Journal of Convergence for Information Technology
Vol. 12. No. 3, pp. 1-10, 2022
e-ISSN 2586-4440
DOI : https://doi.org/10.22156/CS4SMB.2022.12.03.001

앙상블 Voting 기법을 활용한 배추 가격 예측에 관한 연구

이창민[1], 송성광[1], 정성욱[2*]
[1]창원대학교 컴퓨터공학과 학생, [2]창원대학교 컴퓨터공학과 부교수

A Study on the Prediction of Cabbage Price Using Ensemble Voting Techniques

Chang-Min Lee[1], Sung-Kwang Song[1], Sung-Wook Chung[2*]
[1]Student, Department of Computer Engineering, Changwon National University
[2]Associate Professor, Department of Computer Engineering, Changwon National University

요 약 배추와 같은 채소류는 자연재해의 영향을 많이 받기 때문에 폭우나 병해와 같은 재해로 인해 가격 변동이 심해져 농가 경제에 영향을 미치게 된다. 이러한 문제를 해결하기 위해서 농산물 가격 예측을 위한 다양한 노력이 행해졌지만 극심한 가격 예측 변동을 예측하기는 어렵다. 본 연구에서는 단일 분류기를 결합하여 다양한 여러 개의 분류기를 통해 최종 예측 결과를 결정하는 방식인 앙상블 Voting 기법으로 배추 가격을 분석하였다. 또한 시계열 분석 방법인 LSTM과 부스팅 기법인 XGBoost와 RandomForest로 결과 비교를 하였다. 가격 데이터는 일별 데이터를 사용하였고 배추 가격에 영향을 주는 기상정보와 물가지수 등을 사용하였다. 연구 결과로는 실제값과 예측값의 차이를 보여주는 RMSE 값이 약 236 수준이다. 이 연구를 활용하여 농산물 가격 예측과 같은 다른 시계열 분석 연구 모델 선정에 활용할 수 있을 것으로 기대된다.

주제어 : 농산물, 가격예측, 딥러닝, 머신러닝, 앙상블

Abstract Vegetables such as cabbage are greatly affected by natural disasters, so price fluctuations increase due to disasters such as heavy rain and disease, which affects the farm economy. Various efforts have been made to predict the price of agricultural products to solve this problem, but it is difficult to predict extreme price prediction fluctuations. In this study, cabbage prices were analyzed using the ensemble Voting technique, a method of determining the final prediction results through various classifiers by combining a single classifier. In addition, the results were compared with LSTM, a time series analysis method, and XGBoost and RandomForest, a boosting technique. Daily data was used for price data, and weather information and price index that affect cabbage prices were used. As a result of the study, the RMSE value showing the difference between the actual value and the predicted value is about 236. It is expected that this study can be used to select other time series analysis research models such as predicting agricultural product prices

Key Words : Agricultural Product, Price Prediction, Deep Learning, Machine learning, Ensemble Voting

1. 서론

배추는 양귀비목 십자과화에 속하는 식물로 김치의 주재료이며 무, 고추, 파 등과 함께 우리의 식단에서 빠질 수 없는 중요한 식재료 중 하나이다. 또한 배추는 카로틴 및 다양한 비타민 및 미네랄 성분들이 함유되어 있어 우리 몸을 건강하게 하는 탁월한 효능들이 많이 있

*This research is financially supported by Changwon National University in 2021~2022
*Corresponding Author : Sung-Wook Chung (swchung@changwon.ac.kr)
Received February 10, 2022 Revised March 3, 2022
Accepted March 20, 2022 Published March 28, 2022

다. 김치와 배추는 건강식품으로 주목받는데 김치가 현대인들의 고민거리인 암과 노화, 비만 등을 예방하기 때문이다[1]. 2008년 미국 건강 전문지 'Heath'에서는 세계 5대 건강식품 중 하나로 김치를 선정하기도 했다[2].

배추는 수확 시기별 종류가 다르다. 첫 번째, 봄배추는 매년 4월에서 6월에 출하되고 배추에 함유된 수분량은 많지만, 저장성은 가을배추와 겨울 배추보다는 짧다. 두 번째, 여름 배추는 매년 7월에서 10월 사이에 출하되며 가을에 출하되는 배추보다 크기가 작고 무게는 다른 계절 배추보다 가볍다는 특징이 있다. 세 번째, 가을배추는 매년 11월에서 다음 해 1월까지 출하되며 저온에서 배추 결구력이 매우 강하며 봄, 여름에 출하되는 배추에 비해 저장성이 매우 뛰어나다. 마지막으로 겨울 배추는 매년 1월에서 3월 사이에 출하되며 배추의 결구력이 매우 뛰어나고 가을배추와 마찬가지로 저장성이 매우 뛰어나다[3]. 건강하고 싱싱한 배추는 겉잎의 색이 짙고 푸른색을 띠는 것이 좋고 속잎이 노란색을 띠고 있는 것이 좋다고 알려져 있다.

위와 같은 계절별 배추의 특징들로 출하하는 시기에 따라 다른 품종의 배추를 심는다. 따라서 배추는 계절마다 가격이 다르다. 이러한 변동하는 배추의 가격 안정을 위해 농림축산식품부는 도매시장과 산지 조사 및 전문가 분석에 기반하여 배추 수급량을 예측하고 수급 안정 대책을 추진하고 여러 전문가가 농산물 가격 예측을 위해 연구하고 있다.

하지만 위와 같은 노력에도 배추의 가격 변동성에 대한 정확한 예측은 매우 어렵다. 해당 연구에서는 2장에서 설명하는 선행연구에서 사용한 모델들과 달리 앙상블 Voting 모델을 사용하여 비교함으로써 Voting 모델도 시계열 분석 모델들과 같이 좋은 성능을 보여주는 것을 증명하여 다른 가격 예측 연구에서 사용될 모델 결정에 도움을 주고자 한다.

일반적으로 농수산물의 가격 예측 방법은 시간을 통해 차례대로 발생한 시계열 데이터 분석에 좋은 성능을 보이고 RNN (Recurrent Neural Networks)의 관련 정보와 그 정보를 사용하는 지점 사이 거리가 멀어지는 경우 학습 능력이 현저하게 저하되는 장기 의존성 기간의 문제를 해결한 모델인 LSTM (Long ShortTerm Memory)이 있다. 배추의 데이터는 시계열이고 과거의 장기간 데이터를 이용하여 예측해야 하므로 시계열 분석 방법으로 적용할 수 있다. 그런데 LSTM 알고리즘은 은닉층으로만 흐르는 데이터로 인해 학습데이터를 과하게 잘 학습하는 과적합이 일어날 수 있다는 약점이 있다[4]. 이러한 약점은 알고리즘의 오차를 증가시키는 원인으로 작용한다. 그래서 우리의 본 연구에서는 과적합을 방지하기 위해 정규화를 수행한 데이터를 기반으로 LinearRegreession, Ridge Regressor, LassoRegressor 모델들을 하이퍼 파라미터를 차례대로 입력하여 학습하고 측정하면서 가장 좋은 파라미터를 찾아주는 GridSearchCV를 적용한 후 결합하여 앙상블 기법인 Voting 모델에서 예측하려고 한다.

또한 시계열 분석에 높은 성능을 보이는 LSTM 딥러닝 알고리즘과 시계열 분석에 능한 앙상블 boosting 기법의 종류인 RandomForest와 XGBoost 모델과 비교하여 단일 분류기를 결합하여 다양한 여러 개의 분류기를 통해 최종 예측 결과를 결정하는 방식인 앙상블의 Voting 모델을 이용해서, 기상 상황과 자연재해, 병해충을 포함한 변수들로 배추의 가격이 어떻게 변화하는지 예측하는 모델을 만들고, 해당 모델의 정확도를 그래프와 결괏값으로 나타낸다.

우리는 데이터를 기상청 (www.weather.go.kr), 농넷 (www.nongnet.or.kr), 통계청 (kostat.go.kr)에서 배추와 관련된 정보를 수집하였고 이를 토대로 Voting 기법을 적용하였다. 실험 수행 후 딥러닝 평가 지표인 RMSE 측정 결과 약 236 수준이다.

2. 선행연구조사

농산물 가격을 예측하기 위해 가격에 영향을 미치는 다양한 데이터를 분석하는 연구들이 이루어져 왔다. 현재까지 연구되었던 모델들은 시간의 흐름에 따라 관찰된 데이터를 분석해 미래의 값을 예측하고 경향, 주기, 계절성을 파악하는 분석 방법인 시계열 분석으로 대표적인 ARIMA (Autoregressive Integrated Moving Average) 모델이 있고, 신경 회로망을 다층적으로 구성하여 컴퓨터가 다양한 데이터를 통해 마치 사람처럼 생각하고 배울 수 있도록 하는 기술인 딥러닝을 활용하여 농산물 가격에 영향을 미치는 기후, 물가, 과거 가격 등 다양한 변수를 학습시켜 가격을 예측하는 방법의 대표적인 예로는 ANN(Artificial Neural Network), RNN, LSTM을 적용한 연구들이 있다. 또한 ARIMA 모형과 서포트 벡터 머신(Support Vector Machine, SVM)을 결합한 비선형 시계열 분석을 활용한 가격 예측 연구도 있다.

김 위판가격 분석에서는 다양한 시계열 모형 중시차 변수를 이용한 다중회귀모형, ARIMA 모형, VECM 모형을 적용되었다. 사용된 변수는 월별 김 실질 위판가격, 전월의 김 실질 위판가격, 김 수출량, 1인당 월별 양곡 소비량, 꽁치 생산량, 완도지역 수온이 있다[5].

딥러닝을 활용한 농산물 가격 예측 모델로 LSTM 모형을 설정한 연구에서 사용된 변수는 기상변수로는 기온, 강수량, 습도, 풍량, 적설량, 각 변수에 대해 최젓값, 최곳값, 평균값, 중간값 사용되었고 기타 변수로는 경유, 물가 상승률, 전년 수확량, 전년 수입량, 전년 재배면적을 사용하였다[6].

LSTM 모델의 단점인 설명력이 부족하고, 예측 결과를 해석하는 데 어려움을 극복하기 위해 Attention Mechanism을 적용하여 채소 가격에 영향을 미치는 요인과 특정 시간을 찾기 위해 Dual Attention을 사용한 연구에서 사용된 변수는 측정 일, 측정 일의 달, 채소 가격, 총반입물량, 편차, 일 · 월 기상, 주의보, 경보, 보조지표 상관관계 높은 음식의 가격 등이 사용되었다[7].

해당 연구들에서는 가격 데이터와 기상 정보가 공통적인 변수로 사용되었고, 이외에 다양한 변수들은 Table 1에 요약하였다.

Table 1. Variables related to price prediction of agricultural products

Researcher	Agricultural products	Variables
J. O. Nam	seaweed	1. The price of seaweed 2. Seaweed export volume 3. Grain consumption 4. Production of saury 5. Water temperature in Wando
S. H. Shin	green onion, onion, rice, zucchini, spinach.	1. Weather 2. Diesel 3. Inflation rate 4. Harvest 5. Plantation area
H. J. Lim	cabbage, garlic, onion, radish, chili pepper	1. The price of vegetables 2. Weather 3. Price of highly correlated food 4. Supplementary indicators

3. 학습데이터 구축

해당 부분에서는 학습 데이터 구축과정과 데이터 전처리 과정을 설명한다.

3.1 데이터 수집

해당 부분에서는 배추 가격 예측 모델에 사용한 데이터에 관해 설명한다. 배추의 변동하는 가격에 대한 예측을 위해 배추의 가격 데이터와 연도별 생산량 데이터, 배추의 주산지인 해남과 태백의 기상 데이터와 통계청에서 제공하는 소비자 물가 지수와 배추의 물가지수 데이터를 사용하였으며 농넷에서 제공하는 배추의 일별 가격 데이터에서 일주일 전 가격 데이터를 추출하여 입력값 데이터로 이용하였다. 마지막으로 병해충, 폭염, 호우, 태풍들의 주의보와 경보 데이터를 사용하였다. Table 2은 수집한 데이터의 출처를 나타내고 있다. 배추의 데이터 수집 기간과 수집한 데이터의 수는 Table 3과 같다.

Table 2. Data Collection Information

Source	Data
Nongnet (www.nongnet.or.kr)	Cabbage price data
Nongnet (www.nongnet.or.kr)	Cabbage Production data
The Meteorological Administration (www.weather.go.kr)	Haenam, Taebaek Weather Data
National Statistical Office (www.kostat.go.kr)	Consumer index, price index
National Crop Disease Pest Management System (https://ncpms.rda.go.kr/npms/Main.np)	Watch out for pests

Table 3. Data collection period

Section	Period	The number of data
Cabbage price data	2014.01.01. ~ 2021.11.30	2434
Cabbage Production data	2014 ~ 2021	7
Haenam weather	2014.01.01. ~ 2021.11.30	2434
Taebaek weather	2014.01.01. ~ 2021.11.30	2434
Consumer index	2014.01 ~ 2021.11	120
Price index	2014.01 ~ 2021.11	120
Watch out for pests	2014.01 ~ 2021.11	120

3.2 데이터 셋

해당 부분에서는 배추 가격 예측에 사용한 변수들의 선정 이유와 수집 경로를 포함하여 사용된 데이터들의 전처리 과정을 설명한다. 마지막으로 입력 데이터들이 미치는 영향을 알아보기 위해 하나 이상의 독립 변수의 변경이 종속 변수에 영향을 주면 선형관계가 발생하게 되는데 이를 수치화하여 Fig. 1과 같이 나타내었다.

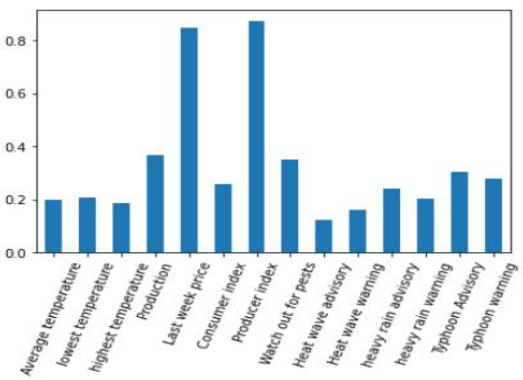


Fig. 1. Voting model

3.2.1 가격 데이터

배추의 과거 가격 데이터는 배추의 가격을 예측하기 위해서 필수적으로 필요한 요소이며 농넷에서 제공하는 도매시장 거래정보 시스템을 이용하였다. 가격 데이터는 일, 순, 월, 분기, 년 총 5가지로 제공하는데 해당 연구에서는 배추의 가격을 일별로 예측하기 때문에 일별 데이터를 수집하였다. 또한 배추의 7일 전 가격은 다음 주 가격에 영향을 미치므로 배추의 가격 데이터에서 7일 전 데이터를 추출하여 입력값으로 사용하였으며 배추의 가격 형성에 영향을 많이 끼치는 생산량도 입력값으로 사용하였다. 해당 부분에서 설명한 데이터의 값은 Table 4와 같다.

Table 4. Price data

date	price	Last week price	Production
2014-01-03	471.24	389.30	2736447
2014-01-04	425.44	373.38	2736447
...			
2021-11-29	797.61	837.841	2156841
2021-11-30	726.36	850.78	2156841

3.2.2 기상 데이터

배추는 계절마다 주로 생산하는 지역이 달라 배추의 계절성을 알아내기 위해 배추 시장의 80%를 차지하는 전라남도 해남과 강원도 태백의 기상 데이터를 이용하였다. 주로 고랭지 배추인 여름 배추는 해발 600m 이상이 되면 한여름에도 배추가 버틸 수 있을 정도로 서늘한 기후를 가진 강원도 태백에서 주로 고랭지 배추가 재배된다. 그래서 고랭지 배추가 출하되는 시기인 7~10월은 강원도 태백의 기상 데이터를 사용하였으며 고랭지 배추를 제외한 봄배추, 가을배추, 월동 배추들이 출하되는 시기인 1~6월, 11~12월은 전라남도 해남지역의 기상 데이터를 사용하였다[3]. 해당 부분에서 설명한 데이터의 값은 Table 5와 같다.

Table 5. Weather data

date	Average temperature	Lowest temperature	Highest temperature
2014-01-03	2.3	-1.1	7.3
2014-01-04	-0.1	-3.7	5.6
...			
2021-11-29	7.6	2.5	13.7
2021-11-30	5.7	0.7	8.9

3.2.3 물가 데이터

배추의 가격 형성에 영향을 미치는 물가지수 데이터와 배추의 물가지수 데이터를 사용하였으며 국가 통계 포털(www.kosis.kr)에서 제공하는 데이터를 사용하였다. 해당 부분에서 설명한 데이터의 값은 Table 6과 같다.

Table 6. Price index data

date	Consumer index	Producer index
2014-01-03	93.73	81.96
2014-01-04	93.73	81.96
...		
2021-11-29	103.87	196.97
2021-11-30	103.87	196.97

3.2.4 자연재해 데이터

배추와 같은 농산물의 가격은 자연재해에 따라 가격 변동이 심하므로 배추의 가격 형성에 영향을 많이 끼치는 병해충, 폭우, 폭염, 호우 주의보와 경보 데이터를 사용하였다. 자연재해 데이터는 월마다 해당 재해가 발생한 횟수를 입력값으로 사용하였다. 해당 부분에서 설명한 데이터의 값은 Table 7과 같다.

3.3 데이터 전처리

해당 부분에서는 전체 데이터에 대한 전처리 방법을 설명한다. 전처리는 Min-Max 정규화를 사용하였다. Min-Max 정규화는 데이터의 분포가 다르면 데이터 해석이 균일하지 않아서 오차 발생률이 높아지기 때문에 Min-Max 정규화로 모든 데이터가 같은 정도의 중요도를 가지도록 해준다.

Table 7. Natural disaster data

date	pest	Heat wave advisory	Heat wave warning	Heavy rain advisory	Heavy rain warning	Typhoon advisory	Typhoon warning
2014-01-03	0	0	0	0	0	0	0
2014-01-04	0	0	0	0	0	0	0
...							
2021-11-29	2	0	0	0	0	0	0
2021-11-30	2	0	0	0	0	0	0

3.3.1 Min-Max 정규화

배추의 가격 데이터와 전체 입력 데이터에 대하여 Min-Max Scaling을 진행하였다. Min-Max Normalization 이라고도 불리며 특성들을 특정 범위로 스케일링을 수행한다. 가장 작은 값은 0, 가장 큰 값은 1로 변환되므로 모든 입력 데이터와 가격 데이터의 값은 0에서 1까지의 범위를 가지게 된다[8].

또한 모든 데이터의 평균을 0, 분산을 1로 만드는 표준화 방법인 StandardScaler보다 회귀에 유용하다는 장점이 있다. Min-Max 정규화의 식은 수식 1과 같다.

$$f(x) = \frac{x - x_{\min}}{x_{\max} - x_{\min}} \quad (1)$$

정규화를 수행하기 전 데이터의 예시로 우리가 사용하는 평균 기온 데이터의 분포는 Table 8과 같다. 평균 기온을 Min-Max 정규화를 수행한 후의 평균 기온 데이터 분포로는 Table 9에서 보이는 것처럼 데이터의 분포가 최소 0에서 최대 1까지의 값을 가지게 된다.

Table 8. Temperature data before normalization

Section	Average Temperature
mean	13.572769
min	-14.9
max	33.7

Table 9. Temperature data after normalization

Section	Average Temperature
mean	0.585859
min	0
max	1

4. 모델 설계

4.1 사용 모델

해당 부분에서는 Voting 모델에서 사용한 모델들의 특징을 설명한다.

4.1.1 LinearRegression

선형회귀는 종속변수 y와 한 개 이상의 독립 변수 x와의 선형 상관 관계를 모델링하는 회귀분석 기법이다. 선형회귀는 선형 예측 함수를 사용해 회귀식을 모델링하며, 알려지지 않은 파라미터는 데이터로부터 추정한다. 이렇게 만들어진 회귀식을 선형 모델이라고 한다[9]. 따라서 가격 예측을 할 때 선형 회귀를 사용해 데이터에 적합한 예측 모형을 개발한다. 개발한 선형 회귀식을 사용해 입력 데이터 x값으로 가격 데이터 y값을 예측할 수 있다. 이러한 선형식을 추정하는 방법으로는 관측값과 예측값 사이의 오차의 제곱 합이 최소가 되는 해를 구하는 방법인 최소 제곱법이 사용된다[10].

4.1.2 LassoRegression

LinearRegression이 적절한 가중치와 편향을 찾아내는 것이 관건이었다면 LassoRegression은 MSE가 최소가 되게 하는 가중치와 편향을 찾는데 동시에 가중치들의 절댓값들의 합이 최소가 되게 한다. 절댓값들의 합은 L1-Norm을 사용한다. 이 방법은 특성값의 계수가 매우 낮다면 0으로 수렴하게 하여 특성을 지워준다. 특성이 모델에 미치는 영향을 0으로 만들어 bias를 증가시켜 과적합을 방지한다. L1-Norm의 수식은 수식 2와 같다.

$$J(\theta) = MSE(\theta) + \alpha \sum_{i=1}^{n} | \theta_i | \quad (2)$$

4.1.3 RidgeRegression

RidgeRegression은 LassoRegeression이 가중치들의 절댓값들을 0으로 만드는 것이었다면 RidgeRegression은 가중치들의 합을 최소로 만드는 것이다. 가중치들의 합을 최소로 만드는 식은 L2-Norm이다. 이 방법은 영향을 거의 미치지 않는 특성에 대하여 0에 가까운 가중치를 주게 된다. L2-Norm의 수식은 수식 3과 같다.

$$J(\theta) = MSB(\theta) + \alpha \frac{1}{2} \sum_{i=1}^{n} \theta_i^2 \quad (3)$$

4.2 Voting 모델

가격 예측을 하는 모델의 알고리즘이 성능이 뛰어나다고 해도 모든 문제에 최적화되는 것은 아니다. 따라서 데이터의 형태나 중요시기에 따라 알고리즘이 선택해야 한다. 그래서 개별 알고리즘이 적절히 혼합된다면 단일 모델보다 성능이 향상될 가능성이 커지게 된다. 그래서 여러 개의 분류기를 생성하고, 그 예측을 결합하여 보다 정확한 예측을 도출하는 기법인 앙상블의 한 종류로 여러 개의 분류기가 투표를 통해 최종 예측 결과를 결정하는 방식인 Voting 기법을 배추 가격 예측을 위한 모델로 선정하였다. 우리가 제시하는 모델의 도안은 Fig. 2와 같다.

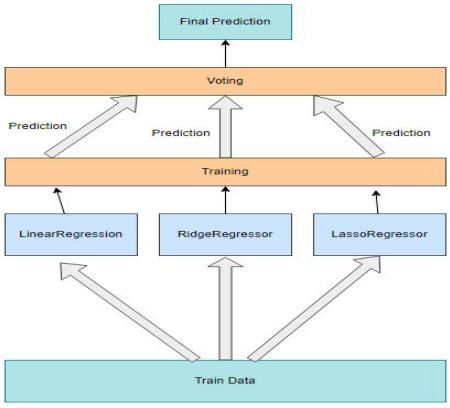


Fig. 2. Voting model

5. 실험

총 데이터는 2014년 1월 1일부터 2021년 12월 31일 까지의 데이터로 총 테이블 인스턴스는 2434개이다. 학습 데이터는 전체 인스턴스의 80%가 사용되었고 테스트를 위해서는 20%의 인스턴스가 사용되었다. Table 10은 학습 및 테스트 데이터 수와 합계를 나타낸다. 예측 모델마다 주말과 공휴일을 제외한 2014년 1월 1일부터 2020년 4월 26일에 해당하는 기간의 3.2장에서 설명한 입력값 데이터로 모델을 학습시켜서 학습된 모델로 2020년 4월 27일부터 2021년 12월 31일까지의 배추 가격을 예측하였다.

5.1 실험 환경

해당 연구에서는 구글에서 제공하는 Colab을 사용하였다. Colab은 구글에서 교육과 과학 연구를 목적으로 개발한 도구로, 구글에서 제공하는 클라우드의 가상 서버에서 작동한다. 또한 Colab은 GPU를 제공하기 때문에 데이터 분석 작업에 용이하다.

Table 10. Number of training and test data

Section	The number of data
Train data	1947
Test data	487
Total	2434

좋은 예측 모델들을 얻으려면 alpha와 같은 하이퍼 파리미터 튜닝이 필요하다. Ridge 모델과 Lasso 모델의 하이퍼 파리미터 튜닝 결과 Lasso의 최적의 alpha 값은 0.5 Ridge의 최적의 alpha 값은 1로 하이퍼 파라미터 튜닝 후 Voting 모델로 조합하여 결괏값을 예측하였다. Voting 모델과 LSTM, XGBoost, RandomForest 모델의 성능을 비교하였다.

LSTM 모델은 순환 신경망 모델인 RNN의 문제점인 적절한 정보와 그 정보가 필요한 곳과의 gap이 클 경우 예측을 잘하지 못한다는 장기 의존성 문제를 해결한 모델이다. cell state는 장기 정보를 전달하는 역할을 하고 forget gate layer는 어떤 데이터를 버릴지 아니면 유지할지를 결정한다. sigmoid를 통해서 나온 값으로 0은 제거 1은 유지를 뜻한다. 또 input gate layer는 어떤 값을 갱신할지 결정하고 tanh layer는 cell state에 더해질 수 있는 새로운 후보 값들의 벡터를 생성한다. LSTM의 모델 구조는 Fig. 3과 같다.

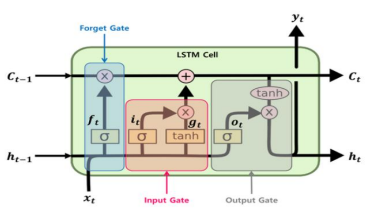


Fig. 3. LSTM model structure

RandomForest모델은 앙상블 머신러닝 모델의 한 종류로 훈련 과정에서 구성한 다수의 결정 트리로부터 평균 예측치 (회귀 분석)를 출력함으로써 동작한다.

XGBoost모델은 Extreme Gradient Boosting의 약자로 Gradient Boosting Algorithm의 과적합과 학습 속도의 문제를 보완한 모델이며 병렬 처리 연산을 수행하고 모델의 최적화를 위해 위치 경사 하강법을 사용한다.

6. 결과

모델의 예측 정확도는 실제값과 예측값의 차이로 모델의 성능을 가늠할 수 있다. 많이 사용하는 평가도구로는 Mean Squared Error (MSE)와 Root Mean Squared Error (RMSE)를 사용한다. MSE는 에러를 제곱하여 평균을 계산한다. 수식은 수식 4와 같다.

$$MSE = \frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y_i})^2 \quad (4)$$

RMSE는 MSE에 Root를 씌운 것으로 MSE 수식을 거의 따르지만 제곱된 에러를 다시 루트로 풀어주기 때문에 에러를 제곱하여 생기는 왜곡된 값이 덜하다는 장점이 있다. RMSE의 수식은 수식 5와 같다. 이러한 장점으로 해당 연구에서는 평가지표로 RMSE를 사용하였다.

$$RMSE = \sqrt{\sum_{i=1}^{n}\frac{(\hat{y_i} - y_i)^2}{n}} \quad (5)$$

6.1 전체 모델 비교

5장에서 말한 학습데이터와 위와 같은 평가 지표로 측정한 모델별 RMSE 값은 LSTM이 299, RandomForest가 286, XGBoost가 260, Voting 모델이 236으로 Voting 모델이 RMSE 값이 가장 낮아 가장 좋은 예측률을 보인다. Fig. 4는 각 모델의 예측 정확도를 시각화 한 것이다.

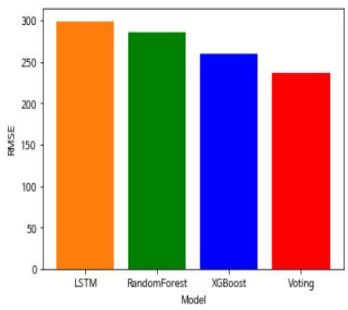


Fig. 4. Comparison of all models

6.2 LSTM

LSTM 모델의 가장 좋은 성능 측정을 위해 파라미터들의 값 설정을 하였다.

모든 학습 데이터 세트를 학습하는 횟수를 결정하는 파라미터인 epoch의 값은 모델을 학습할 때 가장 좋은 성능이 나올 때 중지해야 한다. 해당 연구에서는 실험을 통해 epoch의 값이 200~300회 사이에서 가장 좋은 성능을 보였으며 400회 이상에서는 성능 변화가 더 이상 높아지거나 낮아지지 않았다.

연산 한 번에 들어가는 데이터의 크기를 가리키는 batch size가 너무 큰 경우 한 번에 처리해야 할 데이터의 양이 많아지므로, 학습 속도가 느려지고, 메모리 부족 문제가 발생할 위험이 있다. 하지만 너무 작은 경우 적은 데이터를 대상으로 가중치를 업데이트하므로 자주 발생하게 되어 훈련이 불안정해져 적절한 batch size 설정이 중요하다. 해당 연구에서는 batch size를 20으로 설정하였다.

손실 함수 (loss function)을 최소화하여 학습하는 방법은 어떤 optimizer를 사용하느냐에 따라 달라진다. 해당 연구에서는 가장 일반적으로 많이 사용하며 RMSprop과 momentum의 조합인 adam optimizer를 사용하였다.

6.3 RandomForest

RandomForest 모델의 주요 파라미터는 n_estimators와 max_features가 있다. 가장 좋은 성능 측정을 위해 GridSearchCV를 이용하여 하이퍼 파라미터 설정을 하였다.

최대 선택할 특성의 수를 결정하는 max_features는 전체 특성의 수로 설정하면 모든 특성을 고려하게 되므로 decision tree에서 무작위성이 들어가지 않지만 복원추출의 무작위성은 존재하고 반대로 max_features의 값을 낮추게 되면 각 tree 들은 깊이가 깊어진다는 특성이 있다. 해당 연구에서는 max_features 값을 2로 설정하였다.

생성할 트리의 개수를 결정하는 n_estimators는 50으로 설정하였다.

6.4 XGBoost

XGBoost모델도 RandomForest모델과 같이 좋은 성능 측정을 위해 GridSearchCV를 이용하여 하이퍼 파라미터 선정을 하였다.

최대 깊이를 설정하는 max_depth는 7로 설정하였고 n_estimators는 100으로 설정하였고 데이터 샘플링의 비율을 지정하여 과적합을 제어하는 sub_sample 파라미터는 0.75로 설정하였다.

한 번 학습할 때 얼마만큼 학습해야 하는지 학습량을 의미하는 learning_rate는 너무 작으면 손실이 최적인 가중치를 찾는 데 오랜 시간이 걸리고 반대로 너무 크다면 최적점을 무질서하게 이탈할 가능성이 있다. 그래서 해당 연구에서는 learning_rate를 0.08로 설정하였다.

6.5 Voting

Voting 모델에 들어간 LinearRegression, RidgeRegressor, LassoRegressor 모델들의 파라미터 설정은 5장에서 설명한 것과 같다.

6.6 모델별 가격 변동성 비교

모델별 가격 변동성을 그래프로 나타낸다. x축의 Time(days)는 2020년 4월 27일부터 2021년 11월 30일까지의 총 500일에 해당하는 기간을 의미하고 y축의 Price는 x축의 해당하는 일별 배추의 가격을 의미한다.

Fig. 5는 실제 가격과 LSTM 모델로 예측한 가격의 차이를 시각화하여 변동성과 가격 차이를 나타내었다. 가격의 변동은 잘 따라가는 것으로 보이나 급격한 가격의 변화는 잘 따라가지 못하는 것으로 보인다.

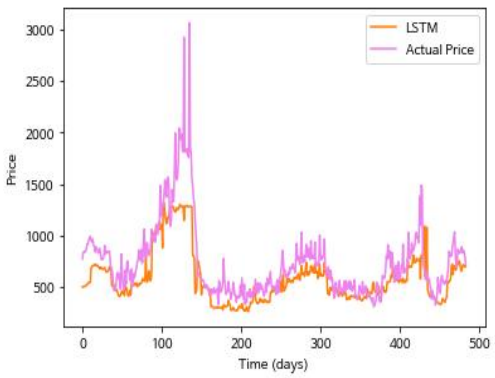


Fig. 5. LSTM model

Fig. 6은 실제 가격과 RandomForest 모델로 예측한 가격의 차이를 시각화하여 가격 변동성과 가격 차이를 나타내었다. 가격의 변동은 잘 따라가는 것으로 보이나 LSTM 모델과 마찬가지로 급격한 가격의 변화는 잘 따라가지 못하는 것으로 보인다.

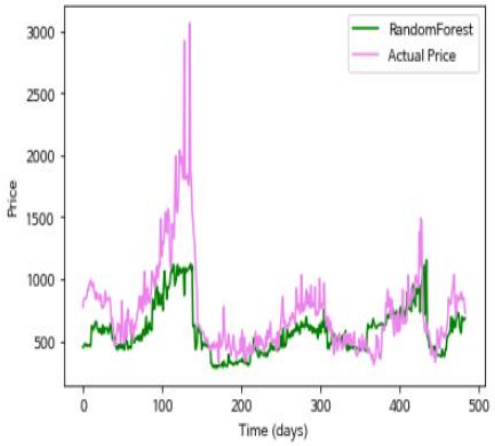


Fig. 6. RandomForest model

Fig. 7은 실제 가격과 XGBoost 모델로 예측한 가격의 차이를 시각화하여 변동성과 가격 차이를 나타내었다. 위에서 언급한 LSTM 모델과 RandomForest 모델과 같이 가격의 변동은 잘 따라간다. 하지만 급격한 가격 변동의 경우 LSTM 모델과 RandomForest 모델에 비해서는 잘 따라가는 것으로 보이나 눈에 띄게 잘 따라가지는 못하는 것으로 보인다.

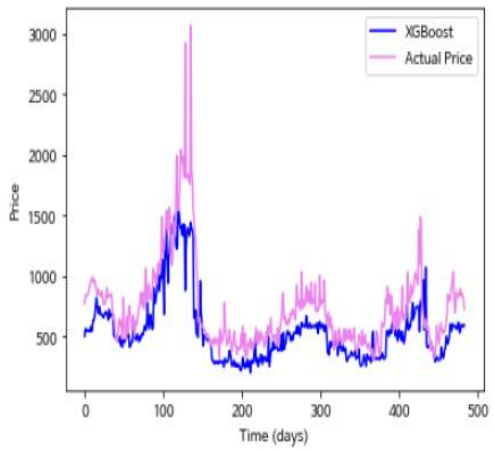


Fig. 7. XGBoost model

Fig. 8은 실제 가격과 Voting 모델로 예측한 가격의 차이를 시각화하여 변동성과 가격 차이를 나타내었다. 가격의 변동성은 잘 따라가며 위 3개의 모델보다 가격의 급격한 변동을 더 잘 따라가는 것으로 보인다.

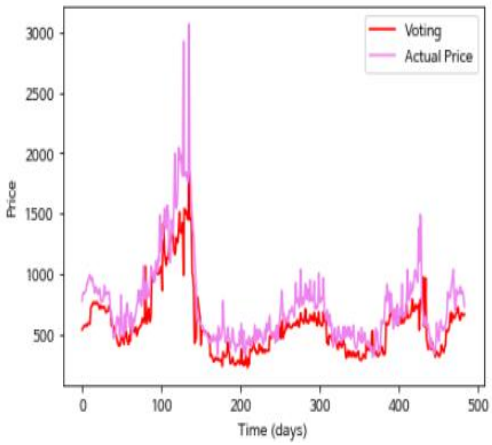


Fig. 8. Voting model

6.7 전체 모델 비교

전체 모델들의 명확한 성능 비교를 위하여 위에서 언급한 총 4개의 모델의 예측값과 실제값을 꺾은 선 그래프로 나타내었다. 그래프는 Fig. 9와 같다.

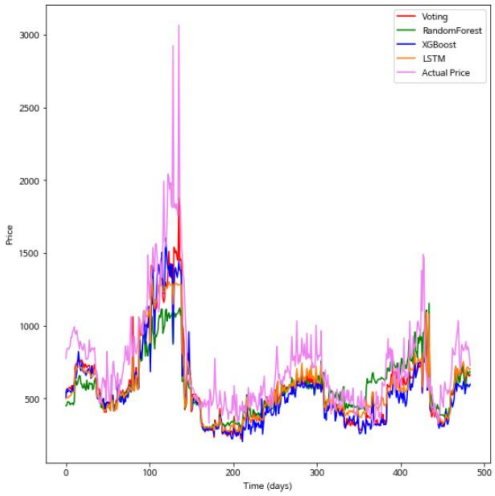


Fig. 9. Comparison graph of all models

7. 결론

본 연구의 목적은 배추에 영향을 미치는 기상요인과 호우주의보와 같은 자연재해나 물가지수와 같은 영향들로부터 가격 변동에 대한 효과적인 방법을 제시하기 위함이다.

배추 가격 예측에 사용된 데이터는 2014년부터 2021년까지 총 8년 치의 데이터가 사용되었다. Voting 모델에 가장 적합하고 배추 가격 예측 정확도가 높게 나온 LinearRegression, RidgeRegression, LassoRegression이 3가지 모델을 이용하여 Voting 모델에 적용하여 다른 LSTM, RandomForest, XGBoost 모델들과 비교한 결과 하나의 모델에 의존하여 예측할 수밖에 없는 3개의 모델과 달리 Voting 모델은 단일 모델 3개를 결합하여 가장 좋은 예측값을 선택하므로 모델에 의한 예측 가격과 실제 가격의 차이를 평가하는 지표인 RMSE가 236으로 실제값과 예측값의 차이가 가장 낮아 가장 높은 정확도를 보였다. 하지만 높은 정확도는 아니기에 Voting 모델의 성능을 개선하기 위해 선행 연구들에서 사용된 변수들을 고려하여 농사를 위해 운용되는 농기계에 필요한 유가의 데이터와 배추의 생산량에 영향을 미치는 재배면적 데이터와 같은 다른 입력값과 시계열 분석에 좋은 성능을 보이는 ARIMA, Prophet 같은 모델들과 비교하여 보팅 모델의 분류기를 교체하거나 개선하여 모델 성능을 높여야 할 점이 있다.

해당 연구에서 제안하는 여러 분류기를 사용한 앙상블 Voting 기법으로 다른 가격 예측 모델 연구에서 다양한 선택지 중의 하나로 사용될 수 있을 것으로 기대된다.

REFERENCES

[1] M. Y. Jin. (2021). *7 Benefits of Cabbage*. Health care NEWS (Online). www.hcnews.or.kr/news_gisa/gisa_view.htm?gisa_category=02010200&gisa_idx=10026

[2] E. G. Pyo. (2006). *Kimchi, selected as one of the world's top 5 health foods.*. SBS news (Online). https://news.sbs.co.kr/news/endPage.do?news_id=N1000090983&plink=COPYPASTE&cooper=SBSNEWSEND

[3] aTKAMIS. (n. d.). *Food Archives*. aTKAMIS (Online). www.kamis.or.kr/customer/archive/archive.do?action=detail&archiveNo=182

[4] D. M. Hawkins. (2004). The problem of overfitting. *J ournal of chemical information and computer sciences, 44(1),* 1-12.

DOI : 10.1021/ci0342472

[5] J. O. Nam. (2014). Forecast for Laver Producer Price Using Time Series Models. *KMI, 29(2),* 271-303.

[6] S. H. Shin, M. K. Lee & S. K. Song. (2018). A Prediction Model for Agricultural Products Price with LSTM Network. *The Korea Contents Association, 18(11),* 416-429.

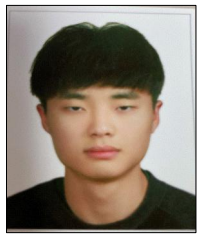
[7] H. J. Lim. (2020). *Dual Attention-based LSTM Model for produce price prediction.* Domestic Master's Thesis. Sejong University. Seoul.

[8] J. H. Park. (2020). *[Park Jung-hyun's Getting Started with Data Science] ⑥ Feature Engineering (2).* AiTimes (Online). www.aitimes.com/news/articleView.html?idxno=134913

[9] wikipedia. (2022). *linear regression.* wikipedia (Online). https://en.wikipedia.org/wiki/Linear_regression

[10] wikipedia. (2021). *Least square method.*wikipedia (Online). https://en.wikipedia.org/wiki/Least_squares

이 창 민(Chang-Min Lee) [학생회원]



· 2017년 3월 ~ 현재 : 창원대학교 컴퓨터공학과 학사과정
· 관심분야 : 인공지능, 딥러닝, 머신러닝
· E-Mail : dlckdals9467@naver.com

송 성 광 (Sung-Kwang Song) [학생회원]

· 2017년 3월 ~ 현재 : 창원대학교 컴퓨터공학과 학사과정
· 관심분야 : 인공지능, 딥러닝, 머신러닝
· E-Mail : sskg0708@naver.com

정 성 욱 (Sung-Wook Chung) [정회원]

· 2010년 8월 : CISE dept. Univ. of Florida, USA, (Ph.D)
· 2010년 10월 ~ 2012년 2월 : KT 종합기술원 중앙연구소 선임연구원
· 2012년 3월 ~ 현재 : 창원대학교 컴퓨터공학과 부교수
· 관심분야 : IoT, 스마트모빌리티, 머신러닝, 실시간 분산 멀티미디어시스템, 홈네트워크
· E-Mail : swchung@changwon.ac.kr