

한강 유역 수위 예측

Team kjh

2022. 09. 05



CONTENTS

I Data flow chart

II 외부데이터찾기

III 결측치 대체

IV 파생변수 생성

V 변수선택

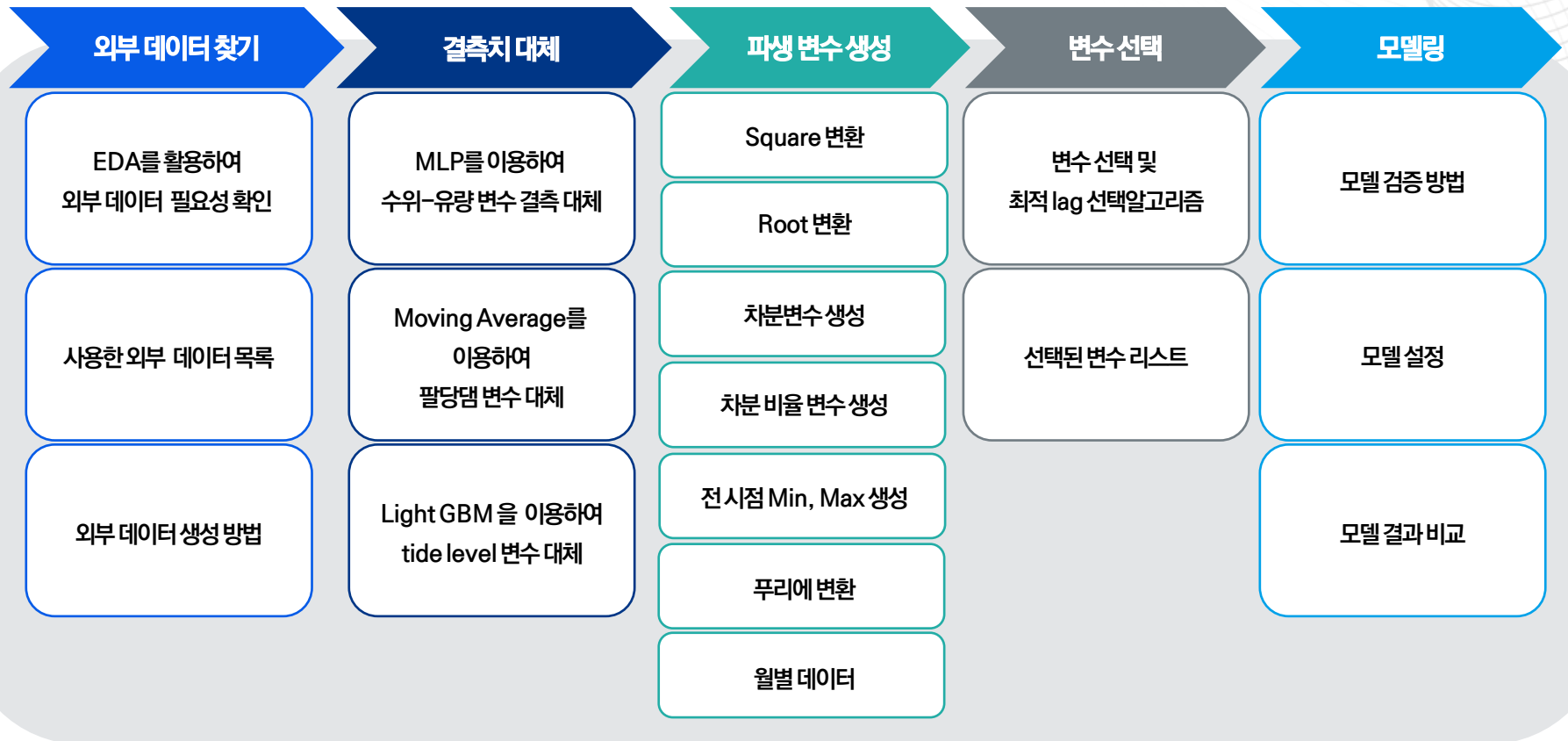
VI 모델링

Chapter

I

Data flow chart

Data flow chart



Chapter

II

외부 데이터 찾기

외부 데이터 찾기

- EDA를 활용하여 외부 데이터 필요성 확인

청담대교, 잠수교, 한강대교, 행주대교의 수위를 EDA를 통해서

확인 해본 결과 오른쪽 그래프처럼 하류(행주대교) 부터 수위의 변동이

일어나는 것을 확인해 볼 수 있었다.

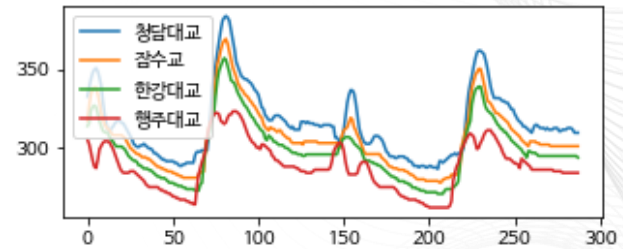
따라서, 행주대교 이전 본류, 지류데이터부터 청담대교

이후 본류, 지류데이터를 이용하여 예측을 진행하면 더욱 더

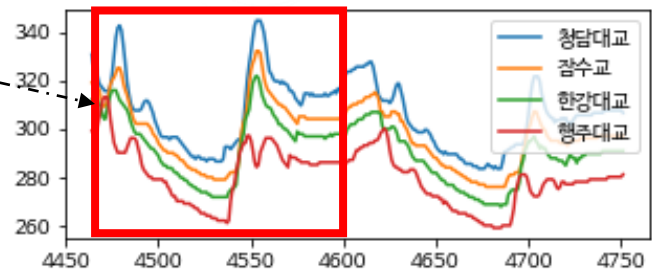
좋은 성능을 보일 수 있을 것이라 판단하였다.

위와 같은 이유로 우리는 6개의 관측소의 수위, 유량 데이터와

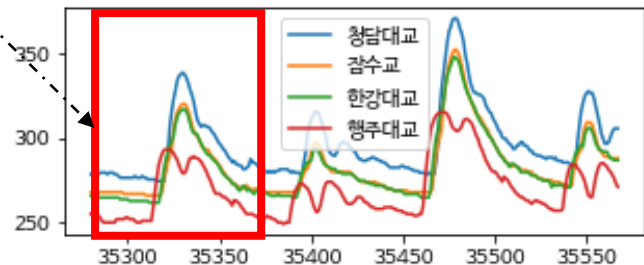
4개의 관측소의 강수량 데이터를 이용하였다.



2013-05-01 ~ 2013-05-02 10분 단위 데이터



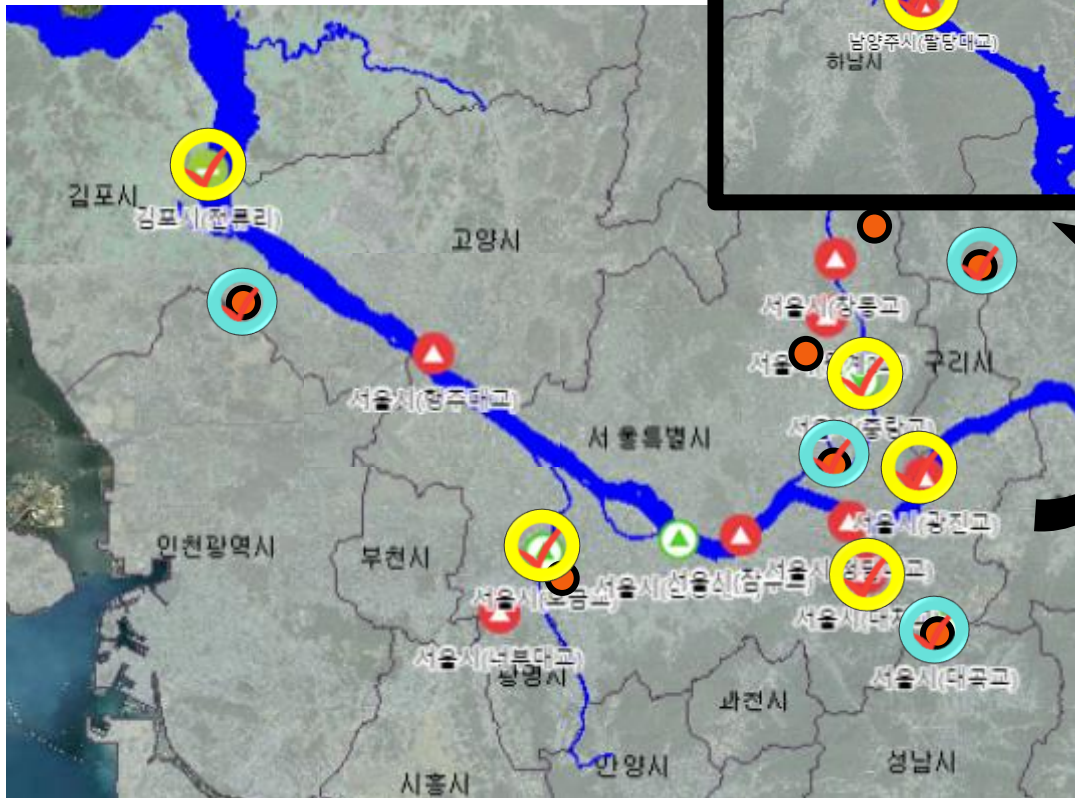
2013-06-01 ~ 2013-06-02 10분 단위 데이터



2019-07-01 ~ 2019-07-02 10분 단위 데이터

외부 데이터 찾기

• 사용한 외부 데이터 목록



: 사용한 외부 데이터 (수위, 유량)

출처: 한강홍수통제소 hrfco.go.kr/sumun/interest.do

2022-09-05 실시간 수문자료



: 사용한 외부 데이터 (강수)

1. 사용한 수위 데이터 관측소

관측소 코드	관측소 명
1018610	팔당대교
1018640	광진교
1018658	대치교
1018675	중랑교
1018697	오금교
1019675	김포시(전류리)

2. 사용한 강수 데이터 관측소

관측소 코드	관측소 명
10184100	대곡교
10184110	남양주시(진관교)
10184140	송정동
10194010	김포시(김포시청)

외부 데이터 찾기

- 사용한 외부 데이터와 데이터 생성 방법

한국홍수통제소 홈페이지([한강홍수통제소](#))에서 api ServiceKey를 받아 데이터를 생성하였다.

- Api를 이용해 받은 데이터는 모두 2012년 ~ 2022년 각각 5월~8월 데이터이다.

- Api URL 자료 요청 포맷

- [수위](#)
- [강수량](#)

- 사용한 외부 데이터 목록

- 강수량 : 김포시(김포시청)
- 수위 : 팔당대교, 광진교, 대치교, 중랑교, 오금교, 김포시(전류리)

- 사용한 외부데이터 링크

https://drive.google.com/file/d/1it6J1ZKj8lpWBELZhjJZdQC5f5L_5Wm/view?usp=sharing

Chapter

III

결측치 대체

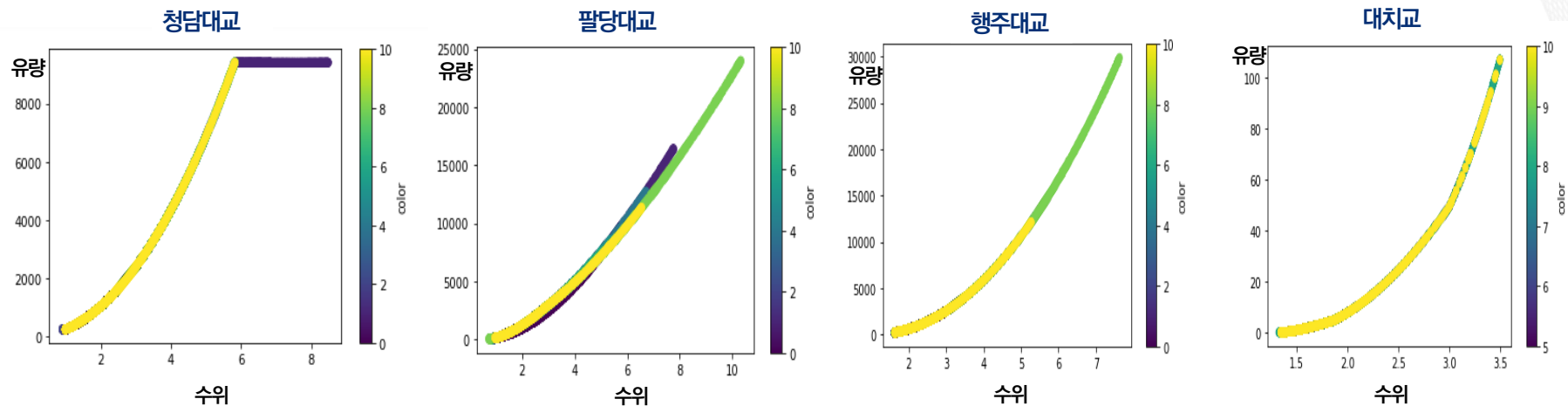
결측치 대체

결측치 대체 방법

1. 결측치의 유무와 개수를 확인한 결과, 잠수교의 유량 (fw_1018680)은 결측값과 0으로만 되어 있어서 결측값을 대체 할 수 없었다. 따라서 잠수교 유량 변수를 제거하였다.
2. **Data leakage**에 주의하여 미래(t+1)시점을 예측하기 위해 현재(t)시점의 결측을 현재(t)시점의 변수들을 이용하여 결측을 대체하였다. (모델 가정 : t를 이용하여 t+1 예측)
3. 관계식 모형 추정시 train data만 사용하여 추정(**data leakage 유의**)
4. 변수 마다 EDA를 통해 서로 다른 결측 대체 방법을 적용 하였다.
 - 유량, 수위 : 유량 수위 곡선 관계식을 딥 러닝(Multi-Layer Perceptron; MLP)으로 추정하여 대체
 - 팔당댐 변수(swl,inf,sfw,ecpc,tototf) : Moving average를 이용하여 대체
 - 팔당댐 변수(tide_level) : Light GBM을 이용하여 대체

결측치 대체

- 유량 수위 곡선관계식과 딥 러닝(MLP)을 이용한 수위, 유량 변수 결측치 대체
 - 유량과 수위의 관계식은 $Q = p(h - e)^\beta - \gamma$ (Chang, Ki-Hwan, & 이재형(2005))를 따른다고 알려져 있었지만 유량과 수위의 scatter plot을 확인해본 결과, 다음과 같이 정확하게 관계식에 맞는 곡선 형태는 아니었다.
 - 수위-유량의 비선형적인 관계식을 추정하기 위해 딥 러닝 방법 중 Multi-Layer Perceptron을 이용하였다.



위 그래프에서 색상은 년도를 나타낸다.(노랑색 : 10 (2022년) , 보라색 : 0 (2012년))

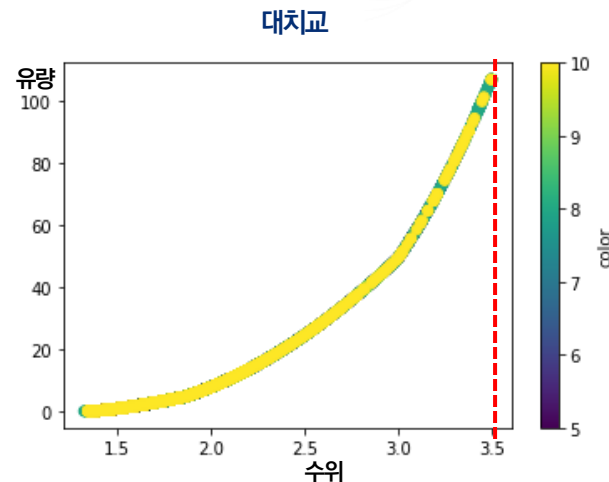
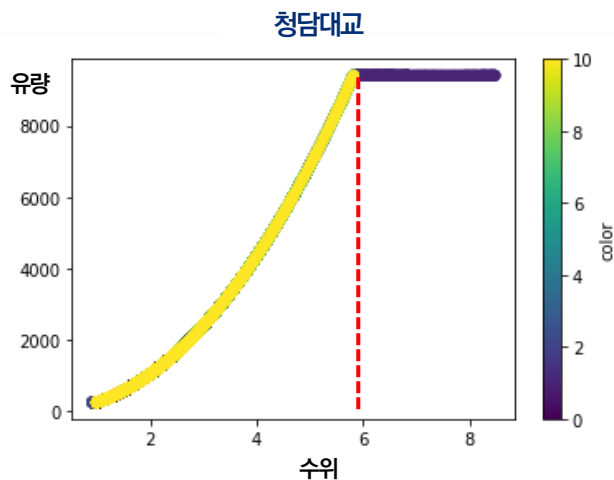
Chang, Ki-Hwan, & 이재형. (2005). Stage-Discharge Rating Curve Model Development and Modification. *Journal of Korea Water Resources Association*, 38(4), 271-280.

<https://doi.org/10.3741/JKWRA.2005.38.4.271>

결측치 대체

- 유량 수위 곡선관계식과 딥 러닝(MLP) 을 이용한 수위, 유량 변수 결측치 대체

test 데이터(2022년)에 결측값들이 존재하는 변수들만 결측대체를 실행하였다. test 데이터의 결측값들은 청담대교, 대치교에만 존재하기 때문에 이를 아래 그래프를 통해 결측을 대체하였다.

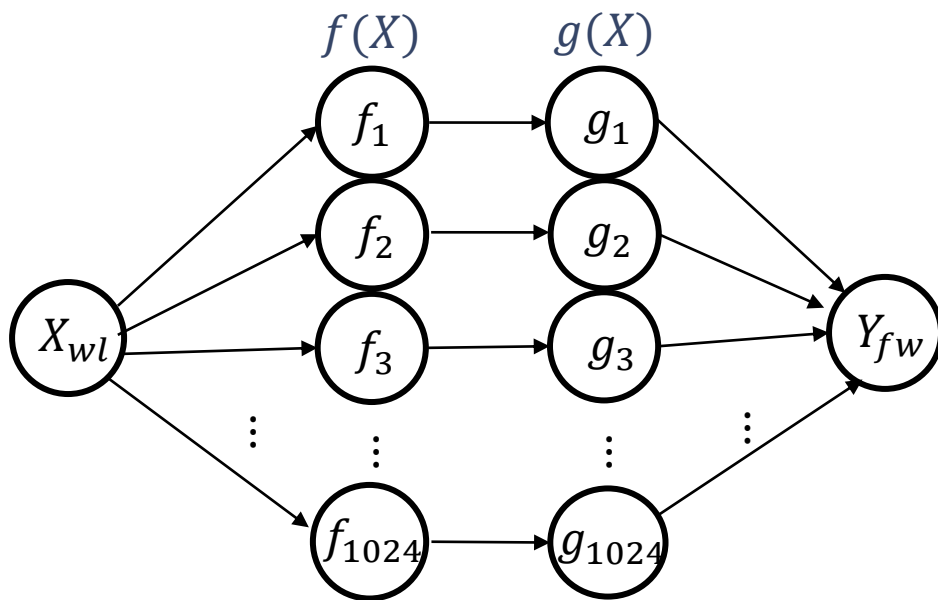


모형 추정에 앞서 청담대교와 대치교의 유량 변수 결측값이 위 그래프의 **빨간 점선** 이후의 값(청담대교 수위 6이상, 대치교 수위 3.5 이상)에서 존재한다면, 각 데이터의 max 값으로 대체 하였다.

결측치 대체

- 유량 수위 곡선관계식과 딥 러닝(MLP)을 이용한 수위, 유량 변수 결측치 대체

청담대교와 대치교의 유량 변수 결측값이 위 그래프의 **빨간 점선** 이전의 값(청담대교 수위 6미만, 대치교 수위 3.5 미만)에서 존재한다면, **MLP**를 이용하여 결측을 대체하였다.



$$Y_{fw} = f(g(X_{wl}))$$

X_{wl} : 유량

Y_{fw} : 수위

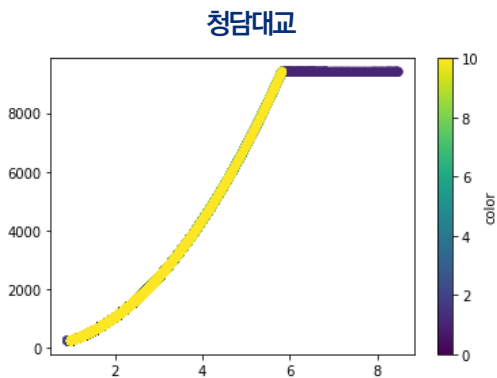
일반적인 linear regression이 아닌 neural network를 이용한 이유는 MLP의 **Universal Approximation Theorem**에 의하여 특정 함수를 근사시킬 수 있기 때문이다.

결측치 대체

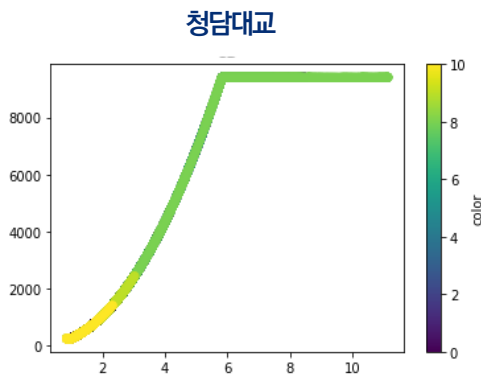
- 유량 수위 곡선관계식과 딥 러닝(MLP) 을 이용한 수위, 유량 변수 결측치 대체

수위, 유량 변수 결측치 대체 결과: train (2012-05-01 ~ 2022-05-31)/ test (2022-06-01~)

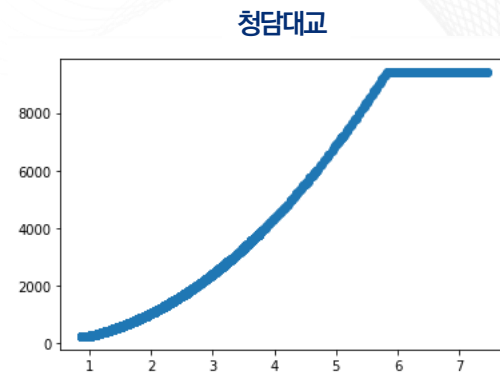
결측치 대체 전



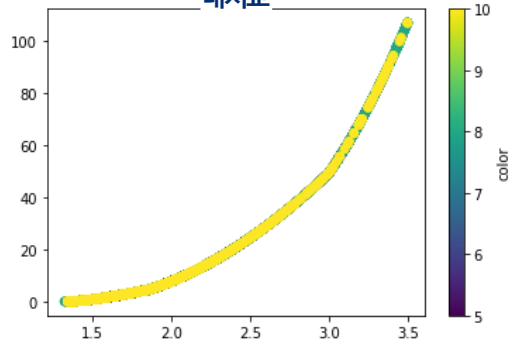
결측치 대체 후(train)



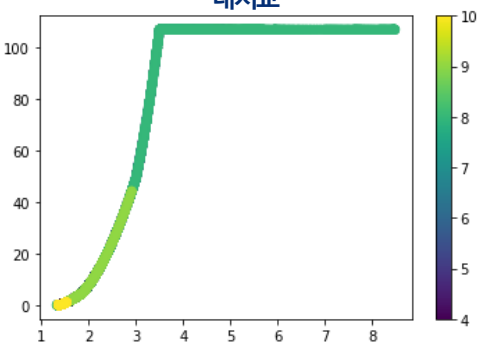
결측치 대체 후(test)



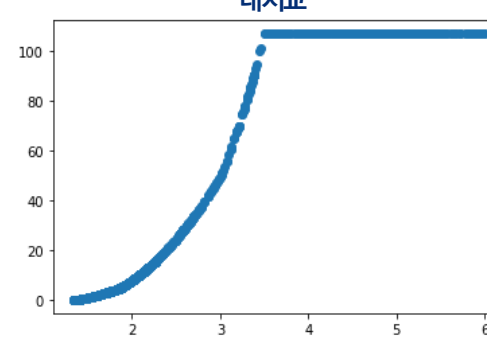
대치교



대치교



대치교



결측치 대체

- Moving Average를 이용한 팔당댐 변수 대체

팔당댐 변수들은 3시점 Moving Average 평균값을 이용하였다.

$$\text{If } x_{t,i} = NaN \text{ Then, } x_{t,i} = (x_{t-1,i} + x_{t-2,i} + x_{t-3,i})/3$$

$$i \in [\text{swl}, \text{inf}, \text{sfw}, \text{ecpc}, \text{tototf}]$$

swl : 팔당댐 현재 수위 (단위: El.m)

inf : 팔당댐 유입량 (단위: m³/s)

ecpc : 팔당댐 공용량 (단위: 백만m³)

sfw : 팔당댐 저수량 (단위: 만m³)

tototf : 총 방류량 (단위: m³/s)

시점	swl	inf	sfw	ecpc	tototf
1	24.85	387	206.01	37.99	387
2	24.85	387	206.01	37.99	387
3	24.85	387	206.01	37.99	387
4	24.86	483	206.37	37.63	384
5	NaN	NaN	NaN	NaN	NaN
6	24.86	384	206.37	37.63	384



시점	swl	inf	sfw	ecpc	tototf
1	24.85	387	206.01	37.99	387
2	24.85	387	206.01	37.99	387
3	24.85	387	206.01	37.99	387
4	24.86	483	206.37	37.63	384
5	24.853	419	206.13	37.87	386
6	24.86	384	206.37	37.63	384

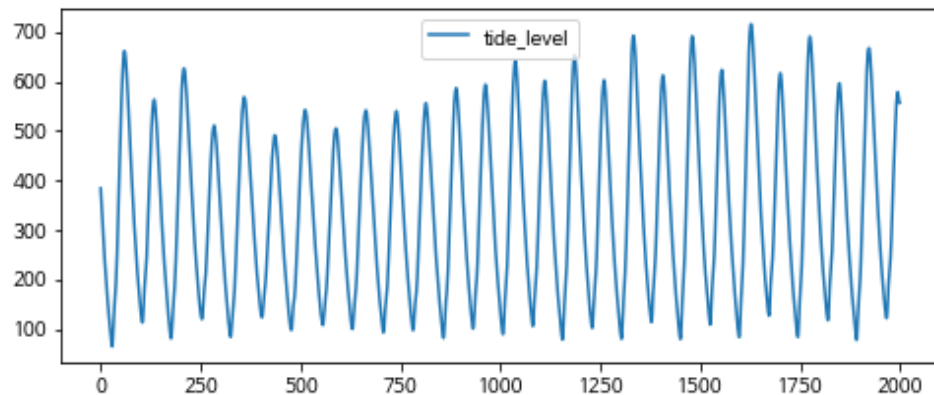
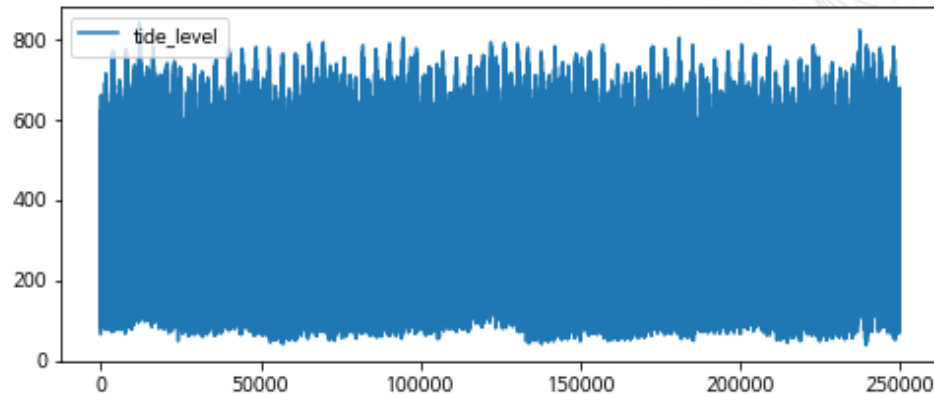
평균



결측치 대체

- Light GBM을 이용한 tide level 변수 대체

강화대교 조위(tide level)는 12시간 간격의 주기성을 가진 데이터이다. 시간 흐름에 따라 주기성이 있는 데이터들은 시계열의 대표적인 모형인 AR, MA 등과 Fourier 변환 또는 wavelet 변환 등과 같은 분해방법 같은 모형들을 이용하여 결측값을 대체할 수 있지만 우리는 Lag 데이터와 Light GBM을 이용하여 결측값을 대체하였다.



결측치 대체

- Light GBM을 이용한 tide level 변수 대체

$$Y_t = f(\tilde{X}_t, \tilde{X}_{t-1}, \tilde{X}_{t-2}, \tilde{X}_{t-3}, Y_{t-1}, Y_{t-2}, Y_{t-3})$$

Y_t : tide level 변수

\tilde{X}_t : 아래 변수들과 그 변수들의 3시점 lag data를 이용

팔당댐 변수

swl : 팔당댐 현재수위 (단위: El.m)

sfw : 팔당댐 저수량 (단위: 만m³)

ecpc : 팔당댐 공용량 (단위: 백만m³)

유량(fw)

팔당대교

광진교

대치교

청담대교

중랑교

한강대교

오금교

행주대교

강수량 변수

rf_10184100 : 대곡교 강수량

rf_10184110 : 진관교 강수량

rf_10184140 : 송정동 강수량

rf_10194010 : 김포시청 강수량

수위(wl)

팔당대교

광진교

대치교

청담대교

중랑교

한강대교

잠수교

행주대교

전류리

오금교

Chapter

IV

파생변수 정의

파생변수 정의

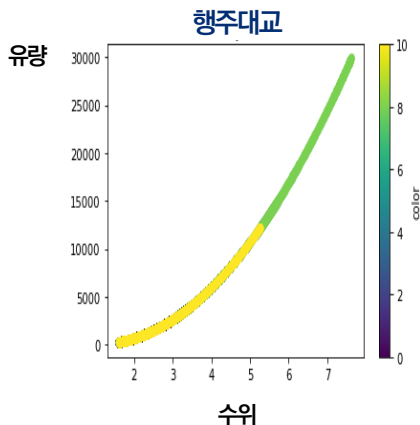
파생변수 정의

파생변수는 기존의 변수를 조합하여 새로운 변수를 만들어 내는 것을 의미한다. 파생변수를 만들어 사용하면 모델 성능 상승의 효과를 기대 할 수 있다. 우리는 다음과 같은 파생변수들을 만들었다.

- Square 변환
- Root 변환
- 차분 변수 생성
- 차분 비율 생성
- 전 시점 Min-Max
- 푸리에 변환(Fourier transform)

파생변수 정의

- Square 변환

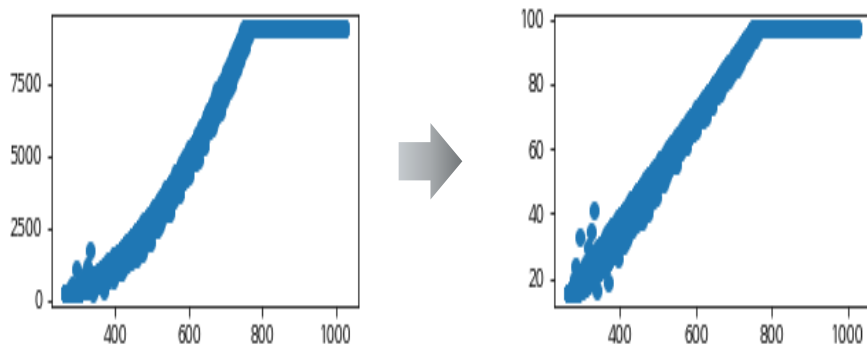


Scatter plot에서 확인해본 결과

현재 시점의 유량과 수위는 2차 다항식 형태로 적합이 가능 하다.

따라서, 전 시점 유량과 현재 시점 수위도 2차 다항식 형태로 적합이 가능하다 생각 되었기 때문에 제곱 변환의 변수를 사용하였다.

- Root 변환



옆 그래프는 현시점 수위와 이전 시점 유량 관계 곡선이다.

제곱근 변환을 통해 좀더 선형적인 관계가 된 것을 확인 할 수 있다.

값이 음수일 때 절대 값의 양의 제곱근으로 대체하였다.

파생변수 정의

- 차분 변수 생성

차분이란 시계열에서 연이은 관측 값의 차이이다. 식으로 나타내면 다음과 같다.

$$y_t^{(1)} = y_t - y_{t-1}$$

우리는 2시점 이전의 차분값도 사용했다.

$$y_t^{(2)} = y_t - y_{t-2}$$

- 차분 비율 생성

차분 된 값에 그 시점의 값으로 나눠준 값을 변수로 사용했다. 식으로 나타내면 다음과 같다.

$$r_t^{(1)} = \frac{y_t - y_{t-1}}{y_{t-1}}$$

마찬가지로 2시점 이전의 비율을 사용했다.

$$r_t^{(2)} = \frac{y_t - y_{t-2}}{y_{t-2}}$$

파생변수 정의

- 전시점 Min, Max 생성

선택된 lag 데이터에서 가장 큰 값과 작은 값을 변수로 만들었다.

- 푸리에 변환(Fourier transform)(주기 함수)

푸리에 변환이란 임의의 입력 신호를 다양한 주파수를 갖는 주기함수들의 합으로 분해하여 표현하는 것이다.
각 주기함수 성분들은 교유의 주파수와 강도를 가지고 있으며 이를 모두 합치면 원본 신호가 된다.

이 아이디어를 시계열 데이터로 접목해서 시계열의 데이터를 임의의 입력 신호로 생각했을 때, 시계열 데이터도
고유의 **주파수**와 **강도**로 나눌 수 있다.

따라서, 이 주파수와 강도를 변수로 사용하였다.

- 월별 데이터 생성

수위 데이터는 계절성이 존재한다고 판단 되었기 때문에, 5~10월은 5~10값으로 지정하여 변수를 만들었다.

Chapter

V

변수선택

변수 제거 및 최적 lag 선택 알고리즘

모델의 성능을 높이기 위해 후진제거법의 아이디어를 차용하여 변수를 제거하였다.

앞에서 언급된 변수를 사용하여 변수선택을 진행하였고, 선택된 변수들의 최적 lag를 선택했다.

변수 선택 알고리즘

1. MSE(mean square error) 임계값 설정
2. 변수를 순서대로 제거한 후 linear regression 을 사용하여 MSE 계산
3. 변수를 제거하고 계산된 MSE와 임계값 MSE를 비교한 후 임계값 보다 작으면 변수 제거
4. 변수의 개수만큼 2~3 반복

변수 제거 및 최적 lag 선택 알고리즘

Inputs: 입력 변수 및 변수 제거 모형(Linear Regression 사용)

Outputs: 제거된 변수명 및 최적 lag 선택

Initialize: ['ymdhm', 'time', 'year', 'wl_1018662', 'fw_1018662', 'wl_1018680', 'wl_1018683', 'fw_1018683', 'wl_1019630', 'fw_1019630', 'fw_out_1018680', 'fw_out_1019675', 'wl_out_1018658', 'fw_out_1018658'] - drop_columns

Algorithms:

best_mse = 10

For i = 0 to # of inputs do

Drop column[i]

For lag = 12, 18, 24 do

Create derived variables ※ 파생변수 차분 비율 변수 lag1, lag2, 푸리에 변환 lag의 1/3개 사용, 제곱 변환

Fit model \hat{f}

Calculate MSE

End

best_model_mse = max(MSE)

best_model_lag = argmin(MSE)

If best_mse > best_model_mse

then best_mse = best_model_mse

drop_columns = drop_columns + column[i]

best_lag = best_model_lag

End

Return drop_columns, best_lag

선택된 변수 리스트

총 41개의 변수 중 15개의 변수가 선택되었다.

'tide_level' : 강화대교 조위

'rf_10184140' : 송정동 강수량

'wl_out_1018610' : 팔당대교 수위

'wl_out_1018640' : 광진교 수위

'fw_out_1018662' : 청담대교 유량

'wl_out_1018662' : 청담대교 수위

'wl_out_1018675' : 중랑교 수위

'wl_out_1018680' : 잠수교 수위

'fw_out_1018683' : 한강대교 유량

'wl_out_1018683' : 한강대교 수위

'wl_out_1018697' : 오금교 수위

'fw_out_1019630' : 행주대교 유량

'wl_out_1019630' : 행주대교 수위

'wl_out_1019675 : 김포시(전류리) 수위

'month' : 월 (5~10)

Chapter

VI

모델링

모델 검증 방법


- K-fold 방법 이용 (k=5 설정)

모델 설정

- 모델: LightGBM(n_estimator=5000), LinearRegression 사용
- 다중 출력 모형 (RegressorChain) 이용
- K-fold 검증 방법을 이용하여 가중평균 결과 생성 (Validation MSE의 역수를 이용)

모델 결과 비교

- K-fold 방법 이용 (k=5 설정)

	Score	Model	
1	Public : 0.9539895956 private : 0.89271	LightGBM(n_estimator=5000), LinearRegression Regression Chain	
2	Public : 0.956403115	1 + LightGBM(n_estimator=1000), LightGBM(n_estimator=500), LinearRegression Regression Chain	



Thank you
