Data Collection

Prices are in Hong Kong Dollar.

We deleted the first row of every stock, because many of them had issues or did not actually have volume for that day.

Invalid Securities when we create a series of numbers: 824, 843, 849, 879, 890.

839 HK was removed because it contained insufficient data (started 2017)
804 HK was removed because it contained insufficient data (started 2015)

Problem: Volume seems to be an issue on a lot of stocks in the early days.
Solution: This probably is an issue from non-computer traded times. We are attempting to filter these days out.

Problem: Missing volume data. We need to decide how to handle this.
Solution: Drop days with missing volume data.

Problem: 899 has no open pricing for some of the data.
Solution: We can use the open of today to closely approximate

Problem: 865 has the same data for 300 rows with different dates.
Solution: We can start the data from after the repeating data.

The paper removed approximately 25 stocks from his testing set. We would like to keep as many as possible, so are only removing stocks which do not have enough data or have data which we simply cannot clean.

Technical Indicators

We used TA-Lib to easily build and replicate the technical indicators in the paper. This library allowed us to pipe in the data and get the indicators with just a few simple function calls.

Initial Model

As of now, we have begun to develop our model. We are using Keras (which is a front for TensorFlow) to build a basic model and run it on stock 833HK. We are currently using the open and close price and feeding this into a model with two LSTM layers and a single Dense layer. Before the data is fed in, it is being normalized. Once we have completed and tested our model, we can easily roll it out for all of the stocks and then add the Attention layer easily.