농산물 가격 예측

소비자와 공급자를 위한 정보 제공과 시장 안정화

2024.07.06

한예강

Contents

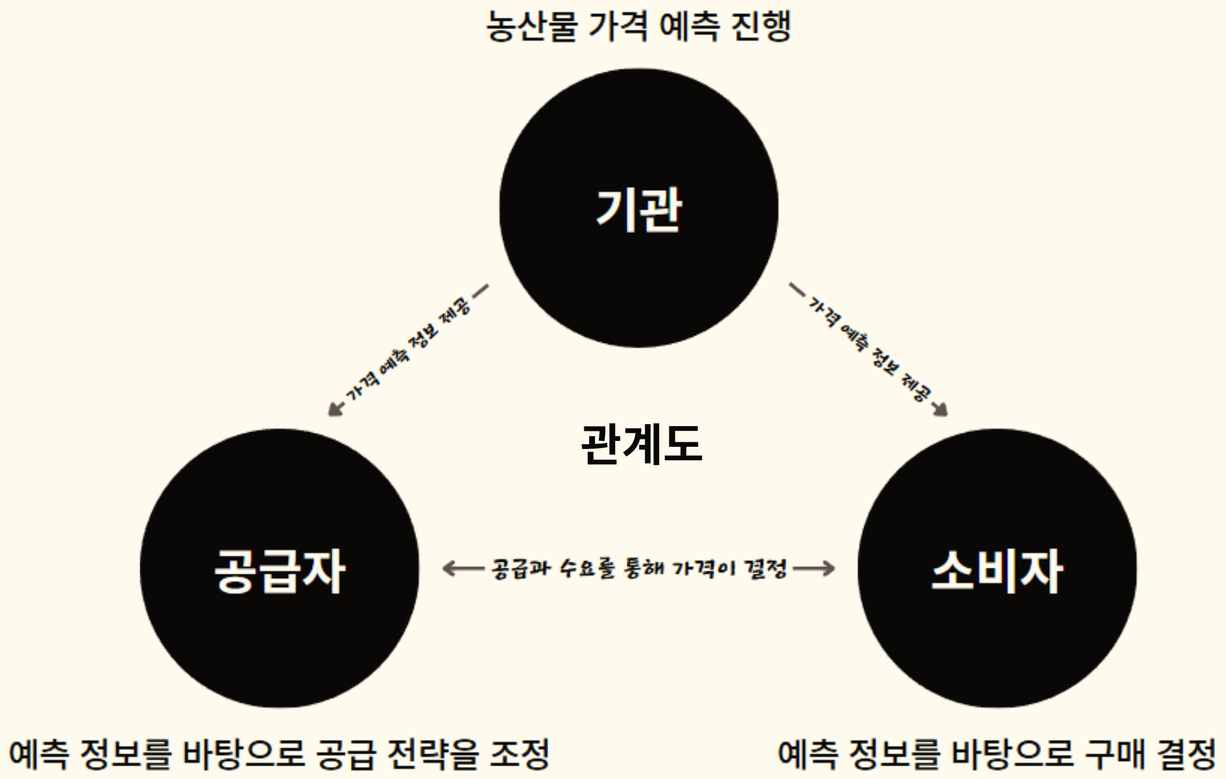
01 문제 정의

주제 소개

농산물 가격의 변동성은 농업 생산자와 소비자 모두에게 큰 영향을 미칩니다. 농산물 가격은 기후 변화, 수요와 공급의 변동, 정책 변화 등 다양한 요인에 의해 좌우됩니다. 이러한 가격 변동을 예측한다면, 생산자는 더 나은 의사 결정을 내릴 수 있으며, 소비자는 보다 합리적인 가격에 농산물을 구매할 수 있습니다.

문제 제기

수요와 공급, 정책 변화 데이터를 완벽하게 수집하기 어려운 상황에서, 기후 변화 데이터를 활용하여 농산물 가격 예측 모델의 성능을 향상시키고자 합니다. 기후 데이터는 공공기관에서 제공하는 신뢰성 높은 자료를 기반으로 하며, 기후 변화와 농업 생산성 간의 상관관계를 분석함으로써 보다 정확한 가격 예측이 가능할 것입니다.



02 문제 해결을 위한 접근법

주요 분석 대상

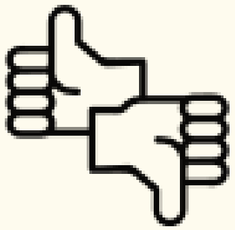
대한민국 정부가 식품 안전 등 다양한 측면에서 관리하고 지원하는 주요 농산물의 가격을 예측하기 위한 데이터를 활용했습니다. 주요 농산물: 배추, 무, 마늘, 양파, 대파, 시금치, 깻잎

데이터 분석

2016/01/01~2020/11/04 데이터 사용
train data: 16/01/01~20/09/28
test data: 20/09/29~20/11/04
요약 통계, 상관 관계, 기술 통계 정보 분석

변수 및 요인 분석

기존 Dataset에서 얻을 수 있는 정보로는 학습이 어렵기에, 기상청 API를 이용하여 날씨 데이터와 Prophet이 계산한 추세와 계절성 데이터를 추가 분석



비교 분석

여러 ML 모델을 비교 분석
사용된 모델 리스트:
linear regression model (Ridge)
tree-based model
(RandomForest, LightGBM, XGBoost)

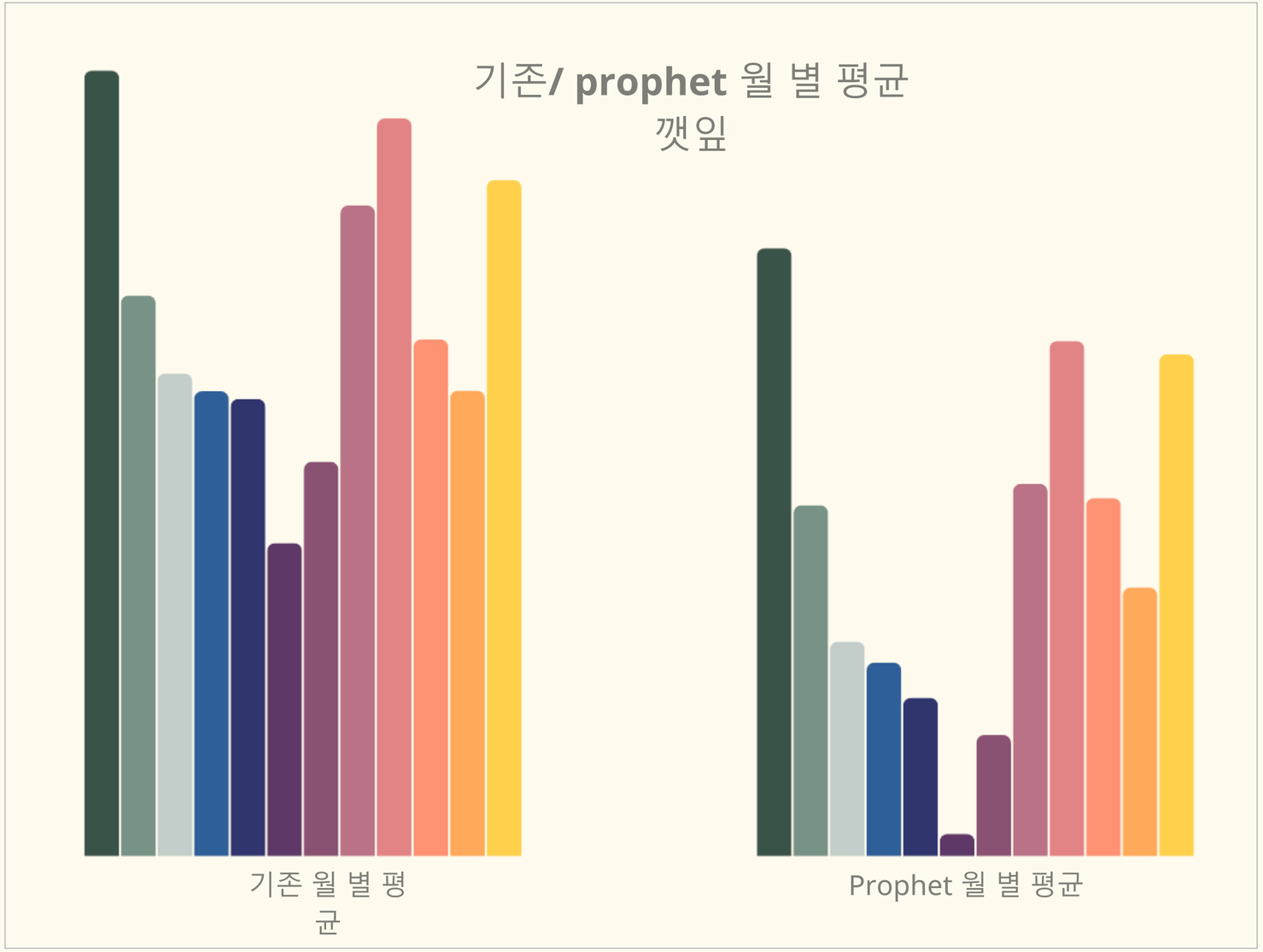
02 문제 해결을 위한 접근법

Prophet 모델의 특징

Prophet은 계절성과 추세를 모델링하는 데 강점을 가지고 있으며, 이 정보들을 **Feature Engineering**을 통해 추출하고 **ML** 모델에 입력함으로써 더 정교한 예측 모델을 구축하고자 합니다.

Prophet 모델의 예측 성능 비교

Prophet 모델의 예측 성능을 비교한 결과, 기존 데이터로 계산한 값과 유사함을 보여줍니다. 이는 **Prophet** 모델이 농산물 가격 예측에 적합한 도구임을 시사합니다.



03 EDA

	cabbage	radish	garlic	onion	daikon	cilantro	artichoke
count	1733	1733	1733	1733	1733	1733	1733
mean	597	490	3433	652	1160	9537	4614
std	410	299	1751	366	645	8757	2733
min	0	0	0	0	0	0	0
25%	335	324	2687	467	853	6015	3387
50%	577	453	3953	638	1188	8607	4492
75%	805	717	4758	925	1551	12500	5942
max	5000	1360	6415	2000	3182	170836	14326

요약 통계 정보

각 품목의 평균 가격을 통해 일반적인 가격 수준을 파악할 수 있습니다. 예를 들어, 고수의 평균 가격이 9537으로 다른 품목에 비해 상당히 높습니다.

	cabbage	radish	garlic	onion	daikon	cilantro	artichoke
cabbage	1.00	0.90	0.69	0.63	0.62	0.53	0.64
radish	0.90	1.00	0.72	0.66	0.85	0.58	0.76
garlic	0.69	0.72	1.00	0.98	0.78	0.91	0.80
onion	0.63	0.66	0.98	1.00	0.77	0.92	0.76
daikon	0.62	0.85	0.78	0.77	1.00	0.69	0.85
cilantro	0.53	0.58	0.91	0.92	0.69	1.00	0.79
artichoke	0.64	0.76	0.80	0.76	0.85	0.79	1.00

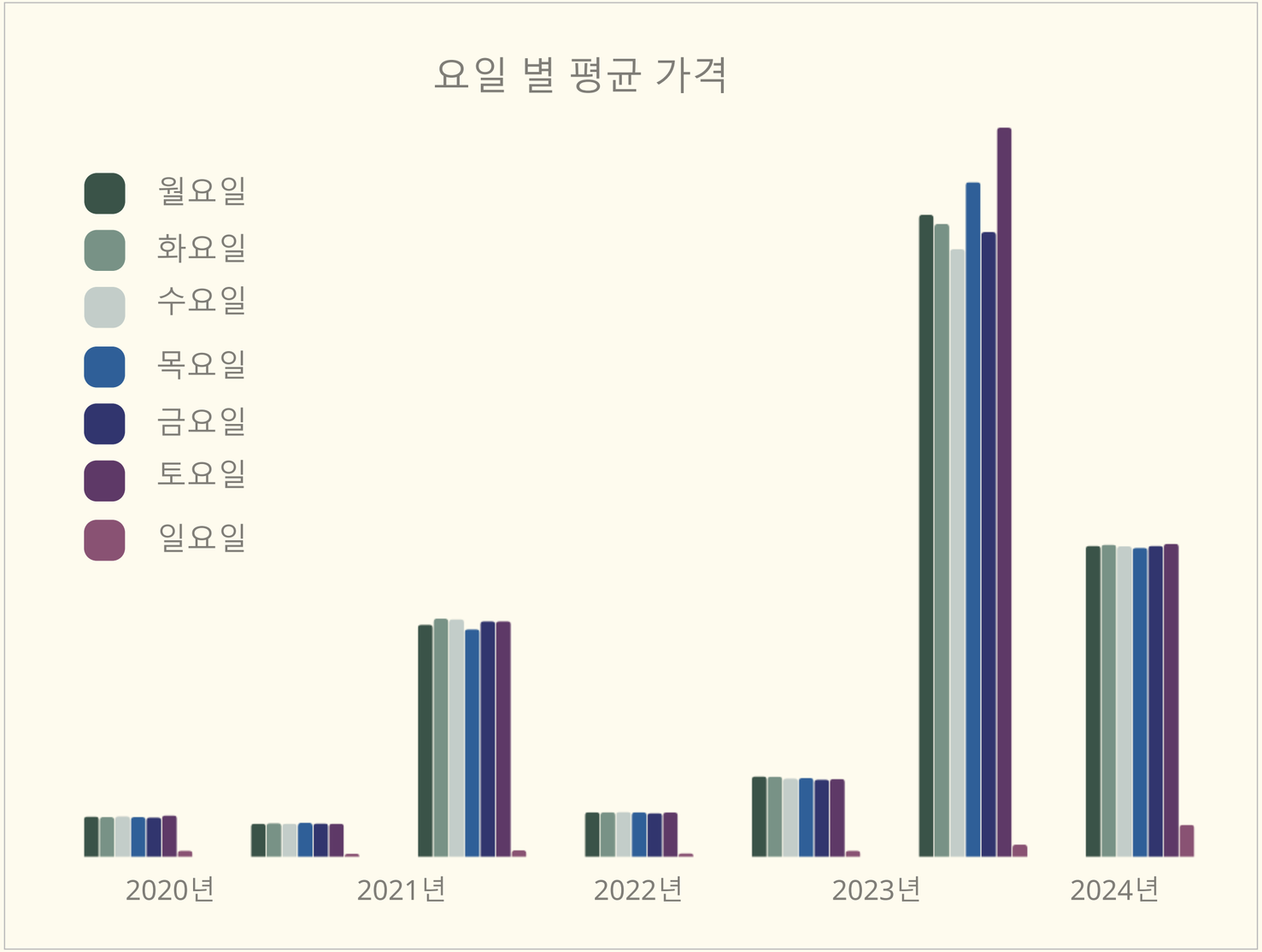
상관 관계

상관 관계가 높을 때, 한 품목의 가격이 상승하면 다른 품목의 가격도 함께 상승하는 경향이 있습니다. 상관 관계가 높은 품목들을 묶어 한 품목의 재고가 부족할 때 다른 품목의 가격 변동을 예측할 수 있습니다.

03 EDA

요일 별 평균 가격 데이터를 분석한 결과, 일요일마다 데이터가 존재하지 않는 것을 확인하였습니다.

이를 통해 농산물 가격 예측 모델을 설계할 때, 일요일을 제외한 데이터를 중점적으로 활용하고 일요일 데이터를 보완하는 방법을 고려할 수 있습니다. 이러한 접근 방식은 보다 정확하고 신뢰성 있는 예측 모델을 구축하는 데 기여할 것입니다.



04 Feature Engineering

가격 예측

- 효율성 증대: 예측된 가격을 바탕으로 농업 생산 및 유통의 효율성을 극대화할 수 있습니다.
- 비용 절감: 가격 변동을 미리 알고 대비함으로써 비용 절감 효과를 기대할 수 있습니다.
- 위험 관리: 예측된 가격 정보를 통해 농업 생산과 관련된 위험 요소를 사전에 파악하고 관리할 수 있습니다.

기본 데이터

- dataset 내에서 얻을 수 있는 정보 추출
- holiday와 rolling mean/std 추가

날씨 데이터

- 기상청 API 허브에서 제공하는 날씨 데이터를 사용
- 주요 특징 : TA (기온), HM (상대 습도), RN (강수량), SD (적설), TS (지면 온도)

Prophet 데이터

- 학습된 Prophet 모델에 저장된 데이터를 사용
- 주요 특징 : trend, yearly, monthly, weekly, holiday

일반 데이터

date	요일	배추_거래량(kg)	배추_가격(원/kg)	무_거래량(kg)	무_가격(원/kg)	양파_거래량(kg)	양파_가격(원/kg)
2016-01-01	금요일	0.0	0.0	0.0	0.0	0.0	0.0
2016-01-02	토요일	80860.0	329.0	80272.0	360.0	122787.5	1281.0
2016-01-03	일요일	0.0	0.0	0.0	0.0	0.0	0.0
2016-01-04	월요일	1422742.5	478.0	1699653.7	382.0	2315079.0	1235.0
2016-01-05	화요일	1167241.0	442.0	1423482.3	422.0	2092960.1	1213.0

데이터 분석

데이터를 확인해보면 날짜, 요일, 각 농산물의 거래량과 가격에 대한 정보가 들어있는 것을 확인할 수 있습니다.
미래의 가격을 예측하려면, 그 시점의 거래량을 미리 알아야 하는데, 이는 불가능합니다. 그렇기에 데이터에서는 날짜 데이터를 이용해 year, month, day, day_of_weak, is_weak 등을 추출하고, holiday라이브러리를 사용해 한국의 휴일 정보도 추가했습니다.

이동 평균/표준편차

이동 평균과 이동 표준 편차를 계산함으로써, 최근 데이터의 패턴과 변동성을 모델에 반영하였습니다.
거래량과 마찬가지로, 이동 평균과 이동 표준 편차도 예측 시점에서는 미래의 값을 알 수 없습니다. 따라서, 최근 일주일 간의 이동 평균과 이동 표준 편차를 사용합니다.

features = ['year', 'month', 'day', 'day_of_week', 'is_weekend', 'is_holiday'] + [각 품목 당 이동 평균/ 표준편차]

날씨 데이터

기상청 API 허브의 지상 관측 자료 데이터를 사용합니다.
용이한 해석을 위해 TA (기온), HM (상대 습도), RN (강수량), SD (적설) 등의 필요한 정보만 사용했습니다.

시간에 따른 변화를 알 수 없기 때문에 제외하였습니다.

features = ['TA', 'HM', 'RN', 'SD' 관련 정보]

변수명	의미(단위)	변수명	의미(단위)
TM	관측시각 (KST)	STN	국내 지점번호
RN_DAY	위 관측시간까지의 일강수량 (mm)	CA_TOT	전운량 (1/10)
WS_AVG	일 평균 풍속 (m/s)	WR_DAY	일 풍정 (m)
WD_MAX	최대풍향	WS_MAX	최대풍속 (m/s)
WS_MAX_TM	최대풍속 시각 (시분)	WD_INS	최대순간풍향
WS_INS	최대순간풍속 (m/s)	WS_INS_TM	최대순간풍속 시각 (시분)
TA_AVG	일 평균기온 (C)	TA_MAX	최고기온 (C)
TA_MAX_TM	최고기온 시가 (시분)	TA_MIN	최저기온 (C)
TA_MIN_TM	최저기온 시각 (시분)	TD_AVG	일 평균 이슬점온도 (C)
TS_AVG	일 평균 지면온도 (C)	TG_MIN	일 최저 초상온도 (C)
HM_AVG	일 평균 상대습도 (%)	HM_MIN	최저습도 (%)
HM_MIN_TM	최저습도 시각 (시분)	PV_AVG	일 평균 수증기압 (hPa)
EV_S	소형 증발량 (mm)	EV_L	대형 증발량 (mm)
FG_DUR	안개계속시간 (hr)	PA_AVG	일 평균 현지기압 (hPa)
PS_AVG	일 평균 해면기압 (hPa)	PS_MAX	최고 해면기압 (hPa)
PS_MAX_TM	최고 해면기압 시각 (시분)	PS_MIN	최저 해면기압 (hPa)
PS_MIN_TM	최저 해면기압 시각 (시분)	SS_DAY	일조합 (hr)
SS_DUR	가조시간 (hr)	SS_CMB	캄벨 일조 (hr)
SI_DAY	일사합 (MJ/m2)	RN_D99	9-9 강수량 (mm)
RN_DUR	강수계속시간 (hr)	SD_NEW	최심 신적설 (cm)
SD_NEW_TM	최심 신적설 시각 (시분)	SD_MAX	최심 적설 (cm)
SD_MAX_TM	최심 적설 시각 (시분)	TE_05	0.5m 지중온도 (C)
TE_10	1.0m 지중온도 (C)	TE_15	1.5m 지중온도 (C)
TE_30	3.0m 지중온도 (C)	TE_50	5.0m 지중온도 (C)
SI_60M_MAX	최대 1시간일사 (MJ/m2)	SI_60M_MAX_TM	최대 1시간일사 시각
RN_60M_MAX	1시간 최다강수량 (mm)	RN_60M_MAX_TM	1시간 최다강수량 시각
RN_10M_MAX	10분간 최다강수량 (mm)	RN_10M_MAX_TM	10분간 최다강수량 시각 (시분)
RN_POW_MAX	최대 강우강도 (mm/h)	RN_POW_MAX_TM	최대 강우강도 시각 (시분)

Prophet 데이터

Prophet 모델은 시계열 예측 알고리즘으로 정확도가 높고 빠르며, 직관적인 파라미터를 통해 모델 수정이 용이하다는 장점을 갖고 있습니다. 또한, 데이터를 학습할 때 Trend, Seasonality, Holiday 등을 계산하여 학습에 반영함으로써 성능을 높입니다.

이렇게 계산된 Trend, Seasonality (yearly, weekly), Holiday 정보를 사용합니다.

features = [각 품목의 'trend', 'yearly', 'weekly', 'monthly', 'holiday']

각 품목의 Trend, Seasonality, Holiday 결과는 아래 링크에서 확인 가능합니다.
(본 프로젝트에서 사용된 코드와는 차이가 있습니다.)

https://github.com/disl9907/ml_study/blob/main/prophet_params.ipynb

05 모델링

사용된 모델 리스트

- Ridge Regression

선형 회귀 모델에 L2 정규화를 적용하여 과적합을 방지하는 모델입니다.

- Random Forest

여러 개의 결정 트리를 앙상블 방식으로 결합하여 예측 성능을 향상시키는 모델입니다.

- LightGBM

대용량 데이터와 높은 효율성을 위해 설계된 Gradient Boosting Machine(GBM) 기반의 모델입니다.

- XGBoost

성능과 속도를 극대화하기 위해 설계된 Gradient Boosting Machine(GBM) 기반의 모델입니다.

성능 측정 방법

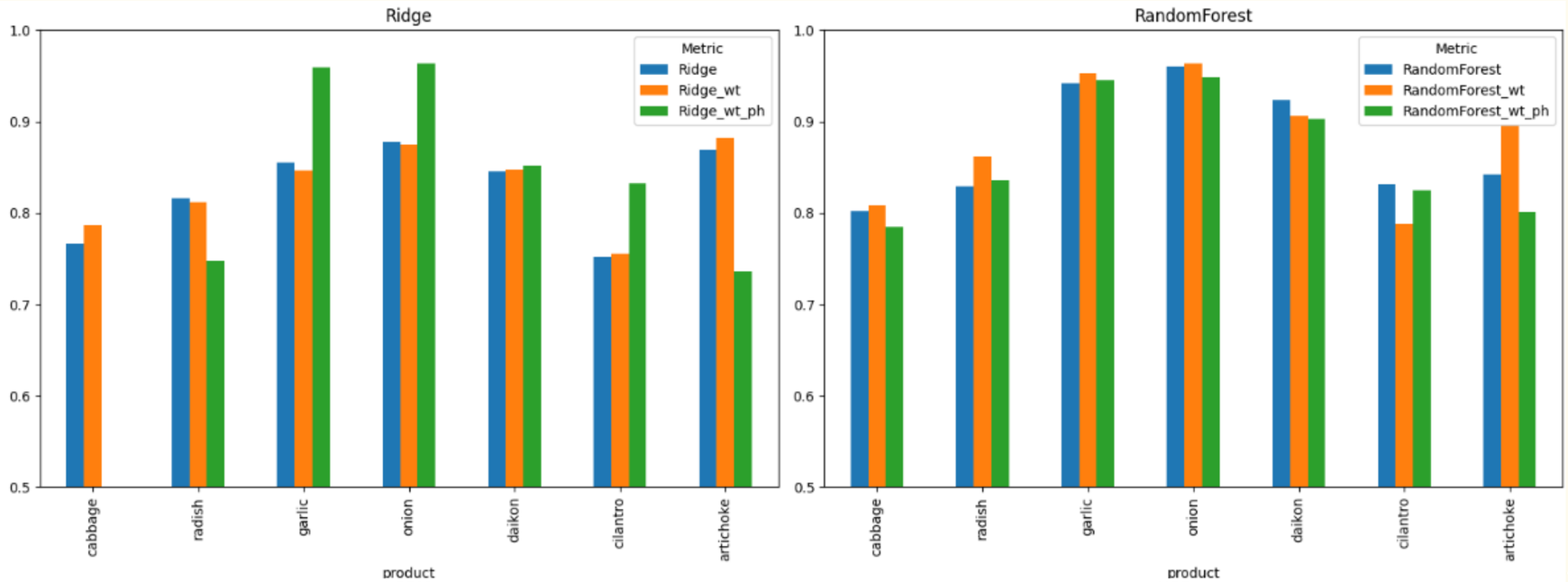
- MdAPE

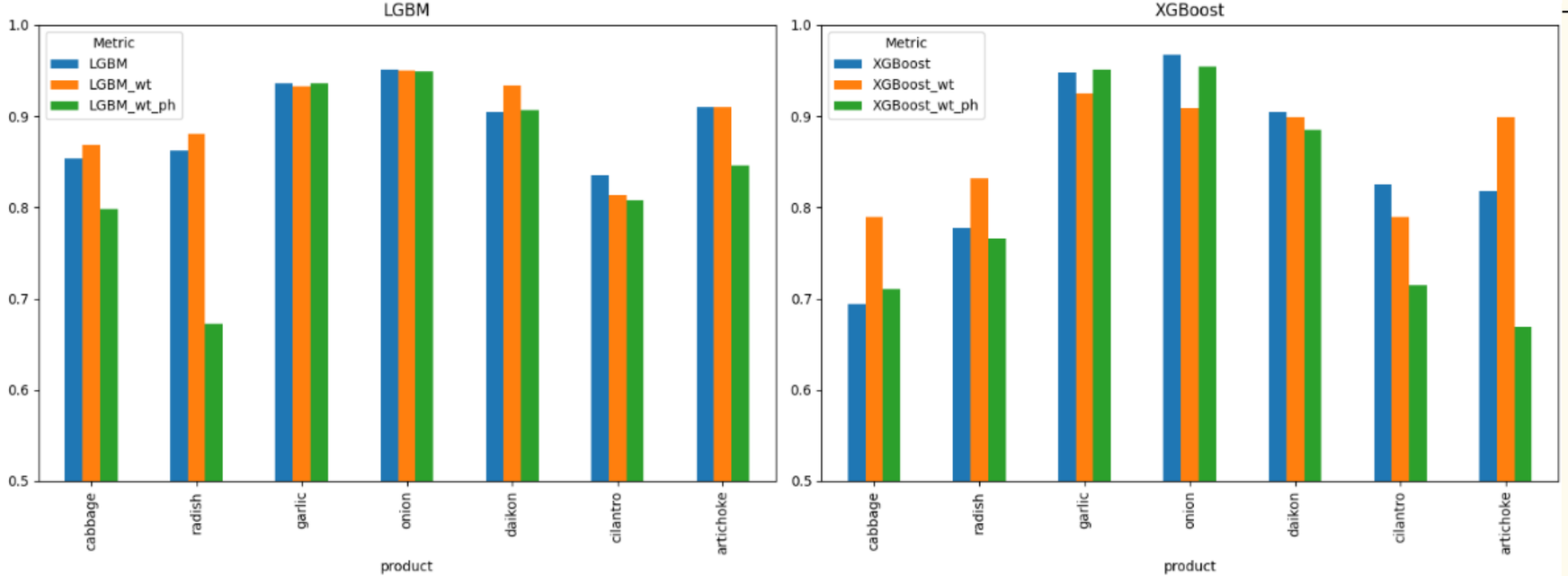
농산물 가격 예측에서는 데이터의 변동성이 크기 때문에, 평균을 사용하는 평가 지표는 특정 기간의 극단적인 오차에 크게 영향을 받을 수 있습니다. MdAPE는 예측 오차의 중앙값을 사용하므로, 예측의 정확성을 보다 직관적으로 이해할 수 있습니다.

프로젝트 수행 방법

- google colaboratory

06 결과 분석





Weather features만 추가했을 때 성능이 향상되었지만, Prophet features를 추가했을 때 좋지 않는 성능을 보이는 것으로 확인됩니다.
그래프 분석 결과, RandomForest가 가격 예측 모델로 가장 적합하다고 판단됩니다.

07 향후 계획 및 발전 방향

이번 프로젝트의 초기 목표는 날씨 데이터를 활용하여 농산물 가격 예측 모델의 성능을 향상시키는 것이었습니다. 실제 날씨 데이터를 포함했을 때 성능이 향상되는 경향이 있음을 확인할 수 있었습니다.

여러 ML 모델을 비교 분석한 결과, RandomForest 모델이 안정적인 예측 정확도를 보여주었습니다. 이를 통해 기관이 시장 가격 변동을 예측하고, 공급자와 소비자에게 정보를 제공하여 적절한 재배 및 판매 전략을 수립하거나 구매 결정을 하는 데 도움이 될 수 있습니다.

향후에는 Overfitting을 예방하기 위한 방안을 고민하고, Feature Selection과 Parameter Tuning을 진행하여 모델의 예측 성능을 높이고자 합니다.