神과함께

데이콘 식수 예측

화이트 박스 모델과
블랙 박스 모델의 활용 방안

박현호 신주영 신하진 최용현 최유현 최하린

CONTENTS

화이트 박스 모델과 블랙 박스 모델

화이트 박스 모델

(회귀분석,의사결정나무 ..)

- 내부 동작 원리가 공개되어 있는 모델
- 모델의 예측 결과를 해석 하고 설명하는데 용이
- 모델의 예측 결과를 이해 할 수 있어 파라미터를 수정하거나 변수를 추가하여 성능 향상 가능

- 복잡한 문제에 대해서는 모델링의 난이도 증가
- 모델이 복잡해질 수록 모델의 해석이 어려워 지는 것은 동일함
- 블랙 박스 모델에 비해 예측 성능이 떨어짐

블랙 박스 모델

(랜덤 포레스트, 부스팅 ...)

- 내부 구조나 작동 원리를 이해하기 어려움
- 높은 예측 성능과 대용량 데이터 처리에 적합
- 이미지, 음성, 자연어 처리 등 다양한 분야에 적용 가능

- 예측 성능은 뛰어나나 해석 및 설명의 어려움
- 과적합 가능성이 높음

데이터 전처리

파생변수

=> 강수량 / 기온 / 공휴일 전일
현재 근무자 수 / 중식메뉴개수 / 석식메뉴개수

범주화

=> 중식, 석식 메뉴 분할 & 메뉴 별 범주화
=> 범주형 변수 원핫 인코딩

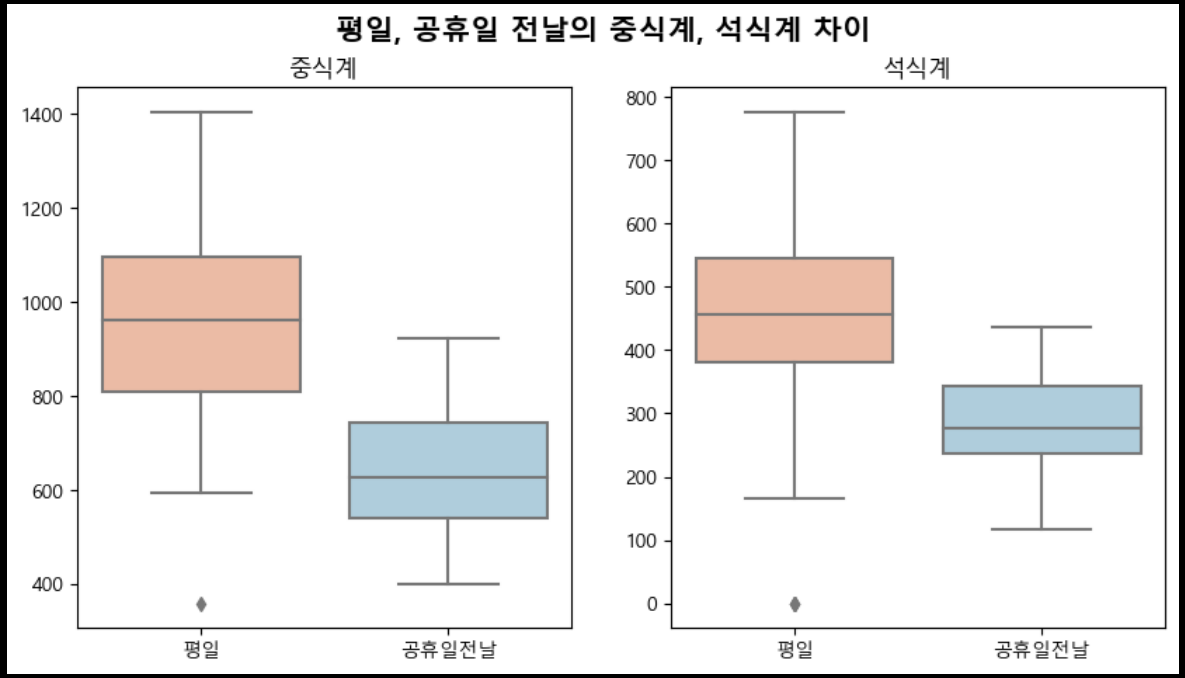
	본사 출장 자수	본사 휴가 자수	본사시간외근무명령서승인건수	현본사소속재택근무자수	공휴일_전날	중식계	석식계	중식_밥_덮밥	중식_밥_밥	중식_밥_볶음밥	...	석식_메인_생선	석식_메인_생선외해산물	석식_메인_소	석식_메인_채소	석식_메인_피자	금	목	수	월	화
0	150	50	238	0	0	1039	331	0	1	0	...	1	0	0	0	0	0	0	0	1	0
1	173	50	319	0	0	867	560	0	1	0	...	0	1	0	0	0	0	0	0	0	1
2	180	56	111	0	0	1017	573	1	0	0	...	1	0	0	0	0	0	0	1	0	0

3 rows × 81 columns

회귀 분석에 사용한 데이터 예시

데이터 시각화

1. 평일, 공휴일 전날

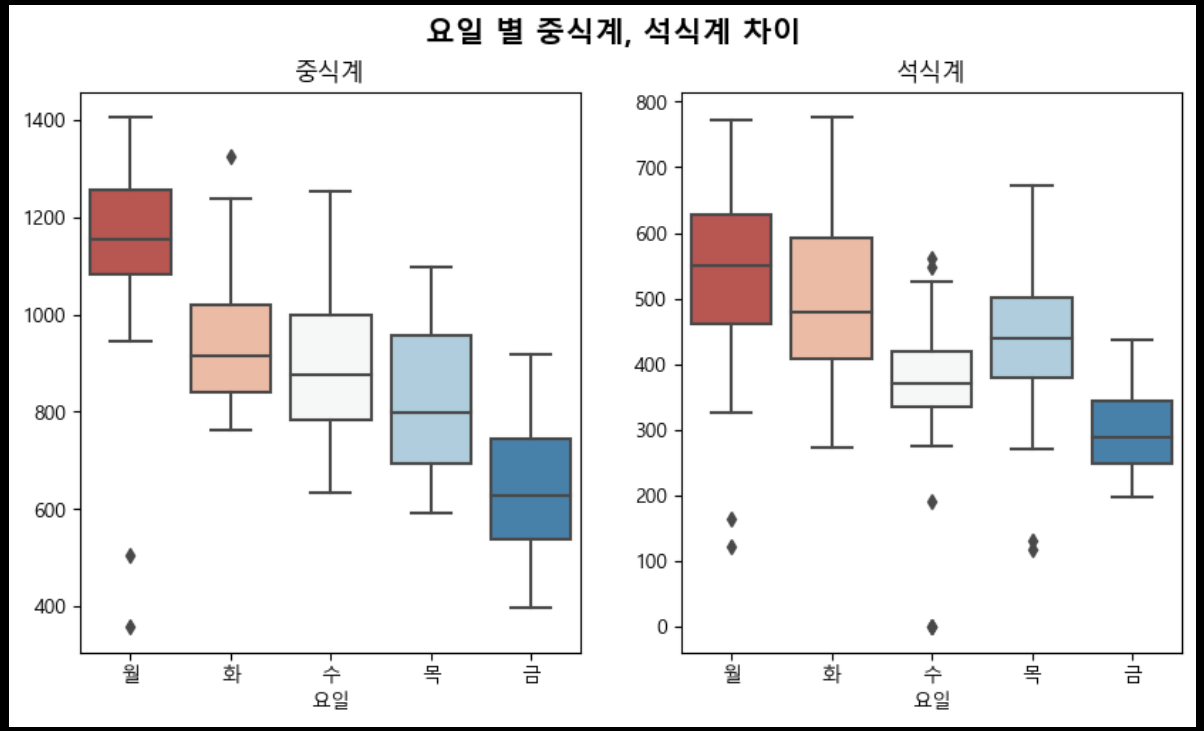


평일 : 월, 화, 수, 목
공휴일 전날 : 금요일 + 공휴일 전일

➡

중식계, 석식계
평균 ***50%*** 이상 차이

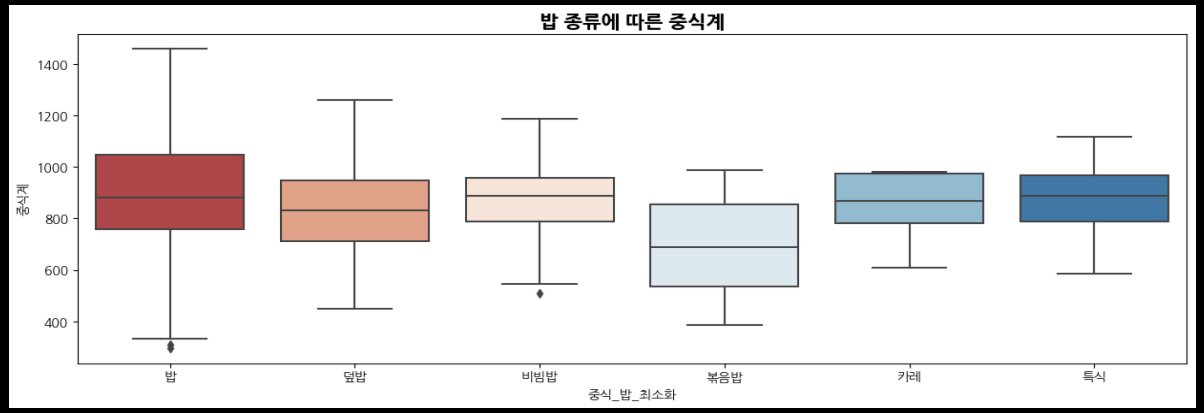
2. 요일

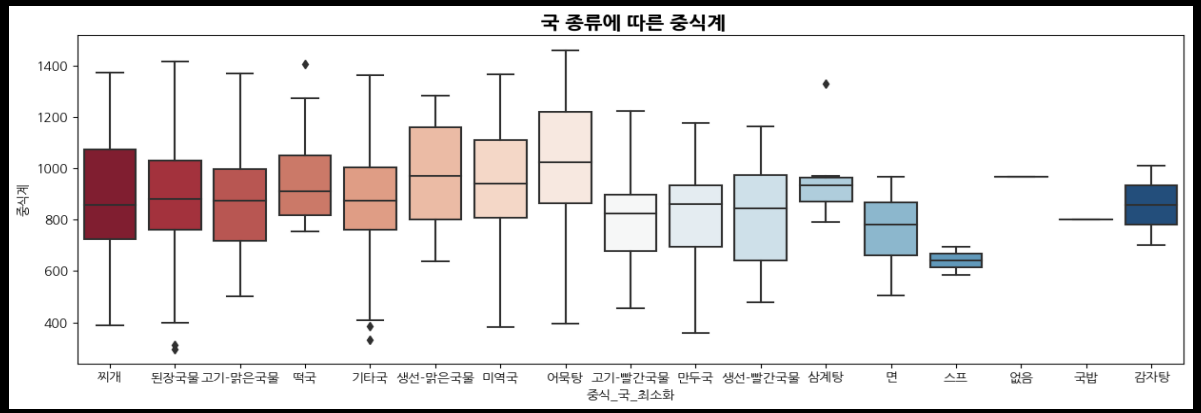


월요일 식수 가장 ▲

금요일의 식수 가장 ▼

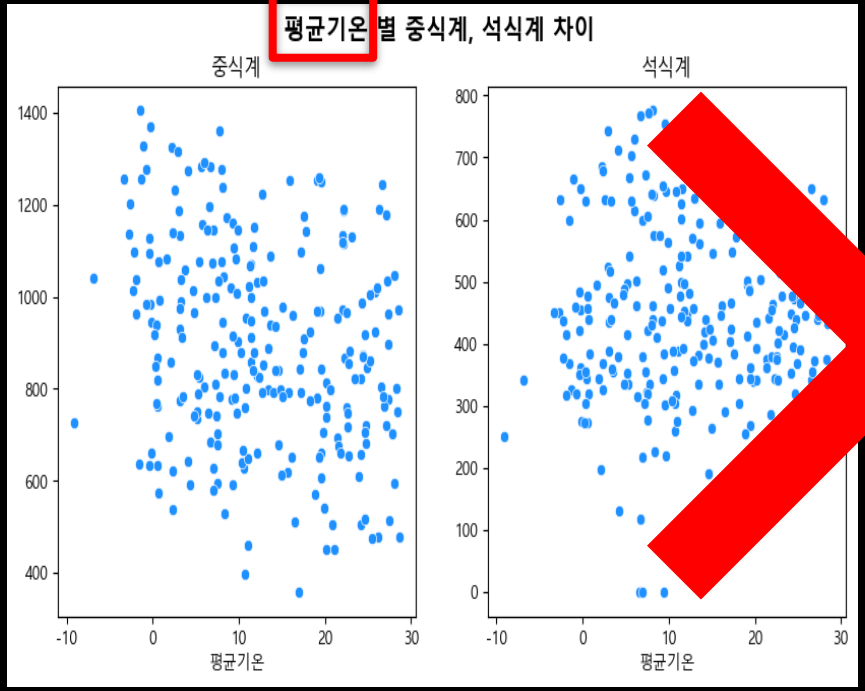
3. 메뉴

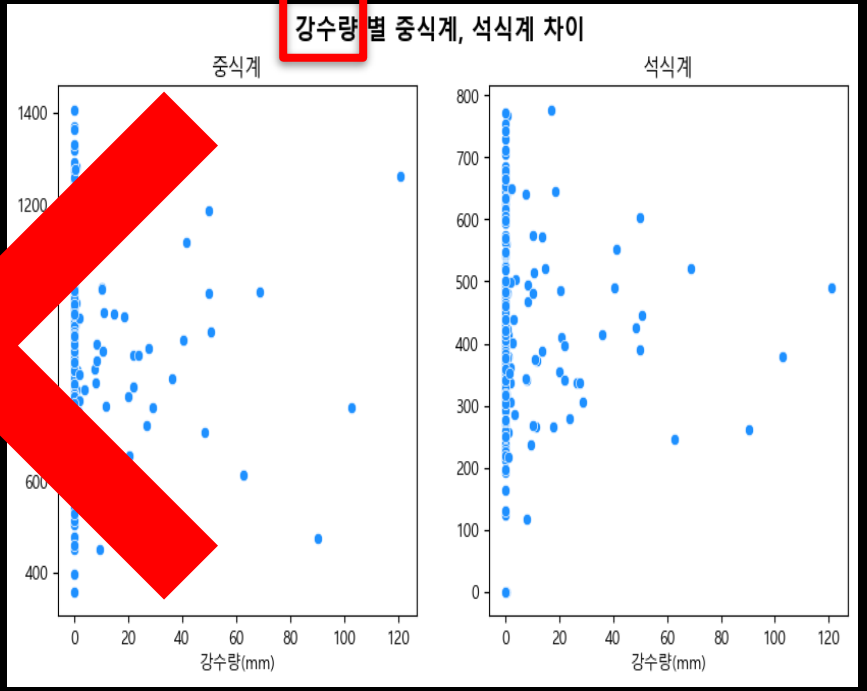




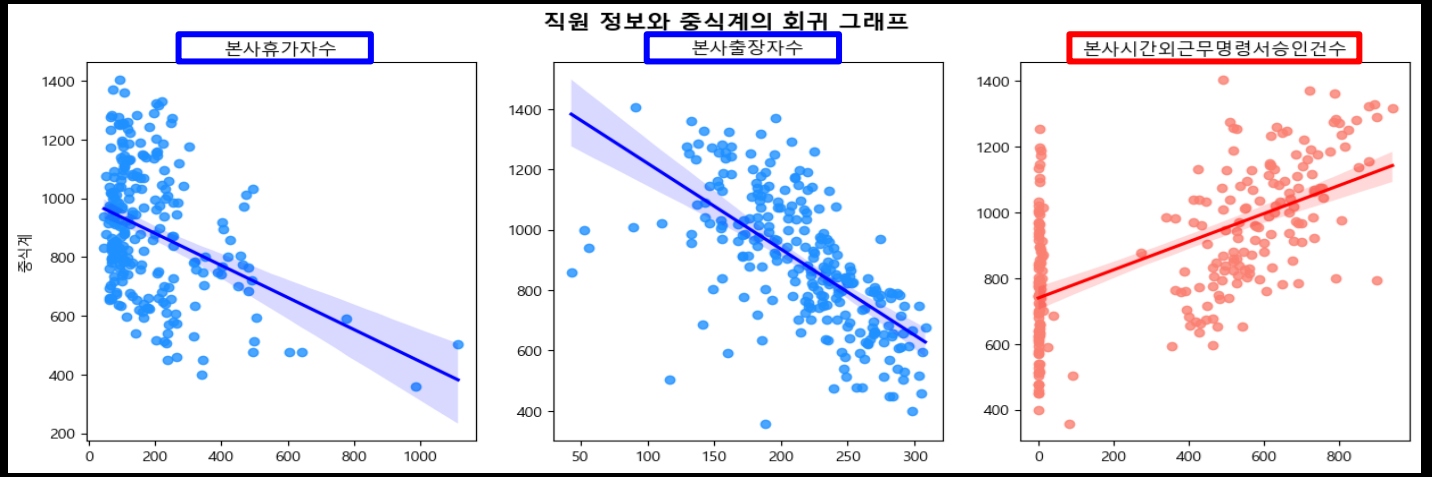
메뉴에 따른 식수 인원의 차이를 보임

4. 평균기온 & 강수량

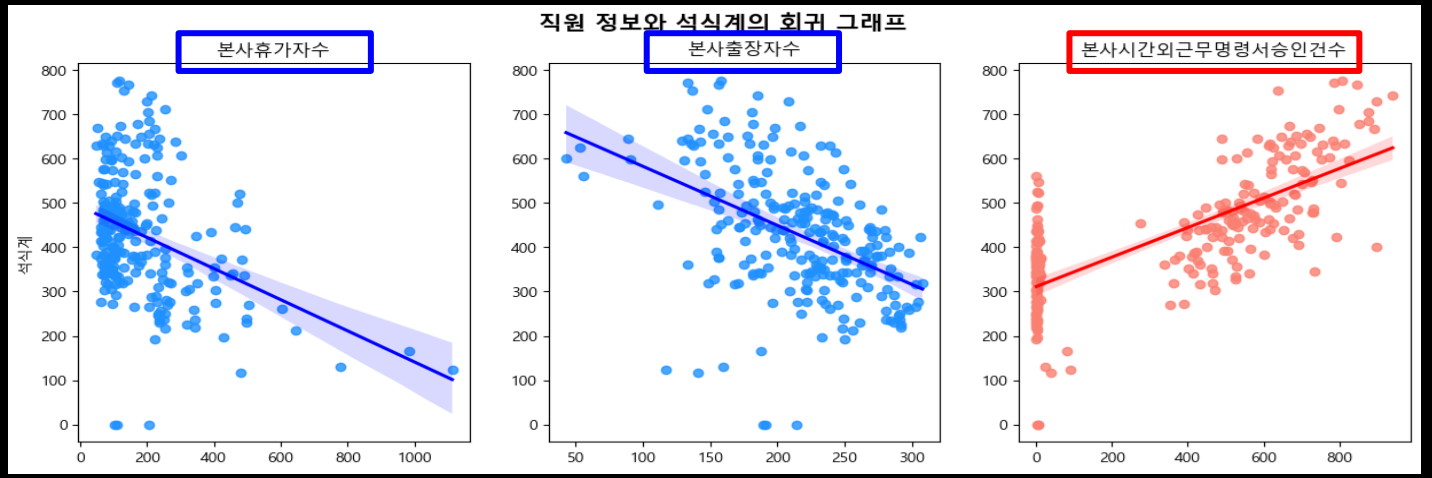

평균기온별 중식계, 석식계 차이
중식계
석식계
평균기온
평균기온


강수량별 중식계, 석식계 차이
중식계
석식계
강수량(mm)
강수량(mm)

중식계


직원 정보와 중식계의 회귀 그래프
본사휴가자수
본사출장자수
본사시간외근무명령서승인건수
중식계

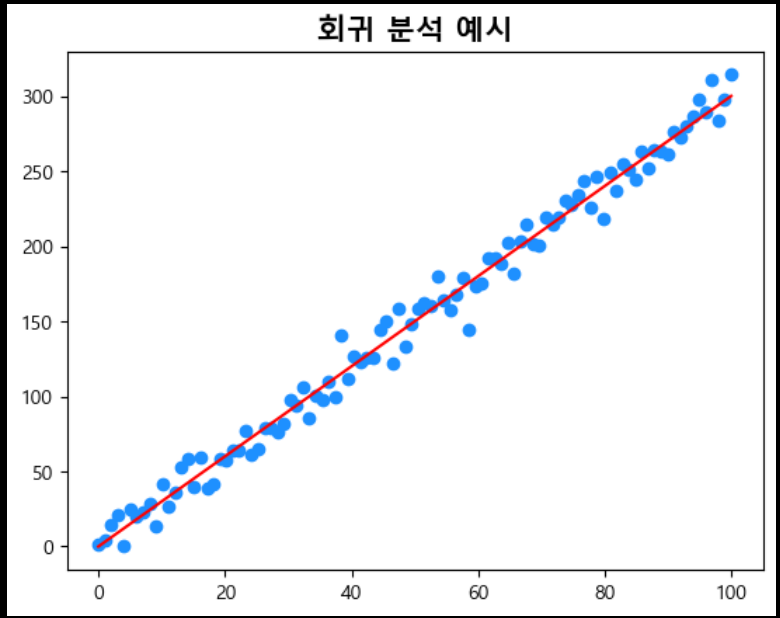
석식계


직원 정보와 석식계의 회귀 그래프
본사휴가자수
본사출장자수
본사시간외근무명령서승인건수
석식계

데이터 모델링

✔ 다중 선형 회귀 분석

하나의 종속 변수를 다양한 독립 변수들을 이용해 선형적인 관계로 예측하는 통계적 방법

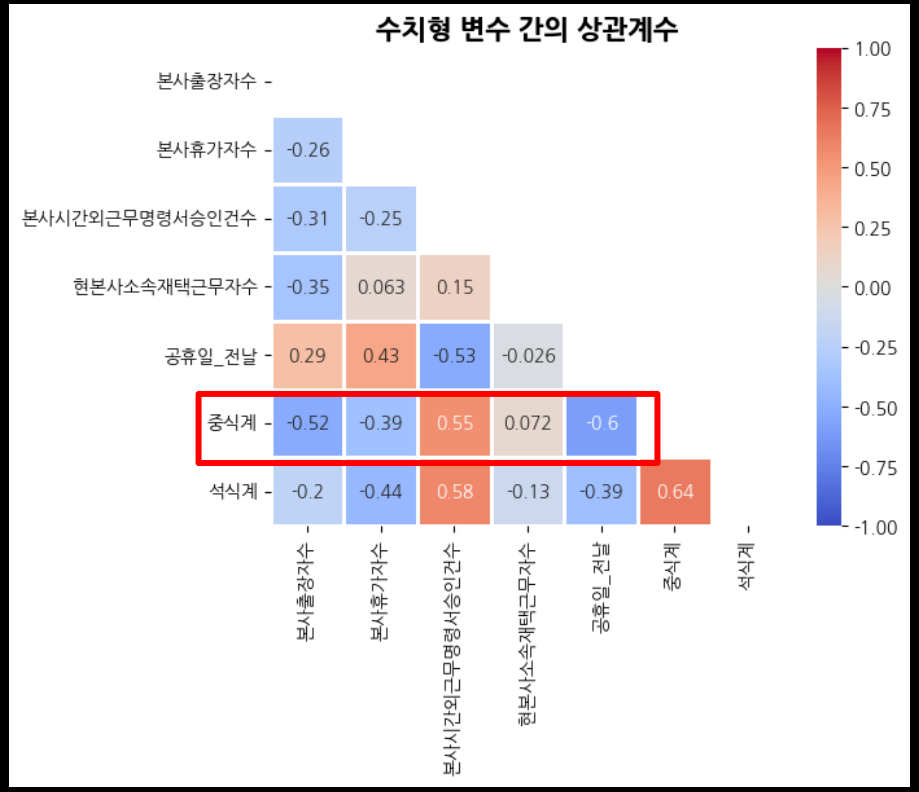


✔ 목표

예측 모델로서 얻은 **회귀 계수**로 해당 데이터를 해석

Ridge, Lasso 회귀 분석을 통해 다중 공선성 해결 및 모델의 간소화

상관관계



재택근무자수를 제외한 다른 변수들과 중식계는
상관 관계를 보임

모델 평가 과정
OLS,Ridge,Lasso 중 모델 선택

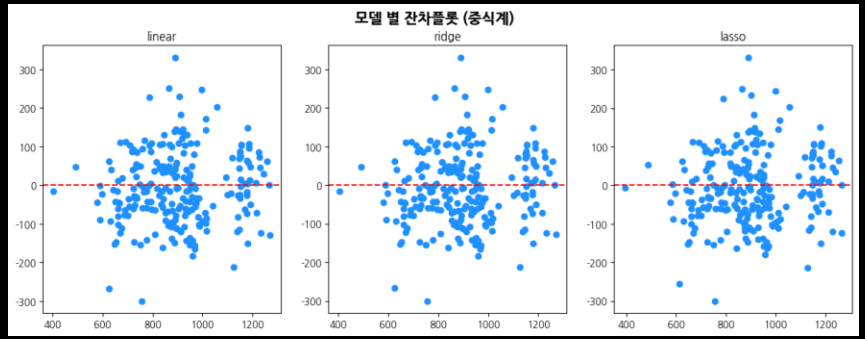
```
OLS
Train set 에 대한 점수: 0.7825346757860546
valid set 에 대한 점수: 0.7823383828482172
----------------------------------------
RIDGE
Train set 에 대한 점수: 0.7824784366918112
valid set 에 대한 점수: 0.7827240270709891
----------------------------------------
LASSO
Train set 에 대한 점수: 0.7783723425740037
valid set 에 대한 점수: 0.7869670914108863
```

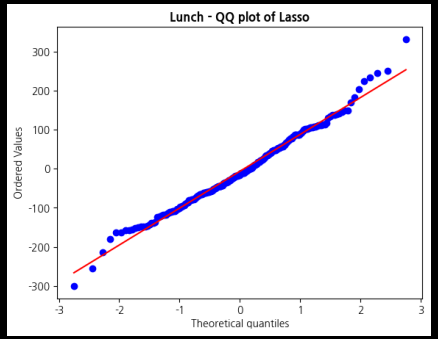
Lasso(alpha = 0.5) 로 시행한 모델의 성능이 가장 높았음

Lasso 회귀 분석 모델을 이용하여 데이터를 설명하도록 함

모델 평가 과정

잔차 플롯으로 모델의 유의성 판단





Residual plot , QQ-plot 을 이용하여 사후검정

회귀 분석 모델 결과 -1

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_k X_k$$

- ✓ 월요일
- ✓ 수요일
- ✓ 목요일
- ✓ 공휴일 전날

해당 O : 1
해당 X : 0

중식계 = 1351

+ **월요일** * (173)
+ **수요일** * (69)
+ **야근자 수** * (0.21)
+ **공휴일 전날** * (-50)
+ **목요일** * (-50)
+ **화요일***(-16)
+ **출장자수** * (-1.84)
+ **휴가자수** * (-0.56)
+ **재택근무자수** * (-0.18)

회귀 분석 모델 결과 -2

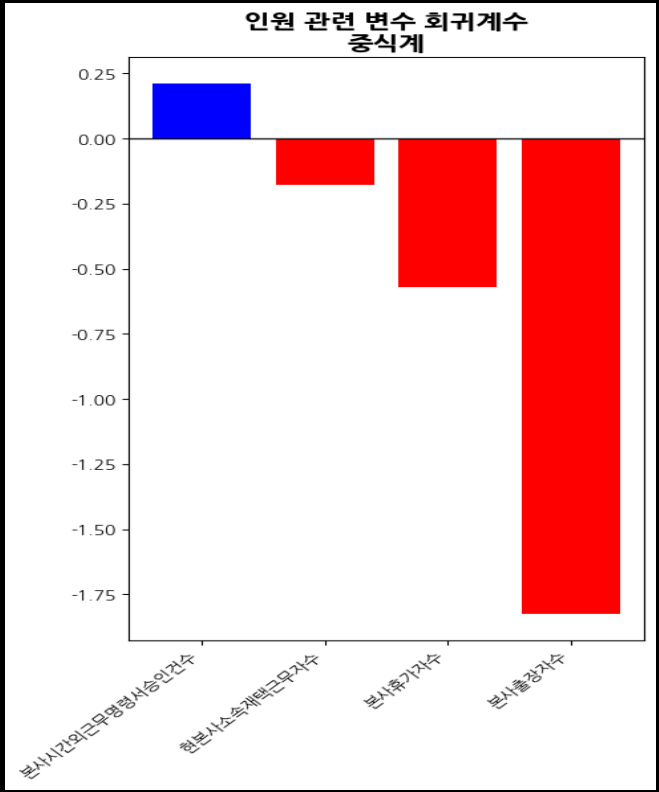
```
절편 = 1351
월요일 = 1
수요일 = 0
야근자수 = 150
공휴일_전날 = 0
목요일 = 0
재택근무자수 = 10
휴가자수 = 100
출장자수 = 120
```

1351 + 월요일 * 173 + 야근자수 * 0.21 + 재택근무자수 * -0.17 + 휴가자수 * -0.56 + 출장자수 * -1.4 = 1330 (예측 된 중식계 인원)

예를 들어
월요일에 야근자수가 150명,
재택 근무자수가 10명,
휴가자수가 100명,
출장자 수가 120명인 날에는
예상 중식 식수 인원이 약 1330 명일 것이라고 예측 할 수 있다.

회귀 계수 - 수치형 변수 해석

(독립변수의 변화에 따른 종속변수의 변화량)

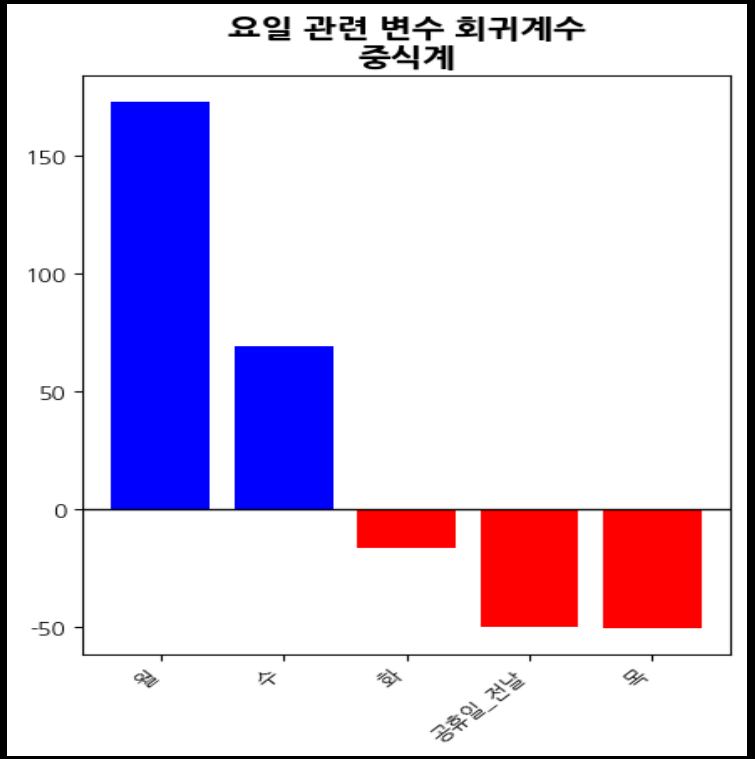


본사 출장자수 회귀 계수(-1.84)

본사 출장자 수가 100명이면 중식 식수 인원이 184명 감소한다

회귀 계수 - 범주형 변수 해석

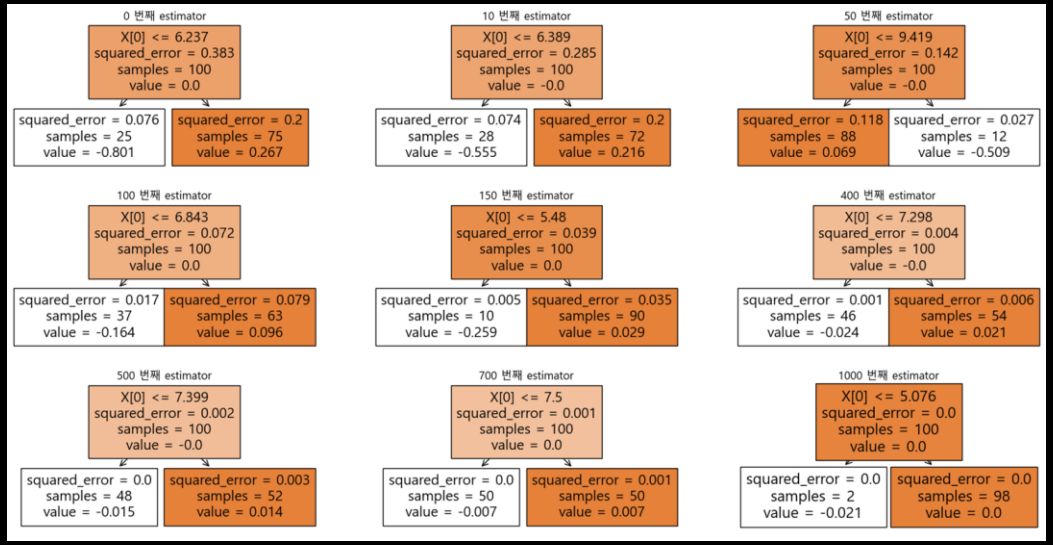
(독립변수의 변화에 따른 종속변수의 변화량)



월요일 회귀 계수 (+173)
공휴일 전날 회귀 계수 (-50)

월요일의 경우 중식 식수 인원 +173▲
공휴일 전날의 경우 중식 식수 인원 -50 ▼

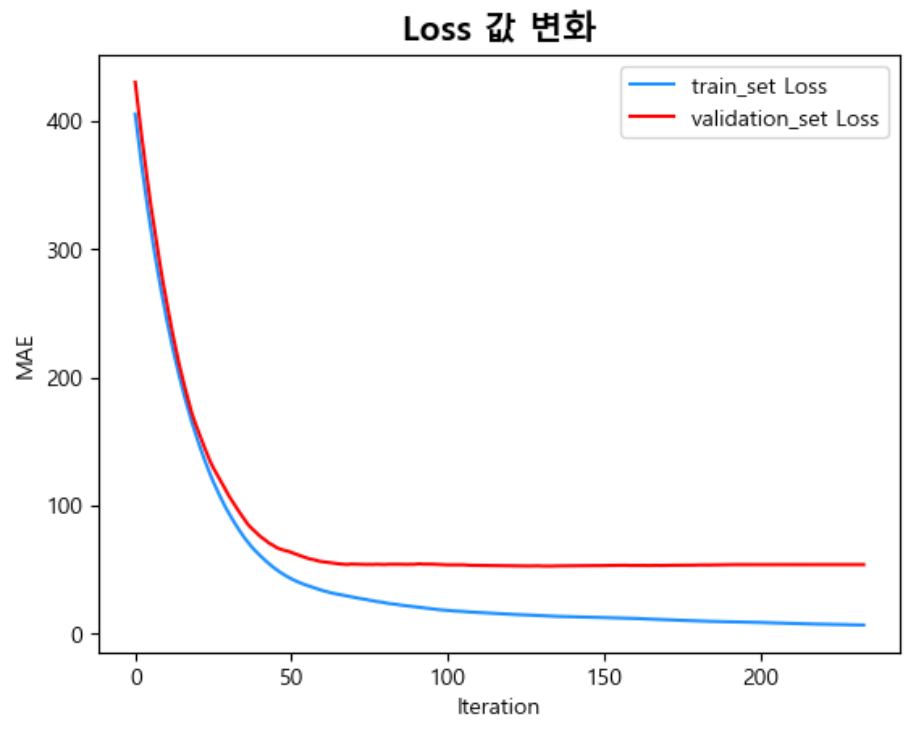
XGBoostRegressor



✔ 순차적으로 약한 학습기를 결합하여 만드는 앙상블 모델

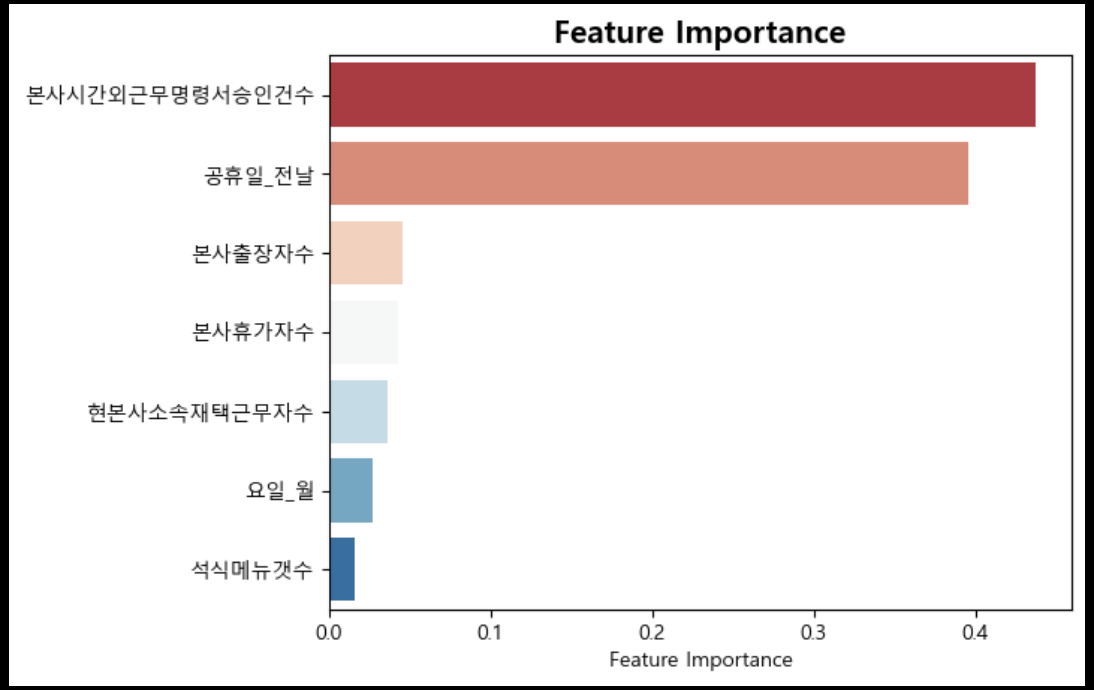
✔ 결과를 해석하는데 용이하지 않지만 높은 예측 성능을 가지고 있음

모델링 과정



10번의 교차 검증을 통해 적절한 estimator 와 iteration 을 찾아 모델 선정

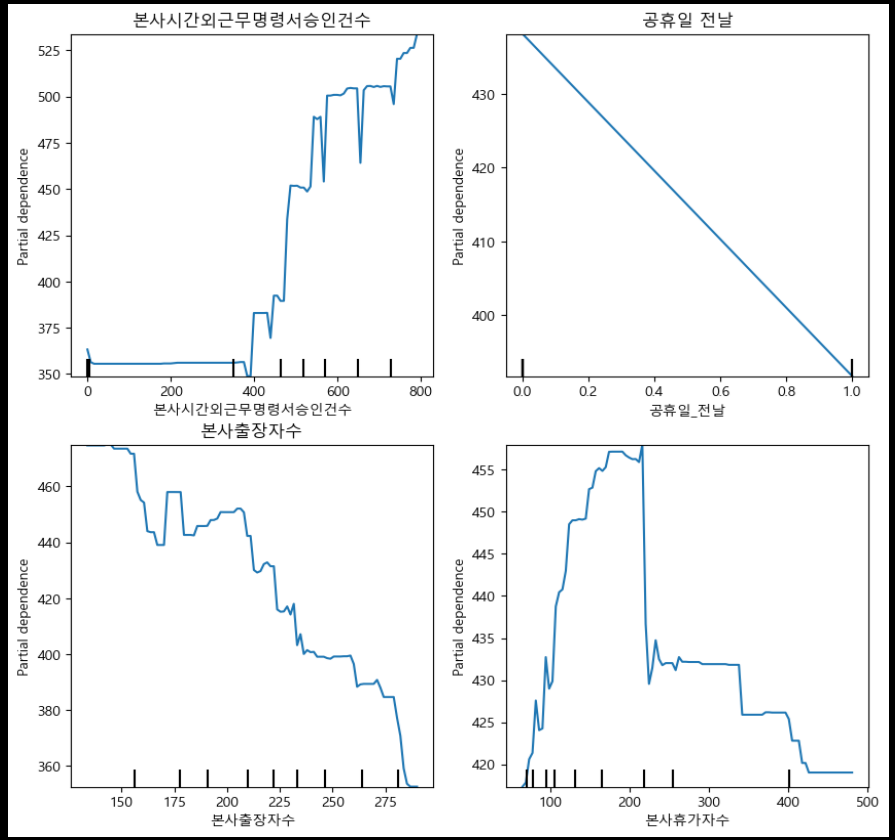
Feature importance



석식계 Feature importance

1. 야근자 수
2. 공휴일 전날
3. 출장자 수

Partial Dependence Plot (PDP)



PDP

→ 해당 **feature의 변화에 따른 예측 값의 변화** 표현

단점

→ 다양한 변수 간의 상호관계를 고려X

변화의 정도만 파악 가능 변화량 설명 부적절

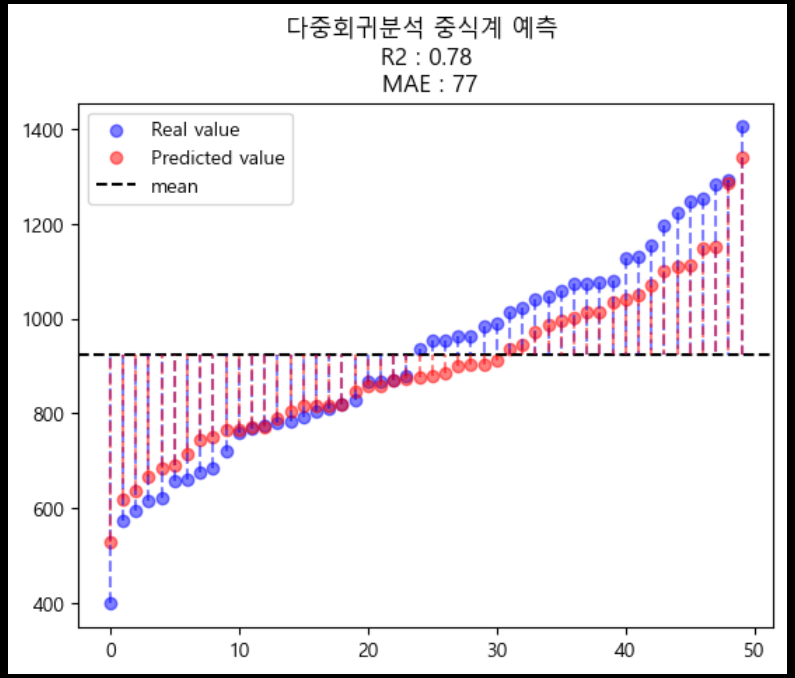
결론

→ 독립 변수와 종속 변수간의 관계를 설명에 있어

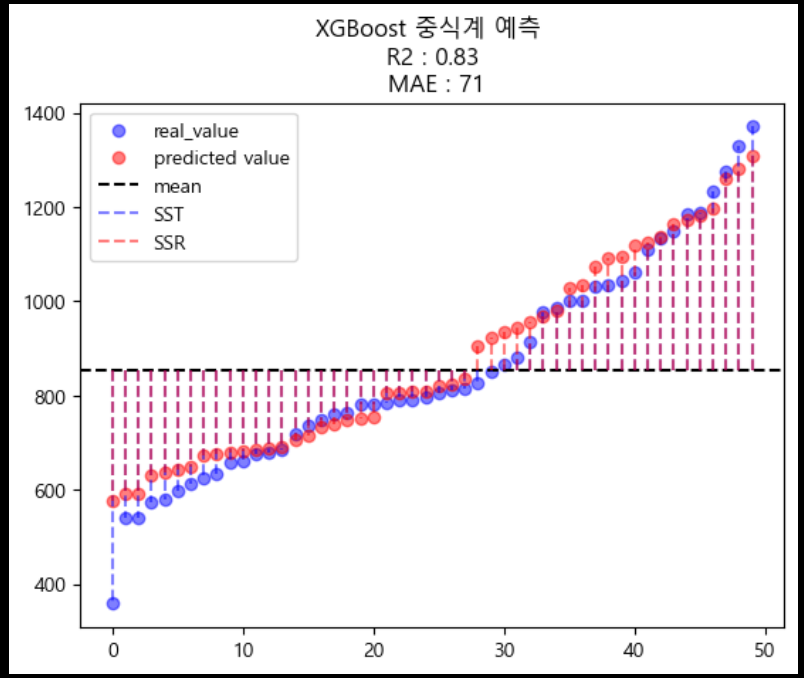
블랙 박스 모델은 ***부적절***

모델의 예측 성능 비교

Lasso Regressor

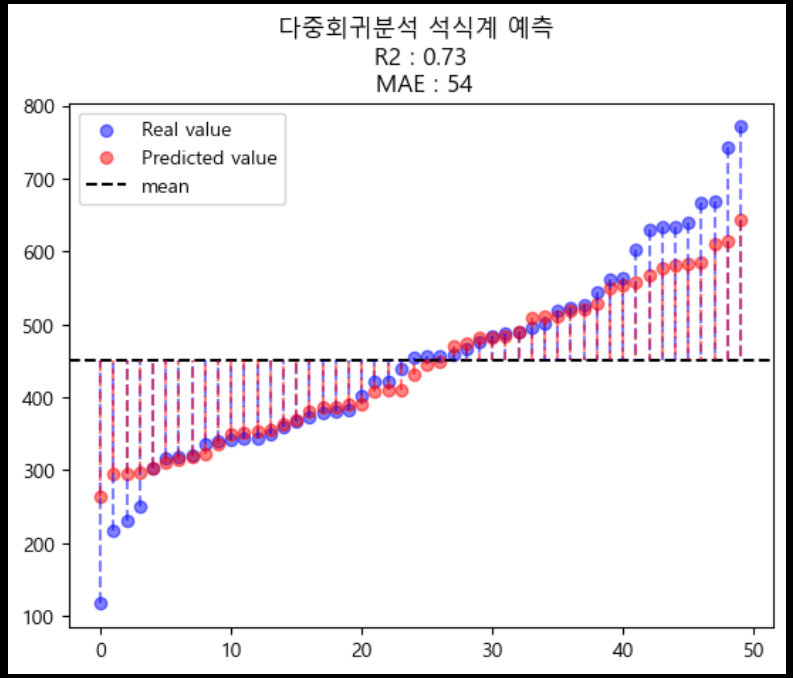


XGBoostRegressor

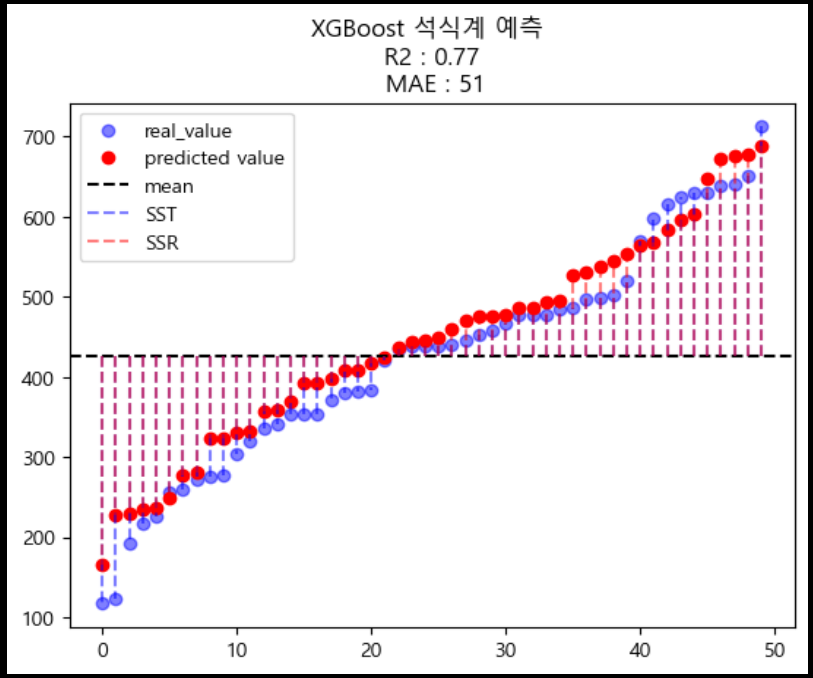


모델의 예측 성능 비교

Lasso Regressor



XGBoostRegressor



결론

다중 선형 회귀 분석

1. 중식계 예측
   R2 score : 0.78
   MAE : 77

2. 석식계 예측
   R2 score : 0.71
   MAE : 54

예측 결과를 통해 해당 데이터를 설명하는데 용이

XGboostRegressor

1. 중식계 예측
   R2 score : 0.83
   MAE : 71

2. 석식계 예측
   R2 score : 0.77
   MAE : 51

예측 성능은 뛰어나나
결과 값에 대한 해석 및 설명이 복잡하고 어려움