Predicting Microsoft Stock Price 'Inefficiencies'

Zachary Mor

Brown University
DATA 1030: Hands on Data-Science
https://github.com/zachmor/data1030-proj

I. Introduction

It would be easier to know whether or not you should buy MSFT stock options if it was possible to predict if MSFT is *undervalued*, *overvalued*, *efficient* or *volatile*, specifically:

Efficient – not undervalued and not overvalued
Undervalued – before the end of the following week it sustains a price 3% higher for 2 days
Overvalued – before the end of the following week it sustains a price 3% lower for 2 days
Volatile – undervalued and overvalued

This information would be helpful for option trading strategies because it can be used to better assess the value of a contract. If it can be deduced that the price is going to be at a value higher or lower than the option contracts net exercise price (strike price plus premium) the contract can be bought or sold accordingly and hopefully realized profitably. Here, trading data from the NYSE from all of 2021 and the first half of 2022 are used to predict the inefficiency class of the price.

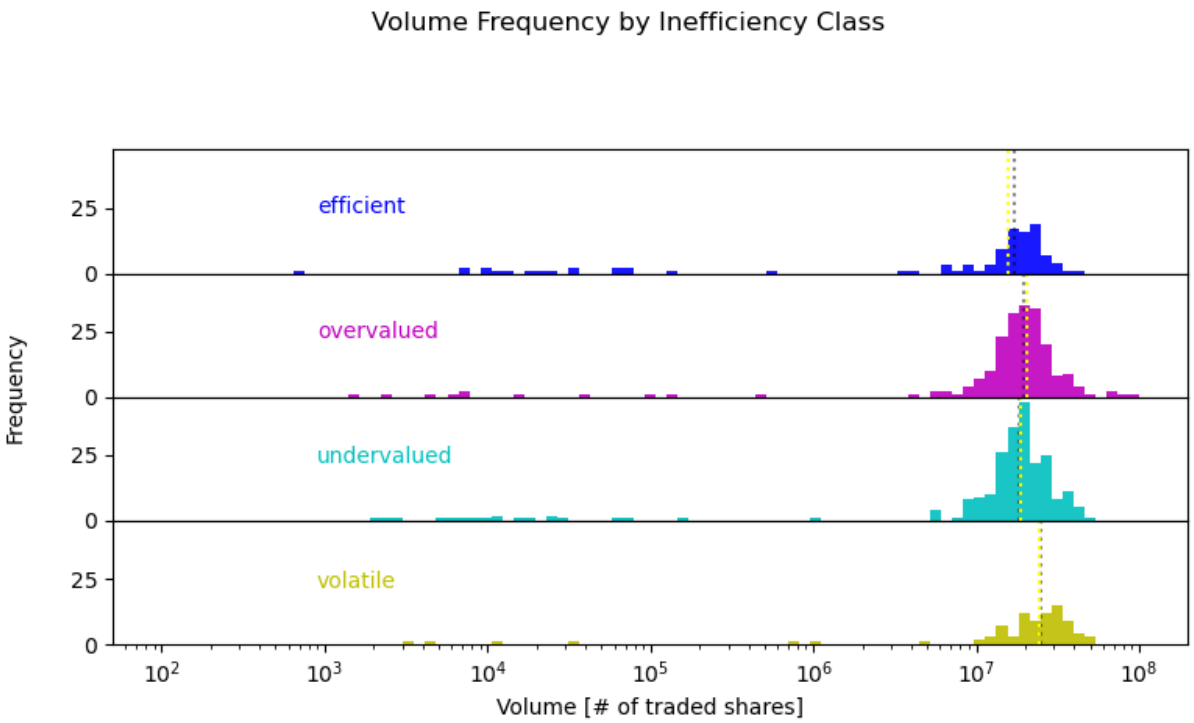
Much research has attempted similar problems, some referenced here are done by (Shen) and Ali, but they use deep learning methods and they classify it more parsimoniously, on just the direction of change rather than *inefficiencies*. Shen achieves > 90% accuracy with some models and Ali achieves similar results. I hope to expect similar results, though with the more complex classification scheme it may worsen.

The dataset contains 1-minute MSFT stock trading data for 231,171 records between 2021-1-1 and 2022-7-1 which is just 30% of 788,400, the number of minutes in a year. The other 70% of the minutes don't have trading data because the market is only open during work-hours on business days.

Each record contains 6 features an opening timestamp, opening price, high price, low price, close price, and volume. The open/close values provide a uniformly distributed random record of last-traded price taken at the beginning and end of the one minute period respectively. The high/low values lends itself to range-wise knowledge which speaks to the volatility of the trading period. Volume, the total number of shares traded, signals the intensity of the trading period.

II. Exploratory Data Analysis

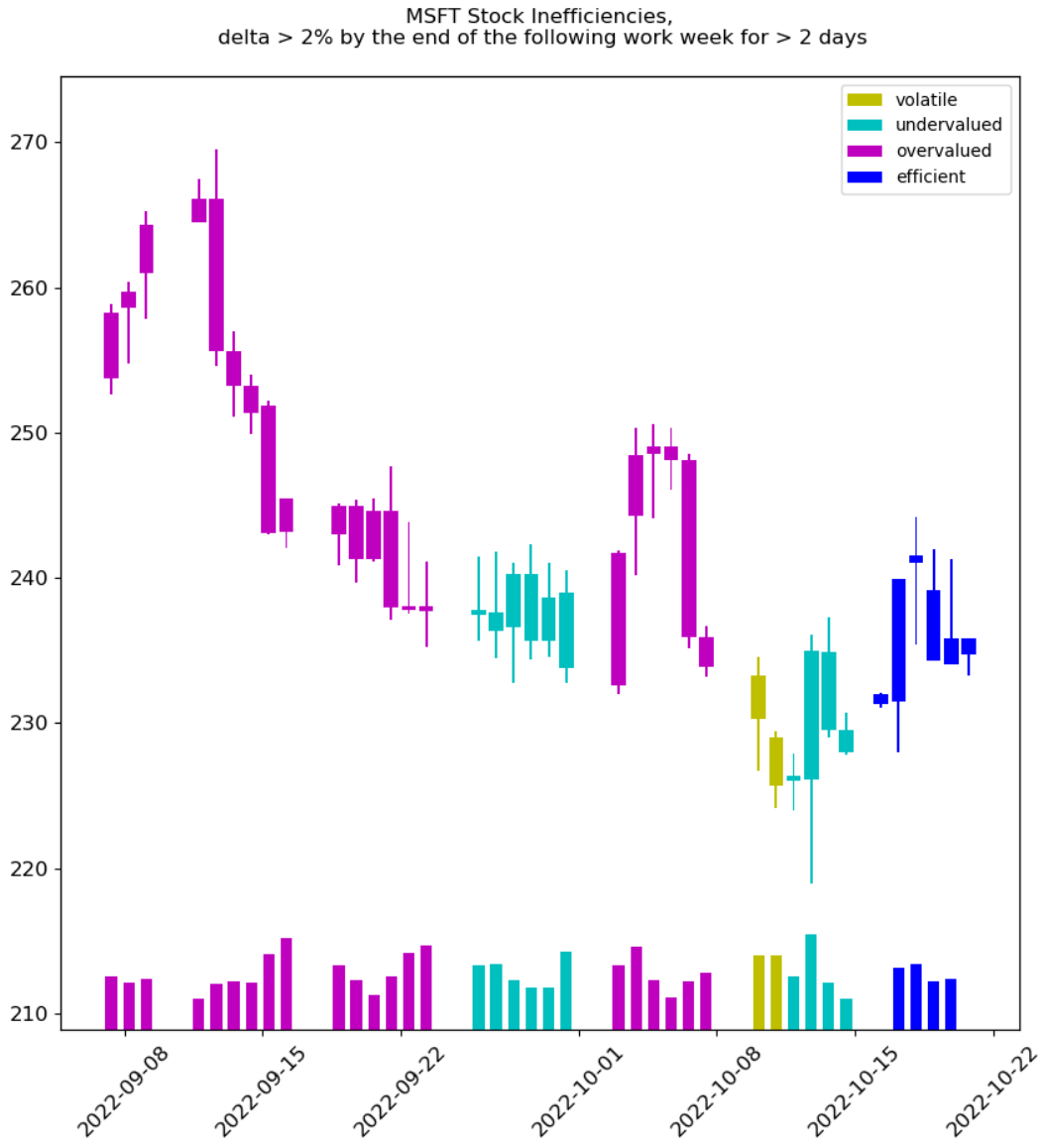
Visualizing multi signal time series requires feature-inclusive plots sometimes because features recorded together must be visualized together to make sense. For example, the high/low prices together are helpful because they describe the trading range of the period, and so plotting the high price alone does not fully capture the significance of the feature. Of all the features, volume is most appropriately visualized alone, because it is not part of a pair. Histograms are used to show the volume distribution for each inefficiency class along with their medians.



The median and mean are shown in black and yellow, respectively. The volatile class has the largest average mean and median volume. and the efficient class has the smallest mean and median values.

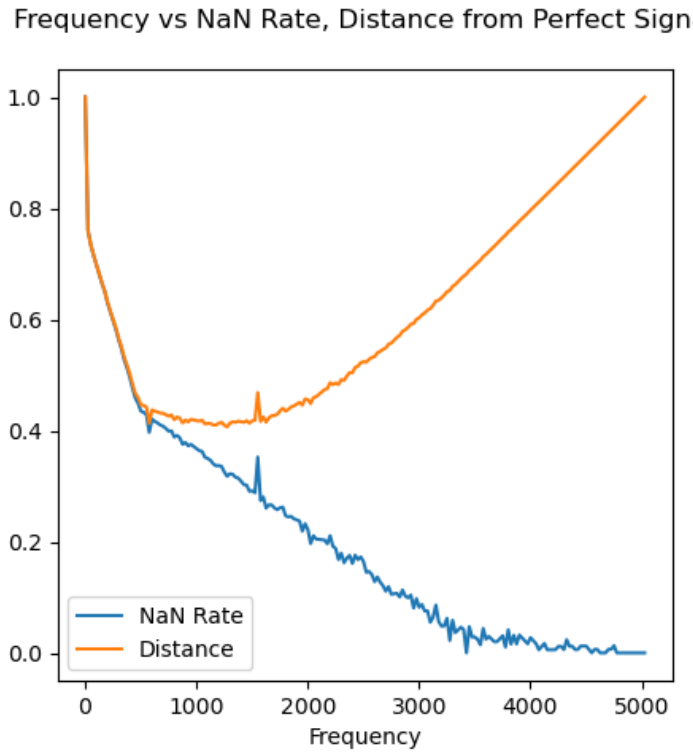
The records' price data can be effectively visualized using candlestick plots which displays all of the features together into one plot. The thicker candle part represents the open/close price bounds and the thin wicks represent the high/low price bounds. Traditionally, the sign of change in price for the period is represented by color, where a green candle means it closed *higher* than it opened, and a red candle means it closed *lower* than it opened. While this price-delta sign is not a feature of the data set (though it is captured in preprocess differencing), it is visualized below as an example of the standard stock visualization plot. Volume is also often visualized, scale-less, alongside the bottom of the candlestick figure with the same color coding.

Instead of coloring the candles according to the price-delta sign, they can be colored according to the inefficiency class of the record. Now we have a plot that includes all of the dataset's features including the target variable.



III. Data Preprocessing

The data is down sampled to a frequency that maximizes frequency rate while minimizing NaN rate. This is done by taking the (frequency, nan rate) pair closest to the 'perfect signal', which would be (0, 0) on the frequency-nan rate axis, representing infinitely fine resolution with no missing values. This value turns out to be 1276 minutes, which is approximately 21.27 hours and so the number of records in the relevant range is reduced to 616 records.



The data is split up chronologically with the first half as training, the next quarter as validation, and the last quarter as test. This is important because this timeseries dataset is not IID, for example the price the stock is trading at today is very dependent on what it was yesterday, and so the model is split strictly chronological so that it is trained on data that is collected before the data used to make the prediction. This emulates how the model would be deployed in production, because it would be impossible to use knowledge that is not available yet.

All the features are continuous and so a standard scaler is applied ubiquitously to standardize the features. Additionally, autoregression and differencing are used to make the records more stationary. Each one requires a lag time, and so this will be included in the hyperparameter CV. I imagine different frequencies may also yield different results and so this will be included in the

hyperparameter grid as well. Each one scales the number of features linearly and so the shape of the dataset goes from

Original (231171, 5) ->
    Preprocessed (231,171/downsample_frequency, 5 * lag1 * lag2)
    = (616, 20),
With a down sample frequency of 1276 and a lag of 2 for both autoregression and differencing.

IV. References

Muhammad Ali, Dost Muhammad Khan, Muhammad Aamir, Amjad Ali, Zubair Ahmad "Predicting the Direction Movement of Financial Time Series Using Artificial Neural Network and Support Vector Machine",
Complexity, vol. 2021, Article ID 2906463, 13 pages, 2021.
https://doi.org/10.1155/2021/2906463

Zhong, X., Enke, D. Predicting the daily return direction of the stock market using hybrid machine learning algorithms.
Financ Innov **5**, 24 (2019).
https://doi.org/10.1186/s40854-019-0138-0