2018 FAST CAMPUS

DATA SCIENCE SCHOOL PROJECT(1)

RECRESSION ANALYSIS

목차

01.

Walmart Recruiting II : Sales in Stormy Weather

Predict how sales of weather – sensitive products are affected by snow and rain

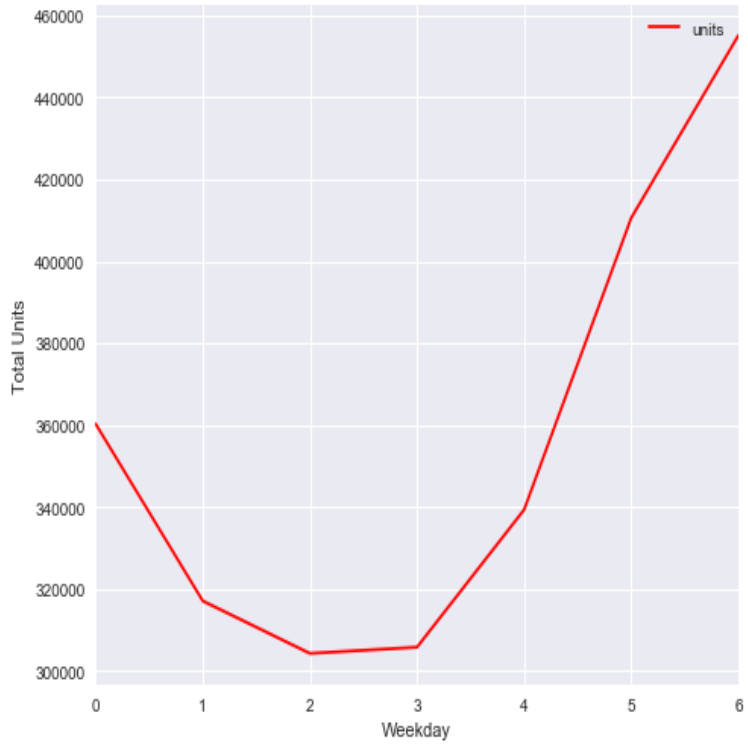
Objective

45개의 store에서 weather-sensitive products are affected by snow and rain

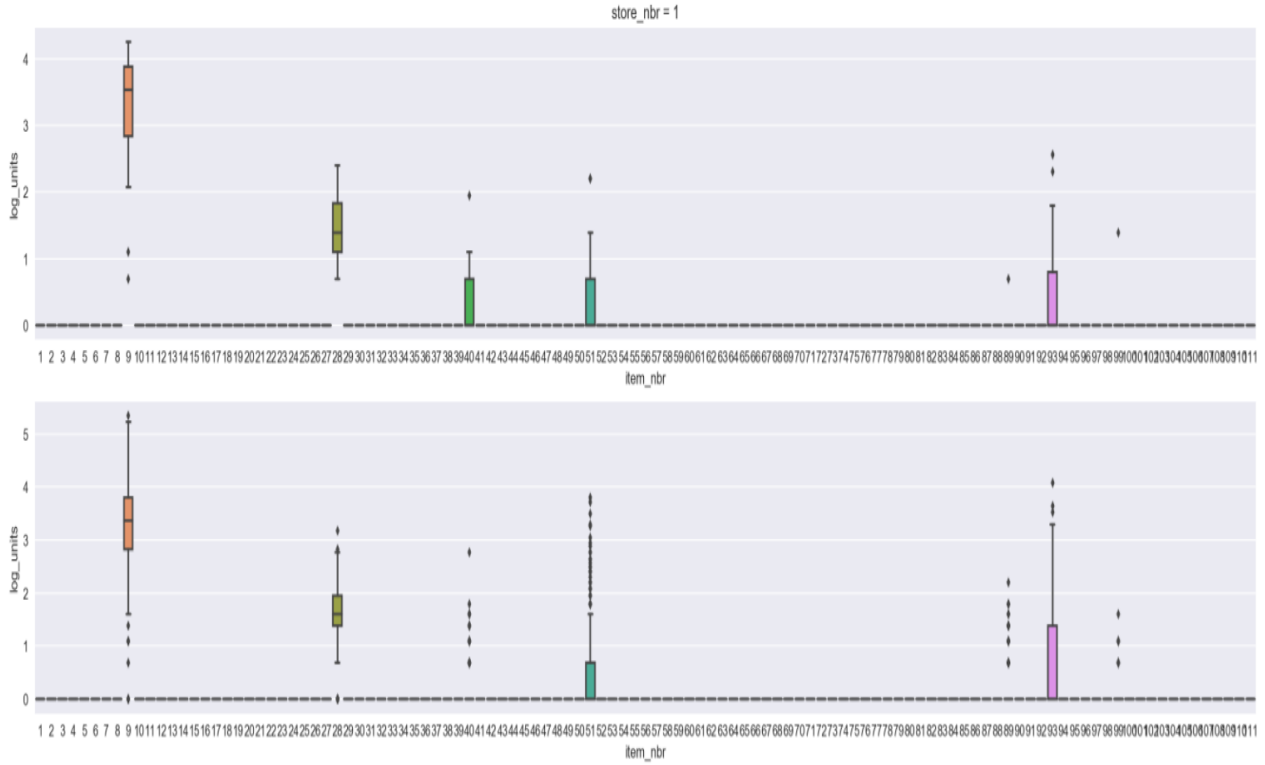
Data Set

- Weather: 2012.01.01 – 2014.10.31의 각 statio날씨
- Key: Store과 Weather Station간의 관계 Mapping
- Train: 2012.01.01 – 2014.10.31의 각 Store, Item 별 Units Data(test날씨 제외)
- Test: 2013.04.01 이후 Weather Event가 발생한 전후 3일

02. EDA



요일별 유닛총량



Store nbr 1의 아이템별 공휴일/비공휴일 판매량

02. EDA

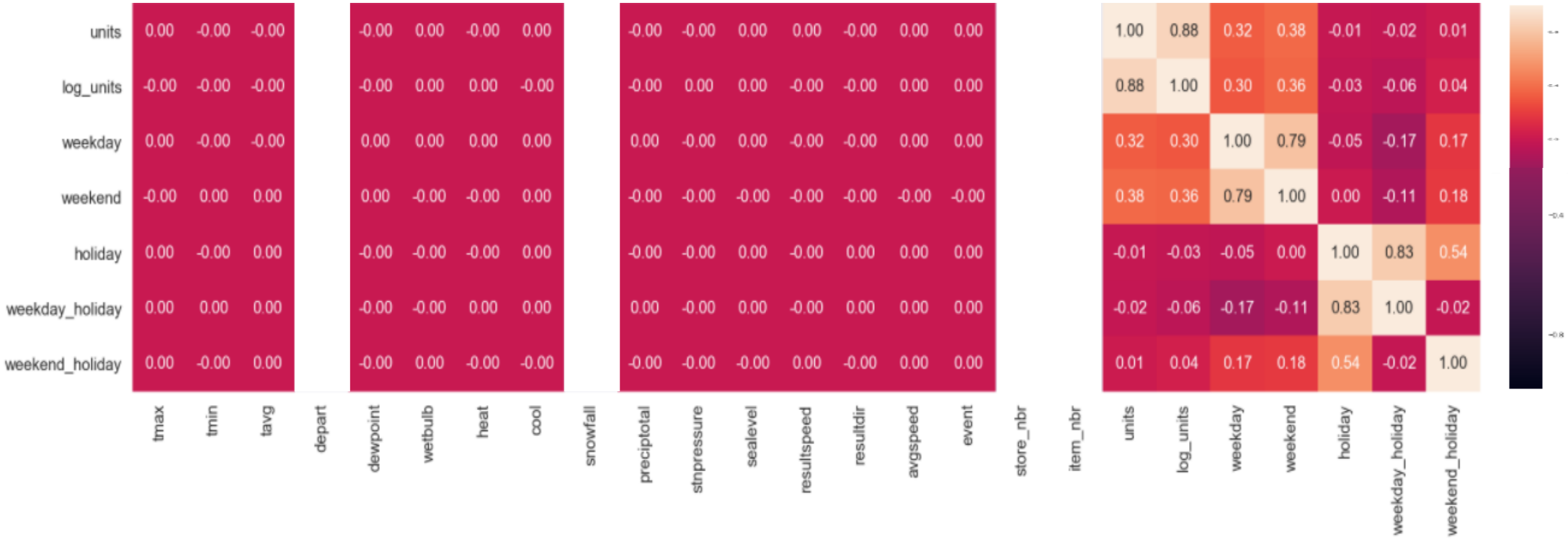
Unit Data

Unit Data event

```
Warning! : 85221     2012-03-19
177573    2012-10-13
Name: date, dtype: object, Match: 2
```

03. CORRELATION PLOT

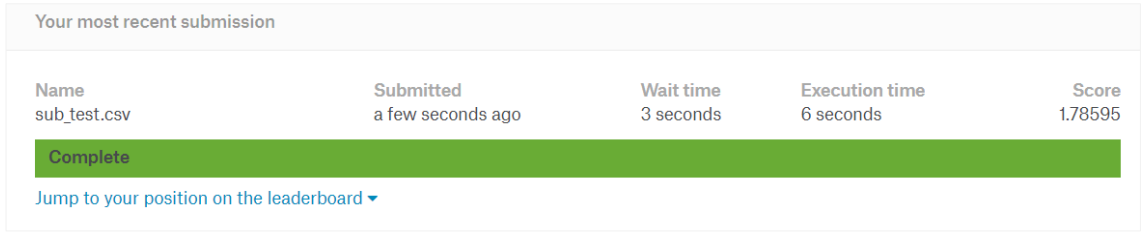
1번 스토어 9번 아이템 Correlation plot



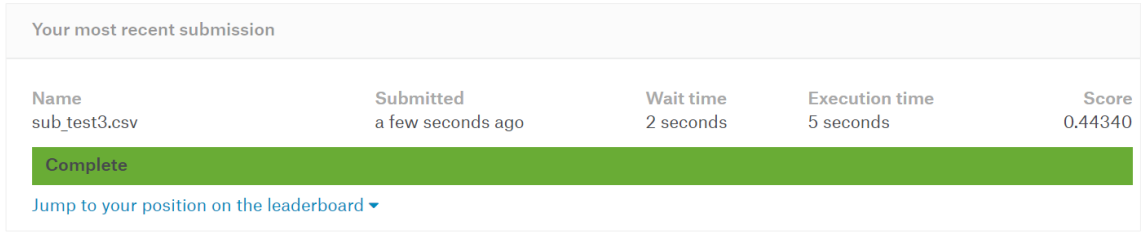
04. FEATURE SELECTION

feature selection
 - 각 store_nbr, item_nbr별로 나누어 modeling 해야한다
 - weather와 log_units(또는 units)는 큰 상관관계가 없어 보인다
 - weekday와 holiday가 log_units(또는 units)와 약간의 상관관계가 있어 보인다
(item_nbr에 따라 다름)

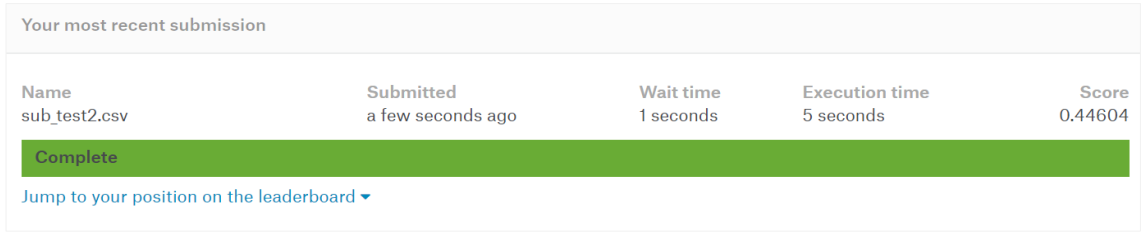
04. OLS MODELING



1’ OLS. 2등(뒤에서)
log_units ~ C(store_nbr) + C(item_nbr) + C(weekday) + C(holiday) + C(event) + 0

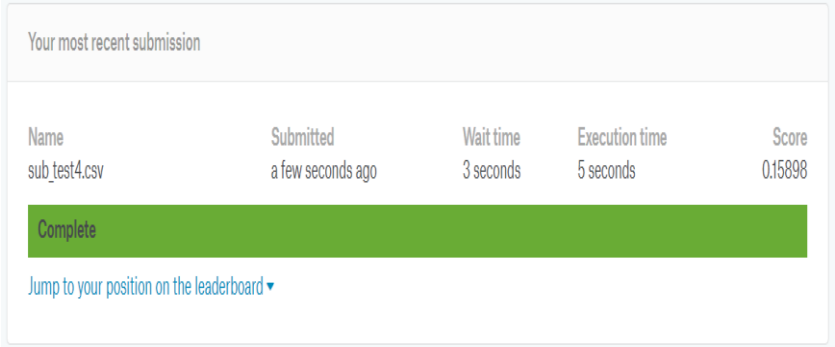


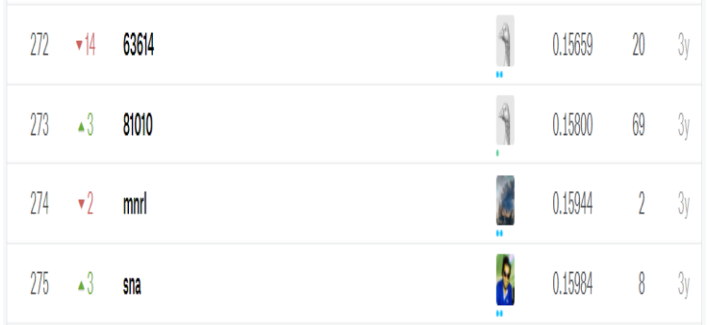
2’OLS. 485명중 350등
C(store_nbr):C(item_nbr) + C(weekday) + C(holiday) + C(event) + 0



2-1’OLS.
C(store_nbr):C(item_nbr) + C(weekday) + C(holiday) + snowfall + preciptotal + 0

04. OLS MODELING



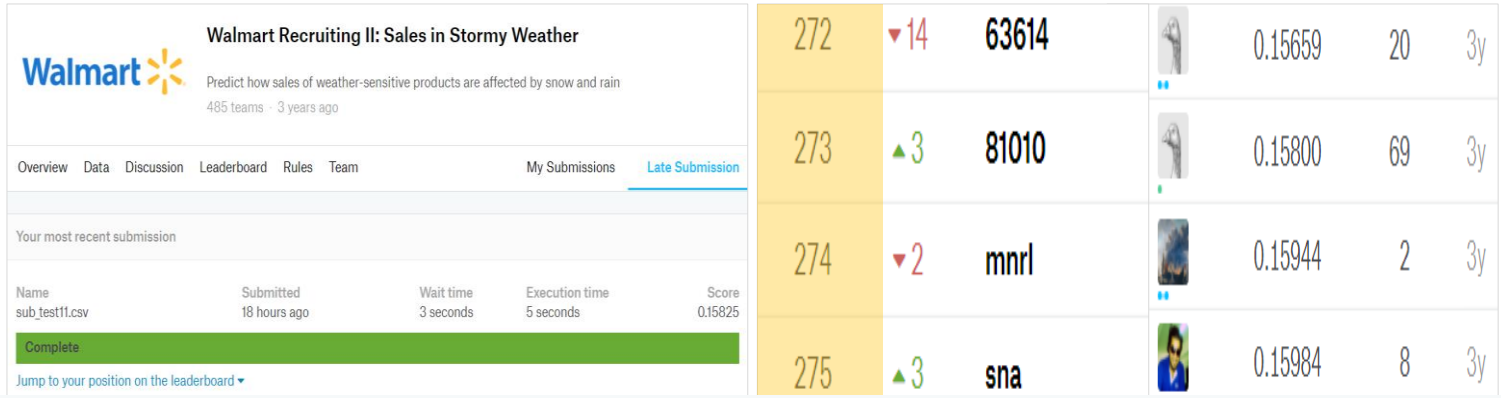


3번
log_units ~ C(station_nbr):C(store_nbr):C(item_nbr) + C(weekday) + C(holiday) + C(event) + 0
memory error 발생
station_nbr별로 나눠서 ols

4번 log_units ~ C(store_nbr):C(item_nbr) + C(weekday) + C(holiday) + C(event) + 0

5번log_units ~ C(store_nbr):C(item_nbr) + C(weekday) + C(holiday) + snowfall + preciptotal + 0
store_nbr별로 나눠서 ols
event 제외
결과안좋음

04. OLS MODELING

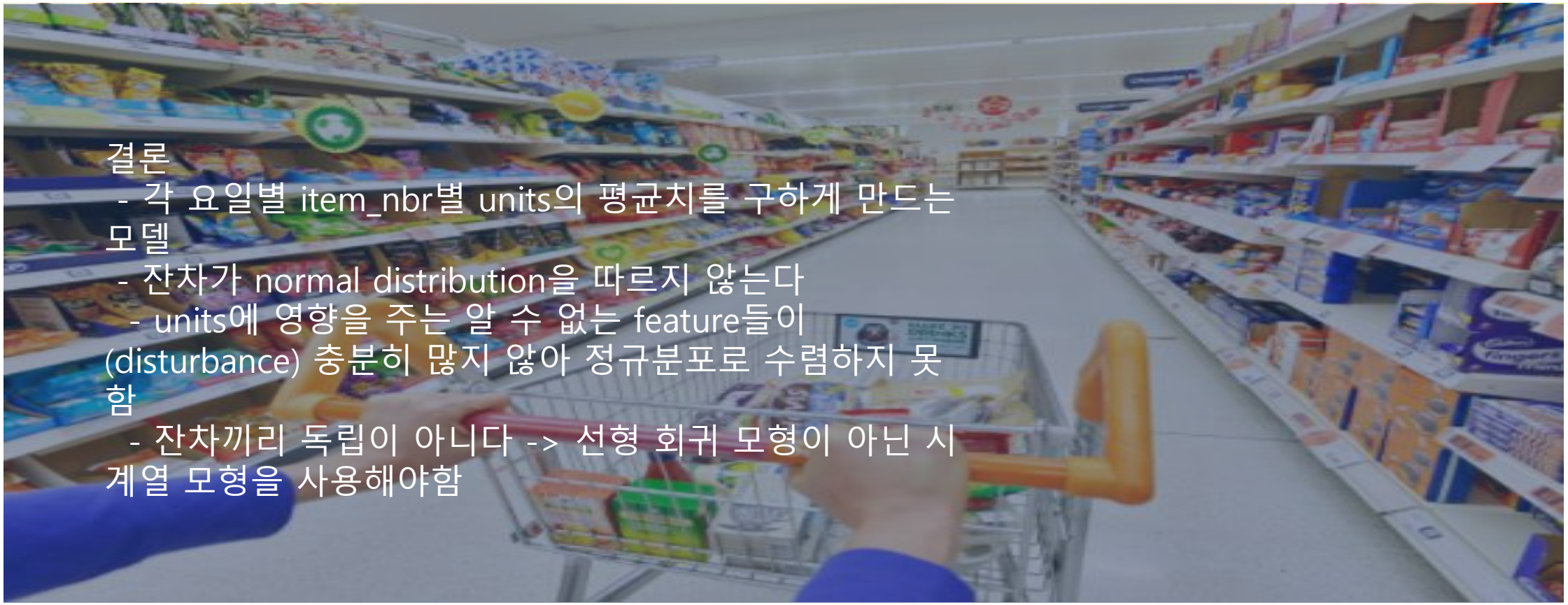


6번
log_units ~ C(item_nbr):C(weekday) + C(item_nbr):C(holiday) + 0

최종
log_units ~ C(item_nbr):C(weekday)+ 0

Walmart Recruiting II: Sales in Stormy Weather
Predict how sales of weather-sensitive products are affected by snow and rain
485 teams · 3 years ago
Overview Data Discussion Leaderboard Rules Team My Submissions Late Submission
Your most recent submission
Name Submitted Wait time Execution time Score

05. CONCLUSION



결론
- 각 요일별 item_nbr별 units의 평균치를 구하게 만드는 모델
- 잔차가 normal distribution을 따르지 않는다
- units에 영향을 주는 알 수 없는 feature들이 (disturbance) 충분히 많지 않아 정규분포로 수렴하지 못함
- 잔차끼리 독립이 아니다 -> 선형 회귀 모형이 아닌 시계열 모형을 사용해야함

06. MOVING AVERAGE