구내식당 식수인원 예측 AI 경진대회®

팀명 : 오늘 식단 : 삼계탕 / 오징어 젓갈

Private 순위 : 13위

차례®

구내식당 식수인원 예측 AI 경진대회

진행과정®

팀 병합 전까지 각자의 데이터셋으로 각자 진행

⇒

팀 병합 후 각자의 데이터셋 합산

⇒

메뉴 피처 사용 유무로 서로 다른 3개의 최종 데이터 셋 사용

⇒

CB, LGBM, NGB 모델 사용

⇒

Public 스코어 기준 일정한 비율로 앙상블

데이터셋 설명(1)® Numeric features + menu features

데이터셋명 : ksh

-일자 : 일자 datetime 변환

-요일 : 요일 오류 기입 있어서 자체적으로 변환

-Month : 월

-Corona(더미) : 재택근무비율 0 초과는 1 아니면 0

-Day_gap : 다음출근날까지의 일수로 예를 들어 오늘이 12월 31일 목요일인 경우 다음 출근날은 1월 4일이기 때문에 값이 4

-야근요일 : 식수 인원이 적은 수, 금요일을 1 나머지는 0

-중식메뉴수 : 중식메뉴를 이루고 있는 메뉴 수

-석식메뉴수 : 석식메뉴를 이루고 있는 메뉴 수

-'해산물', '소', '돼지', '닭', '오리', '채소', '재료_기타' : 중식메뉴를 기준으로 해당 재료가 포함되어있는지 여부

-> 석식 보다는 중식의 메뉴 재료가 중요할 것으로 판단

-중식_메인요리 : 중식의 메인요리

-석식_메인요리 : 석식의 메인요리

데이터셋 설명(2)® only Numeric features

데이터셋명 : Untitled_07-62

1.요일, 일자, 년, 월, 일, 주 일자를 이용하여 date_time으로 구한 열

2.코로나 단계 현본사소속재택근무자수의 사분위수를 이용하여 코로나 단계를 추측한 열

3.휴가퍼센트,출장퍼센트 본사정원수와 출장자수, 휴가자수를 이용하여 비율을 구한 후 비율의 사분위수를 이용하여 퍼센트를 주한 열

4.공휴일전후 공휴일인 날짜를 직접 수작업으로 지정하고 원핫인코딩으로 만든 열

5.중식메뉴수,석식메뉴수 중식, 석식의 메뉴를 split 하여 메뉴의 수가 몇 개인지를 나타낸 열

6.출근퍼센트 본사정원수에서 휴가자수,출장자수,재택근무자수를 빼서 출근한 인원을 만들고 이를 사분위수를 이용하여 만든 열

7.계절 1번에서 만든 월과 주를 이용하여 4개로 나누어 계절을 표현한 열, 문자형으로 만들어 원핫인코딩하였음

8.초중말일자 일자(1~30)을 이용하여 월초, 월중, 월말을 문자형으로 만들어 원핫인코딩한 것

그 후 중식계, 석식계에 영향이 있을만한 열을 각각 뽑은 후 LGBM과 NGB를 앙상블하여 만듬

데이터셋 설명(3)® Numeric features + menu features

데이터셋명 : Untitled_08_63.5

1.메뉴를 이용한 정보

A.먼저 띄어쓰기 및 오타를 수정

B.레시피+기본정보_20210712와 레시피+재료정보_20210712, 그리고 대분류중분류 등의 csv를 불러서 메뉴의 종류를 크게 23가지로 나누어 주었고 일반적으로 인기가 많은 메뉴, 자주 나오지 않는 메뉴 등은 가중치를 크게 해줌

C.가중치가 메뉴마다 다르기 때문에 그 날의 식단이 인기있는지를 알아볼 수 있는 메뉴점수의 합(총점수)를 이용한 열을 만들어 줌

2.년,월,일,주,요일 날짜를 이용하여 년, 월, 일, 주, 요일을 만들어 줌

3.출근 본사정원수에서 휴가자, 출장자, 재택근무자를 제외하여 출근한 인원 수를 만들어 줌

4.휴가비율, 출장비율, 야근비율, 재택비율 휴자가, 출장자, 시간외근무명령서승인, 재택의 수를 정원수로 나누어 비율 열을 만들어 줌

5.계절 월을 이용하여 계절을 지정해줌

6.연도별로 groupby하여 재택, 야근, 출장, 휴가자 수의 평균, 최댓값, 최솟값을 만들어 줌

7.일별로 groupby하여 재택, 야근, 출장, 휴가자 수의 평균, 최댓값, 최솟값을 만들어 줌

8.주별로 groupby하여 재택, 야근, 출장, 휴가자 수의 평균, 최댓값, 최솟값을 만들어 줌

9.요일별로 groupby하여 재택, 야근, 출장, 휴가자 수의 평균, 최댓값, 최솟값을 만들어 줌

10.계절별로 groupby하여 재택, 야근, 출장, 휴가자 수의 평균, 최댓값, 최솟값을 만들어 줌

11.Mean_encoding을 이용하여 요일별 중식계, 석식계의 평균을 열로 만들어 줌

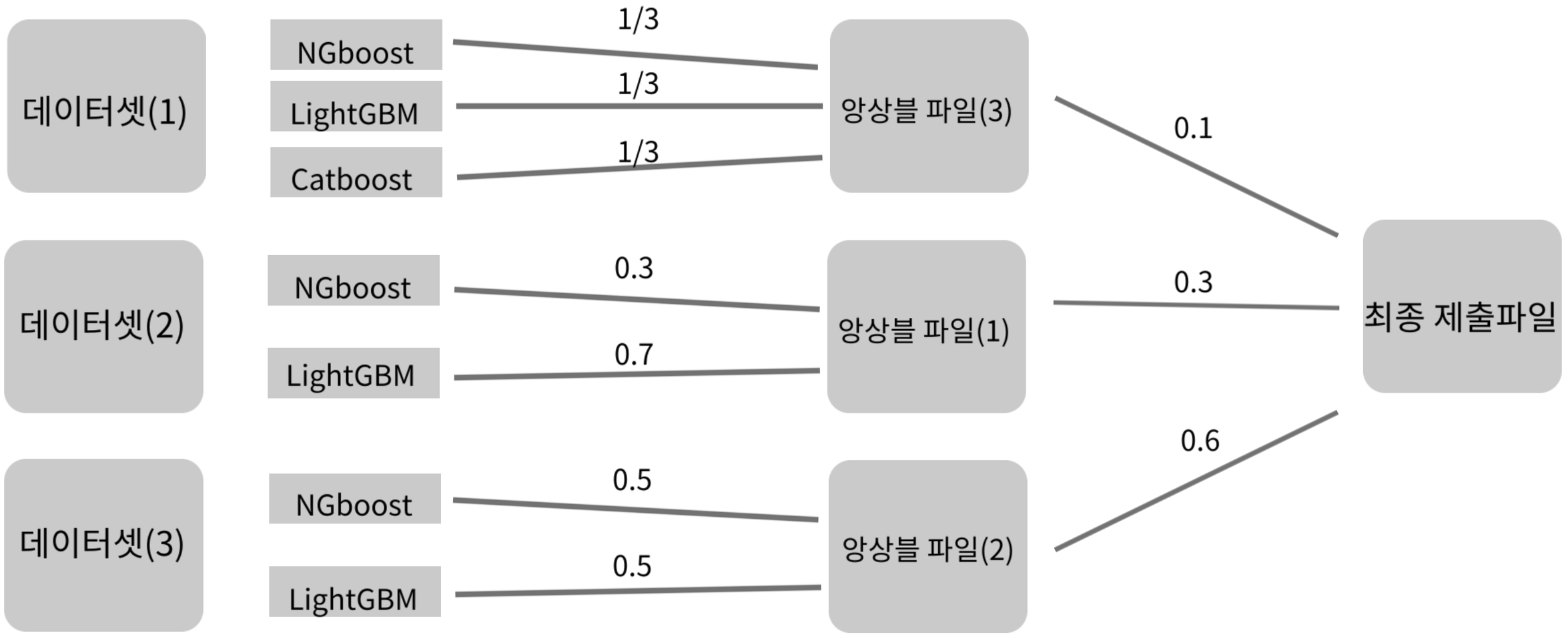
12.Mean_encoding을 이용하여 계절별 중식계, 석식계의 평균을 열로 만들어 줌

'본사정원수', '본사휴가자수', '본사출장자수', '본사시간외근무명령서승인건수', '현본사소속재택근무자수','일' 등 원 데이터에 과적합이 일어날 것으로 보이는 열은 제외시킨 후 진행하였음

LGBM을 이용하여 Feature-Selection을 진행하여 뽑힌 열들 중 상관관계가 높은 열을 제외시킨 후 모델링을 진행

모델링은 1번째의 서브미션과 동일하게 LGBM과 NGB의 앙상블을 이용하여 진행하였음

모델링 과정®


데이터셋(1)
NGboost
LightGBM
Catboost
1/3
1/3
1/3
앙상블 파일(3)
데이터셋(2)
NGboost
LightGBM
0.3
0.7
앙상블 파일(1)
데이터셋(3)
NGboost
LightGBM
0.5
0.5
앙상블 파일(2)
0.1
0.3
0.6
최종 제출파일

더 시도해봤어야하는 것들®

1. 외부데이터 사용 - 코로나 데이터 등 많은 관련된 외부데이터가 있었지만, data leakage를 우려해 아예 사용하지 않았다. 우려했던 부분을 잘 처리하고 사용해봤어야 했다.
2. 메뉴 데이터 처리 - 메뉴를 BPE처리해서 임베딩 처리했었는데 이 방법 외에도 메뉴 데이터를 더 효율적으로 사용할 수 있는 방법을 찾아봤어야 했다.
3. 모델링 - 더 다양한 모델을 찾아보고 모델링을 시도했어야 했다.

아쉬운 점®

1. 적은 데이터에 비해 복잡한 접근법으로 인해 아쉬움이 크다. 과적합 방지는 모델링에서 정말 중요한 요소라는 것을 깨달음.
2. 메뉴 데이터를 좀 더 효율적으로 쓸 수 있는 방법을 못 찾은 부분이 너무 아쉽다.
3. 적은 데이터로 인해 파생변수를 별로 못 만들었다.
4. 데이터를 다루는 것에서 좋은 아이디어가 없어 과적합으로 결과를 마무리한 것이 아쉽다.

느낀 점®

1. Simple is best라는 말이 있다. 다른 분들의 솔루션을 보니 기초부터 차근차근 간단하게 접근하는 것이 중요하다는 것을 깨달았다.
2. 문제에 접근하는 데 있어서 편견을 가지고 하면 안되겠다는 것을 느꼈다.
3. 서브미션의 종류가 다양하여 앙상블하면 효능이 좋다는걸 깨달았다.