BOX A.2: DE-MEANING AND STANDARDIZATION OF FEATURES

Most raw features in the dataset – including prices, volumes, and macroeconomic levels – introduce significant risks if incorporated directly into predictive modelling. They exhibit (a) ***strong drifts*** (long-run systematic upward or downward movements) – which can lead to *spurious trend learning* (i.e., the model interpreting shared trends as causal rather than focusing on deviations that contain predictive signals); (b) ***seasonality*** – which may be misinterpreted as a predictive driver of returns instead of highlighting when values are unusually high or low relative to their seasonally typical baseline; and (c) ***wide scale differences*** across variables and tickers, creating risks of *non-comparability* due to inconsistent units, and *volatility of coefficients* (if features are on very different scales, coefficients will be on wildly different magnitudes, leading to unstable estimation and unfair penalization under regularization).

To mitigate these issues, a two-step transformation was applied:

1. De-meaning (centring): most features were de-meaned using trailing rolling or exponentially weighted averages, shifted by one day to prevent look-ahead. This removes slow-moving drifts and regular seasonal components, forcing the model to focus on deviations from typical values rather than absolute levels.

For these features, de-meaning subtracts a trailing mean, centring the series around zero:

$$\tilde{x_t} = x_t - \mu_{t,w}, \quad \mu_{t,w} = \text{rolling/EWM mean over window } w \ \ \text{(shifted by one day)}$$

Slow-moving macroeconomic and earnings-level variables were normalized using global statistics (full-sample mean and standard deviation) without a rolling de-mean step.

2. Standardization (z-scoring): After centring, most features were standardized by dividing their de-meaned values by trailing volatility (standard deviation), computed on a rolling or exponentially weighted basis and shifted by one day.

$$z_t = \frac{x_t - \mu_{t,w}}{\sigma_{t,w}}, \quad \mu_{t,w} = \text{rolling/EWM mean}, \ \sigma_{t,w} = \text{rolling/EWM std}, \ \text{(all shifted by one day)}$$

This expresses each observation in units of standard deviations from its recent norm, which makes features directly comparable across time and across tickers, regardless of their original scale (price, ratios, macro levels, etc.). For families not de-meaned (macroeconomic levels and earnings levels), standardization was applied using global statistics. The result is that all features are placed on a common, volatility-adjusted scale (typically between −2 and +2 for most observations).

Detailed Methodological Documentation: targets (future returns), event flags (e.g., *event*, earnings_date_*), calendar/timing variables, and binary indicators were auto excluded from transformation. All remaining numeric variables were assigned to families based on their economic and statistical characteristics and for each family, tailored normalization methods were applied:

Rolling windows (≈60 trading days): Rolling normalization was applied to Price levels, Distance-to-trend/Momentum, and Market return features. A 60-day (~3 months) trailing window provides a stable local baseline while remaining responsive to medium-term changes, ensuring deviations are measured relative to a recent but robust benchmark.

Exponentially weighted moving averages (EWM): EWM normalization was applied to Volume/Turnover and Volatility/Beta features. Unlike rolling windows, EWM assigns greater weight to more recent observations, enabling faster adaptation to regime shifts in trading activity or volatility. For volume features, a half-life of 25 days was used, meaning the effective weight of an observation decreases by 50% every ~25 trading days. This choice ensures that the scale reflects the most recent trading patterns while still retaining a meaningful history. For volatility/beta features, a half-life of 40 days was chosen, providing a more gradual adaptation. This reflects the fact that volatility and beta tend to evolve more slowly than trading volumes, requiring a longer memory to establish a stable baseline.

Global statistics (full sample mean and standard deviation). Global normalization was applied to the slow-moving, low-frequency variables of Macroeconomic and Earnings level features, using a constant mean and standard deviation computed once over the full sample. This approach ensures stability for slow-moving, low-frequency series where short windows would generate noisy or artificial fluctuations. For Macroeconomic indicators, global statistics provide a consistent baseline for variables that change slowly and are typically released monthly or quarterly (e.g., CPI, GDP, unemployment). For Earnings levels, global normalization avoids overstating volatility between sparse quarterly updates while still putting different firms and earnings metrics on comparable scales.

Finally, to prevent look-ahead bias, all rolling and EWM statistics were computed on a trailing-only basis and shifted by one trading day. This ensures that the standardized value at time t uses information available only up to $t-1$, preserving the integrity of out-of-sample evaluation.

Contribution to modelling: by transforming features into rolling or exponentially weighted de-meaned and standardized variants, the dataset was rendered more stationary and better aligned for modelling. This process produced cleaner signals, enhanced robustness, and ultimately supported more consistent and reliable predictive performance.

Normalization & De-Meaning Methods by Feature Family

De-Meaning & Normalization Method	Feature Family	Features Included	Outputs
Rolling mean/ std (60d)	Price levels	*open_*, high_*, low_*, close_*	Raw + *_dm, *_z
	Distance-to-trend / Momentum	*bollinger_position_*, macd_*, rsi_*, sma_ratio_20_*, ret_1d_*, ret_5d_*, perf_vs_sp_ret_5d_*, rolling_avg_10d_*, mom_20d_*, mom_63d_*, sharp_20d_*, daily_pct_change_*	Raw + *_dm, *_z
	Market returns	*sp_ret_1d, sp_ret_5d*	Raw + *_dm, *_z
	Sentiment	*sentiment_*, ai_score_*	Raw + *_dm, *_z
	Volume / Turnover	*volume_*, vol_z_60d_*	Raw + *_dm, *_z
EWM mean / std (half-life 25d)	Volatility / Beta	*atr_*, beta_sp_60d_*, bollinger_upper_*, bollinger_lower_*, rv_20d_*, semivol_20d_*, idio_vol_20d_*, dd_60d_*	Raw + *_dm, *_z
Global mean / std (full sample)	Macroeconomic levels	*cpi_raw, cpi_mom, cpi_yoy, exports_raw, exports_mom, exports_yoy, imports_raw, imports_mom, imports_yoy, trade_balance_raw, trade_balance_mom, trade_balance_yoy, unemp_rate, gdp, debt_service_ratio, interest_rate_daily*	Raw + *_gz
	Earnings levels	*reported_eps_*, estimated_eps_*, revenue_million_*, surprise_percent_*, earning_date_*, earnings_date_*, [illegible]*	Raw + *_gz