LG U+ WHY NOT SOFTWARE CAMP 6기
클라우드 기반 빅데이터 분석 부트캠프

# 모델 정의서

**U+콕&홈쇼핑 기반 사용자 맞춤 식재료 및 레시피 추천 서비스**

## Eat's 유혹

**2025.07.23. ~ 2025.09.13**

유혹의 소나타 팀
팀장 이지헌
팀원 강민혁, 서세빈, 장윤수, 정의철, 조선영

# 목차

# 1. 개요

## 1) 프로젝트 개요

- **프로젝트명** : U+콕&홈쇼핑 기반 사용자 맞춤 식재료 및 레시피 추천 서비스
- **프로젝트 개요** : U+콕과 홈쇼핑 양방향 연계를 통한 판매 식재료 및 레시피 추천 웹 애플리케이션(Web Application)
  - ▶ 'U+콕', '홈쇼핑'에서 판매중인 '식품' 카테고리 상품에 대하여 두 서비스의 상호 상품 추천 기능 제공
  - ▶ 관심 상품 구매 편의성 향상을 위한 유사 상품 판매 시의 알림 기능 제공
  - ▶ 사용자 관심 식재료들로 만들 수 있는 레시피 추천 및 미보유 중인 식재료 추천 및 구매 정보 제공을 통한 쇼핑 편의성 향상

## 2) 알고리즘/모델 사용 목적

- 사용자 입력 키워드 기반 레시피 추천 시 조건 일치 추천 후 부족한 개수만큼 유사도 높은 순으로 추천
- '식품' 한정 상품 판매 및 특정 상품을 이용한 레시피 추천을 위한 식품/비식품, 식재료/완제품 분류

## 3) 사용 알고리즘/모델

- 레시피명 기반 레시피 추천
  - ▶ SBERT / SentenceTransformer("paraphrase-multilingual-MiniLM-L12-v2")
- 식품/비식품, 식재료/완제품 분류
  - ▶ Linear SVM

## 4) 적용 분야

- 레시피명 기반 레시피 추천
  - ▶ 만개의 레시피 기본 정보 중 'COOKING_NAME' 컬럼
- 식품/비식품, 식재료/완제품 분류
  - ▶ 홈쇼핑, 쇼핑몰 상품 데이터 중 'PRODUCT_NAME' 컬럼

**SBERT /**

**SentenceTransformer("paraphrase-multilingual-MiniLM-L12-v2")**

# 가. 데이터 정보

## 1) 데이터 출처

만개의 레시피 무료 레시피 데이터(KADX 농식품 빅데이터 거래소)

## 2) 형식 및 특징

- 전처리 전 : CSV
- 전처리 후 :

▶ FCT_RECIPE

| 필드명 | 논리 형식 | 저장 형식 | NULL | 예시 |
|---|---|---|---|---|
| RECIPE_ID | 레시피 고유번호(정수) | INT | N | 7038950 |
| RECIPE_TITLE | 레시피명(문자열) | VARCHAR(200) | Y | 게맛살 명란젓알 김치돌솥비빔밥 한그릇 점심 저녁메뉴추천 |
| COOKING_NAME | 요리명(메인명) | VARCHAR(40) | Y | 명란알게맛살돌솥비빔밥 |
| SCRAP_COUNT | 카운트(정수) | INT | Y | 1 |
| COOKING_CASE_NAME | 상황 분류 (카테고리) | VARCHAR(200) | Y | 영양식 |
| COOKING_CATEGORY_NAME | 카테고리 분류 (카테고리) | VARCHAR(200) | Y | 밥/죽/떡 |
| COOKING_INTRODUCTION | 레시피 소개 (문자열) | VARCHAR(4000) | Y | 게맛살과 명란젓을 이용해서 알밥 만들기 해 볼께요. |
| NUMBER_OF_SERVING | 인분(정수/소수) | VARCHAR(200) | Y | 2인분 |
| THUMBNAIL_URL | 썸네일 이미지 URL (문자열) | VARCHAR(200) | Y | https://recipe1.ezmember.co.kr/cache/recipe/2024/11/20/264faa35fec5b3d0762f59c37ac43ca91.jpg |

▶ FCT_MTRL

| 필드명 | 논리 형식 | 저장 형식 | NULL | 예시 |
|---|---|---|---|---|
| MATERIAL_ID | 재료 고유번호(정수) | INT | N | 1 |
| RECIPE_ID | 레시피 고유번호 (정수) | INT | Y | 7016813 |
| MATERIAL_NAME | 재료명(텍스트) | VARCHAR(100) | Y | 떡국떡 |
| MEASURE_AMOUNT | 계량 수량 (문자열) | VARCHAR(100) | Y | 400 |
| MEASURE_UNIT | 계량 단위 (문자열) | VARCHAR(200) | Y | g |
| DETAILS | 세부 설명 (문자열) | VARCHAR(4000) | Y | NULL |

## 3) 수집 규모

- 전처리 전
  - ▶ 레시피 테이블 총 23,192행

- 전처리 후
  - ▶ 레시피 테이블 총 21,059행
  - ▶ 재료 테이블 총 192,260행

## 4) 전처리 과정

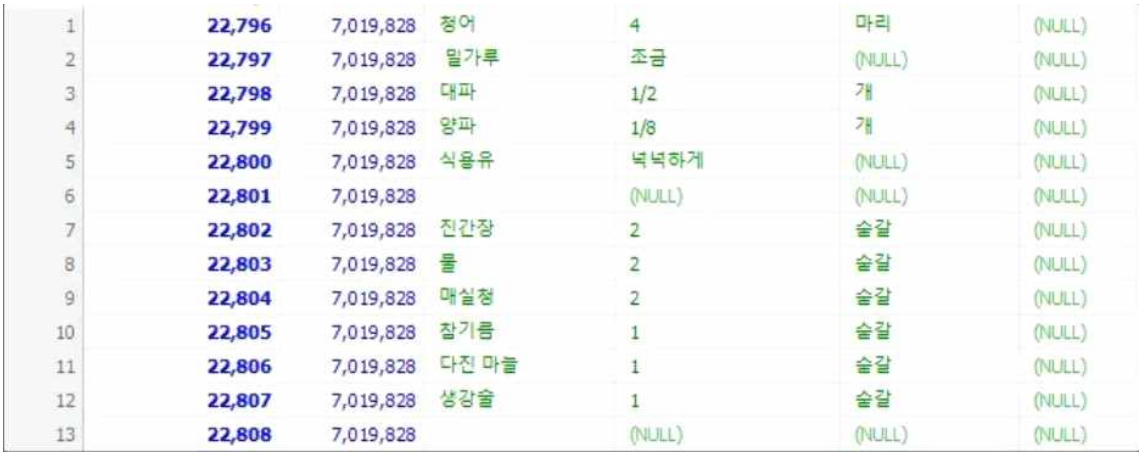
- RAW DATA의 데이터 중 하기 9개 컬럼이 NULL인 경우 전체 삭제
  - ▶ 로우 데이터 예시
    - ▷ 아스키코드 7번의 bell character를 기준 [ 재료명 | 수량 | 단위 | 참고사항 ] 으로 나뉘어져 있음을 확인

| RCP_SNO | CKG_MTRL_ACTO_NM |
|---|---|
| 7019828 | [재료] 청어•4•마리•\| 밀가루••조금•\| 대파•1/2•개•\| 양파•1/8•개•\| 식용유••넉넉하게•\| ••• [양념소스] 진간장•2•숟갈•\| 물•2•숟갈•\| 매실청•2•숟갈•\| 참기름•1•숟갈•\| 다진 마늘•1•숟갈•\| 생강술•1•숟갈•\| ••• |

- 정제된 데이터에 한해 재료 컬럼 전처리
  - ▶ 1차 데이터 정제 예시

| RCP_SNO | CKG_MTRL_ACTO_NM |
|---|---|
| 7019828 | 청어•4•마리•\| 밀가루••조금•\| 대파•1/2•개•\| 양파•1/8•개•\| 식용유••넉넉하게•\| ••• \| 진간장•2•숟갈•\| 물•2•숟갈•\| 매실청•2•숟갈•\| 참기름•1•숟갈•\| 다진 마늘•1•숟갈•\| 생강술•1•숟갈•\| ••• |

  - ▶ 2차 데이터 정제 예시
    - ▷ • 를 기준으로 [ 재료명 | 숫자 | 단위 | 참고사항 ] 으로 컬럼 분할
    - ▷ | 를 기준으로 로우 분할



| | | | | | | |
|---|---|---|---|---|---|---|
| 1 | 22,796 | 7,019,828 | 청어 | 4 | 마리 | (NULL) |
| 2 | 22,797 | 7,019,828 | 밀가루 | 조금 | (NULL) | (NULL) |
| 3 | 22,798 | 7,019,828 | 대파 | 1/2 | 개 | (NULL) |
| 4 | 22,799 | 7,019,828 | 양파 | 1/8 | 개 | (NULL) |
| 5 | 22,800 | 7,019,828 | 식용유 | 넉넉하게 | (NULL) | (NULL) |
| 6 | 22,801 | 7,019,828 | | (NULL) | (NULL) | (NULL) |
| 7 | 22,802 | 7,019,828 | 진간장 | 2 | 숟갈 | (NULL) |
| 8 | 22,803 | 7,019,828 | 물 | 2 | 숟갈 | (NULL) |
| 9 | 22,804 | 7,019,828 | 매실청 | 2 | 숟갈 | (NULL) |
| 10 | 22,805 | 7,019,828 | 참기름 | 1 | 숟갈 | (NULL) |
| 11 | 22,806 | 7,019,828 | 다진 마늘 | 1 | 숟갈 | (NULL) |
| 12 | 22,807 | 7,019,828 | 생강술 | 1 | 숟갈 | (NULL) |
| 13 | 22,808 | 7,019,828 | | (NULL) | (NULL) | (NULL) |

▶ 3차 데이터 정제 예시

▷ 'material_name'의 데이터 중 '밀가루'처럼 앞뒤 공백 있는 경우, 공백 삭제

▷ `ZWSP`, `Word Joiner`포함 확인

▷ `ZWSP`, `NBSP`, `Ideographic Space`, `Word Joiner`, `ZWNBSP` 전부 제거

| | | | | | | |
|---|---|---|---|---|---|---|
| 1 | 22,790 | 7,019,828 | 청어 | 4 | 마리 | (NULL) |
| 2 | 22,791 | 7,019,828 | 밀가루 | 조금 | (NULL) | (NULL) |
| 3 | 22,792 | 7,019,828 | 대파 | 1/2 | 개 | (NULL) |
| 4 | 22,793 | 7,019,828 | 양파 | 1/8 | 개 | (NULL) |
| 5 | 22,794 | 7,019,828 | 식용유 | 넉넉하게 | (NULL) | (NULL) |
| 6 | 22,795 | 7,019,828 | 진간장 | 2 | 술갈 | (NULL) |
| 7 | 22,796 | 7,019,828 | 물 | 2 | 술갈 | (NULL) |
| 8 | 22,797 | 7,019,828 | 매실청 | 2 | 술갈 | (NULL) |
| 9 | 22,798 | 7,019,828 | 참기름 | 1 | 술갈 | (NULL) |
| 10 | 22,799 | 7,019,828 | 다진 마늘 | 1 | 술갈 | (NULL) |
| 11 | 22,800 | 7,019,828 | 생강술 | 1 | 술갈 | (NULL) |

## 나. 알고리즘/모델

### 1) 알고리즘/모델 후보 선정

- **TF-IDF + 코사인 유사도**
  ▶ 레시피명 키워드 빈도 기반 유사도(가장 기본)
- **SBERT**
  ▶ 전체 레시피명을 문장 단위로 임베딩 → 의미 기반 유사도 추천
- **Word2Vec**
  ▶ 레시피명 Word2Vec → BERT 임베딩 후 토큰 유사도 필터링 + 의미 유사도 재정렬
- **Fasttext**
  ▶ 서브워드(n-gram) 임베딩으로 표기 변형/오탈자에 강한 단어 벡터 생성 → 제목 임베딩은 단어 벡터 평균 → 코사인 유사도로 추천
- **Fasttext + SBERT**
  ▶ 1단계(FastText): 표기 변형에 강한 임베딩으로 후보 넓게 수집 → 2단계(BERT/SBERT): 문장 임베딩으로 의미 유사도 재정렬(리랭크)

### 2) 알고리즘/모델 비교

| 알고리즘 | 장점 | 단점 |
|---|---|---|
| **TF-IDF + 코사인 유사도** | 빠름, 구현 간단 | 의미 유사도 약함 |
| **SBERT** | 의미 유사도 우수, 구현 단순<br>(한 번의 임베딩) | 사전학습 의존, 도메인 특화는 추가 튜닝 필요 |
| **Word2Vec** | 의미 유사도 우수, 정밀 필터링 가능 | 구현 복잡, 추론 비용 ↑ |
| **Fasttext** | 오탈자/표기 변형 강건성, 속도 빠름,<br>소규모 코퍼스에서도 안정 | 문장 수준의 의미 표현 한계,<br>다의어 구분 약함 |
| **Fasttext + SBERT** | 강건 리콜(Fasttext) + 의미 정밀도<br>(BERT/SBERT) | 2단계 구조로 복잡·지연↑,<br>운영/캐싱 필요 |

# 다. 알고리즘/모델 선정 및 평가

## 1) 성능 평가 지표

| 항목 | 지표명 | 정의 | 평가 방법 |
|---|---|---|---|
| **적합성** | 추천 적합도 | 유사도 기반 추천 결과가 잘 출력 되는지 | 평가자 3인 이상이 각 레시피에 대해 1~5점 부여 → 평균 산출 |
| **처리 속도** | 추천 결과 응답 시간 | 레시피 추천 소요 시간 | 5회 측정 후 평균값 산출 |

## 2) 평가 방법

- 오믈렛 레시피 20개 추천 결과로 비교(만개의 레시피 내 오믈렛 레시피 15개)
- 유사도 비교 기준 컬럼 : COOKING_NAME

## 3) 결과

- 적합성과 처리 속도 면에서 가상 우수한 성적을 낸 SBERT 선정
- 모델 선정 후 처리 속도는 pgvector를 사용하여 향상

| 모델 | 적합성 | 처리 속도 |
|---|---|---|
| **TF-IDF + 코사인유사도** | 1 | 0.070초 |
| **SBERT** | 5 | 8.805초 |
| **Word2Vec + BERT** | 5 | 35.671초 |
| **FastText** | 1 | 0.710초 |
| **FastText + BERT** | 3 | 36.426초 |

# 라. 선정 알고리즘/모델 적용

## 1) 선정 알고리즘/모델 설명

**SBERT(Sentence-BERT)**

- **분류**
  - ▶ 알고리즘/모델 계열 : 시암(Siamese) 구조의 Transformer Bi-Encoder 기반 문장 임베딩
  - ▶ 용도 : 짧은 텍스트(문장/제목)를 고정 길이 벡터로 변환 → 코사인 유사도 등으로 의미 검색/추천
- **로직**
  - ▶ 동일 가중치의 Transformer 인코더를 사용해 텍스트 쌍을 각각 임베딩, 임베딩 공간에서 가까우면 의미가 유사하도록 학습(대조학습/랭킹 손실)
  - ▶ 온라인 서빙에서는 텍스트 → 벡터 한 번만 계산하면 되어 빠른 검색이 가능
- **장단점**
  - ▶ 장점 : 의미 기반 유사도 우수, 서빙 지연 낮음(Bi-Encoder), 다국어 모델 선택 가능
  - ▶ 한계 : 단일 문장 벡터에 모든 뉘앙스를 담아야 하므로 세부 구분/맥락은 재랭크·규칙 보완이 필요

**paraphrase-multilingual-MiniLM-L12-v2**

- **프레임워크:** sentence-transformers
- **체크포인트:** sentence-transformers/paraphrase-multilingual-MiniLM-L12-v2
- **백본:** MiniLM-L12 Transformer
- **풀링:** Mean Pooling(토큰 평균) → 문장 임베딩 ≈ 384차원
- **특성:** 다국어(한국어 포함) 패러프레이즈/번역쌍 등으로 파인튜닝 → 언어가 달라도 의미가 비슷하면 벡터 근접

## 2) 적용 방법

- 임베딩 생성 : SBERT 알고리즘의 SentenceTransformer("paraphrase-multilingual-MiniLM-L12-v2") 모델로 요리명(COOKING_NAME) 임베딩(384차원)
- 유사도 계산 : 코사인 거리(pgvector <->, 작을수록 유사)
- 정렬/반환: 코사인 거리 오름차순 정렬 → Top-k 추천 결과 생성

## 3) 처리 속도 향상 방법

- 벡터 스토리지/검색 : Postgres + pgvector 사용
  - ▶ 인덱스: ivfflat (vector_cosine_ops) / lists(인덱스 크기)·probes(탐색 폭)로 속도↔정확도 조절
- 파이프라인
  - ▶ pgvector에서 유사도 Top-k ID 조회
  - ▶ 상세 메타는 MariaDB에서 조회/조인
- 정확 일치(LIKE)는 LIMIT를 걸어 소량만 사용, 나머지는 임베딩 검색으로 보완

### 4) 처리 속도 향상 결과

- 모델 인퍼런스 시간 : model.encode() 구간만
- 전체 추천 처리 시간 : 쿼리 수집 + 인퍼런스 + Pgvector 검색 + MariaDB 상세조회/머지

| 구분 | 기준 | MariaDB | Postgre + MariaDB |
|---|---|---|---|
| 로제파스타 | 모델 인퍼런스 시간 | 0.0249초 | 0.0355초 |
| | 전체 추천 처리 시간 | 9.1725초 | 3.5368초 |
| 크림떡볶이 | 모델 인퍼런스 시간 | 0.0233초 | 0.0455초 |
| | 전체 추천 처리 시간 | 10.4827초 | 3.6925초 |

# Linear SVM

# 가. 데이터 정보

## 1) 데이터 출처

- 5개 홈쇼핑사
  - ▶ 홈앤쇼핑
  - ▶ 현대홈쇼핑 / 현대홈쇼핑 + 샵
  - ▶ NS홈쇼핑 / NS샵+
- 쇼핑몰(U+콕)

## 2) 형식 및 특징

- 웹 내 데이터 크롤링 후 DB 적재
- 주요 테이블 데이터 형식
  - ▶ 홈쇼핑

| 필드명 | 논리 형식 | 저장 형식 | NULL | 예시 |
|---|---|---|---|---|
| **FCT_HOMESHOPPING_PRODUCT_INFO** | | | | |
| PRODUCT_ID | 상품 고유번호(정수) | BIGINT(20) | N | 33855745 |
| STORE_NAME | 상품명(문자열) | VARCHAR(1000) | Y | 촉촉한 반건조 오징어 |
| SALE_PRICE | 판매 가격(정수) | BIGINT(20) | Y | 85900 |
| DT_RATE | 할인율(정수) | INT | Y | 12 |
| DC_PRICE | 할인 가격(정수) | BIGINT(20) | Y | 75590 |

  - ▶ 쇼핑몰(U+콕)

| 필드명 | 논리 형식 | 저장 형식 | NULL | 예시 |
|---|---|---|---|---|
| KOK_PRODUCT_ID | 상품 고유번호(정수) | INT | N | 1472282 |
| KOK_STORE_NAME | 판매상점명(문자열) | VARCHAR(100) | Y | 효성어묵 |
| KOK_PRODUCT_NAME | 상품명(텍스트) | VARCHAR(300) | Y | [효성어묵] 부산전통어묵 기획세트 4종 모음전 |
| KOK_THUMBNAIL | 썸네일 URL(문자열) | TEXT | Y | https://d-img.picnique.co.kr/product/normal/admin/16141vsu5447FlsCL581_1615447581_base_1.jpg |
| KOK_PRODUCT_PRICE | 판매 가격(정수) | INT | Y | 27000 |
| KOK_REVIEW_SCORE | 리뷰 점수 | FLOAT | Y | NULL |
| KOK_REVIEW_CNT | 리뷰 수(정수) | INT | | NULL |

## 3) 수집 규모

| ODS | 데이터(테이블)명 | 총 수집 행 수 | 일 평균 수집 행 수 | 제품 별 평균 행 수 |
|---|---|---|---|---|
| 홈쇼핑 | 편성표 | 12,680 | 680.0 | |
| | 제품 정보 | 6,109 | 720.3 | |
| | 이미지 URL | 121,207 | | 19.8 |
| | 상품정보고시 | 93,999 | | 15.4 |
| 콕 | 가격 정보 | 5,071 | 58.7 | |
| | 제품 정보 | 4,532 | 5.7 | |
| | 이미지 URL | 36,661 | | 8.1 |
| | 상품정보고시 | 54,486 | | 12.0 |

| SVC | 데이터(테이블)명 | 총 데이터 행 수 |
|---|---|---|
| 홈쇼핑 | 편성표 | 9,522 |
| | 제품 정보 | 5,044 |
| | 이미지 URL | 59,794 |
| | 상품정보고시 | 78,881 |
| 콕 | 가격 정보 | 5,067 |
| | 제품 정보 | 4,532 |
| | 이미지 URL | 36,008 |
| | 상품정보고시 | 54,331 |

## 4) 전처리 과정

| | 홈쇼핑 | 쇼핑몰(U+콕) |
|---|---|---|
| 서비스 내 사용 항목 | **편성표**<br>• 방영일 / 방영 시작 시간 / 방영 종료 시간<br>• 제품 타입 (메인, 서브)<br>• 제품명<br>• 썸네일 이미지<br><br>**각 제품별 상세 정보**<br>• 원가 / 할인율 / 할인가<br>• 상세 설명 이미지 url<br>• '전자상거래 등에서의 상품 등의 정보제공에 관한 고시' 에 따른 상세정보 | **리스트 페이지**<br>• 할인율 / 할인가<br>• 각 제품별 기본 정보<br>• 제품 썸네일 이미지 url<br>• 판매자 정보<br>• 리뷰 점수 평균 및 점수 별 비율<br>• 평가 항목 별 리뷰 및 비율<br>• 상품 판매 원가<br><br>**각 제품별 상세 정보**<br>• 상품 설명 탭의 이미지 url<br>• '전자상거래 등에서의 상품 등의 정보제공에 관한 고시' 에 따른 상세정보<br>• 상품 리뷰 예시 5개에 대한 개별 평점, 상세 리뷰 내용 |
| 전처리 프로세스 | **공통사항**<br>• 방영일: DATE로 데이터 타입 변경. (%Y%M%d)<br>• 방영시간: TIME으로 데이터 타입 변경 및 LIVE_END_TIME 으로 분할<br>• 미사용 컬럼 미적재<br>• 제품 코드: BIGINT로 데이터 타입 변경<br><br>**홈앤쇼핑**<br>• 크롤링 실행 시간에 방영되는 프로그램의 방영 시간이 '지금 방송 중'으로 표기되는 현상 처리 | • 숫자를 제외한 문자를 제거하는 'STR_TO_NUM' 함수 정의 및 숫자형으로 데이터 타입 변경<br>• 판매자명이 제품명 가장 앞에서 반복되는 현상 처리 |

# 나. 알고리즘/모델

## 1) 알고리즘/모델 후보 선정

- **LightGBM(GBDT)**
  - ▶ 희소/다차원 피처(TF-IDF, BM25, 토큰/문자 n-gram, SBERT 거리/코사인, 카테고리·인기도 등)로 비선형 상호작용을 학습하는 그래디언트 부스팅 트리
  - ▶ 빠른 학습·추론, 피처 중요도 해석 용이, 리랭커로 적합
- **Linear SVM(확률 보정)**
  - ▶ 선형 결정경계로 빠르게 분리(대규모·희소 피처에 강함). 기본 출력은 점수(마진)라서, Platt scaling / Isotonic으로 확률 보정하여 임계값·랭킹에 활용
- **Logistic Regression**
  - ▶ 선형 모델로 확률 출력을 직접 제공(캘리브레이션 부담 ↓). 가중치 해석 가능하고, 희소 피처(TF-IDF 등)와 궁합이 좋음

## 2) 알고리즘/모델 비교

| 알고리즘 | 장점 | 단점 |
|---|---|---|
| **LightGBM(GBDT)** | 비선형 패턴/상호작용 강함, 속도/정확도 균형 좋음, 결측/희소 피처 대응 우수 | 과적합 가능 → 조기종료/규제 필요, 임베딩 원시 벡터는 집약 피처(거리/유사도 등)로 넣는 게 실무적 |
| **Linear SVM(확률 보정)** | 매우 빠름, 대용량/희소 행렬에 최적, 해석(가중치) 용이 | 비선형 관계는 한계 → 폴리커널/XGBoost류 대비 표현력 낮음, 보정 단계가 추가비용 |
| **Logistic Regression** | 가볍고 안정적, 실시간 추론에 적합, **규제(L1/L2)**로 피처 선택/과적합 제어 | 비선형/상호작용 표현 한계 → 교차 피처나 파생피처 필요 |

# 다. 알고리즘/모델 선정 및 평가

## 1) 성능 평가 지표

- **Accuracy** : 전체 예측 중 정답 비율

$$Accuracy = \frac{TP + TN}{TP + FP + FN + TN}$$

- **Marco-F1** : 클래스별 F1을 산술 평균한 값(클래스 불균형에도 각 클래스 동일 기준)

$$Marco - F1 = \frac{1}{C}\sum_{C=1}^{C} F_{1,c}$$

- **AUC(ROC-AUC)** : 임계값 전 범위에서 양성/음성 순위화 품질. 1에 가까울수록 좋음

## 2) 평가 방법

- 데이터 준비
  - ▶ PRODUCT_NAME(상품명)과 IS_FOOD(식품/비식품)
  - ▶ 라벨 : 식품 = 1, 비식품 = 0 반환
- 모델 후보 구성
  - ▶ 각 fold마다 벡터라이저를 새로 학습
  - ▶ 학습 폴드로 모델 학습 → 검증 폴드에 예측(라벨 & 확률)
- OOF(Out-Of-Fold) 예측으로 공정 평가
- 모델별 성능 집계 & 리포트

## 3) 결과

- 평가 데이터 : 총 2391건
- 모델별 정확 예측(정답 개수) / Accuracy / Marco-F1 / AUC
  - ▶ Accuracy 1위 : linear_svm_calibrated (0.9841)
  - ▶ Marco-F1 1위 : linear_svm_calibrated (0.9737)
  - ▶ AUC 1위: logreg (0.9968)

| 모델 | 정답 개수 | 총 데이터 | Accuracy | Macro-F1 | AUC |
|---|---|---|---|---|---|
| linear_svm_calibrated | 2353 | 2391 | 0.9841 | 0.9737 | 0.9951 |
| logreg | 2346 | 2391 | 0.9812 | 0.9686 | 0.9968 |
| lightgbm | 2314 | 2391 | 0.9678 | 0.9450 | 0.9891 |

# 라. 선정 알고리즘/모델 적용

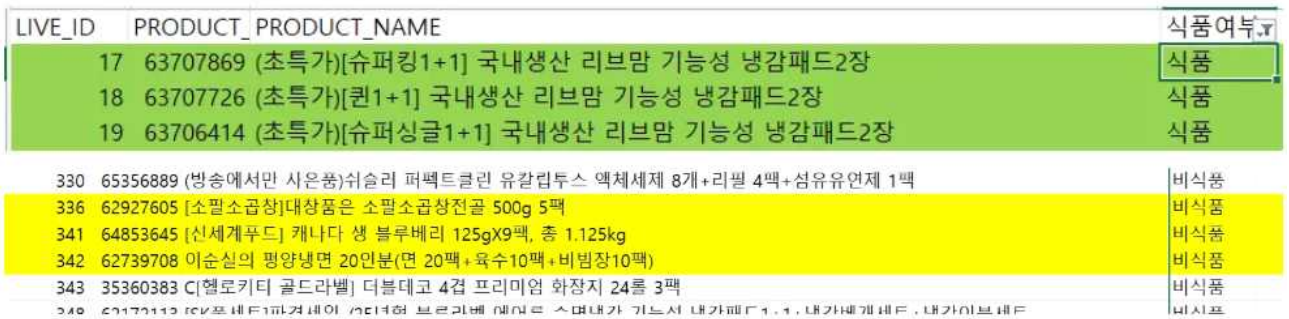
## 1) 선정 알고리즘/모델 설명

**Linear SVM**

- **분류**
  - ▶ 알고리즘/모델 계열 : 선형 서포트 벡터 머신(Linear SVM) – 마진 최대화 기반의 선형 분류기
  - ▶ 용도 : 대규모·희소 텍스트 피처(TF-IDF/BM25/n-gram, SBERT 유사도 스칼라 등)로 이진 분류/리랭크에 활용. 확률 보정을 붙여 Top-k 순위화나 임계값 판단에 사용
- **로직**
  - ▶ 피처 만들기 : 각 쿼리–레시피 쌍을 숫자로 바꿈
  - ▶ 가장 잘 가르는 직선/평면 찾기: 두 클래스를 가장 넓은 간격으로 떨어뜨리는 선을 찾음
  - ▶ 예측: 새 샘플에 대해 선의 어느 쪽에 얼마나 떨어져 있는지 점수(양수/음수)화 → 부호로 클래스 결정
  - ▶ 확률 필요시 보정: 점수를 Platt scaling(또는 Isotonic)으로 0~1 확률처럼 바꿔서 임계값으로 채택/제외를 정하거나, 점수/확률로 Top-k 순위 생성
- **장단점**
  - ▶ 장점
    - ▷ 빠름(대규모 TF-IDF 같은 희소 피처에 특히 강함)
    - ▷ 규제(C)로 일반화 강건, 과적합 제어 용이
    - ▷ 가중치 해석 가능(어떤 피처가 결정에 기여했는지)
  - ▶ 한계 : 직선/평면 하나로 구분하므로 복잡한 비선형 관계는 약함 → 필요시 파생 피처/다른 모델과 조합

## 2) 적용 방법

- 입력 & 전처리
- TF-IDF 벡터화(텍스트 패턴 학습)
  - ▶ 모델이 분류한 라벨링 결과 내 식품임에도 비식품으로 분류되었거나 비식품임에도 식품으로 분류된 경우의 라벨링 직접 수정



| LIVE_ID | PRODUCT_ | PRODUCT_NAME | 식품여부 |
|---|---|---|---|
| 17 | 63707869 | (초특가)[슈퍼킹1+1] 국내생산 리브맘 기능성 냉감패드2장 | 식품 |
| 18 | 63707726 | (초특가)[퀸1+1] 국내생산 리브맘 기능성 냉감패드2장 | 식품 |
| 19 | 63706414 | (초특가)[슈퍼싱글1+1] 국내생산 리브맘 기능성 냉감패드2장 | 식품 |
| 330 | 65356889 | (방송에서만 사은품)쉬슬리 퍼펙트클린 유칼립투스 액체세제 8개+리필 4팩+섬유유연제 1팩 | 비식품 |
| 336 | 62927605 | [소필소곱창]대창품은 소필소곱상전골 500g 5팩 | 비식품 |
| 341 | 64853645 | [신세계푸드] 캐나다 생 블루베리 125gX9팩, 총 1.125kg | 비식품 |
| 342 | 62739708 | 이순실의 평양냉면 20인분(면 20팩+육수10팩+비빔장10팩) | 비식품 |
| 343 | 35360383 | C[헬로키티 골드라벨] 더블데코 4겹 프리미엄 화장지 24롤 3팩 | 비식품 |

- 키워드 부가 피처(도메인 지식 주입)

▶ 사전 기반 피처

▷ '식품' 분류에 사용할 키워드

```
food_keywords = [
    "쌀", "김치", "라면", "즉석", "밥", "국", "탕", "찌개", "반찬", "떡", "과자", "간식",
    "소스", "조미료", "양념", "김", "생선", "정육", "고기", "햄", "어묵", "육포",
    "우유", "치즈", "계란", "달걀", "두부", "요구르트", "커피", "차", "음료", "주스", "홍삼", "분유", "곱창", "블루베리", "냉면", "육수", "유기농", "아이스크림", "다시마", "인분", "두유", "빙수", "오곡", "단팥", "훈제", "장어", "멸치", "해물", "다시팩", "엑스트라버진", "양갱", "만두", "풀무원", "올가", "피자", "고춧가루", "핫도그", "수산물", "오징어", "도넛", "자일리톨", "자숙", "구이", "손질", "아구찜", "코코넛오일", "복분자", "닭발", "매콤", "젓", "명란", "슬라이스", "낙지", "숙성", "고사리", "문어", "데친", "셰프", "소갈비", "굴비",
]
```

▷ '비식품' 분류에 사용할 키워드

```
notfood_keywords = [
"기능성", "접이식", "유산균", "앰플", "찜기", "셋업", "토너", "알로에", "순금", "세라믹", "프라이팬", "쌀통", "쿠션", "LG", "삼성", "우산", "조리기", "푸마", "크로커다일", "브라", "제약", "마스크", "아디다스", "드로즈", "아디다스", "트렁크", "글루타치온", "팬티", "용기", "립스틱", "밍크", "팬츠", "보험", "냉장고", "약품", "프로틴", "루테인", "데비마이어", "혈압", "혈당", "기억력", "개월분", "날씬", "밥솥", "24K", "18K", "한국금자산관리"
]
```

▶ 부가 피처

▷ TF-IDF만 사용할 경우 학습 데이터에 희귀하거나 신조어, 브랜드명이 들어오면 제대로 인지하지 못할 수 있음

ex) "풀무원 유기농 나물 비빔밥" : 학습 데이터에 "풀무원"이 없으면 식품인지 단번에 인식 못할 수도 있음

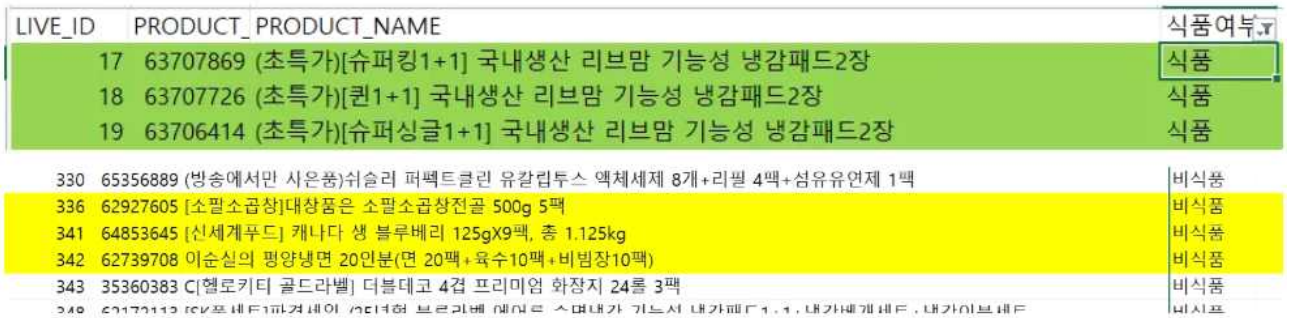
▷ 위 경우를 극복하기 위해 피처들을 TF-IDF 행렬 오른쪽에 hstack으로 붙여서 모델에 같이 학습시킴

◆ food_hit : 식품 키워드 매칭 개수
◆ notfood_hit: 비식품 키워드 매칭 개수
◆ any_food / any_not : 존재 여부(0/1)
◆ conflict : 두 집합 모두 히트(0/1)
◆ score = food_hit - notfood_hit : 음수면 비식품 쪽 신호가 강함

- 학습 & 검증
- 예측

### 3) 학습 및 저장

- 1차 : Linear SVM 활용 사전 피처 + 부가 피처 기반 분류
- 2차 : 모델이 분류한 라벨링 결과 내 식품임에도 비식품으로 분류되었거나 비식품임에도 식품으로 분류된 경우의 라벨링 직접 수정 후 재학습



| LIVE_ID | PRODUCT_ | PRODUCT_NAME | 식품여부 |
|---|---|---|---|
| 17 | 63707869 | (초특가)[슈퍼킹1+1] 국내생산 리브맘 기능성 냉감패드2장 | 식품 |
| 18 | 63707726 | (초특가)[퀸1+1] 국내생산 리브맘 기능성 냉감패드2장 | 식품 |
| 19 | 63706414 | (초특가)[슈퍼싱글1+1] 국내생산 리브맘 기능성 냉감패드2장 | 식품 |
| 330 | 65356889 | (방송에서만 사은품)쉬슬리 퍼펙트클린 유칼립투스 액체세제 8개+리필 4팩+섬유유연제 1팩 | 비식품 |
| 336 | 62927605 | [소팔소곱창]대창품은 소팔소곱창전골 500g 5팩 | 비식품 |
| 341 | 64853645 | [신세계푸드] 캐나다 생 블루베리 125gX9팩, 총 1.125kg | 비식품 |
| 342 | 62739708 | 이순실의 평양냉면 20인분(면 20팩+육수10팩+비빔장10팩) | 비식품 |
| 343 | 35360383 | C[헬로키티 골드라벨] 더블데코 4겹 프리미엄 화장지 24롤 3팩 | 비식품 |

- 모델 저장
  - ▶ tfidf_word.pkl : 단어 단위 TF-IDF 벡터라이저
  - ▶ tfidf_char.pkl : 문자 단위 TF-IDF 벡터라이저
  - ▶ linear_svm_calibrated.pkl : 학습된 분류 모델 (Linear SVM + 확률 보정)
  - ▶ keyword_meta.json : 키워드 피처 생성용 메타데이터

### 4) 예측 결과

- 100개의 더미 데이터 생성

| PRODUCT_NAME | IS_FOOD |
|---|---|
| 용지봉 토시살 구이 170g*9팩 | 식품 |
| [풀무원] 식물성 지구식단 이슬만두 직화불고기맛 180gX7개 | 식품 |
| 해통령 육수링 100알(깊고 진한맛 80gX5봉) | 식품 |
| [베스트] 영산포 숙성 홍어회 총 7팩 (모둠회 150gX2팩+몸살회 150gX3팩) | 식품 |
| [방송특가] 전철우고향랭면 평양냉면 18인분(면 18팩+육수 10팩+비빔장 8팩) | 식품 |
| [오늘만] 한명숙 육미본가 흑염소탕 700gX9팩 (약 20인분) | 식품 |
| [방송특가] (방송중 1팩 더) 국내산 흑산도 홍어회 5+3팩 (모둠회100g+초장20g, 총 6팩) | 식품 |
| [무료배송] 조리기능장 임성근의 특 소곱창전골 800gX7팩 | 식품 |

- 결과 : 3개 False

| PRODUCT_NAME | IS_FOOD | 예측 | 식품확률 | 정답여부 |
|---|---|---|---|---|
| [무료배송] 조리기능장 임성근의 특 소곱창전골 800gX7팩 | 식품 | 비식품 | 0.49 | FALSE |
| [세트] (얼리버드)프롬바이오 유기농 레자몽 6박스(총 84포) | 비식품 | 식품 | 0.51 | FALSE |
| [세트] 자일리톨스톤 10개입(오리지널맛 3개+페퍼민트부스터 3개+레몬향 6개+졸 | 비식품 | 식품 | 0.51 | FALSE |