WASHINGTON STATE UNIVERSITY
School of Economic Sciences

2019 NBER Big Data and High-Performance Computing for Financial Economics Proposal

Predicting Stock Index Movement from News Headline, VIX, Historical Prices

Jikhan Jeong, WSU Economics Ph.D. student, E-mail: jikhan.jeong@wsu.edu

Research objectives

This research work will apply machine learning approaches to predict the daily stock market by using news headlines, VIX, and historical stock prices. This study will provide insight into how to use digitalized big data from unstructured text data and interaction terms between news and VIX to predict the financial market. In particular, the stock market index reflects the future expected financial performance as a leading composite index in economics. Therefore, this study can help us to predict future short-term movements of the financial market.

Research Background

The stock market has high volatility and interacts with diverse external factors. In spite of this complexity, many scholars have tried to predict the future stock market. Wong and Ko (2016) used public emotion from news articles to predict the movement of the stock market index. Kavšek (2017) applied diverse machine learning methods to predict the Dow Jones Industrial Average (DJIA) by using Reddit WorldNews data. Oncharoen, Pisut and Vateekul (2018) predicted the DJIA and S&P 500 by applying deep learning methods including the Convolutional Neural Network (CNN) and Long-Short Term Memory (LSTM) with historical prices, financial technical indicators, and news data. Velay, Marc and Daniel (2018) demonstrated that logistic regression has the highest predictive power compared to other machine methods when forecasting stock index movements using news data. In this paper, the authors applied Word2Vec and Bag of Words methods as the embedding methods. Above all, the papers focused on finding the most predictive machine learning methods for the stock market while using news data. However, the effect of autocorrelation and uncertainty in the stock market on the prediction of future stock prices has not been closely studied. Zhang (2016) showed that the degree of information uncertainty influences the prediction of future stock returns. Mian and Sankaraguruswamy (2012) also found that the investors' sentiments affect the investors' responses to earnings news. In addition, it is possible that the degree of predictive power obtained when using the news differs among sectors because of the different business characteristics and environment in each sector. Therefore, this study will use machine learning methods to predict the DJIA, S&P 500, and sectoral indices. In addition, this study also will cover 500 firm-level stock returns in the S&P 500 to determine how firm-level differences, such as the share of oversee business, may influence the predictive power of the stock return. To predict stock indices and returns, news data, lagged prices based on autocorrelation, VIX, and the interaction term between news and VIX will be used. Above all, this study will reveal the influence of sectoral characteristics and the expected volatility in the stock market, i.e. VIX, on the prediction of the short-term stock market.

Reddit WorldNews will be used as a real-time news source because of its popularity; this analysis will be based on the top 25 headlines in Reddit WorldNews, as chosen by Reddit users in real time. In this regard, this study assumes that the most popular headlines reflect the most interesting and influential topics for a given day for Reddit users. The source of data on Reddit WorldNews is Kaggle, and financial data will be collected from Wharton Research Data Services (WRDS). In order to diversify the source of news headlines and increase the size of the data, another data set from "All the News (Andrew Thompson,2018)" will be used. This dataset contains 204,135 articles that were published from 2013 to early 2018 (the size of this dataset is 1.5gb). Regression-based-machine learning approaches, including regularized linear regressions, Random Forest, and Xgboost, will be applied. To begin, a pilot test using L1, L2 regularized logistic regressions was conducted to predict the ups and downs of the DIJA. In addition, as an embedding method, Word2Vec was conducted as follows:

Figure 1. (left) DIJA prediction from Reddit data (Right) Word2Vec Result

Test prediction accuracy of Logistic Regression (LR)

Prediction Methods	Test Prediction Accuracy
L2 regularized LR	46.8%
L1 regularized LR	48.5%
L2 regularized LR with hyperparameter tuning	54.7%
L1 regularized LR with hyperparameter tuning	**57.9%**

The similar words from "Stock" after Word2Vec

```
In [4]: model.wv.most_similar("stock")
Out[4]:
[('avoidance', 0.9943307638168335),
 ('fleet', 0.9930078983306885),
 ('settlement', 0.9927743077278137),
 ('surge', 0.9921797513961792),
 ('beijings', 0.9919998645782471),
 ('eviction', 0.9909545183181763),
 ('underwater', 0.990685224533081),
 ('mubaraks', 0.9906383752822876),
 ('concentration', 0.9904321432113647),
 ('candidates', 0.990141749382019)]
```

The computing resource and language to be used for the analysis

High Performance Computing (HPC): Kamiak, the HPC Cluster in Washington State University, will be used for this study. Three different partitions in Kamiak including cahnrs, cahnrs_bigmem, and cahrnrs_gpu will be available..

Figure 2: CAHNRS QUEUE and Hardware Specifications

CAHNRS

Queue Name	Node Count	Accelerators per Node	CPU Cores per Node	Memory per Node	Wall Clock Limit per Job	Max CPU Cores per User	Max Memory per User
cahnrs	11	None	20	256GB	7-00:00:00	120	1.5TB
cahnrs_bigmem	1	None	60	2TB	7-00:00:00	N/A	N/A
cahnrs_gpu	1	2 NVIDIA Tesla K80 (4 GPUs total)	24	256GB	7-00:00:00	N/A	N/A

reference: https://hpc.wsu.edu/kamiak-hpc/queue-list/

Computing languages: Python, R, Linux, SQLite

Reference

Branko Kavšek, "Using Words from Daily News Headlines to Predict the Movement of Stock Market Indices," Managing Global Transitions, University of Primorska, Faculty of Management Koper, vol. 15(2 (Summer), pages 109-121, 2017

C. Wong and I. Ko, "Predictive Power of Public Emotions as Extracted from Daily News Articles on the Movements of Stock Market Indices," 2016 IEEE/WIC/ACM International Conference on Web Intelligence (WI), Omaha, NE, pp. 705-708, 2016

G. Mujtaba Mian and Srinivasan Sankaraguruswamy , "Investor Sentiment and Stock Market Response to Earnings News.", The Accounting Review, Vol. 87, No. 4, pp., 2012

Oncharoen, Pisut and Peerapon Vateekul. "Deep Learning for Stock Market Prediction Using Event Embedding and Technical Indicators.", 5th International Conference on Advanced Informatics: Concept Theory and Applications (ICAICTA): 19-24., 2018

Velay, Marc and Fabrice Daniel. "Using NLP on news headlines to predict index trends." CoRR abs/1806.09533, 2018

Zhang, Frank, "Information Uncertainty and Stock Returns". Journal of Finance, Vol. 61, No. 1, 2006