

# 빅콘테스트 데이터분석 홍수ZERO 부문

## 결과 보고서

- 댐 유입 수량 예측을 통한 최적의 수량 예측 모형 도출 -

팀명	경통드림팀
팀장	장성민 (jsm9358@naver.com)
팀원	윤성식 (dbsyoon49@naver.com)
	한보혜 (bohaehan@kookmin.ac.kr)
	유광열 (rhkdduf627@naver.com)

# 목 차

<b>1. 가설.....</b>	<b>6</b>
<b>2. 데이터 수집.....</b>	<b>6</b>
1) K댐의 지역 특징.....	6
2) 기상적 관점.....	8
3) 지리적 관점.....	8
4) 지역적 관점.....	9
<b>3. Feature Making.....</b>	<b>9</b>
1) 기본 피처 만들기.....	9
2) 시간의 흐름에 따른 Lag 피처 만들기.....	9
3) 덴드로그램.....	10
<b>4. 데이터 검증.....</b>	<b>15</b>
1) 데이터 검증의 필요성.....	15
2) 통계적 검증.....	15
3) 모델적 검증.....	18
<b>5. Modeling.....</b>	<b>20</b>
1) 전처리.....	20
2) Feature Selection.....	22
3) Default Model.....	23
4) HyperParameter Tuning.....	23
5) 과적합 검증.....	24
6) 학습의 일반화.....	25
7) Ensemble.....	26

6. 분석결과 및 기대효과.....	27
---------------------	----

7. 참고자료.....	31
--------------	----

## 그림목차

[그림 I -1] 유입량 .....	5
[그림 II-1] A 지역 상관관계 .....	10
[그림 II-2] B 지역 상관관계 .....	10
[그림 II-3] C 지역 상관관계 .....	10
[그림 II-4] D 지역 상관관계 .....	10
[그림 II-5] 수위 D 지역 상관관계 .....	10
[그림 II-6] 수위 E 지역 상관관계 .....	10
[그림 II-7] 수위 관측소 상관관계 .....	11
[그림 II-8] 우량 관측소 상관관계 .....	12
[그림 II-9] 우량관측소 덴드로그램 .....	13
[그림 II-10] 수위관측소 덴드로그램 .....	13
[그림 II-11] 유역평균강수 덴드로그램 .....	14
[그림 III-1] 외부데이터들과 종속 변수 사이의 상관계수 .....	15
[그림 III-2] 종속변수와의 상관계수 P-value 예시 .....	16
[그림 III-3] 독립변수들과 종속변수 사이의 MI score 예시 .....	16
[그림 III-4] 상관계수가 0.9 이상인 독립변수들.....	17
[그림 III-5] Feature Importance.....	19
[그림 IV-1] Kde plot으로 확인한 각 피쳐들의 분포 .....	20
[그림 IV-2] log변환 & Standard Scaling 후 분포의 차이.....	21
[그림 IV-3] 과적합 검증방법 .....	24
[그림 IV-4] 홍수사상번호별 유입량 범위 .....	25
[그림 IV-5] n_split별 모델 성능 .....	25
[그림 V -1] XGB SHAP value importance .....	27
[그림 V -2] LGBM SHAP value importance .....	27
[그림 V -3] Cat SHAP value importance .....	27
[그림 V -4] 해당시간 강우량 force_plot .....	28
[그림 V -5] 자체유입 force_plot .....	28
[그림 V -6] 저수량 force_plot .....	29
[그림 V -7] 방수로 수위 force_plot .....	29
[그림 V -8] 기온 force_plot .....	29

[그림 V-8] 수위 데이터 force_plot .....	30
[그림 V-9] 강우 데이터 force_plot .....	30
[그림 V-10] 유량 데이터 force_plot .....	30

## **표목차**

<표 I-1> 다목적댐 .....	6
<표 I-2> 2006이전 설립 다목적댐 .....	7
<표 I-2> 우량 및 수위관측소가 15개 이상 존재하는 다목적댐.....	7
<표 II-1> 2006이전 설립 다목적댐 .....	16
<표 II-2> 상관관계 기반 제거 피처 .....	17
<표 II-3> 통계적 검증 완료된 데이터 .....	18
<표 II-4> 모델적 검증 성능표 .....	18
<표 III-1> scaling여부 비교 성능표 .....	21
<표 III-2> 모델별 Feature Selection 결과.....	21
<표 III-3> Default Model 성능표.....	23
<표 III-4> RandomSearch Tuning 후 성능표.....	23
<표 III-5> 과적합 검증표 .....	24
<표 IV-1> Averaging Ensemble 검증표.....	26

## 1. 가설

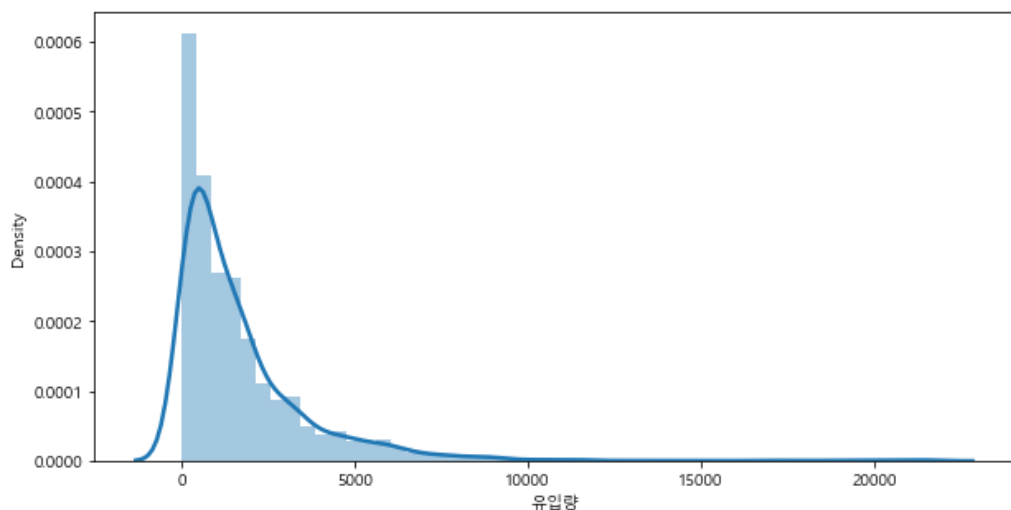
- 자연 현상의 모든 변수를 고려하지 못하더라도 다른 단순한 데이터를 이용하여 유입량 예측이 가능할 것임.
- 수식화 되지 않더라도 데이터 그 자체의 값을 통한 유입량 예측이 가능할 것임.
- 모든 지역에 동일한 계측기가 있지 않더라도 서로 다른 종류의 계측기를 통한 측정값을 이용하여 유입량 예측이 가능할 것임.
  - 즉, 모든 수계에 각기 다른 계측기가 존재할 지라도 각 계측기의 측정값을 각각 하나의 피처로 사용하여 모델을 통한 유입량 예측이 가능할 것이라 생각함.

## 2. 데이터 수집

- 주어진 데이터에서는 A,B,C,D,E 5개 지역에 설치된 **12개의 수위측정기**와 **24개의 우량측정기**를 통해 측정된 수치를 6개의 집단으로 나누어 제공하고 있음.
- 홍수사상시 유입량을 예측하기 위하여 유입량에 수위와 우량을 제외하고 영향을 미칠 수 있는 요인들에 대한 탐색을 크게 **기상적, 지리적, 지역적 관점**으로 나누어 진행함.
- 이때, 해당 관점의 데이터수집을 위해서는 제공된 **K댐에 대한 어느 정도의 지역 특징**이 필요하다고 판단함.

### □ K댐의 지역 특징

- 유입량 탐색



- 주어진 target 값인 유입량의 범위를 탐색해본 결과 적게는  $3.5m^3$ 에서 많게는  $21504.4m^3$ 의 강수가 유입되는 것으로 확인되었음.
- 이는 해당 댐에 상당히 많은 양의 강수가 유입될 수 있다는 것으로 댐의 종류 중 상대적으로 규모가 큰 21개 다목적댐으로 후보지를 좁힐 수 있었음.

댐 종류	댐 이름 (완공년도)
다목적댐	<ul style="list-style-type: none"> <li>• 소양강댐 (1973)</li> <li>• 충주댐 (1985)</li> <li>• 횡성댐 (2000)</li> <li>• 안동댐 (1976)</li> <li>• 임하댐 (1993)</li> <li>• 합천댐 (1988)</li> <li>• 남강댐 (1970)</li> <li>• 밀양댐 (2001)</li> <li>• 군위댐 (2010)</li> <li>• 김천부항댐 (2013)</li> <li>• 영주댐 (2016)</li> <li>• 성덕댐 (2015)</li> <li>• 보현산댐 (2016)</li> <li>• 대청댐 (1981)</li> <li>• 용담댐 (2001)</li> <li>• 섬진강댐 (1965)</li> <li>• 주암댐 (1991)</li> <li>• 부안댐 (1996)</li> <li>• 보령댐 (1999)</li> <li>• 장흥댐 (2006)</li> <li>• 낙동강하굿둑댐 (1987)</li> </ul>

<표 I -1> 다목적댐

○ 주어진 데이터의 수집시기

- 주어진 데이터는 2006년 7월 10일 에서 2018년 7월 7일까지의 데이터를 제공하고 있으므로 다목적댐 중 2006년 이전에 설립된 16개 댐으로 후보지를 좁힐 수 있었음.

댐 종류	댐 이름 (완공년도)
다목적댐	<ul style="list-style-type: none"> <li>• 소양강댐 (1973)</li> <li>• 대청댐 (1981)</li> <li>• 충주댐 (1985)</li> <li>• 용담댐 (2001)</li> <li>• 횡성댐 (2000)</li> <li>• 섬진강댐 (1965)</li> <li>• 안동댐 (1976)</li> <li>• 주암댐 (1991)</li> <li>• 임하댐 (1993)</li> <li>• 부안댐 (1996)</li> <li>• 합천댐 (1988)</li> <li>• 보령댐 (1999)</li> <li>• 남강댐 (1970)</li> <li>• 장흥댐 (2006)</li> <li>• 밀양댐 (2001)</li> <li>• 낙동강하굿둑 (1987)</li> </ul>

<표 I -2> 2006이전 설립 다목적댐

○ 우량 및 수위 관측소의 존재

- 추려진 다목적댐 중 우량 및 수위관측소가 존재하며, 관측소의 개수가 총 15개 이상인 6개의 댐으로 후보지가 좁혀짐.

댐 종류	댐 이름 (완공년도)
다목적댐	<ul style="list-style-type: none"> <li>• 소양강댐 (1973)</li> <li>• 대청댐 (1981)</li> <li>• 충주댐 (1985)</li> <li>• 안동댐 (1976)</li> <li>• 임하댐 (1993)</li> <li>• 남강댐 (1970)</li> </ul>

<표 I -3> 우량 및 수위관측소가 15개 이상 존재하는 다목적댐

○ 6개의 다목적댐 중 주어진 데이터의 관측소 수와 댐의 규모가 가장 비슷한 충주댐을 중심

으로 해당 댐으로 유입되는 수계가 있는 **강원도 및 충청북도 지역**을 특정하고 해당 지역과 충주댐을 기준으로 데이터 수집을 진행함.

## □ 기상적 관점

- 홍수사상시 유입량에 가장 큰 영향을 주는 요소는 강우라 생각하여 기상요소 중 강우에 영향을 미칠 수 있는 11개 요소에 대해 2006 ~ 2018년까지의 데이터를 수집함.
- 강원도와 충청북도에 위치한 기상관측소 각각의 데이터를 수집하여 평균한 값을 사용함.
  - 수집한 데이터 : 기온, 풍속, 습도, 증기압, 이슬점온도, 해면기압, 지면온도, 현지기압, 전운량, 중하층운량, 최저운고
  - 이 중 결측값이 80%가 넘는 전운량, 중하층운량, 최저운고 데이터를 제외함.
  - 해면기압의 산출과정에서 사용되는 수치인 현지기압 데이터를 제외함.
  - 따라서 **기온, 풍속, 습도, 증기압, 이슬점온도, 해면기압, 지면온도 총 7개** 데이터를 사용함.
- 수집된 데이터가 유입량에 영향을 미치는 이론적 근거
  - 온도 : 수문의 기상적변화에 따른 유출량을 연구한 논문 연구결과에 따라 홍수사상에 영향을 미칠 것으로 생각됨.(7-2)
  - 풍속 : 풍속이 강할수록 강우에 따른 유출량이 늘어난다는 논문 연구결과에 따라 홍수사상에 영향을 미칠 것으로 생각됨. (7-1)
  - 증기압 : 물의 증발산에 영향을 주는 인자로 물과 공기의 온도, 기압, 수질, 증발표면 현상에 영향을 받으므로 강수와 관련이 있을 것으로 판단함. (7-1)
  - 이슬점온도 : 기온 하강에 의해 공기가 포화상태에 이르는 온도로 포화상태에 이르게 되면 강수가 발생하므로 관련이 있을 것으로 판단함. (7-2)
  - 해면기압 : 높이에 따른 기압이 다르다는 것을 고려한 수치로 실제 기상상태를 분석하는데 중요하게 이용되는 수치로 관련이 있을 것으로 판단함.
  - 지면기압 : 지면온도에 따라 증발하는 물의 양이 달라지고 날씨의 영향으로 쉽게 바뀌는 수치이므로 관련이 있을 것으로 판단함. (7-2)

## □ 지리적 관점

- 댐은 강수가 유입되는 수계가 존재하며 각 댐마다 연결되어 있는 수계는 다르므로 특정된 지역의 수계에 대한 데이터가 필요하다고 판단함.



- 따라서 수계를 기준으로 특정된 지역 중 세 지점에 대한 관련 데이터를 각각 수집함.
  - 수집한 데이터 : 강우량, 누계 강우량, 방류량, 저수위, 저수량, 저수율
  - 세 지점 중 특정한 지역에 가장 많은 영향을 준다고 판단되는 충주댐지역 데이터를 선택하여 사용함.

## □ 지역적 관점

- 홍수사상은 댐이 존재하는 해당 지역의 강우량과 댐의 수문현황에도 영향을 받을 것이라 판단하여 관련 데이터를 수집함.
- 특정된 지역에 존재하는 5개 지역 유량관측소의 유량데이터를 수집함.
  - 수집한 데이터 : 단양, 영월, 우안, 제천, 충주 5개 지역의 유량관측소 데이터를 수집함.
- 충주댐을 기준으로 해당 지역 댐의 수문현황에 대한 데이터를 수집함.
  - 수집한 데이터 : 댐수위, 방수로 수위, 해당지역의 시간당 강수량, 자체유입량, 총방류량 , 수문의 총 6개 데이터를 수집함.
- 수집된 데이터가 유입량에 영향을 미치는 이론적 근거
  - 기존에 주어진 관측 수치데이터 이외에도 유입량을 관측하는데 도움이 되는 다른 관측소 데이터가 필요하다고 판단함.
  - 단위 시간 동안에 흐르는 유체의 양을 측정하는 유량은 수위, 우량과 함께 유입량을 측정하는데 필요한 데이터라고 판단함.

## 3. Feature Making

### □ 기본 피쳐 만들기

- 확정된 최종적인 데이터에서 파생될 수 있는 '방류량\_대비\_유입량', '강우량\_증감율', 유입량\_대비\_방류량', '댐\_대비\_방수로수위' 피쳐를 생성하였음.
- 해당 피쳐들을 사용하였을 때, 모델의 성능이 크게 향상되지 않아 사용하지 않기로 결정함.

### □ 시간의 흐름에 따른 Lag 피쳐 만들기

- 기상데이터

- 기온, 풍속, 습도, 증기압, 이슬점온도, 해면기압, 지면온도 에 대하여 각각 **1,3,6,9,12,24시간 전, 1,3,6,9,12,24 시간 전과의 차이, 1,3,6,9,12,24 시간 전 대비 증감을 피쳐** 총 105개를 생성함.

#### ○ 유량데이터

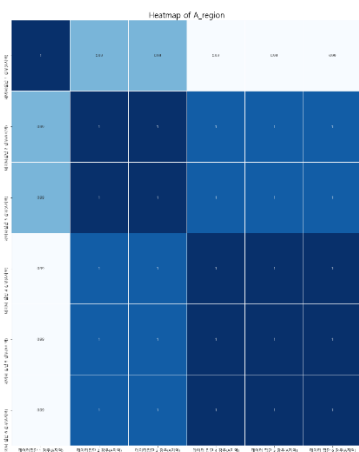
- 앞서 수집한 5개 지역의 유량데이터 및 수위 데이터에 대하여 **1,3,6,9,12,24,48,72 시간 전과의 차이, 1,3,6,9,12,24,48,72 시간 전 대비 증감을 피쳐** 총 95개를 생성함.

○ Lag 피쳐들을 사용하였을 때, 모델의 성능이 크게 향상되지 않아 **사용하지 않기로 결정함.**

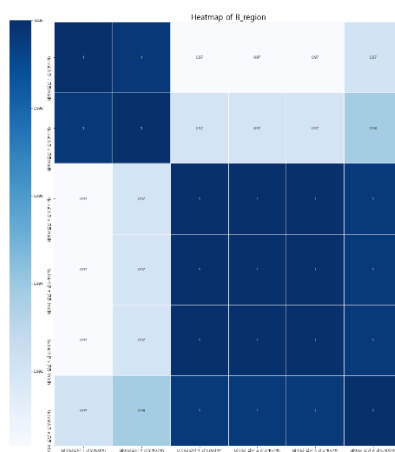
### □ 덴드로그램

○ 6개의 데이터 집단은 모두 동일한 5개 지역의 강우, 수위관측소에서 댐 구간 거리, 시간 등을 달리 설정하여 얻은 독립변수 집단임.

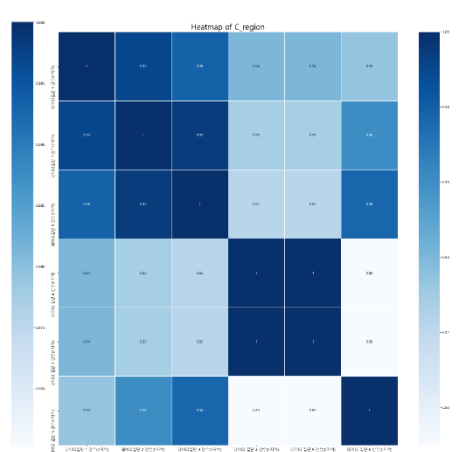
○ **같은 지역에서** 다른 거리, 시간에서 측정된 6개 관측치는 **서로 상관성이 매우 높음.**



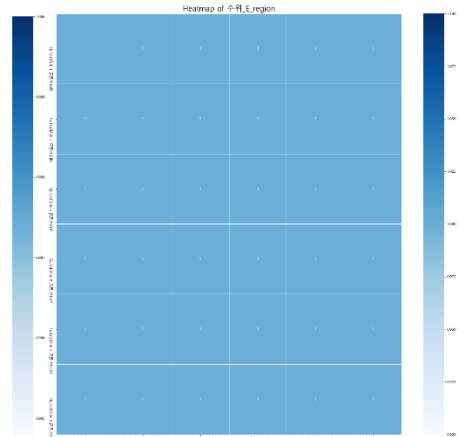
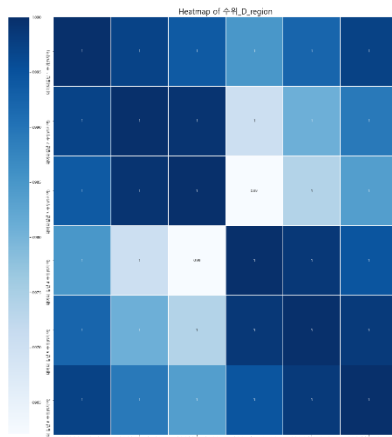
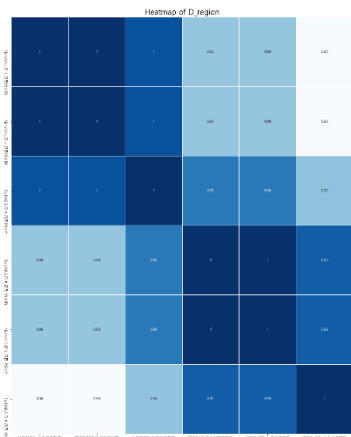
[그림 II-1] A 지역 상관관계



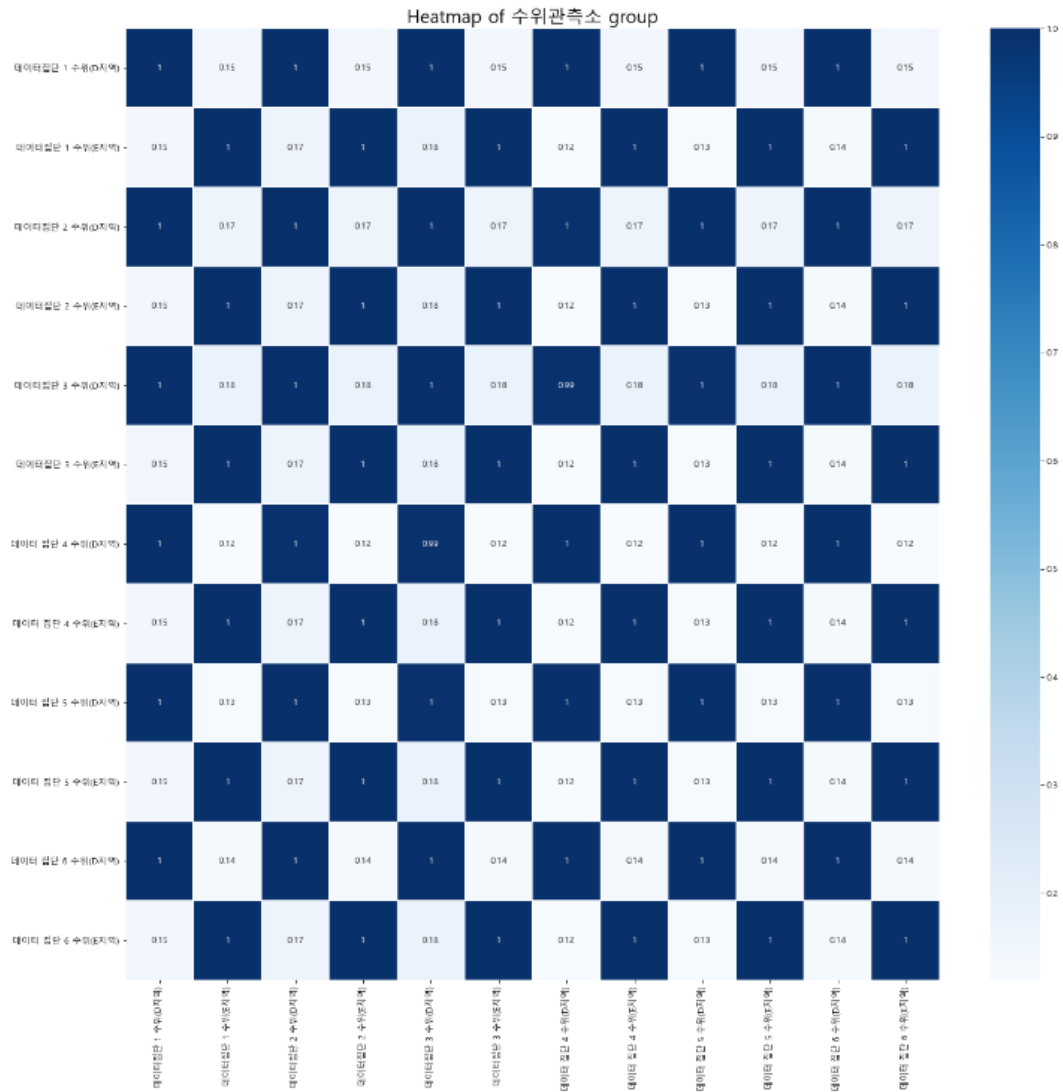
[그림 II-2] B 지역 상관관계



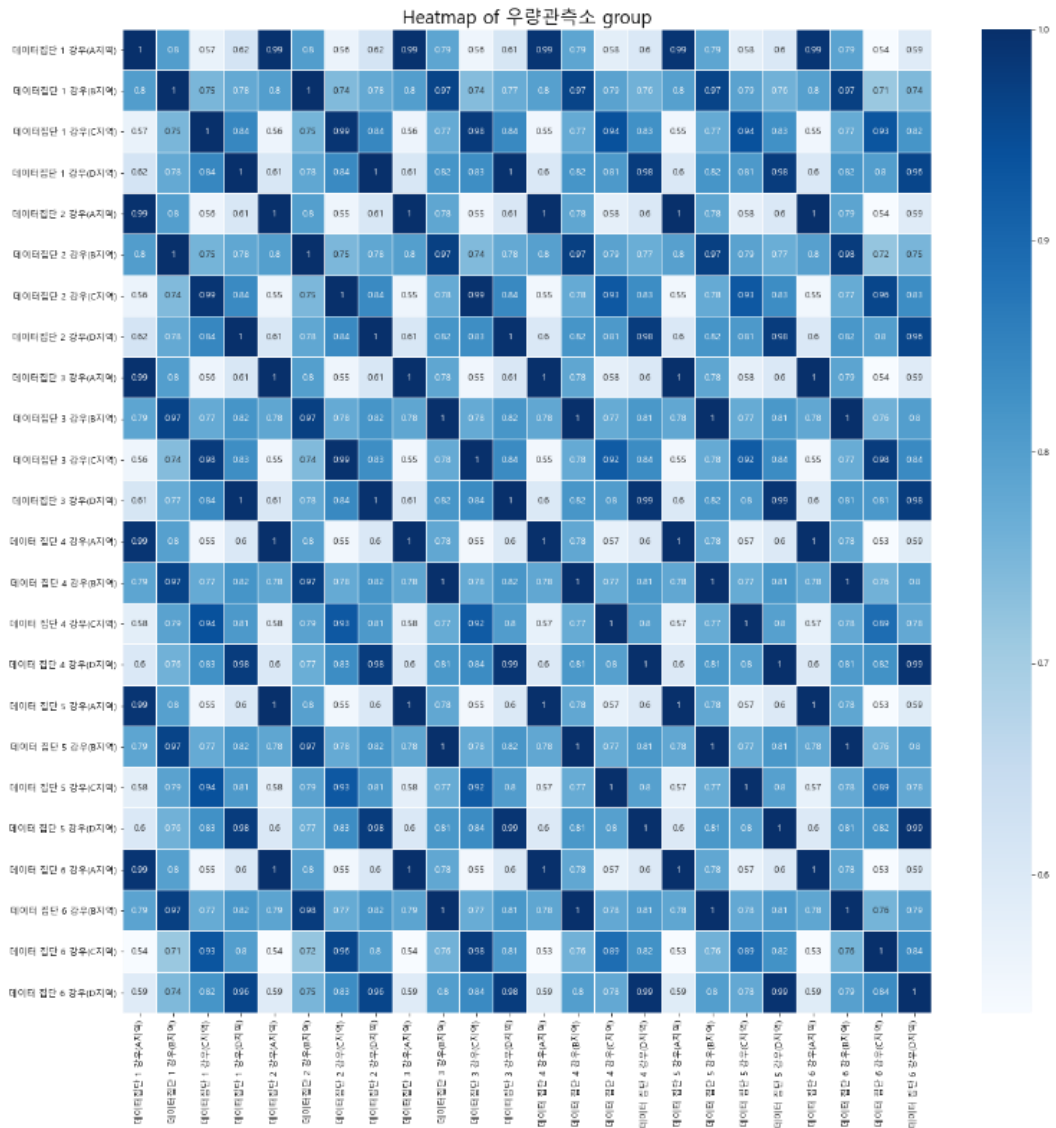
[그림 II-3] C 지역 상관관계



○ 하지만, 24개의 우량관측소와 12개의 수위관측소 관측치들 간의 상관성에서 **지역간 관측치의 상관성은 유의미한 차이를 보임.**

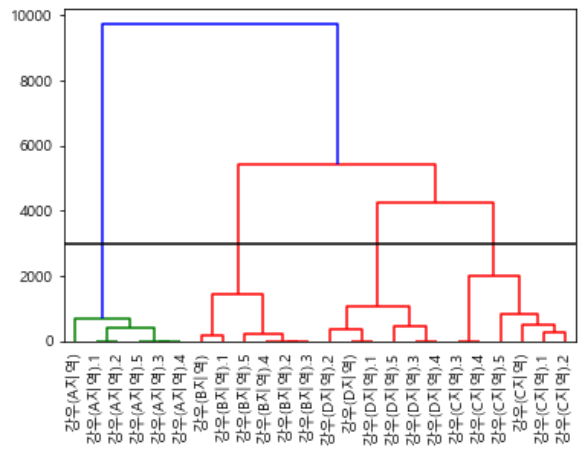


[그림 Ⅱ-7] 수위 관측소 상관관계



- 이를 통해 같은 지역에서 측정된 관측치를 평균하여 하나의 피처로 사용하는 것이 용이한 지 확인하기 위하여 같은 관측소 간의 Hierarchical clustering을 통한 clustering을 진행함.
  - 트리모형을 기반으로 하면서 우리의 주관적 해석을 통해 군집의 개수를 설정할 수 있는 계층적 군집화를 선택하여 clustering을 진행함.
- 우량관측소 24개, 수위관측소 12개, 유역평균강수 6개에 대한 각각의 clustering을 진행하며, 각 관측치 사이의 유사성 계산 방법으로 ward's method를 선택하여 진행함.
  - Feature를 만들기 위해 해당 지역이 가지는 정보 손실을 최소화하는 것이 중요하다고 판단하여 ward method를 선택함.

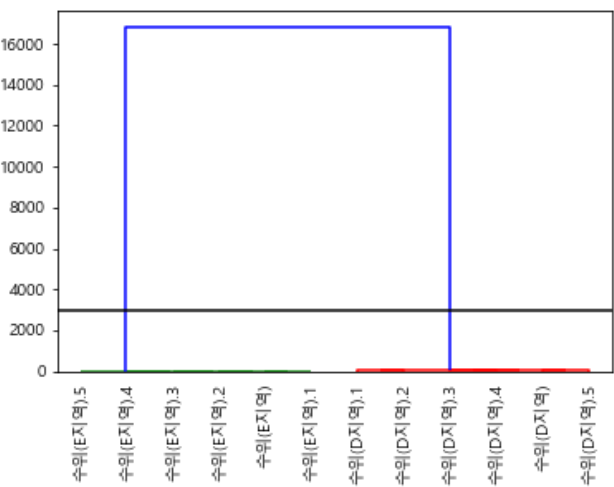
○ 24개 우량관측소



[그림 Ⅱ-9] 우량관측소 덴드로그램

- Clustering 결과 같은 지역이 같은 군집으로 군집화 된 것을 확인할 수 있음.
- 해당 군집화에서 지역끼리 군집화 되면서 큰 집단으로 묶이는 4개의 군집으로 집단을 결정하고 **강우\_0,1,2,3** 피처를 생성함.

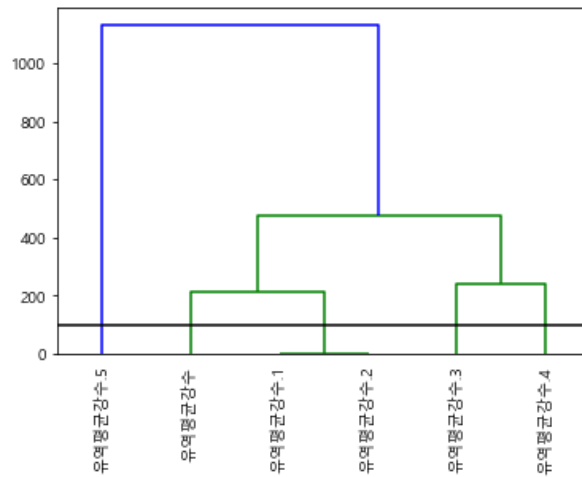
○ 12개 수위관측소



[그림 Ⅱ-10] 수위관측소 덴드로그램

- Clustering 결과 수위관측소 역시 같은 지역끼리 군집화 된 것을 확인할 수 있음.
- 같은 지역끼리 묶이면서 큰 집단으로 묶이는 2개의 군집으로 집단을 결정하고 **수위\_0,1** 집단 피처를 생성함.

○ 6개 유역평균강수



[그림 II-11] 유역평균강수 덴드로그램

- 유역평균강수의 군집화 결과 데이터집단 2,3,이 하나의 군집을 이루는 것을 알 수 있음.
- 유역평균강수는 최대한 각 집단의 관측치를 가지고 가는 것이 좋다고 생각하여 집단을 5개로 결정하고 **유역평균강수\_0,1,2,3,4집단 피쳐**를 생성함.

○ Clustering을 통해 데이터집단에서 같은 관측소와 지역 간의 상관성을 통해 하나의 집단으로 묶어 **총 11개의 피쳐**를 생성함.

## 4. 데이터 검증

### □ 데이터 검증의 필요성

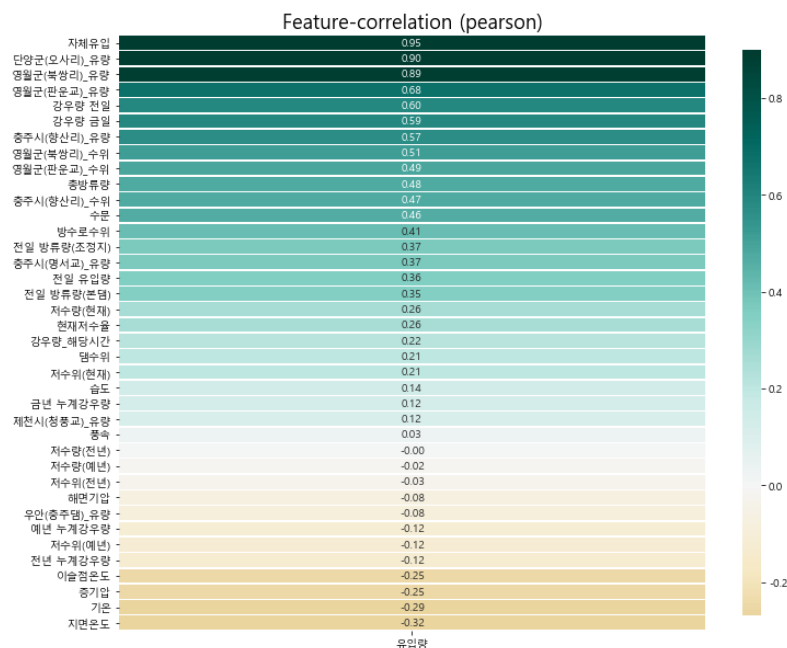
○ 수집된 데이터 및 생성한 피처는 여러가지 분석을 통하여 특정된 지역 및 댐을 기준으로 수집되었으므로 지역이 얼마나 잘 특정되었는가에 따라 예측 오차의 범위가 매우 크게 달라질 것으로 생각됨.

○ 따라서 수집된 데이터와 생성한 피처가 주어진 홍수사상 26번을 예측하기에 적절한 데이터인지 대한 철저한 검증이 필요하다 판단하여 **통계적, 모델적 검증을 진행함.**

### □ 통계적 검증

○ 독립변수와 종속변수 사이의 관계 검증

- 수집된 데이터들이 종속변수(유입량)와 선형의 상관관계를 갖는지를 파악하기 위해 **피어슨 상관계수를 통해 검증을 진행함.**



[그림Ⅲ-1] 외부데이터들과 종속 변수 사이의 상관계수

- 독립변수와 종속변수 사이의 상관계수가 유의미하고 신뢰할 만한 것인지 판단하기 위해 **상관계수에 대한 P-value를 확인함.**

- P-value가 0.05를 넘으면 해당 상관계수가 신뢰할 수 없는 것으로 판단함.

- 상관계수가 높은 독립변수들의 P-value가 충분히 낮은 것을 확인하여 해당 상관계수가 신뢰할 만 하다는 결론을 도출함.

	Feature_names	상관계수	P-value
0	단양군(오사리)_유량	0.879317	0.000000e+00
1	영월군(복상리)_수위	0.497524	1.155465e-180
2	영월군(복상리)_유량	0.865969	0.000000e+00
3	영월군(판운교)_수위	0.511682	1.057282e-192
4	영월군(판운교)_유량	0.713733	0.000000e+00
5	우안(충주댐)_유량	0.405460	7.787278e-115

[그림Ⅲ-2] 종속변수와의 상관계수 P-value 예시

#### ○ Select K Best(MI Score)를 통한 검증

- 앞서 외부데이터와 종속변수 간의 선형관계가 존재하는지 파악했다면, **비선형적인 관계가 존재하는지 파악하기 위해** 선형성 파악보다 더 일반적으로 두 변수 간의 종속관계를 확인할 수 있는 **Mutual Information Regression Score**를 사용해 검증을 진행함.
- Mutual Information Regression Score 가 0.8 이하일 시 해당 독립변수와 종속변수 간의 관계가 없다고 판단함.

	Feat_names	Mutual_Scores
0	단양군(오사리)_유량	1.612358
37	자체유입	1.393806
2	영월군(복상리)_유량	1.340836
4	영월군(판운교)_유량	1.298613
21	금년 누계강우량	1.277805

[그림Ⅲ-3] 독립변수들과 종속변수 사이의 MI score 예시

#### ➔ 검증결과

독립변수들과 종속변수 사이의 선형적 상관관계와 비선형적 상관관계를 모두 파악한 결과 **pearson 상관계수**에서 절댓값이 0.2를 넘지 못하고 **Mutual Information score**에서 0.8을 동시에 넘지 못하는 독립변수를 찾아 제거함.

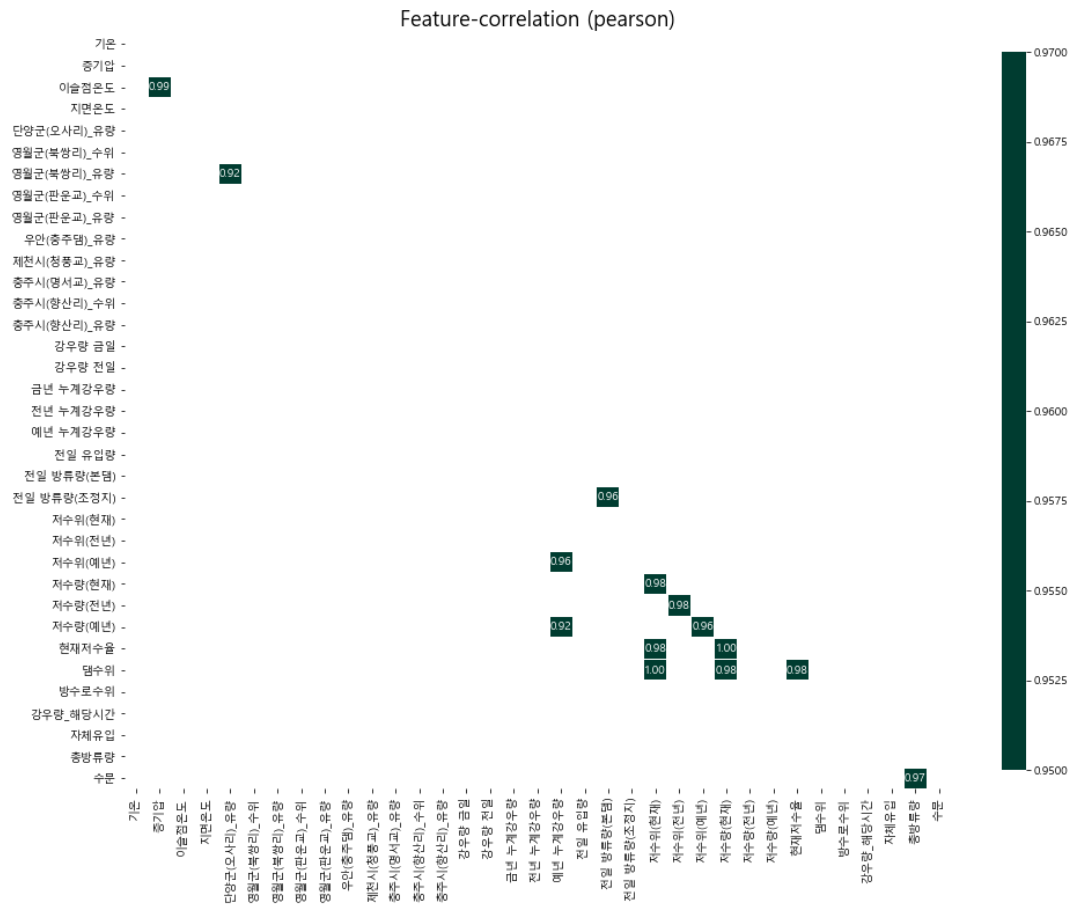
제거 대상 피쳐(3개)
<ul style="list-style-type: none"> <li>· 해면기압</li> <li>· 풍속</li> <li>· 습도</li> </ul>

<표 Ⅱ-1> 2006이전 설립 다목적댐



## ○ 독립변수 간의 상관관계 검증

-독립변수들 간의 상관관계가 너무 높을 경우 발생하는 다중공선성 문제를 해결하기 위해 피어슨 상관계수를 통해 이를 확인하고 제거하고자 함.



[그림III-4] 상관계수가 0.9 이상인 독립변수들

## ➔ 검증결과

상관계수가 0.9 이상인 독립변수 쌍을 다수 발견하여, 둘 중 하나의 변수를 제거함. 제거 시에는 두 독립변수 중 종속변수와 상관관계가 더 낮은 독립변수를 우선적으로 제거함.

제거 대상 피쳐 (10개)	
· 현재저수율	· 수문
· 영월군(북쌍리)_유량	· 저수위(현재)
· 전일 방류량(본댐)	· 예년 누계강우량
· 저수량(전년)	· 저수위(예년)

<표II-2> 상관관계 기반 제거 피쳐

○ 통계적 검증 후 선택된 최종 독립변수

- 독립변수와 종속변수들의 선형성 및 비선형성을 파악하고, 독립변수들 간의 상관관계를 확인하여 유용성이 검증된 23 개의 외부데이터를 분석에 사용하고자 함

통계적 검증이 완료된 외부데이터 (25개)		
• 단양군(오사리)_유량	• 영월군(북쌍리)_수위	• 영월군(판운교)_수위
• 영월군(판운교)_유량	• 우안(충주댐)_유량	• 충주시(명서교)_유량
• 충주시(향산리)_유량	• 기온	• 이슬점온도
• 지면온도	• 강우량 금일	• 금년 누계강우량
• 전년 누계강우량	• 전일 유입량	• 전일 방류량(조정지)
• 저수량(전년)	• 저수량(예년)	• 댐수위
• 방수로수위	• 강우량_해당시간	• 자체유입
• 총방류량	• 강우량 전일	• 증기압
• 제천시(청풍교)_유량	• 충주시(향산리)_유량	

<표II-3> 통계적 검증 완료된 데이터

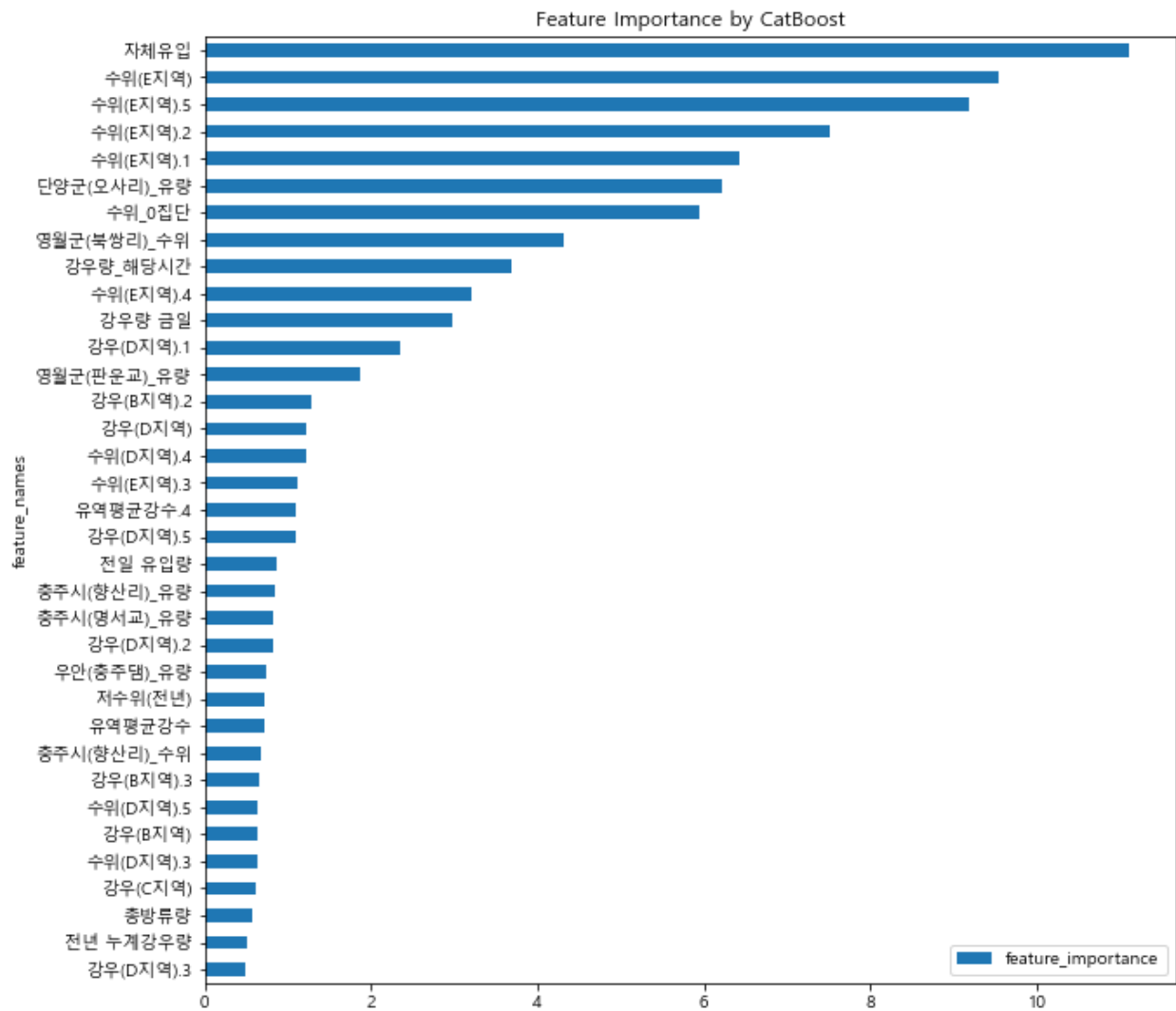
□ 모델적 검증

- 앞서 통계적으로 검증된 데이터를 가지고 실제 모델에서 유입량을 예측하는데 적절한 데이터인지 모델을 통해 검증함.
- 주어진 데이터인 데이터집단 6개, 덴드로그램 알고리즘을 통해 생성된 11개 피처와 통계적검증을 통해 선별된 25개 피처에 대하여 실제 모델에서 유입량을 예측하는데 적절한 데이터인가를 모델을 통해 검증함.
- 사용된 모델 : ExtraTrees , XGB , CatBoost, LGBM 총 4개 모델을 사용함.
  - 예측할 때 불순도를 가장 크게 감소시키는 피처의 중요도를 크게 측정하는 트리계열 모델들을 사용하여 통계적 검증을 거친 피처들이 실제 모델 예측을 진행할 때도 큰 역할을 수행할 수 있는지 확인하기 위해 feature\_importance를 사용하여 검증을 진행함.

모델	ExtraTrees	XGB	CatBoost	LGBM
성능	213.905	245.857	173.496	258.958

<표II-4> 모델적 검증 성능표

○ 가장 성능이 높게 측정된 CatBoost에 대한 feature\_importance를 확인함.



[그림 Ⅲ-5] Feature Importance

- feature importance 를 확인한 결과 수집된 데이터 및 가공된 피처가 유입량 예측시에 상위권에서 영향을 미치고 있음을 알 수 있음.
- 또한, 원본 데이터집단의 데이터들도 상위권에 위치한 것으로 보아 그대로 사용하는 것으로 결정함.

## 5. Modeling

### □ 전처리

#### ○ 피쳐 분포 파악

- 앞서 만들어 놓은 78개의 피쳐들에 대해 KDE plot을 사용하여 분포를 파악함. 파악결과 대다수의 피쳐들이 skewness(왜도)가 높은 것으로 보임.

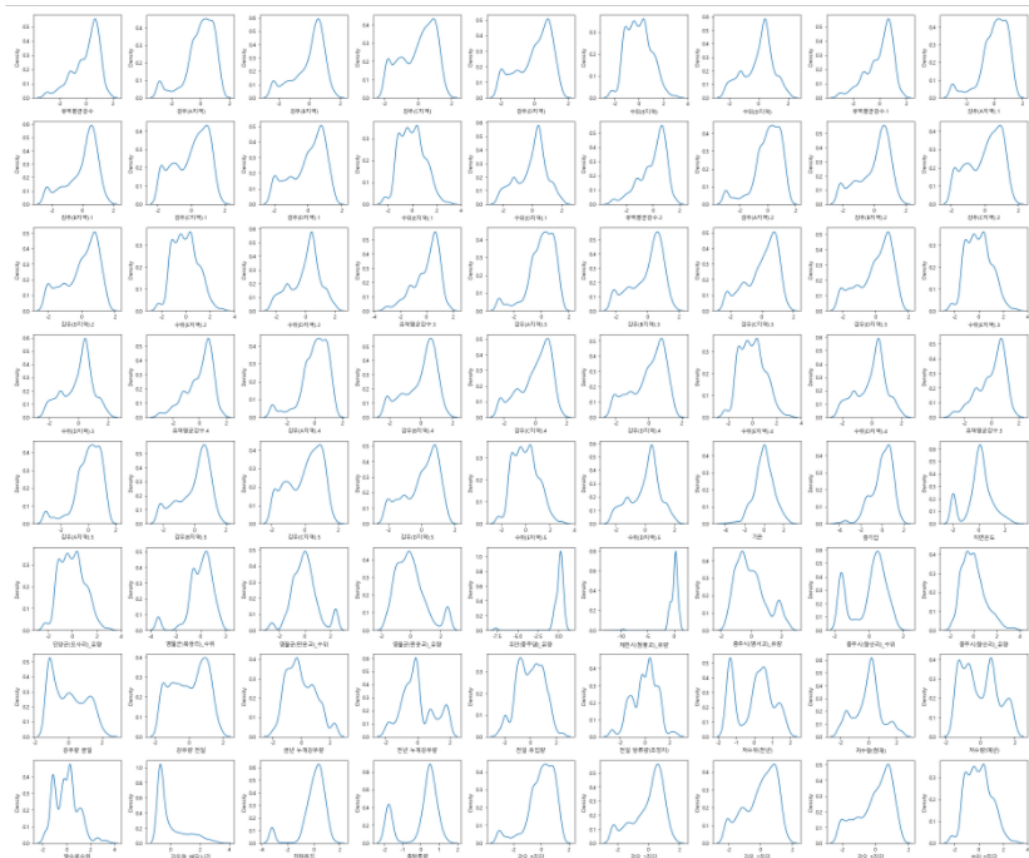


[그림 IV-1] Kde plot으로 확인한 각 피쳐들의 분포

- 데이터의 Skewness가 높다면 모델이 예측을 편향되게 할 것이고, 이 때문에 rmse 값이 더욱 증가할 것이라고 예상함.
- 따라서 **Log transformation**과 **Standard Scaler**를 적용하여 Skewness가 해소되면 모델의 성능이 향상될 것이라 생각함.

## ○ Log Transformation

- Skewness가 1이 넘어가는 피쳐들에 Log transformation을 적용하고 모든 피쳐에 대해 Standard Scaling 을 적용해 Skewness가 잘 해소되었는지 확인함.



[그림 IV-2] log변환 & Standard Scaling 후 분포의 차이

- scaling과 transform의 적용 전후를 ExtraTrees, LGBM, XGB, GBM, CatBoost 4가지의 모델을 활용해 하이퍼파라미터 튜닝 과정을 거치지 않고 성능을 비교함.

RMSE	ExtraTrees	LGBM	XGB	GBM	CatBoost
Scale 전 점수	240.178	348.590	277.771	285.132	204.827
Scale 후 점수	232.663	360.749	278.602	284.561	208.727

<표III-1> scaling여부 비교 성능표

- 검증결과 scaling 전과 scaling 후의 모델 성능이 극명하게 차이가 나지 않음.
- 이는 RMSE값이 평균적으로 낮은 모델들이 **skewness와 scale에 민감하지 않은 Tree Based Model이기 때문**이라 판단함.

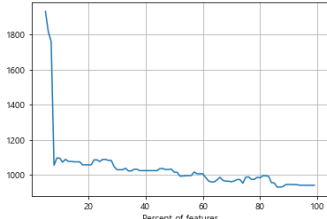
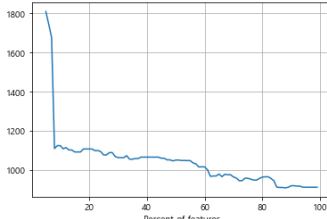
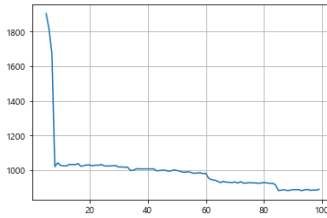
○ 모델링 전 최종 피쳐 선택

- 모델의 전개 과정을 고려했을 때 수식적으로 덜 복잡하고, 원본 데이터를 바로 활용할 수 있는 모델이 홍수 발생 시 유입량을 빠르게 측정하는데 유리할 것이라고 판단함.
- 따라서 **Transformation과 Scaling을 적용하지 않은 피쳐들을 사용하기로 결정함.**

□ Feature Selection

○ 전처리 과정을 거친 피쳐를 전부 다 사용하는 것보다 트리계열 모델을 통한 selection 과정을 거쳐 가장 좋은 성능을 낼 수 있는 피쳐의 개수를 선택하기로 결정함.

○ xgb, lgbm, extra에 대한 모델 selection 과정을 각각 진행하고 성능을 확인함.

Selection 모델명	Selection 시각화
XGB	 <p>(86, 929.989818956301)</p>
LGBM	 <p>(88, 909.2105765427041)</p>
ExtraTrees	 <p>(93, 880.2352814163789)</p>

<표Ⅲ-2> 모델별 Feature Selection 결과

- 가장 많은 피쳐가 선택되면서 모델의 성능이 가장 좋게 측정된 ExtraTrees에서 선택된 93%의 피쳐가 **72개를 가져가기로 결정함.**

## □ Default Model

- 사용된 모델 : **KNeighbor, ExtraTree, GBM, XGB, CatBoost, LGBM 6개 단일모델**을 사용함.
- 선형모델 Ridge, Lasso, Elastic, ARD, BayesianRidge 와 다층퍼셉트론 회귀모델 MLP, 최근접 이웃 모델 KNN, 트리계열 모델 ExtraTree, GBM, XGB, CatBoost, LGBM 중 RMSE가 500 이하로 나오는 위의 6개 모델을 가져가기로 결정함.

KNN	Extra	GBM	XGB	LGBM	CAT
450.350	215.806	263.840	246.192	260.157	173.764

<표Ⅲ-3> Default Model 성능표

## □ Hyperparameter Tuning

- 불필요한 반복 수행 횟수를 줄이면서 정해진 간격 사이에 위치한 값들에 대해 확률적 탐색이 가능하여 시간 대비 효율이 뛰어난 RandomSearchCV를 사용하여 튜닝을 진행함.

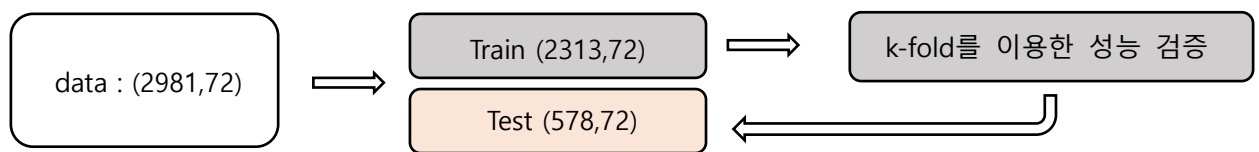
모델명	하이퍼파라미터	튜닝 후 성능
KNN	'n_neighbors' : [3,5,7,9,11], 'weights' : ['uniform','distance']	378.961
ExtraTree	'n_estimators' : [100, 150, 200, 250, 300], 'max_depth' : [10, 12, 15, 17, 20], 'max_features' : [0.8, 0.85, 0.9, 0.95], 'min_samples_split' : [1, 2, 3, 4, 5], 'min_samples_leaf' : [1, 2, 3, 4, 5]	214.514
GBM	'n_estimators' : [100,300,500,1000], 'learning_rate' : [0.01,0.03,0.05,0.1], 'max_depth' : [3,5,6], 'min_samples_leaf' : [3,5,7,9,10], 'min_samples_split' : [2,4,6,8,10], 'subsample' : [0.8,0.9,0.95,1]	187.440
XGB	'n_estimators' : [100,200,300,400,500], 'learning_rate' : [0.01,0.03,0.05,0.1], 'max_depth' : [3,5,6], 'colsample_bytree' : [0.0,0.1,0.3,0.5,0.7,0.9,1], 'min_child_weight' : [1,3,5,6], 'subsample' : [0.8,0.9,0.95,1], 'objective' : ['reg:squarederror']	177.448
LGBM	'n_estimators' : [300,500,700,1000,1100], 'learning_rate' : [0.01,0.03,0.05,0.1], 'max_depth' : [3,5,7,9,10], 'colsample_bytree' : [0.0,0.1,0.3,0.5,0.7,0.9,1], 'subsample' : [0.8,0.9,0.95,1], 'num_leaves' : [30,31,33,35,39,40]	176.825

CAT	'learning_rate': [0.05, 0.1, 0.2, 1, 1.5], 'depth': [3, 5, 7, 9, 10], 'iterations' : [500, 700, 1000, 1200], 'l2_leaf_reg' : [2, 5, 7, 10, 20], 'verbose':[False]	159.890
-----	---	---------

<표Ⅲ-4> RandomSearch Tuning 후 성능표

## □ 과적합 검증

- 튜닝하여 나온 각 모델의 성능이 데이터에 과적합된 것이 아닌지 확인하기 위하여 검증을 실시함.



[그림 IV-3] 과적합 검증 방법

- 전체 데이터를 test 0.2의 비율로 train\_test\_split한 후 k\_fold를 사용하여 학습시킨 train 데이터의 rmse성능과 학습된 모델들로 측정한 test 데이터의 rmse성능을 비교하여 과적합 여부를 판단함.

성능측정	KNN	Extra	GBM	XGB	LGBM	Cat
<b>k-fold</b>	471.629	258.393	217.475	154.802	317.199	148.842
<b>Test data</b>	492.512	240.937	209.904	179.713	304.591	182.722

<표Ⅲ-5> 과적합 검증표

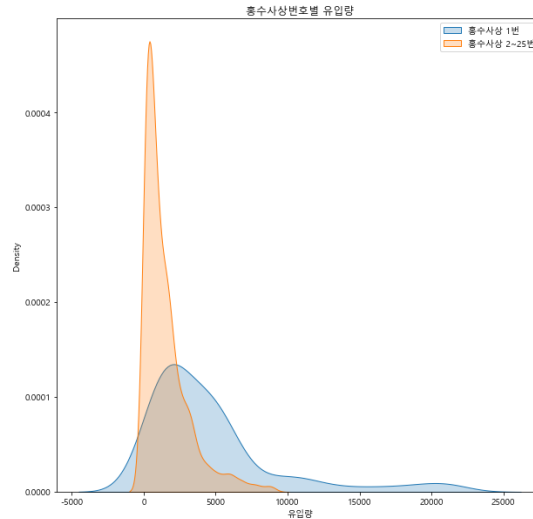
- RMSE 성능이 50 이상 차이가 나 과적합된 양상을 보이는 모델은 없으나 다른 모델들과 큰 성능차이를 보이는 **KNN을 제외하고 학습 및 앙상블을 진행하기로 결정함.**



## □ 학습의 일반화

### ○ 홍수사상번호에 따른 유입량 범위의 불균형

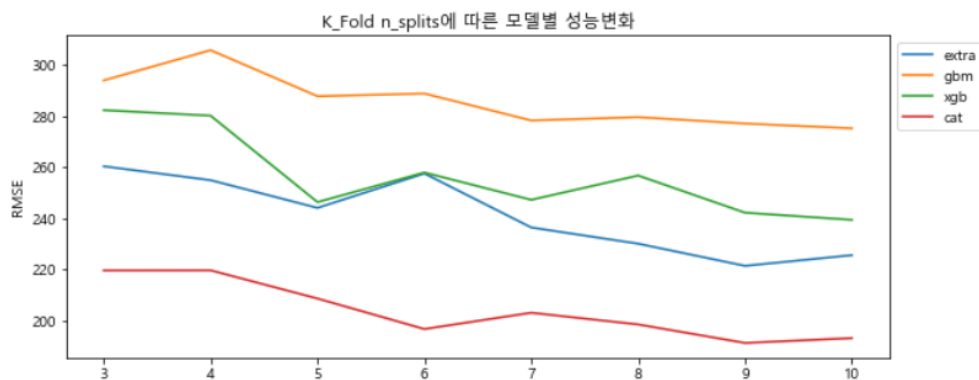
- 총 25개의 홍수사상번호에서 1번 홍수사상의 유입량이 다른 홍수사상번호에 비해 월등히 많음을 알 수 있음.



[그림 IV-4] 홍수사상번호별 유입량 범위

- train\_test\_split을 통해 데이터를 임의로 분할하여 학습할 경우 학습 데이터에 홍수사상번호 1번의 데이터가 고르게 학습되지 않을 수 있으므로 이를 방지하기 위하여 **kfold**를 사용하여 **train전체를 학습할 수 있도록 함**.

### ○ 데이터를 몇 개로 분할할지 결정하기 위하여 n\_split을 3 ~ 10까지 탐색함.



[그림 IV-5] n\_split 별 모델 성능

- n\_split 수가 클수록 RMSE 값이 계속해서 낮아지는 양상을 보이지만, n\_split 이 커질수록 검증되는 데이터의 양이 극명히 적어지게 됨.
- 따라서 대부분의 모델에서 RMSE 값이 크게 낮아지는 양상을 보이고 이후 다시 높아지는 양상을 보이는 **n\_split=5** 으로 결정함.

- 이후 진행되는 모든 학습에서는 **data split 없이 k\_fold를 통해 학습이 진행됨.**

## □ Ensemble

### ○ Averaging Ensemble

- VotingRegressor의 soft voting 을 사용하여 앙상블을 진행함.
- 우리가 최종적으로 사용할 4가지 단일 모델에 대하여 가능한 모든 조합을 voting과 산술 평균을 통해 Averaging 하여 RMSE 값을 측정함.

모델 조합	성능
GBM & XGB	168.880
GBM & LGBM	157.956
GBM & Cat	158.199
XGB & LGBM	153.737
XGB & Cat	151.652
LGBM & Cat	148.526
GBM & XGB & LGBM	153.042
GBM & XGB & Cat	152.930
GBM & LGBM & Cat	146.292
XGB & LGBM & Cat	<b>144.096</b>
GBM & XGB & LGBM & Cat	147.264

<표IV-1> Averaging Ensemble 검증표

### ○ Stacking & Seed Ensemble

- Stacking Transformer를 사용하여 S\_train과 S\_test를 도출하고, S\_train을 5가지 메타모델로 학습시켜 RMSE값을 측정함.

→ 위의 Averaging Ensemble에서 가장 성능이 좋았던 XGB & LGBM & Cat 을 조합하여 학습된 Voting Regressor 모델과 XGB, GBM, LGBM, Cat 의 5가지 모델을 Meta Model로 사용함.

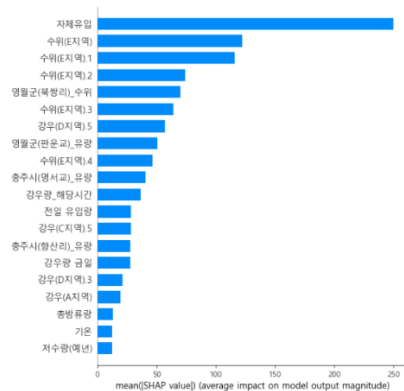
- 각 단일 모델에 대하여 k-fold의 seed와 모델의 seed값을 변경하여 여러 예측 값을 도출한 후 도출된 예측 값을 기하평균을 하여 최종 RMSE 값을 측정함.

- 앞서 진행한 모든 모델링 과정에서 RMSE 값이 가장 낮게 나온 **Averaging Ensemble의 XGB, LGBM, Cat 이 조합된 모델을 사용하여 최종 submission을 도출함.**

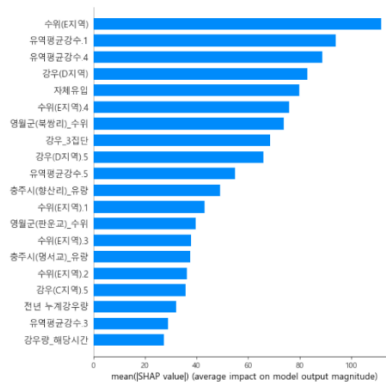
## 6. 분석결과 및 기대효과

○ SHAP 알고리즘을 활용하여 모델이 유입량을 예측할 때 각 데이터가 어떤 영향력 미치고 있는지 분석함.

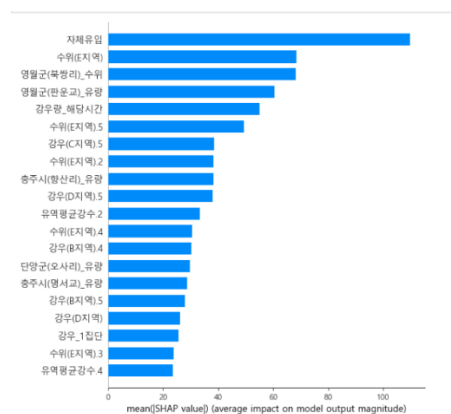
- 모델 종류에 구애 받지 않고 모델이 유입량을 예측하게 된 이유를 이해하기 위해 SHAP 알고리즘을 선택함.
- 실제 submission은 Averaging Ensemble을 통해 학습된 모델의 예측 값이 나왔으므로 **사용된 세 가지 모델 (XGB, LGBM, Cat) 각각의 SHAP value importance를 확인하고 분석함.**



[그림 V -1] XGB SHAP value importance



[그림 V -2] LGBM SHAP value importance



[그림 V -3] Cat SHAP value importance

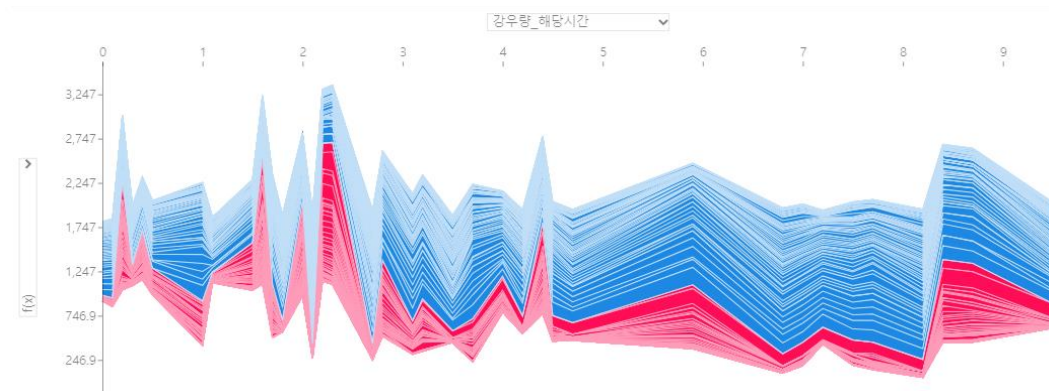
### ❖ SHAP Value 기반 force plot

: 각 피처가 예측값을 예측하는데 양의 영향력을 주었는지 음의 영향력을 주었는지 시각화 된 그래프로 **빨간색은 양의 영향력, 파란색은 음의 영향력**을 나타내고 이때 그래프에서 **x축은 각 피처의 범위** , **y축은 영향력**을 나타냄.

## □ 외부데이터의 영향력

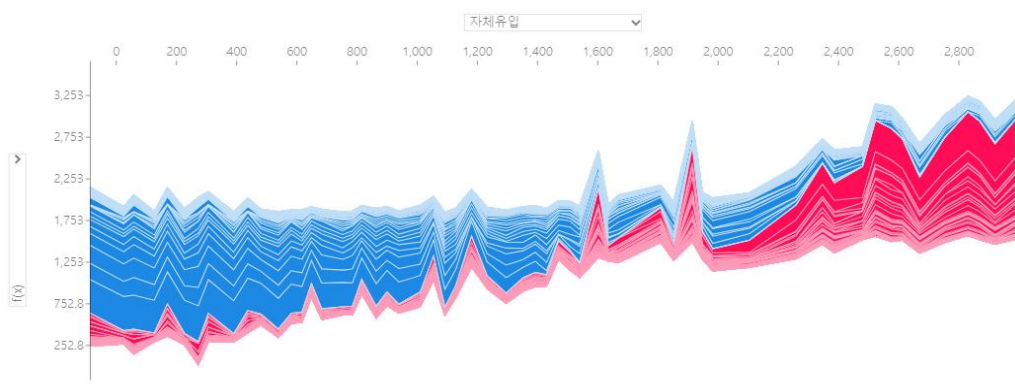
○ 위의 SHAP value importance 그래프를 통해, 수집된 데이터 중에서 날씨와 관련된 **강수량**, **기온** 데이터, 댐과 관련된 **자체유입**, **저수량**, **총방류량** 데이터가 **유입량** 예측에 상위권에서 영향을 미치고 있다는 것을 알 수 있음.

○ 위의 해당 데이터들이 실제로 유입량 예측에 어떤 영향을 미치는지 분석하기 위하여 각각의 데이터에 대하여 force plot를 이용해 시각화를 진행함.



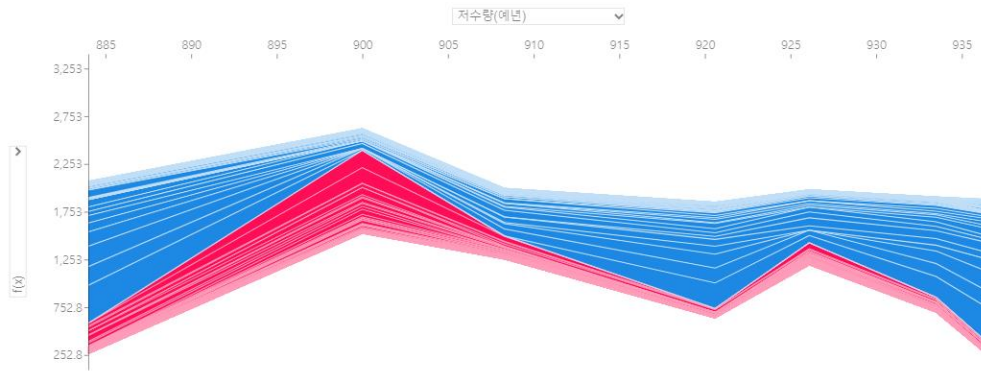
[그림 V -4] 해당시간 강수량 force\_plot

- ➔ 해당시간에 내리는 강우의 양이 적을 때는 대체로 유입량에 양의 영향력을 미치는 비율이 좀더 높고 강우의 양이 늘어날수록 유입량에 음의 영향력의 비율이 높다는 것을 알 수 있음.
- ➔ 이는 실제 우리가 생각하는 인과관계와는 반대되는 양상이지만 강수량이 모델내부에서 위와 같은 영향력을 가지고 있음을 확인함.



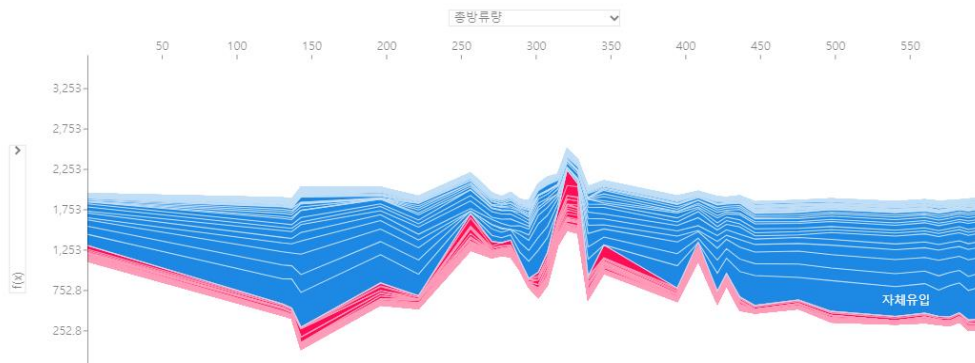
[그림 V -5] 자체유입 force\_plot

- ➔ 댐의 자체 유입량이 늘어날수록 유입량을 예측하는데 양의 영향력을 미치고 있음.



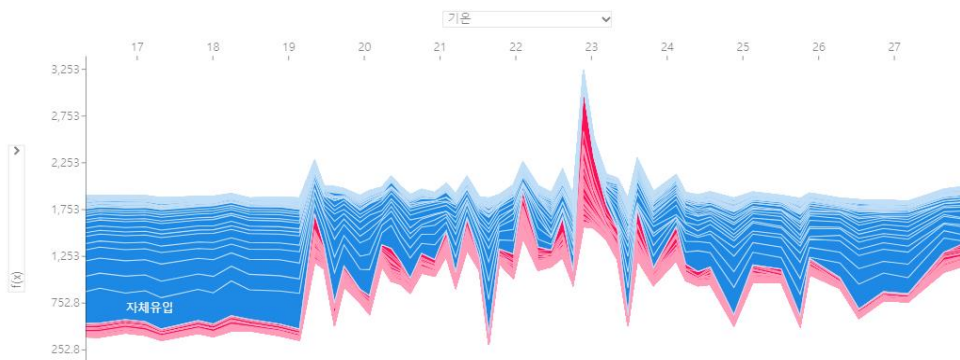
[그림 V -6] 저수량 force\_plot

- ➔ 저수량이 매우 적을 때 유입량에 음의 영향력을 미치고 있음. 이는 실제 저수량이 매우 낮아지는 가뭄사상과 관련이 있을 것으로 판단됨.
- ➔ 반대로 특정 저수량( 약 900  $m^3$ ) 이후로는 유입량에 대한 양의 영향력이 점점 커지는 것을 확인할 수 있음.



[그림 V -7] 총방류량 force\_plot

- ➔ 방류량은 특정 값일 때 (약 320 $m^3$ )를 제외하고 대체로 유입량에 음의 영향력을 미치고 있음.

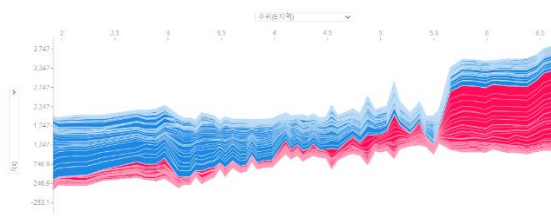


[그림 V -8] 기온 force\_plot

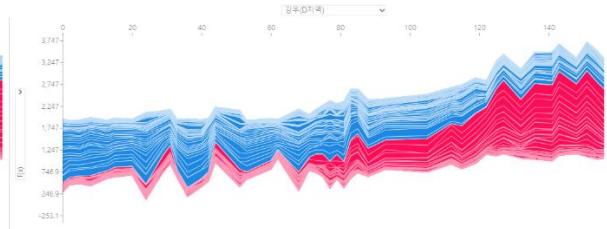
- ➔ 기온은 너무 낮거나 높을 때 유입량에 큰 음의 영향력을 미치고, 약 22 ~ 23도 에서 특히 높은 양의 영향력을 미치고 있음. 이는 특정 날씨일 때의 평균 기온의 영향을 받은 것으로 보여짐.

## □ 우량, 수위 및 유량 데이터

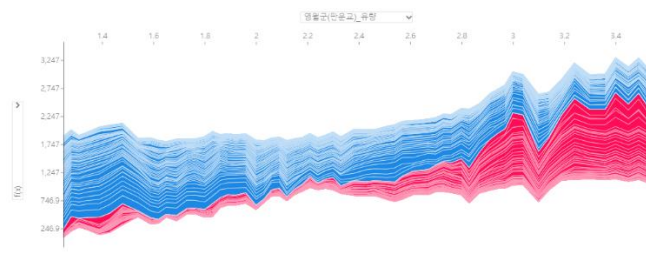
- 수식화 되지 않은 수위, 유량, 우량(강우) 계측 데이터가 상위권에서 함께 유입량 예측에 영향을 주고 있다는 것을 알 수 있음.
- 세 모델의 SHAP value importance에서 가장 상위권에 위치하고 있는 수위, 유량, 우량(강우) 계측 데이터에 대하여 실제 유입량 예측에 어떤 영향을 미치는지 force plot를 이용해 시각화 함.



[그림 V-9] 수위 데이터 force\_plot



[그림 V-10] 강우 데이터 force\_plot



[그림 V-11] 유량 데이터 force\_plot

- 위 세 그래프를 볼 때, 각 계측 데이터들은 양이 많아질수록 유입량 예측에 양의 영향력을 주고 있다는 것을 알 수 있음.
  - 즉 수위, 강우량, 유량이 늘어나게 되면 유입량이 늘어난다는 일반적인 인과관계에 부합하도록 모델에 적용되고 있음.

## □ 결론

- 모델을 통한 유입량 예측에서는 모든 지역에 동일한 계측기가 존재하지 않더라도 각 지역의 서로 다른 계측기를 통해 수집된 데이터를 특별한 가공 없이 함께 사용해 유입량을 예측할 수 있음.
  - 즉, 추가적인 계측기의 구축 없이 기존에 설치되어 있는 계측기만으로도 유입량의 예측이 가능함.
- 자연 현상의 모든 변수를 고려하지 않더라도 계측기를 통해 측정된 데이터와 유입량을 예측할 댐의 현황 및 강우사상과 관련된 데이터를 사용하여 유입량을 예측할 수 있음.
  - 즉, 고려할 수 있는 자연 현상의 변수와 함께 단순한 몇 가지 데이터로 예측이 가능함.

➔ 또한, 해당 모델에서 사용된 데이터는 지형 자료에 구매받지 않으면서 어떠한 댐에서도 수집될 수 있는 데이터로 수집이 어렵지 않고 추가적인 데이터의 가공이 필요하지 않는 장점이 있어 추가적인 기회비용 없이 효율적으로 댐 운영에 기여할 수 있을 것이라 생각됨.

## 7. 참고자료

7-1. 논문) 바람의 효과를 고려한 강우 및 유출 분석

7-2. 논문) 수문기상변화에 따른 유출변화에 관한 연구

7-3. 참고 사이트

- K-water 공공데이터개방포털

[\(<http://opendata.kwater.or.kr/pubdata/dam/exlIncobsrvt.do>\)](http://opendata.kwater.or.kr/pubdata/dam/exlIncobsrvt.do)

- MyWater

[\(\[https://www.water.or.kr/realtime/sub01/sub01/dam/hydr.do?s\\\_mid=1323&seq=1408&p\\\_group\\\_seq=1407&menu\\\_mode=3#this\]\(https://www.water.or.kr/realtime/sub01/sub01/dam/hydr.do?s\_mid=1323&seq=1408&p\_group\_seq=1407&menu\_mode=3#this\)\)](https://www.water.or.kr/realtime/sub01/sub01/dam/hydr.do?s_mid=1323&seq=1408&p_group_seq=1407&menu_mode=3#this)

- 기상청 기상자료개방포털

[\(<https://data.kma.go.kr/data/grnd/selectAsosRltmList.do?pgmNo=36>\)](https://data.kma.go.kr/data/grnd/selectAsosRltmList.do?pgmNo=36)