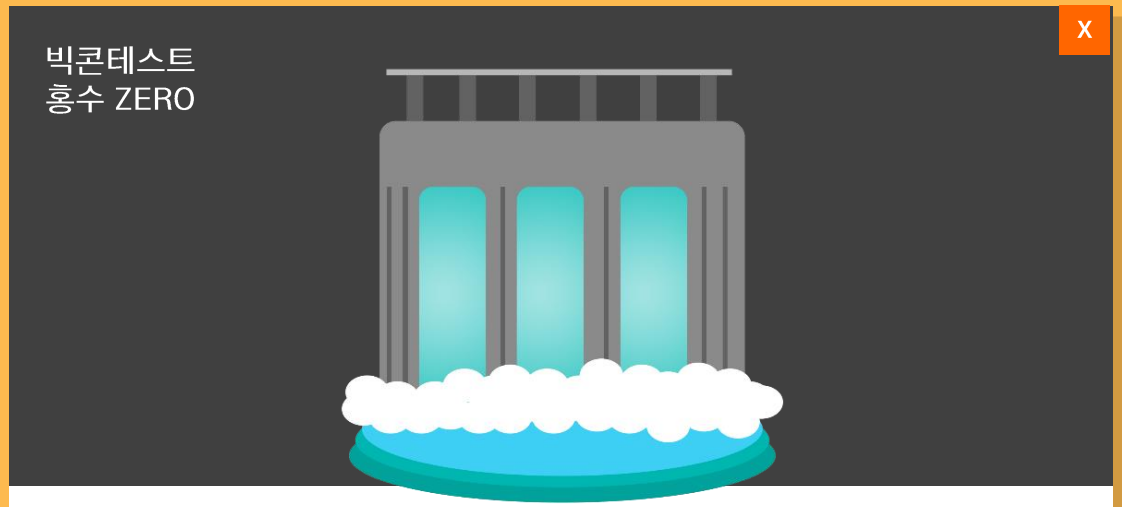


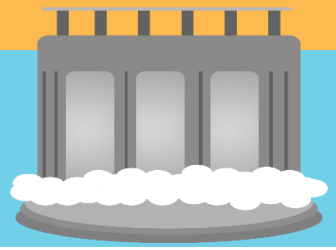
통계적 검증과 앙상블을 활용한 댐 유입량 예측 모형

경통드림팀

장성민 jsm9358@naver.com
한보혜 bohaehan@kookmin.ac.kr
유광열 rhkdduf627@naver.com
윤성식 dbsyoon49@naver.com

Contents

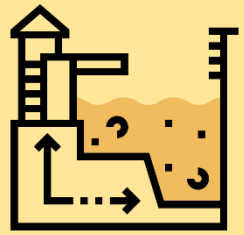
1. 문제 정의



유입량 예측에
자연현상의 모든 변수를
고려할 수 없음

$f(x)$

다양한 변수를 추가하려고
해도 수식화 되지 않은
것들은 추가 불가

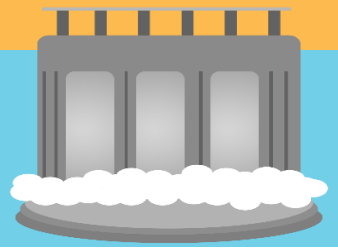


모든 지역에 계측기
구축 부족

솔루션 댐 주변 강우량과 수위 분석 등을 통해 댐의 유입량을 예측해 홍수를 예방할 수 있는 모델 구축

이를 통해 댐 운영을 효율화 해야함

1. 가설 설정

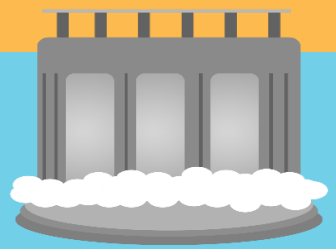


1. 자연 현상의 모든 변수를 고려하지 못하더라도 다른 단순한 데이터를 이용하여 유입량 예측이 가능할 것

2. 수식화 되지 않더라도 데이터 그 자체의 값을 통한 유입량 예측이 가능할 것

3. 모든 지역에 동일한 계측기가 설치 되어 있지 않더라도, 서로 다른 종류의 계측기를 통한 측정값을 이용하여 유입량 예측이 가능할 것

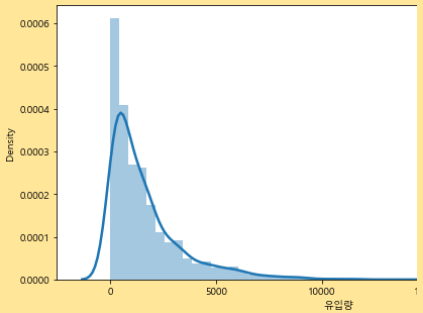
2. 데이터 수집: 지역 특정



지역 특정 이유

- 유입량 예측을 위해 수위와 우량을 제외한 기상적, 지리적, 지역적 특성을 고려하고자 함
- 해당 데이터 수집을 위해 K댐에 대한 지역 특정 필요

지역 특정 고려사항



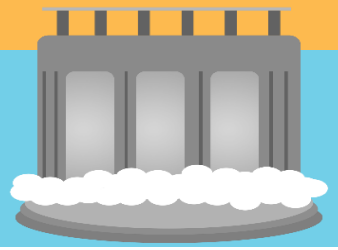
유입량

관측 시기

우량 및 수위관측소

특정 결과: 강원도 및 충청북도 지역을 특정, 해당 지역과 충주댐을 기준으로 데이터 수집 진행

2. 데이터 수집: 관점 별 데이터 수집



기상적 관점

지리적 관점

지역적 관점

- 홍수사상 발생 시 유입량에 가장 큰 영향을 주는 요소는 강우일 것으로 예상
- 논문 등을 통해 특정한 **11**개의 요소에 대해 **2006 ~ 2018**년 까지 데이터 수집
- 강원도와 충청북도에 위치한 기상관측소에서 데이터를 수집하여 평균한 값을 사용

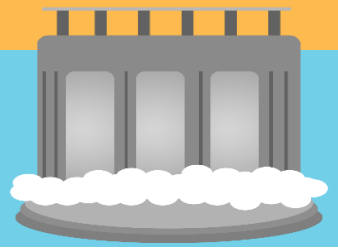
수집한 **11**개 데이터	
기온	전운량
풍속	중하층운량
습도	최저운고
증기압	
이슬점온도	
해면기압	
지면온도	
현지기압	

결측값이 **80%**가 넘는 데이터 제외
(전운량, 중하층운량, 최저운고)

현지기압 데이터 제외

사용할 **7**개 데이터
기온
풍속
습도
증기압
이슬점온도
해면기압
지면온도

2. 데이터 수집: 관점 별 데이터 수집

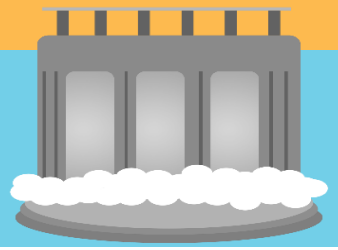


기상적 관점 **지리적 관점** 지역적 관점

- 댐은 강수가 유입되는 수계가 존재하며 각 댐마다 연결되어 있는 수계는 다르므로 특정된 지역의 수계에 대한 데이터가 필요하다고 판단
- 수계를 기준으로 특정된 지역 중 (수계)세 지점에 대한 관련 데이터를 각각 수집

수집한 15개 데이터 종류
강우량
누계 강우량
방류량
저수위
저수량
저수율

2. 데이터 수집: 관점 별 데이터 수집



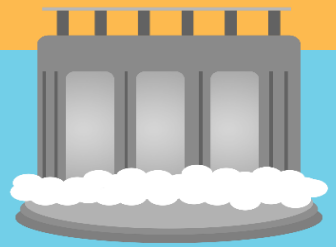
기상적 관점 지리적 관점 **지역적 관점**

- 홍수사상은 댐이 존재하는 해당 지역의 강우량과 댐의 수문현황에도 영향을 받을 것이라 판단해 관련 데이터를 수집
- 기존에 주어진 관측 수치데이터 이외에도 유입량을 관측하는데 도움이 되는 다른 관측소 데이터가 필요하다고 판단

→ 특정된 지역에 존재하는 5개 지역 유량관측소의 유량데이터를 수집

수집한 16개 데이터 종류	
단양_유량, 수위	댐수위
영월_유량, 수위	방수로 수위
우안_유량, 수위	시간당 강수량
제천_유량, 수위	자체유입량
충주_유량, 수위	총 방류량
	수문

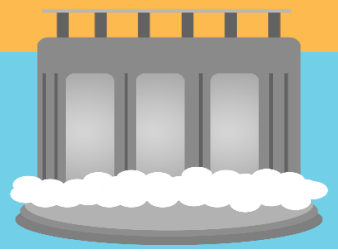
2. 데이터 수집: 외부데이터 목록



기상적 관점(7)	지리적 관점(15)		지역적 관점(16)	
기온	강우량 금일	저수위(현재)	단양(오사리)_유량	충주(향산리)_유량
풍속	강우량 전일	저수위(전년)	영월(북쌍리)_유량	충주(향산리)_수위
습도	금년 누계강우량	저수위(예년)	영월(북쌍리)_수위	댐수위
증기압	전년 누계강우량	저수량(현재)	영월(판운교)_수위	방수로 수위
이슬점 온도	예년 누계강우량	저수량(전년)	영월(판운교)_유량	시간당 강수량
해면기압	전일 유입량	저수량(예년)	우안(충주댐)_유량	자체유입량
지면온도	전일 방류량(본댐)	현재저수율	제천(청풍교)_유량	총 방류량
	전일 방류량(조정지)		제천(청풍교)_수위	수문

총 **38**개의 외부데이터 **Column**을 사용

3. Feature Making

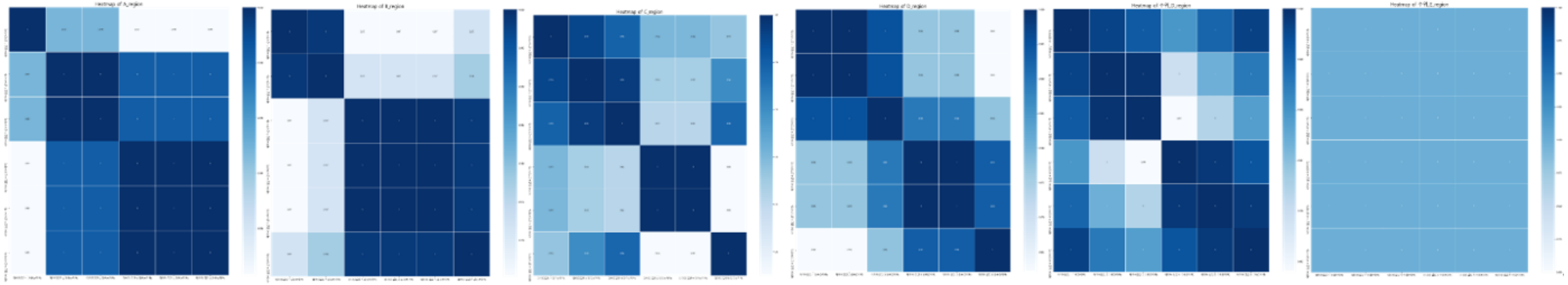


- 덴드로그램을 사용한 **Feature Making**

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

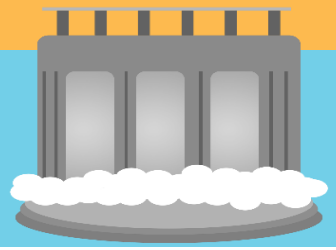
- 같은 지역에서 다른 거리, 시간에서 측정된 6개 관측치는 서로 상관성이 매우 높음

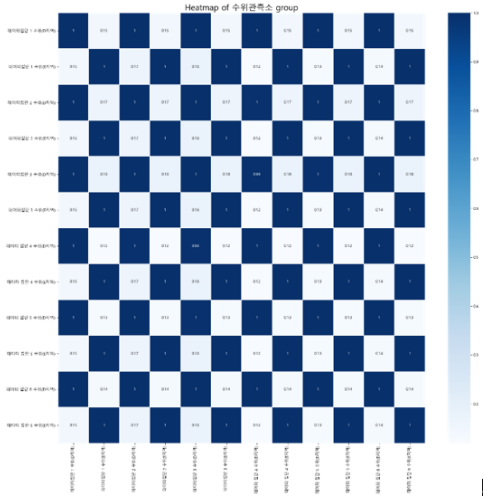
<지역 별 상관관계>



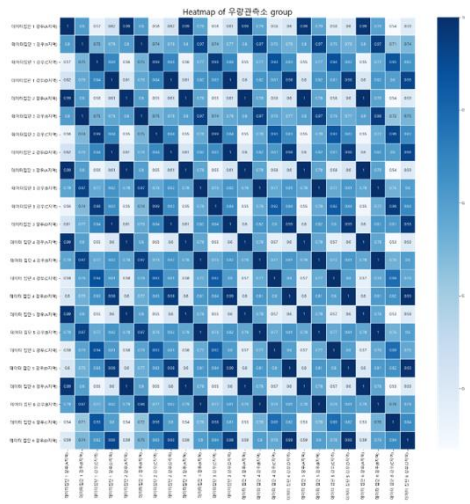
[그림 Ⅱ-1] A 지역 상관관계 [그림 Ⅱ-2] B 지역 상관관계 [그림 Ⅱ-3] C 지역 상관관계 [그림 Ⅱ-4] D 지역 상관관계 [그림 Ⅱ-5] 수위D 지역 상관관계 [그림 Ⅱ-6] 수위E 지역 상관관계

3. Feature Making





<수위 관측소 상관관계>

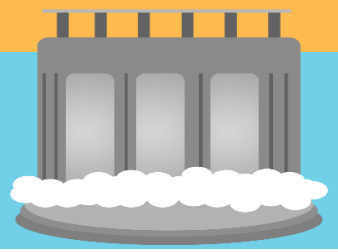


<우량 관측소 상관관계>

24개의 우량관측소와 12개의 수위관측소 관측치들 간의 상관성에서 지역간 관측치의 상관성은 유의미한 차이를 보임

이를 통해 같은 지역에서 측정된 관측치를 평균하여 하나의 피처로 사용하는 것이 용이한지 확인하기 위해 같은 관측소 간의 Hierarchical(계층적) clustering 진행

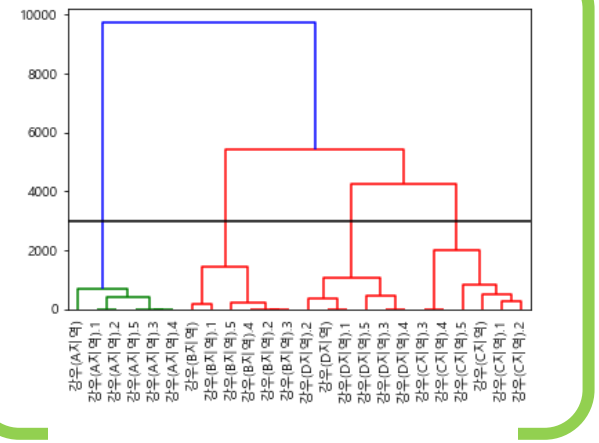
3. Feature Making



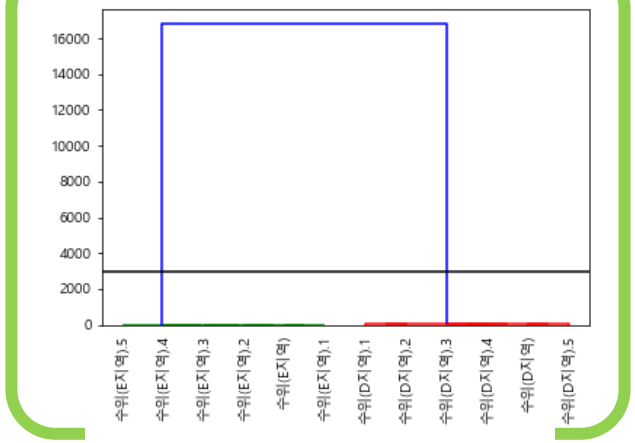
- **Ward Method**

  우량관측소 24개, 수위관측소 12개, 유역평균강수 6개에 대한 각각의 clustering을 진행하며, 해당 지역이 가지는 정보 손실을 최소화 하기 위해 각 관측치 사이의 유사성 계산 방법으로 **ward's method**를 선택하여 진행

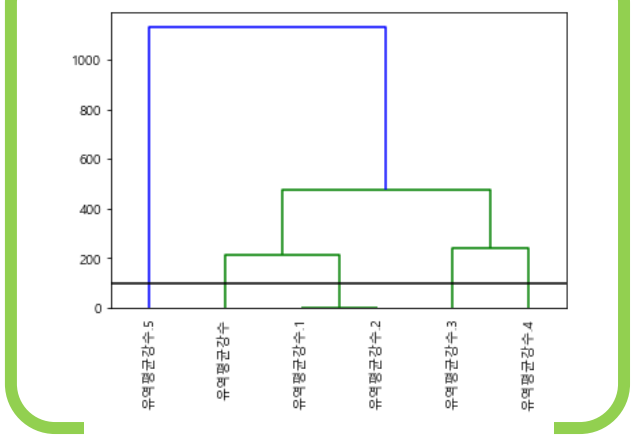
[24개 우량관측소]



[12개 수위관측소]

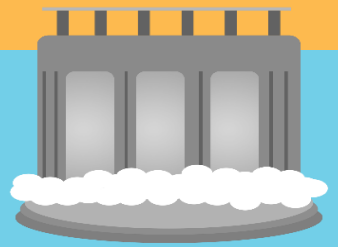


[6개 유역평균강수]



Clustering을 통해 데이터집단에서 같은 관측소와 지역 간의 상관성을 통해 하나의 집단으로 묶어 **총 11개의 피처를 생성**

4. 데이터 검증: 검증 필요성



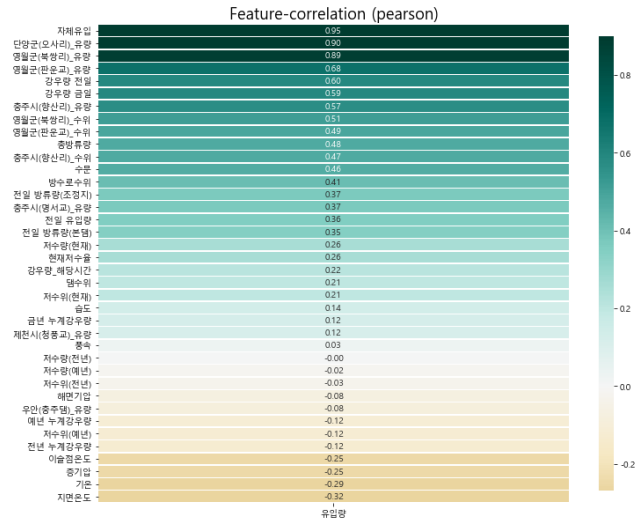
수집된 데이터 및 생성한 피처는 여러가지 분석을 통하여 특정된 지역 및 댐을 기준으로 수집되었음

지역이 얼마나 잘 특정되었는가에 따라 예측 오차의 범위가 매우 크게 달라질 것으로 예상됨

→ 수집된 데이터(38개)와 생성한 피처(덴드로그램) 가 주어진 홍수사상 26번을 예측하기에 적절한 데이터인지에 대한 철저한 검증이 필요하다고 판단하여 **통계적, 모델적 검증을 진행**

4. 데이터 검증: 통계적 검증

- 독립변수와 종속변수 사이의 관계 검증



<선형상관관계 분석>

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

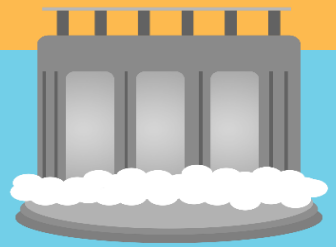
<비선형상관관계 포함 분석>

피어슨 상관계수와 **Mutual Information Score**를 사용해 독립변수와 종속변수 간의 상관관계를 측정한 결과 피어슨 상관계수가 **0.2** 보다 작고, **MI score**가 **0.8**를 넘지 못하는 피처 **3**개를 제거함

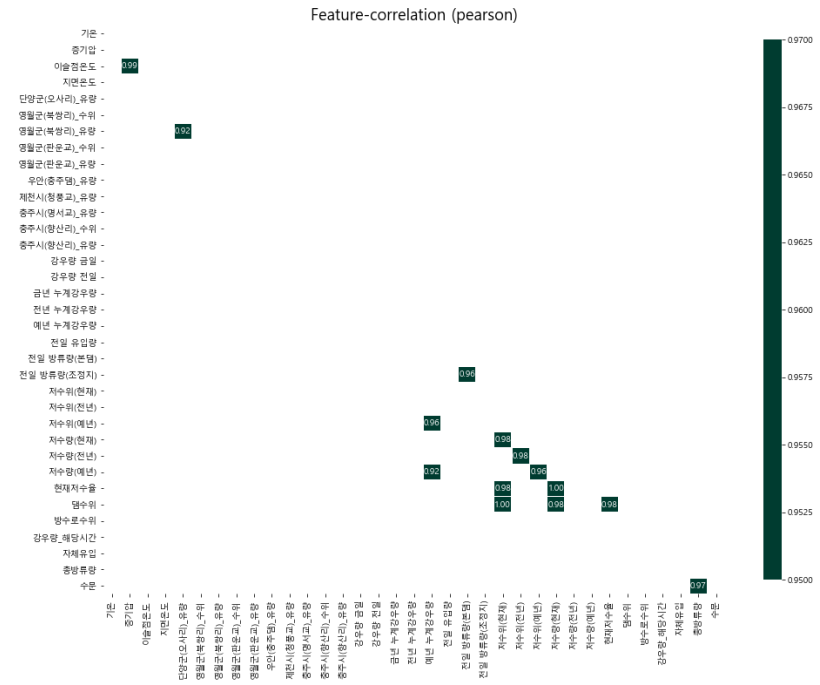
→ **[**‘해면기압’, ‘풍속’, ‘습도’**]**

4. 데이터 검증: 통계적 검증

2021 BIG CONTEST – 홍수ZERO



- 독립변수 간의 상관관계 검증

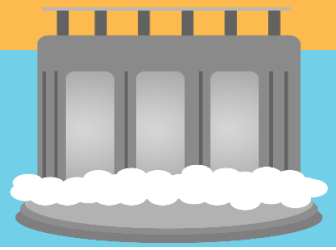


- 독립변수 간의 상관관계가 너무 높을 경우 발생하는 다중공선성 문제를 해결하기 위해 피어슨 상관계수를 이용해 상관성을 확인하고 해결하고자 함
- 독립변수 간 상관계수가 **0.9** 이상인 것들은 둘 중 하나를 제거

→ **['이슬점온도', '영월군(북쌍리)_유량', '전일 방류량(본댐)', '예년 누계강우량', '수문', '댐수위', '저수위(현재)', '저수량(전년)', '저수위(예년)', '현재저수율']**

10개의 피처 제거

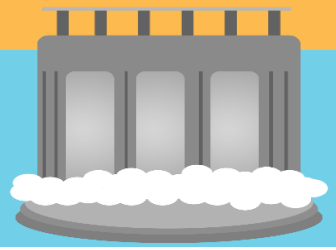
4. 데이터 검증: 통계적 검증



- 통계적 검증 후 남은 외부데이터 목록

검증 후 남은 25개 외부데이터 (피처)			
기온	우안(충주댐)_유량	금년 누계강우량	방수로수위
증기압	제천시(청풍교)_유량	전년 누계강우량	강우량_해당시간
지면온도	충주시(명서교)_유량	전일유입량	자체유입
단양군(오사리)_유량	충주시(향산리)_수위	전일방류량(조정지)	총 방류량
영월군(북쌍리)_수위	충주시(향산리)_유량	저수위(전년)	
영월군(판운교)_수위	강우량 금일	저수량(현재)	
영월군(판운교)_유량	강우량 전일	저수량(예년)	

4. 데이터 검증: 모델적 검증

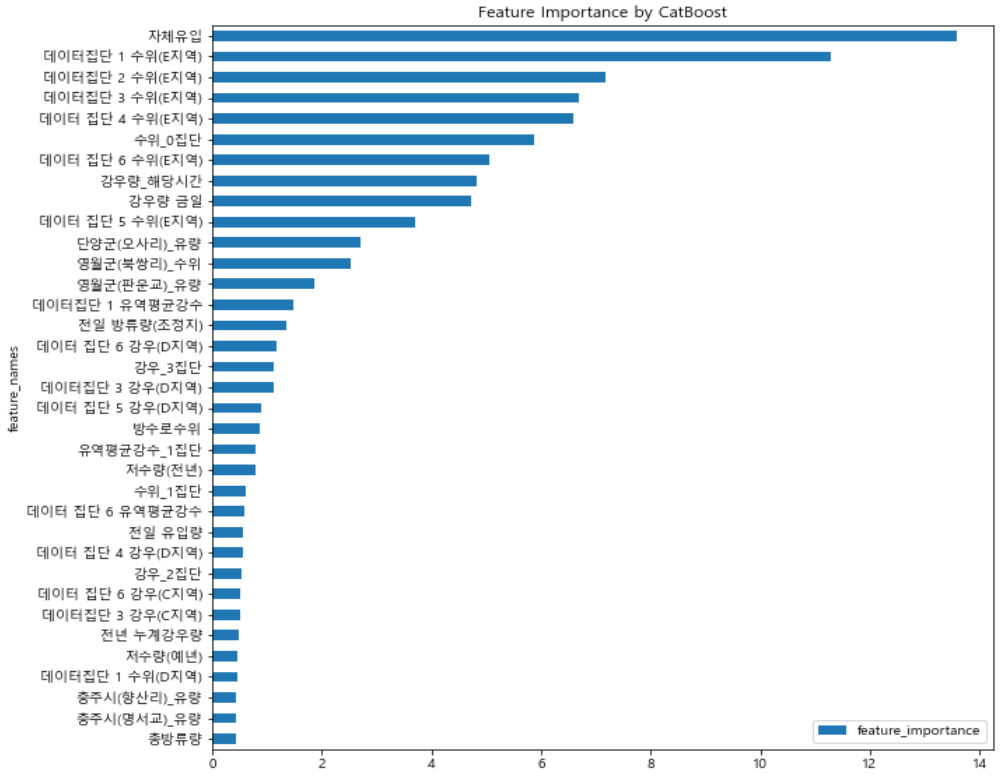


- 통계적으로 검증된 데이터가 실제 모델에서 유입량을 예측하는데 적절한 데이터인지 모델을 통해 검증

  - 트리계열 모델들 중 CatBoost의 Feature_Importance를 사용하여 검증을 진행
  - 수집 및 가공된 피처가 유입량 예측 시 많은 영향을 미침
  - 원본 데이터집단의 데이터들도 상위권에 위치함

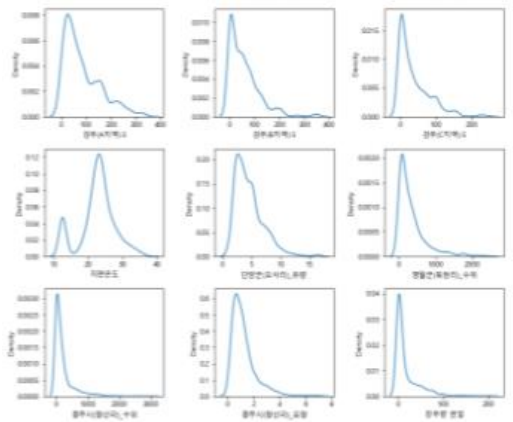
원본데이터 및 검증된 데이터를 사용하는 것으로 결정

모델	Extra Trees	XGB	CatBoost	LGBM
성능	213.905	245.857	173.496	258.958



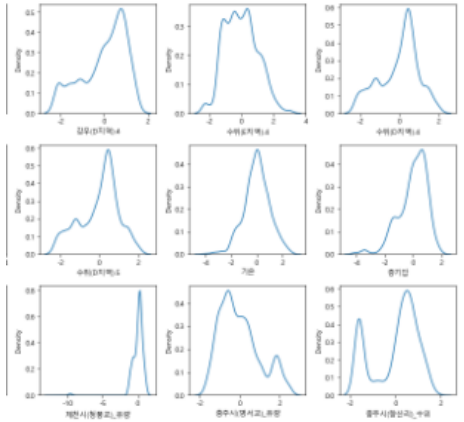
5. Modeling: 전처리

- 데이터 분포파악 및 **Transformer/Scaler** 적용



<원본 분포파악 예시>

log + standard



<변환 후 분포파악 예시>

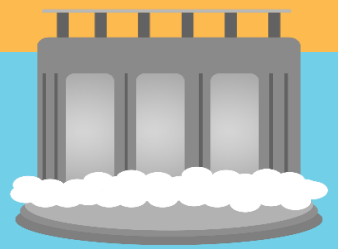
모델	Extra Trees	XGB	CatBoost	LGBM
Scale 전	240.178	277.771	204.827	348.590
Scale 후	232.663	278.602	208.727	360.749

<scaling 전, 후 성능비교>

- 검증결과 **scaling** 전, 후 성능이 극명히 차이 나지는 않음 ← **Tree Based Model**의 특징
- 모델의 전개과정을 고려했을 때 원본데이터를 바로 활용할 수 있는 모델이 유입량을 빠르게 예측하는데 유리할 것이라 판단

Transformation과 **Scaling**을 적용하지 않은 피처를 사용하기로 결정

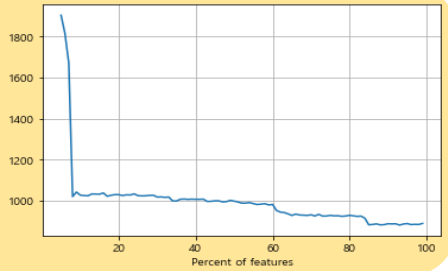
5. Modeling: Feature Selection



트리계열 모델을 통해 성능을 내는 모델 중에서 최적의 피처 개수를 찾기 위해 **Feature Selection** 과정을 진행

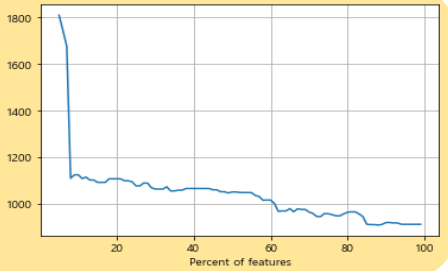
Extra Trees

score: 880.23



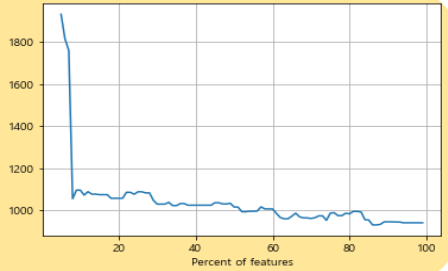
LGBM

score: 909.21



XGB

score:929.98



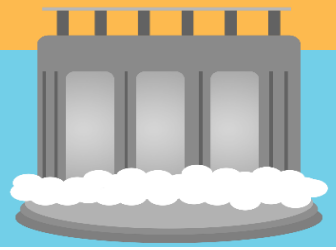
가장 많은 피처가 **Select** 되면서,

모델의 성능이 가장 우수한 **Model: Extra Trees**

Extra Trees Model에서 선택된

93%의 피처 **72**개를 모델링에 사용하기로 결정

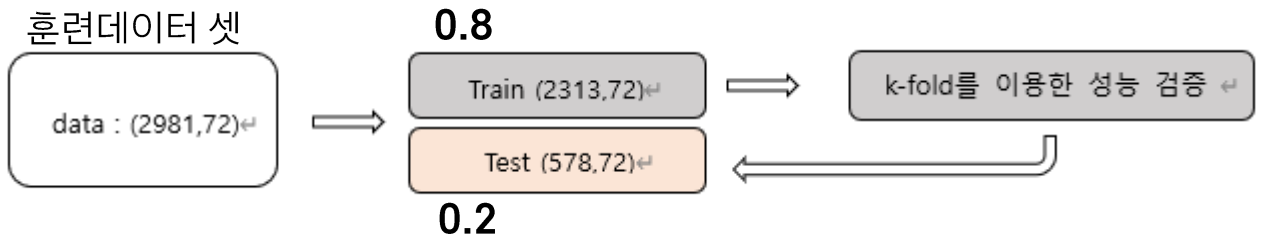
5. Modeling: Hyperparameter Tuning



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

모델명	KNN	Extra Trees	GBM	XGB	LGBM	CatBoost
하이퍼 파라미터	'n_neighbors' : [3,5,7,9,11], 'weights' : ['uniform','distance']	'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],	'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],	'n_estimators' :[100,200,300 ,400,500], 'learning_rate' : [0.01,0.03, 0.05,0.1], 'max_depth': [3,5,6], 'colsample_bytree' :[0,0.1,0.3, 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']	'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' : [0,0.1,0.3,0.5, 0.7,0.9,1], 'subsample' :[0.8,0.9,0.95,1],	'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]
튜닝 후 성능	378.961	214.514	187.514	177.448	176.825	159.890

5. Modeling: 과적합 검증



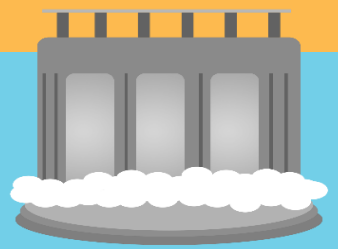
전체 데이터를 test 0.2의 비율로 train_test_split한 후 k_fold를 사용해 학습시킨 train 데이터의 rmse성능과 학습된 모델들로 측정한 test 데이터의 rmse성능을 비교 후 과적합 여부 판단

<과적합 검증표>

성능측정	KNN	Extra	GBM	XGB	LGBM	CatBoost
K-fold	471.629	258.393	217.475	154.802	317.199	148.842
Test data	492.512	240.937	209.904	179.713	304.591	182.722

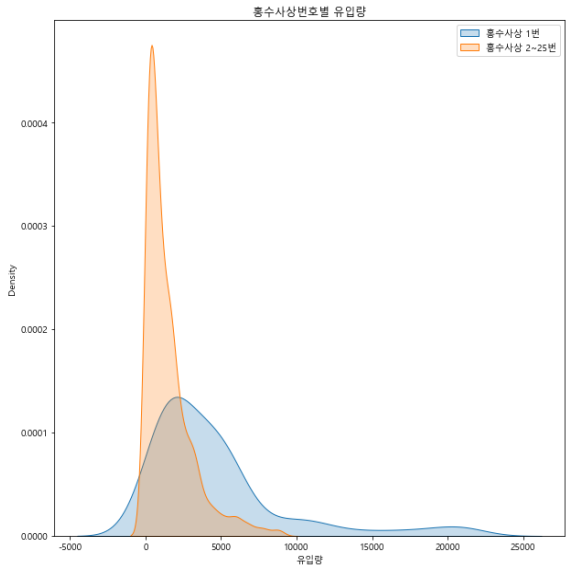
RMSE 성능이 500이상 차이가 나 과적합된 양상을 보이는 모델은 없으나 다른 모델들과 큰 성능 차이를 보이는 KNN을 제외하고 학습 및 앙상블을 진행

5. Modeling: 학습의 일반화



- 학습의 일반화

홍수사상번호에 따라 유입량이 불균형 함을 파악

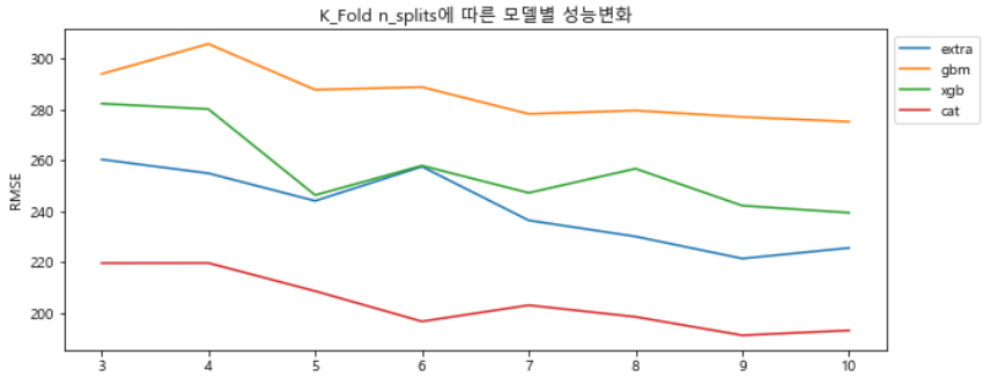


K-fold 를 사용하여 **train set** 전체를

고르게 학습할 수 있도록 함

- n-split 결정

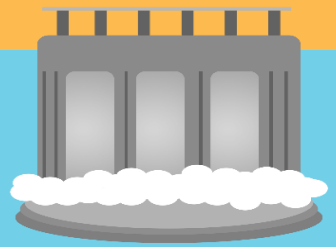
K-fold를 사용하기 위해 데이터를 몇 개**(n)**로 분할할 것인지 탐색함



n_split 수가 클수록 RMSE 값이 계속해서 낮아지는 양상,
하지만 검증에 사용되는 데이터의 양은 극명히 줄어듦

→ 대부분의 모델에서 RMSE 값이 크게 낮아지는 양상을 띄고,
이후 다시 높아지는 양상을 보이는 n_split=5 으로 결정

5. Modeling: 앙상블



- Averaging Ensemble

모델 조합	성능(RMSE)
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	144.096
GBM & XGB & LGBM & Cat	147.264

❖ Stacking & Seed Ensemble

1. Stacking Transformer를 사용해 5가지 메타 모델로 학습
2. 각 단일 모델에 대해 k-fold의 seed와 모델의 seed값을 변경해 여러 예측값을 도출

- 최종적으로 사용할 4가지 단일 모델에 대해 모든 조합을 voting과 산술평균을 통해 Averaging 하여 RMSE 값을 측정
- 진행한 모든 모델링 과정에서 RMSE 값이 가장 낮게 나온 Averaging Ensemble 방법을 사용

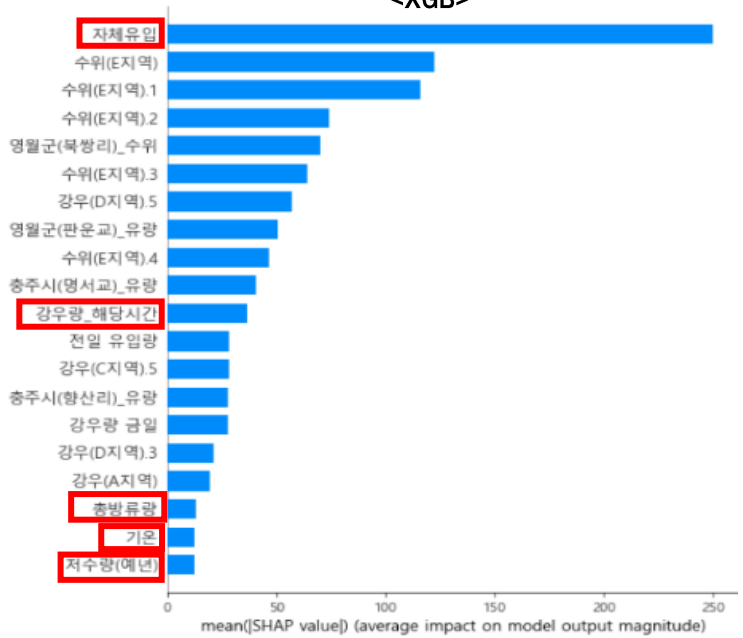
최종선택 **Model**

XGB & LGBM & Cat Average Ensemble

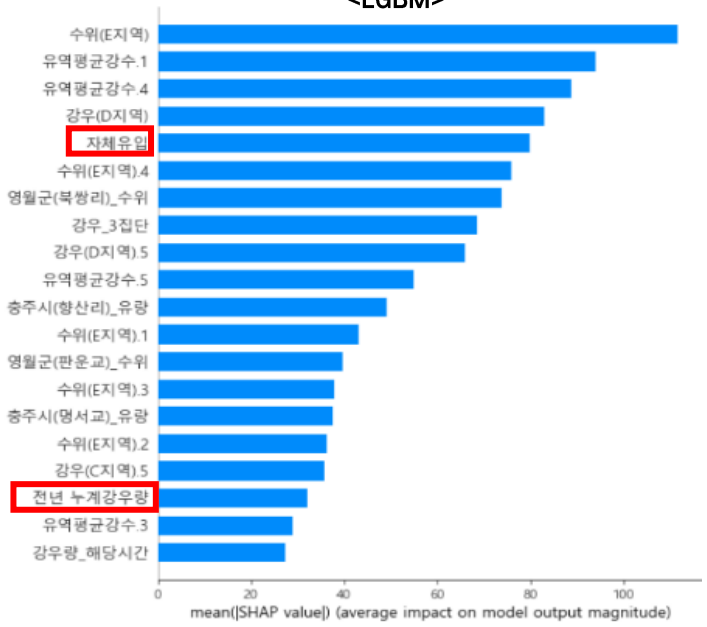
6. 분석결과 및 기대효과

- **SHAP** 알고리즘을 이용한 피처 영향력 분석 - **Shape value** 값을 이용한 각 모델의 **feature importance**

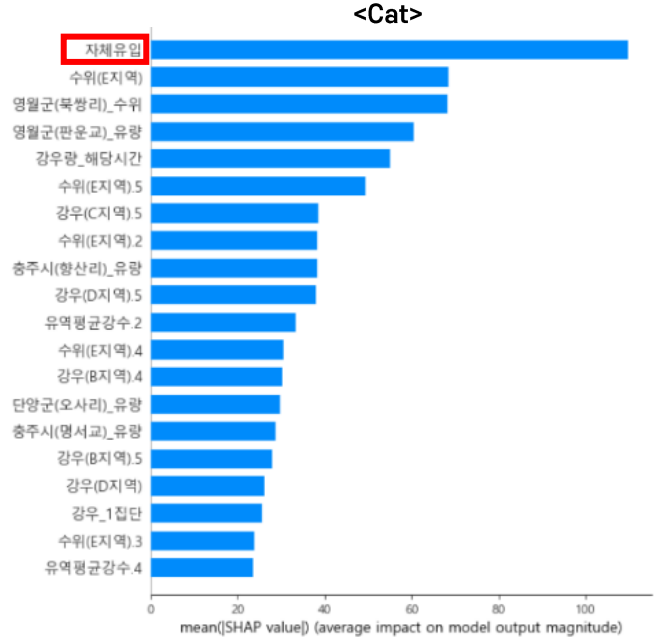
<XGB>



<LGBM>



<Cat>



→ 수집된 데이터 중에서는 기상적 관점의 강우량과 기온, 지리적 관점의 저수량, 지역적 관점의 방류량 데이터가 유입량 예측의 상위권에서 영향을 미치고 있음

→계측기를 통해 측정된 수위, 우량, 유량 데이터가 함께 상위권에서 유입량 예측에 고르게 영향을 미치고 있음을 알 수 있음

6. 분석결과 및 기대효과

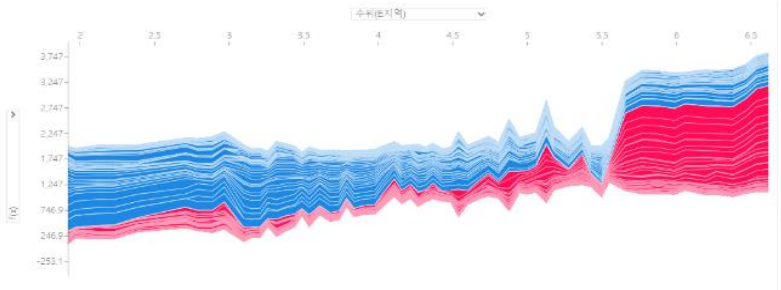
- SHAP value 를 이용한 force plot 분석

❖ SHAP Value 기반 force plot

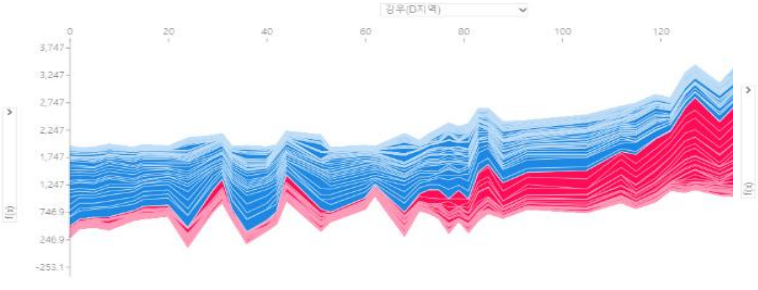
: 예측 값에 대한 피처의 영향력을 보여주는 그래프

빨간색은 양의 영향력, 파란색은 음의 영향력, x축은 각 피처의 범위, y축은 영향력을 나타냄

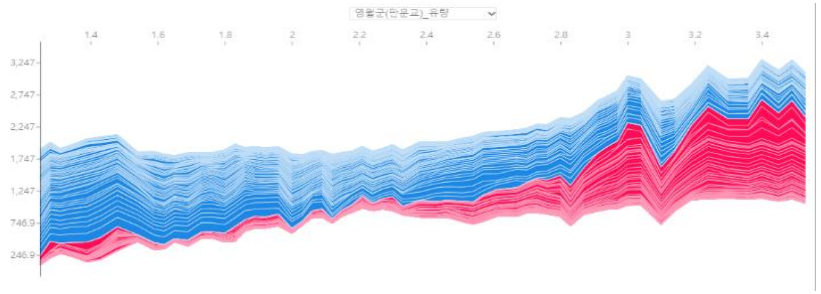
<수위 데이터>



<강우 데이터>



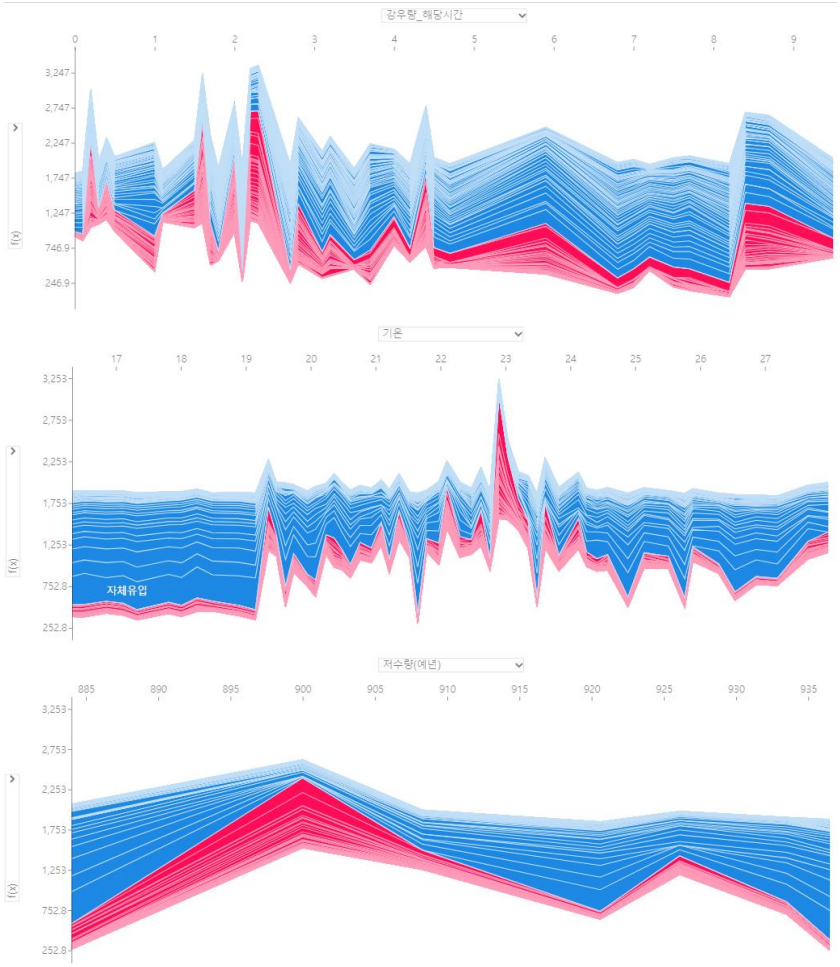
<유량 데이터>



→ 각 데이터들은 양이 많아질수록 모델의 유입량 예측에 양의 영향력을 주고 있다는 것을 확인할 수 있음

→ 즉 수위, 강우량, 유량이 늘어나게 되면 유입량이 늘어난다는 일반적인 인과관계에 부합하도록 모델에 적용되고 있음

6. 분석결과 및 기대효과



[강우량]

→ 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

→ 실제 **우리가 생각하는 인과관계와는 반대되는 양상**이지만 강우량이 모델내부에서 위와 같은 영향력을 가지고 있음을 확인

[기온]

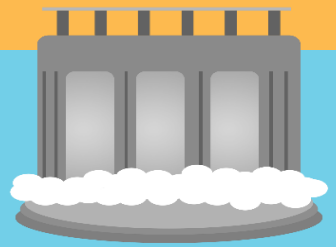
→ 기온은 너무 낮거나 높을 때 유입량에 큰 음의 영향력을 미치고, **약 22 ~ 23도 에서 특히 높은 양의 영향력**을 미치고 있음. 이는 특정 날씨일 때의 평균 기온의 영향을 받은 것으로 보임

[저수량]

→ 저수량이 매우 적을 때 유입량에 음의 영향력을 미치고 있음. 이는 실제 저수량이 매우 낮아지는 가뭄사상과 관련이 있을 것으로 판단

→ 반대로 특정 저수량(약 900 m^3) 이후로는 유입량에 대한 양의 영향력이 점점 커지는 것을 확인

6. 분석결과 및 기대효과



가설1) 자연 현상의 모든 변수를 고려하지 못하더라도 다른 단순한 데이터를 이용하여 유입량 예측이 가능할 것

모든 변수를 고려하지 않더라도 고려할 수 있는 자연 현상의 변수와 함께 단순한 몇 가지 데이터로 예측이 가능함

가설2) 수식화 되지 않더라도 데이터 그 자체의 값을 통한 유입량 예측이 가능할 것

해당 모델에서 사용된 데이터는 지형 자료에 구애 받지 않으면서 수집이 어렵지 않고 추가적인 데이터의 가공이 필요하지 않아 추가적인 기회비용이 필요하지 않음

가설3) 모든 지역에 동일한 계측기가 설치 되어 있지 않더라도, 서로 다른 종류의 계측기를 통한 측정값을 이용하여 유입량 예측이 가능할 것

추가적인 계측기의 구축 없이 기존에 설치되어 있는 계측기만으로도 유입량의 예측이 가능함

분석 초기에 설정한 3가지 가설이 성립하므로 구축된 모델이 댐 운영에 효과적으로 사용될 수 있을 것으로 기대됨.


X
QnA