

# 오산시 주차프로젝트 보고

---

T1F4

구지윤, 송수인, 조성진, 조형진, 최민재

# Contents

---

1. 사용데이터
2. 데이터 전처리
3. 데이터 탐색(EDA)
4. 모델링
5. 결과 시각화

# 사용데이터

데이터 명	기간	대상	설명
오산요청_0821.xlsx	2022-07-01 ~ 2023-06-30(입차기준)  2022-07-01 ~ 2023-07-31(출차기준)	운암공영주차장 오색시장제1공영주차장 오색시장제2공영주차장 오산시청부설주차장	2022년도 하반기 ~ 2023년 상반기  · 주차장 명 · 차량번호(비식별) · 입차시각 · 출차시각

시공단 측에서 제공받은 데이터 (버전3)

# 데이터 전처리: 결측치 (데이터가 없는 부분)

주차장 명	데이터 수	결측치 수	비율
운암공영주차장	694,707	7,263	1.05%
오색시장제1공영주차장	265,849	3,556	1.34%
오색시장제2공영주차장	286,321	5,016	1.75%
오산시청부설주차장	615,594	4,997	0.81%

원시데이터의 결측치 개수와 비율

- 입차는 결측치X
- 차량번호(비식별화)도 결측치X
- 출차에만 20,832개의 결측치를 가지고있기 때문에 출차결측치를 포함한 데이터를 파악하고자하였다.

날짜	주차장명	결측치 수
2022-12-15	오산시청부설주차장	124
2023-06-10	오색시장2공영주차장	112
2022-12-13	오산시청부설주차장	107
2022-12-13	운암공영주차장	98
2022-08-28	오색시장2공영주차장	88
...		

결측치 최대발생 날짜 (상위 5개)

# 데이터 전처리: 결측치

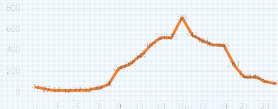
특정한 차량번호의 출차가 모두 결측치인 것을 확인하여 데이터로 파악하고자 하였다.

```
raw['CarNo'] == '3DDAE0BAII4DDgLhThDGWBLvd5+PTg=='
```

- 4곳의 주차장 입출차기록 모두 존재
- 다양한 시간대의 입차기록이 있음. (긴 주차시간이 아님으로 판단)
- 출차는 모두 결측치
- 하루에도 4번 이상 관측되는 것이 대부분 (긴 주차시간이 아님으로 판단)
- 또한 해당 차량의 평균 주차시간을 확인할 수 없음
- 하루에도 주기적으로 주차장을 통과(입차)하기 때문에 주차시간이 길지 않은 것으로 판단하고 분석에서 제외하였다. (총 5957개)

날짜	결측치 수
2022-12-13	285
2022-12-15	282
2023-06-10	117
2023-05-13	113
2022-08-28	95

해당 차량의 날짜별 결측치 수 (내림차순)



해당 차량의 시간대별 결측치 수 (내림차순)

# 데이터 전처리: 결측치

주차장 명	데이터 수	결측치 수	비율 (후)	비율 (전)
운암공영주차장	692591	5147	0.74%	1.05%
오색시장제1공영주차장	265300	3007	1.13%	1.34%
오색시장제2공영주차장	285005	3700	1.30%	1.75%
오산시청부설주차장	613618	3021	0.49%	0.81%

특정 차량번호 제거 후의 결측치: 결측치의 비율이 전체적으로 줄음

다음으로는 차량번호 별 출차결측치를 파악함

출차결측치가 1개인 차량 (9656개)	출차결측치가 2개 이상인 차량 (1545개)
결측치의 분포가 다른 양상을 띠	

출차결측치가 단 한건만 있는 차량과 출차결측치가 두 건 이상있는 차량은 결측치가 관측된 날짜가 다른 것을 확인할 수 있었다.  
이는 다음 슬라이드에서 설명한다.

# 데이터 전처리: 결측치

	주차장 명	일시	count
3847	→ 오색시장2공영주차장	12/14 12시	57
2321	오색시장1공영주차장	9/9 17시	48
3848	→ 오색시장2공영주차장	12/14 13시	45
3846	→ 오색시장2공영주차장	12/14 11시	43
5164	운암공영주차장	6/29 19시	37
5165	운암공영주차장	6/29 20시	23
5171	운암공영주차장	6/30 18시	21
2320	오색시장1공영주차장	9/9 16시	20
1386	오산시청부설주차장	12/15 14시	20
3849	→ 오색시장2공영주차장	12/14 14시	19

출차결측치가 단 1개만 존재하는 차량들의 주차장별 결측치

출차결측치가 단 1개만 존재하는 차량들:

2022년 12월 14일,  
오색2의 결측치가 결측치의 대부분을 차지

→ 12월 14일은 훈련데이터에서 제외.

→ 9월 9일은 추석데이터이므로 제외  
(이유는 후에 서술)

# 데이터 전처리: 결측치

	주차장 명	입시	count
3340	운암공영주차장	6/1 8시	22
3356	운암공영주차장	6/2 8시	20
3382	운암공영주차장	6/5 8시	17
3401	운암공영주차장	6/7 8시	15
3414	운암공영주차장	6/8 8시	13
2629	오색시장2공영주차장	12/14 11시	12
2449	오색시장2공영주차장	10/19 14시	12
2448	오색시장2공영주차장	10/19 13시	11
1147	오색시장1공영주차장	6/3 15시	11
2630	오색시장2공영주차장	12/14 12시	11

출차결측치가 2개 이상 존재하는 차량들의 주차장별 결측치

출차결측치가 단 2개 이상 존재하는 차량들:

운암에서 8시에 지나가는 차량들이 출차가 결측치인 경우가 많다.

운암 8시 입차는 통과차량도 많음  
+ 두 번 이상 결측치가 있음으로 8시에 주기적으로 운암을 통과한다고 판단 가능.

주차대수 예측에 영향력이 없음으로 판단하여 결측치를 데이터에서 제외.

오색시장2공영의 경우 햇빛이 강하거나, 특정 시점(22/12/13-22/12/15)에 차단기에 오류가 있을 것이라고 판단.  
(앞 슬라이드 또한 결측치 다수존재)



# 데이터 전처리: 결측치

차량번호	입차	출차	주차장명
BE79...	2023-06-30 19:02:03	NaT	문암공영주차장
72DA...	2023-06-30 19:01:35	NaT	문암공영주차장
3F1D...	2023-06-30 19:01:20	NaT	문암공영주차장
939F...	2023-06-30 19:00:12	NaT	문암공영주차장
A1A2...	2023-06-30 20:11:45	NaT	오산시청부설주차장
3372...	2023-06-30 20:23:15	NaT	오산시청부설주차장
5D15...	2023-06-30 20:11:28	NaT	문암공영주차장
C4E2...	2023-06-30 20:25:50	NaT	문암공영주차장
CF78...	2023-06-30 20:22:59	NaT	문암공영주차장
E6A8...	2023-06-30 20:53:29	NaT	문암공영주차장
0DE3...	2023-06-30 20:59:33	NaT	문암공영주차장
DDF6...	2023-06-30 20:56:00	NaT	오산시청부설주차장
311D...	2023-06-30 21:03:36	NaT	문암공영주차장
6097...	2023-06-30 21:29:24	NaT	문암공영주차장
A0E6...	2023-06-30 21:09:09	NaT	문암공영주차장
14A5...	2023-06-30 21:07:14	NaT	문암공영주차장
3103...	2023-06-30 21:08:31	NaT	오색시장2공영주차장
311D...	2023-06-30 21:05:15	NaT	문암공영주차장
311D...	2023-06-30 21:02:58	NaT	문암공영주차장
A985...	2023-06-30 22:21:22	NaT	오색시장2공영주차장

( 마지막날 출차 결측 = 아직 출차 하지 않은 차량 )

제공받은 데이터에서 마지막날인 23년 6월 30일에 입차한 차량들 중 아직 출차하지 않은 차량들이 존재할 수 있음 (특히 문암)

따라서, 마지막날의 결측치는 결측이 아닌 데이터가 아직 채워지지 않은 경우로 간주  
-> 제거 X

“

결론적으로는, 입차 기준 23/6/30 입차면서 출차가 결측인 데이터를 제외한 나머지 결측치를 모두 제외

”

# 데이터 전처리: 장기주차

## 장기주차 (24시간 이상 주차)

주차장명	오산시청부설	운암	오색시장1	오색시장2
장기주차(개수)	9,141	2,642	2,411	705

- 오색시장1,2 주차장: 하루 이상 주차하는 차량이 거의 없다는 답변을 받아 24시간 이상 주차하는 차량들의 데이터는 오류로 판단하여 제외.
- 시청 주차장: 새벽동안 관용주차차량이 주차되어있고, 공무원분들의 출근 이후 관용차량을 사용/미사용할 수 있기 때문에 관용주차의 장기주차 가능성 높음. 데이터를 받은 시점이 00시기때문에, 00시에 시청에 주차되어있는 관용차량들이 많고, 이들이 주차면수의 개형에 영향을 끼칠 것으로 판단하여 장기주차를 제외한 대신, 00시에 시청에 주차차량이 많을 것으로 판단하여 초기 주차면수를 조절
- 22년 7월 1일자 이후 입차 기준의 데이터의 한계점, 예를들어 22년 6월 중 입차차량이 7월 1일 이후에 출차한다면 데이터엔 없음, 이로 인해 장기주차차량을 제거하는 것이 맞다고 판단하여 제외
- 운암 또한 장기주차를 고려하여 가능 주차면수 조절 + 장기주차 제외

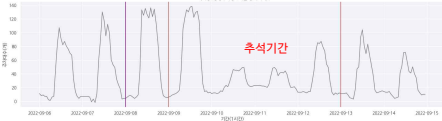
# 데이터 전처리: 추석, 설날



오색시장1,2공영주차장 1시간 당 주차대수



오색시장2공영주차장 1시간 당 주차대수



오색시장1,2공영주차장: 22년9월6일~22년9월15일 추석기간 입출차 개형

## 2) 추석, 설날, 그리고 연휴 전날

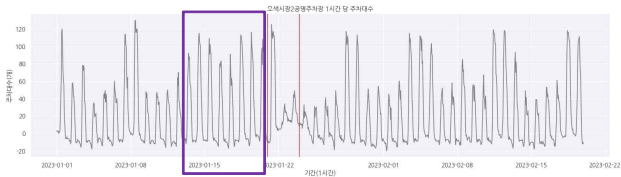
추석(빨간색): 입출차 차단기의 전원을 끈 주차장은 입출차기록X  
입출차가 있으면 무료주차 + 차단기 전원을 끄지 않았음

→ 하지만 1년치의 데이터에서 추석과 설날은 단 두번,  
이상치로 판단하여 훈련에서 제외 (모델에 이상치로 영향을  
끼칠 수 있음)

연휴 전날(보라색)엔 시장에서 장을 보는 인원 증가로, 모델  
훈련에 영향을 끼칠 수 있음.

→ 연휴 전날 또한 항상 혼잡하여 모델링에서 제외

# 데이터 전처리: 추석, 설날



오색시장2공영주차장: 23년1월1일~23년2월22일 설기간 포함 입출차 개형

설날 또한 추석과 비슷한 양상을 보였음.

(연휴 전날에 오색시장2공영주차장 주차대수는 타 날짜, 예를 들어 23년2월1일~23년2월8일과는 다르게 전체적으로 높음을 알 수 있음)

만약 데이터의 기간이 1년이 아닌 3~5년정도가 되었어도, 추석과 설 전 이용객의 증가기간을 파악하여 이들 또한 “설연휴 전 장보기 변수”로 할당하여 훈련하면 적합이 더 잘 됐을 것이라는 아쉬움이 있음.

# EDA: 월별 입출차 건수



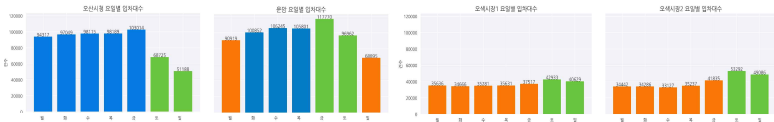
달별로 일수가 다른 것을 고려하면, 월 별 입차대수는 비슷한 것을 알 수 있다.

또한 1년치의 자료로 월별 주기를 확인할 수 없어(각 달이 한 번밖에 없다.) 월별영향은 없는 것으로 판단하였다.

위의 그림에서 X축은 월로 각 숫자는 1월부터 12월을 의미하고 Y축의 숫자는 입차대수를 나타낸다.

4개의 주차장의 입차대수가 크게 차이 나는 것은 주차장의 총 면수가 반영되어 나타난 결과라고 판단된다. 위에서 얘기한 월 별로 일수가 다르다는 것은 2월의 경우 28일, 3월의 경우 31일처럼 차이가 나는 것을 말한다.

# EDA: 요일별 입출차 건수



요일별 입출차 건수를 확인해보면, 각 주차장별로 **요일별 입차대수가 다른 것을 볼 수 있다.**

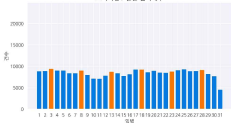
위의 그림에서 X축은 요일로 순서대로 월요일부터 일요일을 의미하고 Y축의 숫자는 입차대수를 나타낸다.

그래프에서의 색깔은 주차장별 요일별로 비슷한 경향을 가진 그룹임을 의미한다.

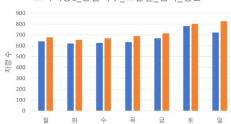
예를 들어, 오색시장1,2와 시청의 경우 평일에 비해 주말의 차이가 발생하여 평일과 주말 2개의 그룹으로 나누어 색칠하였다. 여기서 시청과 운암은 오색시장1,2와 달리 주말에 입출차가 적음을 알 수 있다. 이는 후에 설명할 EDA: 오산시청에서 설명할 오산시청의 출근 그룹과 민원인 차량때문에 일어난 것으로 생각된다. 운암의 경우 평일에서도 월요일, 금요일과 다른 요일에서 차이가 발생하였으므로 다른 주차장처럼 나누는 것이 아닌 좀 더 세밀하게 나뉘어야 할 것으로 생각되어 위의 그림과 같이 3개의 구간으로 나누었다.

# EDA: 일별 입출차 건수

오색시장1 일별 입차대수



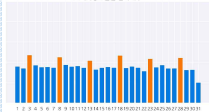
오색시장1\_장날여부\_요일별\_입차\_평균



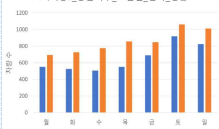
위 그림 두개는 오색시장1의 일별(가로) 입차대수(세로)와 장날여부(주황이 장날)에 따른 요일별(가로) 입차 평균(세로)을 나타낸다

오색시장 1은 장날에 따른 요일별 입차대수의 평균이(아래 그림) 운암, 시청에 비해 차이가 나는 것을 알 수 있다.

오색시장2 일별 입차대수

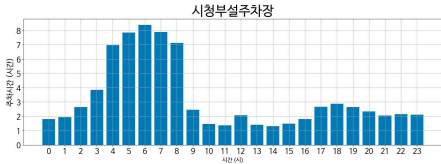
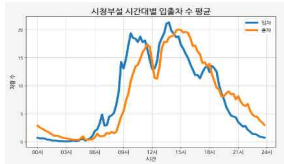


오색시장2\_장날여부\_요일별\_입차\_평균



오색시장 2가 장날에 따른 요일별 입차대수의 평균이 타 주차장에 비해 도드라지게 다른 것을 알 수 있다. 위의 그림을 예시로 들면, 특히 화-목의 입차평균이 장날이 아닌날은 500대정도인 반면에 장날인 날은 700~800대정도가 입차하였다.

# EDA: 오산시청부설주차장



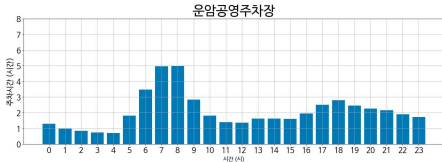
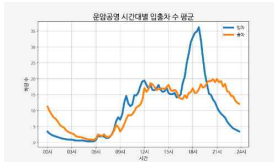
좌측 표의 파란 선이 입차, 주황 선이 출차로, X축이 시간대(0시~23시), Y축이 차량 수이며, 시간대 별 평균 차량수를 개형했다.  
6시 전까지는 출차 차량이 입차 차량보다 많지만, 6시를 기점으로 입차차량이 급격하게 증가하며(출근차량), 20시부터 24시부터는 출차차량이 입차차량보다 많음을 확인할 수 있다.

16시부터 20시 사이는 입차차량이 더 많은데, 이는 퇴근차량과 시청을 저녁에 이용하는 차량이 있어 그러함으로 유추할 수 있다.  
우측 표는 장기주차(24시간)를 제외한 입차차량의 평균 주차시간이다. 확인해보면 4~8시 사이 평균 주차시간이 7시간 이상으로, 출근차량이 많음을 확인할 수 있다.

10시부터 16시사이는 대부분이 민원차량이거나 잠시 주차했던 차량들로 파악할 수 있다.  
17시~19시 사이는 저녁식사나 주변 볼일로 차량을 주차했음을 알 수 있다. (3시간 가량의 평균 주차시간)



# EDA: 운암공영주차장

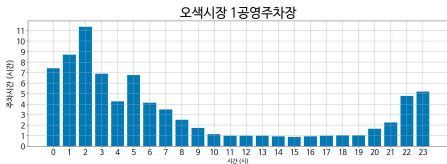
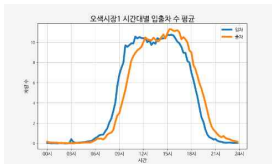


좌측 표의 파란 선이 입차, 주황 선이 출차로, X축이 시간대(0시~23시), Y축이 차량 수이며, 시간대 별 평균 차량수를 개형했다. 6시부터 12시까지는 입차차량이 출차차량보다 더 많지만, 14시부터는 입차차량보다 출차차량이 더 많으며, 16시를 기점으로 입차차량이 가파르게 올라가는 것을 확인할 수 있으며, 18시는 입출차 수가 폭등함을 알 수 있다. 이는 운암공영주차장 주변 상권의 영향을 받은 것으로 사료되는데, 주변에 저녁식사나 여가시간을 즐길거리가 많기 때문이다.

그 후 21시부터 다음날 3시까지는 입차차량이 서서히 빠져나가는 모습을 볼 수 있다.

우측 표는 장기주차(24시간)를 제외한 입차차량의 평균 주차시간이다. 확인해보면 6시~8시 사이의 입차차량의 평균 주차시간이 5시간임을 알 수 있는데, 이는 운암공영주차장이 “환승주차장”이기 때문에 7~8시에 주차장을 일직선으로 통과하는 차량이 최소 전체의 30퍼는 차지하기 때문에(입출차 사이 시간이 5분 이하) 주변 상권 사장님들의 출퇴근 주차시간을 상쇄시킨 것으로 보여진다.

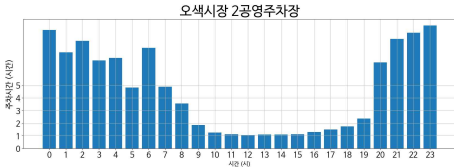
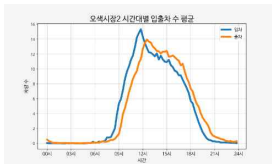
# EDA: 오색시장1공영주차장



좌측 표의 파란 선이 입차, 주황 선이 출차로, X축이 시간대(0시~23시), Y축이 차량 수이며, 시간대 별 평균 차량수를 개형했다. 12시 전엔 입차가 출차보다 많으며, 12시 이후로는 출차가 입차보다 많음을 알 수 있다. 또한 출차가 입차보다 1시간 간격정도로 후행하는 것을 확인할 수 있는데, 이는 우측 표에서의 주차시간으로 확인할 수 있다.

우측 표는 장기주차(24시간)를 제외한 입차차량의 평균 주차시간이다. 22시~5시 사이의 입차차량의 평균 주차시간이 5시간부터 11시간까지 되는 이유로는 표본 수가 적어 평균 11시간이어도 1대~2대정도가 입차한 것이기 때문에(좌측 표 참고) 중요하지 않으며, 입출차가 많은 10시~19시 사이의 차량들의 평균 주차시간이 1시간임을 중요하게 보면 된다.

# EDA: 오색시장2공영주차장



좌측 표의 파란 선이 입차, 주황 선이 출차로, X축이 시간대(0시~23시), Y축이 차량 수이며, 시간대 별 평균 차량수를 개형했다. 10시부터 16시까지 출차가 입차보다 1시간 간격정도로 후행하는 것을 확인할 수 있는데, 이는 우측 표에서의 주차시간으로 확인할 수 있다.

우측 표는 장기주차(24시간)를 제외한 입차차량의 평균 주차시간이다. 오색시장2또한 20시~2시 사이의 입차한 차량의 평균 주차시간이 6시간 이상인 것으로, 주변 주택가의 차량이 밤 중에 주차가 된 것으로 유추할 수 있다. 또한 6시에 입차한 차량의 주차시간도 8시간정도로 오색시장의 주변 상권 상인들이 출근하면서 주차했다는 유추가 가능하다

# 모델링

---

1. 주차예측 모델 정의(회귀/분류)
2. 변수 설정(피쳐/타겟)
3. 모델 선정
4. 훈련/검증/평가 나누기, 기간 나누기
5. Test Set 결과 (회귀/분류-혼잡도)
6. 결과 시각화
7. 모델 자동화

# 모델링: 주차예측모델 정의

오산시의 공영주차장에서 남은 주차면수를 예측하기 위한 목적에 따라 변수 Y는 다양한 유형이 될 수 있다.

## 회귀모델

남은 주차면수를 수치적으로 예측

EX) 현재 오색시장2공영주차장에  
150대의 주차공간이 남았다.

## 분류모델

남은 주차면수를 범주형으로 예측

EX) 현재 오색시장2공영주차장은  
혼잡입니다.

모델을 과대추정하기 위해 회귀모델로 더 섬세한 모델을 만든  
후, **결과값(수치형)을 구간화하여** 출력할 예정

# 모델링: 변수설정 (Y변수)

제공기관: 오산시 시설관리공단			
CarNumber	In	Out	주차장 명
2744A3D3HGzaqjHp	2022-07-01 00:14:47	2022-07-01 00:26:47	오산시청부설주차장
788429D8DddulCDT	2022-07-01 00:21:08	2022-07-01 00:24:27	
9F5C1998pytvGAH+	2022-07-01 00:42:08	NaN	

특정 시점의 남은 주차대수를 알 수 없는 상황에서 입출차 변수를 활용하여  
누적 주차대수 변수(타겟변수)를 정의해야 했음.

# 모델링: 변수설정 (Y변수)

변수 명	의미	참고사항
증감	시간대별 입출차의 총 합	음수: 출차가 더 많음 양수: 입차가 더 많음.
누적	현 시점까지의 입출차의 총 합	

Index	증감	누적(특정 시점의 주차면수)
2022-07-01 18:15:00	12	273
2022-07-01 18:30:00	-7	266
2022-07-01 18:45:00	3	269

누적 주차대수를 파악하기 위해 입출차 기록을 입차와 출차를 각각 +1, -1 로 변환하여 각 시간대별로 리샘플링하였다.

# 모델링: 변수설정 (X변수)

변수 명	의미	참고사항
월	날짜에서 '월' 추출	
일	날짜에서 '일' 추출	
요일	날짜에서 '요일' 추출	
시간	날짜에서 '시간' 추출	
주차시간	차량에서 주차한 시간으로 '출차시각'에서 '입차시각'을 빼서 계산하여 분으로 환산.	장기주차의 기준은 24시간 이상으로 생각.
장날	오색시장의 장날 여부를 나타낸 변수	장날은 매월 3, 8일 오일장.
주말	주말 여부를 나타낸 변수	
공휴일	공휴일 여부를 나타낸 변수	
샌드위치	샌드위치 여부를 나타낸 변수	
3일 이상	3일이상 쉬는 날의 여부를 나타낸 변수	주말과 공휴일이 겹친 경우 Ex) 2022-08-15(월)

다음은 데이터 분석과 모델의 성능향상을 위해 기존의 데이터로부터 생성한 변수이다.  
외부변수로는 강수량과 기온데이터를 사용했다. (출처: 기상자료개방포털의 방재기상관측데이터(AWS))



# 모델링: 모델선정

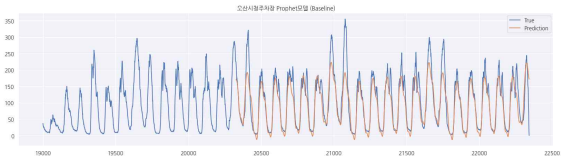
## 모델링에 사용된 기법

- ElasticNet, Logistic Regression과 같은 기본 회귀분석
- Dlinear, Nlinear과 같은 딥러닝 기반 + 전통적인 시계열모델
- Random Forest, LightGBM, CatBoost, XGBoost, Gradient Boosting과 같은 부스팅/배깅 모델

- 기본 회귀분석: 적합이 잘 안됨. 모델의 설명력이 부스팅/배깅모델에 비해 낮음
- 시계열모델: 예측대상까지의 거리가 멀어질수록 정확도가 급격히 낮아짐 (일관된 정확도를 구현하기 위해서는 부적절), 월별 영향(Seasonality) 없음, 주차면적 제한으로 인해 추세(Trend) 없음, 하루 단위(0시~23시)의 사이클이 있지만, 시계열을 시도하기엔 부족함.

“  
딥러닝을 적용하기에는 너무 단순한 데이터와 피쳐, 그리고 짧은 기간으로 성능이 가장 좋았던  
ML 기법인 부스팅/배깅모델 선정  
”

# 모델링: 모델선정 (시계열 제외 이유)

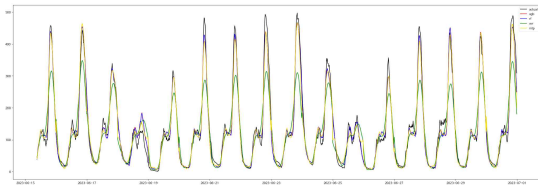


위의 모형은 시계열 모델인데, 극단값(혼잡도)을 잘 따라가지 못하는 것을 알 수 있다.

+

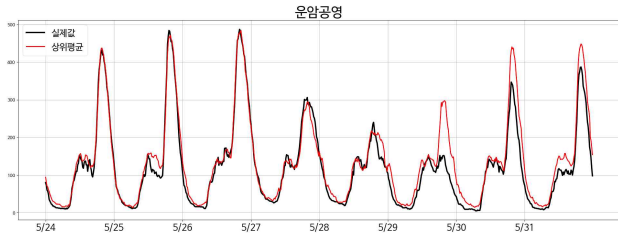
추세가 없으면 적합이 잘 안되는 Dlinear, Nlinear 모델 ...

# 모델링: 모델선정 (SVR 제외 이유)



초록색이 SVR이다.  
타 부스팅 기법들보다 잘 따라가지 못하는 것을 확인할 수 있다.

# 모델링: 모델선정 (선정 이유)



부스팅 중 XGBOOST 모델을 이용하여 운암공영주차장을 적합한 결과다.

# 모델링: 모델선정 (선정 이유 - XGBOOST)

Model	MAE	MAPE
XGBoost	13	0.2245
LightGBM	13.9	0.2505
CatBoost	14.08	0.2463
Random Forest	15.55	0.2562
Gradient boosting	19.66	0.3339

부스팅 모델들의 평가지표 비교

LightGBM: 데이터 개수가 적어 제외

CatBoost: 속도가 느리며, 범주형 자료에 강점이 있어 제외

Gradient Boosting: XGBoost와 속도의 차이가 있으며, XGBoost가 더 성능이 좋음으로 제외

XGBoost는 분류, 회귀 모델 문제에 모두 사용할 수 있는 강력한 모델이며 각 이터레이션에서 맞추지 못한 데이터에 가중치를 부여하여 모델을 학습시키는 부스팅 계열의 트리모델이다.

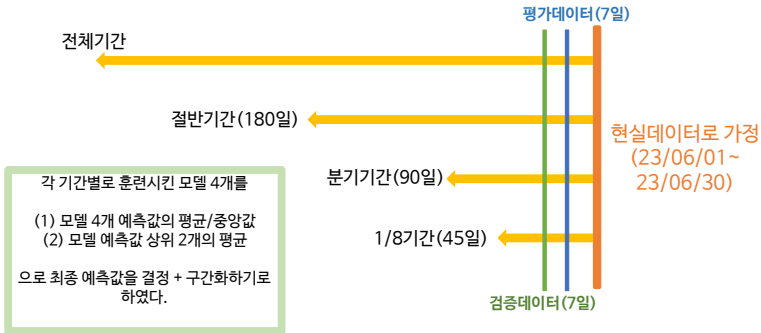
또 강력한 병렬 처리 성능과 자동 가지치기 알고리즘이 적용되어 다른 모델 대비 빠른 속도를 보이며 과적합 규제 기능에서도 이점을 보인다

# 모델링: 모델선정 (XGBOOST 파라미터)

```
model = xgboost.XGBRegressor
param = {
    'max_depth': trial.suggest_int('max_depth', 1,
10),
    'learning_rate':
trial.suggest_float('learning_rate', 0.01, 1.0),
    'n_estimators':
trial.suggest_int('n_estimators', 50, 1000),
    'min_child_weight':
trial.suggest_int('min_child_weight', 1, 10),
    'gamma': trial.suggest_float('gamma', 0.01,
1.0),
    'subsample': trial.suggest_float('subsample',
0.01, 1.0),
    'colsample_bytree':
trial.suggest_float('colsample_bytree', 0.01, 1.0),
    'reg_alpha': trial.suggest_float('reg_alpha',
0.01, 1.0),
    'reg_lambda':
trial.suggest_float('reg_lambda', 0.01, 1.0),
    'random_state':
trial.suggest_int('random_state', 1, 1000),
    'objective': weighted_mse_loss_
}
```

- max\_depth: 최대 Tree 깊이 수 10으로 최대 1024 리프노트 형성
- learning\_rate: 0.01 낮은 비율로 설정하여 과적합방지
- n\_estimators: 높은 값으로 설정하여 과적합방지
- min\_child\_weight: min\_child\_weight를 기준으로 추가 분기 결정(크면 Underfitting)
- gamma: overfitting 방지용이며 너무 높으면 underfitting이 되므로 적당값을 지정
- subsample: 각 트리마다 관측 데이터 샘플링 비율
- colsample\_bytree: 각각의 트리(스텝)마다 사용할 컬럼의 비율
- colsample\_bylevel: 각각의 트리(depth)마다 사용할 컬럼의 비율
- reg\_alpha: noise, outlier와 너무 튀는 데이터들을 잡아준다
- reg\_lambda: 불필요한 가중치를 0으로 만들어서 무시하도록 한다
- objective: weighted mse loss

# 모델링: 훈련/검증/평가, 기간 나누기



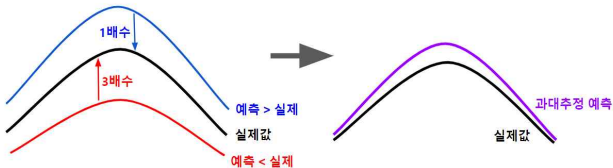
# 모델링: 모델 훈련 특징 - 과대추정

1. Custom Loss (Weighted MSE Loss) 로 모델을 훈련

-> 모델을 데이터로 훈련시킬 때, 모델은 실제값과 예측값의 차이를 최대한 줄이는 방식으로 모델을 만든다.

→ 우리는 예측결과가 실제값보다 낮은 상황(혼잡인데 보통으로 예측하는 경우)을 최대한 피하기 위해 맞춤 손실함수를 정의했다.

추가) Custom Loss뿐만 아니라, Custom Metric 또한 정의했다. (Weighted MAE Metric)



T1F4조 맞춤 손실함수 시각화



# 모델링: 모델 훈련 특징 - 과대추정

주차장명	오산시청	운암공영	오색시장1	오색시장2
과대 : 과소	1 : 20	1 : 25	1 : 2	1 : 3
주차면수	458	593	105	180

(예측결과 < 실제값)의 경우 손실함수를  
일반적인 경우보다 **크게 가중**하여  
모형의 (예측결과 < 실제값)이 되도록 유도한다.

가중치의 크기는 각 주차장의 주차면수 크기를 고려하여 선정했다.

예를 들어, 시청의 경우 예측값이 10이고, 실제값이 20이면,  
손실은 10이 아닌  $10 * 20 = 200$ 이 된다.

# 모델링: 모델 훈련 특징 - 상위평균

상위평균 = 차선의 방법! (과대추정)

최댓값에 비해 과대추정 + **적정 수준의 정확도 유지**

전체기간  
절반기간  
분기기간  
1/8 기간

→ 4개의 모형

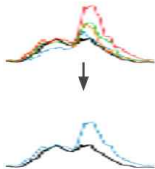
⇒ 4개의 예측결과로 1개의 예측결과를 내야하기 때문에  
가장 실제값에 가까울 수 있게 예측값을 앙상블하여 최종 예측값을 산출.

검은색 = 실제값, 타 색깔 = 예측값

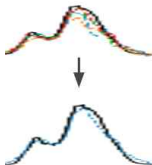
→ [1]의 경우 모델들이 (검은색 제외 4개의 색깔 선) 모두 실제값보다 적은 추정값인 것을 볼 수 있다. 여기서 평균을 낸다면 더 낮아질 수 있기때문에, 상위 2개의 평균값([1]아래 파란선)을 최종예측값으로 선정하여 낮은 예측값을 방지했다.

→ [2]의 경우 모델들이 (검은색 제외 4개의 색깔 선) 모두 실제값보다 높은 추정값이다. 여기서 상위 두개의 평균을 통해 눈에 띄게 높게

[2] 적정수준의 정확도 유지



[1] 낮은 예측을 방지



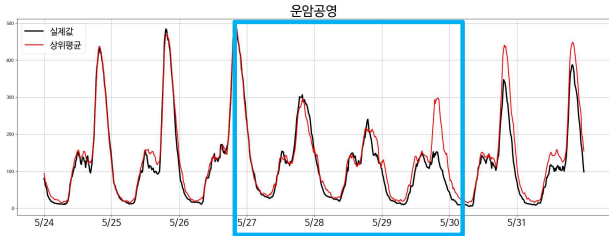
# 모델링: 평가데이터셋 결과

현재 우리의 Test Set은 현실데이터(6/1~6/30) 전인 5/24~5/31이다.

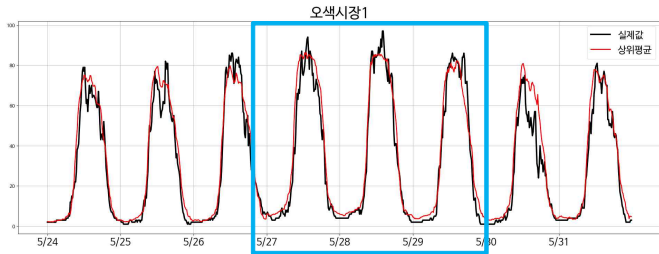
기간 선정의 이유: 공휴일 + 대체공휴일이 있는 연휴기간이라 난이도가 가장 높을 것으로 예상

5 / 24	5 / 25	5 / 26	5 / 27	5 / 28	5 / 29	5 / 30	5 / 31
수	목	금	토	일	월	화	수
			주말	석가탄신일	대체공휴일		

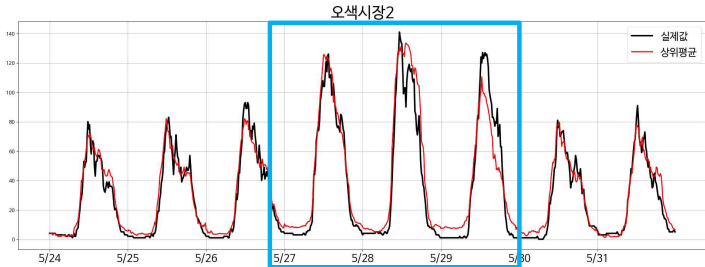
# 모델링: 평가데이터셋 결과 (운암)



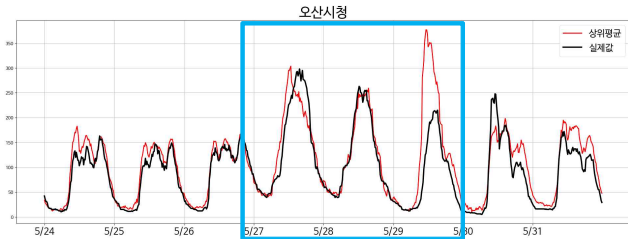
# 모델링: 평가데이터셋 결과 (오색1)



# 모델링: 평가데이터셋 결과 (오색2)



# 모델링: 평가데이터셋 결과 (시청)



시청의 경우가 여러 이유들로 예측이 가장 어려움

향후연구시 필요한 정보들: 행사정보/관용차량정보/광장운용정보/지하주차장개방정보

# 모델링: 혼잡도 정의

EDA과정을 통해 알아낸 주차장별 특징을 반영한 혼잡도를

→ 매우 혼잡/혼잡/보통/원활/매우 원활 총 5단계로 나누었다.

ex) 시장의 경우, 주차면수가 다른 주차장보다 적고 위치적 특성을 반영하여 혼잡의 기준을 높게 하였다.

: 매우 혼잡 80%, 혼잡 : 60%, 보통 40% , 원활 20%





# 모델링: 모델 자동화

MLflow를 이용하여 모델 자동화를 하여 실무에 쉽게 적용할 수 있게 하였다.

각 주차장별, 각 기간별로...

Train, Validation을 이용하여 하이퍼파라미터 Fix ->

Test에 적용 후 Metric 출력 ->

가장 좋은 모델을 MLflow에서 Production Mode로 변환 ->

Real Data (6/1~6/30) 사이에 적용하여 결과 시각화



데이터가 최신으로 들어오도록 함수화하여 코드 자동화 + 원격 레포지토리 활용으로 원할 때 학습 + 저장 가능

								Metrics	
<input type="checkbox"/>	Ⓢ	Run Name	Created	⌵	Duration	Source	Models	MAPE	R2
<input type="checkbox"/>	Ⓢ	capricious-snake-786	8 days ago		17.4s	train.py	xgbr_운암공영_./1	0.263	0.97
<input type="checkbox"/>	Ⓢ	respected-skunk-100	8 days ago		15.9s	train.py	xgbr_운암_en_./6	0.27	0.933
<input type="checkbox"/>	Ⓢ	persistent-calf-467	8 days ago		15.3s	train.py	xgbr_운암_en_./5	0.381	0.919

출처: T1F4의 모델 저장소

# 결과 시각화

아래 그림은 위에서 설명한 모델을 통해 예측한 값과  
사전에 정한 혼잡도 5단계를 시각화한 것이다.



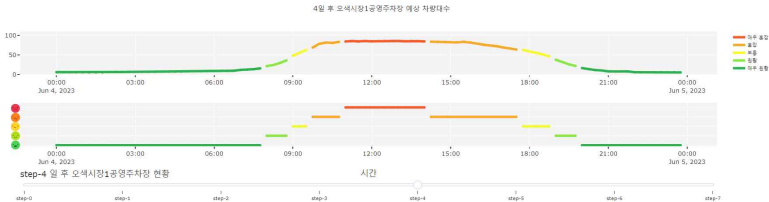
\*Step-1: 오늘  
\*Step-7: 6일 후

## 결과 시각화



# 결과 시각화

아래 그림은 위에서 설명한 모델을 통해 예측한 값과  
사전에 정한 혼잡도 5단계를 시각화한 것이다.



\*Step-1: 오늘  
\*Step-7: 6일 후

# 결과 시각화

아래 그림은 사용빈도가 높은 6시~23시 각 주차장의 혼잡도를 표 형식으로 나타낸 것이다.

좌측의 버튼을 통해 원하는 날(1일 후~7일 후)의 예보를 확인할 수 있다.

오색시장1공영주차장 현황



Remove Images
1일 후
2일 후
3일 후
4일 후
5일 후
6일 후
7일 후

시간	6시	7시	8시	9시	10시	11시	12시	13시	14시	15시	16시	17시	18시	19시	20시	21시	22시	23시
00분	😊	😊	😊	😊	😐	😞	😞	😞	😞	😞	😞	😐	😐	😊	😊	😊	😊	😊
15분	😊	😊	😊	😐	😞	😞	😞	😞	😞	😞	😞	😐	😐	😊	😊	😊	😊	😊
30분	😊	😊	😊	😐	😞	😞	😞	😞	😞	😞	😐	😐	😐	😊	😊	😊	😊	😊
45분	😊	😊	😊	😐	😞	😞	😞	😞	😞	😞	😐	😐	😐	😊	😊	😊	😊	😊

\*1일 후: 오늘  
\*7일 후: 6일 후

# 결과 시각화

아래 그림은 사용빈도가 높은 6시~23시 각 주차장의 혼잡도를 표 형식으로 나타낸 것이다.

좌측의 버튼을 통해 원하는 날(1일 후~7일 후)의 예보를 확인할 수 있다.

오색시장1공영주차장 현황



\*1일 후: 오늘  
\*7일 후: 6일 후

# 결과 시각화

아래 그림은 사용빈도가 높은 6시~23시 각 주차장의 혼잡도를 표 형식으로 나타낸 것이다.

좌측의 버튼을 통해 원하는 날(1일 후~7일 후)의 예보를 확인할 수 있다.

오색시점1공원주차장 현황



시간	6시	7시	8시	9시	10시	11시	12시	13시	14시	15시	16시	17시	18시	19시	20시	21시	22시	23시
00분	😊	😊	😊	😬	😬	😡	😡	😡	😡	😬	😬	😬	😊	😊	😊	😊	😊	😊
15분	😊	😊	😊	😬	😬	😡	😡	😡	😬	😬	😬	😬	😊	😊	😊	😊	😊	😊
30분	😊	😊	😊	😬	😬	😡	😡	😡	😬	😬	😬	😬	😊	😊	😊	😊	😊	😊
45분	😊	😊	😊	😬	😬	😡	😡	😡	😬	😬	😬	😬	😊	😊	😊	😊	😊	😊

\*1일 후: 오늘  
\*7일 후: 6일 후

이상입니다.

---