판담에게 물어봐 !

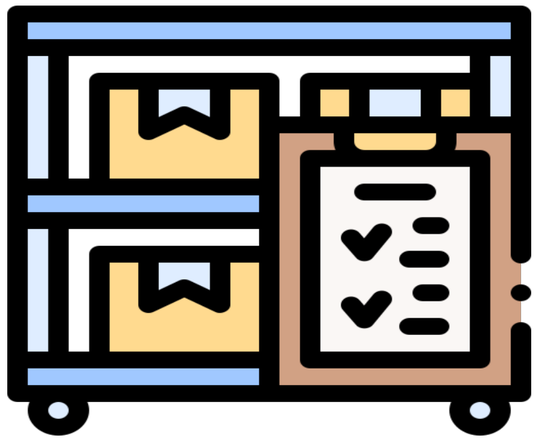
판매량 예측모델, 판담을 이용한
손실 방지 프로젝트

프로젝트 진행

기간	2023.04.28 ~ 2023.05.18 약 3주간 진행
인원	4명
언어	Python
모델	Machine learning

목차

01 프로젝트 소개
주제 선정 이유



01 프로젝트 소개

목표

물건을 "판다" 단어 + 모델의 "ㅁ"을
합쳐 만든 판담

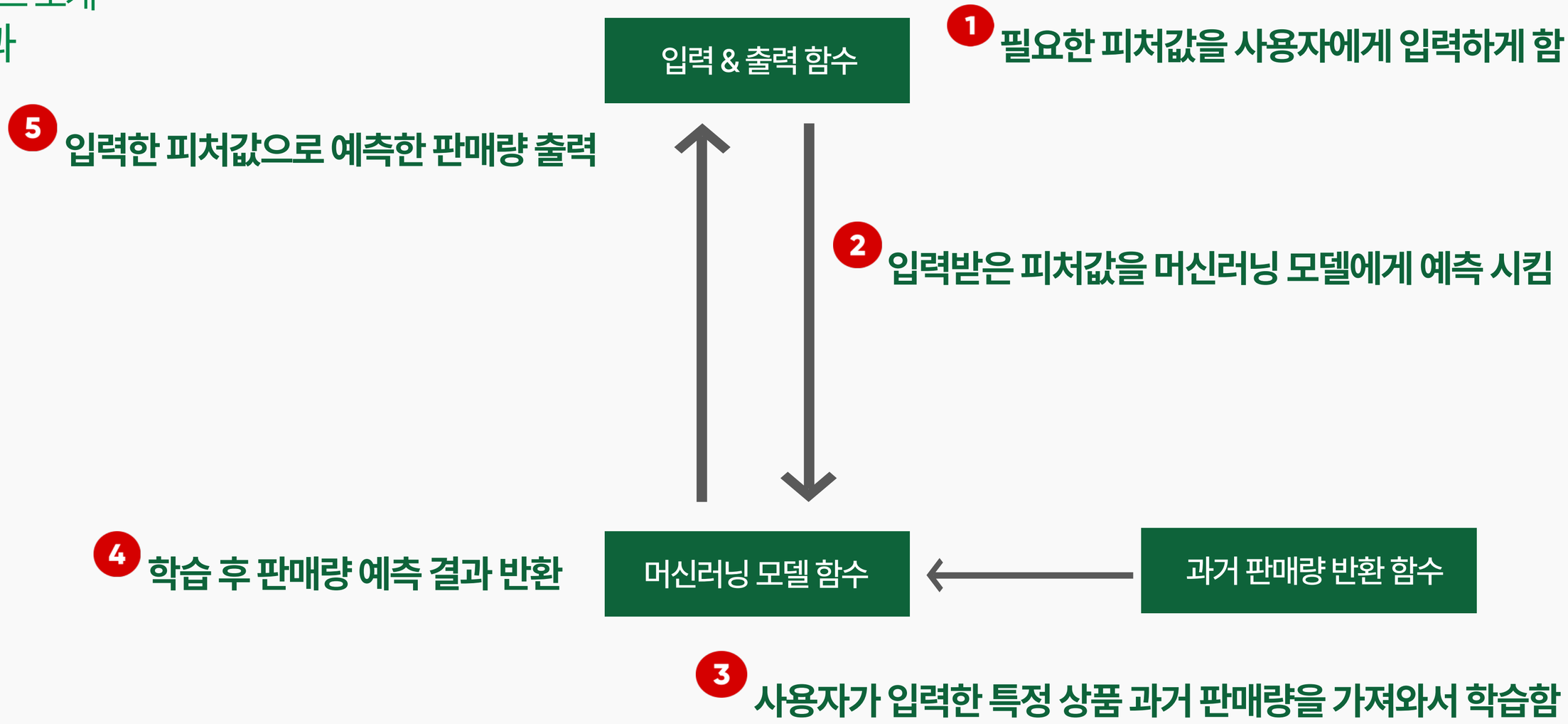
판매량을 예측해
남은 재고를 폐기하는
손실을 방지하자 !

모델 판담에게 기대되는 효과

이커머스	미리 상품별 판매량을 예측해 재고가 부족하지않도록 하여 고객의 만족도는 올리고 재고가 남아 손실이 생기는 사태는 방지
유통	고객사에게 모델을 제공해 그에 맞는 인력 대비 가능
제조	판매량 예측 모델을 적용하여 인기있는 제품을 더 많이 생산 가능
공통	이익 증진 효과가 기대됨

그렇다면 판매량 예측 모델,
판담은 어떻게 결과를
출력할까 ?

01 프로젝트 소개
모델 결과



01 프로젝트 소개
모델 결과

```
========================================
안녕하세요 판매량 예측 모델, 판담입니다.
특정 일의 판매량 예측을 위한 8가지 질문을 시작하겠습니다.

1번) CA, TX, WI 중 원하시는 주를 입력해주세요.
주 입력 : CA
1번) 1 ~ 4 중 원하시는 지점 번호를 입력해주세요.
지점 번호 입력 : 3

2번) 1 ~ 3 중 원하시는 식품 분류 번호를 입력해주세요.

식품 분류 번호 입력 :
```

1 사용자가 직접 피처들을 입력

```
3번) 1 ~ 827 중 원하시는 식품 id를 입력해주세요.
식품 id 입력 : 90

4번) 평일->0, 주말->1 중 원하시는 날짜의 구분 숫자를 입력해주세요.
평일/주말 구분 숫자 입력 : 1

5번) 일반->0, 행사->1 중 원하시는 날짜의 행사유무 숫자를 입력해주세요.
행사 유무 숫자 입력 : 1

6번) 원하시는 날짜의 기온을 입력해주세요.
기온 입력(° C): 30

7번) 원하시는 날짜의 매장 당 인구 수를 입력해주세요.
매장 당 인구 입력(명): 120800

8번) 원하시는 날짜의 실업률을 입력해주세요.
실업률 입력(%): 10

========== 결과 출력 중 잠시만 기다려주세요.===========

Fitting 3 folds for each of 100 candidates, totalling 300 fits
 최적 learning_rate: 0.3
 최적 max_depth: 3
 최적 n_estimators: 100

입력하신 조건에 해당하는 일의 CA_3지점 food_3_90 판매량은 145입니다.

이용해주셔서 감사합니다.
========================================
```

2 사용자가 입력한 피처들로 특정 상품 판매량 예측 결과 출력

판담을 개발한 과정을
보여드리겠습니다.

02 개발과정
라이브러리

02 개발과정
기존 데이터

주제 : 판매량 예측 모델

Walmart 데이터

캘리포니아(4개의 매장)

텍사스(3개의 매장)

위스콘신(3개의 매장)

식품(1437개의 id)

취미용품(565개의 id)

가정용품(1047개의 id)

총 1913일의 데이터 존재(2011-01-29 ~ 2016-04-24)
정확한 지점명, 상품명은 존재하지않는다.

02 개발과정

데이터 수집 및 분석

기존 변수

행사

행사유무에 따른 판매량의 변화

요일

평일/주말에 따른 판매량의 변화

+

추가 변수

기온

세계 기상자료 개방 포털

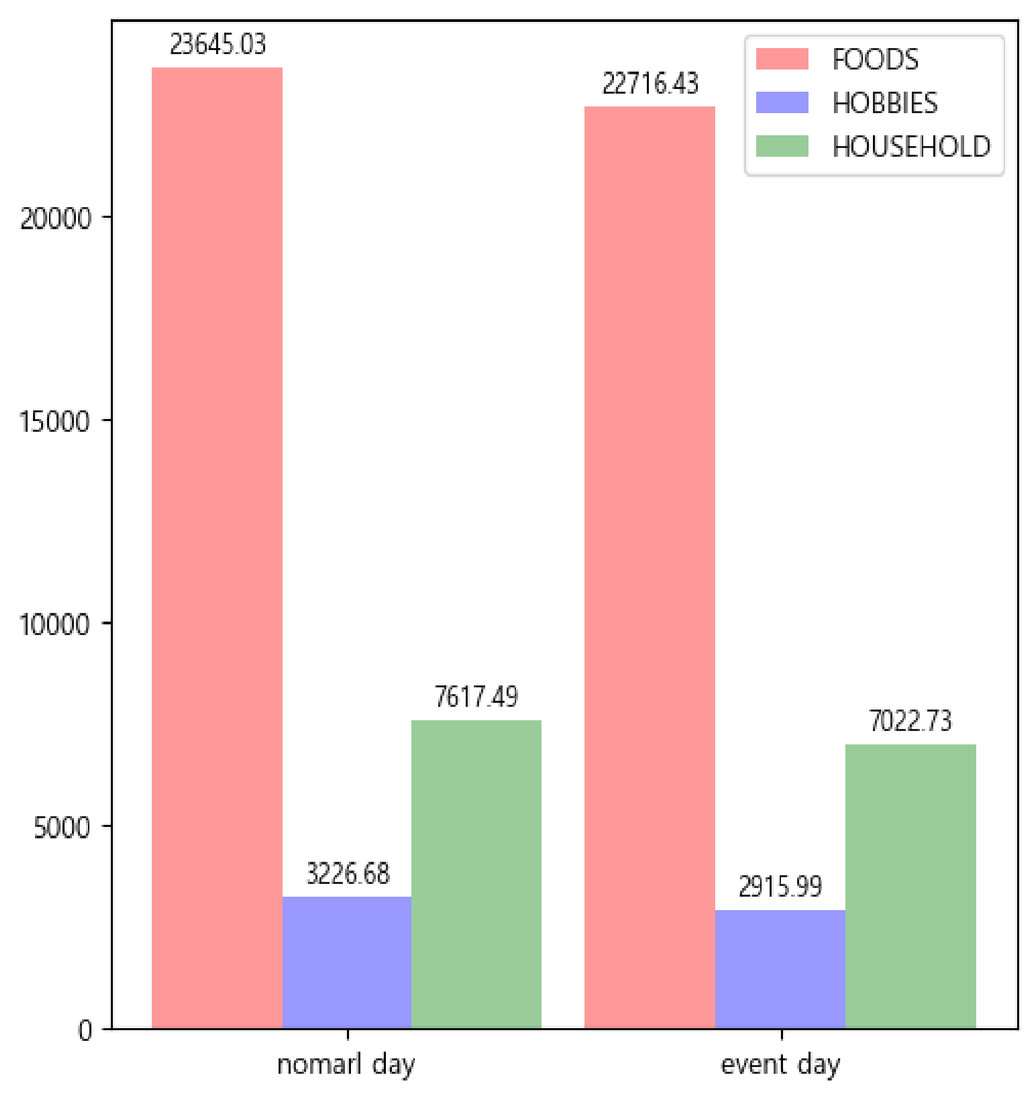
매장당 평균 인구수

미국 인구 데이터(2010,2020)

실업률

미국 실업률, 1990-2016

02 개발과정
데이터 수집 및 분석



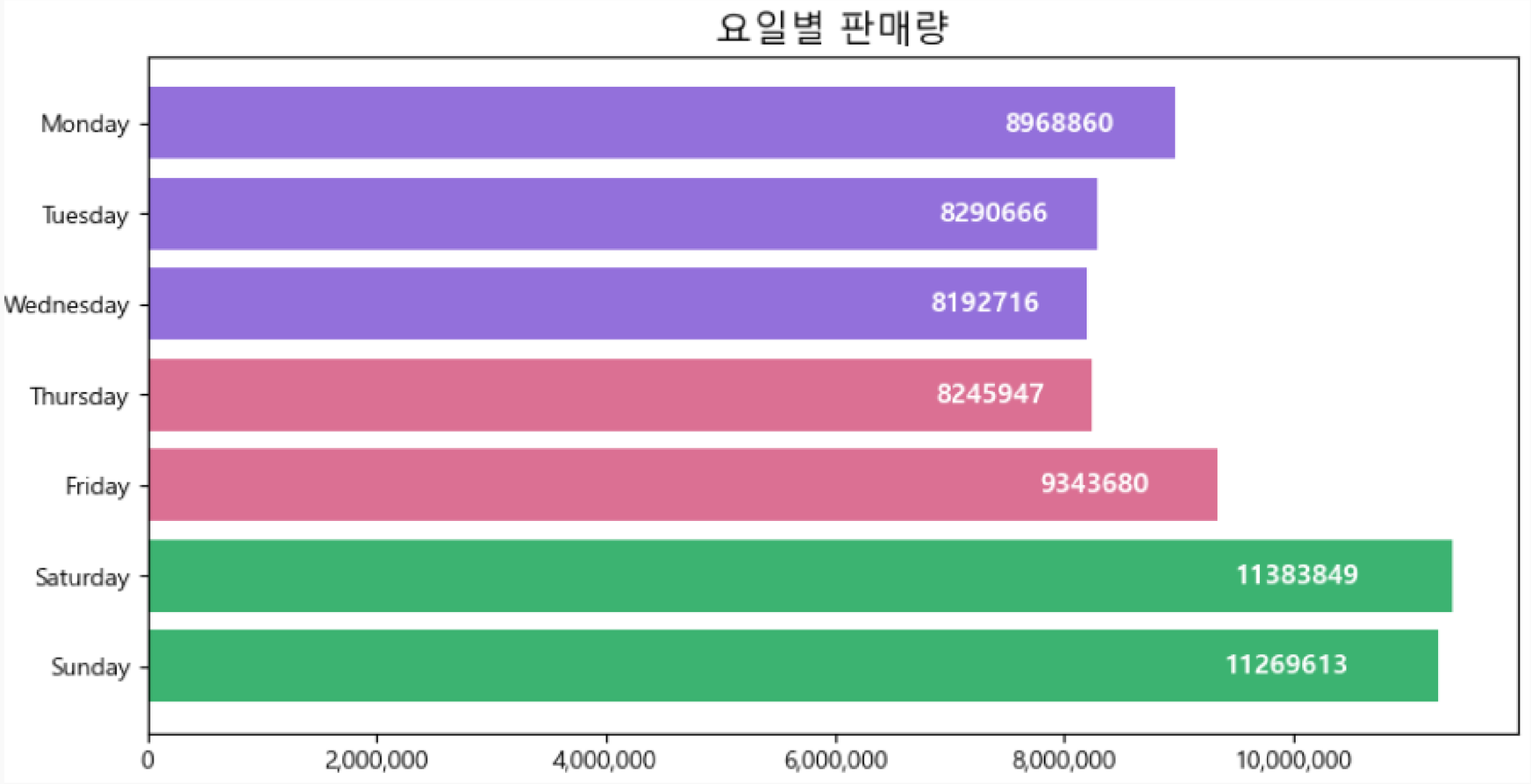
행사에 따른 판매량의 변화

- 행사유무에 따른 카테고리별 판매량
- 3개의 카테고리 모두 행사가 없는날에 판매량이 더 높다.
- 그러므로 행사는 판매량에 영향을 준다고 판단하였다.

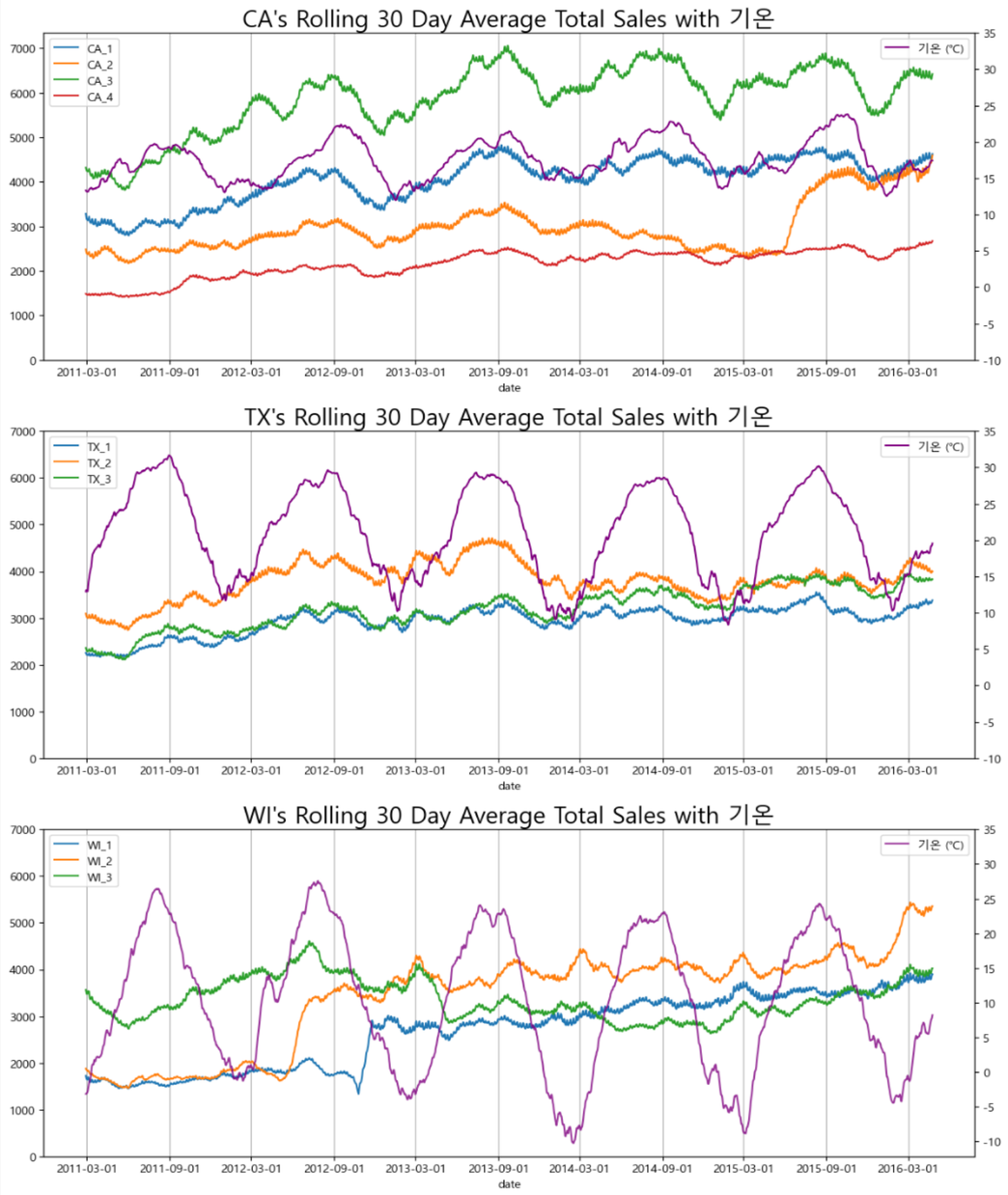
02 개발과정
데이터 수집 및 분석

요일에 따른 판매량의 변화

평일보다 주말 판매량이 많아 요일은 판매량의 영향을 준다고 판단하였다.



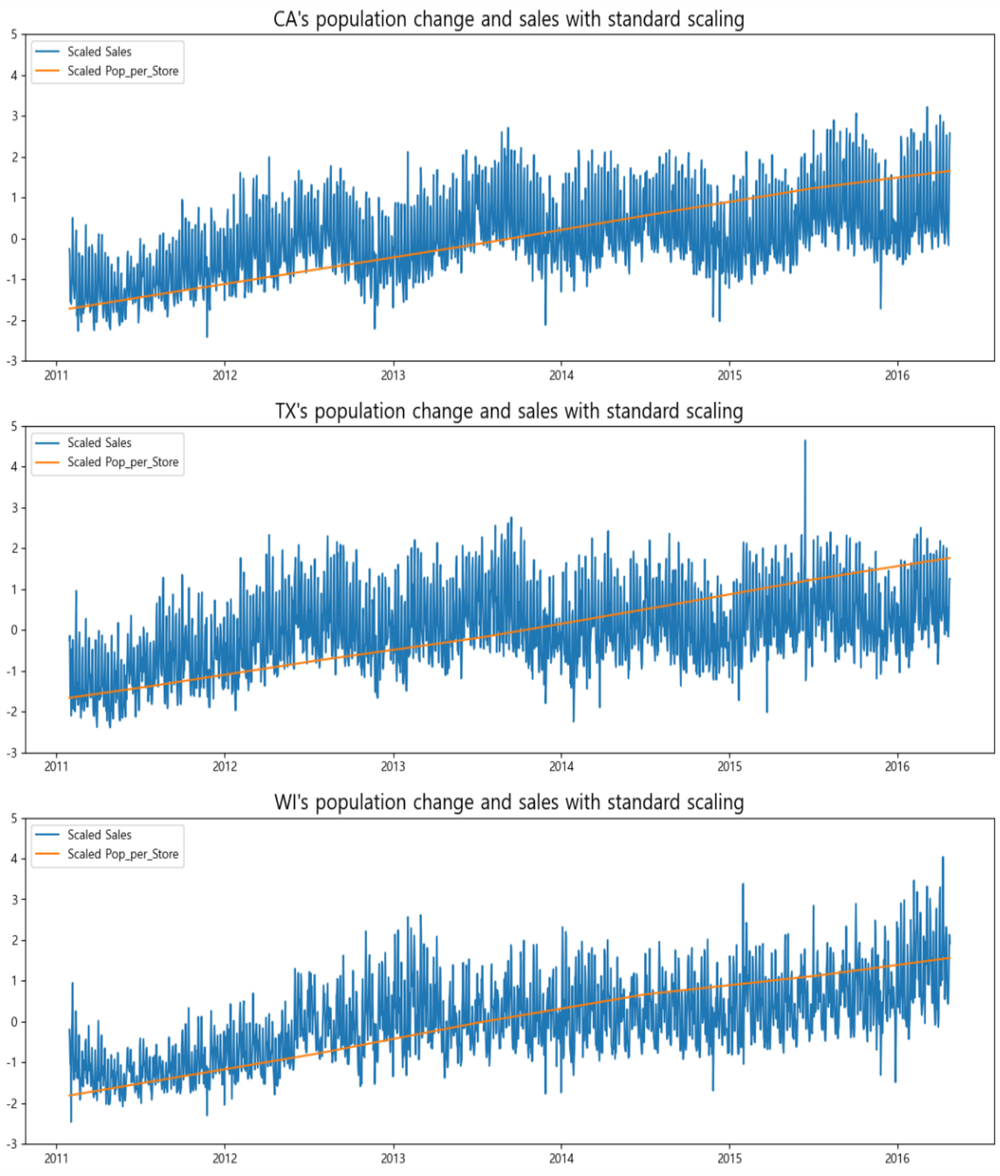
02 개발과정
데이터 수집 및 분석



평균기온에 따른 판매량의 변화

- 지점별 기온과 판매량(30일 단순이동평균)
- 캘리포니아와 텍사스의 경우 평균기온이 올라가면 판매량도 올라가는 변화를 보이고 있지만
- 위스콘신의 경우 온도가 0°C 미만이면 반대로 판매량이 올라가는 변화를 보이고 있다.
  0°C 미만으로 내려가는 부분은 절대값으로 처리
- 평균기온은 판매량에 영향을 준다고 판단하였다.

02 개발과정
데이터 수집 및 분석



매장당 인구수에 따른 판매량의 변화

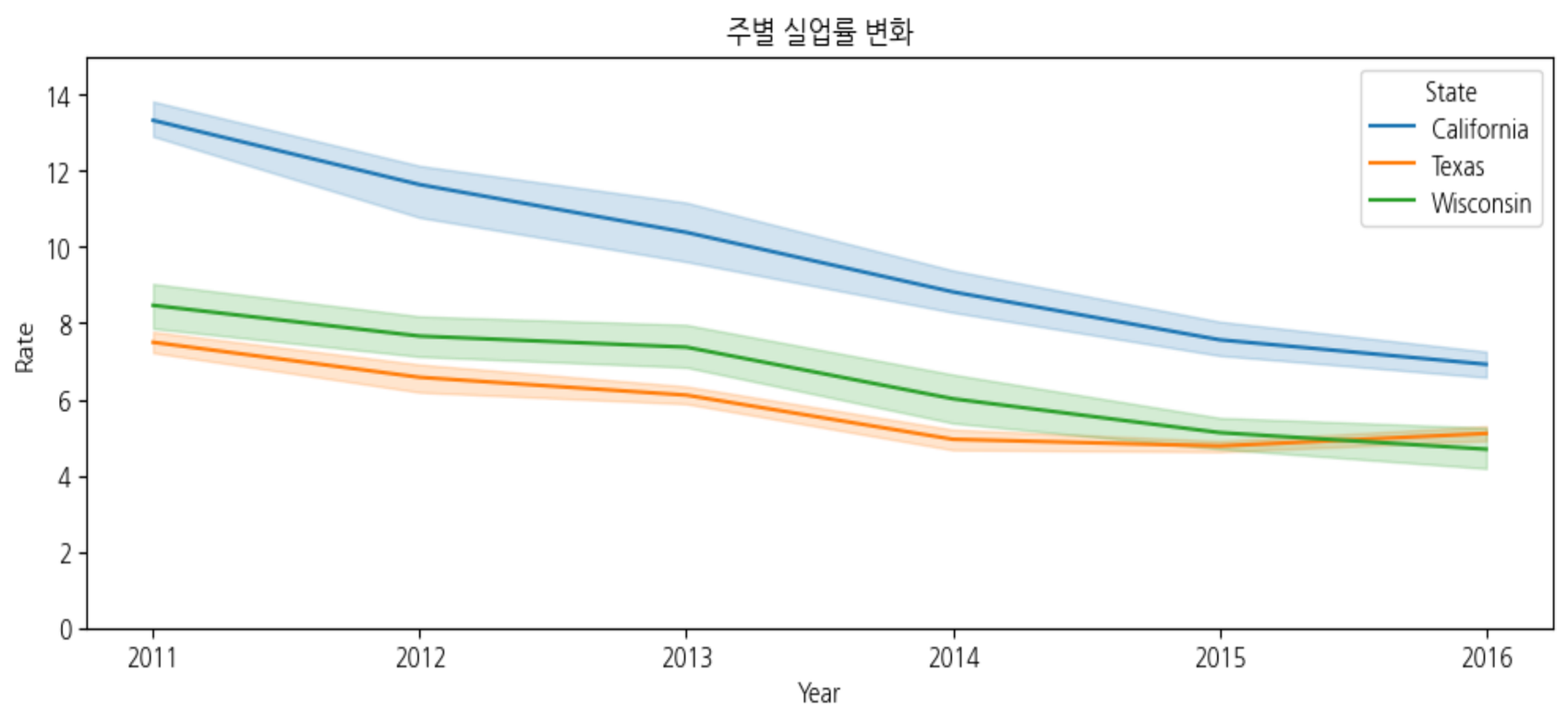
- 한 그래프에 인구수와 판매량의 흐름을 비교하기 위해 스탠다드스케일러 적용
- 3개 주 모두 매장당 인구수가 증가 할 수록 판매량도 증가한다.
- 그러므로 매장 당 인구수는 판매량에 영향을 준다고 판단하였다.

02 개발과정
데이터 수집 및 분석

실업률에 따른 판매량의 변화

전체적으로 실업률이 떨어지는 변화가 보인다.

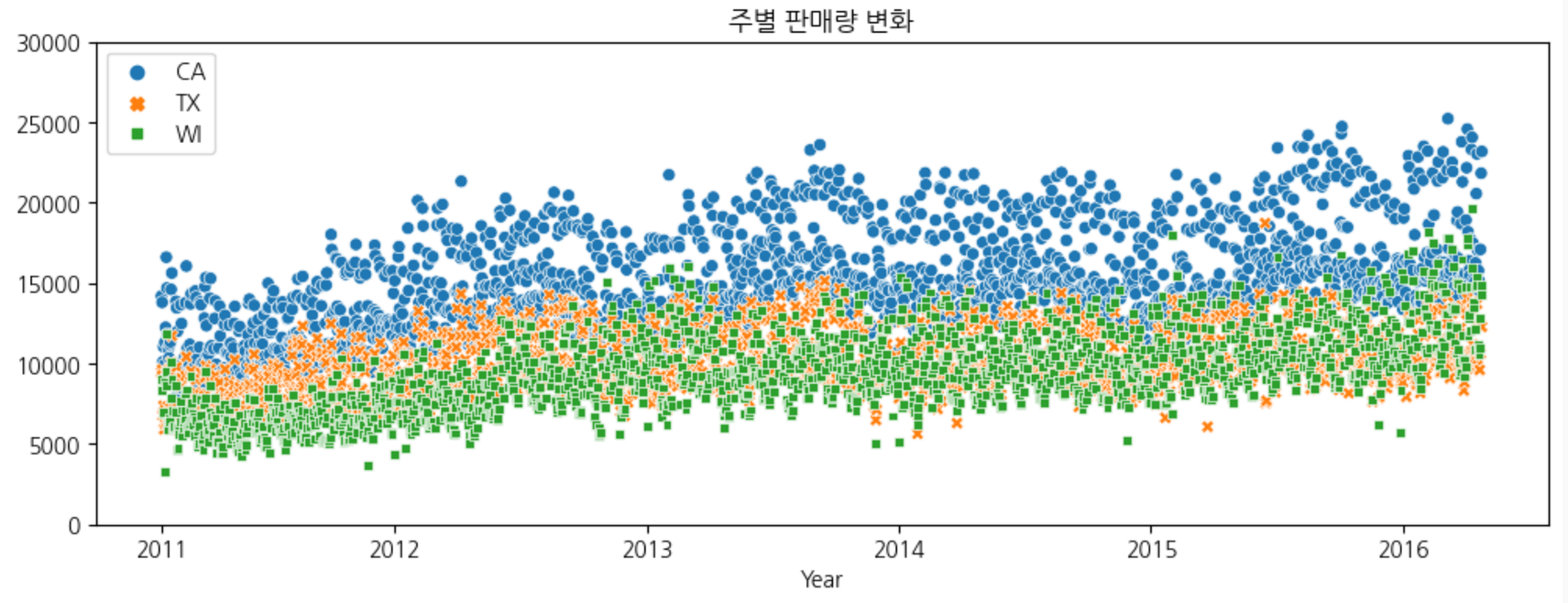
실업률 데이터에서 텍사스에 주목하면 텍사스가 위스콘신보다 실업률이 높아지는 지점이 존재



02 개발과정
데이터 수집 및 분석

실업률에 따른 판매량의 변화

텍사스가 위스콘신보다 실업률이 높아지는 2015년대 중반 지점에서 판매량 또한 위스콘신에게 따라잡히는 것을 보아 실업률은 판매량의 영향을 준다고 판단하였다.



02 개발과정
모델 선정

모델 종류

공통

train : test 비율 8:2로 설정
feature는 평일/주말, 행사여부, 평균 기온, 매장당인구수, 실업률

모델 01 Gradient Boosting Regressor

모델 02 Random Forest Regressor

모델 03 Kernel Ridge

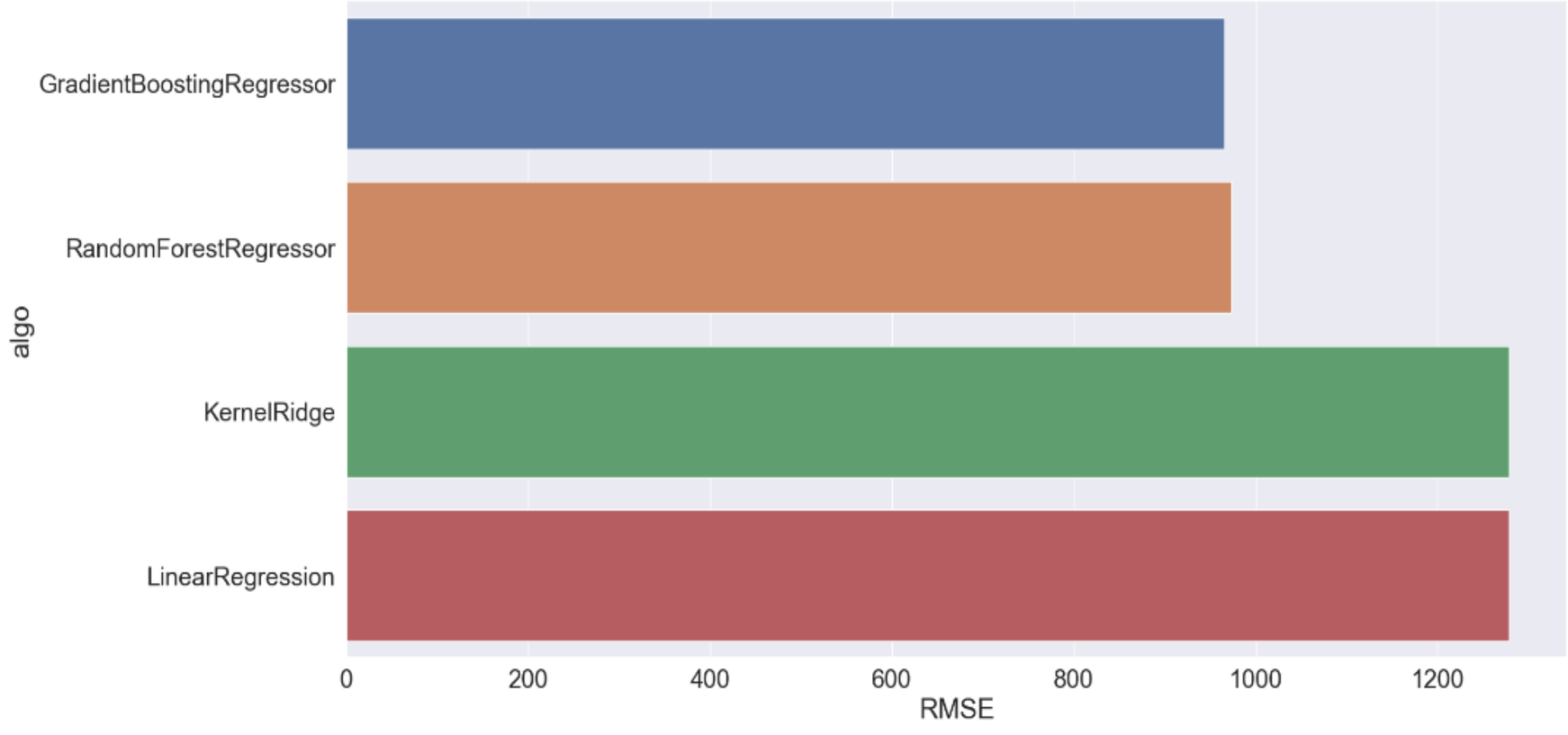
모델 04 Linear Regression

02 개발과정
모델 선정

모델별 RMSE score - 낮을수록 성능 높음

전체 주 판매량 test RMSE score

Gradient Boosting Regressor이 965로 제일 낮다.

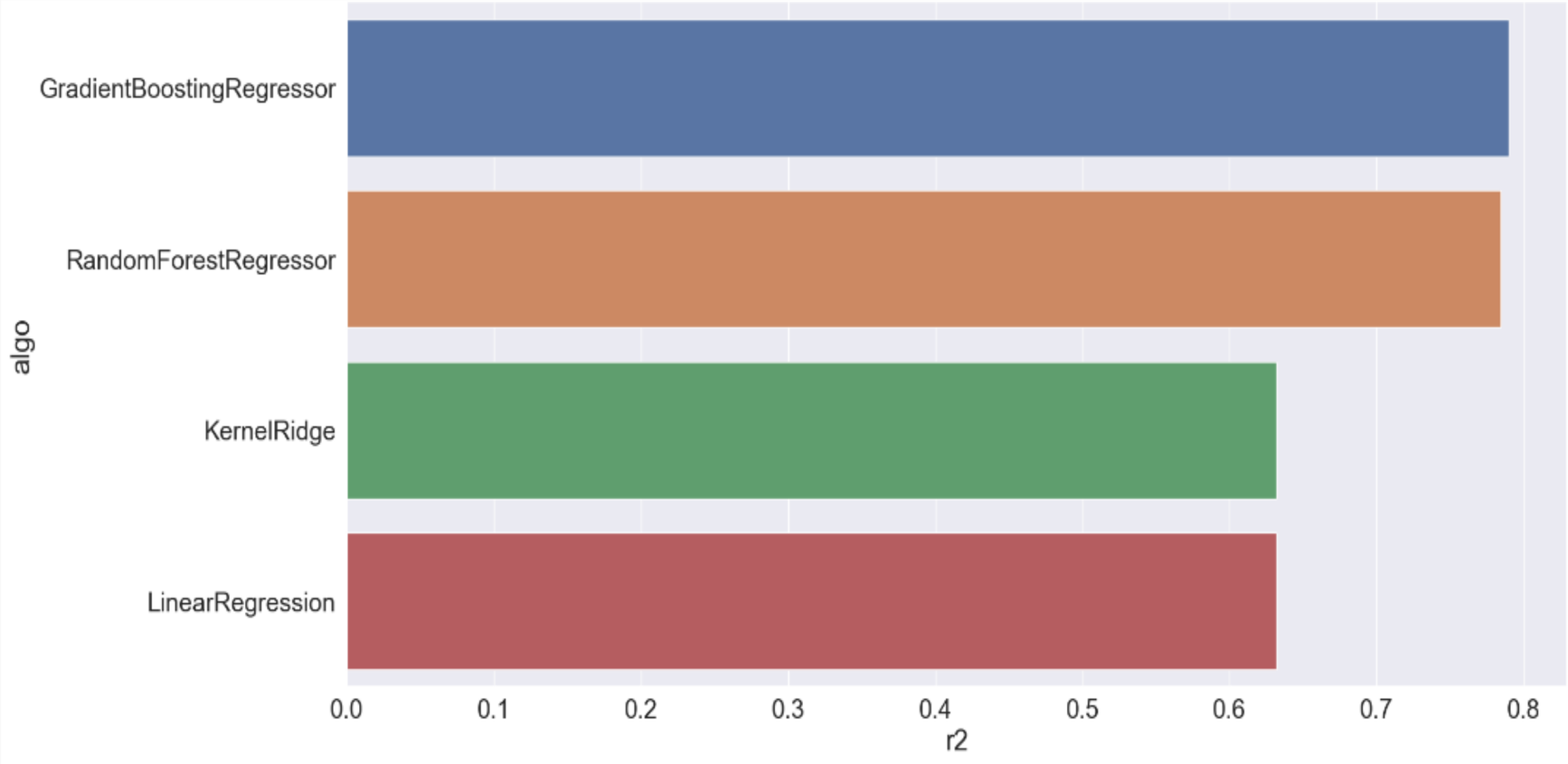


02 개발과정
모델 선정

모델별 R2 결정계수 - 높을수록 성능 높음

전체 주 판매량 test R2

Gradient Boosting Regressor이 0.79로 제일 높다.



GridSearchCV - 최적의 파라미터 찾기

사용자가 선택한 지점별 특정상품의 과거 판매량으로 학습하기 때문에

최적의 파라미터는 선택한 상품에 따라 달라진다.

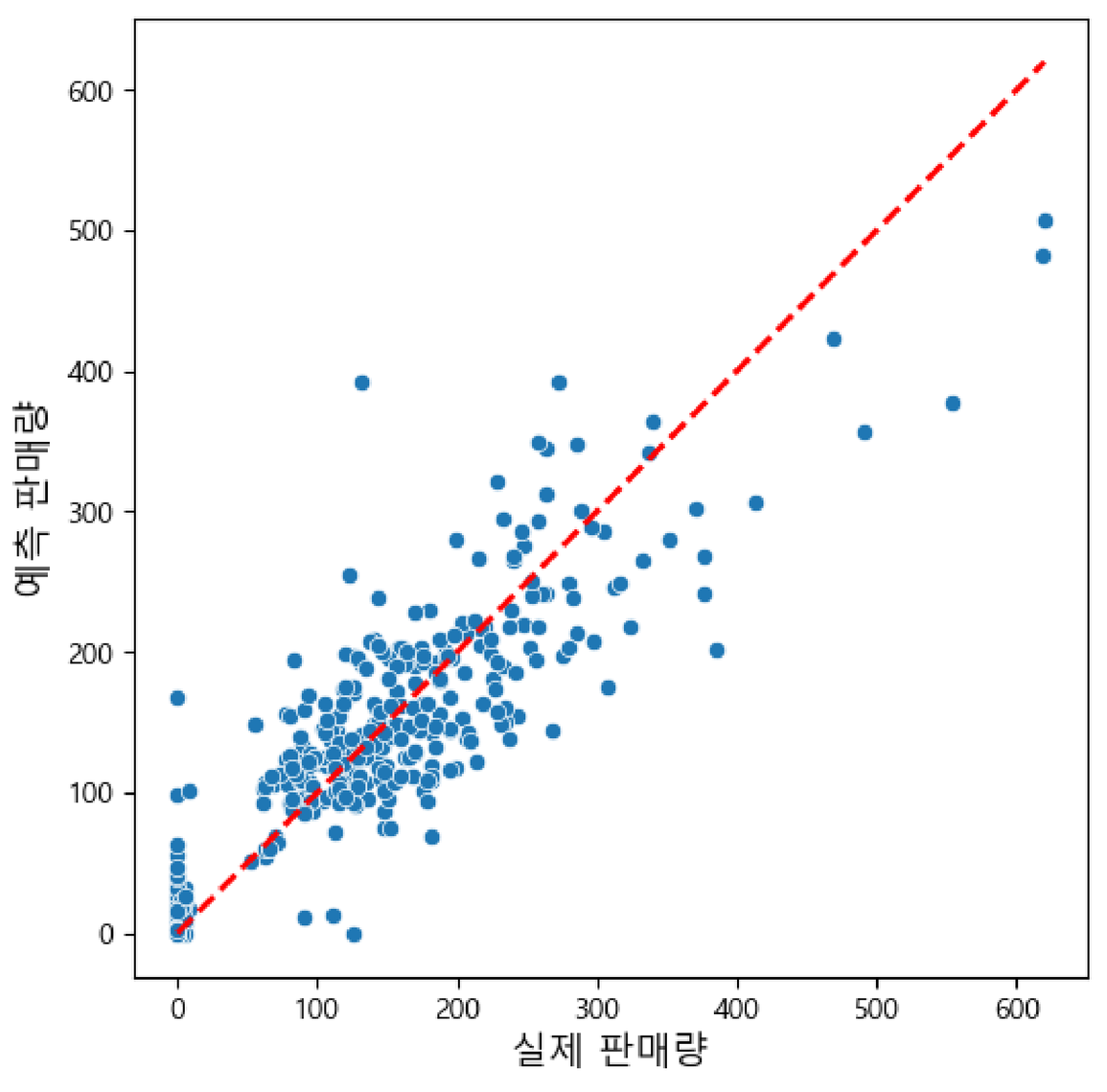
```
========= 결과 출력 중 잠시만 기다려주세요.==========

Fitting 3 folds for each of 100 candidates, totalling 300 fits
 최적 learning_rate: 0.3
 최적 max_depth: 3
 최적 n_estimators: 100

입력하신 조건에 해당하는 일의 CA_3지점 foo최적의 파라미터를입니다.

이용해주셔서 감사합니다.
========================================
```

머신러닝 모델 함수 안에
GridSearchCV로
상품에 맞는
최적의 파라미터를
설정 하는 기능

GridSearchCV 파라미터 범위

- **max_depth : [3,4,5,7,10]**
- **learning_rate : [0.1,0.2,0.3,0.4,0.5]**
- **n_estimators : [50,65,80,100,150]**

02 개발과정
모델 선정



예시) CA_3 food_3_090 상품을 예측한 결과

- 최적의 파라미터

max_depth = 4

learning_rate = 0.3

n_estimator = 65

- 성능

test 데이터 RMSE : 48

test 데이터 R2 : 0.79

최종 모델

모델 : Gradient Boosting Regressor

사용한 파라미터

- max_depth : 트리의 깊이
- learning_rate : 학습률
- n_estimators : 트리의 수

사용한 성능 평가 지표

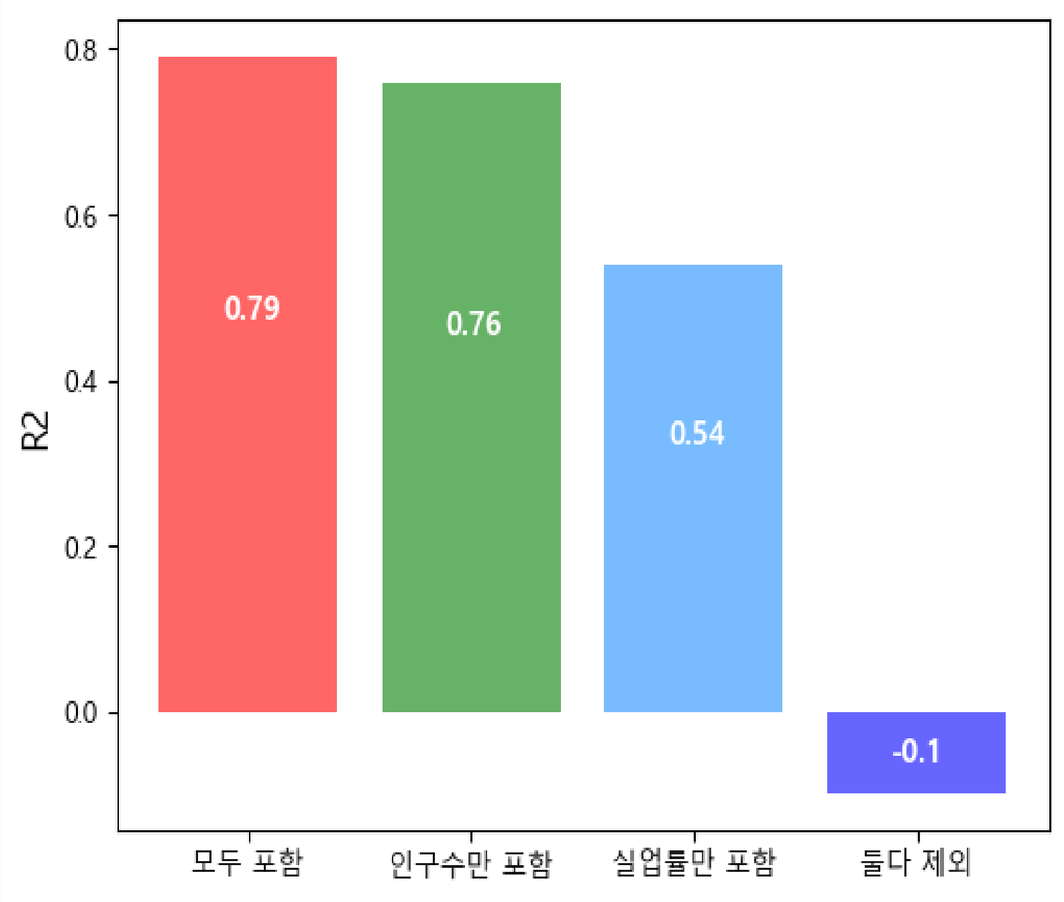
- RMSE
- R2

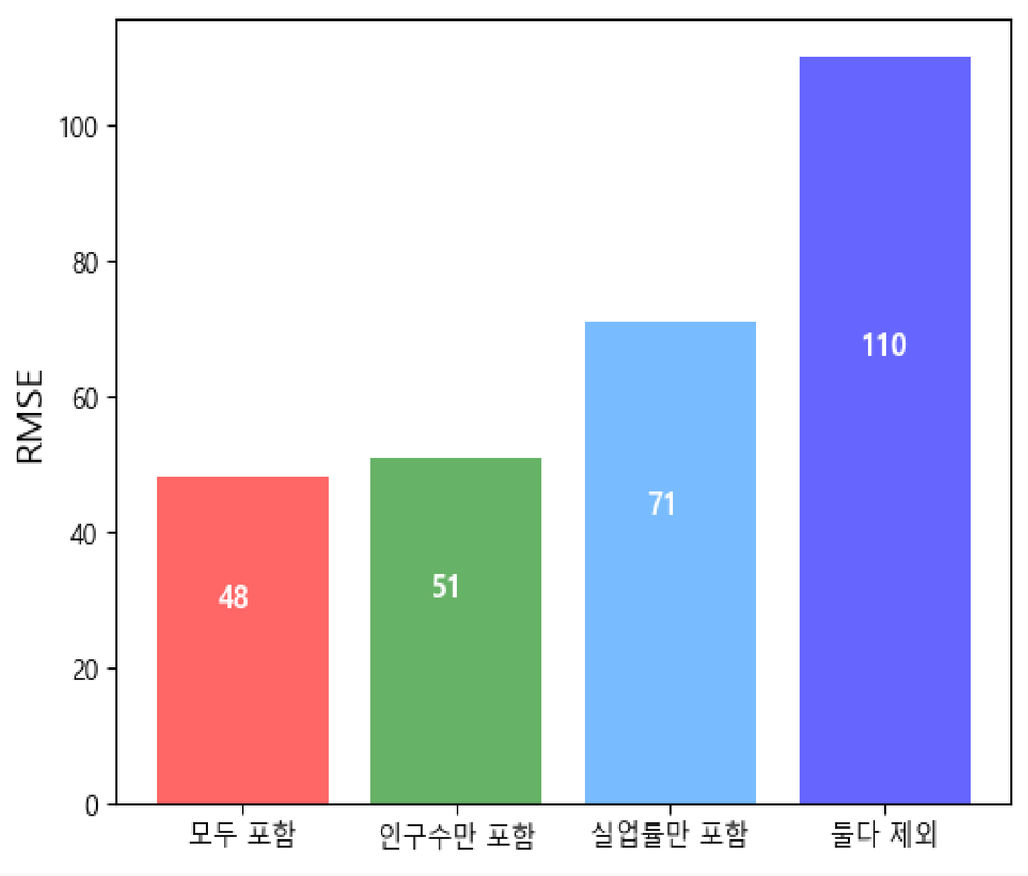
03 회고
모델의 한계점

모델에게 입력해야하는 feature 중
매장당 인구수, 실업률은
사용자가 알기 어렵다.

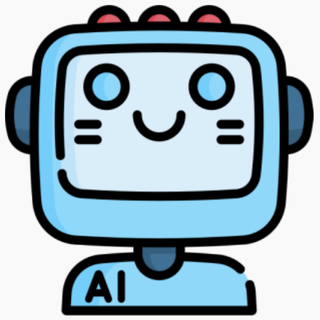
03 회고
모델의 한계점

[평일/주말, 행사여부, 평균기온] +
매장당인구수, 실업률을 모두 포함했을 때 성능이 가장 높았다.





03 회고
모델의 한계점



1. 미래의 매장당 인구수, 실업률을 계산하는 모델을 만들어 추가한다.

2. Prophet을 통한 시계열 예측으로 해당값을 얻을 수 있게 한다.