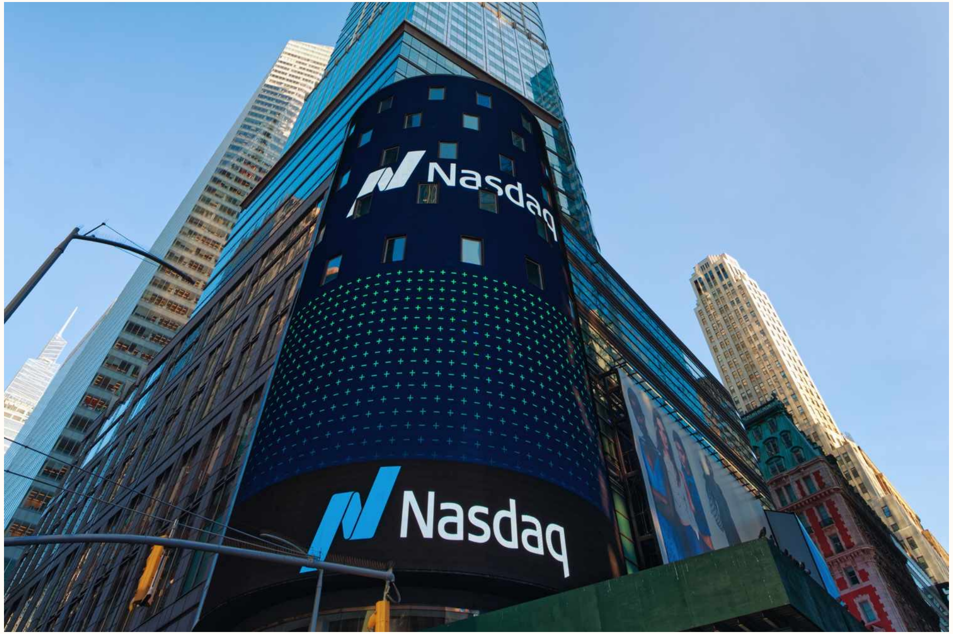
NASDAQ Closing Stock Price Prediction

15.072 Advanced Analytics Edge

Zeki Yan, Meredith Gao, Lucas Goh, Tim Zhou

Introduction

Final 10 minutes of stock market offer critical information to market participants



- Stock exchanges experience **high intensity and volatility**, especially in the **final ten minutes** of the trading day.

- During this time, market makers (i.e. Optiver) **combine data from traditional order books with auction book**.

- Information from these two sources is essential to **offering optimal prices** to all market participants.

Target Variable

Our target variable measures the performance of a stock relative to an index

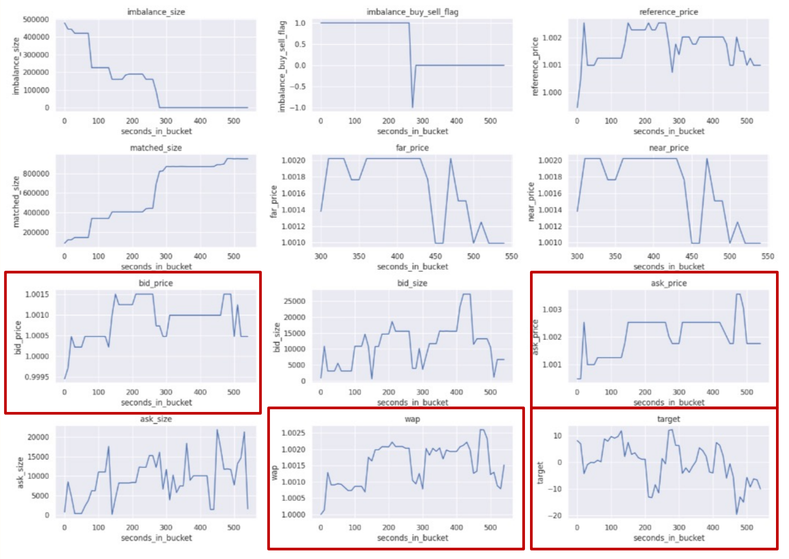
$$\text{Target} = \left(\frac{Stock\ WAP_{t+60}}{Stock\ WAP_{t}} - \frac{Index_{t+60}}{Index_{t}}\right) \times 10{,}000$$

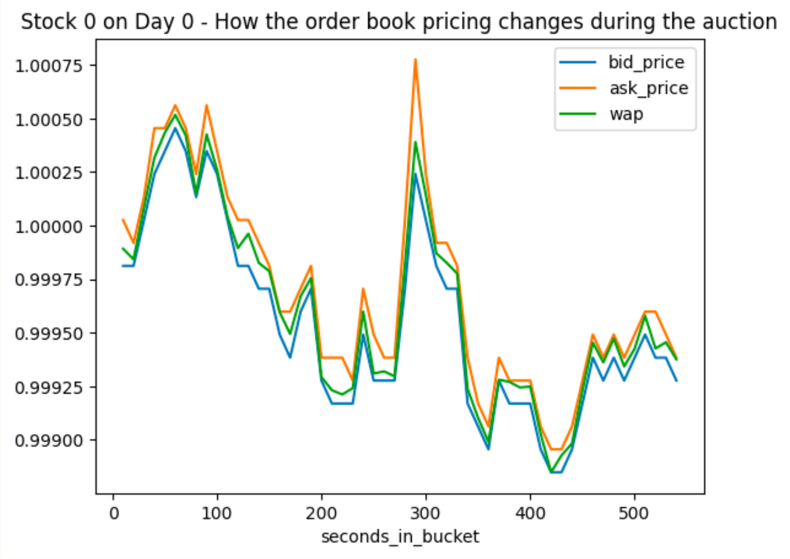
- $Stock\ WAP\ _{t} = Weighted\ Average\ Price\ of\ a\ Stock\ at\ time\ t$
- $Index_{t} = Value\ of\ Stock\ Index\ at\ time\ t$

Positive Target → stock outperforms market in the next minute;
Helps market makers to **predict price movements**

Exploratory Data Analysis

Our target variable measures the performance of a stock relative to an index





WAP is always sandwiched between **bid_price and ask_price**;
Several variables are **highly correlated**

Feature Engineering

Comprehensive feature engineering to power our modeling efforts

Scope	Features Engineered
Row-level	Volume = bid_size + ask_size Liquidity Imbalance = (bid_size – ask_size) / Volume Price Spread = ask_price – bid_price Imbalance Size = bid_size/ask_size Market Urgency = price_spread * liquidity imbalance Lag Prices ……
Stock-specific	Median Volume Median Price Max Spread ……

115 features in total

Finance knowledge leveraged to engineer specific features

Solution Approach #1

One approach is to predict values of Stock WAP at t+60

$$\text{Target} = \left(\boxed{\frac{Stock\ WAP_{t+60}}{Stock\ WAP_t}} - \frac{Index_{t+60}}{Index_t}\right) \times 10{,}000$$

We can opt to only predict Stock WAP at t+60,
only **If we know how Index is computed**

Solution Approach #1

We solved a linear regression to retrieve the formula for *Index*

$$\text{Target} = \left(\frac{Stock\ WAP_{t+60}}{Stock\ WAP_t} - \boxed{\frac{Index_{t+60}}{Index_t}}\right) \times 10{,}000$$

1. We hypothesize that Index is a weighted sum of all *Stock WAPs*
2. This allow us to solve for $Index_t$ using the following linear regression:

$$Index_t = w_0 \cdot Stock\ WAP_{0,t} + \ldots + w_{199} \cdot Stock\ WAP_{199,t}$$

A **R-squared of ~1.0 confirms our hypothesis** as we retrieved the weights with success

```
Sum of Coef: 1.0000000000000002
R2: 0.999999995685508
```

Solution Approach #1

To continue with approach #1, we had to impute values for the last 60 seconds

- We lack data for 10th minute WAP for each day → we can use the 9th minute $WAP, Index,$ and $Target$ to impute missing data

- Solve a Linear Optimization Problem using Gurobi for each ten-second interval t:

Estimated target computed by imputed WAP

WAP should fluctuate around 1

$$\min_{WAP_{t+60}} \boxed{|mean(WAP_{t+60}) - 1|}$$

$$\text{s.t.} \quad \boxed{|(\frac{WAP_{t+60,i}}{WAP_{t,i}} - \frac{\sum_{j=1}^{n} w_j \cdot WAP_{t+60,j}}{Index\ WAP_t})} \cdot 10000 - Target_i| \leq \epsilon, \qquad \forall\ stock\ i$$

- Obtain last-time WAP to include in training set.

Solution Approach #2

To predict for *Target* directly

$$\boxed{\text{Target}} = \left(\frac{Stock\ WAP_{t+60}}{Stock\ WAP_t} - \frac{Index_{t+60}}{Index_t}\right) \times 10{,}000$$

We can also employ a more straightforward approach:
to predict *Target* directly

Results

We achieve decent performance for using both approach; target outperforms still

Table 1: MAE of each model

MAE	Baseline	ARIMA	XGBoost	RF	NN	LightGBM	Ensemble	Kaggle 1^{st}
Predict WAP_{t+60}	6.407	14.566	5.612	5.784	6.192	5.581	5.648	5.308
Predict target	$\boxed{6.407}$	5.452	5.413	5.340	**5.314**	$\boxed{\mathbf{5.314}}$	5.412	$\boxed{5.308}$

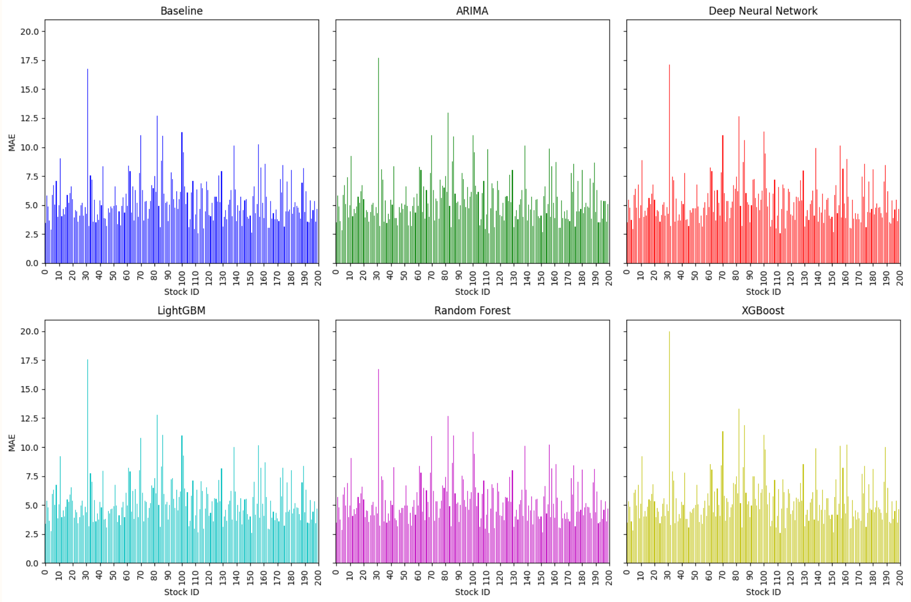
+17%

-0.1%

We achieve **competitive results** with both LightGBM and Neural Network

Results Analysis

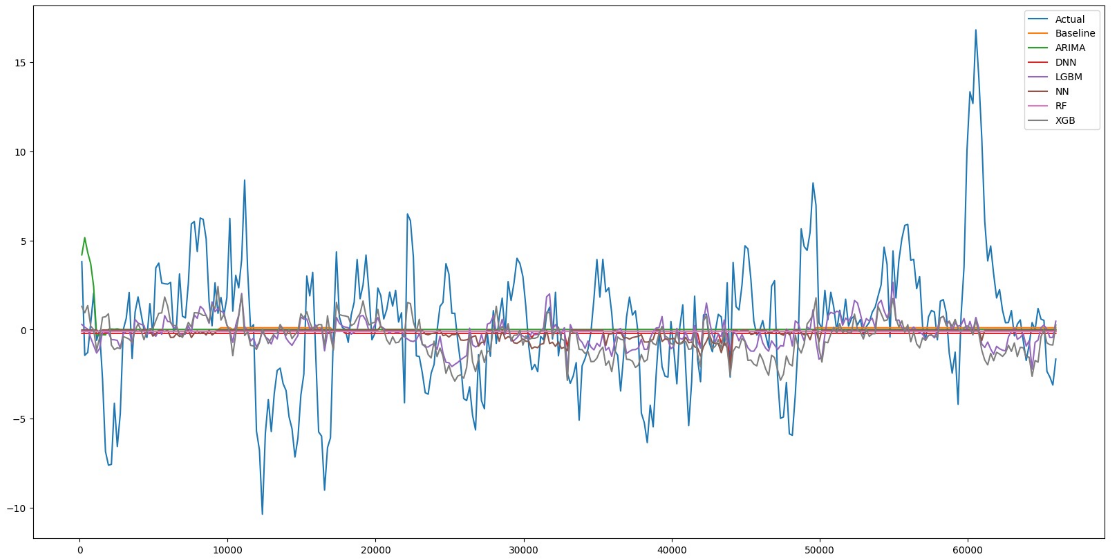
Compare MAE of each model across to detect which stock is the hardest to predict



- **MAE for each stock** using test set data across six models
- The models exhibited **similar patterns**
- Some stocks consistently present challenges in prediction, regardless of the model used

Results Analysis

Models' prediction results vs. actual values Time Series of Stock 151



- The actual target line shows instances of **high volatility**
- The graph reveals that the **NN** and **LGBM** models most closely align with the actual target's high instability

Trading Strategy

Potential trading strategy based on our prediction models and results

For Risk Averse

Prioritize strategy-making on stocks that are easy to predict (e.g. Stock 151)

For Risk Neutral

Trust the model (LightGBM is ~75% accurate in stock's general trend)

For Risk Taking

Short/buy stocks that have drastic fluctuations according to predictions

Conclusion

Our models reduce information gap, enhance market transparency

- Closing prices are **critical for investors, analysts and market stakeholders**, serving as key indicators for assessing securities performance.
- Our model enhances prediction performance by **consolidating signals from auction books** and order books.
- This result helps **reduce information asymmetry**, aiding informed decision making and **enhancing market transparency**.

Thank you!

Zeki Yan, Meredith Gao, Lucas Goh, Tim Zhou