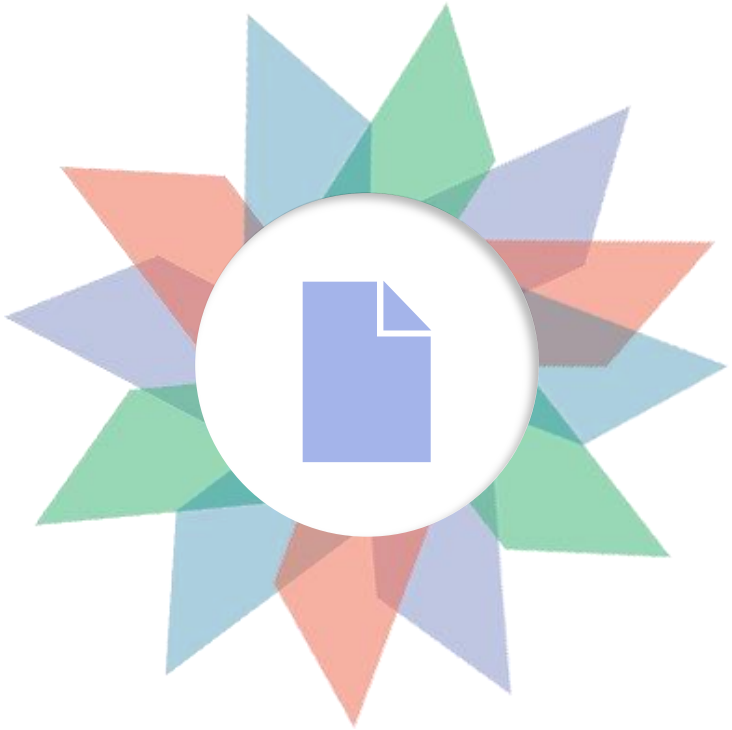
LG AI Research

온라인 채널 제품 판매량 예측 AI 온라인 해커톤

김경호, 김정원, 윤서환, 정민우

Contents

LG AI Research



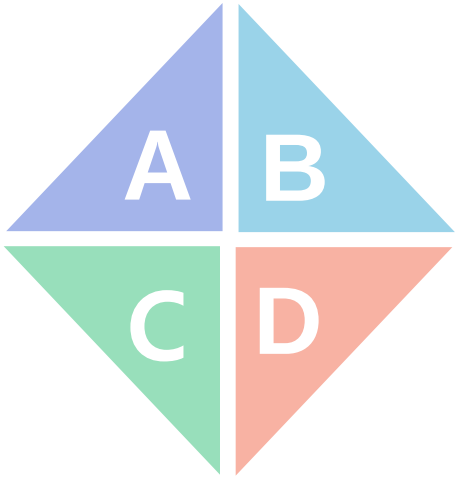
데이터 분석

1-1 구성

데이터셋 구성 - train, sales, brand_keyword_cnt, product_info

train

제품 코드, 대분류, 중분류, 소분류, 브랜드,
쇼핑몰 2022.01.01~2023.04.24 의
실제 일 별 판매량



sales

제품 코드, 대분류, 중분류, 소분류,
브랜드, 쇼핑몰 실제 일 별 총 판매 금액

brand_keyword_cnt

2022.01.01~2023.04.24
브랜드의 연관 키워드 언급량을
정규화한 데이터

product_info

제품 코드, 제품 특성 데이터

1-2 판매량

기간 - 2022.01.01 ~ 2023.04.24

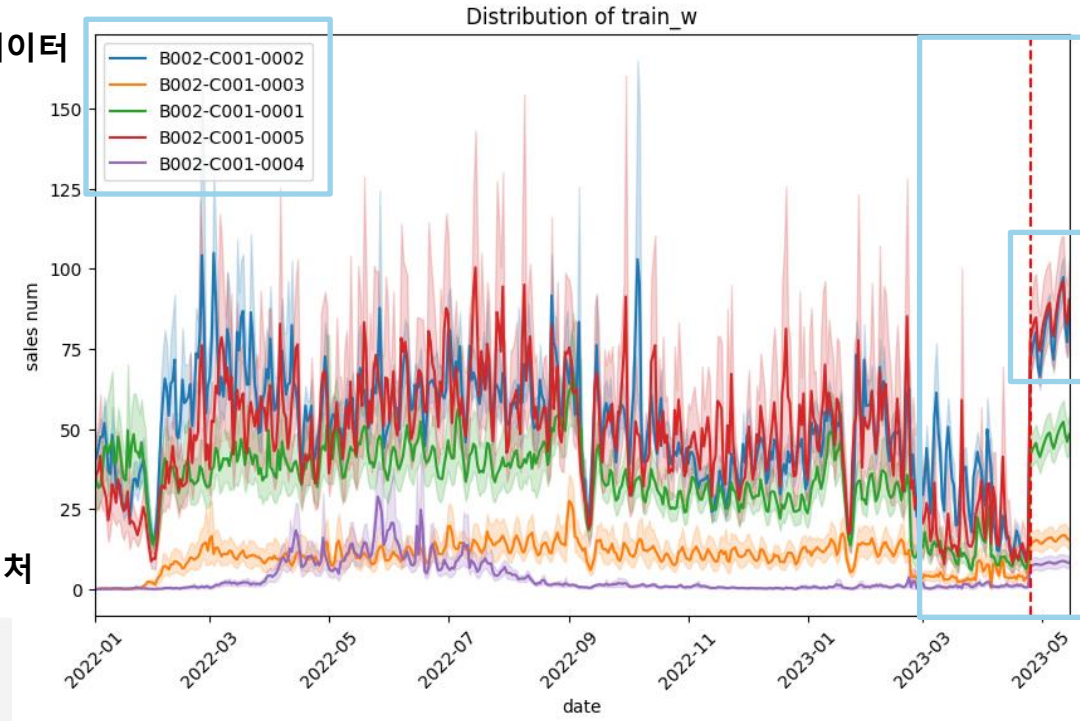
- ☑ 품목 수 - **28894개**
- ☑ 일 수 - **479일**
- ☑ 일자 별 판매량을 나열한 데이터

	제품	대분류	중분류	소분류	브랜드	쇼핑몰	2022-01-01	2022-01-02	2022-01-03	2022-01-04	...	2023-04-15	2023-04-16	2023-04-17	2023-04-18	2023-04-19	2023-04-20	2023-04-21	2023-04-22	2023-04-23	2023-04-24
0	B002-00001-00001	B002-C001-0002	B002-C002-0007	B002-C003-0038	B002-00001	S001-00001	0	0	0	0	...	0	0	0	0	0	0	0	0	0	0
1	B002-00002-00001	B002-C001-0003	B002-C002-0008	B002-C003-0044	B002-00002	S001-00001	0	0	0	0	...	2	0	2	0	2	2	1	0	0	0
2	B002-00002-00002	B002-C001-0003	B002-C002-0008	B002-C003-0044	B002-00002	S001-00001	0	0	0	0	...	0	0	0	0	0	0	0	0	0	0

1-2 판매량

1 다양한 메타 데이터
- 모델 TFT

2 그 밖의 다양한 피처
- 앙상블

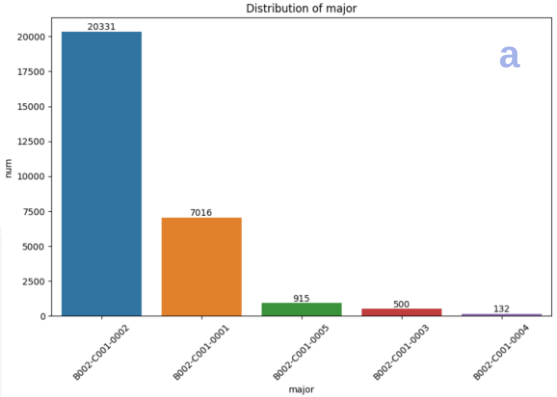


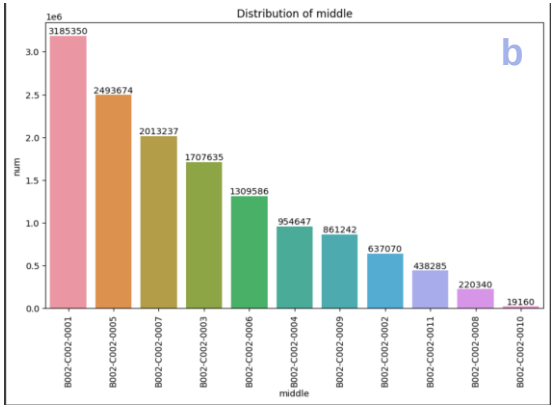
3 Metrics과 Loss와의 관계
- 판매 비중 높은 값을 맞추는 것이 중요

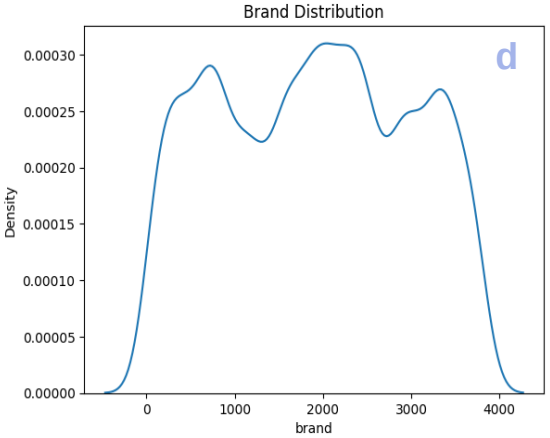
4 모델 검증
- 정성분석과 사후분석

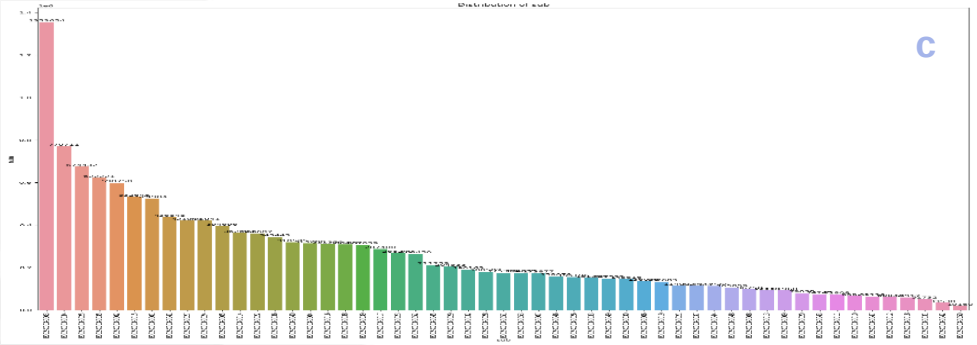
1-2 판매량

분류 별 개수 분포도









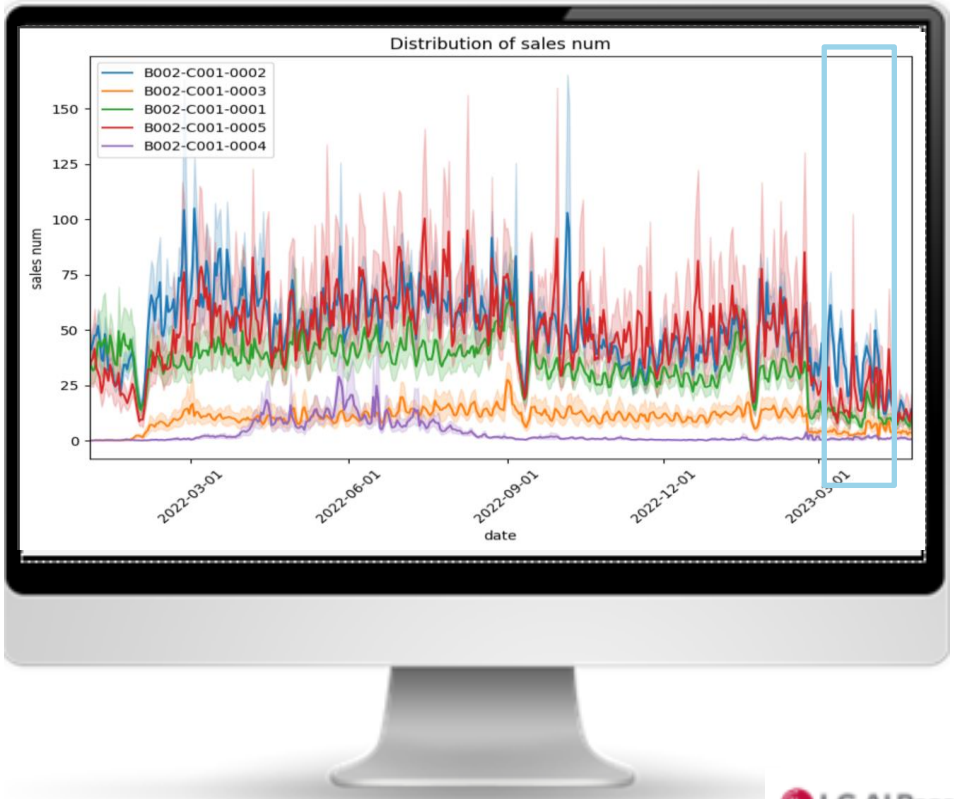
a. 대분류 5 개
b. 중분류 11 개
c. 소분류 53 개
d. 브랜드 2895 개

1-2 판매량

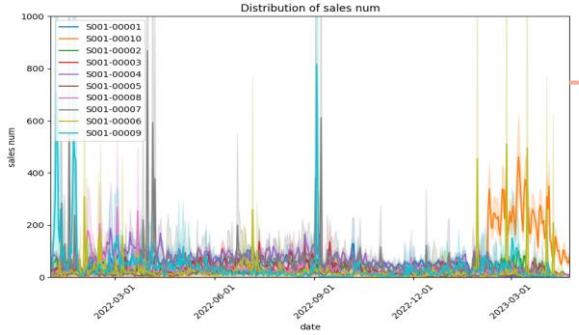
대분류 별 판매량 분포도

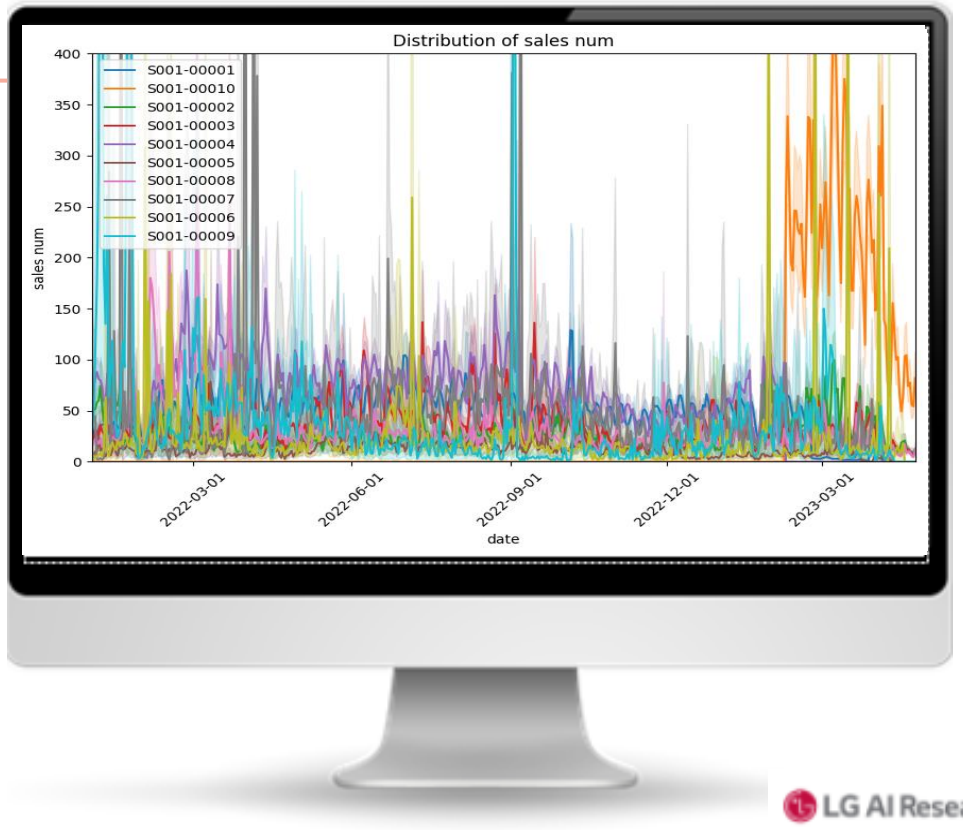
1, 2, 5번 대분류 판매량이 상대적으로 높음

23년 3월 전후로 전반적인 **판매량 급락**



1-2 판매량





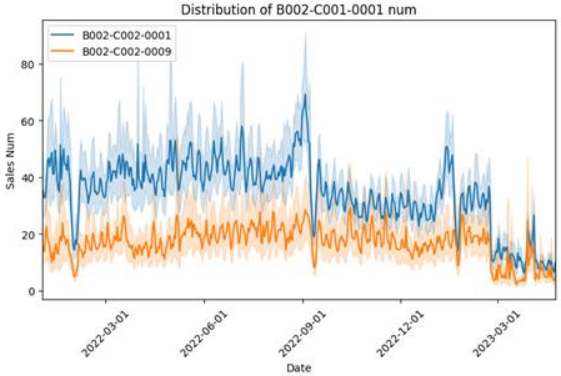
쇼핑몰 별 판매량 분포도

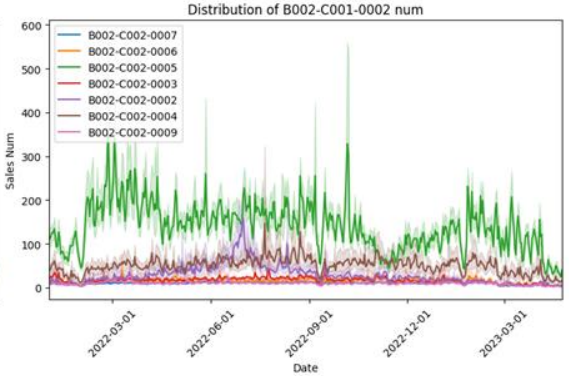
9번 쇼핑몰의 특정 기간 판매량 급증

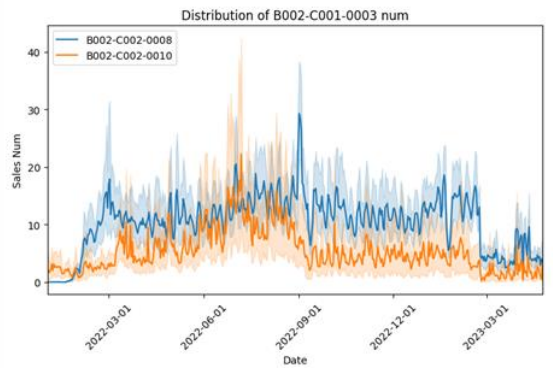
전체적으로 쇼핑몰의 판매량은 대부분 200 이하

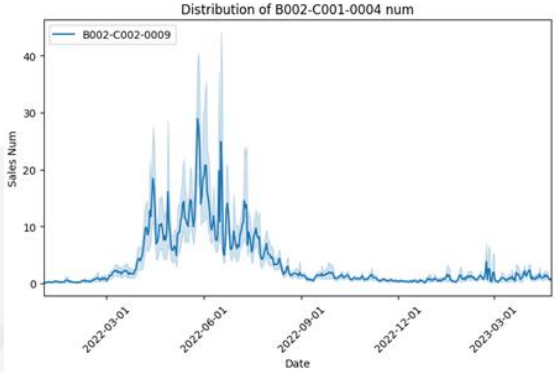
1-2 판매량

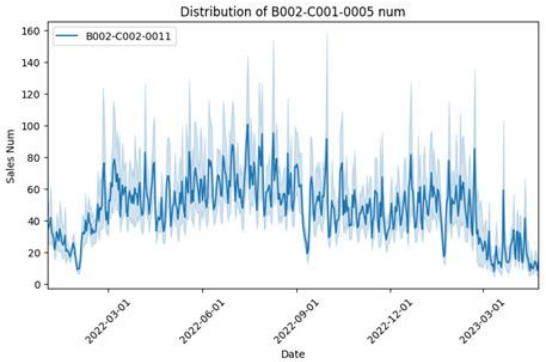
대분류 별 중분류 판매량











LG AI Research

1-3 매출

기간 - 2022.01.01 ~ 2023.04.24

- ☑ 품목 수 - **28894개**
- ☑ 일 수 – **479일**
- ☑ 일자 별 매출을 나열한 데이터

	제품	대분류	중분류	소분류	브랜드	쇼핑몰	2022-01-01	2022-01-02	2022-01-03	2022-01-04	...	2023-04-15	2023-04-16	2023-04-17	2023-04-18	2023-04-19	2023-04-20	2023-04-21	2023-04-22	2023-04-23	2023-04-24
0	B002-00001-00001	B002-C001-0002	B002-C002-0007	B002-C003-0038	B002-00001	S001-00001	0	0	0	0	...	0	0	0	0	0	0	0	0	0	0
1	B002-00002-00001	B002-C001-0003	B002-C002-0008	B002-C003-0044	B002-00002	S001-00001	0	0	0	0	...	44800	0	44800	0	44800	44800	22400	0	0	0
2	B002-00002-00002	B002-C001-0003	B002-C002-0008	B002-C003-0044	B002-00002	S001-00001	0	0	0	0	...	0	0	0	0	0	0	0	0	0	0

추가한 column

판매 가격 = 판매 매출 / 판매량

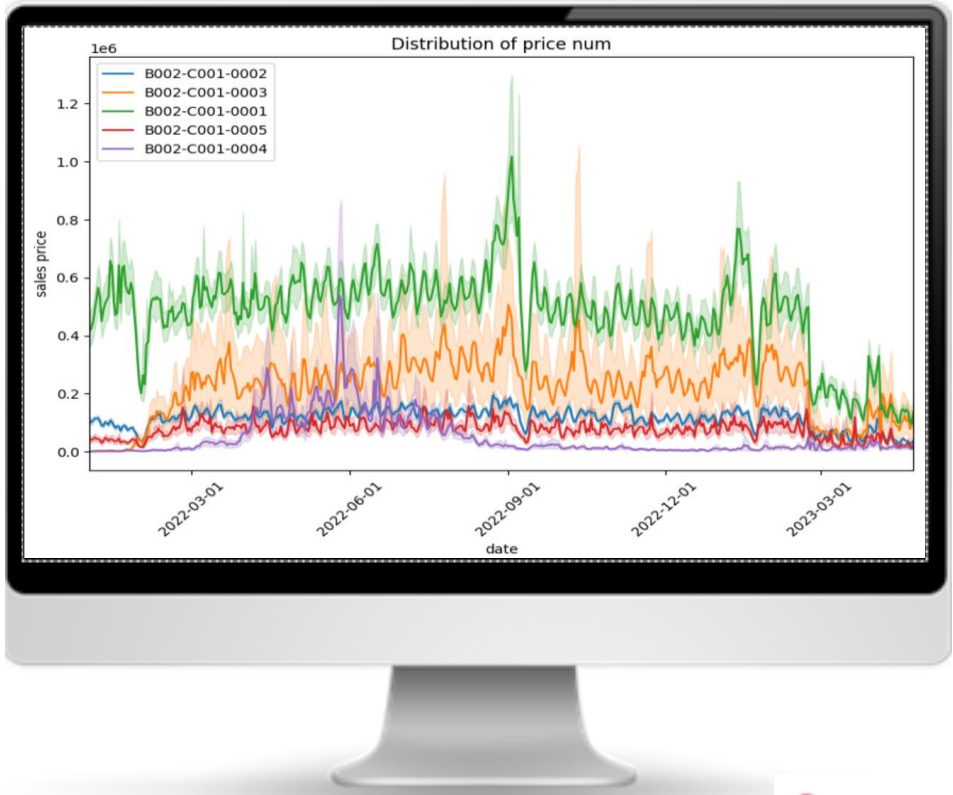
1-3 매출

대분류 별 가격 분포

1, 3번이 가격이 높음
- 낮은 판매량 -> 높은 가격

23년 3월 전후로 전반적인 판매량 급락

2,5번 대분류의 가격은 대체로 일정



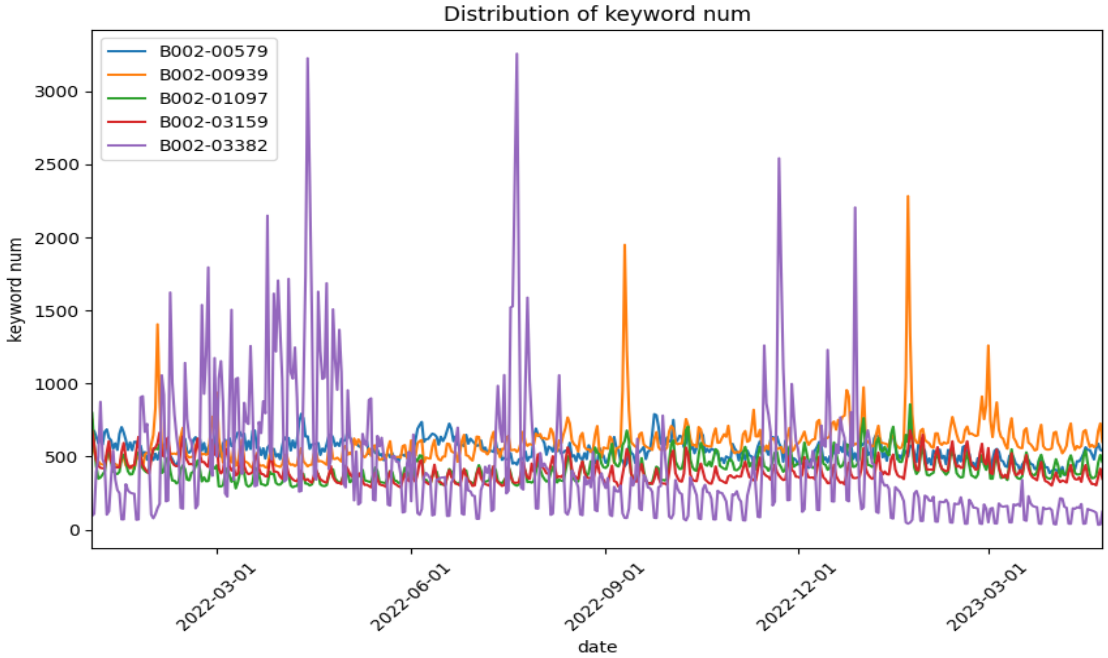
1-4 브랜드 검색 데이터

기간 - 2022.01.01 ~ 2023.04.24

- ☑ 브랜드 수 – **2895 개**
- ☑ 일 수 - **479일**
- ☑ 일자 별 브랜드 언급량을 정규화한 데이터

	브랜드	2022-01-01	2022-01-02	2022-01-03	2022-01-04	2022-01-05	2022-01-06	2022-01-07	2022-01-08	2022-01-09	...	2023-03-26	2023-03-27	2023-03-28	2023-03-29	2023-03-30	2023-03-31	2023-04-01	2023-04-02	2023-04-03	2023-04-04
0	B002-00001	0.84131	0.91383	1.450530	2.422390	1.871190	1.581080	1.232950	1.174930	1.145920	...	0.319110	0.391640	0.377130	0.49318	0.072520	0.29010	0.31911	0.232080	0.333620	0.44966
1	B002-00002	12.64868	20.27850	15.332170	12.750210	13.562510	13.707570	11.937910	15.564250	14.084710	...	10.269790	11.966920	10.646930	10.41485	10.487380	9.48651	9.28343	10.429350	11.154620	11.38671
2	B002-00003	0.33362	0.43516	0.362630	0.174060	0.217580	0.464170	0.420650	0.290100	0.377130	...	0.536690	0.696250	0.449660	0.39164	1.029880	0.49318	0.91383	0.797790	1.015370	0.88482
3	B002-00005	1.07339	1.71163	2.016240	1.914700	1.987230	2.146790	1.682620	1.378000	1.421520	...	2.219320	2.509420	2.872060	2.37888	2.030750	1.53756	1.34899	1.261960	2.320850	2.30635
4	B002-00006	0.00000	0.00000	0.188558	0.246574	0.246574	0.246574	0.377139	0.087012	0.261084	...	0.072526	0.290103	0.087012	0.00000	0.130542	0.00000	0.00000	0.072526	0.217577	0.00000

1-4 브랜드 검색 데이터



대부분의 브랜드는 keyword 0에 가까움 / 상위 5개 브랜드의 키워드 분포

1-5 제품 정보

	제품	제품특성
0	B002-03509-00001	제품유형:일반식품 콜라겐 펩타이드:1000mg 종류:어류 분자량:300Da 섭취대상...
1	B002-02376-00001	700mg x 28정
2	B002-03735-00001	제품타입:정 섭취방법:물과 함께 섭취대상:성인남녀 섭취횟수:하루 두 번 1일 총 섭...
3	B002-03735-00003	제품타입:정 섭취방법:물과 함께 섭취대상:성인남녀 섭취횟수:하루 한 번 1일 총 섭...
4	B002-02769-00001	HCA:900mg 영양소 원료명(식약처고시):비타민B1 영양소 원료명(식약처고시):...
...	...	...
12773	B002-01994-00001	형태:액상형 구성:리필 등급:1종 용도:식기 용도:과일 용도:야채 용도:조리기구 용...
12774	B002-02600-00002	헤어타입:모든 모발용 주요제품특징:머릿결개선 세부제품특징:촉촉함(수분공급) 세부제품...
12775	B002-02372-00095	사용대상:여성용 타입:일자형 흡수량:20ml :1팩 32개
12776	B002-01318-00002	피부타입:모든피부용 향계열:무향 주요제품특징:촉촉함(수분공급) 주요제품특징:풍부한 ...
12777	B002-02473-00064	최소연령:4개월 단계별:2단계 종류:일반분유 제품형태:분말 포장형태:캔 용량:800...

```
# 교집합 구하기
intersection = set(train_product).intersection(set(product_only))

# 교집합의 개수 출력
print(len(intersection))
```

```
9158
```

```
# 차집합 구하기
train_diff = set(train_product) - set(product_only)
product_diff = set(product_only) - set(train_product)

# 차집합의 개수 출력
print('train의 갯수:', len(train_diff))
print('product 갯수:', len(product_diff))
```

```
train의 갯수: 4738
product 갯수: 3620
```

제품 수 – train data와 product_info의 교집합 데이터 **10442개의 교집합 데이터**

1-5 제품 정보

```
Topic 8: 0.053*"어류" + 0.036*"실온보관" + 0.032*"섭취량:1포" + 0.031*"비타민C" + 0.030*"끓는물"
WARNING:gensim.models.ldamodel:too few updates, training might not converge; consider increasing the number of
--- 10 Topics ---
Topic 0: 0.071*"감소" + 0.071*"체지방" + 0.032*"비타민B6" + 0.032*"비타민B1" + 0.031*"판토텐산"
Topic 1: 0.048*"실온보관" + 0.037*"끓는물" + 0.036*"전자레인지" + 0.032*":2분" + 0.027*"분말"
Topic 2: 0.063*"실온보관" + 0.057*"전자레인지" + 0.038*"끓는물" + 0.034*":2분" + 0.026*":10분"
Topic 3: 0.040*"실온보관" + 0.037*"전자레인지" + 0.030*":2분" + 0.029*"끓는물" + 0.025*"음용"
Topic 4: 0.219*"28정" + 0.219*"700mg" + 0.028*"어류" + 0.019*"비오틴" + 0.017*"비타민C"
Topic 5: 0.049*"어류" + 0.047*"섭취량:1포" + 0.045*"액상" + 0.045*"음용" + 0.034*"구미/젤리"
Topic 6: 0.064*"분말" + 0.048*"비타민C" + 0.044*"어류" + 0.036*"비오틴" + 0.030*"식이섬유"
Topic 7: 0.082*"체지방" + 0.080*"감소" + 0.050*"비타민C" + 0.040*"섭취량:3정" + 0.038*"판토텐산"
Topic 8: 0.052*"분말" + 0.046*"음용" + 0.045*"섭취량:1포" + 0.036*"어류" + 0.025*"비타민C"
Topic 9: 0.053*"섞어서" + 0.034*"700mg" + 0.033*"28정" + 0.029*"분말" + 0.027*"섭취량:1스푼"
------------------------1번 제품------------------------
Number of Topics: 1, Coherence Score: (1, 0.552607190124505), Perplexity Score: (1, -5.461715268960762)
Number of Topics: 2, Coherence Score: (2, 0.5254586533444439), Perplexity Score: (2, -5.378511294991995)
Number of Topics: 3, Coherence Score: (3, 0.5142801243967274), Perplexity Score: (3, -5.387834723549957 5)
Number of Topics: 4, Coherence Score: (4, 0.4734137762656722 7), Perplexity Score: (4, -5.303993258423717)
Number of Topics: 5, Coherence Score: (5, 0.5505021688668899), Perplexity Score: (5, -5.1737044378473085)
Number of Topics: 6, Coherence Score: (6, 0.5140663158810131), Perplexity Score: (6, -5.173757789014237)
Number of Topics: 7, Coherence Score: (7, 0.5015699665515055), Perplexity Score: (7, -5.161240677659096)
Number of Topics: 8, Coherence Score: (8, 0.516355278540546), Perplexity Score: (8, -5.2812663536823505)
Number of Topics: 9, Coherence Score: (9, 0.4885069512325352 4), Perplexity Score: (9, -5.224046324640365)
Number of Topics: 10, Coherence Score: (10, 0.4609632649712202 5), Perplexity Score: (10, -5.153062992996711)
Best Topics for Coherence: [1, 5, 2]
Best Topics for Perplexity: 5
```

1. Text processing 처리
   → **word tokenizer 및 regex**
2. Stop word 처리 후 **소분류 별 LDA**
3. **1 ~ 10개의 토픽** 범위 설정
4. **Coherence** 상위 3개 중
   perplexity가 가장 낮은 토픽 수로 분류
5. 소분류별 적절 토픽수로 **split → labeling**
   - 280개로 라벨링(다시 분류화함)

1-5 제품 정보

최종 결과 280개의 label 분류

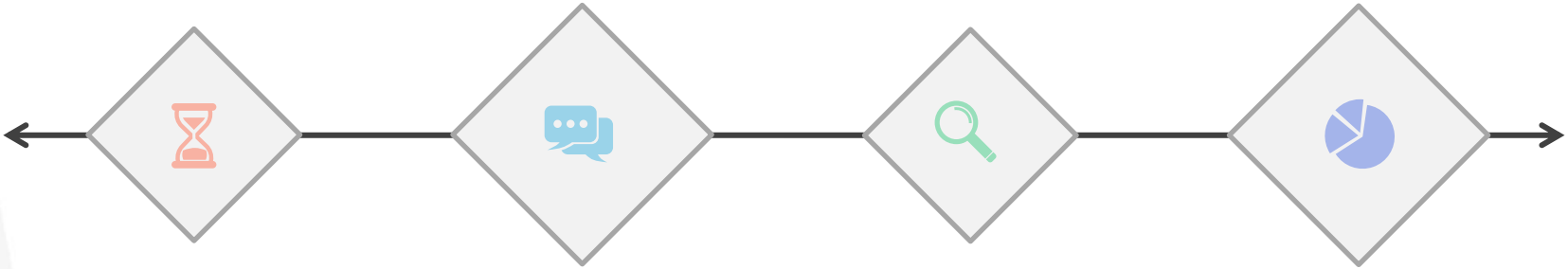
	제품	대분류	중분류	소분류	브랜드	쇼핑몰	제품특성	label
0	B002-00008-00001	B002-C001-0001	B002-C002-0001	B002-C003-0001	B002-00008	S001-00001	['주요', '기능성', '식약처인증', '장건강', '영양소', '원료명', '식...	4
1	B002-00035-00001	B002-C001-0001	B002-C002-0001	B002-C003-0001	B002-00035	S001-00004	['주요소재', '신발', '부가기능발목높이', '폴리에스테르', '상의종류소매기장...	4
2	B002-00037-00001	B002-C001-0001	B002-C002-0001	B002-C003-0001	B002-00037	S001-00001	['700mg', 'x', '28정']	1
3	B002-00043-00001	B002-C001-0001	B002-C002-0001	B002-C003-0001	B002-00043	S001-00003	['제품타입', '캡슐', '섭취방법', '물과', '함께', '섭취대상', '성인...	2
4	B002-00043-00001	B002-C001-0001	B002-C002-0001	B002-C003-0001	B002-00043	S001-00004	['제품타입', '캡슐', '섭취방법', '물과', '함께', '섭취대상', '성인...	2
...	...	...	...	...	...	...	...	...
28889	B002-03528-00004	B002-C001-0005	B002-C002-0011	B002-C003-0053	B002-03528	S001-00001	['최소연령:12개월', '종류', '진밥', '보관방법', '냉장보관', ':6개...	276
28890	B002-03528-00010	B002-C001-0005	B002-C002-0011	B002-C003-0053	B002-03528	S001-00001	['최소연령:12개월', '종류', '진밥', '보관방법', '냉장보관', ':6개...	276
28891	B002-03528-00012	B002-C001-0005	B002-C002-0011	B002-C003-0053	B002-03528	S001-00001	['최소연령:12개월', '종류', '진밥', '보관방법', '냉장보관', ':6개...	276
28892	B002-03709-00001	B002-C001-0005	B002-C002-0011	B002-C003-0053	B002-03709	S001-00001	['최소연령', '기타', '종류', '다짐채소', '알레르기', '유발성분', '...	272
28893	B002-03709-00002	B002-C001-0005	B002-C002-0011	B002-C003-0053	B002-03709	S001-00001	['최소연령', '기타', '종류', '쌀가루', '알레르기', '유발성분', '유...	271

28894 rows × 8 columns

문제 정의

2-1 주요 과제

2022.1.1 ~ 2023.4.24. 데이터를 통해 **2023.4.25. ~ 2023.5.15.** 판매량 예측



01 모델 선정

02 Feature Engineering 및 데이터셋 구성

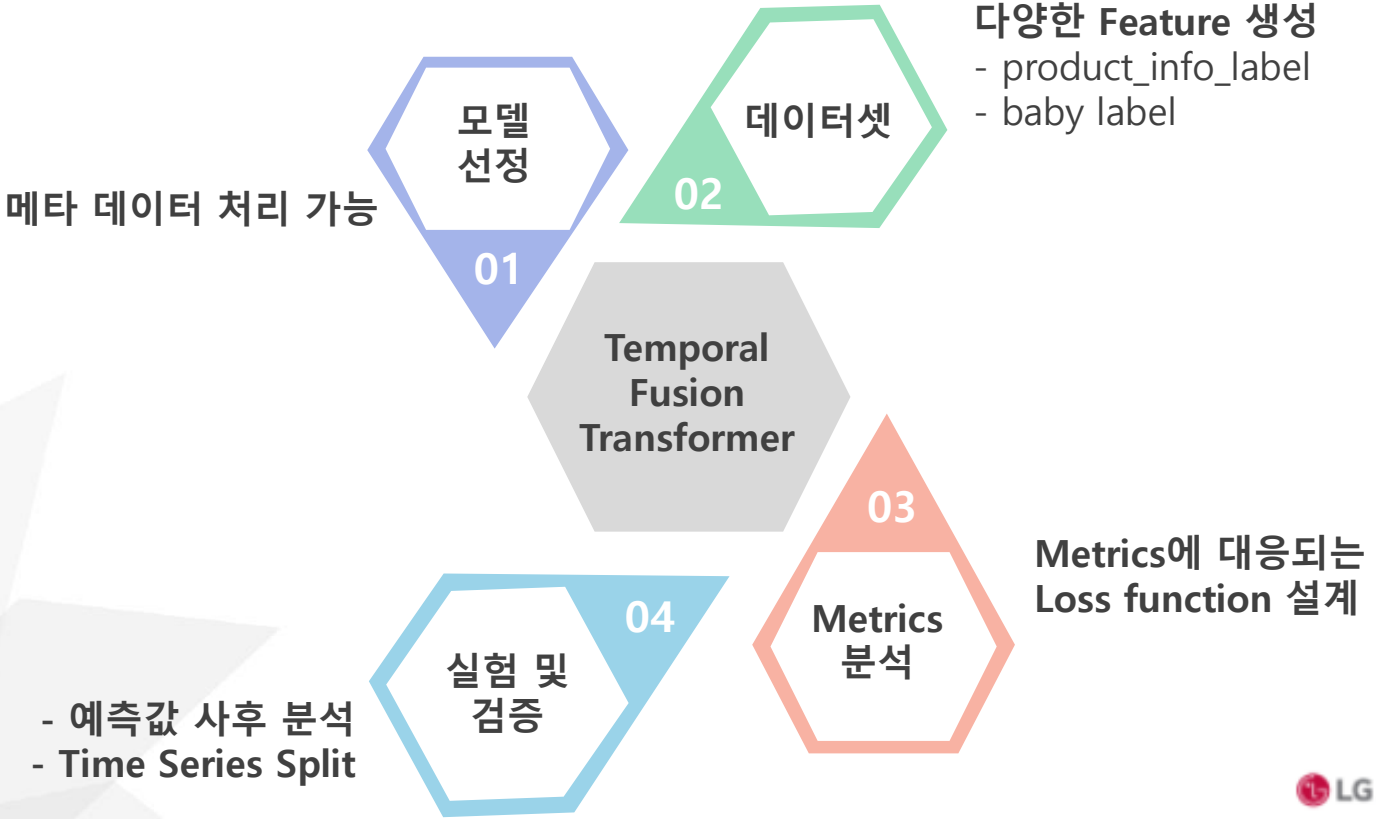
03 Customize Loss Function

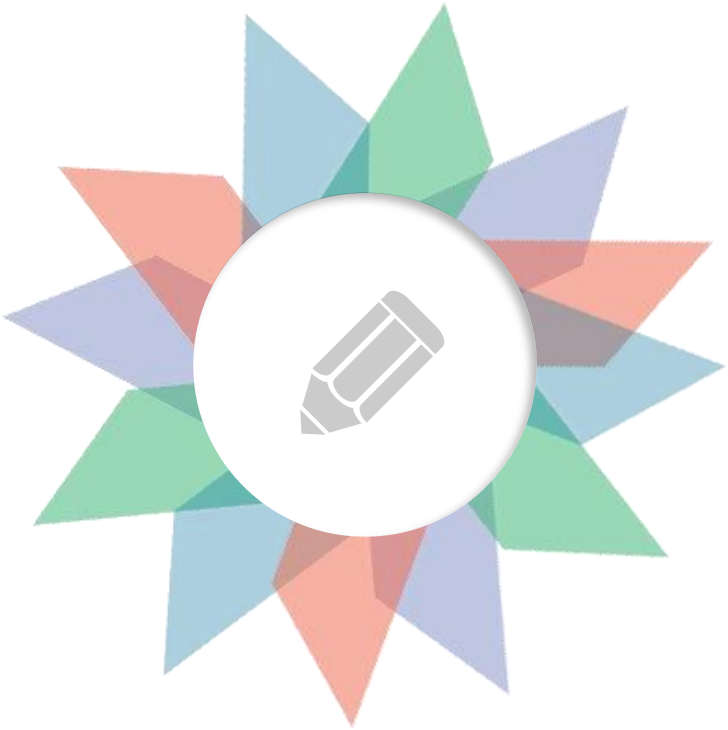
04 실험 및 검증

주안점

- 다양한 feature와 메타 데이터가 공존
- PSFA라는 metrics의 특성

2-1 주요 과제


모델 선정
01
메타 데이터 처리 가능
데이터셋
02
다양한 Feature 생성
- product_info_label
- baby label
Temporal Fusion Transformer
03
Metrics 분석
Metrics에 대응되는 Loss function 설계
04
실험 및 검증
- 예측값 사후 분석
- Time Series Split

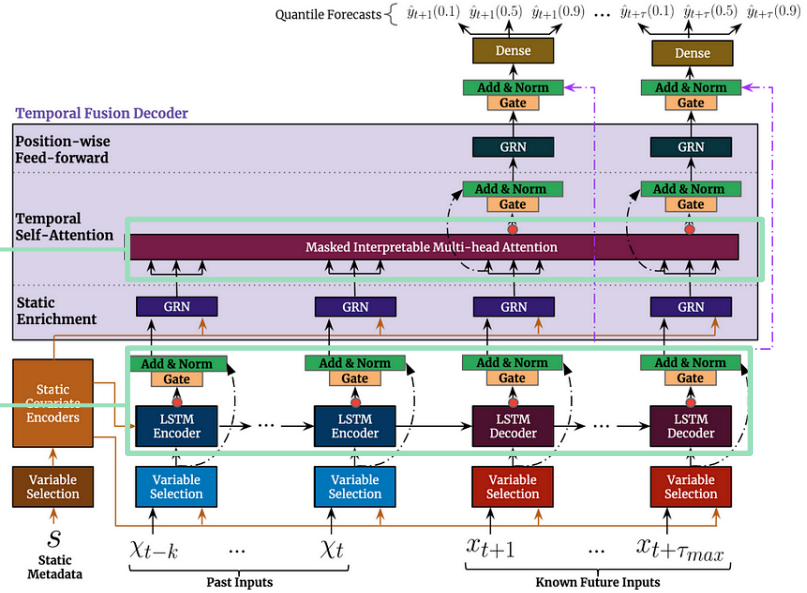


모델 선정

3-1 Architecture

Transformer 기반
시계열 데이터에 Attention 메커니즘을 적용

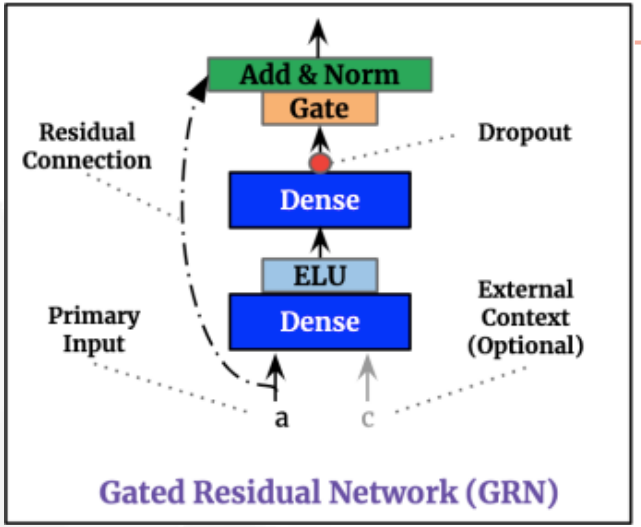
LSTM 기반
Encoder를 통해 시계열 데이터에 맥락 부여
Decoder를 통해 맥락과 미래 정보를 반영해 예측

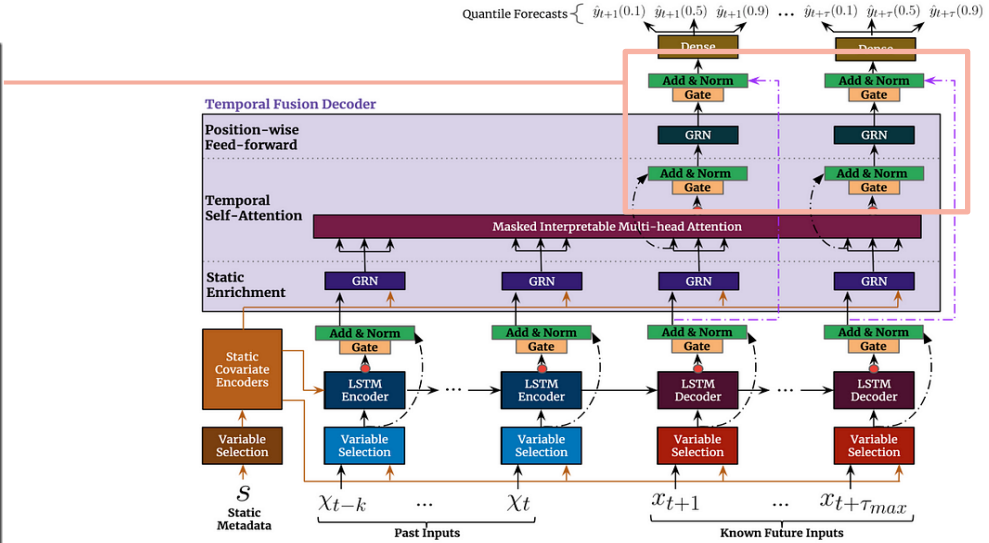


[Temporal FusionTransformers for Interpretable Multi-horizon TimeSeries Forecasting □□ □□]

3-2 세부 특징

1) GRN 구조



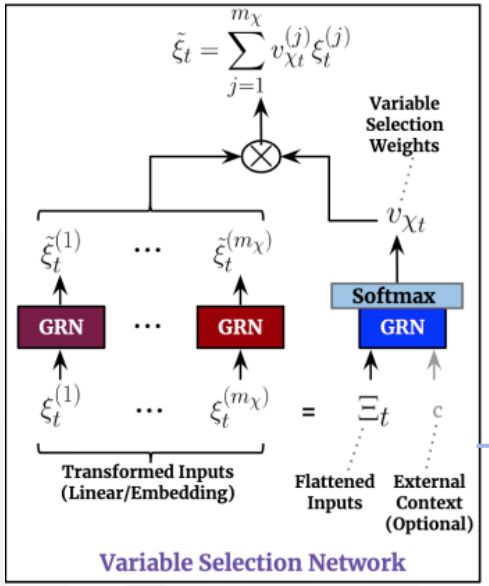


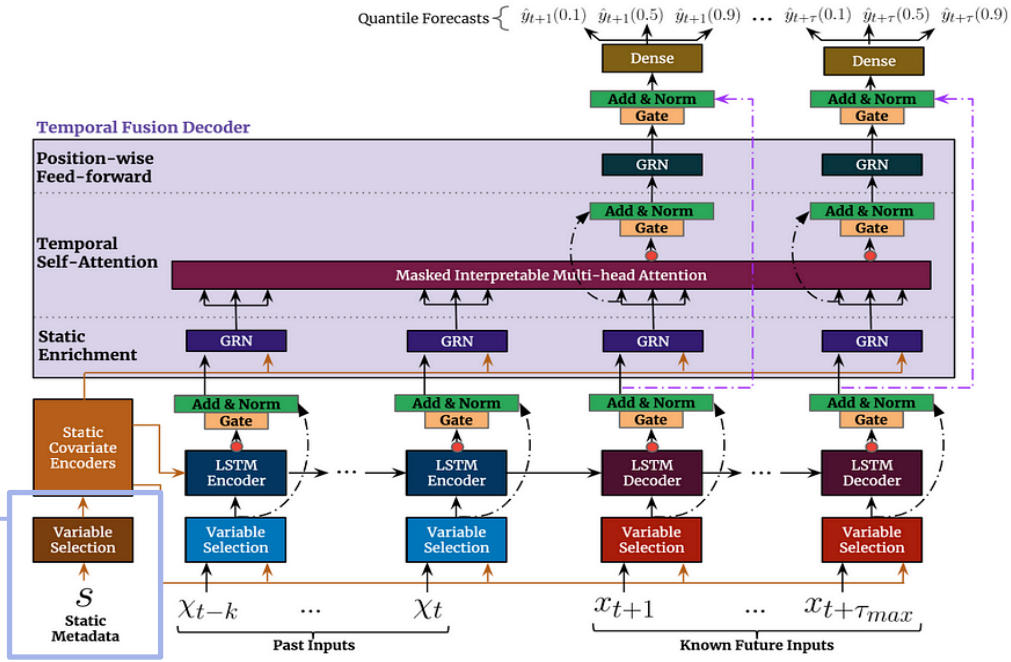
[Temporal FusionTransformers for Interpretable Multi-horizon TimeSeries Forecasting □□ □□]

특징 - Residual Connection이 포함된 최종적인 Dense Layer

3-2 세부 특징

2) VSN 구조





[Temporal FusionTransformers for Interpretable Multi-horizon TimeSeries Forecasting □□ □□]

특징 - Attention 기반으로 영향력 있는 변수를 골라내는 과정

3-3 모델 특징

다양한 데이터 입력 가능

현재까지 알려진 시계열 데이터 (판매량, 매출, 브랜드 키워드 언급량)
외부 범주형 / 정적 변수 (대분류, 중분류, 소분류, 브랜드, 쇼핑몰)

Multi Horizon 예측 가능

2023.4.25. ~ 2023.5.15. 기간 (21일) 예측 가능

	time_idx	product_nums	major	middle	sub	brand	shop	week_weekend	special_day	keyword_cnt	sales_rate	sales
0	0	0	B002-C001-0002	B002-C002-0007	B002-C003-0038	B002-00001	S001-00001	1	1	0.84131	0.0	0.0
1	0	1	B002-C001-0003	B002-C002-0008	B002-C003-0044	B002-00002	S001-00001	1	1	12.64868	0.0	0.0
2	0	2	B002-C001-0003	B002-C002-0008	B002-C003-0044	B002-00002	S001-00001	1	1	12.64868	0.0	0.0

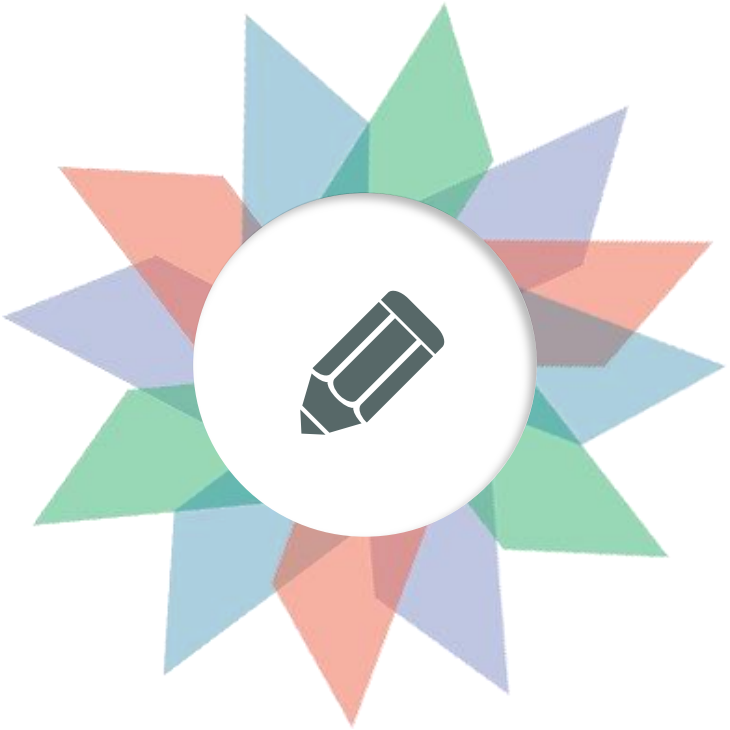
데이터셋 구성 이미지

3-3 사용한 컬럼

time_idx, product_nums, major, middle, Sub, Brand, Shop, special_day, keyword_cnt, sales_rate, Sales, baby

일자 indexing, 제품 ID, 대분류, 중분류, 소분류, 브랜드, 쇼핑몰, 공휴일 여부, 브랜드의 연관 키워드 언급량, 판매량, 유아용품 유무

	time	month	time_idx	week_weekend	day	special_day	product_nums	product	major	middle	...	product_info	product_info_label	sales_rate_log	date_cos	date_sin	day_cos	day_sin	month_cos	month_sin	weight
0	2022-01-01	01	0	1	5	1	0	B002-00001-00001	B002-C001-0002	B002-C002-0007	...	헤어타입:모든 모발용 제품형태:스프레이형 주요제품특징:머릿결개선 주요제품특징:흡수력...	0	0.0	1.0	0.0	1.0	0.0	1.0	0.0	0.296359
1	2022-01-01	01	0	1	5	1	1	B002-00002-00001	B002-C001-0003	B002-C002-0008	...		0	0.0	1.0	0.0	1.0	0.0	1.0	0.0	0.982695
2	2022-01-01	01	0	1	5	1	2	B002-00002-00002	B002-C001-0003	B002-C002-0008	...		0	0.0	1.0	0.0	1.0	0.0	1.0	0.0	0.982695



PSFA 분석

4-1 PSFA 특징

PSFA 수식

SMAPE와 유사

대분류 별 Pseudo 예측 정확도: $$PSFA_m = 1 - \frac{1}{n}\sum_{day=1}^{n}\sum_{i=1}^{N}\left(\underbrace{\frac{|y_i^{day} - p_i^{day}|}{\max(y_i^{day}, p_i^{day})}}_{\text{오차}} \times \underbrace{\frac{y_i^{day}}{\sum_{i=1}^{N} y_i^{day}}}_{\text{(판매)비중}}\right)$$

전체 Pseudo 예측 정확도: $$PSFA = \frac{1}{M}\sum_{m=1}^{M} PSFA_m$$

대분류 별 PSFA 산술평균

높은 값을 예측하는 것이 중요.
낮은 값은 상대적으로 덜 중요.

4-1 PSFA 특징

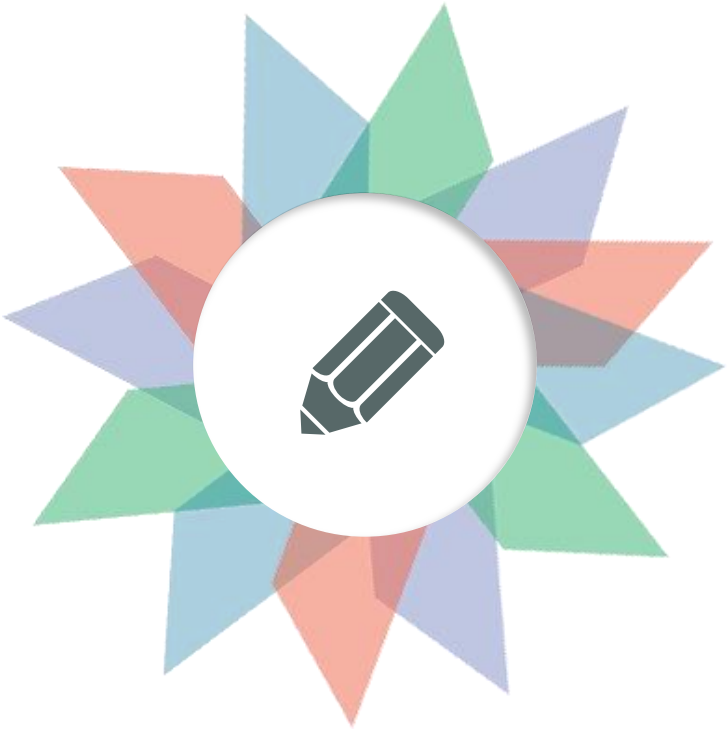
PSFA Code

```python
class PSFA_Row(MultiHorizonMetric):
    def loss(self, y_pred, target):
        y_pred = self.to_prediction(y_pred)
        diff_value = torch.abs(target - y_pred)
        max_value = torch.max(target, y_pred) + 1e-8
        weight_denumerator = torch.sum(target, axis=1).view(y_pred.shape[0], 1) + 1e-8
        weight = target / weight_denumerator
        loss = ((diff_value / max_value) * weight) * (y_pred.shape[1])
        return loss
```

가로 기준 비중

```python
class PSFA_Column(MultiHorizonMetric):
    def loss(self, y_pred, target):
        y_pred = self.to_prediction(y_pred)
        diff_value = torch.abs(target - y_pred)
        max_value = torch.max(target, y_pred) + 1e-8
        weight_denumerator = torch.sum(target, axis=0).view(1, y_pred.shape[1]) + 1e-8
        weight = target / weight_denumerator
        loss = ((diff_value / max_value) * weight) * (y_pred.shape[0])
        return loss
```

세로 기준 비중



Ensemble

5-1 Ensemble의 이유

1 다양한 데이터셋

- Basic, product info label, baby label

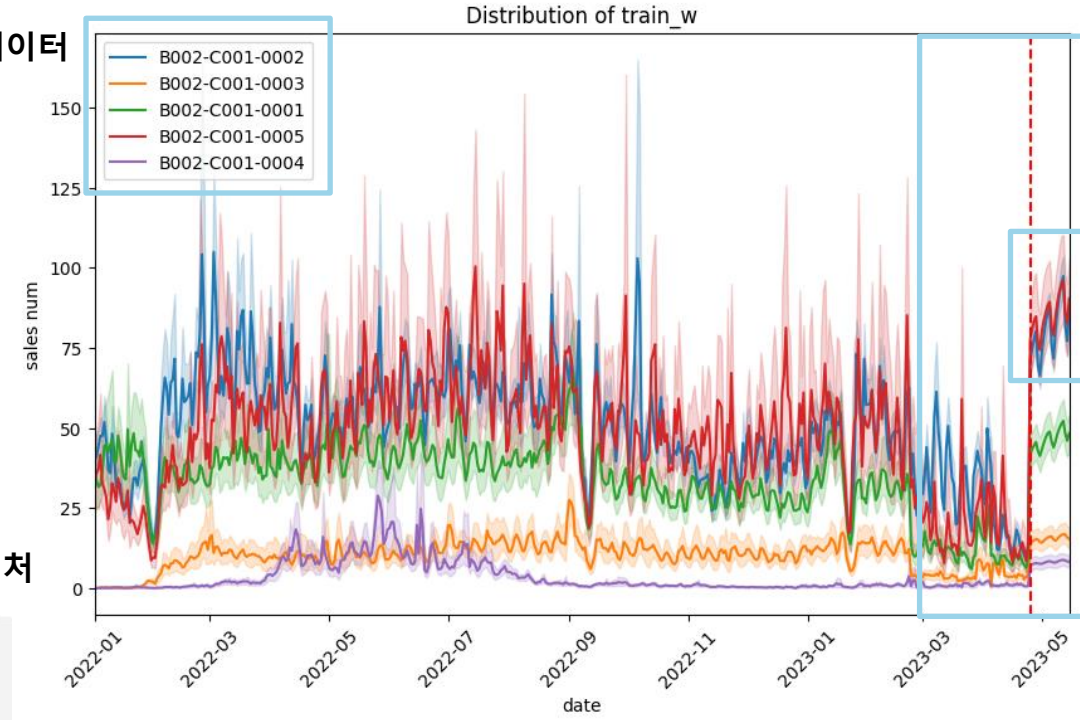
2 다양한 loss 함수

- PSFA_row, PSFA_column

	time	month	time_idx	week_weekend	day	special_day	product_nums	product	major	middle	...	product_info	product_info_label	sales_rate_log	date_cos	date_sin	day_cos	day_sin	month_cos	month_sin	weight
0	2022-01-01	01	0	1	5	1	0	B002-00001-00001	B002-C001-0002	B002-C002-0007	...	헤어타입:모든 모발용 제품형태:스프레이형 주요제품특징:머릿결개선 주요제품특징:흡수력...	0	0.0	1.0	0.0	1.0	0.0	1.0	0.0	0.296359
1	2022-01-01	01	0	1	5	1	1	B002-00002-00001	B002-C001-0003	B002-C002-0008	...		0	0.0	1.0	0.0	1.0	0.0	1.0	0.0	0.982695
2	2022-01-01	01	0	1	5	1	2	B002-00002-00002	B002-C001-0003	B002-C002-0008	...		0	0.0	1.0	0.0	1.0	0.0	1.0	0.0	0.982695

5-1 Ensemble의 이유

1 다양한 메타 데이터
- 모델 TFT



2 그 밖의 다양한 피처
- 앙상블

3 Metrics과 Loss와의 관계
- 판매 비중 높은 값을 맞추는 것이 중요

4 모델 검증
- 정성분석과 사후분석

결과

6-1 결과

모델	Private	Public
Basic		0.567
Product_info_label		0.563
Baby label		0.563
Synchronized		0.571
Ensemble	0.579	0.581

6-2 하이퍼 파라미터

결정 방법 - Optuna 사용

하이퍼 파라미터

Window Size	90
Gradient clip val	0.037
Hidden Size	81
Dropout	0.287
Hidden Continuous Size	12
Attention Head Size	2
Learning Rate	0.006

감사합니다