NS홈쇼핑

B1C4조

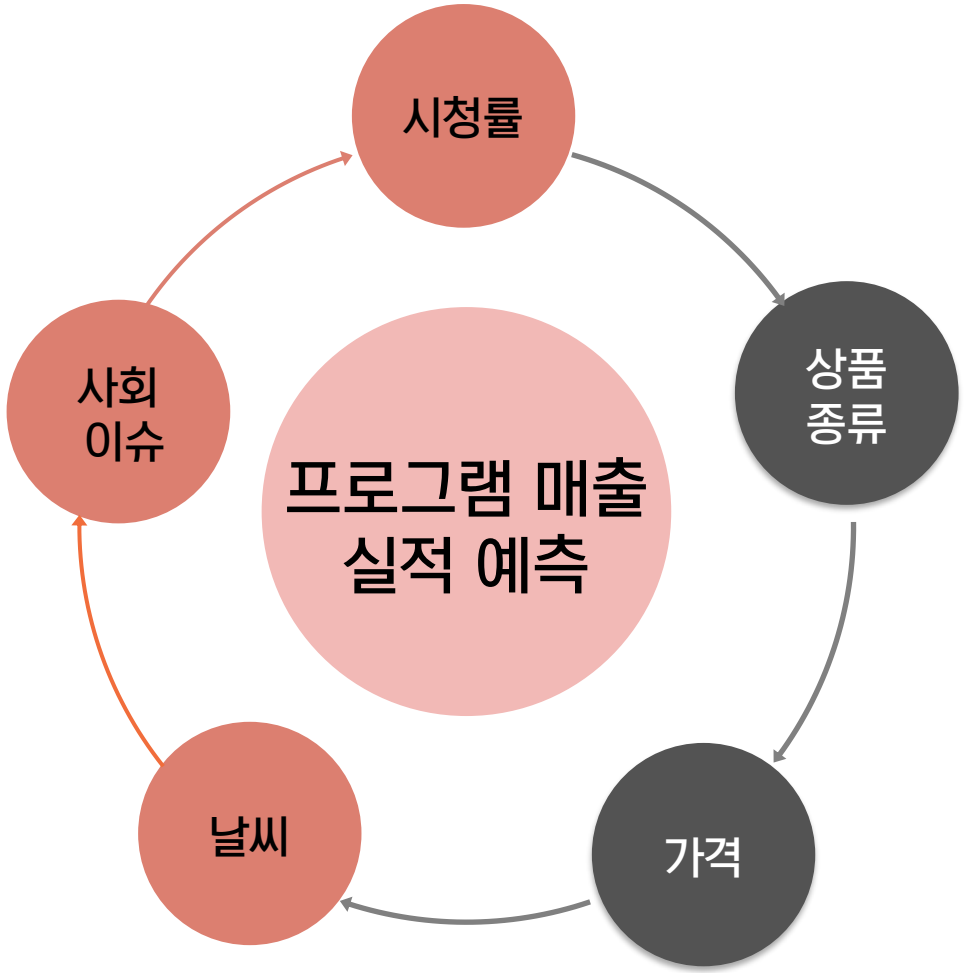
빅콘테스트 챔피언리그

팀장 이재훈(ljw5694@naver.com)
김자영 (1476532@sookmyung.ac.kr)
송재용(seonbinara@daum.net)
윤나은 (na803704@naver.com)
진서연 (sheoyonj0112@naver.com)

NS홈쇼핑

INDEX

프로젝트 목적 및 전체 개요



실적의 변화

내적 요인 : 상품종류, 가격

외적 요인 : 시청률,사회이슈,날씨

시간의 흐름에 따른 변화

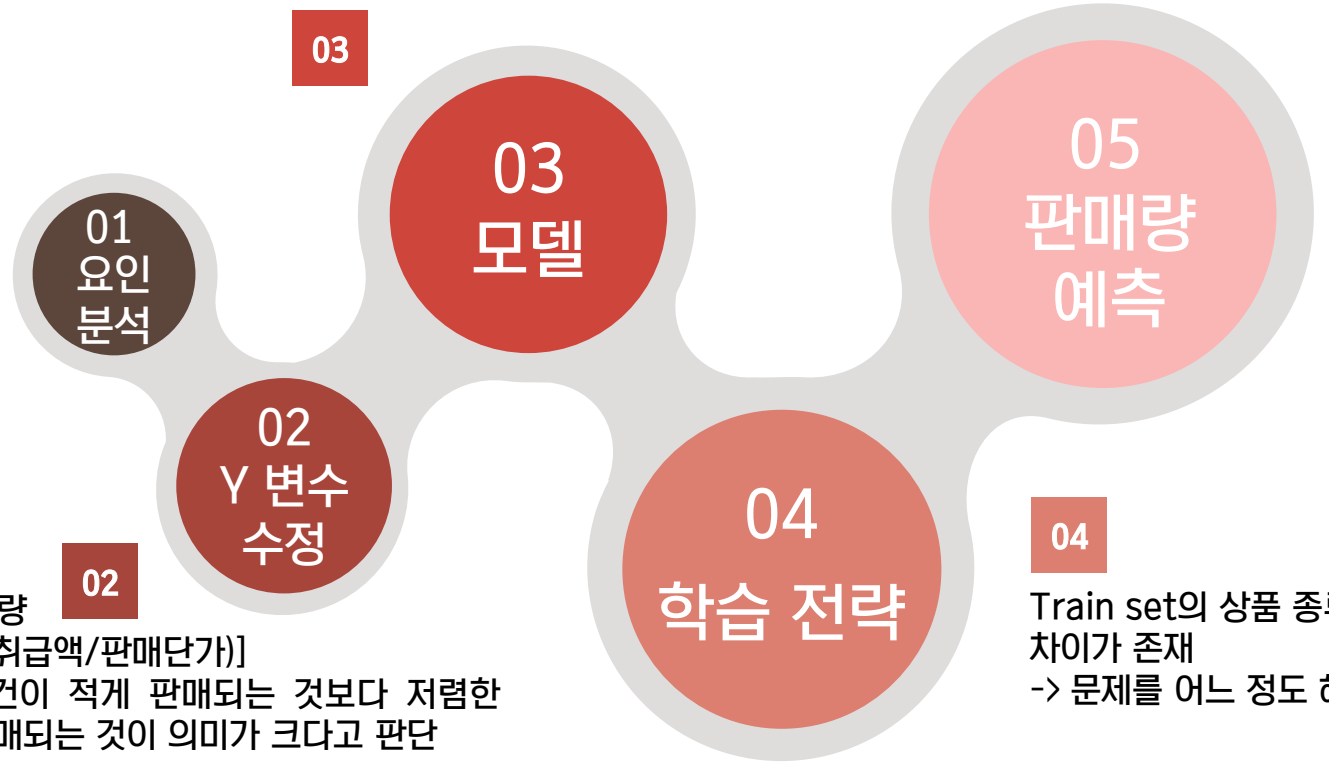
→ 5가지 요소를 고려하여 프로그램 매출 실적 사전 예측, 대응

프로젝트 목적 및 전체 개요

05 학습된 모델을 통해 판매량 예측 및 최적 편성표 도출

시계열 /비시계열 모델

03



01

내부데이터, 외부 데이터 활용 홈쇼핑 상품 매출 실적에 영향을 미치는 요인들을 분석

02

취급액 -> 판매량

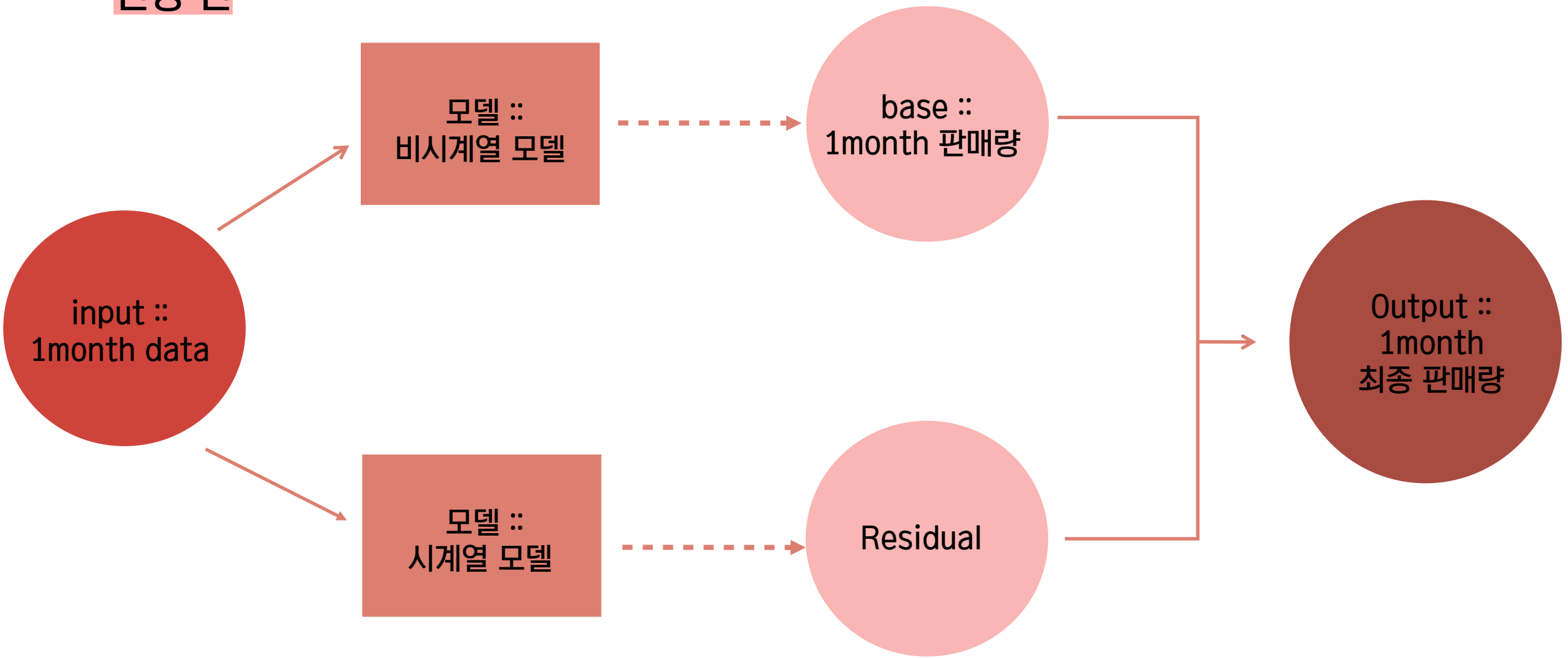
- 판매량 : [ceil(취급액/판매단가)]
- 이유: 고가 물건이 적게 판매되는 것보다 저렴한 물건이 많이 판매되는 것이 의미가 크다고 판단

04

Train set의 상품 종류와 test set의 상품 종류가 꽤 큰 차이가 존재
-> 문제를 어느 정도 해결할 수 있는 학습 전략 선택 필요

프로젝트 목적 및 전체 개요

변경 전



프로젝트 목적 및 전체 개요

변경 후

기존에는 비시계열로 base를 예측하고 시계열 기반 모델(** Convolution, GRU)으로 residual을 조정하는 방식
—> 시계열 기반에서 값이 수렴하지 않는 issue 발생 **

—> 아래 그림과 같이 비시계열 기반의 모델만 사용

** 시계열 기반 모델 구조 정보는 추가자료1(2).pdf에 첨부하였습니다

변수 선정 1 – 날씨

선정 방법

- 102개의 관측소에서 측정한 1시간 단위의 날씨 데이터를 평균 내어 활용
- 가정 내에서 창문 등을 통해 시각적으로 확인하기 쉬울 것이라고 생각되는 날씨 변수 선정하여 반영

변수 명	설명
기온	기온
강수량	강수량
Vs	시정(가시거리)
Lcsch	최저 운고(구름 밑 부분까지의 고도), 값이 낮을 수록 구름이 낮게 위치함
Dc10tca	전운량(하늘에 구름이 덮고 있는 비율에 따라 0-10으로 구분), 전부 구름이면 10 부여
Icsr	일사량(태양의 복사에너지가 지표에 닿는 양), 태양의 고도가 높을수록 일사량 증가
Ss	일조량 (태양 광선이 지면에 얼마 동안 비추었는지)
Pa	기압
Pv	증기압
Hm	습도
Ws	풍속

변수 선정 2 – Japping Time

선정 방법

- 닐슨의 2019년 1월 1일-2019년 12월 31일 일별 시청률 순위 활용
- 인기 프로그램 후 방영되는 홈쇼핑 시청률 및 판매량이 높을 것으로 판단
- 프로그램 별 평균 시청률 계산 후, 평균 시청률이 높은 순으로 sorting 시

변수 명	설명
Popular Program	• 평균 시청률이 7% 이상인 프로그램들이 끝난 직후에 방송하는 경우 1, 아닌 경우 0
Morning Drama	• NS 홈쇼핑 시장의 주 고객층이 40-50대 여성 임을 고려 • KBS 아침마당, SBS 아침드라마 시간대 직후인 경우 1, 아니면 0 부여 MBC, KBS 아침드라마는 작년에 폐지됨

지상파 일일 - TOP 20 LIST FOR TV PROGRAMS

Data Search 2020년 09월 19일

전국 수도권

2020.09.19

가구시청률 TOP 20

(분석기준: 수도권, 가구, 단위:%)

순위	채널	프로그램	시청률
1	KBS2	주말드라마(오삼광빌라)	22.7
2	KBS2	주말드라마(오삼광빌라)	19.0
3	MBC	놀면뭐하니	12.9
4	MBC	놀면뭐하니	10.1
5	SBS	금토드라마(엘리스)	10.0
6	KBS1	KBS9시뉴스	8.0
7	SBS	정글의법칙IN와일드코리아	7.7
7	SBS	금토드라마(엘리스)	7.7
9	KBS2	불후의명곡	7.4
10	KBS1	특파원보고세계는지금	6.9
11	KBS1	시니어토크쇼황금연못	6.8
12	KBS1	김영철의동네한바퀴	6.5

시청자수 TOP 20

(분석기준: 수도권, 개인, 단위:천 명)

순위	채널	프로그램	시청자수
1	KBS2	주말드라마(오삼광빌라)	2,371
2	KBS2	주말드라마(오삼광빌라)	1,899
3	MBC	놀면뭐하니	1,846
4	MBC	놀면뭐하니	1,301
5	SBS	금토드라마(엘리스)	1,186
6	SBS	금토드라마(엘리스)	963
7	SBS	정글의법칙IN와일드코리아	914
8	KBS1	KBS9시뉴스	797
9	SBS	정글의법칙IN와일드코리아	792
10	MBC	전지적참견시점	769
11	KBS1	특파원보고세계는지금	719
12	KBS2	살림하는남자들	696

변수 선정 3 – 날짜, 시간

변수 선정

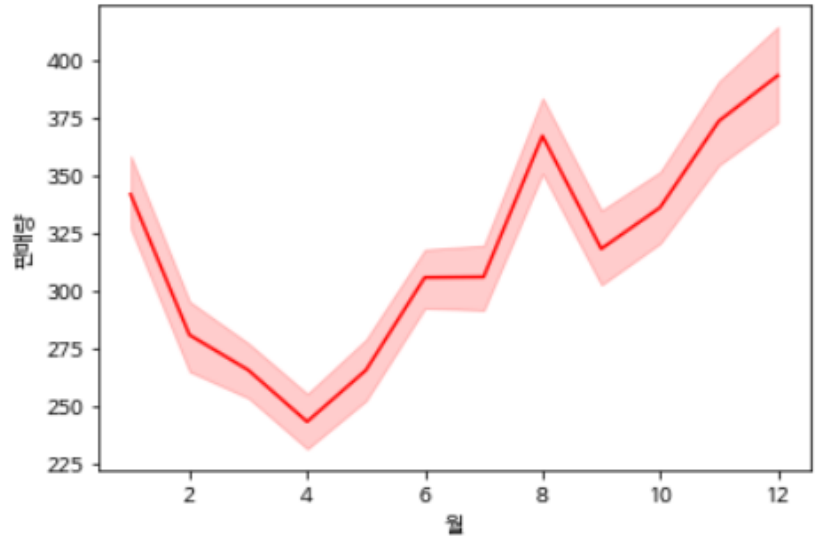
변수 명	설명
월	• 각 방송 날짜의 month
시간	• 각 방송 날짜의 hour
요일	• 각 방송날짜의 weekday
연달아 쉬는 날	• 토요일, 일요일, 공휴일 고려 • 연 이은 휴일의 첫 번째 날은 1, 그 다음은 2, 3… 으로 표시 ex) 전후로 공휴일이 없는 토, 일의 경우 각각 1,2로 표시

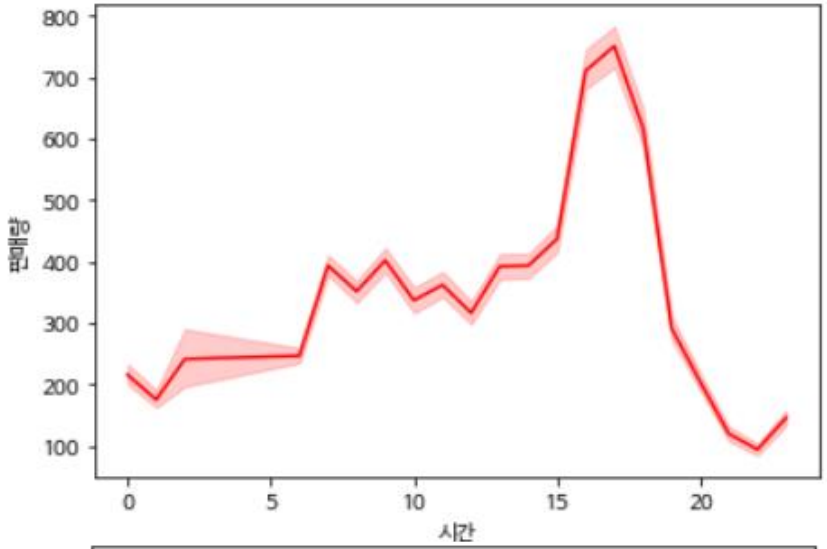
유의사항

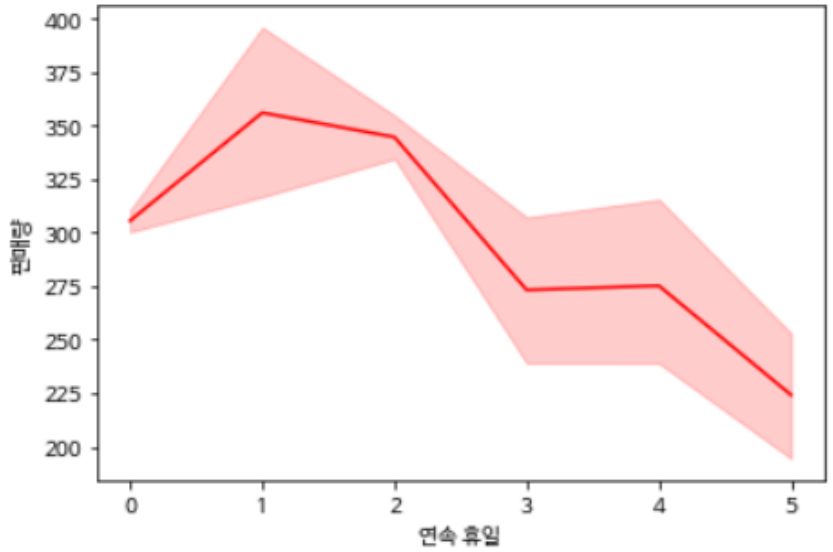
- 하루 방송이 아침 6시~새벽 2시로 구성되어 있으나, 새벽 0~2시 방송은 다음 날로 처리함
  - 일요일에서 월요일로 넘어가는 새벽의 특성이 주말보다는 월요일에 더 가깝다고 판단함.

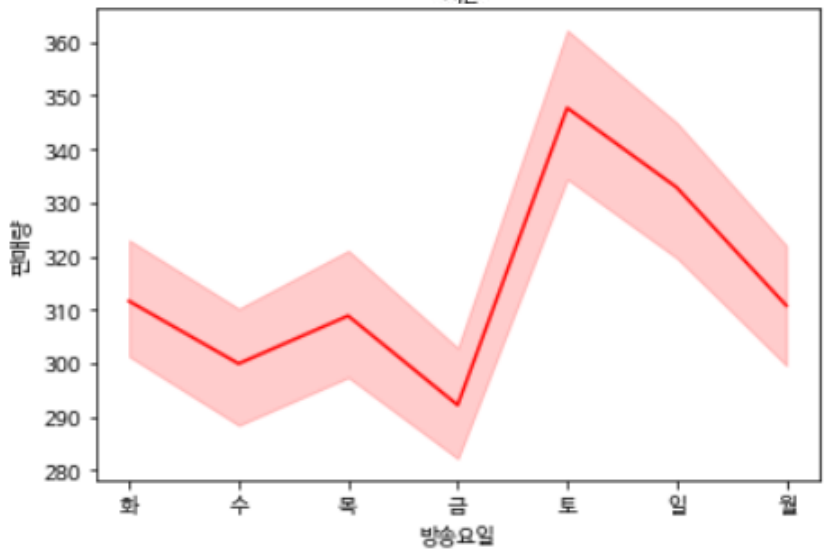
변수 선정 3 – 날짜, 시간

Univariate Analysis









NS홈쇼핑

변수 선정 4 – 상품종류, 상품 명, 가격

변수 선정

변수 명	선정방법 및 설명
브랜드	• 각 제품명을 확인하여 브랜드 추가 • Test data의 경우, 브랜드가 train data에 있는 경우에만 추가
상품군, 마더코드 상품코드	상품군, 마더코드, 상품코드 통일 • 상품명이 동일한데 여러 개의 상품군, 마더코드, 상품코드가 있는 경우, 가장 자주 등장한 것으로 선정 • 띄어쓰기 된 상품명 바탕으로 단어 set이 동일한 경우 동일한 상품으로 판단 • 동일한 상품코드가 없는 경우 대비 상위 카테고리(상품 군, 마더코드)까지 고려
판매단가	• 판매가격

- 아래는 해당하는 단어가 상품명에 포함 시 효과 반영하기 위한 변수임

성별	• 남성, 여성 포함 시 1, 미포함 시 0
일시불/무이자	• 일시불, 무이자 포함 시 1, 미포함 시 0
광고 사람	• 국내 인지도 있는 연예인인 ‘팽현숙‘, ‘이봉원‘, ‘숀리‘ 등 포함 시 1, 미포함 시 0
국내생산	• 국내 생산을 의미하는 단어 포함 시 1, 미포함 시 0 - ‘국내’,’ 동해안’,’ 완도‘,’ 안동‘,’ 영광‘,’ 영산포’ 등이 포함되면 국내 생산임을 의미

변수 선정 5 – 방송편성

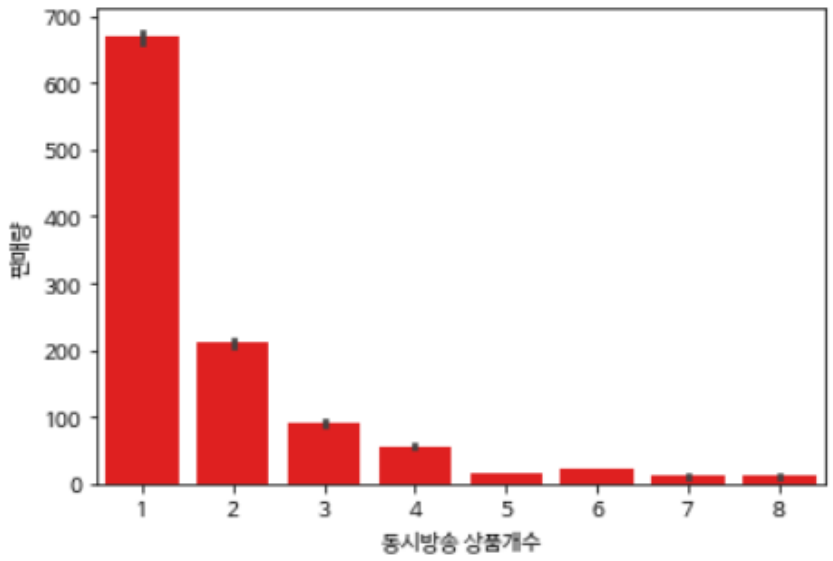
변수 선정

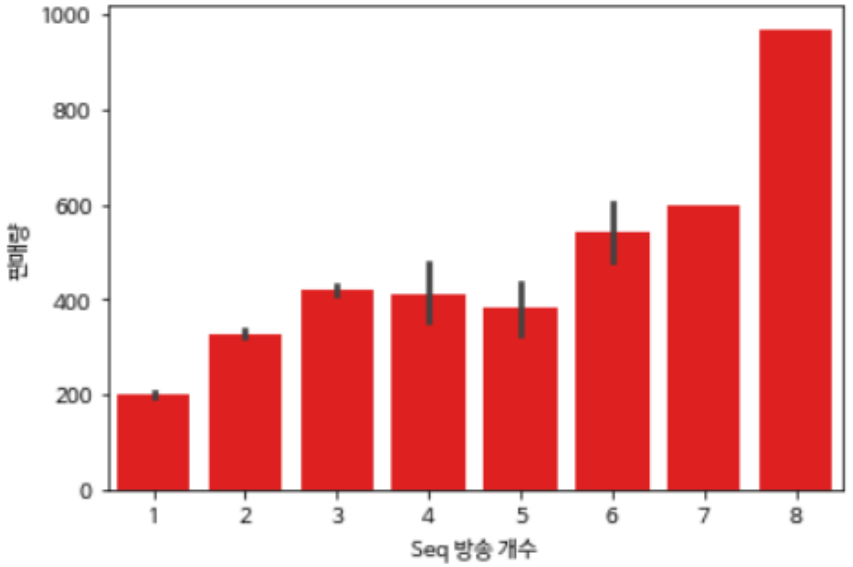
변수 명	선정방법 및 설명
동시방송 상품개수	• 한 방송에 여러 상품이 동시에 판매되는 경우 • 판매량이 분산되는 효과 반영 Ex. 아래 사진처럼 동일 시간에 여성/남성 미들퍼부츠 동시 판매하는 경우 2로 표시
Seq 방송 개수	• 같은 방송이 연달아 방송되는 개수 추가 ex. 1월 1일 7시 A 상품, 7시 20분 A 상품, 20시 A 상품 편성될 때 각각을 1,2,1으로 표시 • 같은 방송 여부: 상품명 단어들의 list 유사성을 바탕으로 판단 • 여러 제품 동시 판매 시: 모든 이름의 concat을 하나의 제품으로 간주하여 처리 • 연속방송 여부: 한 방송이 끝나고 10분 이내로 방송이 되는지
하루 방송 수	• 같은 방송이 하루에 여러 번 방송되는 경우 ex. 1월 1일 7시 A 상품, 7시 20분 A 상품, 20시 A 상품 편성될 때 각각을 1,2,3으로 표시 • 같은 방송 여부, Seq 방송 개수 계산 시 사용된 방법 사용

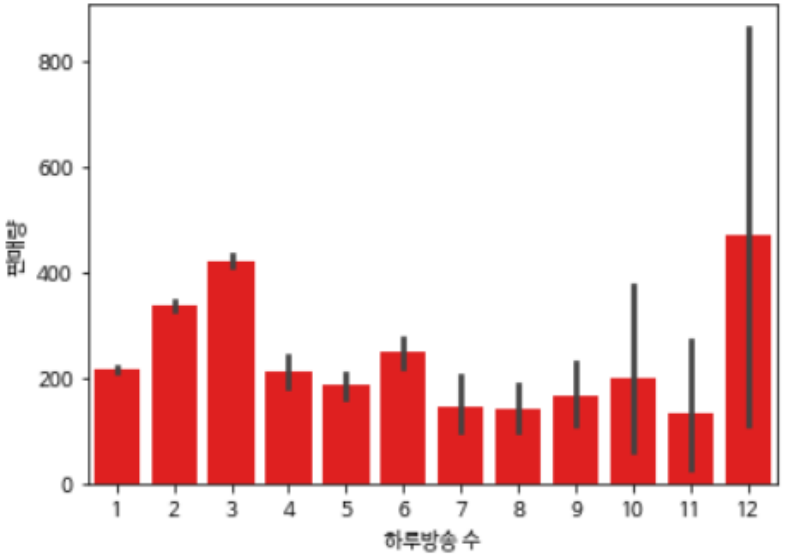
2019-01-03 6:00	20	100781	202292	에펨 여성 미들퍼부츠	잡화	39,800	2,450,000
2019-01-03 6:00		100781	202285	에펨 남성 미들퍼부츠	잡화	49,800	2,645,000

변수 선정 6 – 방송편성

Univariate Analysis



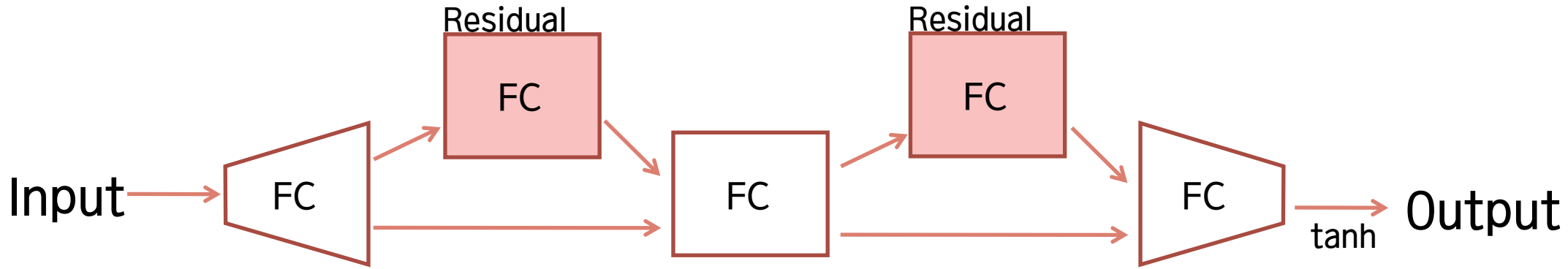




모델 선정 및 학습 방법

모델 구조| Residual Net 기반 모델

- 일반 FC를 이용했을 때 보다 층을 더 깊게 쌓을 수 있음
- 중간중간 ReLU의 변형인 CeLU 활성화 함수와 BatchNorm을 사용하여 학습에 도움을 줌
- 모델의 예측 값이 min,max 사이를 쉽게 움직일 수 있도록 y 변수를 (-1,1) 범위로 min-max transformation 진행한 후, 모델의 예측 값이 (-1,1)이 되도록 마지막에 tanh 함수 사용



Embedding Vector

1. **모든 데이터 embedding vector로 변환**
   numerical의 경우 ((x-mean)/std) * embedding_vector
   모델 input의 차원이 줄어들고, 복잡한 상관관계 반영 가능

2. **Embedding 된 값을 concat , sum 연산을 통해서 모델에 넣음**
   Sum을 할 때는 group별로 sum을 하게되는데, 날씨데이터,상품군,상품코드 등과 같이 동일한 의미를 나타내는 데이터는 Embedding한 차원들도 동일한 의미를 나타내며, 이 값들을 합함으로써 모델 input의 차원을 줄일 수 있다.

모델 선정 및 학습 방법

Model Fitting

1. 2단계 모델 학습:

기존의 train데이터에서 유사도 matrix를 이용해 input의 상품군, 마더코드, 상품코드의 embedding vector를 바꿔서 훈련함

ex. A제품과 B제품의 유사도 0.7, A제품과 C제품의 유사도 0.5, A제품과 D제품과의 유사도 0.3
—> A제품의 상품군, 마더코드, 상품코드를 B제품의 것으로 변경하고 임베딩 벡터를 0.7으로 normalizing.
(0.7 혹은 0.7의제곱을 곱함) training 진행 —> C, D제품을 이용해 같은 방식으로 학습진행

- 단어간의 유사도라서 정확하지 않기때문에 **최소 유사도 기준**을 두어 유사도가 기준 이하 -> 유사도를 0으로 변경
- 상위 k개의 제품을 가지고 유사도 학습을 진행 (k는 하이퍼 파라미터)

→ 임베딩 벡터의 유사도 학습이 보다 잘 이루어질 수 있음, 상품코드, 마더코드가 많이 겹치지 않는 train, test데이터의 문제점을 보완해 줄 수 있다.

✓ 유사도: 전체 단어 대비 단어 중복 개수 계산
ex. '테이트 남성 바지', '테이트 여성 바지' 총 4개의 단어 중 3개가 겹치므로 $\frac{3}{4}$로 표시

2. 브랜드 랜덤선택 전략:

브랜드의 경우 train, test의 겹치지 않는 경우가 많지만 겹치는 경우처럼 유사도 전략을 사용하기 어렵기때문에, 영향력을 줄여야한다.
즉, 모델이 브랜드가 없을때 학습이 잘 되도록 해야한다. training할때 50%의 브랜드 임베딩 벡터는 강제로 0으로 만들어서 학습시킨다.

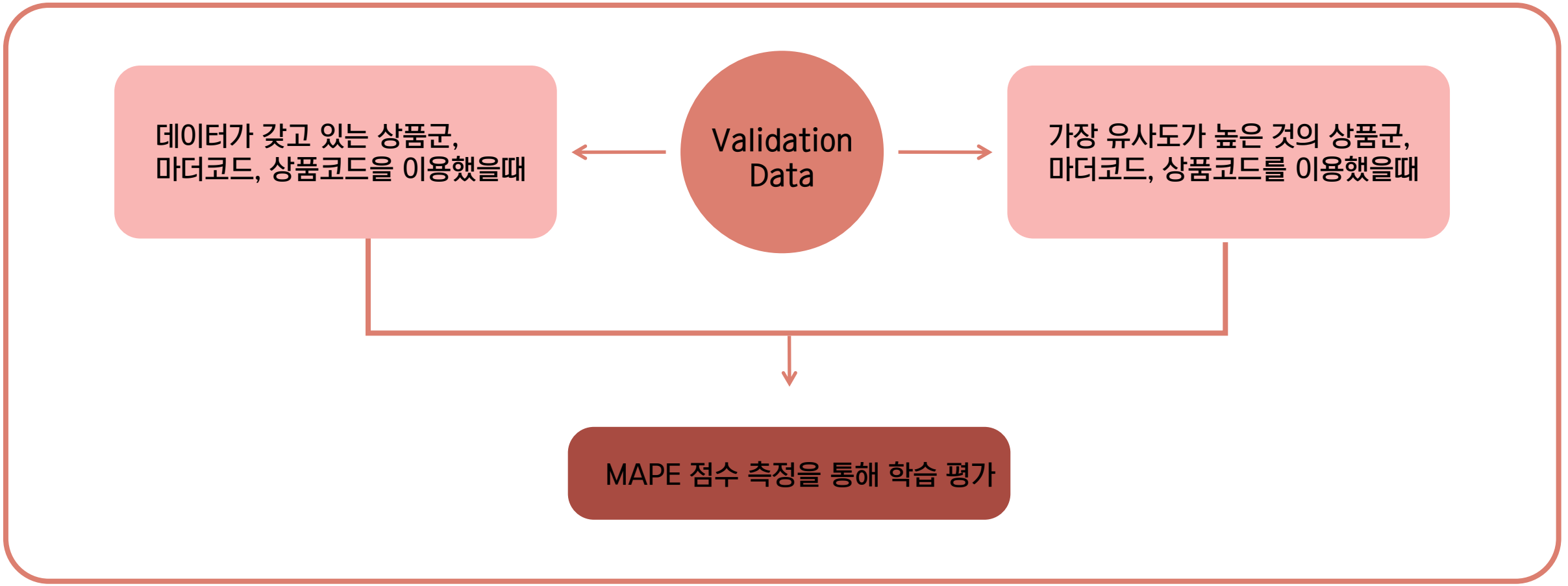
3. 차별화된 Penalty:

상품군, 마더코드의 경우, 임베딩 벡터의 크기에 낮은 페널티를 부여해 변화가 크도록하고, 상품코드, 브랜드는 변화가 작도록 한다.
상품코드, 브랜드는 train dataset과 test dataset 에서 겹치지 않는 경우가 많기 때문에 그 **영향력을 줄여야한다.**

모델 선정 및 학습 방법

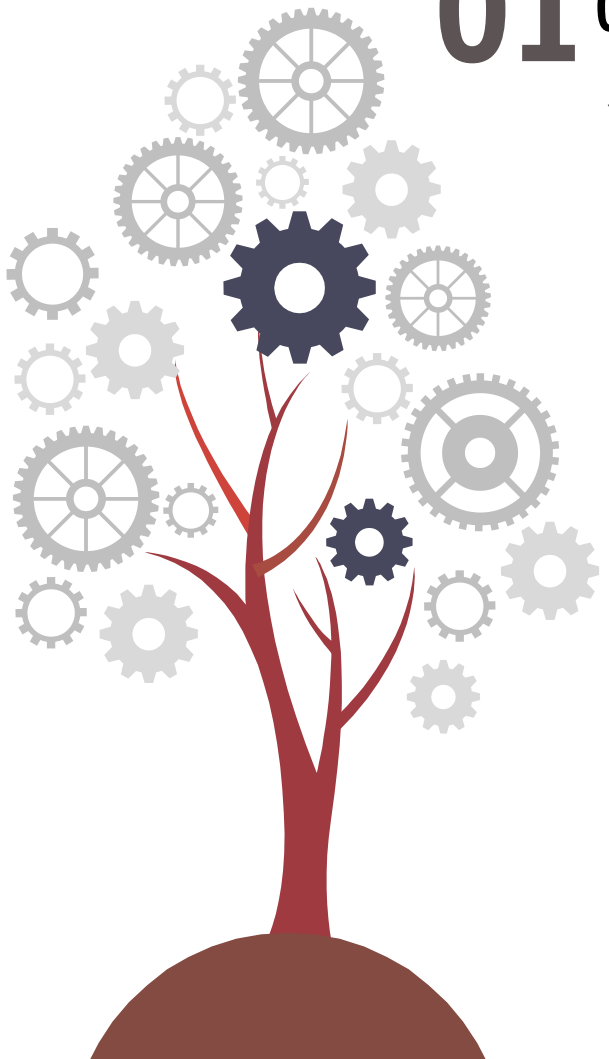
Validation

모델의 학습 평가를 위해 Test MAPE 점수 측정



판매량(취급액) 예측 및 편성 최적화

6월 판매량 예측



01 예측을 위해 사용한 모델

10000 epoch training 한 모델을 사용

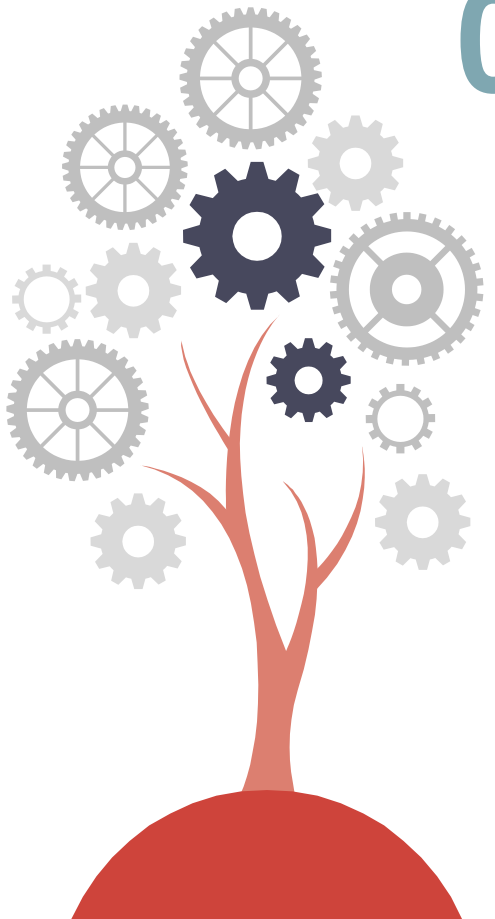
02 예측에 사용한 변수

- 상품종류, 가격, 날짜, 시간, 상품명, 방송편성 변수는 train과 동일한 방법으로 변수 계산
- Train에 없는 상품군, 마더코드, 상품코드의 경우 가장 높은 유사도를 갖는 제품의 상품군, 마더코드, 상품코드 이용
- 타 방송 관련 변수는 현재까지의 프로그램 시청률을 기준으로 함
- 날씨 변수는 6월의 시간대의 날씨를 넣어서 예측

(*날씨는 미리 알 수 있는 것이 아니므로 실제로 사용할때는 날씨의 예측치를 넣으면 됩니다.)

판매량(취급액) 예측 및 편성 최적화

편성 최적화



01 가정

1. 동시간 대에 방송된 제품은 1개의 제품으로 취급
2. 기존의 방송 편성 시간을 동일하게 고정하고 제품만 재배치

02 중복을 제거한 모든 제품 판매량을 모델로 예측한 다음 헝가리안 알고리즘을 이용

⟶ 최대 판매량이 될 수 있도록 배치

03 남은 자리에 2번 중복해서 방송된 제품 판매량을 모델로 예측한 다음 헝가리안 알고리즘으로 배치

04 3번..n번 중복해서 나온 제품을 위의 방식으로 반복해서 배치

05 Train data에서의 빈도수를 고려해서 Seq 방송 개수, 하루방송 수의 최대 개수를 6개로 고정함

› 모든 경우의 수를 고려 시 n!의 복잡도로 하루 치 방송 편성표를 구성 가능 but, 한달 편성 표 구성 시 한계

› 위의 방법을 활용하여 복잡도를 줄이면서 방송 편성 관련 변수(Seq 방송 개수, 하루방송 수)의 영향을 반영 가능

최종 모델선정 및 결과

Model Hyperparameter

Adam_LearningRate= 1e-4;

Adam_beta= (0.5,0.9);

Weight_decay=0.01;

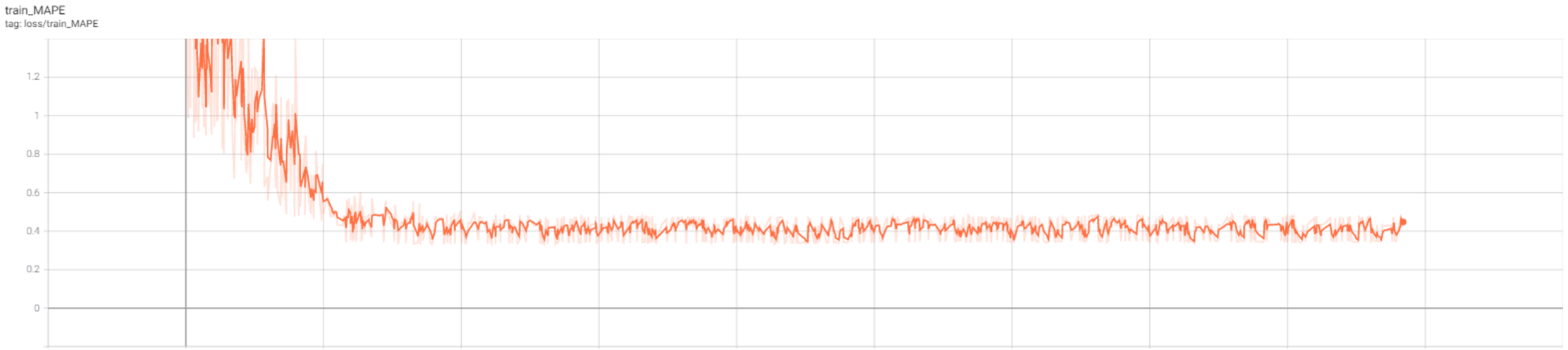
Epoch=10,000

Batch_Size=10,000

Data Hyperparameter

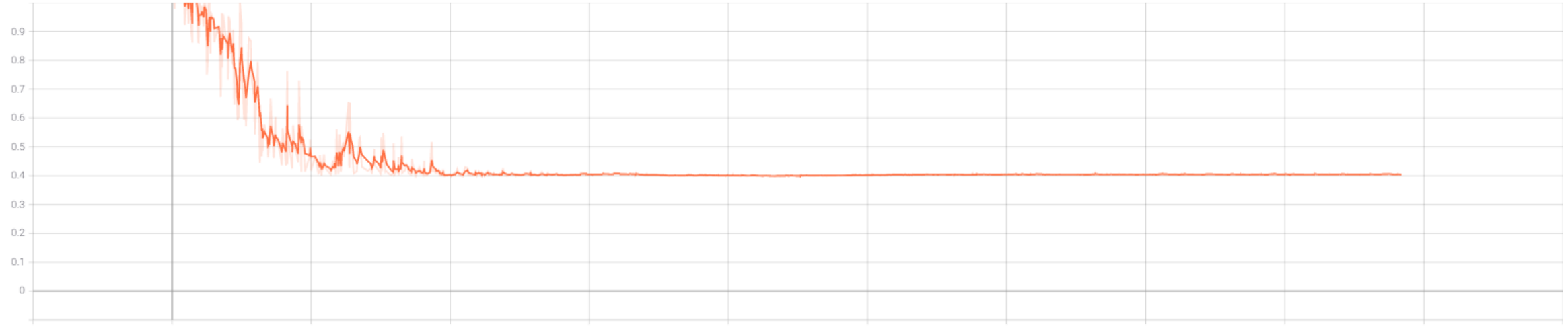
- Y_col_name= '판매량'
- 모든 변수 size 20 벡터로 embedding
- 모든 변수 embedding_regularizer penalty 계수 크기(0.01) 동일
  - 차별화된 페널티 전략 효과 크게 없었음
  - 유사도 바탕 training을 진행한 효과로 보여 짐
- 모든 embedding_vector concat해서 모델에 들어 감
  - Sum 보다 concat을 했을 때 loss 가 안정적인 수렴을 보임
- Seq 방송개수, 하루방송 수 변수를 구할 때, 같은 제품 여부 판단 기준을 80%, 50% 이상의 유사도를 가지는 지로 정함
- 유사도 0.7 이상인 것 중 상위 2개에 대해서 유사도 training 전략을 취함

< Train Mape >

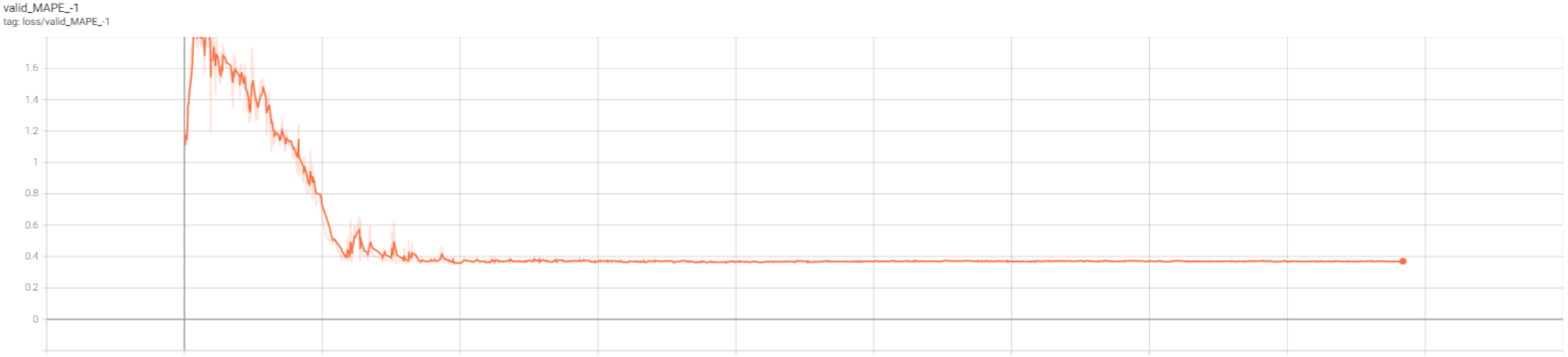


최종 모델선정 및 결과

Validation Mape: 원래 상품 정보 이용



Validation Mape: 유사 상품 정보 이용



한계점

1. 시계열 데이터임에도 불구하고 시계열 데이터 특성을 반영하지 못함
   - 시간에 따른 판매량 변동폭이 커서 시계열 특성을 모델이 잡아내지 못한 것으로 보임

2. 2020년 5,6월 시청률 데이터가 주어지지 않아 시청률 데이터 활용하지 못함

   이전 4주간 해당 요일의 방송 시간 시청률 데이터를 평균내어 해당 날짜 시청률 예측하려고 했으나, test data의 경우 5,6월의 시청률 데이터가 주어지지 않았고, 19년도 시청률을 평균내어 사용하기에는 정확도가 떨어진다고 판단

3. 최적화 알고리즘에서 Seq 방송 개수, 하루 방송 수의 영향 어느 정도만 반영
   - 시간 복잡도를 고려하여, 완전히 같은 제품인지 판단하여 연속, 반복된 방송임을 판단

     > 변수 생성 시에는 일정 크기의 유사도를 가지면 연속, 반복된 방송임을 판단

NS홈쇼핑

감사합니다

- B1C4 조-