2020 빅콘테스트 챔피언 리그

[NS SHOP+ 판매실적 예측을 통한 편성 최적화 방안(모형) 도출]

이병헌 (gijo0104@naver.com)

INDEX

01 분석 배경 및 분석 목표

- 분석 배경
- 분석 목표

분석 배경

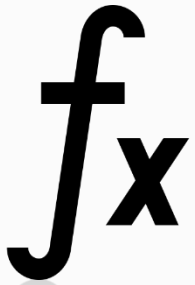
- 최근 빅데이터 기반 인공지능 분석을 TV 홈쇼핑에 적용하는 다양한 사례가 증가하고 있다.

- 과거에는 상품기획자 및 편성 담당자의 경험과 주관에 따라 결정되던 방송 편성 방식이 최근에는 빅데이터 기반 인공지능 기법을 통해 과거 데이터를 통계적으로 분석하여 모델을 만들어 내는 적극적인 시도가 이루어 지고 있다.

- 이러한 흐름에 따라 데이터 분석을 진행하여 판매실적에 영향을 주는 요인을 탐색하고 판매실적을 정확하게 예측할 수 있는 모형을 구축하고자 한다.

분석 목표

매출액 추정



판매 실적에 영향을
주는 요인 탐색

취급액 증대를 위한
정보 및 제안

02 데이터 소개 및 전처리

- 데이터 소개
- 데이터 병합
- 주요 변수 소개
- 전처리 과정

데이터 소개

1. 제공 데이터: 분당 실적 데이터

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
2019-01-01 6:20		100346	201079	테이트 여성 셀린니트3종	의류	39,900	6,955,000
2019-01-01 6:40	20	100346	201072	테이트 남성 셀린니트3종	의류	39,900	6,672,000
2019-01-01 6:40		100346	201079	테이트 여성 셀린니트3종	의류	39,900	9,337,000
2019-01-01 7:00	20	100305	200974	오모떼 레이스 파운데이션 브라	속옷	59,000	6,819,000
2019-01-01 7:20	20	100305	200974	오모떼 레이스 파운데이션 브라	속옷	59,000	15,689,000
2019-01-01 7:40	20	100305	200974	오모떼 레이스 파운데이션 브라	속옷	59,000	25,370,000

38,309행 8개 변수 (분당 실적 데이터 기준)
2019년 1월~12월 데이터

데이터 소개

2. NS shopping mall 사이트 방송 편성표

(http://www.nsmall.com/TComLiveBrodcastingList?tab_gubun=1&tab_Week=1&tab_bord=0&selectDay=2019-01-01)

활용 목적

1. 실제 해당 상품에 대한 후기를 반영한 별점 및 투표수는 매출액에 직접적인 영향을 끼칠 수 있음
2. 브랜드명이 명확히 명시되어 있으므로 이를 활용하여 브랜드 지수화 가능

날짜	방송시간	브랜드명	상품명	별점	별점 투표수	세일가격	판매가격
2019-01-01	오전 2:00 ~ 오전 3:00	[마리끌레르]	[NS Shop+]마리	97	61	38900	39900
2019-01-01	오전 3:00 ~ 오전 4:00	[Cerini by PAT]	[NS Shop+]CER	95	21	68900	69900
2019-01-01	오전 4:00 ~ 오전 5:00	[트레스패스]	[NS Shop+]트레	88	5	48800	49800
2019-01-01	오전 4:00 ~ 오전 5:00	[트레스패스]	[NS Shop+]트레	98	5	48800	49800
2019-01-01	오전 5:00 ~ 오전 6:00	[쿠미투니카]	[NS Shop+]쿠미	97	29	88900	89900
2019-01-01	오전 6:00 ~ 오전 7:00	[테이트]	[NS Shop+]테이	90	8	38900	39900

데이터 소개

2. NS shopping mall 사이트 방송 편성표

(http://www.nsmall.com/TComLiveBrodcastingList?tab_gubun=1&tab_Week=1&tab_bord=0&selectDay=2019-01-01)

방송시간	상품정보	가격	구매하기
오전 2:00 ~ 오전 3:00 [마리끌레르]	 [마리끌레르] [NS Shop+]마리끌레르 파리컬렉션 리얼 하이드로 립스틱 ★★★★★ 61건	~~39,900원~~ 38,900원	방송중 구매가능 방송알림

날짜	방송시간	브랜드명	상품명	별점	별점 투표수	세일가격	판매가격
2019-01-01	오전 2:00 ~ 오전 3:00	[마리끌레르]	[NS Shop+]마리	97	61	38900	39900
2019-01-01	오전 3:00 ~ 오전 4:00	[Cerini by PAT]	[NS Shop+]CER	95	21	68900	69900
2019-01-01	오전 4:00 ~ 오전 5:00	[트레스패스]	[NS Shop+]트레	88	5	48800	49800
2019-01-01	오전 4:00 ~ 오전 5:00	[트레스패스]	[NS Shop+]트레	98	5	48800	49800
2019-01-01	오전 5:00 ~ 오전 6:00	[쿠미투니카]	[NS Shop+]쿠미	97	29	88900	89900
2019-01-01	오전 6:00 ~ 오전 7:00	[테이트]	[NS Shop+]테이	90	8	38900	39900

데이터 소개

3. 기상청 데이터
(https://data.kma.go.kr)

활용 목적

1. 특정한 날씨와 기상 조건은 고객들의 생활습관에 영향을 미칠 것으로 판단
2. 종관 기상관측(ASOS) 자료의 기온, 강수량, 풍속, 습도, 증기압, 이슬점, 일조량 그리고 일사량 등의 일부 지역(서울, 광주, 대구, 대전, 부산, 서울, 인천)의 시간별 자료를 사용

지점명	일시	기온(°C)	강수량(mm)	풍속(m/s)	습도(%)	증기압(hPa)
서울	2018-12-31 22:00	-5.4	NA	1.3	46	1.9
서울	2018-12-31 23:00	-5.2	NA	1.6	47	2
서울	2019-01-01 0:00	-5.5	NA	1	54	2.2
서울	2019-01-01 1:00	-5.9	NA	1.8	56	2.2
서울	2019-01-01 2:00	-6.5	NA	1.2	60	2.3
서울	2019-01-01 3:00	-6.9	NA	2.2	62	2.3

데이터 소개

4. 공휴일 데이터
(https://data.go.kr)

활용 목적

1. 공휴일일 때의 행동 양상이 다를 것이고 이는 소비 패턴에 영향을 미칠 것으로 판단

명칭	날짜
1월 1일	20190101
설날	20190204
설날	20190205
설날	20190206
삼일절	20190301
어린이날	20190505

데이터 소개

5. 소비자 심리지수 데이터
(https://www.index.go.kr)

활용 목적

1. 소비자 심리지수 : 현재 생활형편, 가계수입전망 등 6개 주요 개별지수를 표준화하여 합성한 지수
2. 소비자심리를 종합적으로 판단하는 지수로서 실제 소비에 영향을 미칠 것으로 판단

201812월	201901월	201902월	201903월	201904월	201905월	201906월
96.9	97.5	99.5	99.8	101.6	97.9	97.5

데이터 소개

6. 시청률 데이터
(https://www.nielsenkorea.co.kr)

활용 목적

1. 채널을 돌리는 행위인 재핑(Zapping)은 홈쇼핑 매출에 직접적인 영향을 주는 것으로 알려져 있음
2. 재핑과 관련된 변수로 홈쇼핑 매출에 영향을 미칠 것으로 판단

날짜	지역	분류	순위	채널	프로그램명	값
20190101	전국	시청률	1	KBS1	일일연속극(비켜라운명아)	15.6
20190101	전국	시청률	2	KBS1	KBS9시뉴스	11.8
20190101	전국	시청률	2	KBS2	신년특선영화(관상1부)	11.8
20190101	전국	시청률	4	KBS1	KBS뉴스7	10.8
20190101	전국	시청률	5	KBS2	신년특선영화(관상2부)	10.3
20190101	전국	시청률	6	KBS1	KBS뉴스(09:30)	9.9

데이터 소개

6. 시청률 데이터
(https://www.nielsenkorea.co.kr)

전국 수도권

2019.01.01

가구시청률 TOP 20

(분석기준: 전국13개지역, 가구, 단위:%)

순위	채널	프로그램	시청률
1	KBS1	일일연속극(비켜라운명아)	15.6

시청자수 TOP 20

(분석기준: 전국13개지역, 개인, 단위:천 명)

순위	채널	프로그램	시청자수
1	KBS1	일일연속극(비켜라운명아)	2,446

날짜	지역	분류	순위	채널	프로그램명	값
20190101	전국	시청률	1	KBS1	일일연속극(비켜라운명아)	15.6
20190101	전국	시청률	2	KBS1	KBS9시뉴스	11.8
20190101	전국	시청률	2	KBS2	신년특선영화(관상1부)	11.8
20190101	전국	시청률	4	KBS1	KBS뉴스7	10.8
20190101	전국	시청률	5	KBS2	신년특선영화(관상2부)	10.3
20190101	전국	시청률	6	KBS1	KBS뉴스(09:30)	9.9

데이터 변환

분당 실적 데이터

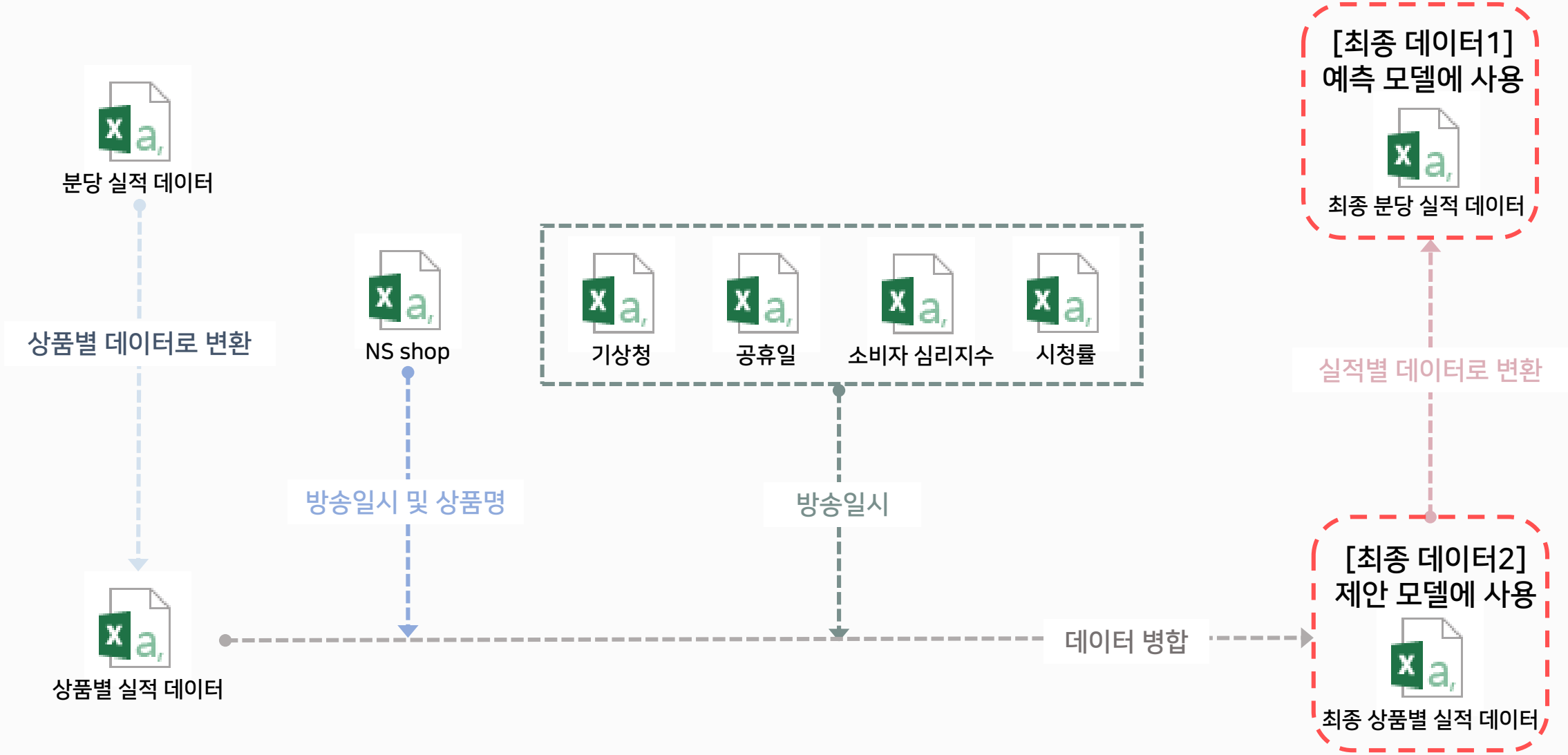
상품별 데이터로 변환

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
2019-01-01 6:20		100346	201079	테이트 여성 셀린니트3종	의류	39,900	6,955,000
2019-01-01 6:40	20	100346	201072	테이트 남성 셀린니트3종	의류	39,900	6,672,000
2019-01-01 6:40		100346	201079	테이트 여성 셀린니트3종	의류	39,900	9,337,000
2019-01-01 7:00	20	100305	200974	오모떼 레이스 파운데이션 브라	속옷	59,000	6,819,000
2019-01-01 7:20	20	100305	200974	오모떼 레이스 파운데이션 브라	속옷	59,000	15,689,000
2019-01-01 7:40	20	100305	200974	오모떼 레이스 파운데이션 브라	속옷	59,000	25,370,000

상품별 실적 데이터

방송일시	노출(분)	마더코드	상품코드	상품명	상품군	판매단가	취급액
2019-01-01 6:00	60	100346	201072	테이트 남성 셀린니트3종	의류	39,900	12,033,000
2019-01-01 6:00	60	100346	201079	테이트 여성 셀린니트3종	의류	39,900	20,663,000
2019-01-01 7:00	60	100305	200974	오모떼 레이스 파운데이션 브라	속옷	59,000	47,878,000

데이터 병합


분당 실적 데이터
상품별 데이터로 변환
상품별 실적 데이터
NS shop
방송일시 및 상품명
기상청
공휴일
소비자 심리지수
시청률
방송일시
데이터 병합
[최종 데이터1]
예측 모델에 사용
최종 분당 실적 데이터
실적별 데이터로 변환
[최종 데이터2]
제안 모델에 사용
최종 상품별 실적 데이터

주요 변수 소개

분당 실적 데이터

방송일시
노출(분)
마더코드
상품코드
상품명
상품군
판매단가
취급액

시청률 데이터

일별_시청률.(채널)
KBS1
KBS2
OCN
SBS
tvN
MBC
MBCevery1
연합뉴스TV
YTN

기상청 데이터

기온
강수량
풍속
습도
증기압
일조량
일사량

X

서울
광주
대구
대전
부산
인천

공휴일 데이터

공휴일 여부

CSI 데이터

소비자 심리지수

NS shop 데이터

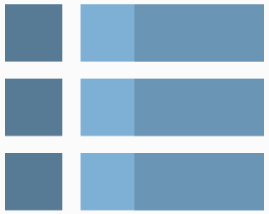
브랜드명
상품명
별점
별점 투표수
세일가격
판매가격

데이터 변환

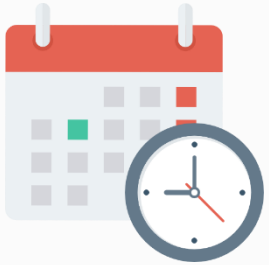
파생변수 생성 전 중요한 변수

취급액

방송시간



상품군



방송일시

브랜드명

별점 투표수

전처리 과정

1. 상품별 실적 데이터 및 NS shop 데이터 : 상품명

- 상품별 실적데이터와 NS shop 데이터를 병합하는 과정에서 상품명이 조금씩 달라서 병합에 힘든 부분이 존재

상품별 실적 데이터 상품명	NS shop 데이터 상품명
(일) 선일금고 이볼브 시리즈 EV-040	일시불 선일금고 이볼브 시리즈 EV-040
보루네오 루나 유로탑 멀티수납형 LED 침대 퀸	보루네오 루나 유로탑 멀티수납형 LED 침대 Q 퀸
무이자 올리고 가스와이드그릴레인지 프리미엄형	(무이자)올리고 가스와이드그릴 프리미엄형
(퀸+퀸)일월 품안애 온수매트	일월 품안애 온수매트 퀸+퀸

1. (일), [일], 일 -> 모두 일시불로 통일, (무), [무], 무 -> 모두 무이자로 통일
2. 특수 기호, 띄어쓰기 및 알파벳의 경우 제거 후 같은 상품명이 되는 경우 병합
3. 이 외 매칭되지 않는 경우 직접 수작업으로 고친 후 작업

전처리 과정

2. NS shop 데이터 : 방송일시

- 상품별 실적데이터와 NS shop 데이터를 병합하는 과정에서 방송일시가 조금씩 달라서 병합에 힘든 부분이 존재

상품명	상품별 실적 데이터 방송일시	NS shop 방송일시
로베르타 디 까메리노 Y밸런스업 지퍼쉐이핑 란쥬	2019-03-31 11:00:00 AM	오전 10:57 ~ 오전 11:57
아리스토우 여성오가닉티셔츠	2019-03-31 11:20:00 PM	오후 11:19 ~ 오전 12:19
[루이띠에] 24K 999 순금 37.5g 골드바 목걸이	2020-06-30 01:20:00 PM	오후 01:23 ~ 오후 02:23

1. 방송일시가 -4분부터 +4분까지 차이가 나는 경우가 존재하고 병합하는 과정에서 Ns shop 방송일시가 더 빠를 경우에 문제가 발생하므로 56분, 57분, 18분, 19분 등에 대하여 정각에 맞게끔 시간을 더함
2. 크롤링 과정에서 오전 00시를 넘어갈 때 날짜가 바뀌지 않으므로 하루를 더하여 이를 해결
3. 오전 2시 ~ 오전 6시까지 새벽 방송 제외

전처리 과정

3. NS shop 데이터 : 브랜드명

- NS shop 데이터 중 브랜드명이 [미정의]인 경우가 존재하지만
  상품명이 동일하면서 브랜드명이 [미정의]가 아닌 데이터가 존재하여 이를 사용하여 대입

4.시청률 데이터

- 2019년 2월 23일의 케이블 방송에 대한 시청률 및 시청자 수 데이터가 결측치이고
  이는 일주일 전인 2019년 2월 16일의 케이블 방송에 대한 시청률 및 시청자 수로 대체
  (방송 프로그램 주기가 일주일인 것을 고려)

전처리 과정

5. 기상청 데이터 : 강수량, 풍속, 일조량

- 단위가 0.1이고 결측치와 0인 값 모두 존재하므로 결측치와 0인 값의 차이를 둠
- 0인 값은 반올림을 하여서 생긴 값으로 간주하고 0~0.05 사이의 값으로 생각하여 0.025를 대입하고 결측치는 0을 대입

6. 기상청 데이터 : 일사량

- 단위가 0.01이고 결측치와 0인 값 모두 존재하므로 결측치와 0인 값의 차이를 둠
- 0인 값은 반올림을 하여서 생긴 값으로 간주하고 0~0.005 사이의 값으로 생각하여 0.0025를 대입하고 결측치는 0을 대입

전처리 과정

7. 분당 실적 데이터 : 상품군

- 상품군이 무형인 경우 판매 단가가 0원으로 취급액 또한 0원이므로 제거

8. 분당 실적 데이터 : 취급액

- 상품군이 무형이 아님에도 취급액이 0인 경우가 존재하고 이는 모델 평가기준인 MAPE를 계산할 때 제외되므로 제거할 수도 있으나 취급액이 0인 것도 정보라고 생각하여 판매단가 대비 작은 값을 대입

(판매단가*0.5)

방송일시	노출(분)	마더코드	상품코드	상품명	상품군	판매단가	취급액
2019-01-04 10:00:00	20	100837	202456	(도냄)무이자 쿠쿠전기밥솥 6	주방	208,000	3,010,000
2019-01-04 10:20:00	20	100837	202456	(도냄)무이자 쿠쿠전기밥솥 6	주방	208,000	0 -> 208000*0.5
2019-01-04 10:40:00	20	100837	202456	(도냄)무이자 쿠쿠전기밥솥 6	주방	208,000	8,176,000
2019-03-03 15:00:00	20	100182	200612	무이자 선일금고 이볼브 시리	생활용품	440,000	0 -> 440000*0.5
2019-03-03 15:20:00	20	100182	200612	무이자 선일금고 이볼브 시리	생활용품	440,000	0-> 440000*0.5
2019-03-03 15:40:00	20	100182	200612	무이자 선일금고 이볼브 시리	생활용품	440,000	0-> 440000*0.5

전처리 과정


상품별 실적 데이터

상품명 조정


Ns shop

상품명, 방송일시 조정

브랜드명 [미정의] 채우기

판매가격 결측치 채우기


기상청

강수량, 풍속, 일조량, 일사량

-> 값 0, 결측치 채우기


시청률

2019.02.23 결측치 채우기

전처리 후 병합


분당 실적 데이터

상품군 무형 제거

취급액 0 -> 판매단가*0.5 대입

03 데이터 탐색 및 파생 변수

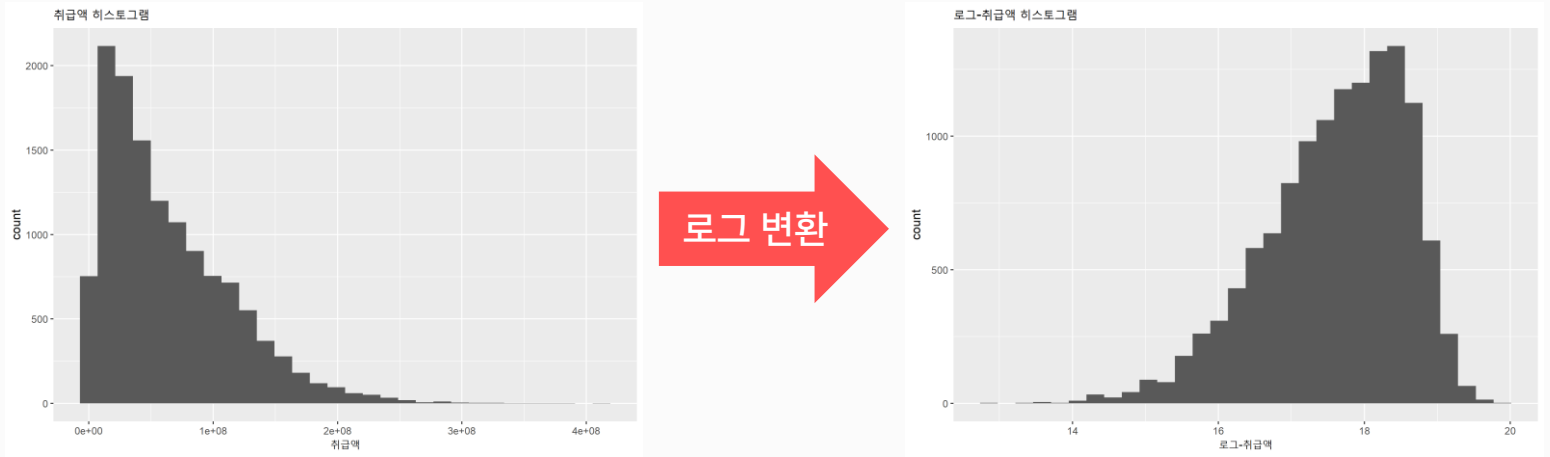
- 종속 변수
- 파생 변수
- 사용 변수 정리

종속 변수

취급액 로그 변환

취급액 → 로그_취급액 = log(취급액) → 로그_취급액

<기존 변수> <파생 변수>



오른쪽으로 꼬리가 긴 데이터를 로그 변환 함으로써 대칭적인 데이터로 변환

평가 기준인 MAPE의 경우 로그 변환을 할 때 더 좋은 결과를 가진다고 알려져 있음

판매수

취급액, 가격 → $판매수 = \frac{취급액}{가격}$ → 판매수

<기존 변수> <파생 변수>

파생 변수

1. 상품특성관련변수
2. 방송일시관련변수
3. 방송시간관련변수
4. 인기관련변수
5. 브랜드관련변수
6. 상품명관련변수
7. 과거실적관련변수
8. 외부변수
9. 분당실적데이터 관련변수

판매단가

판매단가

<기존 변수>

상품군

상품군

<기존 변수>

지불 방식

상품명

<기존 변수>

상품명	지불방식
무이자, 일시불 포함하지 않을시	0
무이자 포함시	1
일시불 포함시	2

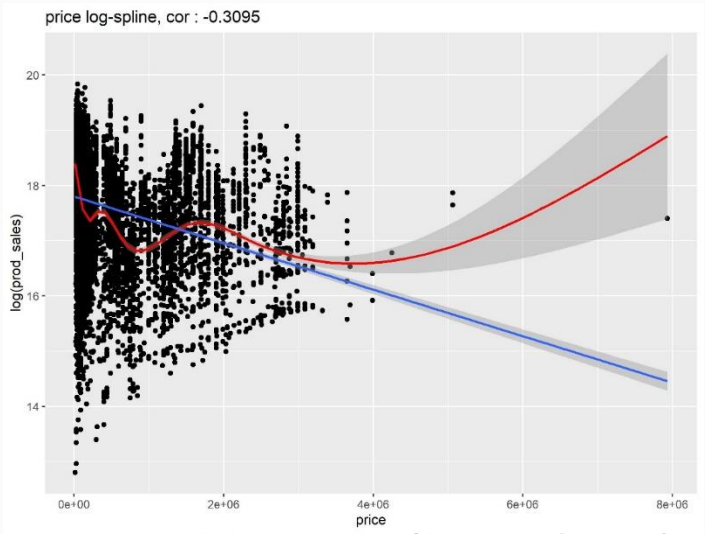
지불 방식

<파생 변수>

파생 변수

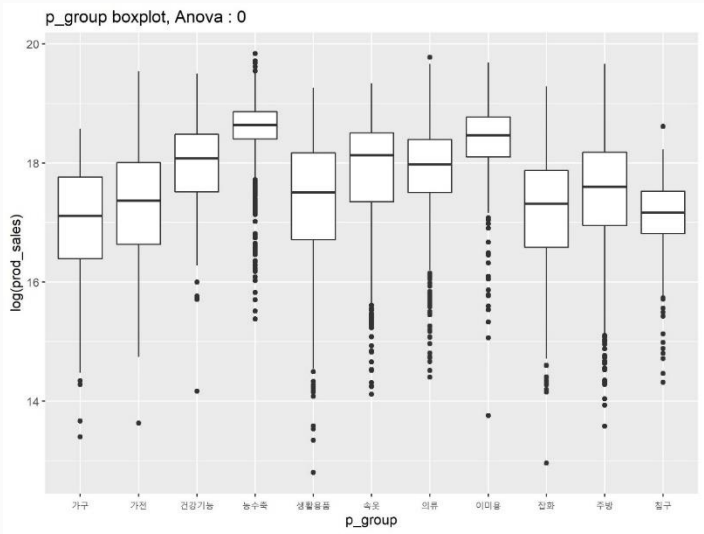
1. 상품특성관련변수
2. 방송일시관련변수
3. 방송시간관련변수
4. 인기관련변수
5. 브랜드관련변수
6. 상품명관련변수
7. 과거실적관련변수
8. 외부변수
9. 분당실적데이터 관련변수

판매단가



빨강 : spline 적합, 파랑 : 회귀 적합

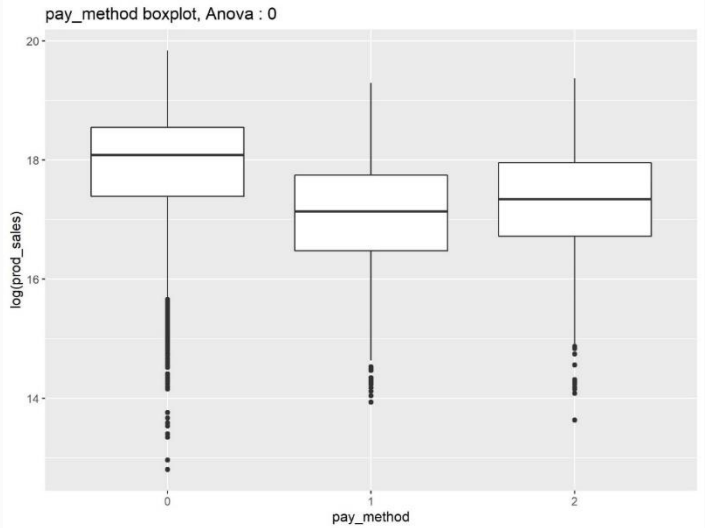
상품군



- 판매단가의 경우 상관계수가 -0.31로 취급액과 음의 상관관계를 가지고 상식적으로도 판매단가는 취급액에 영향을 끼칠 것으로 예상된다.
- 상품군의 경우 박스-상자 그림과 같이 간단히 보아도 각 상품군마다 다른 취급액 분포를 보임을 알 수 있다.

파생 변수

지불방식



왼쪽부터 0(무이자 일시불X), 1(무이자), 2(일시불)

- 지불방식의 경우 무이자나 일시불이 아닌 경우의 상품들의 취급액이 일시불이나 무이자 행사 상품들보다 대략 2.0배가 더 높은 것으로 나타났다. 이는 일시불이나 무이자 행사 상품들의 판매단가가 아닌 상품들의 대략 4.4배나 더 크고 이는 앞서 보았던 판매액과 취급액과의 상관관계가 음인 것이 반영되었다고 보인다.
- 일시불의 평균 취급액은 무이자보다 대략 1.2배만큼 더 크고 이는 가격이 더 저렴하기 때문으로 보인다.

파생 변수

방송시작 시각

방송일시 <기존 변수>

방송일시	방송시작 시각
2019-01-01 14:20:00	14

0, 1, 6 ~ 23까지 범주형 변수

지불 방식 <파생 변수>

요일

방송일시 <기존 변수>

방송일시	요일
2019-01-01 14:20:00	화요일

월,화,수,목,금,토,일요일까지 범주형 변수

지불 방식 <파생 변수>

공휴일

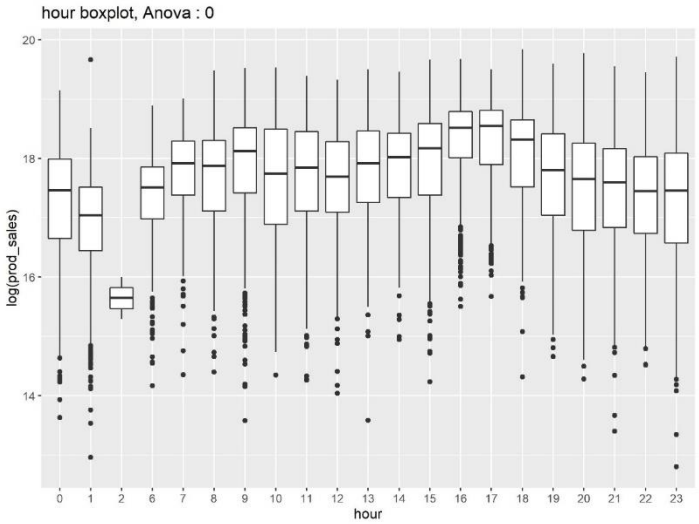
방송일시 <기존 변수>

방송일시	공휴일
공휴일 및 일요일이 아닌 경우	0
공휴일 및 일요일인 경우	1

지불 방식 <파생 변수>

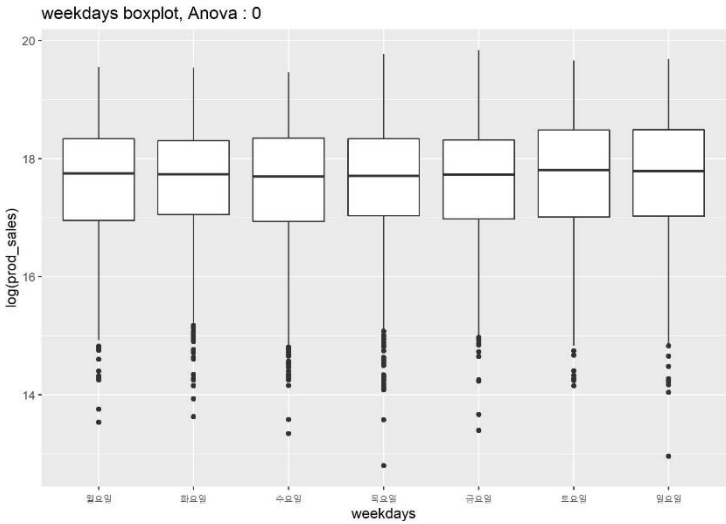
파생 변수

방송시작 시각



왼쪽부터 0시,1시,6시,7시,…,23시

요일

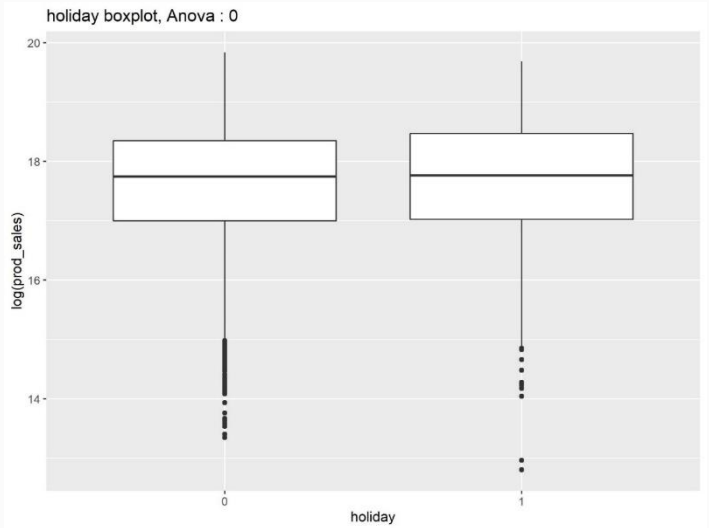


왼쪽부터 월,화,수,목,금,토,일요일

- 방송시작 시각마다 다른 취급액 분포를 가지는 것을 박스-상자 그림으로 알 수 있고 방송시작 시각마다 주 고객층이 다르기 때문에 방송시작 시각 변수는 취급액에 영향을 끼칠 것으로 보인다.
- 요일의 경우 취급액에 log 변환을 취하여 그린 박스-상자 그림이라 차이가 없어 보이지만 Anova 검정을 통하여 p-value가 거의 0에 가까워 취급액에 차이가 있음을 알 수 있고 특히 주말과 평일에서 큰 차이를 보임을 알 수 있다. 대략 1.14배만큼 주말의 취급액이 더 크다.

파생 변수

공휴일



왼쪽부터 공휴일 및 일요일이 아닌 경우, 맞는 경우

- 공휴일 및 일요일인 경우 취급액이 아닌 경우보다 평균 1.1배 더 큰 것으로 나타났고 공휴일과 아닌 경우의 사람들의 행동패턴이 취급액에 영향을 끼친 것으로 보인다.

파생 변수

방송시간(분당 실적 -> 상품별 실적)

노출(분) <기존 변수> ⇢ 방송시간 <파생 변수>

방송일시	노출(분)	상품명
2019-01-01 6:00	20	테이트 남성 셀린니트3종
2019-01-01 6:00	20	테이트 여성 셀린니트3종
2019-01-01 6:20	20	테이트 남성 셀린니트3종
2019-01-01 6:20	20	테이트 여성 셀린니트3종
2019-01-01 6:40	20	테이트 남성 셀린니트3종
2019-01-01 6:40	20	테이트 여성 셀린니트3종

방송일시	방송시간	상품명
2019-01-01 6:00	60	테이트 남성 셀린니트3종
2019-01-01 6:00	60	테이트 여성 셀린니트3종

방송 판매 상품수

방송일시 <기존 변수> ⇢ 방송 판매 상품수 <파생 변수>

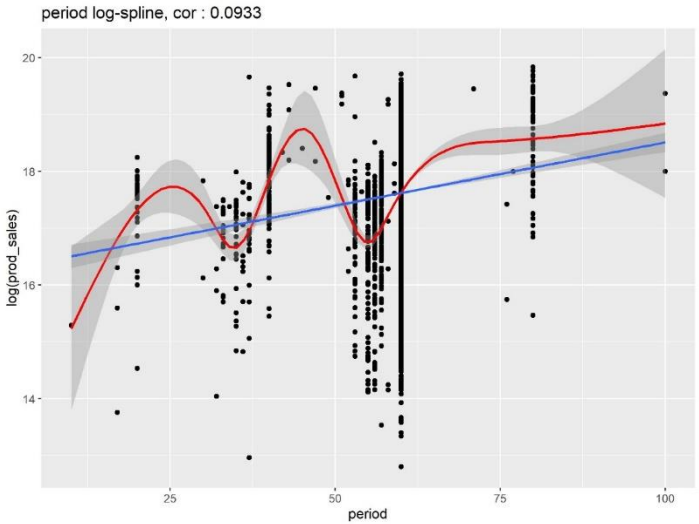
상품명	방송 판매 상품수
테이트 남성 셀린니트3종	2
테이트 여성 셀린니트3종	2

예상 노출시간

방송시간
방송판매 상품수 <파생 변수> ⇢ $\text{예상 노출시간} = \frac{\text{방송시간}}{\text{방송 판매 상품수}}$ ⇢ 예상 노출시간 <파생 변수>

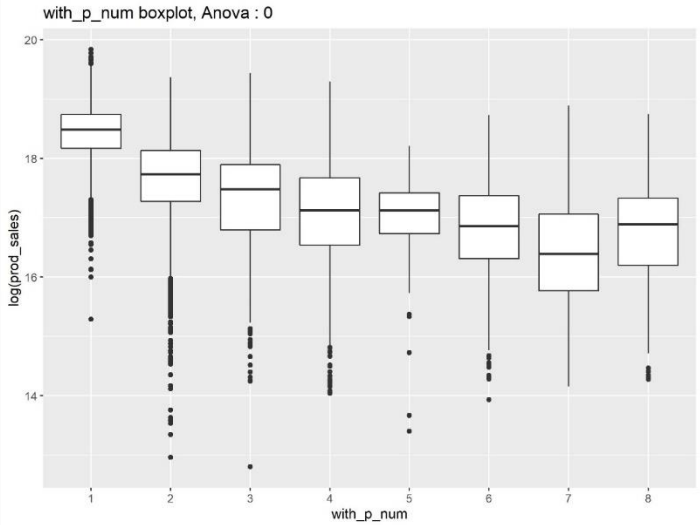
파생 변수

방송시간



빨강 : spline 적합, 파랑 : 회귀 적합

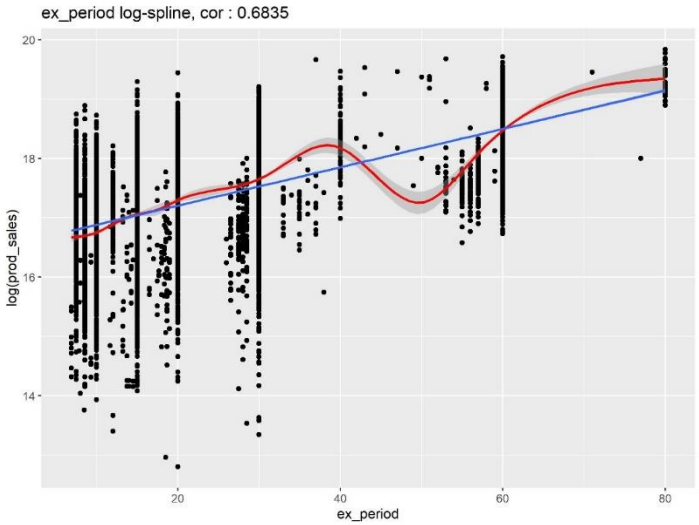
방송 판매 상품수



- 방송시간의 경우 방송시간이 길수록 취급액이 높을 것으로 예상할 수 있지만 상관계수가 0.0933으로 아주 약한 양의 상관관계를 보인다. 이는 방송 판매 상품수와 함께 살펴보아야 할 것으로 보여 예상 노출시간과 취급액의 관계를 다음 장에서 살펴볼 것이다.
- 방송 판매 상품수가 많을수록 낮은 취급액 분포를 보인다. 이는 한 번에 많은 상품을 팔게 되기 때문인 것으로 보이고 숫자형 변수로 변환한 상관계수는 무려 -0.617로 음의 상관관계를 가진다.

파생 변수

예상노출시간



빨강 : spline 적합, 파랑 : 회귀 적합

- 앞서 예상했던 것과 같이 예상 노출시간과 취급액의 상관계수는 0.6835로 강한 양의 상관관계를 보인다. 즉, 방송시간이 아닌 해당 방송에서 상품이 화면에 비춰지는 시간이 취급액과 큰 상관관계를 지닌다는 것을 알 수 있다.
  모델에서는 방송시간은 예측 변수로 사용하지 않고 방송 판매 상품수와 예상 노출시간만 예측 변수로 사용할 것이다.

파생 변수

별점

별점

<외부 변수>

New투표수

별점 투표수 → New투표수

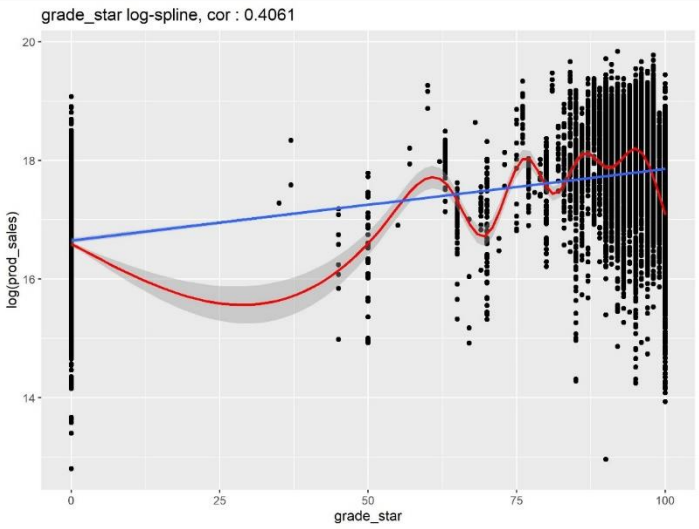
<외부 변수> <파생 변수>

방송일자	상품명	별점 투표수	New투표수
2019-10-08 13:00 – 14:00	[RYN] 린 여성 뉴웨이	50	$50 \div 4 \times 1 = 12.5$
2019-10-11 09:00 – 10:00	[RYN] 린 여성 뉴웨이	50	$50 \div 4 \times 2 = 25.0$
2019-10-15 14:00 – 15:00	[RYN] 린 여성 뉴웨이	50	$50 \div 4 \times 3 = 37.5$
2019-10-20 13:00 – 14:00	[RYN] 린 여성 뉴웨이	50	$50 \div 4 \times 4 = 50.0$

총 누적 별점 투표수가 기록되므로 이전 상품들의 누적 별점 투표수가 일정하게 증가했다는 가정하에 계산하여 새로운 변수 생성

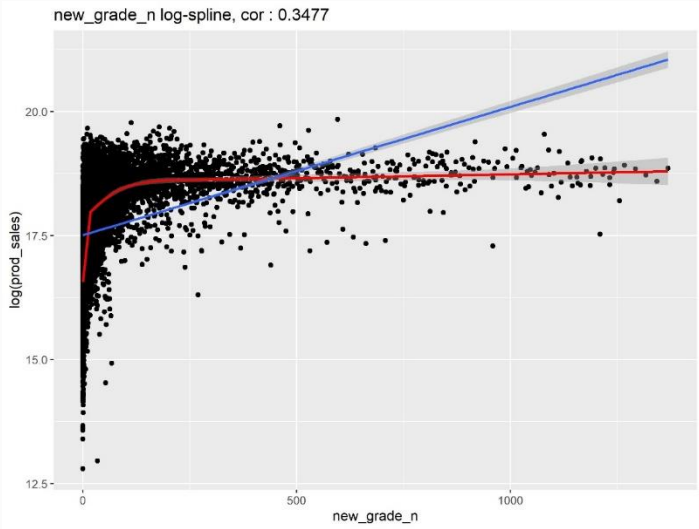
파생 변수

별점



빨강 : spline 적합, 파랑 : 회귀 적합

New투표수



빨강 : spline 적합, 파랑 : 회귀 적합

- 별점은 해당 상품의 인기를 대표할 수 있을 것으로 보이고 상관계수도 0.4061로 조금 강한 양의 상관관계를 보인다. 하지만 투표수가 낮은 경우에는 해당 별점의 신뢰가 낮아질 것으로 보인다.
- New투표수는 spline 빨간색 함수를 보면 알 수 있듯이 일정투표수까지는 취급액과 강한 양의 상관관계를 가지지만 일정 투표수를 넘어서게되면 취급액과 큰 상관이 없음을 알 수 있다.

파생 변수

브랜드 제품 노출수

브랜드명
방송일시
<외부 변수>

→ 브랜드 제품 노출수 <파생 변수>

브랜드 방송 노출수

브랜드명
방송일시
<외부,기존 변수>

→ 브랜드 방송 노출수 <파생 변수>

방송일시	브랜드명	상품명	브랜드 제품 노출수	브랜드 방송 노출수
2019-06-10 01:20	[메이듀]	메이듀 남성 린넨 블렌디드 슬립온	0	0
2019-06-10 01:20	[메이듀]	메이듀 여성 린넨 블렌디드 슬립온	0	0
2019-07-06 06:00	[메이듀]	메이듀 코튼 티블라우스 5종	2	1
2019-09-12 14:00	[메이듀]	메이듀 골드링 펌프스 + 베이직 펌프	3	2
2019-10-01 13:00	[메이듀]	메이듀 골드링 펌프스 + 베이직 펌프	4	3

같은 방송

개수

-1

이 전에 판매된 동일한 브랜드의 제품수와 방송수를 계산

파생 변수

브랜드 판매액 평균, 누적합

브랜드명
취급액
<외부,기존 변수>

→ 브랜드 판매액 평균 <파생 변수>

→ 브랜드 판매액 누적합 <파생 변수>

브랜드 판매수 평균, 누적합

브랜드명
판매수
<외부,기존 변수>

→ 브랜드 판매수 평균 <파생 변수>

→ 브랜드 판매수 누적합 <파생 변수>

방송일시	브랜드명	상품명	판매액
2019-06-10 01:20	[메이듀]	메이듀 남성 린넨 블렌디드 슬립온	6,612,000
2019-06-10 01:20	[메이듀]	메이듀 여성 린넨 블렌디드 슬립온	14,085,000
2019-07-06 06:00	[메이듀]	메이듀 코튼 티블라우스 5종	42,058,000
2019-09-12 14:00	[메이듀]	메이듀 골드링 펌프스 + 베이직 펌프	81,873,000
2019-10-01 13:00	[메이듀]	메이듀 골드링 펌프스 + 베이직 펌프	83,886,000

같은 방송

브랜드 판매액 평균	판매액 누적합
NA	NA
NA	NA
10,348,500	20,697,000
20,918,333	62,755,000
평균 36,157,000	합 144,628,000

파생 변수

브랜드 판매액 평균, 누적합

브랜드명
취급액
<외부,기존 변수>

→ 브랜드 판매액 평균 <파생 변수>

→ 브랜드 판매액 누적합 <파생 변수>

브랜드 판매수 평균, 누적합

브랜드명
판매수
<외부,기존 변수>

→ 브랜드 판매수 평균 <파생 변수>

→ 브랜드 판매수 누적합 <파생 변수>

	방송일시	브랜드명	상품명	판매수
같은 방송	2019-06-10 01:20	[메이듀]	메이듀 남성 린넨 블렌디드 슬립온	221.8792
같은 방송	2019-06-10 01:20	[메이듀]	메이듀 여성 린넨 블렌디드 슬립온	472.6510
	2019-07-06 06:00	[메이듀]	메이듀 코튼 티블라우스 5종	1054.0852
	2019-09-12 14:00	[메이듀]	메이듀 골드링 펌프스 + 베이직 펌프	2051.9550
	2019-10-01 13:00	[메이듀]	메이듀 골드링 펌프스 + 베이직 펌프	2102.4060

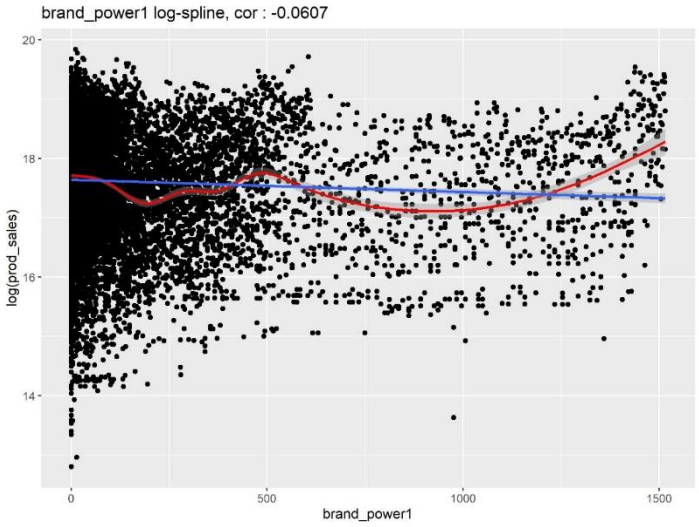
브랜드 판매수 평균	판매수 누적합
NA	NA
NA	NA
347.2651	694.5302
582.8718	1748.6150
950.1426	3800.5700

평균 → 582.8718

합 → 3800.5700

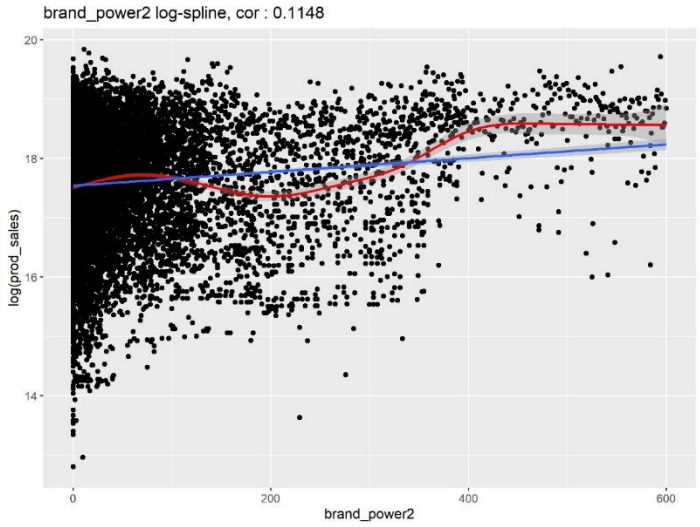
파생 변수

브랜드 제품 노출수



빨강 : spline 적합, 파랑 : 회귀 적합

브랜드 방송 노출수

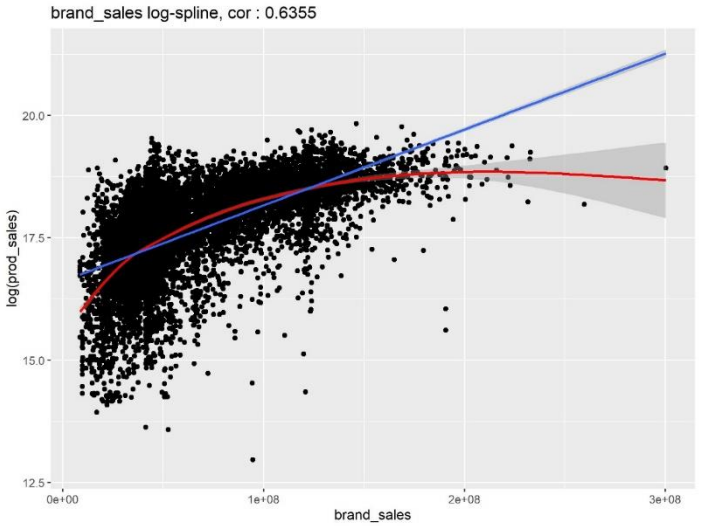


빨강 : spline 적합, 파랑 : 회귀 적합

- 브랜드 제품 노출수는 상관관계수가 -0.0607로 거의 0에 가깝다. 또한 해석도 적절하지 않아 예측 변수로 사용하지 않았다.
- 브랜드 방송 노출수는 상관계수가 0.1148로 아주 약한 양의 상관관계를 가진다. 이는 홈쇼핑의 주 고객층이 40~50대 여성이고 이들의 제품 구매시 중요 결정 요인이 브랜드의 신뢰도인 것이 반영된 것으로 보인다.

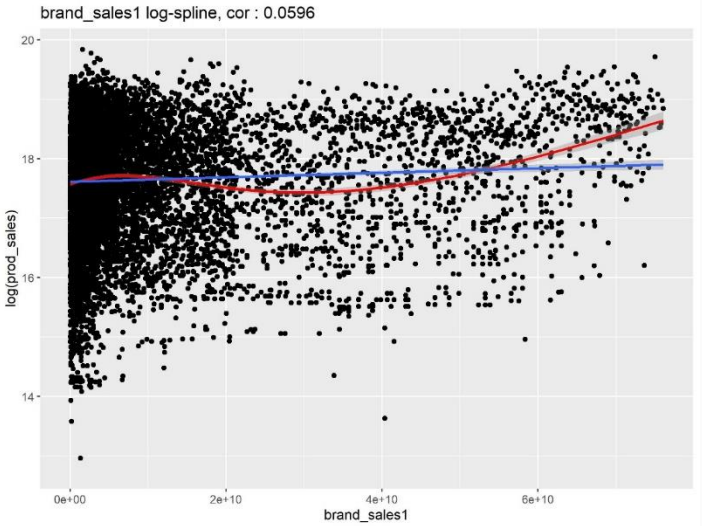
파생 변수

브랜드 판매액 평균



빨강 : spline 적합, 파랑 : 회귀 적합

브랜드 판매액 누적합

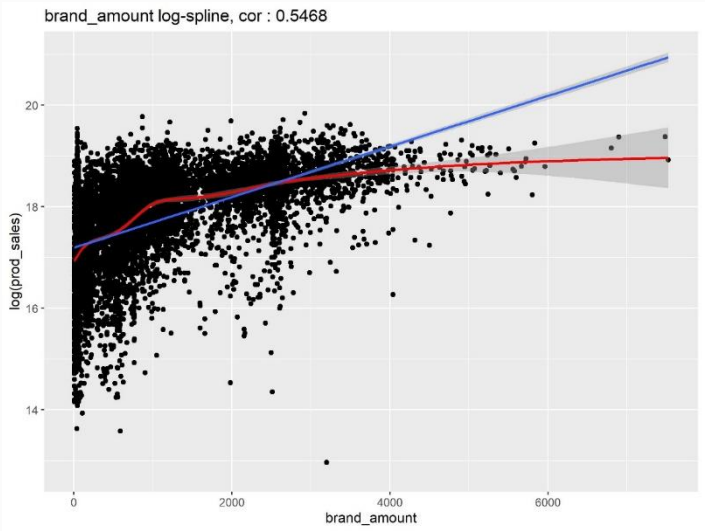


빨강 : spline 적합, 파랑 : 회귀 적합

- 브랜드 판매액 평균은 취급액과 상관계수가 0.6355로 강한 양의 상관관계를 가지고 spline 함수에서 알 수 있듯이 어떤 지점을 넘어가면 취급액과는 크게 상관이 없지만 그 전까지는 취급액과 강한 상관이 있음을 알 수 있다.
- 브랜드 판매액 누적은 상관계수가 0.0596으로 거의 0에 가까우므로 해당 변수는 예측 변수로 사용하지 않았다.

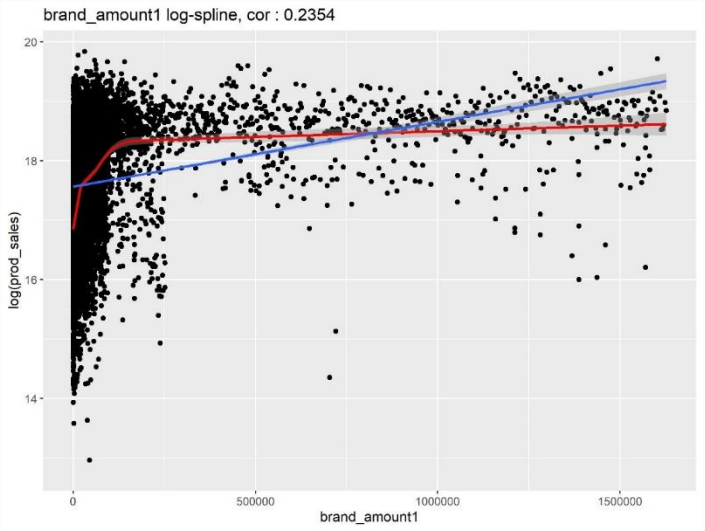
파생 변수

브랜드 판매수 평균



빨강 : spline 적합, 파랑 : 회귀 적합

브랜드 판매수 누적합



빨강 : spline 적합, 파랑 : 회귀 적합

- 브랜드 판매수 평균과 브랜드 판매수 누적합 모두 취급액과 양의 상관관계를 가지고 두 변수 모두 어느 지점까지는 취급액과 큰 상관을 보이지만 어느 지점을 넘어서고는 취급액과 큰 상관을 보이지 않는 것으로 보인다. 즉, 어느 지점보다 작은 변수들의 값은 취급액을 예측하는데 큰 영향을 끼칠 것으로 판단하여 예측 변수로 사용하였다.

파생 변수

18K_단어변수

상품명 <기존 변수>

상품명	18K_단어변수
상품명에 '18K' 포함 X	0
상품명에 '18K' 포함 O	1

18K_단어변수 <파생 변수>

TV_단어변수

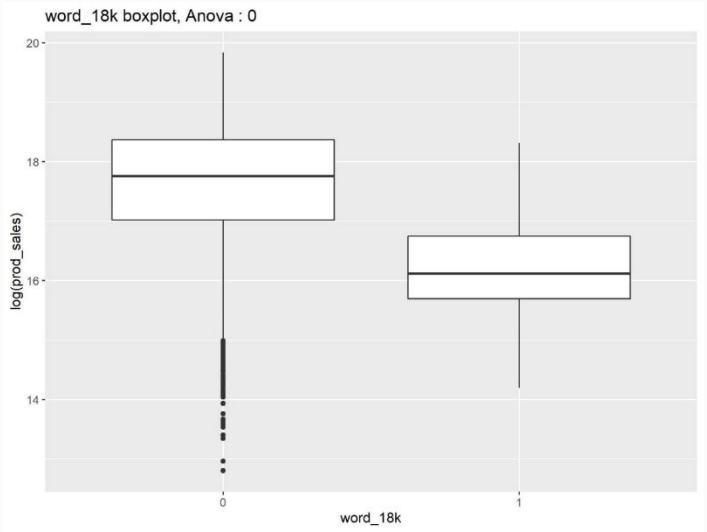
상품명 <기존 변수>

상품명	TV_단어변수
상품명에 'TV' 포함 X	0
상품명에 'TV' 포함 O	1

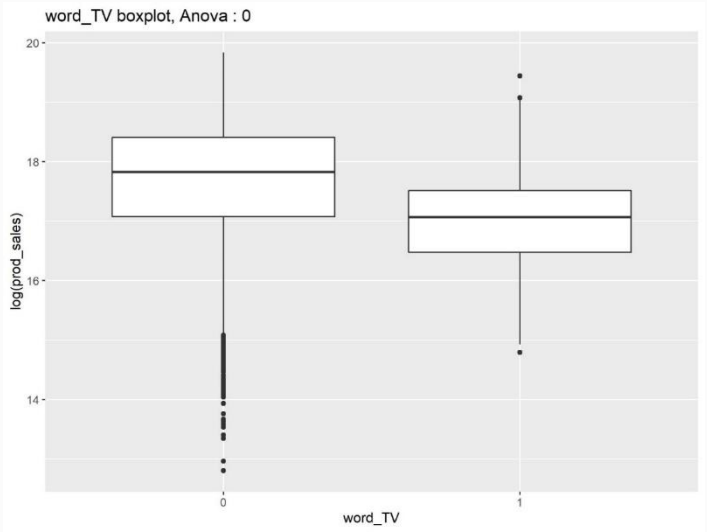
TV_단어변수 <파생 변수>

파생 변수

18K_단어변수



TV_단어변수



- 상품명에 '18k' 단어가 들어간 상품의 개수는 77개로 12807개 중 굉장히 적어 unbalanced한 변수이지만 '18k'단어가 들어가지 않은 경우의 취급액이 들어간 경우보다 대략 3.83배나 더 크므로 해당 변수를 예측변수로 사용하였다.
- 상품명에 'TV'단어가 들어가지 않은 경우의 취급액이 들어간 경우보다 대략 2.29배나 더 크므로 해당 변수를 예측변수로 사용하였다.

파생 변수

취급액_t1, 취급액_t2

상품코드, 상품명
취급액
<기존 변수>

→ 취급액_t1 <파생 변수>

→ 취급액_t2 <파생 변수>

취급액_t평균

상품코드, 상품명
취급액
<기존 변수>

→ 취급액_t평균 <파생 변수>

방송일시	상품코드	상품명	취급액
2019-05-29	200624	코치 노리타 리스틀릿	41,942,000
2019-06-02	200624	코치 노리타 리스틀릿	37,662,000
2019-06-03	200624	코치 노리타 리스틀릿	27,190,000
2019-06-07	200624	코치 노리타 리스틀릿	37,226,000
2019-06-22	200624	코치 노리타 리스틀릿	31,869,000

취급액_t1	취급액_t2
NA	NA
41,942,000	NA
37,662,000	41,942,000
27,190,000	37,662,000
평균 37,662,000	합 27,190,000

파생 변수

취급액_t1, 취급액_t2

상품코드, 상품명
취급액
<기존 변수>

→ 취급액_t1 <파생 변수>

→ 취급액_t2 <파생 변수>

취급액_t평균

상품코드, 상품명
취급액
<기존 변수>

→ 취급액_t평균 <파생 변수>

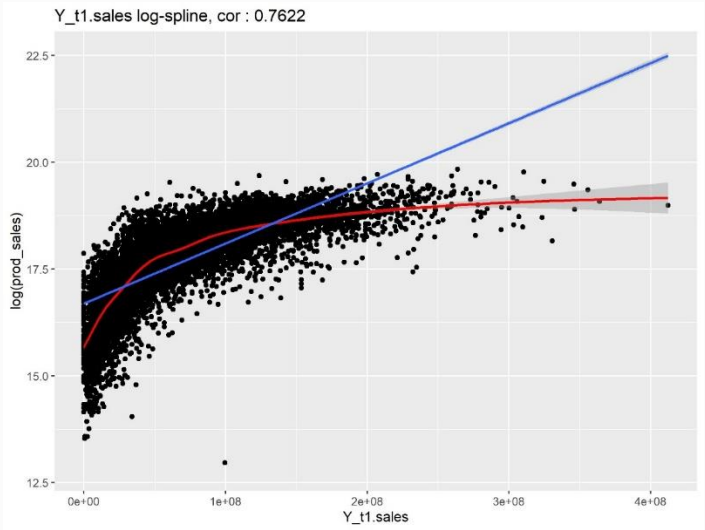
방송일시	상품코드	상품명	취급액
2019-05-29	200624	코치 노리타 리스틀릿	41,942,000
2019-06-02	200624	코치 노리타 리스틀릿	37,662,000
2019-06-03	200624	코치 노리타 리스틀릿	27,190,000
2019-06-07	200624	코치 노리타 리스틀릿	37,226,000
2019-06-22	200624	코치 노리타 리스틀릿	31,869,000

평균

취급액_t평균
NA
41,942,000
39,802,000
35,598,000
36,005,000

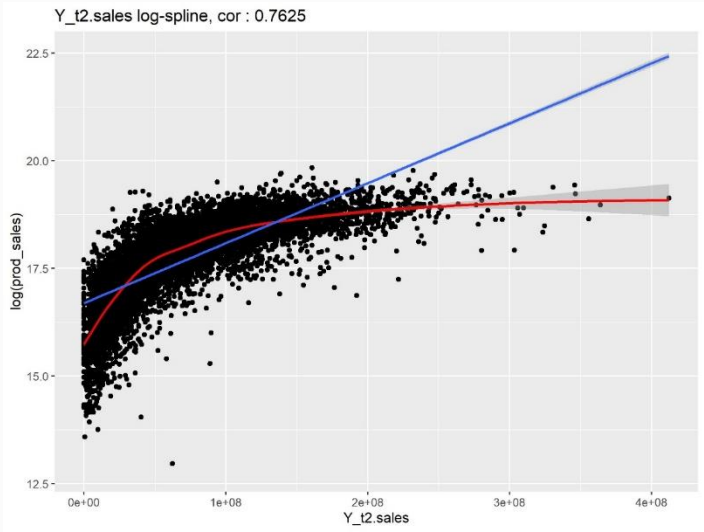
파생 변수

취급액_t1



빨강 : spline 적합, 파랑 : 회귀 적합

취급액_t2

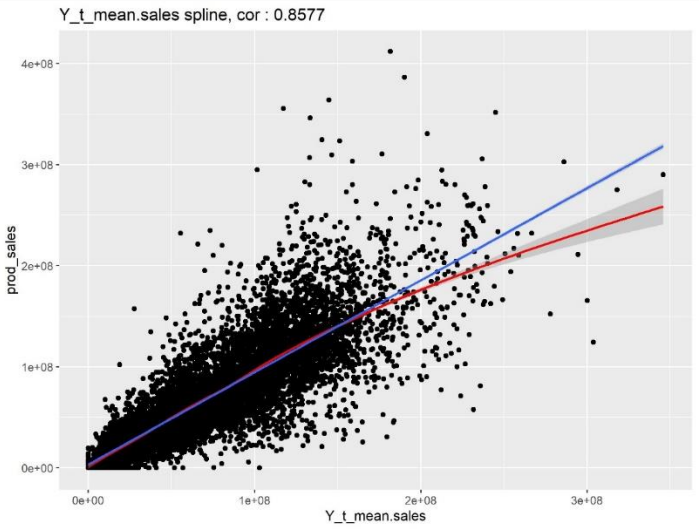


빨강 : spline 적합, 파랑 : 회귀 적합

- 취급액_t1의 경우 상관계수가 0.7622로 강한 양의 상관관계를 가지고 이는 가장 최근에 판매된 취급액이 해당 상품의 취급액과 비슷할 것으로 예상되기 때문인 것으로 보인다.
- 취급액_t2의 경우도 상관계수가 0.7625로 강한 양의 상관관계를 가지고 이는 취급액_t1과 같은 이유이지만 취급액_t1과 취급액_t2 변수를 함께 예측 변수로 사용함으로써 두 변수 간의 차이를 반영할 수 있을 것으로 보인다.

파생 변수

취급액_t평균



빨강 : spline 적합, 파랑 : 회귀 적합

- 취급액_t평균 변수와 취급액의 상관계수는 0.8577로 사실상 가장 높은 상관계수를 보이는 변수이다. 이전 판매된 동일한 상품들의 취급액을 평균한 값이므로 당연한 것으로 볼 수 있다. 하지만 과거 실적이 존재하는 관측치는 test set에서 제한되어 일부 관측치에만 사용할 수 있다는 한계가 있다.

파생 변수

날씨변수

날씨변수 <외부 변수> ---------> PCA1~9 <파생 변수>

변수명(7)	x	지역(6)
기온, 강수량, 풍속, 습도 중기압, 일조량, 일사량		서울, 광주, 대구 대전, 부산, 인천

주성분분석(PCA)

PCA1~9

소비자 심리지수

소비자 심리지수 <외부 변수>

일별시청률.(채널)

일별시청률 <외부 변수> ---------> 일별시청률.채널 <파생 변수>

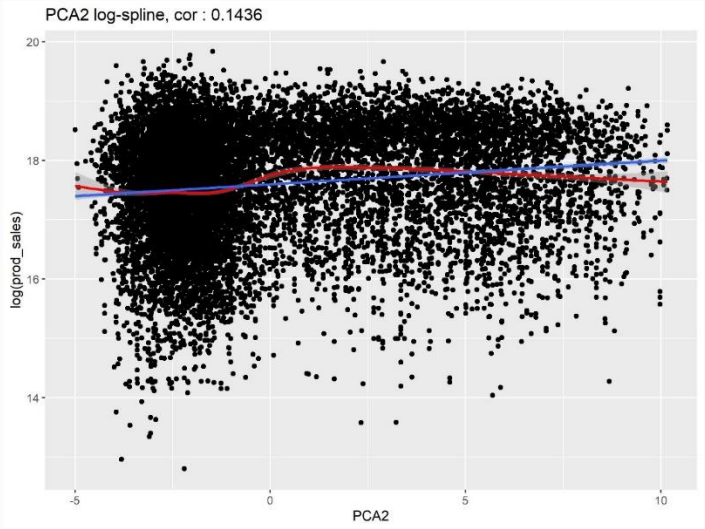
날짜	채널명	시청률
20181231	KBS1	18.4
20181231	KBS2	14.1
20181231	KBS1	11.6
...	...	...
20181231	tvN	1.475

채널별 합

날짜	채널명	시청률
20181231	KBS1	98.2
20181231	KBS2	48.4

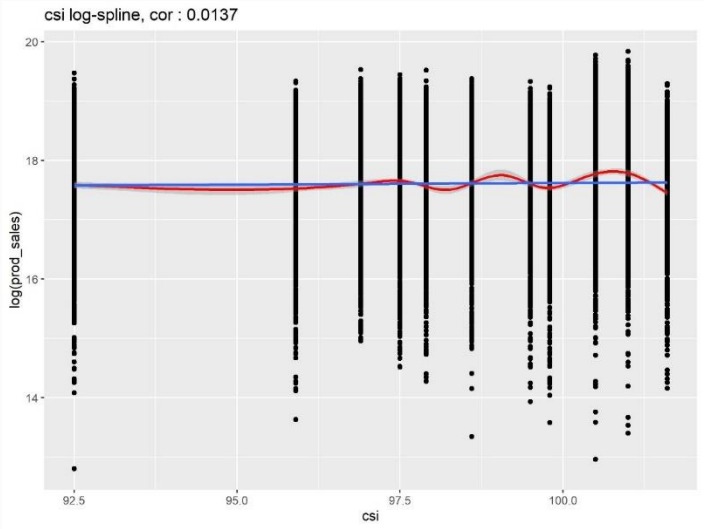
파생 변수

날씨변수



빨강 : spline 적합, 파랑 : 회귀 적합

소비자 심리지수

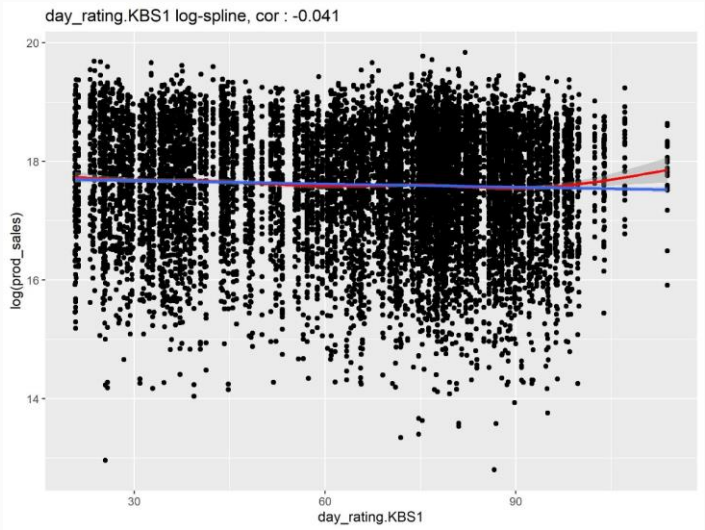


빨강 : spline 적합, 파랑 : 회귀 적합

- 날씨변수는 지역과 변수가 너무 많아서 주성분분석을 실행하였다. 누적 분산 비율이 80%를 넘는 시점인 9개까지만 변수로 활용하였고 이 중 PCA2변수를 제외하고는 취급액과의 상관계수가 너무 낮아 나머지 PCA1, 3~9 변수는 예측 변수에서 제외하고 PCA2 변수만 예측 변수로 사용하였다.
- 소비자 심리지수와 취급액과의 상관계수는 0.0137로 거의 0에 가까우므로 예측 변수에서 제외하였다.

파생 변수

일별 시청률.KBS1



빨강 : spline 적합, 파랑 : 회귀 적합

- 일별 시청률 변수들은 모두 취급액과의 상관계수가 거의 0에 가까우므로 예측 변수에서 제외하였다.

파생 변수

1. 상품특성관련변수
2. 방송일시관련변수
3. 방송시간관련변수
4. 인기관련변수
5. 브랜드관련변수
6. 상품명관련변수
7. 과거실적관련변수
8. 외부변수
9. 분당실적데이터 관련변수

방송시간 비율

노출(분) 방송시간 <기존 변수> → 방송시간 비율 <파생 변수>

방송시간 누적 비율

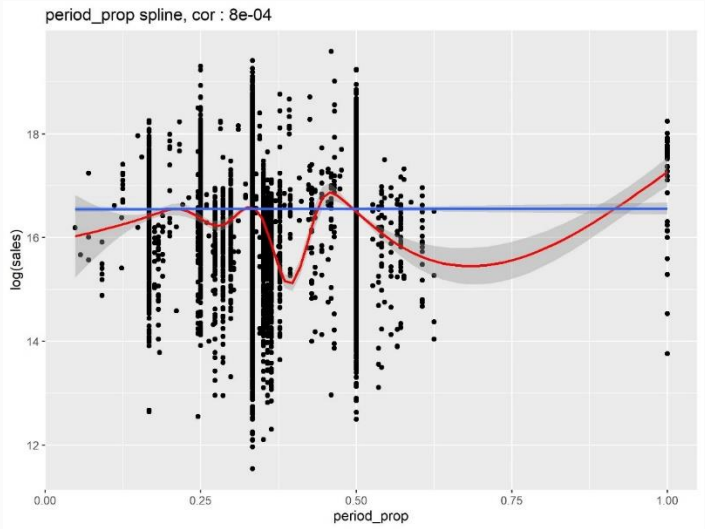
노출(분) 방송시간 <기존 변수> → 방송시간 누적 비율 <파생 변수>

같은 제품 및 같은 방송

방송일시	방송시간	상품명	노출(분)	브랜드 판매수 평균	판매수 누적합
2019-01-01 06:00	60	테이트 남성 셀린니트3종	20	20/60=0.333	0.333
2019-01-01 06:20	60	테이트 남성 셀린니트3종	20	20/60=0.333	0.667
2019-01-01 06:40	60	테이트 남성 셀린니트3종	20	20/60=0.333	1
2019-01-01 11:00	40	크로커다일 The 편안한	20	20/40 = 0.5	0.5
2019-01-01 11:20	40	크로커다일 The 편안한	20	20/40 = 0.5	1
2019-04-17 08:30	30	일시불[안드레아바나]리	30	30/30 = 1	1

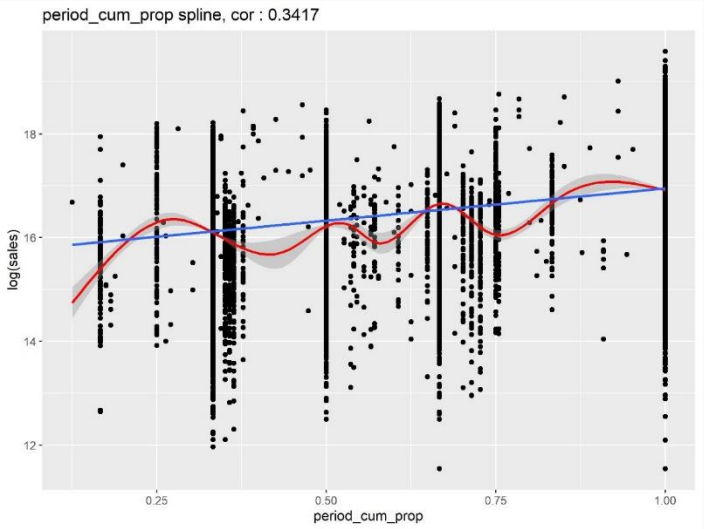
파생 변수

방송시간 비율



빨강 : spline 적합, 파랑 : 회귀 적합

방송시간 누적 비율



빨강 : spline 적합, 파랑 : 회귀 적합

- 방송시간 비율의 상관계수는 0인데 이는 대부분의 상품 방송시간 비율이 일정하게 나뉘어져 있기 때문이다.
- 방송시간 누적 비율은 취급액과의 상관계수가 0.3417로 양의 상관관계를 가진다. 홈쇼핑 방송 특성상 대부분의 매출이 방송 후반에 이루어지기 때문인 것으로 보인다. 위 두 변수를 함께 예측 변수에 사용하여야 정확한 분당 실적 데이터를 만들 수 있으므로 두 변수 모두 예측 변수로 사용하였다.

사용 변수 정리

분류 방식	변수명
1. 상품 특성 관련 변수	판매단가, 상품군, 지불방식
2. 날짜 관련 변수	방송시작시각, 요일, 공휴일
3. 시간 관련 변수	방송시간, 방송 판매 상품수, 예상 노출시간
4. 인기 관련 변수	별점, New투표수
5. 브랜드 관련 변수	브랜드 제품 노출수, 브랜드 방송 노출수, 브랜드 판매액 평균, 브랜드 판매액 누적합, 브랜드 판매수 평균, 브랜드 판매수 누적합
6. 상품명 관련 변수	18K_단어변수, TV_단어변수
7. 과거 실적 관련 변수	취급액_t1, 취급액_t2, 취급액_t평균
8. 외부 변수	날씨변수(PCA1~9),PCA2, 소비자 심리지수, 일별 시청률.(채널)
9. 분당 실적 데이터 관련 변수	방송시간 비율, 방송시간 누적비율

노란색 변수의 경우 예측 변수에서 제외 / 검은색 변수만 예측 변수에 사용

04 데이터 분석 및 모델링

- 변수별 결측치
- 모델 구축
- 모델 비교
- 요인 분석

변수별 결측치

데이터	결측없는 변수	A	B	C	D
1	[전체데이터모델] Train : 36630 (98.01%) Test : 2716 (100%)				
2					
3					
4					NA
5				NA	NA
6			NA	NA	NA
7		NA	NA	NA	NA

데이터	결측없는 변수	A	B	C	D
1	[1번데이터] Train : 34796 (93.11%) Test : 1835 (67.56%)				
2					
3					
4					NA
5				NA	NA
6			NA	NA	NA
7		NA	NA	NA	NA

A: 별점, New투표수, N번째 판매
B: 브랜드 판매액 평균,
브랜드 판매수 평균[누적합]
C: 취급액_t1,
취급액_t평균
D: 취급액_t2

데이터	결측없는 변수	A	B	C	D
1	[2번데이터] Train : 31550 (84.42%) Test : 388 (14.29%)				
2					
3					
4					NA
5				NA	NA
6			NA	NA	NA
7		NA	NA	NA	NA

데이터	결측없는 변수	A	B	C	D
1	[3번데이터] Train : 27904 (74.67%) Test :376 (13.84%)				
2					
3					
4					NA
5				NA	NA
6			NA	NA	NA
7		NA	NA	NA	NA

[예측순서]
전체 데이터 모델로 예측
1번 데이터 모델로 예측 갱신
2번 데이터 모델로 예측 갱신
3번 데이터 모델로 예측 갱신

분당 실적 데이터 - Train set : 37372 / Test set : 2716 (상품군 무형 제거)

변수별 결측치

하늘색 변수 결측치 존재

변수명	1번 데이터	2번 데이터	3번 데이터	4번 데이터
판매단가, 상품군, 지불방식	Train : 36630 (98.01%) Test : 2716 (100%)	Train : 34796 (93.11%) Test : 1835 (67.56%)	Train : 31550 (84.42%) Test : 388 (14.29%)	Train : 27904 (74.67%) Test :376 (13.84%)
방송시작 시각, 요일, 공휴일				
방송 판매 상품수, 예상노출시간				
별점,New투표수,N번째 판매(A)				
18K, TV_단어변수				
방송시간 비율 방송시간 누적 비율				
브랜드 방송 노출수				
브랜드 판매액 평균, 브랜드 판매수 평균[누적합](B)				
취급액_t1, 취급액_t평균(C)				
취급액_t2(D)				

분당 실적 데이터 - Train set : 37372 / Test set : 2716 (상품군 무형 제거)

변수별 결측치

- 각 변수별 결측치가 달라서 변수가 추가될수록 데이터의 수는 줄어들지만 변수의 수는 증가하기 때문에 더 좋은 모델을 만들 것으로 보임
- 앞서 보았던 전체 데이터 및 1번 ~ 3번 데이터 중 수행력은 3번 > 2번 > 1번 > 전체 순으로 보았고 이를 이용하여 예측에 사용
- 전체 데이터로 예측을 한 후 1번, 2번 ,3번 순서대로 해당 관측치에 예측값을 갱신하는 방법으로 예측

- 종속 변수인 취급액은 log 변환을 한 상태로 모델링 후 예측값에 모두 exp($\hat{Y}$)으로 변환하여 예측값 생성

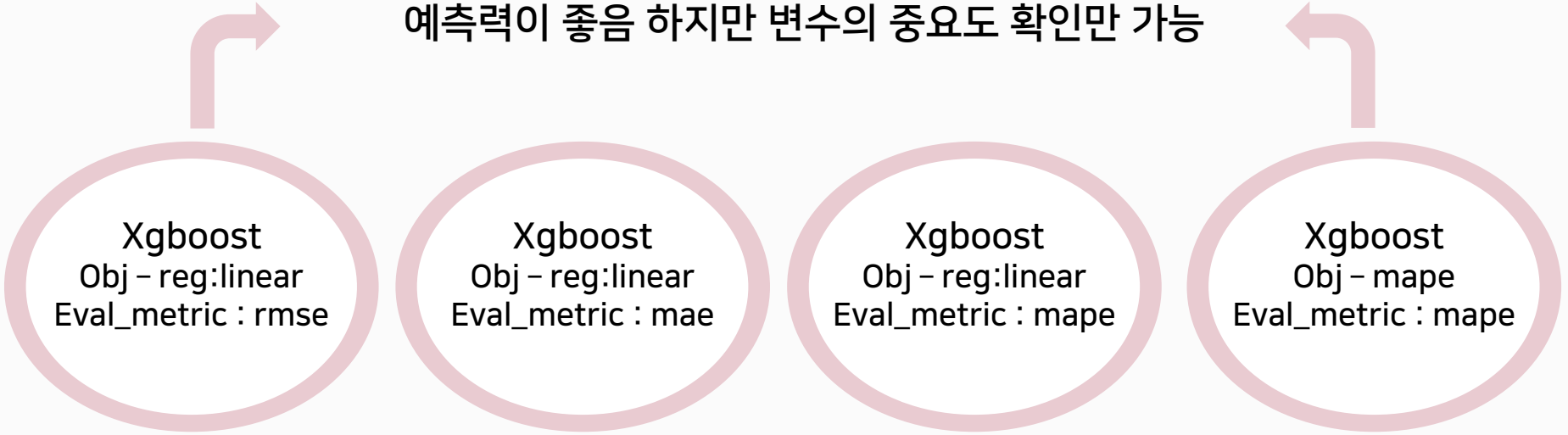
- Train set : 2019년 1월 ~ 2019년 9월 (22126개, 77.0%)
  Valid set : 2019년 10월 ~ 2019년 12월 (8414개, 23.0%)
- 3 FOLD – Cross Validation 을 통한 Hyper Parameter 최적화

모델 구축

해석 가능 & 예측력이 부족

- 해석이 가능
- 계수에 대하여 "선형" 관계 가능

예측력이 좋음 하지만 변수의 중요도 확인만 가능



- 랜덤 포레스트 + boosting + greedy algorithm 으로 과적합 방지
- 예측력 ↑, 사용하기 편리
- 결과 해석이 어려움

- 목적 함수(Objective function) : 해당 함수를 통하여 gradient를 계산하고 해당 함수를 줄이는 방향으로 데이터가 적합된다.
- 평가 기준(Evaluation metric) : Train set으로 적합해 나갈 때 Valid set의 평가 기준을 보고 적합할 TREE의 개수를 정한다.

모델 구축

1. 회귀분석

- Train set 과 Valid set 을 합친 데이터로 모델을 적합하였고 예측 값은 Test set에 대하여 계산
- 각 데이터마다 회귀분석을 실행한 결과의 수정된 결정계수를 보았을 때 3번 > 2번 > 1번 > 전체 순으로의 설명력이 좋음
- 제안사항에서 중요하게 다뤄질 예상 노출시간 변수는 모든 모델에서 유의하였고 계수는 모두 양수

항목	전체 데이터 모델	1번 데이터 모델	2번 데이터 모델	3번 데이터 모델
수정된 결정계수	0.5996	0.6016	0.6851	0.6902
예상 노출시간 계수	양수 (유의O)	양수 (유의O)	양수 (유의O)	양수 (유의O)

모델 구축

2. XGBOOST(목적함수: REGRESSION-LINEAR / 평가기준: RMSE)

- 목적 함수가 REGRESSION일 경우는 RMSE 평가기준을 최소화 시키는 방향으로 gradient descent를 계산 (해당 목적 함수의 Gradient와 Hessian이 필요)
- 평가기준은 Valid set의 평가 기준을 확인하여 얼마나 많은 Tree를 적합할지 결정하는 용도로 사용

$$RMSE = \frac{1}{n}\sum_{i=1}^{n}(\hat{y_i} - y_i)^2$$

모델 구축

3. XGBOOST(목적함수: REGRESSION-LINEAR / 평가기준: MAE)

- 목적함수가 REGRESSION일 경우는 RMSE 평가기준을 최소화 시키는 방향으로 gradient descent를 계산 (해당 목적 함수의 Gradient와 Hessian이 필요, 이전 모델과 목적함수는 동일하지만 평가기준은 다름)
- 평가기준은 Valid set의 평가기준을 확인하여 얼마나 많은 Tree를 적합할지 결정하는 용도로 사용
- 해당 모델의 목적함수는 RMSE 평가 기준을 최소화 시키는 방향으로 계산하지만 Valid set의 MAE 값을 확인하여 Tree의 개수를 정하는 방식

$$MAE = \frac{1}{n}\sum_{i=1}^{n}\left|\hat{y}_i - y_i\right|$$

모델 구축

4. XGBOOST(목적함수: REGRESSION-LINEAR / 평가기준: MAPE)

- 목적함수가 REGRESSION일 경우는 RMSE 평가기준을 최소화 시키는 방향으로 gradient descent를 계산 (해당 목적 함수의 Gradient와 Hessian이 필요, 이전 모델과 목적함수는 동일하지만 평가기준은 다름)
- 평가기준은 Valid set의 평가기준을 확인하여 얼마나 많은 Tree를 적합할지 결정하는 용도로 사용
- 해당 모델의 목적함수는 RMSE 평가 기준을 최소화 시키는 방향으로 계산하지만 Valid set의 MAPE 값을 확인하여 Tree의 개수를 정하는 방식

$$MAPE = \frac{1}{n}\sum_{i=1}^{n}\left|\frac{\hat{y}_i - y_i}{y_i}\right|$$

모델 구축

5. XGBOOST(목적함수: MAPE / 평가기준: MAPE)

- 목적함수가 MAPE인 경우는 MAPE 평가기준을 최소화 시키는 방향으로 gradient descent를 계산 (해당 목적 함수의 Gradient와 Hessian이 필요)
- MAPE는 MAE에 실제 값의 역수의 절댓값을 곱한 값과 같으므로 MAE 목적 함수 Gradient와 Hessian에 해당 값을 곱하여 계산
- 목적함수가 MAE인 경우 해당 목적 함수의 Gradient와 Hessian을 계산하여야 하는데 정확한 미분에 대한 공식이 존재하지 않으므로 MAE loss function과 유사한 fair loss function 사용
- 평가기준은 Valid set의 평가기준을 확인하여 얼마나 많은 Tree를 적합할지 결정하는 용도로 사용

$$MAPE = \frac{1}{n}\sum_{i=1}^{n}\left|\frac{\hat{y}_i - y_i}{y_i}\right| = \left|\frac{1}{y_i}\right| MAE$$

$$Fair\ loss = c^2\left(\frac{|x|}{c} - \ln\left(\frac{|x|}{c}+1\right)\right)$$

해당 모델에서는 c값을 2로 지정

모델 구축

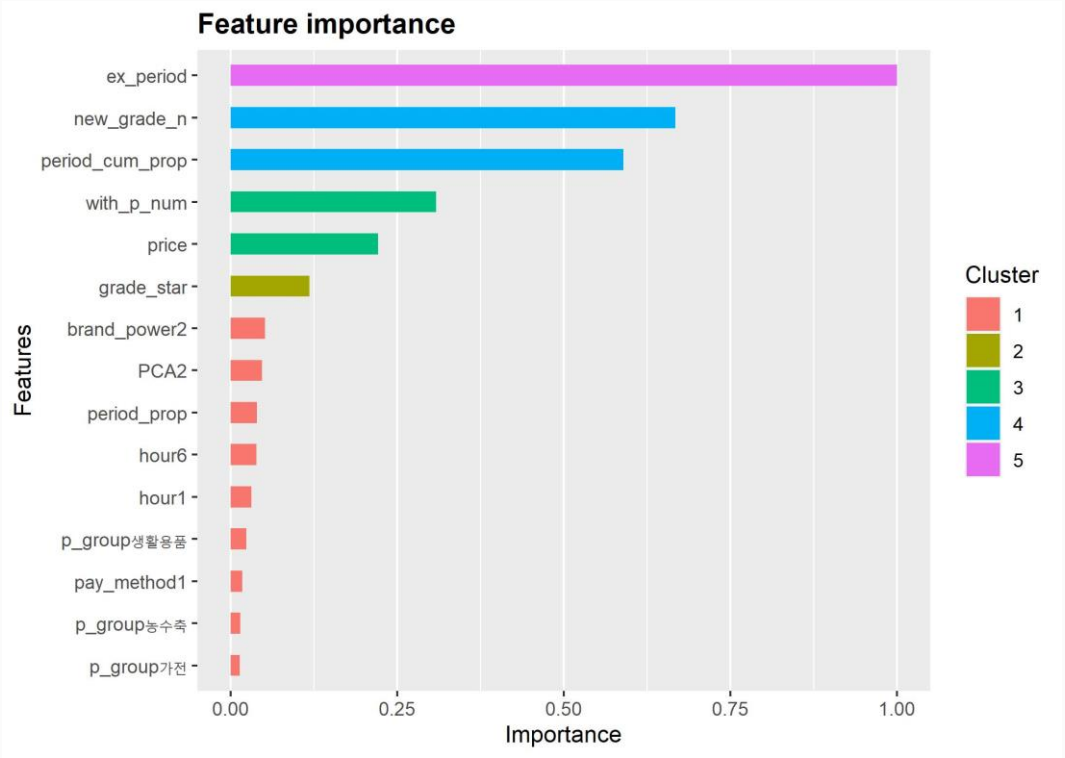
HYPER PARAMETER	XGBOOST(REG, RMSE)	XGBOOST(REG, MAE)	XGBOOST(REG, MAPE)	XGBOOST(MAPE, MAPE)
BOOSTER	Gbtree	Gbtree	최적화를 하지 못하고 Obj – reg:linear Eval_metric : rmse Hyper parameter 그대로 사용	최적화를 하지 못하고 Obj – reg:linear Eval_metric : mae Hyper parameter 그대로 사용
MAX_DEPTH	7	8		
GAMMA	0.992	1.36		
MIN_CHILD_WEIGHT	7.35	7.51		
SUBSAMPLE	0.98	0.972		
COLSAMPLE_BYTREE	0.538	0.569		
ETA	0.048	0.0261		
EARLY_STOPPING_ROUNDS	11	11		
NROUND	10000	10000		

모델 비교

모델명	RMSE	MAE	MAPE
회귀분석	31,980,344	11,916,427	0.5138
XGBOOST(REG, RMSE)	15,753,988	8,599,401	0.3724
XGBOOST(REG, MAE)	15,553,359	8,518,317	0.3730
XGBOOST(REG, MAPE)	15,596,512	8,510,912	0.3701
XGBOOST(MAPE, MAPE)	16,062,439	8,825,321	0.3873
★ XGBOOST 평균 앙상블	15,641,582	8,514,074	0.3701

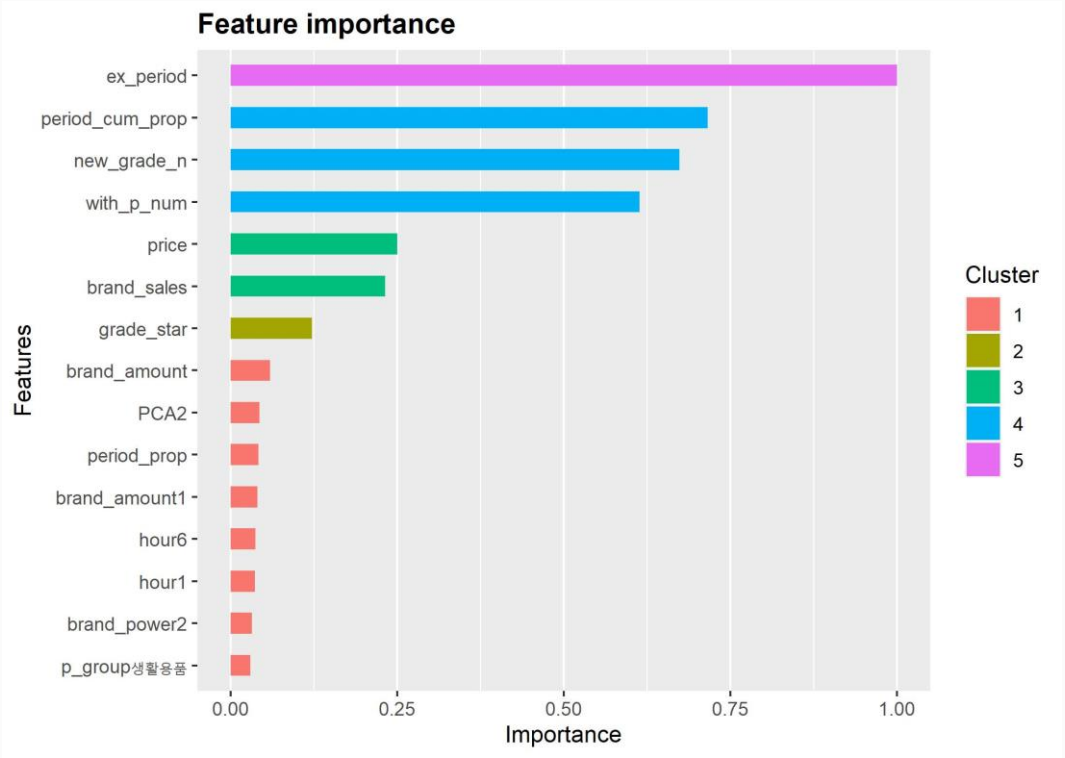
요인 분석

전체 데이터 모델 IMPORTANCE PLOT



상품 노출시간 변수가 해당 XGBOOST 모델에 가장 많은 기여를 하였고
New투표수 변수, 누적 방송시간 비율 변수가 뒤를 이음

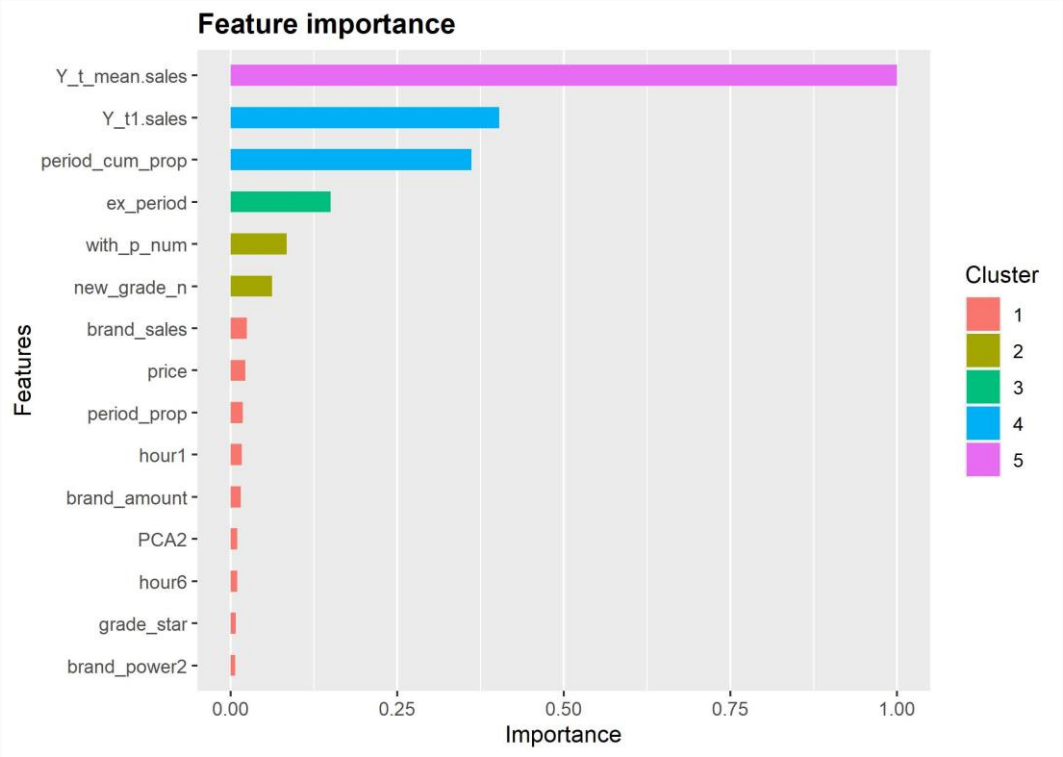
1번 데이터 모델 IMPORTANCE PLOT



상품 노출시간 변수가 해당 XGBOOST 모델에 가장 많은 기여를 하였고
New투표수 변수, 누적 방송시간 비율 변수가 뒤를 이음

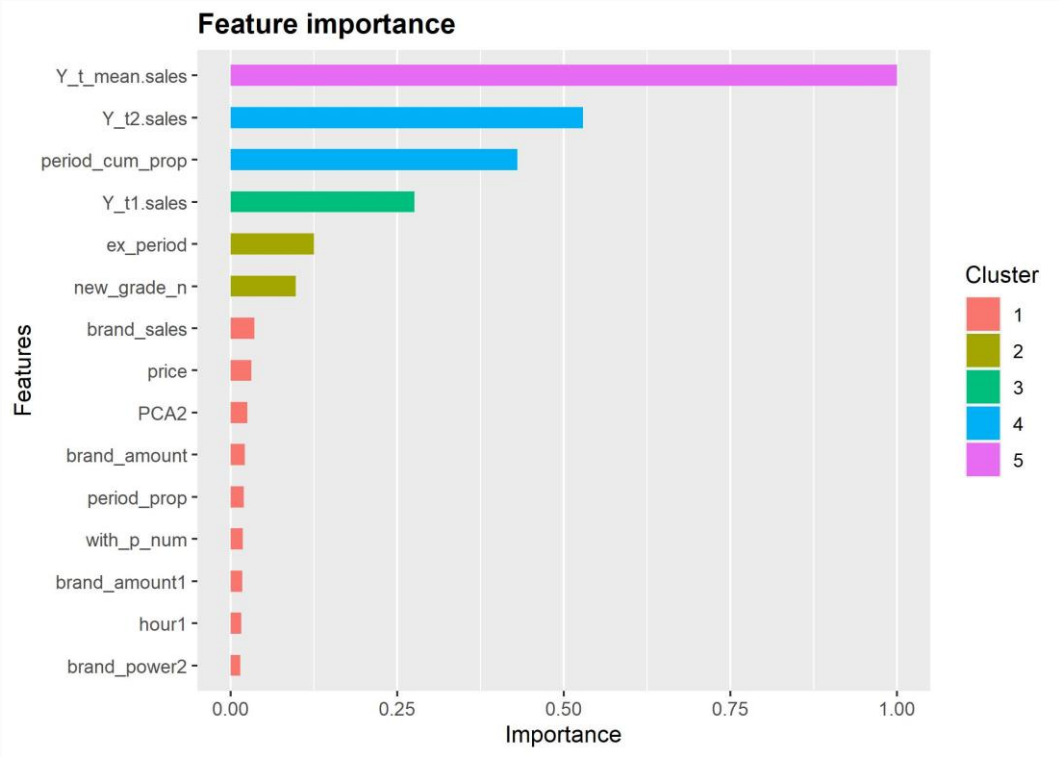
요인 분석

2번 데이터 모델 Importance plot



과거 실적 관련 변수가 가장 많이 기여했고 그 뒤로 누적 방송시간 비율 변수 및 예상 노출시간 변수가 많이 기여함을 알 수 있음

3번 데이터 모델 Importance plot



2번 데이터 모델과 동일하게 과거 실적 관련 변수가 가장 많이 기여했고 그 뒤로 누적 방송시간 비율 변수 및 예상 노출시간 변수가 좋음을 알 수 있음

05 제안 및 결론

- 최적 방송 편성표
- 최적 상품 노출시간
- 결론
- 아쉬운 점

최적 방송 편성표

1. 상품별 실적 데이터를 이용

2. 앞서 만들었던 변수들을 사용하여 XGBOOST 4가지 모델의 평균 앙상블을 이용한 취급액 예측 모델 생성

3. 각 주간 상품들에 대하여 일주일 동안의 모든 방송일시에 대한 취급액을 예측할 데이터 생성

4. 각 주간 상품들에 대한 모든 방송일시에 대한 취급액 예측

5. 헝가리안 알고리즘을 이용하여 각 상품들의 최적 방송일시 할당

최적 방송 편성표

1. 상품별 실적 데이터

방송일시	상품명	판매단가	취급액	시작 시각	요일	방송시간	방송판매 상품수	예상 노출 시간
2019-01-01 6:00	테이트 남성 셀린니트3종	39,900	12,033,000	6	화요일	60	2	30
2019-01-01 6:00	테이트 여성 셀린니트3종	39,900	20,663,000	6	화요일	60	2	30
2019-01-01 7:00	오모떼 레이스 파운데이션 브라	59,000	47,878,000	7	화요일	60	1	60
2019-01-01 8:00	CERINI by PAT 남성 소프트	59,900	99,736,000	8	화요일	60	1	60
2019-01-01 9:00	보코 리버시블 무스탕	79,000	90,973,000	9	화요일	60	1	60
2019-01-01 10:00	CERINI by PAT 남성 풀패키지	79,900	259,678,000	10	화요일	60	1	60

주목할 변수 : 시작 시각, 요일, 예상 노출시간

상품별 실적 데이터에 존재하지 않는 변수 : 방송시간 비율, 방송시간 누적 비율 (분당 실적 데이터에만 존재)

취급액 예측 모델을 생성할 때 추가되는 변수 : 방송시간 / 제거되는 변수 : 방송시간 비율, 방송시간 누적 비율

최적 방송 편성표

2. 상품별 실적 데이터 취급액 예측 모델

데이터	결측없는 변수	A	B	C	D
1	[전체데이터모델] Train : 12546 (98.0%)				
2					
3					
4					NA
5				NA	NA
6			NA	NA	NA
7		NA	NA	NA	NA

데이터	결측없는 변수	A	B	C	D
1	[1번데이터] Train : 11912 (93.0%)				
2					
3					
4					NA
5				NA	NA
6			NA	NA	NA
7		NA	NA	NA	NA

A: 별점, New투표수, N번째 판매
B: 브랜드 판매액 평균,
브랜드 판매수 평균[누적합]
C: 취급액_t1,
취급액_t평균
D: 취급액_t2

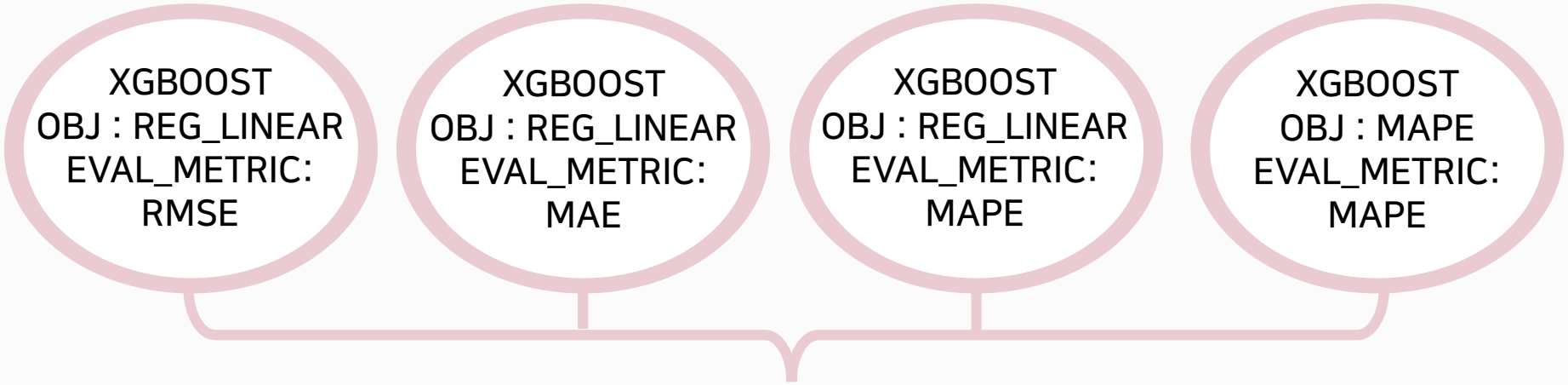
데이터	결측없는 변수	A	B	C	D
1	[2번데이터] Train : 10789 (84.2%)				
2					
3					
4					NA
5				NA	NA
6			NA	NA	NA
7		NA	NA	NA	NA

데이터	결측없는 변수	A	B	C	D
1	[3번데이터] Train : 9524 (74.4%)				
2					
3					
4					NA
5				NA	NA
6			NA	NA	NA
7		NA	NA	NA	NA

[예측순서]
전체 데이터 모델로 예측
1번 데이터 모델로 예측 갱신
2번 데이터 모델로 예측 갱신
3번 데이터 모델로 예측 갱신

최적 방송 편성표

2. 상품별 실적 데이터 취급액 예측 모델



모델명	RMSE	MAE	MAPE
평균 앙상블 모델	36,746,012	21,969,168	0.3583

최적 방송 편성표

3. 각 주간 상품들에 대하여 일주일 동안의 모든 방송일시에 대한 취급액을 예측할 데이터

방송일시	시각	요일	노출시간	마더코드	상품코드	상품명	상품군	판매단가	...
2019-12-02 06:20:00	6	월요일	30	100586	201796	임페리얼 남성 패딩코트세트	의류	99,000	
2019-12-02 07:20:00	7	월요일	30	100586	201796	임페리얼 남성 패딩코트세트	의류	99,000	
2019-12-02 08:20:00	8	월요일	30	100586	201796	임페리얼 남성 패딩코트세트	의류	99,000	
					. . .				
2019-12-08 22:40:00	22	일요일	30	100586	201796	임페리얼 남성 패딩코트세트	의류	99,000	
2019-12-08 23:40:00	23	일요일	30	100586	201796	임페리얼 남성 패딩코트세트	의류	99,000	
2019-12-08 00:40:00	0	일요일	30	100586	201796	임페리얼 남성 패딩코트세트	의류	99,000	

최적 방송 편성표

4. 각 주간 상품들에 대한 모든 방송일시에 대한 취급액 예측

방송일시	시각	요일	노출시간	마더코드	상품코드	상품명	상품군	판매단가	…
2019-12-02 06:20:00	6	월요일	30	100586	201796	임페리얼 남성 패딩코트세트	의류	99,000	
2019-12-02 07:20:00	7	월요일	30	100586	201796	임페리얼 남성 패딩코트세트	의류	99,000	
2019-12-02 08:20:00	8	월요일	30	100586	201796	임페리얼 남성 패딩코트세트	의류	99,000	
					⋮				
2019-12-08 22:40:00	22	일요일	30	100586	201796	임페리얼 남성 패딩코트세트	의류	99,000	
2019-12-08 23:40:00	23	일요일	30	100586	201796	임페리얼 남성 패딩코트세트	의류	99,000	
2019-12-08 00:40:00	0	일요일	30	100586	201796	임페리얼 남성 패딩코트세트	의류	99,000	

취급액
예측

상품명	2019-12-02 06:20	2019-12-02 07:20	2019-12-02 08:20
임페리얼 남성 패딩코트세트 / 임페리얼 여성 패딩코트세트	53,520,738	76,154,766	81,371,144
스튜디오럭스 벤딩팬츠 4종	56,733,549	80,726,281	86,255,794
아르테사노 리얼 카이만악어 바디숄더백 / 테일숄더백 / 토트백	58,234,721	82,862,302	88,538,125

최적 방송 편성표

5. 헝가리안 알고리즘(Hungarian Algorithm)

- 할당 문제를 해결하는 알고리즘으로 1955년 Harold Kuhn에 의하여 개발
- 예를 들어, 각 노동자마다 해당 작업에 따른 비용을 정리한 표가 다음과 같을 때
  최소한의 비용으로 각 노동자마다 1개의 작업을 할당하는 방법을 헝가리안 알고리즘을 활용

	작업1	작업2	작업3
노동자 1	3	8	9
노동자 2	4	12	7
노동자 3	4	8	5

최적 방송 편성표

5. 헝가리안 알고리즘(Hungarian Algorithm)

- 할당 문제를 해결하는 알고리즘으로 1955년 Harold Kuhn에 의하여 개발
- 예를 들어, 각 노동자마다 해당 작업에 따른 비용을 정리한 표가 다음과 같을 때
  최소한의 비용으로 각 노동자마다 1개의 작업을 할당하는 방법을 헝가리안 알고리즘을 활용

	작업1	작업2	작업3
노동자 1	3	**8**	9
노동자 2	**4**	12	7
노동자 3	4	8	**5**

노동자 1 => 작업2
노동자 2 => 작업1
노동자3 => 작업3
최소 비용 : 17

- 헝가리안 알고리즘은 해당 값을 최소화하는 할당 알고리즘이지만 최대화하는 알고리즘으로도 사용 가능

최적 방송 편성표

5. 헝가리안 알고리즘(Hungarian Algorithm)

상품명	2019-12-02 06:20	2019-12-02 07:20	2019-12-02 08:20
임페리얼 남성 패딩코트세트 / 임페리얼 여성 패딩코트세트	53,520,738	76,154,766	81,371,144
스튜디오럭스 벤딩팬츠 4종	56,733,549	80,726,281	86,255,794
아르테사노 리얼 카이만악어 바디숄더백 / 테일숄더백 / 토트백	58,234,721	82,862,302	88,538,125

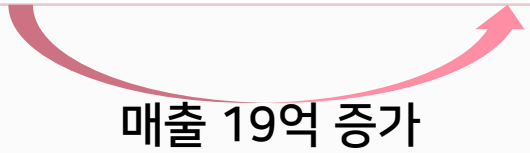
헝가리안
알고리즘 사용

상품명	최적 날짜	예상 취급액
임페리얼 남성 패딩코트세트 / 임페리얼 여성 패딩코트세트	2019-12-04 00:20:00	61,503,680
스튜디오럭스 벤딩팬츠 4종	2019-12-03 07:20:00	81,851,654
아르테사노 리얼 카이만악어 바디숄더백 / 테일숄더백 / 토트백	2019-12-05 08:20:00	87,713,873

최적 방송 편성표

EX) 2019년 12월 02일 ~ 12월 29일, 각 주간마다

	실제 취급액	기존 방송편성 예상 취급액	방송편성 최적화 예상 취급액
2019년 12월 02일 ~ 2019년 12월 08일	18,606,999,881	14,733,256,357	15,278,119,485
2019년 12월 09일 ~ 2019년 12월 15일	16,559,301,997	14,647,337,438	15,066,377,559
2019년 12월 16일 ~ 2019년 12월 22일	16,930,362,379	13,470,897,148	13,694,190,967
2019년 12월 23일 ~ 2019년 12월 29일	17,001,282,326	13,790,036,852	14,497,126,063
Total	69,097,946,583	56,641,527,795	58,535,814,074



최적 상품 노출시간

1. 방송 편성표가 완료

2. 예측 모델은 앞선 모델 그대로 사용

3. 주간 방송 편성표 중 방송 판매 상품 수가 2 이상인 방송들에 대하여 예측할 데이터 만들기

4. 최적 상품 노출시간 구하기

최적 상품 노출시간

1. 방송 편성표가 완료

상품명	최적 날짜	예상 취급액
임페리얼 남성 패딩코트세트 / 여성	2019-12-07 06:20	81,456,134
스튜디오럭스 벤딩팬츠 4종	2019-12-05 09:20	92,337,270
아르테사노 리얼 카이만악어 바디숄더백 / 테일숄더백 / 토트백	2019-12-05 18:20	102,759,386
푸마 드라이셀 쉐이핑 레깅스 3종 / 웜셀 퍼치마레깅스	2019-12-08 15:20	142,355,152
마르엘라로사티 에코무스탕1종	2019-12-02 07:20	81,045,619
무농약레드비트즙 90봉	2019-12-05 07:20	88,932,611
국내산 손질 통오징어 21미	2019-12-03 14:20	138,025,854
[RYN] 린 남성 다이얼락 히트 웜 퍼 방한화 2종 / 여성	2019-12-08 12:20	149,407,060
구스터 티포트 1+1 세트 /1세트	2019-12-02 08:20	94,644,991

빨간색 글씨로 된 상품명은 한 방송에 판매하는 상품 수가 2개 이상으로 방송시간을 나누어 각각 노출시간을 가짐

최적 상품 노출시간

3. 주간 방송 편성표 중 방송 판매 상품수가 2이상인 방송들에 대하여 예측할 데이터 생성

방송일시	시각	요일	방송시간	노출시간	상품코드	상품명	상품군	판매단가	…
2019-12-07 06:20:00	6	토요일	60	10	201796	임페리얼 남성 패딩코트세트	의류	99,000	
2019-12-07 06:20:00	6	토요일	60	50	201804	임페리얼 여성 패딩코트세트	의류	99,000	
2019-12-07 06:20:00	6	토요일	60	11	201796	임페리얼 남성 패딩코트세트	의류	99,000	
2019-12-07 06:20:00	6	토요일	60	49	201796	임페리얼 남성 패딩코트세트	의류	99,000	
					…				
2019-12-07 06:20:00	6	토요일	60	49	201796	임페리얼 남성 패딩코트세트	의류	99,000	
2019-12-07 06:20:00	6	토요일	60	11	201804	임페리얼 여성 패딩코트세트	의류	99,000	
2019-12-07 06:20:00	6	토요일	60	50	201796	임페리얼 남성 패딩코트세트	의류	99,000	
2019-12-07 06:20:00	6	토요일	60	10	201804	임페리얼 여성 패딩코트세트	의류	99,000	

(1) 각 상품들에 최소 노출시간인 10분은 항상 부여

(2) 각 상품들에 부여하고 남은 시간들을 1분 단위로 쪼개어 모든 경우의 수가 될 수 있도록 데이터를 생성

EX) 위의 표를 보면 방송 판매 상품수가 2개, 방송시간은 60분이므로 각 상품에 10분씩 부여하고 남은 40분을 (0분, 40분), (1분, 39분), (2분, 38분), …, (39분, 1분), (40분, 0분)으로 나누어 10분씩 부여된 각 상품 노출시간에 더함

최종적으로 (10분, 50분), (11분, 49분), …, (49분, 11분), (50분, 10분)

최적 상품 노출시간

4. 최적 상품 노출시간 구하기

상품명	최적 날짜	예상 취급액	최적 노출시간	최적 취급액
임페리얼 남성 패딩코트세트 / 여성	2019-12-07 06:20	81,456,134	30 / 30	81,456,134
아르테사노 리얼 카이만악어 바디숄더백 / 테일숄더백 / 토트백	2019-12-05 18:20	102,759,386	20 / 10 / 30	105,533,597
푸마 드라이셀 쉐이핑 레깅스 3종 / 웜셀 퍼치마레깅스	2019-12-08 15:20	142,355,152	30 / 30	142,355,152
[RYN] 린 남성 다이얼락 히트 웜 퍼 방한화 2종 / 여성	2019-12-08 12:20	149,407,060	30 / 30	149,407,060
구스터 티포트 1+1 세트 / 1세트	2019-12-02 08:20	94,644,991	30 / 30	94,644,991

빨간색 글씨로 된 상품명은 한 방송에 판매하는 상품 수가 2개 이상으로 방송시간을 나누어 각각 노출시간을 가짐

최적 상품 노출시간

EX) 2019년 12월 02일 ~ 12월 29일, 각 주간마다

	실제 취급액	기존 편성 예상 취급액	방송 편성 최적화 예상 취급액	노출시간 최적화 예상 취급액
2019년 12월 02일 ~ 2019년 12월 08일	18,606,999,881	14,733,256,357	15,278,119,485	15,340,100,848
2019년 12월 09일 ~ 2019년 12월 15일	16,559,301,997	14,647,337,438	15,066,377,559	15,279,421,974
2019년 12월 16일 ~ 2019년 12월 22일	16,930,362,379	13,470,897,148	13,694,190,967	14,180,159,265
2019년 12월 23일 ~ 2019년 12월 29일	17,001,282,326	13,790,036,852	14,497,126,063	14,661,260,173
Total	69,097,946,583	56,641,527,795	58,535,814,074	59,463,942,260

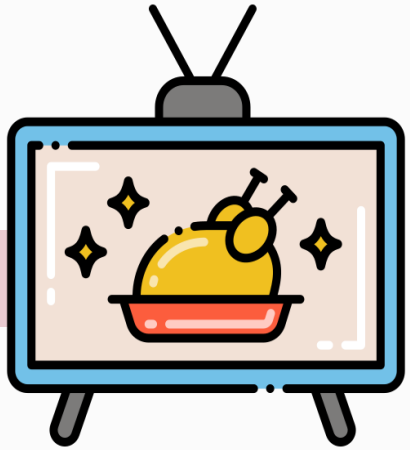
최적 상품 노출시간

EX) 2019년 12월 02일 ~ 12월 29일, 각 주간마다

	실제 취급액	기존 편성 예상 취급액	방송 편성 최적화 예상 취급액	노출시간 최적화 예상 취급액
2019년 12월 02일 ~ 2019년 12월 08일	18,606,999,881	14,733,256,357	15,278,119,485	15,340,100,848
2019년 12월 09일 ~ 2019년 12월 15일	16,559,301,997	14,647,337,438	15,066,377,559	15,279,421,974
2019년 12월 16일 ~ 2019년 12월 22일	16,930,362,379	13,470,897,148	13,694,190,967	14,180,159,265
2019년 12월 23일 ~ 2019년 12월 29일	17,001,282,326	13,790,036,852	14,497,126,063	14,661,260,173
Total	69,097,946,583	56,641,527,795	58,535,814,074	59,463,942,260

결론

최적 방송 편성표



최적 상품 노출시간

예상 취급액 최대화

아쉬운 점

1. 헝가리안 알고리즘을 사용하기 위하여 방송편성이 41~60분인 방송에 대해서만 최적화 실행 가능

   (방송 편성이 41~60분이 아닌 경우는 2.37%)

2. 주간 방송 편성표에 동일한 제품을 2번 이상 판매하는 경우가 존재하는데 헝가리안 알고리즘을 이용하면 두 제품이 비슷한 방송일시에 배치되는 문제점 발생 ➡ 비슷한 시간에 배치될 시 패널티 부여

3. 최적 상품노출시간을 구할 때 사용한 변수는 실제로 존재하는 값이 아닌 방송시간에 방송 상품 수를 나누어 구한 예상노출시간 변수이기 때문에 최적 상품 노출시간을 구하면 많은 경우에 모든 상품이 동등한 시간을 가질 때가 가장 최적이라고 제안하는 한계점 존재

   ➡ 실제 노출시간 변수가 존재한다면 더 좋은 최적 상품 노출시간을 구할 수 있을 것으로 판단

"감사합니다 ☺"