

프로젝트 최종보고서

동서발전 태양광 발전량 예측 AI 경진대회

	팀	원
120210198	윤	동 성
120200369	하	동 균
120200199	김	종 현

목차

1. 프로젝트 개요.....	3
1.1 문제 제기 및 설계 목표	3
1.2 현실적 제한조건	3
2. 요구사항.....	4
2.1 설계 목표 설정.....	4
2.2 합성.....	4
2.3 분석.....	7
2.4 제작.....	15
2.5 시험.....	17
2.6 평가.....	18
2.7 환경.....	18
2.8 이후 private 평가.....	18
3. 기타.....	19
3.1 환경 구성	19
3.2 팀 구성 및 수행기간	20
4. 참조문헌.....	20
5. INDEX.....	21

1. 프로젝트 개요

1.1 문제 제기 및 설계 목표

정부와 관련기관은 중장기 전력수요를 올바르게 전망하고 전력발전을 예측한다. 이에 따른 전력 설비를 확충하거나 원자력 발전을 축소하고 재생에너지 발전을 확대하는 것은 경제적 차원을 넘는 가치를 가져온다. 이에 대한 예측이 가능하다면 보다 원활한 전력 공급 계획이 가능하다. 하지만 전력발전에는 아래와 같은 문제점이 있다.

- 1) 발전량 예측은 기상예보 데이터에 의존하지만 예보데이터는 관측데이터와 오차를 갖는다.
- 2) 발전량에 영향을 주는 기상항목은 매우 다양하며 상호 영향을 주고받는다.
- 3) 발전량은 기상조건 의한 직접적인 영향과 기상조건에 의해 영향을 받아 발생한 요인에 의해 2차 영향을 받는다.

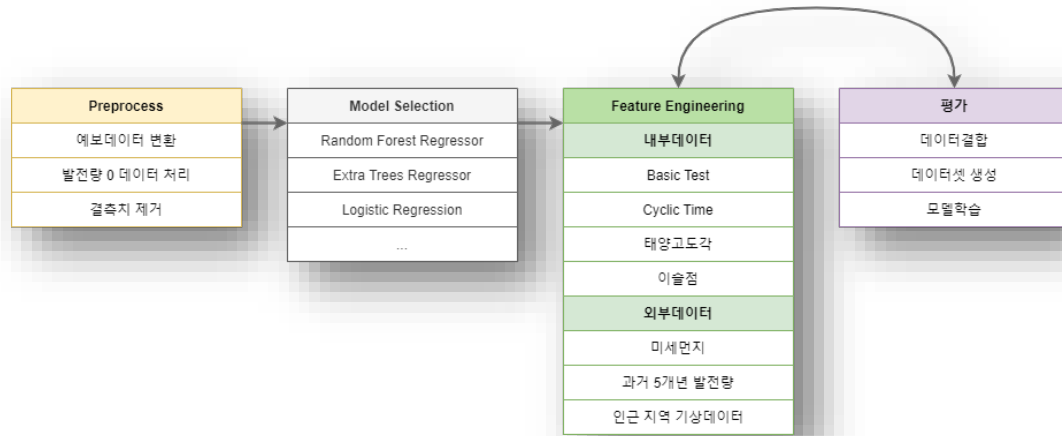
이 문제를 해결하기 위해 대회에서 주어진 기상관측 데이터와 예보데이터를 활용하고, 기상관측에 도움이 된다고 판단하는 외부데이터를 이용하여 최적의 학습모델을 생성하여 발전량을 예측한다.

1.2 현실적 제한조건

발전량의 예측은 기상예측에 의존하고 있으며, 이 기상예측의 경우 전세계 평균 오차율 40%의 불확실성이 높은 데이터이다. 따라서 이 불확실성은 발전량 예측에도 불확실성을 가져오며, 기상조건과 발전량의 불확실성으로 최적의 feature 선정을 하는 것은 불가능하다고 판단하였다. 따라서 개별 feature를 베이스라인 모델에 추가하였을 때와 그렇지 않았을 때 모델의 성능을 비교하여 feature의 추가여부를 결정하였다.

2. 요구사항

2.1 설계 목표 설정



[그림 1] 설계 프로세스 다이어그램

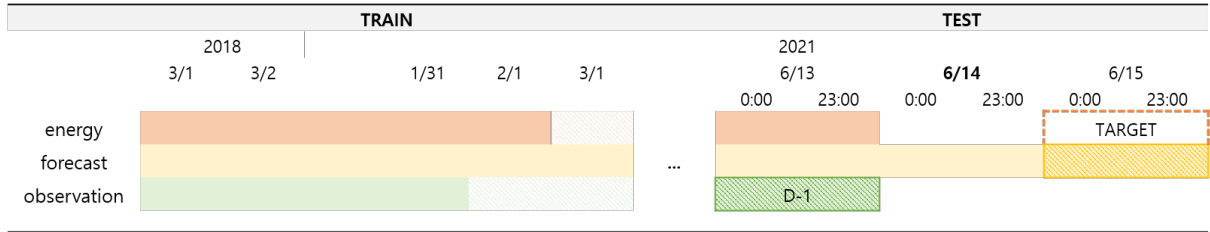
예보 데이터와 모델성능을 향상시키는 외부데이터를 추가하여 발전량 예측 성능을 가장 높게 내는 모델을 생성한다. 주최측에서 제공한 데이터로 베이스라인 모델을 만들고 외부데이터를 하나씩 추가해가며 최종 feature를 결정한다. 최종 feature리스트가 결정되면 최종모델을 학습한다.

단계	세부 목표 및 진행
1	예보데이터 전처리
2	베이스라인 모델 선정
3	예보데이터 feature engineering 및 외부데이터 전처리
4	데이터 결합 및 성능평가
5	최종 모델 선정

[표 1] 세부 목표 및 진행 단계

2.2 합성

대회의 목적은 시간별 태양광 발전량을 예측하는 것이다. 발전량 데이터는 학습에서 타겟 역할을 하기 때문에 학습데이터셋의 기간 범위를 설정할 때 중요한 기준이 된다. 대회에서 제공한 데이터셋은 예측(forecast)데이터는 3월까지 주어졌지만 발전량(energy) 데이터는 2월 1일 까지만 주어졌으므로 발전량이 없는 기간은 사용에 제약이 따른다. 대회 규정상 예측일 전날 자정까지 확인이 가능한 데이터만 학습 및 추론 과정에서 사용 가능하다. 또한 공공데이터 포털의 데이터 제공주기에 따라 당일에 취득이 불가능한 데이터가 있다. 예측일의 예측값을 사용하거나 전날 관측데이터를 포함하여 사용하는 방식으로 학습을 진행하였다.



[그림 2] 각 데이터별 날짜에 따른 적용 가능 범위

2.2.1 기상관측 데이터

대회에서 제공한 기상관측 데이터는 다섯가지 기상정보(기온, 습도, 풍속, 풍향, 전운량)를 가지고 있다. 각 기상정보를 시간별로 측정하였고, 결측치를 갖는 행이 존재한다.

	지점	지점명	일시	기온(°C)	풍속(m/s)	풍향(16방위)	습도(%)	전운량(10분위)
0	152	울산	2018-03-01 00:00	8.2	3.9	340.0	98.0	10.0
1	152	울산	2018-03-01 01:00	7.0	4.1	320.0	97.0	10.0
2	152	울산	2018-03-01 02:00	6.5	5.9	290.0	80.0	NaN
3	152	울산	2018-03-01 03:00	6.2	4.6	320.0	79.0	3.0
4	152	울산	2018-03-01 04:00	6.7	4.5	320.0	73.0	1.0
...
25627	152	울산	2021-01-31 19:00	8.8	2.5	200.0	50.0	5.0
25628	152	울산	2021-01-31 20:00	8.7	3.9	200.0	49.0	1.0
25629	152	울산	2021-01-31 21:00	8.4	2.4	230.0	51.0	7.0
25630	152	울산	2021-01-31 22:00	9.4	3.3	230.0	51.0	8.0
25631	152	울산	2021-01-31 23:00	9.3	3.1	200.0	56.0	9.0
...

[표 2] 울산지역 기상관측 데이터

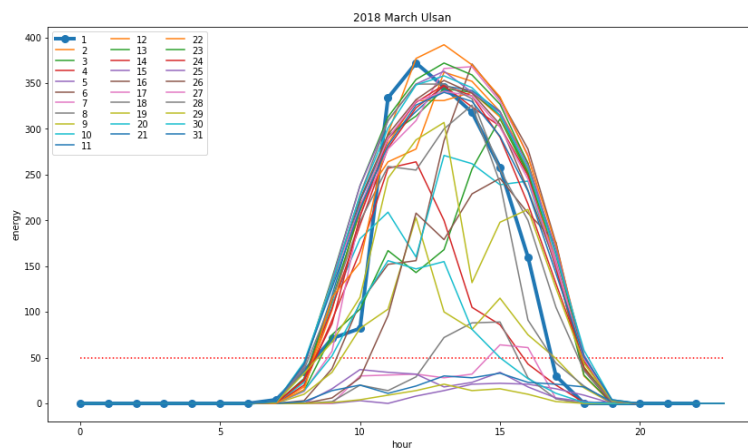
2.2.2 기상예보 데이터

기상예보는 3시간 간격으로 시행되고 4시간 후 부터 61시간 후까지 데이터를 예측하여 제공한다. 이 중에서 특정 시간대(ex: 21시) 데이터를 사용하여 학습을 진행하였다. 예보데이터는 실제 관측일과 가까울수록(14시 데이터보다 21시 데이터가) 좋다. 현업에서는 퇴근시간 같은 현실적인 이유로 17시 데이터를 사용한다. 전운량은 관측에선 0~10 이지만(0맑음, 10구름많음) 예보데이터에선 0~4로 차이점이 있다. (0맑음, 4구름많음)

	Forecast time	Forecast (X시간 후)	기온	습도	풍속	풍향	전운량
	0	4.0	8.0	20.0	14.0	298.0	2.0
	1	7.0	4.0	20.0	4.3	298.0	2.0
	2	10.0	3.0	30.0	1.9	309.0	2.0
	3	13.0	0.0	40.0	1.5	318.0	2.0
	4	16.0	-1.0	45.0	1.8	308.0	2.0
		...					
	15	49.0	13.0	50.0	2.9	172.0	2.0
	16	52.0	14.0	60.0	3.2	167.0	3.0
	17	55.0	12.0	70.0	2.6	178.0	3.0
	18	58.0	11.0	75.0	2.0	200.0	3.0
	19	61.0	10.0	80.0	1.2	275.0	3.0
		...					

[표 3] 울산지역 기상예보 데이터

2.2.3 발전량 데이터



[그림 3] 울산 2018년 3월 일별 시간별 발전량 그래프

태양광 발전량은 총 4곳의 발전량을 예측한다. 태양광 발전소는 당진과 울산 두 지역에 설치되어 있으며 당진에 세 곳, 울산 한 곳에 위치한다. 제공 데이터에는 이 태양광 발전소 정보도 제공되는데 각 발전소별 발전 용량(Capacity)이 다르다. 대회 규정상 발전량이 발전용량의 10%에 못 미치는 시간대는 평가에서 제외한다.

이름	주소	발전용량(MW)
당진수상태양광	충남 당진시 석문면 교로길 30	1.0
당진자재창고태양광		0.7
당진태양광		1.0
울산태양광	울산광역시 남구 용잠로 623	0.5

[표 4] 태양광 발전소 정보

	time	floating	dangjin warehouse	dangjin	ulsan
			...		
6	2018-03-01 7:00:00	0.0	0.0	0	0
7	2018-03-01 8:00:00	0.0	0.0	0	4
8	2018-03-01 9:00:00	36.0	33.0	37	35
9	2018-03-01 10:00:00	313.0	209.0	318	71
10	2018-03-01 11:00:00	532.0	296.0	490	82
11	2018-03-01 12:00:00	607.0	315.0	550	334
12	2018-03-01 13:00:00	614.0	474.0	727	372
13	2018-03-01 14:00:00	608.0	544.0	733	346
			...		

[표 5] 당진 시간별 발전량

2.3 분석

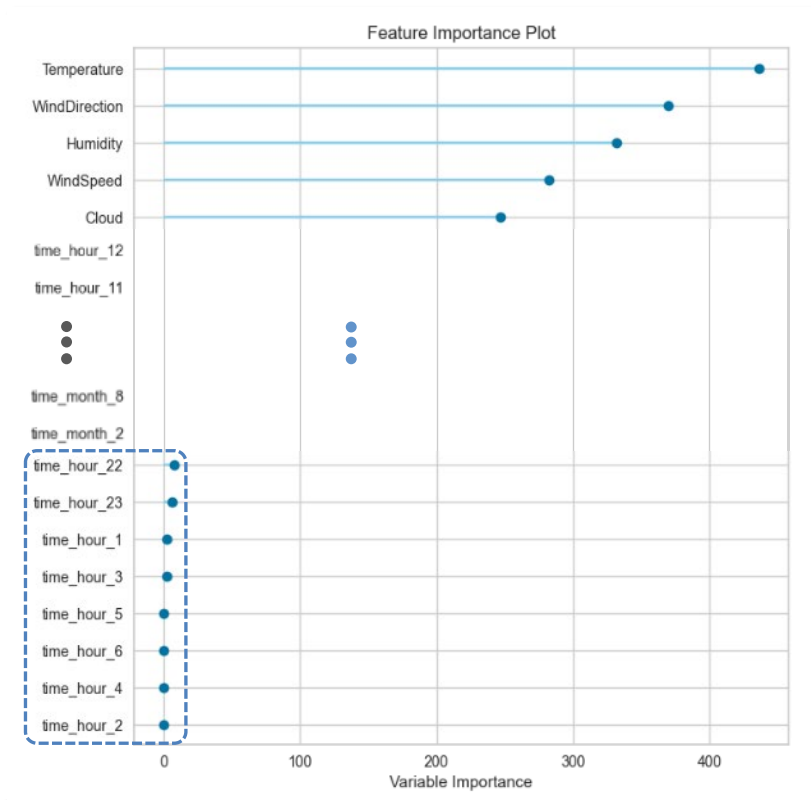
2.3.1 결측치 처리

결측치 비율	처리방법
10% 미만	삭제 또는 대치
10-20%	Hot deck, Regression, model based imputation
20% 이상	Regression, model based imputation

[표 6] 결측치 비율에 따른 처리 가이드라인

기상데이터 결측치 처리의 경우 *Hari et al(2006.6)*에서 제시한에서 제시한 가이드라인에 따른다. 데이터의 결측율이 50% 이상되면 쓰지 않는다. 10% 미만의 경우 삭제 또는 대치, 10-20%에서는 기상청의 경우 전년도 기상데이터를 가져와 Hot deck 방법을 사용한다. 당진, 울산 기상정보 데이터 결측율은 10%미만이라 삭제 처리하였다.

2.3.2 기본 전처리



[그림 4 Feature Importance Plot]

Feature 중요도 그래프는 각 feature가 출력에 주는 중요도를 보여준다. 중요도 차트에서 보면 Temperature, WindDirection, Humidity 같은 feature는 Variable Importance가 300이상의 값을 갖으며 높은 중요도를 보이는 반면 time_hour 1-6, 22, 23 같은 feature는 0에 수렴하여 출력값을 결정하는데 중요도가 매우 낮다는 사실을 확인할 수 있다. 일출시간 이전에는 다른 기상환경에 관계 없이 발전량이 무조건 0이 나오기 때문에 중요도가 낮게 나온것으로 판단하였다.

Dacon에서 제공한 기본 Baseline code에는 feature로 [year, month, day, hour, Temperature, Humidity, Wind Speed, Wind Direction, Cloud]가 사용됐다. 각각의 feature를 제거하여 성능평가를 실시하였다. 이 중에는 발전량 예측과 관련없는 feature도 존재한다. 학습을 방해하는 feature를 제거하여 모델예측 정확도를 올린다.

year	month	day	hour	Temperature	Humidity	Wind Speed	Wind Direction	Cloud	Floating	Warehouse	Dangjin	Ulsan	Average						
X	X	X	X	X	X	X	X	X	8.20	9.64	9.84	6.33	8.50						
									7.94	9.67	9.60	6.31	8.38						
									7.91	9.28	10.11	6.47	8.44						
		X							8.18	9.96	10.04	6.34	8.63						
									11.88	13.31	14.19	13.68	13.27						
	X	X	X	X	X	X	X	X	X	7.84	9.53	9.94	6.57	8.47					
										8.65	10.66	10.34	6.82	9.11					
										7.83	9.48	9.64	6.33	8.32					
										7.91	9.59	10.11	6.18	8.45					
										8.62	10.40	10.82	6.55	9.10					

[표 7] 각 feature별 포함유무에 따른 모델성능 평가 비교

2.3.3 발전량 0

발전량이 0인 경우를 시간(일출시간)에 의한 경우와 기상(악천후)에 의한 경우 두가지로 구분하였다. 시간에 의한 발전량 0은 일출 시간이 지나지 않아 발전량이 0인 경우를 나타내고, 기상에 의한 발전량 0은 일출 시간은 지났지만 기상상황에 의해 발전량이 없는 경우를 나타낸다. 시간에 의한 발전량 0의 경우 어떤 기상상황에서도 발전량이 0이므로 학습에서 제거되어야 한다고 생각하였습니다.

	Data length	floating	warehouse	dangjin	ulsan	Average Score
CASE0: all_time	25,632	8.10	9.43	9.97	6.29	8.45
CASE1: 05-20	17,629	8.19	9.70	10.01	6.33	8.56
CASE2: no_zero	12,113	8.20	9.65	9.84	6.33	8.50

[표 8] 발전량 0 데이터 제거에 따른 지역별 모델성능 비교

all_time CASE0		05-20 CASE2		No_zero CASE1
8.45	<	8.50	<	8.56

[표 9] 발전량 0 데이터 제거에 따른 모델 평균성능 비교

2.3.4 주기성

시계열데이터에서 주기성(계절, 분기, 월, 요일 등)을 갖는 데이터를 모델링할때 삼각함수 변환을 사용한다. Day of year는 0일과 364일의 날씨차이가 거의 없기에 삼각함수로 처리한다. sine과 cosine 중에 데이터와 잘 맞는지 실험을 통해 확인한다. 시계열데이터와 마찬가지로 풍향데이터도 각도라서 0~360인데 사실상 0과 360은 같은 방향을 가리킨다.

	Data length	floating	warehouse	dangjin	ulsan	Average Score
No-time	25,632	8.20	9.65	9.84	6.33	8.51
Non-cycle-time	25,632	8.45	10.36	10.75	5.82	8.85
Cycle-time	25,632	8.22	10.17	10.67	5.72	8.70

[표 10] 시간 데이터 주기성 부여에 따른 지역별 모델성능 비교

Cycle-time (Ulsan)		Non-Cycle-time (Ulsan)		No-time
5.72	<	5.82	<	6.33

[표 11] 시간 데이터 주기성 부여에 따른 모델평균성능 비교

2.3.5 태양고도각

직사광선이 태양광 패널면에 수직으로 내려올 때, 태양광 패널은 최대발전효율을 가진다. 해당 지역의 위도값, 시간대, 연중 일 데이터를 통해 그 시점의 태양의 고도를 알 수 있다. 발전소의 패널 설치각과 태양고도각의 비교를 통해 sun 피쳐를 추가해 일부발전소에서 소폭 성능개선을 하였다.

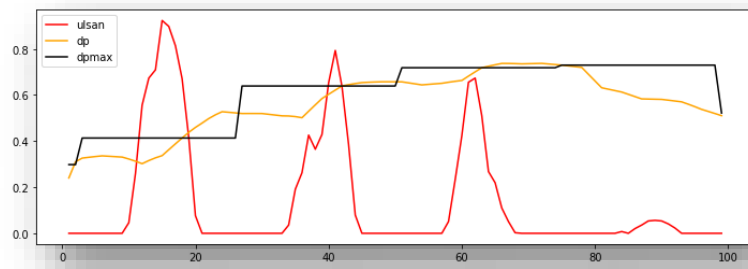
	Data length	floating	warehouse	dangjin	ulsan	Average Score
baseline	25,632	8.20	9.65	9.84	6.33	8.51
Add sun	25,632	7.93	9.66	9.47	6.37	8.36

[표 12] 태양고도각에 따른 지역별 모델성능 비교

Add sun		baseline
8.36	<	8.51

[표 13] 태양고도각에 따른 모델평균성능 비교

2.3.6 이슬점



[그림 5 발전량, 이슬점, 일별 최고온도에서의 이슬점 그래프]

이슬점은 일일간 일정하게 유지되는 편이다. 이슬점의 강하는 공기중의 수증기양이 줄어들었다는 것이고, 줄어든 만큼의 수증기가 응결되어 안개, 전운량에 영향을 주었다고 추론하였다. Magnus 공식을 이용하여 이슬점을 구했으며(이때, 상수는 기상청과 동일한 상수 선정), 하루간 가장 온도가 높을 때의 이슬점, 그리고 각 시간대별 이슬점과의 차이를 파생변수로 추가한 결과 0.04정도의 점수차로 사실상 미미한 차이를 보였다. “태양광발전 단기예측모델 개발”¹에서는 이슬점 강하를 예측시 영향을 미치는 요소로 사용한다.

	floating	warehouse	dangjin	ulsan	Average Score
baseline	8.20	9.64	9.84	6.33	8.50
DPMAX-DP	8.18	9.49	9.90	6.29	8.47

[표 14] 이슬점에 따른 지역별 모델성능 비교

DPMAX-DP		baseline	
8.47	<	8.50	

[표 15] 이슬점에 따른 모델평균성능 비교

¹ 김광득, “태양광발전 단기예측모델 개발”, 한국태양에너지학회 논문집, 2013, pp.62-69

2.3.7 미세먼지

현업에서 태양광 패널 먼지청소에 따라 발전량 효율이 달라진다는 자료를 보고 기상자료개방포털에서 시간별 미세먼지 농도 데이터를 가져와 피쳐로 적용해보니 울산발전소만 소폭 결과향상이 일어났다. 당진에 있는 세 발전소는 근처 기상청 미세먼지 관측소가 당진이 아닌 태안에 있고, 태양광 발전소 근처 당진화력발전소에서 나오는 대기오염물질에 영향을 더 받아 성능향상까지 이뤄지지 않았을 것으로 짐작된다.

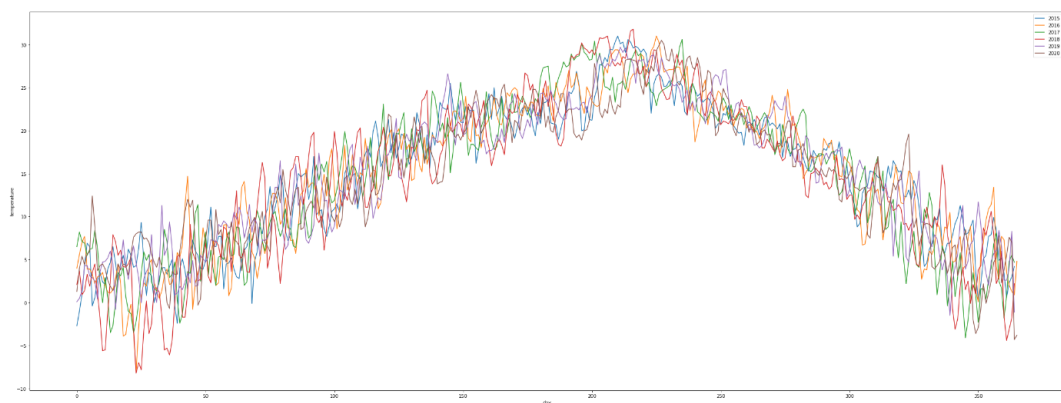
	Data length	Floating	warehouse	dangjin	ulsan	Average Score
baseline	25,632	8.20	9.65	9.84	6.33	8.51
Dust-17h	25,632	7.62	10.67	11.31	5.41	8.75
Dust-23h	25,632	7.61	10.68	11.31	5.38	8.74

[표 16] 미세먼지에 따른 지역별 모델성능 비교

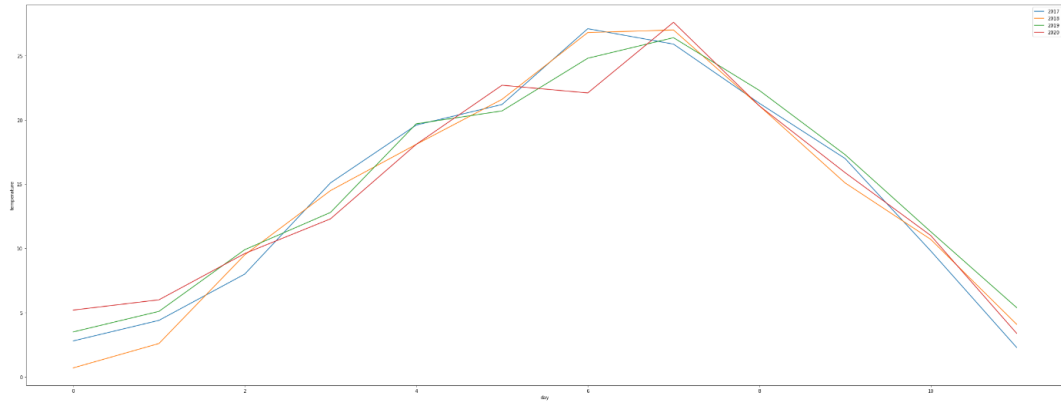
baseline		Dust-23h		Dust-17h
8.51	<	8.74	≤	8.75

[표 17] 이슬점에 따른 모델평균성능 비교

2.3.8 과거발전량



[그림 6] 2015년부터 2020년까지 연간 일별 평균 기온 변화 그래프



[그림 7] 2017년부터 2020년까지 연간 월별 평균 기온 변화 그래프

월별 평균 기온을 비교해본 결과 큰 차이가 없었고, 예보데이터 자체는 오차율이 세계평균 40% 정도인 데이터라 더 많은 데이터를 통해 모델이 잘못된 학습을 하였다. 기후변화로 6년전 기상데이터와 지금이 차이가 많아졌다. 발전량은 해당날짜에 비가 오는지 안 오는지에 따라 급격하게 달라지는데 기간이 길어질수록 강수일이 많아지면서 학습오류가 심해졌다

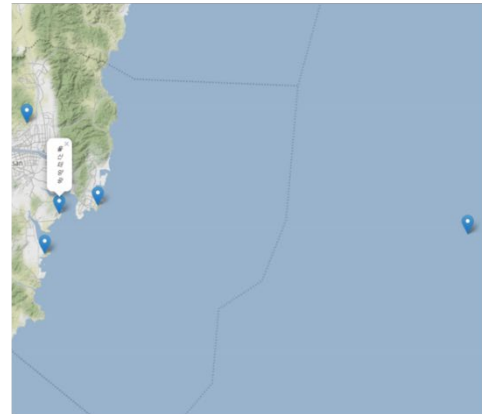
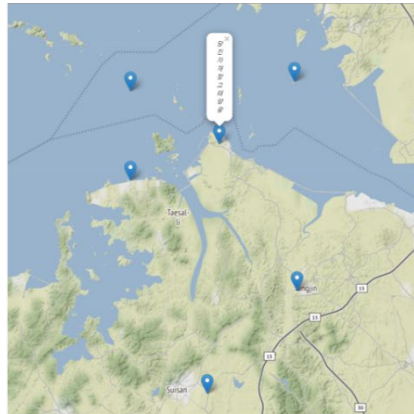
	floating	warehouse	dangjin	ulsan	Average Score
baseline	8.20	9.64	9.84	6.33	8.50
2015~2020	8.05	10.70	11.64	6.78	9.29

[표 18] 과거발전량에 따른 지역별 모델성능 비교

baseline	2015~2020
8.50	< 9.29

[표 19] 과거발전량에 따른 모델평균성능 비교

2.3.9 인접 지역 기상데이터



당진발전소 인근 기상관측소

총관기상관측: 서산
방재기상관측: 당진, 대산, 풍도, 도리도

울산발전소 인근 기상관측소

총관기상관측: 울산
방재기상관측: 울기, 장생포, 온산
해양기상부이: 울산

[그림 8] 당진과 울산발전소 인근 기상관측소

우리나라는 기상관측 데이터를 종합적으로 동시에 관측한다는 의미의 총관기상관측소 102곳과 재난방지를 목적으로 하는 방재기상관측소 510곳을 중심으로 수집한다. 또한 해수면에 장비를 띄워서 관측하는 해양기상관측과 관측장비를 기구에 매달아 하늘에서 측정하는 고층기상관측(레원존대)을 시행하고 있다. 프로젝트에서는 지상관측 데이터와 해양관측 데이터를 추가로 사용했다.

예보데이터만 이용하여 발전량을 예측 하였을때(8.503)보다 실제기상관측 데이터를 추가하여 예보 하였을때(6.975) 모델의 성능이 좋아지는 것을 확인하였다. 다만 발전소에 따라 특정 지역 데이터를 추가 여부가 달라지므로 발전소별로 최적의 조합을 선택하여 모델을 학습시킬 필요가 있다.

당진	도리도	대산	이덕서	울산	온산	풍도	울기	울산 해양	floating	warehouse	dangjin	ulsan	Average score
									8.20	9.65	9.84	6.33	8.503
○									6.60	8.99	8.90		7.703
○	○								6.51	8.76	8.76		7.589
○	○	○							6.69	9.04	8.61		7.666
○	○		○						6.42	8.78	8.66		7.545
○	○	○	○						6.62	9.02	8.56		7.633
				○								4.33	7.023
				○	○							4.23	6.998
				○	○	○						4.23	6.997
				○	○		○					4.29	7.014
				○	○			○				4.26	7.004
				○	○	○	○	○				4.14	6.975

[표 20] 인근 기상데이터 추가에 따른 지역별 모델성능 비교

2.4 제작

2.4.1 주요 함수 설명

단락	함수이름	설명
평가산식	sola_nmae	Validation Metirc
		Input: answer, pred
		output: -nmae
예보데이터를 전처리하여 훈련/검증 데이터셋의 feature와 target을 생성 (기본)	dangjin_train_datast	당진 예보데이터를 입력 받아서 학습/검증 데이터셋을 반환한다.
		input: energy_df, fcst_df, target
		output: train_x, train_y, val_x, val_y
	dangjin_test_datast	당진 예보데이터를 입력 받아서 테스트 데이터셋을 반환한다.
		input: fcst_df
		output: train_x, train_y, val_x, val_y
	ulsan_train_datast	울산 예보데이터를 입력 받아서 학습/검증 데이터셋을 반환한다.
		input: energy_df, fcst_df, target
		output: train_x, train_y, val_x, val_y
	ulsan_test_datast	울산 예보데이터를 입력 받아서 테스트 데이터셋을 반환한다.
		input: fcst_df
		output: train_x, train_y, val_x, val_y
외부데이터를 추가하여 훈련/검증 데이터셋의 feature와 target을 생성	my_train_datast	다른 지역 데이터 추가하여 데이터셋을 반환한다.
		input: energy_df, fcst_df, target, new_df_list
		output: train_x, train_y, val_x, val_y
	concat_dfs	예보 데이터를 전처리한 데이터프레임과 타지역 데이터프레임을 concatenate한다.
		input: new_df_list, _feature_df
		output: feature_df
	train_datast1	예보데이터를 전처리 한다.
		energy_df, fcst_df, target
		feature_df
	train_datast2	concat한 데이터프레임을 feature와 target으로 분리한다.
		input: feature_df, target
		output: train_x, train_y, val_x, val_y

[표 21] 주요 함수 설명

2.4.2 기본 알고리즘



Algorithm 2: Gradient-based One-Side Sampling

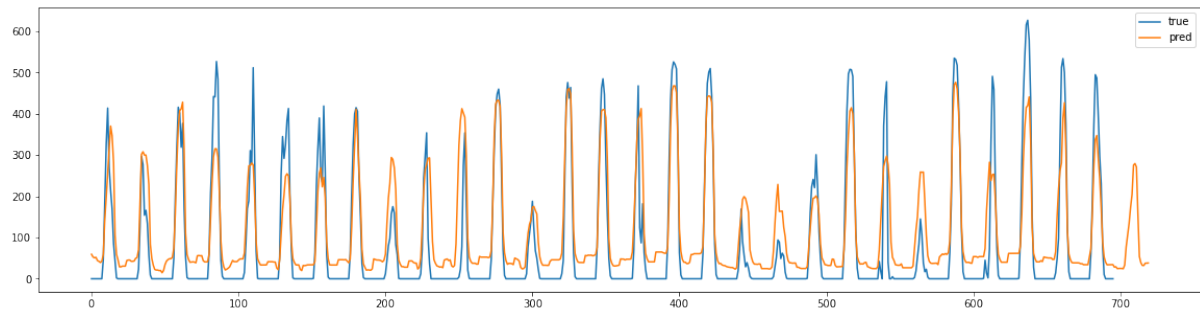
Input: I : training data, d : iterations
Input: a : sampling ratio of large gradient data
Input: b : sampling ratio of small gradient data
Input: $loss$: loss function, L : weak learner
 $models \leftarrow \{\}$, $fact \leftarrow \frac{1-a}{b}$
 $topN \leftarrow a \times \text{len}(I)$, $randN \leftarrow b \times \text{len}(I)$
for $i = 1$ **to** d **do**
 $preds \leftarrow models.predict(I)$
 $g \leftarrow loss(I, preds)$, $w \leftarrow \{1, 1, \dots\}$
 $sorted \leftarrow \text{GetSortedIndices}(\text{abs}(g))$
 $topSet \leftarrow sorted[1:topN]$
 $randSet \leftarrow \text{RandomPick}(sorted[topN:\text{len}(I)], randN)$
 $usedSet \leftarrow topSet + randSet$
 $w[randSet] \times = fact$ \triangleright Assign weight $fact$ to the small gradient data.
 $newModel \leftarrow L(I[usedSet], -g[usedSet], w[usedSet])$
 $models.append(newModel)$

[수식 1] GOSS pseudocode

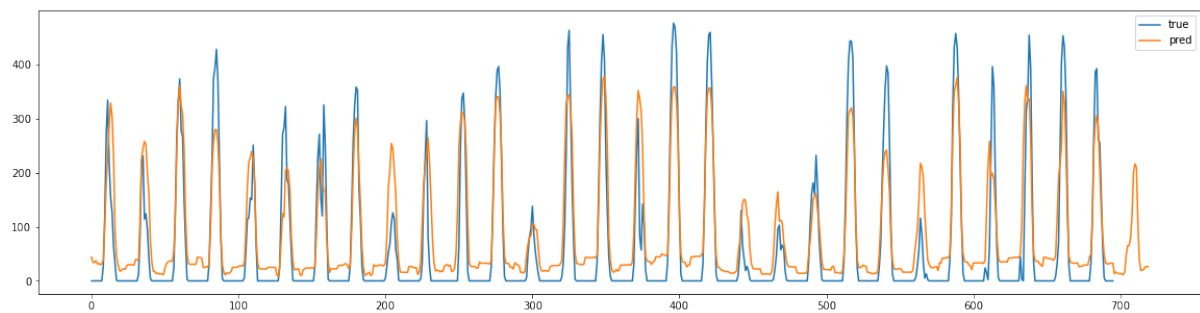
GBDT 가 고차원 변수에 큰 데이터 크기에서 효율성과 확장성이 떨어지는 것을 해결하기 위해 Gradient-based One-Side Sampling (GOSS)과 Exclusive Feature Bundling (EFB)이라는 두 가지 새로운 기술을 더해 새롭게 구현한 것이 Light GBM 이다. 기존 GBDT 대비 훈련과정이 빠르며 정확도는 거의 동일하다.

기상데이터 분석에 있어 기상청의 방식을 많이 찾아보았는데, 불확실한 데이터를 종합해 미래의 날씨를 예측하는 어려움에 있어서 하나의 모델보다 수십개의 모델을 앙상블하여 결과를 도출하고 최종적으로 기상예보관의 선택으로 내일의 날씨예보가 정해지는 식이었다. 팀 커브 또한 ensemble stacking 방식을 사용하기위해 XGB, LGBM, LSTM, CNN, RNN 등을 시도했고, 각각 개별로도 사용을 했을때 결과론적으로 LGBM 하나를 사용한 경우가 리더보드에서 팀내 최고 점수를 받아냈기에 시계열데이터를 학습하는데 있어 LGBM 을 사용하는 것이 나름 적절하다고 내부에서 결론을 지었다.

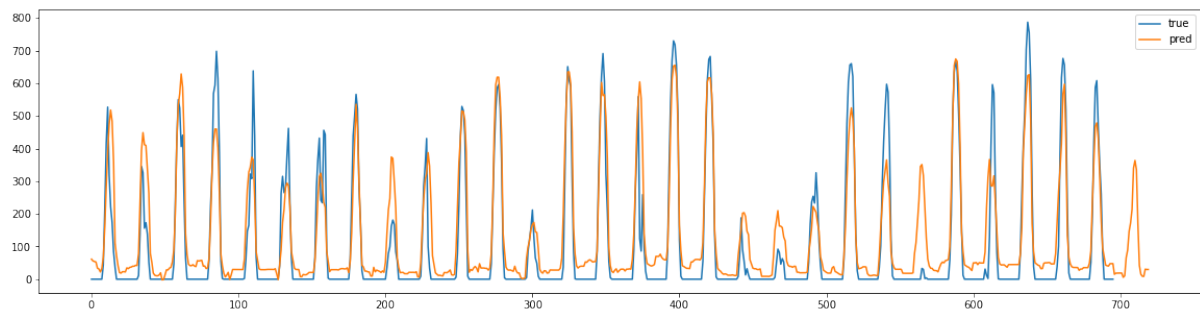
2.5 시험



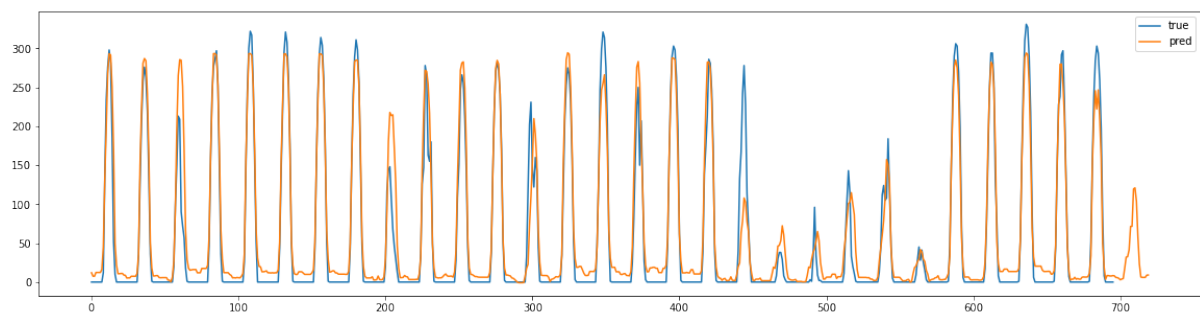
[그림 9] 당진수상태양광 발전량 예측



[그림 10] 당진자재창고태양광 발전량 예측



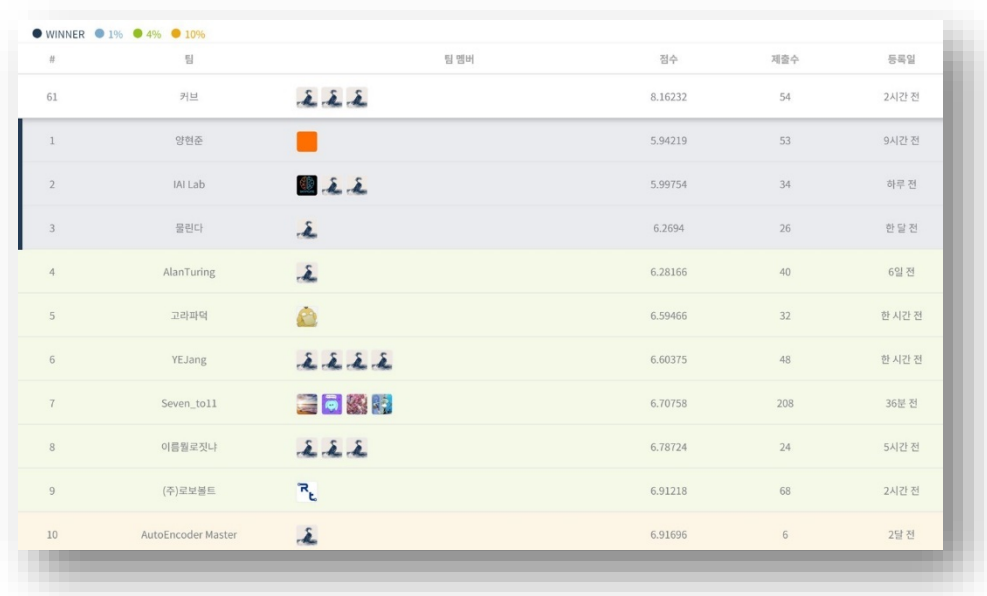
[그림 11] 당진태양광 발전량 예측



[그림 12] 울산태양광 발전량 예측

2.6 평가

public 평가를 기준으로 전체 227개 팀에서 61등(상위 26.67%)의 성적을 거두었다.



#	팀	팀 멤버	점수	제출수	등록일
61	커브		8.16232	54	2시간 전
1	양현준		5.94219	53	9시간 전
2	IAI Lab		5.99754	34	하루 전
3	물린다		6.2694	26	한 달 전
4	AlanTuring		6.28166	40	6일 전
5	고리파덕		6.59466	32	한 시간 전
6	YEJang		6.60375	48	한 시간 전
7	Seven_toll		6.70758	208	36분 전
8	이름필로짓나		6.78724	24	5시간 전
9	(주)로보월드		6.91218	68	2시간 전
10	AutoEncoder Master		6.91696	6	2달 전

[그림 13] Leader Board 26.87% (61/227)

2.7 환경

데이터 전처리와 함수 작성의 작업은 Jupyter notebook에서 작업을 하였고, 모델을 학습시키는 작업은 Google colab을 통해 작업하였다.

2.8 이후 private 평가

Public 평가 종료 시점부터 30일간 다음날의 실제 발전량을 예측한다. 6월 8일부터 7월 9일까지 30일간 매일 발전량을 예측해 제출하고, 매일 채점하여 누적점수를 공개한다.

2.8.1 기상예보 API 구하기

공공데이터 포털 동네예보 API²를 신청하여 인증키를 발급받은 후 익일 기상예보 데이터를 받는다. 예보 데이터는 3시간 간격으로 업데이트가 되며 현업에서 사용하는 17시 데이터, 예보의 정확성을 위한 20시, 23시 데이터를 사용한다.

² <https://data.go.kr/data/15057682/openapi.do>

3. 기타

3.1 환경 구성

User	윤동성
Model	Samsung nt950sbe-x716
CPU	Intel® Core™ i7-8565-U @ Base 1.80GHz Turbo 4.60GHz
RAM	16GB
GPU	Mx150
OS	Windows 10 pro
IDE	Jupyter notebook -Python 3.7.9

User	하동균
CPU	Intel(R) Core(TM) i7-1065G7 CPU @ 1.30GHz 1.5GHz
RAM	16GB
OS	Windows 10 Pro
IDE	Jupyter notebook -Python 3.6.9

User	김종현
CPU	AMD Ryzen Threadripper 1900X
RAM	32GB 2400MHz DDR4
GPU	NVIDIA Quadro P2000 D5 5GB
OS	Microsoft Windows 10 Home 1909 18363.1198
IDE	Jupyterlab 2.2.9

Model	Google Colab
CPU	Intel(R) Xeon(R)
RAM	12GB
GPU	Nvidia K80(12GB)
OS	Linux-5.4.104+-x86_64-with-Ubuntu-18.04-bionic
IDE	Jupyter Notebook

3.2 팀 구성 및 수행기간

학번	이름	기여도(%)
120210198	윤 동 성	30
120200369	하 동 균	35
120200199	김 종 현	35

프로젝트 수행기간 - 2021.04.08 ~ 2020.06.08 (62일)

4. 참조문헌

- Consequences of climatic change for water temperature and brown trout populations in Alpine rivers and streams, Hari et al, 2006
- Comparison of imputation methods for item nonresponses in a panel study, Hyejung Lee, Juwon Songa, 2017
- LightGBM: A Highly Efficient Gradient Boosting Decision Tree, Ke et al. 2017
- Solar Energy Prediction: An International Contest to Initiate Interdisciplinary Research on Compelling Meteorological Problems, MCGovern et al, 2015.
- Dacon Codeshare, Baseline, <https://dacon.io/competitions/official/235720/codeshare/2512>, 2021.06.17
- Dacon Codeshare, OpenAPI, <https://dacon.io/competitions/official/235720/codeshare/2555>, 2021.06.17
- 김광득, "태양광발전 단기예측모델 개발", 한국태양에너지학회 논문집, 2013, pp.62-69
- 예보관 훈련용 기술서 (초급) - 대기물리, 기상청

5. INDEX

표

[표 1] 세부 목표 및 진행 단계	4
[표 2] 울산지역 기상관측 데이터	5
[표 3] 울산지역 기상예보 데이터	6
[표 4] 태양광 발전소 정보	6
[표 5] 당진 시간별 발전량	7
[표 6] 결측치 비율에 따른 처리 가이드라인	7
[표 7] 각 feature별 포함유무에 따른 모델성능 평가 비교	9
[표 8] 발전량 0 데이터 제거에 따른 지역별 모델성능 비교	9
[표 9] 발전량 0 데이터 제거에 따른 모델 평균성능 비교	9
[표 10] 시간 데이터 주기성 부여에 따른 지역별 모델성능 비교	10
[표 11] 시간 데이터 주기성 부여에 따른 모델평균성능 비교	10
[표 12] 태양고도각에 따른 지역별 모델성능 비교	10
[표 13] 태양고도각에 따른 모델평균성능 비교	10
[표 14] 이슬점에 따른 지역별 모델성능 비교	11
[표 15] 이슬점에 따른 모델평균성능 비교	11
[표 16] 미세먼지에 따른 지역별 모델성능 비교	12
[표 17] 이슬점에 따른 모델평균성능 비교	12
[표 18] 과거발전량에 따른 지역별 모델성능 비교	13
[표 19] 과거발전량에 따른 모델평균성능 비교	13
[표 20] 인근 기상데이터 추가에 따른 지역별 모델성능 비교	14
[표 21] 주요 함수 설명	15

그림

[그림 1] 설계 프로세스 다이어그램	4
[그림 2] 각 데이터별 날짜에 따른 적용 가능 범위.....	5
[그림 3] 울산 2018년 3월 일별 시간별 발전량 그래프.....	6
[그림 4 Feature Importance Plot]	8
[그림 5 발전량, 이슬점, 일별 최고온도에서의 이슬점 그래프]	11
[그림 6] 2015년부터 2020년까지 연간 일별 평균 기온 변화 그래프.....	12
[그림 7] 2017년부터 2020년까지 연간 월별 평균 기온 변화 그래프.....	13
[그림 8] 당진과 울산발전소 인근 기상관측소	14
[그림 9] 당진수상태양광 발전량 예측.....	17
[그림 10] 당진자재창고태양광 발전량 예측	17
[그림 11] 당진태양광 발전량 예측	17
[그림 12] 울산태양광 발전량 예측	17
[그림 13] Leader Board 26.87% (61/227).....	18

수식

[수식 1] GOSS pseudocode.....	16
-----------------------------	----