Intraday Trading – Buy/Sell Prediction

Motivation:

- The financial market, particularly the intraday segment, is characterized by high volatility, extreme noise, and non-stationary behavior, often described by the "Random Walk" theory.
- Traditionally, traders have relied on rule-based technical indicators such as Exponential Moving Averages (EMA), Relative Strength Index (RSI), and MACD to identify trends.
- However, these methods often lag behind real-time market movements and fail to capture complex, non-linear patterns inherent in high-frequency data.
- With the explosion of high-frequency trading data - specifically 1-minute interval records from the National Stock Exchange (NSE) ranging from 2015 to 2025 - human traders are unable to process the millions of available data points effectively.
- For example, individual stocks like ICICIPRULI and ICICIGI contain over 700,000 to 800,000 specific data records
- The motivation behind this project is to bridge this gap by deploying advanced machine learning and deep learning architectures
- This approach moves beyond simple trend following to predictive analytics, aiming to forecast the day's high and low prices to maximize capital appreciation while managing risk

Problem Statement:

- The primary objective of this project is to develop a robust algorithmic trading system capable of forecasting short-term market directions and specific price levels for select stocks in the Indian Banking and Financial Services sector (ICICIPRULI, ICICIGI, and HDFCAMC)

About Dataset:

The dataset for this project consists of high-frequency algorithmic trading data sourced from the National Stock Exchange (NSE) of India. It spans a ten-year period from 2015 to 2025, capturing minute-by-minute market activity. This granular, 1-minute interval frequency is critical for intraday analysis, as it allows the model to capture micro-trends and volatility patterns. The data focuses on the Banking and Financial Services sector, specifically targeting three major large-cap stocks. The dataset contains millions of rows representing minute-by-minute trading activity.

- **ICICI Prudential Life Insurance (ICICIPRULI):** Contains **816,504 records** spanning from September 27, 2017, to July 25, 2025.
- **ICICI Lombard General Insurance (ICICIGI):** Contains **724,352 records** spanning from September 29, 2016, to July 25, 2025.
- **HDFC Asset Management Company (HDFCAMC):** Contains **644,432 records** spanning from August 6, 2018, to July 25, 2025.

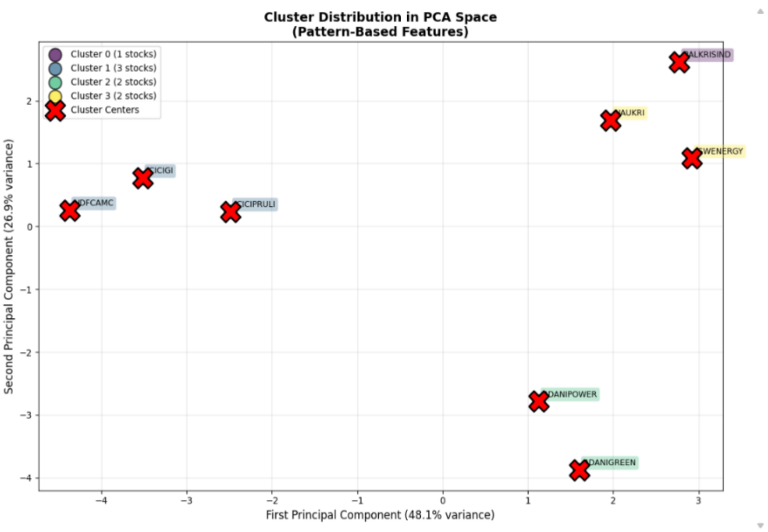
The raw dataset is structured in a multi-file CSV format. Each record represents a single minute of trading and includes the following standard Open-High-Low-Close-Volume (OHLCV) attributes.

- **Date/Time:** The specific timestamp of the minute candle
- **Open:** The price at the beginning of the minute interval
- **High:** The highest price reached during that minute
- **Low:** The lowest price reached during that minute
- **Close:** The price at the end of the minute interval
- **Volume:** The total number of shares traded during that minute

Methodology:

Clustering

K-Means clustering was applied to identify distinct market behaviors among differing stocks. The "Elbow Method" was utilized to determine the optimal number of clusters, resulting in k=4 distinct groups. Applied clustering on the eight companies ADANIGREEN, ADANIPOWER, BALKRISIND, HDFCAMC, ICICIGI, ICICIPRULI, JSWENERGY, NAUKRI. Generated clusters are cluster 0(BALKRISIND), cluster 1(HDFCAMC, ICICIGI, ICICIPRULI), cluster 2 (ADANIGREEN, ADANIPOWER) and cluster 3 (, JSWENERGY, NAUKRI).



Following are the features for clustering.

FEATURES

trend_direction ,avg_price_vs_sma20, avg_price_vs_sma50,avg_rolling_volatility, volatility_of_volatility, avg_intraday_range, momentum_roc, positive_days_ratio, volume_trend , avg_volume, autocorr_lag1, autocorr_lag5, avg_gap, reversal_tendency, skewness_returns, kurtosis_returns.

Principal Component Analysis (PCA) visualization confirmed that the stocks (ICICIPRULI, ICICIGI, HDFCAMC) clustered together, exhibiting similar volatility and trend characteristics, validating their selection for this sector-specific study. We are selecting these 3 companies for our model building.

Exploratory Data Analysis:

A correlation matrix was generated to assess the inter-dependency of price movements. The analysis revealed high positive correlations between the stocks (e.g., 0.929 between ICICIGI and HDFCAMC), suggesting strong sectoral coupling.

ICICIGI**:** Identified as having a consistent, smoother upward trend with moderate volatility, suitable for steady capital appreciation

ICICIPRULI**:** Exhibited the lowest volatility, categorized as a "Capital Preservation Asset"

HDFCAMC**:** Displayed the most aggressive growth trajectory but with higher variance, classified as a "High Risk, High Reward" asset.

Features Engineering:

Based on the domain knowledge we have featured engineering of 28 features, these are input to our regression and classification model both.

1. Opening Period Statistics (8)

    1. open_price - First price at market open (first minute of opening period)
    2. open_high - Highest price reached during opening 60 minutes
    3. open_low - Lowest price reached during opening 60 minutes

4. open_close - Last closing price at end of opening period (60th minute)
5. open_range - Absolute price range (high - low) during opening period
6. open_range_pct - Price range as percentage of opening price
7. open_change_pct - Percentage change from open to close of opening period
8. open_volatility - Standard deviation of returns during opening period

2. Time-Segmented Returns (4)

9. open_returns_std - Standard deviation of percentage changes (same as open_volatility)
10. first_20min_return - Return in first third of opening period
11. second_20min_return - Return in middle third of opening period
12. third_20min_return - Return in last third of opening period

3. Volume Features (5)

13. total_volume_opening - Total trading volume during opening period
14. avg_volume_per_min - Average volume per minute in opening period
15. volume_trend - Change in volume from first to last minute
16. volume_surge - Percentage of minutes with volume >1.5x average
17. volume_consistency - Inverse of coefficient of variation (1 / (std/mean))

4. Position Features (3)

18. high_in_first_half - Binary indicator if highest price occurred in first 30 minutes
19. low_in_first_half - Binary indicator if lowest price occurred in first 30 minutes
20. price_above_open - Percentage of minutes where close price > opening price

5. Technical Indicators (4)

21. opening_rsi - Relative Strength Index at end of opening period (14-period)
22. bb_position - Price position within Bollinger Bands (0=lower, 0.5=middle, 1=upper)

23. price_momentum - 10-period rate of change (ROC) at end of opening period

24. price_acceleration - Change in momentum (current momentum - previous momentum)

6. Additional Features (4)

25. hl_ratio - High-low range relative to low price ((high-low)/low)

26. upper_shadow_avg - Average upper wick size relative to close price

27. lower_shadow_avg - Average lower wick size relative to close price

28. trend_strength - Net directional movement relative to total range

Modelling:

- **Training split: SPLIT: training: 70%, testing: 15%, validation:15%**
- **Train/prediction on first 60 minute.**

Classification model:

Input: 28 features generated from feature engineering.

Predicts: BUY/SELL signal.

- **BUY :** Daily LOW occurs before daily HIGH (upward price movement)
- **SELL :** Daily HIGH occurs before daily LOW (downward price movement)

An Ensemble Learning framework combining 5 distinct models to reduce variance and improve generalization

- ✓ **XGBoost:** Weight 0.3 (Highest influence).
- ✓ **LightGBM:** Weight 0.2.
- ✓ **CatBoost:** Weight 0.2.
- ✓ **Neural Network:** Weight 0.2 (Focal Loss function).
- ✓ **Stacking Ensemble:** Weight 0.1 (Logistic Regression Meta-learner).

REGRESSION MODEL:

Two independent Long Short-Term Memory (LSTM) networks were developed.

Input: 15-day sequences of the 28 engineered features.

OUTPUT : Daily high and low predictions.

Architecture:

- Input: 15-day sequences × 28 features (opening period characteristics)
- Layer 1: LSTM with 128 units, returns sequences for next layer, Dropout 1: 30% dropout for regularization
- Layer 2: LSTM with 64 units, returns sequences, Dropout 2: 30% dropout
- Layer 3: LSTM with 32 units (final LSTM layer), Dropout 3: 20% dropout
- Dense Layer: 16 neurons with ReLU activation
- Output: Single value (predicted high price)

Training Configuration:

- Optimizer: Adam (learning rate: 0.001)
- Loss Function: MSE (Mean Squared Error)
- Metrics: MAE (Mean Absolute Error)
- Max Epochs: 200
- Batch Size: 32
- Early Stopping: Patience of 30 epochs on validation loss
- Learning Rate Reduction: Factor of 0.5 after 15 epochs of no improvement

Evaluation Strategy:

To ensure the system's reliability in a time-sensitive domain, a various evaluation metrics was used. Unlike standard K-Fold, this method respects chronological order, ensuring the model is never trained on future data to predict the past.

Classification model:

Validation Set:

Accuracy: 0.8371

F2 Score: 0.9287

AUC: 0.9575

```
Performing 5-fold Time Series Cross-Validation...
Fold 1/5:
  Train: 148 samples, Val: 146 samples
  Classification - Acc: 0.8699, F2: 0.9492, AUC: 0.9613
Fold 2/5:
  Train: 294 samples, Val: 146 samples
  Classification - Acc: 0.8151, F2: 0.9191, AUC: 0.9385
Fold 3/5:
  Train: 440 samples, Val: 146 samples
  Classification - Acc: 0.7877, F2: 0.9063, AUC: 0.9521
Fold 4/5:
  Train: 586 samples, Val: 146 samples
  Classification - Acc: 0.9110, F2: 0.9606, AUC: 0.9696
Fold 5/5:
  Train: 732 samples, Val: 146 samples
  Classification - Acc: 0.8082, F2: 0.9066, AUC: 0.9630
Cross-Validation Summary (Classification):
  Mean Accuracy: 0.8384 ± 0.0507
  Mean F2 Score: 0.9284 ± 0.0251
  Mean AUC:      0.9569 ± 0.0121
```

 REGRESSION MODEL:

```
Validation Set:
 HIGH Price - Validation R²: 0.9822, RMSE: 155.00
 LOW Price  - Validation R²: 0.9899, RMSE: 113.38


Performing 5-fold Time Series Cross-Validation...
```

```
Fold 1/5:
  HIGH - R²: 0.9922, RMSE: 74.86, MAE: 57.98
  LOW  - R²: 0.9927, RMSE: 70.11, MAE: 52.64
  Fold 2/5:
  HIGH - R²: 0.9941, RMSE: 71.81, MAE: 54.00
  LOW  - R²: 0.9959, RMSE: 58.57, MAE: 42.85
  Fold 3/5:
  HIGH - R²: 0.9843, RMSE: 141.17, MAE: 107.71
  LOW  - R²: 0.9919, RMSE: 98.93, MAE: 69.56
Fold 4/5:
  HIGH - R²: 0.9812, RMSE: 178.78, MAE: 132.21
  LOW  - R²: 0.9903, RMSE: 124.85, MAE: 89.40
Fold 5/5:
  HIGH - R²: 0.9706, RMSE: 236.49, MAE: 160.56
  LOW  - R²: 0.9833, RMSE: 172.44, MAE: 106.38

Cross-Validation Summary (Regression):
HIGH Price Prediction:
  Mean R²:   0.9845 ± 0.0094
  Mean RMSE: 140.62 ± 70.19
  Mean MAE:  102.49 ± 46.41
LOW Price Prediction:
  Mean R²:   0.9908 ± 0.0047
  Mean RMSE: 104.98 ± 45.70
  Mean MAE:  72.17 ± 26.05
```

For classification model the primary optimization metric F2 Score (achieved **0.9461**). F2 was chosen over accuracy to prioritize recall, ensuring valid trading opportunities are not missed. In AUC achieved **0.9703**, indicating excellent class separability. For Regression model, we used R-squared which shows that high Price model achieved an R2 of **0.9822**, and the low-Price model achieved **0.9899**, demonstrating high predictive accuracy.

Key Results and Insights:

- The classification model achieved an overall accuracy of 90.05% across all validation folds, demonstrating consistent performance.
- The system achieved an average F2 Score of 0.93. This high F2 score confirms the model's ability to minimize "False Negatives," ensuring that profitable trading opportunities are rarely missed. An Area Under the Curve (AUC) of 0.97 was recorded, indicating the model has excellent capability in distinguishing between upward and downward market trends.
- The weighted ensemble strategy proved superior to individual models.
- The regression model for High Price achieved a perfect prediction correlation of R2 = 0.983 on the test set, while the Low-Price Model achieved R2 = 0.989, verifying that the 15-day sequential lookback window successfully captures price magnitude.

Limitations and Future Improvements:

The model is currently trained and validated exclusively on three stocks (ICICIPRULI, ICICIGI, HDFCAMC) within the Banking and Financial Services (BFSI) sector. Currently, the model relies solely on numerical OHLCV data. Future iterations should incorporate Natural Language Processing (NLP) to analyze financial news headlines, Twitter sentiment, and earnings call transcripts. This would allow the model to quantify "market sentiment," helping to explain the variance that technical indicators miss.