

농산물 가격 예측 모델 해석

통 벤 저 스

목차



분석배경



데이터

1. 데이터 수집
2. 데이터 EDA
3. 데이터 통합



모델링

1. 선형모델 : OLS, VAR
2. 비선형모델 : LSTM



결론



분석배경

“예측하기 힘든 농산물 가격”

'햄버거에 양상추가 사라졌다', 1주만에 양배추 가격 4배 '경충' 수급 불안정

ⓒ 오수정 기자 | ⓒ 승인 2021.10.27 19:42

갑작스러운 한파로 양상추 가격 급등

[FT스포츠] “햄버거를 시켰는데 햄과 빵만 있어요” 햄버거에 들어가는 양상추 수급이 불안정해지면서 일부 프랜차이즈 햄버거에 양상추가 빠진 채 판매되고 있다.

얼마전 갑작스러운 한파로 인해 채소 재배 농가들이 냉해 피해를 입은 가운데 양상추 냉해 피해가 심해 시중 양상추의 가격이 급등하면서 수급이 불안정해진 탓이다. 양상추가 들어가는 햄버거와 샌드위치 등의 메뉴에 필수인 양상추는 재조달이 어렵게 되어 빠르게 가격이 상승했다. 이에 빠진 채로 판매하는 햄버거를 보며 ‘불고기 마카롱이냐’는 농담이 돌고 있다.

한 샐러드가게 점주가 온라인에 올린 글에는 “지난 19일 양배추 가격이 10kg에 22,250원이었는데 26일 발주 주문서에는 양배추 300g에 7000원으로 1주 전보다 3~4배 차이가 난다”고 밝히며 양배추 가격이 급등했음을 알렸다.

예상치 못한 날씨 탓으로 양배추 외에도 오이 가격도 최근 두배 가까이 올랐으며 상추와 깻잎 등 잎채소도 50~70% 가격상승을 보이며 수확철 한창 재미를 보아야 할 농가들에게 울상을 짓게 만들고 있다.

업체 관계자는 “추운 날씨가 계속 될 경우 햄버거 업계는 큰 타격을 입을 것”이라 말했다. 양상추는 생육기간이 다른 채소에 비해 기간이 더욱 필요하기 때문에 공급 불안 현상이 장기화 될 가능성이 크다.

저작권 © FT스포츠 무단전재 및 재배포 금지

한파로 인해 채소 재배 농가들이 냉해 피해를 입은 가운데 양상추 냉해 피해가 심해 시중 양상추 가격이 급등하면서 수급이 불안정

양배추, 오이, 상추, 깻잎 등 여러 농작물 가격 상승

배추가격 급등 '김장비용은 하락'

(서울=뉴스1) 박세연 기자 | 입력 2021-11-11 13:38:45 | 수정 2021.11.11 13:38:45

11일 오후 서울 송파구 가락농수산물도매시장에서 배추를 판매하고 있다. 김장철을 앞두고 주재료인 배추가 무름병 피해와 늦가을 기습 한파로 가격이 크게 올랐다. 지난 9일 기준으로 전통시장에서 배추 1포기당 가격은 5500원으로 지난해보다 38%가량 올랐다. 올해 김장 비용은 배추를 비롯해 부재료 가격은 인상됐지만 고춧가루 가격이 올해 크게 내리면서 전체적인 금액은 지난해보다 적게 들 것으로 조사됐다. 한국물가정보에 따르면 올 김장비용은 전통시장이 지난해보다 4.9% 내린 31만원, 대형마트는 4.1% 내린 35만7000원이 들 것으로 예상됐다. 2021.11.11/뉴스1

기습 한파로 인해 배추를 비롯해 김장 재료 가격이 인상됐지만 고춧가루 가격이 크게 내려 김장 금액은 적게 들 것

가을장마까지...추석특수 사라질 판, 농산물 가격 빨간불

이소희 기자
입력 2021.08.26 16:39 수정 2021.08.26 16:41



가을장마·추가태풍, 가격변수

평년대비 1.5배·전년 추석보다 1.4배 공급 ↑
공급확대 계단·축산물도 상승가격 여전

“기상변화에 따라 수급이 불안정 해지면서 가격 급등”

한파, 폭염, 장마 등의 기상변화와 가축질병 등으로 인한 농축산물 수급이 원활치 않아 30% 가량의 가격상승에 농민들은 물량수급을, 소비자들은 장바구니 물가를 우려하는 상황이 이어지고 있다.



분석배경

“예측하기 힘든 농산물 가격”

| 턱없이 부족한 농산물 가격 정보

국내 농산물 생산액은 연간 32조원(2019년 기준)에 달한다. 도·소매시장에서 거래되는 금액은 연간 100조원을 넘는다. 자동차와 반도체보다 훨씬 더 큰 시장이다. 그러나 이 시장엔 표준화된 데이터와 가격 기준, 예측 시스템이 없다.

가장 큰 불이익은 농작물을 생산하는 농업인의 몫이다. 정확하게 제공되는 농산물 가격이 없어 매년 폭락하거나 급등하는 상황을 감내해야 한다. 중간 구매자인 기업 역시 정확한 정보 없이 수십억원에서 수조원에 달하는 농산물 구매를 주먹구구식으로 결정해야 한다. 대기업 유통·제조사 구매책임자(MD)들은 표준화된 가격 정보 부재로 일일이 산지 거래처나 도매시장 등에 연락해 최적의 구매 가격을 추정하는 식이다. 농산물의 최종 소비자들도 현재 가격이 높은지 낮은지에 대한 명확한 정보를 찾기 어렵긴 마찬가지다.

농산물 가격 변화의 가장 큰 불이익은
농업인과 소비자의 몫,
하지만 체계적인 예측 시스템이 없다.



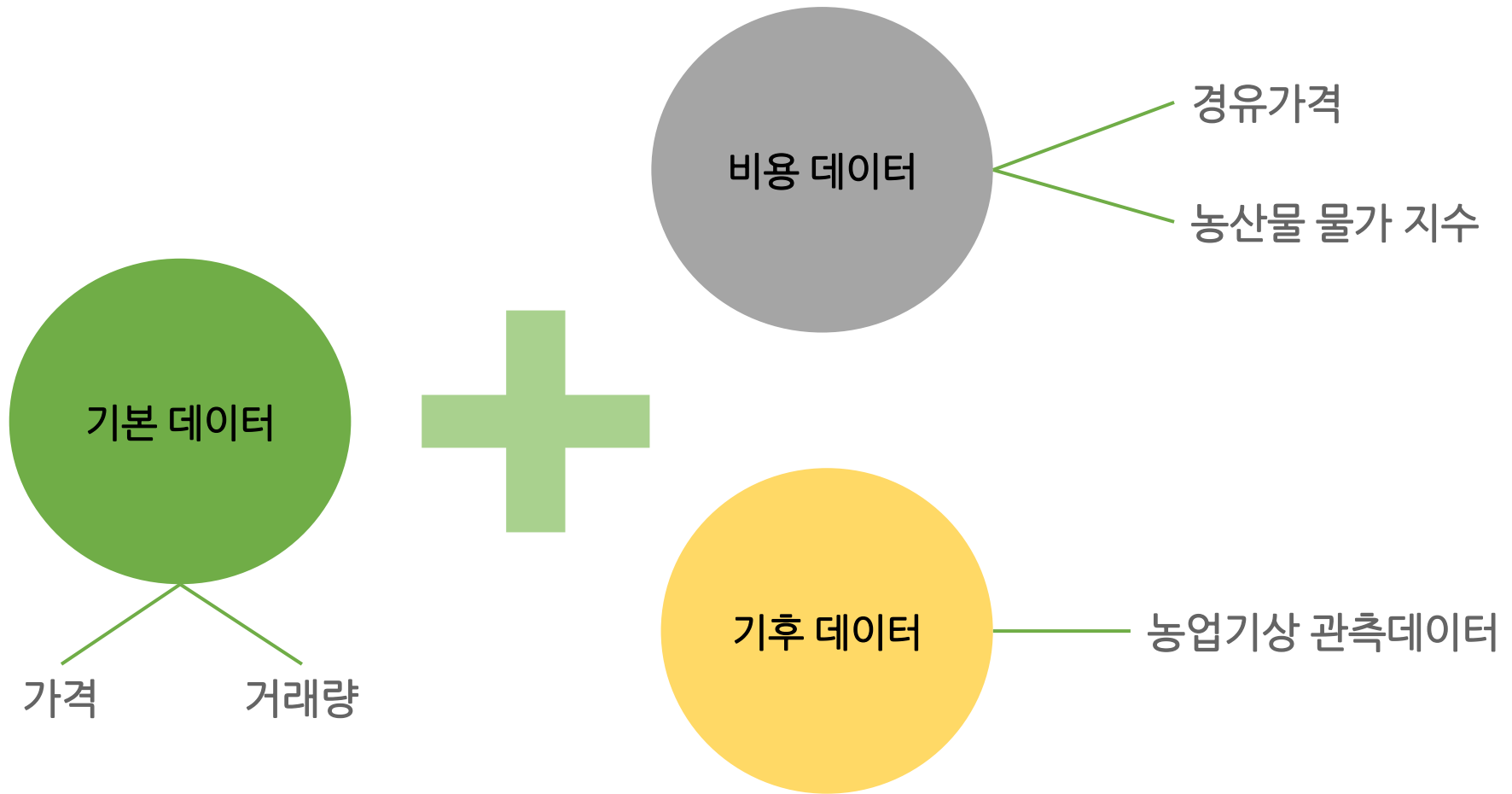
농산물 가격 예측 모델 분석을 통해 기후와 가격 간의 인사이트 도출

데이터

1. 데이터 수집
2. 데이터 EDA
3. 데이터 통합



데이터 수집





데이터 수집

1. 기본 데이터 2. 비용 데이터 3. 기후 데이터

1. 가격

	date	요일	배추_가격 (원/kg)	무_가격 (원/kg)	양파_가격 (원/kg)	건고추_가 격(원/kg)
0	2016-01-01	토요일	0.0	0.0	0.0	0.0
1	2016-01-02	일요일	329.0	360.0	1281.0	11000.0
2	2016-01-03	월요일	0.0	0.0	0.0	0.0
3	2016-01-04	화요일	478.0	382.0	1235.0	4464.0
4	2016-01-05	수요일	442.0	422.0	1213.0	4342.0

...

캠벨얼리_가 격(원/kg)	샤인마스캇 가격(원/kg)
0.0	0.0
2014.0	0.0
0.0	0.0
3885.0	0.0
2853.0	0.0

2. 거래량

	date	요일	배추_거래 량(kg)	무_거래량 (kg)	양파_거래 량(kg)	건고추_기 래량(kg)
0	2016-01-01	토요일	0.0	0.0	0.0	0.0
1	2016-01-02	일요일	80860.0	80272.0	122787.5	3.0
2	2016-01-03	월요일	0.0	0.0	0.0	0.0
3	2016-01-04	화요일	1422742.5	1699653.7	2315079.0	699.0
4	2016-01-05	수요일	1167241.0	1423482.3	2092960.1	1112.6

...

캠벨얼리_기 래량(kg)	샤인마스캇 거래량(kg)
0.0	0.0
880.0	0.0
0.0	0.0
2703.8	0.0
8810.0	0.0

(출처 : 데이콘)

각 품목의 일별 가격(원/kg) 및 거래량(kg)

품목	배추, 무, 건고추, 마늘, 대파, 얼갈이배추, 양배추, 갯잎, 시금치, 미나리, 당근, 파프리카, 새송이, 팽이버섯, 토마토, 청상추, 백다다기, 애호박, 캠벨얼리, 샤인마스캇
기간	2016-01-01 ~ 2020-09-28



데이터 수집

1. 기본 데이터 2. 비용데이터 3. 기후 데이터

1. 경유가격

(출처 : 오피넷)

구분	서울	부산	대구	인천	광주	대전	경남	제주	세종
0 2016년01월01일	1277.10	1175.02	1165.12	1174.31	1174.91	1183.53	1176.24	1203.10	1196.69
1 2016년01월02일	1275.54	1174.62	1164.64	1174.31	1175.90	1182.80	1175.30	1202.65	1196.01
2 2016년01월03일	1275.06	1174.82	1163.69	1173.55	1174.13	1181.91	1174.50	1202.14	1196.01
3 2016년01월04일	1274.66	1173.94	1162.00	1173.35	1172.20	1181.39	1173.37	1201.48	1196.94
4 2016년01월05일	1273.53	1171.30	1160.57	1172.87	1171.27	1178.87	1171.83	1200.02	1195.93

시도별 일별 경유 가격

지역	서울, 부산, 대구, 인천, 광주, 대전, 울산, 경기, 강원, 충북, 충남, 전북, 전남, 경북, 경남, 제주, 세종
기간	2016-01-01 ~ 2020-09-28

2. 농산물 물가 지수

(출처 : 통계청)

시점	농산물
0 2016. 01	101.55
1 2016. 02	109.33
2 2016. 03	106.08
3 2016. 04	104.95
4 2016. 05	101.48

월별 농산물 물가 지수

기간	2016-01 ~ 2020-09
----	-------------------



데이터 수집

1. 기본 데이터 2. 비용데이터 3. 기후 데이터

1. 농업기상 관측데이터

(출처 : 공공데이터포털)

	stn_Code	stn_Name	date	temp	max_Temp	min_Temp	hum	widdir	wind	rain	sun_Time	sun_Qy	condens_Time	gr_Temp	soil_Temp	soil_Wt
0	536824B002	해남군 옥천면	2015-01-01	-1.3	0.6	-2.9	80.0	295.2	2.3	0.8	NaN	7.8	NaN	NaN	3.36	25.9
1	330846A001	천안시 목천읍	2015-01-01	-6.2	-3.8	-8.3	NaN	NaN	0.0	0.0	NaN	NaN	1429.0	NaN	NaN	NaN
2	627911A001	밀양시 상남면	2015-01-01	-3.2	0.2	-7.2	40.1	282.7	2.9	0.0	516.0	11.0	0.0	NaN	2.20	28.5
3	539823A001	진도군 군내면	2015-01-01	-0.8	1.6	-2.8	79.2	257.0	3.5	1.5	217.0	8.2	652.0	NaN	5.02	30.6
4	590823A001	남원시 이백면	2015-01-01	-4.1	-1.3	-6.0	60.7	286.7	2.1	0.5	310.0	7.7	0.0	-4.3	2.16	20.3

관측지점별 일별 기상 데이터

stn_Code	관측지점코드	max_Temp	최고기온(℃)	wind	풍속(m/s)	condens_Time	결로시간(MM)
stn_Name	관측지점명	min_Temp	최저기온(℃)	rain	강수량(mm)	gr_Temp	초상온도(℃)
date	관측일자 (2015-01-01 ~ 2020-09-28)	hum	습도(%)	sun_Time	일조시간(MM)	soil_Temp	지중온도(℃)
temp	기온(℃)	widdir	풍향	sun_Qy	일사량	soil_Wt	토양수분보정값(%)

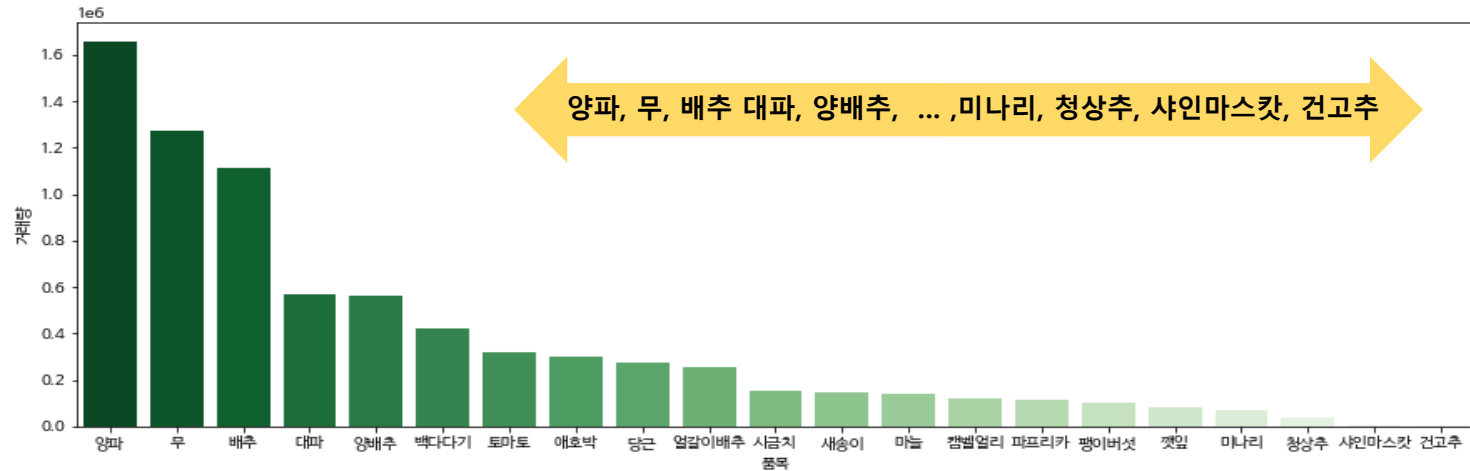


EDA

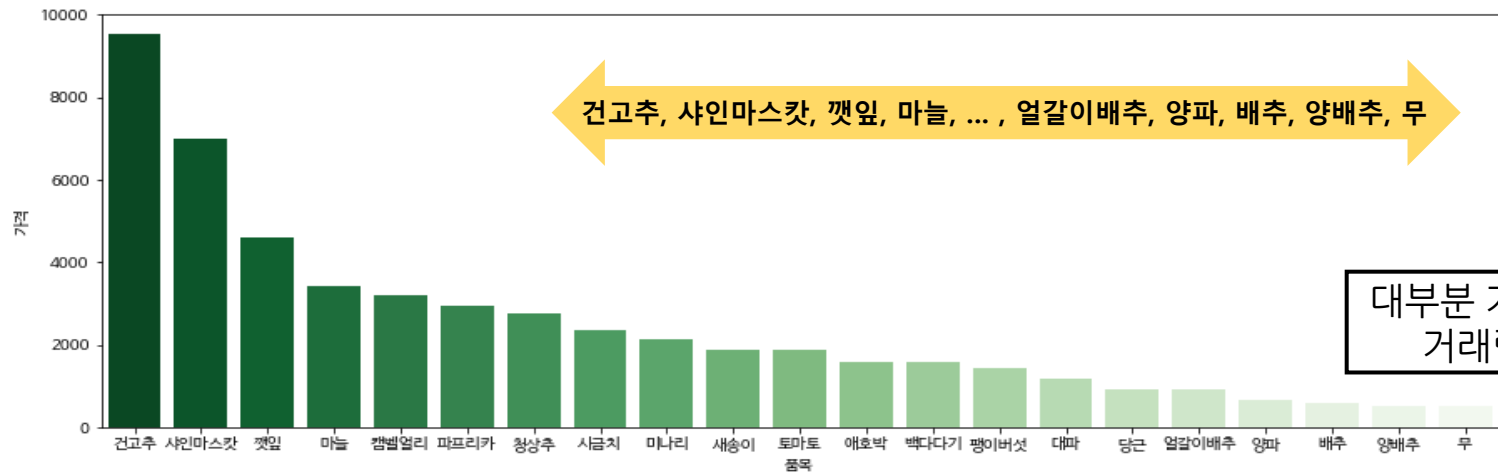
1. 기본 데이터 2. 비용 데이터 3. 기후 데이터 4. 품목 제거

1. 품목별 평균

거래량



가격



대부분 거래량이 많은 품목들이 가격이 낮고
거래량이 적은 품목들이 가격이 높다.

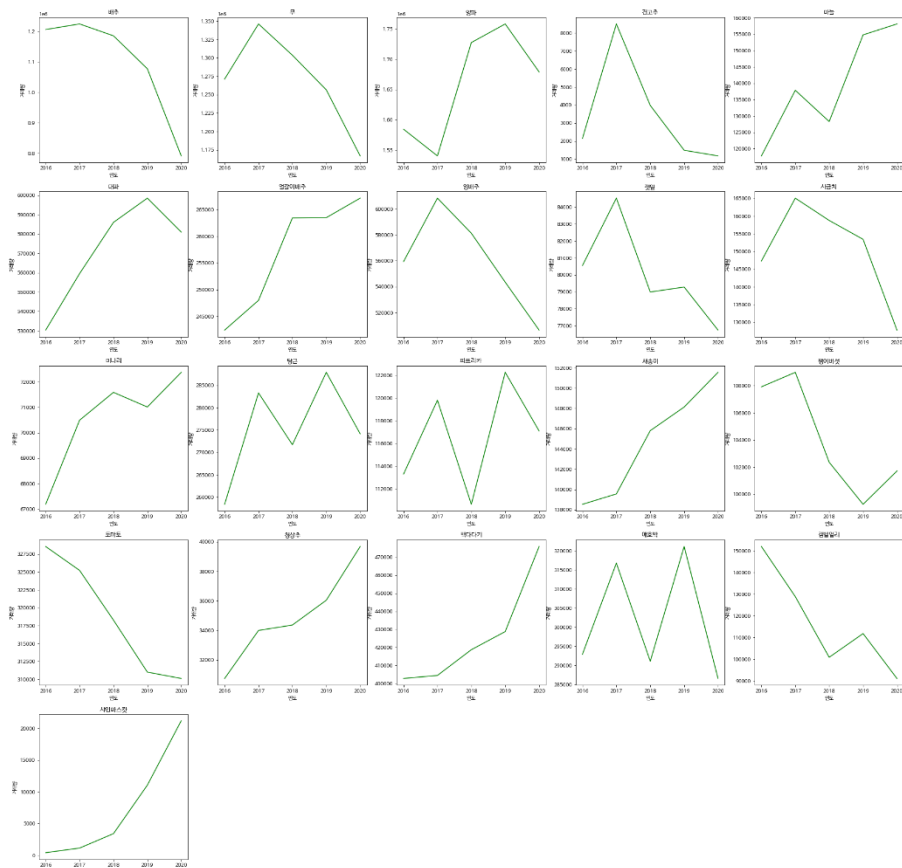


EDA

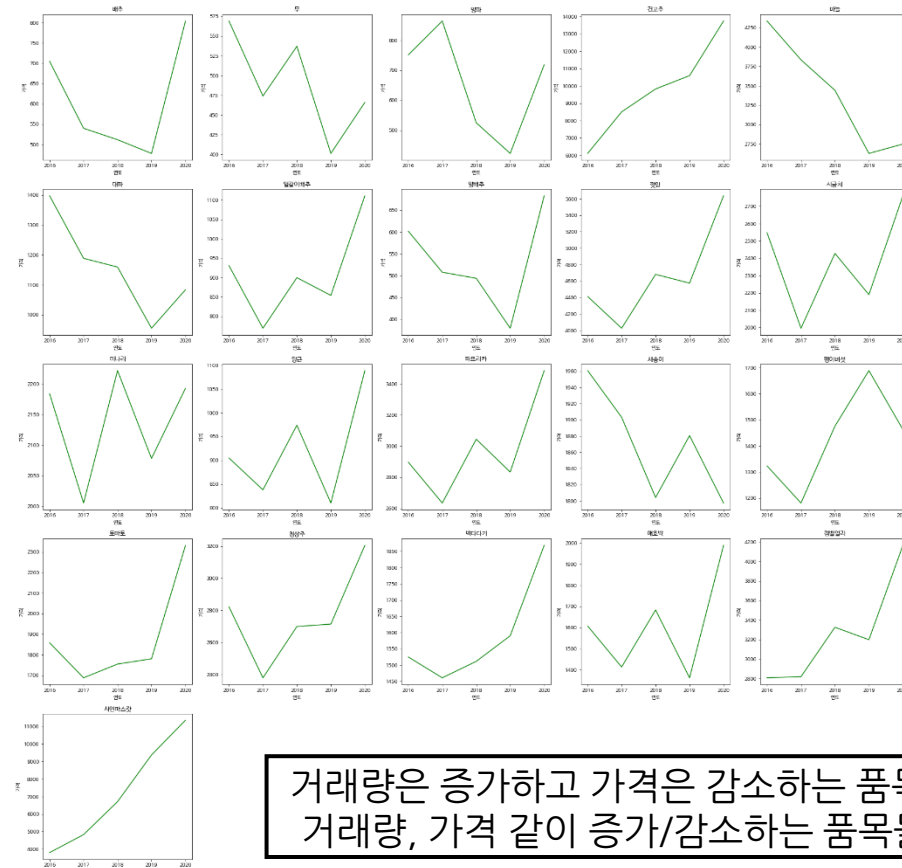
1. 기본 데이터 2. 비용데이터 3. 기후 데이터 4. 품목제거

2. 품목별 연도별 평균

거래량



가격



거래량은 증가하고 가격은 감소하는 품목들도 있고
거래량, 가격 같이 증가/감소하는 품목들도 있다.



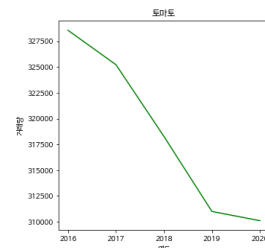
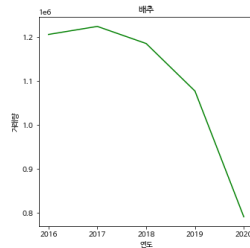
EDA

1. 기본 데이터 2. 비용 데이터 3. 기후 데이터 4. 품목 제거

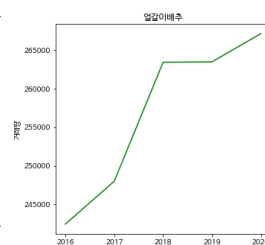
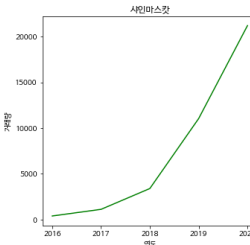
2. 품목별 연도별 평균 _ 특징

거래량

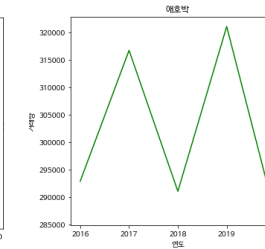
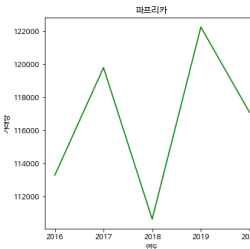
감소추이



증가추이

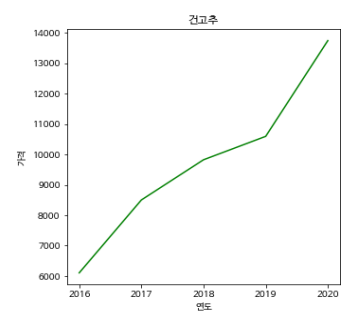
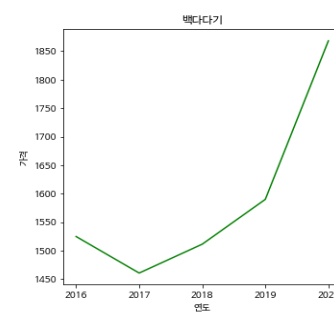
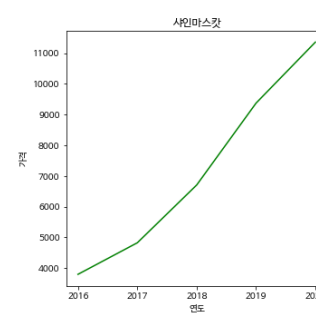


M자모양

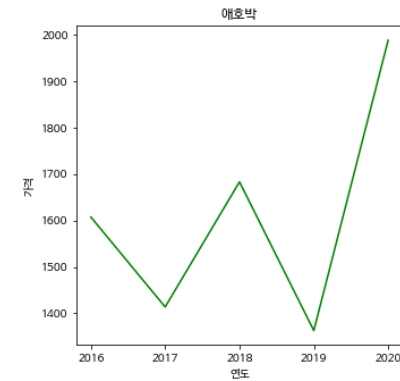
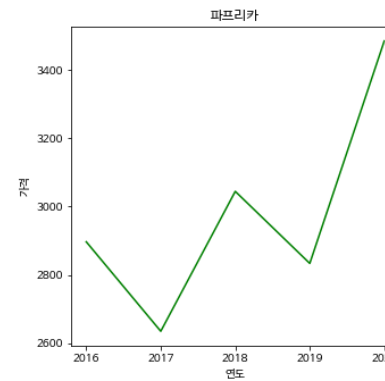


가격

증가추이



W자모양



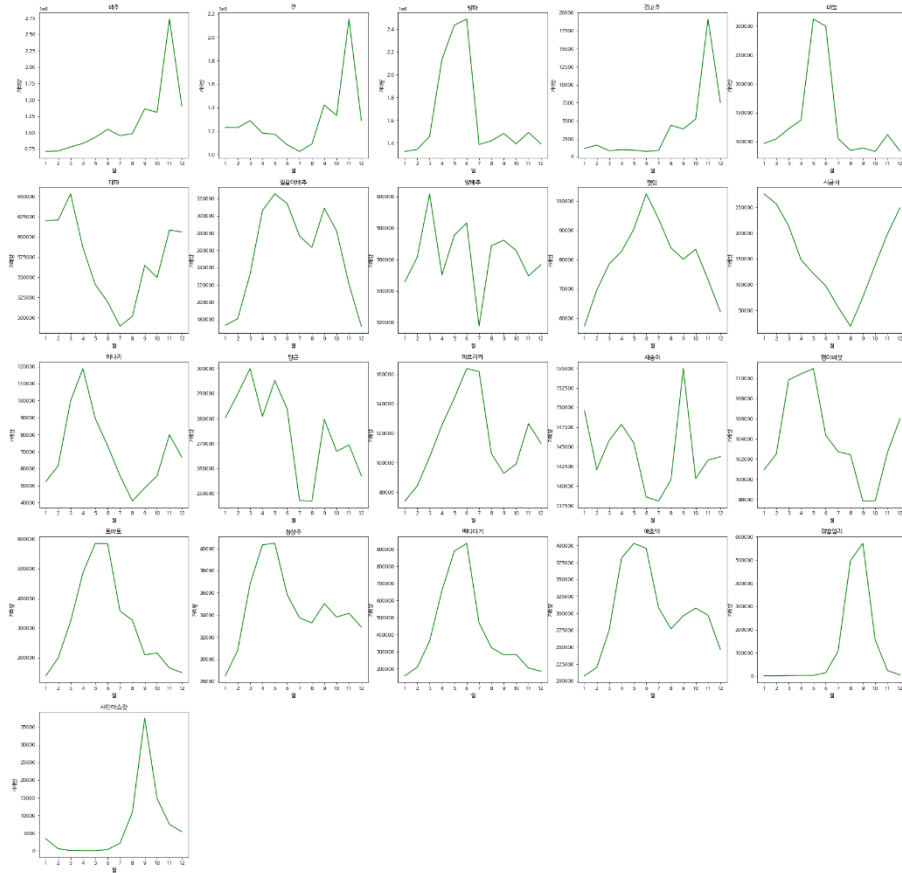


EDA

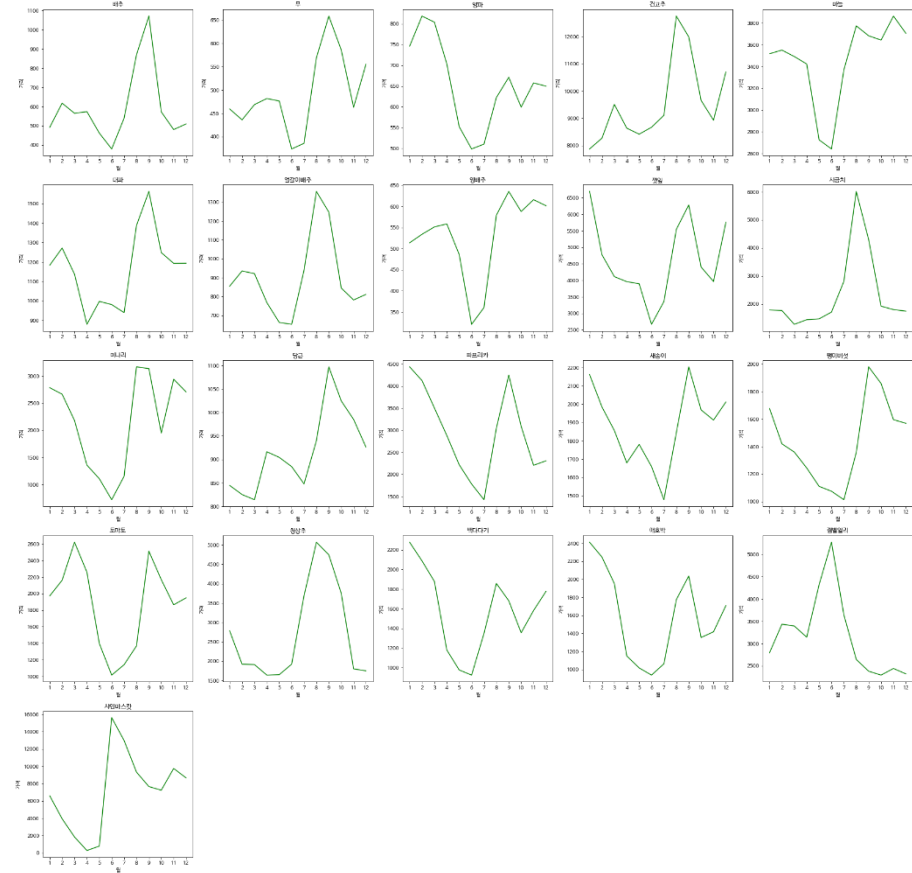
1. 기본 데이터 2. 비용 데이터 3. 기후 데이터 4. 품목 제거

3. 품목별 월별 평균

거래량



가격



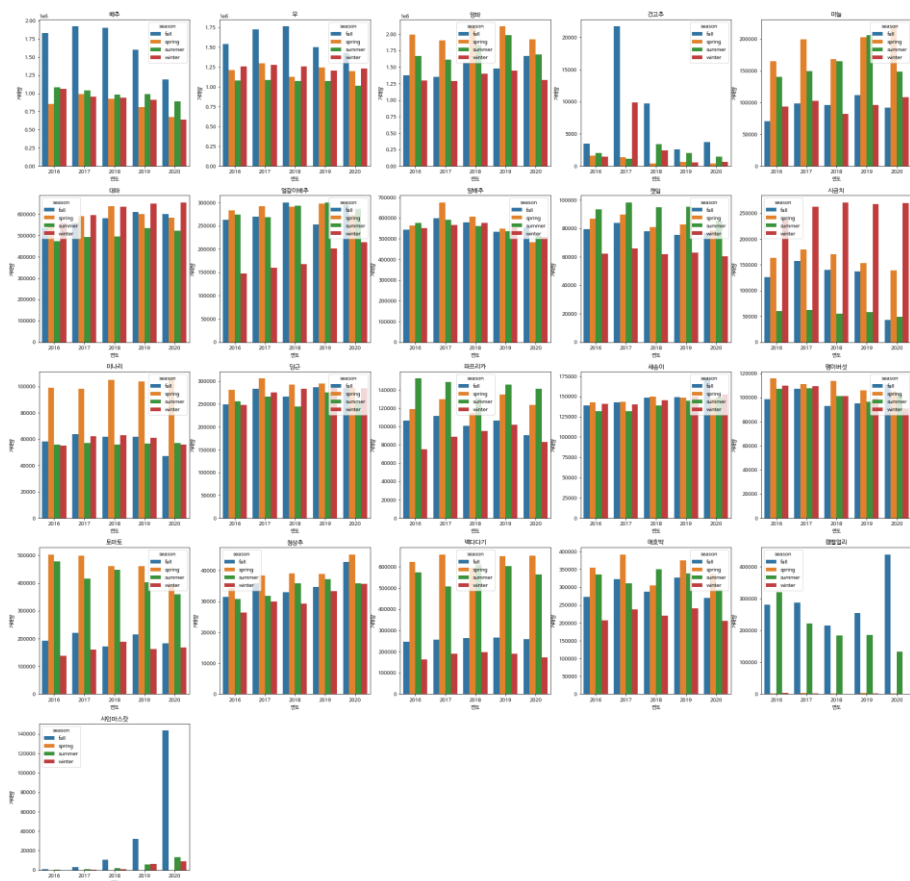


EDA

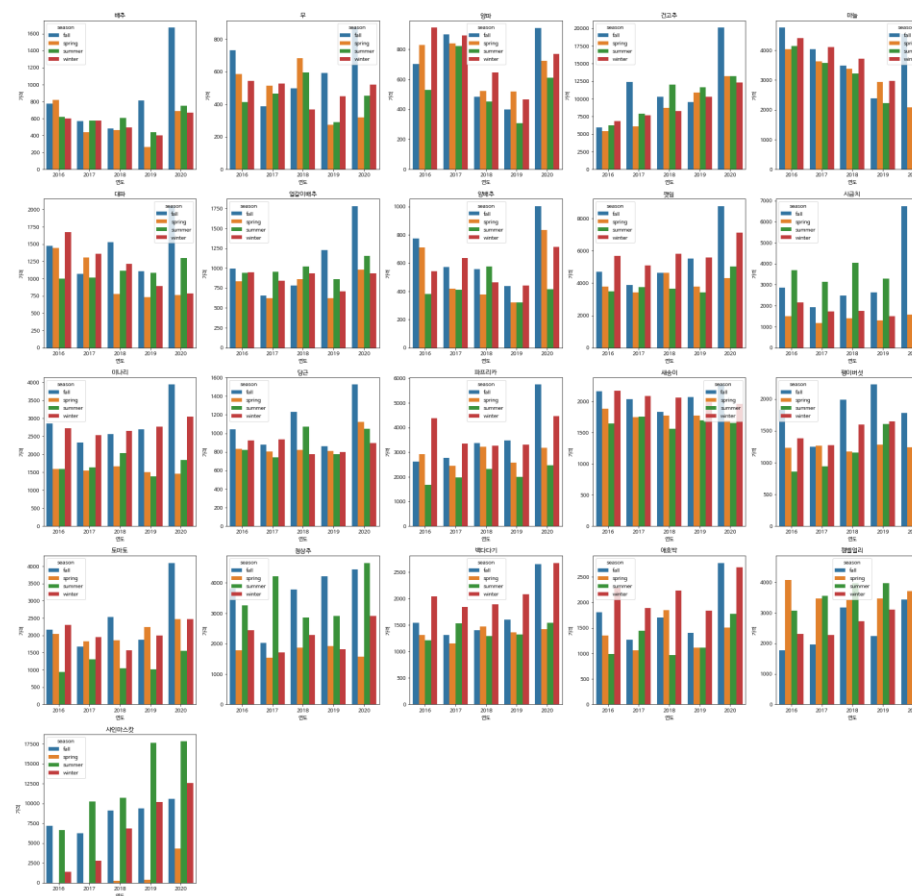
1. 기본 데이터 2. 비용 데이터 3. 기후 데이터 4. 품목 제거

4. 품목별 계절별 평균

거래량



가격





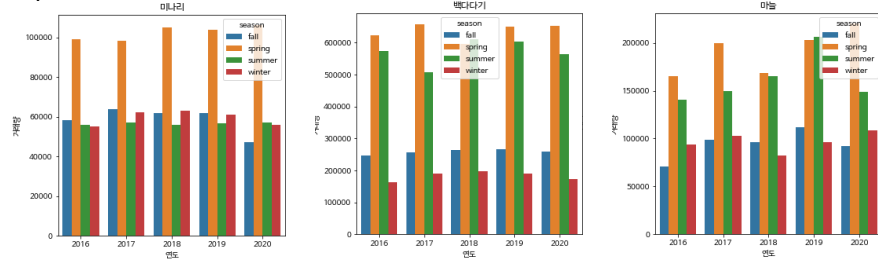
EDA

1. 기본 데이터 2. 비용 데이터 3. 기후 데이터 4. 품목 제거

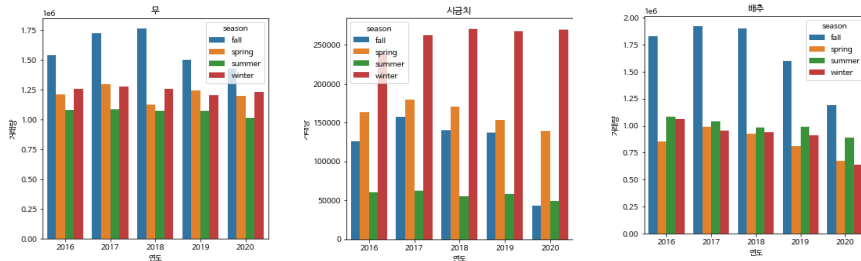
4. 품목별 계절별 평균 _ 특징

거래량

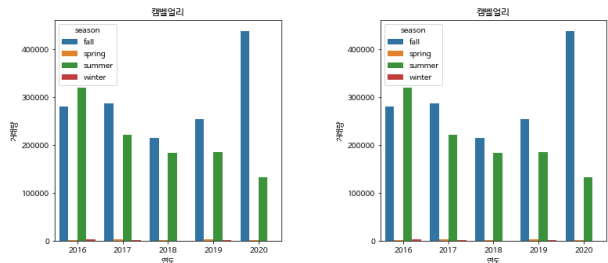
봄, 여름



가을, 겨울

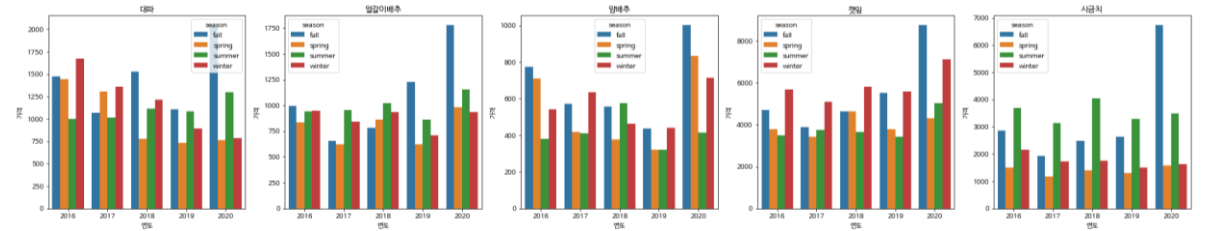


특정 계절



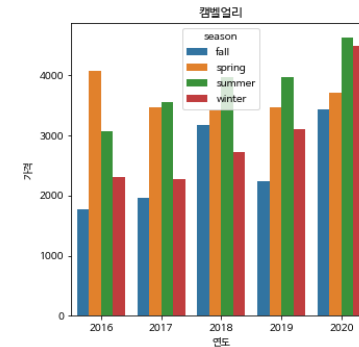
가격

가을

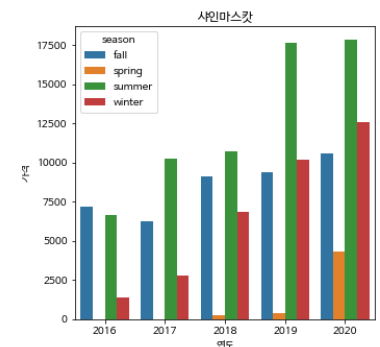


=> 대부분 품목이 가을에 가격이 제일 높다.

그 외 특징



=> 캠벨얼리는 봄, 여름에 가격이 높다.



=> 샤인머스캣은 여름, 가을에 가격이 높다.

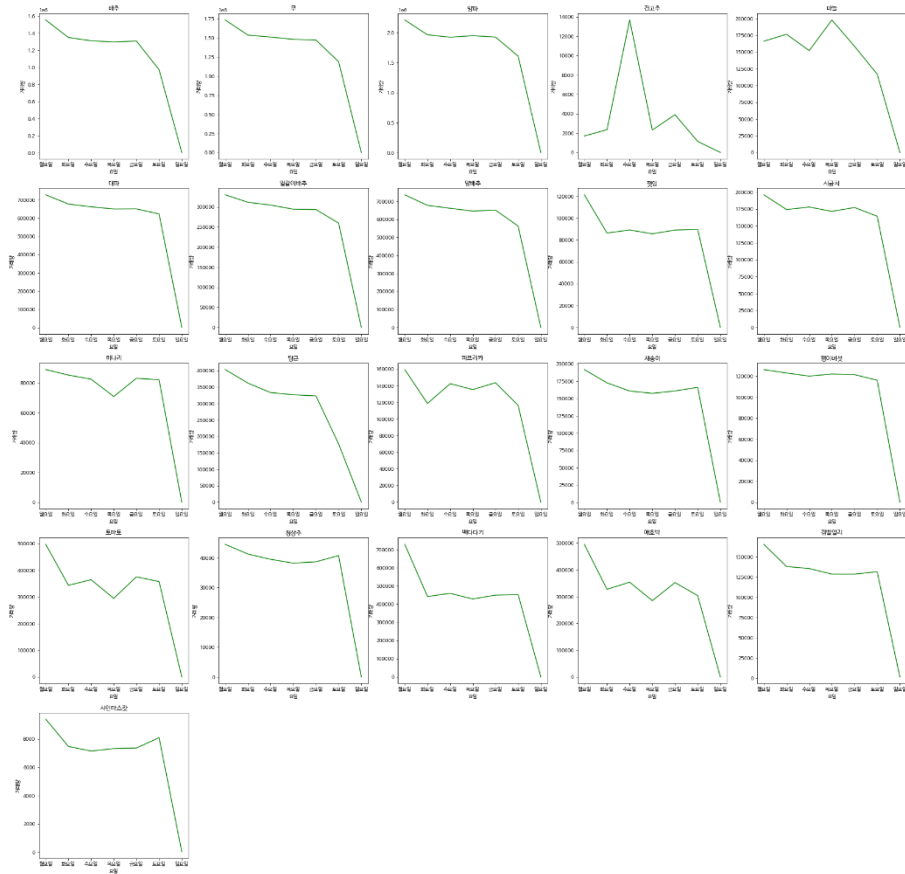


EDA

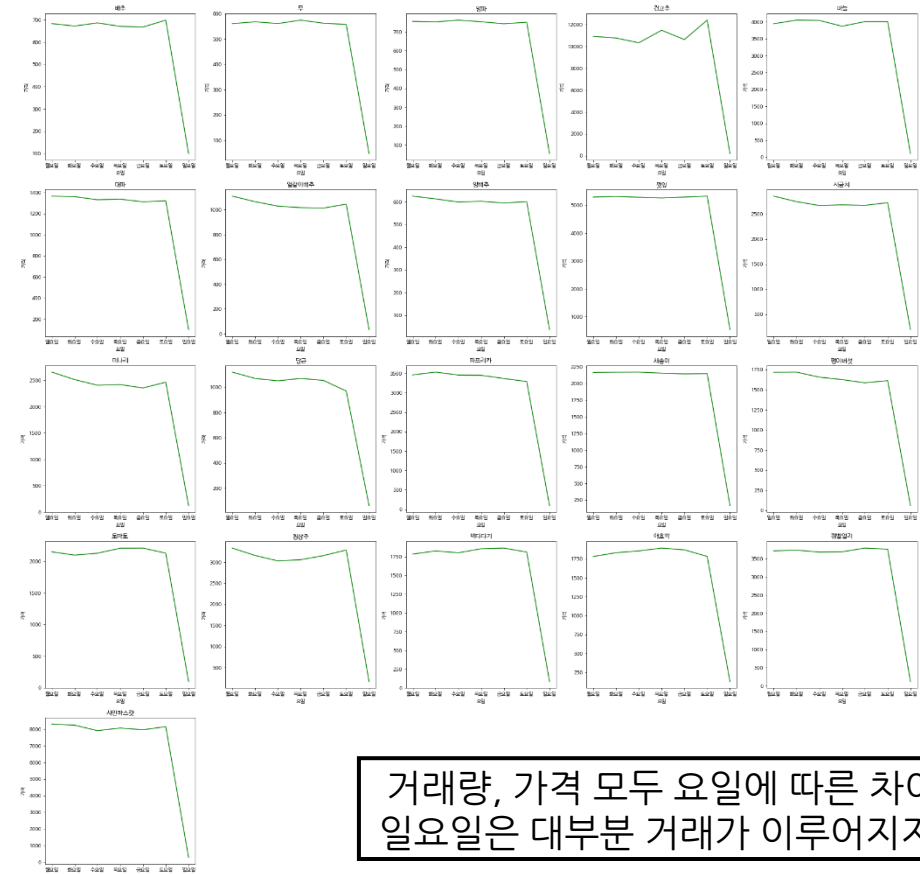
1. 기본 데이터 2. 비용 데이터 3. 기후 데이터 4. 품목 제거

5. 품목별 요일별 평균

거래량



가격



거래량, 가격 모두 요일에 따른 차이는 없고
일요일은 대부분 거래가 이루어지지 않는다.



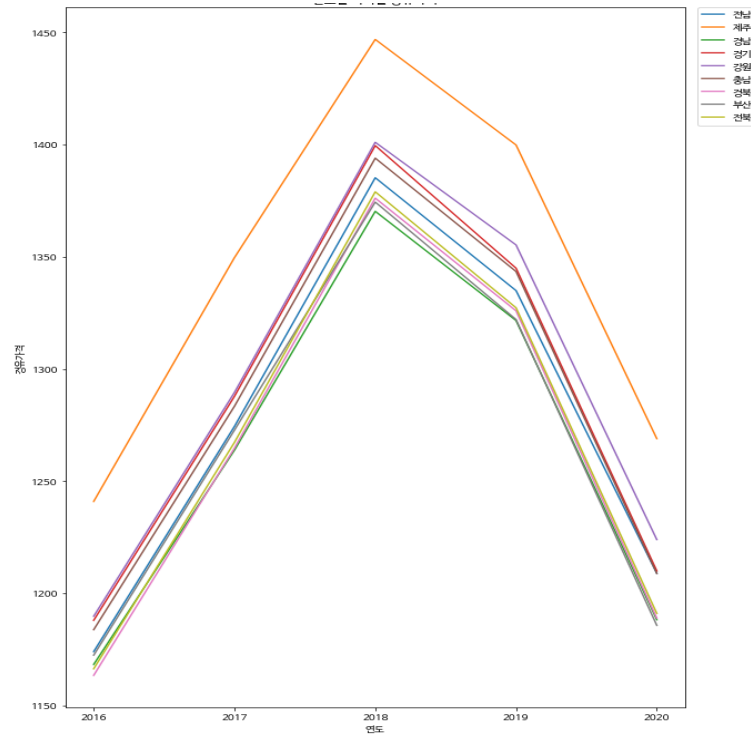
EDA

1. 기본 데이터 2. 비용데이터 3. 기후 데이터 4. 품목제거

1. 경유가격

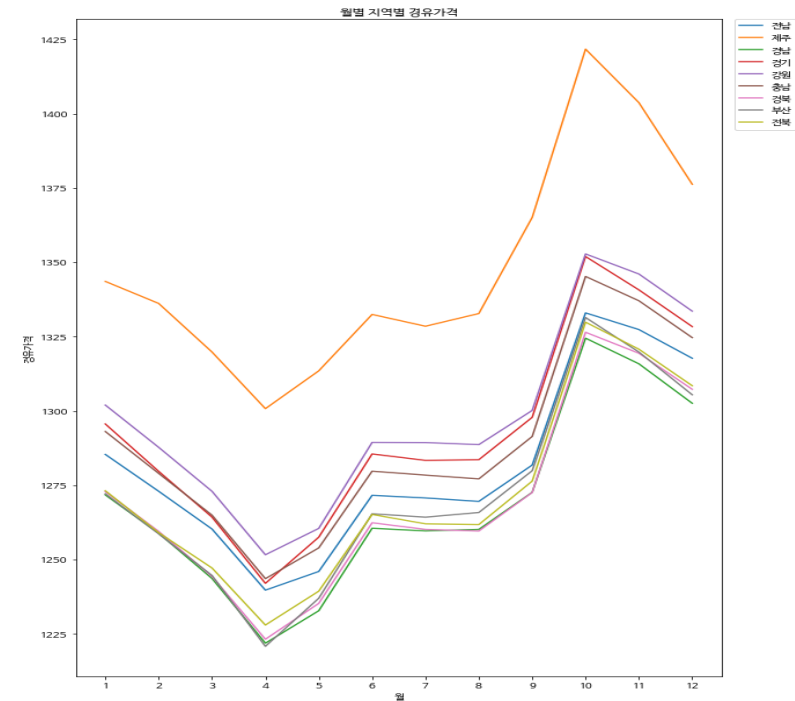
지역별로 거래량, 가격 데이터와 동기간 동안 기록한 일별 데이터

연도별 지역별 경유가격



2018년까지 상승하다가
이후 하락세를 보이고 있다.

월별 지역별 경유가격



제주도가 유독 가격이 높다.

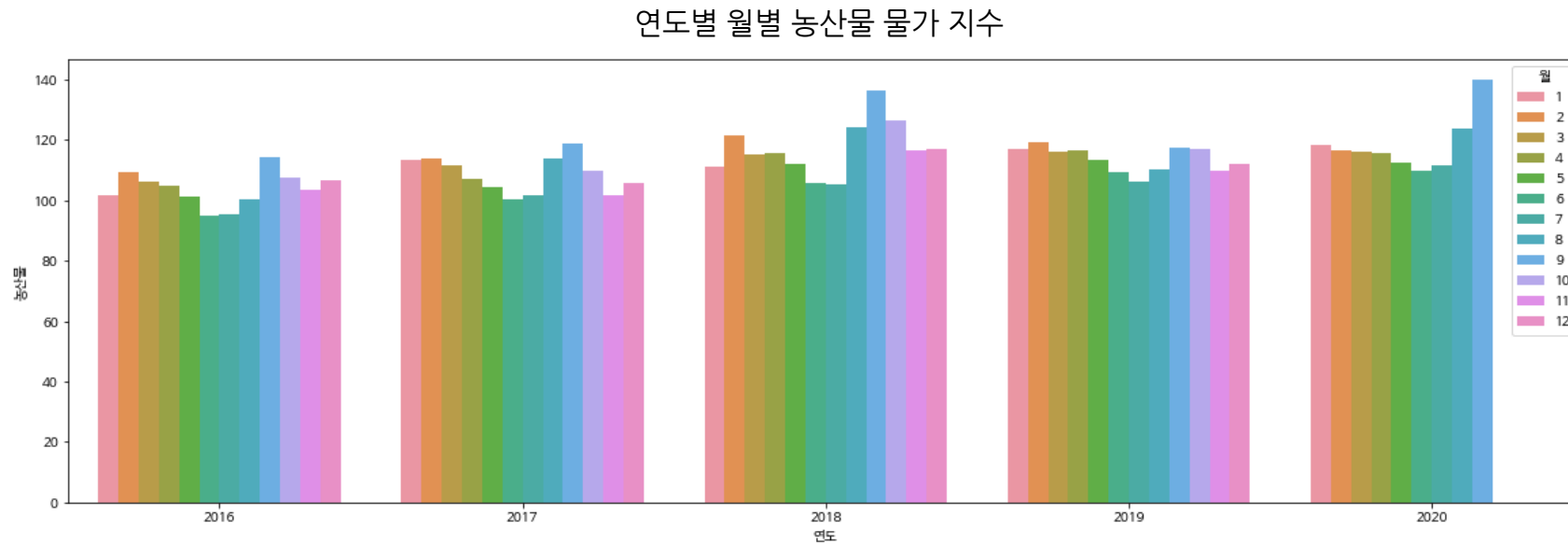


EDA

1. 기본 데이터 2. 비용데이터 3. 기후 데이터 4. 품목 제거

2. 농산물 물가 지수

농산물 분야 소비자 물가지수를 2016년 1월부터 2020년 9월까지 기록한 월별 데이터



상대적으로 9월에 농산물 물가지수가 높다.



EDA

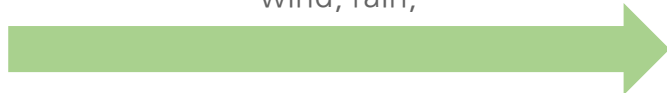
1. 기본 데이터 2. 비용 데이터 3. 기후 데이터 4. 품목 제거

1. 농업기상 관측데이터

- Column별 결측치 확인

no	0
stn_Code	0
stn_Name	0
date	0
temp	5
max_Temp	5
min_Temp	36
hum	745
widdir	3572
wind	0
rain	8
sun_Time	10726
sun_Qy	653
condens_Time	6261
gr_Temp	15368
soil_Temp	4678
soil_Wt	17013

결측치가 500개이하인 column 선택
→ temp, max_temp, min_temp,
wind, rain,



- 가격과 기후 데이터간의 상관계수

	temp	max_Temp	min_Temp	wind	rain	dif_Temp
배추 가격	0.267833	0.254643	0.284399	-0.301400	0.411909	-0.241794
무 가격	-0.033426	-0.045550	-0.013194	-0.015441	0.184418	-0.160946
양파 가격	-0.352335	-0.349387	-0.351849	0.140411	-0.351581	0.132914
전고추 가격	0.214736	0.202080	0.221612	0.142120	0.309928	-0.170724
마늘 가격	-0.151042	-0.168980	-0.124849	-0.107677	0.021093	-0.087088
대파 가격	0.001499	-0.009951	0.021446	-0.211537	0.195147	-0.159302
얼갈이배추 가격	0.243008	0.211028	0.278335	-0.313048	0.384305	-0.485771
양배추 가격	-0.206915	-0.210208	-0.204030	0.180247	0.022150	-0.074945
갯잎 가격	-0.286315	-0.323689	-0.238307	-0.074165	0.075405	-0.170459
시금치 가격	0.545510	0.515502	0.578824	-0.433500	0.458464	-0.553629
미나리 가격	-0.331038	-0.355643	-0.292436	-0.074708	-0.034281	-0.287361
당근 가격	0.107659	0.100760	0.115343	0.198679	0.104546	-0.046690
파프리카 가격	-0.423382	-0.411455	-0.428567	0.088799	-0.094584	0.327962
새송이 가격	-0.436890	-0.445448	-0.411563	-0.055372	-0.388383	-0.092596
팽이버섯 가격	-0.252106	-0.259579	-0.222676	-0.141467	-0.007586	-0.076227
토마토 가격	-0.397976	-0.397101	-0.402390	0.071706	-0.048935	0.329214
청상추 가격	0.468635	0.420051	0.517782	-0.198039	0.568636	-0.631595
백다다기 가격	-0.505473	-0.544357	-0.448152	0.045630	0.047008	-0.392331
애호박 가격	-0.470959	-0.510518	-0.422065	0.046741	-0.017105	-0.070371
캠벨얼리 가격	0.289268	0.300998	0.257374	0.009319	0.144928	0.092524
샤인마스켓 가격	0.307414	0.278644	0.355163	-0.440312	0.381545	-0.494878

→ 전체적으로 뚜렷한 선형관계가 보이지 않는다.

→ 따라서 비선형 모델도 고려하기로 한다.



EDA

1. 기본 데이터 2. 비용 데이터 3. 기후 데이터 4. 품목 제거

- 제주도가 주산지인 품목의 결측치 확인

연도	월	
2016	1	75
	2	84
	3	3
	12	21
2017	1	12
	2	6
	7	78
	8	21
2018	5	3
2020	2	3

제주도의 2016년 1, 2월, 2017년 7월 대부분의 데이터는 결측치



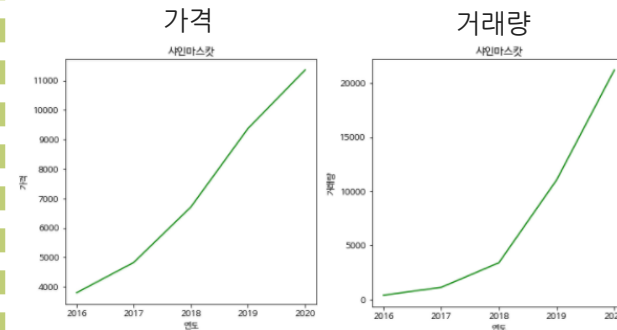
당근 의 결측치	
temp	102
max_Temp	102
min_Temp	133
wind	102
rain	102
dif_Temp	133

양배추 의 결측치	
temp	102
max_Temp	102
min_Temp	133
wind	102
rain	102
dif_Temp	133

무 의 결측치	
temp	102
max_Temp	102
min_Temp	133
wind	102
rain	102
dif_Temp	133

제주도가 주산지인 당근, 양배추, 무 품목은 결측치가 많다.
→ 당근, 양배추, 무 품목은 제외한다.

- 샤인머스켓 가격, 거래량 확인



샤인머스켓은 2016년부터 급격한
가격과 거래량 상승을 보임

date	요일	품목	가격	거래량
2016-01-01	금요일	샤인머스켓	0.0	0.0
2016-01-02	토요일	샤인머스켓	0.0	0.0
2016-01-03	일요일	샤인머스켓	0.0	0.0
2016-01-04	월요일	샤인머스켓	0.0	0.0
2016-01-05	화요일	샤인머스켓	0.0	0.0
...
2016-12-27	화요일	샤인머스켓	0.0	0.0
2016-12-28	수요일	샤인머스켓	0.0	0.0
2016-12-29	목요일	샤인머스켓	0.0	0.0
2016-12-30	금요일	샤인머스켓	0.0	0.0
2016-12-31	토요일	샤인머스켓	0.0	0.0

이는 2016년 모든 데이터의
가격과 거래량이 0이기 때문



0을 대체할 값이 없으므로
샤인머스켓 품목은 제외한다.



1. 기본 데이터 핸들링


가격

	date	요일	매주 가격 (원/kg)	무 가격 (원/kg)	양파 가격 (원/kg)	건고추 가격 (원/kg)
0	2016-01-01	단요일	0.0	0.0	0.0	0.0
1	2016-01-02	바다일	329.0	360.0	1281.0	11000.0
2	2016-01-03	전요일	0.0	0.0	0.0	0.0
3	2016-01-04	화요일	478.0	382.0	1235.0	4464.0
4	2016-01-05	수요일	442.0	422.0	1213.0	4342.0



거래량

	date	요일	배추_거래 량(kg)	무_거래량 (kg)	양파_거래 량(kg)	건고추_거 래량(kg)
0	2016-01-01	금요일	0.0	0.0	0.0	0.0
1	2016-01-02	토요일	80860.0	80272.0	122787.5	3.0
2	2016-01-03	일요일	0.0	0.0	0.0	0.0
3	2016-01-04	월요일	1422742.5	1699653.7	2315079.0	699.0
4	2016-01-05	화요일	1167241.0	1423482.3	2092960.1	1112.6

1. date와 요일 column을
기준으로 melt
 2. melt한 dataframe
합치기
 3. date를 연도, 월, 일로
나누기
- 
4. 각 품목의 **주산지** 매핑

경기: 열갈이배추, 시금치, 미나리
강원: 파프리카
충남: 새송이, 백다다기
전북: 청상추
전남: 배추, 양파, 건고추, 대파
경북: 팽이버섯, 캠벨얼리
경남: 마늘, 깻잎, 애호박
부산: 토마토

[illegible]



데이터 통합

1. 기본 데이터 2. 비용데이터 3. 기후 데이터

2. 비용 데이터 붙이기

경유가격

	구분	서울	부산	대구	인천	광주	대전
0	2016년01월01일	1277.10	1175.02	1165.12	1174.31	1174.91	1183.53
1	2016년01월02일	1275.54	1174.62	1164.64	1174.31	1175.90	1182.80
2	2016년01월03일	1275.06	1174.82	1163.69	1173.55	1174.13	1181.91
3	2016년01월04일	1274.66	1173.94	1162.00	1173.35	1172.20	1181.39
4	2016년01월05일	1273.53	1171.30	1160.57	1172.87	1171.27	1178.87

농산물 물가 지수

	시점	농산물
0	2016. 01	101.55
1	2016. 02	109.33
2	2016. 03	106.08
3	2016. 04	104.95
4	2016. 05	101.48

1. 구분 → 연도, 월, 일로 분리
2. melt
3. 연도, 월, 일, 주산지를 기준으로 merge



1. 시점 → 연도, 월로 분리
2. 연도, 월을 기준으로 merge



	date	요일	품목	가격	거래량	연도	월	일	주산지
0	2016-01-01	금요일	배추	0.0	0.0	2016	1	1	전남
1	2016-01-02	토요일	배추	329.0	80860.0	2016	1	2	전남
2	2016-01-03	일요일	배추	0.0	0.0	2016	1	3	전남
3	2016-01-04	월요일	배추	478.0	1422742.5	2016	1	4	전남
4	2016-01-05	화요일	배추	442.0	1167241.0	2016	1	5	전남

	date	요일	품목	가격	거래량	연도	월	일	주산지	경유가격	농산물
0	2016-01-01	금요일	배추	0.0	0.0	2016	1	1	전남	1191.69	101.55
1	2016-01-02	토요일	배추	329.0	80860.0	2016	1	2	전남	1190.58	101.55
2	2016-01-03	일요일	배추	0.0	0.0	2016	1	3	전남	1189.58	101.55
3	2016-01-04	월요일	배추	478.0	1422742.5	2016	1	4	전남	1187.96	101.55
4	2016-01-05	화요일	배추	442.0	1167241.0	2016	1	5	전남	1185.81	101.55
...	-	-	-	-	-	-	-	-	-	-	-
29456	2020-09-24	목요일	캠벨얼리	3620.0	504242.6	2020	9	24	경북	1128.91	139.93
29457	2020-09-25	금요일	캠벨얼리	3618.0	479683.1	2020	9	25	경북	1128.43	139.93
29458	2020-09-26	토요일	캠벨얼리	3691.0	521493.8	2020	9	26	경북	1127.22	139.93
29459	2020-09-27	일요일	캠벨얼리	3567.0	21717.0	2020	9	27	경북	1126.94	139.93
29460	2020-09-28	월요일	캠벨얼리	3761.0	601841.0	2020	9	28	경북	1126.46	139.93

29461 rows × 11 columns



데이터 통합

1. 기본 데이터 2. 비용 데이터 3. 기후 데이터

3. 기후 데이터 붙이기 (1)

	stn_Code	stn_Name	date	temp	max_Temp	min_Temp	hum	widdir	wind	rain	sun_Time	sun_Qy	condens_Time	gr_Temp	soil_Temp	soil_Wt
0	536824B002	해남군 옥천면	2015-01-01	-1.3	0.6	-2.9	80.0	295.2	2.3	0.8	NaN	7.8	NaN	NaN	3.36	25.9
1	330846A001	천안시 목천읍	2015-01-01	-6.2	-3.8	-8.3	NaN	NaN	0.0	0.0	NaN	NaN	1429.0	NaN	NaN	NaN
2	627911A001	밀양시 상남면	2015-01-01	-3.2	0.2	-7.2	40.1	282.7	2.9	0.0	516.0	11.0	0.0	NaN	2.20	28.5
3	539823A001	진도군 군내면	2015-01-01	-0.8	1.6	-2.8	79.2	257.0	3.5	1.5	217.0	8.2	652.0	NaN	5.02	30.6
4	590823A001	남원시 이백면	2015-01-01	-4.1	-1.3	-6.0	60.7	286.7	2.1	0.5	310.0	7.7	0.0	-4.3	2.16	20.3

2. stn_Name → 시도명으로 바꾼 값으로 '주산지' column 생성

- 해남군 옥천면 → 전남
- 천안시 목천읍 → 충남
- 밀양시 상남면 → 경남
- 진도군 군내면 → 전남
- 남원시 이백면 → 전북
- 시흥시 하중동 → 경기
- 상주시 초산동 → 경북
- 창녕군 대지면 → 경남
- 부산시 강서구 → 부산
- 포천시 신북면 → 경기
- 철원군 동송읍 → 강원
- 청도군 화양읍 → 경북
- 진주시 초전동 → 경남
- 무안군 청계면 → 전남
- 상주시 공성면 → 경북
- 제주시 애월읍 → 제주

1. 결측치가 500개이하인 column 선택
→ temp , max_temp , min_temp ,
wind, rain,

```

no      0
stn_Code 0
stn_Name 0
date     0
temp     5
max_Temp 5
min_Temp 36
hum      745
widdir   3572
wind     0
rain     8
sun_Time 10726
sun_Qy   653
condens_Time 6261
gr_Temp  15368
soil_Temp 4678
soil_Wt  17013

```

	stn_Name	date	temp	max_Temp	min_Temp	wind	rain	주산지
0	해남군 옥천면	2015-01-01	-1.3	0.6	-2.9	2.3	0.8	전남
1	천안시 목천읍	2015-01-01	-6.2	-3.8	-8.3	0.0	0.0	충남
2	밀양시 상남면	2015-01-01	-3.2	0.2	-7.2	2.9	0.0	경남
3	진도군 군내면	2015-01-01	-0.8	1.6	-2.8	3.5	1.5	전남
4	남원시 이백면	2015-01-01	-4.1	-1.3	-6.0	2.1	0.5	전북



데이터 통합

1. 기본 데이터 2. 비용 데이터 3. 기후 데이터

3. 기후 데이터 붙이기 (2)

	stn_Name	date	temp	max_Temp	min_Temp	wind	rain	주산지
0	해남군 옥천면	2015-01-01	-1.3	0.6	-2.9	2.3	0.8	전남
1	천안시 목천읍	2015-01-01	-6.2	-3.8	-8.3	0.0	0.0	충남
2	밀양시 상남면	2015-01-01	-3.2	0.2	-7.2	2.9	0.0	경남
3	진도군 군내면	2015-01-01	-0.8	1.6	-2.8	3.5	1.5	전남
4	남원시 이백면	2015-01-01	-4.1	-1.3	-6.0	2.1	0.5	전북

1. '주산지'와 'date'를 기준으로 그룹으로 묶고
나머지 기후 변수들의 평균 구하기.

2. diff_Temp(일교차) 변수 생성
 $\text{diff_temp} = \text{max_Temp} - \text{min_Temp}$

	주산지	date	temp	max_Temp	min_Temp	wind	rain	dif_Temp
0	강원	2015-01-01	-8.5	-5.7	-10.7	3.0	0.0	5.0
1	강원	2015-01-02	-8.0	-2.8	-12.5	2.0	0.0	9.7
2	강원	2015-01-03	-6.4	1.1	-15.5	1.5	0.0	16.6
3	강원	2015-01-04	0.5	5.2	-4.9	1.1	0.0	10.1
4	강원	2015-01-05	-0.7	6.2	-6.9	0.5	1.0	13.1



데이터 통합

1. 기본 데이터 2. 비용 데이터 3. 기후 데이터

3. 기후 데이터 붙이기 (3)

1. date → 연도, 월, 일로 분리

2. 연도, 월, 일, 주산지를 기준으로 merge

	주산지	date	temp	max_Temp	min_Temp	wind	rain	dif_Temp
0	강원	2015-01-01	-8.5	-5.7	-10.7	3.0	0.0	5.0
1	강원	2015-01-02	-8.0	-2.8	-12.5	2.0	0.0	9.7
2	강원	2015-01-03	-6.4	1.1	-15.5	1.5	0.0	16.6
3	강원	2015-01-04	0.5	5.2	-4.9	1.1	0.0	10.1
4	강원	2015-01-05	-0.7	6.2	-6.9	0.5	1.0	13.1



	date	요일	품목	가격	거래량	연도	월	일	주산지	경유가격	농산물
0	2016-01-01	금요일	배추	0.0	0.0	2016	1	1	전남	1191.69	101.55
1	2016-01-02	토요일	배추	329.0	80860.0	2016	1	2	전남	1190.58	101.55
2	2016-01-03	일요일	배추	0.0	0.0	2016	1	3	전남	1189.58	101.55
3	2016-01-04	월요일	배추	478.0	1422742.5	2016	1	4	전남	1187.96	101.55
4	2016-01-05	화요일	배추	442.0	1167241.0	2016	1	5	전남	1185.81	101.55

	date	요일	품목	가격	거래량	주산지	경유가격	농산물	temp	max_Temp	min_Temp	wind	rain	dif_Temp
0	2016-01-01	금요일	배추	0.0	0.0	전남	1191.69	101.55	2.700000	9.300000	-3.300000	0.366667	0.033333	12.600000
1	2016-01-02	토요일	배추	329.0	80860.0	전남	1190.58	101.55	6.133333	13.800000	0.233333	0.333333	0.000000	13.566667
2	2016-01-03	일요일	배추	0.0	0.0	전남	1189.58	101.55	6.666667	15.233333	2.133333	0.333333	0.000000	13.100000
3	2016-01-04	월요일	배추	478.0	1422742.5	전남	1187.96	101.55	6.133333	11.233333	-0.633333	1.100000	0.000000	11.866667
4	2016-01-05	화요일	배추	442.0	1167241.0	전남	1185.81	101.55	1.700000	4.066667	-1.300000	0.400000	1.466667	5.366667
...
29456	2020-09-24	목요일	캠벨얼리	3620.0	504242.6	경북	1128.91	139.93	18.500000	24.133333	13.500000	0.500000	0.000000	10.633333
29457	2020-09-25	금요일	캠벨얼리	3618.0	479683.1	경북	1128.43	139.93	18.300000	25.100000	12.533333	0.400000	0.000000	12.566667
29458	2020-09-26	토요일	캠벨얼리	3691.0	521493.8	경북	1127.22	139.93	17.500000	25.000000	11.700000	0.433333	0.000000	13.300000
29459	2020-09-27	일요일	캠벨얼리	3567.0	21717.0	경북	1126.94	139.93	17.600000	25.033333	11.200000	0.466667	0.000000	13.833333
29460	2020-09-28	월요일	캠벨얼리	3761.0	601841.0	경북	1126.46	139.93	15.466667	24.600000	9.333333	0.333333	0.000000	15.266667

29461 rows × 14 columns



데이터 통합

1. 기본 데이터 2. 비용 데이터 3. 기후 데이터

4. 일주일 후 가격 붙이기

각 품목의 일주일 후 가격 → 1 week 변수에 추가

	date	품목	가격
0	2016-01-01	배추	0.0
1	2016-01-02	배추	329.0
2	2016-01-03	배추	0.0
3	2016-01-04	배추	478.0
4	2016-01-05	배추	442.0
5	2016-01-06	배추	442.0
6	2016-01-07	배추	448.0
7	2016-01-08	배추	420.0
8	2016-01-09	배추	389.0
9	2016-01-10	배추	0.0
10	2016-01-11	배추	398.0
11	2016-01-12	배추	431.0
12	2016-01-13	배추	429.0

배추의 일주일 후 가격

	date	가격
0	2016-01-01	0.0
1	2016-01-02	329.0
2	2016-01-03	0.0
3	2016-01-04	478.0
4	2016-01-05	442.0
5	2016-01-06	442.0
6	2016-01-07	448.0
7	2016-01-08	420.0
8	2016-01-09	389.0
9	2016-01-10	0.0
10	2016-01-11	398.0
11	2016-01-12	431.0
12	2016-01-13	429.0
13	2016-01-14	441.0
14	2016-01-15	449.0
15	2016-01-16	454.0
16	2016-01-17	0.0
17	2016-01-18	475.0
18	2016-01-19	511.0
19	2016-01-20	511.0

1_week 변수에 추가

	date	품목	가격	1_week
0	2016-01-01	배추	0.0	420.0
1	2016-01-02	배추	329.0	389.0
2	2016-01-03	배추	0.0	0.0
3	2016-01-04	배추	478.0	398.0
4	2016-01-05	배추	442.0	431.0
5	2016-01-06	배추	442.0	429.0
6	2016-01-07	배추	448.0	441.0
7	2016-01-08	배추	420.0	449.0
8	2016-01-09	배추	389.0	454.0
9	2016-01-10	배추	0.0	0.0
10	2016-01-11	배추	398.0	475.0
11	2016-01-12	배추	431.0	511.0
12	2016-01-13	배추	429.0	511.0



5. 최종 분석 데이터 셋

29461 rows x 15 columns

모델링

1. 선형 모델 : OLS , VAR
2. 비선형 모델 : Tree 모델 , 신경망 모델



모델링

목표 : 각 품목의 일주일 후 가격(1_week)을 예측하자 !

모델링 평가 지표

$$\text{NMAE} = \frac{|pred - true|}{true}$$

(Normalized Mean Absolute Error)

Pred : 일주일 후 가격 예측 값
True : 실제 값

Why? 품종별로 가격 스케일에 차이가 존재하기 때문



모델링

1. 선형 모델 : OLS, VAR

2. 비선형 모델 : Tree 모델, 신경망 모델

1. OLS 란 ?

다중 회귀모형 : 한 개의 종속변수, 두개 이상의 독립변수

$$y = \mathbf{X}\beta + \varepsilon$$

$$\begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix} = \begin{pmatrix} 1 & x_{11} & x_{12} & \cdots & x_{1k} \\ 1 & x_{21} & x_{22} & \cdots & x_{2k} \\ \vdots & \vdots & \vdots & & \vdots \\ 1 & x_{n1} & x_{n2} & \cdots & x_{nk} \end{pmatrix} \begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_k \end{pmatrix} + \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{pmatrix}$$

선형 관계를 가장 잘 나타낼 수 있는 회귀 계수 β 추정
그 중 한가지 방법이 바로 OLS

OLS (Ordinary Least Squares)

선형 회귀 모델 추정 기법 중에서 가장 단순하고 보편적인 기법
오차 제곱 합을 최소화하는 회귀계수를 추정하는 기법

statsmodels 패키지에서는 OLS 클래스를 사용하여 선형 회귀분석을 실시 !



모델링

1. 선형 모델 : OLS , VAR 2. 비선형 모델 : Tree 모델 , 신경망 모델

2. 분석과정

Ex) 배추

1. 결측치 처리 : 이전값으로 채우기

2. 배추 데이터 셋 나누기 → train, test 분리

	date	요일	품목	가격	거래량	주산지	농유가격	농산물	temp	max_Temp	min_Temp	wind	rain	dif_Temp	1_week
0	2016-01-01	금요일	배추	0.0	0.0	전남	1191.69	101.55	2.700000	9.300000	-3.300000	0.366667	0.033333	12.600000	420.0
1	2016-01-02	토요일	배추	329.0	80860.0	전남	1190.58	101.55	6.133333	13.800000	0.233333	0.333333	0.000000	13.566667	389.0
2	2016-01-03	일요일	배추	0.0	0.0	전남	1189.58	101.55	6.666667	15.233333	2.133333	0.333333	0.000000	13.100000	0.0
3	2016-01-04	월요일	배추	478.0	1422742.5	전남	1187.96	101.55	6.133333	11.233333	-0.633333	1.100000	0.000000	11.866667	398.0
4	2016-01-05	화요일	배추	442.0	1167241.0	전남	1185.81	101.55	1.700000	4.066667	-1.300000	0.400000	1.466667	5.366667	431.0
...
1728	2020-09-24	목요일	배추	1839.0	1856965.0	전남	1150.27	139.93	20.100000	26.166667	15.333333	0.866667	0.166667	10.833333	0.0
1729	2020-09-25	금요일	배추	1789.0	1880095.5	전남	1149.58	139.93	20.300000	25.800000	14.666667	0.866667	0.000000	11.133333	0.0
1730	2020-09-26	토요일	배추	1760.0	1661090.9	전남	1149.28	139.93	20.033333	25.666667	15.866667	0.566667	1.000000	9.800000	0.0
1731	2020-09-27	일요일	배추	3066.0	25396.0	전남	1149.03	139.93	19.700000	26.066667	14.700000	0.733333	0.000000	11.366667	0.0
1732	2020-09-28	월요일	배추	1867.0	2405051.9	전남	1148.49	139.93	18.600000	25.600000	13.366667	0.600000	0.000000	12.233333	0.0

1733 rows x 15 columns

독립변수 X



1_week값 로그 변환

1_week
6.042633
5.966147
0.000000
5.988961
6.068426
...
0.000000
0.000000
0.000000
0.000000
0.000000

종속변수 Y



모델링

1. 선형 모델 : OLS , VAR

2. 비선형 모델 : Tree 모델 , 신경망 모델

3. 결과해석

배추

OLS Regression Results						
Dep. Variable:	1_week	R-squared (uncentered):	0.912			
Model:	OLS	Adj. R-squared (uncentered):	0.911			
Method:	Least Squares	F-statistic:	1561.			
Date:	Sat, 13 Nov 2021	Prob (F-statistic):	0.00			
Time:	02:44:07	Log-Likelihood:	-2706.7			
No. Observations:	1368	AIC:	5431.			
Df Residuals:	1359	BIC:	5478.			
Df Model:	9					
Covariance Type:	nonrobust					
	coef	std err	t	P> t	[0.025	0.975]
가격	0.0033	0.000	23.129	0.000	0.003	0.004
거래량	7.903e-07	5.96e-08	13.263	0.000	6.73e-07	9.07e-07
경유가격	0.0041	0.001	6.26	0.000	0.003	0.005
농산물	-0.0231	0.007	-3.197	0.001	-0.037	-0.009
temp	0.0133	0.067	0.197	0.844	-0.119	0.145
wind	0.1141	0.095	1.198	0.231	-0.073	0.301
rain	-0.0056	0.004	-1.353	0.176	-0.014	0.003
dif_Temp	0.0103	0.009	1.185	0.236	-0.007	0.027
max_Temp	-0.0087	0.034	-0.253	0.800	-0.076	0.059
min_Temp	-0.0190	0.033	-0.574	0.566	-0.084	0.046
=====						
Omnibus:	475.524	Durbin-Watson:	1.590			
Prob(Omnibus):	0.000	Jarque-Bera (JB):	2062.233			
Skew:	-1.607	Prob(JB):	0.00			
Kurtosis:	8.085	Cond. No.	1.85e+20			
=====						

* Adj R-squared : 0.911

→ 요인들이 종속변수를 약 91.1% 설명한다

* F-statistic : 1561, 유의확률 : 0.00

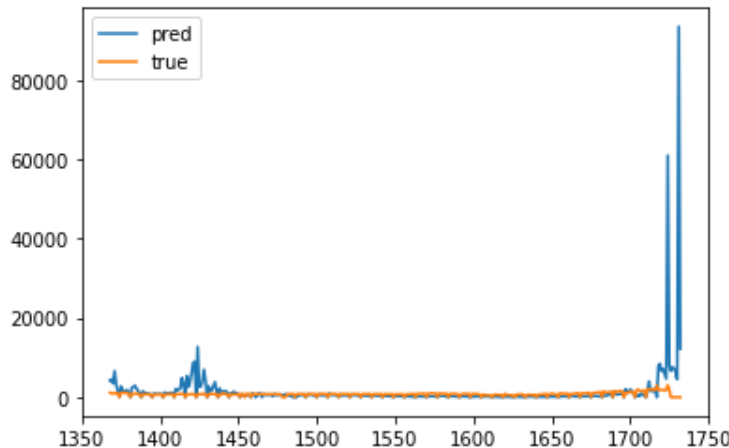
→ 통계적으로 유의한 모형

* P-value값 < 0.05

→ 가격, 거래량, 경유가격, 농산물 변수가 유의하다

* 추정된 회귀계수 $\hat{\beta}$

→ 가격 : 0.0033, 거래량 : 0.0000007903,
경유가격 : 0.0041, 농산물 : -0.0231



NMAE : 1.03



모델링

1. 선형 모델 : OLS , VAR

2. 비선형 모델 : Tree 모델 , 신경망 모델

3. 결과해석

팬이버섯

OLS Regression Results						
Dep. Variable:	1_week	R-squared (uncentered):	0.950			
Model:	OLS	Adj. R-squared (uncentered):	0.950			
Method:	Least Squares	F-statistic:	2879.			
Date:	Sat, 13 Nov 2021	Prob (F-statistic):	0.00			
Time:	02:44:10	Log-Likelihood:	-2503.8			
No. Observations:	1368	AIC:	5026.			
Df Residuals:	1359	BIC:	5073.			
Df Model:	9					
Covariance Type:	nonrobust					
	coef	std err	t	P> t	[0.025	0.975]
가격	0.0008	7.78e-05	10.078	0.000	0.001	0.001
거래량	3.674e-05	1.16e-06	31.592	0.000	3.45e-05	3.9e-05
경유가격	0.0012	0.001	2.279	0.023	0.000	0.002
농산물	-0.0019	0.006	-0.320	0.749	-0.013	0.010
temp	-0.0358	0.067	-0.531	0.595	-0.168	0.096
wind	-0.0284	0.094	-0.303	0.762	-0.212	0.156
rain	0.0022	0.005	0.453	0.651	-0.007	0.012
dif_Temp	-0.0036	0.007	-0.501	0.617	-0.018	0.010
max_Temp	0.0183	0.034	0.538	0.590	-0.048	0.085
min_Temp	0.0219	0.034	0.642	0.521	-0.045	0.089
Omnibus:	378.952	Durbin-Watson:	1.315			
Prob(Omnibus):	0.000	Jarque-Bera (JB):	1190.228			
Skew:	-0.622	Prob(JB):	0.00			
Kurtosis:	17.450	Cond. No.	4.28e+19			

* Adj R-squared : 0.95

→ 요인들이 종속변수를 약 95% 설명한다

* F-statistic : 2879, 유의확률 : 0.00

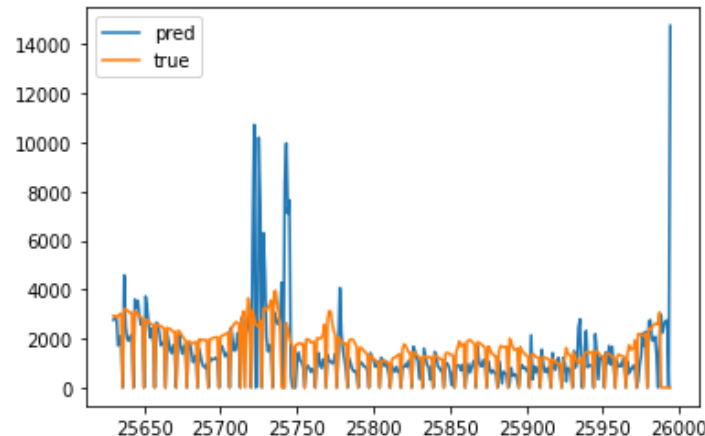
→ 통계적으로 유의한 모형

* P-value값 < 0.05

→ 가격, 거래량, 경유가격 변수가 유의하다

* 추정된 회귀계수 $\hat{\beta}$

→ 가격 : 0.0008, 거래량 : 0.00003674, 경유가격 : 0.0012



NMAE : 0.42



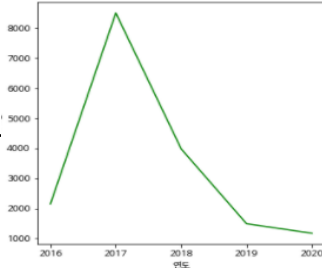
모델링

1. 선형 모델 : OLS , VAR

2. 비선형 모델 : Tree 모델 , 신경망 모델

3. 결과해석

모든품목

품목	배추	양파	건고추	마늘	대파	얼갈이배추	깻잎	시금치
NMAE	1.03	0.53	2517994605959 3.13	0.74	0.54	0.72	1.12	2.31
품목	미나리	파프리카	새송이	팽이버섯	토마토	청상추	백다다기	애호박
NMAE	1.36	1.08	0.61	0.42	1.62	1.96	3.22	1.04
품목	캠벨얼리							
NMAE	3.48							

2017년 거래량이 대략 4배이상 증가

→ 이러한 이상치로 인해 성능이 현저하게 ↓

→ 즉 OLS는 이상치에 민감.

자기상관 고려 X → 전반적으로 낮은 성능
그렇다면 시차를 고려한 시계열 모델은 어떨까?



모델링

1. 선형 모델 : OLS , VAR 2. 비선형 모델 : Tree 모델 , 신경망 모델

1. VAR 이란 ?

회귀모형

설명변수의 영향이 시점(t)이 변하더라도
항상 일정하다는 가정

한계점 : 구조적 변화가 진행되어
설명변수의 영향이 변한 경우 적절한 반
영이 어려움



한계점 보완



ARIMA 모형

모형 설정 용이

한계점 : 변수들 사이의 상호작용 무시
=> 일변량 분석

“ VAR 모형 ” (Vector Auto-Regression)

1. 충격반응분석을 통한 한 변수의 변화가 내생변수에 미치는 동태적 반응 파악
2. 분산분해를 통한 각 내생변수의 변동이 전체 변동에 기여한 부분 분석



모델링

1. 선형 모델 : OLS , VAR 2. 비선형 모델 : Tree 모델 , 신경망 모델

2. 데이터 처리

datetime으로 변환 및 index로 설정

	품목	가격	거래량	경유가격	농산물	temp	max_Temp	min_Temp	hum	wind	rain	sun_Qy	dif_Temp	1_week
date														
2016-01-01	배추	0.0	0.0	1191.69	101.55	2.700000	9.300000	-3.300000	84.133333	0.366667	0.033333	9.033333	12.600000	420.0
2016-01-02	배추	329.0	80860.0	1190.58	101.55	6.133333	13.800000	0.233333	86.900000	0.333333	0.000000	5.933333	13.566667	389.0
2016-01-03	배추	0.0	0.0	1189.58	101.55	6.666667	15.233333	2.133333	89.800000	0.333333	0.000000	9.633333	13.100000	0.0
2016-01-04	배추	478.0	1422742.5	1187.96	101.55	6.133333	11.233333	-0.633333	81.233333	1.100000	0.000000	9.800000	11.866667	398.0
2016-01-05	배추	442.0	1167241.0	1185.81	101.55	1.700000	4.066667	-1.300000	76.566667	0.400000	1.466667	1.900000	5.366667	431.0

Var 모형은 컬럼의 순서도 고려하기 때문에 '결과 - 원인' 순서로 컬럼 지정

	품목	1_week	농산물	가격	거래량	경유가격	dif_Temp	max_Temp	min_Temp	temp	wind	rain
date												
2016-01-01	배추	420.0	101.55	0.0	0.0	1191.69	12.600000	9.300000	-3.300000	2.700000	0.366667	0.033333
2016-01-02	배추	389.0	101.55	329.0	80860.0	1190.58	13.566667	13.800000	0.233333	6.133333	0.333333	0.000000



모델링

1. 선형 모델 : OLS , VAR 2. 비선형 모델 : Tree 모델 , 신경망 모델

배추

정상성 확인 - H0 : 비정상성 vs H1 : 정상성

```

1_week
ADF test statistic: -4.134980921955655
p-value: 0.0008473132427230017
농산물
ADF test statistic: -2.1287035587295784
p-value: 0.23313046089837702
가격
ADF test statistic: -1.8030389974378114
p-value: 0.37895396298178474
거래량
ADF test statistic: -5.816894620552082
p-value: 4.271654200572158e-07
경유가격
ADF test statistic: -2.878108112045438
p-value: 0.047953594019128624
dif_Temp
ADF test statistic: -5.951613894003769
p-value: 2.1405652738582303e-07
max_Temp
ADF test statistic: -3.048780266908163
p-value: 0.030583286374205983
min_Temp
ADF test statistic: -3.0138477562585058
p-value: 0.033621275450113564
temp
ADF test statistic: -2.596454673768401
p-value: 0.0937322304435726
wind
ADF test statistic: -7.636571842458714
p-value: 1.9452978645141846e-11
rain
ADF test statistic: -10.121100165331454
p-value: 9.440360688908869e-18

```

비정상성을 가진
['농산물','가격','경유가격','temp']

1차 차분



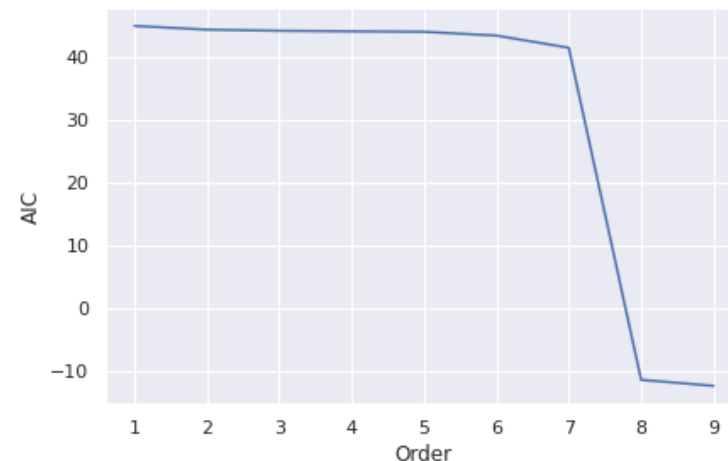
1차 차분 후 비정상성 해결

```

1_week
ADF test statistic: -4.134980921955655
p-value: 0.0008473132427230017
농산물
ADF test statistic: -41.60771758203591
p-value: 0.0
가격
ADF test statistic: -10.096322549232685
p-value: 1.0884548346829093e-17
거래량
ADF test statistic: -5.816894620552082
p-value: 4.271654200572158e-07
경유가격
ADF test statistic: -5.039910776433685
p-value: 1.8491687377966486e-05
dif_Temp
ADF test statistic: -5.951613894003769
p-value: 2.1405652738582303e-07
max_Temp
ADF test statistic: -3.048780266908163
p-value: 0.030583286374205983
min_Temp
ADF test statistic: -3.0138477562585058
p-value: 0.033621275450113564
temp
ADF test statistic: -10.727994152293467
p-value: 3.03048531624028e-19
wind
ADF test statistic: -7.636571842458714
p-value: 1.9452978645141846e-11
rain
ADF test statistic: -10.121100165331454
p-value: 9.440360688908869e-18

```

AIC curve 로 lag 값 채택



시차 9일로 설정



모델링

1. 선형 모델 : OLS , VAR 2. 비선형 모델 : Tree 모델 , 신경망 모델

배추

결과

Results for equation 1_week

	coefficient	std. error	t-stat	prob
const	160.050449	104.757242	1.528	0.127
L1.1_week	0.137003	0.026953	5.083	0.000
L1.농산물	-1.821119	5.395346	-0.338	0.736
L1.가격	-0.013439	NAN	NAN	NAN
L1.거래량	0.000013	0.000017	0.770	0.442
L1.경유품	6.045439	7.994120	0.756	0.450
L1.dif_Temp	37.916069	28.017707	1.353	0.176
L1.max_Temp	-11.377868	40.367326	-0.282	0.778
L1.min_Temp	18.524190	11.309253	1.638	0.101
L1.temp	-65.744572	18.339457	-3.585	0.000
L1.wind	5.154879	14.257002	0.362	0.718
L1.rain	-0.879844	0.540482	-1.628	0.104
L2.1_week	0.087056	0.027502	3.165	0.002
L2.농산물	19.641686	5.373867	3.655	0.000
L2.가격	0.044093	0.050951	0.865	0.387
L2.거래량	0.000007	0.000017	0.399	0.690
L2.경유품	-19.587640	9.996472	-1.959	0.050
L2.dif_Temp	-19.346614	18.958909	-1.020	0.308
L2.max_Temp	12.322805	NAN	NAN	NAN
L2.min_Temp	-13.467857	11.868515	-1.136	0.256
L2.temp	-56.504974	17.775210	-3.179	0.001
L2.wind	-1.999017	15.209702	-0.131	0.895
L2.rain	-0.128257	0.535420	-0.240	0.811

⋮

L9.1_week	-0.010863	NAN	NAN	NAN
L9.농산물	3.776339	5.390144	0.701	0.484
L9.가격	0.044561	0.027064	1.646	0.100
L9.거래량	-0.000030	0.000016	-1.824	0.068
L9.경유품	12.760619	7.909825	1.613	0.107
L9.dif_Temp	32.390882	22.409566	1.445	0.148
L9.max_Temp	-116.698555	27.302145	-4.274	0.000
L9.min_Temp	5.585746	11.238223	0.497	0.619
L9.temp	-32.910354	15.294022	-2.152	0.031
L9.wind	-23.667522	14.650721	-1.615	0.106
L9.rain	0.317620	0.546348	0.581	0.561



변수	Coef	p-value	변수	Coef	P-value
L1.temp	-65.74	0.000	L6.min_Temp	-27.6	0.029
L2.농산물	19.64	0.000	L8.dif_Temp	-104.61	0.000
L2.temp	-56.50	0.001	L8.max_Temp	125.83	0.000
L6.dif_Temp	-37.52	0.019	L9.max_Temp	-116.70	0.000
L6.max_Temp	110.39	0.000	L9.temp	-32.91	0.031

현 시점 6,8일 전의 일 최고 기온이 높을수록
일주일 후 배추의 가격에 양의 영향을 미쳤다



모델링

1. 선형 모델 : OLS, VAR 2. 비선형 모델 : Tree 모델, 신경망 모델

얼갈이배추

정상성 확인 - H_0 : 비정상성 vs H_1 : 정상성

```

1_week
ADF test statistic: -4.792857909485917
p-value: 5.612294997138123e-05
농산물
ADF test statistic: -2.1287035587295784
p-value: 0.23313046089837702
가격
ADF test statistic: -3.734941555489577
p-value: 0.0036476679367053555
거래량
ADF test statistic: -3.564393575447993
p-value: 0.0064793889577664816
경유가격
ADF test statistic: -2.938657596362886
p-value: 0.04103379535073715
dif_Temp
ADF test statistic: -6.195141865190284
p-value: 5.99059448014576e-08
max_Temp
ADF test statistic: -2.8301679037168177
p-value: 0.054092035238827954
min_Temp
ADF test statistic: -2.8369598299665184
p-value: 0.05318518014919115
temp
ADF test statistic: -2.821848901272586
p-value: 0.05521994809800091
wind
ADF test statistic: -3.7068502334127302
p-value: 0.004017824755979362
rain
ADF test statistic: -8.73341934616746
p-value: 3.1467123821027994e-14

```

비정상성을 가진
['농산물', 'max_T', 'min_T', 'temp']

1차 차분



1차 차분 후 비정상성 해결

```

1_week
ADF test statistic: -4.792857909485917
p-value: 5.612294997138123e-05
농산물
ADF test statistic: -41.60771758203591
p-value: 0.0
가격
ADF test statistic: -3.734941555489577
p-value: 0.0036476679367053555
거래량
ADF test statistic: -3.564393575447993
p-value: 0.0064793889577664816
경유가격
ADF test statistic: -2.938657596362886
p-value: 0.04103379535073715
dif_Temp
ADF test statistic: -6.195141865190284
p-value: 5.99059448014576e-08
max_Temp
ADF test statistic: -9.977627688272221
p-value: 2.1564889366319094e-17
min_Temp
ADF test statistic: -13.004010504457566
p-value: 2.6509380700816577e-24
temp
ADF test statistic: -8.452725532759278
p-value: 1.64572158179541e-13
wind
ADF test statistic: -3.7068502334127302
p-value: 0.004017824755979362
rain
ADF test statistic: -8.73341934616746
p-value: 3.1467123821027994e-14

```

AIC curve 로 lag 값 채택



시차 7일로 설정



모델링

1. 선형 모델 : OLS , VAR 2. 비선형 모델 : Tree 모델 , 신경망 모델

얼갈이배추

결과

Results for equation 1_week

	coefficient	std. error	t-stat	prob
const	513.777557	142.804230	3.598	0.000
L1.l_week	0.452547	0.025517	17.735	0.000
L1.농산물	-10.626603	5.314891	-1.999	0.046
L1.가격	-0.365051	0.029391	-12.421	0.000
L1.거래량	-0.000217	0.000082	-2.659	0.008
L1.경유가격	10.458287	6.016970	1.738	0.082
L1.dif_Temp	-53.595390	24.816604	-2.160	0.031
L1.max_Temp	24.271795	15.010453	1.617	0.106
L1.min_Temp	-4.366290	12.154961	-0.359	0.719
L1.temp	8.233491	12.969431	0.635	0.526
L1.wind	45.768165	17.548598	2.608	0.009
L1.rain	0.712789	0.539968	1.320	0.187
L2.l_week	0.034782	0.028421	1.224	0.221
L2.농산물	-7.092061	5.327085	-1.331	0.183
L2.가격	0.013149	0.031222	0.421	0.674
L2.거래량	-0.000103	0.000086	-1.187	0.235
L2.경유가격	-13.491673	11.576477	-1.165	0.244
L2.dif_Temp	-13.907346	26.670295	-0.521	0.602
L2.max_Temp	28.944635	16.313035	1.774	0.076
L2.min_Temp	0.199700	12.456034	0.016	0.987
L2.temp	-1.820367	13.950006	-0.130	0.896
L2.wind	0.841996	19.218461	0.044	0.965
L2.rain	-0.511183	0.543781	-0.940	0.347



변수	Coef	p-value	변수	Coef	P-value
L1.dif_Temp	-53.60	0.031	L5.rain	-1.42	0.009
L1.wind	45.77	0.009	L7.농산물	-11.28	0.035
L3.농산물	16.90	0.002	L7.max_Temp	-42.24	0.004
L4.농산물	-13.06	0.015	L7.rain	-1.21	0.026
L4.dif_Temp	-58.05	0.029			

...

L7.l_week	0.568320	0.025831	22.001	0.000
L7.농산물	-11.278516	5.356671	-2.106	0.035
L7.가격	0.154355	0.027431	5.627	0.000
L7.거래량	0.000640	0.000081	7.885	0.000
L7.경유가격	0.929258	6.005551	0.155	0.877
L7.dif_Temp	42.839800	25.470106	1.682	0.093
L7.max_Temp	-42.243518	14.471667	-2.919	0.004
L7.min_Temp	9.023481	11.730496	0.769	0.442
L7.temp	13.144330	12.945223	1.015	0.310
L7.wind	-25.883608	17.852573	-1.450	0.147
L7.rain	-1.205411	0.541348	-2.227	0.026

1,4일전 일교차가 클수록 일주일 후
얼갈이배추의 가격에 음의 영향을 미쳤다.



모델링

1. 선형 모델 : OLS , VAR 2. 비선형 모델 : Tree 모델 , 신경망 모델

새송이

정상성 확인 - H_0 : 비정상성 vs H_1 : 정상성

```

1_week
ADF test statistic: -6.213816187552035
p-value: 5.42655219377619e-08
농산물
ADF test statistic: -2.1287035587295784
p-value: 0.23313046089837702
가격
ADF test statistic: -5.829736039387242
p-value: 4.001133542699277e-07
거래량
ADF test statistic: -6.9212481079952655
p-value: 1.144980221345938e-09
경유가격
ADF test statistic: -2.823431030903555
p-value: 0.055003974588095104
dif_Temp
ADF test statistic: -6.401146887529275
p-value: 1.9941074473208714e-08
max_Temp
ADF test statistic: -2.840198870178088
p-value: 0.052757104322291866
min_Temp
ADF test statistic: -3.5941742892414195
p-value: 0.005873384540098733
temp
ADF test statistic: -3.454296170490838
p-value: 0.009241439623793563
wind
ADF test statistic: -3.1958268216998285
p-value: 0.020227513512593696
rain
ADF test statistic: -6.843204862746117
p-value: 1.769549476543645e-09

```

비정상성을 가진
['농산물', '경유가격', 'max_T']

1차 차분



1차 차분 후 비정상성 해결

```

1_week
ADF test statistic: -6.213816187552035
p-value: 5.42655219377619e-08
농산물
ADF test statistic: -41.60771758203591
p-value: 0.0
가격
ADF test statistic: -5.829736039387242
p-value: 4.001133542699277e-07
거래량
ADF test statistic: -6.9212481079952655
p-value: 1.144980221345938e-09
경유가격
ADF test statistic: -4.895712112037314
p-value: 3.555081308482447e-05
dif_Temp
ADF test statistic: -6.401146887529275
p-value: 1.9941074473208714e-08
max_Temp
ADF test statistic: -10.234524237043159
p-value: 4.928950294317126e-18
min_Temp
ADF test statistic: -3.5941742892414195
p-value: 0.005873384540098733
temp
ADF test statistic: -3.454296170490838
p-value: 0.009241439623793563
wind
ADF test statistic: -3.1958268216998285
p-value: 0.020227513512593696
rain
ADF test statistic: -6.843204862746117
p-value: 1.769549476543645e-09

```

AIC curve 로 lag 값 채택



시차 8일로 설정



모델링

1. 선형 모델 : OLS , VAR 2. 비선형 모델 : Tree 모델 , 신경망 모델

새송이

결과

Results for equation l1_week

	coefficient	std. error	t-stat	prob
const	874.479718	232.735005	3.757	0.000
L1.l_week	0.429548	0.026778	16.041	0.000
L1.농산물	-10.139116	10.436788	-0.971	0.331
L1.가격	-0.099586	1420482865168.420898	-0.000	1.000
L1.거래량	0.000221	0.000469	0.470	0.638
L1.경유가격	3.773211	15.305022	0.247	0.805
L1.dif_Temp	-58.574780	44.653517	-1.312	0.190
L1.max_Temp	90.506708	36.757767	2.462	0.014
L1.min_Temp	-51.443084	24.860751	-2.069	0.039
L1.temp	-37.385013	24.631391	-1.518	0.129
L1.wind	70.449746	65.068795	1.083	0.279
L1.rain	-0.055065	1.035176	-0.053	0.958
L2.l_week	0.039568	0.028512	1.388	0.165
L2.농산물	9.886765	10.417945	0.949	0.343
L2.가격	-0.074921	0.036651	-2.044	0.041
L2.거래량	0.000310	0.000455	0.683	0.495
L2.경유가격	-20.178166	19.078424	-1.058	0.290
L2.dif_Temp	63.657289	47.946015	1.328	0.184
L2.max_Temp	27.870315	40.112275	0.695	0.487
L2.min_Temp	-3.020883	30.080892	-0.100	0.920

⋮

L8.l_week	-0.099586	1420482865168.425049	-0.000	1.000
L8.농산물	10.331286	10.476040	0.986	0.324
L8.가격	-0.063306	0.034256	-1.848	0.065
L8.거래량	-0.001962	0.000455	-4.310	0.000
L8.경유가격	2.958938	15.130523	0.196	0.845
L8.dif_Temp	-20.589328	41.878646	-0.492	0.623
L8.max_Temp	-97.462588	31.673676	-3.077	0.002
L8.min_Temp	2.853815	25.026515	0.114	0.909
L8.temp	-27.726440	25.903349	-1.070	0.284
L8.wind	-53.140731	66.076401	-0.804	0.421
L8.rain	0.231180	1.025361	0.225	0.822



변수	Coef	p-value	변수	Coef	P-value
L1.max_Temp	90.51	0.014	L5.dif_Temp	-192.76	0.000
L1.min_Temp	-51.44	0.039	L6.temp	87.79	0.002
L3.농산물	31.98	0.002	L8.max_Temp	-97.46	0.002
L4.dif_Temp	117.49	0.015			

5일전 일교차가 심할수록 일주일 후
새송이 가격에 음의 영향을 미쳤다.



모델링

1. 선형 모델 : OLS , VAR 2. 비선형 모델 : Tree 모델 , 신경망 모델

모든품목

품목	배추	양파	건고추	마늘	대파	얼갈이배추	깻잎	시금치
nmae	0.30	0.21	0.38	0.68	0.34	0.26	0.36	0.56
품목	미나리	파프리카	새송이	팽이버섯	토마토	청상추	백다다기	애호박
nmae	0.71	0.31	0.15	0.35	0.31	0.52	0.32	0.38
품목	캠벨얼리							
nmae	0.29							

몇몇 품목의 경우 회귀계수의 부호가 lag 값에 따라 변하는 등 기후 데이터의 영향력의 문제점 발생
 사용되는 변수 및 표본기간, 시차길이에 따라 결과가 달라진다는 VAR 모형의 한계점 고려



좀 더 상세한
 기후 데이터의 분석을 위한
 비선형 모델 고려



모델링

1. 선형 모델 : OLS , VAR 2. 비선형 모델 : Tree 모델 , 신경망 모델

1. TREE 모델

비선형 모델 중 가장 가벼움
모델 해석에도 용이

“ AutoML 패키지 ”

PYCARRET



```
[ ] import pycaret
    from pycaret.regression import *
```



	Model	MAE	MSE	RMSE	R2	RMSLE	MAPE	TT (Sec)
et	Extra Trees Regressor	366.0832	401565.3833	622.5422	0.8890	1.0761	0.1153	1.5890
rf	Random Forest Regressor	398.6124	463163.9662	671.0037	0.8701	1.0823	0.1273	2.9870
lightgbm	Light Gradient Boosting Machine	410.8870	508372.2290	702.4783	0.8577	1.0918	0.1275	0.5460
gbr	Gradient Boosting Regressor	424.1487	482667.0145	685.4968	0.8646	1.0766	0.1396	1.1750
llar	Lasso Least Angle Regression	506.0599	749470.9819	851.9792	0.7929	1.1163	0.1605	0.0170
knn	K Neighbors Regressor	515.1252	666712.7312	810.4949	0.8065	1.1125	0.1717	0.0710
br	Bayesian Ridge	533.8166	736308.0566	846.3047	0.7952	1.0661	0.1842	0.0270
lasso	Lasso Regression	543.3528	736773.3219	848.7537	0.7935	1.0699	0.1900	0.0280
ridge	Ridge Regression	547.9374	735810.6688	848.2306	0.7937	1.0673	0.1934	0.0210
ada	AdaBoost Regressor	547.9480	661213.4778	802.7012	0.8142	1.1095	0.1934	0.4530
dt	Decision Tree Regressor	573.7174	965529.7009	967.6919	0.7183	1.3189	0.1912	0.0770
en	Elastic Net	615.1462	1014228.6281	986.8382	0.7201	1.1371	0.2086	0.0240
lr	Linear Regression	637.2990	911121.6406	942.2554	0.7378	1.1132	0.2374	0.5020
lar	Least Angle Regression	4286362.7456	166192137823537.9375	5726360.3891	-35624290.0444	3.9028	2058.3778	0.0310

mae 기준 가장 성능이 좋았던
Extra Tree 모델을 선택

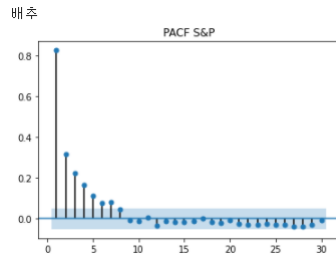


모델링

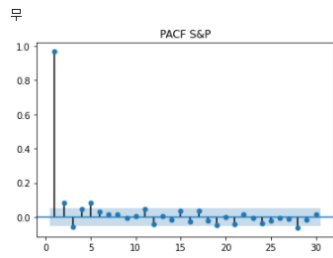
1. 선형 모델 : OLS , VAR 2. 비선형 모델 : Tree 모델 , 신경망 모델

2. 변수 추가 _ 가격 및 거래량

각 품목별 lag 값은 pacf 그래프를 통해 window size 결정

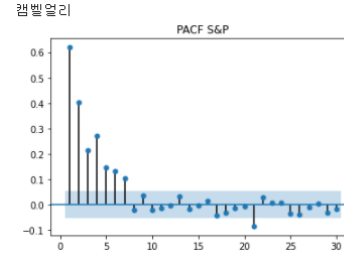


window size = 7



window size = 1

...



window size = 7

p_ : 가격에 대한 추가변수
q_ : 거래량에 대한 추가변수

이전 시점 가격을 현재시점까지
지수가중평균 통계량

date	날짜	품목	가격	거래량	주산지	경유가격	농산물	month_avg	temp	max_Temp	min_Temp
2016-01-01	2016-01-01	배추	0.0	0.0	전남	1191.69	101.55	531.653846	2.700000	9.300000	-3.300000
2016-01-02	2016-01-02	배추	329.0	80860.0	전남	1190.58	101.55	531.653846	6.133333	13.800000	0.233333
2016-01-03	2016-01-03	배추	0.0	0.0	전남	1189.58	101.55	531.653846	6.666667	15.233333	2.133333
2016-01-04	2016-01-04	배추	478.0	1422742.5	전남	1187.96	101.55	531.653846	6.133333	11.233333	-0.633333
2016-01-05	2016-01-05	배추	442.0	1167241.0	전남	1185.81	101.55	531.653846	1.700000	4.066667	-1.300000

...

p_lag_1	q_lag_1	p_lag_2	q_lag_2	p_lag_3	q_lag_3
-1.0	-1.0	-1.0	-1.0	-1.0	-1.0
0.0	0.0	-1.0	-1.0	-1.0	-1.0
329.0	80860.0	0.0	0.0	-1.0	-1.0
0.0	0.0	329.0	80860.0	0.0	0.0
478.0	1422742.5	0.0	0.0	329.0	80860.0

...

p_lag_7	q_lag_7	p_ewma	q_ewma
-1.0	-1.0	-1.000000	-1.000000
-1.0	-1.0	-0.466667	-0.466667
-1.0	-1.0	124.301775	30621.248521
-1.0	-1.0	86.754572	21371.644248
-1.0	-1.0	187.158135	380999.175428



모델링

1. 선형 모델 : OLS , VAR 2. 비선형 모델 : Tree 모델 , 신경망 모델

2. 변수 추가 _ 기후 데이터

시점을 다양하게 고려하여 시차 변수 추가

```

lag['t_lag_1'] = lag.groupby('품목')['temp'].shift(1) # 현재시점 기준 하루전
lag['t_lag_2'] = lag.groupby('품목')['temp'].shift(2) # 현재시점 기준 이틀전
lag['t_lag_3'] = lag.groupby('품목')['temp'].shift(3) # 현재시점 기준 3일전
lag['t_lag_4'] = lag.groupby('품목')['temp'].shift(4) # 현재시점 기준 4일전

lag['t_lag_2w'] = lag.groupby('품목')['temp'].shift(7) # 예측시점 기준 1주일전
lag['t_lag_3w'] = lag.groupby('품목')['temp'].shift(14) # 예측시점 기준 2주일전
lag['t_lag_1m'] = lag.groupby('품목')['temp'].shift(23) # 예측시점 기준 1달전
lag['t_lag_2m'] = lag.groupby('품목')['temp'].shift(53) # 예측시점 기준 2달전

lag['t_lead_1'] = lag.groupby('품목')['temp'].shift(-6) # 예측시점 기준 하루전
lag['t_lead_2'] = lag.groupby('품목')['temp'].shift(-5) # 예측시점 기준 이틀전
lag['t_lead_3'] = lag.groupby('품목')['temp'].shift(-4) # 예측시점 기준 3일전
lag['t_lead_4'] = lag.groupby('품목')['temp'].shift(-3) # 예측시점 기준 4일전

```

['temp' , 'max_Temp' , 'min_Temp' , wind , rain]
변수에도 동일하게 적용

추가한 컬럼 예시

t_lag_1m	t_lag_2m	t_lead_1	t_lead_2	...	max_t_lag_3	max_t_lag_4	max_t_lag_2w
-1.000000	-1.000000	3.000000	3.300000	...	-1.000000	-1.000000	-1.000000
-1.000000	-1.000000	2.833333	2.333333	...	9.300000	-1.000000	-1.000000

⋮

r_lead_1	r_lead_2	r_lead_3	r_lead_4	w_lag_1	w_lag_2	w_lag_3
0.033333	0.0	0.000000	1.466667	0.366667	-1.000000	-1.000000
0.000000	0.0	0.033333	0.000000	0.333333	0.333333	0.366667

autoML 최종 데이터

Target 변수 : 1_week

rows : 24997

features : { numeric 96 , categorical 2 }



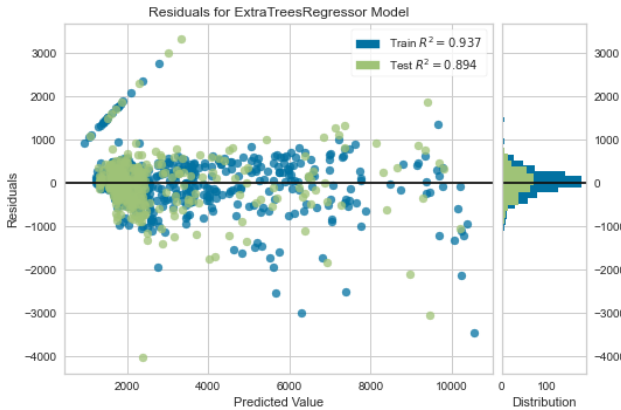
모델링

1. 선형 모델 : OLS , VAR 2. 비선형 모델 : Tree 모델 , 신경망 모델

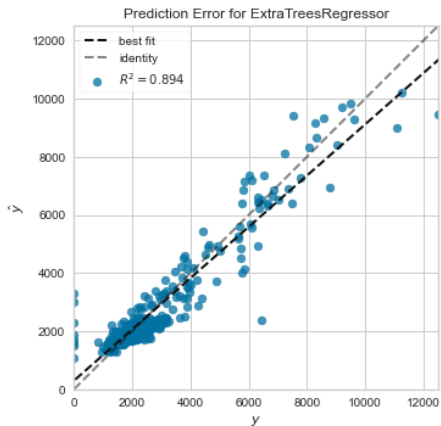
시금치

결과

Residuals plot



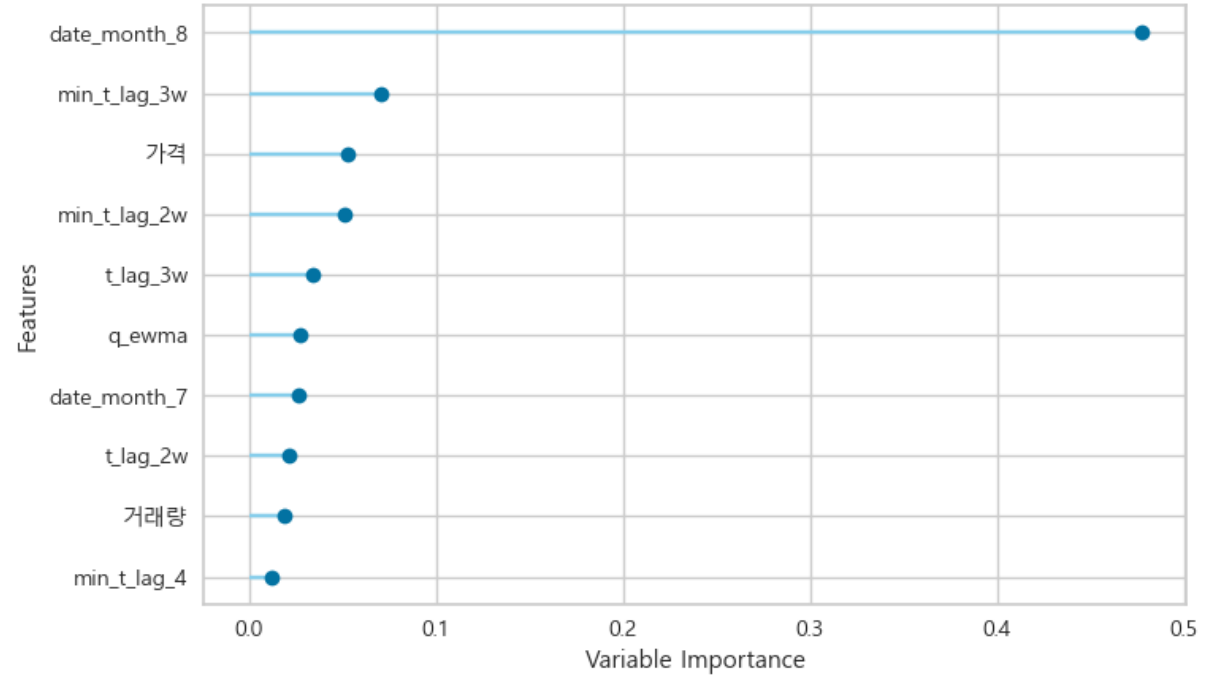
Error plot



Evaluates

nmae	0.38
Train R^2	0.937
Test R^2	0.894

Feature Importance Plot



성능이 가장 좋았던 품목이라는 것을 고려해봤을 때 기후 변수 중요
기후 변수가 중요하게 고려되면 모델의 설명력을 높일 수 있을 것



모델링

1. 선형 모델 : OLS , VAR 2. 비선형 모델 : Tree 모델 , 신경망 모델

모든품목

결과

품목	배추	양파	건고추	마늘	대파	얼갈이배추	깻잎	시금치
nmae	0.37	0.06	0.25	0.09	0.17	0.29	0.34	0.38
Train R^2	0.870	0.877	0.430	0.821	0.806	0.807	0.761	0.937
Test R^2	0.489	0.772	0.301	0.656	0.660	0.683	0.538	0.894
품목	미나리	파프리카	새송이	팽이버섯	토마토	백다다기	애호박	캠벨얼리
nmae	0.29	0.34	0.12	0.21	0.42	0.40	0.30	0.06
Train R^2	0.834	0.881	0.714	0.800	0.855	0.804	0.835	0.696
Test R^2	0.705	0.624	0.560	0.606	0.708	0.557	0.709	0.442
품목	청상추							
nmae	0.18							
Train R^2	0.965							
Test R^2	0.750							

기후 변수의 복잡성을 고려한
딥러닝 기반 모델 적용



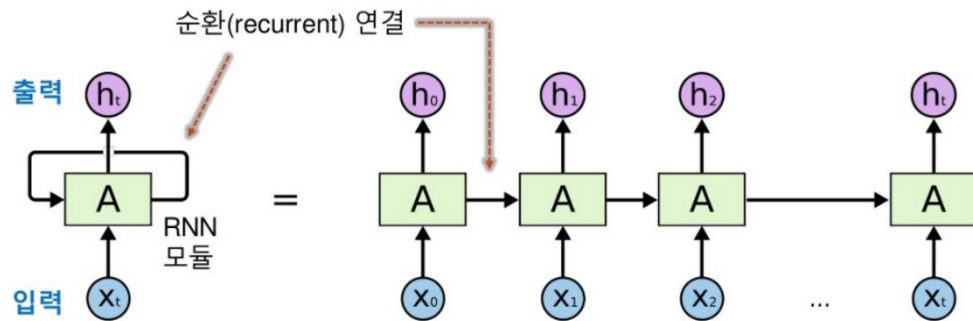
모델링

1. 선형 모델 : OLS , VAR 2. 비선형 모델 : Tree 모델 , 신경망 모델

1. RNN의 한계점

RNN

- 순차적인 데이터를 입력받아 결과값 도출하는 딥러닝 모델
- 이전 입력 값들을 고려 → 현재 입력 값의 출력 값을 결정



RNN의 한계

경사하강법 알고리즘으로 모델 최적화

- 입력 시퀀스가 길어질 수록 곱해지는 미분 값은 늘어남
- Gradient vanishing 문제 발생 : 멀리 떨어져 있는 정보는 현재 셀에 영향을 거의 못미치게 되는 상황

LSTM



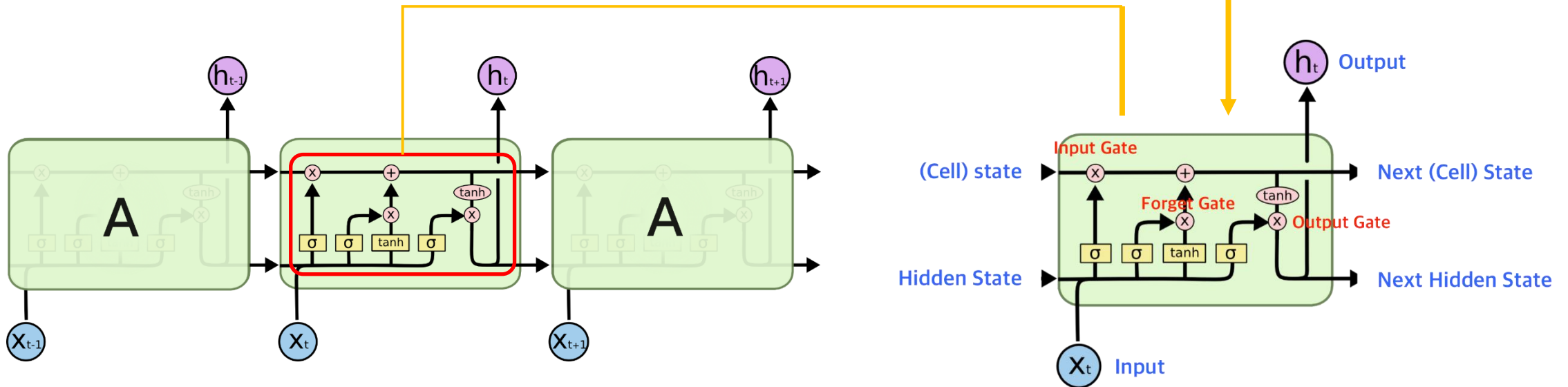
모델링

1. 선형 모델 : OLS , VAR 2. 비선형 모델 : Tree 모델 , 신경망 모델

2. LSTM 이란?

LSTM

- RNN의 한 종류
- 3개의 gate(forget gate, input gate, output gate)와 1개의 cell state 존재
- 잊어야 할 기억은 잊고, 기억해야 할 정보는 기억함으로써 최적 값 도출
- 기존 RNN의 문제인 vanishing gradient 를 방지해줌.
- RNN의 기본 구조는 그대로 사용하되, 기본셀 대신 LSTM셀 사용

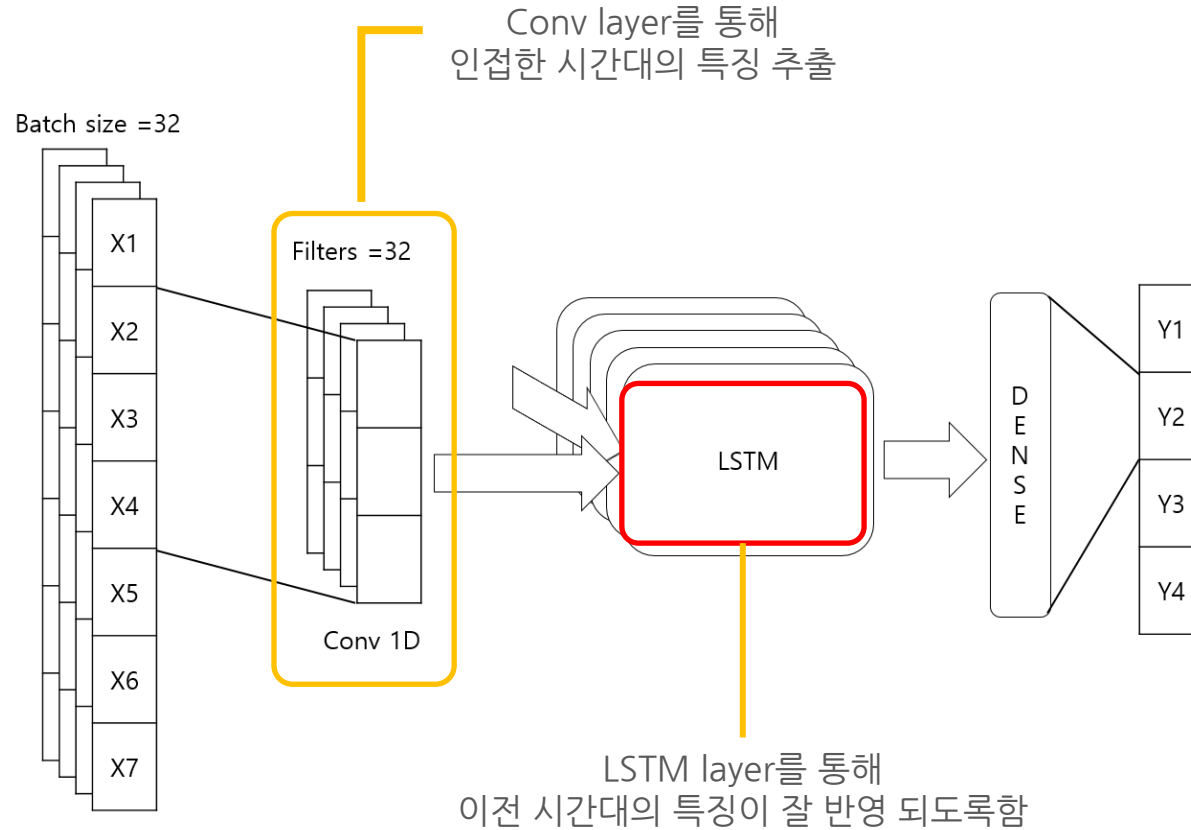




모델링

1. 선형 모델 : OLS , VAR 2. 비선형 모델 : Tree 모델 , 신경망 모델

3. 신경망 모델 구성



최대한 시간의 흐름에 따른 정보를 활용할 수 있도록 구성



모델링

1. 선형 모델 : OLS , VAR 2. 비선형 모델 : Tree 모델 , 신경망 모델

4. 분석과정

- 데이터 전처리

1. 가격, 거래량 0인 데이터 삭제 (대부분 일요일 데이터)

2. 결측치 처리

- Temp, dif_Temp, wind, max temp, min temp → 이전값 으로 대체
- rain → 0값으로 대체

3. MinMaxScaling



거래량	경유가격	농산물	temp	wind	rain	dif_Temp	max_Temp	min_Temp	가격	l_week
0.015514	0.294353	0.148436	0.358860	0.056911	0.000000	0.299914	0.427536	0.317607	0.018079	0.0778
0.273064	0.288071	0.148436	0.358860	0.243902	0.000000	0.256213	0.365539	0.295703	0.049401	0.0796
0.224025	0.282916	0.148436	0.240427	0.073171	0.007184	0.089117	0.192432	0.278854	0.041833	0.0862
0.200661	0.277880	0.148436	0.272484	0.341463	0.000000	0.209512	0.270531	0.242207	0.041833	0.0858
0.199589	0.274547	0.148436	0.283170	0.317073	0.000000	0.044559	0.227858	0.359730	0.043094	0.0882

- 하이퍼 파라미터 설정

- window_size : 과거 며칠의 가격 데이터로 가격을 예측할 것인지를 정하는 parameter

pacf를 바탕으로 각 품목의 window_size 설정

배추 : 8	미나리 : 6
양파 : 4	파프리카 : 4
건고추 : 11	새송이 : 3
마늘 : 7	팽이버섯 : 2
대파 : 6	토마토 : 8
얼갈이배추 : 5	청상추 : 3
깻잎 : 2	백다다기 : 3
시금치 : 3	애호박 : 4
	캠벨얼리 : 8

- Batch size : 32



모델링

1. 선형 모델 : OLS , VAR 2. 비선형 모델 : Tree 모델 , 신경망 모델

5. 결과

품목	배추	양파	건고추	마늘	대파	얼갈이배추	깻잎	시금치
NMAE	0.29	0.16	30.69	0.14	0.17	0.24	0.24	0.25
품목	미나리	파프리카	새송이	팽이버섯	토마토	백다다기	애호박	캠벨얼리
NMAE	0.37	0.24	0.08	0.20	0.28	0.26	0.36	0.48
품목	청상추							
NMAE	0.31							

배추 : 성능 향상된 품목

성능 향상이 기후데이터와 관계가 있는가? → LIME 알고리즘 사용

LIME (locally interpretable model-agnostic explanations)

- 모델의 개별 예측값을 설명하기 위한 알고리즘
- 복잡한 모형을 해석이 가능한 심플한 모형으로 근사시킨다.
- dataset을 MinMaxScaling 했으므로 → 0 ~ 1 사이의 값으로 나타남.
- Batch size = 32 → 시기별로 나뉜 local explanation은 총 32개



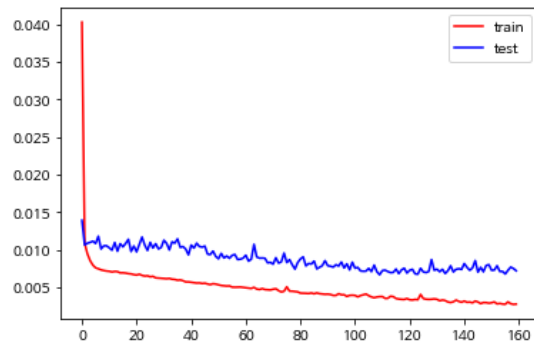
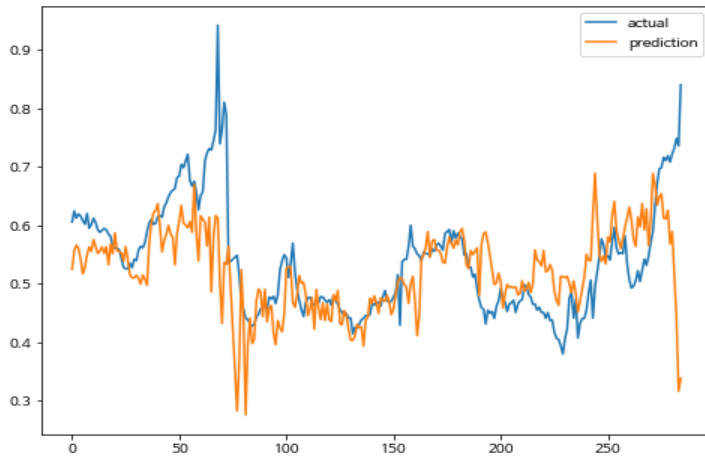
모델링

1. 선형 모델 : OLS , VAR 2. 비선형 모델 : Tree 모델 , 신경망 모델

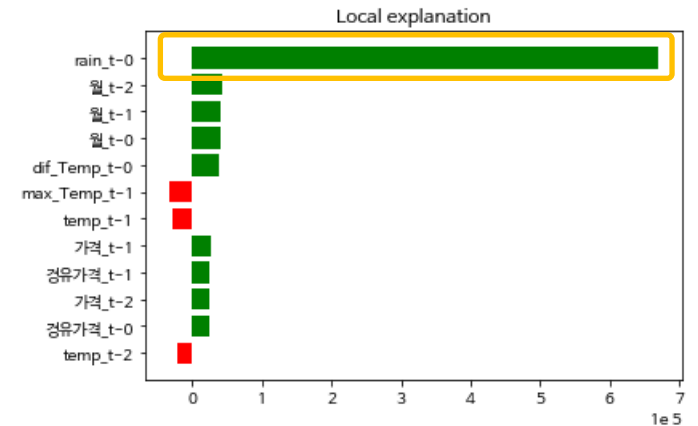
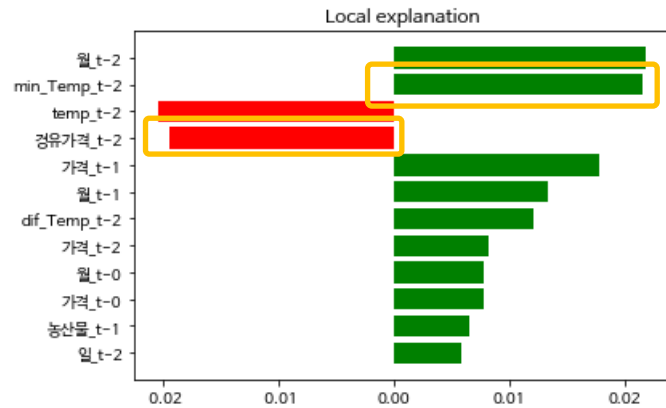
5. 결과

새송이

Window_size : 3
NMAE : 0.08



새송이의 test 데이터 32개의 예측 값 중
1월 중순과 장마기간에 대한 local explanation



일 최저기온 \uparrow \rightarrow 새송이버섯 가격 \uparrow

경유 가격 \uparrow \rightarrow 새송이버섯 가격 \downarrow

- 장마 기간에는 이상치 존재



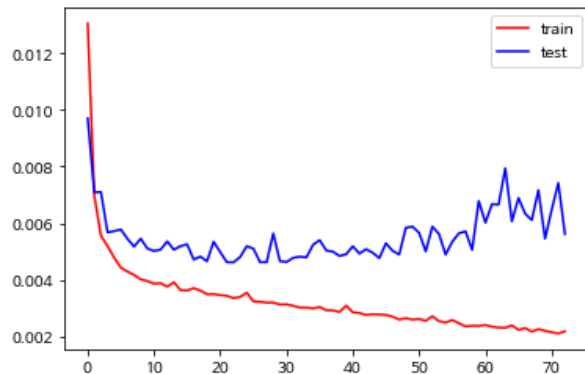
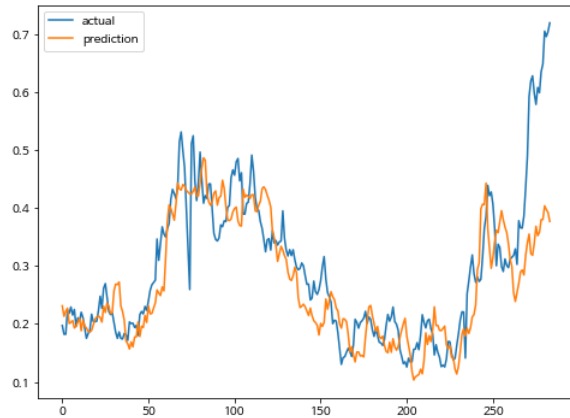
모델링

1. 선형 모델 : OLS , VAR 2. 비선형 모델 : Tree 모델 , 신경망 모델

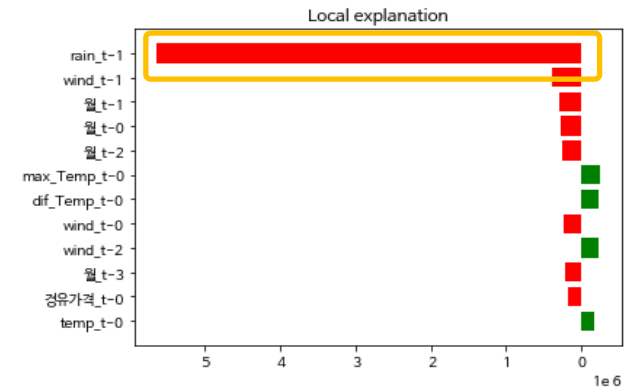
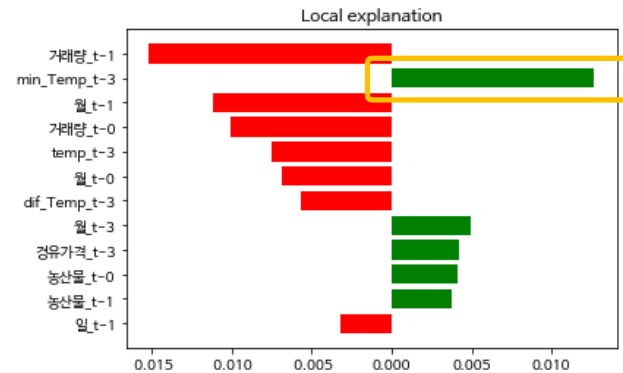
5. 결과

파프리카

Window_size : 4
NMAE : 0.24



파프리카 test 데이터 32개의 예측 값 중
1월 말과 장마 기간의 local explanation



- 일정하지 않은 패턴.
- 다만, 일 최저 기온에 대한 영향은 지속적으로 관찰됨.
- 특정 시기에 강수량이 매우 큰 음의 영향을 보여줌 → 장마기간의 영향으로 예상됨.



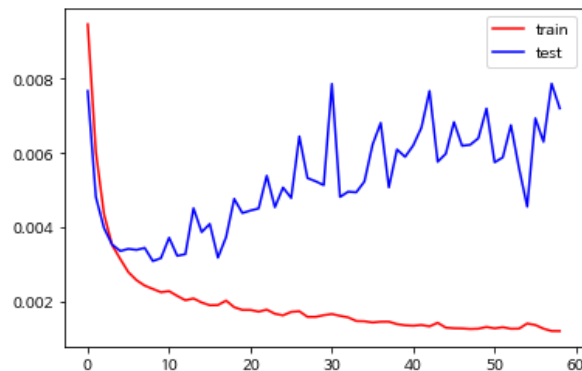
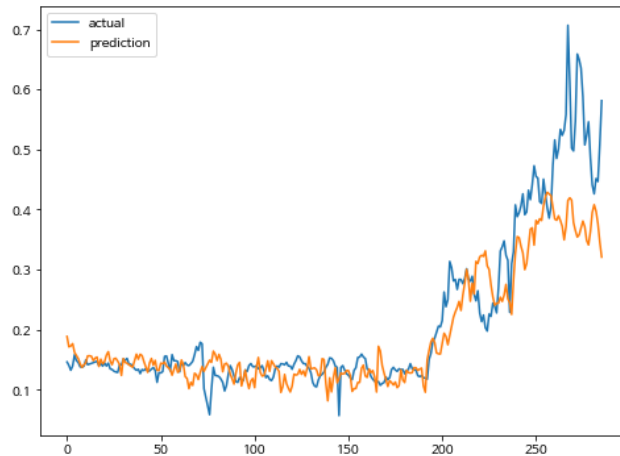
모델링

1. 선형 모델 : OLS , VAR 2. 비선형 모델 : Tree 모델 , 신경망 모델

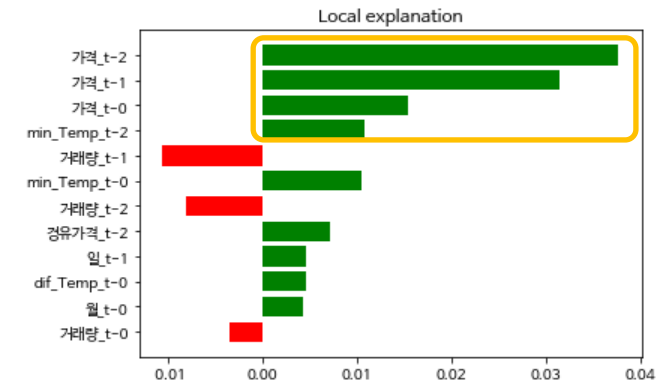
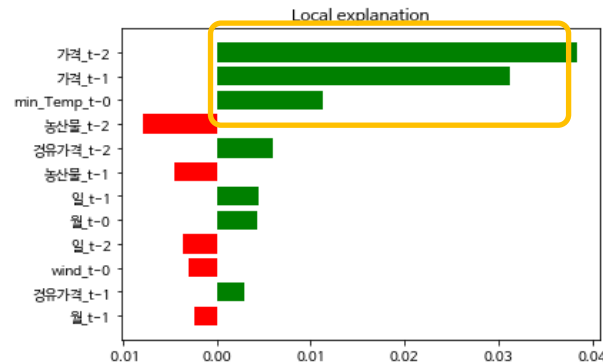
5. 결과

시금치

Window_size : 3
NMAE : 0.25



시금치 test 데이터 32개의 예측 값 중
1월 말 ~ 데이터가 끝나는 시점의 local explanation



- 지속적으로 1-2일전 가격의 영향을 많이 받음.
- 지속적으로 일 최저기온의 영향을 받음
- 앞에서 봤듯이, 시금치는 온도와 상관계수가 높고
Tree 모델의 feature importance에서도 온도 변수와 상대적으로 꽤 높은 값을 가짐.
→ 모델 복잡도 증가에 따른 성능 향상으로 판단됨.



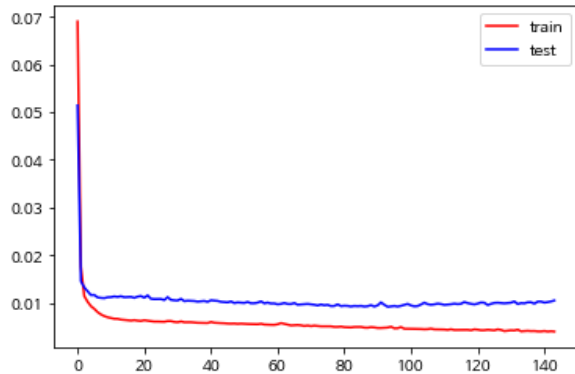
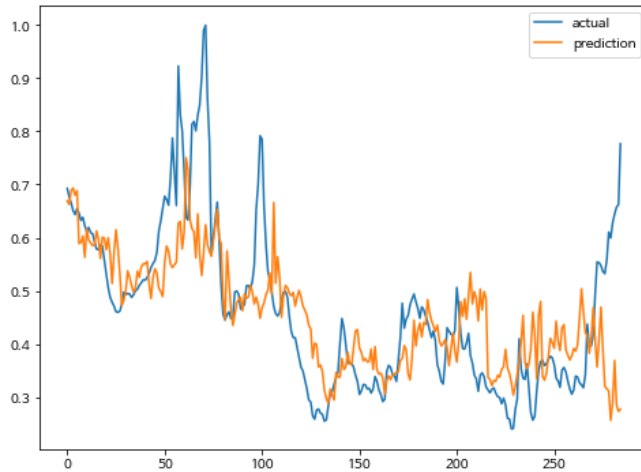
모델링

1. 선형 모델 : OLS , VAR 2. 비선형 모델 : Tree 모델 , 신경망 모델

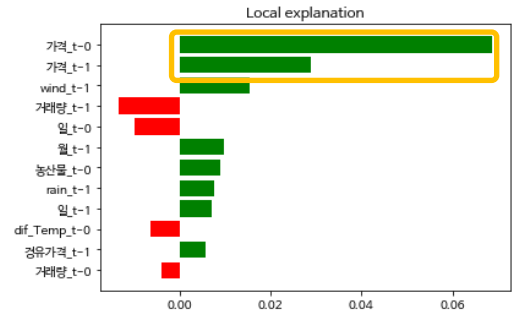
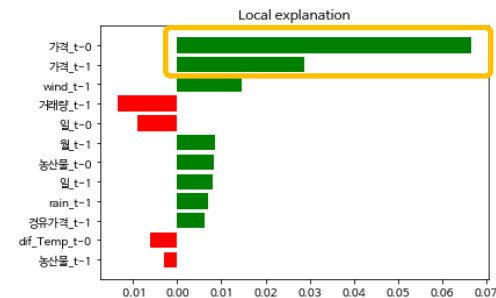
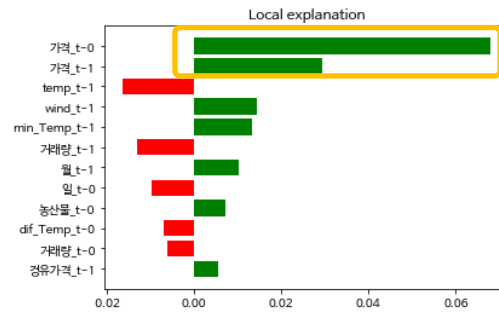
5. 결과

팬이버섯

Window_size : 2
NMAE : 0.20



팬이버섯의 test 데이터 32개의 예측 값 중
1월 local explanation



- 가격의 영향력이 매우 큼.
- 기후에 대한 영향력이 적음.
- 모델 복잡도 증가에 따른 성능 향상으로 판단됨.



모델 분석 결과

건고추

현재 가지고 있는 데이터로는 정확한 예측과 인사이트 도출에 무리

Tree 모델

양파, 마늘, 대파

- 트리 복잡도만으로 충분한 성능을 보여주는 품목들을 LSTM 모델을 적용했을때 성능 저하
- 해당 품목의 가격과 기후변수 간에 큰 연관성이 없어 noise로 작용

LSTM 모델 (LIME 분석)

1. 가격 예측과 기후 변수 간에 복잡한 관계

배추, 얼갈이 배추, 갯잎, 미나리, 파프리카, 새송이, 토마토, 백다다기

2. 가격과/ 거래량과의 관계

시금치, 팽이버섯

결론

1. 한계
2. 의의



결론

1. 한계 2. 의의

“LIME 알고리즘”

- 단순 상관관계 측면에서의 설명 제공
- conditional한 상황에서의 인과성 측정 어려움

Shap의 경우 해당 인과성을 고려할 수 있지만, conv1d layer에 대한 오류가 존재하여 사용하지 못했다.

“비용변수 데이터 수집”

- 재료비, 노무비, 자산구입비 등의 변수를 일별 데이터로 얻지 못해 최종 데이터로 사용하지 못함
- 구체적인 비용 변수와 순수한 기후 변수의 영향력과 함께 더 높은 성능 도출 기대

“기후변수 정보 손실”

- NA 값이 많은 특정 기후 변수를 제거함으로써 정보 손실 발생
- 이로 인해 무,당근,양배추 품목 분석을 하지 못함



결론

1. 한계 2. 의의

“복잡한 모델 해석 용이”

- 보통 복잡한 모델은 해석하기 어렵다는 단점 존재
→ LIME 알고리즘을 이용한 신경망 해석을 통해 이를 극복.
- 복잡한 영향력을 측정할 때
LIME 알고리즘 외에도 다양한 방법으로 확장할 수 있을 것.

“효율적인 공급 운영”

- 분석 결과를 바탕으로 기후의 영향을 받는 품목은 무엇이며 어떠한 영향을 받는지 파악 가능.
→ 품목별로 기후에 따른 공급 스케줄을 작성하여 급격한 가격 변동에 대비.



참고문헌

사용 데이터 목록

데이터명	출처
농산물_가격_거래량 데이터	데이콘
지역별_일별_경유가격	오피넷
농산물물가지수	통계청
농업기상 관측데이터	공공데이터포털

참고문헌

- LSTM 네트워크를 활용한 농산물 가격 예측 모델, 신성호(2018)

분석도구



python



감사합니다



통 벤 저 스