A news-based stock prediction/warning system

The project I was working on is about a stock prediction/warning system for investors who are interested in the semiconductor industry. Based on this system, I can first provide a reference for investors regarding the approximate increase/decrease values. Furthermore, I can help investors build up protection against possible trading curbs based on the information from tech news. For an investor, the biggest concern is to make money and avoid significant collapses in the market. We all know that tech news is highly related to the stock market circumstances and sometimes can even dominate the stock trend. Therefore, I need to delve into those news contents. By examining the historical stock data, we aim to determine which information from tech news can serve as a clear signal to predict an upcoming stock collapse in the semiconductor industry.

The project has 2 parts.

The first part focuses on machine learning and prediction for historical semiconductor industry stock data. In this part, I selected the historical stock value data for the ten largest semiconductor companies on Nasdaq (by market value) over the past five years. For each company, there are 'open,' 'close,' 'high,' and 'low' values for every business day. Each day, the stock values of all selected companies were aggregated and plotted, which served as the primary data for the subsequent analysis.

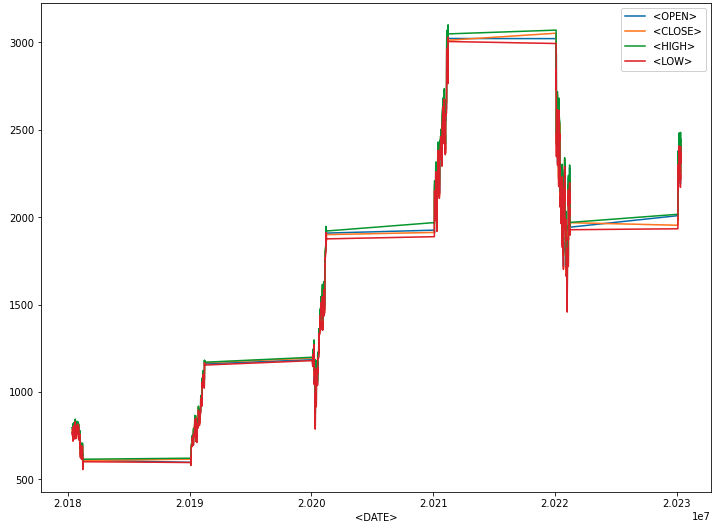


Fig.1. Selected semiconductor companies single-day total stock value (5 years)

To predict the direction of the stock for a new day, rolling averages for windows of 3, 5, and 7 days were calculated. Based on time series analysis, various models including Logistic Regression, Decision Tree, Random Forest, and Gradient Boosting were applied to the train/test data, yielding the highest AUC score of 0.55.



Fig.2. AUC graphs for every model on stock value time series data

The second part focuses on an NLP-based stock collapse warning system. In this part, the news data is parsed from techcrunch.com, which includes all news titles from each day over the past five years. For the days when there is a significant semiconductor stock collapse (>=1%), the news titles were specifically selected and compiled. NLP techniques such as tokenizing and word frequency counting were applied to extract word information. Finally, a special dictionary was constructed, comprising the most frequently occurring single and double words in the news titles preceding the collapses. In the future, each word within the dictionary can serve as a warning alarm for a potential upcoming semiconductor stock collapse.



Fig.3. Dictionary of stock collapse warning word

As a result, in addition to predicting the stock value of the semiconductor industry, the project also compiled a summary of the ten most frequently appearing tech news words that may be highly correlated with collapses. This information can assist semiconductor investors in making decisions for their next steps.