KUBIG
DATA SCIENCE & AI

〈분류예측 2팀〉

유통데이터 수요량 예측 & 상수도 관망 이상치 탐지

Team A.I.M.
19기 최지우, 20기 권민석, 20기 장건호

KUBIG
DATA SCIENCE & AI

CONTENTS

01
유통데이터
수요량 예측

대회 소개
전처리, EDA

02
모델링 및
분석 결과

분석 방법 및 절차
분석 결과, 피드백

03
상수도 관망
이상 탐지

대회 소개
전처리, EDA

04
모델링 및
분석 결과

분석 방법 및 절차
분석 결과, 피드백

KUBIG
DATA SCIENCE & AI

1. 유통데이터 수요량 예측

01. 유통데이터 활용 경진대회

주요일정

- **신청**
  2024.9.19.(목)~**2024.10.15.**(화) 18:00
  홈페이지 신청 및 접수
- **분석자료 제출 마감** **2024.10.18.**(금) 18:00
- **예선**(서류평가)
  2024.10.21.(월)~**2024.10.25.**(금)
  본선 진출 10개팀 선정 (수요예측 부문 5팀, 생성형 AI 활용 부문 5팀)
- **본선**(발표평가) **및 시상식**
  2024.11.6.(수), 신촌 에피소드 369
  최종 10개팀 선정 (수요예측 부문 5팀, 생성형 AI 활용 부문 5팀)

유통데이터 활용 경진대회 [수요예측 부문]에 참가

21년 1월부터 23년 12월까지의 상품별 판매수량 데이터를 바탕으로,

24년 1-6월의 월별 판매수량을 예측

01. 유통데이터 활용 경진대회

구분	항목	설명
1	판매일	상품 판매 일자
2	구분	매출, 반품
3	우편번호	소매점 위치정보
4	매출처코드	상품을 구매한 매출처 번호
5	판매수량	상품이 판매된 수량
6	옵션코드	EA: 최소 단위, CS : 묶음 단위 BX : 박스 단위
7	규격	상품 입고 시 박스에 담겨져있는 수량
8	입수	해당 옵션코드에 상품이 들어있는 EA 수량
9	상품 바코드	상품에 부여되는 코드번호
10	상품명	상품이름
11	대분류	상품 대분류
12	중분류	상품 중분류
13	소분류	상품 소분류

<수요예측 부문> 데이터 상세

- '상품정보' 데이터와 '중소유통물류센터 거래' 데이터
- 데이터상품정보: GTIN, 상품분류코드, 모델명, 중량 등
- 거래 데이터: 판매일, 판매수량, 규격, 상품명 등 정보
- 도매 물류센터에서 소매업자와 이루어진 거래 내역으로, 일반 소비자에게 판매하지 않으며 원가에 상품 판매

1 데이터의 경우 (중분류) 라면, 통조림, 상온즉석

2 데이터의 경우 (대분류) 면류, 라면류에 해당하는 품목의 월별 수요예측 모델을 구축해야 함.

02. 전처리, EDA

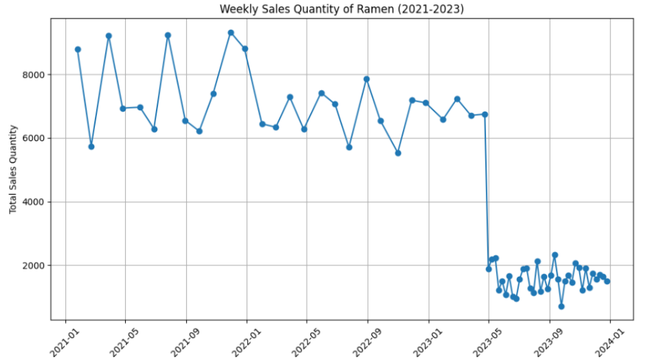
1. 상품정보 데이터 사용(시도)

```
1 na_count = sales1['GTIN'].isna().sum()
2 na_rate = na_count / len(sales1) * 100
3
4 na_rate
```

```
49.78708801167753
```

제공된 상품 정보 데이터를 공용칼럼으로
병합 후 GTIN을 Key로 병합 시도
결측치 비율이 50%, 사용 x

2. 데이터별 수요 특성 확인

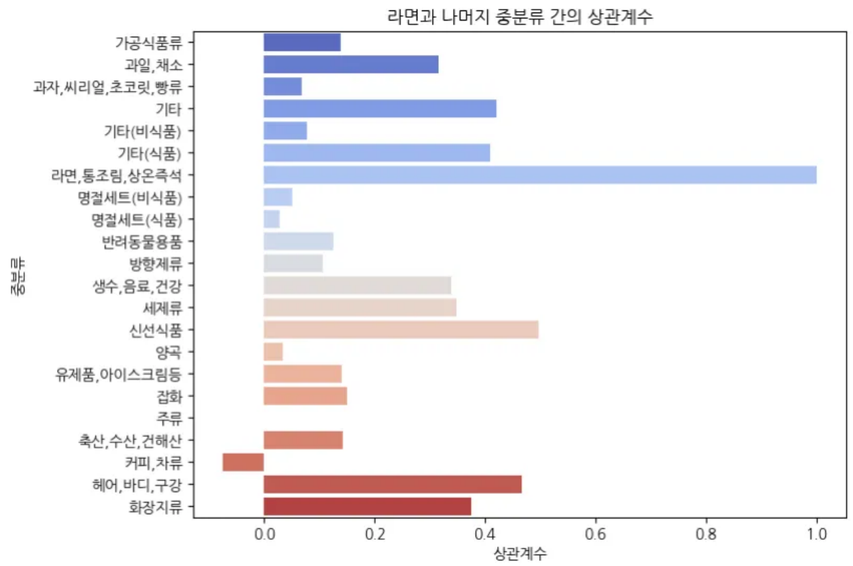


일별 데이터의 변동성을 줄이기 위해
주/월 단위로 grouping

1 데이터 : 주별로 판매수량 집계
2 데이터 : 월별로 판매수량 집계

02. 전처리, EDA

3. 중분류 간 상관계수 분석



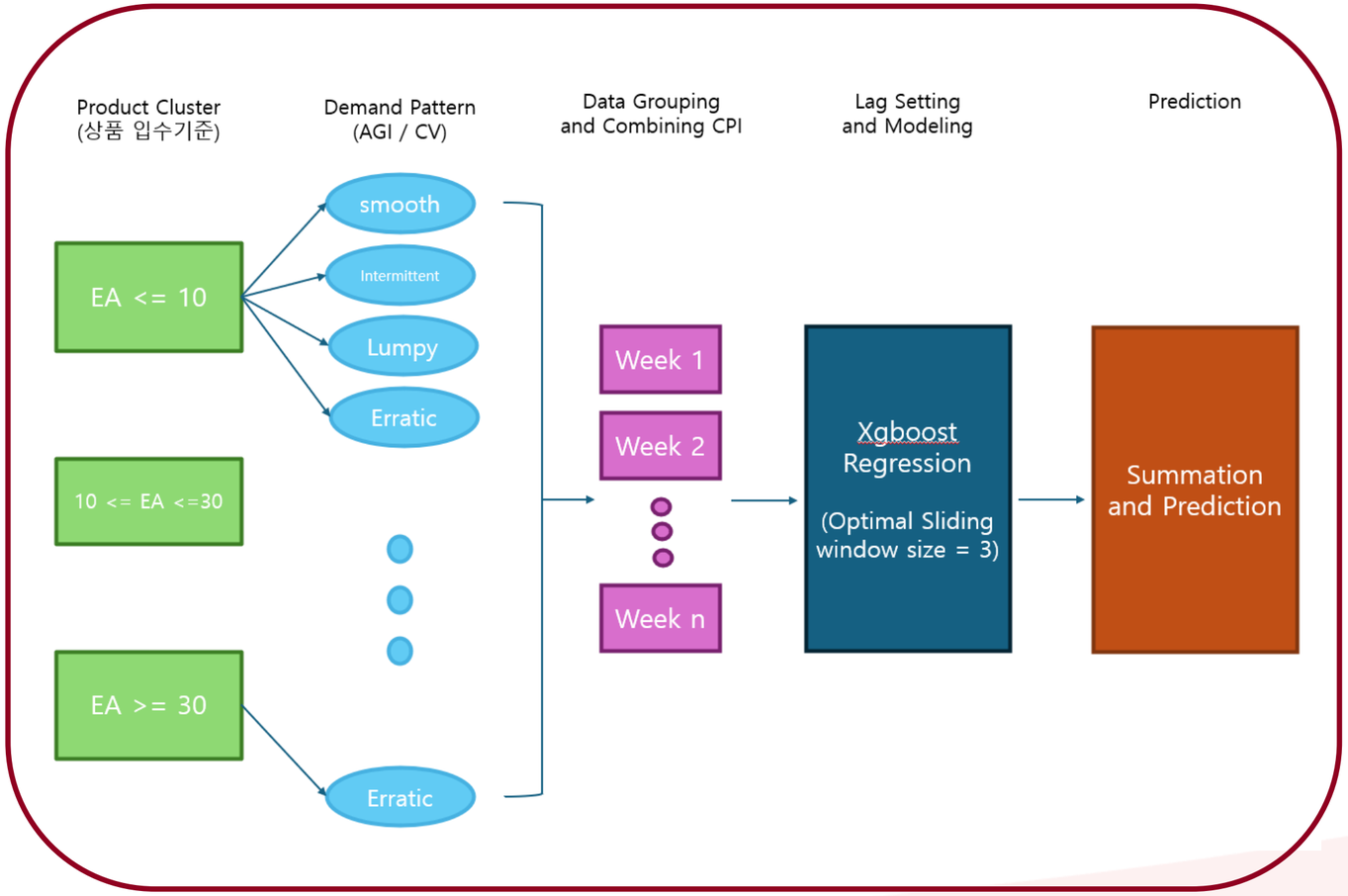
상품 분류별 데이터 상관관계 파악
0.7 이상인 타 상품군이 없어,
라면 데이터만으로 분류

4. 반품 데이터 제거, 이상치 확인

EWMA 차트 생성, 통계량의 3-sigma 기법
적용해 이상치 탐지
구분상 '반품'에 해당하는 데이터 제거

03. 분석 방법 및 절차

참고: 상품 수요 패턴과 클러스터를 고려한 월간 상품 수요 예측 모델 성능 비교(남효연)



- 상품별로 수요 패턴에 따라 군집을 나눈 후, 군집별로 각각 시계열 예측 방법론을 적용

- 개별 품목의 묶음 수를 의미하는 '입수' 값을 기준으로 3개의 군집으로 분류

03. 분석 방법 및 절차

Croston's 및 SyntetosBoylan 근사법의 고정 리드 타임에서의 이론적 배경을 기반으로, ADI와 CV 값을 기준으로 수요 패턴을 구분

ADI	CV	Demand Pattern
0 < ADI < 1.32	0 < CV < 0.49	Smooth
ADI >= 1.32	0 < CV < 0.49	Intermittent
0 < ADI < 1.32	CV >= 0.49	Lumpy
ADI >= 1.32	CV >= 0.49	Erratic

a. Smooth: 일정한 수량의 수요가 규칙적으로 발생

b. Erratic: 일정하지 않은 수량의 수요가 규칙적으로 발생

c. Intermittent: 일정한 수량의 수요가 불규칙적으로 발생

d. lumpy: 일정하지 않은 수량의 수요가 불규칙적으로 발생

03. 분석 방법 및 절차

1. 예측 시점 이전 n개의 값(lag)을 바탕으로 다음 값을 예측, n = 3

2. 각 시기에 대응되는 물가지수 데이터를 추가 설명변수로 선정

3. 각 군집별로 Lag 값과 물가지수를 기반으로 xgboost를 이용해 수요량 예측, 합산

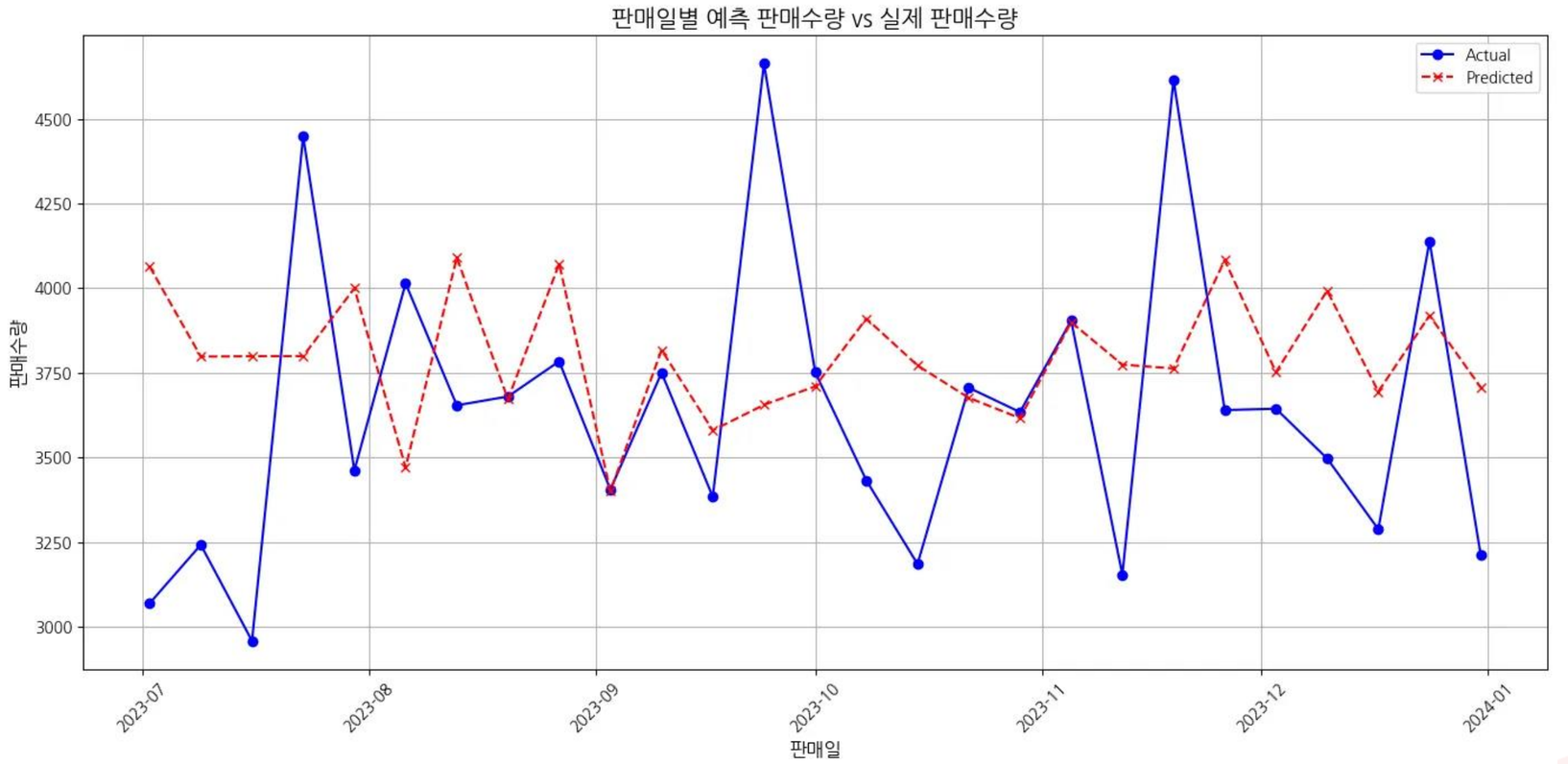
4. 2023년 하반기를 기점으로 train/test split, 성능 평가 수행

a. eval_metric: rmse

b. max_depth: 5

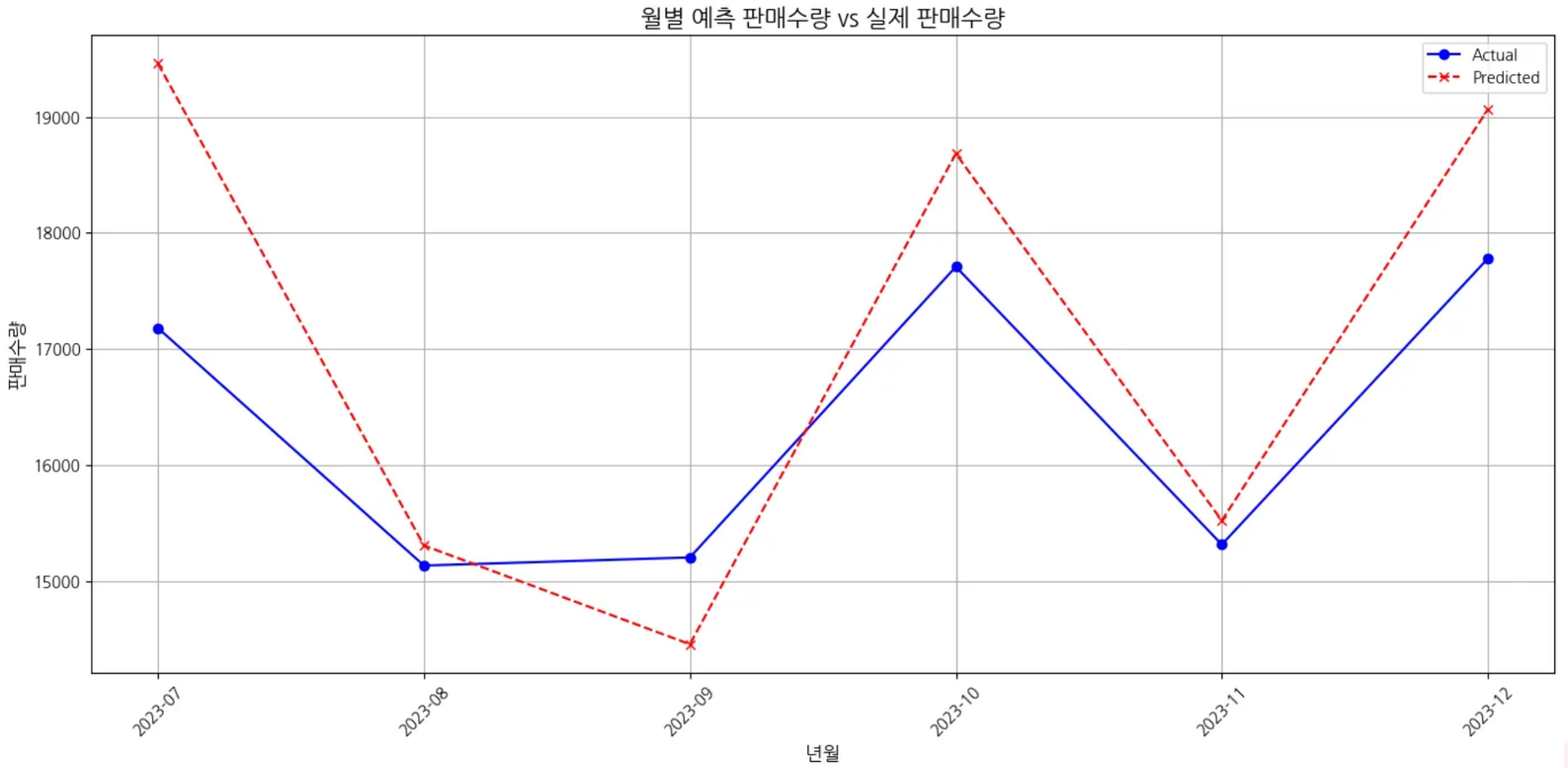
c. learning_rate: 0.1

04. 분석결과



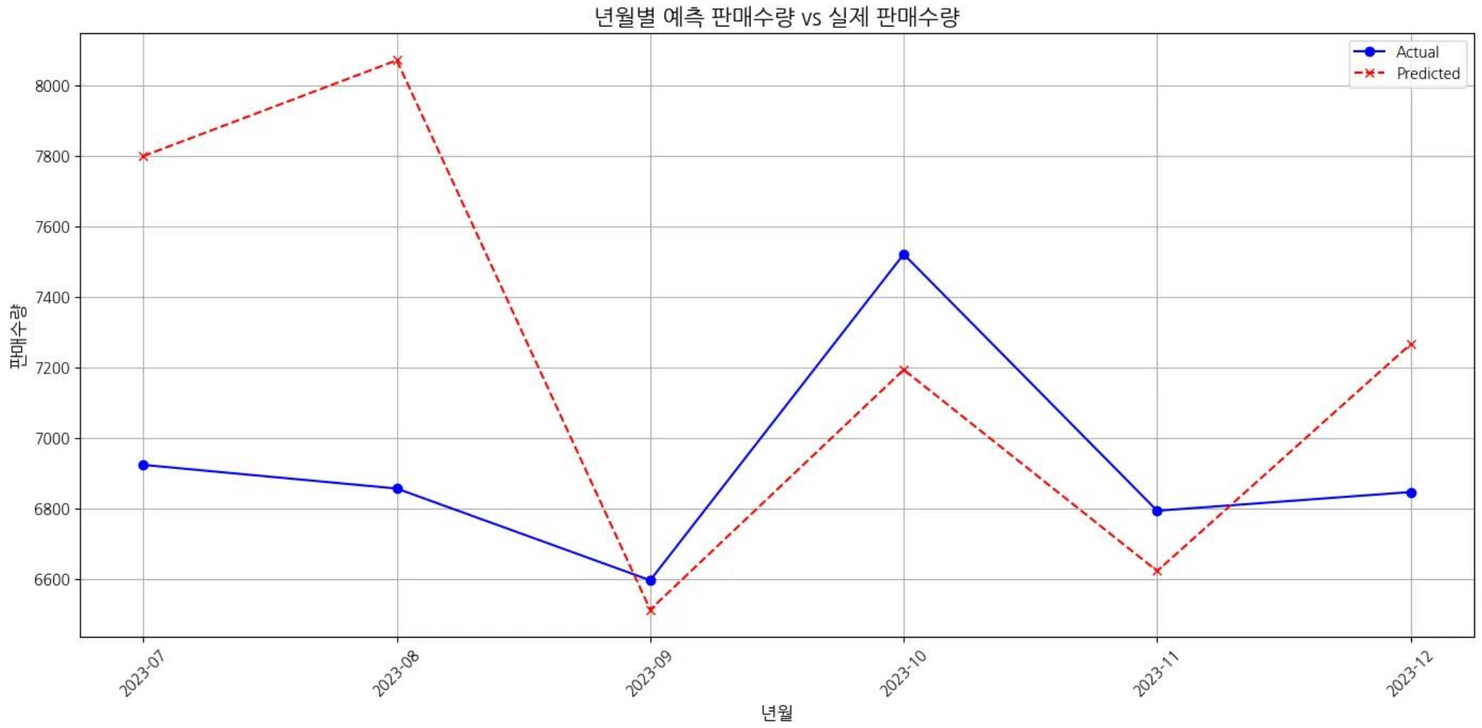
<1데이터 2023.07 ~ 2023.12 예측값과 실제 판매량 시각화 (주별)>

04. 분석결과



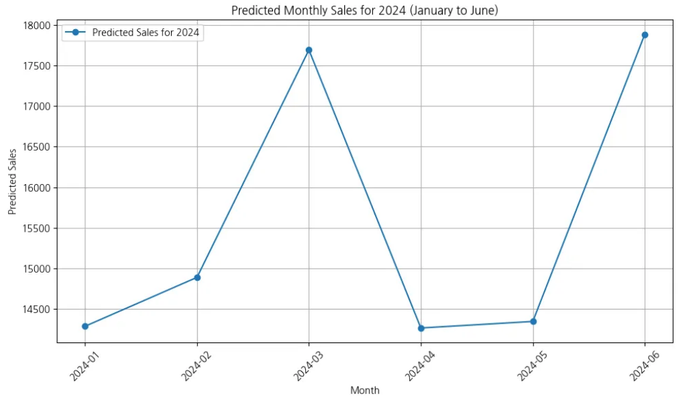
<1데이터 2023.07 ~ 2023.12 예측값과 실제 판매량 시각화 (월별)>

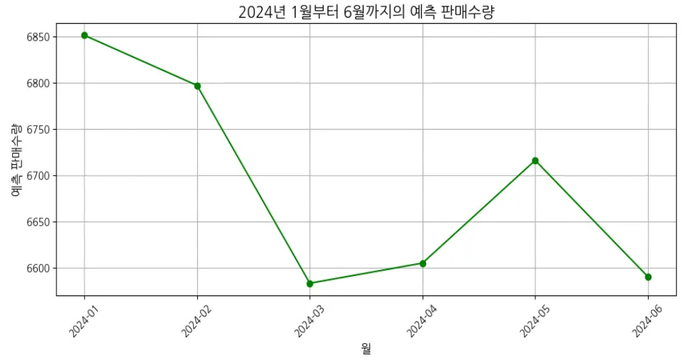
04. 분석결과



<2데이터 2023.07 ~ 2023.12 예측값과 실제 판매량 시각화 (월별)>

04. 분석결과





구분	1데이터 (중분류) 라면,통조림,상온즉석	2 데이터 (대분류) 면류.라면류
1월	14288	6851
2월	14890	6797
3월	17690	6583
4월	14267	6605
5월	14347	6716
6월	17880	6590

동일한 방식으로 23년까지의 데이터를 학습하여,
24년 1-6월 예측에 적용, 답안 제출

05. 활용 방안 & 피드백

활용 방안

- 수요 패턴에 따라 군집화한 모델로서 제품군별 재고 관리 최적화 수행
- 물가지수 외에도 기후, 경제지표, 사회적 이벤트 등 수요변화에 영향을 줄 수 있는 요인 탐색이 필요
- 라면류 카테고리 외 다양한 상품군으로 예측 범위 확대

피드백

- 시계열 예측에 강세를 보이는 모델 (SARIMA, prophet, LSTM 등) 을 시도해 보았으나, 기존 머신러닝 알고리즘 대비 성능이 좋지 못했음.
- '예측', '시계열' 카테고리로 방향성 수립, 추가로 관련 공모전 search

KUBIG
DATA SCIENCE & AI

2. 상수도 관망 이상 탐지

01. 대회 개요

2024 제4회 K-water AI 경진대회

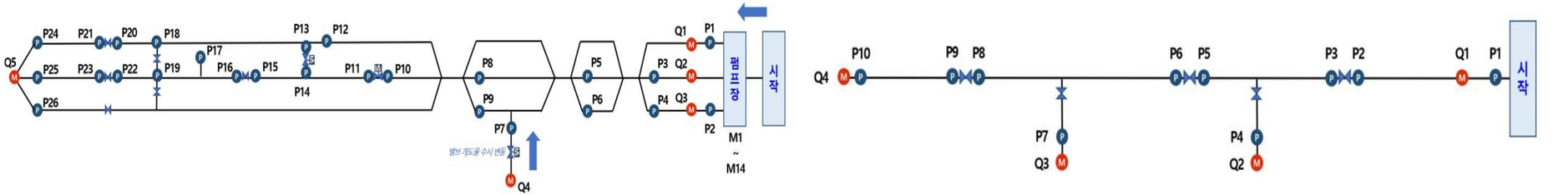
주제: 상수도 관망 이상 감지 AI 알고리즘 개발

[대회 목적]
- 목표: 상수도 관망의 이상 시점과 누수 발생 구간을 정확히 탐지하는 범용 AI 알고리즘 개발

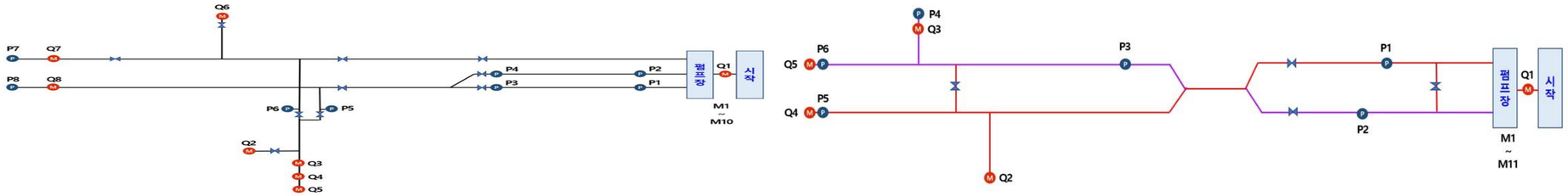
[제공 데이터]
- 학습 데이터:A와 B 구조의 상수도 관망 데이터 + 분 단위 시간 정보
- 평가 데이터:C와 D 구조의 상수도 관망 데이터 + 비식별화된 시간 정보

02. 데이터

Train 관망 구조 A, B



Test 관망 구조 C, D



Q1	Q2	Q3	Q4	P1	P2	P3	P4	P5	P6	P7	P8	P9	P10	anomaly
29277.5	7387.166	12025	9522.872	4.99	3.6862	3.6875	3.9337	4.085	3.9987	3.32	2.2863	3.4975	1.5838	0
28694.53	7378.35	11855	9555.664	4.99	3.6862	3.6925	3.9313	4.0813	3.9825	3.32	2.2887	3.4925	1.5725	0
28814.85	7399.673	12005	9555.664	4.965	3.6875	3.6888	3.9313	4.0837	3.89	3.29	2.3	3.5025	1.5725	0
29249.06	7321.313	12136	9451.097	4.99	3.6875	3.695	3.9288	4.08	3.8837	3.3	2.2825	3.4987	1.5438	0
30138.28	7315.561	12158	9451.097	4.99	3.685	3.7037	3.9331	4.0887	3.8912	3.3	2.2812	3.5013	1.5438	0
29674.53	7340.779	12398	9451.304	4.99	3.6837	3.69	3.9281	4.0875	3.8888	3.3	2.3038	3.49	1.5488	0
29449.22	7352.388	12101	9463.803	4.99	3.685	3.695	3.9181	4.0887	3.8925	3.3	2.3025	3.4913	1.56	0
29449.22	7314.973	12162	9463.803	4.99	3.7013	3.6962	3.9231	4.085	3.89	3.3	2.2863	3.5038	1.56	0
29465.63	7331.286	12171	9503.899	4.99	3.7	3.6913	3.9213	4.0813	3.8925	3.3	2.3025	3.4962	1.555	0

03. 주요과정

1. Rule-based Modeling

Data-driven 아닌 Rule-based 선택 이유

1. 이상치 시점에 해당하는 데이터 부족
2. 다양한 관망 구조에 대한 적용 가능 여부
3. 안정적 운영에 대한 설명 가능성과 신뢰성

➢ Rule-based modeling!!

- 도메인 지식 수집
- 규칙 정의
- 알고리즘 설계

2. Data-driven Modeling

➢ LSTM Autoencoder
- 유량 및 압력 데이터는 시계열 특성이 강함.
- LSTM은 시계열 데이터의 장기 의존성을 학습하여 데이터의 변화를 예측.

➢ GNN + LSTM
- 관망 구조 -> 그래프로 모델링 (노드는 센서, 엣지는 연결성).
- 관망 구조를 그래프로 모델링. 노드 간 상호작용과 연결 정보를 학습하기에 적합.

➢ Anomaly Transformer
- 다변량 이상치 탐지 분야의 SOTA 모델 중 하나.
- 기존의 단기 시계열 특징 뿐만 아니라 self-attention구조를 차용하여 장기 시계열 특징을 반영할 수 있음.

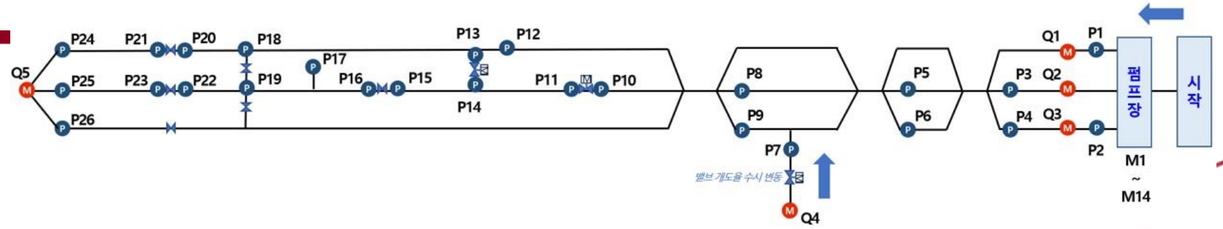
03-1. Rule-based Modeling

도메인 기반 가설

- 상수도의 적절한 운영을 위해서는 유량, 압력, 펌프 작동 여부 등 종합적인 요소가 고려되어야 함.
- 유량 – 총 공급량과 총 사용량의 균형 유지
- 압력 – 압력이 큰 변동폭의 변화를 가져가지 않도록 하는 것이 중요

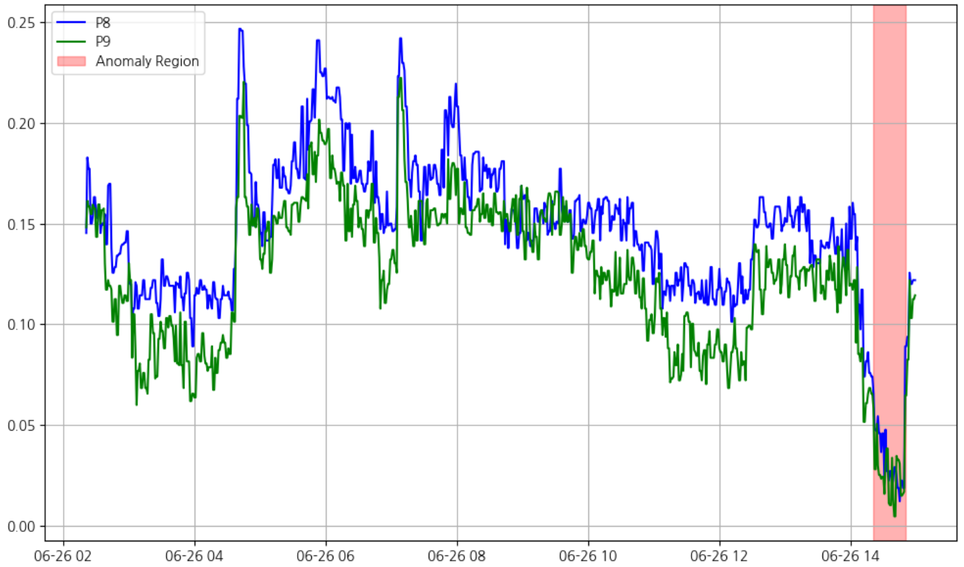
이상 시점 탐지 / 누수 발생 구간 판별!

03-1. Rule-based Modeling



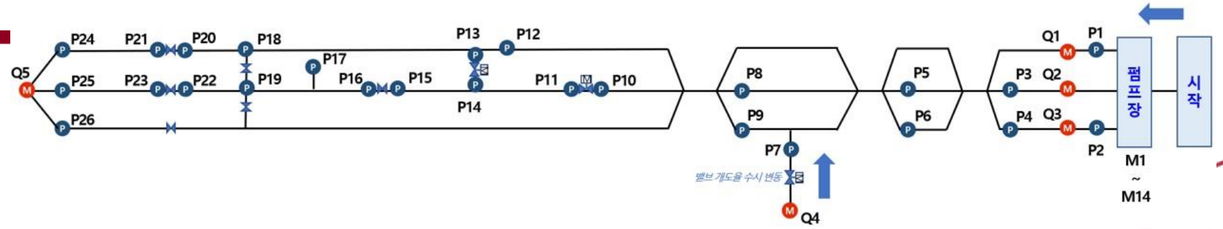
Task 1. 이상 시점 탐지

누수가 발생한 구간의 압력계의 압력 비교



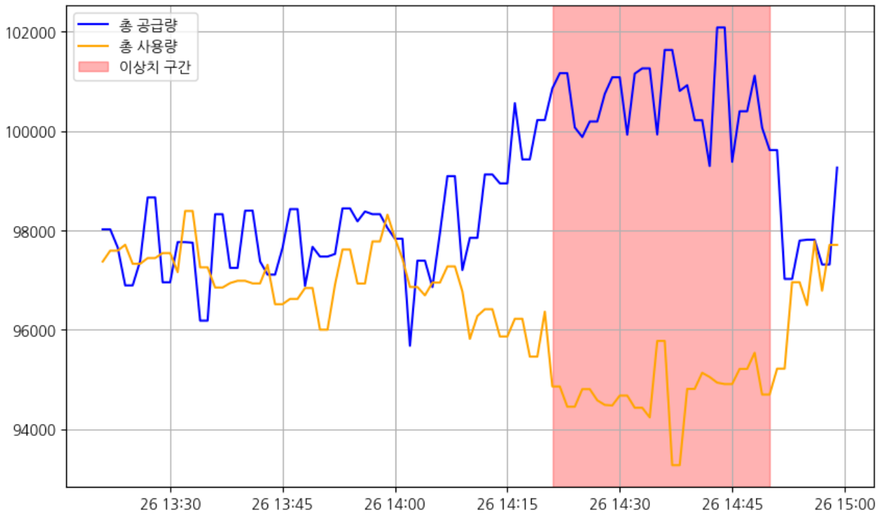
- 이상치 구간에서 두 압력계의 압력이 모두 급격히 낮아진 것을 확인
- 그러나 압력이 급격히 변한 구간이 이상 시점 외에도 다량 발견되어 명확한 기준 확립이 어렵다고 판단

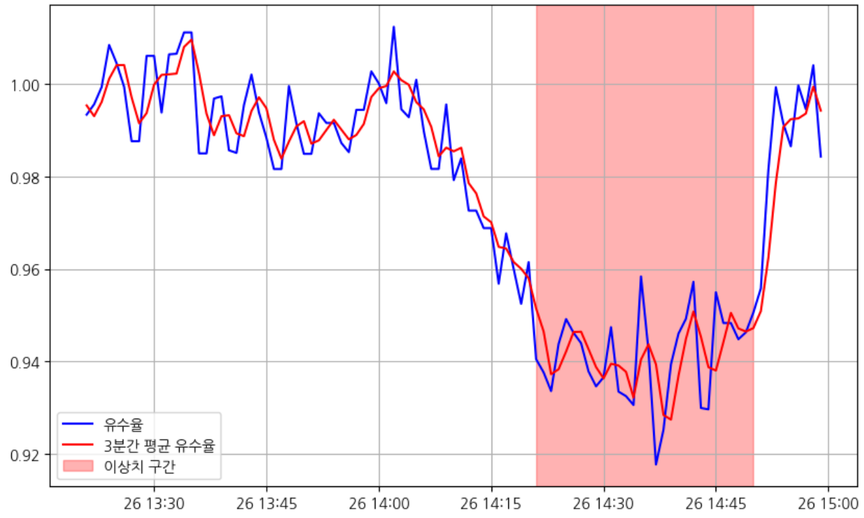
03-1. Rule-based Modeling



Task 1. 이상 시점 탐지

총 공급량 & 총 사용량 비교 (유량)

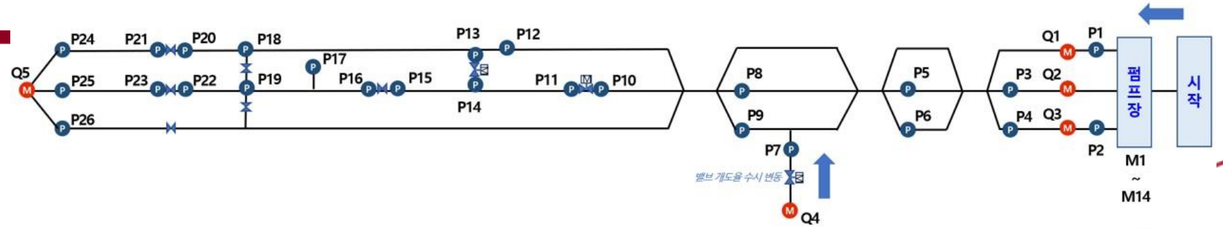


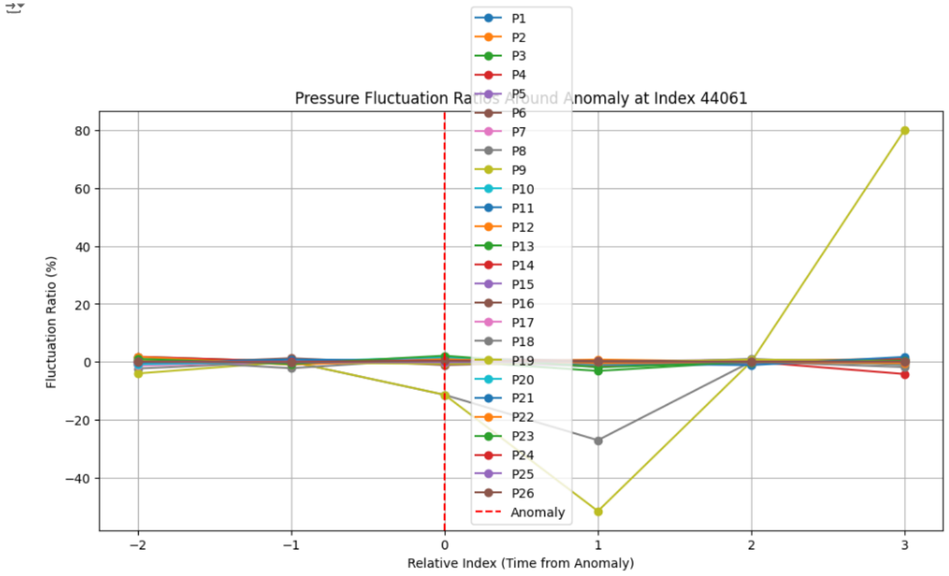


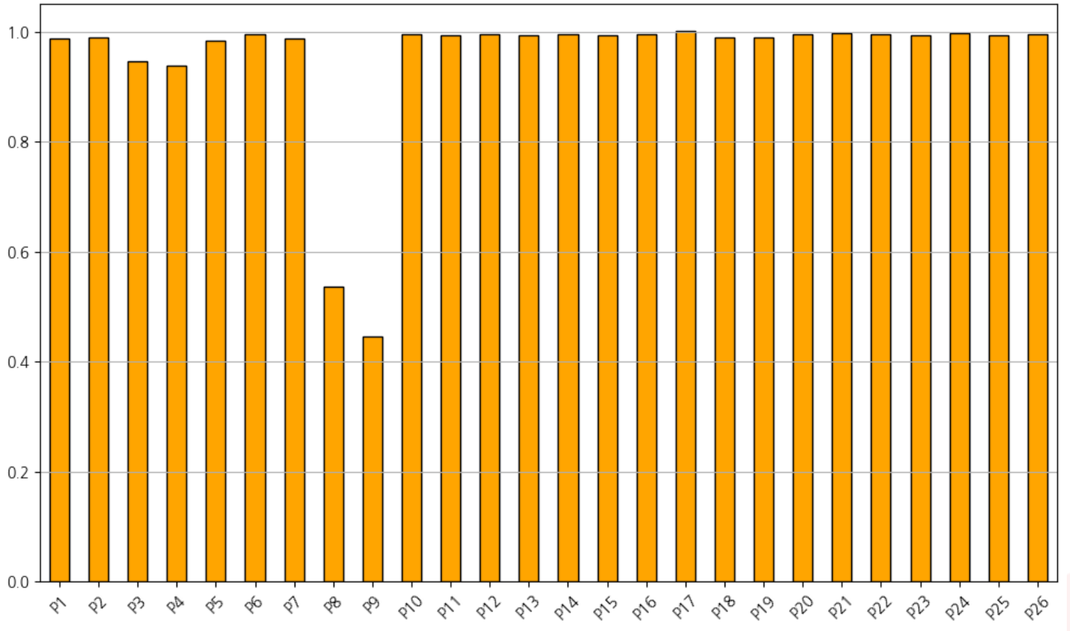
- 이상치 구간에서는 총 공급량과 총 사용량의 간극이 커지는 경향이 있음을 확인
- 이상 시점 탐지는 유량을 활용하기로 결정

03-1. Rule-based Modeling

Task 2. 누수 발생 구간 판별



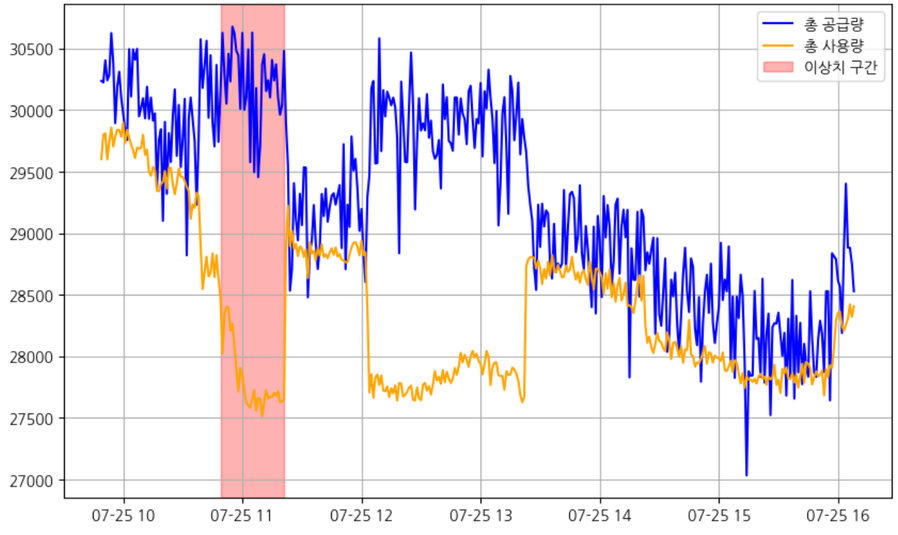


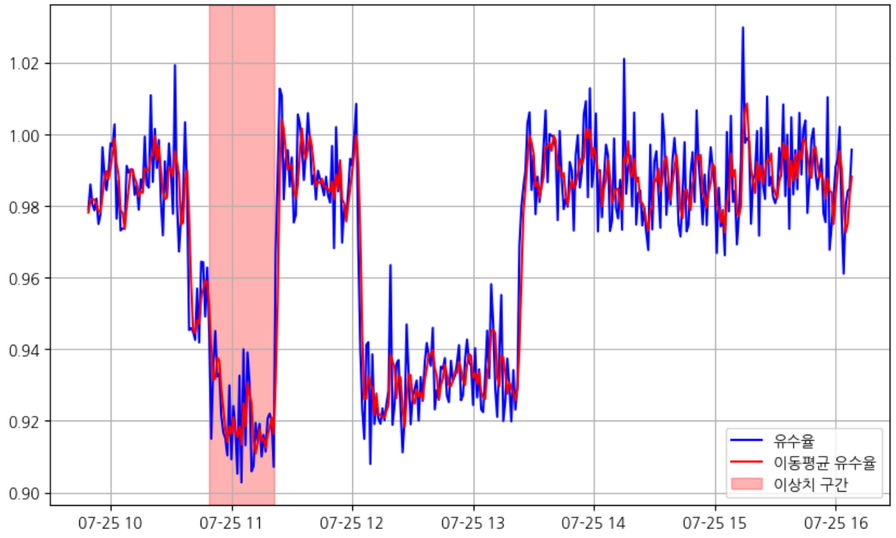


- 누수가 발생한 구간의 압력계의 압력이 낮게 나타나는 것을 확인
- 누수 발생 구간 탐지는 압력 값 활용

03-1. Rule-based Modeling

B 관망 데이터 제외





- 유량의 총 공급량과 총 사용량의 격차가 생기더라도 일부 구간에서는 여전히 anomaly = 0
- 해당 구간들을 분류할 또 다른 단서를 찾기에는 데이터가 한정적이라고 판단
- 주어진 데이터 내에서 적절한 threshold를 선정하기에는 A 관망 데이터만으로도 충분할 것으로 예상

03-1. Rule-based Modeling

유량(Q) 관련 Rule & Threshold

* 3분 평균 유수율 : 1분 단위가 아닌 최근 3분간의 평균 유수율을 계산, range를 늘려 압력 변동폭을 줄인 형태의 feature

이상 시점 탐지

- 3분 평균 유수율 TH_1% 이하 M_1분 지속 & 유수율 TH_2% 이하 M_2분 지속
- 유수율 TH_3% 이하 M_3분 지속
- 최근 M_4분 이내 최대 유수율이 TH_4% 이상

<최적 조건>

TH_1	97.5	M_1	4
TH_2	97.5	M_2	6
TH_3	95	M_3	3
TH_4	100	M_4	6

03-1. Rule-based Modeling

압력(P) 관련 Threshold

누수 발생 구간 탐지

- 전체 평균 변동비 대비 각각의 압력계 변동비가 TH_5 이하인 압력계가 누수 구간
- 해당하는 압력계가 없을 경우 변동비가 제일 낮은 압력계가 누수 구간

* 변동비 산정 방법

변동비 = 누수 발생 구간 압력계 별 압력 평균 / 정상 압력 구간(이상 발생 직전 M_5분간) 동안 압력계별 압력 평균

< 최적 조건 >

TH_5	97.5	M_5	30

03-1. Rule-based Modeling

결과

제목	제출 일시	public점수 private점수
AIM4.csv edit	2024-12-29 20:18:24	0.75 0.4090909091
AIM3.csv edit	2024-12-29 19:48:30	0.7714285714 0.4153846154
AIM2.csv edit	2024-12-29 14:21:32	0.5714285714 0.3076923077
AIM.csv edit	2024-12-29 13:28:33	0.9 0.45
AIM.csv edit	2024-12-29 13:00:48	0 0

Public Score:
전체 테스트 샘플 중 '관망 구조 C' 샘플

Private Score:
전체 테스트 샘플 100%

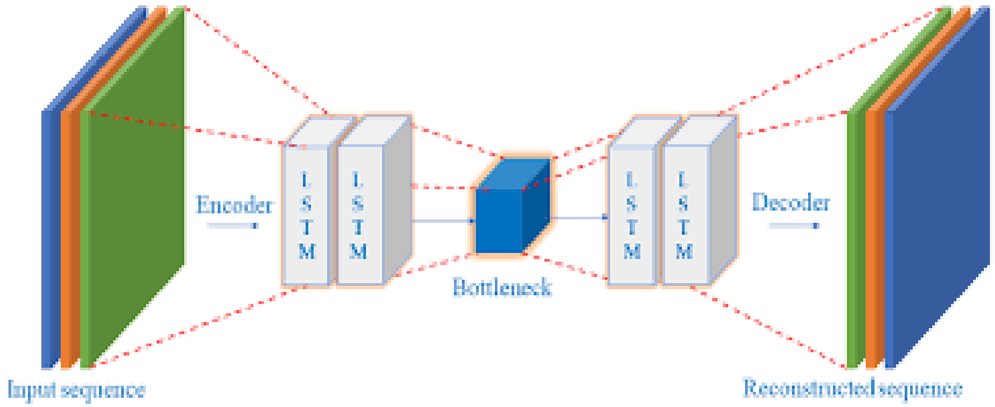
- ✓ 관망 구조 C에 대한 이상 시점 및 누수 구간 탐지 성능은 우수
- ✓ 관망 구조 D에 대한 누수 구간 탐지는 실패
- ✓ 누수 구간 탐지 시 불분명한 압력 조건이 원인으로 추정됨

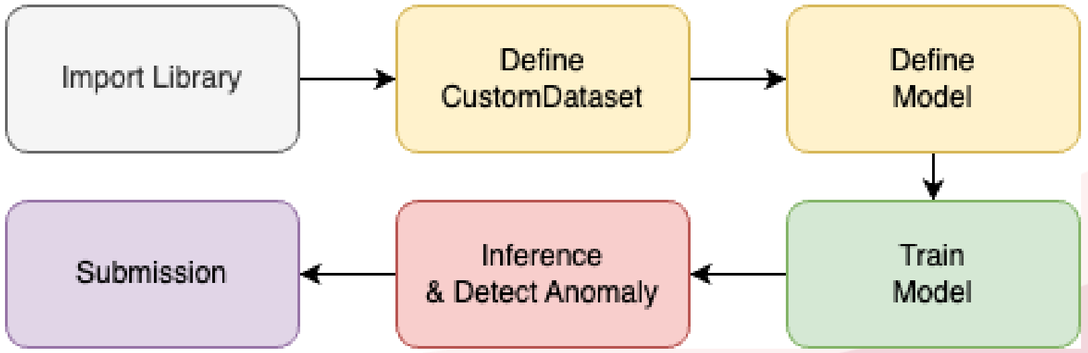
03-2. Data-driven Modeling

LSTM Autoencoder

Baseline Code

- Min-max 정규화
- 관망구조를 반영하지 않고,
  단일 변수의 시퀀스 반영
- Relu & Adam 사용





03-2. Data-driven Modeling

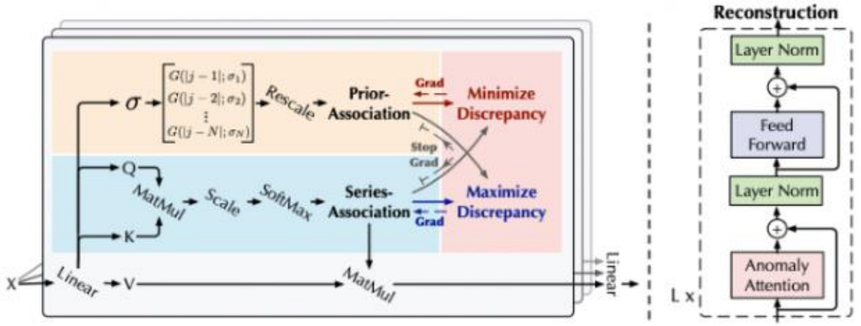
Anomaly Transformer

ANOMALY TRANSFORMER: TIME SERIES ANOMALY DETECTION WITH ASSOCIATION DISCREPANCY

Jiehui Xu,* Haixu Wu,* Jianmin Wang, Mingsheng Long (✉)
School of Software, BNRist, Tsinghua University, China
{xjh20,whx20}@mails.tsinghua.edu.cn, {jimwang,mingsheng}@tsinghua.edu.cn

ABSTRACT

Unsupervised detection of anomaly points in time series is a challenging problem, which requires the model to derive a distinguishable criterion. Previous methods tackle the problem mainly through learning pointwise representation or pairwise association, however, neither is sufficient to reason about the intricate dynamics. Recently, Transformers have shown great power in unified modeling of pointwise representation and pairwise association, and we find that the self-attention weight distribution of each time point can embody rich association with the whole series. Our key observation is that due to the rarity of anomalies, it is extremely difficult to build nontrivial associations from abnormal points to the whole series, thereby, the anomalies' associations shall mainly concentrate on their adjacent time points. This adjacent-concentration bias implies an association-based criterion inherently distinguishable between normal and abnormal points, which we highlight through the *Association Discrepancy*. Technically, we propose the *Anomaly Transformer* with a new *Anomaly-Attention* mechanism to compute the association discrepancy. A minimax strategy is devised to amplify the normal-abnormal distinguishability of the association discrepancy. The Anomaly Transformer achieves state-of-the-art results on six unsupervised time series anomaly detection benchmarks of three applications: service monitoring, space & earth exploration, and water treatment.



$$\mathcal{Z}^l = \text{Layer-Norm}\Big(\text{Anomaly-Attention}(\mathcal{X}^{l-1}) + \mathcal{X}^{l-1}\Big)$$
$$\mathcal{X}^l = \text{Layer-Norm}\Big(\text{Feed-Forward}(\mathcal{Z}^l) + \mathcal{Z}^l\Big),$$

03-2. Data-driven Modeling

Anomaly Transformer

```
adjacency_list_B = {
    'P1': [['P2'], [], ['Q1']],
    'P2': [['P3'], ['Q1'], ['Q2','Q3','Q4']],
    'P3': [['P4','P5'], ['Q1'], ['Q2','Q3','Q4']],
    'P4': [[], ['Q1'], ['Q2']],
    'P5': [['P6'], ['Q1'], ['Q3','Q4']],
    'P6': [['P7','P8'], ['Q1'], ['Q3','Q4']],
    'P7': [[], ['Q1'], ['Q3']],
    'P8': [['P9'], ['Q1'], ['Q4']],
    'P9': [['P10'], ['Q1'], ['Q4']],
    'P10': [[], ['Q1'], ['Q4']]
}
nodes_B = list(adjacency_list_B.keys())
```

대상 P 노드

timestamp	Q1	Q2	Q3	Q4	P1	P2	P3	P4	P5
2024-07-01 0:00	29277.5	7387.166	12025	9522.872	4.99	3.6862	3.6875	3.9337	4.085
2024-07-01 0:01	28694.53	7378.35	11855	9555.664	4.99	3.6862	3.6925	3.9313	4.0813
2024-07-01 0:02	28814.85	7399.673	12005	9555.664	4.965	3.6875	3.6888	3.9313	4.0837
2024-07-01 0:03	29249.06	7321.313	12136	9451.097	4.99	3.6875	3.695	3.9288	4.08
2024-07-01 0:04	30138.28	7315.561	12158	9451.097	4.99	3.685	3.7037	3.9331	4.0887
2024-07-01 0:05	29674.53	7340.779	12398	9451.304	4.99	3.6837	3.69	3.9281	4.0875
2024-07-01 0:06	29449.22	7352.388	12101	9463.803	4.99	3.685	3.695	3.9181	4.0887
2024-07-01 0:07	29449.22	7314.973	12162	9463.803	4.99	3.7013	3.6962	3.9231	4.085
2024-07-01 0:08	29465.63	7331.286	12171	9503.899	4.99	3.7	3.6913	3.9213	4.0813
2024-07-01 0:09	29544.38	7324.926	12184	9503.899	4.99	3.7	3.695	3.925	4.0813
2024-07-01 0:10	29739.06	7389.446	12119	9483.51	4.99	3.6825	3.695	3.92	4.0887
2024-07-01 0:11	29781.72	7310.959	12185	9483.51	4.99	3.705	3.695	3.915	4.08

In Q　　Out Q　　이전 P노드　　이후 P노드

- 관망구조를 전이행렬로 취급
- 단일 P변수에 대한 전후 n개의 노드 및 in-out Q변수를 사용하여 다변량 컬럼을 구성
- **Anomaly Score** = Reconstruction loss + Association Discrepancy

04. 한계점 및 개선방안

1. Rule-based Modeling

관망 구조 데이터 활용 미흡

- 현재 모델은 데이터로 주어진 관망 구조 및 유량계, 압력계 등 관련된 노드들 간의 연결성 등을 반영하지 못함

Rule-based modeling의 한계

- 충분한 도메인 지식을 바탕으로 정확하고 유의미한 가설을 설정하는 것이 중요함.
- 유량 및 압력 데이터를 하나하나 들여다보면서 적정 Threshold 값을 산출하는 과정이 효율적이지 않음
- 개발자의 주관적인 판단에 지나치게 의존함.

규칙 기반 + 알고리즘 융합

- 전략초기 탐지는 규칙 기반으로 수행, 이상 발생 구간의 압력 변화 및 유량 패턴을 ML & DL 모델로 재탐지 및 검증.
- Threshold 수치를 데이터 기반으로 자동으로 학습하도록 개선하여 분석가의 주관적 개입 최소화.

04. 한계점 및 개선방안

2. Data-driven Modeling

다변량 구조에 따른 동적인 In – Output 구성 필요

- 현재 모델은 단일 P변수에 따른 단일 P_flag예측으로 변환하여 훈련 및 예측하는 과정을 거침.
  => 데이터셋의 전체 구조를 반영하지 못하고 비효율적임.
- 데이터 셋에 차이에 따른 변수의 개수 및 유무(ex. M)와 이에 따른 Target 변수의 개수를 고려한 동적인 모델을 구축하는 방향으로 발전할 수 있음.

M변수의 배제와 기타 파생변수의 필요

- M (펌프가동유무) 변수는 데이터셋 별 존재유무가 다르므로 모델의 Feature로 활용하는데 어려움을 겪음.
- 이 밖에도 관망구조에 따른 그룹화나 P-Q 상관성 등 추가적인 분석을 통한 Feature Engineering이 필요함.

KUBIG
DATA SCIENCE & AI

Thank You