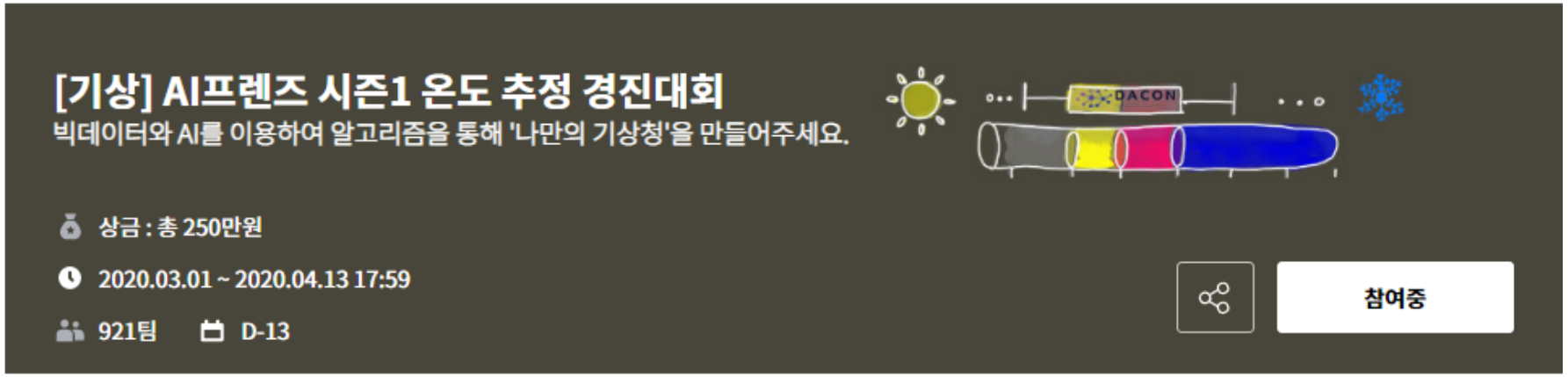
기상청 공공데이터를 활용한
온도 추정 모델 구축

팀원

주재우 안희재 이동준 이윤정 이예지



사이트

DACON 주관 대회

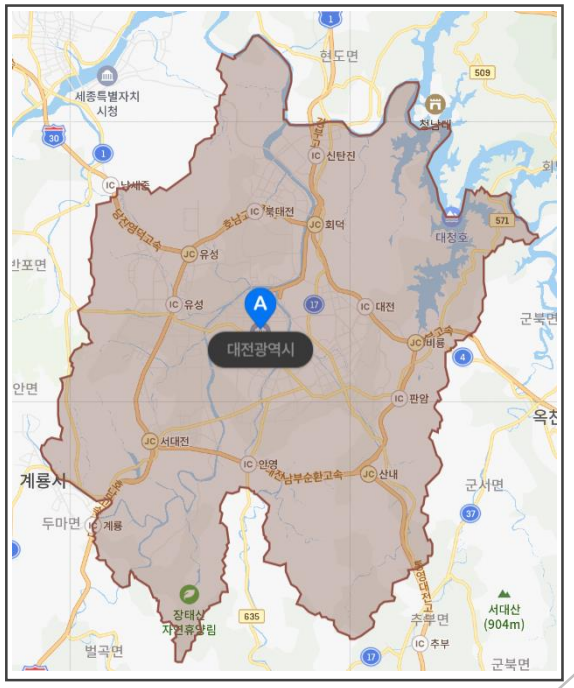
https://dacon.io/competitions/official/235584/overview/

목표

배웠던 RNN 모델을 활용

실력 파악

데이터 특징



대전의 19개의 실내 외 **기온 센서 데이터**로 측정
8개 카테고리의 비 식별화 **기상청 공공데이터** 제공
데이터는 시간 순으로 10분 단위

외부 데이터나 미래 데이터(T시점 이후) 사용 불사

30일치 Y00~Y17 온도 데이터 + 기상청 공공 데이터

↓

3일치 Y18 온도 데이터 + 기상청 공공 데이터

↓

80일치 기상청 공공 데이터 제공

↓

80일치 Y18 온도 예측

기온 (°C, 섭씨)	풍속(ms/s)
습도 (%, 상대 습도)	누적 일사량(MJ/m²)
누적 강수량(mm)	해면 기압(hPa)
풍향 (°, degree)	현지 기압(hPa)

일정

2020.03.13 ~ 2020.03.17 : 데이터 특징과 상관관계 분석

2020.03.18 ~ 2020.03.20 : 모델 구성

2020.03.21 ~ 2020.03.30 : 모델 fine tuning, 데이터 전처리

2020.03.31 ~ : optimizer와 데이터 전처리 가공 방법 도모

환경

IDE : Pycharm, Rstudio, Anaconda, Jupyter Notebook

OS : Windows 10

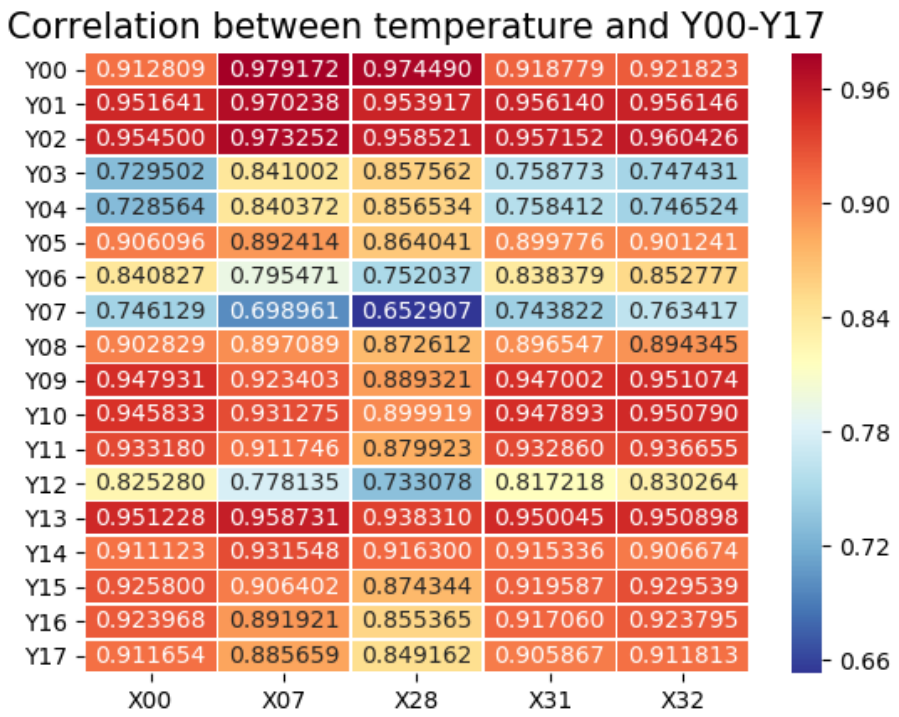
Language : Python 3.7.4, R 3.6.3

Framework : TensorFlow 2.1 GPU, Keras

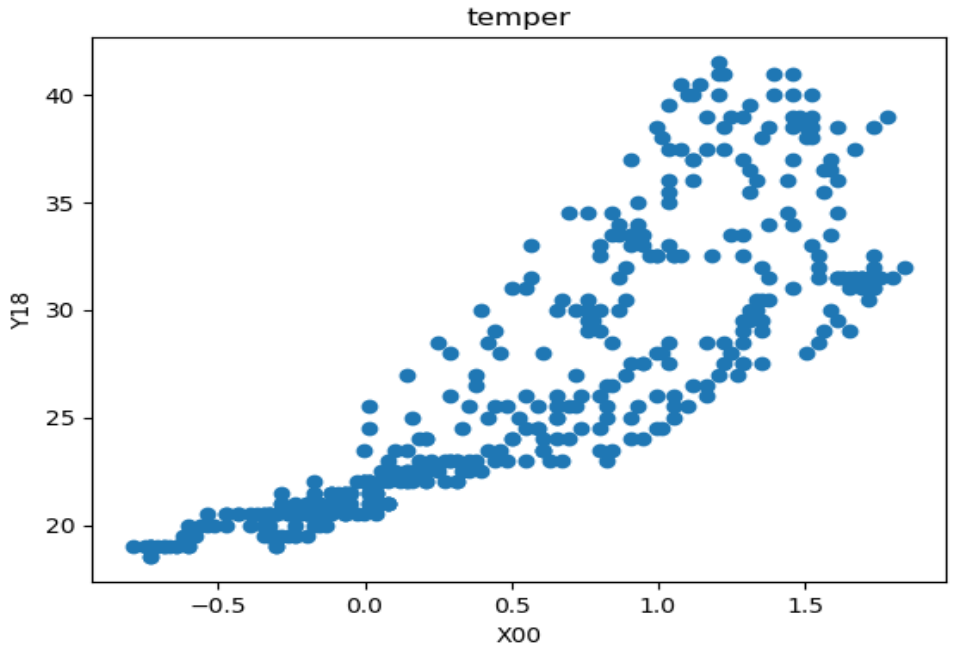
팀원 : 5명

기온

온도와 높은 양의 상관관계를 띄움



피어슨 상관관계 상으로 높은 연관성



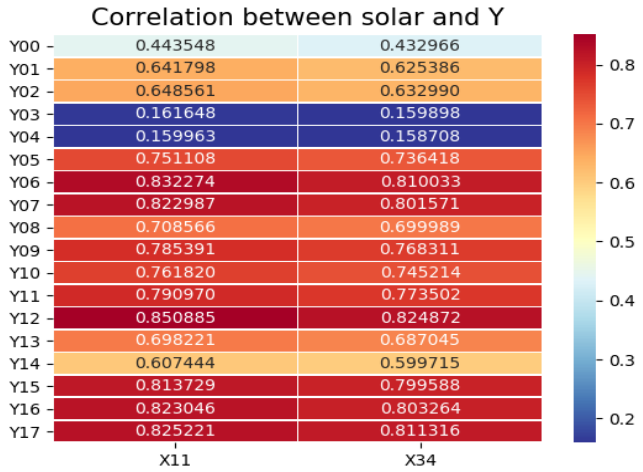
선형관계를 보임

누적 일사량

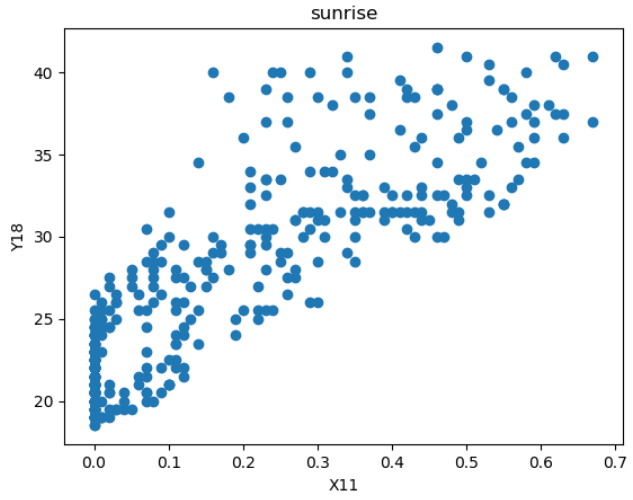
이상치를 제외하고 전처리한 데이터는 온도와 높은 양의 상관관계를 띄움

X11	X14	X16	X19	X34
Min. : 0.00	Min. :0	Min. :0	Min. :0	Min. : 0.00
1st Qu.: 0.02	1st Qu.:0	1st Qu.:0	1st Qu.:0	1st Qu.: 0.03
Median : 7.69	Median :0	Median :0	Median :0	Median : 8.21
Mean :11.01	Mean :0	Mean :0	Mean :0	Mean :11.57
3rd Qu.:22.19	3rd Qu.:0	3rd Qu.:0	3rd Qu.:0	3rd Qu.:22.56
Max. :30.70	Max. :0	Max. :0	Max. :0	Max. :32.24

X14, X16, X19 데이터가 누락되어 있음. **이상치 데이터**로 판단

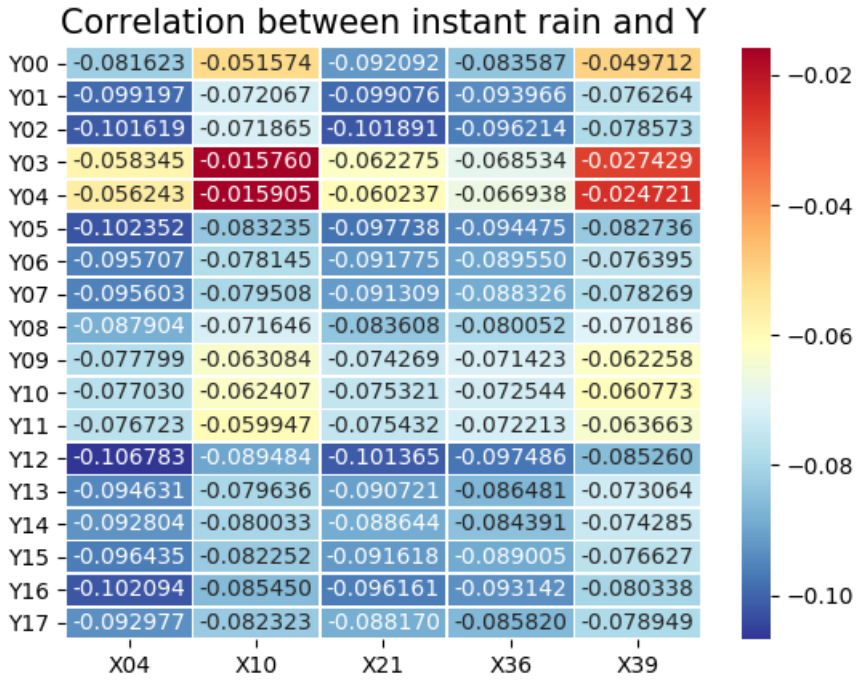


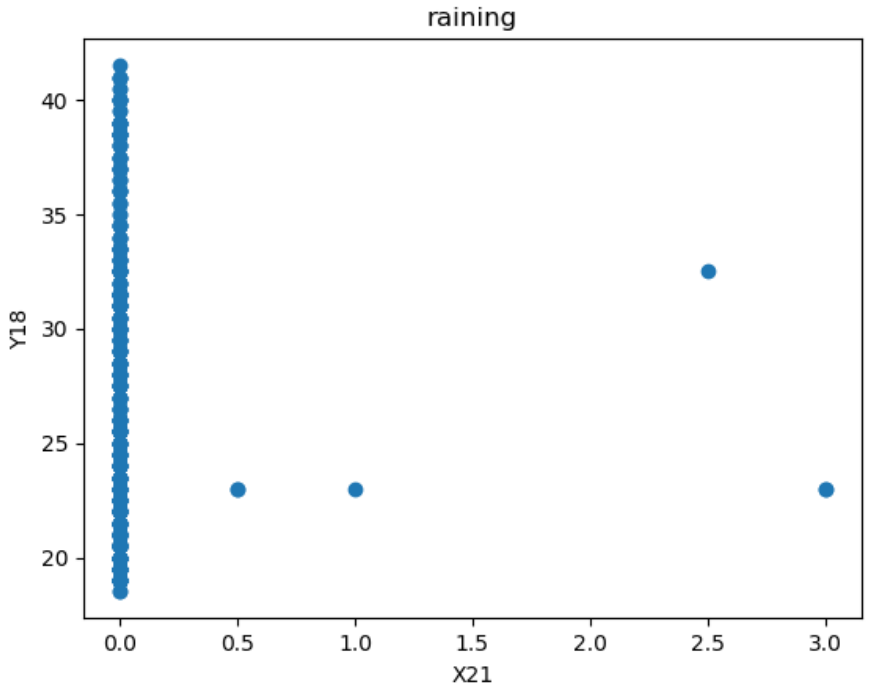
하루치로 등록된 데이터를 **차분**해 10분 단위 데이터로 변경



누적 강수량

온도와 낮은 상관관계를 띄움



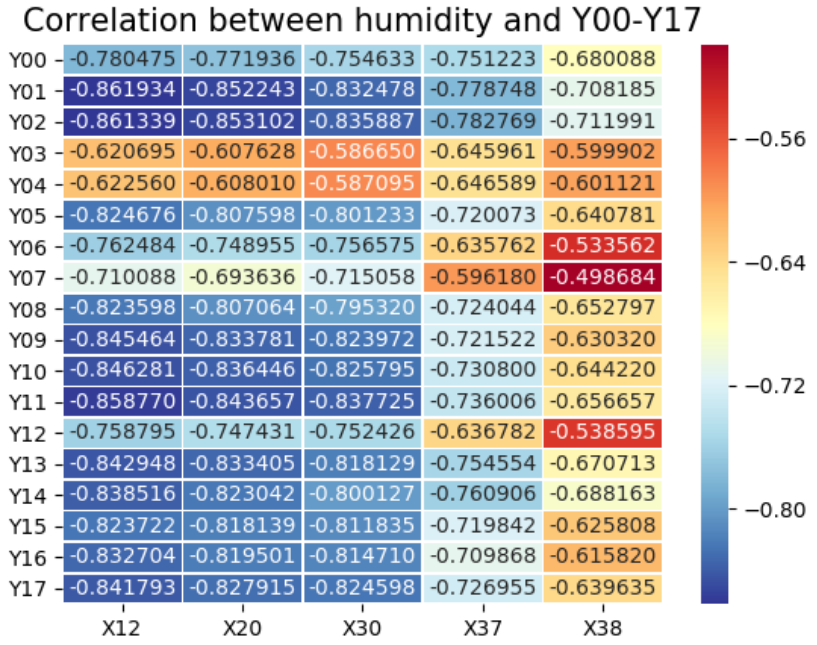


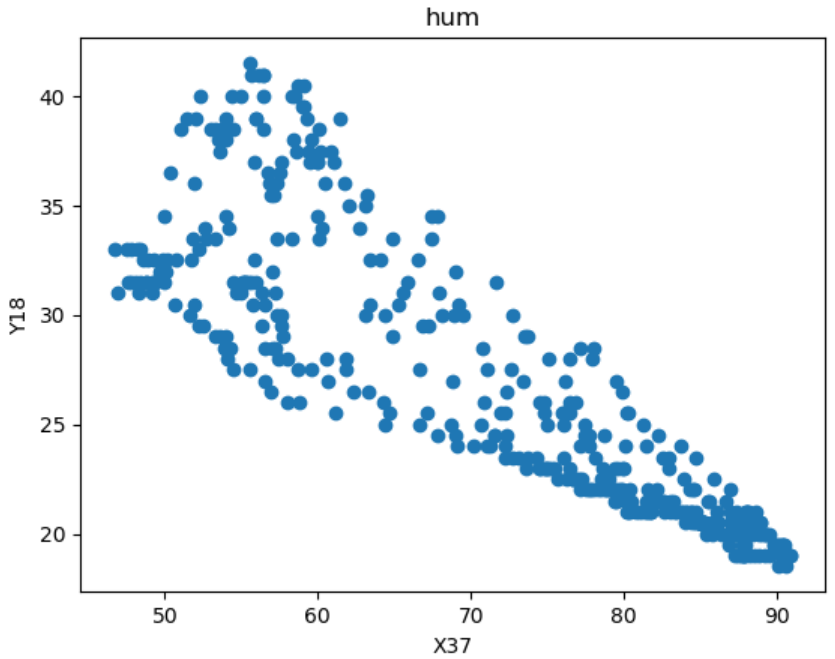
습도

온도와 높은 음의 상관관계를 띄움

1~100%를 가지는 단위라 1/100을 적용해봤지만 결과가 같음

데이터 그대로 사용



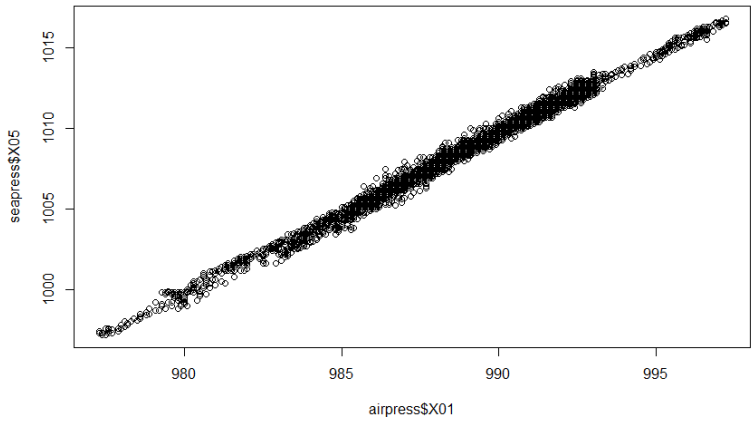


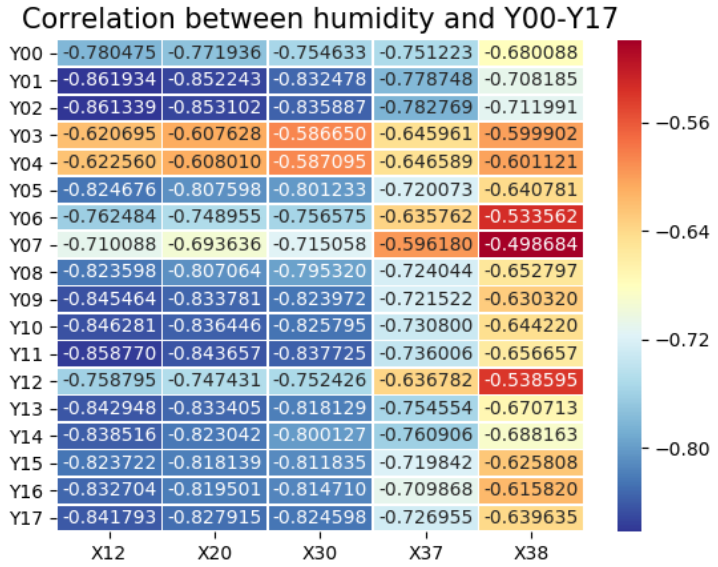
기압

온도와 낮은 상관관계를 띄움

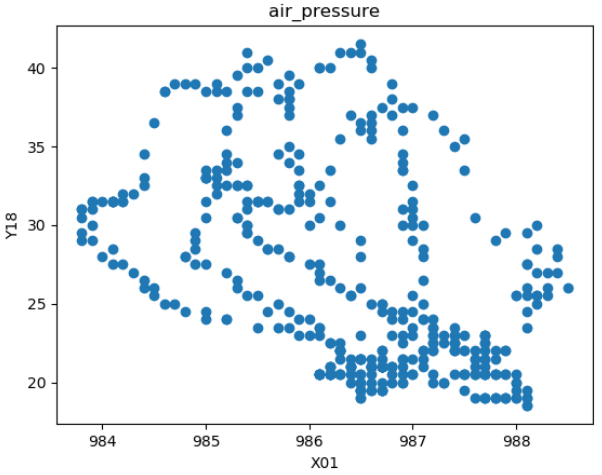
현지 기압 기반으로 계산됨

다중 공선성 고려






```
> summary(airpress)
      X01             X06             X22             X27             X29
 Min.   :977.3   Min.   :978.1   Min.   : 990.8   Min.   : 997.2   Min.   : 989.7
 1st Qu.:986.8   1st Qu.:987.1   1st Qu.: 999.9   1st Qu.:1005.5   1st Qu.: 998.9
 Median :988.6   Median :989.1   Median :1001.9   Median :1007.7   Median :1000.9
 Mean   :988.6   Mean   :989.0   Mean   :1001.8   Mean   :1007.6   Mean   :1000.8
 3rd Qu.:991.0   3rd Qu.:991.6   3rd Qu.:1004.2   3rd Qu.:1010.2   3rd Qu.:1003.4
 Max.   :997.2   Max.   :997.4   Max.   :1010.1   Max.   :1015.4   Max.   :1009.0
> summary(seapress)
      X05             X08             X09             X23             X33
 Min.   : 997.2   Min.   : 997.3   Min.   : 996.3   Min.   : 997.5   Min.   : 998.7
 1st Qu.:1006.5   1st Qu.:1006.7   1st Qu.:1006.3   1st Qu.:1006.9   1st Qu.:1007.1
 Median :1008.5   Median :1008.8   Median :1008.4   Median :1008.9   Median :1009.2
 Mean   :1008.3   Mean   :1008.8   Mean   :1008.4   Mean   :1008.9   Mean   :1009.1
 3rd Qu.:1010.8   3rd Qu.:1011.4   3rd Qu.:1010.9   3rd Qu.:1011.4   3rd Qu.:1011.7
 Max.   :1016.8   Max.   :1017.5   Max.   :1018.0   Max.   :1017.1   Max.   :1016.8
```



풍속

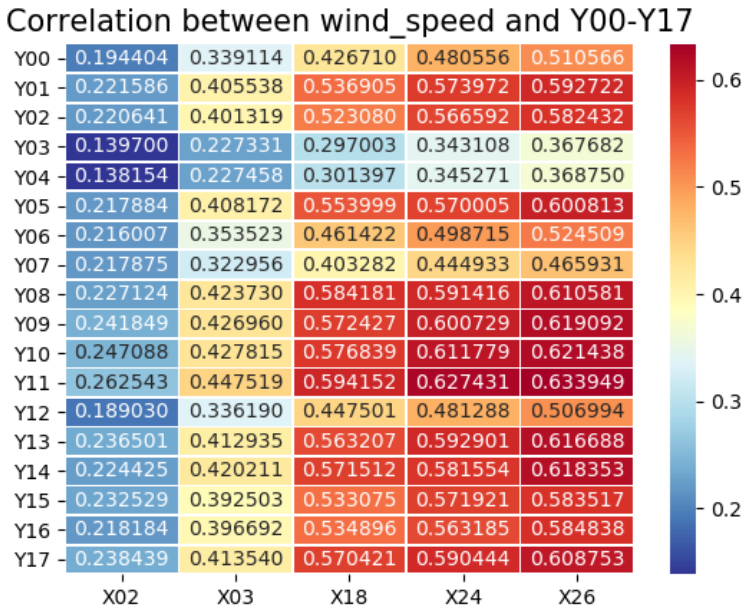
온도와 높은 상관관계를 띄움

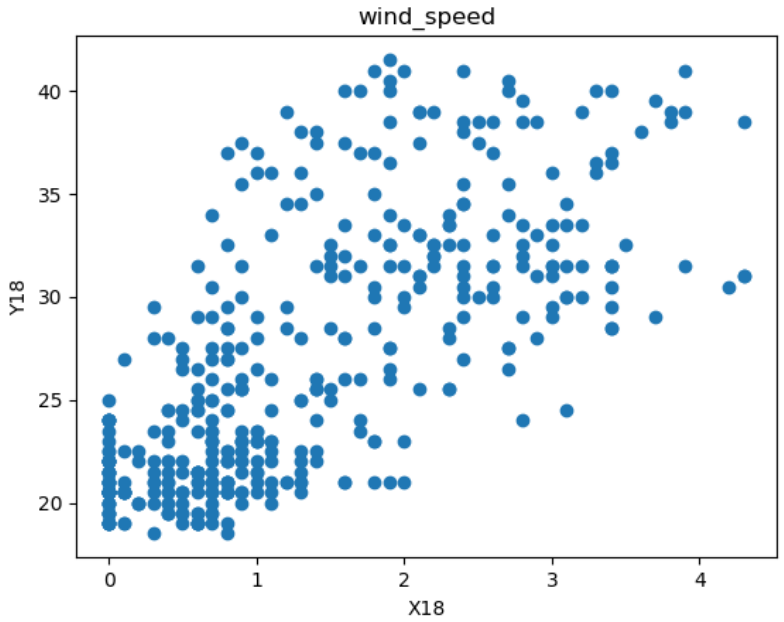
센서 위치에 따라 다른 풍속을 보여줌

전체적으로 높은 상관관계를 보임

분석에 사용하기에는 충분치 않다 판단

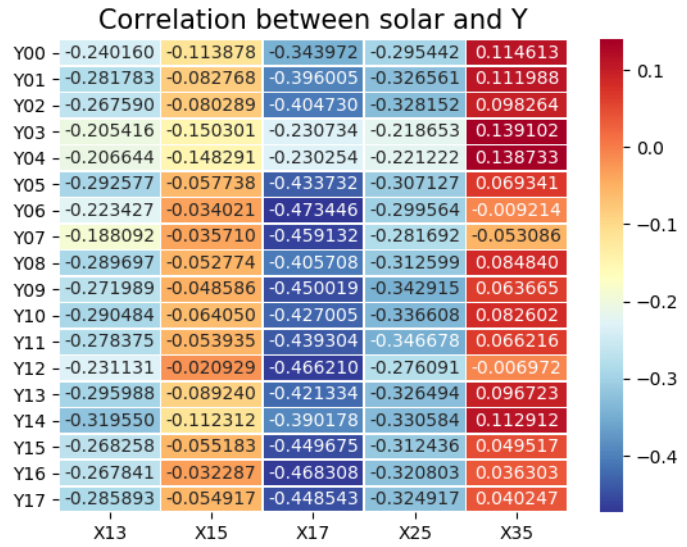
추후 전처리 방법을 알아낸 뒤 적용





풍향

일부 데이터셋은 전치리 후 온도와 높은 상관관계를 띄움

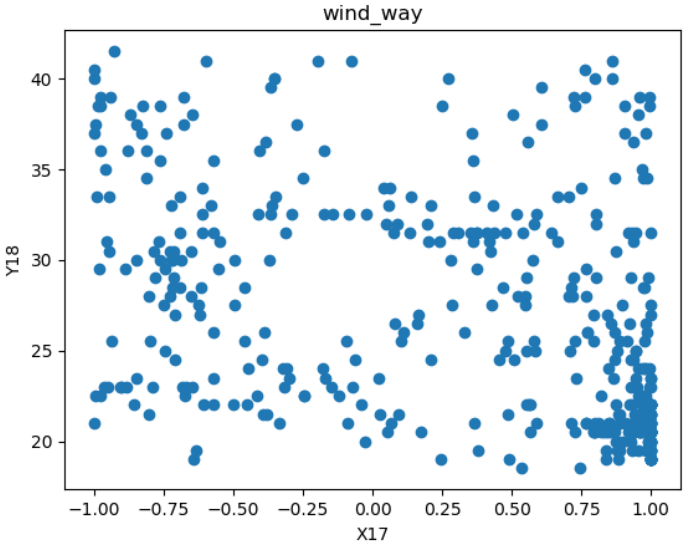


0~360도로 측정된 데이터

Cos함수를 적용해 전처리함

상관계수가 낮은 X15와 X35 제거

```
      X13               X15              X17              X25              X35
 Min.   :  0.00   Min.   :  0.0   Min.   :  0.00   Min.   :  0.0   Min.   :  0.0
 1st Qu.: 69.22   1st Qu.:115.6   1st Qu.: 33.88   1st Qu.:  0.0   1st Qu.:155.3
 Median :182.60   Median :170.2   Median :190.45   Median :134.2   Median :240.8
 Mean   :161.36   Mean   :174.3   Mean   :161.82   Mean   :139.5   Mean   :208.4
 3rd Qu.:228.93   3rd Qu.:258.4   3rd Qu.:253.12   3rd Qu.:270.9   3rd Qu.:277.8
 Max.   :360.00   Max.   :359.9   Max.   :360.00   Max.   :359.9   Max.   :359.9
> summary(cos_dir)
      X13               X15               X17               X25               X35
 Min.   :-1.00000   Min.   :-1.0000   Min.   :-1.0000   Min.   :-1.0000   Min.   :-1.0000
 1st Qu.:-0.64151   1st Qu.:-0.6611   1st Qu.:-0.5529   1st Qu.:-0.2272   1st Qu.:-0.6905
 Median : 0.14898   Median : 0.1332   Median : 0.3607   Median : 0.8504   Median : 0.1104
 Mean   : 0.09048   Mean   : 0.0768   Mean   : 0.1880   Mean   : 0.4033   Mean   : 0.0585
 3rd Qu.: 0.85507   3rd Qu.: 0.8218   3rd Qu.: 0.9634   3rd Qu.: 1.0000   3rd Qu.: 0.7996
 Max.   : 1.00000   Max.   : 1.0000   Max.   : 1.0000   Max.   : 1.0000   Max.   : 1.0000
```

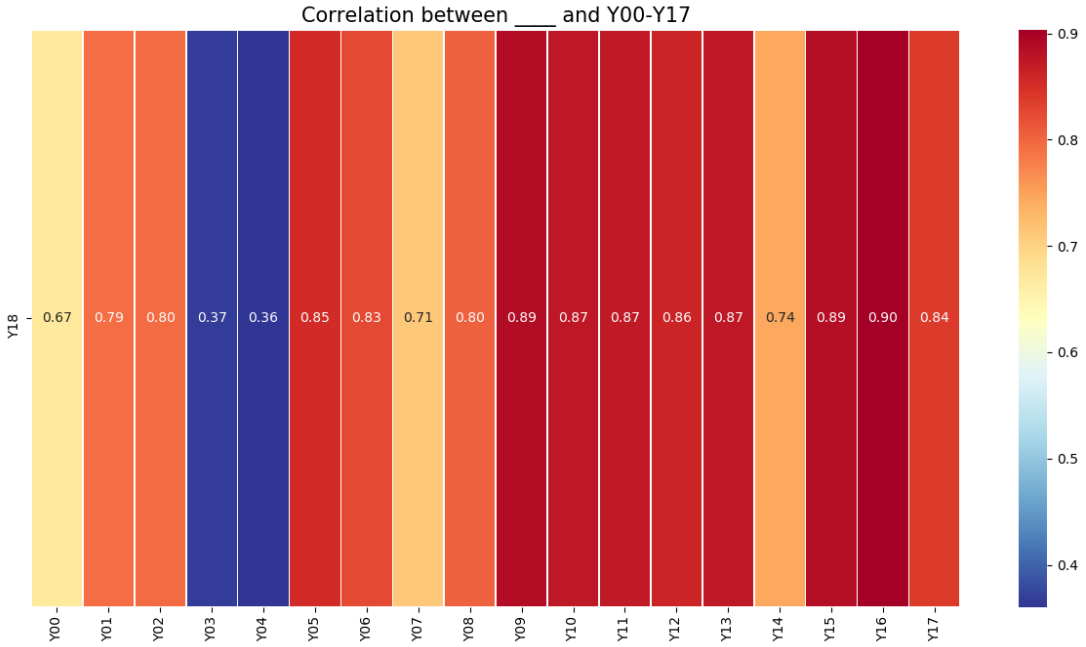


사용할 데이터 셋 선택

Y 데이터 셋

전이 학습을 위해 Y18과 유사한 Y값을 선별하기로 함

상관관계가 높은 Y09, Y15, Y16을 사용



X 데이터 셋

사용

기온, 일사량, 습도

배제

해면 기압, 현지 기압, 강수량

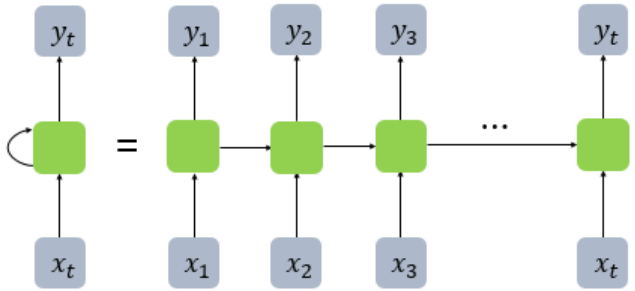
보류

풍향(일부만 사용), 풍속

RNN

RNN에서 개선된 LSTM 신경망을 구성

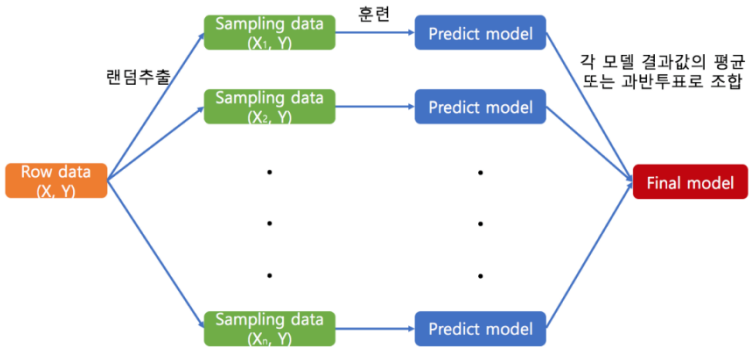
과 적합(Overfitting)과 과 훈련(Overtraining)에 예민



앙상블 모델

여러 개의 모델을 학습시켜 그 예측 값을 종합하여 성능을 개선

과 적합(Overfitting)을 방지하는 알고리즘



전이 학습

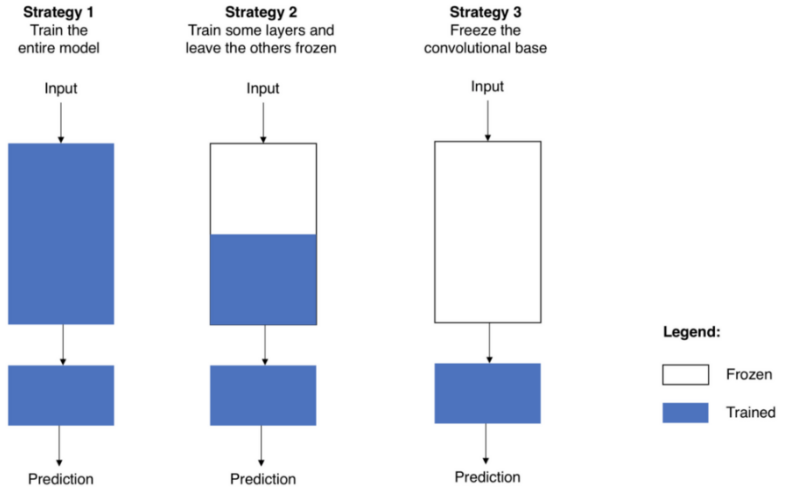
30일 : Y00 ~ Y17 존재 , Y18 누락

3일 : Y00 ~ Y17 누락, Y18 존재

일부 층을 고정시키고 재 학습 시키는 방법을 채택

재 학습 시 epoch는 최대한 적게 설정

이전 학습의 영향을 받을 수 있도록 설계



RNN

simple_lstm_model1
simple_lstm_model2
LSTM와 Dense 레이어로 구성

항등함수로 **선형 회귀** 사용

Xavier 초기값 사용
He보다 나은 결과 산출

앙상블 모델

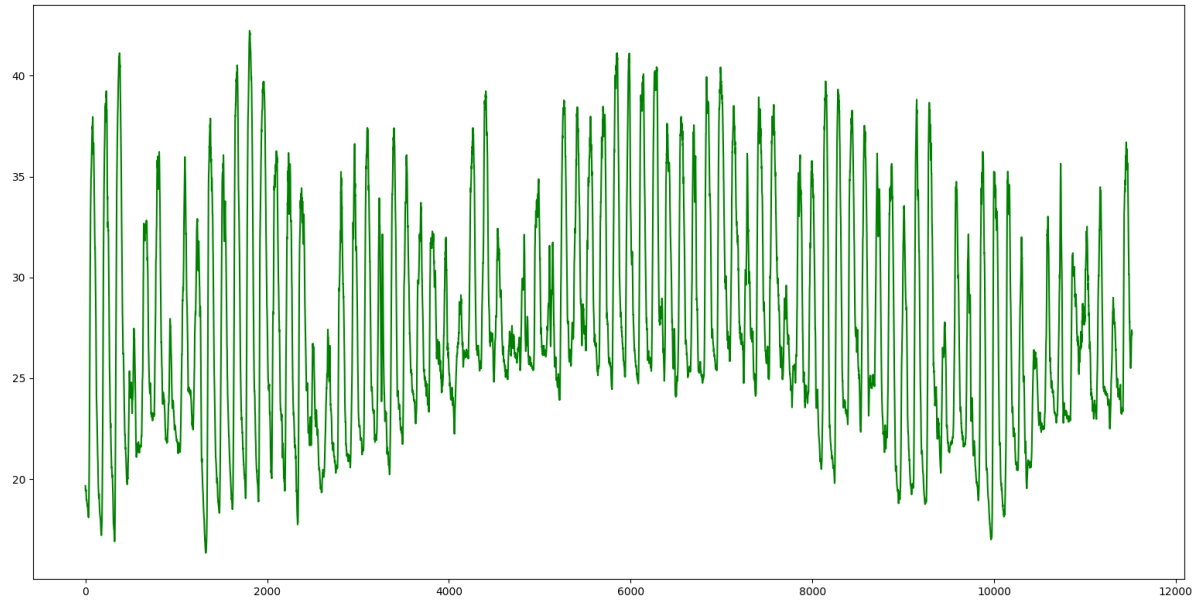
모델 개수 **2개**로 결정

모델 하나에
simple_lstm_model1과
simple_lstm_model2
append

전이 학습

LSTM 레이어만 고정해
전이 학습 실행

Optimizer로 **adamax** 사용
Adam보다 결과가 안정적
으로 나옴



기상청 공공 데이터를 이용해 특정 위치의 온도 예측 모델 생성

적절한 데이터 전처리 수행이 가장 효과적

앙상블 모델 적용을 통한 성능 개선

더 효과적인 **전처리 방법** 모색

Ex) 풍향, 습도

다른 **앙상블 모델** 적용 시도

Ex) 모델 개수 조정, hyper parameter

다른 **알고리즘/optimizer** 시도

Ex) Layer/Batch Normalization,

Nadam/Adaboost/AMS grad 등