Research Article

pISSN : 2287-8920
https://doi.org/10.22716/sckt.2021.9.4.055

# LSTM을 이용한 주가 예측: 기술 지표, 거시 경제 지표, 시장 심리의 조합을 중심으로

한태동[1]
[1]서경대학교 경영학부 조교수

# Stock Price Prediction Using LSTM: Focusing on the Combination of Technical Indicators, Macroeconomic Indicators, and Market Sentiment

Taedong Han[1]
[1]Assistant Professor, Division of Business Administration, Seokyeong University

[1]**Corresponding author:** tdhan@skuniv.ac.kr

**Received** November 16, 2021; **Revised** December 7, 2021; **Accepted** December 8, 2021

**ABSTRACT**

주가 예측 시 가장 큰 장애 요인은 거시 환경 및 시장의 변동성이라 할 수 있겠다. 이론적으로 주가의 무작위성은 딥러닝 알고리즘으로 어느 정도 해소가 가능하지만, 주가를 둘러싼 거시 환경과 시장의 급변과 같은 외부 충격은 주가 예측을 어렵게 만드는 요인이며, 이를 어떻게 해소하는가 하는 것이 주가 예측의 관건이라 할 수 있겠다. 이에 전통적인 기술적 지표나 거시 경제 지표 뿐만 아니라, 시장에 충격을 주는 상황을 이해하기 위한 뉴스 정보나 투자자들의 심리를 반영하기 위해서 주식 종목 관련 단어검색, SNS 주식 커뮤니티에서의 감성 분석 등을 통하여 해결하고자 하는 추세이다. 따라서 본 연구의 목적은 주가에 영향을 미치는 내부 요소와 외부 요소들을 모두 이용하여, 주가 예측 요소의 조합을 통해 보다 높은 예측력을 만드는 딥러닝 모델을 찾는 것이다. 딥러닝 알고리즘으로는 시계열 데이터 학습에 적합한 LSTM을 사용하였으며, 딥러닝 실험 결과 예상대로 모든 종목에 적용할 수 있는 범용 모델은 찾을 수 없었지만, 업종이나 가격대 등에 따라 예측력을 높이는 조합이 존재하며, 현재 시점에서 먼 과거의 데이터보다는 최근의 데이터가 예측력을 높인다는 결과가 나왔다.

In predicting stock prices, the biggest obstacle is macro-environment and market volatility. Randomness of stock prices can be resolved to some extent with deep learning algorithms, but external impacts such as the macro-environment surrounding stock prices and rapid changes in the market are factors that make it difficult to predict stock prices, and how to resolve them is the key to the stock price prediction model. Existing stock price prediction research has often used technical indicators. Recently, macroeconomic indicators have also been taken into consideration, and in order to understand the market impact, news information or investor psychology is reflected. There is also a trend to analyze sentiment in the stock community. Therefore, the purpose of this study is to find a deep learning model that uses both internal and external factors that affect stock prices and creates higher predictive power through the inclusion of specific factors or combinations of factors. As a deep learning algorithm, LSTM suitable for time-series data learning was used. As a result of deep learning experiments, it does not seem possible to find a general-purpose model that can be applied to all stocks, but there are combinations of factors that increase the predictive power depending on the industry or price range, and results also shows that the recent data had better predictive power than the past data.

**Keywords:** Artificial intelligence, Machine learning, Deep learning, LSTM, Stock price prediction

## 1. 서 론

주가 예측을 하는 경우 가장 큰 장애 요인은 거시 환경 및 시장의 변동성이다. 이론적으로 주가의 무작위성은 딥러닝 알고리즘으로 어느 정도 해소가 가능하지만, 주가를 둘러싼 거시 환경과 시장의 급변과 같은 외부 충격은 주가 예측을 어렵게 만드는 요인이며, 이를 어떻게 해소하는가 하는 것이 주가 예측 모델의 관건이라 할 수 있겠다. 이에 전통적인 기술적 지표나 거시 경제 지표 뿐만 아니라, 시장에 충격을 주는 상황을 이해하기 위한 뉴스 정보나 투자자들의 심리를 반영하기 위해서 주식 종목 관련 단어검색, SNS 주식 커뮤니티에서의 감성 분석 등을 통하여 해결하고자 하는 추세이다. 하지만 대부분의 기존 연구들은 개별 지표를 이용하거나 일부 지표만을 이용하여 주가 예측을 하고 있다.

따라서 본 연구의 목적은 주가에 영향을 미치는 모든 요소를 고려하는, 즉 주가 예측에 이용되는 기술 지표, 거시 경제 지표, 그리고 시장 심리 지표 세 가지 모두를 이용하여 가장 효과적인 주가 예측을 하는 딥러닝 모델을 찾는 것이다. 더 나아가서 주가 예측을 위하여 어느 정도의 과거 주가를 이용하는 것이 적절한지, 그리고 모든 주가에 적용할 수 있는 범용 모델을 찾는 것은 쉽지 않으므로 주가의 종류별로 적절한 모델을 찾는 것도 함께 수행한다.

본 연구는 딥러닝 알고리즘, 특히 LSTM(Long Short-Term Memory) 모델을 이용하여 수행되는데, 딥러닝(Deep Learning, 심층 학습)은 컴퓨터가 데이터를 이용해 인공 신경망(Artificial Neural Network, ANN)을 활용하여 스스로 학습할 수 있도록 만들어진 알고리즘으로 머신러닝(Machine Learning, 기계 학습)의 심화된 형태이다. 구글의 인공지능 알파고가 인간 바둑 기사를 물리친 이후로 딥러닝 기술은 급속도로 발전해 오늘날 많은 분야에서 연구의 한 방법론으로 자리 잡아 가고 있으며, 주가 예측에 대한 연구에도 최근 딥러닝 기술이 도입되어 연구가 진행되고 있다.

## 2. 선행 연구

다양한 주가 예측 요소들이 여러 선행 연구에서 이용되어 왔으며, LSTM 이전에도 SVM(Support Vector Machine), MLP(Multi-Layered Perception) 등의 머신러닝 기법을 사용하여 주가 예측을 하거나[1,2], 기술적 지표들을 2차원 그래프로 변환하여 CNN을 적용하는 연구도 있었다[3]. 하지만 최근에는 주가와 같은 시계열 데이터에는 LSTM을 이용하는 추세이다.

Joo & Choi의 연구에서는 주가와 거래량을 사용하여 양방향 LSTM을 제안했다[4]. 본 연구는 예측 요소의 조합을 통한 최적 딥러닝 모델 구축이 주 관심이고, 사전 실험을 통하여 단방향 LSTM 모델도 충분히 만족할만한 예측 결과를 보여주었기 때문에 양방향 LSTM은 본 연구의 범위에서는 제외된다. Jung & Kim 또한 주가와 거래량을 사용하여 LSTM 모델을 실험하였고, 주가의 변화량과 거래량의 변화량을 사용한 것이 특징이다[5].

거시 경제 지표만을 이용한 연구로는 Hong이 있으며[6], Shin et al.은 주가와 Stochastic 1 등의 기술적 지표, 그리고 환율 등의 거시 경제지표를 사용하여 LSTM과 DNN을 비교하였고[7], Fischer & Krauss의 연구 또한 기술적 지표를 사용하여 LSTM과 DNN 등 기존 알고리즘과 비교하였다[8]. 하지만 이 연구들은 최적 주가 예측 요소의 조합보다는 LSTM과 DNN의 비교에 방점을 두었다.

최근에는 투자자의 관심과 감성 등이 주가에 영향을 미치는 척도로서 연구되고 있다. Cho et al.은 키워드 검색량 변화를 통해 주가 예측 모델을 연구하였고[9], Kim et al.은 네이버 주식토론방 데이터를 이용하여 주가의 변동 방향과 변동폭을 예측하였다[10]. 그리고 Zeng et al.은 트위터 키워드 분석을 통하여 투자자의 감성이 주가에 미치는 영향을 연구하였다[11]. 끝으로 Jang은 뉴욕타임즈의 헤드라인으로부터 뉴스 키워드를 입수하고, 국제 금값, 유가, 환율을 거시경제지표로 이용하였다[12]. 그리고 이런 추세에 더하여 Seong & Nam은 온라인 뉴스와 거시경제 지표를 함께 이용하여 주가를 예측하였다[13].

# 3. 실 험

## 3.1 연구 모델 및 분석 방법

### 3.1.1 연구 모델

본 연구의 목적은 기존 연구에서 시도되지 않았던, 주가 예측 요소로서 주가 내재적인 기술 지표, 시장에서 어떤 종목에 대한 투자자의 관심을 나타내는 시장 심리 지표, 그리고 거시 경제 지표의 3가지 성격의 지표를 모두 혼합하여 최적의 조합을 찾는 것이다. 기술적 지표와 거시 경제 지표에는, 많은 연구에서 다양한 특성들이 사용되었지만 본 연구에서는 대표적 특성(feature) 2개씩을 각각 선택하였다 : 2개의 기술적 지표 - 거래량(V)과 코스피 지수(K), 2개의 거시 경제 지표 - 환율(E)과 이자율(I). 그리고 시장 심리 지표는 특정 종목에 대한 인터넷에서의 단어검색 빈도(W)를 이용하였다. 또한 LSTM 모델에서 주가($P_t$)는 과거의 주가($P_{t-n}$)에 영향을 받기 때문에 이를 수식으로 나타내면 식 (1)과 같다.

$$P_t = P_{t-n} + (V_{t-n} + K_{t-n}) + W_{t-n} + (E_{t-n} + I_{t-n}) \quad (1)$$

여기서, $P_t$와 $P_{t-n}$은 각각 현재 t 시점의 주가와 t-n 시점까지의 과거 주가이고, $V_{t-n}$은 t-n 시점까지의 과거 거래량, $K_{t-n}$은 t-n 시점까지의 과거 코스피 지수, $W_{t-n}$은 t-n 시점까지의 과거 단어검색 빈도, $E_{t-n}$은 t-n 시점까지의 과거 환율, 그리고 $I_{t-n}$은 t-n 시점까지의 과거 이자율이다. 그리고 $V_{t-n}$과 $K_{t-n}$은 기술적 지표, $W_{t-n}$은 특정 종목에 대한 시장 심리 지표, $E_{t-n}$과 $I_{t-n}$은 거시 경제 지표이다.

### 3.1.2 알고리즘 선택

본 연구에서는 LSTM(Long Short-Term Memory) 모델을 이용하여 주가 예측에 대한 연구를 실행한다. 딥러닝 알고리즘은 ANN을 기반으로 DNN(Deep Neural Network, 심층 신경망), CNN(Convolutional Neural Network, 합성곱 신경망), RNN (Recurrent Neural Network, 순환 신경망), DBN(Deep Belief Network, 심층 신뢰 신경망), DQN(Deep Q-Network) 등이 있다. 이 중 RNN은 은닉 계층에 과거 데이터의 정보를 기억하기 때문에 시계열 데이터의 학습에 적절하며, 이로부터 발전된 LSTM은 과거 시점이 멀수록 학습 능력이 저하되는 RNN의 문제점을 개선한 알고리즘으로 주가와 같은 비선형 장기 시계열 데이터 예측에 좋은 성능을 보이고 있으며, 최근에 주가 예측 연구에서 주로 사용되는 알고리즘이다. 또한 전통적 시계열 기법인 ARIMA 모델과 비교하여 LSTM의 성능이 더 우수한 것을 확인하는 연구도 있으며[14,15], VAR 모델과의 비교는 LSTM 모델이 항상 우수한 성능을 보이는 것은 아니며 데이터의 특성에 따라 달라진다는 연구도 있다[16]. Cho et al.에 따르면 다변량 시계열 예측을 위한 모델 구축에는 최신의 머신러닝 알고리즘이 항상 우수한 것은 아니며, 비교적 데이터의 길이가 짧은 단계에서는 VAR 모델의 성능이 우수했으며 데이터의 길이가 증가한 경우에는 LSTM 모델의 성능이 우수하게 나타났다[8]. 본 연구에서는 비교적 긴 5년간의 데이터 사용과 딥러닝 알고리즘을 사용한 주가 예측이라는 목적에 따라 LSTM 모델을 적용한다.

### 3.1.3 예측의 정확성 측정

주가 예측의 정확성은 예측값과 실제값의 차이에서 오는 오차값을 통해서 확인할 수 있는데, 본 연구에서는 RMSE(Root Mean Square Error)를 사용한다. RMSE는 평균 제곱근 오차로서, 추정값 또는 예측한 값과 실제 환경에서 관찰되는 값의 차이를 다룰 때 흔히 사용되는 측도이며, 잔차를 제곱화하기 때문에 기울기를 계산할 수 있어 경사하강법을 사용하여 모델의 성능을 최적화하는 머신러닝 알고리즘의 정확도를 평가할 때 가장 널리 쓰이고 있다. RMSE는 예측값에서 실제값을 차감한 값의 제곱 평균을 구한 후에 제곱근 하는 방식이다(식 (2)).

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(actual - predict_i)^2} \quad (2)$$

Figs.1~4는 RMSE 값 차이에 따른 주가 그래프의 예시를 비교해서 보여준다. 실선(actual)은 실제 주가, 점선(predicted)은 예측된 주가를 보여준다. POSCO는 0.02 이하의 RMSE 값이 나타나지 않아서 Fig. 1에서는 신세계의 주가 그래프를 사용하였다.

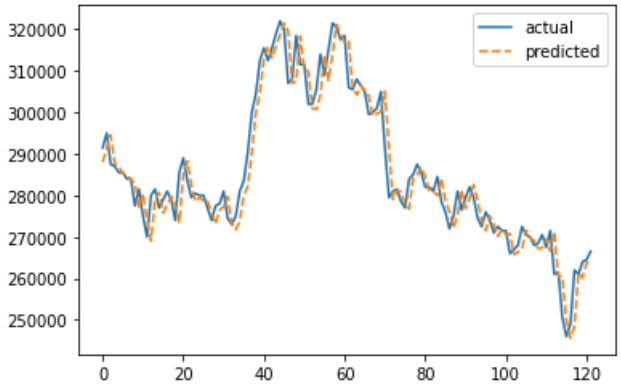


Fig. 1. RMSE=0.0169 (SHINSEGAE)

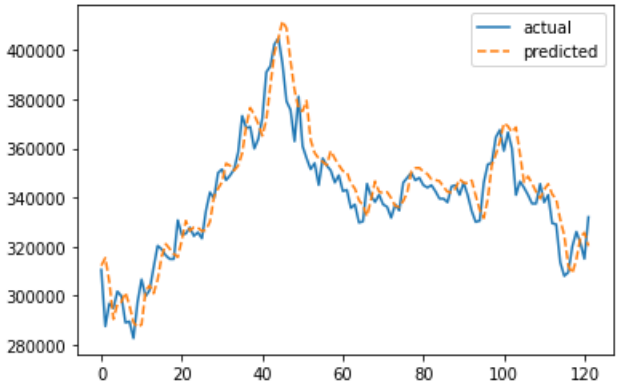


Fig. 2. RMSE=0.0351 (POSCO)

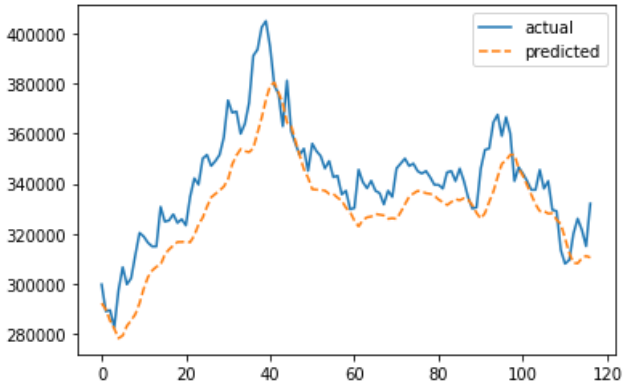


Fig. 3. RMSE=0.0520 (POSCO)

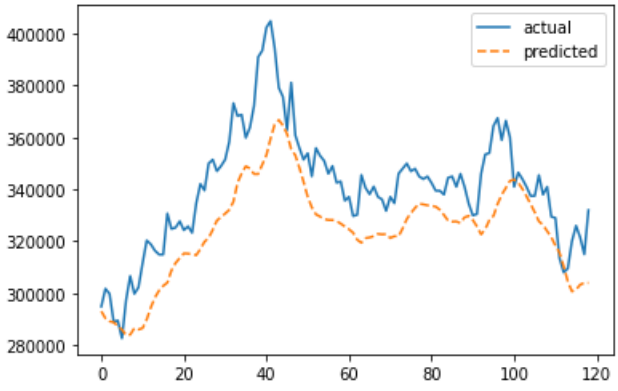


Fig. 4. RMSE=0.0723 (POSCO)

### 3.2 데이터 및 파라메터

본 연구에서는 주가 예측 요소 분석을 위해 2016년 8월 31일부터 2021년 8월 30일까지 전처리 과정에서 제외된 데이터를 제외한 1218일의 일일 주가 데이터를 사용하였다. 일반적으로 훈련 데이터셋과 테스트 데이터셋의 비율을 7:3으로 하는 경우가 많은데, 본 연구의 사전 실험에서 비율을 9:1로 하여 LSTM 모델을 실행할 경우 더 작은 오차값을 보여준 바, 본 연구에서는 1096일의 훈련 데이터셋과 122일의 테스트 데이터셋을 사용하였다. 본 연구는 주가 예측값을 찾아서 사용하는 것이 최종 목적이 아니고, 적절한 주가 예측 요소의 조합과 주가 예측을 위한 적절한 선행 데이터 일수 등을 찾아가는 과정을 검토하는 것이므로 교차 검증을 위한 밸리데이션 데이터셋은 별도로 분류하지 않았다.

주식시장의 업종을 구분하는 데는 다양한 분류가 있는데, 본 연구에서는 네이버 증권에서도 사용하고 있는 WICS(Wise Industry Classification Standard)를 사용하였다[17]. WICS는 총 10개의 대분류, 28개의 중분류, 83개의 소분류로 구성되어 있는

데, 본 연구에서는 대분류를 기준으로 하여 업종별로 대표적인 코스피 종목을 1개씩 임의로 선정하여 주가 데이터를 사용하였다. Table 1은 본 연구에서 사용한 종목들을 보여준다.

**Table 1.** Stock selection by WICS major classification

| 업종 | 종목 | 업종 | 종목 |
|---|---|---|---|
| 에너지 | 한화솔루션 | 건광관리 | 녹십자 |
| 소재 | POSCO | 금융 | 하나금융지주 |
| 산업재 | LS | IT | LG디스플레이 |
| 경기관련소비재 | 신세계 | 커뮤니케이션서비스 | SK텔레콤 |
| 필수소비재 | 하이트진로 | 유틸리티 | SK가스 |

단어검색 빈도는 네이버 데이터랩에서 선정된 주식 종목에 맞는 키워드를 사용하여 상기 주가 데이터와 같은 기간 동안 얻어진 데이터를 사용하였다. 예를 들면, '녹십자'와 관련된 검색 단어로는 '백신', '바이오', '제약', '건강'을 사용하였고, '한화솔루션'에는 '친환경', '태양광', '에너지'를 사용하였다. 주가는 시가, 고가, 저가 종가, 등 여러 가지가 있는데, 수정 종가(adjusted close price)를 사용했다. 환율과 이자율은 한국은행 경제통계시스템에서 제공하는 환율 데이터와 3년 만기 국고채 금리를 이용하였다.

또한 LSTM 모델 실행 시, 입력 변수별로 측정 단위가 다르고, 업종별로 주가 범위가 다르므로 효과적인 학습과 업종별 결과를 비교하기 위하여 입력데이터와 출력 RMSE를 최소-최대 정규화(Min-Max Normalization) 방법을 사용하여 정규화하였다. 또 본 연구의 모든 실험에서 val_loss 값이 에포크 수 50 이전에 충분히 떨어져서 더 이상 낮아지지 않는 모습을 보였기 때문에, 과적합을 피하기 위해 에포크(epochs)는 50으로 설정하면 충분하였다. Fig. 5는 loss와 val_loss 그래프의 예시를 보여준다.

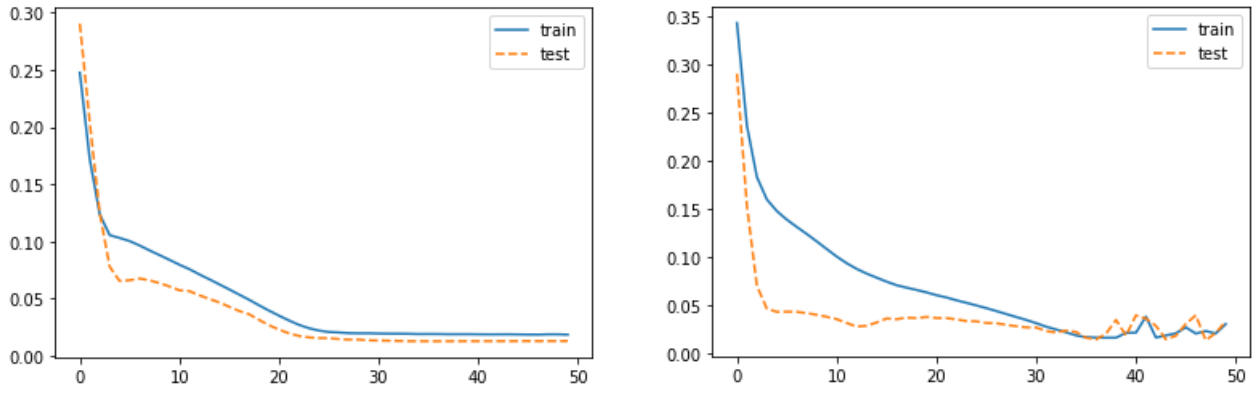


**Fig. 5.** Examples of loss (solid line) and val_loss (dotted line) for setting epochs

### 3.3 실험 결과 및 분석

전체 실험은 10개 종목에 대하여, 각 종목당 29개의 주가 예측 요소들의 조합을 구성하였으며, 주가 예측에 필요한 적정 선행 주가 일수를 찾아보기 위하여 1일(t-1), 2일(t-2, t-1), 3일(t-3, t-2, t-1), 4일(t-4, t-3, t-2, t-1), 5일(t-5, t-4, t-3, t-2, t-1), 10일(t-10, ... , t-1), 20일(t-20, ... , t-1), 30일(t-30, ... , t-1), 40일(t-40, ... , t-1), 50일(t-50, ... , t-1)의 선행 주가를 가지고 총 2900개의 실험이 행해졌다.

Table 2는 업종(종목)별 RMSE 분석을 보여주는데, 전체 10종목 중 최저 RMSE를 가지는 주가 예측 요소 조합에 다른 요소의 조합 없이 자체 과거 주가(P)만 있는 경우는 3 종목 뿐으로, 과거 주가 단일 변량보다는 다른 지표를 포함한 다변량 예측이 더 높은

**Table 2.** RMSE analysis by stock

| 종목 | 주가 예측 요소 조합별 최저 RMSE 범위 | 최저 RMSE 주가 예측 요소 조합 | 분석 |
|---|---|---|---|
| 신세계 | 0.0168 ~ 0.0231 | W+I<br>V+W+I | 0.02 이하: 20 조합, 0.02 이상: 9 조합<br>10 종목 중 가장 낮은 수준의 RMSE. |
| | 0.02 이상 9 조합 중 8 조합에 코스피 지수(K) 포함. 경기 관련 주로서 코스피 지수(기술 지표)의 영향력이 가장 떨어짐.<br>최적의 조합은 시장 심리 지표(W)와 거시 지표 이자율(I) | | |
| LG 디스플레이 | 0.0173 ~ 0.0272 | V+K, K+W<br>V+K+W | 0.02 이하: 21 조합, 0.02 이상: 8 조합 |
| | 0.02 이상 8 조합 모두에 이자율(I) 포함. 거시 지표의 영향력이 가장 떨어진다.<br>최적의 조합은 기술 지표(V, K)와 시장 심리 지표(W) | | |
| 하이트진로 | 0.0179 ~ 0.0278 | V, E<br>V+E | 0.02 이하: 12 조합, 0.02 이상: 17 조합 |
| | 0.02 이상 17 조합 중 14 조합에 코스피 지수(K) 포함. 경기 관련 주로서 코스피 지수의 영향력이 가장 떨어짐.<br>0.02 이하 12 조합 중 7 조합에 거래량(V) 포함. 거래량에 영향을 많이 받음.<br>최적의 조합은 기술 지표 거래량(V)과 거시 지표 환율(E) | | |
| 한화솔루션 | 0.0193 ~0.0317 | P (자체 주가)<br>V, I | 0.02 이하: 5 조합, 0.02 이상: 16 조합, 0.025 이상: 8 조합<br>뚜렷한 특징이 없음 |
| | 최적의 조합은 기술 지표 거래량(V)과 거시 지표 이자율(I) | | |
| SK가스 | 0.0249 ~ 0.0378 | V+W+I | 0.03 이하: 23 조합, 0.03 이상: 6 조합 |
| | 업종은 유틸리티로서 환율(E)이 0.03 이하 23 조합 중 가장 적은 7 조합, 0.03 이상 6 조합 모두에 포함되어 가장 적은 영향을 미침. 최적의 조합은 기술 지표 거래량(V), 시장 심리 지표(W), 거시 지표 이자율(I) | | |
| LS | 0.0252 ~ 0.0423 | P (자체 주가)<br>V+K+W+E | 0.03 이하: 19 조합, 0.03 이상: 9 조합, 0.04 이상: 1 조합 |
| | 업종은 산업재로서 환율(E)이 0.03 이하 19 조합 중 가장 많은 12 조합, 0.03 이상 10 조합 중 단 2 조합에만 포함되어 가장 큰 영향을 미침. 최적의 조합은 기술 지표(V, K), 시장 심리 지표(W), 거시 지표 환율(E). | | |
| 녹십자 | 0.0267 ~ 0.0392 | P (자체 주가)<br>V, W, E, I | 0.03 이하: 13 조합, 0.03 이상: 16 조합 |
| | 코스피 지수(K)가 0.03 이상 16 조합 중 14 조합에 포함되고, 0.03 이하 13 조합에 전혀 포함되지 않음으로 가장 적은 영향을 미침. 최적의 조합은 기술 지표 거래량(V), 시장 심리 지표(W), 거시 지표(E, I). | | |
| 하나 금융지주 | 0.0270 ~ 0.0353 | E, V+K+E<br>V+W+K+E | 0.03 이하: 10 조합, 0.03 이상: 19 조합 |
| | 환율(E)이 0.03 이하 10 조합 중 가장 많은 8 조합, 0.03 이상 19 조합 중 가장 적은 6 조합에 포함되어 가장 큰 영향을 미침.<br>최적의 조합은 기술 지표(V, K), 시장 심리 지표(W)), 거시 지표 환율(E). | | |
| POSCO | 0.0314 ~ 0.0443 | I<br>W+I | 0.04 이하: 23 조합, 0.04 이상: 6 조합<br>뚜렷한 특징이 없음 |
| | 최적의 조합은 시장 심리 지표(W)와 거시 지표 이자율(I) | | |
| SK텔레콤 | 0.0313 ~ 0.0438 | W+E, V+W+E | 0.04 이하: 24 조합, 0.04 이상: 5 조합 |
| | 단어 검색(W)과 코스피 지수(K)가 0.04 이상 5 조합 중에서 각 1회만 포함, 0.04 이하 24 조합 중에서 가장 많은 각 13회 포함되고, 0.04 이하 5 조합 모두에 거래량(V) 포함. 단어검색과 거래량, 코스피 지수에 가장 많은 영향을 받음.<br>최적의 조합은 기술 지표 거래량(V), 시장 심리 지표(W), 거시 지표 환율(E). | | |

예측력을 가진다고 할 수 있겠다. 주가 예측 요소별 최적의 조합에는 업종별 차이나 패턴이 보이지는 않았으며, 전체적으로 거래량(V)과 시장 심리 지표(W)가 전체 10 종목 중 8 종목에 포함되어 있어, 가장 주가에 영향을 미치는 두 요소라고 할 수 있겠다.

종목들을 최저 RMSE에 따라 세 그룹으로 나눠보면 먼저 신세계, LG디스플레이, 하이트진로, 한화솔루션은 종목 최저 RMSE가 0.02 이하로 나타났다. 이는 실험 종목 중 가장 높은 예측력을 보여준다. 신세계와 하이트진로는 경기관련주이고, LG디스플레이는 IT 업종이기는 하나 역시 개인 소비자와 가까운 가전제품도 포함한다는 점에서 경기와 관련이 있다고 보여진다. 주가 예측 요소 조합을 분석해 보면, 특히 신세계와 하이트진로 같은 경기관련주에서는 높은 RMSE를 가지는 조합에 코스피 지수(K)가 압도적으로 포함된 것을 볼 때 시장의 평균보다는 경기에 더 영향을 받는다고 유추해 볼 수 있다. 또한 최저 RMSE 주가 예측 조합에 코스피 지수(K)가 포함되는 종목은 10 종목 중 3 종목 뿐이다. 즉, 코스피 지수의 영향력이 상대적으로 적다는 결과인데, 이는 여러 가지로 해석될 여지가 있지만, 표면적으로는 시장 평균 수익률에 따르는 주식들이 많지는 않은 것으로 보인다. 즉, 우리나라 주식시장은 소수의 종목이 시장 수익률을 선도하는 것으로 볼 수 있겠다.

두 번째 그룹은 최저 RMSE가 0.02~0.03인 종목들로 SK가스, LS, 녹십자, 하나금융지주가 속하며, 거래량(V)과 투자자 관심(W)이 모든 최적 예측 요소 조합에 포함된다. 개별적으로, SK가스는 유틸리티 업종으로 국내 소비가 이루어지므로 환율(E)의 영향을 받지 않는 듯하며, 반대로 LS는 산업재로서 또 하나금융지주는 금융 업종으로서 환율의 영향을 크게 받는 것처럼 보인다.

끝으로 최저 RMSE가 0.03 이상으로 가장 높은 그룹에는 POSCO와 SK텔레콤이 속하는데, SK텔레콤은 30만원 대의 고가 종목으로 기술 지표인 거래량(V), 거시 지표인 환율(E), 그리고 시장 심리 지표인 종목 관련 단어검색(W)에 영향을 받는다.

주가 예측에 필요한 적정 선행주가 일수에 대한 분석을 보면, 총 290개(10종목 × 29 예측 요소 조합) 조합에 대한 실험에서 약 3분의 2에 해당하는 184개의 조합에서 1일을 예측 선행 일수로 사용할 때 가장 낮은 RMSE가 나왔으며, 2일이 그 다음으로 37개가 나왔다(Table 3). 이는 현재 시점에서 먼 과거의 데이터보다는 최근의 데이터가 더 예측력이 높다는 것을 의미한다. 따라서 아래 기술하는 고 가격대의 가격변동이 큰 일부 종목을 제외한다면 최적의 예측 선행 일수는 1~2일이라고 하겠다. 하지만 과거 30일, 40일, 그리고 50일 동안의 과거 주가를 이용한 예측에서도 최저 RMSE를 가지는 상당수의 조합(각각 14, 11, 10개)이 존재하는 것으로 보아, 앞으로 추가 연구를 한다면, 30일 이상 동안의 과거 데이터를 이용한 예측도 종목에 따라 충분히 높은 예측력을 가지는 최적 모델을 찾을 수 있을 것으로 보인다.

**Table 3.** Length of input dates for lowest RMSE

| 예측 선행 일수 | 1 | 2 | 3 | 4 | 5 | 10 | 20 | 30 | 40 | 50 |
|---|---|---|---|---|---|---|---|---|---|---|
| 예측 요소 조합수 | 184 | 37 | 14 | 6 | 6 | 2 | 6 | 14 | 11 | 10 |

특히, 2일부터 50일 사이의 최저 오차 예측 선행 일수 106개 중 POSCO와 SK텔레콤 두 종목에서 약 3분 1인 38개의 조합이 나왔다(Table 4). 두 종목의 58개 예측 요소 조합 중 3분의 2에 해당하는 38개 변수 조합이 이에 해당한다. 두 종목의 가격대는 13만 원대에서 40만 원대이며 이는 상대적으로 높은 가격으로(예를 들면 LG디스플레이의 가격대는 8천 원대에서 3만 8천 원대), 가격대가 높은 주식은 인기가 있고 거래가 활발한 종목으로, 이런 주식은 주가에 영향을 미치는 변수를 많이 고려할 경우, 1일보다는 더 많은 선행 일수의 데이터를 이용하여 예측하는 것이 좋다는 것을 암시하는 것일 수 있다. 가격대 이외의 주가 예측 요소의 조합 등에서는 뚜렷한 패턴이 보이지는 않는다. 이 두 종목의 RMSE는 주가 예측 요소 조합별 최저값이 모두 0.034 이상으로 다른 종목들에 비하여 높다(예를 들면, 신세계의 최저 RMSE는 0.016 대이고 SK가스의 최저 RMSE는 0.024 대이다).

Table 4. Length of input dates over 2 of POSCO and SKT for lowest RMSE

| 업종 | 종목 | 가격대 | 최저 오차 예측 선행 일수 | 최저 RMSE | 주가 예측 요소 조합수 |
|---|---|---|---|---|---|
| 소재 | POSCO | 130,368~404872 | 2 | 0.0329 ~ 0.0373 | 7 |
| | | | 3 | | 3 |
| | | | 4 | | 1 |
| | | | 30 | | 3 |
| | | | 40 | | 4 |
| | | | 50 | | 5 |
| 커뮤니케이션서비스 | SK텔레콤 | 156,252~328,839 | 2 | 0.0342 ~ 0.0438 | 2 |
| | | | 3 | | 2 |
| | | | 4 | | 2 |
| | | | 5 | | 2 |
| | | | 20 | | 3 |
| | | | 30 | | 5 |
| | | | 40 | | 3 |
| | | | 50 | | 1 |

끝으로, 전체 주가 예측 요소 조합별 RMSE 분석을 보면(Table 5), 모든 예측 요소를 다 포함한 모델(V+K+W+E+I)이 전체 평균에서나 최저 RMSE만의 평균에서나 가장 높게 나와 가장 예측력이 떨어짐을 알 수 있다. 반면, 과거 주가(P) 단일 변량이거나 포함된 예측 요소 수가 적을수록 RMSE가 낮아지는 경향을 볼 수 있다. 이는 무조건 많은 예측 요소를 모델에 포함하기보다는 업종이나 종목의 특성을 찾아 적절한 예측 요소를 선별적으로 포함할 필요가 있는 것으로 생각된다.

Table 5. RMSE analysis by combination of features to predict stock price

| | 분석 | 조합 |
|---|---|---|
| 주가 예측 요소 조합별 전체 RMSE 평균 | ~ 0.04 : 8 조합 | P, V, K, W, E, V+K, V+W, V+E |
| | 0.04 ~ 0.05: 19 | I, V+I, K+W, W+E, W+I, K+E, K+I, E+I, V+K+W, V+W+E, V+W+I, V+K+E, V+K+I, V+E+I, K+W+I, K+E+I, V+K+W+E, V+K+W+I, V+W+E+I |
| | 0.05 ~ : 2 | W+K+E, V+K+W+E+I |
| 주가 예측 요소 조합별 최저 RMSE 평균 | 0.025 ~ : 4 조합 | P, E, I, W+E |
| | 0.026 ~ : 7 | V, W, V+W, V+E, W+I, V+W+E, V+W+I |
| | 0.027 ~ : 7 | V+I, E+I, V+K+E, V+K+I, V+K+W+E, V+K+W+I, V+W+E+I |
| | 0.028 ~ : 7 | K, V+K, W+K, K+E, V+K+W, K+W+E, K+W+I |
| | 0.029 ~ : 2 | K+I, V+E+I |
| | 0.030 ~ : 2 | K+E+I, V+K+W+E+I |

## 4. 결 론

이상의 실험으로 우리는 주가 예측에 있어 기술적 지표, 거시 경제 지표, 시장 심리 지표 등을 별도로 사용하기보다는 모두 포함하여 적절한 조합을 찾아서 사용할 필요를 확인하였다. 단, 개별 종목이나 업종 등의 특성에 맞게 적절한 조합을 찾아가야 하

는 것이 앞으로의 과제라 하겠다. 본 연구에서 확인된 바로는 경기 관련 업종이나 종목에는 코스피 지수의 영향력은 떨어진다. 산업재나 금융 업종에는 환율의 영향이 큰 것으로 나타났으며, 반대로 유틸리티 업종은 환율의 영향이 적은 것으로 나타났다. 그리고 고가 종목은 시장 심리 지표인 종목 관련 단어검색에 영향을 많이 받고, 상대적으로 예측 선행 주가 일수를 늘리는 것이 예측력 제고에 도움이 되는 것으로 보인다. 고가 종목을 제외한 종목들은 1~2일 정도의 선행 주가 일수를 가지고 예측하는 것이 좋은 것으로 나타났다.

본 연구는 WICS 업종 분류법에 따라 업종당 1개 종목씩을 분석하였다. 추후 좀 더 다양하고 세밀한 특성으로 업종을 분류하고 더 많은 주식 종목을 포함하여, 연구의 결과를 일반화하여야 할 것이다.

## Acknowledgement

※ 본 연구는 2021년도 서경대학교 교내연구비 지원으로 이루어졌음(This research was supported by Seokyeong University in 2021).

## References

1. J. Heo, and J. Yang, “SVM Based Stock Price Forcasting Using Financial Statements”, KIISE Transactions on Computing Practices, Vol. 21, No. 3, pp. 167-172, 2015.
2. G. Nam, J. Jang, H. Oh, and T. Kim, “Development of Stock Investment System Using Machine Learning”, Proceeding of Information Processing Society Fall Conference, Vol. 24, No. 2, pp. 810-812, 2017.
3. A. Siripurapu, “Convolutional Networks for Stock Trading”, Stanford University, 2015.
4. I. Joo, and S. Choi, “ Stock Prediction Model Based on Bidirectional LSTM Recurrent Neural Network”, Journal of Korea Institute of Information, Electronics, and Communication Technology, Vol. 11, No. 2 , pp. 204-208, 2018.
5. J. Jung, and J. Kim, “A Performance Analysis by Adjusting Learning Methods in Stock Price Prediction Model Using LSTM”, Journal of Digital Convergence, Vol. 18, No. 11, pp. 259-266, 2020.
6. S. Hong, “A Research on Stock Price Prediction Based on Deep Learning and Economic Indicators”, Journal of Digital Convergence, Vol. 18, No. 11, pp. 267-272, 2020.
7. D. Shin, K. Choi, and C. Kim, “Deep Learning Model for Prediction Rate Improvement of Stock Price Using RNN and LSTM”, Journal of KIIT, Vol. 15, No. 10, pp. 9-16, 2017.
8. T. Fischer, and C. Krauss, “ Deep Learning with Long Short-Term Memory Networks for Financial Market Predictions”, European Journal of Operational Research, Vol. 270, No. 2, pp. 654-669, 2018.
9. Y. Cho, K. Sohn, and O. Kwon, “Comparison of Models for Stock Price Prediction Based on Keyword Search Volume according to the Social Acceptance of Artificial Intelligence”, Journal of Intelligence and Information Systems, Vol. 27, No. 1, pp. 103-128, 2021.
10. M. Kim, J. Ryu, D. Cha, and M. Sim, “Stock Price Prediction Using Sentiment Analysis: from "Stock Discussion Room" in Naver”, The Journal of Society for e-Business Studies, Vol. 25, No. 4, pp. 61-75, 2020.
11. M. Zeng, X. Leng, and G. Zhai, “Twitter Mood Predicts the Stock Market”, Journal of Computational Science, Vol. 2, No. 1, pp. 1-8, 2011.
12. E. Jang, “LSTM Combination of BERT Sentiment Analysis and Time Series Macroeconomy Index for Predicting Stock Price”, Master Thesis, Korea University, 2020.
13. N. Seong, and K. Nam, “Combining Macro-Economical Effects with Sentiment Analysis for Stock Index Prediction”, Entrue Journal of Information Technology, Vol. 16, No. 2, pp. 41-54, 2017.

14. S. Siami-Namini, N. Tavakoli, and A. Namin, “A Comparison of ARIMA and LSTM in Forecasting Time Series”, In 17th IEEE International Conference on machine Learning and Applications (ICMLA), pp. 1394-1401, 2018.
15. F. Qian, and X. Chen, “Stock Prediction Based on LSTM under Different Stability”, In 4th IEEE International Conference on Cloud Computing and Big Data Analysis (ICCCBDA), pp. 483-486, 2019.
16. S. Goel, J. Hofman, S. Lahaie, D. Pennock, and D. Watts, “Predicting Consumer Behavior with Web Search”, Proceedings of the National Academy of Sciences, Vol. 107, No. 41, pp. 17486-17490, 2017.
17. FnGuide, https://www.fnguide.com