2022 UOS 빅데이터 알고리즘 경진대회

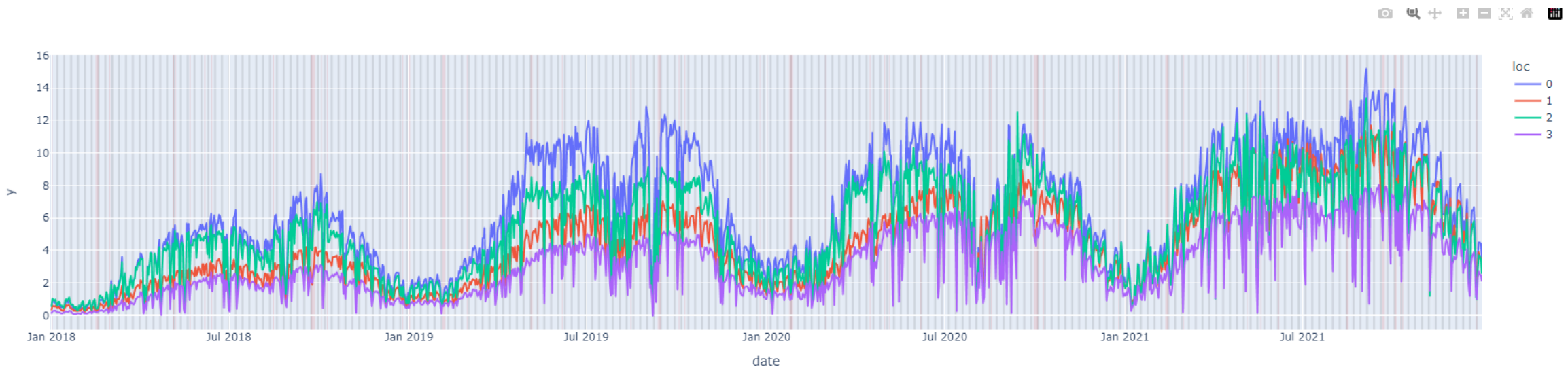
가짜지능 팀

도준형

목차

데이터 분석 및 활용

지역별 차이



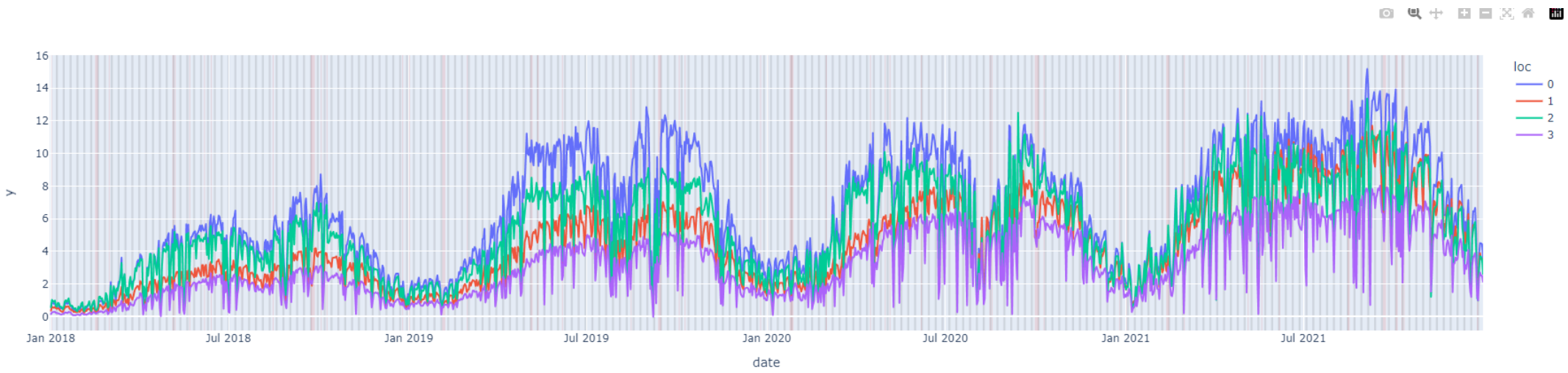
0 : '광진구', 1 : '동대문구', 2 : '성동구', 3 : '중랑구'

지역별 차이

지역	4년간 총 이용량 (1k단위)	2018	2019	2020	2021년
광진구	9071.632	1276.420	2415.974	2398.756	2980.482
동대문구	6185.392	718.708	1342.088	1693.428	2431.168
성동구	7572.292	1116.174	1820.168	2110.926	2525.024
중랑구	4585.710	510.732	977.896	1352.612	1744.470

- 사용량은 모든 년도에서 중랑구 < 동대문구 < 성동구 < 광진구 순으로 많아지는 것을 볼 수 있다.
- 2018년에는 동대문구는 사용량이 중랑구와 가장 차이가 적었으나 2021년에는 성동구와 가장 차이가 적다.
- 광진구를 제외한 세 지역구는 해가 지날수록 증가하는 것을 관찰할 수 있다.
- 광진구는 2020년에 2019년에 비해 약간 감소하였다가 2021년에 증가하는 것을 관찰할 수 있다.

기간별 - 1년


16
14
12
10
8
6
4
2
0
y
Jan 2018
Jul 2018
Jan 2019
Jul 2019
Jan 2020
Jul 2020
Jan 2021
Jul 2021
date
loc
0
1
2
3

기간별 - 1년

- 2019년도 rolling window (20일) 계산 결과

기간별 - 1년

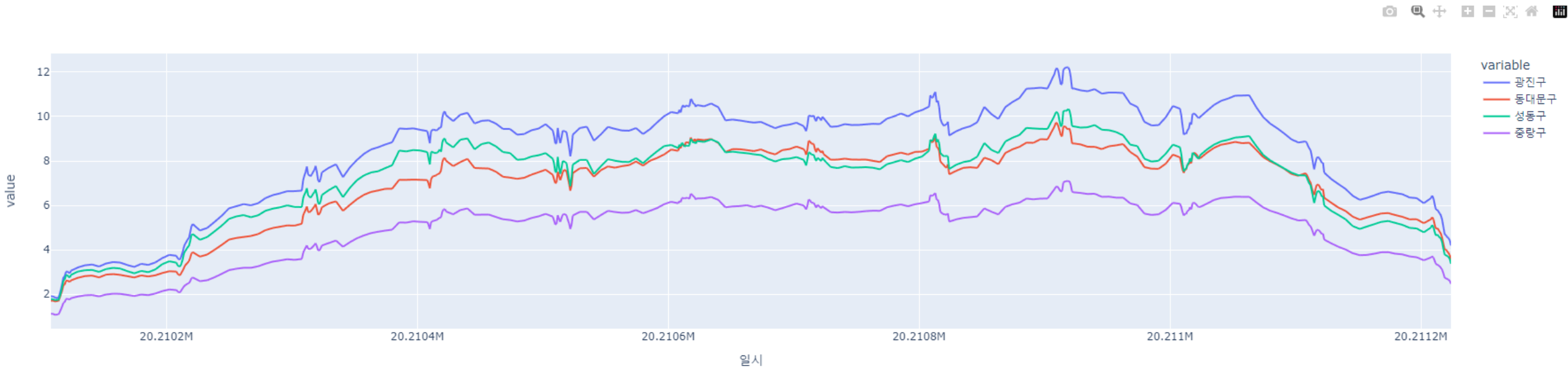
- 2020년도 rolling window (20일) 계산 결과

기간별 - 1년

- 겨울에는 이용자 수가 다른 날보다 상대적으로 적음
- 7, 8월경에 급격히 이용자수가 줄어드는 구간이 있는데, 사용자수가 줄어들기 시작한 날짜와 기상정보를 비교해본 결과 장마철임을 확인함

기간별 - 1년

- 2021년도 rolling window (20일) 계산 결과



기간별 - 1년

- 2021년에는 장마기간이 짧아 7,8월에 이용자수의 감소가 이전에 비해 두드러지게 보이지 않음
- 그럼 2022년도에는?
  - 대회규칙에 따라 이 기간의 기상정보를 사용할 수 없다.
  - 여름 날씨는 변화무쌍하고 미래의 장마기간을 예측하는 작업은 쉽지 않다.
    - 예측을 하기도 어렵다.

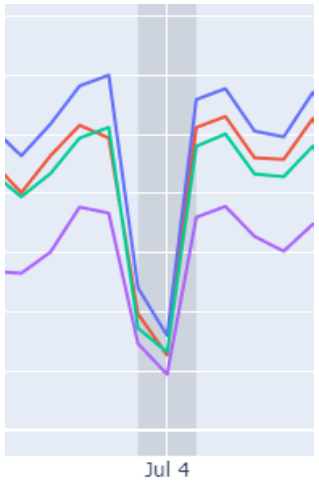
기간별 - 1주일

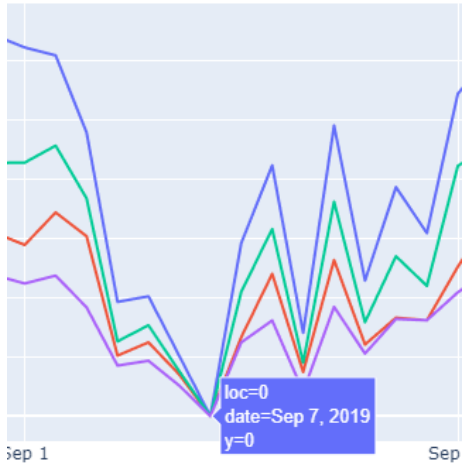
- 평균적으로 토요일, 일요일에 적게 이용하였다.

월	화	수	목	금	토	일
4.65	4.72	4.84	4.81	4.92	4.56	4.30

기타

- 장마 외에도 비가 오는 경우 사용량 감소
- 사용량이 네 지역구 모두 0인 경우가 존재
  - 해당 일자 기사 조회를 해보니 태풍으로 인해 사용이 제한된 경우





모델 학습 전략 및 프로세스

사용한 모델과 전략

- 아래 과정 반복 수행
  1. 학습 데이터 가공
  2. 모델에서 학습할 변수 선택
  3. 2018~2020 데이터를 사용하여 학습, 2021 데이터를 사용하여 validation
  4. Validation 결과 가장 낮은 MAE를 보여주는 모델 선택
  5. 2018~2021 데이터를 사용하여 학습
  6. 2022 데이터로 prediction 수행

사용한 모델과 전략

- Main idea
  - 매년 증가하는 추세를 반영하자
  - 1월 1일부터 12월 31일까지 볼 수 있는 1년간의 패턴을 반영하자
  - 일주일 동안 관찰할 수 있는 패턴을 반영하자
  - 지역별, 3일 연속 공휴일 여부, 주말 여부에 따른 차이를 반영하자
- 위 추세와 패턴을 모두 가산적으로 사용하자

최종 사용 모델

- GAM(일반화 가법모델) 사용
- 각 변수별로 비선형 함수를 fitting한다.
- 학습에 사용된 값들은 아래와 같다.
  - 요일 (weekday), 주말 여부(weekend), 달 (month), 같은 해 1월 1일 기준 일수 차이 (day_of_year), 연도 별 주차 (week_of_year), 3일 이상 연속되는 공휴일 여부 (vacation), 지역구 (loc), 년도(year)
- 공휴일 정보는 한국 천문연구원의 “특일 정보” 사용

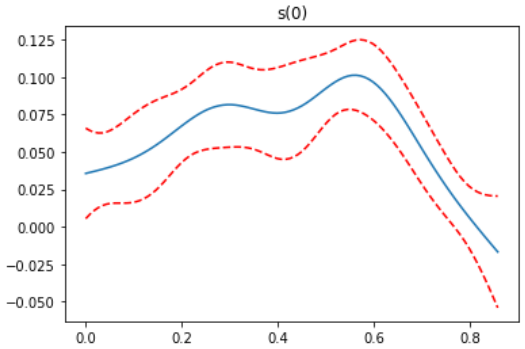
최종 사용 값

- Continuous
  - weekday, month, day_of_year, week_of_year, year
- Categorical
  - weekend, vacation, loc

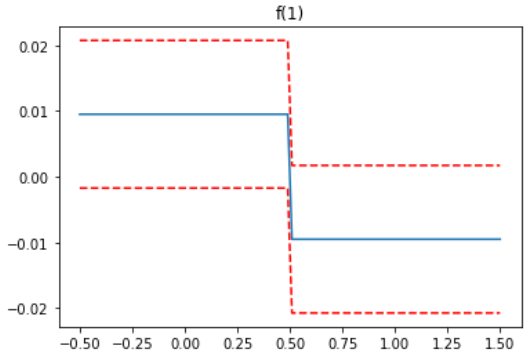
최종 사용 값

- 이용량이 매년 상승하는 추세는 year 에 대해 비선형 적합함수를 fit 하여 적용하자
- 1년 주기로 반복되는 패턴은 month, day_of_year, week_of_year 으로 비선형 적합함수를 fit 하여 학습하자
- 1주일 주기로 반복되는 패턴은 weekday를 사용하여 비선형 적합함수를 fit 하여 학습하자
- 지역별, 주말, 3일 이상 공휴일에 따른 차이는 각각 loc, weekend, vacation을 사용하여 학습하자

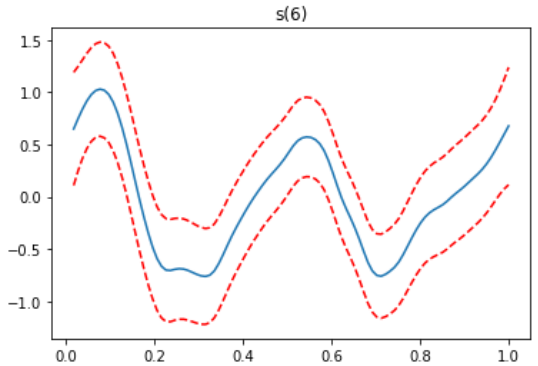
최종 사용 모델의 학습 결과



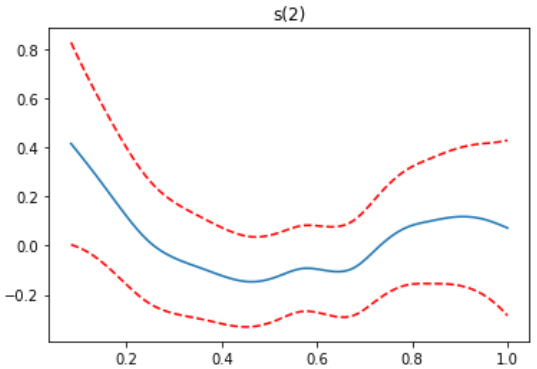
Weekday



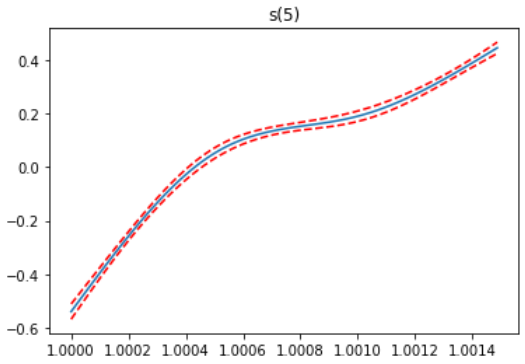
Weekend



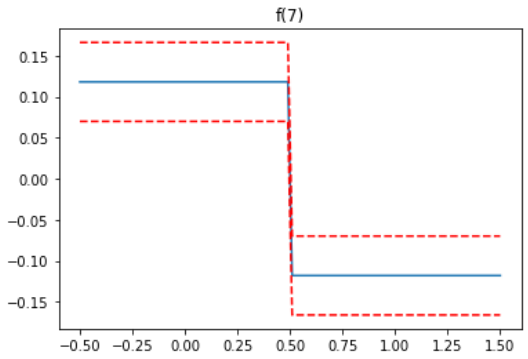
WeekOfyear



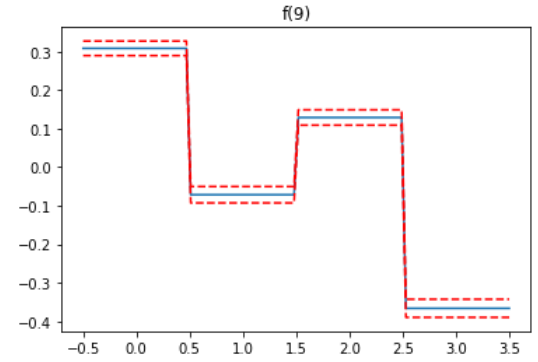
month



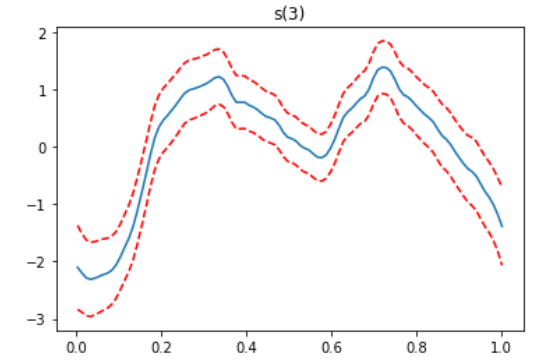
year



vacation



loc



DayOfyear

최종 사용 모델의 학습 결과

- 2018~2020 데이터로 학습 후 2021 데이터로 validation한 MAE
  - 2.07

감사합니다