강우, 수위 데이터를 이용한 댐 유입량 예측 모형

빅콘테스트 퓨처스리그 홍수 zero 분야
김동건

CONTENTS

1. 문제 정의

- 최근 전세계적으로 이상기후 변화를 통해 홍수 피해 증가
- 한국도 홍수 예방, 수해 복구에 연간 3200억원의 예산이 필요
- 홍수 피해 중 가장 규모가 큰 하천 범람을 막기 위해 댐 유입량 예측을 통해 댐 수위조질이 필요

1. 문제 정의

- 강수량과 수위데이터를 통해 앞으로의 댐유입량을 예측하고, 적절한 댐방출량을 정하여 장마, 홍수와 같은 자연재해에 대하여 사전대응할 수 있게 된다.

1. 데이터 분석

2. 최적의 예측 모형 개발

3. 평가 데이터에 적용

2. 데이터

- 빅콘테스트 제공데이터 : 2021 빅콘테스트_데이터분석분야_퓨처스리그_홍수ZERO_댐유입량,강우,수위 데이터._210803 xlsx

테이블설명	2006년부터 2018년까지 댐 유입량			
No.	컬럼ID	타입	NULL	비고
1	홍수사상번호	NUMBER(2)	N	
2	연	NUMBER(4)	N	
3	월	NUMBER(1)	N	
4	일	NUMBER(2)	N	
5	시간	NUMBER(2)	N	
6	유입량	NUMBER(6,1)	N	
7	유역평균강수	NUMBER(4,1)	N	
8	강우(A지역)	NUMBER(4,1)	N	
10	강우(B지역)	NUMBER(4,1)	N	
11	강우(C지역)	NUMBER(4,1)	N	
12	강우(D지역)	NUMBER(4,1)	N	
13	수위(E지역)	NUMBER(4,1)	N	
14	수위(D지역)	NUMBER(4,1)	N	

데이터 전처리

각 데이터세트 별로 6개의 데이터세트 생성

유역평균강수	강우(A지역)	강우(B지역)	강우(C지역)	강우(지역)	수위(E지역)	수위(D지역)

유역평균강수	강우(A지역)	강우(B지역)	강우(C지역)	강우(지역)	수위(E지역)	수위(D지역)

⋮

유역평균강수	강우(A지역)	강우(B지역)	강우(C지역)	강우(지역)	수위(E지역)	수위(D지역)

6개의 데이터 집단은 각각 측정 방식의 차이를 통해 다른 결과값을 가짐

6개의 데이터집단을 나누어 개별적으로 분석에 이용

데이터 전처리

연	월	일	시간

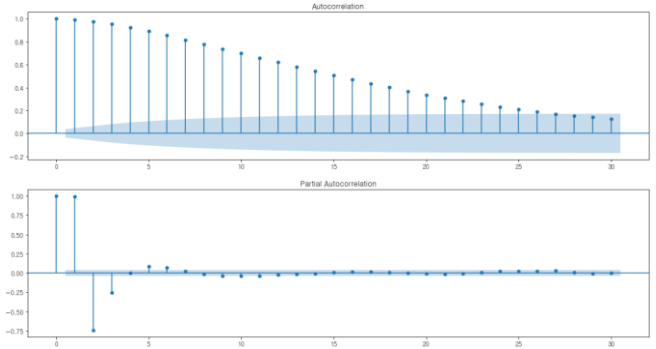
datetime

데이터의 분포와 특성을 파악하기 위해서 시계열 분석을 이용

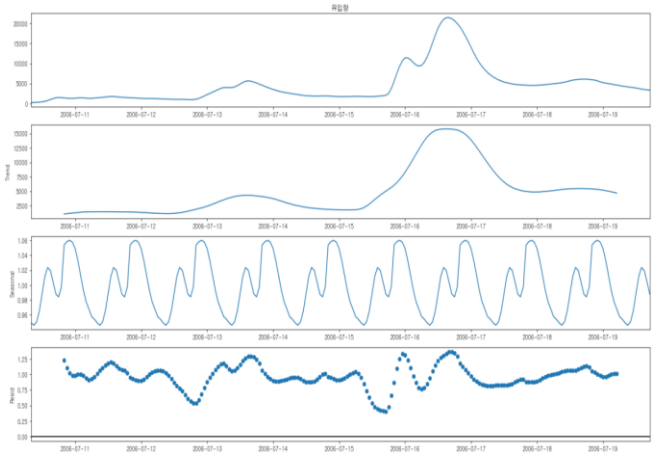
연, 월, 일, 시간 속성을 datetime형식으로 변환 후 index로 지정

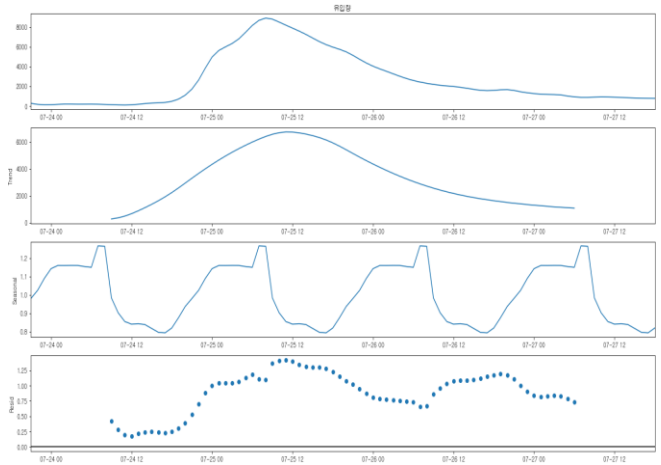
유입량의 시계열 분석을 통해 시계열 분석을 통해 모델을 예측하는 것이 적절한지 판단

시계열 데이터 분석



Result of Dickey-Fuller Test	
Test Statistic	-1.148764
p-value	0.695239
#Lags Used	12.000000
Number of Observations Used	87.000000
Critical Value (1%)	-3.507853
Critical Value (5%)	-2.895382
Critical Value (10%)	-2.584824
dtype: float64	



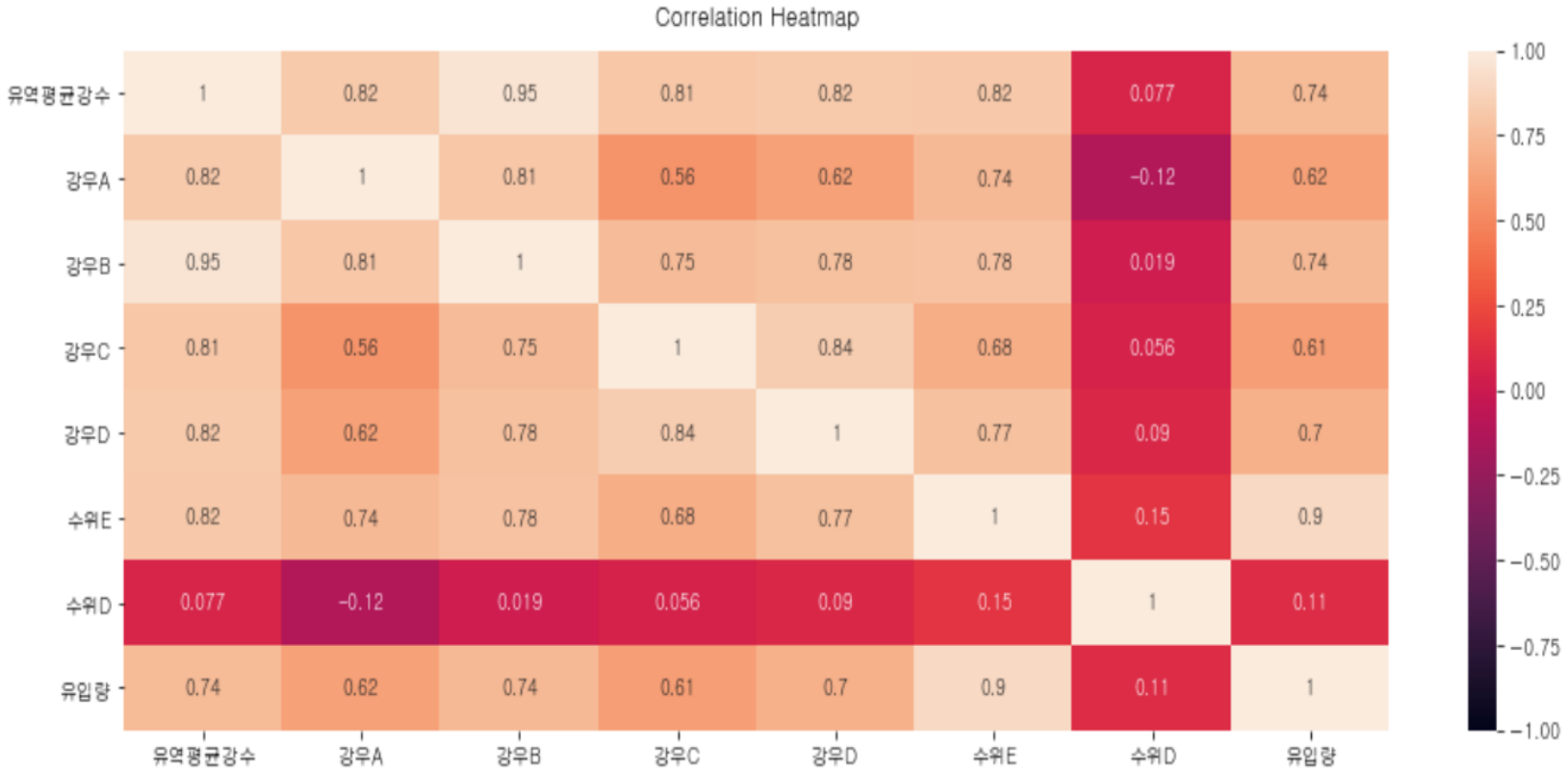


ACF, PACF Dickey-Fuller Test분석을 통해 홍수사상별로 데이터를 분석해본 결과, 각 데이터는 stationarity한 특성을 갖고 있지 않아 시간에 따른 변화를 보이고 있다고 판단할 수 있음

홍사사상번호가 1인 데이터와 홍수사상번호가 5인 데이터를 Timeseries_decompose분석을 통해 분석해본 결과, 유입량은 홍수사상번호에 따라서 다른 주기적 특성을 가지고 있어 시계열 분석모델을 통해 데이터를 예측하는 것은 힘들다고 판단

상관계수 분석

각 feature간 데이터 상관성을 확인하기 위한 correlation분석

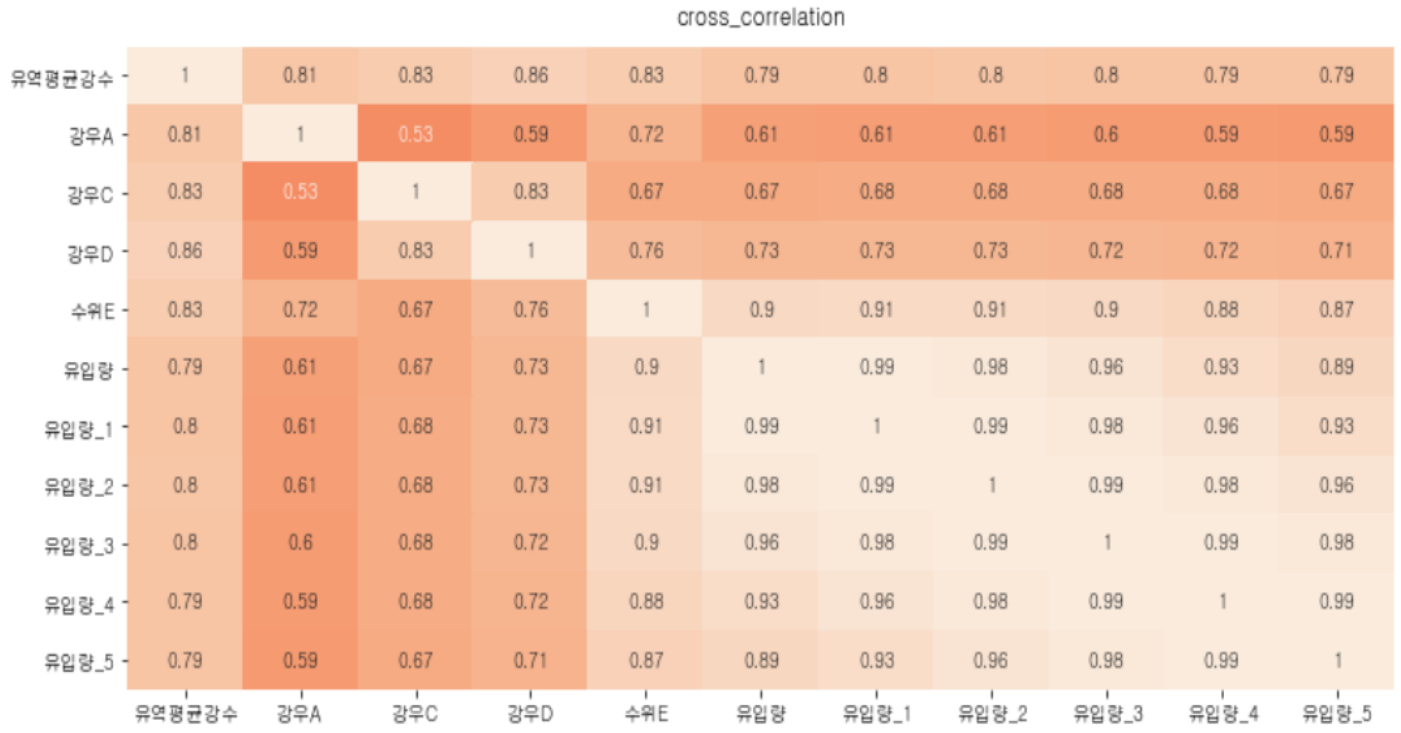


수위D데이터는 유입량과의 상관계수가 다른 feature에 비해 현저히 낮은 값을 가지므로 분석에서 제외

강우B데이터는 유역평균강수와 강한 상관관계를 가져 다중공선성 문제로 분석에서 제외

Cross-Correlation분석

시간변화에 따른 유입량과 각 feature간 cross-correlation분석



유입량은 각 데이터로부터 time lag가 발생한다고 판단하여 유입량의 time lag를 다양히 하여 분석

각 데이터세트의 측정 시점이 세트별로 분석하기 용이할 것으로 판단하여 feature간 cross-correlation분석은 하지 않음

유입량_2와 다른 feature간 correlation이 유입량과의 correlation보다 높은 상관계수를 가지는 것을 볼 수 있음

최종 데이터

각 데이터 세트별 X_features

유역평균강수	강우(A지역)	강우(C지역)	강우(지역)	수위(E지역)

유역평균강수	강우(A지역)	강우(C지역)	강우(지역)	수위(E지역)

⋮

유역평균강수	강우(A지역)	강우(C지역)	강우(지역)	수위(E지역)

Y_label

유입량_2

유입량_2 데이터는 각 데이터 세트별로 홍수사상번호를 기준으로 나눈 뒤 유입량 데이터를 timelag 2만큼 지연시킨 데이터

각 데이터 세트별로 모델링을 통하여 예측값을 구한 후, 상관계수가 높은 데이터세트 순서에 따라 가중치를 부여해 예측

3. 모델링

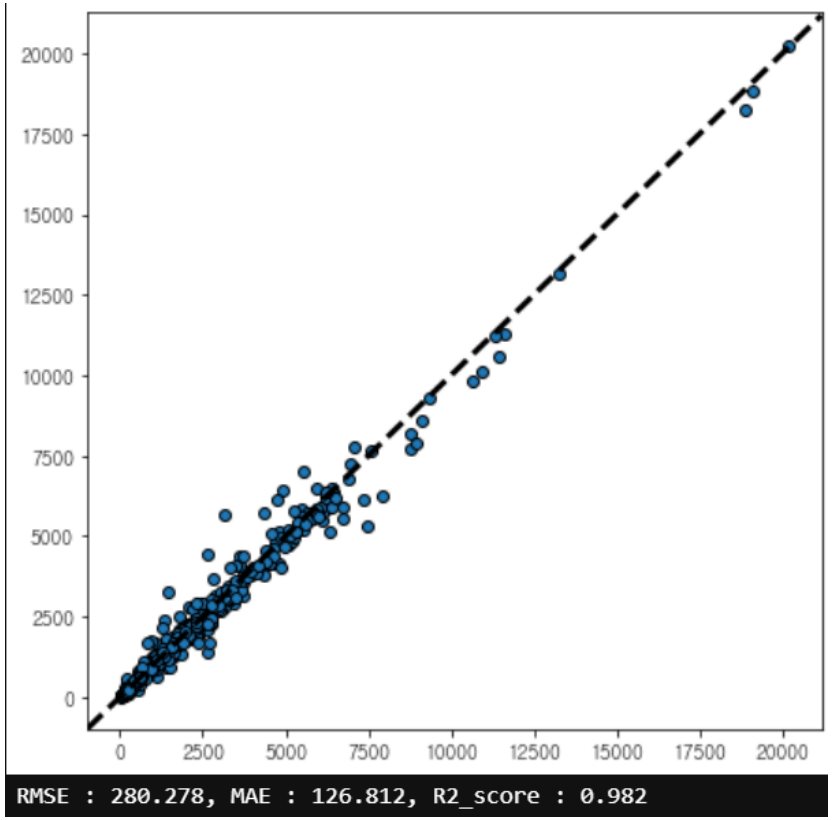
	Model	MAE	MSE	RMSE	R2	RMSLE	MAPE	TT (Sec)
et	Extra Trees Regressor	157.0938	126734.2124	354.6726	0.9720	0.1969	0.1185	0.0740
xgboost	Extreme Gradient Boosting	183.8682	139227.0531	371.9761	0.9699	0.2466	0.1613	0.2040
rf	Random Forest Regressor	193.4063	180633.0013	424.1743	0.9613	0.2296	0.1463	0.1080
gbr	Gradient Boosting Regressor	236.2944	186511.5301	430.9041	0.9596	0.3615	0.2997	0.0380
lightgbm	Light Gradient Boosting Machine	233.6601	289527.7914	532.8804	0.9429	0.2934	0.2188	0.1460

Pycaret 라이브러리를 통한
최적의 모델을 탐색

각 데이터 세트에 대해서
분석해본결과 대부분
ExtraTreeRegressor
ExtremeGradientBoosting
RandomForestRegressor
GradientBoostingRegressor
LightGradientBoostingMachine
모델 순으로 RMSE평가에서
좋은 성능을 보임

이중 Lightgbm모델을 제외한
나머지 4모델에 대하여
분석모델 구현

모델 최적화 전략

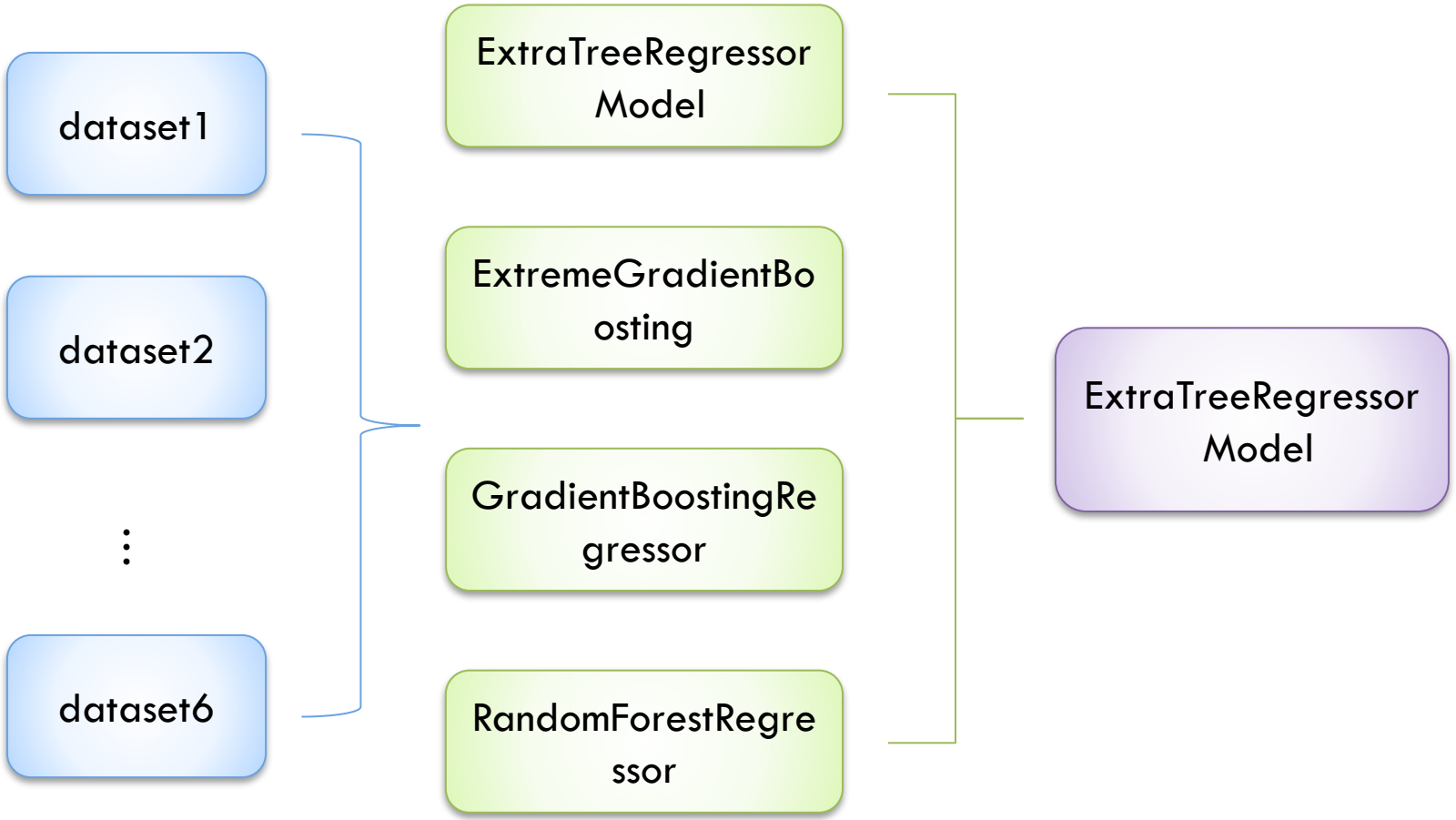


Pycaret 라이브러리를 통한 AutoML분석으로 최적모델을 찾은 후 각 데이터세트에 개별모델로 예측수행

대부분의 모델에서 RMSE값은 250~450의 값을 가지며
데이터세트에 따라 최적의 값을 가지는 모델은 서로 다른 것 으로 판단

스태킹 앙상블 모형을 통하여 오버피팅을 줄이고 다양한 데이터세트에 맞게 모델링을 구성

모델 최적화 전략



데이터 세트별로 CV세트기반의 스태킹모델을 구현하여 구현

폴드수는 5로 정했으며 Test_Set 비율은 0.2로 정하여 모델을 구성

최종모델로는 AutoML분석에서 가장 성능이 좋았던 ExtraTreeRegressor모델을 사용

모델 최적화 전략

각 데이터 세트별로 가중치를 달리하여 최종 결과값을 평균화

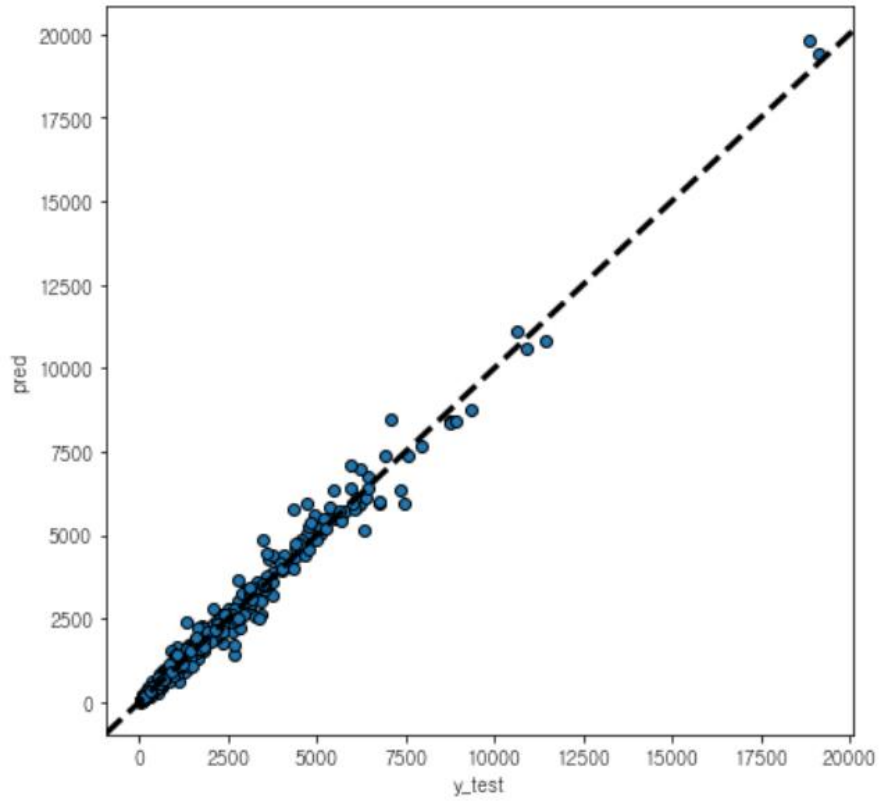
가중치는 각 데이터세트별로 '유입량_2'와의 상관계수 구했을 때, 상관계수가 높은 데이터세트 순서대로 가중치를 부여하였다.

$$\text{최종결과값} = \frac{0.5}{6} \times \text{데이터세트1의 결과값} + \frac{0.7}{6} \times \text{데이터세트2의 결과값} +$$

$$\frac{1}{6} \times \text{데이터세트3의 결과값} + \frac{1}{6} \times \text{데이터세트4의 결과값} +$$

$$\frac{1.3}{6} \times \text{데이터세트5의 결과값} + \frac{1.5}{6} \times \text{데이터세트6의 결과값}$$

모델 최적화 전략




```
RMSE : 269.781, MAE : 147.080, R2_score : 0.984
```

스태킹앙상블 모형을 통해 구현결과

RMSE 값은 269.78
R2_score값은 0.984

개별모델로 수행했던 결과들보다 더 좋은 성능을 보여줌

홍수사상별로도 성능에 큰 영향을 주지 않고, 데이터의 양도 최종예측모델에서는 검증데이터를 두지 않고 학습데이터를 전체 데이터로 사용

4. 예측결과 및 해석

최종 예측모델의 학습데이터로는 각 데이터의 홍수사상번호1~25 까지의 모든 데이터를 사용하였으며, 홍수사상26번의 6개 데이터세트의
유역평균강수, 강우A, 강우C, 강우D, 수위E feature를 이용하여 유입량을 예측하였다.

모델링 결과로 나온 결과는 유입량_2의 결과이므로 최종 유입량의 결과에 맞추어 index가 +2인 값에 넣어주었다.

앞에 생략된 2개의 예측결과는 각 데이터세트의 이전시점 feature데이터가 존재하지 않아 값을 정확히 예측하지 못하였다. 다만 적은 강우량과 수위데이터를 고려하여 예측한 유입량_2의 첫번째 데이터를 생략된 데이터에 넣어주었다.

한계점

모델링을 통해 분석해본 결과,

이전 시점의 각 지역(ABCDE)의 강우, 수위 측정량의 회귀분석으로 유의미한 유입량을 예측할 수 있다. 하지만 이전에 제시되었던 데이터세트에는 측정된 각 지역의 특성이나 거리 등이 반영되지 않아 더 정밀한 분석을 할 수 없었다.

정밀한 분석을 위해서는 각 지역의 측정값과 유입량의 cross_correlation분석을 통해 어느 시점의 feature데이터들이 유입량에 영향을 미치는지 조사할 필요가 있고 제공데이터에 포함되지 않은 지역적 특성을 고려하여 분석에 포함하여야 한다.