DACON
DATA TO VALUE

에너지 빅데이터 활용 데이터 사이언스 경진대회

유 광 남

데이터 파일

DACON
DATA TO VALUE

1. Train.csv
> 1300 세대의 2016.07.26 ~ 2018.06.30 전력 데이터

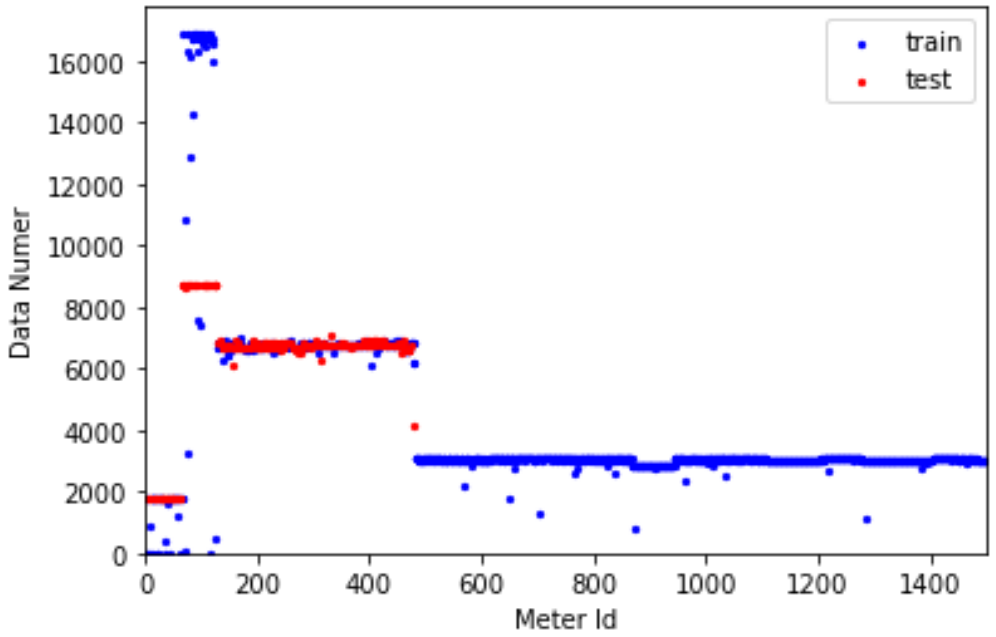
2. Test.csv
> 200 세대의 2017.07.01 ~ 2018.06.30 전력 데이터

3. 인천_시간별__기상자료(16-18)_축소__7월1일.csv (주최측 제공)
> 인천 지역의 2016.07.20 ~ 2018.07.01 날씨 데이터

결측치

1. Train.csv, Test.csv 파일 결측치로 데이터 개수에 차이가 있음
2. ID 481번 기준으로 데이터 개수의 분포가 다름

> 0 ~ 481 번 데이터만 사용하는 것을 고려



결측치 처리

대회 참가하기

DACON
DATA TO VALUE

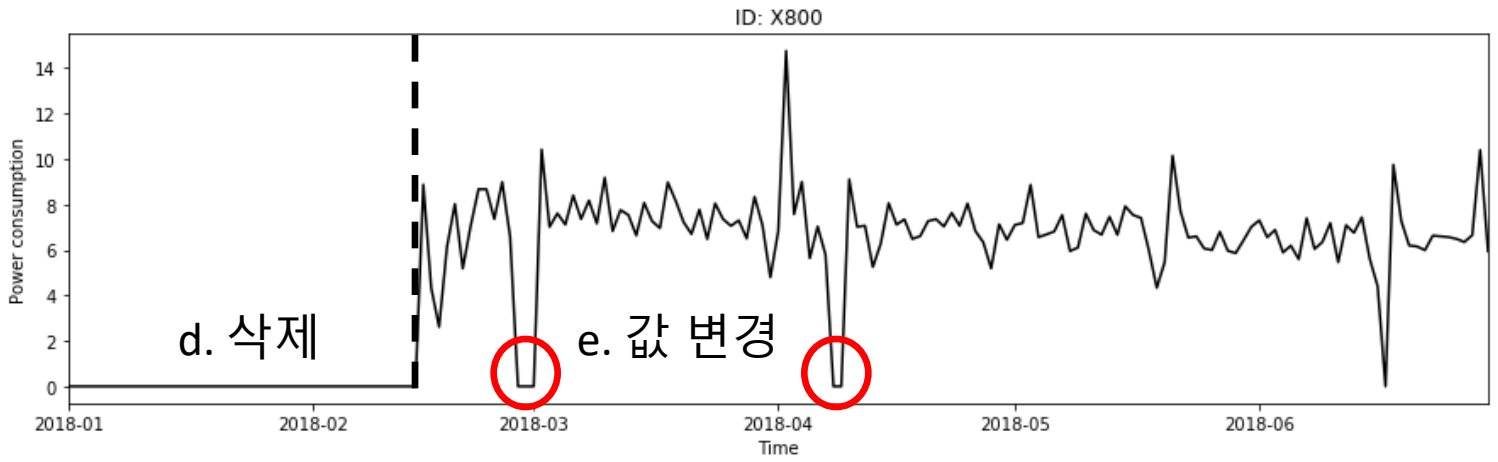
a. 결측치는 모두 0으로 변경

b. 0.003 보다 작은 값은 0으로 변경

c. 시간별 전력을 하루 단위로 합쳐 날짜별 전력으로 변경

d. 유효한 데이터만 사용하기 위해 앞뒤로 0인 구간을 삭제

e. 데이터 중간에 0이 있는 경우 값 변경 (ex. 평균값)



Feature

대회 참가하기

DACON
DATA TO VALUE

1. 전력 데이터 정규화 (*meter id 별로 따로 처리)
> 최대값으로 나눔 > 로그를 씌움 > 평균으로 뺌 > 표준편차로 나눔
2. 전력데이터에 해당하는 날짜의 시간, 요일 정보를 사용
> 시간은 주기가 365일, 최대값이 8월 1일인 사인함수를 사용
3. 2018.07.01 을 예측하는 경우 날씨 데이터 사용을 고려함
4. 2018.07.01 이후를 예측하는 경우 날씨 데이터 사용 X
5. 예측하는 날의 약 10주 전 데이터를 모아 1차원 벡터로 사용

모델

LightGBM

조건에 따라 18개 모델을 생성함

> ID: (0~481 / 0~1500)

> 전력데이터 중 0에 대한 처리를 3 종류로 분류

> 2018.07.01 예측에 날씨 데이터 (사용 / 사용 X)

> 기상 모델 사용시 2018.07.02 예측을 위해 (74 / 75)일 데이터 사용

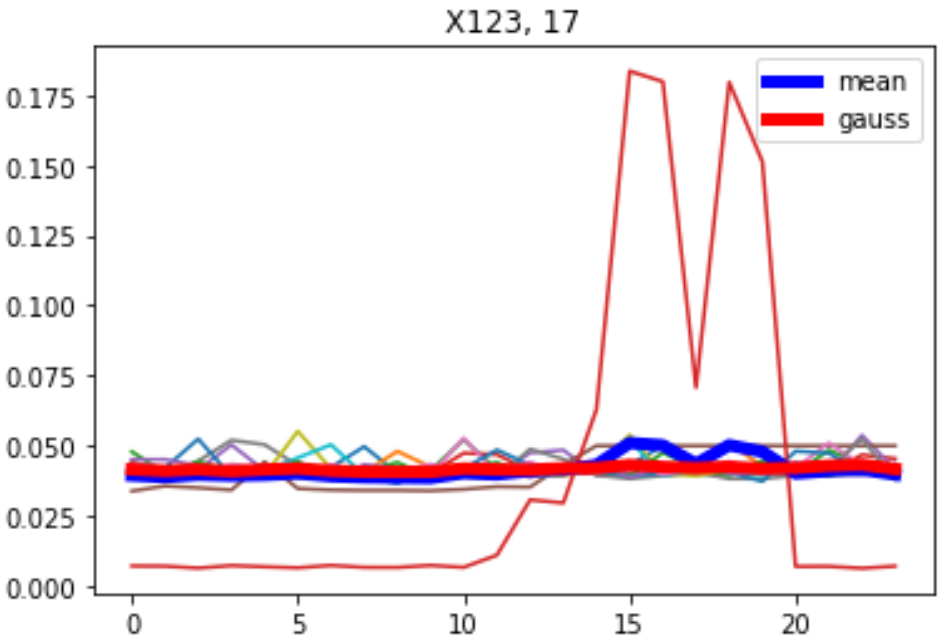
시간별 전력 수요 예측(18.07.01)

대회 참가하기

DACON
DATA TO VALUE

2018.07.01의 시간별 수요 예측을 위해 2018.03.01~2018.06.30 기간 내 일요일에 해당하는 시간별 수요의 분포를 구함

Gauss 함수로 weight를 주어 좀 더 일반적인 대표값을 구함

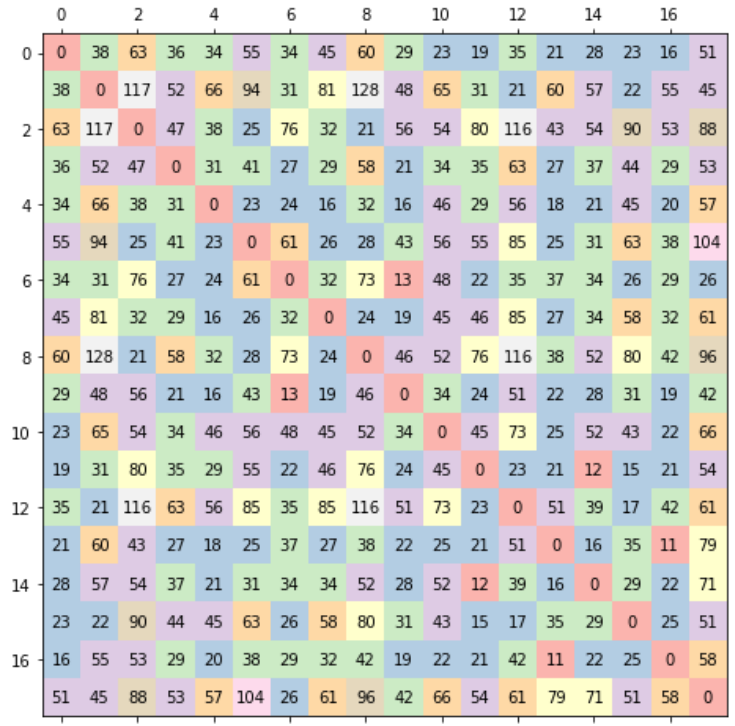


모델 선택

대회 참가하기

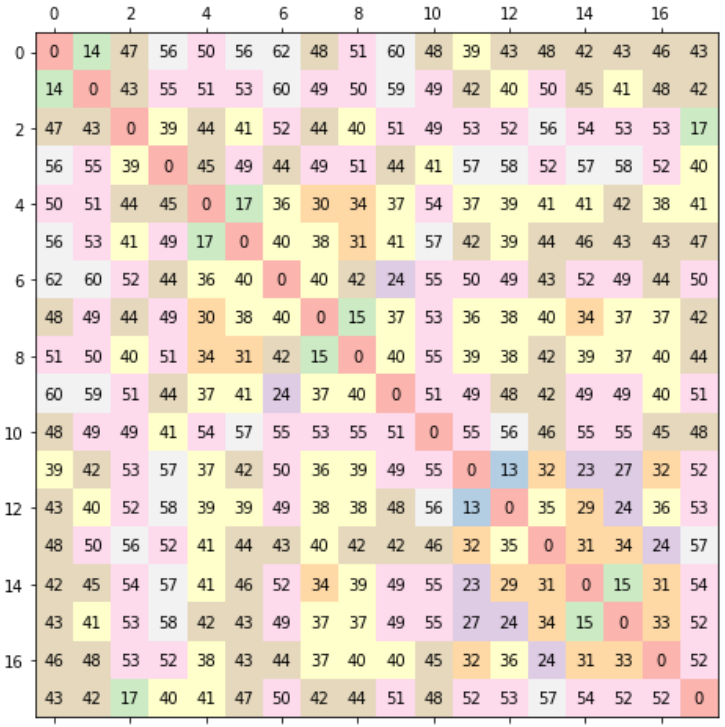
DACON
DATA TO VALUE

1-R2 score



	0	2	4	6	8	10	12	14	16
0	0	38	63	36	34	55	34	45	60	29	23	19	35	21	28	23	16	51
	38	0	117	52	66	94	31	81	128	48	65	31	21	60	57	22	55	45
2	63	117	0	47	38	25	76	32	21	56	54	80	116	43	54	90	53	88
	36	52	47	0	31	41	27	29	58	21	34	35	63	27	37	44	29	53
4	34	66	38	31	0	23	24	16	32	16	46	29	56	18	21	45	20	57
	55	94	25	41	23	0	61	26	28	43	56	55	85	25	31	63	38	104
6	34	31	76	27	24	61	0	32	73	13	48	22	35	37	34	26	29	26
	45	81	32	29	16	26	32	0	24	19	45	46	85	27	34	58	32	61
8	60	128	21	58	32	28	73	24	0	46	52	76	116	38	52	80	42	96
	29	48	56	21	16	43	13	19	46	0	34	24	51	22	28	31	19	42
10	23	65	54	34	46	56	48	45	52	34	0	45	73	25	52	43	22	66
	19	31	80	35	29	55	22	46	76	24	45	0	23	21	12	15	21	54
12	35	21	116	63	56	85	35	85	116	51	73	23	0	51	39	17	42	61
	21	60	43	27	18	25	37	27	38	22	25	21	51	0	16	35	11	79
14	28	57	54	37	21	31	34	34	52	28	52	12	39	16	0	29	22	71
	23	22	90	44	45	63	26	58	80	31	43	15	17	35	29	0	25	51
16	16	55	53	29	20	38	29	32	42	19	22	21	42	11	22	25	0	58
	51	45	88	53	57	104	26	61	96	42	66	54	61	79	71	51	58	0

SMAPE



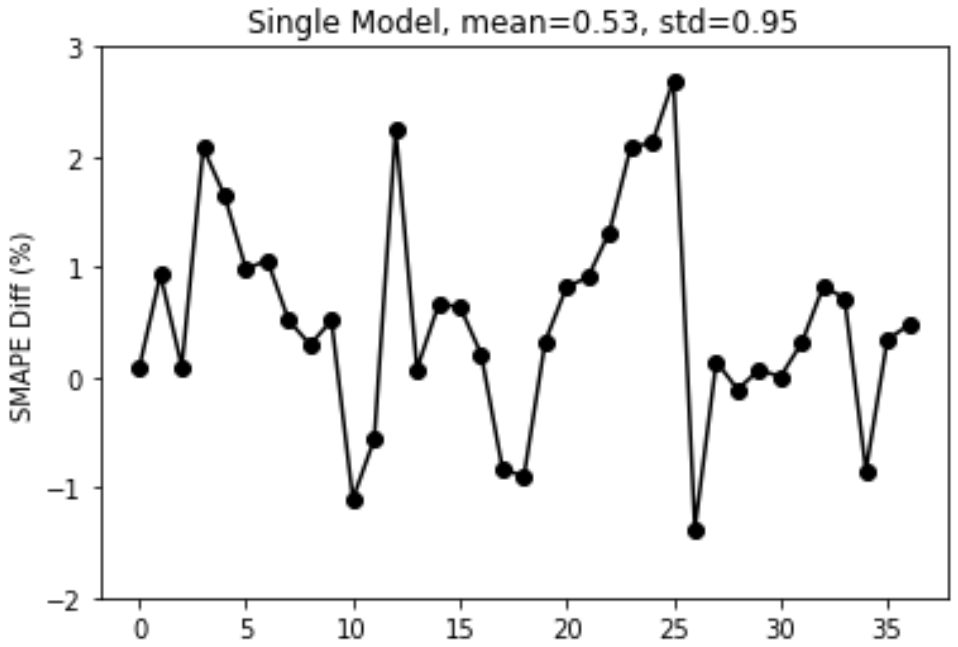
	0	2	4	6	8	10	12	14	16
0	0	14	47	56	50	56	62	48	51	60	48	39	43	48	42	43	46	43
	14	0	43	55	51	53	60	49	50	59	49	42	40	50	45	41	48	42
2	47	43	0	39	44	41	52	44	40	51	49	53	52	56	54	53	53	17
	56	55	39	0	45	49	44	49	51	44	41	57	58	52	57	58	52	40
4	50	51	44	45	0	17	36	30	34	37	54	37	39	41	41	42	38	41
	56	53	41	49	17	0	40	38	31	41	57	42	39	44	46	43	43	47
6	62	60	52	44	36	40	0	40	42	24	55	50	49	43	52	49	44	50
	48	49	44	49	30	38	40	0	15	37	53	36	38	40	34	37	37	42
8	51	50	40	51	34	31	42	15	0	40	55	39	38	42	39	37	40	44
	60	59	51	44	37	41	24	37	40	0	51	49	48	42	49	49	40	51
10	48	49	49	41	54	57	55	53	55	51	0	55	56	46	55	55	45	48
	39	42	53	57	37	42	50	36	39	49	55	0	13	32	23	27	32	52
12	43	40	52	58	39	39	49	38	38	48	56	13	0	35	29	24	36	53
	48	50	56	52	41	44	43	40	42	42	46	32	35	0	31	34	24	57
14	42	45	54	57	41	46	52	34	39	49	55	23	29	31	0	15	31	54
	43	41	53	58	42	43	49	37	37	49	55	27	24	34	15	0	33	52
16	46	48	53	52	38	43	44	37	40	40	45	32	36	24	31	33	0	52
	43	42	17	40	41	47	50	42	44	51	48	52	53	57	54	52	52	0

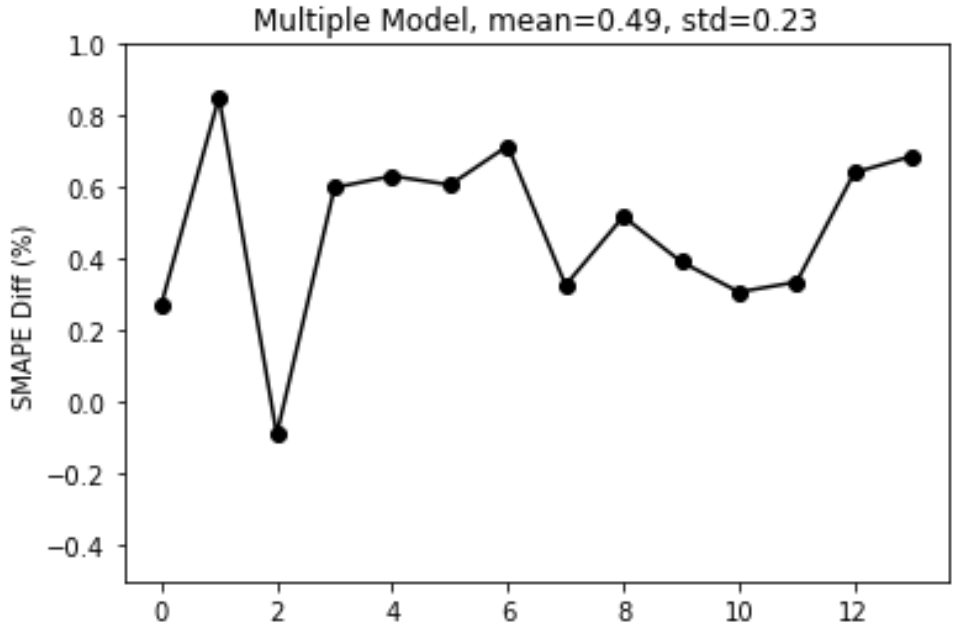
18개 결과에 대해 1-R2 score, SMAPE를 계산, 각 score의 중간값 보다 작은 데이터만 선별하고 평균을 구함

모델 선택

대회 참가하기

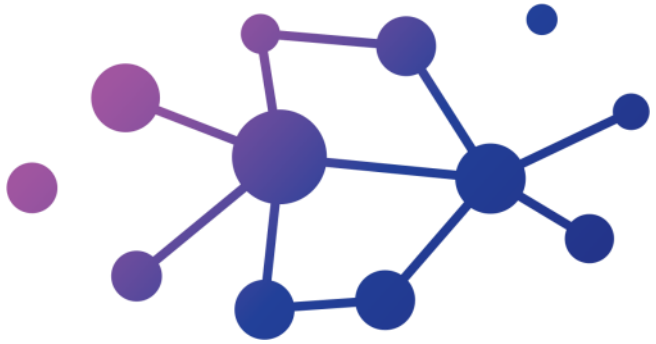
DACON
DATA TO VALUE





Single model, Multiple model 의 private, public score 차이 (백분율)
Single model: mean=0.53, std=0.95, private score가 public score에 비해 약 0.5% 높다.
Multiple model: mean=0.49, std=0.23, single model 처럼 private score가 public score에 비해 약 0.5% 높음, single model에 비해 std가 4배 작아 더 안정적으로 예측가능함

THANK YOU



대회 참가해보기

https://dacon.io