

앙상블 기반 적응적 예측 모델

- 2020 꿈꾸는 아이 (AI) 본선 1라운드 기술경쟁 -

2020.12.05

팀명: FE Lab.

발표자: 송준호

Contents

Part I . 배경

Part II. 제안한 예측 모델

Part III. 예측 성능 평가

Part IV. 결론

Part V. 부록

대회 개요

■ 예측 대상

- 과거 전력 데이터를 이용하여, 200개의 아파트 및 상가의 시간별, 일별, 월별 전력 사용량 예측
 - 시간별 예측: 2018년 7월 1일 00시부터 24시까지 예측
 - 일별 예측: 2018년 7월 1일 부터 10일까지 예측
 - 월별 예측: 2018년 7월 부터 11월까지 예측

■ 데이터

- 인천 특정지역의 모 아파트 및 상가의 전력 사용량
 - Train.csv: 1300세대 (2016.07.26~2018.06.30)
 - Test.csv: 200세대 (2017.07.21~2018.06.30)
- Sampling rate: 1시간

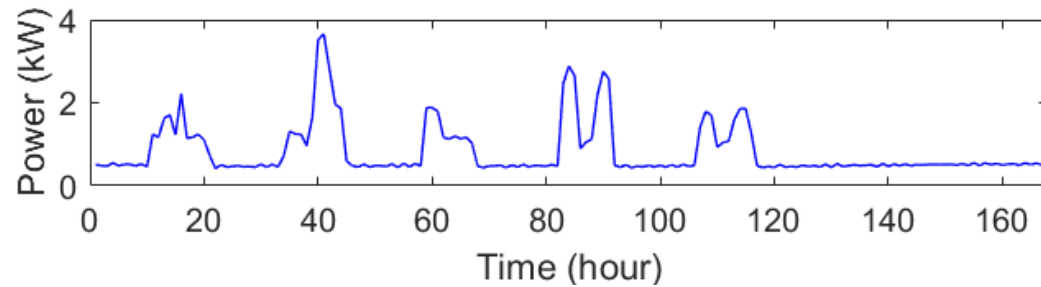
■ 평가 지표

- Symmetric Mean Absolute Percentage Error (SMAPE)

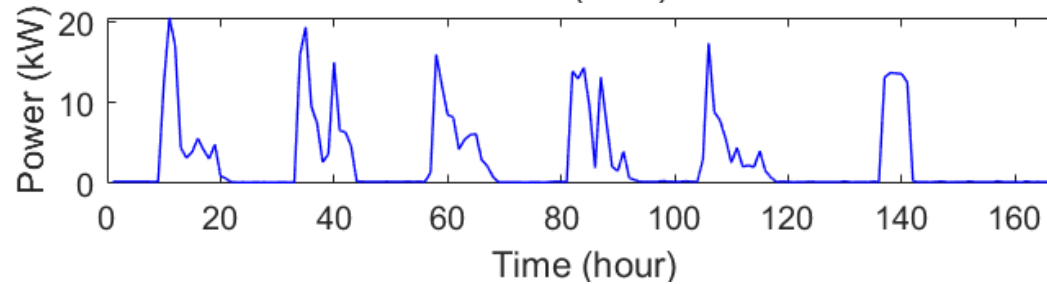
$$\text{SMAPE} = \frac{100\%}{n} \sum_{t=1}^n \frac{|X_t - \widehat{X}_t|}{(|X_t| + |\widehat{X}_t|)/2}$$

데이터 분석

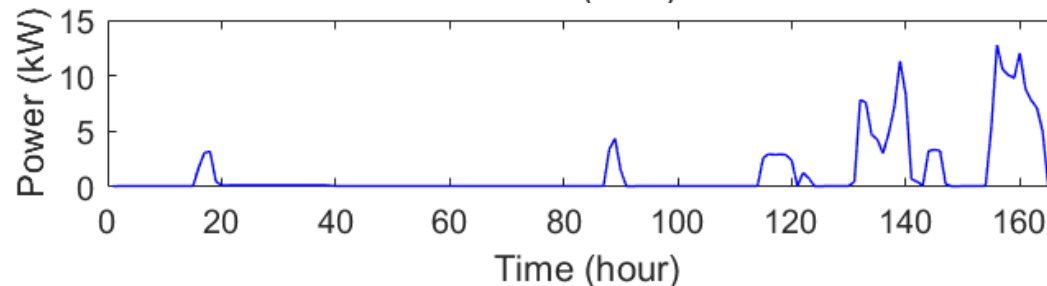
NX1312 →



NX1324 →

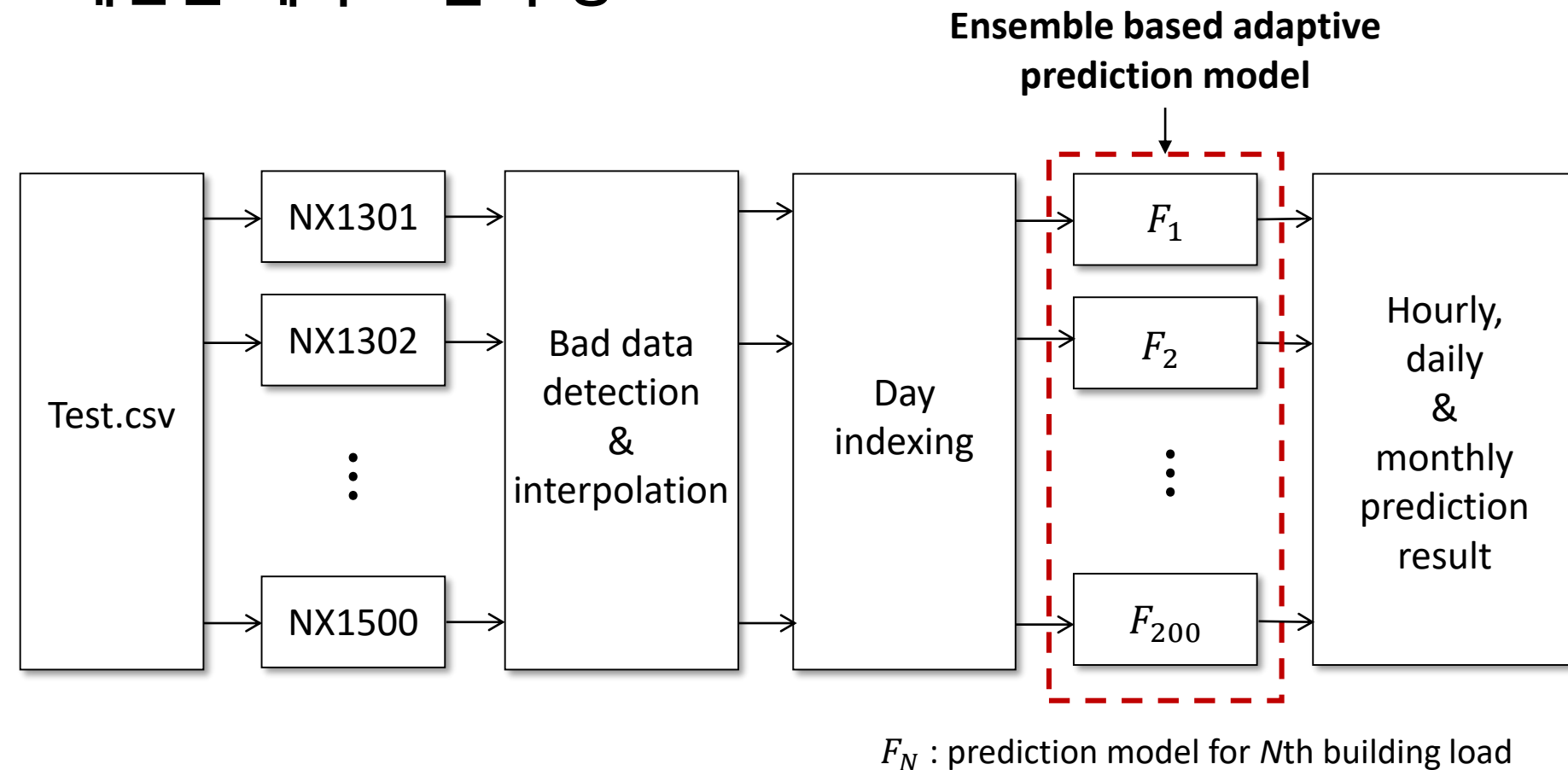


NX1326 →



- 건물에 따라 **다양한 전력 사용량 프로파일**을 보임
- 전력 사용량이 측정되지 못하여 **nan 값**으로 표시되는 부분들이 있음

제안한 예측 모델 구성도



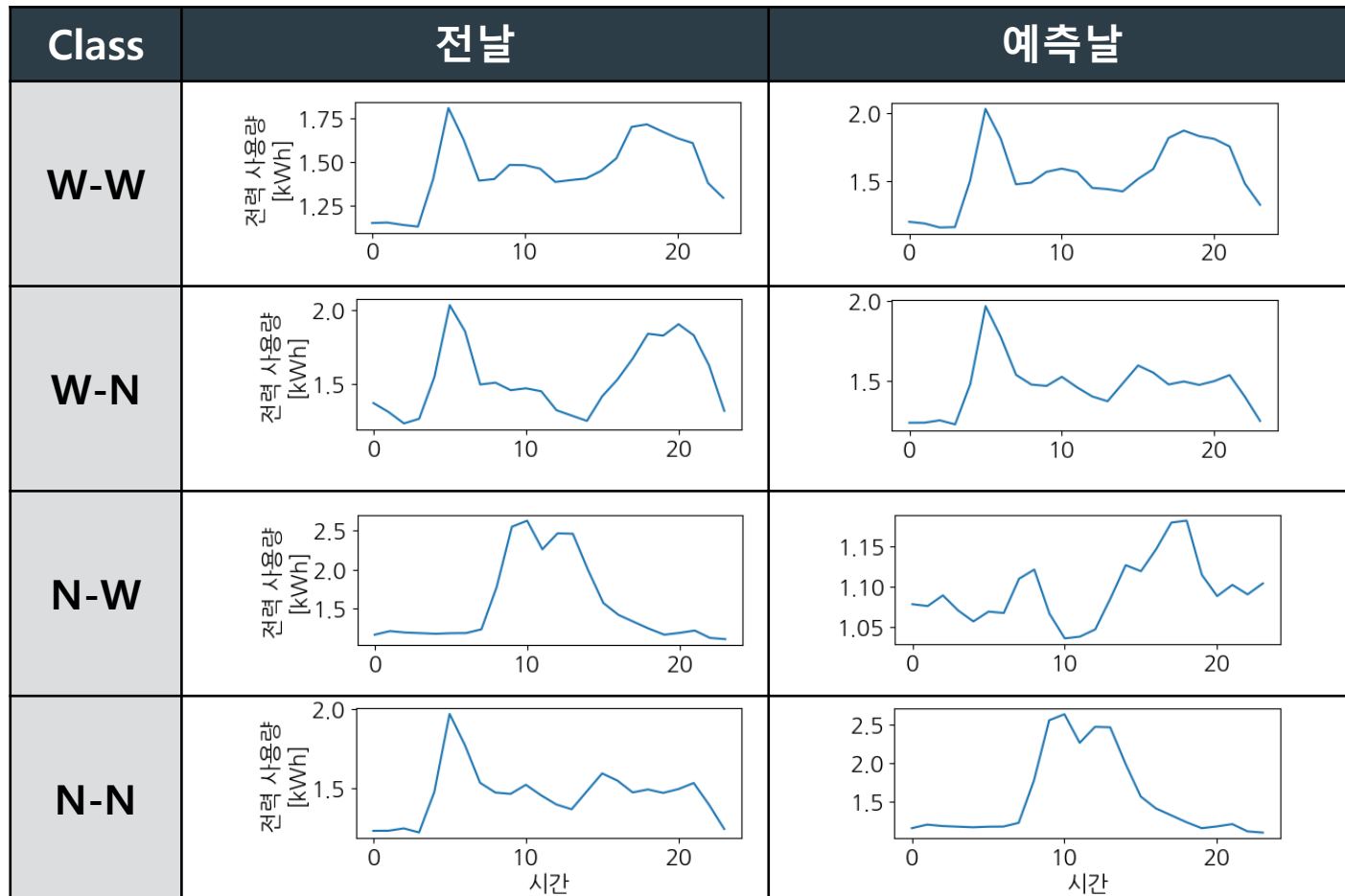
- **각 건물 부하**마다 적합한 예측 모델을 구성하여 시간별, 일별, 월별 예측을 수행

Bad data detection & interpolation

- 각 하루 전력 프로파일에 대해서 nan값 detection 수행
- 하루 전력 프로파일에서 nan 값이 특정 개수 (i.e., 4개) 이상을 넘어가면 그 날은 모델 학습에 사용하지 않음
- 하루 전력 프로파일에서 **nan 값이 특정 개수 이하**일 경우 nan 값이 있는 시간대를 기준으로 앞뒤로 20개 포인트를 이용해서 **spline interpolation**을 수행함

Day indexing

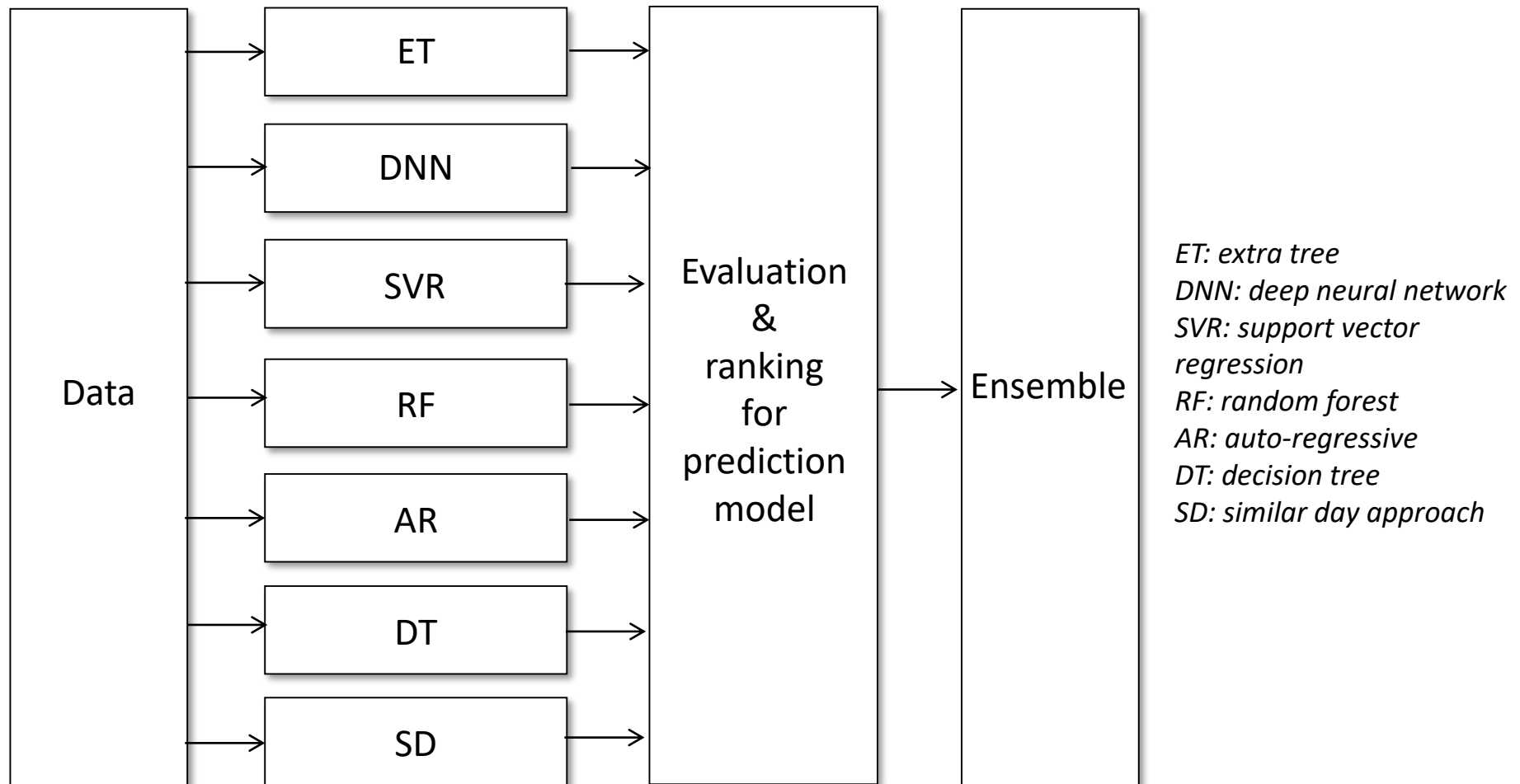
- Day type을 work day와 non-work day로 나눠 학습 및 테스트를 진행



* Work day: 평일

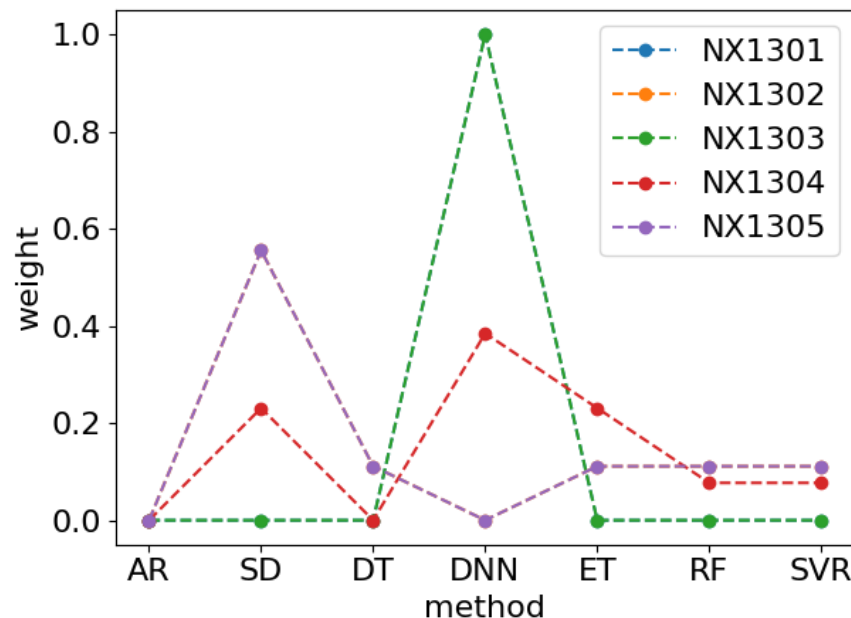
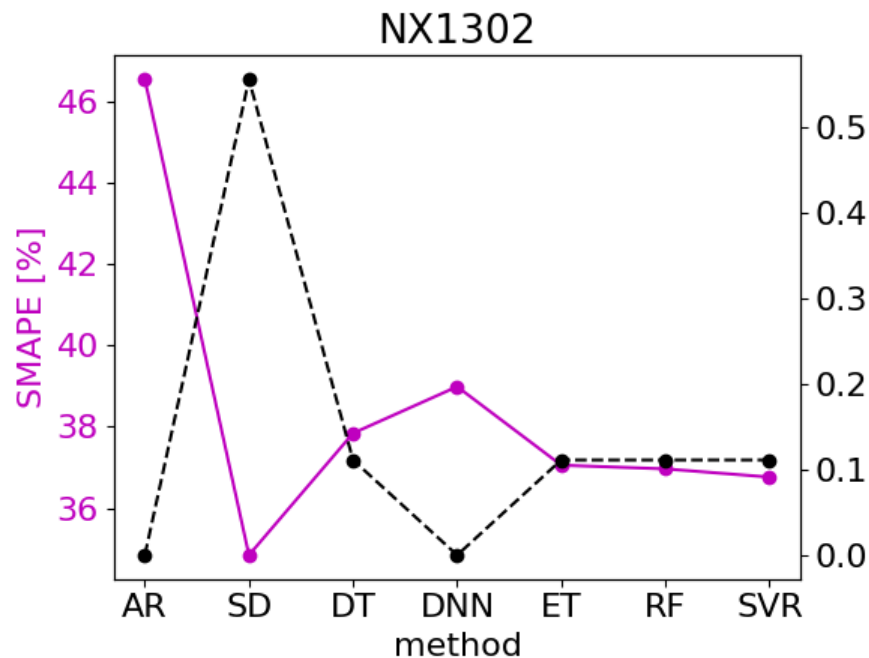
* Non-work day: 주말 및 법정공휴일

Ensemble based adaptive prediction model



- 각 건물 부하의 Validation test에서 7가지 모델을 통해 예측을 수행하고 SMAPE 값 비교를 통해 weight를 할당하여 **앙상블 수행**

Ensemble method



- DNN 모델은 multi-layer perceptron (MLP)과 long short term memory (LSTM) 중 성능이 좋은 모델을 선택함
- 각 모델의 SMAPE 결과에 따라 **weight**를 할당하여 **ensemble** 수행
- 성능이 지나치게 낮은 경우에는 해당 모델을 선택하지 않음
- 각 건물 부하마다 **개별적으로 다른 weight**를 적용함

일별, 월별 예측

■ 일별 예측 모델

- 24시간 데이터를 합하여, 하루 사용량 데이터로 만들어서 예측 모델 구현
- **Similar day approach 예측 모델**로 예측

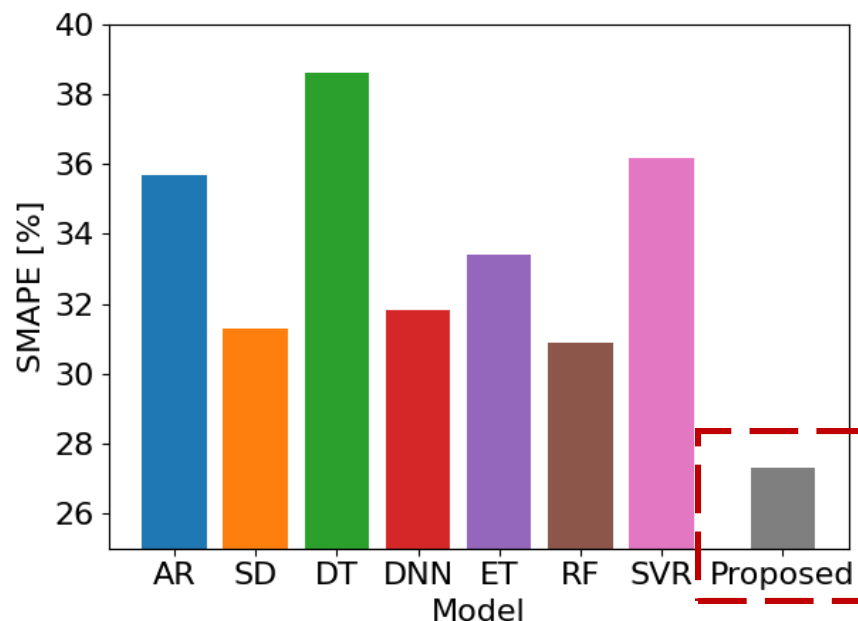
■ 월별 예측 모델

- 예측한 일별 데이터의 합을 통해 결과 예측

* 일별, 월별 예측의 경우, **데이터 수가 부족**하여 예측모델을 충분하게 학습을 수행할 수 없으므로 'similar day approach'를 활용하여 예측을 진행

예측 결과

- 제안한 앙상블 기반 예측 방법과 단일 예측 모델과의 예측 성능 비교



- Test.csv 데이터의 200개 건물 부하에 대한 평균 SMAPE 비교
- 단일 예측 모델에서 가장 예측 정확도가 높은 것은 RF이며 SMAPE는 30.8%로 나타남
- 제안한 **앙상블 기반 예측의 SMAPE는 27.3%**이며 단일 예측 RF 보다 약 **11.6% 개선됨**

DACON 리더보드 결과

Public Private









순위기

WINNER

1%

4%

10%

#	팀	팀 멤버	최종점수	제출수	등록일
1	FE lab.		26.98438	31	16일 전
2	영듀		27.3684	25	18일 전
3	shining_sunny		27.79991	15	17일 전
4	Alhard		28.59153	25	16일 전
5	수요왕		28.63262	18	16일 전
6	ImSoPa		29.37608	10	16일 전
7	WooSeok Shin		29.51326	7	16일 전
8	디엔에이		30.21613	30	16일 전

결론

- 본 대회에서 사용된 데이터는 불확실성이 크고 건물 부하 간의 다양한 특성을 가지고 있음
- 단일 모델로 예측한 결과와 제안한 앙상블 기반 적응적 예측 모델에 의한 예측 결과를 비교한 결과 제안한 방식이 예측 정확도가 더 높음
- 각 건물 부하의 전력 데이터 특성에 따라 적합한 예측 모델을 사용하는 것이 예측에 유리할 것으로 보임

**THANK
YOU**

팀 명: FE Lab.

발표자 송 준 호

제안한 방법에 사용된 예측 모델

▪ Auto-regressive (AR) prediction model

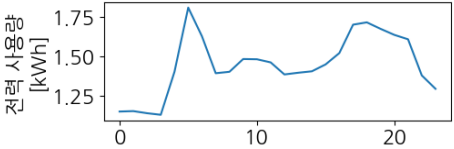
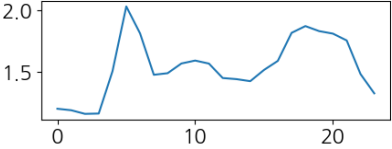
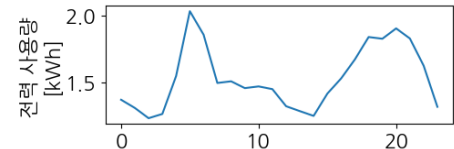
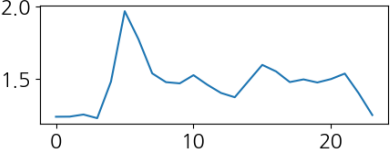
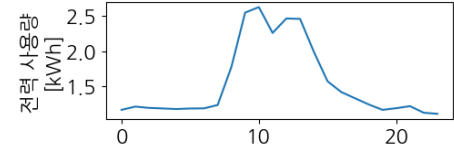
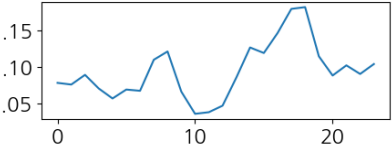
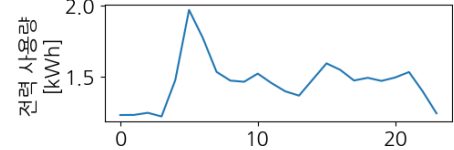
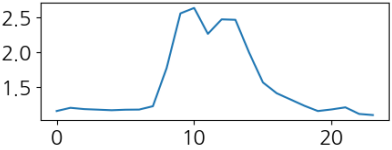
- 데이터 간의 선형 조합을 이용하여 데이터를 예측하는 모델
- Pseudo-inverse method를 이용하여 AR model의 coefficient 계산

Equation	Parameter
$P_t = c + \sum_{i=1}^n P_{t-i} * W_i + \epsilon_t$	<div> <div>[P_t]</div> <div>Power</div> </div> <div> <div>[W_i]</div> <div>Coefficient</div> </div> <div> <div>[c]</div> <div>Constant</div> </div> <div> <div>[ϵ_t]</div> <div>White noise</div> </div>

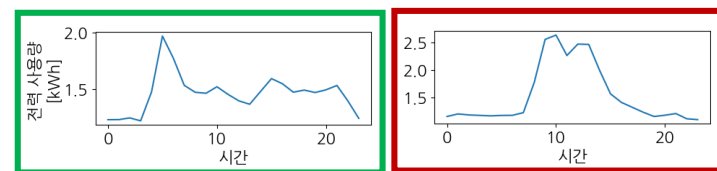
제안한 방법에 사용된 예측 모델

▪ Similar day approach (SD)

- 과거 데이터 셋을 Day type에 따라 전력 프로파일을 분류
- 예측 전날의 전력 프로파일과 가장 유사한 프로파일을 distance 계산하여 추출
- 비슷한 날의 데이터를 여러 개 추출하여 전력 프로파일을 평균하여 예측

Class	전날	예측날
W-N		
W-W		
N-W		
N-N		

Target day



제안한 방법에 사용된 예측 모델

▪ Multilayer perceptron (MLP)

- MAE loss 기반의 학습
- Adaptive Moment Estimation Algorithm (Adam) optimizer 사용
- Activation function: ReLU 사용
- Hidden layer의 개수와 unit의 개수는 예측하고자 하는 smart meter마다 개별적으로 설정

▪ Long short-term memory (LSTM)

- 첫 번째 hidden layer에 LSTM을 사용
- 모든 smart meter에 대해 동일한 hyper parameter 적용

Layer type	Unit	Activation
Input	24	
Hidden layer 1	432	ReLU
Hidden layer 2	168	ReLU
Hidden layer 3	432	ReLU
output	24	Linear

제안한 방법에 사용된 예측 모델

- **Random forest (RF)**

- 다수의 결정 트리를 통해 예측을 수행함
- MAE (mean absolute error) loss를 이용한 학습

- **Extra tree (ET)**

- Random forest보다 random성이 높음

- **Decision tree (DT)**

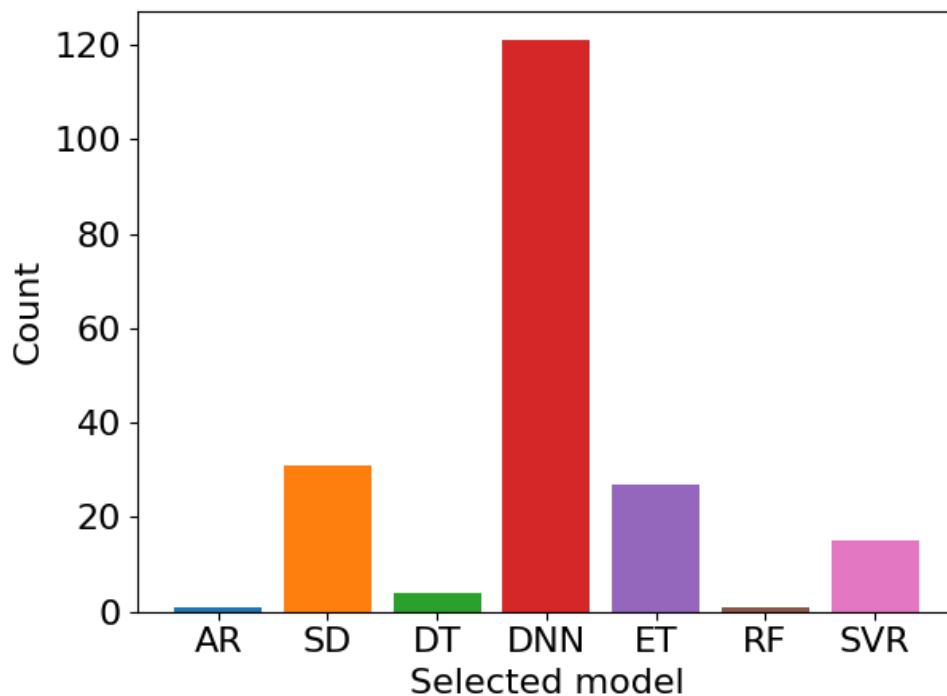
- 간단한 계산을 통해 어떤 항목에 대한 관측 값과 목표 값을 예측

- **Support vector regression (SVR)**

- Support Vector Machine이 가진 예측 능력을 바탕으로, 회귀문제 영역을 해결하기 위해 제안된 모델

예측 결과

- 개별 건물 부하에 대한 모델 선택 횟수



- Test.csv 데이터의 200개 건물 부하에 대하여 예측 모델들을 평가하였을 때 **DNN이 가장 많이 선택됨**
- SD, ET, SVR이 비슷한 정도로 선택됨
- AR, RF는 거의 선택되지 않음