기상청

multicampus

기상청 날씨 빅데이터 콘테스트

‘기상위성 자료를 활용한 여름철 자외선 산출기술 개발’

4조

Contents

대회 소개

- **주제**

  '기상위성 자료를 활용한 여름철 자외선 산출기술 개발'

- **데이터**

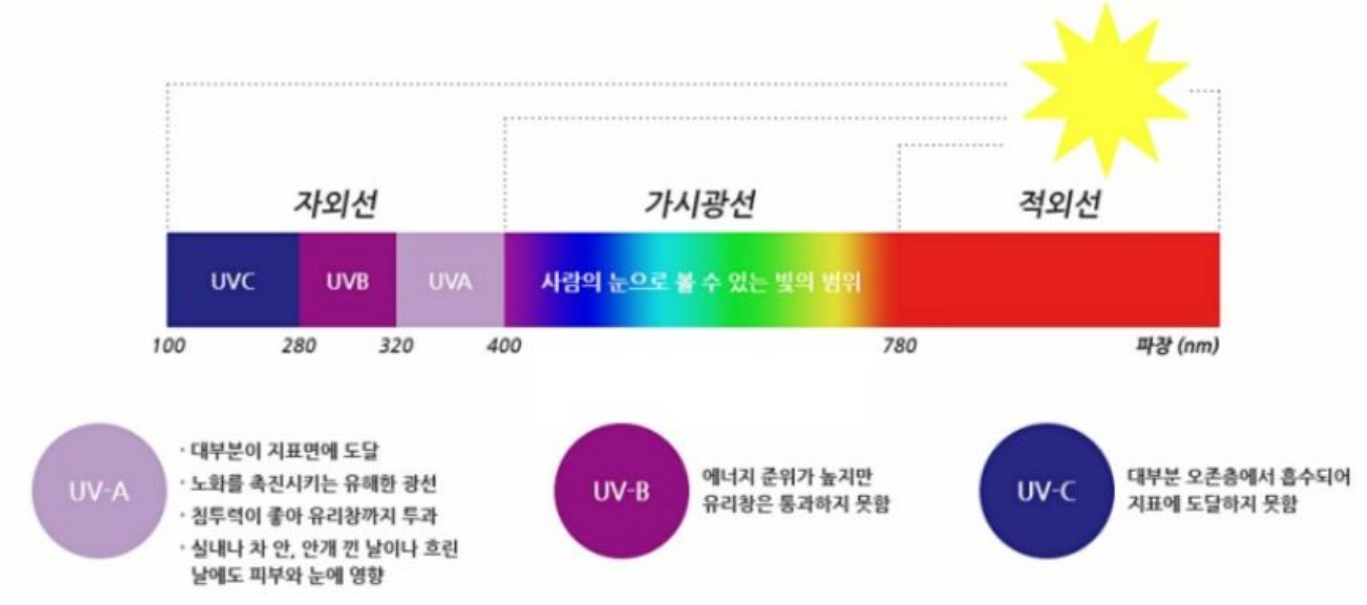
  자외선 데이터, 기상위성 데이터

- **모델 평가 지표**

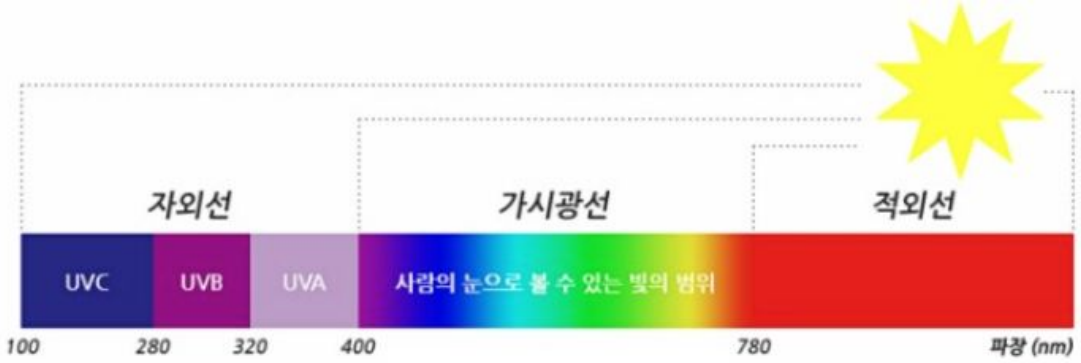
  RMSE(Root Mean Square Error: 평균제곱근오차)

- **주최 및 후원 기관**

자외선(UV: Ultraviolet rays)이란?



자외선(UV)이란?



UV-A
- 대부분이 지표면에 도달
- 노화를 촉진시키는 유해한 광선
- 침투력이 좋아 유리창까지 투과
- 실내나 차 안, 안개 낀 날이나 흐린 날에도 피부와 눈에 영향

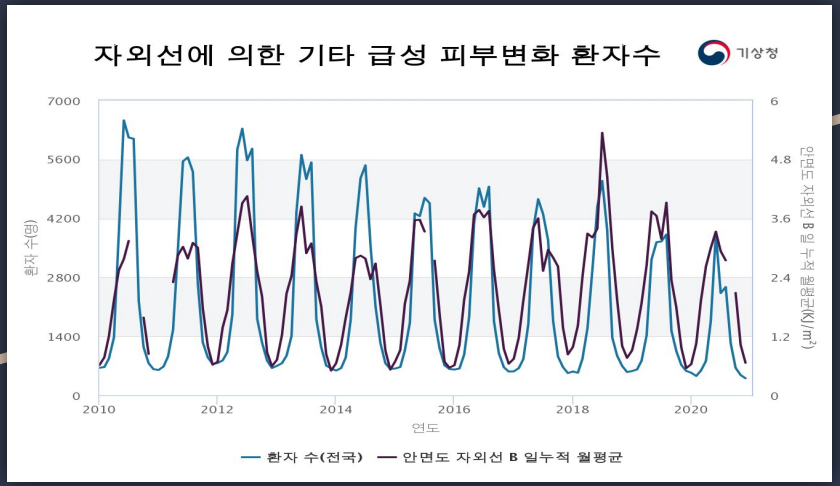
UV-B
에너지 준위가 높지만
유리창은 통과하지 못함

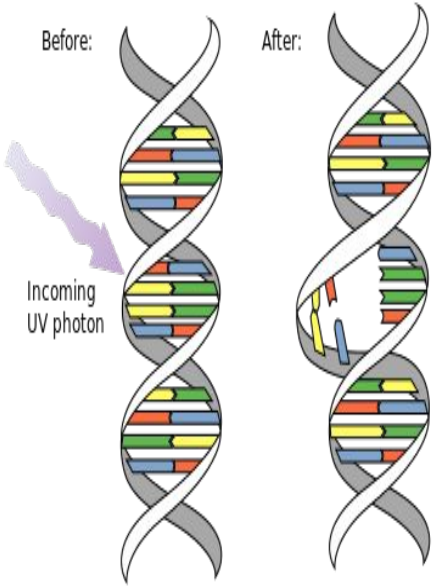
UV-C
대부분 오존층에서 흡수되어
지표에 도달하지 못함

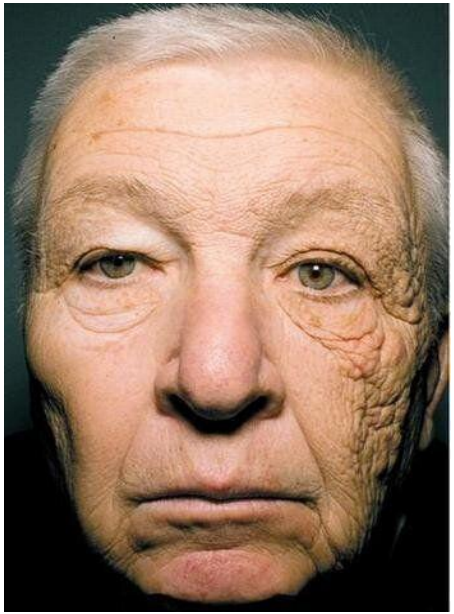
자외선의 유해성

"햇빛에 포함된 자외선을 비롯한 모든 자외선은 발암 유발원이다. IARC(국제 암 연구 기관)는 모든 종류의 자외선의 발암물질 분류를 1군, 즉 암 유발이 확인된 군으로 분류하고 있다."

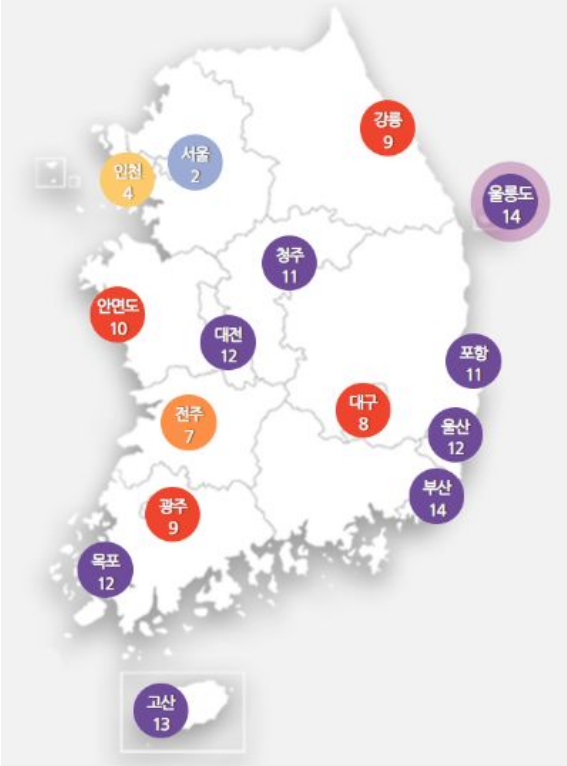
출처: 기상청







자외선 데이터



- 전국 **15**개 관측 지점

  고산센터, 강릉, 서울, 인천, 울릉도독도, 청주,

  안면도, 대전, 포항, 대구, 전주, 울산, 광주, 부산, 목포

- 기상청에서 서비스하는 자외선지수 단계

단계	지수	대책
낮음	~2	안전. 따로 대비하지않아도 무방
보통	3~5	모자, 선글라스 사용 권장
높음	6~7	1-2시간에 피부화상. 긴소매옷과 양산, 자외선 차단제 권장
매우 높음	8~10	1시간 내로 피부화상. 한낮에는 외출자제 권장
위험	11+	수십 분 정도로 피부화상. 가능한 한 실내활동.

기상위성 데이터를 활용한 자외선 산출, 왜 필요할까?

기상위성 데이터

- 천리안위성 2A호의 관측 데이터

YearMonthHourMinute	STN	Lon	Lat	SolarZA	SateZA
년/월/일/시간/분	지점번호	경도	위도	태양천정각	위성천정각
ESR	**LandType**	**band1**	**band2**	**band3**	**band4**
대기외 일사량	지면타입	파랑가시밴드	초록가시밴드	빨강가시밴드	식생가시밴드
band5	**band6**	**band7**	**band8**	**band9**	**band10**
눈/얼음채널	권운밴드	야간안개/하층운밴드	상층수증기밴드	중층수증기밴드	하층수증기밴드
band11	**band12**	**band13**	**band14**	**band15**	**band16**
구름상밴드	오존밴드	대기창밴드	깨끗한대기창밴드	오염된대기창밴드	이산화탄소(CO2)밴드

샘플 데이터

stn	date_time	lon	lat	uv	band1	band2	band3
159	2021-01-29 15:00	129.032	35.10468	1.9	0.11726	0.10436	0.09447

band4	band5	band6	band7	band8	band9	band10	band11	band12
0.09112	0.00121	0.0645	285.274	243.4783	251.0259	256.9	279.0249	250.9546

band13	band14	band15	band16	solarza	sateza	esr	height	landtype
281.0096	281.1729	279.7269	262.8753	61.64925	40.96396	4.49845	69.56	3

샘플 데이터

stn	date_time	lon	lat	uv	band1	band2	band3
159	2021-01-29 15:00	129.032	35.10468	1.9	0.11726	0.10436	0.09447

band4	band5	band6	band7	band8	band9	band10	band11	band12
0.09112	0.00121	0.0645	285.274	243.4783	251.0259	256.9	279.0249	250.9546

band13	band14	band15	band16	solarza	sateza	esr	height	landtype
281.0096	281.1729	279.7269	262.8753	61.64925	40.96396	4.49845	69.56	3

EDA

탐색적 데이터 분석

- **train 데이터 (2020~2021년)**

  : **1,578,960개**

  = 15(개 관측지점) * 6(10분 간격) * 24(시) * 731(366 + 365일)

- **test 데이터 (2019년 8월)**

  : **66,097개**

  = 15(개 관측지점) * 6(10분 간격) * 24(시) * 31(일)

EDA

결측치 확인

train data

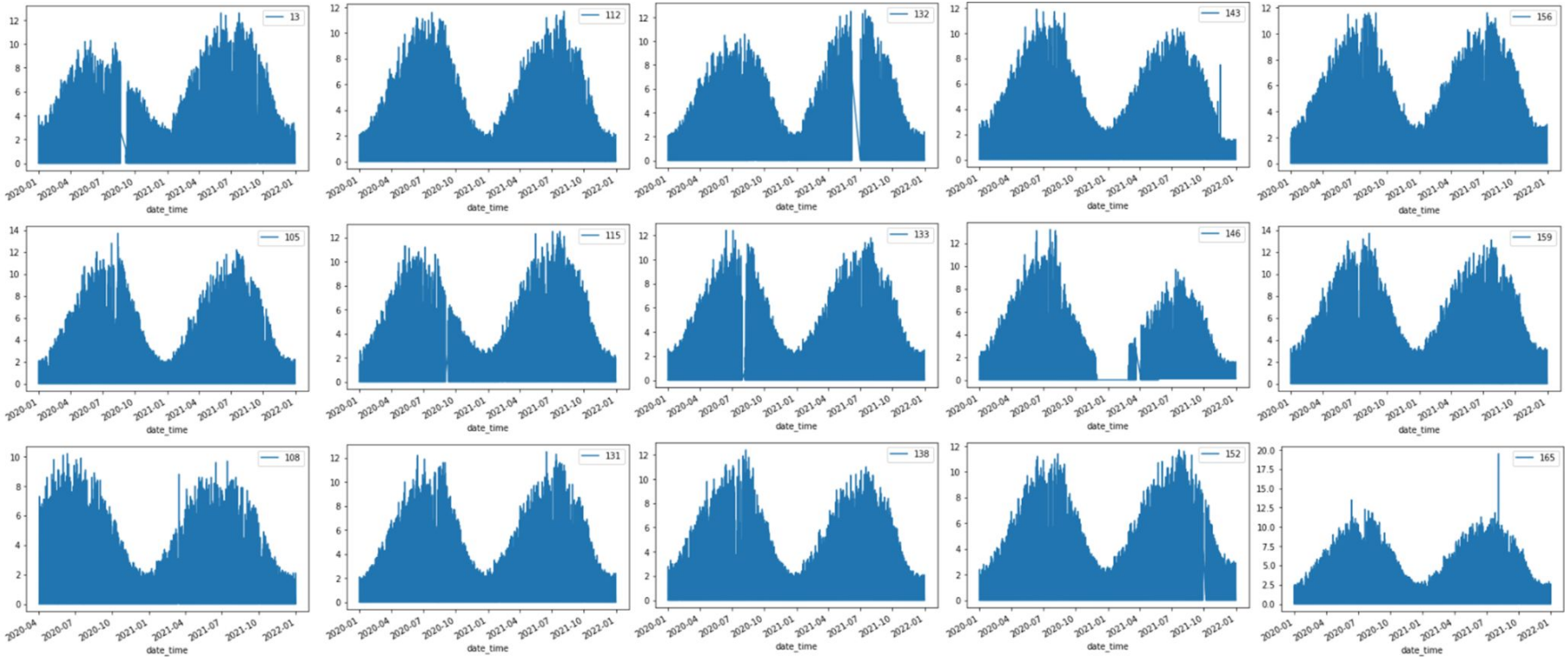
uv	52623	stn	0
band1	18060	date_time	0
band2	18060	lon	0
band3	18060	lat	0
band4	18060	solarza	0
band5	18060	sateza	0
band6	18060	esr	0
band7	18066	height	0
band8	18060	landtype	0
band9	18060		
band10	18060		
band11	18060		
band12	18060		
band13	18060		
band14	18060		
band15	18060		
band16	18060		

test data

uv	0	stn	0
band1	593	date_time	0
band2	593	lon	0
band3	608	lat	0
band4	593	solarza	0
band5	578	sateza	0
band6	578	esr	0
band7	293	height	0
band8	608	landtype	0
band9	608		
band10	608		
band11	593		
band12	578		
band13	608		
band14	593		
band15	578		
band16	578		

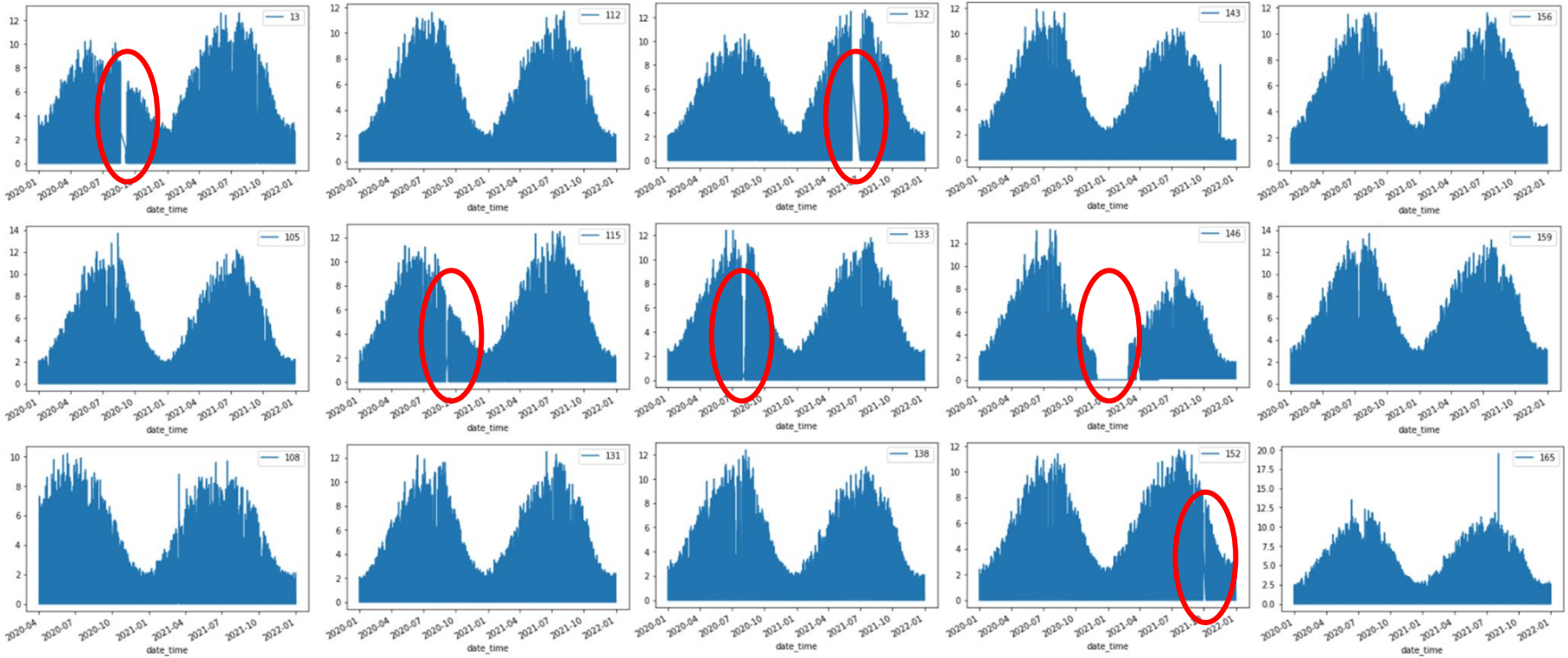
EDA

- UV 변수의 결측 구간



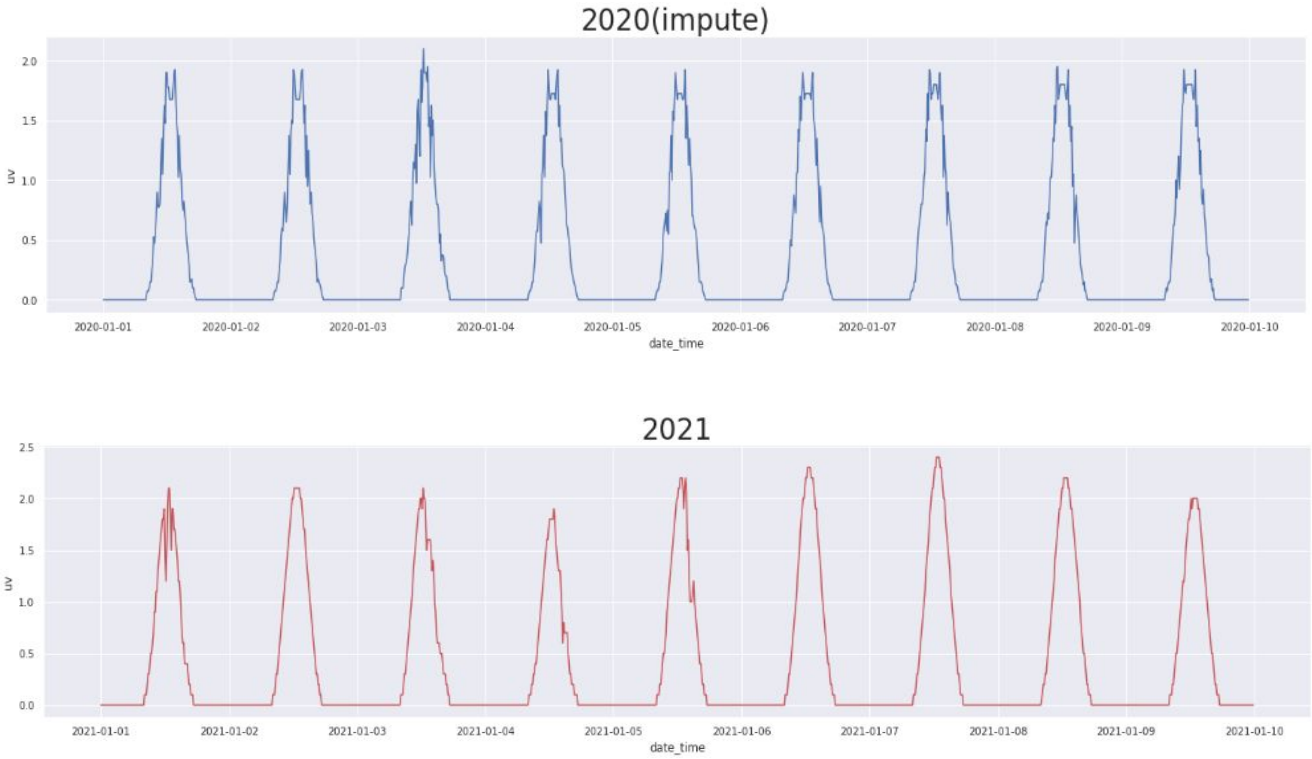
EDA

- UV 변수의 결측 구간



Data Preprocessing

결측치 대체
- sklearn KNN Imputer

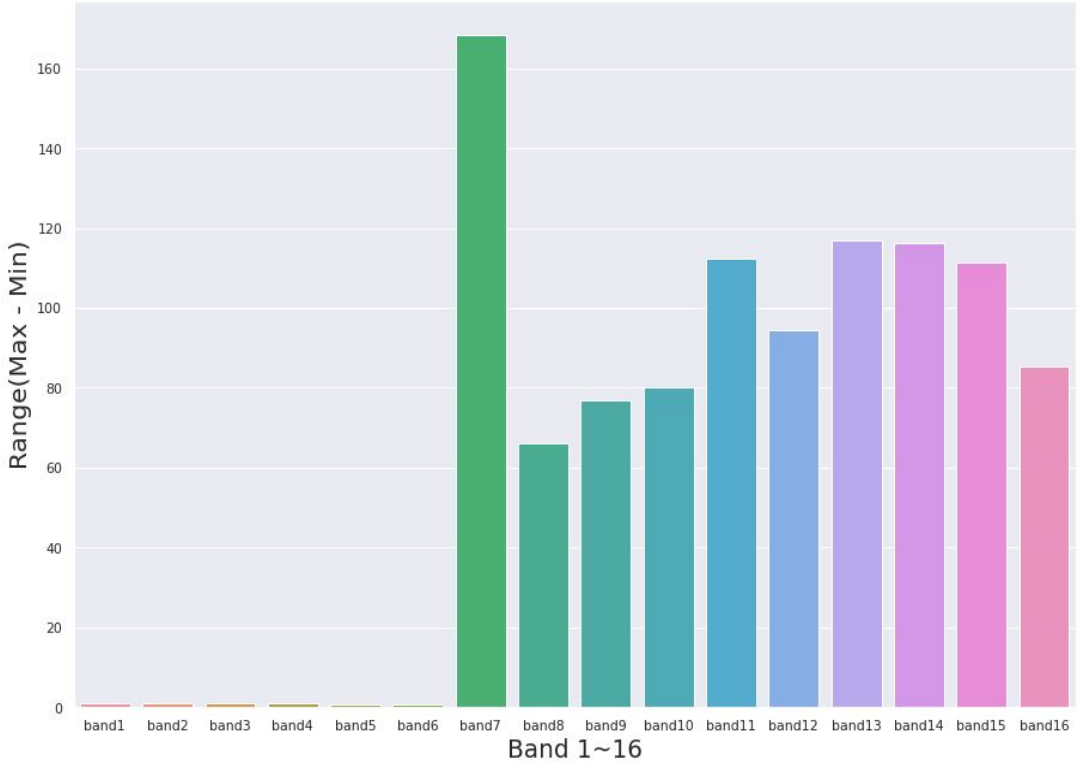


EDA & Preprocessing

변수 변환

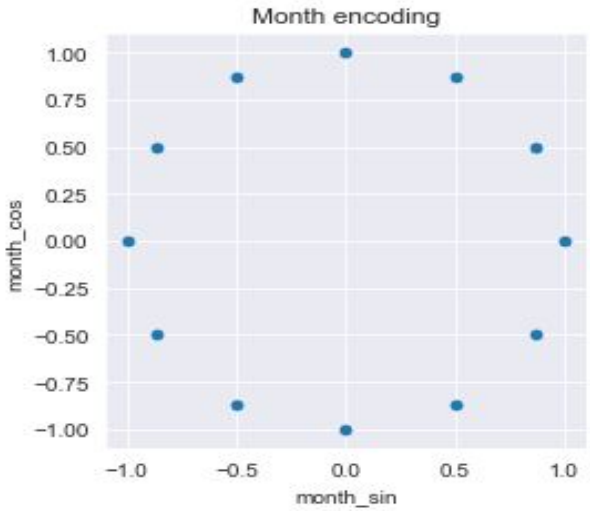
- sklearn MinMaxScaler

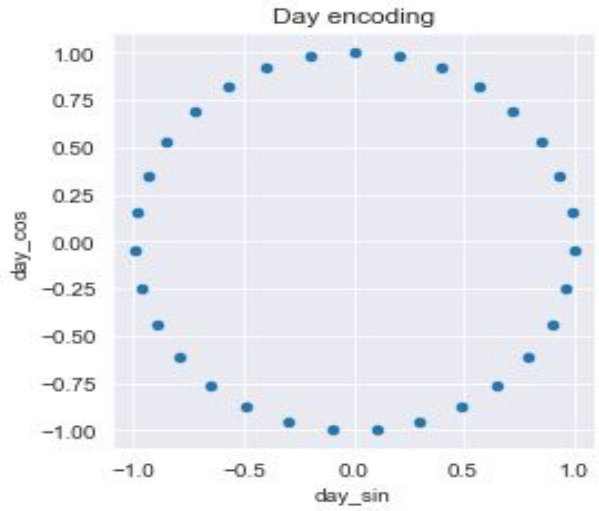
band	min	max	range
band1	-0.01	1.15	1.16
band2	-0.01	1.16	1.17
band3	-0.01	1.12	1.13
band4	-0.01	1.16	1.17
band5	-0.01	0.62	0.63
band6	-0.01	0.71	0.72
band7	152.77	321.23	168.46
band8	196.46	262.62	66.15
band9	195.26	272.13	76.87
band10	195.44	275.44	80.01
band11	194.54	307.00	112.45
band12	212.69	307.22	94.53
band13	194.49	311.30	116.80
band14	193.85	310.27	116.42
band15	193.88	305.17	111.30
band16	196.21	281.53	85.32

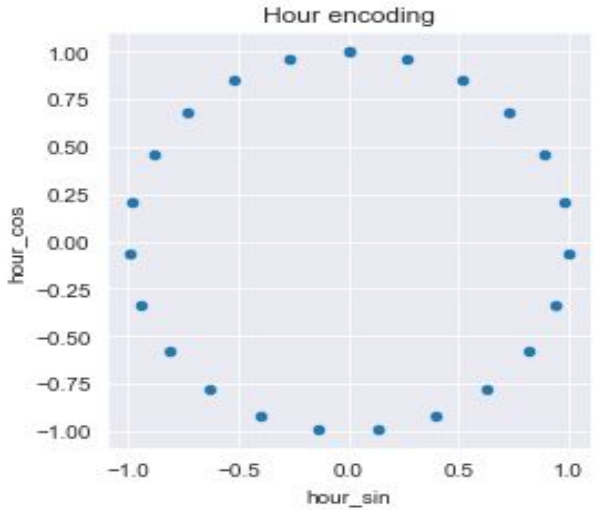


Data Preprocessing

변수 변환
- 월(month), 일(day), 시(hour) 정보 추출하여 sin, cos 변환







Data Preprocessing

파생변수 생성
- 태양천정각(solarza) 변수 활용

uv 근사값 도출

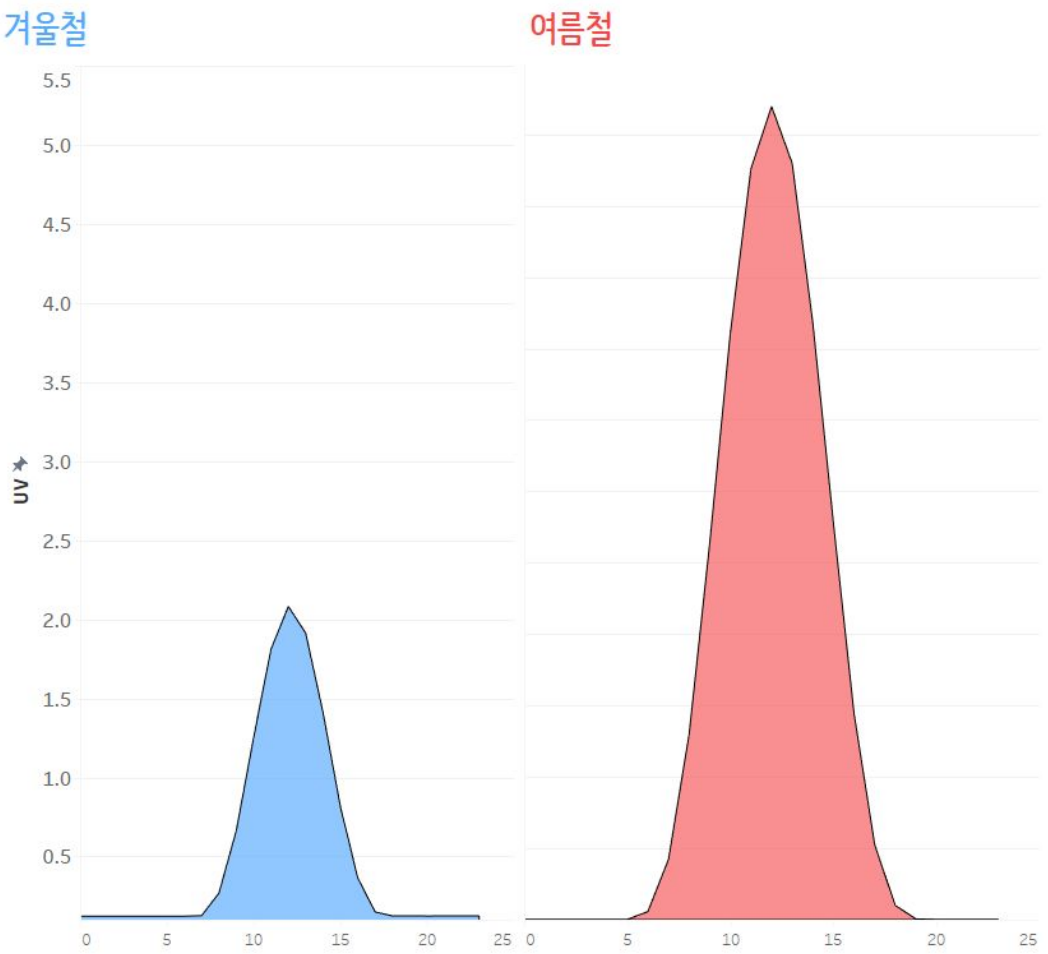
S = solarza(태양천정각)

d = delta

A = amplitude(진폭)

$$A \times \left[\frac{1}{2} \times \left\{\left|\cos\left(S \times 2 \times \frac{\pi}{360}\right) + d\right| + \cos\left(S \times 2 \times \frac{\pi}{360}\right) + d\right\}\right]^2$$

EDA & Preprocessing

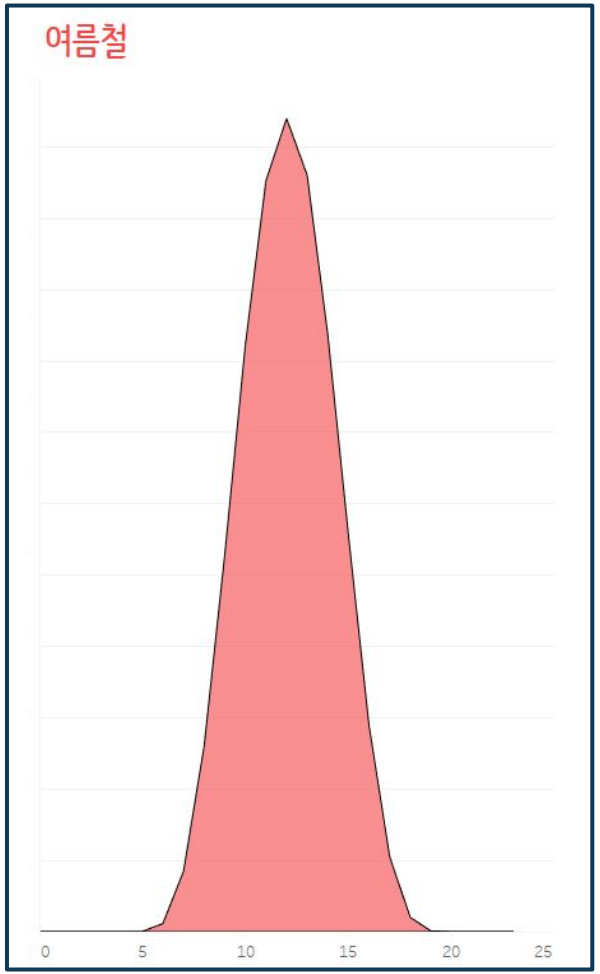


EDA & Preprocessing

전체 2020~2021년(24개월)

데이터 중 7, 8, 9월(여름철)만 사용

<u>즉, 데이터의 25%만 사용</u>

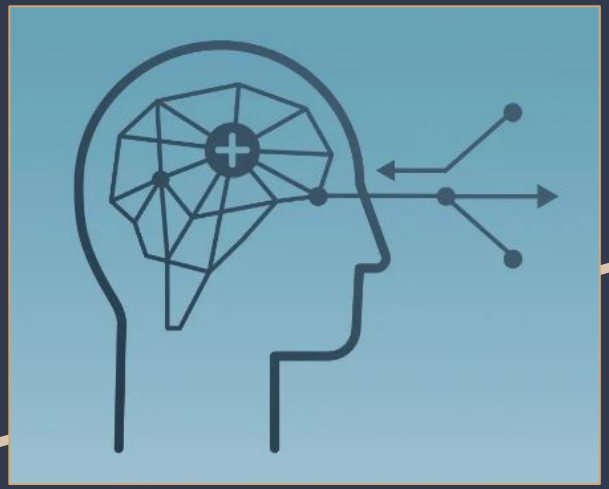


PYCARET

	RMSE
Light Gradient Boosting Machine	0.6938
Cat Boost Regressor	0.6939
Extra Trees Regressor	0.7054
Gradient Boosting Regressor	0.7099
Random Forest Regressor	0.7119
K Neighbors Regressor	0.8899
Bayesian Ridge	1.0312
AdaBoost Regressor	1.0326
Ridge Regression	1.0853
Decision Tree Regressor	1.0879

Deep Learning

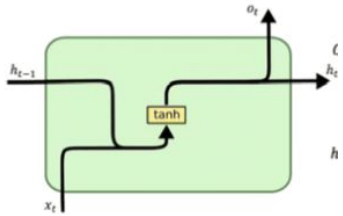
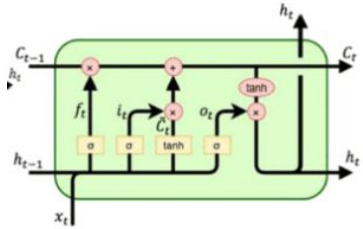
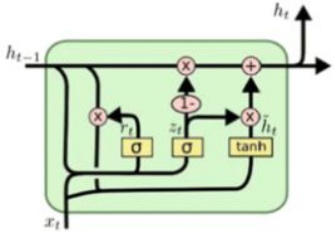
- Keras Sequential 모델



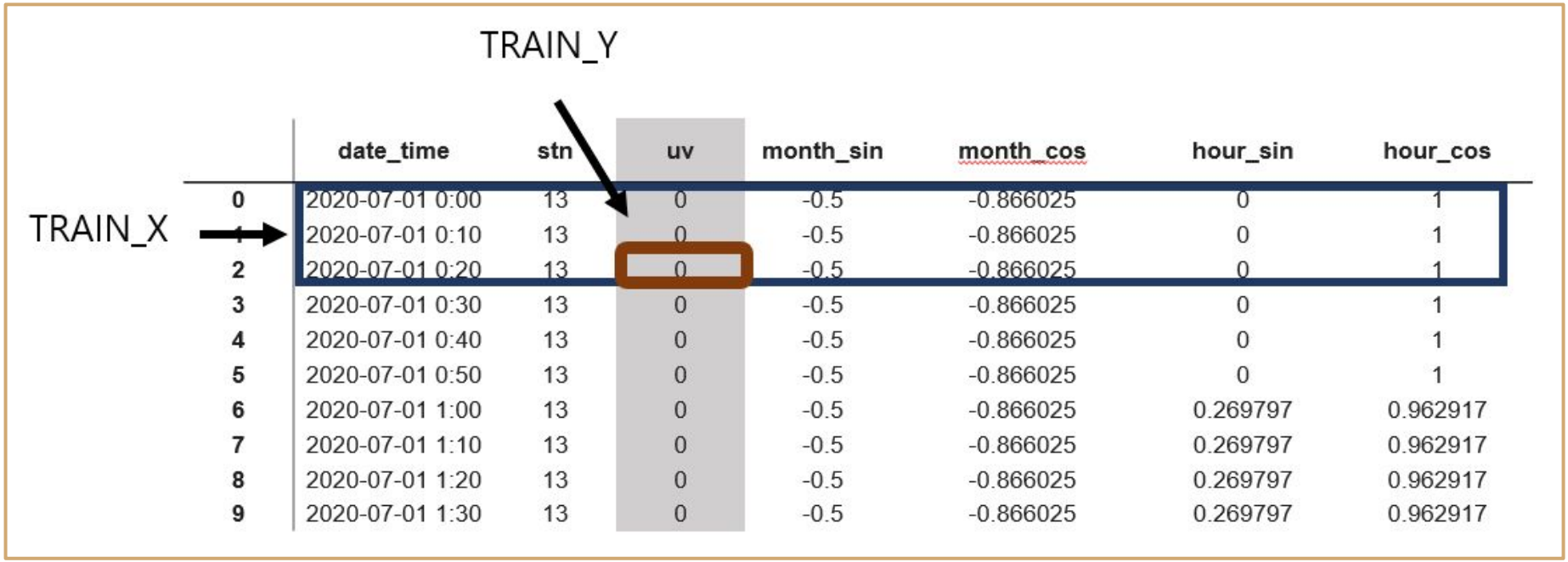
- Layers : **4**
- Parameters: **134,913**
- Optimizer = **Root Mean Squared Propagation**
- Learning rate: **0.001**
- Reduce on plateau로 **모델 개선 없을 시 학습률 0.8배 조정**

RMSE: **0.7223**

RNN

 Simple RNN	 LSTM	 GRU
시계열 데이터 학습에 적합	게이트가 추가된 **RNN.** 장기기억을 가능하게 함	**LSTM**의 개념을 유지하되, 매개 변수가 적고 계산량도 적음
0.6918	**0.6850**	**0.6791**

GRU - 데이터셋 구성



	date_time	stn	uv	month_sin	month_cos	hour_sin	hour_cos
0	2020-07-01 0:00	13	0	-0.5	-0.866025	0	1
1	2020-07-01 0:10	13	0	-0.5	-0.866025	0	1
2	2020-07-01 0:20	13	0	-0.5	-0.866025	0	1
3	2020-07-01 0:30	13	0	-0.5	-0.866025	0	1
4	2020-07-01 0:40	13	0	-0.5	-0.866025	0	1
5	2020-07-01 0:50	13	0	-0.5	-0.866025	0	1
6	2020-07-01 1:00	13	0	-0.5	-0.866025	0.269797	0.962917
7	2020-07-01 1:10	13	0	-0.5	-0.866025	0.269797	0.962917
8	2020-07-01 1:20	13	0	-0.5	-0.866025	0.269797	0.962917
9	2020-07-01 1:30	13	0	-0.5	-0.866025	0.269797	0.962917

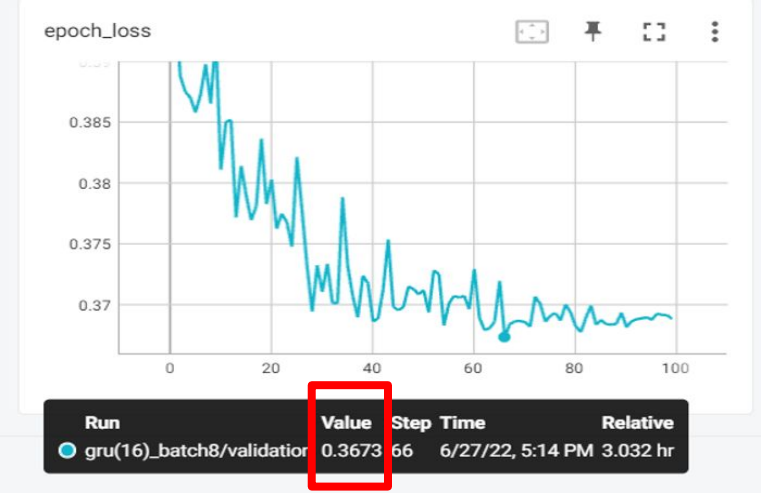
GRU

- Keras Tuner 사용하여 Hyperparameter 조정

- 학습

- ❏ Layers : **2**
- ❏ Unit: **16**
- ❏ Optimizer: **Adam**
- ❏ Learning rate: **0.0001**
- ❏ Epochs: **100**
- ❏ Reduce on plateau로 **모델 개선 없을 시 학습률 0.8배 조정**
- ❏ validation loss: **RMSE**

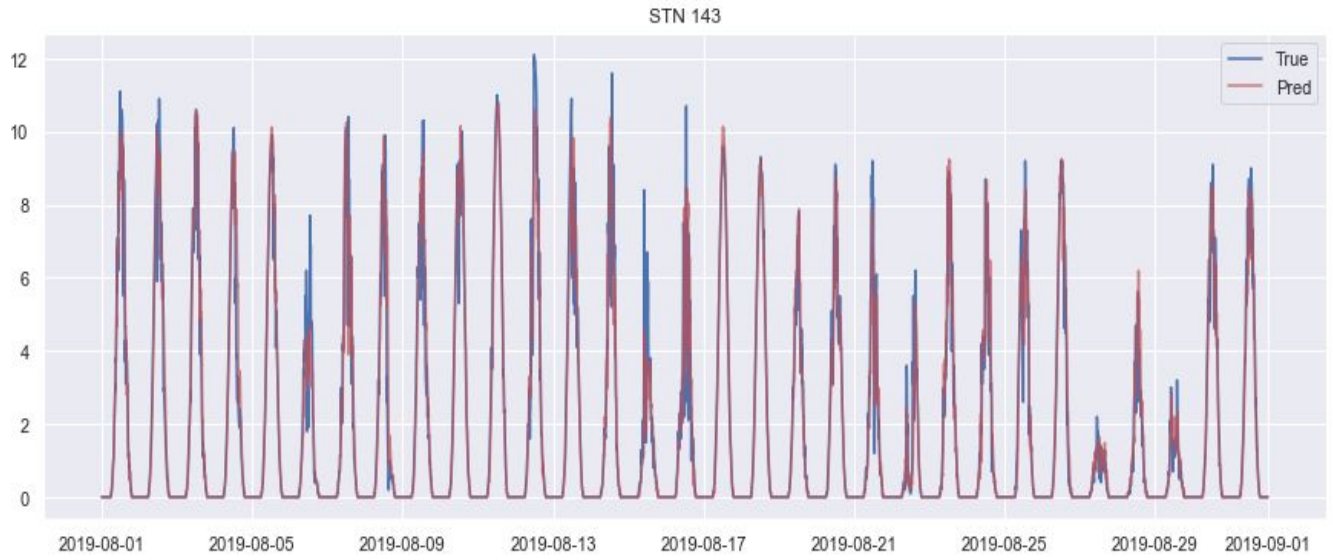
-> **validation loss 값이 가장 낮은 모델 저장**



GRU

- 학습 결과

예측값과 실제값 비교



Results

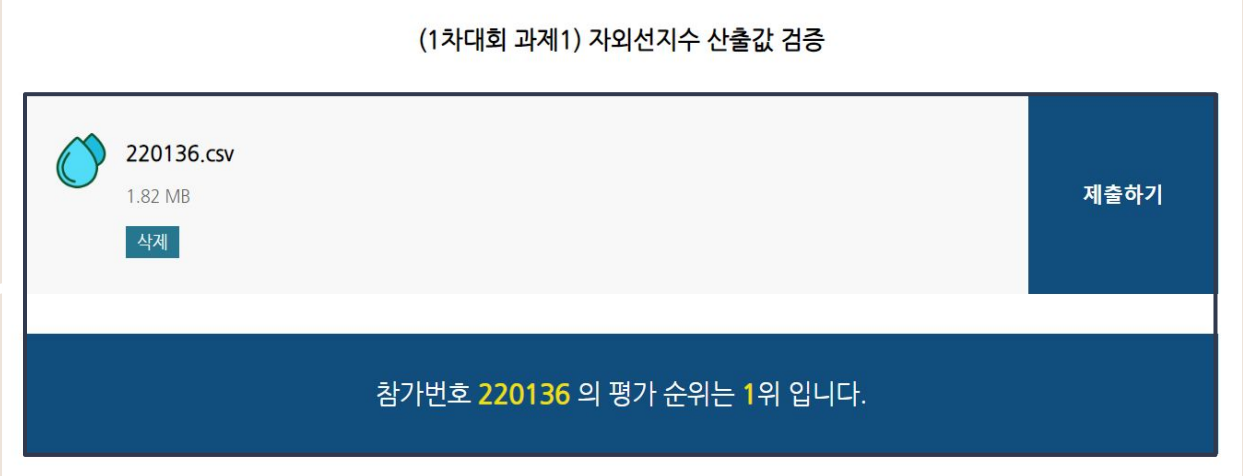
Model : **GRU**

최종 RMSE: **0.6791**

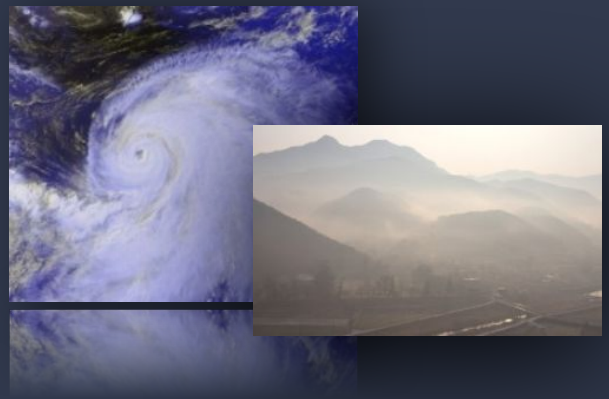
Results

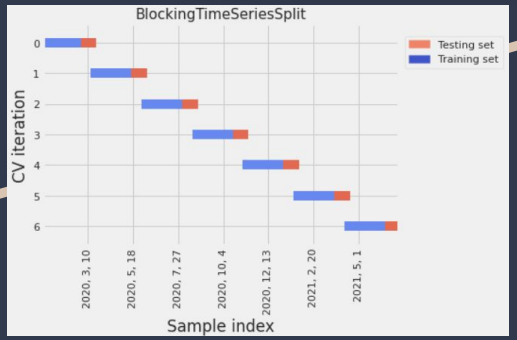
Model : **GRU**

최종 RMSE: **0.6791**



Discussion





1. 변수 추가

주어진 데이터 외 새로운 데이터 탐색 및 추가

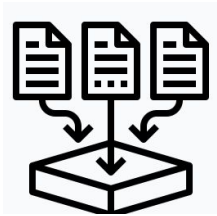
EX) 박무, 연무 현상 유무

EX) 암모늄, 질산염, 황산염 입자 등

2. 다양한 데이터 셋 및 검증 셋 구성

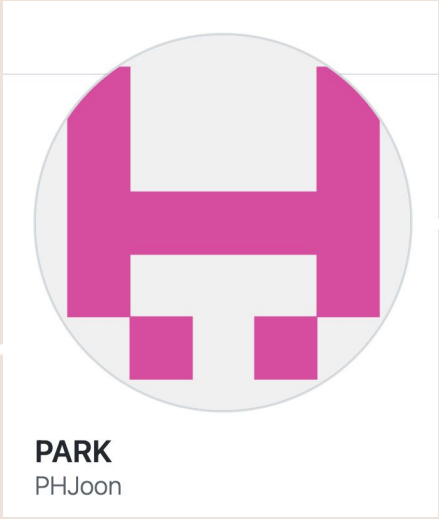
EX) 2022년도 UV에 대한 정보를 21년도에 학습데이터로 사용

EX) 시계열성을 반영한 검증셋

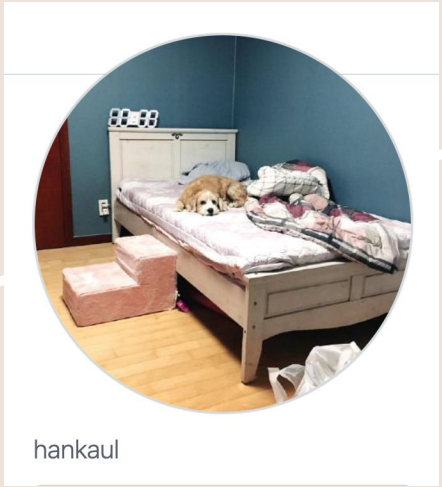


Q&A Time

BaeJjangE
BaeJjangE



PARK
PHJoon



hankaul

Inseo Jeon
eveinseojeon