A Work Project, presented as part of the requirements for the Award of a Master's degree in Business Analytics from the Nova School of Business and Economics.

# EVALUATING TIME-SERIES MOMENTUM AGAINST MACHINE LEARNING IN COMMODITY FUTURES USING MULTI-HORIZON AND TERM-STRUCTURE SIGNALS

Ben Lou Bumann

Work project carried out under the supervision of:

Nicholas Hirschey

29/12/2025

**Abstract**

This thesis replicates the Moskowitz, Ooi, and Pedersen (2012) time-series momentum strategy in commodities, confirming strong pre-2009 performance and weaker results thereafter. Using a walk-forward backtest, it evaluates machine learning predictions based on individual lagged returns and compares linear and nonlinear models. Models using individual lags outperform TSMOM, with linear models performing best, while nonlinearities and interactions add no value. Time-series strategies based on fixed probability thresholds perform poorly, possibly due to miscalibration and limited directional forecasting ability. Adding term-structure and factor features improves Sharpe and accuracy, but gains are largely explained by factor exposure.

**Keywords:** Time-series momentum; Commodity futures; Machine learning; Random Forest; XGBoost; Logistic regression; Term structure; Basis; Basis momentum; Skewness; SHAP interpretability.

The GitHub repository can be found under:
*https://github.com/beloubu/MasterThesis_MLvsTSMOM_CommodityTrading*

This work was funded by Fundação para a Ciência e a Tecnologia (UID/00124/2025, UID/PRR/124/2025, Nova School of Business and Economics) and LISBOA2030 (DataLab2030 - LISBOA2030-FEDER-01314200).

# Introduction

The weak form of the Efficient Market Hypothesis states that prices reflect information contained in past prices, so that past returns should not predict future returns (Fama 1970; Fama 1991). Momentum is a prominent empirical pattern that contradicts this view, since returns exhibit continuation over intermediate horizons (Jegadeesh and Titman 1993). A closely related pattern exists in the time series dimension, where an asset's own past return predicts its future return. Moskowitz, Ooi, and Pedersen (2012, hereafter MOP) show that a simple trading rule based on the sign of past cumulative returns generates attractive risk adjusted performance across multiple asset classes.

Commodity futures are a relevant setting because trend following is widely implemented in futures markets by practitioners. CTA and managed futures fund returns have been shown to exhibit strong exposure to time series momentum factors, suggesting it is an industry relevant practice (Baltas and Kosowski 2011; Hurst, Ooi, and Pedersen 2013). However, the risk adjusted performance of time series momentum has been diminishing in the post 2009 period, particularly in commodities (Georgopoulou and Wang 2016).

This thesis builds on the observation that the standard time series momentum (TSMOM) signal is intentionally simple. It aggregates information from past returns into a single lookback measure and uses only its sign, which disregards information contained in the return history across multiple horizons, including reversal effect after trend horizons (MOP 2012). This raises the question of whether machine learning models that (1) use past returns in a more granular form, allow for nonlinear relationships and interactions, and can adapt to changes in the return dynamics by retraining annually, can extract predictive value beyond what is captured by the simple time series momentum strategy and whether (2) adding term structure and other risk factor based features, which have been shown to exhibit predictive abilities on future commodity returns, can further improve model performance.

## 1.1 Research Design

The thesis is organized to move from validating the setup to testing the two research questions and explaining what drives any observed performance differences. First, the main time series momentum results in MOP (2012) are replicated and the results for commodities will be presented in Section 2.2 to verify that methodology and testing setup are correct and that the effect is present in the sample. Second, the motivation for the two research questions is outlined in Section 3 by summarizing relevant literature and the economic reasoning behind each idea. Third, in Section 4 the empirical setup is described, including the data used, the two feature sets, the out of sample training and backtesting design, and the approach to model interpretation. Fourth, the results are presented in Sections 5 and 0 by comparing each model to the TSMOM benchmark and by comparing the returns only model to the augmented feature model. Factor regressions are then used to test whether the models' performance can be explained by a set of known commodity risk factors or whether the strategies generate significant positive returns beyond these factors. In addition, SHAP values are reported for the tree-based models to show which inputs contribute most to the model's predictions and to support an interpretation of why performance changes when the feature set is expanded. Finally, limitations are discussed in Section 6, and the thesis concludes in Section 7.

# 2 Replication of MOP's (2012) Findings

## 2.1 Methodology of Replication

In order to replicate the original paper, future data for the same cross asset class underlyings as in MOP (2012) has been obtained from Bloomberg. For the replication, contracts on each underlying are selected using Bloomberg's "active future", which is based on liquidity proxies and typically points to the front or second contract, similar to what is stated in MOP. All daily returns are computed from ratio-backadjusted prices and compounded to monthly returns. For FX, spot rates and one-month forward points for nine USD pairs from MOP are used to compute twelve cross forward pairs returns. Summary statistics, notes on data handling and deviations from the original paper are reported in Appendix A1.2. All returns resemble fully collateralized excess returns in dollar terms.

MOP (2012) construct the trading signal as the sign of the past $h$-month cumulative return to predict the sign of the next month's return. Assets are weighted using volatility scaling based on an annualized exponentially weighted moving average of lagged squared daily returns, which is fully observable ex ante. The decay parameter δ is chosen such that the center of mass equals 60 days. The ex-ante annualized variance can be written as:

$$\sigma_t^2 = 261 \sum_{i=0}^{\infty} (1-\delta)\delta^i (r_{t-1-i} - \overline{r_t})^2 \tag{1}$$

The total strategy return at time $t$+1 is then computed as the average volatility-scaled returns across all securities available at time t. Each position is scaled to target an annualized volatility of 40%. Thus, it can be written as:

$$r_{t,t+1}^{TSMOM} = \frac{1}{S_t} \sum_{s=1}^{S_t} \operatorname{sign}\left(r_{t-12,t}^{s}\right) \frac{0.40}{\sigma_t^s} r_{t,t+1}^{s} \tag{2}$$

The passive long benchmark as implemented in MOP (2012) is the volatility scaled long strategy.

Further, MOP (2012) regress monthly excess returns on the sign of their own lagged returns to test for momentum and reversal effects. Here, they find strong and statistically significant return continuation for lags up to 12 months, followed by weaker but persistent return reversals at longer horizons across asset classes. Returns are scaled by their ex-ante volatilities to ensure comparability across futures, and the regression is estimated as a pooled regression with standard errors clustered by month. Thus, the regression function for each lagged month can be written as follows:

$$\frac{r_t^s}{\sigma_{t-1}^s} = \alpha + \beta_h \operatorname{sign}(r_{t-h}^s) + \varepsilon_t^s \tag{3}$$

For simplifying reasons, the authors proposed TSMOM strategy with a lookback of h = 12 months and a holding period of k = 1 will be considered in the following.

## 2.2 Results of the Replication – TSMOM on Commodities

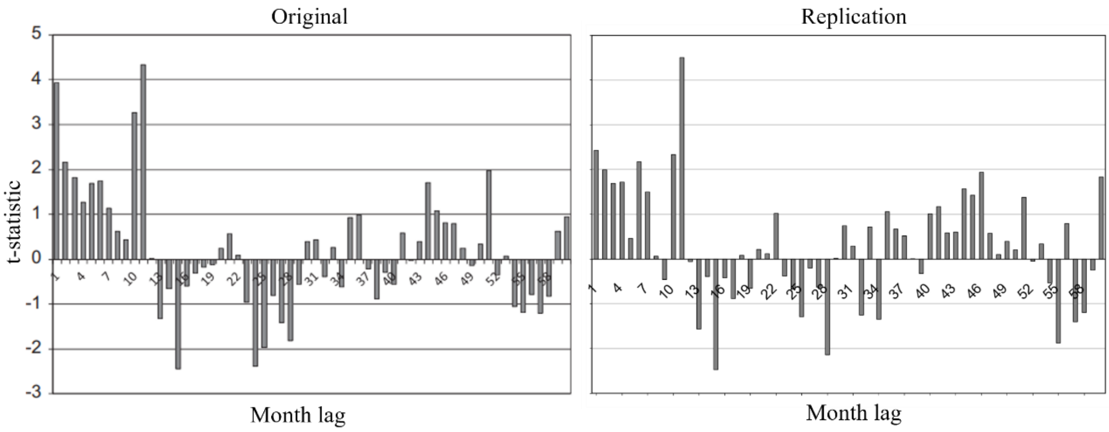


*Figure 1:Lagged Returns t-statistics Commodities - Original vs. Replication*

In the following, the replication results for commodities are reported, because all subsequent analysis is restricted to commodity futures. Replication results for the full cross asset universe are provided in Appendix A1.1. As can be seen on Figure 1, the pooled regressions match the main pattern shown in MOP (2012). The t-statistics on lagged monthly returns show continuation for roughly the first twelve lags and weaker reversal beyond that.

*Table 1: TSMOM Performance - Replication vs. Original - Pre & Post 2009*

| | 1985-08/2025 | | | 1985-2009 | | | 2009-08/2025 | | |
|---|---|---|---|---|---|---|---|---|---|
| Strategy | Mean | Vol. | SR | Mean | Vol. | SR | Mean | Vol. | SR |
| TSMOM | 0,0924 | 0,1488 | 0,6212 | 0,1367 | 0,1386 | 0,9861 | 0,0213 | 0,1621 | 0,1317 |
| Passive Long | 0,0612 | 0,1984 | 0,3083 | 0,0816 | 0,1855 | 0,4398 | 0,0310 | 0,2180 | 0,1423 |
| Equal Weight | 0,0402 | 0,1466 | 0,2745 | 0,0556 | 0,1295 | 0,4295 | 0,0175 | 0,1712 | 0,1020 |
| Original Paper | 0,0954 | 0,1477 | 0,6459 | 0,1424 | 0,1372 | 1,0380 | 0,0184 | 0,1609 | 0,1145 |

The backtest results are also close to the original paper's results and to the extended return series published by AQR Capital Management (2025). Figure 2 compares the growth the replicated strategy with the original series, and Table 1 shows that the performance metrics are very similar.

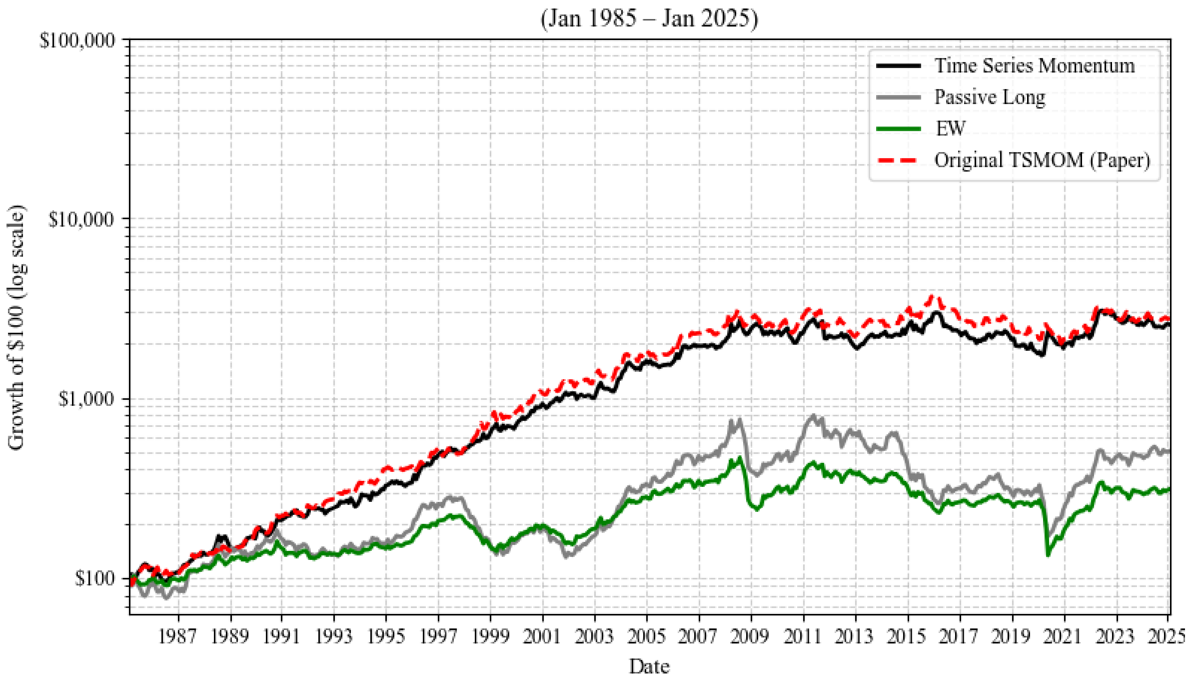


*Figure 2: Cumulative Excess Returns of TSMOM vs Passive Long - Original vs. Replication*

### 2.2.1 Decomposing Predictive Abilities of Past Returns for Commodities

To explain where the predictive abilities of past returns stem from, MOP (2012) decompose the 12-month total future returns into two components: the spot price change and the roll return. Spot price change is measured as the change in the unadjusted front month futures price, while the roll return is defined as the residual component so that total 12-month future return equals: *12m spot price change + 12m roll return*. They then regress next month's total futures returns on the momentum of (i) total futures returns, and separately from (ii) spot price changes and (iii) roll returns, both in univariately and jointly in a multivariate regression. They find that both spot and roll momentum predict next month total futures returns, and that in the multivariate regression the spot and roll coefficients are very similar. Similar findings have been made using data up until 2009 and are presented in Table 2 Panel A. As a result, the multivariate regression explains essentially the same amount of variation as the univariate regression using total momentum, indicating that TSMOM is driven by both spot and term components, while decomposing them doesn't add explanatory powerblue.

*Table 2: Time Series Predictors: Spot Price Change and Roll Returns - Pre and Post 2009*

| **Panel A: Sample Period: 1985 -2009** | | | | | | | |
|---|---|---|---|---|---|---|---|
| Predictors | β-Total | t-Total | β-Spot | t-Spot | β-Roll | t-Roll | $R^2$ |
| Full TSMOM | 0.0246 | 5,16 | - | - | - | - | 0.0092 |
| Spot Only | - | - | 0.0193 | 3,28 | - | - | 0.0042 |
| Roll Only | - | - | - | - | 0.0314 | 4,37 | 0.0048 |
| Spot + Roll | - | - | 0.0209 | 3,53 | 0.0337 | 4,69 | 0.0097 |
| **Panel B: Sample Period: 2010-08/2025** | | | | | | | |
| Predictors | β-Total | t-Total | β-Spot | t-Spot | β-Roll | t-Roll | $R^2$ |
| Full TSMOM | 0.0075 | 1,23 | - | - | - | - | 0.0007 |
| Spot Only | - | - | 0.0134 | 1,80 | - | - | 0.0018 |
| Roll Only | - | - | - | - | -0.0137 | -1,93 | 0.0006 |
| Spot + Roll | - | - | 0.0126 | 1,68 | -0.0109 | -1,57 | 0.0022 |

In contrast, in the post 2009 subsample, shown in Table 2 Panel B, the aggregate total momentum regression has a lower $R^2$ than the joint spot and roll specification, and the spot and roll coefficients no longer have the same sign. While the overall explanatory power remains low in this period, this pattern suggests that aggregating spot and roll into a single total momentum signal can leave out information, which motivates adding term structure measures such as the basis that summarize the implied roll component.

While the replication closely aligns with the original paper's findings and is thus considered validated, Table 1 and Figure 2 show that the strategy performed well prior to 2010, but the risk-adjusted returns have been weak in the years since. Georgopoulou and Wang (2017) suggest that the market interventions of Central Banks, especially quantitative easing, have played a role in distorting the correlation patterns. This raises the question of whether machine learning models that (1) use past returns in a more granular form and can adapt to changes in correlation structure by retraining annually, can extract predictive value out of past returns beyond what is captured by the simple 12-1 TSMOM strategy and whether (2) adding term structure and risk factor based features can improve the performance.

# 3 Motivation and Ideation Behind Features

## 3.1 Model 1: Lagged Returns Only

Prior work suggests that the fixed 12 to 1 TSMOM strategy can be improved when signals from multiple lookbacks are combined or strategies are diversified over different lookbacks. Hurst et al. (2013) show that combining TSMOM strategies with different lookbacks and scaling them by their volatility improves performance relative to single lookback strategies.

Lim, Zohren, and Roberts (2019) show across asset classes that models that use multiple volatility scaled lookbacks and MACD style indicators can outperform the TSMOM benchmark. While an LSTM that directly determines portfolio weights under an expected Sharpe ratio objective delivers the strongest performance, even a simple linear classifier that uses volatility scaling for position sizing outperforms the TSMOM benchmark. This supports the idea that a more granular representation of past returns contains predictive information that is lost when returns are reduced to a single cumulative return signal.

Beckmeyer and Wiedemann (2025) show in equities that the optimal contribution of past returns to the momentum signal varies across days and can be learned through time-varying weights that are set separately for each stock each month, using information from the stock's return history and firm characteristics.

Furthermore, using individual lagged returns allows linear models to attribute signed weights to certain lags, let those change over time and further aggregate potentially both reversal and momentum signals as implied in Figure 1. Additionally, it could consider the magnitude of past returns beyond only the sign. Thus, the idea is that it could attribute high probabilities to assets where returns have been consistently in line with the estimated weights. Additionally, non-linear models could also capture non-linearity and interactions between those lags.

Building on these findings, the first research question tests whether pooled machine learning models can extract predictive information from lagged returns, and can outperform the TSMOM benchmark out-of-sample.

## 3.2 Model 2: Augmenting with Factor Inspired Features

Commodity futures expected returns can be decomposed into spot premia and term premia. Spot premia compensate investors for exposure to spot price risk, while term premia compensate investors for exposure to changes in the futures curve across maturities (Szymanowska et al. 2014). The return time series in this thesis follows the short roll strategy as defined in Szymanowska et al. (2014) and is therefore by definition expected to earn spot premia. However, the term structure is shown to contain information about future spot premia.

Szymanowska et al. (2014) demonstrate that the futures basis serves as a significant indicator of front-month returns, suggesting that the slope of the term structure contains valuable information regarding expected spot premia. Their findings indicate that this forecasting power is primarily driven by structural factors, with approximately 70% of the predictive ability attributed to the historical mean of the basis (the cross-sectional component). The remaining 30% is explained by time-series deviations from this mean.

Boons and Prado (2019) introduce basis momentum, a predictor constructed from differences in average roll returns across maturities and relate it to the slope and the average curvature of the term structure over that horizon. They show that basis momentum contains predictive information on both spot and term premia in the time series and in the cross section.

Further, Fernandez -Perez et al. (2018) show that commodities with more negative return skewness earn higher subsequent futures returns, and that a strategy that goes long the most negatively skewed contracts and short the most positively skewed contracts delivers significant excess returns.

Finally, Szymanowska et al. (2014), citing Dhume (2011), argue that variation in front month returns can proxy scarcity and inventory related conditions and is associated with higher spot premia.

Taken together, these findings motivate augmenting the return-based feature set with term structure measures of the slope of the future curve and basis momentum, as well as measures of skewness and return variation, to test whether this additional information improves predictive accuracy and strategy performance beyond models that use past returns only.

# 4 Empirical Set Up

## 4.1 Data

In contrast to the replication setup, the following backtests switch from Bloomberg's "active future" rolled series to a contract chain that rolls one month prior to expiration. This change is necessary to construct a consistent chain of contracts at multiple maturities using Bloomberg's generic futures, which is required to compute term structure features.

## 4.2 Feature Engineering

### 4.2.1 Lagged Return Features

Model 1 uses a return-only feature set. For each commodity and month, the feature set consists of the most recent 20 lagged monthly excess returns, included as separate features. The 20-month window is chosen to cover both medium-term continuation horizons and longer-horizon reversal effects as stated by MOP (2012) and shown in Figure 1. Each lagged return is volatility scaled using the volatility estimate defined in (1) to make returns comparable across commodities and time.

Lagged returns are not demeaned because Huang et al. (2020) show that TSMOM performance is similar to a strategy based on the sign of the historical mean, which they interpret as indicating that long-run return components can account for a material share of the TSMOM performance. Thus, the features can be written as:

$$x_{j,t}^{(k)} = \frac{r_{j,t-k}}{\sigma_{j,t-k+1}}, \qquad k = 1, \ldots, 20. \tag{4}$$

### 4.2.2 Augmented Feature Set

The second model extends the feature set by adding additional commodity characteristics. Basis momentum is included and is split into its slope and curvature components (Boons and Prado 2019), which are added as separate variables so the models can use them individually and in combination. The basis is computed from the front and the next contract, but contract calendars differ across commodities. Some contracts trade in every calendar month, while others trade only in specific months via bimonthly schedules or contract specific delivery months. As a result, the maturity gap between the front and second contract is not constant and varies across commodities and over time. To make basis measures comparable, the basis is linearly interpolated to a constant one-month maturity gap. This follows the approach of Koijen et al. (2018).

Three basis features are included. The first is the 12-month rolling mean of the basis to capture persistent differences in average basis levels across commodities. Second the current observable basis is included implying the next roll return. Lastly, the z-score of the current basis against the past 12-month is added to capture deviations from recent history (Szymanowska et al. 2014).

Following Fernandez Perez et al. (2018), the skewness is computed as the skewness of daily returns over the past year. The coefficient of variation is computed from the past 12-month front month return series and is included as a simple measure of return stability (Szymanowska et al. 2014). Further details on data handling and computations are reported in Appendix A2.

## 4.3 Classification Models Used

To allow for nonlinearity and interactions between features, tree-based ensembles are used for this prediction task. The problem is formulated as supervised classification on feature-engineered tabular panel data, where return dynamics might depend on nonlinear thresholds. Tree-based models capture these interactions naturally by performing successive splits on different variables (Gu, Kelly, and Xiu, 2020).

The implementation utilizes Random Forest and XGBoost to address this classification task. Random Forests aggregate multiple decision trees trained on bootstrapped subsamples to reduce variance and enhance out-of-sample stability while modeling nonlinearities and interactions (Breiman 2001). In contrast, XGBoost employs a gradient boosting framework to build an additive model by fitting trees sequentially to remaining errors, incorporating regularization to ensure scalable and efficient learning (Chen and Guestrin 2016). Using both methods allows for a consistency check between an ensemble approach focused on variance reduction and one focused on sequential bias reduction. Additionally, it serves as a robustness check to ensure that the findings reflect information content in the features rather than an artifact of backtest overfitting specific to a single model.

In addition to the nonlinear models, a $L2$-regularized logistic regression model is included as a linear benchmark. This allows assessing whether allowing for nonlinearities and interaction effects provides added value beyond a linear model.

All three models are implemented using the scikit-learn library (Pedregosa et al. 2011).

## 4.4 Backtest Framework and Portfolio Construction

### 4.4.1 Walk-Forward Design and Sample Period

A walk-forward backtest is employed to simulate historical strategy performance, ensuring each decision is based exclusively on information available at that point in time. This procedure follows the standard approach for evaluating trading strategies under realistic information constraints (López de Prado 2018). At the end of month $t$, the model is trained only on observations with known outcomes, consisting of feature vectors up to $t-1$ paired with returns realized at t. The trained model then uses features observable at month $t$ to form portfolio positions that realize returns over month $t+1$. This design ensures the backtest is strictly out-of-sample and free of look-ahead bias.

While this design avoids classical lookahead bias, it has limitations regarding generalizability because it produces only one realized historical path. However, serial correlations in commodity futures returns have changed over time. This shifting correlation structure and varying strength of serial dependence motivated the restriction of the training set to available historical information.

The sample begins in January 1985 to ensure sufficient contract coverage, consistent with MOP (2012), and ends in August 2025. Features are constructed using a 20-month lookback and include ex ante volatility estimates that require a minimum estimation period. Because volatility estimates are only available from June 1985 onward, the first month with a complete set of features is January 1987. Using a ten-year initial training window, the training sample therefore starts in January 1987 and ends in December 1996, and the first out of sample prediction is formed for January 1997. After this initial period, models are retrained annually using an expanding training window that includes all data observed up to the retraining date. Here, the models are trained in a pooled manner across all assets and time periods.

### 4.4.2 Portfolio Construction

The pooled training allows the models to characterize features by comparing them across the entire sample, effectively defining thresholds by assessing whether a feature is high or low across assets and time. Consequently, even when predictions are applied within a time-series strategy, they might exploit cross-sectional differences of past returns.

For each model, predicted probabilities are used to form two portfolios. The first follows a time-series approach by going long on all commodities with a positive predicted return and short otherwise, using volatility-scaled weights as defined in MOP (2012) and in equation (1). The second is a cross-sectional strategy that sorts commodities by predicted probability, going long the top 30% and short the bottom 30%. This portfolio is rebalanced monthly and remains dollar neutral.

### 4.4.3 Hyperparameter Tuning

XGBoost and Random Forest require the specification of hyperparameters to, for example, define node-splitting rules, maximum tree depth, subsampling schemes, and the number of estimators. Candidate hyperparameter combinations are evaluated using the area under the ROC curve (AUC). A five-fold time-series cross-validation procedure is used, where folds are constructed from the initial ten-year training window. This metric is chosen because class balance may vary across folds due to market-wide shifts between up and down periods.

Hyperparameter tuning is performed once at the beginning of the sample on the initial training window. Due to computational constraints, the selected hyperparameters are then held fixed throughout the remainder of the walk-forward backtest. The candidate grids and selected hyperparameters are reported in Appendix A6.

## 4.5 Model Interpretability and SHAP Values

To analyze which features drive the predictions of the machine learning models and to interpret the dynamics they capture, Shapley additive explanations (SHAP) are used following Lundberg et al. (2020) to explain the non-linear models. For each observation, SHAP values decompose the model's predicted probability of a positive return into additive contributions of individual features, accounting for interaction effects. The sum of all SHAP values plus the model's expected value equals the model's prediction.

Global feature importance is measured using mean absolute SHAP values, which quantifies the average contribution magnitude of each feature across the sample. In addition, SHAP dependence plots are used to examine nonlinear relationships and interactions learned by the models. For the linear models the features' standardized betas and t-stats will be analyzed.

# 5 Results of Models Trained on Past Return Features Only

To address the first research question, models utilizing only lagged returns as features are compared against the standard time-series momentum strategy. Performance is evaluated using risk-adjusted measures over the same out-of-sample period. Factor regressions are then applied to the portfolio returns to determine if established commodity risk factors can explain the returns or if the strategy delivers significant unexplained returns. Finally, the fitted models are analyzed using SHAP values to identify which specific lags drive the predictions and to examine whether the non-linear models capture interaction patterns across lags that are not reflected in a linear specification.

*Table 3: Performance Metrics - Return Features Only*

Model 1: Using volatility scaled lagged monthly return features only

| Strategy | SR | Ann Mean | Ann. Vol. | Hit Rate | Hit Rate + | Hit Rate - | MDD |
|---|---|---|---|---|---|---|---|
| RF-TS | 0,542 | 0,07 | **0,125** | 52,23% | 52,58% | 51,93% | -32,85% |
| RF-CS | 0,616 | 0,12 | 0,19 | 52,73% | 52,76% | 52,71% | -41,56% |
| XGB-TS | 0,325 | 0,04 | 0,131 | 51,59% | 51,91% | 51,33% | -26,03% |
| XGB-CS | 0,608 | 0,11 | 0,185 | 52,57% | **53,34%** | 51,79% | -48,54% |
| LOG-TS | 0,614 | 0,08 | 0,132 | 52,65% | 52,99% | 52,36% | **-18,27%** |
| LOG-CS | **0,835** | **0,16** | 0,187 | **53,43%** | 53,22% | **53,64%** | -27,12% |
| TSMOM 12M L/S | 0,436 | 0,07 | 0,154 | 51,89% | 52,08% | 51,71% | -41,66% |
| Passive Long | 0,231 | 0,05 | 0,208 | 49,99% | 49,99% | | -78,30% |
| EW-Long | 0,236 | 0,03 | 0,145 | 49,99% | 49,99% | | -60,89% |

Table 3 reports the backtest results for models that use lagged returns as features only. Strategies with a TS suffix apply model predictions directly in a time-series manner using a fixed threshold, with volatility scaling for each asset, while strategies with a CS suffix construct a dollar-neutral portfolio as described before.

Comparing the performance metrics in Table 3 shows that, in terms of the Sharpe ratio, all machine learning strategies outperform the standard TSMOM benchmark, except for the XGBoost time-series strategy. Moreover, all cross-sectional implementations outperform their corresponding time-series variants in both accuracy and risk-adjusted returns.

This pattern could be driven by three factors. First, the cross-sectional strategies exclude the middle range of predicted probabilities, such that only contracts with relatively high model confidence are traded. As a result, capital is allocated to signals with stronger predicted return asymmetry, while low-certainty signals are ignored.

Second, predicting market-wide directional movements using individual asset forecasts might be difficult. The dollar-neutral construction instead focuses on the relative ranking of assets, such that the strategy isolates cross-sectional signals from broad market trends. By maintaining balanced long and short positions, the strategy aims to offset prediction errors related to market-wide shocks, focusing the performance on the model's ability to differentiate between winners and losers. On the other hand, the higher volatility of the CS strategies might be cause of the absence of asset-level volatility scaling and the more concentrated portfolio that invests only in the top and bottom 30 percent of commodities rather than in the full universe.

Third, the time-series strategies utilize a fixed 0.5 probability threshold, which may not be suitable given that the models might not be well-calibrated, especially considering time-varying class imbalances between market up and down movements. Because the proportion of positive realized returns fluctuates over time, a constant threshold can be poorly situated in certain regimes. This can lead to unstable long and short exposures even if the model maintains the correct relative ranking of signals.

## 5.1 Explaining the Models

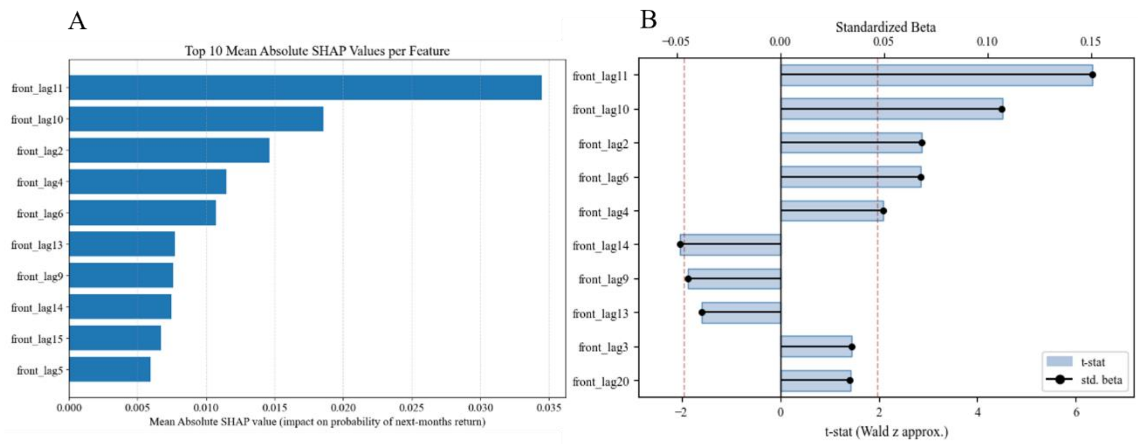


*Figure 3: Model 1 - RF SHAP Feature Importance and Standardized Coefficients with t-Statistics*

To understand how the models make predictions, a training set ending in 2015 is utilized. For the linear model, the feature importance is interpreted directly through the parameters (standardized betas and Wald z-statistics) fitted on this training data. The non-linear model is analyzed by evaluating SHAP values on the out-of-sample period from 2016 to 2025. Figure 3A plots the mean absolute SHAP values for the Random Forest (RF) model, while Figure 3B presents the top 10 features for the linear model.

### 5.1.1 Random Forest

Figure 3A shows that, similar as in MOP (2012), the features with the highest implied importance spread between lag two and fifteen.

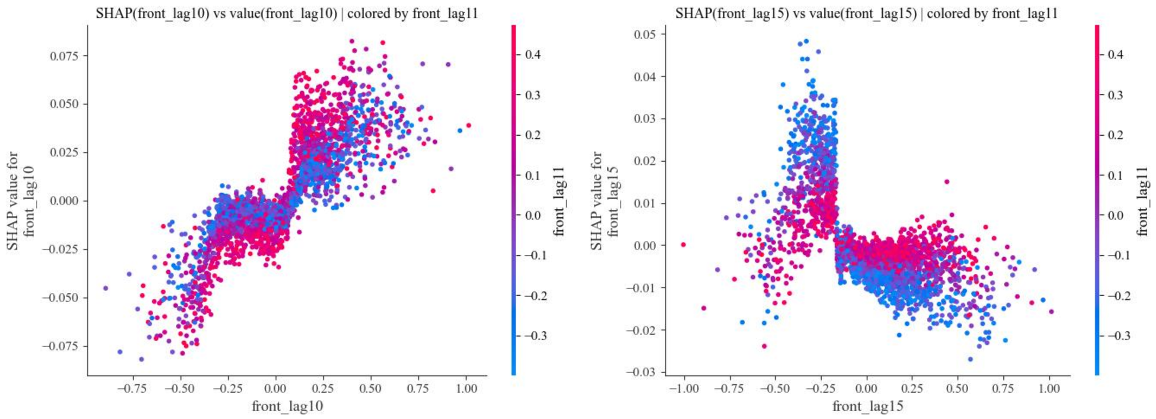


*Figure 4: RF Dependence Plots - Model 1*

The SHAP dependence plots and the beeswarm plot in Appendix A4.1.1 for the Random Forest suggest that the lags behave mostly monotonic, with SHAP values increasing as feature values increase or vice versa. A positive slope indicates that the model links that lag to return continuation relative to the target months return, while a negative slope indicates the opposite. As Figure 4 shows, some lags show slightly nonlinear fitted patterns, for example lag 10 has a S-shaped relation where small deviations around the center have little impact but larger deviations lead to larger SHAP values. Lags 10 and 11 show the strongest interactions with other lags. The plots suggest that lag 15 contributes to a reversal direction for positive values, and that this effect becomes stronger when lag 11 is negative, while a positive lag 11 dampens the reversal pattern attributed to lag 15 for both positive and negative values. Lags 10 and 11 on their own look close to linear continuation signals, and their interaction indicates that lag 11 can strengthen the positive contribution of lag 10 while dampening its negative contribution. Similar patterns are visible in the XGBoost dependence plots, with selected examples reported in Appendix A4.1.2.

However, considering that the linear model has outperformed the non-linear models, it appears that those patterns are rather a matter of overfitting on the training data and do not persist or add value out-of-sample.

### 5.1.2 Logistic Regression

The LOG-CS model is trained on the same sample, and the top 10 features by absolute t-stat and standardized betas are presented in Figure 3B. The most significant features in the linear model mostly align with those in the Random Forest. The results show that the model captures momentum patterns for lags under 12 months, followed by a slight reversal after 12 lags.

## 5.2 Factor Analysis

A factor analysis is conducted to assess whether the strategy's returns are explained by established risk factors or reflect incremental performance beyond conventional factor exposures, benchmarking against commodity risk factors (TSMOM, average commodity market factor (AVG), basis (B), basis momentum (BM) , value (Val), skewness (Skew), Dollar (Dol), cross-sectional momentum (CM) and front month volatility (Vol)). Additional information on the computation of the commodities risk factors can be found in Appendix A5. To account for factor correlations (Figure A 19), the regression specifications are defined using a small set of test variants that vary the included factors. Here, the best-performing linear and non-linear strategies in terms of Sharpe ratio are analyzed

*Table 4: Factor Analysis - Model 1*

| | α ann.% | AR | βAVG | βTM | βCM | βB | βBM | βVal | βVol | βSkew | βDol | $R^2$ |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Panel A: LOG-CS | | | | | | | | | | | | |
| Coef | **13,22** | 0,790 | **0,172** | | **0,344** | | | | -0,031 | | 0,086 | 0,201 |
| T | 4,45 | | 2,760 | | 4,701 | | | | -0,525 | | 1,024 | |
| Coef | **12,27** | 0,740 | **0,183** | | **0,422** | -0,094 | 0,034 | -0,074 | -0,039 | 0,038 | 0,095 | 0,216 |
| T | 3,68 | | 2,954 | | 5,517 | -1,498 | 0,503 | -0,961 | -0,690 | 0,419 | 1,115 | |
| Coef | **11,51** | 0,698 | **0,178** | **0,163** | **0,350** | -0,075 | 0,033 | -0,088 | -0,053 | 0,029 | 0,102 | 0,225 |
| T | 3,52 | | 2,618 | 2,430 | 4,230 | -1,263 | 0,475 | -1,158 | -0,959 | 0,335 | 1,207 | |
| Panel B: RF-CS | | | | | | | | | | | | |
| Coef | **10,12** | 0,574 | 0,049 | | **0,320** | | | | 0,002 | | 0,008 | 0,137 |
| T | 3,28 | | 0,569 | | 4,849 | | | | 0,030 | | 0,121 | |
| Coef | **10,39** | 0,590 | 0,051 | | **0,328** | | -0,038 | 0,000 | 0,004 | 0,002 | 0,004 | 0,138 |
| T | 3,07 | | 0,592 | | 4,373 | | -0,645 | 0,006 | 0,073 | 0,020 | 0,062 | |
| Coef | **10,12** | 0,577 | 0,052 | 0,053 | **0,343** | -0,096 | -0,018 | 0,011 | -0,008 | 0,011 | 0,015 | 0,146 |
| T | 2,86 | | 0,630 | 0,662 | 3,980 | -0,939 | -0,328 | 0,174 | -0,145 | 0,110 | 0,213 | |

Table 4 shows that, for the lagged returns only models, both LOG-CS and RF-CS deliver positive and statistically significant alphas and attractive appraisal ratios across the reported factor specifications, alongside a consistently strong loading on the cross-sectional momentum factor. Since the momentum factor is constructed from a fixed lookback, while the model-based strategies combine information across multiple lags, and overall explanatory power is modest, the intercept should be read as unexplained performance relative to this specific factor set rather than evidence of standalone abnormal returns.

## Results of Models Trained on Additional Features

*Table 5: Performance Metrics - Augmented Features*

Model 2: Using augmented features

| Strategy | SR | Ann Mean | Ann. Vol. | Hit Rate | Hit Rate + | Hit Rate - | MDD |
|---|---|---|---|---|---|---|---|
| RF-TS | 0,253 | 0,03 | 0,135 | 51,90% | 52,53% | 51,50% | -37,52% |
| RF-CS | 0,817 | 0,15 | 0,188 | **54,00%** | 54,06% | **53,93%** | -39,73% |
| XGB-TS | 0,412 | 0,05 | **0,128** | 52,25% | 52,75% | 51,89% | -32,25% |
| XGB-CS | 0,794 | 0,15 | 0,188 | 53,79% | 53,93% | 53,64% | -39,75% |
| LOG-TS | 0,758 | 0,10 | 0,130 | 53,51% | 53,92% | 53,16% | **-18,46%** |
| LOG-CS | **0,882** | **0,17** | 0,196 | 53,70% | **54,14%** | 53,26% | -43,25% |
| TSMOM 12M L/S | 0,436 | 0,07 | 0,154 | 51,89% | 52,08% | 51,71% | -41,66% |
| Passive Long (scaled) | 0,231 | 0,05 | 0,208 | 49,99% | 49,99% | | -78,30% |
| Equal-Weighted Long | 0,236 | 0,03 | 0,145 | 49,99% | 49,99% | | -60,89% |

As shown in Table 5, across the augmented feature strategies, the cross-sectional variants again deliver the strongest risk adjusted performance. RF-CS, XGB-CS, and LOG-CS all achieve Sharpe ratios around 0.8 and annualized mean returns of about 15 to 17 percent, clearly above the TSMOM benchmark. Again, the time series variants are weaker, with RF-TS and XGB-TS producing low Sharpe ratios despite similar volatility. Compared to the first model results in Table 3, it shows that almost all strategies improved their performance, whereas the improvement on the best performing variant across both, the LOG-CS, is only marginal.

## 5.3 Explaining the Models

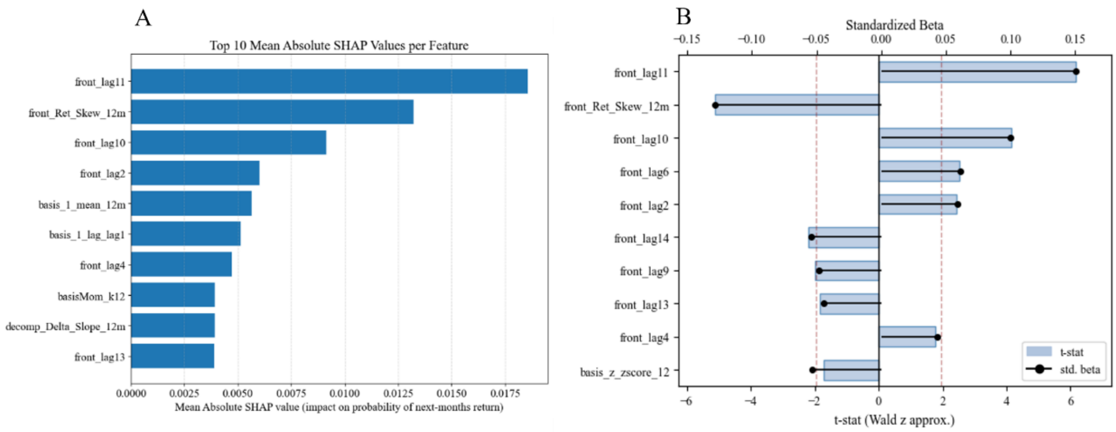


*Figure 5: Model 2 - RF SHAP Feature Importance and Standardized Coefficients with t-Statistics*

To explain the models with the additional features, they have been fitted on the same subperiod as described in Section 5.1. Figure 5A shows the mean absolute SHAP values, and Figure 5B shows the z-statistics and standardized betas. Because RF-CS delivers the strongest Sharpe among the nonlinear models, the analysis focuses on the RF model.

### 5.3.1 Random Forest

The mean absolute SHAP values in Figure 5A indicate that, among the added predictors, skewness and the 12 months mean basis contribute most to the model output, while the relative importance ordering of the lagged return features remains broadly unchanged. Figure 5 also shows that basis momentum might be more informative through its change in slope component, whereas the curvature component has little influence.

The beeswarm plots in Figure A 14 suggest that most features act approximately monotonic, and the relations appear more linear, which is consistent with the shallow tree depth used. As shown in Figure 6, the skewness feature displays a roughly linear negative relation with the prediction contribution, consistent with Fernandez-Perez et al. (2018). For the mean basis it indicates that more negative basis values increase the probability of a positive next month return, while increasing positive basis values have a weaker effect. Additional dependence plots can be found in Appendix A4.1.1.

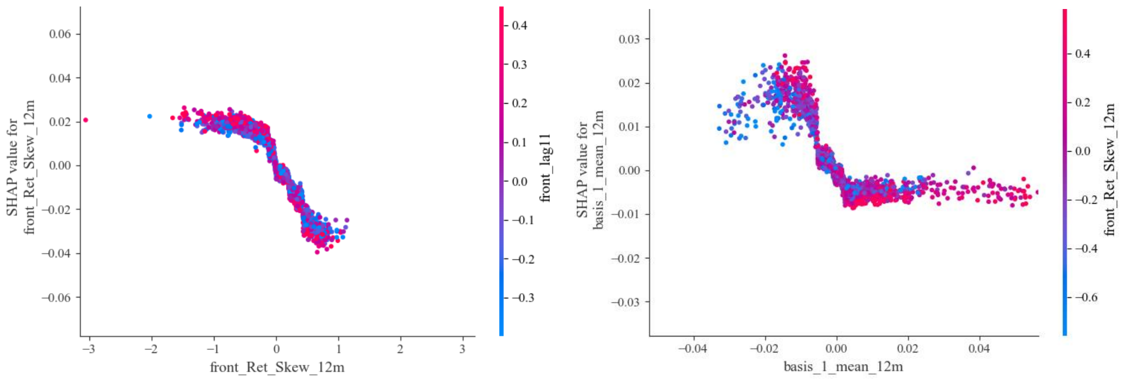


*Figure 6: RF – Dependence - Mean Basis and Skewness*

### 5.3.2 Logistic Regression

Figure 5B shows that the model captures the same momentum patterns for lags under 12 months, followed by a slight reversal after 12 lags, maintaining a similar pattern to the previous model trained on return only. While skewness is attributed a high importance in both models, compared to the Random Forest model, the linear model attributes less importance to the basis and basis momentum features.

## 5.4 Factor Analysis

*Table 6: Factor Analysis - Model 2*

| | α ann.% | AR | βAVG | βTM | βCM | βB | βBM | βVal | βVol | βSkew | βDol | R2 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Panel A: LOG-CS | | | | | | | | | | | | |
| Coef | **14,70** | 0,844 | 0,110 | | **0,368** | | | | -0,032 | | 0,125 | 0,208 |
| T | 4,92 | | 1,153 | | 5,127 | | | | -0,409 | | 1,373 | |
| Coef | 9,07 | 0,578 | 0,102 | | **0,332** | | 0,102 | -0,097 | -0,039 | **0,373** | 0,082 | 0,358 |
| T | 3,02 | | 1,383 | | 4,214 | | 1,385 | -1,444 | -0,630 | 4,124 | 1,175 | |
| Coef | 8,34 | 0,537 | 0,100 | 0,149 | **0,300** | -0,073 | 0,119 | -0,096 | -0,059 | **0,375** | 0,096 | 0,370 |
| T | 2,79 | | 1,203 | 1,807 | 3,265 | -1,070 | 1,741 | -1,487 | -0,975 | 4,337 | 1,315 | |
| Panel B: RF-CS | | | | | | | | | | | | |
| Coef | 12,71 | 0,798 | 0,109 | | **0,421** | | | | -0,081 | | 0,129 | 0,279 |
| T | 5,38 | | 1,430 | | 6,104 | | | | -1,098 | | 1,685 | |
| Coef | 7,53 | 0,549 | 0,083 | | **0,291** | | 0,087 | 0,056 | -0,086 | **0,418** | 0,072 | 0,466 |
| T | 2,76 | | 1,168 | | 3,586 | | 1,484 | 0,799 | -1,412 | 7,361 | 1,314 | |
| Coef | 7,07 | 0,542 | 0,072 | 0,111 | 0,143 | **0,279** | 0,033 | 0,006 | -0,076 | **0,382** | 0,056 | 0,517 |
| T | 2,94 | | 1,195 | 1,454 | 1,733 | 6,860 | 0,528 | 0,100 | -1,446 | 7,226 | 1,107 | |

For the augmented-feature model, the factor regressions in Table 6 suggest that the strategies partly load on the factors implied by the features, but this shows up differently by estimator. As previously observed in the feature importance analysis (see Figure 5), RF-CS exhibits a statistically significant basis loading, whereas LOG-CS does not. In contrast, skewness remains the clearest significant exposure for both models, consistent with its high ranking in both the SHAP values and the t-stats.

Notably, for RF-CS, the cross-sectional momentum exposure becomes insignificant when controlling for basis. This suggests that the non-linear model effectively substitutes or subsumes the traditional momentum signal with the basis factor, potentially due to the high correlation between these factors as shown in Figure A 19. Thus, one explanation for the difference in basis loading might be that the linear model with $L2$ shrinkage simply shrinks away the basis features.

Even when controlling for these risk factors, a positive and significant return remains unexplained for both model setups. However, while the Sharpe ratios for the augmented models improved, the appraisal ratios (AR) remain similar to the first model. This indicates that the additional performance can be explained by exposure to the underlying risk factors.

# 6 Limitations

This study is based on one historical walk forward backtest, so the results depend on the specific path of returns in this sample and may not generalize over to other periods. The reported performance is only one outcome from a noisy process, and may be influenced by backtest overfitting, as conducting many backtests can accidentally pick patterns that are unique to this dataset (López de Prado 2018). In addition, the choice of 20 lagged returns was guided by the t-statistic plot Figure 1 suggesting mild reversals beyond 12 months, which may introduce some lookahead.

Even without intentional bias, the model may rely on relationships that looked predictive in this sample and then happened to persist later, which can inflate risk adjusted returns even if the relationship is not truly stable. Finally, the factor regressions use single lookback momentum factors, while the models learn signals from many lags and features, so the factor analysis may not fully capture what drives the predictions and would benefit from deeper diagnostics across lookbacks and including a short-term reversal factor.

# 7 Conclusion

This thesis first replicates the time series momentum strategy of Moskowitz, Ooi, and Pedersen (2012) in commodities to establish the benchmark and confirm the strong pre 2009 performance and the materially weaker performance after 2009. It then tests two research questions in a walk forward out of sample backtest. Research question one asks whether machine learning models that use individual lagged commodity returns as features outperform the standard TSMOM signal and whether linear or nonlinear specifications perform better. In the backtest, models using individual lags outperformed TSMOM, while the linear models delivered the best performance and allowing for non-linearity and interaction did not improve performance. The time series implementations that trade on a fixed probability threshold performed poorly, which might be due to miscalibrations over time, the models being better at predicting cross sectional performance of commodities than forecasting broad directional moves, and the sorted portfolios benefiting from excluding low confidence predictions. Research question two asks whether adding term structure and other commodity risk factor-based features improves performance beyond the lagged return only model. In the backtest, the augmented feature set improved Sharpe ratio and accuracy, but factor regressions show that most of the incremental gains are explained by exposure to those same risk factors, leaving limited additional unexplained returns relative to the first model.

However, the results should be interpreted with caution as they are based on a single historical backtest path, do not account for implementation frictions such as transaction costs, and may reflect sample specific patterns arising from conducting multiple backtests.

# References


AQR Capital Management, LLC. 2025. "Time Series Momentum: Factors, Monthly." Dataset. January 31, 2025. Accessed December 28, 2025. *https://www.aqr.com/Insights/Datasets/Time-Series-Momentum-Factors-Monthly*.

Asness, Clifford S., Tobias J. Moskowitz, and Lasse Heje Pedersen. 2013. "Value and Momentum Everywhere." The Journal of Finance 68 (3): 929–85. https://doi.org/10.1111/jofi.12021.

Baltas, Akindynos-Nikolaos, and Robert Kosowski. "Momentum Strategies in Futures Markets and Trend-Following Funds." *SSRN Electronic Journal*, 2012. https://doi.org/10.2139/ssrn.1968996.

Beckmeyer, Heiner, and Timo Wiedemann. 2025. "All Days Are Not Created Equal: Understanding Momentum by Learning to Weight Past Returns." *Journal of Banking & Finance* 181 (December): 107565. https://doi.org/10.1016/j.jbankfin.2025.107565.

Boons, Martijn, And Melissa Porras Prado. 2018. "Basis-Momentum." The Journal of Finance 74 (1): 239–79. https://doi.org/10.1111/jofi.12738.

Breiman, Leo. 2001. "Random Forests." Machine Learning 45 (1): 5–32. https://doi.org/10.1023/a:1010933404324.

Bloomberg L.P. *Bloomberg Professional Service*. New York.

Chen, Tianqi, and Carlos Guestrin. 2016. "XGBoost: A Scalable Tree Boosting System." In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '16)*, 785–94. New York: ACM. https://doi.org/10.1145/2939672.2939785.

Della Corte, Pasquale, and Steven Riddiough. 2025. “Currency Speculation.” SSRN Working Paper. https://doi.org/10.2139/ssrn.5756369

Dhume, Deepa. 2011. “Using Durable Consumption Risk to Explain Commodities Returns.” Working paper, Harvard University.

Fama, Eugene F. 1970. “Efficient Capital Markets: A Review of Theory and Empirical Work.” *The Journal of Finance* 25 (2): 383. https://doi.org/10.2307/2325486.

Fama, Eugene F. 1991. “Efficient Capital Markets: II.” *The Journal of Finance* 46 (5): 1575–617. https://doi.org/10.1111/j.1540-6261.1991.tb04636.x.

Fama, Eugene F., and Kenneth R. French. 1993. “Common Risk Factors in the Returns on Stocks and Bonds.” Journal of Financial Economics 33 (1): 3–56. https://doi.org/10.1016/0304-405x(93)90023-5.

Fernandez-Perez, Adrian, Bart Frijns, Ana-Maria Fuertes, and Joelle Miffre. 2018. “The Skewness of Commodity Futures Returns.” Journal of Banking & Finance 86 (January): 143–58. https://doi.org/10.1016/j.jbankfin.2017.06.015.

Georgopoulou, Athina, and Jiaguo (George) Wang. 2016. “The Trend Is Your Friend: Time-Series Momentum Strategies across Equity and Commodity Markets.” *Review of Finance* 21 (4): 1557–92. https://doi.org/10.1093/rof/rfw048.

Gu, Shihao, Bryan Kelly, and Dacheng Xiu. 2020. “Empirical Asset Pricing via Machine Learning.” The Review of Financial Studies 33 (5): 2223–73. https://doi.org/10.1093/rfs/hhaa009.

Huang, Dashan, Jiangyuan Li, Liyao Wang, and Guofu Zhou. 2020. “Time Series Momentum: Is It There?” Journal of Financial Economics 135 (3): 774–94. https://doi.org/10.1016/j.jfineco.2019.08.004.

Hurst, Brian, Yao Hua Ooi, and Lasse Heje Pedersen. “Demystifying Managed Futures.” *SSRN Electronic Journal* 11, no. 3 (September 30, 2013): 42–58. https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2333776.

Jegadeesh, Narasimhan, And Sheridan Titman. 1993. “Returns to Buying Winners and Selling Losers: Implications for Stock Market Efficiency.” *The Journal of Finance* 48 (1): 65–91. https://doi.org/10.1111/j.1540-6261.1993.tb04702.x.

Koijen, Ralph S.J., Tobias J. Moskowitz, Lasse Heje Pedersen, and Evert B. Vrugt. 2018. “Carry.” Journal of Financial Economics 127 (2): 197–225. https://doi.org/10.1016/j.jfineco.2017.11.002.

Lim, Bryan, Stefan Zohren, and Stephen Roberts. 2019. “Enhancing Time Series Momentum Strategies Using Deep Neural Networks.” arXiv, ahead of print, 2019. https://doi.org/10.48550/ARXIV.1904.04912.

López de Prado, Marcos. 2018. *Advances in Financial Machine Learning*. Hoboken, NJ: John Wiley & Sons.

Lundberg, Scott M., Gabriel Erion, Hugh Chen, Alex DeGrave, Jordan M. Prutkin, Bala Nair, Ronit Katz, Jonathan Himmelfarb, Nisha Bansal, and Su-In Lee. 2020. “From Local Explanations to Global Understanding with Explainable AI for Trees.” *Nature Machine Intelligence* 2 (1): 56–67. https://doi.org/10.1038/s42256-019-0138-9.

Moskowitz, Tobias J., Yao Hua Ooi, and Lasse Heje Pedersen. 2012. “Time Series Momentum.” *Journal of Financial Economics* 104 (2): 228–50. https://doi.org/10.1016/j.jfineco.2011.11.003.

Pedregosa, Fabian, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, et al. 2011. "Scikit-learn: Machine Learning in Python." *Journal of Machine Learning Research* 12: 2825–2830.

Szymanowska, Marta, Frans de Roon, Theo Nijman, and Rob Van Den Goorbergh. 2014. "An Anatomy of Commodity Futures Risk Premia." The Journal of Finance 69 (1): 453–82. https://doi.org/10.1111/jofi.12096.

Yang, Fan. 2013. "Investment Shocks and the Commodity Basis Spread." Journal of Financial Economics 110 (1): 164–84. https://doi.org/10.1016/j.jfineco.2013.04.012.

# Appendix

## A1 Additional Information on the Replication

### A1.1 Results of the Full Replication

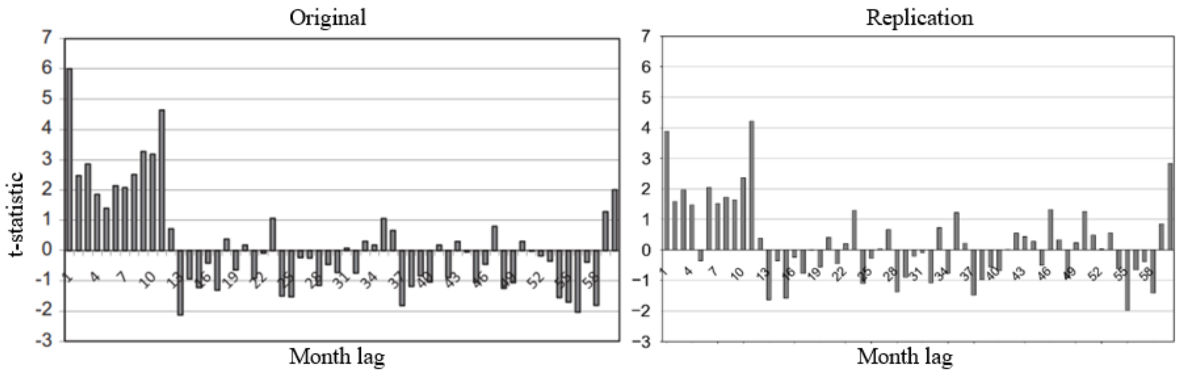


*Figure A 1: Lagged returns beta t-statistics across all assets - Original vs. Replication*

Replicating the TSMOM strategy across the whole investment universe yields very similar results as in MOP (2012). The pooled t-statistics of lagged monthly return betas predicting next-month returns align with the original paper with some minor deviations. As shown on Figure A 1, comparing lagged t-stats of pooled regressions on the asset classes individually show very similar patterns as in MOP (2012), except for FX forwards, which might be attributable to the later start date of contract availability in the data used.

*Table A 1: TSMOM Risk Adjusted Returns across asset classes - Original vs. Replication*

| | 1985/08-2025 | | | 1985-2009 | | | 2009/08-2025 | | |
|---|---|---|---|---|---|---|---|---|---|
| Strategy | Mean | Vol. | SR | Mean | Vol. | SR | Mean | Vol. | SR |
| TSMOM | 0.1178 | 0.1269 | 0.9285 | 0.1603 | 0.1214 | 1.3208 | 0.0502 | 0.1335 | 0.3763 |
| Passive Long | 0.1033 | 0.1325 | 0.7796 | 0.1253 | 0.1375 | 0.9114 | 0.0701 | 0.1238 | 0.5668 |
| Equal Weight | 0.0332 | 0.0729 | 0.4552 | 0.0398 | 0.0690 | 0.5767 | 0.0237 | 0.0790 | 0.2996 |
| Original Paper | 0.1228 | 0.1260 | 0.9744 | 0.1675 | 0.1194 | 1.4031 | 0.0477 | 0.1336 | 0.3567 |

In terms of economic performance, backtesting the TSMOM strategy and comparing it with the original results extended up until 2025 shows a very similar performance. While the replicated strategy initially underperforms in early years, comparing the strategies performance by asset class on Figure A 4 suggests this deviation is attributable to the later starting dates of specific contracts.

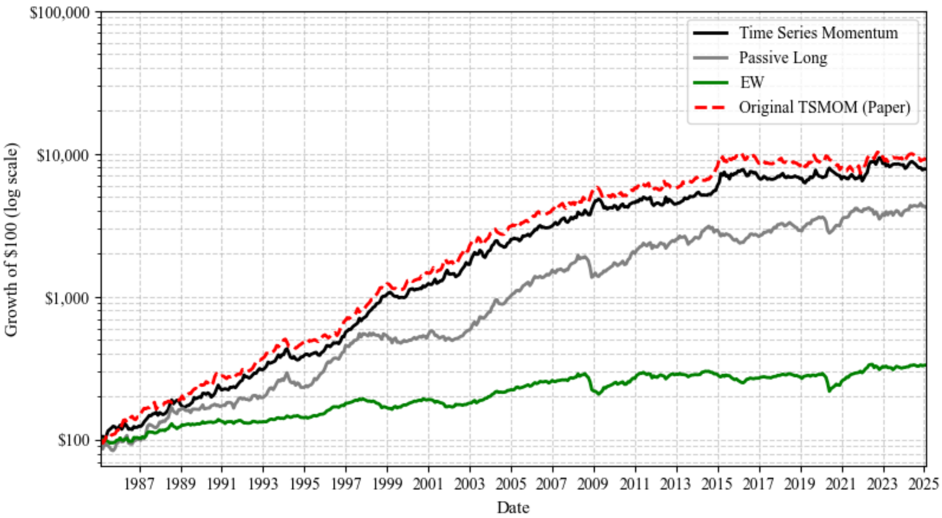


*Figure A 2: Cumulative Returns All Asset Classes - Original vs. Replication*

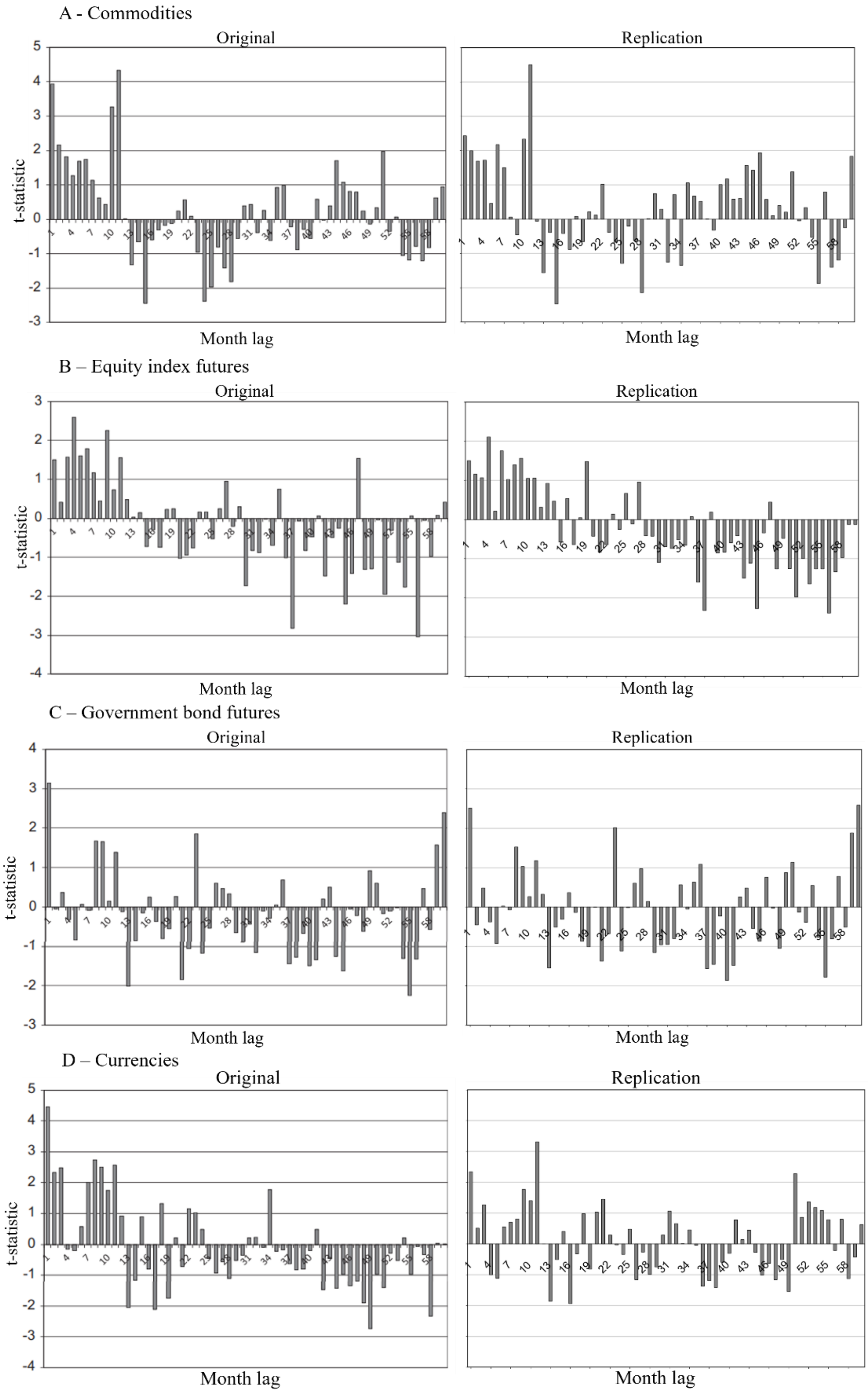


*Figure A 3: Pooled t-stats - All Asset Classes - Replication vs. Original*

### A1.1.1 Cumulative Returns Per Asset Class - Replication Versus Original

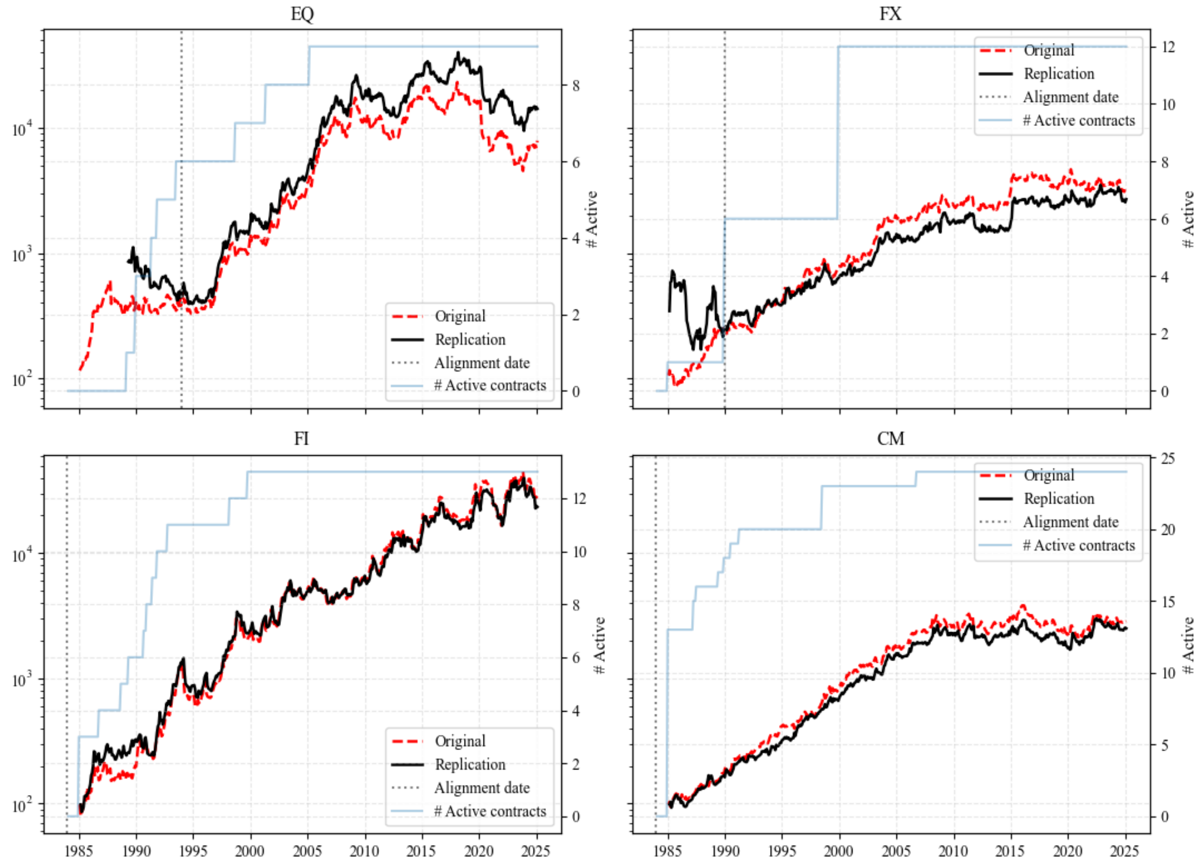


*Figure A 4: Cumulative TSMOM Returns per Asset Class - Original vs. Replication*

Figure A 4 illustrates the number of active contracts for which data is available and the cumulative returns over time of the original paper's TSMOM strategy and the replication. Hereby, the cumulative return indices are aligned at a period, marked by the dotted vertical line, where there is data available for a sufficient number of contracts.

## A1.2 Original Data vs Replication

*Table A 2: Summary Statistics - Original vs. Replication Using Adjusted "Active Future" Returns*

| | | Original | | | Replication | | |
|---|---|---|---|---|---|---|---|
| Asset Class | Asset Name | Start Date | Mean | Volatility | Start Date | Mean | Volatility |
| **Commodities** | Aluminium | Jan-79 | 0,97% | 23,50% | Jul-97 | -3,97% | 19,13% |
| | Brent Crude | Apr-89 | 13,87% | 32,51% | Jun-88 | 19,18% | 35,57% |
| | Cocoa | Jan-65 | 5,61% | 32,38% | Jan-65 | 6,62% | 32,68% |
| | Coffee | Mar-74 | 5,72% | 38,62% | Aug-72 | 7,26% | 40,56% |
| | Copper | Jan-77 | 8,90% | 27,39% | Dec-88 | 9,50% | 26,65% |
| | Corn | Jan-65 | -3,19% | 24,37% | Jan-65 | -1,20% | 23,86% |
| | Cotton | Aug-67 | 1,41% | 24,35% | Jan-65 | 0,46% | 24,76% |
| | Gasoil | Oct-84 | 11,95% | 33,18% | Jul-89 | 17,64% | 35,57% |
| | Gold | Dec-69 | 5,36% | 21,37% | Jan-75 | 1,33% | 19,43% |
| | Heating Oil | Dec-78 | 9,79% | 33,78% | Jul-86 | 18,91% | 37,84% |
| | Lean Hogs | Feb-66 | 3,39% | 26,01% | Apr-86 | -1,05% | 25,33% |
| | Live Cattle | Jan-65 | 4,52% | 17,14% | Jan-65 | 7,19% | 17,61% |
| | Natural Gas | Apr-90 | -9,74% | 53,30% | Apr-90 | 6,91% | 62,44% |
| | Nickel | Jan-93 | 12,69% | 35,76% | Jul-97 | 17,50% | 38,49% |
| | Platinum | Jan-92 | 13,15% | 20,95% | Apr-86 | 7,63% | 22,12% |
| | RBOB Gasoline | Dec-84 | 15,92% | 37,36% | Oct-05 | 9,59% | 44,59% |
| | Silver | Jan-65 | 3,17% | 31,11% | Jan-75 | 2,21% | 31,47% |
| | Soybean Meal | Sep-83 | 6,14% | 24,59% | Jan-65 | 12,02% | 30,70% |
| | Soybean Oil | Dec-90 | 1,07% | 25,39% | Jan-65 | 12,15% | 33,89% |
| | Soybeans | Jan-65 | 5,57% | 27,26% | Jan-65 | 5,85% | 27,22% |
| | Sugar | Jan-65 | 4,44% | 42,87% | Jan-65 | 0,05% | 43,79% |
| | WTI Crude | Mar-83 | 11,61% | 34,72% | Apr-83 | 15,56% | 36,50% |
| | Wheat | Jan-65 | -1,84% | 25,11% | Jan-65 | 0,44% | 25,31% |
| | Zinc | Jan-91 | 1,98% | 24,76% | Jul-97 | 2,55% | 27,48% |
| **Equity Index** | AEX | Jan-75 | 7,72% | 19,18% | Jan-89 | 7,19% | 20,93% |
| | CAC 40 | Jan-75 | 6,73% | 20,87% | Dec-88 | 4,61% | 19,90% |
| | DAX | Jan-75 | 6,33% | 20,41% | Nov-90 | 5,80% | 21,95% |
| | FTSE 100 | Jan-75 | 6,97% | 17,77% | Mar-88 | 3,36% | 15,18% |
| | FTSE MIB | Jun-78 | 6,13% | 24,59% | Mar-04 | 1,98% | 20,04% |
| | IBEX 35 | Jan-80 | 9,37% | 21,84% | Jul-92 | 10,11% | 21,19% |
| | S&P 500 | Jan-65 | 3,47% | 15,45% | Sep-97 | 0,78% | 16,37% |
| | SPI 200 | Jan-77 | 7,25% | 18,33% | May-00 | 4,56% | 13,54% |
| | TOPIX | Jul-76 | 2,29% | 18,66% | May-90 | -3,58% | 20,22% |
| **Bonds** | Australia 10Y Bond | Dec-85 | 3,83% | 8,53% | Sep-87 | 0,49% | 1,30% |
| | Australia 3Y Bond | Jan-92 | 1,34% | 2,57% | Dec-89 | 0,92% | 1,45% |
| | Canada 10Y Bond | Dec-84 | 4,04% | 7,36% | Sep-89 | 3,68% | 6,20% |
| | Euro Bobl | Jan-93 | 2,56% | 3,22% | Oct-91 | 2,58% | 3,21% |
| | Euro Bund | Dec-79 | 2,40% | 5,74% | Nov-90 | 3,50% | 4,95% |
| | Euro Buxl | Dec-98 | 4,71% | 11,70% | Oct-98 | 2,73% | 8,52% |
| | Euro Schatz | Mar-97 | 1,02% | 1,53% | Mar-97 | 0,87% | 1,40% |
| | Japan 10Y Bond | Dec-81 | 3,66% | 5,40% | Oct-85 | 3,43% | 5,82% |
| | UK Long Gilt | Dec-79 | 3,00% | 9,12% | Nov-82 | 2,07% | 7,79% |
| | US 10Y Note | Dec-79 | 3,80% | 9,30% | May-82 | 5,04% | 7,32% |
| | US 2Y Note | Apr-96 | 1,65% | 1,86% | Jun-90 | 1,77% | 1,82% |
| | US 5Y Note | Jan-90 | 3,17% | 4,25% | May-88 | 3,04% | 4,39% |
| | US Long Bond | Jan-90 | 9,50% | 18,56% | Aug-77 | 3,60% | 11,51% |
| **Currencies** | AUDUSD | Mar-72 | 1,85% | 10,86% | Jan-71 | 0,04% | 10,72% |
| | CADUSD | Mar-72 | 0,60% | 6,29% | Jan-71 | 0,09% | 6,12% |
| | CHFUSD | Sep-71 | 1,34% | 12,33% | Jan-71 | 4,40% | 12,16% |
| | EURUSD | Sep-71 | 1,57% | 11,21% | Jan-75 | 0,82% | 10,61% |
| | GBPUSD | Sep-71 | 1,39% | 10,32% | Jan-71 | -0,50% | 10,09% |
| | JPYUSD | Sep-71 | 1,35% | 11,66% | Jan-71 | 4,10% | 11,43% |
| | NOKUSD | Feb-78 | 1,37% | 10,56% | Feb-71 | 1,05% | 10,09% |
| | NZDUSD | Feb-78 | 2,31% | 12,01% | Jan-71 | -0,41% | 11,76% |
| | SEKUSD | Feb-78 | -0,05% | 11,06% | Jan-71 | -0,25% | 10,72% |

### A1.2.1 FX Forward Return Computation Using Bloomberg Forward Points

MOP (2012) likely obtained currency forward returns by computing the spot returns and the implied interest rate differential from one-month forward interest rates.

In contrast, daily excess returns were calculated by combining daily spot rate changes with the interest differential implied by market forward prices. To derive these returns, the nine underlying USD currency pairs were first used to compute 12 specific cross-currency pairs, obtaining both cross-spot and cross-forward prices. Spot returns represent the percentage change in these cross-exchange rates between consecutive days.

The interest differential, or daily carry, was derived from Bloomberg 1-month forward points by constructing outright forward prices and linearizing the resulting forward discount or premium over a 30-day period. The total daily excess return is the sum of this daily carry and the spot return. The returns represent quoted one month forward rates, for which the expiration has been linearized, assuming the interest differential would be realized evenly over time. Thus, it represents opening a new forward position everyday and earning the daily fraction of the interest rate differential.

# A2 Feature Computation and Basis Interpolation

## A2.1.1 Basis Computation and Imputation of Maturity Difference

To make the basis comparable across commodities and to express it as a percentage price difference per unit of maturity, the basis is computed between adjacent contracts as

$$B_{j,t}^{T1,raw} = \frac{F_t^{j,T2}}{F_t^{j,T1}} - 1 \text{ and } B_{j,t}^{T2,raw} = \frac{F_t^{j,T3}}{F_t^{j,T2}} - 1 \tag{5}$$

and then scaled to a constant 30-day maturity gap using

$$B_{j,t}^{T} = B_{j,t}^{T,raw} \cdot 30/\Delta T, \tag{6}$$

where $\Delta T$is the calendar-day difference between the corresponding expiration dates. The estimation of $\Delta T$requires imputation because expiration date fields in Bloomberg generic futures are frequently missing.

If the maturity gap $\Delta T$differs systematically across calendar months (e.g., for commodities that trade only in specific delivery months such as corn), missing gaps are imputed using month-specific medians. If the maturity gap is approximately constant across months, missing gaps are imputed using the global median of that specific commodity.

### A2.1.2 Feature Computation

Basis Mean:

$$\overline{B}_{j,t}^{T1} = \frac{1}{12}\sum_{k=1}^{12} B_{j,t-k}^{T1} \tag{7}$$

Basis z-score

$$Z_{j,t}^{T1} = \frac{B_{j,t}^{T1} - \overline{B}_{j,t}^{T1}}{\sqrt{\frac{1}{11}\sum_{k=1}^{12}\left(B_{j,t-k}^{T1} - \overline{B}_{j,t}^{T1}\right)^2}} \tag{8}$$

Skewness:

$$Skewness_{j,t} = \frac{1}{252}\sum_{d=1}^{252}\frac{\left(R_{t-d,j}^{T1} - \mu_{j,t}\right)^3}{\sigma_{j,t}^3} \tag{9}$$

Value:

$$Value = ln\left(\frac{\frac{1}{13}\sum_{m=54}^{66} F_{t-m}^{j,T1}}{F_t^{j,T1}}\right) \tag{10}$$

Basis Momentum:

$$BM_{j,t} = \prod_{s=t-11}^{t}\left(1 + r_{j,s}^{T1}\right) - \prod_{s=t-11}^{t}\left(1 + r_{j,s}^{T2}\right) \tag{11}$$

Basis Mom decomposition:

$$\text{Average Curvature}_{\mathrm{j,t}} = \sum_{s=t-11}^{t-1} B_{j,s}^{T2} - \sum_{s=t-11}^{t-1} B_{j,s}^{T1} \tag{12}$$

$$\Delta\text{Slope}_{\mathrm{j,t}} = B_{j,t-12}^{T2} - B_{j,t}^{T1} \tag{13}$$

Coefficient of Variation:

$$CV_{j,t}^{(12)} = \frac{\frac{1}{11}\sum_{k=1}^{12}\left(r_{j,t-k} - \overline{r_{j,t}^{(12)}}\right)^2}{\left|\overline{r_{j,t}^{(12)}}\right|} \tag{14}$$

$$\overline{r_{j,t}^{(12)}} = \frac{1}{12}\sum_{k=1}^{12} r_{j,t-k} \tag{15}$$

While not explicitly mentioned in Szymanowska et al. (2014), the coefficient of variation is implemented using the absolute mean.

# A3 Additional Information on the Backtests

### A3.1.1 Model 1 – Using Return Features Only

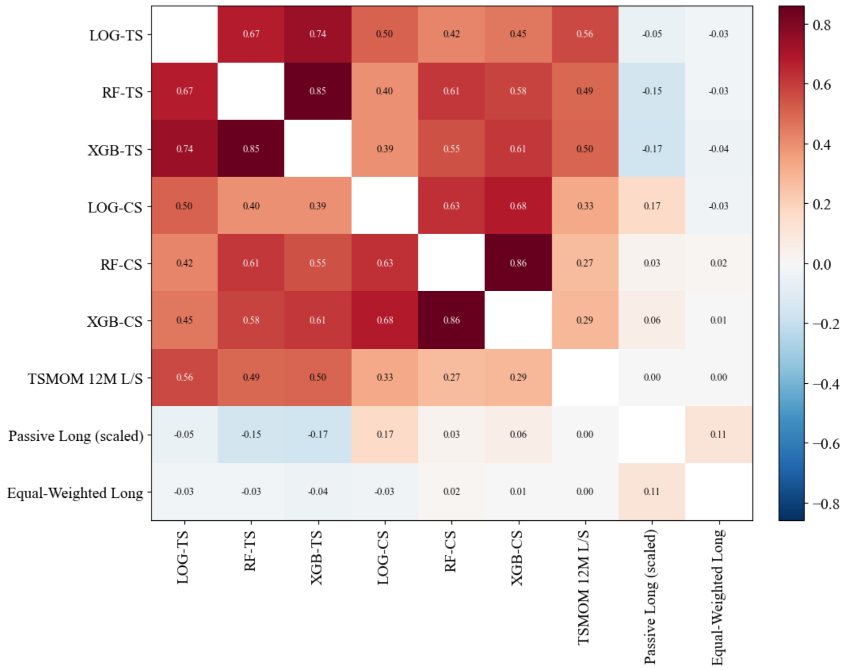


*Figure A 5: Model 1 - Correlation Matrix of Strategy Returns*

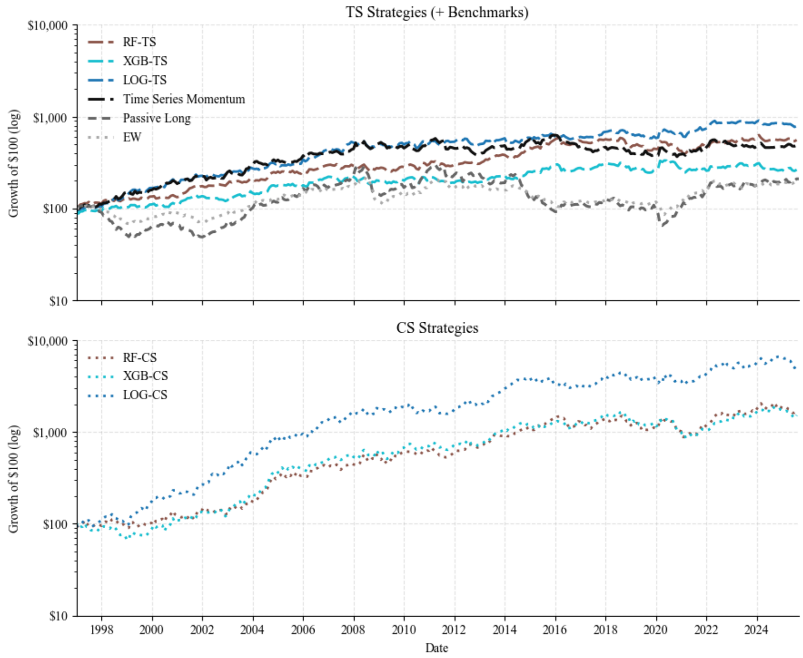


*Figure A 6: Cumulative Returns of All Strategies - Model 1*

### A3.1.2 Model 2 – Augmented Features

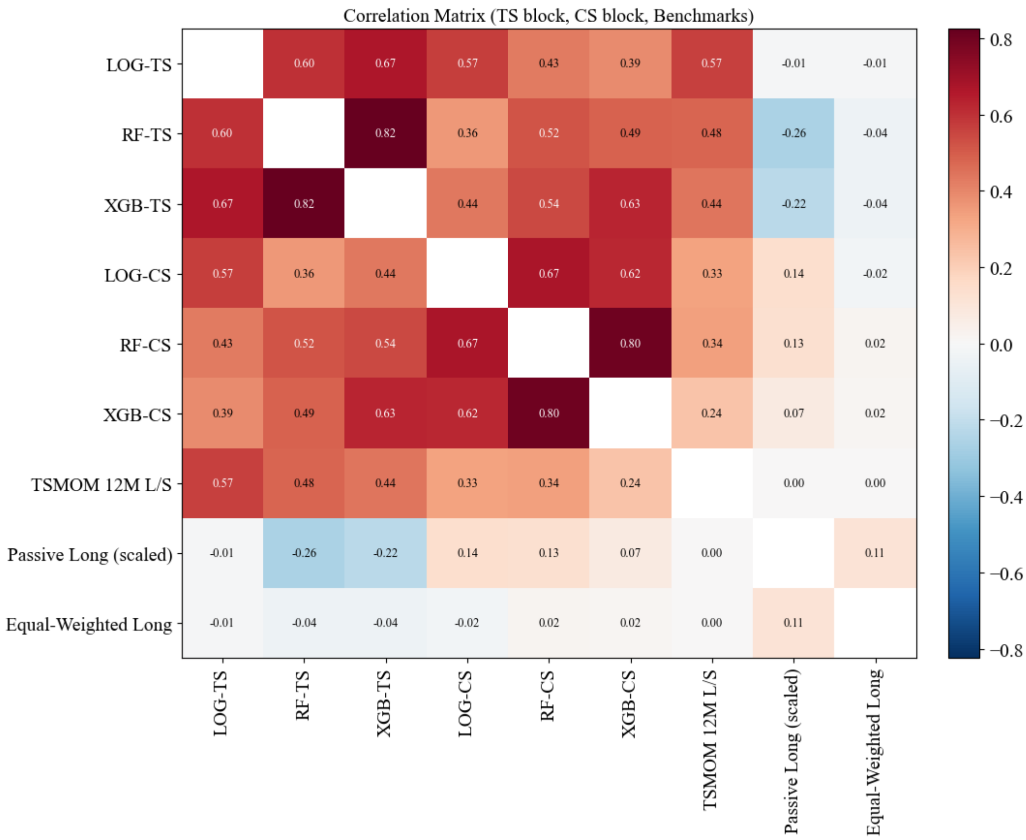


*Figure A 7: Model 2 - Correlation Matrix of Strategy Returns*

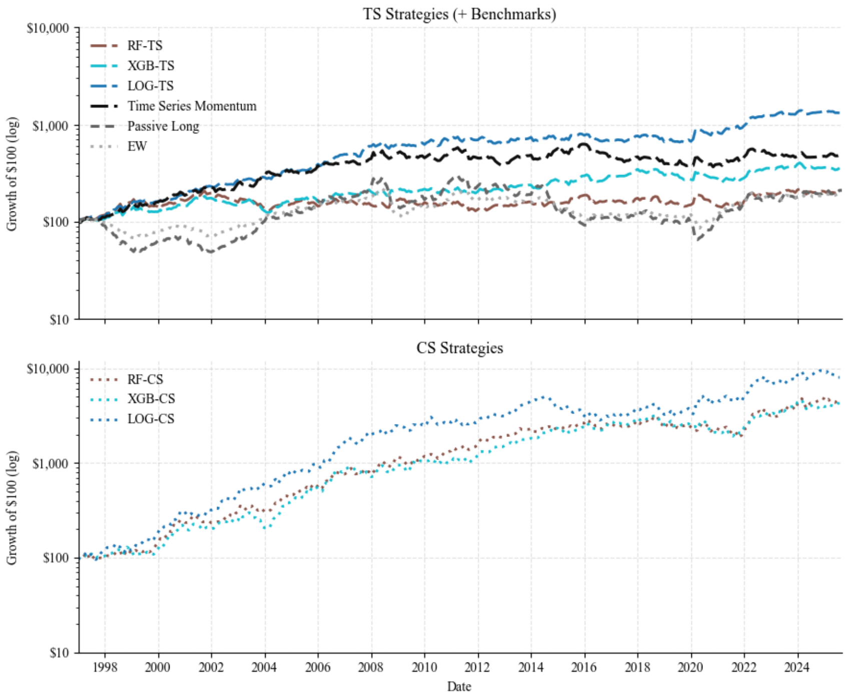


*Figure A 8: Cumulative Returns of All Strategies - Model 2*

# A4 Additional Information on Model Explanations

## A4.1 Model 1- Using Return Features Only

### A4.1.1 Random Forest

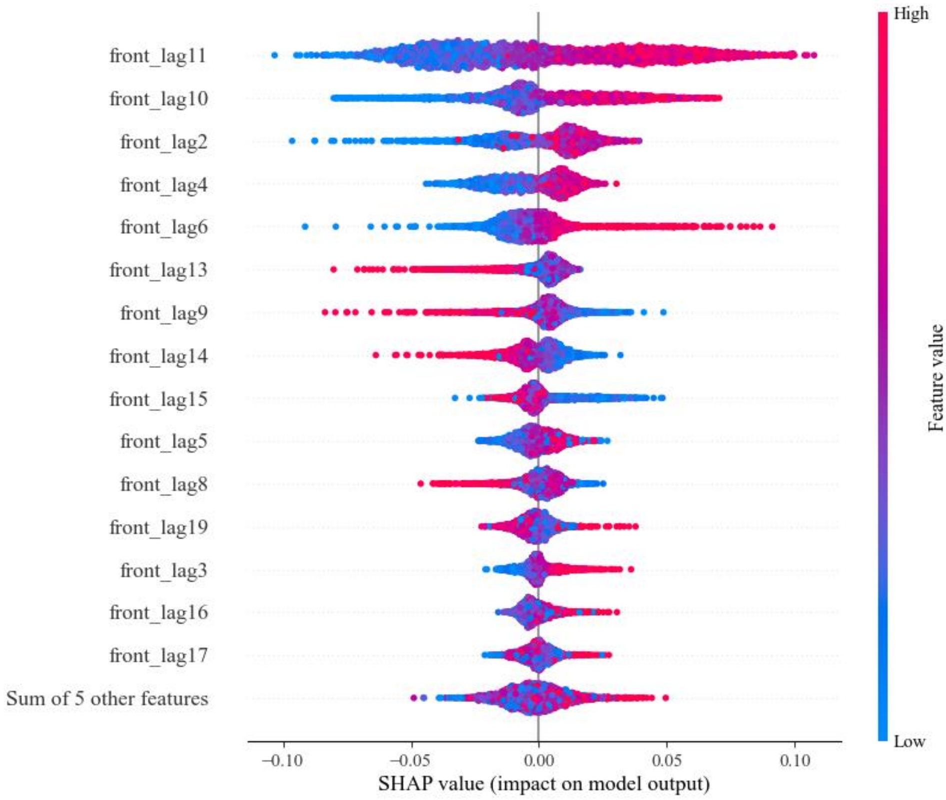


*Figure A 9: Model 1 - RF - SHAP Values Beeswarm Plot*

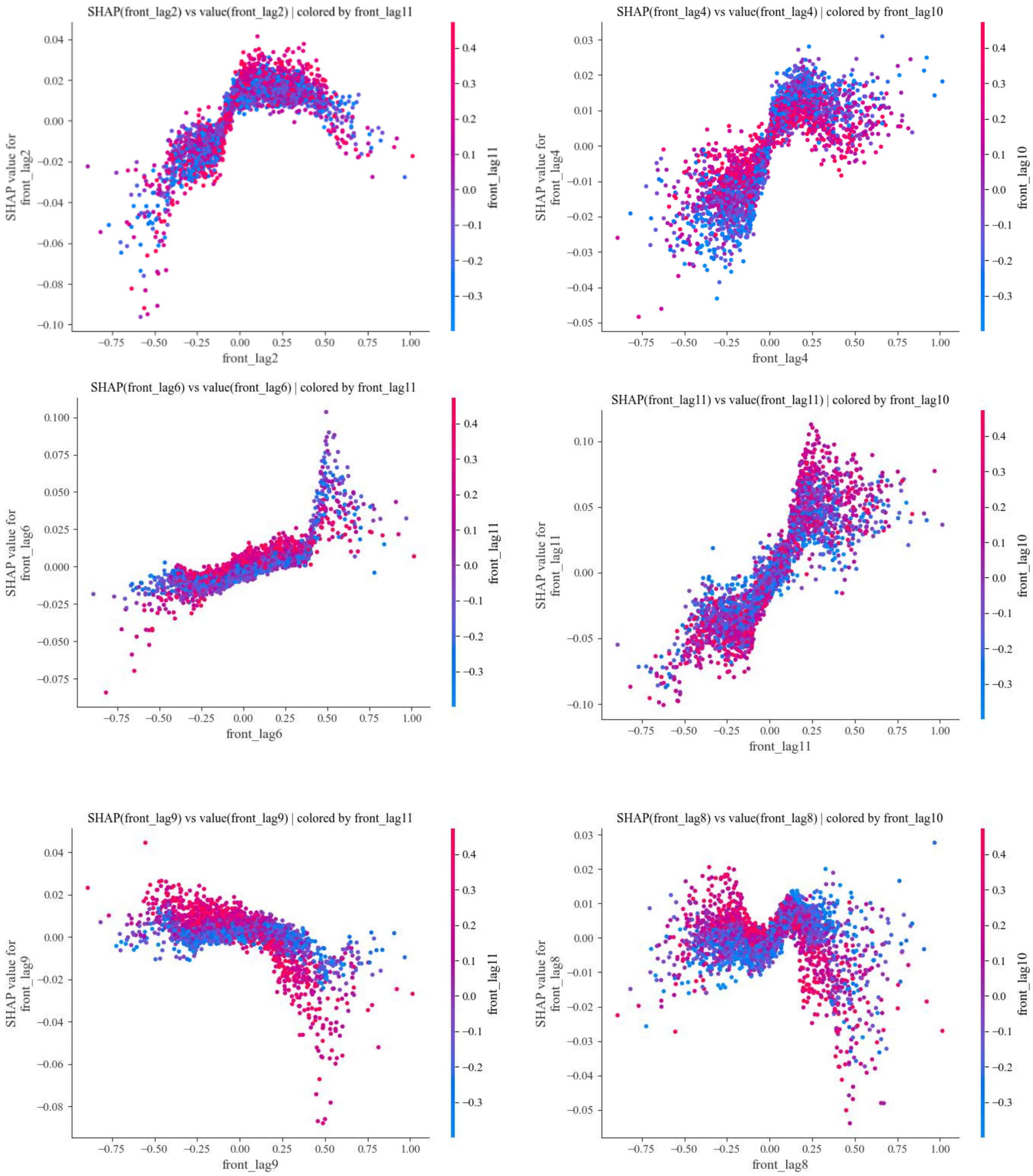


*Figure A 10: Model 1 - RF - SHAP Values Dependence Plots*

### A4.1.2 XGBoost

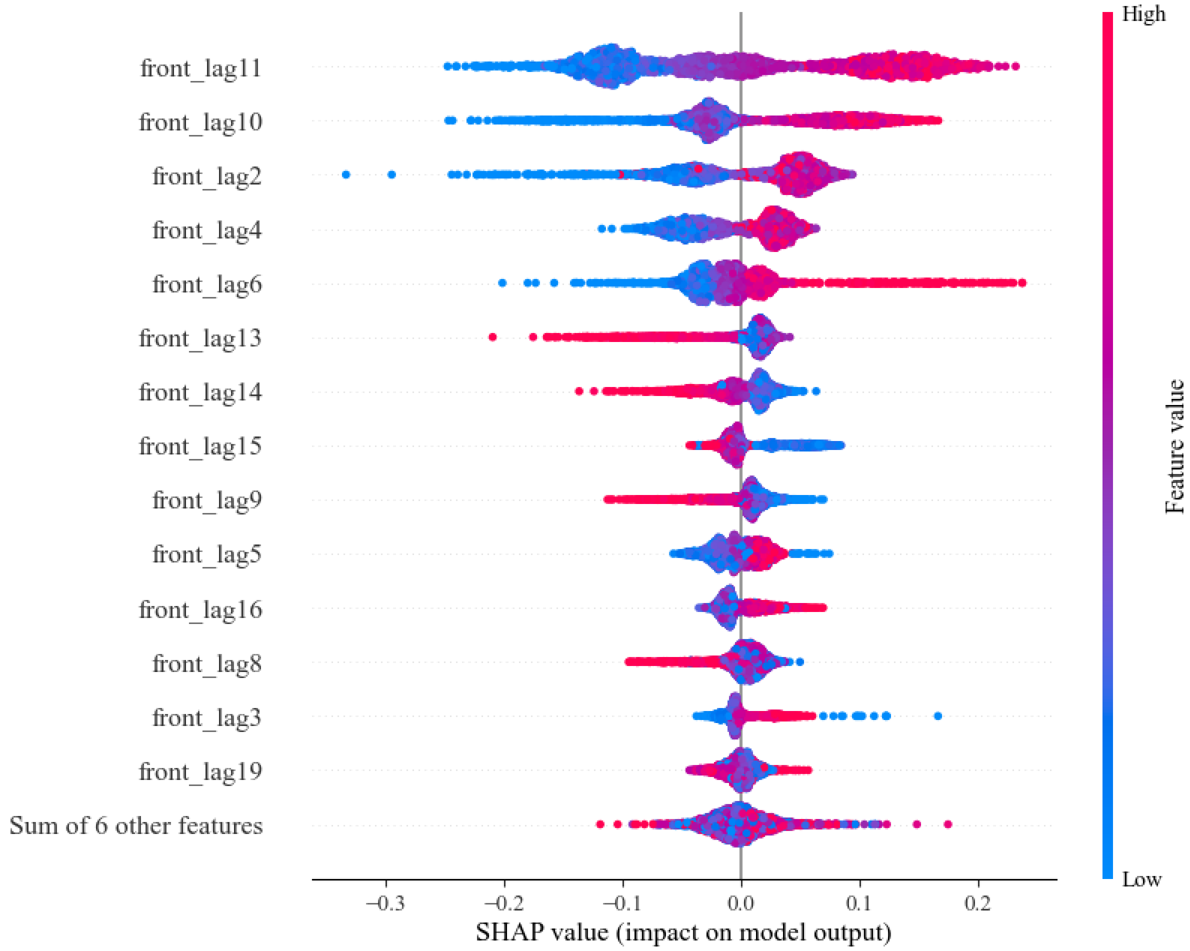


*Figure A 11: Model 1 - XGB - SHAP Values Beeswarm Plot*

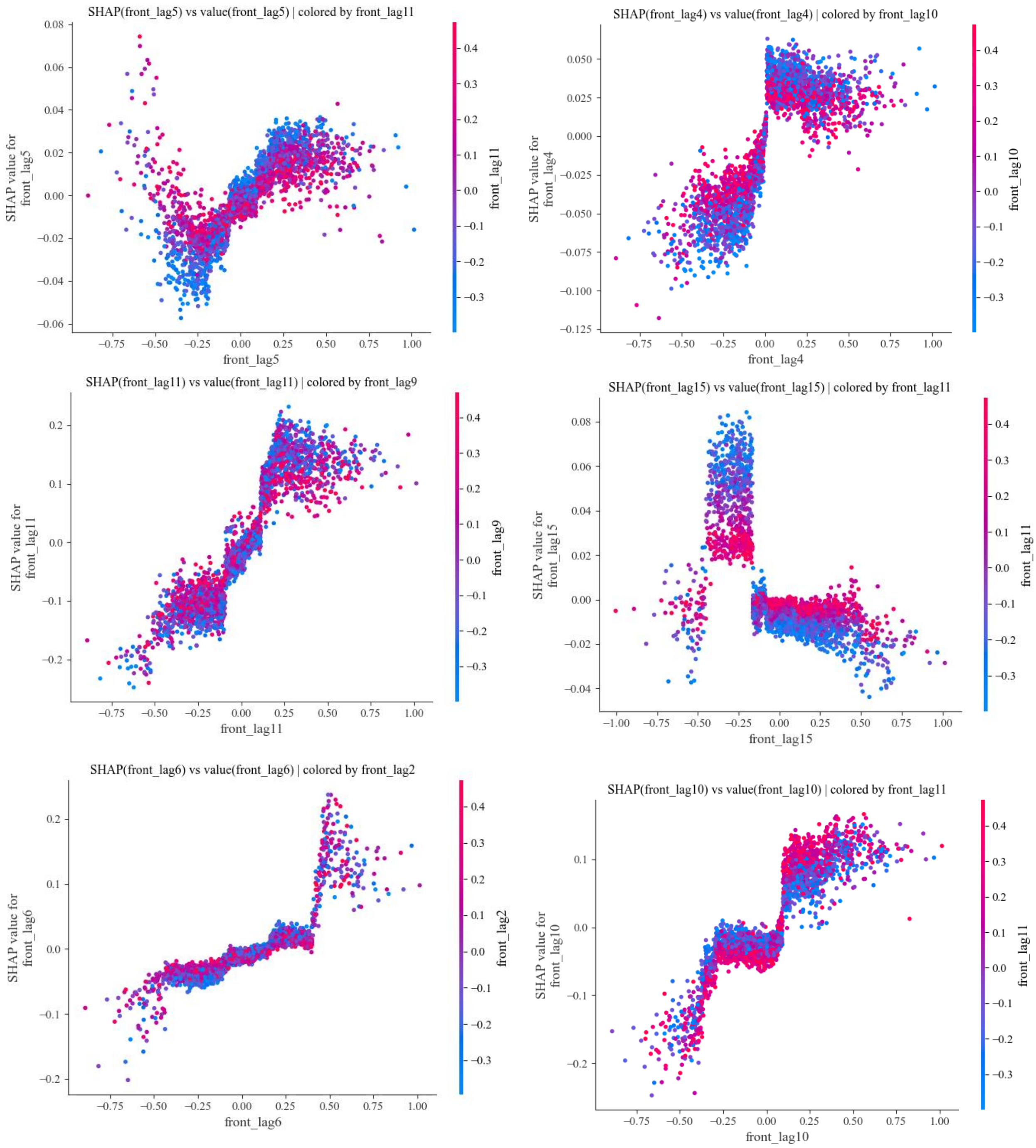


*Figure A 12: Model 1 - XGB - SHAP Values Dependence Plots*

### A4.1.3 Logistic Regression

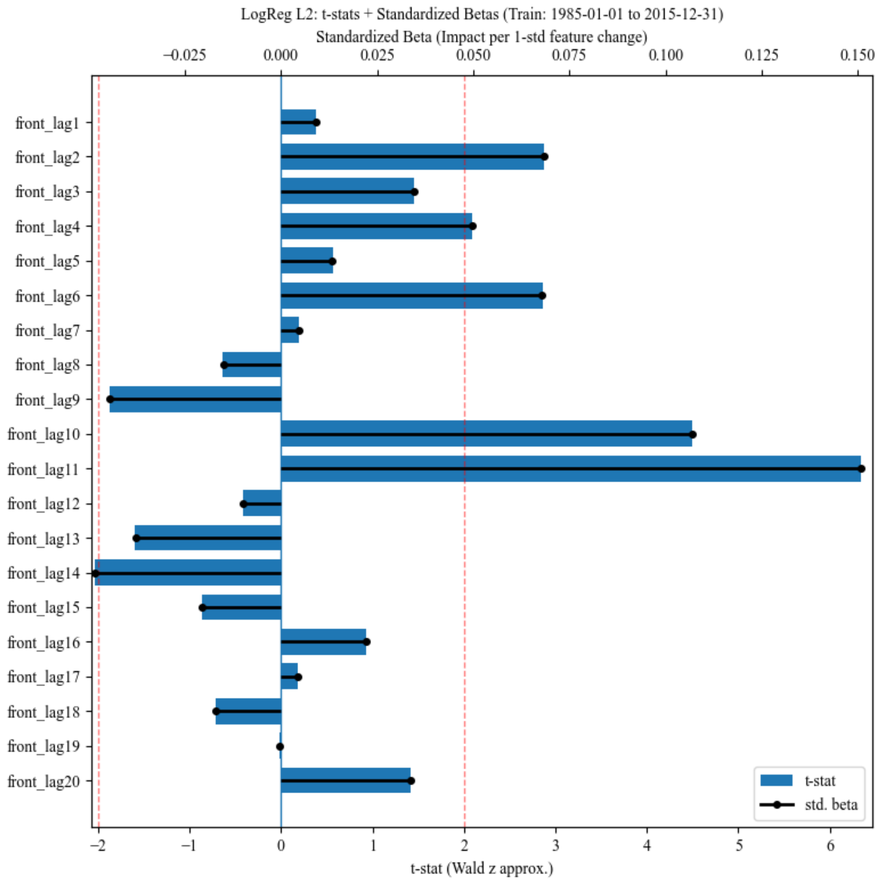


*Figure A 13: Model 1 - LogReg - Standardizes Betas and t-stats*

## A4.2 Second Model – Augmented Features

### A4.2.1 Random Forest

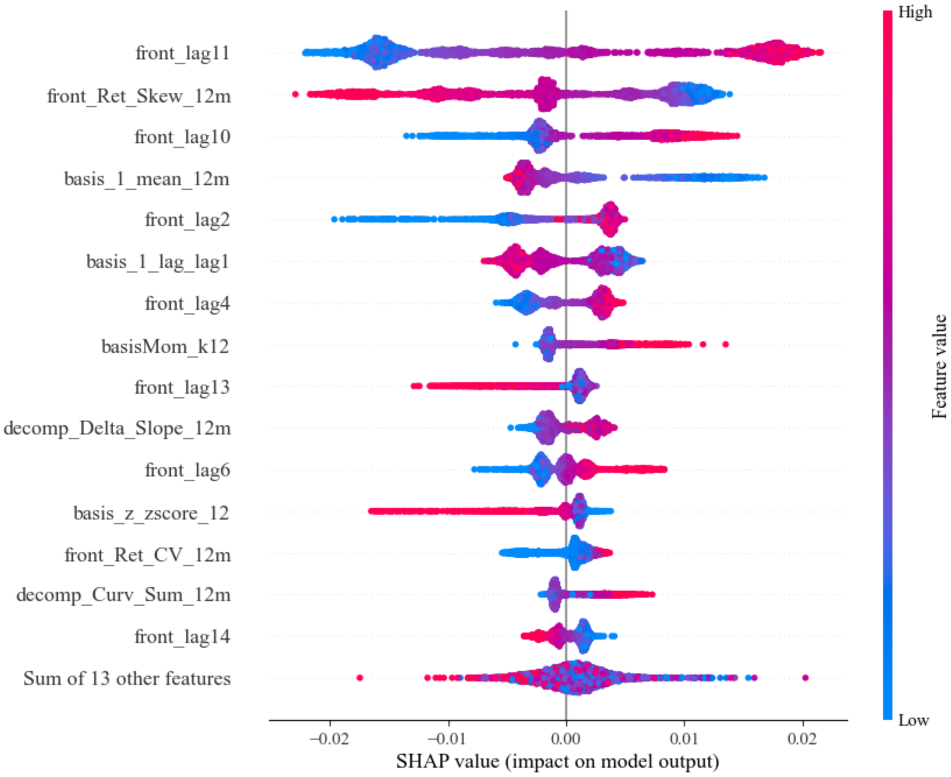


*Figure A 14: Model 2 - RF - SHAP Values Beeswarm Plot*

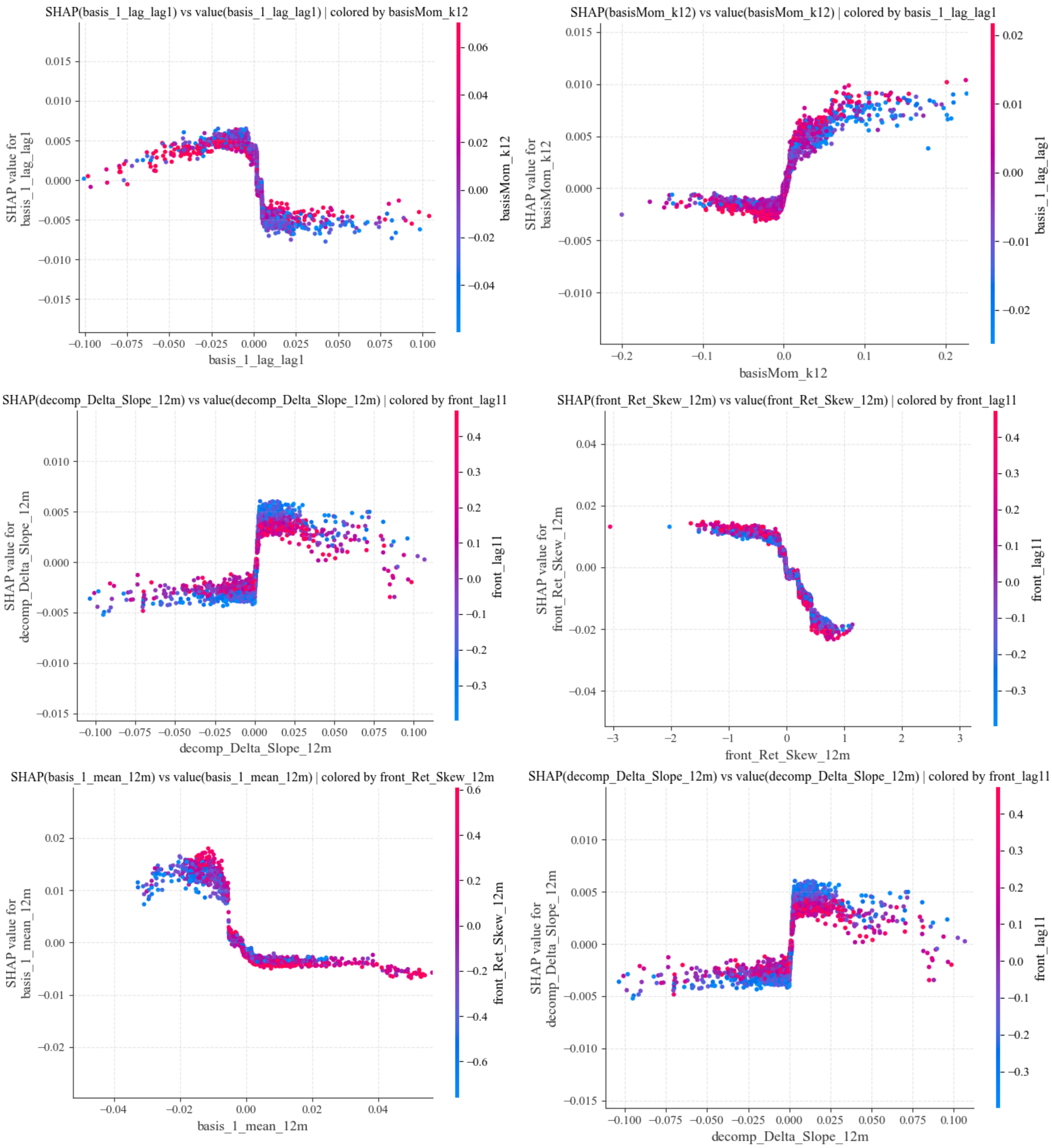


*Figure A 15: Model 2 - RF - SHAP Values Dependence Plots*

### A4.2.2 XGBoost

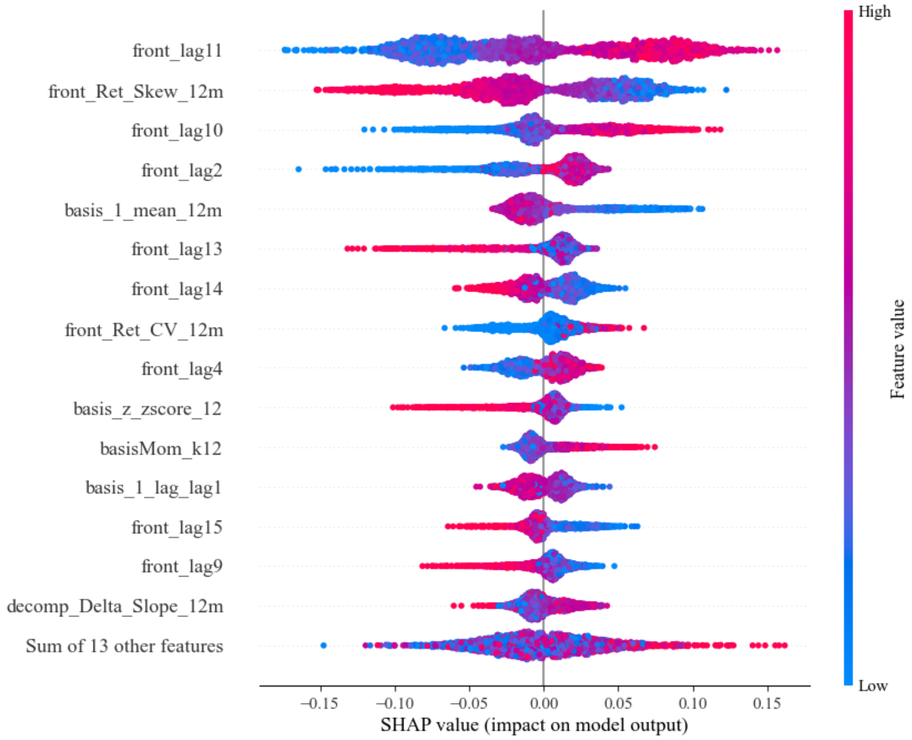


*Figure A 16: Model 2 - XGB - SHAP Values Beeswarm Plot*

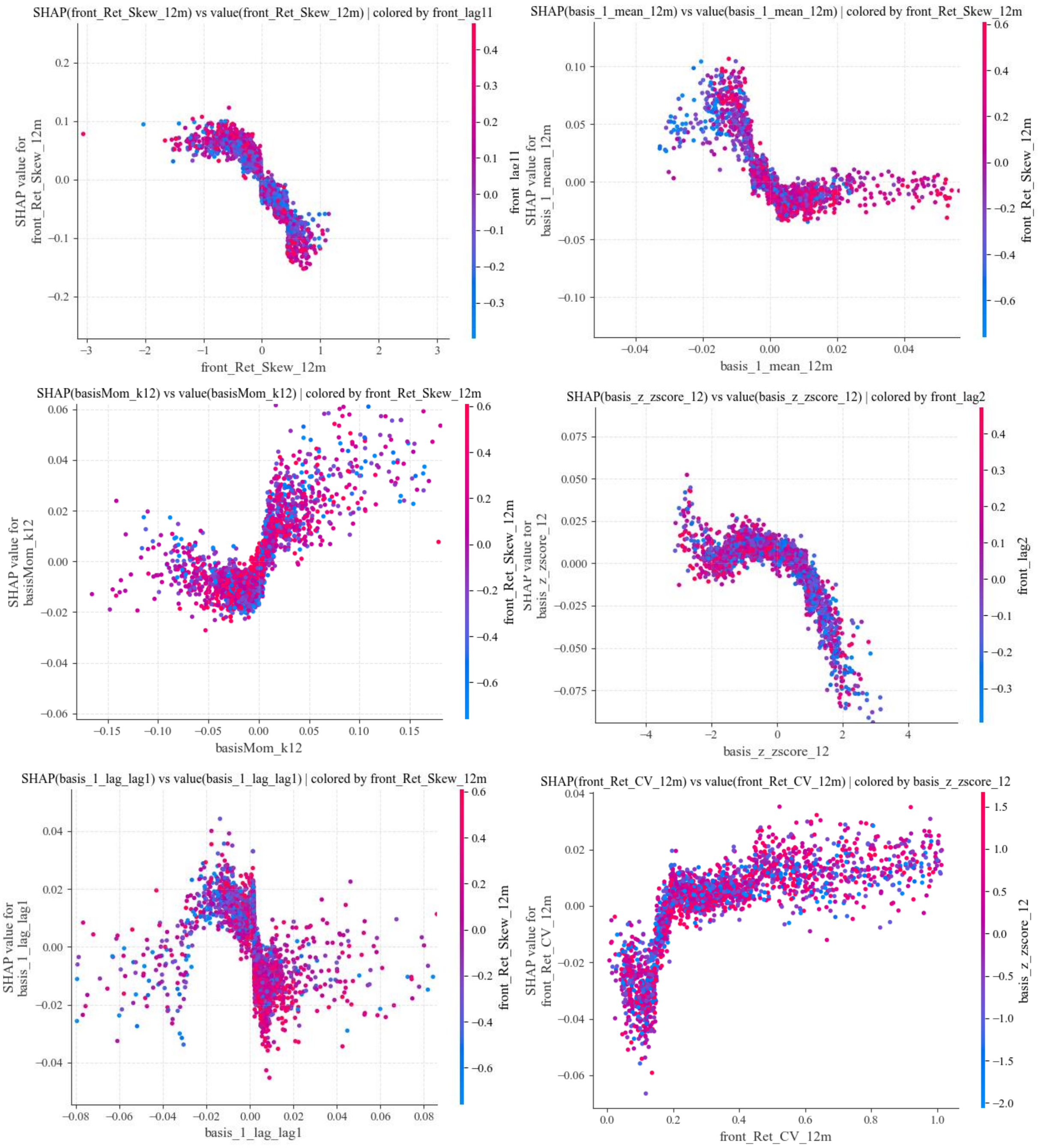


*Figure A 17: Model 2 - XGB - SHAP Values Dependence Plots*

### A4.2.3 Logistic Regression

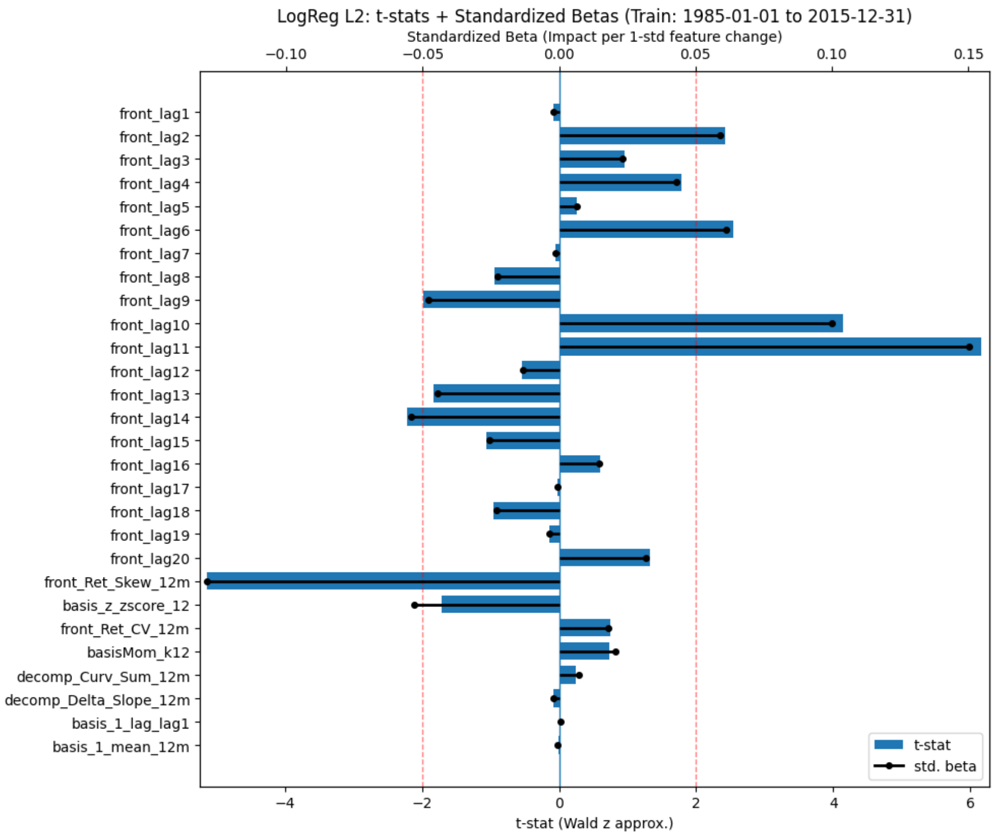


*Figure A 18: Model 2 - LogReg - Standardizes Betas and t-stats*

## A5 Factor Portfolio Construction & Factor Returns

Factor returns in the spirit of Fama (1993) have been constructed to test whether the ML models strategies can be explained by them or if they deliver returns beyond.

The factor portfolios are constructed by sorting commodities on the relevant characteristic at the beginning of month $t$, going long the top 30 percent and short the bottom 30 percent, and computing the equal-weighted long–short return at the end of month $t$.

All the factors below have been computed using the dataset of 24 commodities as described above. Returns are realized on the front contract rolled one month before expiration and all assets available at each month $t$ are considered.

Information on the computation can be found below. Table A 3 shows summary stats on the factor returns and Figure A 19 the correlation matrix of factor returns in the OOS-period.

**AVG (Average Market Factor):**

The commodity "market" factor (AVG) is the monthly rebalanced, equal-weighted portfolio of all commodity futures in the sample, used as a broad proxy for aggregate commodity market returns (Yang, 2013).

**Basis Factor:**

Spot premium basis factor returns following (Szymanowska et al. 2014). Sorted on the basis of the front to the second month contracts price.

**Value Factor:**

Sorted on the log ratio of the average "spot proxy" price about five years ago (average from 4.5–5.5 years earlier) to today's price, so "cheap" commodities are those priced low relative to their longer-run level (i.e., roughly the negative 5-year price change) (Assness, Moskowitz, and Pedersen 2013).

$$Value = ln\left(\frac{\frac{1}{13}\sum_{m=54}^{66} F_{t-m}^{j,T1}}{F_t^{j,T1}}\right) \tag{16}$$

**Skewness Factor:**

Commodities sorted on the skewness of daily returns over the last year (252 days). Goes long in the lowest 30% and short in the highest 30%. (Fernandez-Perez et al. 2018)

**Dollar Beta Factor:**

The dollar factor controls for dollar exposure in commodity futures. Prior evidence shows that some commodities have a significant negative exposure to changes in the U.S. dollar (Erb and Harvey 2006). Each month, commodities are sorted on their dollar beta estimated from a 60-month rolling regression of monthly commodity futures returns on changes in the U.S. dollar against a basket of foreign currencies. The factor goes long the lowest 30% of dollar betas and short the highest 30% (Szymanowska et al. 2014; Erb and Harvey 2006). Data on dollar returns has been obtained from Della Corte and Riddiough (2025).

*Table A 3: Factor Returns - OOS Period: 12/1996 – 08/2025*

| Factor | Ann. Mean | Ann. Vol. | SR |
|---|---|---|---|
| Dollar | 5.86% | 22.41% | 0.262 |
| AVG-Market | 3.41% | 14.49% | 0.235 |
| Basis | 4.45% | 19.65% | 0.226 |
| Basis Momentum | 8.60% | 17.08% | 0.503 |
| Value | 2.49% | 21.95% | 0.114 |
| Skewness | 13.29% | 19.57% | 0.679 |
| CS-Mom | 4.18% | 21.63% | 0.193 |
| CV | 0.66% | 18.11% | 0.036 |
| EW-Factors | 5.38% | 8.07% | 0.667 |

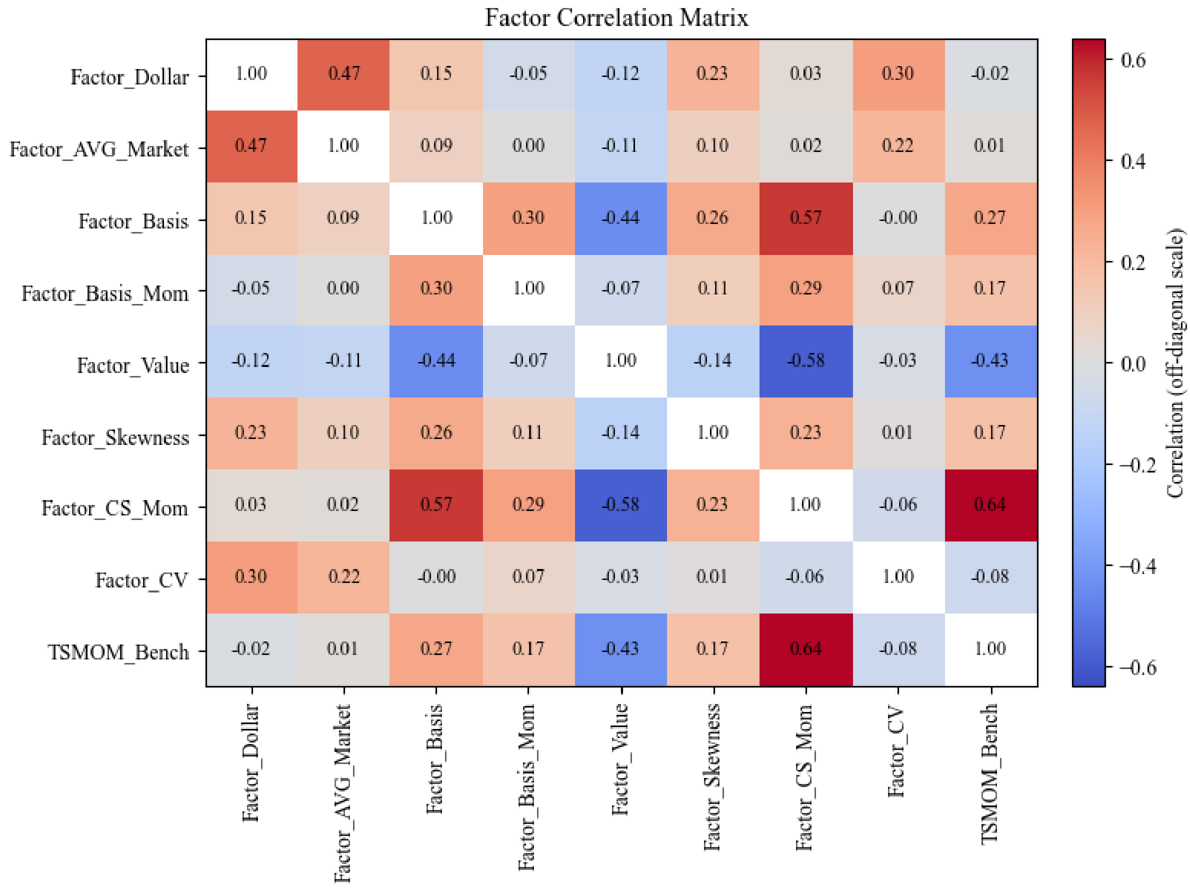


*Figure A 19: Correlation of Factor Returns in the OOS-period*

# A6 Hyperparameter Candidates & Tuned Params

Table A 4 and Table A 5 report the Random Forest and XGB hyperparameter candidates; the parameters obtained in the first model setup (returns only) are marked with parentheses (), and those obtained in the second setup are marked with brackets [].

*Table A 4: Random Forest Hyperparameter*

| RF Hyperparameter | Candidates |
|---|---|
| n_estimators | [(**500**)] |
| max_depth | [**3**], 4, 5, 7, (**10**) |
| min_samples_leaf | (**5**), [**10**], 20 |
| max_features | (**None**), 5, sqrt, [**log2**] |
| criterion | [(**gini**)] |

The number of trees is fixed at 500 due to computational limitations and, as Random Forest generalization error converges with the number of trees and additional estimators yield only marginal gains (Breiman 2001).

*Table A 5: XGB Hyperparameter*

| XGB Hyperparameter | Candidates |
|---|---|
| n_estimators | [**100**], (**200**), 300 |
| max_depth | 3, (**4**), 5, [**6**], 7 |
| learning_rate | [(**0.01**)], 0.05, 0,10 |
| subsample | [**0.6**], (**0.8**) |
| colsample_bytree | [(**0.6**)], 0.8 |