Beyond the bell: Leveraging off-market data for AI-enabled stock directionality forecast

Himadri Mukherjee *
AILABS, India
himadri@ailabs.finance

Suchismita Das *
AILABS, India
suchismita@ailabs.finance

Kaustav Bose
AILABS, India
kaustav@ailabs.finance

Kundan Kumar
AILABS, India
kundan@ailabs.finance

Souren Paul
Drexel University, Philadelphia, USA
souren.paul@gmail.com

Alo Ghosh
AILABS, India
alo@aloghosh.com

*The authors contributed equally

Abstract

Stock directionality forecasts are extremely useful in the financial market aiding in more informed trading decisions. However, it is difficult due to the highly volatile nature of the stock market. The majority of the stock trading takes place during the regular market hours whose data is mostly used for forecasts. Trades are also executed before the market opens (pre-market) and after the market closes (post-market). This off-market trading data is often ignored due to its minute trading volume. Exploration of this data for stock market forecasting is in its nascent state. We forecast the directionality of the end-of-the-day price using this off-market along with regular market hour data. The proposed AI-enabled framework extracts useful features from the off-market data, and 15 *technical indicators based on regular market data followed by a tree-based prediction approach. The obtained results show performance improvements of over* 7% *in closing price directionality forecast when the off-market hour-based features are incorporated.*

Keywords: Pre-market data, Post-market data, AI-enabled directionality forecast, Extended hours trading

1. Introduction

Stock market forecasting refers to the prediction of future values of stock price, return, or directionality. Accurate forecasting of stock prices or directionality enables individual/ institutional investors, and traders to make informed decisions and earn higher returns in the market. Stock market forecasting also helps in risk management, strategy making, economic health analysis, and investment policy implementation. However, forecasting the stock market is an extremely challenging task due to the volatility of the market. The stock market is influenced by several domestic and global factors including economic conditions, political situation, government policies, psychology of the investors, and company-specific variables. As a result, stock price movements are non-stationary, non-linear, noisy, and complex to model (Atsalakis and Valavanis (2009), Dassanayake et al. (2019), Kumbure et al. (2022), and Shah et al. (2019)). Traditionally, the statistical time-series-based methods like moving average, auto-regressive conditional heteroscedasticity (ARCH) model, generalized Auto-Regressive Moving Average (GARCH), auto-regressive moving average (ARMA), autoregressive integrated moving average (ARIMA) models, and Kalman filtering etc. were used for stock market forecasting (Gandhmal and Kumar (2019) and Kumbure et al. (2022)). Over the last few decades, with the surge of artificial intelligence (AI) and machine learning (ML) models, stock market forecasting frameworks have adapted to use the AI models. Most of these models are suitable for handling complex, non-linear data and hence become promising tools for stock market forecasts. Irrespective of the underlying model, features (more commonly termed as 'factors' in finance) play a pivotal role in these AI models. There are mainly 2 types of analytical features- fundamental and technical. Fundamental features for a stock are obtained or derived from the financial statements or reports from the respective company. It includes revenues, expenses, growth rate, earnings, assets, liabilities, turnover, and so on (Kumar et al. (2021) and Kumbure et al. (2022)). As the financial reports and statements are published quarterly and annually by a company, they are naturally preferred for longer-term forecasts (Kumar et al.

URI: https://hdl.handle.net/10125/108981
978-0-9981331-8-8

(2021)). Alternatively, technical features for a stock are technical indicators and charts characterizing the price trends that are computed from the historical stock price and volume data (Dassanayake et al. (2019) and Kumbure et al. (2022)). Rate of Change (ROC), Commodity Channel Index (CCI), and Moving Average Convergence Divergence (MACD) are examples of technical indicators (Boyle and Kalita (2023) and W. Chen et al. (2021)). Technical indicators can be defined for both long and short terms and are therefore useful in both long and short-term forecasts (Ahmed and Goyal (2023)).

In this study, we forecast end-of-the-day (EOD) price directionality (relative to previous EOD prices) using an AI-enabled approach. We employ a decision tree-based model named XGBoost (T. Chen and Guestrin (2016)) for this purpose. Previous studies (Dezhkam and Manzuri (2023) and Vuong et al. (2022)) have shown that XGBoost models are successful in stock market forecasting. As we consider the current day or one day ahead forecast, fundamental features are not likely to have an effective impact. However, technical features would be appropriate for this short-term forecasting. As the task is the daily forecast of EOD prices, our considered features were daily Open, High, Low, Close, and Volume data (OHLCV) (Yearner (2021)) and a set of technical indicators computed on the OHLCV in daily frequency. In standard terminology, OHLCV refers to the price and volume data evoked during the regular market hours of a business day. However, in prominent stock exchanges, trades also take place outside the regular market hours via electronic networks. The period outside the regular market hours is termed as off-market or extended trading hours. Off-market hours are divided into two segments- pre-market trading hours and after-hours or post-market trading hours. Pre-market refers to the period when electronic trading is allowed before the market opens and similarly, post-market refers to the period of a few hours when electronic trading is allowed after the market closes (Langager (2024)). Price and volume data are also evoked during off-market. However, this data is mostly not considered in stock market analysis and forecast. Most of the reported forecasting models that forecast future price or directionality use features computed from the regular market hours data (Bathla et al. (2023), Hoseinzade and Haratizadeh (2019), and Khaidem et al. (2016)). In this study, our objective is to investigate the usefulness of off-market data in directionality forecasting. We employed an XGBoost-based regressor for forecasting the directionality of EOD price. As input, the XGBoost-based regressor was fed a set of useful features extracted from the off-market data along with fifteen technical indicators based on regular market data. We conducted a comparative study between the performances of different models exploiting regular market hour-based features, off-market hours-based features, and both. The results showed a considerable performance improvement when the features extracted from the off-market data were added with the features based on regular market hour data.

Next we present the related work on stock market forecasting in Section 2 followed by the description of the methodology in Section 3. We describe the dataset in Section 4 and present the results in Section 5. We discuss the findings of the study in Section 6 and present our concluding remarks in Section 7.

2. Related Work

Here, we present some prominent works on stock price directionality forecast. The study in (Patel et al. (2015)) provides a comparative analysis of the next day's price movement prediction performances of 4 common classifiers: Artificial Neural Network (ANN), Support Vector Machine (SVM), Random Forest (RF), and Naive-Bayes (NB) on 2 Indian stocks namely Reliance Industries and Infosys Ltd. and 2 Indian stock price indices namely CNX Nifty and S&P Bombay Stock Exchange (BSE) Sensex. The classifiers were trained with 10 technical indicators. Khaidem et al. (Khaidem et al. (2016)) proposed an RF-based scheme using 6 technical indicators on daily price to predict the direction of the closing price after 1, 2, and 3 months. They considered 3 prominent US stocks namely Apple Inc., General Electric Co., and Samsung Electronics Co. Ltd. Chen and He (S. Chen and He (2018)) used a Convolutional Neural Network (CNN) employing a 1-dimensional convolution function to process OHLCV information of stock prices to predict 10 days movement of the stock price for stocks of the Chinese market. Hoseinzade and Haratizadeh (Hoseinzade and Haratizadeh (2019)) proposed a CNN-based framework named 'CNNPred' which processes a 82 features-based input data to predict the direction of the next day's price. They applied the method to 5 US stock indices. Chen et al. (W. Chen et al. (2021)) proposed a model named 'Graph convolutional feature-based CNN' (GC-CNN) for predicting next-day price movement. GC-CNN converts the stock market information and the individual stock information into images. The images are fed into 2 CNN-based networks They experimented with 6 stocks from the Chinese stock market. Zhang et al. (Zhang et al. (2022)) proposed a framework, based on the

transformer model and multiple attention mechanisms, named TEANet for forecasting the directionality of the next day's adjusted closing price. Stock price data were fused with tweets and news headlines data to form the input feature for TEANet. The model was tested on 4 standard datasets comprising stock-related tweets and news. Chandola et al. (Chandola et al. (2023)) proposed a framework to combine stock prices with the features extracted by the 'Word2Vec' model from news headlines and fed the combined features to Long Short Term Memory (LSTM) to predict intra-day price movements of 5 stocks. Campisi et al.(Campisi et al. (2023)) conducted a comparative study on the performances of 5 classification and 6 regression models for predicting the direction of S&P 500 returns after 30 days with only volatility-based features (11 implied volatility indices). Li et al. (Li et al. (2023)) proposed a model named PEN to align text and price streams for joint explainable modeling of the next day's price movement on ACL18 and DJAI datasets. Boyle and Kalita (Boyle and Kalita (2023)), proposed a Spatiotemporal Transformer model for predicting the next day's adjusted closing price movement on the ACL18 and KDD17 datasets. The model was trained with OHLCV data and 18 technical indicators. The discussed works are summarized in Table 1. Most works adhere to OHLCV and technical indicators based on regular market hours. Leveraging input features based on off-market data in AI models is challenging due to less availability of historical data for off-hours trading. In Lam and Mok (2002), the authors used the NASDAQ after-hours and pre-market indexes in 3 regression models- linear regression, partitioned models, and neural network for forecasting NASDAQ price. The studies in C. Chen et al. (2009) and Koehler (2009) used off-market hours-based features but forecast stock volatility. To the best of our knowledge, off-market data has not been extensively used for stock price/ directionality forecast. We attempt to explore this in our study.

3. Methodology

To bridge the lack of studies exploring off-market hour-based features in price directionality forecasting models, we propose a forecasting framework that exploits off-market hour-based features. The features and the predictor are discussed in the following subsections.

3.1. Features

All features are computed daily from historical stock data. We computed 11 features on the off-market data, using 1 min interval intra-day data. These 11 features are computed independently for both pre-market and post-market data. Thus, a total of 22 off-market hour-based features were used. The following are the post-market hour-based features.

- Open_post- Opening price of post-market hours.
- Close_post- Closing price of post-market hours.
- High_post- Highest price of post-market hours.
- Low_post- Lowest price of post-market hours.
- Total_Volume_post- Sum of the volumes of 1 min interval intra-day post-market data.
- Mean_post- Mean of stock prices over 1 min interval intra-day post-market data.
- STD_post- Standard deviation of stock prices over 1 min interval intra-day post-market data.
- Num_of_rise_post- Total number of stock price rises over 1 min interval intra-day post-market data.
- Num_of_fall_post- Total number of stock price falls over 1 min interval intra-day post-market data.
- STD_rise_post- Standard deviation of stock price rises over 1 min interval intra-day post-market data.
- STD_fall_post- Standard deviation of stock price falls over 1 min interval intra-day post-market data.

The similar features were computed for the pre-market data as well.

We computed 15 Technical features (StockCharts (2024) and Thompson (2023)) for parameterizing the market-hours data. The computational procedure of the features is presented in Table 2. To explain the computational procedures easily some notations were used in Table 2.

3.2. Predictor

An XGBoost-based regressor was used for forecasting the EOD price. The predicted EOD price was compared to the previous EOD price to forecast the directionality. We modeled the task as a regression task where the daily EOD price was the target or response variable and the input features computed on daily historical data, are regressor or independent variables. To investigate the usefulness of pre-market and post-market-based features, we train the XGBoost-based regressor under 7 different configurations of input features as described in the following. In all the configurations, EOD price is the response variable. Irrespective of the regressors involving current day data, previous day data or both of them, the EOD directionality is always computed with respect to the previous EOD price. The intra-day

Table 1. Performance of the discussed works targeted towards price directionality forecast in different horizons.

Author	Methods	Input Features	Target	Dataset	Performance
Patel et al., 2015	ANN, SVM, RF, NB	10 technical indicators	Next day's closing price	2 Indices and 2 stocks	Avg. Accuracy- ANN: 74.94%, SVM: 78.71%, RF:83.59%, NB:73.31% Avg. F-score- ANN:0.7659, SVM-0.8029, RF:0.8399, NB: 0.7461
Khaidem et al., 2016	RF	6 technical indicators	1/2/3 month closing price	3 stocks	Avg. accuracy- 1 month: 86.61%, 2 months: 91.51%, 3 months:93.68%
S. Chen and He, 2018	CNN	OHLCV	10 days ahead closing price	Stocks of Chinese stock market	Avg. Accuracy- 74.47%, Avg F-score- 61.13%
Hoseinzade and Haratizadeh, 2019	2D-CNNpred, 3D-CNNpred	82 features based on technical indicators, macro-economic and market data	Next day's closing price	5 indices	Avg. F-score- 2D-CNNpred: 0.4944, 3D-CNNpred: 0.4869
W. Chen et al., 2021	GC-CNN	OHLC, 10 technical indicators	Next day's closing price	6 stocks	Avg. Accuracy- 51.82%
Zhang et al., 2022	TEANet	OHLCV, tweets, news headlines	Next day's adjusted closing price	170 stocks	Avg. accuracy- 64.48%, Avg. MCC- 0.3534
Chandola et al., 2023	LSTM	News headlines, closing prices	Intra-day price	5 stocks	Avg. training accuracy 61.80%, Avg. validation accuracy 52.48%
Campisi et al., 2023	Logistic regression, LDA, Random forest classification, Bagging classification, Gradient boosting classification, Linear regression, Random forest regression, Bagging regression, Gradient boosting regression, Ridge regression, Lasso regression	11 Implied volatility indices	30 days returns	1 Index	Best performer- Accuracy-82.75%, AUC- 0.8495, F-score- 0.8845
Li et al., 2023	PEN	News, Social media post	Next day's closing price	ACL18, DJIA	Avg. Accuracy- 60.21%, Avg.MCC- 0.1880
Boyle and Kalita, 2023	Spatiotemporal transformer	OHLCV, 18 technical indicators	Next day's adjusted closing price	ACL18, KDD17	Avg. Accuracy- 60.29%, Avg. MCC- 0.1980

Table 2. The 15 Technical features used in this experiment along with their computational procedure.

Features	Computational procedure
Volume-based features	
Volume weighted average price (VWAP)	$VWAP = \sum(V \times TP)/\sum V$
Accumulation/Distribution Index (ADI)	$ADI = ADI_{prev} + V \times MFM$; $MFM = ((C-L)-(H-C))/(H-L)$
Force Index (FI)	$FI = EMA_{13}((C - C_{prev}) \times V)$
Money Flow Index (MFI)	$MFI = 100 - (100/(1 + (\sum MF_{+}/\sum MF_{-})))$; $MF_{+} = RMF$ if $RFM \times D > 0$ else $MF_{+} = 0$; $MF_{-} = RMF$ if $RFM \times D < 0$ else $MF_{-} = 0$; $RMF = V \times TP$
Ease of movement (EoM)	$EoM = SMA_{14}(DM/BR)$; $DM = ((H+L)/2 - (H_{prev} + L_{prev})/2)$; $BR = ((V/100,000,00)/(H-L))$
Momentum-based features	
Percentage Price Oscillator (PPO)	$PPO = 100 \times ((EMA_{12}(C) - EMA_{26}(C))/EMA_{26}(C))$
Rate of Change (ROC)	$ROC = 100 \times ((C - C_{12})/C_{12})$
Relative Strength Index (RSI)	$RSI = 100 - (100/(AG/AL))$; $AG = (13 \times AG_{prev} + G)/14$; $AL = (13 \times AL_{prev} + L)/14$; $G = C - C_{prev}$ if $C - C_{prev} > 0$ else $G = 0$; $L = C_{prev} - C$ if $C - C_{prev} < 0$ else $G = 0$;
True strength index(TSI)	$TSI = 100 \times (DS(PC)/DS(\lvert PC \rvert))$; $DS(x) = EMA_{13}(EMA_{25}(x))$;
Williams %R (%R)	$\%R = (-100) \times (Highest_{14}(H) - C)/(Highest_{14}(H) - Lowest_{14}(L))$
Trend-based features	
Mass Index (MI)	$MI = \sum EMAR$; $EMAR = SEMA/DEMA$; $SEMA = EMA_9(H-L)$; $DEMA = EMA_9(SEMA)$
Schaff Trend Cycle (STC)	$STC = 100 \times (MACD - \%K(MACD))/(\%D(MACD) - \%K(MACD))$; $MACD = EMA_{12}(C) - EMA_{26}(C)$; $\%K(x) = 100 \times (x - Lowest_{14}(x))/(Highest_{14}(x) - Lowest_{14}(x))$, $\%D(x) = SMA_3(K(x))$
Percent Rate of Change of Triple Exponential Moving Average (TRIX)	$TRIX = 100 \times (TEMA - TEMA_{prev})/TEMA_{prev}$; $TEMA = EMA_{15}(DEMA)$; $DEMA = EMA_{15}(EMA_{15}(C))$;
Detrended Price Oscillator (DPO)	$DPO = C_{(20/2+1)} - SMA_{20}(C)$;
Commodity Channel Index (CCI)	$CCI = (TP - SMA_{20}(TP))/(0.015 \times MD)$; $MD = \sum_{i=0}^{19} \lvert TP_i - SMA_{20}(TP) \rvert/20$

* $TP = (H+L+C)/3$ is typical price. $EMA_n(x)$ refers to n-period exponential moving average of x. X_{prev} refers to previous day value of X. D symbolizes the directionality of the price with respect to the previous day. $SMA_n(x)$ refers to n-period simple moving average of x. $Highest_n(x)$ is highest value of x in previous n periods. $Lowest_n(x)$ is lowest value of x in previous n periods.

features are computed using 1-minute interval intra-day data.

- Baseline- In this configuration, the input feature set only contains the 15 regular OHLCV-based technical indicators. Here, technical indicators of the previous day are the regressors.

- Post- In this case, the input contains the 11 off-market hour-based intra-day features. Here, post-market hour-based features of the previous day are the regressors.
- Pre- In this case, the input contains the 11 off-market hour-based intra-day features. Here, pre-market hour-based features of the current day are the regressors.
- Post+Pre- In this configuration, the input contains the 22 off-market hour-based intra-day features. Here, 11 features are computed from post-market and pre-market data each.In this case, post-market hour-based features of the previous day and pre-market hour-based features of the current day are the regressors.
- Baseline+Post- Here, the input contains the 15 features from Baseline and 11 features from Post. In this case, technical indicators of the previous day (regular) and post-market hour-based intra-day features of the previous day are the regressors.
- Baseline+Pre- In this configuration, the input contains the 15 features from Baseline and 11 features from Pre. Here, technical indicators of the previous day (regular) and pre-market hour-based intra-day features of the current day are the regressors.
- Baseline+Post+Pre- In this configuration, the input ontains the 15 features from Baseline and 22 features from Post+Pre. Here, technical indicators of the previous day (regular), post-market hour-based intra-day features of the previous day, and pre-market hour-based intra-day features of the current day are the regressors.

4. Dataset

In this study, 500 stocks listed under the $S\&P500$ index were considered. These were considered due to the presence of the largest and most influential stocks under this index [1]. The daily closing price of every stock was considered starting from their inception till "21st May-2024". The tickers originally represented 116 industries and 11 sectors thereby ensuring variety in the dataset.

Information from the regular trading hours ranging from $09:30:00$ Hrs to $16:00:00$ Hrs along with pre and post-market data was considered in this study. The daily open, high, low, close, and volumes were used from the regular hours for the calculation of technical factors while 1 minute data from the pre and post-market hours was used for the calculation of statistical features. Only those tickers were considered, that had at least 600 days of data which reduced the number of tickers to 494. There were certain tickers, where adequate information was not available in the pre and post-market trading hours and those were eliminated as well finally leading to a set of 473 stocks. The details of the sectors along with the number of industries and tickers are presented in Table 3.

Table 3. The number of sectors and industries in the dataset.

Sector	Industry		Ticker	
	C	D	C	D
Basic Materials	7	0	22	0
Communication Services	4	0	20	0
Consumer Cyclical	18	0	56	1
Consumer Defensive	11	0	35	1
Energy	6	0	24	0
Financial Services	11	1	63	2
Healthcare	10	1	59	5
Industrials	19	3	66	8
Real Estate	9	0	29	2
Technology	10	0	71	4
Utilities	5	1	28	4

* C and D denotes the number of considered and discarded entities from the original set due to unavailability of data.

As the number of tickers was very large (for presentation in a Table), they are presented using wordcloud representation in Figure 1. The tickers are listed to provide a brief idea of the stocks because the constituents of the $S\&P500$ index is dynamic and changes asynchronously.[2]

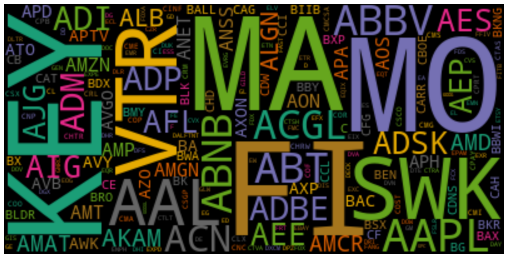

Figure 1. Wordcloud of tickers in the dataset. This presents an idea of the present candidates of the dynamic $S\&P500$ constituents.

[1] https://www.investopedia.com/best-25-sp500-stocks-8550793

[2] https://www.moneydigest.com/1564861/how-s-and-p-500-index-rebalancing-works/

It is noted that there were multiple days for different tickers where no trading was done during pre or post-market hours. Such days were discarded from the dataset. The average number of available data points/ candles (1 minute frequency) for different stocks in a sector along with the standard deviation of the same for regular, pre and post-market hours is presented in Table 4. The metrics were computed using the entire length of the available data as well as for the last 600 days. This was done to demonstrate the recent trend of stocks that might get skewed if an extremely long horizon is considered.

It is noted that the average number of 1-minute candles in a daily level since inception was highest for the "Communication Services" sector. This same trend was observed when only the last 600 days were used for analysis. In the case of the total data, the lowest deviation for the average number of candles was observed for the "Consumer Defensive" sector while for the last 600 days the lowest deviation was obtained for "Utilities". The average number of pre and post-market candles both for the total data as well as the last 600 days was highest for "Communication Services". However, the deviation was also the highest, thereby pointing to extremities. The mean closing price and its deviation across stocks, aggregated across the respective sectors is presented in Table 5.

It is noted from the Table the highest deviation in the closing price was observed for the "Consumer Cyclical" sector. The mean price was also the highest for this. Both the lowest mean price and mean deviation were obtained for the "Utilities" sector. The trend was the same both for the entire length of data as well as the last 600 days. The overall range of deviations was high ranging from 5.59 to 44.76.

5. Results and analysis

In this section, we present the evaluation metric and protocol followed by a discussion of the obtained results.

5.1. Evaluation metric and protocol

The system was evaluated for 300 days based on directional accuracy (%) wherein the change in price from the present day to the next day was used to compute the actual directionality. If the price increased in the next day, a positive label (1) for the directionality was assigned while a decrease in the next day's price w.r.t. the presented day attracted a negative label (0). The forecasted price was used to compute the predicted label using the aforementioned technique and finally, these 2 labels were compared for accuracy computation. The mean accuracy was computed across Sectors and the entire Index. All the reported accuracies are balanced accuracies which aids in avoiding the effect of class imbalance.

5.2. Analysis

The tests were performed in 6 phases excluding the baseline. The same test data of 300 instances were initially modeled using the post-market features and pre-market features in isolation. This was followed by a system that used both the post and pre-market features. The next systems involved fusing the intra-day features from the post and pre-market along with the baseline features. Initially, the post and pre-market features were fused with the baseline system separately, and finally, both the post and pre-market features were combined with the baseline system. The obtained directional accuracies for these systems is presented in Table 6.

It is noted from the Table that all the proposed methods performed better than the baseline system. The best performance was obtained when both the post and pre-market features were fused along with the baseline features. The same trend was observed for both ticker level and sector level aggregation wherein a performance improvement of 7.11% and 7.28% was obtained for the respective aggregations over the baseline system. The second-best performance was obtained when only the pre-market features were used. The ticker and sector-level aggregated performance for this setup was 1.35% and 1.24% less than the best-performing setup. The detailed sector-wise performance for the best result is presented in Table 7. The table presents the number of tickers in each sector that had at least post and pre-market data for at least 2% of the days and the percentage of tickers with performance improvement for such cases is listed in column "Perf". The table also presents the percentage of tickers amidst all the tickers in a sector for whom the directional forecast accuracy improved. It is noted that for certain sectors like "Basic Materials" and "Energy", the performance improvement was obtained for 100% of the tickers for whom post and pre-market data was available for at least 2% of the days. In the case of the "Technology" sector that has the lion's share in the S&P500 index (30.6% as on 10^{th} June 2024 [3]), better performance was obtained for 77.46% of the stocks in this sector. Among the stocks in this sector that were above the 2% threshold in terms of the post and pre-market data, performance improvement was obtained for 96.23% of the stocks. The sensitivity, specificity, balanced and

[3] https://www.spglobal.com/spdji/en/indices/equity/sp-500/#data

Table 4. Average and Standard deviation of the number of candles per day in 1-minute frequency for the tickers computed from the time of inception (Total) and the last 600 days, aggregated across the sectors.

Sector	Total						Last 600 days					
	RM	RS	PRM	PRS	POM	POS	RM	RS	PRM	PRS	POM	POS
Basic Materials	364.47	45.33	5.49	8.02	5.02	5.36	373.79	23.19	8.16	8.69	5.42	5.06
Communication Services	391.78	40.41	18.79	19.55	17.79	17.62	413.01	30.18	35.52	22.54	30.55	20.43
Consumer Cyclical	370.27	51.11	11.16	15.9	10.9	14.35	381.76	26.51	20.59	13.23	17.66	11.89
Consumer Defensive	375.68	36.93	3.85	8.48	5.44	6.12	389.53	18.06	6.92	10.2	6.72	6.61
Energy	385.33	37.94	10.28	16.33	9.2	11.59	404.95	19	24.03	16.82	16.05	11.91
Financial Services	369.81	39.05	6.55	9.85	7.68	9.34	372.57	25.02	9.01	9.46	8.47	7.44
Healthcare	350.55	50.72	4.36	8.75	5.83	7.25	357.7	29.7	5.44	8.08	5.84	5.88
Industrials	357.11	49.02	4.84	8.53	6.15	7.62	358.21	27.75	7.07	7.2	6.7	5.49
Real Estate	331.78	72.11	1.23	4.77	3.01	3.59	363.62	23.15	2.31	3.57	2.72	2.61
Technology	371.78	43.45	11.67	14.59	13.27	15.28	378.5	31.45	21.2	13.2	18.35	13.44
Utilities	370.5	39.01	1.14	4.37	3.33	3.68	384.14	16.75	2.5	4.22	3.14	2.76

* RM refers to the average number of candles in regular trading hours, RS refers to the standard deviation of the number of available samples, PRM refers to the mean for pre-market hours, PRS refers to the standard deviation for pre-market hours, POM refers to the mean for post market hours, and POS refers to the standard deviation for post-market hours.

Table 5. Mean closing price and its deviation across stocks, aggregated w.r.t. the respective sectors.

Sector	TM	TS	LM	LS
Basic Materials	69.29	41.44	139.56	21.83
Communication Services	57.75	40.42	101.75	24.65
Consumer Cyclical	100.1	82.27	245.58	44.76
Consumer Defensive	47.59	33.34	106.95	13.54
Energy	42.56	24.52	78.08	15.45
Financial Services	66.98	50.65	139.5	18.17
Healthcare	79.22	69.53	201.23	31.23
Industrials	72.31	59.95	184.45	29.13
Real Estate	58.94	41.49	125.55	17.05
Technology	74.01	65.56	185.2	40.56
Utilities	32.56	19.1	63.21	5.59

* TM and TS refer to the mean and standard deviation computed across the entire available data while LM and LS refer to the mean and standard deviation that was calculated for the last 600 days. All the values are in ($).

Table 6. Directional accuracy (%) of the different systems aggregated across tickers (Total) and Sectors.

Systems	Total	Sector
Baseline	56.52	56.33
Post	57.23	57.06
Pre	59.72	59.68
Post+Pre	59.38	59.34
Baseline+Post	58.21	58.10
Baseline+Pre	59.49	59.48
Baseline+Post+Pre	60.54	60.43

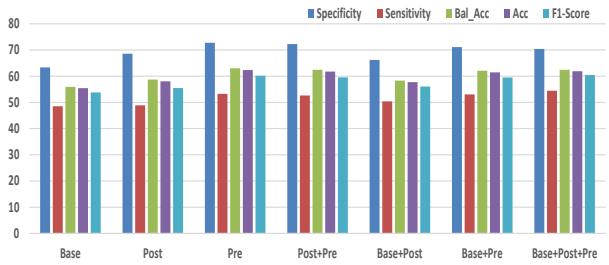

Figure 2. Sensitivity, Specificity, Balanced Accuracy (Bal_Acc), Overall Accuracy (Acc), and F1-Score for the different systems. Here "Base" denotes the baseline system.

overall accuracy, and F1-Scores for the different systems is presented in Figure 2.

It was observed that there were multifarious tickers for whom the performance was the same for the baseline system and the "Baseline+Post+Pre" system. A graphical representation of the percentage of tickers in each sector for whom higher, lower or equal performance was obtained by the proposed system as

Table 7. Performance of the "Baseline+Post+Pre" system. The accuracies (Accur.) are aggregated across the sectors.

Sector	Accur.	Tot.	Cnt.	Perf.	Btr.
Basic Materials	62.46	22	10	100	54.55
Communication Services	61.56	20	14	92.86	70
Consumer Cyclical	62.81	56	45	84.44	71.43
Consumer Defensive	58.82	35	25	96	80
Energy	64.31	24	23	100	100
Financial Services	60.98	63	43	83.72	58.73
Healthcare	63.37	59	30	96.67	59.32
Industrials	58.82	66	35	91.43	62.12
Real Estate	53.78	29	10	50	34.48
Technology	60.17	71	53	96.23	77.46
Utilities	57.70	28	11	90.90	42.86

* The total number of tickers is presented in the column "Tot", the number tickers having post and pre-market data for at least 2% of the total days is presented in "Cnt", "Perf" represents the percentage of tickers among "Cnt" for which performance improvement was obtained over the baseline. "Btr" represents the percentage of tickers among "Tot" for which performance improvement was obtained over the baseline.

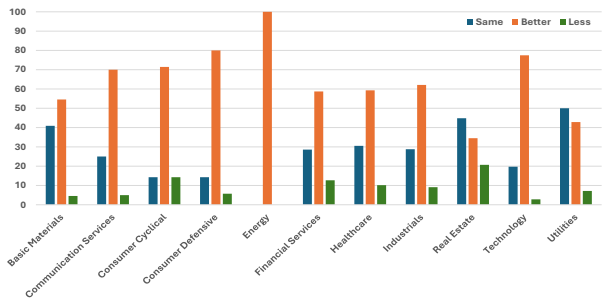

Figure 3. Percentage of tickers in each sector for whom higher, lower, or equal directional accuracy was obtained by the "Baseline+Post+Pre" system as compared to the baseline system.

compared to the baseline system is presented in Figure 3.

It is observed that for the "Utilities" sector, the performance of the proposed system and the baseline system was the same for 50% of the tickers while there was an improvement for 42.86% of the tickers. In the case of "Real Estate" similar performance from both the systems was obtained for 44.83% of the tickers. An improvement was observed for 34.48% of the tickers while the baseline results were better for the remaining 20.69% tickers. In the case of "Basic Materials" similar performance was obtained for 40.91% of the tickers while the performance improved for 54.55% of the tickers. In terms of improvement, the best performance was observed for the "Energy" sector, the directional accuracy improved for 100% of the tickers.

The system's (Baseline+Post+Pre) performance was also compared with the baseline performance in the purview of incremental increase of tickers. The average directional accuracies of both these systems were measured by incrementally increasing the number of stocks with a step of 50. The tickers were initially ordered in descending order based on market cap and were added in batches of 50. It was observed that the proposed system was better in all the steps. This ensures, the system is suitable for scenarios where new stocks are added to the existing stock universe. The performance is graphically presented in Figure 4.

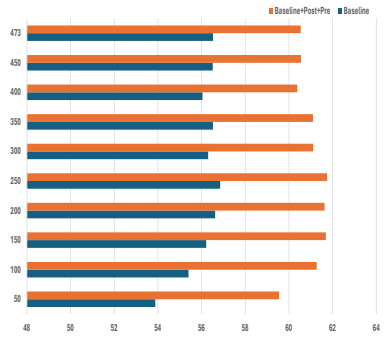

Figure 4. Cumulative directional accuracy of both the systems. The X-axis represents average directional accuracy in (%) while the Y-axis represents number of stocks in each batch.

6. Discussion

It is observed that for certain Sectors like "Real Estate" and "Utilities", the number of tickers demonstrating improved directionality accuracy was relatively less with respect to the other sectors. One probable reason for this is the fewer number of trades in the post and pre-market hours for these Sectors which led to skewed off-market features like standard deviation and mean to name a few. The performance was much better for Sectors like "Technology" where higher off-market trading is observed. In the case of "Technology", "Consumer Defensive", "Consumer Cyclical", and "Communication Services", the performance improved for 77.46%, 80%, 71.43%, 70% of the tickers. These were among the sectors that demonstrated highest improvement. On average, better performance was obtained for 90.64% of the

stocks that qualified over the 2% threshold (percentage of days out of the total number of days where off-market data was available). The system was also tested by gradually increasing the number of stocks from 50 to 473 with a step of 50. The proposed system outperformed the baseline in every step pointing to its suitability for scenarios where new stocks are often added to the existing stock universe. Although the performance of the pre-market-based system is very close to that of the proposed system, the associated extra computation (involving post market and regular hours) is not a hindrance because such computations can easily be completed between the end of post-market and beginning of pre-market (at least 6 hours). Moreover, active systems in this gap also ensure the ability to adjust to any event that might influence the target price.

6.1. Limitations

The performance of this system is dependent on the availability of off-market data. Thus it is not suitable for stocks that are not traded adequately in off-market hours. The study has been performed on the $S\&P500$ stocks. In future, experiments will be performed on a larger set of stocks from the RUSSELL 3000 index, ETFs, commodities, and penny stocks.

6.2. Implications for research and practice

The use of off-market 1-minute data for forecasting EOD prices has been mostly unexplored. This study demonstrates that using features derived from off-market information in addition to regular technical features derived from EOD values enhances directional forecast efficacy. This will serve as a stepping stone for research in this avenue. Further research with different levels of off-market data (1-minute, 5-minute, 15-minute, etc.) has the potential of enhancing the efficacy of stock price forecasting.

Forecasts made using off-market data together with the regular EOD-based features are more accurate than the individual constituents. Such improvement is beneficial for traders as they can get more informed predictions since the input comprises of information that is closer to the opening of the market.

6.3. Future directions

In the future, experiments will be performed on a larger dataset comprising of more stocks both from the US market as well as other markets. The technical feature set as well as the predictor will be further modified and enhanced to improve the baseline performance. Information from other modalities will also be fused including news articles and interview clips for better parameterization of the stock prices. Our plans also encompass a deeper exploration of the pre and post-market data for generating a wider range of handcrafted features. Deep learning techniques will be used on both raw data and extracted features for possible improvement in the forecasts. Finally, this system will be deployed with live data feed to observe its performance and computational overhead in real-time scenario.

7. Conclusion

In this paper, a technique is proposed towards 1-day ahead directionality forecast of stock prices by leveraging pre and post market hour-based features. Exploration of off-market data for training stock price forecasting systems is in its nascent stages till date. The baseline results were engendered using 15 commonly used technical features that were computed using the end-of-day OHLCV values. The proposed technique used these technical features along with statistical features, derived from the post and pre-market data. Mean directional forecast accuracy improved for all the 11 sectors in the dataset. On average, the performance improved for at least 60% of the stocks across all the sectors. The proposed technique also outperformed the baseline in tests where the number stocks in the existing stock universe was gradually increased.

References

Ahmed, E. A. S., & Goyal, S. (2023). Impact of technical parameters for short-and long-term analysis of stock behavior. *Materials Today: Proceedings*, *80*, 1731–1736.

Atsalakis, G. S., & Valavanis, K. P. (2009). Surveying stock market forecasting techniques–part ii: Soft computing methods. *Expert Systems with applications*, *36*(3), 5932–5941.

Bathla, G., Rani, R., & Aggarwal, H. (2023). Stocks of year 2020: Prediction of high variations in stock prices using lstm. *Multimedia Tools and Applications*, *82*(7), 9727–9743.

Boyle, D., & Kalita, J. (2023). Spatiotemporal transformer for stock movement prediction. *arXiv preprint arXiv:2305.03835*.

Campisi, G., Muzzioli, S., & De Baets, B. (2023). A comparison of machine learning methods for predicting the direction of the us stock market on the basis of volatility indices. *International Journal of Forecasting*.

Chandola, D., Mehta, A., Singh, S., Tikkiwal, V. A., & Agrawal, H. (2023). Forecasting directional movement of stock prices using deep learning. *Annals of Data Science*, *10*(5), 1361–1378.

Chen, C., Yu, W., & Zivot, E. (2009). Predicting stock volatility using after-hours information. *Available at SSRN 1324991*.

Chen, S., & He, H. (2018). Stock prediction using convolutional neural network. *IOP Conference series: materials science and engineering*, *435*, 012026.

Chen, T., & Guestrin, C. (2016). Xgboost: A scalable tree boosting system. *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*, 785–794.

Chen, W., Jiang, M., Zhang, W.-G., & Chen, Z. (2021). A novel graph convolutional feature based convolutional neural network for stock trend prediction. *Information Sciences*, *556*, 67–94.

Dassanayake, W., Jayawardena, C., Ardekani, I., & Sharifzadeh, H. (2019). Models applied in stock market prediction: A literature survey.

Dezhkam, A., & Manzuri, M. T. (2023). Forecasting stock market for an efficient portfolio by combining xgboost and hilbert–huang transform. *Engineering Applications of Artificial Intelligence*, *118*, 105626.

Gandhmal, D. P., & Kumar, K. (2019). Systematic analysis and review of stock market prediction techniques. *Computer Science Review*, *34*, 100190.

Hoseinzade, E., & Haratizadeh, S. (2019). Cnnpred: Cnn-based stock market prediction using a diverse set of variables. *Expert Systems with Applications*, *129*, 273–285.

Khaidem, L., Saha, S., & Dey, S. R. (2016). Predicting the direction of stock market prices using random forest. *arXiv preprint arXiv:1605.00003*.

Koehler, P. A. (2009). *Predicting realized volatility: The effects of extended hours trading* (tech. rep.). Citeseer.

Kumar, G., Jain, S., & Singh, U. P. (2021). Stock market forecasting using computational intelligence: A survey. *Archives of computational methods in engineering*, *28*(3), 1069–1101.

Kumbure, M. M., Lohrmann, C., Luukka, P., & Porras, J. (2022). Machine learning techniques and data for stock market forecasting: A literature review. *Expert Systems with Applications*, *197*, 116659.

Lam, K., & Mok, P. Y. (2002). Stock price prediction using intraday and ahipmi data. *Proceedings of the 9th International Conference on Neural Information Processing, 2002. ICONIP'02.*, *5*, 2167–2171.

Langager, C. (2024). Pre-market and after-hours trading [URL: https://www.investopedia.com/ask/answers/06/preaftermarket.asp, Accessed: June 10, 2024].

Li, S., Liao, W., Chen, Y., & Yan, R. (2023). Pen: Prediction-explanation network to forecast stock price movement with better explainability. *Proceedings of the AAAI Conference on Artificial Intelligence*, *37*(4), 5187–5194.

Patel, J., Shah, S., Thakkar, P., & Kotecha, K. (2015). Predicting stock and stock price index movement using trend deterministic data preparation and machine learning techniques. *Expert systems with applications*, *42*(1), 259–268.

Shah, D., Isah, H., & Zulkernine, F. (2019). Stock market analysis: A review and taxonomy of prediction techniques. *International Journal of Financial Studies*, *7*(2), 26.

StockCharts. (2024). Technical indicators & overlays [URL: https://school.stockcharts.com/doku.php?id=technical_indicators, Accessed: June 10, 2024].

Thompson, C. (2023). Schaff trend cycle indicator: How it compares to the macd [URL: https://www.investopedia.com/articles/forex/10/schaff-trend-cycle-indicator.asp, Accessed: June 10, 2024].

Vuong, P. H., Dat, T. T., Mai, T. K., Uyen, P. H., et al. (2022). Stock-price forecasting based on xgboost and lstm. *Computer Systems Science & Engineering*, *40*(1).

Yearner, M. (2021). How to read financial data [URL: https://medium.com/geekculture/what-is-ohlcv-data-in-finance-5f53527cd5ce, Accessed: June 10, 2024].

Zhang, Q., Qin, C., Zhang, Y., Bao, F., Zhang, C., & Liu, P. (2022). Transformer-based attention network for stock movement prediction. *Expert Systems with Applications*, *202*, 117239.