빅데이터 기반

T-commerce 매출액 예측 및 편성 제안

김동규 (rb9928@naver.com)

고광종 (rhkswhdwkd@naver.com)

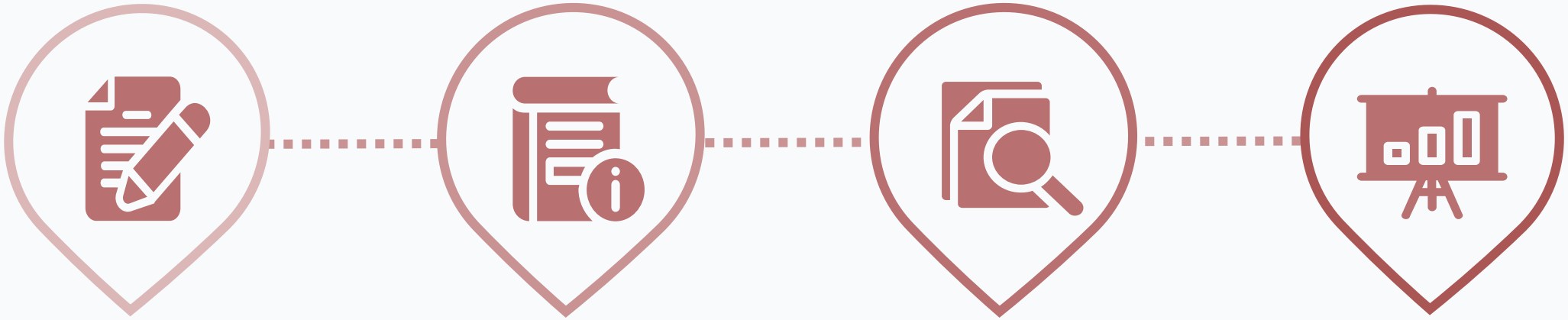
배성은 (noon0131@naver.com)

이정환 (wjdghks9885@naver.com)

채지민 (thrre3214@naver.com)

TEAM 동작하라

목차



프로젝트 개요

분석 배경 및 목적

분석 개요

EDA & 전처리

시계열 특성 변수

방송 특성 변수

상품 특성 변수

날씨 특성 변수

모델링

모델링

스케줄링

결론

시사점

프로젝트 개요

분석 배경 및 목적

분석 개요

(1) 분석 배경 및 목적

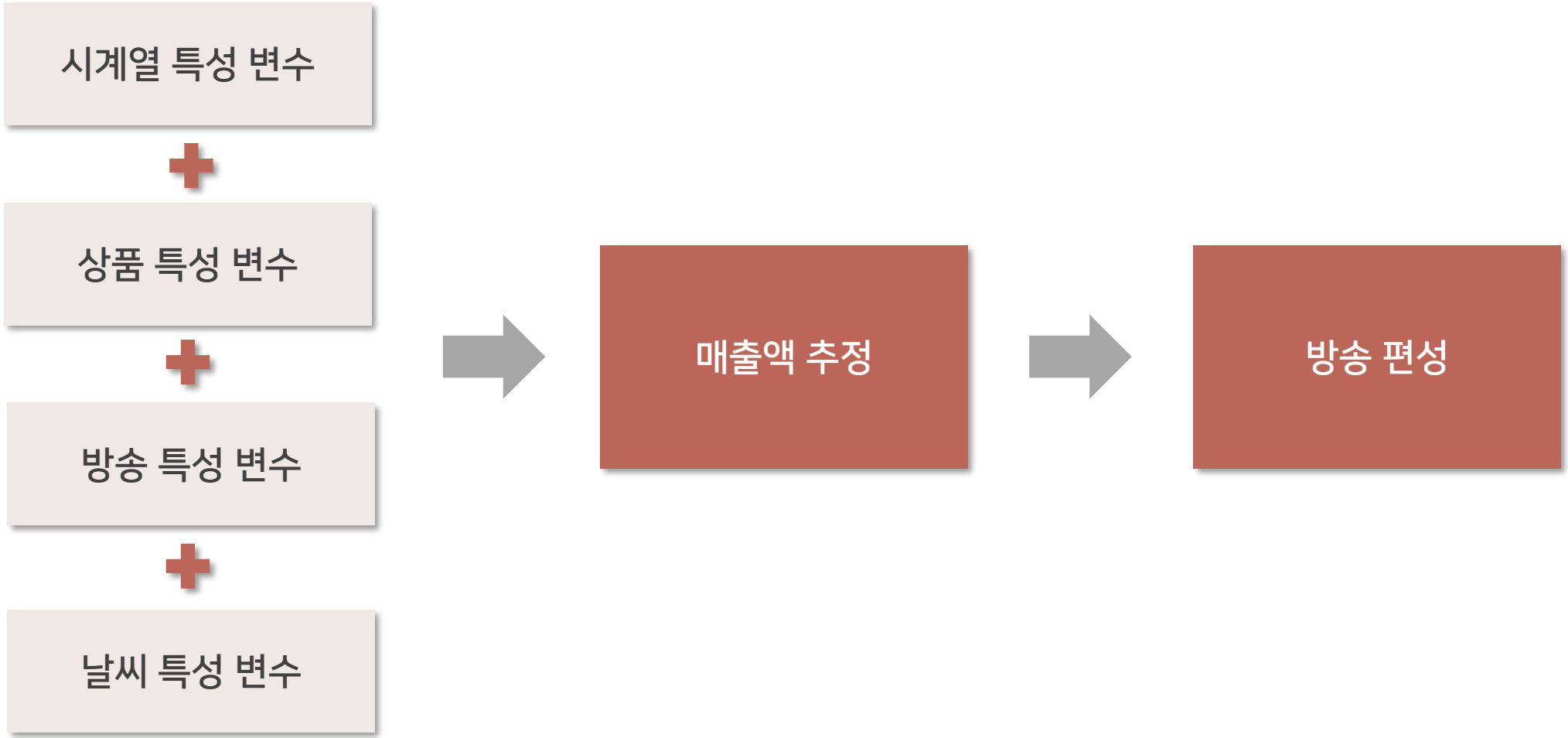
현재

현재 ns shop에서는
운영상품 및 카테고리가
기존 편성표를 바탕으로
반복하고 있는 상황

목표

판매 실적 데이터와 시청률 데이터,
날씨 데이터의 활용과 외부 요인을
바탕으로 상품 매출을 예측하고
미래의 홈쇼핑 편성 최적화

(2) 분석 개요



EDA & 전처리

시계열 특성 변수

방송 특성 변수

상품 특성 변수

날씨 특성 변수

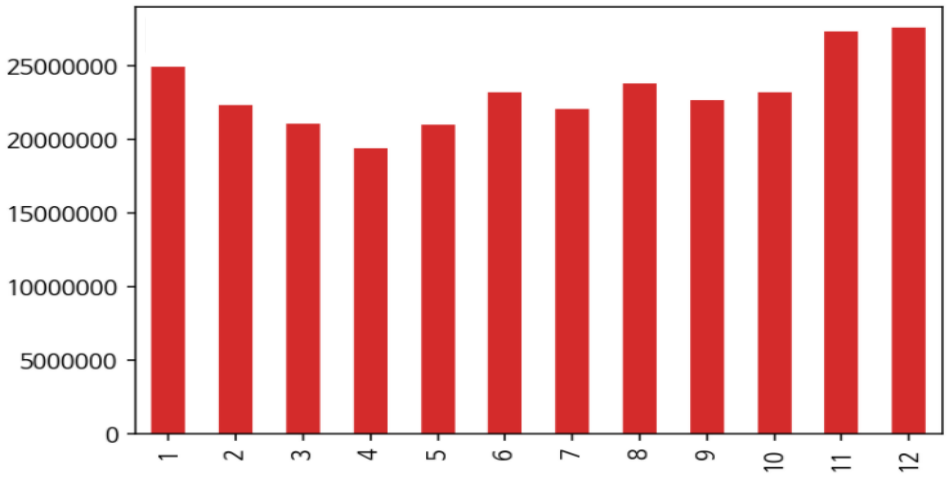
(1) 방송일시 파생변수

Column	Explanation
Year	연도
Month	월 (1 ~ 12)
Day/Businessday	일자/ 영업시간 기준일자 (오전 6시 ~ 다음날 오전 2시)
Hour	시간 (0 ~ 23)
Weekday	요일 (0 : Mon ~ 6 : Sun)
WeekofYear	주차 (0 - 52)
Season	계절 (0 : Spring - 3 : Winter)
Holiday	휴일 여부 (0, 1)

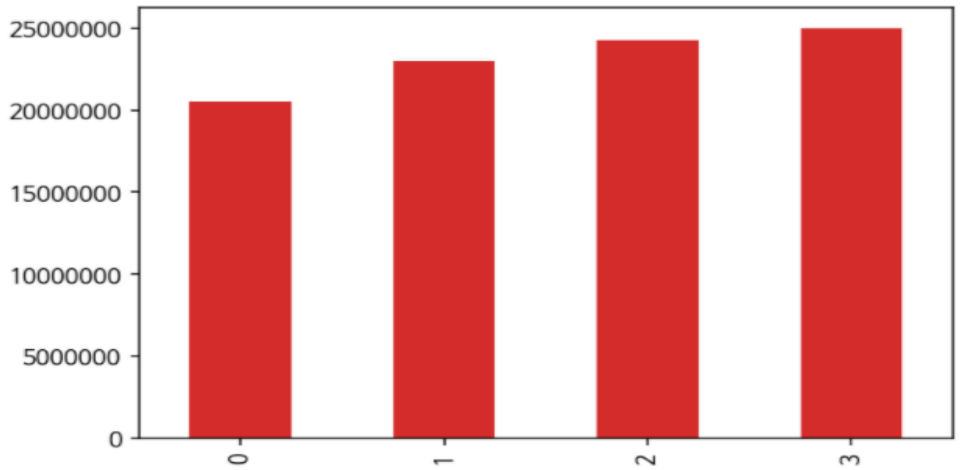
[방송일시] 변수를 활용하여 연도, 월, 일, 시간대, 요일, 주차, 계절, 휴일 변수를 생성

(1) 방송일시 파생변수

월별 매출액 분포



계절별 매출액 분포

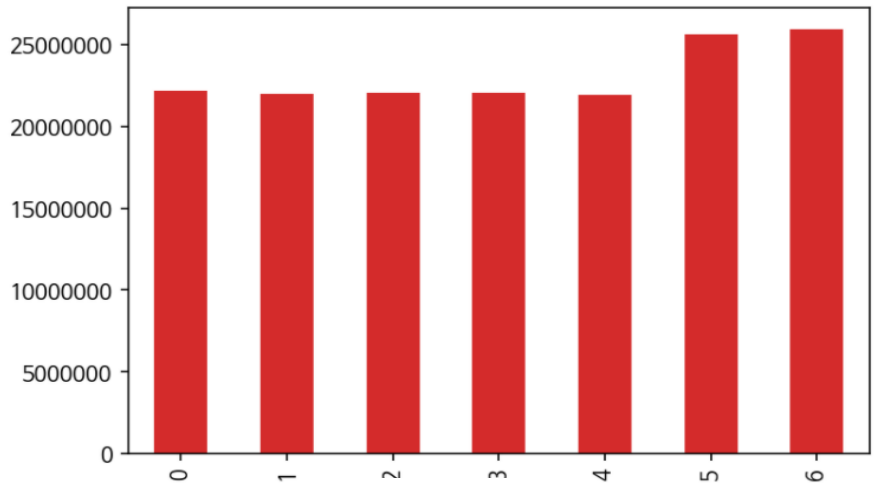


매출액 분포에서 계절성을 확인함 (봄은 매출액이 낮고, 겨울은 높은 편)

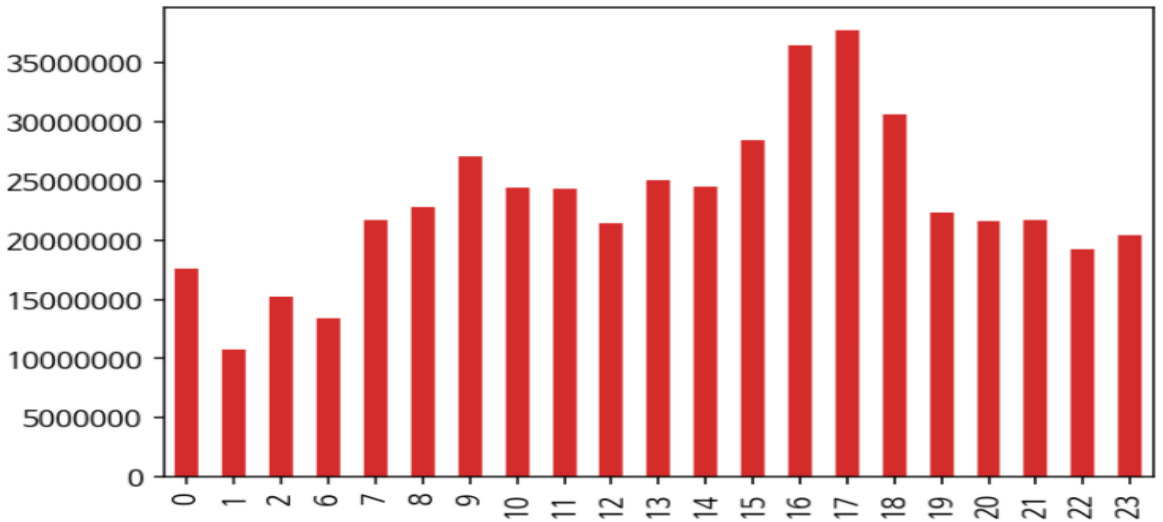
같은 달 내에서도 월초, 월말이 매출액이 높은 편

(1) 방송일시 파생변수

요일별 매출액 분포



시간대별 매출액 분포

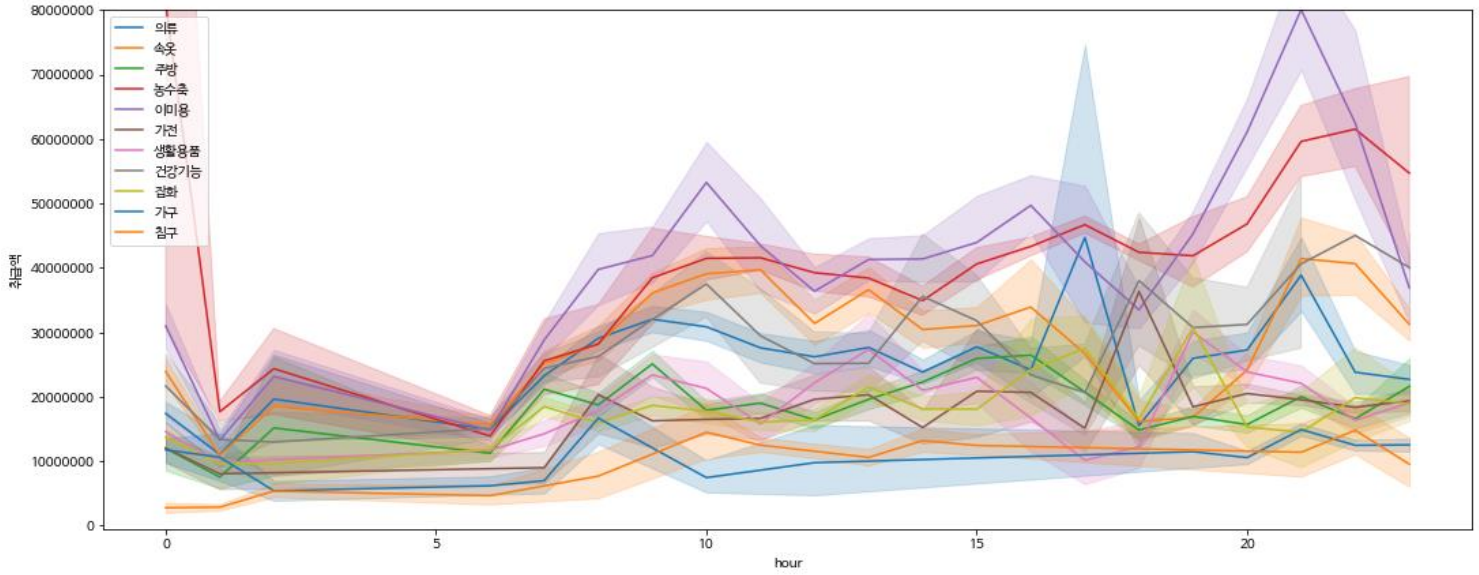


평일 대비 주말이 매출액이 높은 편이며, 공휴일을 포함한 휴일이 평일보다 매출액이 높은 편

평균적으로 16-19시에 매출액이 높은 편

(2) Prime time & Swing time

상품군별 시간대별 매출액 분포

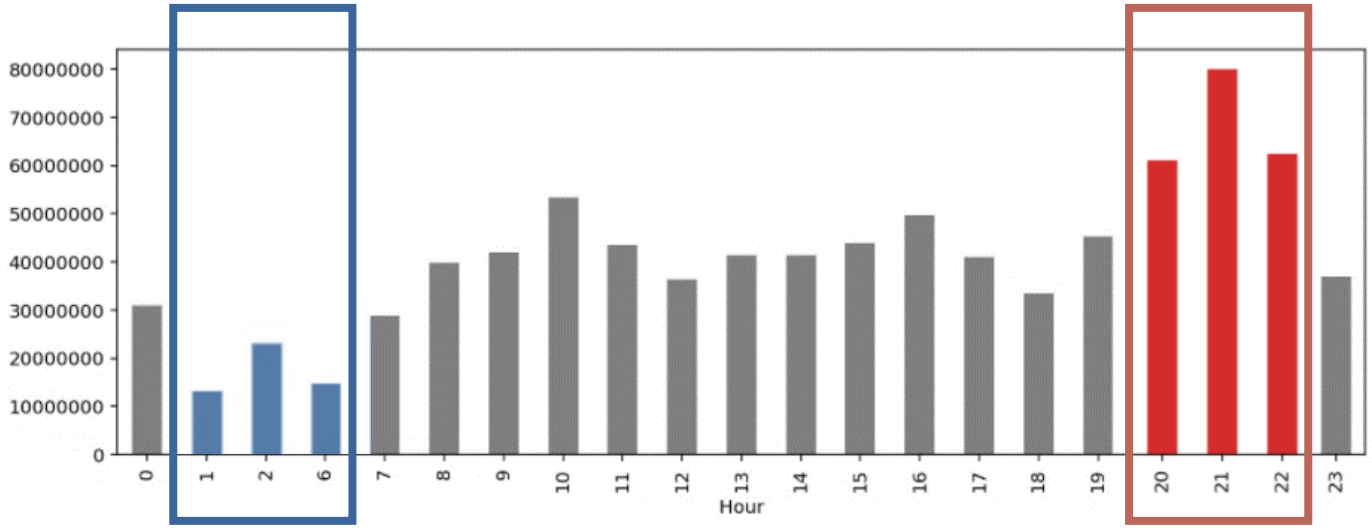


요일별, 상품군별로 잘 팔리는 시간대가 상이함.

따라서 요일별, 상품군별 평균 매출액이 높은 시간대와 낮은 시간대를 각각 Primetime과 Swingtime으로 범주형 변수를 생성함

(2) Prime time & Swing time

Ex. 이미용 시간대별 매출액 분포



Swing time

Primetime

이미용 상품군은 밤시간대 매출액이 높은 편, 심야/새벽시간대 매출액이 낮은 편

(1) 방송 길이 & 방송 순서

Row, 방송길이(노출)는 구분되어 있으나 같은 상품이 연속으로 편성

	방송일시	노출(분)	마더코드	상품코드	상품명	상품군	판매단가	취급액	방송순서	실제노출
0	2019-01-01 06:00:00	20.0	100346	201072	테이트 남성 셀린니트3종	의류	39900	2099000.0	1	10.0
1	2019-01-01 06:00:00	20.0	100346	201079	테이트 여성 셀린니트3종	의류	39900	4371000.0	1	10.0
2	2019-01-01 06:20:00	20.0	100346	201072	테이트 남성 셀린니트3종	의류	39900	3262000.0	2	10.0
3	2019-01-01 06:20:00	20.0	100346	201079	테이트 여성 셀린니트3종	의류	39900	6955000.0	2	10.0
4	2019-01-01 06:40:00	20.0	100346	201072	테이트 남성 셀린니트3종	의류	39900	6672000.0	3	10.0
5	2019-01-01 06:40:00	20.0	100346	201079	테이트 여성 셀린니트3종	의류	39900	9337000.0	3	10.0

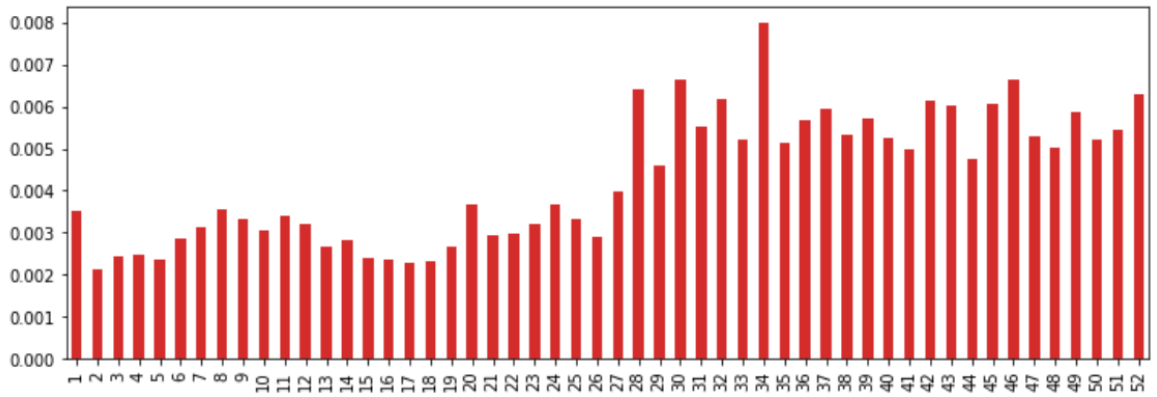
연속으로 편성된 경우를 고려하여 전체 방송 길이를 나타내는 [실제노출] 변수를 생성

또한, 연속으로 편성된 경우 마감시간에 주문이 몰리는 특성때문에 뒤에 편성된 프로그램의 취급액이 높은 경향이 있음

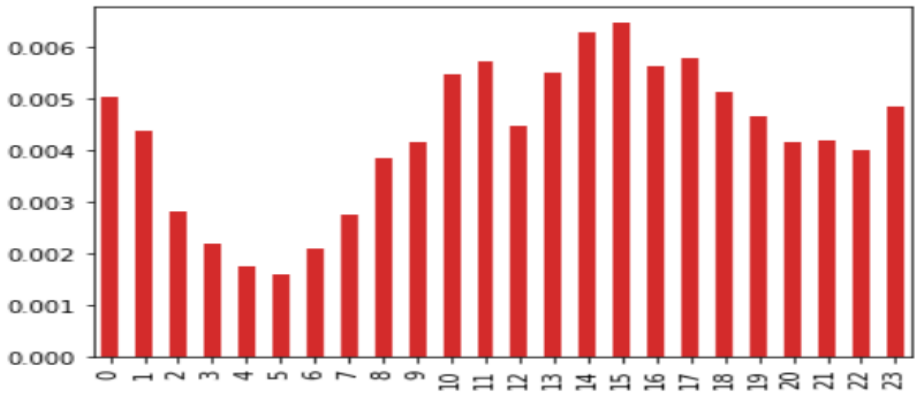
이를 구분할 수 있는 [방송순서] 변수를 생성

(2) 시청률

주차별 평균 시청률 분포



시간대별 시청률 분포



2019년도 상반기보다 하반기에 시청률이 높은 편, 심야시간대 시청률이 낮은 편

전체 요일별로 1시간마다 시청률 합계, 평균, 표준편차 변수를 생성

상품 특성 변수 _ 기존 변수의 한계점

(1) 마더코드 & 상품코드 & 상품명

```
num_list=[]
for num in train_df['마더코드'].unique():
  if num in test_df['마더코드'].unique():
    num_list.append(num)

print('train_df와 test_df와 겹치는 마더코드 개수 :',len(num_list),'개')
print('test_df 마더코드 개수:',len(test_df['마더코드'].unique()),'개')
```

```
train_df와 test_df와 겹치는 마더코드 개수 : 91 개
test_df 마더코드 개수: 225 개
```

겹치지 않는 마더코드 개수 134개

```
num_list=[]
for num in train_df['상품명'].unique():
  if num in test_df['상품명'].unique():
    num_list.append(num)

print('train_df와 test_df와 겹치는 상품명 개수 :',len(num_list),'개')
print('test_df 상품명 개수:',len(test_df['상품명'].unique()),'개')
```

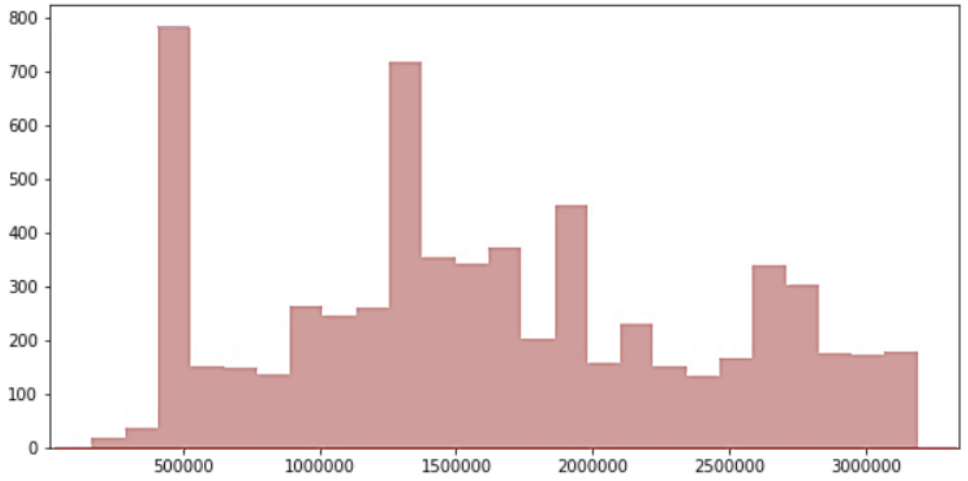
```
train_df와 test_df와 겹치는 상품명 개수 : 46 개
test_df 상품명 개수: 377 개
```

겹치지 않는 상품명 개수 377개

2019년도 데이터에 존재하지 않는
새로운 마더코드/상품코드(상품명)가 등장함

(2) 상품군

가전 판매단가 분포



같은 상품군 내에서도 취급액, 판매 단가의 차이가 큼
브랜드, 상품 특성을 구체적으로 반영하기 어려움

상품 특성 변수 _ 기존 변수의 한계점

```
num_list=[]
for num in train_df['마더코드'].unique():
  if num in test_df['마더코드'].unique():
    num_list.append(num)

print('train_df와 test_df와 겹치는 마더코드 개수 :',len(num_list),'개')
print('test_df 마더코드 개수:',len(test_df['마더코드'].unique()),'개')
```

```
train_df와 test_df와 겹치는 마더코드 개수 : 91 개
test_df 마더코드 개수: 225 개
```

```
num_list=[]
for num in train_df['상품명'].unique():
  if num in test_df['상품명'].unique():
    num_list.append(num)

print('train_df와 test_df와 겹치는 상품명 개수 :',len(num_list),'개')
print('test_df 상품명 개수:',len(test_df['상품명'].unique()),'개')
```

```
train_df와 test_df와 겹치는 상품명 개수 : 46 개
test_df 상품명 개수: 377 개
```

가전 판매단가 분포

1. 상품군을 세분화하는 파생변수 생성

2. (1)의 파생변수를 기반으로 유사한 마더코드/상품코드으로 대체

2019년도 데이터에 존재하지 않는
새로운 마더코드/상품코드가 등장함

같은 상품군 내에서도 취급액, 판매 단가의 차이가 큼
브랜드, 상품 특성을 구체적으로 반영하기 어려움

(3) 중분류

Konlpy

상품명 토큰화

→ 상품군별 핵심 키워드 딕셔너리 생성

Ex. 건강기능

'석류', '유산균', '착즙', '오메가','락토핏', '구미', '루테', '다이어트', '프리바이오틱스', '진액', '홍삼','팥물', '티톡', '비트', '양배추', '전립','두유','보틀', '홍합','비타민', '철갑상어', '히비스커스', '분말','해죽', '바이오'

Word2Vec

토큰 벡터화 → 가중치 매트릭스 X TDM

→ 연관스코어 행렬 구축

Ex. 침대

name	하이바스	침대	가구
(무) 삼익가구 LED 제니비 서랍형 침대 Q	0.127123	0.081164	0.083505
(일) 삼익가구 LED 제니비 서랍형 침대 K	0.127123	0.081164	0.083505
(무) 삼익가구 LED 제니비 서랍형 침대 K	0.127123	0.081164	0.083505
(일)보루네오 유로탑 가죽 침대 슈퍼싱글	0.039567	0.000000	0.039983
(무)보루네오 유로탑 가죽 침대 슈퍼싱글	0.039567	0.000000	0.039983

K-means Clustering

상품명 유사도 기반 군집화

Ex. 암막 커튼

지나송 보노 화이트에디션 암막 커튼(슈퍼특대형)

한스데코 샤를 이중 암막 레이스 커튼(슈퍼특대형)

한스데코 샤를 이중 암막 레이스 커튼(특대형)

한스데코 샤를 이중 암막 레이스 커튼(대형)

상품군을 보다 세분화하기 위해 상품명이 비슷한 상품끼리 군집화한 중분류 변수 생성

(4) 브랜드

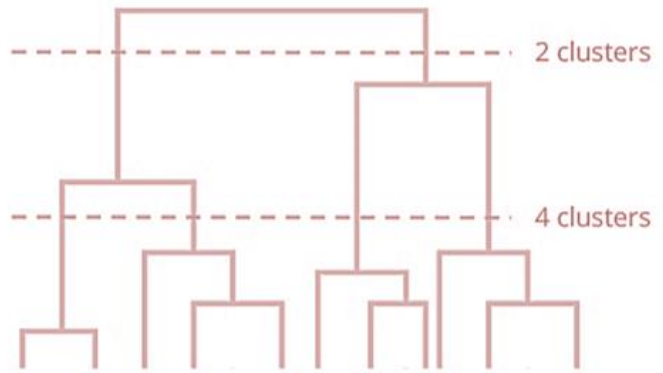
Levenshtein distance

두 문자열 간의 Sequential Distance를 계산하여
유사도를 판단하는 기법

	{}	ㄷ	ㅗ	ㅇ	ㄱ	ㅠ
{}	0	1	2	3	4	5
ㄷ	1	0	1	2	3	4
ㅣ	2	1	1	2	3	4
ㅇ	3	2	2	1	2	3
ㄴ	4	3	3	2	2	3
ㅛ	5	4	4	3	3	3

Hierarchical Clustering

비슷한 특징을 가진 데이터들을 묶어나가며
계층적으로 군집을 구성하는 기법



브랜드 특성을 반영하기 위해 Levenshtein distance를 활용하여 구한 상품명간의 유사도를 바탕으로 클러스터링 변수를 생성

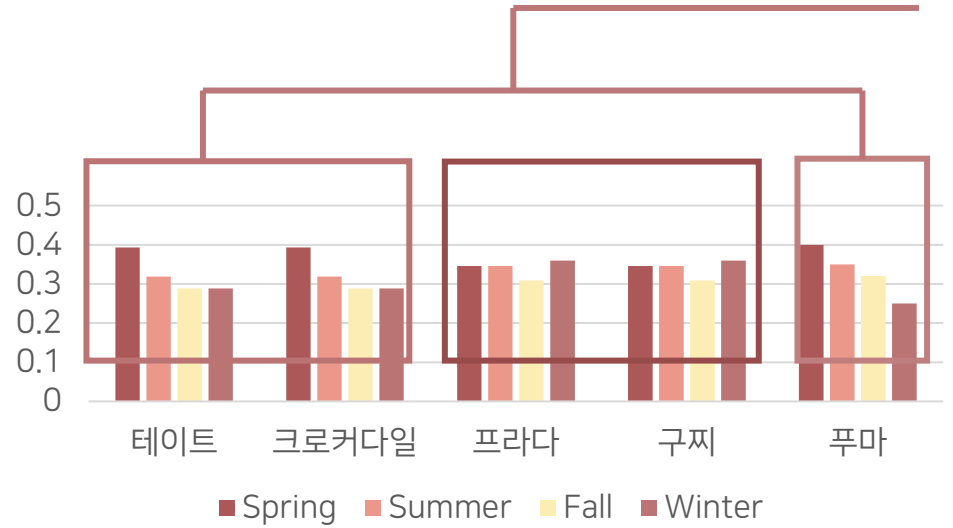
(5) Jenson Shannon Clustering

Jenson-Shannon Divergence

$$JSD(P,Q) = JSD(Q,P)$$

KL divergence를 기반
두 분포의 유사도를 측정

Hierarchical Clustering



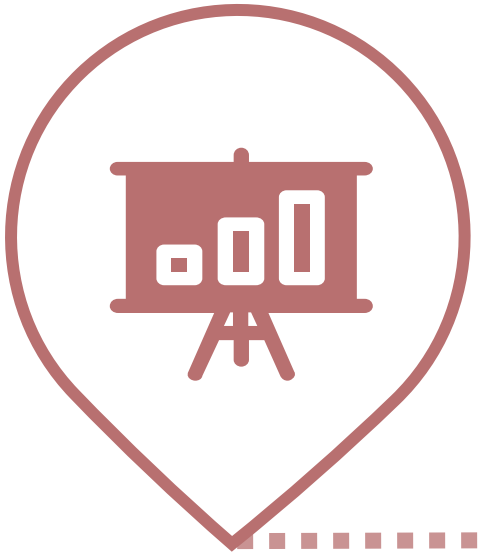
서로 다른 브랜드, 상품이더라도 비슷한 매출액 추이를 보이는 상품끼리 묶어주기 위해
시계열 변수(요일, 시간 등)에 따라 비슷한 매출액 분포를 가진 브랜드, 중분류를
계층적 군집 클러스터링 기법을 사용하여 군집화

날씨 : 기온, 강수량, 풍량, 미세먼지 통계량 변수

미세먼지가 많은 시기에는
마스크, 공기청정기 등이 자주 편성되고
매출액이 높을 것

비가 오는 날에는
외출을 자제하는 경우가 많기 때문에
평소보다 매출액이 높을 것

날짜별/ 시간대별 날씨 데이터의 통계량 변수를 생성
(지역별 분포는 유사하여 구분하지 않음)



모델링

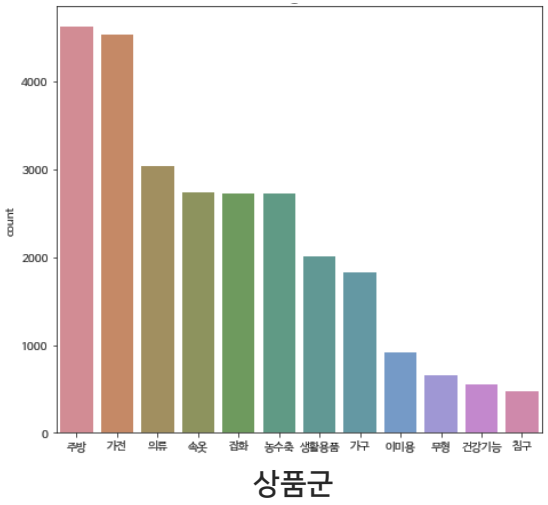
모델링

스케줄링

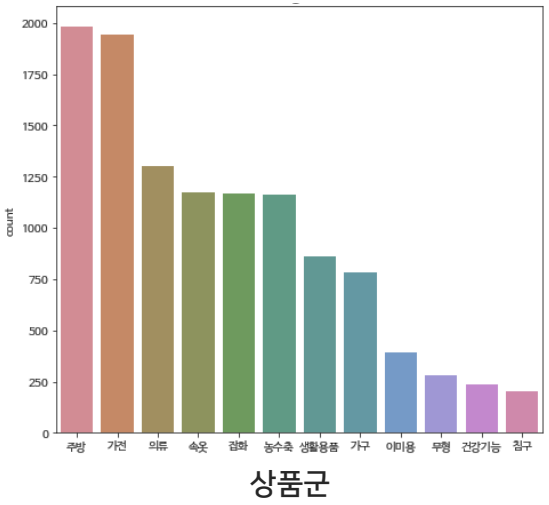
Validation

Train Set : 70% | Validation Set : 30%

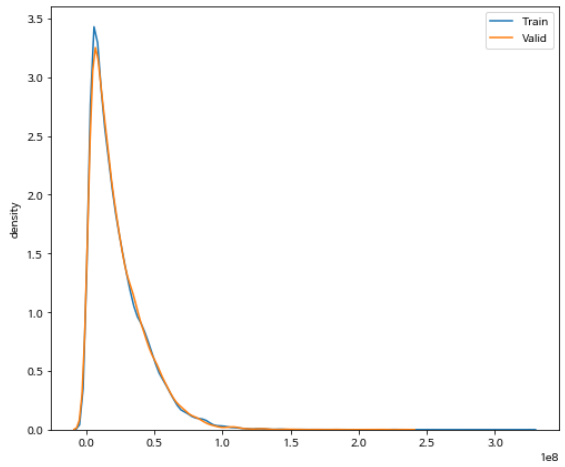
Train Set 상품군 Count



Valid Set 상품군 Count

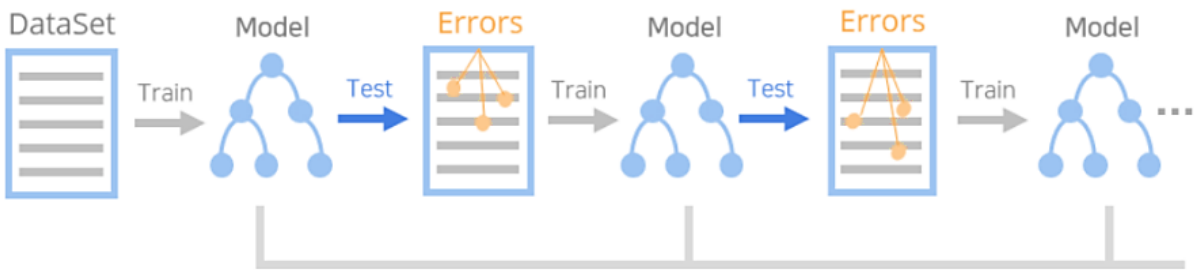


Train & Valid Set 취급액 dist.



상품군, 취급액 등의 분포를 고려하여 Train Set과 Validation Set을 분리

Light GBM (Gradient Boosting Model)



Decision Tree를 순차적으로 학습-예측하면서
잘못된 데이터에 가중치를 부여하고
반복학습을 하며 오류를 개선해 나가는 알고리즘

Why?

많은 변수를 사용하여 예측력을 높이기 위함
변수 간의 다중공선성을 고려하지 않는 모델을 사용하기 위함
기존 Gradient Boosting Model에 비해 더 빠르게 처리할 수 있음

모델 해석

Feature Importance

특정 변수가 트리를 분할하는데 기여하는 정도

Ex. GINI INDEX

$$I_G = 1 - \sum_{j=1}^{c} p_j^2$$

p_j: proportion of the samples that belongs to class c for a particular node

SHAP

예측값에 대한 특성들의 기여도

$$g(z') = \phi_0 + \sum_{j=1}^{M} \phi_j z'_j$$

g : 설명 모델

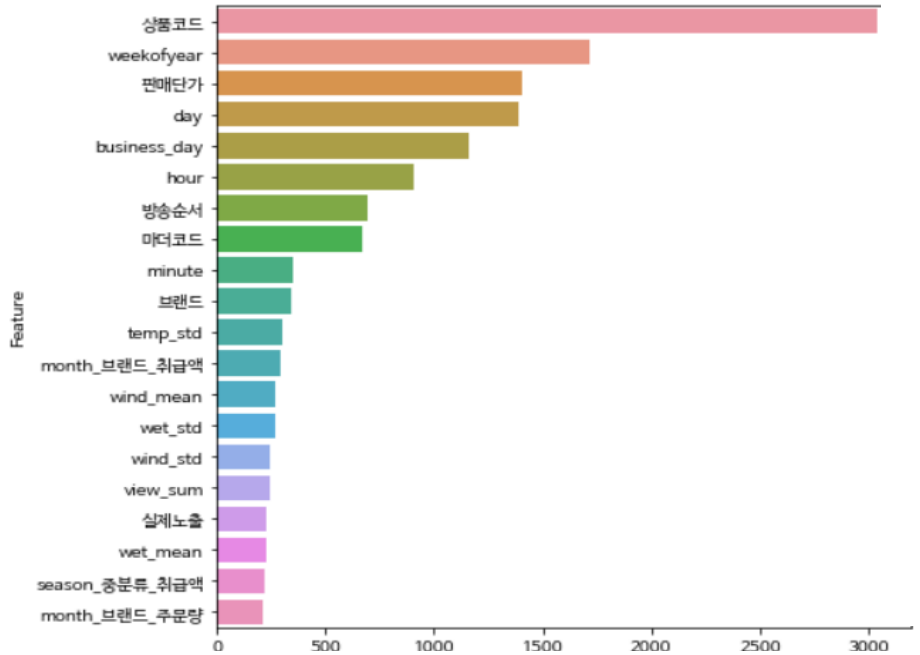
$\emptyset_j \in R$: 특성 j의 특성 기여도 (Shapley value)

$z' \in \{0,1\} M$: 연합 벡터

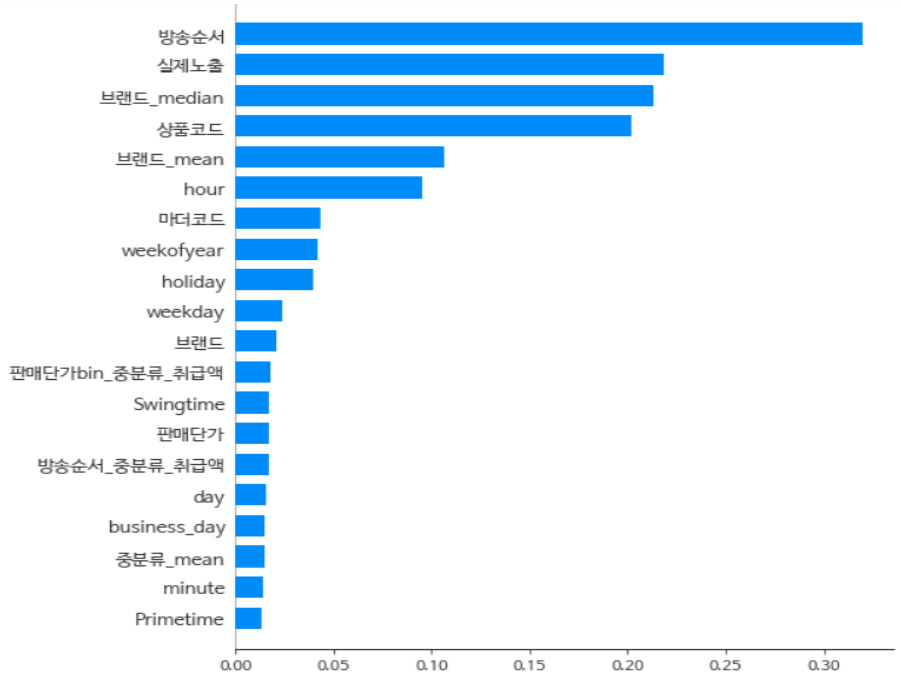
M : 최대 연합 사이즈

모델 해석

Feature Importance



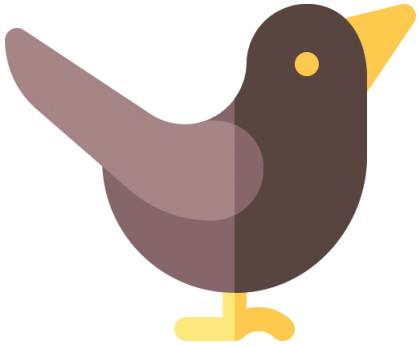
SHAP Value



Feature Importance 에서는 상품 특성 변수와 시계열 특성 변수들이 주요한 영향을 미치는 것으로 나타났다.

SHAP Value 에서는 방송 특성 변수와 시계열 특성 변수들이 주요한 영향을 미치는 것으로 나타났다.

Cuckoo Search를 이용한 방송편성표 최적화



Cuckoo Search

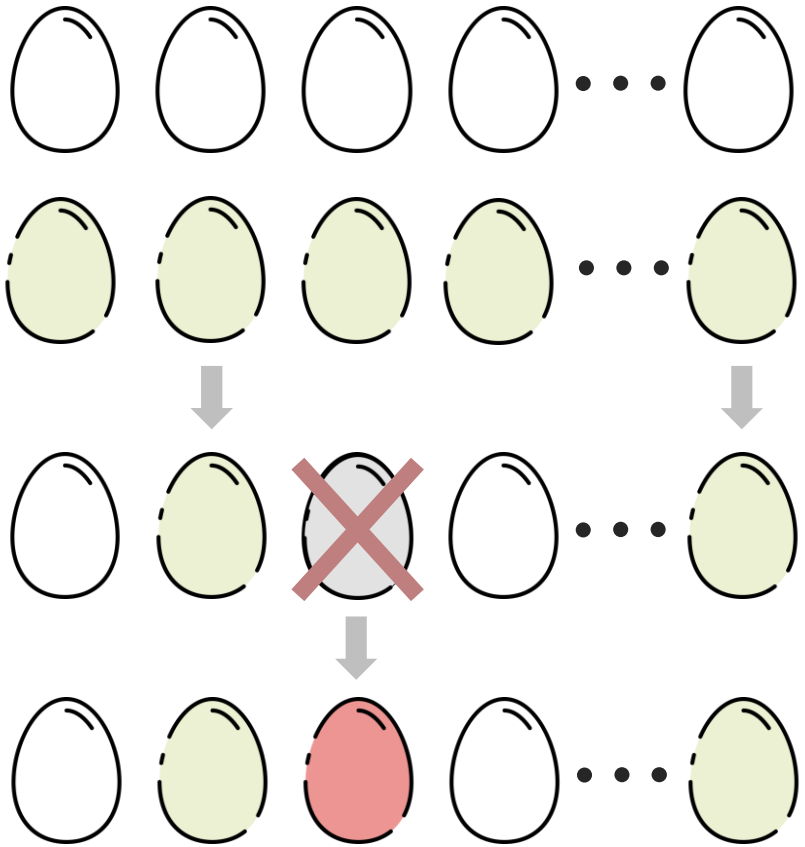
전역최적해를 효과적으로 탐색하기 위해

뻐꾸기의 번식 전략에서 착안한 메타휴리스틱 알고리즘

$$x_i^{(t+1)} = x_i^{(t)} + \alpha \oplus Levy(\boldsymbol{\lambda})$$

New Solution = Current Solution + Transition Probability

Cuckoo Search를 이용한 방송편성표 최적화



Initial Solution

New Solution

Updated Solution

Compare

기존의 알 (Initial Solution)과 비교하여 더 적합도가 높을 경우,
뻐꾸기 알(New Solution)이 선택됨

Evaluation &
Replacement

전체 알 중 적합도가 낮은 하위 알들이 뻐꾸기의 알이라고 판단되어 둥지에서 버려지고 새로운 알을 랜덤으로 생성함

Cuckoo Search를 이용한 방송편성표 최적화

Why?

- 데이터 분석을 통해 만든 모델을 기반으로 최적의 방송편성표를 탐색하는 Cuckoo Search based Scheduler(CSS)를 고안
- Parameter의 수가 적고, 구조가 간단하기 때문에 다른 메타휴리스틱 알고리즘보다 복잡성이 낮고, 세대 당 탐색 속도가 빠른 편
- Levy distribution을 이용하기 때문에 Global optimum을 찾는데 보다 효율적

Asumption

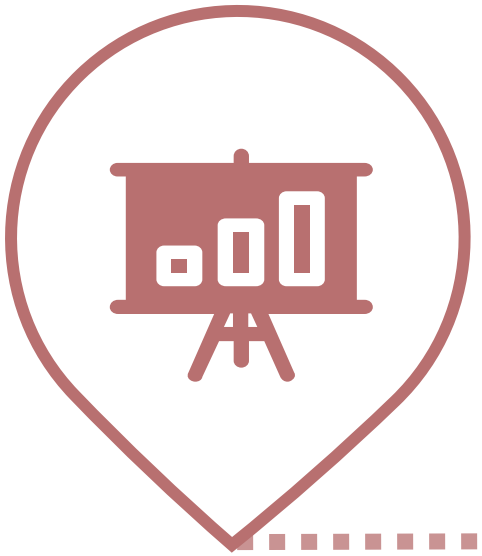
- 취급액을 예측하는 Model의 신뢰성이 중요하기 때문에 Model을 전적으로 신뢰하는 것을 전제로 함.
- 제약조건을 단순화하기 위해 노출시간은 20, 40, 60분만으로, 특정 분포를 이용하여 재구성하였음.
- 실제 데이터(20년 6월 데이터)의 row를 추출해내서 노출시간만 바꾸어 순서를 재구성하기 때문에 해당 데이터로만 편성표가 이루어지는 것을 전제로 함.

Cuckoo Search를 이용한 방송편성표 최적화

	방송일시	노출(분)	마더코드	상품코드	상품명	상품군	판매단가	취급액
0	2020-06-01 06:20:00	20	100731	201962	월드컵 S/S 남성 에어러닝화+패션슬리퍼	잡화	29800	11062000.0
1	2020-06-01 06:40:00	20	100353	202085	마르엘라로사티 린넨 베스트 세트[3월]	의류	39900	14200000.0
2	2020-06-01 07:20:00	40	100097	200762	무이자 올리고 가스와이드그릴레인지 프리미엄형 +버팔로 캠핑쿨러백	주방	129000	12896000.0
3	2020-06-01 08:20:00	60	100117	200996	신일써큘레이터 스탠드 블랙+블랙(SIF-R09DBK)	생활용품	268000	16228000.0
4	2020-06-01 08:40:00	20	100290	201780	클라쎄 벽걸이 에어컨 MKRA06DTB	가전	449000	7468000.0
5	2020-06-01 09:00:00	20	100511	202483	국내산 손질오징어 150g * 15팩	농수축	45900	30528000.0
6	2020-06-01 09:20:00	20	100808	201223	CERINI by PAT 남성 에어홀 팬츠3종	의류	59000	25495000.0
7	2020-06-01 09:40:00	20	100203	201195	(일)[보루네오] 피올레 천연소가죽 소파 6인용	가구	1199000	7138000.0
8	2020-06-01 10:20:00	40	100078	201999	발레리 프론트후크 레이스 브라렛 5세트(18차)	속옷	69900	22956000.0
9	2020-06-01 10:40:00	20	100262	200876	더블모 어성초 샴푸	이미용	59800	25361000.0
10	2020-06-01 11:00:00	20	100099	202000	라쉬반 FC바로셀로나 드로즈 8종	속옷	119000	9062000.0
11	2020-06-01 11:20:00	20	100573	201776	무이자 20년 무풍 슬림 16형 화이트(절전) 스탠드(AF16T5774WZT) + ...	가전	1799000	13408000.0
12	2020-06-01 12:20:00	60	100575	201781	일시불 삼성 UHDTV 75형 KU75UT7000FXKR + 삼성 사운드바HW-T450	가전	2100000	11371000.0
13	2020-06-01 12:40:00	20	100139	201375	보몽드 실루엣 쿨시어서커 침구 풀세트 Q(퀸)	침구	54900	5886000.0
14	2020-06-01 13:00:00	20	100226	202011	*[맥널티] 싱글 오리진 핸드드립 커피 5종 10박스 + 커피포트 1개	농수축	39900	23411000.0

충분히 실현 가능하고 다양한 편성표를 생성해낼 수 있었다.

실제 편성표와는 다소 fitness의 차이가 있었지만, Random Search보다는 약 25%만큼 더 좋은 성능을 내주었다.



결론

시사점

매출액 예측 모델 시사점

- 판매하는 상품과 방송하는 시간이 매출에 영향을 많이 끼친다는 것을 알 수 있음
- 브랜드, 세분화된 카테고리 등의 데이터를 축적함으로써 향후 최적화된 스케줄링 알고리즘을 구현하는 데 기여할 수 있을 것
- 빠르고 가벼운 알고리즘을 활용함으로써 학습과 예측시간을 5분 이내로 단축시킴

방송 편성 알고리즘 추후 연구 및 이용 방안

- EDA를 통해 특정 시간의 주요고객층이 선호하는 상품을 미리 고정한 후 초기화한다면 초기해의 적합도를 끌어올려 전체적인 성능을 향상시킬 수 있을 것
- 더 다양한 실제 세계의 데이터와 자세한 제약조건이 주어진다면 더 정교한 스케줄러를 만들 수 있을 것
- Optimizer들을 방송편성표라는 도메인에 맞는 형태로 변환시킨다면 더 좋은 탐색능력을 가질 것
- NS에서 스케줄 편성에 사용되는 노하우 및 정보와 함께 제안된 CSS를 혼합해서 사용한다면 더 좋은 편성표를 만들 수 있을 것
- CSS를 통해 취급액에 대해 우리가 쉽게 볼 수 없는 Feature를 찾아내서 다양한 데이터분석에 활용할 수 있는 여지가 있음

감사합니다