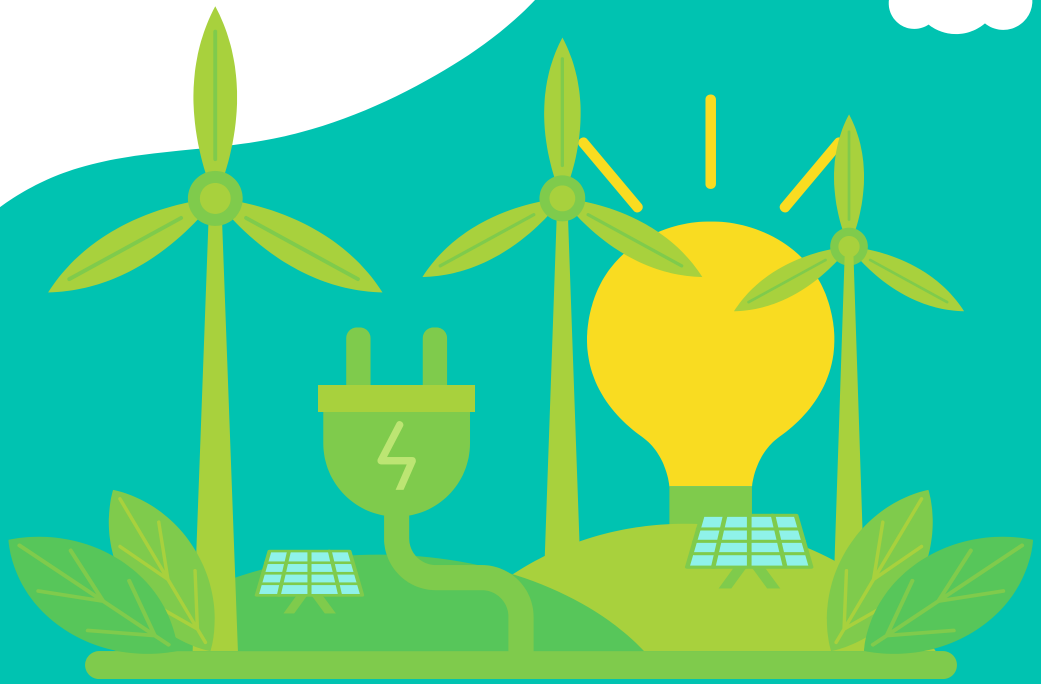
2023 전력사용량 예측 AI 경진대회

[To the mars]
dansama, 화성간다



한국에너지공단

Contents

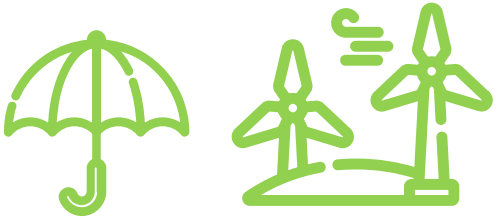
01

Preprocessing

Preprocessing

Building_info

- '태양광용량(kW)','ESS저장용량(kWh)', 'PCS용량(kW)'에서 '-'를 0으로 대체해 주었습니다.
- 이후, train data와 test data에 merge 해주었습니다.



습도,풍속

- 여러가지 시도를 해보았지만 결측치를 통계치로 대체해주는 것이 성능에 가장 긍정적 영향을 주었기에, 해당 월 시간 당 평균으로 결측치를 대체해주었습니다.

02

Feature Engineering

Feature Engineering

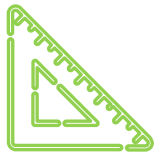
일시

- 시간, 월, 요일, 주에 대한 데이터를 추출하여 feature로 활용하였습니다.

휴일

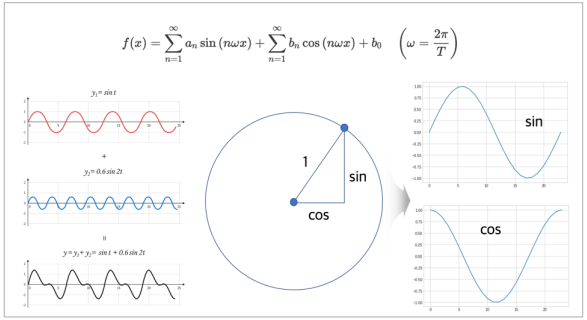
- 일부 건물에서 평일과 휴일의 전력사용량의 차이가 있으므로, 공휴일에 대한 feature를 추가한다면 성능이 더욱 향상 될 것이라고 판단하였습니다.
- 하지만, 이후 clustering에 공휴일을 활용한 뒤 두 feature 모두 사용했을 경우 성능이 하락하여 학습에서는 drop하여 진행하였습니다.

Feature Engineering

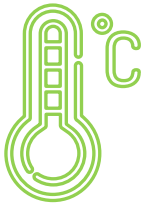


sin,cos time

- 주기성을 갖는 시간에 대해서 모델의 원활한 학습을 위해 hour를 삼각함수로 변환하여 feature로서 활요하였습니다.



출처: https://today-1.tistory.com/55



불쾌지수

- 백화점, 병원 등과 같은 건물유형에서는 불쾌지수가 전력사용량에 유의미한 영향을 줄 수 있다고 판단하여 feature로 활용하였습니다.
- 불쾌지수 단계로 범주화를 진행하였을 때, 더 높은 성능을 보이는 것을 확인하여, 범주형 feature로 활용하였습니다.

DI	℃	불쾌를느끼는 정도
68 이하	20 이하	전원 쾌적
70	21	불쾌를 나타냄
75	24	10% 정도 불쾌
80	26.5	50% 정도 불쾌
83	28.5	전원 불쾌
86	30.0	매우 불쾌

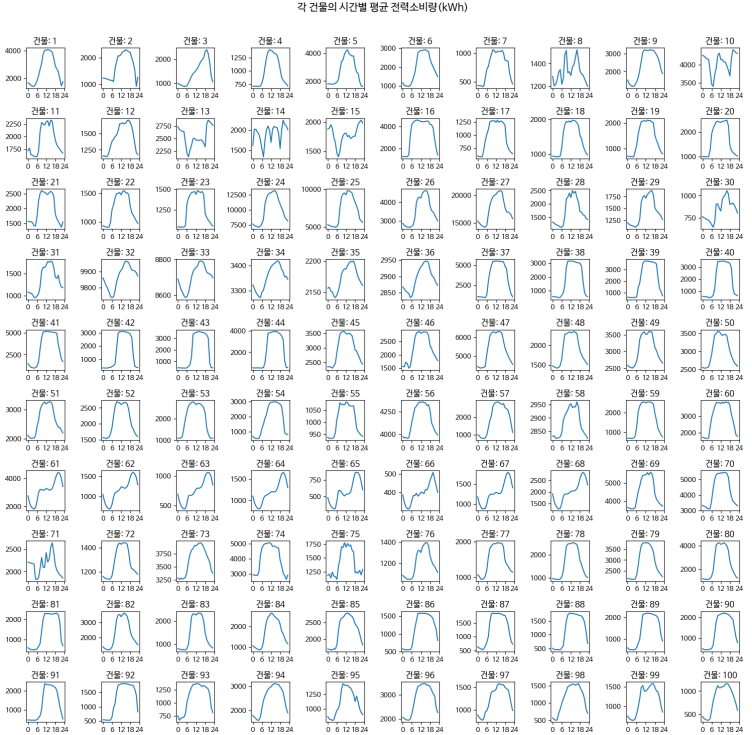
Feature Engineering

파생변수

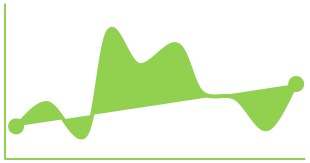
- 각 요일별,시간별,월별 등에 대해서 target의 다양한 통계치를 feature로 추가하여 예측성능을 향상시키고자 했습니다.

Clustering

- 건물 시간별 평균 전력사용량에 대한 정보를 시각화 해본 결과, 주어진 건물유형을 그대로 활용하는 것이 바람직하지 않다고 판단했습니다.
- 따라서, holiday feature를 활용하여 새롭게 clustering을 진행하였습니다.

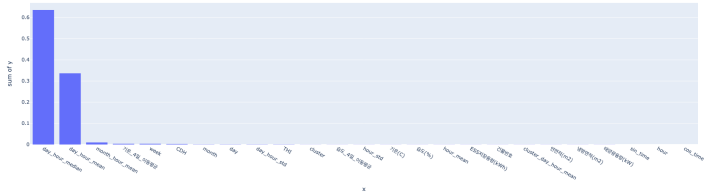


Feature Engineering



이동평균

- 기온, 습도 데이터를 확인해본 결과 현실의 데이터이기에 다소 이상치가 있다는 사실을 발견하였습니다.
- 따라서, 추세에 대한 feature를 추가한다면 이상치의 영향을 줄일 수 있을 것이라고 판단하여 feature로 추가하였습니다.
- 7일과 4일 각각 활용한 모델의 결과물을 Ensemble하는 형태로 일반화 성능을 향상 시키고자 하였습니다.

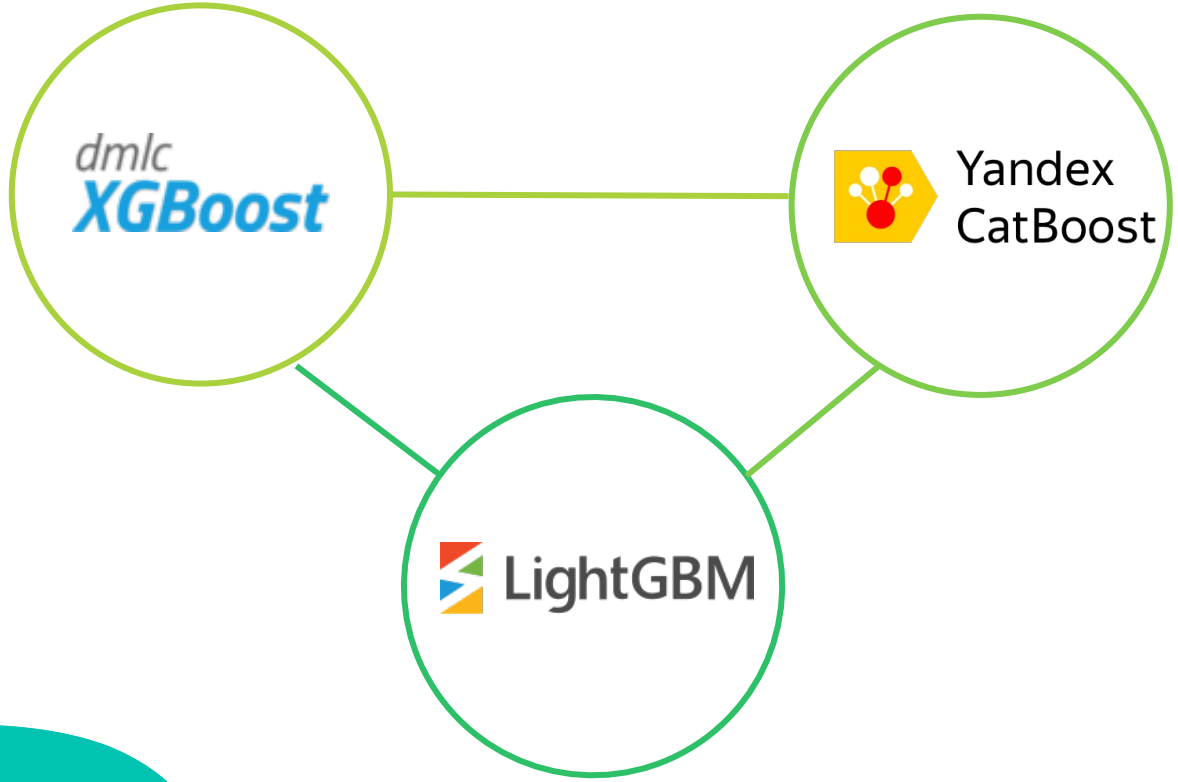


전반의 과정

- 전반적으로 feature engineering과정에서 일반화 성능의 향상을 위해 Feature importance 및 SHAP 라이브러리를 통해서 각각의 feature의 영향을 확인하며 특정 feature에 대해서 과도하게 의존하는 것을 줄이기 위해 노력하였습니다.

03
Train

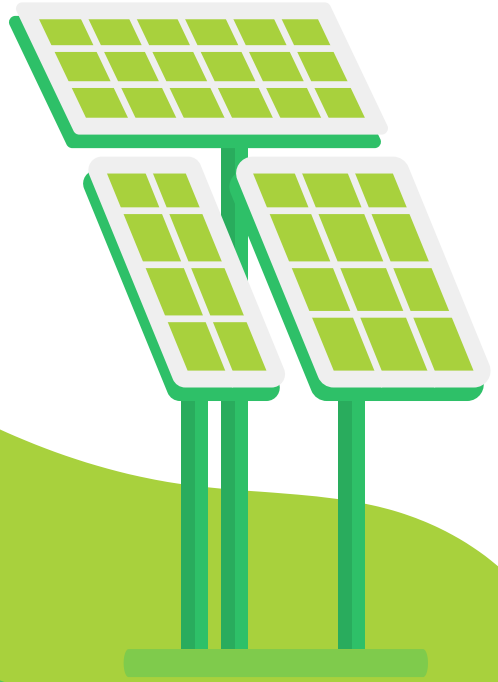
Train



Train

	XGB	LGBM	CAT	Ensemble
2	3.843336253	4.0445853	4.165173434	3.710464195
3	5.510573219	5.592893401	5.827088759	5.296176898
4	5.628413661	5.718504754	5.902353274	5.267299465
5	2.898155401	2.954764369	2.982013994	2.804820347
6	3.740469516	3.987102403	4.207586828	3.53992815
7	2.362135472	2.294428215	2.387399138	2.163817347
8	4.454218739	4.445442912	4.632102079	4.162492456
9	3.24488394	3.219283747	3.389698704	3.111586242
10	2.179696745	2.166383086	2.185099313	1.982034
11	2.348212592	2.358156527	2.353688597	2.185553252
12	2.22126859	2.309164736	2.214539618	2.116306236
13	2.279689278	2.340008251	2.376297472	2.164464232
14	3.088397627	3.091888965	3.222691519	2.944981792
15	9.364872971	9.470495937	9.762942283	9.162595321
16	1.876465964	1.886161639	1.858969748	1.752914307

- Catboost, LGBM, XGBoost 총 3가지의 모델을 활용하였습니다.
  하나의 모델을 활용하였을 때보다 다수의 모델의 결과물을 Ensemble 하였을 때, 예측성능이 더욱 우수하였고, 일반화 가능성도 더욱 향상 시킬 것이라고 판단하여 다수의 모델을 사용하는 방식을 채택하였습니다.
- 다양한 방식으로 학습한 모델들의 결과물을 Ensemble한다면 일반화 성능을 향상시킬 수 있다고 판단하여, 학습과정에서 일부 모델에 대해서 Categorical Feature를 바꿔가며 일반화 성능을 향상시키고자 하였습니다.
- Sklearn을 활용하여 5-Fold 검증 및 Ensemble을 진행하였습니다.
- 학습에서 SEED의 영향력을 줄이기 위해 seed ensemble을 활용하였습니다.



04

Predict

Predict

Predict

예측성능 향상과 일반화 가능성 증대를 위해 3개의 모델에서의 예측을 산술평균하여 진행하였습니다.

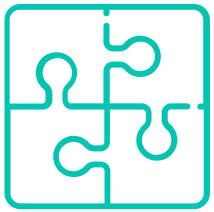
dmlc XGBoost

LightGBM

Yandex CatBoost

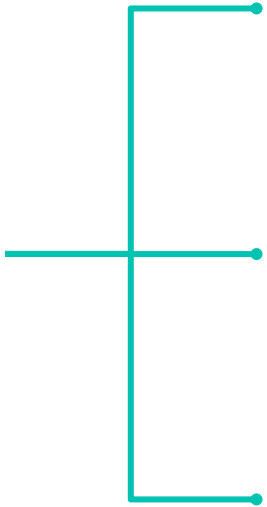
05
Ensemble

Ensemble



Final Submission

- 서로 다른 방식으로 학습한 모델들의 최종결과물에 대해서 가중평균을 진행하여 개별 결과물 대비 높은 성능을 유도하면서 일반화 가능성을 증대시키고자 하였습니다.
- (Public 기준 4.93~ 5.04) 가중평균 결과물 : 4.827)



67번파일

0.35

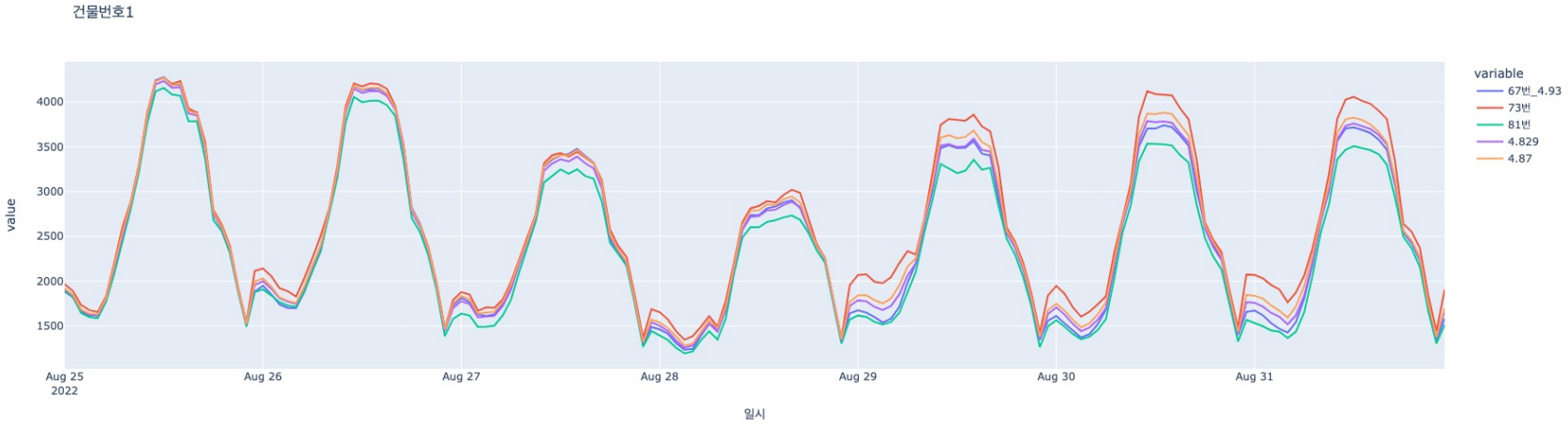
73번파일

0.3

81번파일

0.35

Ensemble



- 최종 예측값과 Public 점수를 활용하여 Ensemble에 활용할 예측모델을 선정하였습니다.
- 각 버전의 최종예측에 대해서 시각화를 통해, 가중평균 비율을 조정하는 것에 활용하였습니다.

THANK YOU