DACON

제주 특산물 가격 예측 AI 경진대회

2023.11.22

B 부터 N 까지

(정유정, 배수연)

Jeju

JTP 제주테크노파크 JEJU TECHNOPARK

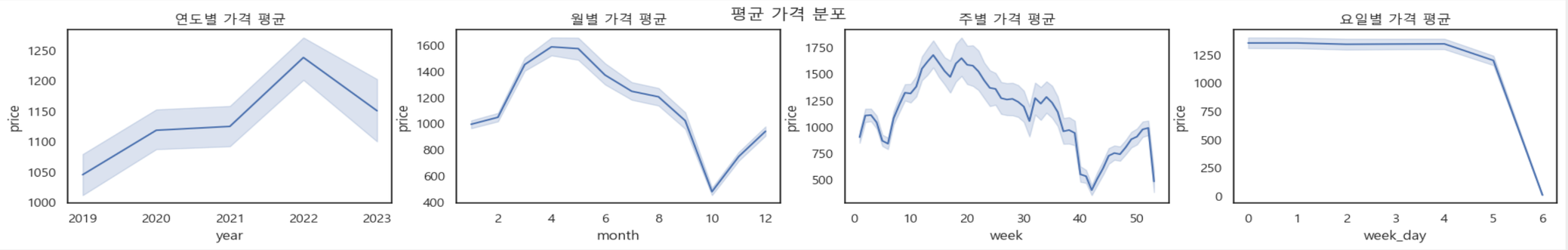
INDEX

데이터 이해

- 기본 정보
  - Train : 2019.01.01 ~ 2023.03.03 / 59,397rows * 7columns - Target : price(원/kg)
  - Test : 2023.03.04 (첫째주 토요일) ~ 2023.03.31 (넷째주 금요일) 28일 / 1,092rows * 5 columns

- 날짜 파생 변수 생성
  - 날짜 데이터에 대한 이해를 위한 'timestamp' 분해 → 년/월/일/요일/주 파생변수 생성

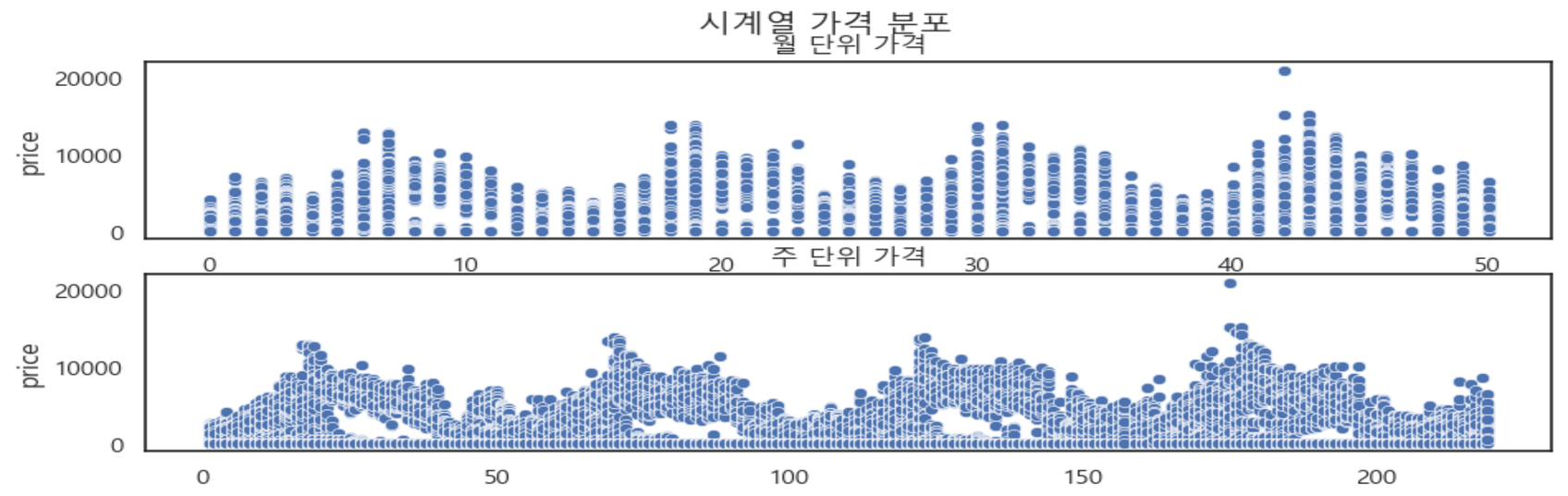
- 날짜 정보 EDA
  1) 연도별 가격 평균을 통해 해가 갈수록 가격 상승 파악 → 시간의 흐름을 파악할 수 있는 분석 및 시계열 파생 변수 생성
  2) 요일별 가격 평균을 통해 일요일에 특산물 값이 0인 것을 발견 공휴일도 거래가 이뤄지지 않을것이라 판단 → 공휴일 파생변수 생성



시계열 파생 변수 EDA

- 시계열 파생 변수 생성
  - 시간의 흐름에 따른 가격 확인을 위해 추가 변수 생성 → year_month (누적 월 : 1개월~50개월), week_num (누적 주차 :1주 ~250주)

- 시계열 정보 EDA
  - 시계열 가격 분포를 통해, 가격변동의 일정한 패턴이 존재함을 파악했으며, 이상치가 존재함을 파악 → 상세 분석 필요

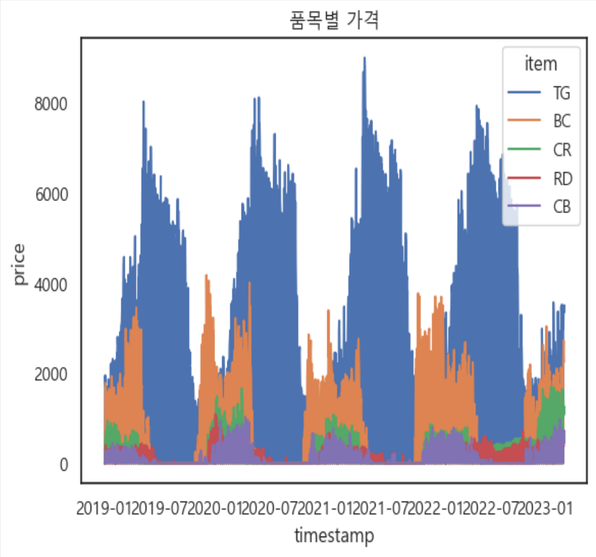


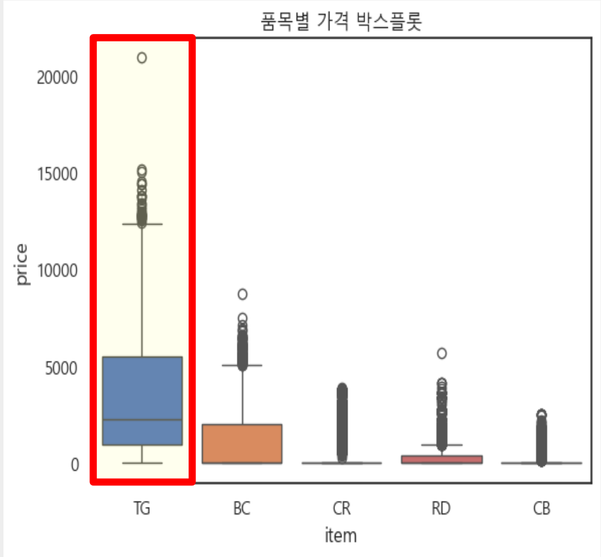
품목별 EDA

- 제주 특산물 품목별 특성의 차이가 있음

- TG의 특성 차이가 두드러짐
  1) 가격의 범위가 넓다
  2) 가격이 0값이 아닌 데이터의 비율이 높아, 다른 품목에 비해 0의 비율이 현저히 낮다.

→ ‘TG’ / ‘TG 외 품목’ 별도 프로세스 필요





● 특성별 전처리 : TG

1) 가격의 범위가 넓다

: 다른 품목보다 가격 편차가 커, 예측의 어려움이 있을것으로 판단

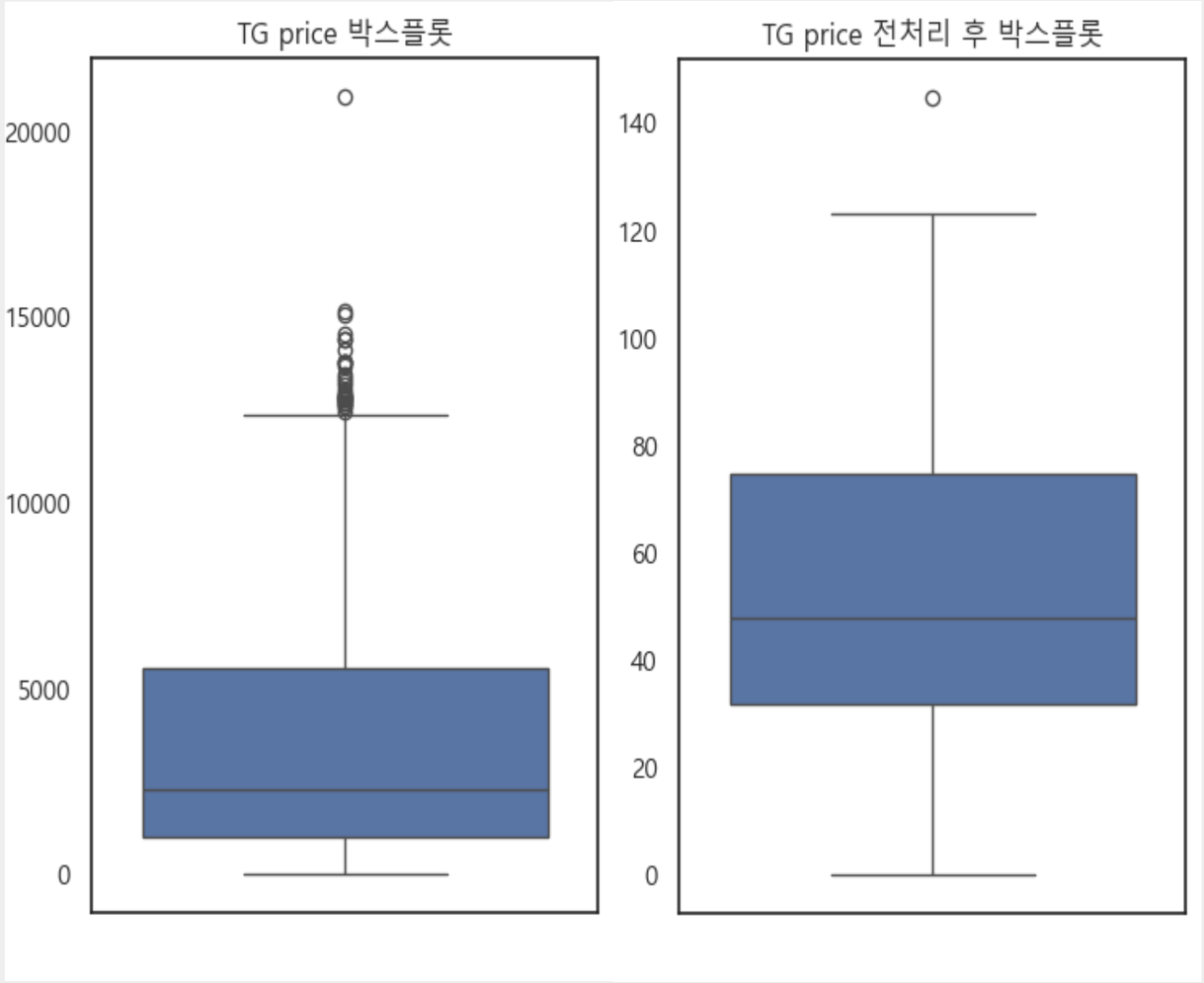
→ Price에 루트, 가격의 범위를 좁힘

2) 다른 품목에 비해 price 0의 비율이 현저히 낮다.

: price 0값/ 0값이 아닌것에 대한 명확한 구분이 필요하다고 판단

→ holidays 수정

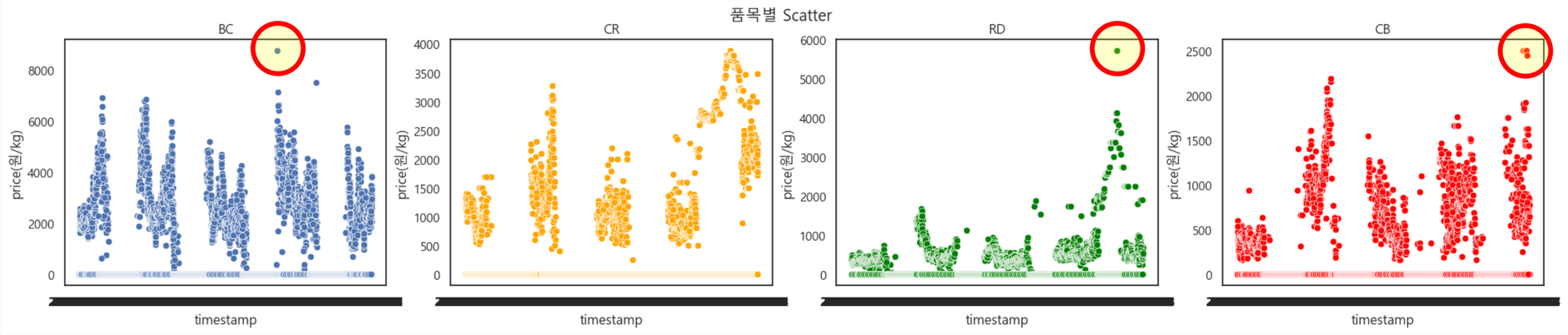
(대체 공휴일, 명절 첫번째 날은 공휴일임에도 가격정보가 존재하기 때문)



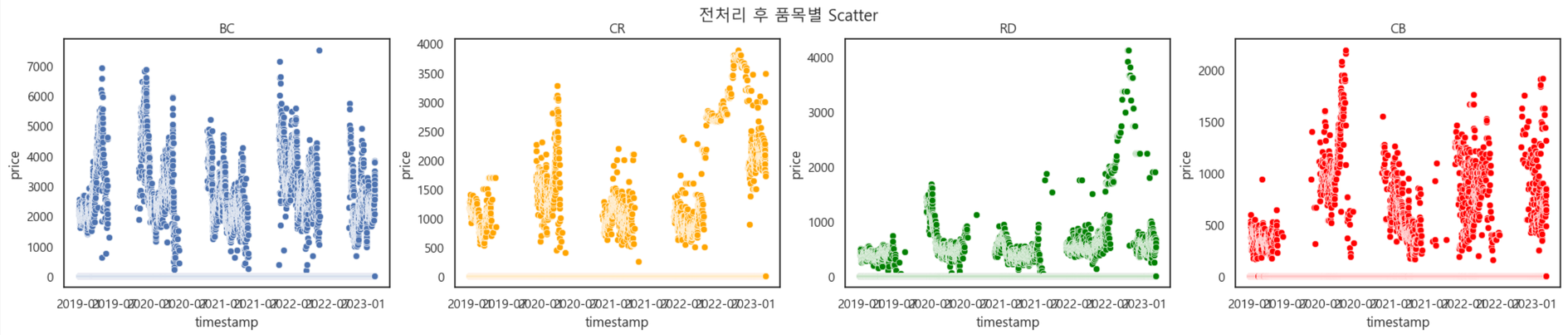
특성별 전처리 : TG 외 품목

- 극이상치 값이 명확한 품목(BC,RD,CB) 을 품목의 평균으로 대체함

BEFORE



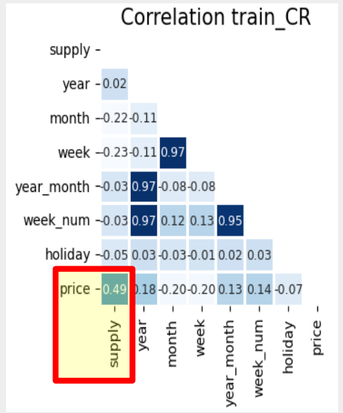
AFTER

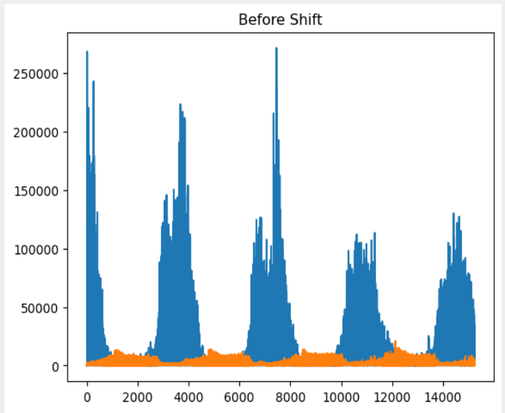


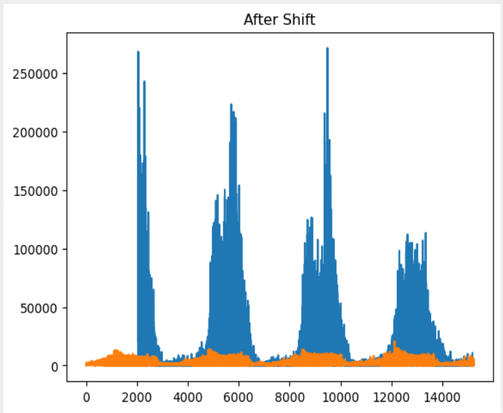
TRY & ERRORS

Supply 예측

: supply 와 price가 높은 상관을 가지기 때문에, supply 예측값으로 price를 예측하고자 시도







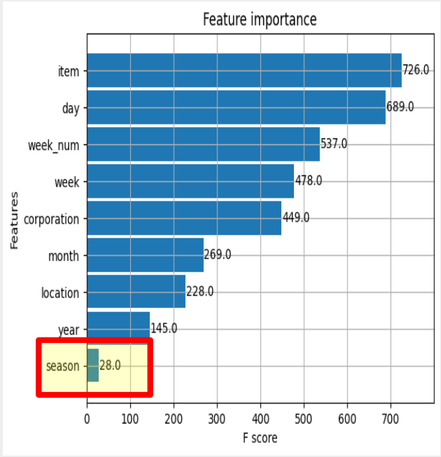
※ Try 2

- Try 1 : Supply 값을 품목별로 범주화 하여 예측하도록 함
- Try 2 : 감귤의 경우 supply 변수를 이동시킬 때 price와 패턴이 비슷해지는 것을 확인했으며, 이동된 supply변수로 price 예측을 시도함

→ Supply 예측값이 정확하지 않아, 성능저하의 원인이 돼 **supply 변수 제외**

Season 파생 변수

: 농작물의 경우 계절의 영향이 클 것으로 판단



- 변수 영향도를 확인한 결과, 가장 낮은 중요도를 가짐

→ **season 변수 제외**

Summary

	Insight	Applied
파생변수의 생성	1. 시계열 예측 및 분석을 위한 기본 변수 필요 2. 유통 물량 및 가격이 존재하지 않는 날 발견 3. 시간의 흐름 파악 중요	1. 년/월/일/요일/주 파생변수의 생성 2. 공휴일 변수 추가 3. Year_num, Week_num 파생 변수 추가
분석 방향	1. ‘TG’와 ‘TG외 품목’의 특성 차이가 남 2. TG 품목의 Price 특성이 두드러짐	1. ‘TG’ 와 ‘TG 외 품목’ 프로세스를 다르게 함 2-1. Price에 루트를 씌워 예측 편차를 줄임. 2-2. 0값과 0값이 아닌 값을 확실히 하기 위해 공휴일이지만 쉬지 않는 날의 공휴일 변수 수정

최종 Train Data

- 훈련데이터로 사용하기 위해 범주형 변수(item, corporation,location)는 one-hot encoding 진행
- 훈련 사용 Feature : ‘TG외 품목 모델’ 20개 / ‘TG 모델’ 16개

  - 'year', 'month', 'day', 'week_day', 'year_month', 'week', 'week_num', 'holiday’, 'corporation_A', 'corporation_B', 'corporation_C’, 'corporation_D', 'corporation_E', 'corporation_F', 'location_J', 'location_S’, ('item_BC', 'item_CB', 'item_CR', 'item_RD’)

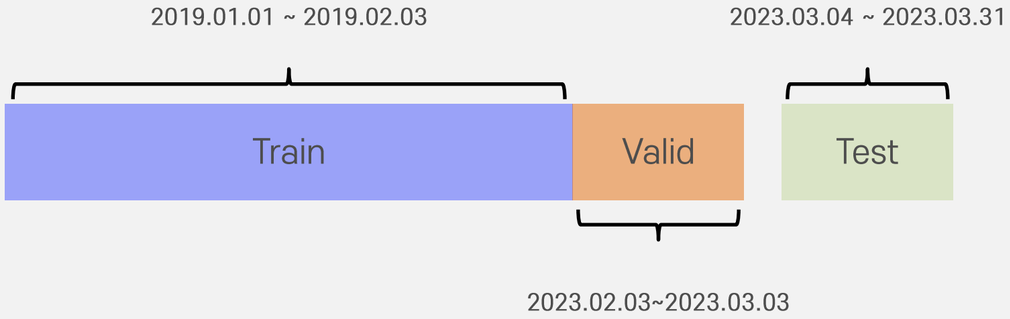
	ID	timestamp	item	corporation	location	price	year	month	day	week_day	year_month	week	week_num	holiday
0	TG_A_J_20190101	2019-01-01	TG	A	J	0.0	2019	1	1	1	0	1	1	1
1	CB_A_S_20190101	2019-01-01	CB	A	S	0.0	2019	1	1	1	0	1	1	1
2	RD_D_J_20190101	2019-01-01	RD	D	J	0.0	2019	1	1	1	0	1	1	1
3	BC_D_J_20190101	2019-01-01	BC	D	J	0.0	2019	1	1	1	0	1	1	1
4	CB_F_J_20190101	2019-01-01	CB	F	J	0.0	2019	1	1	1	0	1	1	1
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
59392	CR_E_S_20230303	2023-03-03	CR	E	S	0.0	2023	3	3	4	50	9	219	0
59393	BC_A_S_20230303	2023-03-03	BC	A	S	2875.0	2023	3	3	4	50	9	219	0
59394	CB_E_J_20230303	2023-03-03	CB	E	J	0.0	2023	3	3	4	50	9	219	0
59395	BC_D_J_20230303	2023-03-03	BC	D	J	3059.0	2023	3	3	4	50	9	219	0
59396	RD_F_J_20230303	2023-03-03	RD	F	J	529.0	2023	3	3	4	50	9	219	0

59397 rows × 15 columns

성능평가 전략 (TSCV)

초반 성능 평가

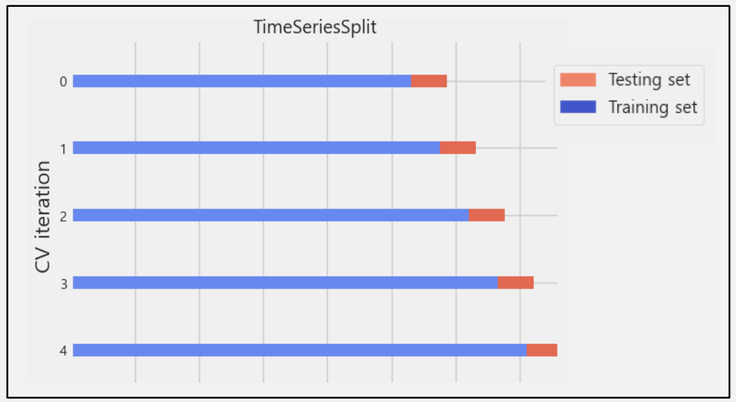
- 1개의 Valid Set(2월)을 이용해 성능 평가
- Valid Score 성능이 높아질수록 Test set 가격 예측 성능이 떨어지는 현상 발생



→ 모델이 2월에 맞춰져 학습되어 일반화 된 성능 평가불가

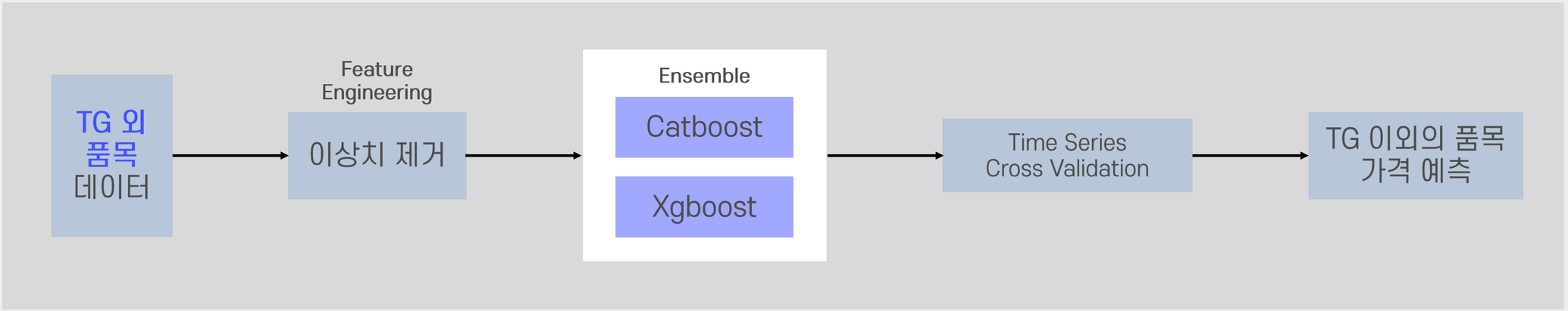
개선된 성능 평가

- 특정 달에 맞춰진 학습을 방지하기 위해 Cross Validation 평가 필요
- 시간의 순서가 보존된 Time Series Cross Validation 수행
- Valid Set을 Test set과 동일한 28일로 고정시키고, Train Set을 점진적으로 늘려가며 Cross Validation 수행

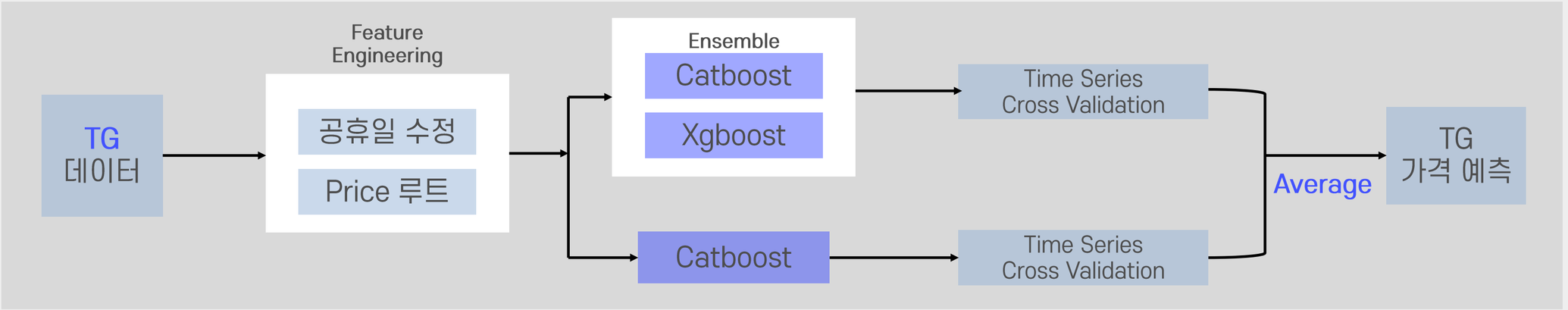


Model Overview

Model 1



Model 2

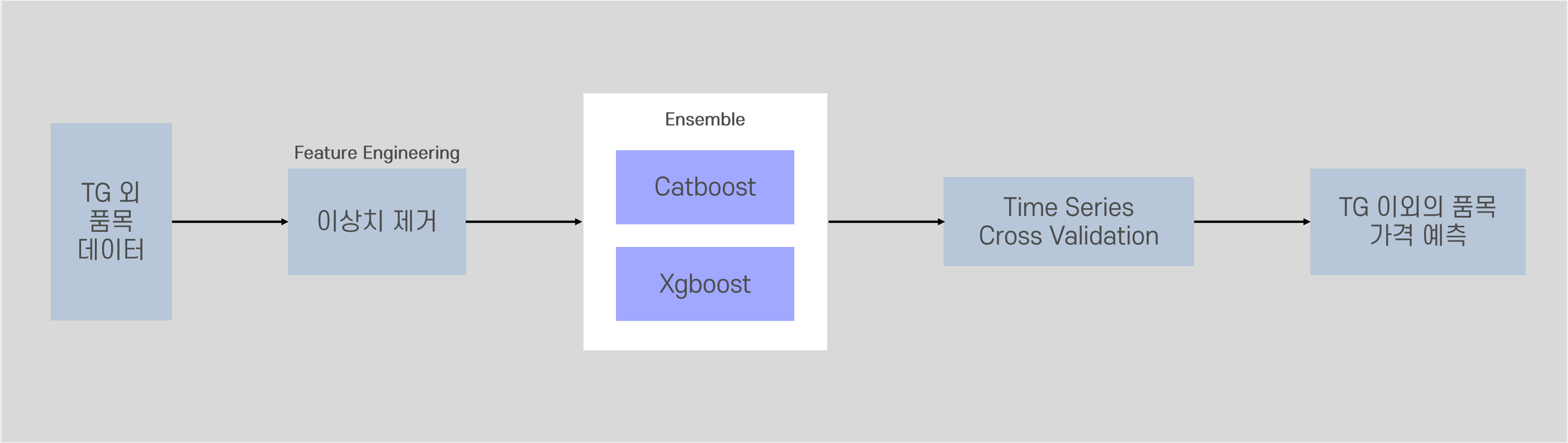


Model 1 : TG 외 품목 모델

- Catboost는 범주형 예측에 탁월한 성능을 보이고, XGBoost는 과적합 방지에 탁월함.
- 단일 모델 보다 앙상블 모델이 0값에 대한 예측 정확도가 높음

- Feature Engineering
  - 이상치 제거

- Modeling
  - Catboost 및 Xgboost 앙상블



Model 2 : TG 모델

TG1 모델

- Catboost + Xgboost 의 앙상블
- 장점 : 2019~2022년 3월의 종합적인 패턴을 가장 잘 반영함
- 단점 : TG2 모델에 비해 Public 리더보드 성능이 떨어짐

```
# 앙상블 모델 정의
cat = CatBoostRegressor(random_state = 2024,
                        n_estimators = 1000,
                        learning_rate = 0.01,
                        depth = 10,
                        l2_leaf_reg = 3,
                        metric_period = 1000)

xgb = XGBRegressor(n_estimators = 1000, random_state = 2024,
                   learning_rate = 0.01, max_depth = 10)


# voting
vote_model = VotingRegressor(estimators =[("cat",cat), ("xgb", xgb)])

vote_model.fit(Xy.drop(columns = ["timestamp", "ID","price"]), Xy["price"])

pred = vote_model.predict(answer_tg1.drop(columns = ["ID"]))
```

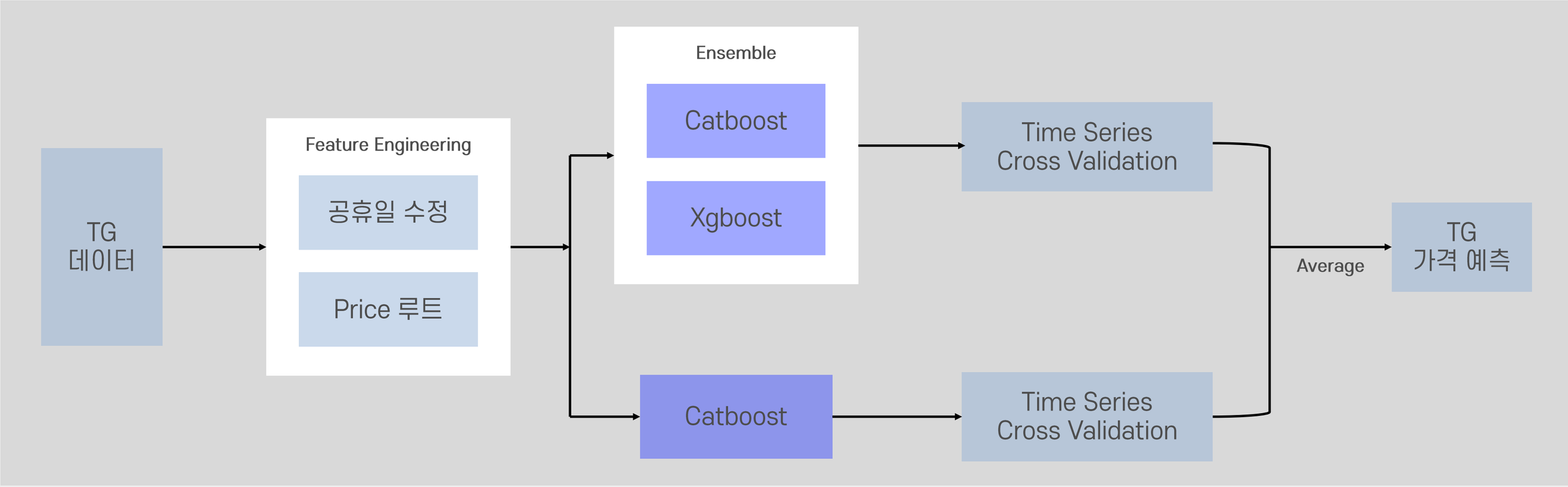
TG2 모델

- Catboost 단일 모델
- 장점 : Public 리더보드 평가 에서 가장 좋은 성능을 보임
- 단점 : TG1 모델과의 예측값 차이가 17일 이후에 몰려 있어, 일반화 성능이 좋지 않을 것으로 판단됨.

```
# 모델 정의 및 훈련 예측
n_estimators =1000
lrs = 0.05
max_depths = 10
l2_leaf_reg = 3

cat = CatBoostRegressor(random_state = 2024,
                        n_estimators = n_estimators,
                        learning_rate = lrs,
                        depth = max_depths,
                        l2_leaf_reg = l2_leaf_reg,
                        metric_period = 1000)

cat.fit(Xy2.drop(columns = ["timestamp", "ID","price"]), Xy2["price"])

pred2 = cat.predict(answer_tg2.drop(columns = ["ID"]))
```

Model 2 : TG 모델

- TG1의 장점과 TG2의 장점을 합치기 위해 두 결과값을 평균 낸다!
- Feature Engineering
  - 공휴일 값 수정
  - TG 값 범위를 조정하기 위해 TG price 루트 씌우기
- Modeling
  - 앙상블(Catboost+Xgboost)결과와 단일모델(Catboost)결과의 평균



0값 처리

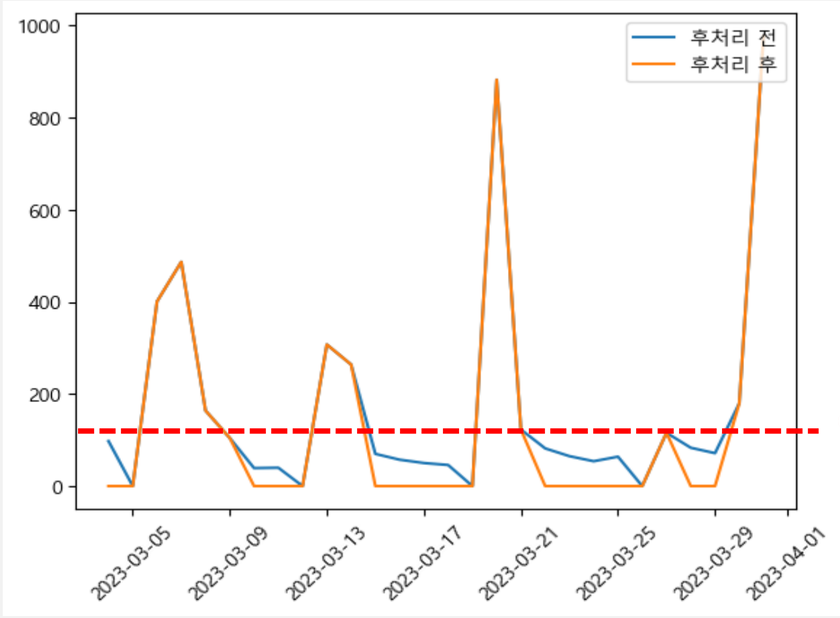
- 품목별 0 값을 제외한 price의 최솟값을 확인하고, 최솟값보다 작은 값을 0으로 처리함

가격 별 최솟값

품목	최소 가격
감귤(TG)	551
브로콜리(BC)	205
당근(CR)	250
양배추(CB)	162
무(RD)	50

모델이 제대로 예측 못한 0값을 확실히 잡기 위해 품목별
최솟값보다 더 작은 값을 임곗값으로 잡아
임곗값 보다 작은 값을 모두 0으로 변환한다.

후처리 결과



임곗값 보다 작은 값이 0으로 변환된 것을 확인할 수 있다.

결과

- Private Score : 812.55943

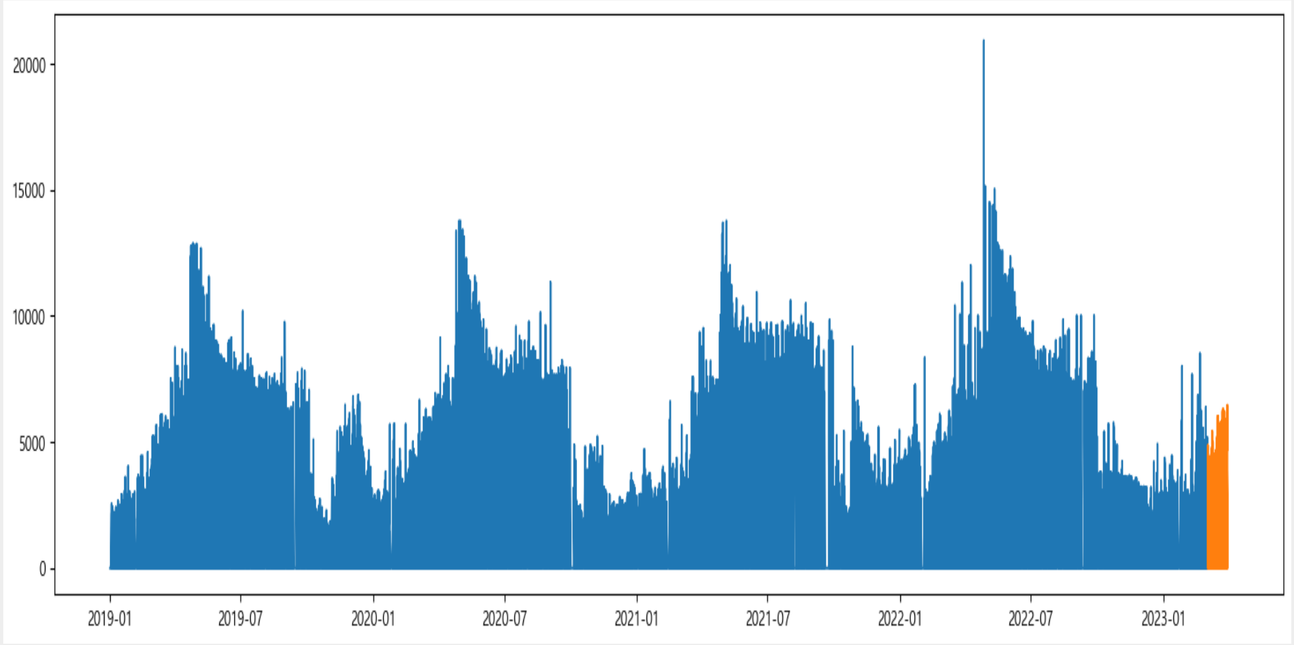
KEYWORD 01 데이터 분석(EDA)을 통한 가격 변동 패턴 파악에 포커스

KEYWORD 02 TG / TG외 품목 특성에 적합한 전처리 및 모델사용

KEYWORD 03 Time Series Cross Validation 성능 평가

CONCLUSION 일반화 된 모델

Test set forecasting



WINNER 1% 4% 10%

#	팀	팀 멤버	최종점수
1	B부터N까지	저어 hh	812.55943

Q & A

감사합니다