Quantitative Stock Forecast with Weather Data

BROWN

Andrew Kent
andrew_kent@brown.edu

Ilija Nikolov
ilija@brown.edu

Jikai Zhang
jikai_zhang@brown.edu

Introduction

The stock market is a complex (and even stochastic) system. Yet, there are certain world occurrences that have a major impact on it, such as weather events. As the ongoing climate crisis worsens, these disturbances will intensify. We aim to investigate the relationship between the weather data and the stock market in order to anticipate the influence on our financial system.

Since both weather and stock data are sequential, we shall work with CNN_LSTM-like models, first to forecast stock prices and then use climate data to refine these predictions. Climate data is mostly periodic, but by using multiple features, such as daily temperature, precipitation, wind direction and pressure, we hope to extract useful predictors.

Our project is mainly inspired by previous LSTM models that only use stock data to predict future price, or only use weather data to forecast the weather. We want to investigate the best way to combine the two and produce a better prediction than any individual model.

Methodology

The models start off with CNN layers to extract the main features from the stock and weather data, such as daily opening/closing price, highest/lowest price, volume and turnover. Then we have LSTM layers that learn sequentially based on the extracted features. Some of the models will have a sequential self-attention as an additional layer of feature cross interaction improvement.

The same models can be used for both the stock and the temperature data once the adequate changes are made. We will be comparing multiple models to see which one performs best, given the task in hand. The data is predicted based on a window size.

Our data preprocessing combines very different sets of data, financial and meteorological. The stock data is obtained from the Yahoo finance library, and the weather data is from the National Oceanic and Atmospheric Administration. The weather data is cleaned by dropping incomplete entries using pandas dataframes.

Prediction using Price ONLY

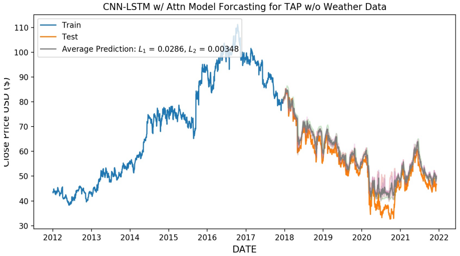


Figure 1. The full-length stock data, where 60% was used for training on a window size of 10 for Molson Coors, based only on stock data for the CNN-LSTM-SelfAttention model. The normalized squared loss for this model is 3.48e-3.

Prediction using Price and Weather

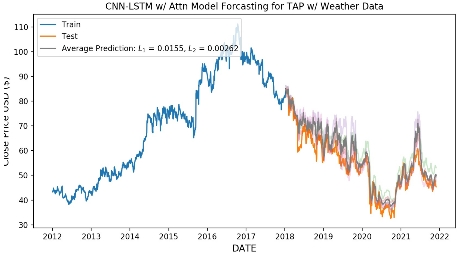
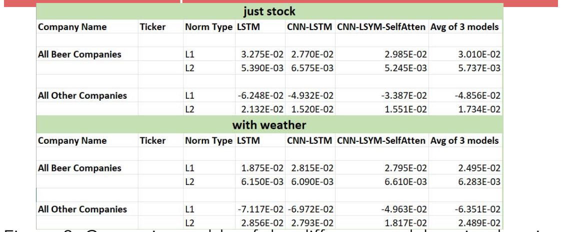


Figure 2. The full-length stock data, where 60% was used for training on a window size of 10 for Molson Coors, based on stock data together with weather data for the CNN-LSTM-SelfAttention model. The normalized squared loss for this model is 2.62e-3. This is lower than Figure 1, training only on the stock data alone, indicative of the success of including the weather data in the model.

Model Comparison



just stock						
Company Name	Ticker	Norm Type	LSTM	CNN-LSTM	CNN-LSYM-SelfAtten	Avg of 3 models
All Beer Companies		L1	3.275E-02	2.770E-02	2.985E-02	3.010E-02
		L2	5.390E-03	6.575E-03	5.245E-03	5.737E-03
All Other Companies		L1	-6.248E-02	-4.932E-02	-3.387E-02	-4.856E-02
		L2	2.132E-02	1.520E-02	1.551E-02	1.734E-02
with weather						
Company Name	Ticker	Norm Type	LSTM	CNN-LSTM	CNN-LSYM-SelfAtten	Avg of 3 models
All Beer Companies		L1	1.875E-02	2.815E-02	2.795E-02	2.495E-02
		L2	6.150E-03	6.090E-03	6.610E-03	6.283E-03
All Other Companies		L1	-7.117E-02	-6.972E-02	-4.963E-02	-6.351E-02
		L2	2.856E-02	2.793E-02	1.817E-02	2.489E-02

Figure 3. Comparison table of the different models trained on just stock price data, and on both stock data with weather data for different, companies. L1 is average distance between training and testing data, and L2 is the squared average distance.

As expected, the stock prices for companies that have some connection to the weather data we considered, which in this case are the beer companies, are predicted better when weather data is included, as compared to companies that do not have strong connections, such as the tech and finance industry.

Discussion

Our model has produced better results whenever training that includes weather data, as opposed to only training on data that only includes the stock price, for the relevant companies. In particular, our model saw improvement in the prediction for the stock price for agricultural companies and not tech & financial companies, which is what we have initially expected of our model.

A lingering problem is how to incorporate different models such that each one captures some of the most important features of the data, and passes them on to the next one, for overall higher perfromance goals. Possible exploration is the introduction of ensemble voting techniques that deals with similar problematic.

The stock prediction of tech companies did not improve much using weather data. To improve this prediction, one might use a more complicated model involving transformers to explore the volatile nature of these stocks. Or maybe think about different easily avaiable data that has some correaltion with tech stock price.