ChatGPT

Feature Specification for Multi-Horizon Stock Return Model

Below is a comprehensive list of predictive features derived from J-Quants API data, structured for direct implementation in a multi-horizon (1d, 5d, 10d, 20d) stock return regression pipeline. Each feature entry includes the feature name, type, data source (J-Quants API endpoint and fields), calculation method, data alignment strategy, justification (theoretical/empirical rationale with citations), and implementation priority.

Price & Technical Indicators

- **Previous Day Price Return** (1-day Momentum/Reversal)
- **Type**: Continuous numeric (time-series, stock-specific)
- **Source**: `/prices/daily_quotes` – uses prior day's close and current day's close price fields.
- **Calculation**: $r_{t,1d} = \frac{\text{Close}_t - \text{Close}_{t-1}}{\text{Close}_{t-1}}$ (one trading day percentage return). Use adjusted prices for corporate actions.
- **Alignment**: Computed at end of day t, used as a feature from day t+1 onward (T+1, i.e. next trading day's prediction input). No interpolation needed – naturally daily.
- **Justification**: Captures short-term return reversal vs. momentum. Short-horizon stock returns often exhibit a reversal effect: stocks with large negative returns one day tend to bounce back over the next few days, and vice versa [1]. This "short-term reversal" anomaly is well-documented (≈ +2%/month strategy profit) [1]. However, if the price move was driven by fundamental news (e.g. earnings), continuation (momentum) can occur instead [2]. Thus, previous-day return is a strong predictor, especially when combined with event indicators (to distinguish news-driven moves from noise).

- **Priority**: **High** – Critical for 1–5 day horizon performance (strong empirical predictive power).

- **5-Day Price Return** (1-week Momentum/Reversion)

- **Type**: Continuous numeric (time-series, stock-specific)
- **Source**: `/prices/daily_quotes` – uses close prices from the last 5 trading days.
- **Calculation**: $r_{t,5d} = \frac{\text{Close}_t - \text{Close}_{t-5}}{\text{Close}_{t-5}}$ (5-trading-day cumulative return). For incomplete weeks (around holidays), use last 5 available trading days.
- **Alignment**: Calculated at end of day t, using data up to t. Used from day t+1 onward. No special interpolation (rolling window naturally shifts daily).
- **Justification**: Captures short-term price trend or reversal over ~1 week. Short-horizon returns often mean-revert, but momentum can emerge if trends persist beyond a day. Empirical studies show weekly return reversals, especially in large stocks [3]. However, reversal strength declines when stocks have high turnover (liquid stocks with heavy trading can exhibit momentum instead) [4]. This feature, combined with volume/turnover features, helps distinguish pure reversal signals from sustained momentum.

- **Priority**: **High** – Important for 5-day ahead predictions; moderate effect on 1-day horizon but contributes to multi-day trend capture.

- **20-Day Price Return** (1-month Momentum)

- **Type**: Continuous numeric (time-series, stock-specific)
- **Source**: `/prices/daily_quotes` – uses close prices ~20 trading days apart.
- **Calculation**: $r_{t,20d} = \frac{\text{Close}_t - \text{Close}_{t-20}}{\text{Close}_{t-20}}$ (20-day cumulative return, ~1 calendar month). Use last 20 trading days.
- **Alignment**: Calculated at end of day t (uses up to day t close), available from day t+1. For newly listed stocks with <20 days data, use available history (or set null).
- **Justification**: Captures medium-term momentum or reversal. Momentum effects are strongest at intermediate horizons (3–12 months), but even 1-month trends can influence short-term returns. For example, stocks near their 1-month high may continue to rise (short-term momentum), while those with extreme 1-month moves might pull back (mean reversion). Including a 20-day trend helps the model differentiate sustained momentum from very short-lived moves. It also overlaps with the “52-week high” effect (stocks near annual highs tend to keep rising short-term) [4].

- **Priority**: **Medium** – Adds context for 10–20d horizon forecasts and cross-sectional momentum; less critical than 1d/5d returns for ultra-short term.

- **Price Distance from Moving Average (20-day z-score)**

- **Type**: Continuous numeric (time-series, stock-specific)
- **Source**: `/prices/daily_quotes` – uses close price; compute 20-day moving average and standard deviation.
- **Calculation**: $Z\text{-score}_t = \frac{\text{Close}_t - \text{MA}_{20,t}}{\sigma_{20,t}}$, where MA${20,t}$ is the 20-day moving average of close prices and σ is the 20-day std. deviation. (Can also compute a 5-day variant similarly.)
- **Alignment**: Computed at end of day t using last 20 days, used from t+1. Forward-fill on days without new price (non-trading days) is not needed if using trading days only.
- **Justification**: Measures how extreme the current price is relative to recent history (a technical “Bollinger band” style indicator). An extreme positive z-score means the stock is far above its recent average – often an overbought signal that may precede short-term mean reversion. An extreme negative z-score indicates oversold conditions that could lead to a bounce. By normalizing, this feature is comparable across stocks and time. It directly quantifies short-term **overreaction** or **breakout** conditions; such extremes often correct or revert in the ensuing days absent new information (consistent with short-term reversal patterns [2]).

- **Priority**: **Medium** – Useful for detecting overbought/oversold states (especially 1–10d horizons), but somewhat redundant with recent return features. Include for robustness.

- **Intraday vs. Overnight Returns** (Open/Close Decomposition)

- **Type**: Two continuous numeric features (time-series, stock-specific)
- **Source**: `/prices/daily_quotes` – uses daily open and close prices (and previous close for overnight).
- **Calculation**: Overnight return: $r_t^{ON} = \frac{\text{Open}_t - \text{Close}_{t-1}}{\text{Close}_{t-1}}$; Intraday return: $r_t^{ID} = \frac{\text{Close}_t - \text{Open}_t}{\text{Open}_t}$. These sum to the total daily return.
- **Alignment**: Both computed for day t after market close. Overnight return uses information from post-market and pre-market moves (e.g. global market influence), intraday return captures within-day momentum. Use as features on day t+1.
- **Justification**: Separating overnight vs. intraday performance provides insight into return drivers. **Overnight returns** often reflect news and global sentiment that emerged when local markets were closed, whereas **intraday returns** reflect price movement given that opening level. Patterns

in these can be predictive: e.g. a strong overnight gap up followed by intraday sell-off could indicate bullish news met with profit-taking, often leading to short-term weakness next day. Conversely, a flat overnight and strong intraday rally might signal genuine buying interest likely to carry into subsequent days. Research shows that overnight price dynamics differ from intraday (e.g. U.S. equities have historically earned most returns overnight, with day session often flat or reversing) [5] [6]. These features let the model capture such effects in the Japanese market context.

- **Priority**: **Medium** – Adds nuance for 1d predictions (if model runs after close, these are fully known). Useful for distinguishing gap-driven moves from intraday momentum.

- **High–Low Range Percent** (Daily Volatility)

- **Type**: Continuous numeric (time-series, stock-specific)
- **Source**: `/prices/daily_quotes` – uses daily high and low price fields (and optionally close).
- **Calculation**: $\text{RangePct}_t = \frac{\text{High}_t - \text{Low}_t}{\text{Close}_t}$, the day's intraday range as a percentage of closing price. Can also use true range (max of High-Low, |High - prevClose|, |Low - prevClose|) for gap-adjusted range.
- **Alignment**: Computed at end of day t, used from t+1. No further alignment needed.
- **Justification**: Intraday range is a proxy for **realized volatility** on that day. A large range indicates high uncertainty or information flow. Unusually high volatility often **mean-reverts** (volatility clustering: a volatility spike may predict elevated risk short-term but also potential reversal as extreme moves settle). Empirically, stocks with extreme high recent volatility tend to underperform going forward (the "volatility effect") [7], possibly because high-volatility episodes are associated with price drops and risk-aversion. Including range percent helps capture short-term volatility shocks that could foretell either continued turbulence or a corrective rebound.

- **Priority**: **Medium** – Important for short-term risk/volatility outlook (1–5d). Could be derived from other vol features; include for direct signal of volatility spikes.

- **5-Day Realized Volatility** (Short-term Volatility)

- **Type**: Continuous numeric (time-series, stock-specific)
- **Source**: `/prices/daily_quotes` – uses past daily returns (close-to-close) over last 5 trading days.
- **Calculation**: $\sigma_{t,5d} = \sqrt{\frac{1}{5}\sum_{i=t-4}^{t}(r_i)^2}$ (sample standard deviation of last 5 daily returns). Alternatively use Parkinson's high-low estimator for higher precision using intraday range. Normalize by annualizing if needed (not necessary for relative comparisons).
- **Alignment**: Computed at end of day t (using returns up to day t), feature from t+1. Use an as-of join: if a trading break occurs, still use last 5 trading days available.
- **Justification**: Measures very recent price volatility. High short-term volatility often persists (volatility clustering), indicating uncertain conditions likely to continue in the next few days. As a predictor, higher recent volatility can signal **lower next-period returns on average** (investors demand a premium or reduce positions) [7]. Also, sudden volatility bursts can indicate **overreaction** that corrects (coupling with reversal signals). Thus, 5-day vol provides a gauge of current risk and potential mean-reversion.

- **Priority**: **High** – Emphasized for short-term forecasts; directly addresses the user's focus on "short-term volatility" signals.

- **20-Day Realized Volatility** (Intermediate Volatility)

- **Type**: Continuous numeric (time-series, stock-specific)
- **Source**: `/prices/daily_quotes` – uses past 20 days of returns.
- **Calculation**: $\sigma_{t,20d} = \sqrt{\frac{1}{20}\sum_{i=t-19}^{t}(r_i)^2}$. (Standard deviation of daily returns over ~1 month.)
- **Alignment**: Computed at end of day t, used from t+1 onward.
- **Justification**: Captures a slightly longer volatility trend. This helps normalize short-term volatility against a baseline. For example, if 5-day vol is high but 20-day vol is also high, volatility is persistently elevated (perhaps due to a regime or sector-wide factor); if 5-day vol spikes far above 20-day, it's a more **transient shock**. By also including 20d vol, the model can infer a volatility **mean-reversion** factor or persistent risk factor. High 20-day volatility is often associated with riskier stocks which on average yield lower future returns (idiosyncratic volatility anomaly [7]).

- **Priority**: **Medium** – Useful for context and for features like volatility spike index; less directly predictive than 5-day vol for immediate moves.

- **Volatility Spike Index** (Volatility Jump vs. Normal)

- **Type**: Continuous numeric (time-series, stock-specific)
- **Source**: Derived from 5d and 20d vol features (no new data required).
- **Calculation**: $\text{VolSpike}_t = \frac{\sigma_{t,5d}}{\sigma_{t,20d}}$ (ratio of short-term vol to month-long vol). Optionally winsorize extreme values.
- **Alignment**: Computed end of day t from vols at t, available from t+1. If 20d window not fully available, handle via shorter window or set null initially.
- **Justification**: Highlights unusually high recent volatility relative to the stock's typical volatility. A VolSpike >> 1 means a **volatility shock** – often accompanying news or market stress – which can predict **short-term mean reversion** as markets calm or **persistent risk** if regime changed. Conversely, a low ratio (<1) means recent trading is calmer than normal, possibly preceding a volatility pickup. This feature encapsulates the concept of volatility mean-reversion (volatility tends to revert to a baseline). It is also a proxy for **recent uncertainty** that might be resolved in coming days.
- **Priority**: **Medium** – Captures a nuanced volatility signal (especially valuable for 5–20d forecasts). Can be derived, but listing explicitly for clarity.

Volume & Liquidity Indicators

- **Trading Volume (Relative to Average)**
- **Type**: Continuous numeric (time-series, stock-specific)
- **Source**: `/prices/daily_quotes` – volume field (number of shares traded). Also uses a rolling average (e.g. 20-day volume).
- **Calculation**: Compute normalized volume: e.g. $\text{VolRel}_t = \frac{\text{Volume}_t}{\text{AVG Volume}_{20d,t}}$ (current day volume divided by the stock's 20-day average volume). Alternatively, use a z-score: $(\text{Volume}_t - \overline{\text{Vol}}_{20})/\sigma(\text{Vol}_{20})$.
- **Alignment**: Known at end of day t (daily total volume). Use as feature from t+1. The rolling average updates daily (forward-fill not needed except skip non-trading days).
- **Justification**: Detects volume surges or droughts. A volume spike (VolRel >> 1) indicates unusual trading activity, often signifying informed trading or event-driven action. High volume confirms that a price move is backed by strong participation, which **tends to validate trends** (price moves on high volume are less likely to reverse immediately) [8] . Conversely, a large price change on low volume may be unreliable and more prone to reversal. Thus, volume relative to normal is a key **contextual feature**: e.g. a +5% return on 5× normal volume is a stronger bullish signal (likely momentum) than +5% on 0.2× volume (which might revert). Empirical studies find short-term

reversals are weaker (or turn into momentum) when turnover is high [4], implying volume moderates the reversal effect.

- **Priority**: **High** – Essential for interpreting price signals (1–5d horizons). Include both raw volume and a normalized variant for scale invariance.

- **Turnover Rate** (Volume as % of Float)

- **Type**: Continuous numeric (time-series, stock-specific)
- **Source**: `/prices/daily_quotes` – volume; `/listed/info` – shares outstanding or float (if provided).
- **Calculation**: $\text{Turnover}_t = \frac{\text{Volume}_t}{\text{Shares Out}_t}$ (daily trading volume divided by total shares outstanding, yielding a percentage of the company's shares traded). If float (free float shares) is available, use that for better accuracy; otherwise total shares gives an approximate turnover. A normalized version can be the turnover relative to its 20-day average.
- **Alignment**: Calculate at end of day t; use from t+1. Shares outstanding from `listed/info` is static or changes only on corporate actions (treat as constant in short term, as-of joined from latest info).
- **Justification**: Standardizes volume by company size. This distinguishes whether a high volume (in absolute terms) is truly significant for that stock. **Turnover** is linked to liquidity and investor attention. Extremely high turnover (a large % of shares changing hands) can indicate **panicked selling or speculative buying**, often short-lived. Low turnover suggests a stock is relatively illiquid or off radar, which can mean larger volatility per unit volume (captured by Amihud illiquidity below). Also, short-term reversal strategies have been found to yield lower profits in high-turnover stocks (as those moves are more likely information-driven) [4]. Including turnover helps the model adjust signals for liquidity differences – e.g. a small-cap stock with 1% of shares traded (high engagement) might react differently than a mega-cap with 0.01% traded.

- **Priority**: **High** – Important for cross-sectional comparability and for features like volume spikes and liquidity effects.

- **Amihud Illiquidity Indicator** (Price Impact of Volume)

- **Type**: Continuous numeric (time-series, stock-specific)
- **Source**: `/prices/daily_quotes` – daily volume (in shares) and value (if available; if not, can approximate value = volume * price); daily return.
- **Calculation**: $\text{Illiquidity}_t = \frac{|r_t|}{\text{Trading Value}_t}$, where trading value = volume * price (in ¥). Usually averaged over a period, e.g. use the past 20 days: $\frac{1}{20}\sum_{i=1}^{20}\frac{|r_{t-i}|}{\text{Value}_{t-i}}$. This gives the classic Amihud illiquidity: how much price moves per unit of trading value (higher = more illiquid). We can also compute a 5-day version for short term.
- **Alignment**: If using a 20-day average, compute at end of day t using data up to t, feature from t+1. For daily (instantaneous) illiquidity, compute each day's ratio as well.
- **Justification**: This feature measures **price impact** – how sensitive the stock's price is to trading volume. Highly illiquid stocks (high Amihud values) have outsized return moves on relatively little volume, indicating limited liquidity. Such stocks often exhibit **stronger reversals** and **idiosyncratic volatility**. Also, large institutional flows avoid illiquid names, so persistent order imbalances can move these prices dramatically (and then mean-revert). By including illiquidity, the model can assess whether a recent price jump is significant: in an illiquid stock, a big jump might be noise (likely to revert), whereas in a very liquid stock, a jump is more likely due to fundamental news. Illiquidity has been shown to predict cross-sectional returns (investors

demand higher returns for illiquid stocks) and to differentiate how stocks respond to volume and volatility shocks [9] [10].

- **Priority**: **Medium** – Improves model's understanding of liquidity conditions (robustness), but not as directly predictive as volume/turnover for short-term returns.

- **Volume-Price Trend Indicator** (Price & Volume Interaction)

- **Type**: Continuous numeric (time-series, stock-specific)
- **Source**: Derived from price and volume features above (no new data; uses recent returns and volume).
- **Calculation**: One example: **Volume-Weighted Momentum** – $\mathrm{VWMOM}_t = r_{t,5d} \times \frac{\mathrm{AVG\ Vol}_{t,5d}}{\mathrm{AVG\ Vol}_{t,20d}}$. This scales recent return by the volume spike over the same period. Alternatively, use a **Price-Volume Trend** measure: cumulative sum of $\Delta\mathrm{Price} \times \mathrm{Volume}$ (used in some technical analysis as a confirmation indicator). We can also include simpler interactions: e.g. Up/Down on high volume flags.
- **Alignment**: Calculated at end of day t, using volumes/returns up to t. Use from t+1.
- **Justification**: Combines price momentum with volume confirmation. A rising price accompanied by rising volume is a stronger bullish signal than a rise on declining volume [8]. This feature gives the model an integrated view: if 5-day return is positive and volume has been above average, the feature will be strongly positive (signaling momentum likely to continue); if 5-day return is positive but volume was weak, the indicator will be weaker or negative (implying a suspect rally likely to revert). Similarly for downtrends: price drop on high volume suggests genuine selling pressure (likely to persist or need significant news to reverse), whereas a price drop on low volume might reverse quickly. This aligns with the principle that **volume confirms price movements** [11].

- **Priority**: **Medium** – Captures a nuanced predictive signal (trend confirmation vs. divergence). It's effectively an engineered feature from others; include if complexity is manageable for the pipeline.

- **Bid-Ask Spread (Proxy)** – if data available

- **Type**: Continuous numeric (time-series, stock-specific)
- **Source**: If intraday or quote data is available (not explicitly listed in J-Quants basic endpoints). Otherwise, a proxy can be estimated from high/low or close/open (higher volatility in prices relative to trading range can imply wider spreads).
- **Calculation**: E.g. use high vs. low vs. close: $\mathrm{PctSpread}_t \approx \frac{\mathrm{High}_t - \mathrm{Low}_t}{\mathrm{Close}_t} - \mathrm{RangePct}_t$ to isolate gap beyond pure volatility. If order book data were available, use quoted bid-ask.
- **Alignment**: Measured daily; if using OHLC proxy, computed at end of day t.
- **Justification**: The bid-ask spread reflects **transaction cost and liquidity**. A widening spread indicates lower liquidity and higher uncertainty. In the short term, very large spreads might predict higher volatility and potential price jumps (as a liquidity event). However, due to data limitations, this is a low-priority proxy feature.
- **Priority**: **Low** – Include only if easily derived; otherwise skip, as other liquidity features cover similar ground.

Sentiment & Positioning Indicators

- **Sector Short-Selling Ratio**
- **Type**: Continuous numeric (time-series, sector-level; mapped to each stock by sector)

- **Source**: `/markets/short_selling` (short selling ratio by sector). This provides the percentage of trading value attributable to short-selling for each sector (likely daily). Also require each stock's sector code from `/listed/info` to map the appropriate ratio.
- **Calculation**: For a stock with sector S, feature = short_sell_ratio<S>_% on day t. If the API provides short-selling ratio as e.g. 20% for sector, use 0.20 as the value. Optionally, one can normalize: subtract the average or compute a z-score relative to that sector's history to highlight unusually high shorting days.
- **Alignment**: The short-selling ratio is presumably published daily after market close. Use a T+1 alignment (feature on day t+1 reflecting day t's ratio) if there's a reporting delay. If provided in real-time daily, can use same day. Typically assume it's end-of-day statistic, so include from next trading day. No interpolation – value stays until updated next trading day.
- **Justification**: This gauges **bearish sentiment and positioning** in the stock's industry. A high sector short-selling ratio means a large fraction of trades in that sector were short sales, indicating widespread bearish bets or hedging in that sector. Empirical evidence shows that stocks with high short interest (i.e. heavy short selling) tend to underperform subsequently [12], as short sellers often possess negative information. By extension, if an entire sector faces heavy shorting, it could signal negative outlook or overvaluation for that sector, pressuring constituent stocks' short-term returns. Alternatively, extremely high shorting can set the stage for short-covering rallies if sentiment shifts. In either case, the ratio provides a predictive signal: rising shorting activity often precedes price declines [13] (information content of shorts), whereas an abnormally high short ratio reaching extremes could mark capitulation (contrarian buy signal). Including this feature helps capture **broad positioning trends** affecting the stock beyond its own fundamentals.

- **Priority**: **High** – Valuable sentiment indicator, especially for 5–20d predictions or during market stress. High priority due to strong theoretical impact of short positioning.

- **Stock-Specific Short Interest** (Short Positions Outstanding)

- **Type**: Continuous numeric (time-series, stock-specific)
- **Source**: `/markets/daily_margin_interest` – daily margin trading balances, which include **margin short sell balance** (shares sold short on margin) for each stock; or `/markets/short_selling_positions` if available (reported short positions $\geq$ certain threshold). Also need shares outstanding from `/listed/info` for normalization.
- **Calculation**: Compute **short interest ratio**: $\text{ShortInt}_t = \frac{\text{ShortShares}_t}{\text{Shares Out}}$ (fraction of float that is short via margin). Also useful is **days-to-cover**: $\frac{\text{ShortShares}_t}{\text{Avg Daily Volume}_{20d}}$ (number of days of normal volume to cover all shorts). We can include either or both. Changes in short interest (week-over-week percent change) can be another variant.
- **Alignment**: Margin short balance is typically reported end-of-day (possibly with T+1 publication delay). Use an as-of join: i.e. use the latest available short interest as of the prediction date. If daily data is promptly updated, use yesterday's figure for today's prediction (T+1). If only weekly, then apply the last known value to all days until a new update (a step function over time).
- **Justification**: **Short interest** is a direct measure of bearish bets on the stock. High short interest has strong predictive power: stocks with unusually high short interest often experience **negative abnormal returns** subsequently, as short sellers' information on overvaluation or bad news tends to be proven correct [12]. Additionally, short interest increases ahead of negative fundamentals (shorts anticipate bad earnings surprises) [14]. Thus, a high or rising ShortInt is generally a bearish signal (priority to fundamental-based down moves). However, extremely high short interest can also imply a potential **short squeeze** scenario if positive news forces shorts to cover, causing sharp short-term rallies. Our model can account for this by also considering momentum and event features (e.g. if short interest is high and an unexpected good news arrives, the upside can be

dramatic). Overall, including short interest helps capture **informational inefficiencies and sentiment** at the stock level.

- **Priority**: **High** – Key positioning indicator for robustness and for 10–20d forecasts (short interest data may move slowly, but its predictive effect is significant).

- **Margin Trading Balance** (Long vs Short Demand)

- **Type**: Two continuous numeric features (time-series, stock-specific)
- **Source**: `/markets/daily_margin_interest` – provides **margin buy balance** (outstanding long positions on margin credit) and **margin sell balance** (outstanding short positions) for each stock. If only weekly data available from `/markets/weekly_margin_interest`, use that with interpolation.
- **Calculation**: We derive: **Margin Long Ratio** = MarginBuyBalance / Market Cap (or / Shares Out) — proportion of the stock's float bought on credit; **Margin Short Ratio** = MarginSellBalance / Market Cap — proportion sold short via margin. Also consider **Margin Long-Short difference** or **ratio** = MarginBuy / MarginSell as an indicator of net retail positioning. Changes in these balances week-to-week can also be features (e.g. surge in margin buying this week).
- **Alignment**: The balance data is typically reported for end-of-day (with possible next-day release). Use as-of join: the latest known balances as of each date (if daily, use yesterday's for today; if weekly, use the last week's value for all days until updated). Interpolate daily values if needed (e.g. assume constant within the week).
- **Justification**: Margin balances reflect **retail investor sentiment and leverage**. High margin buy balance means many investors have taken leveraged long positions, which can signal optimism or speculation. In the short term, a rapid increase in margin buying often happens in rising markets but can precede **price corrections** if those positions become pressured (e.g. margin calls can exacerbate down moves). Conversely, high margin short balance (similarly to short interest) indicates significant bearish bets. The **ratio** of margin longs to shorts can be seen as a **bull/bear sentiment gauge** among retail leveraged traders. Extreme values may be contrarian indicators: e.g. if margin longs are at record highs, the buying capacity might be exhausted (bearish); if margin shorts are at highs, any positive catalyst can trigger short covering (bullish potential). Including these features makes the model aware of **positioning risk** – stocks heavily bought on margin can face forced selling, while heavily shorted stocks can rally on short covering. These are especially relevant for short-term volatility and reversals around market turning points.

- **Priority**: **Medium** – Useful for robustness (particularly in speculative environments). Medium priority due to slower-moving nature, but can materially improve 10–20d forecast accuracy in volatile periods.

- **Net Foreign Buying (Market-Level)**

- **Type**: Continuous numeric (time-series, market-wide; same value applies to all stocks for that day, or perhaps segmented by market section)
- **Source**: `/markets/trades_spec` – investment division info. If this dataset provides daily net purchase amount by investor type (e.g. foreign investors) for the market or exchange section, use the foreign net flow. (If only weekly data is available, use weekly values applied on that week's end or spread across days.)
- **Calculation**: Use the net value (in JPY) of foreign investor stock purchases minus sales on day t (for TSE overall or relevant section). Possibly normalize by total market volume to get a percentage flow, or by market index level. For implementation, one could create an indicator like "foreign flow in billions of JPY". If multiple categories (domestic institutions, individuals, etc.) are available, features can be created for each, though foreign is often most impactful.

- **Alignment**: If daily, use the value from day t as a feature on day t+1 (assuming data confirmed by next day). If only weekly, apply the latest weekly net flow as of the current date (e.g. use the last reported week's flow until a new report). Ensure no look-ahead bias (e.g. if weekly released on Wednesday for prior week, only use it from that day onward).
- **Justification**: **Investor flows** influence short-term price dynamics. In Japan's market, foreign investors' net buying or selling is a well-known driver of index movement and liquidity [15]. A large positive foreign inflow indicates broad optimism and often correlates with near-term price strength in many stocks (a rising tide lifting boats), whereas heavy foreign outflows can presage short-term market weakness. This feature acts as a proxy for **market sentiment and fund flows**. It's effectively a market-level exogenous factor that can help explain why all stocks might have an upward or downward bias on certain days regardless of individual fundamentals. By including it, the model can adjust predictions if, say, there's strong market-wide buying interest that could buoy even weaker stocks in the short run. (If trades_spec has per-stock breakdown by investor, even better – then use stock-specific foreign net volume as a percent of that stock's volume, which would be a very direct flow signal. But absent that, the aggregate still provides signal.)

- **Priority**: **Medium** – Enhances robustness and captures macro sentiment for 1–5 day forecasts. Not stock-specific, but important in broad market moves. Include if data available.

- **Investor Type Rotation Indicators** – optional, if trades_spec granular

- **Type**: Continuous numeric or categorical (time-series, stock or market segment)
- **Source**: `/markets/trades_spec` – if it gives breakdown by investor type (foreign, institutional, retail, proprietary) at the stock or segment level.
- **Calculation**: e.g. **Retail vs Institutional Flow Balance** – difference between retail net buy and institutional net buy (could be market-wide or per stock if available). Or indicator flags like "Foreign heavy buying today" (if foreign net flow ranks high percentile). These could be simplified to dummy signals or continuous values of net flows.
- **Alignment**: Same as above – daily or weekly as reported.
- **Justification**: Different investor classes often exhibit distinct behaviors. For instance, retail investors might chase momentum or mean-reversion differently than institutions. If retail is aggressively buying while institutions are selling, it might signal **unsustainable rallies** (smart money selling into retail demand). Conversely, strong institutional accumulation could predict sustained upward movement. Such features refine the sentiment aspect by identifying **who** is driving the price. They are event-like (when a big shift occurs) and time-series.

- **Priority**: **Low** (if per-stock data isn't available) or **Medium** (if available and known to be impactful). It adds complexity and might be omitted if data is only aggregate.

- **Day-of-Week & Seasonality** (Calendar Effects)

- **Type**: Categorical (one-hot encoded for Monday, Tuesday, …) or binary flags; Continuous (e.g. day index for cyclical pattern)
- **Source**: `/markets/trading_calendar` – provides trading days and holidays. (For day-of-week, no external data needed beyond date.)
- **Calculation**: For each day, generate features: e.g. Monday=1 if day is Monday else 0 (similarly for Tue-Fri). Alternatively, a single categorical feature "DayOfWeek". Also consider end-of-month or pre-holiday flags (if the day before a long weekend).
- **Alignment**: These are known in advance (calendar-based). No alignment issues. Apply to each date as appropriate.

- **Justification**: There are mild but persistent calendar anomalies in stock returns (though weaker in recent years). Historically, Mondays often show lower or negative returns (possibly due to accumulation of bad news over weekend), while Fridays and pre-holidays can have positive bias (optimism heading into time off). Including day-of-week can capture any such systematic effect in the data, improving predictability of the mean return. Also, turning-of-month effects (first few days of month tend to be strong due to fund inflows) could be captured if needed. While these effects are small, they add a bit of predictive power and help the model not to erroneously attribute a pattern to other features.
- **Priority**: **Low** – Minor contribution, but easy to implement. Can be included for completeness (robustness and slight edge in 1-day forecasts).

Derivatives & Market Context Indicators

- **Market Index Return** (TOPIX or Nikkei)
- **Type**: Continuous numeric (time-series, market-level)
- **Source**: `/indices/topix` – TOPIX index OHLC (or `/indices` for other indices like Nikkei 225 if needed). Use index closing values.
- **Calculation**: Daily market return: $r_t^{mkt} = \frac{\text{IndexClose}_t - \text{IndexClose}_{t-1}}{\text{IndexClose}_{t-1}}$. Optionally include 5-day index return as well.
- **Alignment**: Known at end of day t. Use from t+1. (For predicting next-day stock returns, one could also include same-day index return as a contemporaneous factor if modeling close-to-close relationships jointly, but typically we use prior index moves as features.)
- **Justification**: Captures **market-wide movements and beta exposure**. A large portion of a stock's daily move is often driven by market (or sector) factors. Including the recent index return helps the model account for broad market momentum or reversal that could influence all stocks. For example, if the market had a strong rally in the last few days, some mean-reversion at the market level might occur (affecting most stocks' direction). Conversely, a market in steady uptrend can lift stocks (market momentum). By providing the index trend, we allow the model to separate market effects from stock-specific alpha signals. It effectively normalizes stock momentum vs. the market: e.g. a stock up +2% when the market is up +3% is actually relatively weak. We can enhance this by also adding a **stock's excess return vs market** if desired (stock return minus index return).

- **Priority**: **High** – Crucial context feature for all horizons (improves robustness and prevents the model from needing to learn market moves via many individual stocks' behavior).

- **Sector Index Return**

- **Type**: Continuous numeric (time-series, sector-level mapped to stock)
- **Source**: `/indices` – If sector indices (e.g. TOPIX 17 industries or 33 industries indices) are available via the indices endpoint (by index code). Alternatively, compute average return of all stocks in the sector from daily_quotes (though official index is preferred).
- **Calculation**: Similar to market return, but for the stock's specific sector index. $r_t^{sect}(S)$ = return of sector S on day t. Can include 5-day sector return as well.
- **Alignment**: End-of-day t computation, available from t+1 for features.
- **Justification**: Provides **sector-specific momentum/reversal context**. Some price movements are sector-driven (oil price up => oil sector rally, etc.). Including sector return helps explain stock moves due to industry trends. It also allows for features like **relative strength**: e.g. the model can combine stock's return and sector return to identify if a stock is outperforming or underperforming its peers. Empirically, stocks that strongly outperform their sector in the short run might revert (if it was an idiosyncratic overshoot), whereas stocks that lag a strong sector rally might catch up. Additionally, sector momentum itself can persist or reverse: having the sector

trend as a feature lets the model learn these patterns (some sectors exhibit momentum, others mean-reversion in certain regimes).

- **Priority**: **Medium** – Improves cross-sectional predictions and captures industry effects (especially valuable for 5–20d). Include if data readily available.

- **Futures Basis** (Index Futures vs Spot)

- **Type**: Continuous numeric (time-series, market-level)
- **Source**: `/derivatives/futures` – index futures quotes (e.g. Nikkei 225 futures front-month); `/indices` for index level.
- **Calculation**: $\text{Basis}_t = \frac{\text{FuturesPrice}_t - \text{IndexPrice}_t}{\text{IndexPrice}_t}$, using near-month futures closing price and index closing level on the same day. If multiple maturities, use nearest expiration for short-term signal. Can also compute change in basis from yesterday.
- **Alignment**: Futures price at end of day t (which might reflect trading up to early evening if JPX night session considered). Use as feature from t+1. If using night session prices (beyond cash market close), be careful to not include information that wouldn't be known by model at cutoff (unless model runs post-night session). Ideally, use basis as of the cash market close (futures price synchronized with 3pm close).
- **Justification**: The futures basis reflects **expectations and hedging pressure**. A rich (high positive) basis could indicate optimistic outlook or positive carry (low risk-free rate), whereas a **negative basis (backwardation)** often occurs in times of market stress or heavy shorting of futures (paying a premium to hedge). Unusual basis levels can predict short-term corrections: for instance, a significantly negative basis might presage a near-term rebound (as it signals extreme bearish sentiment and demand for hedges), while an excessively positive basis might predict pullback (market over-eagerness). This feature effectively captures the **derivatives market sentiment** relative to spot. It's a subtle signal but can be powerful around turning points.

- **Priority**: **Low** – Advanced feature; include if implementation is easy and data timing is handled. Its predictive value is episodic (most useful in extremes).

- **Index Implied Volatility (Nikkei 225 Option IV)**

- **Type**: Continuous numeric (time-series, market-level)
- **Source**: `/option/index_option` – Nikkei 225 option quotes; possibly provides implied vol or enough data to compute. If not directly given, use option prices to derive a 30-day ATM implied volatility (analogous to VIX for Nikkei). Alternatively, if not feasible, skip or use a proxy like historical vol of index as substitute.
- **Calculation**: Calculate implied volatility from near-the-money options: e.g. take the at-the-money call and put prices for the front month, use the Black-Scholes formula with known strike, underlying (Nikkei index), time to expiration, and risk-free rate to solve for implied vol. Or, if "theoretical price" and "settlement price" are given, the API might directly provide an implied volatility metric. Use that as the feature (in % volatility).
- **Alignment**: Implied vol index for end of day t (the option's closing implied vol or official closing VIX-like measure). Use from t+1. During non-trading days, carry forward last value or compute from any off-session if available.
- **Justification**: Implied volatility is the market's expectation of future volatility and a proxy for **investor fear or complacency**. A spike in implied volatility (like a VIX spike) usually coincides with market drops and **panic**. High implied vol levels often predict **mean-reverting positive returns** (markets tend to rebound after peak fear subsides) – in other words, extremely high volatility is often followed by market gains as fear recedes [16] . Conversely, very low implied vol may precede

weak returns or sudden spikes (complacency before the storm). By including an implied volatility measure, the model can incorporate this **risk sentiment** factor. It helps adjust expectations for stocks: in high-vol regimes, even fundamentally strong stocks can drop, but also the rebound potential is larger. For short-term forecasting, changes in implied vol can signal shifts in regime that affect all stocks' volatility and direction (e.g. rising vol often pairs with momentum crashes or reversals).

- **Priority**: **Medium** – A macro-level feature focusing on volatility forecasting. Medium priority because it's not stock-specific, but it adds predictive power during volatile periods (important for 1d-5d risk-sensitive forecasts).

- **Put/Call Open Interest Ratio** – if data available

- **Type**: Continuous numeric (time-series, market-level or index-level)
- **Source**: `/derivatives/options` – if it includes open interest or volume for calls and puts on the index. If not directly available, this feature may be skipped.
- **Calculation**: $\mathrm{PCR}_t = \frac{\text{Total Put OI}}{\text{Total Call OI}}$ for nearest expiry, or volume-based ratio of put vs call trading. Smooth it over a few days to reduce noise.
- **Alignment**: Compute at end of day t, use from t+1.
- **Justification**: The put/call ratio is a classic **sentiment indicator**. A high PCR (more puts than calls) suggests bearish sentiment or demand for protection, which at extremes can be contrarian bullish (everyone is hedged/pessimistic) or indicative of informed negative outlook. A low PCR (call-dominated) indicates bullish speculation, which at extreme can warn of overbought conditions. Including PCR can improve short-term market timing within the model, particularly for 5–20 day horizons where sentiment mean-reversion plays out.

- **Priority**: **Low** – Include if readily accessible; otherwise skip, as implied vol and short interest cover similar sentiment ground.

- **Futures Trading Volume & Open Interest**

- **Type**: Continuous numeric (time-series, market-level or contract-level)
- **Source**: `/derivatives/futures` – may include volume and OI for futures contracts.
- **Calculation**: Use change in open interest (ΔOI) on the index futures as a feature, and high futures volume days as a feature (flag or continuous). For example, a large increase in OI on a price move indicates new positions driving the move (can foreshadow trend continuation), whereas flat OI on a price move might indicate short covering or long liquidation (existing positions changing, more likely to mean-revert).
- **Alignment**: End of day t futures data, use from t+1.
- **Justification**: **Derivatives flow**: OI rising means money entering the market (new longs/shorts), often reinforcing trends; falling OI means money leaving (position closing), often happening during reversals. Unusual futures volume or OI changes can thus predict **short-term trend persistence or reversal**. For instance, if price rose and OI jumped, new longs likely entered – could sustain the rally short-term. If price rose but OI fell (short covering rally), the move might fade once shorts are done. This nuance from futures market activity complements the cash market signals.
- **Priority**: **Low** – More advanced and subtle; use if data is available and model can benefit from futures market context.

Fundamental & Event Indicators

- **Earnings Announcement Proximity**
- **Type**: Categorical/binary (event indicator) + continuous (countdown)
- **Source**: `/fins/announcement` – earnings announcement schedule (upcoming announcement dates for each stock); `/calendar/markets` for date reference.
- **Calculation**: Create features like: **Upcoming Earnings in 5 Days (Y/N)** – a binary flag that is 1 if the company will announce earnings within the next week (t to t+5 days); **Days to Next Earnings** – an integer or continuous countdown (e.g. 0 if today is announcement day, positive if days until next announcement, negative if days since last announcement). Also, **Recent Earnings Announcement** flag – 1 if an earnings release happened in the past N days (e.g. last 5 days).
- **Alignment**: The schedule is known in advance. For upcoming announcements, mark the feature in the days leading up to the event (and on the event day). For past announcements, the feature becomes active on the day after the announcement (since the actual results are typically released after market close, treat the next trading day as the first day "post-earnings"). Use as-of join to ensure we don't use the knowledge of an announcement before it's officially scheduled.
- **Justification**: Earnings releases are major **events** that drive volatility and returns. In the days before earnings, stocks often experience **elevated volatility and speculative positioning** (some investors buying ahead of expected good news or selling ahead of bad news). This can lead to mean-reverting moves if anticipation was wrong, or continuation if expectations are confirmed. A flag for upcoming earnings can help the model increase predicted volatility or be cautious in directional prediction. More importantly, after earnings, there is the well-known **post-earnings announcement drift (PEAD)**: stocks with positive surprises tend to continue rising for days or weeks after the announcement, while those with negative surprises continue falling [5] [6]. By including a recent announcement flag (and possibly the surprise magnitude as a feature – see next item), the model can capture earnings momentum. For example, if an earnings announcement (past 2 days) was very positive (price jumped), a PEAD-aware model will predict further outperformance short-term [5]. Without these event features, the model might incorrectly label the post-earnings jump as an outlier or pure mean-reversion candidate.

- **Priority**: **High** – Critical for robustness around earnings dates and for capturing earnings momentum in 5-20d horizons.

- **Earnings Surprise / Revision Metrics**

- **Type**: Continuous numeric (event-derived, stock-specific)
- **Source**: `/fins/statements` – financials (to get actual earnings or profit); external consensus data if available (not provided by J-Quants, but we may approximate surprise by price reaction or YoY growth). If no analyst estimate, use **year-over-year EPS growth** or **earnings vs guidance** if available, or **price change on earnings day** as a proxy for surprise.
- **Calculation**: Example features: **Earnings Day Return** = stock's return on the day after earnings announcement (market reaction); this serves as a proxy for surprise (positive if earnings beat expectations, negative if miss). Another: **YoY Quarterly EPS Growth** = (latest EPS – EPS from same quarter last year) / |last year's EPS|. Also, **Analyst Revision Count** if data existed (not in provided API, likely skip or use price-based proxy). These would be updated only when new earnings are announced. The feature can be forward-filled until next announcement (indicating the last earnings surprise magnitude). Normalize surprises (e.g. divide by price or use standardized unexpected earnings if possible).
- **Alignment**: Calculated after an earnings announcement is released. E.g., on the morning after earnings release, compute the surprise metrics and include them as features from that day forward. Use as-of join to keep the same surprise value until a new earnings report replaces it.

Ensure not to use any info that wasn't public at the time (no look-ahead of actual results pre-announcement).

- **Justification**: Quantifies **earnings momentum**. The magnitude of an earnings surprise is directly linked to PEAD: larger positive surprises yield stronger subsequent drift [6]. Including a numerical surprise feature (even if proxied by the earnings-day return) allows the model to scale its expectations: e.g. a mildly positive surprise might lead to a modest drift, while a huge positive surprise (stock up 15% on earnings) often continues to outperform significantly in the following weeks [5] [6]. Similarly, big negative surprises predict continued underperformance as investors underreact initially. Even absent consensus data, the earnings-day return is a reliable indicator of the news' positivity/negativity. Additionally, a feature like YoY growth or earnings revision (if available) captures fundamental momentum – companies with accelerating earnings tend to see upward price momentum, independent of immediate surprise. This ties into the concept that short-term returns are driven not just by technicals but by changes in fundamental expectations.

- **Priority**: **High** – Very important around earnings cycles and for 10–20d predictions (where drift plays out). High priority to accurately model event-driven moves.

- **Analyst/Forecast Signals** – if available via statements or external

- **Type**: Continuous numeric (event-derived, stock-specific)
- **Source**: `/fins/statements` and `/fins/fs_details` – sometimes includes company forecasts or revisions. If the API provides company's projected earnings or analyst forecasts (not sure if included), then features can be built for **earning forecast changes**. If not, this can be skipped.
- **Calculation**: e.g. **Guidance Surprise** = Company's new guidance vs prior guidance or vs analyst consensus (if one can derive it). Or simply **Change in Operating Profit forecast** quarter-over-quarter if the company issues forecasts. Mark significant positive or negative changes.
- **Alignment**: When new forecasts are announced (often at earnings), compute immediately and use from next day onward.
- **Justification**: Captures **forward-looking earnings momentum**. Oftentimes, stock moves are driven more by revised expectations than past results. If a company raises its earnings guidance, that is a strong positive signal that can drive continued stock strength (similar effect as a positive earnings surprise). By quantifying forecast changes, the model can better predict sustained moves beyond the initial announcement reaction.

- **Priority**: **Medium** – If data is available. Otherwise skip or consider low priority relative to other features.

- **Dividend Events** (Ex-dividend Indicator & Yield)

- **Type**: Binary (event flag) + continuous (yield)
- **Source**: `/dividends/dividend` – dividend info (ex-div dates, amounts); `/prices/daily_quotes` (price for yield calc).
- **Calculation**: **Ex-Dividend Day flag** = 1 if today is ex-dividend date for the stock (0 otherwise). **Dividend Yield** = annual dividends per share / price (perhaps not needed for short-term, but include for completeness or to adjust returns). If predicting raw returns, note that on ex-div day the price will drop by roughly the dividend amount – having a flag informs the model of this mechanical effect. One could also adjust the target return to be total return including dividend to avoid a need for the model to learn it. But if not, then this flag is essential.
- **Alignment**: The ex-div date is known beforehand from the calendar. Set the flag on the specific trading day that is ex-div. Dividend yield can be updated when dividend info is announced (usually well in advance or periodically).

- **Justification**: On ex-dividend dates, stock prices typically fall by the dividend amount. This is not a predictive signal per se, but a **data adjustment** issue – the model might otherwise interpret the ex-div price drop as a negative signal. Including an ex-div flag (and/or using dividend-adjusted returns for the target) ensures the model doesn't wrongly attribute that drop to fundamentals or sentiment. Additionally, dividend yield as a feature (though more of a value factor) might have minor short-term effects: high-yield stocks could be more defensive (lower volatility), and around dividend times, some investors buy for yield (supporting the stock before ex-div and then selling after). These are secondary, but the main reason is robustness and correct modeling of those event days.

- **Priority**: **Medium** – The ex-div flag is **high priority for correctness** (to avoid spurious model errors around those days). The dividend yield factor is low priority for predictive power in 1-20 day horizon (more relevant to longer-term), but can be included for completeness or multi-horizon consistency.

- **Corporate Actions & Other Events**

- **Type**: Binary flags (events)
- **Source**: `/listed/info` might list stock splits, mergers; `/fins/dividend` for stock splits or `/calendar/markets` for special trading days. Possibly the API doesn't explicitly list all corporate actions, but major ones like splits can be inferred from price adjustments.
- **Calculation**: Flags for events like **Stock Split day**, **Index inclusion/exclusion** (if known), or **Trading halts** etc. These are rare but can cause abnormal returns around those dates.
- **Alignment**: Known in advance (announcements) or as they occur.
- **Justification**: To capture any out-of-ordinary events that could affect short-term returns or data consistency. For example, stock splits often see price increases before/after due to liquidity perceptions, or index inclusion can cause a temporary price pop due to fund buying. While these are not frequent, including them ensures the model accounts for such event-driven moves when they happen.
- **Priority**: **Low** – These events are infrequent; handle them if convenient, but not critical to daily predictions.

Each feature above should be engineered into the model dataset with careful attention to **alignment** (to avoid look-ahead bias) and appropriate **normalization**. For many continuous features, it's advisable to include a normalized variant: for example, a rolling z-score (time-series normalization) or a percentile rank relative to the stock's history. This was noted for volume and price extremes. Cross-sectional normalization (ranking stocks by a feature each day) could also be beneficial for certain relative features (though our focus is primarily time-series signals per stock).

Implementation Note: In practice, many of these features will be derived from the raw API fields within the data pipeline (e.g. using pandas or SQL transformations). They should be added as new columns in the feature matrix for each stock-date. The model can then be trained on these features to predict 1d, 5d, 10d, 20d forward returns. High-priority features should be developed and validated first (they offer the most bang-for-buck in predictive accuracy), while medium/low priority ones can be added incrementally to test incremental improvements. All features should be checked for leakage and correlation.

By leveraging the **full breadth of J-Quants data**, including price/volume history, sentiment and positioning metrics, derivatives market cues, and fundamental events, this feature set is designed to maximize short-term predictive power (especially for 1–5 day returns) while remaining robust and reproducible for longer horizons up to 20 days. Each feature is aligned to ensure it uses only information available **as of the prediction time**, enabling realistic backtests and live trading usage.

---

[1] [2] [9] [10] A Closer Look at the Short-Term Return Reversal
https://www3.nd.edu/~zda/Reversal.pdf

[3] Institutional Investors and Short-Term Return Reversals
https://journals.sagepub.com/doi/10.1177/0148558X251347864

[4] Short-term momentum and reversals, turnover, and a stock's price-to ...
https://www.sciencedirect.com/science/article/abs/pii/S0927539824000902

[5] [6] Post-Earnings Announcement Effect - Quantpedia
https://quantpedia.com/strategies/post-earnings-announcement-effect

[7] [12] [13] [14] Short Interest Effect - Long-Short Version - Quantpedia
https://quantpedia.com/strategies/short-interest-effect-long-short-version

[8] [11] Volume-Based Trading Methods: A Complete Guide to Success
https://tradewiththepros.com/volume-based-trading-methods/

[15] Japanese stocks log 1.16 trillion yen FPI inflows, biggest in over 4 ...
https://m.economictimes.com/markets/stocks/news/japanese-stocks-log-biggest-weekly-foreign-inflows-in-over-four-months/articleshow/123423329.cms

[16] [PDF] Volatility Timing, Sentiment, and the Short-term Profitability of VIX ...
http://wp.lancs.ac.uk/fofi2020/files/2020/04/FoFI-2020-067-Wenjie-Ding.pdf