DACON

AI프렌즈 시즌1
공공데이터를 활용한
온도추정 경진대회

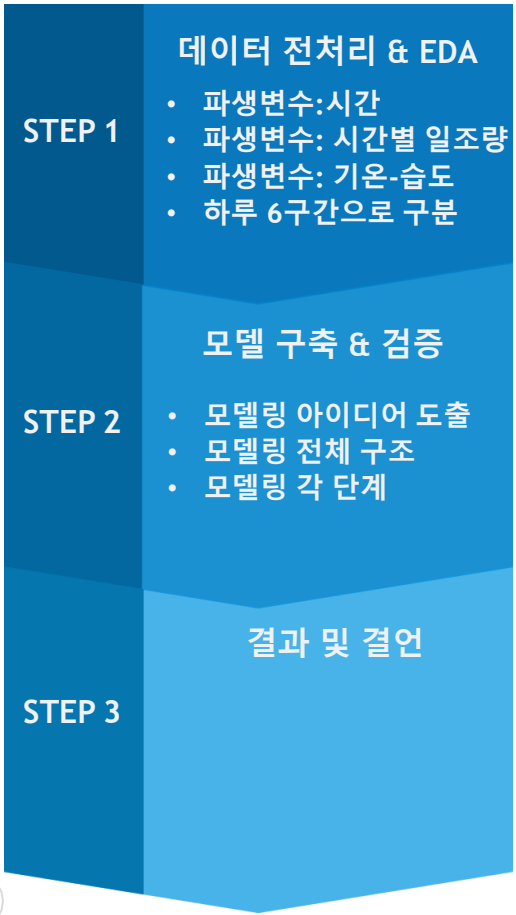
팀명 박꾸디

목차

1 데이터 전처리 및 EDA

2 모델 구축 및 검증

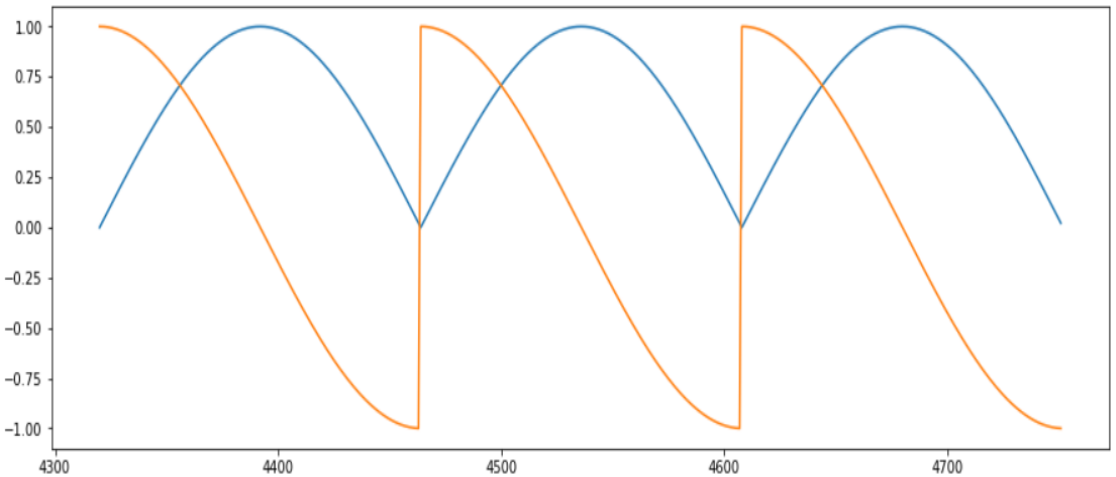
3 결과 및 결언



(1) 파생변수 생성 : 시간

삼각함수 (sin, cos)를 이용해서 시간 변수를 생성

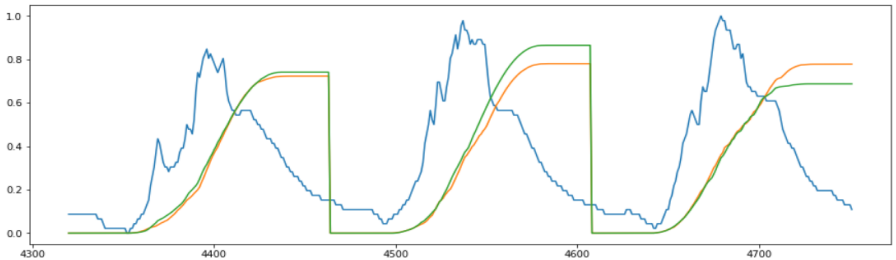
(코드 공유 "**기상캐스터 잔나의 데이터를 만지는 5가지 꿀팁**" 에서 보고 만들었습니다)



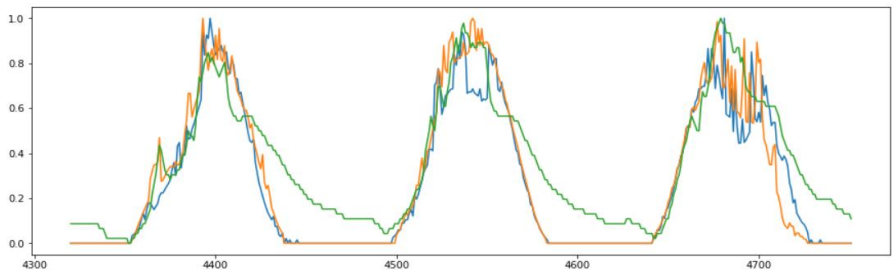
(2) 파생변수 생성 : 각 시간 별 일조량

누적 일조량 변수들로부터 각 시간별 일조량 변수를 생성

누적일조량

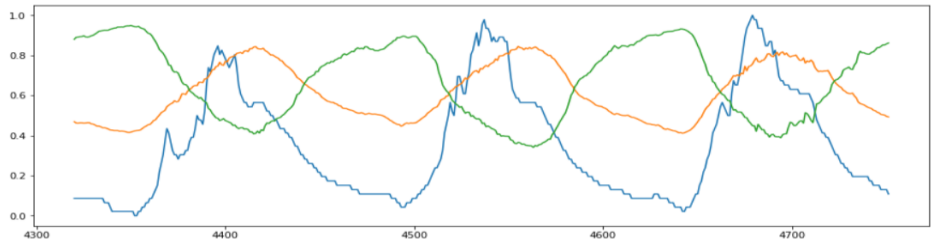


각 시간의 일조량

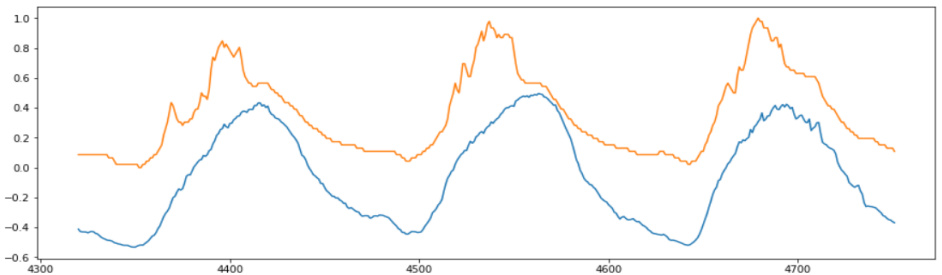


(3) 파생변수 생성 : 기온 변수 – 습도 변수

스케일링하여 시각화할 경우, 각 날에 기온과 습도가 항상 두번 교차하는 시점 존재.
따라서, 기온에서 습도를 뺀 변수들을 생성하였고, 이는 'Y18'과 나름 유사한 패턴을 보임.



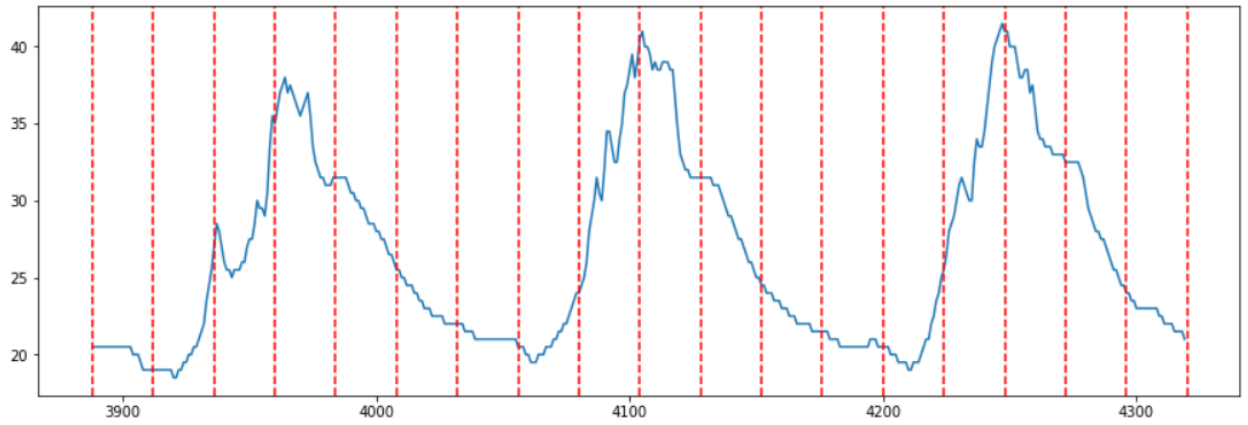
기온 평균
습도 평균
Y18



Y18
기온 평균
- 습도 평균

(4) 파생변수 생성: 하루를 6구간으로 구분

하루를 각 시간대에 따라 6구간으로 구분하였다.



모델링 아이디어 도출

● 어떤 모델을 사용할 것인가?

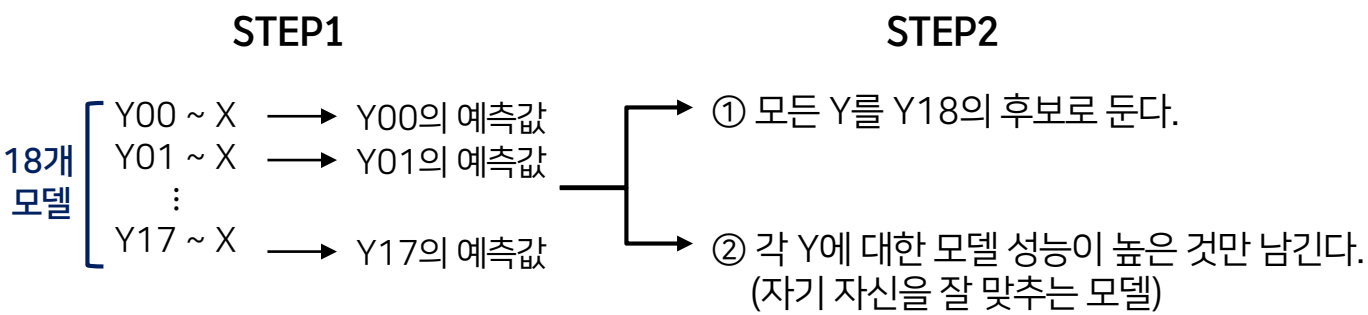
Regression이 제일 적절해 보였지만, 각 변수의 단위에 영향을 많이 받을 것이라 생각.
이 때 각각 변수들을 일 별로 스케일링 하면 최선일 것이지만, 이 경우 Data Leakage가 문제가 됨.
따라서 변수의 단위에 크게 영향을 받지 않는 **Boosting**을 선택.
하지만 Boosting의 경우 Overfitting의 문제가 존재하였고 또한, Y18을 직접적으로 훈련시킬 수 없었기 때문에 다른 Y들을 통해 Y18을 예측하여야 했음.

● 어떻게 이용할 것인가?

Y들을 단순히 앙상블할 경우에는 상당히 많은 오차가 있을 것. EDA 결과 Y18과 가장 유사한 하나의 Y를 찾기는 어려웠기 때문에, 각 시간대에서 Y18과 가장 유사한 Y들을 찾기로 결정.
따라서 **시간대(구간)마다 적절한 Y의 조합을 만들어서 Y18을 만듦**.

2. 모델 구축 및 검증

모델링 전체 구조



STEP3

→ 각 시간대(구간)에서 Y18과 가장 유사한 Y들의 조합을 찾고 앙상블.

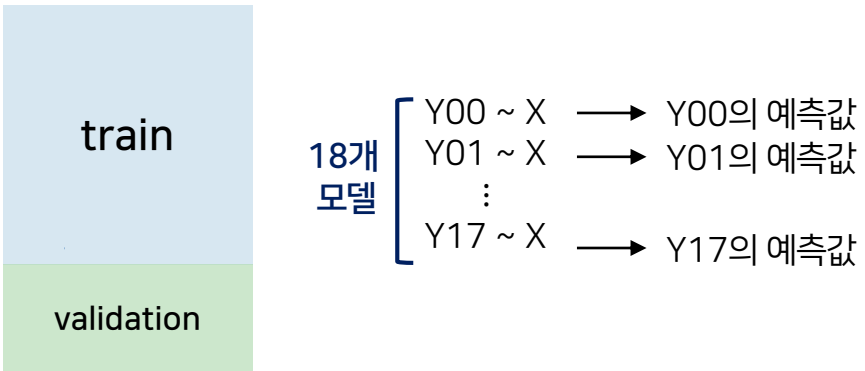
STEP4

→ 스무딩 후 잔차 보정

STEP1

train set에서 두가지 방법으로 validation set을 분리 :
(1) **shuffle=True**(5-fold CV),
(2) **shuffle=False**(train= 앞 데이터 0.8, validation = 뒤 0.2)

두 경우에 대해 Y00부터 Y17까지 각각의 모델을 구성하고,
test set에 대해 각각의 예측값을 뽑고 **앙상블**한다.

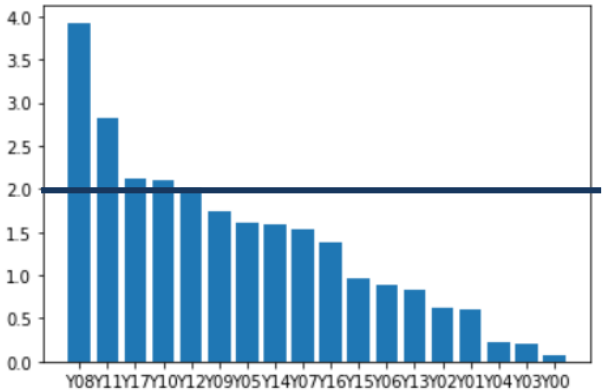


STEP2: 어떤 Y로 Y18을 맞출것인가?

① 모든 Y를 Y18의 후보로 둔다.

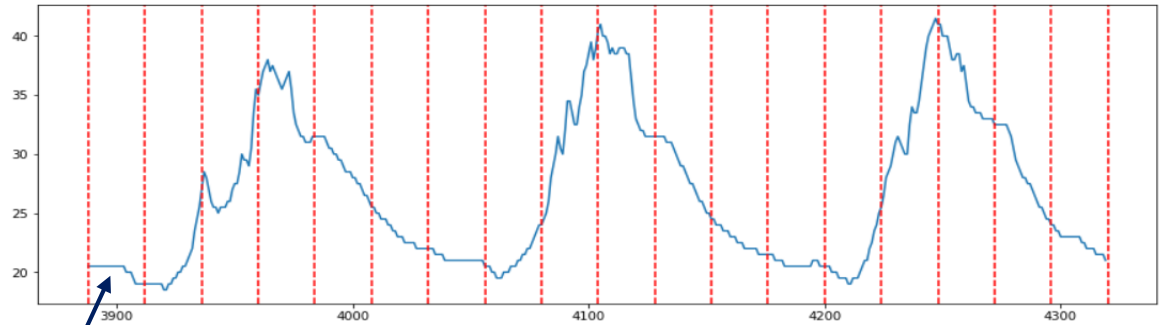
② Y 각각의 모델 중 성능이 높은 것만 남긴다. (자기 자신을 잘 맞추는 모델)

MAE가 2 이하인 Y들만 남기고 이것으로 Y18을 예측한다.
각 Y를 맞추지도 못하는 모델로 Y18을 예측하면 오차가 더 커질 수 있을 것이라 판단.



STEP3: 각 시간대별로 Y를 조합 후 앙상블

하루의 각 시간대(구간)에서 Y18과 가장 유사한 Y들의 평균을 Y18의 예측값으로 한다.
이 때, 구간은 8개, 12개, 144개 등으로 분리해보았다.



Y00, Y10, Y12…

- **Y18과 가장 유사한 Y 찾기**

A_j = Y18과 Y_i 예측값의 MAE가 j 이하인 Y_i들의 평균 $(j = 1, \dots, 30)$
B = Y18

→ $|A_j - B|$가 최소가 되는 A_j 를 이용하여 각 구간을 예측

2. 모델 구축 및 검증

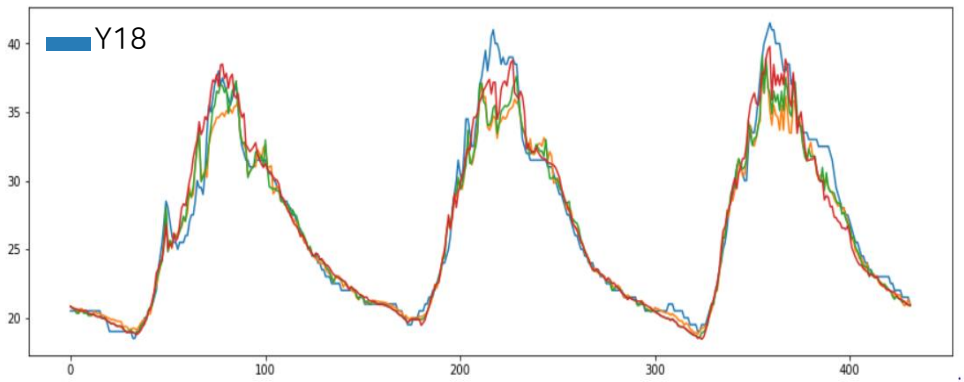
Validation 분리 방법, Y의 선택, 구간 개수에 따라 다음과 같이 3가지 경우로 정리할 수 있고, 이 모든 결과를 앙상블했다.

(1) 5 fold cv + Y(MAE <= 2)를 144구간으로 나눈 Y18을 조합을 짬.

(2) 5 fold cv + Y(ALL)를 144구간으로 나눈 Y18을 조합을 짬.

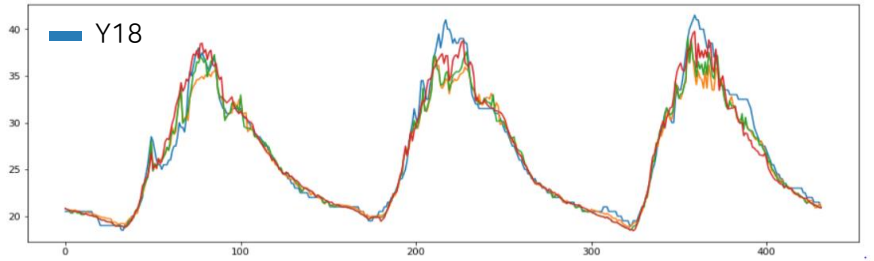
(3) 시간순 0.8, 0.2 + Y(MAE <= 2)를 8, 12 구간으로 나눈 Y18을 조합을 짬.

앙상블

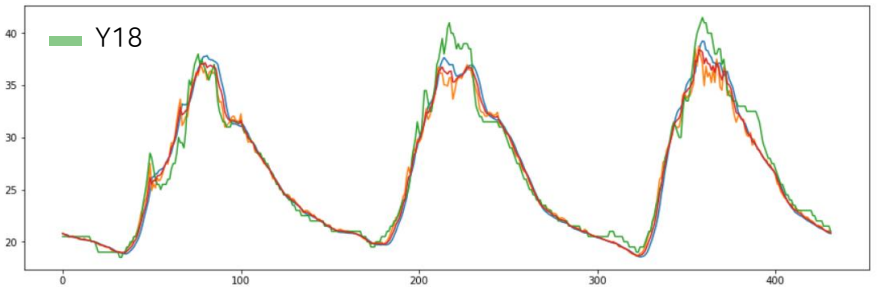


STEP4: 스무딩

예측값이 긴밀하게 연결되도록, **지수 평활법**을 이용하여 스무딩을 진행했다.
스무딩을 통해 성능이 미세하게 증가하였다.



스무딩 전 Y18의
Validation MAE : 1.63

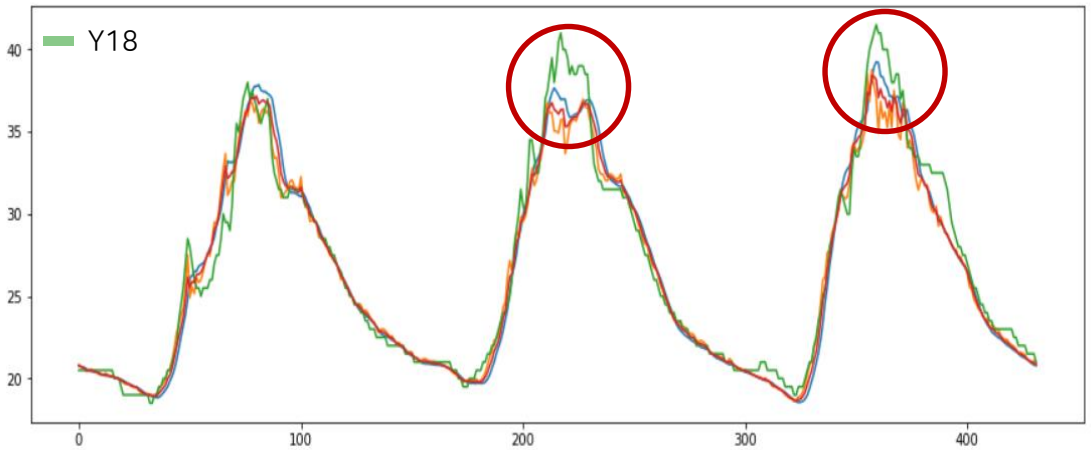


(스무딩 전 Y18, 스무딩 후 Y18)
앙상블의
Validation MAE :1.487
LB MAE : 1.403

STEP4: 잔차 보정

Train data에서 Y18이 존재하는 3일치 데이터를 Validation data으로 하여, RandomForest(depth=1or2)을 이용해서 잔차를 보정했다. Validation set에서는 적합이 잘 되지만, 기간이 멀어질수록 변수의 단위 문제 등으로 인해 적합이 잘 되지 않는 것으로 보인다.

(잔차보정 전 Y18, 잔차보정 후 Y18) 앙상블의 **LB MAE : 1.402**

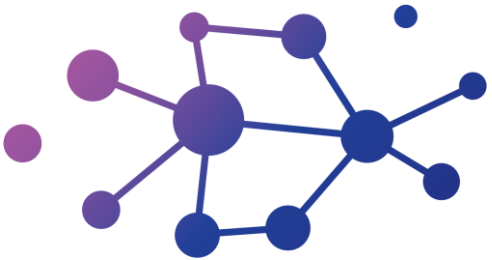


3. 결과 및 결언

1) 3일치의 Y18을 Validation으로 한 결과와 LB의 점수는 유사하였지만, Private에서 굉장한 성능 저하를 보였다. Boosting으로는 과대적합의 가능성이 더욱 커지게 되어, robust한 결과를 얻을 수 없었던 것 같다.

2) Feature engineering 또는 다른 모델을 통해 개별 Yi 자체를 더 잘 예측 할 수 있었다면 Y18을 더 잘 예측할 수 있지 않을까라는 아쉬움이 남았다. Y18과 가깝지 않아서 오차를 보이는 것도 있겠지만, 개별 Yi 자체를 확실하게 설명하고 있지 않아서 오차가 더욱 커진 것이 아닌가 하는 생각이 들었다.

3) Y18 없이 Y18을 예측하는 시도를 이번 대회에서 처음 해보았고, 이에 대한 경험적, 이론적 지식이 부족해서 이러한 케이스에 더욱 적합한 접근 방법을 찾지 못한 것 같아서 아쉬움이 남았다.



THANK YOU