서울특별시 구별 시간대별 배달주문정도 예측 모델 개발

AI-11기 김현승

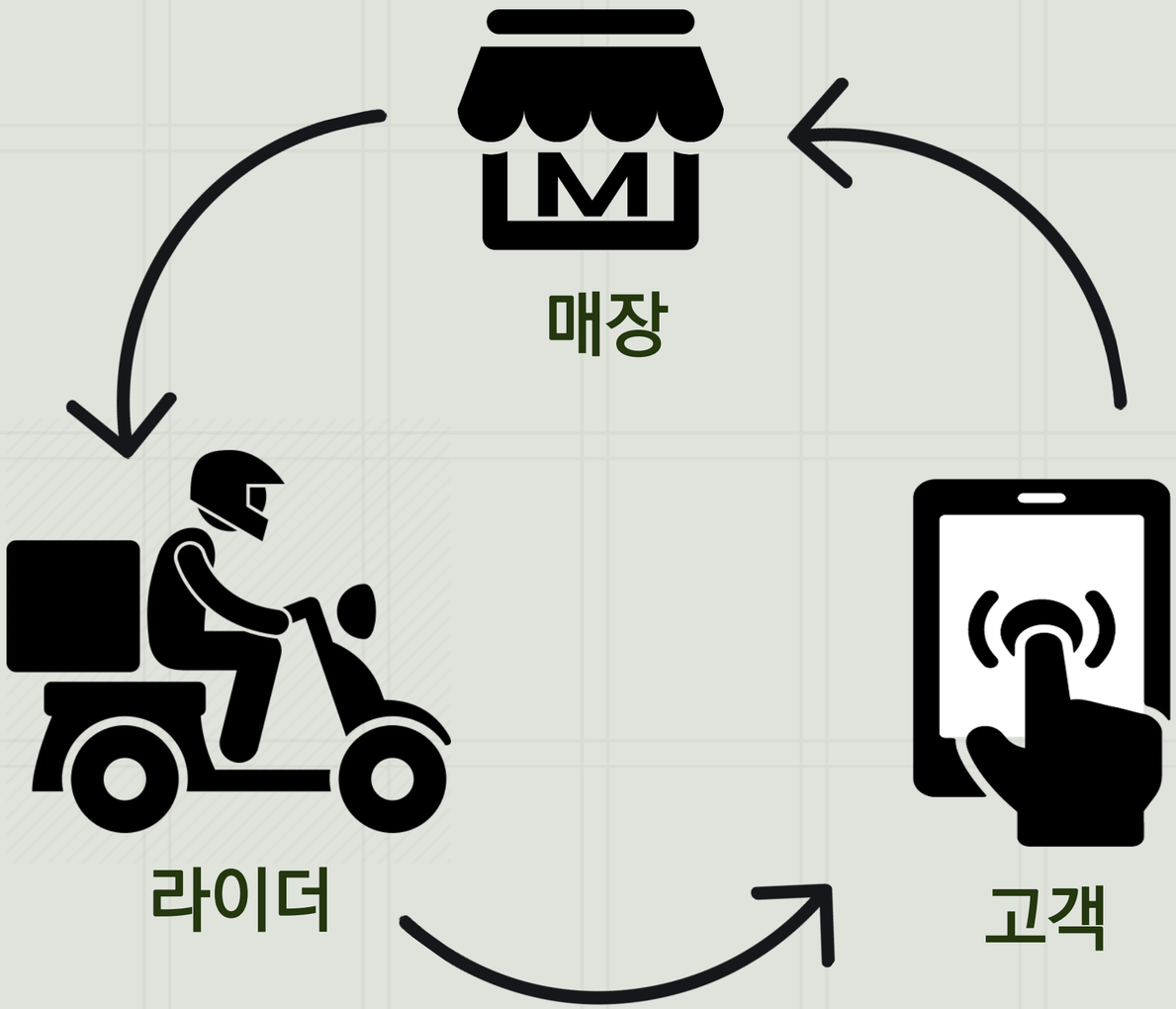
CONTENTS

프로젝트 배경

코로나 시작: 2020. 01 ~
사회적 거리두기: 2020. 02 ~

배달 시장의 증가 → 라이더 직업 증가

이해 관계자 맵



가설

1. 날씨(기온,적설량, 미세먼지 등등) 변수들은 주문정도에 큰 영향을 미칠 것이다.

2. 구별 인구수가 더 많은 구가 주문을 더 많이 시켰을 것이다.

3. 축구경기가 있는 날이면 사람들이 치킨을 더 많이 시킬 것이다.

목적

서울 특별시 구별, 시간대별 배달 주문의 정도(주문 많음, 보통, 적음)을 예측함으로써

어떤 시간대에 어떤 구가 주문이 많은가?
(효율적인 배달 운행)

나의 매장은 어떤 시간대에 주문이 많을까?
(재고관리, 일할 때 시간분배)

02

데이터 전처리

데이터 설명

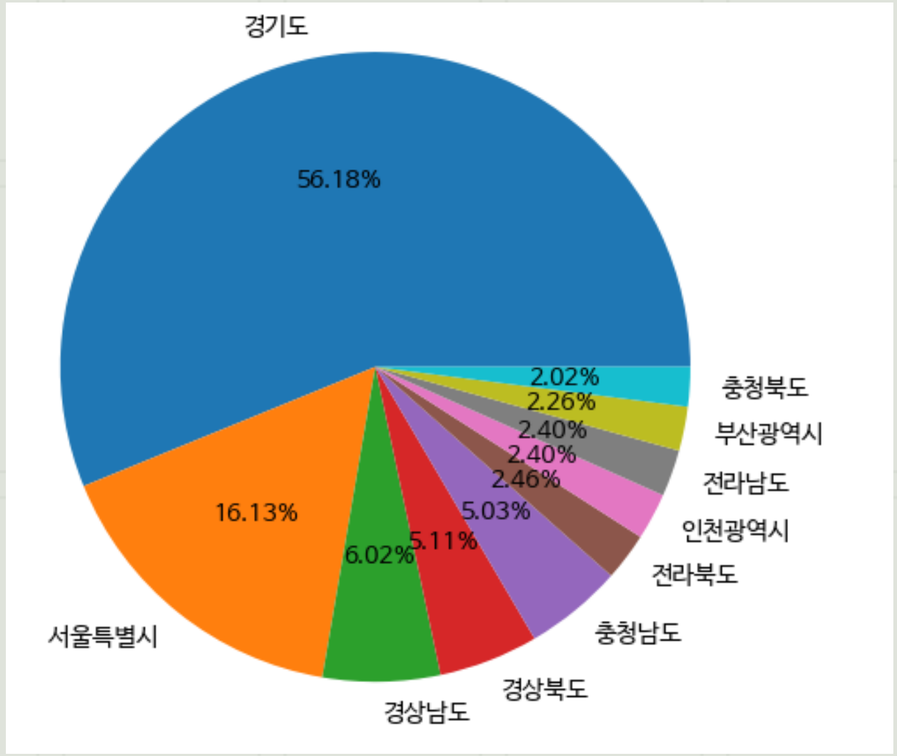
전국 구별, 시간대별 주문건수 데이터 사용(2019~2021.07)

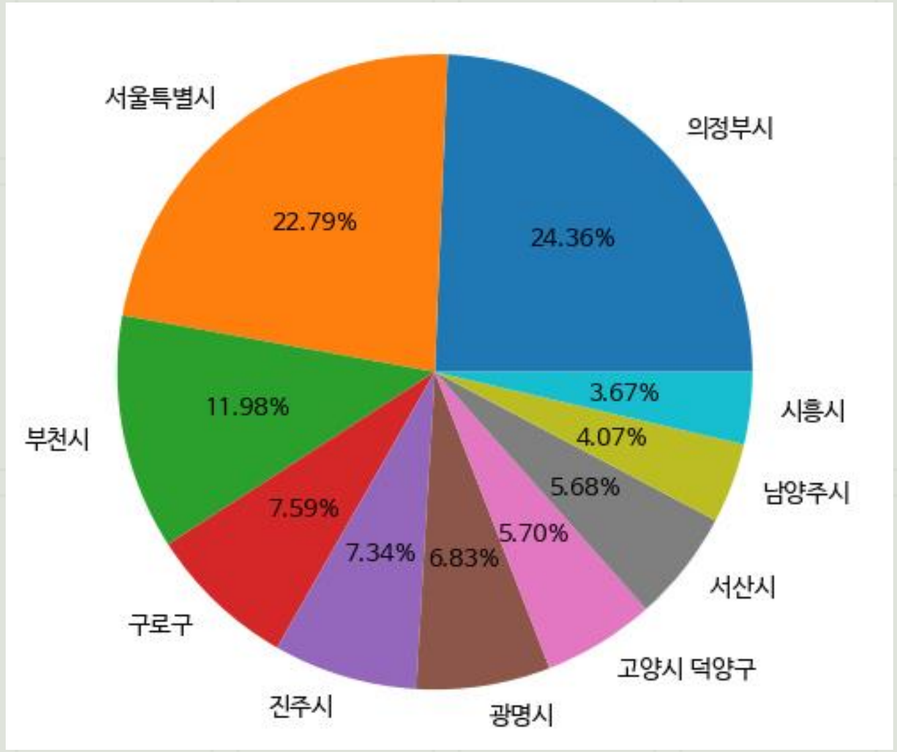
	날짜	시간	업종	시도	구	주문건수
1026041	2021-07-31	23	한식	제주특별자치도	서귀포시	1
1026042	2021-07-31	23	한식	충청남도	천안시 서북구	2
1026043	2021-07-31	23	회	경기도	이천시	2
1026044	2021-07-31	23	회	전라남도	광양시	2
1026045	2021-07-31	23	회	충청남도	천안시 서북구	1

5.7m 행, 6개 열

결측치와 중복 데이터 X

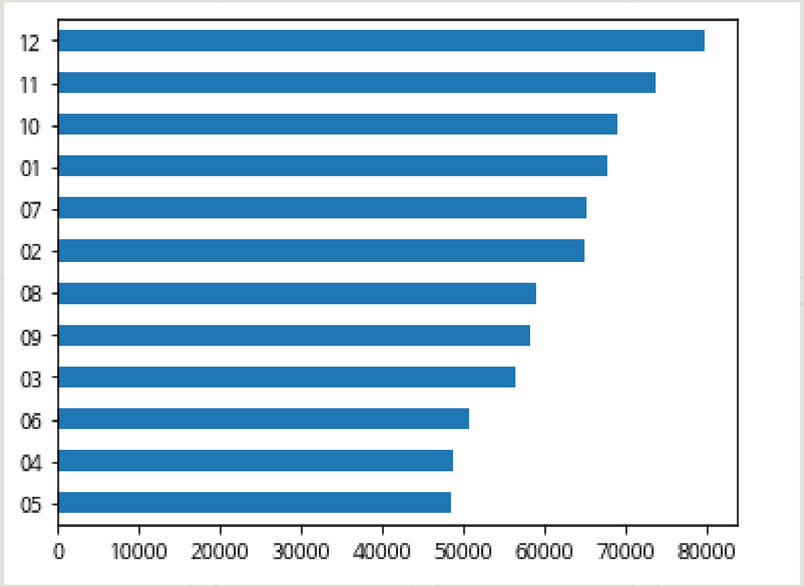
타겟 도시 선정



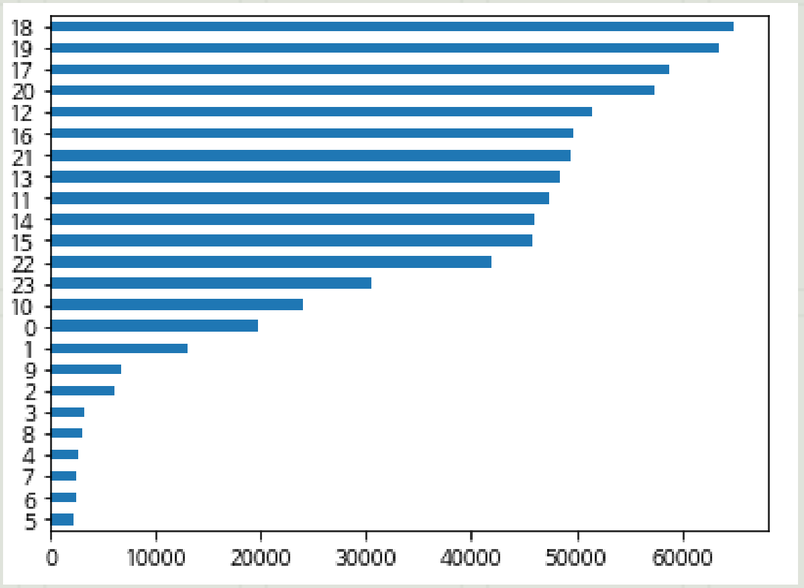


	날짜	시간	업종	시도	구	주문건수
1780	2019-05-19	14	치킨	서울특별시	중구	1
2072	2019-05-21	16	한식	서울특별시	도봉구	1

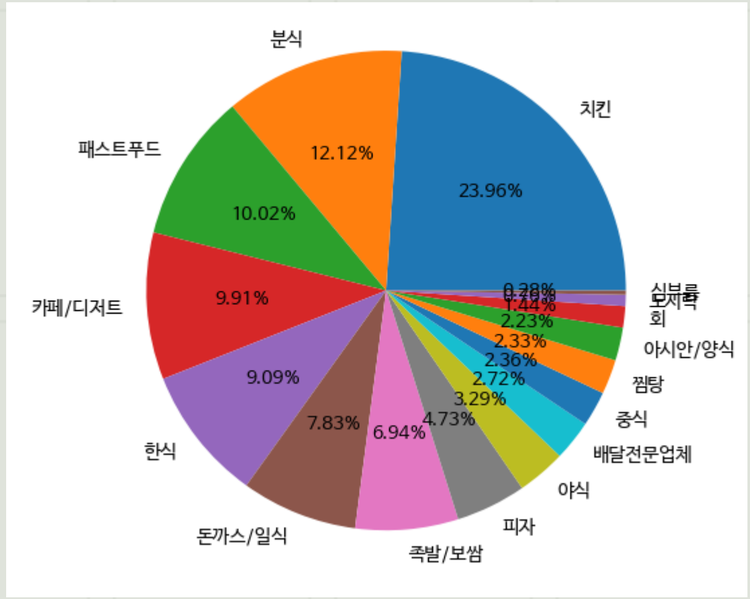
데이터 탐색



월별 데이터수



시간대별 데이터수



업종별 데이터 비율

데이터 전처리

외부에서 추가한 데이터 ▸ 날씨(기상청)
구별 연령별 인구수, 미세먼지(서울시 열린데이터 광장)
축구(네이버)

기존 데이터로 추가한 데이터 ▸ 공휴일, 요일, 주말, **주문 정도(1,2,3)**

0.35	2.0
0.65	4.0
0.85	9.0
0.95	18.0

기존에 있던 주문건수 제거 ▸ **Data Leakage 해소**

	날짜	시간	업종	시도	구	월	time	기온	체감온도	일강수량	...	30대	40대	50대	60대	70대	80대	90대이상	미세	초미세	주문정도
0	2019-05-19	14	치킨	서울특별시	중구	05	2019.05.19.14	21.0	24.1	12.8	...	21749	18632	20811	16464	9514	3846	541	11	8	1
1	2019-05-27	23	치킨	서울특별시	중구	05	2019.05.27.23	15.3	16.5	7.1	...	21749	18632	20811	16464	9514	3846	541	11	5	1

721,437개 행, 30개 열

모델링

초기 모델링(추가 변수 사용 X)

평가지표: Accuracy, AUC

X_train	X_val	X_test
(461719, 4)	(115430, 4)	(144288, 4)

1. 기준 모델(최빈값 사용) ▸

1	0.494538
3	0.321152
2	0.184310

▸ Accuracy 0.4945

2. LogisticRegression ▸ One Hot Encoding ▸ Accuracy 0.5866
AUC 0.71

2차 모델링(모든 변수 사용)

Models
(Using Target Encoder)

- **RandomForest**
  학습 정확도 99.99%, 검증 정확도 68.76%
  학습 AUC 0.9999, 검증 AUC 0.8359
- **XGBoost**
  학습 정확도 72.38%, 검증 정확도 71.80%
  학습 AUC 0.8715, 검증 AUC 0.8602
- **LightGBM**
  학습 정확도 70.67%, 검증 정확도 70.56%
  학습 AUC 0.8495, 검증 AUC 0.8451

Feedbacks

RandomForest 모델은 과적합이 매우 심하다. XGBoost와 LGBM은 하이퍼 파라미터 튜닝 없이도 과적합이 없는 것으로 보인다. 따라서 이 두 가지 모델을 최적화 시켜서 둘 중 더 좋은 성능을 가진 모델을 최종 모델로 사용한다.

3차 모델링(Randomized SearchCV)

XGB Max_depth: 5~9 Learning_rate: 0.025~0.05(0.005 씩) Max_features: 0~1

LGBM Max_depth: 5~8 Learning_rate: 0.1~0.35(0.01 씩) Max_features: 0~1 N_estimators: 150~1000(50씩)

Models

(Using Target Encoder)

XGBoost

학습 정확도 75.30%, 검증 정확도 72.68%
학습 AUC 0.9018, 검증 AUC 0.8716

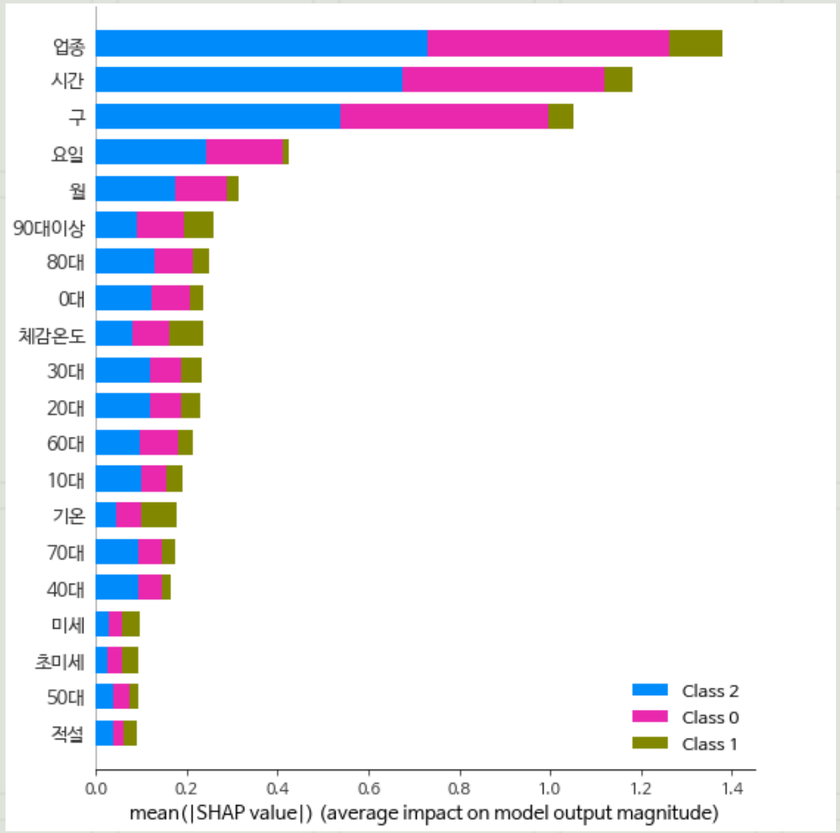
LightGBM

학습 정확도 70.66%, 검증 정확도 70.53%
학습 AUC 0.8492, 검증 AUC 0.8453

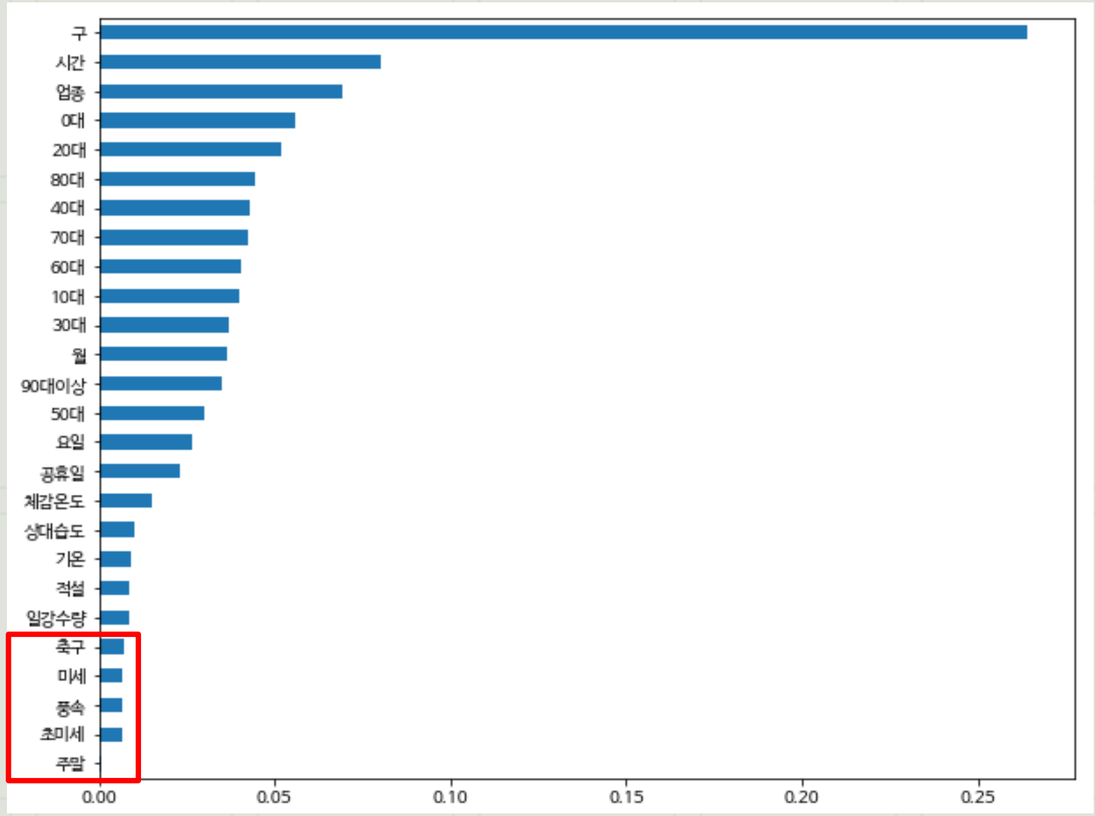
Feedbacks

두 모델을 하이퍼 파라미터 튜닝시킨 결과 XGBoost는 성능은 올라갔지만 과적합이 생겼고, LGBM은 검증 AUC 외에는 성능이 오히려 떨어졌다. 아마도 모든 변수를 다 사용했기 때문에 성능을 떨어뜨리는 변수가 있을 것이고, 모델이 이를 깊게 학습해서 과적합이 생겼을 것 같다.

3차 모델링 결과



XGBoost Shap



XGBoost Feature Importances

4차 모델링(불필요 변수 제거, 수동 하이퍼 파라미터 튜닝)

XGB Max_depth: 6 Learning_rate: 0.4 N_estimators: 120

LGBM Max_depth: 5 Learning_rate: 0.22 N_estimators: 330

Models

(Using **One Hot Encoder**)

XGBoost

학습 정확도 73.07%, 검증 정확도 72.64%
학습 AUC 0.8807, 검증 AUC 0.8708
(시간, 업종, 구, 월, **기온**, 공휴일, 요일, 10대, 20대, 30대, 40대, 60대, 70대, 80대)

LGBM

학습 정확도 73.29%, 검증 정확도 72.71%
학습 AUC 0.8826, 검증 AUC 0.8727
(시간, 업종, 구, 월, **기온**, 공휴일, 요일, 주말, 축구, 0대, 10대, 20대, 30대, 40대, 50대, 60대, 70대, 80대, 90대이상)

Feedbacks

기온을 제외한 모든 날씨 변수를 제거하니까 성능이 올랐다. 날씨가 주문 정도에 영향을 크게 미친다는 가설 1번이 틀렸다는 것을 알 수 있다. 또한, XGBoost에서와 달리 LGBM은 주말, 축구, 모든 나이대의 변수를 다 사용하는 것이 성능이 좋았다. 두 모델 중 더 성능이 좋은 LGBM 모델을 최종 모델로 한다.

LGBM 모델로 테스트 데이터 성능 검증

One Hot Encoder(Transform)

LGBM

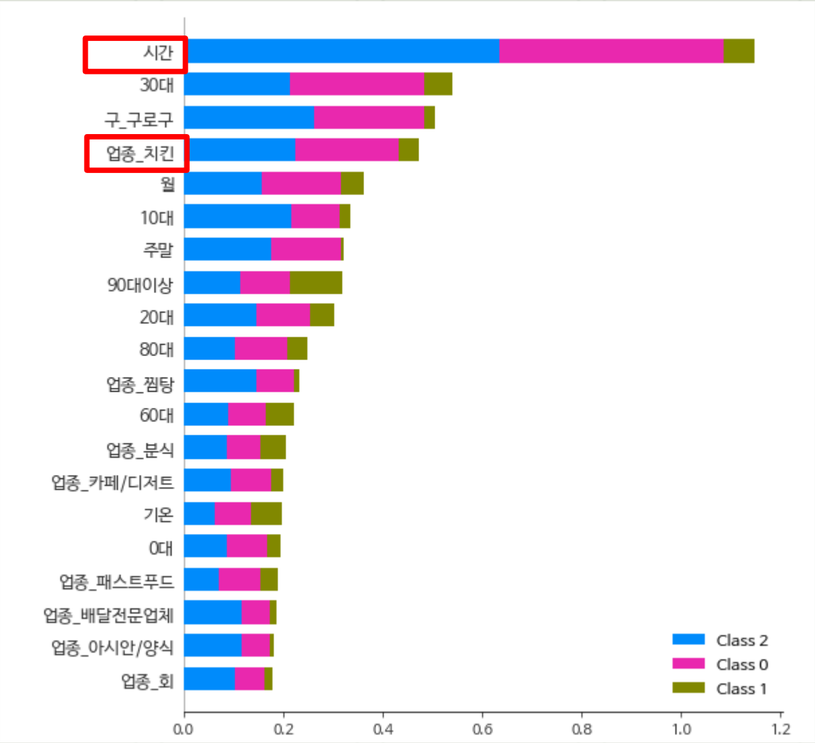
Accracy: 0.7262

AUC: 0.872

04

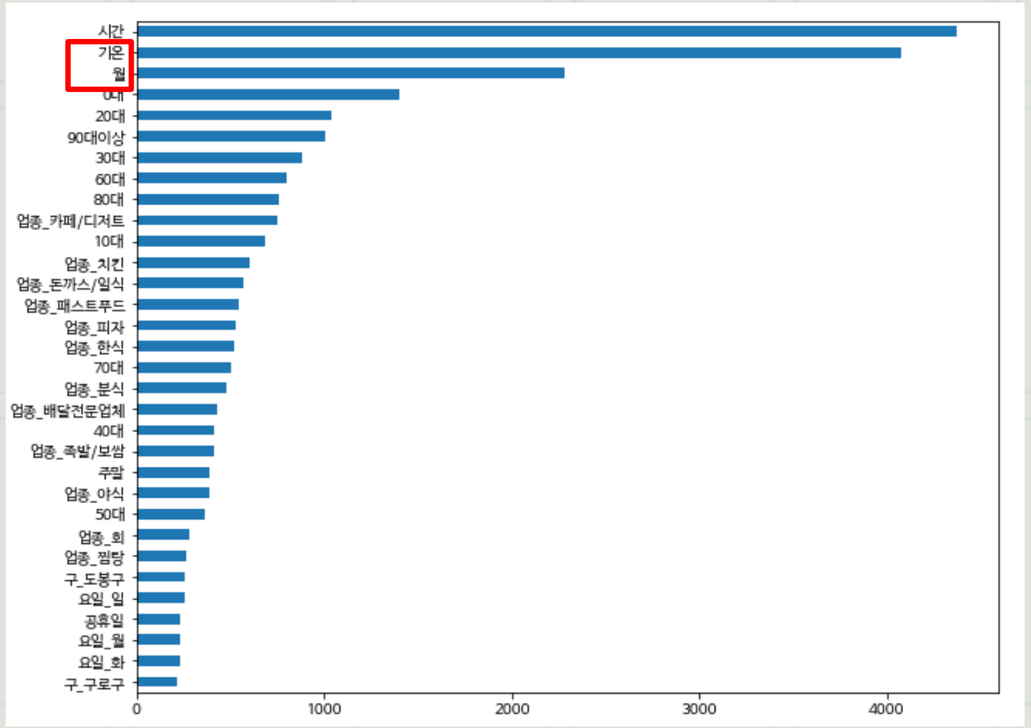
결론

모델 해석



최종 LGBM Sharp

▶ 각 관측치별 타겟에 영향을 미치는 정도의 평균



최종 LGBM Feature Importances

▶ 모델의 성능에 영향을 가장 많이 준 변수들

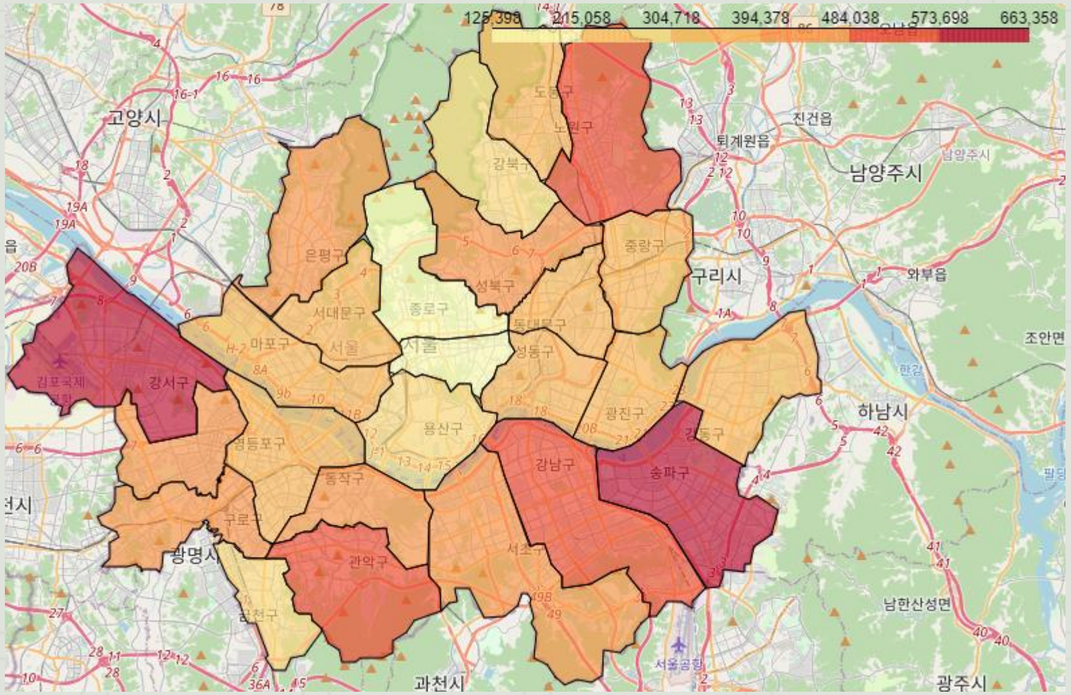
가설 검증

1. 날씨(기온,적설량, 미세먼지 등등) 변수들은 주문정도에 큰 영향을 미칠 것이다.

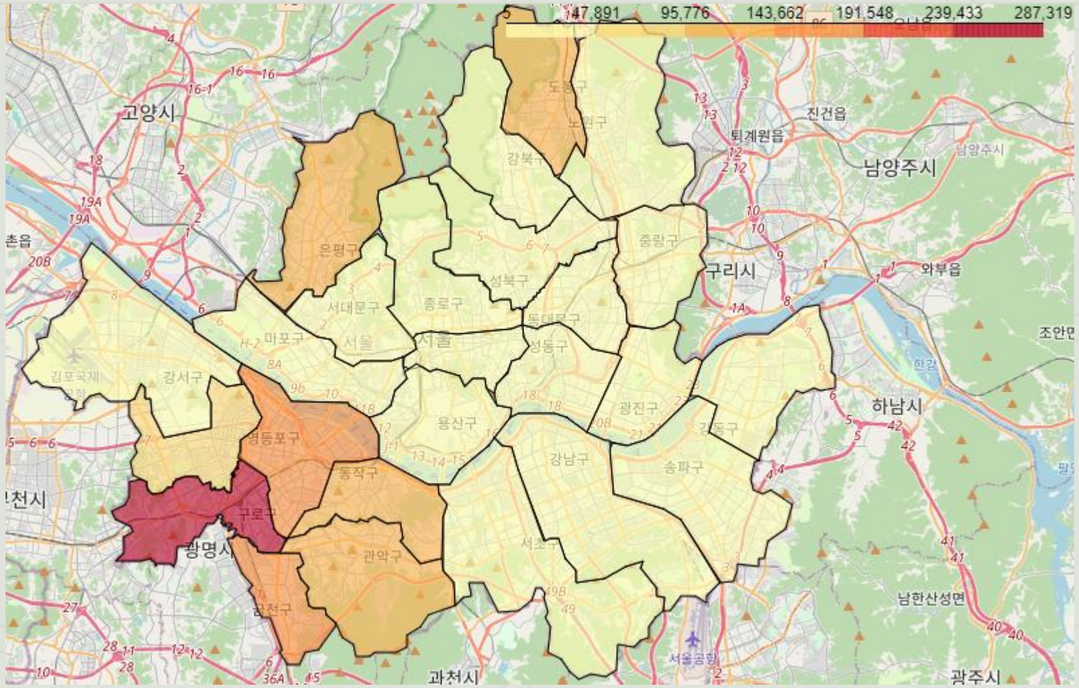
- 기온을 제외한 모든 날씨 변수를 뺐을 때 성능이 더 높았다. 기온을 제외한 날씨변수는 주문정도에 **큰 영향을 미치지 않는다.**

가설 검증

2. 구별 인구수가 더 많은 구가 주문을 더 많이 시켰을 것이다.



구별 인구수 지도

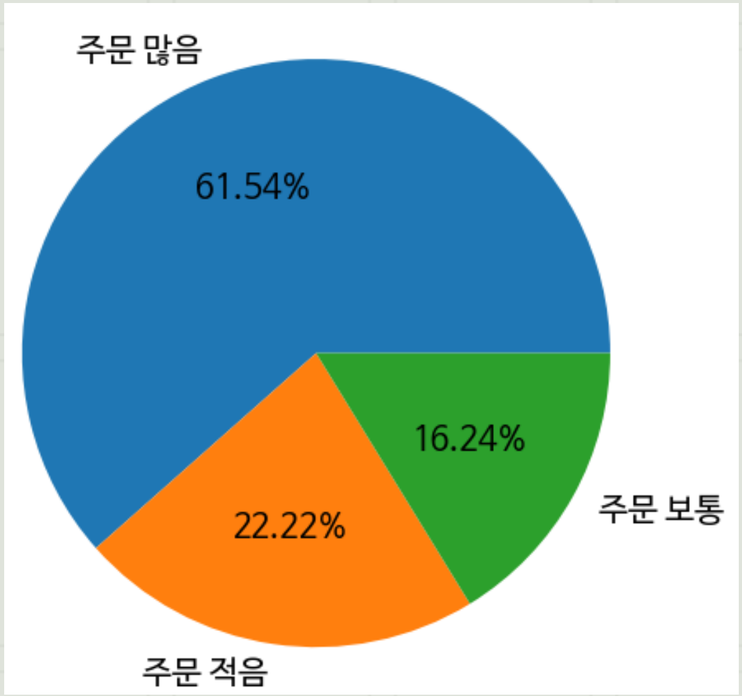


구별 주문량 지도

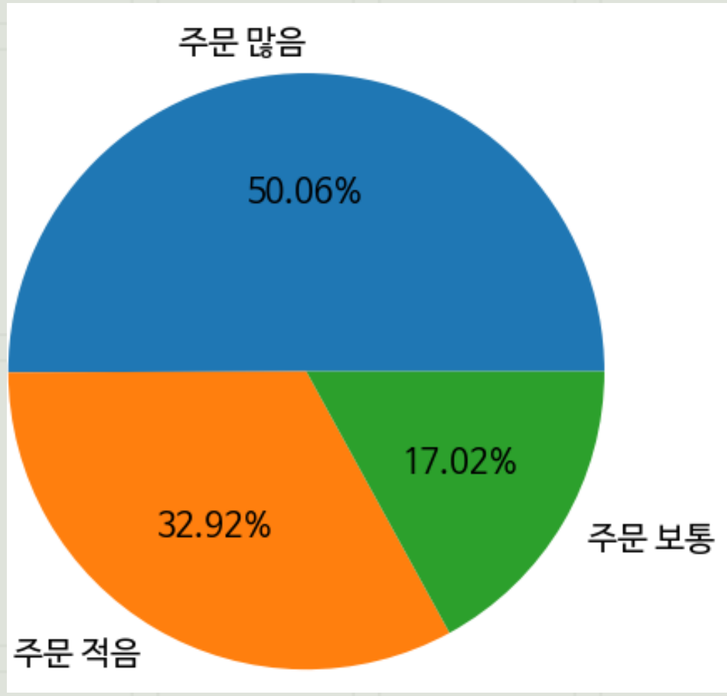
- 인구가 많다고 주문량이 많은 것은 아니다. 인구와 주문량은 큰 상관이 없다.

가설 검증

3. 축구경기가 있는 날이면 사람들이 치킨을 더 많이 시킬 것이다.



축구경기가 있는 날 주문 정도 비율



축구경기가 없는 날 주문 정도 비율

- 축구 경기가 있을 때가 없을 때보다 치킨을 많이 시키는 비율이 약 10퍼센트 정도 더 높다.

한계

1. 데이터가 동별로 되어 있지 않아서 더 세밀한 분석이 어려웠다.

2. 최종 모델에서 사용한 데이터들은 기온 변수를 삭제하면 중복되는 데이터가 많다.
수치형 변수로 쓸 수 있는 변수가 많지 않아서 튜닝이 어려웠다.

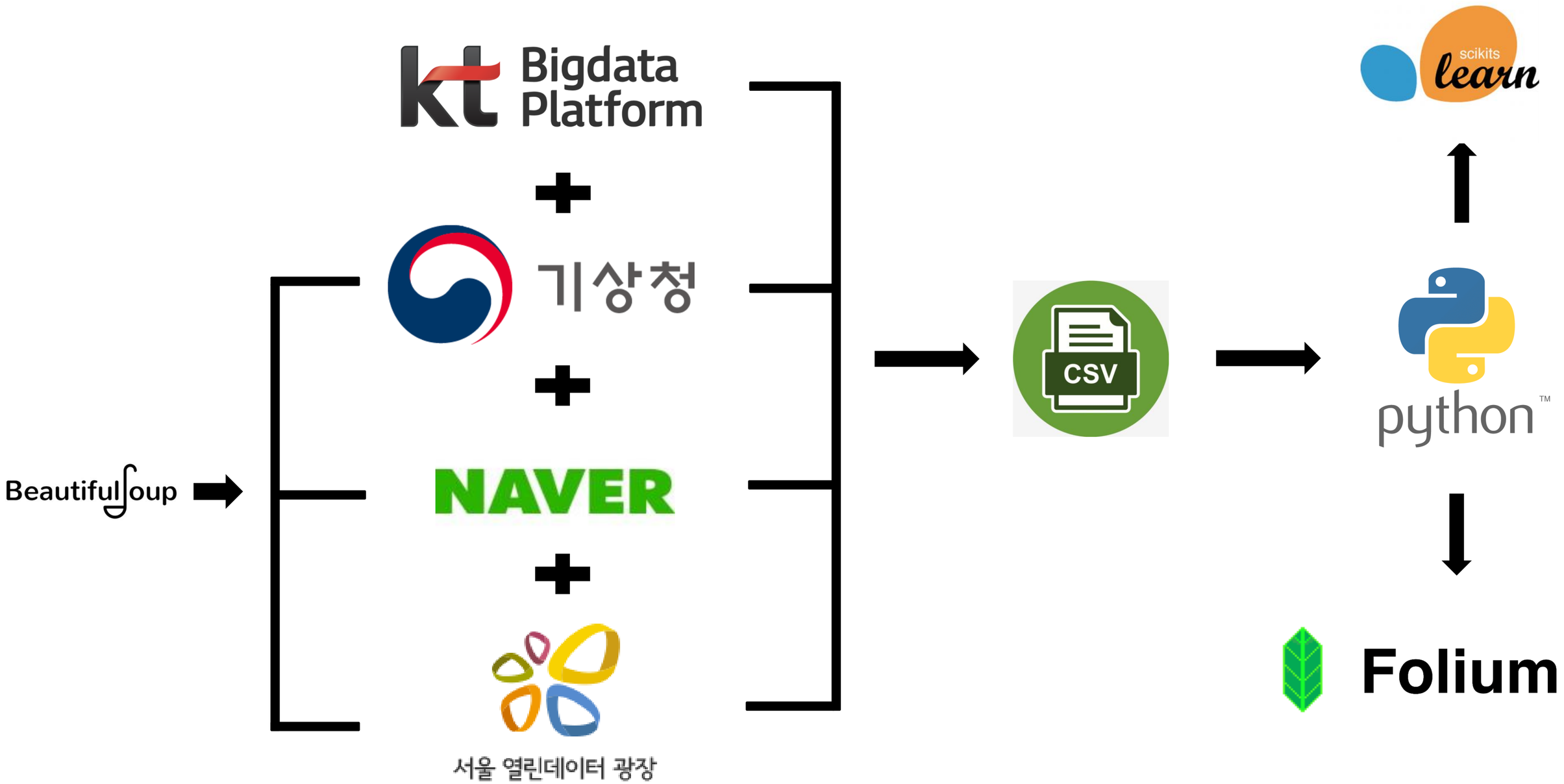
3. Randomized Search CV로 오랜 시간 동안 최적화를 했지만, 오히려 성능이
떨어졌고 결국 수동으로 파라미터를 튜닝해야 했다.

Thank you for Listening!

References

https://www.joongang.co.kr/article/23816720#home

https://www.thinkfood.co.kr/news/articleView.html?idxno=87151


kt Bigdata Platform
+
기상청
+
NAVER
+
서울 열린데이터 광장
Beautiful Soup
CSV
scikits learn
python
Folium


Visual Studio Code


colab


python


BeautifulSoup


scikits
learn


Folium