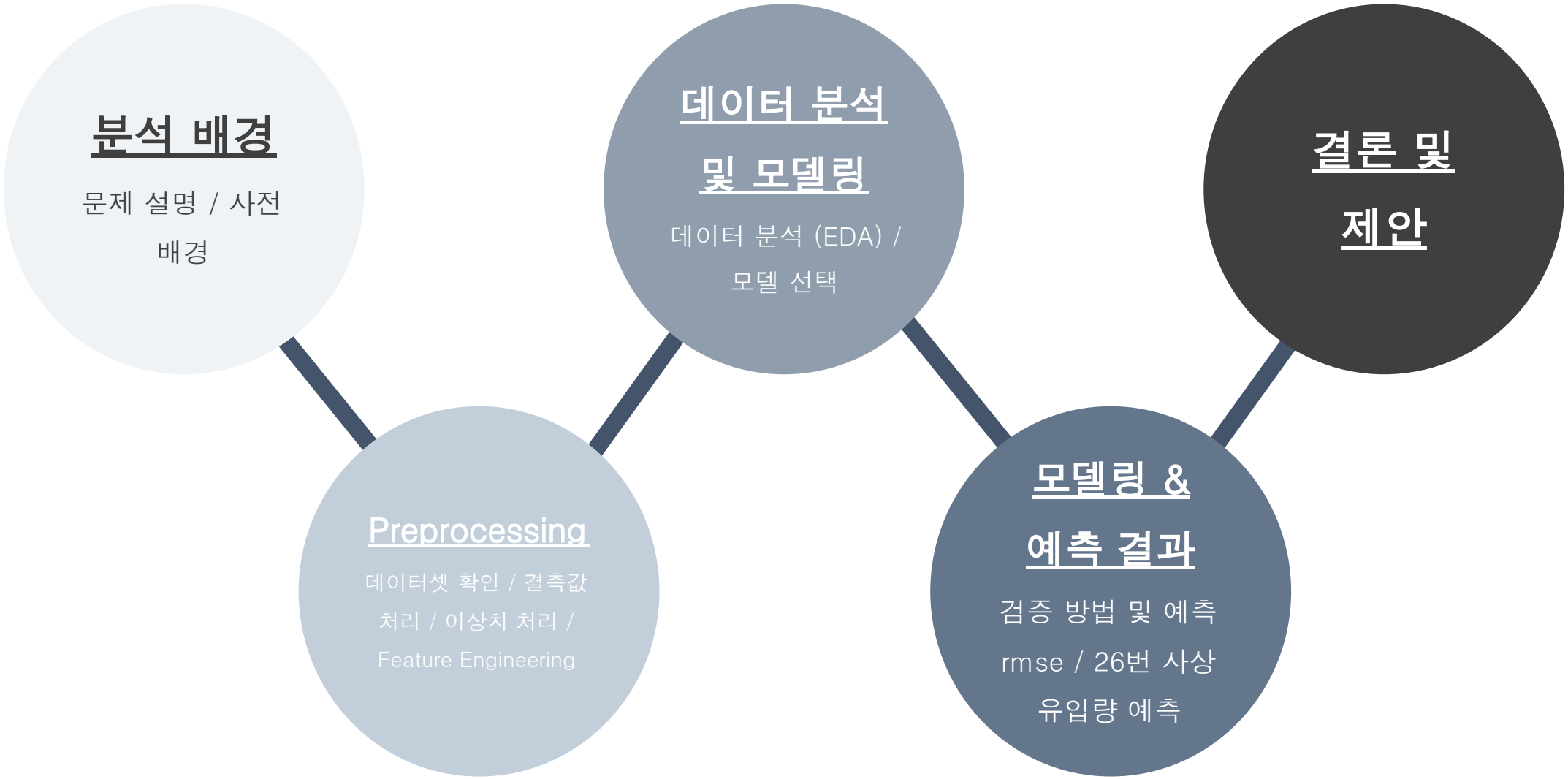
Insightwo : 강동인, 엄다린 박건 박정균

Kalman-Filter & CatBoost를 활용한 댐 유입량 예측

팀명 : insightwo
강동인 : **rkd2016@gmail.com**
엄다린 : edr2@naver.com
박건 : **kun98615@naver.com**
박정균 : park32323@gmail.com

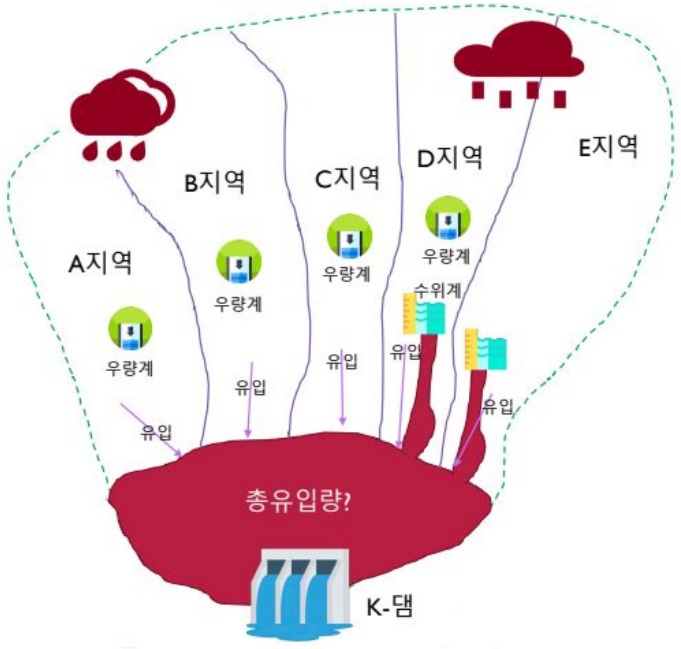
목차


분석 배경
문제 설명 / 사전 배경
Preprocessing
데이터셋 확인 / 결측값 처리 / 이상치 처리 / Feature Engineering
데이터 분석 및 모델링
데이터 분석 (EDA) / 모델 선택
모델링 & 예측 결과
검증 방법 및 예측 rmse / 26번 사상 유입량 예측
결론 및 제안

분석배경

- 문제 설명

최근 10년간 발생했던 홍수사상(총 25개)을 대상으로
유입량에 영향을 미치는 관측소의
강우량 및 수위 데이터를 학습하여
홍수사상 26의 댐 유입수량을 예측하시오.



	시간	유입량	데이터집단 1							데이터집단 2							데이터집단 3							데이터			
			역평균강우량	강우(A지역)	강우(B지역)	강우(C지역)	강우(D지역)	수위(E지역)	수위(D지역)	역평균강우량	강우(A지역)	강우(B지역)	강우(C지역)	강우(D지역)	수위(E지역)	수위(D지역)	역평균강우량	강우(A지역)	강우(B지역)	강우(C지역)	강우(D지역)	수위(E지역)	수위(D지역)	역평균강우량	강우(A지역)	강우(B지역)	강우(C
3	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	6.3	7.0	7.0	7.0	8.0	2.5	122.6	6.4	7.0	7.0	8.0
4	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	6.4	7.0	8.0	7.0	8.0	2.5	122.6	7.3	7.0	8.0	10.
5	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	7.3	7.0	9.0	8.0	8.0	2.5	122.6	8.2	7.0	9.0	10.
6	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	8.2	7.0	10.0	10.0	10.0	2.5	122.6	11.3	9.0	10.0	15.
7	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	11.3	9.0	12.0	10.0	11.0	2.5	122.6	14.4	12.0	12.0	18.
8	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	14.4	12.0	14.0	15.0	14.0	2.5	122.6	16.9	14.0	14.0	20.
9	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	16.9	14.0	17.0	18.0	16.0	2.5	122.6	20.5	15.0	17.0	27.
10	15	952.5	16.9	12.0	24.0	15.0	16.0	2.5	122.6	20.5	15.0	24.0	18.0	16.0	2.5	122.6	20.5	15.0	24.0	20.0	17.0	2.5	122.6	25.8	17.0	24.0	31.
11	16	1207.1	20.5	14.0	33.0	18.0	17.0	2.6	122.6	25.8	17.0	33.0	20.0	17.0	2.6	122.6	25.8	17.0	33.0	26.0	21.0	2.6	122.5	31.7	20.0	33.0	34.
12	17	1392.0	25.8	15.0	40.0	20.0	21.0	2.6	122.5	31.7	20.0	40.0	26.0	21.0	2.6	122.5	31.7	20.0	40.0	30.0	27.0	2.6	122.5	37.9	21.0	40.0	42.
13	18	1454.8	31.7	17.0	46.0	27.0	27.0	2.6	122.5	37.9	21.0	46.0	30.0	27.0	2.6	122.5	37.9	21.0	46.0	32.0	31.0	2.6	122.5	44.1	24.0	46.0	45.
14	19	1426.1	37.9	20.0	51.0	30.0	31.0	2.7	122.5	44.1	24.0	51.0	32.0	31.0	2.7	122.5	44.1	24.0	51.0	39.0	35.0	2.7	122.5	50.0	34.0	51.0	47.
15	20	1357.4	44.1	21.0	56.0	33.0	35.0	2.7	122.5	50.0	34.0	56.0	39.0	35.0	2.7	122.5	50.0	34.0	56.0	42.0	44.0	2.7	122.5	52.9	37.0	56.0	48.
16	21	1293.9	50.0	24.0	57.0	40.0	44.0	2.8	122.5	52.9	37.0	57.0	42.0	44.0	2.8	122.5	52.9	37.0	57.0	44.0	56.0	2.8	122.5	54.6	39.0	57.0	49.
17	22	1255.5	52.9	34.0	58.0	42.0	56.0	3.1	122.5	54.6	39.0	58.0	44.0	56.0	3.1	122.5	54.6	39.0	58.0	45.0	56.0	3.1	122.5	55.9	41.0	58.0	51.
18	23	1256.0	54.6	37.0	59.0	44.0	56.0	3.5	122.5	55.9	41.0	59.0	45.0	56.0	3.5	122.5	55.9	41.0	58.0	46.0	56.0	3.5	122.5	57.0	41.0	58.0	52.
19																											
20																											

분석배경

• 문제 설명

NO	홍수사상번호	연	월	일	시간	유입량
1	26	2018	7	1	6	
2	26	2018	7	1	7	
3	26	2018	7	1	8	
4	26	2018	7	1	9	
5	26	2018	7	1	10	
6	26	2018	7	1	11	
7	26	2018	7	1	12	
8	26	2018	7	1	13	
9	26	2018	7	1	14	
10	26	2018	7	1	15	
11	26	2018	7	1	16	
12	26	2018	7	1	17	
13	26	2018	7	1	18	
14	26	2018	7	1	19	

평가데이터에는 사상번호/연/월/일/시간 이 주어지며 유입량을 예측해야함.

제공 데이터로부터 유입량 **예측이 필요한 시간**의 feature값을 활용할 수 있음.

2888	25	2017	7	18	17	523.0	36.5	10.0	0.0	1.0	1.0	3.1
2889	25	2017	7	18	18	513.4	22.8	6.0	0.0	1.0	1.0	3.0
2890	25	2017	7	18	19	502.8	8.3	2.0	0.0	1.0	1.0	3.0
2891	25	2017	7	18	20	492.0	4.1	1.0	0.0	1.0	1.0	3.0
2892	25	2017	7	18	21	481.1	3.4	1.0	0.0	1.0	1.0	3.0
2893	25	2017	7	18	22	470.5	3.3	1.0	0.0	1.0	1.0	2.9
2894	26	2018	7	1	6		14.3	32.0	0.0	0.0	0.0	1.9
2895	26	2018	7	1	7		11.0	20.0	1.0	0.0	0.0	1.9
2896	26	2018	7	1	8		7.9	11.0	5.0	0.0	0.0	1.9
2897	26	2018	7	1	9		7.9	3.0	11.0	0.0	0.0	1.9
2898	26	2018	7	1	10		13.3	4.0	25.0	1.0	8.0	1.9
2899	26	2018	7	1	11		20.1	13.0	48.0	14.0	24.0	2.0
2900	26	2018	7	1	12		27.0	18.0	58.0	19.0	33.0	2.0
2901	26	2018	7	1	13		34.5	20.0	60.0	22.0	36.0	2.0
2902	26	2018	7	1	14		36.0	22.0	63.0	23.0	37.0	2.0
2903	26	2018	7	1	15		37.7	21.0	67.0	25.0	37.0	2.0
2904	26	2018	7	1	16		39.7	22.0	68.0	25.0	40.0	2.1
2905	26	2018	7	1	17		42.4	21.0	71.0	25.0	41.0	2.2
2906	26	2018	7	1	18		49.6	22.0	74.0	26.0	41.0	2.4
2907	26	2018	7	1	19		57.2	41.0	75.0	29.0	51.0	2.7
2908	26	2018	7	1	20		61.7	64.0	75.0	51.0	62.0	2.8
2909	26	2018	7	1	21		64.8	69.0	78.0	59.0	68.0	2.9
2910	26	2018	7	1	22		68.5	70.0	82.0	65.0	71.0	3.0
2911	26	2018	7	1	23		72.6	72.0	86.0	65.0	77.0	3.0
2912	26	2018	7	1	24		75.8	73.0	86.0	70.0	81.0	3.3
2913	26	2018	7	2	1		76.8	74.0	89.0	73.0	83.0	3.5
2914	26	2018	7	2	2		78.4	75.0	91.0	75.0	85.0	3.7
2915	26	2018	7	2	3		79.4	78.0	92.0	76.0	85.0	3.8
2916	26	2018	7	2	4		85.4	78.0	100.0	77.0	85.0	4.1
2917	26	2018	7	2	5		94.0	80.0	103.0	77.0	93.0	4.4

• 사전 배경

홍수사상

- 토양은 일정 수준 비를 머금지만, 그 수준을 넘게 될 경우 포화상태가 되어 물이 토양 위를 미끄러지듯 흐르게 됨.
- 이 문제에서 홍수 사상임을 명시했으므로, 비가 미끄러지듯 흐르는 것으로 가정.

유입량과 강우량/수위의 시간차

- 비가 땅에 닿고 댐까지 흘러가는 시간차가 생기기에, 유입량과 강우량은 시간차를 두고 상관관계를 가질 것으로 가정
- 물은 밑에서부터 차므로, 유입량이 증가한 후에 수위가 높아질 것으로 추측

온도에 따른 변수

- 주어진 데이터는 6~9월에 한정되어 있으므로, 비 외의 다른 강우(ex.눈)는 발생하지 않았다고 가정

측정 오차 고려

- 유입량은 측정에 오차가 있을 가능성이 높은 점과 오차로 인해 예측 모델 성능 저하를 고려
- Kalman Filter를 통해 오차를 줄임

Preprocessing

• 제공 데이터 확인

결측값 확인

	홍수사상번호	연	월	일	시간	유입량	유역평균강수	강우(A지역)	강우(B지역)	강우(C지역)	...	강우(D지역).4	수위(E지역).4	수위(D지역).4	유역평균강수.5	강우(A지역).5	강우(B지역).5	강우(C지역).5	강우(D지역).5	수위(E지역).5	수위(D지역).5
0	0	0	0	0	0	160	0	0	0	0	...	0	0	0	0	0	0	0	0	0	0

1 rows × 48 columns

int : 홍수사상번호, 연, 월, 일, 시간, 강우
float : 유입량, 수위

→ 연, 월, 일, 시간은 시계열 데이터로 변환

	num	year	hour	target	average_rain_1	a_rain_1	b_rain_1	c_rain_1	d_rain_1	e_level_1	...
0	1	2006	2006-07-10 08:00:00	189.100000	6.4000	7	7	7	8	2.54	...
1	1	2006	2006-07-10 09:00:00	216.951962	6.3000	7	8	7	8	2.53	...
2	1	2006	2006-07-10 10:00:00	251.424419	6.4000	7	9	7	8	2.53	...
3	1	2006	2006-07-10 11:00:00	302.812199	7.3000	7	10	7	8	2.53	...
4	1	2006	2006-07-10 12:00:00	384.783406	8.2000	7	12	8	10	2.53	...
...	...	...	...	...	...	...	...	...	...	...	...
3046	26	2018	2018-07-07 17:00:00	NaN	2.3689	1	0	0	0	3.16	...
3047	26	2018	2018-07-07 18:00:00	NaN	2.3689	1	0	0	0	3.15	...
3048	26	2018	2018-07-07 19:00:00	NaN	2.3689	1	0	0	0	3.13	...
3049	26	2018	2018-07-07 20:00:00	NaN	2.3689	1	0	0	0	3.11	...
3050	26	2018	2018-07-07 21:00:00	NaN	2.3689	1	0	0	0	3.10	...

Preprocessing

• 사용하려 한 외부데이터

엘니뇨 감시 구역의 해수면온도 편차 (°C) → 기상청 기후정보포털
http://www.climate.go.kr/home/05_prediction_new/predict02_02.html

전국 평균 기온 (°C)
전국 평균 강수량 (mm)
전국 평균 풍속 (m/s)
전국 평균 습도 (%)
전국 평균 현지기압 (hPa)
전국 평균 해면기압 (hPa)
전국 평균 일조 (hr)
전국 평균 일사 (MJ/m2)
전국 평균 전운량 (10분위)
전국 평균 지면온도 (°C)

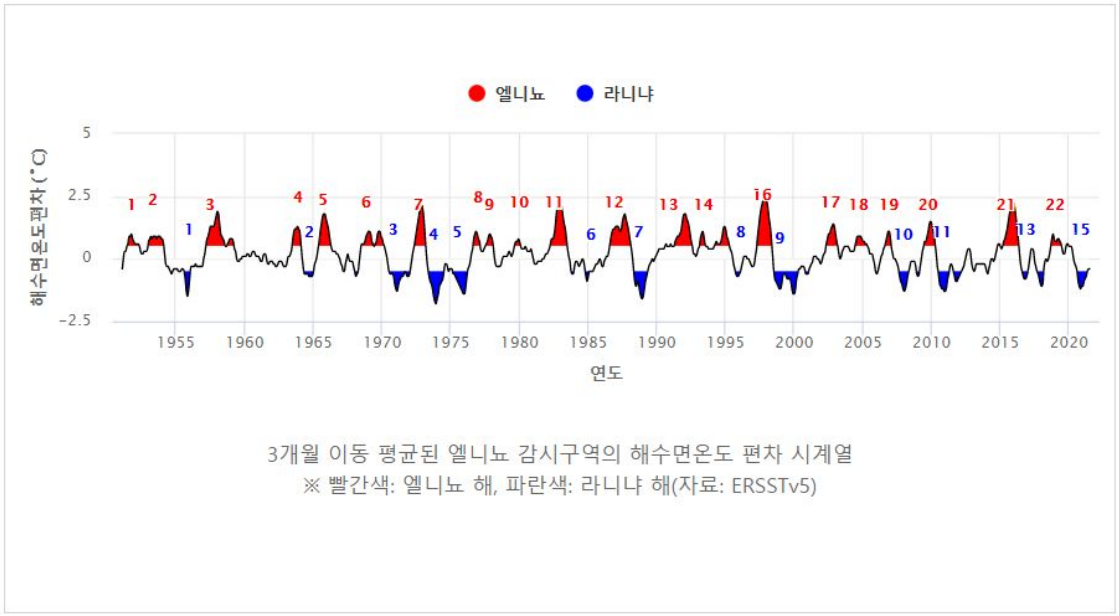
→ 기상청 기상자료개방포털
https://data.kma.go.kr/data/grnd/selectAsosRltmList.do?pgmNo=36

그러나,
외부데이터를 feature로 넣고 빼면서 모델링을 한 결과,
가장 좋은 성능을 내는 모델은
외부데이터를 하나도 쓰지 않는 모델 로 결론이 남

일조 : 태양광선이 땅위를 비추는 시간
일사 : 지표면에 도달하는 태양복사에너지

Preprocessing

• 엘니뇨 감시 구역의 해수면온도 편차 (℃)



3개월 이동 평균된 엘니뇨 감시구역의 해수면온도 편차 시계열
※ 빨간색: 엘니뇨 해, 파란색: 라니냐 해(자료: ERSSTv5)

	YR	MON	DATA
660	2006	1	-0.6
661	2006	2	-0.5
662	2006	3	-0.3
663	2006	4	-0.1
664	2006	5	0.1
665	2006	6	0.2
666	2006	7	0.4
667	2006	8	0.5
668	2006	9	0.7
669	2006	10	0.9
670	2006	11	1.1
671	2006	12	1.1

7월 중순 ~ 8월 중순까지 엘니뇨시기에 우리나라 강수가 증가하고,
라니냐 시기에는 강수가 감소하는 경향이 한반도 남부 지역을 중심으로 나타남.

엘니뇨가 발달하는 해 7월 중순부터 8월 중순에는 기후학적으로 북서태평양 고기압이 강해지는 시기로, **우리나라 남부에서는 강수가 증가하는 경향**이 나타남.

→ 이를 통해서, 유난히 강우량 및 유입량이 큰 사상들을 설명할 수 있을 것이라 판단

DATA의 **절댓값이 큰 양의 값**일수록, **강한 엘니뇨** 발생
DATA의 **절댓값이 큰 음의 값**일수록, **강한 라니냐** 발생

Preprocessing

• 전국 평균 날씨 데이터 – 기온, 강수, 풍속, 습도, 현지기압, 해면기압, 일조, 일사, 전운량, 지면온도

	지점	지점명	일시	기온(°C)	강수량(mm)	풍속(m/s)	습도(%)	현지기압(hPa)	해면기압(hPa)	일조(hr)	일사(MJ/m2)	전운량(10분위)	지면온도(°C)
0	90	속초	2006-07-10 08:00	20.9	NaN	0.7	88	NaN	NaN	0.0	NaN	NaN	NaN
1	90	속초	2006-07-10 09:00	20.8	NaN	1.0	90	1007.8	1009.7	0.0	NaN	10.0	22.8
2	90	속초	2006-07-10 10:00	20.9	NaN	1.0	90	NaN	NaN	0.0	NaN	NaN	NaN
3	90	속초	2006-07-10 11:00	20.9	NaN	0.7	90	NaN	NaN	0.0	NaN	NaN	NaN
4	90	속초	2006-07-10 12:00	20.5	0.5	2.1	92	1006.3	1008.3	0.0	NaN	10.0	NaN
...	...	...	...	...	...	...	...	...	...	...	...	...	...
37123	295	남해	2006-07-29 23:00	25.3	NaN	0.7	87	NaN	NaN	NaN	NaN	NaN	NaN
37124	295	남해	2006-07-30 00:00	25.0	NaN	0.4	89	1008.0	1013.1	NaN			
37125	295	남해	2006-07-30 01:00	24.7	NaN	0.8	91	NaN	NaN	NaN			
37126	295	남해	2006-07-30 02:00	24.2	NaN	0.5	90	NaN	NaN	NaN			
37127	295	남해	2006-07-30 03:00	23.7	NaN	0.1	90	1008.0	1013.1	NaN			

2006년 전국 지점별 날씨 데이터

다양한 전국 날씨 데이터를 feature로 활용하여,
유입량의 변화를 다양하게 설명하려 함.

날짜별로 groupby 후 값을 평균

	일시	지점	기온(°C)	강수량(mm)	풍속(m/s)	습도(%)	현지기압(hPa)	해면기압(hPa)	일조(hr)	일사(MJ/m2)	전운량(10분위)	지면온도(°C)
0	2006-07-10 08:00	189.692308	22.541026	5.332759	5.370513	87.153846	NaN	NaN	0.023077	0.187273	NaN	NaN
1	2006-07-10 09:00	189.692308	22.782051	7.398361	5.694872	86.423077	988.253846	1000.793590	0.019231	0.313182	9.536585	22.843590
2	2006-07-10 10:00	189.692308	22.902564	7.481818	6.108974	85.512821	NaN	NaN	0.029487	0.414091	NaN	NaN
3	2006-07-10 11:00	189.692308	22.846154	8.184615	5.898718	85.461538	NaN	NaN	0.035897	0.461818	NaN	NaN
4	2006-07-10 12:00	189.692308	22.811538	8.247826	6.357692	85.192308	985.739744	998.265385	0.047436	0.519091	9.609756	NaN
...	...	...	...	...	...	...	...	...	...	...	...	...
6534	2018-07-07 17:00	196.673684	23.736842	0.000000	3.004211	63.652632	997.648421	1009.783158	0.531579	1.163023	8.000000	28.777895
6535	2018-07-07 18:00	196.673684	22.954737	0.000000	3.092632	65.715789	997.872632	1010.055789	0.438947	0.712558	8.304348	25.890526
6536	2018-07-07 19:00	196.673684	21.862105	0.000000	2.729474	69.242105	998.196842	1010.430526	0.206316	0.283488	10.000000	23.497895
6537	2018-07-07 20:00	196.673684	20.910526	NaN	2.470526	71.978947	998.613684	1010.903158	0.035789	0.045349	10.000000	21.838947
6538	2018-07-07 21:00	196.673684	20.306316	0.000000	2.471579	73.631579	999.347368	1011.670526	0.000000	0.000000	7.304348	20.940000

Preprocessing

• 전국 평균 날씨 데이터 – 기온, 강수, 풍속, 습도, 현지기압, 해면기압, 일조, 일사, 전운량, 지면온도

average_rain_1	a_rain_1	b_rain_1	c_rain_1	...	기온	강수량	풍속	습도	현지기압	해면기압	일조	일사	전운량	지면온도
6.4000	7	7	7	...	22.541026	5.332759	5.370513	87.153846	988.253846	1000.793590	0.023077	0.187273	9.536585	22.843590
6.3000	7	8	7	...	22.782051	7.398361	5.694872	86.423077	988.253846	1000.793590	0.019231	0.313182	9.536585	22.843590
6.4000	7	9	7	...	22.902564	7.481818	6.108974	85.512821	987.415812	999.950855	0.029487	0.414091	9.560976	22.994444
7.3000	7	10	7	...	22.846154	8.184615	5.898718	85.461538	986.577778	999.108120	0.035897	0.461818	9.585366	23.145299
8.2000	7	12	8	...	22.811538	8.247826	6.357692	85.192308	985.739744	998.265385	0.047436	0.519091	9.609756	23.296154
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
166.8818	232	141	60	...	22.815957	0.787500	0.711702	95.287234	997.530108	1009.855914	0.007447	0.028049	9.000000	23.620213
159.0198	230	136	60	...	23.370213	0.320000	0.659574	93.797872	998.007527	1010.307527	0.032979	0.204146	8.954545	24.843617
153.1347	229	134	60	...	24.417021	0.116667	0.811702	89.744681	998.464516	1010.735484	0.141489	0.540732	8.363636	26.901064
145.1249	214	134	60	...	25.708511	0.058333	1.090426	83.638298	998.619355	1010.832258	0.390426	1.012927	7.954545	29.941489
139.1833	208	134	60	...	26.807447	0.000000	1.365957	77.851064	998.578495	1010.734409	0.546809	1.588049	7.454545	33.018085

NaN값 처리 완료한 전국 평균 날씨 데이터

< NaN값 처리 과정 >

먼저, 제공데이터와 외부데이터를 날짜에 맞춰 merge

사상별로 나눠서, NaN의 값을 pandas의 interpolate 함수를 이용하여 처리

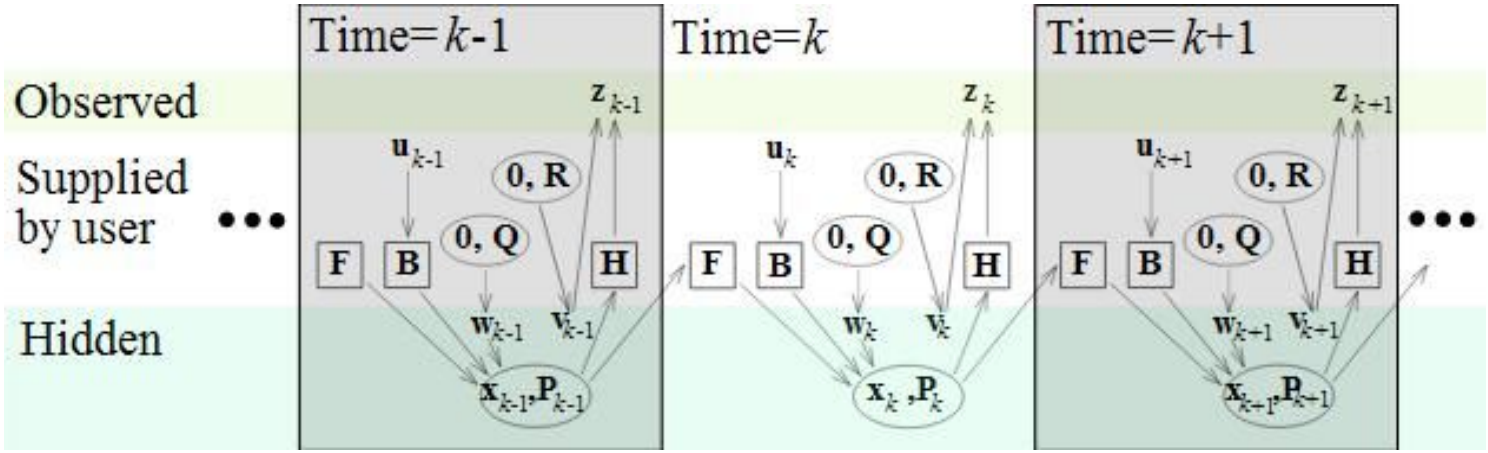
→ 그 중에, 사상별로 첫번째 관측치이면서 NaN인 값들 존재
(interpolate 함수로 처리 불가능)

→ 그런 값들은 NaN값이 아닌 가장 가까운 미래의 값을 그대로 사용

Preprocessing

- **칼만필터란**

$$\mathbf{x}_k = \mathbf{F}_k\mathbf{x}_{k-1} + \mathbf{B}_k\mathbf{u}_k + \mathbf{w}_k$$



순차적으로 나열되어 있는 raw 시계열 데이터에서
이전 데이터의 값을 이용하여 **측정 과정에서 생긴 오차**를 예측
이를 바탕으로 오차를 줄여 더 정확한 데이터로 업데이트
이 과정을 계속해서 반복
활용 예시) 노이즈 캔슬링, 레이더 추적

→ python 의 filterpy 모듈 사용

Preprocessing

- **칼만필터란** – 칼만필터 적용 후

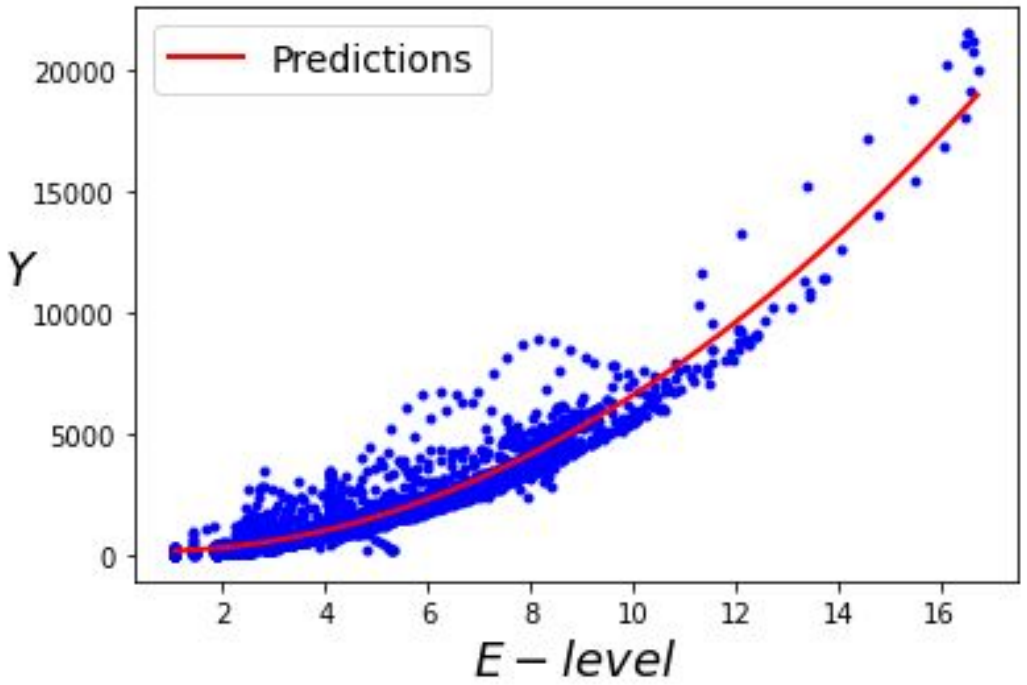
	average_rain_1	a_rain_1	b_rain_1	c_rain_1	d_rain_1	e_level_1	d_level_1	f3	f4	f5	f6	f7	f8	f9
0	6.4000	7	7	7	8	2.54	122.56875	6.389027	6.987531	6.987531	6.987531	7.985037	2.538653	122.268080
1	6.3000	7	8	7	8	2.53	122.56250	6.322118	7.024033	8.014347	7.024033	8.028840	2.532692	123.142091
2	6.4000	7	9	7	8	2.53	122.55625	6.381001	7.016413	9.010653	7.016413	8.019695	2.530129	122.952057
3	7.3000	7	10	7	8	2.53	122.55625	7.029268	7.012387	10.008417	7.012387	8.014864	2.529342	122.853079
4	8.2000	7	12	8	10	2.53	122.55625	7.846669	7.009934	11.606027	7.609010	9.210073	2.529072	122.793303
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
386	22.7836	6	0	1	1	3.03	137.16875	41.474755	12.445231	18.359081	-17.084711	-10.378163	2.932325	137.283189
387	8.2586	2	0	1	1	3.00	137.18125	35.274680	9.830399	13.595709	-17.256941	-10.927851	2.916983	137.293783
388	4.1089	1	0	1	1	2.98	137.18750	29.040882	7.363832	9.393223	-17.165139	-11.237339	2.902000	137.302242
389	3.3854	1	0	1	1	2.96	137.19375	23.336253	5.196495	5.719401	-16.851387	-11.339803	2.887253	137.308867
390	3.2841	1	0	1	1	2.94	137.20625	18.239958	3.310228	2.540139	-16.354134	-11.265968	2.872639	137.314904

제공데이터에서 각 사상별로 따로 칼만필터 적용 후, column 추가

average_rain_1 → f3 / a_rain_1 → f4 / b_rain_1 → f5 / c_rain_1 → f6 / d_rain_1 → f7 / e_level_1 → f8 / d_level_1 → f3

Preprocessing

• E지역 수위를 이용한 pred 값



유입량(target)과 E지역 수위의
correlation 값이 매우 큰 양의 값(0.9)임을 이용함

- E 지역 수위를 변수로 한
  Polynomial Regression 을 통해 유입량 예측
- 이를 inplace하여 하나의 피쳐로 활용

Preprocessing

• 피쳐들의 변화량 구하기

e_level_1(수위 E지역)을 통해 전체적인 유입량의 흐름을 예측하고 2시간전과 1시간전으로부터의 강수량 변화를 측정하여 유입량의 증감을 설명할수 있을것이라는 가설을 세움

A,B,C,D 지역에서 측정된 강우가 댐까지 유입되는 데에는 일정 시간이 걸릴 것이라고 생각함
따라서, 제공데이터에서의 각 지역별 강수량을 1, 2시간씩 미뤄 유입량 예측에 사용

	시간	유입량	데이터집단 1						
			역평균강	강우(A지역	강우(B지역	강우(C지역	강우(D지역	수위(E지역	수위(D지역
3	8	189.1	6.4	7.0	7.0	7.0	8.0	2.5	122.6
4	9	217.0	6.3	7.0	8.0	7.0	8.0	2.5	122.6
5	10	251.4	6.4	7.0	9.0	7.0	8.0	2.5	122.6
6	11	302.8	7.3	7.0	10.0	7.0	8.0	2.5	122.6
7	12	384.8	8.2	7.0	12.0	8.0	10.0	2.5	122.6
8	13	512.5	11.3	7.0	14.0	10.0	11.0	2.5	122.6
9	14	701.5	14.4	9.0	17.0	10.0	14.0	2.5	122.6
10	15	952.5	16.9	12.0	24.0	15.0	16.0	2.5	122.6
11	16	1207.1	20.5	14.0	33.0	18.0	17.0	2.6	122.6
12	17	1392.0	25.8	15.0	40.0	20.0	21.0	2.6	122.5
13	18	1454.8	31.7	17.0	46.0	27.0	27.0	2.6	122.5
14	19	1426.1	37.9	20.0	51.0	30.0	31.0	2.7	122.5
15	20	1357.4	44.1	21.0	56.0	33.0	35.0	2.7	122.5
16	21	1293.9	50.0	24.0	57.0	40.0	44.0	2.8	122.5
17	22	1255.5	52.9	34.0	58.0	42.0	56.0	3.1	122.5
18	23	1256.0	54.6	37.0	59.0	44.0	56.0	3.5	122.5

Preprocessing

• 이상치 제거

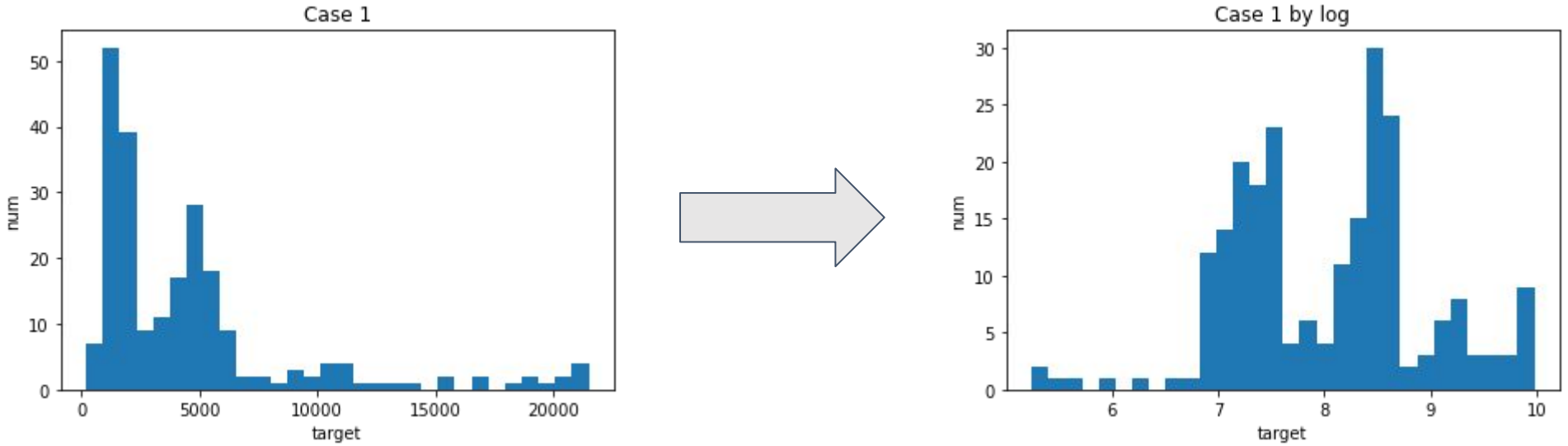
- 강건한 모델학습을 위해 target(유입량 기준으로)을 기준으로 각 사상별 상위, 하위 1프로 제거
- 모델 학습 train_set에만 적용하고 test_set과 26번 예측에는 사용하지 않음

```
삭제하는 데이터 인덱스:  Int64Index([0, 1, 2, 151, 152, 153], dtype='int64')
삭제하는 데이터 인덱스:  Int64Index([244, 309], dtype='int64')
삭제하는 데이터 인덱스:  Int64Index([337, 370], dtype='int64')
삭제하는 데이터 인덱스:  Int64Index([408, 432], dtype='int64')
삭제하는 데이터 인덱스:  Int64Index([455, 476], dtype='int64')
삭제하는 데이터 인덱스:  Int64Index([535, 564], dtype='int64')
삭제하는 데이터 인덱스:  Int64Index([585, 608], dtype='int64')
삭제하는 데이터 인덱스:  Int64Index([645, 663], dtype='int64')
삭제하는 데이터 인덱스:  Int64Index([687, 717], dtype='int64')
삭제하는 데이터 인덱스:  Int64Index([754, 796], dtype='int64')
삭제하는 데이터 인덱스:  Int64Index([844, 845, 901, 902], dtype='int64')
삭제하는 데이터 인덱스:  Int64Index([944, 1003, 1016, 1017], dtype='int64')
삭제하는 데이터 인덱스:  Int64Index([1076, 1105], dtype='int64')
삭제하는 데이터 인덱스:  Int64Index([1131, 1158], dtype='int64')
삭제하는 데이터 인덱스:  Int64Index([1186, 1187, 1188, 1234, 1235, 1236], dtype='int64')
삭제하는 데이터 인덱스:  Int64Index([1463, 1464, 1465, 1486, 1487, 1488], dtype='int64')
삭제하는 데이터 인덱스:  Int64Index([1650, 1651, 1675, 1676], dtype='int64')
삭제하는 데이터 인덱스:  Int64Index([1783, 1810], dtype='int64')
삭제하는 데이터 인덱스:  Int64Index([1868, 1906], dtype='int64')
삭제하는 데이터 인덱스:  Int64Index([1933, 1934, 1935, 2142, 2143, 2144], dtype='int64')
삭제하는 데이터 인덱스:  Int64Index([2199, 2213], dtype='int64')
삭제하는 데이터 인덱스:  Int64Index([2300, 2301, 2397, 2479, 2480, 2481], dtype='int64')
삭제하는 데이터 인덱스:  Int64Index([2618, 2619, 2637, 2638], dtype='int64')
삭제하는 데이터 인덱스:  Int64Index([2709, 2768], dtype='int64')
삭제하는 데이터 인덱스:  Int64Index([2808, 2809, 2836, 2837], dtype='int64')
```

← 각 사상별 삭제된 인덱스

Preprocessing

- log



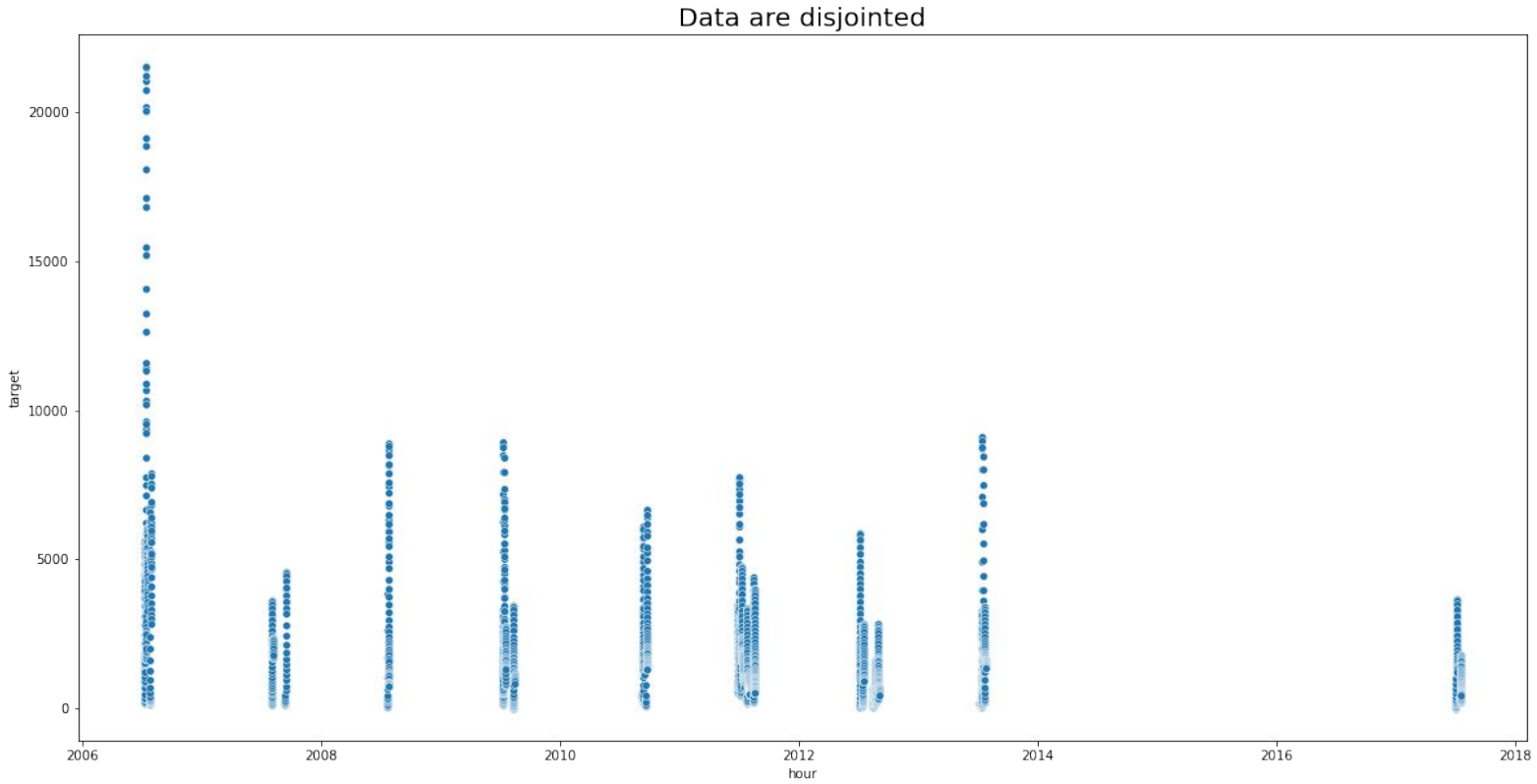
- target값이 편향되어 있기 때문에 skewness값이 매우 높다.
- 왜도를 줄이기 위해 log변환을 사용한다.
- 모델 예측값에 다시 지수함수를 통해 transform해준다.

-> 더 nomarlly 분포된 데이터를 통해 편향되지 않은 학습 가능.

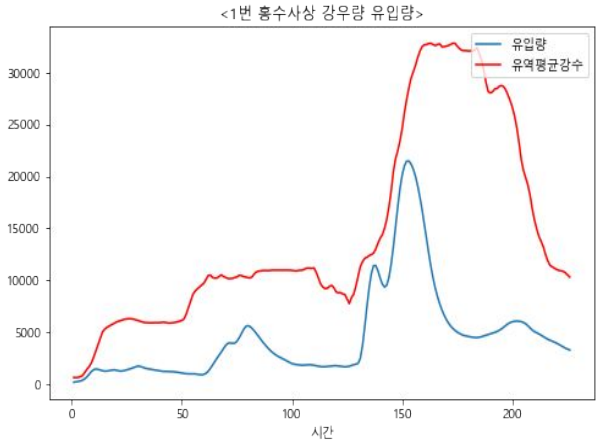
- **제공 데이터 분석** – 특징 1 : 불연속 시계열

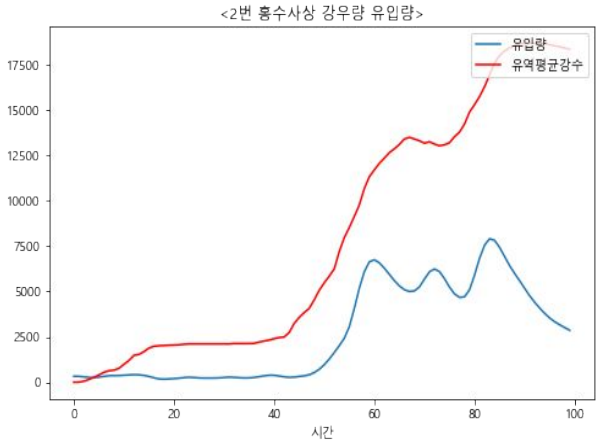
각 홍수사상간 데이터 불연속 발생

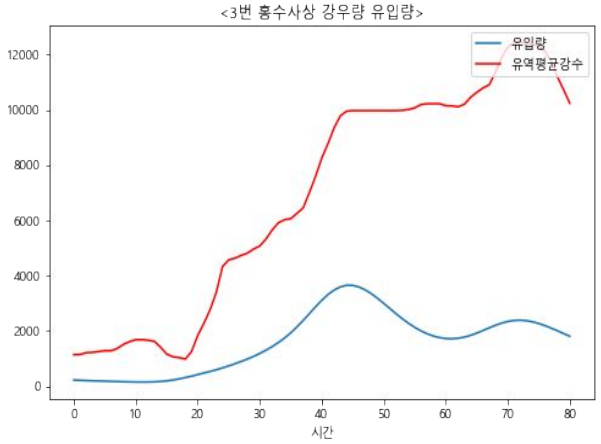
하나의 사상과 다음 사상 사이 데이터가 제공되지 않음.

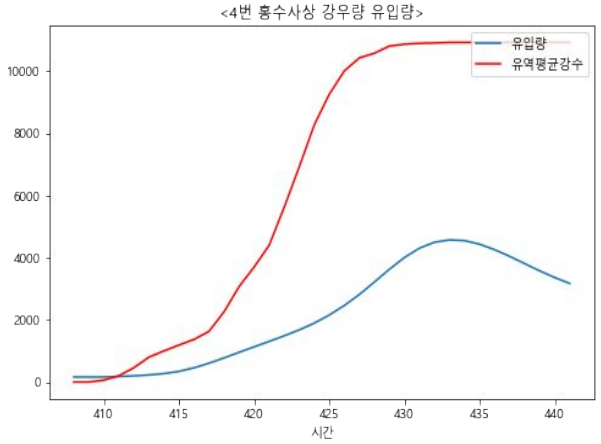


• 제공 데이터 분석 – 특징 2 : 속성간 시간차



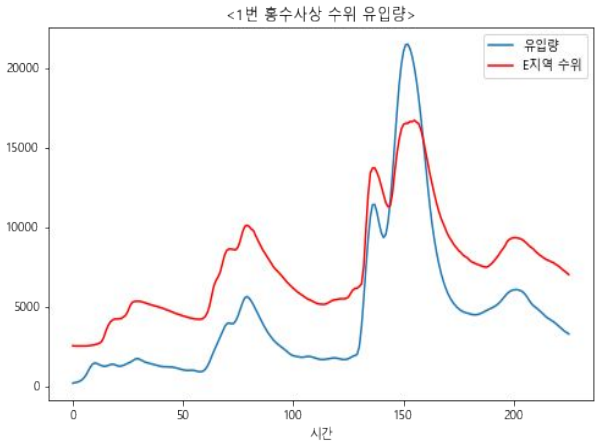


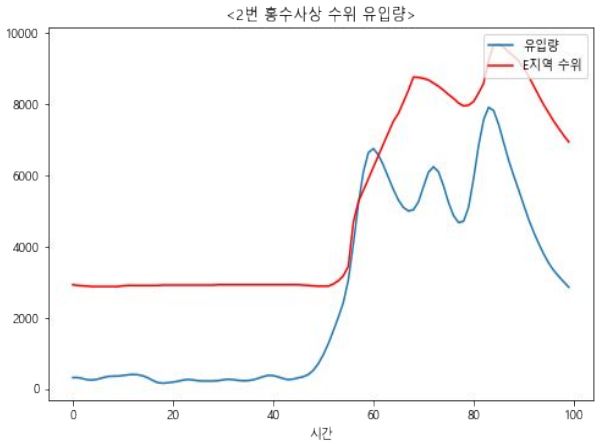


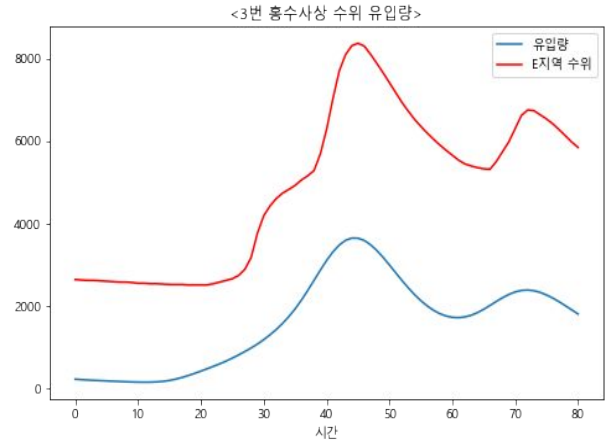


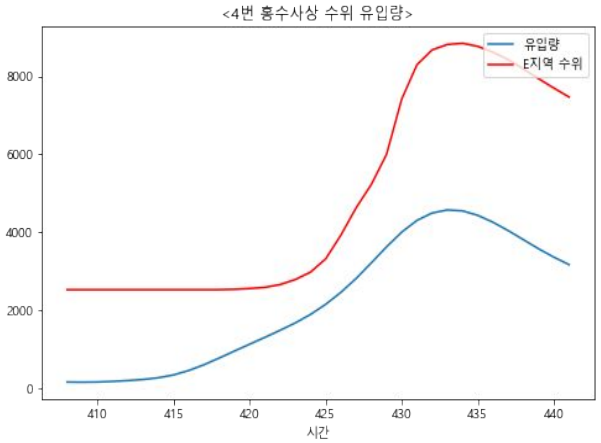
다음은 각 홍수사상 1~4번의 유역평균강수와 유입량을 비교한 것이다. 데이터 크기 차이를 고려하여 유역평균 강수에 100을 곱해주었다. 예상과 비슷하게 유입량보다 평균 강수가 먼저 증가하는 경향을 볼 수 있다.

• 제공 데이터 분석 – 특징 2 : 속성간 시간차









다음은 홍수사상번호 1~4번의 댐 유입량과 E지역 수위를 비교한 것이다. 상관관계 분석 결과 유입량과 D지역 수위는 크게 관련이 없기에, E지역 수위만 비교하였다. 데이터 크기 차이를 고려하여 수위 데이터에 1000을 곱해주었다. 예상과 같이 유입량보다 조금 늦게 수위 데이터가 변하는 경향을 확인할 수 있다.

• 제공 데이터 분석 – 변수간 상관관계 분석

데이터 집단별 평균값에 대한 상관관계 분석

- E 지역 수위가 가장 큰 상관관계(0.9)를 갖는다.
- D 지역의 수위는 전체적으로 상관관계를 갖지 않는다.
- D 지역 수위의 평균은 131.8로 E지역 수위 평균 4.6과 대비된다.
- 상대적으로 D 지형의 규모가 크기 때문에 강우량으로 인한 수위 변화가 작으리라 짐작할 수 있다.
- 반면 E 지역의 수위는 규모가 작기 때문에 조금의 강우에도 민감하게 반응한다.

-> E지역 수위로 **다항회귀분석**을 통해 유입량을 예측해본 후 inplace

- **모델 선택**

  - LSTM

  - CatBoostRegressor

  - LightGBM

  - XGBoost

주어진 과제를 시계열 예측으로서 해결할 전략을 수립.
시계열 데이터를 활용하기 적합한 LSTM을 우선적으로 검토, 구현하였음.

But, LSTM의 성능이 목표치에 부합하지 못하였음.

제공된 **데이터 수 부족** 및 **사상 간의 불연속성** 으로 인한 것으로 판단되어 탈락.

• 모델 선택

- LSTM

- **CatBoostRegressor**

- LightGBM

- XGBoost

Catboost는 ordered boosting으로
각 iteration마다 이전에 썼던 데이터를 반복해서 쓰는 다른 부스팅모델에서 나타는 target leakage로 인한 prediction 문제를 해결하였다.

이에 **과거의 정보**를 바탕으로 **현재의 유입량**을 예측하는 데 효과적일 것이라고 생각했다.

실제 검증결과 다른 부스팅 모델(Xg boost, lgbm)보다 좋은 성능을 보여 최종적으로 해당 모델을 선택했다.

• 검정 방법 – K-Fold Cross Validation

1. 검증 데이터는 26번 사상을 제거하여 사용
2. X_train, y_train 으로 모델 학습
3. X_test 를 통해 prediction 도출
4. prediction과 y_test를 통해 rmse 계산
5. 각 사상을 test_set 으로 했을 때의 rmse의 평균 계산

- X_train: 1~25번 사상 feature
- y_train: 1~25번 사상 유입량

- X_test : 1번 사상 features -> prediction1
- y_test : 1번 사상 유입량

->rmse1(prediction1, y_test)

- X_train: 1, 3~25번 사상 feature
- y_train: 1, 3~25번 사상 유입량

- X_test : 2번 사상 features -> prediction2
- y_test : 2번 사상 유입량

->rmse2(prediction2, y_test)

...

- X_train: 1~2, 4~25번 사상 feature
- y_train: 1~2, 4~25번 사상 유입량

- X_test : 3번 사상 features -> prediction 1
- y_test : 3번 사상 유입량

->rmse3(prediction3, y_test)

average_rmse = (rmse1+rmse2+...+rmse25) / 25

• 검정 결과 및 예상 rmse

다양한 외부데이터와 feature들을 넣고 빼면서
생성한 model들을 비교한 결과,

- 제공 데이터 (num, month, day, hour 제외)
- 제공 데이터에 칼만 필터 적용 feature
- feature들의 변화량
- 이상치 제거

를 포함시킨 데이터로 학습시킨 model의 성능이
가장 뛰어난 것으로 결론남

예상 rmse : 546.8154646460476

```
result
```

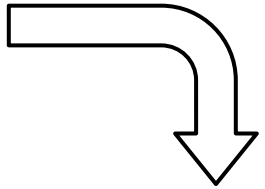
```
[2928.979696165974,
 840.400747552113,
 438.8796675792989,
 367.6393251316601,
 1111.4580813980176,
 234.89197648500607,
 710.2214500148465,
 861.3610755314774,
 357.4109468353351,
 234.78061359106235,
 407.74845434555306,
 339.2141061019587,
 306.0475256704718,
 476.5541173603288,
 315.55370118582243,
 312.5858899975527,
 426.27883753727923,
 530.4768600795812,
 140.2924021595479,
 190.9153989037621,
 515.5424981756258,
 267.37832687165337,
 678.5220722220336,
 506.0065823842164,
 171.24626287101404]
```

```
np.mean(result)
```

```
546.8154646460476
```

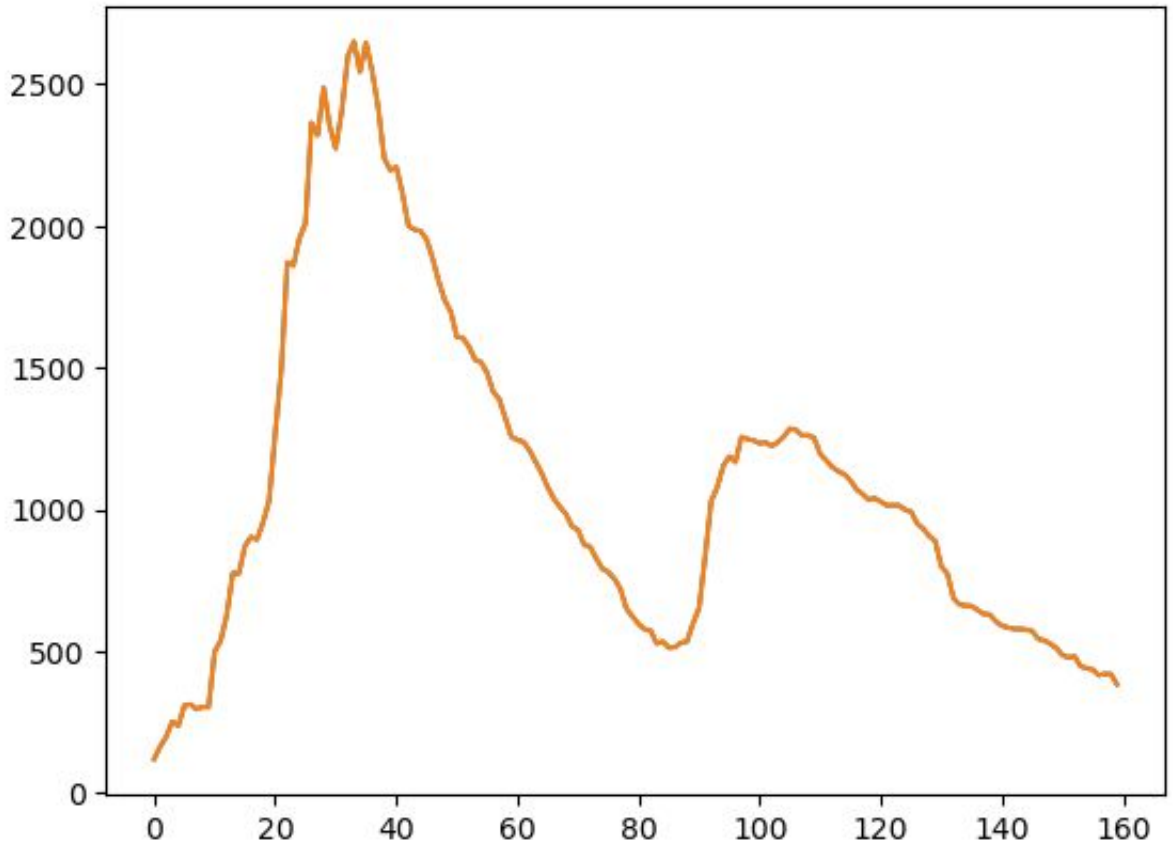
• 26번 사상 댐 유입량 예측 결과

NO	홍수사상번호	연	월	일	시간	유입량
1	26	2018	7	1	6	
2	26	2018	7	1	7	
3	26	2018	7	1	8	
4	26	2018	7	1	9	
5	26	2018	7	1	10	
6	26	2018	7	1	11	
7	26	2018	7	1	12	
8	26	2018	7	1	13	
9	26	2018	7	1	14	
10	26	2018	7	1	15	
11	26	2018	7	1	16	
12	26	2018	7	1	17	
13	26	2018	7	1	18	
14	26	2018	7	1	19	
15	26	2018	7	1	20	
16	26	2018	7	1	21	
17	26	2018	7	1	22	
18	26	2018	7	1	23	
19	26	2018	7	1	24	
20	26	2018	7	2	1	
21	26	2018	7	2	2	
22	26	2018	7	2	3	
23	26	2018	7	2	4	
24	26	2018	7	2	5	
25	26	2018	7	2	6	



array([119.82908498, 163.32505731, 198.50656671, 252.39349557, 237.24592071, 308.4769125 , 312.85721749, 295.85075879,
303.95422509, 302.62709915, 498.20422338, 537.16974312, 627.45157691, 776.10363764, 773.74395575, 870.8206074 ,
903.05153202, 894.10412944, 954.3563796 , 1031.75461577, 1266.8213365 , 1488.30183151, 1871.21712297, 1860.07446406,
1956.08432485, 2008.99000603, 2363.1828904 , 2319.43016445, 2486.75288426, 2352.16610465, 2273.23076044, 2405.32037985,
2598.74522815, 2650.05691298, 2543.2655577 , 2646.57315516, 2547.26476506, 2418.68227975, 2239.72139481, 2196.77304402,
2207.51098602, 2117.34842746, 1999.95500472, 1986.83672446, 1981.89184583, 1952.48466528, 1886.36893903, 1804.59319369,
1737.99401411, 1697.51604313, 1608.51666462, 1604.87306421, 1574.58089069, 1528.0770363 , 1519.18333286, 1481.6880111 ,
1415.20214363, 1389.10453174, 1321.52379387, 1256.71669123, 1244.94693322, 1238.488619 , 1208.65810626, 1169.19531728,
1127.97976559, 1078.63321453, 1038.5369454 , 1009.74802579, 984.52951547, 941.31573677, 928.2085797 , 875.27918645,
869.69332756, 827.76772501, 791.91255349, 779.5602999 , 755.44355 , 720.5139648 , 650.48053507, 622.58293674,
596.5542048 , 577.44462693, 573.15519618, 527.3526179 , 534.30121619, 512.68709914, 514.73397389, 530.0665503 ,
534.61169311, 599.0925443 , 653.11857046, 835.21012071, 1030.33969612, 1078.40572357, 1154.47557713, 1187.20587022,
1167.46850539, 1255.6449739 , 1248.37770573, 1243.41545995, 1233.25759037, 1236.46310164, 1223.29993605, 1237.10239117,
1258.04772096, 1285.06535195, 1280.43179424, 1261.34243605, 1261.72049024, 1252.48041808, 1195.6904345 , 1172.71359602,
1149.24754346, 1134.70562614, 1123.98668609, 1101.90285248, 1070.10472922, 1054.86651758, 1034.14430293, 1040.26423987,
1027.68158152, 1014.733049 , 1014.733049 , 1014.733049 , 998.89898398, 992.31739926, 950.58909867, 933.00882138,
906.13831395, 888.75377765, 799.66203572, 772.83739393, 687.54564101, 664.67829475, 661.03782878, 659.48142631,
645.88503789, 631.34673946, 629.7603587 , 605.46803604, 590.28485076, 583.79603124, 578.60690115, 578.73146948,
576.22990511, 571.56601154, 543.51134675, 538.02325153, 526.46511422, 511.19510252, 485.76524534, 478.50132156,
481.85513023, 446.8583194 , 440.17499776, 435.22236175, 415.71285576, 420.30862264, 419.47001698, 382.99669618])

- **시각화**



예측된 26번 사상 유입량 데이터

• 결론 및 한계점

결론

- 댐 유입량 예측은 실시간으로 이루어져야한다. 따라서 뒤에 나오는 정보들을 이용할 수 없다.
- 이 모델은 예측해야하는 시간의 앞부분의 정보만 사용하기에 현실에 더 적합한 모델이다.
- 검정 결과 26번 사상 예측값의 rmse는 약 546이 될 것이다.

한계

- 사상별 rmse 편차가 크다.
- 1번 사상과 같은 예외적인 상황을 통제할 수 있는 변수 탐색에 실패했다.
- 데이터의 불연속성으로 인해 각 사상별 초기 값의 예측이 어렵다.
- 다만 사상별 데이터가 아닌 연속적 데이터가 주어진다면 더 정확한 모델링이 가능할 것이다. 추가로 lstm 모델을 고려할 수 있다.

감사합니다