Can Search Trend Forecast Stock Returns?

Hyun Ki Kim

The University of Chicago

June 6, 2018

Abstract

This study was undertaken to find arbitrage opportunity using information from search rankings in a major financial website. While stock markets are regarded as efficient in the long run, there is evidence suggesting there are short time-intervals in which prices remain inefficient. As stock markets maintain discrete operating hours, information that arises after markets close will affect prices at the next opening. This paper uses overnight search ranking data in the finance section of Naver, which is the largest web-portal in Korea, to measure investors' attention. Using 8 different machine learning methods, 4 models had higher cumulative returns compared to the market average. Unlike previous studies on financial forecasting, more complex machine learning methods did not have superb performance than the relatively simpler methods. However, the results are highly sensitive to market movement and the random state. Therefore, further research is needed to decompose high variance of the forecast.

I. Introduction

Research on asset pricing theory is based on the random walk hypothesis and the Efficient Market Hypothesis (Fama, 1965). Scholars in this literature assume that every information in the market is instantaneously reflected in the market prices. Although information might occur at random, price does not get adjusted instantaneously in practice. Busse and Green (2002) found that there is a short time interval when stock price remains non-efficient. Furthermore, recent research found evidence of Big Data trend preceding stock market movement (Bollen et al., 2011; Da et al., 2011; Moat et al., 2013; Preis et al., 2013). This implies that there could be a temporary arbitrage opportunity by incorporating enormous number of digital footprints.

Recently, scholars have used online behaviors to predict real-world behaviors. Scholars have found that unemployment (Ettredge et al., 2005) and influenza (Ginsberg et al., 2009; Polgreen et al., 2008) related searches can reveal situations earlier than the government reports. Choi and Varian (2012), for example, drew on the data from Google Trend to predict initial claims for unemployment, automobile demand, and vacation destinations.

In the field of finance, Da et al. (2011) are among the first to use search volume data to measure investors' attention. They analyzed ticker symbol searching patterns in Google Trend and found that the increase in searching behavior is correlated with the increase in stock price 2 weeks later. Recently, many researchers attempt to use internet-based data including Google Trend (Preis et al., 2013), Twitter (Bollen et al., 2011), and Wikipedia (Moat et al., 2013) to forecast stock market.

Although many studies found statistically significant outcomes, higher frequency trading strategy would fare better if the mispricing is motivated by investors' attention. Although the Efficient Market Hypothesis assumes that additional information is immediately incorporated into

the price, this adjustment does not happen instantaneously in reality. Studies have found these price adjustments can take from 15 seconds to 30 minutes (Busse and Green, 2002; Chordia et al., 2005).

People who look up stock tickers in search engines are more likely to be individual investors since institutional investors have access to more sophisticated software such as Bloomberg terminals (Da et al., 2011). Odean (1999) proposed that individual investors choose to focus on attention-grabbing stocks as they are unable to evaluate thousands of stocks that they can potentially purchase. However, this is less problematic when they are selling a stock because most individual investors do not sell short. Thus, Barber and Odean (2008) proposed that individual investors are net buyers of attention grabbing stocks. Therefore, an increase in aggregate demand caused by attention will raise stock price since aggregate supply is inelastic in the short run.

II. Data

The Korea Exchange (KRX) is the world's 13th largest stock market by market capitalization (1,683 billion USD) and 9th largest stock market by monthly trade volume (142 billion USD) as of October 2017. Its regular trading hours are from Monday through Friday from 9:00 a.m. to 3:30 p.m. local time excluding holidays, but before and after-hours trading is available from 7:30 a.m. to 6:00 p.m. Stocks in the KRX are traded in two market divisions, which is the KOSPI (The Korea Composite Stock Price Index) and the KOSDAQ (The Korean Securities Dealers Automated Quotations). The KOSPI is regarded as the representative of nearly 800 large publicly traded companies in Korea, and thus is equivalent to the Dow Jones Industrial Average in the U.S. market, and the KOSDAQ is regarded as the representative of nearly 1,200 small and medium sized companies, and thus is equivalent to the NASDAQ in the U.S. market.

Naver is the largest search engine in Korea that accounts for more than 70% of country's domestic search volume. The web portal provides services such as email, blog, news, finance, map, music and many others in addition to its main service of searching. One of the most unique features of Naver is its real-time ranking. Its main page provides top 20 search key words that people are searching at each moment. Likewise, the finance section of the Naver provides top 30 stocks that people are searching at each time.

This paper analyzes people's searching behavior to forecast stock price fluctuation. In order to measure individual's attention, search rankings in the Naver finance section were used from April 3rd, 2018 to April 27th, 2018. Ranking tables, which includes stock name, and the relative search volume were scraped were scraped every minute (1,440 times per day) throughout this period. However, search ranking is highly correlated with a price fluctuation during trading hours. Therefore, search ranking when the market is closed (from 6:01 pm to 7:29 am) was used to predict stock return day after. While market is closed (799 minutes) there were on average 288 unique stocks appeared in the search ranking. Table 1 shows the number of unique stocks in the search ranking by day of the week. It shows that Saturday-Sunday and Sunday-Monday have more diverse searching pattern as there are no trading hours on that day that focus attention on the stocks with high price fluctuations.

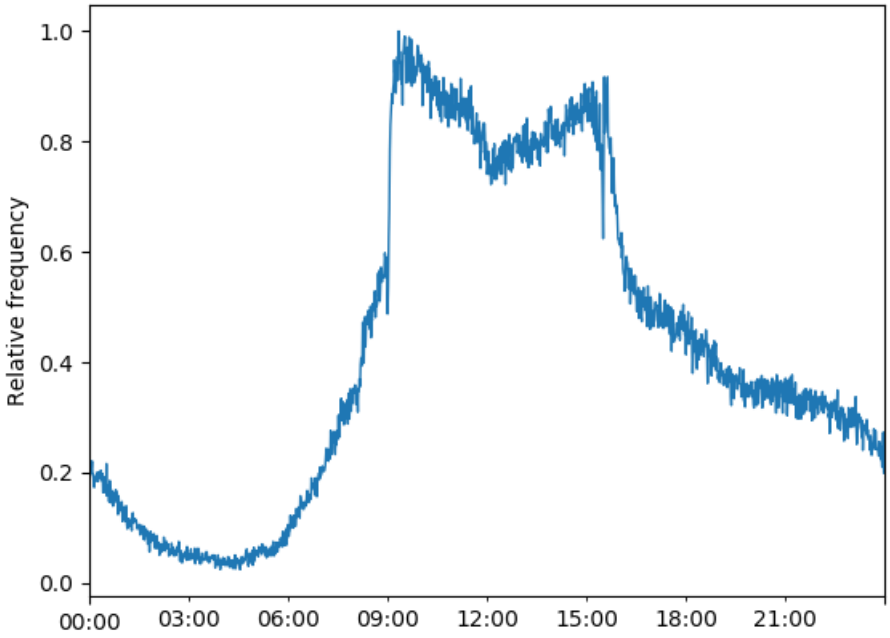
Table 1. Number of unique stocks by day of the week

Sunday - Monday	Monday - Tuesday	Tuesday - Wednesday	Wednesday - Thursday	Thursday - Friday	Friday - Saturday	Saturday - Sunday
317	256	277	260	258	304	386

Note: Observations from April 8th 6:01 p.m. to April 9th 7:29 a.m. were excluded due to technical difficulties.

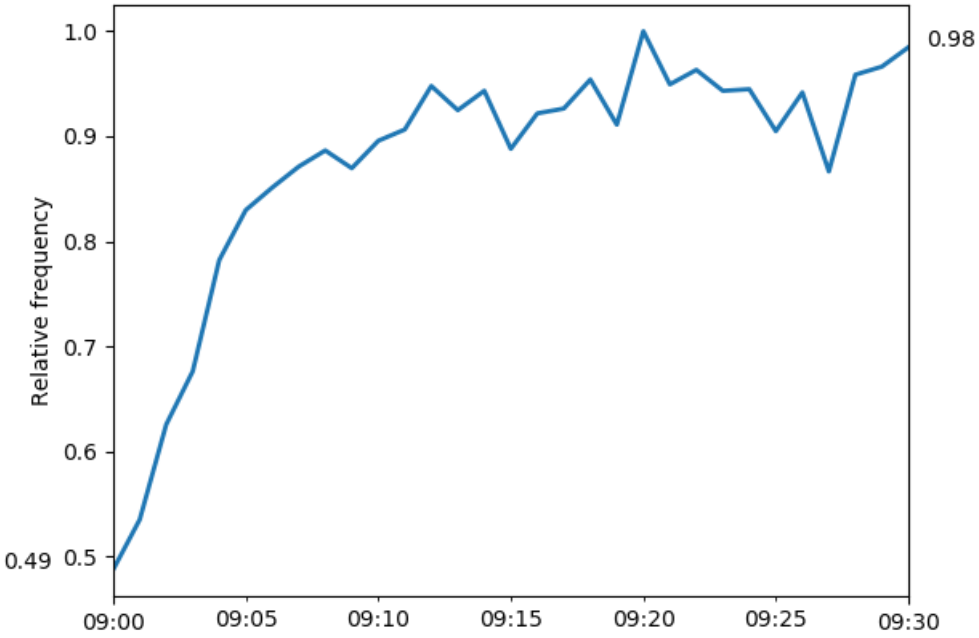
However, it is unlikely that investors pay exactly the same amount of attention to the search ranking throughout the day. Therefore, number of discussion posts posted on the discussion forum was chosen to measure the popularity of the stock in each minute. Naver finance has a separate page for each stock with sections with graph, news, and the discussion forum. Figure 1 shows the minute level count of posts posted on the discussion forum of stocks listed in the KOSPI and the KOSDAQ on weekdays from April 2nd to April 27th, 2018.

Figure 1. Number of posts within a day



Users of Naver finance post more while market is open, especially on near market opening and closing. It also shows U-shape trend from market closing to next market opening with trough around 4 a.m. The largest surge in attention happens around 9:00 a.m., which is right after market opening. Figure 2 shows how the attention doubles from 9:00 a.m. to 9:30 a.m.

Figure 2. Number of posts from 9:00 a.m. to 9:30 a.m.



III. Methodology

Forecasting stock price is regarded as a challenge because stock market is affected by domestic and international political events, macroeconomic conditions, and investors' sentiment. Therefore, a number of machine learning techniques have been widely adopted in financial forecasting to capture such underlying, highly nonlinear processes. This paper focuses on the stock returns in the first 30 minutes after market opening as the evidence suggests that it is the most critical time period when information spreads throughout the market. 8 different machine learning algorithms, namely Bagging, Boosting, K-Nearest Neighbors, Linear Discriminant Analysis, Logistic Regression, Neural Networks, Random Forest, and Support Vector Machines were used to forecast stock returns.

Especially, stock returns between 9:02 a.m. and 9:30 a.m. were classified to compare the performance of each algorithms. Data collection period was split into 3 weeks of training period (April 4th, 2018 to April 20th, 2018) and 3 weeks of testing period (April 23rd, 2018 to May 11th, 2018). However, data on April 9th, 2018 were excluded due to technical difficulties and May 1st, 2018 and May 7th, 2018 were excluded duet to national holiday. 9:02 a.m. was chosen for the buying time to give algorithms 2 minutes period to understand the direction of movement after the market opening. However, setting buying time too late would lose the opportunity to grasp price increases coming from increases in demand.

Returns between 9:02 a.m. and 9:30 a.m. were transformed into classification problem, which divides stocks into those increased more than cutoff versus those did not increase. Setting low cutoff value would include stocks that are naturally fluctuating into the portfolio, rather than selecting stocks that shows increases in demand. However, setting high cutoff value would make sample size for classification to shrink. Therefore, 2%, which includes about 10% of training set was chosen as the cutoff value to compare the models. In total, 32 independent variables were given to each algorithm with default setting in the Scikit-learn package being used without any tunings for parameters. Table 2 shows the description of variables used in the model.

Table 2. Description of independent variables

Name	Description
Last Closing	Closing price for last trading day
Price 09:00	Opening price
Price 09:01	Price at 09:01
Price 09:02	Price at 09:02

Percent 09:00	Percentage increase of opening price compared to last closing
Percent 09:01	Percentage increase of price 09:01 a.m. compared to last closing
Percent 09:02	Percentage increase of price 09:02 a.m. compared to last closing
Weight	Sum of overnight search proportion times the weight at each time
KOSPI	Dummy variable equaling 1 if in the KOSPI, 0 if in the KOSDAQ
Market last closing	Closing market index for last trading day
Market 09:01	Market index at 09:01 a.m.
Market 09:02	Market index at 09:02 a.m.
Percent market 09:01	Percentage increase of market 09:01 compared to market closing
Percent market 09:02	Percentage increase of market 09:02 compared to market closing
Alpha 09:01	Percentage point difference between price and market at 09:01
Alpha 09:02	Percentage point difference between price and market at 09:02
Price maximum	Dummy variable equaling 1 if price met price ceiling (+30%)
Price minimum	Dummy variable equaling 1 if price met price floor (-30%)
Price volatility	Maximum divided by minimum price between 09:00 and 09:02
Price trend	Increase count for price between 09:00 and 09:02
Average price volatility	Price trend times price volatility divided by 2
Market volatility	Maximum divided by minimum index between 09:00 and 09:02
Market trend	Increase count for index between 09:00 and 09:02
Average market volatility	Market trend times market volatility
Volatility ratio	Price volatility divided by market volatility
Average volatility ratio	Average price volatility divided by average market volatility
Average price volatility^2	Average price volatility squared

Average market volatility^2	Average market volatility squared
Average volatility ratio^2	Average volatility ratio squared
Opening increase	Dummy variable equaling 1 if price opening increased
Market opening increase	Dummy variable equaling 1 if market opening increased
Both opening increase	Dummy variable equaling 1 if price and market opening increased

IV. Results

As a baseline for the comparison, the KOSPI had +0.66% and the KOSDAQ had -1.37% cumulative returns from 9:02 a.m. to 9:30 a.m. during 3 weeks of testing period. For the models that depend on the random state, returns were calculated by averaging the cumulative return for 100 simulations at each time. Table 3 shows the returns for 8 models after 3 weeks of testing period.

Table 3. Cumulative return for each machine learning algorithm

Model	Return	Standard deviation	Number of trades
Linear Discriminant Analysis	0.5365	-	73
Bagging	0.1471	0.26	62.43
Random Forest	0.0862	0.21	45.11
Logistic Regression	0.0304	-	4
Support Vector Machines	-0.0211	0.45	352.33
K-Nearest Neighbors	-0.0457	-	27
Boosting	-0.0751	0.06	34.32
Neural Networks	-0.2117	0.46	244.34

Out of total 8 models, 4 models outperformed the market indices after all. Linear Discriminant Analysis model had the best performance with 54% increase after 3 weeks of testing period. The result disagreed with consensus that Support Vector Machines and Neural Networks are among the best in classifying stock returns. Also, most models performed better during the second half of the testing period, suggesting that models are using similar characteristics of market. However, as shown in the standard deviation of the models that depend on random state and are not a reliable strategy. It suggests access returns could in fact be originating from a risk that models are taking.

V. Conclusion

It is generally assumed that all the information on the market is reflected in the stock prices and therefore it is efficient. However, this paper suggests that it is possible to have higher expected returns than the market average, by using machine learning algorithms that capture non-linear relationship. The results disagree with existing literature that compares different machine learning algorithms and claims relatively complex model such as Support Vector Machines and Neural Networks are better at forecasting stock returns. Using 3 weeks each for training and testing period, Linear Discriminant Analysis, Bagging, Random Forest, and Logistic Regression showed better performance than the market average.

Nonetheless, there are a number of limitations of this study. As shown in the standard deviation of the results, models that depends on the random state have large variance in the outcome. This shows that the performance of each model is not a reliable forecast for the stock market and, thus, it is not possible to conclude that models outperformed the market, considering risk-return trade off. Risk adjustment for returns is suggested for further research.

References

Barber, B. M., & Odean, T. (2008). All that glitters: The effect of attention and news on the buying behavior of individual and institutional investors. *Review of Financial Studies,21*, 785–818.

Bollen, J., Mao, H., & Zeng, X. (2011). Twitter mood predicts the stock market. *Journal of Computational Science,2*(1), 1-8.

Busse, J. A. and Green, T. C. (2002). Market efficiency in real time. *Journal of Financial Economics,65*(3), 415-437.

Choi, H., & Varian, H. (2012). Predicting the Present with Google Trends. *Economic Record,88*, 2-9.

Chordia, T., Roll, R. W., & Subrahmanyam, A. (2005). Evidence on the Speed of Convergence to Market Efficiency. *Journal of Financial Economics,76*(2), 271-292.

Da, Z., Engelberg, J., & Gao, P. (2011). In Search of Attention. *The Journal of Finance,66*(5), 1461-1499.

Ettredge, M., Gerdes, J., & Karuga, G. (2005). Using web-based search data to predict macroeconomic statistics. *Communications of the ACM,48*(11), 87-92.

Fama, E. F. (1965). The Behavior of Stock-Market Prices. *The Journal of Business,38*(1), 34-105.

Ginsberg, J., Mohebbi, M. H., Patel, R. S., Brammer, L., Smolinski, M. S., & Brilliant, L. (2009). Detecting Influenza Epidemics Using Search Engine Query Data. *Nature,457*, 1012-1014.

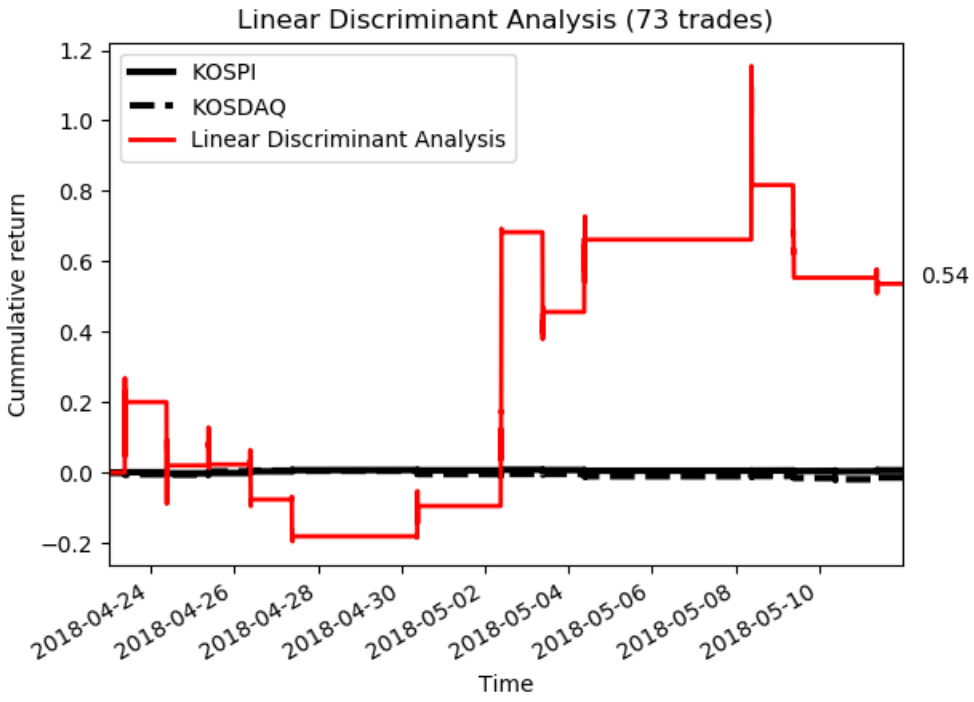
Moat, H. S., Curme, C., Avakian, A., Kenett, D. Y., Stanley, H. E., & Preis, T. (2013). Quantifying Wikipedia Usage Patterns Before Stock Market Moves. *Scientific Reports,3*(1).

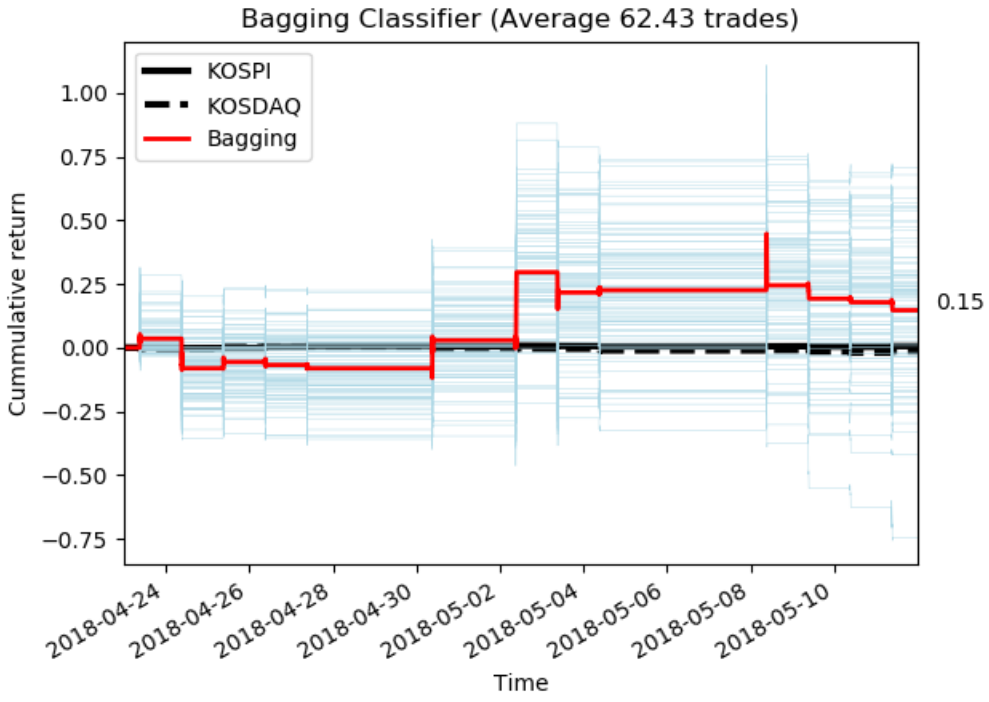
Odean, T. (1999). Do Investors Trade Too Much? *American Economic Review,89*(5), 1279-1298.

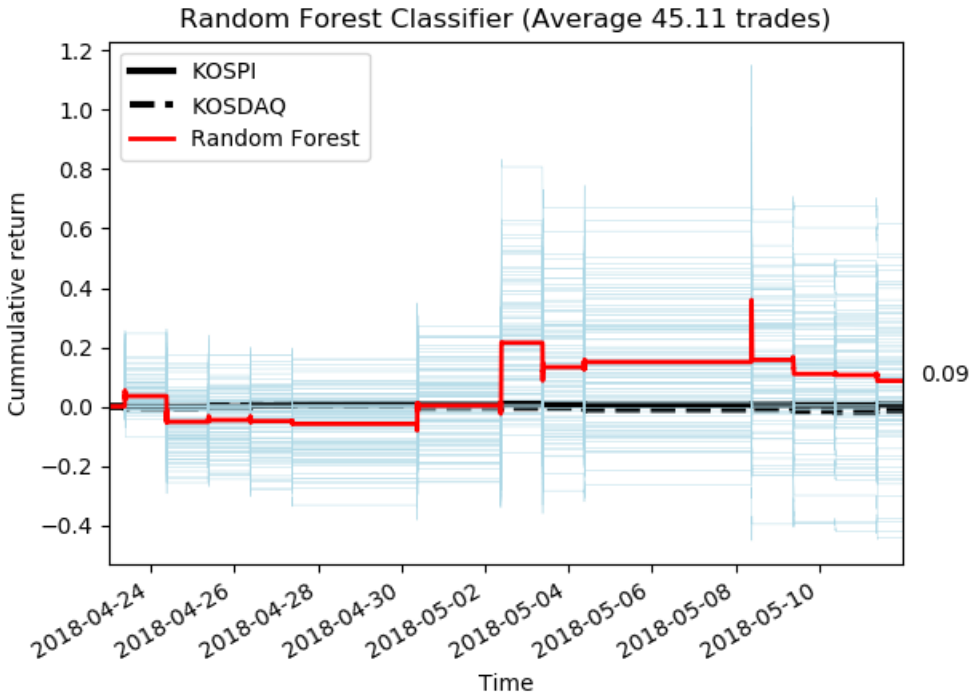
Polgreen, P., Chen, Y., Pennock, D., & Nelson, F. (2008). Using Internet Searches for Influenza Surveillance. *Clinical Infectious Diseases,47*(11), 1443-1448.

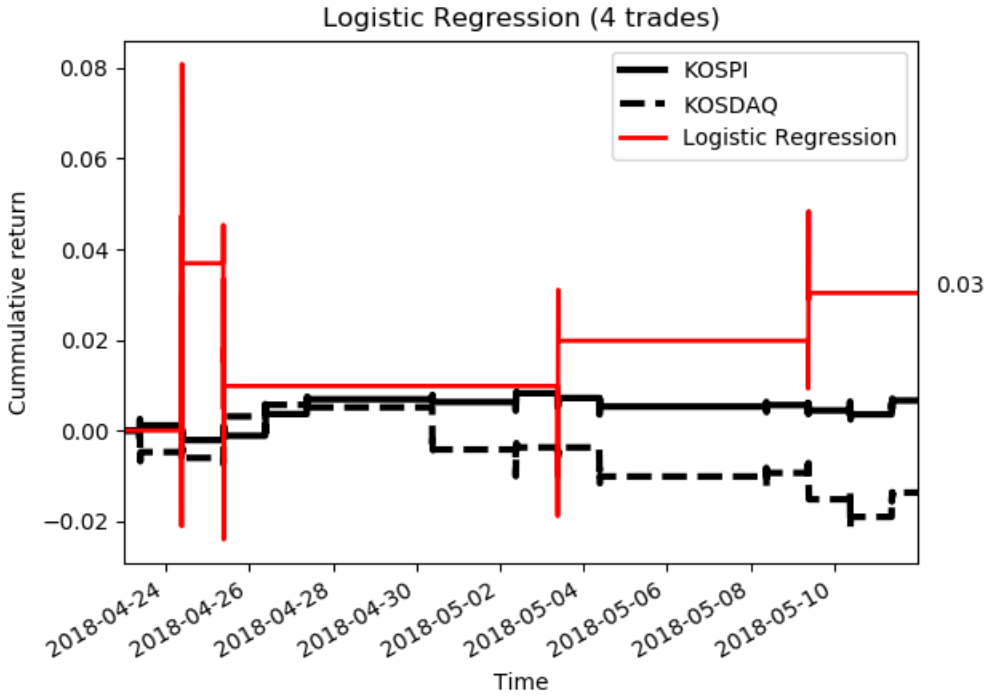
Preis, T., Moat, H. S., & Stanley, H. E. (2013). Quantifying Trading Behavior in Financial Markets Using Google Trends. *Scientific Reports,3*(1).

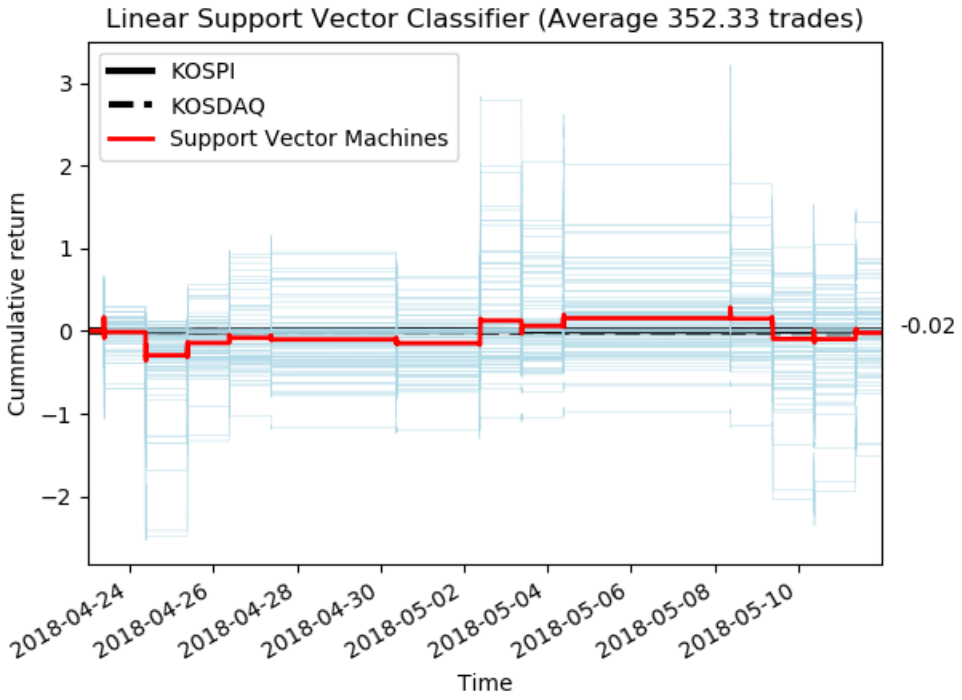
Appendix A. Time series plot of cumulative return

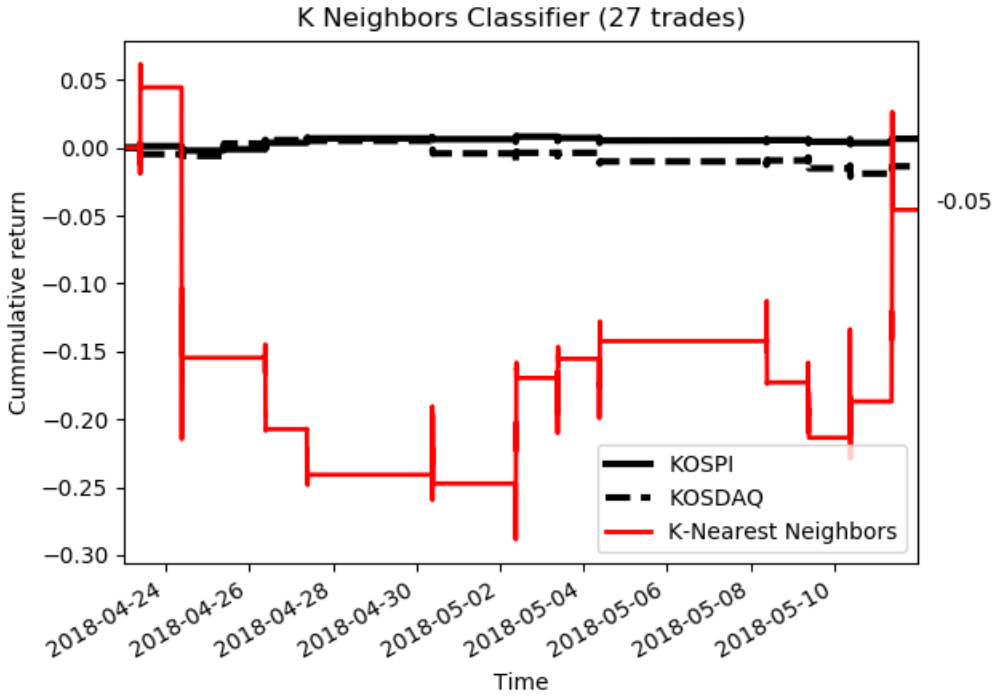


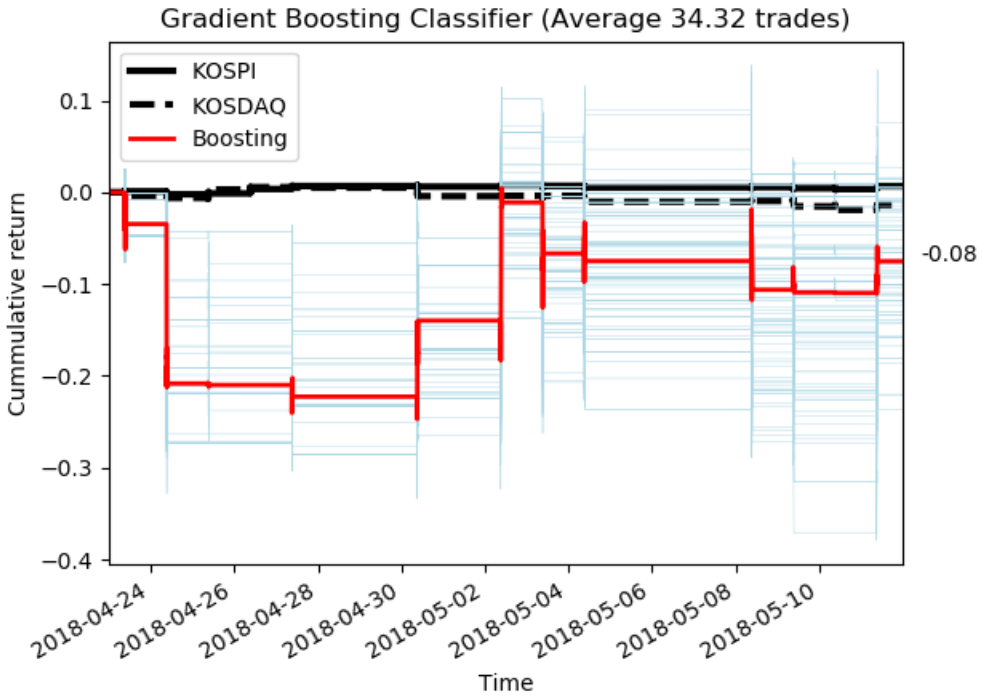


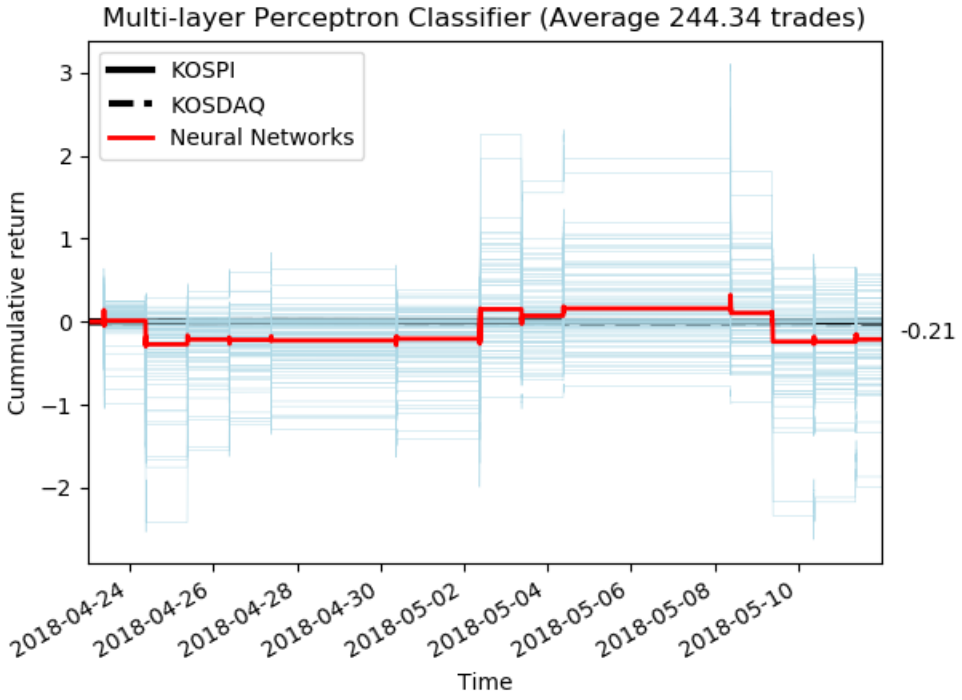

Random Forest Classifier (Average 45.11 trades)
KOSPI
KOSDAQ
Random Forest
Cummulative return
1.2
1.0
0.8
0.6
0.4
0.2
0.0
−0.2
−0.4
0.09
2018-04-24
2018-04-26
2018-04-28
2018-04-30
2018-05-02
2018-05-04
2018-05-06
2018-05-08
2018-05-10
Time


Logistic Regression (4 trades)
KOSPI
KOSDAQ
Logistic Regression
Cummulative return
0.08
0.06
0.04
0.02
0.00
−0.02
0.03
2018-04-24
2018-04-26
2018-04-28
2018-04-30
2018-05-02
2018-05-04
2018-05-06
2018-05-08
2018-05-10
Time


Linear Support Vector Classifier (Average 352.33 trades)
KOSPI
KOSDAQ
Support Vector Machines
Cummulative return
3
2
1
0
−1
−2
-0.02
2018-04-24
2018-04-26
2018-04-28
2018-04-30
2018-05-02
2018-05-04
2018-05-06
2018-05-08
2018-05-10
Time


K Neighbors Classifier (27 trades)
Cummulative return
0.05
0.00
−0.05
−0.10
−0.15
−0.20
−0.25
−0.30
-0.05
KOSPI
KOSDAQ
K-Nearest Neighbors
2018-04-24
2018-04-26
2018-04-28
2018-04-30
2018-05-02
2018-05-04
2018-05-06
2018-05-08
2018-05-10
Time


Gradient Boosting Classifier (Average 34.32 trades)
KOSPI
KOSDAQ
Boosting
Cummulative return
0.1
0.0
−0.1
−0.2
−0.3
−0.4
-0.08
2018-04-24
2018-04-26
2018-04-28
2018-04-30
2018-05-02
2018-05-04
2018-05-06
2018-05-08
2018-05-10
Time


Multi-layer Perceptron Classifier (Average 244.34 trades)
KOSPI
KOSDAQ
Neural Networks
Cummulative return
3
2
1
0
−1
−2
-0.21
2018-04-24
2018-04-26
2018-04-28
2018-04-30
2018-05-02
2018-05-04
2018-05-06
2018-05-08
2018-05-10
Time