



LG Academy
DACON

LG Aimers 오프라인 해커톤

온라인 채널 제품 판매량 예측

TEAM | SteelMate

송준희 | 박지현 | 오수민



온라인 채널 제품 판매량 예측

Part1. 데이터 분석

01 데이터 전처리 및 EDA

02 파생변수 생성

Part2. 모델링

01 모델 검증 : Custom Loss Function

02 모델 알고리즘 : 변수 및 파라미터 선정

Part3. 적용 가능성

01 모델의 실용성 및 활용 가능성

02 실제 현업에서의 적용 가능성

목차

Table of Contents

Before Start

데이터 경로

```
└─ 구글 드라이브
  └─ <LG 해커톤 오프라인>
    └─ data
      └─ brand_keyword_cnt.csv # 메타(Meta) 정보
      └─ product_info.csv # 메타(Meta) 정보
      └─ sales.csv # 메타(Meta) 정보
      └─ product_info.csv # 메타(Meta) 정보
      └─ sample_submission.csv # 파일 제출 양식
      └─ train.csv # 기존 학습 데이터
      └─ train_off1.csv # 가격분류 컬럼
      └─ train_off2.csv # 대량판매 컬럼
      └─ train_off3.csv # 주기성 컬럼
      └─ train_off3.csv # Day_Week 컬럼
      └─ train_TopBrand.csv # 상위브랜드 및 카테고리 파생변수 추가
      └─ PT
      └─ LSTM_fc_Model_earlystop.pt # EarlyStopping
      └─ LSTM_fc_Model(best)_Loss.pt # 최적 Loss(Validation)
      └─ LSTM_fc_Model(PSFA)_Loss.pt # 최고 PSFA(Validation)
      └─ submit
      └─ LSTM_fc_best+.csv # 제출 파일
```

“

Part1.

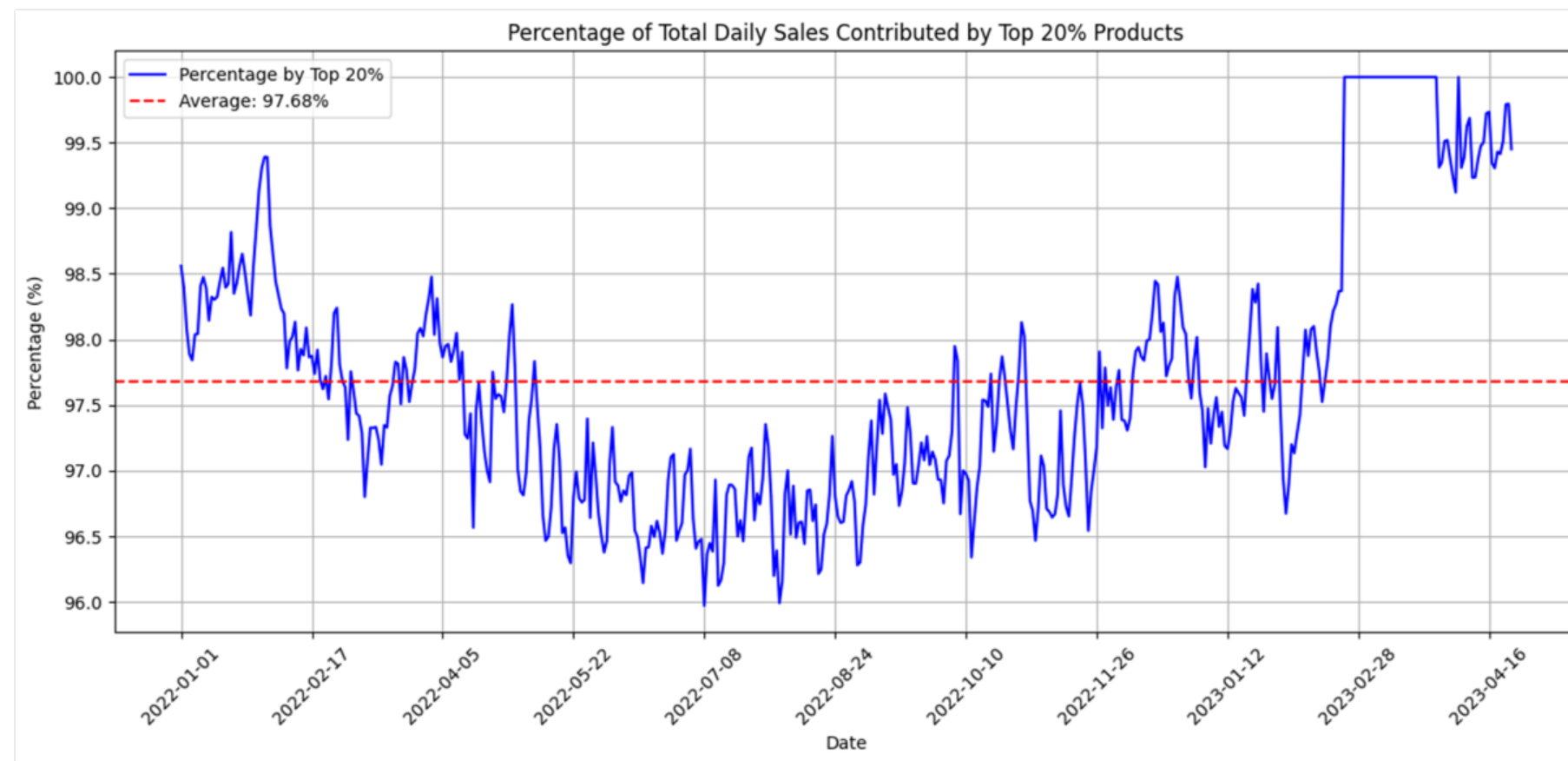
데이터 분석

Contents

01 데이터 전처리 및 EDA

02 파생변수 생성

- 전체 매출의 80%는 전체 제품의 20%가 기여하고 있다.



전체 기간 동안의 누적 판매량을 기준으로,
각 대분류 내에서 상위20%에 속하는 제품이
차지하는 매출 비율을 계산한 결과,
평균 79.10%의 매출을 차지함.

- 대분류 B002-C001-0001: 약 88.64%
- 대분류 B002-C001-0002: 약 91.39%
- 대분류 B002-C001-0003: 약 71.80%
- 대분류 B002-C001-0004: 약 61.75%
- 대분류 B002-C001-0005: 약 81.91%

1. 위의 그래프는 2022-01-1일부터 2023-04-24일까지 상위 20%의 제품이 전체 일별 매출에 차지하는 비율을 나타냄.
2. 빨간색 점선은 해당 기간 동안의 평균 비율(약 97.67)을 표시.

Part1

데이터 전처리 및 EDA

01

Data preprocessing & EDA

- 중분류를 기준으로 각 제품군에 해당하는 카테고리를 생성한다.



▶ 인터넷의 '이마트몰', '엘지 생활건강 공식 홈페이지'를 참고하여 카테고리 속성값을 생성함.

데이터 전처리 및 EDA

- 월별 판매량이 많은 제품들의 종류가 일정하다.

2022-01-31 00:00:00	B002-01069-00002	B002-00113-00001	B002-02920-00006	B002-02723-00004	B002-02463-00007
2022-02-28 00:00:00	B002-02920-00004	B002-02355-00017	B002-02355-00018	B002-02355-00019	B002-02920-00006
2022-03-31 00:00:00	B002-02920-00005	B002-02920-00004	B002-02920-00016	B002-02355-00017	B002-02920-00006
2022-04-30 00:00:00	B002-02920-00005	B002-02920-00004	B002-00113-00001	B002-02920-00016	B002-01069-00002
2022-05-31 00:00:00	B002-00113-00001	B002-02920-00016	B002-01755-00003	B002-01069-00002	B002-02920-00006
2022-06-30 00:00:00	B002-00113-00001	B002-02920-00016	B002-01950-00001	B002-02920-00006	B002-02920-00014
2022-07-31 00:00:00	B002-02920-00014	B002-02052-00019	B002-02920-00016	B002-00113-00001	B002-02920-00006
2022-08-31 00:00:00	B002-02052-00019	B002-02920-00006	B002-03304-00010	B002-02920-00014	B002-01755-00003
2022-09-30 00:00:00	B002-00894-00063	B002-02920-00006	B002-02723-00004	B002-00113-00001	B002-03304-00010
2022-10-31 00:00:00	B002-02920-00005	B002-02920-00004	B002-02920-00006	B002-02920-00025	B002-02723-00004
2022-11-30 00:00:00	B002-00809-00002	B002-02723-00004	B002-00113-00001	B002-02052-00019	B002-03436-00017
2022-12-31 00:00:00	B002-01603-00002	B002-00113-00001	B002-02723-00004	B002-02920-00006	B002-02920-00006
2023-01-31 00:00:00	B002-02355-00010	B002-01603-00002	B002-02920-00005	B002-02723-00004	B002-00113-00001
2023-02-28 00:00:00	B002-02355-00010	B002-01603-00002	B002-00809-00002	B002-02355-00005	B002-02723-00004
2023-03-31 00:00:00	B002-02355-00010	B002-02355-00005	B002-00894-00063	B002-02920-00006	B002-02355-00008

▶ 월별 판매량 Top6 “제품”을 확인

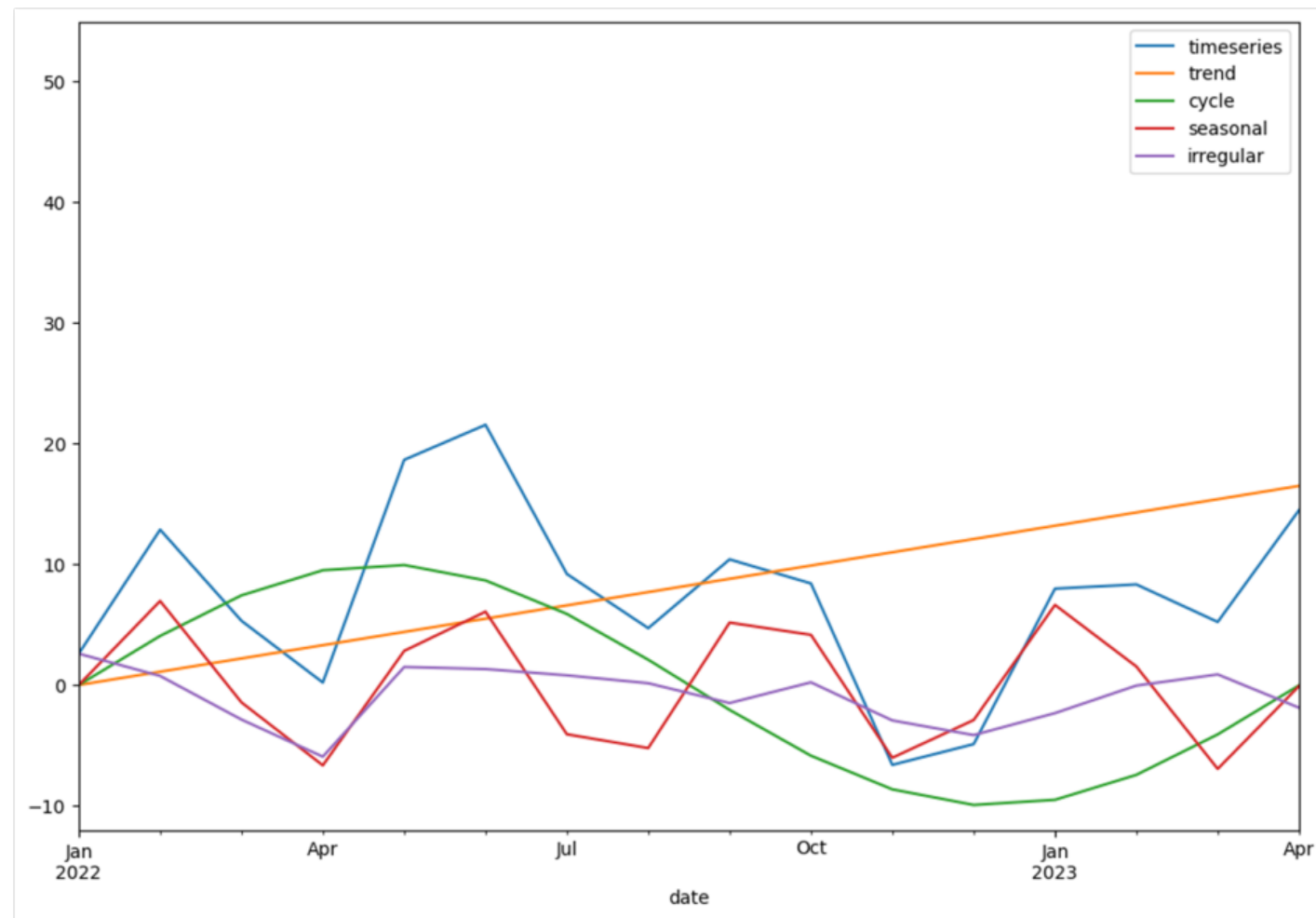
0041, 0003, 0025, 0004, 0001, 0022는 Top 6에 꾸준히 등장함.
제품 특성을 확인해서 어떤 상품인지 확인함.

2022-01-31 00:00:00	B002-C003-0041	B002-C003-0025	B002-C003-0003	B002-C003-0004	B002-C003-0005
2022-02-28 00:00:00	B002-C003-0041	B002-C003-0025	B002-C003-0003	B002-C003-0004	B002-C003-0023
2022-03-31 00:00:00	B002-C003-0041	B002-C003-0025	B002-C003-0003	B002-C003-0004	B002-C003-0052
2022-04-30 00:00:00	B002-C003-0041	B002-C003-0025	B002-C003-0003	B002-C003-0004	B002-C003-0023
2022-05-31 00:00:00	B002-C003-0041	B002-C003-0025	B002-C003-0003	B002-C003-0004	B002-C003-0022
2022-06-30 00:00:00	B002-C003-0041	B002-C003-0025	B002-C003-0003	B002-C003-0004	B002-C003-0011
2022-07-31 00:00:00	B002-C003-0041	B002-C003-0025	B002-C003-0003	B002-C003-0004	B002-C003-0022
2022-08-31 00:00:00	B002-C003-0041	B002-C003-0025	B002-C003-0003	B002-C003-0004	B002-C003-0022
2022-09-30 00:00:00	B002-C003-0041	B002-C003-0025	B002-C003-0003	B002-C003-0004	B002-C003-0022
2022-10-31 00:00:00	B002-C003-0041	B002-C003-0025	B002-C003-0003	B002-C003-0004	B002-C003-0022
2022-11-30 00:00:00	B002-C003-0025	B002-C003-0041	B002-C003-0003	B002-C003-0022	B002-C003-0004
2022-12-31 00:00:00	B002-C003-0041	B002-C003-0025	B002-C003-0003	B002-C003-0004	B002-C003-0022
2023-01-31 00:00:00	B002-C003-0041	B002-C003-0025	B002-C003-0003	B002-C003-0004	B002-C003-0023
2023-02-28 00:00:00	B002-C003-0041	B002-C003-0025	B002-C003-0003	B002-C003-0004	B002-C003-0022
2023-03-31 00:00:00	B002-C003-0041	B002-C003-0025	B002-C003-0003	B002-C003-0022	B002-C003-0023

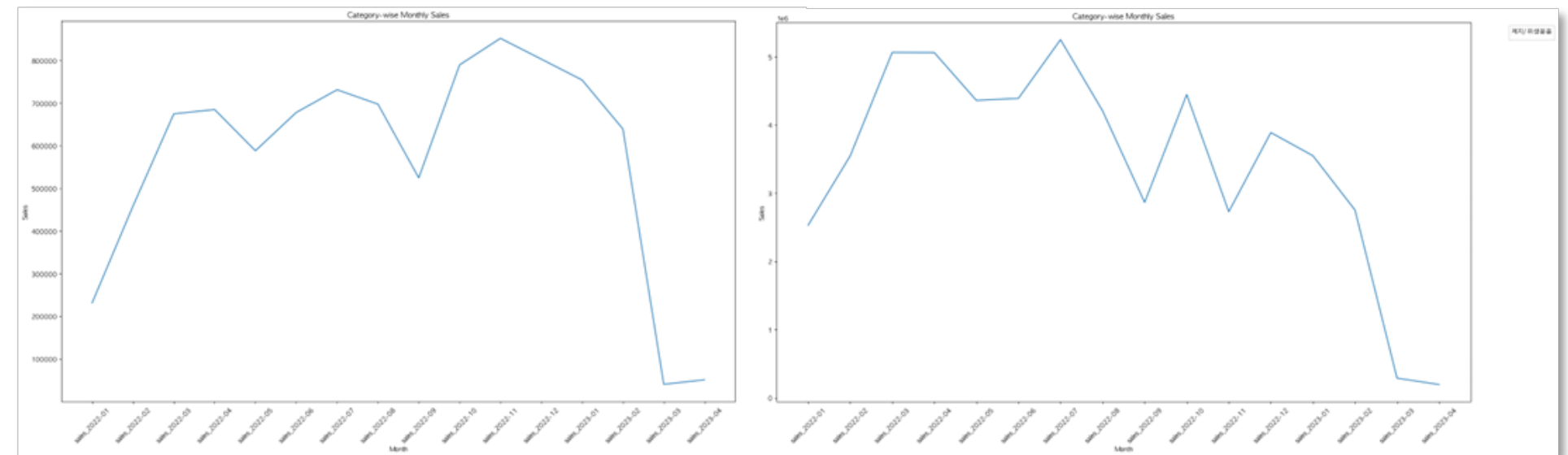
▶ 월별 판매량 Top6 “소분류”를 확인

- 0041 : 유아용 기저귀
- 0003 : 유산균
- 0025 : 물티슈
- 0004 : 단백질보충제
- 0001 : 건강기능식품
- 0022 : 치약

- 제품 판매량의 시계열 패턴을 확인한 결과, 제품 판매에 특정한 주기가 존재한다.



- ▶ 계절성(seasonal)과 불규칙성(irregular)이
전체 시계열에 영향을 미침.



- ▶ 카테고리별 시계열 분석 결과, **고객들의 개별적인 구매 주기 확인 가능.**

- 3개월 : 제지/위생용품
- 4개월 : 욕실용품
- 5개월 : 유아 생활용품, 유아식품, 주방/청소/세탁세제 등
- 1년 : 탈취/방충/살충/제습/방향

* 1년은 특정 계절에 판매량이 높은 카테고리

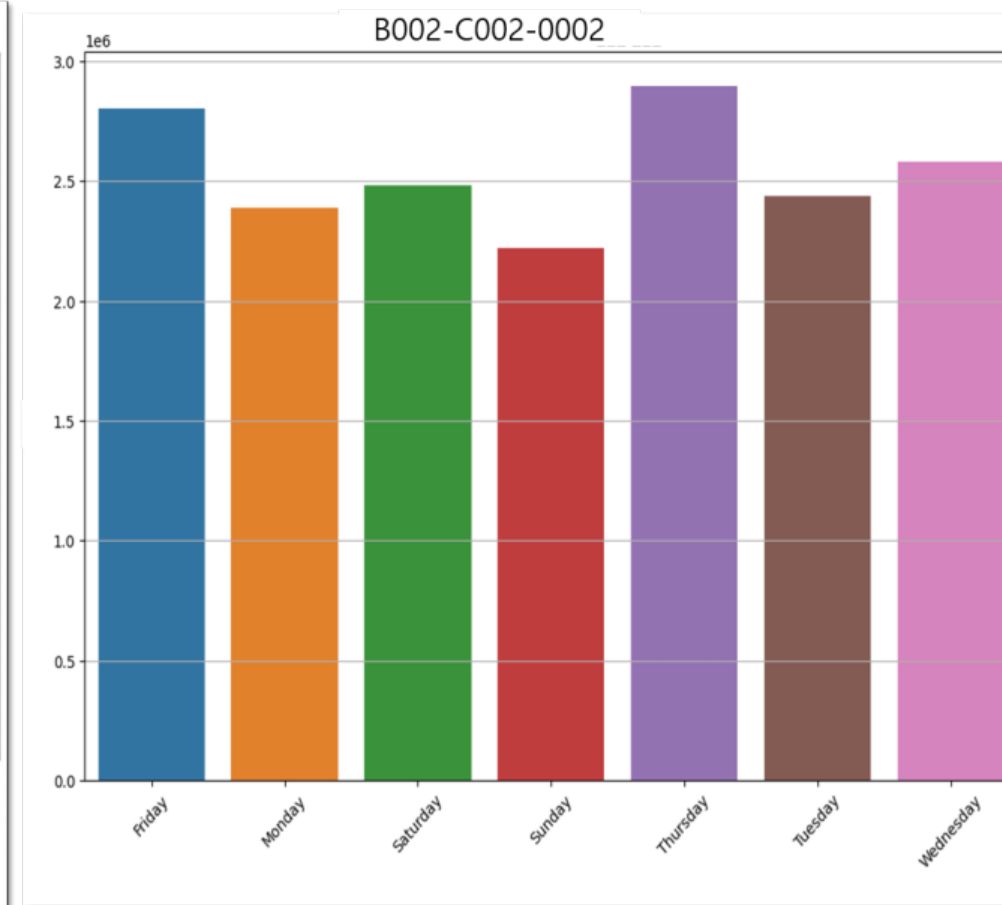
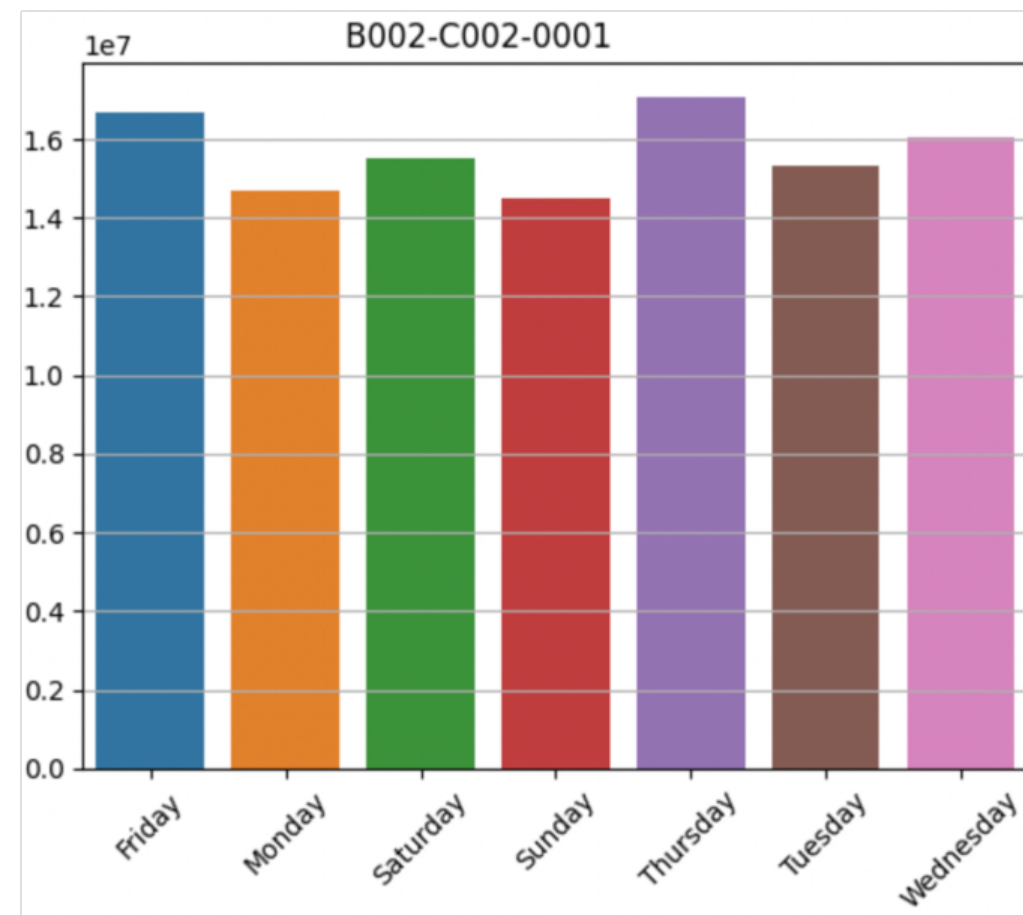
- 중분류를 기준으로 제품의 단가를 확인한 결과, 금액의 폭에 크게 차이가 보인다.

카테고리	가격분류	판매가격	카테고리	가격분류	판매가격
건강기능식품	상	819000	유아식품	상	91900
	중	38400		중	4160
	하	19800		하	1900
뷰티용품	상	33000	제지/위생용품	상	502300
	중	16000		중	13500
	하	9900		하	2875
욕실용품	상	505000	주방/청소/세탁 세제	상	854118
	중	7300		중	12880
	하	2550		하	6225
유아 생활용품	상	389000	탈취/방충/살충/ 제습/방향	상	159000
	중	22900		중	10500
	하	9900		하	4640
유아 위생용품	상	126900	특수헤어용품	상	340000
	중	24100		중	18000
	하	12900		하	9500
유아식품	상	91900	헤어/바디/스킨/ 면도	상	658700
	중	4160		중	22900
	하	1900		하	8100

▶ 일부 중분류에 속한 제품은 높은 가격대에 위치하고,
다른 중분류의 제품은 낮은 가격대에 위치하는 경우가 나타남.

- 건강기능식품 상군집의 최고가격 : 819,000원
 - 하군집의 최고가격 : 19,800원
- 제지/위생용품 상군집의 최고가격 : 502,300원
 - 하군집의 최고가격 : 2,875원
- 헤어/바디/스킨/면도 상군집의 최고가격 : 658,700원
 - 하군집의 최고가격 : 8,100원

- 중분류를 기준으로 요일별 판매량을 분석한 결과, 유의미한 차이를 보인다.



- 화요일 : 주방/청소/세탁세제, 욕실용품
- 목요일 : 건강기능식품, 제지/위생용품,
탈취/방충/살충/제습/방향, 뷰티용품
헤어/바디/스킨/면도, 특수헤어용품,
- 금요일 : 유아 생활용품, 유아 위생용품

▶ 중분류 'B002-C002-0001' & 'B002-C002-0002'의
요일별 판매량

▶ 요일별 판매량이 많은 카테고리 분류

상위 20% 브랜드	전체 기간의 누적 판매량을 기준으로, 각각의 대분류에서 상위 20%에 속하는 제품을 'yes'로, 그 외의 제품은 'no'로 표시하는 "상위 20% 브랜드" 컬럼 생성
대량판매	월별 판매량이 꾸준히 높은 0041, 0003, 0025, 0004, 0001, 0022 제품을 'yes', 그 외의 제품은 'no'로 표시하는 "대량판매" 컬럼 생성
period	시계열 데이터분석을 바탕으로 카테고리별 구매 주기 컬럼 생성 * 주기가 뚜렷하지 않은 카테고리는 steady로 표현
카테고리	중분류를 기준으로 11개의 "카테고리" 컬럼 생성
day_week	중분류를 기준으로 3개의 "화, 목, 금" 속성값을 가지는 컬럼 생성

“

Part2. 모델링

Contents

- 01 모델 검증 : Custom Loss Function
- 02 모델 알고리즘 : 변수 및 파라미터 선정

모델 검증: Custom Loss Function

- 리더보드의 평가 산식은 일별 판매량이 많은 제품을 정확하게 예측할수록 점수가 높아지는 것을 확인함.
- 일반적인 MSE Loss Function이 아닌, 일별 제품의 판매 비중을 추가하여 판매량이 많은 제품을 더 정확히 예측하도록 학습을 유도함.

$$1 - \frac{1}{n} \sum_{day=1}^n \sum_{i=1}^N \left(\frac{|y_i^{day} - p_i^{day}|}{\max(y_i^{day}, p_i^{day})} \right) \times \frac{y_i^{day}}{\sum_{i=1}^N y_i^{day}}$$

[오차] [판매비중]

i : 제품 index

y_i^{day} : i 번째 제품의 day 일의 판매량

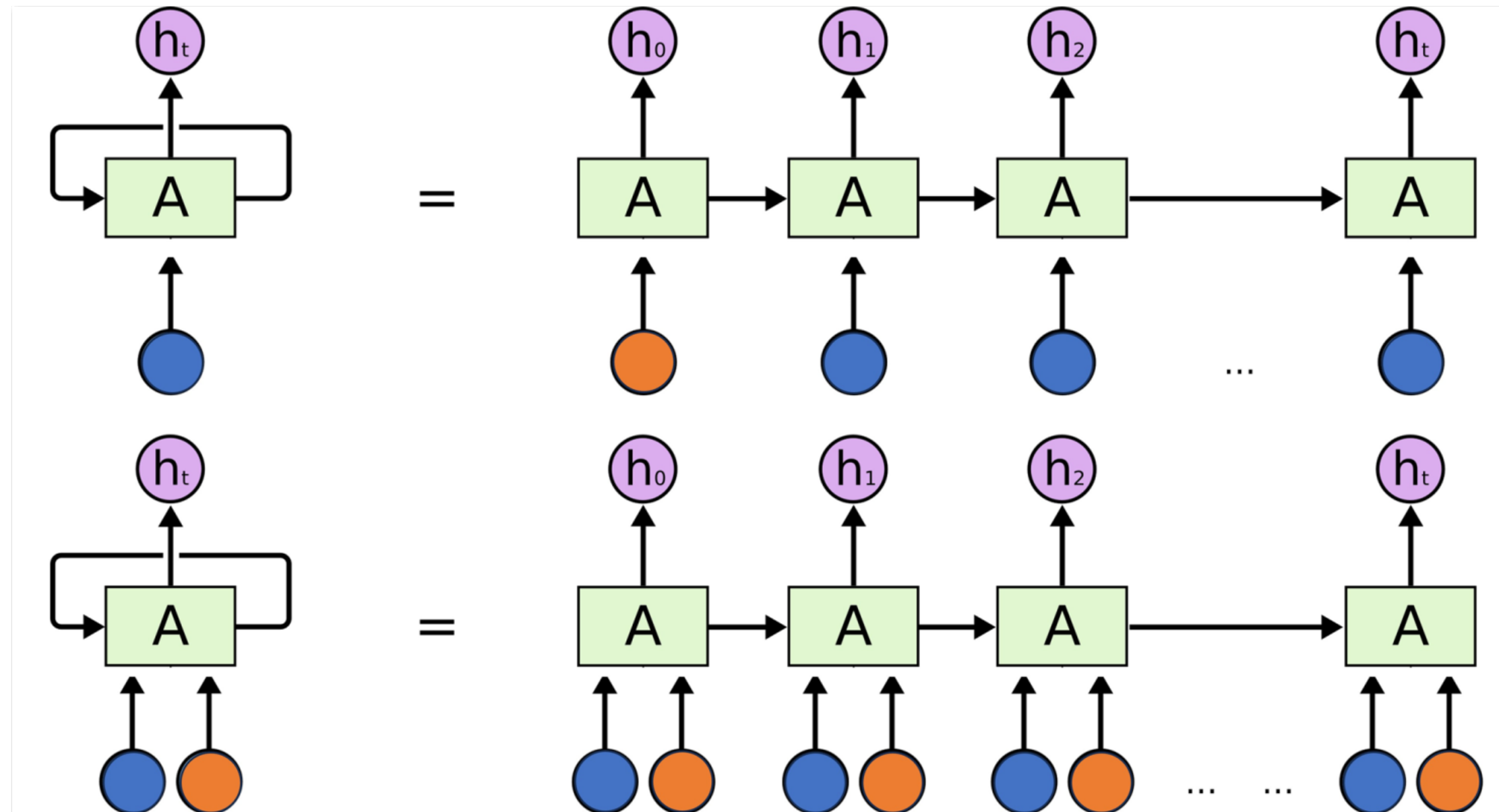
p_i^{day} : i 번째 제품의 day 일의 예측량

```
class CustomLoss(nn.Module):
    def __init__(self):
        super(CustomLoss, self).__init__()

    def forward(self, y_pred, y_true):
        # 일일 판매량의 절대 차이
        abs_diff = torch.abs(y_pred - y_true)
        # max(실제값, 예측값)
        max_values = torch.max(y_true, y_pred)
        # 특정 일의 모든 제품 판매량 합계
        sum_values = torch.sum(y_true, dim=1, keepdim=True)
        # 제품의 일일 판매량 실제값의 비중
        sales_weight = y_true / (sum_values + 1e-10)
        # 최종 손실 계산
        loss = torch.mean(sales_weight * (abs_diff / (max_values + 1e-10)))

        return loss
```

- 활용한 모델 : LSTM
 - 시점마다 제품 특성을 반영한다는 특징을 가짐.



모델 알고리즘: 변수 및 파라미터 선정

- 동일한 LSTM의 구조에서 Window 변화에 따른 성능을 비교한 뒤 최적값을 도출함.

- 조건 1. Predict Size = 21, Step Size = 21로 고정
2. EarlyStopping : Delta = 0.0002, Patience = 20
3. Scheduler : CosineAnnealingLR

▶ 최종 파라미터 선정

- Window Size : 49
- Predict Size : 21
- Step Size : 21
- Split Rate : 0.2

```
100%|██████████| 2709/2709 [00:38<00:00, 70.88it/s]
100%|██████████| 678/678 [00:05<00:00, 123.07it/s]
Epoch : [67] Train Loss : [0.01247] Val Loss : [0.01235] PSFA Score : [0.57382]
조기 종료 카운터: 19 / 20
100%|██████████| 2709/2709 [00:39<00:00, 69.46it/s]
100%|██████████| 678/678 [00:04<00:00, 138.93it/s]
Epoch : [68] Train Loss : [0.01247] Val Loss : [0.01231] PSFA Score : [0.56617]
조기 종료 카운터: 20 / 20
Early stopping
Best saved validation loss: 0.012279
Best PSFA Score : [0.57713]
Best Loss Score : [0.01219]
```

“

Part3.

적용 가능성

Contents

01 모델의 실용성 및 활용 가능성

02 실제 현업에서의 적용 가능성

- 신속성




- 1 에폭 당 학습은 40초, 예측은 4초인 모델은 빅데이터가 자산인 기업에게 필요한 조건이다.
- 가벼운 모델은 대량의 빅데이터가 유입됐을 때, 받아들이기도 쉽고 빠른 결과를 내기 좋다.



```
100%|██████████| 2709/2709 [00:38<00:00, 70.88it/s]
100%|██████████| 678/678 [00:05<00:00, 123.07it/s]
Epoch : [67] Train Loss : [0.01247] Val Loss : [0.01235] PSFA Score : [0.57382]
조기 종료 카운터: 19 / 20
100%|██████████| 2709/2709 [00:39<00:00, 69.46it/s]
100%|██████████| 678/678 [00:04<00:00, 138.93it/s]
Epoch : [68] Train Loss : [0.01247] Val Loss : [0.01231] PSFA Score : [0.56617]
조기 종료 카운터: 20 / 20
Early stopping
Best saved validation loss: 0.012279
Best PSFA Score : [0.57713]
Best Loss Score : [0.01219]
```

- 모델의 일반화 성능

- Public은 10등, Private은 3등으로 30% 양의 데이터보다 전체 데이터에서의 성능이 더 우수한 것으로 나타났다.

▶ Public 순위

#	팀	팀 멤버	점수	제출수	등록일
10	스틸메이트	  	0.57178	52	2시간 전

Private 순위	팀명
1	
2	
3	스틸메이트



LG Academy
DACON

온라인 채널 제품 판매량 예측

감사합니다

TEAM | SteelMate

송준희 | 박지현 | 오수민