Team 902

딥러닝 기반 시간 단위 미세먼지 예보

구자경, 김민섭, 박찬동, 부다정, 임수연, 최동희

2018년 12월 2일

목 차

1. 주제 탐색 과정

주제 : 시간 단위로 미세 먼지 수치를 예보 하자.

- **미세먼지로 인해 사람들의 외부활동이 제한된다.**

미세먼지의 유해성이 널리 알려지면서 국가 예보의 파급 효과는 예상보다 매우 커졌다. 설문조사에 따르면 60세 이상 연령층의 약 70 %가 미세먼지 예보 결과에 따라 외출을 자제한다고 한다.

- **현재의 미세먼지 예보는 '예보' 아닌 '통보'이다.**

2018년 12월 현재 사용중인 베타선 흡수법은 미세먼지를 측정하는 방법일 뿐, 예측의 도구로는 사용이 불가하다.

국내에서 주로 사용되는 방법은 베타선 흡수법으로, 여과지에 먼지를 채취해 베타선을 투과시켜 질량농도를 연속적으로 측정하는 방식입니다. 자동 측정이 가능해 쉽고 편하다는 장점이 있지만 일정 시간 입자 포집 시간이 필요해 실시간 측정이 불가능한 단점이 있습니다.

황사주의보와 미세먼지주의보 차이 자료: 환경부 · 기상청

	발생 원인	예보 기관	특보 발표 기관	주의보(경보) 발령 기준
황사	중국 모래폭풍	기상청	기상청	1시간 평균 미세먼지(PM10) 농도 400㎍/㎥(800㎍/㎥) 이상이 2시간 넘게 지속될 것으로 예상될 때
미세먼지	대기오염물질	환경부 국립 환경과학원	16개 지방자치단체	1시간 평균 미세먼지(PM10) 농도 150㎍/㎥ (300㎍/㎥) 이상이 2시간 넘게 지속됐을 때

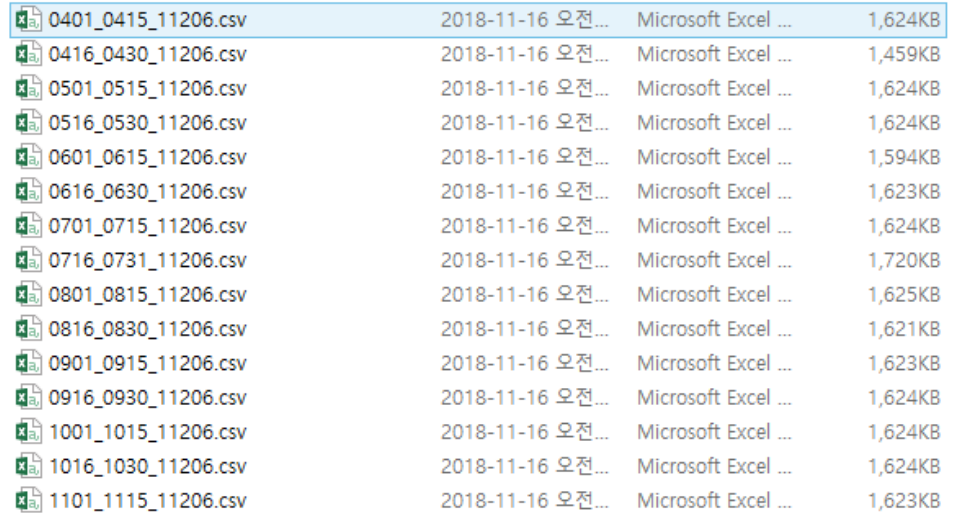
며칠째 미세먼지 예보 아닌 '중계'만…"믿고 나갈 수가 없다"

[JTBC] 입력 2018-11-29 21:02 | 수정 2018-11-29 23:55

〉실시간 미세먼지 수치 예측이 필요

2. 데이터 전처리

1) 16개 공기질 데이터 파일들을 하나의 엑셀 파일로 통합 (180401~181115)



0401_0415_11206.csv	2018-11-16 오전...	Microsoft Excel ...	1,624KB
0416_0430_11206.csv	2018-11-16 오전...	Microsoft Excel ...	1,459KB
0501_0515_11206.csv	2018-11-16 오전...	Microsoft Excel ...	1,624KB
0516_0530_11206.csv	2018-11-16 오전...	Microsoft Excel ...	1,624KB
0601_0615_11206.csv	2018-11-16 오전...	Microsoft Excel ...	1,594KB
0616_0630_11206.csv	2018-11-16 오전...	Microsoft Excel ...	1,623KB
0701_0715_11206.csv	2018-11-16 오전...	Microsoft Excel ...	1,624KB
0716_0731_11206.csv	2018-11-16 오전...	Microsoft Excel ...	1,720KB
0801_0815_11206.csv	2018-11-16 오전...	Microsoft Excel ...	1,625KB
0816_0830_11206.csv	2018-11-16 오전...	Microsoft Excel ...	1,621KB
0901_0915_11206.csv	2018-11-16 오전...	Microsoft Excel ...	1,623KB
0916_0930_11206.csv	2018-11-16 오전...	Microsoft Excel ...	1,624KB
1001_1015_11206.csv	2018-11-16 오전...	Microsoft Excel ...	1,624KB
1016_1030_11206.csv	2018-11-16 오전...	Microsoft Excel ...	1,624KB
1101_1115_11206.csv	2018-11-16 오전...	Microsoft Excel ...	1,623KB

2) 데이터 제공 기준에 맞춰(readme.txt) 날짜(년+월+일+분), PM10, PM2.5만 선택

	A	B	C	D	E	F	G	H	I	J	K	L
1	날짜			PM10								PM2.5
2	2.02E+11	V10O1611	1	35	-9999	-9999	60	1141	54	-999	-999	18
3	2.02E+11	V10O1611	1	30	-9999	-9999	54	1141	54	-999	-999	15
4	2.02E+11	V10O1611	1	28	-9999	-9999	55	1140	54	-999	-999	14
5	2.02E+11	V10O1611	1	44	-9999	-9999	57	1140	54	-999	-999	23
6	2.02E+11	V10O1611	1	31	-9999	-9999	55	1140	54	-999	-999	16

2. 데이터 전처리 (계속)

3) 분 → 시간단위 평균으로 재구성

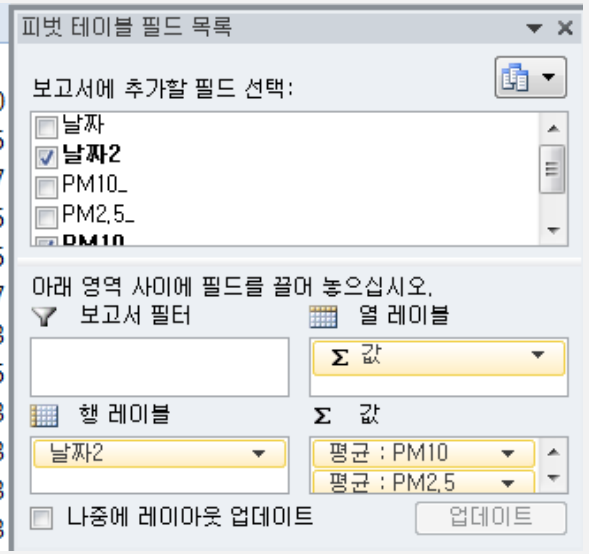
- 날　　짜

FV　=LEFT(A2,10)

	A	B	C	D
1	날짜	날짜2	PM10_	PM2.5_
2	201804010000	=LEFT(A2,10)	35	18
3	201804010001	2(LEFT(text, [num_chars]))0		15
4	201804010002	2018040100	28	14
5	201804010003	2018040100	44	23
6	201804010004	2018040100	31	16
7	201804010005	2018040100	33	17
8	201804010006	2018040100	35	18
9	201804010007	2018040100	18	9
10	201804010008	2018040100	31	16
11	201804010009	2018040100	43	22
12	201804010010	2018040100	37	19
13	201804010011	2018040100	30	15
14	201804010012	2018040100	46	24
15	201804010013	2018040100	24	12
16	201804010014	2018040100	35	18
17	201804010015	2018040100	33	17
18	201804010016	2018040100	43	22
19	201804010017	2018040100	26	13
20	201804010018	2018040100	32	16
21	201804010019	2018040100	41	21
22	201804010020	2018040100	42	22
23	201804010021	2018040100	27	14

- 미세먼지

3	행 레이블	평균 : PM10	평균 : PM2.5
4			
5	?	0	0
6	2018040100	35.9	18.45
7	2018040101	37.88333333	19.26666667
8	2018040102	38.7	19.15
9	2018040103	40.11666667	20.05
10	2018040104	44.51666667	22.96666667
11	2018040105	57.33333333	30.93333333
12	2018040106	53.38333333	29.05
13	2018040107	55.46666667	29.43333333
14	2018040108	46.65	24.08333333
15	2018040109	60.98333333	30.18333333
16	2018040110	66.66666667	29.88333333

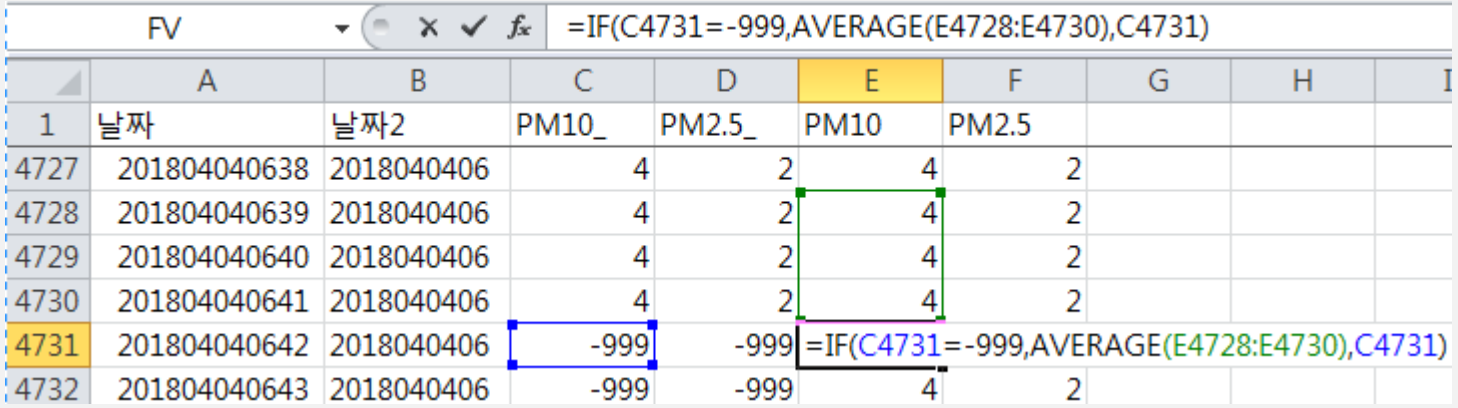


2. 데이터 전처리 (계속)

4) Missing value 처리

- 날　　짜 = 다른 위치 데이터 자료값으로 대체
  단, 8/31은 모든 데이터에 미존재

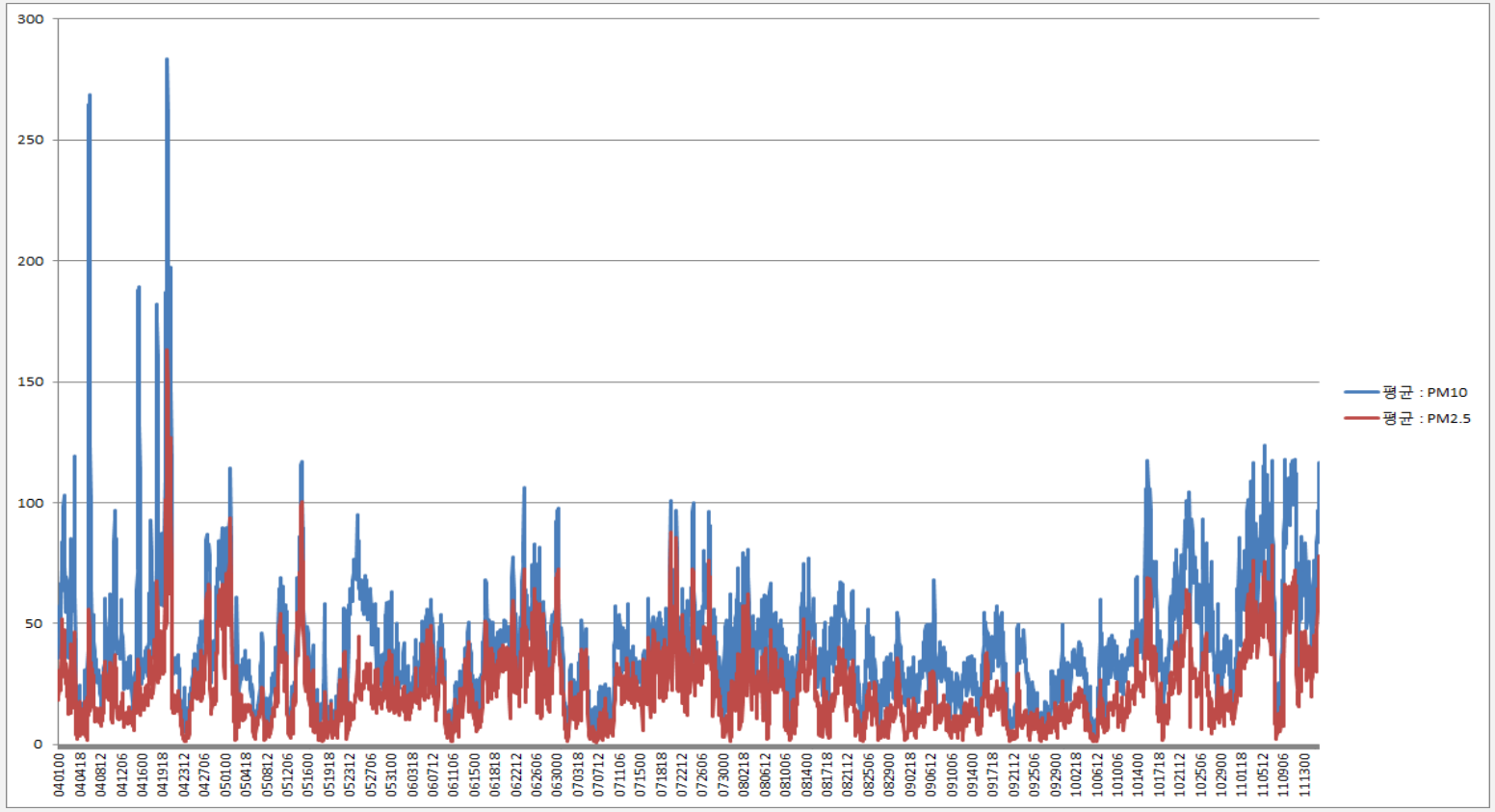
- 미세먼지 = (-999)값 → 이전 3시간 평균으로 대체



	A	B	C	D	E	F	G	H	I
1	날짜	날짜2	PM10_	PM2.5_	PM10	PM2.5			
4727	201804040638	2018040406	4	2	4	2			
4728	201804040639	2018040406	4	2	4	2			
4729	201804040640	2018040406	4	2	4	2			
4730	201804040641	2018040406	4	2	4	2			
4731	201804040642	2018040406	-999	-999	=IF(C4731=-999,AVERAGE(E4728:E4730),C4731)				
4732	201804040643	2018040406	-999	-999	4	2			

2. 데이터 전처리 (계속)

5) '선릉역' 데이터로 train set, 나머지 데이터로 test set 생성



선릉역 주변 측정기(57)에서 측정한 미세먼지 그래프

3. 분석 모델링

- n_steps 최적값으로 '5' 사용
  (n_steps가 낮으면 정확도가 감소하고, 높으면 효율성이 감소하는 경향이 있음)
- 5시간 과거데이터로 1시간 미래의 미세먼지를 예측

```
n_steps = 5
n_features = 1
```

- 데이터를 5시간 단위로 reshape

```
def split_sequence(sequence, n_step):
    X, y = list(), list()
    for i in range(len(sequence)):
        # find the end of this pattern
        end_ix = i + n_steps
        # check if we are beyond the sequence
        if end_ix > len(sequence)-1:
            break
        # gather input and output parts of the pattern
        seq_x, seq_y = sequence[i:end_ix], sequence[end_ix]
        X.append(seq_x)
        y.append(seq_y)
    return array(X), array(y)
```

3. 분석 모델링 (계속)

- 데이터를 5시간 단위로 reshape (계속)

```
x_pm10, y_pm10 = split_sequence(dataset1, n_steps)

print(x_pm10.shape, y_pm10.shape)
```

```
(5467, 5) (5467,)
```

```
x_pm25, y_pm25 = split_sequence(dataset2, n_steps)

print(x_pm25.shape, y_pm25.shape)
```

```
(5467, 5) (5467,)
```

```
x_pm10 = x_pm10.reshape((x_pm10.shape[0], x_pm10.shape[1], n_features))
x_pm25 = x_pm25.reshape((x_pm25.shape[0], x_pm25.shape[1], n_features))
```

- **시계열 분석을 위해 LSTM 활용**

```
model = Sequential()
model.add(LSTM(340, activation='relu', return_sequences=True, input_shape=(n_steps,n_features)))
model.add(LSTM(340, activation='relu'))
model.add(Dense(1))
model.compile(optimizer='adam', loss='mse')
```

3. 분석 모델링 (계속)

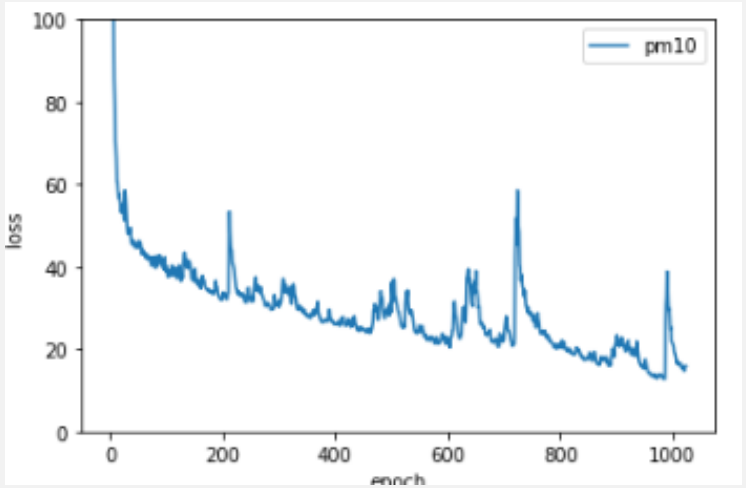
- **적합한 node 수 탐색**
  과적합을 막는 노드 수의 상한선을 계산하는 공식

$$N_h = \frac{N_s}{(\alpha * (N_i + N_o))}$$

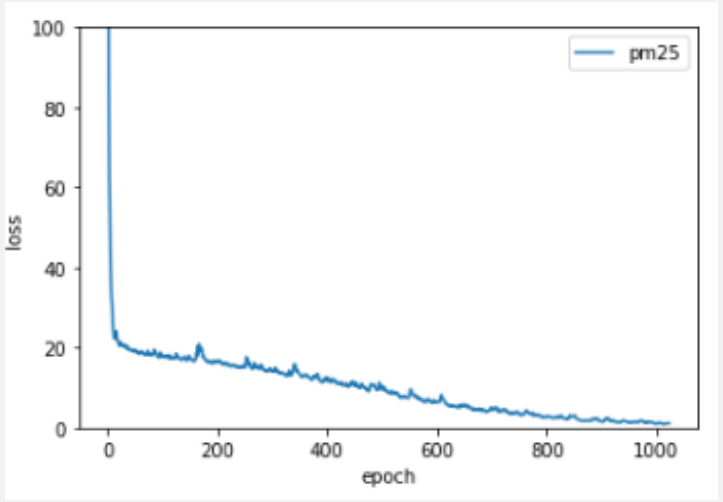
N_i = input nodes. N_o = output nodes. N_s = sample 수. α = 2부터 10까지 임의의 수
-> 최적 node 수를 91~455 범위안에서 heuristic하게 추정

- **모델링**

```
hist = model.fit(x_pm10, y_pm10, epochs=2**10, batch_size=2**8)
```



미세먼지 모델 학습과정

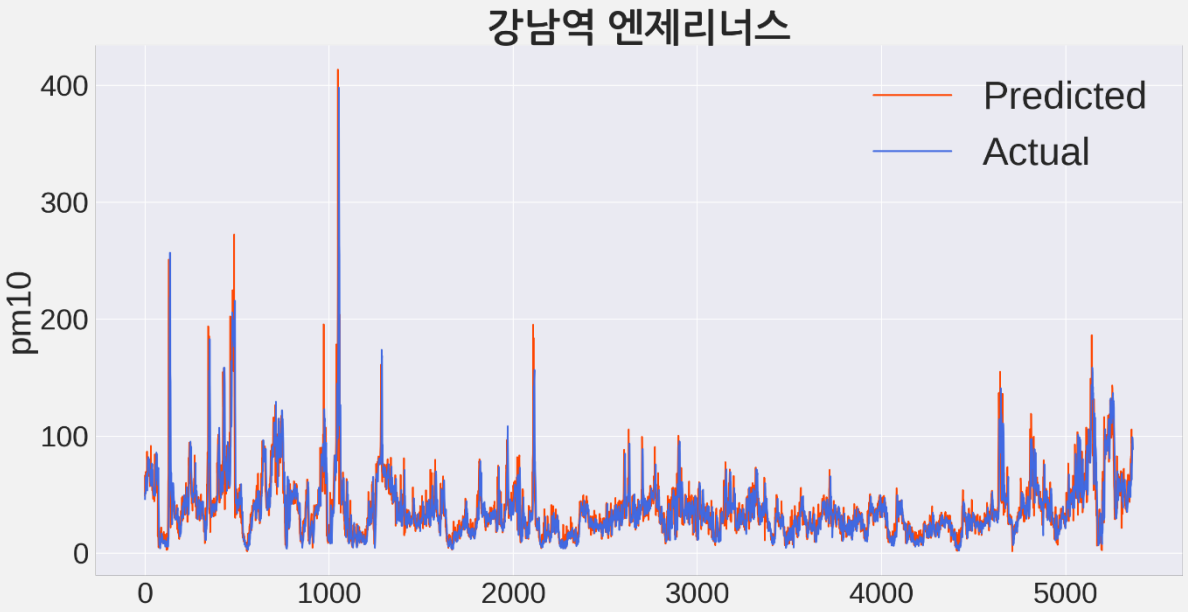


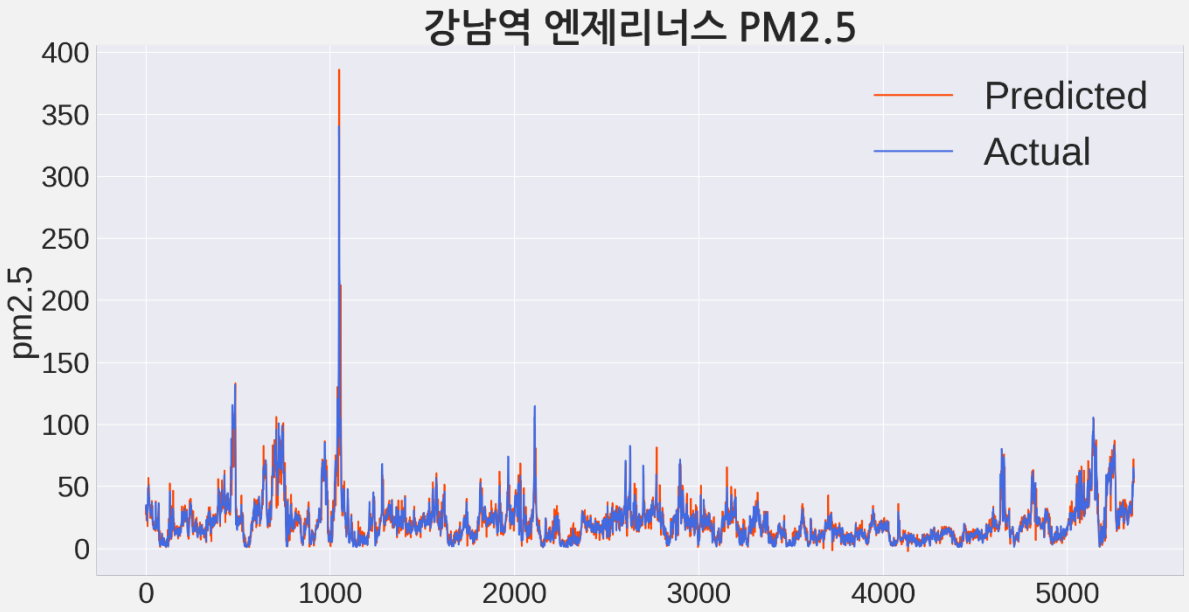
초미세먼지 모델 학습과정

4. 분석 결과 및 평가

- ‘강남역 엔제리너스’ 데이터 예측 결과

단위 : μg/㎥





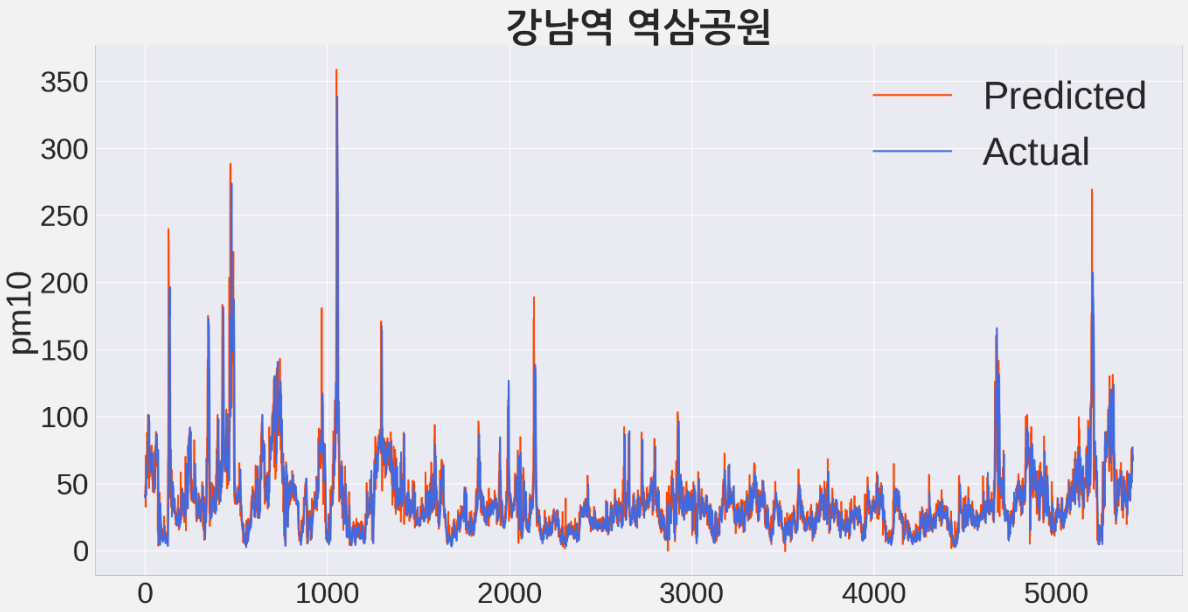
PM10 평균 오차(RMSE) = 9.17

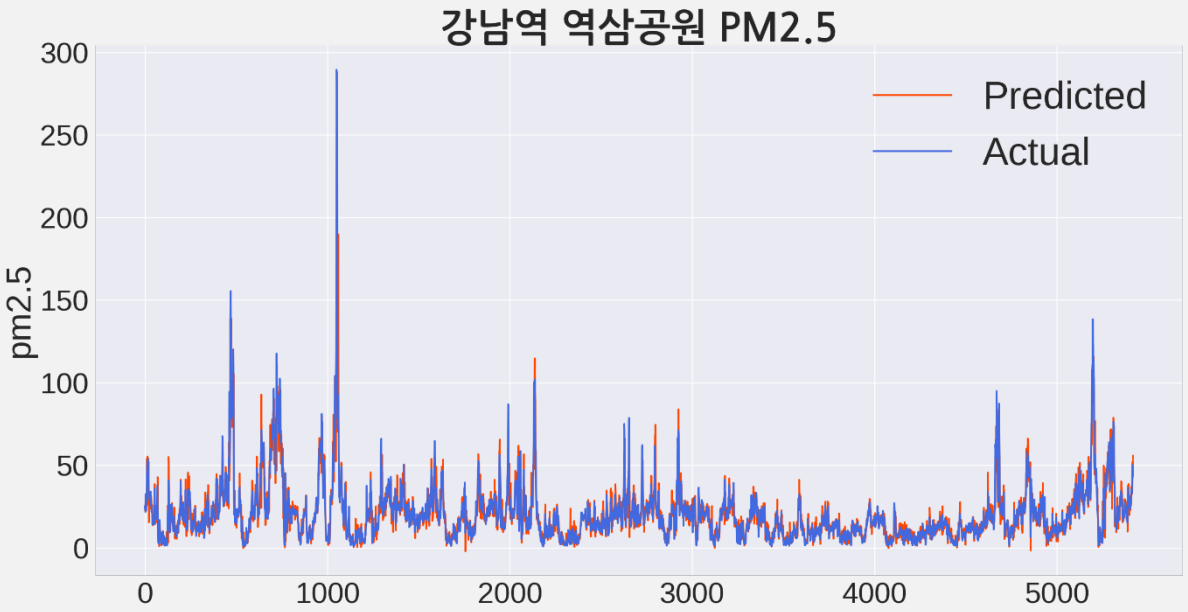
PM2.5 평균 오차(RMSE) = 6.70

4. 분석 결과 및 평가 (계속)

- **‘강남역 역삼공원’ 데이터 예측 결과**

단위 : ㎍/㎥





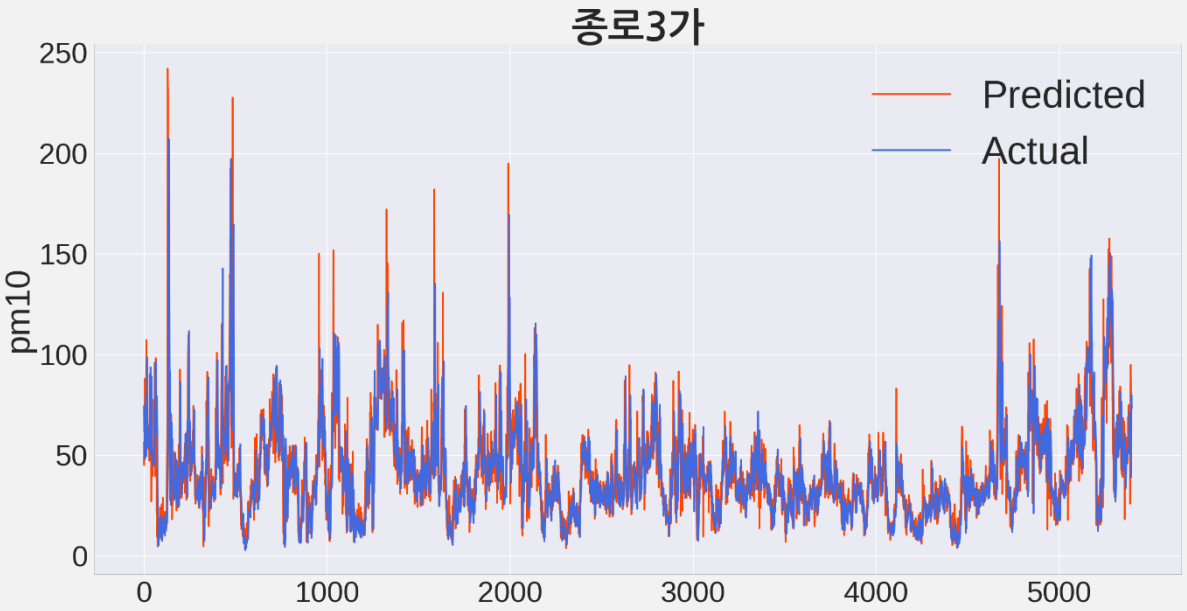
PM10 평균 오차(RMSE) = 9.08

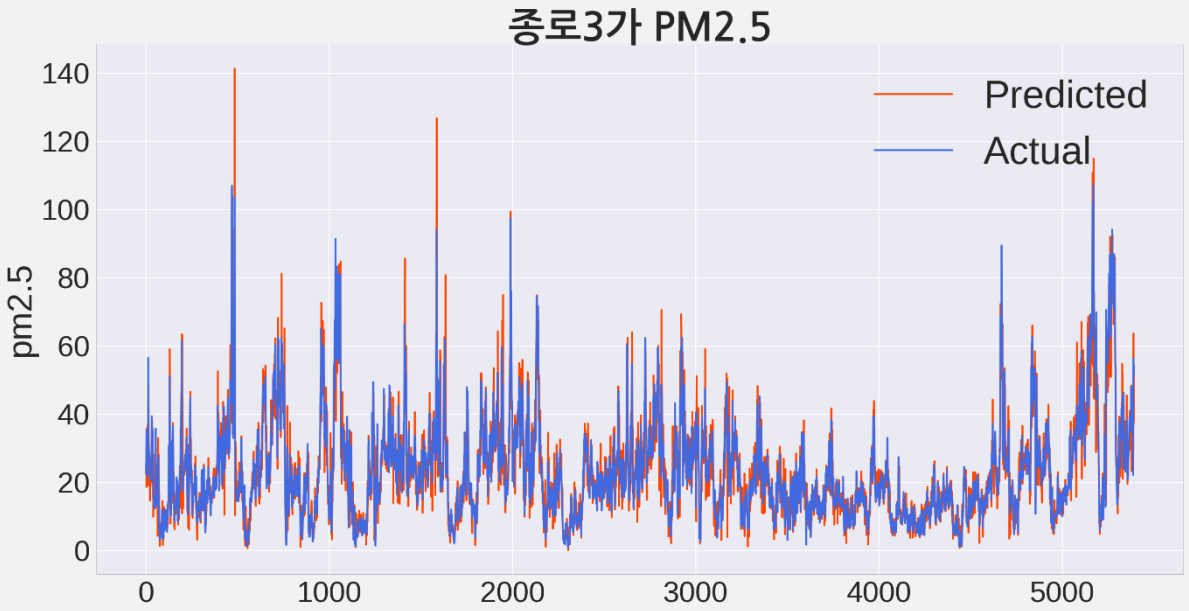
PM2.5 평균 오차(RMSE) = 6.30

4. 분석 결과 및 평가 (계속)

- **'종로3가' 데이터 예측 결과**

단위 : ㎍/㎥





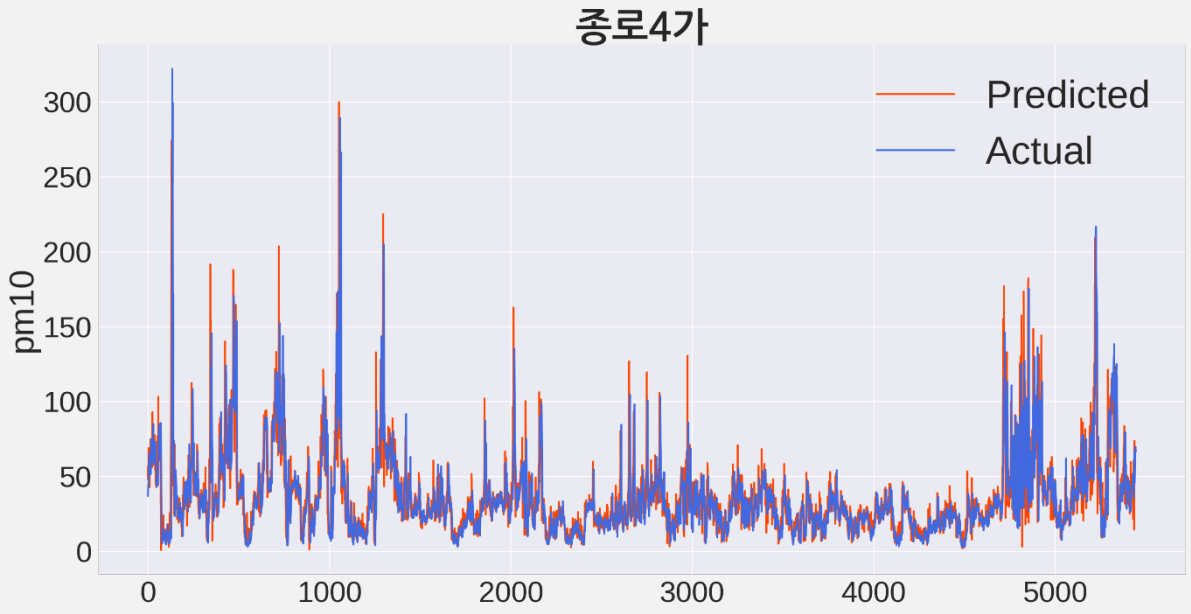
PM10 평균 오차(RMSE) = 8.65

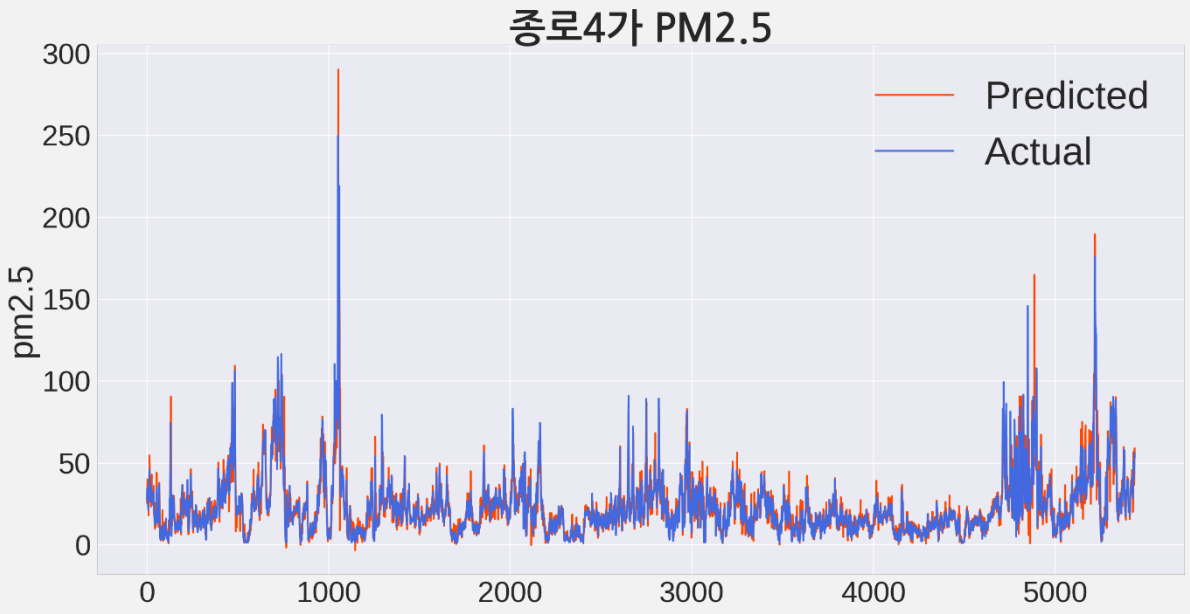
PM2.5 평균 오차(RMSE) = 6.06

4. 분석 결과 및 평가 (계속)

- **‘종로4가’ 데이터 예측 결과**

단위 : ㎍/㎥





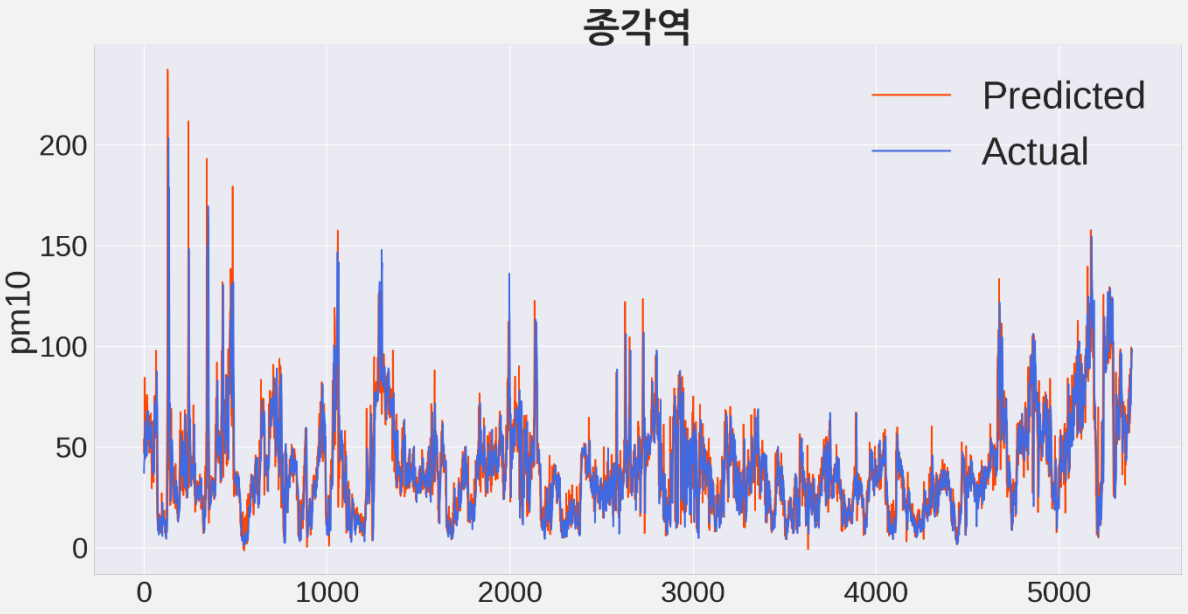
PM10 평균 오차(RMSE) = 11.10

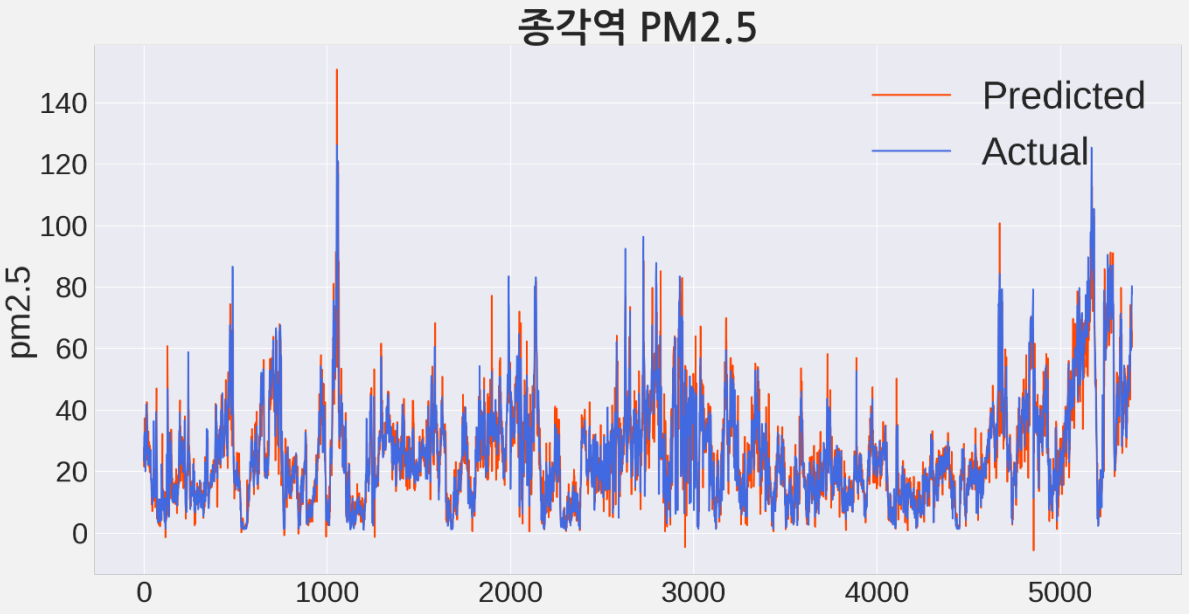
PM2.5 평균 오차(RMSE) = 8.56

4. 분석 결과 및 평가 (계속)

- **‘종각역’ 데이터 예측 결과**

단위 : ㎍/㎥





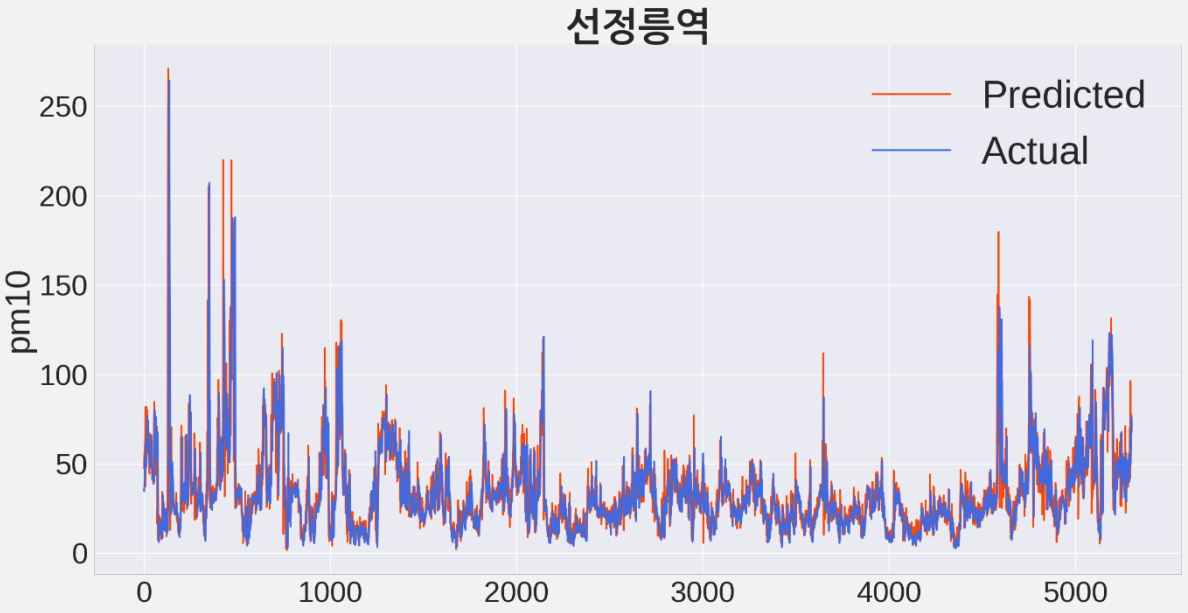
PM10 평균 오차(RMSE) = 8.53

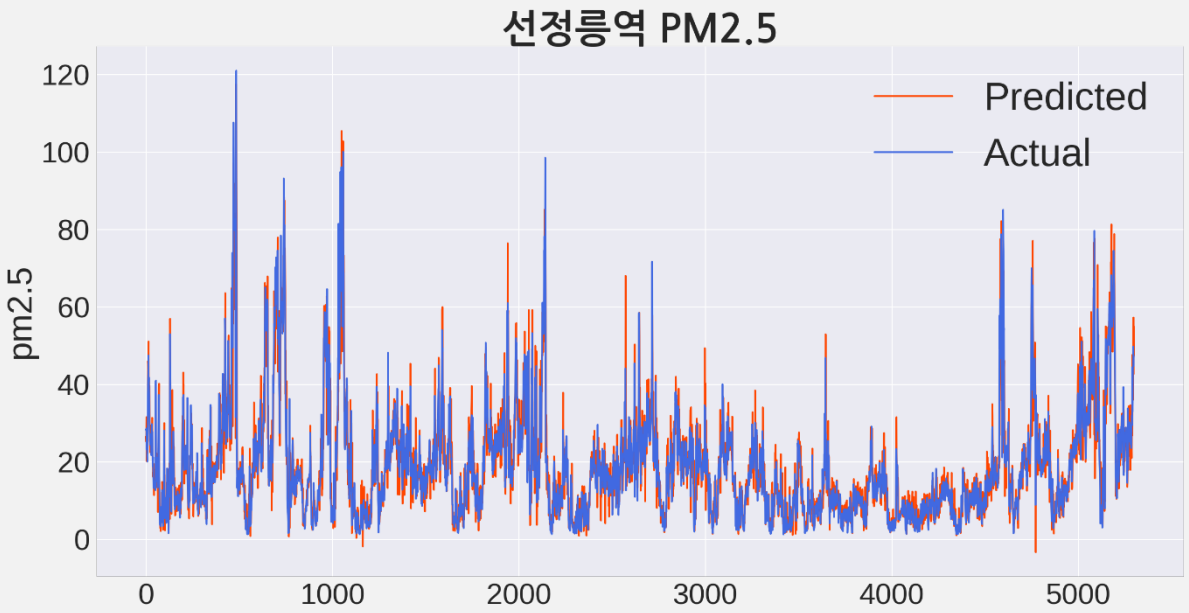
PM2.5 평균 오차(RMSE) = 6.90

4. 분석 결과 및 평가 (계속)

- **‘선정릉역’ 데이터 예측 결과**

단위 : μg/㎥





PM10 평균 오차(RMSE) = 8.42

PM2.5 평균 오차(RMSE) = 5.31

4. 분석 결과 및 평가 (계속)

- **결과 및 평가**

  각 모델 학습시간: i7 CPU 기준 40분

  실제보다 과하게 예측된 경향이 있음

  데이터 양 및 학습시간 대비 우수한 성능의 Model 도출

- **한계점**

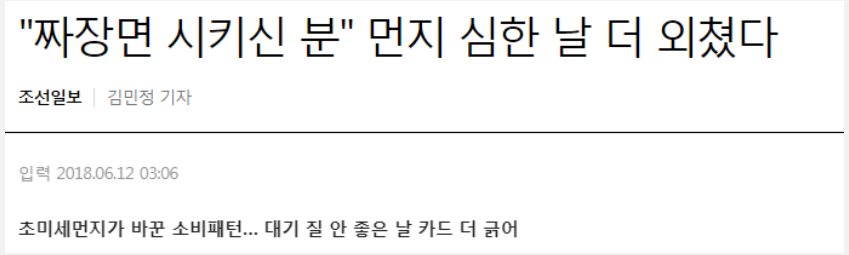
  2018년 이전 데이터 확보 시 성능 개선 가능

  겨울 데이터 확보 시 모델에 계절적 주기성을 보완하여 좀 더 일반화된 모델을 만들 수 있음

  PM 2.5모델의 경우, 학습 시간 추가 확보 시 오차를 감소시킬 수 있음

5. 사업성 분석

- **시간단위 예측으로 Real-Time 의사결정 지원**
  - **살수차 운용 등 정부 미세먼지 관리 대책 수립 및 이행에 시간 및 경제적 비용 감소**
  - **요식업 등 미세먼지에 영향을 받는 산업에 정보를 제공하여, 소비 패턴을 예상하고 대응을 가능하게 함**



"짜장면 시키신 분" 먼지 심한 날 더 외쳤다

조선일보 | 김민정 기자

입력 2018.06.12 03:06

초미세먼지가 바꾼 소비패턴... 대기 질 안 좋은 날 카드 더 긁어

  - **가정, 유치원, 학교 등에 정보를 제공하여 적절한 순간에 공기청정기를 틀거나 외부활동 등의 의사결정을 지원함으로써 에너지 절감 및 헬스 케어에 도움**

유치원에 설치된 `에어가드K 실외공기측정기(OAQ)`는 야외 놀이터의 미세먼지, 휘발성유기화합물(VOCs) 등의 공기질 상태를 정확히 측정해 IoT 기술로 24시간 측정한다. 또 교실 곳곳과 현관에 설치된 실내공기측정기(IAQ)로는 유치원 실내 공기도 감지한다.

- **기존 미세먼지 예보는 필요한 데이터 종류/수가 많고 계산의 복잡성이 높은 단점이 있으나,**

  딥러닝 기반 예보를 활용하면 과거 미세먼지 데이터만을 가지고 저비용 고효율 예측 시스템을 구축할 수 있다

출처

- 며칠째 미세먼지 예보 아닌 '중계'만…"믿고 나갈 수가 없다"
  http://news.jtbc.joins.com/article/article.aspx?news_id=NB11735698

- 황사와 미세먼지 무엇이 다르냐면요…
  http://www.hani.co.kr/arti/society/society_general/733554.html

- 미세먼지 예보는 어떻게 측정할까? 미세먼지 측정 방법과 기상과의 관계
  https://blog.naver.com/mesns/221384758930
  (환경부 공식 블로그)

- 국가 대기질 예보 정확도 향상 연구 (I) 2014, 기후대기연구부 대기질통합예보센터, 국립환경과학원
  발간등록번호 11-1480523-002196-01

- "짜장면 시키신 분" 먼지 심한 날 더 외쳤다
  http://biz.chosun.com/site/data/html_dir/2018/06/12/2018061200016.html

- 케이웨더 "유치원 에어가드K 설치로 야외활동 결정하세요“
  http://www.etnews.com/20160614000111