

2020
빅콘테스트

NS Shop+ 판매실적 예측을 통한 편성 최적화 방안(모형) 도출

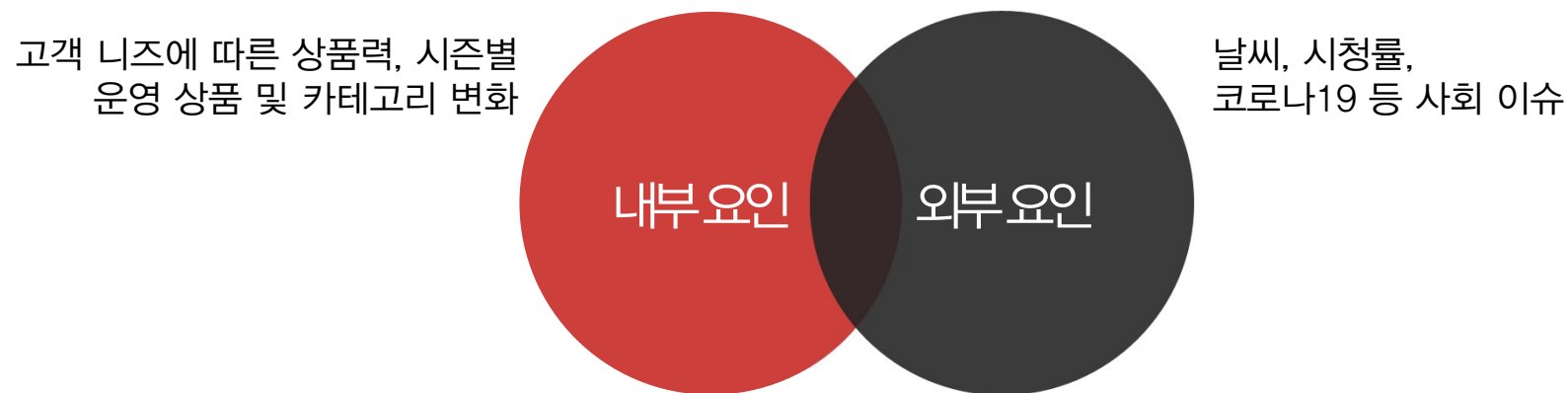
공정의씨앗 팀

이재원 lan4148@naver.com
박세진 sjsj9368@naver.com
최우진 scojin777@naver.com
최호진 patagonistt@gmail.com
황인범 rydn2004@naver.com

목차

1. Data Imputation
2. Feature Engineering
3. Model Set & Predict
4. Optimized Television Schedules

NS Shop+ 판매실적 예측을 통한 편성 최적화 방안(모형) 도출



내부요인과 외부요인을 고려해 '20년 6월 프로그램 매출 실적을 예측하고,
최적 수익을 고려한 요일별 / 시간대별 / 카테고리별 편성 최적화 방안(모형) 제시

The background of the slide consists of a large red parallelogram on the left and a dark gray triangle on the right, both meeting at a diagonal line.

1. Data Imputation

Data Imputation

방송일시	노출(분)	마더코드	상품코드	상품명
2019-01-01 6:00	20	100346	201072	테이트 남성 셀린니트3종
2019-01-01 6:00		100346	201079	테이트 여성 셀린니트3종
2019-01-01 6:20	20	100346	201072	테이트 남성 셀린니트3종
2019-01-01 6:20		100346	201079	테이트 여성 셀린니트3종
2019-01-01 6:40	20	100346	201072	테이트 남성 셀린니트3종
2019-01-01 6:40		100346	201079	테이트 여성 셀린니트3종

➡ 하나의 방송 편성에서 판매하는 상품이 2개 이상일 때, 한 종류의 상품을 제외한 모든 상품에서 '노출(분)'의 결측치 존재



동일한 방송 편성의 '노출(분)'으로 대체

상품명	상품군	판매단가	취급액
일시불 LG 울트라HD TV 70UK7400KNA	가전	2,700,000	50,000
무이자 LG 울트라HD TV 70UK7400KNA	가전	2,990,000	50,000
일시불 LG 울트라HD TV 55UK6800HNC	가전	1,300,000	3,177,000
무이자 LG 울트라HD TV 55UK6800HNC	가전	1,440,000	3,541,000
일시불 LG 울트라HD TV 65UK6800HNC	가전	1,900,000	32,123,000
무이자 LG 울트라HD TV 65UK6800HNC	가전	2,130,000	5,189,000
일시불 LG 울트라HD TV 70UK7400KNA	가전	2,700,000	50,000
무이자 LG 울트라HD TV 70UK7400KNA	가전	2,990,000	50,000



취급액이 0인데, 50,000원으로 잘못 입력된 경우



예측 대상에서 제외

Data Imputation

매주 토요일 18:00~18:20 은
정보방송시간으로 추정 제외

방송일시	노출(분)	마더코드	상품코드	상품명
2019-01-05 17:40	20	100435	201350	우리바다 손질왕꼬막 24팩
2019-01-05 18:20	20	100801	202365	바다먹자 국내산 손질꽃게 7팩
2019-01-05 18:40	20	100801	202365	바다먹자 국내산 손질꽃게 7팩
2019-01-05 19:00	20	100448	202098	일시불 쿠첸 풀스텐 압력밥솥 10인용 (A1)
2019-01-05 19:00		100448	202093	무이자 쿠첸 풀스텐 압력밥솥 10인용(A1)

The background of the slide features a large red parallelogram on the left and a dark gray triangle on the right, both meeting at a diagonal line.

2. Feature Engineering

활용 데이터 정의



NS SHOP+ 실적 데이터

- Train : NS SHOP+ 프로그램 실적 데이터 (2019.1~12월)
- Test : NS SHOP+ 프로그램 실적 데이터 (2020.6월)
- 변수 8개 : 방송일시 / 노출(분) / 마더코드 / 상품코드 / 상품명 / 상품군 / 판매단가 / 취급액
- 레코드 38,309 개



기상청 데이터

- 기상청 공공데이터에서 가져온 외부데이터
- 지역별/시간대별 기온, 강수량, 풍속, 습도, 적설, 전운량 데이터를 가지고 시간대별 각각의 평균 수치형 변수 생성
- 평균 강수량과 평균 적설을 통해 비/눈 여부의 범주형 변수 생성



미세먼지 데이터

- 에어코리아에서 가져온 외부데이터
- 일자별 미세먼지와 초미세먼지 데이터를 가지고 수치형 변수 생성



시간 관련 파생변수 생성

월, 일, 시간hour

원데이터 방송일시를 월, 일,
24시간 기준으로 나눠 변수 생성

168시간

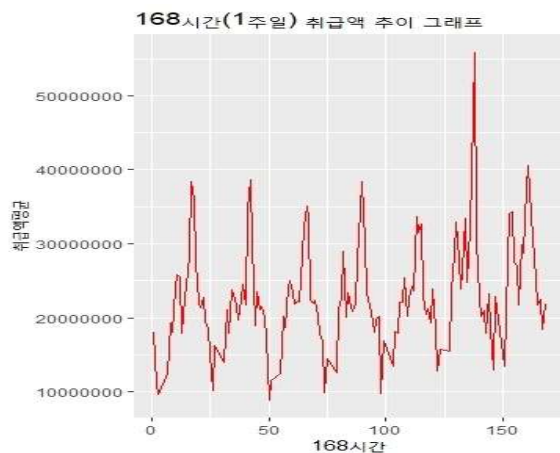
일주일 단위로 $24 * 7 = 168$
시간으로 나눠 변수 생성

방송내_순서, 방송순서

방송순서 변수와, 같은 상품을
판매하더라도 방송 순서가
있는 방송내 순서 변수 생성

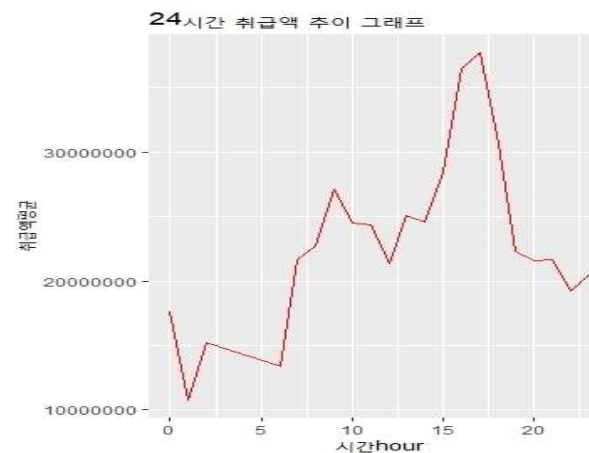


시간 관련 파생변수 생성



168시간_COS, 168시간_SIN

168시간으로 취급액 추이를 그려본 결과, 주기성을 보임. 따라서 주기함수인 COS, SIN 함수를 이용해 데이터 변환



24시간_COS, 24시간_SIN

24시간으로 취급액 추이를 그려본 결과, 주기성을 보임. 따라서 주기함수인 COS, SIN 함수를 이용해 데이터 변환

그룹코드 관련 파생변수 생성

그룹코드_전체횟수

그룹코드별로 전체횟수를 세는 변수 생성

그룹코드_대박횟수

그룹코드별로 취급액의 상위 15%에 포함되는 횟수를 세는 변수 생성

그룹코드_대박확률

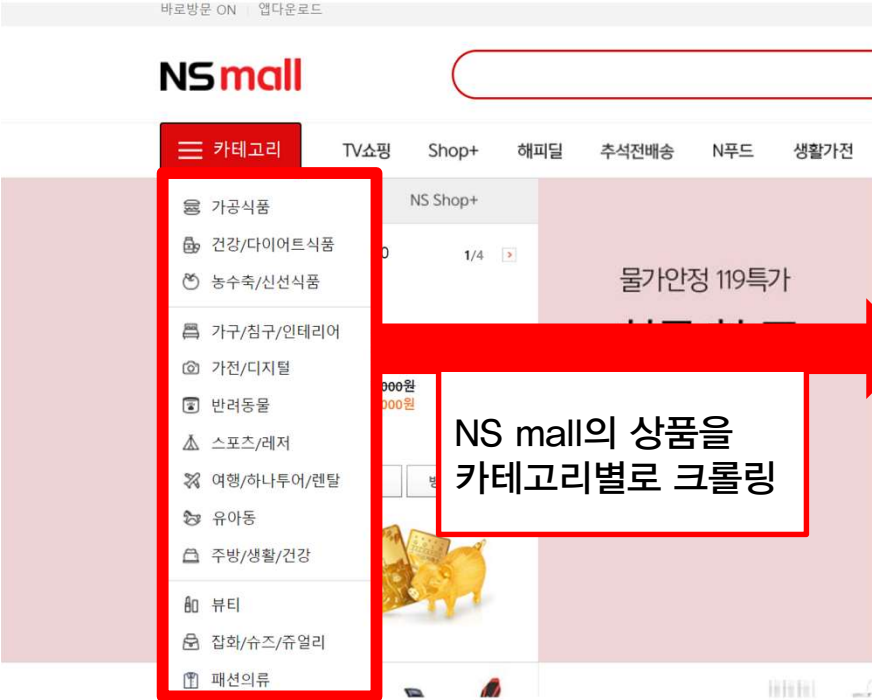
그룹코드_대박횟수를 그룹코드_전체횟수로 나눈 변수 생성

그룹코드_쪽박횟수

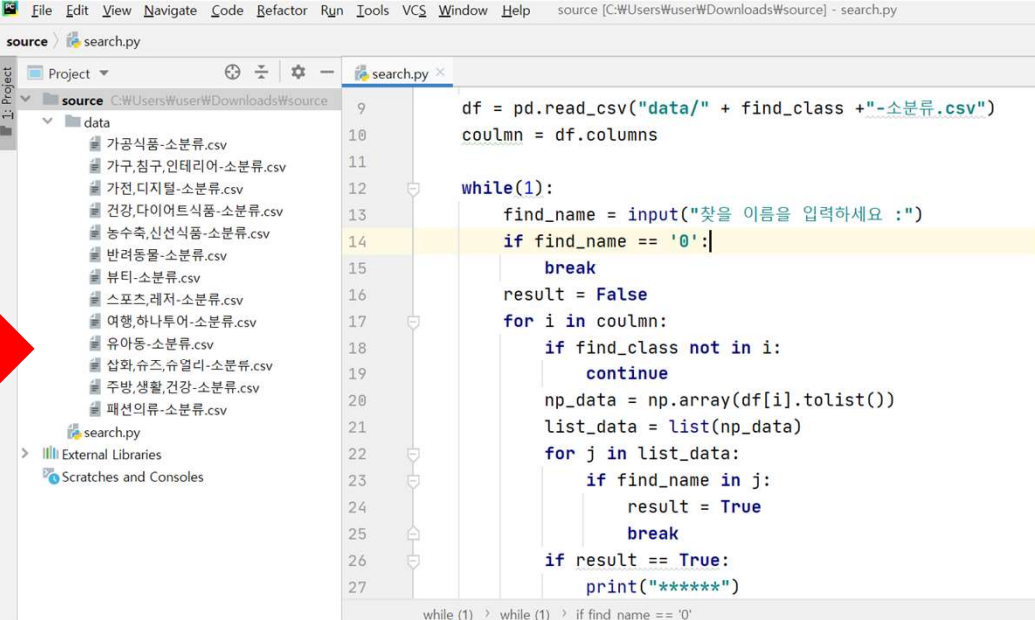
그룹코드별로 취급액의 하위 20%에 포함되는 횟수를 세는 변수 생성

그룹코드_쪽박확률

그룹코드_쪽박횟수를 그룹코드_전체횟수로 나눈 변수 생성



NS mall의 상품을
카테고리별로 크롤링



```

source > search.py
Project
└── source
    ├── data
    │   ├── 가공식품-소분류.csv
    │   ├── 가구,침구,인테리어-소분류.csv
    │   ├── 가전,디지털-소분류.csv
    │   ├── 건강,다이어트식품-소분류.csv
    │   ├── 농수축,선선식품-소분류.csv
    │   ├── 반려동물-소분류.csv
    │   ├── 뷰티-소분류.csv
    │   ├── 스포츠,레저-소분류.csv
    │   ├── 여행,하나투어-소분류.csv
    │   ├── 유아동-소분류.csv
    │   ├── 삽화,슈즈,슈얼리-소분류.csv
    │   ├── 주방,생활,건강-소분류.csv
    │   └── 패션의류-소분류.csv
    ├── search.py
    └── External Libraries
        └── Scratches and Consoles

9      df = pd.read_csv("data/" + find_class + "-소분류.csv")
10
11      coulmn = df.columns
12
13      while(1):
14          find_name = input("찾을 이름을 입력하세요 :")
15          if find_name == '0':
16              break
17          result = False
18          for i in coulmn:
19              if find_class not in i:
20                  continue
21              np_data = np.array(df[i].tolist())
22              list_data = list(np_data)
23              for j in list_data:
24                  if find_name in j:
25                      result = True
26                      break
27              if result == True:
28                  print("*****")

```

while (1) > while (1) > if find_name == '0'

순위 관련 파생변수 생성

중분류순위, 소분류순위

크롤링한 카테고리별 상품들과 원데이터의 상품들을 매칭해 중분류, 소분류별 평균 취급액에 따른 순위형 변수 생성

Run: search

```

*****
*****
가공식품-당, 국, 찌개-당, 국
*****
찾을 이름을 입력하세요 : 0
찾으실 분류 이름을 입력하세요 : 농수축, 선선식품
찾을 이름을 입력하세요 : 손질문어
*****
농수축, 선선식품-수산-해산물
*****
찾을 이름을 입력하세요 : 0
찾으실 분류 이름을 입력하세요 : 뷰티
찾을 이름을 입력하세요 : 볼륨스타일러
*****
뷰티-헤어, 바디-헤어스타일링
*****

```



순위 관련 파생변수 생성

판매단가순위

판매단가별 평균 취급액을 비교해
각 판매단가가 얼마나 실적을
냈는지 평가하는 순위형 변수 생성

판매단가rank

각 판매단가가 가지고 있는
해당분위수로 순위형 변수 생성



기타 파생변수 생성

판매상품종류수

하나의 방송편성에서 몇 개의 상품을 판매하는지 확인하는 변수 생성

방송내_상품종류별_점수

그룹코드마다 평균 취급액을 구하고 분위수와 upper bound / lower bound 를 이용해 6구간으로 변수 생성

방송시간

편성된 '방송일시' 의 차이를 구해 정확한 방송시간을 예측

매진여부

생성한 방송시간 변수를 노출(분)과 비교해 방송시간보다 노출(분)이 짧은 경우는 매진이라고 판단해 변수 생성

결제수단

상품명에 결제수단이 명시되어 있는지 확인한 뒤, 명시돼 있다면 '무이자' , '일시불' 로 구분해 범주형 변수 생성

휴일

토요일, 일요일, 공휴일에 해당하는 '휴일' 범주형 변수를 생성. 단, 명절은 음력을 기준으로 지내 매년 날짜가 바뀌기 때문에 휴일변수에서 제외

시청률 관련 변수

NS SHOP+ 일자별, 시간대별 시청률 (2019년)					
시간대	2019-01-01	2019-01-02	2019-01-03	2019-01-04	2019-01-05
02:18	0	0	0	0.014	0
02:19	0	0	0	0.014	0
02:20	0	0	0	0	0
02:21	0	0	0	0	0
02:22	0	0	0	0	0
02:23	0	0	0	0	0
02:24	0	0	0	0	0

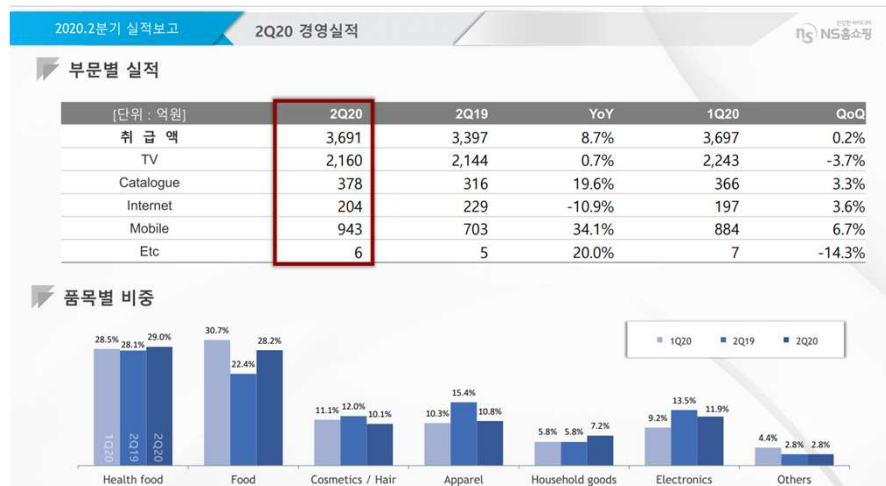
데이터에 결측치인 0 값과
반올림된 값이 다수 존재

시간대	2019-12-21	2019-12-29	2019-12-30	2019-12-31	9-01-01 to 2019-12
01:54	0	0	0	0	0.004
01:55	0	0	0	0	0.004
01:56	0	0	0	0	0.004
01:57	0	0	0	0	0.004
01:58	0	0	0.019	0	0.004
01:59	0	0	0	0	0.004
월화수목금토일02:00-01:59	0.003	0.004	0.005	0.005	0.004

마지막 행과 열에 있는 평균 시청률
값으로 0 값을 대체하려 했으나,
그 값들이 반올림되면서 대체할 값으로
사용하기에 신뢰성이 없다고 판단

시청률 관련 변수 생성하지 않기로 결정

코로나 관련 변수



취급액 (억원)	19년 2Q	20년 2Q	증감량	증감율
Electronics	458.595	439.229	-19.366	-4.2%
Health Food	954.557	1070.39	115.833	12.1%
Food	760.928	1040.862	279.934	36.8%
Household goods	197.026	265.752	68.726	34.9%
Cosmetics/Hair	407.64	372.791	-34.849	-8.5%
Apparel	523.138	398.628	-124.51	-23.8%
Others	95.116	103.348	8.232	8.7%
총합계	3397	3691		

2020 2분기 NS 홈쇼핑 IR 자료에서
NS SHOP+ 취급액과 6월 취급액만
특정할 수 없다는 문제 발생

코로나19로 인한 취급액 영향을 알아보기로
IR 자료의 품목별 실적 비중을 전년 동기와 비교

코로나 관련 변수

*“ 회사는 사업 특성 상 COVID-19의 확산이 매출 및 수요 측면에 유의적인 영향을
미치지는 않을 것으로 판단하고 있으며,
실제로 당분(반)기매출액에 대한 COVID-19의 부정적 영향은 유의적이지 않은 것으로
파악되었습니다.”*

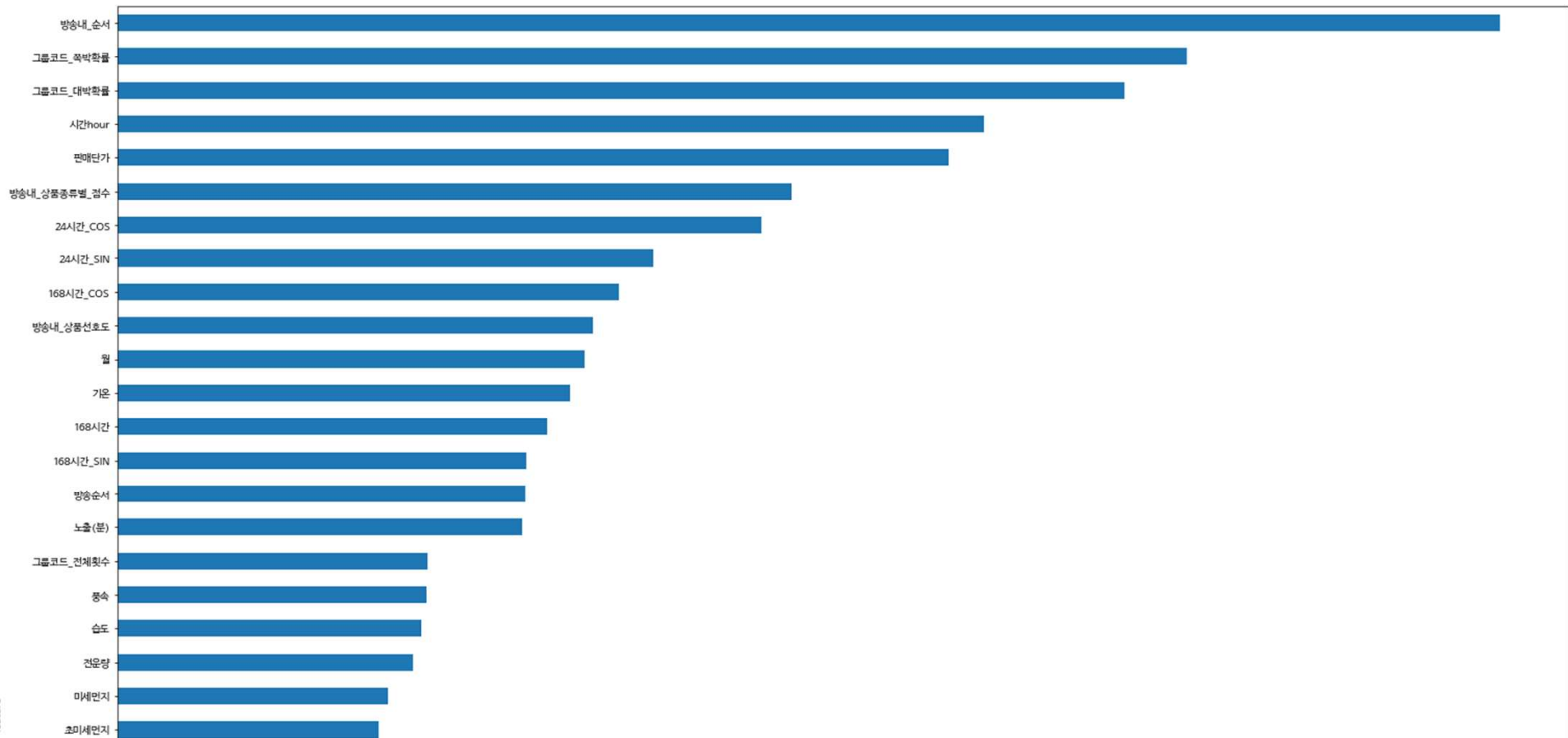
– 금융감독원 전자공시시스템 NS홈쇼핑 반(분)기 보고서 내용 중



최종적으로 코로나 변수 생성하지 않기로 결정

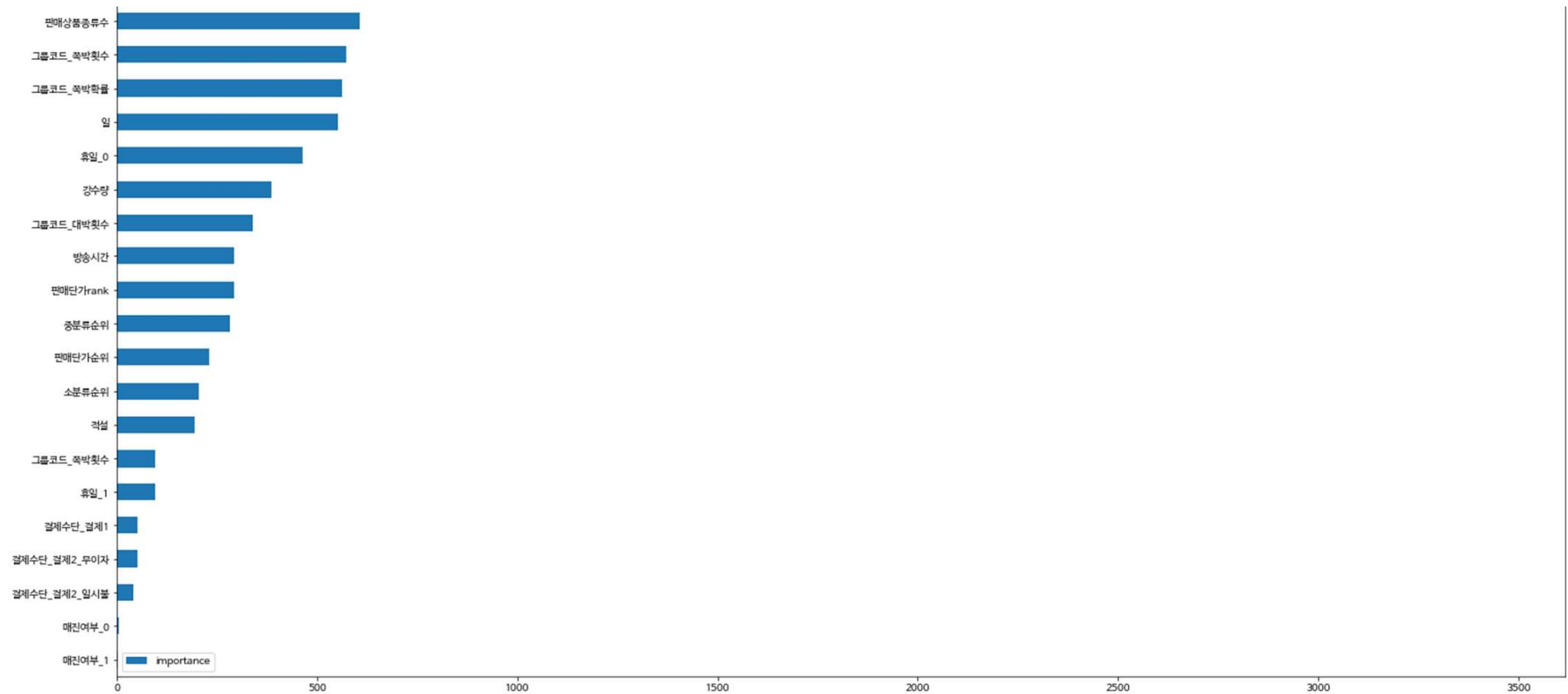
Feature Engineering

변수 중요도



Feature Engineering

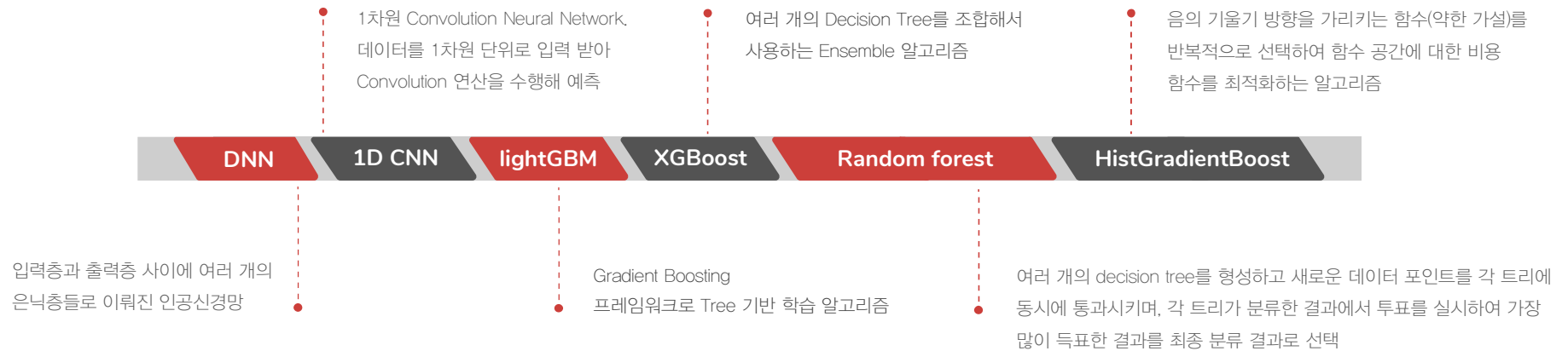
변수 중요도



The background of the slide features a large red parallelogram on the left and a dark gray triangle on the right, both meeting at a diagonal line.

3. Model Set & Predict

취급액 예측 Algorithm





DNN

Predict

- 복잡한 비선형관계 모델링
- 연속형/변수형 변수에 상관 없이 분석 가능
- Feature Extraction이 자동으로 수행

최종 Hyper Parameter

Layer 구조 : Dense Layer(32) – Dense Layer(64)
– Dense Layer(32) – Dense Layer(1)

Drop Out : 0.5

Activation Function : mish – mish – relu

Optimizer : RMSPROP

최종 MAPE : 36.60685898



1D CNN

Predict

- 짧은 구간에 대한 패턴 인식 우수
 - 각 레이어의 입출력 데이터 형상 유지
 - 필터를 공유 파라미터로 사용하므로 일반
- 인공신경망에 비해 학습 파라미터가 매우 적음

최종 Hyper Parameter

Layer 구조 : Convolution – Max Pooling
– Dense Layer(64) – Dense Layer(32) – Dense Layer(1)

Max-Pooling size : 2

Filter : 64

Kernel size : 2

최종 MAPE : 37.2788



lightGBM

Predict

최종 Hyper Parameter

- Tree가 수평적으로 확장
- leaf-wise 방식
- loss(손실)가 적음
- 적은 메모리를 차지

'learning_rate' : 0.1

'num_iterations' : 1200

'max_depth' : 8

'feature_fraction' : 0.7

최종 MAPE : 37.04277276



XGBoost

Predict

최종 Hyper Parameter

- 약한 예측 모형들을 결합해
강한 예측 모형을 만드는 알고리즘
- 무작위성이 없으며 강력한
사전 가지치기 사용
- 배깅과 다르게 순차적

'learning_rate' : 0.1

'num_iterations' : 1200

'max_depth' : 8

'feature_fraction' : 0.7

최종 MAPE : 37.2273



HistGradient Boost

- 비교적 낮은 훈련 및 추론 시간과 결합된 높은 품질을 제공
- 임의의 미분 손실 함수를 최적화하여 일반화

Predict

최종 Hyper Parameter

'learning_rate' : 0.1

'num_iterations' : 1200

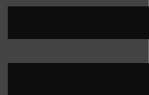
'max_depth' : 8

'feature_fraction' : 0.7

최종 MAPE : 37.0806



VOTING



DNN

Deep Learning



lightGBM

Machine Learning

대수의 법칙 : 모델이 앙상블에 포함된 개별 모델
중 가장 뛰어난 것보다 정확도가 높은 경우가 많음

모델간의 독립성 : 비슷한 모델들이 같은 종류의
오차를 만들기 쉬우므로 서로 독립성이 높은
DNN과 LightGBM 사용

예측 결과, 가장 성능이 좋은 위 2개의 모델을 VOTING,
예측값들을 통합하여 최종 예측값 도출.

최종 예측값 = 21040892.87 , 최종 MAPE = 36.17144



4. Optimized Television Schedules

편성 최적화 방안 도출 : 최적 수익을 고려한 방송일시에 따른 상품 소분류 추천 과정

STEP 1

특정 방송일자(ex. 2019년 1월 1일)에 대한 데이터를 Train으로 설정

STEP 3

상품별 취급액을 기준으로 내림차순 정렬

STEP 5

추출된 상품 데이터들의 소분류별로 개수를 세어 가장 많은 빈도의 소분류를 선택

STEP 7

추천할 최종 소분류 결정

STEP 2

특정 방송일자의 방송시간대별로 그룹화해 상품별 취급액 합을 구함

STEP 4

4분위수로 데이터를 나눠 상위 25%의 상품들을 추출

*가장 큰 값을 선택하지 않은 이유는 하나의 상품군에서 이상치에 가까운 취급액이 나온 데이터를 취하는 것을 막기 위함

STEP 6

소분류의 개수가 같다면 소분류들의 취급액 평균을 구해 더 큰 소분류를 선택

Optimized Television Schedules

상품명	상품군	소분류	lgbm취급액	dnn취급액	평균취급액	기존 취급액과의 차이
헤스떼벨 원 업속옷	보정언더우		2,734,501	12,008,066	7,371,283	5,272,283.41
헤스떼벨 원 업속옷	보정언더우		5,420,365	16,727,188	11,073,776	6,702,776.27
히트용 극세사 의류	잠옷,이지용		4,843,139	26,078,100	15,460,620	12,198,619.71
히트용 극세사 의류	잠옷,이지용		9,084,517	14,430,469	11,757,493	4,802,492.85
히트용 극세사 의류	잠옷,이지용		7,671,401	24,582,228	16,126,814	9,454,814.28
히트용 극세사 의류	잠옷,이지용		13,898,215	28,960,406	21,429,310	12,092,310.38
히트용 극세사 의류	잠옷,이지용		13,017,999	7,279,565	10,148,782	3,329,781.53
코치 엠마 사철잡화	캐주얼가방		21,938,219	13,358,487	17,648,353	1,959,353.20
코치 엠마 사철잡화	캐주얼가방		31,893,833	15,779,014	23,836,423	1,533,576.60
프라다 투웨이'잡화	캐주얼가방		22,419,720	25,605,294	24,012,507	7,879,506.76
일시불 쿠키전: 주방	냄비,압력기		41,662,308	47,235,052	44,448,680	14,387,679.83
일시불 쿠키전: 주방	냄비,압력기		57,337,332	60,262,120	58,799,726	5,257,726.02
일시불 쿠키전: 주방	냄비,압력기		15,979,824	18,942,940	17,461,382	1,386,382.20
프라다 투웨이'잡화	캐주얼가방		29,470,160	27,965,034	28,717,597	3,089,597.07
프라다 투웨이'잡화	캐주얼가방		38,229,747	35,803,848	37,016,797	12,253,202.71

추천 결과

24h

하루 기준

86,279,746

취급액 차이

67개

해당 날짜에 대한 프로그램 개수



감사합니다