Hi Sunny,

Thank you for all the work on our ongoing ML models project. I feel that we’re in a really good place with it right now – especially since we can now undertake some pretty well-informed feature evaluation once the new training data is ready (probably 2-3 more days… I’m processing it right now on my PC).

In the meantime, do you mind if we try and replicate the writeup from the below link using some of the same stock and CSI data?

<https://towardsdatascience.com/granger-causality-and-vector-auto-regressive-model-for-time-series-forecasting-3226a64889a6>

My goal is to be able to predict future values of the CSI Index. For example, I have exported 10+ years of CSI data and corresponding 10+ years of stock data. I’m hoping to segment out the last 15 (+/-) days of the CSI data and set that aside as “test” dataset. Then, using the method outlined in the article, predict the values of the 15 days and measure our error. If 15 days shows poor results, we could also reduce the amount of days. Or, since there are so many years worth of data, try and predict 5-10 days for every single year’s worth of data so we have enough attempts to see if the results are accurate / consistent or just lucky.

I made a quick Jupyter notebook that contains the following:

* Granger Causality function from the webpage
* CSI and stock data
* I ran the CSI and stock data through the function and saved the matrix as a dataframe (it took 1+ hour to run, so thought I’d save us some time here)

What I’m hoping to have replicated from the website is approximately the following:

1 - Can we recreate the stationary check?

2 - Can we use the results from the causality test (I ran this borrowing the function from the webpage) to aid our VAR model? In this case, we see there are about 150 out of the 500 stocks that influence “csi1” with a p value of less than 0.05 (reject null hypothes). My thinking is that these 150 can help the model predict the future values of csi1 while the remaining stock ticker data will just end up being noise. (further, maybe we reduce to only 10-30 of the best stocks with the lowest pvalue against csi1..?)

3 – Can we plot the residuals plot?

4 – Can we run the Durbin-Watson statistic on the data?

5 – Can we “Invert the transformation to get the real forecast”? (probably by using the last ~15 occurrences of the csi data as “test” data for accuracy verification)

6 – And finally, the evaluation step comparing the errors between the predicted values and the actual values (test data)