<https://www.kaggle.com/code/chiangken/introduction-and-explore-data-analysis>

<https://www.kaggle.com/code/tomforbes/optiver-trading-at-the-close-introduction>

<https://www.kaggle.com/competitions/optiver-trading-at-the-close/discussion/441590>

<https://www.kaggle.com/code/mtszkw/xgboost-for-stock-trend-prices-prediction>

<https://towardsdatascience.com/machine-learning-techniques-applied-to-stock-price-prediction-6c1994da8001>

(removed from lit review) M. Derksen, B. Kleijn, R. de Vilder, “Heavy tailed distributions in closing auctions”, Physica A: Statistical Mechanics and its Applications, Volume 593, 2022, 126959, ISSN 0378-4371, <https://doi.org/10.1016/j.physa.2022.126959>.  
(<https://www.sciencedirect.com/science/article/pii/S0378437122000620>)

(removed from lit review) “Clearing price distributions in call auctions”, M. Derksen, B. Kleijn and R. Devilder; Quantitative Finance, 2020 Vol. 20, No. 9, 1475–1493  
<https://doi.org/10.1080/14697688.2020.1744699>

<https://www.kaggle.com/code/j12836/lgbm-submission>

LGBM code

[Submitted on 17 Jun 2020] “Markovian RNN: An Adaptive Time Series Prediction Network with HMM-based Switching for Nonstationary Environments” Fatih Ilhan, Oguzhan Karaahmetoglu, Ismail Balaban, Suleyman Serdar Kozat

<https://arxiv.org/abs/2006.10119>

**Onboarding materials and resources:** <https://www.kaggle.com/competitions/optiver-trading-at-the-close/discussion/441966>

**Understanding the context of the competition:**  
<https://www.kaggle.com/competitions/optiver-trading-at-the-close/discussion/442994>

Debugging: Submission Scoring - <https://www.kaggle.com/competitions/optiver-trading-at-the-close/discussion/455531>

Synthetic Index weights:

Linear Regression for Synthetic Stock Weights: <https://www.kaggle.com/competitions/optiver-trading-at-the-close/discussion/442851>

Interim Progress reports: Oct 11, Nov 1

• Periodic progress reports need to be submitted. Each report should be about 2-3 pages (PDF) in length. Here is a suggested structure.

• Title of project, Names of the two team members

• Introduction: describe the problem you are solving and your approach to the problem.

• Problem statement: Describe your problem precisely using concepts covered in the class, specify the dataset you are using, expected results and model selection and evaluation procedure.

• Literature Review (at least 3 sources with URL references)

• Technical approach: Describe the model building procedures. Provide proof that code is running and your data pipeline is in place. Links to Github repos or other code bases/datasets you are working off of.

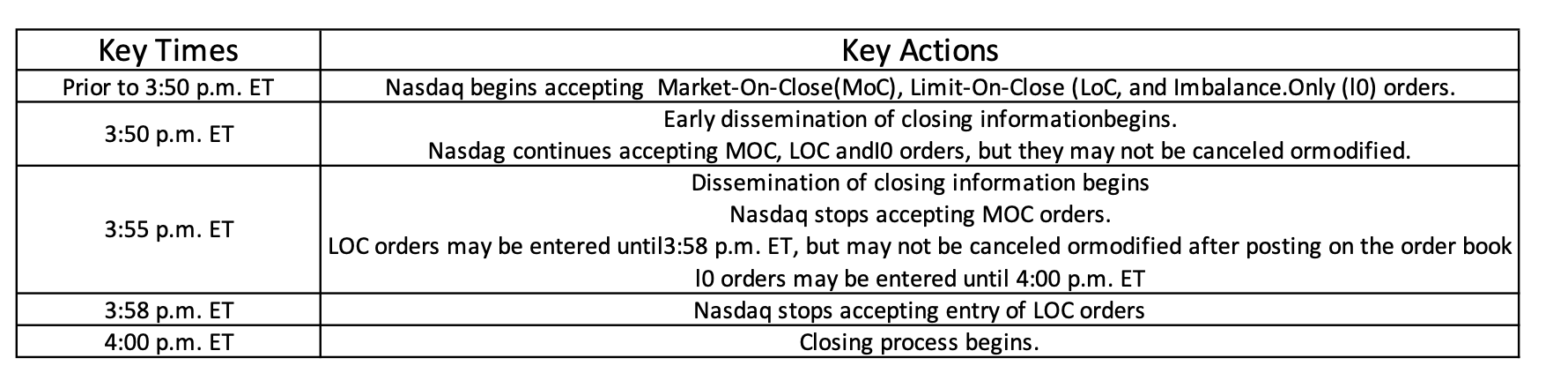
• Results so far: State and evaluate your results obtained thus far, describing progress since the previous update.

• Submission: Please submit your report as a PDF on Gradescope. Only one person on your team should submit. Please have this person add the other team member as collaborator as a "Group Submission".

## Data

There are 200 unique stocks, but not all stock IDs exist in every time bucket.

Data is provided over **481** **days**. There is data for 55 seconds every day. We can check that   
time\_id = 481 ∗ 55 (Number of date ∗ Number of seconds in bucket in a day)



Seconds\_in\_bucket is the number of seconds elapsed since the beginning of the day's closing auction, always starting from 0. It increments in 10 second intervals, i.e. it goes 0, 10, 20, 30, … 540 - 9 minutes (from 3:50 pm to 3:59 pm). We have to predict values for last minute

**Row\_id** is unique for every row. It is constructed as: <date\_id>**\_**<seconds*\_*in\_bucket>**\_**<stock\_id>

E.g. for a random row which is at date\_id 140, seconds\_in\_bucket is 20 and stock\_id is 86, the row\_id is ‘140\_20\_86’

Last Value: This is very similar to the baseline model, except here we predict that the auction price does not change and output the last value available. - since all stocks are not in every time bucket, for a given sample prediction, we need to find the stock\_id on the given date\_id and find the value corresponding to seconds\_in\_bucket equal to 540 and return that value as the last value.

Sample\_prediction is the output dictionary to be returned. Revealed\_target contains the values for the previous date, i.e. the auction period 10 mins

<https://www.kaggle.com/competitions/optiver-trading-at-the-close/discussion/443275>

# Poster

1. Problem Statement
2. Dataset Description & Exploration
   1. Size, features
   2. Basic statistical exploration
   3. Clustering: K-means, GMM
3. Model Selection
4. Training and Hyperparameter Tuning
5. Kaggle Submissions
   1. Current Rank, timeline & results
   2. Difference from Top X (1, 50, 100, 200)
6. Future Work
7. Conclusions
8. References