2024_스마트팜

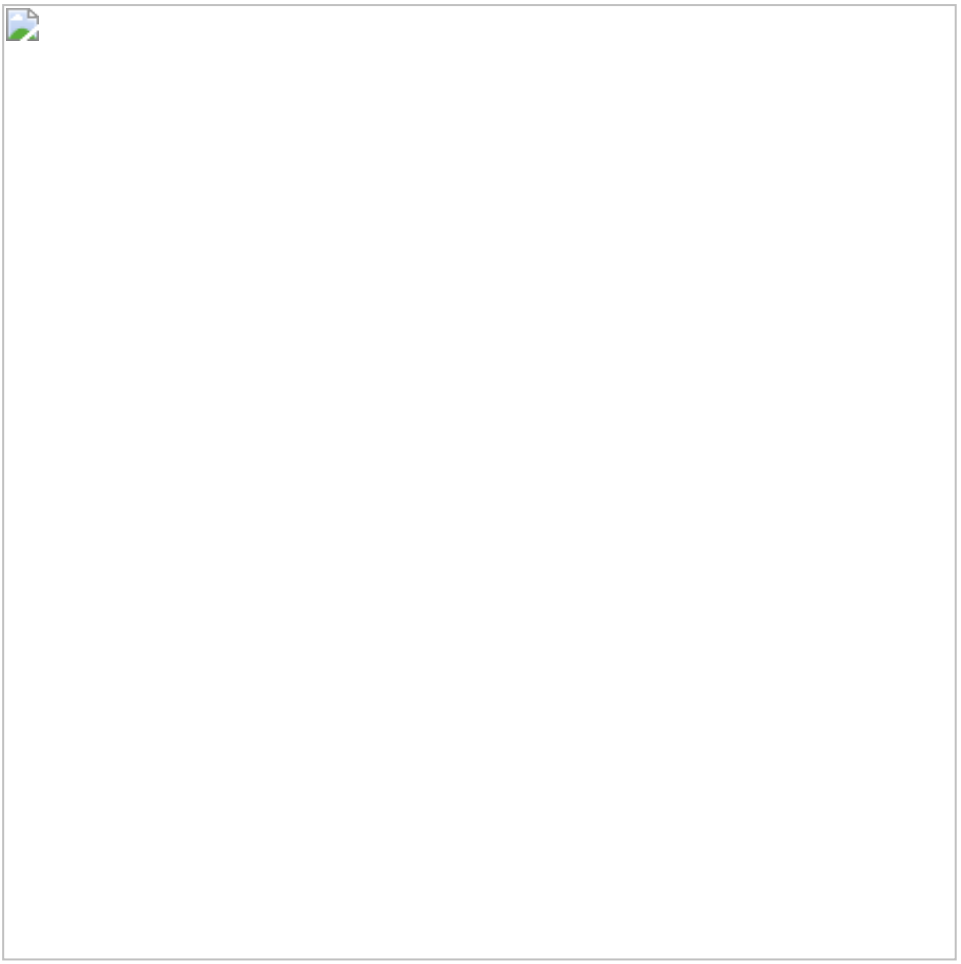
▼ 데이터

데이터 목록

- 내부 환경
  - (스마트팜)참외농장 내부 환경변화에 따른 가격 데이터_데이터
  - (스마트팜)참외농장 내부 환경변화에 따른 생산량 데이터_데이터
  - (스마트팜)참외농장 내부 환경변화에 따른 생육 데이터_데이터

- 외부 환경
  - (스마트팜)참외농장 외부 환경변화에 따른 가격 데이터_데이터
  - (스마트팜)참외농장 외부 환경변화에 따른 내부 환경 데이터_데이터
  - (스마트팜)참외농장 외부 환경변화에 따른 생산량 데이터_데이터
  - (스마트팜)참외농장 외부 환경변화에 따른 생육 데이터_데이터

- Y변수: (스마트팜)참외농장 작물 생육변화에 따른 생산량 데이터_데이터



▼ 외부 데이터 사이트:

- https://www.bigdata-map.kr/
- https://data.mafra.go.kr/main.do
- https://www.n-farm.kr/home/
- https://www.data.go.kr/index.do

최종적으로는 참외의 생산량과 품질(당도나 크기)을 예측(?) → 머신 러닝 모델(?)

일단, 데이터 처리를 열심히... 한 다음에 모델 적합

혹여나 외부 데이터 끌어올 수 있으면 끌어오기

결측값 없는 행들 랜덤 추첨으로 빈칸으로 만들어 놓고 결측값 대체 방법들의 성능 정해서 가장 좋은 성능이 나오는 결측값 대체 방법으로 진행하면 어떨까..?

일단 농장ID와 측정일, 구역명이 같은 건 같은 observation의 값일테니(?) 모든 데이터들을 join해서 결측값들을 대체(?) 하는 형식으로(?)

그리고 날씨 같은 건 기후 데이터 이용해서 결측값 대체해도 될 듯(?)

https://data.kma.go.kr/data/grnd/selectAsosRltmList.do?pgmNo=36 - 시간별 기온, 습도, 풍속 자료 - 지점 상주로 선택하면 됨

일정

Aa Name	Date	Tags
9/29	@2024년 9월 29일	
10/4 회의	@2024년 10월 4일	
10/6 회의	@2024년 10월 6일	

▼ 데이터셋 (연도, 농장코드): 변수

- 내부가격 (2019 at): 내부온도 (기온, 습도, 이산화탄소농도), 출하량, 출하금액
- 내부생산량 (2020 g): 지역온도 (기온, 습도, 풍속), 내부온도, 출하량
- 내부생육 (2020 g): 내부온도, 생육 (엽장, 엽폭, 줄기굵기)
- 외부가격 (2020 g): 지역온도, 외부온도, 출하량, 출하금액
- 외부내부 (2020 r): 지역온도, 외부온도, 내부온도
- 외부생산량 (2019 aa): 지역온도, 외부온도, 출하량
- 외부생육 (2019 at): 지역온도, 외부온도, 생육
- 생육생산량 (2019 u): 생육, 출하량
- 공란인 변수 제외됨 (외부온도 등)

▼ 분석방법 추천 -연주

- 시간대가 2019, 2020으로 다르니까 합치지 말고 따로 분석하는 게 나을 것 같음
- 내부환경 데이터셋이랑 외부환경 데이터셋의 변수들을 내부, 외부로 나누지 말고 내부온도 (습도, 이산화탄소 포함), 지역온도 두개로 나누기 (어차피 외부 온도는 데이터 없음)
- X 변수에 생육도 추가해서 생육, 내부온도, 지역온도 3개로 생각
- **내부가격, 외부생육 데이터 합칠 수 있음 (연도, 농장 동일) → 2019년 내부온도, 지역온도, 생육, 출하량 동시 분석 가능**
- **내부생산량, 내부생육 데이터 합칠 수 있음 (연도, 농장 동일) → 지역온도, 내부온도, 생육, 출하량 동시 분석 가능**
- 위처럼 합치면 2019, 2020 각각 따로 X변수 (내부온도, 지역온도, 생육)랑 Y변수 (출하량) 관계 분석 가능할듯?
- 생산량이 10월부터 2월까지 비는 건 0으로 두기
- 3~9월 결측값은 어떻게 해야할까
- 2019년을 train, 2020을 test로 두고 모델 성능 파악해서 제일 좋은 모델에 다시 2020 train으로 돌리기
- 문제점 1: 2019년 생산량 데이터가 거의 없음. 한 15일정도 있는듯
- 문제점 2: 내부 생산량이랑 내부 생육 데이터가 농장도 같고 연도도 같은데 값들이 조금씩 다름. 내부생육 데이터가 온도랑 습도를 반올림한듯? 그래서 2020 데이터셋의 내부온도 부분은 내부생산량 데이터에서 가져옴 (반올림 안 되고 더 자세함)

2019.xlsx

2020.xlsx

▼ 분석방법 추천 - 형주

일단 솔직히 연도가 같으면 농장 다른건 크게 문제가 아닌거 같음

→ 왜냐면 농장이 달라도 같은 연도 내에서 같은 변수는 값이 같았음

> 년도 분리 data.xlsx

→ 이건 연주가 올린 2019, 2020 데이터에 출하금액 등 빠진 변수들 추가해서 두 개 합친거

→ 여기서 **2019년에는 외부온도 데이터가 있는데 2020년에는 외부온도 데이터가 없음**. 이것만 어떻게 채우면 사실 변수 개수가 2019, 2020가 같아서 위아래로 이어서 2019~2020년도 데이터 한번에 할 수 있긴 함 (combined 탭에 있음)

or 연주가 제안한것 처럼 2019년을 train, 2020을 test로 두고 모델 성능 파악해서 제일 좋은 모델에 다시 2020 train으로 돌리기

▼ 10/4까지 해야 할 것들 정리

일단 지금 해야할 일들:

1. 결측값 처리
    - Y = **생산량(출하량, 출하금액)**
        - 참외 출하는 주로 3월~8월에 진행됨
            - 우리 데이터에서 2019년은 3월~5/28, 2020년은 3/3~8/17까지 있음
            - 근거자료: https://www.yna.co.kr/view/AKR20240223146200030
        - 따라서 1~2, 9~12월은 0으로 둬도 될듯 (출하량, 출하금액 다)
        - 그러면 3~8월 사이의 결측값들은 어떻게 처리?
            - 0으로 둔다
            - 선형 보간, spline 보간법 사용해서 채운다
        - 2019년은 5/28까지밖에 없는데 그럼 6~8월도 0으로 처리할지?
            - 2020년 데이터로 모델을 짜서 2019년 6~8월도 추정값을 채워넣기
    - X = 이건 보간법이든 어떻게든 채우는게 나을거 같음

    - 변수별로 살펴보고 결측치 채울 방안 생각해야 함
    - 외부 온도 데이터 어떻게 할지
    - INDVD_NO - 원본 데이터 보면 얘가 1, 2, 3, 4 이렇게 되어 있는데 시간대가 같음. 근데 지금 우리가 만든 데이터에서는 시간대가 다 다르게 되어 있음. 이거 어떻게 처리?
        - 더미변수??
    - PLANT_AR_SQM
2. EDA, 전처리
3. 외부 데이터 더 찾아서 추가
    a. 토양 성분 데이터 - 돈 내야됨..... 못 씀.

뭔가 지역온도 ~ 외부온도 까지는 시계열로 분석해야할 거 같은데 INDVD_NO 범주 (만약에 만든다면) 랑 길이랑 줄기 관련된 데이터들도 시계열로 봐야할지...? 그래서 지역온도~외부온도 까지 채운다음에 그거에 따라 영향을 받는 어떤 모델로 적합을 시켜서 추정하는 게 맞을지..?

그러니까 결론적으로, **지역온도~외부온도 결측값 채우고 → INDVD_NO 및 참외에 대한 생육정보 예측 → Y변수 예측** 이런 식으로...?

지역온도와 외부온도의 상관관계가 0.7정도로 높았음 → 상관계수 기반의 대체법으로 (다중대체법) 해보는 중. 근데 결측치가 많아도 많아도 너무 많아서 성능 비교를 어케할지...?

방금 뽑아봤는데 겁나 이상한걸로 ㅎ 버려....

~~생산량이 있는 행들만 모아서 모델링을 해 (회귀든 뭐든)~~

~~비어 있는 곳들을 예측값으로 채워~~

~~그 담에 미래 생산량 예측할때는 다 구한다음에 지금 한 해에 예를 들어 10일 정도만 생산량이 있다 그러면 나누기 3을 하던지 이런 방법~~

OR

~~월별 평균치 예측 (그러면 굳이 빈걸 어떻게 해결해야 할지 생각 안해도 됨)~~

▼ 생산량 있는 날짜들

3/5, 12-14, 18-20, 24-26, 28-29, 31

4/2, 3, 8, 15-16, 18-19, 22

5/19, 21-23, 28

3/3, 6, 10, 12-14, 16-21, 23-24, 28, 31

4/1, 6-8, 13-14, 18, 21-25, 27-30

5/1-2, 4-9, 11-12, 15-16, 19, 21, 23, 26, 28, 30

6/1, 3, 5-6, 8-11, 13, 16, 18-20, 22-27, 29-30

7/1, 3, 6-7, 11, 17, 20, 24, 30

8/5-6, 12, 17

2. 데이터가 있는 날만 예측

- 매일 데이터를 예측할 필요가 없다고 판단된다면, **생산량이 기록된 날**의 패턴을 기반으로 예측할 수 있습니다. 이 방식은 실제로 생산된 날만 예측을 하기 때문에, 예측의 정확성을 높일 수 있습니다.
- **모델링 방식**: 과거에 생산량이 존재했던 시점의 패턴을 바탕으로 특정 날짜에 생산이 이루어질지 여부를 먼저 예측한 뒤, 생산이 이루어진 날의 생산량을 예측하는 방식도 가능합니다.
    1. **생산 여부 예측 모델**: 먼저 특정 날짜에 생산이 이루어질지 여부를 분류 문제로 접근합니다. (e.g., 로지스틱 회귀, 랜덤 포레스트 등)
    2. **생산량 예측 모델**: 생산이 이루어지는 날에 한해 생산량을 예측하는 회귀 모델을 사용합니다.

3. 비정형적인 시계열 모델 사용 (Sparse Data Modeling)

- **비정형 시계열 모델**이나 **일부 시점 예측에 강한 모델**을 사용하는 방법도 있습니다. 이러한 모델은 모든 시간에 관측값이 있는 일반적인 시계열 모델과 달리, 불규칙적으로 발생하는 데이터를 잘 처리합니다.
- 예를 들어, **Prophet**이나 **LSTM** 같은 모델을 사용해 불규칙한 시계열을 학습할 수 있습니다. Prophet은 특히 계절성과 휴일 등의 요인을 반영해 불규칙한 시계열에도 잘 맞는 경우가 많습니다.

4. 생산 간격 및 패턴 분석 후 예측

- 데이터를 분석해보면, 생산이 이루어지는 날과 그렇지 않은 날 간에 어떤 **패턴이나 규칙성**이 존재할 수 있습니다. 예를 들어, 특정 날 이후 며칠 동안 생산이 이루어지지 않는 경우가 있을 수 있습니다. 이러한 패턴을 바탕으로 **간격 예측** 모델을 만들어 다음 생산 날짜를 예측할 수 있습니다.
- **생산 간격을 예측**한 후, 그 날의 생산량을 예측하는 두 단계 접근법도 가능합니다.

5. 미래 생산 가능성 확률을 예측

- 매일의 생산 여부를 **확률적으로 예측**하고, 생산 가능성이 높은 날만을 예측 대상으로 삼는 방법입니다. 예를 들어, 어떤 날에 생산될 확률이 70% 이상인 경우만 따로 뽑아 생산량을 예측하는 방식입니다.
- 이는 데이터가 존재하지 않는 날과 비어 있는 날이 무작위적이지 않고 특정 패턴을 따른다면 유용합니다.

다인

너가 말한 방법 괜찮은듯

일단 ~외부온도까지 시계열 모델로 예측해서 채워넣고

이거 바탕으로 참외 특징(잎, 폭 등) 예측해서 채워넣고

이렇게 모인 X들로 생산량 간격 예측 + 값 예측

🎯 최종 결정 사항들

Step 1) 내부 이산화탄소, 외부 온도를 시계열 모델을 이용해서 결측치를 채운다

- VAR, (새벽 3시까자 버텼는데 안 되더라)

Step 2) INDV_NO 개체번호 1, 2, 3, 4에 따라 데이터셋 4개로 나눈다

Step 3) 우선 개체번호 1 데이터셋 사용 → 내부 온도~외부 온도까지 데이터를 '환경변수'로 보고 이를 토대로 잎폭, 옆폭 등 생육 정보 예측해서 결측치 채우기 → 다른 개체들에도 적용

- 우선 정규화 or 표준화 진행해야 할듯
- 환경변수들 5개를 가지고 나머지 여러 Y를 한번에 예측
  - ML: LightGBM, XGBoost
  - DL: LSTM, GRU

Step 4) 이렇게 X변수가 준비되면 위의 GPT가 제안한 2~5 방법으로

1. 2019년의 6~8월에 빈 생산량 데이터 채우고
2. 1~2, 9~12월 생산량 데이터는 0으로 일단 두고
3. 우리의 궁극적 분석 목적인 미래 생산량 예측한다

새로운 문제

1. 개체번호 １２３４가 같은 시간대에 위치함。 그래서 １２３４데이터셋을 나눠서 각각 결측값을 채워봄
2. 문제는 개체번호가 안 쓰여있는 행들이 많은 것
3. 어떤 날짜는 시간대별로 ４개로 나눠져 있고 개체번호가 안 쓰여있는 날짜는 시간대별로 한개 행만 존재해서 분석 어려움
4. 그렇다고 개체번호를 예측하려면 시간대별로 생육 상태를 에측해야 하는데 같은 시간대에서는 환경이 모두 같아서 한 시간대 내에서의 ４개 생육상태를 예측하는 건 불가능
5. 즉 개체번호 예측이 불가능해 보임。 개체번호 결측값도 １１０００개로 많음
6. 결론： 생육상태에 대해서는 １２３４ 개체번호의 평균을 내자。 별로 크게 차이 안 나서 평균이 제일 편리한 방법。 그리고 생산량은 ４배를 해서 시간대별로 행을 １개로만 줄이기。

개체번호, 연도 통합.xlsx

외부온도 변수가 우리가 구하고자 하는 농장에서는 아예 비어있음。그래서 외부생산량 데이터를 통해 지역온도랑 외부온도 사이 관계 구해서 우리 데이터셋의 외부온도 채우기

ｓｔｅｐ１） 시계열 통해 지역온도와 내부온도 결측값 채우기외부생산량 데이터 통해 지역온도와 외부온도 관계 구해서 우리 데이터셋 “개체번호、 연도 통합‘의 외부온도 변수 예측

ｓｔｅｐ２） 외부생산량 데이터 통해 지역온도와 외부온도 관계 구해서 우리 데이터셋 “개체번호、 연도 통합‘의 외부온도 변수 예측

ｓｔｅｐ３） 환경 변수 통해 생육 변수 결측값 채우기 （다중회귀든 뭐든）

ｓｔｅｐ４） 생산량 예측

▼ 외부온도 모델 관련

보간법 말고 ＡＲＩＭＡ 해봤는데 괜찮은 것 같음。 챗지피티가 복잡해서 돌리기 힘들다고 했는데 일단 내 컴으로는 돌아감

ARIMA 모델

장점

- 예측 기반: ARIMA 모델은 시계열 데이터의 패턴과 경향성을 반영하여 결측값을 예측합니다. 과거 데이터를 기반으로 한 예측을 통해 더 정확한 값을 제공할 수 있습니다.
- 시계열의 복잡한 구조 처리: 자기회귀와 이동 평균의 조합으로 데이터의 복잡한 패턴을 잘 설명할 수 있습니다.

단점

- 모델 구축 필요: 모델을 구축하고 파라미터를 설정하는 과정이 필요하며, 이는 시간과 노력을 요구합니다.
- 정상성 가정: ARIMA 모델은 데이터가 정상성을 가져야 하므로, 데이터에 차분을 적용해야 할 수도 있습니다.
- 계산 비용: 대규모 데이터에 대해 ARIMA 모델을 학습시키는 데 시간이 걸릴 수 있습니다.

어떤 방법을 선택해야 할까?

- 데이터의 특성: 시계열 데이터의 패턴이 복잡하고 명확한 경향이 있을 경우 ARIMA 모델이 더 유리할 수 있습니다. 반면, 데이터의 변화가 적고 상대적으로 간단한 경우 보간법이 충분할 수 있습니다.
- 결측값 비율: 결측값이 많고 불규칙하게 분포되어 있는 경우 ARIMA가 더 나은 결과를 제공할 수 있습니다. 보간법은 특정 패턴이 없는 데이터에서는 제한적일 수 있습니다.
- 시간과 자원: 시간적 제약이 있는 경우 보간법이 더 빠르고 효율적일 수 있습니다. ARIMA 모델은 구축하는 데 시간이 소요됩니다.

3. Tensor-based 시계열 모델

- **Tensor Completion**이나 **Matrix Factorization** 기반의 기법을 사용하면, 여러 시계열 변수를 동시에 처리하고 결측값을 채울 수 있습니다. 이 방법들은 특히 시계열 데이터의 다차원적 특성을 잘 반영할 수 있습니다.
- `tensorly`나 `scikit-tensor`와 같은 라이브러리에서 이러한 방식의 모델을 사용할 수 있습니다.

→ 근데 성능이 거지같음 왜 알려준거임;

1. Prophet (페이스북 Prophet 모델)

Prophet은 페이스북에서 개발한 시계열 예측 모델로, **계절성, 추세, 휴일 효과** 등을 쉽게 반영할 수 있는 장점이 있습니다. Prophet은 특히 계절성을 자동으로 탐지하며, 시간별, 일별, 주별, 연별 등의 주기를 다룰 수 있습니다. **시간별 데이터**에도 잘 적용됩니다.

- **장점**:
  - 계절성, 추세, 이벤트(예: 휴일, 기후 변화 등)를 쉽게 다룰 수 있음.
  - 시간 단위, 일 단위, 주 단위 등 다양한 시계열 데이터를 처리 가능.
  - 전에 데이콘할 때 누가 좋다고 했던 것 같음
  - 설치 과정이 좀 지랄맞음

2. LSTM (Long Short-Term Memory)

LSTM은 딥러닝 기반의 **순환 신경망(RNN)** 모델로, 시계열 데이터를 학습하는 데 적합합니다. LSTM은 **장기 의존성**(long-term dependencies)을 잘 처리할 수 있어서 시간별 데이터에서 복잡한 패턴을 학습하는 데 매우 효과적입니다.

- **장점**:
  - 복잡한 시계열 패턴을 잘 학습할 수 있음.
  - 데이터가 비선형적이거나 계절성이 불분명할 때 강력한 성능을 발휘함

5. ARIMA와 머신러닝 모델 결합 (Hybrid Model)

하이브리드 모델은 ARIMA 같은 전통적인 시계열 모델과 XGBoost, LSTM 같은 머신러닝 모델을 결합하여 더 나은 예측 성능을 얻는 방법입니다. 먼저 ARIMA 모델을 적용한 후, 그 잔차(residual)를 머신러닝 모델로 학습하여 두 모델의 장점을 결합할 수 있습니다.

PROPHET 모델

그냥 prophet만 쓰면 다른 것들이랑 큰 차이 없음

근데 예측변수를 따로 추가하면 성능이 엄청 좋아짐

시계열이랑 다른 변수들 한꺼번에 예측변수로 놓을 때 유용함

문제점: 예측변수에 결측값이 존재하면 분석 안 됨

해결법: 예측변수 중 결측값이 가장 작은 것부터 다른 기법들 (보간법 등)로 결측값 채우기. 제일 작은 것의 결측값 채웠으면 그 다음 변수들은 반응변수로 놓고 prophet 실시.

깃허브에 올려둔 파일 (결측값 처리 수정 10-5) 에 자세히 나와있음!

문제점2: MSE 값이 극도로 낮아서 overfitting 우려됨...

MICE

최종 알고리즘

1. MICE 모델로 환경변수들 결측치 채움
2. (다인이 쓴 모델)로 '지역온도'만 고려해서 '외부온도' 예측
    a. 외부 온도는 다 결측치 (데이터가 있는 것도 있는데 농장이 달라서 비우고 시작함)
3. 환경변수 + 외부온도를 X로 두고 생육 정보 결측치 예측
4.

▼ UPDATE

1. 환경정보는 **MICE**로 채우고
2. 외부온도는 다인이가 돌린 **KalmanFilter**로 채웠고

    - 외부온도가 있는 데이터도 있는데 농장ID를 맞추다보니 아예 비어있었음
    - 외부온도는 지역온도로만 돌림
3. 환경정보 + 외부온도로 생육정보 채우기

- plt_lngth만 결측값이 너무 많음
- 네개 생육변수 한번에 채우면 정보 손실이 많음
- plt_lngth만 따로 채우기? —> 나머지 세개를 먼저 채우고 예측변수에 이 채워진 생육변수까지 추가해서 식물길이 예측해봄
- 정보손실 최소화, 반응변수 간의 관계도 반영되는 방법이지 않을까?

4. 이제 채워진 정보들로 생산량 예측 - Transformer? - Temporal Fusion Transformer (TFT),