2021 농산물 가격예측 AI 경진대회

농산물 가격 예측 모형 개발

Team | 가온
팀 장 | 정성문
팀 원 | 김세상, 박민규, 서정인, 정유진

CONTENTS

EDA(탐색적 데이터 분석)와 기본 모델을 만들 때 사용할 데이터 프레임

Train과 Test를 합친 데이터 프레임으로,
요일 정보를 원-핫 인코딩(One-Hot Encoding)으로 병합

2016-01-01부터 2020-11-04까지의
농산물 거래량과 가격으로 구성

	date	요일	배추_거래량(kg)	배추_가격(원/kg)	무_거래량(kg)	무_가격(원/kg)	...	캠벨얼리_거래량(kg)	캠벨얼리_가격(원/kg)	샤인마스캇_거래량(kg)	샤인마스캇_가격(원/kg)	금요일	목요일	수요일	월요일	일요일	토요일	화요일
0	2016-01-01	금요일	0.0	0.0	0.0	0.0		0.0	0.0	0.0	0.0	1	0	0	0	0	0	0
1	2016-01-02	토요일	80860.0	329.0	80272.0	360.0		880.0	2014.0	0.0	0.0	0	0	0	0	0	1	0
2	2016-01-03	일요일	0.0	0.0	0.0	0.0		0.0	0.0	0.0	0.0	0	0	0	0	1	0	0
3	2016-01-04	월요일	1422742.5	478.0	1699653.7	382.0		2703.8	3885.0	0.0	0.0	0	0	0	1	0	0	0
4	2016-01-05	화요일	1167241.0	442.0	1423482.3	422.0		8810.0	2853.0	0.0	0.0	0	0	0	0	0	0	1
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
1765	2020-10-31	토요일	1472058.7	453.0	1966852.1	426.0		34392.5	2920.0	111721.4	9735.0	0	0	0	0	0	1	0
1766	2020-11-01	일요일	0.0	0.0	0.0	0.0		0.0	0.0	0.0	0.0	0	0	0	0	1	0	0
1767	2020-11-02	월요일	1792408.9	441.0	1990362.1	496.0		37043.4	3039.0	88354.3	10205.0	0	0	0	1	0	0	0
1768	2020-11-03	화요일	2015926.5	478.0	2387536.5	465.0		30158.5	3153.0	84795.0	10322.0	0	0	0	0	0	0	1
1769	2020-11-04	수요일	1884530.8	437.0	2637847.2	457.0		26930.0	3171.0	74970.5	10178.0	0	0	1	0	0	0	0

770 rows × 51 columns

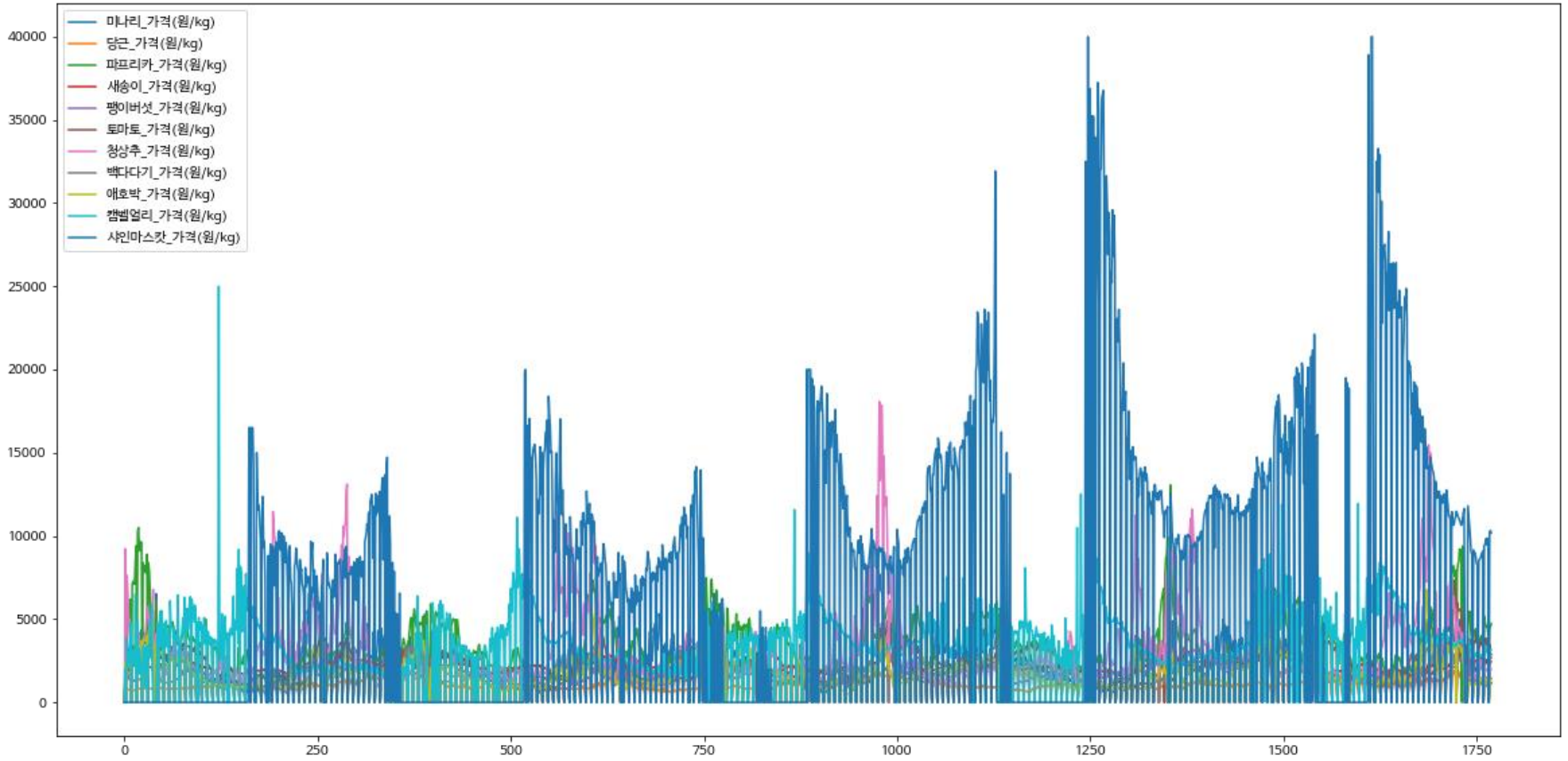
* 대상품목(16): 배추, 무, 양파, 건고추, 마늘, 대파, 얼갈이배추, 양배추, 깻잎, 시금치, 미나리, 당근, 파프리카, 새송이, 팽이버섯, 토마토
* 대상품종(5): 청상추, 백다다기, 애호박, 캠벨얼리, 샤인마스캇

* 가격산출기준: 전국 도매시장 (총 거래금액)/(총 거래량) (원/kg)
※ 거래 취소내역(음수로 집계)은 미반영

농산물 품목별 가격

계절 패턴이 뚜렷하고, 품목별로 확연히 다른 분포를 보이는 것을 확인

- 시계열 반영 모델 생성
- 품목별 모델 생성



농산물 가격 분포

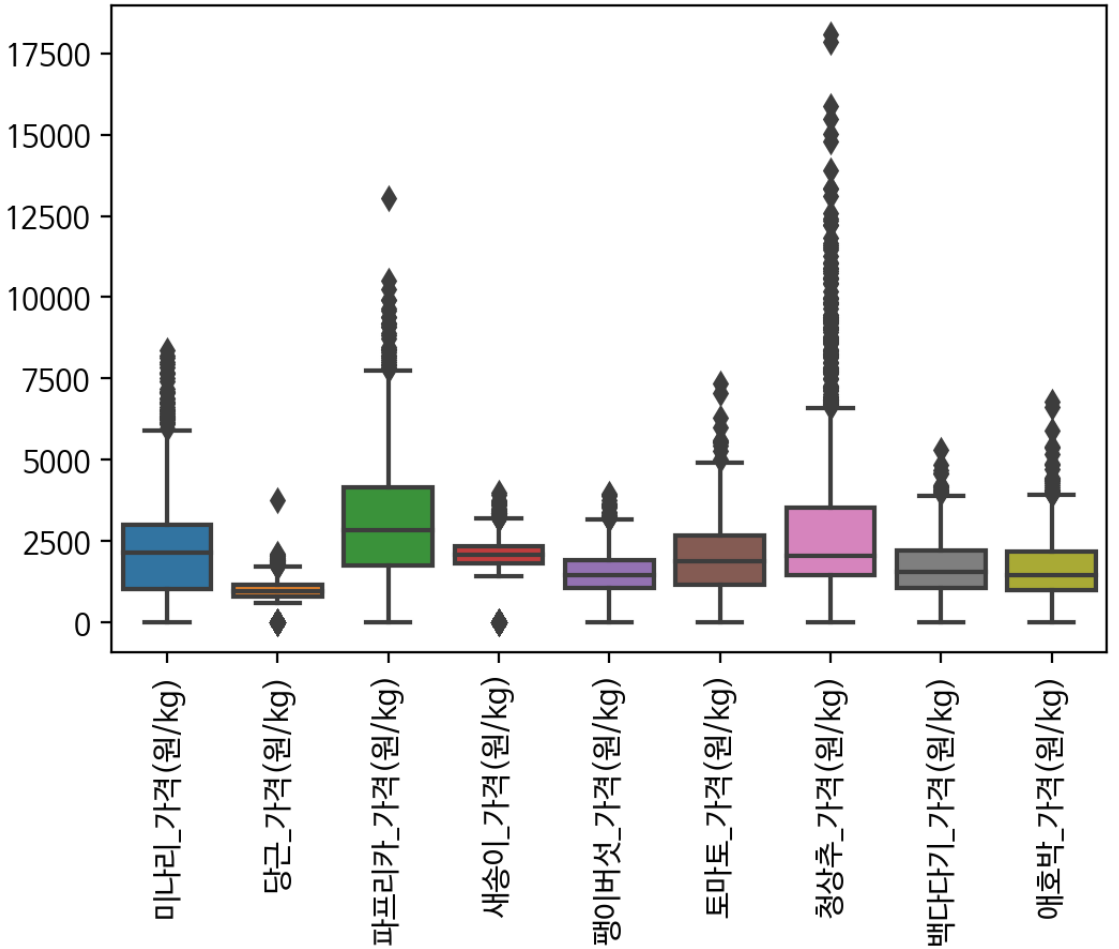
- 이상치가 많이 포함되어 있는 분포이다.
  (이상치는 예측이 쉽지 않고, 모델에 악영향을 주기 때문에 제거해야 하는 값)

- 가격 데이터의 특성상 외부 요인들로 인해,
  가격 급등이 발생할 가능성이 존재한다.

› **이상치가 아닌, 예측해야 할 특이값이라 판단**

- 실제 발생한 데이터이기 때문에 제거가 불가능하다.
  특이값도 예측하기 위한 방안이 필요하다.

› **분해 시계열의 잔차를 통해 특이값 예측이 가능**



농산물 가격 분포

- 이상치가 많이 포함되어 있는 [illegible]
(이상치는 예측이 쉽지 않고, 모델에 악영향을 주기 때문에 제거해야 하는 값)

- 가격 데이터의
가격 급등이 발

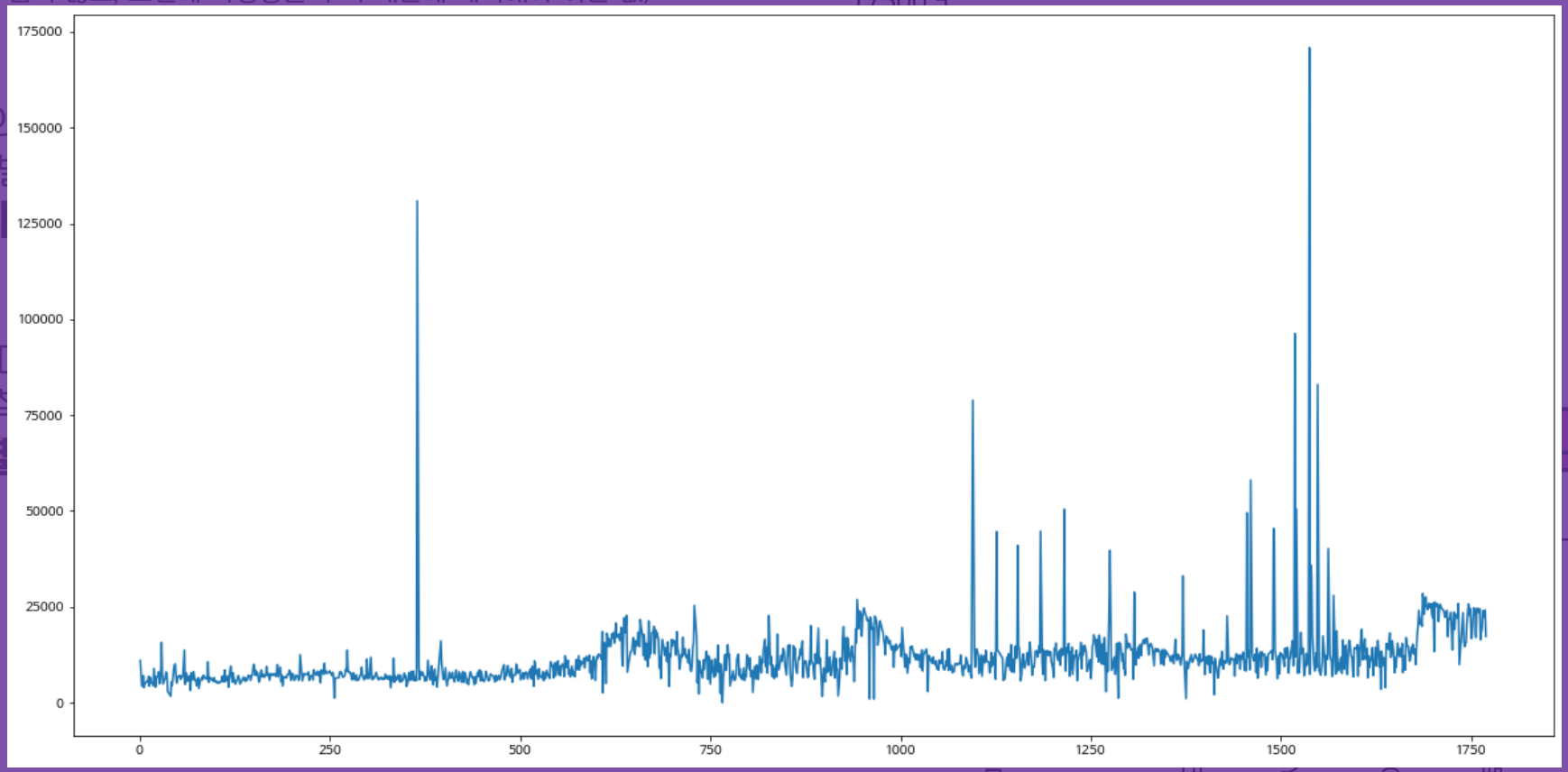
> **이상치가 이**

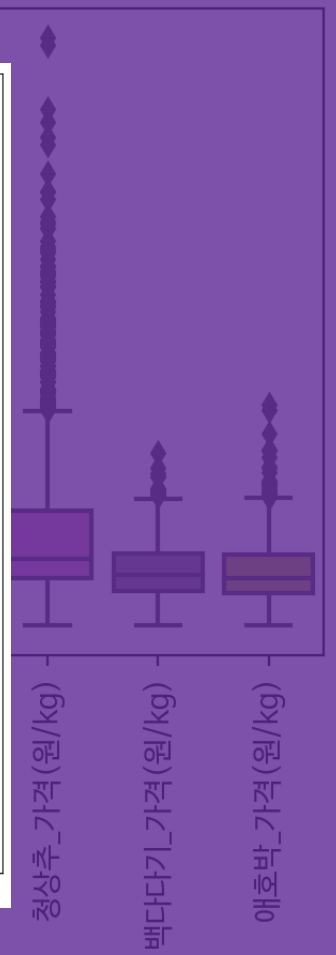
- 실제 발생한 [illegible]
특이값도 예측

> **분해 시계열**

특이값 발생 원인

건고추의 경우 장마 기간이 길어지면,
고추 건조 시에 오랜 기간이 소요되므로 가격 폭등이 발생하는 경우가 존재





초기 모델

대략적인 농산물의 가격 예측 성능을 알아보기 위해
전처리 과정 없이 주어진 데이터를 그대로 사용하는 예측 모델 생성

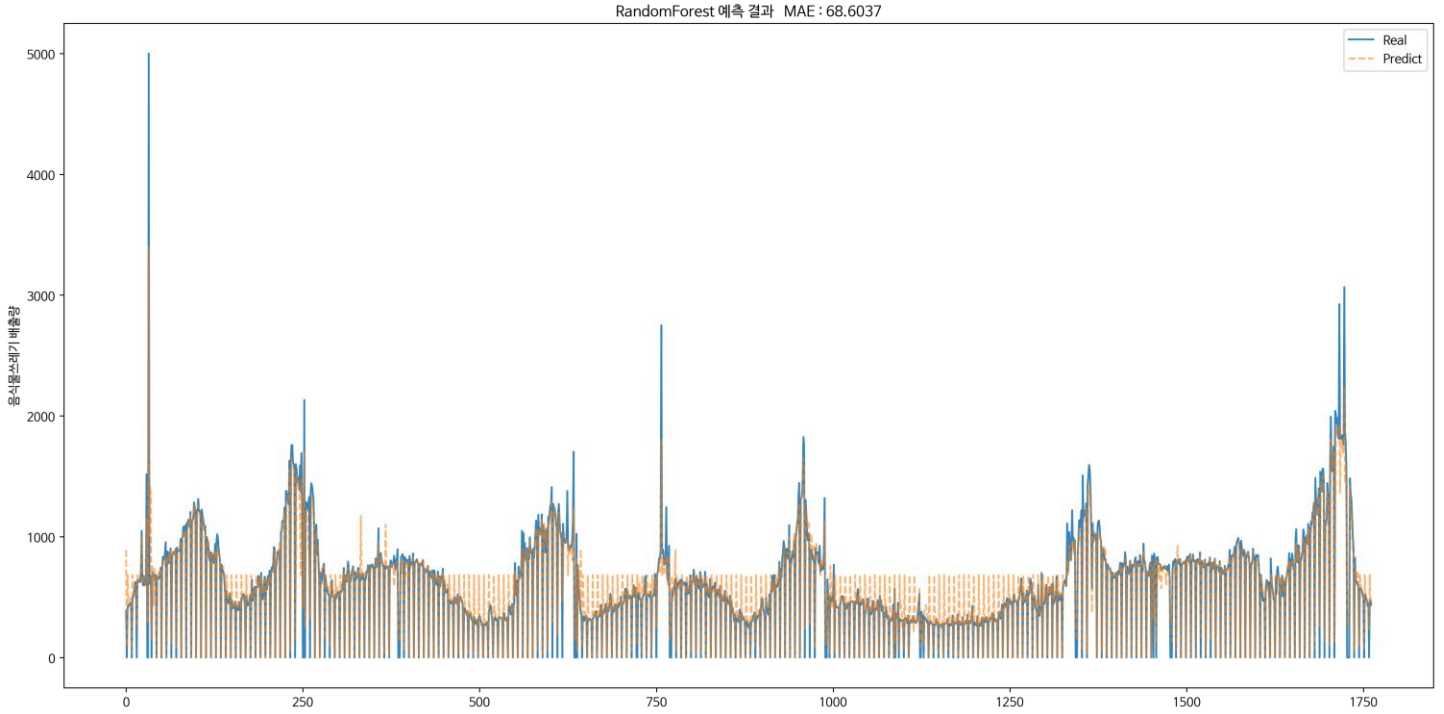
예측 결과

› 일주일 후 배추가격을 예측한 결과, 전반적인 추세는 잘 예측

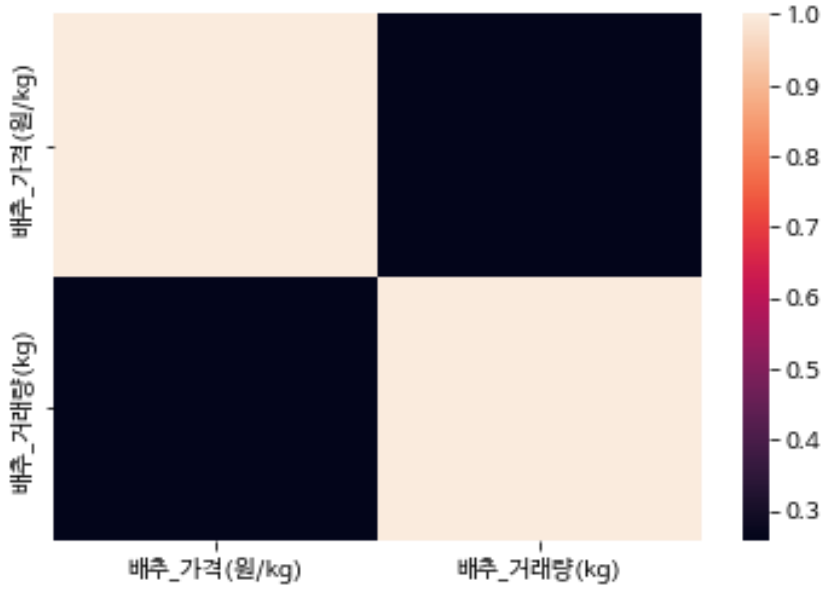
한계점

› 거래가 발생하지 않는 일요일 전후로 예측 성능 저하
› 큰 폭으로 변동되는 가격은 잘 예측하지 못함

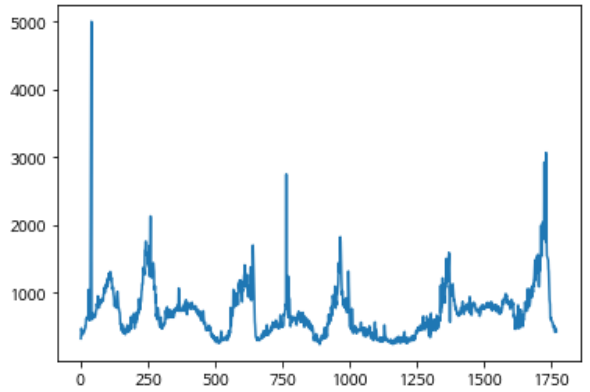
› 상관관계가 있는 변수 도출 필요



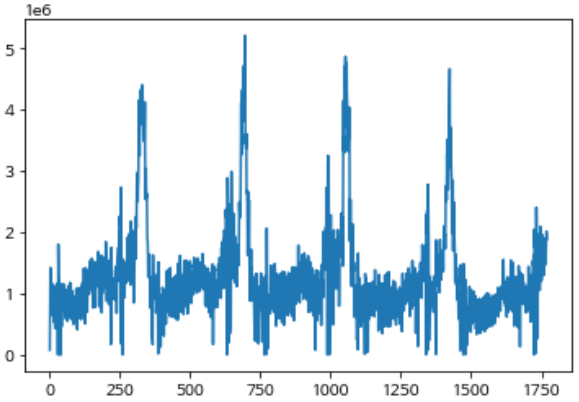
Feature Selection



상관분석 > 거래량은 가격과 큰 상관이 없으므로 feature에서 제외



농산물 가격 : 비정상 시계열 데이터



거래량 : 정상성을 띄는 시계열 데이터

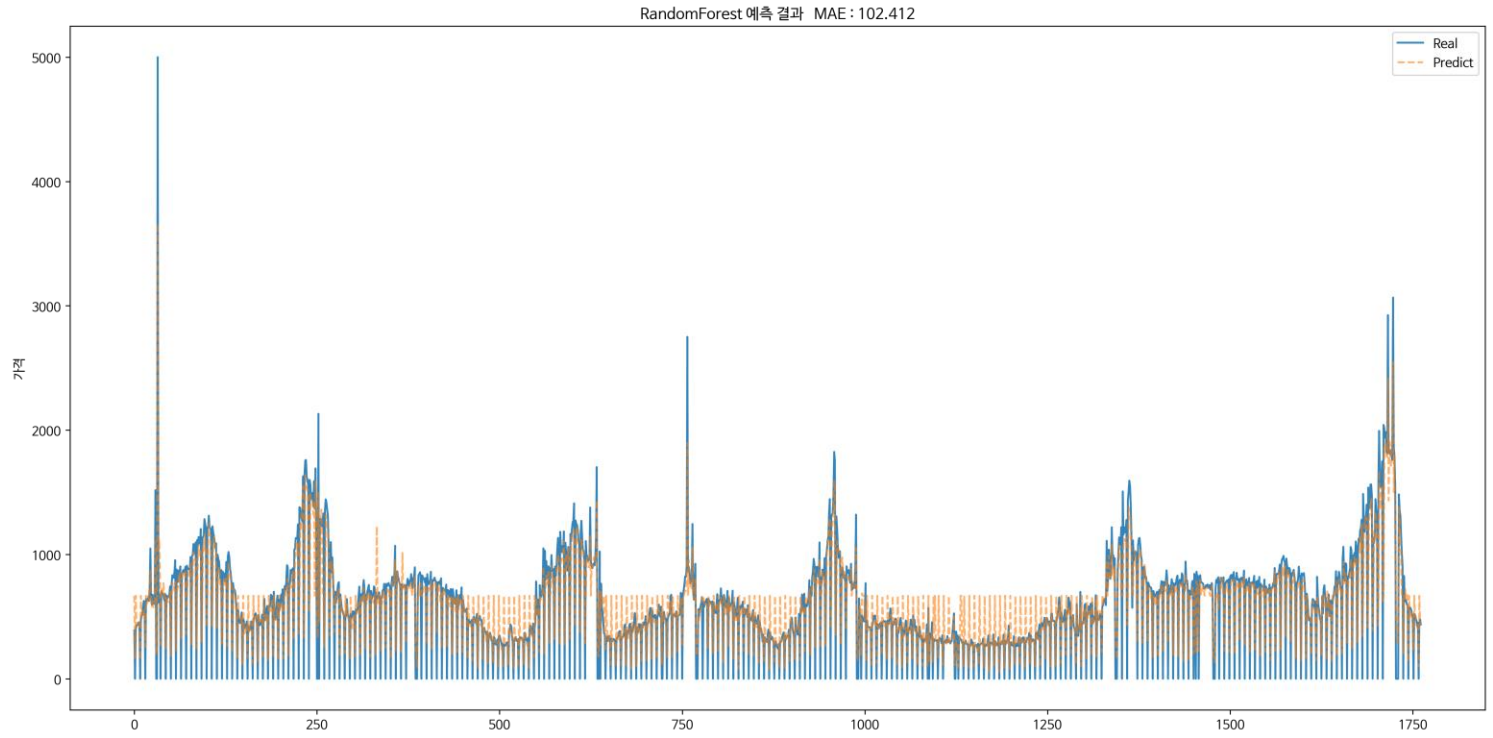
모델 수정 결과 1

Feature에서 거래량을 제외하고 가격만 활용하여 예측 모델 생성

예측 결과
> 거래가 발생하지 않은 날의 전후는 상대적으로 잘 예측

한계점
> 거래가 발생하지 않은 날은 예측 성능이 저하

> 휴일 처리 방안 필요



Feature Engineering 1

거래가 발생하지 않는 휴일이 모델의 성능 저하를 유발
거래가 발생하는 휴일이 존재

> 거래 발생 휴일 예측 필요

	date	요일	배추_거래량(kg)	배추_가격(원/kg)	무_거래량(kg)	무_가격(원/kg)	양파_거래량(kg)	양파_가격(원/kg)	...	백다다기_거래량(kg)	백다다기_가격(원/kg)	애호박_거래량(kg)	애호박_가격(원/kg)	캠벨얼리_거래량(kg)	캠벨얼리_가격(원/kg)	샤인마스캇_거래량(kg)	샤인마스캇_가격(원/kg)
1	2016-01-02	토요일	80860.00	329.0	80272.00	360.0	122787.50	1281.0		434.0	2109.0	19159.0	2414.0	880.0	2014.0	0.0	0.0
2	2016-01-03	일요일	751801.25	403.5	889962.85	371.0	1218933.25	1258.0		250568.0	2077.5	319849.0	2216.0	1791.9	2949.5	0.0	0.0
3	2016-01-04	월요일	1422742.50	478.0	1699653.70	382.0	2315079.00	1235.0		500702.0	2046.0	620539.0	2018.0	2703.8	3885.0	0.0	0.0

Null 값 처리 > 거래가 발생하지 않은 날의 농산물 가격을 하루 전, 하루 후 가격의 평균으로 대체

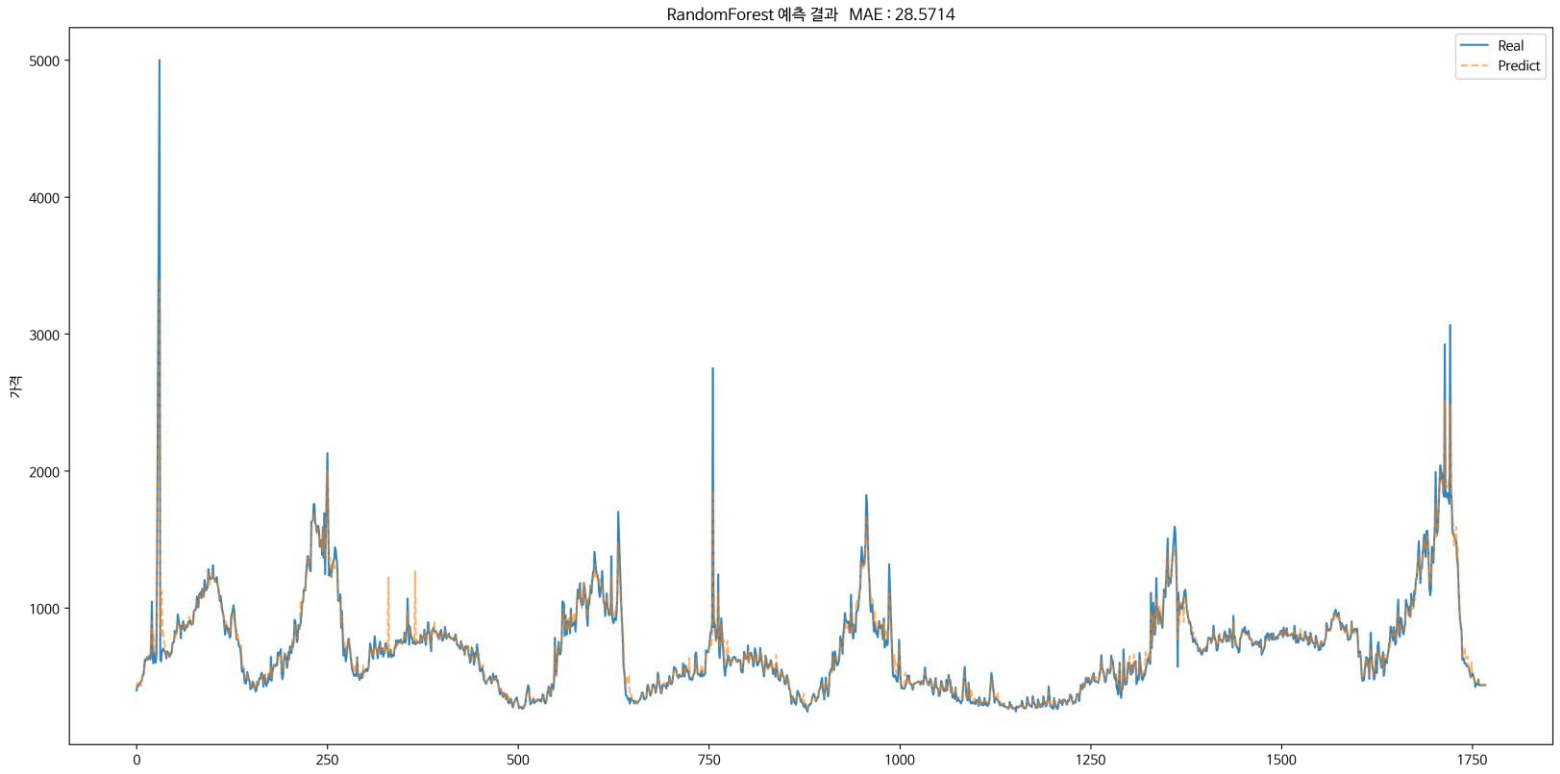
모델 수정 결과 2

Null값을 평균으로 대체한 예측 모델 생성

예측 결과
> 평균 대치법으로 모델에 끼치는 악영향을 제거
> 거래가 발생하는 휴일의 가격 예측 가능
> 전반적 추세는 상당히 높은 수준으로 예측

한계점
> 특이값에 대한 예측 성능은 상대적으로 떨어짐

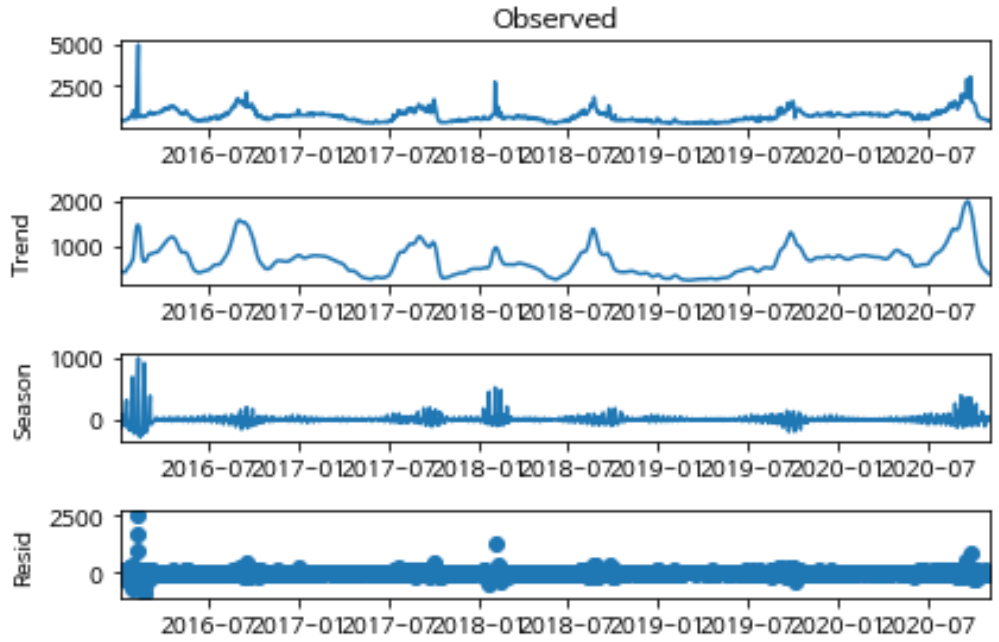
> 잔차 활용 필요



feature engineering 2 - 시계열 분해(STL)

변동성, 계절성, 추세적 특성이 높은 농산물 가격을 예측할 때, 원 데이터를 직접 사용하는 것보다
필터, STL 등의 전처리 과정이 중요한 역할을 하는 것으로 알려져 있다.

> STL을 사용하여 계절성, 추세, 잔차로 분해



Feature 추가 > 분해 시계열의 잔차를 feature로 활용

[출처]
STL-ATTLSTM: Vegetable Price Forecasting Using STL and Attention Mechanism-Based LSTM(mdpi.com)
Helin Yin, Dong Jin, Yeong Hyeon Gu, Chang Jin Park, Sang Keun Han, Seong Joon Yoo, "STL-ATTLSTM: Vegetable Price Forecasting Using STL and Attention Mechanism-Based LSTM", Agriculture 2020

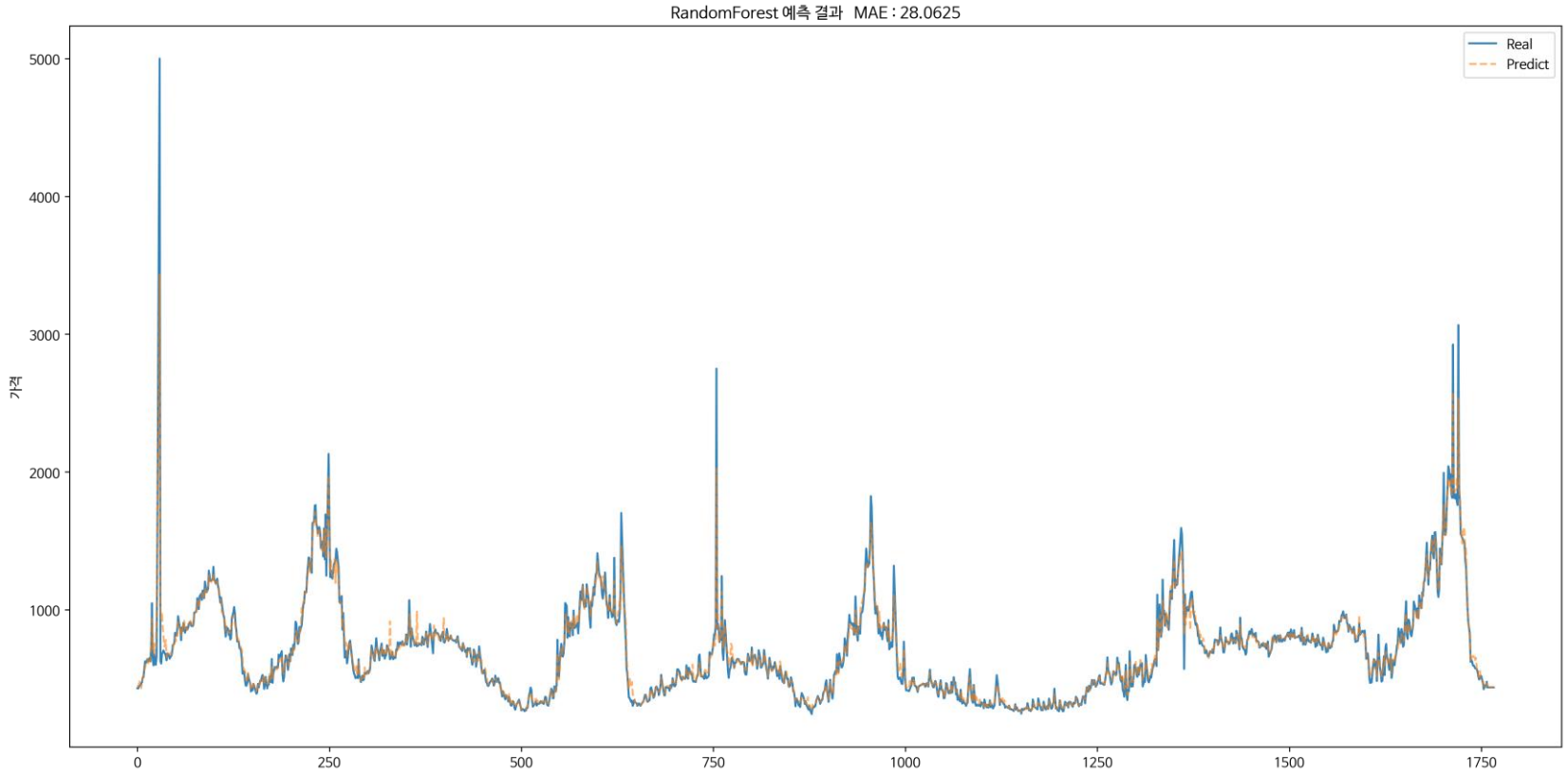
모델 수정 결과 3

분해 시계열의 잔차를 feature로 활용하여 최종 예측 모델 생성

› **전반적인 농산물 가격 뿐만 아니라 특이값도 예측 가능**

분해 시계열의 잔차

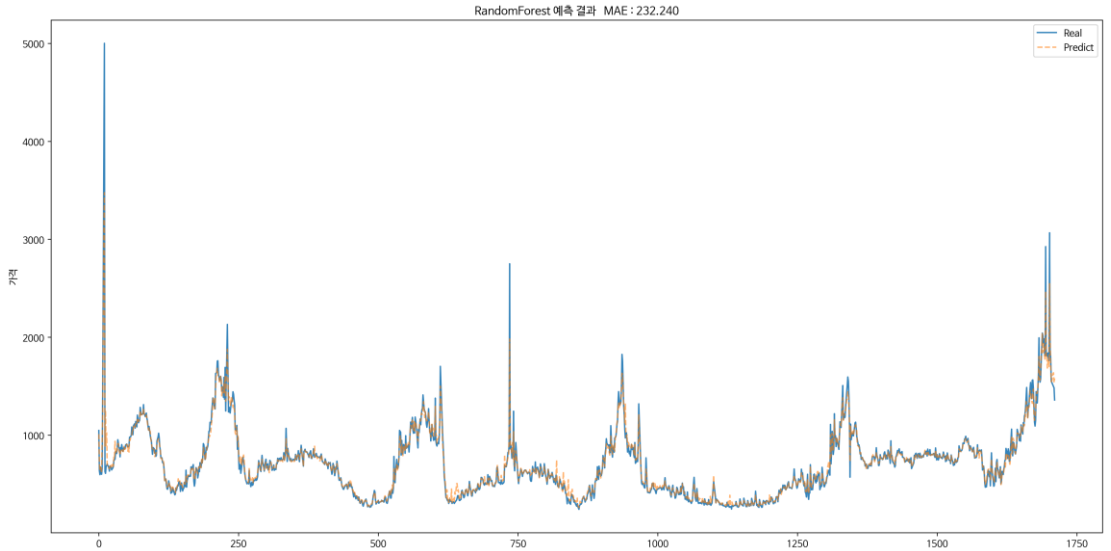
	date	요일	배추_거래량(kg)	배추_가격(원/kg)	...	토요일	화요일	target	resid
1	2016-01-02	토요일	80860.00	329.0		1.0	0.0	0.0	-85.471205
2	2016-01-03	일요일	751801.25	403.5		1.0	0.0	398.0	-1.189226



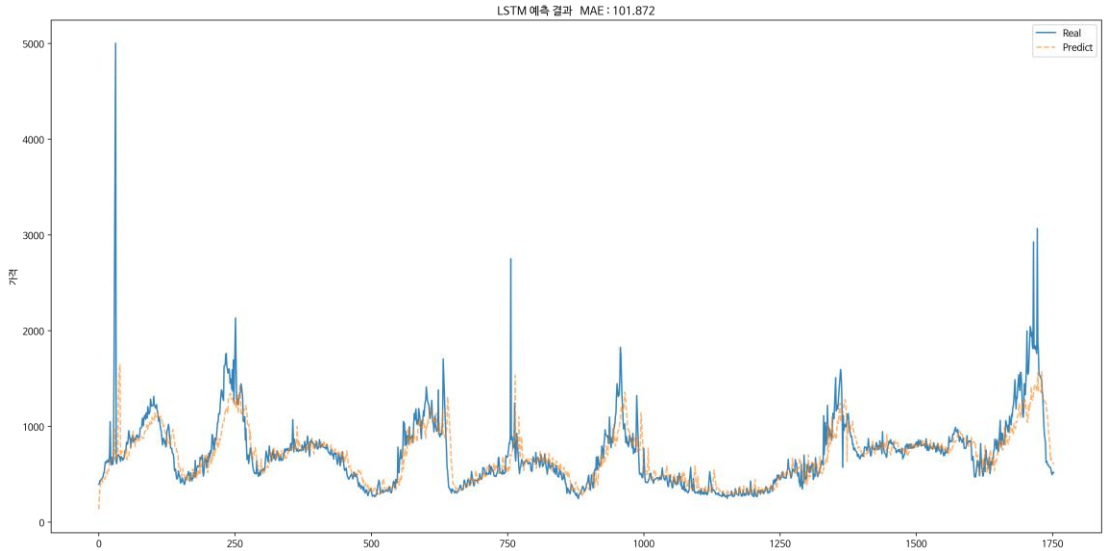
성능평가

성능 평가를 위해 Train 셋과 Test 셋을 구분하여 모델의 성능을 평가
평가 지표: MAE / Target: 4주 후

> 랜덤포레스트 모델은 미래를 잘 예측하지 못함
> 시계열을 반영하기 위해 LSTM 모델 사용
> Feature selection 기능이 없는 LSTM 모델의 특성을 고려하여 예측하고자 하는 target feature의 가격을 feature로 활용



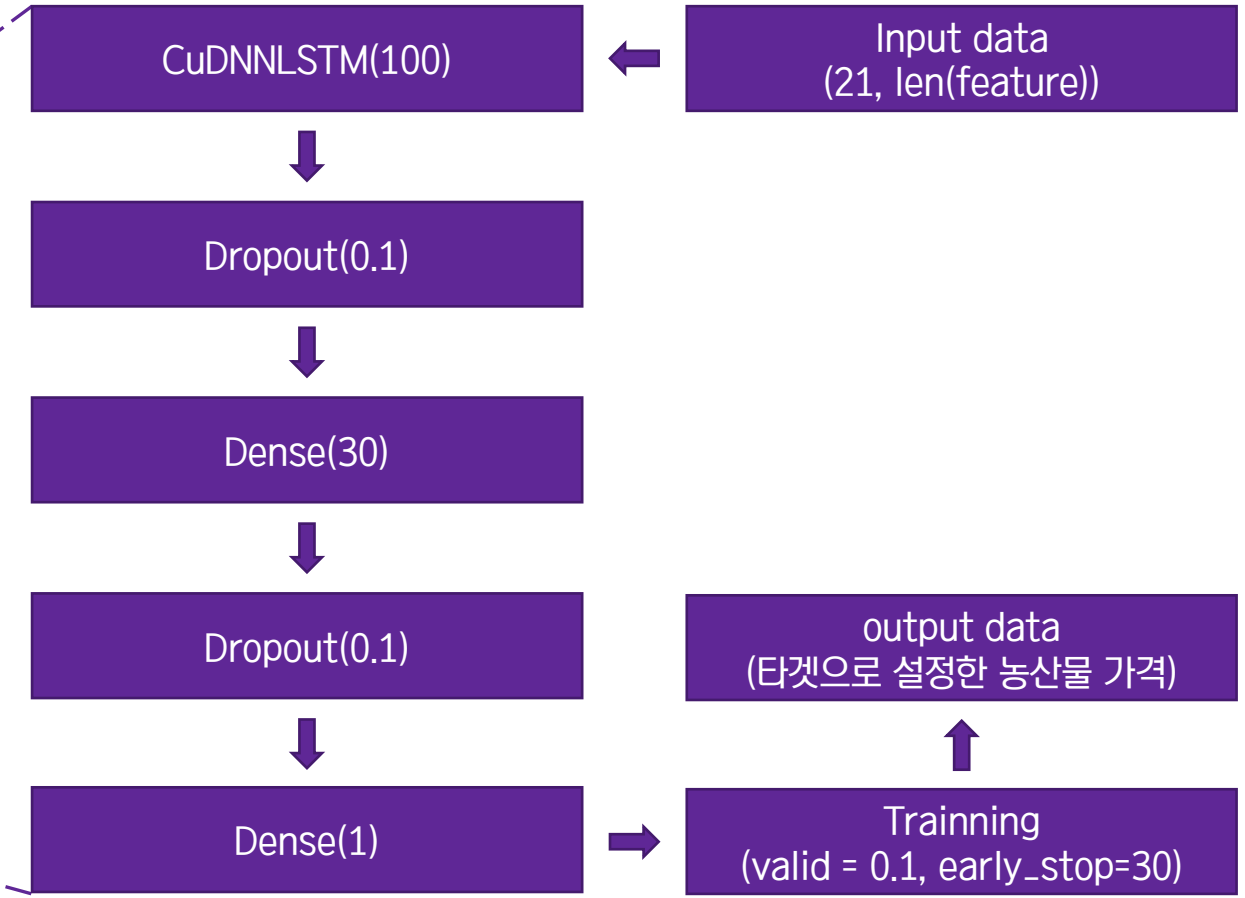
RandomForest MAE: 232.240



LSTM MAE: 101.872

최종 모델

```
class Nong1:

  def __init__(self, df, test):
    데이터프레임 생성
    데이터 전처리
    feature engineering
    feature selection

  def set_feature(self,name):
    target의 feature를 설정

  def set_target(self,week):
    target 설정
    target feature의 시계열 분해

  def set_model(self):
    MinMax스케일링
    학습데이터 프레임 선언
    LSTM에 맞게 reshape
    모델링

  def get_plot(self)
    과거 예측
    평가지표 MAE
    그래프 생성

  def get_price(self):
    타겟으로 설정한 농산물 가격을 받아오는 코드
```



파이프라인

데이터프레임 생성
제출일 기준 과거의 데이터를 받아오는 코드

데이터 로드
제출일 기준 어제의 농산물 거래 데이터를 받아오는 코드

데이터 전처리
농산물 거래 데이터를 일별 거래량, 가격 데이터로 바꾸는 코드

모델 실행
제출일 기준 모든 거래 품목의 1주후, 2주후, 4주후 가격을 예측하는 코드

데이터프레임 업데이트
제출일 기준 어제의 농산물 가격 데이터를 기존의 데이터프레임과 합치는 코드

데이터프레임 생성

예측일 기준 과거의 데이터를 받아오는 코드

```
df1 = pd.read_csv('/content/gdrive/MyDrive/nongsan_data/df1.csv', encoding='utf-8')
df1
```

	date	요일	배추_거래량(kg)	배추_가격(원/kg)	무_거래량(kg)	무_가격(원/kg)	양파_거래량(kg)	양파_가격(원/kg)	...	애호박_거래량(kg)	애호박_가격(원/kg)	캠벨얼리_거래량(kg)	캠벨얼리_가격(원/kg)	샤인마스캇_거래량(kg)	샤인마스캇_가격(원/kg)
0	2016-01-01	금요일	0.0	0.000000	0.0	0.000000	0.0	0.000000		0.0	0.0	0.0	0.0	0.0	0.0
1	2016-01-02	토요일	80860.0	329.000000	80272.0	360.000000	122787.5	1281.000000	...	19159.0	2414.0	880.0	2014.0	0.0	0.0
2	2016-01-03	일요일	0.0	0.000000	0.0	0.000000	0.0	0.000000		0.0	0.0	0.0	0.0	0.0	0.0
...	...	...	...	...	...	...	...	...		...	...	...	...	...	...
2061	2021-09-25	토요일	1642272.0	704.139482	1861274.3	349.233220	1476640.5	832.179434		622795.6	623.0	243256.0	5117.0	441083.3	9507.0
2062	2021-09-26	일요일	0.0	0.000000	0.0	0.000000	0.0	0.000000	...	0.0	0.0	0.0	0.0	0.0	0.0
2063	2021-09-27 00:00:00	월요일	1880289.1	668.382521	2308290.5	309.275538	2040117.6	817.131752		637161.9	853.0	289178.0	4876.0	437645.6	8801.0

2064 rows × 44 columns

데이터 로드

예측일 기준 어제의 농산물 거래 데이터를 받아오는 코드
'date' 에 어제 날짜만 입력하면 이후 코드(학습부터 예측 제출 파일 생성)는 자동으로 실행

```
# date : 제출일 기준 어제 날짜로
date='20210928'
url = 'https://www.nongnet.or.kr/api/whlsIDstrQr.do?sdate='+date

response = urllib.request.urlopen(url).read()
response = json.loads(response)

data = pd.DataFrame(response['data'])
data
```

data

	PUM_NM	LV_NM	TOT_AMT	SAN_NM	SALEDATE	CMP_NM	DAN_NM	WHSAL_NM	SIZE_NM	COST	POJ_NM	TOT_QTY	QTY	KIND_NM	DANQ
0	마늘	특	278010.0	중국	20210928	청주청과	kg	청주도매시장	.	30890	.	72.0	9.0	마늘쫑(수입)	8.0
1	가자미	자연산 하	44000.0	충청남도 보령시	20210928	강북수산	kg	구리도매시장	.	22000	.	14.0	2.0	기타	7.0
2	가자미	자연산 하	132000.0	충청남도 보령시	20210928	강북수산	kg	구리도매시장	.	22000	.	42.0	6.0	기타	7.0
3	가자미	자연산 하	132000.0	충청남도 보령시	20210928	강북수산	kg	구리도매시장	.	22000	.	42.0	6.0	기타	7.0
4	가자미	자연산 하	40000.0	울산 동구	20210928	강북수산	kg	구리도매시장	.	40000	.	10.0	1.0	기타	10.0
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
131527	배추	.	-312000.0	None	20210928	강서청과	kg	서울강서도매	2개	2000	접	-1872.0	-156.0	고냉지배추	12.0
131528	배추	.	1560000.0	None	20210928	강서청과	kg	서울강서도매	.	2500	접	7488.0	624.0	고냉지배추	12.0
131529	배추	.	234000.0	None	20210928	강서청과	kg	서울강서도매	2개	1500	접	1872.0	156.0	고냉지배추	12.0
131530	가지	.	-110000.0	None	20210928	강서청과	kg	서울강서도매	1개(내_뿌리)	5500	상자	-100.0	-20.0	가지(일반)	5.0
131531	가지	.	110000.0	None	20210928	강서청과	kg	서울강서도매	1개(내_뿌리)	5500	상자	100.0	20.0	가지(일반)	5.0

131532 rows × 15 columns

데이터 전처리

농산물 거래 데이터를 일별 거래량, 가격 데이터로 바꾸는 코드

```
for day in days:
    train_dict['date'].append(day)
    for sub in unique_pum:
        # 날짜별, 품목별, 거래량이 0 이상인 행만 선택
        c = tsalet_sample[(tsalet_sample['SALEDATE']==day) & (tsalet_sample['PUM_NM']==sub) &
(tsalet_sample['TOT_QTY']>0)]
        if c.shape[0] == 0:
            train_dict[f'{sub}_거래량(kg)'].append(0)
            train_dict[f'{sub}_가격(원/kg)'].append(0)
        else:
            tot_amt = c['TOT_AMT'].sum().astype(float)
            tot_qty = c['TOT_QTY'].sum().astype(float)
            mean_price = tot_amt/(tot_qty+1e-20)
            train_dict[f'{sub}_거래량(kg)'].append(tot_qty)
            train_dict[f'{sub}_가격(원/kg)'].append(mean_price)
```

```
for sub in unique_kind:
    # 날짜별, 품종별, 거래량이 0 이상인 행만 선택
    c = tsalet_sample[(tsalet_sample['SALEDATE']==day) & (tsalet_sample['KIND_NM']==sub) &
(tsalet_sample['TOT_QTY']>0)]
    if c.shape[0] == 0:
        train_dict[f'{sub}_거래량(kg)'].append(0)
        train_dict[f'{sub}_가격(원/kg)'].append(0)
    else:
        tot_amt = c['TOT_AMT'].sum().astype(float)
        tot_qty = c['TOT_QTY'].sum().astype(float)
        mean_price = round(tot_amt/(tot_qty+1e-20))
        tot_qty = round(tot_qty, 1)
        train_dict[f'{sub}_거래량(kg)'].append(tot_qty)
        train_dict[f'{sub}_가격(원/kg)'].append(mean_price)
```

> df2

	date	배추_거래량(kg)	배추_가격(원/kg)	무_거래량(kg)	무_가격(원/kg)	양파_거래량(kg)	양파_가격(원/kg)	...	백다다기_거래량(kg)	백다다기_가격(원/kg)	애호박_거래량(kg)	애호박_가격(원/kg)	캠벨얼리_거래량(kg)	캠벨얼리_가격(원/kg)	샤인마스캇_거래량(kg)	샤인마스캇_가격(원/kg)
0	2021-09-28	1645627.9	572.767507	1800760.7	280.827541	1892908.7	849.411797		386158.2	1466	448996.1	1000	204770.0	4442	341639.9	8110

모델 실행

예측일 기준 모든 거래 품목의 1주 후, 2주 후, 4주 후 가격을 예측하는 코드

```
weeks = [1,2,4]
features = ['배추', '무', '양파', '건고추', '마늘', '대파', '얼갈이배추', '양배추', '깻잎', '시금치', '미나리', '당근', '파프리카', '새송이', '팽이버섯',
        '토마토', '청상추', '백다다기', '애호박', '캠벨얼리', '샤인마스캇']

week1=[]
week2=[]
week4=[]

for week in weeks:
 print(week)
 for feature in features:
  my_nong1 = Nong1(df1, df2)
  my_nong1.set_feature(feature)
  my_nong1.set_target(week)
  my_nong1.set_model()
  if week == 1:
   week1.append(my_nong1.get_price())
  if week == 2:
   week2.append(my_nong1.get_price())
  if week == 4:
   week4.append(my_nong1.get_price())
  print(feature)
```

데이터프레임 업데이트

예측일 기준 어제의 농산물 가격 데이터(df2)를 기존의 데이터프레임(df1)과 합치는 코드
> 매일 업데이트 되는 데이터를 자동으로 업데이트 가능

```
df1 = pd.concat([df1, df2], axis=0)
df1.to_csv('/content/gdrive/MyDrive/nongsan_data/df1.csv', encoding='utf-8-sig', index=False)
```

date	요일	배추_거래량(kg)	배추_가격(원/kg)	무_거래량(kg)	무_가격(원/kg)	양파_거래량(kg)	양파_가격(원/kg)	...	애호박_거래량(kg)	애호박_가격(원/kg)	캠벨얼리_거래량(kg)	캠벨얼리_가격(원/kg)	샤인마스캇_거래량(kg)	샤인마스캇_가격(원/kg)
2016-01-01	금요일	0.0	0.000000	0.0	0.000000	0.0	0.000000		0.0	0.0	0.0	0.0	0.0	0.0
2016-01-02	토요일	80860.0	329.000000	80272.0	360.000000	122787.5	1281.000000		19159.0	2414.0	880.0	2014.0	0.0	0.0
2016-01-03	일요일	0.0	0.000000	0.0	0.000000	0.0	0.000000		0.0	0.0	0.0	0.0	0.0	0.0
2016-01-04	월요일	1422742.5	478.000000	1699653.7	382.000000	2315079.0	1235.000000		620539.0	2018.0	2703.8	3885.0	0.0	0.0
2016-01-05	화요일	1167241.0	442.000000	1423482.3	422.000000	2092960.1	1213.000000		231958.0	2178.0	8810.0	2853.0	0.0	0.0
...	...	...	...	...	...	...	...		...	...	...	...	...	...
2021-09-23	목요일	20216.0	1464.405125	52127.0	492.965258	14186.0	878.087904		4566.0	2147.0	14383.0	4825.0	9194.0	12065.0
2021-09-24	금요일	2017345.6	803.021436	2384975.3	418.847906	2324538.7	824.078027		898381.7	836.0	227488.0	5338.0	462828.8	10151.0
2021-09-25	토요일	1642272.0	704.139482	1861274.3	349.233220	1476640.5	832.179434		622795.6	623.0	243256.0	5117.0	441083.3	9507.0
2021-09-26	일요일	0.0	0.000000	0.0	0.000000	0.0	0.000000		0.0	0.0	0.0	0.0	0.0	0.0
2021-09-27 00:00:00	월요일	1880289.1	668.382521	2308290.5	309.275538	2040117.6	817.131752		637161.9	853.0	289178.0	4876.0	437645.6	8801.0

date	요일	배추_거래량(kg)	배추_가격(원/kg)	무_거래량(kg)	무_가격(원/kg)	양파_거래량(kg)	양파_가격(원/kg)	...	애호박_거래량(kg)	애호박_가격(원/kg)	캠벨얼리_거래량(kg)	캠벨얼리_가격(원/kg)	샤인마스캇_거래량(kg)	샤인마스캇_가격(원/kg)
2016-01-01	금요일	0.0	0.000000	0.0	0.000000	0.0	0.000000		0.0	0.0	0.0	0.0	0.0	0.0
2016-01-02	토요일	80860.0	329.000000	80272.0	360.000000	122787.5	1281.000000		19159.0	2414.0	880.0	2014.0	0.0	0.0
2016-01-03	일요일	0.0	0.000000	0.0	0.000000	0.0	0.000000		0.0	0.0	0.0	0.0	0.0	0.0
2016-01-04	월요일	1422742.5	478.000000	1699653.7	382.000000	2315079.0	1235.000000		620539.0	2018.0	2703.8	3885.0	0.0	0.0
2016-01-05	화요일	1167241.0	442.000000	1423482.3	422.000000	2092960.1	1213.000000		231958.0	2178.0	8810.0	2853.0	0.0	0.0
...	...	...	...	...	...	...	...		...	...	...	...	...	...
2021-09-24	금요일	2017345.6	803.021436	2384975.3	418.847906	2324538.7	824.078027		898381.7	836.0	227488.0	5338.0	462828.8	10151.0
2021-09-25	토요일	1642272.0	704.139482	1861274.3	349.233220	1476640.5	832.179434		622795.6	623.0	243256.0	5117.0	441083.3	9507.0
2021-09-26	일요일	0.0	0.000000	0.0	0.000000	0.0	0.000000		0.0	0.0	0.0	0.0	0.0	0.0
2021-09-27 00:00:00	월요일	1880289.1	668.382521	2308290.5	309.275538	2040117.6	817.131752		637161.9	853.0	289178.0	4876.0	437645.6	8801.0
2021-09-28 00:00:00	화요일	1645627.9	572.767507	1800760.7	280.827541	1892908.7	849.411797		448996.1	1000.0	204770.0	4442.0	341639.9	8110.0

최종 예측 결과 평가

평가 기간(9월 28일 ~ 11월 4일)동안 매일 농넷 데이터가 업데이트 되면 date를 갱신
→ 파이프라인에 따라 코드를 순차적으로 실행
→ 예측 결과 파일 업로드

최종 평가 점수 : 0.24518

#	팀	팀 멤버	최종점수	제출수
5	가온		0.24518	43

감사합니다

Team | 가온
팀 장 | 정성문
팀 원 | 김세상, 박민규, 서정인, 정유진