회귀분석을 이용한
배추 가격 변동 예측

2018-1 Datamining & Machine Learning

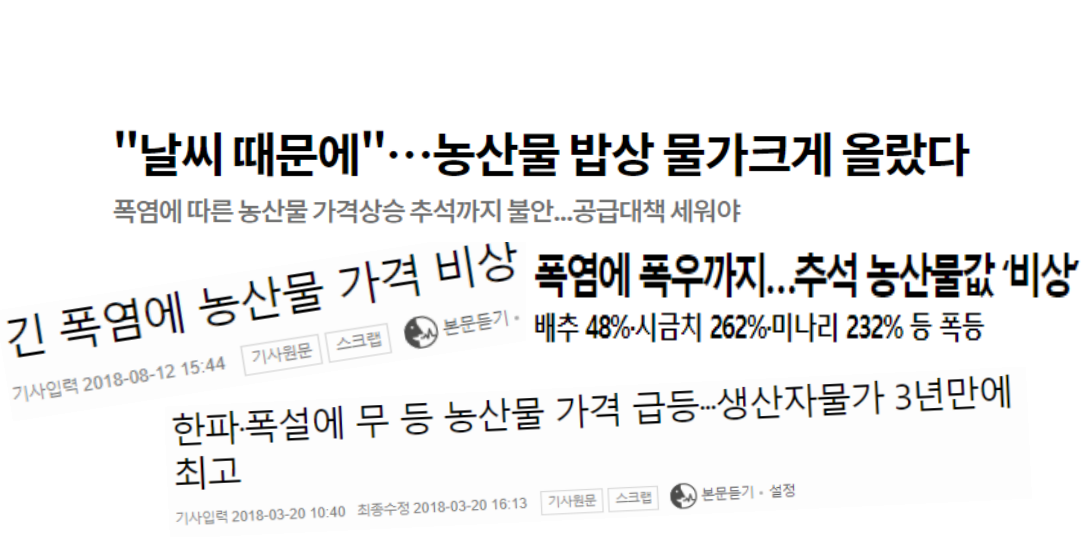
201420874 소프트웨어 이성훈
201420907 소프트웨어 안우일
201420969 소프트웨어 김영운
201421120 소프트웨어 김필선

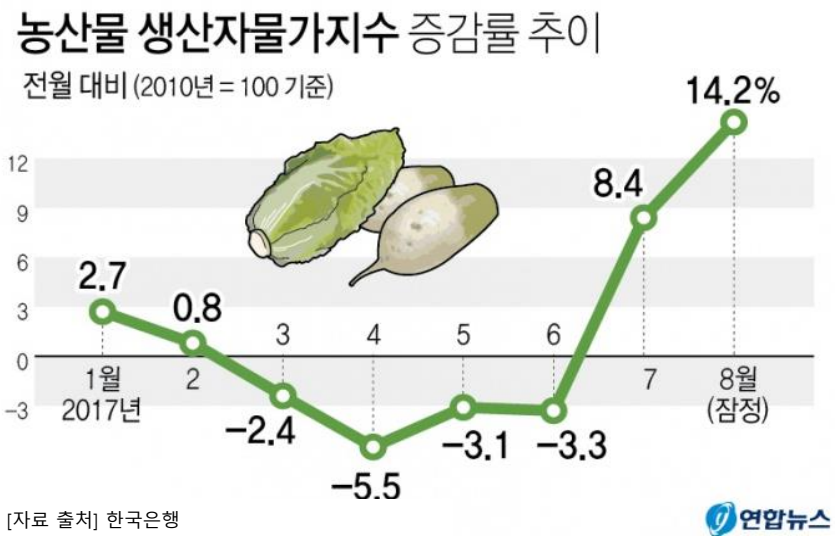
Contents

목차

Motivation

01 주제 선정 동기





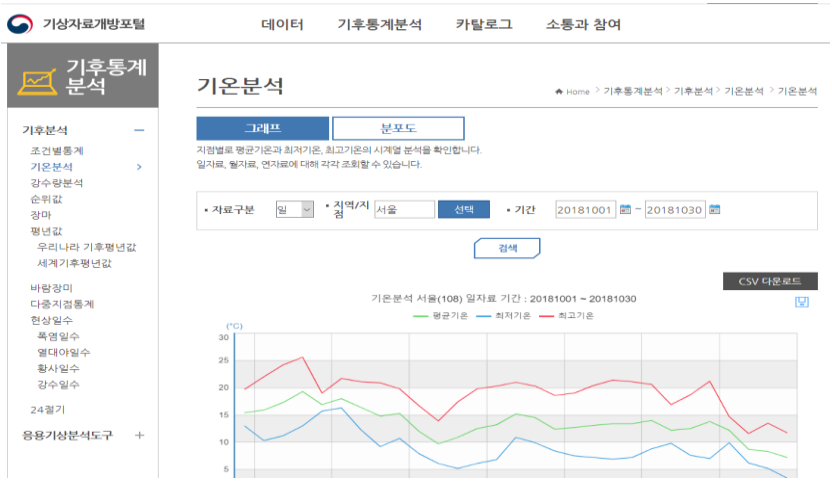
농산물의 공급량, 가격은 기상 조건에 의해 많은 영향을 받고, 그 영향으로 인해 수요량까지 영향을 받게 됨

→ Machine Learning을 이용해 농산물 가격을 예측하여 변화에 대비 할 수 있다!

농산물 가격의 등락폭을 예측해 해당 년도의 농산물 공급량을 안정적으로 유지

Data Processing

02 데이터 전처리



기상청

기상데이터

2008년 ~ 2018년 Data

https://data.kma.go.kr/cmmn/main.do

aTKAMIS

각종 농산물 가격

2008년 ~ 2018년 Data

https://www.kamis.or.kr/customer/main/main.do

Data Processing

02 데이터 전처리

01. 기상청에서 쓰는 데이터의 종류(2008~2018年)

평균기온(˚C)	최저기온(˚C)	최고기온(˚C)	일 강수량(mm)	평균 이슬점온도(˚C)
최대 풍속(m/s)	최대 순간 풍속(m/s)	평균 풍속(m/s)	최소 상대습도(%)	평균 상대습도(%)
평균 증기압(hPa)	평균 현지기압(hPa)	최고 해면기압(hPa)	최저 해면기압(hPa)	평균 해면기압(hPa)
합계 일조 시간(hr)	평균 지면온도(˚C)	최저 초상온도(˚C)	일 최심신적설(cm)	일 최심적설(cm)

02. 유통센터에서 쓰는 데이터의 종류 (2008~2018年)

▪ (도매가격 - 기간별) 채소류 / 배추 / 전체, 등급 : 상품, 단위 : 10kg

(단위 : 원) 데이터저장

구분	11/22	11/23	11/26	11/27	11/28	11/29	11/30
평균	7,000	7,500	7,000	6,400	6,600	6,600	6,800

배추는 모종을 한 뒤 초기 2주의 기간이 중요하다. 모종부터 판매까지 걸리는 시간은 4달

Y-label	X-label(feature)
가격	4달 전 2주 기상 데이터의 Mean OR Median 값

2008.01.02부터 2018.02.28까지 총 2298개의 데이터를 사용(휴일 제외)

Train 2008.01 ~ 2016.12년 data 사용

Test 2017.01 ~2018.02년 data 사용

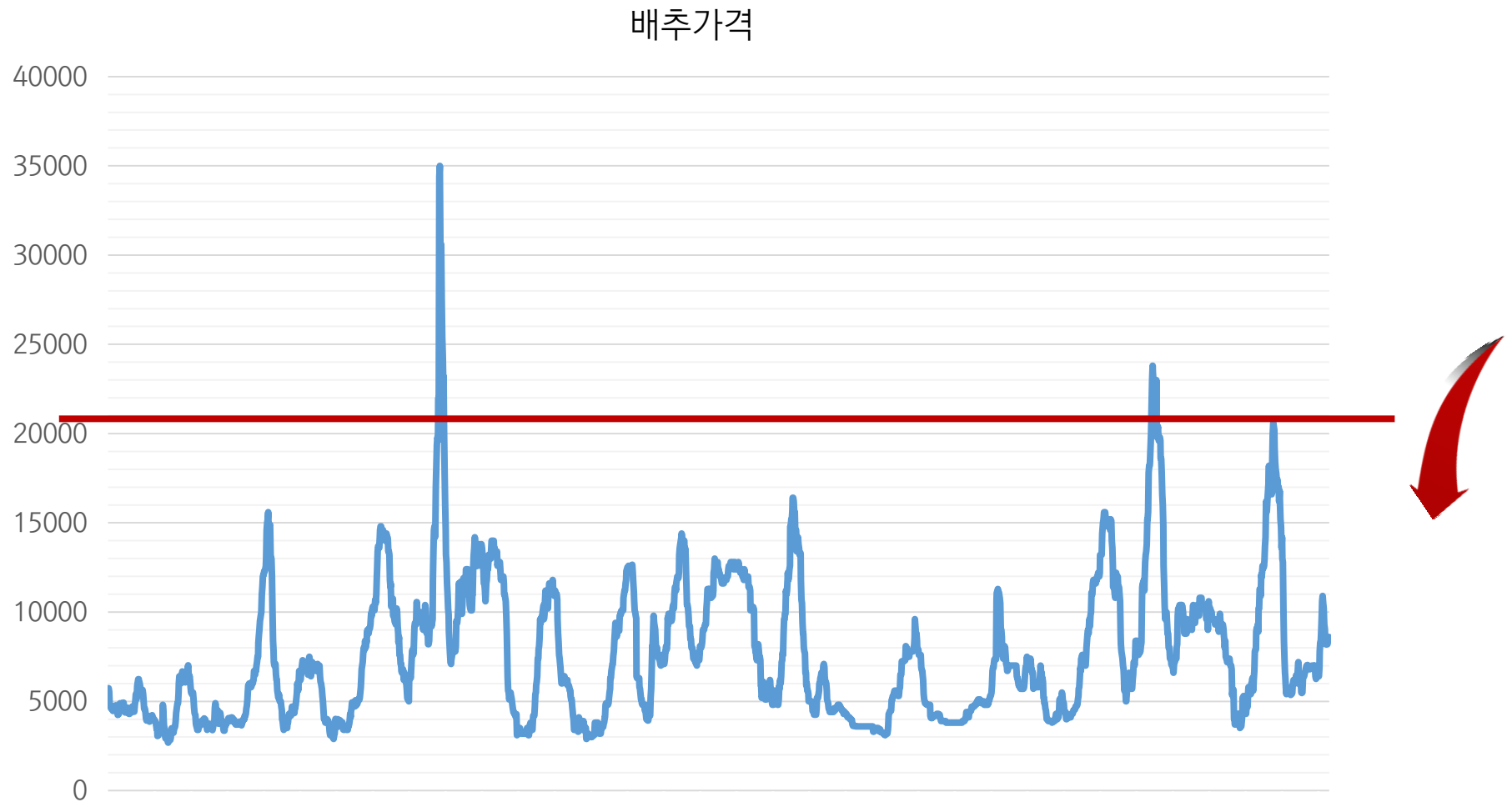
Data Processing

02 데이터 전처리

- Missing value들은 이전 날짜의 값을 채택. (fillna(method= 'ffill', inplace=true))
- 거의 모든 날이 0값인 일 최심신적설(cm)과 일 최심적설(cm)은 제외
- 각 Feature마다 크기의 편차가 커서 Feature 별로 정규분포화를 시행

$$Z = \frac{X - \mu}{\sigma}$$

Threshold를 20,000원으로 하여 Outlier를 제거



03 구현 및 결과

Regression을 위해 사용한 기법

(1) Linear Regression

(2) Lasso Regression

(3) Ridge Regression

(4) Elastic-Net Regression

(5) Support Vector Regression(SVR)

- kernel : linear, RBF, poly

Evaluation

(1) (adjusted) R Square score

$$\bar{R}^2 = \left(R^2 - \frac{p}{n-1}\right)\left(\frac{n-1}{n-p-1}\right)$$

(2) 상승 하락 예측

03 구현 및 결과

1년 단위의 Feature Selection을 위해 Scikit-learn의 RFE를 사용 (Backward 방식)

Recursive Feature Elimination

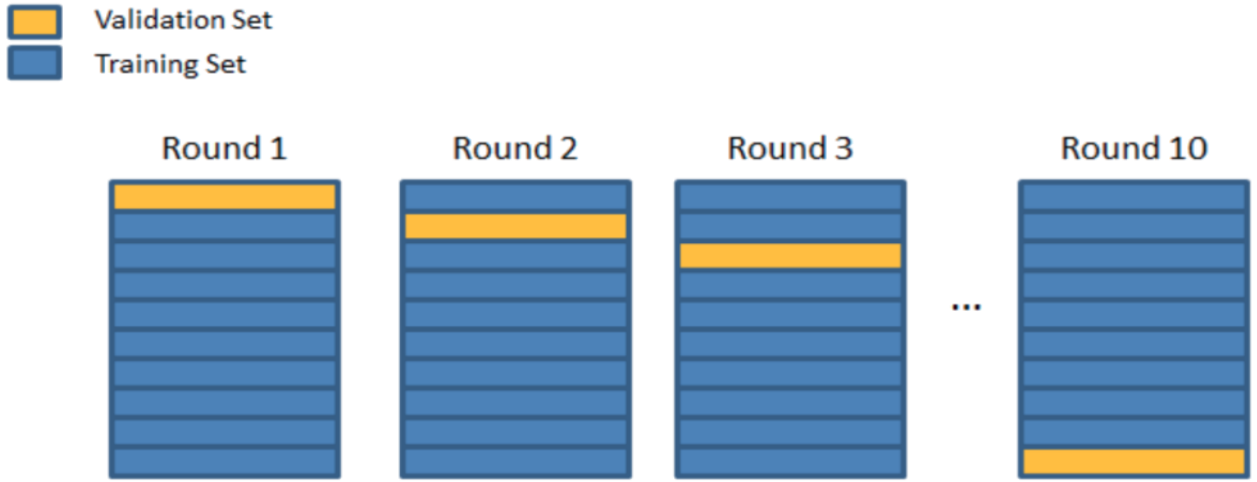
3개	[평균온도, 최고풍속, 이슬점온도]
4개	[평균온도, 최고온도, 최고풍속, 일조량]
5개	[평균온도, 최저기온, 평균 증기압, 일 강수량, 최저 초상온도]
.....	
8개	[평균온도, 최고온도, 최저온도, …, 이슬점온도, 평균 풍속, 최저 초상온도]
9개	[평균온도, 최고온도, 최저온도, … , 이슬점온도, 평균 풍속, 최고풍속, 최대 순간 풍속]
10개	[평균온도, 최고풍속, 이슬점온도, 평균 상대습도, …, 최고풍속, 최대 순간 풍속, 최고 초상온도]

→ RFE를 사용하여 최대로 연관 있는 Feature 3~10개를 Select

03 구현 및 결과

- CV를 통해 Lasso/Ridge/Elastic α와 SVR의 γ 에 따른 변화를 측정함.

- α와 γ 에 따른 변화가 크지 않아 default값인 1.0으로 고정하여 사용함



03 구현 및 결과

1년 Feature Testing

(ex)

	3개	4개	…	8개	9개	10개
Ridge	0.234	0.761	…	0.763	0.125	0.432
Lasso	0.344	0.542	…	0.234	0.235	0.412
Elastic	0.127	0.664	…	0.563	0.174	0.653
SVR(Linear)	0.431	0.454	…	0.903	0.127	0.555
SVR(RBF)	0.332	-0.234	…	0.546	0.165	0.322

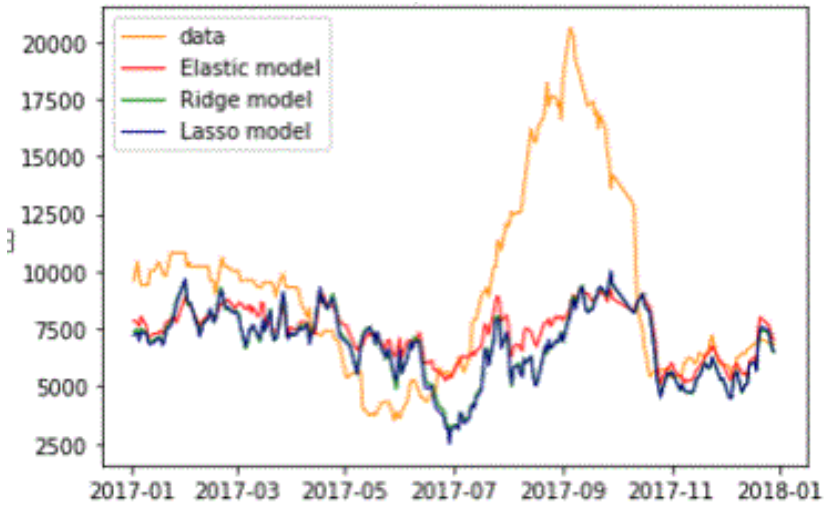
→ (Adjusted) R^2를 사용하여 가장 결과가 좋은 Feature 수를 찾음

Implementation & Result

03 구현 및 결과

1년 단위로 Regression을 시행

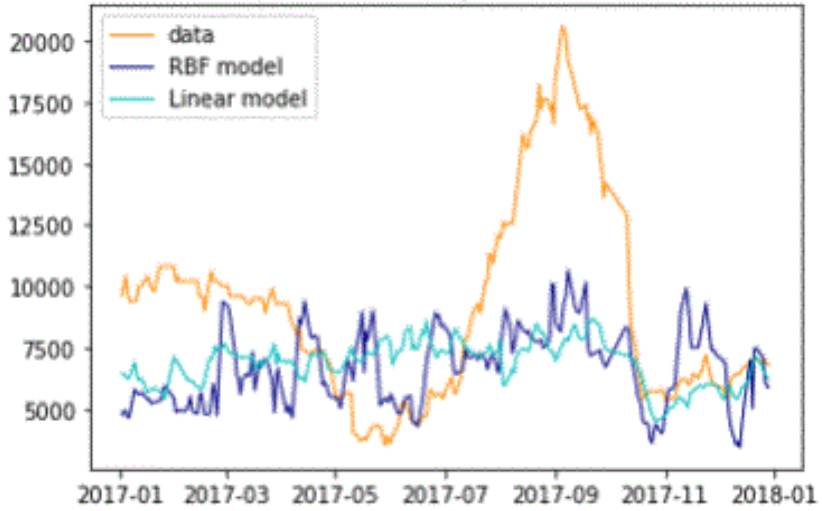
Lasso, Ridge, Elastic Net (Median)



사용 feature 개수 : 10개

Regression	상승/하락	R^2	Adj R^2
Ridge	35.68%	0.101	0.062
Lasso	35.68%	0.097	0.058
Elastic Net	39%	0.115	0.076

SVR(kernel = linear, RBF) (Median)



사용 feature 개수 : 10개

Regression	상승/하락	R^2	Adj R^2
SVR(RBF)	35.5%	0.073	0.028
SVR(Linear)	36%	0.022	0.022

03 구현 및 결과

3개월 간격으로 Feature Selection을 위해 Scikit-learn의 RFE를 사용 (Backward 방식)

Recursive Feature Elimination

1월 ~ 3월	3개 : [평균온도, 최고풍속, 이슬점온도] …, 10개 : [평균온도. … , 일 강수량]
2월 ~ 4월	3개 : [평균온도, 최고풍속, 일조량] …, 10개 : [평균온도. … , 일조량]
3월 ~ 5월	3개 : [평균온도, 최저기온, 평균 증기압] …, 10개 : [평균온도. … , 일 강수량]
.....	
10월 ~ 12월	3개 : [평균온도, 최고풍속, 이슬점온도] …, 10개 : [평균온도. … , 평균 풍속]
11월 ~ 1월	3개 : [평균온도, 최고풍속, 최대 순간 풍속] …, 10개 : [평균온도. … , 최고 기온]
12월 ~ 2월	3개 : [평균온도, 최고풍속, 이슬점온도] …, 10개 : [평균온도. … , 평균 상대습도]

→ 각 분기마다 RFE를 사용하여 최대로 연관 있는 Feature 3~10개를 Select

03 구현 및 결과

3개월 Feature Testing (각 개월마다 시행)

(ex)

	3개	4개	…	8개	9개	10개
Ridge	0.234	0.761	…	0.763	0.125	0.432
Lasso	0.344	0.542	…	0.234	0.235	0.412
Elastic	0.127	0.664	…	0.563	0.174	0.653
SVR(Linear)	0.431	0.454	…	0.903	0.127	0.555
SVR(RBF)	0.332	-0.234	…	0.546	0.165	0.322

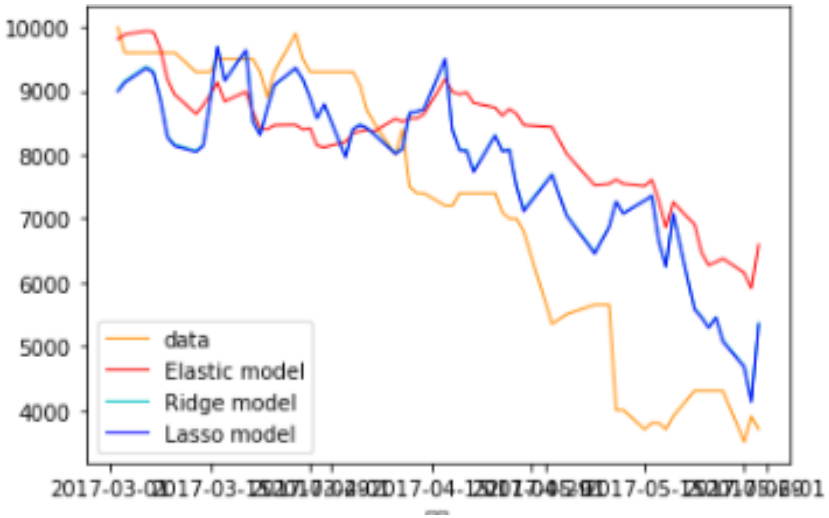
→ (Adjusted) R^2를 사용하여 가장 결과가 좋은 Feature 수를 찾음

Implementation & Result

03 구현 및 결과

3개월 단위로 Regression을 시행(3~5월)

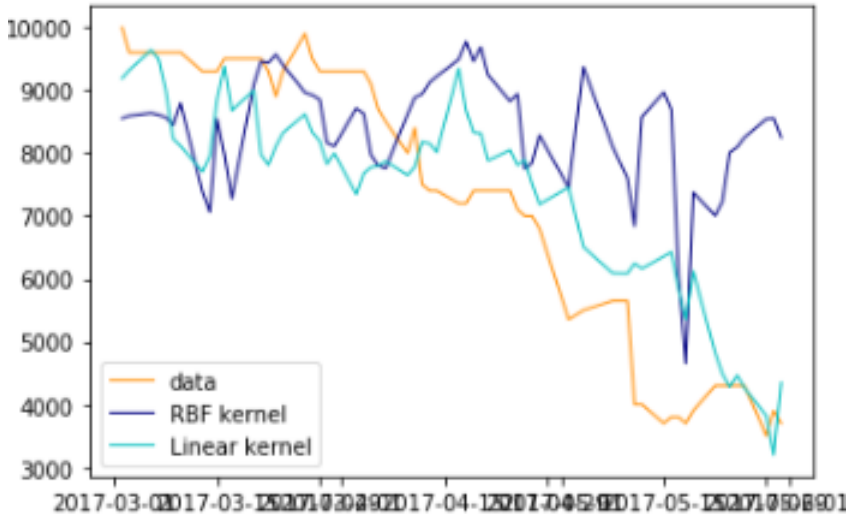
Lasso, Ridge, Elastic Net (Median)



사용 feature 개수 : 9개

Regression	상승/하락	R^2	Adj R^2
Ridge	28%	0.596	0.524
Lasso	28.3%	0.597	0.526
Elastic Net	26.7%	0.338	0.221

SVR(kernel = linear, RBF) (Median)



사용 feature 개수 : 9개

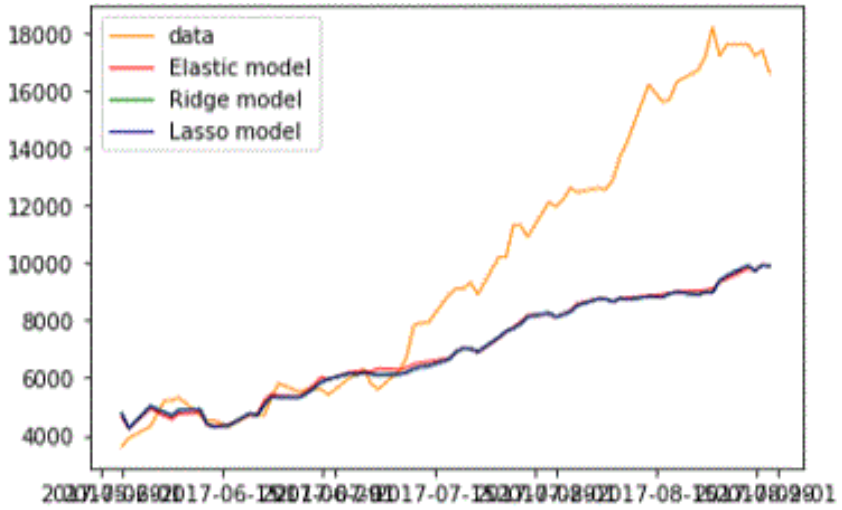
Regression	상승/하락	R^2	Adj R^2
SVR(RBF)	26.6%	-0.290	-0.290
SVR(Linear)	26.7%	0.650	0.650

Implementation & Result

03 구현 및 결과

3개월 단위로 Regression을 시행(6~8월)

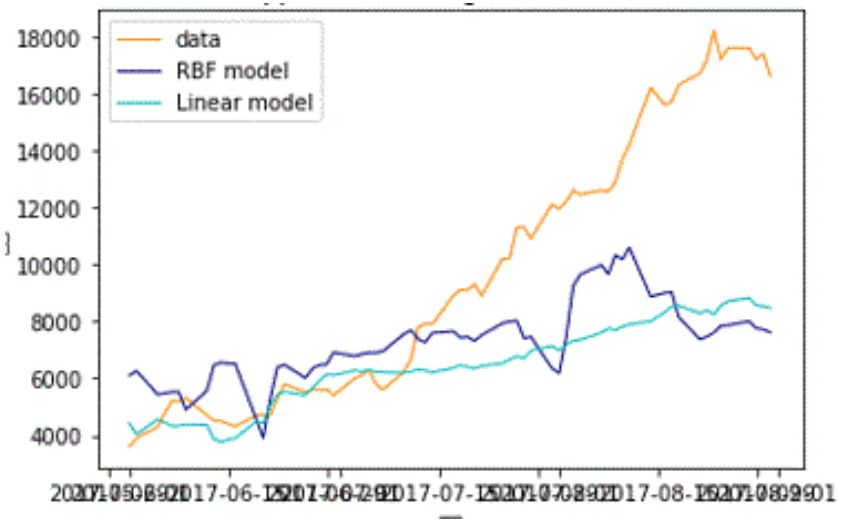
Lasso, Ridge, Elastic Net (Median)



사용 feature 개수 : 3개

Regression	상승/하락	R^2	Adj R^2
Ridge	55.55%	0.289	0.254
Lasso	55.55%	0.289	0.254
Elastic Net	53.96%	0.292	0.257

SVR(kernel = linear, RBF) (Median)



사용 feature 개수 : 5개

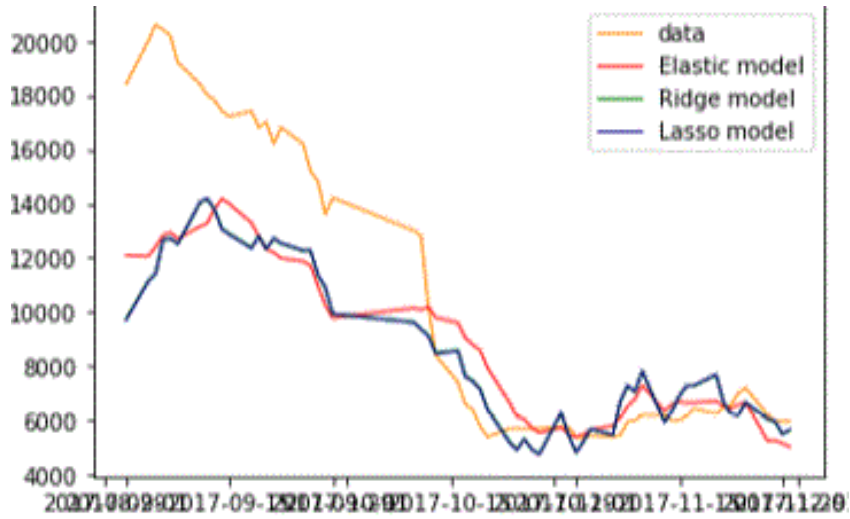
Regression	상승/하락	R^2	Adj R^2
SVR(RBF)	55.6%	0.025	0.011
SVR(Linear)	48.0%	0.053	0.023

Implementation & Result

03 구현 및 결과

3개월 단위로 Regression을 시행(9~11월)

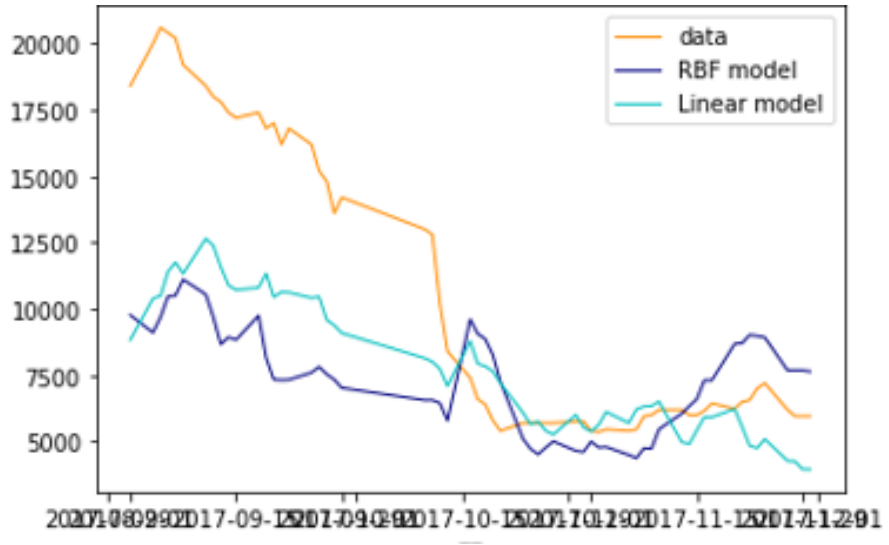
Lasso, Ridge, Elastic Net (Median)



사용 feature 개수 : 7개

Regression	상승/하락	R^2	Adj R^2
Ridge	44%	0.615	0.583
Lasso	44%	0.614	0.583
Elastic Net	48%	0.644	0.592

SVR(kernel = linear, RBF) (Median)



사용 feature 개수 : 8개

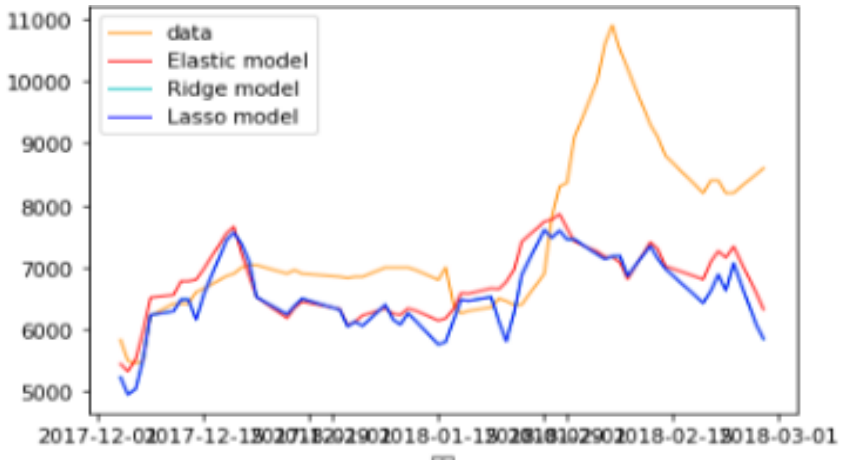
Regression	상승/하락	R^2	Adj R^2
SVR(RBF)	48%	-0.150	-0.175
SVR(Linear)	55%	0.285	0.268

Implementation & Result

03 구현 및 결과

3개월 단위로 Regression을 시행(12~2월)

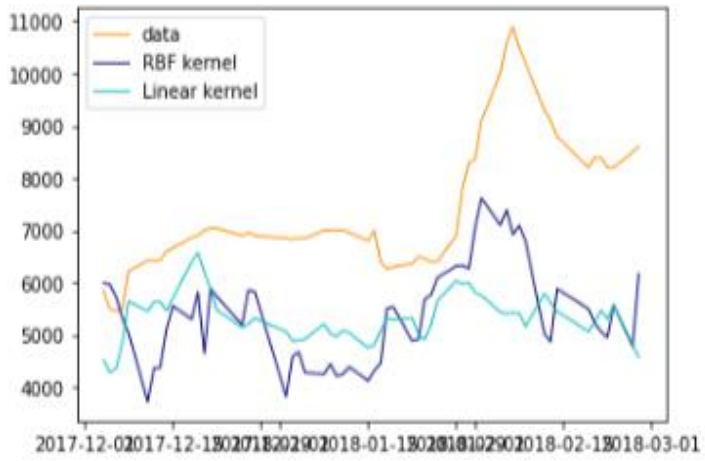
Lasso, Ridge, Elastic Net (Median)



사용 feature 개수 : 8개

Regression	상승/하락	R^2	Adj R^2
Ridge	51.73%	-0.110	-0.293
Lasso	51.76%	-0.110	-0.295
Elastic Net	51.78%	0.092	-0.149

SVR(kernel = linear, RBF) (Median)

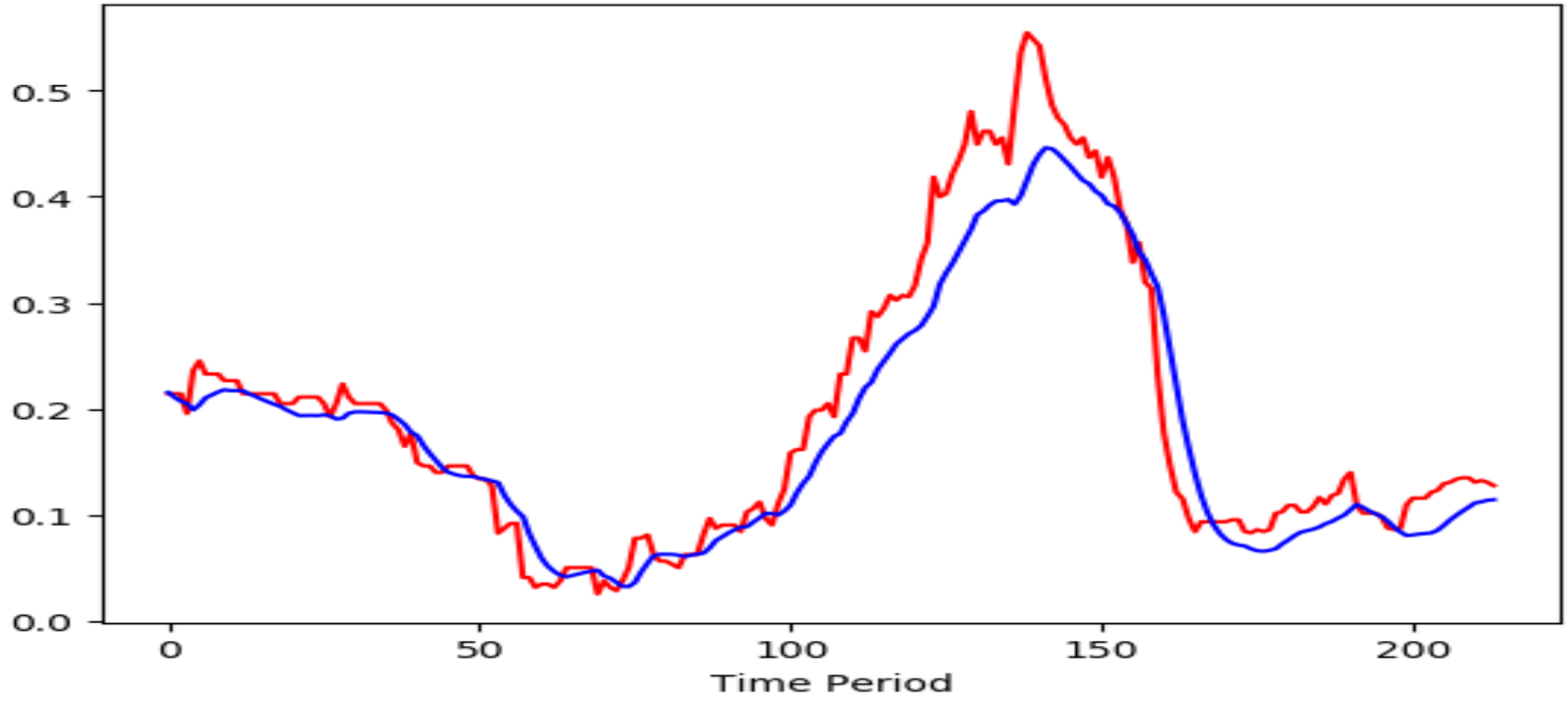


사용 feature 개수 : 8개

Regression	상승/하락	R^2	Adj R^2
SVR(RBF)	48%	-2.097	-2.613
SVR(Linear)	50%	-2.483	-3.603

03 구현 및 결과

RNN 결과 - epoch 100

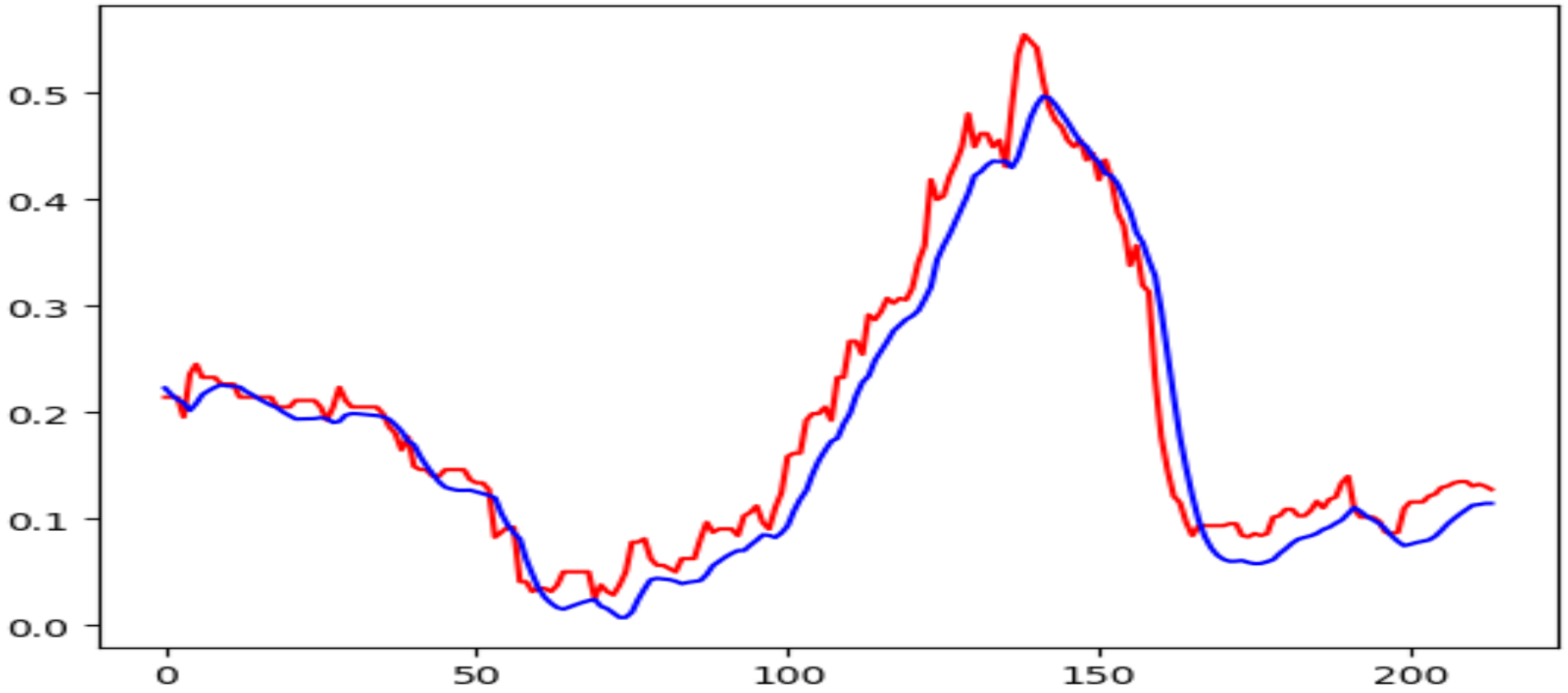


ML lab12-6: RNN with Time Series Data

https://www.youtube.com/watch?v=odMGK7pwTqY&t=106s

03 구현 및 결과

RNN 결과 - epoch 200

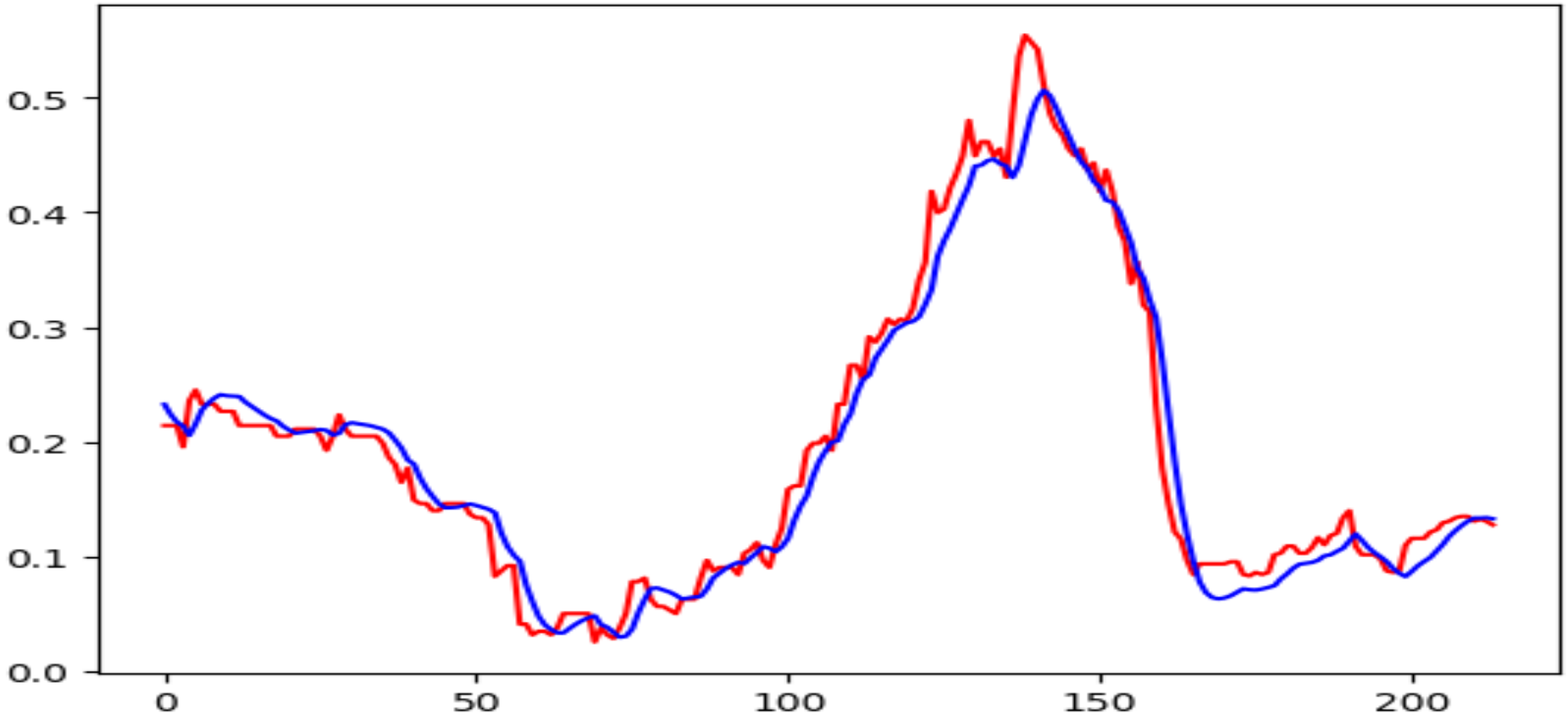


ML lab12-6: RNN with Time Series Data

https://www.youtube.com/watch?v=odMGK7pwTqY&t=106s

03 구현 및 결과

RNN 결과 – epoch 300

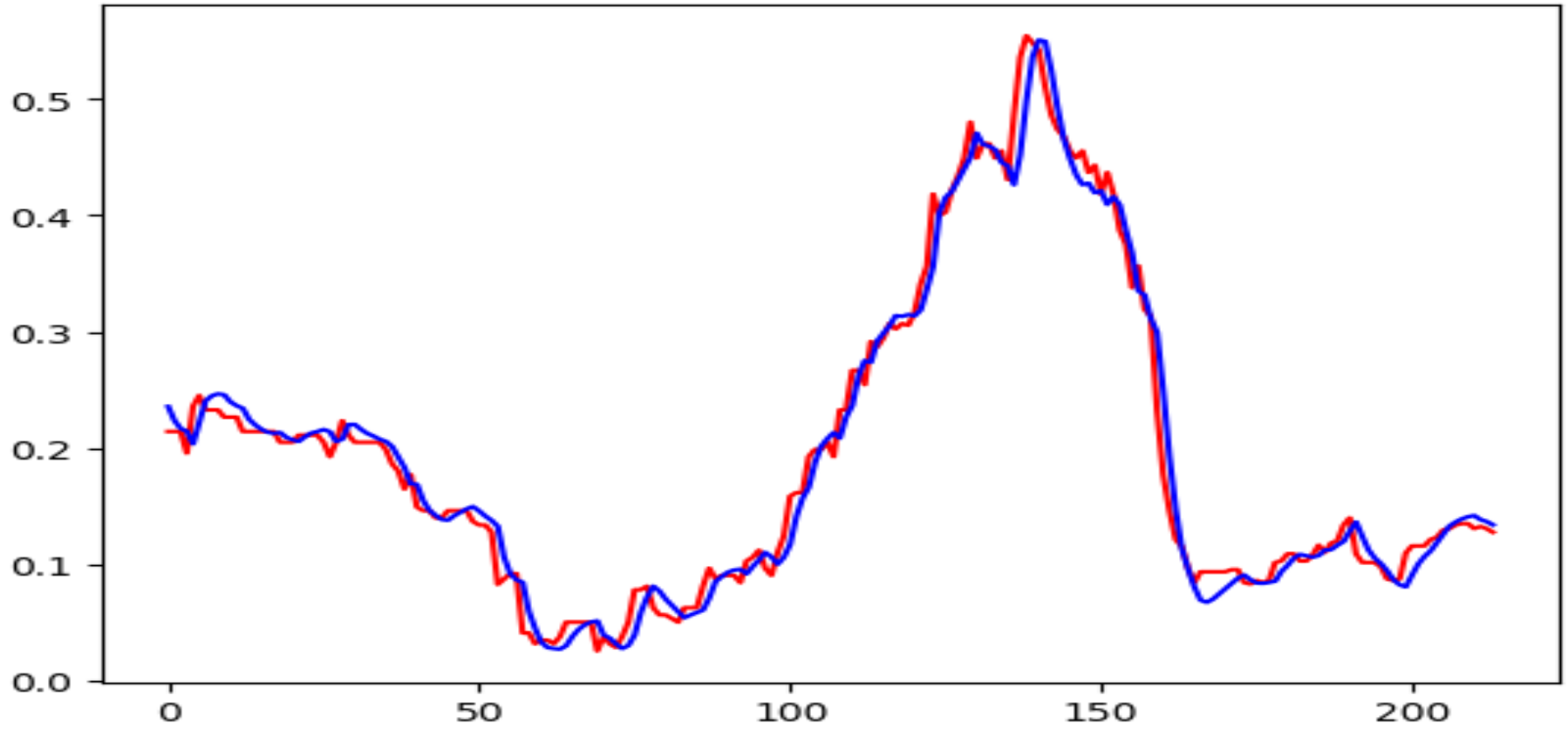


ML lab12-6: RNN with Time Series Data

https://www.youtube.com/watch?v=odMGK7pwTqY&t=106s

03 구현 및 결과

RNN 결과 – epoch 400



ML lab12-6: RNN with Time Series Data

https://www.youtube.com/watch?v=odMGK7pwTqY&t=106s

Implementation & Result

03 구현 및 결과

01

1년 전체의 데이터를
한번에 회귀하는 것보다
3개월 간격으로 데이터를 나눠
회귀할 때, 더 정확했다.

02

농산물 가격변동을 예측할 때
Mean값보다는 Median 값을
사용해서 회귀를 하는 것이
더 정확했다.

03

Max R square Score : 0.644
(9월 ~ 11월 Elastic Net 사용 시)

Max adj R square Score : 0.592
(9월 ~ 11월 Elastic Net 사용 시)

Max predict Up/Down : 60%
(6월 ~ 8월 Ridge/Lasso 사용 시)

Challenge

04 개선사항

01

Feature Selection을
다른 방법을 통해서
해보면 정확도가 올라갈 수 있다.

예) Exhaustive, Forward

02

기상청에 존재하지 않는
다른 요인들을 섞는다면
정확도가 올라갈 수 있다.

예) 물가변동률, 생산량, 태풍여부 등

THANK YOU

Q&A