# Feature Engineering Deep Dive

Understanding the Top 10 Predictive Features

**Technical Reference Guide**

December 2024

# Introduction

This report provides an in-depth analysis of the 10 most important features used in the confidence-filtered multi-class classification trading strategy. Each feature is examined across multiple dimensions: mathematical definition, domain classification, intuitive rationale, and practical interpretation.

These features collectively capture 91.6% of the model's predictive power, spanning three primary domains:

- **Microstructure Analysis:** 61% of total importance
- **Time-Based Patterns:** 18.6% of total importance
- **Momentum & Volume:** 14.5% of total importance

# Feature 1: hl_ratio (26.9% importance)

## Definition

The high-low ratio measures the magnitude of price range during a candle period, normalized by the closing price.

## Domain Classification

- **Primary Domain:** Market Microstructure
- **Sub-Domain:** Volatility Measurement
- **Category:** Range-based volatility estimator

## Mathematical Formula

```
hl_ratio = (High - Low) / Close
```

**Where:**

- High = Highest price during the 15-minute period
- Low = Lowest price during the 15-minute period
- Close = Final price at period end

## Technical Intuition

The hl_ratio quantifies intraperiod volatility by measuring how far prices traveled from their extreme low to extreme high, expressed as a percentage of the closing price. This range-based estimator captures realized volatility without requiring multiple data points, making it computationally efficient and conceptually clean.

High hl_ratio values indicate significant price discovery activity, suggesting either:

1. Strong directional momentum with profit-taking retracements
2. Uncertainty and indecision with multiple price tests
3. Elevated market attention and liquidity provision

## Layman's Explanation

Imagine you're at an auction. The hl_ratio tells you how much the bidding swung between the highest and lowest bids during a 15-minute window, compared to what someone actually paid at the end.

If a stock closes at ₹100, but during those 15 minutes it hit a high of ₹103 and a low of ₹97, the hl_ratio is (103-97)/100 = 6%. This means the price swung 6% around before settling.

A hl_ratio of 1% means calm, controlled trading. A hl_ratio of 5% means volatile, chaotic price action. The model uses this to detect when the market is uncertain (high ratio) versus decisive (low ratio).

## What This Feature Represents

- **Market State:** High values = volatility regime, Low values = consolidation regime
- **Information Flow:** Large ranges suggest new information arriving, narrow ranges suggest equilibrium

- **Predictive Signal:** Volatility clustering means current high hl_ratio predicts future large moves

  *Example: Stock opens at ₹100, touches ₹105 high, drops to ₹98 low, closes at ₹102. hl_ratio = (105-98)/102 = 6.86%. This high ratio signals increased volatility, making the model more cautious about predicting direction.*

## Why It's Important (#1 Feature)

The hl_ratio is the single most important feature because it contextualizes all other signals. A bullish candlestick pattern with hl_ratio of 1% is very different from the same pattern with hl_ratio of 5%. The former suggests confidence, the latter suggests uncertainty.

**Key Insight: Range-based volatility is more informative than return-based volatility for short-term prediction because it captures intraperiod uncertainty that gets lost in open-to-close returns.**

# Feature 2: body_size (13.0% importance)

## Definition

Body size measures the absolute magnitude of the price movement from open to close, representing the core directional component of a candlestick.

## Domain Classification

- **Primary Domain:** Market Microstructure / Technical Analysis
- **Sub-Domain:** Candlestick Pattern Recognition
- **Category:** Momentum magnitude indicator

## Mathematical Formula

$$body\_size = |Close - Open|$$

The absolute value ensures body_size is always positive, measuring magnitude regardless of direction. Direction is captured separately by the sign of (Close - Open).

## Technical Intuition

In candlestick analysis, the body represents the battle between buyers and sellers from open to close. A large body indicates decisive victory for one side—either bulls (green candle) or bears (red candle). A small body (doji) indicates stalemate and indecision.

Body size captures commitment and conviction in price movement. Large bodies suggest strong institutional participation and trend continuation potential. Small bodies suggest uncertainty, exhaustion, or reversals.

## Layman's Explanation

Think of body_size as measuring how much the stock actually moved during trading, ignoring temporary spikes or dips.

If a stock opens at ₹100 and closes at ₹103, the body_size is ₹3. If it opens at ₹100 and closes at ₹100.10, the body_size is only ₹0.10.

A large body_size means traders decisively pushed the price in one direction. A tiny body_size means nobody could win—the price ended close to where it started despite whatever happened in between.

## What This Feature Represents

- **Directional Conviction:** Large bodies = strong trend, Small bodies = consolidation
- **Market Efficiency:** Large bodies suggest efficient price discovery in one direction
- **Momentum Quality:** Sustained large bodies indicate genuine trend, sporadic ones indicate noise

*Example: Stock opens at ₹100, high ₹104, low ₹99, closes ₹103. body_size = |103-100| = ₹3. Combined with hl_ratio of 5%, this shows strong upward momentum (large body) but with volatility (wide range).*

**Key Insight: body_size works in conjunction with hl_ratio. Large body + small range = confident move. Large body + large range = volatile momentum. Small body + large range = indecision.**

# Feature 3: hour_cos (10.5% importance)

## Definition

Cosine transformation of the hour of day, encoding time cyclically to capture intraday patterns without artificial discontinuities at market boundaries.

## Domain Classification

- **Primary Domain:** Temporal Patterns / Market Microstructure
- **Sub-Domain:** Intraday Seasonality
- **Category:** Cyclical time encoding

## Mathematical Formula

$$hour\_cos = \cos(2\pi \times hour / 24)$$

**Where:**

- hour = Current hour in 24-hour format (10 for 10:00 AM, 15 for 3:30 PM)
- 2π = Full rotation (360 degrees) to map 24 hours onto a circle
- cos() = Cosine function, ranging from -1 to +1

## Technical Intuition

Raw hour encoding (10, 11, 12...) creates artificial discontinuities: hour 15 (3:30 PM) and hour 10 (10:00 AM) appear far apart numerically despite being adjacent in the next trading day. Cyclical encoding via cos/sin solves this by mapping time onto a circle.

The cosine component captures the primary phase of intraday patterns. For NSE trading hours (10:00-15:30):

- hour_cos ≈ 0.5 at market open (10:00)
- hour_cos ≈ -0.7 at mid-day (12:30)
- hour_cos ≈ -0.3 at market close (15:30)

## Layman's Explanation

Imagine the trading day as a clock face. Instead of saying '10:00' or '3:30', we describe where we are on the clock using angles.

Why? Because the stock market behaves differently at different times. The first 15 minutes (market open) is chaotic with lots of volatility. The lunch hour is calmer. The last 30 minutes (closing) gets hectic again.

By encoding time as cos/sin, the computer can learn: 'Whenever we're at this angle on the clock (e.g., market open), prices tend to move differently than at that angle (e.g., lunch time).'

## What This Feature Represents

- **Intraday Liquidity Patterns:** Open/close have different liquidity profiles than mid-day
- **Institutional Behavior:** Algos and traders follow time-based schedules
- **Information Flow:** News releases cluster at specific times, affecting volatility

*Example: At 10:00 AM (market open), hour_cos ≈ 0.5. At 12:00 PM (midday), hour_cos ≈ -1.0. The model learns that patterns at 0.5 (open volatility) differ from patterns at -1.0 (midday calm).*

**Key Insight: Time-based features capture market microstructure effects like the 'market open effect' (high volatility 10:00-10:30) and 'closing rush' (elevated activity 15:00-15:30) that are invisible to price/volume features alone.**

# Feature 4: volume_ratio (9.4% importance)

## Definition

Volume ratio measures current trading volume relative to its recent 20-period average, detecting abnormal liquidity events.

## Domain Classification

- **Primary Domain:** Market Microstructure / Liquidity Analysis
- **Sub-Domain:** Volume Anomaly Detection
- **Category:** Relative volume indicator

## Mathematical Formula

$$volume\_ratio = Volume(t) / MA_{20}(Volume)$$

**Where:**

- Volume(t) = Trading volume in current 15-minute bar
- $MA_{20}$(Volume) = 20-period moving average of volume (5 hours of trading)

## Technical Intuition

Volume is the fuel of price movement. High volume confirms the legitimacy of price moves, while low volume suggests weak conviction. However, absolute volume varies widely across stocks and time periods, making raw volume non-comparable.

The volume_ratio normalizes current activity against recent baseline, enabling detection of:

- Volume spikes (ratio > 2.0) indicating new information or institutional activity
- Volume droughts (ratio < 0.5) suggesting market disinterest or pre-announcement quiet periods
- Volume trend shifts that precede price movements

## Layman's Explanation

Think of volume as the number of people shopping at a store. On a normal day, maybe 100 people visit per hour. The volume_ratio tells you if today's hour had way more or way fewer shoppers than usual.

If volume_ratio = 3.0, it means 3x the normal number of shares are trading. Something interesting is happening—maybe good news, bad news, or big institutions buying/selling.

If volume_ratio = 0.3, trading has dried up. The stock is quiet, boring, nobody's interested. These quiet periods often come before big moves when traders are waiting for news.

## What This Feature Represents

- **Information Events:** Volume spikes correlate with news, earnings, or institutional flows
- **Price Move Confirmation:** High volume + large body_size = credible trend

- **Liquidity Regime:** Persistent low volume_ratio suggests illiquidity and higher execution risk

  *Example: Stock typically trades 10,000 shares per 15-min bar. Suddenly, 35,000 shares trade in one bar. volume_ratio = 35,000/10,000 = 3.5. This 3.5x spike flags unusual activity. If price also jumped, it confirms a real move. If price stayed flat, it suggests large hidden orders.*

**Key Insight: Volume ratio is especially important for Indian markets (NSE/BSE) where retail participation creates distinct volume patterns around market open, news events, and F&O expiry days.**

# Feature 5: hour_sin (8.1% importance)

## Definition

Sine transformation of the hour of day, complementing hour_cos to provide complete cyclical time encoding.

## Domain Classification

- **Primary Domain:** Temporal Patterns
- **Sub-Domain:** Intraday Seasonality
- **Category:** Cyclical time encoding (orthogonal component)

## Mathematical Formula

$$hour\_sin = \sin(2\pi \times hour / 24)$$

## Technical Intuition

While hour_cos provides the primary phase encoding, hour_sin provides the orthogonal (perpendicular) component. Together, they uniquely identify any point on the time circle without ambiguity.

Using only cos would create ambiguity: hour 10 and hour 14 might have similar cos values but represent different market regimes. Adding sin resolves this by creating unique (cos, sin) pairs for each hour.

For NSE trading:

- hour_sin ≈ 0.87 at market open (10:00)
- hour_sin ≈ 0.0 at mid-day (12:00)
- hour_sin ≈ -0.97 at market close (15:30)

## Layman's Explanation

If hour_cos is like the X-coordinate on a map, hour_sin is like the Y-coordinate. You need both X and Y to pinpoint exactly where you are.

hour_cos alone might say 'you're somewhere in the middle of the map,' but hour_sin adds 'and you're on the upper half.' Together, they tell you exactly where you are in the trading day.

## What This Feature Represents

The combination of hour_sin and hour_cos (18.6% combined importance) captures:

- Opening volatility effect (10:00-10:30): High activity, wide spreads
- Lunch doldrums (12:00-13:30): Low volume, tight ranges
- Closing rush (15:00-15:30): Institutional rebalancing, elevated volatility

**Key Insight: Time features account for 18.6% of model importance, making them the second most important category after microstructure. This highlights that 'when' you trade matters as much as 'what' you see in price/volume.**

# Feature 6: high_close_diff (7.7% importance)

## Definition

The upper shadow (or upper wick) measures the distance between the period high and the closing price, indicating rejection of higher price levels.

## Domain Classification

- **Primary Domain:** Technical Analysis / Market Psychology
- **Sub-Domain:** Supply/Demand Imbalance Detection
- **Category:** Candlestick shadow analysis

## Mathematical Formula

```
high_close_diff = High - Close
```

## Technical Intuition

The upper shadow represents a failed attempt to sustain higher prices. When the high is significantly above the close, it indicates:

- Selling pressure emerged at higher levels
- Buyers lost conviction and retreated
- Resistance level was tested and rejected

In contrast, a small or absent upper shadow suggests buyers maintained control throughout the period, with closing price near the high—a bullish indication.

## Layman's Explanation

Imagine bidding for a house. At some point during negotiations, you offered your highest price (the 'high'). But by the end, you settled for less (the 'close'). The difference between your peak offer and final price is the high_close_diff—it shows you backed down.

In stocks, if the price hits ₹105 during a 15-minute period but closes at ₹102, the high_close_diff is ₹3. This means buyers pushed up to ₹105 but couldn't hold it—sellers pushed back down. It's a sign of seller strength.

## What This Feature Represents

- **Supply Zones:** Large upper shadows mark price levels where sellers dominate
- **Momentum Exhaustion:** Increasing upper shadows after strong rallies signal trend fatigue
- **Reversal Signals:** Long upper shadow with small body = potential bearish reversal

  *Example: Stock opens at ₹100, rallies to ₹105 (high), then retreats to close at ₹101. high_close_diff = ₹4. The ₹105 level attracted heavy selling, pushing price back down. If this pattern repeats, ₹105 becomes a known resistance level.*

**Key Insight: high_close_diff works asymmetrically with lower_shadow (Feature 10). Large upper shadows are bearish, large lower shadows are bullish. The model learns directional biases from wick asymmetry.**

## Features 7-10: Summary

The remaining four features provide incremental predictive value, collectively accounting for 16% of total importance:

| Feature | Importance | Primary Function |
|---|---|---|
| close_5ma_diff_pct | 5.7% | Short-term mean reversion signal |
| close_log_return_lag1 | 4.1% | Immediate momentum/autocorrelation |
| upper_shadow | 3.2% | Resistance detection, supply zones |
| lower_shadow | 3.0% | Support detection, demand zones |

# Feature 7: close_5ma_diff_pct (5.7% importance)

## Definition

Percentage deviation of current close price from its 5-period moving average, measuring short-term mean reversion pressure.

## Mathematical Formula

$$\texttt{close\_5ma\_diff\_pct} = (\text{Close} - \text{MA}_5(\text{Close})) / \text{MA}_5(\text{Close})$$

## Layman's Explanation

This tells you if the stock price is currently above or below its recent average. If a stock has been trading around ₹100 for the last hour (5 bars × 15 min = 75 minutes), but suddenly jumps to ₹103, the close_5ma_diff_pct is +3%.

Positive values mean the stock is stretched above its average (potential pullback). Negative values mean it's below average (potential bounce). This captures short-term mean reversion tendency.

**Captures: Short-term overextension and mean reversion signals. Most effective in ranging markets, less reliable in strong trends.**

# Feature 8: close_log_return_lag1 (4.1% importance)

## Definition

The most recent bar's logarithmic return, capturing immediate momentum or mean reversion effects.

## Mathematical Formula

$$\texttt{close\_log\_return\_lag1} = \ln(\text{Close}(t-1) / \text{Close}(t-2))$$

## Layman's Explanation

This is simply 'what happened in the last 15 minutes?' If the stock jumped 2% last bar, this feature tells the model 'hey, we just had a 2% move, what does that mean for the next 15 minutes?'

In momentum regimes, positive lag-1 return predicts continued upward movement. In mean-reversion regimes, positive lag-1 return predicts pullback. The model learns which regime is active.

**Captures: Autocorrelation in returns. Indicates if market exhibits momentum (trend continuation) or mean reversion (reversal) tendencies at 15-minute scale.**

# Feature 9: upper_shadow (3.2% importance)

## Definition

Alternative measurement of the upper wick, calculated as distance from the maximum of open/close to the high.

## Mathematical Formula

```
upper_shadow = High - max(Open, Close)
```

## Relationship to high_close_diff

While high_close_diff always measures from high to close, upper_shadow measures from high to the top of the candle body (whichever is higher: open or close).

For green candles (close > open): upper_shadow = high_close_diff

For red candles (close < open): upper_shadow = High - Open

**Captures: Candle-direction-agnostic wick measurement. Provides redundant but complementary information to high_close_diff for model robustness.**

# Feature 10: lower_shadow (3.0% importance)

## Definition

Distance from the period low to the bottom of the candle body, measuring failed attempts to sustain lower prices.

## Mathematical Formula

```
lower_shadow = min(Open, Close) - Low
```

## Layman's Explanation

This is the opposite of upper_shadow. If the stock dropped to ₹97 during the period (the low) but recovered to close at ₹100, the lower_shadow is ₹3. This shows buyers stepped in at ₹97 and pushed the price back up—a bullish sign.

Long lower shadows indicate demand zones (support levels). Short lower shadows indicate sellers maintaining control.

**Captures: Demand zones, buying pressure, support levels. Asymmetric with upper_shadow—models learn directional bias from shadow imbalance.**

# Feature Interactions & Combinations

While individual features provide value, their true predictive power emerges from interactions:

## Interaction 1: hl_ratio + body_size

- **High hl_ratio + Large body_size:** Volatile momentum—strong move with uncertainty
- **Low hl_ratio + Large body_size:** Controlled momentum—confident directional move
- **High hl_ratio + Small body_size:** Indecision—high volatility but no resolution

## Interaction 2: volume_ratio + body_size

- **High volume_ratio + Large body_size:** Confirmed trend—institutional participation
- **Low volume_ratio + Large body_size:** Suspect move—lack of conviction, potential trap

## Interaction 3: hour_cos/sin + hl_ratio

- **Market open (hour ~10) + High hl_ratio:** Expected volatility—normal market behavior
- **Mid-day (hour ~12) + High hl_ratio:** Abnormal volatility—potential news event

## Interaction 4: Shadow Asymmetry

- **Large upper_shadow + Small lower_shadow:** Bearish—rejection of higher prices
- **Small upper_shadow + Large lower_shadow:** Bullish—support found, buyers active
- **Balanced shadows:** Neutral—equilibrium, no directional bias

# Conclusion

The 10 features detailed in this report represent a carefully balanced blend of market microstructure, temporal patterns, and momentum indicators. Their collective 91.6% importance demonstrates that predictive power in financial markets comes not from algorithmic sophistication but from feature engineering grounded in market mechanics.

## Key Takeaways

4. **Microstructure Dominates:** 61% of importance comes from price range, body size, and wick analysis—features that capture market psychology and supply/demand dynamics.
5. **Time Matters:** 18.6% importance from cyclical time encoding proves that 'when' you trade is as critical as 'what' you trade.
6. **Volume Validates:** Volume_ratio (9.4%) confirms that price movements without volume participation are unreliable.
7. **Simplicity Wins:** 10 features achieve 91.6% performance. Complex indicators like RSI and MACD contributed <1% and were eliminated.
8. **Feature Interactions:** Random Forest captures non-linear combinations—high body_size means different things depending on volume_ratio and hl_ratio context.

**Philosophical Insight:**

These features succeed because they measure observable market mechanics—volatility regimes, directional conviction, supply/demand imbalances, liquidity conditions, and temporal patterns—rather than attempting to divine future prices through pattern matching or mystical indicators.

**The strategy works not by predicting the unpredictable, but by detecting when the market reveals exploitable information through these ten measurable dimensions.**

—

*End of Report*