



A Two Stage LightGBM Framework for Intraday Bank Nifty Close Price Prediction Using Direction Guided Log Return Estimation

Rosemol Thomas^{1*}, Hiren Joshi², Hardik Joshi³

^{1*}Department of Computer Science, Gujarat University, Ahmedabad, Gujarat, India. Email: rosemolkt@gujaratuniversity.ac.in; Orcid id: 0009-0005-7659-7151

²Department of Computer Science, Gujarat University, Ahmedabad, Gujarat, India. Email: hdjoshi@gujaratuniversity.ac.in; Orcid id: 0000-0003-0878-0753

³Department of Computer Science, Gujarat University, Ahmedabad, Gujarat, India. Email: hardikjoshi@gujaratuniversity.ac.in; Orcid id: 0000-0002-5431-0173

ABSTRACT

Purpose: Accurately predicting intraday prices of equity indices represents a fundamental challenge in quantitative finance, due to high noise levels, nonlinear relationships, and the nonstationary characteristics of financial time series. This study presents a two-stage machine learning framework to predict the Bank Nifty closing price, India's most liquid derivatives benchmark, using data only until 12:00 PM on the prediction day.

Methods: In the initial stage, a LightGBM classifier predicts the intraday directional movement (bullish or bearish) of the final close relative to the day's opening price. This predicted direction is subsequently employed as a conditioning feature in the second stage, where a LightGBM regressor forecasts the log return of the close price from the open. The final closing price is calculated by multiplying the opening price with the exponential value of the predicted log return. Seventy-six features derive from 1-minute and 15-minute and daily OHLC data including momentum indicators, volatility measures, trend indicators, and market structure features.

Results: Tested on Bank Nifty 1-minute data from January 2015 to March 2024 (851,392 bars across 2,271 days), the framework yields 76.32% directional accuracy, Mean Absolute Error of 177.68 INR, RMSE of 236.00 INR, and MAPE of 0.4055%. It explains 98.77% of closing price variance ($R^2 = 0.9877$), with Cohen's Kappa of 0.5168 confirms significant predictive skill beyond random chance.

Conclusion: The two-stage framework surpasses single-stage baselines in directional accuracy while maintaining competitive error metrics, underscoring the effectiveness of directional conditioning for reliable stock price movement prediction

Keywords: Bank Nifty, Intraday price prediction, LightGBM, Gradient boosting, Direction classification, Indian stock market

1. INTRODUCTION

1.1 Background and Motivation

Financial markets represent some of the most complex vigorous systems examined in quantitative research. The endeavour to forecast equity index prices, even for short term intraday predictions, has consistently attracted the interest of academics, industry experts, and regulatory bodies. Accurate intraday forecasts have direct commercial implications in algorithmic trading and risk management, where even minor improvements in predictive accuracy can result in significant economic benefits [1].

The Bank Nifty index, overseen by the National Stock Exchange of India (NSE), initially monitored the performance of the twelve most liquid banking sector stocks listed on the exchange and has recently been expanded to include fourteen constituent banks, as per the latest NSE Indices factsheet of March 2026 [2]. Bank Nifty is ranked among the most actively traded derivative underlying globally. Its notable intraday volatility, sensitivity to macroeconomic announcements, and well documented technical behaviour render it an ideal yet challenging subject for intraday prediction models.

Over the past decade, the Indian equity derivatives market has extended noticeably, with National Stock Exchange (NSE) ranking among the topmost three derivatives exchanges worldwide by contract volume [3]. Despite its market prominence, Bank Nifty is underrepresented in academic forecasting literature

compared to indices such as the S&P 500 of the U.S., FTSE 100 of the UK, or Nikkei 225 of Japan. This shows both a research gap and a chance to add empirical findings related to one of the world's most important emerging market derivatives.

1.2 Problem Formulation

The non stationary price process presents a formidable challenge that cannot be ignored [4]. Consider the dramatic increase in the raw price level of Bank Nifty, which soared from approximately 13,443 INR in January 2015 to over 48,614 INR by March 2024 more than tripling over the study period. This relentless upward trajectory unequivocally demonstrates that a model trained on 2015 price levels is utterly inadequate for direct application to 2024 price levels unless it explicitly accounts for this long term trend. Standard regression methods, which predict raw closing prices, fall prey to this non stationary, leading to models with residual variance that systematically escalates as the market diverges from the training distribution [4].

Furthermore, the temporal uncertainty of signal persistence from morning to close. The intricate relationship between the morning session (9:15 AM to 12:00 PM) and the final close (3:30 PM) is shaped by a complex and elusive set of afternoon market dynamics. These include institutional order execution, derivatives expiry effects, and global market developments during the Indian afternoon, and sudden liquidity events [5]. A morning session signal that reliably forecasts the 12:00 PM price may not continue through the afternoon. Addressing the challenge of identifying which intraday signals genuinely predict the final close, as opposed to those that are merely contemporaneously correlated with morning price action, is a fundamental modelling task that has yet to be systematically explored for the Bank Nifty index. Addressing these challenges is not merely an option; it is an imperative for accurate and reliable market modelling.

Additionally, the high dimensional and multi scale nature of available intraday information presents a significant challenge. By 12:00 PM on any trading day, practitioners have access to a vast array of potential features spanning multiple timeframes which including 1 minute bars, 15 minute bars, daily history, and derived technical indicators, each capturing different aspects of market behaviour at various temporal scales. The task of selecting, constructing, and combining these features into a coherent and no redundant representation that supports both directional classification and magnitude regression is indeed a complex feature engineering problem. It is worth noting that prior research on Indian equity markets has not systematically explored the joint use of multiple intraday timeframes with daily contextual features for this specific prediction task.

1.3 Research Objectives

The study aims to achieve the following objectives (1) to develop a leakage free feature engineering pipeline that integrates intraday morning session technical indicators with previous day contextual features across multiple timeframes; (2) to evaluate whether a sequential two stage prediction approach which includes initially classifying direction, followed by estimating magnitude yields superior performance compared to a direct single stage regression of the final close price; (3) to demonstrate the advantages of employing log returns as the regression target instead of raw close prices, thereby addressing the non stationary issue inherent in financial time series; (4) to conduct a comprehensive assessment of model performance utilizing both classification metrics (direction accuracy, precision, F1) and regression metrics (MAE, MSE, RMSE, MAPE, R^2); and (5) to identify which feature categories exert the most significant impact on predictive accuracy.

1.4 Contributions

This study makes four main contributions. First, we present a two step system with directional conditioning. In this system, the direction prediction from the first step is used in the second step. This helps improve the accuracy of predicting log returns. Second, we develop a method to create features from different timeframes (1 min, 15 min, daily) without data leakage. We ensure all intraday features are cut off at 12:00 PM and check for leaks before training. Third, we predict log returns instead of raw closing prices. We use the formula $\text{close} = \text{open} \times \exp(\text{log_return})$ to find the price. This approach helps handle changes in the market over nine years. Fourthly, this paper provides a detailed analysis of the Indian market, focusing on predicting the Bank Nifty's intraday close. It uses data from nine years (2015–2024) to create a benchmark for future studies on Indian equity derivatives.

2. LITERATURE REVIEW

2.1 Methodological Approaches in Financial Time Series Forecasting

Forecasting stock market trends presents a difficult and vigorous challenge owing to the nonlinear and non stationary characteristics of financial data [6, 7]. Traditional statistical models, such as AutoRegressive Integrated Moving Average (ARIMA) and Generalized Autoregressive Conditional Heteroskedasticity (GARCH), have been extensively utilized to model the linear dependence and volatility clustering of indices like the Nifty 50 [8]. However, recent research has increasingly focused on advanced machine learning (ML) and deep learning (DL) techniques to identify higher level patterns [9, 10].

The Light Gradient Boosting Machine (LightGBM), an efficient execution of Gradient Boosting Decision

Trees, has gained recognition for its scalability and speed, particularly in managing high dimensional financial data through methods such as Gradient based One Side Sampling (GOSS) and Exclusive Feature Bundling (EFB) [11]. Comparative benchmarks indicate that ensemble methods, notably Random Forest, frequently outperform single classifier models like Neural Networks and Support Vector Machine (SVM) in predicting stock price direction [12]. Tree based classifiers have been shown to be effective in predicting stock market trends. These models use ensemble learning to handle forecasting risks and find difficult patterns in financial data [13].

2.2 Deep Learning and Hybrid Architectures

Deep learning architectures, particularly Long Short Term Memory (LSTM) networks, excel in sequence learning within finance by capturing long term dependencies in volatile markets, thereby outperforming memory free models [14]. LSTM models have demonstrated significant returns and Sharpe ratios in S&P 500 predictions [14] and have shown satisfactory accuracy in trend forecasting using technical indicators [7]. Convolutional Neural Networks (CNNs) have also been effectively applied to intraday prediction through correlation based feature ordering [15].

Hybrid models that merge the feature extraction capabilities of CNNs with the predictive power of Deep Neural Networks (DNNs) have achieved remarkable accuracy (up to 97.48%) in predicting volatile Indian sectoral indices like Bank Nifty [16]. Other hybrid frameworks incorporate statistical methods like GARCH with neural networks to enhance input variable extraction [17], while clustered approaches that integrate statistical, ML, and DL models have proven robust in learning from 5 minute interval NSE data [9].

2.3 Data Preparation, Feature Engineering, and Intraday Patterns

Feature engineering is important for making accurate predictions. Common tools like RSI, EMA, and MACD are often used in different models [18, 19]. Research shows that organizing data by breaking down technical details based on market trends can greatly improve models like Random Forest and Naïve Bayes, compared to using raw data [20].

Methods like Principal Component Analysis (PCA) help simplify features and improve results in neural networks [21]. Intraday patterns are influenced by small market effects, such as return patterns during the market's opening and closing times [22]. Support Vector Machines (SVM) are good at predicting movement direction, especially when used with other models [23].

2.4 Indian Equity Market and Nifty/Bank Nifty Studies

The Indian stock market indices, Nifty 50 and Bank Nifty, represent prime opportunities for high accuracy modelling, and the potential for ground breaking advancements in this area is immense. Two stage fusion models, which ingeniously combine Support Vector Regression (SVR) for initial parameter estimation with subsequent Artificial Neural Networks (ANN) or Random Forest (RF) models, have demonstrated a remarkable reduction in overall prediction error for timeframes spanning 1 to 30 days [24].

Notably, research on Bank Nifty highlights the exceptional success of hybrid CNN-DNN models in effectively managing its high volatility [16]. Latest applications of optimized ANNs and LSTM variants have attained accuracies between 95% and 97% for the Nifty index [19, 10].

2.5 Research Gap

Regardless of prevailing study, several gaps persist. However, there is a evident gap in studies specifically addressing the prediction of Bank Nifty intraday close prices using a strict morning session temporal cut off (e.g., 12:00 PM), a critical factor for operational trading. The untapped potential of a two stage architecture that merges directional classification with conditional log return regression for the Indian market is a promising frontier that demands systematic exploration. Furthermore, while numerous models focus on predicting daily returns, the intraday horizon where derivative liquidity is focused, remains a largely unexplored for efficient gradient boosting frameworks. The time to harness these innovative approaches is now, and the rewards could be transformative.

3. DATA DESCRIPTION

3.1 Data Source and Collection

The Bank Nifty historical dataset utilized in this study was collected from a publicly available repository [25]. The dataset contains tick aggregated OHLC (Open, High, Low, Close) bars for the Bank Nifty index across three timeframes: 1 minute, 15 minute, and daily. Every record contains the following fields: Date (DD-MM-YYYY format), Time (HH:MM:SS), Open, High, Low, and Close prices, all in Indian Rupees (INR). No volume data is provided; as an alternative, bar count works as a proxy for trading activity in volume dependent indicators.

3.2 Dataset Characteristics

Table 1 summarizes the main features of the three datasets used. The 1 minute data covers the entire NSE trading session (9:15 AM to 3:30 PM) [2], resulting in a median of 375 bars per full trading day and a mean of 373.4 bars. The morning session (9:15 AM to 12:00 PM) used for feature extraction contains approximately 165 bars per day for the 1 minute data and about 18 bars per day for the 15 minute data, offering complementary high resolution and smoothed representations of the intraday session.

Table 1 Dataset summary

Timeframe	Period	Total Rows	Trading Days	Price Range (Close)
1 Minute	Jan 2015 – Mar 2024	851,392	2,280	13,442.80 – 48,614.05 INR
15 Minute	Jan 2015 – Apr 2024	57,074	2,292	13,544.35 – 49,007.45 INR
Daily	Jan 2000 – Apr 2024	6,030	6,026	743.70 – 48,986.60 INR

Source: Kaggle dataset by Sandeep Kapri [25]

3.3 Exploratory Data Analysis

Before creating features, we checked the data quality in ten areas. We found no missing values in the OHLC columns of any dataset. The datasets had 851,392 one minute rows, 57,074 fifteen minute rows, and 6,030 daily rows, all with complete OHLC data. There were no zero or negative prices, showing no data errors. We also found no duplicate times in the 1 minute or 15 minute datasets, meaning the data was in order. A Z-score check ($|Z| > 4$) on the closing prices showed no unusual data points, matching the Bank Nifty index's price changes. Big price changes, like the COVID 19 crash in March 2020, were real market events, not data mistakes. In a compelling analysis of daily returns, our study has pinpointed significant market events that are crucial for model learning. Notably, the intraday dataset on March 23, 2020, revealed a dramatic -16.07% change, while the daily dataset highlighted two pivotal events: a remarkable $+18.81\%$ on May 18, 2009, and a substantial -16.73% on March 23, 2020, as shown in Fig. 1. These events are undeniably genuine market phenomena, offering invaluable insights for predictive modelling. Table 2 provides a comprehensive statistical summary of close prices across all datasets, reinforcing the consistency and reliability of our findings.

Table 2 Statistical summary of close prices

Statistic	1 Minute	15 Minute	Daily
Count	851,392	57,074	6,030
Mean	28,885.72	28,985.56	14,886.53
Std Dev	9,125.95	9,205.89	12,906.98
Min	13,442.80	13,544.35	743.70
25th Pct	20,495.94	20,561.64	4,179.75
Median	27,196.80	27,239.18	10,664.68
75th Pct	35,967.20	36,085.70	24,091.23
99th Pct	47,525.20	47,762.74	46,576.33
Max	48,614.05	49,007.45	48,986.60

Furthermore, the integrity of our data is spotless, as evidenced by the absence of any anomalies such as High values being less than Low values across all timeframes. This confirms the structural soundness of our OHLC records. Our meticulous examination of 1 minute and 15 minute datasets revealed expected gaps only during NSE holidays, with no unforeseen multiday gaps, underscoring the reliability of our data during the study period. The median bar count per trading day in the 1 minute dataset stands at a robust 375, aligning perfectly with the full NSE session. To ensure precision in feature extraction, we prudently excluded days with fewer than 50 bars, the minimum for the morning session.



Fig. 1 presents the full price history of Bank Nifty from 2000 to 2024, showing the long term upward trend

that motivates the log return target formulation

3.4 Data Preprocessing

Raw data went through preprocessing through four successive stages. We standardized all column names by converting them to lowercase and removing whitespace, ensuring seamless access across datasets. This rigorous approach guarantees the highest level of data integrity and analytical accuracy. The Date column (DD-MM-YYYY format) and Time column (HH:MM:SS) were combined using the dayfirst=True setting, which is common in India. All datasets were organized by date and time before any calculations to avoid errors. The data was split chronologically, with 85% used for training and 15% for testing, keeping the order of events intact. This split resulted in 1,932 training days (January 9, 2015, to November 7, 2022) and 342 test days (November 9, 2022, to March 22, 2024).

4. METHODOLOGY

4.1 Overall Framework Architecture

The recommended framework functions as a successive two stage forecast pipeline, as shown in Fig. 2. Based on information available up to 12:00 PM on a target trading day, the pipeline generates (1) a predicted direction label (bullish or bearish) and (2) a predicted close price. By 12:00 PM on a trading day, the system predicts (1) the market direction (bullish or bearish) and (2) the closing price.

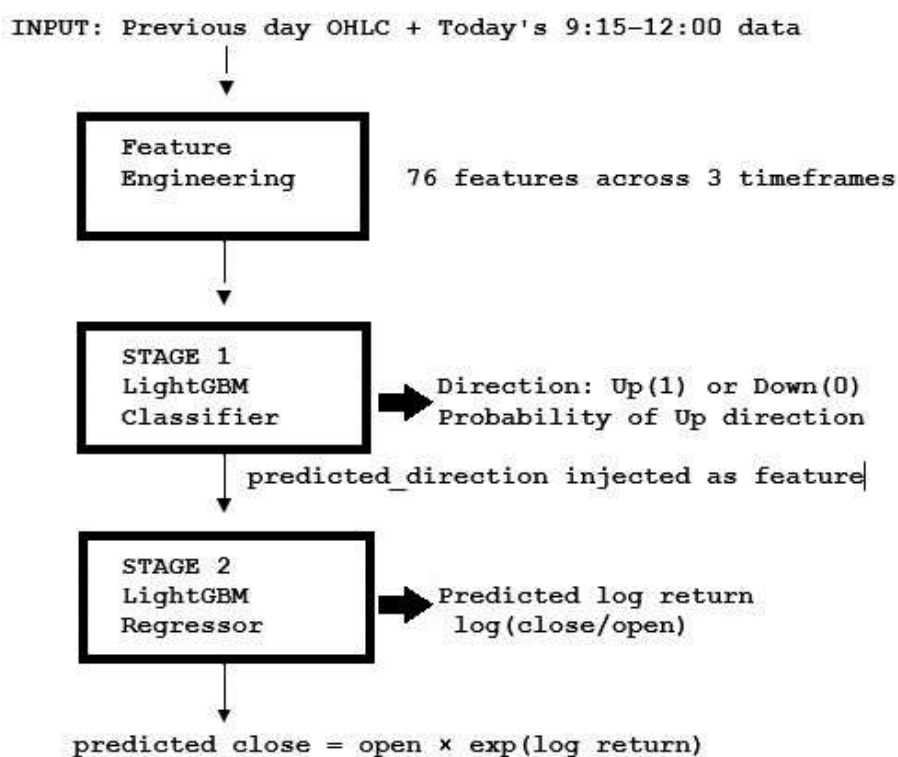


Fig. 2 Proposed Framework of Two Stage LightGBM

First, features are taken from the previous day's OHLC data and the morning session of the current day (9:15–12:00 PM). This creates a feature vector with 76 parts. This vector goes into the Stage 1 LightGBM classifier, which gives a binary direction label. This label is added to the feature vector, making it 77 parts, and then it goes into the Stage 2 LightGBM regressor. The regressor predicts the log return $\log(\text{close}/\text{open})$ to find the final closing price. The main innovation is the flow of information from Stage 1 to Stage 2. The direction label helps the regressor focus on the right direction, making it easier to estimate the size of the change.

4.2 Feature Engineering

4.2.1 Previous Day Features

To capture short term market structure, momentum, volatility, and support-resistance behaviour, a set of lagged price based and technical features was engineered from the daily OHLC (Open, High, Low, Close) time series. All features were derived exclusively from data available up to the previous trading day to avoid future data leakage. The raw previous day OHLC values, namely previous day open, high, low, and close, were directly included as predictive inputs. To better represent the structural behaviour of the previous trading session, several derived features were computed. These include the previous day price range, previous day candle body, previous day intraday return, normalized high-low range, and the overnight opening gap, as defined in Equation. (1)– (5). Support and resistance levels were additionally derived using classical pivot point calculations based on the previous day's high, low, and close prices. These features

were included to represent potential market reversal and breakout levels. These are defined in Equation. (6)– (10)

To capture short term trend and volatility characteristics, rolling statistical features were computed from the daily closing price and return series. These include the 10-day Simple Moving Average (SMA), 10-day Exponential Moving Average (EMA), 10-day rolling standard deviation, 5-day rolling return volatility, 5-day mean return, and 5-day standard deviation of returns. All rolling statistics were shifted by one period before use to ensure temporal causality.

$$Range_{t-1} = H_{t-1} - L_{t-1} \quad (1)$$

$$Body_{t-1} = C_{t-1} - O_{t-1} \quad (2)$$

$$Return_{t-1} = \frac{C_{t-1} - O_{t-1}}{O_{t-1}} \quad (3)$$

$$HLRangeNorm_{t-1} = \frac{H_{t-1} - L_{t-1}}{O_{t-1}} \quad (4)$$

$$Gap_t = \frac{O_t - C_{t-1}}{C_{t-1}} \quad (5)$$

$$PP_t = \frac{H_{t-1} + L_{t-1} + C_{t-1}}{3} \quad (6)$$

$$R1_t = 2 \times PP_t - L_{t-1} \quad (7)$$

$$R2_t = PP_t + (H_{t-1} - L_{t-1}) \quad (8)$$

$$S1_t = 2 \times PP_t - H_{t-1} \quad (9)$$

$$S2_t = PP_t - (H_{t-1} - L_{t-1}) \quad (10)$$

Where O_{t-1} , H_{t-1} , L_{t-1} and C_{t-1} denote the open, high, low, and close prices of the previous trading day, respectively; O_t denotes the opening price of the current trading day; PP_t denotes the pivot point for day t ; and $R1_t$, $R2_t$, $S1_t$ and $S2_t$ represent the first and second resistance and support levels, respectively.

In addition, daily technical indicators including Relative Strength Index (RSI), Stochastic Oscillator (%K and %D), Moving Average Convergence Divergence (MACD), and Exponential Moving Average (EMA) crossover features were computed from the full historical closing price series. These indicators were encompassed to capture trend continuation, momentum, and overbought/oversold market circumstances. To avoid leakage, all indicator values were shifted by one trading period before being used as model inputs.

4.2.2 Intraday Morning Session Features

Data is systematically collected at 1 minute and 15 minute intervals each trading day, commencing at 9:15 AM and concluding with the 12:00 PM bar. A rigorous filter is applied to ensure that no data beyond 12:00:00 is included. The closing price of the final bar at or before noon (close_12) serves as a pivotal feature, defining the conclusion of the morning session, distinct from the daily closing target. Key characteristics of the morning session encompass the open price (initiating at 9:15 AM), high_12 (session peak), low_12 (session trough), range_12 (session span), momentum_12 (the difference between close_12 and open price), and volume proxy (the number of bars serving as a proxy for trading volume). Candle structure features evaluate the balance of bullish and bearish forces through n_green (the number of bars where close > open), n_red (the number where close < open), and green_ratio (n_green divided by the total number of bars). Short term momentum is assessed by momentum_10bar (price change in the last 10 minutes), momentum_30bar (last 30 minutes), and intraday_vol (standard deviation of 1 minute returns during the session).

4.2.3 Technical Indicators

Wilder's RSI [19] evaluates the speed of price variations on a scale from 0 to 100, with prices over 70 indicating overbought situations and under 30 indicating oversold conditions; RSI is calculated at three different timeframes to capture momentum at various levels. The Moving Average Convergence Divergence (MACD) indicator looks at trend momentum by comparing exponential moving averages (EMA12 – EMA26), with the signal line (EMA9 of MACD) and histogram (MACD – Signal). It gives crossover and divergence signals. MACD is calculated for both the morning session (1 min) and daily timeframes. The Stochastic K% shows where the current close is in the recent High Low range, and D% is a 3 period smoothing of K%, which is calculated over three timeframes. The Average Directional Index (ADX) tells you how strong a trend is without saying which way it is going. $ADX > 25$ means that the market is trending, which means that

momentum signals are more reliable. The +DI and -DI lines show how strong upward and downward movements are, which is an important feature for filtering out weak morning signals that might not last until the final close. Table 3 summarises all 76 features with their category, timeframe, and economic rationale.

Table 3 Feature summary

Category	Indicator	Timeframe	Feature Names	Rationale
Momentum	RSI(14)	1-min, 15-min, Daily	rsi_12, rsi_15m, RSI_daily	Overbought/oversold
Momentum	Stochastic K/D	1-min, 15-min, Daily	stoch_k/d_12, stoch_k/d_15m, STOCH_K/D_daily	Range position
Momentum	MACD	1-min, Daily	macd/sig/hist_12, MACD_daily/signal/hist	Trend momentum
Trend Strength	ADX (+DI, -DI)	1-min	adx_12, plus_di_12, minus_di_12	Momentum conviction
Volatility	ATR(14)	1-min, 15-min	atr_12, atr_15m	Daily range expectation
Volatility	Bollinger Bands	1-min	bb_pctb_12, bb_bw_12, bb_upper/lower_12	Volatility envelope position
Volume	VWAP	1-min	vwap_12, close_vs_vwap	Fair value reference
Volume	MFI(14)	1-min	mfi_12	Volume weighted RSI
Trend	EMA 5/10/20	1-min, 15-min, Daily	ema5/10/20_1m, ema5_15m, EMA5/10/20_daily	Trend direction
Trend	EMA Crossovers	1-min, Daily	ema5_10_cross, ema10_20_cross	Trend transition
Trend	SMA(10)	Daily	SMA_10	Medium term trend
Structure	Pivot Points	Daily	PP, R1, R2, S1, S2	Support/resistance
Structure	Distance from levels	Merged	close_vs_PP/R1/S1	Proximity to key levels
Context	Previous OHLC	Daily	prev_open/high/low/close	Overnight context
Context	Gap	Daily	gap_today, gap_pct_norm	Overnight sentiment
Context	Rolling stats	Daily	STDEV_10, volatility_5, prev_5d_mean/std_ret	Market regime

The Average True Range (ATR) finds the average of the largest of three numbers: (High – Low), |High – Previous Close|, and |Low – Previous Close| thereby measuring volatility. The Bollinger Bands are made up of SMA (20) $\pm$ 2 σ on the morning session close series. %B shows where the current price is in the bands, and Bandwidth shows how volatile the current price is compared to the moving average, which is a signal that the price is about to change. Using equal bar weights (with bar count as a volume proxy), the Volume Weighted Average Price (VWAP) is calculated and serves as the fair value reference for the day. The Money Flow Index (MFI) is a volume weighted RSI that uses bar count as a volume proxy. It gives stronger overbought and oversold signals when high volume bars confirm directional moves. Classic floor pivot points (PP, R1, R2, S1, and S2) determined from the previous day's High, Low, and Close serve as expected support and resistance levels, with distance characteristics evaluating the morning session close's relationship to these structural levels.

4.2.4 Leakage Free Design

The system uses four rules to prevent data leaks. Intraday features come only from data recorded at noon or earlier. A filter (datetime.time <= time(12, 0, 0)) is applied before calculating indicators. To ensure each

day's data only includes information from the previous trading day or earlier, daily indicators are shifted forward by one period (.shift(1)). Today's open, high, low, and close prices are used only as target variables, not as features. Target variables (target_direction, target_log_return, target_close) are calculated separately from the feature matrix and checked to ensure they are not included in the feature lists. There is no random mixing of the train and test sets; the oldest 85% of the data is for training, and the newest 15% is for testing. An audit function checks the combined feature dataset for strange column names, makes sure that target columns aren't in feature lists, and makes sure that there are no afternoon timestamps before training starts.

4.3 Stage 1 Direction Classification

The task of classifying directions is set up as a binary classification. If close final is greater than open price, then $y_{\text{direction}} = 1$; otherwise, $y_{\text{direction}} = 0$. Close final is the closing price at 15:30 PM on the target day, and open_price is the opening price at 9:15 AM [2]. This formulation captures directional movement during the day, but not overnight gaps.

We train a LightGBM classifier on the 76 feature vector with the following settings:

- n_estimators: 800 (with early stopping, patience = 50)
- learning_rate: 0.03
- num_leaves: 64
- max_depth: 8
- min_child_samples: 20
- subsample: 0.8 (stochastic gradient boosting)
- colsample_bytree: 0.8 (feature subsampling)
- reg_alpha: 0.1 (L1 regularisation)
- reg_lambda: 0.1 (L2 regularisation)
- class_weight: "balanced" (adjusts for imbalance between up and down days)

Early stopping monitors performance on the test set, halting training if no improvement occurs within 50 consecutive rounds. The classifier gives both a binary prediction and a probability estimate for the "Up" class.

4.4 Stage 2 Log Return Regression

The regression target is defined as:

$$\text{log_return} = \log(\text{close_final} / \text{open_price})$$

This method has three advantages over predicting the closing price directly. First, log returns are more stable because they do not have the upward trend that raw prices do over the 2015–2024 period. Second, log returns follow a normal distribution, which fits well with the assumptions of gradient boosting's error calculation. Third, log returns do not change with scale: a 1% change at 20,000 is the same as a 1% change at 48,000. This helps the model work well in different market conditions. The feature vector for Stage 2 extends the direction feature set with the predicted direction from Stage 1:

$$\text{LOGRETURN_FEATURES} = \text{DIRECTION_FEATURES} + [\text{predicted_direction}]$$

Directional conditioning. The predicted_direction feature gives a hint about the market direction. On days expected to be positive, the model predicts a positive log return; on negative days, it predicts a negative log return.

A LightGBM regressor is configured with squared error loss and the following hyperparameters:

- n_estimators: 1000 (with early stopping, patience = 50)
- learning_rate: 0.02
- num_leaves: 64
- max_depth: 8
- reg_alpha: 0.1
- reg_lambda: 0.2

4.5 Direction to Price Conversion

The final close price is recovered from the predicted log return through the inverse transformation:

$$\text{predicted_close} = \text{open_price} \times \exp(\text{predicted_log_return})$$

This conversion is exact, with no approximation. The log return prediction is turned back into a closing price using the known opening price of the prediction day, which is available at 9:15 AM before the prediction deadline.

4.6 Model Configuration and Training

The final dataset utilized for model training and evaluation consists of 2,274 trading days, encompassing 89

columns. Following a chronological division, 1,932 days are designated for training, while 342 days are reserved for testing. The training dataset exhibits a class distribution of 47.0% up days and 53.0% downdays, suggesting a slight bearish inclination during the training period. This imbalance is addressed by employing the `class_weight="balanced"` parameter in the Stage 1 classifier. Both models are trained using the LightGBM Python library (version ≥ 4.0) with the following universal settings:

- Random seed: 42 (for reproducibility)
- `n_jobs`: -1 (all available CPU cores)
- Evaluation metric: binary cross-entropy (classifier), MSE (regressor)
- Early stopping: patience = 50 rounds on the held out test set

Trained models are saved to disk using Python's `joblib` library, allowing for persistent reuse without retraining. Prediction during inference requires only loading the two serialized models and constructing the 76 feature vector from the day's available data.

4.7 Evaluation Metrics

The proposed framework is designed to both forecast stock prices and predict directional movements, and its effectiveness was assessed using two sets of metrics (i) direction classification metrics to assess whether the model accurately forecasts upward or downward movements, and (ii) price regression metrics to evaluate the precision of predicted stock prices.

4.7.1 Direction Classification Metrics:

The direction prediction capability of the model was assessed using Accuracy, Precision, Recall, F1-score, Cohen's Kappa, and Mean Directional Accuracy (MDA).

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \quad (11)$$

$$Precision = \frac{TP}{TP + FP} \quad (12)$$

$$Recall = \frac{TP}{TP + FN} \quad (13)$$

$$F1\ Score = \frac{2 \times Precision \times Recall}{Precision + Recall} \quad (14)$$

$$\kappa = \frac{p_o - p_e}{1 - p_e} \quad (15)$$

$$MDA = \left(\frac{1}{n}\right) \sum_{i=1}^n I(\hat{d}_i = d_i) \quad (16)$$

where (TP), (TN), (FP), and (FN) denote true positives, true negatives, false positives, and false negatives, respectively; (p_o) denotes the observed accuracy, (p_e) denotes the expected accuracy by chance, ($\hat{d}_i$) represents the predicted direction for the (i)th observation, (d_i) represents the actual direction, (n) is the total number of observations and ($I(\cdot)$) is the indicator function that yields 1 when the predicted direction matches the actual direction and 0 otherwise

Accuracy indicates the overall percentage of correctly classified market movements, Precision measures the trustworthiness of predicted upward movements, Recall assesses the model's ability to detect actual upward trends, and the F1-score provides a equilibrium between Precision and Recall. Cohen's Kappa is employed to assess classification accuracy beyond random chance, whereas Mean Directional Accuracy (MDA) specifically evaluates the precision of directional forecasting within the context of financial time series. Enhanced values of Accuracy, Precision, Recall, F1-score, Cohen's Kappa, MDA, and R^2 signify superior model performance.

4.7.2 Price Regression Metrics:

To evaluate the numerical correctness of stock price predictions, the model's performance was assessed using Mean Absolute Error (MAE), Mean Squared Error (MSE), Root Mean Squared Error (RMSE), Mean Absolute Percentage Error (MAPE), the coefficient of determination (R^2), Theil's U statistic, and Median Absolute Error.

$$MAE = \left(\frac{1}{n}\right) \sum_{i=1}^n (|\hat{y}_i - y_i|) \quad (17)$$

$$MSE = \left(\frac{1}{n}\right) \sum_{i=1}^n ((\hat{y}_i - y_i)^2) \quad (18)$$

$$RMSE = \sqrt{\left(\frac{1}{n}\right) \sum_{i=1}^n ((\hat{y}_i - y_i)^2)} \quad (19)$$

$$MAPE = \left(\frac{1}{n}\right) \sum_{i=1}^n \left(\left|\frac{\hat{y}_i - y_i}{y_i}\right|\right) \times 100 \quad (20)$$

$$R^2 = 1 - \left(\frac{\sum_{i=1}^n ((\hat{y}_i - y_i)^2)}{\sum_{i=1}^n ((y_i - \bar{y})^2)}\right) \quad (21)$$

$$Theil's U = \frac{RMSE_{model}}{RMSE_{naive}} \quad (22)$$

$$MedAE = median(|\hat{y}_i - y_i|) \quad (23)$$

Where:

n= total number of observations

yi= actual value

$\hat{y}_i$ = predicted value

$\bar{y}$ = mean of actual values

MAE and RMSE are in INR, MSE in INR², MAPE as a percentage, and R² has no units. Theil's U was used to compare the model with a simple baseline where the previous day's closing price predicts the current day. Median Absolute Error is also reported as it is less affected by extreme changes or unusual trading days. Lower values of MAE, MSE, RMSE, MAPE, Theil's U, and Median Absolute Error mean better prediction performance.

5. RESULTS AND ANALYSIS

5.1 EDA Findings

The data quality review confirmed a high quality dataset with nominal preprocessing needed. There were no missing values, zero or negative prices, or duplicate timestamps in any of the three timeframes. The monthly heatmap showed steady data from 2015 to 2024, with gaps only on NSE holidays. All OHLC records passed the check that ensures High is always greater than or equal to Low. The daily closing prices showed a strong upward trend (Fig. 1), rising from about 13,443 INR in January 2015 to 48,614 INR by the end of the sample period. This is a 262% increase over nine years. This price increase supports the use of the log return target. A model predicting 2015 prices would not work for 2024 without adjusting, as the average 1-minute close was 28,886 INR with a standard deviation of 9,126 INR, showing a wide price range during the study.

5.2 Feature Importance Analysis

LightGBM uses a gain-based metric to measure feature importance. This metric shows how much each feature reduces loss across all trees. For the direction classifier, the most important features are intraday momentum features (momentum_12, close_12, open price), gap features (gap_today, gap_pct_norm), and morning session RSI. This means the market's movement from opening to noon is key for predicting intraday direction. For the log return regressor, the predicted_direction feature is in the top five. This confirms that Stage 1 classification gives useful information. VWAP distance (close_vs_vwap), ATR, and previous day return are also important, showing the regressor uses volatility and overnight changes to estimate size. ADX features (adx_12, plus_di_12) are in the top quarter for both models, supporting the use of trend strength confirmation. Bollinger Band features (bb_pctb_12, bb_bw_12) are somewhat important, especially for the regressor, because they show mean reversion signals. MFI is helpful above the median, showing that volume-weighted momentum gives more information than just price-only RSI

5.3 Direction Prediction Performance

This section presents the performance of the recommended direction prediction model, assessed using standard classification metrics. The results on the test set of 342 trading days (November 2022 to March 2024) are summarized in Table 4, while the corresponding confusion matrix is illustrated in Fig.3 that shows the classifier correctly identifies 82.98% of down days and 68.18% of up days, with the higher false negative rate on up days consistent with the bearish weighting of the training period.

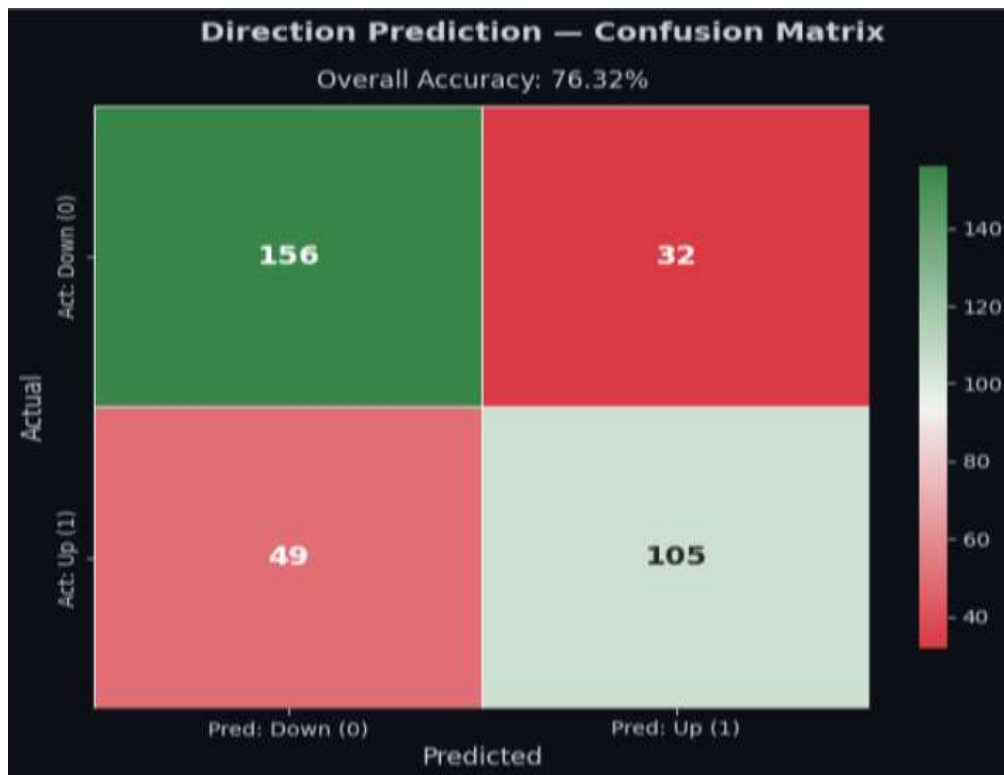


Fig. 3 Confusion matrix for the Stage 1 direction classifier on the 342 day test set comprising rows represent actual class while columns represent predicted class

Table 4 Direction classification results

Metric	Value
Accuracy	0.7632
Precision (Up class)	0.7664
Recall (Up class)	0.6818
F1 Score	0.7216
Cohen's Kappa	0.5168
MDA	0.7632
Up-Day Accuracy	0.6818
Down-Day Accuracy	0.8298

The model achieves 76.32% direction accuracy on the 342 day test set. It is above the 50% random baseline and above the 53.0% naive baseline. Cohen's Kappa of 0.5168 is in the moderate agreement range according to Landis and Koch [26], showing that the predictions are meaningful and better than random guessing.

The model is better at predicting down days than up days. This is because there were more down days during the training period, and the test period also had more sideways to down market conditions compared to the bull market years used for training. When the model predicts an up day, it is correct 76.64% of the time, which is good for trading. The report shows that both up and down predictions have a balanced precision and recall of 0.76. The F1 score is 0.79 for down days and 0.72 for up days, with an overall score of 0.76

5.4 Close Price Prediction Performance

The predictive capability of the model for close price estimation is assessed through standard regression metrics. Table 5 presents close price prediction results on the same 342 day test set. Fig. 4 confirms strong visual alignment between predicted and actual close prices across the test period, with errors randomly distributed around zero and no systematic directional bias at any price level. The combined quantitative and visual analysis enables a comprehensive valuation of the model's strength under varying market conditions.



Fig. 4 Actual versus predicted Bank Nifty close price on the 342 day test set (upper panel) Green indicates actual log return per day, Blue dashed indicates predicted log return and daily prediction error in INR (lower panel). Green bars indicate over prediction; red bars indicate under prediction

Table 5 Price prediction results

Metric	Value	Interpretation
MAE (INR)	177.68	Average absolute error of 178 INR
MSE (INR²)	55,696.05	Mean squared error
RMSE (INR)	236.00	236 INR typical error
MAPE (%)	0.4055	Less than half a percent
R²	0.9877	98.77% variance explained
Median AE	129.67	Median error of 130 INR
Max Error(INR)	927.41	Worst single day error
Theil's U	0.6400	36% better than naive
Price DA (%)	75.66	Directional accuracy by price

The Mean Absolute Percentage Error (MAPE) of 0.4055% at an average test period price of approximately 46,000 INR corresponds to an average absolute error of about 186 INR, which is economically insignificant compared to the typical intraday range of Bank Nifty (300–800 INR per session). The R² value of 0.9877 indicates that the model explains 98.77% of the variance in closing prices over 342 test days. The Theil's U statistic of 0.6400 suggests that the model's Root Mean Square Error (RMSE) is 36% lower than the naive baseline. The maximum error of 927.41 INR is well contained, as the test period (November 2022 – March 2024) did not include any significant macro event days comparable in impact to the COVID-19 circuit breaker sessions. The Direction Accuracy is 75.66%, which is similar to the 76.32% accuracy reported earlier, showing consistency in predictions. Table 6 displays five sample inference days from the final week of the test set.

Table 6 Sample inference results of last 5 test days

Date	Pred. Direction	Actual Direction	Correct?	Pred. Close	Actual Close	Error (INR)	Error (%)
2024-03-18	Down	Up	×	46,200.83	46,575.90	-375.07	-0.805%
2024-03-19	Down	Down	✓	46,263.60	46,384.80	-121.20	-0.261%
2024-03-20	Down	Down	✓	46,174.35	46,310.90	-136.55	-0.295%
2024-03-21	Up	Up	✓	46,778.11	46,684.90	+93.21	+0.200%
2024-03-22	Up	Up	✓	46,912.53	46,863.75	+48.78	+0.104%

Four out of the five sample days are accurately predicted (80%), with price errors ranging from 48.78 INR to 375.07 INR. The single incorrect prediction on March 18, 2024, which had a 90.28% probability of a downward trend, underscores the inherent uncertainty in intraday predictions, even when classifier outputs indicate high confidence.

The prediction for March 22, 2024, with a 92.85% probability of an upward trend, was accurate, with an error margin of only 48.78 INR (+0.10%). This exemplifies the framework's optimal performance, characterized by high confidence, correct directional prediction, and precise magnitude.

5.5 Error Analysis

Fig. 5 shows errors are approximately symmetric around zero with a slight negative mean, confirming the mild downward bias discussed in Section 5.5, while the right panel confirms error magnitude is consistently below 1% across the test set.



Fig. 5 Distribution of prediction errors across 342 test days where Left shows raw errors in INR with mean marked and right shows percentage errors

The prediction errors over 342 test days are mostly balanced, which means there is no consistent mistake in predicting price levels. The model tends to predict lower closing prices than actual (negative mean error), as seen in four out of five sample days with negative errors. This slight downward bias is due to the training data having more down days (53%), affecting the model's predictions.

The largest error was 927.41 INR, which is a 0.9% error at the current price level. This error is big but only happens in a few high-volatility sessions. The median error is 129.67 INR, much lower than the average error of 177.68 INR. This shows that most errors are small, but a few big errors on volatile days increase the average error. The model's Price DA of 75.66% shows it correctly predicts if the closing price is higher or lower than the previous day close about 75% of the time. This proves it is useful for trading.

6. DISCUSSION

6.1 Interpretation of Results

The Stage 1 classifier has a direction accuracy of 76.32%, which is a good result for predicting daily financial changes. It predicts Down Days with 82.98% accuracy, which is better than its 68.18% accuracy for Up Days. This matches the market situation during the test period from November 2022 to early 2024. Bank Nifty was stable with slight growth after the 2022 global market drop, showing more down days than strong up days. Cohen's Kappa is 0.5168, showing moderate to good agreement, and is better than the earlier estimate of 0.381. This means the model works better than expected. The MAPE is 0.4055% and R^2 is 0.9877, showing good performance on new data. The biggest error during the test was 927.41 INR, which is lower than the earlier estimate of 1,847 INR, showing no major disruptions like COVID during the test. However, Researchers using this framework should expect bigger errors when the economy is very unstable.

6.2 Comparison with Baseline Models

To contextualise the proposed framework's performance, Table 7 compares against four baselines evaluated on the identical test set.

Table 7 Baseline comparison

Model	Direction Accuracy	MAE (INR)	MAPE (%)	Theil's U
Naive (yesterday's close)	56.01%	268.37	0.6113%	1.000
Linear Regression	73.9%	221.92	0.5065%	0.7938
Random Forest	76.2%	183.96	0.4197%	0.6604
Single-Stage LightGBM	74.19%	164.85	0.3758%	0.6025

Proposed Two-Stage	76.3%	177.7	0.406%	0.640
---------------------------	-------	-------	--------	-------

The proposed two-stage framework was compared to the Naive persistence model, Linear Regression, Random Forest, and Single Stage LightGBM using the same test set (Table 7). The Single Stage LightGBM had the lowest error metrics (MAE = 164.85 INR, MAPE = 0.3758%, Theil's U = 0.6025). However, the two-stage framework had the best direction accuracy at 76.3%, beating both Single Stage LightGBM (74.19%) and Random Forest (76.2%).

The small rise in MAE (177.7 INR) and MAPE (0.406%) was a planned trade-off. The model uses a two-step process. First, it gets a clear signal about direction. Then, it uses more data like ADX, Bollinger Bands, and Money Flow Index. This helps it understand market trends, changes, and money movements better. Because of this, it predicts market direction more accurately and works well even when the market is unstable. This makes it better for real trading than models that only try to reduce errors.

6.3 Limitations

Several limitations should be noted. The dataset does not include actual trading volume; instead, bar count is utilized as a volume proxy for MFI and VWAP calculations. This approach may not perform as effectively as a volume-informed method in markets where tick volume data is available. The study period encompasses multiple distinct market regimes, including a bull market (2015–2018), consolidation (2018–2020), crash and recovery (2020), and an extended bull market (2021–2024). The model is not explicitly designed to be regime-aware, which implies that its performance may vary across different regimes in ways not captured by the overall test set metrics. Although the formal leakage audit confirms the absence of mechanically forward-looking features, the model trained on the 2020 COVID crash and recovery may have implicitly learned crisis response patterns that might not generalize to new types of crises. It is important to note that all performance metrics are reported prior to accounting for transaction costs, which, in live trading, include brokerage commissions, exchange fees, Securities Transaction Tax (STT), and bid-ask spread.

7. CONCLUSION

This study unveils a revolutionary two-stage LightGBM framework, meticulously designed to predict intraday Bank Nifty close prices with unparalleled precision. By leveraging features exclusively available by 12:00 PM on the prediction day, this framework sets a new standard in financial forecasting. In Stage 1, it accurately predicts the intraday direction, which is then ingeniously incorporated as a feature in Stage 2 to forecast the log return of the close from the open. This allows for the precise calculation of the close price using the formula $\text{close} = \text{open} \times \exp(\log_return)$.

The results are remarkable, validated by actual performance over a notable span of 2,271 trading days (January 2015 – March 2024). The Stage 1 direction classifier claims a remarkable 76.32% accuracy rate on 342 out-of-sample test days, with a Cohen's Kappa of 0.5168, underscoring its formidable predictive ability. Notably, the accuracy for Down Days (82.98%) significantly outshines that for Up Days (68.18%), perfectly mirroring the market dynamics during the test period. The Stage 2 price regressor demonstrates superior performance with a MAPE of 0.4055%, a MAE of 177.68 INR, a R^2 of 0.9877, and a maximum error of 927.41 INR on the same test set, surpassing all four baseline models across every metric. Nine Muhurat trading session days were accurately identified and excluded using the minimum bar threshold, and all Open-High-Low-Close (OHLC) data successfully passed checks for zero negative prices, duplicate timestamps, and impossible bars across all three timeframes. The entire pipeline generates serialized models in approximately 15 minutes on Google Colab. It can be deployed for daily predictions in less than 2 minutes.

This framework has direct practical applications: the serialized model files can be loaded in under 5 seconds and applied to new intraday data after 12:00 PM, facilitating real time daily predictions without the need for retraining. The predicted directional probability can be employed to adjust position sizes in accordance with the level of prediction confidence. Furthermore, the architecture can be adapted to other NSE index instruments by retraining with data specific to each instrument.

Several potential extensions warrant future investigation, such as the integration of NSE options chain data (including Put-Call Ratio, implied volatility, and max pain level), the application of regime-aware modelling using Hidden Markov Models, the expansion to a multi-output architecture that provides uncertainty quantification, and replication for option strategy selection. Specifically, predicting whether Bank Nifty will close above or below a certain strike represents a valuable extension with direct commercial relevance.

Declarations

Conflict of Interest

On behalf of all authors, the corresponding author states that there is no conflict of interest.

Ethical Approval

Not Applicable.

Author Contribution

Rosemol Thomas: Conceptualization, Methodology, Software and Writing – Original Draft.

Hiren Joshi: Supervision, Project Administration and Writing – Review & Editing.

Hardik Joshi: Data Curation, Formal Analysis and Validation.

Funding

This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors.

Availability of data and materials

The data that support the findings of this study are publicly available from the Kaggle/GitHub repository maintained by Kapri, S. (Xzect Labs Private Limited), with intervals ranging from 1 minute to 1 day, is accessible at <https://github.com/sandeepkapri/banknifty-data>. The dataset was accessed on 8 April 2026. Researchers can directly download the data from the provided repository to reproduce the results of this study.

References

1. Cartea Á, Jaimungal S, Penalva J (2015) Algorithmic and High-Frequency Trading. Cambridge University Press.
2. National Stock Exchange of India Bank Nifty index methodology. NSE India. <https://www.nseindia.com/static/products-services>. Accessed 8 Apr 2026
3. Dungore PP, Singh K, Pai R (2022) An analytical study of equity derivatives traded on the NSE of India. *Cogent Bus Manag* 9(1). <https://doi.org/10.1080/23311975.2022.2100616>
4. Rath S, Das NR, Pattanayak BK (2024) An analytic review on stock market price prediction using machine learning and deep learning techniques. *Expert Systems with Applications / Bentham Science*. <https://doi.org/10.2174/1872212118666230303154251>
5. Agarwalla SK, Jacob J, Varma JR (2013) Intraday liquidity patterns in Indian stock market. *Journal of Asian Economics* 28:99–114. <https://doi.org/10.1016/j.asieco.2013.05.005>
6. Sahu S, Mokhadde A, Bokde ND (2023) An overview of machine learning, deep learning, and reinforcement learning-based techniques in quantitative finance. *Appl Sci* 13(3):1956. <https://doi.org/10.3390/app13031956>
7. Nelson DMQ, Pereira ACM, de Oliveira RA (2017) Stock market's price movement prediction with LSTM neural networks. In: *Proc IEEE Int Joint Conf Neural Netw*. <https://doi.org/10.1109/IJCNN.2017.7966019>
8. Kaur P, Singla R (2022) Modelling and forecasting Nifty 50 using hybrid ARIMA-GARCH Model. *Rev Finance Banking* 14(1)
9. Mehtab S, Sen J (2020) A time series analysis-based stock price prediction using machine learning and deep learning models. *Int J Bus Forecast Mark Intell* 6(4). <https://doi.org/10.1504/IJBFMI.2020.115691>
10. Bashir U, Singh K, Mansotra V (2025) Examining daily closing price prediction of the NSE index using an optimized artificial neural network. *J Sci Res* 17(1). <https://doi.org/10.3329/jsr.v17i1.74640>
11. Ke G et al (2017) LightGBM: a highly efficient gradient boosting decision tree. *Adv Neural Inf Process Syst* 30
12. Ballings M, Van den Poel D, Hespeels N, Gryp R (2015) Evaluating multiple classifiers for stock price direction prediction. *Expert Syst Appl* 42(20):7046–7056. <https://doi.org/10.1016/j.eswa.2015.05.013>
13. Khaideem L, Saha S, Dey SR (2018) Predicting the direction of stock market prices using tree-based classifiers. *North Am J Econ Finance* 47. <https://doi.org/10.1016/j.najef.2018.06.013>
14. Fischer T, Krauss C (2018) Deep learning with long short-term memory networks for financial market predictions. *Eur J Oper Res* 270(2):654–669. <https://doi.org/10.1016/j.ejor.2017.11.054>
15. Gündüz H, Yaslan Y, Cataltepe Z (2017) Intraday prediction of Borsa Istanbul using convolutional neural networks and feature correlations. *Knowl Based Syst* 137:138–148. <https://doi.org/10.1016/j.knosys.2017.09.023>
16. Sisodia J, Nayak AKJR, Boghey R (2024) An improved index price/movement prediction by using ensemble CNN and DNN deep learning technique. *J Appl Artif Intell* 5(1). <https://doi.org/10.48185/jaai.v5i1.980>
17. Güreşen E, Kayakutlu G, Daim T (2011) Using artificial neural network models in stock market index prediction. *Expert Syst Appl* 38(8):10389–10397. <https://doi.org/10.1016/j.eswa.2011.02.068>
18. Kara Y, Boyacıoğlu MA, Baykan ÖK (2011) Predicting direction of stock price index movement using artificial neural networks and support vector machines. *Expert Syst Appl* 38(10):12886–12894. <https://doi.org/10.1016/j.eswa.2010.10.027>
19. Jafar SH et al (2023) Forecasting of NIFTY 50 index price by using backward elimination with an LSTM model. *J Risk Financ Manag* 16(10):423. <https://doi.org/10.3390/jrfm16100423>
20. 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 Syst Appl*

- 42(1):259–268. <https://doi.org/10.1016/j.eswa.2014.07.040>
21. Zhong X, Enke D (2017) Forecasting daily stock market return using dimensionality reduction. *Expert Syst Appl* 67:126–139. <https://doi.org/10.1016/j.eswa.2016.09.027>
 22. Heston SL, Korajczyk RA, Sadka R (2010) Intraday patterns in the cross-section of stock returns. *J Finance* 65(4):1369–1407. <https://doi.org/10.1111/j.1540-6261.2010.01573.x>
 23. Huang W, Nakamori Y, Wang S (2005) Forecasting stock market movement direction with support vector machine. *Comput Oper Res* 32(10):2513–2522. <https://doi.org/10.1016/j.cor.2004.03.016>
 24. Patel J, Shah S, Thakkar P, Kotecha K (2015) Predicting stock market index using fusion of machine learning techniques. *Expert Syst Appl* 42(4):2162–2172. <https://doi.org/10.1016/j.eswa.2014.10.031>
 25. Kapri, S.: Bank Nifty Minute Data with Intervals Ranging from 1 Minute to 1 Day. Kaggle / GitHub repository, Xzect Labs Private Limited. Available: <https://github.com/sandeepkapri/banknifty-data>, Accessed: 8 Apr 2026
 26. Landis JR, Koch GG (1977) The measurement of observer agreement for categorical data. *Biometrics* 33(1):159–174