쥬혁이 팀 보고서



목차

활용 데이터 3

데이터 처리 5

데이터 분석 6

데이터 활용 14

예측 모델 17

예측 모델 학습 19

예측 모델 평가 22

1.활용 데이터

* 라이브러리

-os : 운영 체제와 상호작용을 위한 라이브러리. 파일 및 디렉터리 경로 설정, 환경 변수 접근 등 다양한 작업에 사용됨.

-joblib: Python의 데이터 직렬화 및 역직렬화 라이브러리로, 객체를 파일에 저장하거나 불러오는 데 사용됨. 병렬 처리를 쉽게 할 수 있다는 특징이 있음.

-concurrent.futures: 병렬 처리를 위한 라이브러리로, 다중 스레드로 작업을 실행 기능이 존재함.

-glob: 파일 경로와 패턴 매칭을 통해 파일 목록을 찾는 데 사용됨.

-pickle: Python 객체를 바이너리로 저장하고 로드하는 모듈로, 임의의 Python 객체를 파일에 직렬화하거나 메모리에 저장할 수 있음.

-random: 난수 생성 모듈로, 무작위 값을 생성하고 특정 범위 내의 값을 샘플링하는 데 사용됨

-torch: PyTorch의 핵심 데이터 구조로, 딥러닝 모델이 학습할 수 있는 다차원 배열 형태의 데이터를 제공함.

-torch.nn: PyTorch의 신경망 모듈로, 딥러닝 레이어와 활성화 함수 등의 요소를 제공함. 모델 구조를 정의하는 데 필요함.

-numpy: 과학 연산을 위한 Python 라이브러리로, 다차원 배열 및 행렬 연산을 수행할 수 있는 강력한 기능을 제공함

-pandas: 데이터 분석을 위한 라이브러리로, 데이터 필터링, 변환, 집계 기능이 존재함.

-sklearn: 머신러닝, 데이터 분석을 위한 라이브러리로, 전처리, 모델 학습, 예측과 평가가 가능하여 범용성이 높음.

-lightgbm: 그라디언트 부스팅 기반의 머신러닝 모델인 LightGBM을 사용할 수 있고 regression,classification task 수행이 가능함.

-hurst: 데이터의 Hurst 지수를 계산해 시계열의 자기 유사성과 장기적 의존성을 분석하는 데 사용함.

-matplotlib.pyplot: 시각화 라이브러리로 다양한 종류의 그래프를 그릴 때 사용함.

-statsmodels.graphics.tsaplots: ACF(자기상관 함수) 그래프를 생성해 시계열 데이터의 시간에 따른 상관성 시각화에 사용함.

-hmmlearn: Hidden Markov Model(HMM)을 시계열 데이터에 대한 예측과 분류를 수행이 가능함.

* 활용 데이터

-수출입: 2018년 1월부터 2024년 9월까지 품목별 거래 대상국, 거래 금액, 거래 중량 데이터를 농넷(nongnet.or.kr)을 통해 수집

- WTI유 선물: 2018년 1월부터 2024년 10월까지의 종가, 시가, 고가, 저가, 거래량, 변동 정도 데이터를 investing.com을 통해 수집

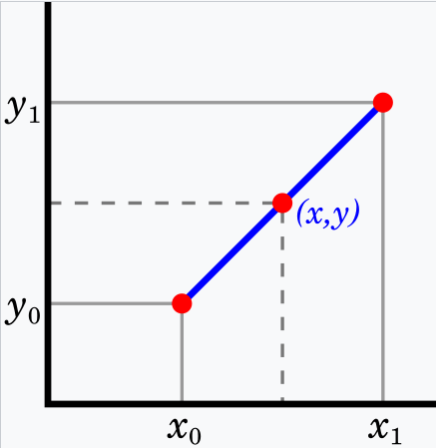
- 가락시장 반입물량:2018년부터 2024년 9월까지의 각 품목별 총 반입량과 전년, 평년 총 반입량을 수집

- 기상: 기상청 기상자료 개방 포털을 통해 강원, 서울, 전라, 경상, 충청, 제주 등 각 지역별 기온, 풍속, 습도, 일조일사, 강수량 데이터를 수집

- 가락 도매 시장 확정거래물량: 농업관측 통계 시스템(oasis.krei.re.kr)을 통해 각 품목별 금액, 전년 반입량, 전년 금액, 평년 금액, 평년 반입량을 수집

2.데이터 처리

* 결측 데이터 처리
* 제공받은 데이터의 결측치 혹은 0 값들이 어느정도 존재했다. 이를 선형 보간을 이용해서 처리했다. 선형 보간법은  끝점의 값이 주어졌을 때 그 사이에 위치한 값을 추정하기 위하여 직선 거리에 따라 선형적으로 계산하는 방법이다.



* 데이터 인코딩 처리
* **Mean Encoding**: 수확 시기별 평균 가격 차이를 반영해 시간적 가격 변동을 모델에 반영한다.
* **Standard Deviation Encoding**: 수확 시기별 가격 변동성을 반영해 타겟값의 분산 차이를 정규화하고, 가격 변동 패턴을 모델에 반영
* 데이터 병합, 형태 변환
* **Transpose**: 시계열 데이터를 Transpose하여 N 시점에 N-8시점까지의 과거 데이터가 피처로 들어가도록 변환하여 머신러닝 모델이 **과거 시점의 데이터**도 학습할 수 있도록 만든다. 예를 들어 배추의 2023년 05월 상순, 중순, 하순의 가격을 예측하는 경우, 배추의 2023년 2월의 상, 중, 하순, 2023년 3월의 상, 중, 하순, 2023년 4월의 상, 중, 하순의 배추 가격을 모델의 피쳐로 사용하는 것이다.
* **동일 품목 다른 품종 Merge**: 동일 품목명에 대한 **다른 품종명**을 갖는 데이터들을 N시점의 데이터에 모두 Merge를 한다. 이때 이전의 Transpose 방식도 활용하며 머신러닝 모델이 다변량 시계열 데이터를 학습하는 효과를 만든다. 예를 들어 배추의 2023년 05월 상순, 중순, 하순의 가격을 예측하는 경우, 배추의 2023년 2월의 상, 중, 하순, 2023년 3월의 상, 중, 하순, 2023년 4월의 상, 중, 하순의 배추 가격 뿐만 아니라 배추와 다른 품종인 얼갈이 배추와 봄동 배추의 2023년 2월의 상, 중, 하순, 2023년 3월의 상, 중, 하순, 2023년 4월의 상, 중, 하순 가격을 추가 피쳐로 사용하는 것이다.

3.데이터 분석

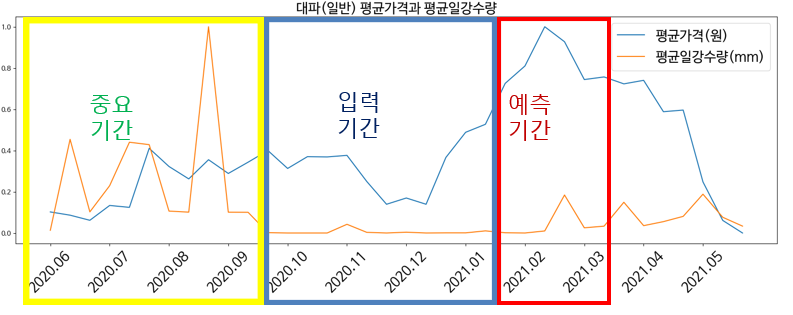
1. 자연적 요인

- 농산물 가격에는 자연적 요인들이 크게 작용한다. 주로 강수량, 온도, 습도, 일조량 등의 요소들이 농작물의 생산량과 품질을 결정하는데, 이들은 곧 농산물의 가격에도 중요한 영향을 미친다. 예를 들어, 농작물의 생육 기간동안 강수량이 평년보다 부족하면 생산량이 줄어들어 공급이 감소하고 가격이 상승할 수 있다. 반대로 강수량이 지나치게 많아도 높은 습도와 병충해로 인해 생산량이 줄어드는 경우도 발생한다. 이런 상황 역시 농산물의 공급 측면에 영향을 미치고, 결국 가격 상승의 원인이 된다.

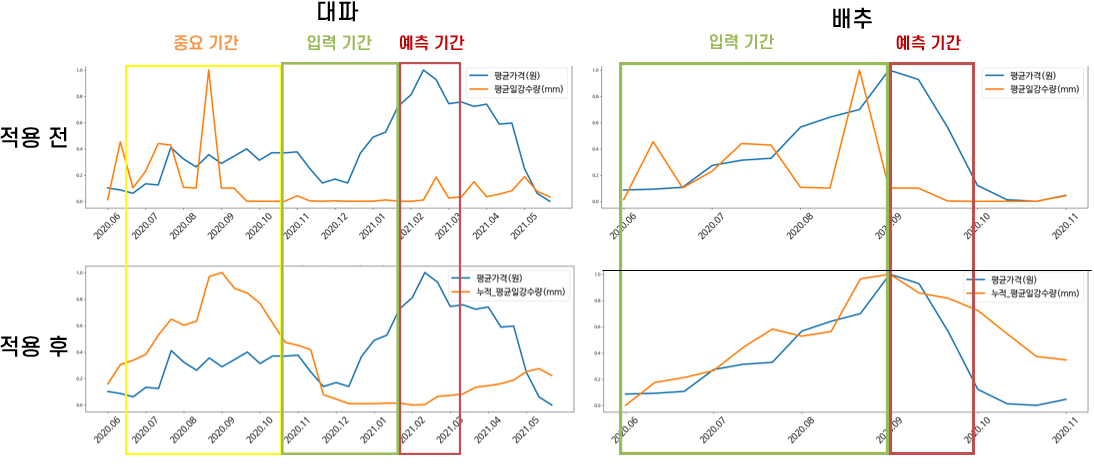
제공받은 데이터의 기간 중에서 위의 사례처럼 자연적 요인으로 인해 가격의 변동이 발생한 사례를 찾아보았고 **2021년 2월 대파**의 가격이 급등한 사례를 발견했다. 이는 **2020년 9월 7일** 발생한 **태풍**에 기인한 것이었다. 이처럼 특정 기간의 농산물 가격 예측을 진행하기 위해서는 해당 기간의 자연적 요인보다 과거 농산물의 성장 시기의 자연적 요인이 필요하다.



하지만, 이번 대회 task 특성상 예측 시점 기준 **최대 3개월 이전**의 데이터만 사용해야 한다는 제약이 존재한다. 게다가 농산물 생산량에 영향을 미치는 자연적 요인은 씨앗의 **파종부터 재배**까지의 기간을 모두 고려해야 한다. 결국 위의 2021년 2월 급등한 대파 가격 예측을 위해선 3개월보다 더 이전의 자연적 요인을 모델이 반영해야 하는 것이다.

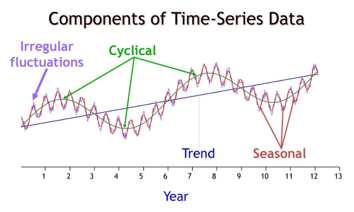


일별 데이터의 단점인 높은 변동성과 과거의 극단 이벤트(태풍)의 미포착 문제를, 누적 데이터를 통해 노이즈를 줄이고 과거의 극단적 기상 상황의 장기적 영향을 반영함으로써 극복했습니다.



평균 일 강수량에 대해 누적 값을 적용하여 장기적 영향을 반영

1. 시계열 요인



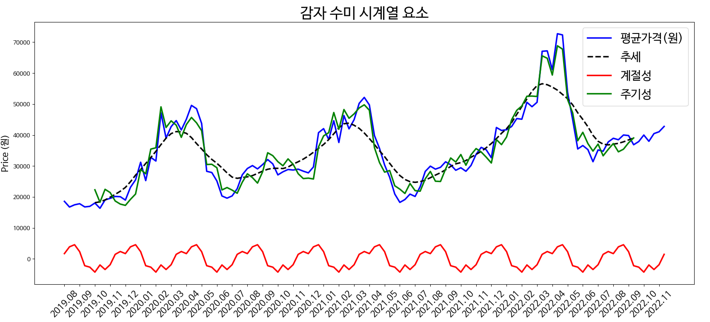
시계열 데이터의 4가지 구성요소

시계열 데이터가 갖는 고유한 특성들인 **트렌드, 계절성, 주기성, 그리고 불규칙 변동** 등의 요인들을 파악하는 것이 중요하다. 왜냐하면 이런 특성들을 파악해서 주어진 시계열 데이터에 대한 깊은 이해가 가능하고 이를 적용한 모델링 과정을 추후에 거쳐야 하기 때문이다.

**트렌드**는 시간에 따른 데이터의 장기적인 상승 또는 하락을 나타내며, 특정 방향으로의 경향성을 설명한다. 이동 평균(moving average)은 일반적으로 이 트렌드를 표현하기 위해 많이 사용되지만, 단순한 **평활화** 방식으로는 노이즈 제거에 한계가 있다. 게다가 이동 평균 방식은 지연성이 강해 가격 변화에 대한 **반응이 느리며**, 노이즈 제거에 한계가 존재한다. 이를 보완하기 위해, 이번 프로젝트에서는 **저역 통과 필터**(Low-Pass Filter)를 활용하여 고주파 노이즈를 제거하고 데이터 변화에 빠르게 반응할 수 있는 방안을 사용했다. 저역 통과 필터를 사용하면 노이즈를 효과적으로 제거할 뿐만 아니라, 트렌드를 더 정확하게 반영할 수 있어 시계열 데이터의 **장기적인 변화**에 민감하게 반응할 수 있게 된다.

**계절성**은 주기적으로 반복되는 패턴을 의미하며, 농산물 가격 예측에서 매우 중요한 요인 중 하나다. 계절성 패턴은 주로 기후의 변화, 품목별 수확 주기 등과 같은 요인이 작용하여 농산물 가격도 계절적으로 변동하게 만든다. 결국 각각 농산물의 **고유 계절성**을 파악하여 모델 학습에 반영하면 가격 예측의 정확도를 높일 수 있다. 이를 계량화하기 위해 sin, cos 변환을 사용하여 주기적 변화를 모델에 반영하고, harvest\_weight(재배기)와 같은 이벤트를 나타내는 추가 변수를 설계했다. 또한, std target encoding을 적용해 변동성을 표준화하여 모델이 계절 변동을 더 명확하게 학습할 수 있도록 했다.

**주기성**은 계절성과 비슷하지만, 경제적 주기나 농산물의 수급 변화와 같은, 비교적 길고 **복잡한 주기적 변동**을 설명한다. 주기성 변동을 반영하기 위해 E\_M\_T(**가격 변화 추이**를 반영하는 Cycle변수)와 같은 파생 변수를 생성하고, 시간에 따라 변화하는 차이를 반영하는 create\_difference\_features(과거와 현재 **가격 차이**를 통해 **단기 변동**과 **중기 추세**를 포착하는 변수)를 추가하여 주기적 변동을 모델이 학습할 수 있도록 계획했다.



감자 수미 품목의 가격 데이터에 대한 추세, 계절성, 주기성 요소들을 추출한 그래프

1. 경제적 요인

앞서 살펴본 요인들에 더해 농산물 가격은 **선 후행 관계**에 의해 나타나는 경제적 요인에 의한 영향도 존재할 것이라 생각했다. 이런 경제적 요인을 파악하기 위해 농산물 가격과 선행, 후행 관계에 있는 변수들을 파악하는 것이 중요하다. 그래서 우선 경제적 요인으로 작용할 수 있는 변수들을 선별했고 이 변수들과 가격 간의 선 후행 관계 파악을 위해 Granger 인과 테스트, 충격 반응 함수 테스트를 진행했다.

\* **Granger 인과 테스트**: 한 시계열 x의 과거 정보가 다른 시계열 Y의 미래 값을 예측하는 데 통계적으로 유의미한 도움을 준다면, 이를 Granger casual 하다고 말한다. 아래는 Granger 인과성을 설명하는 수식이다. P(Y(t+1)∣Y(t),Y(t−1),...,X(t),X(t−1),...)은 식의 왼쪽 항에 해당하며 이는 미래의 Y 값 (즉, Y(t+1)Y(t+1)Y(t+1))이 과거의 Y 값들 뿐만 아니라 과거의 X 값 들에도 조건부 의존성을 가진다고 가정한 확률이다. 즉 Y의 미래 값을 예측할 때 과거 X의 정보를 고려하는 경우이다. P(Y(t+1)∣Y(t),Y(t−1),...)은 식의 우측 항으로 이는 미래의 Y 값이 과거의 Y 값 들만을 고려하여 조건부 의존성을 가진다고 가정한 확률이다.

즉, Y의 미래 값을 예측할 때, 과거의 Y 정보만을 사용하고, X 정보는 사용하지 않는 경우이다. 만약 과거 X의 값이 Y의 미래 예측에 도움이 된다면 두 조건부 확률이 같지 않을 것이라는 가정을 표현한 식이다. 해당 식이 성립하는 경우 X의 과거 정보가 Y의 미래 값을 예측하는데 유의미한 영향을 미칠 것이고 만약 해당 식이 성립하지 않으면 과거 X의 값이 미래 Y값의 예측에 영향이 없는 것을 의미한다.

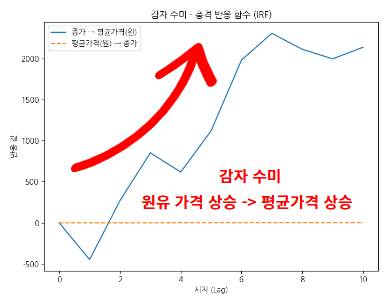
폰트, 텍스트, 타이포그래피, 그래픽이(가) 표시된 사진

자동 생성된 설명

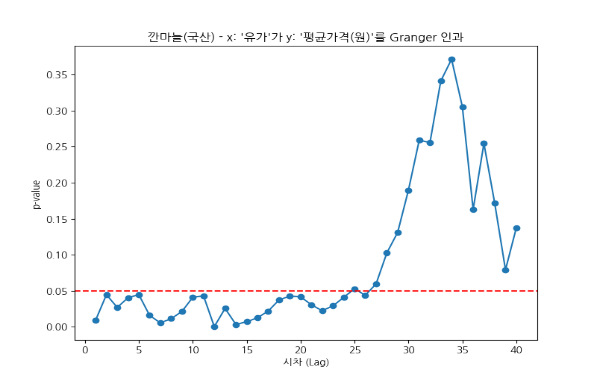
\* **충격 반응 함수**: 한 변수에 충격이 발생했을 때, 다른 변수가 시간의 흐름에 따라 어떻게 반응하는 지 보여주는 함수이다. 시계열 X의 충격으로 시계열 Y의 충격 반응 결과가 양수이면 Y값의 상승을, 시계열 Y의 반응 결과가 음수이면 Y값의 하락을 의미한다. 아래 그래프 예시를 보게 되면 시계열 X(원달러)의 상승(충격)으로 시계열 Y(수출금액/수출물량)의 충격 반응 결과가 양수가 된다. 즉, 이 그래프를 통해 ‘원달러 상승 시 수출 금액이 증가한다’ , ‘원달러는 수출 금액에 영향을 미치는 선행 변수이다’라는 사실을 유추 가능하다.

텍스트, 라인, 그래프, 도표이(가) 표시된 사진

자동 생성된 설명



텍스트, 그래프, 라인, 스크린샷이(가) 표시된 사진

자동 생성된 설명

위 그래프는 각각 깐마늘의 평균가격과 WTI 원유 선물가격(종가)의 충격 반응 함수 그래프, Granger 인과 그래프이다. 원유 선물가격의 충격으로 인해 깐마늘 평균가격의 충격 반응 결과가 양수이고 Granger 인과의 유의확률이 임계값인 0.05 보다 작아지니 **원유 선물가격**이 깐마늘의 평균가격에 **선행하면서** 유의미한 영향을 미친다는 것을 확인 가능하다.

도표, 그래프, 라인, 텍스트이(가) 표시된 사진

자동 생성된 설명텍스트, 라인, 스크린샷, 그래프이(가) 표시된 사진

자동 생성된 설명

위 그래프는 각각 대파의 평균가격과 대파 **총반입량**의 충격 반응 함수와 Granger 인과 그래프이다. 대파의 총 반입량이 평균가격에 **선행 요인**로 작용하여 총 반입량의 증가가 평균가격 하락의 원인이 됨을 확인 가능하다.

텍스트, 그래프, 도표, 라인이(가) 표시된 사진

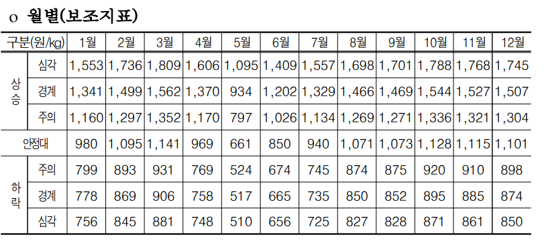
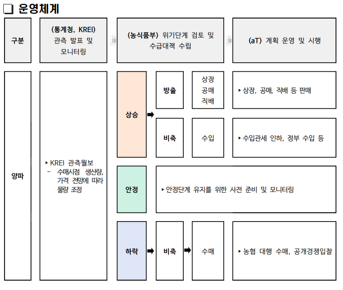
자동 생성된 설명 텍스트, 라인, 그래프, 도표이(가) 표시된 사진

자동 생성된 설명

-대파의 경우와 비슷하게 사과 총 반입량이 사과 평균가격의 선행 요인으로 작용하여 총 반입량 증가가 평균가격 하락의 원인이 됨을 확인 가능하다.

1. 외부적 요인

농산물 가격에는 외부적 요인도 중요한 영향을 미친다. 외부적 요인들은 인위적 개입으로, 주로 **방출, 수매**, 보조금 지급, 수출입 제한 조치, 그리고 가격 지원 정책 등이 포함된다. 이들은 모두 정부나 관련 기관의 정책적 결정으로 주로 농산물 **가격의 안정화**라는 목적을 위해 시행되는 조치들이다. 예를 들어, 농산물의 가격이 급격히 오르거나 내릴 때 **정부**가 물량을 방출하여 수급을 조절하는 것이다. 하지만 현재 농산물 수급관리 운영 체계에는 한계점이 존재한다. 왜냐하면, 각 위기단계별 기준 가격 설정을 **연 1회만** 개정하여 연내에 발생하는 비정상적 가격 변화에 신속한 **대응이 어렵다**. 또한, 기준 가격 설정 시 월별 평년 가격에 고정된 작형별 등락률을 곱해서 정하기 때문에 가격 변동 패턴을 **획일적으로만** 반영한다.



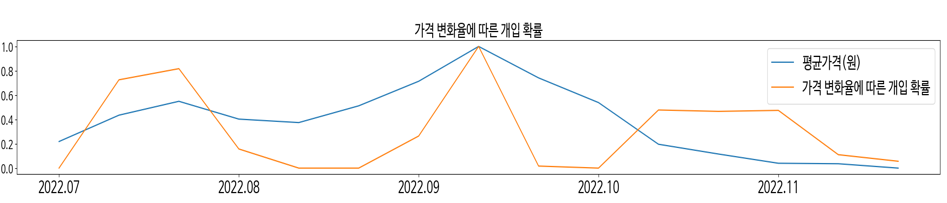
농산물 수급관리 가이드라인(운영체계)과 연 1회 개정되는 위기 단계별 기준가격

데이터 분석하며 외부적 요인이 실제로 농산물 가격에 영향을 미치는 사례를 확인할 수 있었다. 2022년 9월 중순 폭염과 태풍으로 인한 **출하의 어려움으로** 배추 가격이 급등한 시점에 **정부의 방출** 조치가 이루어졌다. 해당 시기에 이러한 방출 조치는 배추 가격을 일정 수준까지 내리는 데 어느정도 기여했음을 확인할 수 있었다.



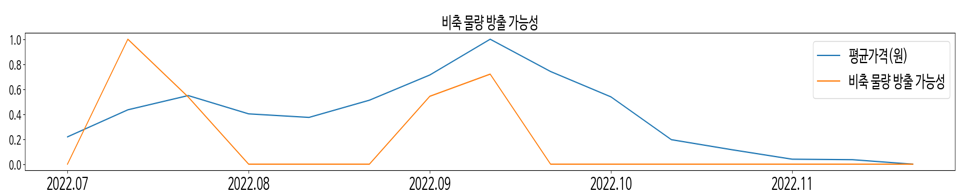
배추 물량의 방출 정책 수행 이후로 안정화된 배추 평균가격 그래프

하지만 외부적 개입은 사람의 의사결정에 따른 **비정기적**, **비 정형적인** 요인으로, 명시적 데이터나 지표가 부족해 기계 학습 모델이 이를 정확히 학습하기 어렵다. 정부의 방출이나 수매 같은 조치들은 특정 조건이 맞춰졌을 때에만 이루어지기 때문에 모델이 이를 예측하기에는 **한계가 존재**한다.

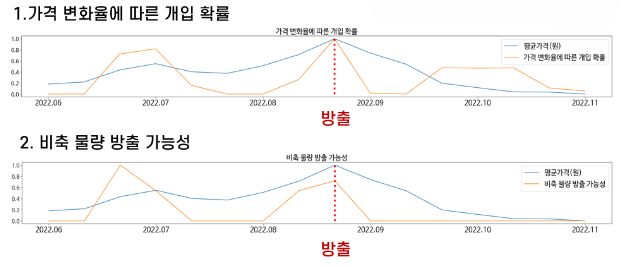
따라서 이번 예측 모델에서는 외부적 요인을 직접 반영하기보다는 "**개입 확률**"을 활용하여 모델에 반영하는 방식을 고려했다. 가격 변동률이 특정 **임계값을 초과**하면 개입 가능성이 높아진다는 점을 이용해, 개입이 이루어질 확률이 높은 구간에서는 이러한 변동을 안정시키는 방식으로 모델에 반영하도록 했다. 

배추 평균가격과 가격 변화율에 따른 개입 확률 그래프

또한, 농산물 **현재 가격과 이동 평균의 차이**가 특정 임계값을 초과하면 그 차이에 비례하여 비축 물량 방출 확률을 반환하여 외부적 요인이 작용할 확률을 모델이 학습하도록 했다.



배추 평균가격과 비축 물량 방출 가능성 그래프



배추 물량 방출이 일어난 시기에 개입확률, 방출 가능성이 높아진 것을 확인 가능하다.

4.데이터 활용

* 특성공학(Feature Engineering)\_ML

앞서 진행한 데이터 분석을 통한 얻은 인사이트를 특성공학 시 충분히 활용하여 농산물 도메인에 알맞은 모델을 만들고자 노력했다.

* 가격 수치 피쳐

1. E\_M\_T: 3개월 전과 현재 시점의 차이 피처로 **가격 변화 추이**를 반영하는 피처이다.
2. 평균\_평년\_diff: 현재 가격이 평년 평균 가격에 비해 얼마나 차이가 있는지 나타내어 **장기적 추세 변화**를 반영하는 피처이다.
3. create\_difference\_features: 각 시점에서 과거와 가격 차이를 통해 **단기 변동과 중기 추세**를 포착한다.
4. cumulative\_increase: 과거 3개월 동안의 가격 누적 변동량 피처이다.

* 시간 정보 피쳐

1. process\_YYYYMMSOON: 날짜 정보를 연도, 월, 순으로 분리하고 각각을 **sin, cos 변환**하여 주기성을 반영한다.
2. determine\_season: 계절 정보를 계산해 주기적 특성을 나타내는 피처이다.
3. harvest\_weight: 품목별 수확 가능한 월에 따라 0 또는 1의 이진 변수를 사용한 피쳐로써 농산물 **계절적 공급** 변화에 따른 가격 변동을 잘 포착한다.
4. determine\_harvest\_weight기반 1\_encoding, 1std\_encoding: 평균이나 표준 편차를 인코딩하여, 데이터의 품목별 수확 가능한 시즌의 성질을 피처로 생성한다.

* 스무딩

1. LPF: 저역통과 필터를 사용해 고주파 노이즈를 제거하여 시계열 데이터의 주요 **트렌드**를 강조한다.

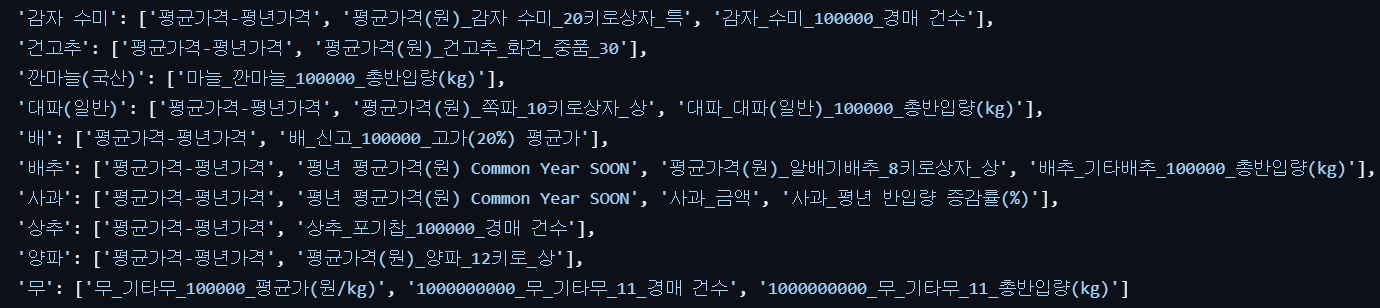
* 방출, 수매

1. add\_intervention\_features(가격 변화율에 따른 개입 확률, 비축 물량 방출 가능성, 변동성 기반의 개입 가능성): 농산물 가격 변화율, 비축 물량 방출 가능성, 변동성 기반 개입 확률을 피처로 추가한다. (**가격의 급격한 변화**나 변동성에 따라 **개입 필요성**을 예측하며, 정책적 개입이 필요한 상황을 반영하여 가격 변동의 영향력을 모델에 반영한다.)
2. calculate\_intervention\_probability(가격 변화율에 따른 개입 확률): 주어진 가격 리스트에서 변화율을 기반으로 개입 확률을 계산한다. (가격이 일정 변화율 이상으로 급등 또는 급락할 경우 개입 확률을 높게 설정하여 **급격한 가격 변동에 대비할** 수 있도록 한다.)
3. release\_probability(비축 물량 방출 가능성): 현재 가격이 이동 평균과 비교해 일정 비율 이상 상승했을 때 **비축 물량 방출 가능성**을 계산한다. (가격 상승이 예상치보다 크면 비축 물량 방출 가능성을 높여 가격 안정화를 유도하는 데 사용한다.)
4. volatility\_intervention(변동성 기반 개입 가능성): 가격의 변동성을 기반으로 **개입 가능성**을 계산하여 변동성이 높은 경우 개입 확률을 높인다. (가격의 변동성이 특정 수준 이상일 때, 개입 가능성을 고려하여 급격한 변동을 완화하는 전략적 판단을 모델에 반영한다)
5. add\_climate\_intervention\_features: 기후 관련 피처에서 특정 값이 평균으로부터 일정 편차를 초과할 경우 **개입 확률**을 계산한다. (기후 변화가 농산물의 수급에 미치는 영향을 반영하여, 기후 이상 현상이 있을 때의 개입 가능성을 예측한다.)
6. calculate\_cumulative\_features\_fixed: **기후 요소의 누적합**을 계산하여 누적된 기후 데이터를 피처로 추가한다. (농산물은 씨앗부터 다 성장될 때까지 3개월 이상이 소요되는데 3개월이라는 한계 때문에 우리는 합계 피처를 사용했고 사람처럼 먹어서 중량이 늘어난다는 개념을 데이터에 추가했다.)

* 특성공학(Feature Engineering)\_DL

DL 모델에서 평균가격과 평년 평균가격의 차이, 시장 총 반입량, 동일 품목 다른 품종의 평균가격 데이터를 다변량 함수들로 추가하여 평균가격 예측을 진행했다.

* 평균가격- 평년 평균가격: 특정 시점에서의 가격이 평년 대비 높은 지 낮은지를 파악하여, 현재 시점에서의 시장 상황을 반영한다. 이 정보는 비정상적인 가격 변동이나 **예외적인 시장 상황**을 반영하는 데 유용하다.
* 시장 총 반입량: 특정 시점에서의 시장 총 반입량은 시장에 유입되는 해당 품목의 **공급량**을 나타내며, 이는 수요와 공급의 균형에 중요한 역할을 한다. 농산물 공급이 많을수록 가격은 하락하고, 공급이 적을수록 가격은 상승하는 경향이 있기 때문에 시장 반입량을 피처로 사용하면 모델이 **공급 변화**에 따른 **가격 변동**을 더 잘 예측할 수 있다.
* 동일 품목 다른 품종 평균가격: 동일 품목 내에서도 품종에 따라 가격대가 다르게 형성된다. 또한, 특정 품종의 가격이 상승은 같은 품목의 다른 품종 수요를 증가시켜 가격의 변동을 유발한다. 그래서 같은 품목 다른 품종의 가격을 피쳐로 추가하면 모델이 농산물의 **수요량 변화와 분산**을 이해하는 데 도움을 줄 수 있다.



DL 모델에서 사용한 품목별 피쳐들

5.모델링

* 모델 알고리즘
* 품목별 Hurst 지수, 장기 의존성의 존재 여부를 통해서 적절한 예측 모델을 선택했다. 아래는 품목별로 가격 예측에 사용한 모델들과 이유이다.

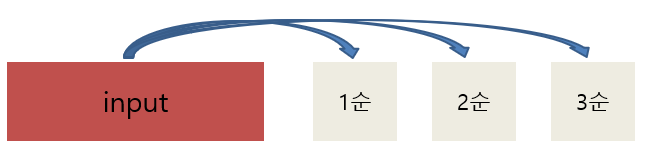
1. 양파, 상추, 감자 수미, 무, 사과, 대파, 배추: Hurst 지수가 0.3~0.4 정도로 0.5에 비해 낮아 가격 장기 의존성이 매우 약하고 **반 추세적** 특성과 **역 상관성**을 띄는 품목들이다. **비선형 패턴**과 단기 변동성을 잘 학습하는 ML을 사용했다.
2. 건고추, 깐마늘(국산), 배: Hurst 지수가 0.7정도로 형성되며 0.5에 비해 높아 데이터에 **장기 의존성**이 존재하는 품목들이다. 과거 패턴이 미래에 지속될 가능성 높기 때문에 순차적 시간 의존성을 보존하고 장기 패턴 학습에 뛰어난 DL을 사용했다.

* 모델 구조 (사용한 모델)

폰트, 그래픽, 로고, 텍스트이(가) 표시된 사진

자동 생성된 설명

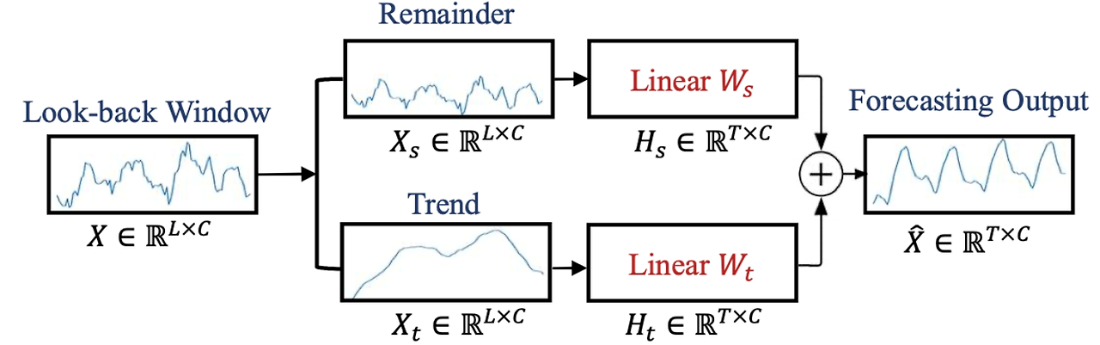
* ML(LightGBM): LightGBM은 결정 트리 기반의 그래디언트 부스팅 알고리즘으로, 균형 성장 방식보다 비선형 관계를 더 잘 포착하는 **leaf-wise** 성장 방식을 사용한다. 또한 데이터의 빈도 분포를 빠르게 계산하는 **Histogram-based** 방식으로 다른 모델에 비해 정확도와 추론 시간에서 이점을 발휘할 수 있기에 해당 모델을 채택했다.



ML 모델의 예측 구조

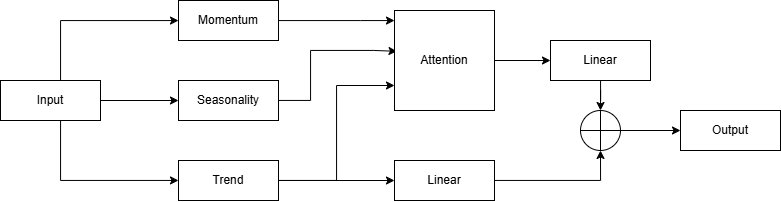
* TimeSeries(Dlinear): 시계열 모델로 선형 구조 기반의 Dlinear 모델을 선택했다. 해당 모델을 택한 이유는 이전에 살펴본 EDA 과정을 거치며 농산물 가격의 **트렌드와 계절성**의 중요성을 파악했기 때문이다.

특히 D(Decomposition)Linear 모델은 시계열 데이터를 **추세와 계절성** 성분으로 분해하여 단기적, 장기적 변화를 동시에 학습할 수 있어 시계열 모델로 택하게 되었다.



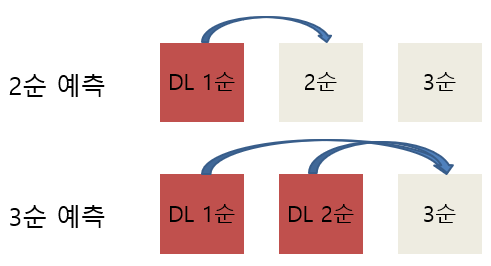
Dlinear 모델의 구조 : 데이터를 Trend, Remainder(seasonality)로 분해하여 예측을 진행한다.

* TimeSeries(Dlinear with Attention): 기존의 Dlinear 모델에서 시계열 데이터의 트렌드 요소, 계절성 요소만 추출하는 것이 아닌 **모멘트 요소를** 추가로 추출하여 각각을 어텐션 메커니즘에 query, key, value로 사용한다.



Dlinear with Attention 모델 구조

* TIME ML: 기존 ML 모델의 2순, 3순 예측을 진행 시 DL의 1순,2순 예측 결과를 추가 피처로 사용하여 더 높은 예측 정확도를 얻는다.



TIME ML 모델 구조

6.모델 학습

* 모델 학습 방식

여러가지 하이퍼 파라미터 조합을 이용해 예측을 진행하고 연도별 kfold 방식을 활용했다. 대회의 평가 산식인 NMAE loss를 구현 후 모델 학습 시에 활용하고 2순과 3순에 대한 loss에 가중치를 주어 비교적 예측이 어려운 2,3순의 예측 성능을 향상시켰다.

이 과정에서 각 품목마다 2018년부터 주어진 데이터 전부를 학습에 활용하는 것이 아닌 ACF를 통해서 품목들의 모델 학습에 사용될 데이터 시작 연도를 정한다.

* + ACF(AutoCorrelation Function): 한 시계열 데이터가 얼마나 이전 시점과 유사한 지 나타내는 지표로 현재 시점과 특정 시차만큼 떨어진 과거 시점 간의 유사도를 통해 구한다. 해당 방법을 이용하여 품목별 주기성 패턴, 장기 자기상관성의 유무를 확인하여 모델 학습 시작 연도를 정했다.

텍스트, 폰트, 라인, 도표이(가) 표시된 사진

자동 생성된 설명텍스트, 스크린샷, 도표, 라인이(가) 표시된 사진

자동 생성된 설명

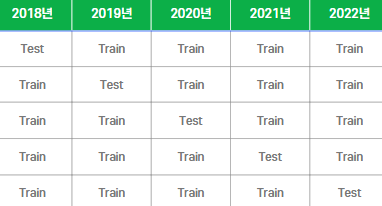
텍스트, 스크린샷, 도표, 그래프이(가) 표시된 사진

자동 생성된 설명 텍스트, 스크린샷, 도표, 그래프이(가) 표시된 사진

자동 생성된 설명

감자처럼 주기적 패턴이 잘 나타나는 품목의 경우 전체 데이터에 대한 학습으로 장기적 트렌드를 모델이 학습하게 했다. 대파처럼 초기에 변동성이 크고 자기 상관성이 빠르게 감소하는 품목의 경우는 시간이 지날 수록 장기 자기상관성이 낮아 2020년 혹은 2021년 이후의 데이터를 활용했다.

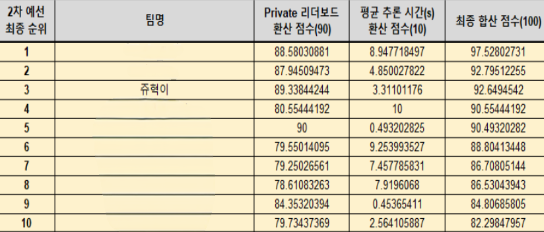
* 검증 방식
* **연도별 KFold 교차검증**: 주어진 훈련 데이터의 기간은 2018년에서 2022년까지 총 5년이다. 년도별 KFold 검증으로 특정 시점 혹은 연도에 대한 과적합 예방하고 이상치에 대한 모델의 민감도를 낮춘다.



* **월별 KFold 교차검증**: 농산물의 **계절성**과 **수확 주기**를 고려해 모델을 평가할 수 있는 점에서 장점이 있다. 또한, 농산물 가격은 계절에 따라 변동폭이 크기 때문에 월별 변동성을 모델이 잘 학습하고 예측하는 지 평가가 가능하다. 연도별 KFold와 유사하게 모델이 특정 시기에 대한 과적합을 방지하고 일반화 성능 향상을 이뤄낼 수 있다.
* **수확 시기별 KFold 교차검증:** 이전에 수행한 데이터 분석을 통해 농산물 수확 시기는 농산물 가격에 큰 영향을 미침을 알 수 있었다. 이에 각 품목별 수확시기별 폴드를 사용해 각 농산물의 재배 기간과 수확 시기의 특성을 더 명확히 반영 가능하다. 또한, 수확 시기별 데이터 분할을 통해 모델이 수확 시기와 그 이후 가격 변화를 잘 학습하게 된다.
* **같은 품목 다른 품종 가격 Validation:** 같은 품목 농산물이라도 품종에 따라 각기 다른 가격 패턴을 갖는다. 왜냐하면 각 품종마다 생육 특성, 소비자의 선호 정도 등이 상이하기 때문이다. 그래서 이와 같은 validation을 진행하면 모델이 **다양한 환경과 특성**을 갖는 농산물에 대해 **강건한 예측** 성능을 갖는지 평가가 가능하며 **모델 범용성** 또한 높일 수 있다.
* **급등, 급락 구간에 오차 가중치를 부여하는 Validation:** 가격이 급변하는 구간에 대한 모델 민감도를 향상시킬 수 있는 방법이라 생각하여 사용해 보았다. 하지만, 이를 모델 학습에 적용 시 모델 안정성 측면에서 과적합 위험이 증가하고 일반화 성능의 저하 현상이 나타나 최종적으로 사용하진 않았다.
* **최종 검증 방법:** 모델의 예측 정확도가 물론 중요하지만 추론 시간 또한 채점 방식에 반영되기 때문에 연도별 KFold 검증 방식을 선택했다.

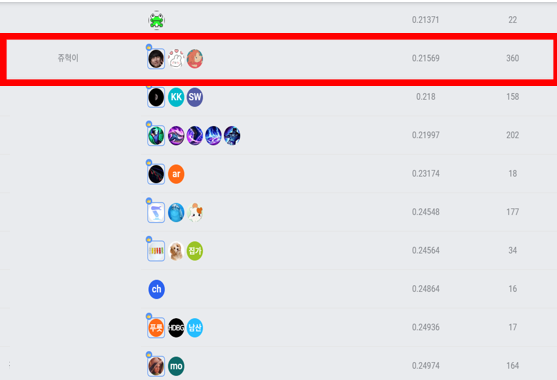
7.결과 및 결언

* 모델링 성능 결과



모델 예측 정확도(리더보드)와 추론 시간을 모두 반영한 점수로 최종 3등을 기록

* 현재 추론 코드에 **print 구문**이 포함되어 있어, 추론 시간 고려 점수 측정에서 다소 **디메리트를** 보였다. 실제로 print구문을 제거한다면, 추론 시간 점수가 올라 0.15점 차이가 나는 2등팀을 제치고 **최종 2등**을 기록할 것이다.



추론 시간을 제외하고 모델 예측 정확도만 고려했을 경우 근소한 차이로 2등을 기록

* 모델링 개선 방향

2차 예선이 진행되는 12일간 총 **360번의 제출**로 제출 수 2등인 팀의 202회보다 1.8배 높을 정도로 본선 진출 팀 중에서 가장 많은 실험을 했고 이 과정에서 여러 인사이트들을 얻었다. 이를 충분히 이용하면 다양한 모델링 개선 방안들을 도출할 수 있다.

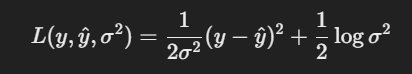
먼저 시계열 모델을 통한 예측 시 농산물 일별 가격 데이터를 이용하여 10일 평균가격이 아닌 **5일 평균** 가격 예측을 수행하는 것이다. 이를 통해 가격 변화 주기를 더 명확히 모델이 학습 가능하다. 특히 어텐션 메커니즘이 추가된 모델의 경우, 더 짧은 주기 단위 데이터가 이전보다 더 많이 제공되면 어텐션 메커니즘이 모멘텀 요소, 계절성 요소에 대한 학습 능력 향상이 예상된다. 또한, 더 과거의 시기의 데이터를 merge 하는 방법을 사용시 모델의 성능이 어느정도 향상시킬 것이라 예상했지만 merge의 수행 과정에서 시간이 오래 걸려 사용하진 않았다. 대회 기간 당시에는 사용하지 못 했지만 농산물의 수출, 수입 데이터와 연관지어 **환율 데이터**를 모델 학습에 사용하면 어느 정도 성능 향상을 기대해볼 수 있을 것이다.

현재 시계열 모델의 경우 예측 품목마다 개별적으로 최적화된 모델을 사용한다. 하지만 이는 모델 학습과 추론 시간이 길다는 단점이 있다. 시계열 모델 이용 시 같은 주기적 특성이나 계절성 패턴을 갖는 품목에 대해서 **many-to-many** 방식으로 예측을 진행하면 서로의 패턴을 보완하며 모델 학습 성능을 향상시키고 학습, 추론 시간의 단축이라는 이점도 얻을 수 있다. 해당 예시로 무와 배추가 있다. 무와 배추의 경우 외부적 요인으로 정부의 방출 수매가 적극적으로 이루어지고 해당 품목들의 가격이 이에 많은 영향을 받는다는 점, 4기작을 통해서 계절마다 생산이 이루어지는 점, 김장의 주 재료가 되면서 자연적 요인이 생산에 많은 영향을 미친다는 점 등 여러 공통점들을 공유한다.

예선을 진행하며 가격이 계절성 패턴을 띄는 배추나 무와 같은 품목은 가격 변동성이 시기에 따라 달라진다는 것을 확인했다. 배추를 예시로 들면 연초인 1,2,3월에서 배추의 가격 변동성은 그리 크지 않지만 9,10월에 이르러서는 김장 물량 확보를 위한 수요 증가와 공급 불안정성이란 요인으로 인해 가격 변동성이 크게 증가한다. 다시 말해 배추의 가격은 **시기**에 따른 **가격의 분산**이 달라지는 **이분산성**이 강하다.

하지만 지금까지 사용했던 NMAE loss는 모든 데이터 포인트의 오차를

동일하게 취급하여 데이터의 분산 변화를 반영하지 못한다. 그래서 데이터 이분산성 고려한 loss인 **Heteroscedastic Loss**를 이용해서 모델이 예측 값과 예측 불확실성인 분산도 학습하도록 한다.



Heteroscedastic Loss의 수식으로 실제 값과 예측 값의 차를 데이터 포인트별 분산으로 나눠준다.

Heteroscedastic Loss를 도입하면 얻을 수 있는 이점으로 시계열 데이터의 비선형적 복잡한 패턴을 더 잘 학습할 수 있다. 변동성이 큰 구간에서 데이터 포인트별 분산이 커지기 때문에 loss를 감소시켜서 모델이 유연한 학습이 진행하도록 유도한다. 이는 농산물 가격 예측 모델이 이분산성, 비선형성 등 다양한 특징을 가진 가격 데이터에 대해 더 효과적인 모델 학습을 유도할 수 있다.

TimeSeries 모델로 Dlinear 모델뿐만 아니라 Nlinear와 Linear모델, 시계열 데이터에 대한 SOTA급 성능을 갖는 PatchTST, SegRNN, iTransformer 와 같은 다양한 모델을 이용한 실험을 진행했었다. 같은 선형 기반 모델인 Nlinear, Linear 보다는 Dlinear 모델이 성능이 우수했다. 또한, 깐마늘(국산)과 배에 대해서는 iTransformer, 무는 segRNN, 사과는 PatchTST 모델을 이용한 경우 예측 정확도가 올라갔다. 하지만 이번 task는 CPU 환경에서 작동해야 하며 모델의 추론 시간 또한 평가 요소에 반영되었다. 그래서 PatchTST, SegRNN,iTransformer와 같은 시간 복잡도가 큰 모델들을 사용하는 대신 Dlinear 모델과 어텐션 메커니즘을 추가한 DlinearwithAttnetion 모델을 TimeSeries 모델로 택했다. 만약 GPU를 사용할 수 있는 컴퓨팅 자원이 제공된다면 각 품목별로 더 다양한, 더 적합한 TimeSeries 모델을 사용하여 예측 성능 향상을 이루어 낼 수 있을 것이다.

머신러닝 모델로 LGBM말고도 Ridge, ExtraTree, RandomForest, CatBoost, XGBoost를 사용하였다. 사용 후 Validation 점수와 리더보드 점수를 복합적으로 고려하였을 때, 정확도 측면에서 LGBM이 최선의 선택은 아니었다. 정확도 측면에서만 10개의 품목을 고려했을 때 ExtraTree가 최선의 선택이었다. 하지만 추론 시간을 고려했을 때, ExtraTree가 LGBM보다 정확도 측면에서 가지는 이점이 추론시간을 고려했을 때는 미비하며 오히려 정확도가 조금 낮은 LGBM을 사용하는 것이 최선의 선택이라 판단하였다. 만약 추론 시간에 대한 제한이 사라진다면 LGBM 이외의 모델들을 택하여 더 정확한 가격 예측이 가능하다.