

2.핵심행사 수요예측 모델 v2.0(2024_phase1)

*모델 개선 사항

1. 카테고리 확장 학습

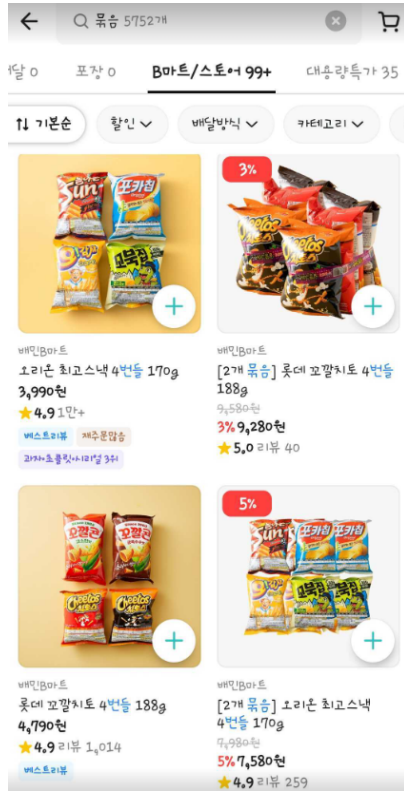
- a. AS - IS: 신선(채소, 정육/수산/계란, 과일), 우유/유제품 카테고리 → TO- BE : 전 카테고리 확장(편리한것 포함)

2. 행사 정보 고도화

- a. 행사 데이터 정보 DB화 → 기대효과: 정확한 행사 정보를 통해 모델의 성능 고도화
 - i. AS - IS: 수기 데이터 기반 학습(프로모션 & 기획전 관리, 핵심행사리스트, 발주 참고등 구글 시트 기반) → TO-BE: B마트 어드민내 등록된 행사 상품 기반 학습
 - 참고 테이블: market.bm_exhibition(기획전), market.pt_recommended_product(번적할인)
 - 행사 기간, 노출 순서, 메인 배너 노출 여부등의 행사 관련 정보 수집
- b. 행사 구좌 유형 변경 → 기대효과: MKT 캠페인등 주요 행사 효과의 정확한 구분을 통해 모델이 세부적으로 학습가능
 - i. AS - IS: 개인적인 기준에 의거한 구분(5일장, 과채고, 박세일, 주말장보기, 신규 첫주문등) → TO- BE: MD실과 협의한 5가지 행사 구좌 유형 구분(번적할인, MKT 캠페인, 광고, 주말장보기, 그외 행사)



































































3. 상품 가격 & 할인을 정보 고도화

- a. 상품의 정상가격(할인전 가격) 과 판매가격 (할인율 가격) 정보 Tracking → 모델 학습에 주요한 원인으로 작용하는 할인율 정보 고도화 가능
 - i. AS - IS: 수기 행사 리스트에 기재되어 있는 상품의 정상가격과 할인가격을 사용하여 정보의 한계 존재 → To - BE: 구매 영수증 기반의 정상가격과 할인 가격을 통해 실제 판매된 상품의 가격, 할인율 정보 학습
 - 상품 마스터 테이블의 정상가격, 구매 영수증 테이블의 상품가격 정보 Tracking
- b. 전시 상품 ID(Product id) 기반하의 학습으로 묶음 상품, 골라담기 정보 필터링 가능
 - i. AS - IS: 물류 상품 ID 기반(sku_code)의 학습으로 묶음 상품, 골라담기 상품들이 모두다 동일한 상품가격 정보 & 할인율로 구분되어 데이터 Noise 존재 → TO - BE: 정확한 구분을 통해 골라담기 상품의 판매량은 제거하고, 묶음 상품은 묶음 상품의 가격을 Tracking



4. 주요 Feature 확장

- a. 날씨 데이터 추가 → 기대효과: 날씨에 민감한 B마트 특성상, 기온과 강수량 여부 구분을 통해 모델의 정확도 상승 효과
 - i. AS - IS: 현재 모델에는 고려되어있지않음 → TO- BE: 하루 최저, 최고 기온 & 강수량(눈/비 포함) 데이터를 통해 모델 학습, 현시점 기준 10일 이후의 날씨 중기 예보 데이터 활용

지역	08일(금)		09일(토)		10일(일)		11일(월)		12일(화)	13일(수)	14일(목)
	오전	오후	오전	오후	오전	오후	오전	오후			
서울 인천 경기도	 10%	 10%	 10%	 10%	 10%	 10%	 10%	 30%	 30%	 40%	 20%
강원도 영서	 10%	 10%	 10%	 10%	 10%	 10%	 10%	 30%	 30%	 30%	 20%
강원도 영동	 10%	 10%	 10%	 10%	 10%	 10%	 10%	 30%	 20%	 20%	 20%
대전 세종 충청남도	 10%	 10%	 10%	 10%	 10%	 10%	 10%	 30%	 20%	 40%	 20%
충청북도	 10%	 10%	 10%	 10%	 10%	 10%	 10%	 30%	 30%	 40%	 20%
광주 전라남도	 10%	 10%	 10%	 10%	 10%	 10%	 20%	 30%	 20%	 20%	 20%

- b. 상품의 이동평균 데이터 변경(표준편차 추가)
- i. AS-IS: 상품 가격, 판매량, 결품(17시) 2주, 5주, 8주 평균 값 활용 → TO-BE: 상품 사격, 판매량, 결품(17시) 1주, 4주,7주 평균, 표준 편차

5. 예측 판매량 보정 작업

- a. 모델에서 예측 성능이 떨어지는 상품의 경우 기존 판매량 정보와 4주 평균 판매량 정보를 통해 판매량 보정

*활용 데이터

*데이터 예시

→ 행사 데이터에 대한 데이터 부족으로 신선 3개 카테고리의 데이터를 모두 활용하나 카테고리 정보를 통해 구분하도록 학습

- 1) 대상 품목: 신선 3개 카테고리(채소, 과일, 정육/수산/계란)
- 기본적으로 상품단위 정보 활용(EX. 과일 SKU의 경우 73개 FC의 합산된 판매량 정보)
- 2) 대상 기간: 2023-01-01 ~ 2024-02-14(결품률이 안정되기 시작한 시점인 10월 이후 데이터 활용)
- 3) 활용 변수: 상품 정보, 가격, 할인, 시계열 정보(이동평균선), 행사, 시간 정보등

index	데이터 유형	컬럼명	컬럼 의미	데이터타입	비고
1	기본정보	date_cd	일자	datetime	
2		category	depth카테고리	varchar	
3		sku_code	sku 코드	varchar	

5	가격, 할인	standard_sale_price	정상가격(할인전 가격)	integer	
6		sale_price	판매가격(할인후 가격)	integer	
7		discount_rate	할인율	float	행사 수기 데이터 활용
8	시계열- 이동평균 (실적, 가격, 결품)	sale_qty	판매수량/판매수량 편차 (1주/4주/7주)	float	
9		sale_price	판매가격/판매가격 편차 (1주/4주/7주)	float	
10		lack_17	17시 결품수/결품편차(1주/4주/7주)	integer	
11	행사	lightning_sale	번쩍할인 행사 여부	binary	
12		mkt	MKT 캠페인 행사 여부	binary	
13		ad_event	광고 행사 여부	binary	
14		waste_sale	과재고 행사 여부	binary	
15		week_sale	주말 장보기 행사 여부	binary	
16		etc_event	기타 행사 여부	binary	
17		no_regist_event	B마트 어드민 행사 여부	binary	
18		not_event	비행사 여부	binary	
19		main_expousre	매인 베너 노출 여부	binary	
20		position_no	노출순서	integer	결측값 다수 존재, 수기 데이터 활용
21	상품 정보	sku_grade	상품등급	categorical	수기데이터 활용
22		fc_storage_method	상품 보관방법	categorical	냉장, 냉동, 상온 구분
23		sale_able_dt	판매가능일수	categorical	상품의 판매가능일수
24		bundle_yn	번들상품여부	binary	
25	시간 정보	y	년도	categorical	미래 데이터 예측을 위해 시간 정보 추가
26		m	월	categorical	
27		d	일	categorical	
28		week_day	요일	categorical	
49	날씨 정보	temp_min	일평균 최저기온	float	API화 작업을 통해 날씨 데이터 & 날씨 중기 예보 데이터 추
50		temp_max	일평균 최고기온	float	
51		rain_yn	비/눈 여부	binary	

- 4) 타겟 변수: 판매량
- 1) 대상 품목: 신선 3개 카테고리(채소, 과일, 정육/수산물/계란)
 - 기본적으로 상품단위 정보 활용(EX. 과일 SKU의 경우 73개 FC의 합산된 판매량 정보)
- 2) 대상 기간: 2023-01-01 ~ 2024-02-14(결품률이 안정되기 시작한 시점인 10월 이후 데이터 활용)

3) 활용 변수: 상품 정보, 가격, 할인, 시계열 정보(이동평균선), 행사, 시간 정보등

index	데이터 유형	컬럼명	컬럼 의미	데이터타입	비고
1	기본정보	date_cd	일자	datetime	
2		category2	depth2 카테고리	varchar	
3		category3	depth3 카테고리	varchar	
4		category4	depth4 카테고리	varchar	
5		category5	depth5카테고리	varchar	
6		sku_code	sku 코드	varchar	
9		bundle_yn	번들상품여부	binary	
10	가격, 할인	standard_sale_price	정상가격(할인전 가격)	integer	
11		sale_price	판매가격(할인후 가격)	integer	
12		discount_rate	할인율	float	구매내역 기반 할인율
13	시계열- 이동평균 (실적, 가격, 결품)	saleqty_7_avg	판매수량 1주 평균	float	
14		sale_qty_28_avg	판매수량 4주 평균	flotat	
15		sale_qty_49_avg	판매수량 7주 평균	float	
16		sale_qty_7_std	판매수량 1주 편차	float	
17		sale_qty_28_std	판매수량 4주 편차	flotat	
18		sale_qty_49_std	판매수량 7주 편차	float	
19		sale_price_7_avg	판매가격 1주 평균	float	
20		sale_price_28_avg	판매가격 4주 평균	flotat	
21		sale_price_49_avg	판매가격 7주 평균	float	
22		sale_price_7_std	판매가격 1주 편차	float	
23		sale_price_28_std	판매가격 4주 편차	flotat	
24		sale_price_49_std	판매가격 7주 편차	float	
25		lack_17_7_avg	17시 결품수 1주 평균	integer	
26		lack_17_28_avg	17시 결품수 4주 평균	integer	
27		lack_17_49_avg	17시 결품수 7주 평균	integer	
28		lack_17_7_std	17시 결품수 1주 편차	integer	
29		lack_17_28_std	17시 결품수 4주 편차	integer	
30		lack_17_49_std	17시 결품수 7주 편차	integer	

31	행사	lightning_sale	번쩍할인 행사 여부	binary	
32		mkt	MKT 캠페인 행사 여부	binary	
33		ad_event	광고 행사 여부	binary	
34		waste_sale	과재고 행사 여부	binary	
35		week_sale	주말 장보기 행사 여부	binary	
36		etc_event	기타 행사 여부	binary	
37		no_regist_event	B마트 어드민 행사 여부	binary	
38		not_event	비행사 여부	binary	
39		main_expousre	매인 베너 노출 여부	binary	
40		position_no	노출순서	integer	결측값 다수 존재, 수기 데이터 활용
41	상품 정보	sku_grade	상품등급	categorical	수기데이터 활용
42		fc_storage_method	상품 보관방법	categorical	냉장, 냉동, 상온 구분
43		sale_able_dt	판매가능일수	categorical	
44		bundle_yn	번들상품여부	binary	
45	시간 정보	y	년도	categorical	미래 데이터 예측을 위해 시간 정보 추가
46		m	월	categorical	
47		d	일	categorical	
48		week_day	요일	categorical	
49	날씨 정보	temp_min	일평균 최저기온	float	API화 작업을 통해 날씨 데이터 & 날씨 증기 예보 데이터 추
50		temp_max	일평균 최고기온	float	
51		rain_yn	비/눈 여부	binary	

4) 타겟 변수: 판매량

*모델링 & 결과

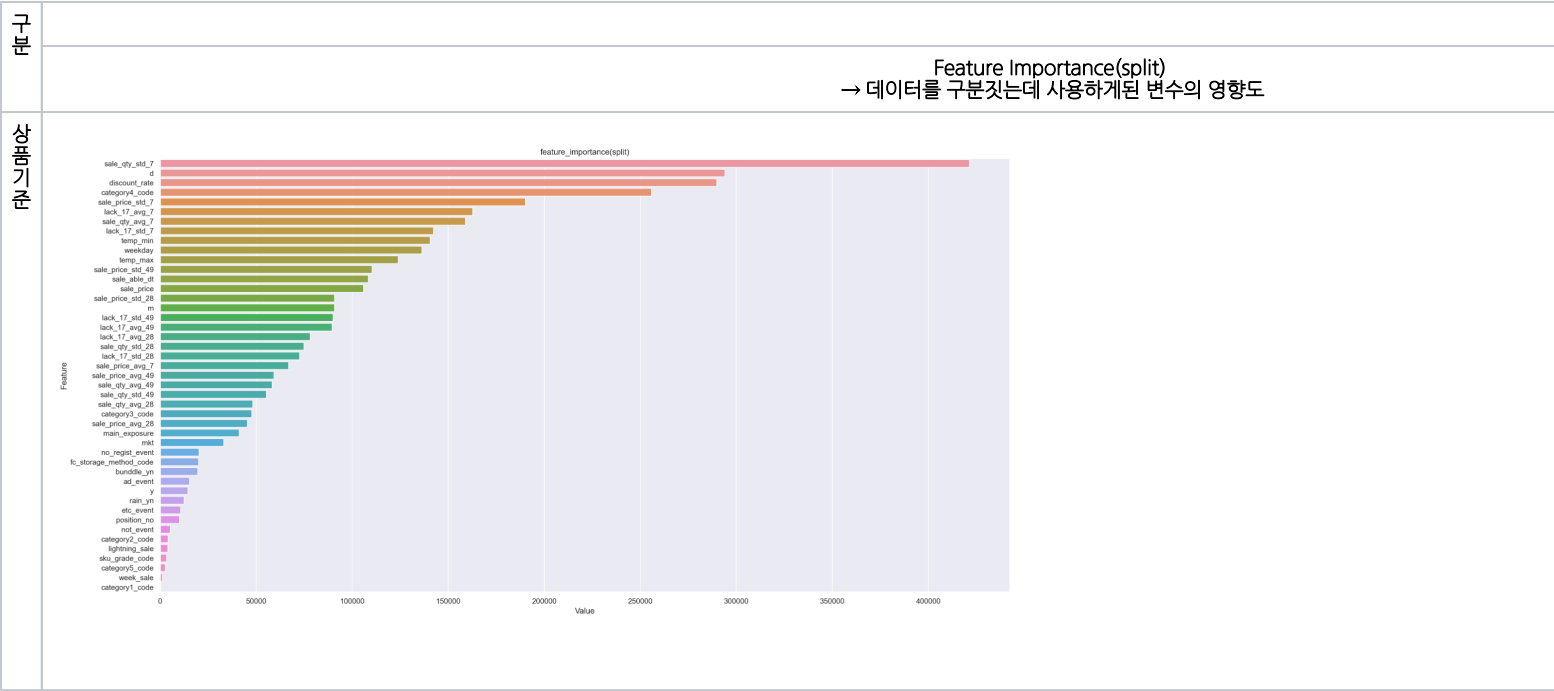
1. 행사 수요예측 version1 모델과 동일한 LightGBM 사용

2. 모델 학습 결과

평가지표	행사 수요예측 모델_v2	행사 수요예측 모델_v1	이마트 수요예측 모델
MAE	14.14	24.53	6~30
MAPE	59.64%	64.5%	70~80%
SMAPE	21%	23%	-

*이마트의 경우 주로 가공 카테고리 대상으로 분석하였으므로 신선과는 평가지표 자체에서 다른 경향성을 보일 수 있음

→ 기존 모델 대비 MAE는 10이상 줄어들며, 정확도 42% 상승, MAPE 정확도 6% 상승, SMAPE는 2% 상승하였음



*feature importance 값이 높을 수록 모델의 판매량 예측에 영향력을 준 변수로 볼 수 있음

3. 시뮬레이션

- 2월 2주차 (2월 12일 ~ 2월 19일)에 핵심 행사 리스트에 해당하는 상품의 예측 판매량 산출후 업데이트 모델, 기존 모델, MD 예상 판매량 오차 비교
→ 해당 주차의 비교 대상 품목이 173개로 비교 표본이 많지않아 추후 정확한 분석 필요
- 판매량 정의

INDEX	판매량 유형	설명	비고
-------	--------	----	----

1	ML 모델 예상 판매량_v2(변경)	고도화된 ML 모델에서 예측한 판매량	
2	ML 모델 예상 판매량_v1(기존)	기존 ML 모델에서 예측한 판매량	
3	MD 예상 판매량	핵심행사 리스트에 기재된 MD 예상 판매량	
4	실제 판매량	실제 해당 기간에 판매된 수량	

• 결과

- 173개 SKU 예측후 SKU의 행사 기간별 예상 판매량 산출(Ex. 하림 닭백숙 1,100g, 행사 기간 4/2 ~ 4/9 동안 1,637개 예상 판매량과 실제 판매량 비교)
- 173개중 90개의 ML 모델v2 오차율이 가장 뛰어난 것으로 판단(약 52%)

예측 성능 순위				
기준	ML 모델 예상 판매량_v2(변경)	ML 모델 예상 판매량_v1(기존)	MD 예상 판매량	합계
전체	52%	27%	21%	100%
순위(종합)	1위	3위	2위	

*각 예상 판매량과 실제 판매량의 차이 비교

*MD 예상 판매량과 핵심 행사 타겟치의 경우 값이 작성되어 있지 않은 경우가 존재하여, 정확한 판단을 위해 없는 경우를 제거한 Case에 대해서도 오차율 측정

• 단일성능 비교

- 1) ML VS MD

예측 성능 순위			
기준	ML모델	MD 예상 판매량	합계
전체	73%	27%	100%

- ML모델과 MD 예상 판매량중 정확도가 더욱 높은 유형은 ML 모델로 73%를 차지함
- ML 모델의 성능이 MD 예상 판매량 대비 약 2.7배정도 높은것을 확인할 수 있음

- 2) ML(v2) VS ML(v1)

예측 성능 순위			
기준	ML모델	MD 예상 판매량	합계
전체	62%	38%	100%

- 새롭게 개발한 ML모델과 ML모델 이전버전중 정확도가 더욱 높은 유형은 ML 모델로 62%를 차지함
- ML 모델의 성능이 이전 모델 대비 약 1.63배정도 높은것을 확인할 수 있음

raw_set: [비교데이터_0304.xlsx](#)