#BIG_DATA_PROGRAMMING

Solar - Prediction

#강예진 60201658
#박선진 60180520
#장유진 60201703

목차

A table of Contents

Part 1.

데이터 설명

데이터 설명

Solar Radiation Prediction

- NASA HI-SEAS Project
- 화성과 비슷한 환경의 Hawaii 마우나로아 화산
- 2016년 9월 ~ 12월 동안 측정한 data
- 30000개 정도 데이터
- *Kaggle: Solar Radiation Prediction*

Part 1

데이터 설명

SolarPrediction.csv (2.96 MB)

Detail Compact Column — 11 of 11 columns

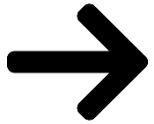
UNIXTime	Data	Time	Radiation	Temperature	Pressure	Humidity	WindDirection(De...	Speed	TimeSunRise	TimeSunSet
1.47b – 1.48b	1Sep16 – 31Dec16	20Sep19 – 21Sep19	1.11 – 1.6k	34 – 71	30.2 – 30.6	8 – 103	0.09 – 360	0 – 40.5	20Sep19 – 20Sep19	21Sep19 – 21
	AM									
1475228421	9/29/2016 12:00:00 AM	23:40:21	1.21	48	30.46	60	137.71	3.37	06:13:00	18:13:00
1475228124	9/29/2016 12:00:00 AM	23:35:24	1.17	48	30.46	62	104.95	5.62	06:13:00	18:13:00
1475227824	9/29/2016 12:00:00 AM	23:30:24	1.21	48	30.46	64	120.2	5.62	06:13:00	18:13:00
1475227519	9/29/2016 12:00:00 AM	23:25:19	1.2	49	30.46	72	112.45	6.75	06:13:00	18:13:00
1475227222	9/29/2016 12:00:00 AM	23:20:22	1.24	49	30.46	71	122.97	5.62	06:13:00	18:13:00
1475226922	9/29/2016 12:00:00 AM	23:15:22	1.23	49	30.46	80	101.18	4.5	06:13:00	18:13:00

- UNIXTime >> 유닉스 타임
- *Date >> 일자*
- *Time >> 하와이 기준 시간*
- *Radiation >> 태양 복사*
- *Temperature >>기온*
- *Pressure >> 기압*
- *Humidity >> 습도*
- *WindDirection >> 풍향*
- *Speed >> 풍속*
- *TimeSunRise >> 일출 시간*
- *TimeSunSet >> 일몰 시간*

Part 1

데이터 설명

# UNIXTime	Data	Time	TimeSunRi...	TimeSunSet
1475229326	9/29/2016 12:00:00 AM	23:55:26	06:13:00	18:13:00
1475229023	9/29/2016 12:00:00 AM	23:50:23	06:13:00	18:13:00
1475228726	9/29/2016 12:00:00 AM	23:45:26	06:13:00	18:13:00
1475228421	9/29/2016 12:00:00 AM	23:40:21	06:13:00	18:13:00
1475228124	9/29/2016 12:00:00 AM	23:35:24	06:13:00	18:13:00



Time
23:55:26
23:50:23
23:45:26
23:40:21
23:35:24

hour	minute	second
23	55	26
23	50	23
23	45	26
23	40	21
23	35	24

Time >> 시 / 분 / 초

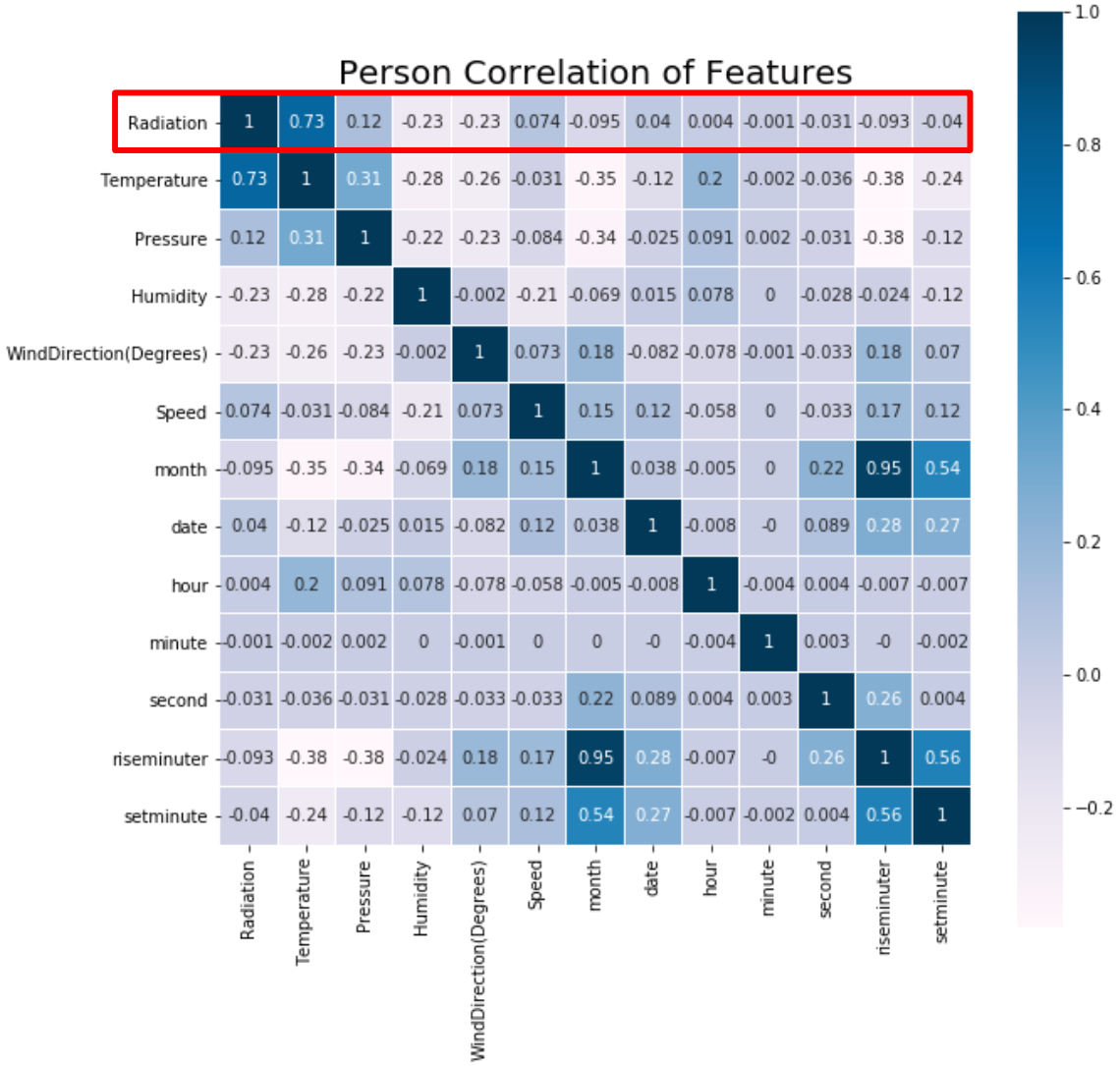
Part 2.

EDA

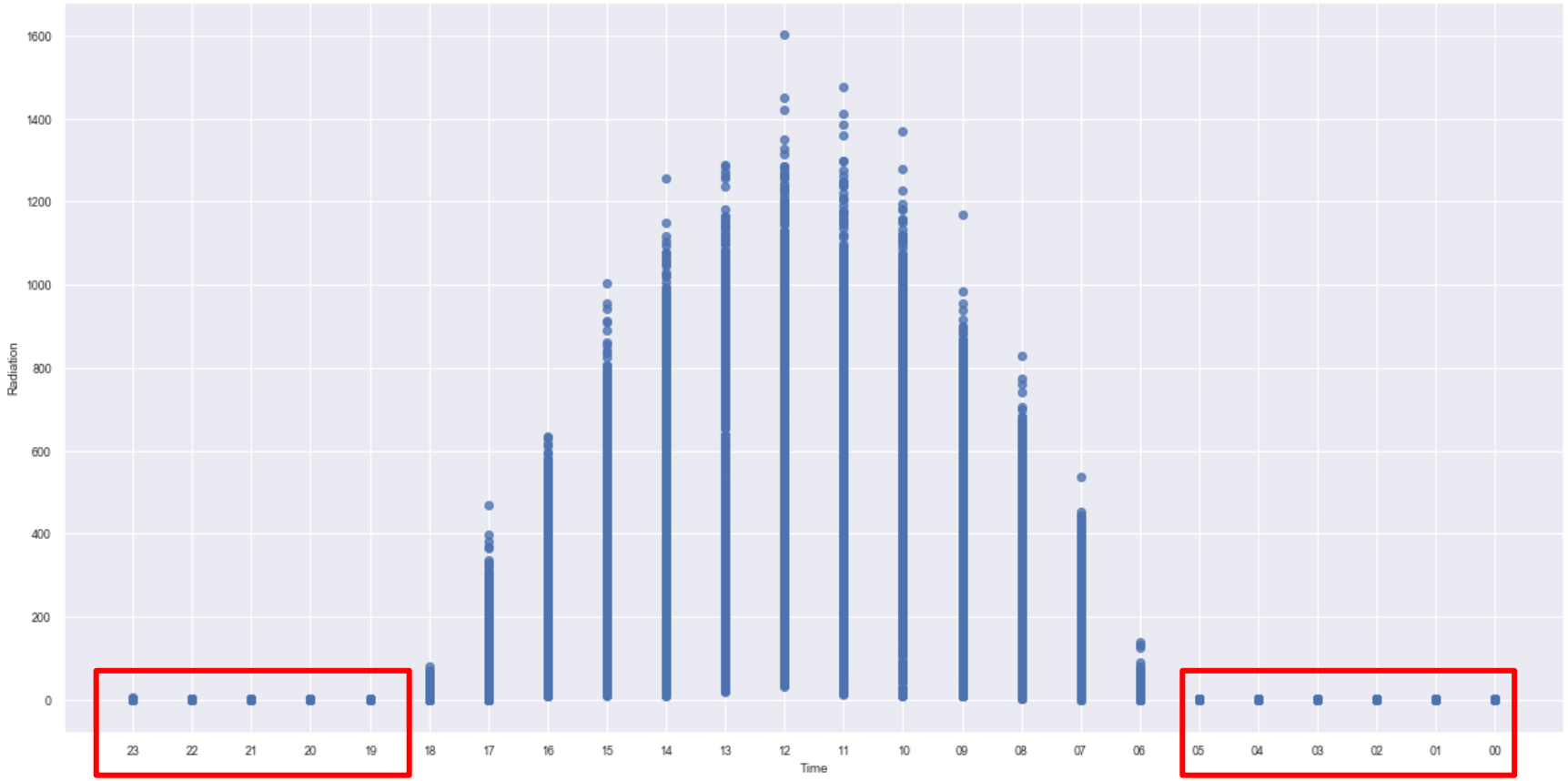
EDA

상관관계 분석

- Radiation 일사량과의 correlation
- 기온, 기압, 습도에서 유의미함을 찾음
- **시간과의 상관관계 값이 낮은 이유?**



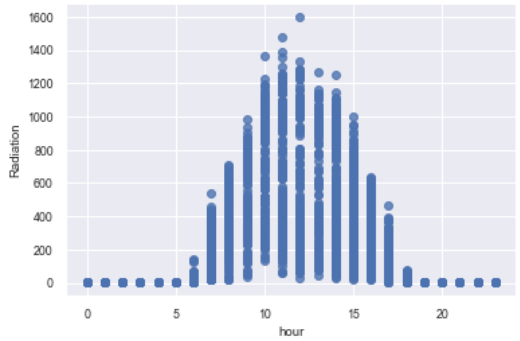
EDA



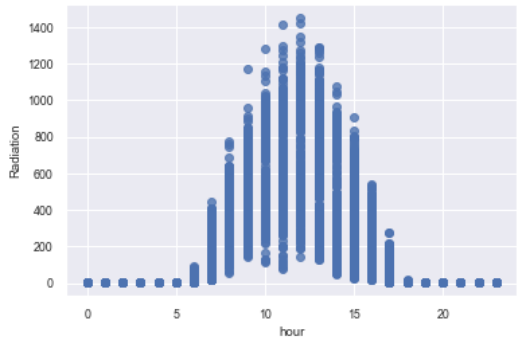
시간별 Scatter plot

EDA

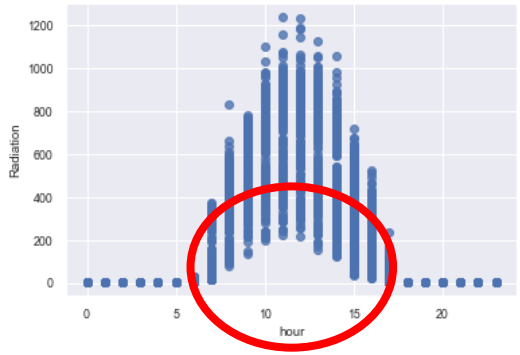
월별 Scatter plot



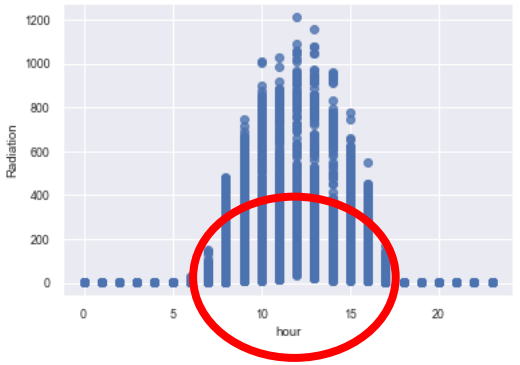
9월



10월



11월



12월

- 월별 Radiation 분포 양상의 차이
- 날씨와 관련이 있지 않을까?

EDA

NATIONAL WEATHER SERVICE
NATIONAL OCEANIC AND ATMOSPHERIC ADMINISTRATION

- 2016년 Hawaii 가까운 지역의 날씨 자료 검색
- 월별 평균 강수량 11월이 가장 낮고,

  12월이 가장 높음
- Column의 연관성 도출

Observed Weather | Climate Prediction and Variability | Local Data/Records

NOWData - NOAA Online Weather Data Enlarge results Print

Monthly Total Precipitation for HILO INTERNATIONAL AP 87, HI

Click column heading to sort ascending, click again to sort descending.

Year	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec	Annual
2016	1.29	3.28	4.89	9.64	5.83	11.19	7.32	24.68	10.74	19.13	10.04	20.40	128.43
Mean	1.29	3.28	4.89	9.64	5.83	11.19	7.32	24.68	10.74	19.13	10.04	20.40	128.43
Max	1.29 2016	3.28 2016	4.89 2016	9.64 2016	5.83 2016	11.19 2016	7.32 2016	24.68 2016	10.74 2016	19.13 2016	10.04 2016	20.40 2016	128.43 2016
Min	1.29 2016	3.28 2016	4.89 2016	9.64 2016	5.83 2016	11.19 2016	7.32 2016	24.68 2016	10.74 2016	19.13 2016	10.04 2016	20.40 2016	128.43 2016

EDA

pyspark dataframe aggregation
-평균

```
>>> avg_temp.show()
+-----+-----------------+
|month|         avg_temp|
+-----+-----------------+
|    9| 53.68113792638533|
|   10| 52.46865434757964|
|   11| 50.78500724287784|
|   12|47.608892699657034|
+-----+-----------------+
```

- 기온

```
>>> avg_hum.show()
+-----+-----------------+
|month|          avg_hum|
+-----+-----------------+
|   12| 79.5264576188143|
|    9|79.48577592018336|
|   10|78.94637796168234|
|   11|62.38495895702559|
+-----+-----------------+
```

- 습도

```
>>> avg_prs.show()
+-----+-----------------+
|month|          avg_prs|
+-----+-----------------+
|   11| 30.44577981651425|
|   10|30.438462759324807|
|    9|  30.4320978832418|
|   12|30.374427976481897|
+-----+-----------------+
```

- 기압

EDA

Feature selection

- 기온(Temperature)
- 기압(Pressure)
- 습도(Humidity)
- 풍향(WindDirection)
- 풍속(Speed)
- month
- date
- year
- hour
- minute
- second
- seminuter
- setminute

```
['Temperature', 'Pressure', 'Humidity', 'WindDirection(Degrees)', 'Speed', 'month', 'date', 'year', 'hour', 'minute', 'second',
seminuter', 'setminute']
```

Part 3

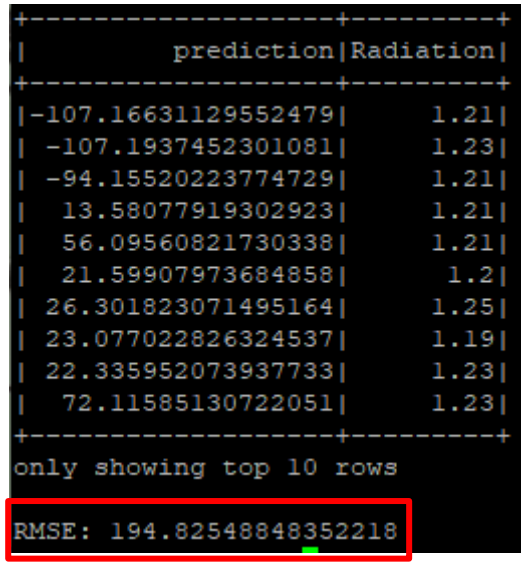
모델링

모델링

Pyspark Mllib

Linear regression

=>



```
+------------------+---------+
|        prediction|Radiation|
+------------------+---------+
|-107.16631129552479|     1.21|
| -107.1937452301081|     1.23|
| -94.15520223774729|     1.21|
|  13.58077919302923|     1.21|
|  56.09560821730338|     1.21|
|  21.59907973684858|      1.2|
| 26.301823071495164|     1.25|
| 23.077022826324537|     1.19|
| 22.335952073937733|     1.23|
|  72.11585130722051|     1.23|
+------------------+---------+
only showing top 10 rows

RMSE: 194.82548848352218
```

=> Random Forest

- RMSE 값 194.8254 >> 너무 높다 판단

모델링

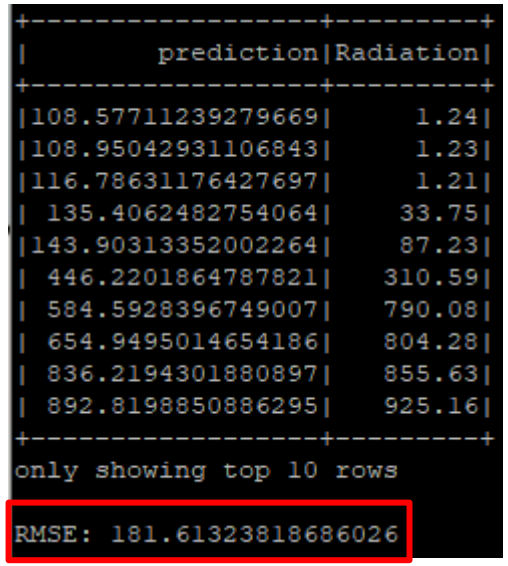
전체 데이터 RF

```
+------------------+---------+
|        prediction|Radiation|
+------------------+---------+
| 20.64930932403363|     1.21|
| 20.64930932403363|     1.23|
| 21.88778013672256|     1.21|
|20.127617574793227|     1.21|
|20.127617574793227|     1.21|
|19.474924950520755|      1.2|
|16.578980361202103|     1.25|
|18.405581801487337|     1.19|
|18.595186123626906|     1.23|
|20.342016086796868|     1.23|
+------------------+---------+
only showing top 10 rows

RMSE: 133.6635324751923
```

모델링

전체 시간대 RF

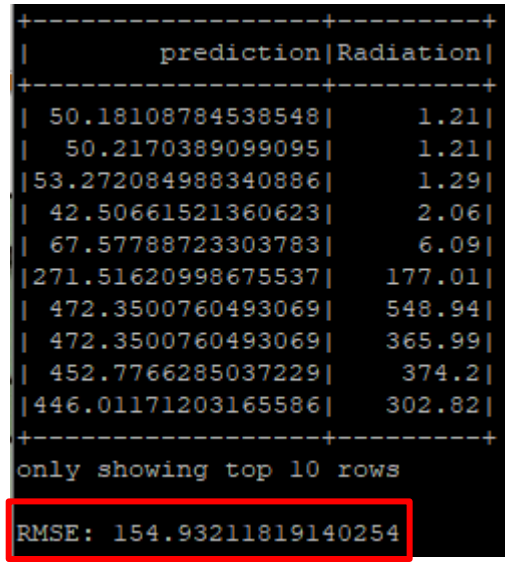


prediction	Radiation
108.57711239279669	1.24
108.95042931106843	1.23
116.78631176427697	1.21
135.4062482754064	33.75
143.90313352002264	87.23
446.2201864787821	310.59
584.5928396749007	790.08
654.9495014654186	804.28
836.2194301880897	855.63
892.8198850886295	925.16

only showing top 10 rows

RMSE: 181.61323818686026

9월

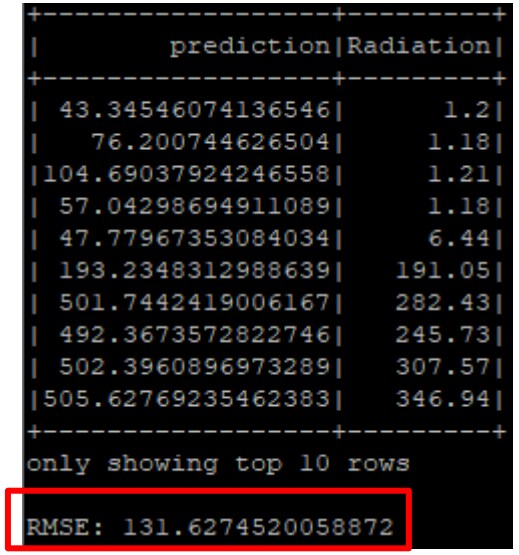


prediction	Radiation
50.18108784538548	1.21
50.2170389099095	1.21
53.272084988340886	1.29
42.50661521360623	2.06
67.57788723303783	6.09
271.51620998675537	177.01
472.3500760493069	548.94
472.3500760493069	365.99
452.7766285037229	374.2
446.01171203165586	302.82

only showing top 10 rows

RMSE: 154.93211819140254

10월

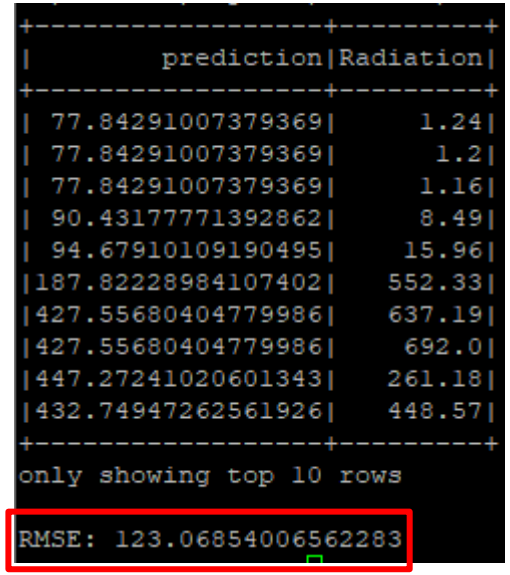


prediction	Radiation
43.34546074136546	1.2
76.200744626504	1.18
104.69037924246558	1.21
57.042986949110 89	1.18
47.77967353084034	6.44
193.2348312988639	191.05
501.7442419006167	282.43
492.3673572822746	245.73
502.3960896973289	307.57
505.62769235462383	346.94

only showing top 10 rows

RMSE: 131.6274520058872

11월



prediction	Radiation
77.84291007379369	1.24
77.84291007379369	1.2
77.84291007379369	1.16
90.43177771392862	8.49
94.67910109190495	15.96
187.82228984107402	552.33
427.55680404779986	637.19
427.55680404779986	692.0
447.27241020601343	261.18
432.74947262561926	448.57

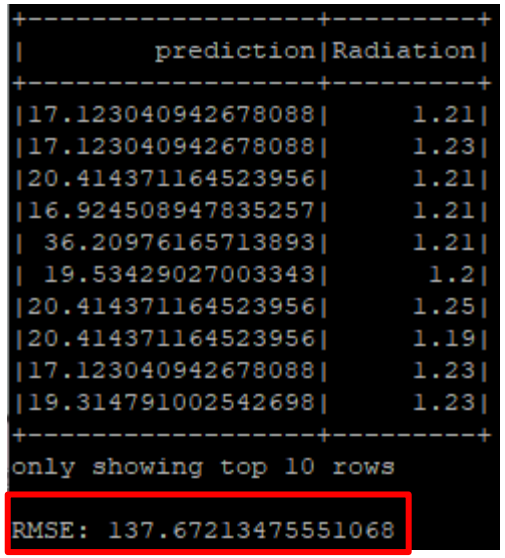
only showing top 10 rows

RMSE: 123.06854006562283

12월

모델링

일사량 변동이 큰 시간대 별 RF

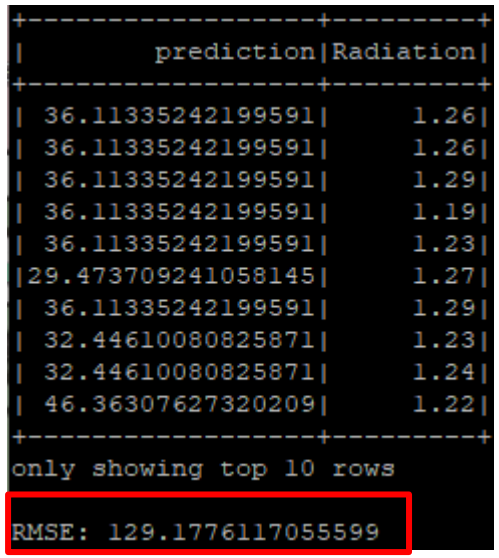


prediction	Radiation
17.123040942678088	1.21
17.123040942678088	1.23
20.414371164523956	1.21
16.924508947835257	1.21
36.20976165713893	1.21
19.53429027003343	1.2
20.414371164523956	1.25
20.414371164523956	1.19
17.123040942678088	1.23
19.314791002542698	1.23

only showing top 10 rows

RMSE: 137.67213475551068

9월

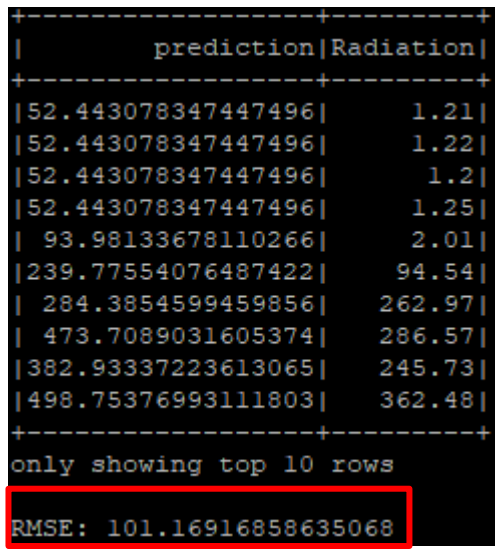


prediction	Radiation
36.11335242199591	1.26
36.11335242199591	1.26
36.11335242199591	1.29
36.11335242199591	1.19
36.11335242199591	1.23
29.473709241058145	1.27
36.11335242199591	1.29
32.44610080825871	1.23
32.44610080825871	1.24
46.36307627320209	1.22

only showing top 10 rows

RMSE: 129.1776117055599

10월

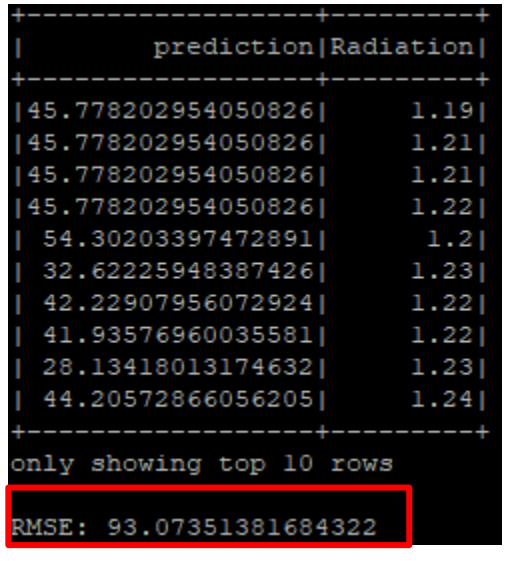


prediction	Radiation
52.443078347447496	1.21
52.443078347447496	1.22
52.443078347447496	1.2
52.443078347447496	1.25
93.98133678110266	2.01
239.77554076487422	94.54
284.3854599459856	262.97
473.7089031605374	286.57
382.93337223613065	245.73
498.75376993111803	362.48

only showing top 10 rows

RMSE: 101.16916858635068

11월



prediction	Radiation
45.778202954050826	1.19
45.778202954050826	1.21
45.778202954050826	1.21
45.778202954050826	1.22
54.30203397472891	1.2
32.62225948387426	1.23
42.22907956072924	1.22
41.93576960035581	1.22
28.13418013174632	1.23
44.20572866056205	1.24

only showing top 10 rows

RMSE: 93.07351381684322

12월

Part 4

결과 분석

Part 4

결과 분석

데이터 분석 결과

- 시간대별, 월별 데이터 분포가 다르게 나타남
- 영향을 주는 column의 특성, 일사량과의 연관성 발견

모델링 결과 분석

- Random Forest 사용시 정확도 높아짐
- 시간대별, 월별 일사량 예측 결과

  전체 데이터로 예측한 결과보다 RMSE 감소

개선 방향

- 기온이 높은 월의 데이터 모델 정확도 향상 필요

감사합니다.

Q & A