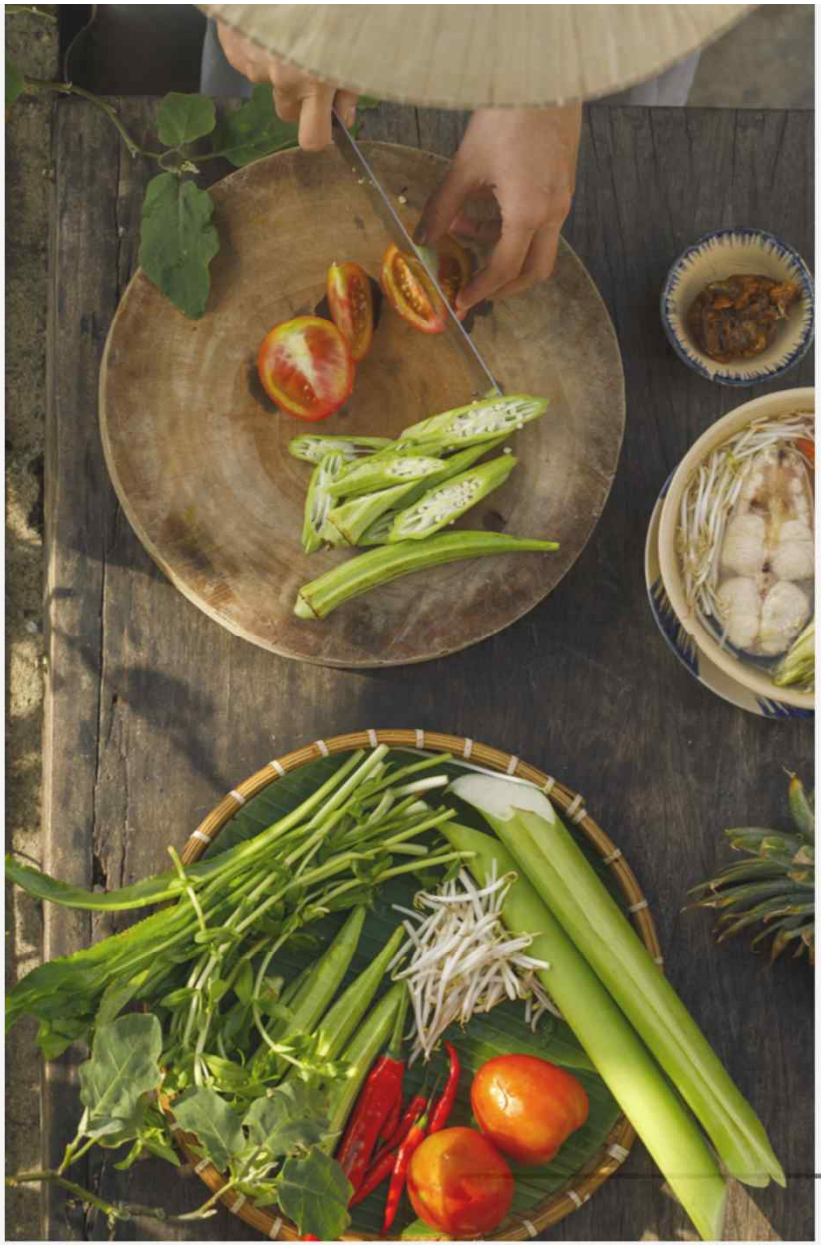
# 식사 배달 업체 원자재 수요 예측 모델

## Codestates AI Bootcamp 2nd Project: Machine Learning



# 목차

## 01

### 문제 정의, 데이터 소개

해결할 문제를 정의하고
모델링에 사용할 데이터를
소개합니다.

## 02

### 가설과 평가지표,
### 데이터 전처리 설명

가설과 모델 평가지표를 선정하고
데이터 전처리 방식을 설명합니다.

## 03

### 머신러닝 적용과 검증,
### 모델 해석

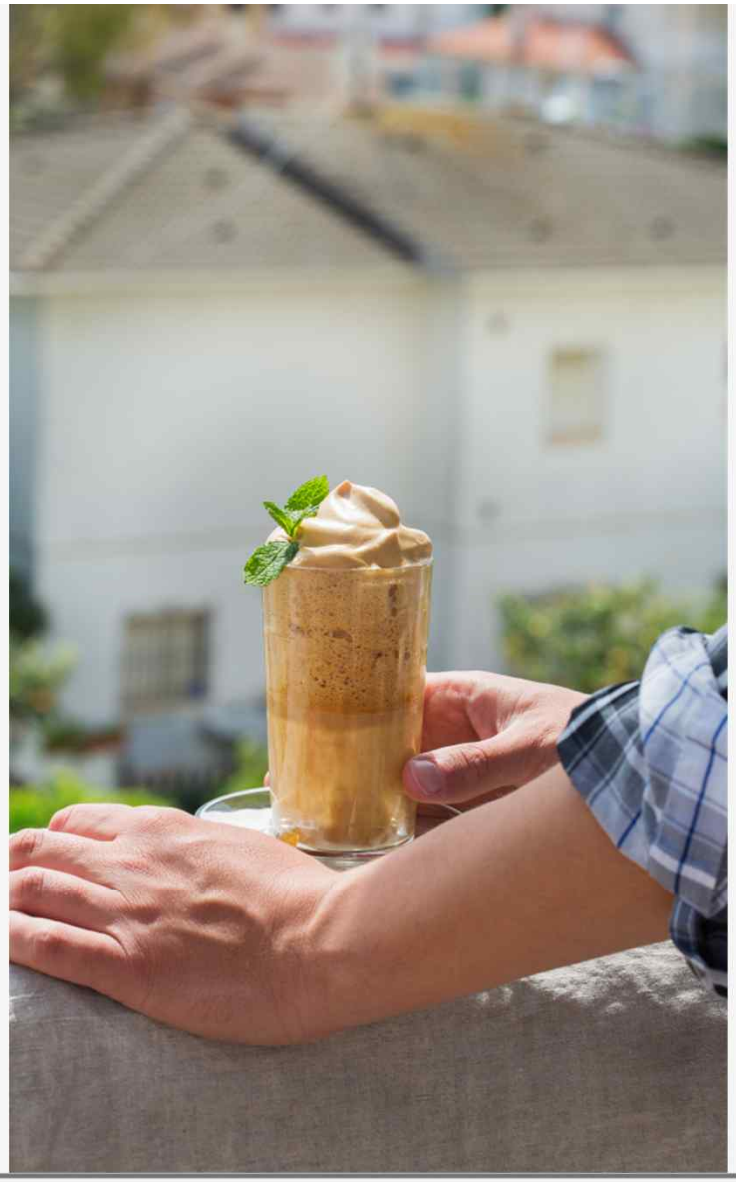
모델을 적용하고, 검증한 뒤
결과를 해석합니다.

## 04

### 한계와 보완 가능성

데이터 및 모델의 한계와
개선 방안을 논의합니다.

03



# 문제 정의와 질문

*"과거 145주간의 데이터를 통해서*
*다음 10주의 배달식 수요를*
*예측할 수 있을까?"*

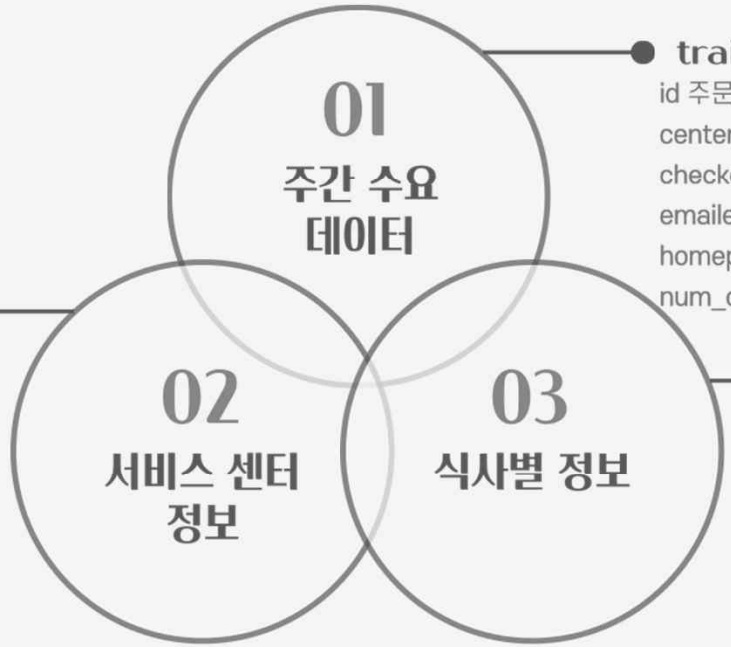
- 목표(target)는 무엇인가?
  어떻게 구할까?

- 어떤 데이터를 쓸 수 있을까?
  파생되는 정보는 없을까?

- 어떤 모델링을 해야
  정확하게 예측할 수 있을까?

# 데이터 소개

Meal delivery company dataset
(Saptarshi Ghosh provides, Kaggle, 2018.12.)




**train.csv**

id 주문 식별넘버 / week 주문 주간(1~145)
center_id 센터 식별넘버 / meal_id 식사 식별넘버
checkout_price 최종 결제금액 / base_price 기본 결제금액
emailer_for_promotion 이메일 프로모션 대상 여부
homepage_featured 홈페이지 추천 여부
num_orders 주문수

**fulfilment_center_info.csv**

center_id 센터 식별넘버
city_code 도시 고유코드
region_code 지역 고유코드
center_type 센터 종류(A~C)
op_area 배달 가능한 거리

**meal_info.csv**

meal_id 식사 식별넘버
category 식사 종류(음료, 샐러드, 디저트 등등)
cuisine 요리 종류

# 02

# 가설과 평가지표, 데이터 전처리

# 본격적인 모델링 전,
# 무엇이 필요할까요?

## 01

### 가설 설정
### (변수 간 관계 추론)

"배달 가능한 거리와 주문량 간에는 양의 상관관계가 있지 않을까?" etc.

## 02

### 기준 모델,
### 평가 지표 설정

기준모델로 '평균' 설정
평가지표 MAE, Rsqaure 설정

## 03

### 데이터 전처리
### (EDA, 특성공학 등)

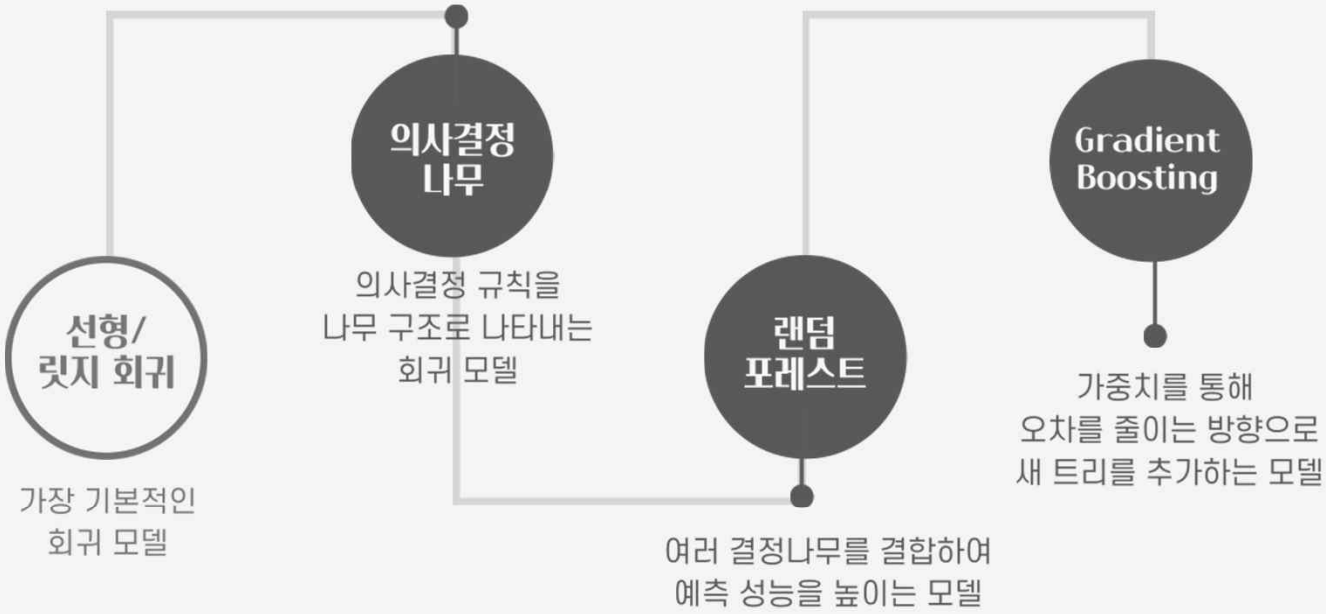
결측치 확인, 처리
파생변수 생성,
이상치(상위 5%) 제거,
데이터 누출(leakage) 방지

# 03

# 머신러닝 적용과 검증, 모델 해석

# 모델 적용, 검증

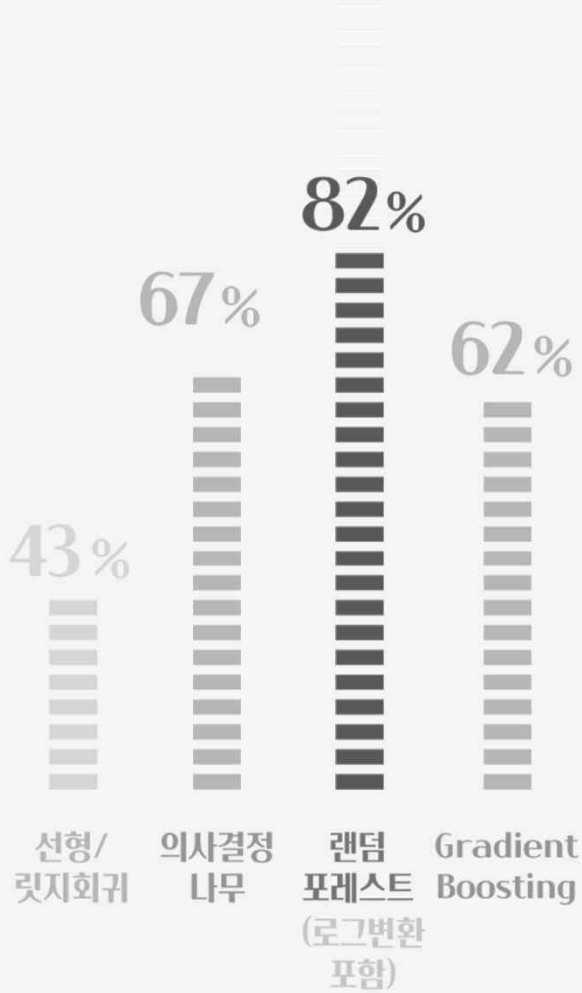
## 4가지 회귀모델 중 어느 것이 가장 잘 예측했을까요?



# 검증 정확도 평가 결과

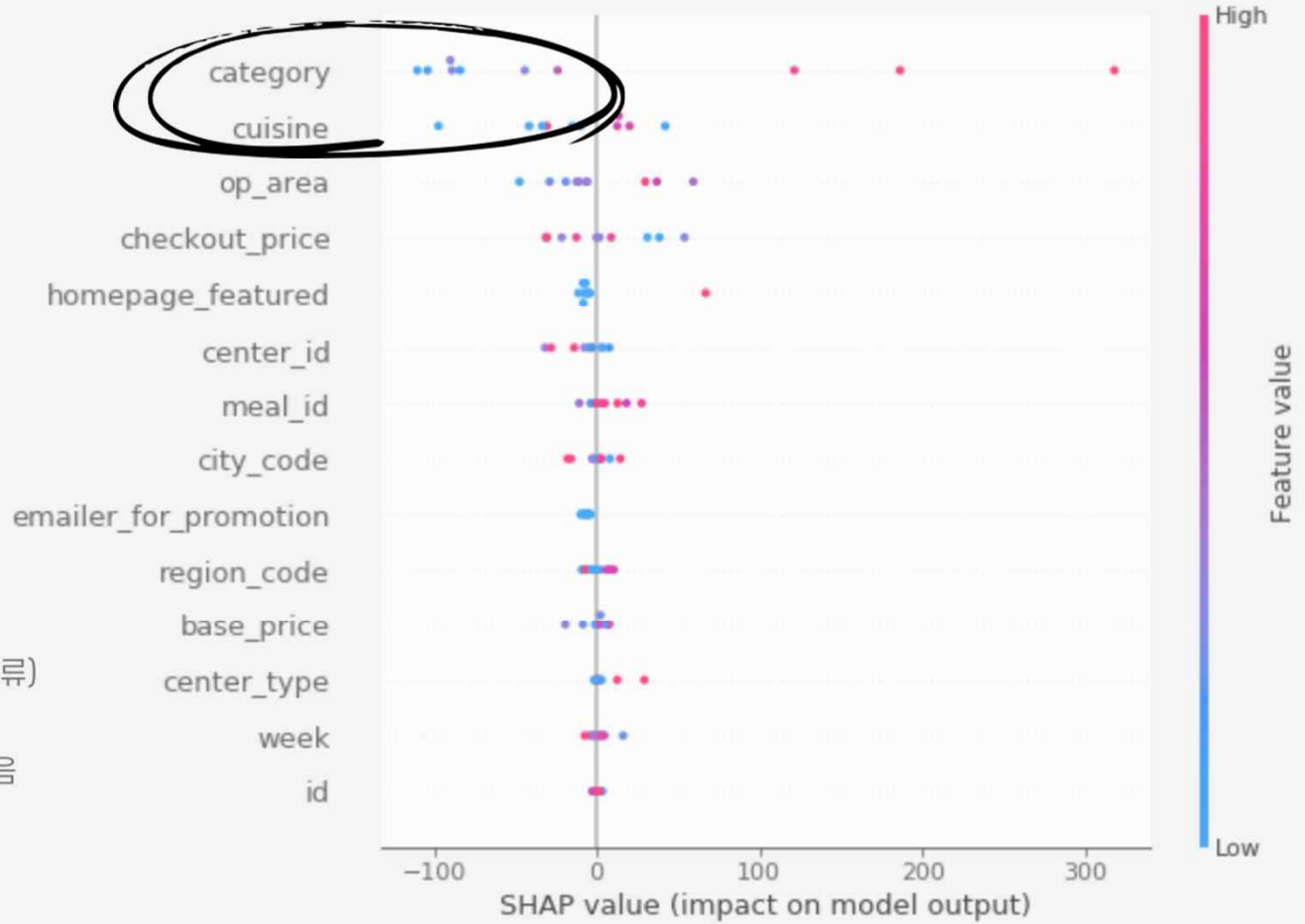
## 훈련한 모델의 정확도를 검증 세트로 확인

《검증 정확도》
선형/릿지 회귀 0.429
의사결정나무 0.668
랜덤포레스트 0.822(로그변환시 0.816)
Gradient Boosting 0.62



# 시각화를 통한 모델 해석

## 어떤 변수가 식사 주문량에 가장 큰 영향을 미칠까요?




category (식사의 범주) / cuisine (식사의 종류)
→ 다른 변수에 비해
식사 주문량에 음의 영향을 끼치는 경우가 많음

# 시각화를 통한 모델 해석

## 랜덤한 1개의 데이터를 추출해 모델이 판단한 결과를 살펴봅니다.


cuisine / meal_id (식사의 종류/고유번호)

op_area (배달 가능한 거리)

→ 해당 데이터의 식사 주문량에 양의 영향을 주고 있음

# 04

# 한계와 보완 가능성

모델링 과정에서의 한계와
보완 가능성을 살펴봅니다.

# 한계와 보완 가능성

## 모델링 과정에서의 한계와 보완 가능성을 살펴봅니다.

### 01 특성공학의 어려움

→ 지역 특성 등의 추가적인 정보로
특성공학 과정을 개선할 수 있습니다.

### 02 데이터의 불균형과 이상치 문제

→ 분포의 편향을 줄이면
모델이 더 잘 예측하게 할 수 있습니다.

### 03 모델의 다양성 부족

→ 다양한 모델을 구축해봄으로써
가장 좋은 모델을 찾아낼 수 있습니다.

### 04 하이퍼파라미터 튜닝 문제

→ 최적 하이퍼파라미터를 발견, 적용하여
모델의 성능을 개선할 수 있습니다.

### 05 현업에 최적화된 모델 선택의 문제

→ 현실적으로 어떤 모델이 가장 적합할지
논의를 통해 선택할 수 있습니다.

# 감사합니다

**Special Thanks to:**

Saptarshi Ghosh (data)

Miricanvas, Pixabay (free imgs & presantation format)