The Korean Journal of Applied Statistics
2021, Vol. 34, No. 5, 745–760

DOI: https://doi.org/10.5351/KJAS.2021.34.5.745
Print ISSN1225-066X / Online ISSN 2383-5818

Prediction of the direction of stock prices by machine learning techniques

Yonghwan Kim[a], Seongjoo Song[1,a]

[a]Department of Statistics, Korea University

Abstract

Prediction of a stock price has been a subject of interest for a long time in financial markets, and thus, many studies have been conducted in various directions. As the efficient market hypothesis introduced in the 1970s acquired supports, it came to be the majority opinion that it was impossible to predict stock prices. However, recent advances in predictive models have led to new attempts to predict the future prices.

Here, we summarize past studies on the price prediction by evaluation measures, and predict the direction of stock prices of Samsung Electronics, LG Chem, and NAVER by applying various machine learning models. In addition to widely used technical indicator variables, accounting indicators such as Price Earning Ratio and Price Book-value Ratio and outputs of the hidden Markov Model are used as predictors. From the results of our analysis, we conclude that no models show significantly better accuracy and it is not possible to predict the direction of stock prices with models used. Considering that the models with extra predictors show relatively high test accuracy, we may expect the possibility of a meaningful improvement in prediction accuracy if proper variables that reflect the opinions and sentiments of investors would be utilized.

Keywords: prediction, machine learning, classification, stock price

1. 서론

금융시장 참여자들에게 있어서 주식 가격의 움직임은 직접적인 수익으로 이어질 수 있기에 매우 중요하다. 주식가격이 어느 방향으로 움직일 지를 잘 예측할 수 있다면 안정적으로 높은 수익을 얻을 수 있을 것이므로, 주가에 대한 예측은 학계 및 업계의 많은 관심을 받아왔다. 과거에는 주식 가격의 예측이 불가능하다는 의견이 많았는데, 예를 들면 Cowles (1933)는 당시 24개의 금융회사의 전문가들에게 6개월 간격으로 투자종목을 추천받아 투자를 진행했지만, 무작위로 선택한 것 보다 통계적으로 유의하게 좋은 결과를 얻지 못했다. Fama (1970)는 주식시장이 효율적 시장 가설을 따른다고 주장하였다. 효율적 시장 가설이란 주식의 가격이 모든 정보를 이미 반영하고 있다는 것으로, 주식시장에서 얻을 수 있는 정보를 이용하여 시장포트폴리오보다 더 나은 수익을 일관되게 얻을 수 없으며 주식 가격의 예측은 불가능하다는 것을 함축한다. 여기서 더 나아가 Umstead (1977)는 주식 시장의 예측을 위한 연구가 진행되어 예측에 성공한다 하더라도, 이 정보가 시장 참여자에게 알려져 효율적 시장 가설을 더욱 잘 따르게 할 것이고, 예측에 도움이 되는 설명변수의 발견은 다시 무의미하게

This work is based on the Master thesis of the first author, Yonghwan Kim, from Korea University.
This work was supported by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIT) (No. 2021R1F1A1048986).
[1] Corresponding author: Department of Statistics, Korea University, 145 Anam-Ro, Seongbuk-Gu, Seoul 02841, Korea. E-mail: sjsong@korea.ac.kr

Published 31 October 2021 / journal homepage: http://kjas.or.kr

될 것이라 주장하였다. 이밖에도 Bussière (2013)는 선형 모형으로는 금융시장의 변동성을 잡아내지 못하고, 외표본(out-of-sample)의 금융 위험 예측이 불가능하다고 주장하였다.

한편, 최근 기술의 발달로 인해 새로운 방법론을 이용하여 주식 시장에서 미래가격을 예측해보고자 하는 시도들이 이어지고 있다. Dsouze와 Mallikarjunappa (2015) 그리고 Durusu 등 (2019)은 주식시장의 효율성을 부정하고 주가의 예측가능성을 제시하였다. 이 외에도 주식 가격을 직접 예측하여 높은 예측력을 제시한 연구들도 많이 있는데, 이는 2절에서 소개하도록 하겠다. 새로운 방법론을 이용하면 주식의 미래가격을 예측할 수 있다는 주장들을 확인해 보기 위해 이 논문에서는 주가의 방향성 예측에 대한 선행연구의 방법과 결과들을 소개하고, 대표적으로 사용되는 기술적 지표 등을 변수로 하는 통계 모형과 기계학습 모형을 혼합 적용하여 한국유가증권시장의 개별 주식 가격의 방향을 예측해 보았다. 또한 예측 모형을 통하여 현실적인 수익으로 이어지는 투자 전략을 수립할 수 있는지 판단해보고자 하였다.

주식 가격 예측의 가능성을 알아보기 위해 사용된 모형은 근래에 많이 사용되는 분류 기법인 이차판별분석(quadratic discriminant analysis, QDA), support vector machine (SVM), random forest (RF), extreme gradient boost (XGB), 심층신경망(deep neural network, DNN) 등이다. 사용된 설명변수들은 선행연구에서 보편적으로 사용되었던 기술지표들을 비롯하여 회계지표인 price earning ratio (PER), price book-value ratio (PBR), 그리고 은닉마르코프모형(hidden Markov model, HMM)의 출력값 등이다.

2절에서는 선행연구를 요약하여 기존의 연구 방향성에 대하여 살펴보고, 각 결과를 활용한 투자전략을 수립할 수 있는지 판단해 본다. 3절에서는 개별 주식 가격 예측에 사용된 분석 기법에 대해 설명하고, 이를 토대로 4절에서는 삼성전자, LG화학, NAVER 등 3개 기업의 실제 개별 주식 자료에 적용하여 결과를 살펴볼 것이다. 5절에서 결론 및 향후 연구방향을 서술한다.

2. 선행연구

주식 가격 및 방향성을 예측하기 위해 사용된 통계적 방법론으로는 간단한 다중선형회귀모형에서부터 시계열모형, 마르코프 모형이나 신경망까지 다양하다. 특히 최근에는 기계학습 분야가 폭발적으로 발전하면서 복잡한 비선형 모형을 이용한 주식 가격 예측에 대한 연구가 이어지고 있으며, 예측가능성에 대한 기대도 높아지고 있다. 이를테면, Gu 등 (2020)은 기계학습 모형으로 수익률을 예측하고 이를 활용해 투자한다면, 시장 수익률을 상회하는 수익률을 얻을 수 있다고 주장하였다. 1957년 부터 2016년까지 60년간의 개별 주식에 대해 선형모형, 의사결정나무 기반 기계학습 모형, 신경망 모형 등 다양한 기법을 활용해 실험을 진행했는데, 결정계수 등의 측도로는 찾을 수 없지만 Sharpe ratio를 측도로 삼았을 경우 유의하게 우월한 투자전략을 찾을 수 있으며 이를 투자에 활용한다면 초과 수익을 달성할 수 있다고 주장하였다. 지난 30년간 예측 결과를 토대로 모의투자를 진행하였을 때 시장 수익률보다 최대 약 2배 이상의 수익률을 달성 할 수 있었다는 점을 근거로 제시하였다.

이 절에서는 전통적인 선형회귀모형부터 기계학습 방법론까지 주식 가격 및 방향성 예측에 사용되어 온 다양한 방법론과 결과를 정리해 보고자 한다. 모든 선행연구를 다 다룰수는 없기에 예측 결과의 평가방법별로 몇 가지 연구를 소개하기로 하겠다.

2.1. 결정계수

Umstead (1977)는 월별로 계산되는 National Bureau of Economic Rasearch (NBER) 경기 선행 지수(composite leading indicators index)를 이용해 S&P500 지수의 월별 마지막 거래일 종가를 예측하였다. 1948년에서 1960년을 training set으로, 1961년에서 1974년을 test set으로 하여 모형을 적합했으며, S&P500 지수의 자기상관 모형과 NBER 경기 선행 지수의 비교상관 모형을 혼합하여 사용하였다. t월의 S&P500 지수를 추정하기 위해 $t-16$월의 S&P500 지수와 함께 $t-10$월 부터 $t-6$월 까지의 NBER 경기 선행 지수가 사용되었다. training

set의 결정계수는 약 0.623으로 좋은 성능을 보이는 듯 했지만, test set의 결정계수는 0.066으로 나타나, 실제 예측은 성공적이지 못했다.

Cheng 등 (1990)은 주가 지수에 영향을 주는 선행 지수가 존재하며, 예시로는 실업률, 무역 수지, 소비자 물가 지수, 광의통화, 전체 통화량, 저축 예금액 등이 있다고 하였다. 여기서 광의통화는 협의통화에 정기 예적금, 시장형 금융상품, 실적배당형 금융상품, 금융채 등을 더한 것으로, 협의통화에 비해 크고 유동성이 떨어지는 통화량을 의미한다. 이러한 선행지수를 이용해 hang seng index (HSI)를 예측하기 위해 다중선형회귀모형을 적합하였다. training set의 결정계수는 약 0.7026 이었고, 이를 기반으로 예측이 가능하다고 주장했지만, test set에 대한 결과는 제시하지 않았다.

2.2. Mean absolute percentage error (MAPE)

Akgiray (1989)는 ARCH와 GARCH 모형을 이용하여 1963년 1월 1일부터 1986년 12월 31일까지 기간의 center for research in security prices (CRSP) 시장 지수의 변동성을 예측하였다. 이때, 6년 단위 4개의 기간 (1963년–1968년, 1969년–1974년, 1975년–1980년, 1981년–1986년)으로 분류하여 각 기간의 마지막 240거래일을 test set으로, 나머지를 training set으로 사용했으며, test set의 20거래일 단위 분산을 예측해 실제 분산과 비교하였는데, 이렇게 변동성을 예측할수 있다면, 위험 정도를 예상해 투자자들의 투자에 도움이 될 수 있을 것이라 주장하였다. 각 test set의 MAPE는 GARCH모형이 더 좋았지만, 모든 수치가 30%를 상회하는 등 예측오차가 크게 계산되어 유의미한 예측가능성을 얻지는 못하였다.

Kim과 Lee (1997)는 신경망(neural network, NN)과 사례기반추론(case-based reasoning, CBR) 등을 이용해 S&P500 지수의 예측을 시도하였다. 사례기반추론은 과거에 있었던 사례들의 결과를 바탕으로 새로운 사례의 결과를 예측하는 기법이다 (Kolodner, 1993). 사용된 변수는 과거의 S&P500 지수, 연방 자금, 광의통화, 주택 신축 착공, 산업 생산, 소비자 물가 지수 등으로, 이들을 표준화한 후 로그를 취해 사용하였다. 이때, 1 거래일 단위로 예측하여 6 거래일 후를 예측하는 web prediction과 6 거래일 후를 직접 예측하는 jump prediction을 모형별로 비교하였다. 예측 방식별 MAPE의 모형별 평균은 약 7.3%과 5.5%로, web prediction 보다 jump prediction이 더 좋은 성능을 보였다. 하지만 이는 traning set에서의 수치로서, test set에 대한 구체적인 예측 성능은 제시되지 않았다.

Ikoku 등 (2010)은 ARIMA와 추세를 추정한 structural arima (SARIMA) 기법 등을 활용해 2000년 1월 1일부터 2009년 12월 31일까지 기간의 월별 NSE all share index 종가를 적합하였다. 이때, price-earning-ratio (PER)를 이용해 추세를 추정한 후 시계열 기법을 적용한 모형과 추세를 추정하지 않고 시계열 기법을 적용한 모형의 결과를 비교하였는데, augmented Dickey-Fuller unit root (ADF) 검정과 Phillips-Perron unit root (PP) 검정을 통해 PER과 all share index의 자기상관 계수를 각각 확인하여 1차 차분된 자료를 사용하였다. 이후 두 자료의 Granger causality 검정을 이용한 공적분 검정을 수행하고, 이를 통해 PER과 all share index 두 변수의 상호간에 유의한 관계가 있음을 보였다. 결과적으로 SARIMA 모형으로 적합함이 가장 좋다고 주장했지만, MAPE가 약 10%이상으로 크게 계산되었다. 해당 기간동안 NSE all share index의 평균 월 변화율은 약 6.49% 였고, 이 값과 비교하면 예측의 오차가 상대적으로 크다는 것을 알 수 있다. 즉, 이 결과를 기반으로 투자를 진행할 때 실질적 수익으로 이어질 수 있을지 의문이 드는 결론이었다.

Sopipan 등 (2012)은 거래일 기준 $t-1$ 시점의 미국, 영국, 일본, 중국 등 주요 국가의 주가지수와 환율, 금, 석유 가격 등 18개의 변수로 주성분분석과 다중선형회귀분석을 사용해 t 시점의 일별 SET50 지수의 종가를 예측했다. SET50는 태국의 주가지수이다. 2007년 1월 4일부터 2011년 2월 28일까지의 기간을 training set으로, 2011월 3월 1일부터 2011년 3월 30일까지의 기간을 test set으로 사용했으며, 주성분 분석 결과 각 3개, 4개, 18개의 성분을 설명변수로 하는 총 3개의 모형을 비교하였다. 세 모형 모두 0.97을 상회하는 결정계수 수치를 나타냈고, 모든 성분을 사용 했을 때 test MAPE는 약 0.9%, MSE는 약 65.75 정도로 나타났다. 그러나

이 경우, $t-1$ 시점의 시장 종료 전에 투자 전략을 수립해야 적절한 투자가 가능함에도 불구하고 각 국가의 주요 주가 지수 등의 정보를 얻는 시점이 그보다 늦을 수 있기 때문에 현실적인 수익률로 이어지는 것은 어려울 것이라고 판단된다.

Gupta와 Dhingra (2012)는 2002년 8월 12일에서 2011년 9월 23일까지 기간의 Tata Steel, 2003년 2월 10일부터 2005년 1월 21일까지 기간의 Apple Inc., IBM, 그리고 Dell 회사의 일별 주식 가격 자료를 변수로 하는 HMM을 이용해

$$\frac{\text{시가} - \text{종가}}{\text{시가}}, \frac{\text{고가} - \text{시가}}{\text{시가}}, \frac{\text{시가} - \text{저가}}{\text{시가}},$$

등을 예측하였다. 사전 확률 분포는 정규 분포를 사용했으며, 4개의 은닉 상태를 가정하였다. 이 결과를 ARIMA, 인공신경망 등의 결과와 비교했는데, 모든 기업에서 ARIMA나 인공신경망 모형을 사용한 결과보다 HMM을 사용했을 때 더 좋은 성능을 보였다. Tata Steel과 Apple Inc.의 경우 약 1.5%, IBM과 Dell의 경우 약 0.6%의 MAPE 수치를 나타냈는데, 이는 당일 시가에 대한 비율적 고가, 저가, 종가를 예측한 것이다. 그러나 고가 및 저가가 관찰되는 시점은 미리 알 수 없으므로 현실적인 투자 전략을 수립하는 데 사용되기는 어려워 보인다.

Hwang (2018)은 심층신경망(DNN)을 이용해 KOSPI의 일별 종가를 예측하였다. 3개의 신경망 모형을 이용 및 연결하여 최종 예측값을 산출하였는데, 첫 번째 모형은 auto-encoder를 이용해 encoding된 출력 값을 DNN의 변수로 사용하였다. 설명변수는 1, 2 거래일 전의 고가, 저가, 종가와 5일, 10일, 15일 이동평균 등 총 12개를 사용했으며, 1개의 은닉층을 가진다. 두 번째 모형은 첫 번째 모형에서 encoding된 출력값을 이용해 1개의 은닉층을 가지고 고가, 저가 및 종가를 출력하는 신경망을 학습하도록 하였다. 마지막 모형은 두 번째 모형의 출력값을 이용해 1개의 은닉층을 가지고, 종가를 출력하는 모형이다. 이렇게 학습된 세 개의 신경망 모형을 연결하여 최종 모형으로서 KOSPI의 종가를 예측하도록 하였다. 제시된 최종모형의 test MAPE는 약 0.6672%로, test 기간의 KOSPI 평균 일 변화율이 0.9%였던 것에 비교하더라도 상당히 좋은 예측력을 보였다. 그러나 구간 추정치 등을 제시하지 않고 있어, 통계적으로 유의미하게 좋은 예측이 가능한지 확인하기는 어려웠다.

2.3. Root mean aquare error (RMSE)

Kim과 Seong (2016)은 기존에 사용된 ARIMA 모형과 경험적모드분해법(empirical mode decomposition, EMD)을 혼합 활용하여 일별 KOSPI 종가를 예측하였다. EMD 방법은 시계열 자료를 서로 직교하는 내재모드함수(intrinsic mode function, IMF)들로 분해하는 방법으로 (Kim 등, 2009), EMD에 의하여 생성된 IMF들은 원 시계열을 다양한 주파수를 가지는 성분들로 분해한 것이다. 이 IMF들을 고주파와 저주파로 그룹화하여 통합된 IMF와 잔차의 성분들은 그 성질에 따라서 ARIMA 모형 및 지수평활법과 결합된 혼합모형으로 예측을 수행하는데에 변수로 사용되었다. 가장 좋은 모형의 test기간(2018년 8월 1일에서 2018년 9월 30일까지)의 1-step(다음 거래일의 종가 예측) RMSE는 약 21.97로 제시되었다. 해당기간 KOSPI의 일별 수익에 대한 표본 표준편차는 약 14.057로, 표준편차에 비해 RMSE의 값이 상당히 크므로 이 결과가 실제 수익으로 이어질 수 있을지 판단하기는 어렵다.

2.4. 정분류율

Huang 등 (2005)은 S&P500 지수, JPY 지수를 변수로 하여 SVM을 이용해 다음 거래일에 해당하는 NIKKEI 225 지수의 이동 방향을 예측하였다. 이 외에도 선형판별분석(linear discriminant analysis, LDA), 이차판별분석(QDA) 등의 모형과 성능을 비교하였다. 그 가운데 SVM의 성능이 가장 좋았는데, test set에 대한 0.73의

정분류율을 보였다. 그러나 무정보적 정분류율과 신뢰구간 등의 정보가 주어져 있지 않아 통계적으로 유의미한 예측력을 가지는지 확인하기는 어렵다.

Zhao 등 (2008)은 GARCH를 이용해 조건부 표준편차를 추정한 후, 이를 위험 정도를 나타내는 추정값으로 보고 변수로 활용해 conditional arch (CARCH) 모형을 적합하여 주주들의 심리를 추정하고, 더 나아가 주식 시장에서의 수익을 예측할 수 있다고 주장하였다. 상해종합주가지수(Shanghai stock exchange, SEC)의 2007년 자료에 모형을 적합하여 2008년 3월의 거래일별 종가의 이동 방향 및 변화량을 예측하였는데, 방향 예측에 대한 정분류율이 0.8235이었으며 95% 신뢰구간은 (0.5657, 0.962)로 매우 높았지만, 같은 시기의 무정보적 정분류율(no informative accuracy: 모든 값을 같은 방향 즉, 상승 또는 하락으로 예측한 결과의 정분류율 중 높은 수치)이 0.7059로 나타나, 통계적으로 유의미한 차이를 보여주지 못하였다.

Farshchian과 Jahan (2015)은 HMM을 이용해 이란 소재의 Jaber Ebne Hayyan Pharmaceutical, Shargh Cement, Shiraz Petrochemical 등 각 3개 기업의 개별주식의 방향을 대상으로, 거래량, 시장지수, 산업지수를 변수로 사용해 일별 종가의 상승, 하락 등 방향을 예측했다. 각 회사 일별 종가의 상승, 하락을 기준으로 하여 자료를 나누어 상승, 하락 2개의 상태로 가정한 후 모형을 적합했고, 출력 결과를 다양한 결정모형의 설명변수로 활용해 주식 가격의 방향을 예측하였다. 약 0.7의 test accuracy를 보였으나, 여기서도 무정보적 정분류율과 신뢰구간 등의 정보를 확인할 수 없어, 통계적으로 유의미한 예측력을 보이는지 알 수 없다.

또, Ren 등 (2019)은 기업 관련 인터넷 뉴스 기사의 긍정적 단어 수와 부정적 단어 수를 이용해 투자자의 심리를 추정하고, 이와 함께 시가, 고가, 저가, 종가, 거래량 등을 변수로 하여 SVM을 적합하였다. 다음 거래일의 Shanghai stock exchange 50 (SSE50) 지수의 이동방향을 예측한 결과 0.7 이상의 높은 정분류율을 보여주었다. 그러나 부정적 단어와 긍정적 단어의 추출 방식 등의 방법이 모호하며 명확한 알고리즘을 제시하지 않았고, 무정보적 정분류율과 신뢰구간 등의 정보 역시 제시하지 않아, 예측력의 유의성을 확인하기 어렵다.

Basak 등 (2019)은 주식 가격의 이동량 및 방향을 이용한 relative strength index (RSI), stochastic oscillator (SO), williams percentage range (W%R), moving average convergence and divergence (MACD), price rate of change (PROC) 등과 거래량을 이용한 on balance volume (OBV) 등의 기술적 지표들을 변수로 하여 random forest (RF)와 XGBoost (XGB)를 이용해 Apple, FaceBook의 주식 가격의 이동 방향을 예측하였다. 시계열 분석에서 자주 쓰이는 rolling window 방식을 사용하고, 예측 대상에 해당하는 시점을 달리하여 정분류율을 비교했는데, 최대 0.93에서 0.94정도의 매우 높은 수치를 보였다. 이는 90 거래일 후를 예측하는 것이었지만, 5 거래일 후를 예측할 때에도 정분류율은 0.7을 상회는 수치를 보였고, 이는 높은 수익으로 직결될 수 있어 보였다. 그러나 예측을 수행하는 시점과 예측 대상 시점 사이의 정보를(예측 수행시점 기준으로 미래의 정보) 예측에 활용했기 때문에 현실적인 투자 전략을 수립하여 수익으로 이어가기에는 무리가 있다.

이 결과를 현실적으로 적용 가능한지 알아보기 위해 KOSPI의 대표 종목 중 하나인 삼성전자의 2007년 7월 2일부터 2020년 11월 6일까지 기간의 주가에 대해 해당 연구를 재현해보았다. Distance가 1일 때, 즉 5 거래일 후를 예측하기 위해 예측시점으로부터 4거래일 후까지의 정보를 알 때에는 약 0.7을 상회하는 test set에 대한 정분류율을 보이지만, 현실에 적용이 가능한 경우로서 distance가 5일 때에는(예측시점까지의 정보만 알 때) 약 0.5001로 무정보적 정분류율인 0.5243보다도 낮은 수치를 보였다. 앞으로 5거래일 후의 주가를 예측하려면 실제로 사용할 수 있는 정보는 distance가 5 이상이어야 하고, distance가 5 미만인 모형은 적용 불가능하다. 1부터 5까지에 해당하는 distance 별로 test set에 대한 정분류율은 Figure 1을 통해 확인할 수 있고, 예측에 적용가능한 distance에 대하여는 test accuracy가 높지 않음을 볼 수 있다.

3. 방법론

이 절에서는 4절에서 실제 주식 가격 예측을 하는데에 사용한 기계학습 방법론을 간략히 기술하고자 한다. 비교적 약한 예측 모형(weak prediction model)과 강한 예측 모형(strong prediction model)을 모두 고려하여

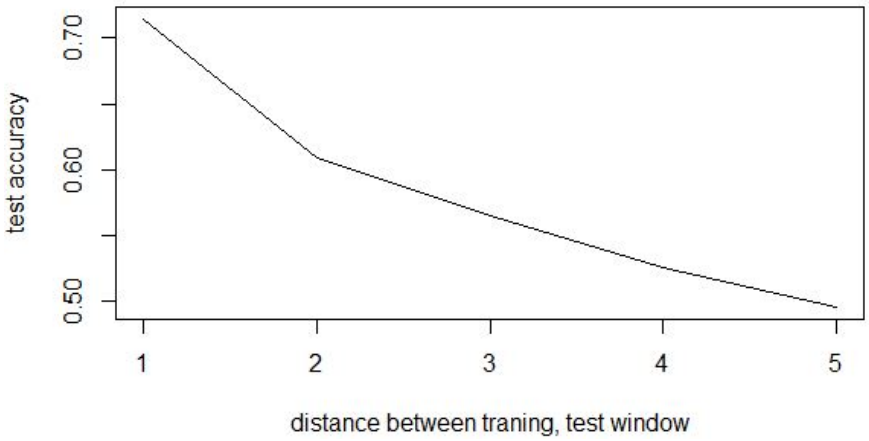

Figure 1: *Accuracy for test set by distance between training set and test set.*

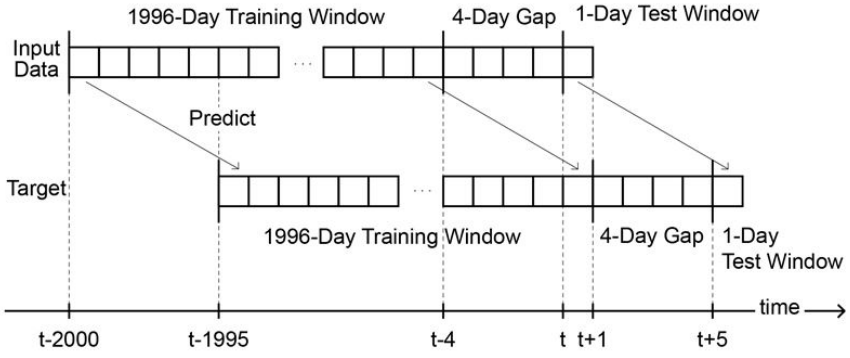

Figure 2: *Rolling window.*

보편적으로 자주 사용되는 모형을 선정하였다. 4절에서는 5거래일 후의 등락을 예측해 볼 것이다. 자산의 가치는 장기적으로 우상향의 형태를 띄는 것이 보통이기에 장기예측은 크게 의미가 없을 수 있으며, 예측 결과를 토대로 투자를 통해 초과 수익을 얻고자 한다면 단기예측에 더욱 관심이 있게 된다. 그러나 1거래일 후와 같이 너무 짧은 미래의 예측은 백색소음의 영향을 받기 쉽기 때문에 이 영향을 줄여 예측 가능성을 높이기 위해 5거래일 후인 $t+5$ 시점의 등락예측을 목표로 하였다. 거래일 기준으로 5일 후의 예측은 일주일 후의 가격을 예측하는 것이 된다.

자료는 rolling window 방식으로 적용했는데, training window의 크기는 1996년(설명변수의 경우 $t-2000$ 부터 $t-5$, 반응변수의 경우 $t-1995$ 부터 t)이며 test window의 크기는 1(설명변수의 경우 t, 반응변수의 경우 $t+5$)로, $t+5$ 시점의 종가를 예측한다(Trading window: 5). 이때, 예측을 수행하는 t 시점에서 투자에 대한 의사결정을 해야 하기 때문에 현실적으로 $t+1$ 시점부터 $t+4$ 시점까지의 정보는 사용할 수 없다. 이 값들은 학습과정에 사용하지 않도록 했으며, 이 기간을 Gap이라 한다. Figure 2에서 예측을 수행하는 시점과 예측 대상에 해당하는 시점 및 Gap의 크기 등을 가시적으로 확인할 수 있다.

3.1. 은닉마르코프모형(hidden Markov model)

HMM은 관찰될 수 없는 상태 확률변수가 존재하고 이 상태로부터 영향을 받는 관찰가능한 확률변수가 존재하며, 상태 확률변수는 관찰가능한 확률변수의 값을 통해서만 파악될 수 있는 이중 구조를 가정하고, 상태 확률변수 및 두 확률변수의 관계를 관찰가능한 변수를 통해 파악하도록 모형화하는 기법이다 (Rabiner와 Juang, 1986).

주식 가격의 상승과 하락 상태를 추정하기 위하여 t 시점과 $t+5$ 시점의 종가를 비교해 '상승'과 '하락'이라는 두 개의 은닉 상태를 가정했으며, t 시점의 일별 로그수익률을 관측값으로 사용하였다. 사전 확률 분포는 정규분포를 가정하였고, 최적화 방식으로는 Baum-Welch 알고리즘을 사용하였다. 이를 통해 도출된 전이 확률 행렬을 이용해 $t+5$ 시점의 상승 확률을 구하고, 이 예측된 확률을 4절의 사례연구에서 추가적인 설명변수로서 사용하고자 한다.

3.2. 이차판별분석(quadratic discriminant analysis)

QDA는 이차항을 사용하는 모형이므로 선형모형보다는 복잡하지만 다른 기계학습 모형보다는 단순하여, 변수 사이의 관계파악이 상대적으로 용이하면서 test set에서의 성능이 많이 떨어지지 않을 것으로 기대되어 비교모형에 포함하였다. 설명변수에 존재하는 c개의 집단 $\omega_1, \omega_2, \ldots, \omega_c$에 대해, 각 집단의 사전 확률 $P(\omega_i) = n_i/N$을 고려한 베이즈 정리를 이용해 집단별 사후확률을 추정하는 방식을 사용하기 때문에 (Tharwat, 2016), 설명변수 간의 공분산을 고려할 수 있다는 장점도 있다. 이때, N은 총 표본의 수, n_i는 i번째 집단에 속하는 표본의 수이다. 설명변수의 확률분포는 정규분포를 가정한다.

3.3. Support vector machine (SVM)

SVM을 이용한 결정모형은 support vector를 이용해 각 집단을 분리하는 초평면(hyperplane)을 찾는 것인데, 이 때 각 집단과 초평면 사이의 거리를 이용하는 학습 알고리즘이다. 초평면과 각 집단 사이의 거리의 최솟값인 margin을 최대화 하는 초평면을 찾는 과정에서 margin을 결정하는 관측 값들이 support vector에 해당한다. Kernel 함수를 이용해 고차원 접근이 가능하다 (Cristianini와 Shawe-Taylor, 2000).

Support vector를 사용하기 때문에 이상치 혹은 오류 자료에 영향을 받지 않는 장점을 갖는다고 평가받는 SVM은, 주식 가격을 잡음을 포함한 정보로 간주하여 예측하는 데에 유용할 것이라 판단해 연구모형으로 사용하였다. 이번 연구에서는 선형 모형보다는 복잡한 관계를 예상해 비선형 kernel 함수인 'Radial' 옵션을 사용했다.

3.4. Random forest (RF)

의사결정나무를 이용한 앙상블(ensemble)모형인 RF (Breiman, 2001)는 가장 많이 사용되는 결정모형 중 하나이다. Bagging의 과정을 거치고, 추가로 변수 일부분을 복원 추출하여 의사결정나무를 생성하는 작업을 반복하여 앙상블하는 기법으로, M 개의 변수가 있다면 $m = \sqrt{M}$ 개의 변수를 랜덤 추출하여 사용한다. 각 의사결정나무는 Gini계수가 0이 될 때까지 모형을 구체화하는데, 이렇게 생성된 의사결정나무들의 출력 값을 종합하여 최종 예측 값을 결정한다.

이번 연구에서는 사용된 변수가 많지 않기 때문에 50개의 의사결정나무를 사용하여 RF를 적합하였다. 이때 각 의사결정나무별로 최소 1개의 노드를 가지며, 각 자식 노드에 한 가지 집단만 남을 때까지 노드의 개수를 늘릴 수 있도록 제한을 두지 않았다.

3.5. Extreme gradient boost (XGB)

Boosting은 많은 수의 약한 예측 모형을 종합하여 강한 예측 모형으로 만드는 것으로, 최적화 방법에 대한 많은 연구가 있었다. Gradient boosting (GB)은 경사 하강법(gradient descent)을 이용한 의사결정나무에 기반을 둔 Boosting모형으로, i 번째 학습모형에서의 가장 높은 오분류 대상을 $i+1$ 번째 모형에서 정분류 하도록 학습하는 방식으로, 높은 성능을 보여주는 대표적인 모형으로 평가받고 있다. XGB (Chen과 Guestrin, 2016)는 의사결정나무를 만드는 시간을 최소화 하는데 목적을 둔 GB모형으로, 각 의사결정나무를 만드는데 변수의 일부분만 사용하도록 설계되었다.

3.6. 심층신경망(deep neural network)

DNN은 인공신경망(artificial neural network)에 기반하여 설계된 개념이며, 입력층과 출력층 사이에 여러 개의 은닉층들로 이루어진 인공신경망으로, 복잡한 비선형 관계를 모형화할 수 있다는 장점이 있다. 주식 시장에는 다양한 자산들과 여러 종류의 상품들이 있으며 이들은 많은 시장 참여자들과 다양한 변수들의 복잡한 관계 속에서 거래되는데, 이 복잡한 관계를 모두 이해하고 표현한다는 것은 매우 어려운 일이다. 신경망은 이러한 복잡한 관계를 해석하지 않고도 적합할 수 있다는 점에서, 신경망을 이용한 주가예측이 유용할 것이라 판단하였다.

이번 연구에서 주가의 방향성 예측을 위해 4개의 은닉층과 ($m = 4$), 은닉층의 활성화 함수 'relu'를 사용하였고, 출력층의 활성화 함수는 'sigmoid'를 사용하였다. Cross-entropy를 손실함수로 사용했으며, 최적화 방식으로는 Adam 알고리즘을 사용하였다.

4. 사례연구

이 절에서는 3절에서 제시된 방법론을 이용하여 한국유가증권시장에서 거래되는 삼성전자, LG화학, NAVER 등 3개 종목에 대해 주식 가격의 방향을 직접 예측해보고자 한다. 시가총액, 거래량, 인지도 등을 고려하여 각각 다른 업종에서 3가지 종목을 선택하였다.

4.1. 변수

각 종목의 2007년 7월 2일부터 2020년 11월 6일까지 기간에 해당되는 일별 종가, 일별 거래량 등을 활용해 기술적 지표를 생성하여 설명변수로 사용하였다. 여기에 추가적으로 PER, PBR 등 회계 정보를 내포한 변수와 HMM의 출력값 등을 사용하였다. 반응변수를 제외한 모든 기술적 지표 변수들을 만들 때 사용된 일별 종가는 그대로 사용하지 않고, 지수 평활(exponentially smoothing)한 S를 사용하였다.

4.1.1. 반응변수

예측 대상은 5 거래일 후 주식 가격의 방향성이므로, 식 (4.1)과 같이 각 종목에 대한 거래일 $t+5$의 종가와 t의 종가를 비교하여 t 시점의 반응변수 y_t를 주식 가격의 방향성을 나타내는 이항변수로 정의하였다.

$$y_t = \begin{cases} 1, & \text{if close}_t < \text{close}_{t+5} \\ 0, & \text{if close}_t \geq \text{close}_{t+5} \end{cases} \tag{4.1}$$

close_t : t시점의 종가

4.1.2. 단순지수평활법(simple exponential smoothing)

기술적 지표 변수를 만들때 사용되는 각 종목의 일별 종가는 그대로 사용하지 않고 식 (4.2)와 같이 지수 평활을 거쳐 생성하였다. Wang 등 (2012)에 따르면 주식 가격을 지수평활하는 경우 보다 효과적인 예측이 가능하다. 평활계수(smoothing factor) α는 0.095를 사용하였다.

$$S_t = \begin{cases} close_1, & \text{if } t = 1 \\ \alpha \times \text{close}_t + (1-\alpha)S_{t-1}, & \text{if } t > 1. \end{cases} \tag{4.2}$$

4.1.3. 설명변수

설명변수로 사용한 기술적 지표들은 널리 사용되는 PROC, OBV, MACD, SL의 네 가지로, 자세한 설명은 다음과 같다. 아래에 사용되는 S는 4.1.2절의 지수평활로 생성된 값들이다.

- Price rate of change (PROC)

  한 시점에서 다음 시점까지의 가격 변화율을 측정한 것으로, 현재 가격과 d거래일 전 시점의 가격을 비교한다. 흔히, 12거래일 전의 가격을 사용하며, 이번 논문에서도 같은 값을 사용하였다. Larson (2015)에서 약 90%의 주가 등락 예측 성공률을 입증한 매우 강력한 지표라고 지목받았으며, Abouloula 등 (2018), Kim 등 (2006)의 논문에서 실험 및 가격 예측 등에 사용되었다.

$$\text{PROC}_t = \left(\frac{S_t - S_{t-d}}{S_{t-d}} \right), \quad d = 12$$

- On balance volume (OBV)

  거래량의 변동을 이용해 주식 가격을 추정하는 지표로, 누적 거래량으로 개별 주식의 매수와 매도 경향을 지수화한 것이다 (Granville, 1976). 상승할 때 거래량을 더하고 하락할 때 거래량을 빼는 방식을 이용한다. Kim 등 (2006), Metghalchi 등 (2012)의 논문에서 가격 예측 및 분석 등에 사용하였다.

$$\text{OBV}_t = \begin{cases} 0, & \text{if } t = 0 \\ \text{OBV}_{t-1} + \text{Volume}_{t-1}, & \text{if } S_t > S_{t-1} \\ \text{OBV}_{t-1} - \text{Volume}_{t-1}, & \text{if } S_t < S_{t-1} \\ \text{OBV}_{t-1}, & \text{if } S_t = S_{t-1} \end{cases}$$

  Volume_t : t 시점의 거래량

- Moving average convergence & divergence (MACD)

  장기 이동평균과 단기 이동평균을 비교하는 모멘텀 지수로, 26일 지수이동평균과 12일 지수이동평균의 차이로 정의된다 (Appel, 2005). Chong과 Ng (2008), Wang과 Kim (2018) 등의 논문에서 주식 가격 분석 및 예측 등에 사용되었다. 평활계수 α는 0.095를 사용하였다.

$$\text{MACD}_t = \text{EMA}^{12}_{S,t} - \text{EMA}^{26}_{S,t}$$

  $\text{EMA}^d_{S,t}$: t 시점에서 S의 d일간의 지수평활

- Signal line (SL)

MACD의 9일 지수이동평균으로 SL과 MACD는 함께 사용된다. MACD와 값을 비교하며, 두 값이 교차하는 시점에 매수 혹은 매도를 결정하는 방식으로 사용된다. 평활계수 α는 0.095를 사용하였다.

$$\mathrm{SL}_t = \mathrm{EMA}^{9}_{\mathrm{MACD},t}$$

$\mathrm{EMA}^{d}_{textMACD,t}$: t 시점에서 MACD의 d일간의 지수평활

이 논문에서는 여러 참고문헌에서 활용된 이와 같은 기술적 지표 외에도 회계지표와 HMM의 출력값을 주가 방향성 예측을 위한 추가변수로 사용하고자 한다.

- 회계 정보

  개별 회사의 회계정보를 이용한 지표로, Price Earning Ratio (PER)와 Price to Book-value Ratio (PBR)를 고려한다. 이들은 회사의 가치 대비 평가된 정도를 지수화한 지표로, PER은 주당 순이익을, PBR은 주당 순자산가치를 나타낸다.

$$\mathrm{PER}_t = \frac{\mathrm{close}_t}{\mathrm{Earning}_t}$$

$$\mathrm{PBR}_t = \frac{\mathrm{close}_t}{\mathrm{Book\_value}_t}$$

close_t : 종가

$\mathrm{Earning}_t$: 주당 순이익

$\mathrm{Book\_value}_t$: 주당 순자산

- HMM 출력값

  3.1절에서 소개된 HMM의 출력 값을 이용해 t 시점의 상승 확률을 계산하여 여러 결정모형의 설명변수로서 사용해보려고 한다. 이를 통해 모멘텀 투자 전략의 가치를 간접적으로 확인할 수 있을 것이다.

4.2. 예측 결과

4.1.3절에 소개된 기술적 지표 변수들을 기본으로 하고 PER, PBR, HMM 출력값을 추가해 가면서 QDA, SVM, RF, XGB, DNN 등의 모형을 사용하여 삼성전자 (005930), LG화학 (051910), NAVER (035420)의 주가 방향성을 예측해 보았다. 5거래일 이후 주가방향을 예측하였고, test set에 대한 정분류율을 측도로 결과를 평가하였다.

Tables 1–3은 기술적 지표 변수들을 기본으로 하고 설명변수를 추가하며 test set에 대한 정분류율을 비교한 표로, 가장 좋은 결과값은 굵은 숫자로 표시하였다. “None” 아래의 열은 추가변수가 없는 모형의 정분류율이고, “HMM” 아래의 열은 HMM의 출력값이 추가변수로 들어간 모형의 정분류율 등이다. 모든 종목에서 기술적 지표만을 사용한 모형이 아닌 추가변수를 사용한 모형이 좋은 결과를 보여주었다.

삼성전자와 NAVER의 경우 HMM, PER, PBR 모두를 사용한 QDA모형이 가장 높은 정분류율을 보였고, SVM, RF, XGB, DNN 등 다른 모형에 대해서도 모든 추가변수를 고려한 경우 상대적으로 높은 예측력을 유지함을 알 수 있다. 모형간의 비교에서는 QDA, SVM 등 상대적으로 약한 예측 모형의 성능이 전반적으로 좋았다. LG화학의 경우에는 HMM의 출력값을 사용한 SVM모형이 가장 높은 정분류율을 보였는데, 회계지표를 활용한 PER과 PBR을 추가했을 때는 예측력이 오히려 하락했으며 QDA, SVM, RF의 모형에서는 정분류율이 0.5보다 낮게 나타났다. 삼성전자와 NAVER의 경우와 달리 LG화학은 XBG나 DNN 등 강한 예측모형이 전반적으로 높은 정분류율을 보였다.

Table 1: Test accuracy of Samsung

No informative accuracy			0.5243	
	None	HMM	PER,PBR	HMM,PER,PBR
QDA	0.5104	0.5139	0.5277	**0.5402**
SVM	0.4993	0.5014	0.5118	0.5007
RF	0.4993	0.4965	0.5042	0.4937
XGB	0.4820	0.5055	0.5208	0.5062
DNN	0.5250	0.5215	0.5236	0.5229

Table 2: Test accuracy of LGchem

No informative accuracy			0.5278	
	None	HMM	PER,PBR	HMM,PER,PBR
QDA	0.5097	0.4993	0.4861	0.4785
SVM	0.4917	**0.5381**	0.4958	0.4924
RF	0.5111	0.5243	0.4951	0.4986
XGB	0.5118	0.5042	0.5264	0.5243
DNN	0.5236	0.5250	0.5250	0.5291

Table 3: Test accuracy of Naver

No informative accuracy			0.5062	
	None	HMM	PER,PBR	HMM,PER,PBR
QDA	0.4986	0.4958	0.5118	**0.527**
SVM	0.4708	0.4806	0.5222	0.5215
RF	0.5013	0.5048	0.5	0.5051
XGB	0.4743	0.5111	0.4909	0.5097
DNN	0.4619	0.4840	0.4674	0.4639

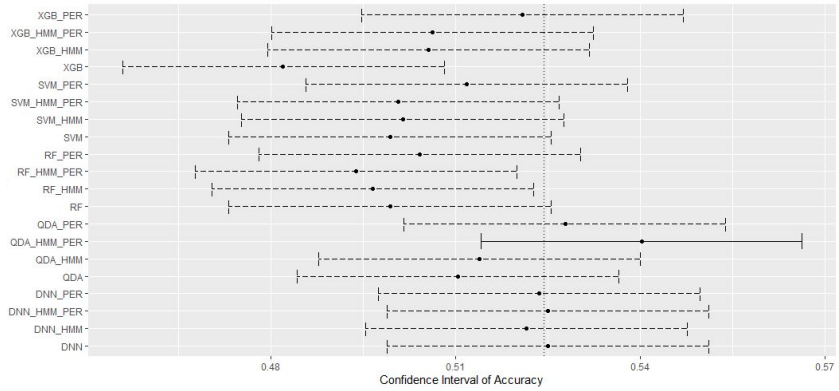

Figure 3: *Confidence interval of test accuracy for Samsung.*

Figures 3–5는 test set에 대한 정분류율의 95% 신뢰구간을 종목별, 모형별, 추가변수별로 표현한 그림으로, 가장 높은 정분류율을 보이는 경우를 실선으로 표현하였다. 가장 높은 정분류율을 갖는 모형의 경우 점 추정치에서는 무정보적 정분류율보다 높은 수치를 나타냈지만, 구간 추정을 통해 살펴보면 통계적으로

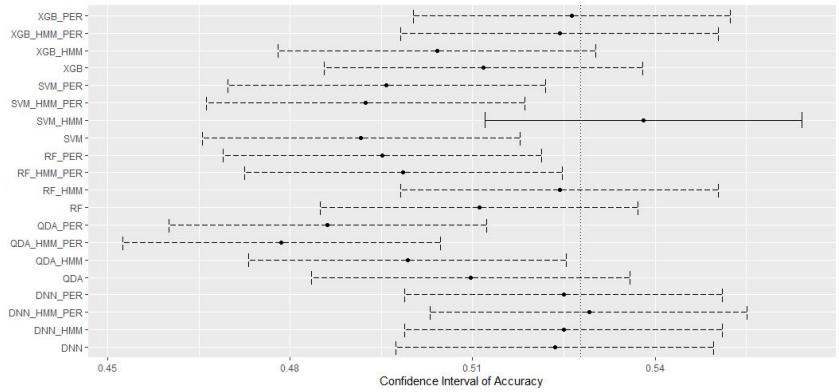

Figure 4: *Confidence interval of test accuracy for LGchem.*

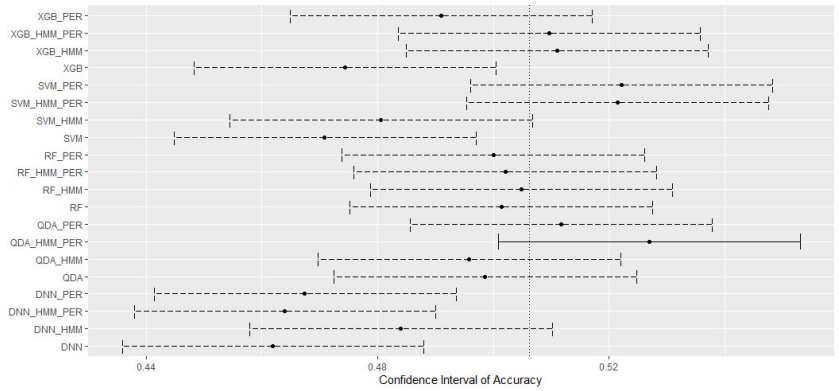

Figure 5: *Confidence interval of test accuracy for Naver.*

유의미한 예측력 개선을 보여주지는 않았다. 또한 예측 기법 및 변수를 변화시키며 비교해보았을 때, 많은 신뢰구간들이 서로 겹치고 있어서, 정분류율이 상대적으로 나은 모형을 찾을 수는 있으나 타 모형들과의 차이가 크지 않다는 것을 확인할 수 있었다.

삼성전자와 LG화학, NAVER의 주가에 대한 이 방향성 예측은 정분류율을 평가측도로 사용하였으므로, 2.4절에서 소개된 선행연구의 정분류율과 비교해볼 수 있다. 물론 자료와 사용변수를 비롯한 세부적인 분석구조가 모두 다르기 때문에 정분류율을 직접적으로 비교하여 차이에 의미를 두기는 어려운 점이 있다. 그럼에도 불구하고 정분류율을 비교한다면, 선행연구들의 정분류율은 모두 이번 예측의 정분류율보다 높았고, 그 중에 무정보적 정분류율을 제공한 Zhao 등 (2008)은 시계열모형을 이용한 심리추정을 활용하여 정분류율을 무정보적 정분류율에 비해 16.7% 개선한 것으로 보고하였다. 삼성전자에서 약 3%, LG화학에서 약 2%, NAVER에서 약 4%의 개선을 보인 이 논문의 결과와 비교하여 높은 수치이다. 이 연구에서는 주로 기계학습 모형의 성능을 확인하고자 했기 때문에 Zhao 등 (2008)에서 활용한 시계열모형을 적용하지 않았지만, 추후 심리변수를 추가한 시계열모형을 이용한 예측과도 비교해 볼 필요가 있어 보인다. Basak 등 (2019)의 경우 현실적으로 활용이 불가능한 자료를 활용하였고, Farshchian과 Jahan (2015)이나 Huang (2005), Ren 등 (2019)에서는 무정보적 정분류율을 제공하지 않았기 때문에 이번 예측과 직접 비교하기는 어렵다.

5. 결론

이 논문에서는 금융시장에서 주식의 가격 및 가격의 상승과 하락의 방향성을 예측할 수 있는가에 대한 과거 연구들을 정리하고, 최근의 연구경향인 기계학습을 활용한 예측가능성을 타진해보기 위해 사례연구를 진행하였다. 선형회귀분석, 자기회귀 누적 이동평균, 자기회귀 조건부 이분산성, 은닉마르코프 모형, support vector machine, 의사결정나무 기반 모형, 신경망 등을 이용한 선행연구들을 정리하였고, 은닉마르코프모형, 이차판별분석, support vector machine, random forest, extreme gradient boost, 심층신경망 등의 방법론을 이용하여 삼성전자, LG화학, NAVER의 주가 방향성을 예측해보았다. 선행 연구에서 널리 사용된 기술적 지표들을 설명변수로 하고, 회계지표와 HMM 출력값을 추가하여 모형을 적합하였다. 무정보적 정분류율과 비교했을 때, 통계적으로 유의미하게 예측이 가능한 모형은 확인할 수 없었으나, 점 추정치에서 무정보적 정분류율을 상회하는 모형들은 발견할 수 있었다.

모형별로 비교했을 때, 통계적으로 유의미한 수치 차이는 아니지만, RF, XGB, DNN 등의 강한 예측 모형보다 QDA 등 약한 예측 모형이 상대적으로 더 좋은 성능을 보였다. 이는 주식 가격에 영향을 주는 변수와 주식 가격 간의 관계를 모형화 하는 데 있어서 test set에서는 비교적 단순한 모형의 결과가 더 나을 수 있음을 나타낸다. 변수별로 비교해보면, 각 개별 주식에 대해 가장 높은 예측력을 보인 모형들은 모두 PER, PBR 등 회계지표와 HMM 출력 값을 활용한 변수가 추가된 경우였다. 이 역시 통계적으로 유의미한 수치 차이를 나타낸 것은 아니지만, 각 기업에 대한 수치적 평가 및 주식 가격의 흐름이 투자자들의 결정에 영향을 주고 있다고 판단된다.

선행 연구의 결과들과 이 논문에서 수행한 3개 종목에 대한 사례연구의 결과를 통해서 볼 때, 주식 가격 방향성의 예측가능성은 기계학습 방법을 활용한다 하더라도 불투명해 보인다. 그러나 향후 주식 가격 예측에 있어서, 개별 주식에 대한 투자자들의 평가를 대변할 수 있는 광범위한 변수들(예를 들어, 뉴스 및 댓글, 블로그, 예측보고서 등을 분석하여 투자자들의 심리를 추정한 변수, 혹은 관련된 다른 종목의 과거에 해당하는 주식 가격 등)을 선택하여 접근한다면, 더 좋은 결과를 얻을 수도 있지 않을까 기대해 본다.

References

Abouloula K, Habil BE, and Krit S (2018). Money management limits to trade by robot trader for automatic trading , *International Journal of Engineering, Science and Mathematics*, **7**, 195–206.

Akgiray V (1989). Conditional heteroscedasticity in time series of stock returns: evidence and forecasts, *Journal of Business*, **62**, 55–80.

Appel G (2005). *Technical Analysis Power Tools for Active Investors*, FT Press, Upper Saddle River.

Basak S, Kar S, Saha S, Khaidem L, and Dey SR (2019). Predicting the direction of stock market prices using tree-based classifiers, *North American Journal of Economics and Finance*, **47**, 552–567.

Breiman L (2001). Random forests, *Machine Learning*, **45**, 5–32.

Bussière M (2013). In defense of early warning signals, *Banque De France Working Paper*, 420.

Chen T and Guestrin C (2016). XGBoost: A scalable tree boosting system. In *KDD'16: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 785–794.

Cheng TCE, Lo YK, and Ma KW (1990). Forecasting stock price index by multiple regression, *Managerial Finance*, **16**, 27–31

Chong TT and Ng W (2008). Technical analysis and the London stock exchange: testing the MACD and RSI rules using the FT30, *Applied Economics Letters*, **15**, 1111–1114.

Cowles A (1933). Can stock market forecasters forecast, *Econometrica*, **1**, 309–324.

Cristianini N and Shawe-Taylor J (2000). *An Introduction to Support Vector Machines and Other Kernel-based*

Learning Methods, Cambridge University Press, Cambridge.

Dsouze JF and Mallikarjunappa T (2015). Do the stock market indices follow random walk?, *Asia-Pacific Journal of Management Research and Innovation*, **11**, 251–273.

Durusu-Ciftci D, Ispir MS, and Kok D (2019). Do stock markets follow a random walk? New evidence for an old question, *International Review of Economics and Finance*, **64**, 165–175.

Fama EF (1970). Efficient capital markets: A review of theory and empirical work, *Journal of Finance*, **25**, 383–417.

Farshchian M and Jahan MV (2015). Stock market prediction with hidden Markov model, *International Congress on Technology, Communication and Knowledge*, 473–477.

Fletcher R (1987). *Practical Methods of Optimization*, John Wiley & Sons Ltd, Chichester.

Granville JE (1976). *Granville's New Strategy of Daily Stock Market Timing for Maximum Profit*, Simon & Schuster, New York.

Gu S, Kelly B, and Xiu D (2020). Empirical asset pricing via machine learning, *Review of Financial Studies*, **33**, 2223–2273.

Gupta A and Dhingra B (2012). Stock market prediction using hidden Markov models, *Students Conference on Engineering and Systems*, 16–18.

Huang W, Nakamori Y, and Wang SY (2005). Forecasting stock market movement direction with support vector machine, *Computers and Operations Research*, **32**, 2513–2522.

Hwang H (2018). Daily stock price forecasting using deep neural network model, *Journal of the Korea Convergence Society*, **9**, 39–44.

Ikoku AE, Hosseini A, and Okany CT (2010). Can price-earnings ratios predict stock prices?, *The International Journal of Finance*, **4**, 6581–6611.

Kim H and Seong B (2016). EMD based hybrid models to forecast the KOSPI, *The Korean Journal of Applied Statistics*, **29**, 525–537.

Kim M, Min S, and Han I (2006). An evolutionary approach to the combination of multiple classifiers to predict a stock price index, *Expert Systems with Applications*, **31**, 241–247.

Kim S and Lee D (1997). Extended forecasts of a stock index using learning techniques : A study of predictive granularity and input diversity, *Asia Pacific Journal of Information Systems*, **7**, 67–83.

Kolodner JL (1993). *Case-Based Reasoning*, Morgan Kaufmann Publishers, Massachusetts.

Larson M (2015). *Price Rate of Change: 12 Simple Technical Indicators: That Really Work*, Wiley, New Jersey

Metghalchi M, Chang Y, and Garza-Gomez X (2012). Technical analysis of the Taiwanese stock market, *International Journal of Economics and Finance*, **4**, 90–102.

Kim D, Oh H, and Suh J (2009). A multi-resolution approach to non-stationary financial time series using the Hilbert-Huang transform, *Korean Journal of Applied Statistics*, **22**, 499–513.

Rabiner L and Juang B (1986). An introduction to hidden Markov models, *IEEE ASSP Magazine*, **3**, 4–16.

Ren R, Wu DD, and Liu T (2019). Forecasting stock market movement direction using sentiment analysis and support vector machine, *IEEE Systems Journal*, **13**, 760–770.

Sopipan N, Kanjanavajee W, and Sattayatham P (2012). Forecasting SET50 index with multiple regression based on principal component analysis, *Journal of Applied Finance & Banking*, **2**, 271–294.

Tharwat A (2016). Linear vs. quadratic discriminant analysis classifier: a tutorial, *International Journal of Applied Pattern Recognition*, **3**, 145–180.

Umstead DA (1977). Forecasting stock market prices, *Journal of Finance*, **32**, 427–441.

Wang J and Kim J (2018). Predicting stock price trend using MACD optimized by historical volatility, *Mathematical Problems in Engineering*, ID 9280590, 1–12

Wang JJ, Wang JZ, Zhang ZG, and Guo SP (2012). Stock index forecasting based on a hybrid model, *Omega*, **40**, 758–766.

Zhao Y, Yang M, and Qi C (2008). Forecast stock market returns based on risk anticipation, *International Conference on Information Management, Innovation Management and Industrial Engineering*, **2**, 377–380.

Received June 19, 2021; Revised July 21, 2021; Accepted July 21, 2021

기계학습을 활용한 주식 가격의 이동 방향 예측

김용환[a], 송성주[1,a]

[a]고려대학교 통계학과

요 약

금융시장에서 주식 가격 자체 또는 가격의 방향성에 대한 예측은 오래 전부터 관심의 대상이 되어 왔기에 여러 방면에서 다양한 연구가 이어져 왔다. 특히 1960년대에 들어서며 많은 연구가 진행되었고 예측가능성에 대해 찬반의 의견들이 있었는데, 1970년대에 나타난 효율적 시장 가설이 지지를 받으면서 주식 가격의 예측은 불가능하다는 의견이 주를 이루었다. 그러나 최근 기계학습 등 예측기술의 발달로 인해 주식 시장에서 미래를 예측해 보려는 새로운 시도가 이어져, 주식시장의 효율성을 부정하고 높은 예측력을 주장하는 연구들이 등장하고 있다.

이 논문에서는 과거 연구들을 평가방법 별로 정리하고, 새로운 주장의 신빙성을 확인하기 위해 이차판별분석, support vector machine, random forest, extreme gradient boost, 심층신경망 등 다양한 기계학습 모형을 적용하여 한국유가증권시장에 상장된 종목 중 삼성전자, LG화학, Naver 주식 가격의 방향성을 예측해보았다. 이때, 널리 사용되는 기술적 지표 변수들과 더불어 price earning ratio, price book-value ratio 등 회계지표를 활용한 변수와, 은닉마르코프모형의 출력값 변수를 사용하였다. 분석결과, 이번 연구의 조건 하에서는 통계적으로 유의미한 예측력을 제시하는 모형이 존재하지 않았고, 현 시점에서 단기 주가 방향성의 예측은 어렵다고 판단되었다. 비교적 단순한 이차판별분석 모형과 회계지표를 활용한 변수를 추가한 모형이 상대적으로 높은 예측력을 보였다는 점에서, 복잡한 모형을 시도하기 보다는 주식 가격에 대한 투자자들의 의견 및 심리가 반영될 수 있는 다양한 변수를 개발하여 활용한다면 향후 유의미한 예측이 가능할 수도 있을 것이다.

주요용어: 예측, 기계학습, 분류, 주식 가격

이 논문은 제1저자 김용환의 석사학위 논문의 일부를 발췌한 것임.
이 논문은 정부(과학기술정보통신부)의 재원으로 한국연구재단의 지원을 받아 수행된 연구임(No. 2021R1F1A1048986)
[1]교신저자: (02841) 서울특별시 성북구 안암로 145, 고려대학교 통계학과. Email: sjsong@korea.ac.kr