Project Proposal - Optiver’s Trading at the Close

# Problem

In this project, we will tackle the Kaggle competition “Optiver - Closing at the Trade”, which involves building a model that can predict the closing price movements for 200 Nasdaq listed stocks using data from the order book and the closing auction of the stock.1

The dataset provided by Optiver for this competition contains time series data from the daily Nasdaq Closing Cross Auctions.

In addition to this dataset, we will potentially explore other data sources such as mentions, and expert predictions on social media. By tracking sentiments and events in the news, we can get additional context on how the company stock is likely to do.

# Background

There are several prior works for this subject. A few sample implementations that we will take a look at are:

1. Tutorial Notebook for the competition by Optiver:- - <https://www.kaggle.com/competitions/optiver-trading-at-the-close/discussion/441590>   
   As we are adapting this project from the Kaggle competition, we will also be referring to competitive implementations that utilize only the dataset provided here.
2. S. Qiu and Z. Jia, "Forecasting the Trend of NASDAQ Stock Based on Machine Learning and Tween Analysis," 2022 2nd International Conference on Big Data, Artificial Intelligence and Risk Management (ICBAR), Xi'an, China, 2022, pp. 74-78, doi: 10.1109/ICBAR58199.2022.00022. <https://ieeexplore.ieee.org/document/10108281>
3. Market prediction using machine learning based on social media specific features

Original Article Open Access Published: 22 February 2023 volume 28, pages 410–417 (2023) <https://link.springer.com/article/10.1007/s10015-023-00857-z>

1. Social Media and Stock Market Prediction: A Big Data Approach – by Mazhar Javed Awan1,2,\*, Mohd Shafry Mohd Rahim1, Haitham Nobanee3,4,5, Ashna Munawar2, Awais Yasin6 and Azlan Mohd Zain7 <https://www.techscience.com/cmc/v67n2/41320/html>

In addition to this, we hope to utilize some of the prior work from <https://informationtracer.com> in gathering data from social media.

# Data

Our primary dataset will be taken from the Kaggle competition hosted by Optiver:

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

This dataset contains Orderbook/Auction data containing bid/ask price and bid/ask size information for 200 stocks over a period of 481 days. (Not all stocks have data on each day). The size of the dataset is 5237980 rows × 17 columns.

In addition to this, we may look at combining the above dataset with the data we will acquire from social media trends corresponding to the dates in the Optiver dataset. We are yet to identify reliable, exhaustive and influential sources for the same. Once identified, we will modify the dataset to include any events that occurred related to a particular stock or industry. We will extract the associated sentiment polarity which will be float values in the closed range -1 to +1, and include it in our analysis.   
The social media data may be acquired using the tool <https://informationtracer.com> developed by Zhouhan Chen, or through web scraping methods.

# Method

For this project, we will be taking a multi-modal approach. The data set taken from the kaggle competition will be preprocessed and imputed into a supervised ensemble model such as random forest, XGBoost, LightGBM, etc. based on model performance.

In the first phase, we will implement a baseline model which, as per the Optiver tutorial:   
“assumes that there is no valuable information about the direction any stock moves, which translates to a predicted value of 0 for all observations.”

For the second phase, we will approach this problem using gradient boosting models - XGBoost and LightGBM. We will also build Pytorch neural net models to handle the sequentiality in the time series data.

Parallely, we will narrow down 1-2 industry domains for social media analysis and extract events for sentiment analysis. For this information obtained from social media or finance blogs, we will process them through a separate preprocessing pipeline, such as the ones mentioned in the work presented in the “Background” section and processed through a neural network or an ensemble model based on the nature of the dataset obtained.

The final result will be obtained by ensembling the results of the 2 datasets.

An alternative approach to this problem would be to preprocess the 2 different datasets and combine them to obtain a uni-model prediction.

# Results

Quantitatively, the metrics mentioned for model evaluation will be Mean Squared Error as it is the standard metric recommended in the competition. However, given the nature of stock market data, we will perform additional statistical testing such as t-tests, in order to ascertain the stability of the model.

Qualitatively, we believe that a multi-model approach to stock market predictions would make predictions more accurate than either of the models stand alone as this method helps us to quantify and assess more “uncertainties” which are usually not recorded in a typical stock market model.

# References

1. [https://www.kaggle.com/competitions/optiver-trading-at-the-close](https://www.kaggle.com/competitions/optiver-trading-at-the-close/data)