Master's thesis

Duration timing on Government Bond Indices with Factor, Macro and Market signals

Edoardo Tarcisio Visconti

Supervisors: David Anderson (OLZ), Patrick Walker (OLZ), Dr. Urban Ulrych (ETH)

Department of Mathematics

Fall Semester 2025

Contents

1 Problem setting and introduction

Factor models, particularly value and momentum, have been shown to exhibit predictive power across asset classes. After reviewing the relevant literature, the first part of this thesis examines whether these signals are informative for duration timing, defined as selecting the most attractive maturity bucket within a fixed-income index. We focus on the WGBI, a globally diversified sovereign bond benchmark, which reduces exposure to country-specific macroeconomic effects and allows for a cleaner assessment of factor signals.

The second part shifts to a single-country setting, focusing on the United States. Here we test whether the same signals remain informative when macroeconomic dynamics are country specific, and we extend the analysis to macroeconomic and market variables using a range of machine-learning methods. This setting is more suitable for evaluating the predictive content of variables such as policy rates, inflation, and growth indicators, which are inherently country dependent.

in the third part: to do IPCA?

Duration timing refers to the selection of the most attractive maturity bucket within a fixed-income index. In this study, the available buckets are 1–3, 3–5, 5–10, and 10+ years. The analysis is conducted from the perspective of a professional asset manager seeking to improve returns while controlling risk. Accordingly, performance is primarily evaluated using Sharpe ratios, with cumulative returns playing a secondary role. We also report Sharpe ratios adjusted for turnover-related transaction costs; in practice, these differ only marginally from the unadjusted Sharpe ratios, indicating that turnover does not materially affect the conclusions in this setting. Turnover is computed at each rebalancing date as the sum of absolute changes in portfolio weights relative to the pre-rebalancing (price-drifted) allocation, and transaction costs are incorporated using a linear penalty with $c = 0.0001$, which reflects a conservative estimate of trading costs in highly liquid government bond markets.

The benchmark is an equal-weighted portfolio of the four duration buckets, rebalanced monthly and fully agnostic to market conditions. This provides a neutral, no-forecast allocation that any duration-timing strategy should outperform in terms of risk-adjusted performance.

We adopt a holding period of 21 business days, corresponding to a monthly investment horizon commonly used in professional fixed-income management. The analysis is restricted to long-only allocations, consistent with institutional mandate constraints that typically preclude short positions in sovereign duration buckets. Cash holdings are also excluded, as cash-equivalent instruments at scale behave similarly to short-duration sovereign bonds in both risk and return terms.

The market data is provided by OLZ AG, while selected macroeconomic series are sourced from the Federal Reserve Economic Data (FRED) database. Although data are available from 2000 to the end of 2024, the analysis starts in 2005 to ensure sufficient history for reliable signal construction (which is necessary since some strategies require feature engineering or threshold calibration. The period from 2005 to 2014 is used

as the training window, with the objective of promoting robustness rather than maximizing in-sample Sharpe ratios. Performance is then evaluated strictly out of sample over 2015–2024. Given the strong regime dependence of bond markets, these choices are made to obtain strategies that generalize across market environments rather than overfitting historical conditions.

To characterize the evaluation period, we compute Sharpe ratios separately for 2014–2019 and 2019–2024. This analysis is purely ex post and is not used in strategy design or tuning, in order to avoid forward-looking bias.

The cumulative returns of the four CHF-hedged WGBI duration buckets (Figure 1) highlight the contrast between these regimes. The 2014–2019 period is relatively stable and favorable to long-duration exposure, whereas 2019–2024 is marked by severe drawdowns driven by inflation shocks and aggressive monetary tightening. As a result, strategies naturally perform differently across the two environments.

This implies that a strategy calibrated on the training period (2005–2014) may perform well in one part of the evaluation sample but poorly in another. The analysis therefore emphasizes the importance of robustness and motivates a focus on strategies that generalize across regimes rather than overfit specific historical conditions.

Now follows a concise introduction to bonds, written from the perspective of a STEM student with limited prior knowledge of finance, as is the case for the author.

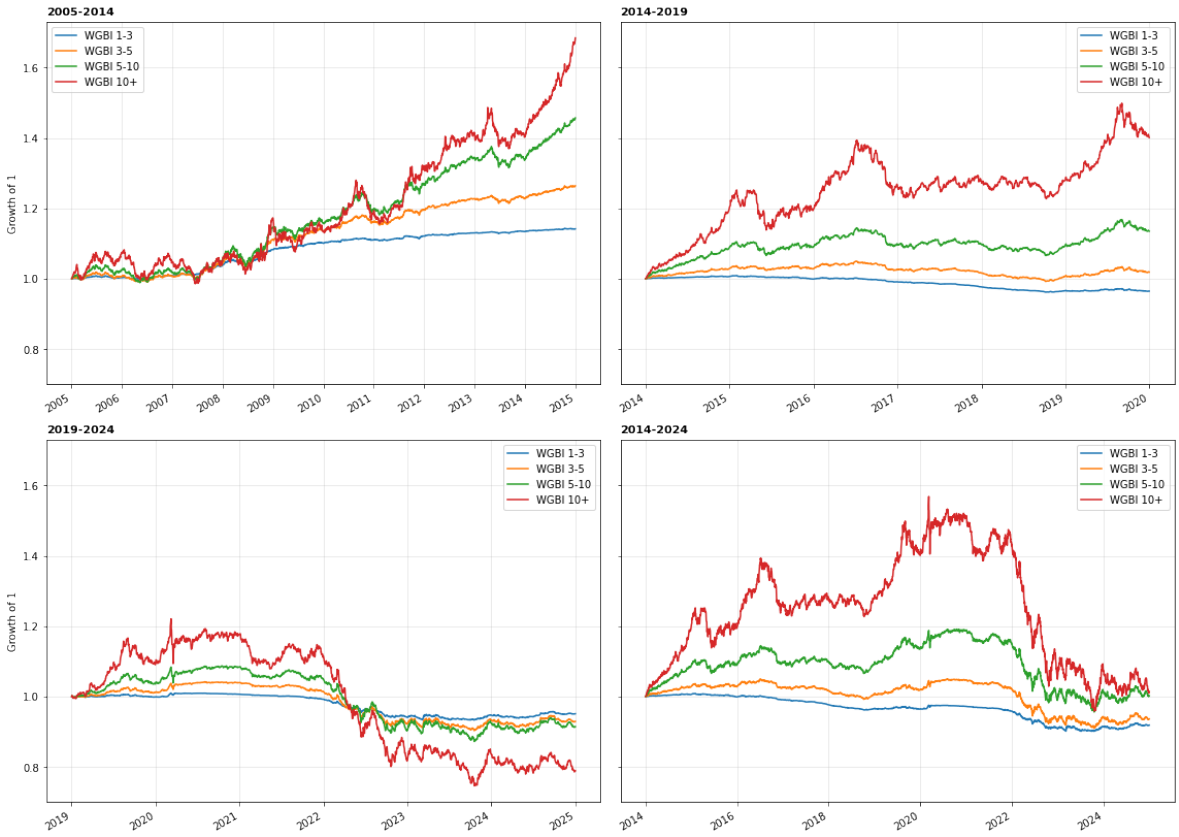

Figure 1: Cumulative returns of the four WGBI duration buckets (1–3, 3–5, 5–10, 10+) hedged into CHF. Top-left: "training" sample (2005–2014), top-right: early test period (2014–2019), bottom-left: late test period (2019–2024), bottom-right: full evaluation period (2014–2014).

2 Government Bonds and the Macroeconomic Picture for Mathematicians

In this introductory section we explain the basic mechanics of government bonds, how price and yield are mathematically linked, and how these quantities respond to monetary policy and macroeconomic conditions. We introduce the concepts of duration and clarify the economic meaning and shape of the yield curve.

2.1 Government Bonds

A bond is a debt instrument in which the issuer promises to pay fixed coupon interest at regular intervals (annual or semiannual) and to repay the principal at maturity. Government bonds (or sovereign bonds) are issued by national governments to borrow from the market and finance public spending.

At issuance, a bond has:

- a redemption value R (the amount repaid at maturity);

- a sequence of cash flows $\{C_t\}_t$, usually a fixed percentage of the face value (the coupon).

The bond price equals the present value of future payments discounted at the market **yield to maturity (YTM)**:

$$P = \sum_{t=1}^{T} \frac{C_t}{(1+\text{YTM})^t} + \frac{R}{(1+\text{YTM})^T}.$$

The YTM is the bond's internal rate of return if held to maturity: the single discount rate that makes the present value of all coupons and principal equal to the market price. Plotting yields against maturities gives the *yield curve*. In this project we use bond indexes, which aggregate bonds across maturities. Each index's yield is the market-value-weighted average of its constituents' yields.

The primary market is the market for newly issued bonds, where governments or corporations sell debt directly to investors through auctions (typically among institutional investors such as primary dealers, pension funds, or large asset managers). These investors submit bids indicating the price or yield at which they are willing to purchase the bonds. The auction clears when the total demand equals the issuance amount, and the yield of the last accepted bid becomes the issue yield.

After issuance, the same securities trade among investors in the secondary market, where prices fluctuate daily with supply, demand, and expectations about future interest rates and inflation.

In the primary market (issuance stage), the yield emerges from investor demand; the price adjusts so that the fixed coupon rate and market yield are consistent. In the *secondary market*, yields continue to fluctuate dynamically as bonds are traded and market conditions evolve.

From the equation above, price and yield move inversely:

$$\text{Yield rises} \Rightarrow \text{Price falls}, \qquad \text{Yield falls} \Rightarrow \text{Price rises}.$$

Investors trade based on *price*, but what truly matters to them is *yield*. Price is simply the mechanism that adjusts until the desired yield is achieved.

Investors decide what yield they require, considering:

- prevailing market rates and the yield curve,
- credit and duration risk,
- alternative investment opportunities.

They then act on prices:

$$\begin{cases} \text{If a bond's yield is too low} & \Rightarrow \text{sell} \Rightarrow \text{price drops} \Rightarrow \text{yield rises}, \\ \text{If a bond's yield is too high} & \Rightarrow \text{buy} \Rightarrow \text{price rises} \Rightarrow \text{yield falls}. \end{cases}$$

The market converges to an equilibrium price where the bond's yield equals the market yield for its maturity and credit risk. This keeps the market arbitrage-free: all bonds with the same risk and maturity must offer the same yield.

When a new bond is issued, it establishes a benchmark yield for that maturity. Existing bonds with similar maturities reprice to align with that benchmark, including any liquidity premia, and local yield changes propagate through the entire yield curve.

A bond's sensitivity to changes in yield is measured by its duration. The *Macaulay duration* D_{mac} is defined as the weighted average time of the discounted cash flows:

$$D_{\text{mac}} = \frac{1}{P} \sum_{t=1}^{T} t \cdot \frac{C_t + I_{t=T} R}{(1 + \text{YTM})^t}.$$

It measures the average time (in years) required to recover the bond's price through its discounted payments.

The *modified duration* adjusts for compounding:

$$D_{\text{mod}} = \frac{D_{\text{mac}}}{1 + \text{YTM}}.$$

For a small change in yield Δy,

$$\frac{\Delta P}{P} \approx -D_{\text{mod}} \, \Delta y.$$

Thus, bonds with longer duration are more sensitive to yield movements. Intuitively, the Macaulay duration represents the time at which all discounted cash flows could be concentrated into a single equivalent payment.

The *convexity* refines this linear approximation by capturing curvature in the price–yield relation:

$$\frac{\Delta P}{P} \approx -D_{\text{mod}} \, \Delta y + \tfrac{1}{2} C_x (\Delta y)^2,$$

where C_x denotes the convexity coefficient. Higher convexity implies that price gains from yield decreases are larger than price losses from equivalent yield increases.

This framework provides the mathematical foundation to interpret bond prices, yields, and their sensitivity to macroeconomic and policy shifts.

2.2 Yield Curves and Monetary Policies

a. Yield Curve Dynamics

The yield curve, i.e. the term structure of interest rates, describes how bond yields vary with their maturities at a given point in time.

The (zero-coupon) yield curve is widely employed as a benchmark for pricing fixed-income securities and as an indicator of economic expectations. Its configuration provides insights into anticipated economic growth, inflation, and monetary policy direction:

- A **normal** (upward-sloping) curve, where longer maturities exhibit higher yields, indicates a positive term premium. This is the most common configuration. Longer-term bonds incorporate both expectations of future interest rate increases in an expanding economy and a higher yield to compensate investors for greater liquidity risk over time.

- An **inverted** (downward-sloping) curve, where short-term yields exceed long-term yields, reflects expectations that current restrictive monetary policy will lead to slower growth and eventual rate cuts. It typically signals market anticipation of an economic downturn or recession, as observed before major crises such as 2008.

The curve can also be **flat**, reflecting uncertainty or a transitional phase in the economic cycle, where short- and long-term yields converge due to ambiguous expectations about future interest rates. Alternatively, a **humped** curve, with yields peaking at intermediate maturities, may indicate temporary market imbalances in supply and demand or differing risk perceptions across maturities.

Table 1: Yield Curve Shapes and Historical Examples

Yield Curve Shape	**Examples and Context**
Normal (upward-sloping)	U.S. (2004–2006): typical of an expansion phase. Short-term rates were low while long-term yields were higher, reflecting expectations of future growth, inflation, and monetary tightening.
Inverted (downward-sloping)	U.S. (2006–2007): short-term yields exceeded long-term yields as the Federal Reserve raised policy rates. The inversion anticipated the 2008 financial crisis.

The Nelson–Siegel model provides a parsimonious parametric representation of the yield curve. It expresses the yield at maturity τ as:

$$y(\tau) = \beta_0 + \beta_1 \frac{1 - e^{-\tau/\lambda}}{\tau/\lambda} + \beta_2 \left(\frac{1 - e^{-\tau/\lambda}}{\tau/\lambda} - e^{-\tau/\lambda} \right)$$

The three parameters β_0, β_1, and β_2 (obtained in practice by minimising the MSE at each time) correspond respectively to the *level*, *slope*, and *curvature* of the yield curve, while λ controls the exponential decay rate that determines the maturity at which the maximum curvature occurs. These three factors (level, slope, and curvature) are also those typically identified by applying PCA to yields across maturities, confirming that both methods capture the same fundamental sources of variation in the term structure.

A more flexible extension of this specification is the Nelson–Siegel–Svensson (NSS) model, which adds an additional curvature term governed by a second decay parameter. This introduces an extra coefficient that allows the yield curve to capture more complex shapes, particularly at long maturities. In the empirical analysis that follows, we adopt this extended specification to better fit the cross section of yields while retaining a clear economic interpretation of the underlying factors.

b. Monetary policies

Central banks use policy rates to steer inflation, employment, and overall economic stability. By influencing short-term borrowing costs, policy rates affect aggregate demand, credit conditions, and inflation expectations, with the ultimate goal of keeping inflation close to target while stabilizing output.

Rate hikes are implemented to slow economic activity and contain inflation. Higher borrowing costs reduce consumption and investment, short-term yields rise, and the yield curve often flattens or inverts as markets anticipate weaker growth or future easing. Rate cuts, instead, aim to stimulate the economy: borrowing becomes cheaper, demand recovers, short-term yields fall, and the curve typically steepens. When policy is stable and inflation is near target, yield movements are driven mainly by expectations, and the curve remains relatively normal. Inflationary or deflationary shocks shift yields across maturities, with the short end reacting first to changes in policy expectations.

As a result, different duration buckets respond to different economic forces:

- **1–3 years (short duration):** Primarily driven by near-term monetary policy, with low interest-rate sensitivity and relatively low volatility. Usually performs better in rising-rate environments.
- **3–5 years (short–intermediate):** Influenced by policy expectations and medium-term macro dynamics, offering a balanced risk–return profile.
- **5–10 years (intermediate):** Sensitive to inflation expectations, growth outlook, and term-premium movements. Performs well during slowdowns and easing cycles, but with higher volatility.
- **10+ years (long duration):** Most sensitive to yield changes, driven by long-run inflation expectations, real rates, and global risk sentiment. High potential returns in sustained easing regimes, but vulnerable during inflationary periods.

3 Relevant Literature and motivation

3.1 Motivation for factor strategies on the WGBI

As discussed in the introduction, this thesis focuses on duration timing, defined as the selection of the most attractive maturity bucket within a fixed-income index. The key idea is that different segments of the yield curve do not offer the same risk–return trade-off at all times, since each maturity responds differently to changes in market conditions. Short maturities are mainly driven by expectations about near-term policy rates and react quickly to central bank communication, while long maturities are more sensitive to long-run inflation expectations, term premia, and shifts in risk aversion.

The objective is therefore to choose, among the 1–3, 3–5, 5–10, and 10+ year buckets, the segment that offers the most favorable compensation for bearing duration risk at a given point in time. The problem is inherently comparative: rather than forecasting absolute bond returns, it aims to identify which part of the yield curve is expected to outperform the others over the next holding period on a risk-adjusted basis.

A key motivation for pursuing this framework comes from recent research on yield-curve premia. Brooks and Moskowitz [4] provide a comprehensive international study showing that government bond returns are strongly shaped by style characteristics traditionally used in other asset classes, such as value, momentum, and carry. Their analysis covers both the time series and cross-section of returns across countries and

maturities, and their central result is that these style characteristics explain yield-curve premia more effectively than the usual principal-component factors (level, slope, curvature) that dominate the term-structure literature.

For duration timing, their findings are particularly relevant for several reasons:

- **Information beyond the yield curve.** Style factors contain pricing information that is not fully captured by the yield curve itself, implying that yield movements alone do not summarize expected returns across maturities and that factor signals can help identify relative outperformance along the curve.
- **Heterogeneous exposure across maturities.** Value, momentum, and carry load differently on level, slope, and curvature portfolios, implying that short, intermediate, and long maturities respond asymmetrically to the same underlying signal.
- **Robustness across markets and regimes.** Style signals remain effective across a wide range of markets, regimes, and datasets, including both synthetic zero-coupon yields and live traded bonds, suggesting that they are not sample-specific or tied to a single country's dynamics.

Other works also show that value and momentum signals carry strong predictive power in fixed income, consistent with what has been documented for equities and other asset classes.

For example, Asness, Moskowitz, and Pedersen [1] demonstrate that value and momentum effects are persistent and robust across global markets, including government bonds. In their setting, value captures whether yields or term premia are high relative to their historical levels, while momentum measures the strength and direction of recent price trends. Both signals predict bond excess returns, and combining them improves performance because the two contain complementary information.

We build on these ideas by constructing value, momentum, and carry signals following the same principles used in their paper. We then combine them into a single indicator, using both a simple average and correlation-adjusted weighting based on their historical co-movement.

Baltussen, Martens, and Penninga [2] extend this evidence using an exceptionally long historical sample of sovereign bond markets. They confirm that value and momentum signals remain effective across countries, regimes, and transaction-cost assumptions. Importantly, their results show that signal strength varies across the yield curve, reinforcing the idea that different maturities react differently and that factor-driven rotation across duration buckets is economically meaningful.

Additional support comes from the broader trend-following and momentum literature. Kolanovic and Wei [10] document that momentum strategies applied across asset classe, including bonds, consistently identify periods when price trends persist, improving tactical allocation decisions. Similarly, Hurst, Ooi, and Pedersen [8] analyse more

than a century of data and conclude that trend-following delivers persistent excess returns across markets and time horizons. Their findings highlight that momentum is not market-specific but a general feature of asset-price dynamics, including fixed income.

We therefore test several variants of momentum-style signals, and in the next section we also allow our machine-learning models to incorporate a form of momentum by assigning higher weights to more recent observations.

Taken together, this line of research provides a strong justification for examining whether similar signals can help identify which duration bucket is most attractive at a given point in time. This motivates the first part of the thesis, where we test whether factor signals hold predictive power on the WGBI index before turning to the U.S. market for more granular macro-driven analysis.

3.2 Motivation for strategies on the US markets using macro and market signals with ML

In 1997, Ilmanen [9] provides one of the earliest comprehensive studies on U.S. bond return predictability. He shows that excess returns on different maturities can be anticipated using simple economic and market indicators such as term spreads, real yields, market levels, mometum signals and measures of market volatility. Importantly for duration timing, Ilmanen documents that these signals affect short, intermediate, and long maturities differently. Short maturities react mostly to near-term policy expectations, while long maturities respond to long-run inflation and term-premium dynamics. This heterogeneity directly supports the idea that rotating across maturity buckets can be beneficial.

Ludvigson and Ng [11] extend the view of Ilmanen by introducing macroeconomic information extracted from a very large panel of economic indicators. They show that a small number of latent macro factors, summarizing broad economic activity and inflation conditions, contain substantial predictive power for bond excess returns. Their results demonstrate that macroeconomic information can improve forecasts beyond what yield-curve variables alone provide, establishing one of the most influential macro-based frameworks in fixed-income predictability.

This provides background and motivation for introducing macroeconomic and market variables into the analysis of U.S. government bonds.

However, it is not clear how to integrate such a large amount of time-series information directly into a practical strategy. For this reason, we turn to machine learning, which allows us to handle many features simultaneously and to engineer signals that interact with each other in a systematic way.

Bianchi, Buchner, and Tamoni [3] further develop this direction by applying machine-learning models to forecast U.S. Treasury excess returns. They show that non-linear methods such as random forests and neural networks extract additional predictive structure compared to traditional linear regressions, especially when dealing with many macro and financial predictors. Their results highlight several features that are important for our purposes: 1. predictive signals differ across maturities, 2. interactions between variables matter, and 3. the structure of expected returns is regime-dependent.

In their work, they alos show that non-linearities within macroeconomic categories are more important than interactions across categories. This aligns with the structure of our problem: because predictors within each dataset are often correlated and the number of observations available at each rebalancing date is limited, our analysis focuses primarily on within-category interactions. This point will be made precise in the relevant methodological (see Section 5.3). What is a constraint for us is supported by their empirical evidence. They also document that ensembling shallow networks can substitute for additional network depth; analogously, we examine whether ensembles of simple, lightweight models improve stability and predictive performance in our setting.

Eventually, Caruso and Coroneo [5] add an important dimension by working with real-time macroeconomic data instead of revised series. They show that the information available to investors at the time of the forecast, incorporating asynchronous releases, revisions, and survey expectations, improves interest-rate predictions, especially at short maturities. Their mixed-frequency framework shows that the predictive content of macro variables changes depending on the state of policy and the information set available at the time. We take inspiration from their data construction but adapt it to our setting: we work exclusively with real-time macroeconomic series, never with revised vintages, because our perspective is that of a practitioner intending to deploy the model in live conditions.

Taken together, these papers motivate the second part of the thesis: to study whether richer sets of macroeconomic and market predictors, combined with machine-learning methods, can improve duration-timing decisions in the U.S. Treasury market, where signals are numerous, noisy, and maturity-specific.

3.3 IPCA for gov bonds

4 Factor strategies on WGBI index

4.1 Overview

We analyze duration-timing strategies on the WGBI using returns fully hedged into CHF. This removes exchange-rate effects that could dominate short-horizon returns and obscure the role of duration. It also reflects the perspective of OLZ as a Swiss asset manager, for whom CHF-hedged returns are the relevant investment measure.

Following the literature reviewed in Section 3.1, we evaluate factor-based strategies such as carry, value, and momentum, together with simple combinations and transformations of these signals. At this stage, we deliberately exclude macroeconomic and broad market variables to isolate the predictive content of factor signals alone.

4.2 Carry, Value and Momentum

a. Signal Definitions and Economic Motivation

- **Carry**

$$s_{b,t}^{(\text{carry})} = \bar{y}_{b,t}^{(5d)}$$

Higher yields imply higher expected carry and roll-down if the curve remains stable. This signal favours buckets with relatively higher yields and is equivalent to betting that the yield curve will not change over the next 21 days. Yields are smoothed using a 5-day moving average to reduce short-term noise and the influence of transient outliers.

- **Value**
  - *Rolling historical value*

    $$s_{b,t}^{(\text{value\_hist})} = y_{b,t} - \bar{y}_{b,t}^{(1y)}$$

    A yield above its one-year average suggests that the bucket is "cheap" relative to its recent history and its price may revert upward. This signal favours buckets whose yield level exceeds their own past average.
  - *NSS curve-based value*

    $$s_{b,t}^{(\text{value\_NSS})} = y_{b,t} - \hat{y}_{b,t}^{\text{NSS}}$$

    Here the observed yield is compared to the yield implied by a Nelson–Siegel–Svensson curve fitted daily to the full cross-section of maturities. The curve is estimated by ordinary least squares using the average life of each maturity segment available (i.e. 1–3y, 3–5y, 5–7y, 7–10y, 10–15y, 15–20y, and 20+y) as a proxy for its time to maturity. A yield above the smooth NSS curve indicates that the bucket is "cheap" relative to the structural shape of the curve and may mean revert through a price increase.

  In essence, the two value signals bet on mean reversion but capture different forms of mispricing. The rolling historical measure is a *time-series* signal: it assumes that each bucket has its own typical yield level and that deviations from this level will revert. The NSS-based measure is a *cross-sectional* signal: it assumes that the yield curve should be smooth across maturities and detects distortions relative to that shape. Using both allows the strategy to target mispricings arising either from bucket-specific dynamics or from broader deformations of the yield curve itself.

- **Momentum signals**
  - *Directional momentum*

    $$s_{b,t}^{(\text{mom1})} = \text{sign}\left(r_{b,t}^{(1w)}\right) + \text{sign}\left(r_{b,t}^{(2w)}\right) + \text{sign}\left(r_{b,t}^{(1m)}\right)$$

    Captures the consistency of short-term returns but may become uninformative when all buckets move together. Indeed, in such regimes, all buckets may rise or fall together for extended periods, causing the directional signal to assign the same score to every bucket. This effectively collapses the strategy into an equal-weighted allocation, providing no useful cross-sectional information. For this reason, we introduce additional momentum variants that extract trend strength over different horizons

– *1-month cumulative momentum*

$$s_{b,t}^{(\text{mom2})} = r_{b,t}^{(1m)}$$

A short-term trend signal that assigns magnitude rather than discrete votes.

– *6-month cumulative momentum*

$$s_{b,t}^{(\text{mom3})} = r_{b,t}^{(6m)}$$

A medium-term trend indicator that reacts more slowly and captures smoother directional movements.

– *Blended momentum*

$$s_{b,t}^{(\text{mom4})} = \tfrac{1}{2}\, r_{b,t}^{(1m)} + \tfrac{1}{2}\, r_{b,t}^{(6m)}$$

A combination of short- and medium-term horizons that stabilises the signal and reduces noise.

In all cases, the underlying idea is the same: we bet on a positive relationship between recent performance and future performance. These momentum indicators are therefore purely time-series signals, designed to capture persistent trends within each duration bucket. Although they are constructed in a time-series manner, they are ultimately used to form cross-sectional allocations across buckets at each rebalancing date.

b. Portfolio Construction

To transform the signals into portfolio weights, we first standardise each factor cross-sectionally at every rebalancing date:

$$\tilde{s}_{b,t}^{(f)} = \frac{s_{b,t}^{(f)} - \mu_t^{(f)}}{\sigma_t^{(f)}}.$$

This ensures that only the *relative* strength of each bucket's signal matters at time t, and provides a stable scale for mapping signals into portfolio weights.

Since all strategies are fully invested and long-only, we convert the composite scores into weights using a softmax transformation:

$$\omega_{b,t}^{(f)} = \frac{\exp\left(s_{b,t}^{(\text{f})}/\tau\right)}{\sum_j \exp\left(s_{j,t}^{(\text{f})}/\tau\right)},$$

where we set $\tau = 0.5$, chosen during the training phase as a value that balances concentration and diversification in the resulting allocations.

We adopt the softmax transformation rather than simple rescaling for two reasons. First, softmax remains well-defined even if all standardised signals are negative: it assigns greater weight to the least negative bucket rather than collapsing the strategy into an equal-weight allocation. Second, alternatives such as zero-clipping negative signals followed by normalization discard useful information and delivered inferior performance in the training.

c. Results

We now comment on the performance of these signals by analyzing the tables below (Tables 2, 3, 4, and 5). We also note that the results obtained in the evaluation periods and presented here (or in the following paragraph) are not used in any way when combining strategies in the next section.

In all tables, the blue row denotes the benchmark allocation, dark green highlights strategies that outperform the benchmark in both cumulative return and Sharpe ratio, light green indicates outperformance in Sharpe ratio only, and yellow indicates higher cumulative returns only. The same color-coding convention is used consistently in the tables presented in the remainder of the paper.

Training Sample (2005–20114)

Strategy	Cum.	Sharpe	Sharpe/turn
WGBI 1–3y	0.142	1.779	1.779
WGBI 3–5y	0.264	1.407	1.407
WGBI 5–10y	0.457	1.276	1.276
WGBI 10+y	0.685	0.945	0.945
equal weighted	0.375	1.223	1.223
carry	0.640	0.961	0.961
value	0.398	1.241	1.237
nss	0.257	1.159	1.141
momentum	0.380	1.258	1.217
momentum2	0.589	1.213	1.181
momentum3	0.531	1.073	1.059
momentum4	0.558	1.138	1.116

Table 2: Pure factor strategies, 2005–2014 (training sample).

Evaluation Period I (2014–2019)

Strategy	Cum.	Sharpe	Sharpe/turn
WGBI 1–3y	−0.035	−1.193	−1.193
WGBI 3–5y	0.020	0.249	0.249
WGBI 5–10y	0.137	0.835	0.835
WGBI 10+y	0.403	0.989	0.989
equal weighted	0.121	0.784	0.784
carry	0.362	0.993	0.991
value	0.038	0.549	0.539
nss	0.129	0.771	0.752
momentum	0.196	0.979	0.951
momentum2	0.362	1.276	1.249
momentum3	0.247	0.787	0.777
momentum4	0.279	0.919	0.902

Table 3: Pure factor strategies, 2014–2019 (evaluation period I).

Evaluation Period II (2019–2024)

Strategy	Cum.	Sharpe	Sharpe/turn
WGBI 1–3y	−0.049	−0.666	−0.666
WGBI 3–5y	−0.070	−0.429	−0.429
WGBI 5–10y	−0.084	−0.300	−0.300
WGBI 10+y	−0.208	−0.389	−0.389
equal weighted	−0.102	−0.410	−0.410
carry	−0.077	−0.202	−0.205
value	−0.098	−0.590	−0.597
nss	−0.083	−0.284	−0.288
momentum	−0.111	−0.415	−0.434
momentum2	−0.108	−0.340	−0.367
momentum3	−0.041	−0.128	−0.141
momentum4	−0.094	−0.288	−0.308

Table 4: Pure factor strategies, 2019–2024 (evaluation period II).

Full Evaluation Sample (2014–2024)

Strategy	Cum.	Sharpe	Sharpe/turn
WGBI 1–3y	−0.081	−0.792	−0.792
WGBI 3–5y	−0.063	−0.259	−0.259
WGBI 5–10y	0.004	0.027	0.027
WGBI 10+y	0.015	0.055	0.055
equal weighted	−0.027	−0.055	−0.055
carry	0.198	0.318	0.316
value	−0.077	−0.319	−0.327
nss	−0.011	−0.007	−0.017
momentum	0.008	0.037	0.016
momentum2	0.059	0.128	0.103
momentum3	0.083	0.172	0.160
momentum4	0.057	0.125	0.106

Table 5: Pure factor strategies, 2014–2024 (full out-of-sample period).

Across the training period (2005–2014), all strategies appear to perform well, but this outcome must be interpreted with care. The results are structurally biased because the signals were designed and selected using this very window, and also because the regime itself is exceptionally favourable: yields decline almost monotonically for a decade, lifting returns across the entire curve. In such an environment even a naïve equal-weighted portfolio achieves a relatively good Sharpe ratio, which already indicates that limited timing skill is required to perform well.

In the 2014–2019 window, performance becomes more heterogeneous. Long-duration buckets lead the market, and momentum (especially momentum2) continues to perform well, whereas value and NSS weaken materially. Their mean-reversion logic fails in a regime where yields move smoothly in one direction. Carry also remains effective due to its systematic tilt toward higher-yielding segments.

The 2019–2024 period is characterized by inflation shocks and rapid tightening cycles. All duration buckets generate negative returns, and most factor signals follow the same pattern. Momentum breaks down as trends repeatedly reverse; value and NSS remain weak throughout. Only long-horizon momentum (momentum3) shows limited relative resilience, though still negative overall. In a broad rates sell-off, long-only duration strategies struggle to produce positive returns, so a meaningful outcome can simply be to design a strategy that reduces the drawdown.

Over the full evaluation horizon (2014–2024), combining the benign early subperiod with the highly adverse later years, most signals average out to flat or mildly negative performance. Carry is the best performing factor (in terms of Sharpe ratio and cumulative returns), while long-horizon momentum again appears the least fragile (with the best sharpe from 2019 to 2024). The returns and allocations of these two strategies are shown in detail in Figures 2 and 3.

Across the training window (2005–2014), the strong performance of value and momentum should be viewed in the context of the regime itself. Much of the empirical literature documenting the strength of these signals was built on long samples dominated by yield declines. The training period here shares exactly that structure, so the strong performance of value and momentum is unsurprising and partly reflects the same historical bias: these signals were never stress-tested against a decade like the post-2014 period.

That said, even after accounting for this regime effect, certain conclusions remain robust. Carry and momentum strategies consistently outperform the equal-weighted benchmark both in the training window and in the more challenging evaluation period. They allocate more effectively across the curve, avoid the structural weaknesses of simple static weighting, and preserve relative performance even when conditions turn unfavourable for duration overall. For this reason, carry and momentum emerge as the most reliable pure-factor approaches and clearly preferable to an equal-weight allocation.

Figure 2: Top row: cumulative returns of the carry portfolio and the four WGBI buckets plus the equal–weighted benchmark. Bottom row: portfolio weights, showing persistent allocation to the highest–yielding bucket except when relative yields compress or invert.

Figure 3: Top row: cumulative returns of momentum3 portfolio versus benchmarks. Bottom row: portfolio weights implied by the 6-month momentum signal, with slow regime shifts and delayed reactions around turning points.

4.3 Combination of signals

As discussed in Section 4.2, each factor reflects a distinct economic mechanism. Carry assumes the yield curve remains locally stable; value exploits mean reversion in yield levels; momentum captures the empirical persistence of recent price trends.

For the composite model, we retain the following three signals:

- *carry*;
- *value* (rolling historical version, not NSS, since it delivers stronger performance in the training sample and is more robust);
- *momentum2* (1-month cumulative return), which we select in place of the directional-vote momentum or the long-horizon variants. Momentum1 suffers from the issue that it might assign identical signals to all maturities and the very Momentum2 offers the most attractive combination of Sharpe ratio and cumulative return in the training window.

a. Equal-weighted composite

The baseline combination simply averages the three standardised signals:

$$s_{b,t}^{\text{total}} = \frac{1}{3}\Big(\tilde{s}_{b,t}^{\text{carry}} + \tilde{s}_{b,t}^{\text{value}} + \tilde{s}_{b,t}^{\text{momentum2}}\Big).$$

This procedure allows the model to automatically favour whichever signal is strongest at a given time.

b. Adaptive Combination via Information Coefficients

The equal-weighted composite assumes that the economic interpretation of each factor is stable through time. However, this need not be the case. For example, value relies on mean reversion, but during persistent trending regimes deviations from the average may actually widen rather than revert. To account for these possible sign reversals or regime dependencies, we build an adaptive weighting scheme using daily *Information Coefficients* (which we will refer to as ICs). The procedure consists of the following steps:

1. **Forward 21-day returns.** For each bucket b, compute the forward cumulative 21-day return:
$$r_{b,t}^{(21)} = \prod_{k=1}^{21}(1 + r_{b,t+k}) - 1.$$
2. **Daily cross-sectional IC.** The predictive ability of each factor f is measured by the daily cross-sectional correlation between its signal and the subsequent returns:
$$IC_t^{(f)} = \operatorname{corr}\big(s_{b,t}^{(f)},\, r_{b,t}^{(21)}\big).$$
3. **Rolling IC averages.** To extract persistent information and reduce noise, compute rolling averages (e.g., over 100 days):
$$\overline{IC}_t^{(f)} = \frac{1}{N}\sum_{i=1}^{N} IC_{t-21-i}^{(f)}.$$

4. **IC-based factor weights.** Normalise the averaged ICs so that the composite depends on each factor's relative predictive strength:

$$\varepsilon_t^{(f)} = \frac{\overline{IC}_t^{(f)}}{\sum_g |\overline{IC}_t^{(g)}|}.$$

5. **Adaptive composite signal.** For each bucket, form an IC-weighted composite:

$$s_{b,t}^{IC} = \varepsilon_t^{\text{carry}}\, \tilde{s}_{b,t}^{\text{carry}} + \varepsilon_t^{\text{value}}\, \tilde{s}_{b,t}^{\text{value}} + \varepsilon_t^{\text{momentum2}}\, \tilde{s}_{b,t}^{\text{momentum2}}.$$

c. Combination via Logistic regression

Some predictive content may lie in the short-term dynamics of value, momentum, and carry themselves. Their recent changes (for example over 5 or 20 days) can contain information that is not visible in the raw levels, and as noted earlier certain signals may relate inversely to subsequent returns. For this reason we also test a logistic regression with light feature engineering. The features used are deliberately simple and constructed only from the two extreme buckets (1–3y and 10y+): the raw relative spreads $\text{carry}_{10y+} - \text{carry}_{1\text{–}3y}$, $\text{value}_{10y+} - \text{value}_{1\text{–}3y}$, $\text{mom}_{10y+} - \text{mom}_{1\text{–}3y}$, together with their short-term changes computed over 5 days and over 20 days. These quantities capture both the cross-sectional gap between the long- and short-duration buckets and the recent evolution of that gap.

The model is trained to predict which of the two duration buckets will outperform over the subsequent 21 days. The portfolio weights are then obtained directly from the model's output probabilities, so that the allocation reflects the estimated likelihood that each bucket will deliver the higher forward return.

A broader exploration of feature sets and more sophisticated machine learning models is intentionally left for later, since the aim of this section is to evaluate the baseline predictive content of the underlying factors. We apply logistic regression only to the two extreme buckets, the lowest-duration and the highest-duration portfolios, rather than using a multinomial model across all four buckets. The rationale for this choice is discussed in Section 5.3 as the details of the training.

d. Results

We now evaluate the performance of these strategies using Tables 6, 7, 8, and 9. We also include carry, value, and momentum2 as reference points for combined strategies.

Training Sample (2005–2014)

Strategy	Cum.	Sharpe	Sharpe/turn
equal weighted	0.375	1.223	1.223
carry	0.640	0.961	0.961
value	0.398	1.241	1.237
momentum2	0.589	1.213	1.181
signal comb	0.572	1.279	1.248
IC	0.499	1.070	1.058
logistic reg	0.363	0.851	0.835

Table 6: Combined factor strategies, 2005–2014 (training sample).

Evaluation Period I (2014–2019)

Strategy	Cum.	Sharpe	Sharpe/turn
equal weighted	0.121	0.784	0.784
carry	0.362	0.993	0.991
value	0.038	0.549	0.539
momentum2	0.362	1.276	1.249
signal comb	0.249	1.243	1.208
IC	0.214	0.750	0.739
logistic reg	0.227	0.887	0.870

Table 7: Combined factor strategies, 2014–2019 (evaluation period I).

Evaluation Period II (2019–2024)

Strategy	Cum.	Sharpe	Sharpe/turn
equal weighted	−0.102	−0.410	−0.410
carry	−0.077	−0.202	−0.205
value	−0.098	−0.590	−0.597
momentum2	−0.108	−0.340	−0.367
signal comb	−0.044	−0.187	−0.208
IC	0.023	0.108	0.094
logistic reg	−0.117	−0.344	−0.353

Table 8: Combined factor strategies, 2019–2024 (evaluation period II).

Full Evaluation Sample (2014–2024)

Strategy	Cum.	Sharpe	Sharpe/turn
equal weighted	−0.027	−0.055	−0.055
carry	0.198	0.318	0.316
value	−0.077	−0.319	−0.327
momentum2	0.059	0.128	0.103
signal comb	0.120	0.307	0.279
IC	0.112	0.254	0.241
logistic reg	−0.020	−0.014	−0.026

Table 9: Combined factor strategies, 2014–2024 (full out-of-sample period).

Signal combined shows a stable and intuitive pattern across all periods. In the training sample it performs close to the strongest individual signals, with Sharpe just below momentum2 but above carry and value. In 2014–2019 it improves further: Sharpe 1.24, broadly in line with momentum2 and much higher than value and carry. In the 2019-2024 it is able to limit the losses compared to the single factors. This confirms that averaging the three signals produces a robust profile without depending on a single factor.

IC behaves differently because its weights are recalibrated based on which factor is actually predictive in each window. In the training and early-evaluation periods it is slightly weaker than signal combined but still competitive, with Sharpe ratios around 1.07 (training) and 0.75 (2014–2019). The important point is the 2019–2024 cycle: this is the only sample where every duration bucket and every static factor turns negative. Here signal combined also suffers (Sharpe –0.19), while IC remains slightly positive (Sharpe 0.11). The table makes this clear: IC is the only approach with a positive Sharpe in the sell-off regime.

Over the full 2014–2024 sample, the two approaches converge. Signal combined has a Sharpe of 0.31 and IC 0.25, both well above value and momentum2, and slightly

Figure 4: IC–based composite strategy on WGBI duration buckets. Top row: cumulative returns versus benchmarks. Second row: portfolio weights implied by the adaptive IC signal. Third row: violin plot of IC–weighted signals across buckets. Bottom row: time series of rolling information coefficients for carry, value, and momentum, which determine the dynamic factor weights used in the composite signal.

below carry. This shows that (i) static diversification across carry, value and momentum is already strong and stable, and (ii) the IC mechanism adds resilience in stressed periods by down-weighting failing signals and leaning on whichever still contains useful information.

We can use the third and fourth row of Figure 4 to diagnose the predictive performance of value, carry and momentum. Carry exhibits consistently positive information coefficients across all periods, confirming that its relative-yield signal maintains a stable relation with forward returns. Momentum also remains mostly positively correlated with future performance, with only moderate variation in stressed regimes. Value, by contrast, flips sign repeatedly: in some periods it correlates positively, in others nega-

tively. This instability indicates that value is not a reliable standalone predictor and that its forecasting direction is regime-dependent.

The time-series of rolling ICs in the Fourth row reinforces the same point. Carry's IC stays above zero almost everywhere, momentum fluctuates more but remains predominantly positive, especially during trending phases. Value, however, oscillates around zero with extended negative stretches, reflecting reversals in its economic interpretation. These dynamics justify the need for an adaptive mechanism like the very IC-based weighting.

Finally, the logistic regression behaves worse than expected. By construction, one would anticipate that a model that updates weights based on recent factor behaviour should at least limit drawdowns relative to static combinations, yet this is not what we observe. The strategy fails to protect capital in the 2019–2024 regime and does not deliver a clear improvement in the earlier periods either. This suggests that, in its current form, the machine learning layer is either too weak to extract additional structure from the inputs, or it is overfitting noise given how light the feature set is. In the next section we will see that the main issue is not the use of logistic regression per se, but the lack of sufficiently rich and informative features.

4.4 Predictive Power of PCA

In this subsection, we test whether principal components extracted from the yield curve contain incremental predictive information. As established in the literature (e.g., [7]), PCA applied to yields typically produces factors interpreted as *level*, *slope*, and *curvature*. These components summarise the geometry of the term structure and provide a compact description of whether the curve is normal, inverted, or humped.

To exploit this information, we compute PCA using a rolling window of one year (approximately 250 trading days) of yield data and project the yield vector observed on the day prior to each rebalancing date. We deliberately rely on a short rolling window rather than the full available history, as the objective is to capture short-horizon changes in curve shape rather than long-run average yield levels.

Interpretation of the PCA scores is standard:

- a negative PC2 is typically associated with an inverted curve (short rates above long rates);
- a positive PC3 is indicative of a humped curve, where intermediate maturities yield more than both short and long ends.

An alternative would have been to fit the NSS curve at each date using the available information, but we did not explore this option due to practical and computational constraints, and because it provides only point-in-time information without linking the curve's shape to past observations, making it not directly comparable to our approach.

a. Modification of Carry

We use the information extracted from PCA to construct a simple modification of the carry signal. Whenever the second principal component falls below -0.2 (a threshold selected on the training sample), we flip the sign of the carry signal.

The intuition is that a normally sloped yield curve tends to persist, whereas inversions typically arise during stress episodes and represent an unstable configuration that policymakers often attempt to reverse. By inverting carry in such regimes, the strategy explicitly positions for a re-normalisation of the yield-curve slope.

b. PCA threshold strategy

This second PCA–based strategy uses the shape of the yield curve to guide duration allocation, without relying on traditional factor signals. At each rebalancing date, we observe the values of PC2 and PC3 (computed from the previous one year of yield data) and classify the yield curve into three simple regimes:

- **Humped curve (**$\mathrm{PC3} > 0.2$**).** When the curve is humped, intermediate maturities typically offer the most attractive risk–return trade-off. The strategy allocates 50% to the 3–5y bucket and 50% to the 5–10y bucket.
- **Inverted curve** ($\mathrm{PC2} < -0.2$ and $\mathrm{PC3} < 0.2$). Under inversion, short maturities dominate: they carry higher yields and usually benefit most when rates eventually decline. The portfolio is fully allocated to the 1–3y bucket.
- **Normal curve (otherwise).** When the curve is upward sloping, long maturities generally deliver the strongest carry. The portfolio allocates entirely to the 10y+ bucket.

In essence, this strategy converts PCA information into a *curve–shape switch*: depending on whether the yield curve is normal, inverted, or humped, duration exposure is shifted toward the segment that historically performs best in that environment. This results in a simple and highly interpretable regime–based allocation rule driven directly by the geometry of the term structure.

We then extend this PCA–regime strategy by allowing the portfolio weights within each curve configuration to be parametrised rather than fixed. Each regime is associated with a vector of weights of the form:

- *Humped curve* ($\mathrm{PC3} > 0.2$): weights $[0,\ \alpha,\ 1-\alpha,\ 0]$.
- *Inverted curve* ($\mathrm{PC2} < -0.2$ and $\mathrm{PC3} < 0.2$): weights $[\gamma,\ 1-\gamma,\ 0,\ 0]$.
- *Normal curve* (otherwise): weights $[0,\ 0,\ 1-\delta,\ \delta]$.

A grid search over (α, γ, δ) is performed on the training sample. For each parameter triplet, the full strategy is simulated and its Sharpe ratio (optionally adjusted for turnover) is computed. The parameter set that maximises performance is retained. In the tables below, we refer to the unoptimised and optimised versions of this approach as *Full PCA* and *Full PCA**, respectively.

c. Logistic Regression with yield curve information

We expect signals to exhibit similar behaviour when the yield curve is in a comparable state. To exploit this, we split the sample into two regimes based on the sign and magnitude of the second principal component. PCA is computed exactly as before, and at each rebalancing date we classify the current curve using the previous day's PCA values. We then train the logistic regression on a restricted subsample consisting only of days in which the curve was in the same regime. Concretely, if today's state satisfies $\text{PC2} < -0.2$, the model is trained only on past observations with $\text{PC2} < -0.2$; the same logic applies symmetrically when the curve is on the opposite side.

d. Results

We now evaluate the performance of these strategies using Tables 10, 11, 12, and 13. We also include carry and logistic regression strategies as reference points for comparison.

Training Sample (2005–2014)

Strategy	Cum.	Sharpe	Sharpe/turn
equal weighted	0.375	1.223	1.223
carry	0.640	0.961	0.961
carry mod	0.613	1.149	1.131
Full PCA	0.507	1.049	1.021
Full PCA*	0.515	1.137	1.109
logistic reg	0.363	0.851	0.835
log reg PCA	0.392	0.883	0.865

Table 10: PCA-based strategies, 2005–2014 (training sample).

Evaluation Period I (2014–2019)

Strategy	Cum.	Sharpe	Sharpe/turn
equal weighted	0.121	0.784	0.784
carry	0.362	0.993	0.991
carry mod	0.359	1.226	1.200
Full PCA	0.298	1.105	1.070
Full PCA*	0.274	1.117	1.083
logistic reg	0.227	0.887	0.870
log reg PCA	0.294	1.116	1.097

Table 11: PCA-based strategies, 2014–2019 (evaluation period I).

Evaluation Period II (2019–2024)

Strategy	Cum.	Sharpe	Sharpe/turn
equal weighted	−0.102	−0.410	−0.410
carry	−0.077	−0.202	−0.205
carry mod	−0.025	−0.049	−0.060
Full PCA	−0.144	−0.329	−0.343
Full PCA*	−0.132	−0.334	−0.348
logistic reg	−0.117	−0.344	−0.353
log reg PCA	−0.197	−0.604	−0.616

Table 12: PCA-based strategies, 2019–2024 (evaluation period II).

Full Evaluation Sample (2014–2024)

Strategy	Cum.	Sharpe	Sharpe/turn
equal weighted	−0.027	−0.055	−0.055
carry	0.198	0.318	0.316
carry mod	0.220	0.383	0.365
Full PCA	0.238	0.395	0.371
Full PCA*	0.189	0.346	0.324
logistic reg	−0.020	−0.014	−0.026
log reg PCA	0.038	0.093	0.079

Table 13: PCA-based strategies, 2014–2024 (full out-of-sample period).

The modification of Carry (see Figure 5) is consistently more robust than the standard carry strategy. In the training window it achieves a higher Sharpe ratio, in the first evaluation period it remains competitive or superior, and—most importantly—in the 2019–2024 sell-off regime it cuts losses substantially relative to plain carry. The PCA-based curve-shape adjustment allows the strategy to avoid the typical failure of carry

Figure 5: Top row: cumulative returns of the carry modified portfolio and the four WGBI buckets plus the equal–weighted benchmark. Bottom row: portfolio weights of the carry-modified strategy. Unlike standard carry, the allocation shifts frequently and tracks the regime indicated by PC2.

when the curve inverts or flattens, producing a far more stable behaviour across regimes.

PCA-based strategies deliver the highest full-sample Sharpe ratio over 2014–2024. However, their weakness is evident in the second evaluation window: both specifications react too slowly to the rapid rate-hiking cycle and experience deep drawdowns. This shows that, although highly informative on average, they are not robust in fast macro rotations.

Despite this, the underlying PCA signal remains powerful: it cleanly captures level, slope, and curvature dynamics and provides structural predictive information about the term structure. The poor short-horizon robustness does not diminish its value; rather, it suggests that a PCA-driven signal can serve as a foundation for more sophisticated and potentially superior strategies.

The optimised PCA strategy performs worse than the naïve PCA version because, ex post, we know that regimes change abruptly. This reinforces the point that robustness matters more than squeezing out marginal in-sample gains. A strategy tuned too closely to historical conditions collapses as soon as the environment shifts, whereas a more restrained, less overfit specification degrades more slowly and remains usable across regimes.

Finally, logistic regression with PCA performs slightly better than the standard logistic model in the training sample and over the full 2014–2024 period. The improvement is small, and the Sharpe ratios should be read cautiously: when returns hover around zero, minor changes in volatility can move the Sharpe up or down without indicating real predictive power. The method is not robust in 2019–2024, where per-

formance drops sharply. The post-pandemic sell-off regime has few precedents in the historical sample, so a regime-based classifier has little reliable structure to learn from.

4.5 Comments on Factor strategies

Factor strategies show clear predictive content: several of them outperform the benchmark, even if most do not generate strong Sharpe ratios. With only yields and past returns, achieving a positive Sharpe during broad fixed-income sell-offs is extremely difficult; in our tests, only the IC approach manages to do so. Momentum appears explicitly as a signal and implicitly inside both logistic models and the IC procedure, where the link between the most recent observation and the current state drives the forecast. This reinforces that momentum is a genuinely informative feature.

The final section highlights that PCA-based strategies carry meaningful predictive power. PC2, in particular, is an effective indicator of the curve's shape and underlying regime. For an asset manager, the implication is straightforward: valuable information is embedded in individual factors, but using them in isolation is insufficient. Combining signals, through IC weighting, signal averaging, or PCA-based curve adjustments such as the modification of carry, produces materially stronger and more reliable behaviour. These combined methods deliver the best full-period performance relative to the benchmark and remain notably more resilient in major sell-offs, showing smaller drawdowns and clearly superior Sharpe ratios compared with an equal-weighted allocation.

5 Strategies on the US Government Bond Index

5.1 Overview

In this section we examine duration-timing strategies in the US Treasury market. Returns are now taken in dollars to eliminate currency exposure. We first evaluate the performance of the factor signals developed in the previous section. Because the US market embeds far more macroeconomic information in yields than the global WGBI universe, where the US accounts for only about 44% of the index, we expect the standalone factors to be less effective here. This motivates extending the analysis by incorporating macro and market signals into additional strategies.

5.2 Analysis of Factor Strategies

By inspecting Tables 14, 15, 16, and 17, the shortest-duration bucket emerges as the best-performing segment in the US data, displaying the highest Sharpe ratio. At first sight this contrasts with the WGBI results, where the shortest bucket seemed to be the weakest. The comparison is misleading: in both datasets the shortest bucket follows a largely monotonic path, but in the WGBI case the appreciation of the Swiss franc suppresses the observed returns on low-duration bonds (since we were using returns fully hedged in CHF), and their extremely low volatility makes the Sharpe ratio appear artificially unfavourable. In reality, the shortest bucket is exactly the segment that preserves capital during broad fixed-income selloffs in both setups.

Training Sample (2005–2014)

Strategy	Cum.	Sharpe	Sharpe/turn
US 1–3y	0.281	1.813	1.813
US 3–5y	0.474	1.126	1.126
US 5–10y	0.667	0.873	0.873
US 10+y	1.077	0.664	0.664
equal weighted	0.612	0.900	0.900
carry	0.999	0.690	0.690
carry mod	0.512	0.589	0.575
value	0.697	0.766	0.765
nss	0.821	0.718	0.710
momentum	0.599	0.837	0.819
momentum2	0.721	0.752	0.735
momentum3	0.586	0.560	0.555
momentum4	0.767	0.695	0.688
signal comb	0.720	0.744	0.731
IC	0.758	0.761	0.753
logistic reg	0.589	0.676	0.671
log reg PCA	0.741	0.819	0.813
Full PCA	0.379	0.430	0.415
Full PCA*	0.409	0.497	0.482

Table 14: US factor strategies, 2005–2014 (training sample).

Evaluation Period I (2014–2019)

Strategy	Cum.	Sharpe	Sharpe/turn
US 1–3y	0.077	1.422	1.422
US 3–5y	0.133	0.878	0.878
US 5–10y	0.217	0.775	0.775
US 10+y	0.542	0.683	0.683
equal weighted	0.236	0.791	0.791
carry	0.488	0.686	0.686
carry mod	0.403	0.851	0.842
value	0.181	1.135	1.130
nss	0.195	1.087	1.068
momentum	0.229	0.659	0.641
momentum2	0.391	0.755	0.738
momentum3	0.149	0.297	0.293
momentum4	0.152	0.306	0.299
signal comb	0.255	0.722	0.704
IC	0.310	0.624	0.617
logistic reg	0.432	0.852	0.847
log reg PCA	0.403	0.788	0.783
Full PCA	0.255	0.519	0.503
Full PCA*	0.235	0.544	0.528

Table 15: US factor strategies, 2014–2019 (evaluation period I).

Evaluation Period II (2019–2024)

Strategy	Cum.	Sharpe	Sharpe/turn
US 1–3y	0.110	1.023	1.023
US 3–5y	0.078	0.347	0.347
US 5–10y	0.039	0.131	0.131
US 10+y	−0.121	−0.055	−0.055
equal weighted	0.030	0.106	0.106
carry	0.019	0.085	0.084
carry mod	0.038	0.110	0.102
value	0.055	0.212	0.208
nss	0.076	0.344	0.331
momentum	−0.006	0.021	0.006
momentum2	−0.083	−0.090	−0.104
momentum3	0.123	0.241	0.238
momentum4	0.079	0.176	0.171
signal comb	0.092	0.210	0.199
IC	0.263	0.538	0.530
logistic reg	0.062	0.152	0.148
log reg PCA	0.114	0.253	0.249
Full PCA	−0.101	−0.071	−0.079
Full PCA*	−0.072	−0.052	−0.061

Table 16: US factor strategies, 2019–2024 (evaluation period II).

Full Evaluation Sample (2014–2024)

Strategy	Cum.	Sharpe	Sharpe/turn
US 1–3y	0.154	0.960	0.960
US 3–5y	0.161	0.437	0.437
US 5–10y	0.177	0.294	0.294
US 10+y	0.179	0.175	0.175
equal weighted	0.182	0.291	0.291
carry	0.360	0.301	0.300
carry mod	0.264	0.305	0.296
value	0.159	0.376	0.372
nss	0.231	0.573	0.558
momentum	0.212	0.304	0.289
momentum2	0.218	0.233	0.220
momentum3	0.141	0.173	0.169
momentum4	0.186	0.212	0.206
signal comb	0.128	0.205	0.189
IC	0.312	0.357	0.350
logistic reg	0.374	0.386	0.382
log reg PCA	0.481	0.460	0.455
Full PCA	0.271	0.279	0.266
Full PCA*	0.257	0.293	0.280

Table 17: US factor strategies, 2014–2024 (full out-of-sample period).

Conversely, in the US valuation sample the strong performance of the shortest bucket can obscure the true contribution of timing strategies: a naïve rule that always selects the lowest-duration bucket would look strong on the full sample while delivering little value in terms of active allocation. Since the objective is to maximise returns for a given level of risk, one cannot evaluate a timing strategy by inspecting the Sharpe ratio in isolation; one must control for the structural dominance of the low-duration bucket before assessing whether a strategy genuinely extracts information rather than merely reverting to the safest segment.

Given this, the only single-factor strategies that appear to outperform the benchmark in Sharpe terms are value and NSS. This outperformance, however, is largely mechanical: both strategies generate an almost static allocation concentrated in the shortest-duration buckets, as illustrated for NSS in Figure 6, with value exhibiting a very similar pattern. Such positioning provides protection during sell-offs but prevents the strategies from capturing returns in Evaluation Period I, when longer durations dominate performance. As a result, these approaches do not represent genuinely investable signals; they behave instead like static underweights to duration rather than adaptive timing strategies.

Figure 6: Top row: cumulative returns of the NSS portfolio and the benchmarks. Bottom row: portfolio weights. As we see the allocation is almost static on the lowest duration buckets, making this strategy not worth choosing despite the relative high sharpe ratio.

Carry and momentum do not outperform the benchmark in this setting. While both signals delivered clear timing value when applied to the full WGBI, their informational content weakens substantially when restricted to the US segment. Signal amplitudes compress, and the resulting strategies behave close to the benchmark, with limited duration-timing ability. This is expected: when focusing on a single market, much of the information exploited by carry and momentum at the global level is already

embedded in prices, reducing their marginal predictive power.

Because individual signals are weak, their combinations inherit the same limitation. With low single-factor informativity, aggregation alone cannot extract meaningful additional structure from the US curve.

The IC approach remains robust. It delivers strong full-sample performance and reacts appropriately during broad sell-offs, achieving higher Sharpe ratios and cumulative returns. PCA-based information is also valuable: the modified carry strategy outperforms both standard carry and the benchmark, and logistic-regression models augmented with PCA inputs dominate their counterparts without PCA. Among all approaches, the PCA-enhanced logistic models perform best, exhibiting the highest Sharpe ratios and the greatest robustness.

Overall, strategies are clearly less effective than in the WGBI setting. The US curve exhibits less structural dispersion, causing most signals to collapse toward benchmark-like behaviour. Among all methods, the IC approach stands out as the most reliable and economically interpretable: it performs consistently across regimes, is particularly resilient during sell-offs, and improves upon the equal-weighted benchmark even in favourable periods. Logistic regression achieves higher Sharpe ratios on average, but its advantage is less stable.

These results suggest that yield-curve information alone is insufficient to generate robust duration-timing value on the US curve. This naturally motivates extending the analysis beyond term-structure signals, introducing macroeconomic and market-based time-series information in the next section.

5.3 Incorporating Market and Macro signals with Machine Learning

a. Methodological choices and ML models

A natural way to incorporate additional signals and features is through machine-learning methods. Since there is no unambiguous target, we frame the problem as a classification task and train models to predict the bucket with the highest return over the next 21 trading days.

At each rebalancing date t, the target is constructed using 21-day forward returns computed only up to t-21, ensuring that no forward-looking information is used. Before selecting the models, we analyze the empirical properties of this target. Figure 7 reports the histogram of the best-performing buckets in the training sample. We can clearly see that the shortest-duration and the longest-duration buckets dominate the distribution, while intermediate maturities are selected far less frequently.

This effect becomes even more pronounced once we exclude borderline observations, defined as cases in which the return difference between the best and second-best buckets over the 21-day horizon is smaller than 20 basis points. After applying this margin filter, the intermediate-duration buckets are almost never selected, indicating that economically meaningful signals tend to correspond to a clear preference for either very short or very long duration.

For this reason, rather than training a multi-class classifier over all buckets, we focus on binary classification between the lowest-duration and highest-duration buckets.

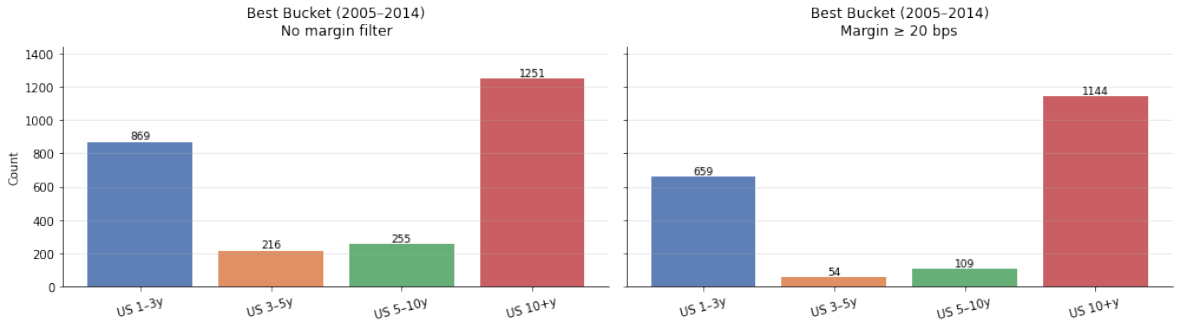

Figure 7: Histogram of the best-performing duration buckets based on 21-day forward returns in the training sample; excluding observations where the return spread between the best and second-best bucket is below 20 bps removes marginal cases and highlights that only the shortest- and longest-duration buckets remain dominant.

At each rebalancing date, models are retrained using the previous four years of daily data, corresponding to roughly 1,000 observations. However, this raw count is misleading: financial time series are highly correlated, and a substantial fraction of the information is therefore redundant. Once serial dependence is taken into account, together with the fact that some predictors are observed only at a monthly frequency, the effective number of independent observations is much smaller. Heuristically, it is closer to 50–100 data points per prediction, which imposes a severe constraint on the complexity of any model that can be reliably estimated.

In this setting, working with 15–20 time series and multiple transformations quickly leads to a feature space of roughly 75–100 variables, assuming that each time series is engineered into about 5 features. This places the problem in a regime in which even aggressive regularization is typically insufficient to prevent overfitting. For this reason, the only viable approach is to restrict attention to carefully selected subsets of factors rather than using the full feature set at once; in practice, we limit each subset to at most ten features. This choice is a necessity in our framework and is consistent with the empirical findings of [3], who adopt a similar strategy to address the low effective dimensionality induced by correlated financial time series. While this approach may initially discard interactions across different subsets, we later attempt to recover and combine this information at a subsequent stage of the analysis.

Given these constraints, each model is trained to predict a single rebalancing decision, and the resulting probability outputs are directly used to construct portfolio weights.

- **Logistic Regression (LR).** Used as a simple and transparent benchmark. It is interpretable, and relatively robust in small-sample settings.
- **Random Forests (RF).** Included to capture nonlinear effects and interactions across features, at the cost of higher model complexity and an increased risk of overfitting.

- **Gradient Boosting (XGB).** Designed to model complex nonlinearities and feature interactions more aggressively than RF, but also more sensitive to overfitting in data-constrained environments.

- **Support Vector Machines (SVM, Gaussian kernel).** Provide an intermediate level of flexibility, allowing for nonlinear decision boundaries while remaining more controlled than tree-based ensembles.

- **Stacked model.** Combines the probability outputs of RF, XGBoost, and SVM to aggregate their complementary strengths, with the caveat that stacking is intrinsically harder to regularize and more prone to overfitting given the limited effective sample size.

All hyperparameters are selected using the training sample only. In addition, models are estimated with exponentially decaying observation weights, so that more recent data receive higher importance. This choice is meant to improve adaptability to regime changes and allow the models to react more quickly to shifts in the underlying market environment.

b. Feature Engineering and Selection

As discussed, we partition the full information set into coherent subsets of predictors, based on similar economic interpretation. The used sets are:

- **Factors**: yield-curve and return-based signals capturing carry, value, momentum, and term-structure information (built as we did in the previous section).

- **FX CHF**: USD–CHF exchange-rate.

- **FX EUR**: USD–EUR exchange rate.

- **Gold**: gold price.

- **Macro**: mostly low-frequency macroeconomic variables capturing inflation dynamics and real-activity conditions. Specifically, we include CPI and PPI[1], the Federal Funds target rate (DFEDTAR) [2], survey-based expected inflation at the 1-, 5-, and 10-year horizons (EXPINF1YR, EXPINF5YR, EXPINF10YR), and a market-implied one-year inflation expectation change[3].

- **Market**: broad market indicators including the NASDAQ 100 (NDX), S&P 500 (SPX), VIX index, gold price (XAU), and the S&P GSCI commodity index, capturing global risk appetite, equity-market conditions, volatility, and commodity-cycle dynamics.

[1]Consumer Price Index and Producer Price Index. For each month, both series are aligned to the release date of the later of the two announcements (relative to the same month).

[2]We use the effective target (DFEDTAR) when available (up to 2008), and thereafter proxy it by the midpoint of the upper and lower bounds (DFEDTARU and DFEDTARL).

[3]It is constructed as the difference between the 1-year nominal Treasury yield (DGS1) and the yield on a 1-year inflation-linked zero-coupon bond.

- **Gold & Commodities**: gold price and the S&P GSCI commodity index, capturing safe-haven demand, inflation-hedging properties, global growth conditions, and commodity-cycle dynamics.
- **VIX**: equity-market volatility measure capturing changes in market uncertainty.
- **MOVE**: bond-market volatility index capturing uncertainty in interest-rate markets.
- **PCA**: principal components of the yield curve providing a low-dimensional representation of term-structure movements.

Each subset is then expanded through feature engineering, including trend and volatility measures, anomaly indicators, and simple interactions. Feature selection is carried out in two steps:

- **Correlation filtering**: highly collinear variables are first removed by excluding features with pairwise correlations above 0.90, leading to a more stable and interpretable feature set. See for example Figure 8.
- **Random Forest screening**: on the reduced feature set, a Random Forest is trained on realized returns over the training sample and the ten most important variables are retained according to feature-importance scores. At this stage, the Random Forest is used purely as a selection device rather than as a predictive model; flexibility and limited forward-looking (only on the training set) are allowed, since the objective is variable ranking rather than out-of-sample forecasting. See for example Figure 9.

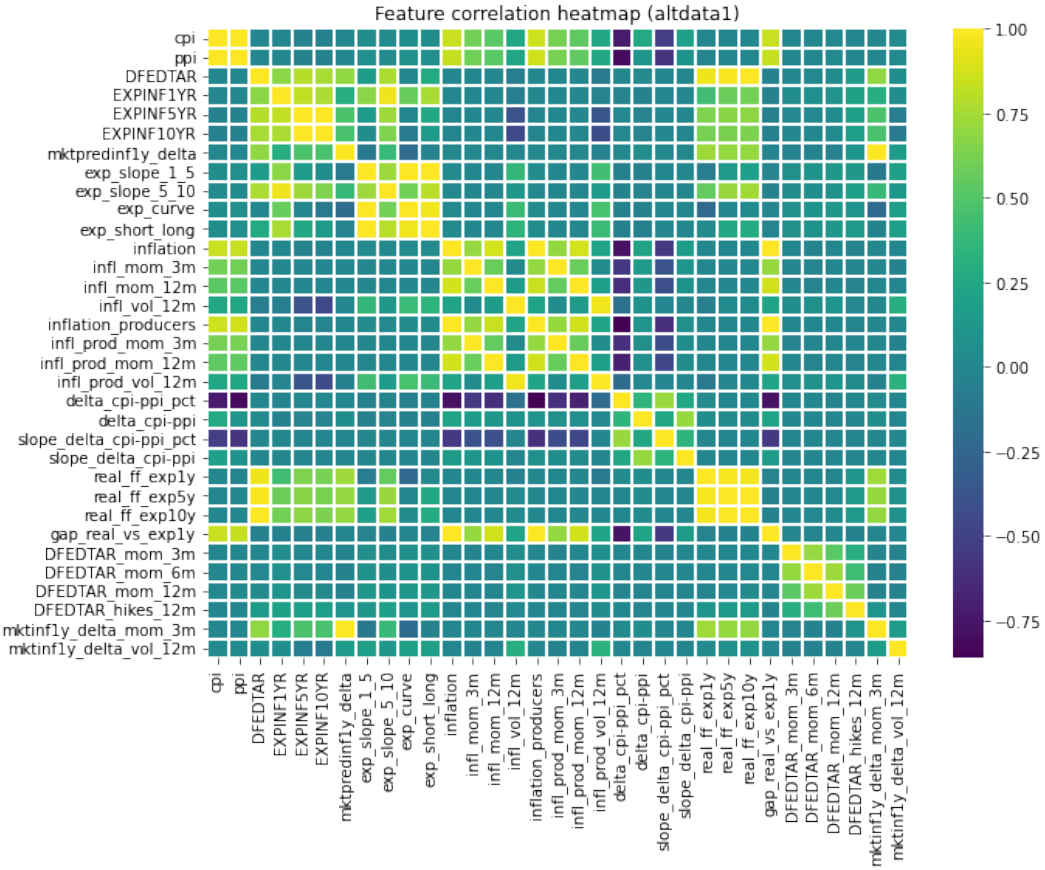

Figure 8: Correlation heatmap of the engineered macro/alternative features. Strong dependencies across variables motivate a correlation-based pruning step prior to model training.

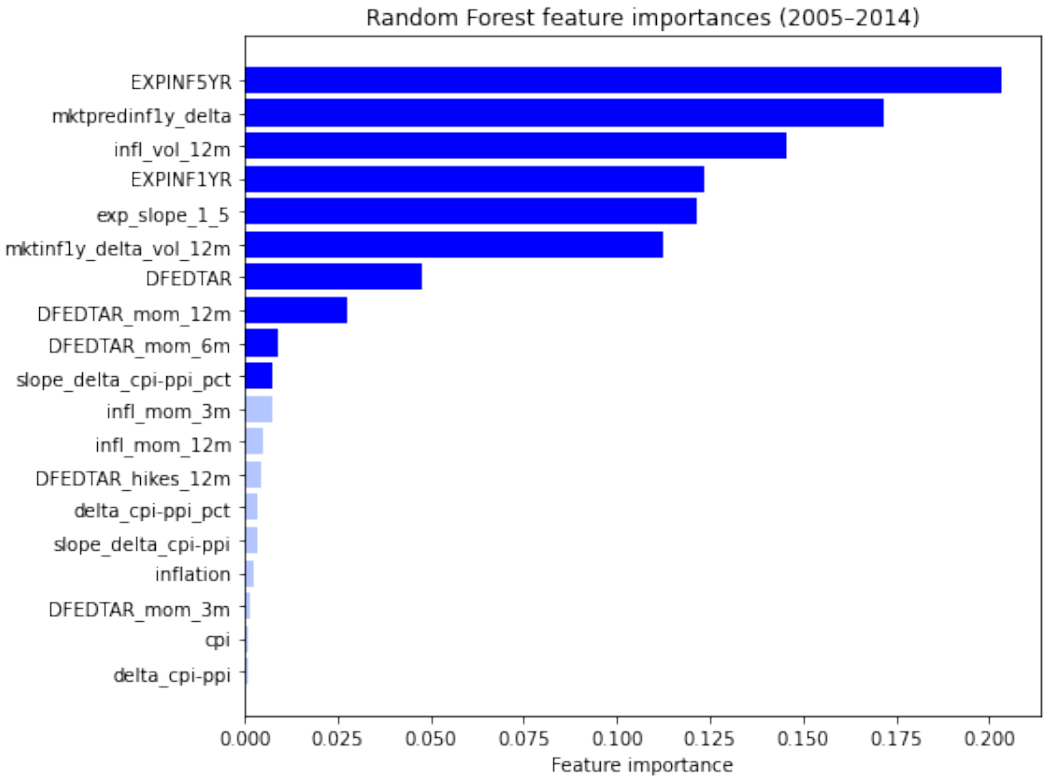

Figure 9: Random Forest feature importances estimated on the training sample (2005–2014) using only the subset of non-collinear features. Dark blue bars correspond to the features selected for subsequent modeling.

c. Results and Comments

Here we turn to the out-of-sample evaluation and analyze how the different strategies perform over the full evaluation period. Table 18 reports cumulative returns and Sharpe ratios computed over 2014–2024, allowing a direct comparison across models and information sets under identical market conditions. The table highlights which datasets and learning algorithms generalize best beyond the training window, and provides a clear benchmark against the equal-weighted strategy.

Full Evaluation Sample (2014–2024)

Table 18: Performance comparison across datasets and models, sorted by Sharpe ratio. Row colors: green = Forex chf, violet = Factor, yellow = Macro, blue = equal-weighted baseline.

Strategy	Dataset	Cum. Return	Sharpe
xgb	forex chf	0.505	0.479
rf	forex chf	0.410	0.457
svm	forex chf	0.363	0.405
lr	factor	0.417	0.405
svm	move	0.381	0.386

Continued on next page

Strategy	Dataset	Cum. Return	Sharpe
svm	macro	0.349	0.376
svm	factor	0.328	0.363
stack	gold_comm	0.314	0.347
stack	macro	0.278	0.321
xgb	forex eur	0.282	0.320
lr	macro	0.285	0.311
xgb	move	0.274	0.306
stack	market	0.264	0.305
equal weighted	baseline	0.182	0.291
lr	forex chf	0.232	0.301
stack	gold	0.245	0.298
svm	gold	0.232	0.280
stack	factor	0.232	0.271
svm	vix	0.214	0.267
svm	pca	0.228	0.267
rf	move	0.214	0.262
rf	vix	0.194	0.260
rf	factor	0.198	0.252
stack	forex chf	0.204	0.252
xgb	macro	0.223	0.251
rf	macro	0.215	0.249
stack	vix	0.200	0.245
stack	forex eur	0.187	0.240
stack	pca	0.190	0.240
lr	vix	0.172	0.232
lr	move	0.151	0.209
xgb	factor	0.162	0.203
stack	move	0.153	0.199
rf	forex eur	0.146	0.198
lr	gold	0.137	0.196
rf	market	0.138	0.193
lr	forex eur	0.130	0.181
xgb	vix	0.125	0.179
rf	gold	0.124	0.175
xgb	market	0.125	0.172
svm	forex eur	0.124	0.170
rf	pca	0.117	0.164
xgb	gold_comm	0.096	0.150
lr	pca	0.091	0.136
lr	market	0.081	0.128
svm	gold_comm	0.074	0.127
rf	gold_comm	0.055	0.102
svm	market	0.053	0.098
xgb	pca	0.002	0.045
lr	gold_comm	0.006	0.044
xgb	gold	−0.001	0.041

The main insights derived from Table 18 are summarised below:

- **Outperformance relative to the benchmark.** A large number of algorithm–dataset combinations outperform the equal-weighted portfolio, confirming that duration timing benefits from conditioning on additional information and from systematic decision rules rather than static allocations.

- **Strength of USD/CHF Forex signal.** USD/CHF-based datasets deliver consistently strong performance across different algorithms. This robustness suggests that FX information contains genuine predictive power that is largely invariant to the choice of learning algorithm.

- **CHF versus EUR and Gold signals.** The strong performance of CHF-based FX strategies stands in contrast to both EUR FX and gold-based signals. The EUR exchange rate is more directly exposed to euro-area macroeconomic conditions and monetary policy, causing EUR-based signals to reflect regional rather than global risk dynamics, which limits their relevance for U.S. duration timing.

  Gold-based signals, by contrast, tend to react with delay to macroeconomic and financial stress. As documented by [**?**], gold prices typically adjust after inflation surprises or policy shifts, behaving more as a contemporaneous hedge than a forward-looking predictor.

- **Overfitting risk in Factor and Macro datasets.** Factor- and Macro-based signals appear more prone to overfitting. Their strongest results are obtained with simpler or more controlled models (e.g. Logistic Regression and SVM), while more complex approaches (Random Forests, Gradient Boosting, stacking) tend to perform poorly, consistent with the limited effective sample size. Nevertheless, both Factor and Macro datasets consistently show predictive power and tend to function as intended: they capture economically meaningful information and deliver robust signals when model complexity is kept under control, confirming their usefulness as core inputs.

- **Ambiguous role of MOVE.** The MOVE dataset exhibits mixed predictive performance. It performs relatively well when paired with SVM and Gradient Boosting, but delivers weak or inconsistent results under Logistic Regression and Random Forests. Overall, its predictive content appears fragile and model-dependent, making its economic role less clear than that of FX or factor-based signals.

- **Weak performance of Market, VIX, and Commodities.** Market-wide indicators, volatility measures (VIX), and commodity-based datasets—especially gold—exhibit relatively weak performance. This suggests limited incremental information for duration timing, despite their traditional role as risk or safe-haven indicators.

- **Sharpe ratios and economic significance.** Higher Sharpe ratios generally coincide with higher cumulative returns. This indicates that performance improvements are not driven solely by volatility compression but correspond to economically meaningful excess returns.

When paired with the appropriate information sets and sufficiently constrained algorithms, machine learning methods add value relative to using only handcrafted factor signals. In particular, USD/CHF FX, Factor, and Macro datasets deliver the strongest and most consistent performance, while other datasets provide limited incremental information. Simpler models tend to be preferable: moderate non-linear flexibility (e.g.

SVMs) improves results, while Logistic Regression remains competitive because most interactions are already encoded through feature engineering.

Predictive power is nonetheless state-dependent. The relevance of a given dataset varies across macroeconomic regimes, and some information sets perform primarily during stress episodes while underperforming simpler factor signals in benign environments. Since ML models may adapt slowly to such regime shifts, we also introduce threshold-based strategies that condition directly on current indicator levels rather than on rolling performance. The next sections first present these threshold strategies and then introduce a systematic procedure for selecting datasets and algorithms based on recent performance (see Section 5.5).

5.4 Incorporating Market and Macro Signals with Threshold-Based Strategies

a. Threshold-Based Strategies

These strategies are intentionally simple and rule-based, aiming to limit overfitting and maintain economic interpretability. Given the limited data availability, more complex specifications cannot be backtested reliably; threshold-based rules therefore offer a clear and robust way to incorporate market and macro information without relying on in-sample noise. They can be viewed as the straightforward counterpart to the machine-learning strategies discussed earlier, serving as a low-complexity benchmark based on explicit rules.

Their design entails clear limitations. First, these strategies perform poorly in benign, falling-yield environments (as we see from the results on the training period), where carry dominates and defensive positioning is penalised. This is structural rather than accidental: the rules are intended for protection and regime identification, not for systematic carry harvesting. Second, thresholds are calibrated once on the training sample and then kept fixed, introducing calibration risk as regime boundaries may drift over time.

Accordingly, these strategies are most effective during sharp sell-offs or periods of elevated stress, and tend to underperform in sustained falling-yield regimes where simpler factor-based signals prevail.

We tested the following threshold-based strategies:

- **Fed-rate threshold strategy.**
  - Compute the 3-month change in the policy rate, Δr_{3m}.
  - If $\Delta r_{3m} > 0$: allocate 100% to 1–3y.
  - If $\Delta r_{3m} \leq -0.5$: allocate 50% to 3–5y and 50% to 5–10y.
  - Otherwise: allocate 100% to 10+y.

  Economic rationale: policy tightening raises short-end risk and favors short duration; strong easing supports intermediate maturities, while stable policy environments favor long duration through carry.

- **CPI–PPI sign strategy.**
  - Compute YoY inflation rates CPI_{YoY} and PPI_{YoY}.
  - If $\text{sign}(\text{CPI}_{\text{YoY}}) = \text{sign}(\text{PPI}_{\text{YoY}})$: allocate 100% to 10+y.
  - Otherwise: allocate 100% to 1–3y.

  Economic rationale: consistent CPI and PPI signals indicate coherent inflation dynamics, supporting long duration; divergence increases inflation uncertainty and favors short duration.

- **CPI level–momentum threshold strategy.**
  - Let π_t denote CPI YoY inflation and $\overline{\pi}_t^{(3)}$ its 3-month moving average.
  - Define momentum $z_t = \pi_t - \overline{\pi}_t^{(3)}$.
  - If $\pi_t \leq 2.0$ and $z_t \leq 0$: allocate 100% to 10+y.
  - If $\pi_t \geq 3.0$ and $z_t \geq 0$: allocate 100% to 1–3y.
  - Otherwise: allocate 50% to 3–5y and 50% to 5–10y.

  Economic rationale: low and decelerating inflation supports long duration, while high and accelerating inflation increases tightening risk and favors short duration.

- **FX CHF z-score threshold strategy.**
  - Define USD/CHF returns $r_t = \Delta(\text{USDCHF})_t$ and CHF strength

$$s_t = -\sum_{j=1}^{20} r_{t-j}.$$

  - Standardize using a 252-day rolling window to obtain a lagged z-score z_t.
  - If $z_t \geq 0.5$: allocate 100% to 10+y.
  - If $z_t \leq -0.5$: allocate 100% to 1–3y.
  - Otherwise: allocate 50% to 3–5y and 50% to 5–10y.

  Economic rationale: CHF strength signals risk-off conditions and declining yields, favoring long duration; CHF weakness reflects risk-on regimes and supports shorter duration.

- **VIX level threshold strategy.**
  - Let VIX_t denote equity-market volatility.
  - If $\text{VIX}_t \leq 18$: allocate 100% to 10+y.
  - If $\text{VIX}_t \geq 25$: allocate 100% to 1–3y.
  - Otherwise: allocate 50% to 3–5y and 50% to 5–10y.

  Economic rationale: low volatility reflects stable financial conditions, supporting long duration; elevated volatility signals stress and favors short duration.

- **MOVE level threshold strategy.**
  - Let MOVE_t denote bond-market volatility.
  - If $\text{MOVE}_t \leq 70$: allocate 100% to 10+y.
  - If $\text{MOVE}_t \geq 100$: allocate 100% to 1–3y.
  - Otherwise: allocate 50% to 3–5y and 50% to 5–10y.

  Economic rationale: low rate volatility supports long duration, while high uncertainty about interest rates favors short duration.

- **MOVE level + spike overlay strategy.**
  - Apply the MOVE level rule above.
  - Define a k-day change $\Delta_{10}\text{MOVE}_t = \text{MOVE}_t - \text{MOVE}_{t-10}$.
  - If $\Delta_{10}\text{MOVE}_t \geq 10$: override and allocate 100% to 1–3y.

  Economic rationale: sudden jumps in bond-market volatility signal abrupt repricing of rate risk and justify an immediate defensive shift to short duration.

- **PCA (PC2/PC3) threshold strategy.**
  - Use lagged yield-curve principal components $(\text{PC2}_{t-1}, \text{PC3}_{t-1})$.
  - If $\text{PC3}_{t-1} > 0.2$: allocate to the belly of the curve (3–5y and 5–10y).
  - If $\text{PC2}_{t-1} < -0.2$: allocate to the short end (1–3y).
  - Otherwise: allocate mostly to long duration (10+y).

  Economic rationale: PC2 and PC3 capture slope and curvature regimes of the yield curve, guiding allocation across short, intermediate, and long maturities.

- **Expected-inflation slope threshold strategy.**
  - Define $d_t = \text{EXPINF1YR}_t - \text{EXPINF10YR}_t$.
  - If $d_t > 0$: allocate 100% to 1–3y.
  - If $d_t < -0.3$: allocate 100% to 10+y.
  - Otherwise: allocate 50% to 3–5y and 50% to 5–10y.

  Economic rationale: front-loaded inflation expectations increase tightening risk and favor short duration, while declining short-term expectations support long duration.

b. Results and comments

Training Sample (2005–2014)

Strategy	Cum.	Sharpe
US 1–3y	0.281	1.813
US 3–5y	0.474	1.126
US 5–10y	0.667	0.873
US 10+y	1.077	0.664
Equal-weighted	0.612	0.900
MOVE+Spike	0.671	0.816
CPI–PPI	1.297	0.785
Fed-rate	0.983	0.668
VIX	0.615	0.666
CPI Level–Mom	1.077	0.664
MOVE	0.452	0.643
Exp. Infl.	0.216	0.328
PCA	0.409	0.497
FX CHF	0.398	0.523

Table 19: Threshold-based strategies, training sample (2005–2014).

Evaluation Period I (2014–2019)

Strategy	Cum.	Sharpe
US 1–3y	0.077	1.422
US 3–5y	0.133	0.878
US 5–10y	0.217	0.775
US 10+y	0.542	0.683
Equal-weighted	0.236	0.791
MOVE+Spike	0.599	0.869
CPI–PPI	0.775	0.933
Fed-rate	0.460	0.660
VIX	0.466	0.678
CPI Level–Mom	0.542	0.683
MOVE	0.456	0.781
Exp. Infl.	0.277	0.691
PCA	0.234	0.541
FX CHF	0.080	0.227

Table 20: Threshold-based strategies, evaluation period I (2014–2019).

Evaluation Period II (2019–2024)

Strategy	Cum.	Sharpe
US 1–3y	0.110	1.023
US 3–5y	0.078	0.347
US 5–10y	0.039	0.131
US 10+y	−0.121	−0.055
Equal-weighted	0.030	0.106
MOVE+Spike	0.226	0.391
CPI–PPI	−0.377	−0.477
Fed-rate	0.158	0.256
VIX	0.227	0.346
CPI Level–Mom	−0.121	−0.055
MOVE	0.207	0.361
Exp. Infl.	0.176	0.714
PCA	−0.072	−0.052
FX CHF	0.010	0.061

Table 21: Threshold-based strategies, evaluation period II (2019–2024).

Full Evaluation Sample (2014–2024)

Strategy	Cum.	Sharpe
US 1–3y	0.154	0.960
US 3–5y	0.161	0.437
US 5–10y	0.177	0.294
US 10+y	0.179	0.175
Equal-weighted	0.182	0.291
MOVE+Spike	0.570	0.509
CPI–PPI	0.656	0.407
Fed-rate	0.471	0.356
VIX	0.366	0.338
CPI Level–Mom	0.179	0.175
MOVE	0.445	0.449
Exp. Infl.	0.406	0.619
PCA	0.256	0.293
FX CHF	0.106	0.143

Table 22: Threshold-based strategies, full evaluation sample (2014–2024).

The results highlight a clear asymmetry between regimes (Tables 19–22). In the training sample (2005–2014, Table 19), several threshold strategies deliver high cumulative returns but generally lower Sharpe ratios than the equal-weighted benchmark, indicating that their gains come at the cost of higher volatility and less stable performance. This already signals that these rules are not designed to dominate in smooth, carry-driven environments.

In Evaluation Period I (2014–2019, Table 20), when yields were broadly falling and volatility was low, the equal-weighted portfolio remains a strong benchmark. Only a small subset of strategies, most notably MOVE+Spike and CPI–PPI, outperform it on both Sharpe and returns, while most others add limited incremental value.

In Evaluation Period II (2019–2024, Table 21), characterized by sharp repricing, inflation shocks, and volatility spikes, the picture reverses. Several threshold strategies (MOVE+Spike, Fed-rate, VIX, MOVE, Expected Inflation) clearly dominate the equal-weighted benchmark on both dimensions. This confirms that these rules are effective precisely in stress and sell-off regimes, where rapid changes in uncertainty and policy expectations matter most.

Over the full evaluation period (2014–2024, Table 22), the same pattern persists: volatility- and inflation-based strategies systematically outperform the benchmark, while FX-based and PCA-based rules remain weaker.

Overall, these results support the interpretation that threshold strategies are not substitutes for factor timing, but rather complements: they underperform in stable, trend-driven markets, yet provide substantial value in turbulent regimes, justifying their use as regime-conditional overlays rather than standalone allocations.

5.5 An ensemble of all strategies

a. Methodology

Ex ante, it is unclear which of the strategies presented so far will perform best in the next period. While this is inherently difficult, we rely on the empirical evidence in the literature suggesting that momentum can be informative. We therefore construct an ensemble approach that selects and combines strategies based on their recent performance. Concretely, at each rebalancing date we summarize recent performance through a single score, rank all strategies accordingly, and select the top 3 strategies whose portfolio weights are then combined to form the allocation for the subsequent period. This, in theory, would also ensure implicit protection against big drawdowns by mixing the weights.

To construct this ranking, at each rebalancing date t, for each strategy i we compute its cumulative return C_i and Sharpe ratio S_i over a rolling evaluation window (of 4 Years). Sharpe ratios are truncated below at zero to avoid rewarding strategies with poor risk-adjusted performance. We then define the normalized quantities

$$\text{cum\_norm}_i = \frac{C_i}{\sum_j C_j}, \qquad \text{sharpe\_norm}_i = \frac{\max\{S_i, 0\}}{\sum_j \max\{S_j, 0\}},$$

and construct the overall score

$$\text{Score}_i = 0.7 \cdot \text{sharpe\_norm}_i + 0.3 \cdot \text{cum\_norm}_i.$$

This score places primary emphasis on risk-adjusted performance while still accounting for absolute returns. Relying exclusively on Sharpe ratios would be misleading, as illustrated by the training sample: the shortest-duration buckets exhibit the highest Sharpe

ratios but generate relatively small cumulative returns. Incorporating a return component therefore helps prevent systematically favoring strategies that remain persistently exposed to low-duration buckets.

To avoid any selection on the evaluation set, we fix the metric as it is (i.e., the 70–30 weights) and always select the top three performing strategies. Also the evaluation window is kept fix. An ex post analysis, not discussed here, shows that the performance results presented are not heavily influenced by the choice of these parameters.

b. Results and comments

In Table 23 we show the perfomance of the ensemle of strategie. Factor refers to all single-signal factor strategies introduced earlier (e.g. carry, value, momentum, NSS), explicitly excluding both signal combinations and PCA-based strategies. ML includes all machine-learning strategies obtained from the combination of datasets and algorithms, excluding stacking methods. Threshold-based strategies constitute a separate class and include all rule-based allocations driven by predefined levels or regime indicators.

Full evaluation Sample (2014-2024)

Strategy	Cum. Return	Sharpe	Sharpe (net)
Equal-weighted	0.182	0.291	0.291
Threshold	0.630	0.533	0.525
ML	0.178	0.223	0.218
Factor	0.093	0.163	0.153
Threshold + Factor	0.344	0.380	0.381
All	0.311	0.364	0.355

Table 23: Performance comparison across strategy groups, full evaluation sample (2014–2024).

At first glance, the aggregate *All* portfolio appears to perform better than the equal-weighted benchmark, suggesting that combining all strategy classes is effective (see Figure 10). A closer inspection, however, reveals that this performance is not driven by a balanced contribution across components. Instead, it is largely attributable to the threshold-based strategies (see Figure 11), which display substantially higher Sharpe ratios and cumulative returns over the evaluation period.

Machine-learning strategies (see Figure 12) contribute relatively little in this setting, likely because they do not adapt quickly enough to regime changes and therefore struggle when market conditions shift abruptly. Factor strategies also underperform within the ensemble, not because of a lack of signal, but rather due to the ensemble methodology itself: factor-based approaches such as IC perform well in isolation, indicating that aggregation methods exploiting the cross-sectional correlation structure of signals are more suitable than simple performance-based selection. Overall, the results suggest that the apparent success of the full ensemble mainly reflects the effectiveness of threshold-based rules in stress and sell-off regimes, rather than uniformly strong performance across all strategy classes.

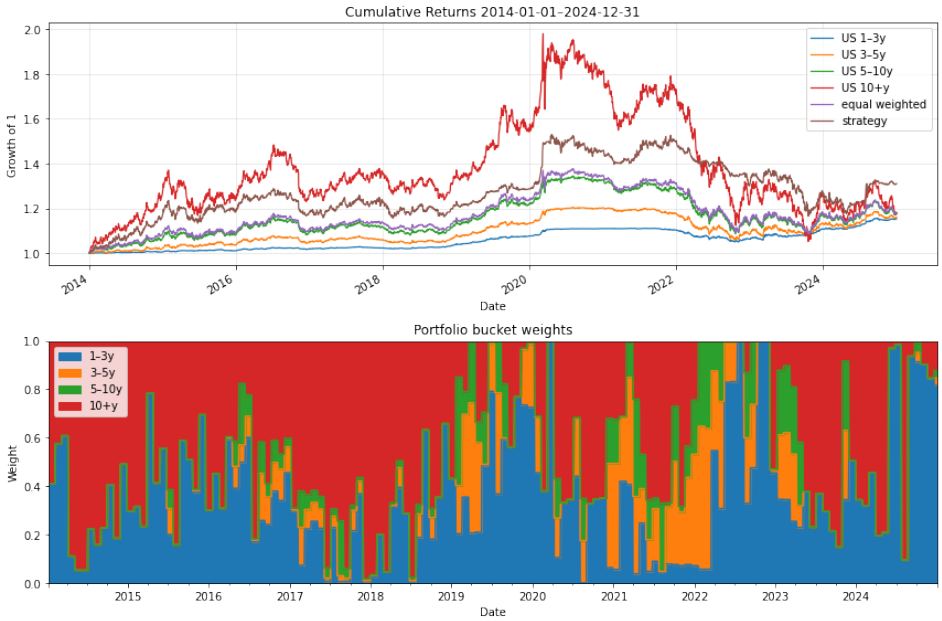

Figure 10: Full ensemble combining threshold-based, factor, and ML strategies. Top panel: cumulative returns compared with the benchmarks. Bottom panel: portfolio weights over time.

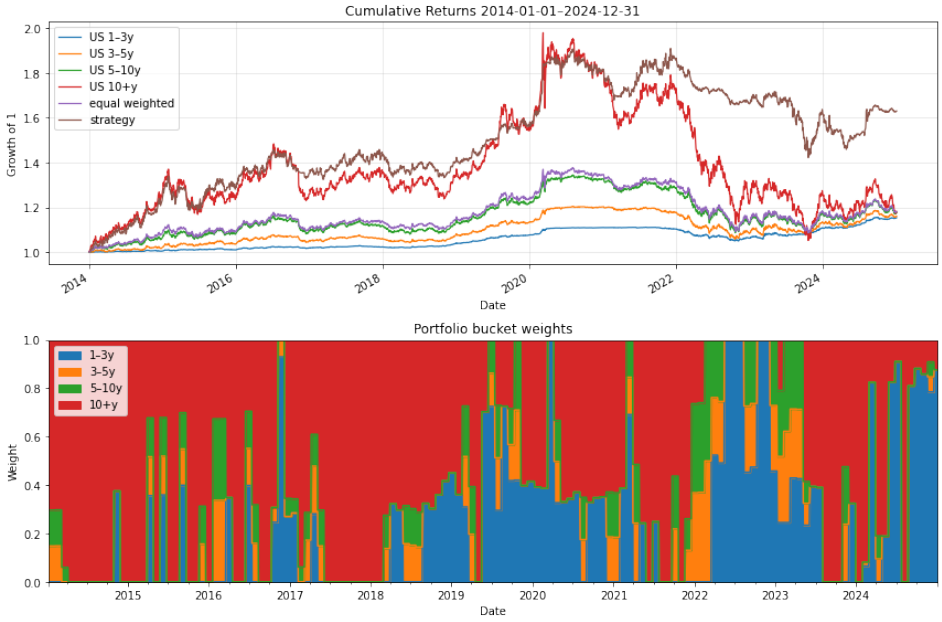

Figure 11: Threshold-based ensemble. Top panel: cumulative returns of the ensemble portfolio compared with the benchmarks. Bottom panel: portfolio weights over time.

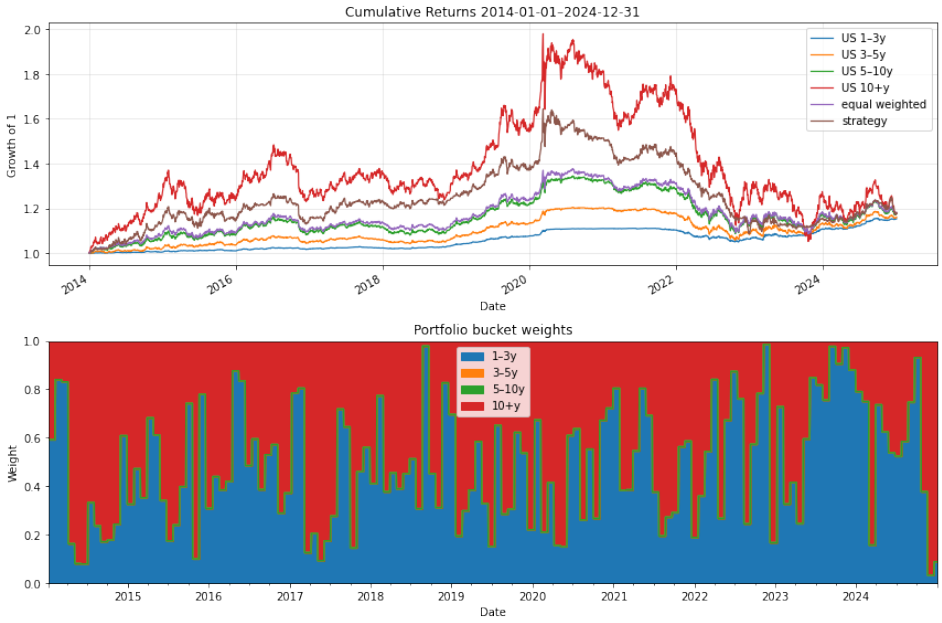

Figure 12: Machine-learning ensemble. Top panel: cumulative returns of the ML-based ensemble compared with the benchmarks. Bottom panel: portfolio weights over time.

6 IPCA

to do

7 Conclusions

to do

References

[1] Clifford S. Asness, Tobias J. Moskowitz, and Lasse Heje Pedersen. Value and momentum everywhere. *Journal of Finance*, 68(3):929–985, 2013. doi: 10.1111/jofi.12021.

[2] Guido Baltussen, Menno Martens, and Oliver Penninga. Factor investing in sovereign bond markets: Deep sample evidence. Ssrn working paper, Robeco Institutional Asset Management, 2021. URL: https://ssrn.com/abstract=3873863.

[3] Diego Bianchi, Markus Buchner, and Alberto Tamoni. Bond risk premiums with machine learning. *Review of Financial Studies*, 34(2):1046–1089, 2021. URL: https://doi.org/10.1093/rfs/hhaa062.

[4] Jordan Brooks and Tobias J. Moskowitz. Yield curve premia. *SSRN Working Paper*, 2017. URL: https://ssrn.com/abstract=2956411.

[5] Antonio Caruso and Luca Coroneo. Does real-time macroeconomic information help to predict interest rates? *Journal of Money, Credit and Banking*, 55(8):2028–2058, 2023. URL: https://doi.org/10.1111/jmcb.13021.

[6] John H. Cochrane and Monika Piazzesi. Bond risk premia. *American Economic Review*, 95(1):138–160, 2005. URL: https://doi.org/10.1257/0002828053828581.

[7] Francis X. Diebold and Canlin Li. Forecasting the term structure of government bond yields. *Journal of Econometrics*, 130(2):337–364, 2006. URL: https://doi.org/10.1016/j.jeconom.2005.03.005.

[8] Brian Hurst, Yao Hua Ooi, and Lasse Heje Pedersen. A century of evidence on trend-following investing. *Journal of Portfolio Management*, 40(1):15–29, 2014. URL: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2993026.

[9] Antti Ilmanen. Forecasting u.s. bond returns. *The Journal of Fixed Income*, 1997. URL: https://www.aqr.com/Insights/Research/Journal-Article/Forecasting-US-Bond-Returns.

[10] Marko Kolanovic and Zhenyu Wei. Momentum strategies across asset classes — risk factor approach to trend following. Technical report, J.P. Morgan, 2015. URL: https://www.cmegroup.com/education/files/jpm-momentum-strategies-2015-04-15-1681565.pdf.

[11] Sydney C. Ludvigson and Serena Ng. Macro factors in bond risk premia. *The Review of Financial Studies*, 22(12):5027–5067, 2009. URL: https://doi.org/10.1093/rfs/hhp081.

Acknowledgments

This project has been conducted with the help of artificial intelligence tools, which have supported technical writing and code debugging/writing. However, all ideas, interpretations, mathematical derivations, methodological choices, and conclusions presented in this work remain the sole responsibility of the author.