

2021 POSTECH OIBC CHALLENGE

# 태양광 발전량 예측 경진대회

로제핌닭

# 목차

1 분석 목표

2 데이터 소개

3 분석 과정

4 분석 결과

5 한계점

### 가상 발전소의 다음 날 24시간 동안의 매시간 발전량을 입찰하여 예측 성과에 따른 인센티브 획득

- ❖ 대회 시간 동안의 실제 발전량 중, 시간당 발전량이 발전용량의 10%(2.0247MWh)이상 발전된 시간대에 대해서 10시, 17시에 입찰된 양과 실제 발전량의 차이를 발전용량으로 나눈 값을 10시, 17시 입찰 양에 대한 예측 오차로 산정
- ❖ 두 시간대 입찰 양에 대한 예측 오차의 평균을 최종 입찰 오차로 산정
  - 최종 입찰 오차가 6% 이하이면 해당 시간대의 발전량(kWh) X 4원의 인센티브 획득
  - 최종 입찰 오차가 6% 초과, 8% 이하이면 해당 시간대의 발전량(kWh) X 3원의 인센티브 획득
  - 최종 입찰 오차가 8% 초과이면 인센티브 없음

## 제공 데이터

발전소별 발전량

gens

221761 rows x 3 columns

기상 실측 데이터

weathers1, 2, 3

209339 rows x 13 columns

351232 rows x 9 columns

200982 rows x 7 columns

기상 예측 데이터

forecasts1, 2, 3

6557869 rows x 14 columns

11247269 rows x 10 columns

563952 rows x 11 columns

발전소별 인근  
ASOS 번호

pv\_sites

24 rows x 10 columns

## 사용 데이터

발전소별 발전량

gens

221761 rows x 3 columns

기상 실측 데이터

weathers1, 2

209339 rows x 13 columns

351232 rows x 9 columns

발전소별 인근  
ASOS 번호

pv\_sites

24 rows x 10 columns

기상청 기상자료개방포털

일사량 데이터

new\_insol

250798 rows x 6 columns

## 데이터 병합

### 발전량과 기상 실측 데이터 병합

각 발전소별로 가장 가까운 기상관측소의 기상 실측 데이터만을 사용

발전소 id 1~23 : weathers1의 id 1~23과 병합

발전소 id 24 : weathers3의 id 24가 가장 가깝지만 결측치가 많으므로 weathers2의 id 10과 병합

	id	capacity	wth1_id	wth1_dist	wth2_id	wth2_dist	wth3_id	wth3_dist	asos_station	asos_dist
0	1	811	1	713	1	2620	1	2616	131	25836
1	2	839	2	521	2	3137	2	2061	252	5552
2	3	819	3	236	3	3015	3	2858	119	13163
3	4	819	4	1061	4	2351	4	2198	156	6842
4	5	838	5	633	5	2297	5	2382	108	10633
5	6	838	6	845	6	12850	6	3068	165	14634
6	7	838	7	1597	2	8368	7	2043	252	6771
7	8	919	8	133	7	2522	8	466	244	2251
8	9	838	9	1415	8	4934	9	2059	129	2607
9	10	838	10	190	9	22494	10	1525	174	16968
10	11	839	11	355	10	5562	11	2391	279	17267
11	12	819	12	574	11	2026	12	2014	108	13931

12	13	999	13	975	12	16066	13	1935	93	13185
13	14	945	14	78	13	11304	14	1221	156	9336
14	15	839	15	429	14	7929	15	2054	279	8111
15	16	838	16	551	6	19213	16	2215	268	18020
16	17	838	17	844	15	10285	17	2586	260	26012
17	18	838	18	1169	16	11039	18	1253	156	19369
18	19	832	19	735	17	10778	19	2748	254	11515
19	20	818	20	618	2	11707	20	1312	172	13196
20	21	803	21	182	1	8358	21	711	131	16148
21	22	803	22	378	18	1536	22	1992	131	2878
22	23	800	23	1341	19	8190	23	2688	203	31227
23	24	839	11	6925	10	1712	24	985	279	11120

## 데이터 병합

### 일사량 데이터 병합



“태양광 발전량에 1차적으로 영향을 미치는  
기상요소는 **지면 도달 일사량**이다. 1)”



**일사량 변수 추가**



**기상청 기상자료개방포털**

데이터 - 기상관측 - 지상 - 종관기상관측(ASOS)

일사량 데이터 다운로드

2020.06.01 - 2021.07.31 관측 데이터

[ 각 발전소별로 가장 가까운 ASOS 지점의 일사량 데이터와 병합 ]

1) 이순환, 국지 기상 요소에 의한 태양광 발전량 변동특성에 관한 연구,  
Journal of Environmental Science International(한국환경과학회지), Vol.23, No.11 , pp.1947

## 데이터 탐색

하루동안 발전량이 0인 날이 존재

이상치로 판단?

year	month	day	id	Amount
2020	12	13	23	0
2021	1	8	7	0
2021	1	18	2	0
"	"	"	7	0
"	"	"	23	0

Case1. 2020년 12월 13일 이천



13일

평균기온:-2.2℃  
최고기온:0.8℃  
최저기온:-6.1℃  
평균운량:6.1  
일강수량:3.7mm

Case2. 2021년 1월 18일 영광, 이천

오늘 기상 정보		서울	인천	수원	춘천
최저/최고 기온(℃)		-4/3	-2/3	-4/3	-7/2
관측하는 강수 확률(오전/오후, %)		(70/60)	(60/60)	(70/60)	(80/70)
강릉	청주	세종	대전	전주	광주
-2/5 (60/70)	-2/5 (70/60)	-4/4 (70/60)	-3/5 (70/60)	-2/6 (70/60)	-2/7 (60/10)
목포	대구	포항	울산	부산	제주
0/7 (30/0)	-4/5 (60/60)	-2/5 (10/60)	-3/7 (0/10)	-2/8 (10/10)	4/10 (30/10)

18일

평균기온:0.9℃  
최고기온:6.3℃  
최저기온:-2.6℃  
평균운량:8.4  
일강수량:3.2mm

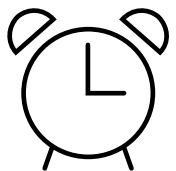
18일

평균기온:-3.6℃  
최고기온:1.5℃  
최저기온:-7.5℃  
평균운량:5.6  
일강수량:1.3mm

발전량이 0인 날에 해당 지역에는 눈이 내렸으며  
운량을 통해 구름이 많이 끼었음을 확인  
존재할 수 있는 값이라고 판단

\*운량 기준 6≤운량≤8 : 구름많음, 9≤운량≤10 : 흐림

## 1. 시간변수 통일



weathers데이터의 time

2020-05-31T15:01:00+00:00  
2020-05-31T16:01:00+00:00  
2020-05-31T17:01:00+00:00  
...  
2021-06-30T12:00:00+00:00  
2021-06-30T13:00:00+00:00  
2021-06-30T14:00:00+00:00

시간대 UTC +0  
str형식의 데이터



2020-06-01 00:01:00  
2020-06-01 01:01:00  
2020-06-01 02:01:00  
...  
2021-06-30 21:00:00  
2021-06-30 22:00:00  
2021-06-30 23:00:00

한국 시간대인 UTC+9  
datetime형식으로 변환



hour	day	month	year
0	1	6	2020
1	1	6	2020
2	1	6	2020
...	...	...	...
21	30	6	2021
22	30	6	2021
23	30	6	2021

'분'을 기준으로 반올림 후  
year, month, day, hour 변수 생성



## 2. 풍속변수 통일

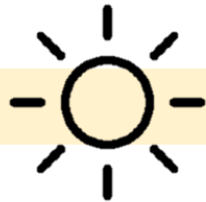


weathers1 데이터의 wind\_spd  
풍속 단위 km/h

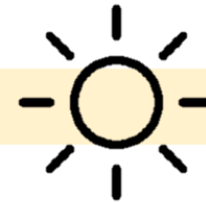


weathers2 데이터의 wind\_spd  
풍속 단위 m/s **X 3.6**  
km/h

## 3. 일사량변수 통일



데이터셋에 포함된 insolation  
일사량 단위 MJ **X 278**  
W/m<sup>2</sup>



예측에 사용할 insolation  
일사량 단위 (W/m<sup>2</sup>)

## 4. 이상치 제거

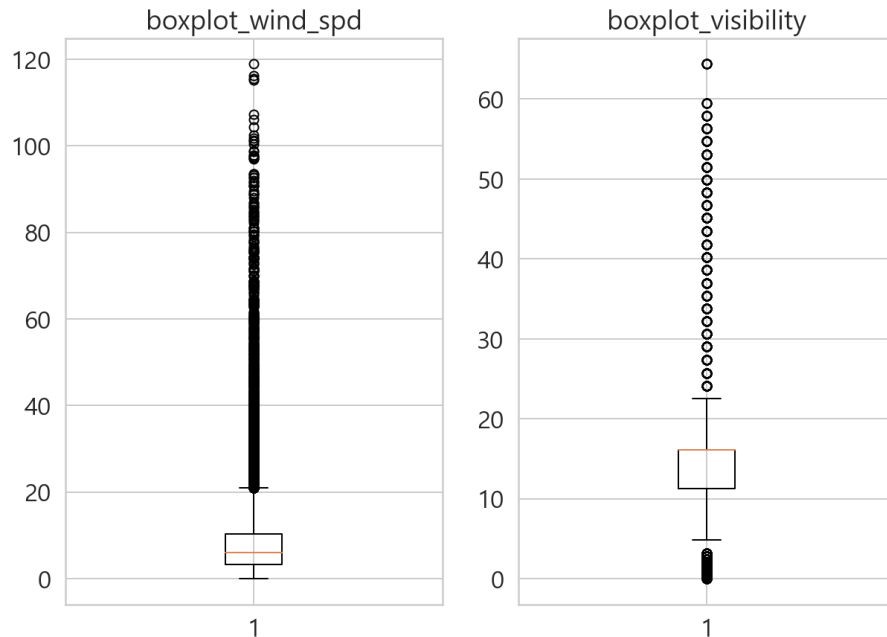
분포가 불균등하며 이상치가 다수 존재하는 변수 **wind\_spd, visibility**

IQR 방법으로 이상치 제거

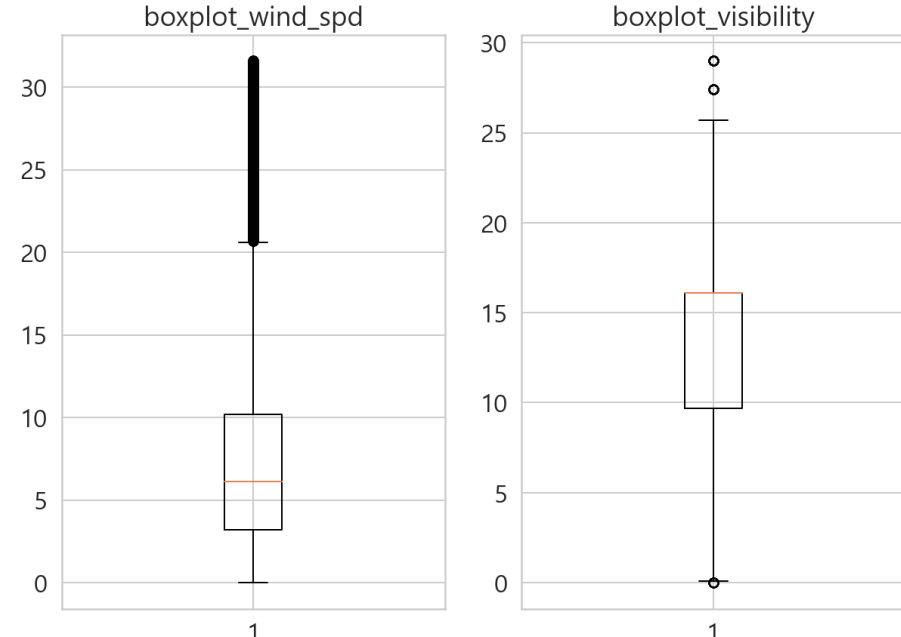
### IQR 방법

$IQR = Q3(75\%) - Q1(25\%)$   
Q3+3\*IQR과 Q1-3\*IQR을 벗어나는  
값을 제거하는 것

이상치 제거 전



이상치 제거 후



## 4. 이상치 제거

# 흐리지 않고 비가 오지 않으며 일사량이 200 이상인데  
발전량이 10 미만인 경우

```
amount<10  
7<hour<19  
precip_1h==0  
cloudiness<90  
insolation>=200
```

발전량이 존재해야하는 시간대와 기상상황인데  
이에 부합하지 않는 이상치라고 판단하여 제거

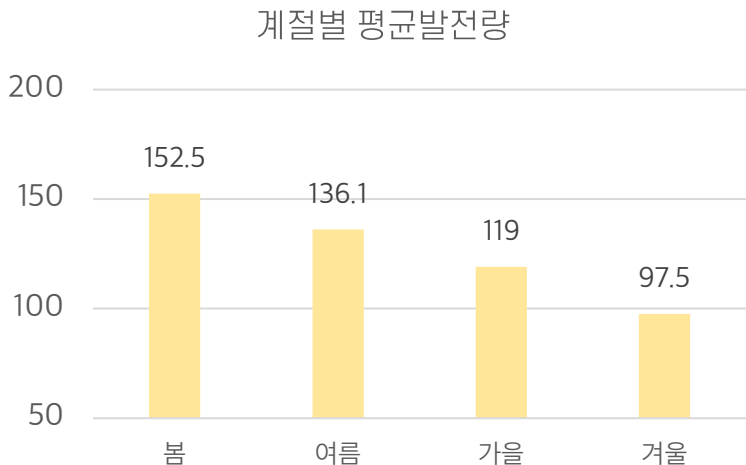
# 일사량이 0인데 발전량이 100 이상인 경우

```
Amount>=100  
insolation==0
```

발전량에 많은 영향을 주는 일사량이 0임에도 불구하고  
발전량이 특정량을 넘어서는 것은  
이해하기 어렵다고 판단하여 제거

## 5. 파생변수 생성

### 1) summer



그래프상에서 계절별 평균 발전량의 눈에 띄는 차이 존재  
예측하려는 계절은 여름이므로 여름 여부를 나타내는 'summer' 변수 생성

**summer변수가  
유의미한 변수인지 확인**

여름과 다른 계절의 평균 발전량이  
다르다고 할 수 있는지 검정

**ANOVA**

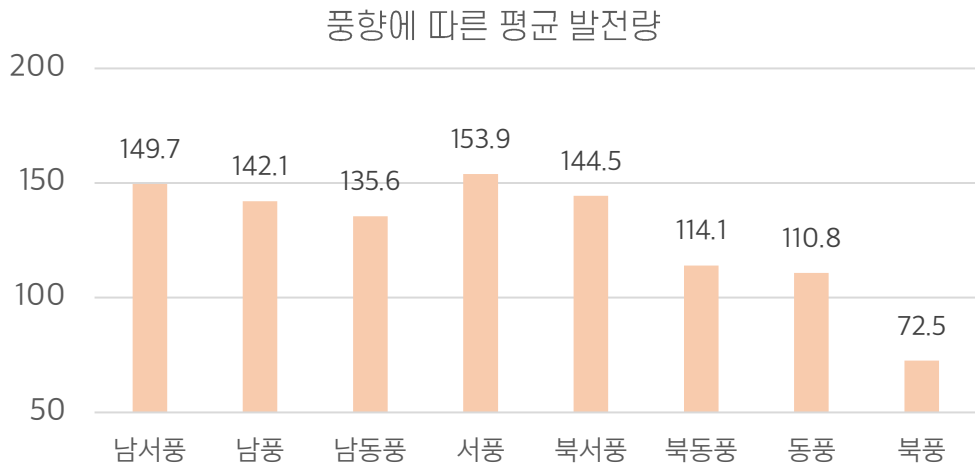
검정통계량  
215.80  
p-value  
7.85e-49

**T-test**

검정통계량  
14.83  
p-value  
9.64e-50

### 2) wind\_dir\_cat


각도	풍향	각도	풍향
0°, 360°	북풍	0°초과 90°미만	북동풍
90°	동풍	90°초과 180°미만	남동풍
180°	남풍	180°초과 270°미만	서풍
270°	서풍	270°초과 360°미만	북서풍



풍향에 따른 발전량의 차이가 존재해 범주화한 풍향변수 사용

## 6. 더미변수 생성

모델의 성능을 높이기 위한 더미변수 생성  
hour변수만 더미변수로 채택했을 때의 성능이 가장 뛰어남



hour\_1 ~ hour\_23

23개의 변수 추가적으로 생성

## 7. 최종 데이터셋

변수명	변수설명	변수타입	변수명	변수설명	변수타입
id	발전소ID	category	ceiling	최저운고(m)	int
year	년	category	precip_1h	강수량(mm)	float
month	월	category	capacity	발전설비용량(kW)	int
day	일	category	asos_station	종관기상관측소(ASOS) 지점번호	int
temperature	기온(°C)	float	insolation	일사량	float
humidity	습도(%)	int	summer	여름 여부	int
dew_point	이슬점온도(°C)	float	amount	발전량	float
wind_spd	풍속(km/h)	float	wind_dir_cat	풍향 범주	category
uv_idx	지수(0~12)	int	hour_1	1시 여부	unit8
visibility	시경(km)	float	...	...	...
cloudiness	운량(%)	int	hour_23	23시 여부	unit8

## 데이터 분리

전처리를 모두 마친 데이터를 8:2의 비율로 학습, 테스트 데이터 셋으로 분리

x\_train

id	year	month	day	temperature	humidity	dew_point	wind_spd	uv_idx	visibility	...	hour_14	hour_15	hour_16	hour_17	hour_18	h
24413	3	2020	12	27	2.6000000	86	0.5000000	11.2000000	0	8.0000000	...	0	0	0	0	0
214184	22	2021	5	12	21.7000000	25	0.8000000	9.3000000	7	19.3000000	...	0	0	0	0	0
182101	19	2021	1	20	6.0000000	38	-7.4000000	6.8000000	3	16.1000000	...	0	0	0	0	0
191669	20	2021	1	8	-15.9000000	83	-18.1000000	2.9000000	0	20.9000000	...	0	0	0	0	0
139149	15	2020	8	8	24.6000000	94	23.7000000	8.8000000	0	6.4000000	...	0	0	0	0	0
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
176963	19	2020	6	6	21.6000000	73	16.5000000	11.1000000	0	9.7000000	...	0	0	0	0	0
117952	13	2020	6	3	18.4000000	77	14.3000000	3.7000000	1	9.7000000	...	0	0	0	0	0
173685	18	2021	3	22	6.2000000	54	-2.3000000	8.3000000	0	20.9000000	...	0	0	0	0	0
43567	5	2020	12	6	4.0000000	60	-3.1000000	7.9000000	0	16.1000000	...	0	0	0	0	1
199340	21	2020	10	5	11.1000000	66	5.0000000	3.7000000	0	16.1000000	...	0	0	0	0	0

184264 rows × 41 columns

y\_train

```
24413    0.0000000
214184   468.0686997
182101   577.0320854
191669    0.0000000
139149    0.0000000
...
176963    0.0596416
117952   42.3975586
173685   14.7962084
43567     0.6094548
199340    0.0000000
Name: amount, Length: 184264, dtype: float64
```

학습 데이터셋

x\_test

id	year	month	day	temperature	humidity	dew_point	wind_spd	uv_idx	visibility	...	hour_14	hour_15	hour_16	hour_17	hour_18	h
1194	1	2020	7	22	23.8000000	87	21.5000000	4.0000000	0	11.3000000	...	0	0	0	0	0
100380	11	2020	9	1	28.7000000	68	22.2000000	14.8000000	3	16.1000000	...	0	0	0	0	0
80202	9	2020	8	10	26.2000000	87	23.9000000	10.4000000	0	16.1000000	...	0	0	0	0	0
176885	19	2020	6	3	26.5000000	63	18.9000000	9.6000000	2	12.9000000	...	0	1	0	0	0
3078	1	2020	10	12	18.1000000	81	14.7000000	2.3000000	2	9.7000000	...	0	0	0	0	0
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
75536	8	2021	3	27	7.3000000	55	-1.3000000	0.7000000	0	19.3000000	...	0	0	0	0	0
18721	2	2021	6	22	31.3000000	37	15.1000000	5.4000000	9	20.9000000	...	0	0	0	0	0
145573	15	2021	5	17	16.0000000	91	14.6000000	1.4000000	0	4.8000000	...	0	0	0	0	0
34796	4	2021	1	21	10.1000000	60	2.6000000	2.9000000	2	16.1000000	...	0	0	0	0	0
146476	15	2021	6	24	19.8000000	80	16.3000000	3.7000000	3	19.3000000	...	0	0	0	0	0

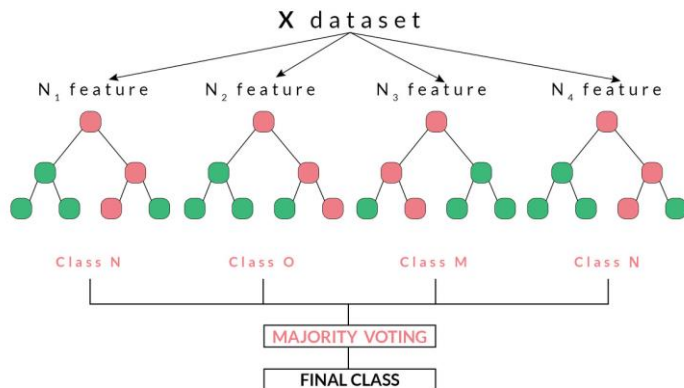
46066 rows × 41 columns

테스트 데이터셋

y\_test

```
1194    1.4057316
100380   447.0283876
80202     0.0680194
176885   434.0116197
3078    145.0066170
...
75536    0.0000000
18721    563.9422687
145573    0.7390592
34796    184.8319265
146476    315.1415470
Name: amount, Length: 46066, dtype: float64
```

## 모델링



### 랜덤 포레스트

- 부트스트래핑(bootstrapping) 분할 방식을 통해 데이터 세트를 분리한 후 학습
- 의사결정나무에 비해 과적합 위험이 낮음
- 소요시간 대비 성능이 뛰어남

### 하이퍼 파라미터 튜닝

사이킷런에서 제공하는 GridSearchCV 함수를 이용하여 하이퍼 파라미터 튜닝 진행

n\_estimator  
300

min\_samples\_split  
12

min\_samples\_leaf  
2

max\_features  
0.6

## 모델 성능

RMSE

42.8229

실제 값과 예측 값의 차이를 제곱해 평균한 MSE 값에 루트를 씌운 것

$$\text{RMSE} = \sqrt{\frac{1}{n} \sum_{i=1}^n (Y_i - \hat{Y}_i)^2}$$

MAE

16.3806

실제 값과 예측 값의 차이를 절댓값으로 변환해 평균한 것

$$\text{MAE} = \frac{1}{n} \sum_{i=1}^n |Y_i - \hat{Y}_i|$$

R-squared

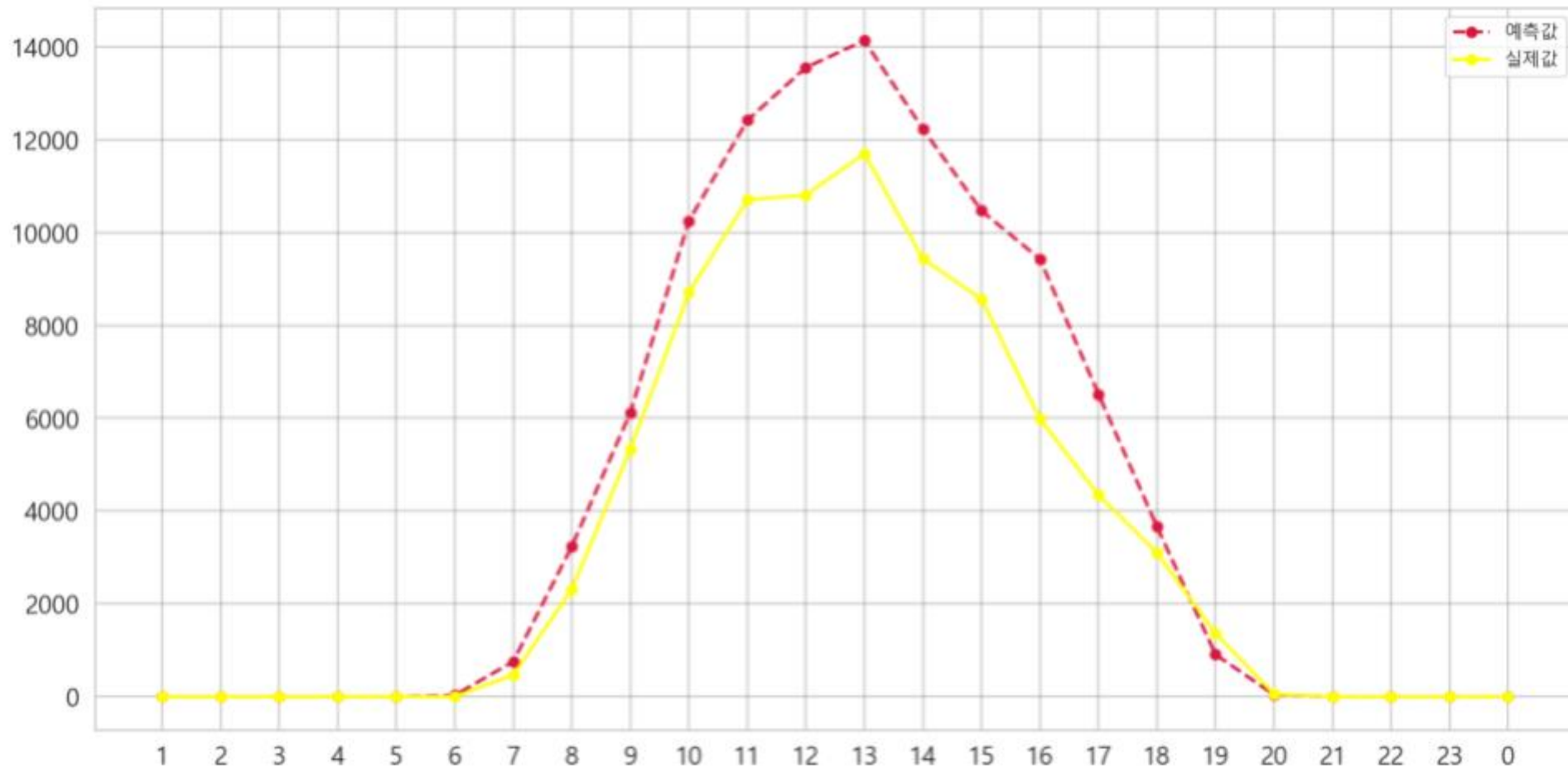
0.9518

실제 값의 분산 대비 예측 값의 분산 비율  
1에 가까울수록 예측 정확도가 높음

$$R^2 = \frac{\text{예측값 Variance}}{\text{실제값 Variance}}$$



## 실제 발전량과의 비교 – 8월 19일



## 오차율 - 8월 19일



## 1. 사용 데이터의 부정확성

각 발전소에서 가장 가까운 ASOS에 일사량 관측 정보가 존재하지 않는 경우가 있어 인근의 다른 ASOS 데이터를 이용하였다. 실제 해당 발전소의 일사량과 분석에 활용한 일사량이 다를 수 있다.

## 2. 기상 예측데이터의 부정확성

태양광 발전량 예측은 전날 기상 예측데이터를 바탕으로 이루어지므로 예측데이터가 정확하지 않다면 예측값의 정확성에도 큰 영향을 줄 수 있다.

# 감사합니다

로제핌닭