2024 날씨 빅데이터 콘테스트 과제 4

시공간 복합데이터를 활용한

전력 수요 예측 개선

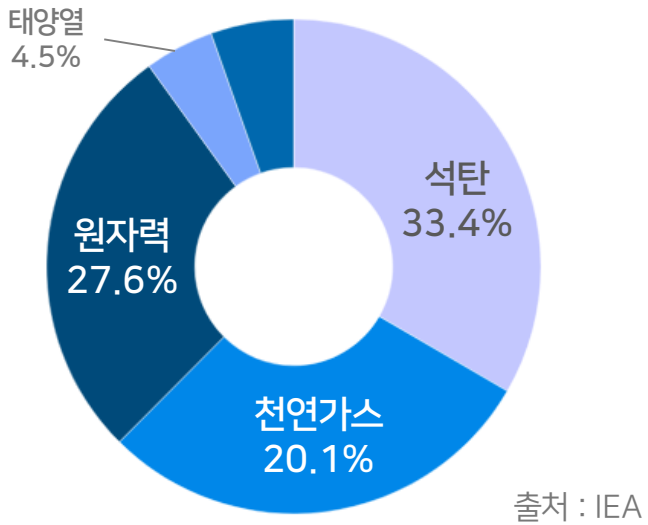
기상청
Korea Meteorological Administration

참가번호 240555
팀명 카이제곱분포

전기는 국산이지만 원료는 수입입니다

인당 전기 소비량과 전기 원료 수입 의존도가 높은 대한민국

대한민국 전기 생산 원료 비율 (2022)

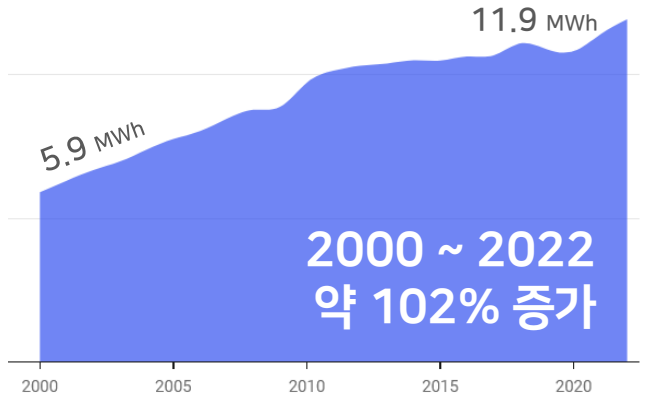


출처 : IEA

대한민국 인당 전력 소비량 통계

순위	국가명	인당 전력 소비량
1	대만	11.826 MWh
2	브루나이	11.656 MWh
3	대한민국	11.402 MWh
4	싱가포르	10.22 MWh

2021년 아시아/호주 인당 전력 소비량 통계 (IEA)



2022 ~ 2022 대한민국 인당 전력 소비량 추이(IEA)

인당 전력 소비량이 높으면서 전기 생산 원료가 대부분 수입에 의존한다면, **선제적으로 전력 수요량을 예측할 필요**가 있음

기본 위치 변수

기본적으로 제공된 기온, 상대습도와 같은 **시계열 기상 관측 데이터**에 관측지점의 **공간적인 정보를 포함한 위치 변수**를 도입

지역별 연간 가정 전기 평균 사용량 (2015, 미국)

출처 : EIA

지역별로 전기 소비량이 다를 수 있다는 통계를 확인,

공간적인 특징을 반영할 수 있는 변수 도입 하기로!

AWS

기본 제공된 기상 전력 데이터의

STN에 대응하는 격자 번호 변수

출처 : 기상청 날씨누리
기상청 자료개방포털

위도 **경도** **고도**

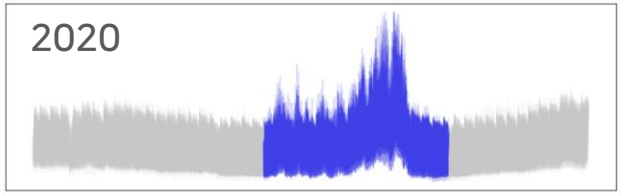
그리고, 위 세 변수를 활용한 파생변수 도입

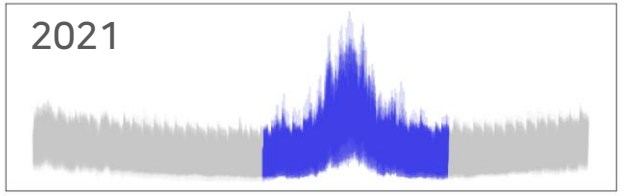
구성된 기본적인 데이터셋을 바탕으로 파생변수를 생성해봅시다!

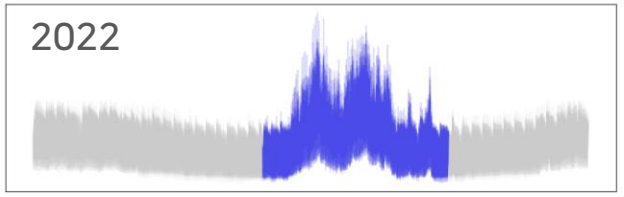
파생 변수 | 삼각함수 변환

전력량을 나타내는 변수 elec의 변동주기를 시각화한 결과 **일별, 연별 주기성**이 나타나고 있음을 확인할 수 있음

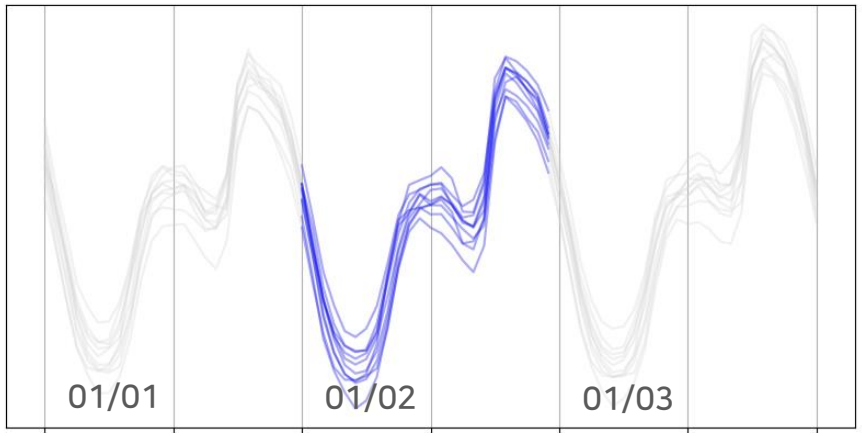
전력량 변수의 연도별 추이







전력량 변수의 일별 추이



특히, 6~9월 **냉방이 활발한 여름**을 중심으로 전력량 변수가 동일한 개형으로 크게 변동하는 모습이 관찰됨

파생 변수 | 삼각함수 변환

주기성을 가지는 변수들(일, 월)이 **주기적 특징을 강화**할 수 있도록 **월 변수** (년 주기성) 와 **시간 변수** (일 주기성) 을 **삼각함수로 변환함**

일별 주기 변환 **전** (시간 단위)					
22시	23시	0시	1시	…	21시

머신러닝 모델은 23시와 0시가 1시간 간격의 **연속된 값이라고 인식할 수 없음**

sin, cos 함수를 통한 주기성 변수 변환

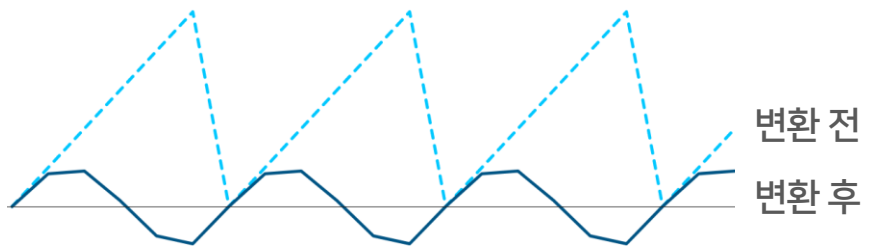
일별 주기 변환 **후** (시간 단위)					
0	-0.84	0	0.84	…	-0.90

불연속적인 주기를 연속적으로 변환해 모델이 이를 주기로 해석할 수 있도록 처리

Elec의 주기적 변동에 영향을 미칠 것으로 추측되지만,

변수 값 자체가 연속적으로 변동하지 않는 시계열이 변환 대상

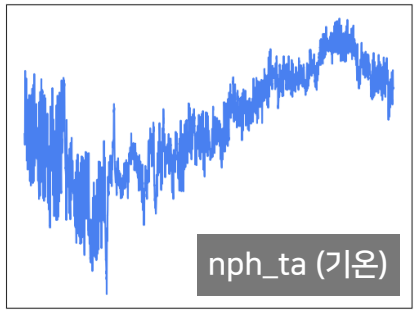
- 시간 (0~23)
- 월 (1~12)
- 여름 (6~8월)

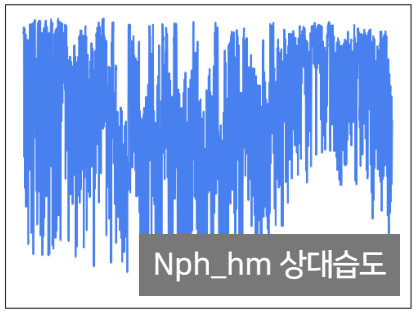


파생변수 | 푸리에 변환을 통한 노이즈 제거

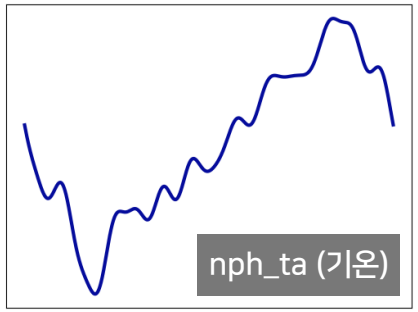
푸리에 변환은 복잡한 시계열 데이터를 **여러 개의 주기함수로 분해**하는 기법으로,
푸리에 변환을 응용하여 시계열 데이터의 **노이즈를 제거**할 수 있음

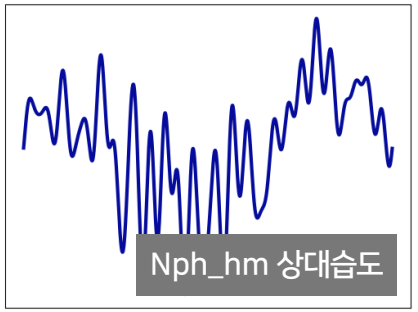
푸리에 변환 **전** > 푸리에 변환 **후**





복잡하고 노이즈가 많이 낀 형태. 전처리 없이 모델에 학습시킨다면 정확도가 떨어질 위험이 있음





노이즈가 크게 제거되어 원 시계열 데이터의 파동이 강조됨.

노이즈가 큰 시계열 데이터인 **기온, 상대습도, 10분 평균 풍속, 체감온도 변수**를 푸리에 변환하여 파생변수 생성

파생변수 | 기타 파생변수

01

오후 (12 ~ 15시) 를 나타내는 이진 파생변수

낮 시간대 (12 ~ 15시) 격자별 전력량 변수의 변동이 크게 두드러지는 것을 발견. 이를 반영할 수 있도록 데이터가 해당 시간대에 속하는지의 여부를 나타내는 파생변수 생성

02

불쾌지수 및 불쾌 여부 파생변수

여름철 급격하게 변동하는 전력 사용량의 원인을 냉방과 관련한 요인으로 추측. 이를 반영할 불쾌지수 관련 파생변수 생성

03

행정구역 파생변수

고도, 위도, 경도 외에도 AWS 관측지점의 지리적인 영향력을 반영할 수 있도록 상위 계층의 행정구역(15개) 파생변수로 생성

04

산지 여부, 수도권 여부 파생변수

AWS 관측지점의 지형적인 특징(산지 여부)을 반영하기 위해 산지 여부를, 수도권과 비수도권의 사회문화적인 이질성을 반영하기 위해 수도권 여부 파생변수 생성

모델링 설계 | 복합 모델링

Train Set : 2020 ~ 2021년의 데이터 (70%) / Validation Set : 2022년의 데이터 (30%)

목적변수의 시계열성을 고려해 random split이 아닌 시간 순서에 따라 train / valid set 구성

1차 모델링	>	2차 모델링
LGBM 모델 : 시계열성 적합		CatBoost 모델 : 시계열성이 제거된 잔차에 대한 적합
시계열 설명변수를 사용해 비선형 추세 & 계절성 적합		일부 시계열 변수와 비시계열 변수를 통해 Residual Fitting
LGBM 모델은 Leaf-wise 트리 분기를 통한 높은 예측 정확도가 특징. 그러나 고차원에서 데이터 과적합의 우려가 있음		CatBoost 모델은 범주형 변수에 대해 강력한 성능을 가짐. Level-wise 트리 분기를 통해 과적합에 상대적으로 강건함

모델링 설계 | 1차 모델

공간적 특성과 무관한 비선형 시계열 추세 & 계절성 적합 및 추정치 잔차 시계열성 최소화 모델

본 데이터는 시간적 특성과 공간적 차이가 혼재된 경시적 자료의 성격을 보여, 일반적인 시계열 모델의 적용에 무리가 있음을 판단.
Leaf-wise 분기를 통해 변수의 비선형적 특성의 파악에 강력한 성능을 보이며
각 변수의 중요도 파악 및 병렬 계산을 통한 대용량 데이터셋의 빠른 처리가 가능한 LGBM 사용

01. 변수 선택 기준 선정

- 기상 변수 및 시간 변수 포함
- 시계열 기상 변수를 통한 시간적 특성 반영
- 삼각함수 기반의 파생변수를 통한 주기성 반영

02. 데이터 학습 및 검증

LGBM
모델링

03. 모델 성능 평가

1차 모델링 성능			
R^2	MSE	MAE	ELEC_VAR
0.906	62.43	5.80	-

모델링 설계 | 2차 모델

1차 모델의 잔차를 지역 특성 변수와 일부 시계열 변수를 통해 Residual Fitting

시간적 추세에 더해 지역적 특성 및 일부 시점의 영향력을 추가 반영

1차 모델의 지역 무관한 시간적 추세의 과대·과소 추정 여부에 대한 해석을 제공

Level wise한 트리 분기를 통해 과적합에 강건하고 다양한 범주형 변수의 처리에 뛰어난 CatBoost 사용

01. 변수 선택 기준 선정

- 관측 지점 별 지역적 특성 및 기상 차이 반영
- 전력 사용량의 추세가 상이한 일부 시점의 특성 반영
- 오차에서 나타나는 주기성 제거

02. 데이터 학습 및 검증

CatBoost 모델링

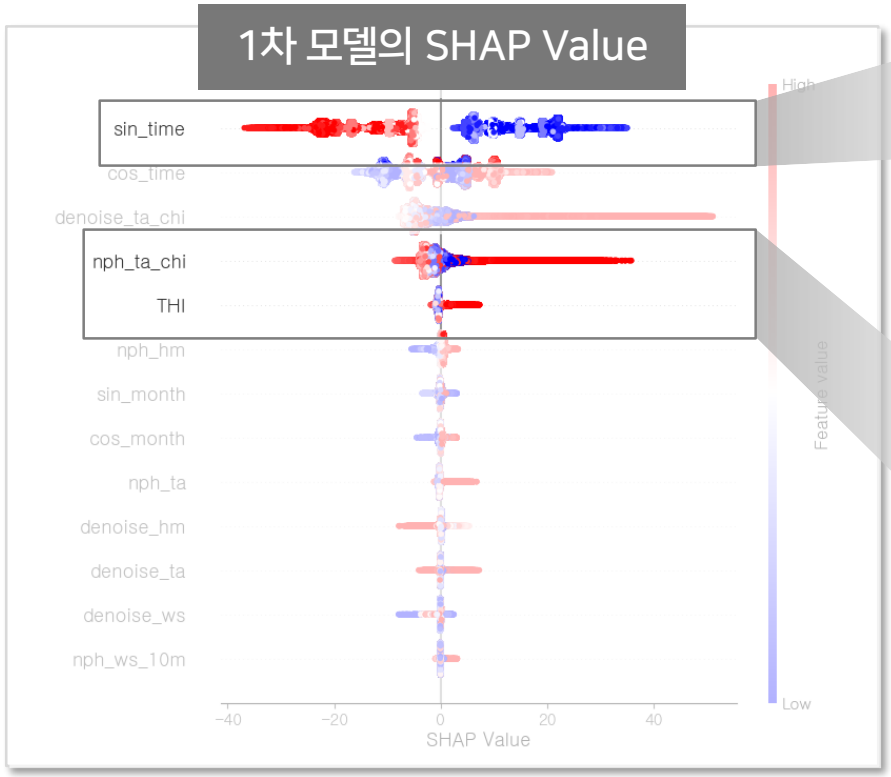
03. 모델 성능 평가

2차 모델링 성능			
R^2	MSE	MAE	ELEC_VAR
0.940	39.33	4.37	0.978

SHAP을 이용한 변수의 결과 해석

SHAP은 개별 예측값에 대한 각 변수의 영향력을 누적 배분하여

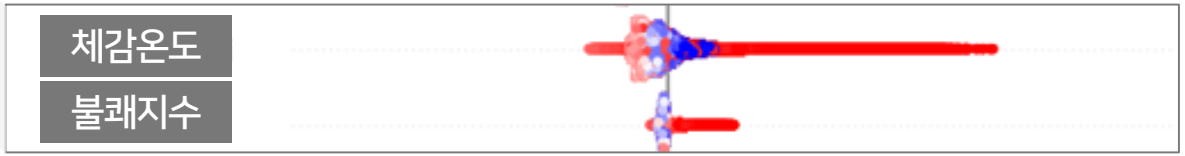
변수의 중요도와 방향성 파악 가능한 모델로, 인간의 직관과 유사한 해석을 제공



시간을 나타내는 변수인 sin_time에서 전력량과 **강한 음의 관계**가 관찰됨

전력량의 변동이 시간의 흐름에 크게 의존하고 있으며,

오전에 비해 오후에 전력 소비량이 더 많다는 뜻으로 해석가능함



체감온도, 불쾌지수와 같이 전력 소비의 주체가 체감하는 주관적 느낌과 관련된 변수에서 전력량과 **강한 양의 관계**가 관찰됨

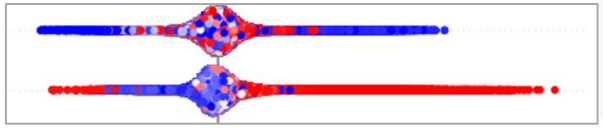
여름철 냉방으로 인한 전력 소비가 전체 전력 소비의 큰 비중을 차지하고 있음을 추측 가능

SHAP을 이용한 변수의 결과 해석

요일 (0~6)

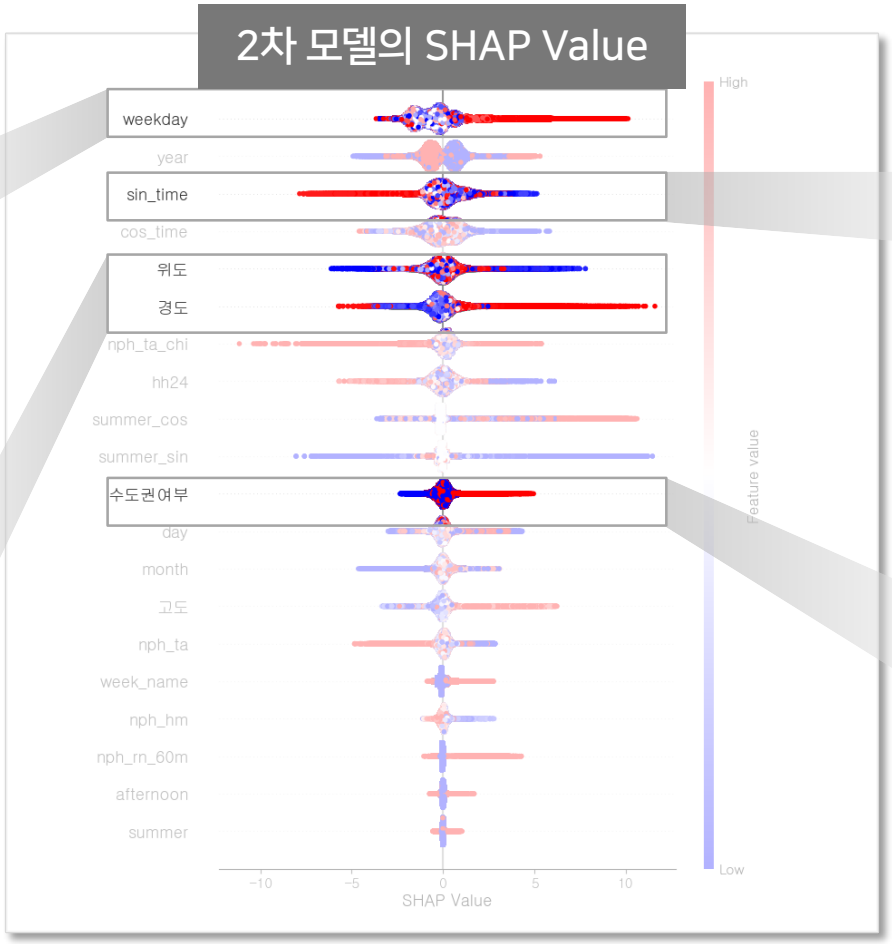
일반적인 전력 사용량 추세에 비하여
주말의 전력 사용량 추세가 과소추정됨

위도 / 경도

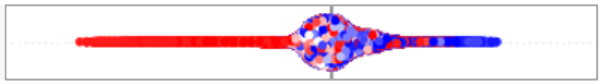


남쪽에 위치할수록, 동쪽에 위치할수록
전력 사용량이 높은 변동성을 보임

2차 모델의 SHAP Value

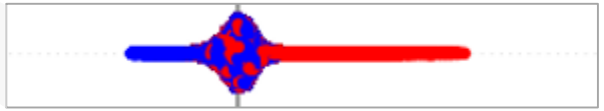


sin(시간)



오후에 비해 오전의 전력 사용량이
과소추정됨

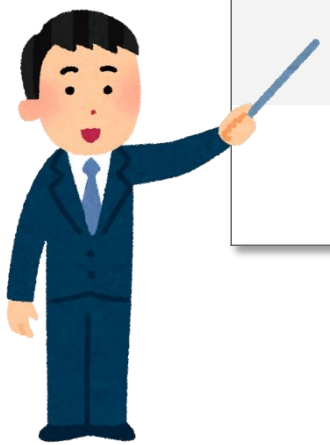
수도권 여부



수도권은 비수도권에 비해
전력 사용량이 과소추정됨

기상 변수와 전력량의 관계 해석

기상 변수	1차 모델 결과 해석	2차 잔차 모델 결과 해석
체감온도 (ta_chi)	전력 사용량과 양의 관계를 가짐	**높은 체감온도**에서 1차 모델 전력량 **추정치에 대한 변동성 심화**
기온 (ta)		**높은 기온**에서 1차 모델 전력량 **추정치를 과대추정 경향**
습도 (hm)		**낮은 습도**에서 1차 모델 전력량 **추정치를 과소추정 경향**
풍속 (ws)		변수 자체의 중요도 낮음



기상변수들은 전반적으로 전력 사용량과 양의 관계를 가지고 있으며,

그 중에서도 특히 체감온도의 영향력이 두드러짐

활용 방안의 실용성 및 실현 가능성

01

예측 모델 자체의 실용성

고차원 데이터에 강건한 머신러닝 모델 (LGBM, CatBoost) 의 특성상, 현실의 다양한 데이터를 유연하게 추가할 수 있어 높은 가변성을 보임

LGBM과 CatBoost의 GPU를 활용한 연산을 통해 현실의 대용량 데이터를 신속하게 처리 가능

지역 변수를 통해 관측 지점에 따른 지역적 영향력 반영 가능

02

변수 영향력의 해석을 통한 문제 해결

모델링을 통해 도출된 변수 영향력을 확인하는 과정에서,
전력 소비의 주체가 체감하는 불쾌와 같은 감정이 전력 소비에 큰 영향을 미침을 확인

‘사람이 느끼는 더위를 줄일 수 있는 방법’ 을 구심점으로 삼아 여름철 전력 소비량 감소 방안을 고안할 수 있음

예) 여름철 불쾌감을 극대화하는 열섬현상 해소를 위해 녹지를 조성하여 불필요한 냉방으로 낭비되는 전력을 줄일 수 있도록 함

기대효과

시공간 정보를 반영한 **고성능 전력 수요 예측 모델**의 구축 및

전력 수요에 대한 다각도의 원인 분석 & 전력난 해소 인사이트 제시

소비자

- 전력 사용의 안정화로 인한 삶의 질 향상
- 안정적 전력 공급으로 인한 전기세 절감
- 전력 소비 요인 인사이트를 통한 생활습관 개선

한국전력공사

- 선제적 수요예측을 통한 전력 공급의 안정화
- 안정적인 전력 공급을 통한 에너지 계획 수립 용이
- 기상 정보를 바탕으로 한 친환경 에너지 정책 수립

정부

- 저비용의 전력 절감 정책을 통한 에너지 효율 제고
- 에너지 효율화를 통한 온실가스 감축 및 국제사회 위상 제고
- 전력에 대한 지역 균형 발전 방안 마련

기상청

- 기상 변수의 활용 다양화로 인한 기상산업분야 활용 확대
- 전력 데이터와의 연계를 통한 에너지 기상 분야의 발전 가능성
- 다양한 데이터 연계를 통한 기상 문제 해결방안 모색