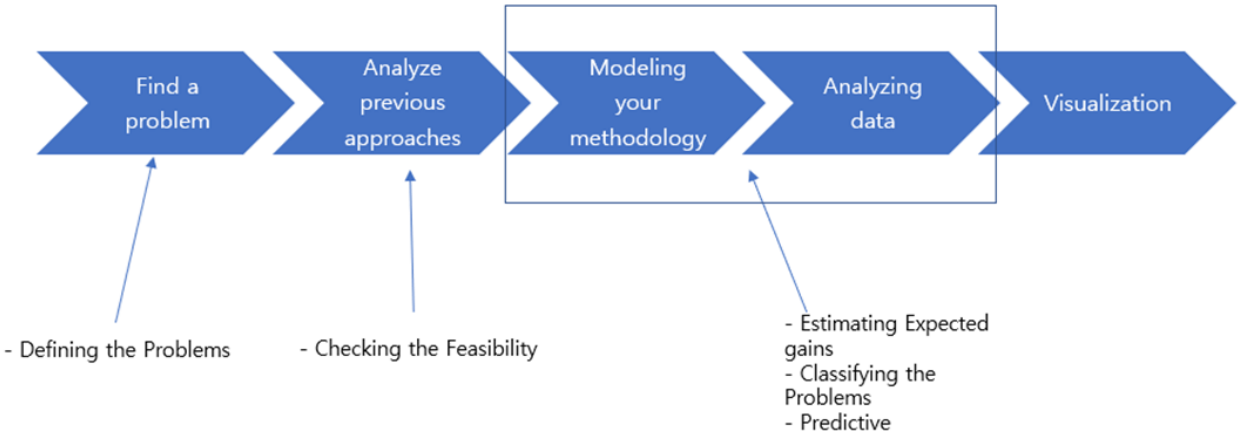
# 빅데이터 프로젝트 - 카페 추천시스템 Report

조 민수 **20161851**
이 상진 **20162191**
김 재민 **20172861**

## 0. Section



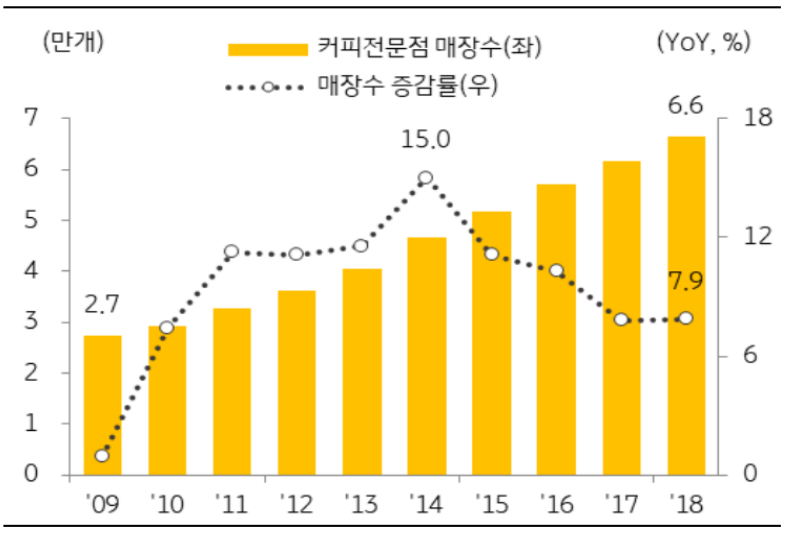

[ fig1. Index ]

## Ⅰ. Find a problem

### 주제 : Cafe Recommendation System

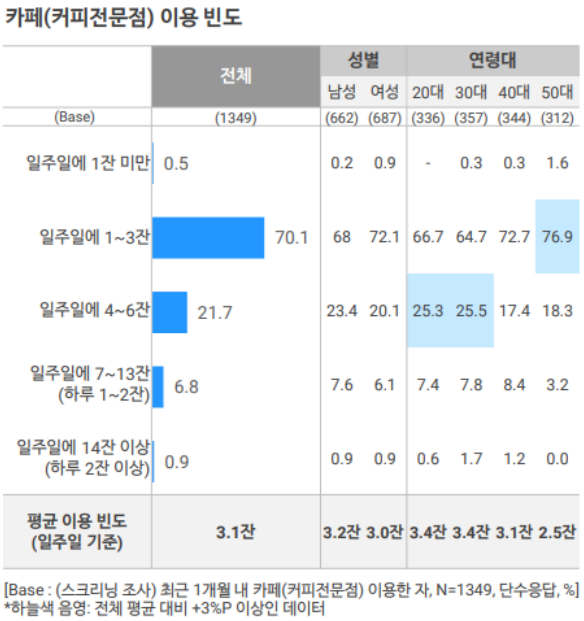
### 왜 카페인가?

21세기 현대 사회에서 카페는 대인관계에 있어 빠지지 않는 요소가 되었다.
카페 방문 목적도 점차 다양해지고 있으며, 그에 따라 사용자 맞춤식 카페추천 시스템을 고안하게 되었다.




[ fig2. 국내 커피 전문점 수 추이, 2019, KB금융지주, 커피전문점 현황 및 시장여건 분석 ]

2022년, 카페는 더 이상 커피를 마시는 공간이 아닌, 생활문화적 공간이라 칭해도 어색하지 않다.
2019년 기준 8만개에 달하는 커피 전문점들이 국내에서 영업을 하고 있으며, 종사자의 수도 20만명을 능가할 정도로 커피 산업의 양적 성장이 이루어졌다. 양질의 생활 문화공간을 원하는 수요의 증가로 카페의 수는 계속하여 늘어나고 있다(fig2).

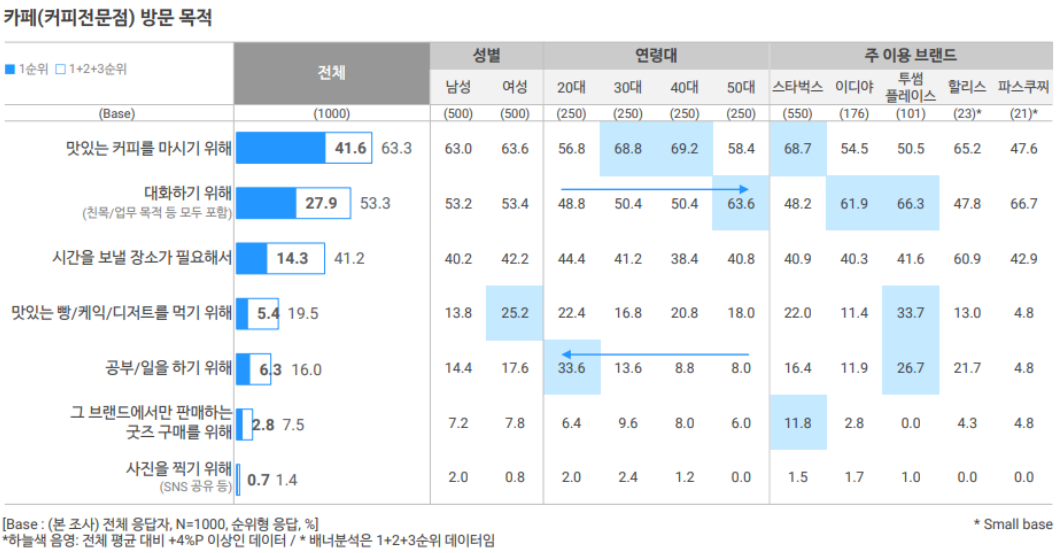


**카페(커피전문점) 이용 빈도**

| | 전체 | 성별 | | 연령대 | | | |
|---|---|---|---|---|---|---|---|
| | | 남성 | 여성 | 20대 | 30대 | 40대 | 50대 |
| (Base) | (1349) | (662) | (687) | (336) | (357) | (344) | (312) |
| 일주일에 1잔 미만 | 0.5 | 0.2 | 0.9 | - | 0.3 | 0.3 | 1.6 |
| 일주일에 1~3잔 | 70.1 | 68 | 72.1 | 66.7 | 64.7 | 72.7 | 76.9 |
| 일주일에 4~6잔 | 21.7 | 23.4 | 20.1 | 25.3 | 25.5 | 17.4 | 18.3 |
| 일주일에 7~13잔 (하루 1~2잔) | 6.8 | 7.6 | 6.1 | 7.4 | 7.8 | 8.4 | 3.2 |
| 일주일에 14잔 이상 (하루 2잔 이상) | 0.9 | 0.9 | 0.9 | 0.6 | 1.7 | 1.2 | 0.0 |
| **평균 이용 빈도 (일주일 기준)** | **3.1잔** | **3.2잔** | **3.0잔** | **3.4잔** | **3.4잔** | **3.1잔** | **2.5잔** |

[Base : (스크리닝 조사) 최근 1개월 내 카페(커피전문점) 이용한 자, N=1349, 단수응답, %]
*하늘색 음영: 전체 평균 대비 +3%P 이상인 데이터

[ fig3. 카페 방문 통계, 2020, OpenSurvey, 카페 이용 트렌드 리포트 (20~50대 표본 800인) ]



**카페(커피전문점) 방문 목적**

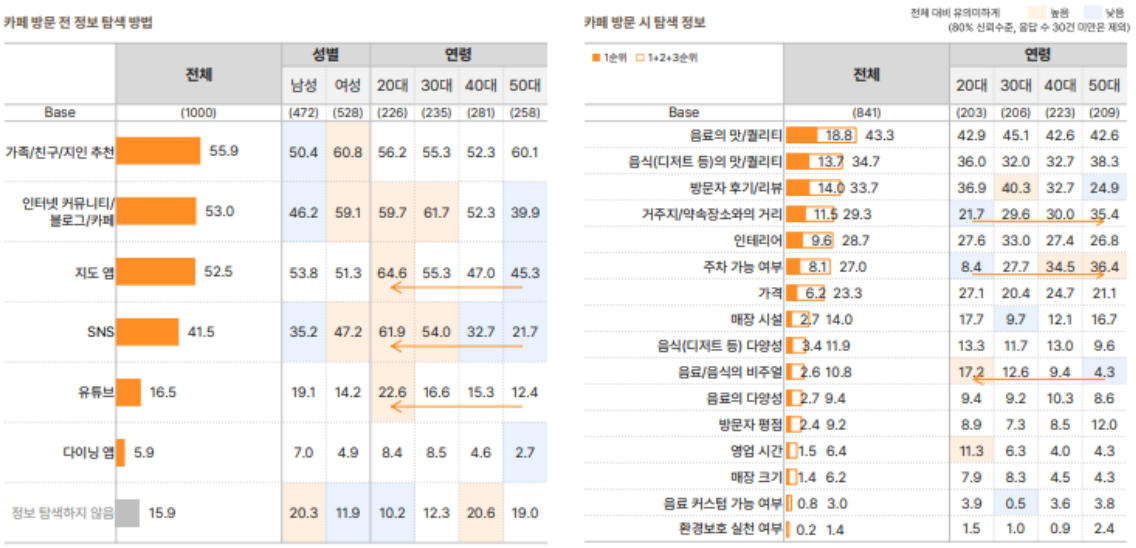
| ■1순위 □1+2+3순위 | 전체 | 성별 | | 연령대 | | | | 주 이용 브랜드 | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | 남성 | 여성 | 20대 | 30대 | 40대 | 50대 | 스타벅스 | 이디야 | 투썸 플레이스 | 할리스 | 파스쿠찌 |
| (Base) | (1000) | (500) | (500) | (250) | (250) | (250) | (250) | (550) | (176) | (101) | (23)* | (21)* |
| 맛있는 커피를 마시기 위해 | 41.6 63.3 | 63.0 | 63.6 | 56.8 | 68.8 | 69.2 | 58.4 | 68.7 | 54.5 | 50.5 | 65.2 | 47.6 |
| 대화하기 위해 (친목/업무 목적 등 모두 포함) | 27.9 53.3 | 53.2 | 53.4 | 48.8 | 50.4 | 50.4 | 63.6 | 48.2 | 61.9 | 66.3 | 47.8 | 66.7 |
| 시간을 보낼 장소가 필요해서 | 14.3 41.2 | 40.2 | 42.2 | 44.4 | 41.2 | 38.4 | 40.8 | 40.9 | 40.3 | 41.6 | 60.9 | 42.9 |
| 맛있는 빵/케익/디저트를 먹기 위해 | 5.4 19.5 | 13.8 | 25.2 | 22.4 | 16.8 | 20.8 | 18.0 | 22.0 | 11.4 | 33.7 | 13.0 | 4.8 |
| 공부/일을 하기 위해 | 6.3 16.0 | 14.4 | 17.6 | 33.6 | 13.6 | 8.8 | 8.0 | 16.4 | 11.9 | 26.7 | 21.7 | 4.8 |
| 그 브랜드에서만 판매하는 굿즈 구매를 위해 | 2.8 7.5 | 7.2 | 7.8 | 6.4 | 9.6 | 8.0 | 6.0 | 11.8 | 2.8 | 0.0 | 4.3 | 4.8 |
| 사진을 찍기 위해 (SNS 공유 등) | 0.7 1.4 | 2.0 | 0.8 | 2.0 | 2.4 | 1.2 | 0.0 | 1.5 | 1.7 | 1.0 | 0.0 | 0.0 |

[Base : (본 조사) 전체 응답자, N=1000, 순위형 응답, %]
*하늘색 음영: 전체 평균 대비 +4%P 이상인 데이터 / * 배너분석은 1+2+3순위 데이터임

* Small base

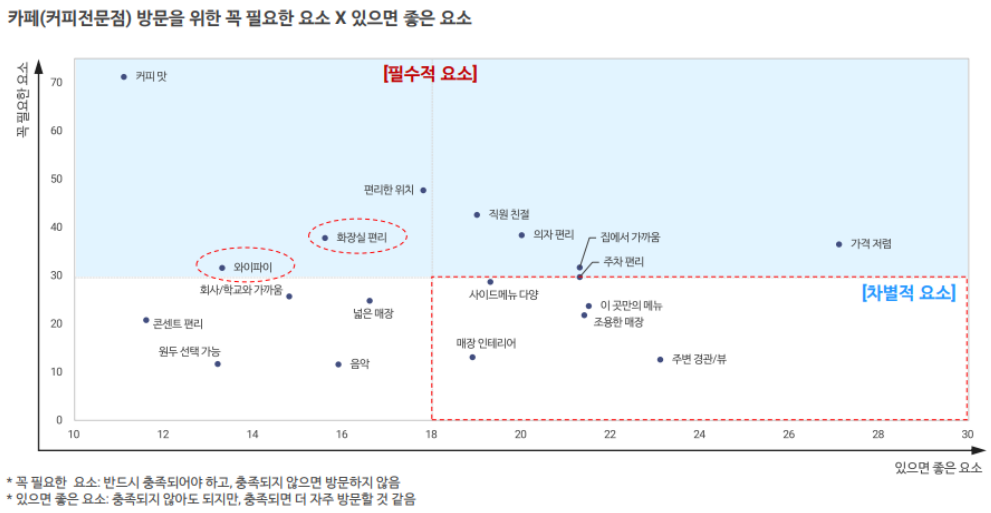
[ fig4. 카페 방문 목적, 2020, OpenSurvey, 카페 이용 트렌드 리포트 (20~50대 표본 800인) ]

20 ~ 50대를 대상으로 진행한 2020년의 설문에서 확인할 수 있듯, 설문모집단의 91.8%는 일주일에 최대 6잔의 커피를 카페에서 구매하는 것으로 나타난다(fig3).
위와 같은 커피 산업, 다시말해 커피 전문점의 양적 성장은 본래의 기능인 '커피'를 즐기는 것을 넘어, 모임, 데이트, 공부, SNS포스팅 등 여러 사회적 내인들을 배경으로 이루어졌음을 알 수 있다(fig4).




[ fig5. 카페 방문을 위한 정보 탐색 경로, 2022, OpenSurvey, 20~50대 표본 800인 ]

위의 통계에서 확인할 수 있듯이 카페는 대한민국 생활에 있어 빠질 수 없는 요소가 되었다. 카페가 하나의 생활문화적 공간으로 대표되며, 카페에서 기대하는 요구 사항의 다변화가 이루어졌다. 때문에, 기존의 브랜드 커피를 주로 소비하는 시장에서 자신의 요구에 부합하는 카페를 찾는 경우가 많아졌으며, 2022년의 통계자료를 통해 이용자 주도적인 카페 방문을 위한 다변화된 정보 탐색이 이루어지고 있음을 확인할 수 있다(fig5). 특히 온라인 상에서 검색하여 SNS상의 리뷰, 포털 사이트에서의 평점 정보, 유튜브 등의 매체를 통해 찾게 되는 경우가 다수이다.
하지만 현재 사용자의 세부적 요구에 부합하는 특정한 카페를 추천해 줄 수 있는 추천시스템이 구축되어 있지 않다. 따라서 사용자의 목적에 맞는 카페를 추천할 수 있도록 하는 카페 추천시스템을 제안한다.




[ fig6. 카페 방문시 판단 항목 설문, 2020, OpenSurvey, 20~50대 표본 800인 ]

# Ⅱ. Analyze previous approaches

## 기존에 사용되는 일반적인 접근방법

워드 임베딩과 이미지 임베딩을 통한 Hybrid filtering 카페 추천 시스템

Hybrid filtering = content-based filtering + collaborative filtering

이미지 기반 카페 추천 시스템 (Cafe-in)

## 기존에 사용되는 일반적인 알고리즘

KNN collaborative filtering

Content-Based Filtering

Collaborative Filtering

DCNN;Deep Convolutional Neural Network (image processing & tagging)

## 선행연구 정리(Reference)

* **LightFM (Hybrid filtering)**
  - Github Repo: https://github.com/lyst/lightfm
  - Paper: Kula, Maciej. "Metadata embeddings for user and item cold-start recommendations." *arXiv preprint arXiv:1507.08439* (2015).
  - Cold start 상황에서 수용될 수 있는 수준의 추천을 마련해 주는 추천시스템을 고려하기 위해 작성된 논문으로 Lyst라는 국제 온라인 패션 회사에서 고안한 기술이다. 많은 item이 존재하며 새로운 상품이 자주 등록되고(cold-start가 잦고), 다수의 고객이 새로 가입한 신규 고객(cold-start)이라는 점에서 LightFM은 좋은 성능을 내는 것으로 주목받고 있다. 본 모델은 content-based filtering과 collaborative filtering의 장점을 결합한 모델이며, 아래와 같은 특징들을 지닌다.

1) 학습 데이터에서 collaborative한 데이터들과 user-item feature을 모두 사용하였다.
2) 생성된 embedding 벡터가 feature들에 대한 핵심적인 의미 정보를 포함하며 tag를 통한 추천 등에서 중요하게 사용될 수 있다.

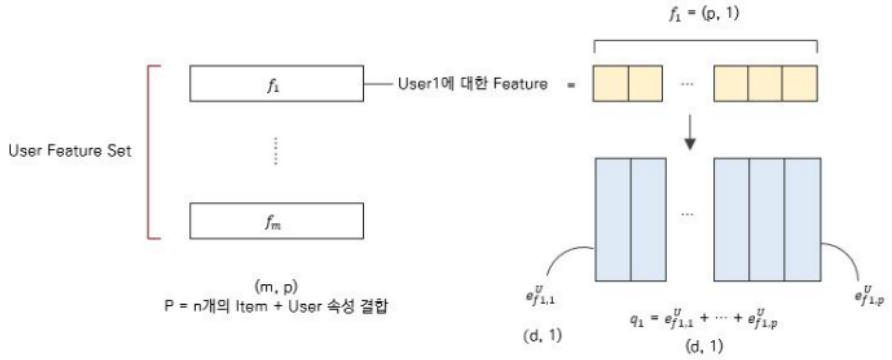
모델은 아래와 같이 구성된다.

| | |
|---|---|
| $U$ | User의 집합 |
| $I$ | Item의 집합 |
| $F^U$ | User Feature의 집합 |
| $F^I$ | Item Feature의 집합 |
| $f_u$ | $u$라는 User의 features, $f_u \subset F^U$ |
| $f_i$ | $i$라는 Item의 features, $f_i \subset F^I$ |
| $e_f^U$ | $f_u$의 각 User feature들에 대한 d-차원 Embedding 벡터 |
| $e_f^I$ | $f_i$의 각 Item feature들에 대한 d-차원 Embedding 벡터 |
| $b_f^U$ | $u$라는 User의 features, $f_u \subset F^U$ |
| $b_f^I$ | $i$라는 Item의 features, $f_i \subset F^I$ |

$$q_u = \sum_{j \in f_u} e_j^U \quad p_i = \sum_{j \in f_i} e_j^I$$

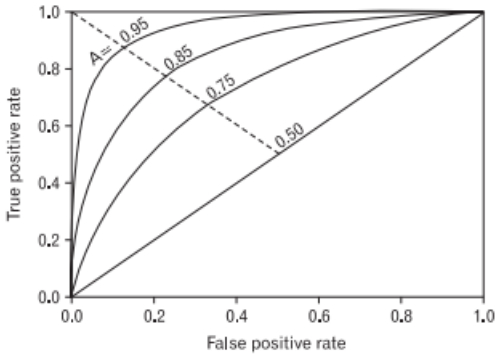
$$b_u = \sum_{j \in f_u} b_j^U \quad b_i = \sum_{j \in f_i} b_j^I$$

유저(user)에 대한 잠재 벡터는 해당 유저가 가지는 feature들의 잠재 벡터들의 합으로 구성되며 아이템(item)에 대한 잠재 벡터 또한 유저의 경우와 동일한 방식으로 계산된다. Bias의 경우는 위 수식과 같이 계산된다.




모델의 구조를 모식화 해 본다면 위의 그림과 같다. 해당 그림은 유저 feature에 대한 예시만을 든 것이다. 유저와 아이템 각각의 feature와 둘 간의 상호작용을 모두 고려할 수 있는 구조이며, 이 특징을 통해 cold-start 문제를 완화할 수 있다.

평가지표로는 AUC(Area Under the Curve)를 일반적으로 사용하며, ROC(Receiver Operating Charateristic) 곡선 아래의 면적을 뜻한다. AUC는 이진 분류(binary classification) 또는 다중 클래스 분류(multi-class classification)와 같은 문제에서 참/거짓 라벨의 분류 성능을 측정하는 지표이다.



# Ⅲ. Modeling your methodology

## Our Project's Methodology

< 크롤링 진행 이전 목표 >

- LightFM (Hybrid filtering)을 적용하고 AUC를 통해 정량적인 평가를 진행한다.

< 크롤링 진행 이후 목표 >

- Content-based filtering을 다양한 유사도 metric을 적용하여 진행.
  적용하는 유사도 metric은 Euclidean distance vs Pearson correlation vs cosine similarity.
- Cosine similarity 기반 정량 평가 진행.
- RMSE(Root Mean Squared Error) 기반 정량 평가 진행.

생각보다 크롤링 작업에 소요된 시간이 많아 크롤링 진행 이전 목표에서 수정이 이루어졌습니다.

## 기존의 추천시스템과 무엇이 다른가?

카페에 대한 전문적인 추천시스템이라는 컨텐츠를 찾아보기 어려웠다.
다소 흔하거나 찾아보기 쉬운 일반적인 컨텐츠보다 독특한 항목에 대하여 오로지 카페에 집중된
추천시스템이라는 점에 대하여 차별성을 가진다고 할 수 있다.
단순히 '인기많은 카페' 라기보다 '지금, 당신이 원하는 카페가 이런것인가요?'라는 목적성을 갖고있다.

# Ⅳ. Data Processing

## 데이터 수집(Crawling)

네이버 지도 내에, 별점 외에 사용자들이 평가해 놓은 지표데이터가 있었다.
지표자체가 labeling되어있어, 이를 적극 활용하기로 결정하였다.

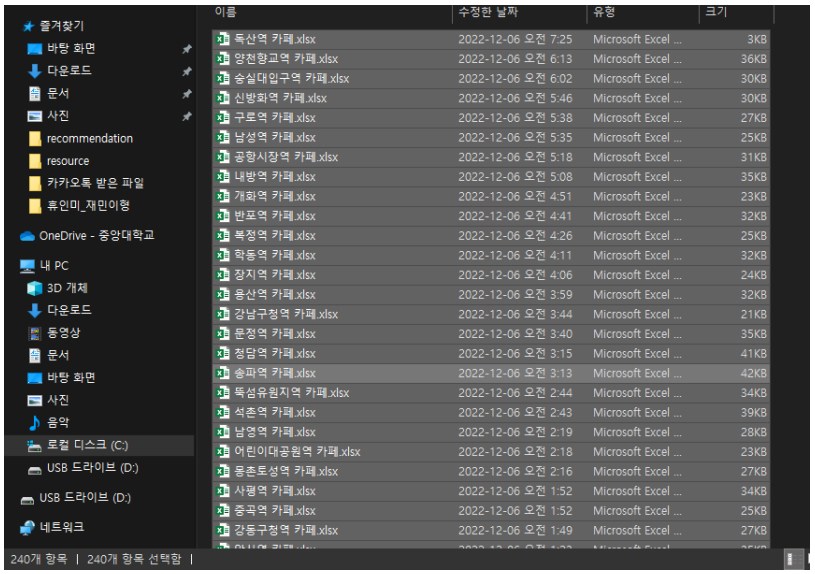

[ fig7. 활용할 라벨링 데이터의 예시, 네이버 지도, 상도동 리릿커피 ]

서울 전역에 대한 카페를 크롤링 하는데에 있어서, 지하철역을 기준으로 크롤링하였다.
크롤링 대상 웹사이트는 네이버 지도이다. 네이버 지도 내에서 카페 데이터를 크롤링 후,
데이터 내 Labeling 된 평가지표를 추출하였다.

```
keyword = pyautogui.prompt("검색어를 입력하십시오")
wb = openpyxl.Workbook()
ws = wb.create_sheet()
ws.append(["영업중", "가게명", "특징1", "특징1수", "특징2", "특징2수", "특징
"특징8", "특징8수", "특징9", "특징9수", "특징10", "특징10수", "특징11", "특
"특징16", "특징16수", "특징17", "특징17수", "특징18", "특징18수", "특징19",

browser = webdriver.Chrome("./chromedriver.exe")
browser.get("https://map.naver.com/v5/")
browser.implicitly_wait(10)
browser.maximize_window()

search = browser.find_element_by_css_selector("input.input_search")
search.click()
time.sleep(1)
search.send_keys("강남역 카페")
time.sleep(1)
search.send_keys(Keys.ENTER)
time.sleep(2)
```

[ fig8. 크롤링 코드 일부 ]

위 스크린샷은 크롤링 코드의 일부분으로 실제 크롤링은 서울시에 위치한 역사명을 받아, 자동화된 크롤링을 수행하게 된다.




[ fig9. 기본적인 크롤링이 진행된 이후 raw data 집합체의 모습 ]

서울시 내에 존재하는 총 279개 역사 중 240개 역사에 대한 크롤링을 성공적으로 진행하였다. 현재 수집된 데이터 만으로도 충분히 추천 시스템을 구현하는데에 무리가 없지만, 추천의 정확도를 높이기 위해 남은 39개의 역사에 대하여 웹 클래스에 대한 구체화된 구분을 추가하여 자동화된 크롤링을 진행중이다.

인접한 역에서 중복된 데이터가 발생하지만, 후에 중복되는 데이터에 대한 처리를 진행하였다.

기본적인 크롤링이 진행된 이후의 데이터 모습은 아래의 그림과 같다.

| | A | B | C | D | E | F | G | H | I | J | K | L | M | N | O | P | Q | R | S | T | U | V | W |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 0 | 히피스베이 | 빵이 맛있( | 681 | 친절해요 | 249 | 특별한 메 | 245 | 가성비가 | 121 | 매장이 청 | 74 | 음료가 맛 | 25 | 커피가 맛 | 15 | 인테리어 | 11 | 대화하기 | 6 | 좌석이 편 | 4 | 사진이 잘 |
| 2 | 0 | 아띠랑스 | 커피가 맛 | 470 | 인테리어 | 284 | 디저트가 | 271 | 음료가 맛 | 199 | 매장이 청 | 190 | 친절해요 | 156 | 사진이 잘 | 120 | 대화하기 | 104 | 화장실이 | 88 | 뷰가 좋아 | 86 | 좌석이 편 |
| 3 | 0 | 솔로핸드 | 커피가 맛 | 117 | 친절해요 | 91 | 음료가 맛 | 54 | 인테리어 | 54 | 매장이 청 | 37 | 가성비가 | 35 | 화장실이 | 30 | 디저트가 | 29 | 대화하기 | 28 | 사진이 잘 | 19 | 아늑해요 |
| 4 | 0 | 다정도 병( | 음료가 맛 | 135 | 디저트가 | 102 | 특별한 메 | 92 | 친절해요 | 79 | 인테리어 | 63 | 매장이 청 | 54 | 대화하기 | 54 | 좌석이 편 | 44 | 커피가 맛 | 39 | 뷰가 좋아 | 23 | 화장실이 |
| 5 | 0 | 쉬즈베이글 | 가성비가 | 37 | 친절해요 | 35 | 디저트가 | 32 | 커피가 맛 | 31 | 특별한 메 | 18 | 음료가 맛 | 17 | 매장이 청 | 14 | 인테리어 | 4 | 대화하기 | 3 | 사진이 잘 | 2 | 뷰가 좋아. |
| 6 | 0 | 블랙다운커 | 커피가 맛 | 503 | 친절해요 | 302 | 음료가 맛 | 290 | 특별한 메 | 178 | 매장이 청 | 98 | 가성비가 | 68 | 디저트가 | 39 | 인테리어 | 39 | 대화하기 | 38 | 집중하기 | 31 | 좌석이 편 |
| 7 | 0 | 일구구일스 | 빵이 맛있( | 87 | 친절해요 | 59 | 특별한 메 | 43 | 매장이 청 | 33 | 가성비가 | 19 | 인테리어 | 15 | 음료가 맛 | 4 | 커피가 맛 | 3 | 사진이 잘 | 2 | 집중하기 | 2 | 주차하기 |

[ fig10. 크롤링 직후 데이터 ]

```python
import openpyxl
import os

def getpath():
    _path = os.getcwd()
    _path = os.chdir('./files')
    return _path

if __name__ == "__main__":
    file_list = os.listdir(getpath())
    number_files = len(file_list)
    print("#of files to integrate: ", number_files)

    wb_new = openpyxl.Workbook()
    ws_new = wb_new.active

    for _fi, _file  in enumerate(file_list):
        print(_fi+1, "/", len(file_list), "번째 파일 합치는 중...")
        wb_old = openpyxl.load_workbook(_file)
        ws_old = wb_old.worksheets[0]

        for row in ws_old.iter_rows(min_row=2):
            row_with_values = [cell.value for cell in row]
            # print(type(row_with_values))  # list
            # print(type(row_with_values[0]))
            if row_with_values[0] is None:      # 공백 데이터 제거 (NoneType으로 나옴)
                continue
            else:                               # 크롤링 해온 데이터라면 합쳐줌
                ws_new.append(row_with_values)


    wb_new.save("integrated.xlsx")
```

[ fig11. 크롤링을 통해 추출된 데이터들을 한 개의 파일로 integrate 하는 코드 ]

위 스크린샷은 크롤링을 통해 추출된 데이터들을 한 개의 파일로 합쳐주는 코드부의 일부이다.
이후 한글로 입력된 feature들의 간소화를 위해 데이터 라벨의 재가공 작업(relabeling)을 진행하였다.
데이터 라벨의 재 가공은 엑셀의 바꾸기 기능을 통해 진행하였다.

| | | | | |
|---|---|---|---|---|
| 2 | | 친절해요 | | friendly |
| 3 | | 디저트가 맛있어요 | | dessert |
| 4 | | 인테리어가 멋져요 | | interior |
| 5 | | 매장이 청결해요 | | clean |
| 6 | | 커피가 맛있어요 | | coffee |
| 7 | | 특별한 메뉴가 있어요 | | special |
| 8 | | 빵이 맛있어요 | | bread |
| 9 | | 사진이 잘 나와요 | | photo |

[ fig12. 데이터 라벨 재가공 작업을 위한 mapping table의 일부 ]

위 그림과 같은 매핑 테이블을 통해 크롤링 직후 다소 불필요하게 긴 데이터의 Feature를 구분짓기가 쉽도록 재구성하였다.
매핑 테이블을 통해 재가공된 데이터들을 기반으로 크롤링 이후의 데이터에 대하여 중복 데이터제거를 포함한 데이터 전처리 작업을 진행하였다. 데이터 전처리 작업에는 호텔이나 팝업스토어와 같은 노이즈 데이터를 제거하는 작업도 포함되었다.

| 열1 | friendly | dessert | interior | clean | coffee | special | bread | photo | non_co | tea | fresh | seat | talk | concen | view | toilet | parking | plenty |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 히피스베이글 | 249 | 0 | 11 | 74 | 15 | 245 | 681 | 4 | 25 | 0 | 0 | 4 | 6 | 1 | 4 | 3 | 1 | 0 |
| 아띠랑스 | 156 | 271 | 284 | 190 | 470 | 45 | 0 | 120 | 199 | 0 | 0 | 85 | 104 | 15 | 86 | 88 | 84 | 0 |
| 슬로핸드 | 91 | 29 | 54 | 37 | 117 | 9 | 0 | 19 | 54 | 5 | 0 | 4 | 28 | 17 | 13 | 30 | 3 | 0 |
| 다정도 병인 양 | 79 | 102 | 63 | 54 | 39 | 92 | 0 | 12 | 135 | 0 | 0 | 44 | 54 | 10 | 23 | 19 | 2 | 0 |
| 쉬즈베이글 먹대점 | 35 | 32 | 4 | 14 | 31 | 18 | 0 | 2 | 17 | 0 | 0 | 1 | 3 | 1 | 2 | 0 | 0 | 0 |
| 블랙다운커피 | 302 | 39 | 39 | 98 | 503 | 178 | 0 | 6 | 290 | 0 | 0 | 27 | 38 | 31 | 9 | 5 | 3 | 0 |
| 일구구일스콘 | 59 | 0 | 15 | 33 | 3 | 43 | 87 | 2 | 4 | 0 | 0 | 0 | 0 | 2 | 0 | 0 | 1 | 0 |
| 몽브루 | 153 | 234 | 175 | 0 | 368 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 297 | 0 | 0 | 0 |

[ fig12. 라벨링 후 데이터 ]

카페의 종류에 따라 많은 feature가 존재했다.(feature수 : 103개) - 일반카페, 스터디카페, 테마카페 etc...
결과적으로 우리가 구현하고자하는 시스템에 적합하지 못하다고 판단되어 수작업을  통해 카페의 종류에 따라 지표를 통폐합하였다.

| | A | B | C | D | E | F | G | H |
|---|---|---|---|---|---|---|---|---|
| 1 | | | | | | | | |
| 2 | | | 친절해요 | 디저트가 맛있어요 | 인테리어가 멋져요 | 매장이 청결해요 | 커피가 맛있어요 | 특별한 메뉴가 있어요 |
| 3 | | | 설명이 자세해요 | | | | | |
| 4 | | | 조보자에게도 적합해요 | | | | | |
| 5 | | | 추천을 잘해줘요 | | | | | |

[ fig13. feature 통폐합 ]

데이터 분석을 위해 데이터 값을 별도로 정규화하였다. 데이터 값의 정규화를 위해 각 카페의 최대값을 기준으로 각 카페의 데이터를 나누어 주었다.
이를 통해 카페가 가질 수 있는 최대 값이 1이 되도록 만들어 주었다.

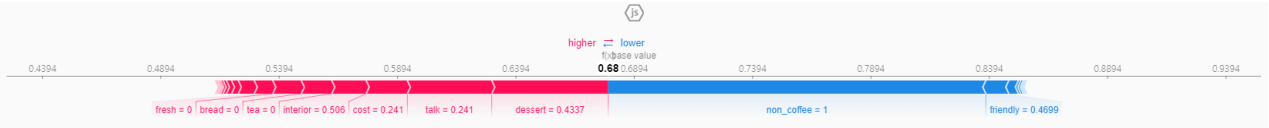
| | A | B | C | D | E | F | G | H | I | J | K | L | M | N | O | P | Q | R | S | T | U | V | W |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | name | friendly | dessert | interior | clean | coffee | special | bread | photo | non_coffe | tea | fresh | seat | talk | concentra | view | toilet | parking | plenty | food | big | cost | concept |
| 2 | 히피스베이 | 0.365639 | 0 | 0.016153 | 0.108664 | 0.022026 | 0.359765 | 1 | 0.005874 | 0.036711 | 0 | 0 | 0.005874 | 0.008811 | 0.001468 | 0.005874 | 0.004405 | 0.001468 | 0 | 0 | 0 | 0.17768 | 0 |
| 3 | 아띠랑스 | 0.331915 | 0.576596 | 0.604255 | 0.404255 | 1 | 0.095745 | 0 | 0.255319 | 0.423404 | 0 | 0 | 0.180851 | 0.221277 | 0.031915 | 0.182979 | 0.187234 | 0.178723 | 0 | 0 | 0 | 0.06383 | 0 |
| 4 | 슬로핸드 | 0.777778 | 0.247863 | 0.461538 | 0.316239 | 1 | 0.076923 | 0 | 0.162393 | 0.461538 | 0.042735 | 0 | 0.034188 | 0.239316 | 0.145299 | 0.111111 | 0.25641 | 0.025641 | 0 | 0 | 0 | 0.299145 | 0 |
| 5 | 다정도 병 | 0.585185 | 0.755556 | 0.466667 | 0.4 | 0.288889 | 0.681481 | 0 | 0.088889 | 1 | 0 | 0 | 0.325926 | 0.4 | 0.074074 | 0.17037 | 0.140741 | 0.014815 | 0 | 0 | 0 | 0.059259 | 0 |
| 6 | 쉬즈베이글 | 0.945946 | 0.864865 | 0.108108 | 0.378378 | 0.837838 | 0.486486 | 0 | 0.054054 | 0.459459 | 0 | 0 | 0.027027 | 0.081081 | 0.027027 | 0.054054 | 0 | 0 | 0 | 0 | 0 | 1 | 0 |
| 7 | 블랙다운커 | 0.600398 | 0.077535 | 0.077535 | 0.194831 | 1 | 0.353877 | 0 | 0.011928 | 0.576541 | 0 | 0 | 0.053678 | 0.075547 | 0.06163 | 0.017893 | 0.00994 | 0.005964 | 0 | 0 | 0 | 0.135189 | 0 |
| 8 | 일구구일스 | 0.678161 | 0 | 0.172414 | 0.37931 | 0.034483 | 0.494253 | 1 | 0.022989 | 0.045977 | 0 | 0 | 0 | 0 | 0.022989 | 0 | 0 | 0.011494 | 0 | 0 | 0 | 0.218391 | 0 |
| 9 | 몽브루 | 0.415761 | 0.63587 | 0.475543 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0.807065 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 10 | 본죽&비빔 | 0.392523 | 0 | 0.056075 | 0.411215 | 0 | 0.093458 | 0 | 0 | 0 | 0 | 0.280374 | 0 | 0 | 0 | 0.018692 | 0.009346 | 0.009346 | 0.17757 | 1 | 0.252336 | 0.084112 | 0 |
| 11 | 크림 디저 | 0.726457 | 1 | 0.130045 | 0.475336 | 0.264574 | 0.363229 | 0 | 0.040359 | 0.192825 | 0 | 0 | 0.004484 | 0.017937 | 0.008969 | 0.004484 | 0.008969 | 0.013453 | 0 | 0 | 0 | 0.403587 | 0 |
| 12 | 카페 라티 | 0.461538 | 0.461538 | 0.230769 | 0.410256 | 1 | 0.025641 | 0.076923 | 0.025641 | 0.538462 | 0 | 0 | 0.128205 | 0.051282 | 0.025641 | 0 | 0.153846 | 0.051282 | 0 | 0 | 0 | 0 | 0 |
| 13 | 도너즈윤 | 0.486842 | 1 | 0.039474 | 0.210526 | 0.197368 | 0.342105 | 0 | 0.078947 | 0.078947 | 0 | 0 | 0.026316 | 0 | 0 | 0.013158 | 0 | 0 | 0 | 0 | 0 | 0.157895 | 0 |
| 14 | 쿤데라 | 0.744681 | 0.574468 | 0.06383 | 0.425532 | 1 | 0.148936 | 0 | 0.106383 | 0.404255 | 0 | 0 | 0.212766 | 0.212766 | 0.042553 | 0 | 0.12766 | 0.021277 | 0 | 0 | 0 | 0.255319 | 0 |
| 15 | 비르케 | 0.429907 | 0.457944 | 0.336449 | 0.345794 | 1 | 0.074766 | 0 | 0.140187 | 0.364486 | 0 | 0 | 0.11215 | 0.345794 | 0.093458 | 0.35514 | 0.17757 | 0.065421 | 0 | 0 | 0 | 0.046729 | 0 |
| 16 | 메가MGC | 0.444853 | 0.205882 | 0.073529 | 0.316176 | 1 | 0.079044 | 0 | 0.025735 | 0.545956 | 0 | 0 | 0.220588 | 0.172794 | 0.082721 | 0.036765 | 0.064338 | 0.007353 | 0 | 0 | 0 | 0.854779 | 0 |
| 17 | 크림앤버터 | 0.473684 | 1 | 0.210526 | 0.447368 | 0.421053 | 0.5 | 0 | 0.052632 | 0.184211 | 0 | 0 | 0.026316 | 0.026316 | 0 | 0.026316 | 0.026316 | 0.026316 | 0 | 0 | 0 | 0.236842 | 0 |
| 18 | 오븐, 비스 | 0.690476 | 1 | 0.095238 | 0.452381 | 0.238095 | 0.285714 | 0 | 0.02381 | 0.190476 | 0 | 0 | 0.02381 | 0.047619 | 0.02381 | 0 | 0 | 0 | 0 | 0 | 0 | 0.214286 | 0 |
| 19 | 커피배우 | 0.386364 | 0.25 | 0.204545 | 0.068182 | 1 | 0.386364 | 0 | 0.045455 | 0.568182 | 0 | 0 | 0.045455 | 0.090909 | 0 | 0 | 0.022727 | 0 | 0 | 0 | 0 | 0.204545 | 0 |
| 20 | 무너미 | 0.209677 | 1 | 0.258065 | 0.185484 | 0.516129 | 0.354839 | 0 | 0.040323 | 0.217742 | 0 | 0 | 0.016129 | 0.193548 | 0.056452 | 0.016129 | 0.016129 | 0.016129 | 0 | 0 | 0 | 0.040323 | 0 |
| 21 | 이디야 4.1 | 0.44186 | 0.232558 | 0.069767 | 0.232558 | 1 | 0.116279 | 0 | 0.023256 | 0.534884 | 0 | 0 | 0.116279 | 0.232558 | 0.116279 | 0.069767 | 0.093023 | 0.069767 | 0 | 0 | 0 | 0.395349 | 0 |
| 22 | 파리바게 | 0.497608 | 0 | 0.038278 | 0.210526 | 0.090909 | 0.064593 | 1 | 0.023923 | 0.088517 | 0 | 0 | 0.016746 | 0.028708 | 0.014354 | 0.014354 | 0.009569 | 0.002392 | 0 | 0 | 0 | 0.086124 | 0 |

[ fig14. data 값 정규화 ]

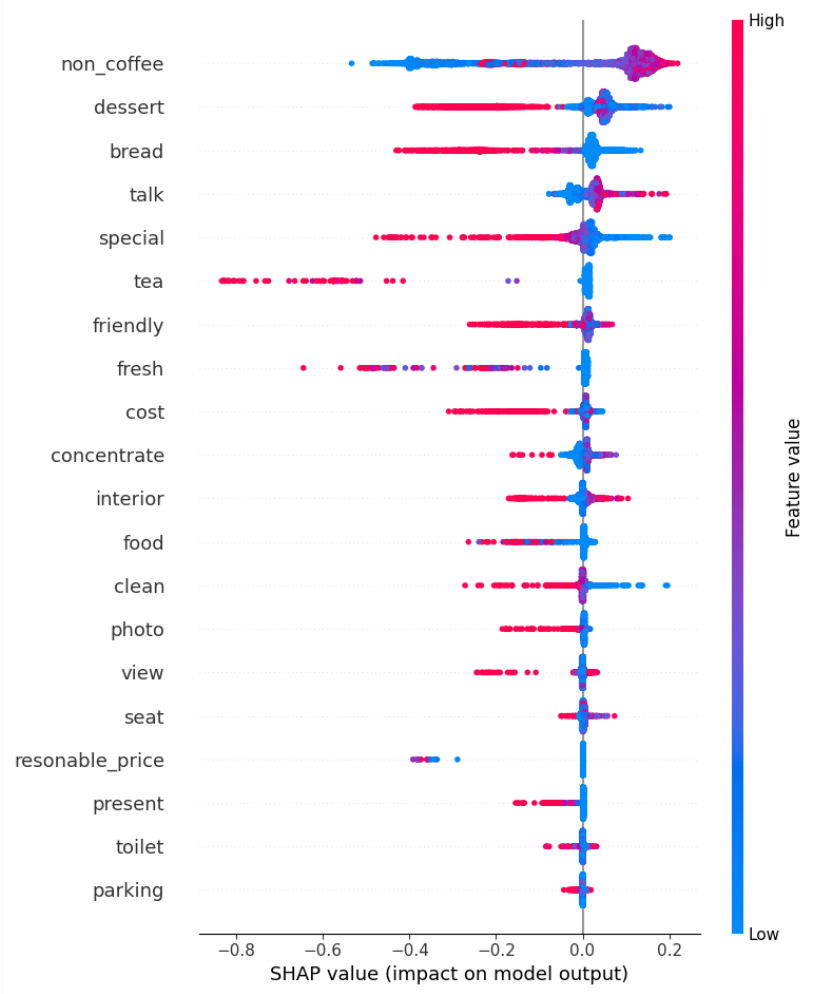
# Ⅴ. Analysis & Visualization

## 데이터 분석 및 시각화

데이터 처리 완료 후 예측값에 대한 각 변수들의 중요도와 예측값과 영향력을 측정하기 위해 SHAP를 사용하였다. SHAP(SHapley Additive exPlanations)은 일반적으로 기계 학습 모델의 결과치(출력치)를 설명하기 위한 접근 방식이며, 게임이론에서 유래하였다.
학습 모델에 상관없이 특정 변수의 존재 유무에 따른 개별적인 예측치를 실제 존재하는 변수들의 영향력의 합계로서 표현할 수 있다고 알려져있다. 또한, 영향력을 측정하는 값으로 Shap Value(Shapley Value)를 사용한다.




[ fig15. 커피의 맛을 기준으로 하는 한 개 변수의 Shap Value 요약 ]




[ fig16. 커피의 맛을 기준으로 하는 모든 변수의 Shap Value 요약 ]

위의 그림은 커피의 맛을 기준으로 모든 변수의 Shap Value를 요약한 것이다.

붉은 색을 띌 수록 항목은 커피의 맛을 선호한 경우와 양의 상관관계를 지니는 것이고,

파란색을 띌 수록 음의 상관관계를 지니는 것이다.

Coffee를 기준으로한 (Coffee vs ~), 몇 가지 변수에 대한 해석을 진행해본다면 다음과 같다.

vs cost : 메뉴의 가성비가 좋다고 대답한 사람들은 커피의 맛에 대한 만족감을 표현하지 않는 경향이 있다.

vs non_coffee : 커피 외 음료를 선호한 사람들은 커피의 맛에 대한 만족감을 표현하는 경향이 낮았다.

vs bread : 빵이 맛있다고 표현한 경우에는 커피가 맛있다고 표현하지 않았다.

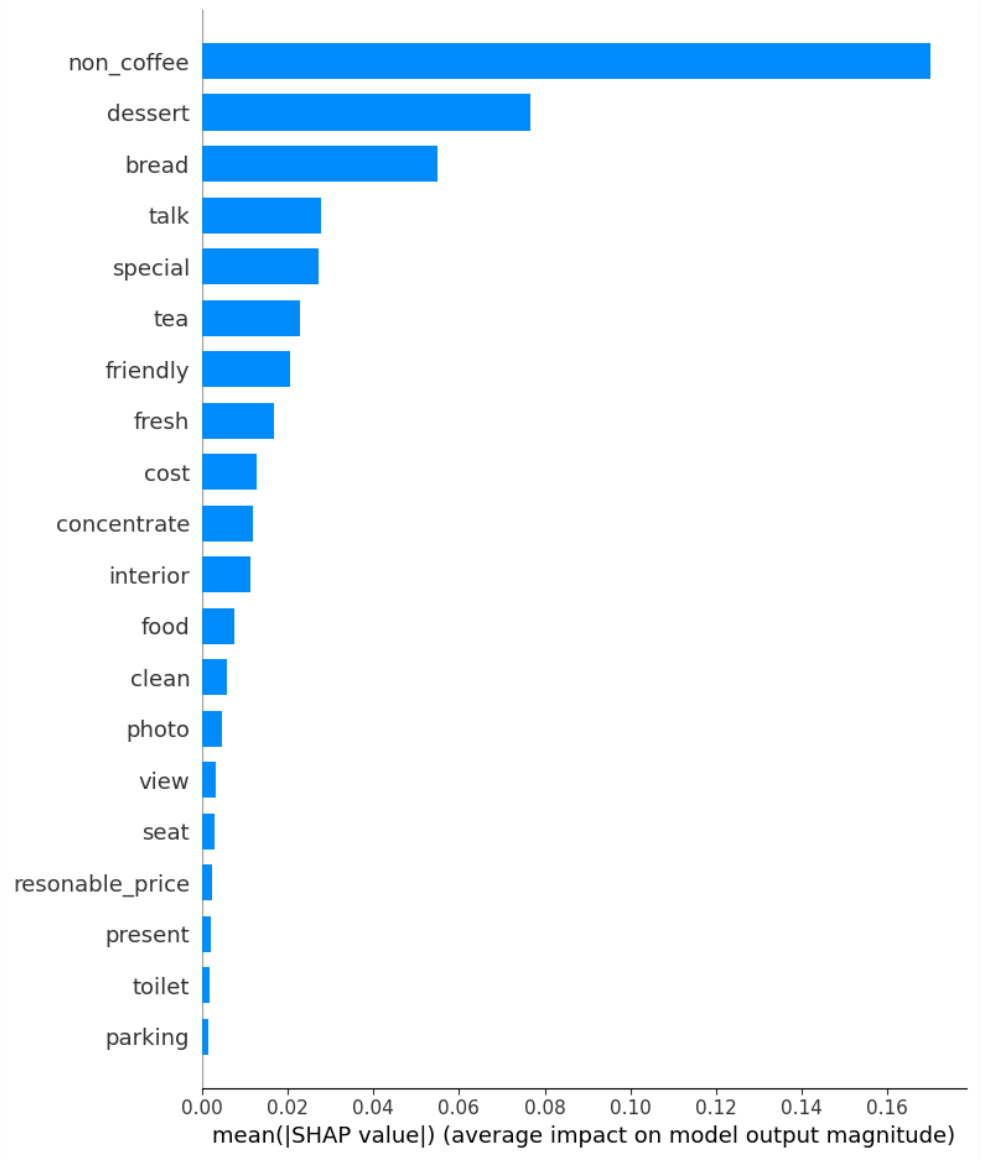
vs interior : 인테리어는 커피의 맛이 좋다고 판단하는 것에 큰 영향을 미치지 않았다.

vs clean : 매장이 청결하다고 대답한 경우에는 커피의 맛에 대한 만족감을 표현하는 경향이 낮았다.

vs special : 특별한 메뉴가 있다고 대답한 경우에는 커피의 맛에 대한 만족감을 표현하는 경향이 낮았다.

우리가 수집한 네이버 지도 내의 평가지표는 다중선택이 가능함에도 불구하고,

위와 같은 결과값을 얻어낼 수 있었다.




[ fig17. 커피의 맛에 대한 평가에 영향을 주는 변수들의 중요도 ]

추가적으로 커피의 맛 평가에 대해 가장 영향을 주는 특성으로 non_coffee의 맛 평가가 예측되었고,
이에 대한 절대값은 평균 17% 가량으로 갖는다.

## 수행결과(Result)

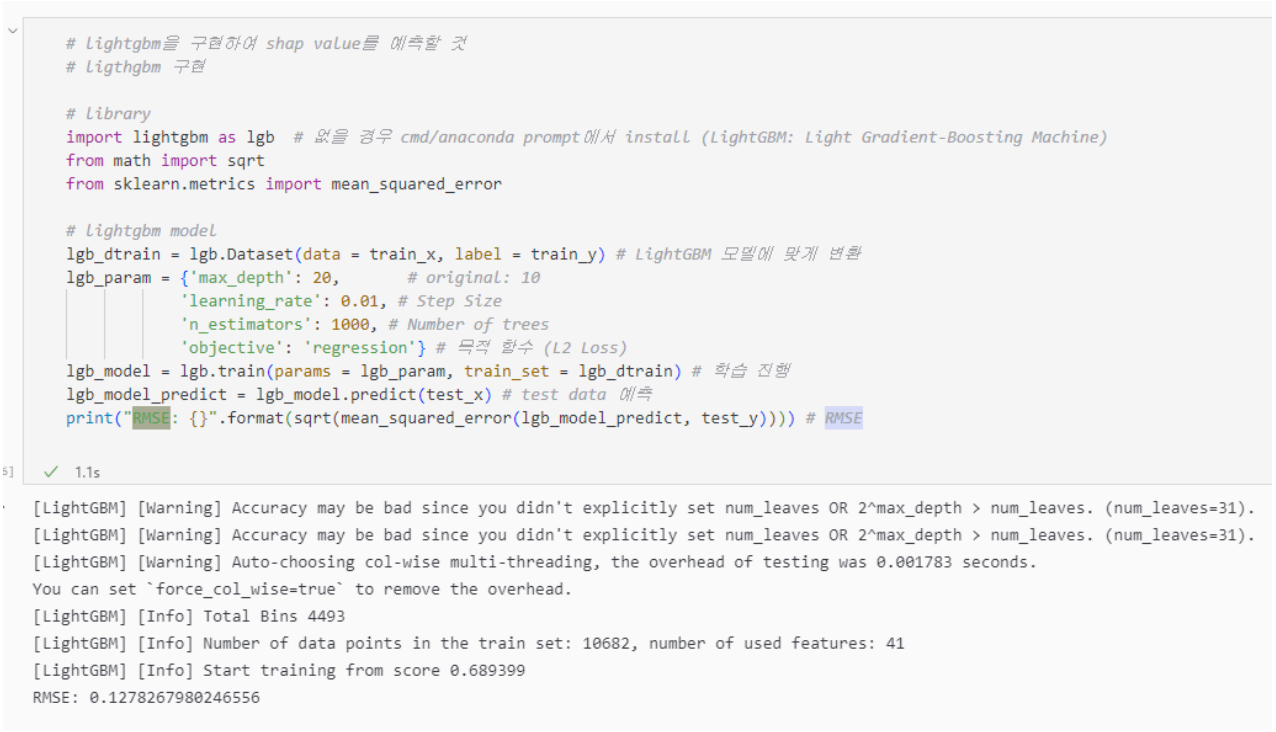
```
cafename = '서촌금상고로케'

result = recommend_cafe_list(data,cafe=cafename)


# data['name'] = data['열1']
# data.set_index('열1', inplace=True)
data.rename(columns={'열1':'name'},inplace=True)
result.rename(columns={'열1':'name'},inplace=True)
index = data.index[(data['name'] == cafename)]

user = data.iloc[index]

print(result['name'])
```

✓ 28.8s

```
14997      꽈페 홍대점
8497        부트브레드
14900    꿀넹쿠키 연남점
1037        일팔공일오
11737       빵미제빵소
Name: name, dtype: object
```

[ fig18. 고안한 추천시스템을 통해 추천된 목록, '서촌금상고로케' 기준]

입력값으로 사용자가 좋았던 카페 명을 입력받으면, 해당 카페와 비슷한 카페들 상위 5개를 추천해주었다. 현재는 임의로 '서촌금상고로케'를 넣었지만 데모에선 실제 존재하는 카페 명을 검색해서 넣어 유사한 카페들을 추천할 것입니다.



```
# lightgbm을 구현하여 shap value를 예측할 것
# ligthgbm 구현

# library
import lightgbm as lgb  # 없을 경우 cmd/anaconda prompt에서 install (LightGBM: Light Gradient-Boosting Machine)
from math import sqrt
from sklearn.metrics import mean_squared_error

# lightgbm model
lgb_dtrain = lgb.Dataset(data = train_x, label = train_y) # LightGBM 모델에 맞게 변환
lgb_param = {'max_depth': 20,       # original: 10
             'learning_rate': 0.01, # Step Size
             'n_estimators': 1000, # Number of trees
             'objective': 'regression'} # 목적 함수 (L2 Loss)
lgb_model = lgb.train(params = lgb_param, train_set = lgb_dtrain) # 학습 진행
lgb_model_predict = lgb_model.predict(test_x) # test data 예측
print("RMSE: {}".format(sqrt(mean_squared_error(lgb_model_predict, test_y)))) # RMSE
```

✓ 1.1s

```
[LightGBM] [Warning] Accuracy may be bad since you didn't explicitly set num_leaves OR 2^max_depth > num_leaves. (num_leaves=31).
[LightGBM] [Warning] Accuracy may be bad since you didn't explicitly set num_leaves OR 2^max_depth > num_leaves. (num_leaves=31).
[LightGBM] [Warning] Auto-choosing col-wise multi-threading, the overhead of testing was 0.001783 seconds.
You can set `force_col_wise=true` to remove the overhead.
[LightGBM] [Info] Total Bins 4493
[LightGBM] [Info] Number of data points in the train set: 10682, number of used features: 41
[LightGBM] [Info] Start training from score 0.689399
RMSE: 0.1278267980246556
```

[ fig19. 고안한 추천시스템을 통해 추천된 목록의 평가, 데이터셋에서 임의로 추출된 데이터 기준]

RMSE를 통한 정확도 평가(prediction accuracy)를 진행하였을 때 12 ~ 13% 가량의 정확도로 평가되었다.