Report

Most of my modeling decisions are in the notebooks and as inline comments. The Readme gives general instructions on how to create the models and process the data.
The makefile should take care of everything.

There are exploratory notebooks that I used to decide on modeling features.

Executive Summary

I modeled forward price movements, but the model did not yield significant results.

I then modeled the forward liquidity of the best bid/ask price level using a lasso regression. The model was largely based on momentum of features and price volatility.

Overall, the model was not significant, but shows promise as a causal model.

Notes on Modeling

Based on the existing research I read, it seemed that the most interesting predictive features would be the **bid/ask spread, price volatility, liquidity, and supply/demand pressures**.

For modeling purposes the changes in those features are more appropriate, as it transforms them into more stationary signals, and thus more appropriate as predictive targets.

I wanted to create a causal model, not just coincidental.
Because many of the forces mentioned above are typically autocorrelated and the causal pathway is not always one-way, this made the modeling task difficult.

So although I made the predictive model for liquidity, an autoregressive VAR model may be more appropriate for this task.

The Lasso model will by its nature illuminate which features are additive to our predictive model.

Sampling
I chose to use a 10 percent sample of one of the orderbooks for the exploratory data analysis in each case. I chose this approach to reduce the amount of bias in the analysis. By limiting the data that I analyzed and used in the model training and then assessing for predictive power on an unseen dataset, I reduce the risk of creating a model overfit to a small amount of data.

I also only took one observation for each timestamp. So the predictive modeling is for each timestep, not intra timestep.

Models

From the exploratory data analysis, and my experience in financial time series modeling, I knew that momentum or autocorrelation would be an important factor in this modeling.

Spread Modeling

It looked like the spread at each price level was virtually invariant - e.g. the spread between ap0 and bp0 was always appx 5 dollars in the 20190612 dataset. This being the case I decided to pursue a more interesting predictive feature.

Price Movement Modeling

A principle feature we would have interest in is future price movement.
I looked at price movements 5000 steps into the future.

As the steps vary in their time interval, I needed to control for this. To normalize this feature, we will want to look at the **change in price over a change in time**.

The theoretical causal model in this case would be forward price movement as a function of supply/demand pressure, price momentum, and market liquidity. The supply/demand pressure is measured here as the quantity of bids less the quantity of asks (i.e. a positive number is higher demand, a negative number is higher supply), momentum is price change over a defined period of time, and liquidity is the spread at each price level.

Unfortunately, lasso regression minimized all of the feature importance to zero, and returned a constant as the best predictor.

So I moved on to modeling liquidity.

Liquidity Modeling

The next model is on the future liquidity of the best bid and ask. This would be `bq0 - aq0`. This also acts as a proxy for the supply/demand pressure at this price level.

The causal model is forward change in liquidity (bq0 - aq0) as a function of past changes in liquidity across all levels (`bqx - aqx`), rolling price volatility, and rolling measures of price momentum. All past windows are equal to the forward prediction window. The forward and rolling window was 5000 steps in my analysis.

I added a feature of bid and ask vwap. The motivation was that the bid and ask price movements are slightly independent; and the vwap is a better measure of long term price movements and pressures.

Interestingly the most important feature in the initial modeling was the past change in ;iquidity for price level 0. This would suggest that the liquidity has a momentum. The next most impotrant features we the bid and ask vwap momentum, and the price volatility (standard deviation). These too suggest that liquidity is mainly a function of price momentum and the volatility of that momentum.

Unfortunately, this model did not return significant results; however, it does provide a useful causal model that can be improved upon.

Potential Improvements

I would like to combine the models from each day into a voting machine. I think my sampling method could be improved by taking an agregation from each timestamp rather than just the first observation.

I did not hypertune the Lasso model parameters (alpha) because the purpose of the Lasso model is to suss out which features are the most significant. If I were to choose the Lasso model as a final model, I would spend more time tuning the hyper parameters.

Incorporating data from other assets would likely improve model predictive performance, as assets tend to move in clusters.

Calculating more features of liquidity across all bids and asks would likely improve model performance as well. This could be the change in total quantity of bids less total quanity of asks.

The project would be improved with analysis of the Lasso model assumptions. Although Lasso model can handle multicollinearity, there may be outliers in the data causing leverage points.

Normalization or standarization of some of the features would likely improve model performance and increase the ability ot generalize the model.

I would have also liked to create a modeling pipeline in sklearn to handle all of the data cleaning and feature engineering.

The use of tensorflow rather than sklearn would also allow the the fine-tuning of the model. By using sklearn as the modeling framework, I had to make separate models for each day. I would have preferred to create a single model, and updated it as time went on.