제2회 유통데이터 활용 경진대회
(데이터모델링부문) 분석보고서

1. 팀명 : 원볼원 (정순주, 조수익)

2. 주제

Prediction-SSSD 생성형 인공지능 모형을 활용한 수요 예측

3. 세부내용

I. 서론

본 프로젝트는 지역 물류센터 내 판매 수량 데이터를 활용하여 향후 판매 수요를 예측한다. 이를 위해 생성형 인공지능 모형인 Diffusion기반 SSSD 모형을 활용한다[1]. SSSD는 시계열 데이터 예측에 있어 State-of-The Art (SOTA)를 기록한 모형으로 현재 가장 강력한 시계열 예측 모형이다. 이 모형을 통해 본 프로젝트에서는 외부 데이터 없이 최소한의 데이터로 수요를 예측할 수 있는 모형을 구현하였으며, 데이터가 적은 산업 현장에서도 도입 가능함을 입증하였다. 또한 지역 물류 센터 내 판매 수량 데이터에서 MSE기준 0.015를 기록하며 높은 정확도를 기록하였다.

II. 본론

1. 분석 목표

   실제 산업에 유용하게 활용될 수 있도록 최소한의 데이터를 학습하여 높은 예측 결과를 도출.

2. 데이터 분석 개요

   1) Dataset 정의

      - 2021년 1월 ~ 2022 6월까지의 지역 물류센터 내 판매 수량 데이터셋

   2) Dataset 개괄

      - 데이터 개수: 총 521,995개

- 데이터 Columns

이름	설명	예시	유형
판매일	2021년 1월 4일 ~ 2022년 7월 1일	2021-01-04	Datetime
구분	하위 카테고리로 판매유형을 구분	매출/반품	Object
우편번호	지역별 우편번호	37542	Object
판매수량	해당 판매 수량	5	Int64
옵션코드	상품별 포장단위 약어 구분	EA	Object
규격	포장단위 내 상품 용량 및 개수	50g*16*1	Object
입수	포장단위 내 상품 개수	1	Int64
상품 바코드	상품 바코드	8801077334102	Float64
상품명	상품 이름	면도기]프레쉬 708<10입>	Object

3) Dataset 분석 과정

- 우편번호, 옵션코드, 규격, 입수 컬럼 제외
- 결측치 데이터 제외
- 상품명 기준, 최종 타겟 품목명과 관련 없는 상품 제외
- 판매주차 칼럼 추가
- 이상치 데이터 제외
  - ♦ 판매주차 기준, 판매기록이 2번 이하인 데이터 제외
  - ♦ 판매주차 기준, 판매수량의 총합이 0인 데이터 제외

	판매일	구분	판매수량	상품 바코드	상품명	판매주차
9	2021-01-04	매출	3	1701001521813	신라면 컵 6入 XX	2021-01
22	2021-01-04	매출	3	8801043045674	농심]안성탕면 컵 6入	2021-01
56	2021-01-04	매출	1	18801045522286	오뚜기]진라면매운멀티<120g>	2021-01
72	2021-01-04	매출	1	8801043015028	농심]너구리 얼큰멀티<40>	2021-01
73	2021-01-04	매출	1	8801043014847	농심]신라면 멀티<40>	2021-01
77	2021-01-04	매출	1	8801043015721	농심]신라면 컵<30>	2021-01
83	2021-01-04	매출	1	1701006157383	코카콜라<500ml*24>	2021-01
90	2021-01-04	매출	1	8801043015271	농심]짜파게티 멀티<40>	2021-01
91	2021-01-04	매출	1	68801056290308	레쓰비<175ml*30>	2021-01
104	2021-01-04	매출	5	8808244208044	삼다수2L	2021-01

[그림 1] 데이터 분석 후 데이터셋 예시

3. 개발 모형

1) SSSD 모형

Structured State Space Diffusion(SSSD)은 본 프로젝트의 개발 모형인 Prediction-SSSD 모형의 베이스 모형으로써, 생성형 인공지능 모형 Diffusion을 기반으로 한다. SSSD는 Convolution layer를 통하여 시계열 데이터의 특성을 추출한다. 추출된 특성은 잠재영역에서 Diffusion 모형을 통해 분석되어 Decoder 역할의 Inference를 통해 데이터를 생성한다. SSSD는 여러 시계열 데이터 벤치마크들에서 강력함과 강건함을 보이며 가장 높은 성능을 기록하고 있다고 발표되었다.

2) Predicition-SSSD 모형

본 프로젝트에서는 SSSD의 Masking을 Customizing하여 지역 물류센터 내 판매수량 예측 모형을 구현하였다. Prediction-SSSD는 전체 데이터의 세그먼트를 분할하여 임의로 추출된 세그먼트를 Masking하지 않고, 대회의 문제에 따라 상/하반기로 구분된 데이터 중 하반기만을 Masking하여 절반의 데이터로 나머지 데이터를 예측하도록 설계되었다. 이를 통해 지역 물류센터와 같이 중소규모의 센터에서도 정교하게 물류를 예측할 수 있는 모형을 구현하였다.

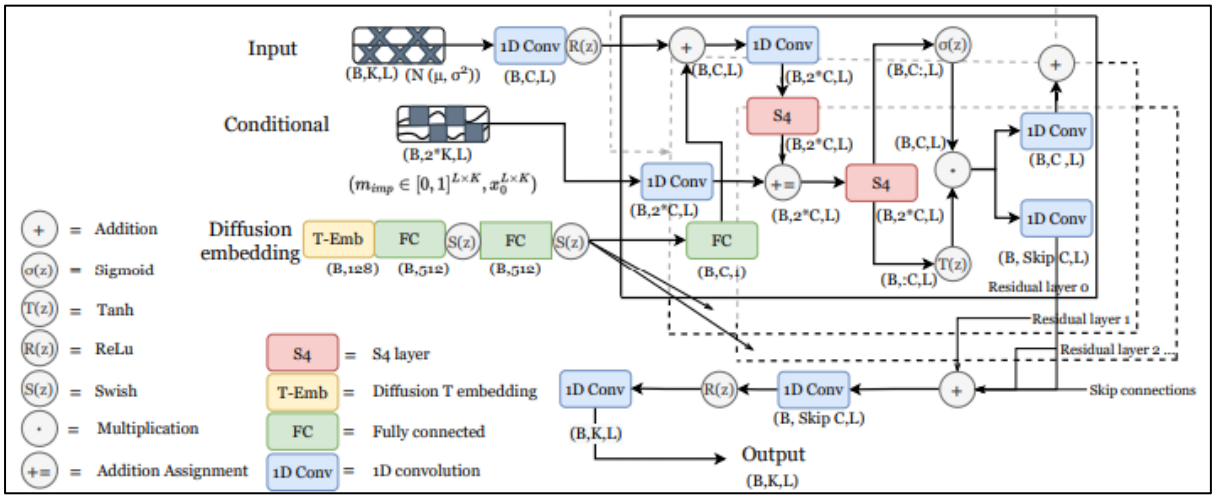
3) 입력 데이터

모형의 입력데이터의 형식은 배열 형식의 시계열 데이터이다.

```
array([[[ 6.30365146e-01, -8.41320441e-02, -2.41321426e-01,
          -1.62726735e-01, -2.69722689e-02],
        [ 4.75134930e-01, -1.17803816e-01, -1.80518491e-01,
          -1.40609152e-01, -1.00699813e-01],
        [ 8.24792444e-02, -1.37710062e-01, -1.23030775e-01,
          -2.18446141e-01, -2.76154089e-02],
        ...,
        [ 7.94856517e-01, -1.38829291e-01, -2.06596809e-01,
          -1.91537361e-01,  1.02121885e-01],
        [ 6.19524112e-01, -2.70277126e-01,  4.11533073e-02,
          -3.14767188e-01,  1.65725481e-01],
        [ 3.55674296e-01, -2.78301929e-01, -2.03315494e-01,
          -2.98752775e-01,  4.30660731e-01]],
```

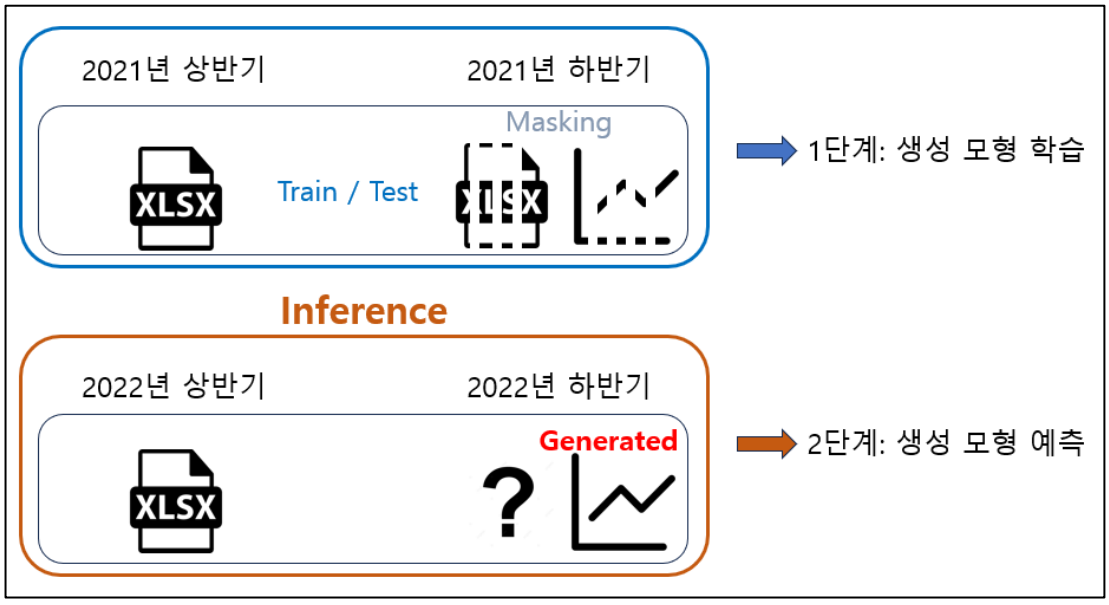
[그림 2] 입력 데이터의 예시

4) SSSD 모형 구조도



[그림 3] SSSD 모형 구조도

5) Prediction-SSSD 모형 학습 구조도

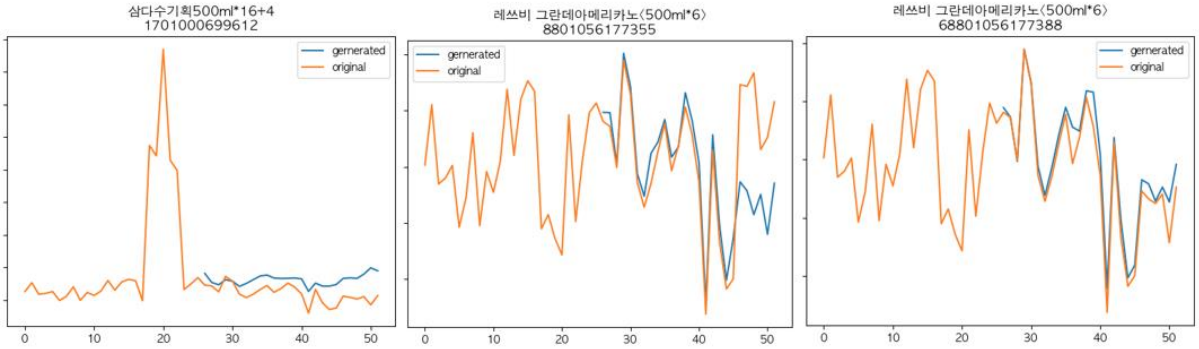


[그림 4] Prediction-SSSD 모형 학습 구조도

III. 결론

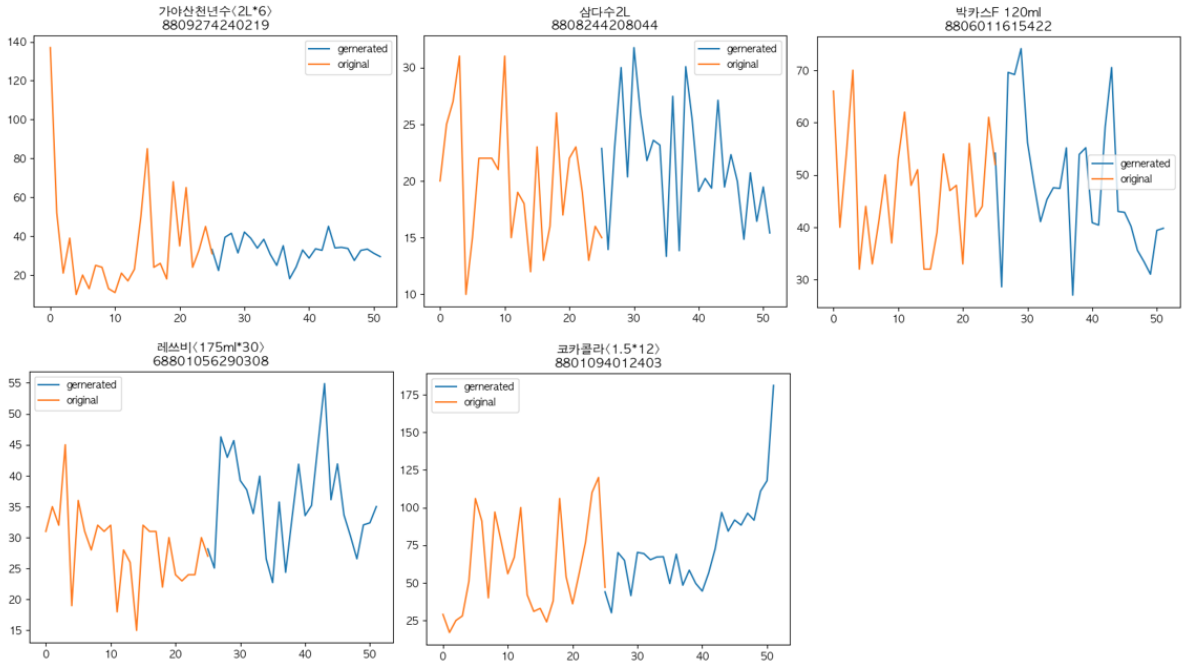
1. 품목별 결과 그래프

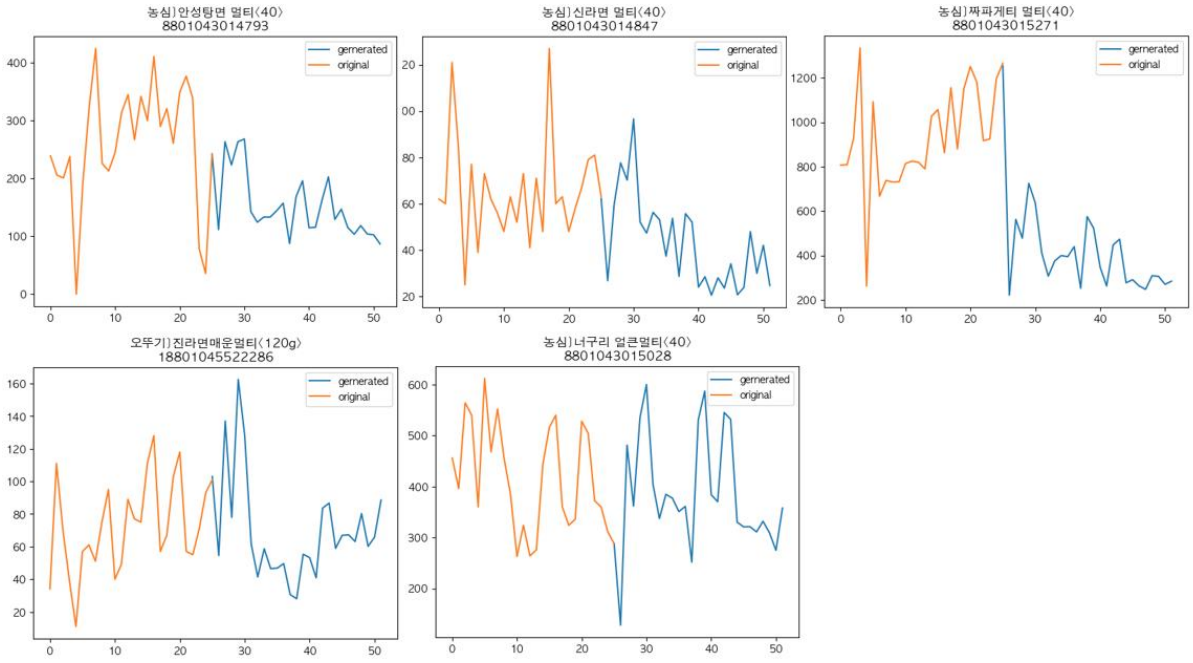
1) 1단계 모형 학습 결과



[그림 5] 품목 중 임의 추출된 3개 상품의 그래프

2) 2단계 모형 예측 결과





[그림 6] Target 품목 10개의 2022년 하반기 수요 예측 생성 그래프

2. 성능 평가

```
output directory ./results/contest/150000/valid/T200_beta00.0001_betaT0.02
SSSDS4Imputer Parameters: 48.035333M
Successfully loaded model at iteration 150000
Data loaded
begin sampling, total number of reverse steps = 200
generated 2 utterances of random_digit at iteration 150000 in 19 seconds
saved generated samples at iteration 150000
########################################
########################################
########################################
########################################
prediciton-SSSD valid dataset MSE : [0.015236701]
```

[그림 7] 1단계 모형 평가 결과 (MSE: 0.015)

2021년 데이터를 활용한 1단계 모형 학습 과정에서 상반기 데이터의 특성으로 하반기 데이터를 정교하게 예측하였다. 이는 모형의 평가 결과인 MSE (0.015)와 평가 결과 그래프를 통해 확인할 수 있다.

1) 2021년 데이터를 활용한 모형 학습에서 하반기 데이터에 Masking을 사용함으로써 견고성을 높인다. Prediction-SSSD에서 Masking은 수요데이터의 중간이 유실되거나 비정상적인 값이 있어도 이에 민감하지 않고 안정적으로 학습할 수 있는 역할을 하였다.

2) 그래프 계형 또한 자연스럽고 잘 따라가는 양상을 보인다. 이는 모형이 의미 있는 패턴을 학습하고, 이를 그래프 형태로 자연스럽게 표현할 수 있음을 보여주

며, 이러한 결과는 모형이 안정적이며 효과적으로 작동함을 시사한다.

3. 모형 특징

1) 외부 데이터 불필요

이 모형은 외부 데이터에 의존하지 않고, 주어진 내부 데이터만을 활용하여 효과적으로 작동한다. 이는 데이터 수집의 번거로움을 줄이고 독립적인 예측 모형을 구축할 수 있게 한다.

2) 소량 데이터 활용 가능

품목별 데이터가 소량일 경우에도 예측하고자 하는 품목의 수요를 모형을 통해 빠르고 실용적으로 예측 가능하다. (모형에 사용된 품목 수: 80개)

3) 단기 데이터 활용 가능

짧은 기간 동안의 데이터를 활용하여 수요 예측을 수행할 수 있기 때문에 신속한 의사 결정 및 비용 절감이 가능하다. (모형에 사용된 데이터의 기간: 18개월)

4. 모형 활용 방안

1) 재고 관리 비용 절감

해당 모형은 미래의 수요량 예측만 아니라 판매 수량을 숫자로 제시함으로써 재고 관리 비용을 절감하는데 도움을 준다. 이를 통해 재고 조정과 구매 계획을 최적화할 수 있다.

2) 그래프 자료 활용

모형이 생성한 그래프를 활용하여 품목별 주기성 관측 및 특정 이벤트 (예: 코로나)와 같은 특정 이벤트에 따른 급격한 수요 변화를 시각적으로 확인 분석할 수 있다. 이를 통해 새로운 비즈니스 기회 창출 및 실시간 모니터링과 조정에 기여할 수 있다. 그래프를 활용하여 시각화가 뚜렷하기 때문에 품목별 주기성 관측 및 코로나와 같은 특정 이벤트에 대한 수요 영향을 예측할 수 있다.

참고문헌

[1] Juan M.L. Nils S., Diffusion-based Time Series Imputation and Forecasting with Structured State Space Models, TSMR, February. 2023