



2018 **FAST CAMPUS**

DATA SCIENCE SCHOOL PROJECT(1)

REGRESSION ANALYSIS

TEAM: FORTUNETELLER

정영조
나영민
조하닉

DATA SCIENCE SCHOOL

목차

01. INTRODUCTION

02. EDA

03. FEATURE SELECTION

04. OLS MODELING

05. CONCLUSION

06. 시계열 모형 (MOVING AVERAGE)

07. SUMMARY

1. INTRODUCTION

Walmart Recruiting II : Sales in Stormy Weather



Objective

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

Data Set

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

1. INTRODUCTION (Continued)

Walmart Recruiting II : Sales in Stormy Weather



Weather Columns

- tmax (최고 기온, °F), tmin (최저 기온, °F), tavg (평균 기온, °F), depart (30년간 평균 기온과 tavg의 차이, °F),
- dewpoint (이슬점, °F), wetbulb (습구온도, °F), heat (65°F - tavg), cool (tavg - 65°F)
- sunrise (일출 시간), sunset (일몰 시간), codesum (RA:rain, SN:snow 등 36개의 특이 날씨)
- snowfall (적설량, inches), preciptotal (강수량, inches)
- stnpressure (평균 기압, inchHg), sealevel (해수면 기압, inchHg)
- resultspeed (합성풍속, mph), resultdir (합성 풍향, 00:북, 09:동, 18:남, 27:서) , avgspeed (평균 풍속, mph)

1. INTRODUCTION (Continued)

Walmart Recruiting II : Sales in Stormy Weather



Rules

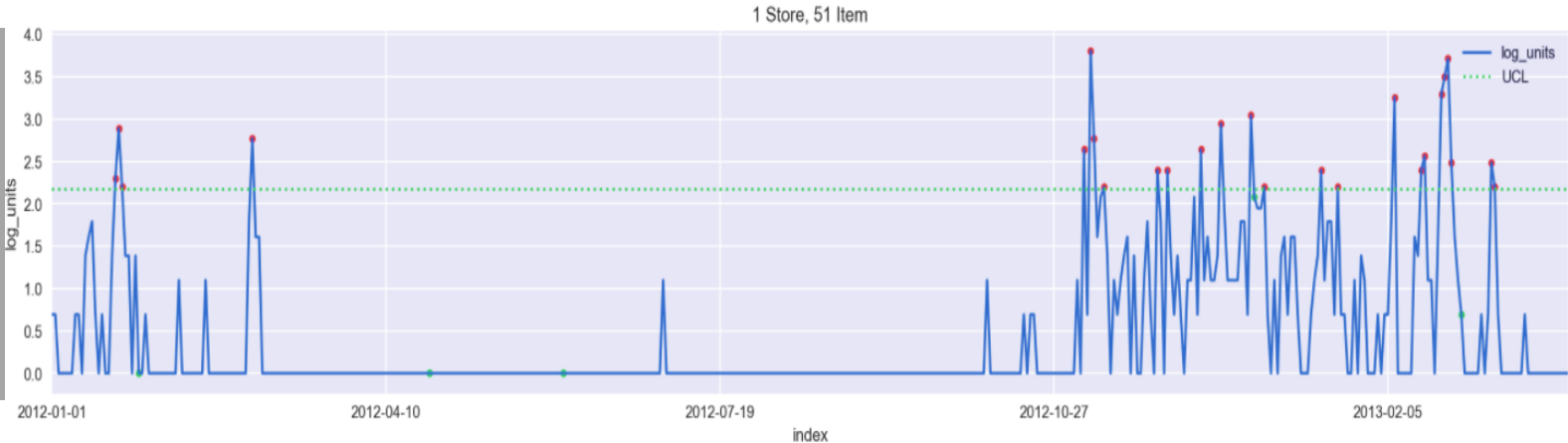
- 외부데이터 사용 금지
- Train data set에서 2013-04-01 이전 데이터를 Training data로 정의한다
- You do not need to forecast weather in addition to sales (it's as though you have a perfect weather forecast at your disposal)

Assumptions

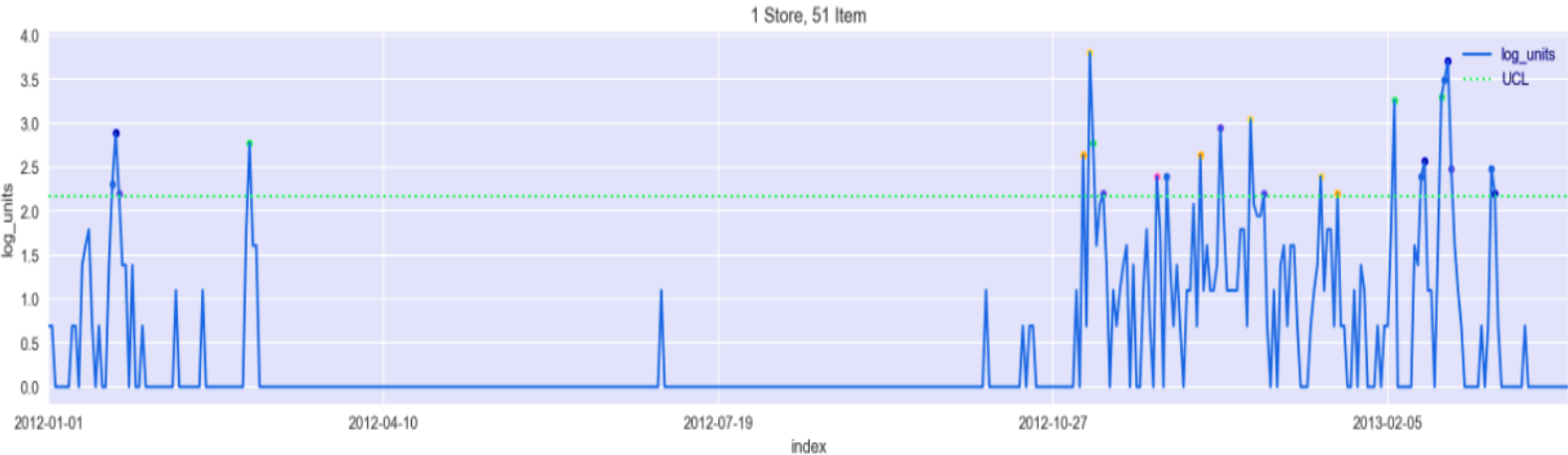
- Weather event는 문제에서 정의한 Preciptotal > 1 inch, Snowfall > 2inch 를 따른다
- Target value "units"는 독립변수(들)의 선형조합이다

2. EDA(Exploratory data analysis)

Unit vs .Date
(Mean + 2 * Sigma Highlighted)

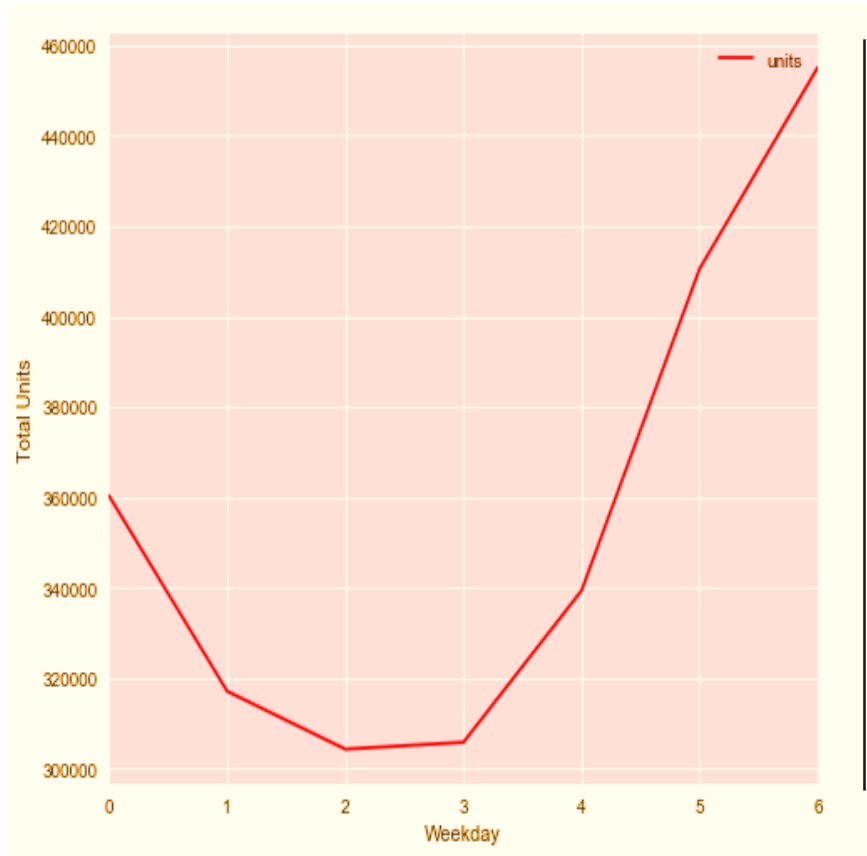


Unit vs. Date
(Weekday Highlighted)

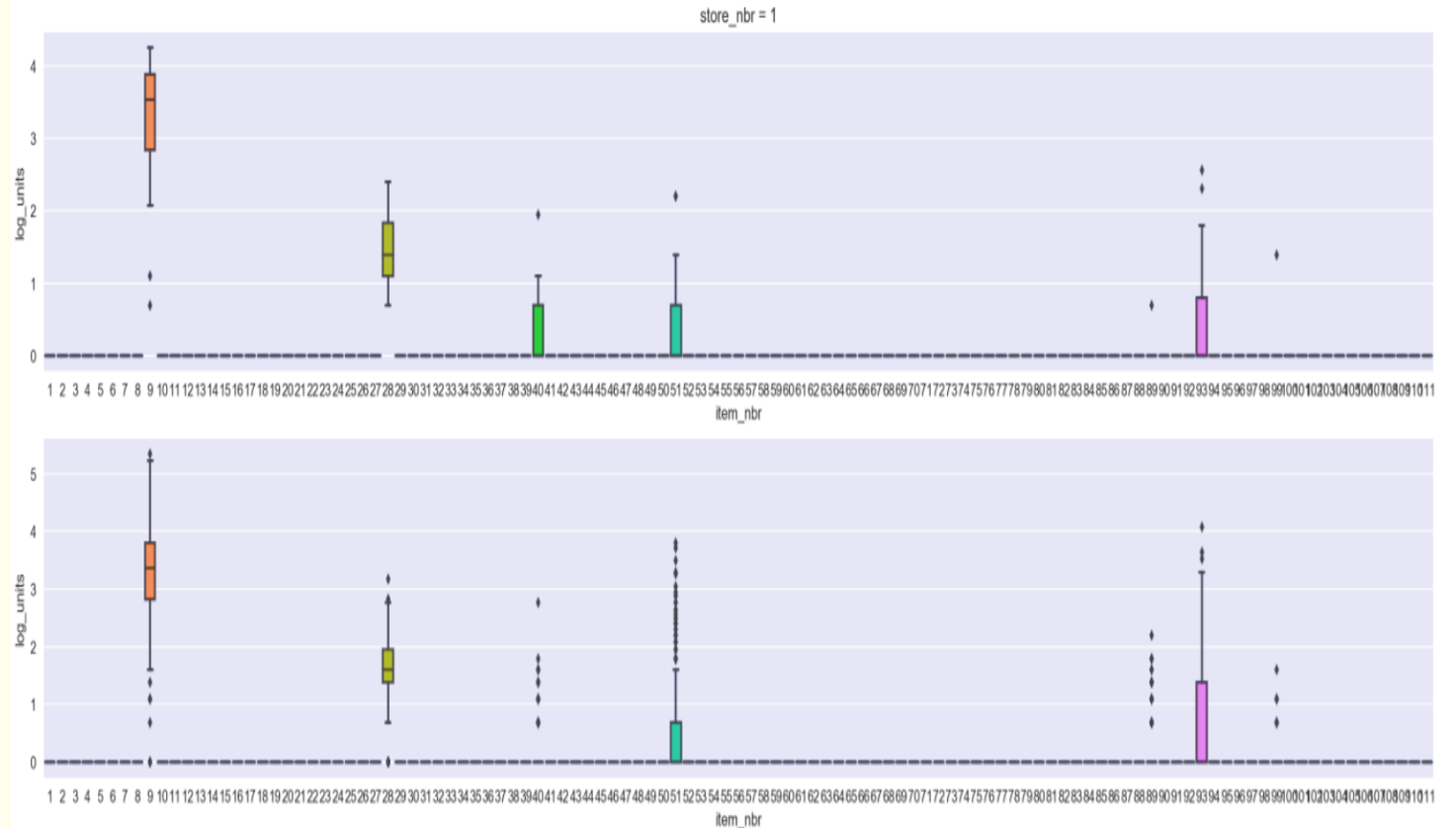


Monday: 3
Tuesday: 1
Wednesday: 3
Thursday: 4
Friday: 5
Saturday: 4
Sunday: 5

2. EDA(Exploratory data analysis)



[1] 요일 별 유닛 총 판매량

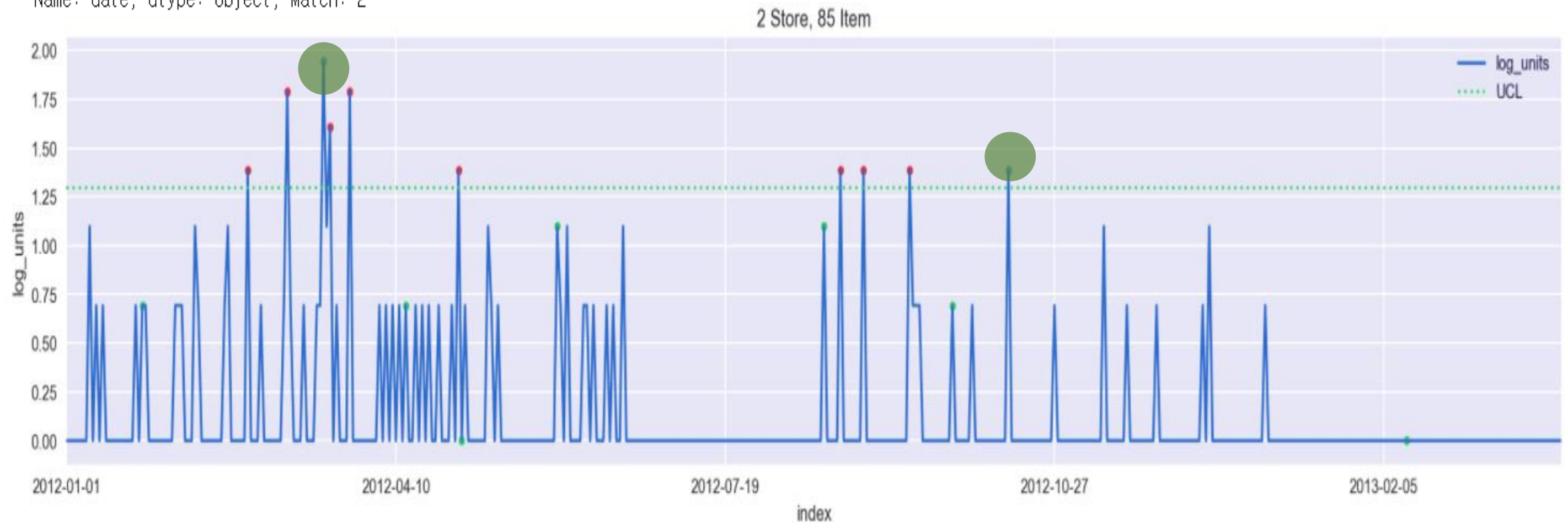


[2] Store_nbr 1의 아이템별 공휴일/비공휴일 판매량

2. EDA(Exploratory data analysis)

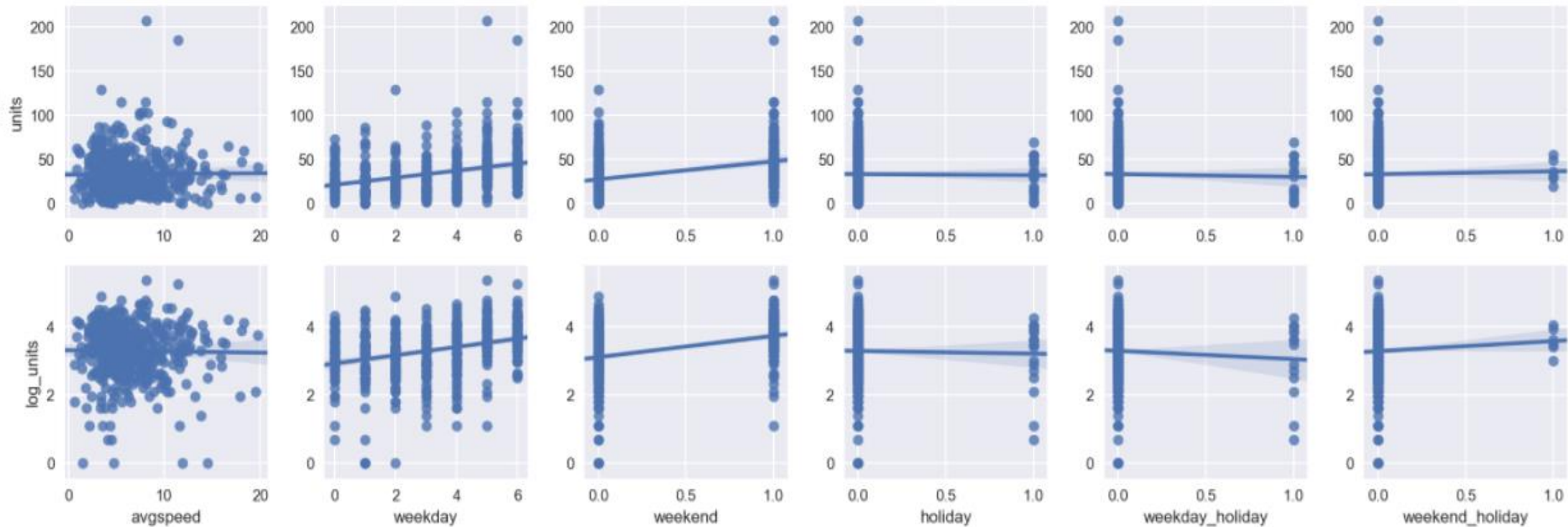
Unit vs Date (Weather Event Highlighted)

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



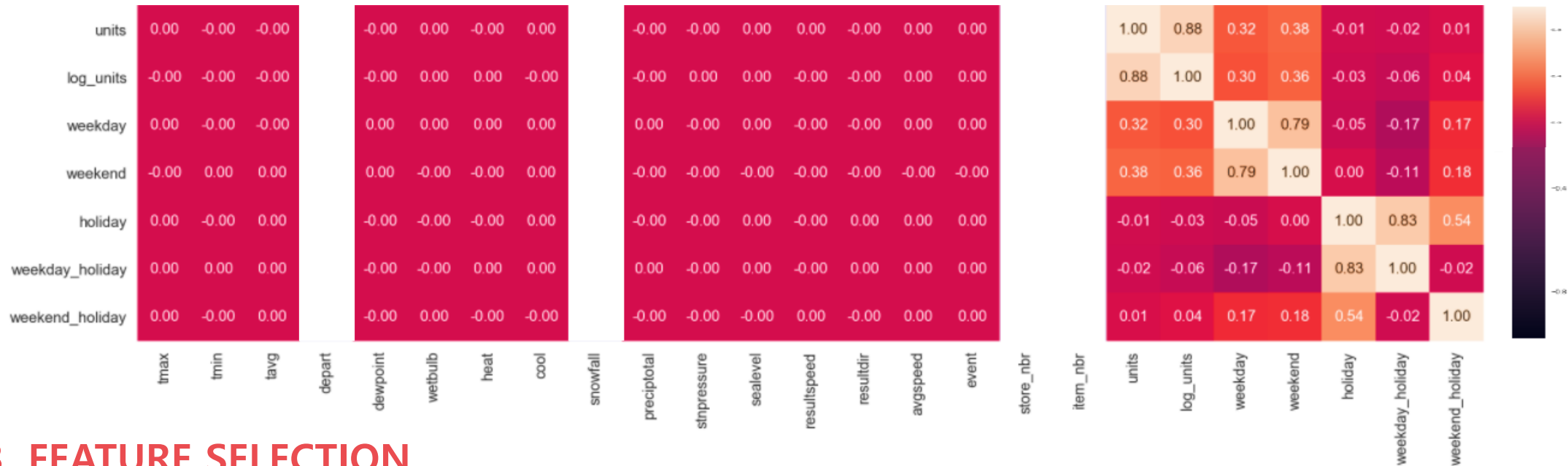
2. EDA(Exploratory data analysis)

1번 스토어 9번 아이템 Pair Plot



2. EDA(Exploratory data analysis)

1번 스토어 9번 아이템 Correlation Plot



3. FEATURE SELECTION

- 각 Store_nbr, Item_nbr별로 나누어 Modeling해야 한다
- 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)

4. OLS MODELING (Trial & Error)

Trial Test

Your most recent submission				
Name	Submitted	Wait time	Execution time	Score
sub_test.csv	a few seconds ago	3 seconds	6 seconds	1.78595
Complete				
Jump to your position on the leaderboard ▾				

Your most recent submission				
Name	Submitted	Wait time	Execution time	Score
sub_test3.csv	a few seconds ago	2 seconds	5 seconds	0.44340
Complete				
Jump to your position on the leaderboard ▾				

Your most recent submission				
Name	Submitted	Wait time	Execution time	Score
sub_test2.csv	a few seconds ago	1 seconds	5 seconds	0.44604
Complete				
Jump to your position on the leaderboard ▾				

Trial 1. 2등(뒤에서)

$\log_units \sim C(\text{store_nbr}) + C(\text{item_nbr}) + C(\text{weekday}) + C(\text{holiday}) + C(\text{event}) + 0$

Trial 2. 485명중 350등





$C(\text{store_nbr}):C(\text{item_nbr}) + C(\text{weekday}) + C(\text{holiday}) + C(\text{event}) + 0$

Trial 2-1

$C(\text{store_nbr}):C(\text{item_nbr}) + C(\text{weekday}) + C(\text{holiday}) + \text{snowfall} + \text{preciptotal} + 0$

4. OLS MODELING (Trial & Error)

Trial Test

Your most recent submission						
Name	Submitted	Wait time	Execution time	Score		
sub_test4.csv	a few seconds ago	3 seconds	5 seconds	0.15898		
Complete						
Jump to your position on the leaderboard ▼						
272	▼14	63614		0.15659	20	3y
273	▲3	81010		0.15800	69	3y
274	▼2	mnrl		0.15944	2	3y
275	▲3	sna		0.15984	8	3y

Trial 3.

$$\log_units \sim C(station_nbr):C(store_nbr):C(item_nbr) + C(weekday) + C(holiday) + C(event) + 0$$


Memory Error 발생
Station_nbr별로 나눠서 OLS실행

Trial4.

$$\log_units \sim C(store_nbr):C(item_nbr) + C(weekday) + C(holiday) + C(event) + 0$$

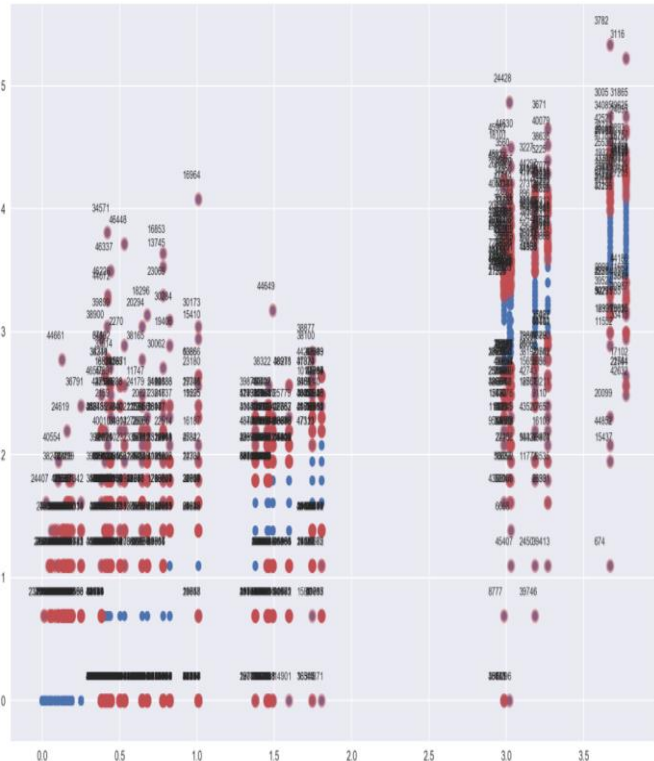
Trial5.

$$\log_units \sim C(store_nbr):C(item_nbr) + C(weekday) + C(holiday) + snowfall + preciptotal + 0$$


Store_nbr별로 나눠서 OLS실행, Event 제외
결과 좋지 않음

4. OLS MODELING (Final)

Final Model Result



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

OverviewDataDiscussionLeaderboardRulesTeam

My SubmissionsLate Submission

Your most recent submission

Name	Submitted	Wait time	Execution time	Score
sub_finally.csv	just now	2 seconds	4 seconds	0.15794
Complete				

Jump to your position on the leaderboard ▾

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





OverviewDataDiscussionLeaderboardRulesTeam

My SubmissionsLate Submission

Your most recent submission

Name	Submitted	Wait time	Execution time	Score
sub_test1.csv	18 hours ago	3 seconds	5 seconds	0.15825
Complete				

Jump to your position on the leaderboard ▾

272	▼14	63614		0.15659	20	3y
273	▲3	81010		0.15800	69	3y
274	▼2	mnrl		0.15944	2	3y
275	▲3	sna		0.15984	8	3y

Trial 6.
 $\log_units \sim C(item_nbr):C(weekday)$
 $+ C(item_nbr):C(holiday) + 0$

최종:

$\log_units \sim C(item_nbr):C(weekday)$
 $+ 0$

4. OLS MODELING (Final) (Store_nbr 1) Final Model Result Summary

OLS Regression Results							OLS Regression Results						
Dep. Variable:	log_units	R-squared:	0.842				Dep. Variable:	log_units	R-squared:	0.995			
Model:	OLS	Adj. R-squared:	0.839				Model:	OLS	Adj. R-squared:	0.995			
Method:	Least Squares	F-statistic:	340.5				Method:	Least Squares	F-statistic:	1.219e+04			
Date:	Thu, 15 Mar 2018	Prob (F-statistic):	0.00				Date:	Thu, 15 Mar 2018	Prob (F-statistic):	0.00			
Time:	18:37:53	Log-Likelihood:	23410.				Time:	18:38:27	Log-Likelihood:	1.1342e+05			
No. Observations:	50505	AIC:	-4.527e+04				No. Observations:	50505	AIC:	-2.253e+05			
Df Residuals:	49728	BIC:	-3.840e+04				Df Residuals:	49728	BIC:	-2.184e+05			
Df Model:	776						Df Model:	776					
Covariance Type:	nonrobust						Covariance Type:	nonrobust					
	coef	std err	t	P> t	[0.025	0.975]		coef	std err	t	P> t	[0.025	0.975]
C(item_nbr)[1]:C(weekday, contrast_weekday)[custom0]	0	0.019	0	1.000	-0.037	0.037	C(item_nbr)[1]:C(weekday, contrast_weekday)[custom0]	0	0.003	0	1.000	-0.006	0.006
C(item_nbr)[2]:C(weekday, contrast_weekday)[custom0]	0	0.019	0	1.000	-0.037	0.037	C(item_nbr)[2]:C(weekday, contrast_weekday)[custom0]	0	0.003	0	1.000	-0.006	0.006
C(item_nbr)[3]:C(weekday, contrast_weekday)[custom0]	0	0.019	0	1.000	-0.037	0.037	C(item_nbr)[3]:C(weekday, contrast_weekday)[custom0]	0	0.003	0	1.000	-0.006	0.006
	●							●					
	●							●					
	●							●					
C(item_nbr)[109]:C(weekday, contrast_weekday)[custom6]	0	0.019	0	1.000	-0.037	0.037	C(item_nbr)[109]:C(weekday, contrast_weekday)[custom6]	0	0.003	0	1.000	-0.006	0.006
C(item_nbr)[110]:C(weekday, contrast_weekday)[custom6]	0	0.019	0	1.000	-0.037	0.037	C(item_nbr)[110]:C(weekday, contrast_weekday)[custom6]	0	0.003	0	1.000	-0.006	0.006
C(item_nbr)[111]:C(weekday, contrast_weekday)[custom6]	0	0.019	0	1.000	-0.037	0.037	C(item_nbr)[111]:C(weekday, contrast_weekday)[custom6]	0	0.003	0	1.000	-0.006	0.006
Omnibus:	46753.784	Durbin-Watson:	2.000				Omnibus:	33790.919	Durbin-Watson:	2.000			
Prob(Omnibus):	0.000	Jarque-Bera (JB):	23826967.194				Prob(Omnibus):	0.000	Jarque-Bera (JB):	39853352.566			
Skew:	3.588	Prob(JB):	0.00				Skew:	1.767	Prob(JB):	0.00			
Kurtosis:	109.165	Cond. No.	1.02				Kurtosis:	140.571	Cond. No.	1.02			

DATA SCIENCE SCHOOL

5. CONCLUSION


[3] 1번 Store Item_nbr 51



DATA SCIENCE SCHOOL

- 각 요일, Item_nbr별 Units의 평균치를 구하는 모델
- 잔차분포가 Non-normal함
 - 요인)
 - 1) Units에 영향을 주는 알 수 없는 Feature들이(Disturbance)충분히 많지 않아 정규분포로 수렴하지 못함
 - 2) 잔차끼리 독립이 아님(선형 회귀 모형이 아닌 시계열 모형을 사용하는 것이 적합해 보임)

6. 시계열 모형 (MOVING AVERAGE)

**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






OverviewDataDiscussionLeaderboardRulesTeamMy SubmissionsLate Submission

Your most recent submission

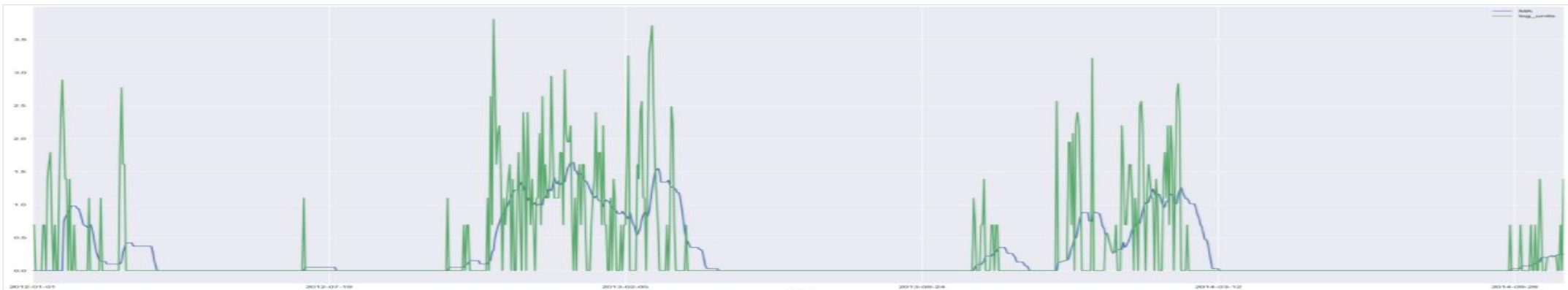
Name	Submitted	Wait time	Execution time	Score
ma.csv	a few seconds ago	5 seconds	4 seconds	0.10568

Complete

Jump to your position on the leaderboard ▾

96	▲6	35957		0.10521	26	3y
97	▲13	anaef		0.10530	7	3y
98	▲4	Victor Mayrink		0.10553	37	3y
99	▲1	TheAnalyticProphet		0.10570	23	3y
100	▲3	99247		0.10586	11	3y

- 1번 Store의 51번 Item의 MA를 이용한 Fit 그래프



7. SUMMERY

- 종속변수 Units는 독립변수 Store_nbr, Item_nbr, Weekday, Holiday, Event의 선형조합으로 결정되는 기대 값과 고정된 분산을 가지는 정규분포를 따른다 라고 가정함
- 각 스토어의 아이템별 Units vs. Date 플롯은 판매량이 Random함을 확인
- 그 Random함이 Weekday, Holiday, Event에 영향을 받은 것인지 알아보기 위해 다수의 OLS를 시행함
- Holiday와 Event의 영향이 없다고 판단, Item_nbr와 Weekday의 Interaction을 모수로 사용하는 OLS를 최종 모델로 채택
- 최종 모델에서 잔차의 분포가 정규분포를 따르지 않음을 확인
- 그 이유는 잔차 간의 독립이 성립되지 않기 때문이라고 생각
- 잔차 간 독립이 성립하지 않음은 종속변수 간의 관계가 있음을 나타내고 따라서 시계열 모형 (이동평균선)을 이용하여 종속변수를 재추정하였고, 회귀모형보다 좋은 결과 값을 확인

Thank You
