JKSCI

한국컴퓨터정보학회논문지
Journal of The Korea Society of Computer and Information
Vol. 28 No. 8, pp. 67-75, August 2023
https://doi.org/10.9708/jksci.2023.28.08.067

Comparison of Stock Price Prediction Using Time Series and Non-Time Series Data

Min-Seob Song*, Junghye Min*

*Student, Dept. of Computer Science, Inha Technical College, Incheon, Korea
*Professor, Dept. of Computer Science, Inha Technical College, Incheon, Korea

[Abstract]

Stock price prediction is an important topic extensively discussed in the financial market, but it is considered a challenging subject due to numerous factors that can influence it. In this research, performance was compared and analyzed by applying time series prediction models (LSTM, GRU) and non-time series prediction models (RF, SVR, KNN, LGBM) that do not take into account the temporal dependence of data into stock price prediction. In addition, various data such as stock price data, technical indicators, financial statements indicators, buy sell indicators, short selling, and foreign indicators were combined to find optimal predictors and analyze major factors affecting stock price prediction by industry. Through the hyperparameter optimization process, the process of improving the prediction performance for each algorithm was also conducted to analyze the factors affecting the performance. As a result of feature selection and hyperparameter optimization, it was found that the forecast accuracy of the time series prediction algorithm GRU and LSTM+GRU was the highest.

▸ **Key words**: stock price prediction, non-time series, time series, deep learning, optimal predictors, hyperparameter

[요 약]

주가 예측은 금융시장에서 중요하게 다뤄지고 있는 주제이지만 영향을 미칠 수 있는 다수의 요소들로 인해 어려운 주제로 고려되고 있다. 본 논문에서는 시계열 예측 모델 (LSTM, GRU)과 데이터의 시간적 의존성을 고려하지 않는 비 시계열 예측 모델 (RF, SVR, KNN, LGBM)을 주가 예측에 적용하여 성능을 비교하고 분석하였다. 또한 주가 데이터와 기술적 분석 보조지표, 재무제표 지표, 매수매도 지표, 공매도, 외국인 지표 등 다양한 데이터를 조합 및 활용하여 최적의 예측 요소를 찾아내고 업종별로 주가 예측에 영향을 미치는 주요 요소들을 분석했다. 하이퍼파라미터 최적화 과정을 통해 알고리즘별 예측 성능을 향상 시키는 과정도 진행하여 성능에 영향을 주는 요인을 분석하였다. 변수 선택과 하이퍼 파라미터 최적화 과정을 거친 결과, 시계열 예측 알고리즘인 GRU, 그리고 LSTM+GRU의 예측 정확도가 가장 높은 것으로 나타났다.

▸ **주제어**: 주가 예측, 비 시계열, 시계열, 딥러닝, 예측 요소, 하이퍼파라미터

• First Author: Min-Seob Song, Corresponding Author: Junghye Min
*Min-Seob Song (magnet9805@naver.com), Dept. of Computer Science, Inha Technical College
*Junghye Min (jhmin@inhatc.ac.kr), Dept. of Computer Science, Inha Technical College
• Received: 2023. 08. 03, Revised: 2023. 08. 17, Accepted: 2023. 08. 22.

http://www.ksci.re.kr pISSN:1598-849X | eISSN:2383-9945

I. Introduction

주가 예측은 금융 분야에서 중요한 연구 주제이지만 동시에 정확도 확보가 어려운 주제이기도 하다. 주가 데이터는 시간적 패턴과 불안정성을 갖는 대표적인 시계열 데이터로서, 정확한 예측을 위해서는 심층적인 분석과 예측 모델들의 적용이 필요하다. 최근에는 데이터의 순서와 시간적 의존성을 고려하여 패턴을 파악하고 예측하는 방법을 사용하는 시계열 예측 알고리즘과 데이터 간의 순서와 시간적 의존성을 고려하지 않고 독립적으로 분석하는 비 시계열 예측 알고리즘이 다양한 분야에서 사용되고 있으며, 이들의 장점과 한계를 이해하는 것이 주가 예측 연구에 있어서 중요한 과제이다. 특히 2020년부터 시작된 코로나 상황에서는 주가의 변동성이 크기 때문에 정확한 주가 예측이 어려울 것으로 예측되었다.

본 연구에서는 코로나 시기를 포함한 기간의 주가 예측 정확도 분석을 위하여 시계열 예측 알고리즘과 비 시계열 예측 알고리즘을 적용하여 연구를 수행했다. 이러한 두 가지 접근 방법의 상이점을 탐구하고, 어떤 상황에서 어떤 알고리즘이 더 적합한지에 대한 연구를 위해 주가 데이터를 활용했다. 주가 데이터는 시계열 데이터의 대표적인 예시로, 그 특성상 다양한 변동성과 패턴을 보여준다. 주가 데이터는 기본적으로 시가, 저가, 고가, 종가, 거래량 등의 데이터가 있으며 최근에는 이를 이용해 기술적 분석의 보조지표로 사용되는 이동평균, WMA(Weighted Moving Average), RSI(Relative Strength Index) 등이 주가 예측에 활용된다[1-2]. 주가 예측에 있어 BPS(Book-value Per Share), PBR(Price on Book-value Ratio)과 같은 재무제표지표 또한 중요한 예측변수 중 하나로 담당하고 있다[3-4].

해당 연구에서는 기본적인 데이터와 보조지표, 재무제표지표, 매수매도 지표, 외국인지표, 공매도지표 등을 추가로 활용하여 시계열과 비 시계열 예측 알고리즘들 간의 주가 예측 성능의 차이를 제시한다. 실험 결과를 통해 코로나 시기의 더욱 다양하고 격동적인 변동성과 패턴에 대해서 데이터의 시계열성이 예측에 미치는 영향을 파악하고, 더 나아가 데이터 분석과 예측 분야에 적용할 수 있는 유용한 정보를 제공하는 것을 목적으로 한다. 업종별 영향을 분석하기 위해 IT(삼성전자), 소재(LG 화학), 금융(KB 금융) 분야의 주가 예측 데이터를 분석하여 업종별 변수의 중요도를 비교했으며, 전체적인 알고리즘 성능 예측을 비교하기 위해서는 IT 업종인 삼성전자 주가 데이터를 이용하였다.

본 논문은 2장에서 시계열, 비 시계열 예측 알고리즘의 연구 동향과 주가의 예측 요소 데이터, 주가 예측 모델, 변수 선택법, 그리고 평가 지표를 제시한다. 3장에서는 제안한 알고리즘을 적용하고, 업종별 변수 선택을 적용한 과정과 결과에 대해 나타낸다. 4장에서는 하이퍼파라미터 최적화 및 결과를 기술하고, 마지막으로 결론을 맺는다.

II. Preliminaries

1. Related works

최근 딥러닝 기반 시계열 알고리즘 연구는 활발하게 진행되고 있으며, 특히 RNN(Recurrent Neural Network)의 장기 의존성 문제를 완화하기 위한 LSTM(Long Short-Term Memory)과 GRU(Gated Recurrent Unit)와 같은 모델들이 널리 사용되고 있다. 이러한 알고리즘들은 데이터의 시간적 특성을 잘 파악하고, 기존의 RNN 모델과 비교하여 기울기 소실(vanishing gradient) 문제를 해결한 장점을 갖고 있다. 또한 금융 분야뿐만 아니라, 자연어 처리, 음성인식, 이미지 분석 등 다양한 분야에서도 딥러닝 시계열 알고리즘이 적용되어 유의미한 결과를 얻고 있다.

비 시계열 정형 데이터 예측에는 RF(Random Forest) 등의 머신러닝 알고리즘들이 널리 활용되고 있어 데이터 분석과 예측 분야에 더욱 폭넓은 가능성과 유용성을 제공하고 있다.

Kun은 LSTM-CNN 알고리즘으로 인간 활동 인식 문제를 성공적으로 해결했다[5]. 이 알고리즘은 LSTM 레이어와 CNN(Convolutional Neural Nnetwork) 레이어를 유기적으로 결합하여 활동 특징을 자동으로 추출하고 분류하는 방법을 제시했으며, 고정된 활동 인식 문제에서 높은 정확도와 일반화 능력을 보여주었다.

Masooma는 지속 가능한 환경 개발을 위한 딥러닝 모델인 Spatial Feature Attention Based LSTM 모델을 제안한다[6]. 이 모델은 여러 기상 요소들의 공간적 및 시간적 관계를 정확하게 파악하여 온도를 예측하는 데에 활용되며, 기상 예측 및 기상 분석 분야에 기여한다.

곽재진은 한국의 전기차 수요를 다루고, 주요 요인을 밝히기 위해서 머신러닝을 활용하여 예측 모델을 구축했다[7]. RF 모델을 사용한 전기차 수요 예측은 다중선형회귀 분석 모델보다 높은 예측 정확도를 보였으며, 변수들의 상대적인 중요도를 연구하여 전기차 수요에 영향을 미치는 요인들을 더욱 정확하게 파악할 수 있었다.

본 논문의 연구 주제인 주가 예측에도 많은 연구가 진행되었다.

최훈은 딥러닝 및 머신러닝을 활용한 주식 예측 분석에 대해 연구했다[8]. 양방향 LSTM 순환 신경망을 이용한 주식 예측 모델이 기존의 다른 모델들보다 높은 예측률을 보였다.

유한정은 순환신경망을 이용한 S&P 500 지수의 예측에 대해 연구했다[9]. LSTM, GRU가 높은 예측률을 보였으며, 데이터 전처리와 정규화 방법, 적층 수에 따라 예측 성능이 크게 달라졌다.

Xiurui는 주가 예측을 위해 기업 간 공간 의존성을 고려한 GCN(Graph Convolutional Network)-LSTM을 사용했고, 분 단위 주가 데이터에서 다른 방법을 능가하는 성능을 보였다[10].

기존 주가 예측 논문들은 한정된 변수 다양성을 가지고 있으며, 알고리즘별 최적의 예측 요소와 하이퍼파라미터 최적화의 결과를 비교하지 않는다. 본 연구에서는 다양한 변수 활용과 비 시계열, 시계열 예측 알고리즘 비교를 통해 주가 예측에 새로운 관점을 제시한다.

2. Research Data

본 연구에서는 IT, 소재, 금융 분야의 KOSPI 상위 1위 기업인 삼성전자, LG화학, KB금융의 2018년 5월부터 2023년 3월까지의 5년 치 데이터를 사용했다.

우선, 주가 데이터, 재무제표지표, 매도매수, 외국인, 공매도 지표를 수집하기 위해 Naver, KRX 등 다양한 웹 사이트에서 주가 정보를 스크래핑하여 제공하는 API인 pykrx를 사용했다[11]. Table 1에 수집한 재무제표지표를 정리했다.

Table 1. Selected financial ratios

Name	Formula
BPS	(Shareholder's equity - Preferred equity) / Total common shares outstanding
EPS	Net earnings / Total shares outstanding
PER	Share price / EPS
PBR	Share price / BPS
DIV	Dividend per share / Share price
DPS	Dividends / Shares

또한, 기술적 분석의 보조지표를 수집하기 위해, TA-Lib API를 사용했다[12]. 기술적 분석은 주식 시장과 금융 시장을 분석하고 예측하는 기법 가운데 하나이다. 수집한 보조지표들은 추세를 파악하는 추세 지표와 주가 추세의 속도를 보여주는 모멘텀 지표이다. 추세 지표로는 MA(Moving Average), MOM(Momentum), WMA, SIG(Signal(n)t), CCI(Commodity channel index)를, 모멘텀 지표로는 Stochastic(Slow% K, Slow% D), LWR(Larry William's R%), ADO(Accumulation/Distribution Oscillator), RSI(Relative Strength Index), CCI(Commodity channel index)를 수집했다.

Table 2에 수집한 기술적 보조지표를 도출하는 공식들을 나타냈다.

전체적으로 수집한 데이터는 주가 데이터 9개, 기술적 보조지표 12개, 재무제표지표 6개, 매도매수 지표 9개, 외국인 지표 5개, 공매도 지표 3개로 총 45개의 변수가 있다. 이후 연구에는 날짜, 업종, 종목을 제외한 42개의 변수를 사용하였다.

3. Algorithms for non-time series prediction

정형 데이터 예측에는 다양한 알고리즘들이 널리 활용되며, 데이터의 구조와 특성에 따라 적절히 선택되어 사용되고 있다. 아래 알고리즘 에서는 데이터간 시간적 의존성은 고려되지 않은다.

3.1 SVM (Support Vector Machine)

데이터를 고차원 공간으로 매핑하여 클래스 간의 최대 마진을 찾는 분류 및 회귀 알고리즘이다. 결정 경계를 찾기 위해 클래스 간의 서포트 벡터를 활용하며, 이상치에 강하고 선형 및 비선형 문제를 처리 가능하다. 데이터가 선형적으로 구분되지 않는 경우 커널 기법을 이용하여 비선형 문제를 해결한다.

3.2 RF (Random Forest)

다수의 결정 트리를 앙상블 하여 분류 및 회귀를 수행하는 모델이다. 각 트리는 부트스트랩 샘플링과 무작위 특성 선택을 통해 생성되며, 다수의 트리 결과를 결합하여 과적합을 줄이고 예측의 안정성을 높인다.

3.3 KNN (K-Nearest Neighbors)

인접한 K개의 이웃 데이터를 활용하여 분류 또는 회귀하는 모델이다. 데이터 간 유사도를 측정하여 가장 가까운 이웃들을 선택하고, 분류 문제에서는 다수 클래스로 분류하며 회귀 문제에서는 주변 이웃들의 평균값을 예측값으로 사용한다. 학습 데이터의 분포에 따라 적응적으로 예측하며, 간단하고 직관적인 특성을 가진다.

3.4 LGBM (LightGBM)

Leaf-wise 분할 방식을 채택하여 속도와 성능을 향상시킨 그래디언트 부스팅 모델이다. Leaf-wise는 최대 손실 값의 리프를 먼저 분할하여 효율적인 트리를 구성한다. 이러한 방식은 더 정확한 예측을 가능케 하면서도 작은 리프에서부터 분할하기 때문에 효과적인 학습과 예측이 가능하며, 대용량 데이터 처리에 효과적이다.

3.5 XGBoost (Extreme Gradient Boosting)

규제화와 병렬 처리를 통해 높은 예측력을 가진 그래디언트 부스팅 모델이다. 작은 리프로부터 시작하여 깊은 트리를 생성하여 예측을 수행하며, 학습 단계에서 과적합을 방지하는 규제화를 적용한다. 트리의 깊이나 분할에 대한 제약을 부여하여 모델을 일반화하고 안정적으로 만든다.

3.6 AdaBoost (Adaptive Boosting)

약한 학습기들을 순차적으로 결합하여 성능을 향상시키는 앙상블 학습 모델이다. 각 학습기는 이전 학습기에서 잘못 분류된 샘플에 가중치를 부여하여 다음 학습에서 더 강한 모델을 생성한다. 이러한 방식으로 앙상블된 모델은 개별 모델보다 높은 정확도를 제공하며, 다양한 학습기의 결합으로 인해 더 강력한 예측 성능을 나타낸다.

4. Algorithms for time series prediction

최근 시계열 데이터를 분석하고 예측하는 데에 딥러닝 알고리즘들이 주로 활용되고 있으며, 기존의 머신러닝 모델들과 비교했을 때, 길이가 긴 시퀀스에 걸쳐있는 패턴을 효과적으로 인식하고 추출할 수 있으며, 더욱 효과적인 다중 입력 및 출력 작업을 가능하게 한다.

4.1 LSTM (Long Short Term Memory)

장기 기억 메커니즘을 갖춘 시계열 딥러닝 알고리즘으로, 게이트 메커니즘으로 현재 정보와 이전 정보를 조절하여 기억력을 강화하고, 기울기 소실 문제를 완화하여 장기적 패턴을 학습한다. 여러 개의 층(layer)으로 구성되며, 각 층은 입력 데이터를 처리하고 은닉 상태를 전달함으로

Table 2. Selected technical indicators

Indicator	Formula
MA (Simple n-day Moving Average)	$\frac{1}{n} \times \left(\sum_{i=t-n+1}^{t} C_i \right)$
MOM (Momentum)	$C_t - C_{t-n}$
WMA (WeightedMoving Average)	$\frac{n \times C_t + (n-1) \times C_{t-1} + \cdots + C_{t-n+1}}{n + (n-1) + \cdots + 1}$
Stochastic K%	$\frac{(C_t - LL_{t\_n})}{(HH_{t\_n} - LL_{t\_n})} \times 100$
Stochastic Slow% K (Stochastic D%)	$\frac{\sum_{i=0}^{n-1} K_{t-i}\%}{n}$
Stochastic Slow% D	$\frac{\sum_{i=0}^{m-1} SK_{t-i}\%}{m}$
SIG(Signal(n)t) (Parabolic Sar)	$SIG_{t-1} + AF(EP - SIG_{t-1})$
LWR (Larry William's R%)	$\frac{HH_{t\ t-n+1} - C_t}{HH_{t\ t-n+1} - LL_{t\ t-n+1}} \times 100$
ADO (Accumulation/DistributionOscillator)	$\frac{H_t - C_t}{H_t - L_t}$
RSI(Relative Strength Index)	$100 - 100 \div (1 + \frac{\sum_{i=1}^{n-1} N\_UP_{t-i}}{\sum_{i=1}^{n-1} N\_DW_{t-i}})$
CCI (Commodity channel index)	$\frac{M_t - SM_t}{0.015 D_t}$

C_t means the closing price at the time of t, H_t means the high price, L_t means the low price, HH_{t_n} means the maximum value for the period of n at the time of t, and LL_{t_n} means the minimum value. The AF of the SIG indicator is the acceleration factor, and the EP is the Extreme Price, which means an important market price, and if the highest and lowest prices do not come out, the previous EP is used as it is. M_t represents the average of closing price, high price, and low price, SM_t represents the simple moving average for the t period of M, and D_t represents the simple average for the t period of the absolute value of the difference between M_t and SM_t. In this thesis, variables were derived by setting AF as 0.02 and EP high of 0.2, MA as 5 and 20 days, SMA, WMA, and MOM as 10, Stochastic Slow% K and Stochastic Slow% D, RSI as 14 days, and CCI as 20 days.

써 복잡한 시계열 데이터의 특성을 학습하며 예측한다.

4.2 GRU (Gated Recurrent Unit)

LSTM보다 더 간결한 구조로 구현된 시계열 딥러닝 알고리즘으로, LSTM의 게이트 메커니즘을 간소화하여 연산 비용을 줄이면서도 장기 의존성 문제를 처리한다. Fig. 1은 GRU의 네트워크 구성도이다.

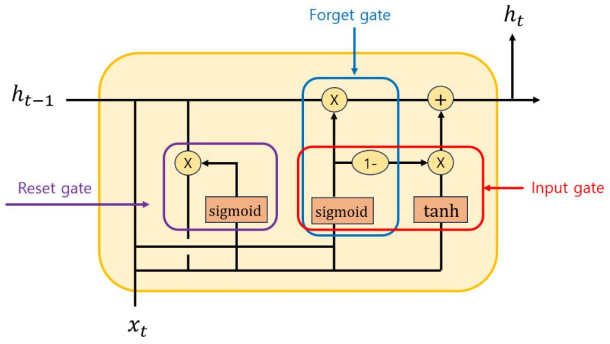

Fig. 1. GRU Network

GRU 네트워크는 재귀적인 구조를 가진 시계열 딥러닝 알고리즘으로, h(t-1)과 x(t)를 입력으로 받는다. 이전 시점의 은닉 상태인 h(t-1)과 현재 시점의 입력 데이터 x(t)를 각각 곱 연산과 합 연산을 통해 조합하여 sigmoid 함수와 tanh 함수를 거쳐서 reset gate, forget gate, input gate를 생성한다. Reset gate는 이전 은닉 상태를 얼마나 무시할지를 결정하고, forget gate는 이전 정보를 얼마나 버릴지를 결정하며, input gate는 현재 정보를 얼마나 반영할지를 결정한다. 이러한 게이트들을 활용하여 은닉 상태 h(t)를 업데이트하여 시계열 데이터의 장기 의존성을 파악하고 예측에 활용한다.

4.3 LSTM+GRU

LSTM과 GRU를 결합한 LSTM+GRU 모델은 두 모델의 강점을 조합하여 시계열 데이터를 효과적으로 학습하는 시계열 딥러닝 알고리즘이다. LSTM의 장기 의존성 파악과 GRU의 간결한 구조를 결합하여 모델의 성능을 향상시킨다. 주가 예측과 같은 시계열 데이터 분석에 높은 성능을 발휘한다.

5. Feature selection

이 연구에서는 앞서 언급한 연구 데이터들의 효율적인 조합을 위해 다섯 가지 변수 선택 기법을 활용했다.

5.1 RF feature_importances_

Random Forest (RF) 모델에서 제공하는 속성으로, 각 특성(Feature)이 모델의 예측에 얼마나 중요한 역할을 하는지를 나타내는 값이다.

5.2 RFE (Recursive Feature Elimination)

재귀적으로 특성을 제거하여 모델의 성능을 평가하는 방법이다. 초기에 모든 특성을 사용하여 모델을 학습하고, 중요하지 않은 특성들을 제거하면서 성능을 평가하고 최종적으로 가장 중요한 특성들만 남기는 방식이다.

5.3 RFECV (Recursive Feature Elimination with Cross-Validation)

RFE와 유사하지만 교차 검증을 추가로 수행하여 특성 선택의 안정성을 높인 방법이다. 교차 검증을 통해 여러 번의 모델 학습과 평가를 수행하며, 각 특성의 중요도를 평균하여 최종적으로 가장 중요한 특성들을 선택한다.

5.4 Lasso regression

L1 규제를 적용하는 선형 회귀 모델로, L1 규제는 일부 특성들의 계수를 0으로 만들어 변수 선택(Feature Selection)에 활용된다.

5.5 Lidge regression

L2 규제를 적용하는 선형 회귀 모델로, L2 규제는 모든 특성의 계수를 0에 가깝게 조절하여 과적합을 완화하는 데 사용한다.

6. Metrics

이 연구에서는 회귀 문제인 종가 예측을 수행하였으며, 이를 평가하기 위해 평균 제곱근오차 RMSE(Root Mean Squared Error)와 결정계수 R2(R-squared)를 사용하였다.

Table 3에 해당 평가지표들의 공식을 나타냈다.

Table 3. Formula of metrics

Name	Formula
RMSE	$\sqrt{\frac{\sum_{i=1}^{n}(Actual_i - Predicted_i)^2}{n}}$
R2	$1 - \frac{\sum_{i=1}^{n}(Actual_i - Average)^2}{\sum_{i=1}^{n}(Actual_i - Predicted_i)^2}$
$Actual_i$: Actual value at time i $Predicted_i$: Predicted value at time i	

III. Selection of Models and Features

이 연구에서는 실험을 위해 Jupyter Notebook을 사용하였으며, Python 언어와 그에 포함된 라이브러리인 pandas, scikit-learn, 그리고 Tensorflow를 활용했다. 알고리즘 선택을 위해 삼성전자의 종가를 이용하여 모델별 test set의 RMSE, R2를 비교하였으며, Min-Max Scaling을 통해 (0,1) 사이의 값으로 구현하였다. train set, test set의 비율을 70:30으로 2018년 5월부터 2021년 9월 까지의 데이터를 학습 데이터에, 2021년 10월부터 2023년 3월 까지의 데이터를 테스트 데이터로 사용하였다. 학습과 예측을 마친 후에는 RMSE의 용이한 확인을 위한 inverse transformation을 진행했다. 비시계열 예측에 사용된 알고리즘은 SVR, RF, KNN, XGB, LGBM, ADABOOST 총 6가지 모델이다. 시계열 데이터를 이용한 예측에는 알고리즘은 LSTM, GRU, LSTM+GRU 모델을 사용했다.

성능이 좋은 알고리즘을 채택하기 전에, 먼저 시간의 의존성을 고려않는 SVR, RF, KNN, XGB, LGBM, ADABOOST 알고리즘에도 n일 치 종가와 같이 시계열 관점에서의 데이터를 입력하는 것이 유의미한가를 살펴보기 위해 1일 전, n일 치 종가를 입력하여 예측 성능을 비교했다. n은 시계열 예측 알고리즘에 10, 15, 20일 치 종가를 입력하여 예측 성능을 비교하여 평균적으로 우수한 일 수를 채택하였다. Fig. 2.에 그 결과를 나타냈다.

평균적으로 15일 치 종가를 입력한 예측 오차가 낮았다. 이어서, Fig. 3.에 비 시계열 예측 알고리즘별 1일 전, 15일 치 종가 일수에 대한 test set의 RMSE 결과에 대해서 나타냈다.

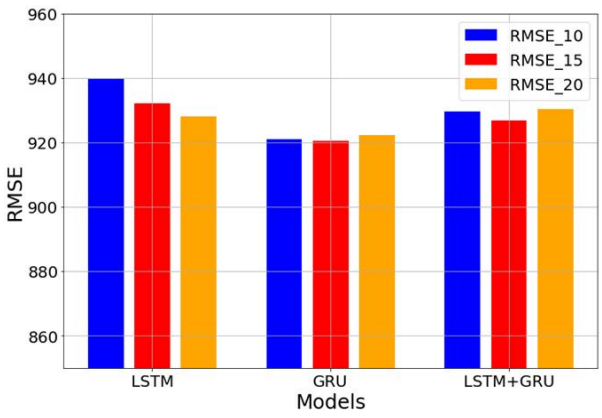

Fig. 2. RMSE Comparison of 10 days, 15 days, 20 days close of time-series models

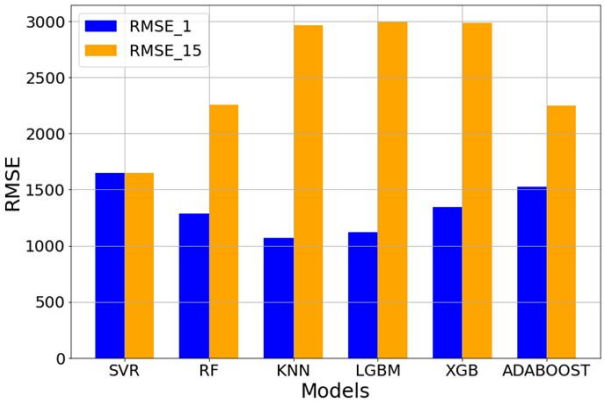

Fig. 3. RMSE of non-time-series models using the close of 15 days or 1 day

SVR을 제외한 모든 비 시계열 예측 모델에서 15일치 종가를 입력으로 한 모델보다 1일 전 종가를 입출력으로 한 모델의 예측력이 확연히 우수한 모습을 보인다. 비 시계열 예측 알고리즘에 시계열 관점에서의 데이터를 적용하는 것은 유의미한 영향을 미치지 않는다는 것을 보이며 이후 알고리즘 비교 과정에 SVR을 제외한 모든 비 시계열 예측 알고리즘에 1일 치 종가를 입출력하여 비교한다.

Fig. 4, Table 4는 비 시계열, 시계열 예측 알고리즘별 1일 전 또는 15일 치 종가를 입력하여 test set의 RMSE를 비교하는 결과를 나타냈다. 알고리즘 비교 결과, 비교적 좋은 성능을 보인 시계열 예측 알고리즘 LSTM, GRU, LSTM+GRU와 비 시계열 예측 알고리즘 RF, KNN, LGBM이 이후 과정에 사용될 알고리즘으로 채택되었다.

이 후 변수 선택 과정으로 2.2에 설명된 42개의 변수를 아래의 5가지 변수 선택법에 적용하였다. 변수 선택법으로는 (1)RF feature_importances_ 기준 상위 10개 변수, (2)RF RFECV 기준 10개 변수, (3)LinearSVR RFE 기준 10개 변수, (4)Lasso Regression, (5)Ridge Regression coef 기준 상위 10개 변수로 두었다. 3가지 이상 변수 선택법에서 공통으로 선택된 변수들을 업종별로 선별하였고, 이 결과를 Table 5에 나타냈다.

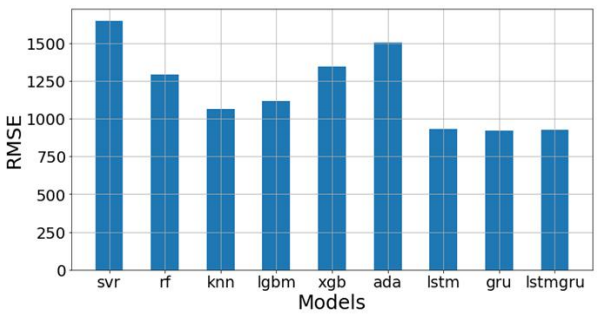

Fig. 4. Comparison of RMSE between algorithms using the close of 15 days or 1 day

Table 4. RMSE between algorithms using the close of 15 days or 1 day

	Model	Input	RMSE
NON TIME SERIES	SVR	close of 15 days	1646.7199
	RF	close of 1 day	1284.5359
	KNN		1066.9414
	LGBM		1118.5131
	XGB		1343.2103
	ADA BOOST		1522.4091
TIME SERIES	LSTM	close of 15 days	932.2000
	GRU		920.6104
	LSTM+GRU		926.7996

42개의 변수 중 모든 업종에서 9개의 변수가 채택되었다. 기본적으로 사용되는 종가, 저가, 시가, 고가 외에도 기술 보조지표인 이동평균, WMA, SMA, MOM과 재무제표지표인 PER, PBR, 외국인 보유수량이 주가 예측에 유의미함을 보인다. 종가, 저가, 고가, 시가, MA5는 모든 업종에서 영향력이 큰 방면, 외국인 보유수량은 IT, Finance 업종에서만 선택됐으며, SMA, PBR은 Material 업종에서만 선택됐으며, MOM 또한 Finance 업종에서만 선택됐다.

매개변수의 최적 조합을 찾는 과정 역시 feature selection 시 고려해야 하는 중요한 기법 중 하나이다[14]. 그러므로, 우선적으로 Table 6에는 각각의 알고리즘별로 모든 변수의 조합을 삼성전자 데이터에 입력하여 예측오차가 가장 낮은 조합을 나타내었으며, 이는 최적의 변수 조합을 찾는 과정에서의 결과를 요약한다.

또한, Fig. 5에서 GRU의 최적의 변수 조합을 입력하여 종가를 출력하는 GRU 네트워크의 구조를 시각적으로 보여주었으며, 이는 모델의 구성과 흐름을 이해하는 데 도움이 되도록 나타냈다. 15일 치 종가, 고가, MA5, MA20, 외국인 보유수량을 입력하여 다음 날 종가를 예측하는 구조이다.

Table 5. Result of feature selection

Industry	Selected feature
IT (Samsung Electronics Co., Ltd)	close, low, high, open, MA5, MA20, WMA, PER, foreign_holding
Material (LG Chem Ltd.)	close, low, high, open, MA5, WMA, SMA, PER, PBR
Finance (KB Financial)	close, low, high, open, MA5, WMA, PER, foreign_holding, MOM

Table 6. Optimal combination by model

Model	Optimal Combination	RMSE
RF	close, high, low, PER of 1 day	1130.7106
KNN	close, high of 1 day	1009.3184
LGBM	close of 1 day	1118.5131
LSTM	close, high, WMA, foreign_holding of 15 days	908.9118
GRU	close, high, PER, foreign_holding of 15 days	905.4588
LSTM+ GRU	close, high, MA5, MA20, foreign_holding of 15 days	909.3614

IV. Optimization and Results

본 연구에서 주가 예측을 위해 활용한 LSTM과 GRU와 같은 시계열 예측 딥러닝 알고리즘의 층(layer)과 노드(node) 수는 모델의 예측 성능에 큰 영향을 미치는 중요한 요소 중 하나이다[13]. 층의 수는 모델이 시계열 데이터의 복잡한 시간적 의존성을 파악하는 데 영향을 미친다. 층 수가 많으면 더 복잡한 패턴을 학습할 수 있으나, 적은 데이터로는 과적합(overfitting) 가능성이 커지며, 노드 수는 모델의 표현력과 학습 능력에 영향을 주며, 많은 노드 수는 층 수와 마찬가지로 복잡한 패턴 학습에 도움이 되지

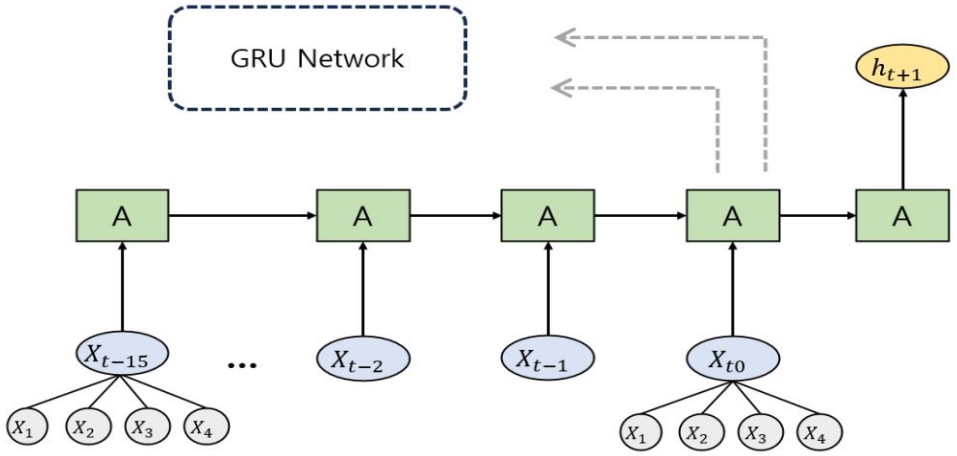

Fig. 5. GRU Network with optimal combination

만, 훈련 시간이 증가하고, 데이터가 적을 때는 과적합의 가능성이 존재한다. 따라서, LSTM과 GRU 모델의 성능을 최적화하기 위해서는 적절한 층과 노드 수를 찾는 것이 중요하다.

Table 7에는 다양한 층과 노드 구성으로 LSTM과 GRU 모델을 실험하고, 검증 데이터를 활용하여 최소 RMSE를 도출한 최적의 하이퍼파라미터를 기록했다.

이어서, 최종적으로 Table 8에는 모든 알고리즘의 최적 하이퍼파라미터와 주요 평가지표인 RMSE와 R2 결과를 제시했다. 결과적으로, RF와 LGBM과 같은 트리 계열 모델은 max_depth와 n_estimators와 같은 하이퍼파라미터의 변화가 모델 성능에 큰 영향을 미친다는 것을 확인했다. 반면에 GRU와 LSTM과 같은 딥러닝 시계열 모델에서는 노드의 수와 층의 깊이가 성능 향상에 영향을 미쳤으며, 최적화 함수, batch_size와 같은 다른 하이퍼파라미터들은 큰 영향을 미치지 않았다. Fig. 6은 삼성전자에 위의 실험 결과들을 적용 후 요약하여 알고리즘별 성능을 보여준다. RMSE_close는 알고리즘별 종가만을 이용하여 입력한 RMSE 결과, RMSE_feature_selection은 feature selection 이후 매개변수 최적 조합의 입력 RMSE 결과, RMSE_HPO는 노드, 층 최적화와 하이퍼파라미터 최적화 RMSE 결과를 의미한다.

비 시계열 예측 알고리즘과 비교하여, 시계열 예측 딥러닝 알고리즘인 LSTM, GRU, LSTM+GRU 모델이 Fig. 6에 표현한 모든 경우에서 더 낮은 예측 오차를 보였다. 이로써 본 연구는 시계열 예측 딥러닝 알고리즘들이 주가 예측에 높은 성능을 제공함을 보여준다. 또한 시계열 예측 알고리즘 중에서도 GRU가 사용된 알고리즘이 더 높은 예측 정확도를 보여준다.

Table 7. RMSE of hidden layer and node by model

Model	Node	Layer	RMSE
LSTM	16	2	910.9001
	32	2	914.8317
	32	3	913.2780
	64	2	913.1321
	64	3	916.9996
GRU	16	2	910.4560
	16	3	955.4798
	32	3	901.8101
	64	3	901.8013
LSTM+ GRU	32	2 / 2	911.9075
	64	2 / 2	904.9332
	64 / 32	2 / 2	909.5059

V. Conclusions

본 연구에서는 주가 예측에 있어서 비 시계열 예측 알고리즘과 시계열 예측 알고리즘의 성능을 비교하고, 다양한 주가 예측 요소의 중요도와 최적의 조합을 연구했다. 코로나로 인한 변동성 크고 불안정한 상황에도 불구하고, 시계열 예측 딥러닝 알고리즘이 더 낮은 예측 오차를 보였으며, 이는 시계열 관점의 데이터 분석이 높은 변동성과 불안정한 환경에서도 유의미한 결과를 도출함을 나타낸다. 연구에서 확인된 feature selection의 성능 향상은 변수 선택의 중요성을 강조하며, 이는 이후 연구에서의 의사결정에 도움을 줄 것으로 예상된다. 또한, 시계열 예측 딥러닝 알고리즘 노드 수와 레이어 깊이의 조정은 유의미한 성능 향상을 보였으며, 모델 튜닝에서 중요한 지표로 활용되었다. 향후 연구에서는 글로벌 시장의 동향과 국제적인 영향 요소들을 고려하여 전 세계 주가 데이터에 관한 예측 연구를 수행하여 예측 성능을 다각화하고 향상시킬 예정이다.

Table 8. Hyper parameters by model

Model	Optimized hyper parameter	RMSE	R2
RF	n_estimators:200, max_depth:5, min_samples_leaf:1, min_samples_split:2	**1060.8211**	**0.9761**
KNN	n_neighbors:10, weights:uniform, metric:manhattan, leaf_size:10	**1008.5801**	**0.9784**
LGBM	n_estimators:500, max_depth:20, num_leaves:20, learning_rate:0.01	**987.7137**	**0.9792**
LSTM	optimizer:nadam, batch_size:64, dropout:None	**911.9774**	**0.9825**
GRU	optimizer:adam, batch_size:32, dropout:None	**901.6247**	**0.9829**
LSTM+GRU	optimizer:adamax, batch_size:16, dropout:None	**907.0604**	**0.9827**

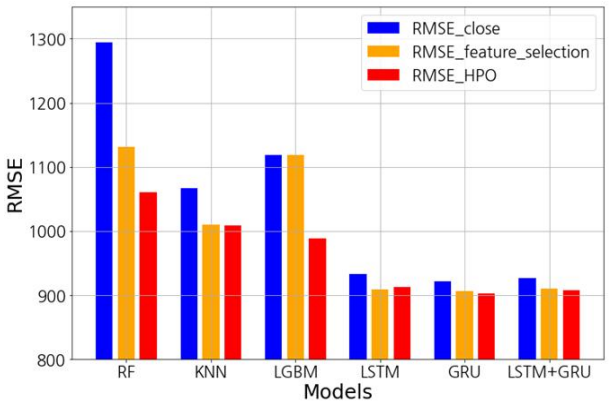

Fig. 6. RMSE of feature selection and hyperparaeter tuning

REFERENCES

[1] M. Nabipour et al., "Predicting Stock Market Trends Using Machine Learning and Deep Learning Algorithms Via Continuous and Binary Data; a Comparative Analysis," IEEE Access, vol. 8, pp. 150199-150212, Aug 2020. DOI: 10.1109/ACCESS.2020.3015966

[2] Y. Kara et al., "Predicting direction of stock price index movement using artificial neural networks and support vector machines: The sample of the Istanbul Stock Exchange," Expert Systems with Applications, vol. 38, pp. 5311-5319, May 2011. https://doi.org/10.1016/j.eswa.2010.10.027

[3] L. Pei Fun et al, "Price Earnings Ratio and Stock Return Analysis (Evidence from Liquidity 45 Stocks Listed in Indonesia Stock Exchange)," Jurnal Manajemen dan Kewirausahaan, pp. 1-10, May 2012. DOI: 10.9744/jmk.14.1.7-12

[4] K. Hyung gyu, "An Empirical Study on the Relationship between Price-Earnings Ratios and Stock Values," Korean Industrial Economic Association, pp. 725-743, Apr 2019. DOI: 10.22558/jieb.2019.04.32.2.725

[5] K. XIA et al., "LSTM-CNN Architecture for Human Activity Recognition," IEEE Access, vol. 8, pp. 56855-56866, Mar 2020. DOI: 10.1109/ACCESS.2020.2982225

[6] M. Ali raza suleman et al., "Short-Term Weather Forecasting Using Spatial Feature Attention Based LSTM Model," IEEE Access, vol. 10, pp. 82456-82468, Aug 2022. DOI: 10.1109/ACCESS.2022.3196381

[7] K. Jaejin et al., "Demand Analysis and Forecasting of Battery Electric Vehicles in Korea," Journal of the Korean Society of Supply Chain Management, vol. 20, pp. 24~35, May 2020. https://doi.org/10.25052/KSCM.2020.05.20.1.24

[8] H. Choi et al., "Stock prediction analysis through artificial intelligence using big data," Journal of the Korea Institute of Information and Communication Engineering. Vol. 25, No. 10, pp. 1435-1440, Oct 2021. DOI: 10.6109/jkiice.2021.25.10.1435

[9] Y. Han Jeong., "A Study on the Characteristics of Recurrent Neural Networks Using the S&P 500 Index," Graduate School of Sogang University, Feb 2020. UCI: I804:11029-000000065193

[10] X. Hou et al., "ST-Trader: A Spatial-Temporal Deep Neural Network for Modeling Stock Market Movement," IEEE/CAA Journal of Automatica Sinica, Vol. 8, No. 5, pp. 1015-1024, May 2021. DOI:10.1109/JAS.2021.1003976

[11] Pystock, KRX stock information scraping, https://github.com/sharebook-kr/pykrx

[12] TA-Lib, ta-lib-python, https://github.com/TA-Lib/ta-lib-python

[13] N. Gorgolis et al., "Hyperparameter Optimization of LSTM Network Models through Genetic Algorithm," 2019 10th International Conference on Information, Intelligence, Systems and Applications (IISA). Jul 2019. DOI: 10.1109/IISA.2019.8900675

[14] Y. Jiang et al., "Combination Features and Models for Human Detection," 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). , Jun 2015. DOI: 10.1109/CVPR.2015.7298620

Authors

Min-Seob Song received the A.S degress in Department of Computer Science from Inha Technical College in 2022 and is currently a B.S. in Department of Computer Science at Inha Technical College, Incheon, Korea. He is interested in artificial intelligence, deep learning, data analysis.

Junghye Min received the B.S. degree in mathematics from Ewha Women's University, Seoul, Korea, in 1995, and the M.E. and Ph.D. degrees in computer science and engineering from Pennsylvania State University, U.S.A. in 2003 and 2005, respectively. From 2005 to 2021 she was a Principal Enginner with Samsung Research, Samsung Electronics, Seoul, Korea. Dr. Min joined the faculty of the Department of Computer Science at Inha Technical College, Inchoen, Korea in 2022 and is currently an assistant Professor. Her research interests include image enhancement, image style transfer, and pattern recognition.