We first created a baseline model for predicting the time series data using Autoregressive integrated moving average(ARIMA) model. An ARIMA model is specifically designed for time series, which can forecast and understand time series data without feature extraction. In such model, the value and error of a certain time is canulated by a combination of its previous terms. A normal ARIMA model takes the following form:

ϕ(B)(1−B)dYt​=θ(B)εt​

ϕ(B)=1−ϕ1​B−ϕ2​B2−⋯−ϕp​Bp  
𝜃(𝐵)=1+𝜃1𝐵+𝜃2𝐵2+⋯+𝜃𝑞𝐵𝑞θ(B)=1+θ1​B+θ2​B2+⋯+θq​Bq

BkYt​=Yt−k  
Where(1-B)d represents performing d difference to the time series Yd, ϕ(B) is an autoregressive polynomial, θ(B) is the moving average polynomial, and B is the backshift operator. The number of past values used in autoregression and moving average, together with the difference order, are adjusted and evaluated using AIC and BIC to find the best fit for the data.

In order to improve the performance of the model, we added seasonal patterns to it, making it the SARIMA model. The form of the seasonal part of the model is similar to ARIMA, but the backshift operator is BS, where S is the length of the seasonal period. The final SARIMA model takes the following form:

Φ(Bs)ϕ(B)(1−Bs)D(1−B)dYt​=Θ(Bs)θ(B)εt​  
We followed the Box-Jenkins procedure when implementing the model, including ADF test to check if the data is stationary, plotting ACF and PAC for parameter estimation, calculating AIC and BIC to evaluate different sets of parameters, and diagnostic checking after model fitting.

The ARIMA model was first used on the price time series in the Tapes data to grasp the mainstream trend of the stock market. The main reason to start with Tapes data is that the raw LOB data contains long lists of ask and bid quotes, which needs specific design to be transformed into a single timeseries with minimum information loss. The Tapes dataset, however, contains a single time series which can act as the summary of the asks, bids, and the stock market. The model parameter was set to (1,1,1) through ADF testing and evaluation metrics with AIC, BIC and HQIC. The model performance is in figure(), which shows overfitting in the in-sample prediction part and unsatisfied further prediction. The prediction proved to be fluctuating in a far smaller range than the train data, which shows that the model didn’t capture the transaction trend over time.

After further exploration towards the data, we discovered a periodicity on certain part of the asks and bids data, which gave us insight about implementing SARIMA model. To compare the performance between models more reliably, we chose to use LOB dataset as data source on all 3 models. The time series data of asks and bids were adjusted by taking the average of all values for each hour to meet the data structure requirement of the SARIMA model. As we can see from the figure(), SARIMA performs better than ARIMA, with lower overfitting and a basic ability to predict the upcoming price. However, the model shows a clear tendency of repeating a single period. Such feature makes the model hard to perform well enough on time series which can be easily influenced by external factors like the stock market, and is difficult to perform trading strategy on, because it cannot create diverse scenes for the strategy.