

2020  
빅콘테스트  
데이터분석  
챔피언리그

지난 월요일 7시에 팔았던 그 니트,  
이번 토요일 3시에 팔았다면?

팀 엄덕구:{

팀장: {김민수: zhddhkdn@naver.com}

팀원: {정혜인: heianjung@gmail.com}

팀원: {선형주: sawo101@naver.com}

팀원: {신왕수: thinpig99@gmail.com}}

# 목차

비즈니스 이해

01



데이터 이해

02



데이터 처리

03



예측 모델링

04



편성 최적화

05



비즈니스 적용

06



# 01

## Business Understanding

분석 배경

분석 목표와 방법론



# 분석배경

## 온디맨드 서비스 제공을 위한 편성 차별화

- 최근 마이크로 미디어와 오픈마켓의 성장으로 **TV쇼핑 성장세 둔화**
- 4년간 TV쇼핑 **송출 수수료 인상률은 302%**에 달하고 있는 상황
- 포스트 코로나 시대에 적절한 On-Demand 상품 기획 및 편성 필요**
- 내·외부 변수에 적절히 대처할 수 있는 편성 자동화 모델 구축 필요**
- 나아가 AI기반 개인화 맞춤형 쇼핑 환경을 구축할 수 있는 기반 마련



[위기의 홈쇼핑 탈출구 없나] ① 해마다 수직상승 '송출수수료' 이대로 괜찮나

'코로나19 전략'에 희비 엇갈린 홈쇼핑...'MD가 갈랐다'

□현대홈쇼핑, 1분기 영업이익 나란히 마이너스  
코로나19 사태 수혜 기대 불구 식품 외에는 감소세  
MD전략에 따라 각자 수익성취급고 엇갈려

기사입력 2020-05-11 14:17:12 | 강필성 기자 | feel@

신종 코로나 감염증(코로나19) 사태에 최대  
비가 엇갈리는 성적표를 받았다. 전반적으로  
이 극명하게 엇갈린 탓이다.



마트·홈쇼핑이 '날씨' 분석하자, 매출이  
뛰었다



11일 홈쇼핑 [위기의 홈쇼핑 탈출구 없나] ③진화하는 홈쇼핑

GS홈쇼핑은 메트로신문 신원선 기자 | 2020-08-13 15:08:09  
영업이익은 3

같은 기간  
익은 289억



롯데홈쇼핑 리얼피팅 서비스 시연 장면/롯데홈쇼핑

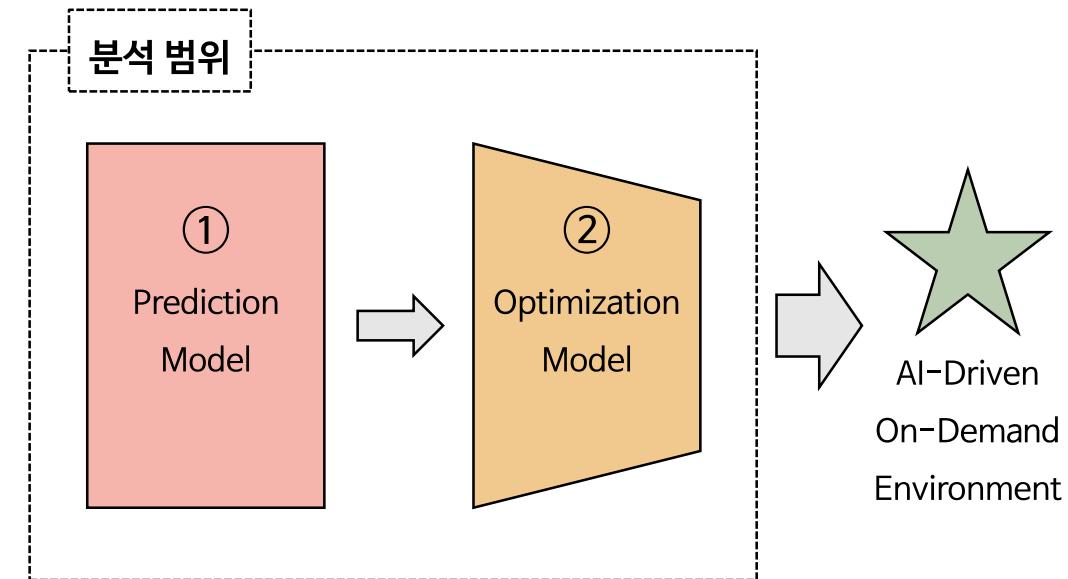
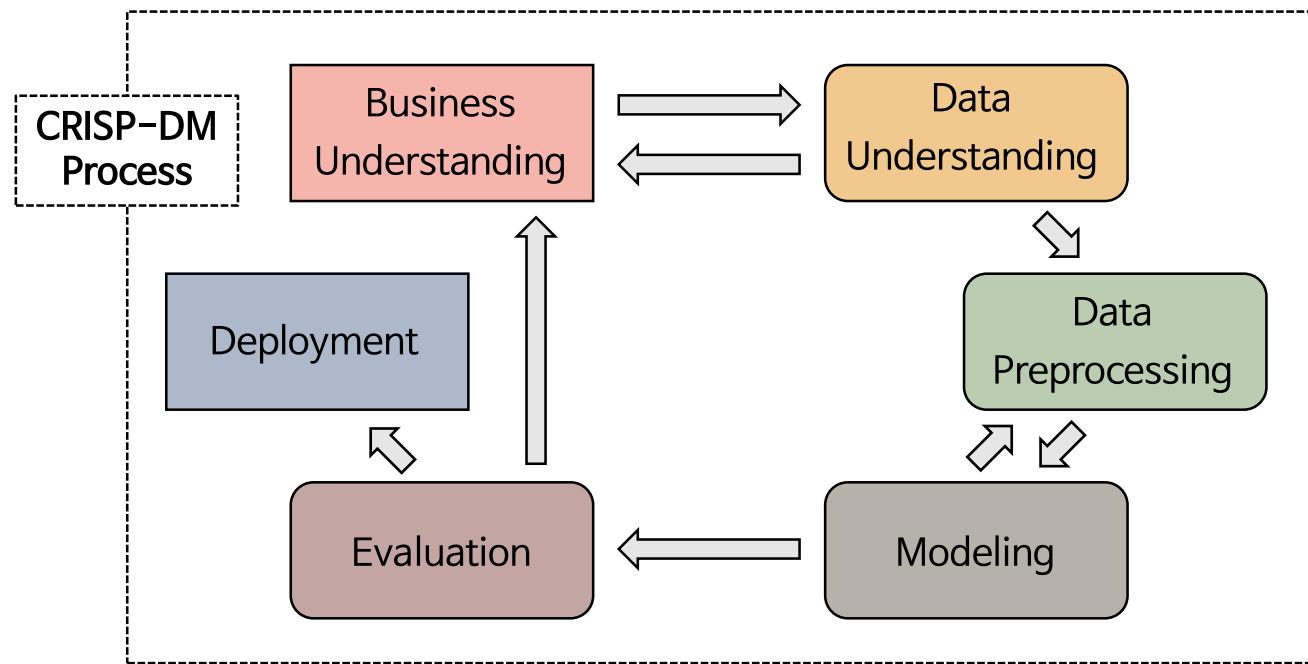
[위기의 홈쇼핑 탈출구 없나] ③진화하는 홈쇼핑

TV홈쇼핑이 거듭 진화하고 있다. 인공지능(AI)과 증강·가상현실(AR-VR), 빅데이터 등 정보통신기술을 접목해 소비자  
들의 쇼핑 편의성을 높이고 있다. 여기에 모바일에 익숙한 MZ세대를 겨냥해 모바일 경쟁력 또한 강화하고 있다.

# 분석목표와 방법론

예측 모델과 편성 자동화 모형 구축  
CRISP-DM 프로세스 적용

- 내·외부 데이터를 활용하여 특정 시점의 특정 아이템에 대한 취급액 예측 모델 구축
- 취급액 예측 모델을 기반으로 TV쇼핑 방송 편성표 최적화 모형 구축
- 이를 통해 내·외부 변화에 적절히 대응하는 편성 자동화 환경 마련
- 분석 방법론은 CRISP-DM 프로세스 활용



# 02

## Data Understanding

데이터 확보하기

데이터 살펴보기

데이터 탐색하기(EDA)



# 데이터 확보하기

날씨, 검색량, 가격비교, 주가지수 등

- 기본 제공 데이터를 비롯하여 다양한 외부 데이터 확보
- NS홈쇼핑: 기본 제공 데이터(2019.01 ~ 2019.12)
- 시청률 데이터: 기본 제공 데이터(2019.01 ~ 2019.12)
- 네이버 데이터랩: TV쇼핑 업체 검색량 지수 데이터 활용
- 네이버 쇼핑: 편성 상품 가격비교 데이터 활용
- 기상청 기상자료개발포털: 시간별 지상 기온 활용(서울 관측소 기준)
- Investing.com: KOSPI 지수 일일 종가 활용



# 데이터 살펴보기

## Data Shape, Data Type, 결측치

- Data Shape: Train Data(38309, 8) / Test Data(2891, 7)
  - 기본 제공 학습 데이터는 **약 3만 8천개의 데이터**가 있으며, 타겟 값인 **취급액**을 제외한 총 **7개의 변수**가 존재(Figure2-1)
  - 테스트 데이터는 **2020년 6월의 편성 테이블**이며, 제공된 학습데이터와 시계열적 간극이 있음(Figure2-2)
  - Nan Value: ‘**노출(분)**’, ‘**취급액**’ 변수에서 결측치가 다양 존재하며, ‘**노출(분)**’의 결측치는 **동일 방송일시 이종상품의 경우 Nan값으로 처리되었으며, ‘취급액’의 결측치는 취급액이 0원인 경우임**
- 〈Figure2-3〉〈Figure2-4〉

	방송일시	노출(분)	마더코드	상품코드	상품명	상품군	판매단가	취급액
0	2019-01-01 6:00	20.0	100346	201072	테이트 남성 셀린니트3종	의류	39,900	2,099,000
1	2019-01-01 6:00	Nan	100346	201079	테이트 여성 셀린니트3종	의류	39,900	4,371,000
2	2019-01-01 6:20	20.0	100346	201072	테이트 남성 셀린니트3종	의류	39,900	3,262,000
3	2019-01-01 6:20	Nan	100346	201079	테이트 여성 셀린니트3종	의류	39,900	6,955,000
4	2019-01-01 6:40	20.0	100346	201072	테이트 남성 셀린니트3종	의류	39,900	6,672,000

〈Figure 2-1〉

	방송일시	노출(분)	마더코드	상품코드	상품명	상품군	판매단가	취급액
0	2020-06-01 6:20	20.0	100650	201971	잭필드 남성 반팔셔츠 4종	의류	59,800	NaN
1	2020-06-01 6:40	20.0	100650	201971	잭필드 남성 반팔셔츠 4종	의류	59,800	NaN
2	2020-06-01 7:00	20.0	100650	201971	잭필드 남성 반팔셔츠 4종	의류	59,800	NaN

〈Figure 2-2〉

방송일시	0
노출(분)	16784
마더코드	0
상품코드	0
상품명	0
상품군	0
판매단가	0
취급액	2930

〈Figure 2-3〉

#	Column	Non-Null Count	Dtype
0	방송일시	38309	non-null object
1	노출(분)	21525	non-null float64
2	마더코드	38309	non-null int64
3	상품코드	38309	non-null int64
4	상품명	38309	non-null object
5	상품군	38309	non-null object
6	판매단가	38309	non-null object
7	취급액	35379	non-null object

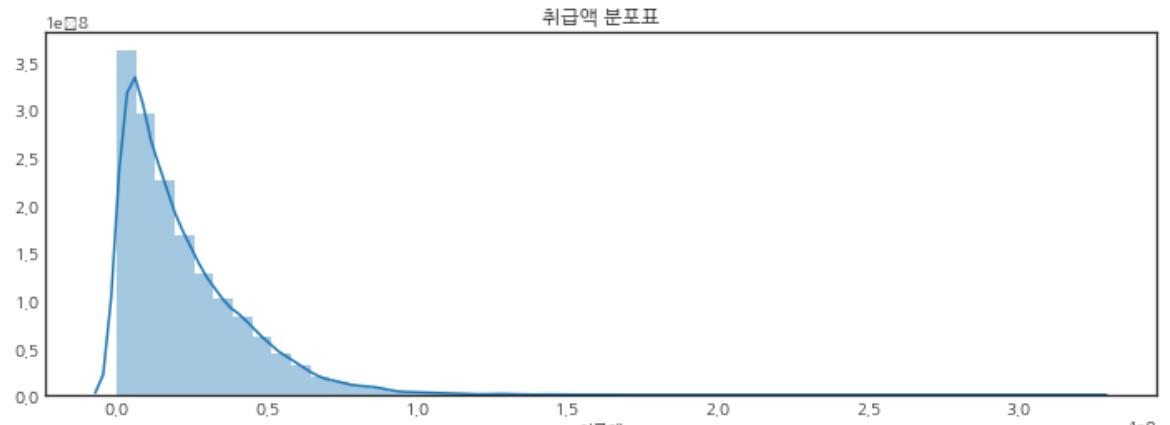
dtypes: float64(1), int64(2), object(5)

〈Figure 2-4〉

# 데이터 살펴보기

## 데이터 기술 통계량 확인

- 취급액에 대한 히스토그램: 눈에 띄게 불규칙한 분포는 보이지 않으나, **왜도 (Skewness)**가 매우 높게 형성되어 있음. **log화**를 통해 왜도(Skewness)를 조정해줄 경우 성능 개선을 기대할 수 있음 (Figure2-5)
- 기술통계량 분석: 평균 취급액 약 2,200만원이며, 중앙값은 1,600만원으로 형성 되어 있음. 3분위 값(75%)이 3,100만원인데 비해 최대값이 3억2천만 원으로 이상치가 매우 크게 형성되어 있음. 이상치 처리를 통해 성능 개선을 기대할 수 있음 (Figure2-6)



〈Figure 2-5〉

기술통계	취급액
Count	37,372
Mean	21,870,390
Std	20,194,290
Min	0
25%	6,880,750
50%	16,129,500
75%	31,631,250
max	322,009,000

〈Figure 2-6〉

# 데이터 살펴보기

## Train/Test Column별 Values 비교

- 범주형 변수 중 마더코드와 상품코드에 대한 Train/Test-set 비교
- 마더코드 비교:** Train-set에 존재하는 고유한 마더코드 개수 716개 중, Test-set에 **공통으로 존재하는 마더코드가 91개 밖에 되지 않음**  
(Figure2-7)
- 상품코드 비교:** Train-set에 존재하는 고유한 상품코드 개수 2,124개 중, Test-set에 **공통으로 존재하는 상품코드가 27개 밖에 되지 않음**  
(Figure2-8)
- 즉, 전년도에 판매한 상품을 그대로 당해에 재판매하는 경우가 매우 흔치 않으며, 두 변수가 모델의 예측 성능에 긍정적인 영향일 미칠 가능성 또한 적음**

Train-set에 있는 고유한 마더코드 개수 : 716  
Test-set에 있는 고유한 마더코드 개수 : 225

=====  
Train/Test-set에 공통으로 포함되어 있는 마더코드 개수 : 91

=====  
Train-set에만 있는 마더코드는 총 625개 입니다.

=====  
Test-set에만 있는 상품코드는 총 134개 입니다.

〈Figure 2-7〉

Train-set에 있는 고유한 상품코드 개수 : 2124  
Test-set에 있는 고유한 상품코드 개수 : 417

=====  
Train/Test-set에 공통으로 포함되어 있는 상품코드 개수 : 27

=====  
Train-set에만 있는 상품코드는 총 2097개 입니다.

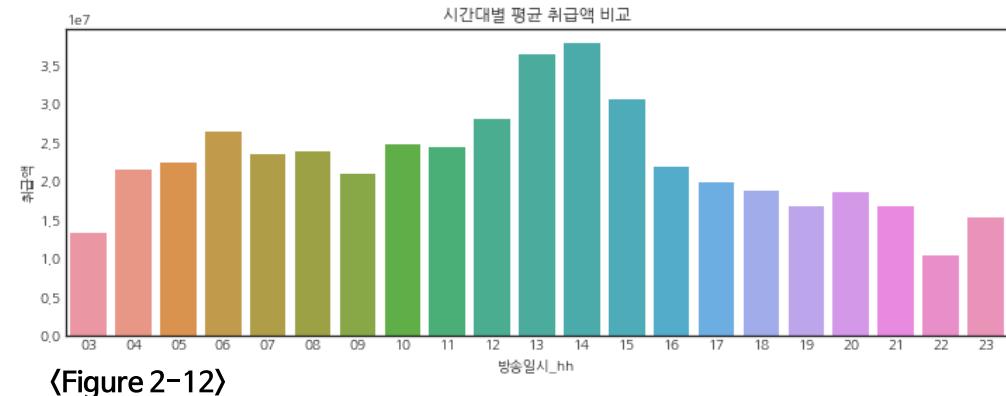
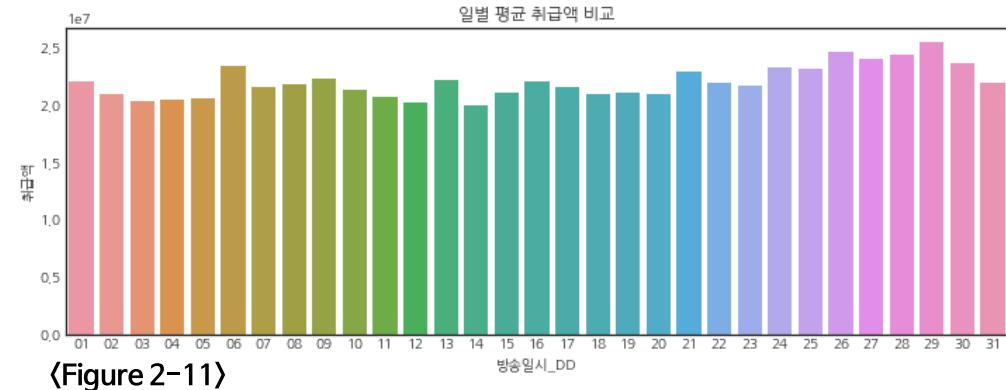
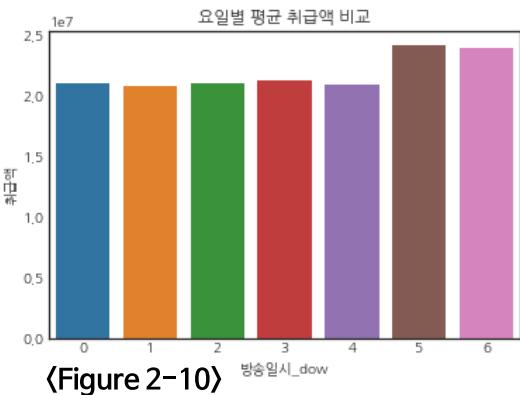
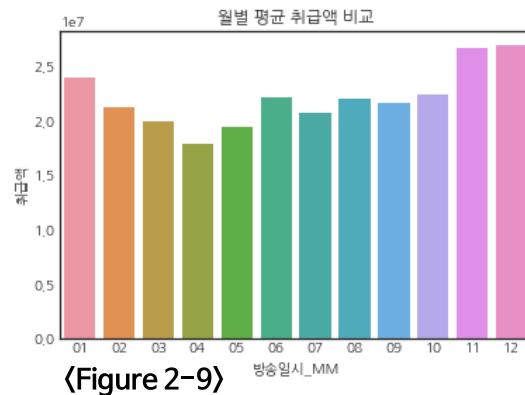
=====  
Test-set에만 있는 상품코드는 총 390개 입니다.

〈Figure 2-8〉

# 데이터 탐색하기

## 방송시간별 취급액 비교

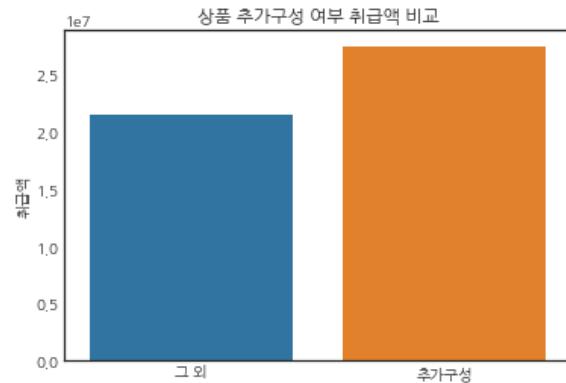
- 방송일시 변수를 `to_datetime()` 함수를 활용하여 월, 일, 시, 요일로 나눠 취급액 비교
- 월별: 겨울 시즌인 1, 11~12월에서 평균 취급액이 가장 높게 형성하고, 4월에 저점을 찍고 여름~가을에는 고른 분포를 보이는 것으로 확인  
(Figure2-9)
- 요일별: 평일보다 주말, 특히나 토요일에 높게 형성 (Figure2-10)
- 일별: 월말로 갈수록 취급액이 소폭 상승하는 경향을 보이며, 일주일 단위로 하루가 돋보이는 것으로 보아 요일의 영향이 클 것으로 확인  
(Figure2-11)
- 시간대별: 17시에 가장 높은 취급액을 형성하였고, 저녁 시간대보다 오전 시간대가 높게 형성(일 단위로 보기 위해 기준 시간에 -3h 적용되어 있음)  
(Figure2-12)



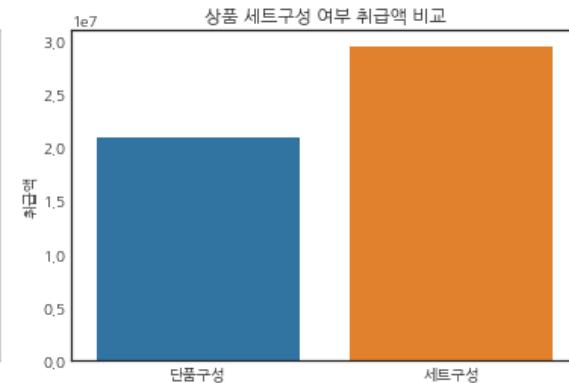
# 데이터 탐색하기

## 상품명별 취급액 비교

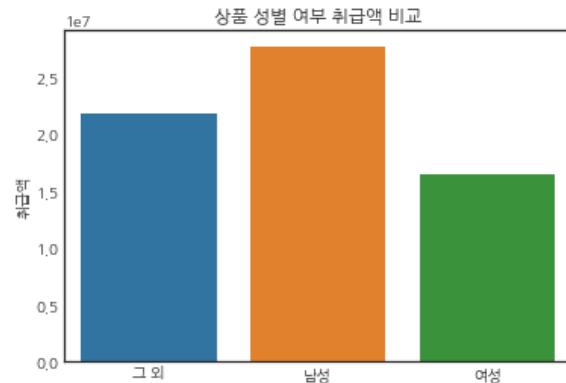
- 상품명 변수를 `.str.contains()` 함수를 활용하여 추가 변수를 생성하고 취급액 비교
- 추가구성 여부: 상품명에 '+'가 존재하는 상품이 그렇지 않은 상품보다 평균 취급액이 높게 형성 [\(Figure 2-13\)](#)
- 세트구성 여부: 상품명에 '세트'가 존재하는 상품이 그렇지 않은 상품보다 평균 취급액이 높게 형성 [\(Figure 2-14\)](#)
- 성별 구분: 상품명에 '여', '남' 구분이 되어 있는 상품이 그렇지 않은 상품과 다른 평균 취급액 형성 [\(Figure 2-15\)](#)
- 할부 플랜 구분: 상품명에 '무이자', '일시불' 구분이 되어 있는 상품이 그렇지 않은 상품보다 매우 낮은 평균 취급액 형성 [\(Figure 2-16\)](#)



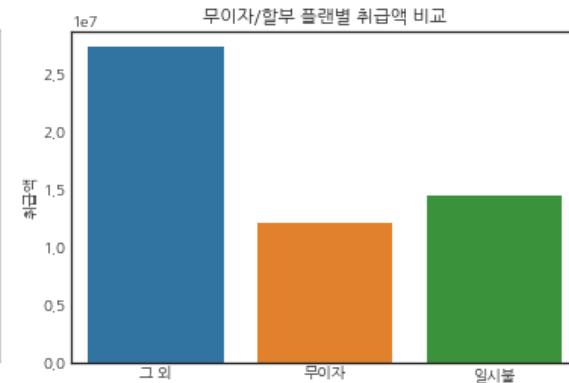
〈Figure 2-13〉



〈Figure 2-14〉



〈Figure 2-15〉

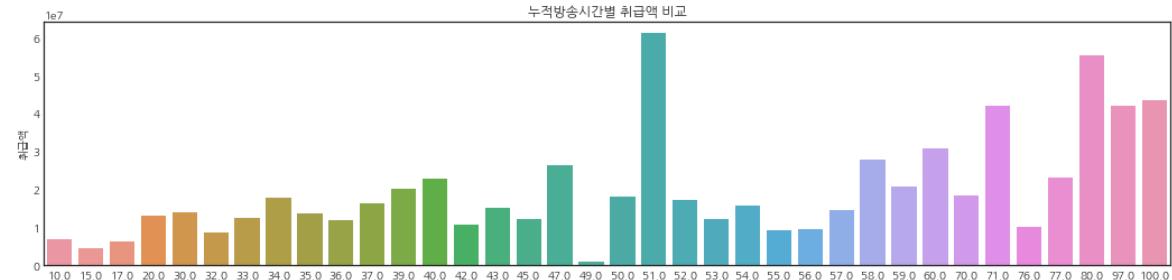


〈Figure 2-16〉

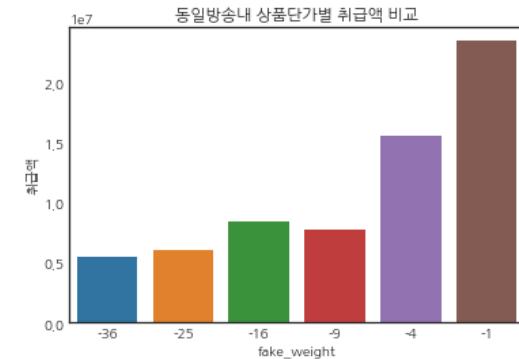
# 데이터 탐색하기

## 노출시간, 판매상품에 따른 취급액 비교

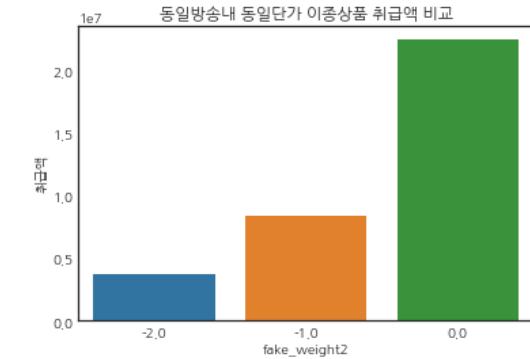
- 제공된 데이터 테이블에서는 한 편성이 보통 ‘20분 X 3타임’, ‘30분 X 2타임’으로 나뉘어져 있으며, 취급액도 타임별로 나뉘어져 있음
- 따라서 노출 시간을 누적하여 누적 시간별로 평균 취급액을 비교,  
누적 노출 시간이 길어질수록 취급액이 증가함 <Figure2-17>
- 제공된 데이터 테이블에서는 동일방송 내 판매상품이 다른 경우가 존재하며,  
단가가 같은 경우와 다른 경우가 존재함
- 동일방송 내 판매단가별 판매상품의 평균 취급액 비교 결과,  
판매단자가 낮을수록 평균 취급액이 낮게 형성되는 것을 확인 <Figure2-18>
- 동일방송 내 동일단가 이종상품의 평균 취급액 비교 결과,  
특정 상품의 취급액이 매우 높게 형성되는 것을 확인 <Figure2-19>



<Figure 2-17>



<Figure 2-18>



<Figure 2-19>

# 03

## Data Preprocessing

데이터 전처리

파생 변수 생성

외부 데이터 활용

변수 인코딩



# 데이터 전처리

## Log Transform, Outlier, Missing Value

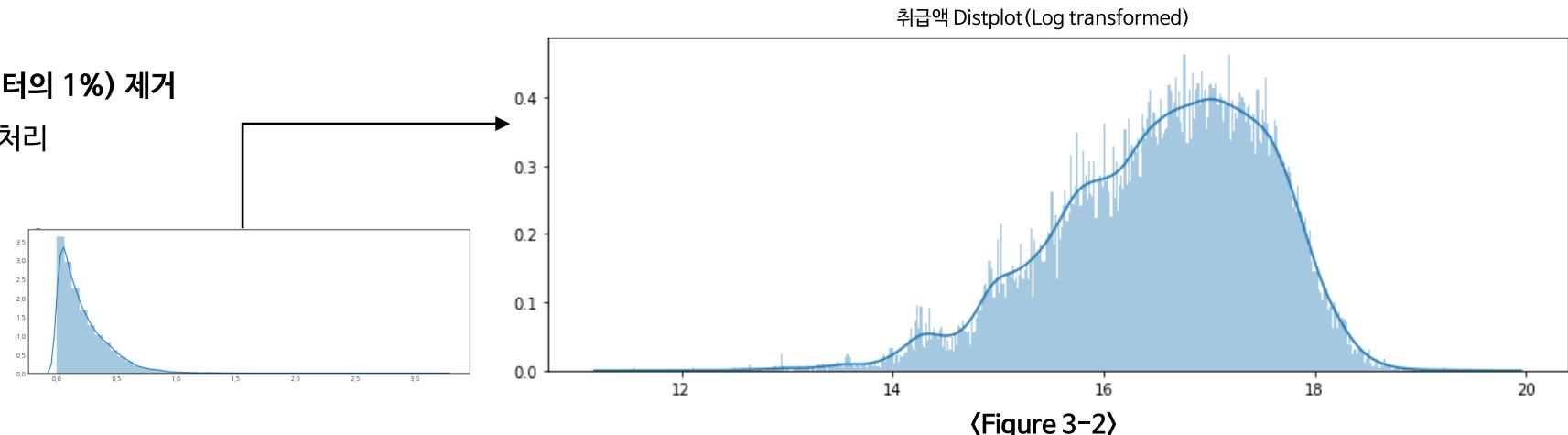
- Time adjustment: 자정~익일 새벽 2시까지 방송은 연·월·일이 다르게 할당되는 문제가 발생하여 일괄적으로 -3시간씩 시간을 조정 <Figure3-1>
- Log Transform: 취급액의 왜도(Skewness)를 조정해주기 위해 로그화 적용 <Figure3-2>
- Outlier: 취급액 1억원 이상 값(전체 데이터의 1%) 제거
- Missing Value: Forward Fill 방식으로 처리



	방송일시	방송일시_MM	방송일시_DD
0	2019-01-01 06:00:00	01	01
1	2019-01-01 06:00:00	01	01
2	2019-01-01 06:20:00	01	01
3	2019-01-01 06:20:00	01	01
4	2019-01-01 06:40:00	01	01
...	...	...	...
38304	2020-01-01 00:20:00	01	01
38305	2020-01-01 00:40:00	01	01
38306	2020-01-01 01:00:00	01	01
38307	2020-01-01 01:20:00	01	01
38308	2020-01-01 01:40:00	01	01

	방송일시	방송일시_MM	방송일시_DD
0	2019-01-01 03:00:00	01	01
1	2019-01-01 03:00:00	01	01
2	2019-01-01 03:20:00	01	01
3	2019-01-01 03:20:00	01	01
4	2019-01-01 03:40:00	01	01
...	...	...	...
37367	2019-12-31 20:40:00	12	31
37368	2019-12-31 21:00:00	12	31
37369	2019-12-31 21:00:00	12	31
37370	2019-12-31 21:00:00	12	31
37371	2019-12-31 21:00:00	12	31

<Figure 3-1>



# 파생 변수 생성

## 상품명 파생 변수 생성

- `.str.contains()`를 활용한 특정 문구 파생 변수 생성
- 상품명\_plan: 무이자/일시불 여부
- 상품명\_add: 추가 구성 여부
- 상품명\_maker: LG/삼성 여부
- 상품명\_set: 세트 구성 여부
- 상품명\_sex: 성별 구분 여부
- new\_상품명: 상품명 정규표현식 처리
- 상품명\_brand: new\_상품명의 앞 2글자 추출

	방송일시	노출(분)	마더코드	상품코드	상품명	상품군	판매단가	취급액	상품명_plan	상품명_add	상품명_maker	상품명_set	상품명_sex	new_상품명	상품명_brand
1400	2019-01-16 8:00	20.0	100088	200236	에코라믹 통주물 스톤 냄비 세트	주방	60900	17793000	그 외	그 외	그 외	세트 구성	그 외	에코라믹 통주물 스톤 냄비 세트	에코
7256	2019-03-15 16:00	20.0	100330	201038	LG생활건강 대크 헌장빨래 화이트업(30매X9박스)	생활용품	59800	15944000	그 외	그 외	그 외	LG 단품 구성	그 외	LG생활건강 테크 한장빨래 화이트업	LG
540	2019-01-07 8:00	20.0	100816	202404	보코 리버시블 무스탕	의류	79000	7153000	그 외	그 외	그 외	단품 구성	그 외	보코 리버시블 무스탕	보코
11220	2019-04-20 8:20	20.0	100448	201387	일시불 39년 신제품 쿠첸 풀스텐 압력밥솥 10인용(A1)	주방	168000	11539000	일시 불	그 외	그 외	단품 구성	그 외	쿠첸 풀스텐 압력밥솥	쿠첸
27799	2019-09-23 14:30	30.0	100845	202498	알비에로 마르티니 1A클라쎄 지오맵 라이트 백팩	잡화	279000	6950000	그 외	그 외	그 외	단품 구성	그 외	알비에로 마르티니 1A클라쎄 지오맵 라이트 백팩	알비
32721	2019-11-12 22:00	20.0	100148	200447	무이자 LG 울트라HD TV 65UM7900BNA	가전	1800000	36005000	무이 자	그 외	그 외	LG 단품 구성	그 외	LG 울트라HD TV 65UM7900BNA	LG
3548	2019-02-07 17:20	20.0	100708	202076	천수봉명인 선재 전통 메주 세트	농수축	96000	51963000	그 외	그 외	그 외	세트 구성	그 외	천수봉명인 선재 전통 메주 세트	천수
14584	2019-05-19 12:40	20.0	100251	200853	K-SWISS 여성PK 티셔츠 4종	의류	69900	16972000	그 외	그 외	그 외	단품 구성	여성	K SWISS 여성PK 티셔츠	K
182	2019-01-03 8:40	20.0	100774	202261	엘렌실라 달팽이크림(콜라겐+파트너)	이미용	79000	36208000	그 외	그 외	그 외	추가 구성	그 외	엘렌실라 달팽이크림	엘렌
23033	2019-08-08 6:00	20.0	100237	200807	이지엔 슬라이더 지퍼백	주방	39800	8100000	그 외	그 외	그 외	단품 구성	그 외	이지엔 슬라이더 지퍼백	이지

〈Figure 3-3〉

# 파생 변수 생성

## 방송시간

- `to_datetime()` 를 활용한 시간 관련 파생 변수 생성
- 방송일시\_dow: **요일 구분**(0~6, 월~일)
- 방송일시\_dow2: **주말, 평일 구분**(0=주말, 1=평일)
- 방송일시\_MM, DD, hh: **월, 일, 시 구분**(MM=월, DD=일, hh=시)
- 방송일시\_MMDD, DDhh, hhmm: **월일, 일시, 시분 구분**(MMDD=월일, DDhh=일시, hhmm=시분)
- 방송일시\_mmmm1, 2, 3: **일, 월, 년 단위 누적 시간(분 단위)**  
(mmmm\_1=일단위 누적, mmmm\_2=월단위 누적, mmmm\_3=연단위 누적)

	방송일시	방송일시_dow	방송일시_MM	방송일시_DD	방송일시_hh	방송일시_mm	방송일시_MMDD	방송일시_DDhh	방송일시_hhmm	방송일시_MMDDhh	방송일시_mmmm_1	방송일시_mmmm_2	방송일시_mmmm_3	방송일시_dow2
9055	2019-04-01 16:40:00	0	04	01	16	40	0401	0116	1640	040116	38400	38400	153600	1
3397	2019-02-06 06:20:00	2	02	06	06	20	0206	0606	0620	020606	7200	43200	86400	1
25815	2019-09-01 06:20:00	6	09	01	06	20	0901	0106	0620	090106	7200	7200	64800	0

〈Figure 3-4〉

# 파생 변수 생성

## 누적 노출시간, 동일방송 판매상품 구분

- 노출(분)을 활용한 파생변수 생성 <Figure3-5>
- cast time, cast count: 누적 노출 시간, 누적 노출 횟수 구분
- cast\_time\_sum, cast\_count\_sum: 해당 판매 상품의 총 누적 노출 시간, 총 누적 횟수 구분
- cast\_time\_ratio: 해당 판매 상품의 총 누적 노출 시간 대비 노출 시간 비율
- 동일방송 판매상품 구분을 활용한 파생변수 생성
- fake weight: 동일방송 내 상품단가별 역 가중치 부여 (-36~0) <Figure3-6>
- fake weight2: 동일방송 내 동일단가 이종상품별 역 가중치 부여 (-2~0) <Figure3-7>

방송일시	노출 (분)	마더코 드	상품코 드	상품명	상 품 군	판매단 가	취급액	cast_time	cast_count	cast_time_sum	cast_count_sum	cast_time_ratio
2019-01-01 6:00	20.0	100346	201072	테이트 남성 셀린니트3종	의류	39900	2099000	20.0	1	60.0	3.0	0.333333
2019-01-01 6:20	20.0	100346	201072	테이트 남성 셀린니트3종	의류	39900	3262000	40.0	2	60.0	3.0	0.666667
2019-01-01 6:40	20.0	100346	201072	테이트 남성 셀린니트3종	의류	39900	6672000	60.0	3	60.0	3.0	1.000000
2019-01-03 0:00	20.0	100346	201072	테이트 남성 셀린니트3종	의류	39900	7329000	20.0	1	40.0	2.0	0.500000
2019-01-03 0:20	20.0	100346	201072	테이트 남성 셀린니트3종	의류	39900	10481000	40.0	2	40.0	2.0	1.000000

<Figure 3-5>

방송일시	노출 (분)	마더코 드	상품코 드	상품명	상 품 군	판매단 가	취급액	fake_weight	fake_weight2	
15653	2019-05-29 08:00:00	20.0	100838	202461	무이자 쿠쿠 블랙스톤 전기밥솥 10인용	주 방	278000	16532000	-1	0.0
15654	2019-05-29 08:00:00	20.0	100838	202474	일시풀 쿠쿠 블랙스톤 전기밥솥 10인용	주 방	268000	3027000	-4	0.0
15655	2019-05-29 08:00:00	20.0	100838	202462	무이자 쿠쿠 블랙스톤 전기밥솥 6인용	주 방	248000	7405000	-9	0.0
15656	2019-05-29 08:00:00	20.0	100838	202475	일시풀 쿠쿠 블랙스톤 전기밥솥 6인용	주 방	238000	1815000	-16	0.0

<Figure 3-6>

방송일시	노출 (분)	마더코 드	상품코 드	상품명	상 품 군	판매 단가	취급액	fake_weight	fake_weight2	
1717	2019-01-19 16:00:00	20.0	100597	201832	컬럼비아 올니워크드로 즈 패키지 10종	속옷	79000	11564000	-1	0.0
1718	2019-01-19 16:00:00	20.0	100597	201836	컬럼비아 올니하트통드 로즈 패키지 7종	속옷	79000	5273000	-1	-1.0
1719	2019-01-19 16:00:00	20.0	100597	201834	컬럼비아 올니워크드로 즈 언더셔츠 6종	속옷	79000	3321000	-1	-2.0

<Figure 3-7>

# 외부 데이터 활용

## 네이버 API를 통한 상품 카테고리화

- 네이버 쇼핑 오픈 API 활용하여 상품명 검색을 통한 세부 카테고리화 <Figure3-8>
  - ① ‘상품명’ 변수 정규표현식 전처리 실시 -> ‘new\_상품명’
  - ② ‘new\_상품명’을 통해 카테고리 분류 실시
  - ③ ②에서 분류하지 못한 상품은 자카드 유사도를 통해 유사 카테고리로 할당
- cat1: 대분류로서 총 9개의 분류로 할당 <Figure3-9>
- cat2: 중분류로서 총 87개의 분류로 할당 <Figure3-10>
- cat3: 소분류로서 총 249개의 분류로 할당 <Figure3-11>

	상품명	new_상품명	상품군	cat1	cat2	cat3
7144	NNF SS트레이닝 세트	NNF SS트레이닝 세트	의류	패션의류	남성의류	트레이닝복
35789	고칼슘검은콩두유48팩+호두아몬드 두유48팩	고칼슘검은콩두유 호두아몬드 두유	농수축	식품	가공식품	두유
12833	2019 S/S 기라로쉬 올인원 선글라스	기라로쉬 올인원 선글라스	잡화	패션잡화	선글라스/안경테	선글라스
13440	일시불 삼성 UHD TV UN55NU7050F	삼성 UHD TV UN55NU7050F	가전	디지털/가전	영상가전	TV
8791	보루네오 루나 유로탑 멀티수납형 LED 침대 K 킹	보루네오 루나 유로탑 LED 침대	가구	가구/인테리어	침실가구	침대
33674	푸마 코튼 언더탑 9종	푸마 코튼 언더탑	속옷	패션의류	남성언더웨어/잠옷	러닝

<Figure 3-8>

네이버쇼핑API_대분류	
0	패션의류
1	생활/건강
2	식품
3	화장품/미용
4	디지털/가전
5	스포츠/레저
6	패션잡화
7	가구/인테리어
8	출산/육아

<Figure 3-9>

네이버쇼핑API_중분류	
0	남성의류
1	여성의류
2	여성언더웨어/잠옷
3	주방용품
4	냉동/간편조리식품
...	...
82	교재/서적
83	장갑
84	서재/사무용가구

<Figure 3-10>

네이버쇼핑API_소분류	
0	니트/스웨터
1	브라팬티세트
2	바지
3	코트
4	냄비/솥
...	...
244	의자
245	섬유유연제
246	복근운동기구

<Figure 3-11>

# 외부 데이터 활용

## 네이버 쇼핑, 기상청, 코스피 지수

- 네이버 쇼핑 크롤링을 통한 파생 변수 생성 <Figure3-12>
- review\_counts: 네이버 쇼핑 리뷰 개수
- internet\_price: 네이버 쇼핑 기준 단가
- price\_minus: 판매단가 대비 네이버 쇼핑 기준 단가 차액
- search\_naver: 네이버 쇼핑 판매 여부
- 기상청 데이터 활용하여 파생 변수 생성 <Figure3-13>
- temperature: 서울 관측소 기준 시간별 지상 기온
- 코스피 지수 활용하여 파생 변수 생성 <Figure3-14>
- 변동 %: 코스피 지수 일일 변동률

		상품명	상품군	판매단가	취급액	review_counts	internet_price	price_minus	search_naver
9870	비버리힐스풀로클럽 남성기초세트(골드+플라)	이미용	39800	24603000	2697	32200	7600	1	
429	무이자 LG 울트라HD TV 65UK6800HNC	가전	2130000	46402000	2	1894000	236000	1	
22807	로베르타 디 까메리노 풀 바스트 업 지퍼브라팬티	속옷	89900	13863000	0	0	89900	0	
1494	키친플라워 참소쿠리 2세트	주방	39800	13073000	21	32940	6860	1	
23926	임성근의 녹용도가니탕 풀세트	농수축	55900	21142000	294	54900	1000	1	
28230	노와 멀티 만능 다지기 1+1세트	주방	59800	19532000	0	0	59800	0	

<Figure 3-12>

	방송일시	노출(분)	마더코드	상품코드	상품명	상품군	판매단가	취급액	temperature
18100	2019-06-22 08:00:00	20.0	100188	200640	코치 아웃라인 파일백	잡화	138000	6511000	39.2
9719	2019-04-07 10:00:00	20.0	100837	202468	무이자 쿠쿠전기밥솥 6인용	주방	208000	3918000	31.1
8744	2019-03-30 11:00:00	20.0	100271	200889	해스티지 그레이스 양가죽 코트	의류	199000	13162000	12.3
34575	2019-12-02 18:20:00	20.0	100327	201031	[VONIN]보닌 남성기초세트	이미용	54000	61430000	0.0
37155	2019-12-29 09:40:00	20.0	100638	201956	램프룩 자동회전냄비	주방	109000	71768000	0.4
5077	2019-02-22 12:30:00	30.0	100427	201340	일시불 프로파룩 에어프라이어	주방	94000	34384000	17.0

<Figure 3-13>

	방송일시	노출(분)	마더코드	상품코드	판매단가	상품명	변동 %
36502	2019-12-21 19:00:00	20.0	100837	202470	208000	무이자 쿠쿠전기밥솥 6인용(QS)	0.00
2187	2019-01-24 07:00:00	20.0	100026	200039	40900	궁중 손질새우 200미 + 동태포 200g	0.81
22795	2019-08-05 19:20:00	20.0	100837	202470	208000	무이자 쿠쿠전기밥솥 6인용(QS)	-2.56
9724	2019-04-07 10:20:00	20.0	100837	202471	198000	일시불 쿠쿠전기밥솥 6인용	0.00
15625	2019-05-28 22:00:00	20.0	100715	202090	49000	젠틀월 스트레치 통풍 메쉬 자켓 1종	0.23

<Figure 3-14>

# 변수 인코딩

## Label Encoder 활용

- 모델의 원활한 학습을 위해 범주형 변수 인코딩
- Label Encoder를 통해 범주형 변수를 정수 인덱싱 처리

⟨Figure3-15⟩⟨Figure3-16⟩

상품명	상품군	상품명_brand	cat1	cat2	cat3	new_상품명
테이트 남성 셀린니트3종	의류		테이	패션의류	남성의류	니트/스웨터 테이트 남성 셀린니트
테이트 여성 셀린니트3종	의류		테이	패션의류	여성의류	니트/스웨터 테이트 여성 셀린니트

⟨Figure 3-15⟩

encoding_상품명:	encoding_상품군:	encoding_상품명_brand:	encoding_cat1:	encoding_cat2:	encoding_cat3:	encoding_new_상품명:
373	3	196	4	42	236	659
1434	8	323	7	48	219	1002

⟨Figure 3-16⟩

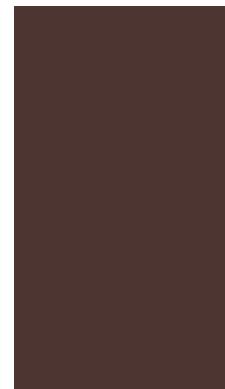
# 04

## Modeling

바닐라 모델링

모델 튜닝 결과

예측 모델 모형



# 바닐라 모델링

## 모델별 기본 성능 비교

- 트리기반, 선형기반, 부스팅 기반의 모델들을 튜닝 없이 바닐라 모델로 기본 성능 비교
- 기본 성능이 가장 좋은 상위 2개의 모델을 선정해 채택 후 파라미터 튜닝 실시
- RF(Random Forest Regressor), XGB(XGBoost Regressor)까지 2개의 모델 선정

모델	MAPE	비고
RF	36.63	채택
XGB	39.67	채택
h2o RF	41.42	기각
lgbm	52.71	기각
LR	57.85	기각
GBR	58.12	기각
Ridge	58.20	기각
Lasso	81.28	기각

# 모델 튜닝 결과

## 베이즈 서치를 통한 모델 파라미터 튜닝

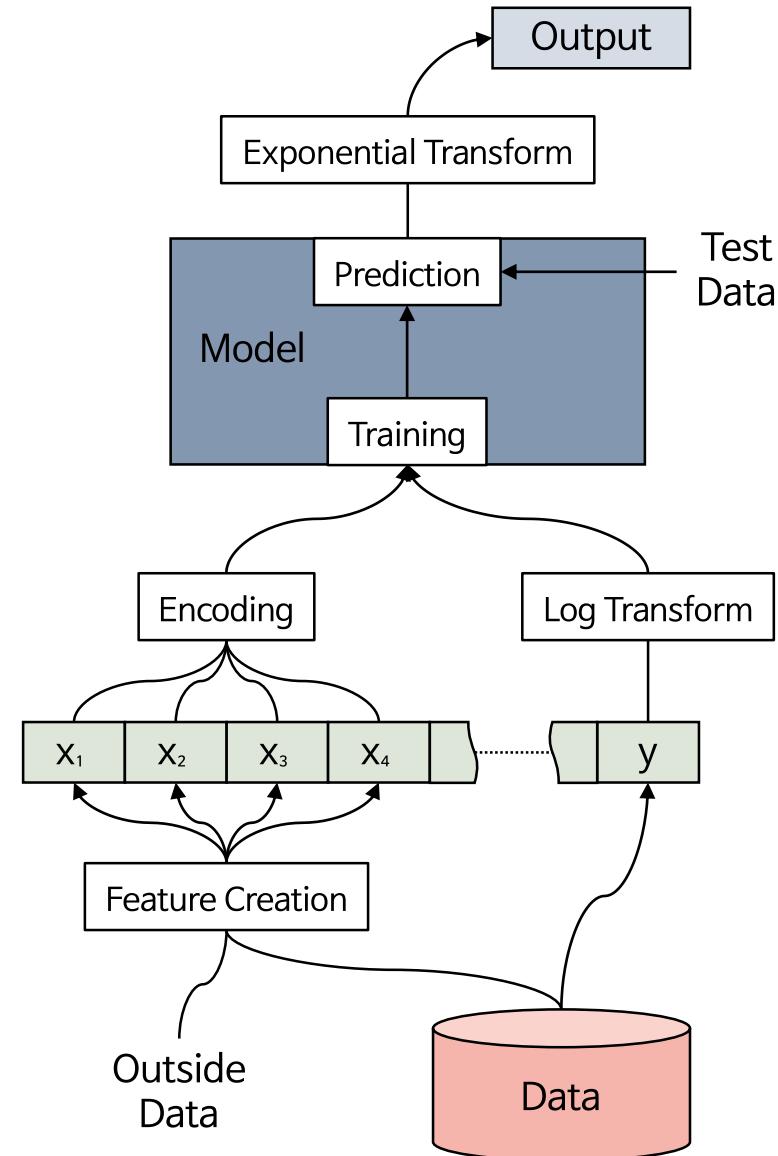
- ‘베이즈 서치’를 통한 모델 최적의 파라미터 값 추출
- 해당 파라미터 값을 통한 validation\_score 출력
- XGB 모델이 RF보다 더 높은 성능을 보임
- 최종 validation\_score : 29.68 기록

모델	최적 파라미터	MAPE
RF	{           max_depth = 20, max_features = ‘auto’, min_samples_leaf = 1, min_samples_split = 2, n_estimators = 1000         }	33.64
XGB	{           learning_rate = 0.247, max_depth = 7, n_estimators = 300, colample_bytree = 0.31         }	29.68

# 예측 모델 모형

## TV쇼핑 편성 상품 취급액 예측 모형

- ① 내·외부 데이터에 대한 파생 변수 생성
- ② 독립변수에 대한 encoding 작업
- ③ 종속 변수에 대한 log transform 작업
- ④ 예측 모델 학습
- ⑤ Test Data에 대한 prediction 작업
- ⑥ Prediction에 대한 exponential transform 작업
- ⑦ 최종 예측 결과 추출



# 05

## Optimization Modeling

모델링 아이디어

편성 최적화 모형

샘플 최적화 결과

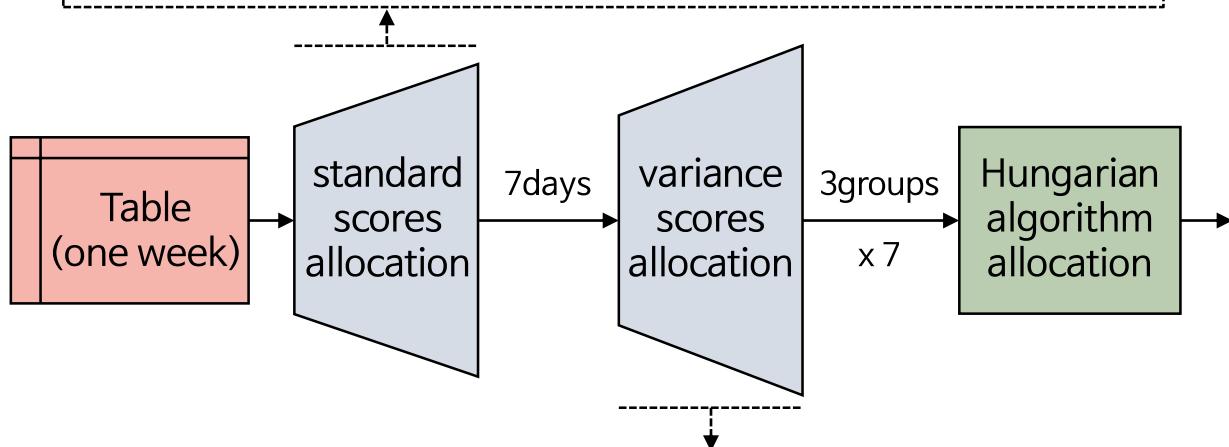


# 모델링 아이디어

## 연산량을 줄여 전수조사를 가능하게 하자

- 가장 정확한 방법은 모든 경우의 수를 조합하여 취급액을 비교하는 것  
→ 경우의 수가 기하급수적으로 늘어나 **연산이 불가능**
- 할당 최적화 알고리즘인 헝가리안 알고리즘을 사용하여 최적화 진행  
→ 높은 시간 복잡도( $n$ 의 4승)에 의해 **연산이 불가능**
- 스테이지를 통해 후보군을 할당해가며 병렬적으로 헝가리안 최적화를 진행  
→ 병렬 연산이 가능해지며 전수조사에 가까운 성능을 기대할 수 있음

- 표준점수 할당 알고리즘
  - 상품별 취급액의 표준점수 추출
  - 상품별 가장 높은 표준점수를 가진 요일로 할당
  - 각 요일은 할당된 상품의 표준점수 내림차순으로 상품 확정
  - 할당되지 못한 상품은 다음으로 높은 표준점수를 가진 요일로 재할당
  - ③~④를 반복하여 모든 요일에 균등하게 할당

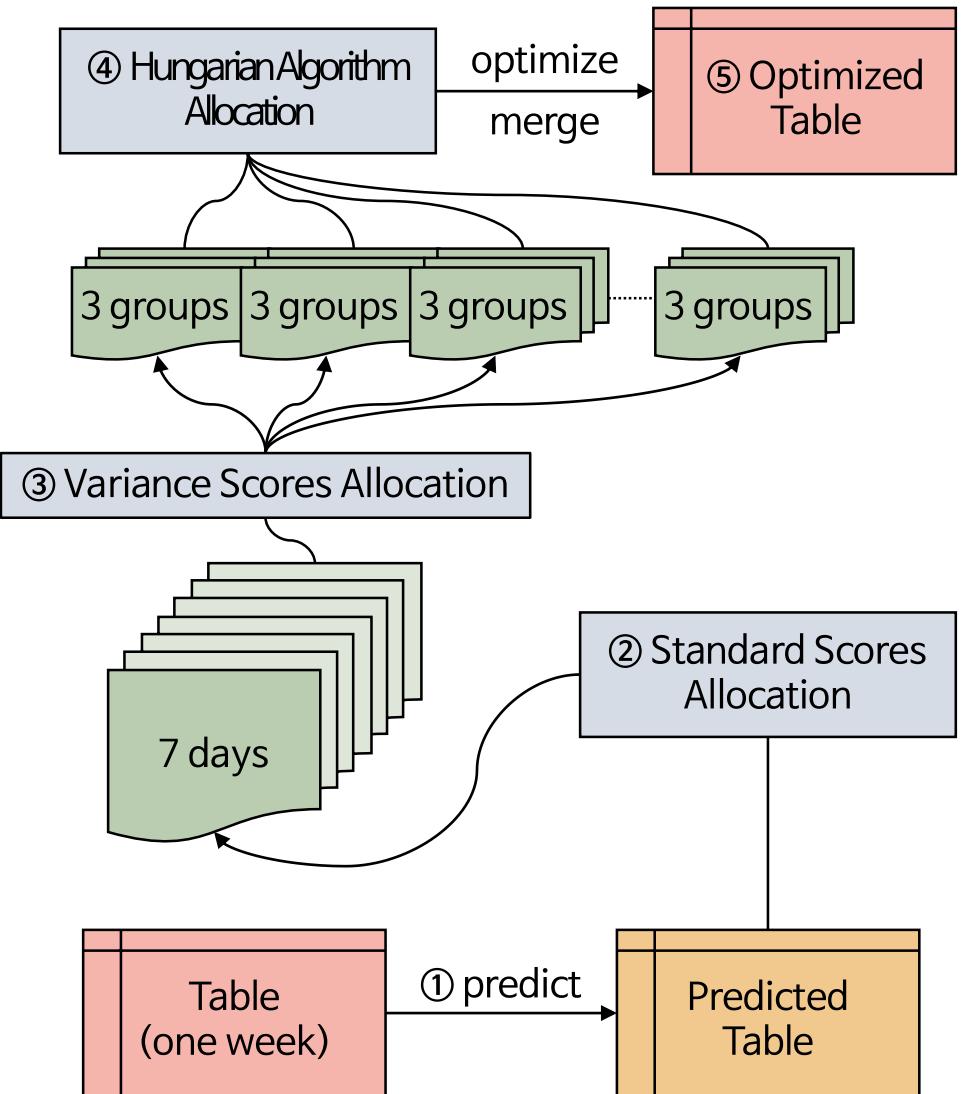


- 분산점수 할당 알고리즘
  - 시간별·상품별 분산 값 추출
  - 분산 값을 기준으로 내림차순, 상단부터 순차적으로 3그룹에 할당
  - 시간대별 그룹과 상품별 그룹을 1:1 매칭
  - (6, 6) or (7, 7) 크기의 (시간, 상품) 행렬 완성

# 편성 최적화 모형

## 3-Stage Allocation Model

- ① 일주일간 편성되어 있는 상품을 전부 특정 시간(분산이 큰 시간)기준으로 예측모델을 활용하여 예상 취급액 추출
- ② 표준점수 할당 알고리즘을 통해 각 상품을 최적의 요일로 균등 할당
- ③ 분산점수 할당 알고리즘을 통해 요일별로 상품 후보를 3그룹으로 균등 할당
- ④ 시간그룹과 상품그룹을 1:1 매칭하여 헝가리안 알고리즘으로 최적 할당
- ⑤ 최적화 결과를 합쳐서 주간 편성표 확정



# 샘플 최적화 결과

## 기존 편성 대비 2.21배 취급액 상승<sup>1)</sup>

- 기준에 제공된 테스트 데이터(2020년 6월분)에 대한 예측 취급액과 편성 최적화 모델(3-Stage Allocation Model)을 통해 최적화된 취급액 비교
- 원활한 비교를 위해 6월 1일 ~ 6월 7일(월~일)의 일주일간의 샘플데이터 활용
- 해당 기간의 상품 중 총 1시간의 방송을 하며 20분 단위로 노출되는 상품만 활용
- (시간:일주일 x 상품:132개)의 편성표를 편성 최적화 모델을 통해 재편성
- 기준을 벗어나는 예외 상품, 미방송 시간대는 최적화 제외

구분	총 예측 취급액(원)	비율(배)
기존 편성표	5,462,293,510	-
최적화된 편성표	12,112,904,770	-
비교	+ 6,650,611,259	2.21

1) 특정 기간의 샘플 분석 결과로 전체 기간의 편성 최적화 경우 오차가 발생할 수 있음

# 06

## Deployment

비즈니스 적용 및 기대효과



# 비즈니스 적용 및 기대효과

## 날씨, 주가 등의 외부 요인을 고려한 방송 편성

- 기상 데이터를 활용하여 날씨에 따른 전략적 상품 배치 가능
- 주가 데이터를 활용하여 경제 지표에 따른 소비 심리 변화 반영
- 그 외 추가적인 외부 변수를 통해 코로나19 등의 사회 변화에 따른 소비 변화 선제적 대응 가능



# 비즈니스 적용 및 기대효과

## 신규 상품 기획 및 신규 편성에 대한 비용 감소

- 신규 상품 기획 시 고려해야하는 상품의 구성, 가격, 프로모션 등에 대한 취급액을 예측하여 최적의 상품 기획 가능
- 예측 모델에 활용된 주요 변수(상품명, 방송 구성 등)을 상품 기획 시 정성적 고려
- 데이터 기반의 상품기획을 통해 타당성 있는 업무 진행과 높은 ROI 기대
- 상품에 대한 취급액 증가로 협력사와 NS홈쇼핑의 동반 성장 가능



# 비즈니스 적용 및 기대효과

## 단계별 최적화 모델의 적은 연산량을 통한 편성 최적화

- 3단계의 최적화 모델을 통해 연산량을 현저히 줄여 빠른 편성 최적화 가능
- 긴급한 방송 편성의 변경 시 빠른 긴급 편성 가능
- 적은 연산량을 기반으로 장기 편성 기획 가능



# 감사합니다.

## 팀 엄덕구:{

팀장:{김민수: zhddhkdn@naver.com}

팀원:{정혜인: heianjung@gmail.com}

팀원:{선행주: sawo101@naver.com}

팀원:{신왕수: thinpig99@gmail.com}}