2023 전력사용량 예측 AI 경진대회

똥안

Contents

CV strategy

부분적인 월간 데이터

데이터셋 작음

안정적인 CV – Public Score correlation

일반화 가능한 변수선택

CV strategy

안정적인 CV – Public Score correlation

일반화 가능한 변수선택

안정적인 CV – Public Score correlation

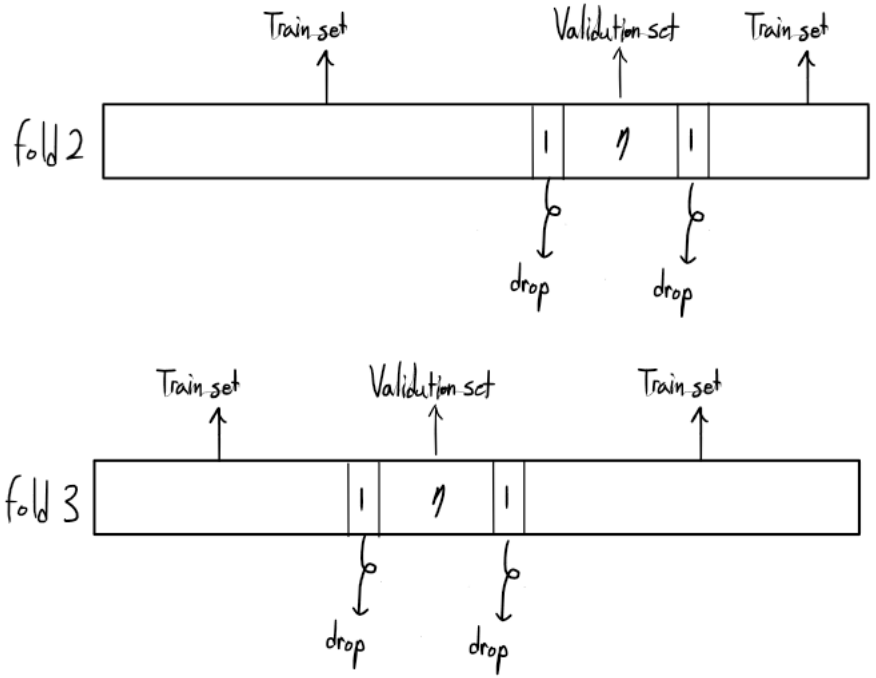
충분한 fold 확보

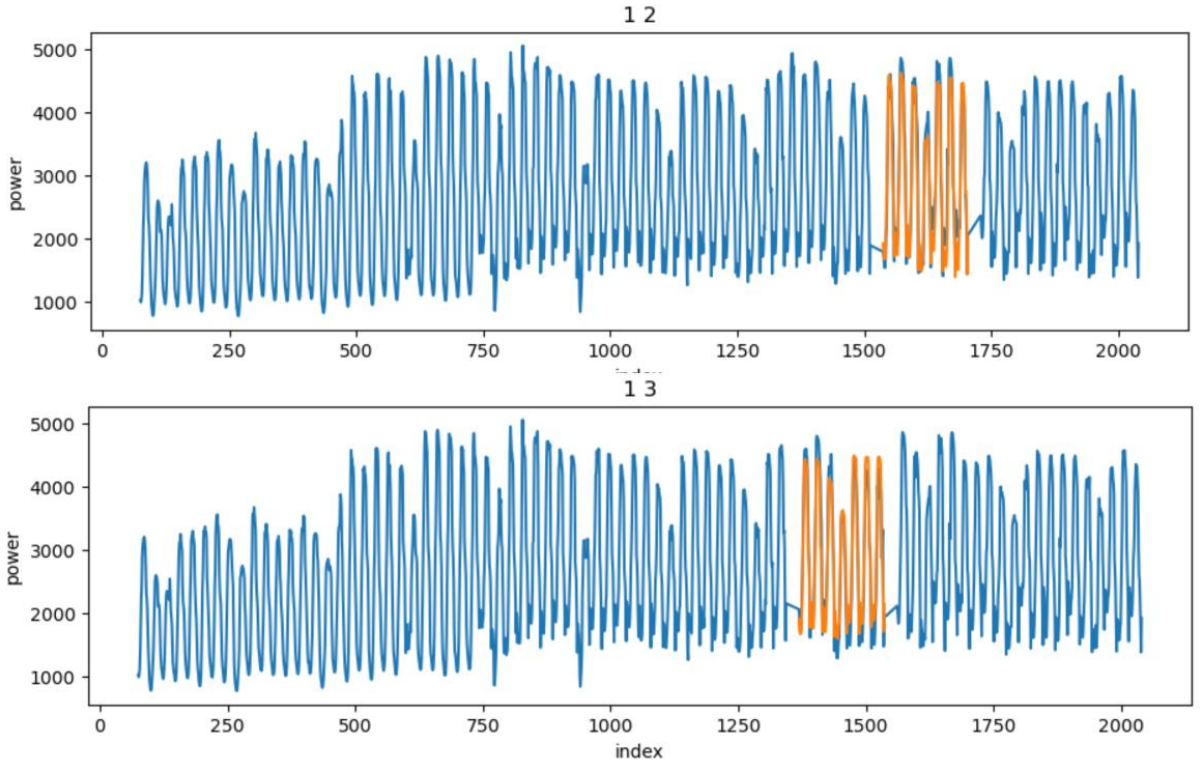
코드 작성, 유지, 보수 용이성 고려

→ Validation set

7일 간격으로 앞, 뒤 1일 드랍

→ 11fold 확보





CV strategy

안정적인 CV – Public Score correlation

일반화 가능한 변수선택

Target Statistic (TS) 변수 사용 시 빠르게 overfitting

특정 범주형 변수와 TS 함께 사용 시 CV – Public Score correlation 급격히 하락

→

범주형 변수의 정보를 TS에 포함시키면서 그와 동질적인 범주형 변수들을 제거하여 CV – Public Score correlation을 높이는 전략

마감 2주 전까지 TS를 제외하고, 필수적인 범주형 변수, 기본 변수만 사용

모든 변수는 CV, Public Score 최소 1% 개선 후에 추가

-1% ~ 1% 내에서 변수를 제거할 수 있다면 혹은 범주형 변수의 차원을 줄일 수 있다면 제거

CV strategy

안정적인 CV – Public Score correlation

일반화 가능한 변수선택

		dhc + weekend	whc
One-Hot Encoding 적용시 차원의 수		12	**5**
전체 CV 평균		4.772	**4.759**
submission_1	Public Score	5.295	**5.270**
	Private Score	**6.093**	6.105
submission_2	Public Score	5.344	**5.279**
	Private Score	6.114	**6.109**

CV – Public Score correlation을 높이는 전략을 바탕으로 dhc + weekend 조합 대신 whc선택

비슷한 논리로 여타 범주형 변수들(month, university_vacation, school_vacation, dhc, weekend)를 제거

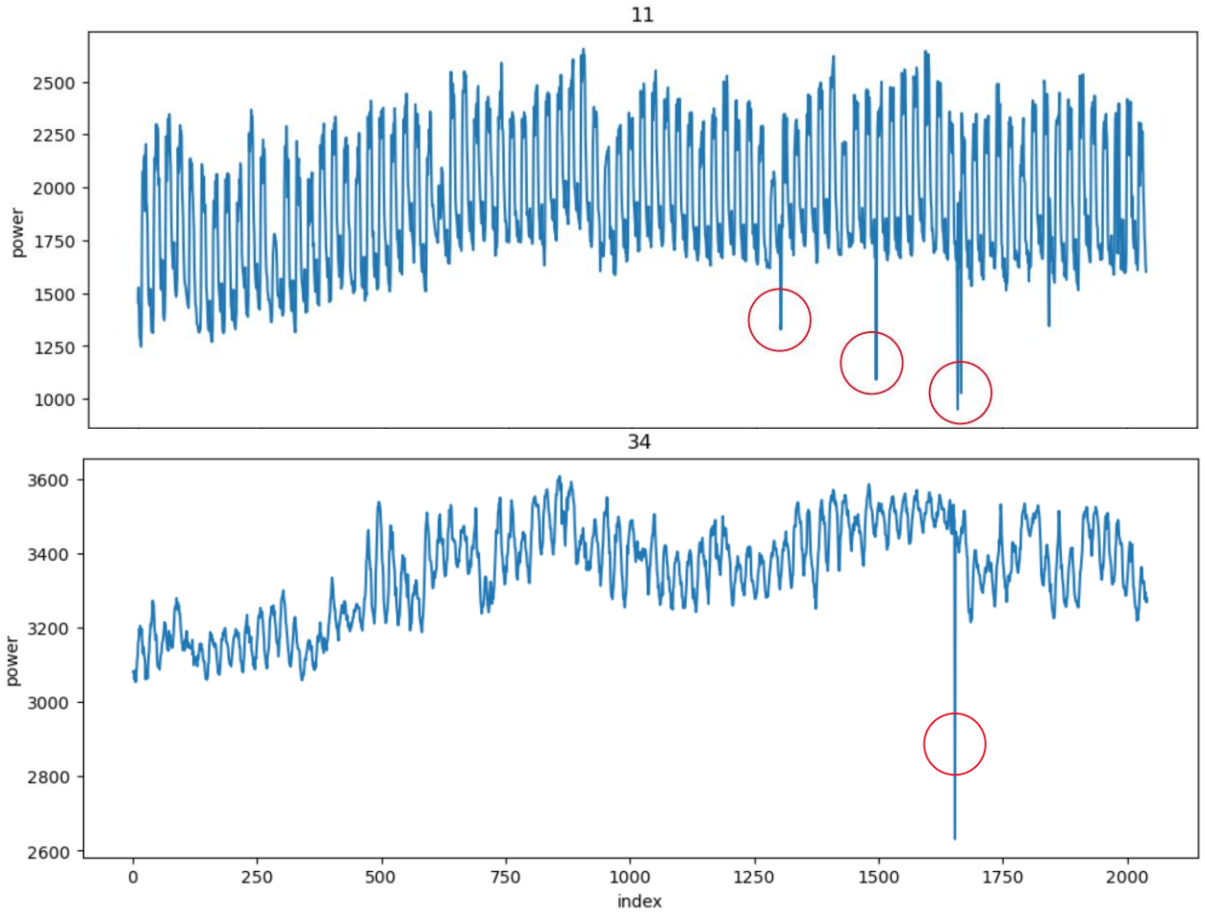
1개의 범주형 변수(whc)만을 사용

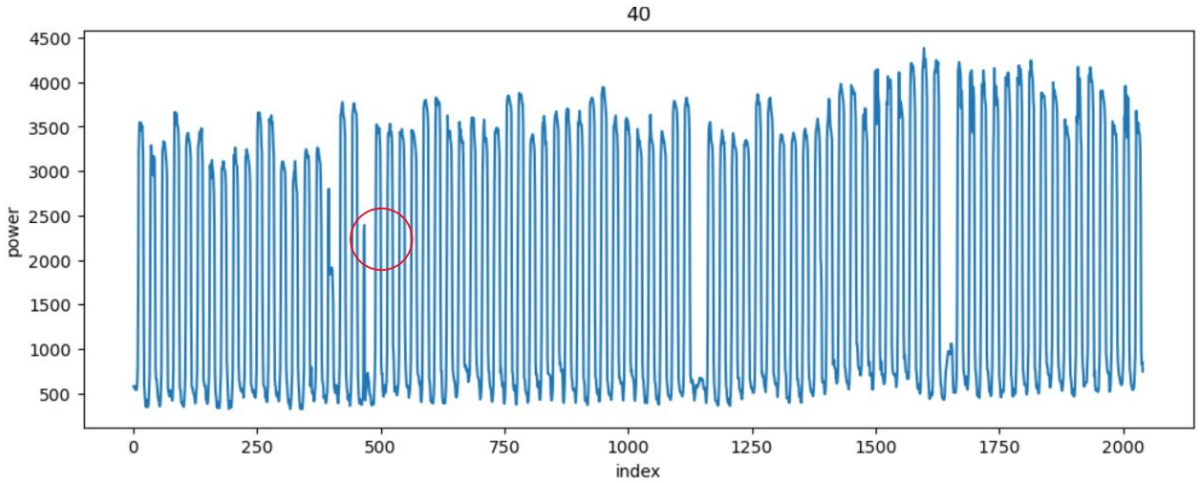
Remove outliers

안정적인 CV – Public Score correlation

일반화 가능한 변수선택

Type1. 긴 꼬리



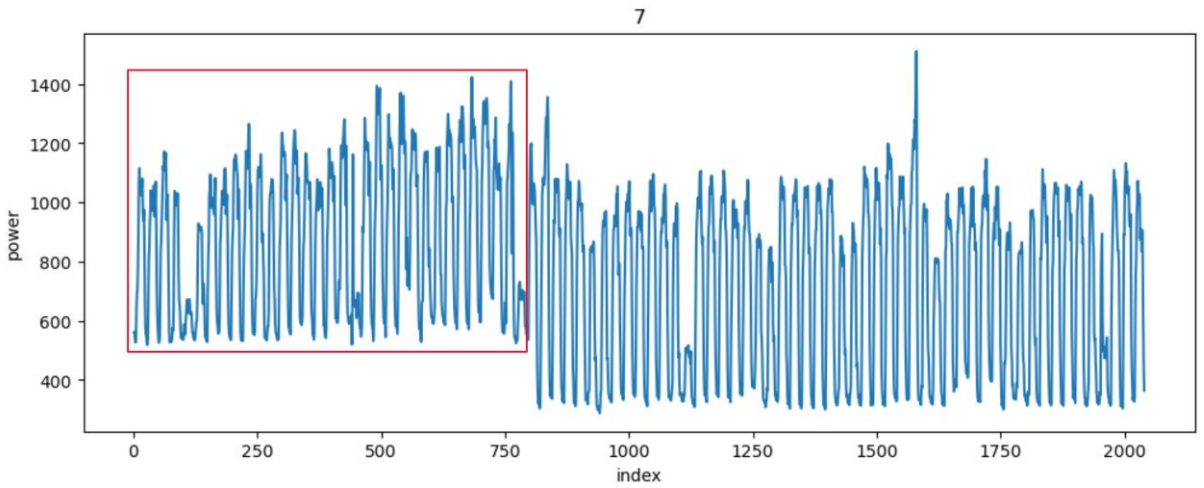


Remove outliers

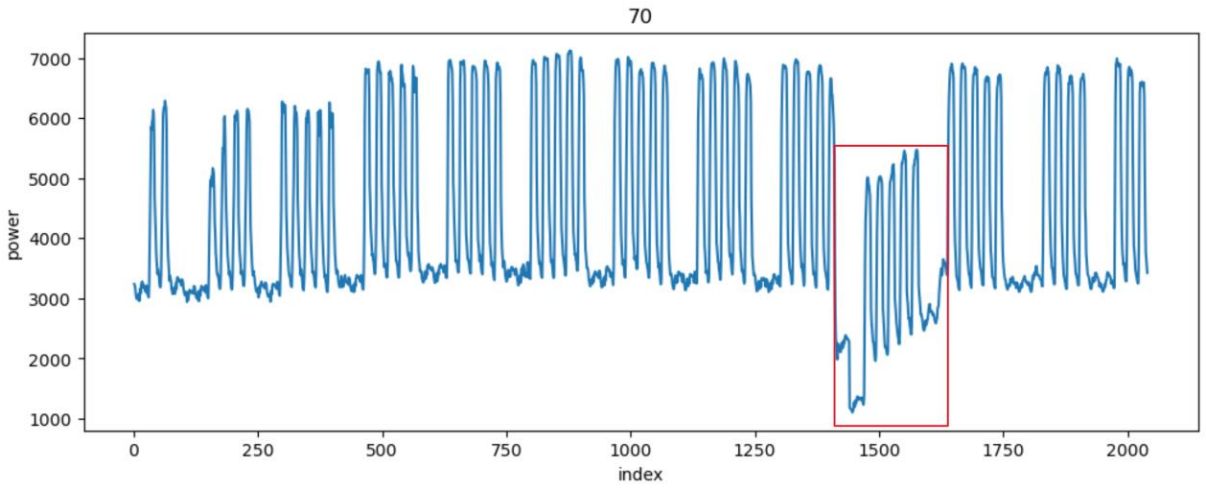
안정적인 CV – Public Score correlation

일반화 가능한 변수선택

Type2. 범위 이동



Type3. 휴가

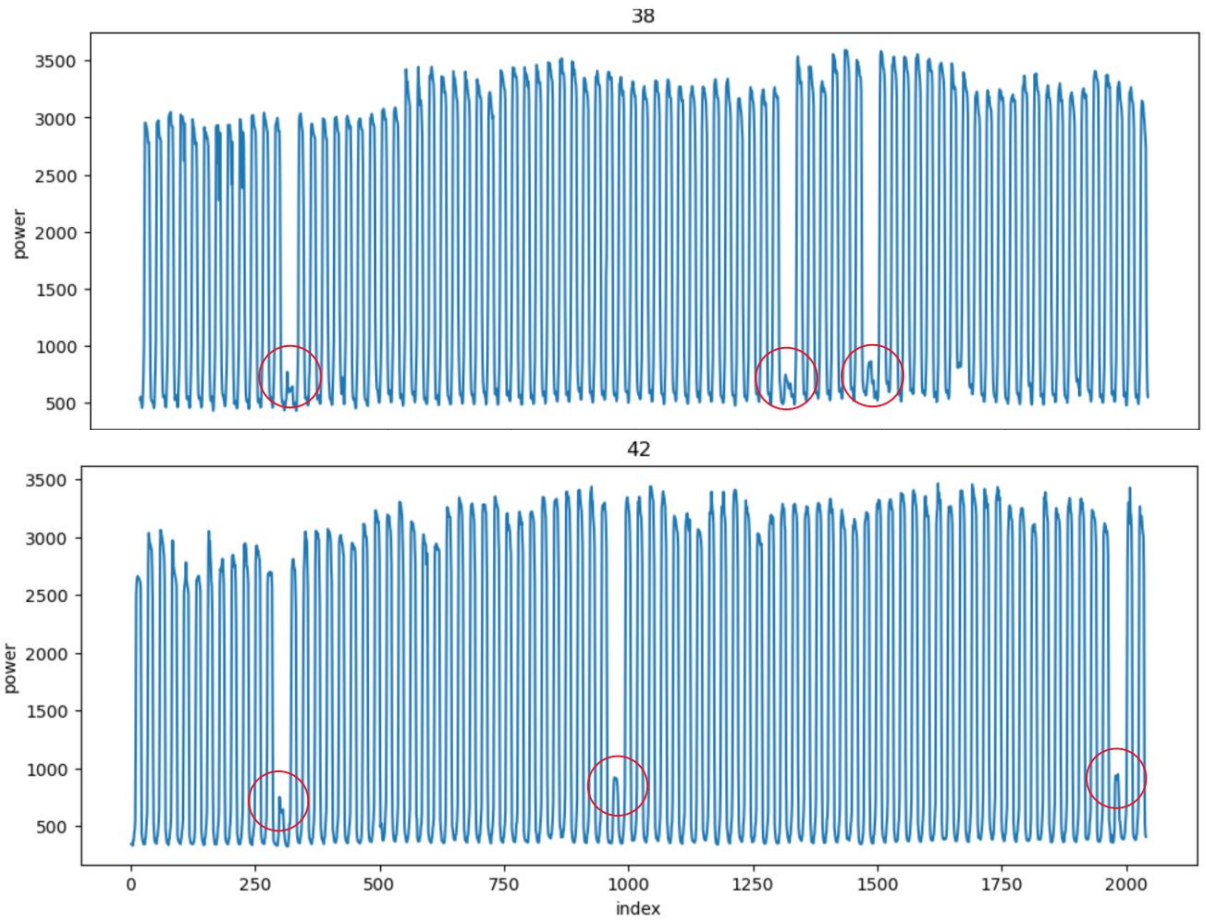


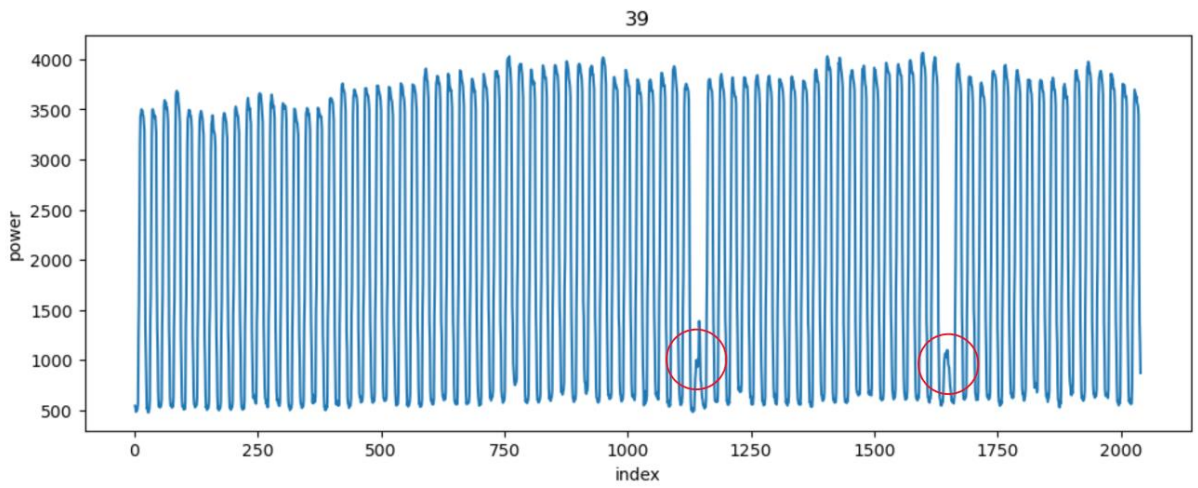
Remove outliers

안정적인 CV – Public Score correlation

일반화 가능한 변수선택

Type4. 불규칙적 휴무일



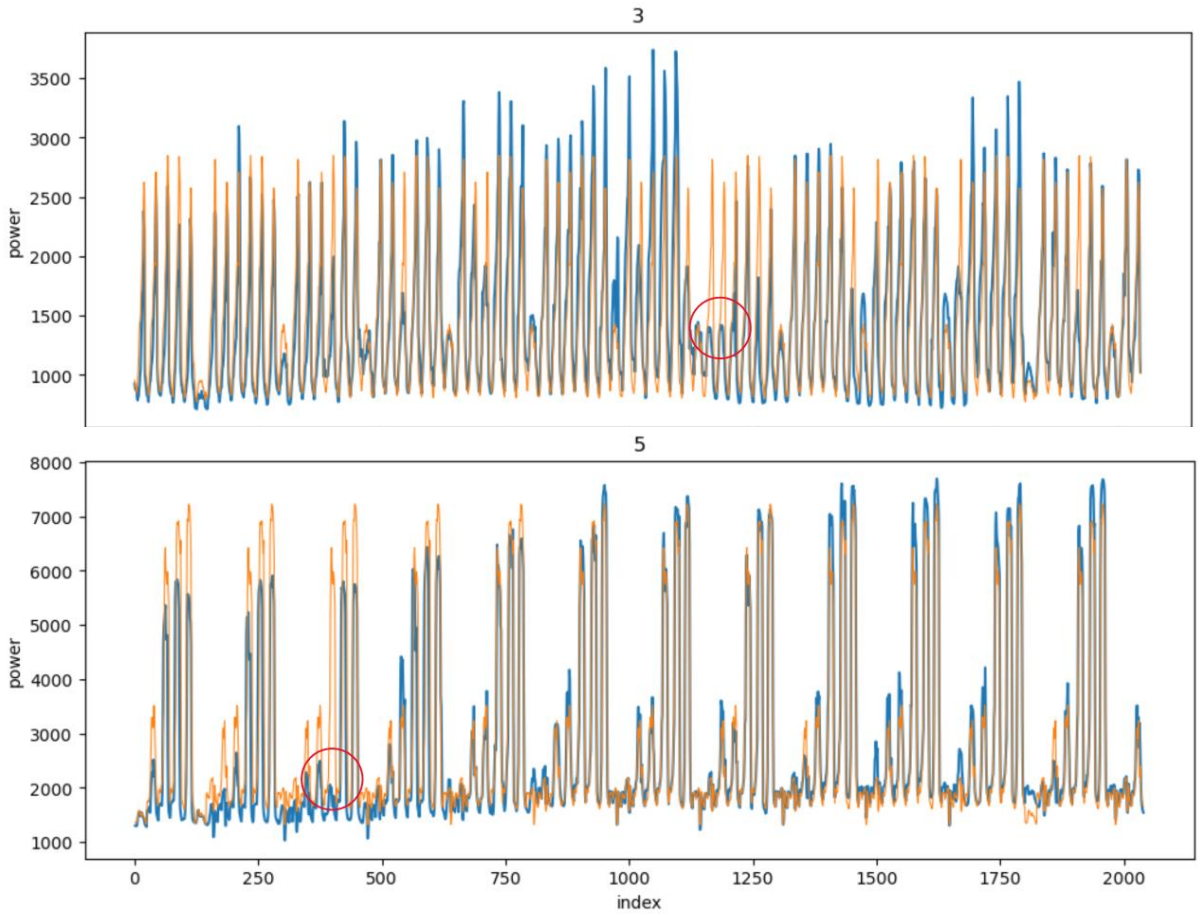


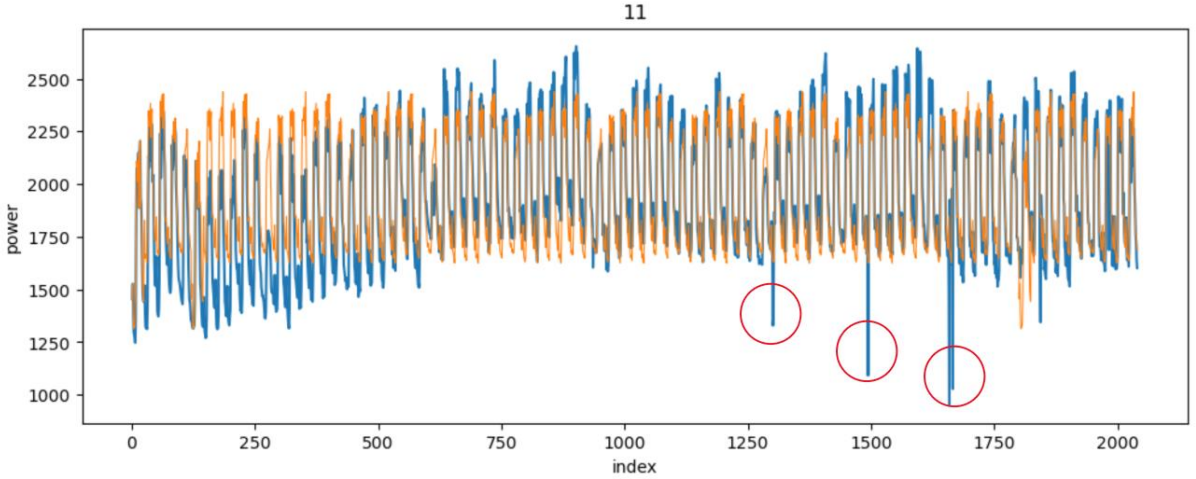
Remove outliers

안정적인 CV – Public Score correlation

일반화 가능한 변수선택

탐지방법1. dhc, hour별 전력량 평균 이용



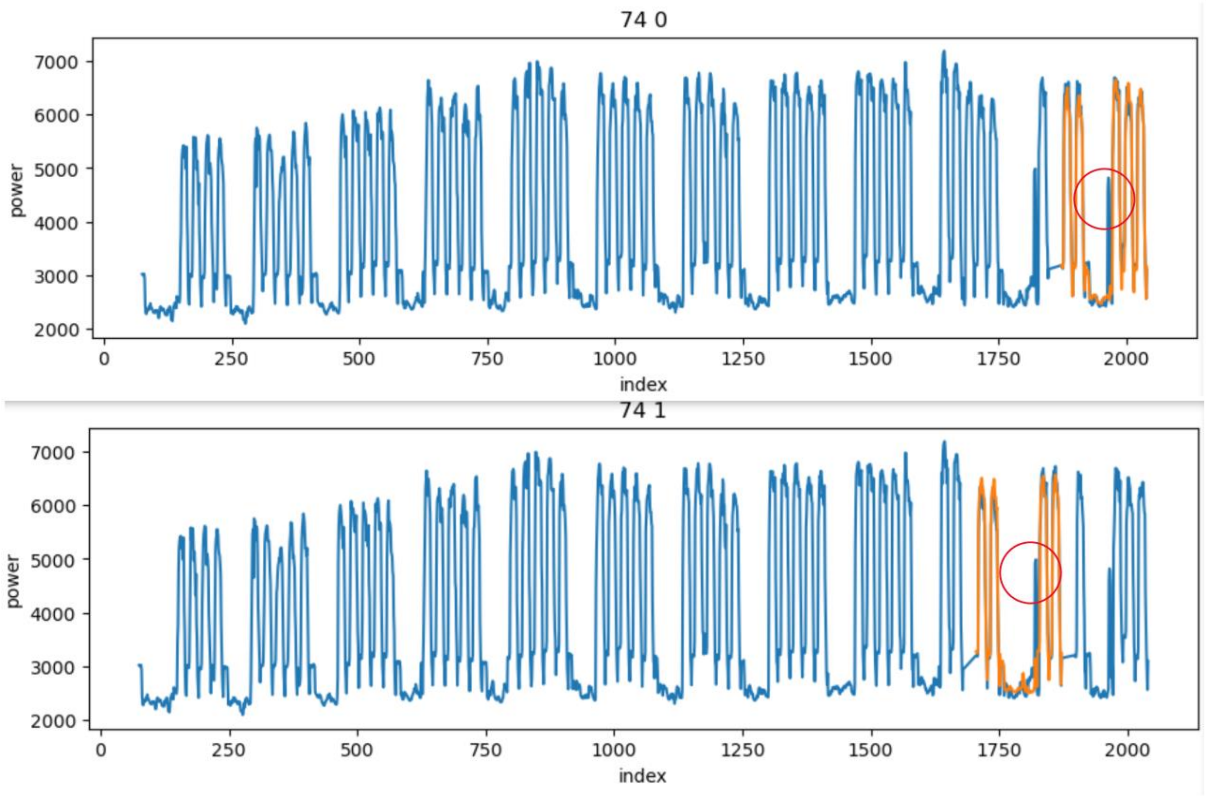


Remove outliers

안정적인 CV – Public Score correlation

일반화 가능한 변수선택

탐지방법2. Validation set 예측값 모니터링



Feature

안정적인 CV – Public Score correlation

일반화 가능한 변수선택

공통변수

whc,

cos_hour, sin_hour,

temperature_squared, THI, humidity_squared,

temperature_squared_mean, THI_mean,

power_log1p_stdd_mean,

power_log1p_stdd_shift

EE_TRAIN_0.ipynb

공통변수,

power_log1p_stdd_cumweek_mean_shift,

power_log1p_stdd_thisweek_mean_shift

EE_TRAIN_1.ipynb

공통변수,

power_log1p_stdd_cumweek_mean_shift

EE_TRAIN_2.ipynb

공통변수

Feature

안정적인 CV – Public Score correlation

일반화 가능한 변수선택

범주형 변수

whc

One-Hot Encoding을 통해 5개로 변환

	dhc	weekend	whc
0	Monday	Workday	Workday
1	Tuesday	Weekend1	Weekend1
2	Wednesday	Weekend2	Weekend2
3	Thursday		Holiday
4	Friday		Closed
5	Saturday		-
6	Sunday		-
7	Holiday		-
8	Closed		-

Feature

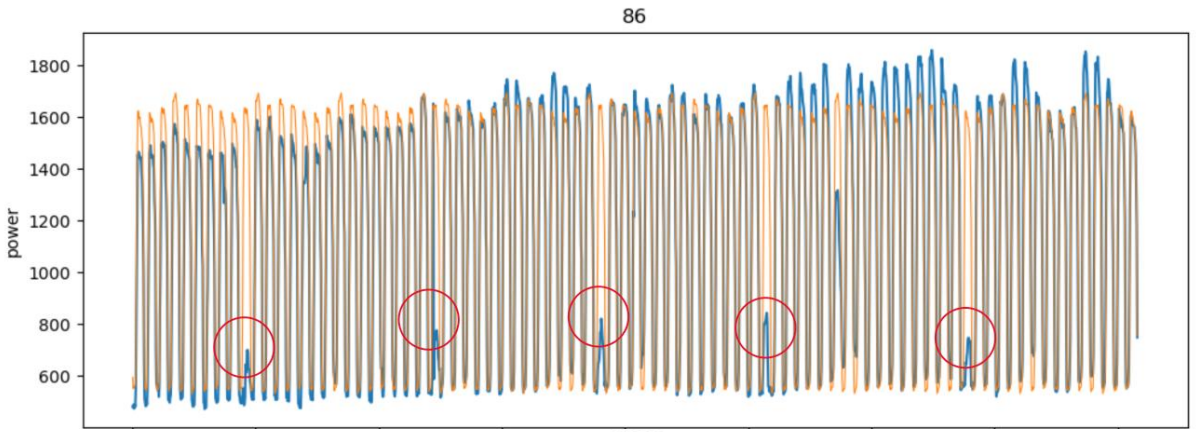
안정적인 CV – Public Score correlation

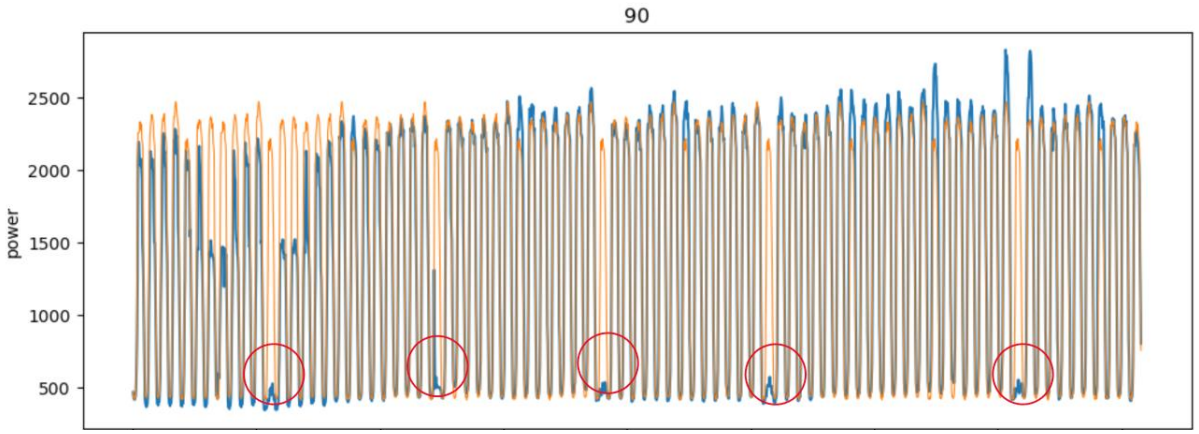
일반화 가능한 변수선택

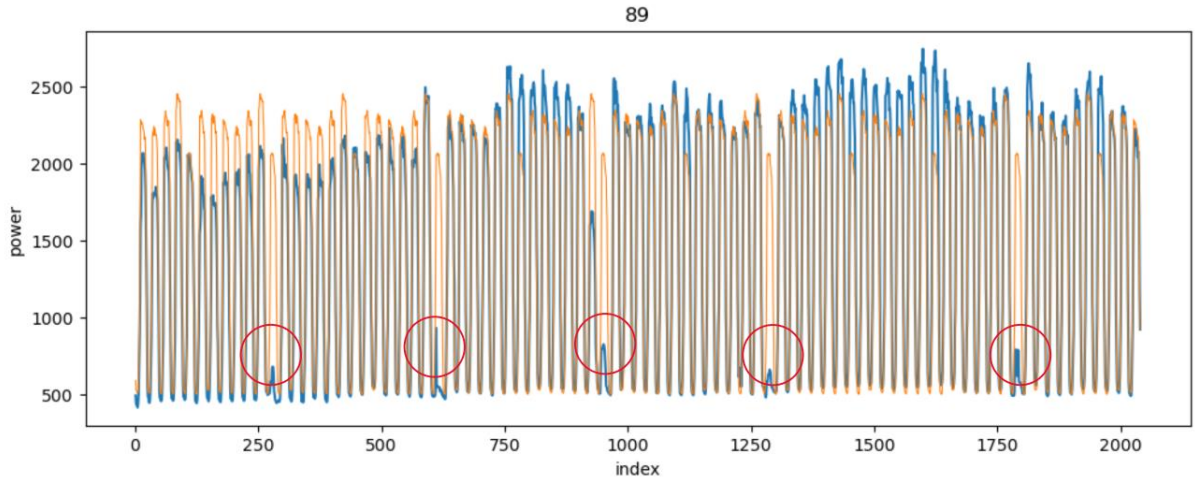
범주형 변수

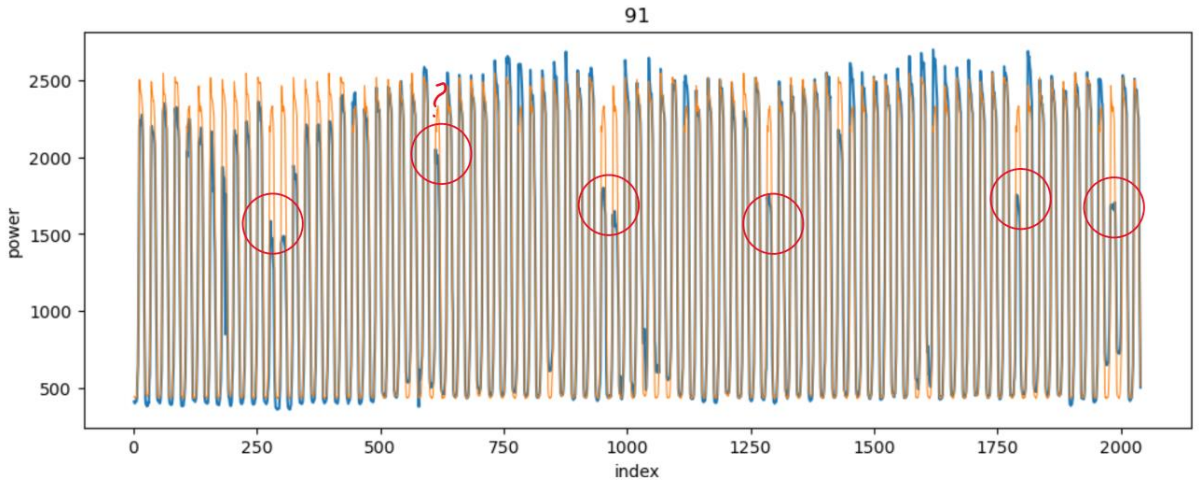
Closed

할인마트 휴무일









Feature

안정적인 CV – Public Score correlation

일반화 가능한 변수선택

범주형 변수

Closed

할인마트 휴무일

86번 마트의 경우 매월 10일과 4번째 일요일을 휴무일로 지정

87, 88, 89, 90, 91, 92번 마트의 경우 매월 2, 4번째 일요일을 휴무일로 지정

규칙적인 휴무일 외 불규칙적인 휴무일(91번 마트의 경우 6월 13일, 7월 11일, 8월 22일)을 휴무일로 포함

2022년 8월 28일 일요일이 4번째 일요일이므로 Private Score에 영향을 미칩니다.

Feature

안정적인 CV – Public Score correlation

일반화 가능한 변수선택

수치형 변수

모든 수치형 변수는 NN을 염두에 두고 생성

cos_hour, sin_hour

temperature_squared, THI, humidity_squared

Feature

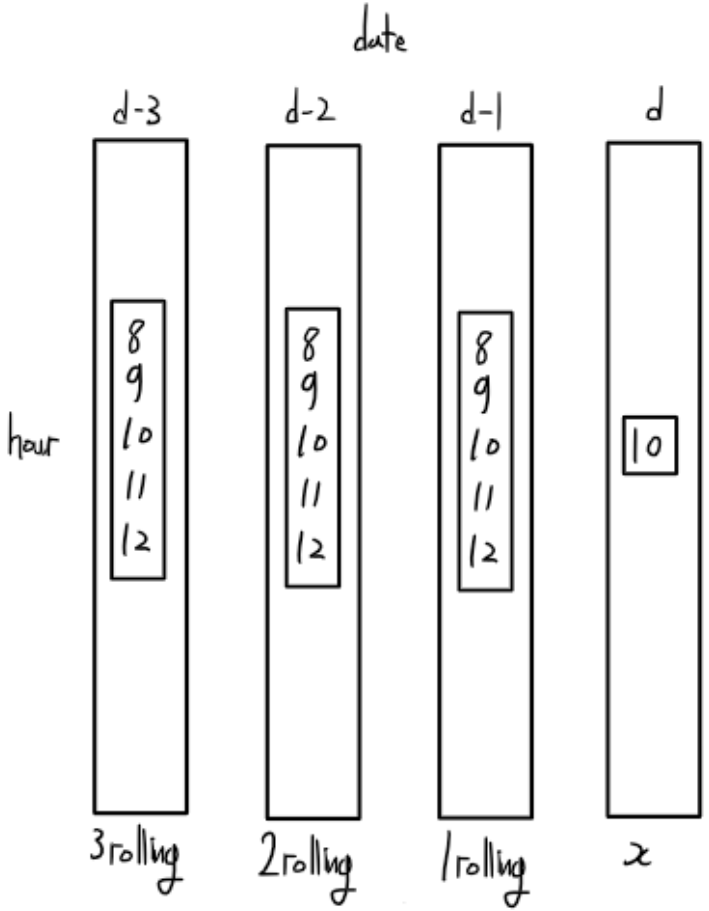
안정적인 CV – Public Score correlation

일반화 가능한 변수선택

수치형 변수 (rolling)

모든 수치형 변수는 NN을 염두에 두고 생성

temperature_mean, THI_mean



$$i\ rolling_h^d = \frac{x_{h-2}^{d-i} + x_{h-1}^{d-i} + x_h^{d-i} + x_{h+1}^{d-i} + x_{h+2}^{d-i}}{5}$$

$$2\ rolling_{10}^d = \frac{x_8^{d-2} + x_9^{d-2} + x_{10}^{d-2} + x_{11}^{d-2} + x_{12}^{d-2}}{5}$$

rolling: 어제 혹은 2, 3일 전 비슷한 시간대(5window)에 느꼈던 온도, 습도의 정보

$$1\ rolling\_mean_h^d = \frac{x_h^d + 1\ rolling_h^d}{2}$$

$$12\ rolling\_mean_h^d = \frac{x_h^d + 1\ rolling_h^d + 2\ rolling_h^d}{3}$$

$$123\ rolling\_mean_h^d = \frac{x_h^d + 1\ rolling_h^d + 2\ rolling_h^d + 3\ rolling_h^d}{4}$$

$$mean_h^d = \frac{x_h^d + 1 rolling\_mean_h^d + 12 rolling\_mean_h^d + 123 rolling\_mean_h^d}{4}$$

Feature

안정적인 CV – Public Score correlation

일반화 가능한 변수선택

수치형 변수 (rolling)

모든 수치형 변수는 NN을 염두에 두고 생성

temperature_mean, THI_mean

	temperature_squared	THI	humidity_squared
vanilla	0.521688	0.487043	-0.275728
1rolling	0.523406	0.483777	-0.273691
2rolling	0.533213	0.487919	-0.279840
3rolling	0.536625	0.490460	-0.268420
1rolling_mean	0.554868	0.505015	-0.314470
12rolling_mean	0.578910	0.517876	-0.338940
123rolling_mean	0.596524	0.527870	-0.350416
mean	0.576859	0.518419	-0.333736
power_log1p	1.000000	1.000000	1.000000

상관계수 증가

변수들과 power_log1p 간 상관계수

1rolling_mean
12rolling_mean
123rolling_mean

정보함축 →

{feature}_mean

Feature

안정적인 CV – Public Score correlation

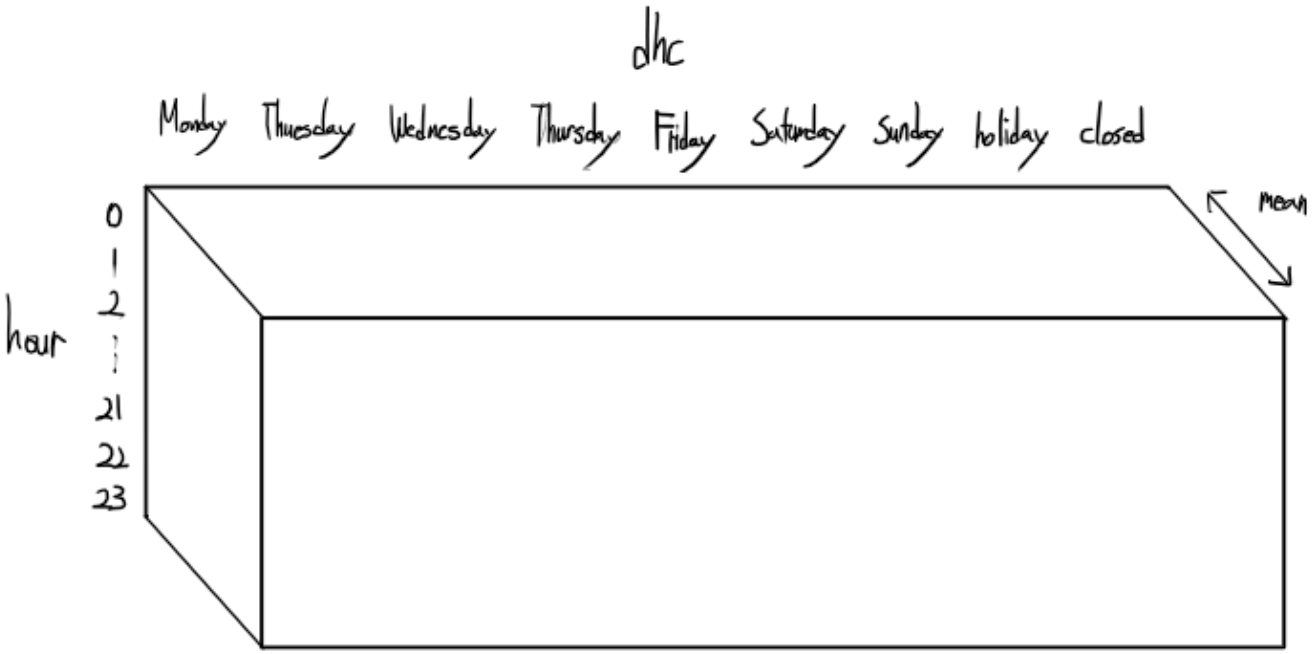
일반화 가능한 변수선택

수치형 변수 (Target Statistic)

모든 수치형 변수는 NN을 염두에 두고 생성

power_log1p_stdd_mean

예) 월요일 1시 전력량들의 평균을 월요일 1시의 변수로 사용



dhc 정보 포함

↓

dhc + weekend 조합 제거

Feature

안정적인 CV – Public Score correlation

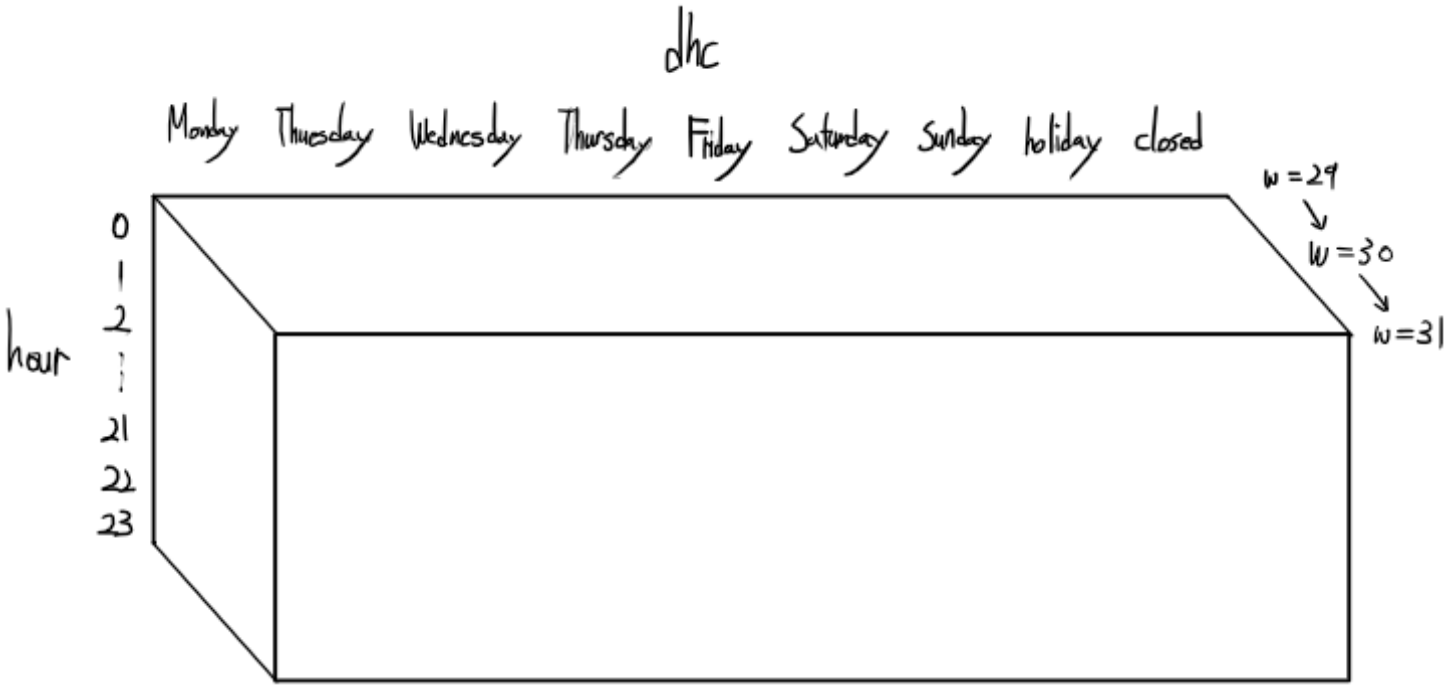
일반화 가능한 변수선택

수치형 변수 (Target Statistic)

모든 수치형 변수는 NN을 염두에 두고 생성

power_log1p_stdd_shift

예) 지난주 월요일 1시의 전력량을 이번주 월요일 1시의 변수로 사용



dhc 정보 포함

month, vacation 정보 일부 포함

dhc + weekend 조합 제거

month, vacation 제거

Feature

안정적인 CV – Public Score correlation

일반화 가능한 변수선택

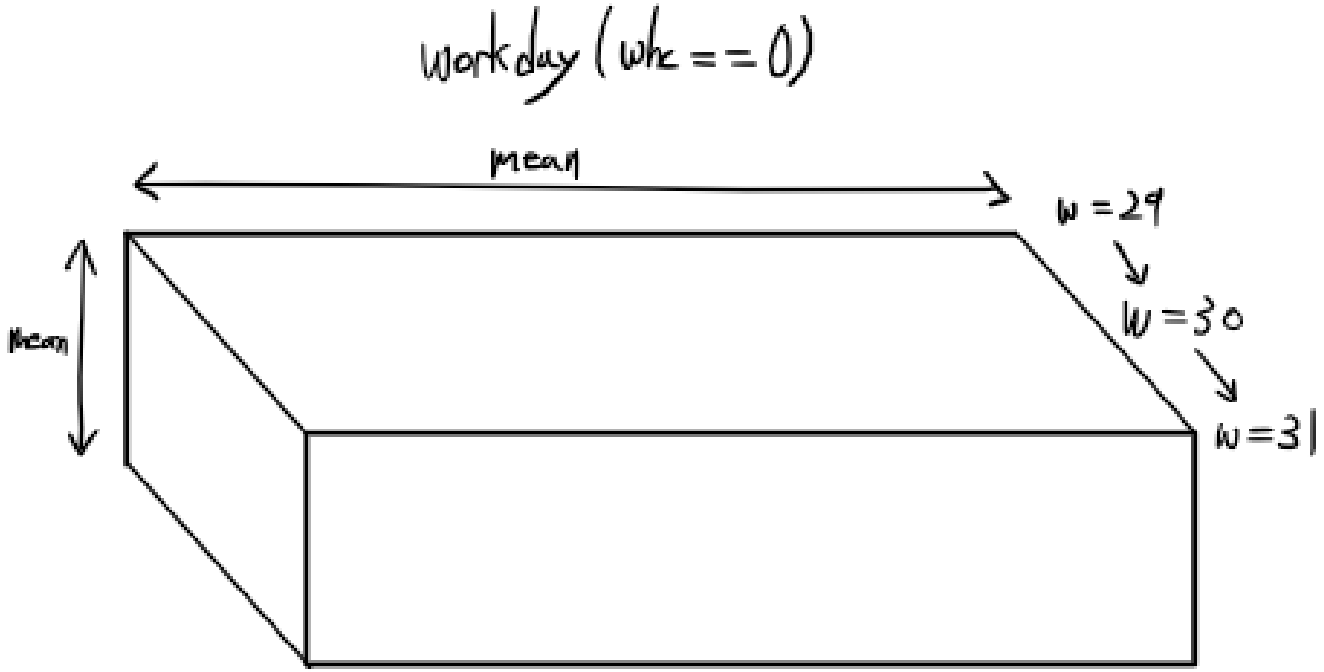
수치형 변수 (Target Statistic)

모든 수치형 변수는 NN을 염두에 두고 생성

power_log1p_stdd_cumweek_mean_shift

예) 지난주 workday들의 전력량 평균을 이번주 workday의 변수로 사용

예) 14번 건물의 경우 지난주 workday(월, 화, 수, 목, 금)들의 전력량 평균을 이번주 수요일 변수로 사용



workday 정보 포함

month, vacation 정보 일부 포함

week 정보 일부 포함

month, vacation 제거

week 제거

Feature

안정적인 CV – Public Score correlation

일반화 가능한 변수선택

workday 정보 포함

month, vacation 정보 일부 포함

week 정보 일부 포함

↓

month, vacation 제거

week 제거

수치형 변수 (Target Statistic)

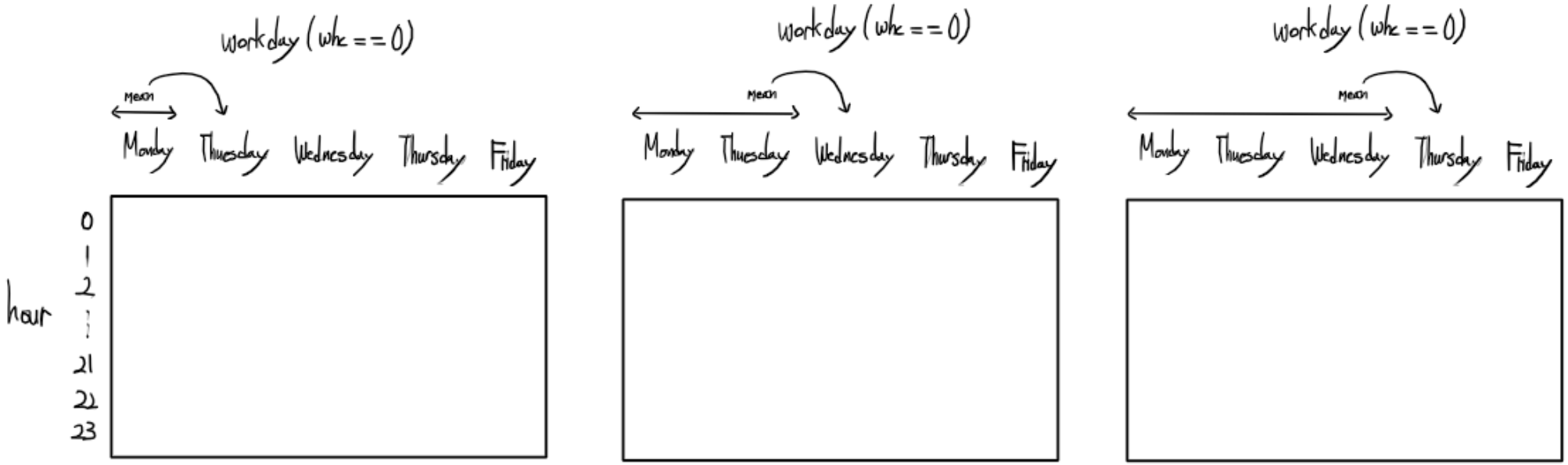
모든 수치형 변수는 NN을 염두에 두고 생성

power_log1p_stdd_thisweek_mean_shift

예) 이번주 workday인 월요일의 전력량 평균을 이번주 화요일 변수로 사용

예) 이번주 workday인 월, 화 전력량 평균을 이번주 수요일 변수로 사용

예) 이번주 workday인 월, 화, 수 전력량 평균을 이번주 목요일 변수로 사용



Feature

안정적인 CV – Public Score correlation

일반화 가능한 변수선택

공통변수

whc,

cos_hour, sin_hour,

temperature_squared, THI, humidity_squared,

temperature_squared_mean, THI_mean,

power_log1p_stdd_mean,

power_log1p_stdd_shift

EE_TRAIN_0.ipynb

공통변수,

power_log1p_stdd_cumweek_mean_shift,

power_log1p_stdd_thisweek_mean_shift

EE_TRAIN_1.ipynb

공통변수,

power_log1p_stdd_cumweek_mean_shift

EE_TRAIN_2.ipynb

공통변수

Feature

안정적인 CV – Public Score correlation

일반화 가능한 변수선택

EE_TRAIN_0.ipynb

	mean
power_log1p_stdd_mean	11.225888
power_log1p_stdd_shift	9.112626
power_log1p_stdd_thisweek_mean_shift	3.538538
cos_hour	2.691489
THI_mean	2.219360
temperature_squared_mean	1.367848
THI	1.304061
sin_hour	1.113261
power_log1p_stdd_cumweek_mean_shift	0.710931
temperature_squared	0.629741
humidity_squared	0.380160

EE_TRAIN_1.ipynb

	mean
power_log1p_stdd_mean	11.323575
power_log1p_stdd_shift	7.429943
cos_hour	4.397443
THI_mean	2.449407
sin_hour	2.018548
temperature_squared_mean	1.713289
THI	1.206053
power_log1p_stdd_cumweek_mean_shift	1.076983
temperature_squared	0.527321
humidity_squared	0.410157

EE_TRAIN_2.ipynb

	mean
power_log1p_stdd_mean	11.834591
power_log1p_stdd_shift	7.895171
THI_mean	2.056753
temperature_squared_mean	1.577904
sin_hour	1.500297
cos_hour	1.499326
THI	1.268647
temperature_squared	0.618974
humidity_squared	0.498841

수치형 변수의 Feature Importance

Feature

안정적인 CV – Public Score correlation

일반화 가능한 변수선택

EE_INFERENCE.ipynb

submission_1.csv

8월 27일 이전	8월 27일 당일, 이후
EE_TRAIN_0 EE_TRAIN_1	EE_TRAIN_2
Public Score	5.2707265718
Private Score	6.1053683645

submission_2.csv

8월 27일 이전	8월 27일 당일, 이후
EE_TRAIN_1	EE_TRAIN_2
Public Score	5.2786695632
Private Score	6.1087725037

submission_1과 submission_2 간 Score 차이가 크지 않으므로 더 적은 변수를 사용한 submission_2 를 선택하는 것이 바람직해 보임

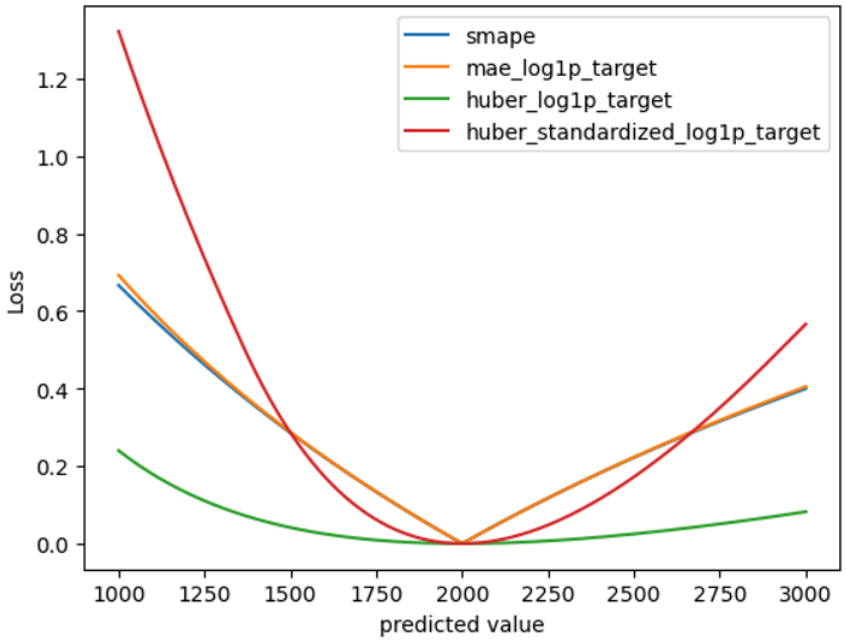
Model

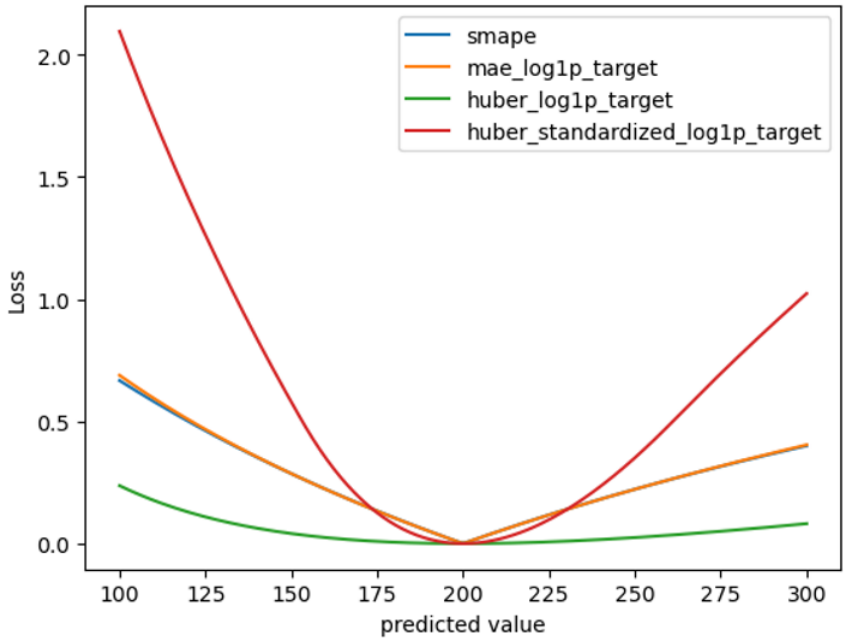
안정적인 CV – Public Score correlation

일반화 가능한 변수선택

건물별 XGBoost 단일모형

Loss function: huber_standardized_log1p_target





	smape	mae_log1p_target	huber_log1p_target	huber_standardized_log1p_target
diffrentiable	X	X	O	O
stable	X	X	X	O

코드공유 게시판
huber loss & standardized_log1p_target 참조

Model

안정적인 CV – Public Score correlation

일반화 가능한 변수선택

건물별 XGBoost 단일모형

Hyper parameter tuning: 없음

Post processing: 없음