#2022 UOS 빅데이터 알고리즘 경진대회

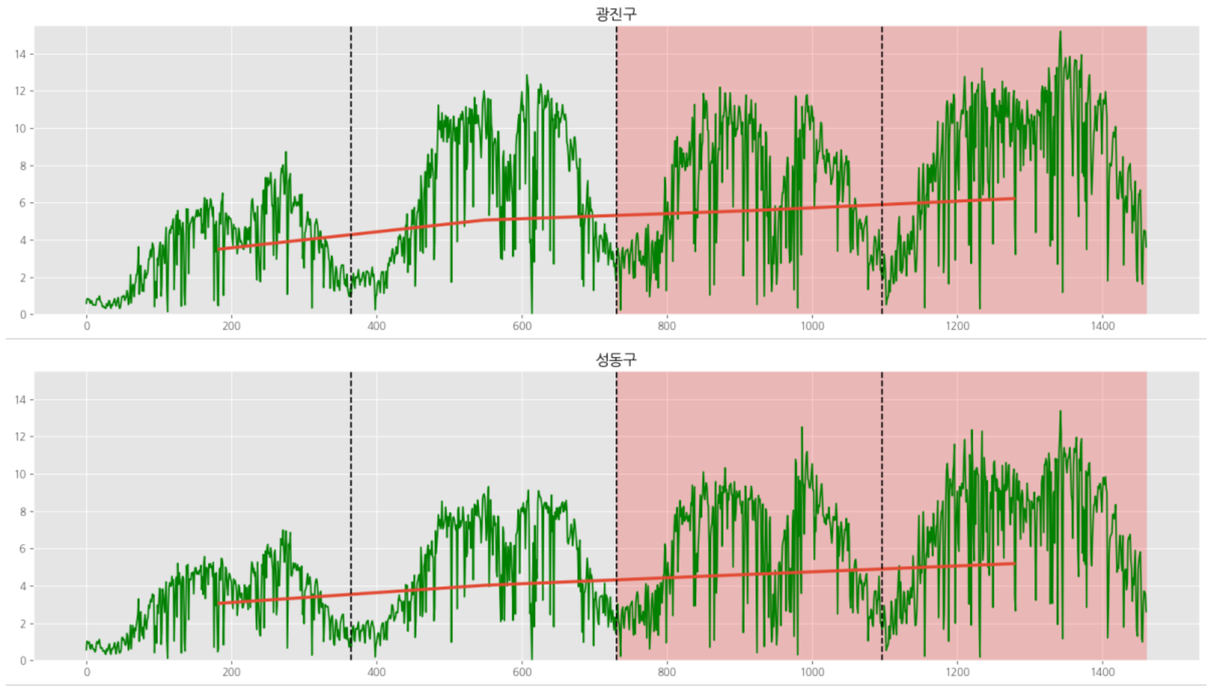
서울시 지역구별
따릉이 대여량 예측 모델 개발

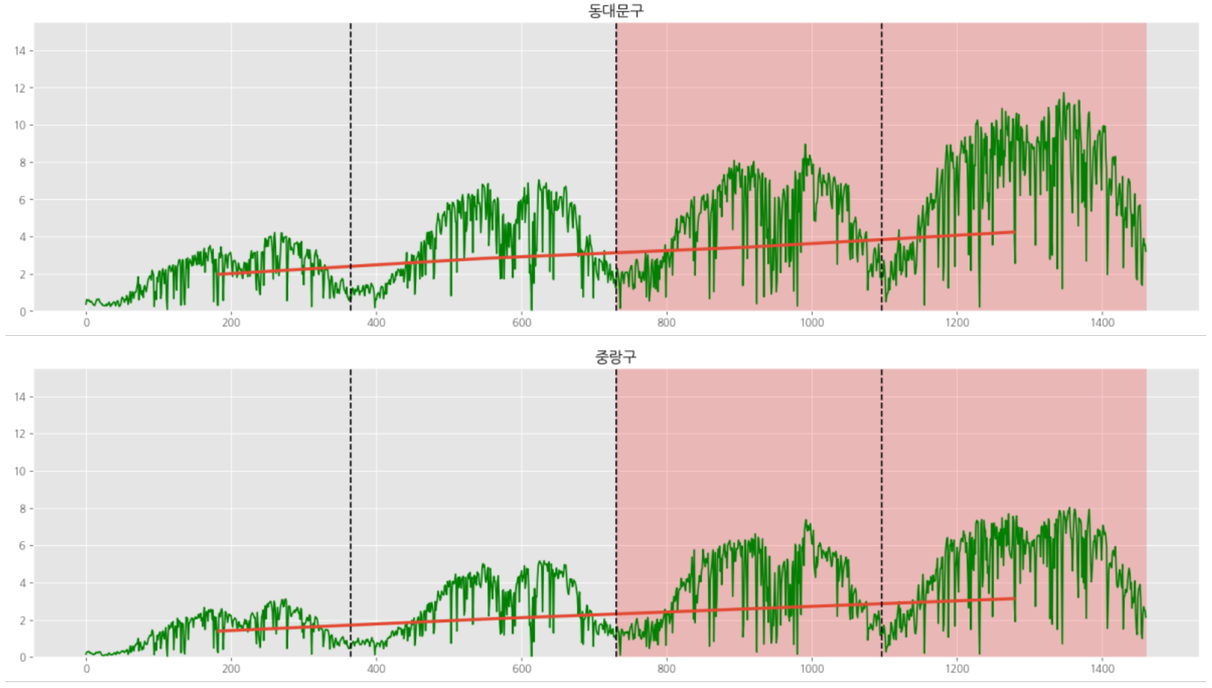
ij._.nim (국민대학교 AI빅데이터융합경영학과)

목차

01 제공 데이터 EDA

구별 따릉이 대여량 추이

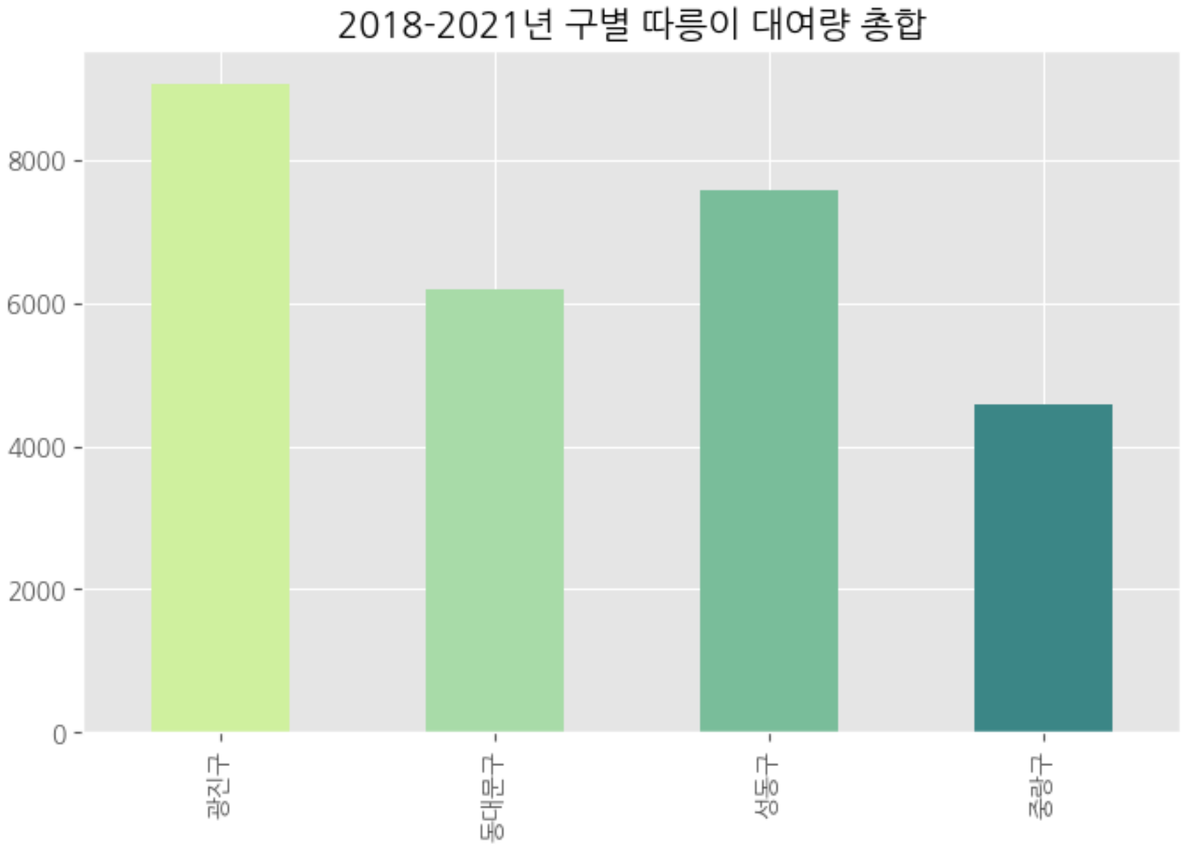




구별 따릉이 대여량이 증가 추세를 보이고 있으며 특정한 주기를 갖는 시계열 데이터임을 확인함.
특히, 코로나가 발생한 2020년 이후 대여량이 크게 증가하는 것을 확인할 수 있음.

01 제공 데이터 EDA

구별 따릉이 대여량 총합



따릉이 대여량 총합은 광진구, 성동구, 동대문구, 중랑구 순으로 많으며
각 구별 대여량 scale의 차이가 있음을 확인함.

02 Feature Engineering

따릉이 대여량에 영향을 미칠 것으로 생각되는 날짜와 날씨 데이터를 활용해 feature engineering 진행

날짜 → datetime으로 변경

- ✔ year, month, day, weekday 생성
- ✔ month, day, weekday에 cyclical encoding을 적용하여 주기성 부여
- ✔ 2020년을 기준으로 코시국 여부 생성
- ✔ 평일/주말 생성

02 Feature Engineering

따릉이 대여량에 영향을 미칠 것으로 생각되는 날짜와 날씨 데이터를 활용해 feature engineering 진행

날씨

외부데이터 mapping : [기상청 기상자료개방포털] 기후통계분석 조건별통계

✔ 평균기온, 최고기온, 최저기온, 일교차, 강수량 feature 생성

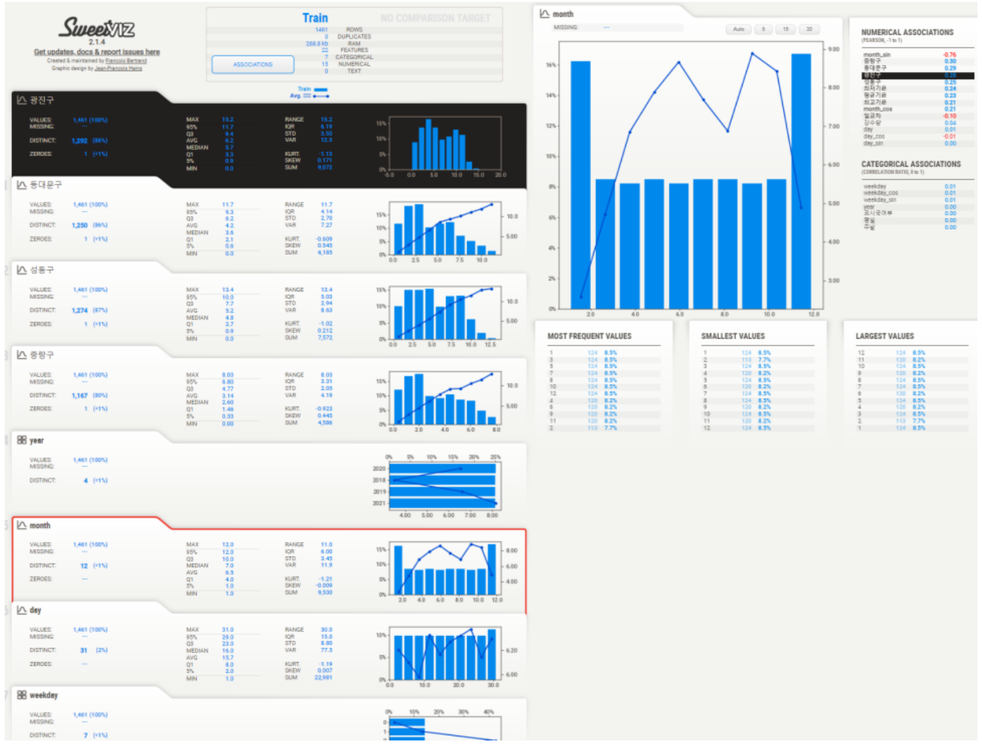
✔ 2022년에 해당하는 test 데이터에는 2000년부터 2017년까지의 통계값을 활용하여 생성

	광진구	동대문구	성동구	중랑구	year	month	day	weekday	month_sin	month_cos	...	weekday_sin	weekday_cos	코시국여부	평일	주말	평균기온	최고기온	최저기온	일교차	강수량
0	0.592	0.368	0.580	0.162	2018	1	1	0	5.000000e-01	0.866025	...	0.000000	1.000000e+00	0	1	0	-1.3	3.8	-5.1	8.9	0.0
1	0.840	0.614	1.034	0.260	2018	1	2	1	5.000000e-01	0.866025	...	0.500000	8.660254e-01	0	1	0	-1.8	1.8	-4.3	6.1	0.0
2	0.828	0.576	0.952	0.288	2018	1	3	2	5.000000e-01	0.866025	...	0.866025	5.000000e-01	0	1	0	-4.7	-0.4	-7.1	6.7	0.0
3	0.792	0.542	0.914	0.292	2018	1	4	3	5.000000e-01	0.866025	...	1.000000	6.123234e-17	0	1	0	-4.7	-0.7	-8.7	8.0	0.0
4	0.818	0.602	0.994	0.308	2018	1	5	4	5.000000e-01	0.866025	...	0.866025	-5.000000e-01	0	1	0	-3.0	1.6	-5.6	7.2	0.0
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
1456	3.830	3.416	2.908	2.350	2021	12	27	0	-2.449294e-16	1.000000	...	0.000000	1.000000e+00	1	1	0	-7.6	-3.9	-12.9	9.0	0.0
1457	4.510	3.890	3.714	2.700	2021	12	28	1	-2.449294e-16	1.000000	...	0.500000	8.660254e-01	1	1	0	-4.1	-0.9	-8.5	7.6	0.0
1458	4.490	3.524	3.660	2.524	2021	12	29	2	-2.449294e-16	1.000000	...	0.866025	5.000000e-01	1	1	0	0.4	5.9	-3.8	9.7	0.2
1459	4.444	3.574	3.530	2.506	2021	12	30	3	-2.449294e-16	1.000000	...	1.000000	6.123234e-17	1	1	0	-3.9	0.2	-6.8	7.0	0.0
1460	3.616	3.210	2.620	2.146	2021	12	31	4	-2.449294e-16	1.000000	...	0.866025	-5.000000e-01	1	1	0	-6.7	-3.9	-8.8	4.9	0.0

1461 rows × 22 columns

03 생성 feature EDA

SWEETVIZ를 활용하여 최종 feature의 분포 및 상관관계 파악



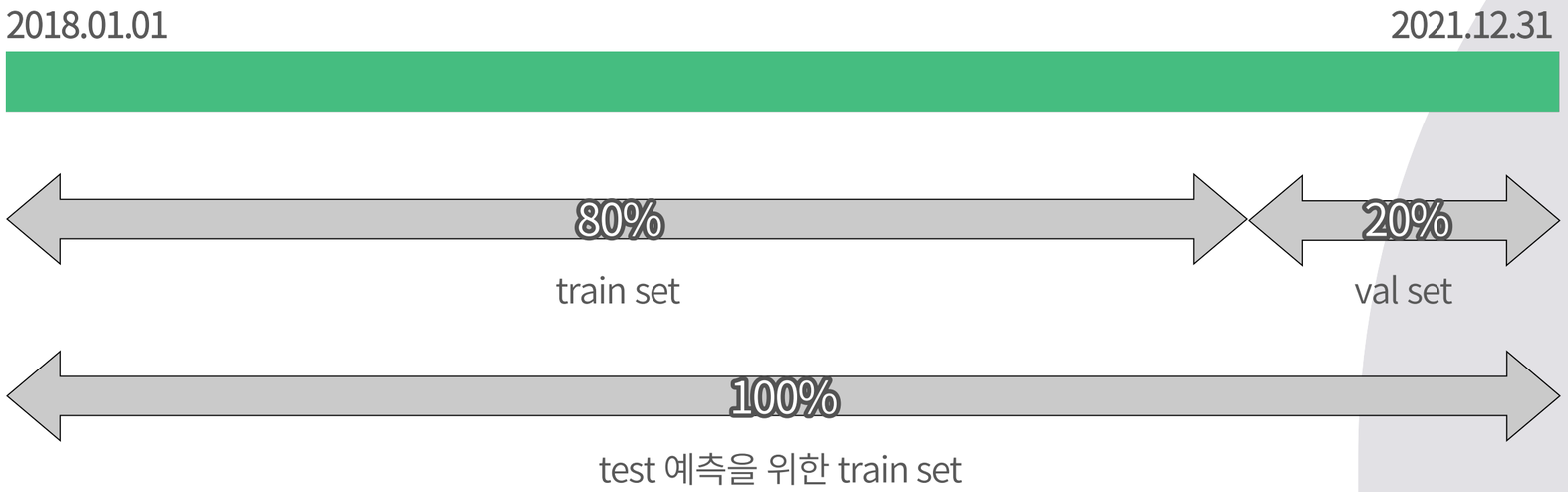
월별 따릉이 대여량을 살펴보면, 장마의 영향을 크게 받아 8월 경에 대여량이 감소함을 확인할 수 있음.
또한, 평균기온이 높을수록 대여량이 높은 양상을 보이나 20도를 넘어가면 다시 감소하는 추세를 확인함.

04 Modeling

모델 학습 및 검증

ExtraTree

✔ Validation set 구축



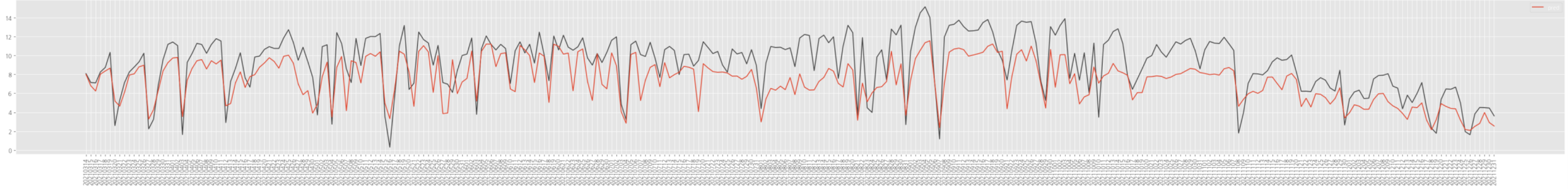
데이터의 양이 많지 않기 때문에, validation set의 비율을 20%로 설정.
시계열 특성을 갖는 데이터이기 때문에, train_test_split에서 shuffle=False로 설정.
test 데이터 예측 시에는 전체 데이터로 재학습 진행.

04 Modeling

모델 학습 및 검증

ExtraTree

✔ Validation set 예측 결과



다소 underestimate하는 경향.

모델이 대여량이 적은 8월에 큰 영향을 받고 underestimate하는 것으로 판단하여, 월별 예측값 보정 진행.

04 Modeling

모델 학습 및 검증

ExtraTree

✔ 예측값 보정

- 겨울에 해당하는 달 '1월' → 예측값 보정 X
- 여름에 해당하는 달 '8월' → 예측값 + 4
- 이외 달 → 예측값 + 2

광진구	2.082923
동대문구	2.247106
성동구	1.547463
중랑구	1.262330

보정 전 MAE

광진구	1.203583
동대문구	1.130758
성동구	1.327704
중랑구	1.215476

보정 후 MAE

04 Modeling

모델 학습 및 검증

ExtraTree

✔ Submission 생성 : 전체 데이터로 재학습한 후 test 데이터를 예측한 뒤 예측값 보정을 거쳐 submission 생성

	일시	광진구	동대문구	성동구	중랑구
0	20220101	2.397068	1.839152	1.952300	1.397532
1	20220102	2.513020	2.018176	1.985996	1.524736
2	20220103	3.075212	2.492724	2.808660	1.668940
3	20220104	2.818048	2.340704	2.566944	1.532928
4	20220105	2.631508	2.174476	2.341000	1.445724
...	...	...	...	...	...
329	20221126	7.459264	6.896708	6.243572	5.495960
330	20221127	7.486600	6.846020	6.296204	5.467764
331	20221128	7.530212	7.126860	6.257124	5.394940
332	20221129	8.204032	7.626084	7.172860	5.842708
333	20221130	7.925444	7.485084	6.616016	5.715720

334 rows × 5 columns

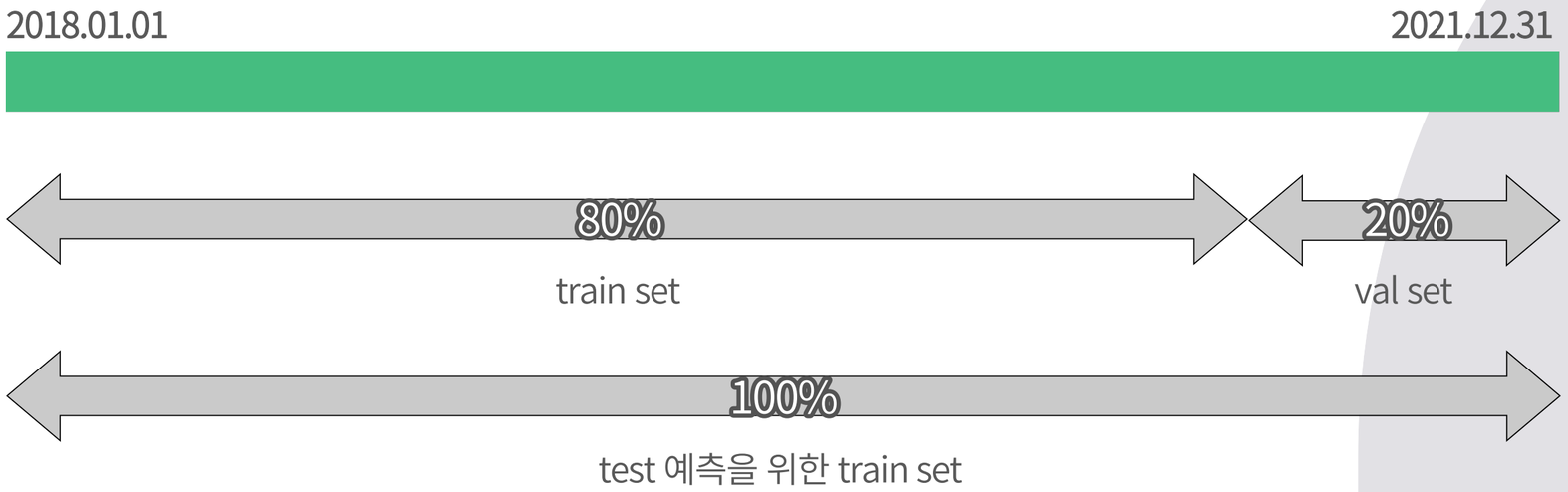
Public Score : 1.8440434751

04 Modeling

모델 학습 및 검증

Prophet

✔ Validation set 구축



데이터의 양이 많지 않기 때문에, validation set의 비율을 20%로 설정.
test 데이터 예측 시에는 전체 데이터로 재학습 진행.

04 Modeling

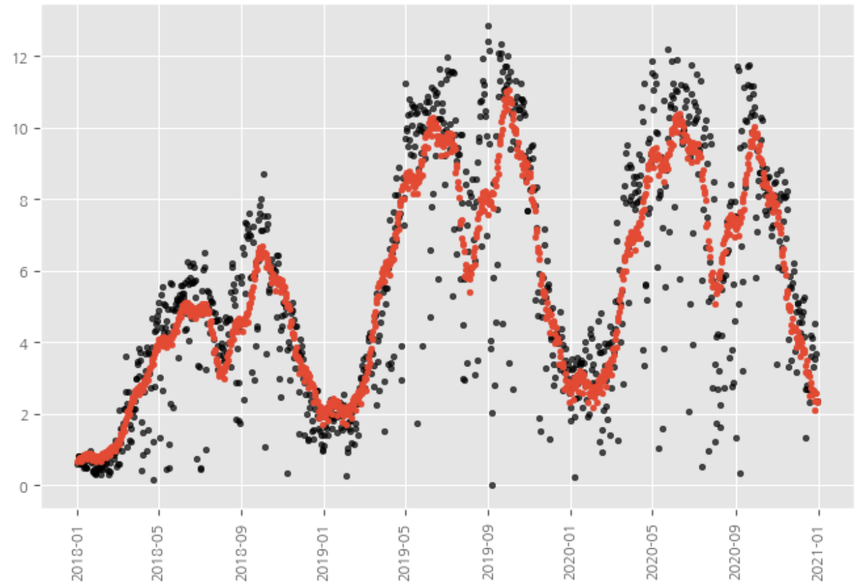
모델 학습 및 검증

Prophet

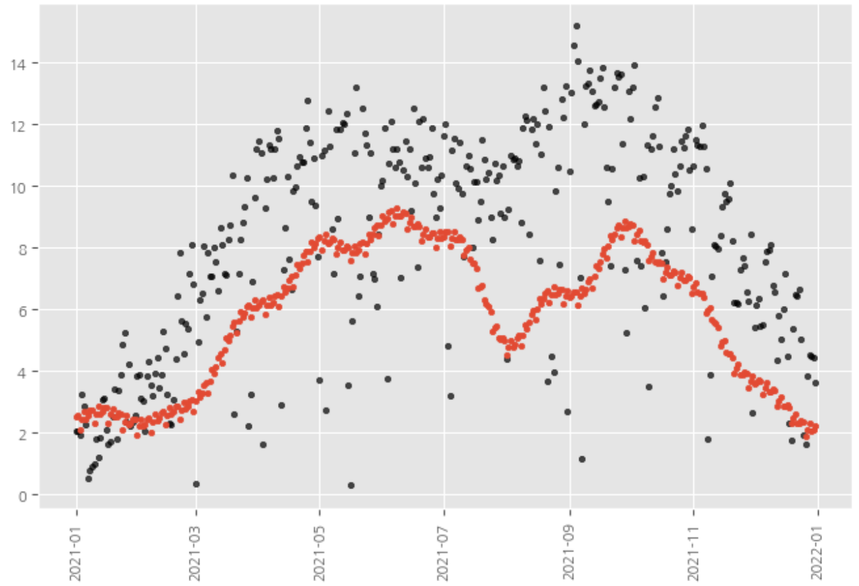
✔ 모델 구축 후 '광진구' 학습 및 예측 테스트

```
# Prophet에 시계열성 부여
m = Prophet(seasonality_mode='multiplicative', yearly_seasonality='auto', weekly_seasonality='auto',
            daily_seasonality='auto', holidays_prior_scale=10.0)
```

2018~2020 학습 결과



2021 예측 결과

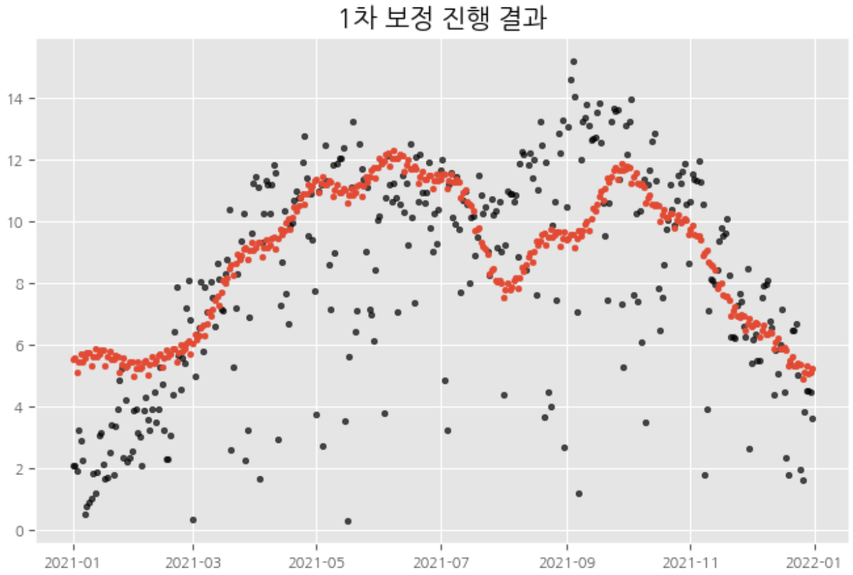


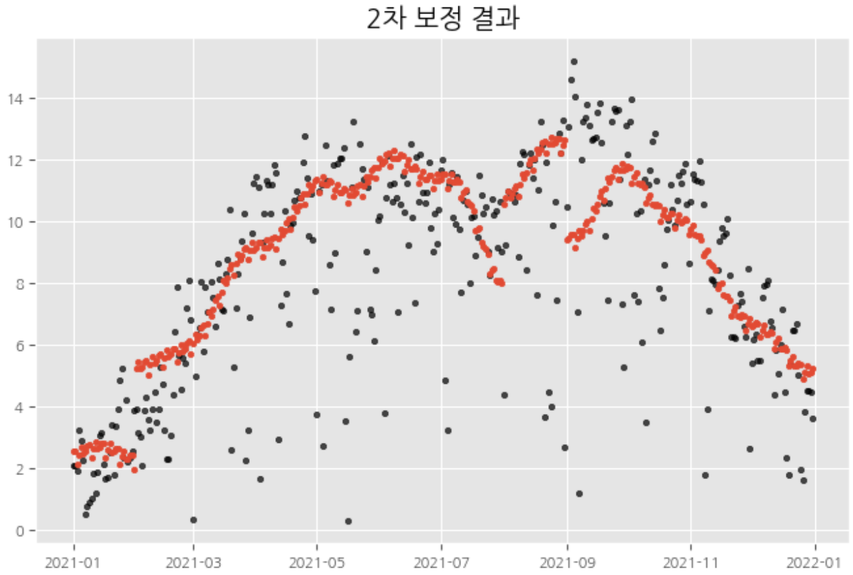
04 Modeling

모델 학습 및 검증

Prophet

✔ 예측값 보정





결과 그래프를 보고 값을 수정하며 최적의 보정값 탐색.

2차 보정 결과, 오차가 많이 줄어들고 적절한 예측값을 뽑는 것을 확인함.

04 Modeling

모델 학습 및 검증

Prophet

✔ 예측값 보정

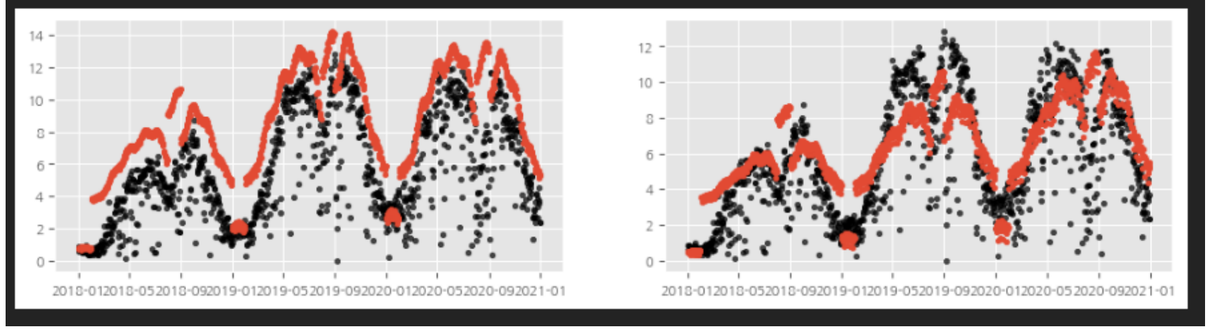
구별 따릉이 대여량의 scale이 다르므로, 보정값의 scale을 달리하기 위해 광진구와 동대문구, 성동구와 중랑구로 나누어 보정 진행.

```
15:16:50 - cmdstanpy - INFO - Chain [1] start processing
15:16:50 - cmdstanpy - INFO - Chain [1] done processing


광진구: MAE 1.8063463914878255


15:16:51 - cmdstanpy - INFO - Chain [1] start processing
15:16:51 - cmdstanpy - INFO - Chain [1] done processing


동대문구: MAE 2.0277179262550713
```



[광진구/동대문구]

- 겨울에 해당하는 달 '1월' → 예측값 보정 X
- 여름에 해당하는 달 '8월' → 예측값 + 6
- 이외 달 → 예측값 + 3

04 Modeling

모델 학습 및 검증

Prophet

✔ 예측값 보정

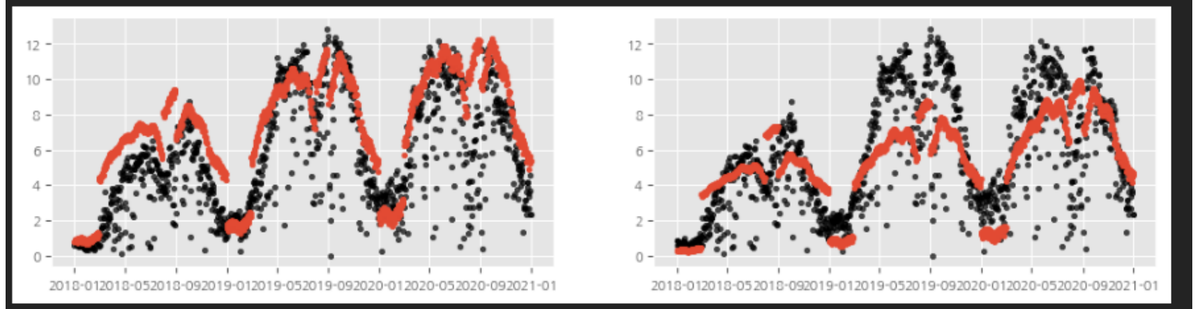
구별 따릉이 대여량의 scale이 다르므로, 보정값의 scale을 달리하기 위해 광진구와 동대문구, 성동구와 중랑구로 나누어 보정 진행.

```
15:16:52 - cmdstanpy - INFO - Chain [1] start processing
15:16:52 - cmdstanpy - INFO - Chain [1] done processing


성동구: MAE 2.408873750406698


15:16:53 - cmdstanpy - INFO - Chain [1] start processing
15:16:54 - cmdstanpy - INFO - Chain [1] done processing


중랑구: MAE 2.792291774161852
```



[성동구/중랑구]

- 겨울에 해당하는 달 '1월','2월' → 예측값 보정 X
- 여름에 해당하는 달 '8월' → 예측값 + 5.5
- 이외 달 → 예측값 + 3

04 Modeling

모델 학습 및 검증

Prophet

✔ **Submission 생성 :** 전체 데이터로 재학습한 후 test 데이터를 예측한 뒤 예측값 보정을 거쳐 submission 생성

	일시	광진구	동대문구	성동구	중랑구
0	20220101	2.235738	2.194571	2.388499	2.055112
1	20220102	1.838252	1.589996	2.044017	1.497861
2	20220103	2.105466	2.601973	2.599862	1.899331
3	20220104	2.331849	3.039997	2.888089	2.316883
4	20220105	2.131580	2.660507	2.489096	2.056484
...	...	...	...	...	...
329	20221126	6.457441	7.771698	7.403075	7.226829
330	20221127	6.051856	7.001936	6.958452	6.483654
331	20221128	6.232266	8.069549	7.471391	6.868727
332	20221129	6.378125	8.487887	7.703525	7.276792
333	20221130	6.150093	7.980419	7.209836	6.894246

334 rows × 5 columns

Public Score : 1.9791359471

04 Modeling

모델 학습 및 검증

Ensemble

✔ **Ensemble** : 더 좋은 성능을 낸 ExtraTree에 가중치를 크게 두고 submission ensemble 진행

ExtraTree 예측값

	일시	광진구	동대문구	성동구	중랑구
0	20220101	2.397068	1.839152	1.952300	1.397532
1	20220102	2.513020	2.018176	1.985996	1.524736
2	20220103	3.075212	2.492724	2.808660	1.668940
3	20220104	2.818048	2.340704	2.566944	1.532928
4	20220105	2.631508	2.174476	2.341000	1.445724
...	...	...	...	...	...
329	20221126	7.459264	6.896708	6.243572	5.495960
330	20221127	7.486600	6.846020	6.296204	5.467764
331	20221128	7.530212	7.126860	6.257124	5.394940
332	20221129	8.204032	7.626084	7.172860	5.842708
333	20221130	7.925444	7.485084	6.616016	5.715720

334 rows × 5 columns

60%

+

Prophet 예측값

	일시	광진구	동대문구	성동구	중랑구
0	20220101	2.235738	2.194571	2.388499	2.055112
1	20220102	1.838252	1.589996	2.044017	1.497861
2	20220103	2.105466	2.601973	2.599862	1.899331
3	20220104	2.331849	3.039997	2.888089	2.316883
4	20220105	2.131580	2.660507	2.489096	2.056484
...	...	...	...	...	...
329	20221126	6.457441	7.771698	7.403075	7.226829
330	20221127	6.051856	7.001936	6.958452	6.483654
331	20221128	6.232266	8.069549	7.471391	6.868727
332	20221129	6.378125	8.487887	7.703525	7.276792
333	20221130	6.150093	7.980419	7.209836	6.894246

334 rows × 5 columns

40%

최종 submission

	일시	광진구	동대문구	성동구	중랑구
0	20220101.0	2.332536	1.981320	2.126780	1.660564
1	20220102.0	2.243113	1.846904	2.009204	1.513986
2	20220103.0	2.687314	2.536424	2.725141	1.761096
3	20220104.0	2.623569	2.620421	2.695402	1.846510
4	20220105.0	2.431537	2.368888	2.400239	1.690028
...	...	...	...	...	...
329	20221126.0	7.058535	7.246704	6.707373	6.188308
330	20221127.0	6.912702	6.908386	6.561103	5.874120
331	20221128.0	7.011033	7.503935	6.742831	5.984455
332	20221129.0	7.473669	7.970805	7.385126	6.416341
333	20221130.0	7.215304	7.683218	6.853544	6.187130

334 rows × 5 columns

Public Score : 1.759994145

Private Score : 2.17002

감사합니다