|  |
| --- |
| **요 약 문(풍력 예측 분야)** |

작성자 : MakinaPebbles(유현우, 전민규)

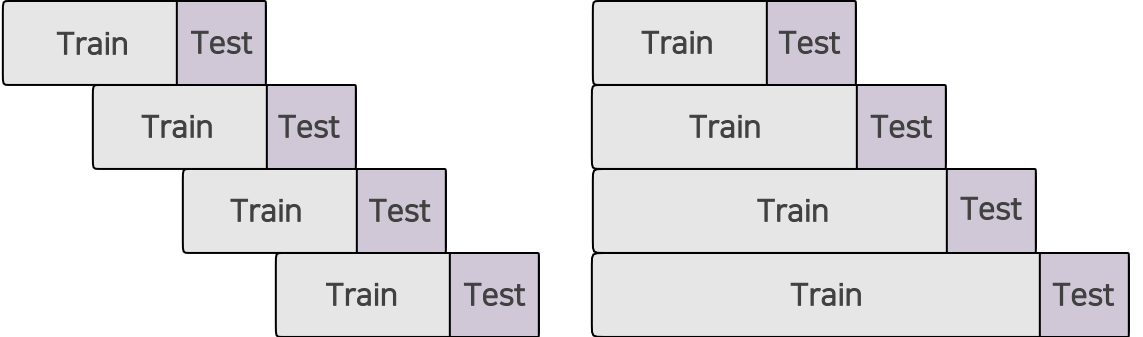
연락처 : +82-10-2720-7948(유현우)

**가. 문제 정의**

1. **Time Series Forecasting**

시간에 따라 연속적으로 변하는 를 예측하는 문제. 주어진 변수 로 를 설명하여 미래를 예측하는 문제이다. **Train Set과 Test Set의 데이터 분포(Distribution)이 다른 경우가 많으며**, 이는 Time Series Forecasting을 일반적인 Regression문제와 구분 짓는 요인이 된다. 시간에 따라 기상 예보의 분포와 발전량의 분포가 변하는 태양광 발전량 예측은 Regression이 아닌 Time Series Forecasting 문제이다.

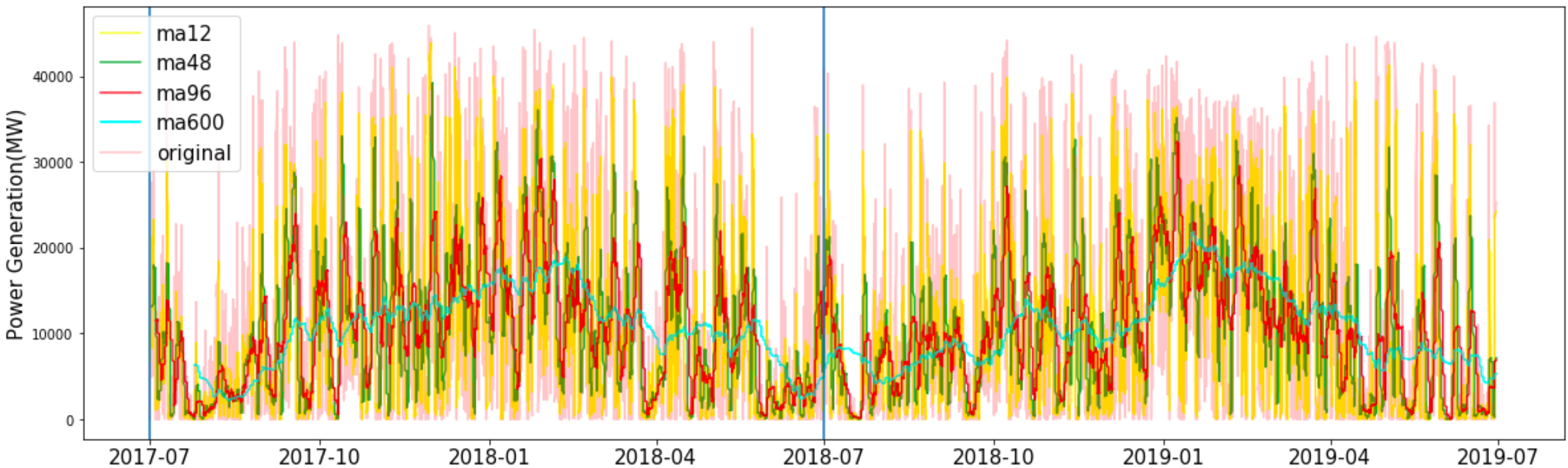
1. **Time Series Forecasting Model Validation**

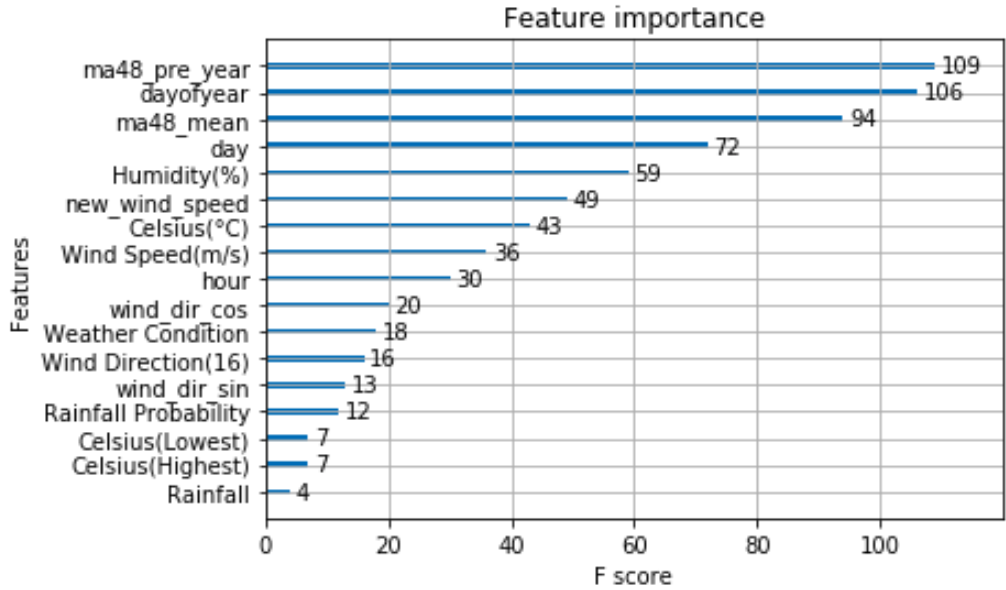


Walk-Forward Optimization Nested Cross Validation

**나. 발전량 데이터의 경향 확인**

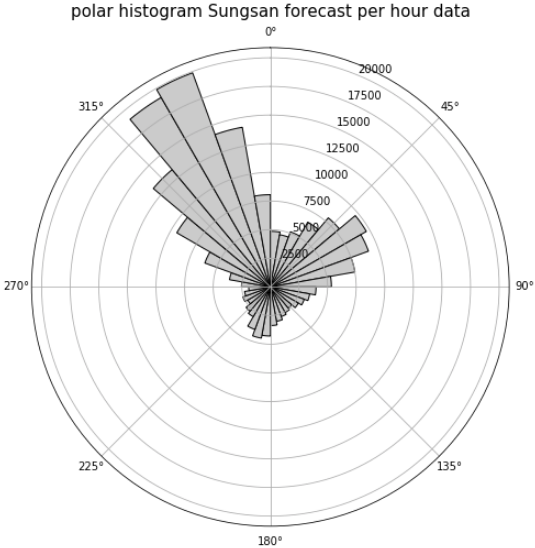
불규칙적이고 노이즈가 많이 껴있는 기상정보와 발전량 데이터를 보고 일정한 규칙성을 확인하였다.



실제로 Auto Normalized Mutual Information과 Auto Correlation값을 통해 확인하였고 이전 년도의 발전량 데이터의 Moving Average를 새로운 변수로 사용함으로써 성능 향상을 가져올 수 있었다.

또한 모든 변수들 중 가장 높은 Feature Importance를 갖는것을 확인할 수 있었다. 다른 지역에서도 비슷한 경향을 가지므로, 다른 지역의 이전의 발전량 데이터로 더 큰 성능 향상을 가져올 수 있을 것이라 기대할 수 있다.

**다. 새로운 풍속 정보 인코딩**



실제로 발전량이 풍속에 많은 영향을 받는다는 사실은 알려져있다. 그러나 실제 풍속은 풍력 터빈의 방향에 따라 발전량에 영향을 줄 수 있다. 터빈이 바라보는 쪽에서 바람이 온다면 같은 풍속이어도 다른 방향보다 더 강한 효과를 가져올 것이라고 예상할 수 있다. 따라서 이 풍향 정보를 고려한 풍속을 새롭게 인코딩하여 변수를 생성하였다.

**다. Windowing**

나.에서 발전량의 주기가 말해 주는 것은 기상정보의 주기성도 유추해볼 수 있다. 발전량은 기상정보의 영향을 받기 때문이다. 따라서 이전 시간대의 feature들을 그대로 가져와서 현재에 사용하는 방법으로 성능을 높였다. 가장 높은 성능 개선을 이룬 요인 중 하나이다.

**라. 1차 제출 예측 모델**

(1) 1st-level Model : Bayesian Optimization으로 모델별 최적의 Feature Engineering 방법, Hyperparameter 탐색

(2) 2nd-level Model : Fully Connected Neural Network로 Stacking , 최종 예측결과를 출력한다.

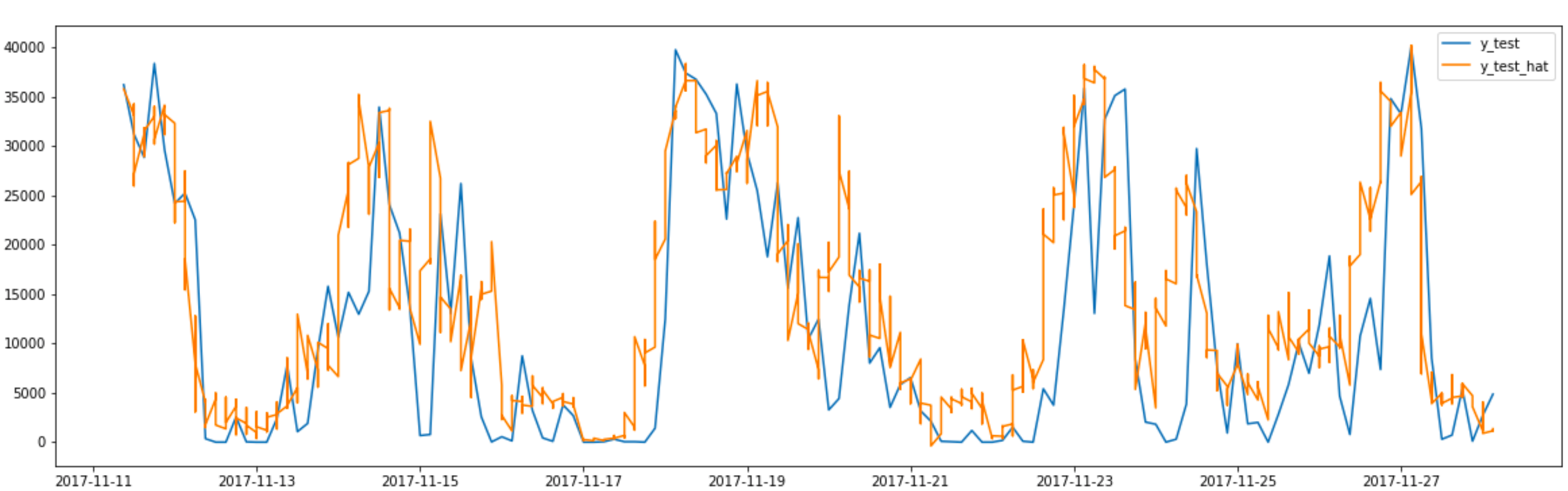
(3) Train Dataset : 전력거래소 데이터

(4) Test Dataset : 전력거래소 데이터

(5) 사용변수 : 기상예보에서 제공되는 변수, 앞서 제시된 Feature Engineering으로 만들어진 파생변수, 시간과 관련된 변수(년, 월, 일)

(6) 검증결과 (Walk Forward Validation)

Walk Forward Validation을 이용하여 2017-07부터 2019-07까지 2주 정도를 1Fold로 하여 10구간(20주)을 training하고2 주를 맞추는 방식으로 검증하였다.

<구간별 발전량과 예측값 분포>