동서발전 태양광 발전량 예측 AI 경진대회

Data Analysis Assistant

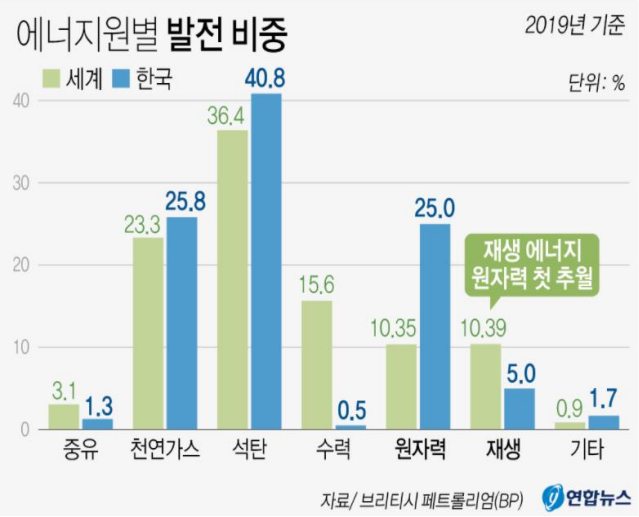
통계학과 ASMR

201611501 강경준
201611510 박민규
201611523 안성빈
201811526 이은주

Data Analysis Report

Contents

Project Purpose



2019년 역사상 처음으로 전 세계 재생에너지의 발전량이 원자력 발전량을 넘어섰다.[1] 우리나라 역시 2030년까지 재생에너지 발전량 비중 20%를 목표로 하는 이행계획을 발표하였으며,[2] 이는 재생에너지가 나날이 중요해질 것임을 뜻한다. 특히 태양광 발전량은 2019년 기준 재생에너지 발전량에서도 67%의 비율을[3] 차지하므로 효율적인 태양광 발전 전력 수급 계획을 세우는 것은 환경적 효과에 더불어 상당한 경제적 효과를 가져올 것으로 예상이 된다.

따라서 한국동서발전(주)에서 주최하는 “동서발전 태양광 발전량 예측 AI 경진대회”에 참여하여 인공지능 기반 태양광 발전량 예측 모델을 만들어, 보다 원활한 전력 수급 계획을 가능하게 하고자 한다.

Project Summary

[동서발전 태양광 발전량 예측 AI 경진대회]

해당 공모전은 울산 태양광 발전소 1곳과 당진의 태양광 발전소 3곳(당진 태양광, 당진 수상태양광, 당진 자재창고태양광)에서의 태양광 발전량을 예측하기 위한 모델을 생성하는 공모전이다. 각 발전소별 발전용량은 500,1000,1000,700 이며, 총 발전용량은 3200이 된다.

Train Data : 2018.03~2021.01의 발전소 지역 기상예보 및 태양광 발전량 데이터
Test Data : 2021.02의 기상예보 데이터

Train Data를 통해 기상예보 데이터(설명변수)를 이용하여 태양광 발전량(반응변수)을 예측하는 모델을 생성한다. 이후 Test Data를 통해 2021.02의 태양광 발전량을 예측함으로써 NMAE-10을 기준으로 모델 성능을 평가하여 순위를 결정한다.

하지만, 우리는 프로젝트의 목적에 부합하도록 모델 성능 평가 기준을 달리 하였다.

Project Summary

1. 모델 성능 평가 측도 - RMSE를 기준으로 모델 학습

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|Y_i - \hat{Y}_i| \qquad \text{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}$$

NMAE-10 : 총 발전용량의 10% 이상 발전한 데이터만을 활용하여 계산된 MAE를 총 발전용량으로 나누어 정규화한 평가측도

(Ex. 총 발전용량이 1000으로 측정된 당진 발전소의 경우 1000의 10%인 100 이상의 발전을 한 데이터 만을 평가기준으로 설정하여 밤 시간, 비 오는 날 등 발전량이 적거나 없는 경우는 평가 대상이 되지 않는다)

RMSE : 평균 제곱 오차(MSE)의 제곱근이다.

RMSE를 평가 측도로 사용하는 이유

(1) **이상치**에 민감하다. 즉, NMAE를 측도로 하는 것 보다 이상치에 더욱 민감한 모델을 생성할 수 있다.

(2) RMSE를 최소화 하는 회귀계수는 **LSE**(Least Squares Estimators)가 된다. 이때 Gauss-Markov Assumption이 만족한다면 해당 회귀계수는 **BLUE**(Best Linear Unbiased Estimators)가 된다. 단순히 Error를 최소화 하는 것이 아닌 BLUE라는 뛰어난 추정치를 얻을 수 있으므로 RMSE를 평가기준으로 선정을 한다.

2. Train, Test Set 설정

모델링을 할 때, 모델의 성능 평가를 위해 전체 데이터를 train set, test set으로 나누는데, 통상적으로 75%:25% 비율을 기준으로 한다.

하지만 본 대회의 목적은 2018년 3월부터 2021년 1월까지의 데이터를 통해 다음 달인 2021년 2월의 발전량을 예측하는 것이기 때문에 임의로 전체 데이터를 train set, test set로 나눠 모델링하지 않고 다음과 같이 설정하였다.

Train set : 2018년 3월-2020년 12월 / Test set : 2021년 1월

공모전과 다른 성능 평가 지표를 사용하므로 공모전의 성적이 다소 낮을 수 있다. 하지만 프로젝트의 목적에 부합하도록 모든 발전소를 같은 방식으로 분석하여 결과를 도출하였다. 중복을 피하기 위해 발전량이 가장 큰 당진 태양광 발전소에 대한 분석내용 위주로 보고서를 작성하였다.

Data Description

기상예보 데이터를 설명변수로 하여, 태양광 발전량 수치를 예측하는 모델을 생성한다.
기상 예보 데이터의 변수는 다음과 같다.

- ✓ Forecast time : 예보 발표 시점
- ✓ forecast : 예보 시간

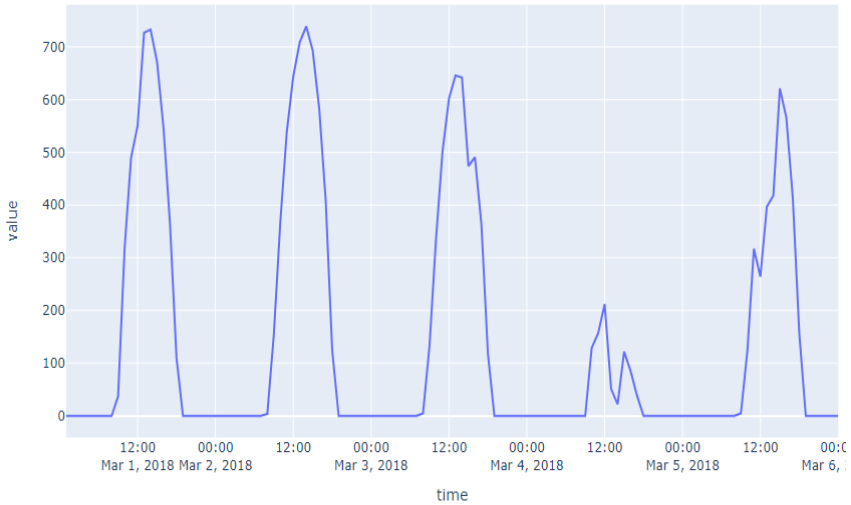
Ex) Forecast time:2018-03-01 11:00:00, forecast:4.0
=> 2018-03-01 11:00:00에 발표한 2018-03-01 15:00:00 예보

- ✓ Temperature(℃) : 예보 발표 시점 'forecast' 시간 후 온도에 대한 기상 예보
- ✓ Humidity(%) : 예보 발표 시점 'forecast' 시간 후 습도에 대한 기상 예보
- ✓ WindSpeed(m/s) : 예보 발표 시점 'forecast' 시간 후 풍속에 대한 기상 예보
- ✓ WindDirection(º) : 예보 발표 시점 'forecast' 시간 후 풍향에 대한 기상 예보
- ✓ Cloud : 예보 발표 시점 'forecast' 시간 후 하늘정보에 대한 범주형 기상 예보(1~4) - (1-맑음, 2-구름보통, 3-구름많음, 4-흐림)

	Forecast time	forecast	Temperature	Humidity	WindSpeed	WindDirection	Cloud
0	2018-03-01 11:00:00	4.0	0.0	60.0	7.3	309.0	2.0
1	2018-03-01 11:00:00	7.0	-2.0	60.0	7.1	314.0	1.0
2	2018-03-01 11:00:00	10.0	-2.0	60.0	6.7	323.0	1.0

예측변수(predictor variable)가 되는 에너지 변수는 다음과 같다.

2018-03-01 06:00:00	0.0
2018-03-01 07:00:00	0.0
2018-03-01 08:00:00	0.0
2018-03-01 09:00:00	36.0
2018-03-01 10:00:00	313.0
2018-03-01 11:00:00	532.0
2018-03-01 12:00:00	607.0
2018-03-01 13:00:00	614.0
2018-03-01 14:00:00	608.0
2018-03-01 15:00:00	641.0



매시간별 발전량이 측정되어 있으며, 매일 어느정도 반복적인 패턴을 보인다.
이때 태양광이 약한 저녁시간부터 다음 날 아침까지는 발전량이 0으로 기록된다.

EDA

탐색적 데이터 분석은 수집한 데이터가 들어왔을 때, 이를 다양한 각도에서 관찰하고 이해하는 과정이다. 이를 통해 데이터를 한 층 더 직관적으로 이해할 수 있으며, 결측치 등의 데이터의 문제점도 파악해 볼 수 있다. 뿐만 아니라 분석 계획 또한 세울 수 있기에 꼭 필요한 과정이다.

① 데이터 변환

설명변수(일기 예보) 데이터

- 예보시간을 기준으로 정렬 => 반응변수(energy) 데이터의 형태와 다르다.
- 동일 시점에 대해 다른 시간에 예보한 경우 존재 => 중복 값 존재

  Ex) 15:00에 예측한 다음날 09:00의 데이터와 23:00에 예측한 다음날 09:00의 데이터 중복

그러므로, 더 정확하고 수월한 EDA를 진행하기 위해 데이터 변환을 실시한다.
데이터 변환은 다음과 같은 방식으로 진행한다.

- 공모전의 목적이 에너지 발전이 이루어지기 전에 예측을 하는 것이므로 당일 예측 데이터는 사용하지 않는다.
- 중복된 값은 제외하고 가까운 시간에 예보된 데이터일수록 정확하다고 판단하여 제일 가까운 시간에 예측된 데이터를 사용한다. 즉, 23시가 가장 늦은 시간 예보 시간이므로 23시에 예측된 다음 날 기상예보 데이터를 사용하고 나머지는 제외한다.
- 반응변수(energy 데이터)와 동일한 형태로 정렬될 수 있도록 Forecast time + forecast를 하여 예보 데이터의 예측되는 시간을 나타내는 Forecast_time 변수를 생성한다. 그리고 Forecast_time 변수를 기준으로 energy 데이터와 동일하게 매시간 나열되게끔 데이터를 정리한다.

Forecast_time	Forecast time
2021-02-01 03:00:00	2021-01-31 23:00:00
2021-02-01 06:00:00	2021-01-31 23:00:00
2021-02-01 09:00:00	2021-01-31 23:00:00
2021-02-01 12:00:00	2021-01-31 23:00:00
2021-02-01 15:00:00	2021-01-31 23:00:00

< Forecast_time 변수 생성 >

Forecast_time	Temperature	Humidity	WindSpeed	Cloud
2021-02-01 01:00:00	NaN	NaN	NaN	NaN
2021-02-01 02:00:00	NaN	NaN	NaN	NaN
2021-02-01 03:00:00	7.0	90.0	3.4	4.0
2021-02-01 04:00:00	NaN	NaN	NaN	NaN
2021-02-01 05:00:00	NaN	NaN	NaN	NaN

< 1 시간 단위로 기상 예보 데이터를 정렬 >

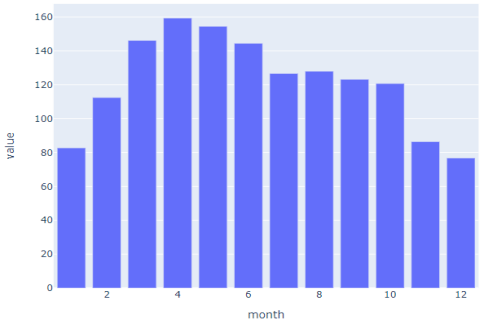
② 결측치 확인

Target이 되는 energy 데이터에는 따로 결측치가 존재하지 않으나 기상예보 데이터는 결측치가 존재한다. 이는 기상예보가 매 시간에 대해서 발표하는 것이 아니라 3시간 단위의 시간에 대한 예보만을 발표하기에 발생한 문제로 생각된다.

즉, 03시~24시 중에 3의 배수에 해당하는 시간에 대한 예보 데이터만이 존재한다. 이에 대한 해결책을 고민한 끝에 **보간법**을 활용해 NaN 값을 대체하기로 하였다.

③ Energy Data 탐색 – Target이 되는 energy data를 살펴보도록 한다.

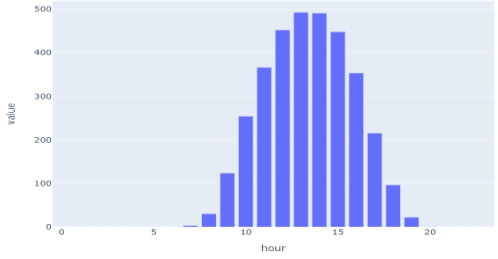
(1) 월별 평균 발전량 그래프



일반적으로 해가 가장 높게 뜨고 기온이 높은 여름(6월~8월)에 발전량이 가장 많을 것 같지만 평균적으로 봄(3월~5월)의 발전량이 더 높다. 이는 다음의 이유를 생각해보면 타당해 보인다.

- ✓ 비가 오는 날에는 태양광이 거의 존재하지 않기에 강수량이 많은 여름의 평균 발전량이 낮을 수 있다.
- ✓ 태양광 판넬의 입사각이 20~30도 이기에 무조건 태양이 높게 뜬다고 발전량이 늘어나지 않을 수 있다.

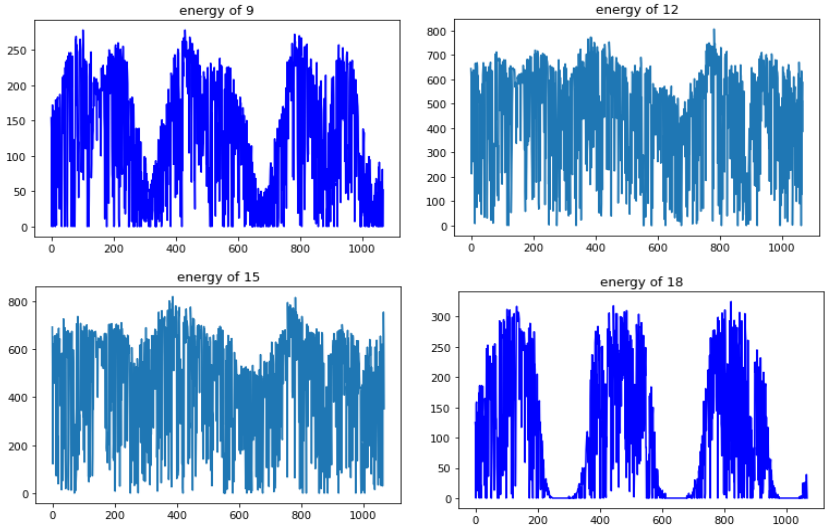
(2) 시간별 평균 발전량 그래프



20:00 ~ 06:00 : 발전량이 없음
07:00 :공모전의 평가기준을 넘기는 데이터가 존재하지 않음
13:00, 14:00 : 가장 크게 발전량이 측정됨

=> 발전량이 0인 20:00~06:00의 영향으로 월별 그래프가 평준화되어 나타났음이 의심된다. 그러므로, 각 시간별로 에너지에 대한 그래프를 살펴보기로 한다.

(3) 각 시간별 발전량 그래프



비교를 위해 NaN이 아닌 9시, 12시, 15시, 18시에 대하여 살펴보았다.

09시, 18시: energy 데이터에서 아주 높은 계절성을 발견할 수 있다. 이는, 계절에 따른 낮의 길이 등을 고려했을 때 타당한 결과로 보인다.

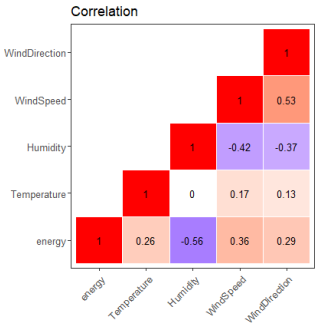
12시, 15시 : 비교적 약해 보이지만 이 역시 계절성을 포함하고 있음을 생각할 수 있다.

④ 설명변수와 반응변수의 상관관계 탐색

상관관계 : 2개의 변수가 선형 관계가 있는 범위를 표현하는 통계적 측도
상관계수 : 상관관계를 설명해주는 척도로 -1~1 범위에 있으며
절대값이 1에 가까울수록 강한 상관관계를 가지고 있음을 뜻한다.

▶ 범주형인 Cloud 변수는 제외하고 상관관계를 살펴본다.

(1) 전체 데이터내에서의 상관계수

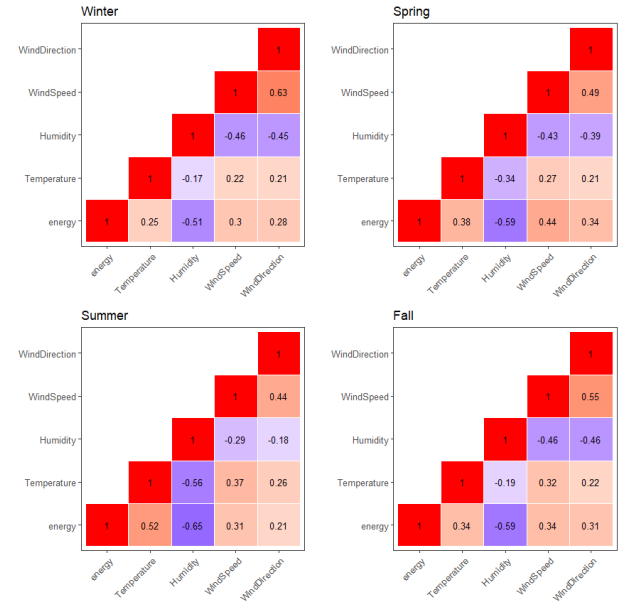


상관계수 ↑ - (에너지, 습도), (에너지, 풍속),
(습도, 풍속), (습도, 풍향), (풍속, 풍향)

상관계수 ↓ - (에너지, 온도)

상관계수 0 - (습도, 온도)

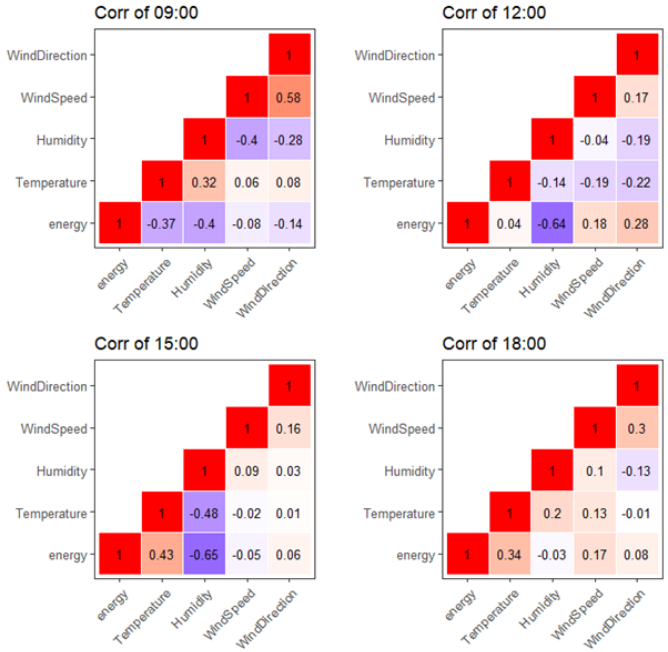
(2) 계절별 데이터내에서의 상관계수



[습도, 온도] 상관관계 : 여름에 상관계수가 높다.
[에너지, 온도] 상관관계 : 봄, 여름에는 높으나
겨울, 가을에 낮다.

전체 데이터에서의 상관계수와 계절별 데이터에서의 상관계수가 상당히 상이한 것을 고려할 때 데이터 내부에 계절성이 있음을 고려할 수 있다. 즉, 분석을 진행함에 있어 시계열적 요소를 다루는 방법을 고민해볼 필요가 있다.

(3) 시간대별 데이터내에서의 상관계수



[에너지, 온도] 상관관계
09시, 18시 : 상당히 높은 상관계수를 보인다.
12시 : 0에 가까운 상관계수가 나타난다.
⇒ 앞의 에너지 그래프에서 9시와 18시에
높은 계절성을 보였던 사실과 연관 지을 수 있다.

이외에도 시간대별로 변수 간의 상관관계가 상당히 다르게 나타나는 것을 볼 수 있다.

전체 데이터내에서의 상관관계와 계절별 또는 시간별 데이터내에서의 상관관계가 상당히 다르게 나타나므로 데이터를 분리하여 모델을 생성하는 것을 고려해본다.

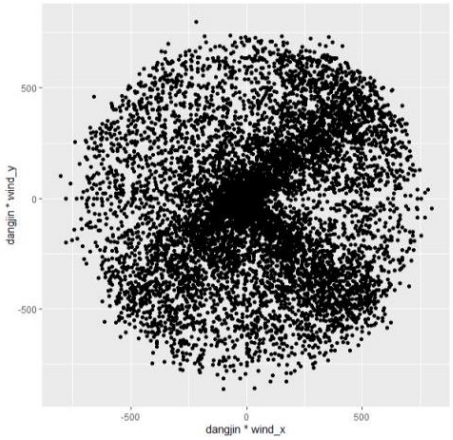
EDA 2

⑤ 설명변수의 EDA

(1) 풍향 변수에 대한 EDA

풍향변수는 각도(º)로 입력이 되어있다. 이는 방향을 나타낼 뿐 수치적인 의미는 없다.
Ex) 1º와 359º는 수치상으로는 차이가 많이 나지만 사실 거의 동일한 방향이다.

즉, 풍향 변수를 다른 형태로 변환하여 반응변수(energy)와의 관계를 알아보도록 한다.

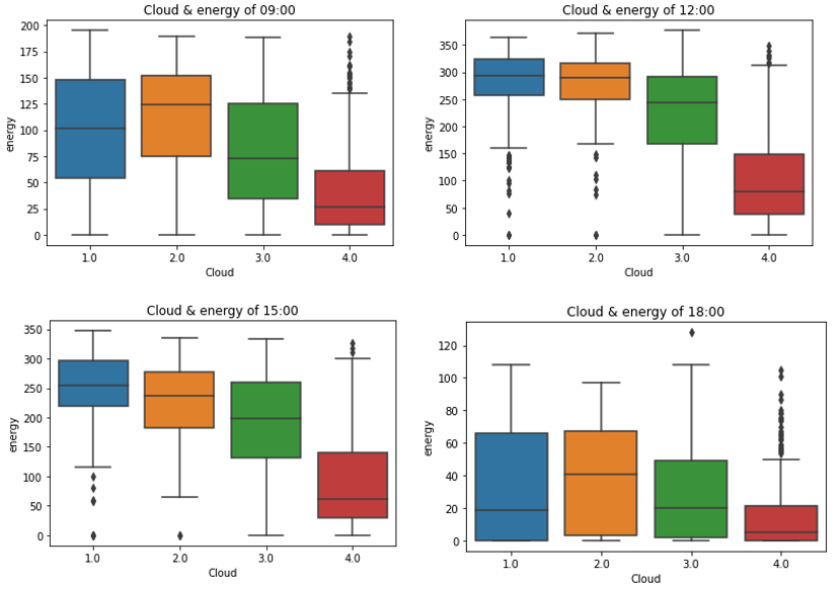


[변환 방법]
▶ 각도(º)를 Sin, Cos함수로 전환하여 2차원 평면에 표시
▶ 에너지 발전량을 중심(0,0)에서의 거리로 사용

풍향에 따라 발전량에 차이가 있다면 특정한 방향에서만 중심으로부터 멀리 떨어진 점이 많이 존재하는 형태여야 한다. 하지만 그래프를 보면, 원의 형태로 각 점이 전체적으로 골고루 분포하기 때문에 풍향과 발전량은 상관관계가 존재하지 않는다.

(2) 구름 변수에 대한 EDA

구름 변수는 1,2,3,4 로 표현된 범주형 변수이나 숫자가 클수록 구름이 많음을 뜻하므로 수치적인 의미가 있다. 우선, Boxplot으로 시간별 구름과 에너지 변수의 관계를 살펴본다.



모든 시간대에서 구름이 많아질수록 발전량이 감소하는 패턴을 보인다.
단, 시간별로 약간 다른 관계를 보인다.

[낮 시간(12:00 / 15:00)]
단조적인 감소 형태를 보인다.

[아침(09:00) / 저녁(18:00)]
단조적인 감소보다는 4(흐림)일 때 급격한 감소를 보인다.

하지만, 구름 변수가 분석에 있어 상당히 중요한 변수임은 틀림없어 보인다..
이에 따라 구름변수는 원-핫 벡터로 전환하여 분석에 사용하는 것을 고려해본다.

[원-핫 벡터란?]

표현하고 싶은 범주에 1을 부여하고, 다른 인덱스에 0을 부여하는 범주의 표현 방식

[원-핫 벡터를 사용하는 이유]

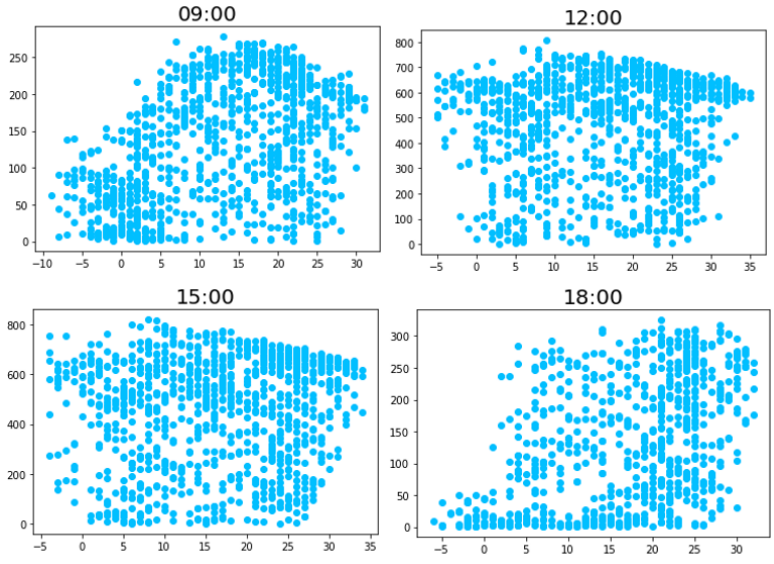
선형회귀 모델을 사용할 계획이고 이 경우, label 형태의 변수는 수치적인 의미를 가진다. 하지만, 앞서 확인한 결과 18:00 같은 경우 수치적으로 단조적인 형태가 아닌 갑자기 4에서 감소하는 패턴으로 label 형태보다는 원-핫 벡터 형태로 변수를 전환해주는 것이 변수의 효과를 더욱 잘 설명할 것으로 판단된다.

Cloud
1.0
2.0
3.0
4.0
3.0

Cloud_1.0	Cloud_2.0	Cloud_3.0	Cloud_4.0
1	0	0	0
0	1	0	0
0	0	1	0
0	0	0	1
0	0	1	0

(3) 온도 변수와 Energy 변수의 관계

앞서 살펴본 상관계수 MAP에서 아침(09:00)과 밤(18:00)에는 온도변수와 energy의 상관계수가 큰 반면 낮(12:00,15:00)에서는 상관계수가 0에 가까운 것을 확인할 수 있었다. 이를 그래프를 통해 더욱 자세히 살펴본다.



[12:00 / 15:00]
특정온도(15º정도)까지는 증가하다가 그 이후로는 감소하는 패턴

[09:00 / 18:00]
단순증가 패턴이라기 보다는, 특정구간(20º)을 지나면서 오히려 감소하는 패턴으로 보임

이 그래프를 토대로 시간별 데이터내에서의 온도 변수와 energy변수가 특정온도를 극값으로 가지는 비선형적 관계를 가짐을 고려해볼 수 있다.

Preprocessing

전처리는 앞서 확인한 EDA를 바탕으로 분석 목적에 적합하게 데이터를 가공하는 작업이다.

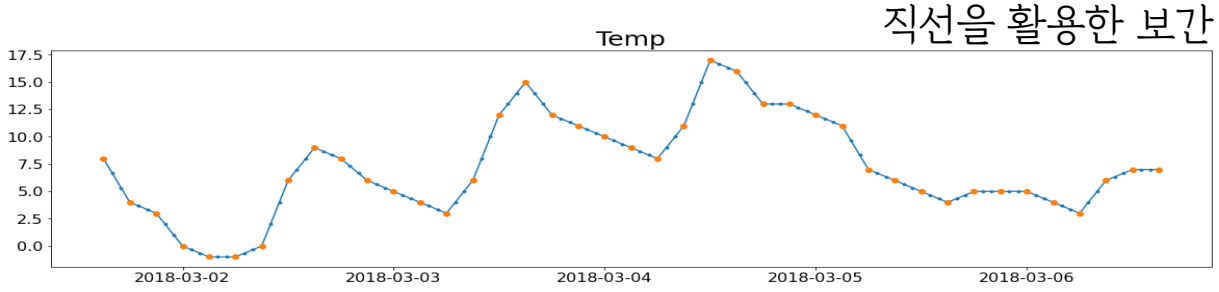
① 보간법을 통한 결측치 대체

Energy 변수: 1시간 단위 ↔ 기상예보 데이터: 3시간 단위

보간법을 통해 예보되지 않은 시간의 데이터를 대체한다.
보간은 다음과 같은 절차로 진행한다.

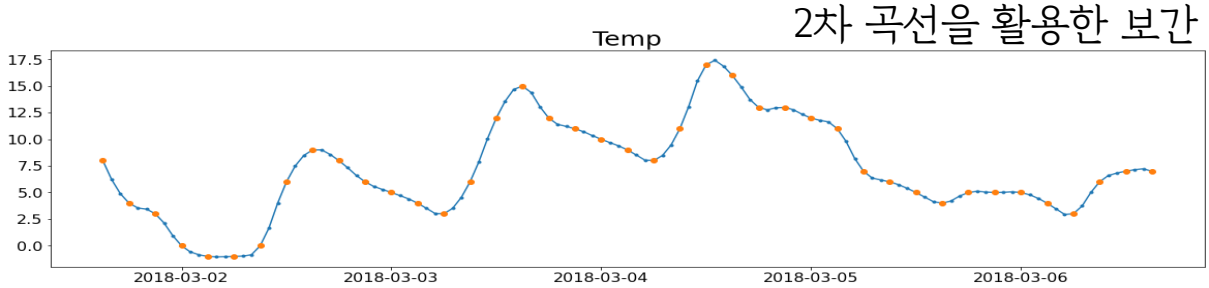
- ✓ 보간 방법은 linear, quadratic, cubic 방법 중 채택을 한다.
  특히 변수마다 시간에 따른 흐름이 다르므로 변수마다 가장 적절한 방법 적용한다.
- ✓ 기상 관측 데이터를 기상 예보 데이터와 동일한 형태로 변형하여
  이를 기준으로 보간 방법을 선정한다.
- ✓ 3가지 보간법으로 생성된 데이터 중 원본 데이터와 가장 작은 차이를 보이는
  방법을 해당 변수에 대한 보간법으로 선정한다.

직선을 활용한 보간



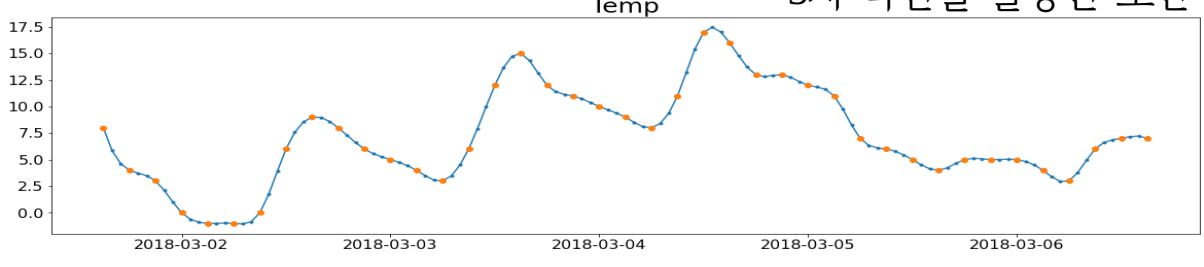
linear

2차 곡선을 활용한 보간



quadratic

3차 곡선을 활용한 보간



cublic

※ 기상 관측 데이터는 과거의 기상자료를 1시간 단위로 기록한 데이터이다.

Preprocessing 2

< 기상 관측 데이터를 활용한 보간법 선정 >

time	Temperature
2018-03-01 10:00:00	9.6
2018-03-01 11:00:00	10.5
2018-03-01 12:00:00	11.2
2018-03-01 13:00:00	10.7
2018-03-01 14:00:00	10.5
2018-03-01 15:00:00	9.9
2018-03-01 16:00:00	9.2
2018-03-01 17:00:00	8.3
2018-03-01 18:00:00	6.6
2018-03-01 19:00:00	5.6

3시간 단위 외의 데이터는 결측 처리

time	Temperature
2018-03-01 10:00:00	NaN
2018-03-01 11:00:00	NaN
2018-03-01 12:00:00	11.2
2018-03-01 13:00:00	NaN
2018-03-01 14:00:00	NaN
2018-03-01 15:00:00	9.9
2018-03-01 16:00:00	NaN
2018-03-01 17:00:00	NaN
2018-03-01 18:00:00	6.6
2018-03-01 19:00:00	NaN

Data_1
선형 보간

Data_2
2차 보간

Data_3
3차 보간

Temperature	Temp_linear	Temp_quadratic	Temp_cubic
9.6	10.200000	10.334490	10.353159
10.5	10.700000	10.929852	10.904243
11.2	11.200000	11.200000	11.200000
10.7	10.766667	11.097254	11.119667
10.5	10.333333	10.635716	10.672439
9.9	9.900000	9.900000	9.900000
9.2	8.800000	8.904208	8.864468
8.3	7.700000	7.700297	7.709703
6.6	6.600000	6.600000	6.600000
5.6	5.733333	5.655274	5.655793

원래 데이터와 비교하여 MAE 도출

temp linear error: 0.289706
temp quadra error: 0.272799
temp cubic error: 0.274704

가장 낮은 error의 방법을 선택

기상 관측 데이터를 기준으로 선정한 각 변수의 보간법은 다음과 같다

Temperature	Humidity	Wind Speed	Wind Direction	Cloud
quadratic	quadratic	linear	linear	linear

Preprocessing 2

② 변수제거 및 변환

풍향 변수와 구름 변수의 EDA 결과를 바탕으로 변수제거 및 변환을 해준다.

1) Wind Direction 변수 제거
2) Cloud 변수를 원-핫 벡터로 변환

③ 변수추가

온도 변수와 energy변수의 비선형적 관계를 해결하기 위하여 $Temp^3$ 을추가한다.

[$Temp^2$이 아닌 $Temp^3$를 추가한 이유]

특정 온도(15~20)를 극대 값으로 가지기는 하나 대칭적으로 보이지 않으므로 $Temp^3$ 항이 필요로 하다. $Temp^2$와 $Temp^3$ 모두 추가하는 방법 또한 있었으나 이는 VIF문제를 야기한다. 이에 따라 $Temp^3$ 만을 추가한다.

④ Data(Observation) Selection

프로젝트의 목적은 태양광으로 인해 전기가 발전될 때의 발전량을 예측하는 것이다. 하지만, 일조량이 없는 20:00부터 다음날 06:00까지 발전량이 없다고 볼 수 있으며 07:00 또한 없다고 보아도 무방함을 EDA를 통해 알 수 있었다.

이에 따라 더욱 정확한 모델 예측을 위해 20:00-07:00까지의 관측치를 제거하며, 동시에 남은 데이터 중에서 날씨 등의 영향으로 발전량이 0인 관측치도 제거한다.

⑤ Data Division

같은 온도라도 시간에 따라 다른 계절을 의미할 수 있다는 점과 시간마다 변수의 상관관계가 다르다는 점 등의 EDA 결과를 고려해보면 데이터를 시간 별로 분리하여 분석할 필요가 있다. 이러한 분석을 진행하기 위해 데이터를 시간 별로 분리하였다.

최종 데이터 예시(12시)

day_int	hour	Temperature	Temp3	Humidity	WindSpeed	Cloud_2.0	Cloud_3.0	Cloud_4.0	energy
20180302	12	2.0	8.0	45.0	1.6	0	0	0	644
20180303	12	9.0	729.0	50.0	0.9	1	0	0	602
20180304	12	9.0	729.0	100.0	2.4	0	0	1	212
20180305	12	6.0	216.0	75.0	3.9	0	0	1	264

Analysis 3

Regression Analysis

회귀 분석이란, 설명변수(기상예보)가 반응변수 (발전량)에 미치는 영향을 확인하는 방법으로 설명변수를 통해 반응변수를 예측하고자 한다.

lm(formula = energy ~ Temp + $\mathbf{Temp^3}$ + Humidity + WindSpeed + factor(Cloud))

9시 OLS

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	244.5	11.21	21.819	< 2e-16
Temperature	7.77	0.37	20.608	< 2e-16
Temp3	-0.005	0.0005	-9.955	< 2e-16
Humidity	-1.824	0.136	-13.43	< 2e-16
WindSpeed	-2.157	0.777	-2.778	0.0056
factor(Cloud)2	3.17	5.63	0.563	0.573
factor(Cloud)3	-35.11	4.324	-8.119	1.46e-15
factor(Cloud)4	-80.21	5.022	-15.971	< 2e-16

12시 OLS

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	738.5	21.83	33.84	< 2e-16
Temperature	4.127	1.09	3.755	< 2e-16
Temp3	-0.004	0.001	0.277	0.78
Humidity	-4.15	0.322	-12.89	< 2e-16
WindSpeed	4.144	1.914	2.165	0.03
factor(Cloud)2	-24.7	14.85	-1.664	0.096
factor(Cloud)3	-117.8	11.31	-10.411	1.46e-15
factor(Cloud)4	-264.5	14.34	-18.447	< 2e-16

17시 OLS

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	324	22.26	14.555	< 2e-16
Temperature	9.475	0.91	10.356	< 2e-16
Temp3	0.004	0.001	0.451	0.65
Humidity	-3.125	0.302	-10.337	< 2e-16
WindSpeed	7.63	1.682	4.539	6.35e-06
factor(Cloud)2	-27.92	10.83	-2.578	0.01
factor(Cloud)3	-92.3	10.18	-9.071	< 2e-16
factor(Cloud)4	-160	12.17	-13.145	< 2e-16

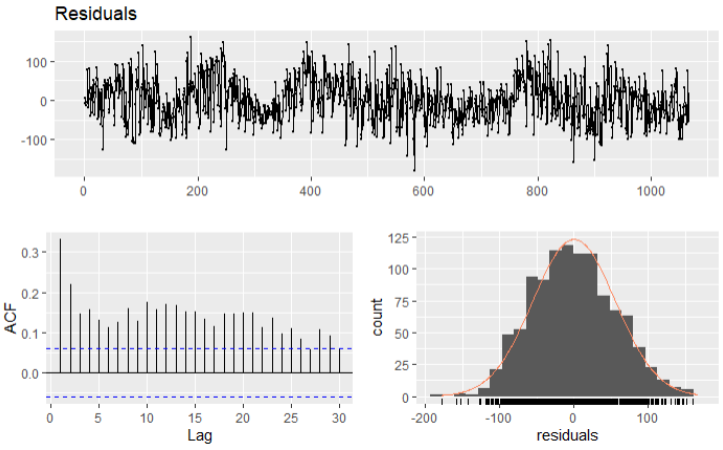
OLS RMSE										
08:00	09:00	10:00	11:00	12:00	13:00	14:00	15:00	16:00	17:00	18:00
26.02	52.47	90.82	118.3	140.5	154.1	147.3	149	138.7	118.6	77.38

Checking Improvement Points 3

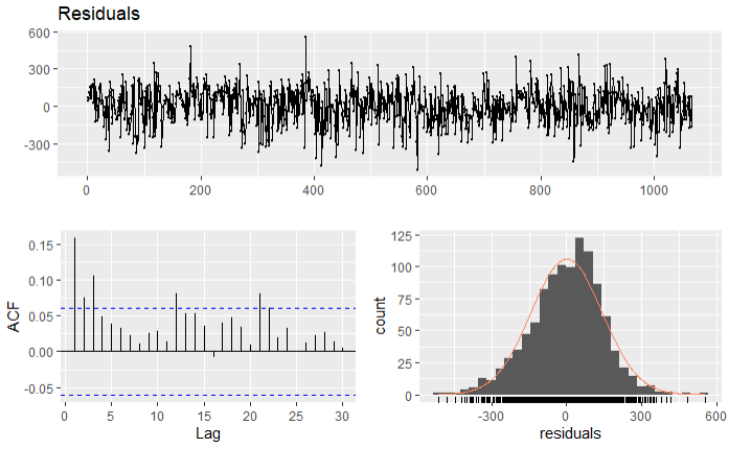
회귀 가정 확인

Regression Assumption – 선형성 / 독립성 / 등분산성 / 정규성

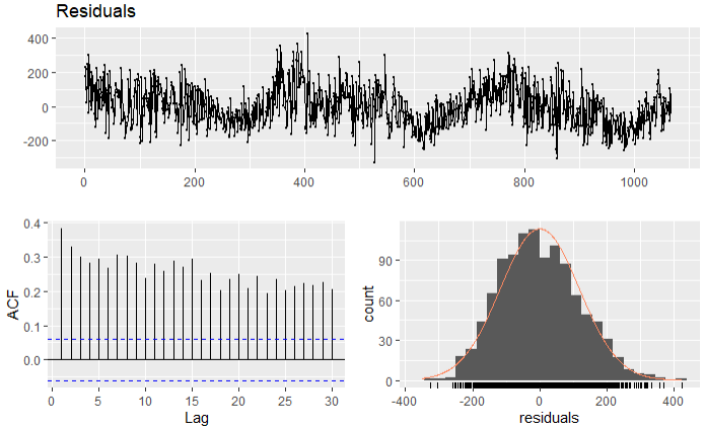
▶ 9시 OLS



▶ 12시 OLS



▶ 17시 OLS



✓ 잔차의 그래프를 확인 시,
약간의 패턴을 보이는 등
정상성을 만족하지 않는다.

✓ lag - t autocorrelation function의
값이 아주 높게 나온다.

✓ 공분산 구조가 등분산성 및 독립성을
만족하지 않는다

↓

Gauss-Markov Assumption
만족하지 않음

↓

모델 보완 필요

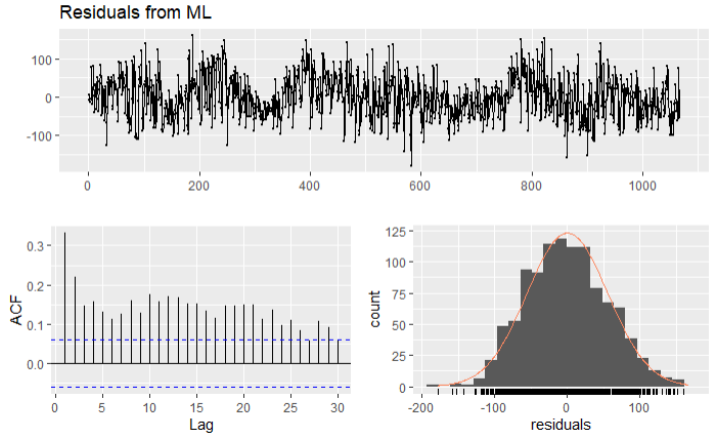
Improved Analysis 4

앞서 진행한 분석의 문제점을 해결을 위해 3가지 보완 방법 진행

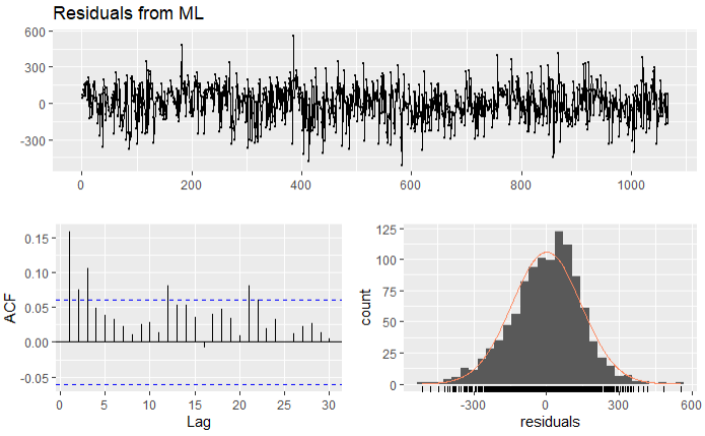
① GLS

위의 잔차 검증에서 Ordinary Least Squares Estimator를 활용했을 때 정확성(분산)의 측면에서 best인 estimator로 추정되지 않음을 알 수 있다. 따라서 이러한 문제를 해결하기 위해 Generalized Least Squares Estimator를 활용한 모형을 고려해본다.

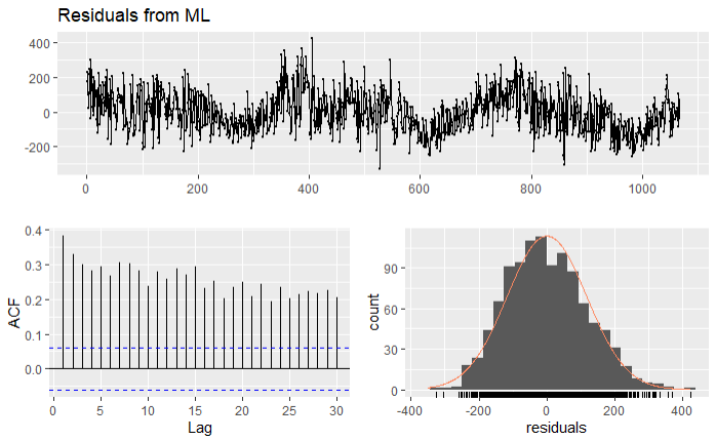
▶ 9시 GLS



▶ 12시 GLS



▶ 17시 GLS



공분산 구조를 추정하는 데 한계가 있어 정확히 GLS 모형을 적합하는데 제한사항이 있다. 특히 여전히 Auto Correlation이 존재한다. 이를 해결하기 위해 AR 모형, ARMA모형을 사용하였음에도 해결하지 못했다.

GLS RMSE										
08:00	09:00	10:00	11:00	12:00	13:00	14:00	15:00	16:00	17:00	18:00
15.24	52.44	108.19	128.25	134.91	127.97	176.61	190.66	140.66	75.81	39.91

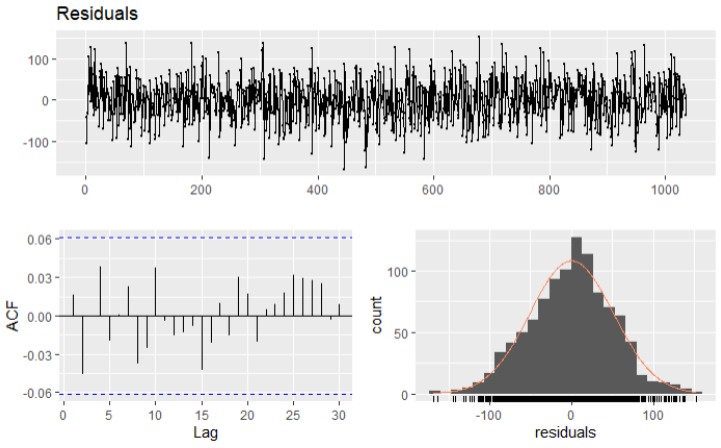
GLS모형의 시간 별 RMSE를 계산한 결과,
OLS보다 전체적으로는 성능이 증가함을 알 수 있다.
하지만 아직 대부분의 데이터가 보인 심각한 Auto Correlation 문제를 해결할 필요가 있다.

Improved Analysis

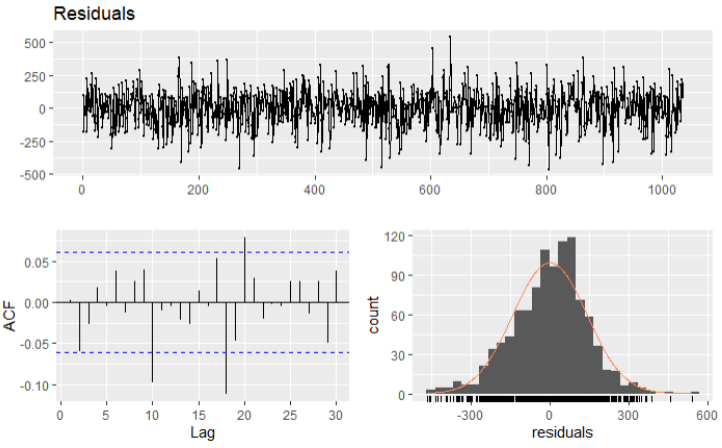
② Shuffle 및 Month 변수 추가

앞서 GLS에서 여전히 Auto Correlation의 문제가 있음을 확인하였다. 그 문제를 해결하기 위해서 Shuffle을 하여 데이터 간의 독립성을 생성하였다. 또한, EDA에서 드러난 계절성을 제거하여 등분산성을 만들기 위해서 월을 더미 변수로 추가하였다.

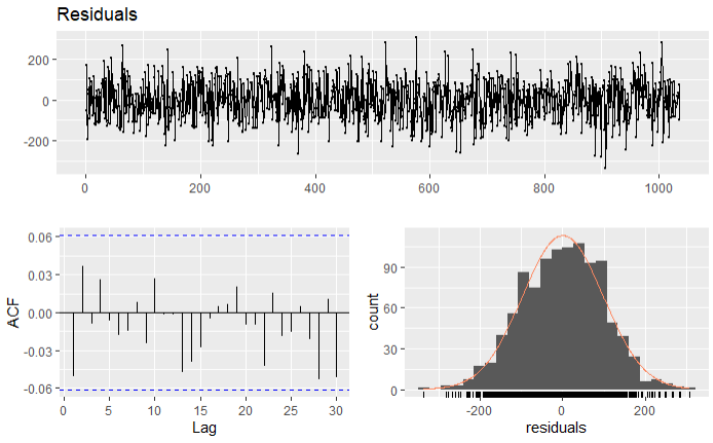
▶ 9시 Month추가 & shuffle



▶ 12시 Month추가 & shuffle



▶ 17시 Month추가 & shuffle



잔차 분석 결과, 대부분의 시간대 데이터가 Gauss-Markov Assumption을 만족한다. 하지만 여전히 가정을 만족하지 못하는 데이터가 존재하므로 해당 데이터를 따로 다룬다.

Improved Analysis 4

③ 비선형 머신러닝 적용

가정을 크게 벗어나 보이는 데이터의 경우, Gauss-Markov Assumption의 만족 여부를 확인하고자 BP-test와 Shapiro-test를 활용하여 검정을 실시하였다. 그 결과, 다음의 데이터들은 가정을 크게 벗어나기에 회귀를 적용할 수 없다고 판단한다.

울산	당진	당진 수상	당진 자재창고
17:00, 18:00	17:00, 18:00	17:00, 18:00	18:00, 19:00

이 데이터들은 주로 저녁 시간대의 데이터로 다른 시간대의 데이터들에 비해 일조시간(낮의 길이)에 영향을 많이 받는 데이터이다. 하지만, 일조시간에 대한 예보 데이터는 구할 방법이 없으므로 비선형 머신러닝을 통해 발전량을 예측한다.

데이터 별로 Cross Validation 및 Hyper-Parameter Tunning을 활용하여 Light GBM Regressoion, Random Forest Regression, KNN Regression, Support Vector Regression 중 가장 성능이 좋은 모델을 선정하였으며, 결과는 다음과 같다

데이터	울산	당진	당진 수상	당진 자재창고
17:00	LGB Regressor	RF Regressor	RF Regressor	해당 x
18:00	RF Regressor	RF Regressor	LGB Regressor	LGB Regressor
19:00	해당 x	해당 x	해당 x	LGB Regressor

앞서 진행한 분석보다 성능이 향상되었음을 확인할 수 있다.

Improved Analysis 4

[모델별 성능 비교]

모든 분석은 과적합의 위험이 있다. 이를 해결하기 위해 Validation 데이터를 설정하여 Train 데이터로부터 학습된 데이터가 얼마나 잘 예측을 하는지 확인할 필요가 있는데. 이때 통상적으로 전체 데이터를 3 : 1의 비율로 임의로 지정하여 Train_data와 Validation_data를 나눈다.

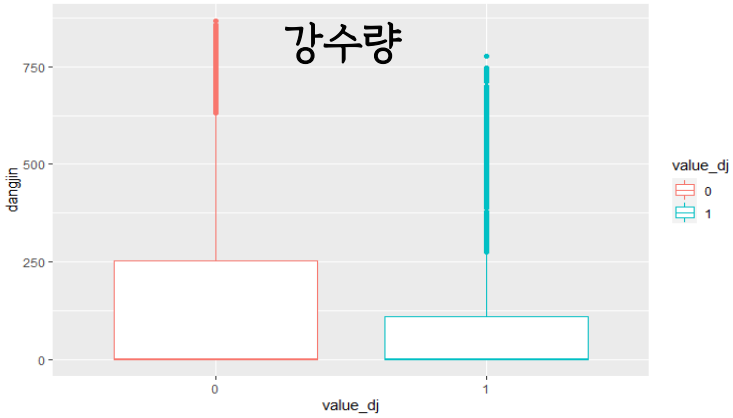
하지만 본 대회의 목적은 과거의 데이터(2018년 3월 ~ 2021년 1월)를 활용해서 미래의 값(2021년 2월)을 예측하는 것이기 때문에 마지막 달인 2021년 1월 데이터를 Validation 데이터로 나머지 과거의 데이터를 Train 데이터로 지정해서 모델의 성능을 평가한다.

Validation RMSE			
Ordinary Least Squares Estimator	Generalized Least Squares Estimator	Month & Shuffle OLS	Machine Learning 추가
133.1495	131.1131	126.1979	125.8902

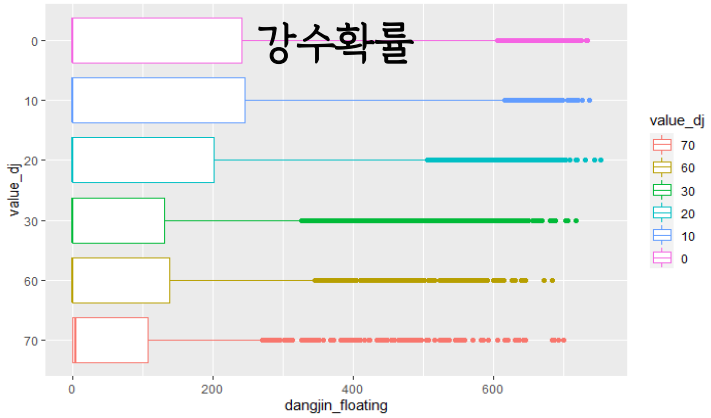
Improved Analysis

외부data 추가

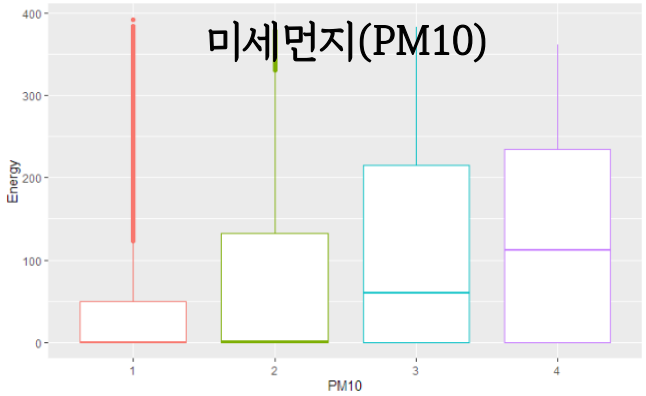
태양광 발전량 예측과 관련된 논문[5]들을 참고한 결과 기존 변수 이외에도 미세먼지, 강수량 등의 변수를 추가적으로 사용했을 때 성능이 더 개선될 수 있다는 결론을 얻었다. 해당 공모전에서 외부 데이터를 활용하는 것이 허용되므로 추가적인 변수를 고려하여 다양한 공공 데이터를 이용하였다. EDA, 데이터 전 처리, 데이터 분석의 과정은 앞의 기존 과정과 동일하게 진행했으며 최종적으로 미세먼지, 강수확률, 강수량 예보데이터를 추가 변수로 선정하여 분석을 진행했다.

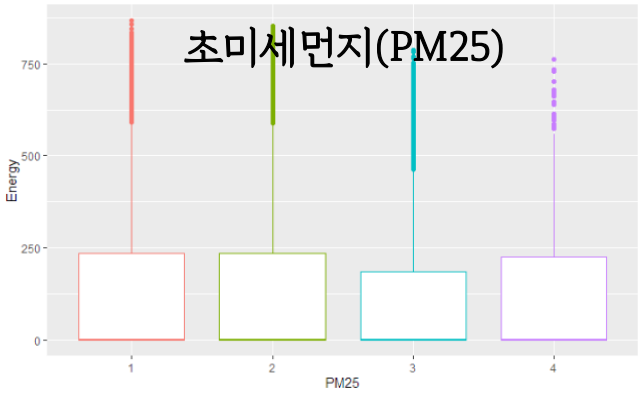


강수 유무에 따른 평균 발전량의 차이가 존재하는 것으로 보인다. 이에 따라 비가 오는 경우와 오지 않는 경우로 나누어 더미 변수로 입력했다.



강수확률에 따른 평균 발전량의 차이가 존재하는 것으로 보인다. 따라서 강수 확률을 3가지(-20%, 20%-60%, 70%-)로 나누어 더미 변수로 입력했다.



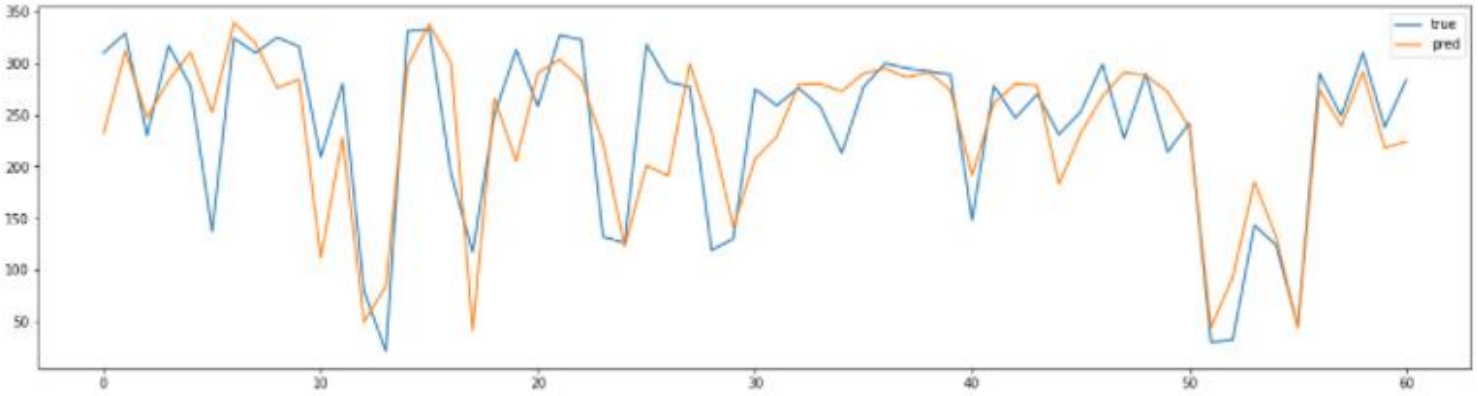


미세먼지 데이터를 EDA한 결과, PM10은 태양광 발전량에 영향이 있는 것으로 판단하였다. 그러나 PM25는 영향력이 있다고 판단이 되지 않기에 PM10만 외부변수로 채택하였다.

Validation RMSE	
외부변수 추가 전	외부변수 추가 후
125.8902	124.9791

Final Result

[최종 예측 모델]



[공모전 결과]

DACON

동서발전 태양광 발전량 예측 AI 경진대회

태양광 | 한국동서발전(주) | 개인/스타트업 | 시계열 | NMAE

상금 : 총 1,600만원

2021.06.09 ~ 2021.07.09 18:00 + Google Calendar

1,124명 D-11

17	ASMR		7.2318	60

Reference

1 https://www.yna.co.kr/view/GYH20200714000600044
〈박영석 기자〉 2020.07.14

2 https://www.lse.ac.uk/GranthamInstitute/wp-content/uploads/2018/02/Implementation-plans-for-renewable-20-by-2030.pdf

3 국가승인통계 제388003호
"2019년도 발전설비현황"
발간자료

4 기상청 – 기상관측
https://data.kma.go.kr/data/grnd/selectAsosRltmList.do?pgmNo=36

5 기계학습을 이용한 태양광
발전량 예측 및 결함 검출
시스템 개발 – 이승민, 이우진

『 Hands on Machine Learning 』
- 오렐리앙 제롱

『 실전 시계열 분석 』
- 에일린 닐슨 저

『 데이터 전 처리 대전』
- 모토하시 도모미쓰 저

『 회귀분석 강의노트 』
- 부산대학교 통계학과 김충락 교수님