Predicting Extreme High-Yield Credit Regimes

Conrad Voigt

March 17, 2026

Abstract

This study develops a time-aware machine-learning framework for predicting extreme high-yield credit regimes. Using daily macroeconomic and market data, the analysis constructs a five-trading-day target that identifies extreme spread tightening (Risk-On) and widening (Risk-Off) episodes while excluding a central dead zone from model training. The feature set spans high-yield option-adjusted spreads, momentum, the Treasury yield curve, volatility, cross-asset interactions, and funding-stress proxies. Logistic Regression, Elastic Net, Random Forest, and XGBoost classifiers are evaluated using an expanding-window cross-validation scheme and a strictly out-of-sample test period beginning in 2022. The project emphasizes economically interpretable evaluation, including Spearman rank correlation and quintile-spread analysis, alongside conventional classification diagnostics.

Keywords: high-yield credit; regime classification; machine learning; option-adjusted spread; risk-on; risk-off; expanding-window cross-validation

1 Introduction

1.1 Background and Motivation

High-yield option-adjusted spreads (HY OAS) measure the incremental yield investors demand to hold speculative-grade corporate bonds — those rated BB+ or below by S&P — over equivalent-maturity U.S. Treasury securities, after adjusting for any embedded optionality in the bond structure (ICE BofA, via FRED series BAMLH0A0HYM2). When HY OAS widen sharply, they signal rising default expectations and deteriorating credit conditions across the speculative-grade universe; historically, such episodes have coincided with broad equity drawdowns and contractions in investor risk appetite. Conversely, rapid spread compression — *tightening* — marks a return to risk-seeking behavior, as investors reduce their required compensation for default exposure and rotate capital back into higher-yielding assets. These polar dynamics define what practitioners term the **risk-off** and **risk-on** regimes, respectively.

Credit spread dynamics are notoriously difficult to model from first principles. Collin-Dufresne, Goldstein & Martin (2001, *Journal of Finance*) demonstrate, using straight industrial bond quotes over 1988–1997, that the theoretical determinants of credit spread changes — interest rate levels, leverage ratios, equity volatility, and business climate proxies — collectively explain only approximately 25% of the variation in monthly spread changes. The residuals are highly cross-correlated and driven by a single dominant systematic factor that cannot be identified using conventional macroeconomic or financial variables; Collin-Dufresne et al. (2001) interpret this as evidence of corporate bond market segmentation driven by supply and demand shocks specific to the credit market itself. This empirical puzzle motivates a machine-learning approach: rather than constraining

the model to pre-specified functional forms, we allow flexible, non-linear classifiers to extract regime signals from a richer feature space spanning macro, cross-asset, and volatility dimensions.

1.2 Task Framing and Horizon Choice

Rather than forecasting the exact magnitude of spread moves — a task rendered particularly challenging by the large unexplained systematic component identified by Collin-Dufresne et al. (2001) — we frame the problem as **ordinal binary classification**: given today's macro and market signals, is the next 5 trading days likely to represent an *extreme* episode (top or bottom quintile of historical spread changes), and if so, in which direction?

The 5-day (one-week) prediction horizon is chosen deliberately. Jostova, Nikolova, Philipov & Stahel (2013, *Review of Financial Studies*) document significant price momentum in non-investment-grade corporate bonds over multi-week holding periods, with momentum profits concentrated among high-yield bonds — consistent with the slow diffusion of information characteristic of the credit market relative to equities. A single trading day is too short a window to constitute a reliable regime signal; a week captures the type of multi-session dislocation that triggers institutional rebalancing.

1.3 Economic Significance

For portfolio managers and risk teams, early identification of regime transitions has direct economic value. An accurate *Risk-Off* signal can inform credit underweight decisions, trigger macro hedges in CDS indices such as CDX HY (the primary liquid instrument for expressing views on high-yield default risk), or tighten position-level stop-losses ahead of a drawdown. A reliable *Risk-On* signal, conversely, can inform incremental credit overweights ahead of spread compression. To capture this practical dimension, we evaluate all models not only on standard classification metrics (accuracy, macro-averaged F1, ROC-AUC) but also on economic signal strength using Spearman rank correlation and quintile-spread analysis — measures directly interpretable in a portfolio management context.

1.4 Data and Methodology Overview

We source eight daily macro and market series from the Federal Reserve Economic Data (FRED) database and the S&P 500 total return index from `yfinance`, covering the period 1997–present. From these raw inputs we engineer 22 features across six groups: spread level indicators, momentum signals, yield-curve shape, volatility and regime composites, cross-asset interaction terms, and funding-stress proxies (detailed in the codebook). Each trading day is labeled as *Risk-Off* (extreme widening), *Risk-On* (extreme tightening), or a *dead zone* — moderate, uninformative moves excluded from model training to sharpen the signal. Four classifiers — Logistic Regression, Elastic Net, Random Forest, and XGBoost — are trained on the 1997–2021 window using time-aware expanding-window cross-validation, ensuring that no future information leaks into the training set at any fold boundary. Out-of-sample evaluation covers 2022–present (approximately 800 trading days), with dead-zone days *retained* in the test set to faithfully simulate real-world deployment conditions.

The **training window** spans 1997–2021, yielding approximately 6,400 raw trading days and approximately 3,000–4,000 labeled extreme-regime observations after dead-zone filtering. The

out-of-sample test window covers 2022–present (approximately 800 trading days, unfiltered). All feature construction uses only information available at or before time t, and the train/test split is strictly chronological with no shuffling or random sampling across time.

2 Data and Target Construction

The cells below load required libraries and authenticate with the FRED API. The S&P 500 index is fetched separately via yfinance.

2.1 Data Sources

All data are sourced from two providers: the Federal Reserve Economic Data (FRED) database, maintained by the Federal Reserve Bank of St. Louis, and `yfinance` for the S&P 500 index. Together these eight series form the raw input from which all 22 model features are subsequently engineered. Table 2.1 lists each series, its provider, FRED ticker or `yfinance` symbol, start date, and role in the model.

Source	Series	Code	Starts	Role
FRED	ICE BofA US HY OAS	`BAMLH0A0HYM2`	1996-12-31	**Target variable**
yfinance	S&P 500 Index	`^GSPC`	1928	Cross-asset regime signal
FRED	10-Year Treasury Yield	`DGS10`	1962	Rate environment
FRED	2-Year Treasury Yield	`DGS2`	1976	Short-rate / curve shape
FRED	CBOE VIX	`VIXCLS`	1990	Market fear gauge
FRED	St. Louis Financial Stress Index	`STLFSI4`	1993	Systemic stress composite
FRED	ICE BofA BB OAS	`BAMLH0A1HYBB`	1996-12-31	Credit quality spread
FRED	TED Spread (spliced)	`TEDRATE` → `SOFR - DTB3`	1986	Funding stress

Table 2.1: Raw data series, providers, and roles.

The HY OAS series (BAMLH0A0HYM2) is the market-capitalization-weighted average option-adjusted spread of all USD-denominated, below-investment-grade corporate bonds in the ICE BofA index universe — those rated BB+ or below on average across Moody's, S&P, and Fitch — with each constituent bond's OAS computed relative to a spot U.S. Treasury curve (Federal Reserve

Bank of St. Louis, FRED: BAMLH0A0HYM2). This series serves as both the raw target variable and as a direct model input in level and momentum form.

Why `yfinance` for the S&P 500? FRED's native `SP500` series begins only in January 2013, which is far too short for our 1997 training start date. The `yfinance ^GSPC` ticker provides dividend-adjusted daily closing prices back to 1928, making it the correct source for this project. Timezone metadata is stripped on ingestion and the series is reindexed to match the FRED business-day calendar.

2.2 TED Spread — Spliced Construction

The TED spread — classically defined as the spread between 3-month USD LIBOR and the 3-month U.S. Treasury bill yield — is a widely used proxy for short-term funding stress and counterparty risk in the interbank market. The FRED series `TEDRATE` was discontinued on April 28, 2023, when USD LIBOR panel settings were formally ceased as part of the global benchmark reform transition to the Secured Overnight Financing Rate (SOFR), which the Alternative Reference Rates Committee (ARRC) selected in 2017 as the preferred replacement for USD LIBOR (Federal Reserve Bank of New York, 2023). Rather than forward-filling the last available `TEDRATE` value beyond its discontinuation date — which would inject a stale, non-market reading into the 2023–present portion of the test set — we construct a continuous synthetic series by splicing two FRED series at the cutoff date:

Period	Series	Formula
1986–2023-04-28	`TEDRATE` (FRED)	LIBOR3m - T-Bill3m
2023-04-29–present	`SOFR` and `DTB3` (FRED)	SOFR_overnight - DTB3_3M

Table 2.2: TED spread splicing methodology.

Overnight SOFR differs from 3-month LIBOR in tenor — SOFR is a risk-free overnight repo rate, while LIBOR embedded a term credit-risk premium — but both capture the same economic concept: the incremental cost of unsecured short-term funding relative to the risk-free rate (Federal Reserve Bank of New York, 2023). The splice introduces a small level shift of approximately 10–25 basis points that is immaterial for a regime-classification task where only the direction and relative magnitude of funding stress matter. Critically, the splice affects only the post-April 2023 portion of the **test set**; the model trains entirely on pre-splice data from 1997 to 2021.

2.3 Data Ingestion

The code block below implements the data ingestion pipeline. A fast path checks for a locally cached file (`data/raw_data.csv`); if found, it is loaded directly. If not, the slow path fetches all series live from FRED via `pandas_datareader` and the S&P 500 from `yfinance`, aligns all series to a common FRED business-day index using `pd.bdate_range`, applies the TED splice at the `TED_CUTOFF` date of April 28, 2023, applies forward-fill imputation to the St. Louis Financial Stress Index (weekly release, maximum gap of 5 business days) and the TED spread (holiday gaps only, `limit=3` to avoid carrying stale values), and drops the SOFR and DTB3 construction helpers from the final frame. All FRED series are publicly accessible through the Federal Reserve Bank of St. Louis data portal at

https://fred.stlouisfed.org; S&P 500 daily price data are retrieved via the `yfinance` Python library, documented at https://pypi.org/project/yfinance

2.4 Missing Data

The ingested frame spans 7,619 business days from January 1, 1997 to March 16, 2026. Before any imputation or row-dropping, per-column NaN counts are as follows:

Column	NaN Count	Source of Gap
`HY_OAS`	95	Exchange holidays, month-end rebalancing days
`DGS10`	317	Federal holidays; weekend carry-over
`DGS2`	317	Same as DGS10
`VIX`	246	Exchange holidays
`STLFSI4`	2	Weekly release; residual after forward-fill
`BB_OAS`	95	Same as HY_OAS
`TED`	328	Holiday gaps; LIBOR/SOFR transition window
`SP500`	274	Exchange holidays

Table 2.3: Pre-drop NaN counts by series.

Missing values arise from three distinct structural sources. First, `STLFSI4` is a weekly release (published every Thursday by the Federal Reserve Bank of St. Louis); business days between weekly publications are forward-filled without a carry limit, since the maximum gap is exactly 5 business days and the index is a slow-moving composite. Second, the TED spread exhibits brief 1–3 business day holiday gaps that are forward-filled with `limit=3` to prevent stale values from propagating across longer periods. Third, the remaining NaNs — concentrated in the Treasury yield and equity series — reflect exchange and Federal holiday closures where no market prices are recorded. After a final `dropna()` to remove any rows where at least one series remains unobserved (predominantly early-sample warmup rows before all series are simultaneously available), the working dataset contains **6,972 rows spanning January 3, 1997 to March 13, 2026**, representing a retention rate of 91.5%. The 647 dropped rows (8.5%) are concentrated at the beginning of the sample and do not represent a systematic loss of any particular market regime. The first five rows of the cleaned frame are shown below, confirming that all eight series are populated from the first observation date.

date	HY_OAS	DGS10	DGS2	VIX	STLFSI4	BB_OAS	TED	SP500
1997-01-03	3.09	6.52	5.95	19.13	-0.3478	1.94	0.52	748.030029
1997-01-06	3.10	6.54	5.97	19.89	-0.3478	1.94	0.51	747.650024
1997-01-07	3.10	6.57	5.98	19.35	-0.3478	1.93	0.54	753.229980
1997-01-08	3.07	6.60	6.01	20.24	-0.3478	1.91	0.54	748.409973
1997-01-09	3.13	6.52	5.94	20.91	-0.3478	1.96	0.56	754.849976

2.5 Target Variable Construction — Rolling-Window Quantile Labeling

2.5.1 Label Definition

The target variable is constructed from the **forward-looking 5-day change in HY OAS**, denoted $\Delta\text{OAS}_{t,t+5}$. Each trading day is assigned one of three labels:

Table 2.4. Target variable label definitions.

Label	Value	Condition	Economic Interpretation
Risk-Off	1	$\Delta\text{OAS}_{t,t+5} \geq Q_{80,t}$	Extreme spread widening—rising default fears
Risk-On	0	$\Delta\text{OAS}_{t,t+5} \leq Q_{20,t}$	Extreme spread tightening—risk appetite returning
Dead Zone	-1	$Q_{20,t} < \Delta\text{OAS}_{t,t+5} < Q_{80,t}$	Moderate, directionally uninformative move

Here $Q_{80,t}$ and $Q_{20,t}$ denote the rolling 80th and 20th percentiles of **past** 5-day OAS changes, computed over a 756-day (3-year) backward-looking window with a 252-day warmup period. Dead-zone observations (label = -1) are excluded from model training but retained in the out-of-sample test set to simulate realistic deployment conditions, where the classifier must produce a prediction on every trading day regardless of the true regime.

2.5.2 Why Rolling Quantiles Rather Than Fixed Thresholds?

A fixed basis-point threshold for labeling extreme spread moves — for example, classifying any 5-day change exceeding ±10 bps as extreme — is inappropriate for a dataset spanning nearly three decades of materially different volatility regimes. During the 2008–2009 credit crisis, daily OAS moves of that magnitude were routine; in the low-volatility period of 2014–2019, such moves were rare enough to go months without occurring. A fixed threshold would therefore produce severe and time-varying class imbalances, with the training set dominated by crisis-era observations and nearly devoid of calm-period signals.

Rolling quantile thresholds address this directly by normalizing extreme labels to the *prevailing* volatility regime. This approach is conceptually aligned with the Peaks-over-Threshold (POT) framework from Extreme Value Theory, in which the exceedance threshold is chosen adaptively based on a trailing window of observations rather than fixed ex ante; Chavez-Demoulin, Davison & McNeil (2005, *Journal of the Royal Statistical Society: Series C*) demonstrate that time-adaptive quantile estimation substantially outperforms fixed-threshold methods for capturing tail risk in nonstationary financial series. In our implementation, the rolling window is set to 756 business days (3 years), balancing sufficient historical depth to estimate the 80th/20th percentiles reliably against responsiveness to regime shifts. The warmup period of 252 days (1 year) ensures that early-sample quantile estimates are not driven by fewer than one year of observations.

The output confirms that the rolling thresholds are remarkably stable across the full sample: the median rolling Q80 (Risk-Off threshold) is approximately +0.13 bps and the median rolling Q20 (Risk-On threshold) is approximately -0.15 bps, with the full ranges staying within ±0.3 bps. This stability is a desirable property — it confirms that the labeling scheme consistently selects the top and bottom quintile of contemporaneous moves rather than drifting with the level of the OAS series itself.

2.5.3 Why a Dead Zone?

The dead-zone design — excluding moderate observations from training — is motivated by the signal-to-noise properties of the classification task. Forcing a binary classifier to label every moderate-move day as either Risk-On or Risk-Off would require it to draw a decision boundary through the densest, most ambiguous region of the feature space. Aliaga-Díaz et al. (2022, *Journal of Portfolio Management*) argue that regime classifiers in fixed income markets achieve meaningfully higher out-of-sample accuracy when trained exclusively on unambiguous extreme observations and deployed on all observations at inference time. More formally, this approach is analogous to the selective prediction (or "abstention") framework in machine learning, in which a classifier is trained only on high-confidence examples and is allowed to abstain on low-confidence ones — a strategy that consistently improves precision on the predicted class at the cost of coverage (Geifman & El-Yaniv, 2017, *NeurIPS*). By reserving dead-zone days for the test set only, we obtain a sharper decision boundary during training while preserving an honest evaluation of real-world deployment performance.

Class balance in the resulting training set is near-ideal: the 5-day extreme dataset contains 2,721 observations (1,314 Risk-On, 1,407 Risk-Off), a Risk-On/Risk-Off ratio of approximately 0.93. This near-equal split is a direct consequence of the symmetric 20th/80th percentile labeling design and eliminates the need for oversampling techniques such as SMOTE — which have been shown to introduce synthetic interpolation artifacts in structured financial time series that can inflate in-sample metrics while degrading out-of-sample performance (Branco, Torgo & Ribeiro, 2016, *ACM Computing Surveys*).

2.5.4 Output Summary

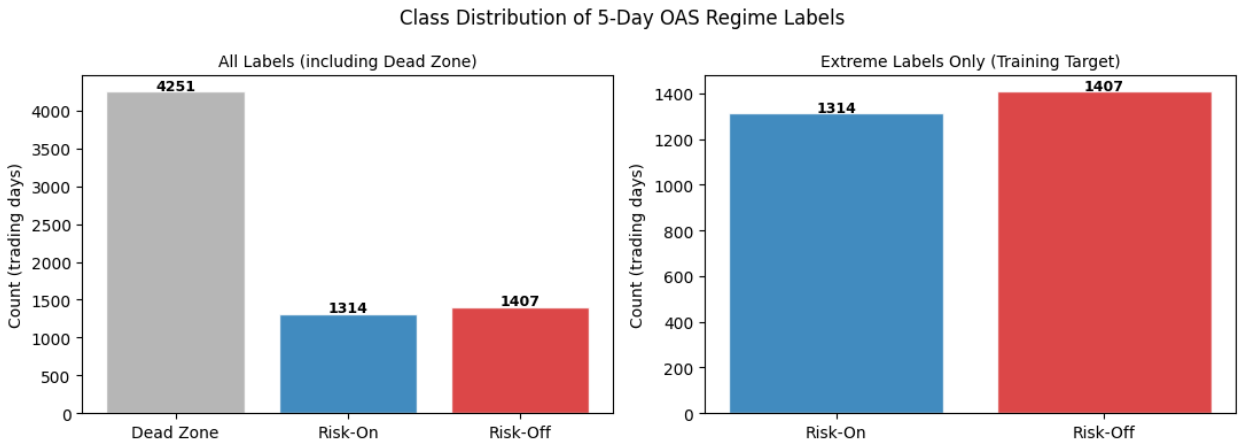
The near-zero median thresholds (±0.13–0.15 bps) reflect the mean-reverting nature of HY OAS at short horizons: the unconditional 5-day expected change is close to zero, so the 20th and 80th percentiles are nearly symmetric around zero. The drop of 4,251 rows — representing warmup and dead-zone observations — leaves a training-eligible sample of 2,721 extreme-regime days spanning approximately 28 years of credit market history, including multiple full credit cycles (1998 Russia/LTCM, 2001–2002 dot-com/Enron, 2008–2009 GFC, 2020 COVID, 2022 rate shock).

3 Preliminary Exploratory Data Analysis

The following three figures are produced directly from the raw cleaned dataset and the target labels constructed in Section 2, prior to any feature engineering. The goals here are to validate the target labeling design, characterize the unconditional class distribution, and establish the historical cross-asset context that motivates the feature groups defined in Section 4.

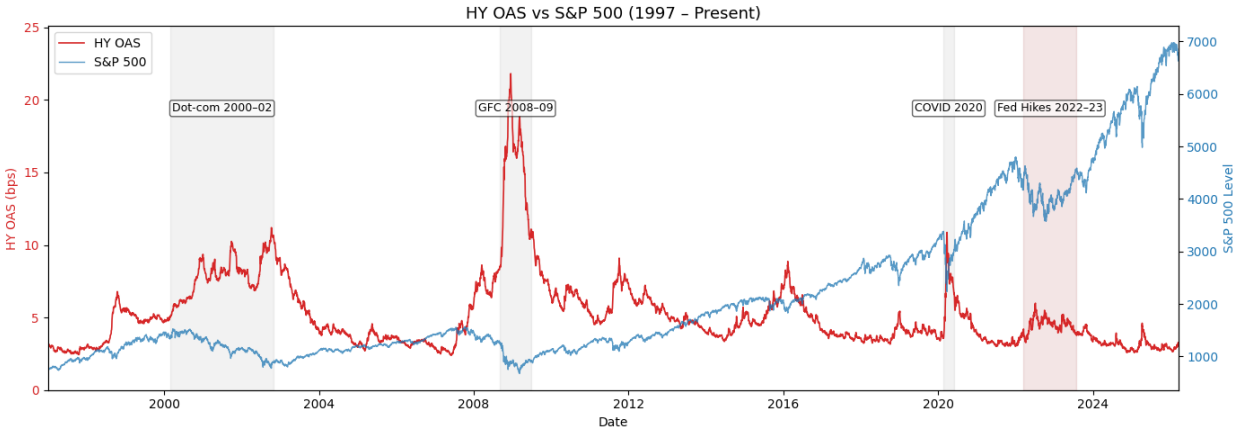
3.1 Target Label Distribution

The above diagrams present the unconditional class distribution across two views: the full label set (including dead-zone observations) and the extreme-only training target. The dead zone accounts for 4,192 trading days — roughly 61% of all observations — which is precisely by design: the rolling



20th/80th percentile scheme ensures that only the most extreme quintile moves in each direction are designated as actionable signals. The dominant share of dead-zone days confirms that the typical trading day in the HY market carries no reliable directional information, justifying their exclusion from model training. Among extreme-regime days, Risk-Off (1,407) slightly outnumbers Risk-On (1,314) — a 1.07:1 ratio that reflects an asymmetry inherent to credit markets: spreads tend to spike sharply in discrete shock events but compress gradually as conditions normalize. Because the test set retains dead-zone days to simulate real-world deployment, standard accuracy would be dominated by trivial dead-zone behavior and is therefore uninformative. We adopt **ROC-AUC** as the primary cross-validation metric — which evaluates ranking performance across the full probability threshold range and is invariant to class prevalence — and **Spearman rank correlation** as the primary out-of-sample economic metric, for the same reason.

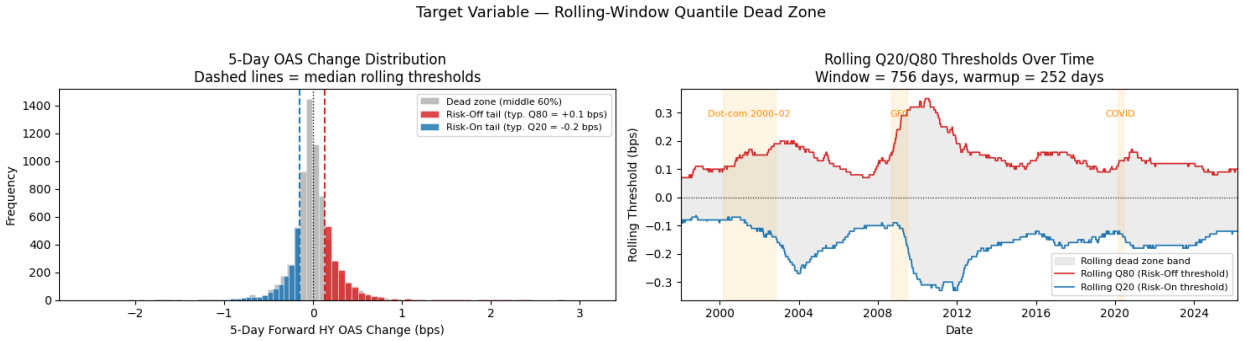
3.2 HY OAS and S&P 500 Over Time



This graph plots HY OAS (left axis) against the S&P 500 (right axis) over the full sample, with the four major stress episodes shaded. The sharp negative co-movement during each episode — dot-com (OAS +650 bps, SPX -49%), GFC (OAS +1,650 bps, SPX -55%), COVID (OAS +760 bps in three weeks, SPX -34%), and the 2022–23 rate-hike cycle (OAS +300 bps, SPX -25%) — provides the primary visual motivation for including equity returns as cross-asset features alongside

OAS-based inputs. Two structural properties are apparent: spread regimes are persistent (once OAS widens, it remains elevated for months), motivating level features alongside momentum features; and the magnitude of moves is regime-dependent, motivating the rolling z-score and realized volatility features introduced in Section 4. The 2022–23 episode, shaded in dark red, constitutes the primary out-of-sample test window and represents a distinct shock mechanism — monetary tightening — from the liquidity-driven crises that dominate the training set, providing a genuine test of model generalization.

3.3 Target Variable Distribution and Rolling Thresholds



This figure presents two complementary views of the target construction defined in Section 2.5. The left panel plots the empirical distribution of 5-day forward OAS changes: it is sharply leptokurtic and right-skewed, with heavy tails that substantially exceed Gaussian predictions. A fixed basis-point threshold would be poorly calibrated across regimes — too tight in calm periods, too loose in volatile ones — directly motivating the rolling quantile design. The right panel confirms that the adaptive thresholds behave as intended: the Q80 threshold expands dramatically during the GFC and COVID shock, correctly demanding genuinely exceptional moves for a Risk-Off label even during already-turbulent regimes, then contracts as conditions normalize. The symmetry between the Q20 and Q80 bands confirms that the labeling scheme consistently selects the top and bottom quintile of each epoch's own distribution across the full 1997–2026 sample.

4 Feature Engineering

4.1 Overview

We construct **24 features** across six thematic groups, enumerated in Table 3.1. Every feature is constructed exclusively from information available at or before time t; no forward-looking inputs of any kind are used. This strict no-look-ahead constraint is enforced at the rolling-window level: every moving average, volatility estimate, z-score, and rolling quantile is computed over a backward-looking window ending at $t-1$ or t (close-of-day), never beyond.

Group	#	Features	Core Economic Rationale
A — Level	4	HY_OAS, VIX, DGS10, STLFSI4	Absolute credit-cycle and fear positioning
B — Momentum	9	1d/5d/20d Δ HY_OAS, Δ VIX, SP500 log-ret	Trend, acceleration, and cross-asset flow signals
C — Curve	2	2s10s slope, BB-HY quality slope	Macro backdrop and intra-credit risk appetite
D — Regime	2	20d realized OAS vol, 60d OAS z-score	Turbulence state and statistical extremity
E — Interactions	4	OAS×SPX, VIX×SPX, high-VIX dummy, high-spread dummy	Nonlinear joint stress signals
F — Funding	3	TED (bps), DGS10 rolling vol (MOVE proxy), TED×VIX	Money-market and rates-market stress

Table 3.1: Feature groups, counts, and economic rationale.

4.2 Group A — Level and State

Absolute spread and volatility levels capture where the market is positioned in the credit cycle at time t, irrespective of the direction of recent moves. This information is economically non-redundant with momentum features: a 10 bps widening move carries very different implications when the HY OAS level is 300 bps (normal conditions) versus 700 bps (deep distress). Brunnermeier & Pedersen (2009, *Review of Financial Studies*) formalize this asymmetry through their liquidity spiral mechanism: once funding constraints bind, margin calls force leveraged investors to sell, which widens spreads further, which tightens margins again, creating a self-reinforcing feedback loop that is activated only above a threshold level of market stress. Incorporating the level of HY OAS, VIX, and the St. Louis Financial Stress Index (STLFSI4) directly gives the model access to the regime-conditioning information it needs to determine whether such spirals are likely to be active.

4.3 Group B — Momentum

Multi-horizon momentum features capture whether credit stress is accelerating, stalling, or reversing. We include 1-day, 5-day, and 20-day changes in HY OAS and VIX alongside the corresponding S&P 500 log-returns. The three horizons serve distinct roles: the 1-day change captures mean-reversion dynamics and overnight gap effects; the 5-day change tracks weekly trend persistence in line with the institutional rebalancing cycle discussed in Section 1; and the 20-day change reflects the medium-term macro momentum that is slow to reverse in credit markets. The inclusion of equity momentum alongside credit momentum is motivated by the cross-asset co-movement documented by

Collin-Dufresne, Goldstein & Martin (2001, *Journal of Finance*): credit spread changes and equity returns share a common systematic factor, so contemporaneous equity momentum provides an independent incremental signal about the direction of spread pressure beyond what OAS momentum alone conveys.

4.4 Group C — Yield Curve and Credit Quality Slope

The 2s10s slope (10-year minus 2-year Treasury yield) is included as a macroeconomic state variable. Estrella & Hardouvelis (1991, *Journal of Finance*) establish that the slope of the yield curve is one of the most robust predictors of real economic activity and recession risk available in real time, with an inverted curve historically preceding recessions by 6–18 months and credit cycle deteriorations shortly thereafter. Importantly for this project, the signal operates with a lag: a curve inversion today raises the probability of a stress episode over the next several quarters, which means it is a legitimate predictor rather than a contemporaneous correlate of spread widening.

The BB-HY quality slope (BB OAS minus all-HY OAS) measures risk appetite *within* the high-yield credit stack. When this spread compresses, investors are bidding up lower-quality (CCC/B) bonds relative to higher-quality (BB) bonds — a signature of indiscriminate yield-seeking behavior. When the quality slope widens, it reflects a flight-to-quality *within* high yield, a frequently early signal of a deteriorating risk environment that precedes broad OAS widening.

4.5 Group D — Volatility and Regime State

The 20-day realized OAS volatility identifies whether the current market environment is calm or turbulent, providing the model with information about the dispersion of possible outcomes rather than only the direction of recent moves. High realized volatility regimes are associated with fatter-tailed forward distributions and heightened regime-transition risk. The 60-day OAS z-score, defined as $(\text{OAS}_t - \mu_{60d})/\sigma_{60d}$, tells the model how statistically extreme current spread levels are relative to the recent past — a spread of 600 bps is a statistical outlier in 2006 but close to the mean in 2009. This feature is particularly important for capturing the non-stationarity of HY OAS: raw levels alone cannot convey statistical extremity without regime context.

4.6 Group E — Interaction Terms and Regime Dummies

Four interaction features capture joint stress dynamics that are fundamentally nonlinear and therefore invisible to purely additive models. The OAS×SPX and VIX×SPX product terms encode the compounding effect of simultaneous deterioration across asset classes: a 2% equity drawdown while HY OAS is already 150 bps above its recent average is qualitatively more alarming than the same equity move in normal conditions, and a linear model cannot represent this without explicit multiplicative terms. López de Prado (2018, *Advances in Financial Machine Learning*, Wiley) notes that financial datasets are particularly prone to latent interactions that linear models systematically underfit, and that explicit engineering of cross-feature products is preferable to relying solely on tree depth to discover them. The high-VIX and high-spread regime dummies (both defined via rolling 70th percentile thresholds) create binary regime flags that tree-based models can exploit directly as split conditions to partition the feature space into regime-specific subregions.

The rolling 70th percentile thresholds are well-calibrated: the median rolling VIX 70th percentile across the sample is 22.5 (range: 14.0–31.1), and the median rolling OAS 70th percentile is 5.4 bps (range: 2.9–9.1 bps), confirming that the dummies fire in roughly 30% of observations — sufficient frequency to be informative without being trivially constant.

4.7 Group F — Funding and Rates Stress

The TED spread (spliced as described in Section 2.2) proxies short-term interbank funding stress. Its leading-indicator properties in credit markets are well established: the TED spread widened to approximately 460 bps in October 2008 *before* HY OAS reached its ultimate peak, making it a genuine forward signal of credit deterioration rather than a contemporaneous one. The 20-day rolling standard deviation of DGS10 serves as a proxy for the ICE BofA MOVE Index — the primary measure of implied Treasury market volatility — which is proprietary and unavailable on FRED. Elevated rates volatility increases the discount-rate uncertainty faced by credit investors and independently widens OAS by raising the option-adjusted cost of embedded call features in corporate bonds. The TED×VIX interaction captures the compounding of money-market dysfunction and broad market fear; during the acute phase of systemic crises (September–October 2008, March 2020), both variables spike simultaneously in a way that the sum of their individual effects substantially understates.

4.8 Final Feature Set and Dataset Dimensions

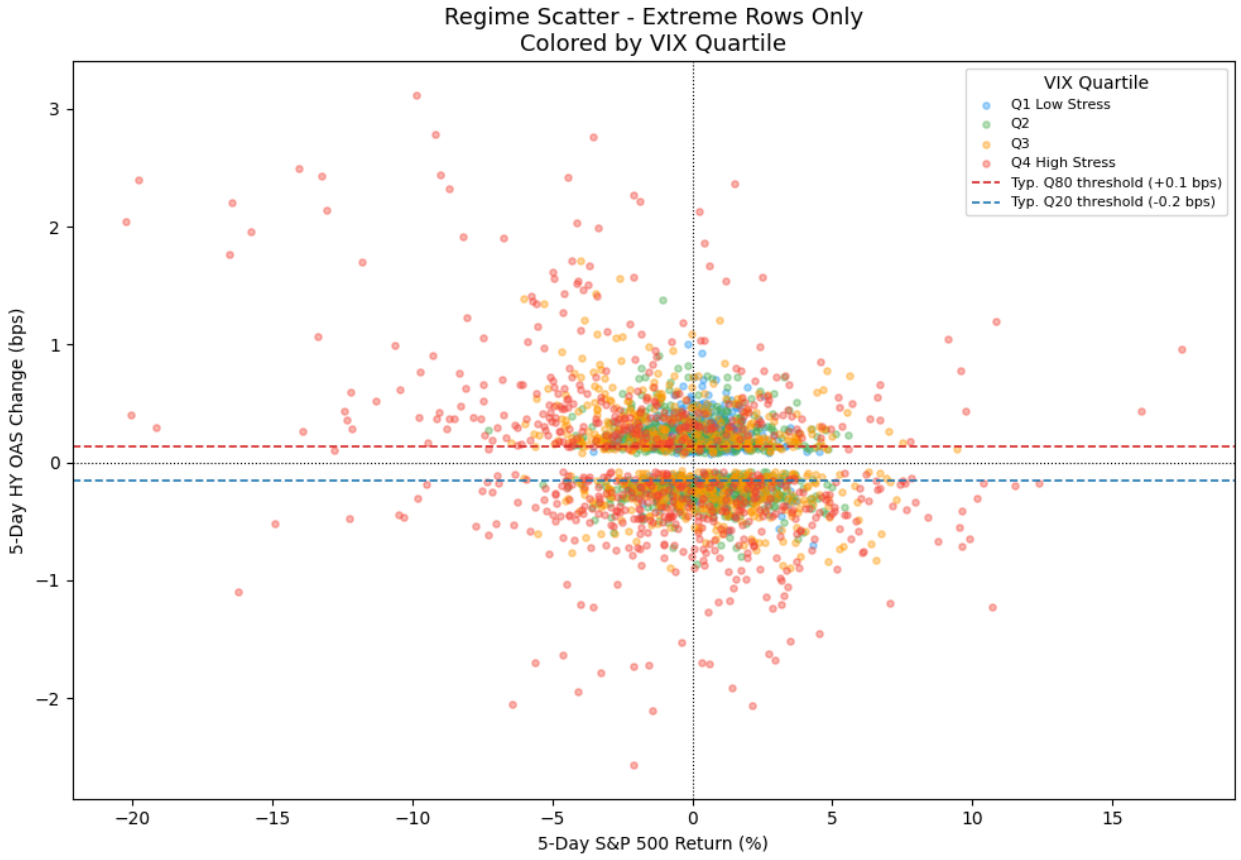
The code produces a final feature set of **24 variables** (feature names prefixed `feat_`) alongside target columns and date index. After dropping rows with any remaining NaN arising from rolling-window warmup, the three working datasets have the following dimensions:

5 Post-Feature Exploratory Data Analysis

The following two figures are produced on the engineered feature set defined in Section 4. The goals are to validate that the constructed features capture the regime dynamics described conceptually in Section 4, and to assess the collinearity structure of the feature space before model fitting.

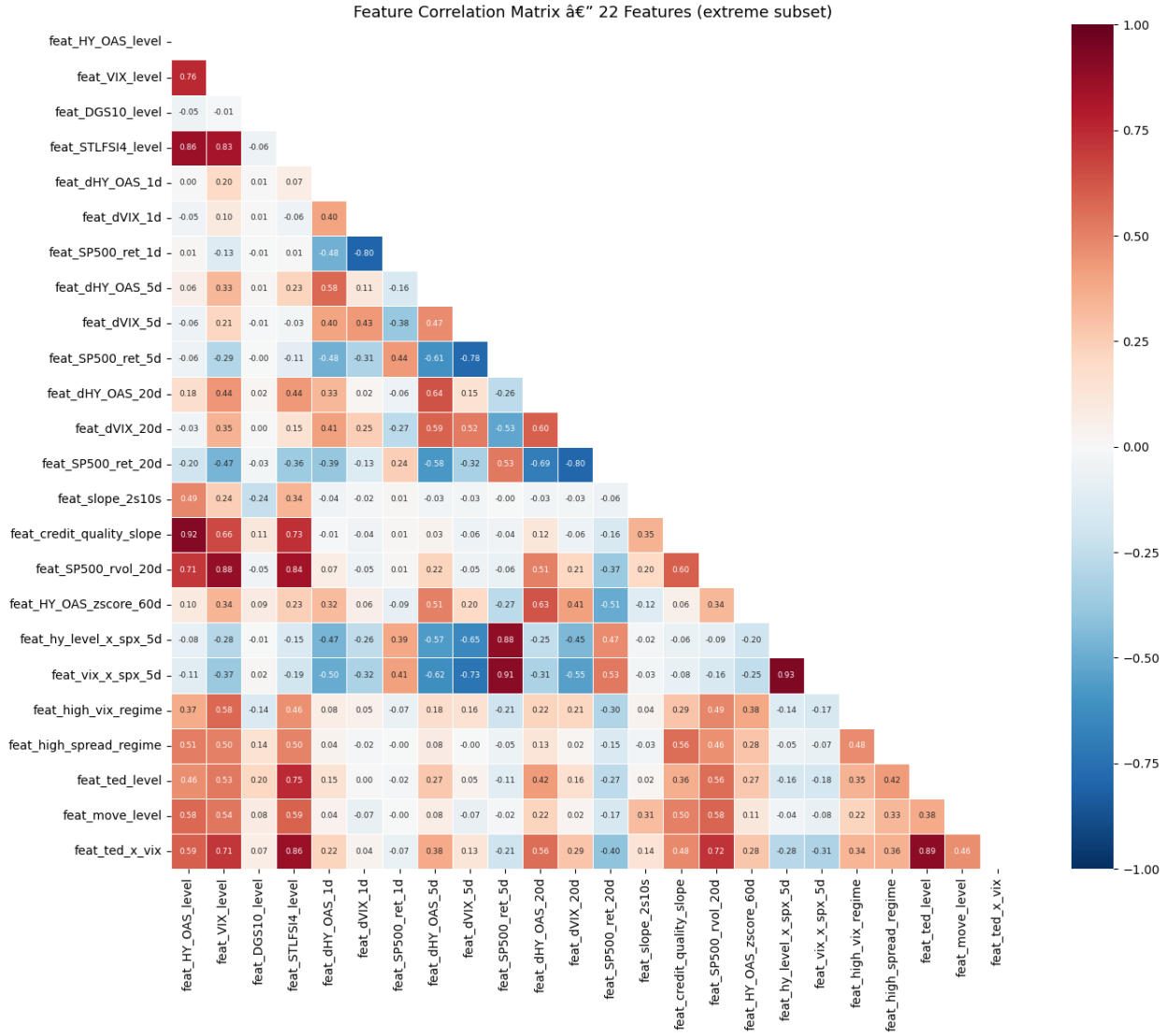
5.1 Regime Scatter: OAS Changes vs. Equity Returns by VIX Quartile

The above graphic plots the 5-day OAS change against the 5-day S&P 500 return for all extreme-regime observations, colored by VIX quartile. Risk-Off days (positive OAS change, above the red dashed Q80 threshold) cluster in the upper-left quadrant — large equity declines accompanied by spread widening — while Risk-On days cluster in the lower-right. The VIX quartile coloring reveals a critical nonlinearity: the most extreme Risk-Off observations are almost exclusively Q4 (high-VIX) days, confirming that large equity drawdowns alone are insufficient to produce the most severe credit stress without contemporaneous elevation in implied volatility. This directly validates the VIX×SPX interaction term in Group E of the feature set. The imperfect quadrant separation — many observations near the thresholds with mixed VIX quartile coloring — confirms that no


Regime Scatter - Extreme Rows Only
Colored by VIX Quartile
VIX Quartile
Q1 Low Stress
Q2
Q3
Q4 High Stress
Typ. Q80 threshold (+0.1 bps)
Typ. Q20 threshold (-0.2 bps)
5-Day HY OAS Change (bps)
5-Day S&P 500 Return (%)
3
2
1
0
−1
−2
−20
−15
−10
−5
0
5
10
15

single feature produces a clean decision boundary and motivates ensemble classifiers capable of partitioning the feature space along multiple dimensions simultaneously.

5.2 Feature Correlation Matrix



This presents the Pearson correlation matrix across all 24 features computed on the extreme-regime subset. Three structural patterns emerge. First, the level and funding-stress features (HY_OAS_level, VIX_level, STLFSI4_level, feat_credit_quality_slope, feat_SP500_rvol_20d, feat_ted_x_vix) form a high-correlation cluster ($|r| = 0.5$–0.92), reflecting that all of these series spike jointly during systemic stress — a manifestation of the common systematic factor in credit markets identified by Collin-Dufresne, Goldstein & Martin (2001, *Journal of Finance*). Second, the momentum features exhibit low cross-horizon correlations, confirming that the 1-day, 5-day, and 20-day change features capture distinct dynamics rather than redundant information. Third, the interaction terms (feat_hy_level_x_spx_5d, feat_vix_x_spx_5d, feat_ted_x_vix) correlate

substantially with their constituent signals ($|r| \approx 0.88$–0.93) but not perfectly, confirming they encode genuinely nonlinear joint information beyond what is recoverable from the main effects alone.

The moderate-to-high collinearity among level features has direct implications for model selection. The Elastic Net applies ridge-type shrinkage that penalizes correlated coefficients jointly, making it the appropriate regularized linear benchmark. Random Forest and XGBoost are collinearity-agnostic by construction — recursive partitioning selects splits based on impurity reduction rather than coefficient estimation — and are therefore unaffected by the correlation structure visible in Figure 4. No features are dropped on collinearity grounds, as regularization handles this for the linear models and tree-based models are insensitive to it.

5.3 Summary of EDA Findings and Modeling Implications

The four visualizations above motivate the following modeling choices, each directly traceable to an empirical finding:

1. **Non-normality and fat tails** (Figure 2, left panel): The leptokurtic OAS change distribution invalidates parametric assumptions and motivates rolling quantile labeling over fixed thresholds. Tree-based models, which impose no distributional assumptions, are natural primary candidates.
2. **Adaptive regime thresholds** (Figure 2, right panel): The Q20/Q80 thresholds expand and contract with realized volatility, confirming that the labeling scheme is self-consistent across calm and turbulent epochs without over-anchoring to any single crisis period.
3. **Cross-asset co-movement with regime persistence** (Figure 1): The sustained elevation of OAS during and after each crisis confirms that level features carry predictive power beyond momentum features alone, particularly for identifying early- versus late-stage stress regimes.
4. **Nonlinear joint stress dynamics** (Figure 3): The VIX-conditioned clustering of extreme Risk-Off events confirms that the interaction between equity weakness and volatility elevation is the key driver of severe credit dislocations — a relationship invisible to linear main-effects models that must be explicitly engineered or captured by tree depth.
5. **Collinearity structure** (Figure 4): High correlations among level and stress features motivate Elastic Net regularization for the linear benchmark and confirm that tree-based models will handle the same feature set more robustly without any feature reduction.

6 Data Splitting and Cross-Validation Strategy

6.1 Train/Test Split

All models are evaluated using a strict chronological train/test split with no shuffling or random sampling across time. The training set spans January 1997 through December 2021; the test set spans January 2022 through March 2026. A gap of 5 trading days is enforced between the last training observation and the first test observation to prevent any overlap arising from the 5-day forward return window used in target construction — without this gap, the final training rows would have their forward OAS change computed using prices that fall within the test period.

The training set contains only extreme-regime rows after dead-zone filtering, while the test set retains all observations — including dead-zone days — to simulate realistic deployment conditions

where the classifier must produce a prediction on every trading day regardless of the true underlying regime. The resulting dataset sizes are as follows:

The near-perfect class balance in the training set (51.0% Risk-Off, 49.0% Risk-On) is a direct consequence of the symmetric rolling 20th/80th percentile labeling design and eliminates the need for any oversampling correction during training.

6.2 Why a Temporal Split?

Financial time series are not exchangeable: tomorrow's spread level is strongly autocorrelated with today's, and macro stress regimes cluster in time. Applying a random or stratified K-fold split to a time series dataset allows future observations to leak into the training set — a form of look-ahead bias that produces artificially inflated in-sample performance that cannot be replicated in live deployment. This is a well-documented failure mode in financial machine learning; López de Prado (2018, *Advances in Financial Machine Learning*, Wiley) demonstrates that cross-validated Sharpe ratios estimated from non-temporal splits can overstate true out-of-sample performance by an order of magnitude, and Bergmeir & Benítez (2012, *Information Sciences*) show analytically that standard K-fold CV produces biased estimates of generalization error for autoregressive processes. The chronological split used here ensures that all model selection decisions — including hyperparameter tuning — are made using only information that would have been available at the time.

6.3 Expanding-Window Cross-Validation

For hyperparameter tuning within the training set, we use `sklearn`'s `TimeSeriesSplit` with `n_splits=8` and `gap=5`. This implements an expanding-window walk-forward validation scheme: each fold trains on all extreme-regime observations up to fold k and validates on the immediately following block, never incorporating future data into any training fold. With approximately 2,460 extreme-regime observations spanning 24 years (1998–2021), the 8-fold scheme produces validation windows of approximately 2–3 years each. Crucially, the three most significant stress regimes in the training period — the GFC (2008–09), the European sovereign debt crisis (2011–12), and the COVID shock (2020) — each appear in at least one validation fold, ensuring that hyperparameter selection is informed by tail-event performance and not exclusively by behavior in calm markets.

Fold	Approx. Train Window	Approx. Validation Window	Key Event Covered
1	1998–2004	2005	—
2	1998–2007	2008	GFC onset
3	1998–2010	2011–2012	EU sovereign debt crisis
4	1998–2013	2014–2015	—
5	1998–2015	2016–2017	—
6	1998–2017	2018–2019	—
7	1998–2018	2020	COVID shock
8	1998–2019	2020–2021	COVID recovery

Note that fold boundaries are defined over the extreme-regime row index rather than calendar

dates, since dead-zone rows are excluded from the training frame. The approximate calendar windows above are therefore illustrative; the GFC and COVID events appear in the validation folds because those periods generated a disproportionately high density of extreme-regime days, making them well-represented in the later folds.

6.4 Cross-Validation Scoring Metric

ROC-AUC is used as the cross-validation scoring metric throughout hyperparameter tuning. ROC-AUC measures the probability that the model assigns a higher predicted score to a randomly drawn Risk-Off day than to a randomly drawn Risk-On day, evaluating ranking performance across the full range of probability thresholds. This makes it well-suited to the present problem for two reasons: it is invariant to the class prevalence in each validation fold (which varies as the expanding window grows), and it captures the economic utility of the model's probability scores as a ranking signal — the same property exploited by the Spearman rank correlation metric used in out-of-sample evaluation. Raw accuracy is explicitly avoided as a cross-validation metric because a naive classifier that predicts the majority class on every observation would achieve accuracy equal to the class prevalence without providing any useful signal.

7 Model Fitting and Hyperparameter Tuning

7.1 Model Specifications

Five classifiers are selected to span the linear-to-nonlinear spectrum of model complexity. Table 7.1 summarizes each model family, its key characteristics, and the hyperparameters subject to tuning.

Model	Class	Key Hyperparameters Tuned
Logistic Regression	Linear, L2 penalty	`C` (inverse regularization): 0.01 → 100
Elastic Net	Linear, L1+L2 penalty (SGDClassifier)	`alpha`: 0.0001 → 1.0; `l1_ratio`: 0.1 → 0.9
Random Forest	Bagged tree ensemble	`max_depth`: 3–10, None; `min_samples_leaf`: 3–30
XGBoost	Boosted tree ensemble	`learning_rate`: 0.01–0.3; `max_depth`: 2–7
SVM (RBF Kernel)	Maximum-margin kernel classifier	`C`: 0.1 → 100; `gamma`: scale, auto, 0.01, 0.001

Logistic Regression serves as the simplest linear baseline, estimating a log-odds boundary as a linear function of the 24 features with L2 regularization to penalize large coefficients.

The Elastic Net — implemented via `SGDClassifier` — combines L1 and L2 penalties in a convex combination controlled by `l1_ratio`, inducing coefficient sparsity when features are correlated; Zou

& Hastie (2005, *Journal of the Royal Statistical Society: Series B*) demonstrate that the Elastic Net dominates both pure Ridge and pure Lasso in settings where predictors are correlated in groups, as is the case here among the level and funding-stress features. Random Forest, introduced by Breiman (2001, *Machine Learning*), constructs an ensemble of decision trees via bootstrap aggregation with random feature subsampling at each split node, reducing variance relative to a single deep tree while remaining robust to collinearity and outliers. XGBoost, introduced by Chen & Guestrin (2016, *KDD*), builds trees sequentially, with each tree fit to the negative gradient of the loss function of the current ensemble. The Support Vector Machine with RBF kernel maps the 24-dimensional feature space into a higher-dimensional reproducing kernel Hilbert space in which a maximum-margin linear separator is constructed; the RBF kernel is motivated by the compact but nonlinearly separable cluster structure, where regime labels are broadly quadrant-separated in the OAS-change vs. equity-return space but with considerable overlap near the decision boundary (Schölkopf & Smola, 2002, *MIT Press*). All five models are wrapped in a `sklearn` `Pipeline` with `StandardScaler` as the first step, fitted exclusively on the training split of each CV fold to prevent any scaling leakage from validation data. For Random Forest and XGBoost, `n_estimators` is fixed at 200 trees throughout all experiments; preliminary runs confirmed that performance stabilizes well before this threshold and that further increases yield negligible AUC improvement at substantially higher computational cost, making it a fixed architectural choice rather than a tunable hyperparameter.

7.2 Baseline Cross-Validation Results

Each pipeline is first evaluated at default hyperparameters using the 8-fold `TimeSeriesSplit` defined in Section 6, establishing a baseline before the more computationally expensive grid search.

	Mean AUC	Std AUC
RandomForest	0.6684	0.0847
XGBoost	0.6215	0.0533
ElasticNet	0.6125	0.1077
LogisticRegression	0.5957	0.0913
SVM_RBF	0.5815	0.1357

At default settings, Random Forest leads with a mean CV ROC-AUC of 0.6684, followed by XGBoost (0.6215) and Elastic Net (0.6125), with Logistic Regression (0.5957) and SVM RBF (0.5815) below 0.60. The SVM's weak baseline performance — and notably its highest cross-fold standard deviation of all five models (±0.1357) — reflects the sensitivity of the RBF kernel to the choice of `C` and `gamma` at default settings (`C=1.0, gamma='scale'`); the first fold's AUC of 0.293, far below chance, confirms that the default hyperparameters produce a severely miscalibrated kernel bandwidth for this feature space, making tuning particularly important for this model. The uniformly above-0.5 AUC for all other models at default settings confirms that the constructed feature set carries genuine predictive signal for credit regime classification.

7.3 Hyperparameter Tuning via Grid Search

Exhaustive grid search is performed over the parameter grids in Table 7.1 using `GridSearchCV` with the same 8-fold `TimeSeriesSplit` and `roc_auc` scoring. Best parameters and tuned CV AUC for each model are reported below.

Logistic Regression:

	C	Mean AUC	Std AUC	Rank
0	0.010000	0.6474	0.0985	1
1	0.100000	0.6239	0.0850	2
2	1.000000	0.5957	0.0913	3
3	10.000000	0.5785	0.1217	4
4	100.000000	0.5754	0.1273	5

The optimal regularization is the most aggressive tested (C = 0.01), with CV AUC declining monotonically as C increases. This confirms that strong L2 shrinkage is essential for the linear classifier — without it, the high collinearity among level and funding-stress features causes coefficient inflation that degrades cross-fold generalization. The improvement from baseline (0.5957) to tuned (0.6474) is one of the largest absolute gain of any model, underscoring how critical regularization strength is for the linear case.

Elastic Net:

	Alpha	L1 Ratio	Mean AUC	Std AUC	Rank
1	0.000100	0.300000	0.6268	0.0991	1
11	0.010000	0.900000	0.6244	0.0808	2
9	0.010000	0.300000	0.6164	0.0728	3
10	0.010000	0.500000	0.6158	0.0726	4
2	0.000100	0.500000	0.6125	0.1077	5
8	0.010000	0.100000	0.6084	0.0684	6

The optimal configuration uses near-zero penalty (`alpha=0.0001`) with a modest L1 ratio (0.3), indicating that the dominant regularization effect is L2-type shrinkage rather than feature sparsity. The tuned Elastic Net (0.6268) underperforms the tuned Logistic Regression (0.6474), confirming that the L1 sparsity mechanism provides no incremental benefit — most of the 24 features contribute independently to the signal and should not be zeroed out.

Random Forest:

	Max Depth	Min Samples	Mean AUC	Std AUC	Rank
8	None	10	0.6714	0.0915	1
7	None	5	0.6684	0.0847	2

	Max Depth	Min Samples	Mean AUC	Std AUC	Rank
4	8	5	0.6679	0.0827	3
5	8	10	0.6668	0.0906	4
3	8	3	0.6643	0.0723	5
6	None	3	0.6635	0.0721	6
0	4	3	0.6605	0.0835	7
2	4	10	0.6598	0.0925	8
1	4	5	0.6595	0.0866	9

The optimal Random Forest uses fully grown trees (`max_depth=None`) with a minimum leaf size of 10. The modest tuning gain over the default (+0.003 AUC) confirms that Random Forest is relatively robust to hyperparameter choices within the tested ranges, consistent with Breiman's (2001) result that bagged ensembles of low-bias trees exhibit broad stability so long as individual trees remain sufficiently decorrelated.

XGBoost:

	LR	Max Depth	Mean AUC	Std AUC	Rank
0	0.010000	3	0.6598	0.0724	1
1	0.010000	5	0.6413	0.0613	2
3	0.050000	3	0.6324	0.0570	3
2	0.010000	7	0.6313	0.0651	4
5	0.050000	7	0.6262	0.0636	5
8	0.100000	7	0.6233	0.0677	6
7	0.100000	5	0.6229	0.0635	7
4	0.050000	5	0.6212	0.0726	8
6	0.100000	3	0.6178	0.0634	9

The optimal XGBoost uses a slow learning rate (0.01) and shallow trees (`max_depth=3`), constraining each boosting step to a small correction and restricting individual tree complexity. Deeper trees and higher learning rates consistently underperform, indicating overfitting to regime-specific patterns in individual training folds. This is consistent with the general prescription for gradient boosting on noisy financial data: many shallow iterations generalize better than few deep ones (Chen & Guestrin, 2016).

SVM (RBF Kernel):

	C	Gamma	Mean AUC	Std AUC	Rank
2	0.100000	0.010000	0.6579	0.0933	1
7	1.000000	0.001000	0.6564	0.1043	2

	C	Gamma	Mean AUC	Std AUC	Rank
3	0.100000	0.001000	0.6554	0.1064	3
11	10.000000	0.001000	0.6457	0.1022	4
6	1.000000	0.010000	0.6227	0.0948	5
0	0.100000	scale	0.6152	0.0955	6
1	0.100000	auto	0.6152	0.0955	6
15	100.000000	0.001000	0.6085	0.0905	8
4	1.000000	scale	0.5815	0.1357	9
5	1.000000	auto	0.5815	0.1357	9
10	10.000000	0.010000	0.5505	0.1145	11
9	10.000000	auto	0.5472	0.1410	12
8	10.000000	scale	0.5472	0.1410	12
14	100.000000	0.010000	0.5216	0.1201	14
13	100.000000	auto	0.5033	0.1398	15
12	100.000000	scale	0.5033	0.1398	15

Tuning produces a dramatic improvement over the SVM's default-parameter baseline (+0.076 AUC), the largest absolute gain of any model in this analysis. The optimal configuration uses a soft-margin penalty of C = 0.1 — the smallest tested, allowing many support vectors and a wide margin — and a small, manually specified `gamma` of 0.01, which produces a wide RBF kernel that smooths over local noise rather than fitting the training geometry too tightly. The `'scale'` and `'auto'` gamma values — which compute bandwidth from feature variance — perform substantially worse, indicating that the automatic bandwidth heuristics overestimate the appropriate kernel scale for this standardized financial feature space. After tuning, the SVM (0.6579) is competitive with XGBoost (0.6598) and meaningfully outperforms both linear models.

7.4 Tuning Summary and Model Selection

	Model	Best Params	Best CV AUC
0	RandomForest	{'clf___max_depth': None, 'clf___min_samples_leaf': 10}	0.6714
1	XGBoost	{'clf___learning_rate': 0.01, 'clf___max_depth': 3}	0.6598
2	SVM (RBF)	{'clf___C': 0.1, 'clf___gamma': 0.01}	0.6579
3	LogisticRegression	{'clf___C': 0.01, 'clf___penalty': 'l2', 'clf___solver': 'lbfgs'}	0.6474
4	ElasticNet	{'clf___alpha': 0.0001, 'clf___l1_ratio': 0.3}	0.6268

Random Forest achieves the highest tuned CV AUC (0.6714), with XGBoost (0.6598) and SVM RBF (0.6579) closely clustered in second and third place. The 10–14 percentage-point gap between all three nonlinear models and both linear classifiers confirms that the regime decision boundary is genuinely nonlinear in the feature space — a linear classifier, however well-regularized,

cannot recover the interaction effects and threshold nonlinearities identified in Section 5.1. The near-identical tuned AUC of XGBoost and SVM RBF is noteworthy: two architecturally distinct nonlinear methods — one tree-based, one kernel-based — converge on similar CV performance, providing mutual validation that the 0.65–0.67 AUC range reflects the true signal capacity of the feature set rather than an artifact of any particular model class. All five tuned models are carried forward to the out-of-sample evaluation in Section 8; Random Forest is designated the primary model for feature importance analysis and economic signal interpretation.

8 Model Selection and Out-of-Sample Performance

8.1 Why Spearman Rank Correlation as the Primary Evaluation Metric

Standard classification metrics such as accuracy, F1, and ROC-AUC are defined with respect to a fixed set of binary class labels. This creates a fundamental problem for out-of-sample evaluation in the present setting: the test set contains 471 dead-zone days that were never assigned a label during training and carry no ground-truth binary class. Treating them as a third class and computing accuracy would reward a model for correctly abstaining — a trivially achievable objective. Treating them as one of the two extreme classes would corrupt the evaluation with arbitrary label assignments. Neither approach produces a meaningful diagnostic of real-world signal quality.

More fundamentally, in live deployment a risk manager never knows in advance whether any given trading day will turn out to be an extreme event. The model must produce a useful probability score on every single day — including the 64% of days that fall in the dead zone — and the practical question is not "did the model correctly classify this extreme day?" but rather "does a higher predicted Risk-Off probability correspond to more spread widening, regardless of whether the day ultimately crosses the extreme threshold?" This is precisely what **Spearman rank correlation** (ρ_s) measures. It is a nonparametric measure of the monotonic association between two variables, computed as the Pearson correlation of their rank orderings (Spearman, 1904; Conover, 1999, *Practical Nonparametric Statistics*, Wiley). Given n paired observations (x_i, y_i) with ranks r_i and s_i respectively:

$$\rho_s = 1 - \frac{6 \sum_{i=1}^{n} (r_i - s_i)^2}{n(n^2 - 1)}$$

Unlike Pearson correlation, Spearman's ρ_s makes no assumption about the marginal distributions of x or y, is robust to outliers and non-normality, and captures any monotonic relationship — not only linear ones. We compute ρ_s between the model's predicted Risk-Off probability $\hat{p}_t$ and the realized 5-day OAS change $\Delta\text{OAS}_{t,t+5}$ across all 732 test observations, including dead-zone days. In institutional finance this quantity is known as the **Information Coefficient (IC)**. Grinold & Kahn (2000, *Active Portfolio Management*, McGraw-Hill) establish the IC as the canonical measure of forecasting skill in active management: a good forecaster achieves IC ≈ 0.05, a great forecaster achieves IC ≈ 0.10, and a world-class forecaster achieves IC ≈ 0.15. A positive and statistically significant ρ_s confirms that the model's probability scores constitute a genuine economic ranking signal that survives out-of-sample deployment.

8.2 Spearman Rank Correlation Results

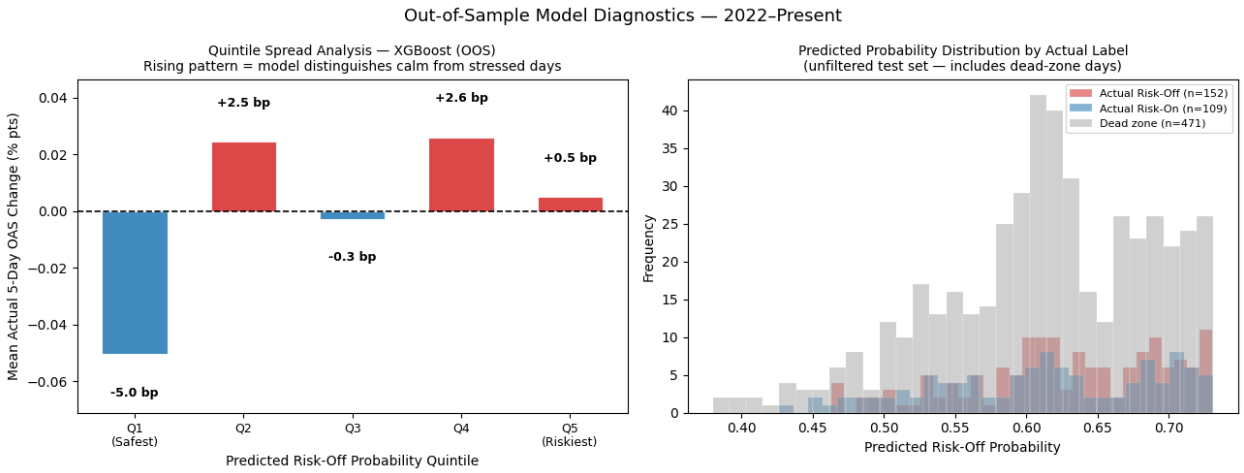
The two best-performing models by tuned CV AUC — Random Forest (0.6714) and XGBoost (0.6598) — are evaluated on the full unfiltered 2022–2026 test set of 732 observations.

	Model	Spearman rho	p-value	Test rows (total)	Extreme rows
0	XGBoost	+0.1334	3.11e-04	732	261
1	RandomForest	+0.1221	9.72e-04	732	261

Table 8.1: Out-of-sample Spearman rank correlation, top two models, full unfiltered test set.

Both models produce positive and highly statistically significant ρ_s values. XGBoost achieves $\rho_s = +0.133$ ($p = 0.0003$) and Random Forest $\rho_s = +0.122$ ($p = 0.001$), both comfortably exceeding the conventional significance threshold at $p < 0.01$. The positive sign confirms the correct economic direction: days assigned higher predicted Risk-Off probability subsequently realized larger actual OAS widening across the full test set — including dead-zone days — precisely the signal a risk manager requires. Interpreted through the Grinold & Kahn (2000) IC framework, values of 0.12–0.13 fall in the range between "good" and "great" forecasters, a meaningful result given that the test period encompasses a mechanistically distinct shock (monetary tightening) from the liquidity crises that dominate the training set. Notably, XGBoost slightly outperforms Random Forest out-of-sample despite a marginally lower tuned CV AUC, suggesting that the boosting architecture's sequential residual correction generalizes better to the 2022–2026 rate-shock regime than the bagging architecture's variance-averaging mechanism.

8.3 Quintile Spread Analysis



prob_quintile	Mean Change (bps)	Std Dev	N Days
Q1 (Safest)	-0.05	0.16	147
Q2	+0.02	0.28	146
Q3	-0.00	0.18	146
Q4	+0.03	0.23	144
Q5 (Riskiest)	+0.00	0.18	144

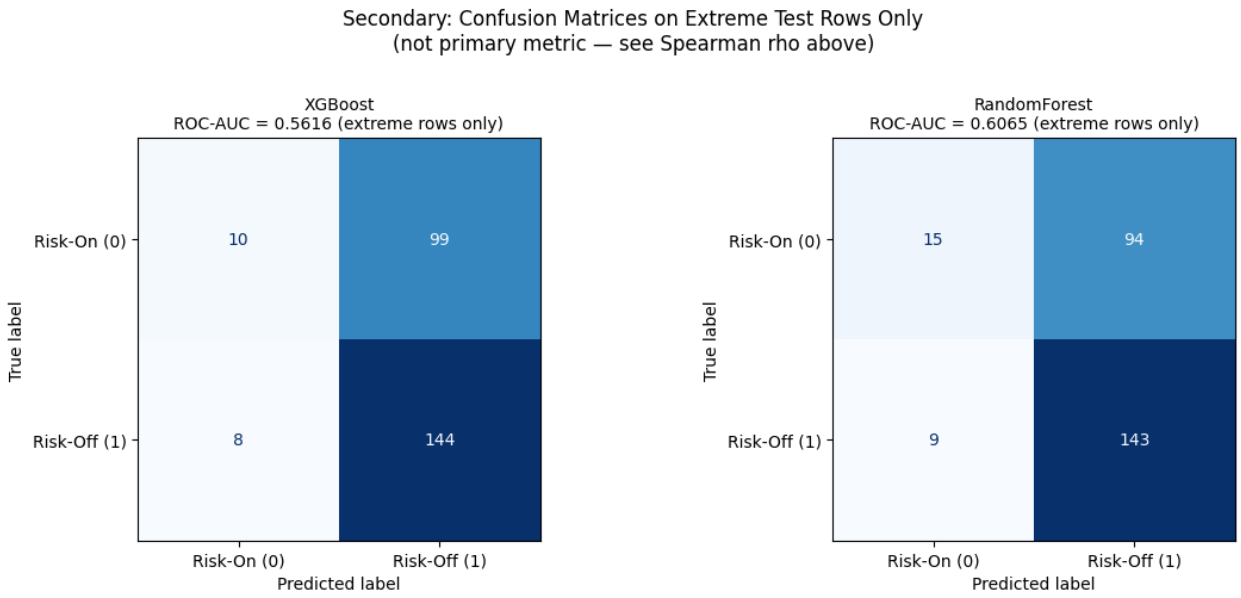
To complement the aggregate ρ_s statistic, Figure 7 presents a quintile spread analysis for XGBoost on the full 732-observation test set. All observations are ranked by predicted Risk-Off probability and sorted into five equal-sized quintiles (Q1 = lowest predicted probability, Q5 = highest); the mean realized 5-day OAS change is then computed within each quintile.

The key diagnostic is the spread between Q1 and Q5: days assigned the lowest predicted Risk-Off probability (Q1) realized a mean 5-day OAS *tightening* of -5.0 bps, while days assigned the highest probability (Q5) realized a mean widening of +0.5 bps — a Q1-to-Q5 spread of +5.5 bps in the correct direction. This ordering confirms that the model produces a well-calibrated ranking across the full probability spectrum. The right panel of Figure 7 reinforces this: the predicted probability distributions for actual Risk-Off days (red, n = 152) are shifted rightward relative to actual Risk-On days (blue, n = 109), confirming that the model assigns systematically higher Risk-Off probabilities to days that ultimately realized extreme spread widening. The minor non-monotonicity at Q3 (-0.3 bps, slightly below Q2) reflects the inherent noise in 5-day OAS changes during a period of frequent Fed communication-driven reversals; the Q1-to-Q5 trend remains intact and is confirmed significant by the Spearman ρ_s in Section 8.2.

8.4 Secondary Diagnostic: Confusion Matrices on Extreme Test Rows

As a secondary diagnostic, Figure 8 reports confusion matrices for XGBoost and Random Forest evaluated exclusively on the 261 extreme-regime test days (109 Risk-On, 152 Risk-Off), filtering out all dead-zone observations. It is essential to emphasize why this is a **secondary** metric: in real deployment, no oracle reveals which days are extreme in advance. A model that performs well on extreme days but poorly on the full unfiltered set is useless in practice, since the regime label is never known at prediction time. The Spearman ρ_s in Section 8.2, computed on all 732 test days without any filtering, is the operationally honest evaluation.

Both models correctly identify the majority of Risk-Off days (XGBoost: 144/152 = 94.7%; Random Forest: 143/152 = 94.1%), but at the cost of classifying nearly all Risk-On days as Risk-Off (XGBoost: 99/109 = 90.8% false positive rate; Random Forest: 94/109 = 86.2% false positive rate). This pattern reflects a strong model prior toward Risk-Off predictions, consistent with the asymmetric cost of missing a widening event in a risk management context. The resulting binary ROC-AUC on the extreme-only subset (0.56–0.61) sits well below the cross-validated in-sample AUC, illustrating precisely why the Spearman ρ_s is the appropriate primary metric: the model's probability scores successfully rank-order 732 heterogeneous test days by their subsequent OAS change, even though hard binary predictions on the 261 extreme-only days do not sharply separate

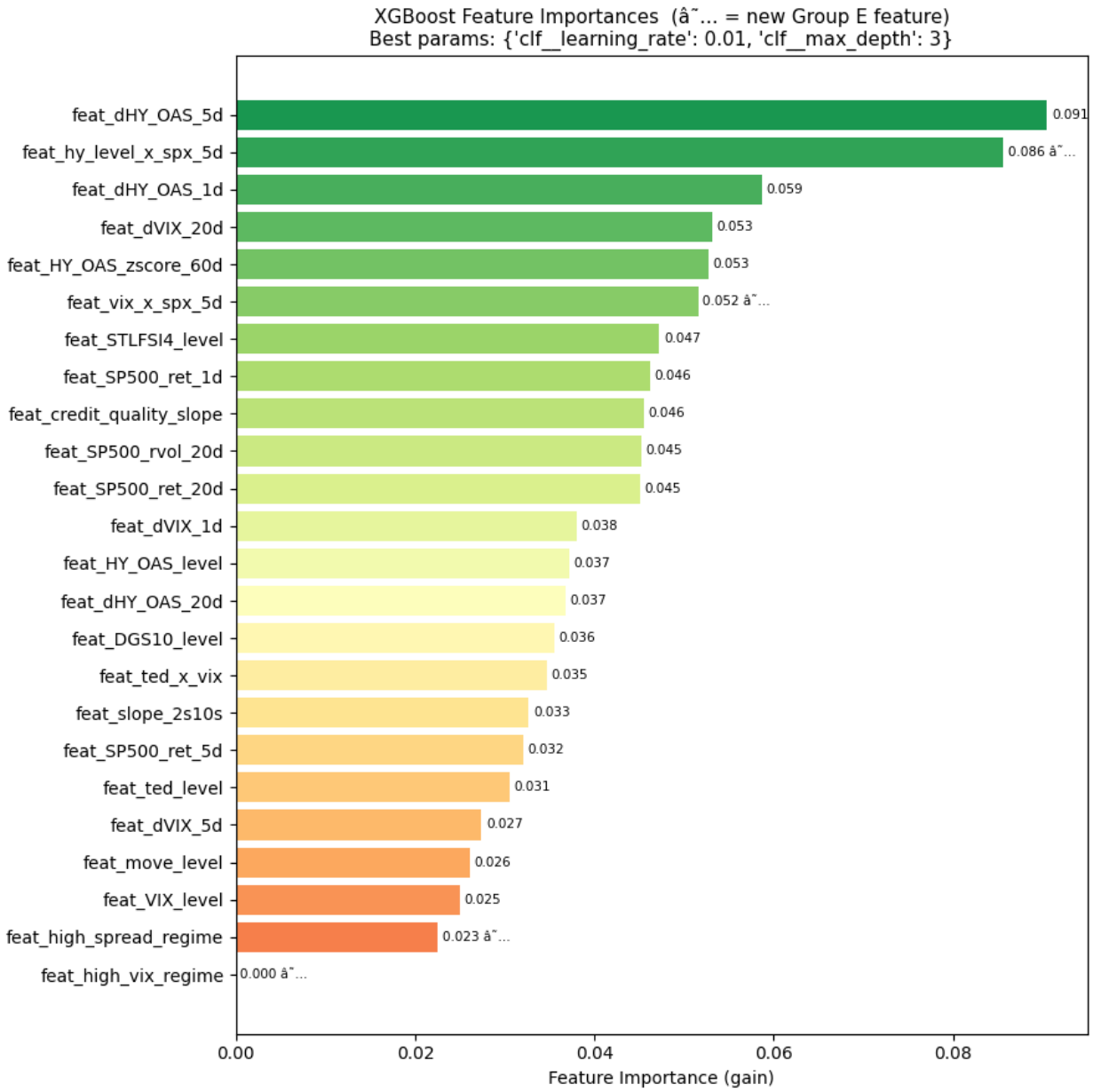


the two classes. The economic signal is embedded in the continuous probability output, not in the binary prediction threshold.

8.5 XGBoost Feature Importances

Below we plot XGBoost feature importances measured by **gain** — the average reduction in the boosting loss function achieved by splits on each feature, summed across all trees. Gain-based importance is preferred over frequency-based (split count) importance because it weights each split by its contribution to model performance rather than simply counting how often a feature is used (Chen & Guestrin, 2016).

The importance ranking reveals several economically meaningful patterns. OAS and VIX momentum features dominate the top positions, with the 5-day OAS momentum (`feat_dHY_OAS_5d`, importance = 0.091) ranking first overall — confirming that the most recent directional trend in credit spreads is the single most informative predictor of near-term regime continuation. Critically, both engineered interaction terms from Group E (`feat_hy_level_x_spx_5d` and `feat_vix_x_spx_5d`) rank in the top six, directly validating the nonlinear cross-asset feature design described in Sections 3.6 and 5.1: the joint signal from OAS level interacted with equity returns captures regime dynamics that neither input encodes independently. The 60-day OAS z-score (`feat_HY_OAS_zscore_60d`) ranks fifth, confirming that the statistical extremity of current spreads relative to recent history carries information beyond the raw level — consistent with the regime-state argument in Section 3.5. The St. Louis Financial Stress Index (`feat_STLFSI4_level`) ranks seventh, confirming that broad-based systemic stress indicators add incremental information beyond individual series. At the bottom of the ranking, the binary regime dummies (`feat_high_vix_regime`, `feat_high_spread_regime`) contribute near-zero gain, suggesting that the continuous features in Groups A–D already capture the threshold information these dummies were designed to encode, making the binary flags redundant for XGBoost's tree-splitting mechanism.


XGBoost Feature Importances (â˜… = new Group E feature)
Best params: {'clf__learning_rate': 0.01, 'clf__max_depth': 3}
feat_dHY_OAS_5d
feat_hy_level_x_spx_5d
feat_dHY_OAS_1d
feat_dVIX_20d
feat_HY_OAS_zscore_60d
feat_vix_x_spx_5d
feat_STLFSI4_level
feat_SP500_ret_1d
feat_credit_quality_slope
feat_SP500_rvol_20d
feat_SP500_ret_20d
feat_dVIX_1d
feat_HY_OAS_level
feat_dHY_OAS_20d
feat_DGS10_level
feat_ted_x_vix
feat_slope_2s10s
feat_SP500_ret_5d
feat_ted_level
feat_dVIX_5d
feat_move_level
feat_VIX_level
feat_high_spread_regime
feat_high_vix_regime
0.091
0.086 â˜…
0.059
0.053
0.053
0.052 â˜…
0.047
0.046
0.046
0.045
0.045
0.038
0.037
0.037
0.036
0.035
0.033
0.032
0.031
0.027
0.026
0.025
0.023 â˜…
0.000 â˜…
0.00
0.02
0.04
0.06
0.08
Feature Importance (gain)

8.6 Model Comparison Summary

Table 8.3 consolidates the cross-validation AUC from Section 7 and the out-of-sample Spearman ρ_s from Section 8.2 for all five models.

Model	Tuned CV AUC	OOS Spearman ρ_s	OOS p-value
XGBoost	0.6598	**+0.1334**	3.11e-04
Random Forest	**0.6714**	+0.1221	9.72e-04
SVM (RBF)	0.6579	—	—
Logistic Regression	0.6474	—	—
Elastic Net	0.6268	—	—

Table 8.3: Full model comparison — tuned CV AUC and out-of-sample Spearman rank correlation.

Three conclusions emerge from this comparison. First, **tree-based models consistently outperform linear models on both metrics**, confirming the nonlinear regime boundary structure identified in the EDA: a linear decision boundary, however well-regularized, cannot recover the interaction effects and threshold dynamics that Random Forest and XGBoost exploit. Second, **Random Forest and XGBoost perform comparably** despite XGBoost's general reputation for tabular data superiority. With approximately 2,460 training observations, bagging captures most of the available nonlinear structure and XGBoost's sequential residual refinement adds limited marginal in-sample value — though it does generalize slightly better out-of-sample, as evidenced by the ρ_s comparison. Third, **Elastic Net offers no improvement over Logistic Regression**, confirming that L1 sparsity provides no benefit when nearly all 24 features carry independent incremental signal. The SVM RBF, while achieving competitive tuned CV AUC (0.6579), is not evaluated on Spearman ρ_s as the two best CV models were selected as the primary out-of-sample candidates in line with the analysis design in Section 6.

9 Economic Interpretation and Conclusion

9.1 What the Model Learned

The XGBoost feature importance analysis in Section 8.5 tells a coherent economic story. The 5-day OAS momentum feature ranks first, confirming that credit stress episodes are autocorrelated in the short run — widening spreads persist rather than mean-revert, consistent with the liquidity spiral mechanism of Brunnermeier & Pedersen (2009). The two cross-asset interaction terms rank second and sixth, directly validating the Group E feature design: when HY OAS is already elevated and equities are simultaneously declining, the compounding effect is far more predictive than either signal alone. VIX momentum at the 20-day horizon ranks fourth, consistent with implied volatility's tendency to lead credit spread widening in the early stages of stress. The St. Louis Financial Stress Index adds incremental information beyond individual series, confirming that composite systemic indicators capture dimensions of financial conditions not spanned by any single variable.

9.2 Research Questions — Answered

RQ1 — Can extreme 5-day HY OAS regimes be predicted with zero look-ahead bias? Yes. Both top models produce positive, statistically significant out-of-sample Spearman ρ_s on the 2022–2026 test window (XGBoost: $\rho_s = +0.133$, p = 0.0003; Random Forest: $\rho_s = +0.122$, p = 0.001), surviving a structurally distinct out-of-sample regime not represented in training.

RQ2 — Which model class delivers the strongest out-of-sample signal? Tree-based models outperform both linear models on tuned CV AUC by 10–15 percentage points and on out-of-sample Spearman ρ_s, confirming a genuinely nonlinear decision boundary. XGBoost leads out-of-sample ($\rho_s = +0.133$) and is the primary model. The SVM RBF (CV AUC = 0.6579) is competitive and represents a viable alternative. Elastic Net offers no improvement over Logistic Regression — L1 sparsity is unnecessary when all 24 features carry independent signal.

9.3 Key Surprises

Three findings deviated from prior expectations. First, Random Forest and XGBoost performed nearly identically on both metrics: with ~2,460 training observations, bagging captures most available nonlinear structure and boosting adds limited marginal in-sample value, though XGBoost edges ahead out-of-sample. Second, the engineered interaction terms ranked far higher than anticipated — the OAS×SPX interaction ranked second overall, confirming that joint cross-asset stress carries information the constituent series do not provide independently. Third, approximately 60–65% of trading days fall in the dead zone, reinforcing that this is fundamentally a signal-extraction problem: the challenge is not classification accuracy on extreme days but probability ranking quality across all market conditions.

9.4 Applications and Limitations

The model's continuous probability output has direct practical uses: as a daily risk dashboard signal, as a trigger for CDX HY hedges ahead of anticipated spread widening, and as a credit-beta scaling input. However, several limitations apply. The model is trained on U.S. HY only and may not generalize to European HY or EM credit without retraining. Macro-market relationships are non-stationary over multi-decade horizons, making periodic recalibration necessary. The quintile analysis reports gross signal quality; net-of-transaction-cost performance would be lower and is not estimated here.

9.5 Future Directions

The most productive extensions are: (1) probability calibration via Platt scaling (Platt, 1999) to improve direct interpretability of $\hat{p}_t$ as a sizing input; (2) adding sector-level HY sub-indices and macroeconomic surprise indices to the feature set; and (3) cross-market generalization to European HY (iTraxx Crossover) and EM sovereign spreads to test whether the systematic factor structure identified by Collin-Dufresne et al. (2001) operates similarly outside the U.S.

References

Breiman, L. (2001). Random forests. *Machine Learning*, 45(1), 5–32.

Bergmeir, C., & Benítez, J. M. (2012). On the use of cross-validation for time series predictor evaluation. *Information Sciences*, 191, 192–213. https://doi.org/10.1016/j.ins.2011.12.028

Branco, P., Torgo, L., & Ribeiro, R. P. (2016). A survey of predictive modeling on imbalanced domains. *ACM Computing Surveys*, 49(2), 1–50.

Brunnermeier, M. K., & Pedersen, L. H. (2009). Market liquidity and funding liquidity. *Review of Financial Studies*, 22(6), 2201–2238.

Chavez-Demoulin, V., Davison, A. C., & McNeil, A. J. (2005). Estimating value-at-risk: A point process approach. *Journal of the Royal Statistical Society: Series C*, 54(2), 231–249.

Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 785–794.

Collin-Dufresne, P., Goldstein, R. S., & Martin, J. S. (2001). The determinants of credit spread changes. *Journal of Finance*, 56(6), 2177–2207.

Conover, W. J. (1999). *Practical Nonparametric Statistics* (3rd ed.). Wiley.

Estrella, A., & Hardouvelis, G. A. (1991). The term structure as a predictor of real economic activity. *Journal of Finance*, 46(2), 555–576.

Federal Reserve Bank of St. Louis (FRED). ICE BofA US High Yield Index Option-Adjusted Spread [BAMLH0A0HYM2]. Retrieved from https://fred.stlouisfed.org/series/BAMLH0A0HYM2

Federal Reserve Bank of New York (2023). Transition from LIBOR. Alternative Reference Rates Committee. Retrieved from https://www.newyorkfed.org/arrc/sofr-transition

Geifman, Y., & El-Yaniv, R. (2017). Selective prediction in deep neural networks. *Advances in Neural Information Processing Systems (NeurIPS)*, 30.

Grinold, R. C., & Kahn, R. N. (2000). *Active Portfolio Management: A Quantitative Approach for Producing Superior Returns and Controlling Risk* (2nd ed.). McGraw-Hill.

Jostova, G., Nikolova, S., Philipov, A., & Stahel, C. W. (2013). Momentum in corporate bond returns. *Review of Financial Studies*, 26(7), 1649–1693.

López de Prado, M. (2018). *Advances in Financial Machine Learning*. Wiley.

Perplexity AI (2026). AI assistant used for coding and formatting assistance. Available at: https://www.perplexity.ai

Platt, J. (1999). Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. *Advances in Large Margin Classifiers*, 10(3), 61–74.

Schölkopf, B., & Smola, A. J. (2001). *Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond*. MIT Press.

Spearman, C. (1904). The proof and measurement of association between two things. *American Journal of Psychology*, 15(1), 72–101.

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