# 시계열 데이터를 활용한 건물의 전력 사용량 예측

데이터과학부 김동현

## 배 경

**여름철 전력 사용량** 증가로 인해 블랙아웃과 같은 사회적 문제가 대두되면서, 전력 사용량 예측의 중요성이 점점 커지고 있습니다.

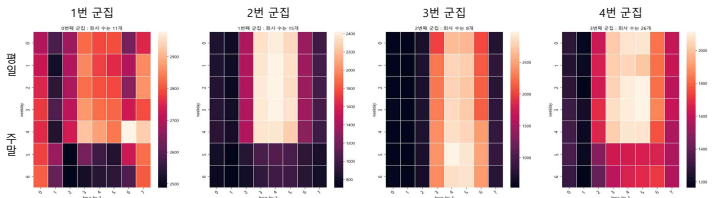
**전력사용량을 정확히 예측하면** 적절한 전력 배분과 함께 요금을 사전에 계산할 수 있어 **효율적인 에너지 관리**가 가능합니다.

이에 따라, **2020년 6월 1일부터 8월 24일까지의 시간별 데이터를** 활용하여 **60개 건물의 기상정보, 건물정보, 전력 사용량을 바탕으로 전력 사용량을 예측하는 모델을 개발**하고, 예측 정확도를 높이는데 중점을 두었습니다..

※ 본 프로젝트는 데이콘에서 개최한 전력 사용량 예측 AI 경진대회의 데이터를 활용하였습니다.

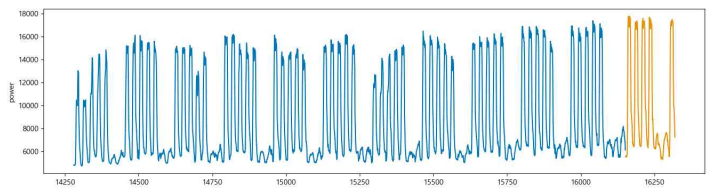
## EDA(Exploratory Data Analysis)

1. 건물의 용도에 따라 전력 사용량에 차이가 있을 것으로 판단하여, **시간과 요일별로 전력 사용량 패턴이** 유사한 건물들을 **군집화**하여 **4개의 군집**을 생성하였습니다. **x축은 시간, y축은 요일**을 나타내며, **색이 연할수록 전력 사용량이 많습니다**.




**1번 군집**: **평일에 전력소모가 크며, 특히 새벽과 낮에 전력 소모가 큼**(공장)
**2번 군집**: **평일 주간에 전력 소모가 크며**, 주말과 밤에는 전력 소모가 적은 건물(학교)
**3번 군집**: **낮에만 전력 소모가 큰 건물,** 평일과 주말 구분 없음.(식당, 카페)
**4번 군집**: 밤에 전력 소모가 적고, **낮 시간대에 전력 소모가 큰 건물**(쇼핑몰)

2. 특정 건물의 전력 사용량 데이터에서 **시간대별, 요일별 주기성을 확인**하기 위해 분석을 진행하였습니다. x축은 시간, y축은 전력 사용량을 나타내며, **노란색 부분은 일주일간의 전력 사용량**을 강조하여 보여줍니다.




**평일**에는 규칙적으로 **전력 사용량이 증가와 감소를 반복하는 주기**성을 보였으나, **주말**에는 **전력사용량이 평일 대비 낮은 수준**을 유지하는 경향이 나타났습니다.

이러한 주기적인 전력 사용량 패턴을 분석함으로써, **시간대, 요일, 주말 여부 등의 특성을 반영한 데이터 전처리 및 모델링 전략**을 수립할 수 있습니다.

## 모 델 설 계

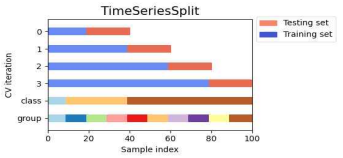
모델의 성능을 향상시키기 위해 다음과 같은 세 가지 접근 방식을 적용했습니다.

1. 전체 데이터를 기반으로 학습한 모델 (기본 모델)
2. 군집화 기법을 사용해 유사한 건물들을 그룹화한 후, 각 그룹에 대해 별도의 모델을 학습 (총 4개의 모델)
3. 군집화 없이 각 건물 번호별로 개별 모델을 학습 (총 60개의 모델)

3가지 모델에 대해 **LGBM 모델**을 사용하여 예측을 진행하였습니다.

1. **K-Fold 방식**

시계열 데이터의 특성을 반영하기 위해



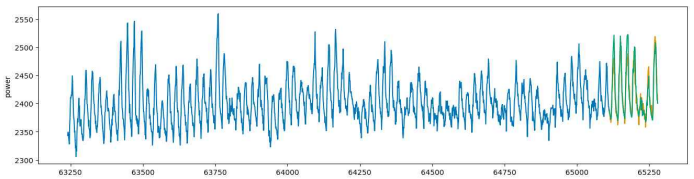

**Time Series split K-fold** 교차 검증 방법을 적용하였으며, 데이터는 기업별로 스택 되어 있기 때문에 날짜순으로 정렬한 후 time series split 방식으로 모델을 나누었습니다.

Validation 세트는 1주일 단위로 설정하였고, 모델의 성능 비교를 위해, 평가지표로 SMAPE를 사용하여 세 가지 모델을 평가 하였습니다.
※ SMAPE는 MAPE(Mean absolute percentage error)의 개량형으로 오차율을 %로 표현하며 작은 값을 가질수록 모델의 성능이 좋다.

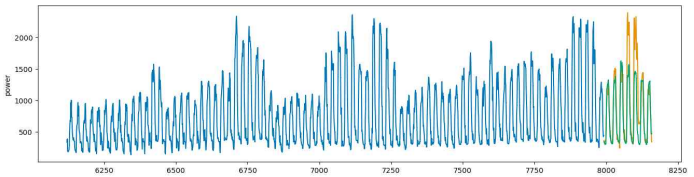
2. **XGBoost 모델 사용**

방법2에서 건물별로 60개의 모델을 생성할 경우, 각 모델이 학습할 수 있는 데이터의 양이 줄어들어 예측 정확도가 떨어질 수 있다는 가설을 세웠습니다. 이를 해결하기 위해 LGBM보다 상대적으로 작은 데이터셋에서도 성능을 잘 유지하고 과적합을 방지할 수 있는 XGBoost 모델을 사용하였고, 예측 향상이 이루어졌습니다.

아래 그림은 SMAPE값이 낮은 건물과, 높은 건물의 예측값과, 실제값을 시각화한 것입니다.




성능이 가장 좋은 건물에서는 예측값이 실제값과 거의 일치하는 모습을 보여 모델이 잘 학습 되었음을 알 수 있었습니다.




반면, 성능이 가장 좋지 못했던 건물에서는 예측값이 실제값을 대체로 따라가며 주기적인 패턴을 학습하였으나, 값이 변동하는 구간에 예측 성능이 저하되는 모습을 보였습니다.

## 결 론

이 task의 핵심은 **시계열 데이터의 특성을 이해하고 이를 효과적으로 반영해 예측 성능을 극대화**하는 것입니다.

· **Time Series split 방식**은 각 건물의 독자적인 패턴 학습에 효과적이었으며, 일부 건물(예: 32번 건물)에서는 **예측값과 실제값이 거의 일치하는 높은 성능**을 보였습니다.

· 그러나 일부 건물(예: 4번 건물)은 **급격한 변동 구간에서 큰 오차**가 발생하여 **특수한 패턴 학습에 한계**가 드러났습니다.

· **군집화를 통해 유사한 건물들을 그룹화**한 뒤 각 그룹을 개별 모델로 분석한 방식은 기본 모델보다 우수한 성능을 보였습니다. 이는 **유사한 패턴을 가진 데이터의 공통 특성을 효과적으로 학습**했기 때문입니다.

이처럼 **다 범주 시계열 데이터를 분석할 때는 각 범주의 데이터 특성을 고려**하여, **모델링 기법과 데이터 분리 전략을 세심히 설계하는 것**이 중요합니다.