| 따릉이 대여 수 예측

비어플 프로젝트 6조

김기호 배정민 이가영

2021.10.23

목 차

따릉이 대여 수 예측

분석목적

- 서울시 따릉이의 이용률이 증가함에 따라 서울시 대응책이 필요해짐
- 따릉이 이용현황에 대한 데이터를 분석하여 서울시민들의 편의성 증대 목적
- 2017년 4~5월 마포구 따릉이 정보를 통해 6월의 따릉이 대여 수를 예측

따릉이 대여 수 예측

데이터 소개

변수 소개

Train/Test Data

- Id : 고유 id
- Hour : 시간
- Hour_bef_temperature : 1시간 전 기온
- Hour_bef_precipitation : 1시간 전 비가 오지 않았으면 0, 비가 왔으면 1
- Hour_bef_windspeed : 1시간 전 평균 풍속
- Hour_bef_humidity : 1시간 전 습도
- Hour_bef_visibility : 1시간 전 가시성
- Hour_bef_ozone : 1시간 전 오존
- Hour_bef_pm10 : 1시간 전 미세먼지(pm10)
- Hour_bef_pm2.5 : 1시간 전 미세먼지(pm2.5)
- Count : 따릉이 대여 수 (목적변수)

기상청 외부 데이터

- 지점 : 지점 번호
- 지점명 : 지점 도사 아름
- 일시 : 연도-월-일 시간
- 기온 : 해당 일시의 기온
- 강수량: 해당 일시의 강수량
- 풍속 : 해당 일시의 풍속(평균)
- 습도 : 해당 일시의 습도
- 시정: 해당 일시의 시정

따릉이 대여 수 예측

데이터 소개

Train Data

id	hour	hour_bef_ temperature	hour_bef_ precipitation	hour_bef_ windspeed	hour_bef_ humidity	hour_bef_ visibility	hour_bef_ ozone	hour_bef_ pm10	hour_bef_ pm2.5	count
3	20	16.3	1.0	1.5	89.0	576.0	0.027	76.0	33.0	49.0
6	13	20.1	0.0	1.4	48.0	916.0	0.042	73.0	40.0	159.0
…	…	…	…	…	…	…	…	…	…	…
2178	21	20.7	0.0	3.7	37.0	1395.0	0.082	71.0	36.0	216.0
2179	17	21.1	0.0	3.1	47.0	1973.0	0.046	38.0	17.0	170.0

1459개의 행과 11개 변수

따릉이 대여 수 예측

데이터 소개

Test Data

id	hour	hour_bef_ temperature	hour_bef_ precipitation	hour_bef_ windspeed	hour_bef_ humidity	hour_bef_ visibility	hour_bef_ ozone	hour_bef_ pm10	hour_bef_ pm2.5
0	7	20.7	0.0	1.3	62.0	954.0	0.041	44.0	27.0
1	17	30.0	0.0	5.4	33.0	1590.0	0.061	49.0	36.0
…	…	…	…	…	…	…	…	…	…
2166	16	27.0	0.0	1.6	46.0	1956.0	0.032	40.0	26.0
2177	8	22.3	0.0	1.0	63.0	1277.0	0.007	30.0	24.0

715개의 행과 10개 변수

따릉이 대여 수 예측

데이터 소개

기상청 외부 데이터

지점	지점명	일시	기온	강수량	풍속	습도	시정
108	서울	2017-03-31 23:00	5.3	NA	2.2	79	2000
108	서울	2017-04-01 0:00	4.9	NA	1.5	81	2000
…	…	…	…	…	…	…	…
108	서울	2017-06-30 22:00	24.6	NA	1.9	70	701
108	서울	2017-06-30 23:00	24.2	NA	1.9	70	670

2185개의 행과 8개 변수

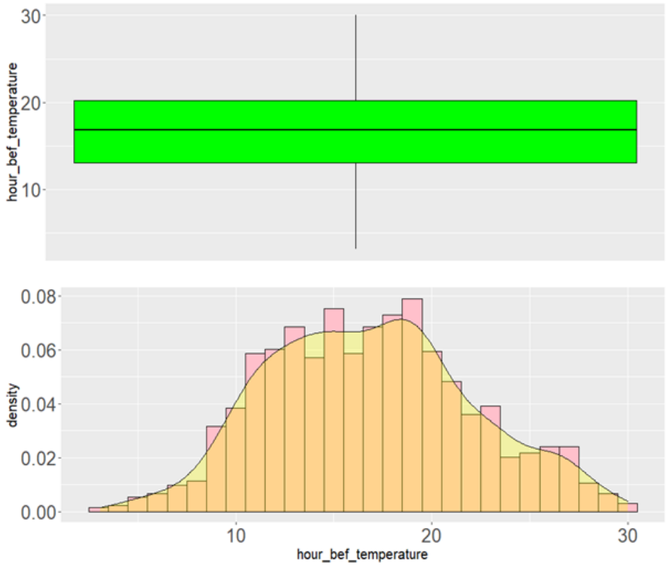
따릉이 대여 수 예측

데이터 전처리

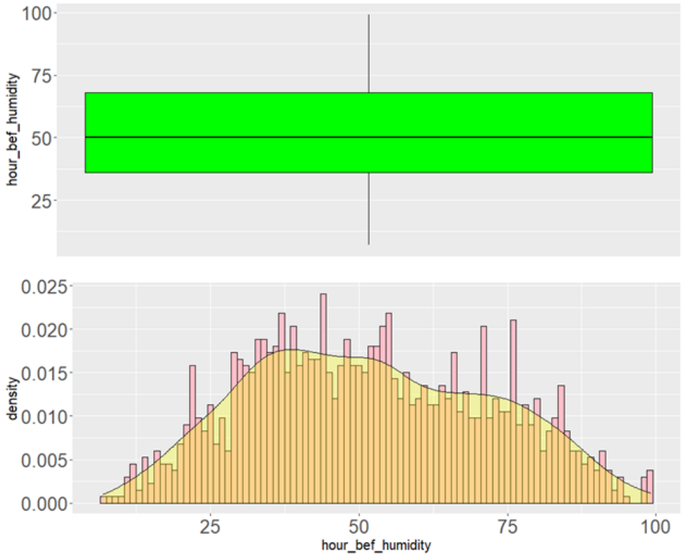
단일 변수 시각화

기온과 습도는 종형 분포이며, 기온의 경우 정규분포의 형태와 유사하다.

기온



습도

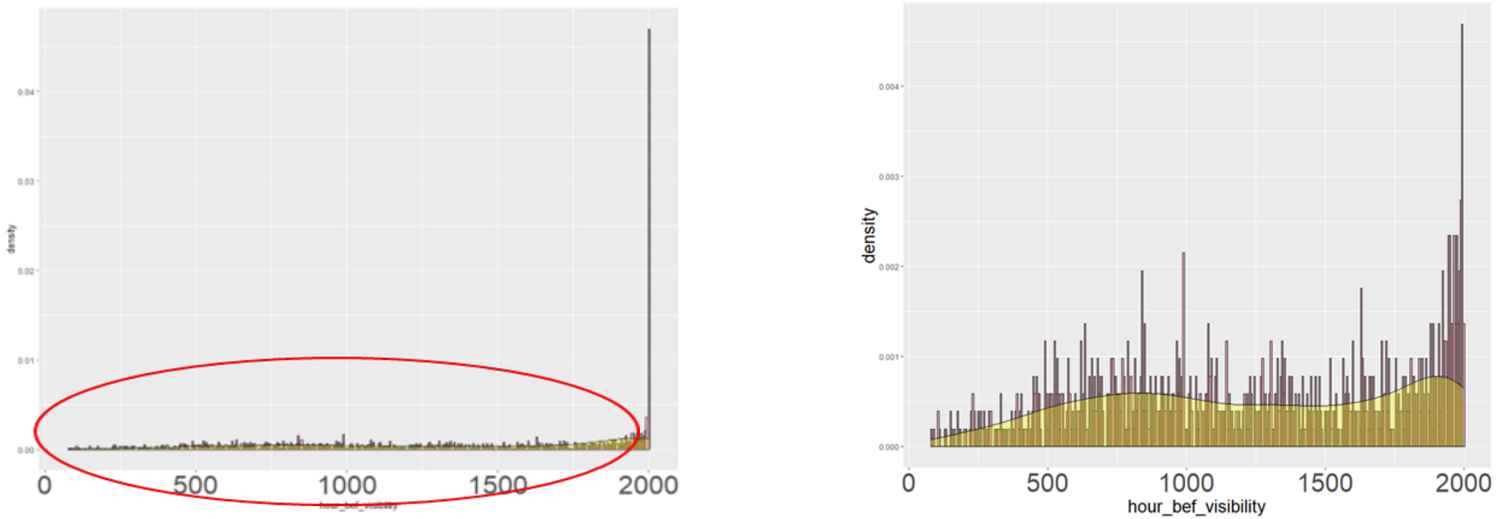


따릉이 대여 수 예측

데이터 전처리

단일 변수 시각화

시정은 많은 값이 2000에 몰려 있다.



시정

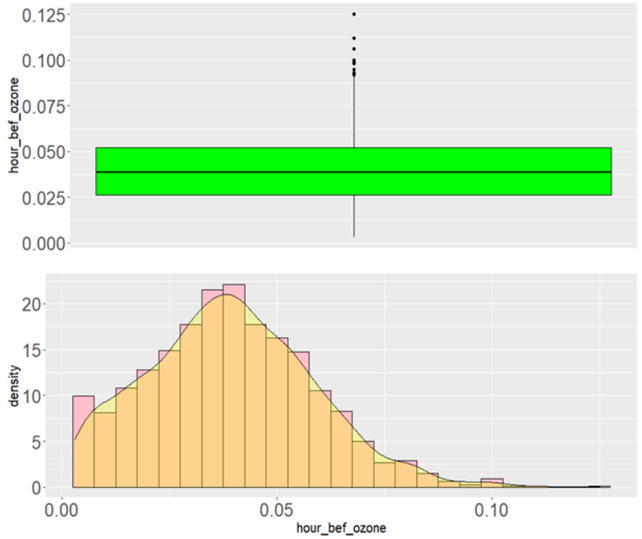
따릉이 대여 수 예측

데이터 전처리

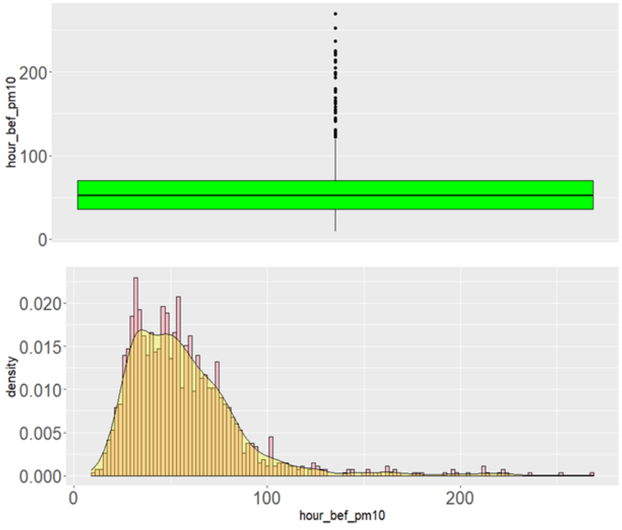
단일 변수 시각화

세 변수 모두 오른쪽으로 꼬리가 긴 형태의 분포를 나타낸다.

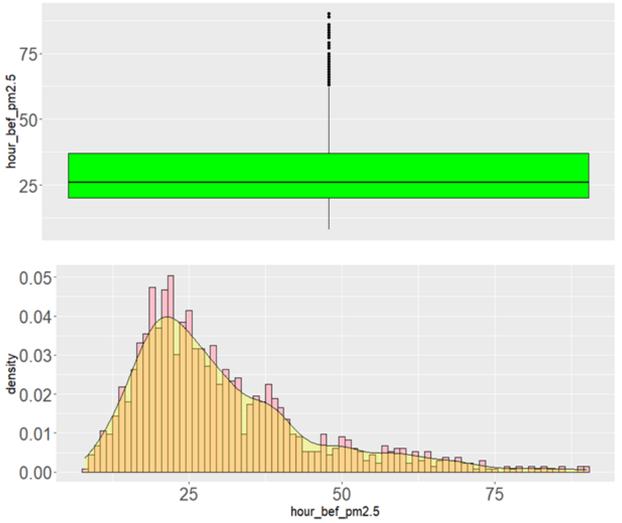
오존



pm10



pm2.5

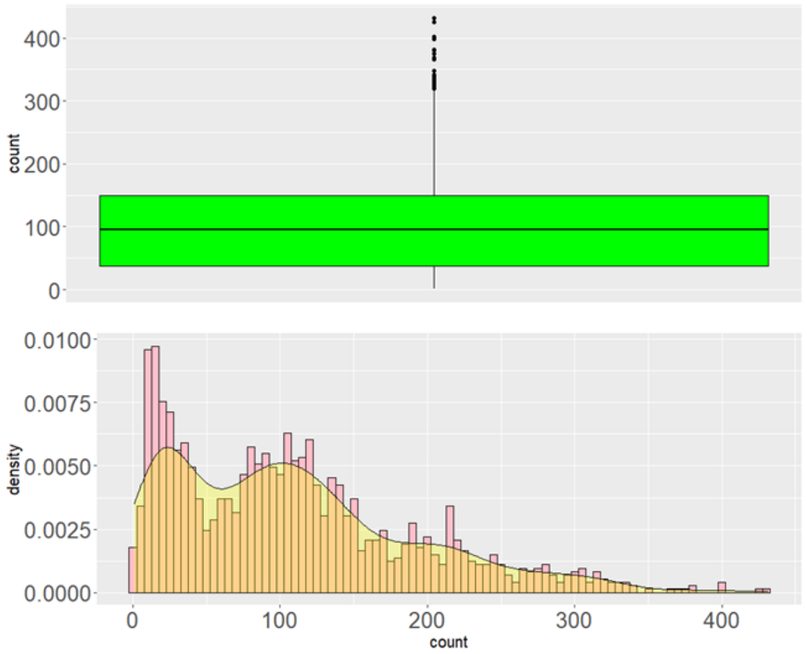


따릉이 대여 수 예측

데이터 전처리

단일 변수 시각화

따릉이 대여 수



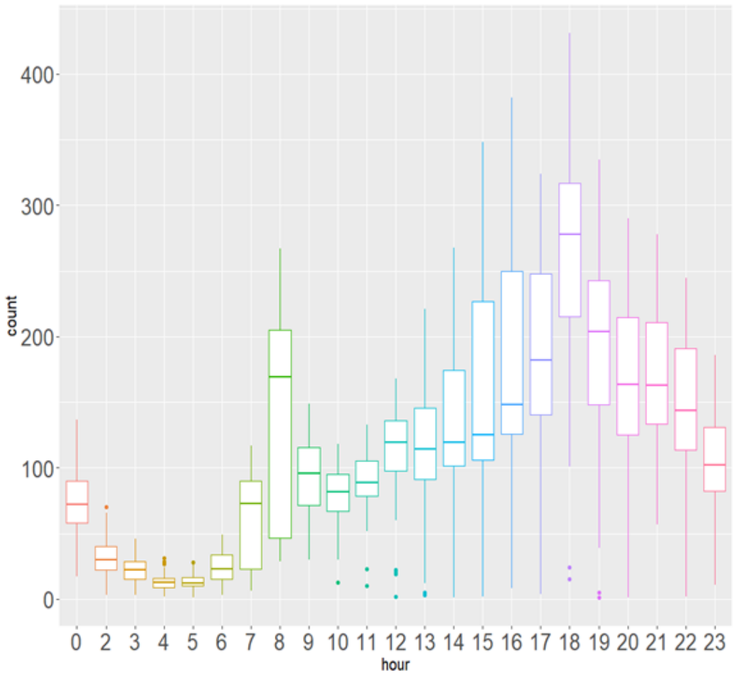
따릉이 대여 수의 단일 분포는 쌍봉 형태를 띈다.
오른 쪽으로 꼬리가 긴 형태의 분포이다.

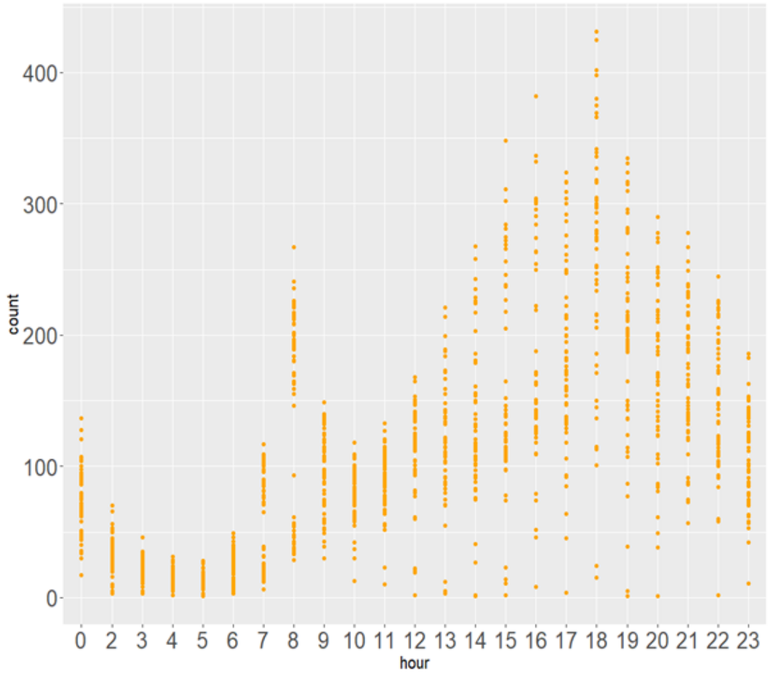
따릉이 대여 수 예측

데이터 전처리

다중 변수 시각화

따릉이 대여 수





시간에 따른 따릉이 대여 수 분포를 보면
8시와 18시를 기준으로
대여 수 증감이 두드러진다.

따릉이 대여 수 예측

데이터 전처리

문제상황-결측치

Test Data

id	hour	hour_bef_temperature	hour_bef_precipitation	hour_bef_windspeed	hour_bef_humidity	hour_bef_visibility	hour_bef_ozone	hour_bef_pm10	hour_bef_pm2.5
1943	19	NA	NA	NA	NA	NA	NA	NA	NA

Test Data 행 중에서 날씨와 기상상황이 누락된 행이 존재

정보가 누락된 Train Data를 제거하지 않고
활용해야 한다고 판단

기상청 외부 데이터로 대응 시켜
누락된 학습용 데이터 결측치 대체

따릉이 대여 수 예측

데이터 전처리

문제상황-결측치

기상청 외부 데이터로 대응 시킨 후

id	hour.x	hour.y	date	day	hour_bef_temperature	hour_bef_windspeed	hour_bef_humidity	hour_bef_visibility	hour_bef_precipitation.x	hour_bef_precipitation.y	hour_bef_ozone	hour_bef_pm10	hour_bef_pm2.5	count	일시
1420	0	99	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	39	NA
1553	18	99	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	1	NA
…	…	…	…	…	…	…	…	…	…	…	…	…	…	…	…
1451	17	17	2017-05-30	화요일	29.8	3.9	12	2000	0	NA	0.058	40	11	215	2017-05-30 17:00
983	16	16	2017-05-03	수요일	30	3.2	16	1183	0	NA	0.1	70	38	304	2017-05-03 16:00

대응이 되지 않은 시간대 일부 존재

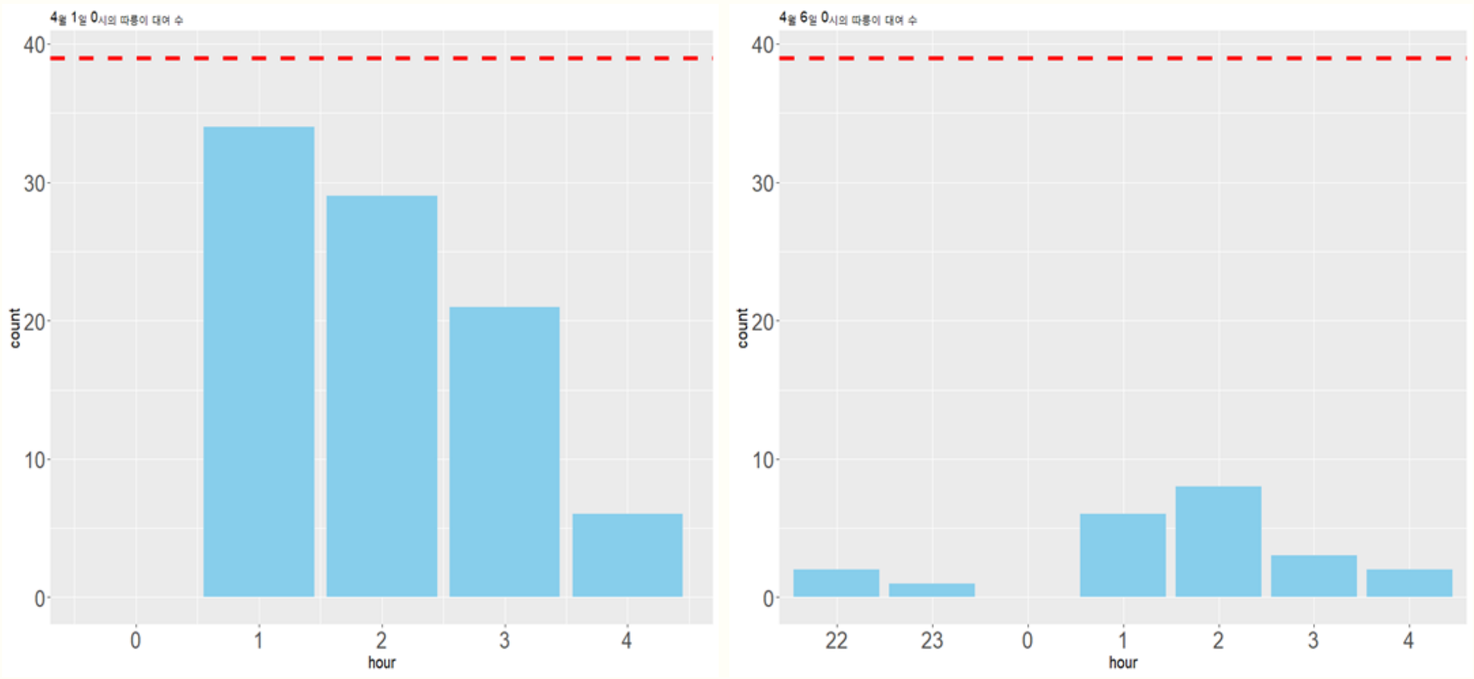
기온, 습도, 시정, 풍속 변수 값의 누락으로
대응되지 않은 시간대는 0시와 18시임

따릉이 대여 수 예측

데이터 전처리

문제해결

0시가 존재하지 않는 4월 1일과 4월 6일의 따릉이 대여 수



전후 시간대와 함께 비교한 결과,
누락된 0시는
4월 1일의 0시임을 유추할 수 있다

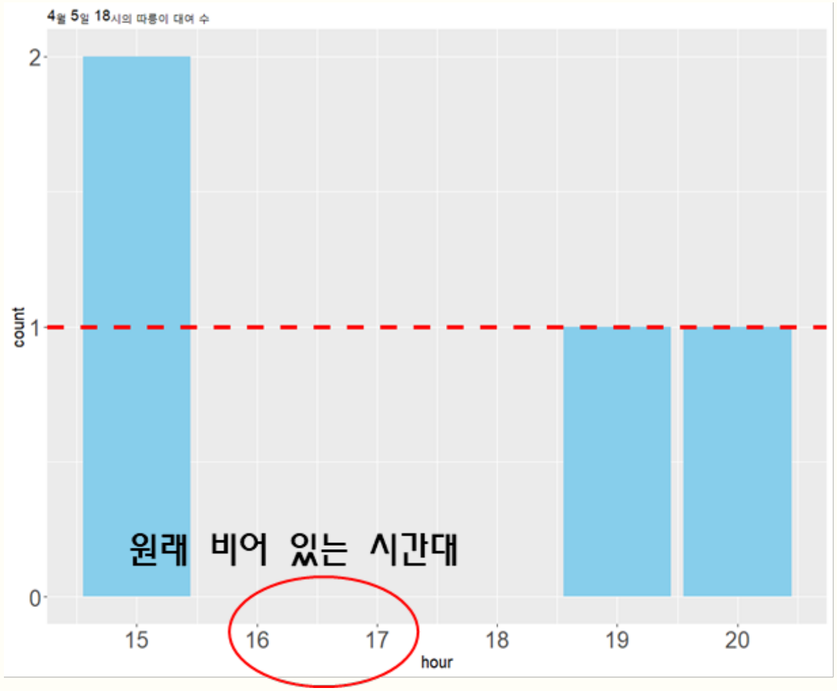
따라서, 직접 대체

따릉이 대여 수 예측

데이터 전처리

문제해결

18시가 존재하지 않는 4월 5일의 따릉이 대여 수



전후 시간대와 함께 비교한 결과,
누락된 18시 역시
4월 5일의 18시임을 유추할 수 있다

따라서, 직접 대체

따릉이 대여 수 예측

데이터 전처리

문제상황-결측치

특정 변수에 결측치 다수 존재

결측변수와 상관계수가 가장 높은 변수로 결측변수 대체

이 때, 그룹별 통계량 값을 이용하여 결측치 대체

단변량 상관계수

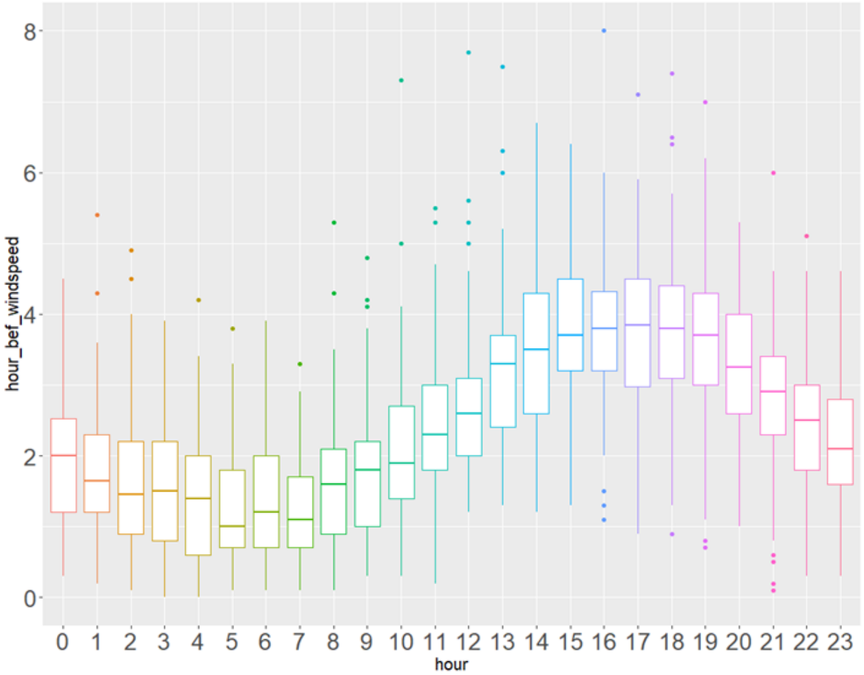
- hour_bef_windspeed와 hour의 상관계수 : 0.46
- hour_bef_ozone와 hour_bef_temperature의 상관계수 : 0.53
- hour_bef_pm2.5와 hour_bef_visibility의 상관계수 : -0.63
- hour_bef_pm10와 hour_bef_pm2.5의 상관계수 : 0.53

따릉이 대여 수 예측

데이터 전처리

문제해결

시간에 대한 풍속 분포



hour_bef_windspeed와 hour의 상관계수 : 0.46
hour 그룹 별 hour_bef_windspeed의 중앙값으로 대체

hour		hour_bef_windspeed
그룹 번호	구간	중앙값
1	0	2
2	1	1.65
3	2	1.45
4	3	1.50
…	…	…
21	20	3.25
22	21	2.90
23	22	2.50
24	23	2.10

따릉이 대여 수 예측

데이터 전처리

문제해결

온도에 대한 오존 분포

hour_bef_ozone와 hour_bef_temperature의 상관계수 : 0.53
hour_bef_temperature그룹 별 hour_bef_ozone의 중앙값으로 대체

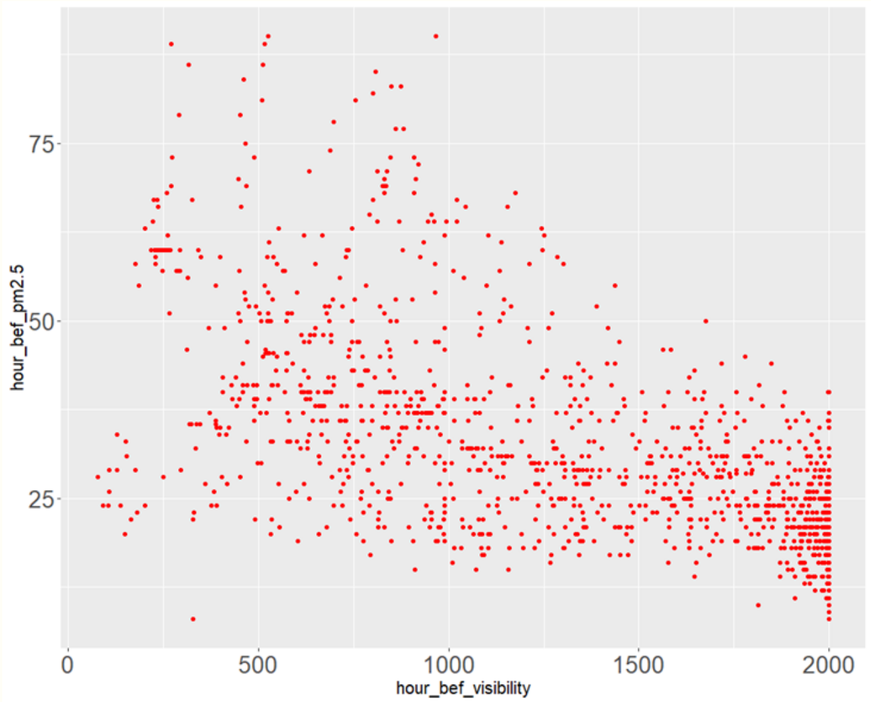
hour_bef_temperature		hour_bef_ozone
그룹 번호	구간	중앙값
1	10 이하	0.0235
2	10 초과 15 이하	0.0340
3	15 초과 20 이하	0.0380
4	20 초과 25 이하	0.0510
5	25 초과	0.0650

따릉이 대여 수 예측

데이터 전처리

문제해결

시정에 대한 pm2.5 분포



hour_bef_pm2.5와 hour_bef_visibility의 상관계수 : -0.63
hour_bef_visibility그룹 별 hour_bef_windspeed의 중앙값으로 대체

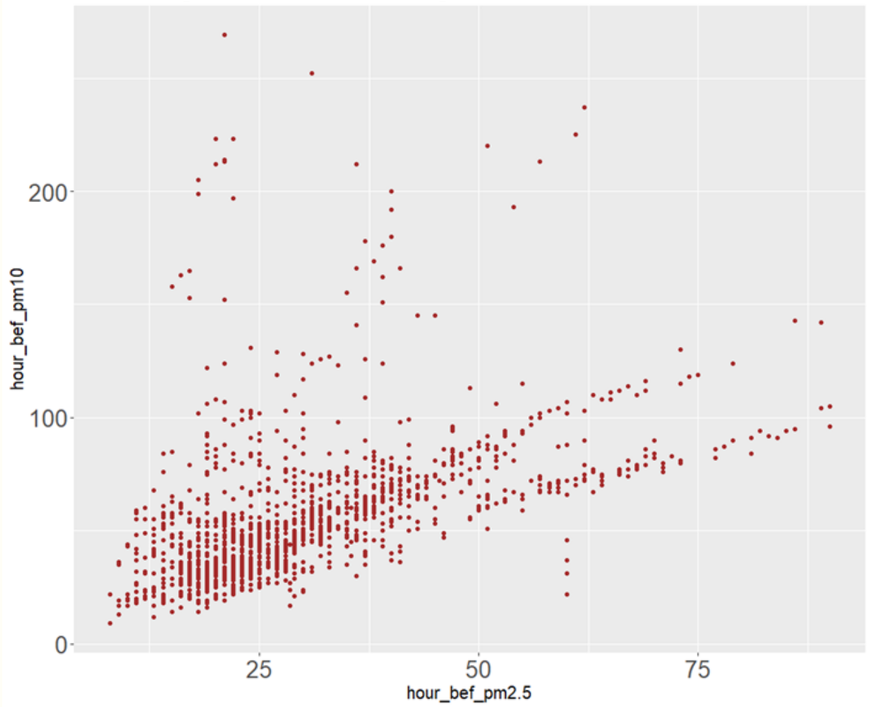
hour_bef_visibility		hour_bef_pm2.5
그룹 번호	구간	중앙값
1	100 미만	26
2	100 초과 200 이하	29
3	200 초과 300 이하	60
4	300 초과 400 이하	35.5
…	…	…
18	1700 초과 1800 이하	28.5
19	1800 초과 1900 이하	24
20	1900 초과 2000 미만	21
21	2000	19

따릉이 대여 수 예측

데이터 전처리

문제해결

pm10에 대한 pm2.5의 분포



hour_bef_pm10와 hour_bef_pm2.5의 상관계수 : 0.53
hour_bef_pm2.5그룹 별 hour_bef_pm10의 중앙값으로 대체

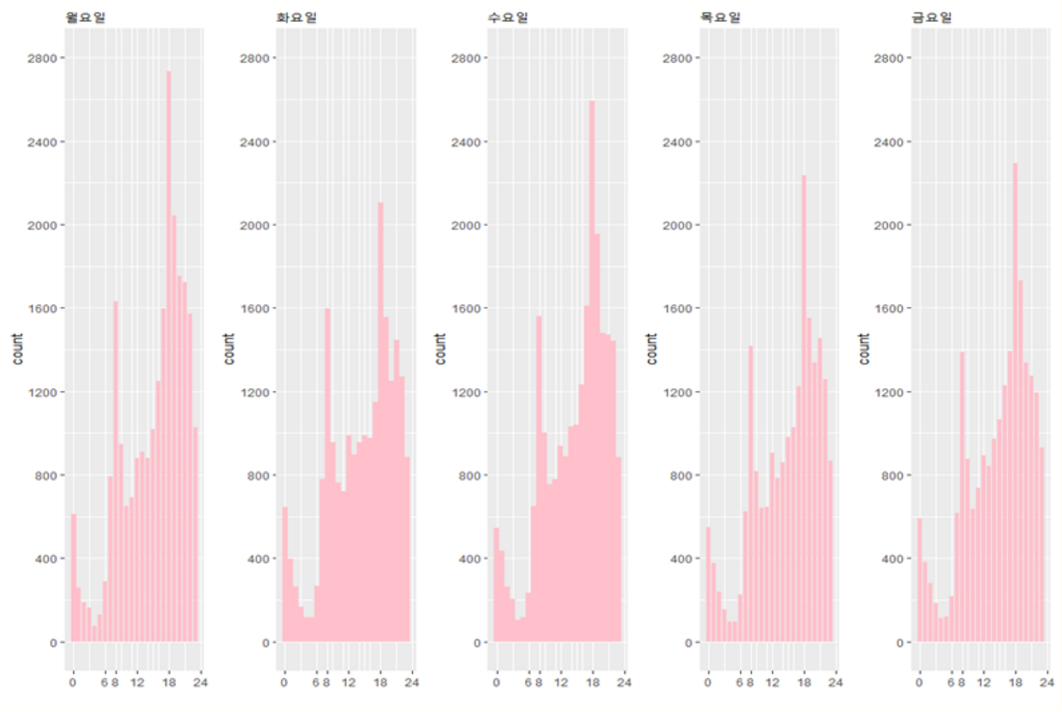
hour_bef_pm2.5		hour_bef_pm10
그룹 번호	구간	중앙값
1	10 미만	20
2	10 초과 20 이하	35
3	20 초과 30 이하	44
4	30 초과 40 이하	60
…	…	72
18	60 초과 70 이하	78.5
19	70 초과 80 이하	86
20	80 초과	94

따릉이 대여 수 예측

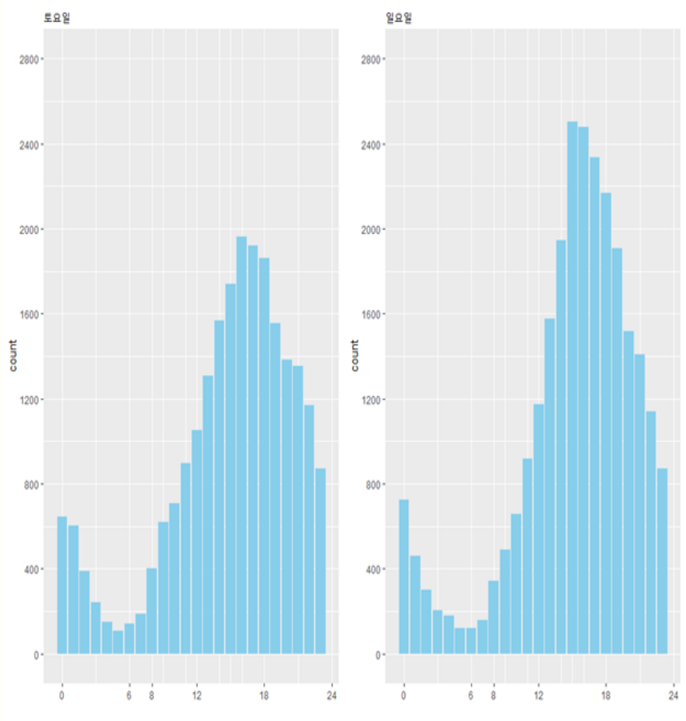
데이터 전처리

파생변수 생성

평일



주말



외부데이터 반영으로
요일을 특정이 가능해짐

평일/주말 여부에 따라 특징이 존재

평일은 출퇴근 시간에 급증
주말은 오후 시간대에 증가했다가 감소

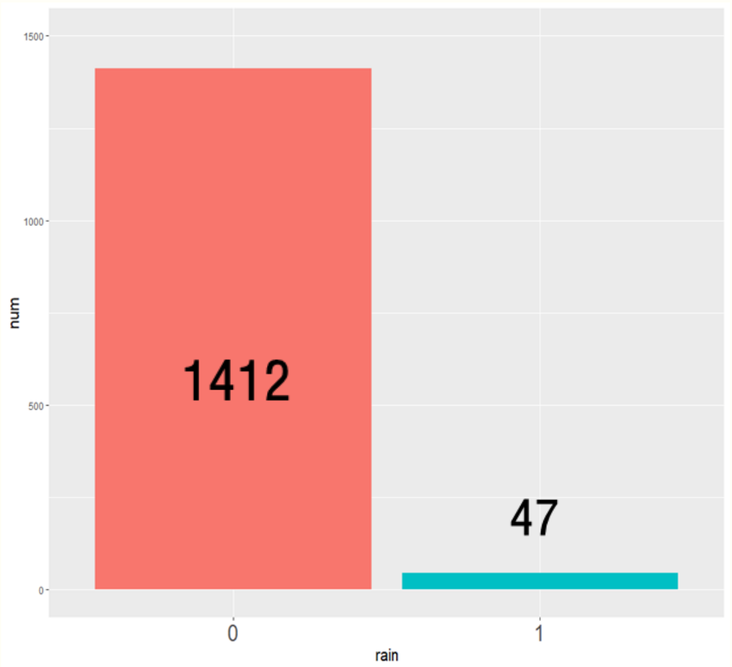
파생변수 weekend 생성

따릉이 대여 수 예측

데이터 전처리

파생변수 생성

기존의 강수 여부 변수



기상청 외부데이터로 대응시킨 결과
이진 변수 형태가 아닌 수치형 변수 형태로 강수 정보가 존재함

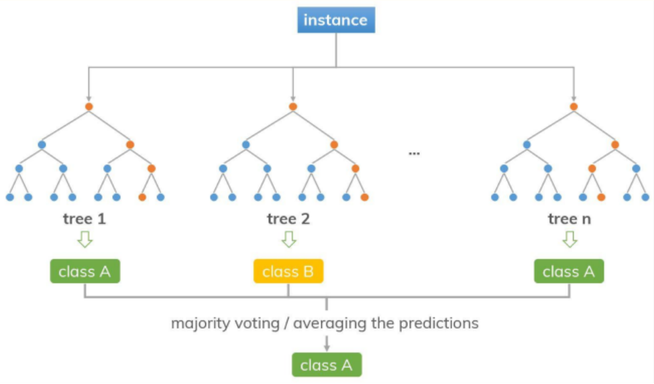
더 많은 정보를 담고 있는 수치형 변수를 대신 활용

따릉이 대여 수 예측

모델링

사용한 알고리즘

Random Forest



- 앙상블 기법의 대표 모델
- 결정트리와 bagging 결합 모델
- 과적합을 방지하고 안정성 높음
- 스케일에 구애받지 않음
- 다중공선성 영향이 적음

XGBOOST

- 경사 부스팅 알고리즘을 기반으로 하여 강력한 예측 성능을 제공
- 대용량 데이터셋에 대해서도 빠르고 효율적인 학습이 가능
- 모델의 복잡도를 제어하기 위해 정규화 기법과 조기 종료(early stopping) 기능을 제공하여 과적합 방지

따릉이 대여 수 예측

모델링

사용한 알고리즘

SVM

- SVM은 주어진 데이터를 가장 잘 분리하는 경계를 찾는 것을 목표로 함
- 과적합에 대한 저항력이 강함
- 이상치에 대한 영향력이 작으며, 데이터 분포에 크게 의존하지 않음
- 고차원 데이터를 처리할 수 있고,
  커널(kernel) 함수를 통해 다양한 형태의 비선형 함수를 사용가능
- SVM은 볼록 최적화(convex optimization) 문제로 정식화되어 있어 결과가 최적값에 근접

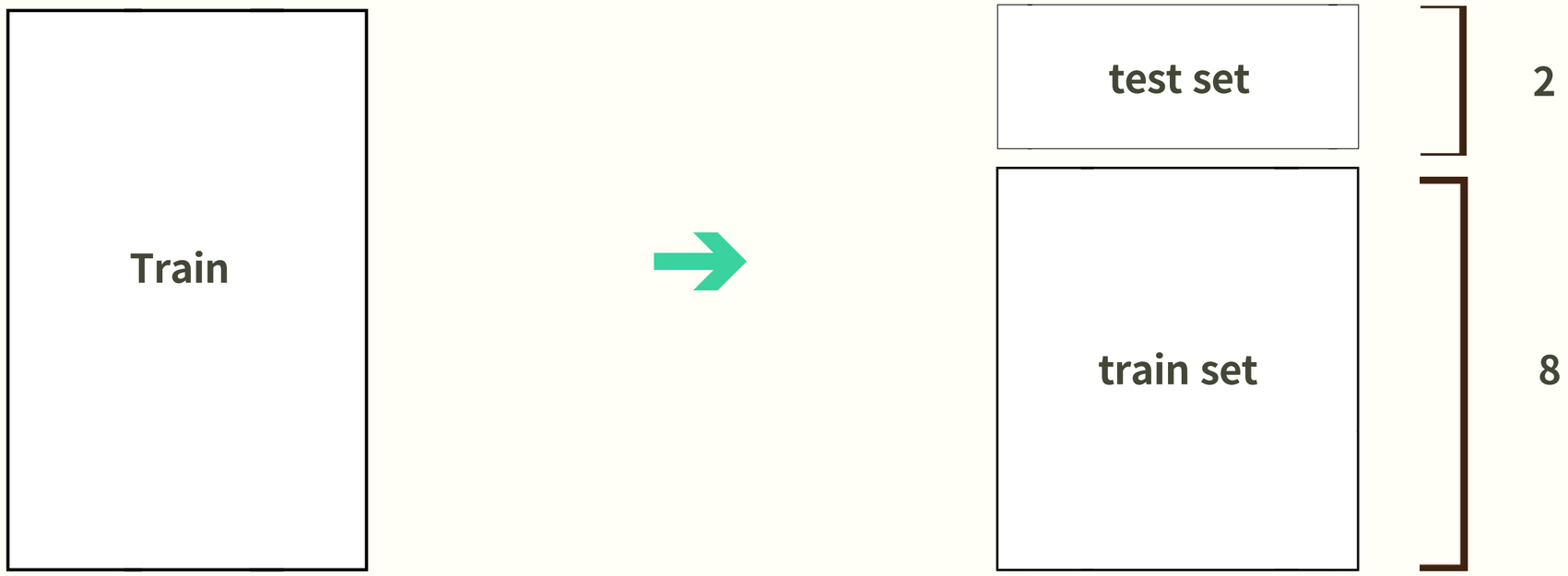
LGBM

- 경사 부스팅 알고리즘을 기반으로 하여 강력한 예측 성능을 제공
- 잎(leaf)-wise 방식을 사용하여 대규모 데이터셋에서도 빠른 학습 속도를 제공
- 히스토그램 기반의 결정 트리 학습 알고리즘을 사용하기 때문에
  기존의 알고리즘보다 훨씬 효율적인 메모리 사용량을 가지며 성능을 제공
- Gradient-Based One-Side Sampling (GOSS)와
  Exclusive Feature Bundling (EFB)이라는 두 가지 혁신적인 기술을
  사용하여 높은 정확성 유지

따릉이 대여 수 예측

모델링

모델 구성

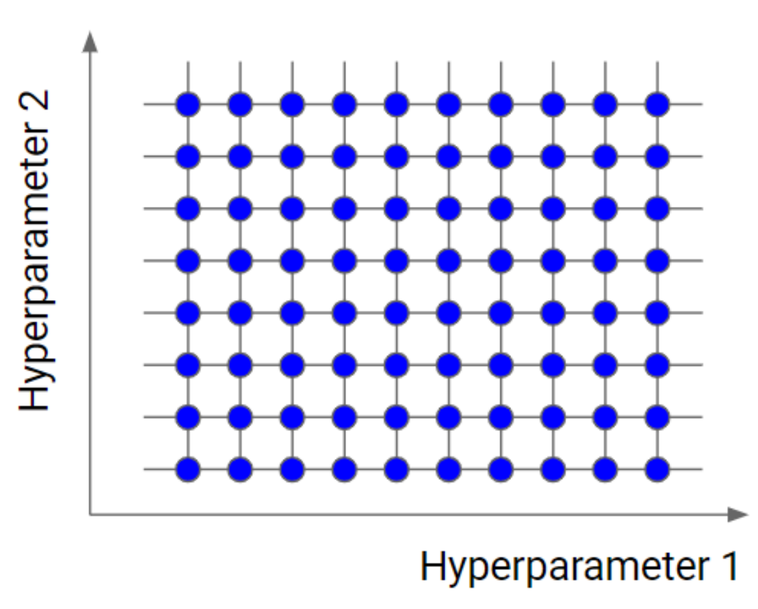


모델 평가를 위해 Train Data를 다시 train set과 test set으로 분리

따릉이 대여 수 예측

모델링

하이퍼 파라미터



Grid Search 기법

조절가능한 매개변수들로 하여금
가능한 모든 조합을 시도하여
최적의 하이퍼파라미터를 찾는 방법

따릉이 대여 수 예측

모델링

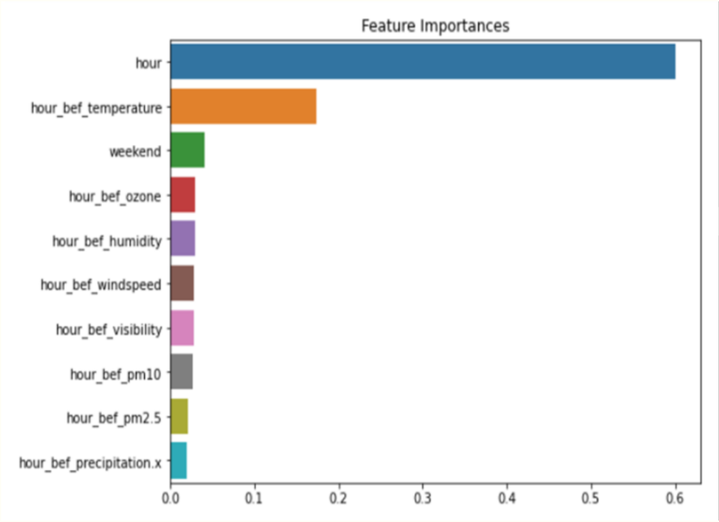
최종 모델 선정

Dacon에 제출한 RMSE 결과로 평가

모델 종류	RANDOM FOREST REGRESSOR	XGBOOST	SVM
데이콘 결과	30.58	30.87	53.42
순위	1	2	3

RANDOMFOREST REGRESSOR로 최종 모델 선정

변수중요도 결과



hour(시간),
hour_bef_temperature(기온),
weekend(주말여부) 순으로
모델에 반영됨

따릉이 대여 수 예측

결론 및 활용방안

결론1

평일 출퇴근, 등하교 시간대에
상대적으로
따릉이 대여 수요가 많다

결론2

주말에는 낮 시간대에
따릉이 수요가 많다

결론3

기온이 높아질수록
따릉이 수요가 올라가는 경향이 있다.

따릉이 대여 수 예측

결론 및 활용방안

활용방안1

평일 출퇴근, 등하교 시간대 전후로
따릉이 점검시간 확보(주요 시간은 피하기)

활용방안2

주말 낮 시간에는
다중이용시설 근처에
따릉이 배치 늘리기

활용방안3

온도에 높아짐에 따라
따릉이 이용객의 수가 증가하므로
여름에는 사설업체와 협력하여
사설 자전거 추가 배치

따릉이 대여 수 예측

감사합니다