model	RMSE	MAE	MAPE	R2	Exact	Within 1
Fitch	OLS baseline	2.338	1.815	0.124	0.556	0.162	0.500
	SVM	1.391	0.807	0.053	0.843	0.575	0.837
	XGBoost	**0.109**	**0.075**	**0.005**	**0.999**	**1.000**	**1.000**
	Random Forest	0.541	0.359	0.025	0.976	0.747	0.978
	Decision Tree	1.845	1.249	0.083	0.723	0.407	0.688
	Neural Network	0.746	0.470	0.030	0.955	0.682	0.956
	Stacking	0.859	0.602	0.040	0.940	0.562	0.920
Moody's	OLS baseline	2.436	1.870	0.129	0.563	0.151	0.503
	SVM	1.456	0.825	0.056	0.844	0.580	0.820
	XGBoost	**0.128**	**0.087**	**0.005**	**0.999**	**0.995**	**1.000**
	Random Forest	0.905	0.603	0.041	0.940	0.579	0.893
	Decision Tree	2.050	1.400	0.094	0.690	0.409	0.622
	Neural Network	0.663	0.421	0.027	0.968	0.730	0.972
	Stacking	0.944	0.606	0.042	0.934	0.574	0.902
S&P	OLS baseline	2.553	1.994	0.158	0.589	0.163	0.446
	SVM	1.492	0.855	0.062	0.859	0.597	0.807
	XGBoost	**0.148**	**0.100**	**0.007**	**0.999**	**0.993**	**1.000**
	Random Forest	0.590	0.377	0.032	0.978	0.744	0.956
	Decision Tree	2.062	1.325	0.096	0.731	0.456	0.686
	Neural Network	0.824	0.548	0.040	0.957	0.622	0.944
	Stacking	0.982	0.644	0.055	0.939	0.531	0.899

Note: RMSE = root mean squared error; MAE = mean absolute error; MAPE = mean absolute percentage error; R2 = coefficient of determination; Exact = share of exact rating predictions; Within 1 = share of predictions within one notch.

Table 6 repeats the variable-importance exercise separately for Fitch, Moody's, and S&P. The first-order message is convergence: the same broad fundamentals dominate across agencies, which supports the paper's common-core interpretation. The second-order message is agency heterogeneity. Moody's places relatively more weight on cyclical indicators, especially the output gap and unemployment, while S&P assigns relatively more weight to the external position. Fitch sits closer to the average ordering.

Table 6. Variable-importance rankings by agency: Fitch, Moody's, and S&P.

Variable	Fitch	Moody's	S&P
Taxburden	0.249 (1)	0.259 (1)	0.246 (1)
Debt	0.182 (2)	0.183 (2)	0.142 (3)
Current	0.176 (3)	0.136 (3)	0.193 (2)
Old ratio	0.135 (4)	0.110 (6)	0.122 (4)
Output gap	0.095 (5)	0.111 (5)	0.109 (5)
Unemployment	0.080 (6)	0.119 (4)	0.108 (6)
Inflation	0.075 (7)	0.068 (7)	0.073 (7)
Fr_bbalance	0.007 (8)	0.013 (8)	0.008 (8)

Note: the number in parenthesis shows the ordinal position of every feature per CRA.

Table 7 provides the SHAP counterpart to Table 6. Its contribution is to show that the agency-specific ranking is not driven solely by the random-forest importance metric. Revenue capacity remains the strongest feature across the three agencies, while the current account is especially important for S&P. Demographics as described by the old ratio variable, appear in the top three variables for all agencies. Unemployment, output gap, inflation and the budget balance are ranked consistently, in this order, across all three agencies at the last four places and with an importance only a fraction of the top ranked variable, the tax burden.

Table 7. SHAP rankings by agency: Fitch, Moody's, and S&P.

Variable	Fitch	Moody's	S&P
Taxburden	1.23 (1)	1.42 (1)	1.19 (1)
Old ratio	0.79 (2)	0.69 (3)	0.73 (3)
Current	0.71 (3)	0.63 (4)	0.81 (2)
Debt	0.61 (4)	0.73 (2)	0.50 (4)
Unemployment	0.31 (5)	0.46 (5)	0.42 (5)
Outputgap	0.28 (6)	0.37 (6)	0.37 (6)
Inflation	0.19 (7)	0.16 (7)	0.21 (7)
Fr_bbalance	0.07 (8)	0.12 (8)	0.07 (8)

Note: the number in parenthesis shows the ordinal position of every feature per CRA.

The alignment between Tables 6 and 7 strengthens the interpretation that agency discretion is expressed mainly through differences in weights, not through an entirely different set of fundamentals. Across both the random forest variable-importance rankings (Table 6) and the SHAP decompositions (Table 7), the agency-level results display a consistent hierarchical structure. Revenue capacity (taxburden) ranks first for all three agencies under both metrics, confirming its position as the dominant fundamental in the rating function regardless of the

method used to measure it. At the other end of the distribution, inflation and the budget balance rule (fr_bbalance) are consistently the two least important variables, occupying the seventh and eighth positions across all agencies and both metrics. In the middle tier, debt and the current account always fall within positions two to four, though their relative order varies by agency (Moody's VIM places debt above the current account while S&P's SHAP reverses that order) reflecting agency-specific sensitivity to external versus fiscal-stock signals. Output gap is consistently in the range of positions four to six. The two variables that shift most across agencies and metrics are old-age dependency and unemployment, which exchange positions depending on whether the method is VIM or SHAP and on which agency is considered, suggesting that demographic and labour-market pressures are weighted differently across rating committees.

Figure 9 adds the observation-level direction of the SHAP effects. Across agencies, high debt, high unemployment, high inflation, and demographic pressure tend to push predicted ratings downward, while stronger revenue capacity and external balances push ratings upward. The dispersion of the dots also matters: the same variable can have a larger or smaller contribution depending on the surrounding macro-fiscal state. The figure therefore supports a central argument of the paper: sovereign ratings are ordinal and threshold-based, but their empirical mapping from fundamentals is state-contingent rather than purely linear.

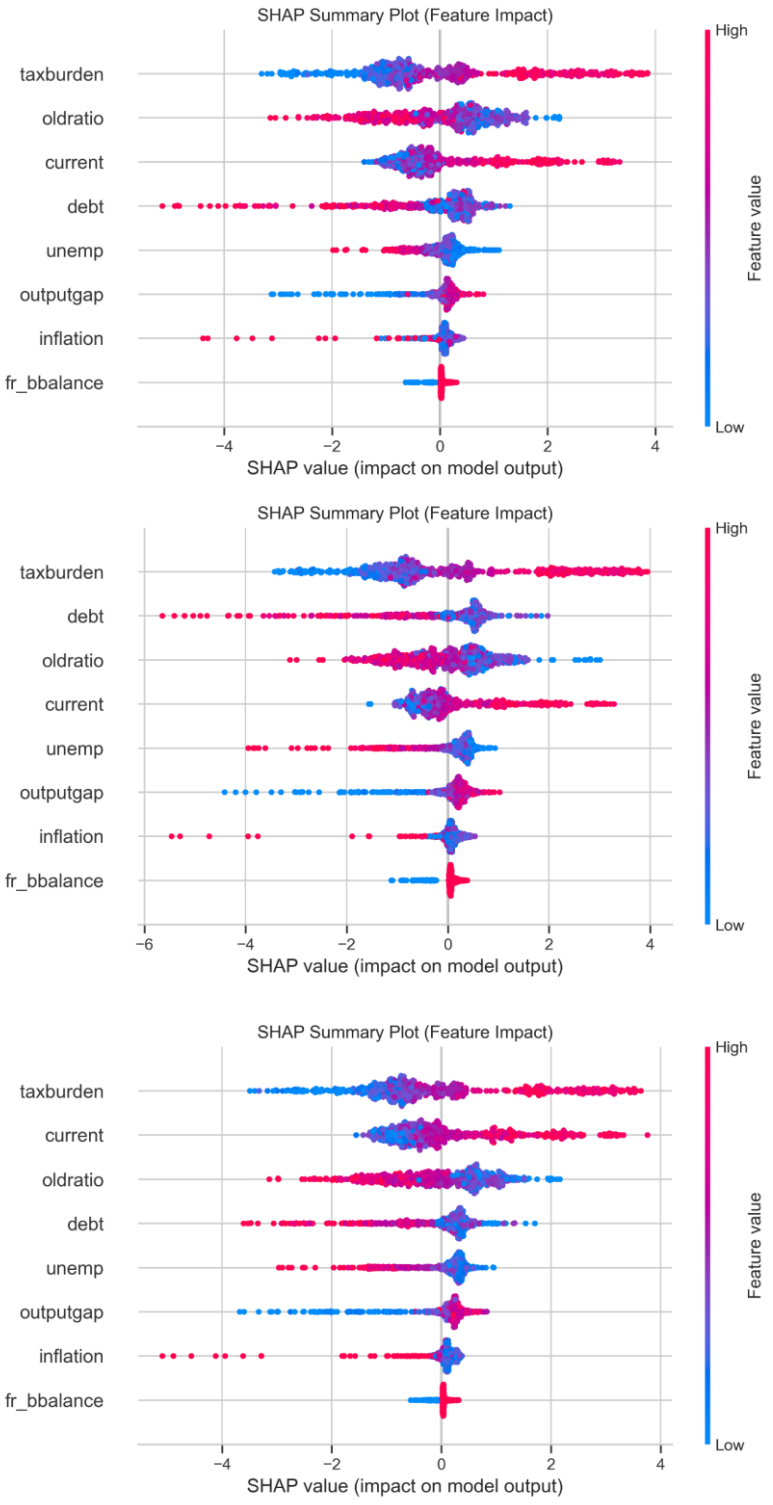


Figure 9. SHAP summary plots by agency: Fitch, Moody's, and S&P.

Figure 10 reports agency-specific partial-dependence profiles for four high-leverage determinants: current-account balance, public debt, tax burden/revenue capacity, and old-age dependency. The cross-agency similarity is useful because it shows that the nonlinear structure is not an artefact of averaging across agencies. At the same time, differences in slope and

threshold location are consistent with the paper's interpretation that agency-specific deviations are overlays around a common fundamentals-based benchmark.

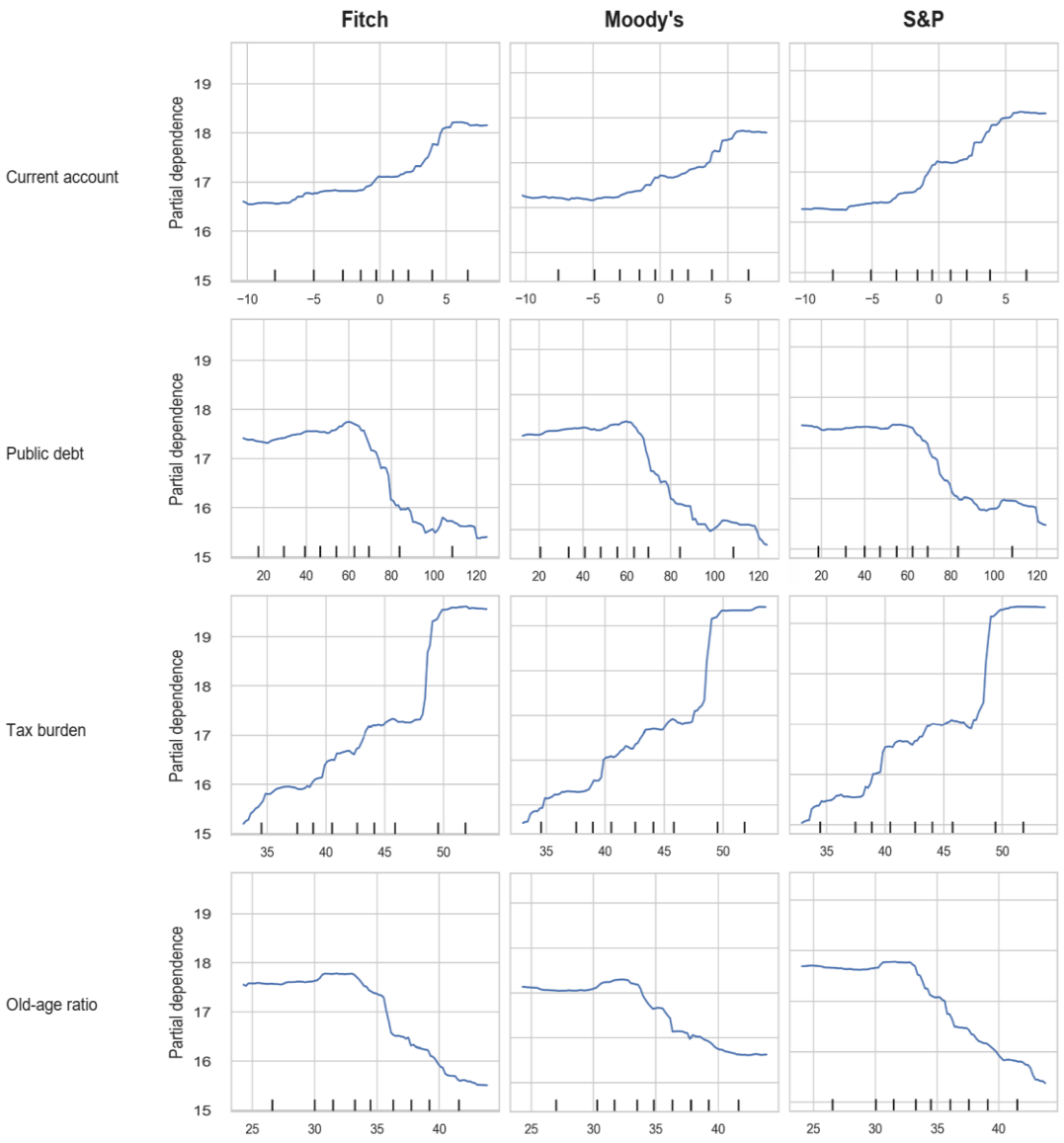


Figure 10. Agency-specific partial-dependence plots: current-account balance, public debt, tax burden/revenue capacity, and old-age dependency.

Figure 11 closes the diagnostic evidence by showing agency-specific two-way partial-dependence surfaces. The surfaces are remarkably similar across Fitch, Moody's, and S&P: high revenue capacity mitigates the rating cost of debt, external surpluses are most valuable when fiscal fundamentals are already strong, and the combination of high debt with external weakness is consistently penalized. The figure therefore reconciles the two sides of the evidence. Agencies share a common macro-fiscal rating technology, but the intensity of the interaction terms varies enough to leave economically meaningful agency-specific deviations.

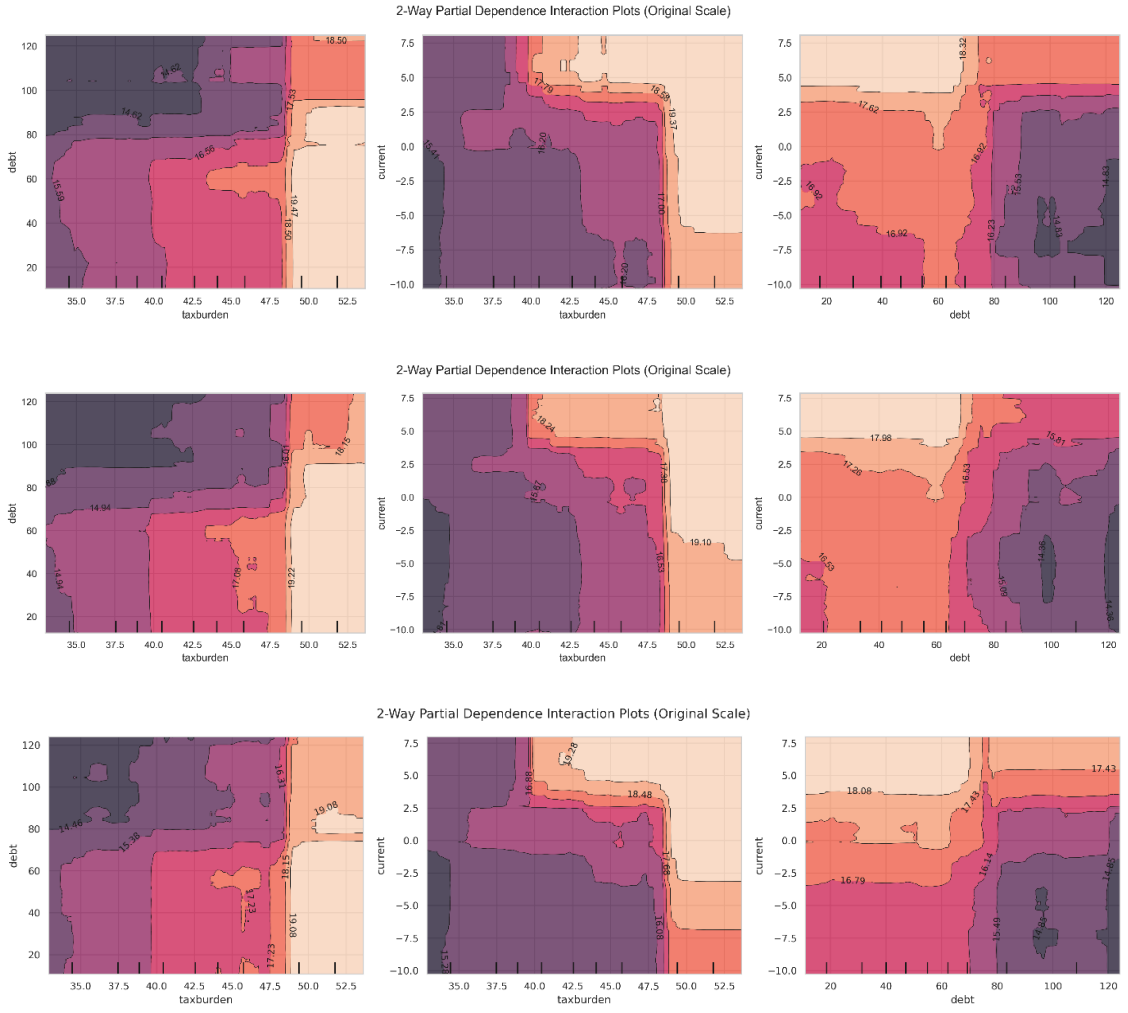


Figure 11. Two-way partial-dependence interaction plots by agency: Fitch, Moody's, and S&P.

5.4. Market Pricing

The market-pricing exercise asks whether bond markets respond to the full observed rating or whether they separately price its two components: the part explained by common macro-fiscal fundamentals and the agency-specific residual above or below that benchmark. The decomposition follows directly from the ordered-probit estimates: for each sovereign-year, the fundamentals-implied shadow rating is the expected rating score given the observed macro-fiscal covariates, and the agency deviation is the difference between the observed rating and that benchmark. Using annual 10-year sovereign bond yields from the ECB, the dependent variable is the annual average spread versus the German Bund in basis points. The main sample is an unbalanced panel of 19 euro-area sovereigns over 1999–2024. All specifications include country and year fixed effects; the preferred specification adds lagged macro controls to limit reverse causality. These estimates are not meant to deliver structural identification, since they

show whether ratings behave in markets as purely fundamentals-based summaries or as signals that also contain agency-specific overlays.

Table 8. Sovereign Bund spreads, fundamentals-implied ratings, and agency deviations, 1999-2024.

Variable	*Observed rating*	*Shadow rating*	*Decomposition*	*Lagged decomposition*	*Asymmetric deviations*
Observed average rating	-51.570***				
	(13.843)				
Shadow average rating		-49.304*	-51.419**	-61.102	-49.144**
		(27.639)	(18.688)	(47.076)	(18.707)
Average agency deviation			-52.001***	-30.769***	
			(11.831)	(7.724)	
Positive deviation					-42.885***
					(12.021)
Negative deviation					-64.041***
					(20.949)
Observations	380	361	361	379	361
Countries	19	19	19	19	19
Within R-squared	0.630	0.448	0.633	0.544	0.638
Country FE	Yes	Yes	Yes	Yes	Yes
Year FE	Yes	Yes	Yes	Yes	Yes
Controls	No	No	No	Yes, lagged	No
Sample	Euro-area sovereign-years with Bund spreads	Euro-area sovereign-years with Bund spreads	Euro-area sovereign-years with Bund spreads	Euro-area sovereign-years with Bund spreads	Euro-area sovereign-years with Bund spreads

Notes: Dependent variable is the annual average 10-year sovereign yield spread versus Germany, in basis points. Robust standard errors clustered by country in parentheses. Better ratings should lower spreads, so negative coefficients indicate tighter funding conditions. ***, **, and * denote statistical significance at the 1%, 5%, and 10% levels, respectively.

The estimates confirm that markets price more than the common fundamentals core. In the reduced-form specification (column 1), a one-notch improvement in the observed average rating is associated with roughly 52 basis points lower Bund spreads. Once the rating is decomposed (column 3), the shadow rating and the agency deviation are both large and negative: in the contemporaneous specification, a one-notch more favourable shadow rating is associated with about 51 basis points lower spreads, while a one-notch more favourable agency deviation is associated with about 52 basis points lower spreads. The within R^2 rises from 0.448 in the shadow-rating-only specification (column 2) to 0.633 in the decomposition, matching the fit of the observed-rating baseline and confirming that the two components together account for the same market information as the observed rating. The lagged specification (column 4) shows that the agency deviation remains significant at −31 basis points even after controlling for lagged macro fundamentals, addressing the concern that the deviation proxies for current conditions already visible to markets. Column 5 splits the deviation into positive and negative components: negative deviations, where agencies are more conservative than fundamentals imply, are priced more strongly (−64 basis points) than positive ones (−43 basis points), suggesting that the markets respond asymmetrically to agency conservatism. This is a very interesting finding, suggesting that bond markets are highly sensitive to "precautionary" or

"unjustifiably conservative" rating stances by CRAs, potentially due to regulatory triggers (e.g., ECB collateral eligibility haircut thresholds, etc). Given an average Bund spread of about 104 basis points in the estimation sample, these are economically large effects.

Appendix evidence helps interpret that pricing result. Appendix Table A5 and Appendix Figure A1 show that the spread effect of agency deviations becomes more negative as debt rises, and a discrete high-debt-state split indicates that once debt exceeds 100% of GDP the per-notch effect increases from about 44 to 64 basis points. Appendix Table A6 adds a short crisis-period comparison for 2008-2013. That interaction has the same sign but is estimated less precisely, so the debt-state channel is the more stable mechanism. Markets therefore do not appear to respond only to fundamentals; they also respond to whether agencies are unusually harsh or unusually lenient relative to those fundamentals. This evidence should nevertheless be read as associational rather than causal. Even with lagged and first-difference specifications, the annual panel cannot fully separate agency information from market anticipation, omitted common shocks, or simultaneous responses by ratings and yields to the same news. The main Bund-spread sample also trades breadth for comparability: non-euro sovereigns are analysed through yield-level robustness checks rather than through a direct spread against Germany.

The market-pricing exercise is built around the ordered-probit shadow rating rather than the ML-based shadow rating for a specific reason. The nonlinear diagnostics in Section 5.3 show that flexible models absorb nearly all systematic variation between fundamentals and ratings ($R^2 \approx 0.999$), leaving a residual agency deviation with near-zero variance (analytical results of the machine learning based exercise can be found in Table A7). A variable with almost no variance cannot be precisely estimated in a panel regression; the standard errors on the ML-based deviation are large enough to preclude inference, not because agency judgment is economically negligible but because the ML fit is too tight for the residual to be identifiable. The ordered-probit deviation retains enough cross-country and time-series variation to be identified precisely, and the near-perfect ML fit retroactively validates it as a genuine measure of agency discretion rather than functional-form misspecification. The ML exercise therefore supports the pricing test indirectly by confirming the cleanliness of the probit residual, rather than entering it directly.

5.5. Discussion

Three points stand out. The ordered-probit and fixed-effects evidence indicates that EU sovereign ratings share a common macro-fiscal core, but within-country rating changes are much narrower than pooled cross-country estimates suggest. The fixed-effects average-rating panel is useful precisely because it separates slow-moving cross-sectional structure from the subset of fundamentals that move ratings within sovereigns over time.

The market-pricing evidence shows that both the shadow rating and the agency deviation matter for sovereign funding conditions, and the deviation term survives the most informative lagged and first-difference checks. This is consistent with a setting in which ratings operate as macro-financial state variables through collateral frameworks, regulation, benchmark mandates, and portfolio rebalancing.

The mechanism evidence sharpens that interpretation. Agency deviations are priced more strongly when sovereign balance sheets are more fragile, and the discrete high-debt-state split shows that the effect intensifies further once debt rises above 100% of GDP. A supplementary crisis-period comparison is directionally similar but less precise, again pointing to balance-sheet fragility as the clearer mechanism.

What markets are pricing in the agency-deviation term remains open to interpretation. It may reflect informational advantages of rating committees, certification effects around salient rating thresholds, regulatory salience, or common reactions of agencies and investors to the same news. The annual panel cannot cleanly separate those channels. The narrower conclusion supported by the evidence is that deviations from the fundamentals benchmark are not treated by bond markets as irrelevant residuals.

Taken together, the ordered-response estimates, market-pricing tests, and XAI (Explainable AI) diagnostics should be read as complementary pieces of evidence: the ordered model supplies the structural benchmark, the pricing regressions ask whether deviations from that benchmark matter for spreads, and the nonlinear diagnostics test whether the benchmark misses economically relevant thresholds or interactions.

The nonlinear evidence adds an interpretive layer to these results. Debt and fiscal capacity display threshold behaviour, while agency differences are present but modest relative to the common structure. Taken together, the findings suggest that ratings combine a common macro-fiscal core with discretionary overlays that matter for sovereign-risk pricing.

The agency-specific diagnostics add nuance without overturning the common-core result. Moody's appears relatively more sensitive to cyclical indicators such as the output gap and unemployment, S&P gives somewhat greater weight to external-position variables, and Fitch

sits closer to the common macro-fiscal ordering. These differences are economically useful because they identify where discretion enters the rating process: not as a wholesale alternative to fundamentals, but as variation in how agencies weight business-cycle, external, and fiscal-capacity signals.

6. Conclusions

Overall, the findings show that EU sovereign ratings are not well described as mechanical summaries of macro-fiscal fundamentals alone. Using an ordered-probit benchmark and a fixed-effects panel for the cross-agency average, we find that inflation, debt-to-GDP, current account balance, the budget balance rule indicator, output gap, old-age dependency, and revenue capacity are the most stable correlates of ratings across Fitch, Moody's, and S&P, while within-country variation is narrower and concentrated in fewer variables. The relevance of the tax burden and of the debt-to-GDP ratio conveys important policy messages for fiscal policy makers. The country-and-year-dummy ordered probits are considerably less stable and are best interpreted as stress tests rather than as clean fixed-effects estimators.

ECB 10-year sovereign yield data further show that both the fundamentals-implied shadow rating and the agency deviation are priced in euro-area Bund spreads. A one-notch more favourable value of either term is associated with about 50 basis points lower spreads in the baseline decomposition, and the deviation term remains especially robust in lagged, year-end, first-difference, and broader EU yield-level specifications. Appendix Table A5 and Appendix Figure A1 show that this pricing effect steepens as debt rises, while the discrete high-debt-state split indicates that once debt exceeds 100% of GDP the per-notch effect of deviations increases by about 20 basis points. Appendix Table A6 shows a crisis-period interaction with the same sign but less precision.

Overall, agency deviations are priced more strongly when government balance sheets are more fragile, and the effect intensifies further once the debt-to-GDP ratio rises above 100%. A supplementary crisis-period comparison is similar but less precise, again pointing to balance-sheet fragility as the clearer mechanism.

Two limitations are worth emphasizing. First, the spread regressions do not identify causal effects; they show pricing associations in a setting where ratings and market conditions co-evolve over time. Second, the main pricing sample is restricted to euro-area sovereign-years to preserve a clean Bund-spread interpretation, while the annual frequency inevitably smooths within-year rating actions and market reactions.

More broadly, sovereign ratings appear to combine a common fundamentals core with discretionary overlays that markets still treat as relevant information. That reading is consistent with the sovereign-bank nexus, collateral frameworks, regulation, and portfolio mandates that give ratings macro-financial force. Agency judgment is therefore not simply residual noise around fundamentals; it is part of the information that sovereign bond markets appear to price.

References

1. Abad, P., Alsakka, R., & ap Gwilym, O. (2018). The influence of rating levels and rating convergence on the spillover effects of sovereign credit actions. *Journal of International Money and Finance, 85*, 40-57. https://doi.org/10.1016/j.jimonfin.2018.03.005

2. Adelino, M., & Ferreira, M. A. (2016). Bank ratings and lending supply: Evidence from sovereign downgrades. *The Review of Financial Studies, 29*(7), 1709-1746. https://doi.org/10.1093/rfs/hhw004

3. Afonso, A. (2003). Understanding the determinants of sovereign debt ratings: Evidence for the two leading agencies. *Journal of Economics and Finance, 27*(1), 56-74. https://doi.org/10.1007/BF02751590

4. Afonso, A., Gomes, P., & Rother, P. (2011). Short- and long-run determinants of sovereign debt credit ratings. *International Journal of Finance & Economics, 16*(1), 1-15. https://doi.org/10.1002/ijfe.416

5. Afonso, A., Furceri, D., & Gomes, P. (2012). Sovereign credit ratings and financial markets linkages: Application to European data. *Journal of International Money and Finance, 31*(3), 606-638. https://doi.org/10.1016/j.jimonfin.2012.01.016

6. Aizenman, J., Binici, M., & Hutchison, M. (2013). Credit ratings and the pricing of sovereign debt during the euro crisis. *Oxford Review of Economic Policy, 29*(3), 582-609. https://doi.org/10.1093/oxrep/grt036

7. Athey, S., & Imbens, G. W. (2017). The state of applied econometrics: Causality and policy evaluation. *Journal of Economic Perspectives, 31*(2), 3-32. https://doi.org/10.1257/jep.31.2.3

8. Athey, S., & Imbens, G. W. (2019). Machine learning methods that economists should know about. *Annual Review of Economics, 11*, 685-725. https://doi.org/10.1146/annurev-economics-080217-053433

9. Bar-Isaac, H., & Shapiro, J. (2013). Ratings quality over the business cycle. *Journal of Financial Economics, 108*(1), 62-78. https://doi.org/10.1016/j.jfineco.2012.11.004

10. Becker, B., & Milbourn, T. (2011). How did increased competition affect credit ratings? *Journal of Financial Economics, 101*(3), 493-514. https://doi.org/10.1016/j.jfineco.2011.03.012

11. Bissoondoyal-Bheenick, E. (2005). An analysis of the determinants of sovereign ratings. *Global Finance Journal, 15*(3), 251-280. https://doi.org/10.1016/j.gfj.2004.03.004

12. Bolton, P., Freixas, X., & Shapiro, J. (2012). The credit ratings game. *The Journal of Finance, 67*(1), 85-111. https://doi.org/10.1111/j.1540-6261.2011.01708.x

13. Boot, A. W. A., Milbourn, T. T., & Schmeits, A. (2006). Credit ratings as coordination mechanisms. *The Review of Financial Studies, 19*(1), 81-118. https://doi.org/10.1093/rfs/hhj009

14. Breiman, L. (2001). Random forests. *Machine Learning, 45*(1), 5-32. https://doi.org/10.1023/A:1010933404324

15. Cantor, R., & Packer, F. (1996). Determinants and impact of sovereign credit ratings. *The Journal of Fixed Income, 6*(3), 76-91. https://doi.org/10.3905/jfi.1996.408185

16. Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 785-794). https://doi.org/10.1145/2939672.2939785

17. Chen, Z., Matousek, R., Steward, C., & Webb, R. (2019). Do rating agencies exhibit herding behaviour? Evidence from sovereign ratings. *International Review of Financial Analysis, 64*, 57-70. https://doi.org/10.1016/j.irfa.2019.05.003

18. Cortes, C., & Vapnik, V. (1995). Support-vector networks. *Machine Learning, 20*(3), 273-297. https://doi.org/10.1023/A:1022627411411

19. de Haan, L., & Vermeulen, R. (2021). Sovereign debt ratings and the country composition of cross-border holdings of euro area sovereign debt. *Journal of International Money and Finance, 119*, 102473. https://doi.org/10.1016/j.jimonfin.2021.102473

20. De Moor, L., Luitel, P., Sercu, P., & Vanpée, R. (2018). Subjectivity in sovereign credit ratings. *Journal of Banking & Finance, 88*, 366-392. https://doi.org/10.1016/j.jbankfin.2017.12.014

21. European Central Bank. (2026). Long-term interest rate statistics. ECB Data Portal. https://data.ecb.europa.eu/methodology/long-term-interest-rate-statistics

22. Eurostat. (2026). Exchange and interest rates: information and data. https://ec.europa.eu/eurostat/web/exchange-and-interest-rates/information-data

23. Ferri, G., Liu, L.-G., & Stiglitz, J. E. (1999). The procyclical role of rating agencies: Evidence from the East Asian crisis. *Economic Notes, 28*(3), 335-355. https://doi.org/10.1111/1468-0300.00016

24. Fuchs, A., & Gehring, K. (2017). The home bias in sovereign ratings. *Journal of the European Economic Association, 15*(6), 1386-1423. https://doi.org/10.1093/jeea/jvx009

25. Friedman, J. H. (2001). Greedy function approximation: A gradient boosting machine. *The Annals of Statistics, 29*(5), 1189-1232. https://doi.org/10.1214/aos/1013203451

26. Gande, A., & Parsley, D. C. (2005). News spillovers in the sovereign debt market. *Journal of Financial Economics, 75*(3), 691-734. https://doi.org/10.1016/j.jfineco.2003.11.003

27. Golbayani, P., Florescu, I., & Chatterjee, R. (2020). A comparative study of forecasting corporate credit ratings using neural networks, support vector machines, and decision trees. *The North American Journal of Economics and Finance, 54*, 101251. https://doi.org/10.1016/j.najef.2020.101251

28. Gu, S., Kelly, B., & Xiu, D. (2020). Empirical asset pricing via machine learning. *The Review of Financial Studies, 33*(5), 2223-2273. https://doi.org/10.1093/rfs/hhaa009

29. Hashimoto, R., Miura, K., & Yoshizaki, Y. (2023). Application of machine learning to a credit rating classification model. Bank of Japan Working Paper No. 23-E-6.

30. Hilscher, J., & Nosbusch, Y. (2010). Macroeconomic fundamentals and the pricing of sovereign debt. *Review of Finance, 14*(2), 235-262. https://doi.org/10.1093/rof/rfq005

31. Huang, Z., Chen, H., Hsu, C.-J., Chen, W.-H., & Wu, S. (2004). Credit rating analysis with support vector machines and neural networks: a market comparative study. *Decision Support Systems, 37*(4), 543-558. https://doi.org/10.1016/S0167-9236(03)00086-1

32. International Monetary Fund, Fiscal Affairs Department. (2025). Fiscal rules dataset: 1985-2024 update.

33. Kaminsky, G., & Schmukler, S. L. (2002). Emerging market instability: Do sovereign ratings affect country risk and stock returns? *The World Bank Economic Review, 16*(2), 171-195. https://doi.org/10.1093/wber/16.2.171

34. Kim, K. S. (2005). Predicting bond ratings using publicly available information. *Expert Systems with Applications, 29*(1), 75-81. https://doi.org/10.1016/j.eswa.2005.01.007

35. Kisgen, D. J., & Strahan, P. E. (2010). Do regulations based on credit ratings affect a firm's cost of capital? *The Review of Financial Studies, 23*(12), 4324-4347. https://doi.org/10.1093/rfs/hhq077

36. Kleinberg, J., Ludwig, J., Mullainathan, S., & Obermeyer, Z. (2015). Prediction policy problems. *American Economic Review, 105*(5), 491-495. https://doi.org/10.1257/aer.p20151023

37. Lundberg, S. M., & Lee, S.-I. (2017). A unified approach to interpreting model predictions. *Advances in Neural Information Processing Systems, 30*.

38. Lundberg, S. M., Erion, G., Chen, H., DeGrave, A., Prutkin, J. M., Nair, B., Katz, R., Himmelfarb, J., Bansal, N., & Lee, S.-I. (2020). From local explanations to global understanding with explainable AI for trees. *Nature Machine Intelligence, 2*, 56-67. https://doi.org/10.1038/s42256-019-0138-9

39. Mathis, J., McAndrews, J., & Rochet, J.-C. (2009). Rating the raters: Are reputation concerns powerful enough to discipline rating agencies? *Journal of Monetary Economics, 56*(5), 657-674. https://doi.org/10.1016/j.jmoneco.2009.04.004

40. Molnar, C. (2022). *Interpretable machine learning: A guide for making black box models explainable* (2nd ed.). https://christophm.github.io/interpretable-ml-book/

41. Mullainathan, S., & Spiess, J. (2017). Machine learning: An applied econometric approach. *Journal of Economic Perspectives, 31*(2), 87-106. https://doi.org/10.1257/jep.31.2.87

42. Opp, C. C., Opp, M. M., & Harris, M. (2013). Rating agencies in the face of regulation. *Journal of Financial Economics, 108*(1), 46-61. https://doi.org/10.1016/j.jfineco.2012.10.011

43. Reusens, P., & Croux, C. (2017). Sovereign credit rating determinants: A comparison before and after the European debt crisis. *Journal of Banking & Finance, 77*, 108-121. https://doi.org/10.1016/j.jbankfin.2017.01.006

44. Rudin, C. (2019). Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence, 1*(5), 206-215. https://doi.org/10.1038/s42256-019-0048-x

45. Rumelhart, D. E., Hinton, G. E., & Williams, R. J. (1986). Learning representations by back-propagating errors. *Nature, 323*(6088), 533-536. https://doi.org/10.1038/323533a0

46. Shapley, L. S. (1953). A value for n-person games. In H. W. Kuhn & A. W. Tucker (Eds.), *Contributions to the Theory of Games II* (pp. 307–317). Princeton University Press.

47. Skreta, V., & Veldkamp, L. (2009). Ratings shopping and asset complexity: A theory of ratings inflation. *Journal of Monetary Economics, 56*(5), 678-695. https://doi.org/10.1016/j.jmoneco.2009.04.006

48. Varian, H. R. (2014). Big data: New tricks for econometrics. *Journal of Economic Perspectives, 28*(2), 3-28. https://doi.org/10.1257/jep.28.2.3

Appendix

Table A1. Quantitative rating scale.

	Ordinal scale	*S&P*	*Moody's*	*Fitch*
Highest quality	21	AAA	Aaa	AAA
High quality	20	AA+	Aa1	AA+
	19	AA	Aa2	AA
	18	AA-	Aa3	AA-
Strong payment capacity	17	A+	A1	A+
	16	A	A2	A
	15	A-	A3	A-
Adequate payment capacity	14	BBB+	Baa1	BBB+
	13	BBB	Baa2	BBB
	12	BBB-	Baa3	BBB-
Likely to fulfil obligations. ongoing uncertainty	11	BB+	Ba1	BB+
	10	BB	Ba2	BB
	9	BB-	Ba3	BB-
High credit risk	8	B+	B1	B+
	7	B	B2	B
	6	B-	B3	B-
Very high credit risk	5	CCC+	Caa1	CCC+
	4	CCC	Caa2	CCC
	3	CCC-	Caa3	CCC-
Near default with possibility of recovery	2	CC	Ca	CC
	1	C	C	C
Default	0	SD/D		DDD/DD/D

Table A2. Correlation matrix.

	Inflation	*Debt-to-GDP*	*Current account*	*Unemployment*	*Budget balance rule*	*Output gap*	*Old-age dependency*	*Revenue capacity*	*RM*	*RSP*	*RF*	*RRAVG*
Inflation	1											
Debt-to-GDP	-0.279	1										
Current account	-0.264	0.029	1									
Unemployment	-0.175	0.325	-0.232	1								
Budget balance rule	-0.283	0.231	0.242	-0.212	1							
Output gap	0.291	-0.393	-0.139	-0.542	-0.031	1						
Old-age dependency	-0.067	0.495	0.095	0.035	0.239	-0.154	1					
Revenue capacity	-0.257	0.350	0.372	-0.115	0.205	-0.115	0.328	1				
RM	-0.206	-0.291	0.342	-0.464	0.204	0.328	-0.234	0.399	1			
RSP	-0.213	-0.256	0.376	-0.457	0.184	0.301	-0.253	0.411	0.954	1		
RF	-0.222	-0.277	0.355	-0.449	0.182	0.304	-0.272	0.384	0.968	0.971	1	
RRAVG	-0.216	-0.278	0.362	-0.464	0.194	0.316	-0.255	0.403	0.987	0.986	0.991	1

Table A3 – Robustness Results for ordered probit with contemporaneous and one-year lag explanatory variables, with and without fixed effects

	No-fixed effects			*Fixed effects*					
Explanatory Variable lags	*One-lag*			*Contemporaneous*			*One-lag*		
Agencies	Fitch	Moody's	S&P	Fitch	Moody's	S&P	Fitch	Moody's	S&P
INFLATION	-0.035***	-0.048***	-0.003***	-0.040***	-0.105***	-0.049***	-0.044***	-0.070***	-0.002*
	(0.006)	(0.010)	(0.001)	(0.008)	(0.016)	(0.008)	(0.008)	(0.014)	(0.001)
DEBT	-0.010***	-0.013***	-0.009***	-0.060***	-0.076***	-0.031***	-0.050***	-0.059***	-0.028***
	(0.002)	(0.002)	(0.001)	(0.005)	(0.006)	(0.005)	(0.005)	(0.005)	(0.005)
CURRENT	0.071***	0.059***	0.081***	-0.006	-0.002	-0.011	0.029**	0.030**	0.023*
	(0.008)	(0.009)	(0.008)	(0.013)	(0.014)	(0.013)	(0.013)	(0.014)	(0.013)
UNEMP	-0.053***	-0.061***	-0.059***	-0.168***	-0.190***	-0.269***	-0.153***	-0.171***	-0.189***
	(0.012)	(0.012)	(0.012)	(0.026)	(0.027)	(0.026)	(0.025)	(0.026)	(0.025)
FR_BBALANCE	0.656***	0.736***	0.758***	1.019***	1.314***	0.706***	0.877***	1.454***	0.983***
	(0.141)	(0.139)	(0.124)	(0.220)	(0.228)	(0.203)	(0.210)	(0.219)	(0.197)
OUTPUTGAP	0.077***	0.086***	0.072***	-0.010	0.002	-0.034	0.003	-0.014	-0.003
	(0.015)	(0.015)	(0.015)	(0.024)	(0.025)	(0.023)	(0.023)	(0.024)	(0.022)
OLDRATIO	-0.103***	-0.097***	-0.107***	-0.065**	0.031	0.013	-0.029	0.050	0.099***
	(0.009)	(0.009)	(0.009)	(0.031)	(0.034)	(0.031)	(0.031)	(0.032)	(0.029)
TAXBURDEN	0.123***	0.146***	0.135***	0.019	0.121***	0.006	0.028	0.101***	0.035
	(0.008)	(0.009)	(0.009)	(0.024)	(0.026)	(0.023)	(0.024)	(0.024)	(0.023)
cut point1	-3.438***	-3.035***	-2.563***	-20.885***	-15.985***	-21.064***	-17.544***	-12.977***	-12.863***
	(0.538)	(0.644)	(0.473)	(1.691)	(1.741)	(1.742)	(1.587)	(1.635)	(1.484)
cut point2	-3.238***	-2.631***	-2.022***	-20.604***	-15.708***	-20.211***	-17.326***	-12.674***	-12.129***
	(0.511)	(0.555)	(0.425)	(1.680)	(1.726)	(1.713)	(1.579)	(1.618)	(1.458)
cut point3	-2.568***	-1.957***	-1.705***	-19.556***	-15.033***	-19.635***	-16.475***	-12.087***	-11.655***
	(0.455)	(0.478)	(0.408)	(1.636)	(1.710)	(1.697)	(1.557)	(1.609)	(1.447)
cut point4	-2.426***	-1.858***	-1.566***	-19.327***	-14.884***	-19.391***	-16.283***	-11.969***	-11.443***
	(0.448)	(0.471)	(0.403)	(1.630)	(1.707)	(1.693)	(1.555)	(1.607)	(1.445)
cut point5	-2.251***	-1.596***	-1.384***	-19.021***	-14.454***	-19.024***	-16.025***	-11.638***	-11.140***
	(0.441)	(0.454)	(0.399)	(1.621)	(1.697)	(1.687)	(1.551)	(1.602)	(1.442)
cut point6	-2.021***	-1.400***	-1.031***	-18.565***	-14.089***	-18.251***	-15.600***	-11.374***	-10.522***
	(0.432)	(0.443)	(0.393)	(1.610)	(1.690)	(1.675)	(1.540)	(1.598)	(1.438)
cut point7	-1.881***	-1.302***	-0.728*	-18.287***	-13.899***	-17.573***	-15.318***	-11.233***	-9.946***
	(0.428)	(0.439)	(0.389)	(1.606)	(1.688)	(1.668)	(1.535)	(1.597)	(1.436)
cut point8	-1.543***	-0.986**	-0.263	-17.659***	-13.016***	-16.446***	-14.678***	-10.631***	-8.922***
	(0.421)	(0.429)	(0.386)	(1.599)	(1.673)	(1.655)	(1.527)	(1.588)	(1.430)
cut point9	-1.016**	-0.829*	0.050	-16.542***	-12.499***	-15.641***	-13.596***	-10.280***	-8.161***
	(0.415)	(0.425)	(0.385)	(1.586)	(1.658)	(1.647)	(1.517)	(1.582)	(1.427)
cut point10	-0.566	-0.603	0.336	-15.356***	-11.880***	-14.867***	-12.489***	-9.793***	-7.510***
	(0.412)	(0.421)	(0.383)	(1.573)	(1.645)	(1.635)	(1.510)	(1.574)	(1.419)
cut point11	-0.190	-0.253	0.749**	-14.406***	-10.914***	-14.001***	-11.646***	-8.981***	-6.730***
	(0.410)	(0.417)	(0.381)	(1.561)	(1.628)	(1.623)	(1.502)	(1.563)	(1.411)
cut point12	0.244	0.346	1.157***	-13.446***	-9.579***	-13.201***	-10.744***	-7.753***	-5.965***
	(0.407)	(0.413)	(0.380)	(1.547)	(1.617)	(1.616)	(1.490)	(1.553)	(1.404)
cut point13	0.567	0.623	1.478***	-12.716***	-8.961***	-12.473***	-10.055***	-7.179***	-5.245***
	(0.406)	(0.411)	(0.381)	(1.537)	(1.613)	(1.608)	(1.481)	(1.551)	(1.397)
cut point14	0.942**	0.854**	1.680***	-11.910***	-8.419***	-11.938***	-9.247***	-6.702***	-4.711***
	(0.407)	(0.409)	(0.384)	(1.530)	(1.608)	(1.606)	(1.474)	(1.548)	(1.397)
cut point15	1.284***	1.169***	2.216***	-11.061***	-7.768***	-10.237***	-8.473***	-6.068***	-3.067**
	(0.411)	(0.407)	(0.391)	(1.530)	(1.600)	(1.599)	(1.473)	(1.541)	(1.401)
cut point16	1.729***	1.660***	2.612***	-9.852***	-6.629***	-8.943***	-7.290***	-5.024***	-1.775
	(0.416)	(0.408)	(0.395)	(1.526)	(1.591)	(1.598)	(1.473)	(1.533)	(1.407)
cut point17	2.007***	2.028***		-9.113***	-5.596***		-6.519***	-4.088***	
	(0.418)	(0.413)		(1.519)	(1.585)		(1.469)	(1.527)	
cut point18		2.258***			-4.905***			-3.400**	
		(0.417)			(1.588)			(1.528)	
cut point19		2.637***			-3.418**			-2.027	
		(0.423)			(1.591)			(1.533)	
cut point20		2.976***			-2.103			-0.698	
		(0.426)			(1.602)			(1.548)	
Observations	733	728	755	724	710	746	733	728	755
$Pseudo\ R^2$	0.195	0.204	0.198	0.495	0.527	0.513	0.479	0.499	0.487

Robust standard errors clustered by country in parentheses. ***, **, and * denote statistical significance at the 1%, 5%, and 10% levels, respectively.

Table A4. Market-pricing robustness checks.

Variable	*Year-end Bund spread*	*First-difference spread*	*EU yield level*	*EU lagged yield level*
Shadow average rating	-49.689**		-0.514***	-0.561*
	(20.812)		(0.125)	(0.306)
Average agency deviation	-41.990***		-0.419***	-0.334***
	(6.605)		(0.101)	(0.072)
Change in shadow average rating		-67.214*		
		(35.675)		
Change in average deviation		-72.470**		
		(30.998)		
Observations	361	359	621	631
Countries / clusters	19	19	26	26
R-squared	0.606	0.525	0.811	0.763
Country FE	Yes	No, first differences	Yes	Yes
Year FE	Yes	Yes	Yes	Yes
Controls	No	Yes, differences	No	Yes, lagged
Yield sample years	1999-2024	1999-2024	1995-2024	1995-2024

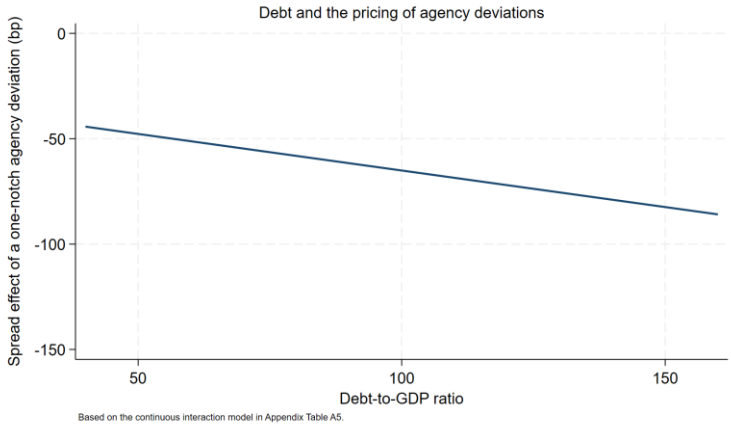
Notes: Column B uses year-end Bund spreads; column C uses first differences of annual-average Bund spreads; columns D and E use annual average 10-year sovereign yields in levels for the broader EU sample. Robust standard errors clustered by country in parentheses. ***, **, and * denote statistical significance at the 1%, 5%, and 10% levels, respectively.

Table A5. Debt-burden heterogeneity in the pricing of agency deviations.

Variable	*Debt interaction*	*High-debt state*
Shadow average rating	-61.238**	-48.842**
	(28.671)	(18.963)
Average agency deviation	-56.349***	-44.133***
	(18.363)	(8.834)
Debt burden (centered, per 10 p.p. of GDP)	-23.509	
	(19.496)	
Deviation x debt burden	-3.471***	
	(0.621)	
High-debt state (debt >= 100% of GDP)		-47.782
		(37.594)
Deviation x high-debt state		-20.340***
		(6.829)
Observations	361	361
Countries	19	19
Within R-squared	0.675	0.655
Country FE	Yes	Yes
Year FE	Yes	Yes
Controls	No	No
Sample years	1999-2024	1999-2024
Sample	Euro-area sovereign-years with Bund spreads	Euro-area sovereign-years with Bund spreads

Notes: Column B uses the continuous debt burden centered at the mean of the euro-area Bund-spread sample and scaled so that one unit equals 10 percentage points of GDP. Column C uses a discrete high-debt state equal to one when debt exceeds 100% of GDP. The interaction terms test whether the pricing effect of agency deviations changes with sovereign indebtedness. Robust standard errors clustered by country in parentheses. ***, **, and * denote statistical significance at the 1%, 5%, and 10% levels, respectively.

Figure A1. Debt and the pricing of agency deviations.



Notes: The line plots the implied spread effect of a one-notch agency deviation from the continuous interaction model in Appendix Table A5, with a 95% confidence band. More negative values indicate a stronger spread response to agency deviations.

Table A6. Stress-period comparison in the pricing of agency deviations, 1999-2024.

Variable	*Crisis-period interaction*
Shadow average rating	-50.512**
	(18.552)
Average agency deviation	-47.412***
	(10.596)
Deviation x crisis period (2008-2013)	-12.426
	(12.073)
Obs.	361
Within R-squared	0.639
Country FE	Yes
Year FE	Yes
Controls	No
Sample	Euro-area sovereign-years with Bund spreads

The crisis-period interaction equals one in 2008-2013, spanning the global financial crisis and the euro-area sovereign debt crisis. The interaction term tests whether the pricing effect of agency deviations differs in stress years. Robust standard errors are clustered by country. ***, **, and * denote statistical significance at the 1%, 5%, and 10% levels, respectively.

Table A7. Sovereign Bund spreads, fundamentals-implied ratings, and agency deviations, 1999-2024 using Machine Learning.

Variable	*Observed rating*	*Shadow rating*	*Decomposition*	*Lagged decomposition*	*Asymmetric deviations*
Observed average rating	-51.570***				
	(13.843)				
ML Shadow average rating		-52.785***	-52.475***	-66.662***	-52.523***
		(11.910)	(12.010)	(21.005)	(12.335)
ML Average agency deviation			-46.204	-58.920	
			(38.380)	(36.952)	
ML Positive deviation					-38.140
					(120.016)
ML Negative deviation					-54.324
					(88.521)

Notes: Dependent variable is the annual average 10-year sovereign yield spread versus Germany, in basis points. Robust standard errors clustered by country in parentheses. Better ratings should lower spreads, so negative coefficients indicate tighter funding conditions. ***, **, and * denote statistical significance at the 1%, 5%, and 10% levels, respectively. The adopted ML model is the XGBoost.