BDA & 동서발전 기상데이터 공모전

EDA에 근거한 풍력 발전 예측 모델링
: 속도, 방향, 온도를 활용한 분석

TEAM 선데이

60Hertz wai B.D.A BIG DATA ANALYSIS EWP 국립기상과학원 KIER 한국에너지기술연구원 KOREA INSTITUTE OF ENERGY RESEARCH

Contents

01

INTRO

02

EDA

- 속도/방향/온도

03

Data 전처리

- 경주풍력과 영광풍력
- 데이터 병합
- 변수 drop

04

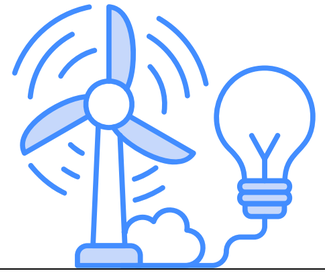
Modeling

- 모델 선택
- 하이퍼파라미터튜닝
- 다른 방식 시도

05

마무리

- 팀원들 소감
- 참고문헌



Title 01. Intro

1. 대회 주제

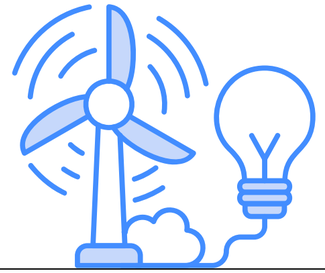
기상, 환경 데이터를 활용한 하루 전 풍력 발전량 예측

2. 제공 데이터

- 메타 데이터 : 풍력 발전소 및 터빈의 상세 정보 (발전용량, 위/경도, 터빈 모델 등)
- 분석용 데이터 : 풍력 터빈 SCADA 데이터 (터빈별 발전량, 풍속, 풍향, 출력제어량 등)
- Feature : 기상예보 LDAPS 데이터 (풍속, 기온, 기압 등)
- Target : 풍력 발전단지의 총발전량 (train_y) [kWh]

3. 규칙

- 2020 ~ 2022년 데이터를 활용하여 2023년 시간단위의 풍력 발전량 예측
- 분석용 데이터인 SCADA 데이터는 예측에 사용 X

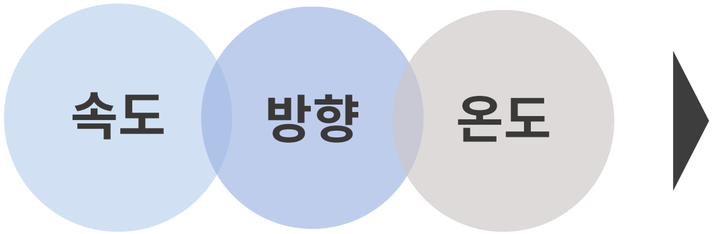


Title 02. EDA

풍력 발전량 형성

풍력 발전 예측에 필요한 데이터는 주로 **풍속**, **풍향**, **온도** 등과 같은 기상 요소에 기반

→ 풍력 터빈의 출력에 직접적인 영향을 미침



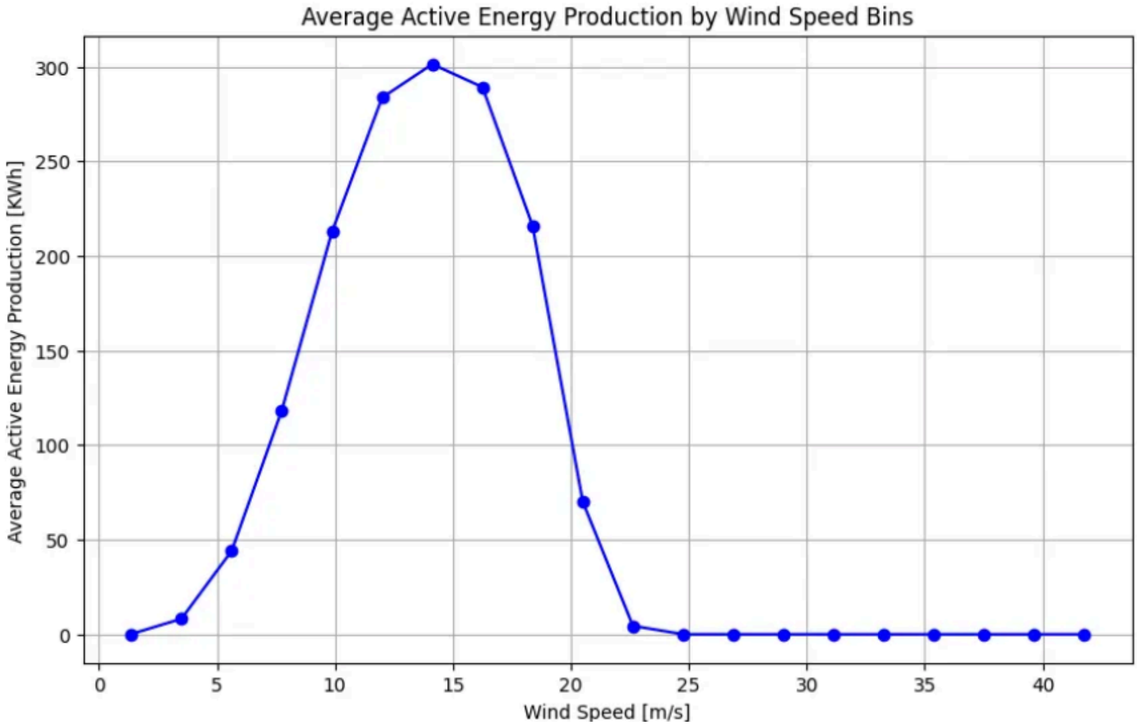
이 3가지에 집중하여 EDA를 진행함

제공 데이터 중 SCADA 데이터는 예측 feature로 사용 불가능

따라서 SCADA 데이터 **EDA를 통해 풍력 발전 에너지와의 연관성을 파악**하여
이를 기반으로 파생변수를 생성하고자 함.

속도 | 방향 | 온도

✓ 풍속이 크면 에너지 생산량이 많을 것이다.



SCADA 데이터의 풍속과 에너지 생산량

☑ 풍속이 약 15m/s 까지는 에너지 생산량이 **증가하다가 이후로 감소**하는 모습

→ 풍속이 높다고 무조건 에너지량이 많아지지 않음

☑ 풍속이 약 25m/s 부터는 더 이상 **에너지가 생산되지 않음**

→ 풍속이 25m/s 를 넘어가면 태풍으로 간주해 터빈 가동을 멈추기 때문에 에너지 발전량이 측정되지 않음

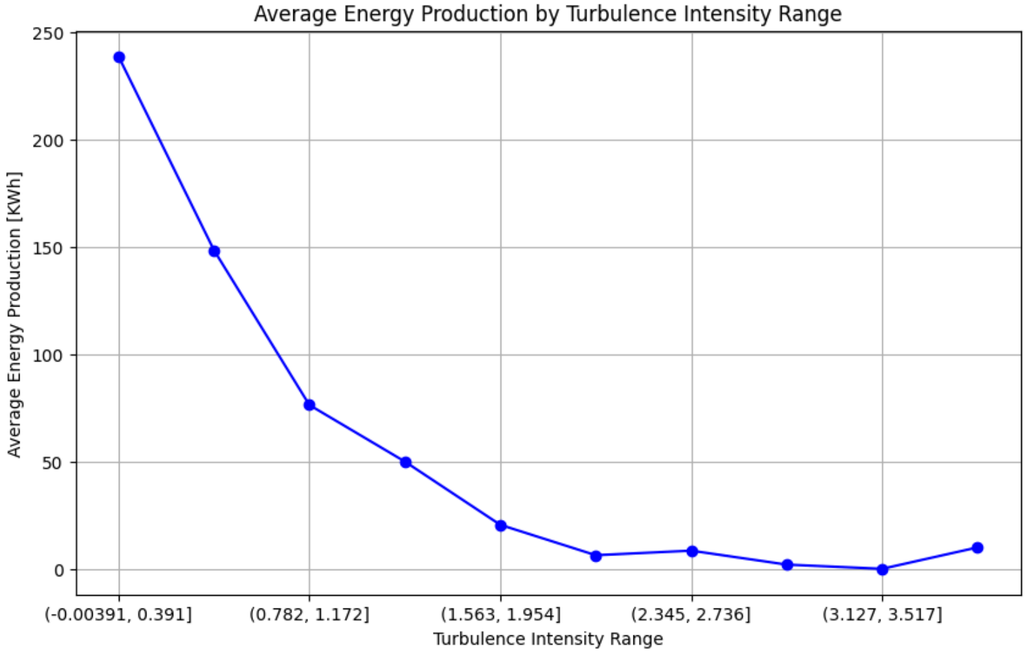
*EDA에서 그려진 SCADA 그래프는 경주풍력의 그래프로, 영광도 비슷한 양상을 가짐

→ wind 라이브러리 함수를 통해 **풍속 파생변수 생성**

+ 풍속 제곱 파생변수 생성

✓ 난류가 낮을수록 에너지 생산량이 높을 것이다.

난류 : 바람의 최대 속도와 최소 속도 차이를 바람의 평균 속도로 나눈 값



SCADA 데이터의 풍속을 이용한 난류와 에너지 생산량

☑ 낮은 난류 강도

→ 바람이 일정하고 안정적일 때, 터빈은 지속적으로 높은 효율의 에너지를 생산 가능

☑ 높은 난류 강도

→ 바람의 변동이 크면 터빈이 갑작스러운 바람 변화에 반응해야 해 터빈이 제대로 동작하지 않음

→ 최대 풍속, 최소 풍속 데이터를 이용해 **난류 파생변수 생성**

그 외 파생 변수 생성

- 폭풍의 속도

풍속에 따라 에너지 생성량이 달라지는 것처럼 폭풍의 속도도 영향을 끼칠 것이라 생각

LDAPS 데이터의 ‘storm_u_5m’와 ‘storm_v_5m’ 변수를 활용하여 ‘storm_speed’를 계산

```
LDAPS[‘storm_speed’]
= np.sqrt(LDAPS[‘storm_u_5m’]**2 + LDAPS[‘storm_v_5m’]**2)
```

- 바람의 운동 에너지

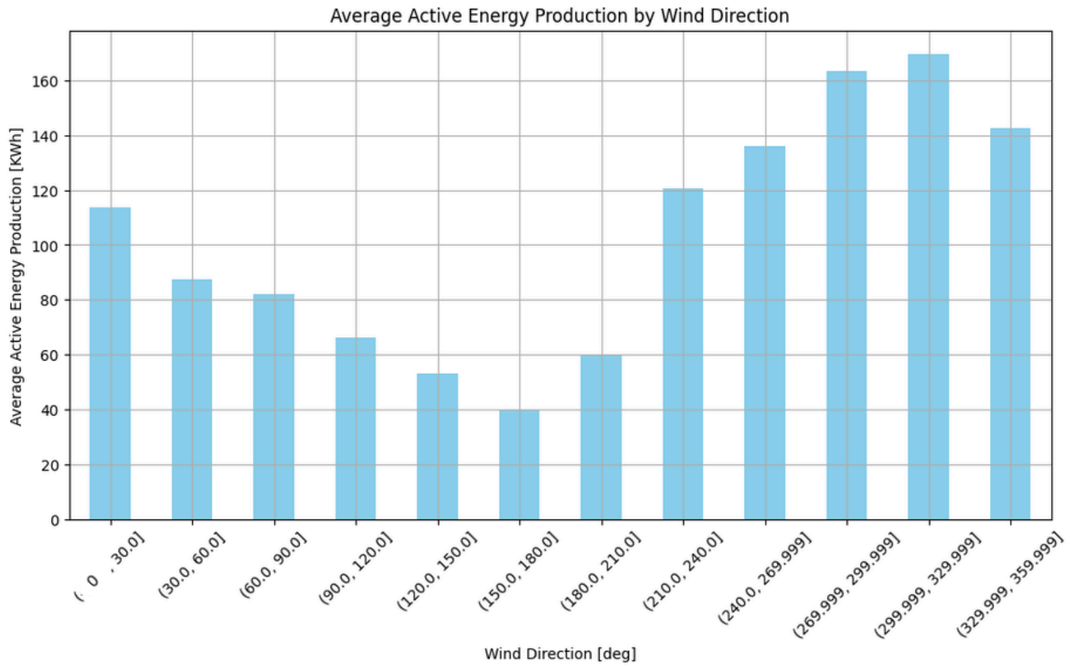
이론적으로 터빈이 얻을 수 있는 에너지

파생변수로 생성한 풍속과 밀도 변수를 활용하여 ‘wind_energy’를 계산

```
LDAPS[‘wind_energy’]
= 0.5 * (LDAPS[‘air_density’] * turbine_area * LDAPS[‘wind_speed’]**3)
```

속도 | 방향 | 온도

✓ 에너지가 많이 생성되는 특정 바람 방향이 있을 것이다.



SCADA 데이터의 풍향과 에너지 생산량

☑ 낮은 에너지 생성량

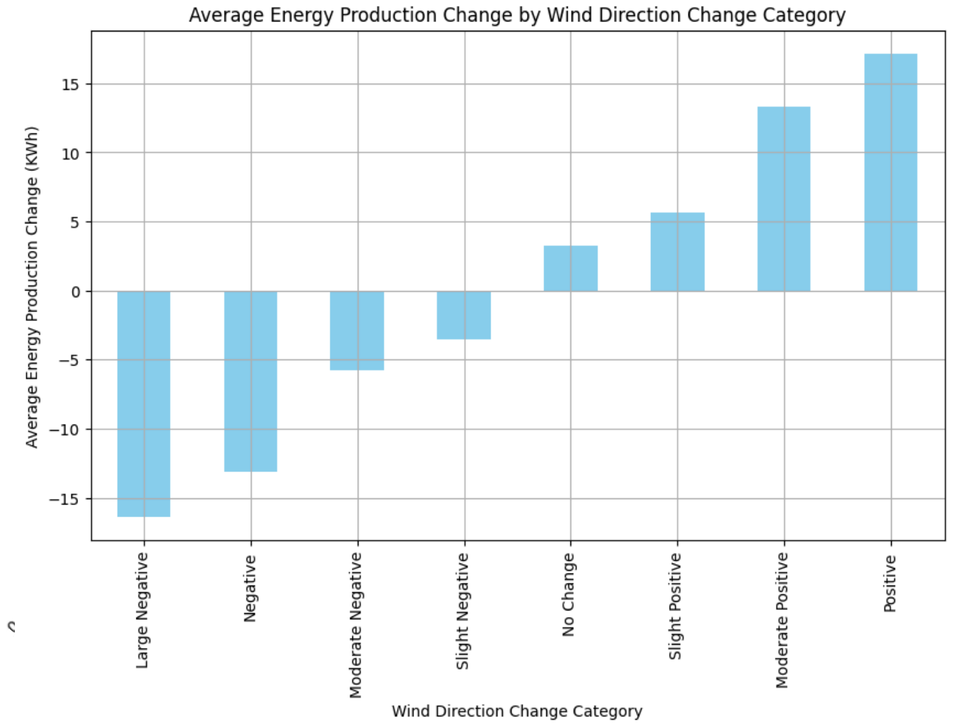
→ 120도~210도 사이에서 에너지 생산량이 50 이하로 떨어짐

☑ 높은 에너지 생성량

→ 240도 이상에서 상대적으로 높은 에너지 생산량
해당 방향으로 바람이 불 때 터빈의 효율이 좋아짐

→ wind 라이브러리 함수를 통해 **풍향 파생변수 생성**

✓ 풍향이 급격하게 변화하면 에너지량도 급격하게 변할 것이다.



SCADA 데이터의 풍향 변화량과 에너지 생산량

☑ 풍향이 반시계 방향으로 변화

→ 에너지 생산량이 떨어지는 경향

☑ 풍향이 거의 변화 X

→ 에너지 생산량의 변화가 거의 없거나 미미한 증가

☑ 풍향이 시계 방향으로 변화

→ 에너지 생산량이 증가하는 경향

→ 차분을 통해 **풍향 변화율 파생변수** 생성

그 외 파생 변수 생성

- 폭풍 방향

풍향에 따라 에너지 생성량이 달라지는 것처럼 폭풍의 방향도 영향을 끼칠 것이라 생각

LDAPS 데이터의 ‘storm_u_5m’와 ‘storm_v_5m’ 변수를 활용하여 ‘storm_direction’를 계산

```
LDAPS[‘storm_direction’]
= np.degrees(np.arctan2(LDAPS[‘storm_u_5m’], LDAPS[‘storm_v_5m’]))
```

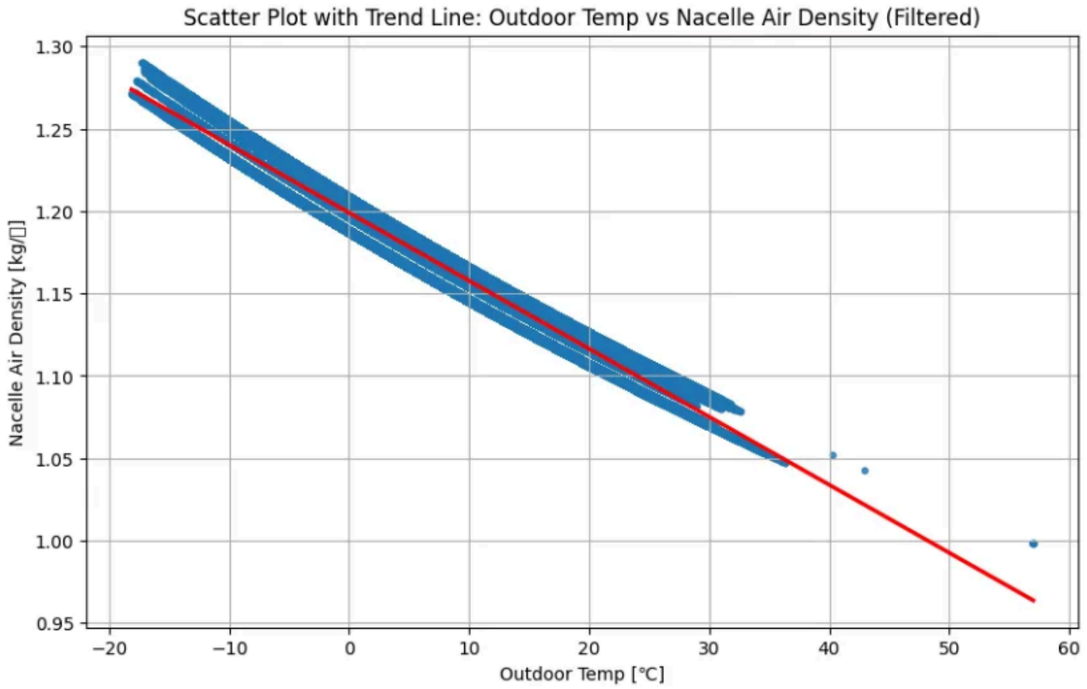
- 풍속 & 풍향 상호작용

풍속과 풍향은 풍력 발전량에 가장 큰 영향을 끼침

파생변수로 생성한 풍속과 풍향 변수를 활용하여 ‘wind_speed_direction_interaction’를 계산

```
LDAPS[‘wind_speed_direction_interaction’]
= LDAPS[‘wind_direction’] * LDAPS[‘wind_speed’]
```

✓ 온도가 증가하면 밀도가 감소할 것이다.

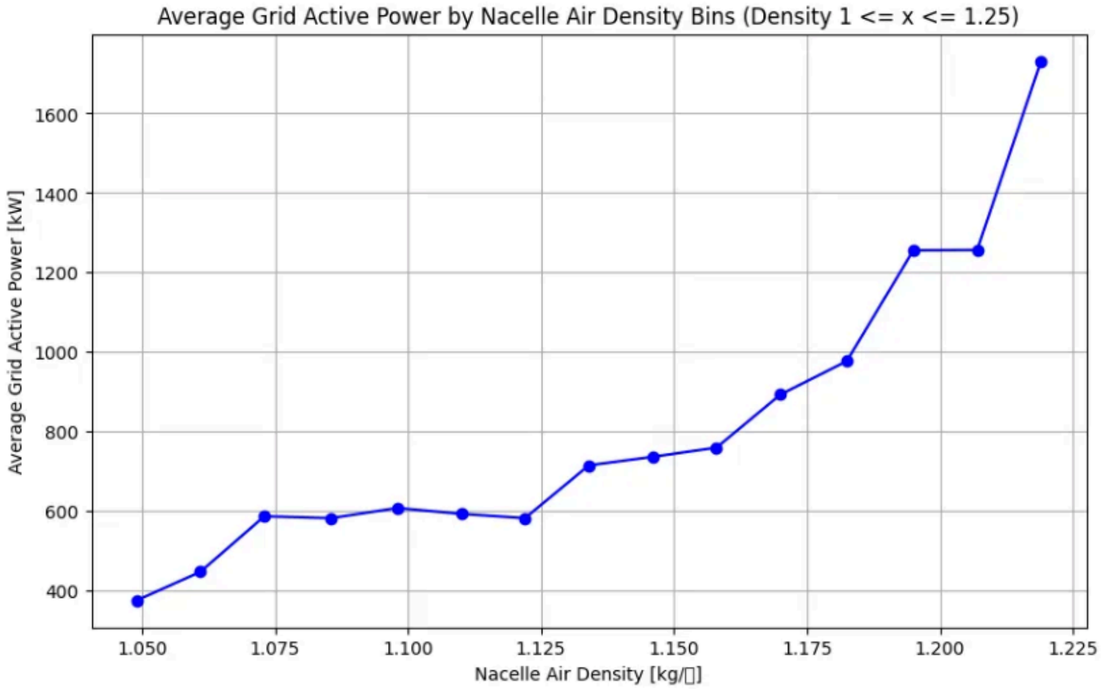


SCADA 데이터의 외부온도와 나셀 공기 밀도

☑ **반비례** 관계

→ 외부 온도가 증가하면 공기 밀도는 감소

✓ 밀도가 증가하면 에너지 생산량이 증가할 것이다.



SCADA 데이터의 나셀 공기 밀도와 에너지 생산량

☑ **밀도가 증가하면 에너지 생산량이 증가하는 경향**

→ 밀도를 생성해 풍력 발전량 공식에 활용

파생 변수 생성

- 절대 온도

LDAPS 데이터의 공기 온도 변수를 이용해 절대온도 계산

공기밀도 계산에 사용, 기온에 따라 공기밀도가 변동하기 때문에 절대 온도를 기반으로 한 밀도 계산은 발전량 예측의 정확도를 높일 것이라 생각

```
LDAPS['temp_air_K'] = LDAPS['temp_air'] + 273.15
```

- 공기 밀도

LDAPS 데이터의 압력 변수와 파생변수로 생성한 절대온도를 이용해 공기 밀도 계산

공기 밀도는 압력에 비례하고 온도에 반비례

앞서 소개한 풍력 발전량 공식에 활용

```
LDAPS['air_density'] = LDAPS['pressure'] * 100 / (R_air * LDAPS['temp_air_K'])
```

✓ 최종 데이터

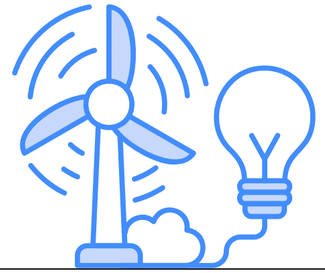
제공 변수 (LDAPS)

pressure	temp_air
relative_humid	elevation
specific_humid	land_cover
storm_v_5m	surf_rough
storm_u_5m	turbine_id
wind_u_10m	wind_v_10m
frictional_vmin_50m	frictional_vmax_50m

+

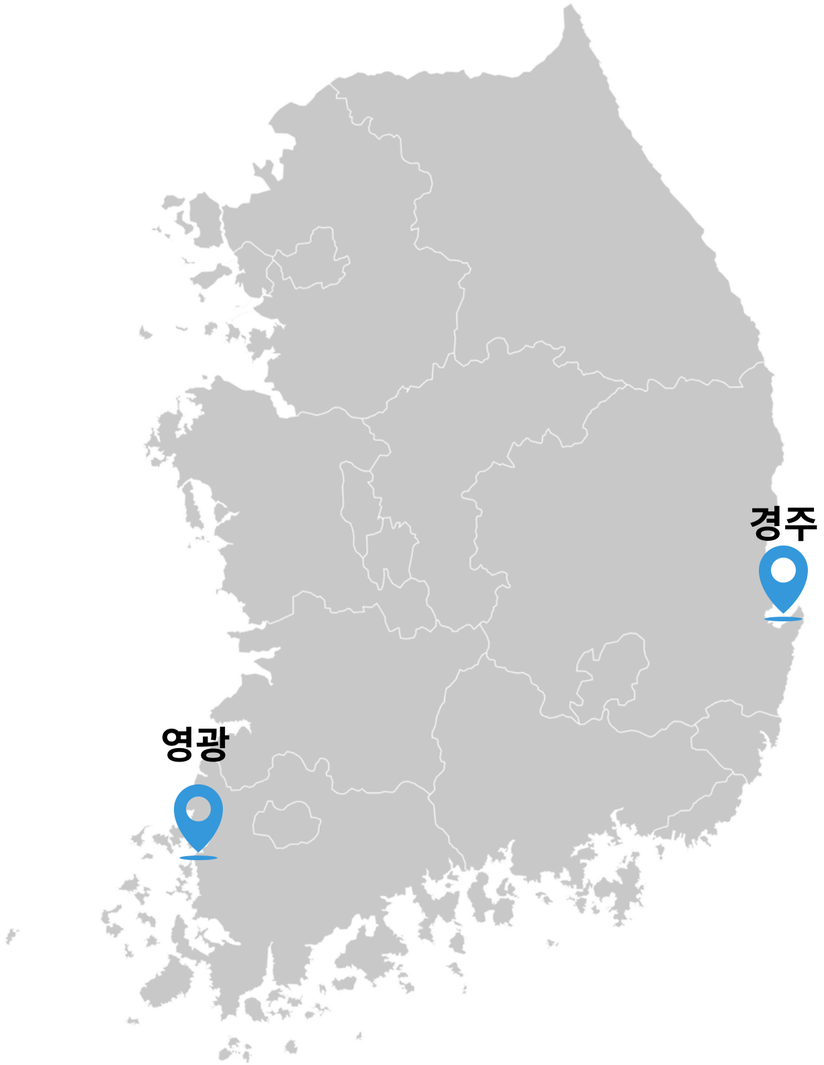
파생 변수

storm_direction	wind_speed_squared
wind_speed	air_density
storm_speed	temp_air_K
wind_energy	wind_direction_diff
wind_direction	turbulence_intensity
wind_speed_direction_interaction	



Title 03. Data 전처리

경주 풍력과 영광 풍력



경주, 영광의 지리적 차이

☑ 경주는 동해, 영광은 서해

- 동해는 깊은 수심과 급격한 해저 지형을 가지고 있어 바람이 더 강하고 일정하게 불 수 있는 경향. 특히, 동해안은 동쪽에서 불어오는 해풍과 겨울철 강한 북서풍의 영향을 더 많이 받음
- 서해는 상대적으로 완만한 해안선과 얕은 수심이 특징, 계절에 따라 몬순(북서풍, 남동풍)의 영향을 받으며, 바람의 세기와 방향이 주로 계절적 변화에 따라 달라짐

→ 풍속, 풍향, 기후 등 **지리적 차이로 인한 풍력 발전량이 다를 것**으로 판단.

경주 풍력과 영광 풍력

☑ 메타데이터에서 경주와 영광에 따른 차이 존재

발전단지 명	발전단지 용량	실제 용량[MW]	송전계약용량[MW]	송전전압 [kV]	대표 위도	대표 경도	터빈 개수
경주풍력2단지	20.7	20.7	20	22.9	35.73088	129.3673	9
영광풍력	79.6	79.6	79.6	22.9	35.25258	126.3423	35

☑ 경주와 영광의 풍력 발전량 차이

예측에 사용하는 LDAPS 데이터는 크게 경주풍력(총 9개의 turbine)과 영광풍력(총 35개의 turbine)으로 나뉨

planet_name	평균 energy_kwh
경주풍력	**6735.45**
영광풍력	**16234.78**

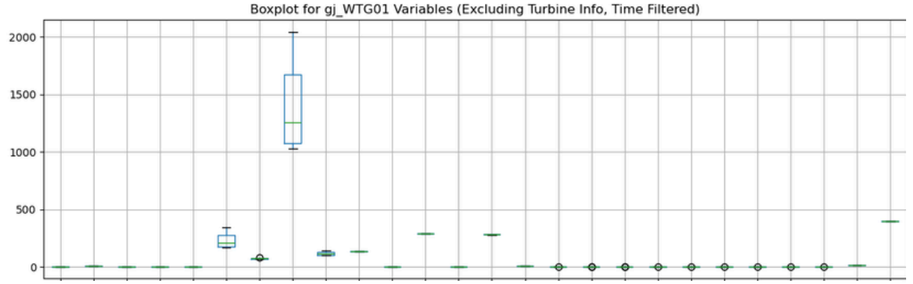
실제로 train_y에서 발전단지 별 에너지발전량 평균을 보면,
영광풍력이 경주풍력의 3배 가까이 되는 것을 확인

→ 이러한 근거로, **경주와 영광을 따로 예측**하기로 결정!

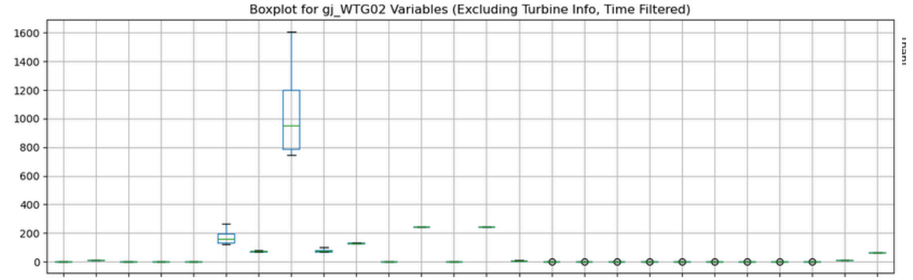
Data 전처리

데이터 병합

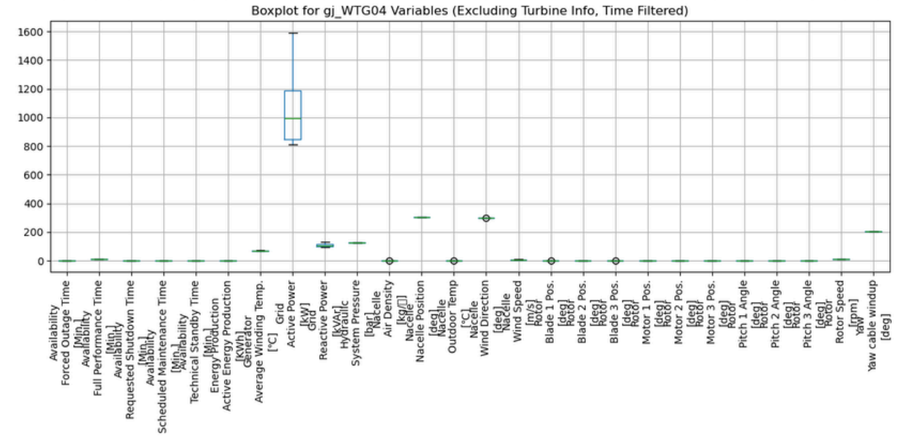
터빈 1



터빈 2



터빈 4



LDAPS 데이터와 train_y 데이터를 날짜 기준으로 병합 필요

But, LDAPS 데이터는 같은 날짜에 여러 터빈 ID 존재

같은 시간대에 여러 터빈들 중에서 가장 대표할 수 있는 값은 무엇일까?

☑ 터빈 별 같은 시간 동안의 변수 값 비교

→ Box plot을 보면 모두 **비슷한 분포**를 가짐

터빈 별 특징보다 **환경적 요인**이
풍력 발전에 끼치는 영향이 더 크다고 판단

※ 이 그래프는 경주에 대한 그래프이며,
경주, 영광 모든 터빈에 대하여 여러 Box plot을 그려보았으나 같은 결론을 내림

데이터 병합

같은 시간에 여러 개 있는 터빈 ID별
데이터를 하나로 통합

EX)

날짜	터빈 ID	풍속	온도
2시	1	15	20
2시	2	16	19.5
2시	3	14	21.5
2시	4	15	22

↓ max값

날짜	풍속	온도
2시	16	22

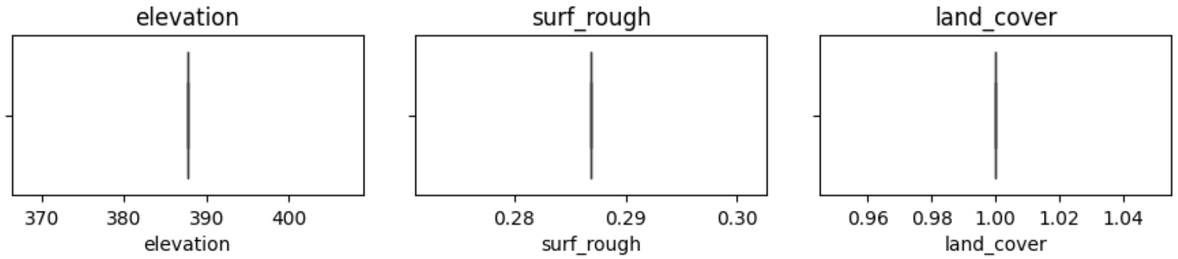
같은 시간대에 여러 터빈들 중에서 가장 대표할 수 있는 값은 무엇일까?

최종적으로 **같은 시간대**에서 **대표할 수 있는 값**을 찾기 위해
각 시간대 별 LDAPS의 변수를 **'max' 값으로 병합**

(평균값도 고려해보았으나 max 값으로 병합했을 때 더 좋은 성능이 나타남.
이는 에너지 생산 예측에 피크값이 더 중요한 영향을 끼치기 때문으로 보임.)

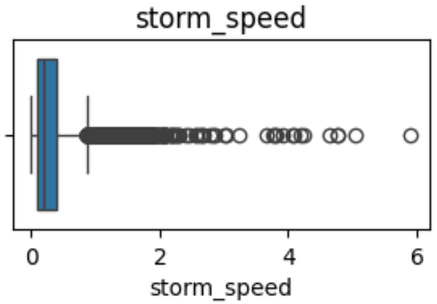
변수 drop

✓ ‘elevation’, ‘surf_rough’, ‘land_cover’, ‘turbine_id’



max 값으로 합병되면서
3개의 변수 **모두가 같은 값**
=> **drop**

✓ ‘storm_speed’



오른쪽으로 매우 많은 **이상치**
=> **drop** 후 성능 개선

✓ ‘temp_air_K’

temp_air_K = temp_air + 273.15로 이는 **대칭 이동**에 해당.

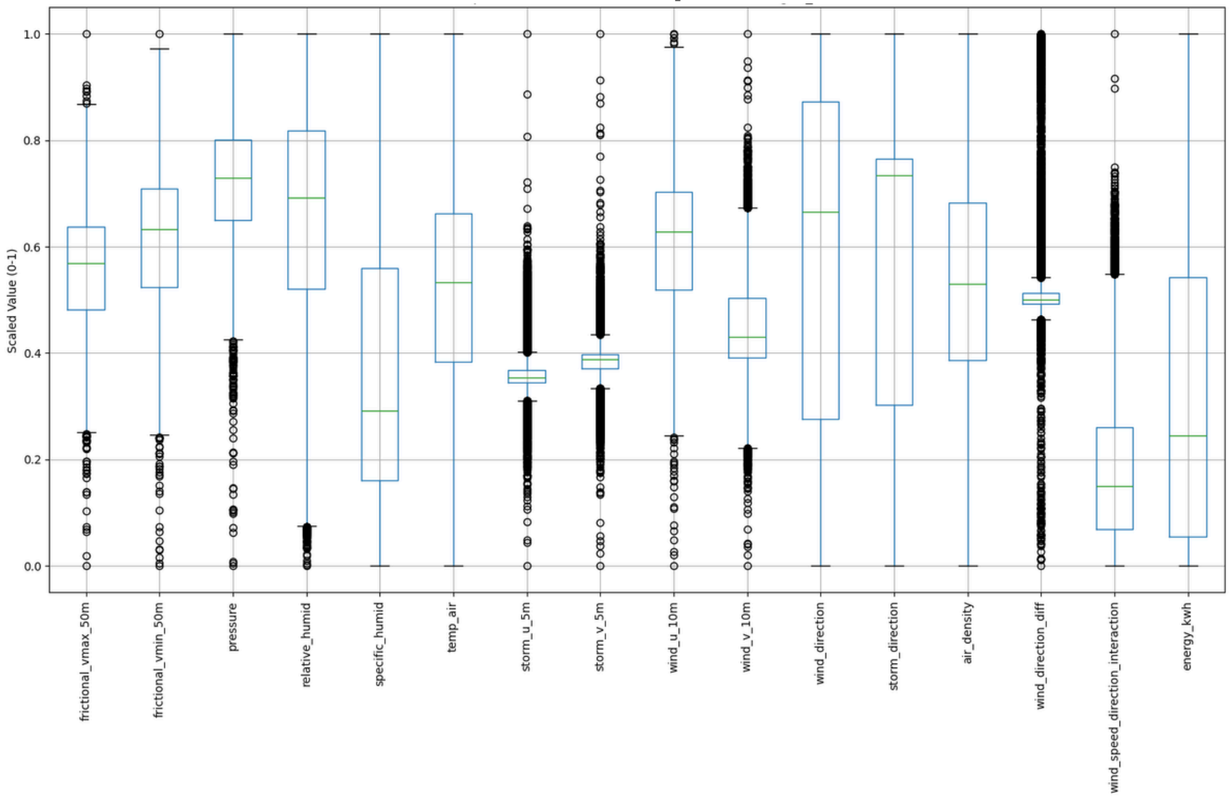
즉, 두 변수의 상대적인 차이나 순서는 변하지 않음.

두 변수는 **실질적으로 동일한 역할**을 하기 때문에

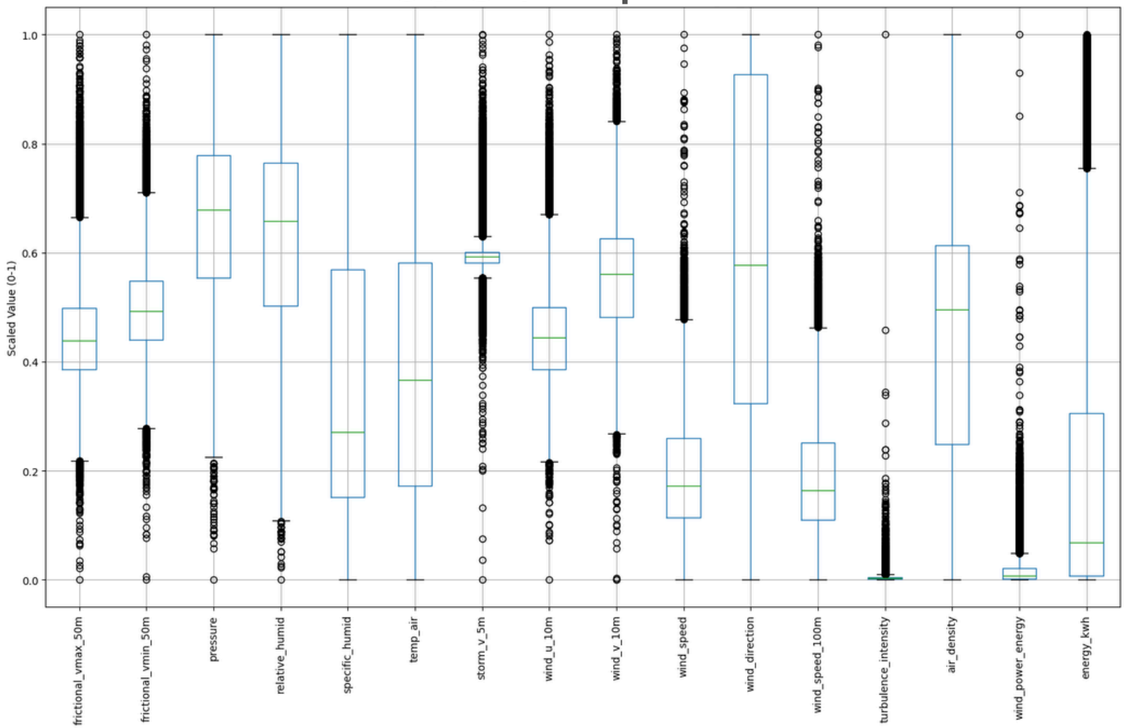
‘temp_air_K’ drop

이상치 확인

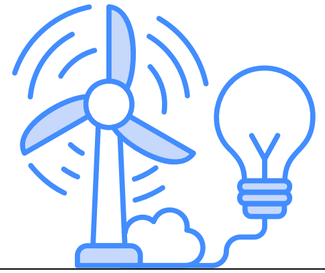
경주 풍력 box-plot



영광 풍력 box-plot



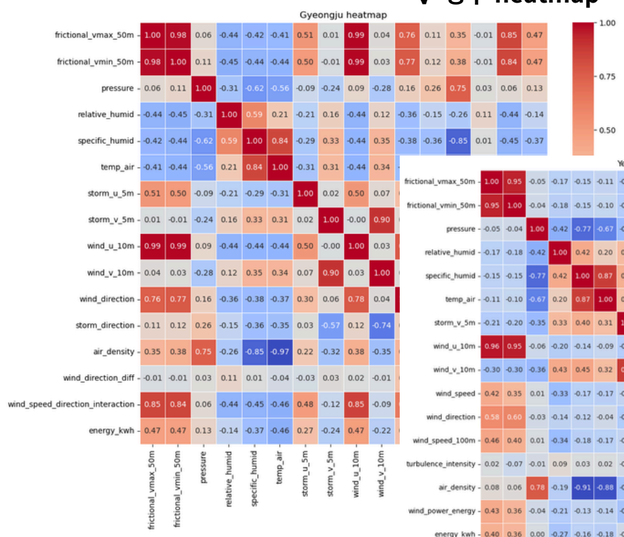
기상 데이터는 자연적으로 **변동성**이 크다고 판단 → 이상치 따로 처리 X



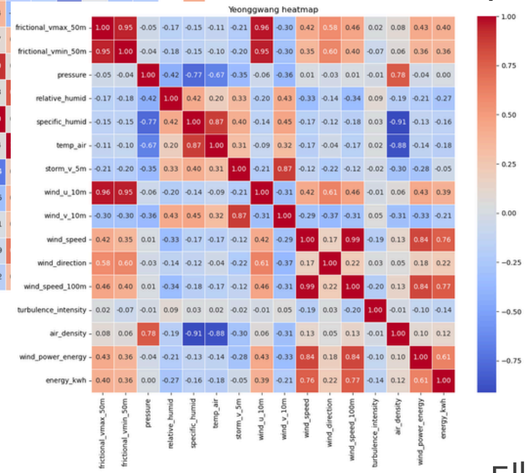
Title 04. Modeling

Modeling

✓ 경주 heatmap



✓ 영광 heatmap



데이터 특징

- ☑ 시계열 데이터
- ☑ 변수마다 다른 스케일
- ☑ 다수의 이상치
- ☑ 데이터 간 다중 공선성 문제
- ☑ 대용량 데이터

데이터의 특징을 고려한 모델 선택

XGBoost	LightGBM	LSTM	AutoGluon
트리 기반 모델로, 데이터 스케일 및 다중공선성에 민감 X	트리 기반 모델로, 데이터 스케일 및 다중공선성에 민감 X, 대용량 데이터에 적합	시계열 데이터로, 장기 의존성 및 비선형 패턴을 고려	WeightedEnsemble_L2 모델 적용

모델 간 NMAE값 비교

	XGBoost	LGBM	AutoGluon	LSTM
경주	**12.50**	12.91	13.38	13.49
영광	9.85	**9.70**	10.60	9.97

☑ **경주는 XGBoost,**
영광은 LGBM이
가장 좋은 성능을 보임
=> 최종 모델로 선택

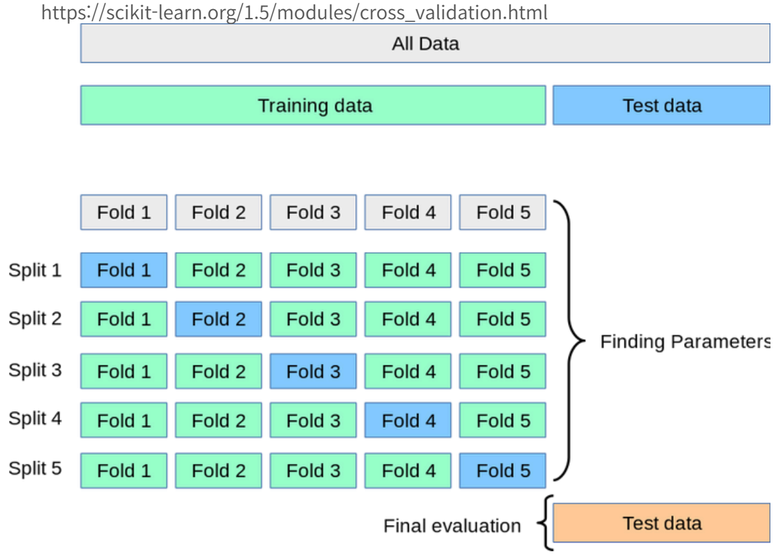
성능 개선을 위한 변수 선택

경주	영광
'wind_speed', 'wind_speed_squared', **'wind_energy'** 변수를 제거했을 때 NMAE 값이 12.50 에서 **12.46** 로 성능이 좋아짐 => 제거하기로 결정	**'wind_direction_diff' , 'storm_u_5m',** **'wind_speed_direction_interaction',** **'storm_direction', 'wind_speed_squared'** 변수를 제거했을 때 NMAE 값이 9.70 에서 **9.48** 로 감소 => 제거하기로 결정

하이퍼파라미터 튜닝

	튜닝 전	튜닝 후
경주	12.46 →	**12.37**
영광	9.48 →	**9.35**

https://scikit-learn.org/1.5/modules/cross_validation.html



Grid search로 하이퍼파라미터 튜닝 진행

결과를 보면,
경주의 경우 NMAE 값이 약 0.09,
영광의 경우 NMAE 값이 약 0.13 감소

이때, **K - Fold Cross Validation** 사용

K - Fold Cross Validation *k = 5 사용

k개의 fold를 만들어서 교차 검증하는 방법론

- 한 번의 split로 평가하는 것보다 **일관된 성능 평가** 가능
- grid search 과정에서 사용하여 **특정 분할에만 성능이 최적화될 위험 방지**

다른 방식 시도

모델	XGB	LGBM	XGB + LGBM
경주	13.36	13.68	**12.87**
영광	9.54	9.55	**9.48**

앞서 설명한 파생변수 및 전처리 방식과 다른 방법으로 데이터 처리 시도
=> 총 3가지 결과물 제출

- ☑ 경주와 영광에 대해 **XGB 모델**로 예측한 결과
- ☑ 경주와 영광에 대해 **LGBM 모델**로 예측한 결과
- ☑ XGB와 LGBM 결과값을 평균 내어 만든 **앙상블 모델**로 예측한 결과

→ 단일 모델로 예측하는 것보다, 앙상블 모델의 성능이 더 우수

제출 기회 횟수 제한으로, 앞서 설명한 전처리 방식으로 앙상블 모델을 만들어 제출해보지는 못함
=> 시도해본다면 **성능 개선의 가능성** 존재

도식화

	발전단지	발전기 번호	제작사	turbine ID	터빈 모델	정격 출력 [kW]
0	경주풍력	WTG01	UNISON	U113	2300
1	경주풍력	WTG02	UNISON	U113	2300
2	경주풍력	WTG03	UNISON	U113	2300
3	경주풍력	WTG04	UNISON	U113	2300
4	경주풍력	WTG05	UNISON	U113	2300
5	경주풍력	WTG06	UNISON	U113	2300
6	경주풍력	WTG07	UNISON	U113	2300
7	경주풍력	WTG08	UNISON	U113	2300
8	경주풍력	WTG09	UNISON	U113	2300

< meta 데이터 >

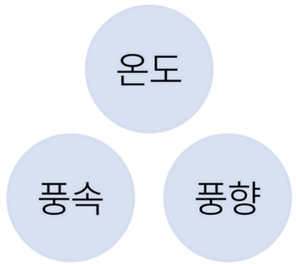
dt	elevation	land_cover	surf_rough	frictional_vmax_50m
2023-01-01 00:00:00+09:00	387.640625	1.0	0.286911	11.116076
2023-01-01 01:00:00+09:00	387.640625	1.0	0.286911	10.945148
2023-01-01 02:00:00+09:00	387.640625	1.0	0.286911	11.457981
2023-01-01 03:00:00+09:00	387.640625	1.0	0.286911	11.583426
2023-01-01 04:00:00+09:00	387.640625	1.0	0.286911	11.028671
...	...	...	...	...
2023-12-31 19:00:00+09:00	291.109375	1.0	0.297867	4.330050
2023-12-31 20:00:00+09:00	291.109375	1.0	0.297867	3.073275
2023-12-31 21:00:00+09:00	291.109375	1.0	0.297867	2.087387
2023-12-31 22:00:00+09:00	291.109375	1.0	0.297867	1.802292

< LDAPS 데이터 >

Pre processing

- 경주 & 영광 지리적 차이 고려
- 파생변수 생성
- 변수 drop

EDA



python

Train model

- 경주 XGB
- 영광 LGBM
- 하이퍼파라미터튜닝

Data verification

Prediction

Download predicted data in CSV format

CSV

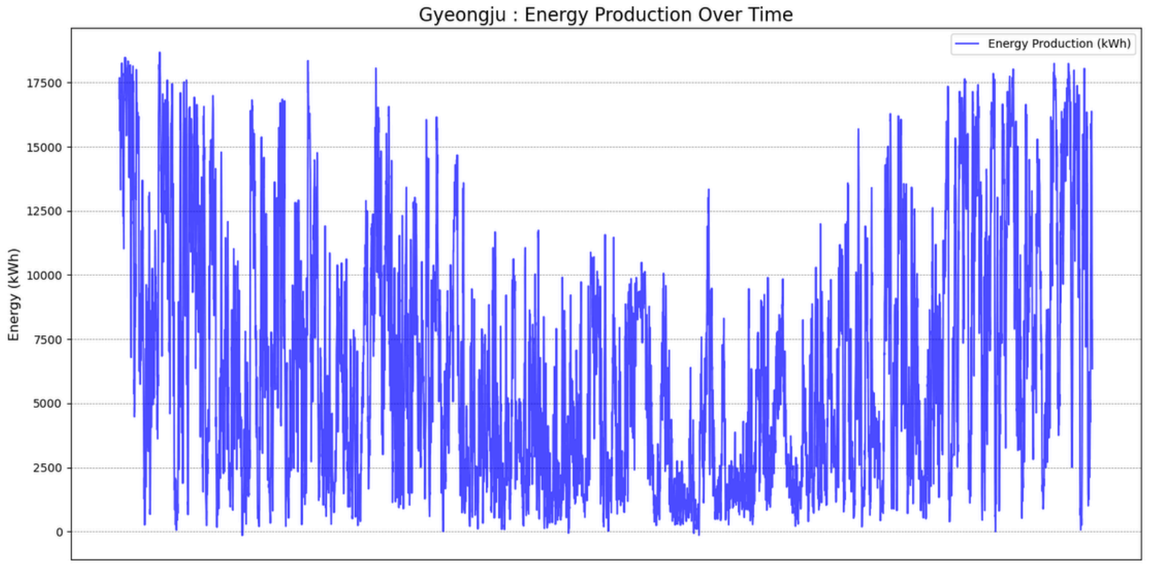
Visualize predicted data graphs

최종 제출 결과 & 시각화

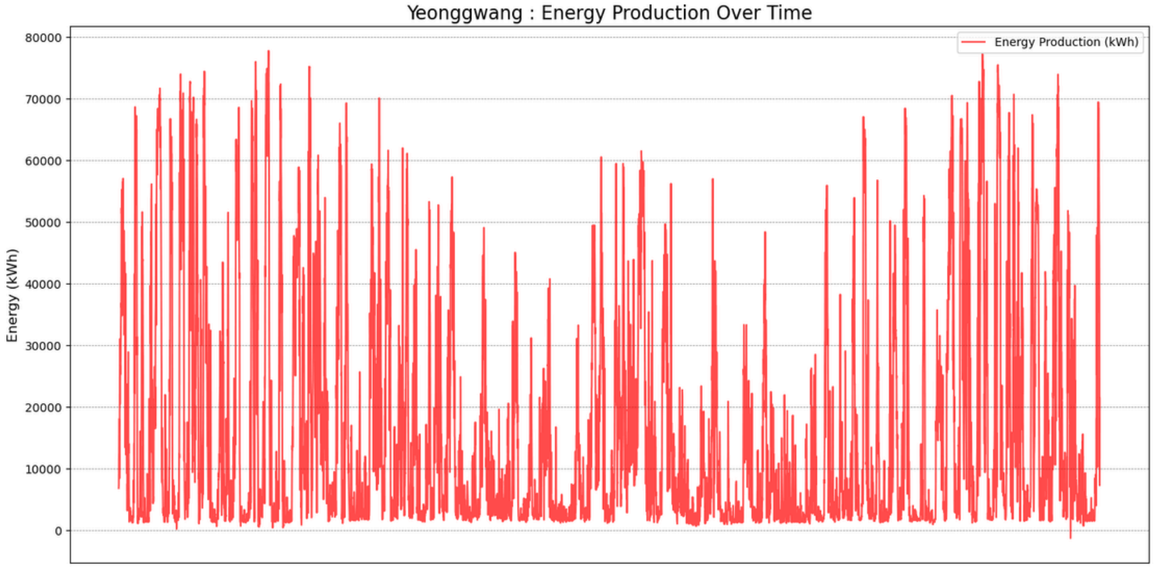
< 최종 제출 결과 >

	NMAE
경주	12.37
영광	9.35

✓ 경주 시각화



✓ 영광 시각화



팀원들 소감

"공모전을 통해 원본 데이터를 활용한 모델링 과정을 체험하며 데이터 분석의 실질적인 중요성을 깨달았습니다. 특히 기상 요소가 풍력 발전에 미치는 영향을 분석하며 변수 생성과 전처리의 중요성을 느꼈고, 이를 통해 현장에 적용 가능한 실무적 역량을 키울 수 있었습니다."

"데이터 전처리와 변수 생성이 예측 모델 성능에 결정적인 영향을 미친다는 점을 배울 수 있었습니다. 기상 데이터를 활용한 풍력 발전 분석 과정에서 데이터의 특성을 파악하고 이를 모델링에 반영하며 데이터 분석의 실질적인 깊이를 체험할 수 있어 매우 뜻깊은 시간이었습니다."

"기상 데이터를 분석하며 풍력 발전 예측 모델을 구축하는 과정이 매우 흥미로웠습니다. 실제 데이터를 기반으로 문제를 해결하는 방법을 배우며 실무적인 감각과 데이터를 다루는 기술적 역량을 크게 향상시킬 수 있었습니다. 협업 과정에서 팀원들과 유익한 경험도 공유했습니다."

"기상 데이터와 풍력 발전의 연관성을 분석하며 데이터를 이해하고 적절히 전처리하는 과정의 중요성을 느꼈습니다. 원본 데이터를 다뤄볼 기회가 적었던 만큼 이번 공모전이 실무적 경험을 쌓는 데 매우 뜻깊었고, 협업을 통해 데이터 분석의 다양한 접근법도 배울 수 있었습니다."

참고 문헌

☑ 논문

변동민, “영광풍력발전량 예측오차 감소 알고리즘 연구”, (석사학위 한양대학교 일반대학원, 2022)

☑ 대회 수상작

데이콘 풍력 발전량 예측 AI 헤커톤
https://dacon.io/competitions/official/236066/codeshare/7720?page=1&dtype=recent

데이콘 전기 발전량 예측 헤커톤
https://dacon.io/competitions/official/235736/overview/description

☑ 사이트

windlibpackage 모듈 설명 사이트
https://windpowerlib.readthedocs.io/en/stable/

BDA & 동서발전 기상데이터 공모전

감사합니다

TEAM 선데이

60Hertz wai B.D.A BIG DATA ANALYSIS EWP 국립기상과학원 KIER 한국에너지기술연구원 KOREA INSTITUTE OF ENERGY RESEARCH