전력 데이터

분석 완료 보고서

LS 빅데이터 스쿨 4조

오재현 박수현 임혜빈 황보우

2023.11.03 (금)

0. 배경

비즈니스 배경

전기사용량 지난해보다 10% 늘면 요금은 50% '껑충'…누진제 주의

2023.08.07 15:24 입력

박상영 기자

한전은 "전력 소모량이 큰 에어컨을 많이 사용할수록 요금부담이 큰 폭으로 증가할 수 있다"고 강조했다. 한전에 따르면 평소 한 달 전기 사용량이 283kWh인 가구가 여름철 에어컨을 평균 수준(월 160kWh)을 추가 사용한다면 요금은 약 4만1000원 증가한다. 그러나 에어컨 사용량이 평균보다 30%나 증가할 경우(208kWh) 요금 부담은 약 6만4000원 늘어난다.

수요가 역대 여름 중 가장 높았음에도 정부는 이보다 많은 104GW의 전력공급 능력을 확보함에 따라 예비전력은 10GW가 넘었다.

만일 예상보다 수요가 더 많거나 일부 발전소 고장 등으로 공급 능력이 줄어들어 예비력이 5.5GW 밑으로 떨어지면 전력 수급 경보가 발령된다. 가장 낮은 단계인 '준비'를 시작으로 추가 예비력 감소에 따라 '관심'(예비력 3.5~4.5GW), '주의'(2.5~3.5GW), '경계'(1.5~2.5GW), '심각'(1.5GW 미만)으로 격상된다.

* 출처 : 경향신문

비즈니스 계획안

전기 사용 요금

- 한국 전기요금의 경우, 주택용/산업용(일반용) 등 건물 용도별 요금 체계 상이함
- 전기요금 = 기본요금 + 전력량요금 (+ 기후환경요금 ± 연료비조정요금)

< 주택용 전기 요금 누진제 - 하계 (7~8월) >

전력사용량(kWh)	기본요금 (원/호)	전력량요금 (원/kWh)
300 kWh 이하	910	120.0
301 ~ 450 kWh	1,600	214.6
450 kWh 초과	7,300	307.3

< 산업용 (일반용) 전기 요금 >

산업용		기본요금 (원/kWh)	전력량 요금 (원/kWh, 여름철(6 ~ 8월))
저압전력		5,550	107.7
고압 A	선택 I	6,490	116.3
	선택 II	7,470	111.5
고압 B	선택 I	6,000	115.1
	선택 II	6,900	110.4

비즈니스 계획안

- **비즈니스 목표**

  : 과거의 전력사용량 데이터를 활용해 다음 달 전력 사용량 예측과 사용 요금 안내

- **기대 효과**

  1. 소비자 입장

  - 가정용 누진세 적용 (전력 사용량 비례 요금제 x, 기준치보다 조금만 더 써도 큰 금액 청구)

    → 예상 전력 사용량과 요금을 미리 알 수 있다면 전기 사용 절약 가능

  2. 공급자 입장

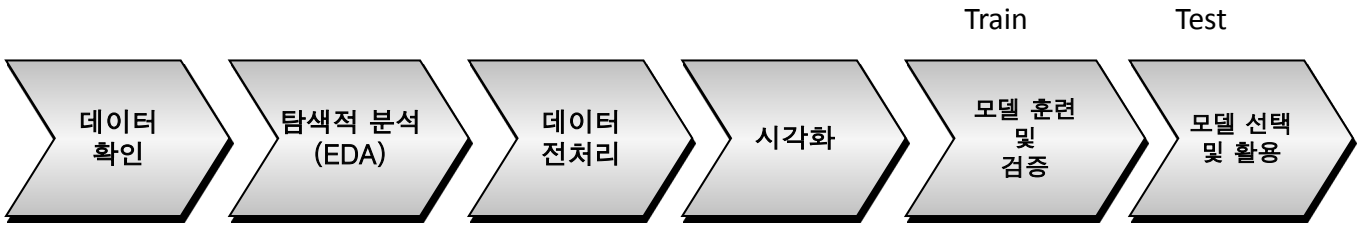
  - 다음 달 예상 전력 사용량 데이터에 맞춰 전력 생산 가능

    (블랙아웃, 전력 과잉공급 등 문제 해결 도움)

1. 비즈니스 요구사항

전력 사용량 예측 모델 생성 계획

▪ 훈련 데이터(Train)와 검증 데이터(Test)를 분석하여 전력사용량 예측 모델 생성



- 모델 평가 기준

  = SMAPE (Symmetric Mean Absolute Percentage Error)

  * SMAPE : 백분율(또는 상대적) 오차를 기반으로 하는 정확도 측정값

각 데이터 이상치 및 결측치 처리	각 데이터의 상관성 분석 및 새로운 변수 생성	전력 데이터 예측 모델 생성 및 분석

2-1. 데이터 명세

전력 데이터 명세 및 결측치 파악

- **전력 데이터 명세**

  - 훈련 데이터 (Train) : 122,400행 / 10열
  - 검증 데이터 (Test) : 10,080행 / 9열

	변수명									
	건물	날짜	전력사용량 (kWh)	기온	풍속	습도	일조	강수량	비전기냉방설비	태양광보유
훈련 데이터 (Train)	1 ~ 60번	2020.06.01 ~ 08.24 (13주)	1시간 *	1시간					각 건물의 해당 시설 보유 여부	
검증 데이터 (Test)	1 ~ 60번	2020.08.25 ~ 08.31 (1주)	X	3시간				6시간	각 건물의 해당 시설 보유 여부	

* 데이터 측정 단위 시간

전력 데이터 명세 및 결측치 파악

▪ **데이터 별 결측치 수 파악**

변수명	검증 데이터 결측치 (Test)
기온 (°C)	66.7% (6,720개 / 10,080개)
풍속 (m/s)	66.7% (6,720개 / 10,080개)
습도 (%)	66.7% (6,720개 / 10,080개)
강수량 (mm, 6시간)	83.3% (8,400개 / 10,080개)
일조 (hr, 3시간)	66.7% (6,720개 / 10,080개)
비전기냉방설비운영	77.2% (7,784개 / 10,080개)
태양광보유	83.9% (8,456개 / 10,080개)

- 날씨 변수의 경우, 훈련 데이터와 달리
  검증 데이터의 측정 단위가 3시간 / 6시간으로
  전체 데이터의 60% 이상 결측치 발생

- 검증 데이터에 결측치가 많기 때문에
  처리 방식에 따라 예측 결과에 유의미한 차이 예상

건물번호, 날짜

건물번호 (num)

- 총 61개의 건물
- 건물 번호 : 1번 ~ 60번

  (추가 데이터 61번 건물 - 청주)
- 각 건물 별로 용도가 다를 것으로 추정

날짜 (date_time)

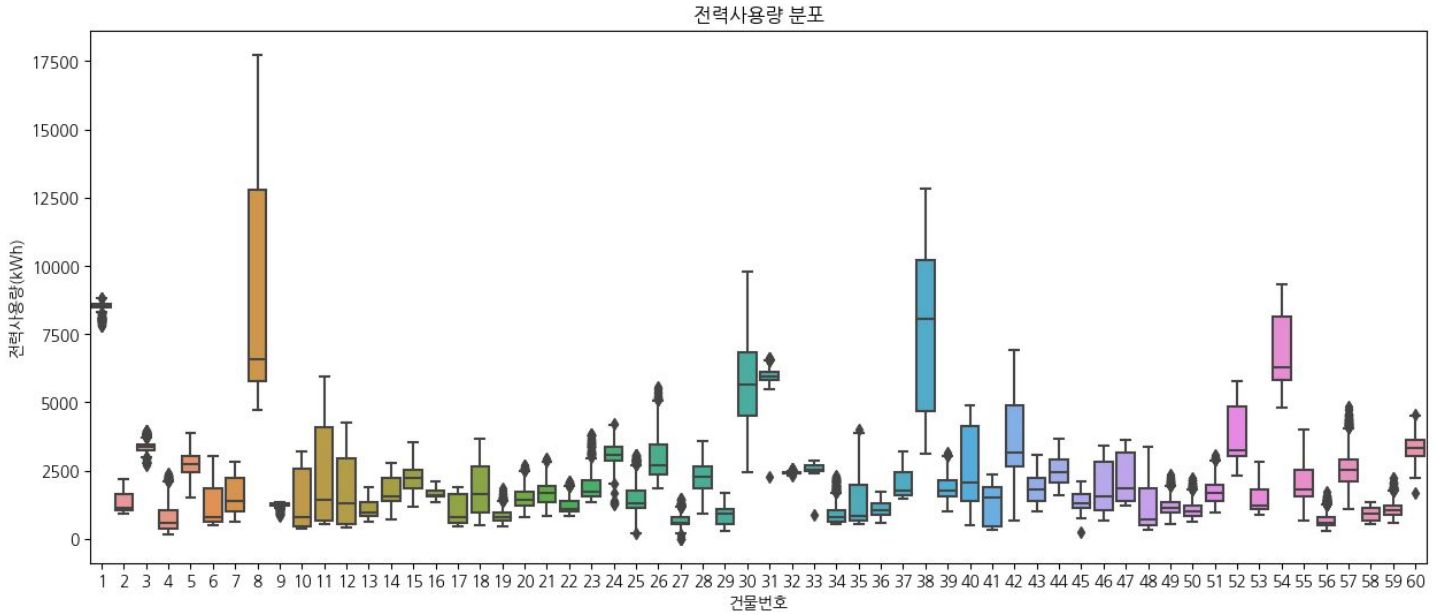
- 시계열 데이터
- 변수 구성 = 연-월-일, 시간 (1시간 단위)
- 2020년 6월 1일 00시 ~ 8월 24일 23시
- 날짜를 구성하고 있는 '월, 일, 시간'을 새로운 변수로 추출
- 일주일 단위로 날짜를 구분할 수 있도록 '요일' 파생변수 생성

전력사용량 (kWh)

- 전력(W) = 전류(A) x 전압(V)
- 전력사용량(Wh) = 전력(W) x 시간(h)

  : 1시간 동안 사용한 전력량 (1 kWh = 1000 Wh)

- 건물별 용도와 면적이 다르기 때문에 전력사용량이 다양함 (넓은 범위의 값)
- 일상 생활에서 전력사용량이 0이 되는 경우는 극히 일부 → 정전 또는 전기 점검 추정

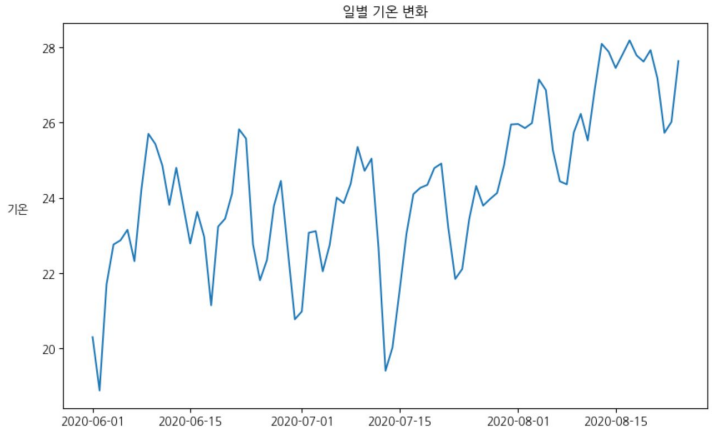


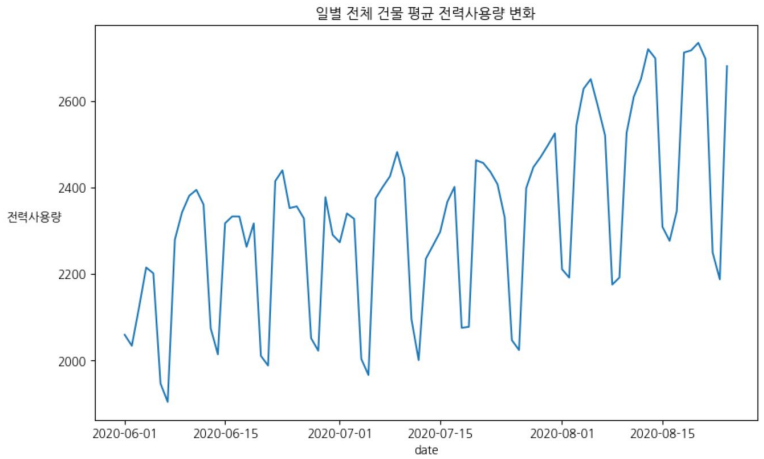
	전력사용량 (kWh)
최대값	17,739.225
중간값	1,700.352
최소값	0

기온(°C), 습도(%)

기온 (°C)

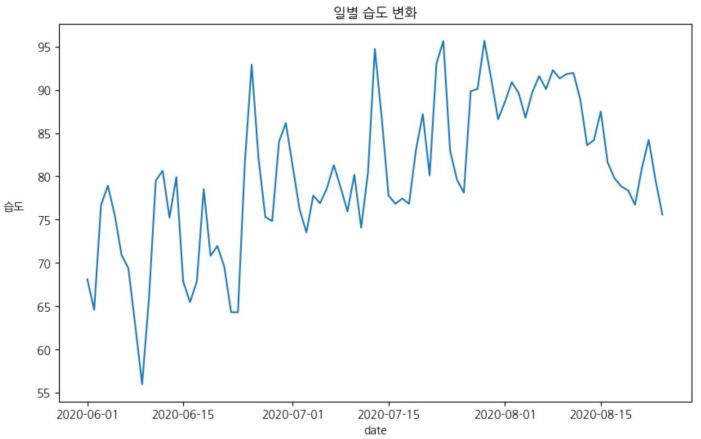
- 각 건물이 위치한 지역의 공기의 온도
- 날짜가 지날수록 기온 상승





습도 (%)

- 각 건물이 위치한 지역의 습도
- 날짜가 지날수록 습도 상승



- 기온 / 습도 상승 → 냉방설비 운영 상승
  → 전력사용량 ↑
- 따라서, 기온과 습도 모두 중요한 변수

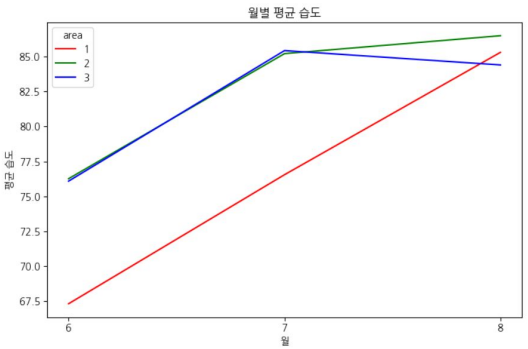
⇨ 기온과 연관있는 냉방도일 파생변수 생성

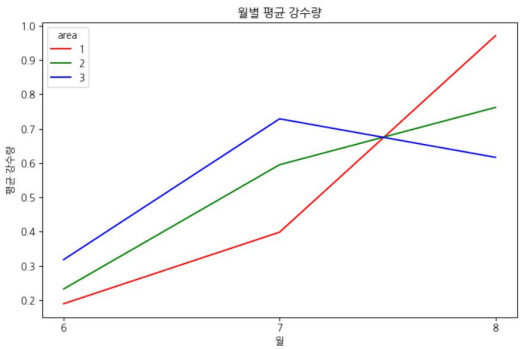
강수량(mm), 비전기 냉방 설비

강수량 (mm)

- 각 건물이 위치한 지역의 시간당 강수량
- 지역에 따른 월별 강수량 변화

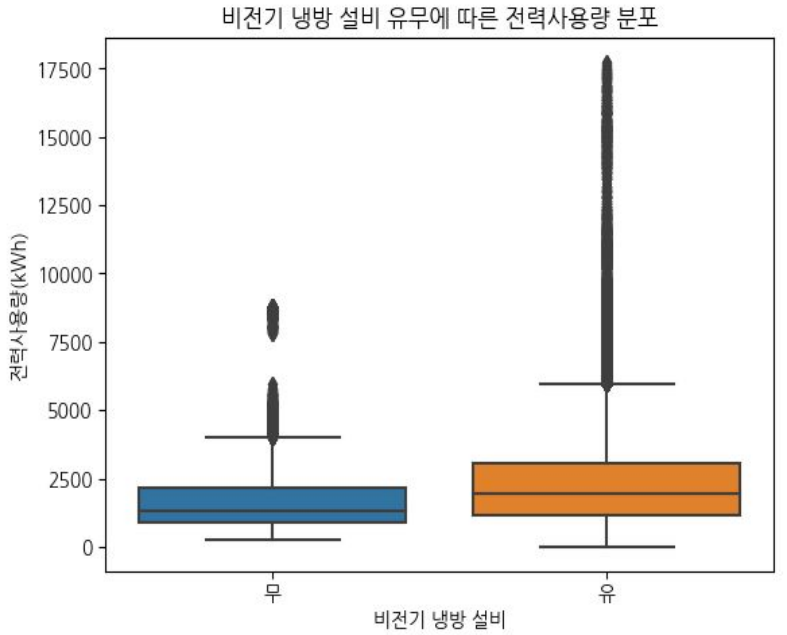
	6월	7월	8월
지역 1	0.19 mm	0.40 mm	0.97 mm
지역 2	0.23 mm	0.59 mm	0.76 mm
지역 3	0.32 mm	0.72 mm	0.62 mm





비전기 냉방 설비

- 전기가 아닌 다른 연료로 작동하는 냉방시설
- 각 건물의 비전기 냉방 설비 운영 여부
- 대체로 전력사용량이 높은 곳에서 비전기 냉방 설비를 사용

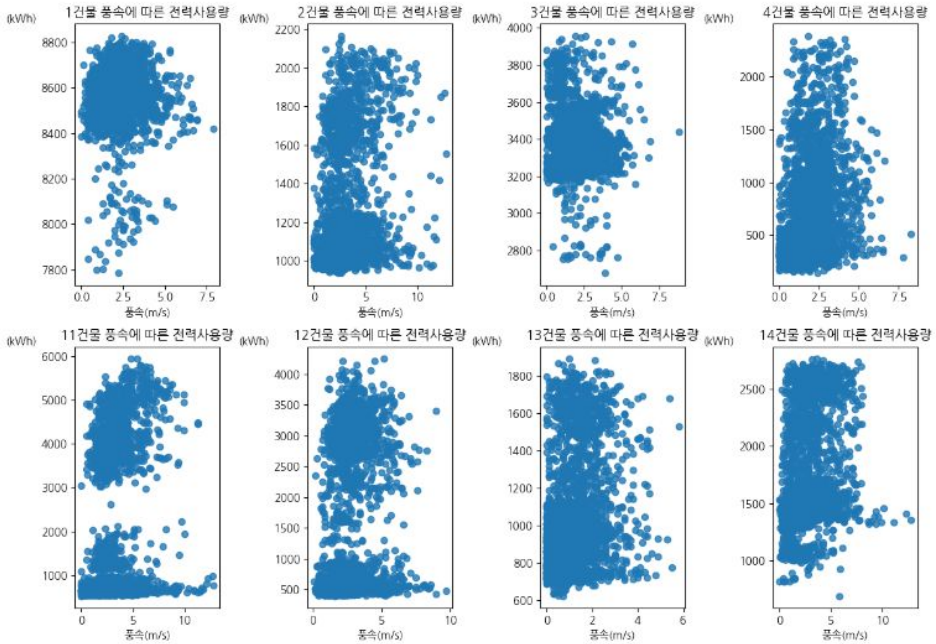


풍속(m/s)

풍속 (m/s)

- 각 건물이 위치한 지역의 풍속
- 지상 10미터 지점에서 측정
- 건물 별 전력사용량과 풍속 사이 직접적 연관성

  X
- 바람이 많이 불면 체감 온도 및 불쾌 지수 감소



체감온도

$$= -0.2442 + 0.55399Tw + 0.45535T - 0.0022Tw^2 + 0.00278Tw*T + 3.0$$

* T : 기온(℃), Tw : 습구온도

불쾌지수

$$= (9/5)T - 0.55(1 - RH)((9/5)T-26) +32$$

* T : 기온(℃), RH : 상대습도(%)

- 여름철 체감온도, 불쾌지수 공식에 풍속 포함

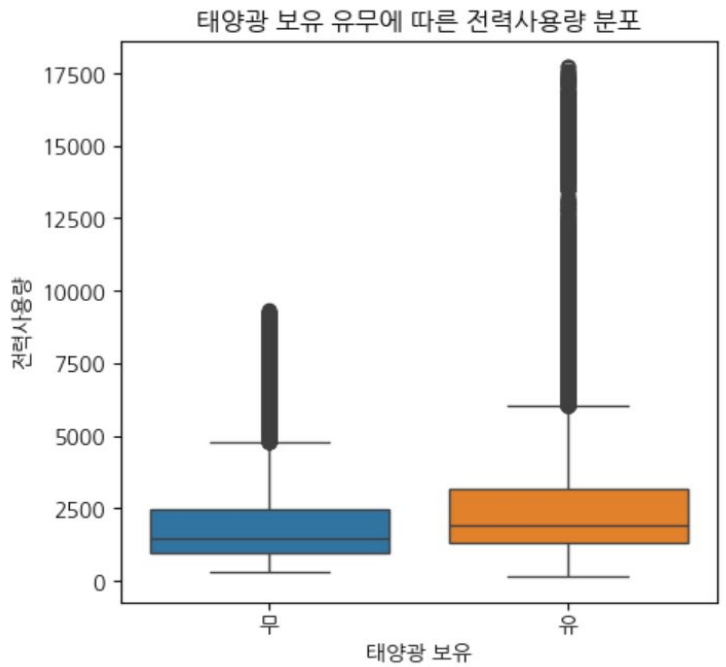
  X

➪ **여름철 풍속은 전력사용량에 큰 영향을 미치지 않음**

태양광보유, 일조(hr)

태양광 보유

- 태양광을 통해 전력 생산이 가능한 패널
- 각 건물의 태양광 패널 보유 여부
- 전력 사용량이 높은 건물의 태양광 패널 보유율이 높음



일조 (hr)

- 태양광선이 구름이나 안개의 방해없이 지표면을 비춘 시간
- 각 건물이 위치한 지역의 일조시간

태양광 발전량

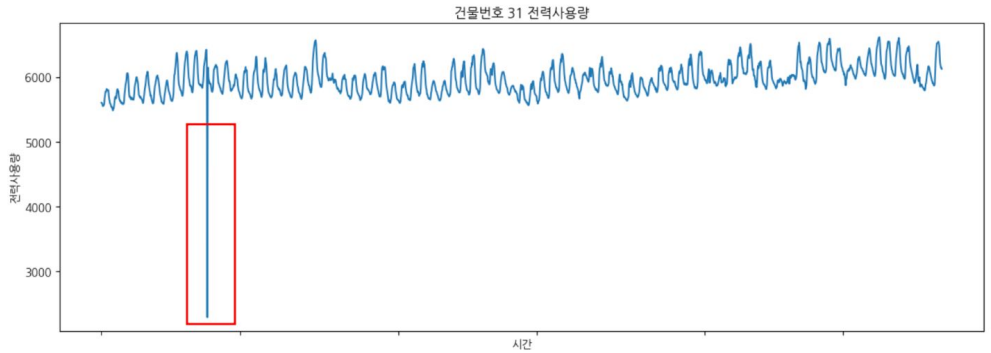
- 국내 일일 평균 일조시간 = 6시간
- 국내 평균 일일 태양광 발전 시간 = 3.6시간
- 일일 태양광 발전량
  = 태양광 설치 총 용량 x 일일 발전량 *

* 일일 발전량 = 지역별 일일 일조시간 * 0.6

이상치 처리

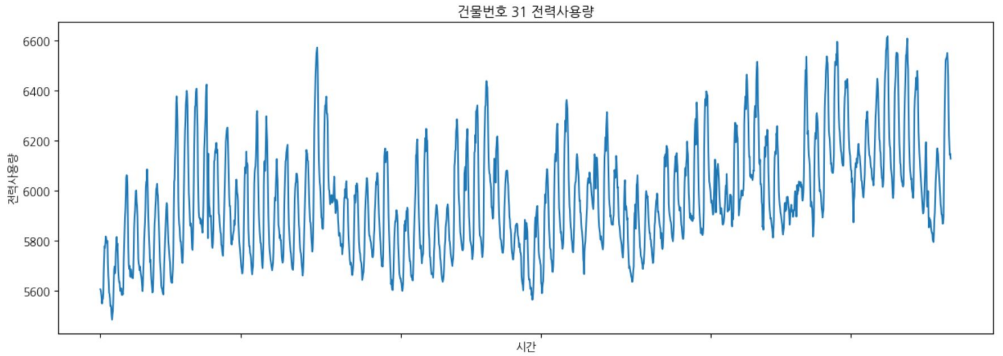
전력사용량

- 이상치 판단 기준 : 20% 이상 차이나는 전력사용량
- 처리 방법 : 이상치 해당 시간의 앞뒤 평균으로 대체
- 이상치 보유 건물 : 건물 9 25 27 31 33 36 45 55 60
- 예시) 31번 건물



이상치 처리 전

이상치 처리 후



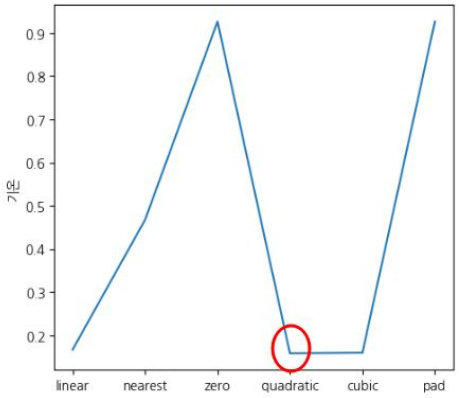
결측치 처리 (1/2)

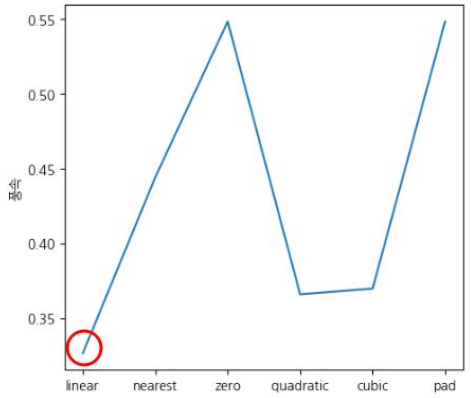
기온, 풍속, 습도, 강수량

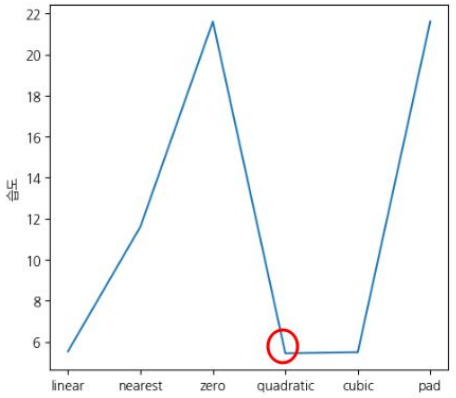
- 데이터 측정 단위 : 훈련 데이터 1시간 / 검증 데이터 3시간 (강수량 : 6시간)

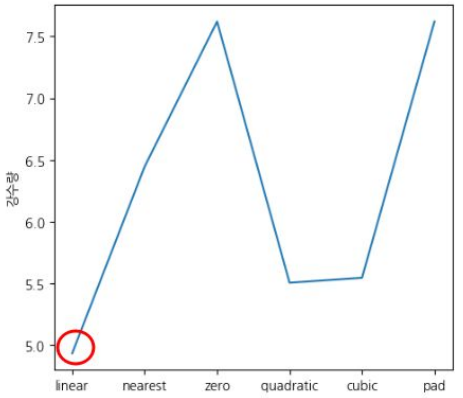
  → 검증 데이터 매 단위 당 2시간씩 결측치 발생 (강수량 : 5시간씩)

- **결측치 처리 방법(보간법)**
  - 학습데이터에 똑같은 단위로 임의의 결측치 생성
  - 가장 오차가 적은 보간법을 선택해 결측치 처리
    - **기온 : quadratic**
    - **풍속 : linear**
    - **습도 : quadratic**
    - **강수량 : linear**









결측치 처리 (2/2)

비전기냉방설비, 태양광보유

- 각 건물 번호에 맞게 훈련 데이터의 값을 검증 데이터의 결측치에 대입

일조

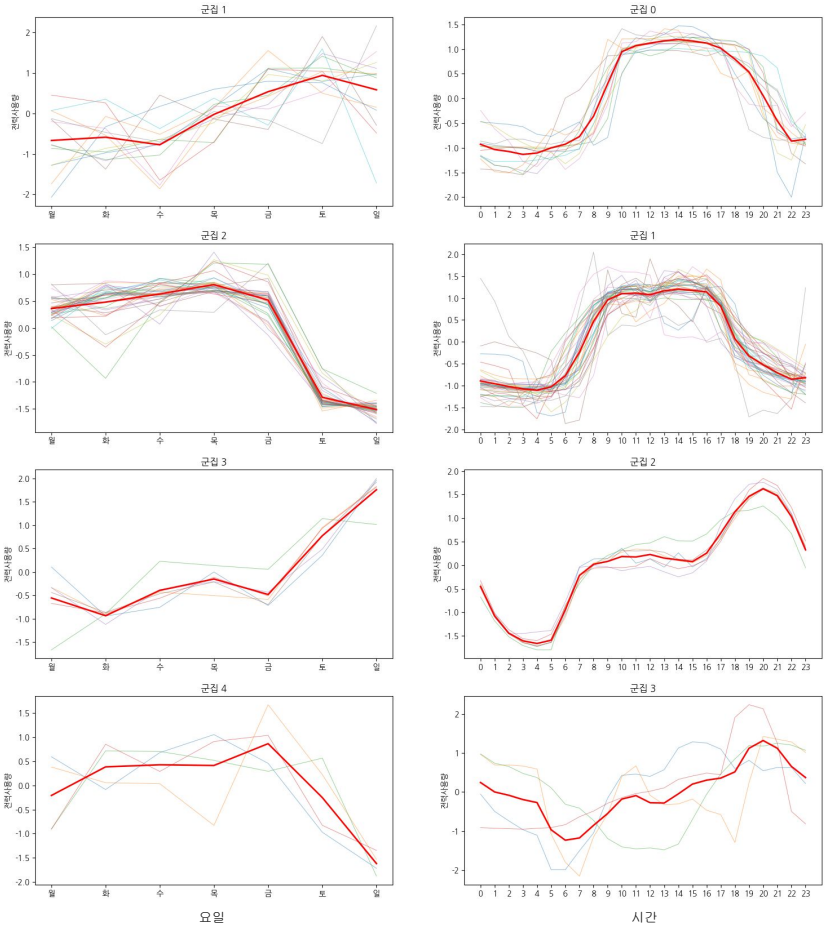
- 데이터 측정 단위 : 훈련 데이터 1시간 / 검증 데이터 3시간

  → 검증 데이터 매 단위 당 2시간씩 결측치 발생
- 훈련 데이터 0.0 ~ 1.0 / 검증 데이터 0.0 ~ 3.0 (0.1 단위 수치형 변수)
- 일조시간 시간당 최대 1.0 → 검증 데이터 표준화

- **결측치 처리 방법**
  - 결측치 앞뒤 시간의 데이터를 활용하는 것이 가장 효율적
  - 시계열 데이터 **보간법**의 가장 기본값(선형 비례 방식)으로 결측치 처리

건물 군집화

- 가정) 건물의 용도에 따라 평균적인 전력 사용량 분포가 비슷할 것
- 각 건물 별 요일과 시간에 따른 평균 전력사용량을 기준으로 군집화 → **총 4개의 군집**



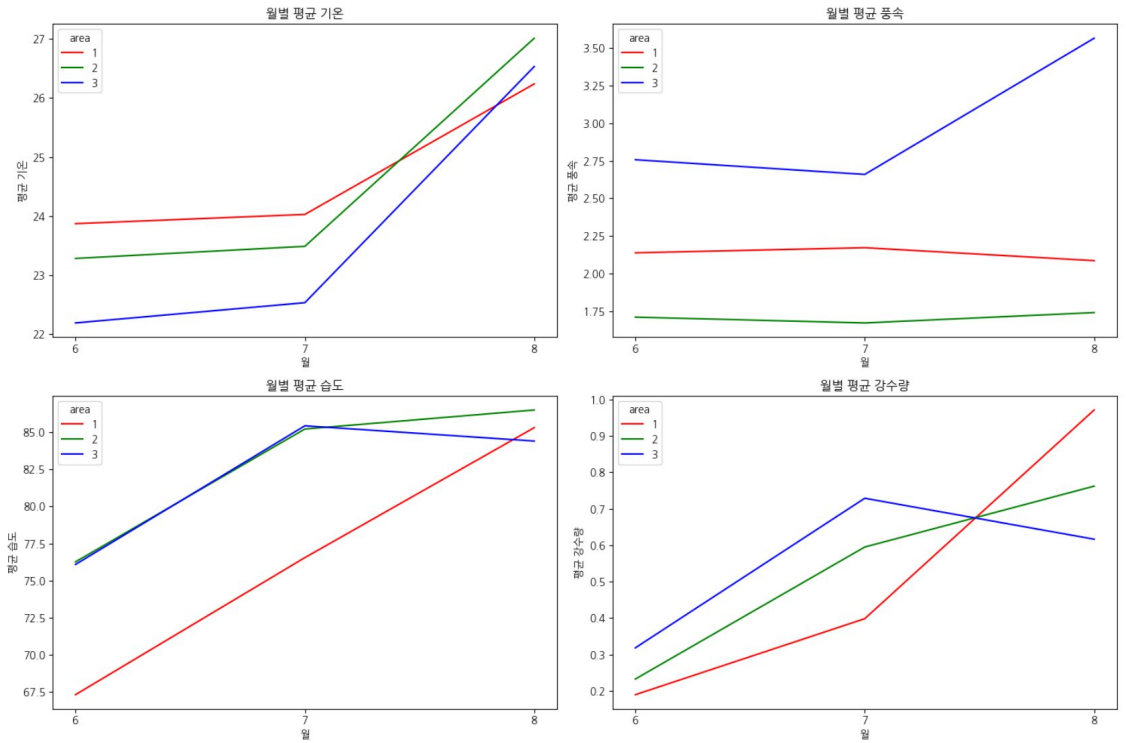
< 군집 별 전력 사용 시간대 및 요일 >

	주 사용 시간대	주 사용 요일	용도
군집 1	주간 (9~20시)	평일, 주말	상가 건물
군집 2	주간 (7~18시)	평일	사무실, 학교 등
군집 3	야간 (17시 이후)	주말	유흥시설, 주점 등
군집 4	야간 (18 ~ 00시)	평일	기타

지역 군집화 (1/2)

지역 군집화

- 가정) 날씨가 같으면 같은 지역일 것
- 기온, 습도, 강수량, 일조, 강수량 모두 같은 건물이 많았음 → **총 3개의 군집**
- 각 지역에 속한 건물 개수 (지역 1 : 19개, 지역 2 : 15개, 지역 3 : 26개)



< 지역 별 평균 날씨 >

	지역 1	지역 2	지역 3
평균 풍속(m/s)	2.1	1.7	2.9
평균 기온(℃)	24.579	24.410	23.540
평균 습도(%)	75.758	84.424	81.850
평균 일조	0.187	0.216	0.243
평균 강수량(mm)	0.486	0.514	0.552

- 습도 / 강수량 : 연관성이 높음
- 풍속 : 관련 없음
- 지역별로 날씨 추세 다름

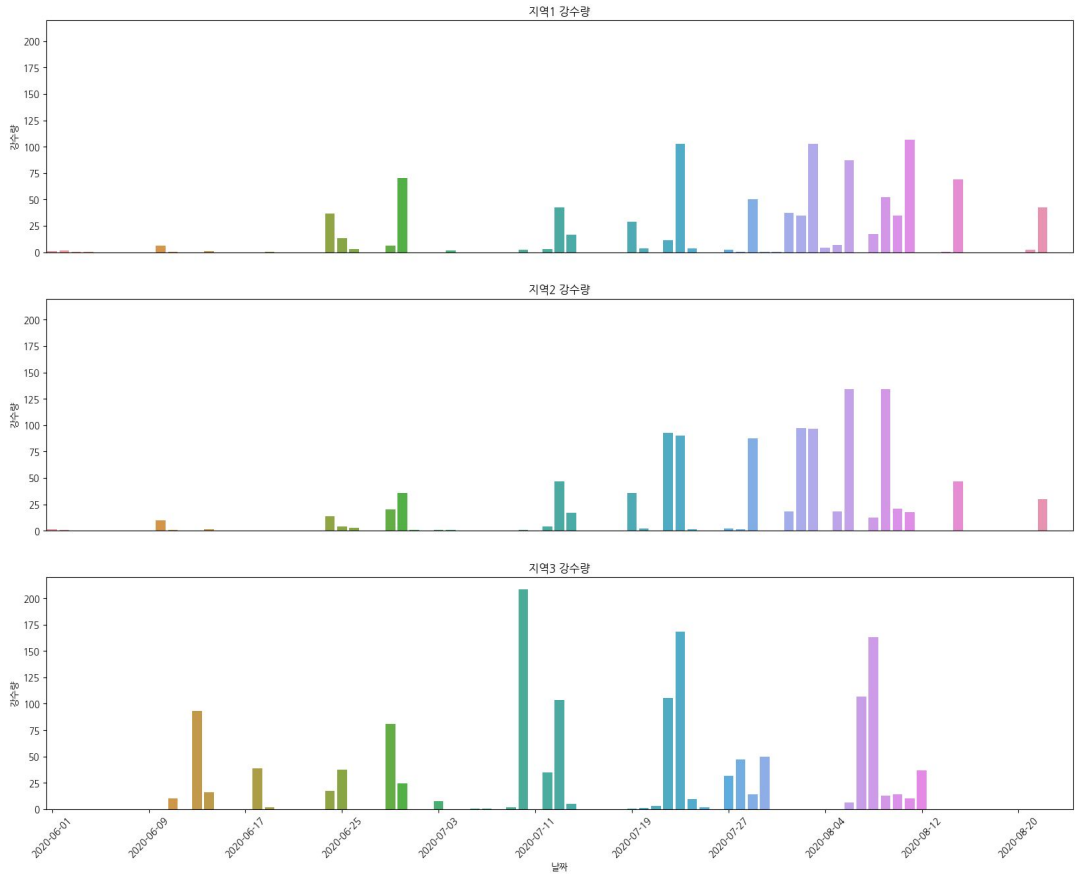
지역 군집화 (2/2)

지역 추정

- **지역 1**
  - 기온 높음 / 습도, 강수량, 일조 낮음
  - 구름으로 인해 흐린 날이 많았을 것
  - 비가 타 지역에 비해 많이 내리지 않음

- **지역 2**
  - 습도 높음 / 강수량, 일조, 기온 낮지 않음
  - 장마 기간에 많이 비가 자주 왔을 것

- **지역 3**
  - 강수량, 일조, 습도 높음 / 기온 낮음
  - 비가 한번에 많이 내렸을 것



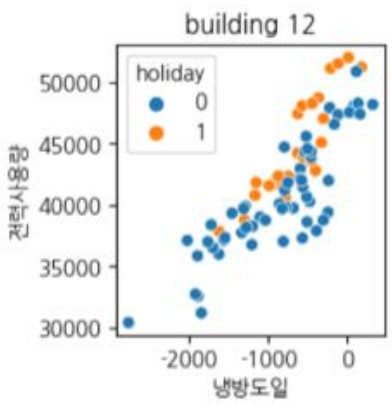
냉방도일

냉방도일

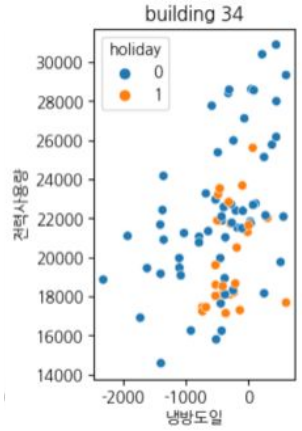
- 일평균 기온 24℃ 이상의 값에 대한 1년간의 적산 온도를 도일수(degree day)를 단위로 나타낸 것
- 도일(degree day) : 외부 기온과 실내 기온의 차와 이에 따라 소요되는 연료 소비를 고려한 기후 수치
  - 도일이 크면 연료 소비량이 많아짐
  - 실내온도가 같아도 외부온도가 다르기 때문에 지역 또는 건물마다 다름
- 기온보다 **냉방 수요**에 대한 설명력이 높음

건물유형별 휴일에 따른 냉방도일 변화

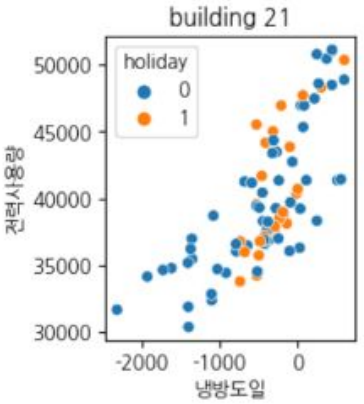
유형 1



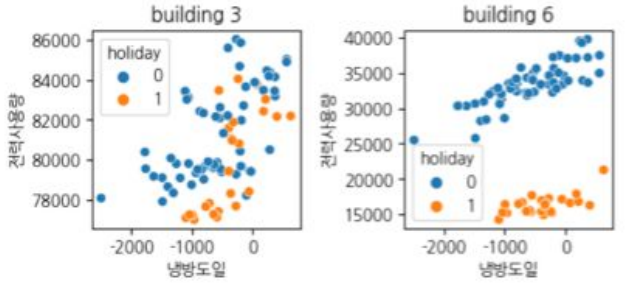
유형 2



유형 3



유형 4



➪ 일평균 냉방도일과 일평균 전력소요 유의미한 관계 존재
(유형 4 제외)

체감온도, 불쾌지수

체감온도

- 인간이 느끼는 더위/추위를 수량적으로 나타낸 것
- 구분 : 여름철(5~9월) / 겨울철(10~익년 4월)

불쾌지수

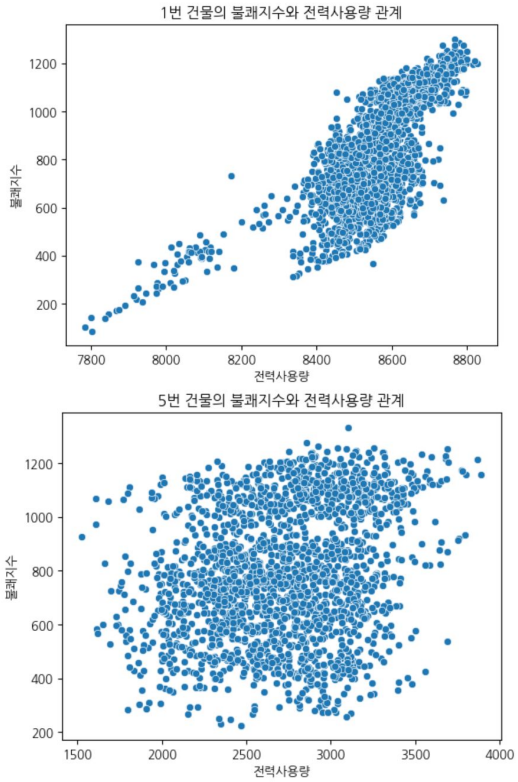
- 날씨에 따라 인간이 느끼는 불쾌감의 정도를 기온과 습도를 조합해 나타낸 수치

단계	지수 범위	설명 및 주의사항
매우 높음	80 이상	전원 불쾌감을 느낌
높음	75 ~ 80 미만	50% 정도 불쾌감을 느낌
보통	68 ~ 75 미만	불쾌감을 나타내기 시작함
낮음	68 미만	전원 쾌적함을 느낌

< 불쾌지수 단계 별 정의 >

불쾌지수 분석

- 불쾌지수 80 이하 데이터 187개 (약 0.15%)
- 불쾌지수가 높으면 전력사용량이 높을까?



➪ 건물 별 관계 상이

모델링 (1/3)

단변량 모델링

- 사용 변수 : 전력사용량 (kWh)
- 시계열 모델 선택
- 모델 학습 데이터 분할
  - 시계열 데이터 특성 고려, 앞 80% = 학습 데이터 / 뒤 20% = 검증 데이터
- 건물 별 최적의 차분 수 탐색 후 전력 사용량 예측

Auto ARIMA	Prophet	ARIMA	ARIMA + LSTM
- 추세 반영 o - 계절성 반영 x - 계절성 반영 예측 모델 사용 필요	- 추세, 계절성 등 변동 요인 반영 - Auto ARIMA의 계절성 반영 여부 단점 해결 가능	- 건물 별 AR(q)값과 MA(p) 값을 찾아 차수를 적용 - 차수 (5, 1, 1) 고정한 것이 성능이 더 우수하였음	- ARIMA를 사용하여 구한 예측 결과값(test)을 활용하여 단변량 LSTM 적용

모델링 (2/3)

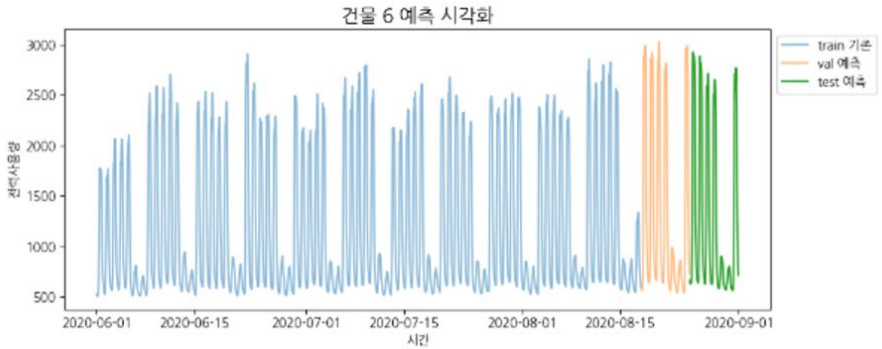
다변량 모델링

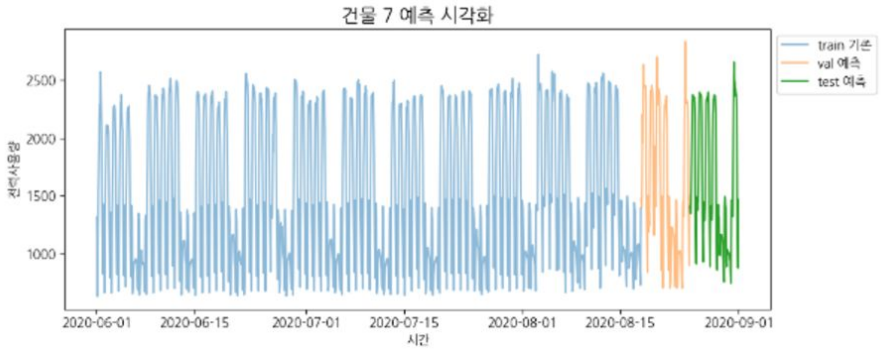
- 사용 변수 : 시간(date_time) 제외 모든 변수
- 회귀 모델 선택
- 하이퍼 파라미터 탐색
  - 각 건물에 대해 그리드 서치 (랜덤 7 : 3 - SMAPE 성능 기준)
- 모델 학습 데이터 분할
  - 시간 마지막 일주일 = 검증 데이터 / 나머지 = 학습 데이터

XGBoost	건물 별 최적 모델	XGBoost + 최적 모델
– 하이퍼 파라미터 3개 선택 – 전력 사용량 데이터(Public) 학습 성능 우수	- 각 건물 별 최적 모델 탐색 - XGB, LGBM, CatBoost, Extra Tree - 전체 변수 데이터 (Private) 학습 성능 우수 모델 선택	– 두 모델의 최종 결과값의 산술 평균을 최종 모델 예측값으로 사용

모델링 (3/3)

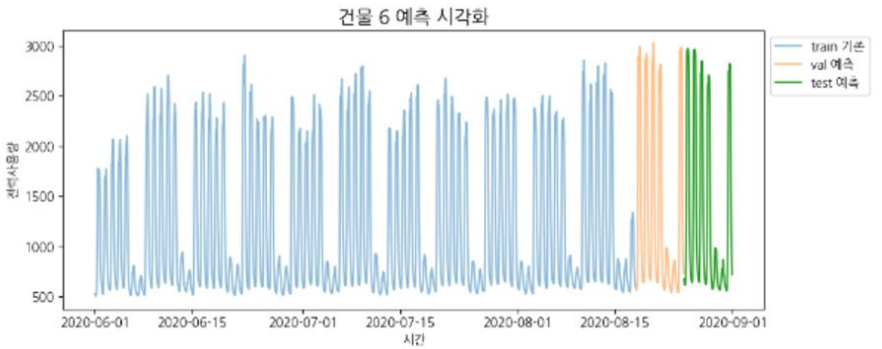
XGboost

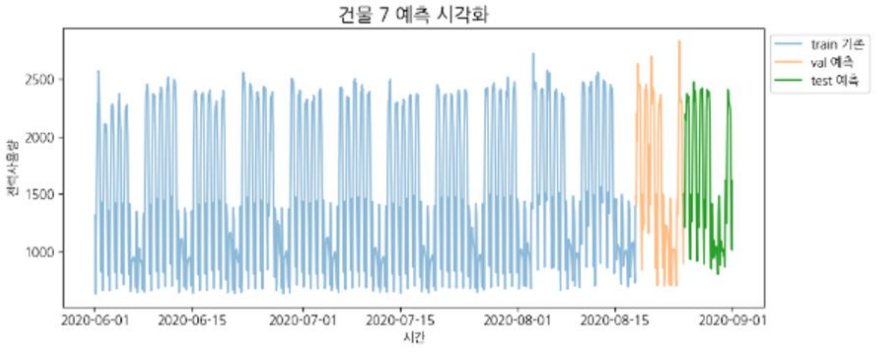




건물별 최적 모델

건물 6번 : lgbm / 건물 7번 : catboost





모델 선택

▪ **모델 성능 비교**

변수	사용 모델	점수 *	변수	사용 모델	점수
단변량	ARIMA	8.5	다변량	XGBoost	6.48
	Auto ARMIA	30.26		건물 별 최적모델	6.005
	Prophet	12.74			
	ARIMA + LSTM	8.59		**XGB + 최적모델**	**6.007**

* 점수 = DACON 제출 Private 점수 기준

➪ XGB와 건물 별 최적모델의 **앙상블 모델** 선택

최종 성능 = 6.007

End Of Documents