Capstone Three: LSTM Neural Network Candlestick Forecasting for Short-term Trading Opportunities

Problem Statement Formation

To reduce risk and maximize returns on short-term (1-5 trading days) trades, what are the expected open, high, low, and close (OHLC) prices for a given index, stock, ETF, forex, or futures contract tomorrow? What are the expected OHLC prices for the following week? What is the most effective rolling window size for training data?

Context

Given that stock prices can be modeled by a random walk, forecasting prices may seem to be an exercise in futility. However, using neural networks to identify patterns in OHLC data may provide some insight for potential trading opportunities. In trading, a popular approach in technical analysis is the use of candlestick patterns to identify potential reversals or continuation in price changes over time. This relies on price over time reflecting collective psychology that tends to repeat itself and the tendency for prices to regress to a longer-term mean. Further, incorporating information other than price itself, such as volume, moving averages, and RSI, may yield more reliable models in practice as traders and algorithms rely on these for making decisions. In conjunction with some trading rules, such as avoiding shorts on indices (since they tend to go up over time) and stop-loss orders based on volume profile, a model or set of models may provide an edge to grow an account over time. Based on this paper, a short long-term memory (LSTM) neural network model will be used for this application.

Criteria for Success

The primary criterion for success is a model that provides a forecast with reasonable MSE, that is, a model with an MSE that is not orders of magnitude larger than a typical day's trading range of the asset.

Deliverables for this project include a project report, slide deck, and at least one LSTM neural network model.

Scope of Solution Space

While the goal is to create a model that is generalized for any asset with usable OHLC and technical indicator data, training data will be limited to daily and hourly data on SPY and QQQ. The primary scope is a 1-candle forecast, but a label dataset for a 5-candle forecast will also be attempted. The 5-candle forecast label will be the open, high, low, and close for the 5-candle period, not for each individual candle of the 5-candle period. Further, three sliding window sizes will be evaluated: 5, 10, and 20 candles. In addition to OHLC data, the volumes, 21-candle moving average, 50-candle moving average, and default Relative Strength Index will be added as features.

Constraints

The constraint for this project is the amount of data available for training. Ideally, daily data would be used for each asset to train a model for that particular asset. This would ensure the model is trained to any peculiarities of the asset itself and remove some noise of the lower timeframes. However, that greatly reduces the amount of data that can be gathered for training. Two approaches will be attempted

to address this: 1) the standardized distributions of each asset's daily data (percent change from Candle 1 Open Price) will be compared, and if they are not drastically different, the z-scores will be blended, and 2) the standardized distributions of hourly data (percent change from Candle 1 Open Price) will also be compared, and if they are not drastically different, they will also be blended. Since the data is not expected to be normally distributed around zero (since they increase more over the long-term), a test for normality will be conducted and a log transform considered for both approaches as well. If these approaches demonstrate feasibility, future work could involve incorporating data from other index ETFs and/or forex data.

Data Sources

The primary source for training data will be the Alpha Vantage API.

General approach: The general approach will start by pulling all available daily and hourly data for the two big index ETFs, SPY and QQQ. The three rolling window sizes will then be generated for these four datasets, and subsequent calculation of percent changes and standardization will be performed on the 12 new datasets – three for each ETF's daily data and three for each ETF's hourly data. It is expected that QQQ will have a higher standard deviation of percent changes, so this will be a good evaluation for the two standardization approaches listed above.

Then, all available daily and hourly technical indicator data for the two symbols will be pulled, standardized, and merged with the respective OHLC datasets by date or time. Finally, the four datasets for each rolling window will be concatenated.

References

D. Shah, W. Campbell and F. H. Zulkernine, "A Comparative Study of LSTM and DNN for Stock Market Forecasting," 2018 IEEE International Conference on Big Data (Big Data), 2018, pp. 4148-4155, doi: 10.1109/BigData.2018.8622462.