빅데이터 페스티벌

음식점 만족도 요인 분석과 리뷰 기반의 추천시스템 : 부산 음식점을 중심으로

2020150406 최민경, 2020150473 김지후

A PRESENTATION ABOUT

1 주제 선정 이유 및 필요성

2 음식점 만족도 예측 모델

3 리뷰기반의 추천시스템

4 의의 및 발전방향

주제 선정 이유 및 필요성

자영업자의 대출규모
887조 5000억원
전년 동기 대비 14.2%증가

포스트 코로나 시대 지속 가능한 외식산업

(1) 음식점 관점:
소상공인의 생태계를 보호하고 지역의 상권의 활성화를 위해서 음식점 만족도에 영향을 미치는 요인을 분석

(2) 소비자의 관점:
소비자가 적절하고 편리하게 음식점을 선택할 수 있도록 돕기 위해 음식점을 추천해주는 추천 시스템을 구현

음식점 만족도 예측 모델

1) 데이터 설명
2) 전처리
3) 모델링
4) 분석 결과 및 한계

데이터 설명

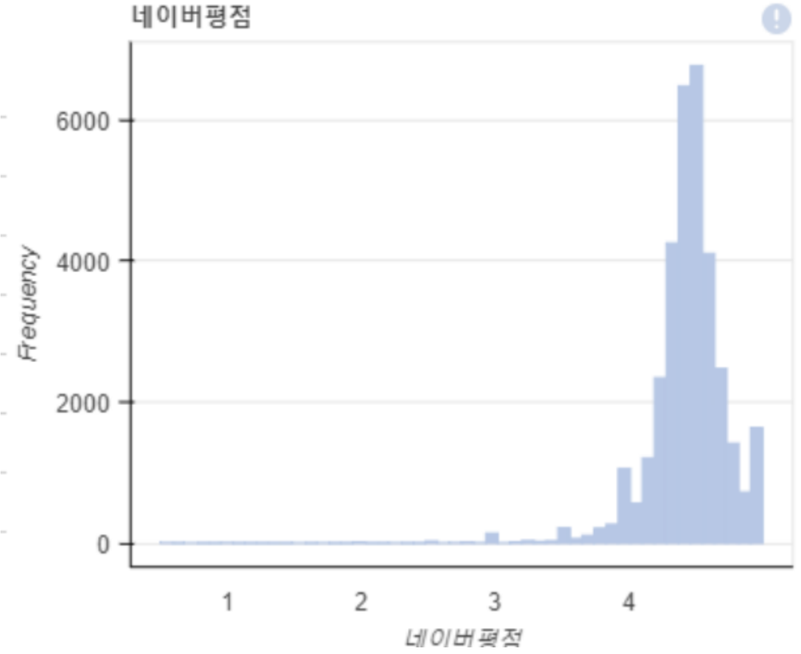
데이터 설명:
Data Explanation

- 공공데이터포털 부산관광공사 음식테마거리 음식점 정보 (https://www.data.go.kr/data/15096715/fileData.do)

- 부산시의 구별 인구수, 호텔수, 문화재 개수, 관광사업체수, 자동차 수, 공원 수 등의 추가 수집 데이터

식당명	지역명_y	식당위도	식당경도	영업신고증업태명	영업인허가명	음식점소개내용	주차가능여부	와이파이제공여부	놀이방유무	반려동물입장가능	다국어메뉴판제공	화장실정보내용	택배판매유무
커피탱크	사하구	35.081	128.9682	다방	휴게음식점	부산광역시 사하	0	0	0	0	0	0	
명지횟집	강서구	35.1131	128.9316	한식	일반음식점	"명지횟집"은 부	0	0	0	0	0	0	
미니스톱(연산시	연제구	35.1813	129.0771	기타(편의점)	일반음식점	"미니스톱(연산시	0	0	0	0	0	0	
썬더치킨	동구	35.1171	129.0395	호프/통닭	일반음식점	부산광역시 동구	0	0	0	0	0	0	
장터한식	동래구	35.2056	129.0872	기타	일반음식점	부산광역시 동래	0	0	0	0	0	0	
용두산식당	중구	35.1002	129.0344	한식	일반음식점	부산광역시 중구	0	0	0	0	0	0	
섹션	사상구	35.1725	128.9825	한식	일반음식점	부산광역시 사상	0	0	0	0	0	0	
목포녹동세발낙	사상구	35.1696	128.9808	회집	일반음식점	부산광역시 사상	0	0	0	0	0	0	
조은데이	사상구	35.1693	128.9792	한식	일반음식점	"조은데이"는 부	0	0	0	0	0	0	
해동숯불생고기	사상구	35.1427	128.9842	한식	일반음식점	부산광역시 사상	0	0	0	0	0	0	
개코포자	동구	35.1293	129.0479	호프/통닭	일반음식점	부산광역시 동구	0	0	0	0	0	0	
카페루미(경성대	남구	35.1369	129.1005	기타	일반음식점	어디 가야 할지 고	0	0	0	0	1	0	
유성커피숍	사하구	35.1039	128.9719	다방	휴게음식점	부산광역시 사하	0	0	0	0	0	0	
백구당	중구	35.1048	129.0357	제과점영업	제과점영업	부산광역시 중구	0	0	0	0	1	0	
청탑	중구	35.0979	129.0332	경양식	일반음식점	부산광역시 중구	0	0	0	0	0	0	
경복	중구	35.0975	129.0292	다방	휴게음식점	부산광역시 중구	0	0	0	0	0	0	
유앤아이커피 ?	중구	35.099	129.0289	다방	휴게음식점	부산광역시 중구	0	0	0	0	0	0	
세명	중구	35.099	129.0289	다방	휴게음식점	부산광역시 중구	0	0	0	0	0	0	
까페그루	서구	35.1162	129.013	경양식	일반음식점	부산광역시 서구	0	0	0	0	0	0	
원조18번완당	서구	35.1057	129.0205	분식	일반음식점	2021년 06월 28	1	0	0	0	1	1	
홍화반점	부산진구	35.1602	129.0649	중국식	일반음식점	부산광역시 부산	1	0	0	0	1	0	
하동집	서구	35.0994	129.0226	한식	일반음식점	부산광역시 서구	0	0	0	0	1	0	
시민제과	서구	35.0967	129.0236	제과점영업	제과점영업	부산광역시 서구	0	0	0	0	1	0	
묘향정	영도구	35.0909	129.0628	중국식	일반음식점	부산광역시 영도	0	0	0	0	0	0	
김밥천국	영도구	35.0915	129.041	분식	일반음식점	부산광역시 영도	0	0	0	0	0	0	
한성	영도구	35.0898	129.0381	다방	휴게음식점	부산광역시 영도	0	0	0	0	0	0	
통학	중구	35.1055	129.0238	다방	휴게음식점	부산광역시 중구	0	0	0	0	0	0	
중화각	동구	35.1274	129.0403	중국식	일반음식점	어디 가야 할지 고	0	0	0	0	0	0	
밀양보신탕	동구	35.1291	129.0462	한식	일반음식점	무엇을 먹을지 고	0	0	0	0	0	0	
만춘반점	동구	35.1356	129.0594	중국식	일반음식점	어디 가야 할지 고	0	0	0	0	0	0	
덕화원	동구	35.1298	129.0486	중국식	일반음식점	부산광역시 동구	0	0	0	0	0	0	
동흥 중화요리	동구	35.1197	129.0412	중국식	일반음식점	어디 가야 할지 고	0	1	0	0	0	0	

데이터 설명



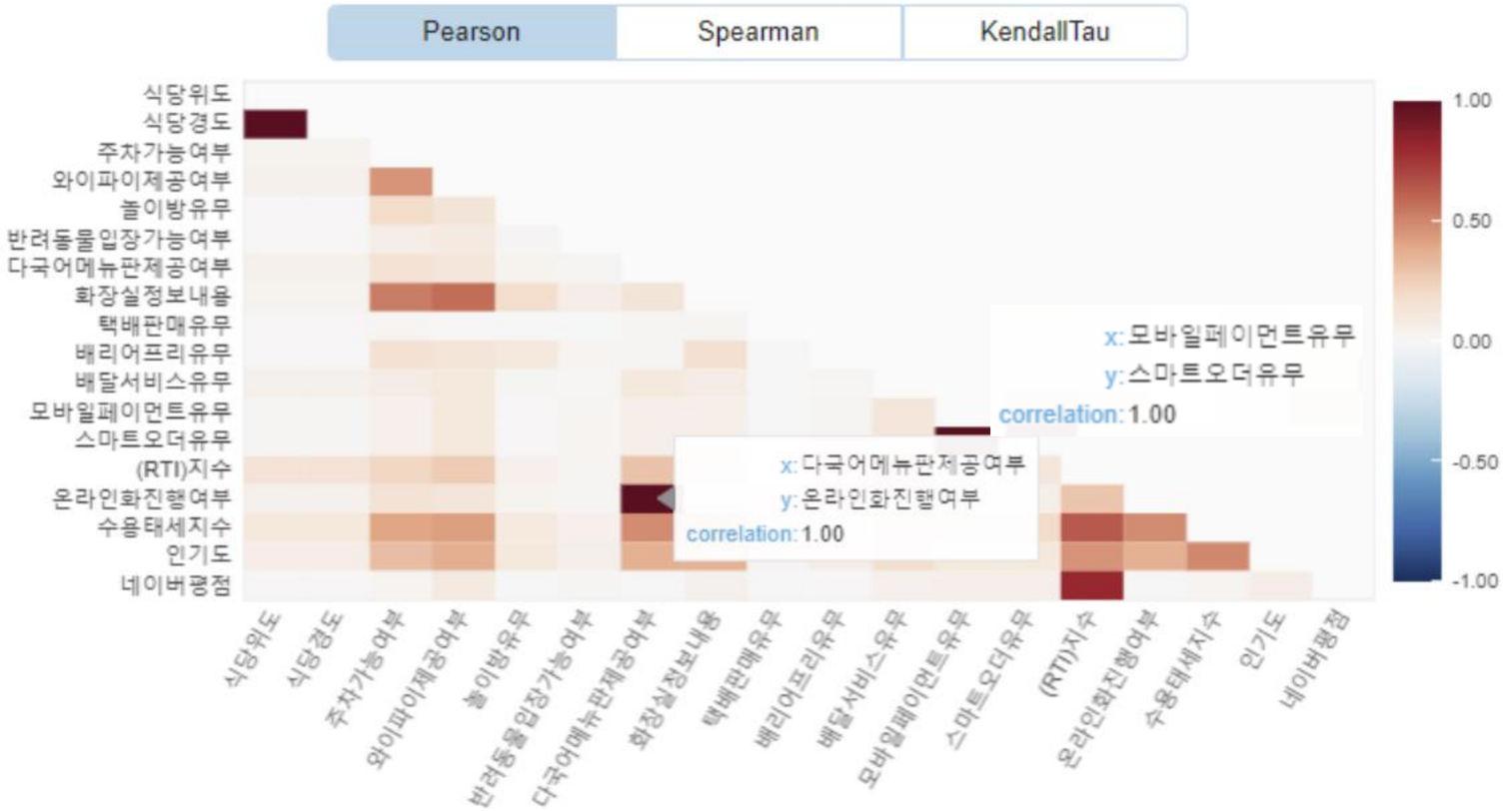
목적변수 : 네이버평점

- 약간 오른쪽으로 치우쳐진 분포

- train set: 목적변수가 결측치가 아닌 음식점
- test set: 목적변수가 결측치인 음식점

Approximate Distinct Count	191	Mean	4.4384
Approximate Unique (%)	0.6%	Minimum	0.5
Missing	16321	Maximum	5
Missing (%)	32.3%	Zeros	0
Infinite	0	Zeros (%)	0.0%
Infinite (%)	0.0%	Negatives	0
Memory Size	534.2 KB	Negatives (%)	0.0%

변수 선택:
Feature Selection



- correlation coefficient가 1인 변수 중 하나 제거
  : 온라인화진행여부, 스마트오더 유무 변수 제거

- 모델링에 필요없는 변수 제거
  : 식당명, 지역명, 영업신고증업태명, 음식점소개내용 변수 제거

전처리:
Preprocessing

전처리: Preprocessing

1. 범주형 변수 0,1변환

주차가능여부	와이파이제공여부	놀이방유무	반려동물입장가능여부	다국어메뉴판제공여부	화장실정보내용	택배판매유무	배리어프리유무	배달서비스유무	모바일페이먼트유무
0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	1	0
0	0	0	0	0	0	0	0	0	0
0	0	0	0	1	0	0	0	0	0
0	0	0	0	1	0	0	0	0	0

2. 표준화

```
#표준화
from sklearn.preprocessing import StandardScaler
from sklearn.decomposition import PCA
scaler = StandardScaler()

X_train_num = X_train[['인구','호텔','문화재','관광사업체','자동차','공원']]
X_val_num = X_val[['인구','호텔','문화재','관광사업체','자동차','공원']]

X_train_std = scaler.fit_transform(X_train_num)
X_val_std = scaler.transform(X_val_num)
```

모델링:
Modeling

모델링: Modeling

Linear Regression
Random Forest Regressor
Decision Tree Regressor
XGB Regressor
LGBM Regressor
CatBoost Regressor

Model	RMSE
Linear Regression	0.288861
Random Forest Regressor	0.290685
Decision Tree Regressor	0.291403
XGB Regressor	0.288593
LGBM Regressor	0.289186
CatBoost Regressor	0.289680

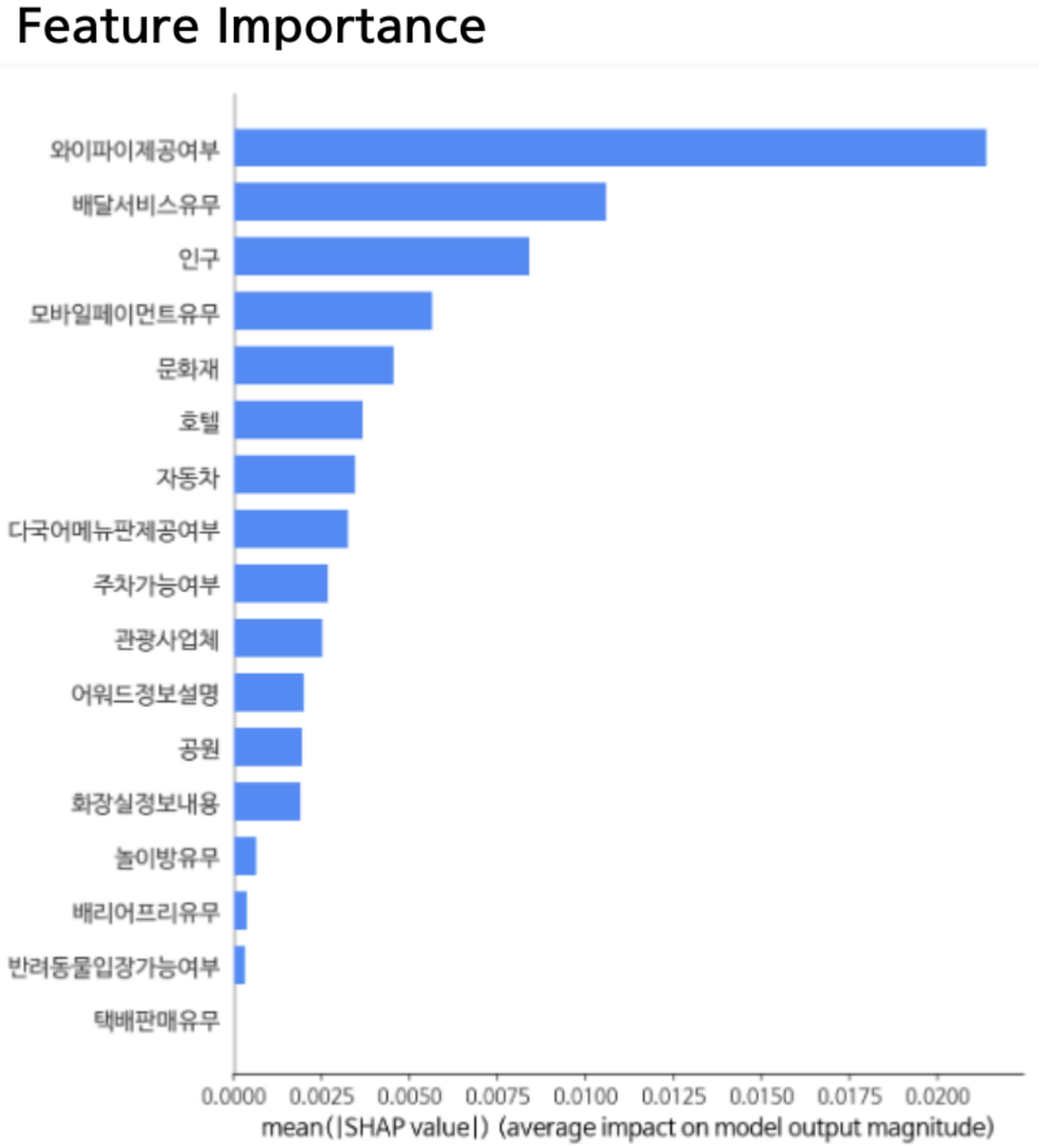
모델링: Modeling

Linear Regression
Random Forest Regressor
Decision Tree Regressor
XGB Regressor
LGBM Regressor
CatBoost Regressor

Model	RMSE
Linear Regression	0.288861
Random Forest Regressor	0.290685
Decision Tree Regressor	0.291403
XGB Regressor	0.288593
LGBM Regressor	0.289186
CatBoost Regressor	0.289680

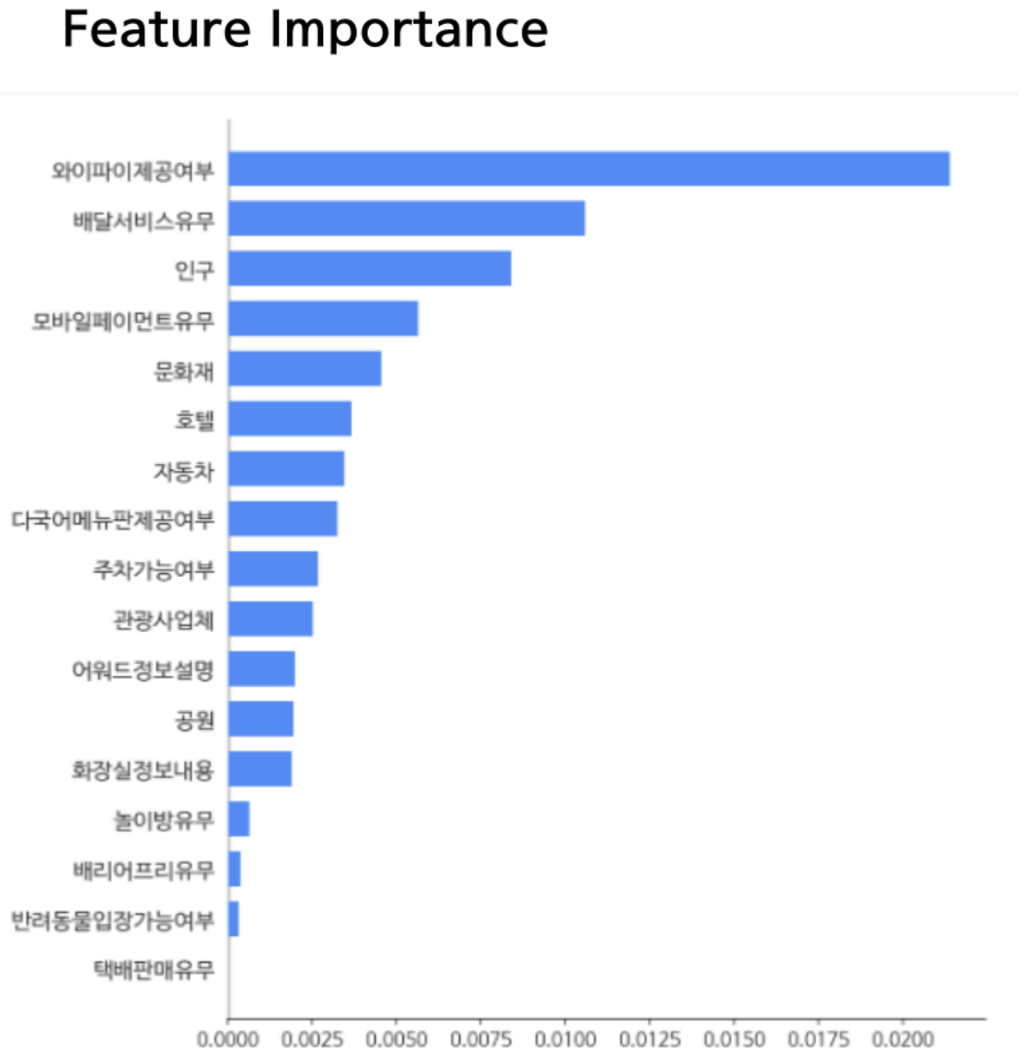
모델링: Modeling

XGB Regressor



모델링: Modeling

XGB Regressor



- 와이파이제공여부, 배달서비스유무, 모바일 페이먼트 유무 등 음식점의 서비스와 관련된 변수들이 만족도에 큰 영향력을 미친다.

- 근처 인구, 문화재의 개수 등 주변 환경 역시 만족도에 영향력을 미친다.

예측결과

name	네이버평점	주차가능여부	와이파이제공여부	놀이방유무	반려동물입장가능여부	다국어메뉴판제공여부	화장실정보내용	택배판매유무	배리어프리유무	배달서비스유무	모바일페이먼트유무	어워드정보설명	인구	호텔	문화재	관광사업체	자동차	공원
새진주집(중구)	4.346982	0	0	0	0	0	0	0	0	0	0	2	42609	8	6	30	35408	4
풍배식당(동구)	4.357748	0	0	0	0	0	0	0	0	0	0	1	89712	7	16	42	52490	5

예측한 네이버 평점이 낮은 두 식당의 경우, 서비스적 부분이 아예 제공되지 않았고 인구도 적은 편이이었다.

name	네이버평점	주차가능여부	와이파이제공여부	놀이방유무	반려동물입장가능여부	다국어메뉴판제공여부	화장실정보내용	택배판매유무	배리어프리유무	배달서비스유무	모바일페이먼트유무	어워드정보설명	인구	호텔	문화재	관광사업체	자동차	공원
팻그라운드(북구)	4.612503	1	1	0	1	0	1	0	0	0	0	0	285390	0	8	0	103651	79
밀회관(부산덕천역점)(북구)	4.580340	0	1	0	0	0	1	0	0	0	1	0	285390	0	8	0	103651	79

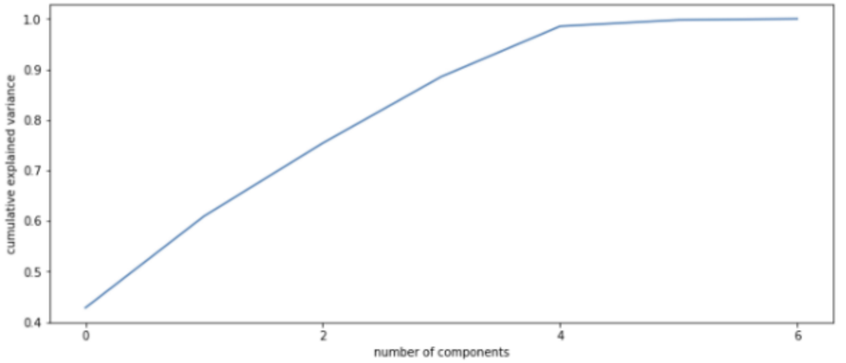
예측한 네이버 평점이 높은 두 식당의 경우, 서비스적 부분이 많이 제공되는 편이었고 인구도 많다.

추가 시도:
Further Try

추가 시도:
Further Try

1. PCA

: 과대적합 방지 및 변수의 개수를 줄이기 위해 PCA진행



2. Hyperparameter -OPTUNA

```
#lgbm regressor optuna hyperparameter 설정 - rmse 값 가장 낮은 것 기준
def objective_ET(trial):
    param = {
        "n_estimators": trial.suggest_int("n_estimators", 90, 700, step = 10),
        "num_leaves": trial.suggest_int("num_leaves", 31, 40, step = 1),
        "max_depth" : trial.suggest_int("max_depth", -1, 30, step = 2),
        "learning_rate": trial.suggest_float("learning_rate",0.05, 0.5),
        "objective": trial.suggest_categorical("objective", ['regression']),
        "random_state": trial.suggest_categorical("random_state", [2022]),
        "subsample": trial.suggest_discrete_uniform('subsample', 0.5, 1, 0.05),
        'reg_alpha': trial.suggest_loguniform('reg_alpha', 1e-3, 10.0),
        'reg_lambda': trial.suggest_loguniform('reg_lambda', 1e-3, 10.0),
        "colsample_bytree": trial.suggest_discrete_uniform('colsample_bytree', 0.5, 1, 0.1),
        'min_child_samples': trial.suggest_int('min_child_samples', 1, 300)
    }
    model = LGBMRegressor(**param)
    cc = cross_val_score(model, X_train, y_train, scoring = 'neg_mean_squared_error', cv = 5)
    avg_rmse = np.mean(np.sqrt(-cc))


    return avg_rmse
```

3. Voting Ensemble

: 투표를 통해 최종 예측결과를 결정

```
from sklearn.ensemble import VotingRegressor
estimators = []
rf_model =RandomForestRegressor(n_estimators = 640, max_depth = 13, random_state = 2022, min_samples_split = 3,
                                min_samples_leaf = 1)
estimators.append(('randomforest', rf_model))
xg_model = xgboost.XGBRegressor(n_estimators = 200, max_length = 11, learning_rate = 0.2265403685269807,
                                subsample = 0.8 ,random_state = 2022,
                                colsample_bytree = 0.8, min_child_samples = 8)
estimators.append(('xg', xg_model))
cat_model = CatBoostRegressor()
estimators.append(('cat', cat_model))
lgbm_model = LGBMRegressor(n_estimators = 220, num_leaves = 40, max_depth = -1, learning_rate = 0.05020711323769501,
                           objective = 'regression', subsample = 0.8 ,
                           reg_alpha = 0.5034144173716503, reg_lambda =  0.06815185964124228, random_state = 2022,
                           colsample_bytree = 0.8, min_child_samples = 7)
estimators.append(('lgbm', lgbm_model))
```

```
ensemble = VotingRegressor(estimators)
ensemble.fit(X_train, y_train)
pred = ensemble.predict(X_val)

rmse = mean_squared_error(y_val, pred, squared = False)
print(rmse)
```

0.2708012953368 49

성능개선!

추가 시도:
Further Try

한계

1) 데이터의 한계:

우리가 사용한 변수는 주변 환경, 가게의 서비스, 시스템 등의 내용을 담고 있다.
하지만 맛, 분위기, 가격 등 음식점을 판단하는데 있어 중요한 변수들을 고려하지 못해서 성능이 좋지 않았다.

2) 분석 목적의 한계:

분석 목적이 음식점 만족도에 영향을 미치는 요인을 분석하는 것이었는데, 여기에 머신러닝을 적용하여 성능을 높이기 위해 복잡한 모델, 방법을 사용할수록 해석에는 어려움이 있었다.

리뷰기반의 추천시스템

1) 데이터 설명
2) 전처리
3) 모델링
4) 분석 결과

데이터 설명

데이터 설명

맛집 리뷰 사이트 "망고플레이드"에서 음심점 리뷰 993개 크롤링

	리뷰	맛	식당명	가고싶다	전체평점	주소	음식종류
0	₩n 두번째 방문인데 요번에도 4시간 웨이팅토요일 10:30 테이블...	맛있다	톤쇼우	1,624	4.7	부산시 수영구 민락동 181-20	까스 요리
1	₩n 부산 최고의 웨이팅 핫플서울토박이로 살다가 간만에 부산을 놀러...	맛있다	톤쇼우	1,624	4.7	부산시 수영구 민락동 181-20	까스 요리
2	₩n 제 목숨을 바쳐도 좋어요.충청도 사는 소녀는 톤쇼우를 먹으러 ...	맛있다	톤쇼우	1,624	4.7	부산시 수영구 민락동 181-20	까스 요리
3	₩n 부산가서 무슨 돈까스 먹냐는 약간의 공격을 당했지만후기보니 너...	맛있다	톤쇼우	1,624	4.7	부산시 수영구 민락동 181-20	까스 요리
4	₩n 캬 사진만으로도 너무나 가고싶었지만 웨이팅이 엄두가 안났던 톤...	맛있다	톤쇼우	1,624	4.7	부산시 수영구 민락동 181-20	까스 요리

각 음식점 당 10개의 리뷰를 결합
식당의 특징을 나타내는
하나의 글로 사용

총 193개의 식당

	식당명	unique
0	1984나폴리	[₩n 위치가 골목 안이라 분위기가 새롭고 하몽피자 새롭다 새로와₩...
1	303화덕	[₩n 그냥 그냥 맛있는건 아니구요₩n , ₩n ...
2	Cafe de 220VOLT	[₩n 커피(?)원두(?)머신(?)회사인 것 같았다.전문적으로 보이...
3	가온밀면	[₩n 해운대역 근처에 위치한 '가온밀면'- 밀면, 만두부산에서 먹...
4	거대갈비	[₩n 부산 해운대에 위치한 "거대갈비". 해운대에서 유명한 고기집...
...	...	...

전처리:
Preprocessing

전처리

1) 정규표현식을 사용해 **부호 제거**

2) Spacing 패키지를 사용해 **뛰어쓰기 교정**

	식당명	unique	text
0	1984나폴리	[₩n 위치가 골목 안이라 분위기가 새롭고 하몽피자 새롭다 새로와₩...	위치가 골목 안이라 분위기가 새롭고 하몽피자 새롭다 새로와 ...
1	303화덕	[₩n 그냥 그냥 맛있는건 아니구요₩n , ₩n ...	그냥 그냥 맛있는건 아니구요 화덕피자는 첨이예요...
2	Cafe de 220VOLT	[₩n 커피(?)원두(?)머신(?)회사인 것 같았다.전문적으로 보이...	커피 원두 머신 회사인 것 같았다 전문적으로 보이는 많은 부품과 머신들 ...
3	가온밀면	[₩n 해운대역 근처에 위치한 '가온밀면'- 밀면, 만두부산에서 먹...	해운대역 근처에 위치한 가온밀면 밀면 만두부산에서 먹은 밀면 중에 가장 청결...
4	거대갈비	[₩n 부산 해운대에 위치한 "거대갈비". 해운대에서 유명한 고기집...	부산 해운대에 위치한 거대갈비 해운대에서 유명한 고기집 중 한 곳이지요 와...

최종 데이터

모델링:
Modeling

형태소 분석기

Mecab

한국어 형태소 분석기로 일본어용 형태소 분석기를 한국어를
사용할 수 있도록 수정한 것입니다.

위치가 골목 안이라 분위기가 새롭고 하몽 피자 새롭다

['위치', '가', '골목', '안', '이', '라', '분위기', '가', '새롭', '고', '하몽', '피자', '새롭', '다']

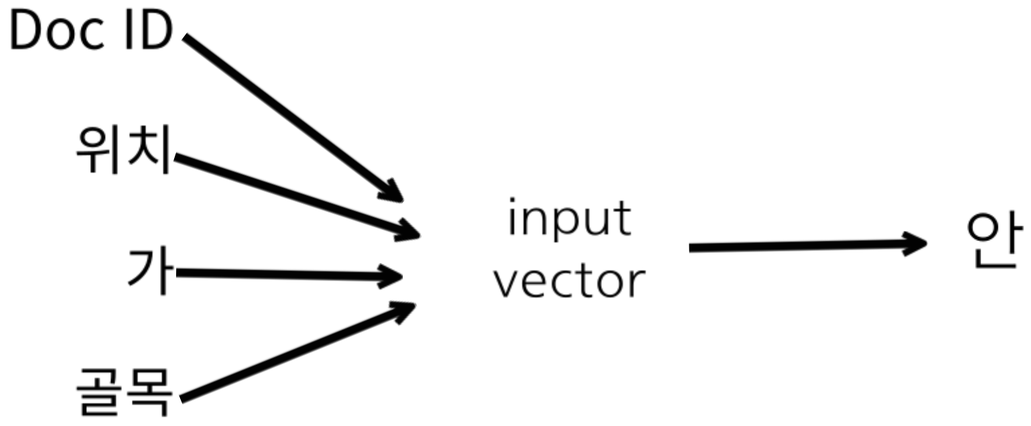
모델링

Doc2Vec

Word2Vec은 단어를 임베딩하는 워드 임베딩 알고리즘,

Doc2Vec은 Word2Vec을 변형하여 문서의 임베딩을 얻을 수 있도록 한 알고리즘입니다.

[**'위치', '가', '골목', '안',** '이', '라', '분위기', '가', '새롭', '고', '하몽', '피자', '새롭', '다']



k개 단어와 함께 문서 ID를 넣어 다음 단어를 예측하는 과정으로 학습

분석 결과

결과

모델을 불러와 입력 받은 문장을 벡터화한 후

기존의 벡터화된 문장들(태그)과와 코사인 유사도를 비교하여 상위 10개의 태그와 유사도를 출력

추천 결과

호찐빵

('마가만두', 0.4616379737854004)
('비앤씨', 0.3801164031028747 6)
('고래사', 0.366252839565277 1)
('남천동 보성녹차 팥빙수', 0.362138301134109 5)
('전포명가떡집 (휴업중)', 0.34528601169586 18)
('흰여울비치', 0.34382781386375 43)
('노홍만두', 0.34201702475547 79)
('봉샌드', 0.33881527185440063)
('딤타오', 0.33691710233688354)
('해운대소문난암소갈비집', 0.32140094041824 34)

가온밀면

('본가제일면가', 0.51917445659637 45)
('해운대밀면', 0.51121455430984 5)
('대성밀냉면', 0.50524812936782 84)
('원조밀면', 0.49640083312988 28)
('원조부산밀면', 0.46773278713226 32)
('국제밀면', 0.46408554911613464)
('내호냉면', 0.46064314246177673)
('노홍만두', 0.44128662347793 58)
('공원칼국수', 0.41909927129745483)
('해운대31cm해물칼국수', 0.40066802501678467)

결과

실제 리뷰 비교

호찐빵

음 찐 빵 맛집으로 소문난 호빵집 방문 가는 날이 장날이라 비가 생각보다 많이 왔고 그래서인지 대기하는 사람이 많지는 않았던 거 같다 그래도 차들은 줄 서서 대기하고 있어 맛집 인증 찐빵과 왕만두를 구매했고 망플의 가격보다 500원 인상된 것을 확인할 수 있었다 사자마자 맛보았는데 단팥은 생각보다 달지 않았고 팥의 양은 꽤 넉넉했다 그래 봐야 찐빵 아니겠냐며 갈까 말까 하던 나 자신 반성합니다 통팥소와 결대로 부드럽게 찢어지는 외 피미친 회전율 너무 착한 가격까지 뭐 하나 흠잡을 것도 잡고 싶지도 않아 그냥 다 완전 짱이야 정말 맛있다 여태 찐빵 먹어 본 것 중에선 이집이 가장 맛있었습니다 가격도 착하고 다음에 기장을 지나갈 일이 있으면 또 들리고 싶은 집이였습니다 (중략)

마가만두

만두와 요리가 메인이다 이 집 만두는 옛날 스타일 도 톰하고 투박한 만두 찐만두와 군 만두 그리고 유산 슬밥을 우리는 주문했다 군만두는 피가 도톰하고 안에 육즙이 잘 들어 있다 인위적인 게 아니라 적당한 양의 육즙 갓 튀겨 나와 바삭한 맛이 좋다 유산 슬밥도 버섯 오징어 새우 청경채 죽순 등 다양하게 들어갔는데 양이 적지 않다 꽤나 배가 부르다 찐만두가 가장 늦게 나왔는데 군만두보다 육즙이 더 가득 호호 불어 먹는 게 꽤 맛있다 오늘도 마가 만두를 맛 나게 먹었지만 다음번엔 신발원 가고 싶다 먹히영 만두는 당연하고 다른 요리도 맛있었어요 깔끔 담백했다 부산에 올 때마다 갈 집이다 (중략)

의의 및 발전방향

의의

1) **음식점, 소비자 각각의 관점**에서의 외식 산업을 분석했습니다.

2) 예측모델을 통해 **정형데이터**를 사용하고 추천시스템을 통해서 **비정형데이터**를 다뤄볼 수 있었습니다.

3) 성능향상을 위해 **PCA, 앙상블, 초모수 조절** 등 다양한 시도를 했습니다.

4) **가공되지 않은 데이터**를 다루고 추가 **데이터 탐색과정**을 통해 예측모델의 설명력을 높이기 위해 노력했습니다.

발전방향

1) 예측모델의 경우, 음식점 가격, 맛, 분위기 등의 추가적인 데이터 수집을 통해 **더 정확한 예측이 가능한 모델**로 발전할 수 있습니다.

2) 추천시스템의 경우, 소**비자에 대한 정보를 추가로 반영해 소비자 기반의 추천시스템**을 구현하면 소비자 맞춤형 추천시스템으로 발전할 수 있습니다.

3) 부산 지역에서만 그치지 않고 전국의 데이터를 활용하여 **전지역에서 활용가능한 예측 모델**이 될 것으로 기대됩니다.

THANK YOU!