스마트 빌딩 빅데이터 분석 경진대회

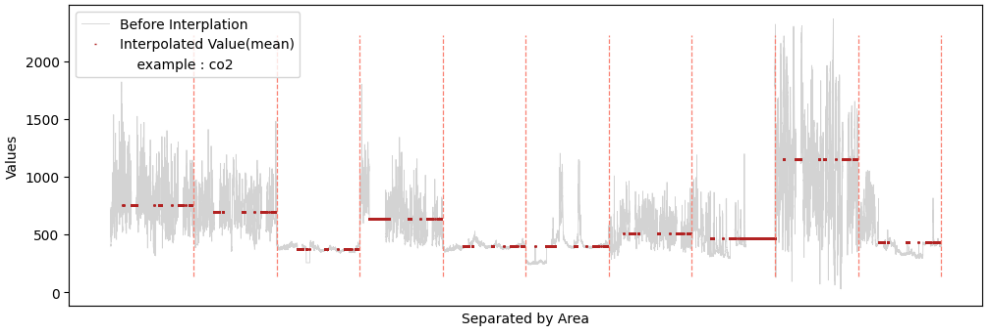
팀원 : 이윤섭, 이동현, 윤예지

1. 데이터 개요 및 진행 방향

데이터가 시계열 자료를 포함하고 있지만 시간이 반복되는 부분을 각 다른 공간으로 가정하였습니다. 데이터는 결측값을 포함하고 있으며 기온(tmp)과 습도(humi)를 제외한 변수는 첨도가 크고 오른쪽으로 꼬리가 긴 분포를 하고 있습니다.

인원 수를 세는 문제이므로 트리기반 앙상블, 인공신경망으로 회귀분석했으며 시계열을 더 반영할 수 있고 가중치를 업데이트 하는 LSTM, GRU, CNN을 최종 모델에 채택했습니다.

2. 데이터 전처리



시간을 나타내는 'regdate' 변수는 데이터셋에서 공간을 구분하고 결측값 이전 시간에서 1분씩 더하여 유실된 시간을 채웠습니다. 그 외의 변수는 공간마다 다른 특성을 가졌을 것이라고 판단해 각 측정 위치에 따른 해당 변수의 대표값으로 결측치를 채웠습니다.

'regdate' 변수는 datetime 타입으로 분석시 범주형 변수로 계산되므로 '요일'과 '시간(하루/분)'으로 나누어 파생변수를 생성했습니다. 범주형 변수인 '요일'은 분석 결과에 영향을 주지 않거나 많은 차원을 생성해 성능을 낮추기도 하였으므로 최종모델에서 제외하였습니다. 공간을 나타내기 위해 생성한 파생변수인 'area'도 같은 이유로 제외했습니다.

변수간의 단위, 범위 차이가 크기 때문에 정규화(Normalization)을 적용해 변수들의 분포를 일정하게 만들었습니다. 데이터누설(Data Leakage)를 방지하기 위해 테스트 데이터를 학습에 직접 이용하거나 날짜를 임의로 입력하는 경우를 최소화했습니다.

3. 모델 구축

	CNN (1D)	LSTM	GRU
Drop out	0.654	0.746	0.760
Hidden layers	0.915	1.715	2.376
Batch size	805.0	708.8	843.0
Neurons	14.49	16.22	23.04

각 신경망의 하이퍼 파라미터를 Bayesain Optimization으로 탐색·튜닝하였고 튜닝 범위와 반복횟수는 동일하게 설정했습니다. 은닉층의 활성함수는 ReLu를 사용하였고 출력은 항등함수($y = x$)로 나타냈습니다.

4. 모델 평가

	CNN (1D)	LSTM	GRU
MSE	**0.7947**	1.603	2.086
RMSE	**0.891**	1.266	1.444

모델의 Validation은 MSE(Mean Squared Error)로 하였으며 해석의 용이함을 위해 RMSE(Rooted Mean Squared Error)도 평가에 포함하였습니다. 학습한 딥러닝 인공신경망 중에서 CNN이 가장 높은 성능으로 미지의 데이터를 예측할때 대략 0.891명의 오차가 생길 수 있습니다.

CNN이 다른 다른 신경망보다 성능이 높게 나온 원인을 학습 데이터가 시계열을 띄고 있지만 각기 다른 공간이 중복되어있기 때문에 순차적인 구조에 의존하는 RNN 계열보다 전체 데이터를 1차원으로 인식하고 각 관측값의 특징을 추출하는 CNN이 더 유연하게 학습했기 때문이라고 평가하였습니다.