

Assignment #4

In this assignment, you will use the provided dataset to practice implementing some of the concepts we have discussed in class. You must submit your completed results and some brief commentary about your results and decisions via a Dropbox submission folder on the course D2L website. You must also submit your code (as separate files, not attached in the report) so that your results can be entirely recreated.

1. Explore the attached dataset and find additional information from the resources listed below. Note that this is the same dataset as was used in assignment 2.

- a) The dataset comprises NY Stock Exchange with several additional predictors, as explained in the following paper (especially sections 5 and 6):
<https://www.sciencedirect.com/science/article/abs/pii/S0957417419301915>
- b) The original dataset was obtained from:
<https://archive.ics.uci.edu/ml/datasets/CNNpred%3A+CNN-based+stock+market+prediction+using+a+diverse+set+of+variables>
- c) Although the original dataset includes 5 different files, we will only use the NYSE.csv file, which includes values from 2010 to 2017.

2. Hold out the last 3 months of 2017 for out-of-sample prediction and implement the following:

- a) Begin by forecasting with a single LSTM-layer and optimize its performance by varying the various hyperparameters, for example: LSTM units, no. of layers, batch size, learning rate, etc. Plot and analyze the model performance (accuracy and loss) vs. the number of epochs. You should also test different techniques to avoid model overfitting. Be sure to include and contrast examples of the different approaches – don't only show "the best". Plot the predictions and report the forecasting error using appropriate metrics.
- b) Repeat the forecasting using a 1D CNN-based forecasting model. Evaluate and optimize its performance similar to point 2 (a). This may require you to go beyond course material, but there are plenty of resources online that you can leverage.
- c) Now, implement a hybrid ConvLSTM forecasting model by combining the CNNs and LSTMs together. Evaluate and optimize its performance in comparison to 2 (a) and (b). Plot the predictions and report the forecasting error using appropriate metrics in comparison to 2(a) and (b).

3. Finally, compare your best model performance with the best statistical model found in Assignment 2 (ARIMA, SARIMA, VAR, etc.). Report and discuss your findings why certain models would outperform the others?