2020 빅콘테스트
2020 BIG CONTEST

NS홈쇼핑

2020 빅콘테스트 데이터분석 분야 챔피언리그

NS SHOP+ 판매실적 예측을 통한 편성 최적화 방안 도출

insomNIA _ 기세현(jape908@naver.com), 민규선(vkvkvk2892@naver.com), 하진용(eevee7075@naver.com)

CONTENTS

1 분석배경

- 분석 배경
- 데이터 소개
- 분석 주제 및 과정

2 데이터탐색

- 데이터 전처리
- EDA

3 변수생성

- 파생변수
- 외부변수

4 모델링

- 데이터 분할
- 모델링
- 성능비교

5 결론 및 활용방안

- 결론
- 활용방안

분석 배경

1

분석 배경

코로나 바이러스(COVID-19)의 확산을 방지하기 위해
사회적 거리두기 캠페인이 진행되면서
'언택트(Untact)'라는 트렌드가 생겨남

- 언택트(Untact) = 부정의 의미를 담은 접두사(Un-) + 접촉하다(Contact) = 사람과 접촉하지 않는 행태

홈쇼핑 또한 비대면으로 물품을 구매하는 언택트 소비에 해당

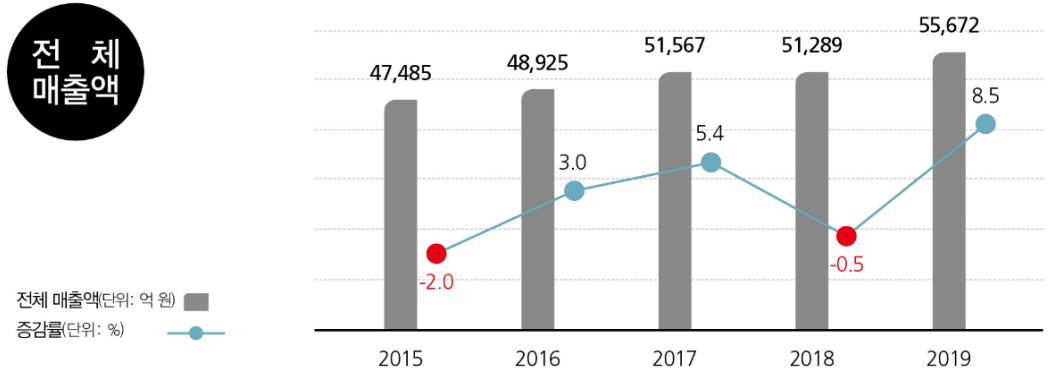
O2O IoT AI + COVID-19

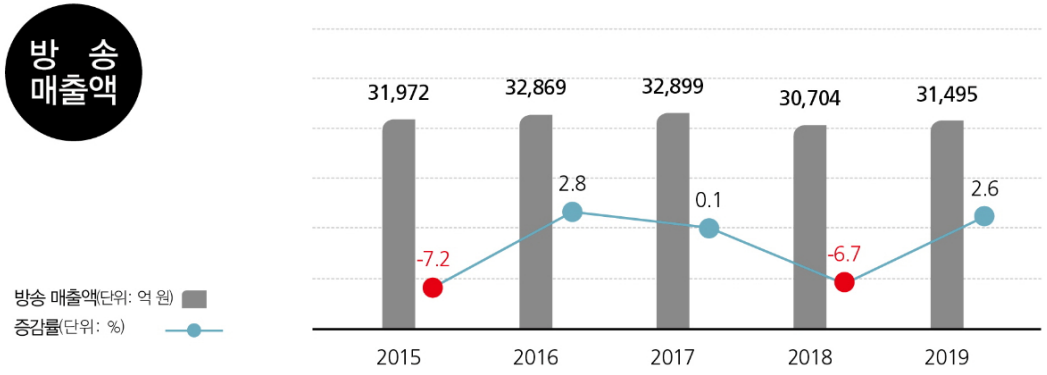
= 언택트 마케팅의 중요성 ⬆

분석 배경

매출액 : TV 홈쇼핑사의 직매입 상품 매출과 협력사로부터 받는 판매수수료 매출 등 매출 합계

출처 : 한국TV홈쇼핑협회





전체 매출액이 증가하는 추세이긴 하지만 방송 매출액은 아직 안정되지 않은 모습

방송 편성 최적화로
방송 매출액을 극대화할 수 있는 방안 마련 필요

데이터 소개

방송일시	노출(분)	마더코드	상품코드	상품명	상품군	판매단가	취급액
2019-01-01 6:00	20	100346	201072	테이트 남성 셀린니트3종	의류	39,900	2,099,000
2019-01-01 6:00		100346	201079	테이트 여성 셀린니트3종	의류	39,900	4,371,000
2019-01-01 6:20	20	100346	201072	테이트 남성 셀린니트3종	의류	39,900	3,262,000
⋮							
2020-01-01 1:20	20	100490	201478	더케이 예다함 상조서비스(티포트)	무형	-	
2020-01-01 1:40	17	100490	201478	더케이 예다함 상조서비스(티포트)	무형	-	

2019년 1개년 NS SHOP+ 판매실적 데이터

8개 변수, 38309개 행 존재

노출(분), 판매단가, 취급액 변수에 결측값 존재

idea

- 취급액 예측 문제
- 취급액은 매우 큰 분산을 가지고 있음. 또한 상품별로 판매단가가 상이하기 때문에 취급액의 단위가 동등하지 않다고 판단.

  ➡ '판매단가 × 판매개수 = 취급액'을 이용하여 취급액 예측 문제를 판매개수 예측 문제로 전환.

분석 주제 및 과정

주제1 NS SHOP+ 판매실적 예측

제공된 변수 및 파생변수, 외부변수를 탐색하여 매출에 큰 영향을 주는 요인을 파악하고 판매실적 예측

주제2 방송 편성 최적화 방안 도출

요일별, 시간대별, 카테고리별 시청률 및 판매실적을 탐색하여 최적 수익을 고려한 편성 최적화 방안 제시

분석과정요약

분석을 통해 매출액에 큰 영향을 주는 요인들을 파악하고 다양한 방송 편성 방안 제시

데이터 탐색

2

데이터 전처리

〈 결측값 처리 〉 ※ 예측 상품 중 판매가 0인 프로그램 실적은 예측에서 제외함 – 데이터설명서

	노출(분)	판매단가	취급액
결측값 개수	16784	937	2930

취급액 0 제거 →

노출(분)	판매단가	취급액
14976	0	0

- 노출(분)

방송일시	노출(분)	마더코드	상품코드	상품명
2019-01-02 10:00	20	100448	202098	일시불 쿠첸 풀스텐 압력밥솥 10인용 (A1)
2019-01-02 10:00	20	100448	202093	무이자 쿠첸 풀스텐 압력밥솥 10인용(A1)
2019-01-02 10:00	20	100448	202100	일시불 쿠첸 풀스텐 압력밥솥 6인용(A1)
2019-01-02 10:00	20	100448	202095	무이자 쿠첸 풀스텐 압력밥솥 6인용(A1)

– 동일 방송일시 기준 첫 번째 행의 노출(분)에만 값이 존재하여 방송일시가 같으면 같은 값의 노출(분)으로 기입

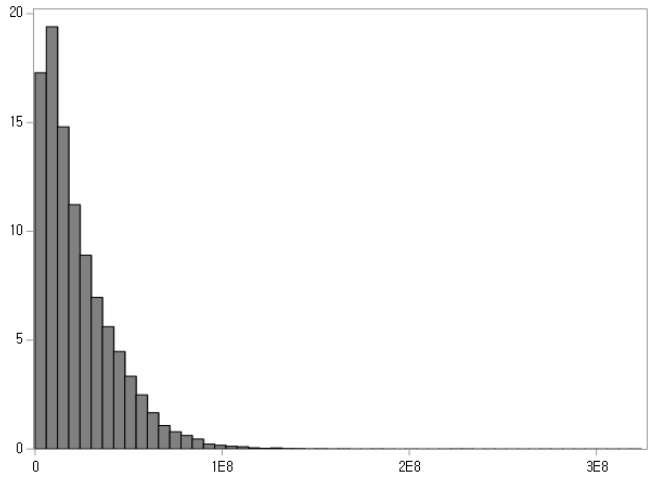
〈 새벽 시간대 처리 〉 방송시간대가 0~2시면 요일 – 1요일

방송일자	방송시간	요일
2019-11-30	23:40	토요일
2019-12-01	0:00	일요일
2019-12-01	1:40	일요일
2019-12-01	2:00	일요일

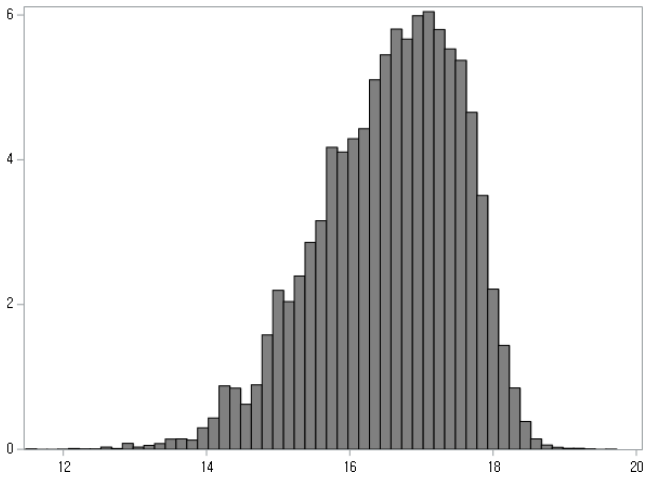
→

방송일자	방송시간	요일
2019-11-30	23:40	토요일
2019-12-01	0:00	**토요일**
2019-12-01	1:40	**토요일**
2019-12-01	2:00	**토요일**

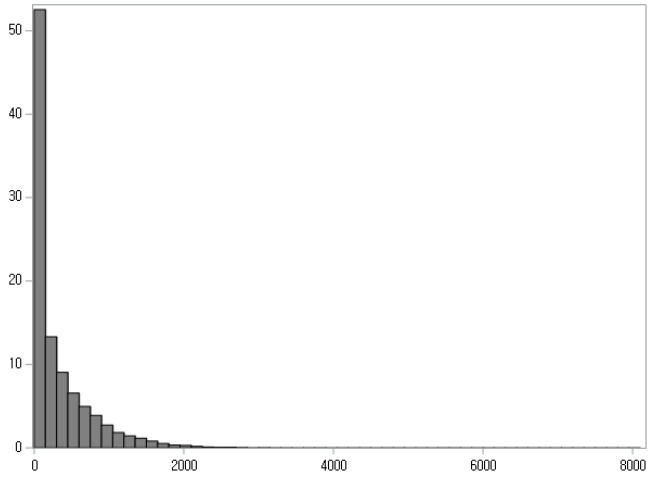
EDA – 취급액



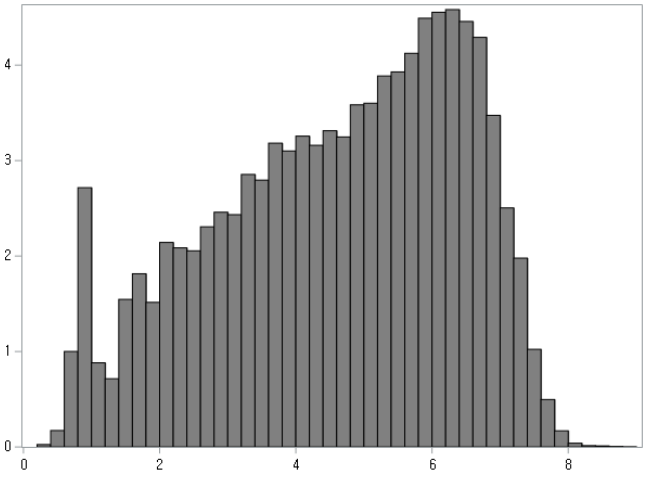
〈 취급액 〉



〈 로그_취급액 〉



〈 판매개수 〉



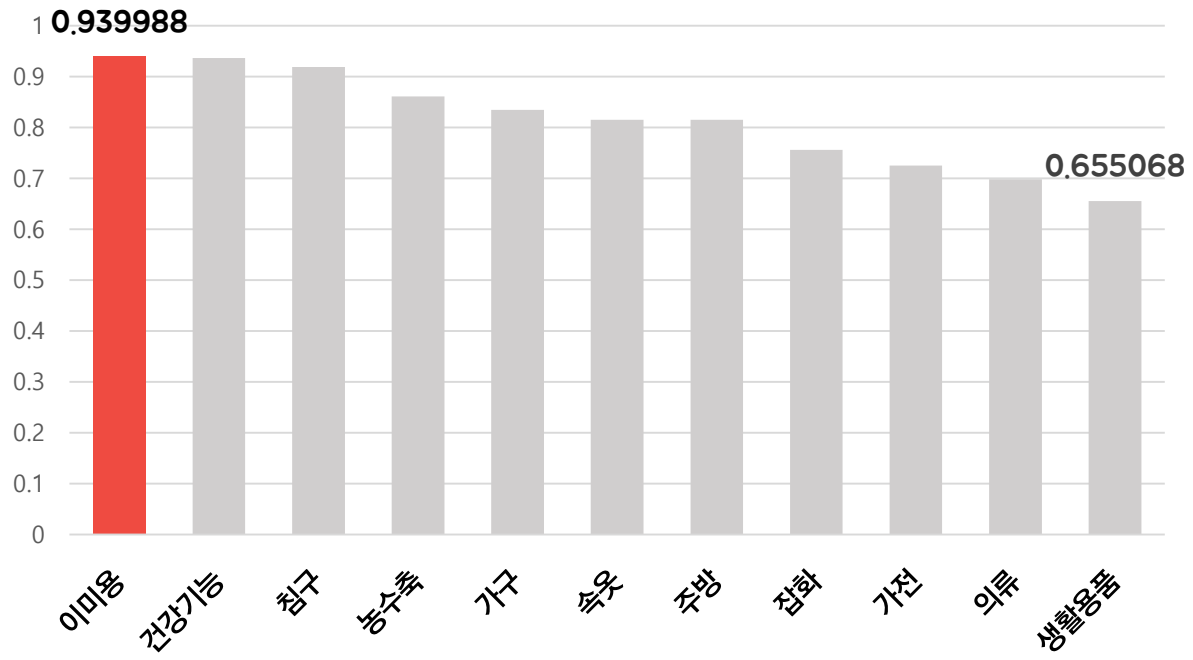
〈 로그_판매개수 〉

	취급액	판매개수
최솟값	103000.00	1.30
평균	23102409.23	314.80
중앙값	17326000.00	130.00
최댓값	322009000	8070.40
분산	4.0231955E14	183035.52
표준편차	20057904.94	427.83
왜도	1.84	2.48
첨도	7.29	11.72

- 취급액은 분산이 매우 크기 때문에 '판매단가 × 판매개수 = 취급액 ' 을 이용하여 취급액 대신 판매개수 예측 문제로 전환.
- 취급액과 판매개수의 분포가 비슷한 것을 확인.
- 분포가 왼쪽으로 치우쳐져 있기 때문에 로그를 취해준 판매개수를 예측.

EDA – 취급액

상품군별 판매개수와 취급액 상관계수

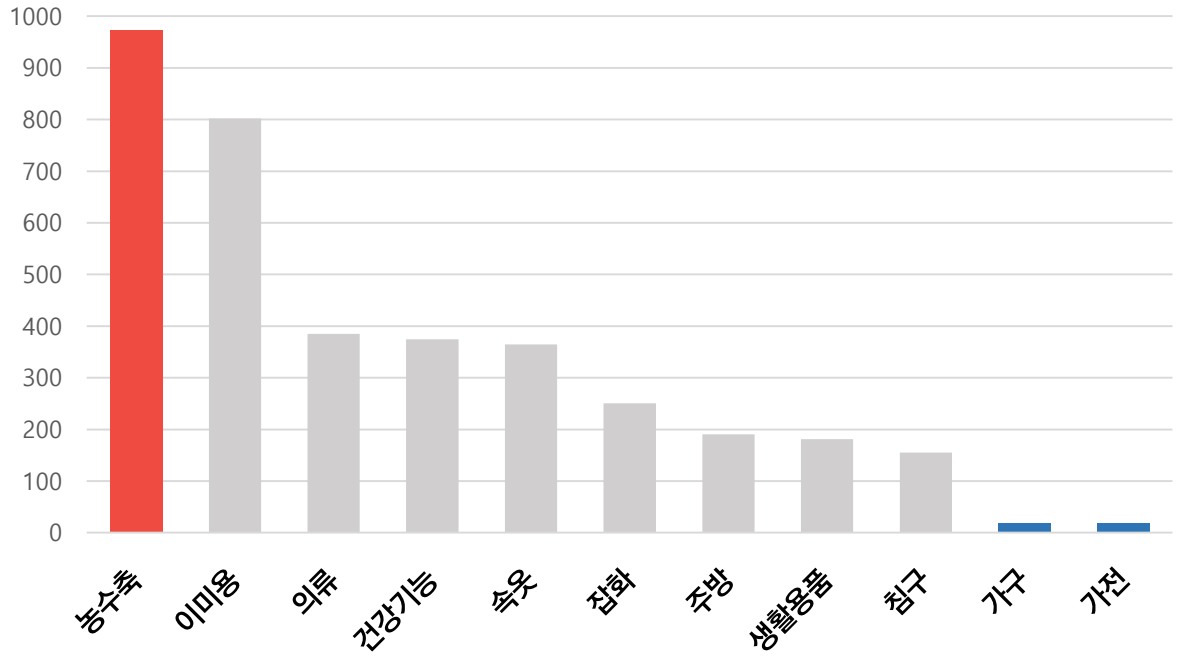


상품군	상관계수
가구	0.834417
가전	0.724849
건강기능	0.936521
농수축	0.860817
생활용품	0.655068
속옷	0.814799
의류	0.698089
이미용	0.939988
잡화	0.755835
주방	0.814799
침구	0.918597

- 전체에서 판매개수와 취급액 간의 상관계수는 0.7631로 매우 높음.
- 이미용 상품군에서 판매개수와 취급액의 상관계수는 0.94로 매우 높고, 상관계수가 가장 작은 상품군인 생활용품과도 0.66의 작지 않은 양의 상관성을 보임.
- 상관성이 작은 상품군일수록 고객들이 할인혜택을 많이 받고 구매하는 것으로 추측.

EDA – 판매개수

상품군별 평균판매개수

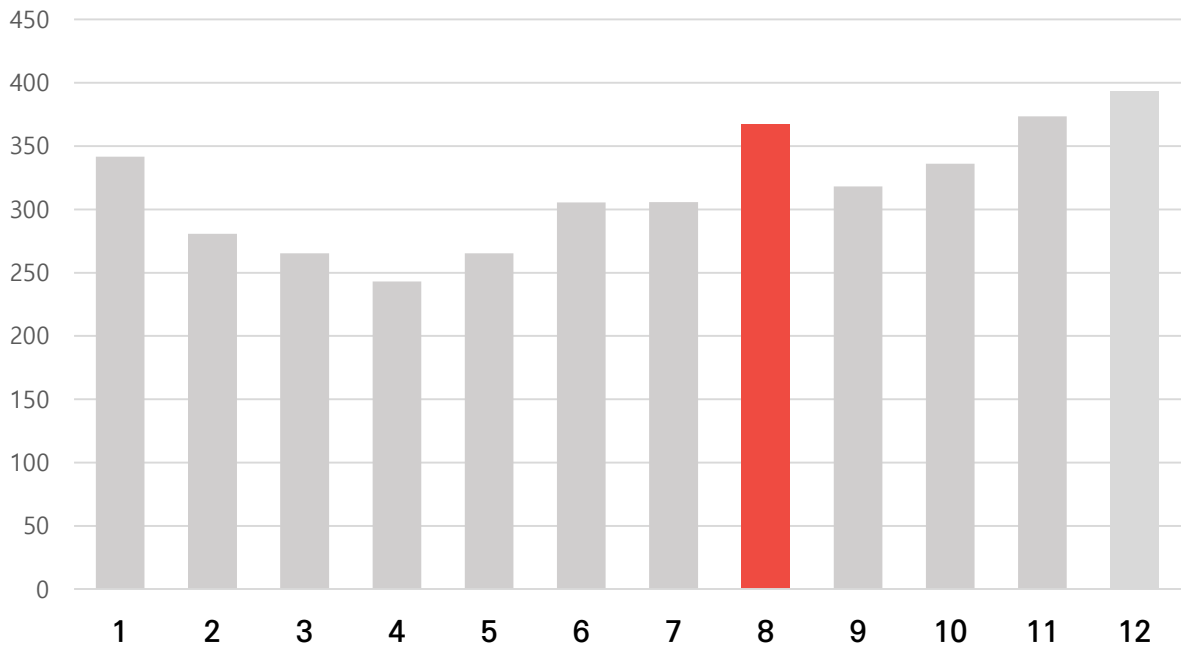


상품군	방송횟수	총판매개수	평균판매개수
가구	2302	45413.56	19.72787
가전	5163	100156.7	19.39893
건강기능	786	294461.8	374.6334
농수축	3884	3776003	972.1944
생활용품	2769	501484.7	181.1068
속옷	3910	1425039	364.46
의류	4331	1667274	384.9629
이미용	1305	1046214	801.6963
잡화	3694	925214.5	250.4641
주방	6571	1252989	190.6847
침구	664	102963.5	155.0655

- 농수축 상품군이 압도적으로 높은 판매량을 보임.
- 농수축 상품군의 평균판매개수는 972.19, 가전의 평균판매개수는 19.73으로 상품군별로 평균판매개수의 편차가 심함.
- 주방 상품군의 방송횟수는 6571, 침구 상품군의 방송횟수는 664로 상품군별로 방송횟수의 편차도 심함.

EDA - 판매개수

월별 평균판매개수



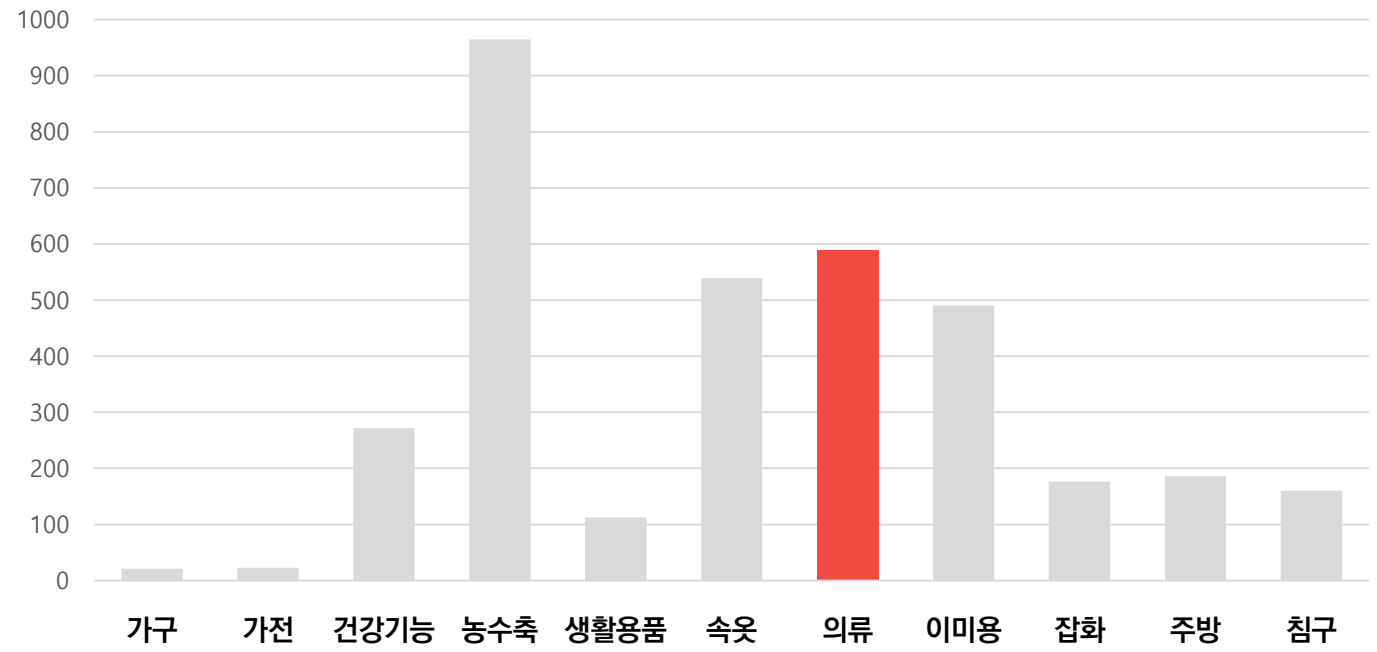
월	방송횟수	총판매개수	평균판매개수
1	2765	944670.8	341.6531
2	2674	750488.5	280.6614
3	3073	815562.6	265.3962
4	3132	761012.1	242.9796
5	3280	870279.8	265.3292
6	2860	874016.5	305.6002
7	3132	957821.4	305.8178
8	2902	1065108	367.0256
9	2982	948371.4	318.032
10	2974	999160.2	335.9651
11	2707	1011109	373.5165
12	2898	1139613	393.2413

- 연말로 갈수록 판매량이 많아짐.
- 8월에 판매량이 많은 이유는 휴가철에 해당하고, 9월에 있는 추석을 준비하기 때문에 평소보다 지출이 많은 시기.

EDA – 판매개수

월별 상품군별 평균판매개수

〈 8월 〉



- 모든 달에서 판매량 1위는 농수축 상품군, 2위는 이미용 상품군이지만 8월만 의류 상품군 판매량이 두 번째로 많음.
- 8월은 휴가와 추석 준비로 평소보다 지출이 많은 달이기 때문에 이미용 상품군에 지출을 줄이고 명절선물세트에 해당되는 농수축 상품군이나 휴가를 준비하며 의류 상품군에 지출하는 것으로 추측.

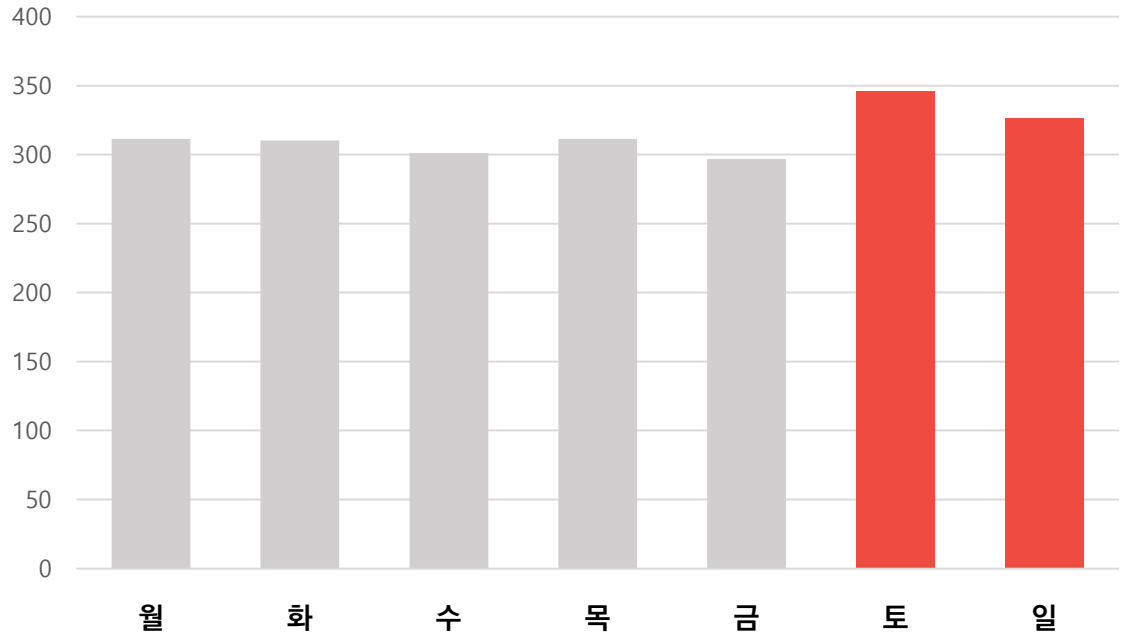
EDA – 판매개수

주차별 평균판매개수

- 추석이나 설날과 같은 명절 2주 전에는 판매량이 매우 높고 명절이 있는 주에는 판매량이 급격히 떨어짐.
- 명절 준비를 2주 전부터 하고, 홈쇼핑에서 명절 선물을 많이 구매하는 것으로 추측.
- 명절 이후에는 판매량이 꾸준히 증가하는 형태.

EDA – 판매개수

요일별 평균판매개수

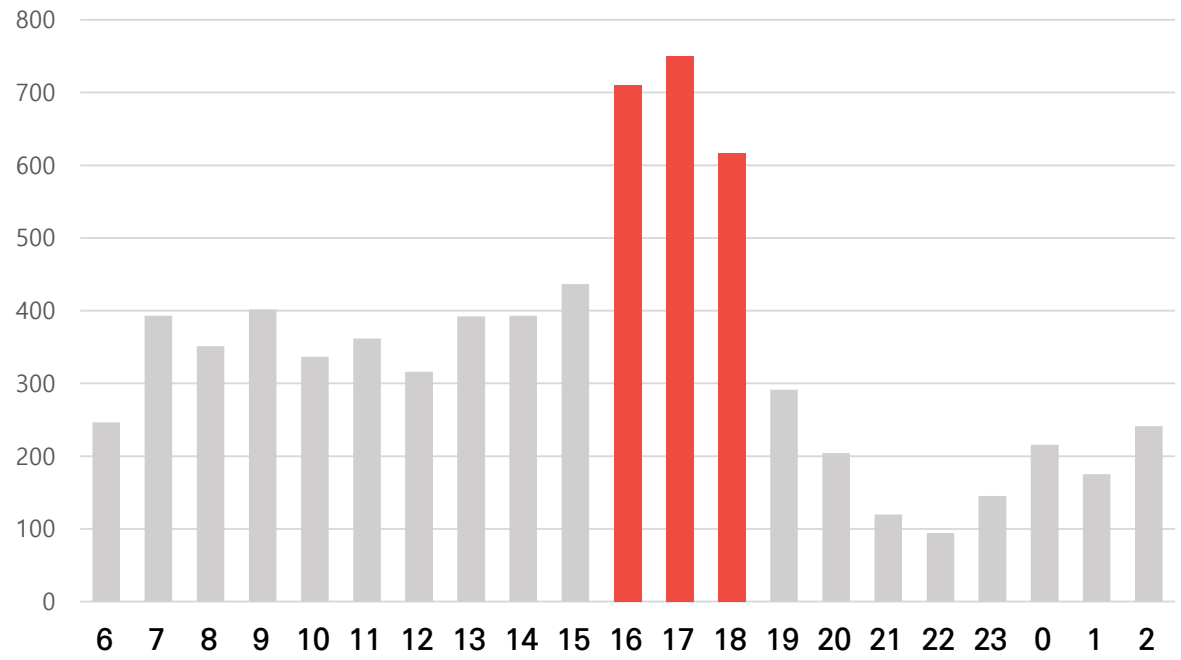


요일	방송횟수	총판매개수	평균판매개수
월	4967	1546362	311.3272
화	5252	1629370	310.2381
수	5070	1527230	301.2288
목	5032	1566577	311.323
금	4853	1439863	296.6955
토	4892	1692879	346.0505
일	5313	1734932	326.5446

- 평일보다 주말에 판매량이 많고 토요일에 판매량이 가장 많음.
- 홈쇼핑의 주 고객층은 주부로 예상되는 40~50대 여성이지만 주말에는 평일에 일하는 직업을 가진 직장인과 같은 고객층까지 유입되어 평일에 비해 판매량이 더 많은 것으로 추측.

EDA – 판매개수

시간대별 평균판매개수

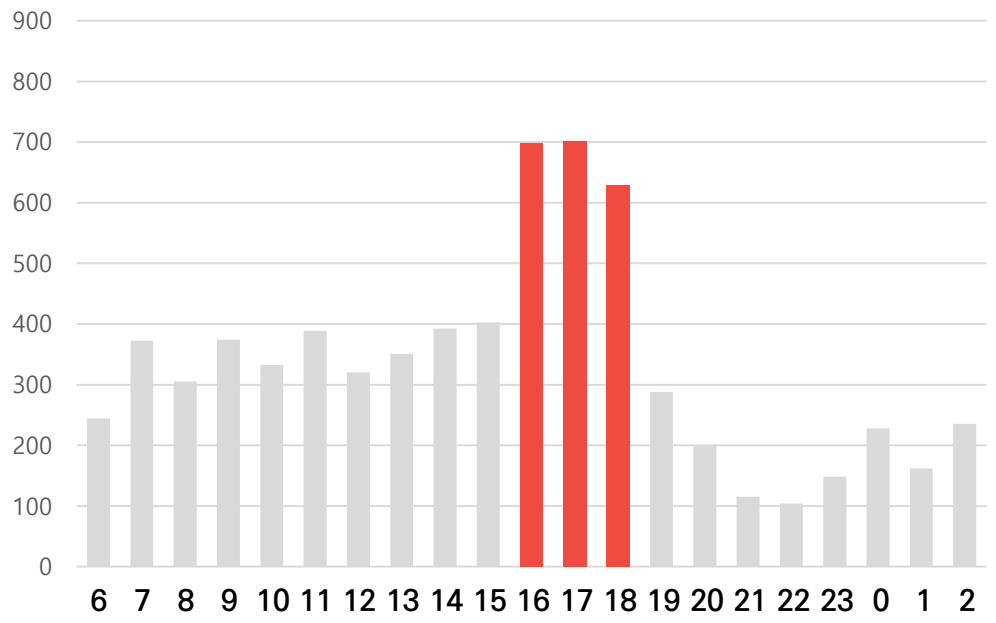


- 다른 시간대에 비해 16~18시에 판매량이 압도적으로 높음.
- 16~18시의 방송 편성을 확인해보니 대부분이 농수축 상품군.
- 16시 이전에는 다양한 상품군이 골고루 편성되어 있고, 19시 이후에는 가전 상품군이 많이 편성되어 있음.

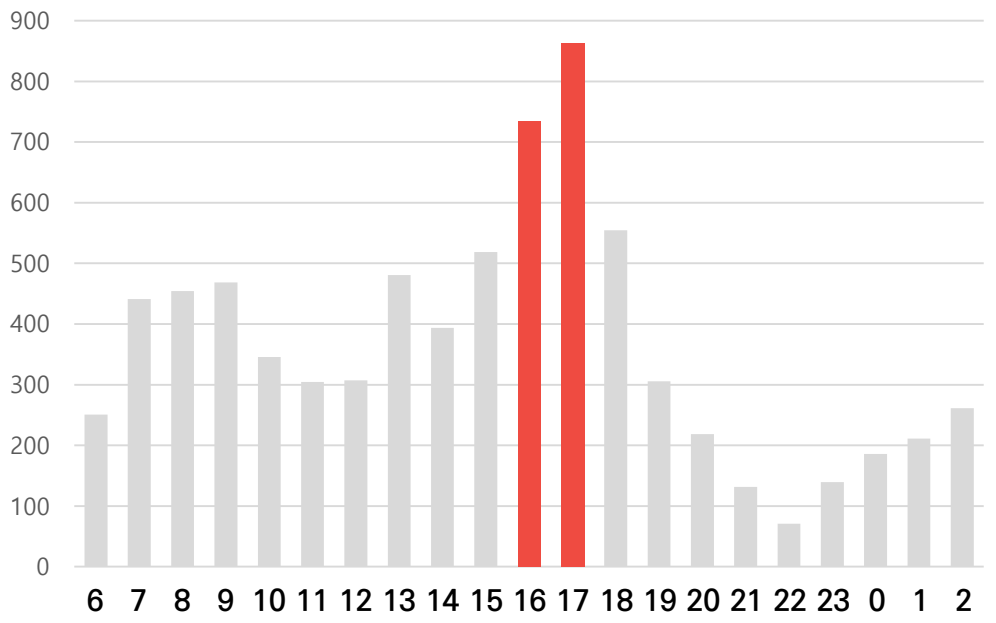
시간대	방송횟수	총판매개수	평균판매개수
6	1329	327511.8	246.4348
7	1460	574005.7	393.1546
8	1599	561824.7	351.36
9	1553	623696.8	401.6078
10	1875	631146.4	336.6114
11	1823	659549.8	361.7937
12	1746	552032.7	316.1699
13	1649	646384.4	391.9857
14	1654	650371.7	393.2115
15	1592	694867.7	436.4747
16	1372	974120.7	710.0005
17	1350	1012707	750.1537
18	1162	717429.1	617.4089
19	1746	509021	291.5355
20	2262	461746.6	204.132
21	3013	360433.7	119.6262
22	3153	296701.1	94.10119
23	2066	300660.5	145.5278
0	1407	303715.9	215.8606
1	1498	262402.5	175.1686
2	70	16883.35	241.1907

EDA - 판매개수

평일 시간대별 평균판매개수



주말 시간대별 평균판매개수



- 평일과 주말 시간대별 판매량 분포가 크게 다르지는 않지만 주말의 판매량이 전체적으로 높은 편임.
- 평일은 16~18시에 판매량이 많고 주말은 16~17시에 판매량이 많음.
- 평일은 16~18시가 고르게 판매량이 많은 반면 주말은 17시가 16시보다 평균 판매량이 100개 이상 많으며 평일에 비해 18시의 판매량이 적은 편임.

EDA – 시청률

시청률 데이터셋

시간대	2019-01-01	2019-01-02	2019-01-03	⋯	2019-12-30	2019-12-31	2019-01-01 to 2019-12-31
02:00	0	0	0		0	0	0.003
02:01	0	0	0		0	0.012	0.003
02:02	0	0	0		0	0	0.004
⋮				⋱			
01:58	0	0	0		0.019	0	0.004
01:59	0	0	0		0	0	0.004
월화수목금토일 02:00-01:59	0.004	0.006	0.002		0.005	0.005	0.004

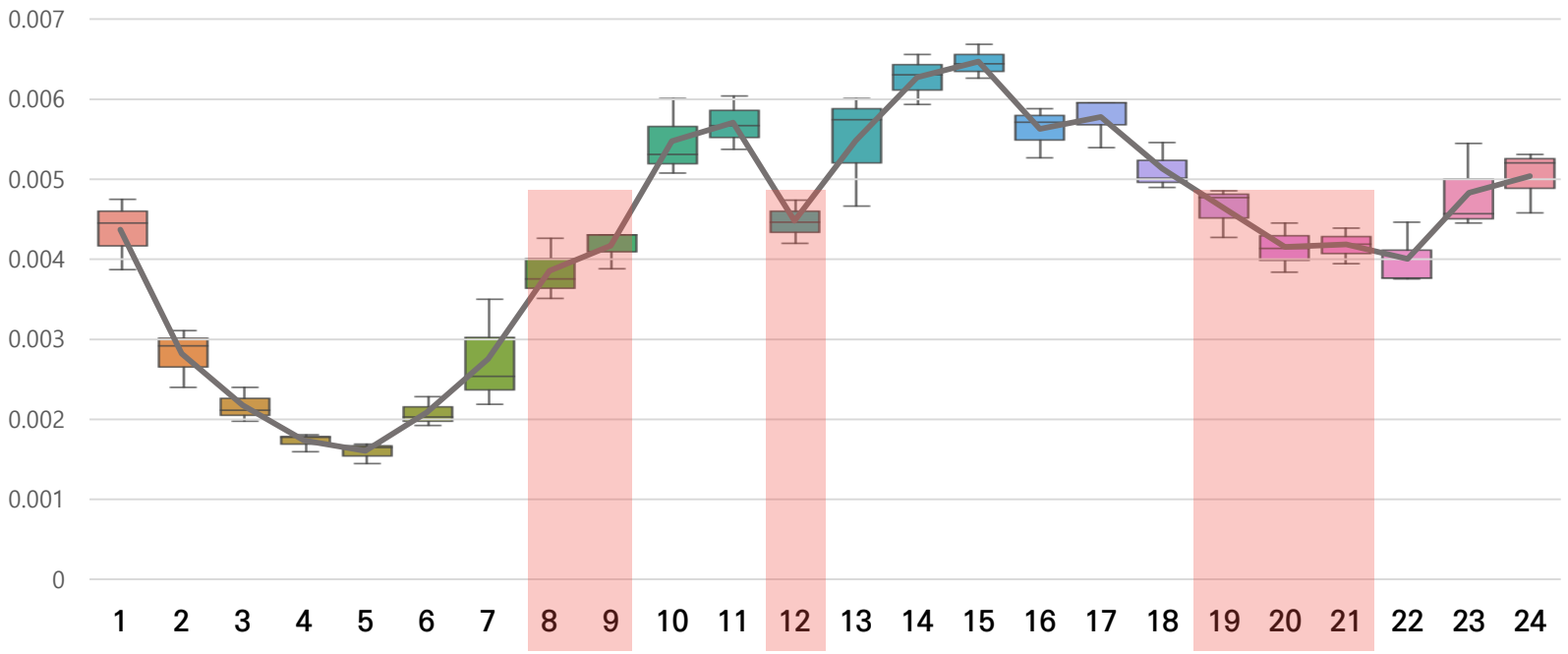
일자	시간	시청률
2019-01-01	02:00	0
2019-01-01	02:01	0
2019-01-01	02:02	0
⋮		
2019-12-31	01:57	0
2019-12-31	01:58	0
2019-12-31	01:59	0

왼쪽과 같이 제공된 데이터를 오른쪽과 같이 변형하여 탐색 진행

1440행(시간) 365열(날짜) → 525600행 3개 변수 (일자, 시간, 시청률)

EDA – 시청률

시간대별 평균 시청률

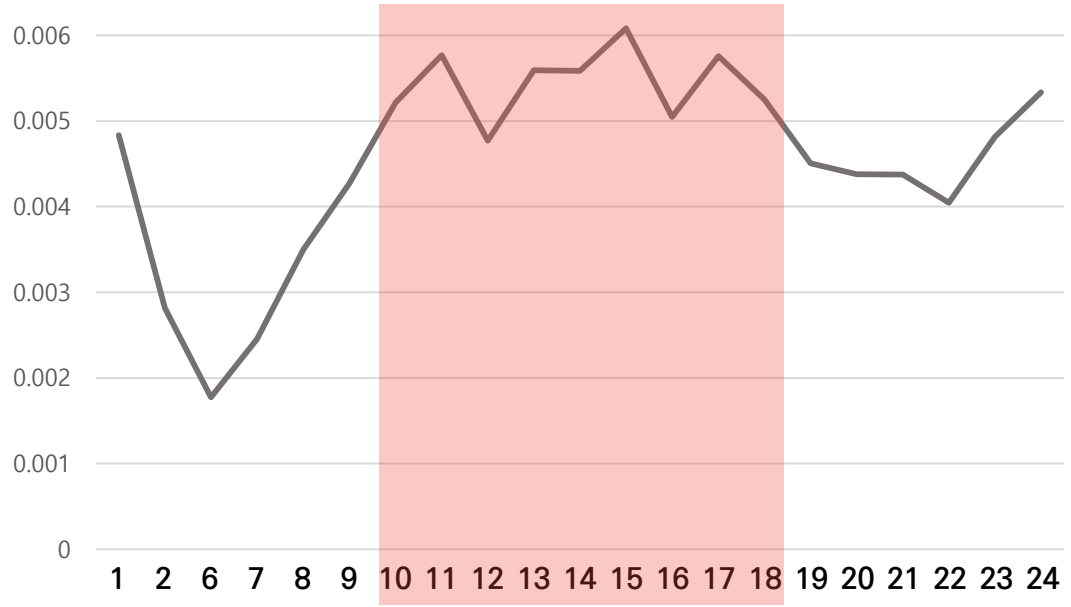


- 14~15시에 시청률이 가장 높음.
- 홈쇼핑 주 시청층이 40~50대 여성인 것을 고려했을 때 식사 후 잠시 휴식하며 홈쇼핑을 보는 시간을 갖는 것으로 추측.
- 식사 시간대에 시청률이 비슷하게 나타나고 분산이 작은 것으로 보아 주로 보는 사람들이 계속 본다고 추측할 수 있으므로 더 많은 사람들이 방송을 시청할 수 있도록 접근성을 높이는 방안이 필요.

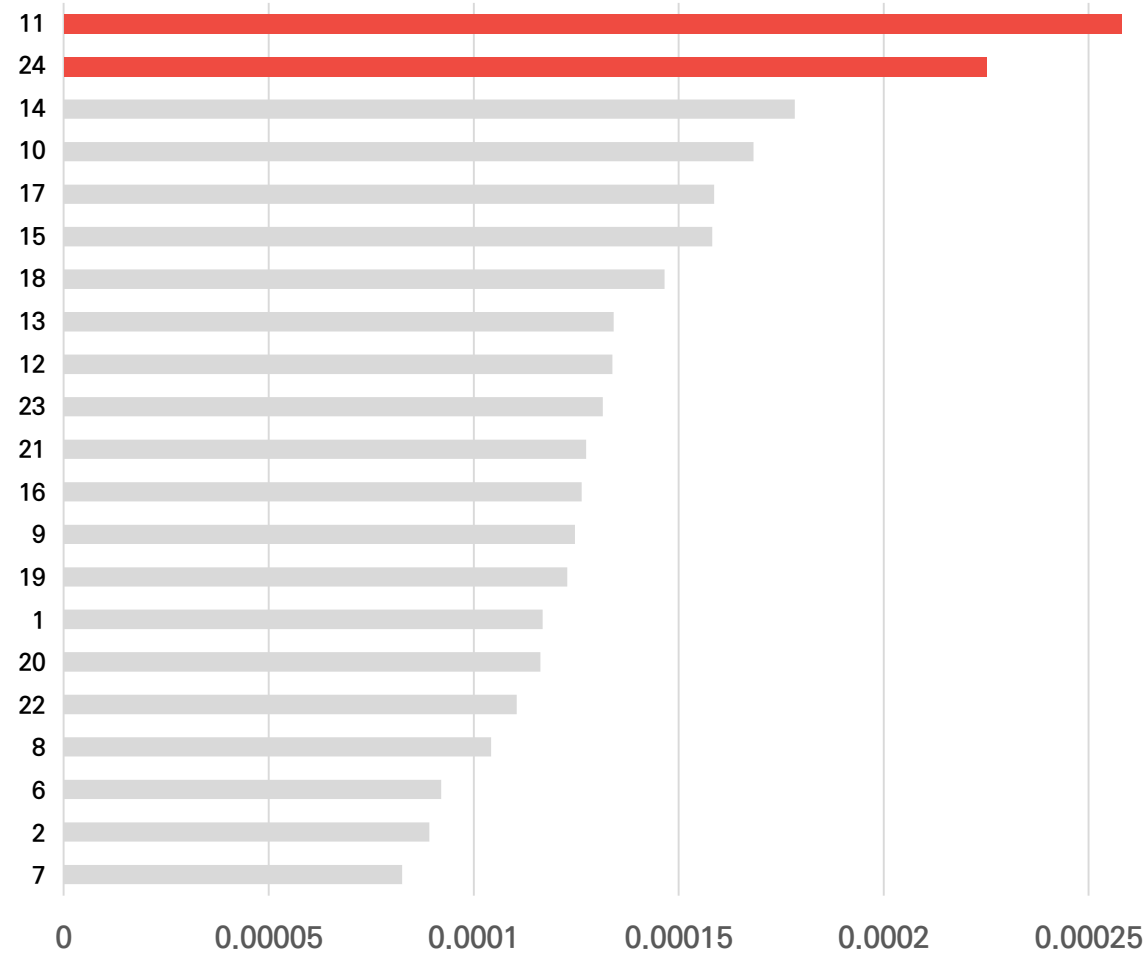
EDA – 시청률

평일 시간대별 시청률

〈 평균 〉



〈 분산 〉

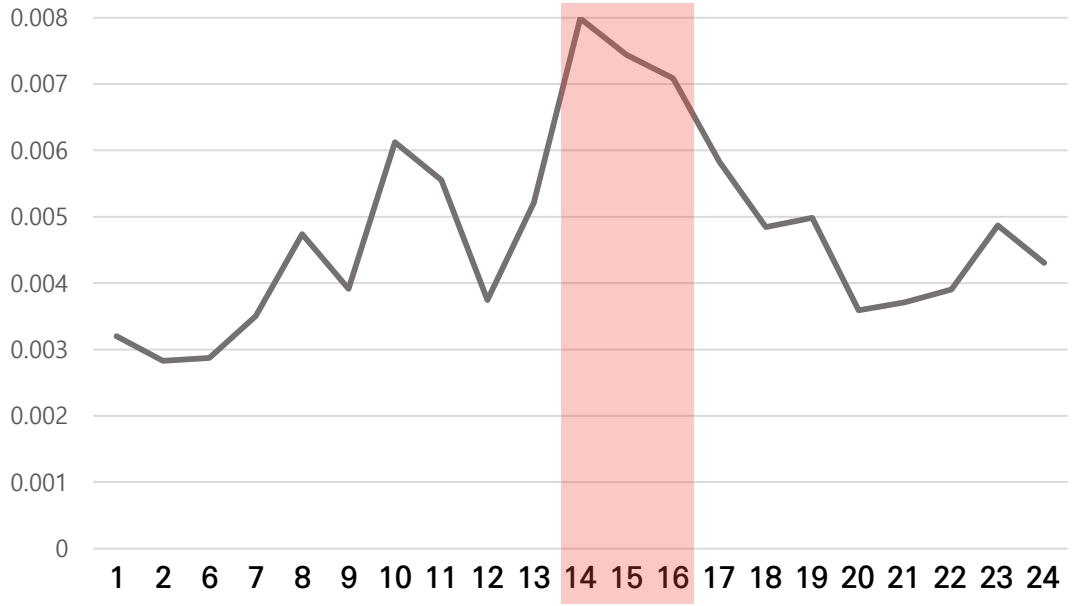


- 평일에는 오전부터 오후까지 시청률이 0.0055 정도로 균일하게 높음.
- 분산이 큰 시간대는 사람들이 채널을 이리저리 돌려보는 '재핑'과 연관이 있을 것으로 추측.

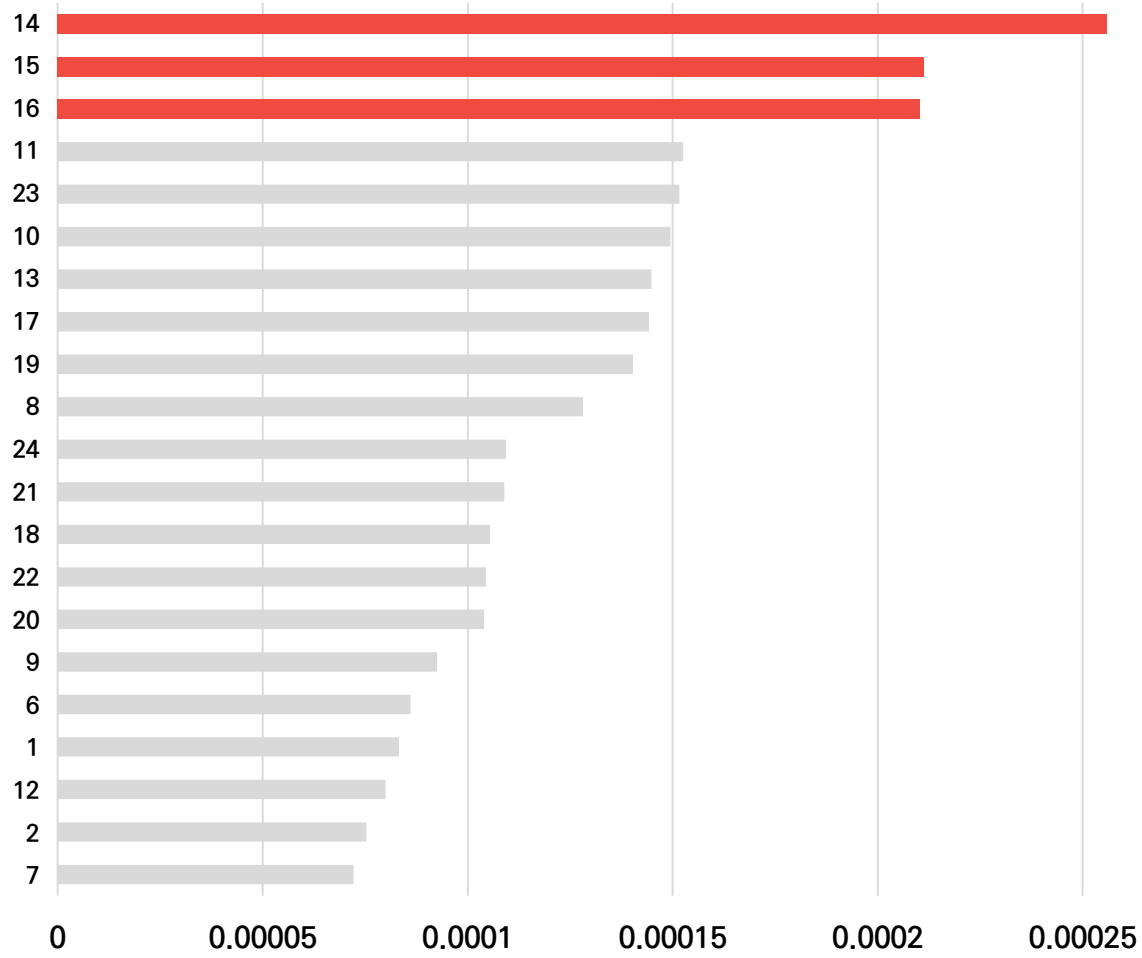
EDA – 시청률

주말 시간대별 시청률

〈 평균 〉



〈 분산 〉



- 평일과 달리 주말에는 14~16시에 시청률이 약 0.0075로 잠깐 높고 분산도 가장 큼.
- 분산이 큰 시간과 평균이 높은 시간이 동일한 것으로 보아 주말 14~16시에 항상 홈쇼핑을 보는 층과 채널을 이리저리 돌리다가 보는 층이 함께 있는 시간대인 것으로 추측.

변수 생성

3

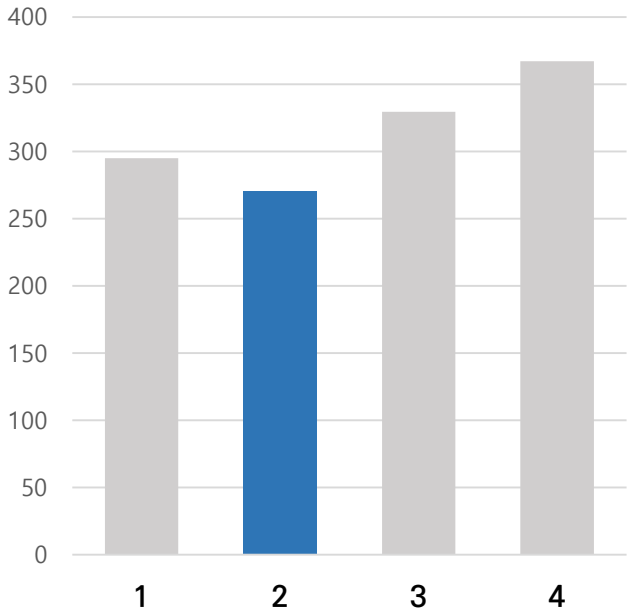
파생변수 – 방송일자

- **월 :** 방송일자에서 추출. (1 ~ 12)
- **요일 :** 방송일자에서 추출. (월 ~ 일)
- **주차 :** 방송일자에서 추출. 1년 중 몇 번째 주인지 나타내는 변수. (1 ~ 53)
- **분기 :** 월에서 추출. (1 : 1, 2, 3월 / 2 : 4, 5, 6월 / 3 : 7, 8, 9월 / 4 : 10, 11, 12월)
- **계절 :** 월에서 추출. (봄 : 3, 4, 5월 / 여름 : 6, 7, 8월 / 가을 : 9, 10, 11월 / 겨울 : 12, 1, 2월)
- **평일주말 :** 요일에서 추출. (평일 : 월, 화, 수, 목, 금 / 주말 : 토, 일)
- **공휴일 :** 방송일자에서 추출. (1 : 공휴일 / 0 : 공휴일 아님)
- **공휴일일주일전 :** 방송일자에서 추출. 공휴일 변수의 일주일 전 날짜부터 당일까지. (1 : 공휴일 일주일 전 / 0 : 해당없음)

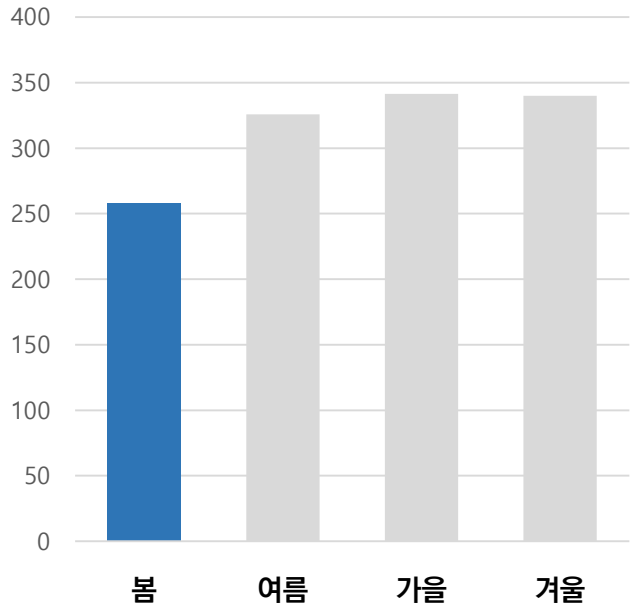
방송일자	월	주차	요일	분기	계절	평일주말	공휴일	공휴일일주일전
2019-01-01	1	1	화	1	겨울	평일	1	1
2019-01-02	1	1	수	1	겨울	평일	0	0
2019-01-03	1	1	목	1	겨울	평일	0	0
2019-12-30	12	53	월	4	겨울	평일	0	0
2019-12-31	12	53	화	4	겨울	평일	0	0

파생변수 – 방송일자

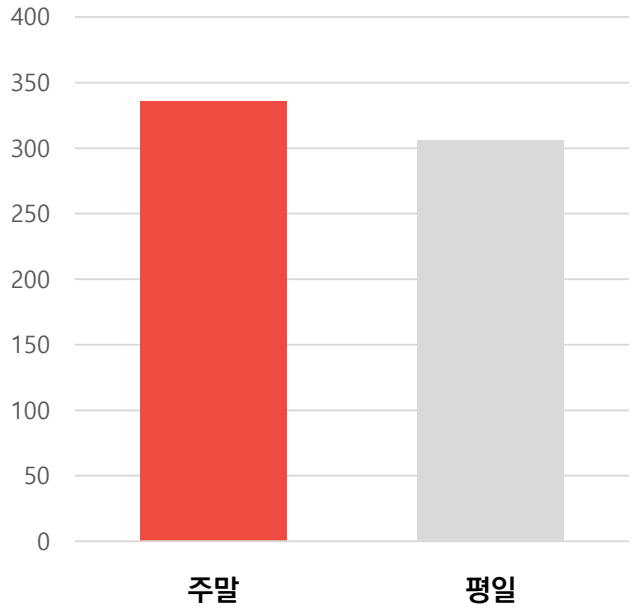
분기별 평균판매개수



계절별 평균판매개수



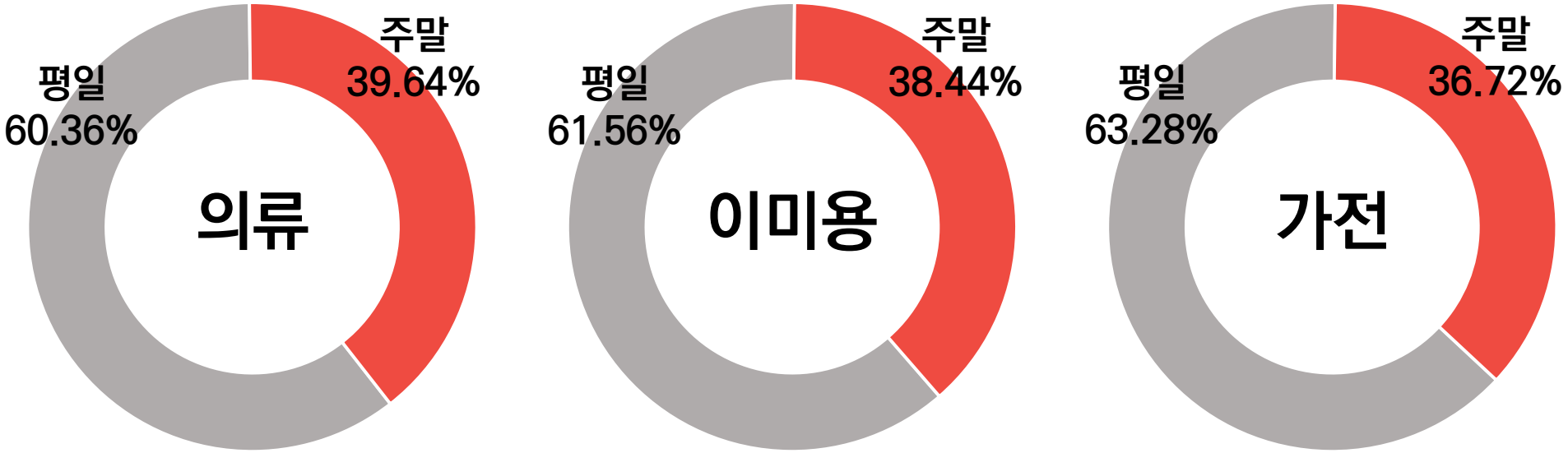
평일주말별 평균판매개수



- 2분기에 판매량이 가장 적고 4분기에 판매량이 가장 많음. 다른 분기에 비해 2분기에는 특별한 이벤트가 없기 때문으로 추측.
- 봄이 다른 계절에 비해 판매량이 적음. 분기와 마찬가지로 봄이 다른 계절에 비해 특별한 이벤트가 없기 때문으로 추측.
- 주말이 평일에 비해 판매량이 많음. 주말에는 평일에 일하는 직업을 가진 직장인과 같은 고객층까지 유입되어 평일에 비해 판매량이 더 많은 것으로 추측.

파생변수 – 방송일자

평일주말별 판매비율



- 평일 주말 판매횟수 비교.
- 의류, 이미용, 가전은 다른 상품군보다 주말에 많이 구매.

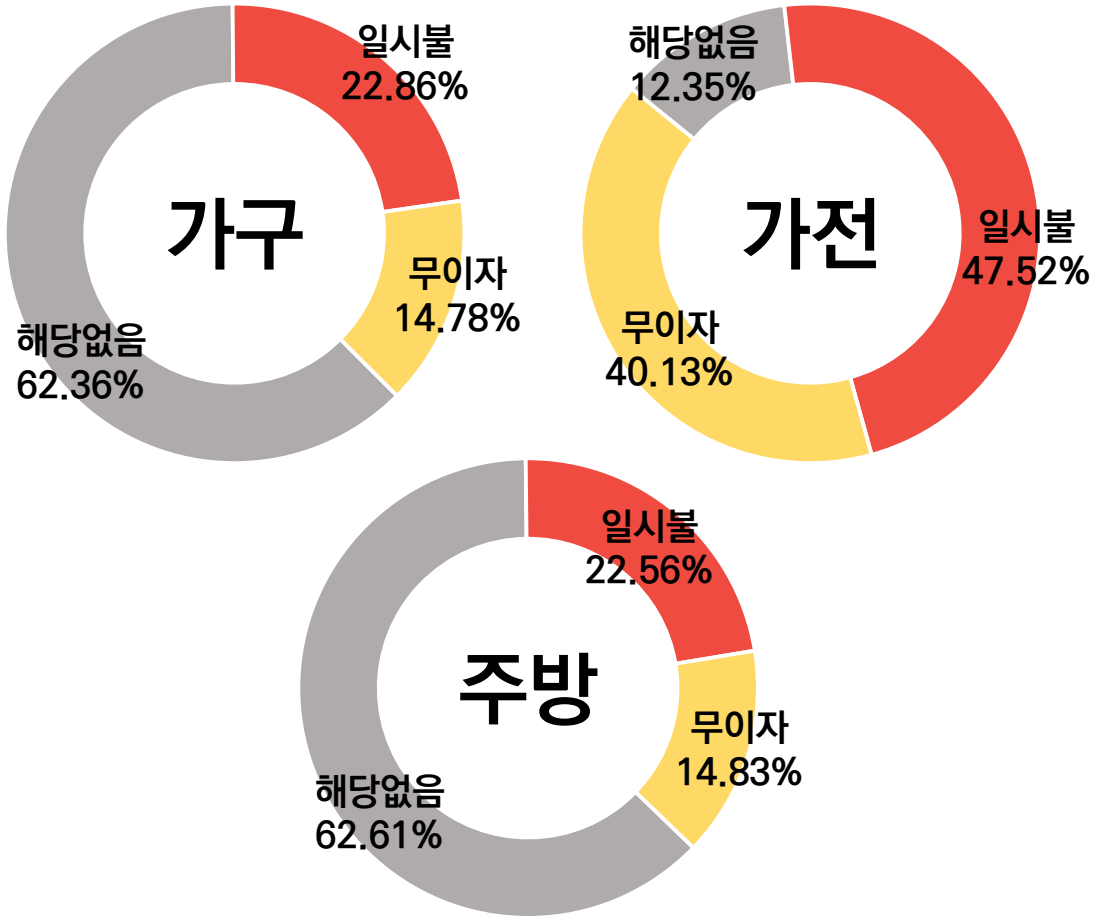
파생변수 – 상품명

- 결제방법 변수
  - 일시불, 무이자, 해당없음
  - 상품명에 '일시불', '(일)', '일)'이 들어가면 일시불, '무이자', '(무)', '무)'가 들어가면 무이자, 둘 다 아니면 해당없음.

상품명	결제방법
일시불 삼성 노트북 9 메탈 기본형 NT900X5J-K14	일시불
무이자 삼성 노트북 9 메탈 기본형 NT900X5J-K14	무이자
(일) 삼익가구 LED 제니비 서랍형 침대 SS	일시불
(무) 삼익가구 LED 제니비 서랍형 침대 SS	무이자
일) 한샘 하이바스 내추럴 기본형	일시불
무) 한샘 하이바스 내추럴 기본형	무이자
멋진밥상 흥양농협 쌀 20kg	해당없음

- 결제방법이 일시불이거나 무이자인 경우는 가전에서 가장 많이 발견.
- 판매단가가 높은 상품군에 무이자, 일시불이 많이 나타남.
- 무이자보다 일시불의 판매량이 더 많음.

결제방법별 판매비율

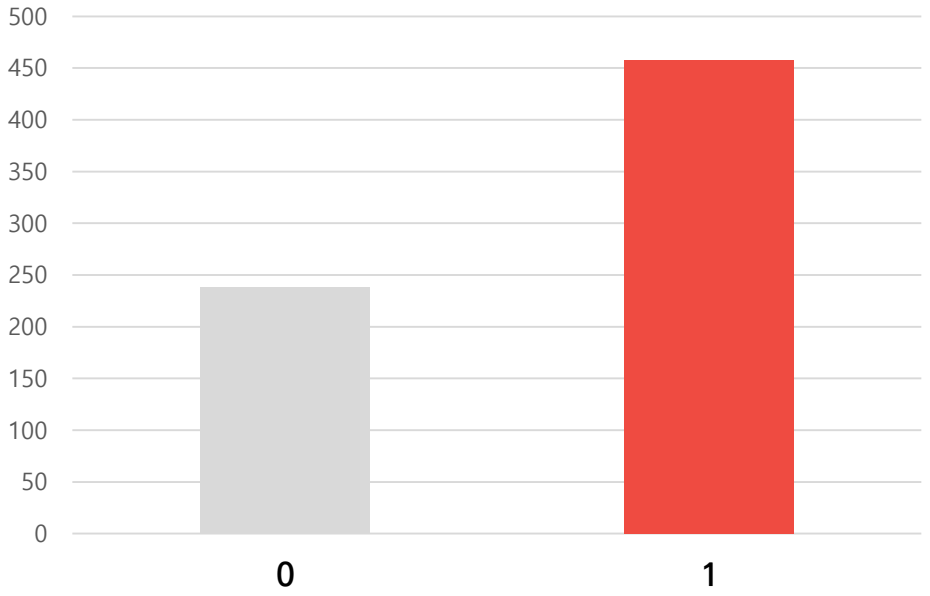


파생변수 – 상품명

– 세트상품 변수

– 상품명에 ‘세트‘, ‘패키지‘, ‘+’, ‘–종’(1종 제외) 포함 여부

상품명	구성
일시불[가이거] 제니스시계 주얼리**세트**	**1**
오모떼 360도 텐션업 레이스 **패키지** 시즌4	1
NNF 쿠션퍼자켓**+**베스트	1
테이트 남성 셀린니트**3종**	**1**
도스문도스 카이만 엠보 소가죽 핸드백 **1종**	**0**

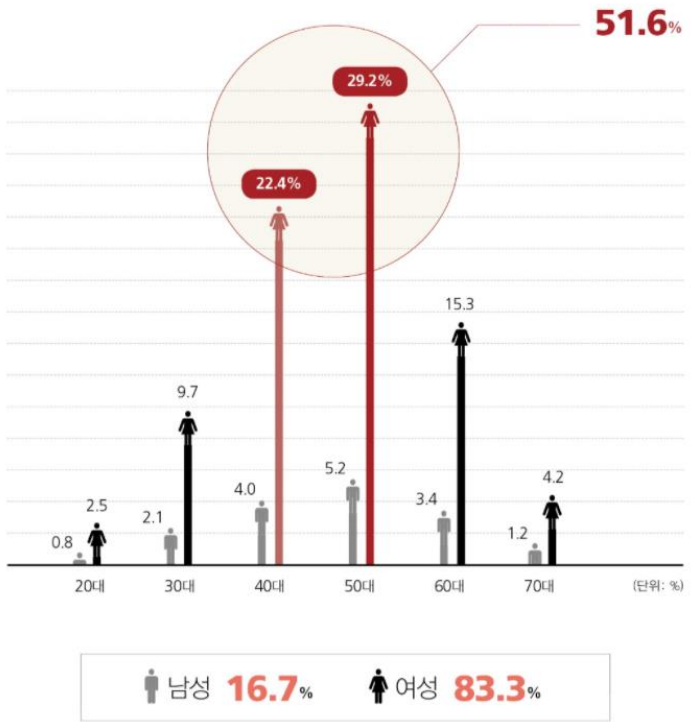


– 세트로 상품을 파는 경우, 단일상품보다 판매량이 약 두 배 많음.
– 세트로 방송이 편성된 횟수는 12357회, 단일상품으로 편성된 횟수는 23022회.
– 세트상품이 단일상품보다 잘 팔리기 때문에 세트로 편성을 더 많이 하는 방향 고려.

파생변수 – 상품명

- **성별 변수**
  - 여성, 남성, 해당없음.
  - 상품명에 '여성', '브라', '란쥬', '블라우스', '밍크', '립스틱', '퍼 베스트', '원피스', '귀걸이', '목걸이', '반지', '팔찌', '고데기'가 들어가면 **여성**, '남성', '면도'가 들어가면 **남성**, 나머지는 **해당없음**.

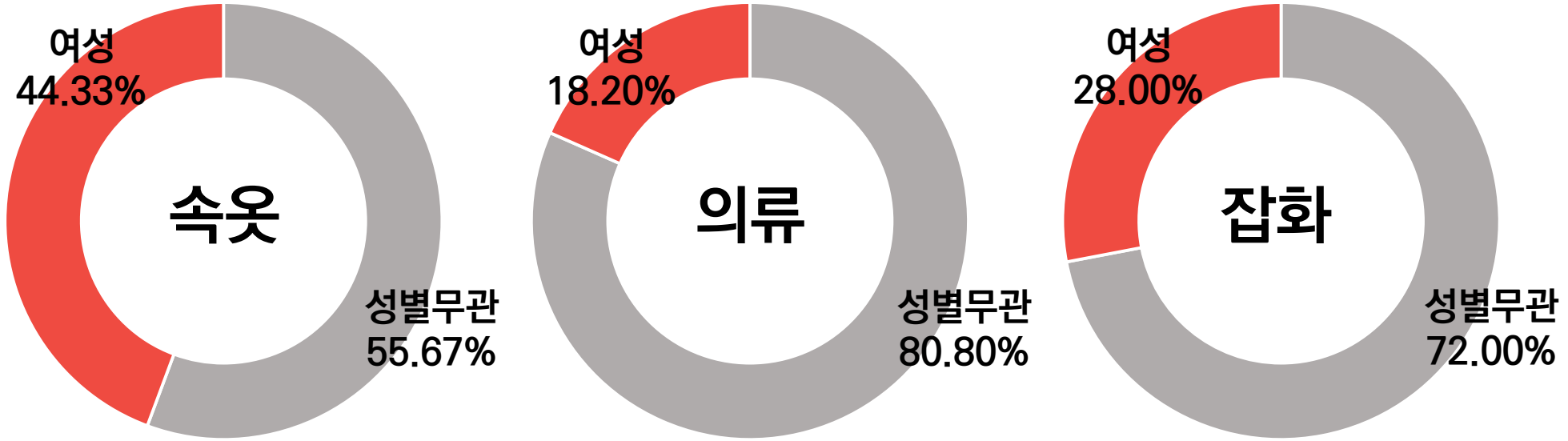
상품명	성별
오모떼 오리지널 쉐이핑 **브라**팬티 시즌2	여성
루시헨느 레이스 홀리데이 **란쥬** 패키지	여성
코몽트 **남성** 프린트티셔츠8종	남성
도루코 페이스5 **면도**기 세트 (1세트)	남성
디비노 선글라스 세트	해당없음



- 한국TV홈쇼핑에 따르면 TV홈쇼핑 주 고객층은 40, 50대 여성.
- 홈쇼핑 성별 구매 비중이 여성이 높다는 점을 참고하여 성별 변수 생성.

파생변수 – 상품명

성별 판매비율



- 위의 그래프에서는 여성과 그 외로 비교하여 나타냄.
- 속옷 상품군의 경우 여성 관련 상품의 판매량이 전체 속옷 판매량의 절반 가까이 차지.
- 속옷, 의류, 잡화 상품군에서 여성 관련 상품이 주요한 상품이 될 수 있음.

파생변수 – 상품명

– 브랜드 변수

- 상품명에서 브랜드에 해당하는 부분을 추출하여 브랜드 변수 생성.
- 추가적으로 브랜드파워를 알 수 있는 변수 생성.
  - 데이터가 2019년도 1개년밖에 없기 때문에 브랜드의 판매실적을 담고 있는 브랜드파워 변수를 사용하려면 미래참조에 유의해야 함.
  - 브랜드의 가치는 판매개수 그대로 사용하지 않고 판매개수 기준 오름차순으로 순위를 매겨서 브랜드 가치를 나타내는 변수로 활용.

* 추가했을 때 모델 성능이 나빠져 사용하지 않음.

상품명	브랜드
일시불 삼성 노트북 9 메탈 기본형 NT900X5J-K14	삼성
무이자 삼성 노트북 9 메탈 기본형 NT900X5J-K14	삼성
(일) 삼익가구 LED 제니비 서랍형 침대 SS	삼익
(무) 삼익가구 LED 제니비 서랍형 침대 SS	삼익
일) 한샘 하이바스 내추럴 기본형	한샘
무) 한샘 하이바스 내추럴 기본형	한샘

파생변수 – 상품당노출시간

- 상품당노출시간 변수
  - 같은 방송일시에서 여러 개의 상품을 판매하면 그만큼 한 상품 당 화면에 노출되는 시간이 짧을 것이라고 생각.
  - 노출(분) 변수와 한 방송에서 판매하는 상품의 개수를 고려하여 상품당노출시간 변수 생성.

방송일시	노출(분)	상품명	상품개수	상품당노출시간
2019-01-01 6:00	20	테이트 남성 셀린니트3종	2	10
2019-01-01 6:00	20	테이트 여성 셀린니트3종	2	10
2019-01-02 10:00	20	일시불 쿠첸 풀스텐 압력밥솥 10인용 (A1)	4	5
2019-01-02 10:00	20	무이자 쿠첸 풀스텐 압력밥솥 10인용(A1)	4	5
2019-01-02 10:00	20	일시불 쿠첸 풀스텐 압력밥솥 6인용(A1)	4	5
2019-01-02 10:00	20	무이자 쿠첸 풀스텐 압력밥솥 6인용(A1)	4	5

파생변수 – 편성비율

- 산출 방식 예시

방송주차	브랜드	편성횟수	주차브랜드편성비율
1	A	8	0.4
1	B	5	0.25
1	C	7	0.35

$$\frac{\text{주차브랜드편성횟수}}{sum(\text{주차편성횟수})} = \frac{8}{8+5+7}$$

- 주차브랜드

방송주차	브랜드	주차브랜드편성비율
1	테이트	0.016
1	오모떼	0.022
1	CERINI_BY_PAT	0.007

- 날짜브랜드

방송날짜	브랜드	날짜브랜드편성비율
2019-01-01	테이트	0.162
2019-01-01	오모떼	0.067
2019-01-01	CERINI_BY_PAT	0.081

- 월브랜드

방송월	브랜드	월브랜드편성비율
01	테이트	0.008
01	오모떼	0.008
01	CERINI_BY_PAT	0.004

- 월마더코드

방송월	마더코드	월마더코드편성비율
01	100346	0.008
01	100305	0.008
01	100808	0.003

(브랜드별로 방송주차, 날짜, 월마다 / 월별 마더코드) 편성비율에서 차이가 난다

파생변수 – 판매상품비율

– 산출 방식 예시

방송월	마더코드	월마더코드판매상품수	월마더코드판매상품비율
01	10	3	0.15
01	100	4	0.2
01	1003	9	0.45
01	138	2	0.1
01	238	2	0.1

$$\frac{\text{월마더코드판매상품수}}{sum(\text{월판매상품수횟수})}$$

$$= \frac{3}{3+4+9+2+2}$$

– 월마더코드

방송월	마더코드	월마더코드판매상품비율
01	100346	0.017
01	100305	0.021
01	100808	0.004

– 주차마더코드

방송주차	마더코드	주차마더코드판매상품비율
1	100346	0.016
1	100305	0.022
1	100808	0.005

파생변수 – 구성비율

- 산출 방식 예시

방송날짜	상품군	브랜드	편성횟수	날짜상품군브랜드구성비율
2019-01-01	속옷	BYC	6	0.6
2019-01-01	속옷	오모떼	4	0.4
2019-01-01	의류	CERINI_BY_PAT	1	1

$$\frac{\text{날짜상품군브랜드편성횟수}}{sum(\text{날짜상품군편성횟수})} = \frac{6}{6+4}$$

- 날짜상품군브랜드
  - 해당 날짜의 판매되는 상품군 편성 횟수 중에서 각 브랜드가 차지하는 비율

방송날짜	상품군	브랜드	날짜상품군브랜드구성비율
2019-01-01	의류	테이트	0.352
2019-01-01	속옷	오모떼	0.384
2019-01-01	의류	CERINI_BY_PAT	0.176

파생변수 – 구성비율

– 시간브랜드상품군

– 해당 주차 및 시간에 편성된 상품군 편성중에서 개별 브랜드 편성이 차지하는 비율

방송주차	방송시간	상품군	브랜드	시간브랜드상품군구성비율
18	14	의류	뱅뱅	0.285714
22	11	잡화	코치	0.260870
46	17	농수축	안동	0.142857

– 계절시간브랜드상품코드

– 해당 계절 및 시간에 편성된 제품 중에서 개별 브랜드 제품들의 편성이 차지하는 비율

방송계절	방송시간	브랜드	상품코드	날짜상품군브랜드구성비율
여름	23	삼성	201682	0.005964
겨울	2	밸런스파워	201236	0.034483
봄	23	K-SWISS	200856	0.039286

파생변수 – 랭크

같은 방송이어도 방송시간이 지날수록 상품이 더 많이 팔리는 경향이 있음

- **상품코드 :** 해당 방송일자에서 방송되는 상품코드들의 시간 순서

상품코드	방송일시	상품명	상품코드랭크
201072	2019-01-01 6:00:00	테이트 남성 셀린니트3종	1
201079	2019-01-01 6:00:00	테이트 여성 셀린니트3종	1
201079	2019-01-01 6:20:00	테이트 여성 셀린니트3종	2
201072	2019-01-01 6:20:00	테이트 남성 셀린니트3종	2
201079	2019-01-01 6:40:00	테이트 여성 셀린니트3종	3

- **마더코드 :** 해당 방송일자에서 방송되는 마더코드들의 시간 순서

마더코드	방송일시	상품명	마더코드랭크
100346	2019-01-01 6:00:00	테이트 남성 셀린니트3종	1
100346	2019-01-01 6:00:00	테이트 여성 셀린니트3종	1
100346	2019-01-01 6:20:00	테이트 여성 셀린니트3종	2
100346	2019-01-01 6:20:00	테이트 남성 셀린니트3종	2
100346	2019-01-01 6:40:00	테이트 여성 셀린니트3종	3

파생변수 – 랭크

- **상품군 :** 해당 방송일자에서 방송되는 상품군들의 시간 순서

상품군	방송일시	상품명	상품군랭크
의류	2019-01-01 6:00:00	테이트 남성 셀린니트3종	1
의류	2019-01-01 6:00:00	테이트 여성 셀린니트3종	1
의류	2019-01-01 6:20:00	테이트 여성 셀린니트3종	2
의류	2019-01-01 6:20:00	테이트 남성 셀린니트3종	2
의류	2019-01-01 6:40:00	테이트 여성 셀린니트3종	3

- **판매단가 :** 해당 방송일자에서 방송되는 판매단가들의 시간 순서

판매단가	방송일시	상품명	판매단가랭크
39,900	2019-01-01 6:00:00	테이트 남성 셀린니트3종	1
39,900	2019-01-01 6:00:00	테이트 여성 셀린니트3종	1
59,900	2019-01-01 7:00:00	오모떼 레이스 파운데이션 브라	1
59,900	2019-01-01 7:20:00	오모떼 레이스 파운데이션 브라	2
59,900	2019-01-01 8:20:00	CERINI BY PAT 남성 소프트 기모 릴렉스팬츠	3

파생변수 – 범주형 변수 더미화

- 범주형 정보를 분석에 활용하기 위해 각종 범주형 변수 one-hot encoding 진행
- 6개월 이후 및 1개월 이후 예측 문제로 최대 시간 정보를 개별 월 단위로 한정

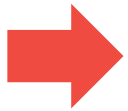
성별	상품군	방송요일	방송계절	일시불무이자	시즌성
남성	의류	Monday	겨울	해당없음	시즌성2
해당없음	잡화	Saturday	여름	해당없음	시즌성3
해당없음	주방	Tuesday	겨울	무이자	시즌성4
남성	의류	Sunday	봄	해당없음	시즌성3
여성	속옷	Sunday	가을	해당없음	시즌성4

남성	여성	해당없음	의류	잡화	...	일시불	무이자	시즌성2	시즌성3	시즌성4
0	1	0	0	0	...	0	0	0	0	1
0	0	1	0	0	...	0	1	0	1	0

외부데이터 – 중기예보(기상)

- 중기예보 데이터

발표시각	예보시각	최저기온	최고기온
2018-12-24 06시	2019-01-01	-7	0
2018-12-24 06시	2019-01-02	-6	1
2018-12-25 06시	2019-01-03	-6	2
2018-12-26 06시	2019-01-04	-5	3



방송날짜	예보시각	최저기온	최고기온
2019-01-01	2019-01-01	-7	0
2019-01-02	2019-01-02	-6	1
2019-01-03	2019-01-03	-6	2
2019-01-04	2019-01-04	-5	3

홈쇼핑 판매실적에는 날씨가 영향을 주는 것으로 알려져 있기 때문에 외부에서 수집하여 사용.

미래의 데이터를 사용하면 문제가 생기기 때문에 기상청에서 기상예보 데이터를 사용.

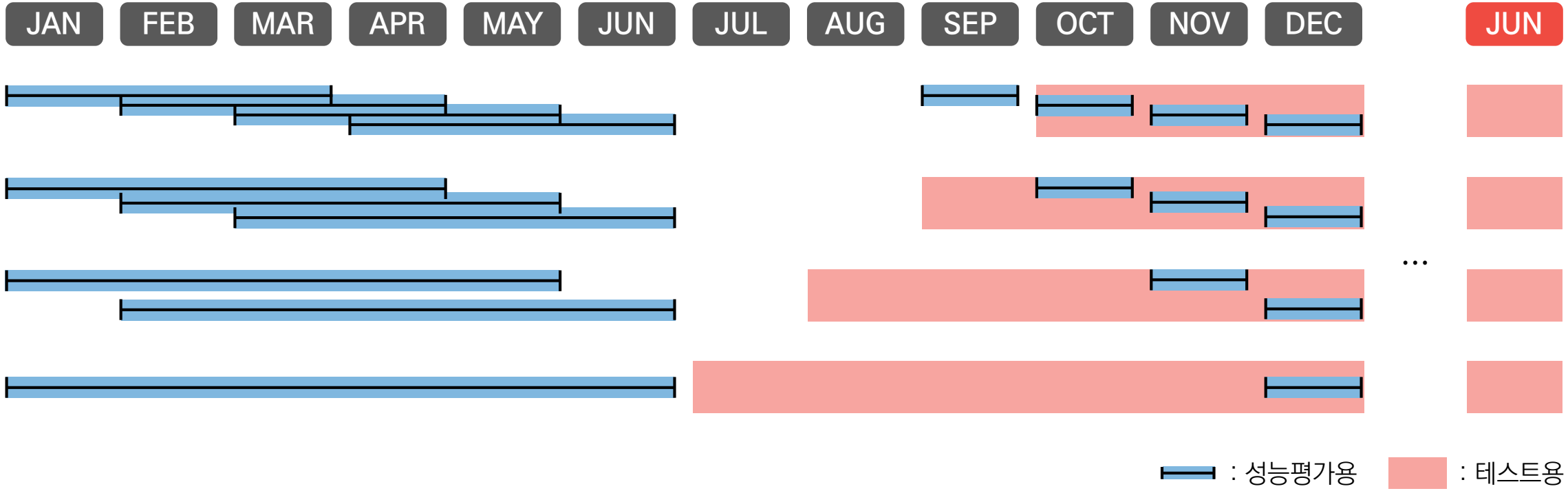
1주일 후의 최저, 최고기온을 예측한 데이터 (출처: 기상청)

모델링

4

데이터 분할

- 2019년 1~12월의 데이터로 2020년 6월의 판매실적을 예측하는 것이 문제.
- train set과 test set 사이에 5개월의 공백이 존재하기 때문에 모델의 성능을 평가할 때에도 동일한 환경에서 진행해야 한다고 판단.



학습 기간에 따라 validation set을 나눌 수 있음

'(3, 4, 5, 6)개월 학습 ➔ 6개월 후 예측' 을 통해 성능 평가

데이터 분할

- 2019년 1~12월의 데이터로 2020년 6월의 판매실적을 예측하는 것이 문제.
- train set과 test set 사이에 5개월의 공백이 존재하기 때문에 모델의 성능을 평가할 때에도 동일한 환경에서 진행해야 한다고 판단.

	dataset	Validation				test
1	train	1~6				7~12
	test	12				6
2	train	1~5	2~6			8~12
	test	11	12			6
3	train	1~4	2~5	3~6		9~12
	test	10	11	12		6
4	train	1~3	2~4	3~5	4~6	10~12
	test	9	10	11	12	6

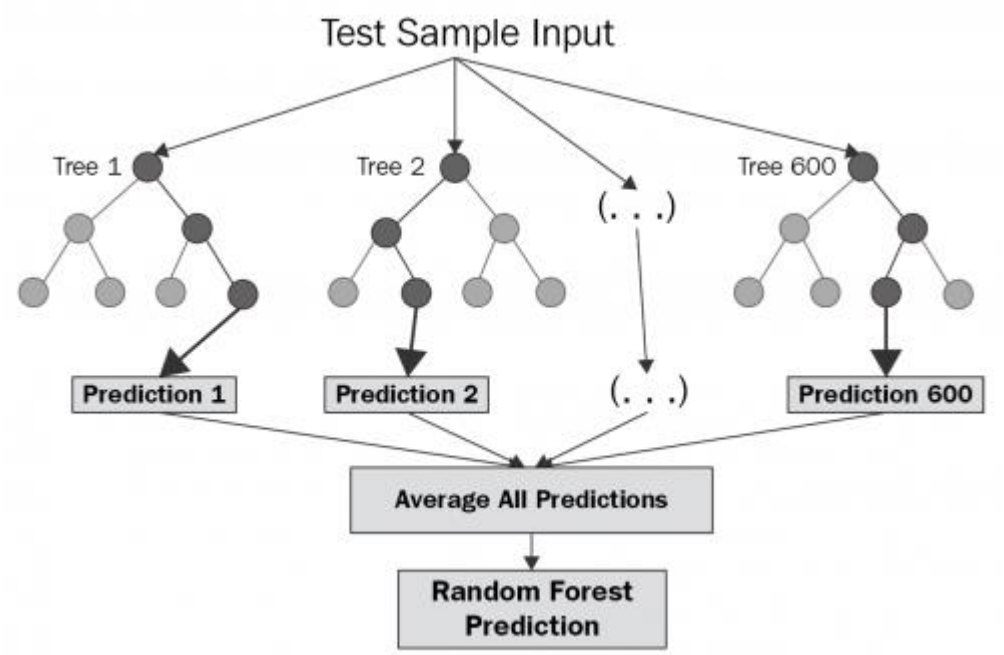
학습 기간에 따라 validation set을 나눌 수 있음

‘(3, 4, 5, 6)개월 학습 ➔ 6개월 후 예측’ 을 통해 성능 평가

모델링 – 학습 알고리즘

RandomForest

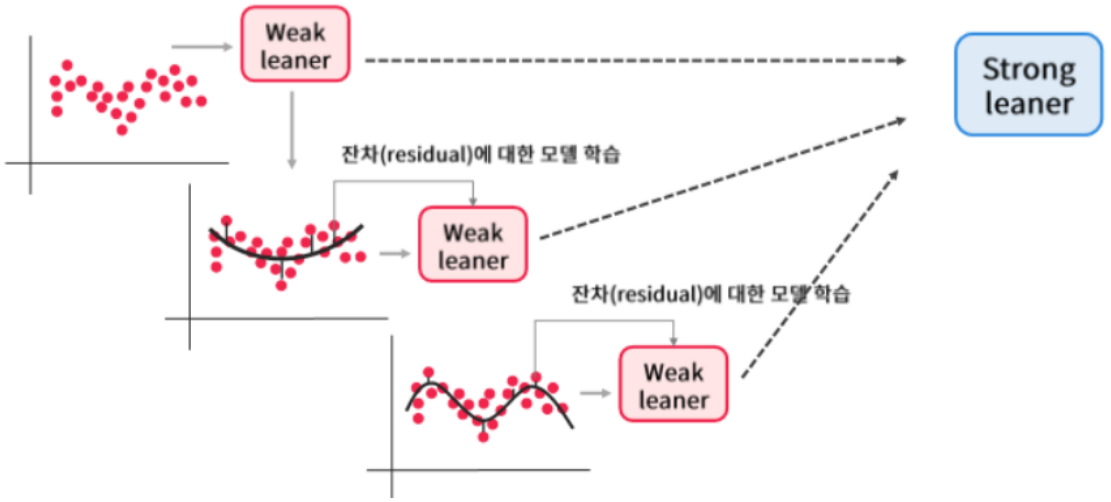
- 다수의 결정 트리들을 학습하는 앙상블 방법
- 월등히 높은 정확성, 간편하고 빠른 수행속도, 임의화를 통한 좋은 일반화 성능



출처 : IT위키

Gradient Boosting

- Gradient descent와 boosting이 결합된 방법
- 이전 모델이 예측한 데이터의 오차를 가지고 이 오차를 예측하는 학습기를 만드는 것

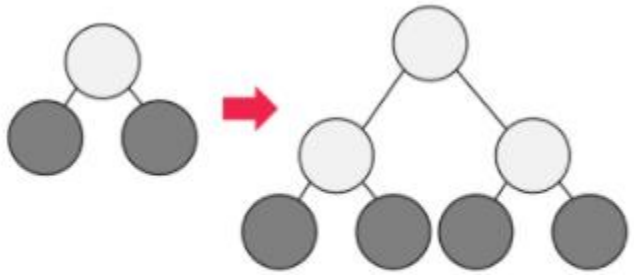


출처 : https://heung-bae-lee.github.io/

모델링 – 학습 알고리즘

XGBoost

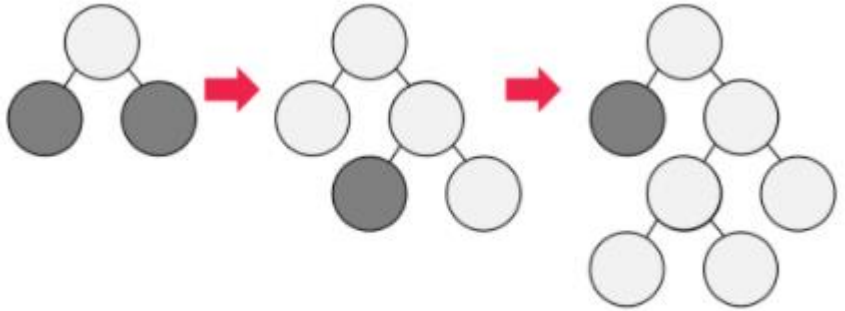
- 그래디언트 부스팅 프레임워크를 사용하는 의사결정 트리 기반 앙상블 머신러닝 알고리즘
- 장점으로는 GBM 대비 빠른 수행시간, 과적합 규제 기능 존재, 뛰어난 예측 성능 등이 있음.
- Level-wise (수평 확장)



Level-wise growth

LightGBM

- 그래디언트 부스팅 프레임워크를 사용하는 의사결정 트리 기반 앙상블 머신러닝 알고리즘
- 대용량 데이터 처리 가능, 다른 모델들보다 적은 자원 사용, 빠른 수행시간, GPU 지원, 높은 정확도
- Leaf-wise (수직 확장)



Leaf-wise growth

출처 : https://heung-bae-lee.github.io/

모델링 – 최종 변수 선정

최종 변수 (

〈범주형 변수〉

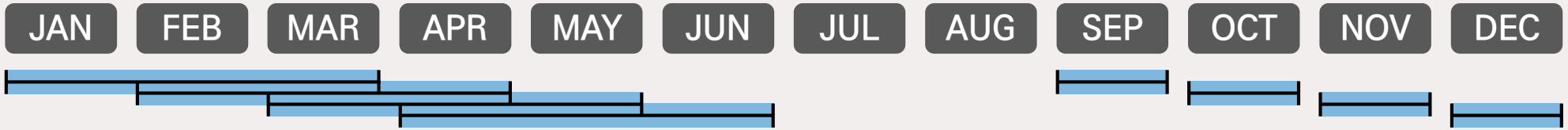
세트구성
방송계절
일시불무이자여부
방송요일
상품군
성별
브랜드변화
평일주말

〈수치형 변수〉

판매단가
노출(초)
기온(최저, 최고)
방송시간
주차브랜드편성비율
날짜브랜드편성비율
월브랜드편성비율
월마더코드편성비율
월마더코드판매상품비율
주차마더코드편성비율
상품코드랭크
마더코드랭크
상품군랭크
날짜상품군브랜드구성비율
판매단가랭크
계절시간브랜드상품코드구성비율
시간브랜드상품군구성비율
계절브랜드구성비율

)

모델링 – 3개월



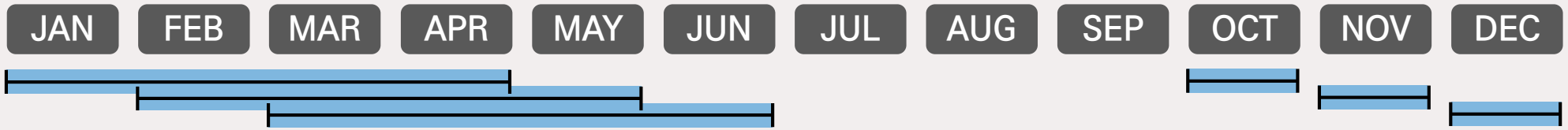
모델	평균 MAPE	표준편차 MAPE	평균 RMSE	표준편치 RMSE
XGB	58.568389	3.409890	264.195651	71.843168
LGBM	57.550793	4.753960	258.905919	75.254204
GradientBoosting	57.565202	4.749907	257.078558	76.019645
RandonForest	58.035075	4.687558	260.100521	77.184436

* 파라미터 튜닝없이 기본 모델로 성능 평가

1~3월 → 9월 예측 / 2~4월 → 10월 예측 / 3~5월 → 11월 예측 / 4~6월 → 12월 예측

모델별 평균 MAPE가 약 57-58로 나타남

모델링 – 4개월



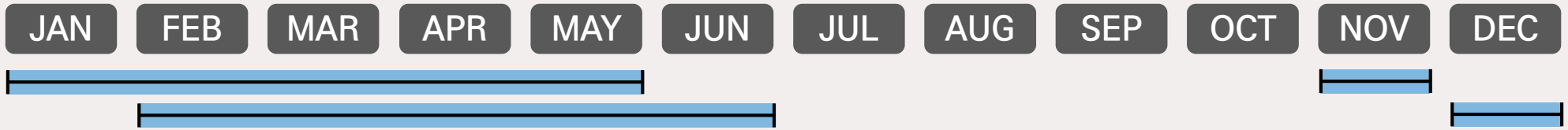
모델	평균 MAPE	표준편차 MAPE	평균 RMSE	표준편치 RMSE
XGB	59.049538	6.785847	270.397618	88.262519
LGBM	57.150615	6.495811	266.680704	89.276245
GradientBoosting	56.951273	6.136884	266.729352	89.949705
RandonForest	57.496623	5.957642	268.987503	89.861343

* 파라미터 튜닝없이 기본 모델로 성능 평가

1~4월 → 10월 예측 / 2~5월 → 11월 예측 / 3~6월 → 12월 예측

모델별 평균 MAPE가 약 56-59로 나타남

모델링 – 5개월



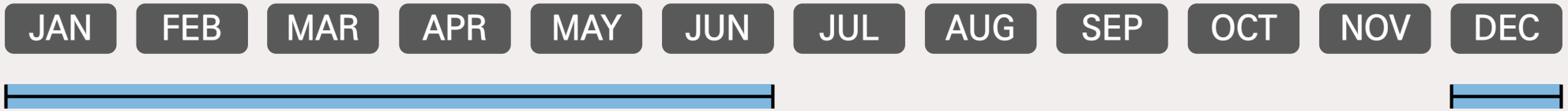
모델	평균 MAPE	표준편차 MAPE	평균 RMSE	표준편치 RMSE
XGB	55.928789	8.282093	287.322327	95.108201
LGBM	55.808508	8.186330	285.363119	97.180959
GradientBoosting	55.922405	7.948343	284.429031	95.280068
RandonForest	56.331602	7.900177	286.836407	96.412994

* 파라미터 튜닝없이 기본 모델로 성능 평가

1~5월 → 12월 예측 / 2~6월 → 12월 예측

모델별 평균 MAPE가 약 55-56로 나타남

모델링 - 6개월



모델	MAPE	RMSE
XGB	57.661173	261.153322
LGBM	56.031349	255.675154
GradientBoosting	55.888261	257.096290
RandonForest	56.275740	259.094053

* 파라미터 튜닝없이 기본 모델로 성능 평가

1~6월 → 12월 예측

모델 MAPE가 약 55-57로 나타남

최종모델링은 MAPE가

가장 낮게 나온 5개월로 진행

모델링 – 랜덤포레스트

1~5월 Train → 11월 Validation

파라미터	최적 파라미터
Max_depth	10
Max_features	0.8575162278238667
Min_sample_leaf	8
N_estimators	156

55.47889060406135

2~6월 Train → 12월 Validation

파라미터	최적 파라미터
Max_depth	10
Max_features	0.8
Min_sample_leaf	1
N_estimators	250

64.98934102892477

모델링 – 그래디언트 부스팅

1~5월 Train → 11월 Validation

파라미터	최적 파라미터
Max_depth	10
Subsample	0.8
Min_sample_leaf	5
N_estimators	185
Learning_rate	0.05
Min_sample_split	5

56.81519509615555

2~6월 Train → 12월 Validation

파라미터	최적 파라미터
Max_depth	10
Subsample	0.8
Min_sample_leaf	5
N_estimators	184
Learning_rate	0.05
Min_sample_split	5

68.43806396730682

모델링 – XGBoost

1~5월 Train → 11월 Validation

파라미터	최적 파라미터
Colsample	0.7
Gamma	5.0
Learning_rate	0.05
Max_depth	10
N_estimator	232
Subsample	0.8

54.61635110142286

2~6월 Train → 12월 Validation

파라미터	최적 파라미터
Colsample	0.7
Gamma	5.0
Learning_rate	0.05
Max_depth	10
N_estimator	232
Subsample	0.8

64.42541586317887

모델링 – LightGBM

1~5월 Train → 11월 Validation

파라미터	최적 파라미터
colsample_bytree	0.8121268904290354
learning_rate	0.12243251547182638
max_depth	7
min_child_weight	4.982767871318733
n_estimators	153
num_leaves	10
subsample	0.8879832560752989

53.27905335512161

2~6월 Train → 12월 Validation

파라미터	최적 파라미터
colsample_bytree	0. 8121268904290354
learning_rate	0.12243251547182638
max_depth	7
min_child_weight	4.982767871318733
n_estimators	153
num_leaves	10
subsample	0.8879832560752989

61.6731801144362

모델링 – 앙상블

1~5월 Train → 11월 Validation

모델	검증
랜덤포레스트	55.47889060406135
그레디언트부스팅	56.81519509615555
XGBoost	54.61635110142286
LightGBM	53.27905335512161

2~6월 Train → 12월 Validation

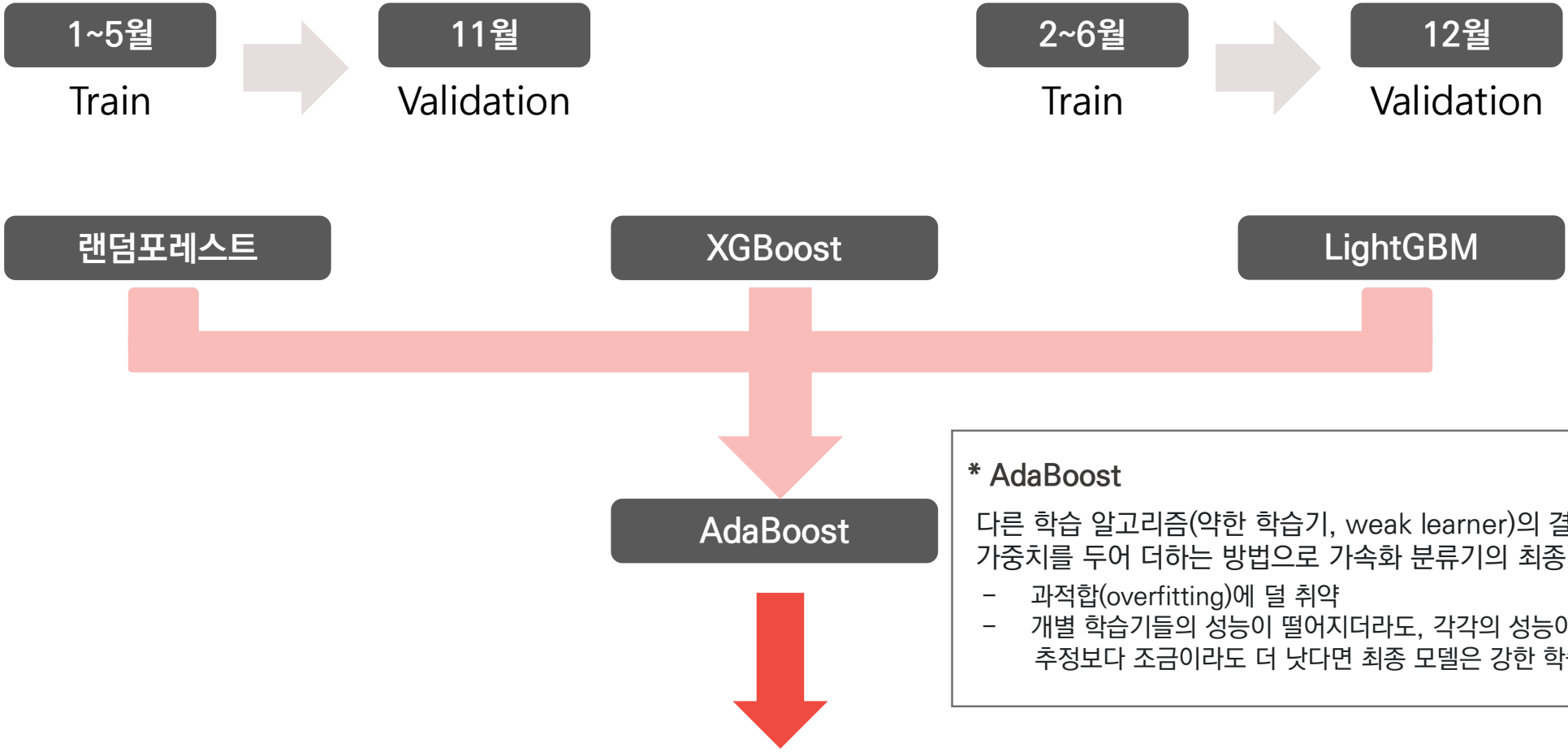
모델	검증
랜덤포레스트	64.98934102892477
그레디언트부스팅	68.43806396730682
XGBoost	64.42541586317887
LightGBM	61.6731801144362

상위 3개 모델을 선택하여 기하평균 진행

53.09658968136469

62.49316789678772

모델링 – 스태킹



*** AdaBoost**

다른 학습 알고리즘(약한 학습기, weak learner)의 결과물들을
가중치를 두어 더하는 방법으로 가속화 분류기의 최종 결과물을 표현

- 과적합(overfitting)에 덜 취약
- 개별 학습기들의 성능이 떨어지더라도, 각각의 성능이 무작위 추정보다 조금이라도 더 낫다면 최종 모델은 강한 학습기로 수렴

54.94934208464506

64.43287730614112

모델링 – 최종모델

1. 앙상블과 스태킹 비교결과
》 앙상블이 더 좋은 성능을 보임
》 최종 모델은 앙상블로 결정

2. 1-5월과 2-6월 비교결과
》 1-5월 성능이 더 우수
》 1-5월 하이퍼 파라미터로 결정

3. 5개월을 사용하여 모델예측
》 학습 데이터 중 8-12월을
사용하여 2020년 6월 예측

결론 및 활용방안

5

최적화 방안 – 신규 고객 마케팅

부대찌개 먹는 부대찌개 끝판왕

자카드 거리 계산

소들녘 소 갈비탕
하늘내린 용대리황태10미
피시원 국내산 손질 대구 8팩 * 매운탕 양념 8팩
멋진밥상 흥양농협 햅쌀 20KG
우리바다 손질왕꼬막 24팩

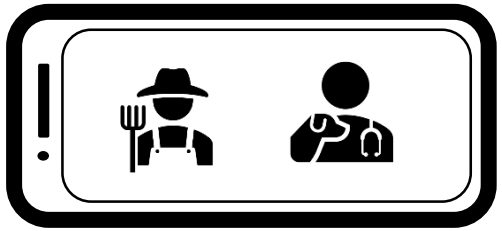
비슷한 이웃 기반 추천
상위 2개

소들녘 더 맛있는 돼지왕구이 팩
[맛있는 제주]손질 생선 대세트광어갈치고등어

유명 영상 크리에이터가 올린 영상 제목과 NS홈쇼핑에서 판매한 상품 간 자카드 거리를 계산하여 비슷한 이웃 기반 추천서비스를 도입한다.

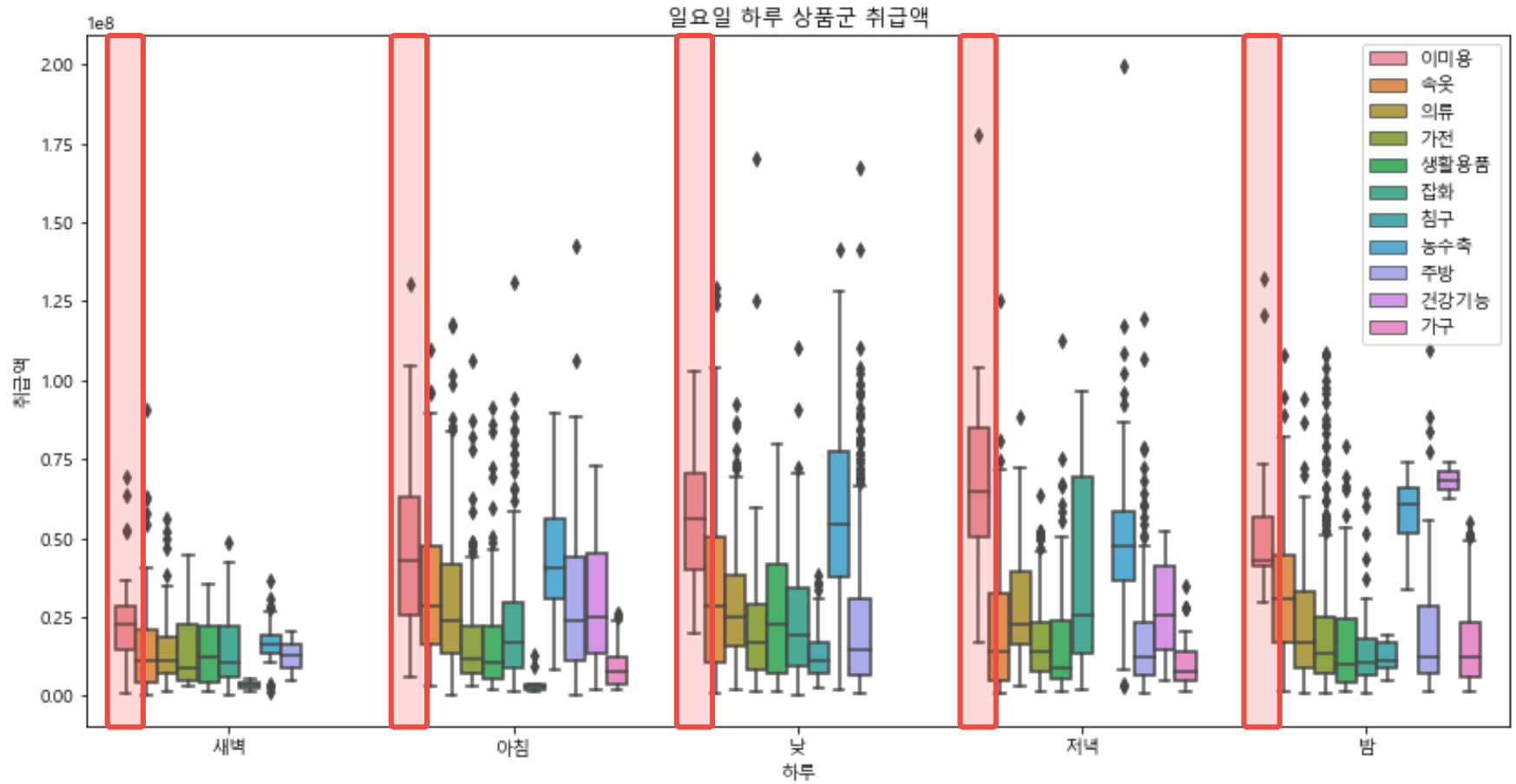
각 상품군별 유명 영상 크리에이터들과 협업을 통해서 NS홈쇼핑의 상품의 노출 및 홍보 채널을 확보하고 고객 유입의 확실한 효과를 누릴 수 있다.

예를 들어 시청자가 100만이 넘는 영상 크리에이터와 콜라보를 통해 각 지역의 특산물 등을 판매하는 방송을 진행한다.



NS홈쇼핑

최적화 방안 – 집중도 있고 생생한 프로그램 구성

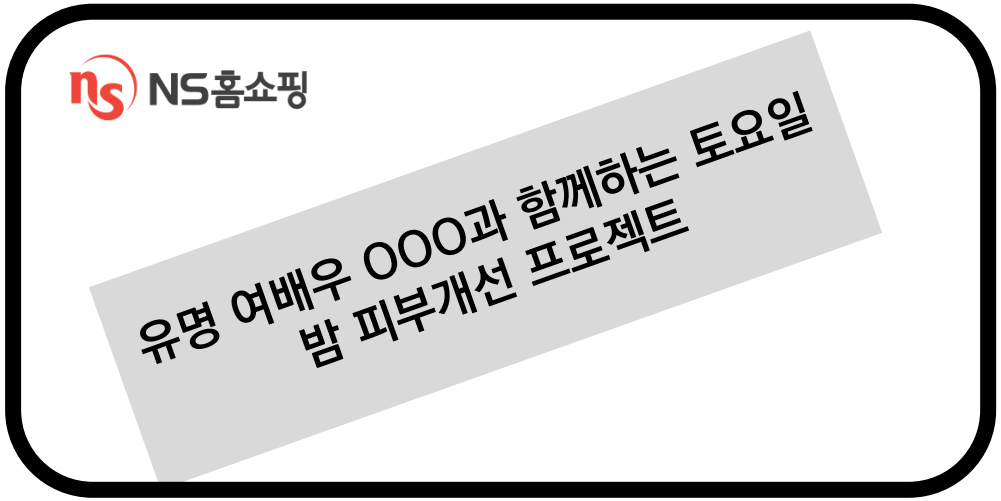


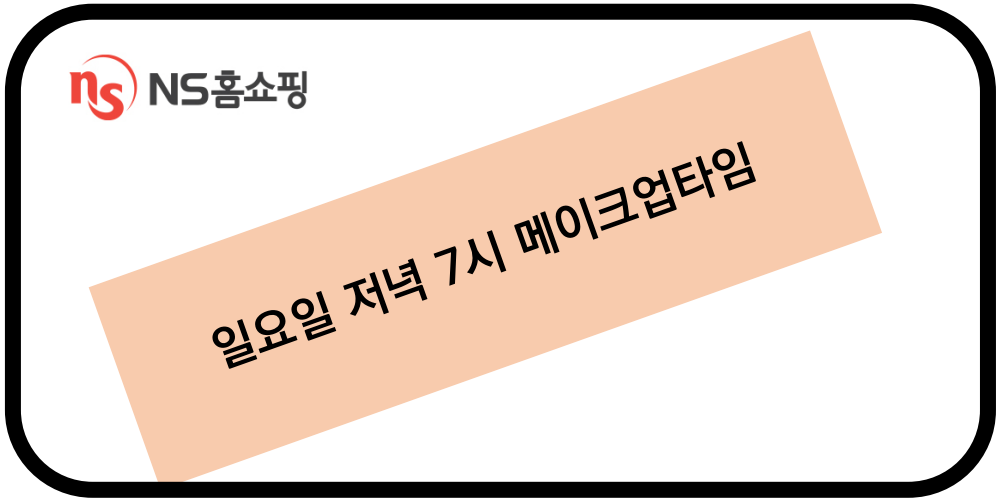
새벽(0-5시)
아침(5-12시)
낮(12-17시)
저녁(17-21시)
밤(21-24시)

〉〉 일요일에 이미용은 모든 시간대에 항상 잘 팔린다.

〉〉 이처럼 요일별로 잘 팔리는 상품군을 파악하여 방송배치 하는 것이 효과적이다.

최적화 방안 – 집중도 있고 생생한 프로그램 구성





최근 다양한 채널을 통해 연예인들이 콘텐츠 소비자들과 활발한 교류를 하고 있다. 이렇게 소비자들과 연예인 사이의 거리가 가까워지는 추세를 활용하여 지상파 방송의 고정 프로그램 처럼 주기적으로 다양한 연예인들을 방송에 초대해 고객들의 집중도를 높인다.

예를들어 이미용 상품의 경우 실제로 연예인들이 방송에서 before after를 방송 진행 과정에서 보여주면서 보다 생생한 경험을 고객들에게 간접적으로 전달해 줄 수 있다. 더불어 매주 정규프로그램처럼 프로그램을 편성하여 고정 시청층을 끌어들일 수 있다.

최적화 방안 – 짜임새 있는 편성 시간대 활용

판매실적 부진 제품

상품명	상품군	판매량	판매단가
도루코 페이스5 면도날 4개입	생활용품	8.046875	12800
1세트 센티멘탈 디퓨저골드	생활용품	5.774193	31000
기본구성 바로톡 싱크대 거름통	생활용품	4.9388753	40900
여자를 위한 빨강 팥물, 레드빈 티톡 50포	건강기능	4.976190	42000
디키즈 아동 스테디움점퍼세트	의류	4.074074	54000

매진제품

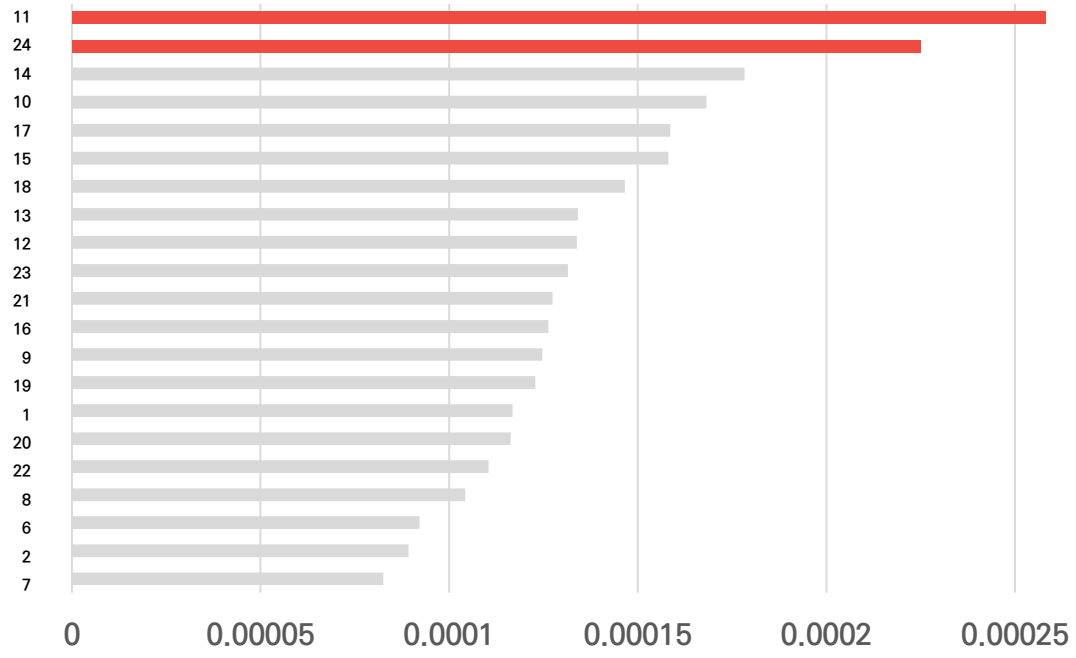
상품명	노출(분)
레이프릴 무빙 맥시풀커버 브라팬티	20
레이프릴 무빙 맥시풀커버 브라팬티	20
레이프릴 무빙 맥시풀커버 브라팬티	17.1

〉〉 2.9분 동안 보여줄 광고성 영상제작

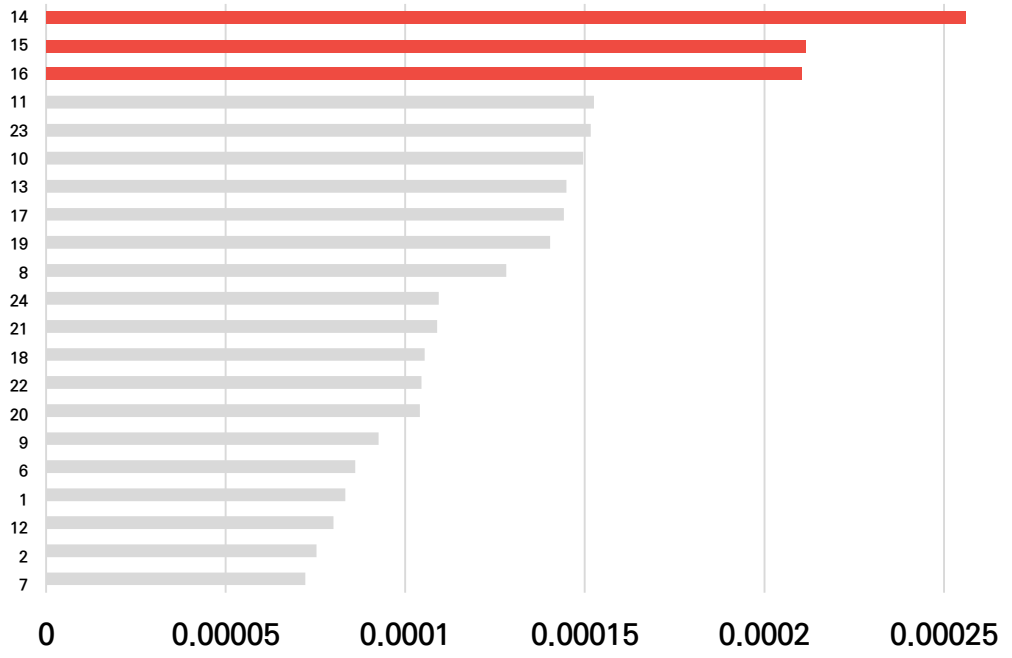
데이터를 살펴보면 중간중간 편성시간보다 방송이 일찍 끝나는 경우들이 빈번하게 등장한다. 이러한 시간에는 뒤의 시간대 방송을 앞당겨서 하는 것이 아니라 이전 주 혹은 더 과거에 판매실적이 부진했던 제품들을 여러 기준으로 필터링하여 할인을 대폭 적용하거나 게릴라성 이벤트를 실시하여 재고도 활용하고 소비자로 하여금 특권을 누리는 감정을 느끼게 해 충성도를 높일 수 있다.

최적화 방안 – 신상품 소개 시간대

〈 평일 시청률 분산 〉



〈 주말 시청률 분산 〉



평일과 주말 시간대 시청률을 보면 분산이 유독 큰 시간대가 존재한다.
이 시간대에 새로운 상품을 소개하여 수시로 채널을 돌리는 시청자들에게 홍보효과를 가져다 줄 것이다.
만약 상품이 자주 팔린다면, 상품이 많이 팔리는 주말 낮 시간대나 평일 저녁 시간대로 옮길 수 있다.

2020 빅콘테스트
2020 BIG CONTEST

NS홈쇼핑

감사합니다.

insomNIA _ 기세현, 민규선, 하진용