NASDAQ Closing Stock Price Prediction
Project Abstract

Meredith Gao, Lucas Goh, Zeki Yan, Tim Zhou

Project Summary

In this project, we aim to develop models that are able to predict the closing movements of the NASDAQ Closing Cross auction. Our dataset contains historical data for the daily ten-minute closing auctions on the NASDAQ stock exchange for the 200 selected stocks at ten-second intervals. Specifically, we want to predict *Target*[1] - the difference between the future movements in Weighted Average Price (*WAP*) of a stock and a synthetic index.

Methodoloy

Approach 1: Calculating Target by predicting its core components — WAP_{t+60} , Index WAP_{t+60}

- **Retrieving Stock Weights w_i For Index *WAP*[1]:** Index *WAP* is a weighted sum of all 200 stock *WAP*. By applying linear regression to given *WAP* of all stocks, we successfully determined the precise weights constituting the index, with R-squared of 1.0.
- **Predicting WAP_{t+60} :** We imputed the missing *WAP* for the last trading minute of each day in the training data through solving a series of linear equations so that we have a complete set of WAP_{t+60} as labels for prediction.
- **Limitation on Predictive Accuracy:** Efficacy of this method was constrained, as reflected by a Mean Absolute Error (MAE) of over 10 in all models. This suboptimal performance could be attributed to the intricate interactions among various stocks, which cannot be assumed to operate independently.

Approach 2: Directly predicting Target

- Our second approach revolves around a simpler, yet more effective approach, which is to predict *Target* directly, results of this approach are listed below.

Feature Engineering

Sample Clustering: Cluster stocks into two clusters using k-means, the clustering label serves as a new feature.

Feature Generation: Involves four main categories of features: stock summary features (including median and standard deviation of price/size), cross-combined features (calculating ratios/products of price/size attributes), combinations of price features (various two/three-price combinations), and advanced features (shift/return, time, and stock-related attributes). This comprehensive approach provides a rich set of data for informed decision-making in the stock market.

Models

Data: Our data is a 30-day sample of the original dataset due to computing resource limitation, using an 80-20 train-test split.

Individual Model Training: 1) ARIMA 2) XGBoost 3) Random Forest 4) Neural Network 5) LightGBM were trained.

Model Ensembling: A linear regression model was trained to ensemble these models by assigning weights to each model.

Evaluation Metric: Mean absolute error (MAE)

Results

Our best model improves MAE by **17.1%** compared to baseline model, and is only **1.1%** worse than Kaggle 1st Place.

	Baseline[2]	ARIMA	XGBoost	Random Forest	Neural Network	LightGBM	Ensemble	Kaggle 1st Place
MAE	6.407	5.452	5.413	5.340	**5.314**	**5.314**	5.412	**5.308**

Impact

Closing prices are pivotal for investors, analysts, and other market stakeholders to assess the performance of specific securities and the broader market landscape. By enhancing the consolidation of signals from auctions and order books, our model can help improve market efficiency and accessibility.

Sources

https://www.kaggle.com/competitions/optiver-trading-at-the-close

[1] $Target = (\frac{Stock\ WAP_{t+60}}{Stock\ WAP_{t}} + \frac{Index\ WAP_{t+60}}{Index\ WAP_{t}}) \times 10000$, where $Index\ WAP_{t} = \sum_{i=1}^{200} w_{i} \times WAP_{t}$

[2] The baseline model predicts a 0.1 target for +ve imbalance and a -0.1 target for -ve imbalance based on '*imbalance_buy_sell_flag*'.