2018-1 Machine Learning & Data Mining

기상데이터 회귀 모델을 통한 배추 가격 예측

201420874 이성훈
201420907 안우일
201420920 김영운
201421120 김필선

1. 주제선정 동기

한국인에게 김치는 없어서는 안 될 친근한 음식이다. 날씨가 추워지고 겨울이 다가오는 때가 되면, 가정에서는 한 해 동안 먹을 김치를 만들기 위해 김장을 하는 풍경이 그려진다.

이에 따라 10월 정도가 되면, 김치의 원재료인 배추에 대한 수요량은 다른 때보다 높아지게 된다. 이때, 만약 배추의 공급이 원활하지 못한다면 그 가격은 평소보다 증가하게 된다. 이처럼 농산물의 공급량이 이러한 수요량을 만족시키지 못한다면 가격이 증가하는 현상이 발생하게 된다. 이러한 농산물의 가격 변동 현상은 상품을 직접 구매해야 하는 소비자에게 직접적으로 영향을 주게 된다.

<그림1> 2013년 배추가격변화

그렇다면 농산물의 공급량 변동이 발생하는 이유는 무엇일까. 현상을 분석해본 결과, 우리는 길게 지속되는 장마, 강도 높은 폭염, 돌풍과 폭우를 수반하는 태풍 등이 있었던 해에 가격 급등 현상이 발생하는 것을 쉽게 찾아볼 수 있었다. 인공적으로 재배 환경을 조성할 수 있는 비닐하우스 등의 실내에서 자라는 농산물도 있겠지만 이는 한정적인 범주이며, 사실상 대부분의 농산물은 실외에서 심어지고 자라게 된다. 이에 기초하여 우리는 농산물의 재배 환경과 수확량의 측면에서, 날씨가 미치는 영향이 매우 큰 것으로 판단하였고 이는 결국 공급량에 직접적인 영향을 주게 된다고 생각하였다. 그래서 과거의 여러 날씨 데이터를 이용하여 회귀 분석을 한다면, 미래의 농산물의 가격을 예측할 수 있다는 생각에 착안하였다.

우리는 여러 농산물 중, 한국에서 김치의 원재료로 사용되어, 매해 수요가 꾸준한 배추를 주된 대상으로 특정하였다. 먼저 배추의 생산량 대부분을 차지하며, 주요 재배지인 전라남도 해남지역의 날씨 데이터를 수집할 것이다. 그 후에 필요에 맞게 정제하고, 최종적으로 모델에 사용할 데이터를 얻어낼 것이다. 배추의 모종부터 판매시기까지의 기간을 고려하여 학습 데이터를 정하고, 이를 학습하여 미래의 배추 가격 변동을 예측할 수 있는 모델을 만들 것이다. 이후에는 모델의 예측 성능을 높이기 위하여 가격에 영향을 미치는 요소들을 조정하고 여러 도구들을 사용해보면서 최적의 모델을 선택할 것이다. 또한, 농산물의 가격을 직접 예측하는 것이 주된 목표지만 부가적으로 농산물의 가격이 상승할지 하강할지에 대한 예측도 진행해 볼 것이다.

2. 데이터 전처리

기상데이터를 통한 회귀모델을 구축하기 위해서는 기상데이터와 농산물 가격데이터가 필요했다. 기상 데이터로는 기상자료개방포털(https://data.kma.go.kr/cmmn/main.do)에서 얻었고, 농업기상관측 데이터를 통해 얻을 수 있는 약 20가지 기상 데이터들을 활용하였다. 배추의 가격 데이터는 유통센터(http://kostat.go.kr/wnsearch/search.jsp)를 통해 얻을 수 있었다. 각각 2008~2017년까지의 데이터를 모았으며, 결측값(Missing Value)이 있는 경우 1일 전의 데이터로 채웠다.

<그림 1> 기상자료개방포털과 유통센터

각 년도의 기상과 배추가격 데이터를 결합하여 최종 데이터로 사용하였으며, 특정 날의 가격의 feature들로 사용할 기상데이터로 4달전 2주 기상데이터들의 mean값과 median값을 사용하였다. 회귀식 구성을 위한 Train Data로 2008~2016년까지의 기상, 가격 데이터를 사용했으며, 마지막 최종 성능을 측정하기 위한 Test Data로 2017년의 기상, 가격 데이터를 사용하였다.

3. 모델 선택 과정

3.(1) 회귀 기간 선정

가장 효율적으로 배추 가격예측이 쓰이기 위해서는 계절에 관계없이 모든 데이터로 회귀를 하는 것이 좋을 것이다. 이에 따라 2008~2016년까지의 모든 데이터로 회귀하는 것을 시도해봤다. 하지만, 배추가 계절에 따라 키우는 방법과 영향을 받는 요인이 다르다고 생각되었다. 따라서 정확도를 높이는 측면에서 3개월의 분기별로 데이터를 모아 회귀를 시도하는 방법을 시도해봤다.(ex. 1~3월을 예측할 때 2008~2016년의 1~3월 데이터로 회귀를 시도함)

3.(2) Feature Selection

Feature Selection은 Regression과정에서 매우 중요하다고 할 수 있다. 사용하는 Feature의 수와 종류에 따라서 회귀의 정확도가 달라질 수 있기 때문이다. 회귀 모델에서 Feature Selection을 하는 방법은 크게 3가지로, Exhaustive, Forward, Backward가 있다.

이번 프로젝트에서는 scikit-learn에서 제공하는 backward 방식의 RFE(Recursive Feature Elimination)을 사용하여서 3개~10개까지의 Feature들을 고를 때, 어떤 Feature들을 선택해야 하는지 결정하였다. 결과는 아래 표와 같다.

Feature	기호	Feature	기호
평균기온	A	최저기온	B
최고기온	C	일 강수량	D
평균 이슬점온도	E	최대 풍속	F
최대 순간풍속	G	최소 상대습도	H
평균 상대습도	I	평균 증기압	J
평균 현지기압	K	최고 해면기압	L
최저 해면기압	M	평균 해면기압	N
합계 일조시간	O	평균 지면온도	P
최저 초상온도	R		

전체 기간(1월 ~ 12월) Feature	
3	N, Q, R
4	L, N, Q, R
5	L, N, O, Q, R
6	L, M, N, O, Q, R
7	K, L, M, N, O, Q, R
8	E, K, L, M, N, O, Q, R
9	E, I, K, L, M, N, O, Q, R
10	D, E, I, K, L, M, N, O, Q, R
11	D, E, F, I, K, L, M, N, O, Q, R
12	D, E, F, G ,I, K, L, M, N, O, Q, R
13	D, E, F, G ,I, K, L, M, N, O, P, Q, R
14	C, D, E, F, G ,I, K, L, M, N, O, P, Q, R
15	C, D, E, F, G H, I, K, L, M, N, O, P, Q, R
16	A, C, D, E, F, G H, I, K, L, M, N, O, P, Q, R
17	A, C, D, E, F, G H, I, J, K, L, M, N, O, P, Q, R
18	A, B, C, D, E, F, G H, I, J, K, L, M, N, O, P, Q, R

1~3월의 Feature	
3	A, C, J
4	A, C, J, N
5	A, C, H, J, N
6	A, C, H, I, J, N
7	A, C, H, I, J, N, R
8	A, C, D, H, I, J, N, R
9	A, C, D, E, H, I, J, N, R
10	A, C, D, E ,H, I, J, K, N, R

2~4월의 Feature	
3	C, I, R
4	C, E, I, R
5	C, E, H, I, R
6	C, E, H, I, N, R
7	C, E, H, I, L, N, R
8	C, E, H, I, L, N, O, R
9	C, E, H, I, J, L, N, O, R
10	C, D, E, H, I, J, L, N, O, R

3~5월의 Feature	
3	E, I, L
4	B, E, I, L
5	B, E, I, L, R
6	B, E, I, L, Q, R
7	B, D, E, I, L, Q, R
8	B, D, E, I, L M, Q, R
9	B, D, E, F, I, L M, Q, R
10	B, C, D, E, F, I, L, M, Q, R

4~6월의 Feature	
3	I, L, N
4	D, I, L, N
5	D, I, J, L, N
6	D, I, J, L, N, O
7	D, I, J, L, N, O Q
8	D, I, J, L, N, O Q
9	D, E, I, J, L, N, O Q
10	D, E, F, I, J, L, N, O, Q

5~7월의 Feature	
3	E, I, Q
4	E, I, L, Q
5	E, I, L, Q, R
6	E, F, I, L, Q, R
7	B, E, F, I, L, Q, R
8	B, E, F, I, L O, Q, R
9	B, E, F, I, L O, Q, P, R
10	B, E, F, I, L, N, O, P, Q, R

6~8월의 Feature	
3	D, E, Q
4	D, E, N, Q
5	D, E, F, N, Q
6	B, D, E, F, N, Q
7	B, D, E, F, K, N, Q
8	B, D, E, F, K, N, P, Q
9	B, C, D, E, F, K, N, P, Q
10	B, C, D, E, F, I, K, N, P, Q

7~9월의 Feature	
3	E, F, Q
4	E, F, I, Q
5	E, F, I, Q, R
6	D, E, F, I, Q, R
7	D, E, F, I, P, Q, R
8	D, E, F, I, K, P, Q, R
9	D, E, F, I, K, L, P, Q, R
10	D, E, F, I, J, K, L, P, Q, R

8~10월의 Feature	
3	D, J, R
4	B, D, J, R
5	B, D, J, M, R
6	B, D, F, J, M, R
7	B, D, E, F, J, M, R
8	B, D, E, F, J, K, M, R
9	B, D, E, F, J, K, L, M, R
10	B, D, E, F, J, K, L, M, P, R

9~11월의 Feature	
3	L, M, R
4	E, L, M, R
5	E, L, M, O, R
6	E, L, M, N, O, R
7	E, F, L, M, N, O, R
8	E, F, L, M, N, O, Q, R
9	D, E, F, L, M, N, O, Q, R
10	C, D, E, F, L, M, N, O, Q, R

10~12월의 Feature	
3	I, N, Q
4	I, K, N, Q
5	I, J, K, N, Q
6	I, J, K, N, P, Q
7	I, J, K, N, P, Q, R
8	G, I, J, K, N, P, Q, R
9	G, I, J, K, M, N, P, Q, R
10	C, G, I, J, K, M, N, P, Q, R

11~1월의 Feature	
3	K, M, Q
4	I, K, M, Q
5	I, J, K, M, Q
6	C, I, J, K, M, Q
7	C, D, I, J, K, M, Q
8	C, D, I, J, K, M, N, Q
9	C, D, I, J, K, M, N, Q, R
10	C, D, I, J, K, M, N, P, Q, R

12~2월의 Feature	
3	D, M, Q
4	D, M, Q, R
5	D, M, P, Q, R
6	D, J, M, P, Q, R
7	D, I, J, M, P, Q, R
8	D, I, J, K, M, P, Q, R
9	A, D, I, J, K, M, P, Q, R
10	A, D, I, J, K, M, N, P, Q, R

3.(3) Regression Model 선정

3.(3).1 Scikit-learn의 Lasso, Ridge, Elastic Net Regression을 사용

Lasso는 L_1norm으로, manhattan distance와 형태의 함수를 제공한다. Lasso의 경우 최적 값은 모서리 부분에서 나타날 확률이 Ridge에 비해 높아 몇몇 유의미하지 않은 변수들에 대해 계수를 빠르게 0으로 만들어, 의미 있는 변수들을 추정할 수 있게 만들어준다.

$$\min\left(||Y - X\theta||_2^2 + \lambda||\theta||_1\right)$$

<그림 3> Lasso 식

Ridge는 L_2 norm으로, L_2일 경우 euclidean distance 형태의 함수를 제공한다. Lasso에 비해 계수들이 0으로 가는 속도가 느리며, 상관성이 있는 변수들에 대해서 적절한 가중치 배분을 하게 된다. 즉, 실제 영향력 있는 정보만큼 압축해주기 위해서 Penalty를 부여하는 방법이라고 할 수 있다.

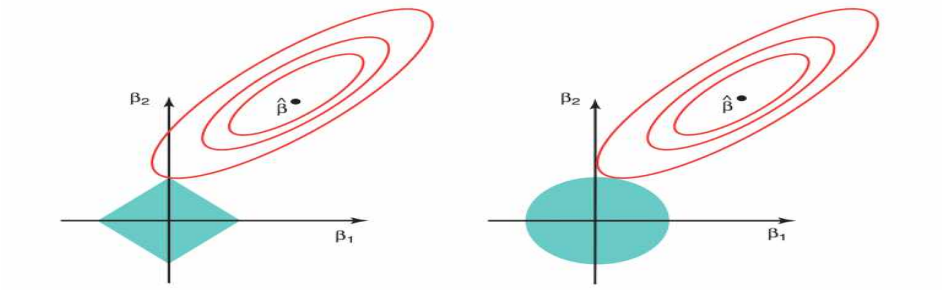
$$\min\left(||Y - X(\theta)||_2^2 + \lambda||\theta||_2^2\right)$$

<그림 4> Ridge 식

Elastic-Net의 경우는 Lasso와 Ridge를 결합한 형태라고 할 수 있다. 각각 λ_1과 λ_1는 Lasso와 Ridge의 파라미터와 같은 역할을 한다고 할 수 있다. 파라미터의 값에 따라 어느 것에 더 가중치를 두고 정규화를 할지 결정할 수 있다.

$$\min\left(||Y - X\theta||_2^2 + \lambda_1||\theta||_1 + \lambda_2||\theta||_2^2\right)$$

<그림 5> Elastic-net 식



<그림 4> Lasso(좌)와 Ridge(우)

3.(3).2 SVR을 사용(kernel = linear, RBF) 사용

SVR은 Support Vector Regression을 의미한다. SVR은 SVM과 같은 개념의 알고리즘이나, SVM은 Classifier로, SVR은 Regression으로 사용된다는 점에서 차이점을 가진다. scikit-learn에서 사용하려면 'from sklearn.svm import SVR'로 SVR을 import한 뒤, SVR()을 통해 model을 만들면 된다.

SVR의 대표적인 parameter로는 kernel, gamma, C가 있다. kernel의 경우 SVR의 kernel type을 정의하는 것으로, 'linear', 'rbf', 'sigmoid', 'poly'등이 쓰일 수 있다. gamma는 Kernel coefficient이고, C는 error term에 대한 Penalty parameter이다.

4. 분석 결과 및 평가

회귀 모델을 평가하는 방식은 여러 가지가 있다. 이번 프로젝트에서는 크게 두 가지 방법으로 모델을 평가했다. 첫째로, R Square Error)를 통해 모델을 평가했다. MSE로 모델을 평가할 때, scikit-learn의 R2_score와 adjusted_R2 score 함수를 정의해 통해 계산하였다. 두 번째로는 가격의 상승과 하락을 얼마나 맞췄는가를 기준으로 모델을 평가했다. 상승/하락을 맞춘 정확도로 모델을 평가할 때, python code로 함수를 정의하여 계산하였다.

```python
def adjusted_r2(real_val, pred_val, sample, independent_var):
    r2 = r2_score(real_val, pred_val)
    adj_r2 = 1 - (1 - r2) * (sample - 1) / (sample - independent_var - 1)

    return adj_r2
```

```python
def check_trends(array_true, array_pred):
    total = len(array_true)
    index = 0
    same = 0

    while index < total:
        if array_true[index] == array_pred[index]:
            same = same + 1
        index = index + 1

    print(">>> 일치하는 원소의 개수는", total, "개 중", same)
    check = (same / total) * 100
    print(">>> 일치율 :", check, "%")
    print('=' * 100)
```

<그림6> R Square(좌)와 상승/하락정확도 계산 코드

회귀 분석에 앞서, cross-validation을 시행하여 hyper-parameter를 조사했다. scikit-learn에서 제공하는 회귀 방법들 중 총 5가지 방법을 사용하였고, Linear Regression, Lasso, Ridge, Elastic을 한 그룹으로, SVR(linear), SVR(rbf)를 한 그룹으로 나눠서 회귀 분석을 진행하여 표를 작성했다

- Cross-Validation 결과 (li: Linear Regression, L: Lasso, R: Ridge, E : Elastic)
(상승/하락 일치율 (%))

n / α	3	4	5	6	7	8	9	10
0.05	li: 32.88 L: 32.92 R: 32.88 E: 32.92	li: 32.79 L: 32.82 R: 32.79 E: 33.04	li: 33.32 L: 33.28 R: 33.32 E: 33.28	li: 33.65 L: 33.65 R: 33.66 E: 33.69	li: 33.61 L: 33.73 R: 33.61 E: 33.69	li: 33.29 L: 33.33 R: 33.29 E: 33.25	li: 32.99 L: 33.15 R: 32.99 E: 32.99	li: 33.43 L: 33.19 R: 33.43 E: 33.82
0.25	li: 32.88 L: 32.88 R: 32.88 E: 32.92	li: 32.79 L: 32.79 R: 32.79 E: 32.79	li: 33.32 L: 33.32 R: 33.36 E: 33.36	li: 33.65 L: 33.65 R: 33.65 E: 33.12	li: 33.61 L: 33.73 R: 33.61 E: 33.36	li: 33.29 L: 33.33 R: 33.29 E: 33.08	li: 32.99 L: 33.15 R: 32.99 E: 33.20	li: 33.43 L: 33.15 R: 33.47 E: 33.32
0.5	li: 32.88 L: 32.88 R: 32.88 E: 33.00	li: 32.79 L: 32.83 R: 32.79 E: 32.59	li: 33.32 L: 33.32 R: 33.32 E: 33.36	li: 33.65 L: 33.65 R: 33.65 E: 33.04	li: 33.61 L: 33.69 R: 33.61 E: 33.32	li: 33.29 L: 33.37 R: 33.29 E: 33.12	li: 32.99 L: 33.10 R: 32.99 E: 33.12	li: 33.43 L: 33.15 R: 33.48 E: 33.56
1.0	li: 32.88 L: 32.88 R: 32.88 E: 33.04	li: 32.79 L: 32.83 R: 32.79 E: 32.64	li: 33.32 L: 33.32 R: 33.36 E: 33.12	li: 33.65 L: 33.61 R: 33.65 E: 32.87	li: 33.61 L: 33.65 R: 33.61 E: 33.49	li: 33.29 L: 33.37 R: 33.29 E: 33.16	li: 32.99 L: 33.11 R: 32.99 E: 33.16	li: 33.46 L: 33.15 R: 33.43 E: 33.52
2.0	li: 32.88 L: 32.88 R: 32.88 E: 32.88	li: 32.79 L: 32.83 R: 32.83 E: 32.80	li: 33.32 L: 32.87 R: 33.36 E: 32.87	li: 33.65 L: 33.61 R: 33.61 E: 32.87	li: 33.61 L: 33.65 R: 33.58 E: 33.20	li: 33.29 L: 33.41 R: 33.24 E: 33.24	li: 32.99 L: 33.11 R: 33.07 E: 33.16	li: 33.43 L: 33.15 R: 33.43 E: 33.32
5.0	li: 32.88 L: 32.88 R: 32.88 E: 32.92	li: 32.79 L: 32.87 R: 32.83 E: 32.84	li: 33.32 L: 33.28 R: 33.36 E: 32.63	li: 33.65 L: 33.52 R: 33.61 E: 32.75	li: 33.61 L: 33.65 R: 33.61 E: 33.48	li: 33.29 L: 33.45 R: 33.29 E: 33.40	li: 32.99 L: 33.03 R: 33.07 E: 33.11	li: 33.43 L: 33.03 R: 33.47 E: 33.16

n : # of features

Lasso, Ridge, Elastic CV 결과 (r2) (L: Lasso, R: Ridge, E : Elastic)

α \ n	3	4	5	6	7	8	9	10
0.05	L:-2.954 R:-2.954 E:-2.938	L:-2.945 R:-2.945 E:-2.924	L:-2.800 R:-2.800 E:-2.773	L:-2.746 R:-2.745 E:-2.719	L:-2.840 R:-2.840 E:-2.808	L:-2.662 R:-2.662 E:-2.646	L:-2.760 R:-2.763 E:-2.728	L:-2.895 R:-2.895 E:-2.857
0.25	L:-2.880 R:-2.954 E:-2.953	L:-2.944 R:-2.945 E:-2.852	L:-2.799 R:-2.800 E:-2.696	L:-2.745 R:-2.745 E:-2.643	L:-2.839 R:-2.839 E:-2.715	L:-2.662 R:-2.662 E:-2.599	L:-2.760 R:-2.762 E:-2.664	L:-2.895 R:-2.895 E:-2.777
0.5	L:-2.953 R:-2.954 E:-2.818	L:-2.944 R:-2.945 E:-2.781	L:-2.799 R:-2.799 E:-2.628	L:-2.745 R:-2.745 E:-2.577	L:-2.838 R:-2.839 E:-2.634	L:-2.662 R:-2.662 E:-2.551	L:-2.761 R:-2.762 E:-2.614	L:-2.896 R:-2.895 E:-2.709
1.0	L:-2.952 R:-2.953 E:-2.718	L:-2.943 R:-2.944 E:-2.673	L:-2.798 R:-2.799 E:-2.531	L:-2.744 R:-2.745 E:-2.485	L:-2.837 R:-2.839 E:-2.524	L:-2.662 R:-2.662 E:-2.475	L:-2.761 R:-2.762 E:-2.539	L:-2.897 R:-2.894 E:-2.609
2.0	L:-2.950 R:-2.953 E:-2.582	L:-2.941 R:-2.944 E:-2.533	L:-2.796 R:-2.799 E:-2.415	L:-2.742 R:-2.744 E:-2.378	L:-2.834 R:-2.838 E:-2.400	L:-2.662 R:-2.662 E:-2.377	L:-2.763 R:-2.761 E:-2.441	L:-2.898 R:-2.893 E:-2.484
5.0	L:-2.945 R:-2.952 E:-2.390	L:-2.935 R:-2.943 E:-2.344	L:-2.791 R:-2.797 E:-2.275	L:-2.737 R:-2.743 E:-2.256	L:-2.826 R:-2.836 E:-2.265	L:-2.662 R:-2.661 E:-2.262	L:-2.768 R:-2.758 E:-2.316	L:-2.902 R:-2.890 E:-2.332

Cross-validation 결과, 함수 내의 hyper-parameter로 정확도가 유의미하게 변화하지 않는 것으로 파악되었다. 따라서 이후에 Regression을 진행할 때, (Linear Regression, Lasso, Ridge, Elastic)의 Hyper-parameter인 alpha값을 1.0으로, SVR의 Hyper-parameter값인 gamma값을 1.0으로 설정하여 Regression을 진행하였다.

4.(1) 1년 단위로 Regression을 시행했을 때

-Lasso, Ridge, Elastic, 결과 (r2, adjusted r2, 일치율)

	3	4	5	6
Elastic	r2: 0.036 adj r2: 0.024 일치율: 36.92%	r2: 0.023 adj r2: 0.006 일치율: 36.51%	r2: 0.018 adj r2: -0.002 일치율: 36.92%	r2: 0.020 adj r2: -0.005 일치율: 36.51%
Ridge	r2: 0.056 adj r2: 0.044 일치율: 36.51%	r2: 0.006 adj r2: -0.011 일치율: 34.85%	r2: -0.040 adj r2: -0.062 일치율: 36.92%	r2: -0.025 adj r2: -0.051 일치율: 37.34%
Lasso	r2: 0.056 adj r2: 0.044 일치율: 36.51%	r2: 0.006 adj r2: -0.011 일치율: 34.85%	r2: -0.039 adj r2: -0.061 일치율: 36.51%	r2: -0.028 adj r2: -0.054 일치율: 37.34%

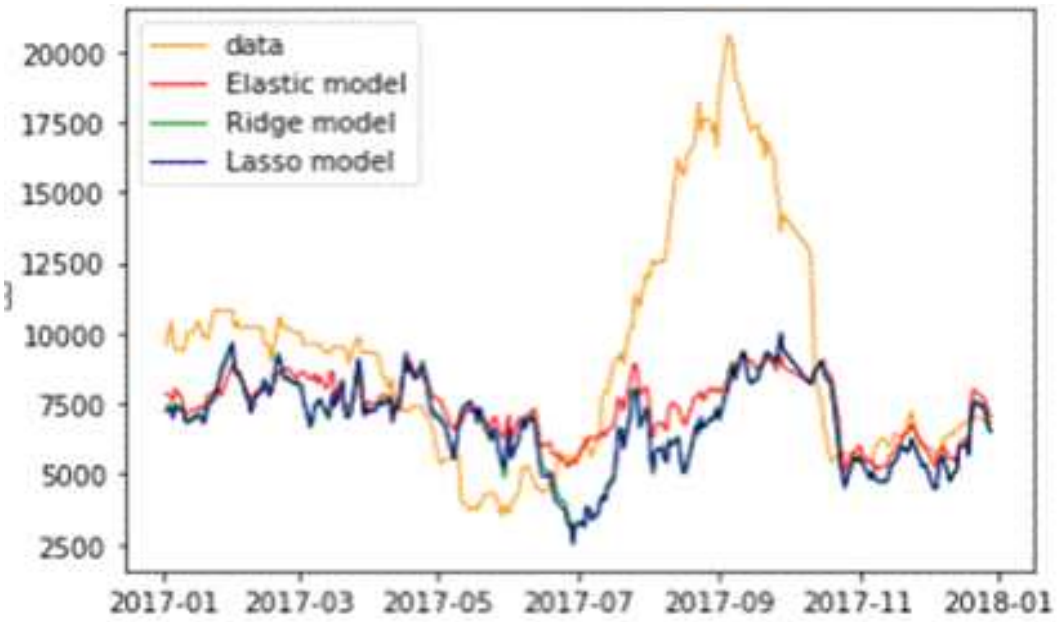
	7	8	9	10
Elastic	r2: 0.066 adj r2: 0.038 일치율: 39.41%	r2: 0.058 adj r2: 0.026 일치율: 37.34%	r2: 0.109 adj r2: 0.075 일치율: 39.41%	**r2: 0.115** **adj r2: 0.076** **일치율: 39.00%**
Ridge	r2: -0.011 adj r2: -0.041 일치율: 36.92%	r2: -0.011 adj r2: -0.046 일치율: 37.34%	r2: 0.087 adj r2: 0.051 일치율: 35.26%	**r2: 0.101** **adj r2: 0.062** **일치율: 35.68%**
Lasso	r2: -0.013 adj r2: -0.043 일치율: 38.17%	r2: -0.013 adj r2: -0.048 일치율: 37.75%	r2: 0.083 adj r2: 0.048 일치율: 35.68%	**r2: 0.097** **adj r2: 0.058** **일치율: 35.68%**

-SVR(kernel=linear), SVR(kernel=rbf) (r2, adjusted r2, 일치율)

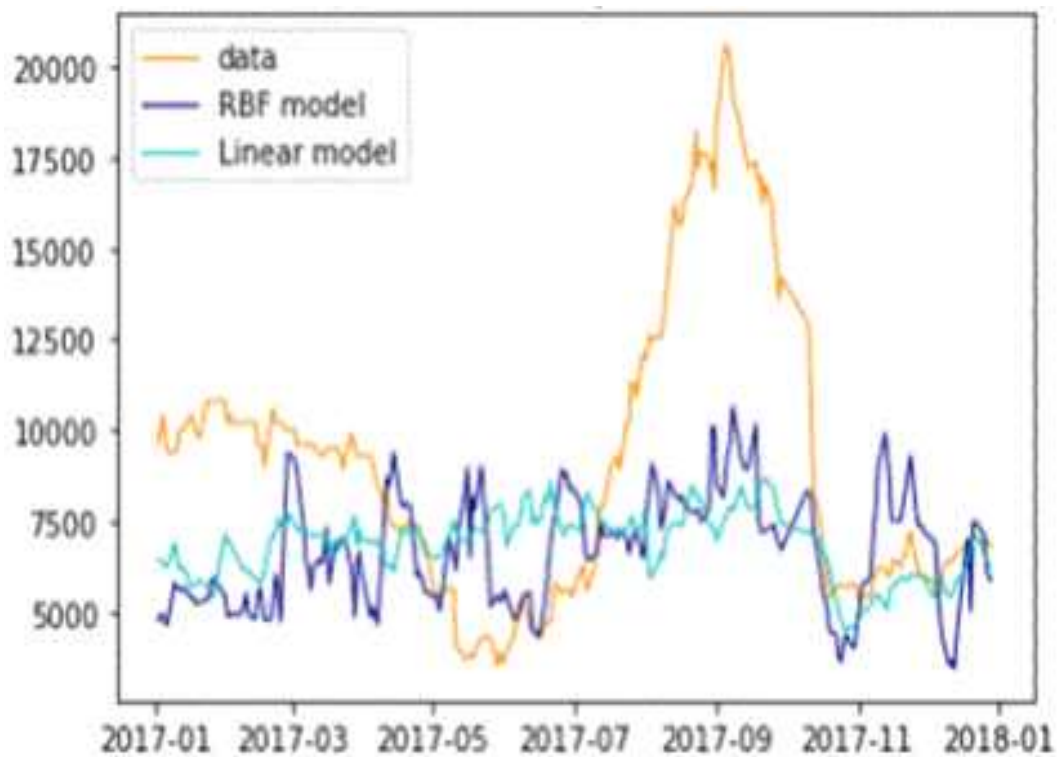
	3	4	5	6
SVR(Linear)	r2: -0.277 adj r2: -0.277 일치율: 36.51%	r2: -0.155 adj r2: -0.155 일치율: 35.26%	r2: -0.179 adj r2: -0.179 일치율: 34.85%	r2: 0.330 adj r2: -0.330 일치율: 36.09%
SVR(RBF)	r2: 0.303 adj r2: -0.303 일치율: 38.58%	r2: -0.326 adj r2: -0.326 일치율: 35.68%	r2: -0.341 adj r2: -0.341 일치율: 36.51%	r2: -0.379 adj r2: -0.379 일치율: 39.00%

	7	8	9	10
SVR(Linear)	r2: -0.340 adj r2: -0.340 일치율: 35.68%	r2: -0.299 adj r2: -0.299 일치율: 39.00%	r2: -0.488 adj r2: -0.488 일치율: 36.09%	**r2: 0.022** **adj r2: 0.022** **일치율: 36.09%**
SVR(RBF)	r2: -0.432 adj r2: -0.432 일치율: 40.66%	r2: -0.446 adj r2: -0.446 일치율: 36.92%	r2: 0.100 adj r2: 0.061 일치율: 35.68%	**r2: 0.073** **adj r2: 0.028** **일치율: 35.68%**

-Lasso, Ridge, Elastic, 결과 (그래프)



-SVR(kernel=linear), SVR(kernel=rbf) (그래프)



4.(2) 3개월 단위로 시행했을 때

4.(2).1 결과 3~5월

-Lasso, Ridge, Elastic, 결과 (r2, adjusted r2, 일치율)

	3	4	5	6
Elastic	r2: 0.197 ad r2: 0.155 일치율: 26.6%	r2: 0.183 ad r2: 0.125 일치율: 31.6%	r2: 0.159 ad r2: 0.082 일치율: 31.6%	r2: 0.235 ad r2: 0.150 일치율: 28.3%
Ridge	r2: 0.200 ad r2: 0.158 일치율: 26.6%	r2: 0.183 ad r2: 0.125 일치율: 31.6%	r2: 0.131 ad r2: 0.052 일치율: 31.6%	r2: 0.213 ad r2: 0.126 일치율: 28.3%
Lasso	r2: 0.200 ad r2: 0.158 일치율: 26.6%	r2: 0.184 ad r2: 0.12 일치율: 31.6%	r2: 0.132 ad r2: 0.053 일치율: 31.6%	r2: 0.213 ad r2: 0.126 일치율: 28.3%

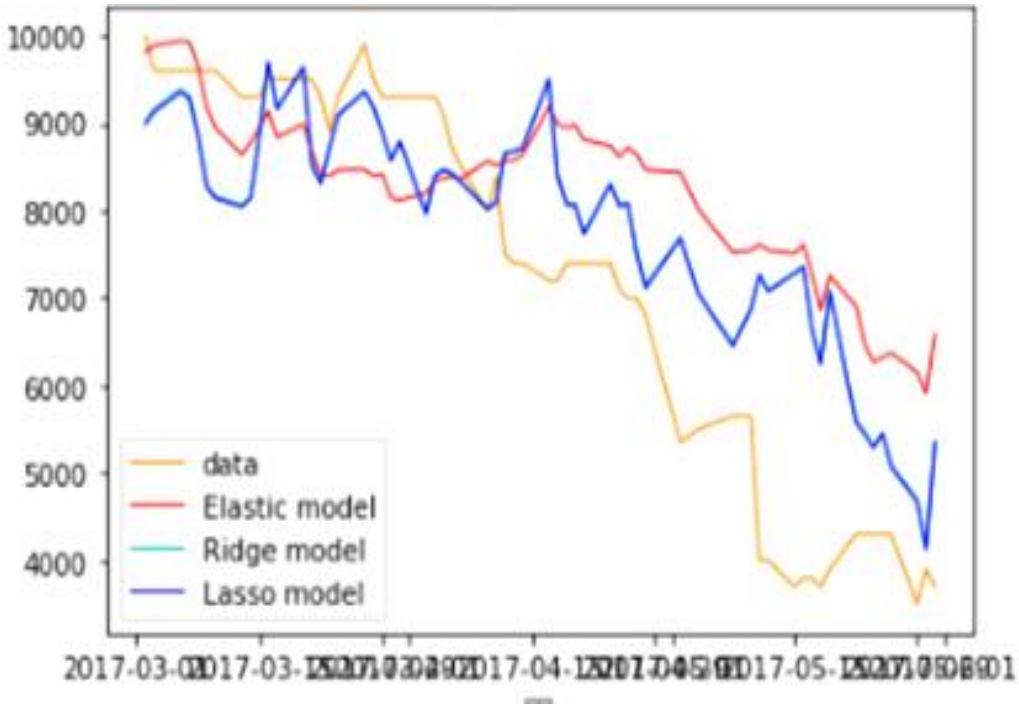
	7	8	9	10
Elastic	r2: 0.284 ad r2: 0.189 일치율: 28.3%	r2: 0.297 ad r2: 0.189 일치율: 26.6%	**r2: 0.338** **ad r2: 0.221** **일치율: 26.7%**	r2: 0.318 ad r2: 0.181 일치율: 30.0%
Ridge	r2: 0.325 ad r2: 0.236 일치율: 28.3%	r2: 0.342 ad r2: 0.241 일치율: 28.3%	**r2: 0.596** **ad r2: 0.524** **일치율: 28%**	r2: 0.536 ad r2: 0.443 일치율: 26.6%
Lasso	r2: 0.325 ad r2: 0.236 일치율: 28.3%	r2: 0.342 ad r2: 0.240 일치율: 28.3%	**r2: 0.597** **ad r2: 0.526** **일치율: 28.3%**	r2: 0.537 ad r2: 0.444 일치율: 26.6%

-SVR(kernel=linear), SVR(kernel=rbf) (r2, adjusted r2, 일치율)

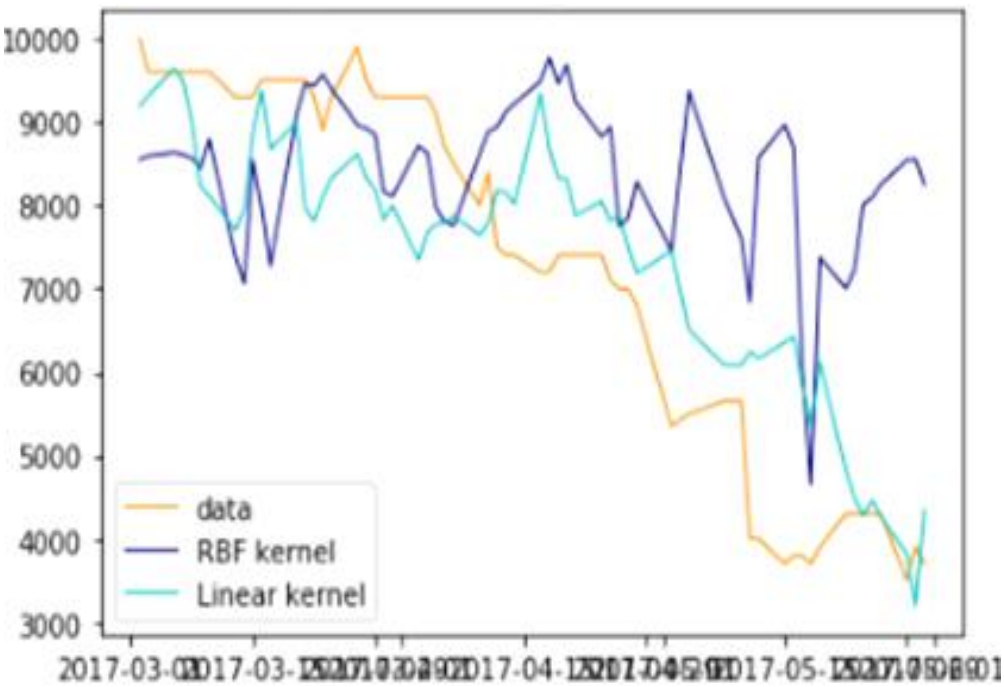
	3	4	5	6
SVR(Linear)	r2: 0.182 ad r2: 0.182 일치율: 25.0%	r2: 0.179 ad r2: 0.179 일치율: 25.0%	r2: -0.024 ad r2: -0.024 일치율: 30.0%	r2: 0.051 ad r2: 0.051 일치율: 25.0%
SVR(RBF)	r2: -0.321 ad r2: -0.321 일치율: 28.3%	r2: -0.158 ad r2: -0.158 일치율: 30.0%	r2: -0.361 ad r2: -0.361 일치율: 28.3%	r2: -0.375 ad r2: -0.375 일치율: 28.3%

	7	8	9	10
SVR(Linear)	r2: 0.110 ad r2: 0.110 일치율: 26.0%	r2: 0.250 ad r2: 0.250 일치율: 25.0%	**r2: 0.650** **ad r2: 0.650** **일치율: 26.7%**	r2: 0.303 ad r2: 0.303 일치율: 23.3%
SVR(RBF)	r2: -0.441 ad r2: -0.441 일치율: 28.3%	r2: -0.450 ad r2: -0.450 일치율: 30.0%	**r2: -0.290** **ad r2: -0.290** **일치율: 26.6%**	r2: -0.502 ad r2: -0.502 일치율: 28.3%

-Lasso, Ridge, Elastic, 결과 (그래프)



-SVR(kernel=linear), SVR(kernel=rbf) (그래프)



4.(2).2 결과 6~8월

-Lasso, Ridge, Elastic, 결과 (r2, adjusted r2, 일치율)

	3	4	5	6
Elastic	**r2: 0.292** **ad r2: 0.257** **일치율: 53.96%**	r2: 0.288 ad r2: 0.239 일치율: 55.5%	r2: 0.287 ad r2: 0.226 일치율: 57.14%	r2: 0.287 ad r2: 0.212 일치율: 52.38
Ridge	**r2: 0.289** **ad r2: 0.254** **일치율: 55.55%**	r2: 0.281 ad r2: 0.232 일치율: 53.9%	r2: 0.281 ad r2: 0.219 일치율: 53.96%	r2: 0.283 ad r2: 0.208 일치율: 52.83%
Lasso	**r2: 0.289** **ad r2: 0.254** **일치율: 55.55%**	r2: 0.281 ad r2: 0.232 일치율: 53.9%	r2: 0.281 ad r2: 0.219 일치율: 53.96%	r2: 0.283 ad r2: 0.208 일치율: 52.38%

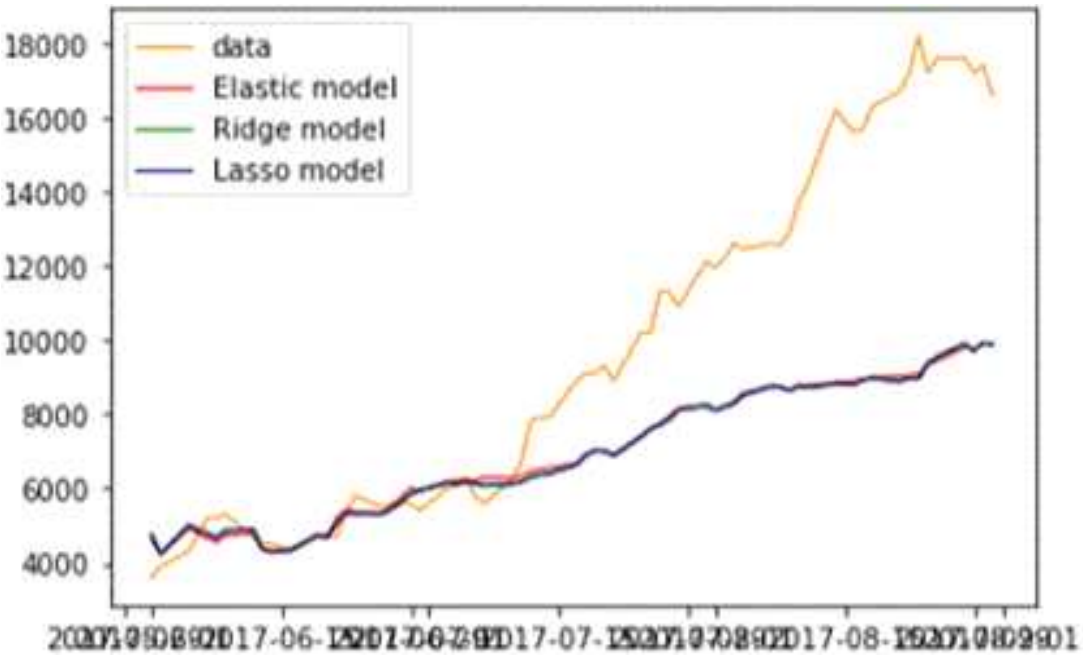
	7	8	9	10
Elastic	r2: 0.265 ad r2: 0.173 일치율: 60.3%	r2: 0.263 ad r2: 0.156 일치율: 57.14%	r2: 0.260 ad r2: 0.130 일치율: 57.14%	r2: 0.260 ad r2: 0.120 일치율: 57.14%
Ridge	r2: 0.259 ad r2: 0.167 일치율: 57.14%	r2: 0.258 ad r2: 0.150 일치율: 53.96%	r2: 0.251 ad r2: 0.126 일치율: 57.14%	r2: 0.259 ad r2: 0.119 일치율: 50.79%
Lasso	r2: 0.259 ad r2: 0.167 일치율: 57.14%	r2: 0.258 ad r2: 0.156 일치율: 53.96%	r2: 0.251 ad r2: 0.126 일치율: 57.14%	r2: 0.259 ad r2: 0.119 일치율: 52.38%

-SVR(kernel=linear), SVR(kernel=rbf) (r2, adjusted r2, 일치율)

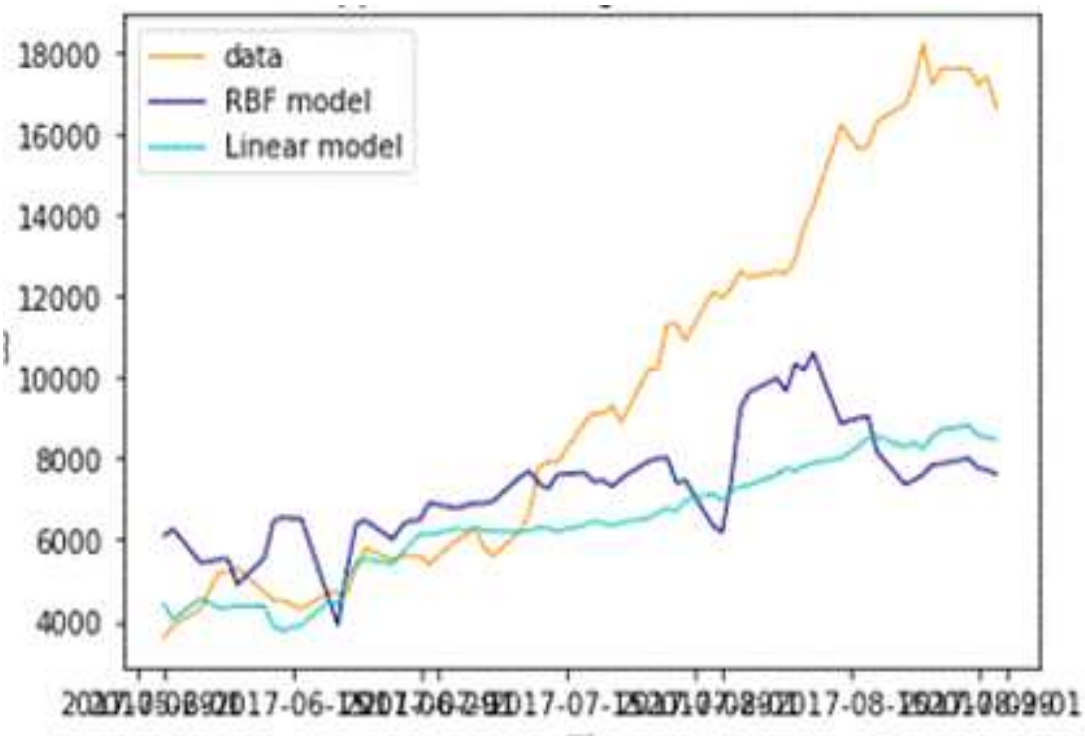
	3	4	5	6
SVR(Linear)	r2: -0.053 ad r2: -0.053 일치율: 55.5%	r2: -0.172 ad r2: -0.172 일치율: 52.3%	**r2: 0.053** **ad r2: 0.023** **일치율: 48.0%**	r2: -0.083 ad r2: -0.083 일치율: 49.2%
SVR(RBF)	r2: 0.068 ad r2: 0.068 일치율: 42.85%	r2: 0.023 ad r2: 0.023 일치율: 46.0%	**r2: 0.025** **ad r2: 0.011** **일치율: 55.6%**	r2: -0.072 ad r2: -0.072 일치율: 44.44%

	7	8	9	10
SVR(Linear)	r2: -0.151 ad r2: -0.151 일치율: 53.96%	r2: -0.058 ad r2: -0.058 일치율: 46.03%	r2: 0.049 ad r2: 0.049 일치율: 46.61%	r2: -0.271 ad r2: -0.271 일치율: 41.26%
SVR(RBF)	r2: -0.299 ad r2: -0.299 일치율: 44.44%	r2: -0.364 ad r2: -0.364 일치율: 41.26%	r2: -0.453 ad r2: -0.453 일치율: 46.04%	r2: -0.519 ad r2: -0.519 일치율: 41.26%

-Lasso, Ridge, Elastic, 결과 (그래프)



-SVR(kernel=linear), SVR(kernel=rbf) (그래프)



4.(2).3 결과 9~11월

-Lasso, Ridge, Elastic, 결과 (r2, adjusted r2, 일치율)

	3	4	5	6
Elastic	r2: 0.549 ad r2: 0.525 일치율: 44.82%	r2: 0.478 ad r2: 0.440 일치율: 48.27%	r2: 0.474 ad r2: 0.424 일치율: 48.27%	r2: 0.475 ad r2: 0.414 일치율: 48.27%
Ridge	r2: 0.534 ad r2: 0.509 일치율: 43.10%	r2: 0.431 ad r2: 0.388 일치율: 48.27%	r2: 0.298 ad r2: 0.231 일치율: 43.10%	r2: 0.352 ad r2: 0.277 일치율: 50.0%
Lasso	r2: 0.534 ad r2: 0.509 일치율: 43.10%	r2: 0.431 ad r2: 0.388 일치율: 48.27%	r2: 0.294 ad r2: 0.227 일치율: 43.10%	r2: 0.349 ad r2: 0.274 일치율: 48.27%

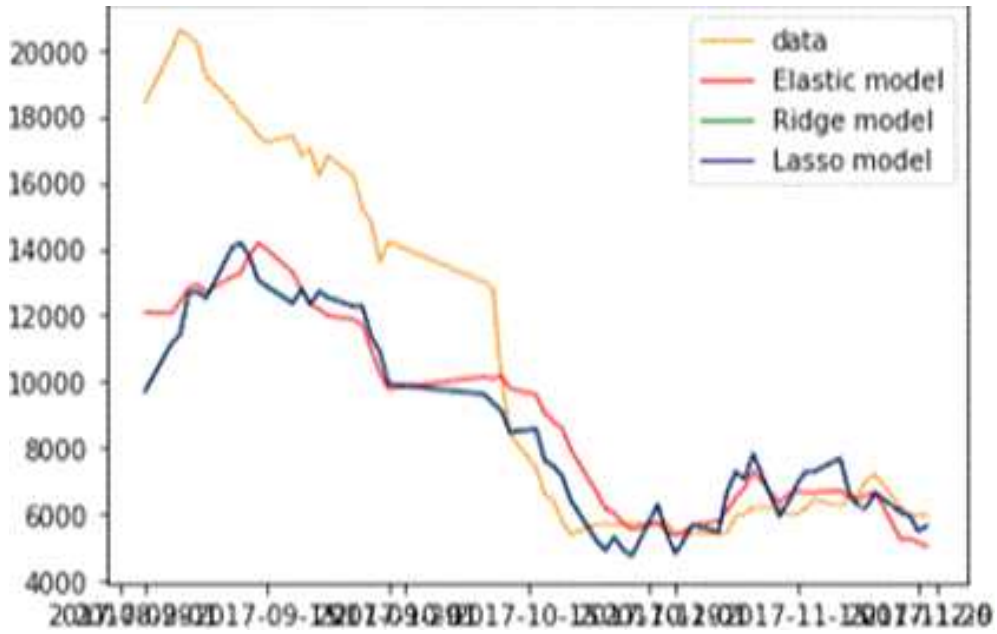
	7	8	9	10
Elastic	**r2: 0.644** **ad r2: 0.592** **일치율: 48.22%**	r2: 0.610 ad r2: 0.548 일치율: 48.27%	r2: 0.570 ad r2: 0.491 일치율: 46.55	r2: 0.570 ad r2: 0.480 일치율: 50.0%
Ridge	**r2: 0.615** **ad r2: 0.583** **일치율: 44.27%**	r2: 0.624 ad r2: 0.563 일치율: 51.72%	r2: 0.568 ad r2: 0.489 일치율: 51.72%	r2: 0.568 ad r2: 0.478 일치율: 51.72%
Lasso	**r2: 0.614** **ad r2: 0.583** **일치율: 44.35%**	r2: 0.624 ad r2: 0.563 일치율: 51.72%	r2: 0.568 ad r2: 0.489 일치율: 51.72%	r2: 0.568 ad r2: 0.478 일치율: 51.72%

-SVR(kernel=linear), SVR(kernel=rbf) (r2, adjusted r2, 일치율)

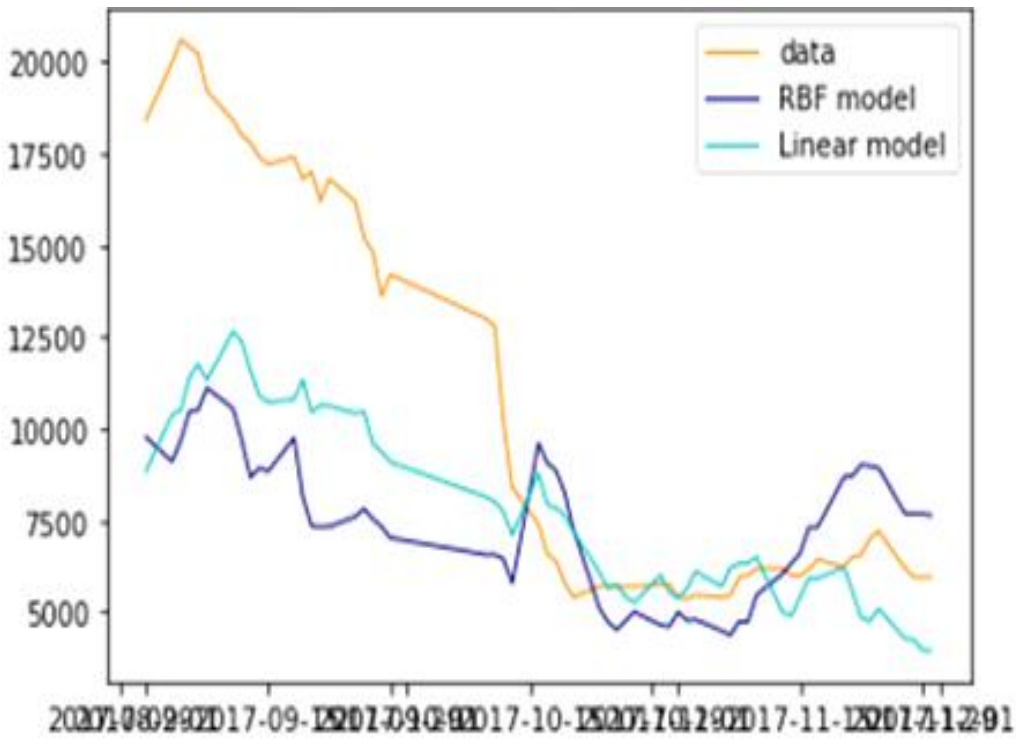
	3	4	5	6
SVR(Linear)	r2: 0.260 ad r2: 0.260 일치율: 43.10%	r2: -0.107 ad r2: -0.107 일치율: 43.10%	r2: -0.250 ad r2: -0.250 일치율: 46.55%	r2: -0.250 ad r2: -0.250 일치율: 48.27%
SVR(RBF)	r2: 0.245 ad r2: 0.245 일치율: 37.93%	r2: -0.035 ad r2: -0.035 일치율: 43.10%	r2: -0.109 ad r2: -0.109 일치율: 41.37%	r2: -0.128 ad r2: -0.128 일치율: 41.37%

	7	8	9	10
SVR(Linear)	r2: 0.109 ad r2: 0.109 일치율: 50.0%	**r2: 0.285** **ad r2: 0.268** **일치율: 55.17%**	r2: 0.159 ad r2: 0.159 일치율: 53.44%	r2: 0.221 ad r2: 0.221 일치율: 55.17%
SVR(RBF)	r2: 0.017 ad r2: 0.017 일치율: 36.20%	**r2: -0.150** **ad r2: -0.175** **일치율: 48.27%**	r2: -0.421 ad r2: -0.421 일치율: 44.82%	r2: -0.473 ad r2: -0.473 일치율: 44.82%

-Lasso, Ridge, Elastic, 결과 (그래프)



-SVR(kernel=linear), SVR(kernel=rbf) (그래프)



4.(2).4 결과 12월~2월

-Lasso, Ridge, Elastic, 결과 (r2, adjusted r2, 일치율)

	3	4	5	6
Elastic	r2: 0.33 ad r2: -0.022 일치율: 42.85%	r2: -0.001 ad r2: -0.078 일치율: 44.64%	r2: -0.001 ad r2: -0.099 일치율: 44.64%	r2: 0.039 ad r2: -0.077 일치율: 46.42%
Ridge	r2: 0.016 ad r2: -0.040 일치율: 42.85%	r2: -0.044 ad r2: -0.125 일치율: 44.64%	r2: -0.100 ad r2: -0.208 일치율: 44.64%	r2: -0.069 ad r2: -0.198 일치율: 42.85%
Lasso	r2: 0.016 ad r2: -0.040 일치율: 42.85%	r2: -0.044 ad r2: -0.125 일치율: 44.64%	r2: -0.100 ad r2: -0.207 일치율: 46.42%	r2: -0.069 ad r2: -0.197 일치율: 42.85%

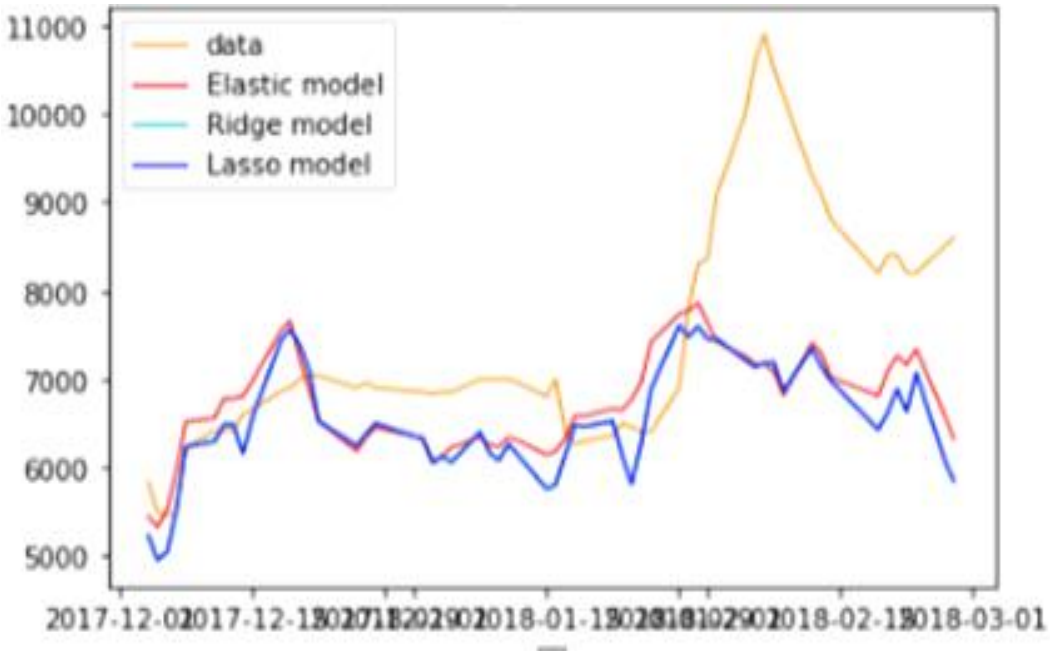
	7	8	9	10
Elastic	r2: 0.060 ad r2: -0.074 일치율: 42.85%	**r2: 0.092 ad r2: -0.149 일치율: 51.78%**	r2: -0.192 ad r2: -0.420 일치율: 44.64%	r2: -0.372 ad r2: -0.670 일치율: 39.28%
Ridge	r2: -0.053 ad r2: -0.203 일치율: 44.64%	**r2: -0.110 ad r2: -0.293 일치율: 51.73%**	r2: -1.320 ad r2: -1.764 일치율: 44.64%	r2: -1.704 ad r2: -2.292 일치율: 42.85%
Lasso	r2: -0.053 ad r2: -0.203 일치율: 44.64%	**r2: -0.110 ad r2: -0.295 일치율: 51.76%**	r2: -1.329 ad r2: -1.775 일치율: 44.64%	r2: -1.711 ad r2: -2.300 일치율: 41.07%

-SVR(kernel=linear), SVR(kernel=rbf) (r2, adjusted r2, 일치율)

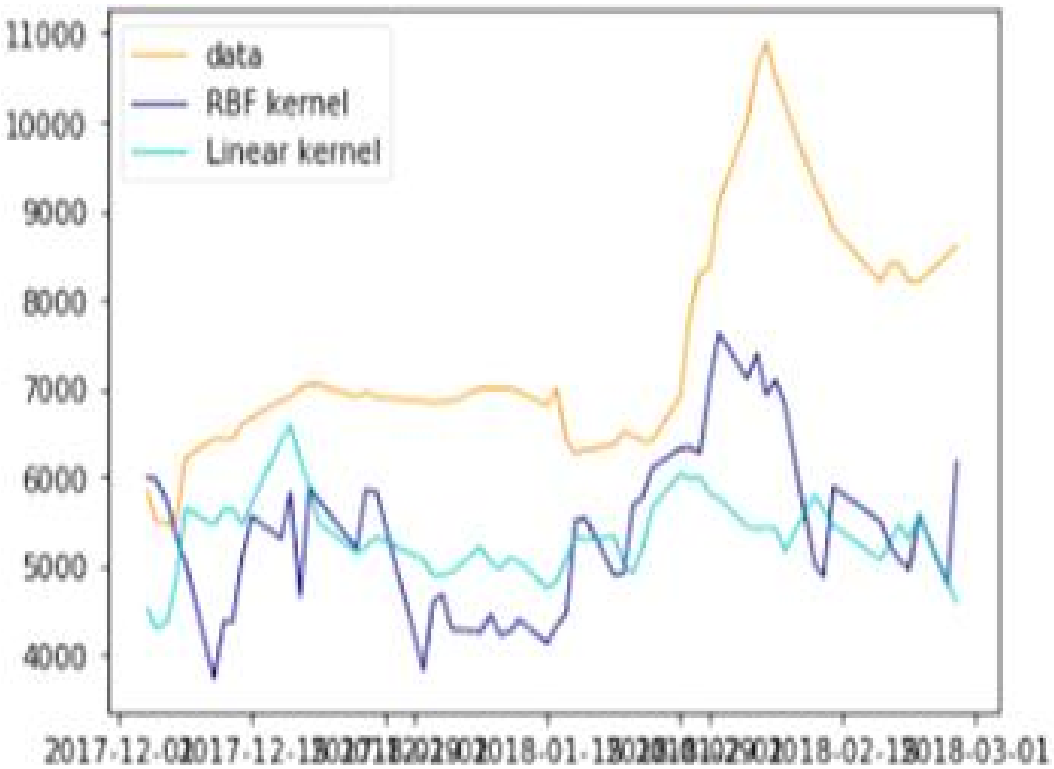
	3	4	5	6
SVR(Linear)	r2: -2.707 ad r2: -2.917 일치율: 41.07%	r2: -2.874 ad r2:-3.172 일치율: 42.85%	r2: -2.629 ad r2: -2.985 일치율: 44.64%	r2: -2.047 ad r2: -2.413 일치율: 46.42%
SVR(RBF)	r2: -3.567 ad r2: -3.826 일치율: 28.57%	r2: -1.951 ad r2: -2.178 일치율: 33.92%	r2: -1.500 ad r2: -1.745 일치율: 42.85%	r2: -3.232 ad r2: -3.740 일치율: 46.42%

	7	8	9	10
SVR(Linear)	r2: -2.331 ad r2: -2.807 일치율: 46.42%	**r2: -2.483 ad r2: -3.063 일치율: 50.0%**	r2: -1.760 ad r2: -2.289 일치율: 44.64%	r2: -3.106 ad r2: -3.999 일치율: 41.07%
SVR(RBF)	r2: -2.318 ad r2: -2.792 일치율: 44.64%	**r2: -2.097 ad r2: -2.613 일치율: 48.21%**	r2: -2.264 ad r2: -2.889 일치율: 44.64%	r2: -1.992 ad r2: -2.642 일치율: 44.64%

-Lasso, Ridge, Elastic, 결과 (그래프)



-SVR(kernel=linear), SVR(kernel=rbf) (그래프)



5. 요약(Summary)

본 팀은 배추의 가격을 예측하는 모델을 만들기 위해 설계와 구현을 진행하였다. 가격을 예측하는 요소에는 기상청을 통해 해남지역의 2008년부터 2017년까지의 날씨 데이터를 얻어왔으며, 가격은 농수산물유통정보 사이트에서 배추의 전국 평균 도매가격을 얻어왔다. 두 개의 데이터를 통합하고 Missing Value를 처리하며 Outlier를 제거하는 등, 필요한 정제과정을 거쳤다. 배추의 모종부터 출하까지 평균 4개월 정도의 시간이 걸리는 것을 고려하여, 예측하고자 하는 날짜의 4개월 전의 날짜부터 2주 동안의 각 날씨 요소의 중앙값(Median)을 feature로 사용하였다. Scikit-learn library에 있는 RFE (Recursive Feature Elimination)를 사용하여, feature의 개수가 3개~10개가 될 때 어떤 feature들이 가장 관계가 높은지 각 경우에 대해서 구해주었다. 그런 후에 Cross-Validation을 통하여 최적의 모델 파라미터와 feature갯수를 정하여 모델을 선정하였다. 2008년부터 2016년까지의 데이터는 학습하는 용도로 사용하고, 2017년 데이터를 성능을 테스트하는데 사용하여 1년 단위로 Regression을 시행해 보았다. 그런 후에 더 좋은 성능을 내기 위하여 Overfitting을 방지하기 위한 Lasso, Ridge, Elastic Net 방식을 모델에 적용하여 각 경우의 성능을 측정해 보았다. 또한 부가적으로 가격 변동에 대한 일치율도 계산해 보았다.

예상보다 높은 성능이 나오지 않아 1년 동안 일정 기간에 따라 가격에 영향을 미치는 feature가 다르다고 생각하고, 1년을 3개월씩 4개의 구간으로 나누어 4개의 모델로 만들어 보기로 하였다. 최적의 구간 분할을 위하여 기존의 데이터를 1월~3월, 2월~4월, … , 10월~12월까지로 데이터를 새롭게 정리하고, 앞서 사용한 방법처럼 RFE를 통해 각 구간에서 feature의 개수마다의 상관계수가 높은 feature들을 정해 주었다. Linear Regression과 SVR(Support Vector Regression) 모델을 사용하여 성능을 측정하고 기록하였다. 마찬가지로 가격 변동에 대한 일치율도 계산해 보았다.

성능을 측정함으로써 알 수 있었던 점은, 1년 단위로 Regression을 하는 것보다 3개월 간격으로 데이터를 나누어 Regression 하는 것이 더 높은 성능을 보였다는 것이다. 다만, 9~12월의 결과가 좋지 않은 것은 Test set이 기존의 9~12월의 기온과 가격간의 관계와 다르게 높았기 때문으로 해석된다. 기존의 또한 배추 가격을 결정짓는 요소에 날씨 외에도 여러 변인들이 작용하기 때문에, 날씨에 대한 데이터로 배추의 가격이 아닌 생산량을 예측하였다면 조금 더 밀접한 상관관계와 정확도를 얻어낼 수 있었을 것이다.

6. 전망 및 기대효과

이번 프로젝트를 진행하면서, 배추의 날씨 데이터와 더불어 전년도의 생산량이나 유통가격, 수입량 등의 경제적인 요소를 추가한다면 배추의 가격을 예측하는 것이 충분히 가능하다는 것을 알 수 있었다. 배추의 가격을 예측할 수 있다면 이는 상당한 의미와 실용가치가 존재한다. 일반적으로 배추는 수확이 되면 각 지방의 농협에 판매되어 축적된다. 농협은 매일 배추를 얼마나 출하할지 정하여 물량을 조절하게 되는데, 만약 다음날 배추 가격 변동을 예측할 수 있다면 이는 출하량을 조절 하는데 큰 도움이 될 것이다. 농협뿐 아니라 정부에서도 배추를 대량으로 구매하여 비축하거나, 필요할 때 물량을 풀어 공급에 도움을 주는데, 가격 변동이 미리 예상된다면 정부는 상황을 미리 대처 하는게 가능해진다. 이는 모두 물가 폭등이나 등락 현상을 방지하여 사람들이 배추를 사는데 문제가 없도록 하기 위함이므로, 이러한 가격 예측 모델의 실용가치가 뛰어나다고 할 수 있다.

이 모델은 확장성 또한 뛰어나다. 배추뿐 아니라 대부분의 농산물이 날씨에 영향을 받기 때문에 앞서 구현하는 과정에서처럼 재배기간 등을 고려하여 데이터를 조금씩만 수정해준다면, 날씨 데이터를 재사용하여 얼마든지 양파나 상추, 시금치, 사과 등의 수요가 높은 작물이나 과일의 가격도 예측 가능하다. 이러한 방식으로 확장이 가능하다면, 농협이나 농산물에 직결된 기관 등에서의 높은 사용률을 기대해 볼 수 있다.