곡물 생산량 예측 데이터 관련 작업

생성일	@2023년 5월 13일 오전 8:20
태그	

일조시간

- 비어 있는 데이터는 어떻게 핸들링 할 것인가?
  - 한 주에 여러 도시들에 대한 데이터가 있음.
  - 각 도시에 대해 데이터가 없을 때도 있는데 평균을 내서 주를 치환
  - 평균을 낼 때 비어있거나 잘못된 데이터는 제외하고 평균을 냄.
- 결측치는 어떻게 보완할 것인가?
  - 일조 시간 데이터가 1900년부터 1987년까지 밖에 없음.
  - 1900년부터 1987년까지의 데이터를 학습하여
  - 1988년부터 2022년 까지는 regression으로 예측하여 사용.

육류소비, 채소소비, 과일소비 - 홍준님

- 기호식품 빼고는 1961년부터 2021년까지 데이터가 있음.
- 1900년부터 1960년까지 결측치가 있었는데 linear regression으로 대체했음.
- 그래프를 그려보니 선형으로 나와서 1900년부터 1960년까지를 linear regression으로 대체해도 될 듯함.
- 소비량을 단순히 linear regression으로 하면 1900년대 초반까지 가면 음수로 나왔음. 1920년 밑으로는 1920년 값으로 통일해서 대체를 함.

곡물데이터 - 성준님

- 여러 column이 있었는데 필요한 값들만 놔두고 데이터 정리 완료함.
- 공통적으로는 1929년부터 다 있음.
- csv 형태로 다 나와있음.
- LSTM 모델도 만들었음.

온도, 강수량 데이터 → 동준님

- 비어있는 값들이 있었음. 역사가 긴 도시나 농촌 도시만 뽑아서 넣었음.
- 일부 결측치가 있음 : 숫자가 아님. 숫자가 아닐 경우 0으로 바꾸면 됨.
- 1960년부터~현재까지 데이터가 확실함
- 1800년대 후반부터 데이터가 있긴한데 측정 시점이 달라 데이터가 불안정함.
- 옛날 데이터는 월별로 평균치를 채워넣는 방식으로 해결 시도.

비료 가격 → 동준님

- 1960년부터 다 있음.
- regression으로 결측치를 채워넣어야 할 것 같음.

연도별 작물 가격, 인구수 → 혜림님

- 밀, 쌀 개별 가격을 찾아보니 데이터가 일정하지 않은 것 같아 곡물 가격으로 대체 1910~2022년까지 있음.
- 인구수 55년부터 22년까지 데이터가 있음. 결측치 빼고는 선형적으로 증가하는 듯함.
- 인당 칼로리가 문제. 2010년대부터 2020년대까지 밖에 없음.
- 인당 칼로리 데이터는 제외해도 괜찮을 듯함.

유가 → 병무님

- 1860~2021년 국제 유가 데이터가 있음. 연도별로 데이터가 나와있음. 한 연도에 있는 월은 그 연도의 평균 유가로 만듦.

모델

- LSTM으로 먼저 만들었음.
  - 모델 자체의 구조도 너무 간단함.
  - 쌀 생산량 예측을 LSTM으로 하는 것이 맞는가? → 이전 년도의 데이터가 현재 결과값에 영향을 미치지 않은 것 같음.
- 모델 개선 방법 후보
  1. LSTM으로 기후 예측 모델을 만들고, 그 기후에 따라 input data를 넣어서 작물 생산량 예측을 만든다면 의미 있을 것 같음.

    a. 기후 예측 모델과, 작물 생산량 예측 모델을 별개로 가져감
    b. 기후 예측 모델은 시계열 기반으로 만들고(LSTM, RNN), 작물 생산량 예측 모델은 MLP 기반으로 제작.
    c. “작물 생산 예측 모델을 먼저 학습시키고, 어떤 feature가 중요한지 가중치를 보고 해당 가중치를 보고 중요한 Feature에 대한 예측을 하겠다” 는 스토리
2. 각각의 데이터들이 서로 영향을 미치다 보니 단순히 input으로 넣어서 하나의 output으로 뽑아내는게 성능에 유의미한 영향을 줄 것 같지 않음.
    a. 복잡한 상호작용을 포착하려면 transformer 모델을 사용하면 좀 더 잘 포착할 수 있을 것 같음.
    b. 원래 text data 같은 sequential data에 사용하는 모델이기 때문에 우려가 있음.
3. Feature Engineering
    a. PCA(Principal Component Analysis) 사용해서 차원 축소
    b. 모델 개선 보다는 전처리 용도임

- 모델의 Input Output 형태를 어떻게 가져갈 것인가?
    - 토지 면적을 input으로 넣을지? 면적 당 생산량을 예측할 것인지?
    - 전체 생산량을 output으로 넣으면 소비량에 영향을 많이 받을 것이고, 면적 당 생산량을 output으로 잡으면 기후의 영향을 많이 받을 것 같음.
    - 이번 주에는 MLP에 Feature들을 다 넣고 각 Feature에 대한 weight를 확인해보려는 목적이기 때문에 일단 전체 생산량을 output으로 잡아보는게 좋을 것 같음.
    - Input이 무엇인가?
        - output → 2022년 곡물 생산량
        - input : (1월 습도, 2월 습도, … , 12월 습도, 1월 강수량, 2월 강수량, … , 12월 강수량, 유가, 소비량 ,)
        - decision tree로 어떤 feature가 중요한 지 구해봄.

누가 무엇을 할 건인가?

- 데이터 전처리 → 동준님. 혜림님, train test data split

- 발표자료 만들기 → 정훈님.
- 곡물 생산량 예측 모델 만들기 → 성준님.
- Decision Tree 통해서 입력 Feature 중요도 확인하기 → 홍준님. 데이터가 준비돼야 함.
- 발표 → 병무님