|  |
| --- |
| **요 약 문(태양광 예측 분야)** |

작성자 : MakinaPebbles(유현우, 전민규)

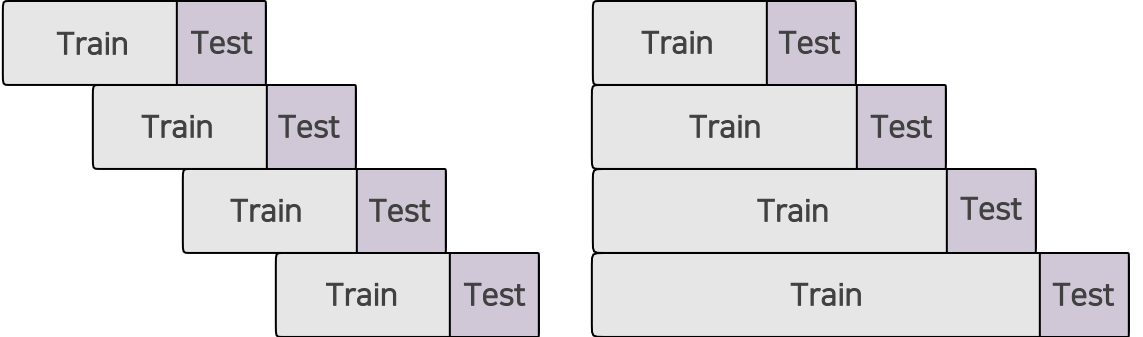
연락처 : +82-10-2720-7948(유현우)

**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. Data Augmentation (타 발전소 데이터 활용)**

|  |  |  |  |
| --- | --- | --- | --- |
| **Train Set** | **Test Set** | **Train R2** | **Test R2** |
| kpx | kpx | 0.8205 | 0.7456 |
| kpx+타발전소 | kpx | 0.7992 | 0.7585 |
| kpx+타발전소 | kpx 14시예보 | 0.7992 | 0.7603 |

빠른 실험을 위해 타 발전소 발전량을 kpx데이터의 발전량과 동일한 스케일로 스케일링하고, 카테고리 변수 Location을 추가하여 학습시켰다. 다른 조건과 설비를 가진 발전소의 데이터를 Condition(Location)을 통해 구분하며 학습하여Test Score가 향상되었다.

**다. 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.

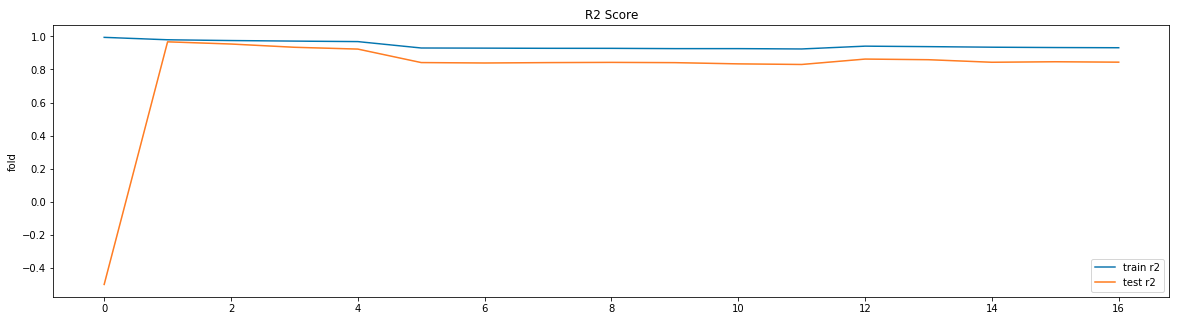
(3) Train Dataset : 전력거래소, 한국농어촌공사, 한국남부발전 데이터

(4) Test Dataset : 전력거래소 데이터

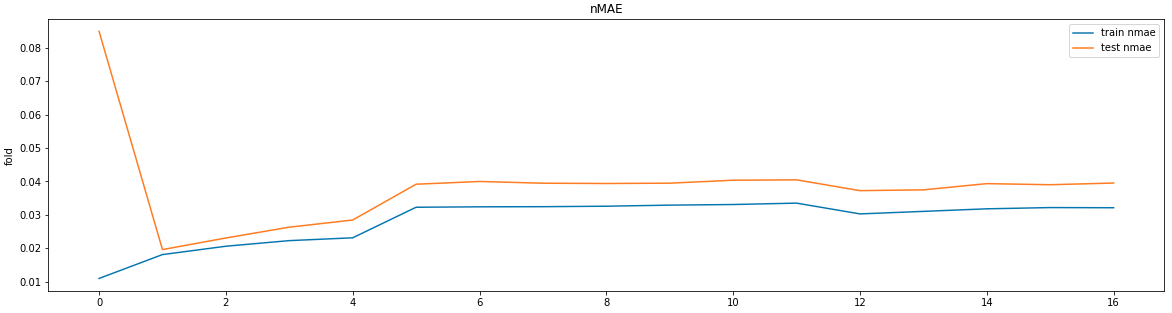
(5) 사용변수 : 기상예보에서 제공되는 변수, 앞서 제시된 Feature Engineering으로 만들어진 파생변수, 시간과 관련된 변수(년, 월, 일)

(6) 검증결과 (Nested Cross Validation)

Nested Cross Validation을 이용하여 2017-07부터 2019-07까지 1개월을 1Fold로 하여 검증하였다.



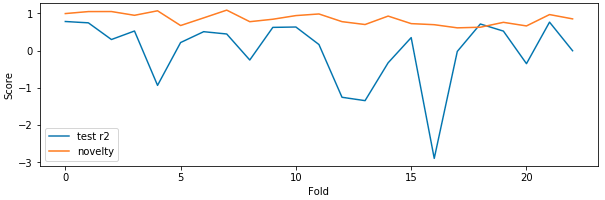
<Train R2 Score와 Test R2 Score>



<Train nMAE와 Test nMAE>

**라. 흐린날 발전량 예측이 잘 안되는 이유**

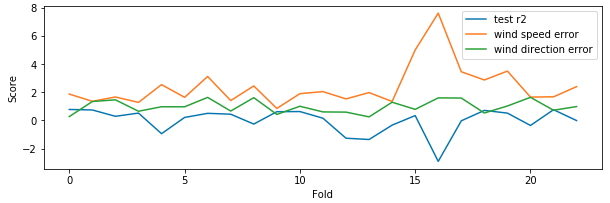
(1) Novelty (Unseen Pattern)



<Fold에 따른 Novelty Score와 Test Score>

**Stationary한 맑은날 기상의 Distribution에 비해, 흐린날의 기상은 훨씬 복잡하고 다양한 패턴을 가지고 있다**. **Unseen인 Pattern이 많을수록 예측성능이 떨어지는 것을 확인할 수 있다.** 모델을 실제로 운용할 때에는, **Novelty Score를 발전량 예측의 신뢰도로 사용하여 모델의 사용편의성을 높인다.**

(2) 기상예보의 부정확성

****

<Fold에 따른 풍속, 풍향의 관측치-예보간 오차와 Test Score>

**모델이 기상예보와 발전량의 관계를 제대로 학습하였으나, Test Set의 Input인 기상예보가 부정확하여 예측에 실패하는 경우이다**. 위의 그래프는 과거의 기상 관측치와 기상예보가 얼마나 틀렸는지를 나타낸 것이다. 풍속과 풍향의 Error가 증가할때, Test Score가 크게 떨어지는 것을 확인할 수 있다. Correlation을 보았을때도 Test Score와 풍속, 풍향의 값이 각각 -0.5336, -0.2129로 Error가 클수록 Test Score가 떨어지는 것을 알 수 있다.

반대로 **Train Set에 틀린 예보가 들어가 모델이 잘못된 관계를 학습하기도 한다. 이런 경우는 Train Set의 관측치-예보간 오차 Score를 조사하여 예측이 크게 틀린 예보를 Train Set에서 배제함으로써 방지할 수 있다.**