

2021 기상청 빅데이터 콘테스트 민간협력형

1. 공모 요약서

2. 공모안 PPT



1. 공모 요약서

1. 공모 제목

날씨 데이터 분석을 통한 온라인 시장 마케팅
: 앵거리즘과 SNS 노출광고를 중심으로

2. 공모 배경

날씨 데이터는 비즈니스 가치 데이터로 진화중이며, 온라인 시장 규모는 계속적으로 커지고 있다. 따라서 온라인 시장에서의 날씨 데이터 활용은 큰 잠재력을 갖고 있을 것으로 판단된다.

이에 날씨데이터와 온라인 판매 데이터 분석을 통해 효과적인 날씨 마케팅 전략을 수립할 수 있을 것으로 기대된다.

3. 분석데이터 정의

분석에 사용한 데이터는 크게 일별 기상 정보 데이터, 일별 미세먼지 데이터, 소셜 데이터, 일별 온라인 구매이력 데이터 4가지이다.

모든 데이터는 2018년, 2019년 2년간의 일별 데이터이다.

1) 기상 정보 데이터 : 기상자료 개방 포털에서 제공하는 강수량, 기온, 바람, 습도, 일조일사량 등 15개의 변수에 대한 데이터를 사용하였다. 온라인 구매 데이터를 이용하기 때문에 기상정보 역시 지역을 한정하지 않고 전국 지역의 기상 데이터를 이용하였다. 공식을 이용하여 불쾌지수와 체감온도 수치를 추가하였다.

2) 미세먼지 데이터 : Air Korea에서 제공하는 전국 지점별 일별 평균 대기오염도 데이터를 이용하였다. 미세먼지, 초미세먼지의 모든 지점 평균값을 기상정보 데이터에 추가하여 사용하였다.

3) 소셜 데이터 : 바이브 컴퍼니에서 제공하는 상품 383개에 대한 일별 SNS 언급량 데이터를 이용하였다. 블로그, 커뮤니티, 인스타그램의 문서 10만건당 건수를 이용하였다.

4) 온라인 구매이력 데이터: 상품 383개에 대한 일자별 구매자 성별, 연령대, 구매 건수 데이터를 이용하였다.

* 소셜데이터와 온라인 구매이력 데이터는 온라인 쇼핑몰 쿠팡과 지마켓의 상품 분류 카테고리 참조하여 중분류 카테고리를 추가하였다. 일자별 로 중분류를 기준으로 SNS 언급량과 구매 건수를 각각 합하여 cnt_sum 과 qty_sum 변수를 만들고, 이를 기상정보와 일자를 기준으로 결합하여 wth_sns 데이터프레임, wth_qty 데이터프레임을 만들어 분석에 이용했다.

4. 활용 분석기법 및 모델링 결과

1. XGBoost와 Elastic Net 회귀모델을 이용한 '날씨-SNS 언급량' 모델과 '날씨-구매량' 모델 비교

: SNS 언급량과 온라인 구매량 중 어떤 것이 더 당일의 날씨에 민감하게 반응하여 밀접하게 연관되어 있는지 알아보기 위해 with_sns, with_buy 데이터와 XGBoost, Elastic Net 회귀 모델을 이용하였다. XGBoost 모델에서 독립변수는 각 일자별 19개의 기상요소와 one-hot encoding을 통해 생성한 35개의 label dummy 변수로 하였다. 종속 변수는 중분류별 언급량의 합인 cnt_sum 변수, 중분류별 구매량 합인 qty_sum으로 하여 각각 모델을 학습하고 그 성능을 결정계수 R square 값으로 성능을 비교하였다. XGBoost 모델 학습 결과 '날씨-SNS 언급량'의 결정계수 0.9914, '날씨-SNS 언급량'의 결정계수 0.9197로 '날씨-SNS 언급량' 모델의 성능이 더 좋았다. 따라서 SNS 언급량이 당일의 날씨에 더 민감하게 반응한다고 판단하여 상품군 분석에 SNS 언급량을 사용하였다. 추가적으로, 전체 데이터가 아닌 중분류별 비교를 보기 위해 모델 학습 결과 중분류 중 약 60%가 '날씨-SNS 언급량' 모델의 성능이 더 높은 것이 나타났다.

2. Elastic Net 회귀 모델을 이용한 날씨 민감 상품군 탐색

: 중분류 중 어떤 상품군이 날씨에 민감하게 반응하는지 알아보기 위해 Elastic Net 회귀 모델의 R square 값을 이용하였다. 중분류 별로 데이터를 나누어 XGBoost 모델을 학습하였으나 데이터의 크기가 줄어들어 과적합 문제가 발생하여 다중선형회귀 모델을 사용하기로 결정하였다. 앞서 EDA와 변수 간 상관관계 분석 결과 독립 변수 간의 높은 상관관계가 존재하였다. 이로 인해 다중 선형 회귀 모델에서 다중 공선성으로 인한 문제가 우려되었다. 이를 해결하기 위해 회귀모델 중 Ridge와 Lasso 회귀의 장점을 모두 가지고 있어 변수 선택 효과를 낼 수 있고 variance를 줄일 수 있는 Elastic Net 회귀 모델을 선택하였다. 회귀모델의 성능은 종속변수의 정규성 전체에 영향을 많이 받기 때문에 EDA 결과 왜도가 높았던 종속변수를 로그화하여 분석에 이용하였다. R square 값이 0.35 이상이면 기상 요소들로 SNS 언급량을 설명할 수 있다고 보고 날씨 민감 상품군으로 지정하였다. 분석결과 '공기정화기기, 과일, 난방매트, 난방기기, 생수/음료, 선케어, 에어컨, 장/소스/드레싱식초' 8개의 상품군이 날씨민감상품군으로 분류되었다.

3. Permutation Importance를 이용한 상품군별 중요 기상 변수 탐색

: Permutation Importance를 이용하여 3번에서 구한 날씨 민감 상품군의 Elastic Net 회귀 모델의 변수 중요도를 구하였다. 가장 큰 중요도를 보인 변수 두 개를 해당 상품군에 영향을 많이 미치는 중요 기상 변수로 정의한다. 분석 결과 날씨 민감 상품군별 중요 기상 변수는 다음과 같다.

[공기 정화기기-PM10, PM25], [과일-THI(불쾌지수), avg_max_tem],

[난방매트-THI, avg_max_tem], [냉방기기-THI, avg_max_tem],

[생수/음료-avg_max_tem, THI], [선케어-sr_sum, avg_max_tem],

[에어컨-avg_max_tem, THI], [장/소스/드레싱식초-THI, sr_sum]

4. 중심화 이동평균과 선형 회귀모델을 이용한 SNS 언급과 구매 사이의 기간 분석

: SNS에 언급된 날씨 민감 상품군이 구매로 이어지기까지 걸리는 시간을 알아보기 위해 중심화 이동평균과 선형회귀 모델을 이용하였다. 시계열 데이터인 일자별 언급량의 불규칙 성분을 제거하기 위해 이동평균값을 구하였다. 이동평균법 중 데이터가 계절성을 띄는 경우에 적합한 중심화 이동평균을 이용하였다. 7일치의 데이터를 중심화이동평균 낸 값과 며칠 뒤의 구매량이 회귀분석 시 가장 큰 R square 값을 갖는지 비교하였다. 1,3,5,7,14,21,28,35,60 일 뒤의 구매량과 각각 회귀분석을 하였고 1일 뒤의 값과 R square 0.5 이상을 갖는 상품군은 공기정화기기, 냉방기기, 난방매트, 선케어, 에어컨이다. 이는 해당 상품군들을 구매하기 1일 전, 사람들이 SNS를 작성하기 위해 SNS에 접근한다는 것으로 해석될 수 있다. 이 외의 품목은 작은 R square 값을 가져 SNS에서 표현된 관심이 실제 구매로 이어지지 않는 것으로 파악되었다.

5. 서비스 활용 방안

- 분석 결과 다수의 상품군에서 불쾌지수가 변수 중요도 상위권에 존재하고, 중심화 이동평균 분석결과 구매 전 SNS 접근이 이루어지는 상품군과 SNS의 관심이 실제 구매로 이어지지 않는 상품군이 존재하여 이를 마케팅에 활용하기로 결정하였다.
- 불쾌지수와 양의 상관관계인 상품군은 일명 '앵거리즘' 전략을 활용한다. 앵거리즘은 스니커즈사의 '헝거리즘'을 벤치마킹한 것으로 불쾌지수 정도에 따라 할인율을 차등 적용하여 불쾌지수가 높아질수록 높은 할인율을 적용하는 방식이다. 해당 전략을 통해 불쾌지수가 높아 상품에 대한 수요가 높아졌을 때 할인율을 높게 적용함을 광고하여 더욱 그 구매를 늘릴 수 있다.
- 특히 구매 전 sns에 접근하는 것으로 분류된 상품군은 수요가 높아질 때 SNS에 광고를 내보내면 구매 의사가 있는 사람들이 광고에 효과적으로 노출되고 자사 기업의 제품 구매로 유도할 수 있다. sns 관심표현이 구매로 이어지지 않는 경우에도 영향을 미치는 날씨 요소가 상승할 때 광고에 지속적으로 노출되어 실제 구매 이어지는 경우가 증가할 수 있다.

6. 기대 효과

- 앵거리즘을 활용함으로써 소비자의 감성지수에 따라 가격을 결정한다는 점에서 소비자와의 공감대 형성 및 브랜드 친숙도가 향상될 것으로 기대한다.
- 소비자가 물품을 구매하기 전 SNS에 접근한다는 점을 이용하여 보다 효과적인 시기의 SNS 광고를 통해 마케팅 비용을 효율적으로 사용할 수 있다.

2. 공모안 PPT

날씨 데이터 분석을 통한 온라인 시장 마케팅 : 앵거리즘과 SNS 노출광고를 중심으로

엘 리 온

류지민
이나연
이선주
이동우

CONTENTS

1. 공모배경

2. 분석 데이터 정의

- 1. 사용 데이터 소개
- 2. 데이터 전처리

3. 활용 분석 기법 및 모델링 결과

4. 서비스 활용 방안 및 기대효과

CN A P T E R 1

공모배경

1. 경영/마케팅/판매에 있어서 날씨 데이터의 중요성



날씨는 단순히 일기예보가 아니라
상품의 수요를 예측하고, 기업경영 계획까지 바꾸는
‘비즈니스 가치 데이터’로 진화중

=> 기상정보는 사회, 경제적 효용가치가 높은 빅데이터

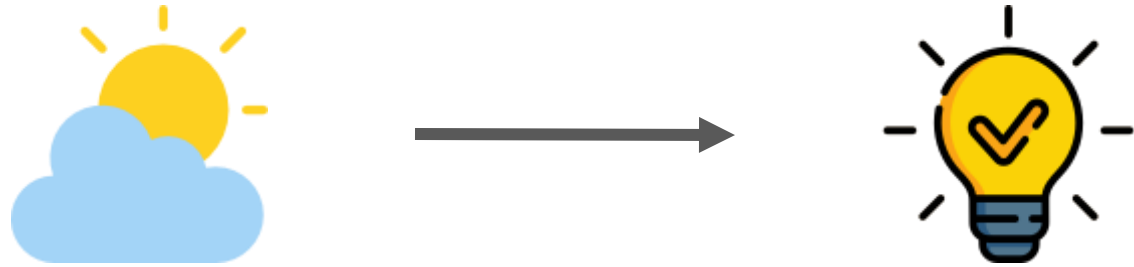
스타벅스의 Rainy Day BOGO 쿠폰 : 비오는 날 친구와 함께 가서
쿠폰을 내면 친구의 음료까지 주는 마케팅

나뭇루의 'Rainy Day', 'Sunny Day' 캠페인 : 비오는 날에는
아이스크림 가격 할인, 맑은 날에는 아이스크림 한 스쿱 더 올려줌



온라인 구매 정보 데이터와 기상정보 데이터를 분석하여
온라인 시장에서 날씨와 연관된 구매행태를 파악할 필요가 있다.

⇒ 유통, 판매를 포함한 전 분야에 걸친
더욱 효과적인 경영, 판매 전략을 세울 수 있을 것이다



기상 데이터를 분석하여 얻은 인사이트를 이용

=> 날씨에 따른 특정 제품에 대한 판매율을 상승시키는 마케팅 전략



특히, 소셜 미디어 상의 정보는 소비자의 구매의사결정에 중요한 요소로 작용



이를 활용한 마케팅 전략은 온라인 상의 판매율/수익율을 높이는데 효과적

CHAPTER 2

분석 데이터 정의/소개

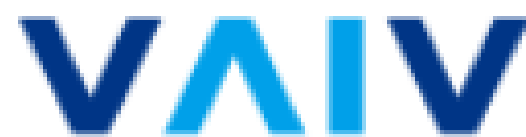
분석 데이터 정의 | 1. 사용 데이터

전국 기상 변수 데이터:
2018-01-01~2019-12-31
강수량, 기온, 바람, 습도, 일조일사량



미세먼지 데이터:
2018-01-01~2019-12-31
전국 일별 평균 데이터

온라인 구매 데이터:
2018-01-01~2019-12-31
분석 상품군/뷰티, 식품, 냉난방가전



소셜 데이터;
2018-01-01~2019-12-31
분석 상품군/뷰티, 식품, 냉난방가전
블로그, 커뮤니티, 인스타그램

분석 데이터 정의 | 2. 데이터 전처리 및 EDA : 날씨 데이터 셋 만들기/변수 생성

☀ 온라인 구매와 함께 분석-> 지역 특징이 불가!
=> 전국 기준으로 데이터셋을 생성

Max_tem-min_tem으로 tem_range의 결측치 대체시

☀ 최다 일강수량, 일교차에서 결측치 발견!

결측치를 제외한 행에서 차이가 0에 수렴

=> 이 값으로 tem_range의 결측치를 대체하기로 결정!

최다일강수량, 일교차 - null 존재

column:	avg_rn	Percent of NaN value: 0.00%
column:	max_rn	Percent of NaN value: 10.82%
column:	avg_tem	Percent of NaN value: 0.00%
column:	avg_max_tem	Percent of NaN value: 0.00%
column:	max_tem	Percent of NaN value: 0.00%
column:	avg_min_tem	Percent of NaN value: 0.00%
column:	min_tem	Percent of NaN value: 0.00%
column:	tem_range	Percent of NaN value: 0.27%
column:	avg_wd	Percent of NaN value: 0.00%
column:	max_wd	Percent of NaN value: 0.00%
column:	max_inst_wd	Percent of NaN value: 0.00%
column:	avg_rhm	Percent of NaN value: 0.00%
column:	min_rhm	Percent of NaN value: 0.00%
column:	ss_sum	Percent of NaN value: 0.00%
column:	sr_sum	Percent of NaN value: 0.00%

강수량) Nan 값-> 강수량이 0인 경우 -> 0으로 모두 대체

일교차)

date	
2018-01-01	-13.0
2018-01-02	-12.7
2018-01-03	-20.0
2018-01-04	-19.9
2018-01-05	-13.9
...	
2019-12-27	-12.6
2019-12-28	-10.9
2019-12-29	-17.2
2019-12-30	-13.3
2019-12-31	-18.0

avg_max_tem-avg_min_tem

date	
2018-01-01	0.000000e+00
2018-01-02	0.000000e+00
2018-01-03	0.000000e+00
2018-01-04	0.000000e+00
2018-01-05	0.000000e+00
...	
2019-12-27	0.000000e+00
2019-12-28	-3.552714e-15
2019-12-29	3.552714e-15
2019-12-30	0.000000e+00
2019-12-31	0.000000e+00

max_tem-min_tem

분석 데이터 정의

2. 데이터 전처리

: 날씨 데이터 셋 만들기/변수 생성

 불쾌지수(THI)

*불쾌지수(THI) 공식: $(0.81 \times \text{avg_tem}) + 0.01 \times \text{avg_rhm} \times ((0.99 \times \text{avg_Tem}) - 14.3) + 46.3$

*체감온도(sen_tem) 공식: $13.12 + 0.6215 \times \text{avg_tem} - 11.37 \times (\text{avg_wd} \times 3.6)^{**0.16} + 0.3965 \times (\text{avg_wd} \times 3.6)^{**0.16} \times \text{avg_tem}$

 체감온도





 month

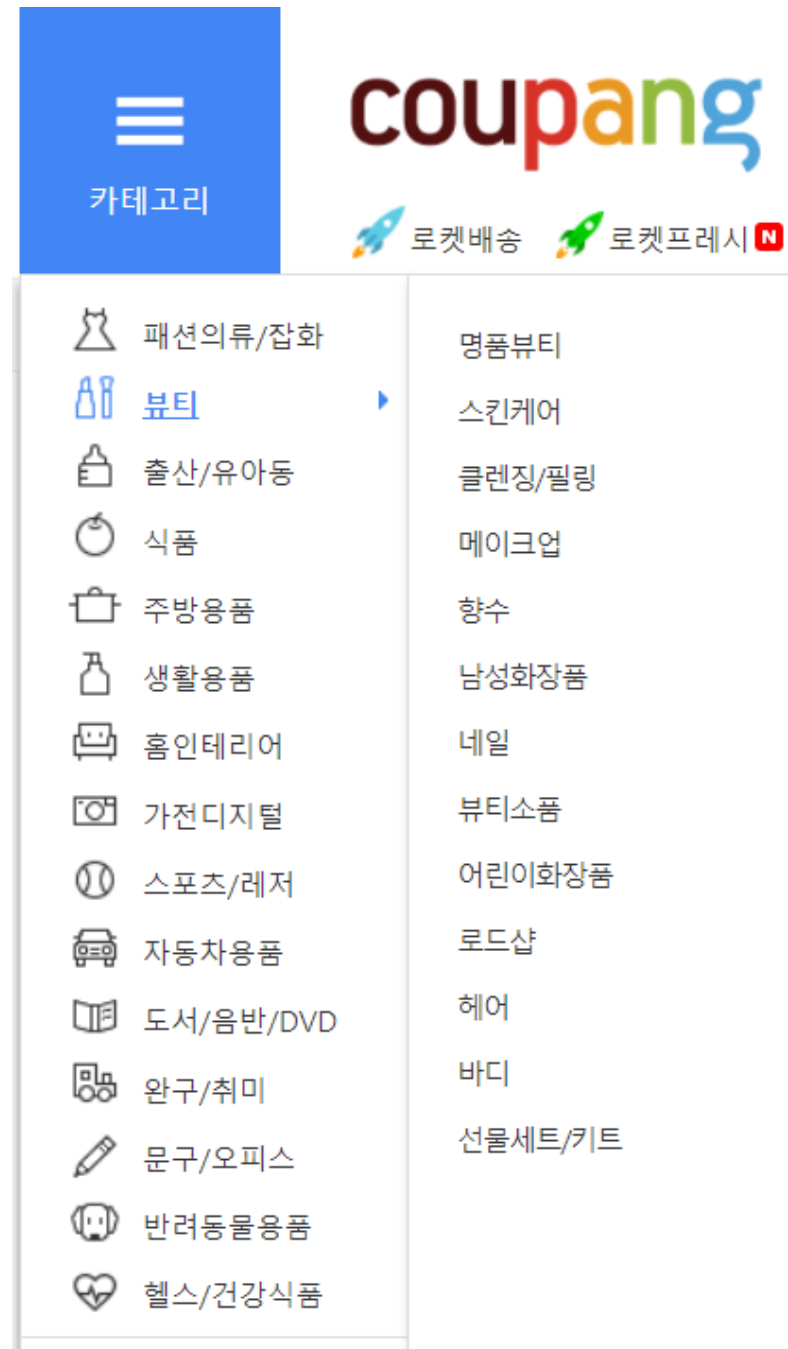
 미세먼지

지점별 데이터만 존재

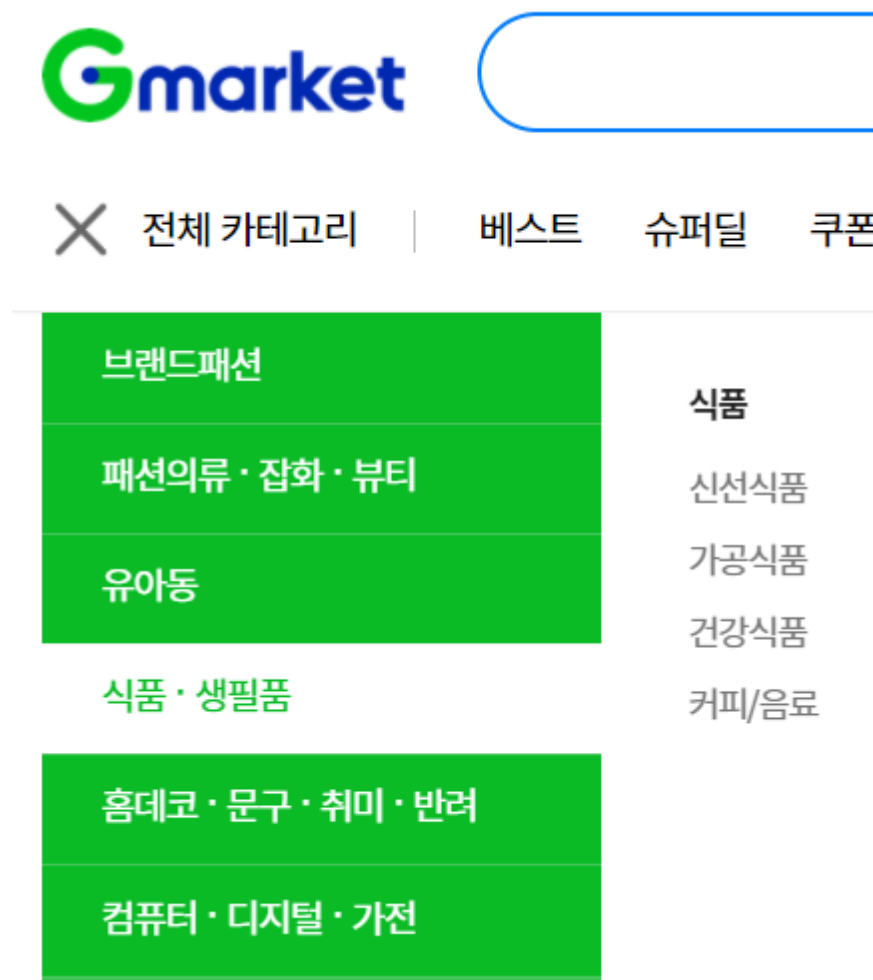
-> 전국 지점별 일별 평균 대기 오염도 데이터에서 미세먼지, 초미세먼지만 일별 모든 지점 평균값으로 전국 미세먼지 dataset 생성, concat

	PM10	PM25	month	avg_rn	max_rn	avg_tem	avg_max_tem	max_tem	avg_min_tem	min_tem	tem_range	avg_wd	max_wd	max_inst_wd	avg_rhm	min_rhm	ss_sum	sr_sum	THI	sen_tem
2018-01-01	44.247625	21.767793	1	0.0	0.0	-0.4	5.2	10.0	-4.9	-13.1	23.1	1.9	12.5	18.1	51.0	12.0	8.7	9.31	38.48104	-2.810026
2018-01-02	55.835057	31.722031	1	0.0	0.0	0.2	5.8	11.5	-5.3	-12.3	23.8	1.8	11.0	17.8	57.0	14.0	6.6	6.89	38.42386	-1.981246
2018-01-03	33.682890	18.207946	1	0.0	3.8	-1.6	2.8	11.3	-5.3	-16.8	28.1	2.4	12.1	17.8	46.0	13.0	8.6	9.35	37.69736	-4.824906
2018-01-04	41.252011	23.891078	1	0.0	5.0	-1.7	2.3	9.3	-5.6	-18.5	27.8	1.6	10.9	16.2	51.0	12.0	5.0	6.98	36.77167	-3.874782
2018-01-05	49.815903	32.736032	1	0.0	4.5	-0.1	4.7	9.9	-3.4	-12.1	22.0	2.0	16.2	21.5	54.0	10.0	6.7	7.90	38.44354	-2.589670
...
2019-12-27	28.883411	17.549153	12	0.0	5.9	0.6	5.1	9.6	-3.2	-11.3	20.9	2.4	21.5	25.9	53.0	14.0	8.2	10.36	39.52182	-2.225897
2019-12-28	40.172877	22.655049	12	0.0	1.0	0.5	7.5	11.6	-4.9	-11.7	23.3	1.1	7.3	11.7	60.0	20.0	8.2	10.07	38.42200	-0.492879
2019-12-29	44.382107	28.309295	12	6.2	34.3	2.6	6.6	17.3	-2.4	-8.9	26.2	1.2	18.1	21.0	73.0	26.0	0.2	3.33	39.84602	1.669374
2019-12-30	31.393231	22.152415	12	0.9	3.5	5.0	9.8	15.5	0.1	-7.5	23.0	2.1	17.4	23.8	79.0	15.0	2.2	3.96	42.96350	3.252319
2019-12-31	21.731749	11.849556	12	0.1	10.2	-3.8	0.4	10.7	-6.7	-14.4	25.1	3.6	21.9	28.4	46.0	13.0	8.8	11.31	34.91348	-8.642578

-  주어진 온라인 구매 데이터: 대분류, 소분류로 구분된 온라인 구매 데이터
-  소분류로 분석/전략 수립: 전략의 범용성 DOWN. 새로운 제품군이 추가되었을 시 대응이 번거로움.
-  대분류로 분석/전략 수립: 각 범주 안에 속하는 너무나 다양한 상품들이 존재. 이를 해당 수준의 범주에서 분석/전략 수립 하기에는 제품들의 각각 특성을 너무 간과할 것이라는 판단
-  중분류라는 카테고리를 신설!
: 대분류 내 제품들의 특성 차이를 반영하고,
새로운 제품군이 추가되었을 때 범용성이 높은 전략 수립



쿠팡, 지마켓의 분류 카테고리를 참조,
중분류 카테고리를 신설



분석 데이터 정의

2. 데이터 전처리

: 온라인 구매 데이터/ 데이터 셋 만들기



중분류로 데이터 추가

```
#buy_last 데이터에 중분류 데이터를 추가  
buy_new_last=pd.merge(buy_last,mid_cat_last,how='left',on='sm_cat')
```



중분류별 라벨링 진행

```
from sklearn.preprocessing import LabelEncoder  
encoder = LabelEncoder()  
buy_new_last['mid_label']=encoder.fit_transform(buy_new_last['mid_cat'].values)
```

	date	sex	age	big_cat	sm_cat	qty	month	mid_cat	mid_label
0	2018-01-01	F	20	식품	가공란	37	1	축산/계란	30
1	2018-01-01	F	30	식품	가공란	16	1	축산/계란	30
2	2018-01-01	F	40	식품	가공란	9	1	축산/계란	30
3	2018-01-01	F	50	식품	가공란	3	1	축산/계란	30
4	2018-01-01	M	20	식품	가공란	13	1	축산/계란	30
...
2056894	2019-12-31	M	20	냉난방가전	히터	8	12	난방기기	9
2056895	2019-12-31	M	30	냉난방가전	히터	22	12	난방기기	9
2056896	2019-12-31	M	40	냉난방가전	히터	38	12	난방기기	9
2056897	2019-12-31	M	50	냉난방가전	히터	23	12	난방기기	9
2056898	2019-12-31	M	60	냉난방가전	히터	10	12	난방기기	9



중분류를 기준으로 일별 판매량 합하기

```
# 중분류를 기준으로 일별 판매량을 합한다.  
buy_agg=buy[['date','qty','mid_label']].groupby(['date','mid_label']).sum().reset_index()
```



날짜/중분류에 대한 행 데이터로 buy_agg 결측값 확인, 0으로 대체

```
buy_agg=pd.merge(wth_sns[['date','mid_label']],buy_agg,on=['date','mid_label'],how='outer')
```

```
buy_agg.fillna(0,inplace=True)
```

	date	mid_label	qty_sum
0	2018-01-01	0	5523.0
1	2018-01-01	1	123.0
2	2018-01-01	2	821.0
3	2018-01-01	3	226.0
4	2018-01-01	4	1636.0
...
25545	2019-12-31	30	4693.0
25546	2019-12-31	31	2687.0
25547	2019-12-31	32	1130.0
25548	2019-12-31	33	0.0
25549	2019-12-31	34	2503.0

날씨 데이터와 생성된 buy데이터를 일자별로 merge!

분석 데이터 정의

2. 데이터 전처리

: SNS 데이터/ 데이터 셋 만들기

중분류로 데이터 추가

```
#중분류 행 추가
mid_cat_last=pd.read_csv('mid_cat_last.csv',encoding='euc-kr')
sns_new_last=pd.merge(sns_last,mid_cat_last,how='left',on='sm_cat')
```

중분류별 라벨링 진행

```
#중분류 Label 추가
from sklearn.preprocessing import LabelEncoder
encoder = LabelEncoder()
sns_new_last['mid_label'] = encoder.fit_transform(sns_new_last['mid_cat'].values)
```

	date	big_cat	sm_cat	cnt	month	mid_cat	mid_label
0	2018-01-01	뷰티	기능성 링클케어 화장품	12.154295	1	기능성 화장품	6
1	2018-01-01	뷰티	기능성 모공관리 화장품	36.000828	1	기능성 화장품	6
2	2018-01-01	뷰티	기능성 아이케어 화장품	0.895782	1	기능성 화장품	6
3	2018-01-01	뷰티	기능성 영양보습 화장품	14.868175	1	기능성 화장품	6
4	2018-01-01	뷰티	기능성 트러블케어 화장품	48.819391	1	기능성 화장품	6
...
279585	2019-12-31	냉난방가전	가스온수기	0.084023	12	난방기기	9
279586	2019-12-31	냉난방가전	산림욕기	0.172214	12	공기정화기기	3
279587	2019-12-31	냉난방가전	에어커튼	0.336094	12	냉방기기	12
279588	2019-12-31	냉난방가전	신발건조기	1.224592	12	건조기	1
279589	2019-12-31	냉난방가전	의류건조기	3.168111	12	건조기	1

중분류를 기준으로 일별 SNS 언급량을 합하기

```
# 중분류를 기준으로 일별 sns 언급량을 합한다.
sns_agg=sns[['date','cnt','mid_label']].groupby(['date','mid_label']).sum().reset_index()
```

날씨 데이터와 생성된 sns데이터를 일자별로 merge!

분석 데이터 정의

2. 데이터 전처리

: SNS 데이터/ 데이터 셋 만들기

중분류로 데이터 추가

```
#중분류 행 추가
mid_cat_last=pd.read_csv('mid_cat_last.csv',encoding='euc-kr')
sns_new_last=pd.merge(sns_last,mid_cat_last,how='left',on='sm_cat')
```

중분류별 라벨링 진행

```
#중분류 Label 추가
from sklearn.preprocessing import LabelEncoder
encoder = LabelEncoder()
sns_new_last['mid_label'] = encoder.fit_transform(sns_new_last['mid_cat'].values)
```

	date	big_cat	sm_cat	cnt	month	mid_cat	mid_label
0	2018-01-01	뷰티	기능성 링클케어 화장품	12.154295	1	기능성 화장품	6
1	2018-01-01	뷰티	기능성 모공관리 화장품	36.000828	1	기능성 화장품	6
2	2018-01-01	뷰티	기능성 아이케어 화장품	0.895782	1	기능성 화장품	6
3	2018-01-01	뷰티	기능성 영양보습 화장품	14.868175	1	기능성 화장품	6
4	2018-01-01	뷰티	기능성 트러블케어 화장품	48.819391	1	기능성 화장품	6
...
279585	2019-12-31	냉난방가전	가스온수기	0.084023	12	난방기기	9
279586	2019-12-31	냉난방가전	산림욕기	0.172214	12	공기정화기기	3
279587	2019-12-31	냉난방가전	에어커튼	0.336094	12	냉방기기	12
279588	2019-12-31	냉난방가전	신발건조기	1.224592	12	건조기	1
279589	2019-12-31	냉난방가전	의류건조기	3.168111	12	건조기	1

중분류를 기준으로 일별 SNS 언급량을 합하기

```
# 중분류를 기준으로 일별 sns 언급량을 합한다.
sns_agg=sns[['date','cnt','mid_label']].groupby(['date','mid_label']).sum().reset_index()
```

날씨 데이터와 생성된 sns데이터를 일자별로 merge!

CHAPTER 03

활용 분석 기법 및 모델링 결과

3. 활용 분석 기법 및 모델링 결과

분석 과정

- 1 XGBoost 회귀모델을 이용한 '날씨-SNS 언급량' 모델과 '날씨-구매량' 모델 비교
- 2 Elastic Net 회귀 모델을 이용한 날씨 민감 상품군 탐색
- 3 Permutation Importance를 이용한 상품군별 중요 기상 변수 탐색
- 4 중심화 이동평균과 선형 회귀모델을 이용한 SNS 언급과 구매 사이의 기간 분석

3. 활용 분석 기법 및 모델링 결과

1

XGBoost 회귀모델을 이용한 '날씨-SNS 언급량' 모델과 '날씨-구매량' 모델 비교

< XGBoost 회귀모델 >

- 중분류 원-핫 인코딩 : 새로운 데이터프레임 생성
 - > 'wth_buy_one'
 - > 'wth_sns_one'
 - => 변수 추가 : label_0, label_1 . . . label_34
- Dateset을 7:3 비율로 train, test dateset 나눔
- 전체 dateset으로 XGBoost 회귀모델 돌림
 - > n_estimators=500, learning_rate=0.1,
max_depth=4

	buy	sns
Train_r^2	0.9567	0.9954
Test_r^2	0.9197	0.9914



'날씨-SNS 언급량' 모델의 설명력이 더 높음

3. 활용 분석 기법 및 모델링 결과

2 Elastic Net 회귀 모델을 이용한 날씨 민감 상품군 탐색

$r^2 > 0.35$ 이상인 중분류 선택



라벨	중분류	r^2
03	공기정화기기	0.417620075
04	과일	0.382079197
10	난방매트	0.394932342
12	냉방기기	0.621953347
20	생수/음료	0.694241031
21	선케어	0.727865951
26	에어컨	0.508298716
28	장/소스/드레싱식초	0.459458556

3. 활용 분석 기법 및 모델링 결과

3 Permutation Importance를 이용한 상품군별 중요 기상 변수 탐색

- Permutation Importance를 이용해 2번에서 구한 날씨 민감 상품군의 Elastic Net 회귀모델의 변수 중요도를 구함

<날씨 민감 상품군 별 중요 기상 변수>

라벨	중분류	변수1	변수2
03	공기정화기기	PM10	PM25
04	과일	THI(불쾌지수)	avg_max_tem
10	난방매트	THI(불쾌지수)	avg_max_tem
12	냉방기기	THI(불쾌지수)	avg_max_tem
20	생수/음료	avg_max_tem	THI(불쾌지수)
21	선케어	sr_sum	avg_max_tem
26	에어컨	avg_max_tem	THI(불쾌지수)
28	장/소스/드레싱식초	THI(불쾌지수)	sr_sum

Weight	Feature
0.8705 ± 0.0455	THI
0.2867 ± 0.0318	avg_max_tem
0.1861 ± 0.0204	avg_min_tem
0.0957 ± 0.0302	PM25
0.0920 ± 0.0117	PM10
0.0548 ± 0.0124	min_tem
0.0376 ± 0.0053	ss_sum
0.0275 ± 0.0104	max_rn
0.0262 ± 0.0140	avg_rhm
0.0107 ± 0.0026	max_tem
0.0086 ± 0.0052	min_rhm
0.0004 ± 0.0016	max_inst_wd
0.0001 ± 0.0002	avg_rn
0.0000 ± 0.0000	avg_tem
0.0000 ± 0.0000	sen_tem
0.0000 ± 0.0000	avg_wd
0.0000 ± 0.0000	max_wd
0.0000 ± 0.0000	sr_sum
0.0000 ± 0.0000	tem_range

Ex) 냉방기기

3. 활용 분석 기법 및 모델링 결과

4 중심화 이동평균과 선형 회귀모델을 이용한 SNS 언급과 구매 사이의 기간 분석

SNS에 언급된 날씨 민감 상품군이 구매로 이어지기까지 얼마나 걸릴까?

X=불규칙 성분을 줄인 일별 SNS 언급량의 7일 중심화 이동평균값

Y= n(=1,3,5,7,14,21,35,60) 일 뒤의 구매량

N=1
R² 값이 0.5 이상

라벨	상품군	R ²
3	공기정화기기	0.7280
10	난방매트	0.6007
12	냉방기기	0.5555
21	선케어	0.5909
26	에어컨	0.5388



회귀분석

R² 값이 0.5 미만

과일

생수/음료

장/소스/드레싱식초

SNS에 표현된 관심이
실제 구매로 이어지지 않는 상품군

구매자들이 구매 1일 전 SNS 에 접근하는 상품군

C H A P T E R 4

서비스 활용방안/기대효과

4. 서비스 활용방안/기대효과



불쾌지수를 마케팅에 적용할 순 없을까?

Concept: 불쾌지수를 토대로 사람들의 마음을 알아준다.

4. 서비스 활용방안/기대효과

선례 탐구/사례 분석

헝거 리즘[HUNGER-ISM]: 미국 Mars사의 SNICKERS 초콜릿 바 마케팅 전략



- > 허기가 질 때 SNS에 사람들의 감정적 격양/ 안좋은 반응을 감지
- > 허기를 달래기 위해 소비자들의 초콜릿 바 구매 의욕 증가
 - 소비 의욕이 증가할 것으로 예측되는 시점에 사람들의 감정 수준을 10단계로 나누어 평가, 차등적 할인을 적용
- > 초콜릿을 더 싼 가격에 제시함으로써 짜증을 조금 덜어보라는 메시지를 전달
- > 매출 67% 이상 성장, 사이트 방문자 수 1000% 이상 증가 효과



4. 서비스 활용방안/기대효과

[STRATEGY]



[SNS]



[ANGER-ISM]

4. 서비스 활용방안/기대효과

앵거리즘

불쾌 지수:

날씨에 따라 인간이 느끼는 불쾌감의 정도를
기온과 습도를 이용해 나타낸 수치

💡 무더위, 습함 으로부터 오는 짜증과 불쾌함을, 품목들을
조금 더 싸게 구매하여 더 기분 좋은 소비를 할 수 있다면?

💡 불쾌지수로 사람들의 감정을 짐작,
정도에 따라 할인율을 달리하는 마케팅 전략 수립

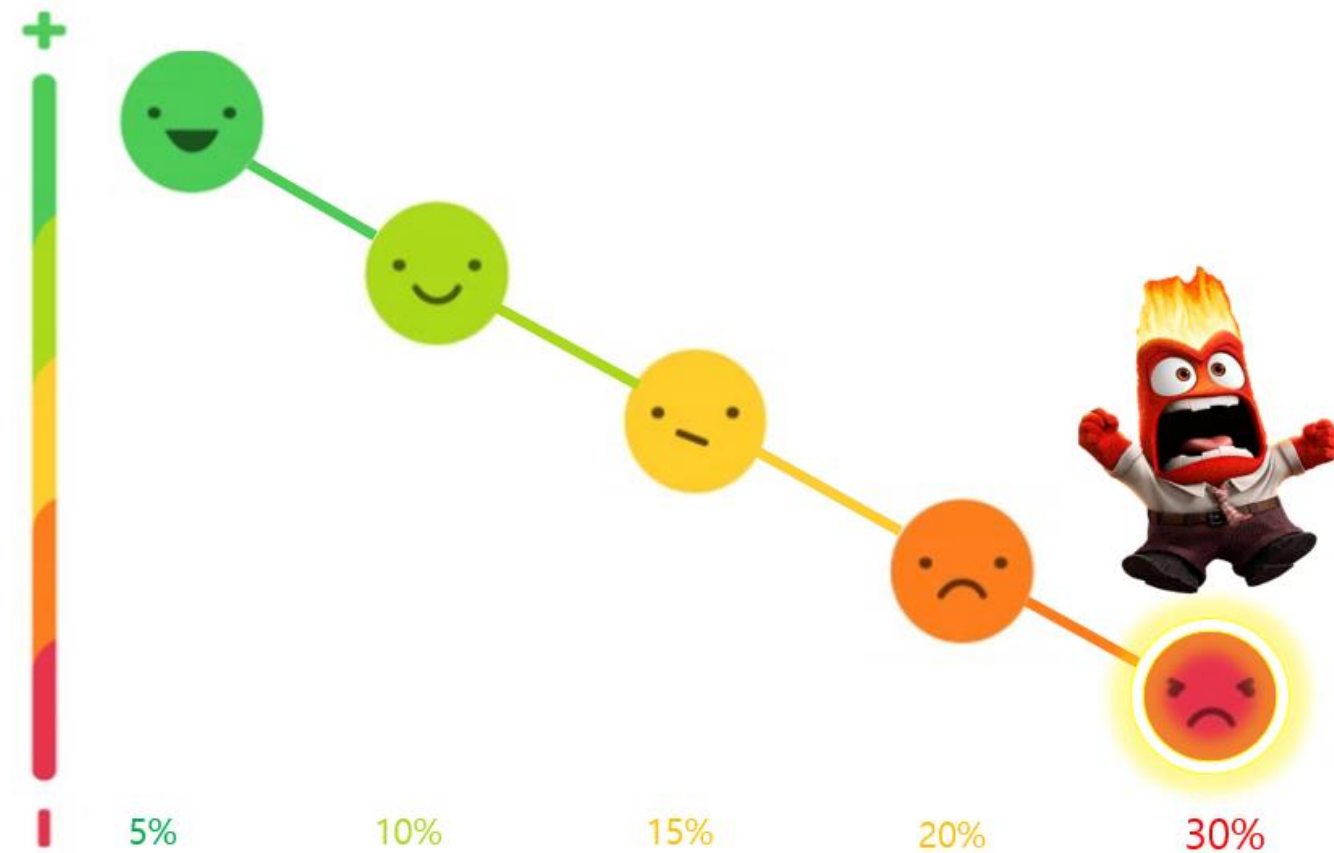


[ANGER-ISM]

4. 서비스 활용방안/기대효과

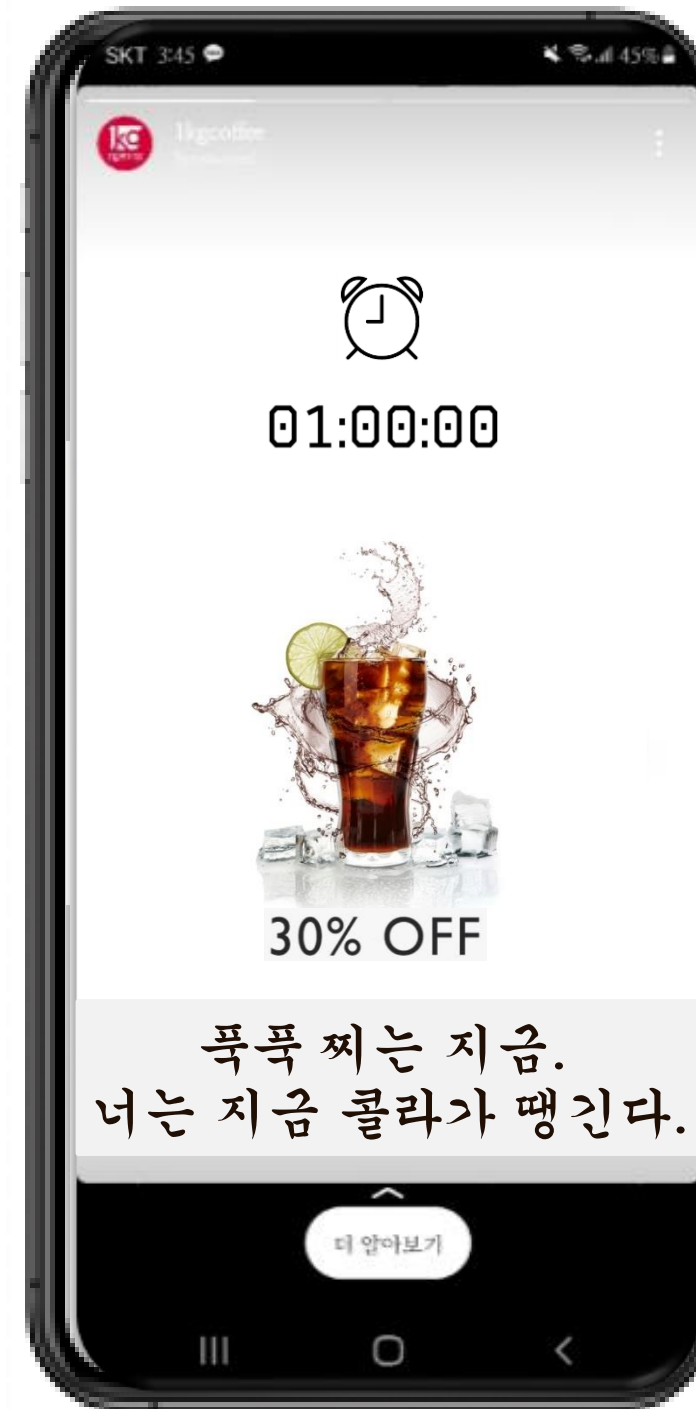
앵거리즘/작동원리

불쾌지수와 SNS 언급량이
양의 상관관계가 존재하는 품목들을 분류,



불쾌 지수의 정도에 따라 차등적인 할인율을 정하고,

사람들의 수요가 증가할 것으로 예상되는 시간에 타임 딜 삽입,
하단 <더 알아보기> 버튼으로 자사 앱/ 회원가입 유도



불쾌 지수의 정도에 따라 도출된
할인율을 반영한 바우처 제공

4. 서비스 활용방안/기대효과

앵거리즘/ 추가 활용 방안

🔍 불쾌지수와 SNS 언급량이 양의 상관관계를 가지지 않는 품목들?

💡 음의 상관관계를 가지는 품목

💡 불쾌지수가 낮아지면
SNS언급량이 많아지는 품목

💡 불쾌지수에 반비례하는
할인율을 적용

💡 상관관계 존재 X


💡 불쾌지수 외 다른 날씨 요소와
SNS언급량이 그 상관관계를 가지는 품목


💡 SNS언급량과 관련있는 날씨 요소를 모니터링,
증가할 것이라고 예측되는 시점에서
프로모션 진행

4. 서비스 활용방안/기대효과

앵거리즘/기대효과

 매출 증가 효과

 소비자와의 공감대 형성,
브랜드 친숙도 UP

 이미지 제고 효과,
고객 충성도 증대 기대

4. 서비스 활용방안/기대효과

SNS 활용방안



소비자들이 구매 전에 SNS에 접근하는 품목들에 대한 SNS 상 광고 집행 전략

- 자사 제품에 대한 주목도를 증가시키는 데에 집중
- SNS 언급 정도가 실제 구매로 이어지는 경우를 늘리고, 구매를 늘리는 데에 집중
- 커뮤니티 별 주 이용 타겟을 설정, 각각에 대한 마케팅 전략 수립

4. 서비스 활용방안/기대효과

SNS 활용방안/Instagram



시간대별 활성 ?

피크 타임

10:00 PM



출처: App Ape

📄 주 사용 계층/시간대 파악: 2/30대 여성, 오후 10시

📄 사진, 영상의 퀄리티, 설명 및 해시 태그가 제품 구입에 중요한 영향

💡 해당 품목에 대한 수요가 증가할 때 , 할인프로모션 광고를 SNS 사용 피크타임에 송출

💡 콘텐츠 제작, 인스타그램 셀럽 등을 통해 광고

4. 서비스 활용방안/기대효과

SNS 활용방안/맘 카페



📄 주 사용 계층: 3/40대 여성

📄 성공사례 분석: 마켓 컬리



- 🔍 충분한 구매력을 갖춘 30-40대 여성을 타겟, 맘 카페 등의 커뮤니티 적극 활용하여 마켓컬리만의 메시지 전달
- 🔍 인적 네트워크로 정보 공유하는 성향, 리뷰 적극 활용 후 구매
=> 리뷰 이벤트 시행

4. 서비스 활용방안/기대효과

SNS 활용방안/맘 카페



- 💡 가족 모두를 위해 제품을 구입하는 경우를 고려, 테마 별/컨셉 별 선착순 무료 샘플 증정 이벤트, 할인 프로모션 진행
- 💡 리뷰 이벤트 적극 시행, 자사의 메시지를 지속적으로 고객층에게 노출
- 💡 각 제품 군별 EDA를 진행, 수요에 영향을 많이 받는 날씨 요소를 모니터, 구매가 예측되기 전 일정 기간 지속적으로 노출, 구매 소요 시간 단축



닥터지 >
닥터지 그린마일드 업 선 에센스 한정기획(50ml+10ml+미니 에코백)
29,000원 **20,300원** 혜택 정보
세일 | 오늘드림
78명이 보고있어요
배송정보
일반배송 | 2,500원 (20,000 원 이상 무료배송)
올리브영 배송 | 평균 3일 이내 배송
오늘드림 | 2,500원 또는 5,000원
결제혜택
THE CJ 카드 추가10%할인
CJ ONE 포인트 최대 2% 적립 예상
그래스라

감사합니다.