Predicting the price of cabbages through time series of price and external factors

15146312 Park, Gangmin
15146302 Kang, Euihyeon
15146304 Ku, Seunghyo

Abstract

This study introduces the process for predicting the cabbage price. For the purpose of forecasting, the data are based on cabbage and radish prices and time series data of weather, import volumes and import prices. Data collected over four years from 2015 to 2018 are averaged over a week to predict a week ago from now. Prediction models are evaluated using Gradient Boosting Regression, KNN Regression, Support Vector Regression, Linear Support Vector Regression, Linear Regression, Random Forest Regression, and Decision Tree Regression. For the best-predicted model, the R_square was approximately 64% and this result of study is expected to help farmers and consumers through pricing forecasts.

1. Introduction

The cabbage kimchi is a Korean traditional food. According to a 2013 National Health Nutrition Survey analysis, cabbage kimchi is still loved by Koreans as the third most popular food of all [1]. Figure 1 shows the monthly change in cabbage prices from 2015 to 2018. Cabbage, an essential ingredient for making cabbage kimchi, has maintained steady demand, but the price of cabbage can be seen to change irregularly.

The factors that change the price of cabbage would be kimchi-making before winter, weather factors such as heavy rain or heavy snow, the price of radish, the amount of cabbage imported and the import price.

A related study of predicting wholesale prices of onions used data such as past wholesale prices, production and cultivation areas of onion[2]. In the study [2], prices were estimated using the ARDI (autoregressive and distributed lags) model, but in this study, we wanted to select more accurate model among seven models including Linear Support Vector Regression.

Another study used data of past cabbage prices and the frequency of agricultural weather keywords associated with crop growth in documents containing cabbage as key data [3]. In this study, however, weather data such as temperature and precipitation considering the growth of crops were used, as well as amount of imported agricultural products and radish prices that could affect the demand and supply of cabbages.

2. Analytics

2.1 Data

Table 1 shows the independent variables for predicting the price of cabbage, and the data for past cabbage prices are agricultural trade data from 2015 to 2018 collected by KAMIS. The data include the average transaction price and the **Normal Year** price of cabbage. **Normal Year** means that average price expending maximum and minimum values in the data for the five-year period based on this year's data. In addition,

we also collected price data for cabbages and radishes due to their high importance in the agricultural market [4]. As shown in Figure 1, the agricultural trade data are time series data and therefore the price of cabbage can be predicted by this data alone, but additional weather data and imported agricultural product data have been collected to take into account other variables that do not include time series data. Table 2 is the weather data and was collected over the period from 2015 to 2018. Agricultural trade data and weather data were collected in Seoul, Daejeon, Daegu, Gwangju and Busan.

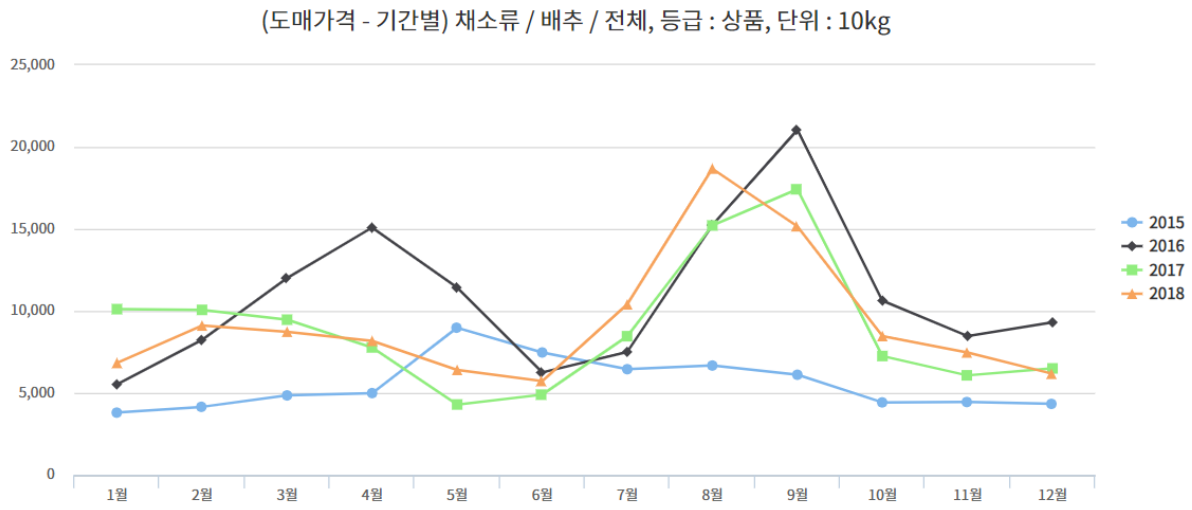


Figure 1. Monthly cabbage transaction price (10kg) graph from 2015 to 2018

Cabbage Average
Cabbage Normal year
Radish Average
Radish Normal year

Table 1. Agricultural Trade Data

Average Temperature	Duration of Bright Sunshine
Lowest Temperature	Solar Radiation Quantity
Highest Temperature	Maximum Depth of Snow Cover
Precipitation	Maximum Depth of New Snowfall
Wind Speed	Amount of Cumulus
Dew Point	Average Temperature of Ground
Relative Humidity	Total Low Evaporation
Vapor Pressure	Total Large Evaporation
Spot Atmospheric Pressure	Amount of Middle and Lower Cloud
Sea-level Pressure	

Table 2. Weather Data

Cabbage import weight
Cabbage amount of import

Table 3. Imported agricultural product data

The data set used in this study contains missing values. The processing of missing values is defined according to the characteristics of the variables, and each is handled differently. The first missing value

type is appeared when there are no events occurred. For example, in the case of the **Maximum Depth of snow cover**, the amount of snow is measured, so there is no value in the day when it is not snowing. In this case, fill the missing value with zero. The second type is appeared when there are not recorded. For example, in the case of an **amount of cumulus** and an **amount of middle and lower cloud** in 2017, there is no measurement of approximately 160 days. Therefore, the average of the other year's measurements is obtained and filled. The last type is the case where it cannot be determined. If the variable's values are missing without a particular pattern, it is impossible to determine. Weather data are time series data that maintains the trend of previous data, so in this case the missing values are filled by referring to the data from the previous one.

Agricultural trade data and weather data exist five data based on each region. Since this study is aimed at predicting universal cabbage prices, an average of five regional data is obtained and used. It will then generate data sets from 2015 to 2018 by integrating agricultural trade data, imported agricultural product data and weather data.

The purpose of this study is to predict the price of cabbage in a week, so the average of the time series data over a given period of time is obtained and used as an independent variable. Because all independent variables are time series data, we averaged all independent variables over a week and estimated the price at the present time through the variables from two weeks to a week before the present time.

Data filled with missing values are a total of 981 data with an even distribution of 248, 246, 243, 244 data per year. The data are all numeric value. Figure 2 shows the mean temperature distribution for 2015. Figure 3 shows the distribution of average temperatures from 2015 to 2018. Most weather data have been found to be highly correlated with average temperature data. Weather data also reflect seasonal characteristics on a yearly basis. Figure 4 shows the average price of cabbages and radishes from 2015 to 2018. Through Figure 3 and Figure 4, we can verify a similar graph shape and seasonal characteristics every year.

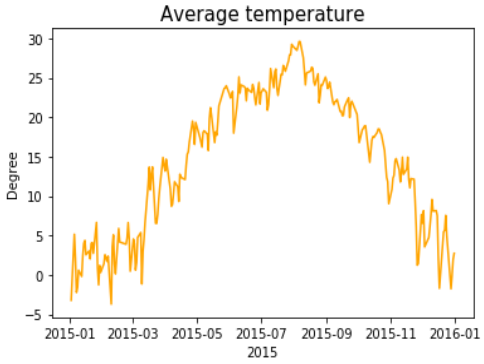


Figure 2. Average temperature in 2015

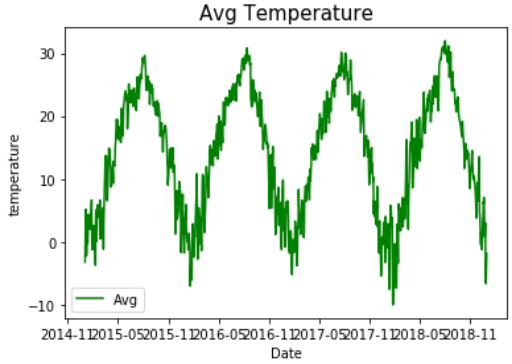
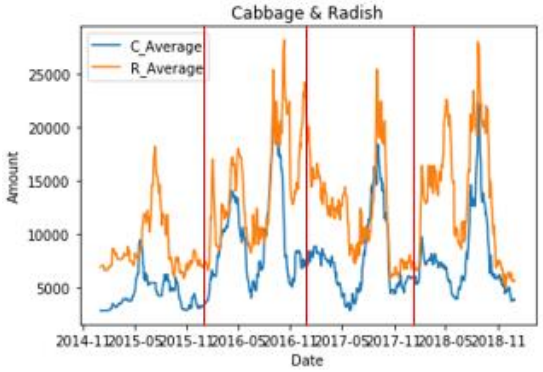


Figure 3. Average temperature from 2015 to 2018



	VIF Factor	features
0	48084.703408	Sea-level pressure_W
1	45841.794674	Spot atmospheric pressure_W
2	36819.492206	Average temperature_W
3	26730.655590	Dew point_W
4	7551.560260	Highest temperature_W
5	7321.667346	Lowest temperature_W
6	1559.323067	Relative humidity_W
7	1043.492891	Average temperature of ground_W
8	258.881311	Duration of bright sunshine_W
9	177.727575	Total large evaporation_W
10	163.865282	Total low evaporation_W
11	109.182668	Vapor pressure_W
12	104.904881	Amount of cumulus_W
13	99.553236	Solar radiation quantity_W
14	72.459196	Amount of middle and lower cloud_W
15	36.622435	Import price_W
16	35.608547	Import weight_W
17	25.681921	Wind speed_W
18	14.610881	Radish_Normalyear_W
19	9.989151	Cabbage_Normalyear_W
20	8.530164	Cabbage_Price_W
21	8.049111	Radish_Price_W
22	5.727877	Maximum depth of snow cover_W
23	5.098014	Maximum depth of new snowfall_W
24	4.571857	Precipitation_W

Figure 4. Average price of cabbages and radishes from 2015 to 2018

Figure 5. VIF of independent variables

Multicollinearity is measured in this study to identify independent variables to be used for pricing forecasting.

Figure 5 shows the result of measuring the multicollinearity of each variable about all independent variables. From **Sea-level pressure** (index 0) to **Radish_NormalYear** (index 18), the VIF factor exceeds 10. These variables are not suitable for models to predict accurate results. Therefore, the variables are removed one by one to measure their multicollinearity, taking into account the effects between each independent variable.

	VIF Factor	features
0	48063.591381	Sea-level pressure_W
1	45826.389579	Spot atmospheric pressure_W
2	26554.342402	Dew point_W
3	14348.627680	Average temperature_W
4	4757.213022	Lowest temperature_W
5	1543.887022	Relative humidity_W
6	1043.442711	Average temperature of ground_W
7	257.114050	Duration of bright sunshine_W
8	175.882920	Total large evaporation_W
9	162.328363	Total low evaporation_W
10	104.809545	Vapor pressure_W
11	102.633627	Amount of cumulus_W
12	99.309199	Solar radiation quantity_W
13	72.458978	Amount of middle and lower cloud_W
14	36.378661	Import price_W
15	35.237003	Import weight_W
16	25.049877	Wind speed_W
17	14.480439	Radish_Normalyear_W
18	9.932128	Cabbage_Normalyear_W
19	8.530113	Cabbage_Price_W
20	8.029454	Radish_Price_W
21	5.725926	Maximum depth of snow cover_W
22	5.075288	Maximum depth of new snowfall_W
23	4.568344	Precipitation_W

(a)when **Highest temperature** was removed

	VIF Factor	features
0	46127.276419	Sea-level pressure_W
1	44077.058682	Spot atmospheric pressure_W
2	23449.078326	Average temperature_W
3	7476.805882	Highest temperature_W
4	7218.550163	Lowest temperature_W
5	1070.303525	Dew point_W
6	1032.642719	Average temperature of ground_W
7	247.626138	Duration of bright sunshine_W
8	177.727521	Total large evaporation_W
9	163.740135	Total low evaporation_W
10	104.849669	Amount of cumulus_W
11	97.784590	Vapor pressure_W
12	95.589953	Solar radiation quantity_W
13	72.099589	Amount of middle and lower cloud_W
14	35.369003	Import price_W
15	34.417530	Import weight_W
16	25.675541	Wind speed_W
17	14.464391	Radish_Normalyear_W
18	9.987378	Cabbage_Normalyear_W
19	8.530088	Cabbage_Price_W
20	8.020822	Radish_Price_W
21	5.714485	Maximum depth of snow cover_W
22	5.074014	Maximum depth of new snowfall_W
23	3.862030	Precipitation_W

(b) when **Highest Relative Humidity** was removed

Figure 6. VIF factor of variable sets when a variable who has more than 10 multicollinearity was removed

Figure 6 shows the result of measuring the multicollinearity when each (a)**Highest temperature** and (b)**Highest Relative Humidity** are removed from a set of variables with greater than 10 multicollinearity. In this study, all variables in a set of variables whose multicollinearity is greater than 10 were removed one by one to confirm the multicollinearity, but no apparent change was found. This is because, as identified in the EDA step, each independent variable has a similar distribution shape.

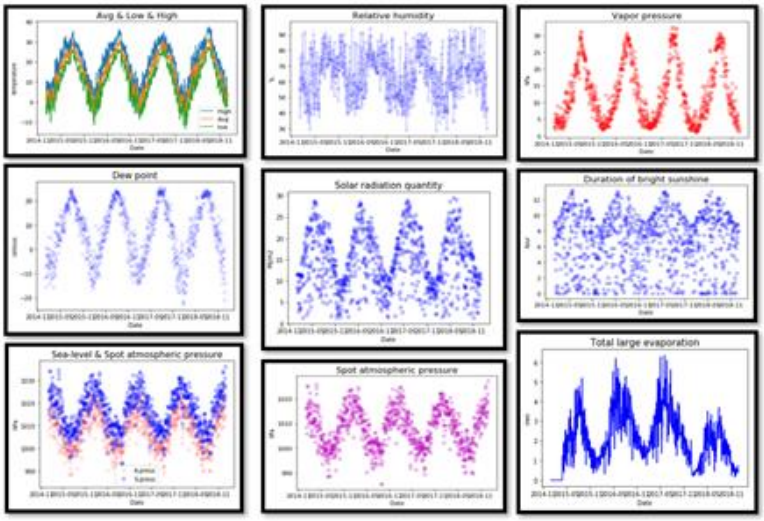


Figure 7. the graphs of independence variables whose multicollinearity is greater than 10

Figure 7 shows a graphical representation of some of the independent variables whose multicollinearity is greater than 10. Each independent variable shows a similar graph shape. Therefore, in this study, the method of measuring the multicollinearity was used by removing all independent variables with greater than 10 multicollinearity and adding them again one by one. Figure 8 shows a pseudo algorithm that selects variables that keep the multicollinearity less than 10.

```
variableList = all Independece variable;
remainList = new remainList();
for(variable : variableList):
        if(variable.VIF > 10):
                variableList.remove(variable);
                remainList.add(variable);
for(variable : remainList)
        if(variableList.add(variable).VIF_all <10 && variable == minimum VIF increasing):
                        variableList.add(variable)
```

Figure 8. Pseudo algorithm of VIF variable selection

2.2 model

The independent variables used in this study are all time series data. Therefore, we separate validation into 2018 data and train into 2017 data from 2015.

Backward elimination technique is used for feature selection in separated train set. When the first time the radish normal year was removed, the R2_square increased. But when the other variable was removed sequentially, R2_square was decreased. Therefore, only the radish normal year was removed from the train set.

Because the variables used in this study are all numerical variables, Linear Support Vector Regression, Linear Regression, KNN Regression, Gradient Boosting Reform, Random Forest Regression, Support Vector Regression, and Decision Tree Regression were used for evaluation. A model was run using the

cabbage price at the current time as a target variable and the data averaged from two weeks ago to a week ago as an independent variable.

2.3 result

Table 4 show the results of running each model.

	Method	Score
1	Gradient Boosting Regression	0.64004
2	KNN Regression	0.617634
3	Support Vector Regression	0.595854
4	Linear Support Vector Regression	0.571686
5	Linear Regression	0.565908
6	Random Forest Regression	0.503584
7	Decision Tree Regression	0.213532

Table 4. Results of running each model

Gradient Boosting Regression had the highest score of 64%, followed by 61.8% of the K-NN Regression and 59.6% of the Support Vector Regression.

3. Conclusion

This study shows that the price of cabbages after one week can be predicted using the time series data of weekly average data consisted of the cabbage price, radish price, weather data, and cabbage import data.

In 2.1 it shows processes for data collection and preprocessing. In 2.2 it shows several used models and evaluations of them. In 2.3 it shows predicted results.

With cabbage prices, weather data and agriculture import data, we could predict the price of cabbages at about 64% R_square, but we expect better prediction results when factors such as the production, consumption or production area of cabbages are given which is not considered in this study.

reference

[1]: Korea Health Industry Development Institute(KHIDI). National Food & Nutrition Statistics 2013: based on 2013 Korea National Health and Nutrition Examination Survey. Korea Health Industry Development Institute(KHIDI). 2015

[2]: Kuk-Hyun Nam, Young Chan Choe, A Study on Onion Wholesale Price Forecasting Model, 한국농촌지도학회, 2018

[3]: Jang, SooHee ,Chun, Heuiju ,Cho, Inho ,Kim, DongHwan, A study on cabbage wholesale price forecasting model using unstructured agricultural meteorological data, 한국데이터정보과학회지, 2017

[4]: Ji-Yun, Park, Young-Gu, Park, The Development of Chinese Cabbage and Radish Forecast Models, 한국농촌경제연구원, 2013