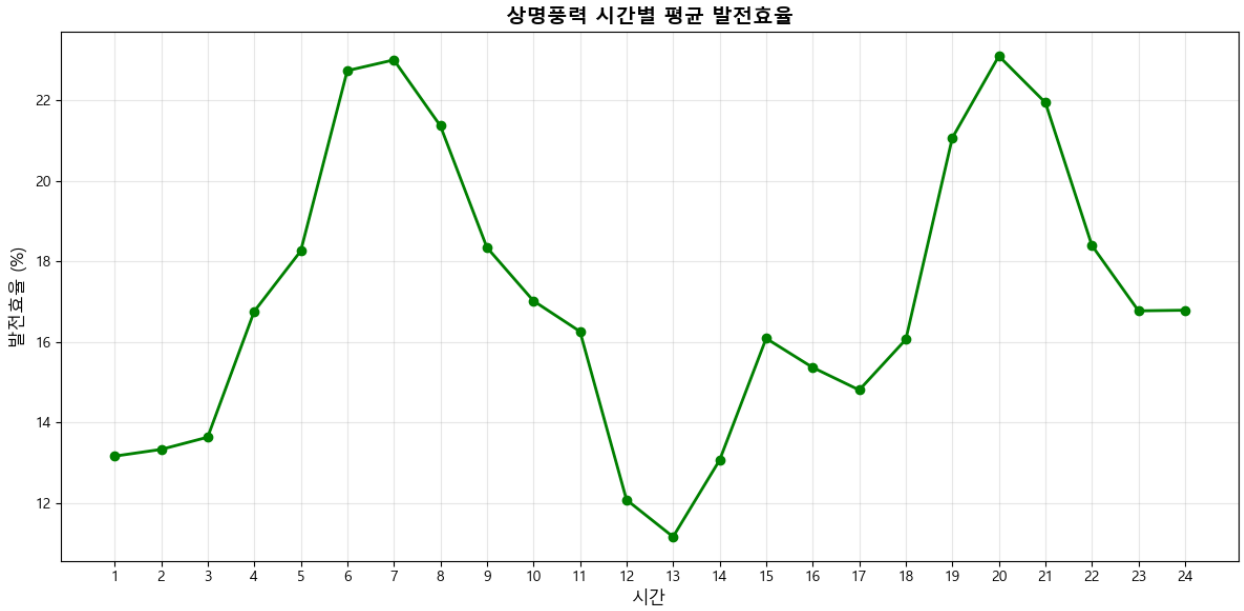
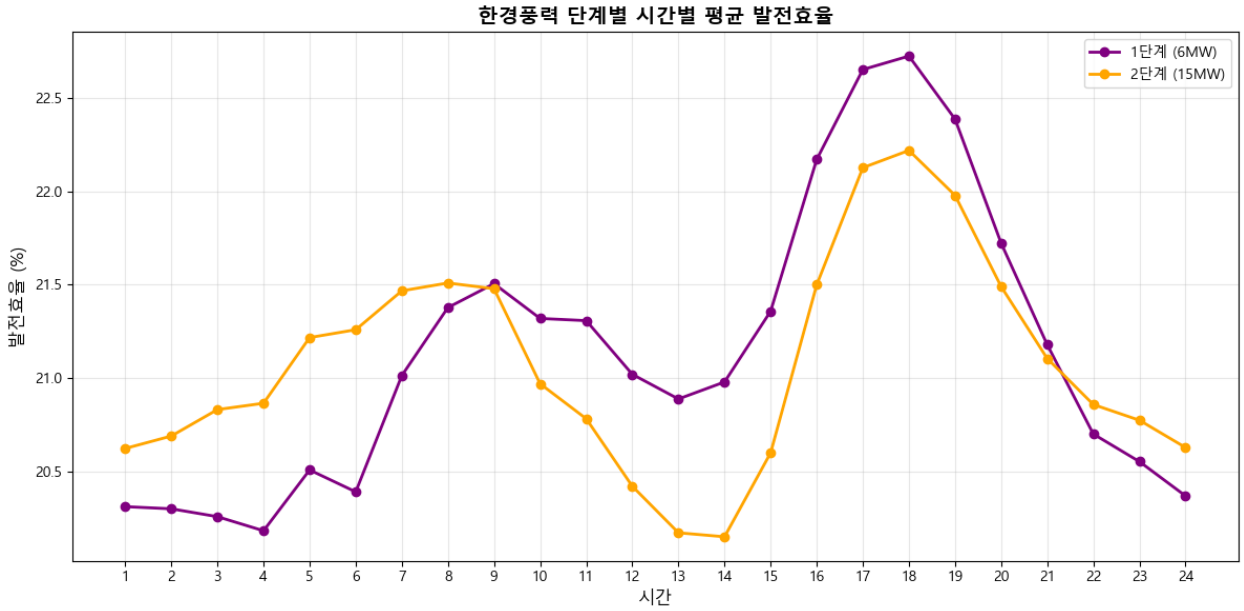
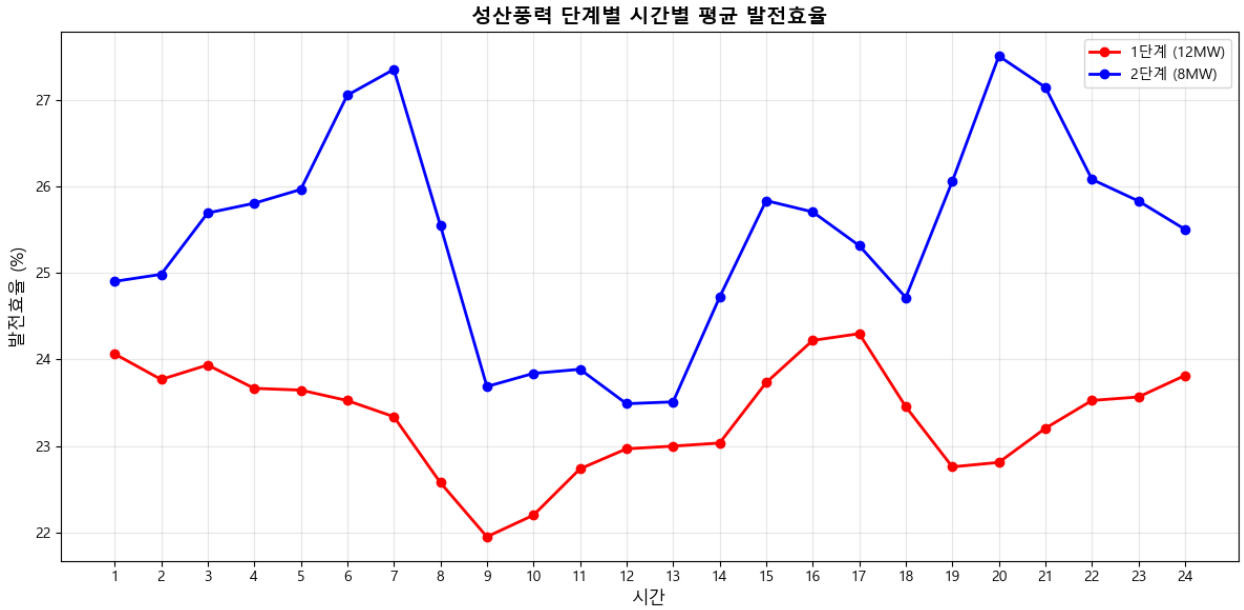
XGBoost 기반 풍력발전량 예측 모델 분석 보고서 추가 자료

1. 커트인·커트아웃 기준 기반의 이상치 판별 및 정제 전략



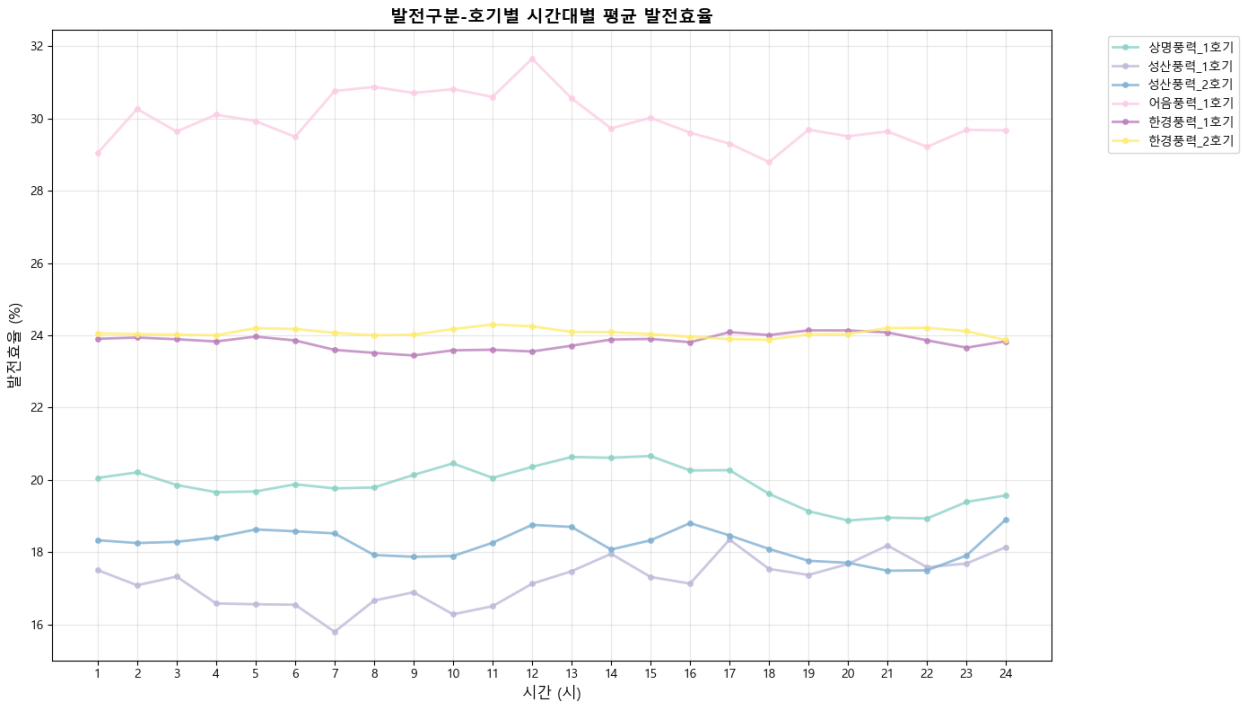




초기 데이터 분석 단계에서 발전 구분별 효율을 분석한 결과, 특정 시간대에 효율이 비정상적으로 낮은 패턴이 관측되었습니다. 이를 통해 일부 발전소에서 특정 시간에 출력 제어가 이루어지고 있음을 확인하였습니다.

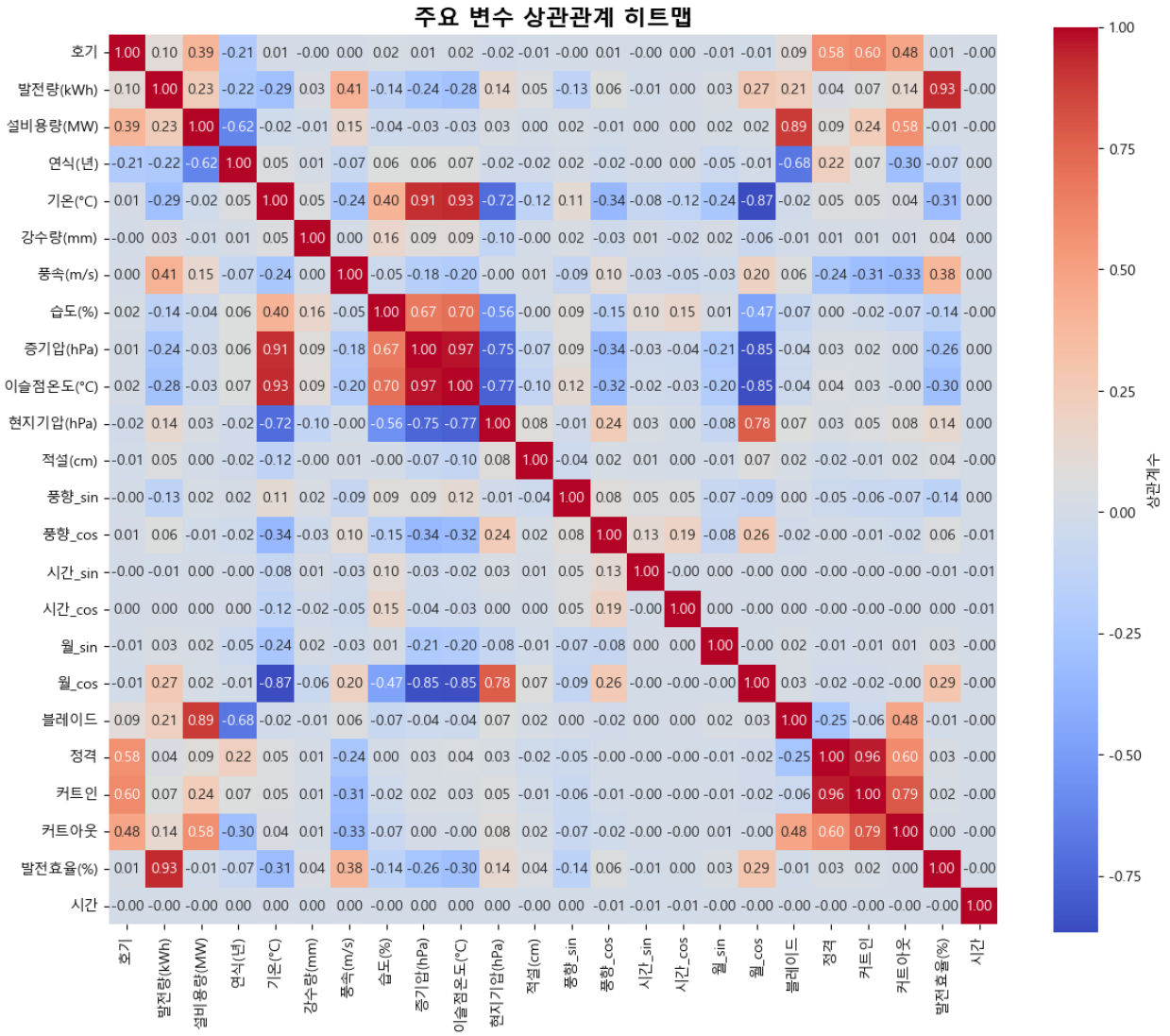
이에 따라, 풍력 터빈의 모델별 커트인(Cut-in), 커트아웃(Cut-out), 정격 풍속(Rated Wind Speed) 등의 운전 조건을 참고하여, 풍속이 커트인~커트아웃 범위에 있음에도 발전량이 0 인 경우를 이상치로 간주하였습니다. 해당 조건에 부합하는 데이터는 출력 제어로 인한 비정상적인 정지 상태로 판단하고, 전처리 과정에서 필터링하였습니다.

이러한 과정을 통해, 외부 요인에 의한 비가동 구간을 제외하고, 실제 터빈이 가동 중일 때의 순수한 발전 데이터를 확보할 수 있도록 하였습니다.



결과적으로, 출력 제어로 인한 비정상적인 발전량 0 값을 제거함으로써, 보다 신뢰도 높은 가동 구간 중심의 학습용 데이터를 구축하였습니다.

2. 상관계수 기반 발전량 결정 요인 탐색



모델링 이전 단계에서 설비용량, 풍속 등과 발전량 간의 상관관계를 파악하기 위해 상관계수 히트맵을 통해 주요 변수들의 연관성을 시각적으로 분석하였습니다.

---

3. XGBoost 기반 최종 모델 결정 근거

Model	MAE	RMSE	R^2 Score	nMAE (%)	nRMSE (%)	MBE	Eff. MAE	Eff. RMSE	Eff. Median Err	Negative Predictions
XGBoost	79.4870	152.9992	0.9374	1.9456	3.7450	0.7167	34.6491	63.4402	13.7715	5410
LightGBM	80.9317	155.0967	0.9357	1.9810	3.7963	1.2912	35.4355	64.7037	14.2178	5675
CatBoost	90.7126	174.9963	0.9181	2.2204	4.2834	-2.7206	39.4667	72.0469	15.0051	5970
RandomForest	92.2621	176.4665	0.9168	2.2583	4.3194	0.5174	40.1654	73.0006	15.7357	0

1. 모델 선정의 제 1 원칙: 예측 정확도

본 프로젝트의 핵심 목표는 발전량을 가장 정확하게 예측하는 것입니다. 따라서 모델 선정의 제 1 원칙을 예측 오차(RMSE, MAE)를 최소화하는 것으로 설정했습니다. 4 가지 주요 모델(XGBoost, LightGBM, CatBoost, RandomForest)에 대한 홀드아웃 검증 결과, XGBoost 가 거의 모든 핵심 성능 지표에서 우수한 성능을 보였습니다.

2. RandomForest 에 대한 고찰 및 기각 사유

특히 RandomForest 모델은 예측 과정에서 음수값을 생성하지 않는 안정적인 특성을 보였습니다. 이는 별도의 후처리 없이 결과를 해석할 수 있다는 장점이 있습니다.
그러나 본 프로젝트에서는 발전량의 음수 예측값을 사후적으로 0 으로 클리핑한 상태에서 지표를 계산하였으며, 그럼에도 불구하고 XGBoost 가 주요 성능 지표에서 RandomForest 를 상회하였습니다.
따라서, 예측값의 물리적 타당성이라는 부수적 안정성보다는 정확도 중심의 핵심 성능을 우선하여 XGBoost 단일 모델을 최종 모델로 채택하였습니다. 별도의 앙상블은 필요하지 않다고 판단하였습니다.

4. K-Fold 교차검증 결과 및 결과 비교

```
[INFO] K-Fold 교차검증 시작 (평가지표: nMAE, R²)...
   Fold 1 nMAE (range): 1.95%, R²: 0.9376
   Fold 2 nMAE (range): 1.93%, R²: 0.9400
   Fold 3 nMAE (range): 1.95%, R²: 0.9391
   Fold 4 nMAE (range): 1.97%, R²: 0.9369
   Fold 5 nMAE (range): 1.96%, R²: 0.9395

[교차검증] 평균 nMAE (range): 1.95% ± 0.01%
[교차검증] 평균 R²: 0.9386 ± 0.0012

[INFO] 최종 성능 평가를 위한 모델 학습 중...
  - 결정계수 (R² Score): 0.9374
  - 정규화 MAE (nMAE): 1.95%
  - 정규화 RMSE (nRMSE): 3.74%
  - 평균 편향 오차 (MBE): 0.7167
==================================================
```

1. 교차검증을 통한 모델 안정성 입증

데이터 분할 방식에 따라 성능이 좌우되는 우연성을 배제하고, 모델이 가진 본연의 일반화 성능을 측정하기 위해 5-fold 교차검증을 실시했습니다.

- 평균 R^2 Score: 0.9386 (± 0.0012)
- 평균 nMAE: 1.95% (± 0.01%)

가장 주목할 점은 표준편차(±) 값이 낮다는 것입니다. 이는 학습 데이터셋이 어떻게 구성되더라도 모델의 성능이 거의 변하지 않고 일관되게 유지됨을 의미합니다. 이 결과는 본 모델이 특정 데이터에만 과적합(overfitting)되지 않았으며, 어떠한 데이터가 입력되어도 꾸준한 성능을 기대할 수 있는 매우 안정적이고 신뢰도 높은 모델임을 증명합니다.

2. 최종 모델 성능 평가

전체 데이터의 20%를 별도로 분리한 검증 데이터셋(Validation Set)으로 최종 모델의 예측 성능을 평가한 결과는 다음과 같습니다.

평가지표	값	의미 및 해석
결정계수 (R^2 Score)	0.9374	모델이 전체 발전량 데이터 변동의 약 93.7%를 설명해냄을 의미합니다. 이는 모델이 데이터의 패턴을 매우 탁월하게 학습했다는 것을 보여주는 높은 수치입니다.
정규화 MAE (nMAE)	1.95%	예측 오차의 평균이 전체 발전량 범위의 1.95% 수준임을 나타냅니다. 모델의 예측이 매우 정밀하다는 것을 의미합니다.
정규화 RMSE (nRMSE)	3.74%	실제 값과 예측 값의 차이가 클수록 더 큰 패널티를 부여하는 지표로, 이 값 또한 매우 낮게 유지되어 모델의 강건성을 보여줍니다.
평균 편향 오차 (MBE)	0.7167	예측값들이 실제값보다 전반적으로 높은지 낮은지를 나타내는 지표입니다. 0 에 가까운 이 값은, 우리 모델이 과대 또는 과소 예측으로 치우치지 않고 매우 균형 잡힌 예측을 하고 있음을 시사합니다.

결론

교차검증을 통해 입증된 모델의 높은 안정성과, 최종 평가에서 확인된 뛰어난 예측 정확도를 종합해 볼 때, 본 XGBoost 모델은 풍력 발전량을 예측하는 데 있어 매우 신뢰할 수 있으며 실용적인 솔루션이라 할 수 있습니다. 특히, 교차검증과 최종 평가의 성능 지표가 거의 일치한다는 점은 모델의 신뢰성을 다시 한번 뒷받침하는 결과입니다.

```
중요 특성:
    1. 풍속(m/s)_cubed: 0.3658
    2. 설비용량(MW): 0.1739
    3. 회전체면적: 0.1186
    4. 풍향_sin: 0.0622
    5. 월_sin: 0.0311
    6. 풍향_cos: 0.0304
    7. 월_cos: 0.0294
    8. 연식(년): 0.0288
    9. air_density: 0.0221
   10. 기온(°C): 0.0194
   11. 강수량(mm): 0.0192
   12. 현지기압(hPa): 0.0179
   13. absolute_humidity: 0.0157
   14. 습도(%): 0.0137
   15. 시간_cos: 0.0135
   16. 증기압(hPa): 0.0134
   17. 이슬점온도(°C): 0.0133
   18. 시간_sin: 0.0114
```

1. 핵심 물리 법칙의 정확한 학습

모델이 가장 중요하게 판단한 상위 3 개 특성은 풍력 발전의 핵심 물리 법칙과 완벽하게 일치합니다.

- 풍속(m/s)_cubed (중요도: 0.3658): 가장 압도적인 중요도를 보인 변수입니다. 이는 발전량이 바람 속도의 세제곱에 비례한다는 풍력 발전의 원칙을 모델이 매우 정확하게 이해하고 있음을 보여주는 증거입니다.
- 설비용량(MW) (중요도: 0.1739): 터빈의 최대 출력 용량이 클수록 발전량이 커지는 것은 당연한 사실입니다
- 회전체면적 (중요도: 0.1186): 바람을 받는 면적이 넓을수록 더 많은 에너지를 포착할 수 있습니다. 모델이 이 또한 주요 변수로 학습한 것은 물리적으로 매우 타당합니다.

2. 기상 정보를 넘어선 현실 세계의 복합적 요인 반영

본 모델의 진정한 강점은 단순히 이론적인 물리 법칙과 기상 정보만 학습한 것이 아니라는 점입니다.

- 터빈의 고유 스펙 고려: 설비용량과 회전체면적은 단순히 숫자가 아니라, 각 발전소에 설치된 터빈의 고유한 하드웨어 스펙입니다. 모델이 이를 변수로 활용한 것은, 개별 터빈의 특성까지 고려하여 맞춤형 예측을 수행하고 있음을 의미합니다.

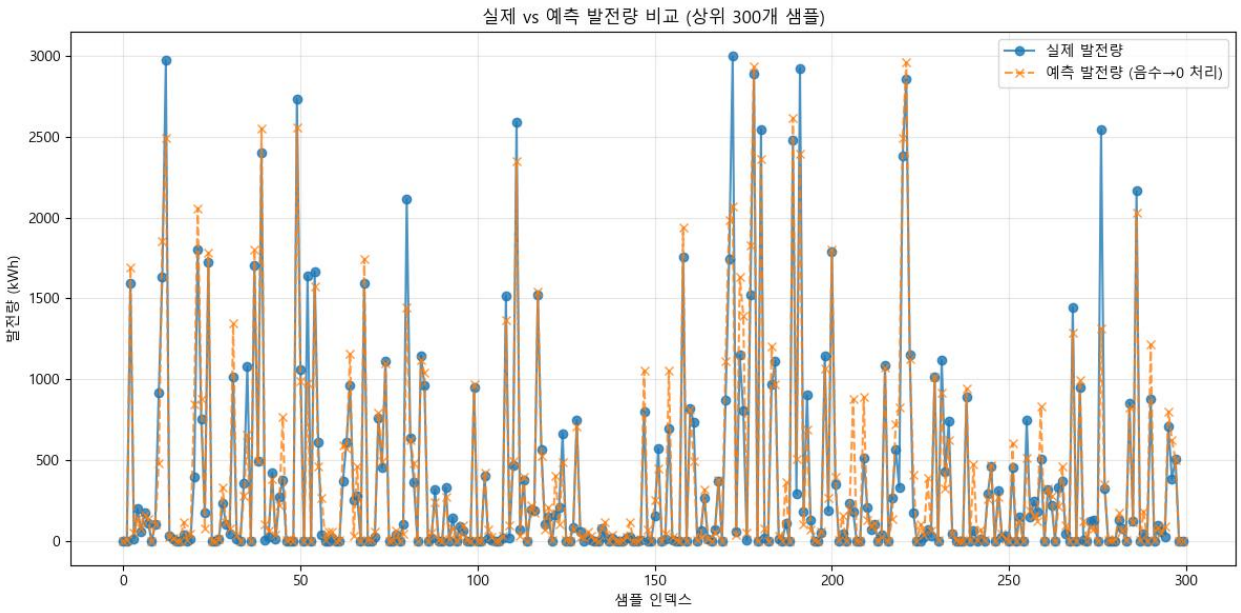
3. 정교한 기상 및 시간적 패턴 학습

상위 변수 외에도, 모델은 다음과 같은 다양한 기상 및 시간적 패턴을 정교하게 예측에 활용했습니다.

- 방향성 및 계절성: 풍향_sin/cos, 월_sin/cos 을 이용하여, 바람의 방향과 계절에 따른 발전량 패턴의 차이를 이해하고 있음을 보여줍니다.
- 대기 상태 변수: air_density(공기 밀도), 기온, 현지기압, 습도 등 공기의 상태에 영향을 미치는 다양한 기상 요소를 종합적으로 고려하여 예측의 정밀도를 높였습니다.

---

5. 최종 결과 비교

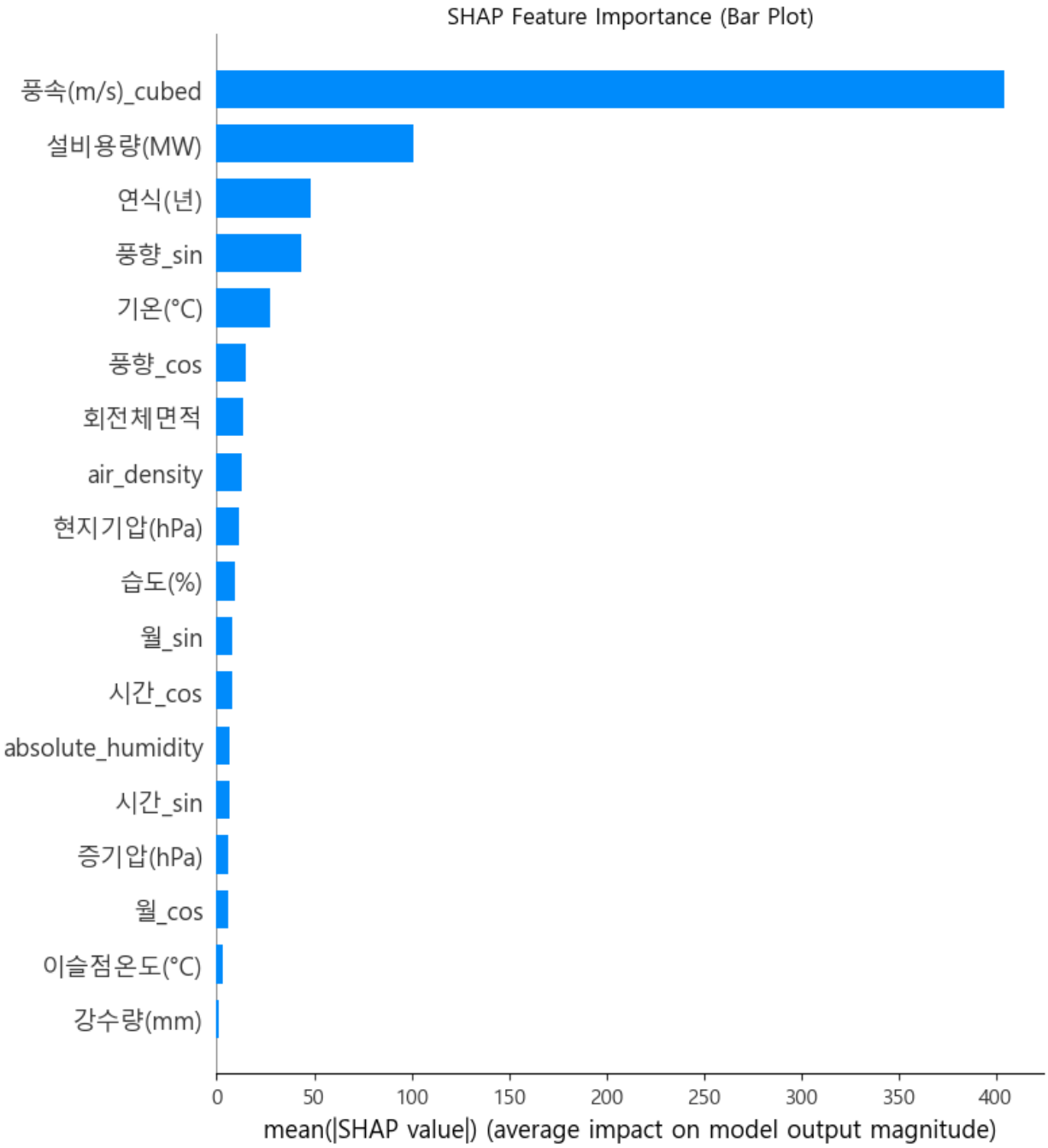


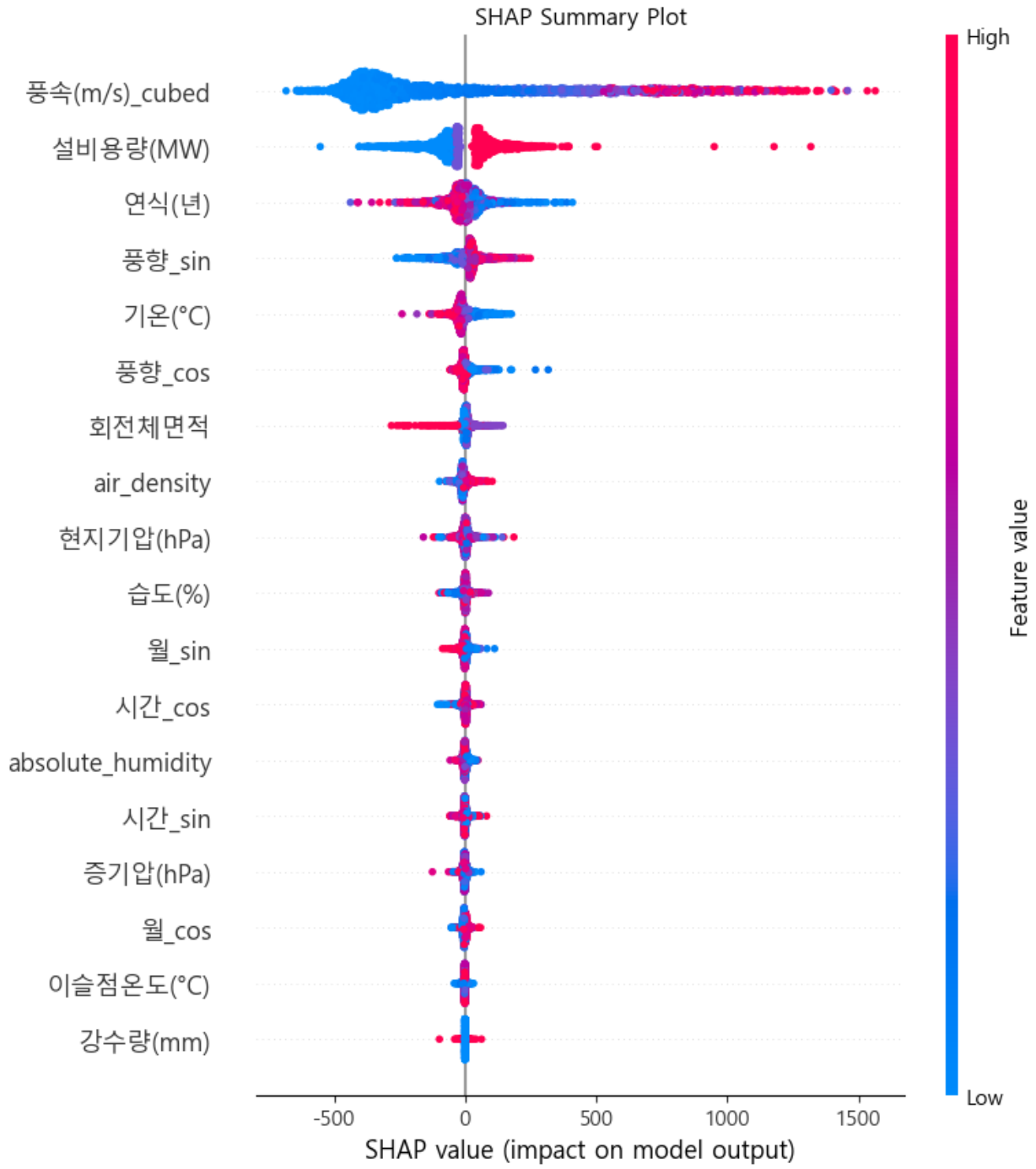
분석 및 해석

- 그래프 전반에 걸쳐 모델의 예측값(주황색 점선)이 실제 발전량(파란색 실선)의 흐름을 매우 근접하게 따라가고 있음을 확인할 수 있습니다. 모델이 발전량이 급증하는 피크구간과 발전량이 거의 없는 저점구간을 모두 효과적으로 포착하고 있습니다.

---

6. 모델 중요 피처 분석 (SHAP)





모델의 예측 근거: 물리 법칙의 재현

- SHAP 플롯을 보면, 풍속 값이 높을수록(붉은색 점) SHAP 값이 양(+)의 방향으로 크게 증가하고, 풍속 값이 낮을수록(푸른색 점) 음(-)의 방향으로 크게

감소합니다. 이는 바람이 강해질수록 발전량이 기하급수적으로 증가한다는 물리 법칙을 모델이 잘 학습했음을 보여줍니다.

- 연식(년) 변수를 보면, 연식이 오래될수록(붉은색) SHAP 값이 음(-)의 방향으로, 연식이 최신일수록(푸른색) 양(+)의 방향으로 분포하는 뚜렷한 경향을 보입니다. 이는 터빈의 노후화가 진행될수록 발전 효율이 감소하는 현실 세계의 패턴을 모델이 성공적으로 포착했음을 의미합니다.
