

NS SHOP+

판매실적 예측을 통한 편성 최적화 방안 도출

FC성수

[팀장] 김희주 : wngmlrla0320@naver.com

[팀원] 이태헌 : taeheon7753@gmail.com

임상우 : samcjswok@gmail.com

최은비 : p5653ceb@gmail.com

CONTENTS



1 프로젝트 개요

- 1) 프로젝트 진행 배경
- 2) 프로젝트 목적 및 개요

2 프로젝트 세부 내용

- 1) 분석 프로세스
- 2) 데이터 전처리 및 eda
- 3) 데이터 분석
- 4) 최종 예측 결과



01

프로젝트 개요

01. 프로젝트 개요

프로젝트 진행 배경

T커머스에서의 경쟁력 강화를 위해서는 상품 및 내부요인과 함께 기상, 휴일 등 다양한 외부요인들을 충분히 고려하여 최적 수익을 창출하도록 방송을 편성해야 합니다. 본 프로젝트에서는 T커머스 매출에 영향을 주는 내/외부 요인을 충실히 반영하여 최적 편성 모델을 제안하고자 합니다.

“T-커머스의 매출은 방송 채널 및 상품 자체의 경쟁력 외에도 하루 24시간의 방송 시간 중 어떤 시간에 어떤 상품을 편성하였는가에 큰 영향을 받는다.”

- 『딤러닝과 통계 모델을 이용한 T-커머스 매출 예측』 논문 중

“백화점, 마트, 홈쇼핑, 편의점 등 소매 유통 산업 업체들은 이미 주단위, 혹은 월단위로 **기상을 예측하고 매출을 분석**해 다음 판매 전략 수립에 활용하고 있다.”

- 『마트·홈쇼핑이 '날씨' 분석하자, 매출이 뛰었다』 기사 중

내부요인의 강화와 함께 **외부요인을 충분히 고려한 최적 편성 모델 개발 필요**

상품요인

- 재고확보 가능여부
- 판매가격
- 상품 경쟁력 및 특징

내부요인

- 우수채널 보유여부
- 우수상품 소싱능력
- 품질관리 능력
- 고객관리 능력

외부요인

- 날씨
- 편성일(휴일 등)/빈도
- 동시간대 타채널 판매상품
- 동시간대 타방송 프로그램
- 코로나 등 이슈 및 이벤트
- 사회/경제 트렌드 변화

* 출처: 딤러닝과 통계 모델을 이용한 T-커머스 매출 예측, 김인중 외 6명, 정보과학회 논문지 44권 8호 2017. 08.
『마트·홈쇼핑이 '날씨' 분석하자, 매출이 뛰었다』, 머니투데이, 2020.8.5
엔에스쇼핑공사 반기보고서(2019.06).

프로젝트 목적 및 개요

본 프로젝트의 취지와 홈쇼핑의 특수성을 충분히 고려하여 본 프로젝트를 다음과 같이 진행하였습니다.

프로젝트 명	NSshop+ 판매실적 예측을 통한 편성 최적화 방안 도출	
프로젝트 목적	T커머스의 내/외부 요인을 반영하여 최적 수익을 도출하는 편성을 제안함으로써 NS Shop+의 경쟁력 강화를 지원하는 것을 목적으로 함	
프로젝트 진행 기간	2020년 8월 3일 ~ 2020년 9월 28일 (약 2개월)	
프로젝트 범위	판매실적 예측	2019년 1년의 판매 데이터를 기반으로 2020년 6월 판매실적 예측
	편성 최적화 방안 제시	최적 수익을 고려한 요일별/ 시간대별/ 카테고리별 편성 최적화 방안(모형)제시

02

프로젝트 세부내용

02

프로젝트 세부내용

2-1

분석 프로세스

2-2

데이터 전처리 및 EDA

2-3

데이터 분석

2-4

예측 및 성능개선

2-1. 분석 프로세스

분석 프로세스 개요

본 프로젝트는 '데이터 전처리 및 EDA', '데이터 모델링', '예측 및 성능 개선'의 세 단계로 진행하였습니다.



02

프로젝트 세부내용

2-1

분석 프로세스

2-2

데이터 전처리 및 EDA

2-3

데이터 분석

2-4

예측 및 성능개선

데이터 전처리

본 프로젝트에서 제공받은 데이터는 2019년도 실적데이터와 시청률 데이터입니다.

제공데이터

정보 보안상의 문제로 공개하지 않습니다.

데이터 전처리

제공 데이터 중 실적데이터의 결측치와 2020년 1월 데이터, 취급액이 50,000원인 이상치를 제거하였으며, 노출(분)의 결측치는 동일 방송의 노출(분) 데이터와 동일하게 삽입하였습니다.

결측치 제거

정보 보안상의 문제로 공개하지 않습니다.

방송의
노출(분)

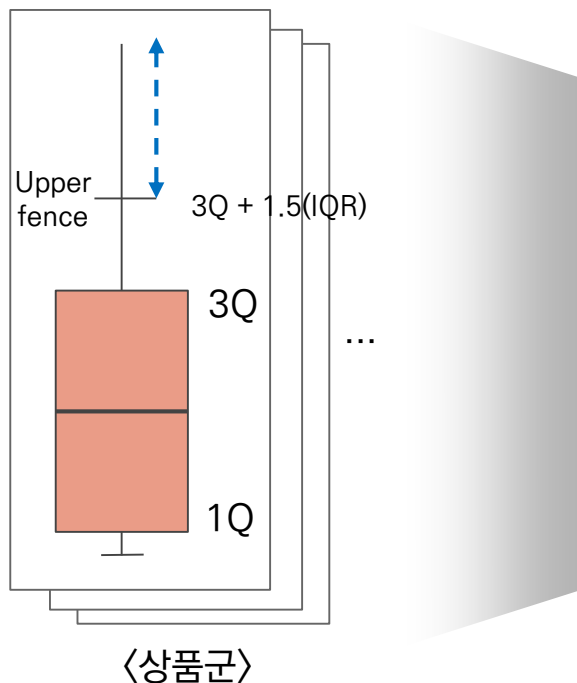
데이터 전처리

전체 데이터를 상품군으로 구분하였고, 각 상품군별로 upper fence를 벗어난 이상치를 제거하였습니다. 가구, 가전, 침구, 잡화에서 이상치가 발견되었으며 총 28개를 제거하고 분석을 진행하였습니다.

이상치 제거

- 상품군별 취급액 기준 upper fence를 벗어난 이상치 데이터 제거
- 전체 35,375 건의 데이터 중 총 28개의 이상치 데이터 제거

〈매출액 분포〉



상품군별 이상치

정보 보안상의 문제로 공개하지
않습니다.

데이터 전처리

제공데이터인 실적 데이터와 시청률 데이터를 함께 분석하기 위해 하나의 dataframe으로 병합하였습니다.

시청률 데이터 merge

정보 보안상의 문제로 공개하지 않습니다.

변수 생성 및 탐색

예측의 정확성을 높이기 위하여 내부변수와 외부변수에서 변수를 추가하고 생성하는 활동을 진행하였습니다.

생성 변수 정리

내부변수		외부변수	
변수 생성		변수 추가	변수 생성
방송일시 세분화 <ul style="list-style-type: none"> • 방송요일 • 연휴 일차 • 연휴 길이 • 방송 시간 sin함수 변환 • 프라임 시간 • 상품군별 프라임 시간 		코로나 반영 <ul style="list-style-type: none"> • 서비스업 생산지수 • 마스크 검색량 	계절 특성 반영 <ul style="list-style-type: none"> • 전국 단위 최고, 최저 기온 • 상대온도
		경제적 측면 반영 <ul style="list-style-type: none"> • 생활물가지수 	
판매 상품 세분화 <ul style="list-style-type: none"> • 상품군별 가격 구분 • 상품 구성 		계절 특성 반영 <ul style="list-style-type: none"> • 최고, 최저 기온 • 강수확률 	

변수 생성 및 탐색

예측의 정확도를 높이기 위해 제공 데이터 내에서 파생변수를 생성하였습니다.

내부변수

구분	생성변수 명	변수 설명
방송일시 세분화	B_date	방송 일자의 요일
	holiday_nums	연휴 중 해당 일의 일차
	holiday_duration	연휴가 지속된 총 일수
	Shour	방송시간의 sin함수 변환 값
	prime_hour	평일/ 주말 총매출 비중 50%를 차지하는 시간대 - 평일과 주말을 구분하여 생성
	Items_prime_hour	상품군 별로 프라임 시간대 설정
판매상품 세분화	price_category	상품군별 가격대를 4개(저가, 중가, 고가, 초고가)로 구분
	Item_nums	상품 구성 (n종)

변수 생성 및 탐색

방송 일시와 관련하여 방송 요일과 휴일 및 연휴길이를 추가하였고, 연속적이고 반복적인 시간의 특성을 반영하기 위해 sin함수로 변환한 방송 시간 변수를 생성하였습니다.

파생 변수 생성

방송일시 관련 변수

방송 요일

요일별로 매출 양상이 다를 것으로 판단하여 추가하였음

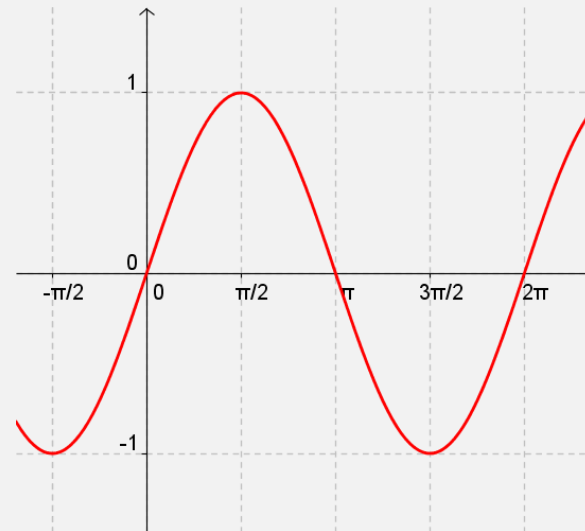
휴일

연휴 길이 별로 일차별 매출 양상이 다를 것으로 판단하여 추가하였음

- holiday_nums : 연휴 중 해당 일의 일차
- holiday_duration : 연휴가 지속된 총 일수

방송 시간

방송 시간의 주기성과 연속성을 반영하기 위해 sin함수로 변환하여 추가하였음



변수 생성 및 탐색

방송요일 별 취급액의 합과 평균을 살펴보니 평일보다는 주말에 취급액이 높았으며, 토요일보다는 일요일에 더 많은 매출을 올리는 것을 확인하였습니다.

파생 변수 탐색

방송일시 관련 변수

1e11	방송요일 별 취급액 합	방송요일 별 취급액 평균
------	--------------	---------------

정보 보안상의 문제로 공개하지 않습니다.

변수 생성 및 탐색

연휴 길이와 일차 별 취급액 추이를 살펴보면 연휴의 길이, 연휴 일차 모두 1일에서 가장 높은 취급액을 보이는 것을 확인하였습니다. 연휴 첫째날 홈쇼핑 소비가 가장 많고 지속되는 연휴가 아닌 단일 휴일일 때 가장 많은 취급액을 나타냈습니다.

파생 변수 탐색

방송일시 관련 변수

연휴 길이 별 취급액 합

1e11

정보 보안상의 문제로
공개하지 않습니다.

연휴 일차 별 취급액 합

1e11

정보 보안상의 문제로
공개하지 않습니다.

연휴길이 별 취급액 평균

1e

정보 보안상의 문제로
공개하지 않습니다.

연휴 일차 별 취급액 평균

1e7

정보 보안상의 문제로
공개하지 않습니다.

정보 보안상의 문제로
공개하지 않습니다.

변수 생성 및 탐색

상품의 구성이 취급액에 영향을 줄 수 있다고 생각하여 상품 명에서 구성을 추출하여 구성 컬럼을 별도로 생성하였습니다.

파생 변수 생성

상품 구성

“상품 구성에 따라 매출이 달라질 것이다.”

정규표현식을 이용하여 상품 구성 추출(1차)

```
for i in range (len(df)):
    target = df["상품명"][i]
    pt = '[0-9]+[가-힣]'
    df["구성"][i] = str(re.findall(pt, target))

for i in range (len(df)):
    m = df["상품명"][i]
    p = '\d+[+]\d+'
    df["구성1"][i] = str(re.findall(p, m))
```

n+m 구성, n매 m개 등 확인 후 구성 지정

```
[("15종4종", 19), ('7종7종', 14), ("3종20봉", 20),
 ("30매9박", 9), ("8박8주", 8), ("3종4팩2팩2팩8팩", 16),
 ("100개70개30개", 100), ("90개70개20개", 90),
 ("2019년1세16조4조", 20), ("2019년8조", 8), ("7세14종", 7),
 ("7종1차", 7), ("106차2세", 2), ("4종6월", 4),
 ("6종7월", 6), ("4종2종1종1종", 4), ("400매2개", 400),
 ("200매1개", 200), ("10종3차", 10), ("30봉1봉31봉", 31),
 ("10종7월", 10), ("8종7월", 8), ("2종8월", 2),
 ("5종5종", 5), ("4종8월", 4), ("10종8월", 10),
 ("130개80개50개", 130), ("12박12개", 12), ("60개5개", 65),
```

변수 생성 및 탐색

상품 구성 수량에 따른 취급액을 확인하였으나, 수량과 취급액의 관계가 뚜렷한 상관을 나타내지는 않는 것을 확인하였습니다.

파생 변수 탐색

상품 구성

상품구성과 매출액 관계

1e7

⋮
|
⋮

정보 보안상의 문제로 공개하지 않습니다.

0

탐색적 자료 분석

취급액 상위 50%를 차지하는 시간대를 평일/ 주말로 구분하여 프라임시간대를 설정하였고, 카테고리별로도 별도의 프라임시간을 확인하여 적용하였습니다.

파생 변수 생성

프라임 시간

“매출이 높은 시간대는 주중, 주말, 상품군별로 다를 것이다.”

구분별 매출의 상위 약 50%를
차지하는 시간대

시간대	매출
정보 보안상의 문제로 공개하지 않습니다.	
...	...

프라임
시간
(prime_
hour)

기본 프라임 시간

: 평일과 주말의 프라임 시간을 구분하여 지정

평일

주말

상품군별 프라임 시간(item_prime_hour)

정보 보안상의 문제로 공개하지
않습니다.

2-2. 데이터 전처리 및 EDA 변수 생성 및 탐색

평일 프라임 시간과 주말의 프라임 시간은 거의 유사하였으나, 주말은 오전 8시가, 평일은 오전 11시가 포함되었습니다.

파생 변수 탐색

프라임 시간

1e7

평일 시간대별 취급액

정보 보안상의 문제로 공개하지 않습니다.

평일 프라임 시간

평일 매출액 상위 50%에 해당하는
시간대

정보 보안상의 문제로
공개하지 않습니다.

주말 프라임 시간

주말 매출액 상위 50%에 해당하는
시간대

정보 보안상의 문제로
공개하지 않습니다.

hour

변수 생성 및 탐색

평일, 주말의 구분에서는 프라임 시간의 차이가 크지 않았지만 카테고리별 구분은 카테고리별로 프라임 시간이 다르게 도출되었습니다.

파생 변수 탐색

프라임 시간

카테고리

프라임 시간

정보 보안상의 문제로 공개하지
않습니다.

정보 보안상의 문제로 공개하지
않습니다.

변수 생성 및 탐색

1년 간의 판매 양상은 예측하고자 하는 취급액의 분포에 영향을 미칠 것으로 가정하여 상품군별, 월별, 시간대별, 계절별 취급액의 양상을 확인하였으며, 단가기준으로 판매양상을 보기 위해 주문 수량에 대한 분포를 확인하여 새로운 변수를 생성하였습니다.

파생 변수 생성

판매 양상

“1년 간 판매 양상은 예측 취급액의 분포에 영향을 미칠 것이다”

상품군별 판매 양상 (7)

상품군별 취급액 최대값, 최소값, 평균, 중위수, 25%, 75%, 분산

월별 판매 양상 (4)

월별 취급액 최대값, 최소값, 평균, 중위수

시간대별 판매 양상 (4)

시간별 취급액 최대값, 최소값, 평균, 중위수

계절별 판매 양상 (7)

계절별 상품군별 취급액 최대값, 최소값, 평균, 중위수, 25%, 75%, 분산

1년간 **상품 판매 양상을 가장 잘 반영하는 데이터**를 변수로 생성

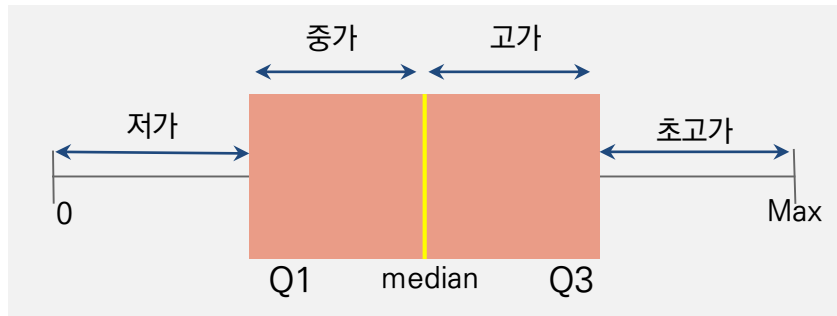
변수 생성 및 탐색

동일 카테고리 내에서도 단가의 차이가 크고 그에 따른 구매 패턴, 취급액이 다른 경향을 보이기 때문에 카테고리 세분화를 진행하였습니다.

파생 변수 생성

price_category

“가격대 별로 매출의 추이는 다를 것이다”



- 11개 카테고리 가격대별 구분
 - 0원- 1분위수 : 저가
 - 1분위수 - 중위수 : 중가
 - 중위수 - 3분위수 : 고가
 - 3분위수 - 최대값 : 초고가
- 총 44개 카테고리로 세분화

정보 보안상의 문제로 공개하지 않습니다.

2-2. 데이터 전처리 및 EDA

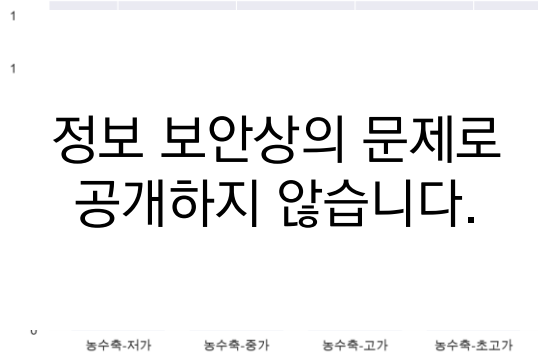
변수 생성 및 탐색

최고 주문량을 나타내는 가격대가 상품군 별로 다를 수 있음을 확인하였습니다. 공통적인 경향으로는 저가상품의 주문량이 가장 많음을 확인하였습니다.

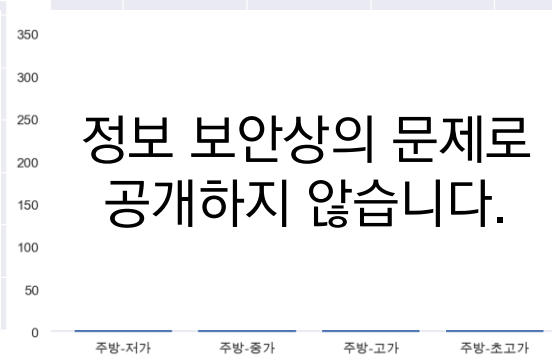
파생 변수 탐색

price_category

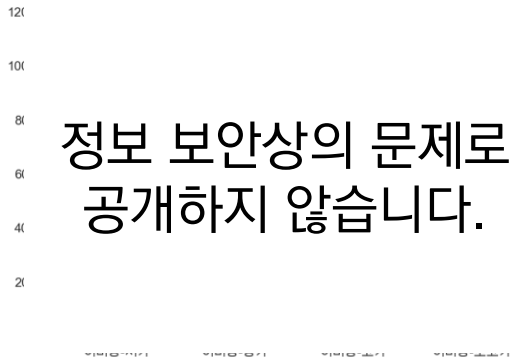
농수축 가격대별 주문량



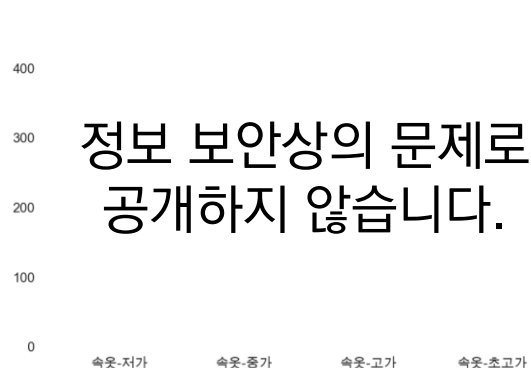
주방 가격대별 주문량



이미용 가격대별 주문량



속옷 가격대별 주문량








정보 보안상의 문제로
공개하지 않습니다.

변수 생성 및 탐색

내부 변수와 함께 외부변수를 추가함으로써 매출예측의 정확도를 높이하고자 하였습니다.

외부변수

구분	외부변수 명	변수 설명	출처
코로나 영향 반영	service_index	• 서비스업 생산 지수	 - 나라지표
	mask	• 마스크 검색량	
경제적 측면 반영	life_costing_index	• 생활물가지수	 - 나라지표
계절 특성 반영	day_MinTemp day_MaxTemp	• 도시별 일평균 최고, 최저기온 예보 정보를 • 지역별 인구수에 비례하여 계산	 기상청 기상자료개방포털
	Rain	• 도시별 강수 확률을 지역별 인구수에 비례하여 계산	 기상청 기상자료개방포털
	Relative_temp	• 월별 평균온도 기준으로 상대온도 변수 생성 - 봄/여름 : 일 최고기온 - 월 평균 최고기온 - 가을/겨울: 월 평균 최저기온 - 일 최저기온	 기상청 기상자료개방포털
	tv_rating	• 지상파 top20 시청률 데이터	

2-2. 데이터 전처리 및 EDA 변수 생성 및 탐색

코로나 19 이전과 이후의 사회, 경제적 요인을 반영하기 위하여 다양한 외부변수를 추가하였습니다.

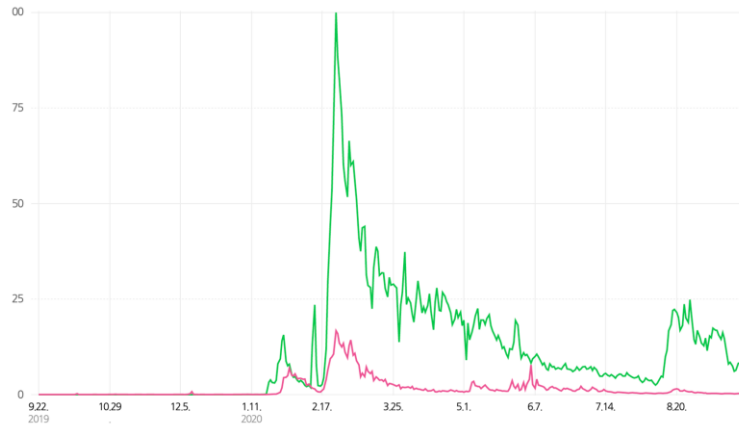
외부 변수 추가

코로나 영향 반영

사회요인

마스크 검색량

- 마스크 검색량은 코로나 확진자 추이와 유사한 등락을 보임
- 2019년의 경우, 미세먼지의 영향으로 마스크 검색량이 존재했기 때문에 2020년 코로나의 영향을 반영하기에 적합한 변수로 고려함



서비스업 생산 지수

- 서비스업의 성장세를 가늠하는 핵심지표로, 서비스업 전체 및 개별업종의 생산 활동을 종합적으로 파악하기 위한 지수
- 취급액과 가장 상관관계가 높은 서비스업 생산지수의 계절요인 지수를 변수로 활용함

신규 확진자 수	1	-0.45	-0.46	-0.63
서비스업 생산지수 (경상)	-0.45	1	1	0.66
서비스업 생산지수 (불변)	-0.46	1	1	0.66
서비스업 생산지수 (계절)	-0.63	0.66	0.66	1
신규 확진자 수	서비스업 생산지수 (경상)	서비스업 생산지수 (불변)	서비스업 생산지수 (계절)	

변수 생성 및 탐색

코로나 19 이전과 이후의 사회, 경제적 요인을 반영하기 위하여 다양한 외부변수를 추가하였습니다. 코로나 19 전후의 경제상황을 반영하기 위해 다양한 경제지표를 살펴보고, 취급액과 상관이 가장 높았던 생활물가지수를 경제관련 지표로 반영하였습니다.

외부 변수 추가

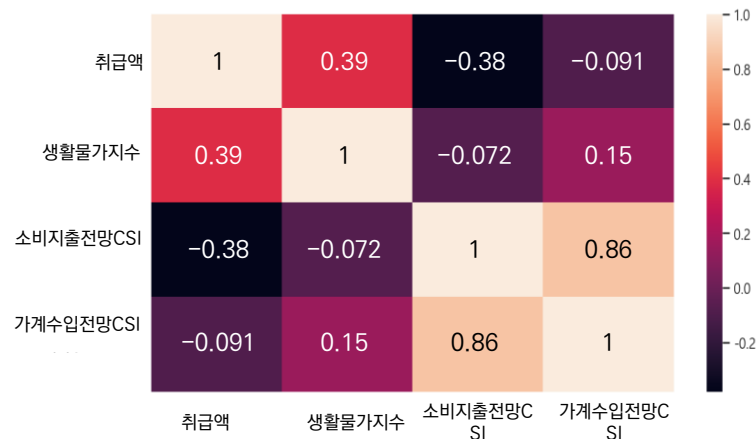
경제적 측면 반영

경제요인

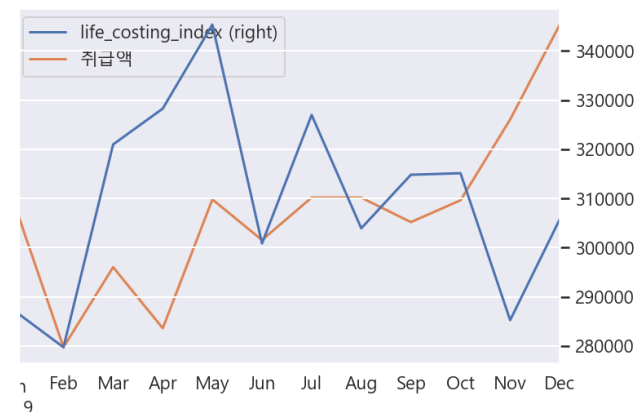
생활물가 지수

- 일상생활에서 소비자들이 자주 많이 구입하는 생활필수품을 대상으로 작성된 소비자물가지수의 보조지표
- 경제요인을 반영하는 많은 지수들 중, 취급액과의 상관관계가 가장 높은 생활물가지수를 변수로 활용함

취급액과 경제지표 상관관계



월별 취급액과 생활물가 지수



변수 생성 및 탐색

집에 오래 머물수록 매출이 늘어날 것이라는 가정으로 예보데이터와 지상파 TV 시청률 변수를 추가하였으며, 상대온도 변수를 생성해서 집에 머무르게 될 가능성을 수치로 표현하고자 하였습니다.

외부 변수 추가

날씨요인 & 시청률

날씨요인

예보데이터 추가

- 기온, 강수확률을 예보 데이터로 반영
: 실제 현업에서 활용 시, 예측일의 날씨데이터는 예보데이터만 수집이 가능하기 때문
- 지역 인구에 비례하여 평균 기온, 강수확률로 반영

상대온도 변수 생성

- 월 평균기온과의 차이로 상대온도 변수 생성
- 봄/여름 : 월 평균 온도보다 일 최고 온도가 높을 때 집에 머무를 것
- 가을/겨울: 월 평균 온도보다 일 최저 온도가 낮을 때 집에 머무를 것

날씨 요인에 따른 재택율

집에 머무는 시간 증가



TV시청 확률 증가



홈쇼핑 시청 및 구매 확률 증가

추가 변수

일별 지상파TV 시청률 상위 20개 프로그램 시청률 합

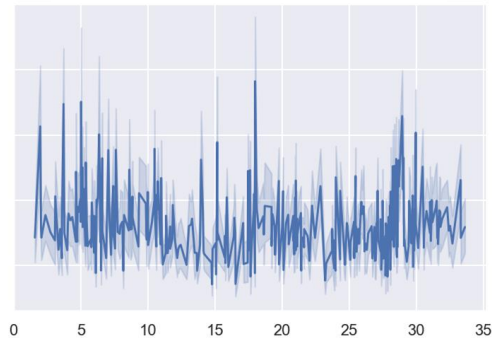
변수 생성 및 탐색

외부변수로 날씨요인과 시청률의 주문량과의 상관관계를 확인하고자 하였으나, 눈에 띄는 직접적인 상관관계는 확인하지 못하였습니다.

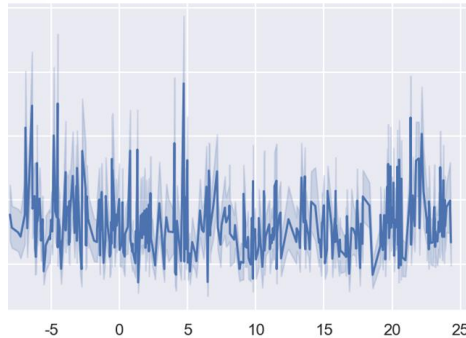
외부 변수 탐색

날씨요인 & 시청률

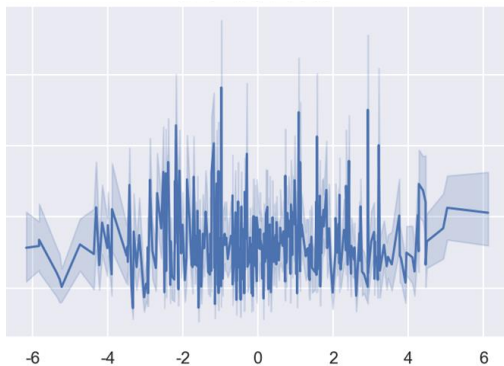
최고 기온과 주문량



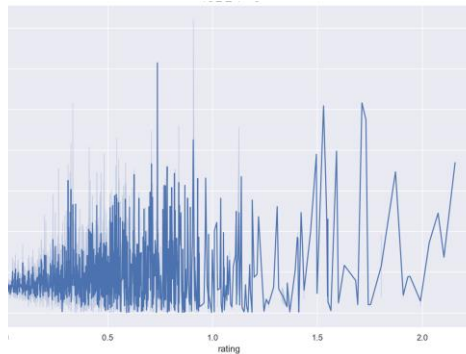
최저 기온과 주문량



상대 기온과 주문량



시청률과 주문량



기온과 주문량

- 최고, 최저 온도와 상품 주문량은 뚜렷한 경향을 나타내지는 않았음
- 시청자가 재택에 머무르는 정도를 확인하고자 생성한 상대온도 또한 주문량과 큰 상관을 나타내지 못하였음

시청률과 주문량

- 시청률과 주문량의 관계를 함께 살펴보았을 때, 직접적인 상관관계를 확인하지는 못하였음

02

프로젝트 세부내용

2-1

분석 프로세스

2-2

데이터 전처리 및 EDA

2-3

데이터 분석

2-4

예측 및 성능개선

학습알고리즘 선정 및 분석

상품 편성의 반복 주기를 활용하여 19년 6월 데이터를 기준으로 모델을 생성하였으며, 학습데이터와 예측 데이터의 비연속성을 효과적으로 반영하기 위해 시계열 모델이 아닌 모델을 선정하여 예측에 활용하였습니다.

매출 예측

예측 목표

- ns 홈쇼핑 20년 6월 매출 예측
- 제공 데이터 : 19년 12개월 데이터

상품 편성 반복 주기 활용 예측



- ▶ 월별 평균 매출액
 - 월별, 계절별 매출액의 평균값이 다름
 - 월 단위의 주기성은 보이지 않음

계절에 민감한 홈쇼핑 산업의 특성상, 1년 단위로 상품 주기성을 발견할 수 있을 것

⇒ 20년 6월과 가장 유사한 편성을 가질 것으로 예상되는 **19년 6월을 기준으로 모델 생성**

학습 데이터와 예측 데이터의 비연속성 반영

학습데이터

2019년 1월 ~ 12월

예측데이터

2020년 6월

▶ 2020년 1월 ~ 5월 데이터 공백

시계열 데이터 분석 방법으로 접근 시, 예측값 이전 5개월 데이터의 공백을 예측값으로 채워서 분석을 진행하게 됨

⇒ **학습/예측 데이터 간 비연속성**을 반영할 수 있는 모델 선정

학습알고리즘 선정 및 분석

학습알고리즘은 모델의 측면에서는 예측 정확도와 학습 시간을 고려하였고, 데이터의 측면에서는 보유한 데이터의 비선형성을 효과적으로 반영할 수 있는 모델로 선정하였습니다.

매출 예측

모델 활용

모델 선정 기준

- 모델측면 : 높은 예측 정확도, 짧은 학습시간
- 데이터 측면 : 데이터의 비선형성을 효과적으로 반영

양상블 계열 알고리즘

- Light GBM
 - 그라디언트 부스팅 프레임 워크.
 - 결정 트리 알고리즘을 기반으로 하며 **순위 지정**, 분류 및 기타 머신러닝에 활용
 - **많은 양의 데이터에 적합**하며 속도가 빠름
- Catboost
 - 기존 부스팅 모델의 느린 학습 속도와 overfitting문제를 해결
 - **범주형 변수 처리 속도가 빠름**

딥러닝 알고리즘

- DNN(Deep Neural Network)
 - 은닉층이 2개 이상인 학습방법
 - **복잡한 비선형 관계 모델링에 활용**
 - 출력층의 활성화함수를 Linear로 설정해 수치 예측
 - 데이터가 루프백 없이 입력에서 출력으로 흐르는 피드포워드 네트워크

학습알고리즘 선정 및 분석

데이터의 오차 값을 낮추기 위하여 데이터 셋을 5개 그룹으로 분리하여 분석을 진행하였으며, 각 카테고리에 적합한 변수선정과 하이퍼 파라미터 튜닝을 거쳐 1차 예측값을 도출하였습니다.

매출 예측

모델 활용

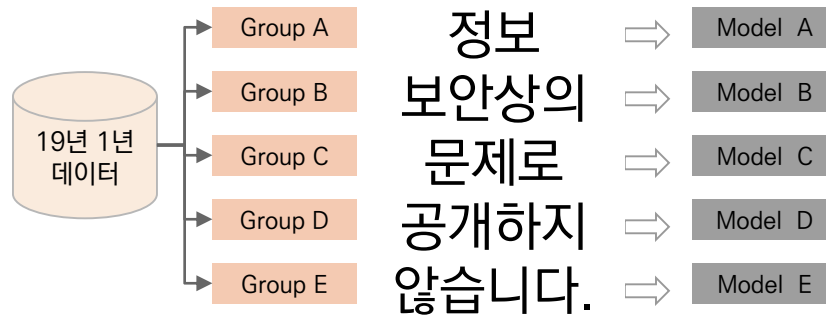
1차 모델링

기계학습모델

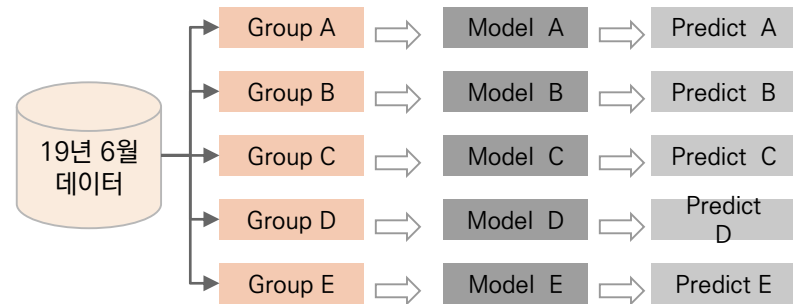
- 데이터의 비선형성을 반영할 수 있으며, 데이터 셋 당 MAPE 값을 가장 낮게 하는 모델을 선정해 1차적으로 매출을 예측한다.

1. 데이터셋 분리 및 모델 선정

2. 모델 통한 1차 매출 예측(19년 6월)



주간 매출 양상이 비슷한 상품군들을 묶어 5개의 데이터셋을 생성한 후, 그룹별 mape 가 가장 낮은 모델을 선정한다.



생성한 모델에 19년 6월 데이터를 대입해 19년 6월 매출을 예측한다.

학습알고리즘 선정 및 분석

시간대별 매출액 평균값과 세부 상품군별 매출액 평균값을 적정 가중치로 반영하여 실제값과 가장 가까운 값을 도출하도록 세부카테고리별로 다른 가중치를 도출하였습니다.

매출 예측

모델 활용

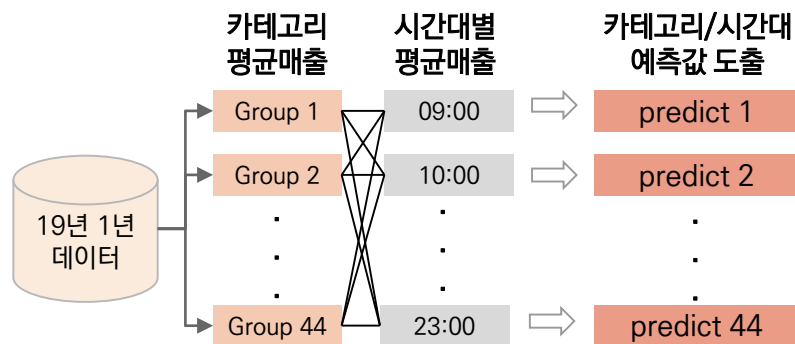
2차 모델링

통계적 예측

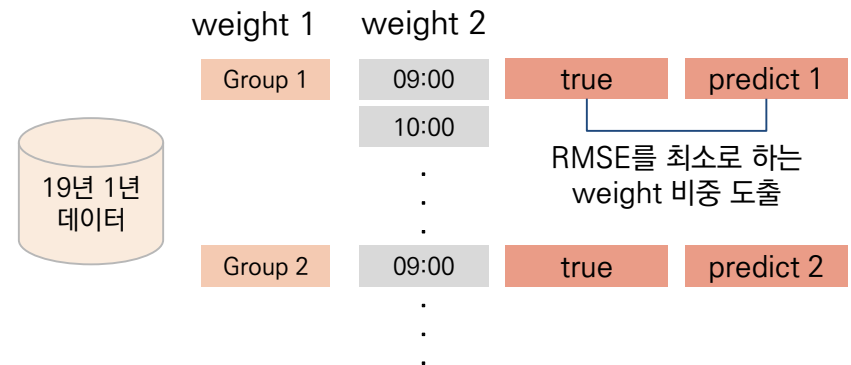
- RMSE값을 가장 낮게 하는 가중치를 선정하여 통계 모델을 생성한다
- 예측 모델과 통계 모델에 가중치를 부여하여 예측&통계 모델을 생성한다

3. 통계 모델 생성

4. 통계적 접근 최적 가중치 설정



시간대별 매출액 평균값과 상품군 평균값을 반영하여 통계적 모델을 생성하고 한 번도 판매된적 없던 시간대*상품군 조합에 예측값 추가



실제 값과 예측값의 오차가 최소가 되는 상품군과 시간대의 가중치 설정

학습알고리즘 선정 및 분석

학습알고리즘은 모델의 측면에서는 예측 정확도와 학습 시간을 고려하였고, 데이터의 측면에서는 보유한 데이터의 비선형성을 효과적으로 반영할 수 있는 모델로 선정하였습니다.

매출 예측

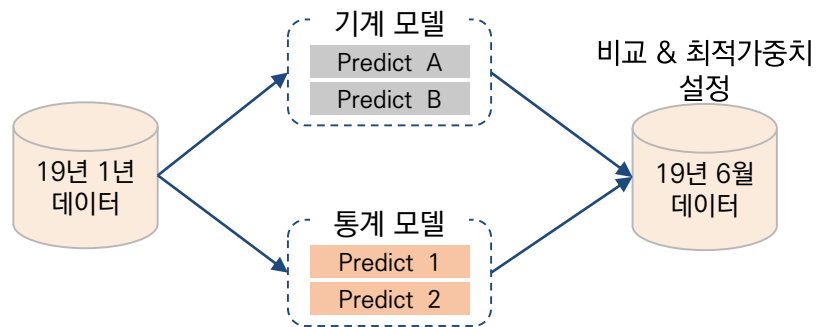
모델 활용

3차 모델링

기계 & 통계모델

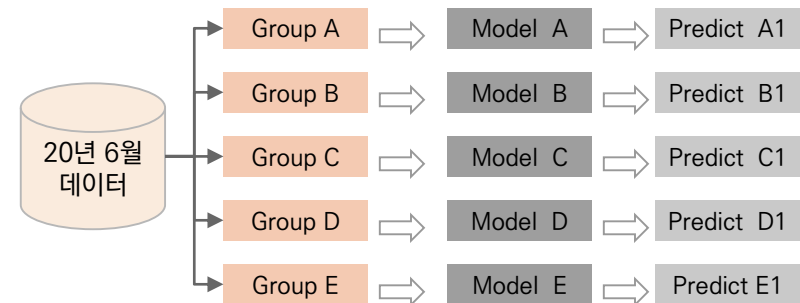
- RMSE값을 가장 낮게 하는 가중치를 예측&통계 모델을 통한 예측값과 모델 예측값에 부여하여 19년 6월 매출액을 예측한다

5. 기계&통계 모델을 통한 2차 매출 예측(19년 6월)



19년 1년 데이터로 구한 통계적 예측값과 모델 예측값이 19년 6월의 데이터와 가장 가까워지는 가중치 도출

6. 기계&통계 모델을 통한 20년 6월 매출 예측



19년 6월 편성표에 최적화된 가중치를 기준으로 20년 6월 매출 예측

학습알고리즘 선정 및 분석

통계적 예측의 정확도를 높이기 위해 카테고리별 평균 취급액과 시간대별 평균취급액의 가중치를 카테고리별로 다르게 적용하였습니다.

매출 예측

통계적 예측

실제 값과 예측값의 RMSE를 최저로 만드는 가중치 조합 도출

$$Sales = w_1 * SalesPerCaterogy + w_2 * SalesPerHour$$

* SalesPerCategory : 카테고리별 평균 취급액 , * SalesPerHour: 시간대별 평균 취급액 (147개 시간, 21h*7일)

구분	category1	...	44개 카테고리	sales/ hour
06:00				
07:00				
...				
147개 시간대 (21h * 7일)				
sales/category				

판매기록이 있는
Category와 시간대 조합의 **실제값**



오차 최소화

0부터 1까지 0.1 단위로 w_1 , w_2 를
조정해서 얻은 **예측값**

학습알고리즘 선정 및 분석

예측한 매출을 바탕으로 1일 최대 매출 달성을 위한 방송 시간대별 최적의 카테고리를 헝가리안 알고리즘을 활용하여 할당하였습니다.

편성 방안 도출

편성 알고리즘

활용 알고리즘

헝가리안 알고리즘

알고리즘 특징

할당문제 최적화에 활용하는 알고리즘

- 문제 해결을 위한 최소 비용을 구하는 알고리즘
- 본 프로젝트에서는 예측 데이터를 기반으로 최대값을 산출

편성단위

1주일 주간 편성 (147개 시간대별 추천 상품군 제시)

도출 방법

- 주간 편성과 일간 편성 동시에 진행
- 주간, 일간 편성에서 시간대별 중복 상품 우선 배치
- 중복되지 않는 시간대의 경우, 높은 예상매출액의 상품 우선 배치

02

프로젝트 세부내용

2-1

분석 프로세스

2-2

데이터 전처리 및 EDA

2-3

데이터 분석

2-4

예측 및 성능개선

최종 예측 결과

모델을 활용한 예측에서는 앙상블계열 모델과 딥러닝 모델을 최종비교 하였고, 앙상블계열의 CatBoost Regressor가 가장 좋은 성능을 보였습니다. 2020년 6월의 예측 정확도와 오차는 구할 수 없어 2019년 6월의 예측치로 정확도와 오차를 확인하였습니다.

최종 결과

그룹별 모델 예측

모델 예측 성능 비교

- 앙상블 모델과 딥러닝 모델을 비교함
- 다섯 가지 모델 중 앙상블 모델인 catboost와 lgbm 성능이 가장 좋았음

기계 모델 예측 성능

model_name	R-squared	MAPE
XGBRegressor	0.61	68.80
LGBMRegressor	0.68	57.88
CatBoostRegressor	0.69	53.52
DNN	0.49	91

모델 자체의 성능으로 확인하였을 때, CatBoostRegressor의 R^2 값이 0.69로 가장 높았으며, MAPE값 또한 53.52로 가장 낮게 분석됨

기계 모델 19년 6월 예측

model_name	R-squared	MAPE
XGBRegressor	-	-
LGBMRegressor	0.73	80.22
CatBoostRegressor		
DNN	-	-

2019년 6월 데이터의 예측치와 실제 값을 비교하였을 때 기계모델을 혼합하여 사용한 모델에서는 R^2 가 0.73, MAPE가 80.22로 분석됨

최종 예측 결과

모델을 활용한 예측에서는 앙상블계열 모델과 딥러닝 모델을 최종비교 하였고, 앙상블계열의 CatBoost Regressor가 가장 좋은 성능을 보였습니다. 2020년 6월의 예측 정확도와 오차는 구할 수 없어 2019년 6월의 예측치로 정확도와 오차를 확인하였습니다.

최종 결과

그룹별 모델 예측

모델 예측 성능 비교

- 최종 모델은 grid search cv로 하이퍼 파라미터 튜닝 진행

하이퍼파라미터 튜닝

파라미터 \ 모델	LGBM Regressor	Catboost Regressor
n_estimators	1000	1000
learning_rate	0.01	0.01
depth	6	6

최종 예측 결과

기계학습 모델만으로는 정확한 예측이 어렵다고 판단하여 통계적 예측기를 함께 사용하여 최종 예측을 진행하였습니다.

최종 결과

모델 예측 합계

모델 예측
성능 비교

- 통계적 접근으로 2019년 6월의 매출을 예측함
- 기계학습과 통계적 예측기를 혼합하여 2019년 6월의 매출을 예측한 결과

통계적 예측 결과

model_name	R-squared	MAPE
통계적 예측기	0.51	85.47

통계적 예측기로 2019년 6월의 예측값을 도출하였고,
실제값과 비교하였을 때, R^2 값은 0.51, MAPE값은 85.47로
기계학습 모델보다는 다소 성능이 떨어짐

기계&통계 결과

model_name	R-squared	MAPE
기계 & 통계 예측기	0.25	52.89

기계학습과 통계적 예측기를 함께 사용한 모델의 결과는
 R^2 값이 0.25, MAPE값은 52.89로 도출되었음

02. 프로젝트 세부내용

최종 예측 결과

2020년 6월의 데이터를 다음과 같이 예측하였습니다.

최종 결과

6월 매출 예측

연월일시	노출(회)	이력코드	상품코드	상품명	상품군	판매가격	금액(원)
2020-06-01 6:20	20	100650	201971	책필드 남성 반팔셔츠 4종	의류	59,800	
2020-06-01 6:40	20	100650	201971	책필드 남성 반팔셔츠 4종	의류	59,800	
2020-06-01 7:00	20	100650	201971	책필드 남성 반팔셔츠 4종	의류	59,800	
2020-06-01 7:20	20	100445	202278	쿠미투니카 쿨 레이스 라쥬쉐이퍼&팬티	속옷	69,900	
2020-06-01 7:40	20	100445	202278	쿠미투니카 쿨 레이스 라쥬쉐이퍼&팬티	속옷	69,900	
2020-06-01 8:00	20	100445	202278	쿠미투니카 쿨 레이스 라쥬쉐이퍼&팬티	속옷	69,900	
2020-06-01 8:20	20	100381	201247	바비리스 퍼펙트 볼볼스타일러	이미용	59,000	
2020-06-01 8:40	20	100381	201247	바비리스 퍼펙트 볼볼스타일러	이미용	59,000	
2020-06-01 9:00	20	100381	201247	바비리스 퍼펙트 볼볼스타일러	이미용	59,000	
2020-06-01 9:20	20	100638	201956	벨레즈온 심리스 원피스 4종 패키지	속옷	59,900	
2020-06-01 9:40	20	100638	201956	벨레즈온 심리스 원피스 4종 패키지	속옷	59,900	
2020-06-01 10:00	20	100638	201956	벨레즈온 심리스 원피스 4종 패키지	속옷	59,900	
2020-06-01 10:20	20	100348	201091	벨레즈온 심리스 원피스 4종 패키지	속옷	59,900	
2020-06-01 10:40	20	100348	201091	벨레즈온 심리스 원피스 4종 패키지	속옷	59,900	
2020-06-01 11:00	20	100348	201091	벨레즈온 심리스 원피스 4종 패키지	속옷	59,900	
2020-06-01 11:20	20	100012	200016	AAC 삼채포기김치 10kg	농수축	40,900	
2020-06-01 11:40	20	100012	200016	AAC 삼채포기김치 10kg	농수축	40,900	
2020-06-01 12:00	20	100012	200016	AAC 삼채포기김치 10kg	농수축	40,900	
2020-06-01 12:20	20	100080	200217	아키 라이크라 릴렉스 보정브라 패키지(뉴아키28차)	속옷	99,900	
2020-06-01 12:40	20	100080	200217	아키 라이크라 릴렉스 보정브라 패키지(뉴아키28차)	속옷	99,900	
2020-06-01 13:00	20	100080	200217	아키 라이크라 릴렉스 보정브라 패키지(뉴아키28차)	속옷	99,900	
2020-06-01 15:20	20	100362	201150	에이윌러스 슈퍼선스틱 1004(최저가)	이미용	39,900	

정보 보안상의 문제로 공개하지 않습니다.

세부내용은 prediction 컬럼을 참고해주세요.

최종 예측 결과

2020년 6월 1주차의 방송을 다음과 같이 편성하였습니다. 편성은 주간, 일간 편성을 모두 진행하여 중복 상품을 우선 배치하며, 중복되지 않을 경우 예상매출이 높은 상품을 할당하였습니다.

최종 결과

편성 제안

주간 편성 결과

day	hour	category	sales
monday	0	low_price	
	1	mid_price_l	
	2	extra_high_p	
	6	extra_high_p	
	7	low_price_l	
	8	high_price_livi	
	9	extra_high_pri	
	10	extra_high_p	
	11	low_price	
	12	high_price	
	13	low_price_l	
	14	high_price_l	
	15	mid_price	
	16	low_price	
	17	mid_price	
	18	mid_price	
	19	low_price_livi	
	20	mid_price	
	21	high_price	
	22	low_price	
	23	extra_high_p	

주별 편성표

일간 편성 결과

day	hour	category	sales
monday	0	low_price	
	1	mid_price_l	
	2	low_price_h	
	6	low_price	
	7	mid_price	
	8	high_price_livi	
	9	extra_high_pri	
	10	extra_high_p	
	11	high_price_l	
	12	high_price	
	13	low_price_h	
	14	extra_high_pri	
	15	mid_price	
	16	mid_price	
	17	mid_price	
	18	high_price_l	
	19	low_price_livi	
	20	mid_price	
	21	high_price	
	22	low_price	
	23	extra_high_p	

요일별 편성표

최종 편성 안

day	hour	category	sales
monday	0	low_price	
	1	mid_price_l	
	2	low_price_h	
	6	low_price	
	7	mid_price	
	8	high_price_livi	
	9	extra_high_pri	
	10	extra_high_p	
	11	high_price_l	
	12	high_price	
	13	low_price_h	
	14	extra_high_pri	
	15	mid_price	
	16	mid_price	
	17	mid_price	
	18	high_price_l	
	19	low_price_livi	
	20	mid_price	
	21	high_price	
	22	low_price	
	23	extra_high_p	

최종 편성표

: 주간, 일간 중복 카테고리

A still life composition featuring a variety of fresh produce and kitchen items. In the top left, a dark bowl is filled with strawberries and blackberries. Below it, a halved blood orange reveals its red segments. To the right of the orange, several small yellow cherry tomatoes are scattered. In the center, a bunch of vibrant red chard stalks lies horizontally. Below the chard, two orange tulips with green leaves are positioned. At the bottom left, a small dark bowl contains red peppercorns. Next to it, a pair of scissors with dark red handles is visible. Several small, square photographic prints are scattered throughout the arrangement. The entire scene is set against a plain white background.

감사합니다.