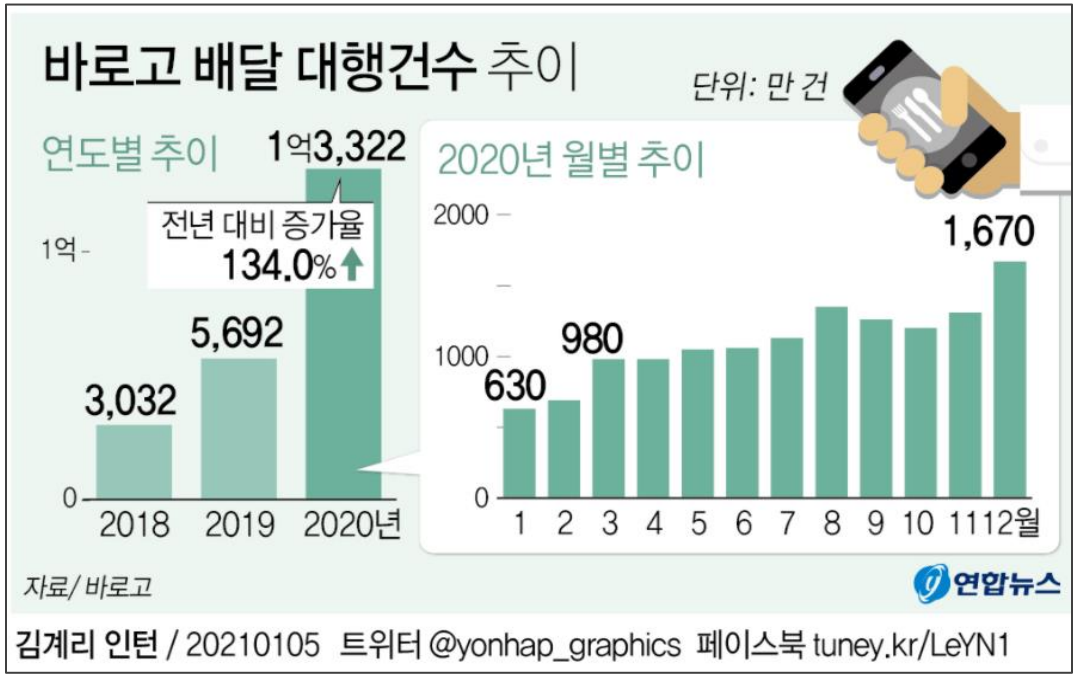
빅데이터 서비스 개발 전공 프로젝트

시도별 배달 주문건수 예측

코딩쟁이들 김동현 김은영 박선익

목차

1_분석 배경



2020년 배달 건수 증가율 전년대비 134%로 폭증!

일부 지역 소비자의 경우 배달이 몰리는 특정 시간대에는 배달 서비스 이용이 제한

일부 업종의 경우 주문이 몰리는 시간대에는 라이더들에게 배달 요청을 거부당하기도…

전국에서 가장 배달 수요가 많은 곳으로 손꼽히는 강남·서초 지역에서 점심시간에 주문하는 소비자는 자신의 위치에서 1㎞ 이상 떨어진 음식점을 이용할 수 없게 됐다.

1_분석 배경

일별, 시간대별 배달 주문건수를 예측할 수 있다면

업주와 소비자 모두 주문 폭증에 대비할 수 있지 않을까?

업주

언제 주문이 몰리는 지 알고
주문량 증가에 미리 대비

소비자

주문량이 몰리는 시간대를 피해 주문

1_분석 배경

주문건수에 영향을 주는 변수는 무엇이 있을까?

1_분석 주제

외부 변화에 따른 시도별 배달 주문건수 예측

주제 1 날씨, 미세먼지, 코로나 확진자수를 고려한 주문건수 최적 예측 모델 선정

주제 2 데이터 분석 및 주문건수 예측 결과를 웹에 리포트 형식으로 구현

2_데이터 명세

출처	데이터이름	제공 형태	요약
KT-빅데이터센터	시간-지역별 배달 주문건수	csv	지역-시간-업종별 주문배달건수 기록 자료 (1957315 × 6) 기간 : 2019.07.17 – 2020.09.30
기상자료개방포털	종관기상관측 (ASOS)	csv	시간대별 기상 요소 관측 자료 (275231 × 15)
에어코리아	미세먼지 데이터	csv	시간대별 미세먼지 관측 자료 (21216 × 9)
공공데이터 서울시 확진자 현황 경기도 감염병관리지원단	코로나 확진자수	csv/API	일별, 지역별 코로나 확진자수 기록 자료 (827 × 3)

2_데이터 전처리_데이터 결측치

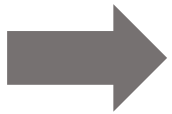
변수 이름	결측치 개수
기온	13
강수량	187,887
풍속	96
일조	84,859
적설	224,881
운량	431
확진자수	110,048

기온 · 풍속 · 운량

시간대별 시간 데이터이므로 선형적인 관계를 고려하여 결측치 대체

`pandas.interpolate()`

기온	강수량	풍속	습도	일조	적설	운량
18.6	0.0	1.1	83.0	NaN	0.0	0.0
18.1	0.0	1.1	83.0	NaN	0.0	0.0
17.9	0.0	NaN	82.0	0.1	0.0	0.0
20.8	0.0	0.6	64.0	1.0	0.0	2.0
23.5	0.0	1.8	57.0	1.0	0.0	3.0
24.7	0.0	NaN	46.0	1.0	0.0	0.0
26.3	0.0	3.2	40.0	1.0	0.0	0.0
27.3	0.0	3.1	31.0	1.0	0.0	0.0
27.9	0.0	3.0	31.0	1.0	0.0	0.0
28.5	0.0	2.7	34.0	1.0	0.0	0.0



기온	강수량	풍속	습도	일조	적설	운량
18.6	0.0	1.10	83.0	0.083333	0.0	0.0
18.1	0.0	1.10	83.0	0.091667	0.0	0.0
17.9	0.0	0.85	82.0	0.100000	0.0	0.0
20.8	0.0	0.60	64.0	1.000000	0.0	2.0
23.5	0.0	1.80	57.0	1.000000	0.0	3.0
24.7	0.0	2.50	46.0	1.000000	0.0	0.0
26.3	0.0	3.20	40.0	1.000000	0.0	0.0
27.3	0.0	3.10	31.0	1.000000	0.0	0.0
27.9	0.0	3.00	31.0	1.000000	0.0	0.0
28.5	0.0	2.70	34.0	1.000000	0.0	0.0

2_데이터 전처리_데이터 결측치

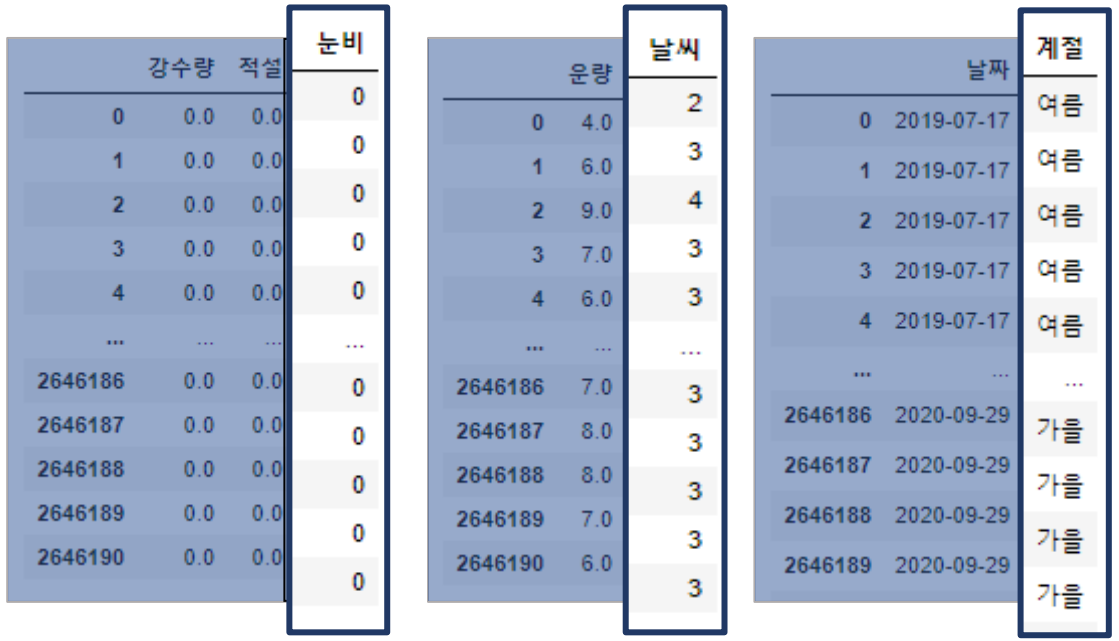
변수 이름	결측치 개수
기온	13
강수량	187,887
풍속	96
일조	84,859
적설	224,881
운량	431
확진자수	110,048

강수량 · 적설
NULL 값은 0으로 대체

일조
결측치 값이 너무 많아, 사용하지 않기로 결정

확진자수
결측치 값을 0으로 대체(확진자수가 없는 날)

2_데이터 전처리_파생변수 생성



	강수량	적설	눈비
0	0.0	0.0	0
1	0.0	0.0	0
2	0.0	0.0	0
3	0.0	0.0	0
4	0.0	0.0	0
...	...	...	...
2646186	0.0	0.0	0
2646187	0.0	0.0	0
2646188	0.0	0.0	0
2646189	0.0	0.0	0
2646190	0.0	0.0	0

	운량	날씨
0	4.0	2
1	6.0	3
2	9.0	4
3	7.0	3
4	6.0	3
...	...	...
2646186	7.0	3
2646187	8.0	3
2646188	8.0	3
2646189	7.0	3
2646190	6.0	3

	날짜	계절
0	2019-07-17	여름
1	2019-07-17	여름
2	2019-07-17	여름
3	2019-07-17	여름
4	2019-07-17	여름
...	...	...
2646186	2020-09-29	가을
2646187	2020-09-29	가을
2646188	2020-09-29	가을
2646189	2020-09-29	가을

눈비

강수량 적설 결합 ➔ 값의 존재 유무에 따라
1, 0에 해당하는 범주형 변수 생성

날씨

운량 값을 기준으로 순서형 변수(1, 2, 3, 4) 생성

운량값	0 – 1	3 – 5	6 – 8	9 – 10
날씨	맑음	구름조금	구름많이	흐림
변수	1	2	3	4

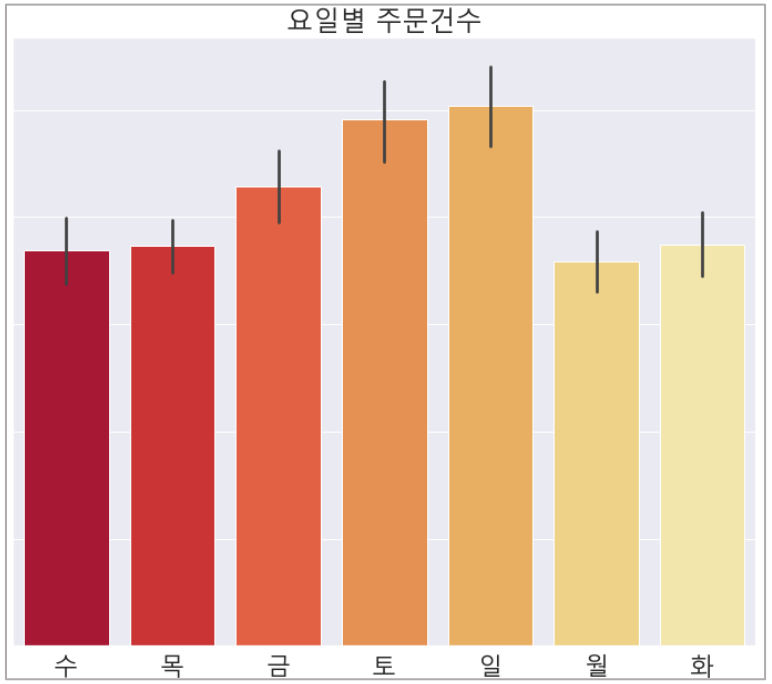
※ 기상청 하늘상태 표현을 참고하여 세분화하였음

계절

날짜	3 – 5월	6 – 8월	9 – 11월	12 – 2월
계절	봄	여름	가을	겨울

※ 한국민족대백과사전 기준으로 세분화하였음

2_데이터 전처리_파생변수 생성



	날짜	요일
0	2019-07-17	수
1	2019-07-17	수
2	2019-07-17	수
3	2019-07-17	수
4	2019-07-17	수
...	...	...
2646186	2020-09-29	화
2646187	2020-09-29	화
2646188	2020-09-29	화
2646189	2020-09-29	화
2646190	2020-09-30	수

요일

특정 요일(금, 토, 일)에 주문건수가 집중되는 경향
이를 반영할 수 있는 각 날짜에 맞는 요일 변수 생성

2_데이터 전처리_파생변수 생성

2020년 공휴일 (총 67일) | 2020년 달라지는 제도 | 2020년 달력보기

명칭	날짜	요일	명칭	날짜	요일
신정	1월 1일	수	현충일	6월 6일	토
설날 (연휴) 대체공휴일	1월 24일 ~ 26일 1월 27일	금 ~ 일 월	광복절	8월 15일	토
삼일절	3월 1일	일	추석 (연휴)	9월 30일 ~ 10월 2일	수 ~ 금
제21대 국회의원선거	4월 15일	수	개천절	10월 3일	토
부처님오신날	4월 30일	목	한글날	10월 9일	금
어린이날	5월 5일	화	크리스마스	12월 25일	금

기념일	날짜
초복, 중복, 말복	2019년 7월 22일, … 2020년 8월 15일
스포츠 국제 경기 (월드컵 / 올림픽)	2019년 10월 10일, … 2020년 1월 26일
블랙데이	2020년 4월 14일

	날짜	요일	기념일	공휴일
0	2019-07-17	수	0	0
1	2019-07-17	수	0	0
2	2019-07-17	수	0	0
3	2019-07-17	수	0	0
4	2019-07-17	수	0	0
...	...	...	...	...
2	2020-09-29	화	0	2
3	2020-09-29	화	0	2
4	2020-09-29	화	0	2
5	2020-09-29	화	0	2
6	2020-09-30	수	0	1

공휴일

네이버 검색으로 나오는 연도별 공휴일 기준 변수 설정

공휴일 전날	공휴일	나머지
2	1	0

기념일

초복, 중복, 말복, 블랙데이, 스포츠 국제 경기 등이
배달 주문건수가 많은 날임을 확인

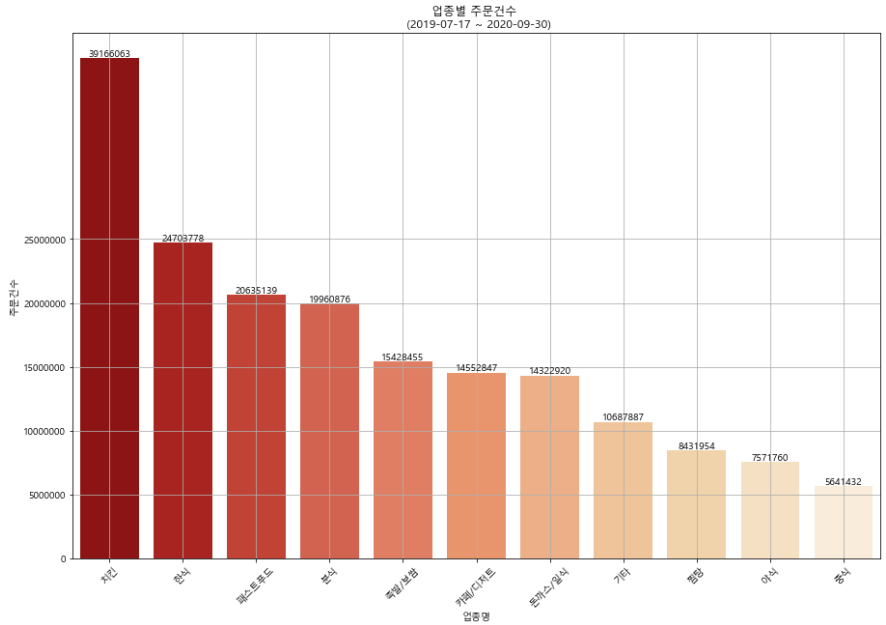
기념일	나머지
1	0

2_데이터 전처리_제거변수

적설

눈 오는 날이 적음 ➔ 0 값이 많음 ➔ 강수량 컬럼과 병합하여 ‘눈비’ 파생변수 생성 후 제거

2_데이터 전처리_업종명




```
final_data['업종명'].unique()
```

executed in 32ms, finished 13:22:09 2021-04-22

```
array(['도시락', '돈까스/일식', '배달전문업체', '분식', '심부름', '아시안/양식', '야식', '족발/보쌈',
       '찜탕', '치킨', '패스트푸드', '피자', '한식', '회', '카페/디저트', '중식'], dtype=object)
```

```
final_data['업종명'].unique()
```

executed in 160ms, finished 13:21:38 2021-04-22

```
array(['기타', '돈까스/일식', '분식', '야식', '족발/보쌈', '중식', '찜탕', '치킨', '카페/디저트',
       '패스트푸드', '한식'], dtype=object)
```

심부름

배달음식과 무관한 업종이므로 포함된 Record 모두 삭제

도시락 · 배달전문업체 · 아시안/양식

주문건수가 적고 업종 구분 불분명 ➔ 기타로 편입

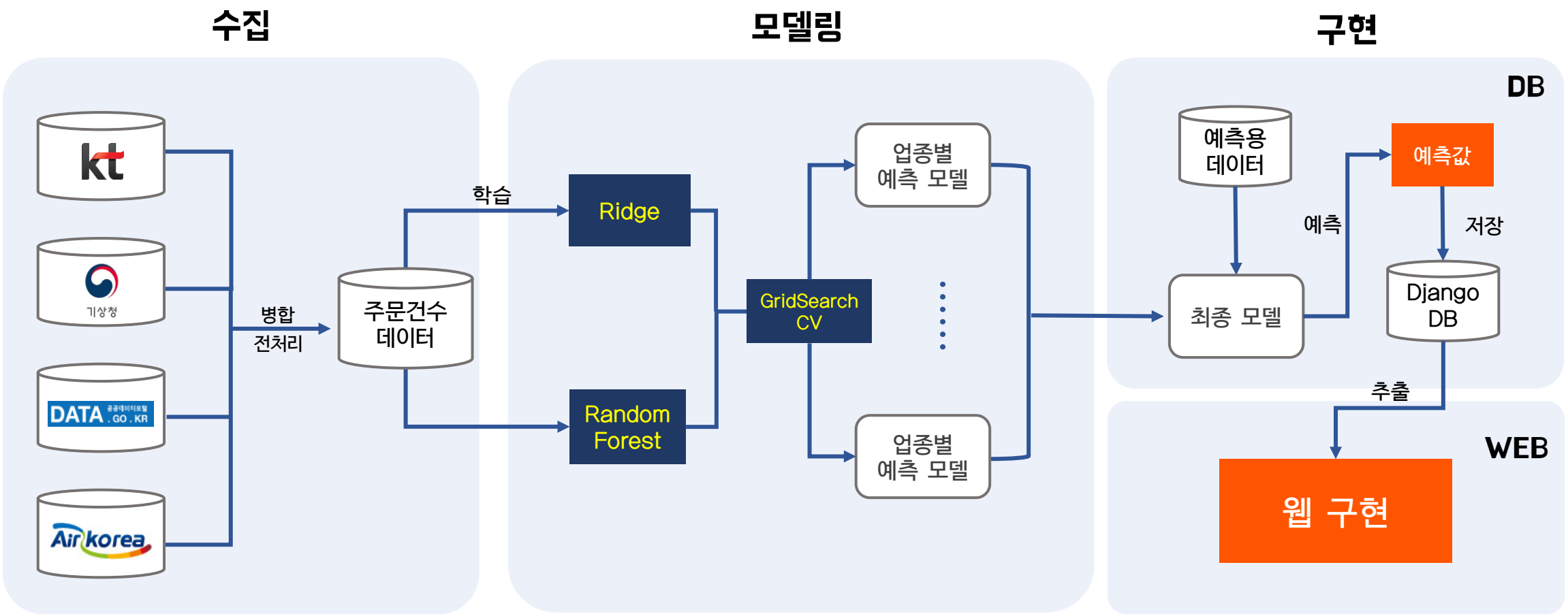
피자

동일 범주인 패스트푸드로 편입

회

동일 범주인 돈까스/일식으로 편입

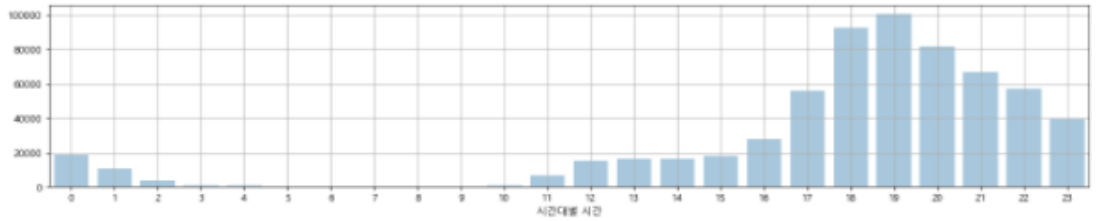
2_Workflow


수집
모델링
구현
kt
기상청
DATA .GO.KR
Airkorea
병합
전처리
주문건수
데이터
학습
Ridge
Random
Forest
GridSearch
CV
업종별
예측 모델
업종별
예측 모델
DB
예측용
데이터
예측값
예측
저장
최종 모델
Django
DB
추출
WEB
웹 구현

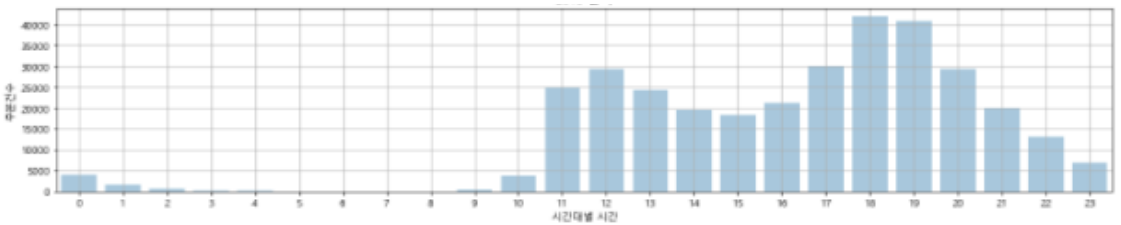
2_Modeling

업종 및 시간대별 주문건수

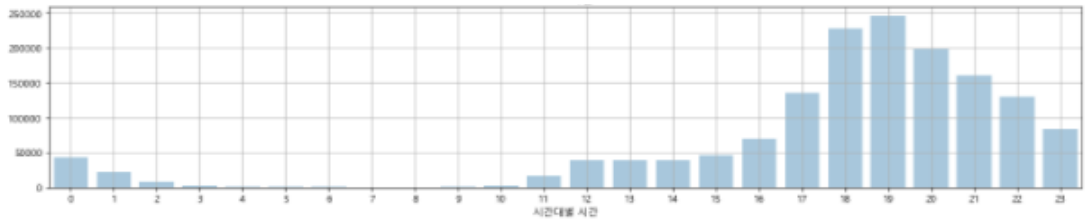
2019 치킨



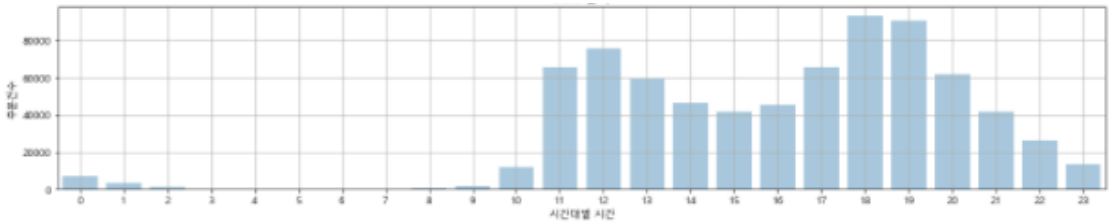
2019 분식



2020 치킨



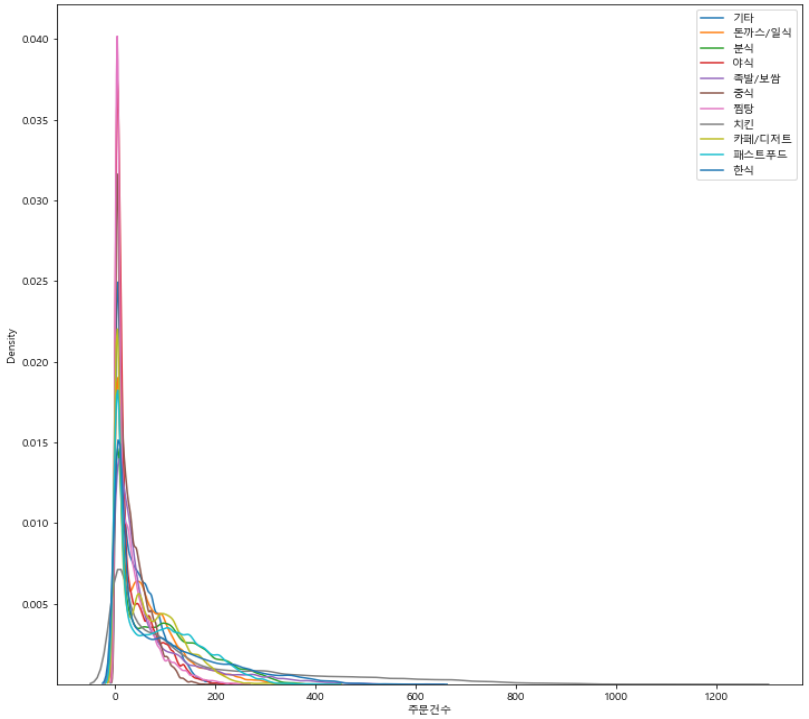
2020 분식

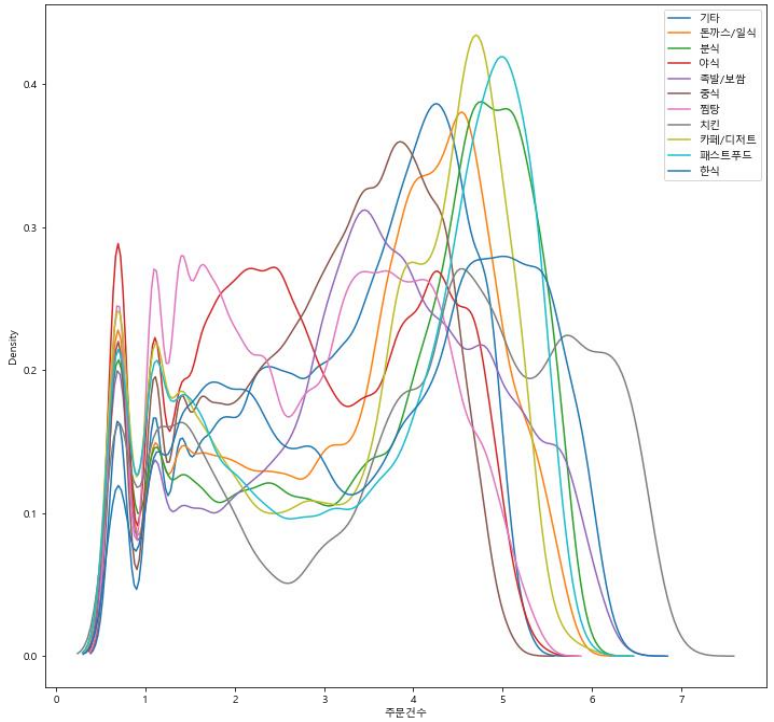


치킨과 같이 저녁에 자주 먹는 음식은 저녁 시간에 몰리는 경향

분식, 한식 등 점심이나 저녁을 가리지 않는 음식은 점심, 저녁시간에 분포

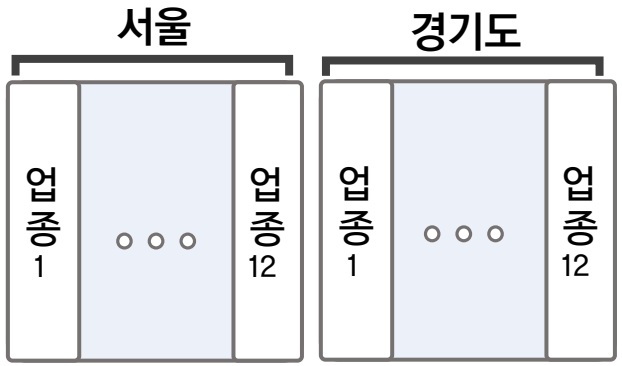
2_Modeling





업종별 주문건수 분포의 정규성을 위해 주문건수 값에 log1p 적용

2_Modeling_모델 선택



1) 지역별/업종별 데이터로 분리

Data Set
Train 8 : Test 2

2) Train 8 : Test 2 으로 데이터 분할

Ridge Regression

Linear Regression

RandomForestRegressor

GradientBoostingRegressor

3) 회귀모형에 적합

```
==========치킨==========
업종명 : 치킨    model: LinearRegression         Fold : 3       cross_val_score : 0.924
업종명 : 치킨    model: LinearRegression         Fold : 5       cross_val_score : 0.925
업종명 : 치킨    model: LinearRegression         Fold : 7       cross_val_score : 0.924
업종명 : 치킨    model: LinearRegression         Fold : 9       cross_val_score : 0.924
업종명 : 치킨    model: Ridge    Fold : 3        cross_val_score : 0.924
업종명 : 치킨    model: Ridge    Fold : 5        cross_val_score : 0.924
업종명 : 치킨    model: Ridge    Fold : 7        cross_val_score : 0.924
업종명 : 치킨    model: Ridge    Fold : 9        cross_val_score : 0.925
업종명 : 치킨    model: RandomForestRegressor    Fold : 3       cross_val_score : 0.942
업종명 : 치킨    model: RandomForestRegressor    Fold : 5       cross_val_score : 0.943
업종명 : 치킨    model: RandomForestRegressor    Fold : 7       cross_val_score : 0.942
업종명 : 치킨    model: RandomForestRegressor    Fold : 9       cross_val_score : 0.943
업종명 : 치킨    model: GradientBoostingRegressor        Fold : 3       cross_val_score : 0.903
업종명 : 치킨    model: GradientBoostingRegressor        Fold : 5       cross_val_score : 0.903
업종명 : 치킨    model: GradientBoostingRegressor        Fold : 7       cross_val_score : 0.903
업종명 : 치킨    model: GradientBoostingRegressor        Fold : 9       cross_val_score : 0.904
```

4) 모델 평가와 예측력 확인

모델명	스코어
RandomForestRegressor	0.876
LinearRegression	0.832
Ridge	0.832
GradientBoostingRegressor	0.828

Linear 보다 계수가 안정적

5) 각 모델별 Score 합계 산출하여 스코어 상위 2개 모델을 선택

Ridge, RandomForestRegressor 모델 선택

2_Modeling_GridSearch_Fold값 확인

	광역시도명	업종명	모델명	fold 수	스코어
4	서울	기타	Ridge	3	0.663
5	서울	기타	Ridge	5	0.663
6	서울	기타	Ridge	7	0.663
7	서울	기타	Ridge	9	0.662
8	서울	기타	RandomForestRegressor	3	0.744
...	...	...	...	...	...
343	경기도	한식	Ridge	9	0.932
344	경기도	한식	RandomForestRegressor	3	0.952
345	경기도	한식	RandomForestRegressor	5	0.953
346	경기도	한식	RandomForestRegressor	7	0.953
347	경기도	한식	RandomForestRegressor	9	0.954

1) 지역 / 업종별 Ridge & RF Model Fold 값, Score 값 추출

	fold 수	스코어
0	3	0.852591
1	5	0.854182
2	7	0.854591
3	9	0.854977

2) 전체 Ridge & RF Model Fold 별 Score 값의 평균

Fold 수가 9인 경우 Model의 성능이 좋음

2_Modeling_GridSearch_Best Parameter 값 확인

$$\hat{\beta}_{Ridge} = (X^T X + \lambda I)^{-1} X^T y$$

$$R^2 = 1 - \frac{Unexplained\ Variation}{Total\ Variation}$$

$$RMSE(\theta_1, \theta_2) = \sqrt{\frac{\sum_{i=1}^{n}(x_{1,i} - x_{2,i})^2}{n}}$$

Ridge

지역 / 업종별 Best Parameter 추출 결과 및 RMSE, R^2 score

서울						경기					
업종명	Best Parameter		평가지표			업종명	Best Parameter		평가지표		
	Alpha	Normalize	Train R^2 score	Test RMSE	Test R^2 score		Alpha	Normalize	Train R^2 score	Test RMSE	Test R^2 score
치킨	0.1	True	0.667	0.454	0.671	치킨	0.01	False	0.862	0.478	0.848
분식	0.01	False	0.685	0.461	0.65	분식	0.01	True	0.926	0.38	0.927
패스트푸드	0.1	True	0.890	0.35	0.875	패스트푸드	0.01	False	0.941	0.397	0.934
한식	0.1	True	0.819	0.411	0.807	한식	0.01	True	0.909	0.393	0.906
카페/디저트	10	False	0.828	0.431	0.83	카페/디저트	0.1	True	0.928	0.391	0.925
족발/보쌈	0.1	True	0.568	0.519	0.555	족발/보쌈	0.01	True	0.722	0.599	0.723
야식	0.1	True	0.568	0.478	0.548	야식	0.01	True	0.891	0.44	0.877
돈까스/일식	1	True	0.926	0.366	0.924	돈까스/일식	0.1	True	0.953	0.391	0.954
찜탕	0.1	True	0.779	0.454	0.795	찜탕	0.01	True	0.919	0.438	0.918
중식	0.1	True	0.894	0.358	0.897	중식	1	True	0.951	0.355	0.951
기타	0.1	True	0.820	0.441	0.814	기타	0.1	True	0.933	0.415	0.93

2_Modeling_GridSearch_Best Parameter 값 확인

Random Forest Regressor
지역 / 업종별 Best Parameter 추출 결과

서울				경기			
업종명	Best Parameter			업종명	Best Parameter		
	Max depth	Min samples split	N estimators		Max depth	Min samples split	N estimators
치킨	30	8	150	치킨	30	8	150
분식	30	8	150	분식	30	8	50
패스트 푸드	30	20	100	패스트 푸드	30	8	150
한식	30	8	150	한식	30	8	150
카페/디저트	30	8	150	카페/디저트	30	8	150
족발/보쌈	30	8	150	족발/보쌈	30	8	50
야식	20	20	150	야식	30	20	150
돈까스/일식	30	8	150	돈까스/일식	30	8	150
찜탕	30	8	150	찜탕	30	8	100
중식	30	8	150	중식	8	8	100
기타	30	8	100	기타	30	8	100

※ random_state = 156 고정

2_Modeling_GridSearch_Best Parameter 값 확인

Random Forest Regressor

지역 / 업종별 RMSE, R^2 score

서울				경기도			
업종명	평가지표			업종명	평가지표		
	Train R^2 score	Test RMSE	Test R^2 score		Train R^2 score	Test RMSE	Test R^2 score
기타	0.837	0.404	0.740	기타	0.929	0.425	0.880
돈까스/일식	0.838	0.400	0.736	돈까스/일식	0.952	0.405	0.917
분식	0.924	0.386	0.848	분식	0.965	0.354	0.947
야식	0.891	0.385	0.831	야식	0.953	0.355	0.923
족발/보쌈	0.914	0.374	0.872	족발/보쌈	0.967	0.325	0.948
중식	0.798	0.435	0.688	중식	0.926	0.415	0.867
찜탕	0.747	0.455	0.592	찜탕	0.941	0.388	0.905
치킨	0.963	0.314	0.944	치킨	0.979	0.323	0.969
카페/디저트	0.875	0.447	0.801	카페/디저트	0.945	0.469	0.906
패스트푸드	0.94	0.334	0.910	패스트푸드	0.968	0.393	0.939
한식	0.907	0.392	0.854	한식	0.963	0.366	0.946

2_Modeling_최종모델선정결과

모델 비교

모델명	Ridge	Random Forest Regressor
Train R^2	0.835	0.915
Test RMSE	0.427	0.388
Test R^2	0.830	0.862
학습시간(초)	0.050	7.629

※ 학습시간은 치킨(업종)을 기준으로 산출하였음

Train Score의 경우 RF가 우수한 성능

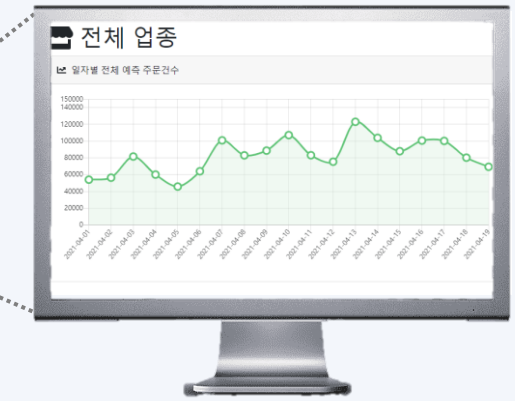
웹 구현까지 고려해봤을 때, 준수한 성능과 빠른 시간을 보이는

Ridge 모델을 적용하는 것이 현실적으로 타당

3_활용방안

API로 수집된
기상/미세먼지/코로나

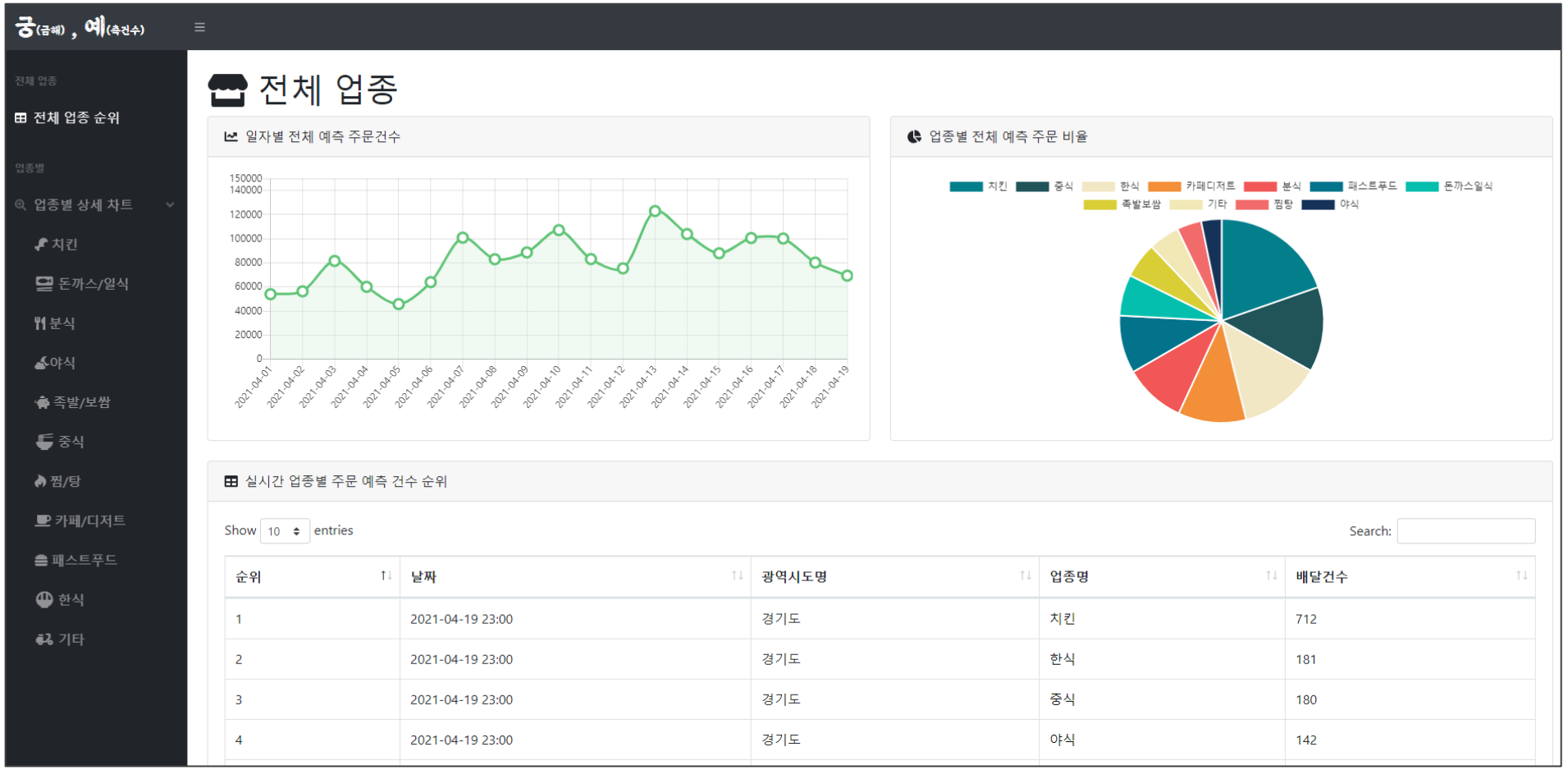
배달음식 주문건수
예측모델



궁(금해), 예(측건수)

최신 데이터를 반영한
각 시도 업종 시간대별
예상 배달주문 건수 정보 제공

3_서비스 시연


궁(금해) , 예(측건수)
전체 업종
전체 업종 순위
업종별
업종별 상세 차트
치킨
돈까스/일식
분식
야식
족발/보쌈
중식
찜/탕
카페/디저트
패스트푸드
한식
기타
전체 업종
일자별 전체 예측 주문건수
150000
140000
120000
100000
80000
60000
40000
20000
0
2021-04-01
2021-04-02
2021-04-03
2021-04-04
2021-04-05
2021-04-06
2021-04-07
2021-04-08
2021-04-09
2021-04-10
2021-04-11
2021-04-12
2021-04-13
2021-04-14
2021-04-15
2021-04-16
2021-04-17
2021-04-18
2021-04-19
업종별 전체 예측 주문 비율
치킨
중식
한식
카페디저트
분식
패스트푸드
돈까스일식
족발보쌈
기타
찜탕
야식
실시간 업종별 주문 예측 건수 순위
Show 10 entries
Search:
순위
날짜
광역시도명
업종명
배달건수
1
2021-04-19 23:00
경기도
치킨
712
2
2021-04-19 23:00
경기도
한식
181
3
2021-04-19 23:00
경기도
중식
180
4
2021-04-19 23:00
경기도
야식
142

3_기대효과

업주

- 주문량 예측으로 재료 낭비 감소
- 날씨에 따른 주문량 고려할 수 있어 보다 정확한 재고 관리 가능
- 주문 폭증에 대비 가능
- 추가 데이터(시간대별 평균 소요시간)를 통한 추가 확장 가능

고객

- 주문 폭증 시간대 예측으로 인한 대비 가능
- 배달 지연으로 인한 문제 발생 최소화

3_한계점

데이터 편차의 문제

특정 배달 대행 업체의 배달 데이터
→ 점유율에 따른 지역구별 데이터 편차 발생

데이터 특성상 경기도권에 양이 집중된 경향

모델 관련 한계

신경망 모형과 같은 딥러닝 모델 성능 미확인

4_참고문헌

이태수, “주문 급증에 라이더 부족…"점심땐 1㎞ 이내만 배달합니다"”, 연합뉴스, , 2020-11-29, https://www.yna.co.kr/view/AKR20201127060500030, 2021-04-10

이계리, “[그래픽] 바로고 배달 대행건수 추이”, 연합뉴스, 2021-01-05, https://www.yna.co.kr/view/GYH20210105000500044, 2021-04-10

권재영, 김시내, 박은지, 송종우, “국내 배달음식 이용건수 분석 및 예측”, vol.28, no.5, 2015, pp. 977-990 (14 pages)

김대룡, 김다영, 변수지, “날씨에 따른 배달음식 주문건수 예측”, 2016.10, 480-481(2 pages)