## Task 2. Business understanding

**1. Identifying Business Goals:**

**Background:**

This project focuses on developing a predictive model for Nasdaq-listed stocks' closing price movements, utilizing data from the order book and the closing auction. The initiative is crucial for market makers like Optiver, operating in the dynamic environment of stock exchanges, especially during the intense final ten minutes of trading.

**Business Goals:**

The primary objective is to enhance Optiver's trading strategies by identifying trends and parameters influencing stock price movements. The central goal is to construct a deep learning model capable of predicting which stocks will rise or fall in the coming days/weeks.

**Business Success Criteria:**

Success will be measured by the model's ability to accurately predict stock price movements, outperforming baseline models. Additionally, success criteria include improved trading strategies, increased market efficiency, and enhanced accessibility during the critical final ten minutes of trading.

**2. Assessing the Situation:**

**Inventory of Resources:**

The project leverages historical data from the order book and closing auction, computational resources for model development, and Optiver's expertise in electronic market making.

**Requirements, Assumptions, and Constraints:**

Requirements include access to accurate and timely order book and auction data. Assumptions are based on the persistence of market patterns observed in historical data, while constraints involve potential limitations in data availability and unforeseen disruptions in the Nasdaq trading environment.

**Risks and Contingencies:**

Risks encompass model overfitting due to market anomalies and potential data discrepancies. Contingency plans involve robust model validation, continuous monitoring for emerging risks, and adaptability to changing market conditions.

**Terminology:**

Standard financial and trading terminologies will be employed to ensure clear communication and understanding among project stakeholders.

**Costs and Benefits:**

Costs involve computational resources, data acquisition, and potential regulatory compliance. Benefits include improved trading efficiency, enhanced market competitiveness, and the potential for increased profits through informed trading decisions.

**3. Defining Data-Mining Goals:**

**Data-Mining Goals:**

The primary goal is to construct a deep learning model capable of predicting stock price movements based on order book and closing auction data. Secondary goals include feature engineering to capture relevant market dynamics and ensuring model interpretability to align with Optiver's understanding.

**Data-Mining Success Criteria:**

Success will be determined by the model's ability to outperform baseline benchmarks, demonstrate robustness across varying market conditions, and provide actionable insights for trading strategies. Compatibility with Optiver's existing technological infrastructure is also a success criterion.

In conclusion, this business understanding report outlines the project's goals and assesses the situation in alignment with the CRISP-DM framework. The success of the project will not only benefit Optiver by enhancing its trading strategies but also contribute to improved market efficiency and accessibility during critical trading moments, thus reinforcing Optiver's position as a leading global electronic market maker.

## Task 3. Data understanding

**1. Gathering Data:**

**Outline Data Requirements:**

The dataset comprises daily ten-minute closing auction data from the NASDAQ stock exchange. Essential variables include stock\_id, date\_id, imbalance\_size, imbalance\_buy\_sell\_flag, reference\_price, matched\_size, far\_price, near\_price, [bid/ask]\_price, [bid/ask]\_size, wap, seconds\_in\_bucket, and the target variable representing the 60-second future move in the wap.

**Verify Data Availability:**

The primary data files are train.csv and test.csv, containing the auction data. Both training and test datasets are expected to be delivered via the API. Additional files such as sample\_submission, revealed\_targets, public\_timeseries\_testing\_util.py, example\_test\_files, and optiver2023/ are available for reference and API functionality.

**Define Selection Criteria:**

All necessary features for predicting stock price movements, especially those contributing to the target variable, will be included. Any stock\_id and date\_id combinations lacking essential information will be excluded from the analysis.

**Describing/Exploring/Verifying data:**There are a total of 17 variables in the train.csv file and over a million rows of data. Variables *far\_price* and *near\_price* seem to have a lot of missing values. We will have to test and see which variables have the highest correlation with the *target* variable, as that is the one we are going to want to predict. Overall there is a lot of data to use to try and predict the *target* value.

## Task 4. Planning your project

Since we haven't fully divided each task between team members we will estimate the time by member 1, member 2 etc.

**Tasks:**

1. **Cleaning data** - We have a lot of data at our disposal, so the first task will be to clean up missing values and converting data to a proper structure. Since we got the idea and data from Kaggle, then the data is relatively well cleaned. For example the dates are uniform across all data points and currencies are converted into USD, but there is still work to be done to make the data useful for our project.

**Estimated time: Member 1 → 6 hours, member 2 → 6 hours**

1. **Identifying variables -** The essential part of our project is to find out which variables have a high correlation with the *target* variable. We have a lot of data and quite a number of variables, so finding the correct variables is crucial.

**Estimated time: member 3 → 6 hours**

1. **Developing a model -** Having found the proper variables, it is time to construct a model. We will also have to test which model works best for our project. Since we are trying to predict a numerical value and not a boolean value, we will most likely go for a regression based model. We will experiment with different models like random forests and neural networks. Since we are working with time series data we want to experiment with an LSTM network.

**Estimated time: member 1 → 15 hours, member 2 → 15 hours, member 3 → 15 hours**

1. **Reflecting on the model -** Having found the proper variables and constructing a model, the next task is to see if our model with its input has predicted the *target* variable well and if there are some optimizations we can make for the model to run faster and keep its good prediction score. It’s important to do tasks 2 and 3 well, so we don’t have to keep on coming back to them and repeating the tasks over and over again. This process includes tuning the model, for example adding or removing layers, changing the number of nodes, changing weights in case we go the neural network route. But the process more or less applies to other models.

**Estimated time: member 1 → 8 hours, member 2 → 8 hours, member 3 → 8 hours**

1. **Predicting what stocks to invest in -** Now after constructing our model and being satisfied with it, we should identify which stocks will be rising in the coming days and which stocks will be falling. If our model has correctly predicted which stocks have indeed risen and we feel confident in the model, we will try investing in them and see how much of a profit we can make.

**Estimated time: member 1 → 3 hours, member 2 → 3 hours, member 3 → 3 hours**