빅콘테스트 2021 결과 보고서

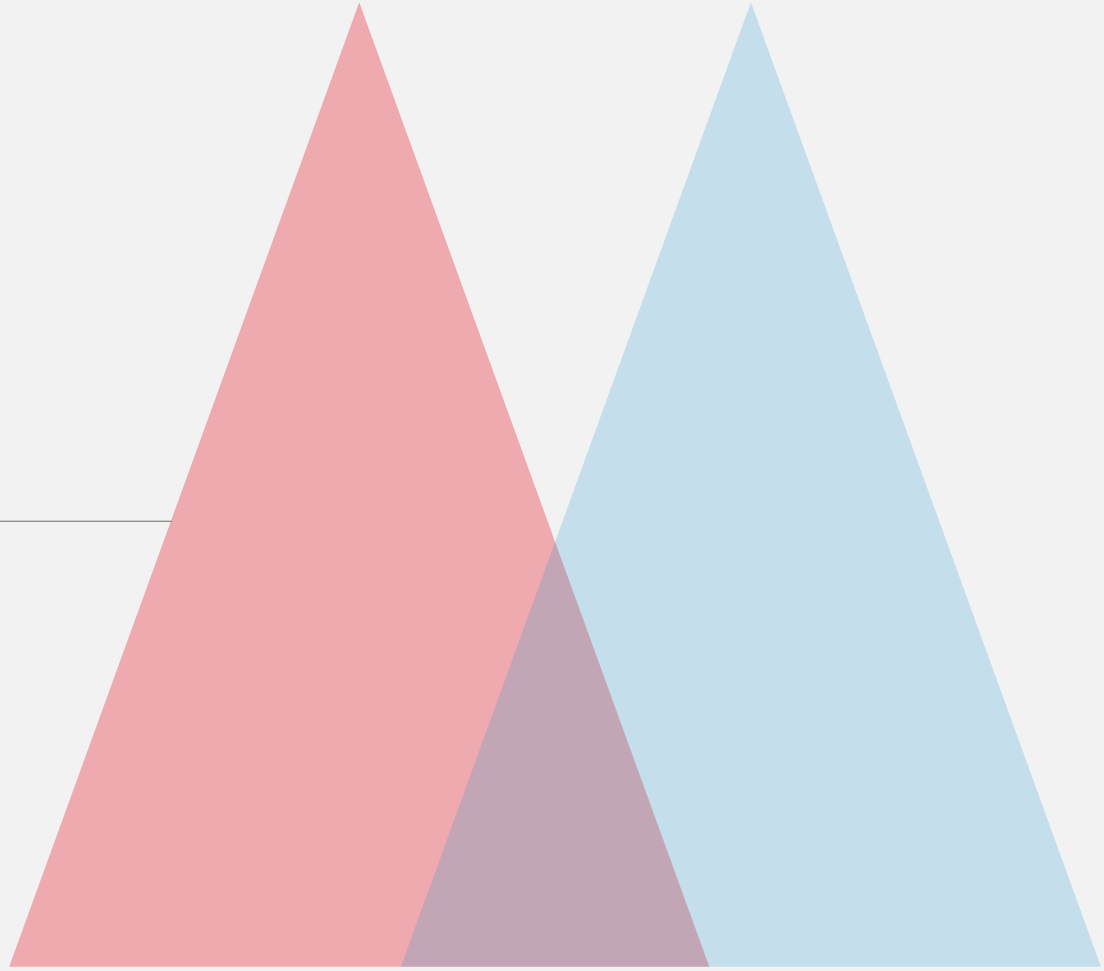
머신러닝을 이용한 댐의 유입량 예측

부문 : 퓨처스리그 홍수ZERO
팀명 : 최적 00
팀장 : 최대원(daeone0920@korea.ac.kr)
팀원 : 임정섭(qaws1000@naver.com)

Contents

001

대회 개요



1

001 대회 개요

문제 설명

주제	댐 유입 수량 예측을 통한 최적의 수량 예측 모형 도출
목적	장마 및 태풍기간 동안 댐 주변 지역 강우량, 수위데이터 분석을 통해 댐에 유입되는 수량(水量)을 예측하여 홍수기 댐 운영 효율화
데이터	제공 데이터: 1~25번 홍수사상 기간 동안 K-댐의 유입량과 관측소의 강우량 및 수위 데이터 외부 데이터: 1~26번 홍수사상 기간 동안 전국 28개 댐의 실제 관측 데이터 평가 데이터: 26번 홍수사상 기간 동안 관측소의 강우량 및 수위데이터
예측 대상	평가 데이터의 26번 홍수사상의 종속변수인 유입량

데이터 설명

홍수사상번호	연	월	일	시간	유입량	데이터집단 1							데이터집단 2						
						유역평균강수	강우(A지역)	강우(B지역)	강우(C지역)	강우(D지역)	수위(E지역)	수위(D지역)	유역평균강수	강우(A지역)	강우(B지역)	강우(C지역)	강우(D지역)	수위(E지역)	수위(D지역)
1	2006	7	10	8	189.1	6.4	7.0	7.0	7.0	8.0	2.5	122.6	6.3	7.0	7.0	7.0	8.0	2.5	122.5
1	2006	7	10	9	217.0	6.3	7.0	8.0	7.0	8.0	2.5	122.6	6.4	7.0	8.0	7.0	8.0	2.5	122.6
1	2006	7	10	10	251.4	6.4	7.0	9.0	7.0	8.0	2.5	122.6	7.3	7.0	9.0	7.0	8.0	2.5	122.6
1	2006	7	10	11	302.8	7.3	7.0	10.0	7.0	8.0	2.5	122.6	8.2	7.0	10.0	8.0	8.0	2.5	122.6
1	2006	7	10	12	384.8	8.2	7.0	12.0	8.0	10.0	2.5	122.6	11.3	9.0	12.0	10.0	10.0	2.5	122.6
1	2006	7	10	13	512.5	11.3	7.0	14.0	10.0	11.0	2.5	122.6	14.4	12.0	14.0	10.0	11.0	2.5	122.6
1	2006	7	10	14	701.5	14.4	9.0	17.0	10.0	14.0	2.5	122.6	16.9	14.0	17.0	15.0	14.0	2.5	122.6

인덱스
홍수사상번호: 홍수 고유번호
연/월/일/시간: 관측 일시

종속변수
유입량: K-댐에 흘러 들어오는 유량

독립변수
유역평균강수: 전체 유역 평균 누적 강수량
강우(A/B/C/D지역): A/B/C/D 관측소 누적 강수량
수위(D/E지역): D/E 관측소 수위

데이터집단
종속변수(유입량)와 상관성이 높은 데이터 집단으로 총 6개로 구성

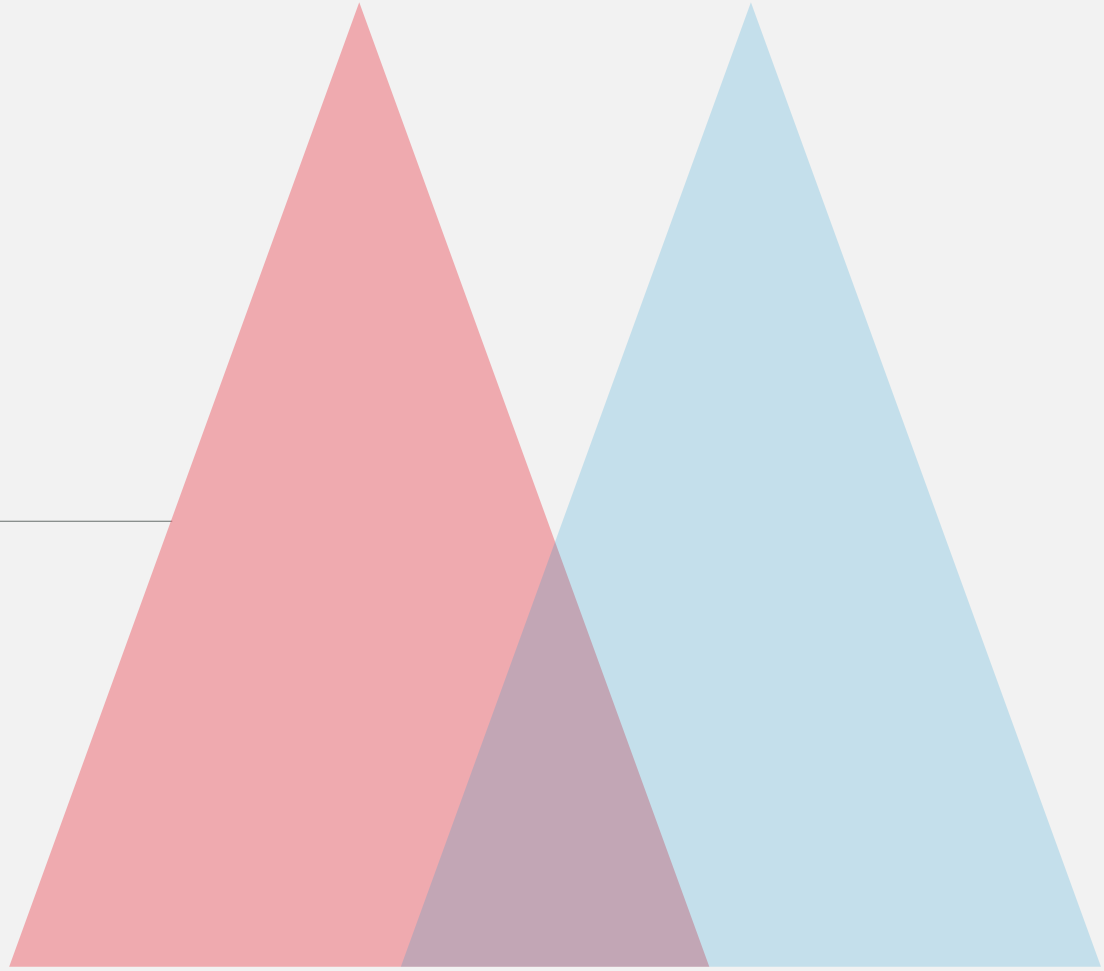
평가 방법

$$RMSE = \sqrt{\sum_{i=1}^{160}(\text{예측값}_i - \text{실제값}_i)^2}$$

- 평가 데이터 기반의 댐 유입 수량에 대한 예측의 정확도 평가
- 26번 홍수사상(2018.7.1. ~ 2018.7.7.)에 대한 예측결과 160개 데이터에 대하여 RMSE(Root Mean Square Error) 평가

002

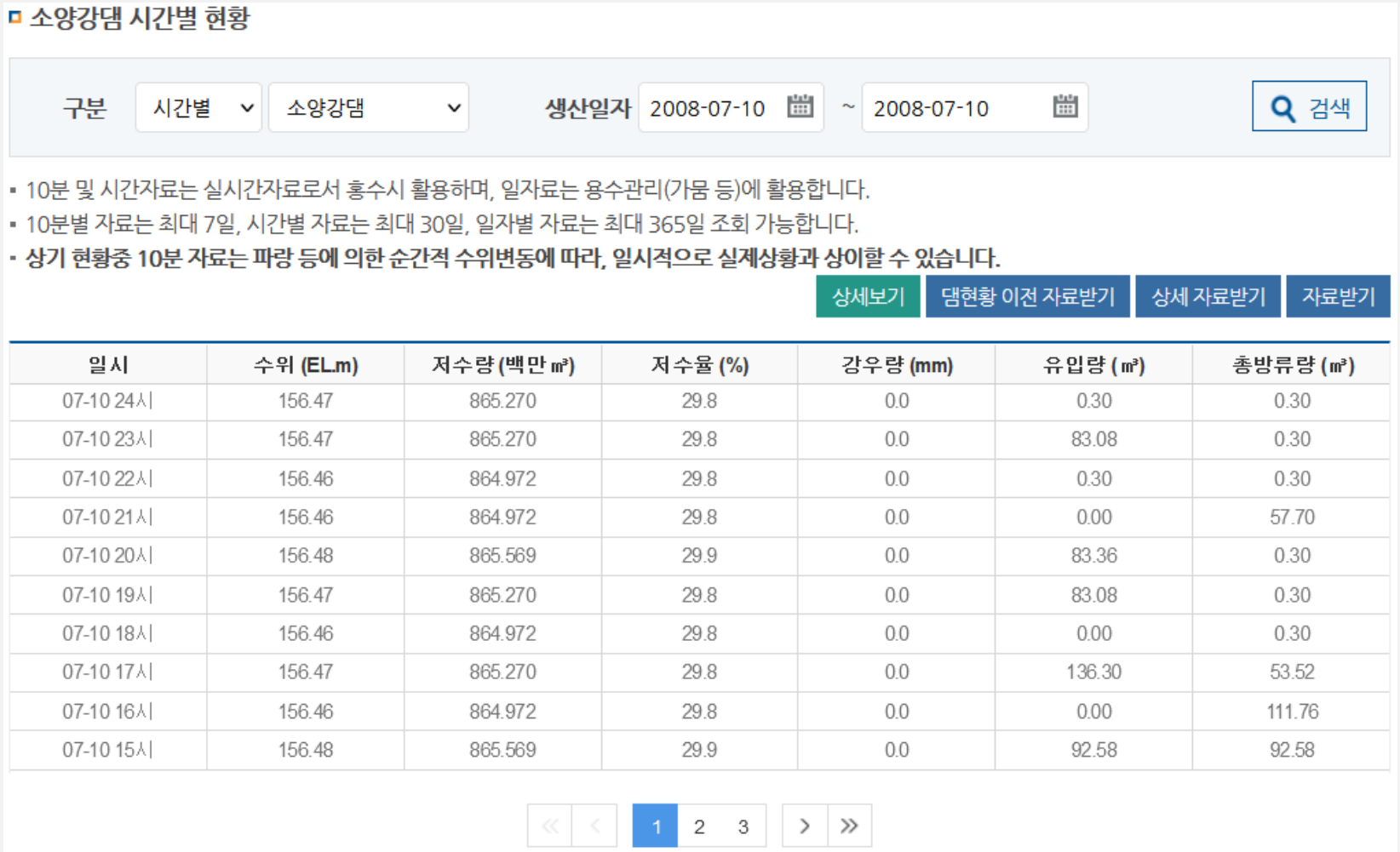
특징 공학



외부 데이터

MyWater(물정보포털) 댐/보 실시간 현황

https://www.water.or.kr/realtime/sub01/sub01/dam/hydr.do?seq=1408&p_group_seq=1407&menu_mode=2



소양강댐 시간별 현황

구분 시간별 소양강댐 생산일자 2008-07-10 ~ 2008-07-10 검색

- 10분 및 시간자료는 실시간자료로서 홍수시 활용하며, 일자료는 용수관리(가뭄 등)에 활용합니다.
- 10분별 자료는 최대 7일, 시간별 자료는 최대 30일, 일자별 자료는 최대 365일 조회 가능합니다.
- **상기 현황중 10분 자료는 파랑 등에 의한 순간적 수위변동에 따라, 일시적으로 실제상황과 상이할 수 있습니다.**

상세보기 | 댐현황 이전 자료받기 | 상세 자료받기 | 자료받기

일시	수위 (EL.m)	저수량 (백만 ㎥)	저수율 (%)	강우량 (mm)	유입량 (㎥)	총방류량 (㎥)
07-10 24시	156.47	865.270	29.8	0.0	0.30	0.30
07-10 23시	156.47	865.270	29.8	0.0	83.08	0.30
07-10 22시	156.46	864.972	29.8	0.0	0.30	0.30
07-10 21시	156.46	864.972	29.8	0.0	0.00	57.70
07-10 20시	156.48	865.569	29.9	0.0	83.36	0.30
07-10 19시	156.47	865.270	29.8	0.0	83.08	0.30
07-10 18시	156.46	864.972	29.8	0.0	0.00	0.30
07-10 17시	156.47	865.270	29.8	0.0	136.30	53.52
07-10 16시	156.46	864.972	29.8	0.0	0.00	111.76
07-10 15시	156.48	865.569	29.9	0.0	92.58	92.58

« ‹ 1 2 3 › »

수집 외부댐 목록 (28개)

- 소양강댐
- 횡성댐
- 안동댐
- 임하댐
- 합천댐
- 남강댐
- 밀양댐
- 충주댐
- 대청댐
- 보령댐
- 용담댐
- 섬진강댐
- 주암댐
- 주암조절지댐
- 부안댐
- 장흥댐
- 광동댐
- 달방댐
- 영천댐
- 안계댐
- 운문댐
- 대곡댐
- 대암댐
- 선암댐
- 연초댐
- 구천댐
- 수어댐
- 평화의댐

수집 외부댐 변수 (6개)

- 수위
- 저수량
- 저수율(%)
- 강우량
- 유입량
- 총방류량

데이터 전처리(1) – 데이터 저장 및 병합

제공 데이터
(평가 데이터)

flood_id	period	연	월	일	시간	데이터집단 1						
						유역평균강수	강우 (A지역)	강우 (B지역)	강우 (C지역)	강우 (D지역)	수위 (E지역)	수위 (D지역)
1	2006-07-10 08:00	2006	7	10	8	6.4	7.0	7.0	7.0	8.0	2.5	122.6
1	2006-07-10 09:00	2006	7	10	9	6.3	7.0	8.0	7.0	8.0	2.5	122.6

외부 데이터

flood_id	period	소양강댐						횡성댐					
		수위	저수량	저수율	강우량	유입량	총방류량	수위	저수량	저수율	강우량	유입량	총방류량
1	2006-07-10 08:00	162.39	1055.024	36.4	0	0.5	0.5	166.73	29.678	34.2	0	0	7.16
1	2006-07-10 09:00	162.4	1055.366	36.4	0	95.5	0.5	166.73	29.678	34.2	0	7.17	7.17

데이터 정보

- 제공 데이터의 행 개수는 총 2891개이다. 평가 데이터의 행 개수는 총 160개이다.
- 제공(평가) 데이터의 독립변수는 데이터집단 6개 별 측정값 7개와 연, 월, 일, 시간으로 총 46개이다.
- 외부 데이터는 시간(hour) 단위로 총 28개 댐의 6개 변수에 대해 수집했으며, 열은 총 168개이다.

데이터 저장 및 병합

1. 제공(평가) 데이터와 외부 데이터를 모두 Pandas DataFrame으로 변환한다.
2. DataFrame의 인덱스는 flood_id(홍수사상번호)와 period(일시)로 설정한다.
3. 두 DataFrame을 flood_id와 period를 기준으로 (내부)조인한다.

데이터 전처리(2) – 이상치 제거

[표] 이상치 판단 기준과 처리 방법에 따른 모델 성능(RMSE)

	IQR	Z-score
이상치 제거	257.74	250.77
이상치 대체*	444.19	426.08

IQR = (Q3 - Q1)일 때, (Q1 - 1.5×IQR)보다 작거나 (Q3 +1.5×IQR)보다 큰 값을 이상치로 정한다.

Z-score의 임계값은 2이다. Z-score의 절댓값이 임계치를 초과하면 이상치로 분류한다.

* 이상치의 대체 값으로 IQR 기준일 때 중앙값을, Z-score 경우에는 평균을 사용했다.

이상치 판단 기준 및 처리 방법 결정

이상치 판단 기준과 처리 방법을 달리한 4가지 경우에 대해 XgBoost 모델로 성능을 평가해보고 가장 우수한 방안을 채택한다. 실험 결과 Z-score와 이상치 제거를 사용할 때 가장 성능이 좋았다.

이상치 제거 과정

1. 병합된 데이터의 독립변수 중에서 종속변수(유입량)과의 상관계수가 가장 높은 11개 열을 선택한다.
2. 선택된 변수 각각을 표준화하여 Z-score를 구한다.
3. Z-score의 절댓값이 2를 초과하면 이상치로 분류한다.
4. 한 개 이상의 변수에 대해 이상치를 포함하는 행은 데이터셋에서 제거한다. (행 개수 3051 → 2796)

특징 추출(1) – 파생 변수

flood_id	period	홍수사상 진행도	홍수사상 번호
1	2006-07-10 08:00	0	1
	2006-07-10 09:00	0.0044	1
	2006-07-10 10:00	0.0089	1
	...	...	...
	2006-07-19 16:00	0.9956	1
	2006-07-19 17:00	1.0000	1
26	...	...	...
	2018-07-07 17:00	0.9748	26
	2018-07-07 18:00	0.9811	26
	2018-07-07 19:00	0.9874	26
	2018-07-07 20:00	0.9937	26
	2018-07-07 21:00	1.0000	26

홍수사상진행도
각 홍수사상별 시간의 경과율을 변수로 추가한다.
0부터 1까지의 실수 범위이다.

홍수사상번호
홍수사상의 번호를 변수로 추가한다.
1부터 26까지의 자연수 범위이다.

- 제공 데이터는 **홍수사상마다 유입량의 변화가 반복되는 시계열 데이터**의 특성을 지닌다.
- 이전에 독립변수에 포함된 연/월/일/시간 값과 이번에 추가한 홍수사상진행도 및 홍수사상번호를 통해 이러한 데이터의 성질을 반영한다.

002 특징 공학

특징 추출(2) – 변수 목록

출처	변수(개수)	변수 개수
제공 데이터	데이터집단(6) × {유역평균강수, 강우 A/B/C/D, 수위 D/E}(7)	42
	연, 월, 일, 시간	4
외부 데이터	외부댐(28) × {수위, 저수량, 저수율, 유입량, 강우량, 총방류량}(6)	168
파생 변수	홍수사상진행도, 홍수사상번호	2
합계	-	216

특징 변환(1) – One-Hot 인코딩, 로그 변환

flood_id
1
2
3
…
26

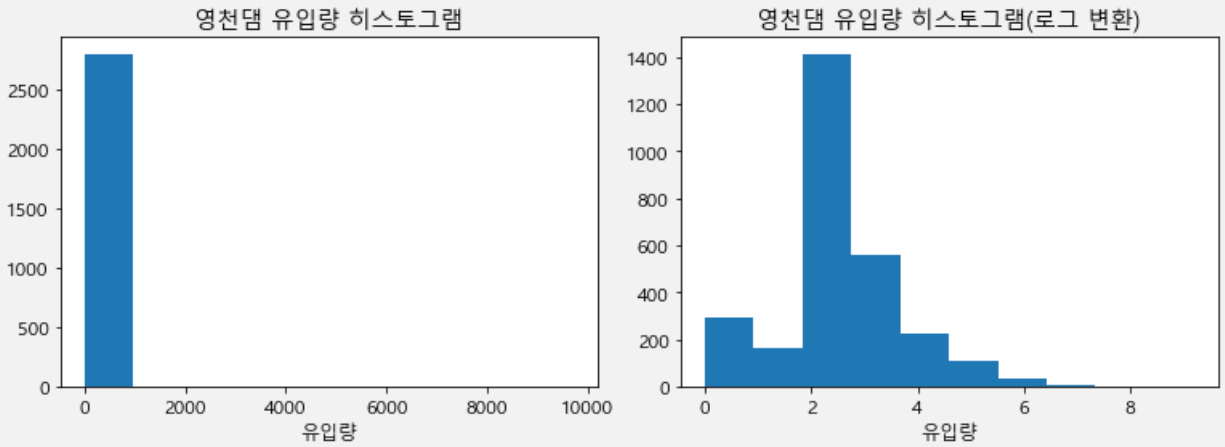
flood_id_1	flood_id_2	…	flood_id_26
1	0		0
0	1		0
0	0		0
…	…		…
0	0		1

One-Hot 인코딩

1~26 범위의 값을 가지는 범주형 변수인 홍수사상번호를 0또는 1의 값을 가지도록 칼럼을 추가하고 수치형 변수로 변환한다.

** 총 특징 개수가 241개로 증가한다.*

[그림] 왜도가 가장 큰 영천댐 유입량의 로그 변환 전후 히스토그램 변화



로그 변환

왜도의 절댓값이 1보다 큰 변수(총 111개)는 1을 더하고 자연로그를 취하여, 치우치거나 편중된 분포를 평탄하게 만든다.

- 연, 월, 일, 시간, 홍수사상번호, 홍수사상진행도는 로그 변환 대상에서 제외한다.

특징 변환(2) – 정규화

스케일링 필요성

- 데이터의 변수들은 단위와 범위가 모두 다르므로, 스케일링(scaling) 과정을 통해 회귀모델에서 각 변수가 동일한 영향을 끼치도록 조정해야 한다.

- 스케일링 기법에는 정규화(normalize)와 표준화(standardize)가 있다.

스케일링 기법 결정

- Xgboost 모델로 성능을 측정한 결과 정규화가 더 우수했으므로, 이를 채택했다.

- 정규화는 어떤 변수의 최댓값과 최솟값을 이용해 0과 1사이의 값으로 변환하는 작업이다. 자세한 공식은 좌측에 있다.

$$x_i\_new = \frac{x_i - min(x)}{\max(x) - \min(x)}$$

정규화 적용 과정

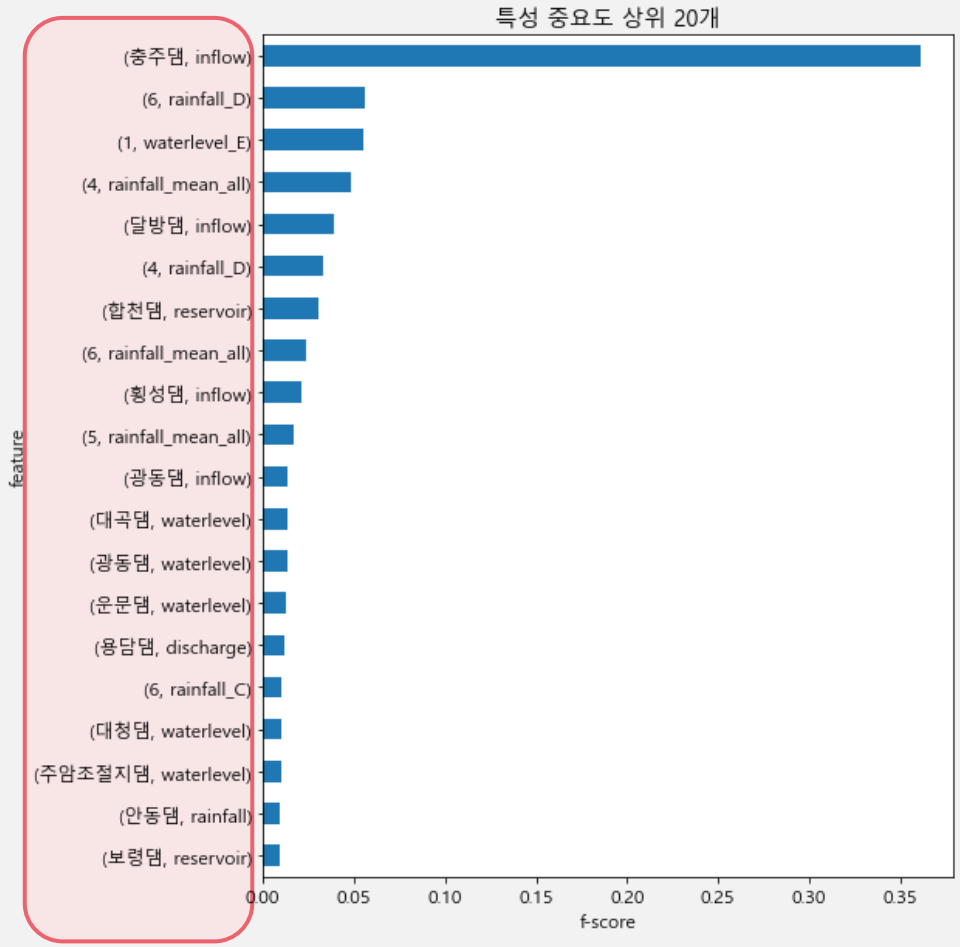
- 예측시점인 26번을 제외한 1~25번 홍수사상 데이터의 분포를 기준으로 최댓값과 최솟값을 구한다.

- 공식을 적용하여 26번 홍수사상을 포함한 전체 데이터를 정규화한다.

2

002 특징 공학

특징 선택



상위 78개 선택!

특징 선택 알고리즘(모델 기반 선택)

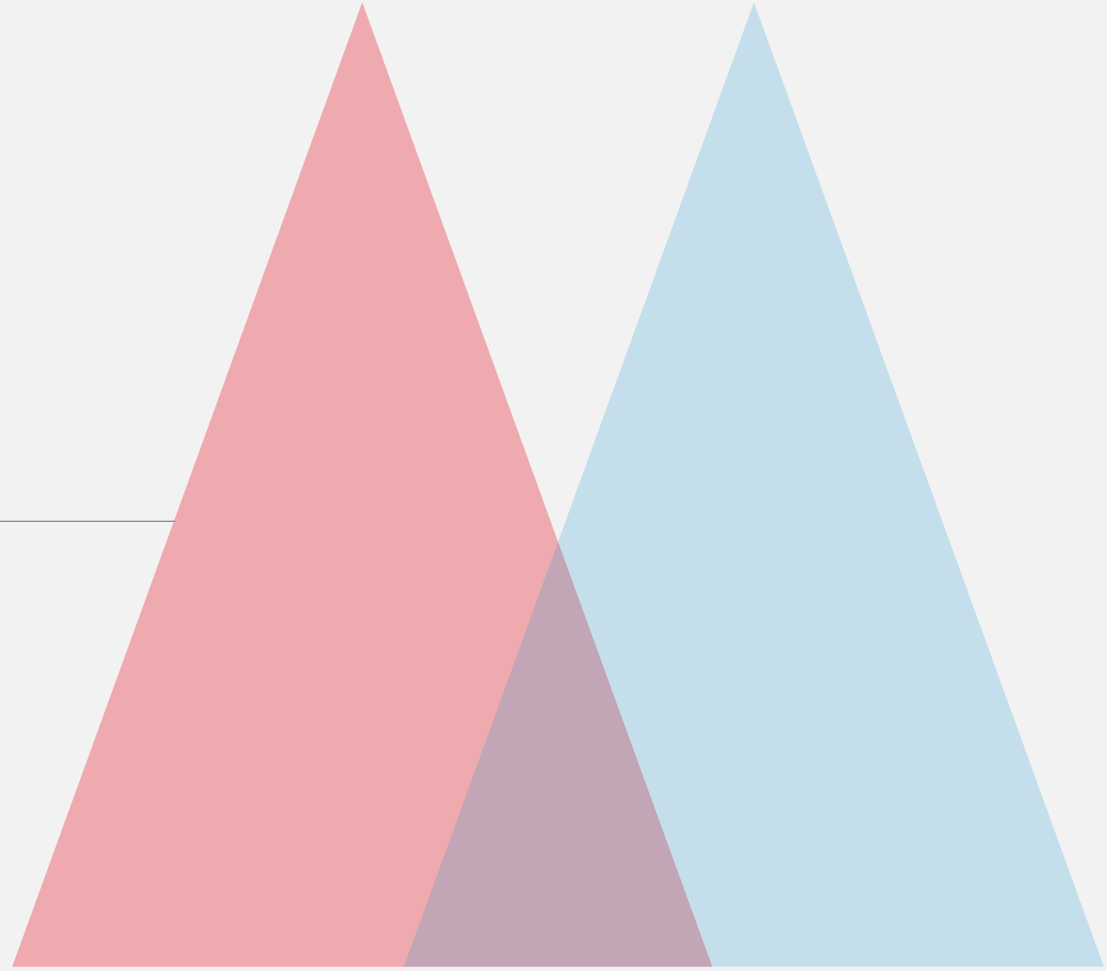
1. 병합된 데이터(1~25번 홍수사상)를 train set과 test set으로 분리한다.
2. Xgboost 모델에 train set를 학습시키고, 각 특징의 중요도를 계산하여 정렬한다.
3. 특징 중요도가 높은 순서대로 변수를 하나씩 추가해가며 train set를 학습시키고, 교차검증을 통해 성능을 평가한다.
4. RMSE가 가장 낮은 경우(최우수 성능일 때)의 변수 조합을 선택한다.

특징 선택 결과

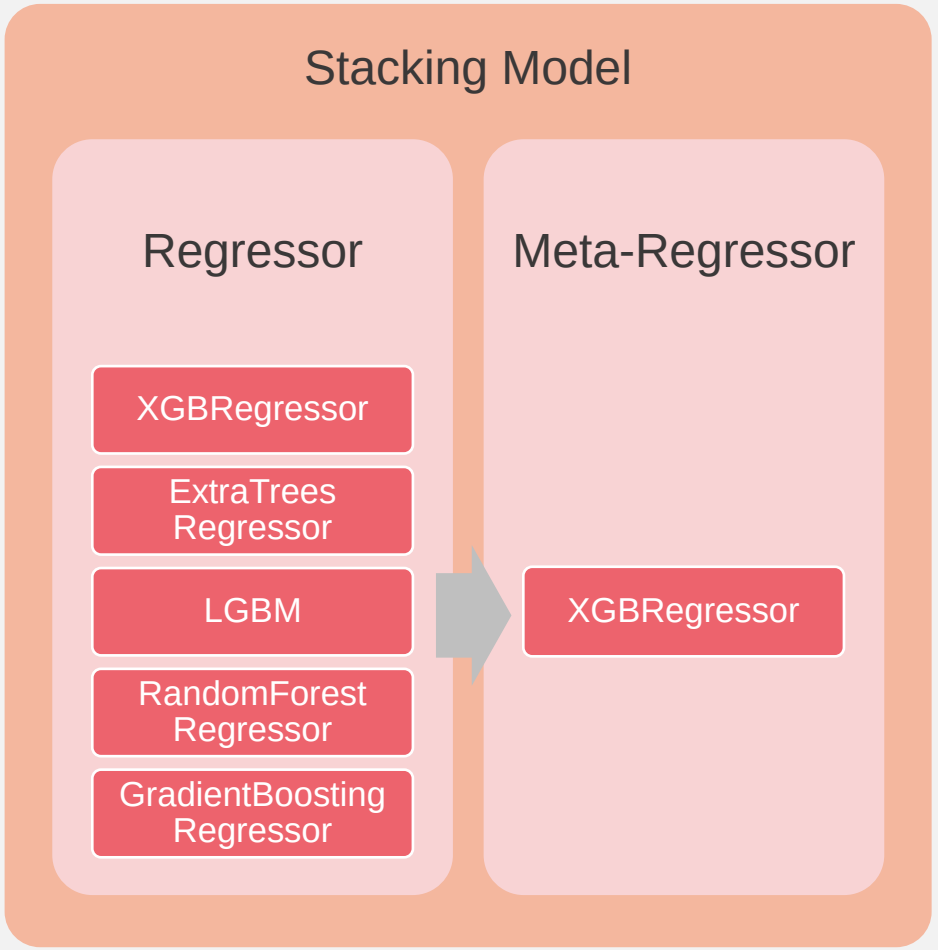
241개의 독립변수 중 78개가 선택되었다.

003

모델링



모델 구축



스태킹(Stacking) 모델

단일 모델을 사용하는 것이 아니라, 여러 모델들을 결합하는 스태킹을 활용한다.

스태킹 모델의 구성

- 여러 regressor 모델이 학습을 통해 예측한 데이터를 취합하고, 최종적으로 Meta-regressor 모델을 사용하여 예측한다.
- 앙상블, 부스팅 계열의 모델을 내부 모델로 사용한다.
- Meta-regressor는 모델 최적화 이후에 선정했으며, 결론적으로 XGBRegressor가 선택되었다.

모델 평가

1단계: [train set] 학습 및 평가

↓

2단계: [train set] 교차검증(5 fold)

↓

3단계: [test set] 평가

모델 평가 과정

- *특징 선택* 과정에서 분리한 train set과 test set을 그대로 사용한다.

- train set은 모델 학습과 교차검증에 사용된다.
- test set은 모델에 학습되지 않으며 최종적인 성능 확인을 위해서만 쓰인다.

- 1단계와 2단계의 결과를 비교하여 모델의 과적합(overfit, underfit)을 파악하고 모델 성능의 판단척도로 삼는다.
- 충분한 모델 최적화 후에 만들어진 최종모델은 마지막으로 3단계 결과를 활용해 실전에서 나타날 최종적인 성능을 판단했다.

모델 최적화(1) – 단일 모델 최적화

모델 최적화 과정(1)

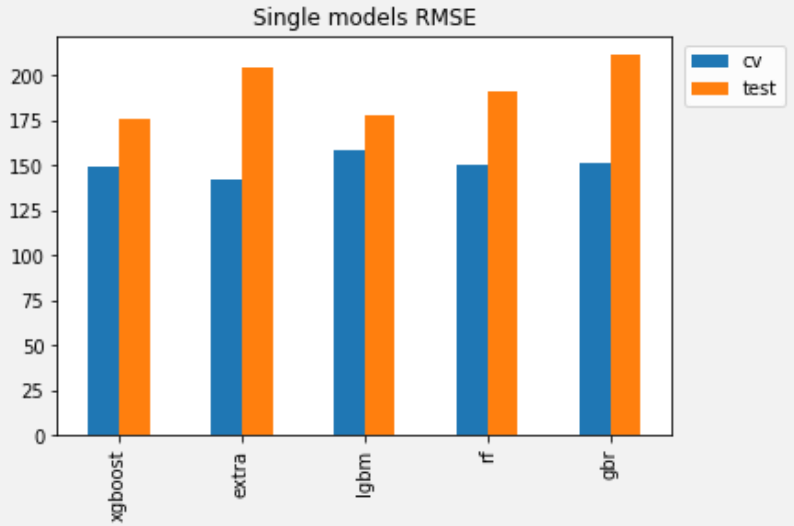
- 스태킹에 앞서 내부 모델 각각을 최적화한다.
- 단일 모델 최적화에는 교차검증을 적용한 Grid Search를 사용했다. (sklearn 라이브러리의 GridSearchCV 함수)

단일 모델 별 최적화에 따른 RMSE 변화 (교차검증)

- Xgboost 214 → 149
- ExtraTree 163 → 143
- LGBM 196 → 158
- RandomForest 232 → 150
- GradientBoosting 218 → 152

train set 교차검증 및 test set 평가 결과(RMSE)

	cv	test
xgboost	149.378102	175.748694
extra	142.591362	204.667971
lgbm	158.190462	178.434810
rf	149.952729	191.106729
gbr	151.834310	211.646004



모델 최적화(2) – Meta-Regressor 결정

모델 최적화 과정(2)

- 최적화된 단일 모델들을 regressor로 모두 스태킹 모델에 포함한다.
- Meta-regressor를 선택하기 위해 5개 모델을 각각 채용했을 때의 결과를 비교한다.
- XGBRegressor를 Meta-regressor로 사용했을 때 성능이 가장 우수하므로 채택한다.

최종모델 결정

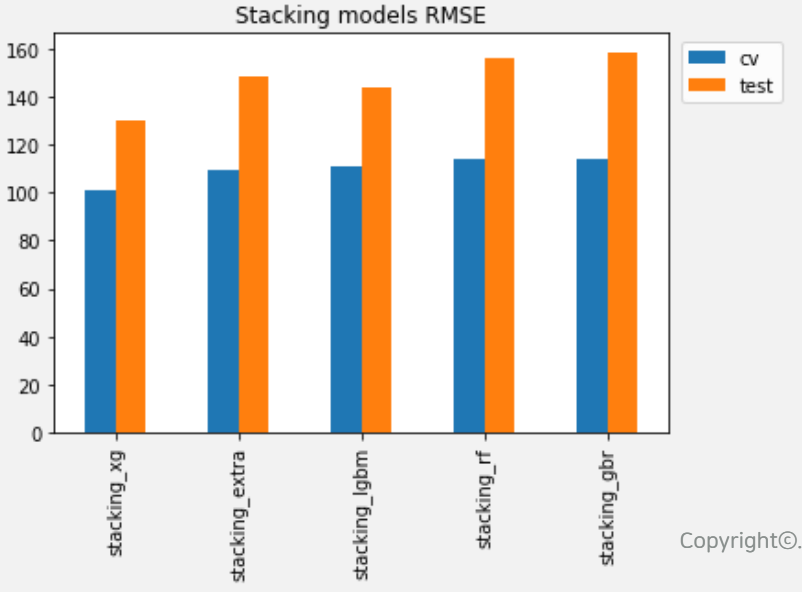
최종모델은 앞서 설명한 스태킹 모델에 최적화된 단일 모델 5개를 regressor로 하고, XGBRegressor를 Meta-regressor로 사용한 모델이다.

test set 예측 결과

최종 모델의 test set 평가에 대한 RMSE는 **130.4**를 기록했다.

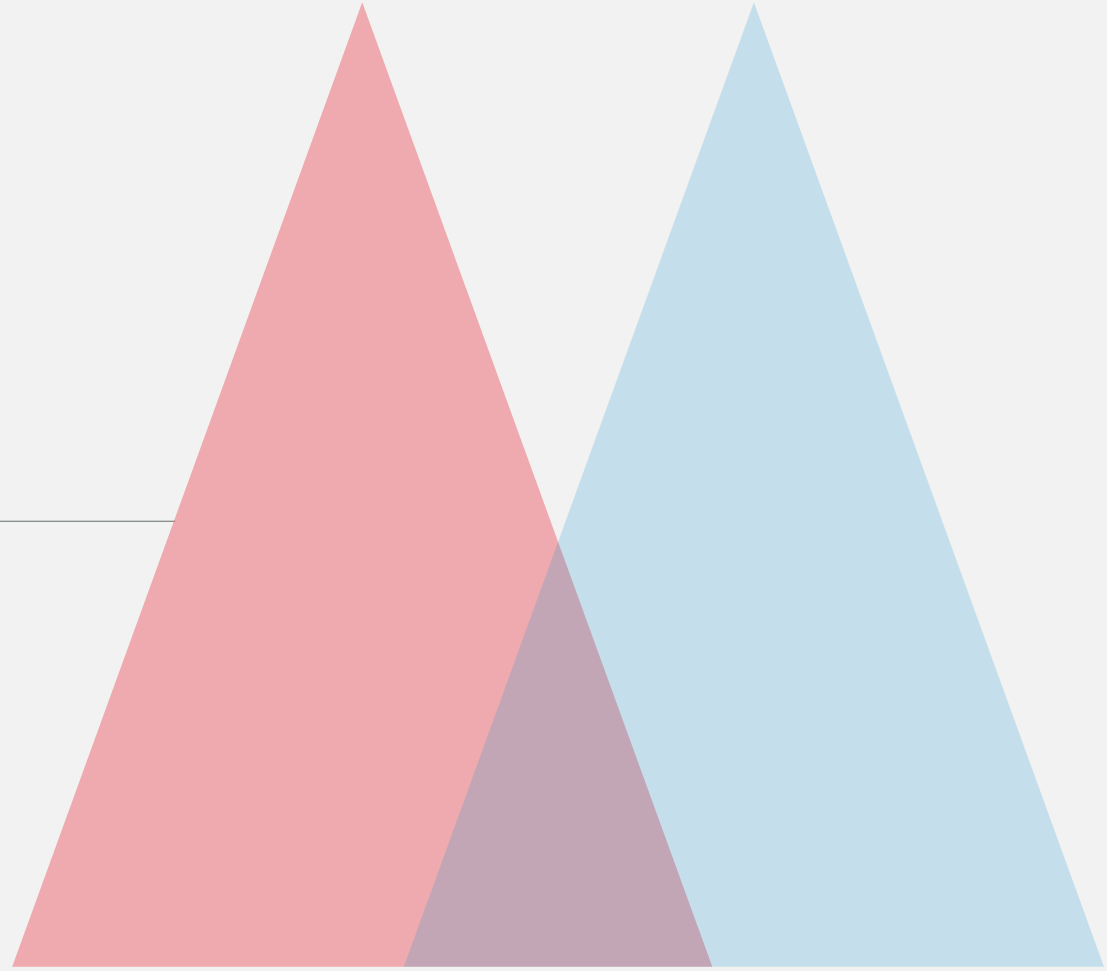
Meta-regressor 변화에 따른 교차검증 및 test set 평가(RMSE)

	cv	test
stacking_xg	101.290417	130.393954
stacking_extra	109.728971	148.373156
stacking_lgbm	111.315662	144.349711
stacking_rf	113.990601	156.635513
stacking_gbr	114.025729	158.957407



004

결론



예측 결과 및 결론

평가 데이터 예측

병합된 데이터(26번 홍수사상)를 전처리하고, 최종 모델을 이용해 종속변수인 유입량을 예측한다.

홍수사상 번호	연	월	일	시간	유입량
26	2018	7	1	6	347.5
26	2018	7	1	7	281.5
26	2018	7	1	8	403.7
26	2018	7	1	9	635.8
26	2018	7	1	10	549.3
26	2018	7	1	11	512.2
26	2018	7	1	12	723.7
26	2018	7	1	13	693.8
26	2018	7	1	14	684.2
26	2018	7	1	15	695.0
26	2018	7	1	16	678.4
...	...	...	...	...	...
26	2018	7	7	21	545.1

결론

머신러닝 모델을 구축하고, 제공데이터와 수집한 외부 데이터를 학습시켜 평가 데이터의 유입량을 예측했다.