🍴

AI 기반 집단급식소 식수 예측 시스템 구축

키워드	ML	NLP	Python	Python Flask	Web

✔️ 분석 개요

분석 배경

- 집단 급식소에서의 식수 예측은 비용 문제로 인해 영양사의 경험과 직관에 의한 예측이 주를 이루고 있음
- 데이터 기반 식수 예측 모델링을 통해 예측의 정확성과 객관성, 보편타당성을 보충

분석 목적

- 미배식 되어서 나오는 음식물 쓰레기 최소화

분석 목표

- 지자체에서 제공한 식단표, 카드 데이터와 공공데이터 포털의 데이터를 활용하여 AI 기반의 집단급식소 식수 인수 인원을 예측
- 실무에서 바로 활용할 수 있도록 모델 배포 및 서비스 웹사이트 구현

분석 내용

- 지자체에서 제공한 약 6년간의 식단표, 카드 데이터와 공공데이터 포털의 데이터를 활용하여 AI 기반의 집단급식소 식수 인수 인원 예측
- 한식 메뉴의 특성을 반영한 비정형 데이터(식단표) 처리를 통한 모델 성능 개선
- Gradient Boosting, Random Forest, CatBoost 모델을 블렌딩하여 최종 앙상블 모델 도출
- 웹 서비스 구현

분석 결과

- 최종 모델의 MAPE는 7.64
- **식수 예측 서비스 실무 적용 후 식수 예측 오차율 약 10%P 개선**
  - 기존 오차율 (예측 모델 사용 전) : 15~20%
  - 개선 오차율 (예측 모델 사용 후): 5~10%

✔️ 분석 프로세스

0. 선행 연구 조사

- 선행 연구를 통해 메뉴, 날짜, 날씨 등의 변수가 식수 인원을 추정하는 주요 예측 변수임을 확인

▼ 참고한 선행 연구

- Cheng L, Yang IS, and Baek SH (2003). Investigation on the performance of the forecasting model in university foodservice. Journal of Nutrition and Health, 36, 966-973.
- Baek OH, Kim MY, and Lee BH (2007). Menu satisfaction survey for business and industry foddservice workers - Focused on food preferences by gender. Journal of The Korean Society of Food Culture, 22, 511-519.
- Lim JY (2016). (Analysis of forecasting factors affecting meal service in business foodservice (Master's thesis)) , Yonsei University, Seoul.

1. 데이터 수집

분석 데이터 목록 및 설명

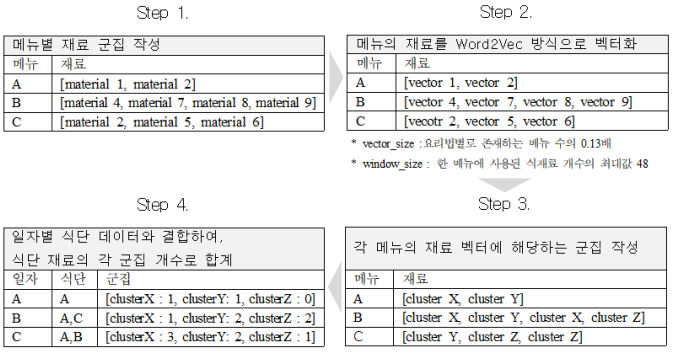
- 대구광역시 A구청으로부터 약 6년 간의 일일 식단표 및 구내식당 카드 결제 데이터 내역을 제공받음
- 이 외에 예측에 필요한 데이터는 공공데이터 포털에서 수집

카테고리	활용데이터 목록	구성 내용	데이터 규모	출처
지자체 데이터	구내식당 일일 식단표	일자별 식단	1,448건	대구시 O구청
	메뉴별 식재료	메뉴, 메뉴별 식재료	1,065건	
	메뉴별 조리법	메뉴, 메뉴별 조리법	917건	
	일자별 구내식당 카드결제 내역	날짜, 성명, 수량, 단가	385,482건	
기상청 데이터	대구광역시 기상 데이터	일 최고/최저기온, 일 강수량 등	-	공공 데이터 포털

2. 변수 생성

- 선행 연구와 EDA를 기반으로 기상 변수(체감온도, 폭염 여부, 강우 여부, 적설 여부)와 시계열 변수(월, 연도, 직전일 식수 인원, 연휴 전날 여부) 생성
- **데이터 검토 중 동일한 메뉴가 다르게 표기된 문제를 발견**
  이에 예측 정확도를 높이기 위해 메뉴의 특성을 반영한 파생변수를 생성하여 예측에 반영함.
  - ex) 돈육김치찌개, 돼지김치찌개
- 메뉴의 이름은 달라도 식재료, 조리법과 같은 메뉴의 본질적인 요소는 유사하다는 아이디어에서 착안하여 두 정보를 활용하여 파생 변수를 생성함.
  1. 조리법 정보(국, 김밥/주먹밥, 김치, 무침/샐러드)를 활용한 파생 변수

     → 조리법을 기준으로 메뉴를 분류
  2. 재료 정보 활용를 활용한 파생 변수
     - 재료의 종류가 약 1,200개로 매우 다양했기 때문에 분류를 통해 범주를 줄임.
     - 재료 분류 시 word2vec과 spherical k-means 방법을 사용하여 연구자의 주관적인 개입을 최소화 함.
     - 아래 프로세스를 통해 데이터 가공



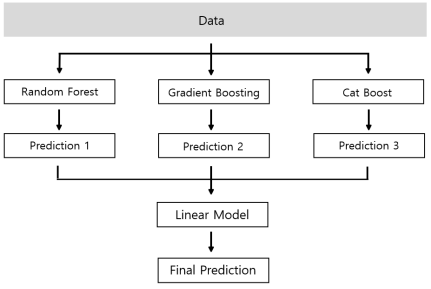
- 앞서 만든 변수들을 일자 기준으로 결합하여 최종 데이터셋을 생성

날짜 변수 / 기상 변수 / 시계열 변수 / 조리법 정보를 반영한 대분류 변수 (메뉴 특성 변수) / 재료 정보를 반영한 재료 군집 변수 (메뉴 특성 변수) / 종속 변수

Date	요일	월	년	연휴 전날	체감 온도	폭염 여부	비/눈	전날_식수인원	과일/음료/과자	구이/볶음	...	김밥/주먹밥_5	김밥/주먹밥_6	빵/떡_0	빵/떡_1	빵/떡_2	빵/떡_3	빵/떡_4	빵/떡_5	빵/떡_6	식수인원
2016-01-18	0	1	2016	0	-8	0	0	265	0.0	1.0	...	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	314
2016-01-19	1	1	2016	0	-13	0	0	314	0.0	2.0	...	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	294
2016-01-20	2	1	2016	0	-11	0	0	294	0.0	0.0	...	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	312
2016-01-21	3	1	2016	0	-10	0	0	312	0.0	1.0	...	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	259
2016-01-22	4	1	2016	0	-7	0	0	259	0.0	0.0	...	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	226
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...

3. 모델링

- 여러 예측 모형(Gradient Boost, XGB, LGBM, CatBoost)중 MSE와 MAE를 기준으로 가장 유의한 모델 3가지(Gradient Boosting, Random Forest, CatBoost)를 블렌딩하여 최종 앙상블 모델 도출
  - <스태킹 앙상블 모형 구조>



  - 스태킹 앙상블을 적용한 결과, MSE가 865.28로 기존 단일 모델들에 비해 약 8~12% 개선되었음.

Model	MSE	MAE
Random Forest	1,056.61	24.85
GradientBoosting	972.41	24.12
CatBoost	943.59	23.88
Ensemble	**865.28**	**22.26**

- **최종 모델의 MAPE는 7.64로 예측값은 실제값에서 7% 정도의 오차만 존재**

4. 분석 결과

웹 서비스

- python flask를 사용하여 웹 서비스 구현
- 날짜, 메뉴를 입력하면 예측값을 도출

실무 적용 후 식수 예측 오차율 약 10%P 개선

- 기존 오차율 (예측 모델 사용 전) : 15~20%
- 개선 오차율 (예측 모델 사용 후): 5~10%

5. 본 프로젝트의 차별점

- 기존 연구는 연구자의 주관적인 판단하에 메뉴를 분류하였으나, 본 프로젝트는 텍스트 마이닝에서 일반적으로 사용되는 단어 임베딩 방법론을 적용하여 연구자의 부분적인 개입을 최소화하고 성능을 향상시킴
- 또한, 한식은 메뉴의 종류가 많고 본질적으로 같은 메뉴라도 다른 이름으로 표기되는 경우가 많음. 이러한 한식의 특성을 반영하여 조리법과 요리법 정보를 활용한 파생변수를 생성함.
  그 결과 기존 데이터를 그대로 사용했을 때보다 성능이 향상됨.

- **메뉴명을 그대로 사용**

Model	MSE
Ensemble	**967.83**

- **조리법, 요리법 정보를 활용**

Model	MSE
Ensemble	**865.28**