고급데이터분석 PROJECT

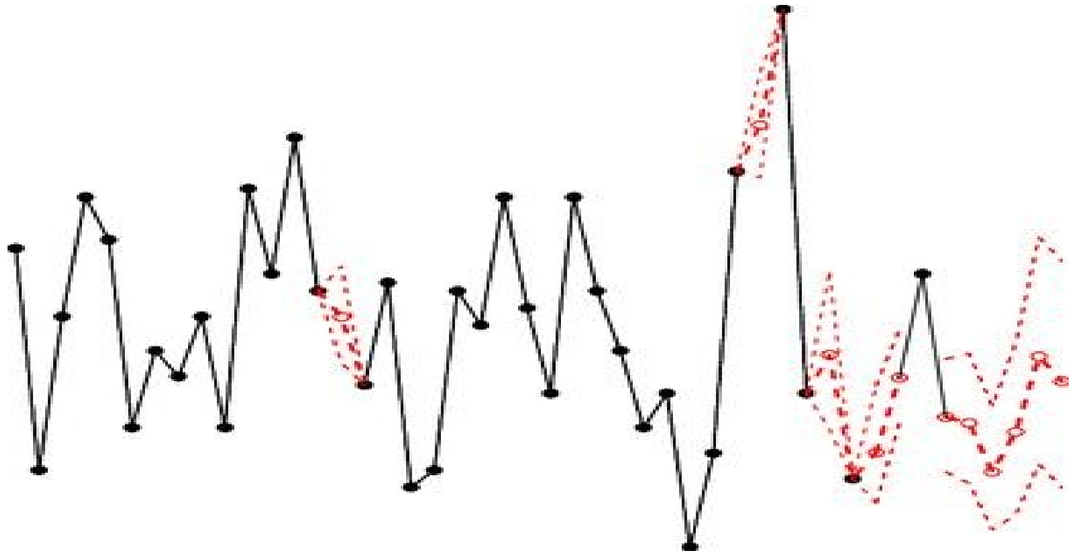
Time-series Imputation

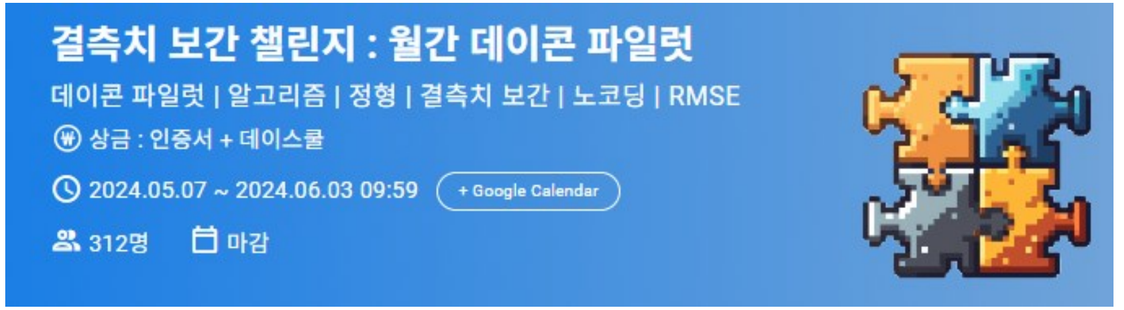
목차

연구 주제 및 선정 배경

Time-series data Imputation

시계열 데이터는 시간의 흐름에 따라 순차적으로 수집된 데이터.
시계열 데이터는 경제, 기후, 금융, 의료 등 다양한 분야에서 쓰임.





배경 요인 1

시간의 제약이 있는 시계열 데이터는 특성상
실제로 수집되는 시계열 데이터는 불완전하거나
결측치가 포함되어 있어 분석이 비교적 어려움.

배경 요인 2

결측치 처리는 데이터의 연속성을 유지하고 분석의 정확성을
높이는 데 필수적인 작업으로 적절한 결측치 처리는
데이터의 품질 및 모델 성능을 향상시키는 중요한 역할.

연구 목표

Time-series data Imputation

목표 1

결측치 처리의 이론적 이해

시계열 데이터 결측치 처리의 기본 개념과
다양한 기법들을 학습하고 이해.

목표 2

기법 비교 및 평가

여러 결측치 처리 방법을 실제 시계열
데이터에 적용해보고,
직접 각 기법의 성능을 비교 및 평가.

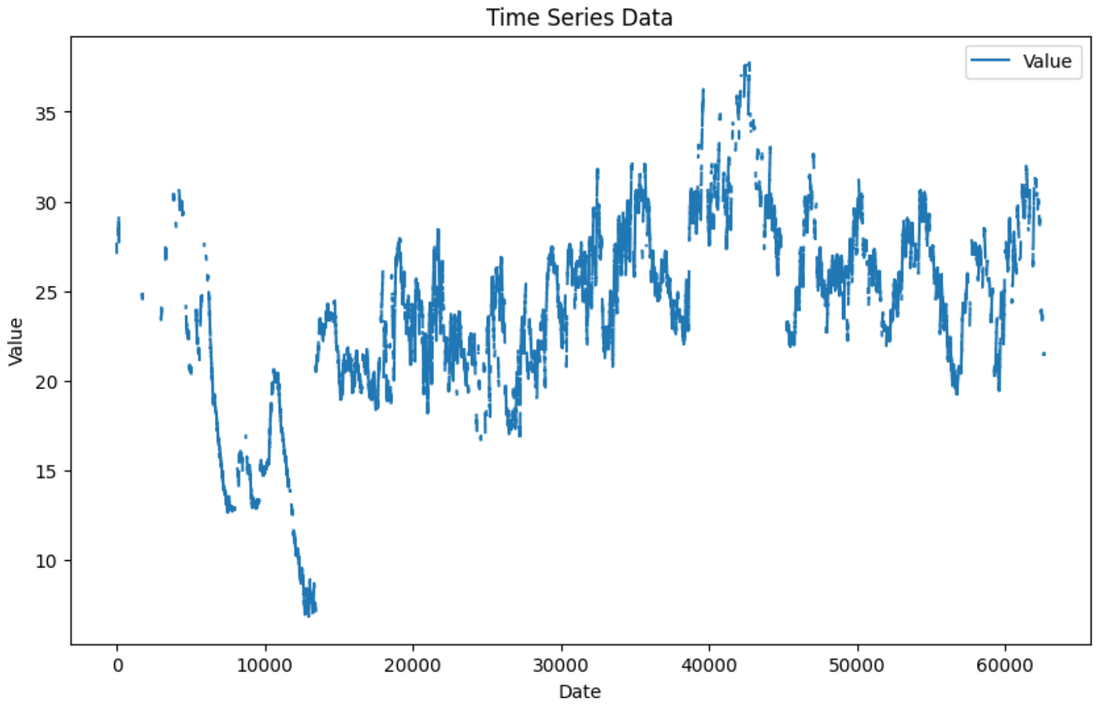
목표 3

결측치 처리 방법 인사이트

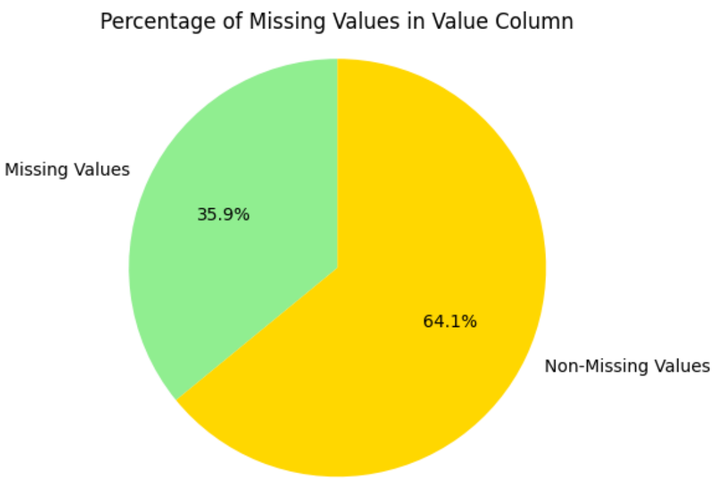
특정 시계열 데이터의 특성에 따라
최적의 결측치 처리 방법을 알아내고,
그에 따른 인사이트 얻고자 함.

단변량 시계열 데이터

단변량 시계열 데이터



id	Value
SAMPLE_00000	27.63677025
SAMPLE_00001	27.25081825
SAMPLE_00002	27.15434837
SAMPLE_00003	
SAMPLE_00004	
SAMPLE_00005	
SAMPLE_00006	
SAMPLE_00007	
SAMPLE_00008	
SAMPLE_00009	



데이터셋 구성

- 일정한 텀 마다 센서에서 온도를 측정한 데이터.
- 이 과정에서 기기/통신 결함으로 인한 결측이 발생.

- 변수
  - id : 샘플 고유 ID (SAMPLE + 숫자)
  - Value : 일정한 텀 마다 센서에서 측정된 온도 값 (°C)
- 단변량 시계열 데이터이므로 결측치를 단순 삭제 불가.
- 총 결측률이 35.9%

이전 결과 요약

	방법론 설명	RMSE
평균값	결측치를 해당 열의 전체 데이터의 평균값으로 대체하는 방법.	6.0482555382
중앙값	측치를 해당 열의 전체 데이터의 중앙값으로 대체하는 방법. 중앙값은 데이터를 크기순으로 정렬했을 때 중간에 위치한 값.	5.9534007587
앞의 값	결측치를 해당 결측치 이전의 값으로 대체.	3.142049155
선형 보간법	인접한 두 데이터 포인트를 직선으로 연결하여 결측값을 추정하는 방법.	**2.7877574619**
다항식 보간법	인접한 여러 데이터 포인트를 이용해 다항식을 구성하여 결측값을 추정하는 방법.	10.5082777801
스플라인 보간법	구간별로 다항식을 이용하여 매끄러운 곡선을 만들어 결측값을 추정하는 방법 각 구간에서 다항식을 이용하여 결측치를 추정하며, 각 구간의 경계에서 연결성을 보장.	3.0326187835
SVR	서포트 벡터 머신을 이용한 회귀 분석 기법으로, 데이터의 패턴을 학습하여 결측치를 추정. 수식적으로는 마진 안에 있는 데이터를 이용해 회귀식을 만듦.	3.0328705384
EWMA	지수 가중 이동 평균을 이용하여 결측치를 추정하는 방법. 최근 데이터에 더 큰 가중치를 부여.	2.7877654275

ARIMA Method

AR 모델 (AutoRegressive Model)

• 자기상관성을 시계열 모형으로 구성한 것으로, 변수의 과거 관측값의 선형 결합을 통해 변수의 미래값을 예측하는 모델.
• 이전의 관측값이 이후의 관측값에 영향을 주는 것을 반영.

$$y_t = c + \phi_1 y_{t-1} + \phi_2 y_{t-2} + \cdots + \phi_p y_{t-p} + \epsilon_t$$

y_t는 t시점의 관측값, c는 상수, ϕ는 가중치, ϵ_t는 오차항을 의미합니다.

- 오차항은 평균이 0, 분산이 1인 표준정규분포를 따르는 백색잡음입니다.

MA 모델

- 과거 백색 잡음을 반영하여 미래값을 예측하는 모델

$$y_t = c + \theta_1 \epsilon_{t-1} + \theta_2 \epsilon_{t-2} + \cdots + \theta_q \epsilon_{t-q} + \epsilon_t$$

ARMA 모델

p개의 자기 자신의 과거값과 q개의 과거백색 잡음의 선형조합을 의미

$$y_t = AR(p) + MA(q)$$

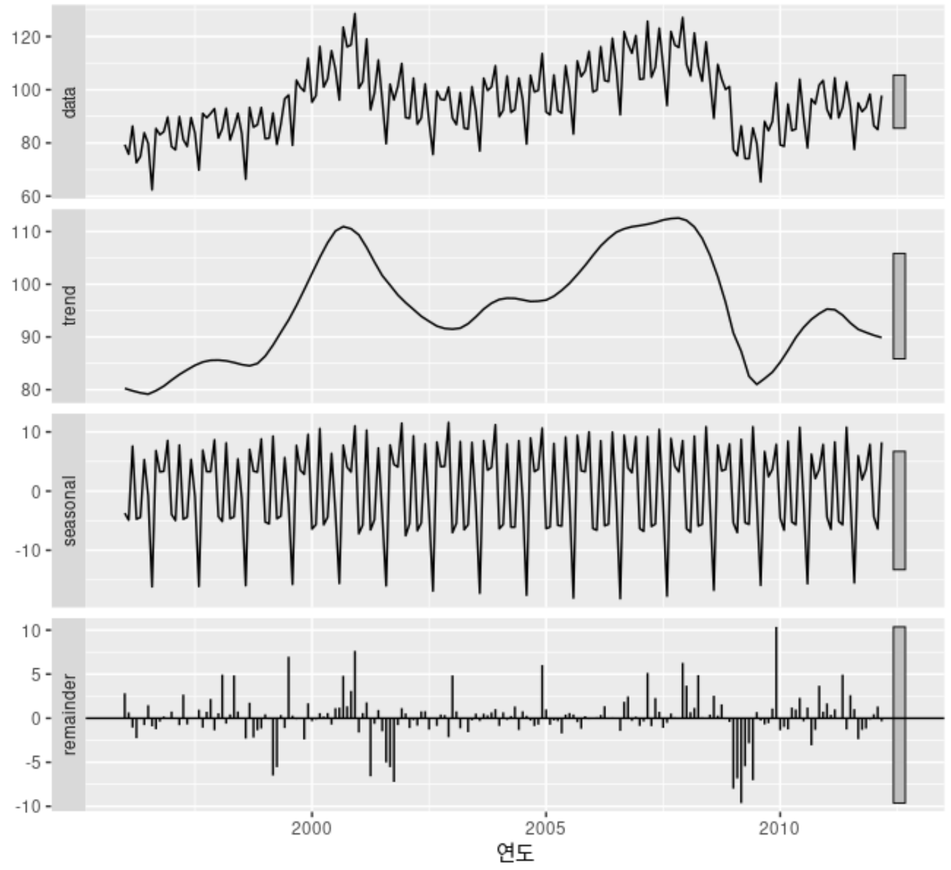
$$y_t = c + \phi_1 y_{t-1} + \phi_2 y_{t-2} + \cdots + \phi_p y_{t-p} + \epsilon_t + \theta_1 \epsilon_{t-1} + \theta_2 \epsilon_{t-2} + \cdots + \theta_q \epsilon_{t-q}$$

ARIMA 모델

- ARMA 모델에서 d차 차분을 추가적으로 적용시킨 모델로
- 비정상시계열에도 바로 적용 가능
- ARIMA(p,d,q) : d차 차분한 데이터에 AR(p)모형과 MA(q)모형을 합친 모델

$$y'_t = y_t - y_{t-m},$$

Holt-winters Method



Holt-Winters 기법

- 시계열 데이터에서 추세와 계절성을 모두 고려하여 예측하는 방법
  - 추세(Trend): 시계열 데이터가 장기적인 변화
  - 계절성(Seasonality): 데이터가 일정한 주기로 반복되는 패턴

•Level (수준): 시계열의 기본 수준
•Trend (추세): 시계열의 추세를 나타내는 값
•Seasonality (계절성): 시계열의 계절적 패턴을 나타내는 값
•이 기법은 각 요소의 가중치를 조정하여 시계열 데이터의 다양한 패턴을 반영

$$\hat{y}_{t+h|t} = \ell_t + hb_t + s_{t+h-m(k+1)}$$
$$\ell_t = \alpha(y_t - s_{t-m}) + (1-\alpha)(\ell_{t-1} + b_{t-1})$$
$$b_t = \beta^*(\ell_t - \ell_{t-1}) + (1-\beta^*)b_{t-1}$$
$$s_t = \gamma(y_t - \ell_{t-1} - b_{t-1}) + (1-\gamma)s_{t-m},$$

총 결과 정리

	방법론 설명	RMSE
평균값	결측치를 해당 열의 전체 데이터의 평균값으로 대체하는 방법.	6.0482555382
중앙값	결측치를 해당 열의 전체 데이터의 중앙값으로 대체하는 방법.	5.9534007587
앞의 값	결측치를 해당 결측치 이전의 최근 값으로 대체.	3.142049155
선형 보간법	인접한 두 데이터 포인트를 직선으로 연결하여 결측값을 추정하는 방법.	**2.7877574619**
다항식 보간법	인접한 여러 데이터 포인트를 이용해 다항식을 구성하여 결측값을 추정하는 방법.	10.5082777801
스플라인 보간법	구간별로 다항식을 이용하여 매끄러운 곡선을 만들어 결측값을 추정하는 방법	3.0326187835
SVR	서포트 벡터 머신을 이용한 회귀 분석 기법으로, 데이터의 패턴을 학습하여 결측치를 추정.	3.0328705384
EWMA	지수 가중 이동 평균을 이용하여 결측치를 추정하는 방법.	2.7877654275
Linear+Noise	선형 보간법에 랜덤하게 Noise 추가.	2.9549151818
ARIMA	시계열 데이터의 특성을 설명하는 자기 회귀(AR, Autoregressive), 적분(I, Integrated), 이동 평균(MA, Moving Average)을 고려하는 방법	2.7887965992
Holt-winters	시계열 데이터에서 추세와 계절성을 모두 고려하는 방법	**2.7874594914**

대회 최종 결과

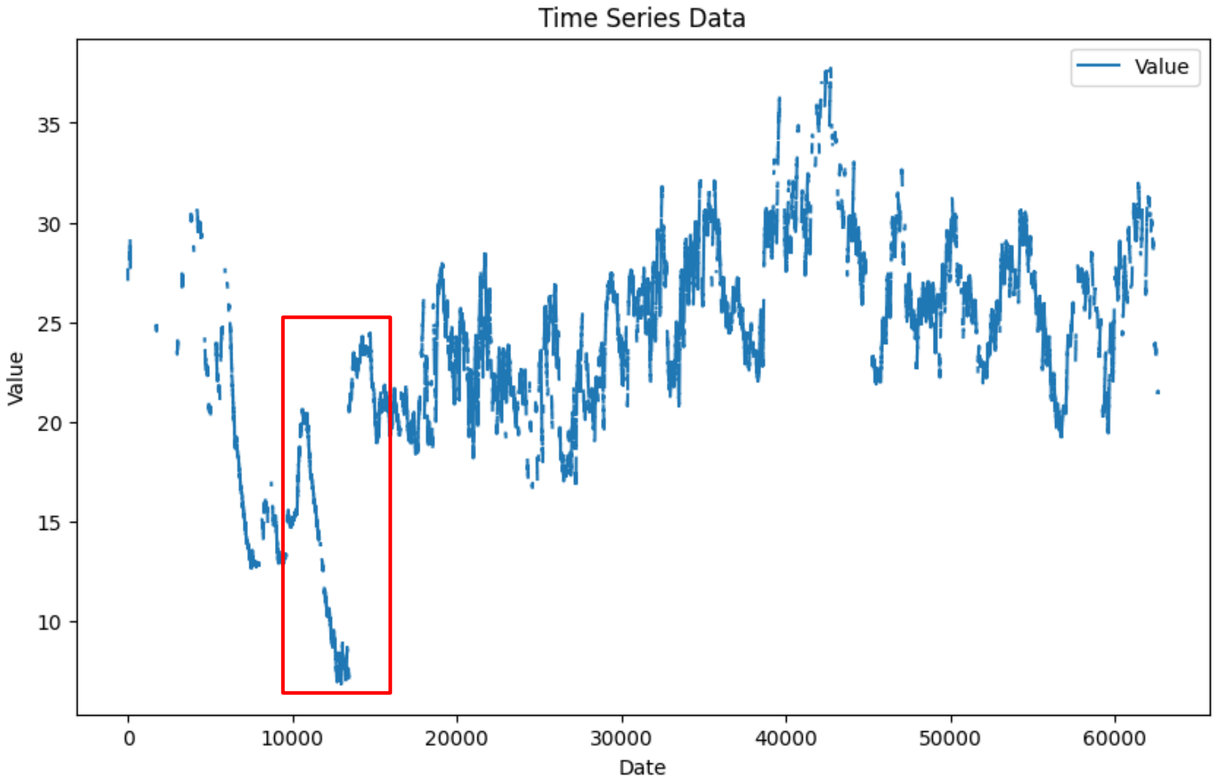
● WINNER ● 1% ● 4% ● 10%

전체 랭킹 >

#	팀	팀 멤버	점수	제출수	등록일
22			2.7874	7	17일 전
1			2.74103	9	한 달 전
2			2.74404	17	한 달 전
3			2.74451	18	한 달 전

- 총 312명 중에 22등으로 10%안의 성적을 기록

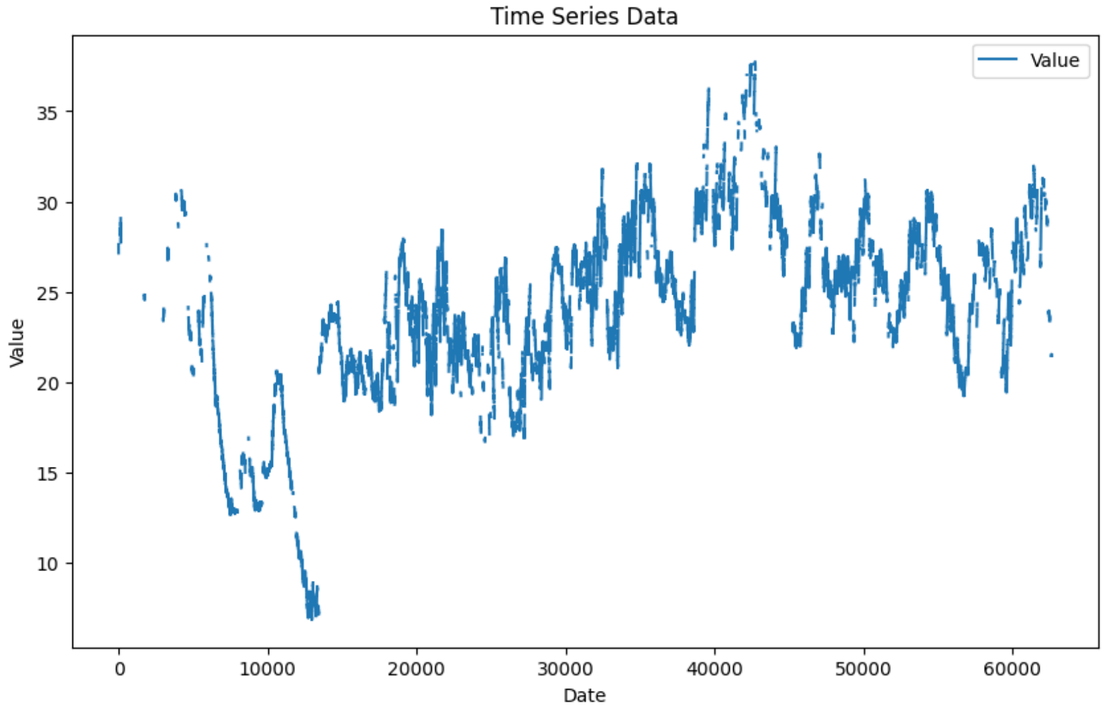
최적의 보간법 분석 결과

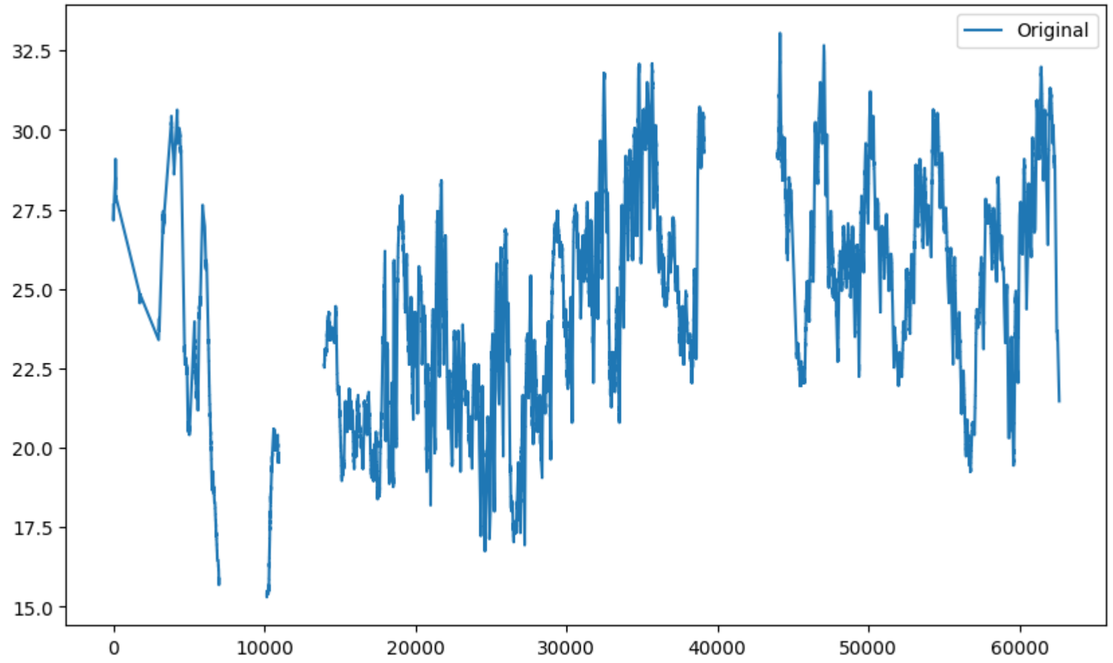


비교적 선형보간법이 높은 성능의 변화를 이끌어냄.

- 데이터의 결측구간이 굉장히 단순함.
- 일자로 연결되는 형태가 많음.
- 그러므로 선형보간법이 좋은 성능을 보임.

데이터 수정 및 진행



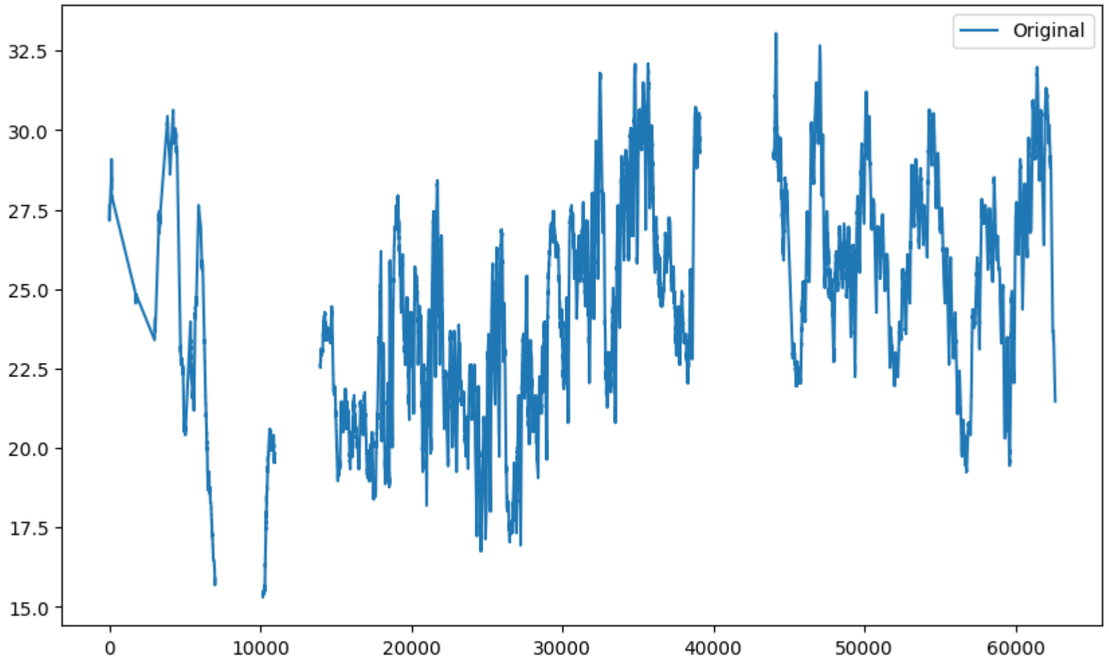


- 단순한 일자 구간에는 선형 보간법으로 미리 채운 후
- 극대 극소 지점 몇 개를 선정하여 Nan값으로 처리함.

Masking data 결과 정리

	방법론 설명	RMSE
평균값	결측치를 해당 열의 전체 데이터의 평균값으로 대체하는 방법.	7.978891485426411
중앙값	결측치를 해당 열의 전체 데이터의 중앙값으로 대체하는 방법.	8.040489111428966
앞의 값	결측치를 해당 결측치 이전의 최근 값으로 대체.	0.5046316634075125
선형 보간법	인접한 두 데이터 포인트를 직선으로 연결하여 결측값을 추정하는 방법.	0.48567587520823075
다항식 보간법	인접한 여러 데이터 포인트를 이용해 다항식을 구성하여 결측값을 추정하는 방법.	170.20093901288683
스플라인 보간법	구간별로 다항식을 이용하여 매끄러운 곡선을 만들어 결측값을 추정하는 방법	0.735781024694459
SVR	서포트 벡터 머신을 이용한 회귀 분석 기법으로, 데이터의 패턴을 학습하여 결측치를 추정.	7.519170356595184
ARIMA	시계열 데이터의 특성을 설명하는 자기 회귀(AR, Autoregressive), 적분(I, Integrated), 이동 평균(MA, Moving Average)을 고려하는 방법	0.5021135531917638
Holt-winters	시계열 데이터에서 추세와 계절성을 모두 고려하는 방법	0.488226295210856

Masking data 분석 결과



- 선형 보간법이 성능이 좋게 나온 이유는 다음과 같음.

- 다른 보간법(평균값, 중앙값 등등)에 비해 선형 보간법이 실제로 성능이 좋기 때문.
- 봉우리를 제거하면서 일자 구간도 어느 정도 Nan값 처리되어 선형 보간으로 채웠을 때 오차가 적게 발생.

다변량 시계열 데이터

다변량 시계열 데이터

- 지금까지 단변량의 시계열 데이터를 보간함.
- -> 보통 데이터는 다변량으로 이루어짐.

- 다변량 시계열 데이터의 경우 앞서 단변량 시계열 데이터 달리 다른 방법이 존재함.

- 1. 단변량 시계열 데이터의 경우와 같이 각 열마다 따로 적용할 수 있음.
- ->각 열의 정보를 반영할 수 있으나 각 행의 정보를 반영할 수 없음.

- 2. 딥러닝 기법의 다변량 시계열 데이터 보간 모델을 사용
- -> 변수들의 관계도 파악하여 결측치 처리 가능

다변량 시계열 데이터

	Building Number	Date	Temperature (C)	Precipitation (mm)	Wind Speed (m/s)	Humidity (%)	Sunshine (hr)	Solar Radiation (MJ/m2)	Power Consumption (kWh)
0	1	2022-06-01 00:00:00	18.6	NaN	0.9	42.0	NaN	NaN	1085.28
1	1	2022-06-01 01:00:00	18.0	NaN	1.1	45.0	NaN	NaN	1047.36
2	1	2022-06-01 02:00:00	17.7	NaN	1.5	45.0	NaN	NaN	974.88
3	1	2022-06-01 03:00:00	16.7	NaN	1.4	48.0	NaN	NaN	953.76
4	1	2022-06-01 04:00:00	18.4	NaN	2.8	43.0	NaN	NaN	986.40
...	...	...	...	...	...	...	...	...	...
203995	100	2022-08-24 19:00:00	23.1	NaN	0.9	86.0	0.5	NaN	881.04
203996	100	2022-08-24 20:00:00	22.4	NaN	1.3	86.0	0.0	NaN	798.96
203997	100	2022-08-24 21:00:00	21.3	NaN	1.0	92.0	NaN	NaN	825.12
203998	100	2022-08-24 22:00:00	21.0	NaN	0.3	94.0	NaN	NaN	640.08
203999	100	2022-08-24 23:00:00	20.7	NaN	0.1	95.0	NaN	NaN	540.24

204000 rows × 9 columns

- 100개 건물들의 2022년 06월 01일부터 2022년 08월 24일까지의 데이터
- 일시별 기온, 강수량, 풍속, 습도, 일조, 일사 정보 포함
- 전력사용량(kWh) 포함

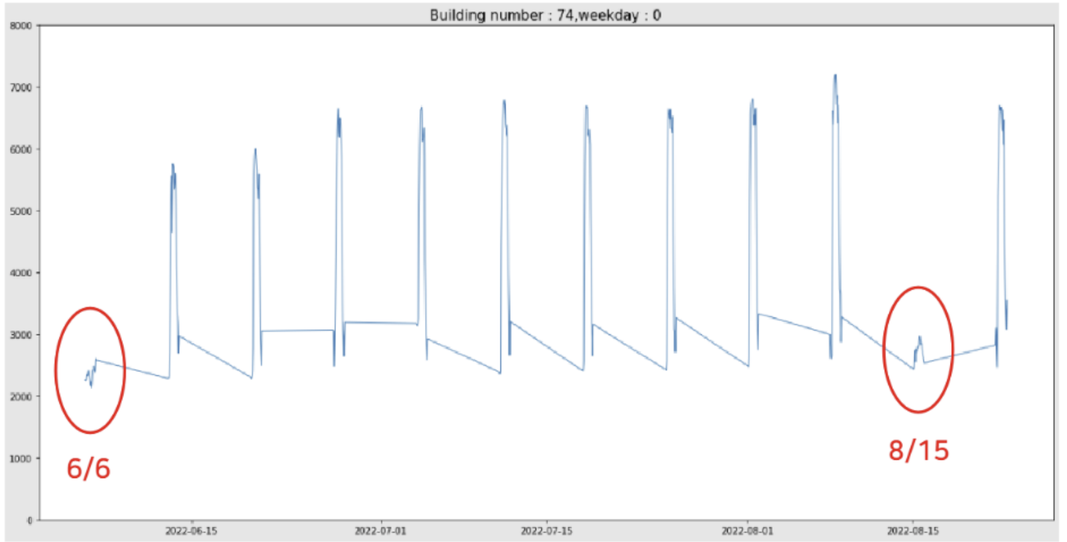
한국에너지공단
KOREA ENERGY AGENCY

데이터 결측값

	실제 결측값	MASK처리 후 결측값
Building Number	0	0
Date	0	0
Temperature (C)	0	20447
Precipitation (mm)	160069	160069
Wind Speed (m/s)	19	20232
Humidity (%)	9	20497
Sunshine (hr)	75182	75182
Solar Radiation (MJ/m2)	87913	87913
Power Consumption (kWh)	0	20449

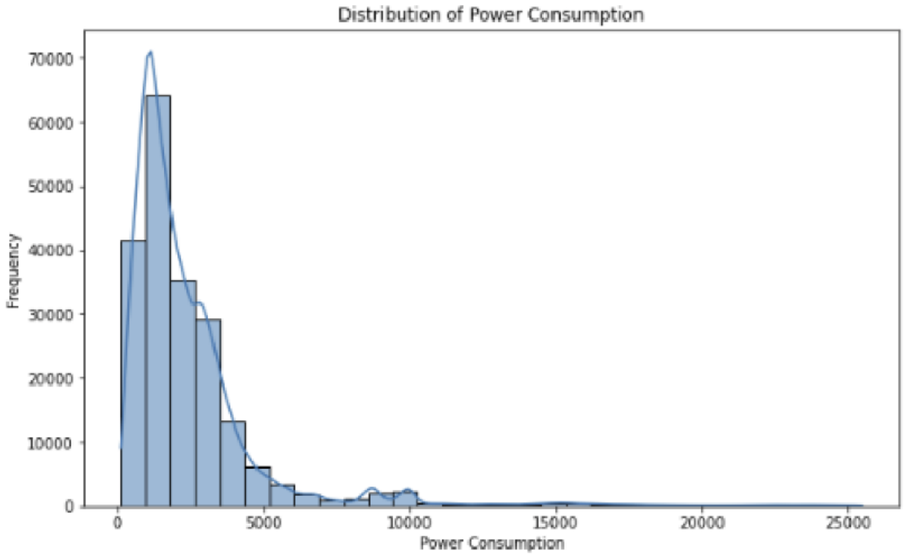
- 결측값이 이미 많이 존재하는 열 제외하고 임의로 결측값이 아닌 값들의 10%를 Nan값으로 대체
- Nan값으로 대체하기 전의 데이터를 라벨로 쓰고 보간 모델을 돌린 뒤 나온 값들과의 오차를 구함.
- 이 방식은 후에 참고한 논문에서 행한 방법과 동일하게 진행함

데이터 분석- 공휴일

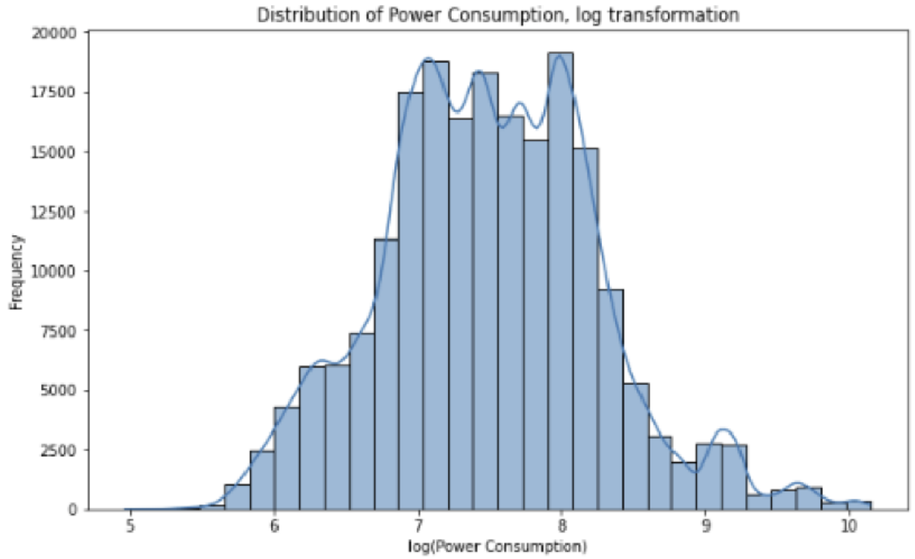


- **6/1(**지방선거**), 6/6(**현충일**), 8/15(**광복절**)**의 경우 평일이지만 휴일과 같이 전력소비량이 낮음
- 해당 공휴일과 주말을 **holiday** 변수로 통합하여 날짜와 전력소비량의 관계를 학습에 반영

데이터 분석 - 전력소비량



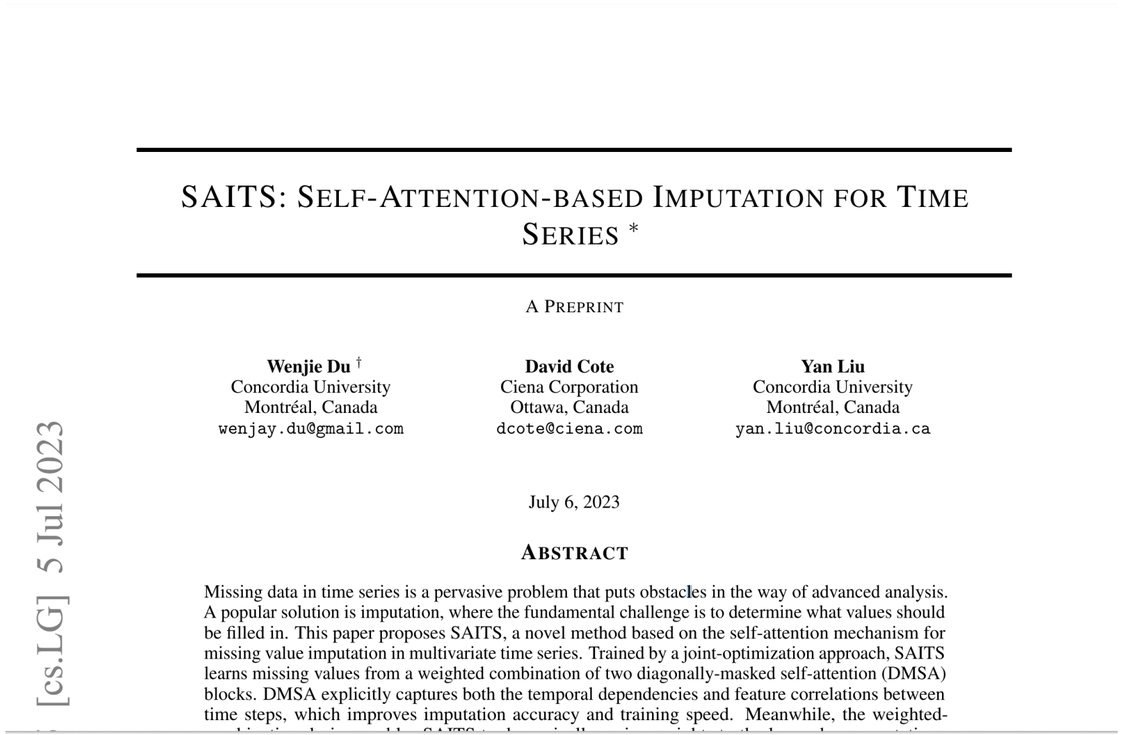
skewness : 3.7430



skewness : 0.2432

- 전력 소비량의 데이터분포가 **skew**값이 **3.7**정도로 높은 **positive skew**형태를 띔
- 값의 범위를 줄이고 정규분포에 가깝기 하기 위해 **log transformation** 진행

Saits 모델



SAITS: SELF-ATTENTION-BASED IMPUTATION FOR TIME SERIES *

A PREPRINT

Wenjie Du †
Concordia University
Montréal, Canada
wenjay.du@gmail.com

David Cote
Ciena Corporation
Ottawa, Canada
dcote@ciena.com

Yan Liu
Concordia University
Montréal, Canada
yan.liu@concordia.ca

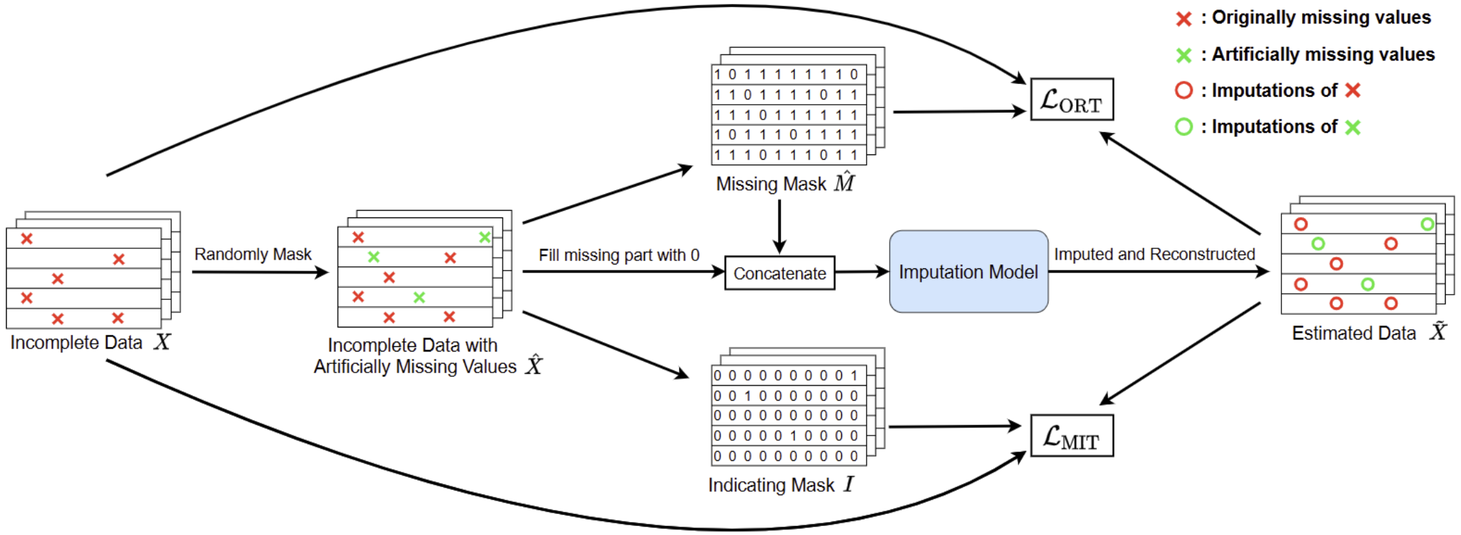
July 6, 2023

ABSTRACT

Missing data in time series is a pervasive problem that puts obstacles in the way of advanced analysis. A popular solution is imputation, where the fundamental challenge is to determine what values should be filled in. This paper proposes SAITS, a novel method based on the self-attention mechanism for missing value imputation in multivariate time series. Trained by a joint-optimization approach, SAITS learns missing values from a weighted combination of two diagonally-masked self-attention (DMSA) blocks. DMSA explicitly captures both the temporal dependencies and feature correlations between time steps, which improves imputation accuracy and training speed. Meanwhile, the weighted-

- SAITS는 셀프 어텐션 메커니즘에 기반하여 결측값을 대체하는 방법으로, 공동 최적화 접근법을 통해 훈련.
- SAITS는 두 개의 대각선-마스킹 셀프 어텐션(DMSA) 블록의 가중 조합에서 결측값을 학습.
- DMSA는 시간 단계 간의 시간적 의존성과 특징 상관성을 명시적으로 포착하여 대체 정확도와 훈련 속도를 향상.
- 동시에 가중 조합 설계를 통해 SAITS는 어텐션 맵과 결측 정보에 따라 두 DMSA 블록에서 학습된 표현에 가중치를 동적으로 할당 가능.

Saits 모델 학습 과정



$$\ell_{\text{MAE}}\left(estimation, target, mask\right) = \frac{\sum_{d=1}^{D}\sum_{t=1}^{T}|(estimation - target) \odot mask|_t^d}{\sum_{d=1}^{D}\sum_{t=1}^{T} mask_t^d}$$

$$\mathcal{L}_{\text{MIT}} = \ell_{\text{MAE}}\left(\tilde{X}, X, I\right)$$

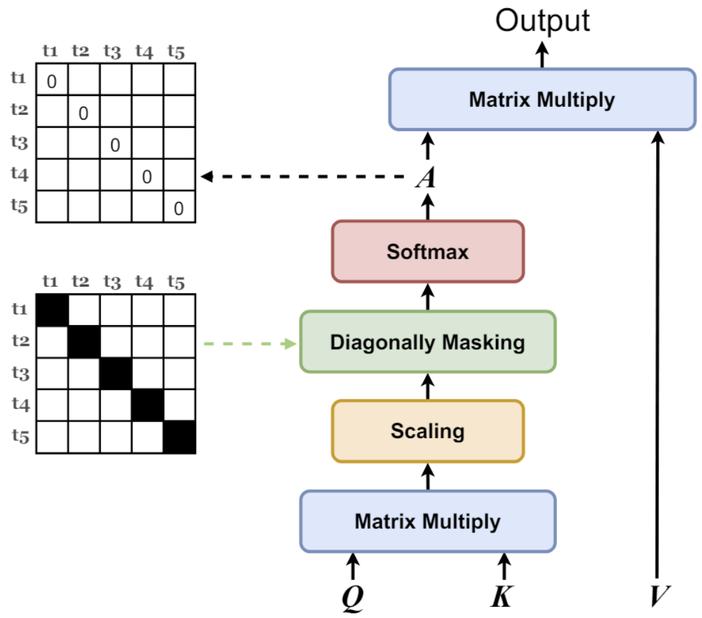
$$\mathcal{L}_{\text{ORT}} = \ell_{\text{MAE}}\left(\tilde{X}, X, \hat{M}\right)$$

- 결측값이 있는 다변량 시계열에서 셀프 어텐션 기반 대체 모델을 잘 훈련시키기 위해, 대체와 재구성의 공동 최적화 훈련 접근법이 설계됨.
- 여기서 "대체"는 모델이 주어진 샘플의 결측 부분을 채우는 과정을 의미하며, "재구성"은 모델이 처리 후에 관측된 값을 가능한 정확하게 복원하는 것을 의미함.
- 이 공동 최적화 접근법은 두 가지 학습 작업으로 구성. : 마스크된 대체 작업 (MIT)와 관측된 재구성 작업 (ORT).
- 이에 따라 훈련 손실은 MIT의 대체 손실과 ORT의 재구성 손실에서 누적됨.

Saits 모델 구조

$$[\text{DiagMask}\,(x)]\,(i,j) = \begin{cases} -\infty & i = j \\ x\,(i,j) & i \neq j \end{cases}$$

$$\text{DiagMaskedSelfAttention}\,(Q,K,V) = \text{Softmax}\left(\text{DiagMask}\left(\frac{QK^{\mathsf{T}}}{\sqrt{d_k}}\right)\right)V$$
$$= AV, \quad \text{where } A \text{ is attention weights}$$



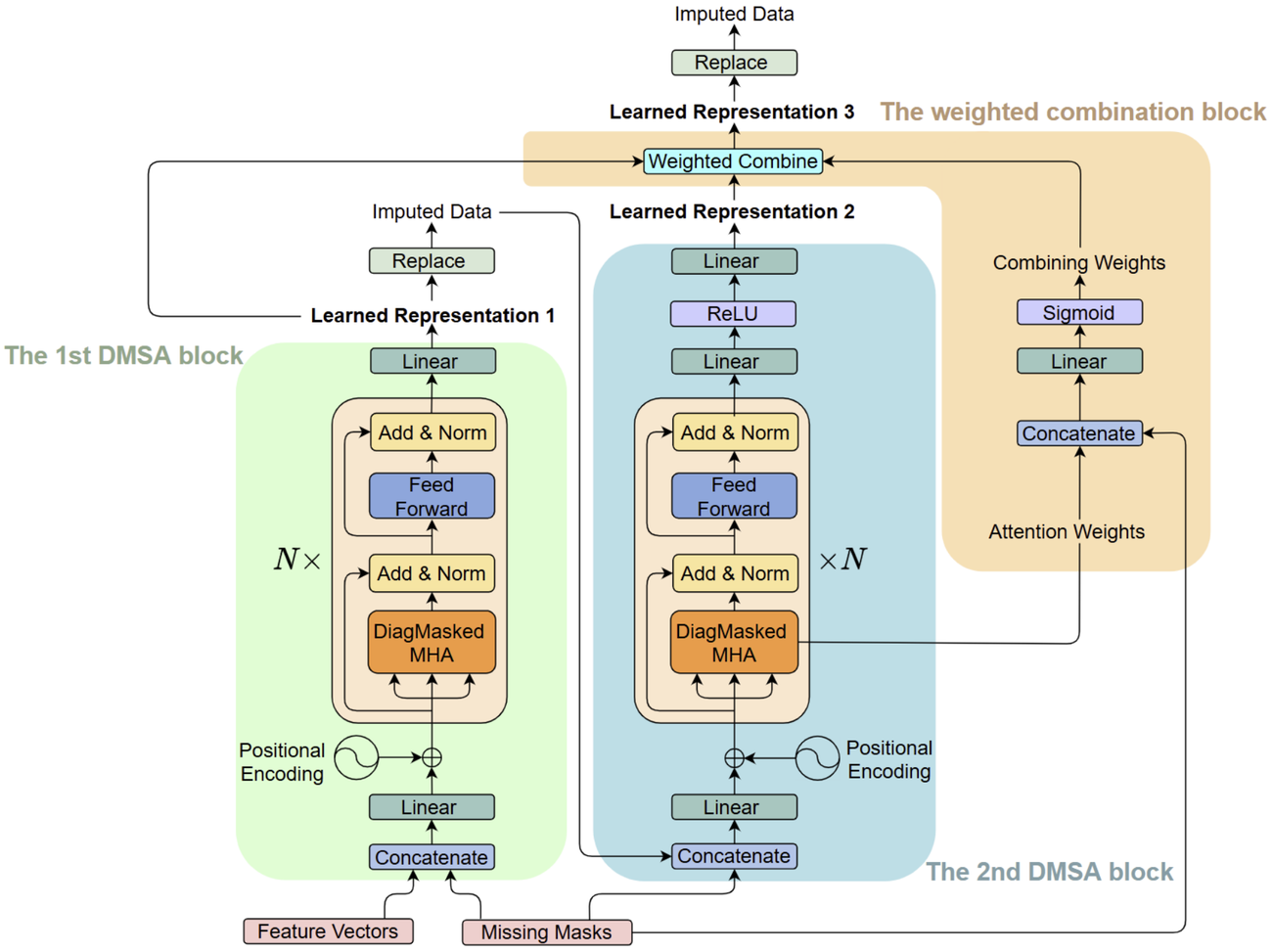


Figure 3: The SAITS model architecture.

평가 지표

3가지의 평가 지표를 사용
1. MAE (Mean Absolute Error)
2. RMSE (Root Mean Square Error)
3. MRE (Mean Relative Error)

$$\text{MAE}\left(estimation, target, mask\right) = \frac{\sum_{d=1}^{D}\sum_{t=1}^{T}\left|\left(estimation - target\right) \odot mask\right|_t^d}{\sum_{d=1}^{D}\sum_{t=1}^{T} mask_t^d}$$

$$\text{RMSE}\left(estimation, target, mask\right) = \sqrt{\frac{\sum_{d=1}^{D}\sum_{t=1}^{T}\left(\left(\left(estimation - target\right) \odot mask\right)^2\right)_t^d}{\sum_{d=1}^{D}\sum_{t=1}^{T} mask_t^d}}$$

$$\text{MRE}\left(estimation, target, mask\right) = \frac{\sum_{d=1}^{D}\sum_{t=1}^{T}\left|\left(estimation - target\right) \odot mask\right|_t^d}{\sum_{d=1}^{D}\sum_{t=1}^{T}\left|target \odot mask\right|_t^d}$$

다변량 시계열 데이터 결과

MAE/MRE/RMSE.

	Saits			Transfomer			BRITS		
	MAE	MRE	RMSE	MAE	MRE	RMSE	MAE	MRE	RMSE
Power Consumption Log + Hoilday	0.52733	0.20955	1.07222	0.55852	0.23265	1.07762	0.78466	0.37012	1.76028
Power Consumption Log	0.53856	0.25665	1.04153	0.55295	0.21054	1.10991	0.82686	0.55712	1.83849
기존 데이터	54.5258	0.65336	256.12882	81.28097	0.62163	627.4905	244.9088	4.71916	1167.7156

결론

01. 단변량 시계열 데이터

기본적으로 잘 알려진 보간법들을 직접 적용하며 성능을 확인할 수 있었음.
➔ 데이터의 분포와 특성에 따라 최적의 보간법이 달라짐.

통계적으로 시계열 데이터를 분해하고 그 요소를 활용하는 모델을 적용할 수 있었음.
➔ 데이터가 쉽게 디자인된 탓에 그 성능을 제대로 확인할 수 없었음.

02. 다변량 시계열 데이터

각 열의 정보 뿐만 아니라 변수들의 관계도 포함하여 결측값을 보간하는 모델을 사용할 수 있었음
➔ 효과적으로 다변량 시계열 데이터를 보완가능함.

이외에도 다양한 보간 모델이 존재하여 후에 데이터 특성에 따라 맞는 다른 보간 모델을 적용해보고자 함.

감사합니다.