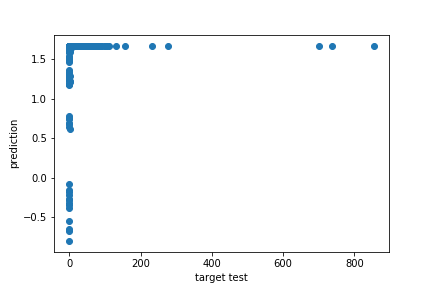
Jackson Poole

04.23.2019

ECON 498: Problem Set 1

In this data scraping and analysis exercise, I decided to continue exploring the data from coinmarketcap.com and uncover whether there was a distinct relationship between coin price volatility and volume. I hypothesized that coin brands with very high volumes would correlate with higher short-term volatility because of speculative practices among a larger population of users or traders. Because volume reflects the total amount of equity in the respective coin types, I thought this would be more appropriate than using supply, which is only measured in units. In measuring short-term volatility, the absolute value of the one-hour currency percent change (of price) was the best variable to use because it was the most reactive variable; it enabled us to see results even though the data was only collected over several days. To test this theory, I thought a linear regression model would be most appropriate, so I chose to use Sklearn’s linear model to help fit, predict, and analyze the scrapped data recorded every two hours.

After running the coinmarketcap.com data through Sklearn’s model, my results were startling, given my hypothesis and general expectations. I found that almost no correlation existed between the two variables. With the test size set at .25 and random state set at 0, the r2 score was 1.4531E-5. The correlation measure indicated there was almost no positive association among the variables at all like I had expected. Without fixing a random state, the r2 showed very similar results to the fixed state of zero with the r2 score fluctuating around 0. The following scatterplot depicts the algorithm’s prediction and the target test at a random state of zero.

 As we can see in the graph above, using only one variable, volume, it proved to be far too uncorrelated for the Sklearn algorithm to accurately predict hourly percent change. I believe the lack of variation in volume, especially at higher levels, contributed to this lack of correlation where we would expect to see at least some variation. Coins like Bitcoin may serve as disruptive outliers with very high volumes when a majority of the roughly two thousand coins have volumes which cluster around the lower end of the spectrum.

Furthermore, I noticed my dataset had a low degree of variation overall. One aspect contributing to this issue is the fact that the data was only gathered over several days instead of a longer time. Increasing the length of time scrapping and increasing the wait times between scrapping may lead to higher degrees of variation. Though the lack of variation was especially noticeable in variables like market cap and circulating supply (which may not be updated as often), percentage change over 24 hours or 7 days may yield a different result when looking at slightly longer-term volatility. I did not choose these longer measures because they did not reflect the very short-term volatility I wanted to test. Also, I decided not to use other algorithms like random forest because of the low degree of variation, and I did not think they could perform any better.

Even though my results did not yield an impressive linkage, the results actually are quite useful because they hint that hourly volatility is not dependent on the total volume. This insight can have significant consequences in how an individual investor might choose to allocate his or her capital based on risk preferences. If the individual is risk-averse, they might make very different choices if they understand short-term volatility is not correlated with the volume (like it might be in other investments).