

Prediction of Closing Prices for NASDAQ Listed Stocks: A Comparative Study Based on Gradient Boosting Models

Jingxuan Huang *

School of transportation Science and Engineering, Civil Aviation University of China, Tianjin, China

* Corresponding author: irishuang_jingxuan@ldy.edu.rs

Abstract. The forecasting of trends in U.S. stocks remains a focal point in the financial domain. Within the realm of fine-grained forecasting tasks, predicting price trends with pronounced volatility in the last ten minutes of the closing cross-auction presents a distinctive challenge. This study develops four forecasting models—LightGBM (LGB), XGBoost (XGB), CatBoost (CBT), and a weighted fusion model—utilizing data from the final-market order book and closing auction of 200 NASDAQ stocks. The research aims to conduct a comprehensive comparison of the performance of these four models on the same prediction task. The study results reveal Mean Absolute Error (MAE) values of 5.195, 5.197, 5.200, and 5.202 for the respective models, signifying the integrated model as the most precise predictor. This comparative analysis facilitates a nuanced understanding within the research community regarding the performance of each model, offering a valuable reference for future investigations. Simultaneously, it underscores the inherent value of machine learning in financial forecasting, aiding investors in enhancing decision-making efficiency during critical moments on trading days.

Keywords: NASDAQ, gradient boosting, prediction.

1. Introduction

At the close of each trading day on the Nasdaq Stock Market, the final ten minutes of the Closing Cross auction dictate the official closing prices of listed securities, serving as a pivotal metric for evaluating the performance of individual securities and the overall market. Consequently, accurately predicting fluctuations in stock closing prices during this critical moment holds significant implications for investors, analysts, and other market participants. Built upon this backdrop, this study integrates order book and auction data to establish four predictive models (LightGBM, XGBoost, CatBoost, and a weighted fusion model), intending to thoroughly investigate their performance on the same prediction task. While these models demonstrate comparable performance in terms of Mean Absolute Error, contrasting their outcomes contributes to a more comprehensive understanding of the strengths and similarities of different models in handling financial time series forecasting tasks.

Through this research, we not only provide investors with more reliable tools for stock price prediction but also offer deeper market signal analysis for decision-makers in the financial market. The study seeks to advance the practical application of machine learning in the financial domain, enhance market efficiency, and provide valuable guidance for future research in this field.

2. Literature Review

The pursuit of accurately predicting stock market trends has long been an enticing yet immensely challenging objective. Although the "Efficient Market Hypothesis" proposed by economist Eugene Fama in the 1960s cast a shadow over the feasibility of such practices, the concurrent emergence of "quantitative investing" laid the groundwork for analyzing mathematical patterns to forecast stock trends [1]. The advent of XGBoost (XGB), LightGBM (LGB), and CatBoost (CBT) in 2016-2017 facilitated machine learning on a large scale and in high-dimensional data settings. Consequently, they not only swiftly gained attention in the quantitative research field but continue to be widely employed in the industry over a decade later.

For instance, He et al. utilized data provided by Jane Street to fit a CBT model for predicting stock indices and trading strategies [2]. Ye et al. employs data from Jane Street and the LGB algorithm, trained a stock prediction model [3]. Yun et al. employed a hybrid GA-XGBoost algorithm to fit data from 4,203 trading days of the Korean Composite Stock Price Index 200 (KOSPI) to predict the direction of the index the following day [4]. Zhu used data from the Kaggle platform, and a weighted fusion model based on XGB and LGB to predict the volatility of 112 stocks [5]. Durga et al. utilized five years of historical stock prices from Yahoo Finance, employing linear regression, random forests, and XGB to forecast the stock price of Adani Green Energy Limited, with XGB achieving the highest accuracy [6].

However, in studies predicting stock price trends using the aforementioned three popular gradient boosting models, typically only one or two of these models are involved, lacking comprehensive comparative research on the performance of all three models.

3. Data

The data utilized in this study originates from the Kaggle platform and is provided by the market-making firm Optiver [7]. The dataset encompasses dynamic historical data of the order book and stock closing auction for the last ten minutes of trading on each trading day over a 480-day period for 200 stocks listed on the Nasdaq.

The dataset spans 480 trading days, comprising order book and closing auction data for the final ten minutes of each day, resulting in a total of 5,237,980 unique values. The training set incorporates 4,742,980 data points from days 0 to 435 (constituting 90.5% of the total dataset). The validation set comprises 462,000 data points from days 436 to 477 (constituting 8.8% of the total dataset). The test set covers the last 22 days from days 478 to 480, incorporating 33,000 data points (constituting 0.6% of the total dataset).

4. Methodology

In this study, three widely applied gradient boosting models in financial time series forecasting were selected: XGBoost (XGB), LightGBM (LGB), and CatBoost (CBT). These models have demonstrated excellent performance in handling nonlinear, high-dimensional data, as well as financial data with intricate temporal and spatial relationships. All three models belong to the gradient-boosting algorithm family, where they iteratively train weak learners and combine them, effectively capturing nonlinear relationships and intricate patterns within the data.

4.1. Gradient Boosting

In general, XGBoost (XGB), LightGBM (LGB), and CatBoost (CBT) are all grounded in the Gradient Boosting algorithm and have been further enhanced based on it. The core principle of Gradient Boosting involves iteratively integrating weak models to progressively refine predictive performance, minimize loss functions, and bring model predictions on training data closer to actual values.

During each iteration, a weak model is introduced, focusing on correcting the residuals of the previous model on the training data, which represent the differences between actual values and the predictions of the preceding model. In each iteration, the predictive output of the newly introduced weak model (typically a tree model) is added to the predictions of the previous model, with the addition's weight controlled by a learning rate (η). This process can be expressed as:

$$F_t(x) = F_{t-1}(x) + \eta \cdot h_t(x) \quad (1)$$

Here, the predicted value of the model $F_t(x)$ is the sum of the predicted values of the previous model $F_{t-1}(x)$ and the product of the learning rate η and the new model $h_t(x)$.

Simultaneously, after each iteration, the residual for each sample is computed, defined as the difference between the actual observed value y_i and the current model's prediction $F_{t-1}(x_i)$. These residuals become the training targets for the subsequent model, aiding the model in focusing on correcting the portions that the preceding model failed to predict accurately. The residual calculation is expressed as:

$$r_{it} = y_i - F_{t-1}(x_i) \quad (2)$$

The loss function is utilized in each iteration to quantify the model's performance on the training data, aiming to iteratively adjust model parameters to minimize the value of the loss function. It serves as a measure of the model's predictive errors on the training set and can be represented as:

$$\text{Loss} = \sum_{i=1}^N l(y_i, F_{t-1}(x_i) + \eta \cdot h_t(x_i)) \quad (3)$$

Here, l specifies the specific form of the loss function.

4.2. XGBoost (XGB) Model

XGBoost, introduced by Tianqi Chen in 2014, addresses some of the limitations in Gradient Boosting through regularization and distributed computing methods, earning renowned for its efficient performance and robustness. In the financial domain, the XGB model is frequently employed for handling high-dimensional, complex-structured data. Its handling of missing values and regularization mechanisms contribute to its outstanding performance in financial time series forecasting tasks.

Compared to Gradient Boosting, XGBoost introduces regularization terms into the objective function, which can be expressed as follows:

Objective Function:

$$\text{Objective} = \text{Loss} + \Omega(f) \quad (4)$$

Regularization Term:

$$\Omega(f) = \lambda \cdot \left(\frac{1}{2} \sum_{k=1}^K \omega_k^2\right) + \alpha \cdot \sum_{k=1}^K |\omega_k| \quad (5)$$

Here, λ and α serve as hyperparameters adjusting the strength of regularization.

4.3. LightGBM (LGB) Model

Introduced by Microsoft in 2016, LightGBM enhances the efficiency of gradient boosting for large-scale and high-dimensional datasets. It is a fast, distributed machine learning model based on the gradient boosting framework. Its advantages include efficient training speed, low memory usage, and exceptional handling capabilities for large-scale data. In the financial domain, the LGB model has become a popular choice due to its highly tunable parameters and outstanding performance, providing robust performance for dynamic stock price prediction tasks.

The computational efficiency of LGB is attributed to improvements in the handling of continuous features and the tree-growing strategy compared to traditional Gradient Boosting.

Firstly, LGB departs from the conventional method of sorting features to find split thresholds. Instead, it discretizes continuous features into intervals and constructs histograms based on these intervals. Split calculations are then performed on the histograms. This approach significantly reduces time complexity, resulting in improved computational speed.

Secondly, unlike the depth-first layer-wise growth in traditional methods, LGB adopts a leaf-wise growth strategy. It selects the child node with the maximum gain for the current leaf node to split.

While traditional Gradient Boosting uses a level-wise tree growth approach, splitting all leaves on the current level without considering the gain of newly generated leaves and the gain of leaves not scheduled for splitting, the leaf-wise strategy calculates the gain for all current leaves and selects the leaf with the maximum gain for splitting. This leads to faster fitting speeds. The principle of selecting the optimal split node is as follows in Fig. 1.

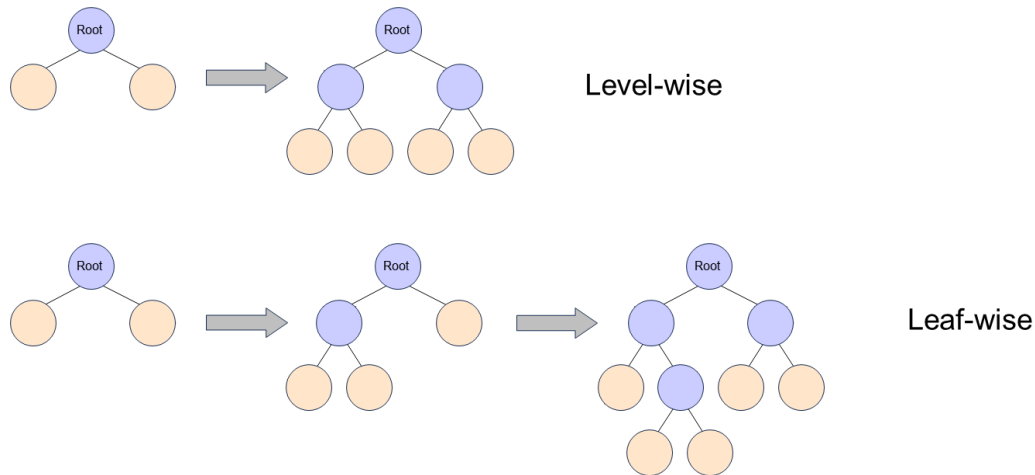


Figure 1. Optimal split node

4.4. CatBoost (CBT) Model

CatBoost is an outstanding gradient-boosting model renowned for its adept handling of categorical features, alleviating the burden of feature engineering by automatically managing these features. In the realm of financial time series prediction, where categorical features are prevalent, the CBT model excels in this task. Its robust ability to handle categorical features has made it highly favored in modeling financial data.

The strength of CBT lies in its unique approach to handling categorical features. Unlike conventional methods that employ one-hot encoding, which increases data dimensions, CBT utilizes target encoding to avoid the "curse of dimensionality." The specific transformation can be expressed as follows:

$$\text{Target Encoding}(x_i) = \frac{\sum_{j=1}^n I(y_j=1) \cdot x_i}{\sum_{j=1}^n I(y_j=1)} \quad (6)$$

Here, x_i represents the value of the i -th categorical feature, y_i is the target variable for the corresponding i -th sample (typically the label in binary classification, 0 or 1), and I is an indicator function, equal to 1 when $y_i = 1$ and 0 otherwise. The formula calculates the proportion of occurrences of the category value for which the target variable $y_i = 1$. This ratio represents the target-encoded value.

4.5. Weighted Fusion Model

To further enhance model performance, this study incorporates a weighted fusion model. By judiciously combining the predictions of LGB, CBT, and XGB with appropriate weights, the research aims to achieve a more robust and accurate predictive outcome by leveraging the strengths of each model. This fusion approach often proves pivotal in practical applications when model selection is uncertain or performances are closely matched.

A fusion model is a technique that combines multiple independent models to obtain a more stable and superior predictive performance than a single model. It has found widespread application in improving the performance of time-series prediction tasks [8]. Common ensemble methods include Bagging, Boosting, Stacking, Voting, Averaging, among others. Despite XGB, LGB, and CBT being exemplary single ensemble models, each employing boosting methods to integrate multiple rounds

of decision trees, the weighted fusion of these three models still demonstrates an enhanced upper limit in predictive performance.

In this study, applying weights of 0.4, 0.3, and 0.3 to LGB, XGB, and CBT, respectively, resulted in the Ensemble model. The MAE score of the Ensemble model slightly outperformed the best-performing individual model, LGB.

5. Results

MAE (Mean Absolute Error) is a metric used to assess the predictive performance of a model. It calculates the average absolute error between the model's predicted values and the actual observed values, providing a measure of the model's average prediction accuracy across all samples. A smaller MAE value indicates more accurate predictions by the model. In comparison to other evaluation metrics, MAE is relatively insensitive to outliers, making it widely applicable in various domains, including finance, medicine, and engineering. The calculation can be expressed as follows:

$$MAE = \frac{1}{n} \sum_{i=1}^n |y_i - x_i| \quad (7)$$

Where n is the total number of data points, y_i is the predicted value for data point i , x_i is the observed value for data point i [7]. This study adopts MAE as the primary performance metric for the models, given the Laplace-distributed residuals, the widespread popularity of Mean Absolute Error (MAE) in assessing regression problems, the natural interpretability of MAE in quantifying average error magnitude, and the Kaggle official evaluation metric being MAE for this dataset [9-10]. Residual Histogram and Quantile-Quantile Plot are shown in the following Fig. 2 and Fig. 3.

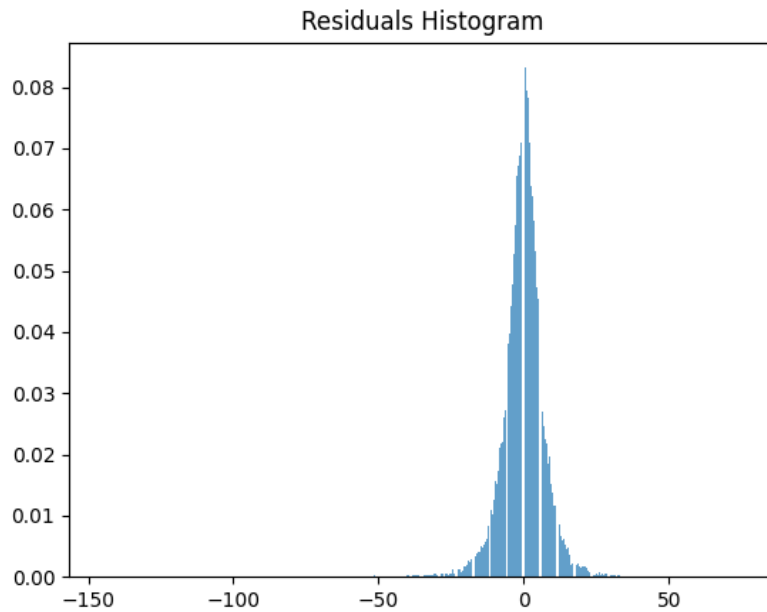


Figure 2. Residuals Histogram

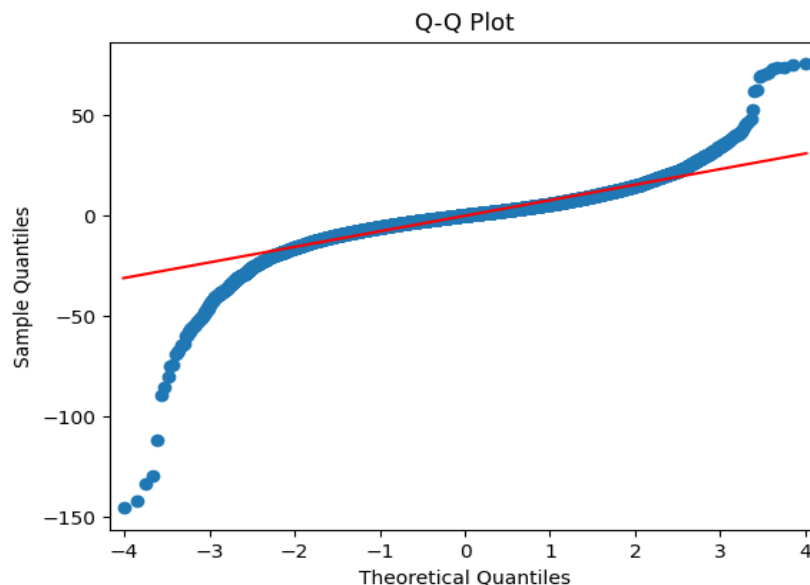


Figure 3. Q-Q Plot

The peaked residual histogram and the curved Q-Q Plot imply that MAE is the optimal evaluation metric for this dataset [9]. The specific scores for different metrics are presented in the Table 1 below.

Table 1. MSE scores for different models

Model	MAE
Ensemble Model	5.1952
LightGBM	5.1970
CatBoost	5.1996
XGBoost	5.2022

From the table, it is evident that the ensemble model performs the best when MAE is used as the primary evaluation metric, aligning with the textbook conclusion that "an ensemble model can achieve better results than any individual model." Therefore, under the premise of prioritizing MAE as the main scoring metric, the ensemble model is the optimal choice for fitting this dataset.

However, it is noteworthy that the scores of the four models are very close. This suggests that although the ensemble model has a slight advantage, the performance of the four models on this dataset is highly similar. This conclusion is consistent with the trend chart weighted across 200 stocks (See Fig. 4).

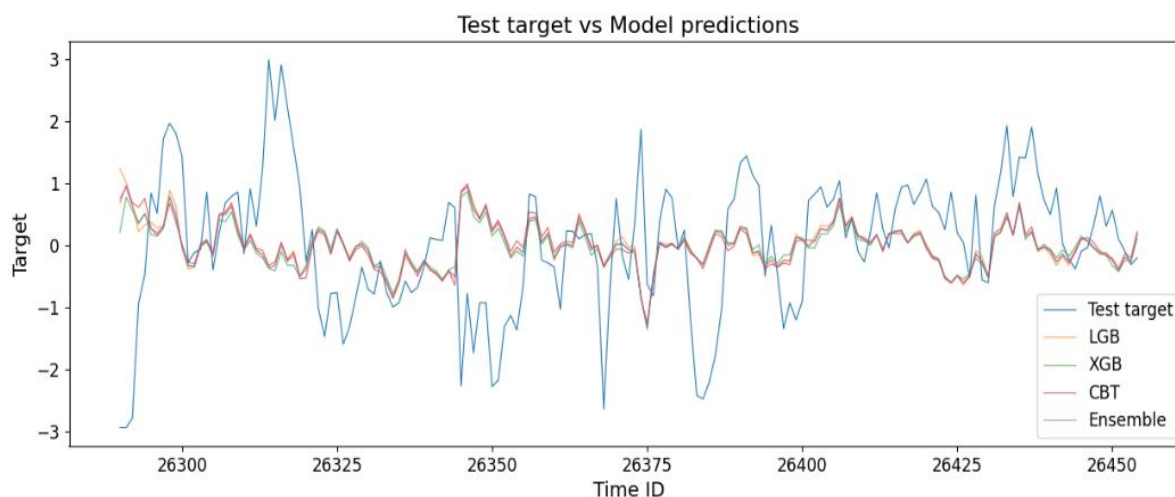


Figure 4. Trend Chart of the Target Variable and Predictions from Different Models

6. Conclusion

This study focuses on predicting trends in the U.S. stock market, with particular attention to forecasting price fluctuations in the last ten minutes of the closing auction. Four prediction models, namely LightGBM, XGBoost, CatBoost, and a weighted fusion model were constructed using closing auction data and order book information from 200 NASDAQ-listed stocks. The study results indicate MAE values of 5.195, 5.197, 5.200, and 5.202 for the four models, highlighting the superior predictive accuracy of the ensemble model. This comparative research provides the research community with a comprehensive understanding of the performance of gradient-boosting tree models in stock price prediction tasks, emphasizing the potential value of machine learning in enhancing efficiency in financial forecasting.

References

- [1] Guo J, Wang S, Ni L M, et al. Quant 4.0: Engineering Quantitative Investment with Automated, Explainable and Knowledge-driven Artificial Intelligence. 2022, arXiv preprint arXiv: 2301.04020.
- [2] He Y, Jingze O, Li M. CatBoost based Jane Street Market Forecast Model. 2021 IEEE International Conference on Artificial Intelligence and Industrial Design (AIID). IEEE, 2021: 51-54.
- [3] Ye F, Wang J, Li Z, et al. Jane Street Stock prediction model based on LightGBM. 2021 6th International Conference on Intelligent Computing and Signal Processing (ICSP). IEEE, 2021: 385 - 388.
- [4] Yun K K, Yoon S W, Won D. Prediction of stock price direction using a hybrid GA-XGBoost algorithm with a three-stage feature engineering process. Expert Systems with Applications, 2021, 186: 115716.
- [5] Zhu D. Predicting stock volatility based on weighted fusion model of XGBoost and LightGBM. 2022 3rd International Conference on Computer Vision, Image and Deep Learning & International Conference on Computer Engineering and Applications (CVIDL & ICCEA). IEEE, 2022: 65 - 67.
- [6] Durga P, Sudhakar T. Machine Learning Classification Methods for Stock Market Forecasting. 2023 International Conference on Innovative Data Communication Technologies and Application (ICIDCA). IEEE, 2023: 14 - 21.
- [7] Tom Forbes, John Macgillivray, Matteo Pietrobon, Sohier Dane, Maggie Demkin. Optiver - Trading at the Close. Kaggle, 2023, available at: <https://kaggle.com/competitions/optiver-trading-at-the-close>.
- [8] Bojer C S, Meldgaard J P. Kaggle forecasting competitions: An overlooked learning opportunity. International Journal of Forecasting, 2021, 37 (2): 587 - 603.
- [9] Hodson T O. Root-mean-square error (RMSE) or mean absolute error (MAE): When to use them or not. Geoscientific Model Development, 2022, 15 (14): 5481 - 5487.
- [10] Willmott C J, Matsuura K. Advantages of the mean absolute error (MAE) over the root mean square error (RMSE) in assessing average model performance. Climate research, 2005, 30 (1): 79 - 82.