2020 Big Contest

길치

김용환 (k01110111@nate.com)

조윤영 (double_y22@naver.com)

Contents

00. 분석 목적

4차 산업혁명 양방향 서비스 T-Commerce

* 출처 : 메일경제 T커머스 덕분에…"매출이 2배 늘었어요

- TV 홈쇼핑보다 진화된, TV를 통해 발생하는 고객 경험 중심의 양방향 상거래 서비스
- 정보성 : 쇼호스트의 자세한 설명
- 유용성 : 시공간 제약 X

T-Commerce 방문 경위 및 요인별 중요도

	습관적	목적성	즉흥적
T커머스	0.274	0.360	0.365
홈쇼핑	0.221	0.071	0.708
인터넷쇼핑	0.308	0.481	0.211

	상품	서비스	정보	신뢰
T커머스	0.336	0.201	0.253	0.210
홈쇼핑	0.312	0.292	0.096	0.300
인터넷쇼핑	0.395	0.175	0.276	0.154

* T 커머스 이용 및 구매 결정요인 분석 T커머스, TV 홈쇼핑, 인터넷쇼핑 비교를 중심으로, 2017, 박지은

- T 커머스는 홈쇼핑과 인터넷 쇼핑 그 중간에 존재
- 능동적으로 상품을 찾아보는 목적성, 그리고 방송하는 상품 즉흥성을 함께 가지고 있음

고객이 채널을 찾았을 때 알맞은 상품을 추천하고 방송하는 것

딥러닝 자동화 편성 시스템 구축

00. 분석 목적

내부 요인
- 고객 니즈 상품력
- 시즌별 상품 변화

외부 요인
- 시청률
- 날씨
- 사회적 이슈

빅데이터 분석 및 딥러닝으로 매출데이터 학습!

자동 편성과 상품추천!

시청률

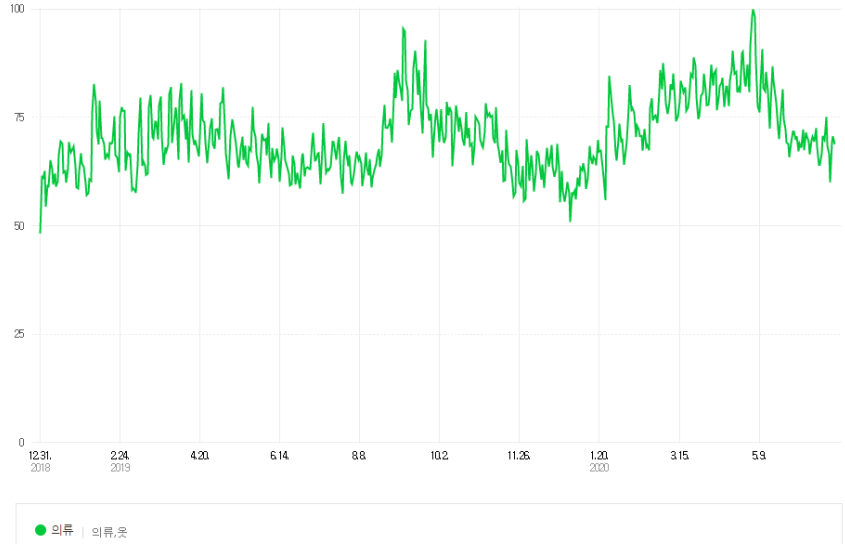
시간별 시청률

- 상품 컨텐츠와 상관관계
- 날씨

사회적 이슈

네이버 상품군별 통합검색량

소비자가 가장 관심있는 제품을 의미



* 출처 네이버랩 검색어 트렌드

전처리 및 EDA

파생변수 생성

00. 분석 순서

변수생성 | 분석시행 | 결과

STEP0 Data Handling → **STEP1** Rating 예측 → **STEP2** 취급액 예측 모델 → **STEP3** Ensemble → **STEP4** 최종 예측

전처리 및 파생변수 생성

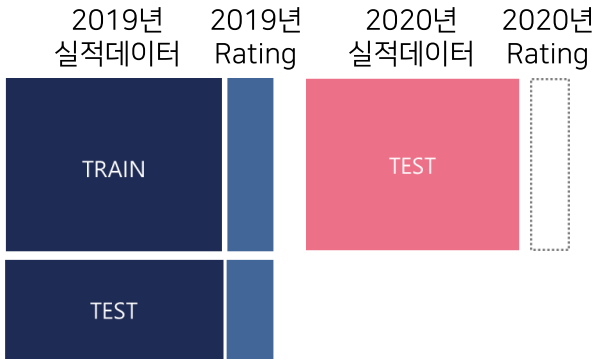
- 월, 일, 시간 변수
- 요일 변수 생성
- 공휴일 여부 변수
- Rating 변수 생성

- 방송순서 변수(number)
- Charge = 일시불, 무이자
- Brand
- Gender
- Package
- 날씨 변수
- 네이버 상품군 별 검색량

Rating 추정

- 2019년 실적데이터를 이용하여 Rating을 추정하는 모델을 만들어 2020년의 시청률을 예측

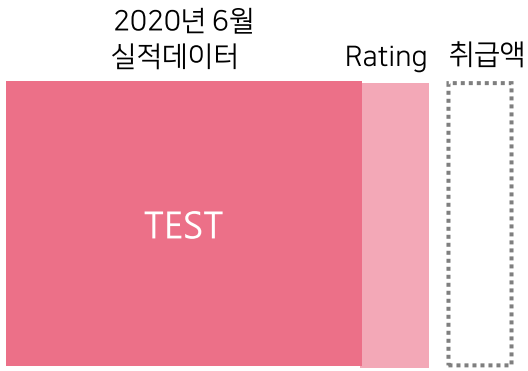
- LGBM, CatBoost, DNN 앙상블



취급액 예측

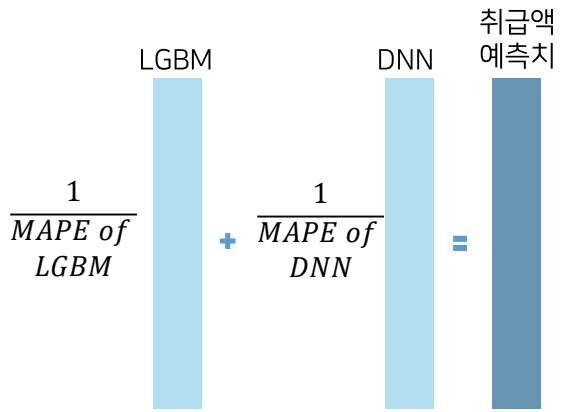
- 예측된 Rating을 변수로 추가하여 2020년 6월의 취급액을 예측

- LGBM, DNN 앙상블



Ensemble

- 각 모델의 test MAPE의 역수를 Weight로 하여 취급액 예측치를 Ensemble



* 2019데이터와 2020 데이터의 상품코드, 마더코드, 상품군 매칭 진행

01. 전처리 & 파생변수

상품군 무형 삭제

- 상품군이 '무형'인 데이터 삭제

Package 변수 생성

- 세트 상품여부

노출(분) NA 처리

- 함께 방송된 상품의 노출(분)으로 대체

방송일시 변수

- 월, 일 ,시간 변수
- 요일 변수 생성
- 공휴일 여부 변수

시청률 데이터 – Ratings 변수 생성

- 시청률 데이터를 방송일시, 노출(분) 단위에 맞춰 구간안의 max 시청률을 대입

	방송일시	마더코드	상품코드	상품명	rating
1000	2019-01-12 08:40:00	100293	200949	[가이거] 제니스시계 주얼리세트	0.011
1001	2019-01-12 09:00:00	100808	202377	CERINI by PAT 남성 소프트 기모 릴렉스팬츠	0.004
1002	2019-01-12 09:20:00	100808	202377	CERINI by PAT 남성 소프트 기모 릴렉스팬츠	0.023
1003	2019-01-12 09:40:00	100808	202377	CERINI by PAT 남성 소프트 기모 릴렉스팬츠	0.023
1004	2019-01-12 10:00:00	100271	200896	헤스티지 엘레나 라쿤양가죽 콤비 롱코트	0.024

교호작용 항 변수 생성

- Brand & 마더코드
- Brand & 상품코드
- 상품군 & 판매단가
- 마더코드 & 판매단가
- 상품군 & Naver_{t}
- 상품군 & Naver_{t-1}

01. 전처리 & 파생변수

Charge 변수 생성

- 상품명에서 일시불, 무이자를 변수로 생성함

Gender 변수 생성

- 상품명에서 남성/여성의 상품을 뽑아 변수 생성

네이버 상품군 별 검색량 변수 추가

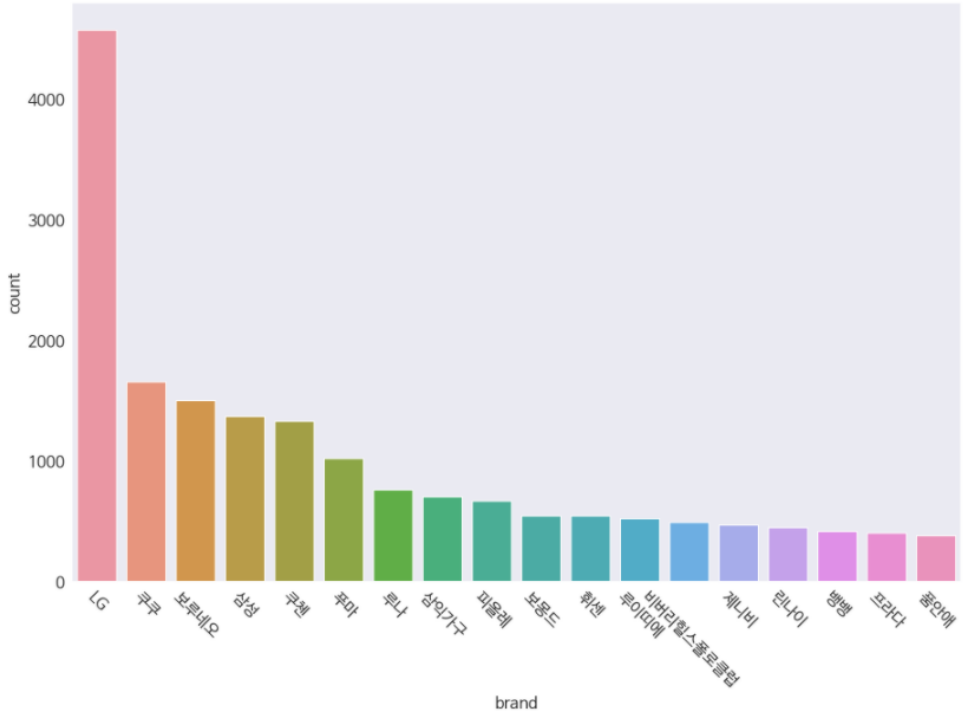
- 네이버 상품군 별 검색량을 당일, 전날 기준으로 추가

날씨 변수 추가

- 당일의 기온, 강수량, 풍속, 습도, 적설, 전운량을 변수로 추가

브랜드 변수 추가

- 상품명에서 브랜드 명을 변수로 생성 (상위 18개)



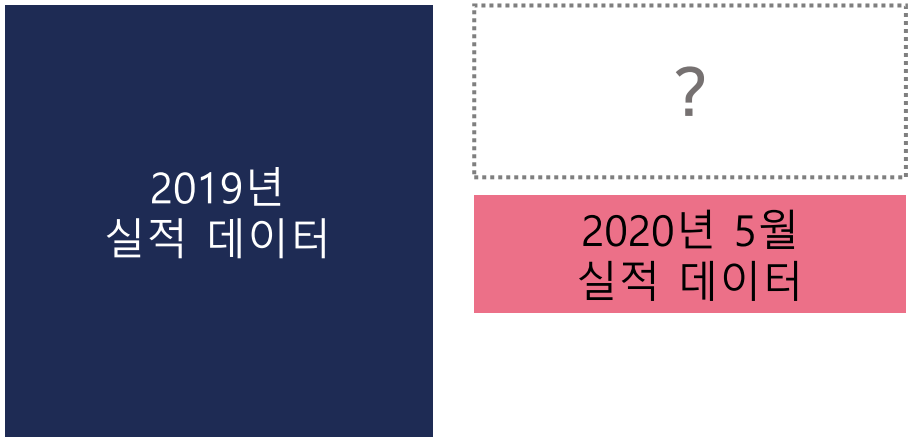
가전

01. 전처리 & 파생변수

Number 변수 생성 (시계열처리)

문제점

- 2020년 1월~5월까지의 데이터가 비어있기 때문에 Auto-Correalation을 추정하는 전통적인 시계열 분석은 불가능
- 2020년 6월의 첫 번째 자료에 대해 Jump prediction 후 Auto-corellaion을 이용한 Web Prediction 가능,
- but 추정 오차가 점점 커진다는 약점



위의 문제점을 보완하기 위해 넘버링 진행

01. 전처리 & 파생변수

Number 변수 생성 (시계열처리)

보완

① '방송일, 상품코드' 단위로 데이터를 분할

② 방송 시간을 순서대로 1, 2, 3…의 넘버를 지정하여 int로 넣어줌

③ 이를 통해 같은 상품코드, 같은 방송일 경우의 순서를 어느정도 보정하는 효과를 기대함

④ 만약 하루에 여러 번 같은 상품 방송을 했다면, 연속적인 단위로만 구성

	방송일시	노출(분)	마더코드	상품코드	상품명	판매단가	취급액	number
0	2019-01-01 06:00:00	20.0	100346	201072	테이트 남성 셀린니트3종	39900.0	2099000.0	1
1	2019-01-01 06:00:00	NaN	100346	201079	테이트 여성 셀린니트3종	39900.0	4371000.0	1
2	2019-01-01 06:20:00	20.0	100346	201072	테이트 남성 셀린니트3종	39900.0	3262000.0	2
3	2019-01-01 06:20:00	NaN	100346	201079	테이트 여성 셀린니트3종	39900.0	6955000.0	2
4	2019-01-01 06:40:00	20.0	100346	201072	테이트 남성 셀린니트3종	39900.0	6672000.0	3
5	2019-01-01 06:40:00	NaN	100346	201079	테이트 여성 셀린니트3종	39900.0	9337000.0	3

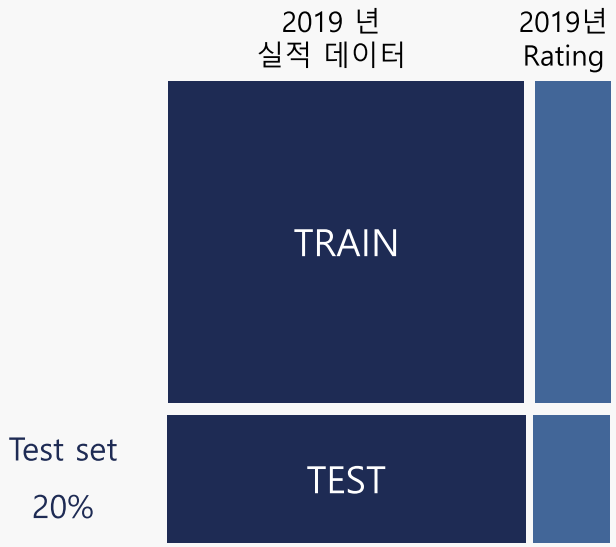
시청률 & 취급액 예측

LGBM, DNN, Cat Boost

02. 시청률 데이터

- 상품코드, 마더코드, 상품군, 월, 요일, 시간, number, charge, 브랜드, gender를 dummy변수화 하여 적용
- LGBM은 Category화 하여 이용
- DNN은 Normalize한 자료 이용

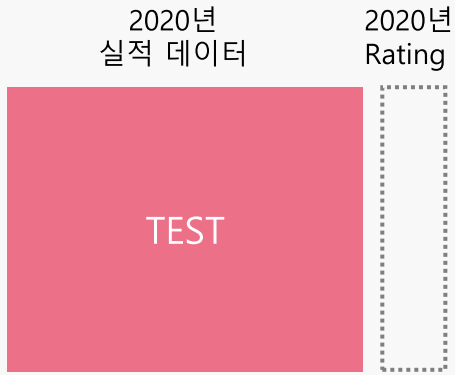
STEP 1



❶ Test set RMSE 기준

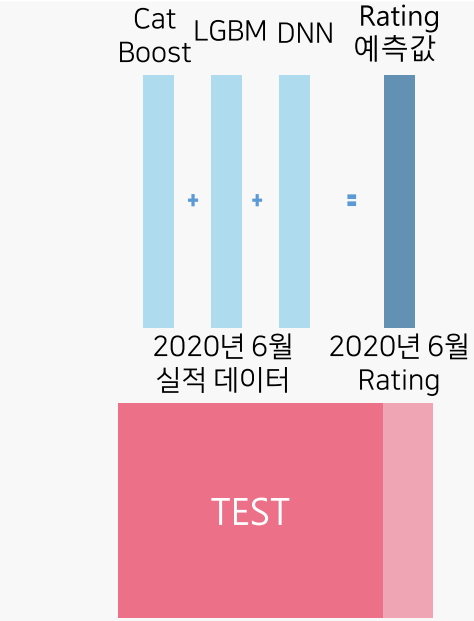
❷ LGBM, Catboost DNN 모델

STEP 2



❶ 앞서 만든 모델을 2020년 6월 데이터에 적용

STEP 3



❶ 3가지 모델의 결과를 RMSE의 역수로 weight두어 ensemble

02. 시청률 데이터

LGBM

```
{'learning_rate': 0.01,
   'max_depth': 32,
   'boosting': 'gbdt',
   'objective': 'regression',
   'metric': 'mse',
   'is_training_metric': True,
   'num_leaves': 128,
   'feature_fraction': 0.9,
   'bagging_fraction': 0.7,
   'bagging_freq': 5,
   'seed':2018}
```

CatBoost

```
(loss_function='RMSE',
 eval_metric = 'MAPE',
 depth=10,
 learning_rate=0.1,
 verbose=False,
 iterations=1000)
```

DNN

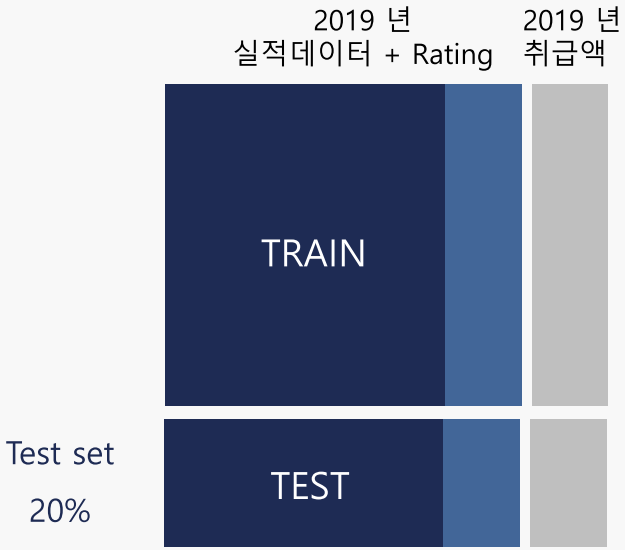
Layer (type)	Output Shape	Param #
dense_7 (Dense)	(None, 256)	409600
dropout_4 (Dropout)	(None, 256)	0
dense_8 (Dense)	(None, 256)	65792
dropout_5 (Dropout)	(None, 256)	0
dense_9 (Dense)	(None, 128)	32896
dropout_6 (Dropout)	(None, 128)	0
dense_10 (Dense)	(None, 64)	8256
dropout_7 (Dropout)	(None, 64)	0
dense_11 (Dense)	(None, 32)	2080
dense_12 (Dense)	(None, 16)	528
dense_13 (Dense)	(None, 1)	17

Total params: 519,169
Trainable params: 519,169
Non-trainable params: 0

02. 취급액 예측

- 앞서 예측한 Rating을 추가하여 모델링
- CatBoost 가 좋지 않은 성능을 보여 제외

STEP 1

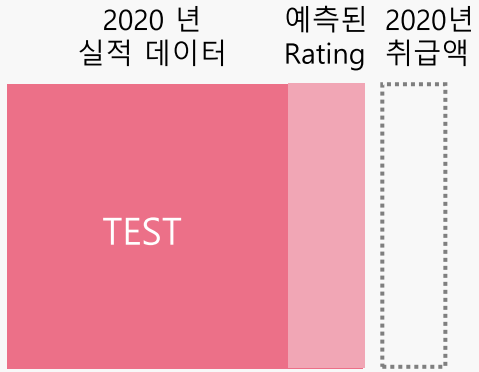


❶ Test set MAPE 기준

❷ Test에 예측된 Rating 사용

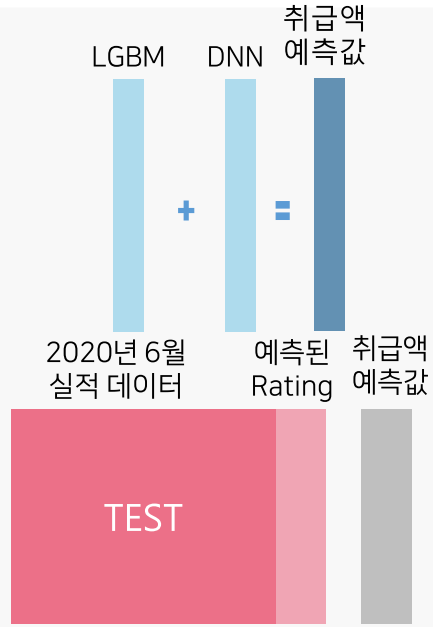
❸ DNN, LGBM

STEP 2



❶ 앞서 만든 모델을 2020년 6월 데이터에 적용

STEP 3



❶ 2가지 모델의 결과 값을 MAPE의 역수로 weight두어 ensemble

02. 취급액 예측

LGBM

```
{'learning_rate': 0.05,
   'max_depth': 32,
   'boosting': 'gbdt',
   'objective': 'regression',
   'metric': 'mse',
   'is_training_metric': True,
   'num_leaves': 128,
   'feature_fraction': 0.9,
   'bagging_fraction': 0.7,
   'bagging_freq': 5,
   'seed':2018}
```

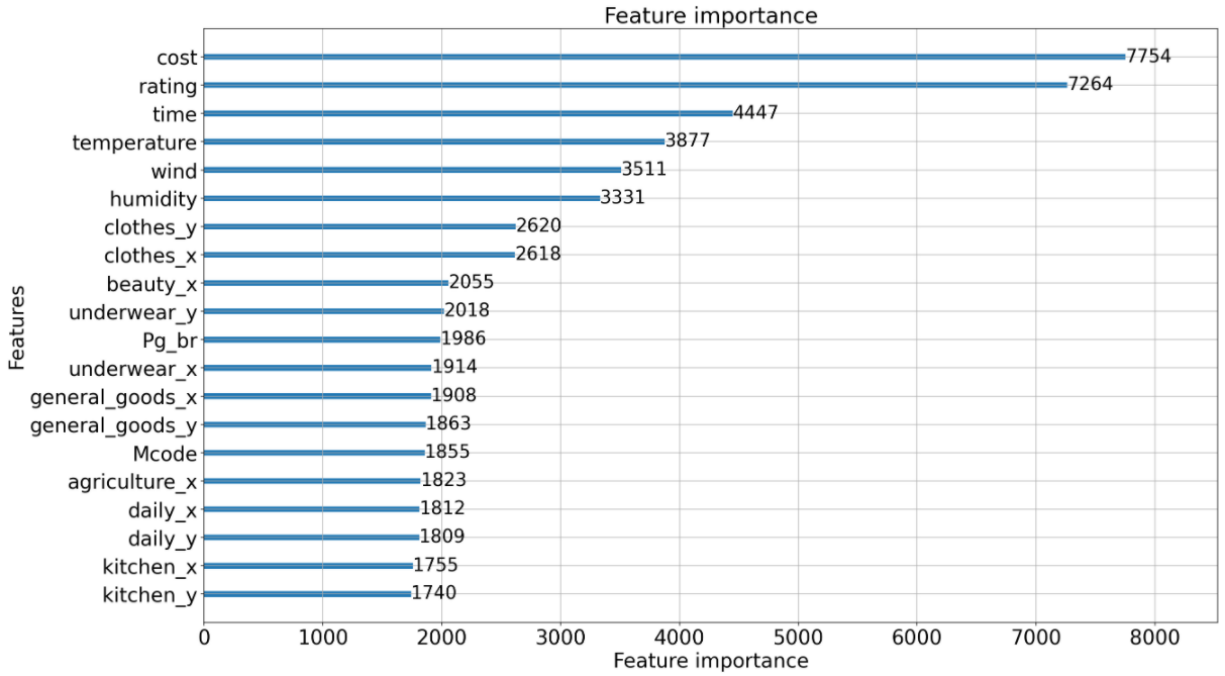
DNN

Layer (type)	Output Shape	Param #
dense_7 (Dense)	(None, 256)	409600
dropout_4 (Dropout)	(None, 256)	0
dense_8 (Dense)	(None, 256)	65792
dropout_5 (Dropout)	(None, 256)	0
dense_9 (Dense)	(None, 128)	32896
dropout_6 (Dropout)	(None, 128)	0
dense_10 (Dense)	(None, 64)	8256
dropout_7 (Dropout)	(None, 64)	0
dense_11 (Dense)	(None, 32)	2080
dense_12 (Dense)	(None, 16)	528
dense_13 (Dense)	(None, 1)	17

Total params: 519,169
Trainable params: 519,169
Non-trainable params: 0

02. 취급액 예측

변수 중요도 Plot



* LGBM 의 imporatance plot

- Cost / Rating / time /
- Temperature / Wind / humidity
- 네이버 검색량
- 상품군 & 브랜드 교호작용 (Pg_br)
- Mcode

➡ 외부변수, 파생변수들의 효과 확인

➡ 상위 Rank
- 네이버변수 → 트렌드에 민감하게 반응
- 날씨 변수

편성표 최적화

매출 예측과 헝가리안 알고리즘

03. 편성표 최적화

모든 경우의 상품코드의 시간별 날짜별 취급액을 예측하여 할당

예측

group	(1, 00)	(1, 01)	(1, 06)	(1, 07)	(1, 08)
Pcode					
200003	2.006274e+07	2.066555e+07	1.877079e+07	2.043867e+07	2.009243e+07
200004	2.075504e+07	2.073577e+07	1.970928e+07	2.159332e+07	2.132799e+07
200005	1.957742e+07	2.027404e+07	1.786776e+07	1.973428e+07	1.972516e+07
200006	2.228165e+07	1.797493e+07	1.980777e+07	2.429553e+07	2.402929e+07
200007	2.422010e+07	2.021227e+07	2.177290e+07	2.593238e+07	2.583620e+07
...	...	...	...	...	...
202485	8.826062e+06	9.271648e+06	1.037542e+07	1.256575e+07	1.267359e+07
202486	1.307647e+07	1.376533e+07	1.366663e+07	1.599830e+07	1.588506e+07
202506	2.935261e+07	2.741556e+07	2.298816e+07	2.810163e+07	2.713606e+07
202507	2.935261e+07	2.741556e+07	2.298816e+07	2.810163e+07	2.713606e+07
202511	4.581381e+07	4.339621e+07	4.122678e+07	4.711308e+07	4.541339e+07

389 rows × 607 columns

상품코드 ←

→ (일, 시간)

20년 6월에 편성된 모든 상품코드를
시간별 날짜별로 예측하여
모든 경우의 수의 취급액을 추출

헝가리안 할당

	Pcode	day	time
0	200003	1	7
1	200004	26	15
2	200006	8	0
3	200007	15	23
4	200010	22	00
..	...	..	...
508	202428_2	26	09
509	202484_2	27	13
510	202506_2	24	00
511	202507_2	16	18
512	202511_2	25	22

- 헝가리안 할당을 이용하여 매출을 최대화하도록 재편성
- 1개월간 최대 2번 편성할 수 있도록 함

03. 편성표 최적화

재편성된 예시 모습

시간	06.08 (월)	06.09 (화)	06.10 (수)	06.11 (목)	Today 06.12 (금)	06.13 (토)	06.14 (일)
14:00	14:00~15:00 (60) [보루네오] 델루나 유로탑 슬라이딩 LED침대 퀸 상품보기	14:00~15:00 (60) [신일] 써큘레이터 스탠드 블랙(SIF-PC30DCC) 상품보기	14:00~15:00 (60) [클라쎄] 벽걸이 에어컨 TDOZ-S10JK 상품보기	14:00~15:00 (60) [헨리코튼] 골프 여성 반바지3종세트 상품보기	14:00~15:00 (60) [LG전자] 매직스페이스 냉장고 무이자 상품보기	14:00~15:00 (60) [래쉬톡] 원터치 속눈썹 상품보기	14:00~15:00 (60) [한일] 대용량 레드 스텐 분쇄믹서기 상품보기
15:00	15:00~16:00 (60) [르젠] 1+1세트 무선 써큘레이터 상품보기	15:00~16:00 (60) [보국] 1+1세트 왕팬 윈큘레이터 BKF-0F435 상품보기	15:00~16:00 (60) [LG전자] 매직스페이스 냉장고 일시불 상품보기	15:00~16:00 (60) [예작] 남성 썸머셔츠 3종 상품보기	15:00~16:00 (60) [삼성] 청소기 파워모션 VC33M31B0LD 일시불 상품보기	15:00~16:00 (60) [CERINI] by PAT 남성 올데이 기능성 반팔 티셔츠 8종 상품보기	15:00~16:00 (60) [매직쉐프] 플랫타입 전자레인지 일시불 상품보기

03. 편성표 최적화

기존 편성표

방송일시	노출(분)	마더코드	상품코드	상품명	상품군	판매단가	취급액
2020-06-01 6:20	20	100650	201971	잭필드 남성 반팔셔츠 4종	의류	59,800	2679121
2020-06-01 6:40	20	100650	201971	잭필드 남성 반팔셔츠 4종	의류	59,800	8653764
2020-06-01 7:00	20	100650	201971	잭필드 남성 반팔셔츠 4종	의류	59,800	24256594
2020-06-01 7:20	20	100445	202278	쿠미투니카 쿨 레이시 란쥬쉐이퍼&팬티	속옷	69,900	25217092
2020-06-01 7:40	20	100445	202278	쿠미투니카 쿨 레이시 란쥬쉐이퍼&팬티	속옷	69,900	42578798
2020-06-01 8:00	20	100445	202278	쿠미투니카 쿨 레이시 란쥬쉐이퍼&팬티	속옷	69,900	50248929
2020-06-01 8:20	20	100381	201247	바비리스 퍼펙트 볼륨스타일러	이미용	59,000	18974862
2020-06-01 8:40	20	100381	201247	바비리스 퍼펙트 볼륨스타일러	이미용	59,000	32623024
2020-06-01 9:00	20	100381	201247	바비리스 퍼펙트 볼륨스타일러	이미용	59,000	45192741

기존 편성표에서
예상되는 6월 총 매출
₩71,879,737,749

새로운 편성표

시간	06.08 (월)	06.09 (화)	06.10 (수)	06.11 (목)	Today 06.12 (금)	06.13 (토)	06.14 (일)
14:00	14:00~15:00 (60) [보루네오] 멜루나 유로탑 슬라이딩 LED침대 퀸 상품보기	14:00~15:00 (60) [신일] 써큘레이터 스탠드 블랙(SIF-PC30DCC) 상품보기	14:00~15:00 (60) [클라쎄] 벽걸이 에어컨 TDOZ-S10JK 상품보기	14:00~15:00 (60) [헨리코튼] 골프 여성 반바지3종세트 상품보기	14:00~15:00 (60) [LG전자] 매직스페이스 냉장고 무이자 상품보기	14:00~15:00 (60) [래쉬톡] 원터치 속눈썹 상품보기	14:00~15:00 (60) [한일] 대용량 레드 스텐 분쇄믹서기 상품보기
15:00	15:00~16:00 (60) [르젠] 1+1세트 무선 써큘레이터 상품보기	15:00~16:00 (60) [보국] 1+1세트 왕팬 윈큘레이터 BKF-0F435 상품보기	15:00~16:00 (60) [LG전자] 매직스페이스 냉장고 일시불 상품보기	15:00~16:00 (60) [예작] 남성 썸머셔츠 3종 상품보기	15:00~16:00 (60) [삼성] 청소기 파워모션 VC33M31B0LD 일시 상품보기	15:00~16:00 (60) [CERINI] by PAT 남성 올데이 기능성 반팔 티셔츠 상품보기	15:00~16:00 (60) [매직쉐프] 플랫타입 전자레인지 일시불 상품보기

새롭게 편성한 편성표에서
예상되는 6월 총 매출
₩78,977,772,660

감사합니다