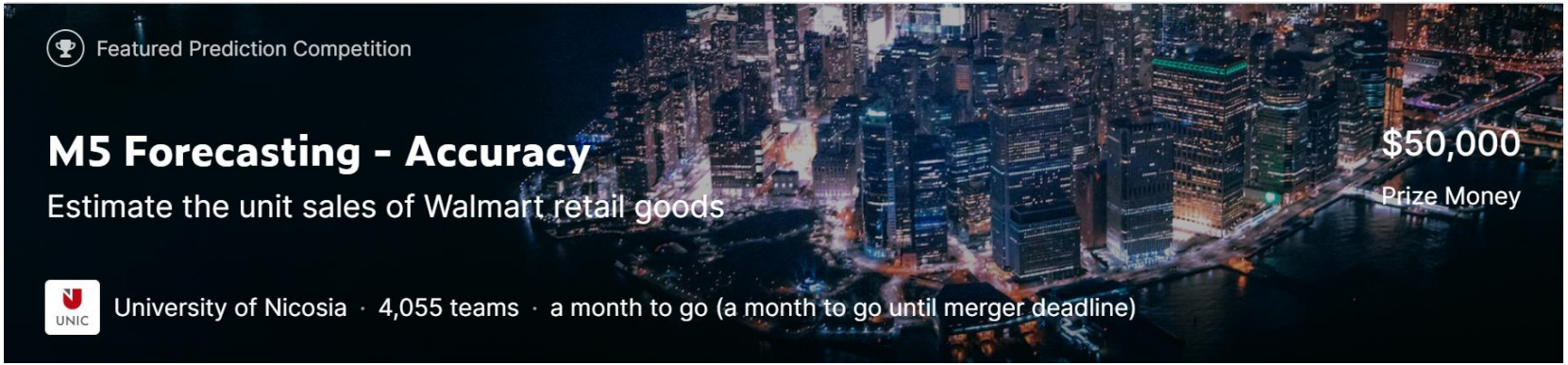
M5 Forecasting
- Accuracy

현예성 김민석 최성웅
강호석 문구영 이노아

KU-BIG

Contents



Files

- `calendar.csv` - Contains information about the dates on which the products are sold.
- `sales_train_validation.csv` - Contains the historical daily unit sales data per product and store `[d_1 - d_1913]`
- `sample_submission.csv` - The correct format for submissions. Reference the Evaluation tab for more info.
- `sell_prices.csv` - Contains information about the price of the products sold per store and date.
- *`sales_train_evaluation.csv` - Available once month before competition deadline. Will include sales `[d_1 - d_1941]`*

KU-BIG

Data - sales_train_validation

	id	item_id	dept_id	cat_id	store_id	state_id	d_1	d_2	d_3	…	d_1908	d_1909	d_1910	d_1911	d_1912	d_1913
1																
2																
3																
…																
30489																
30490																

ID information

Sales information

Data - calendar

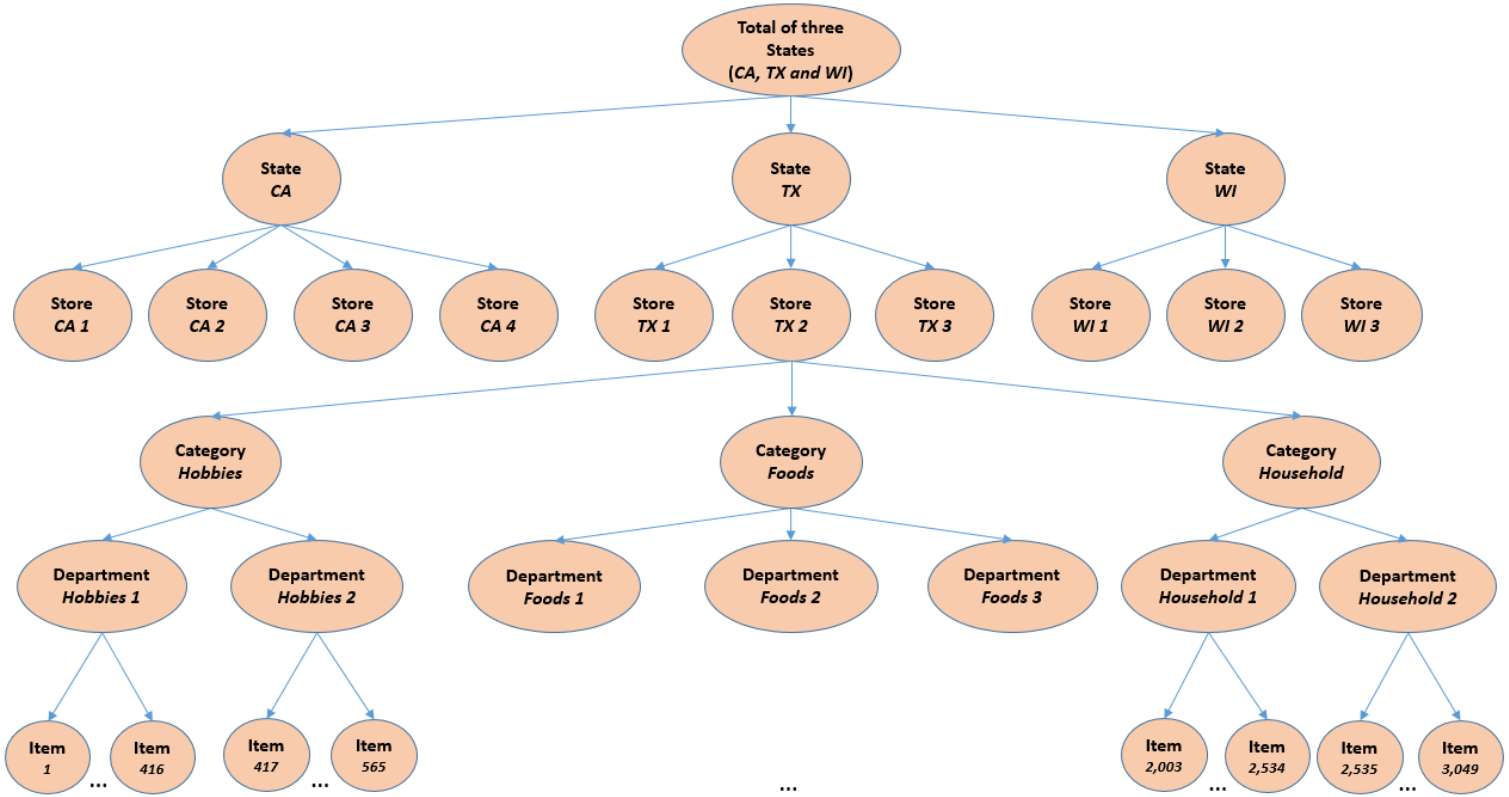
date	wm_yr_wk	weekday	wday	month	year	d	event_nam	event_type	event_nam	event_type	snap_CA	snap_TX	snap_WI
2011-01-29	11101	Saturday	1	1	2011	d_1					0	0	0
2011-01-30	11101	Sunday	2	1	2011	d_2					0	0	0
2011-01-31	11101	Monday	3	1	2011	d_3					0	0	0
2011-02-01	11101	Tuesday	4	2	2011	d_4					1	1	0
2011-02-02	11101	Wednesda	5	2	2011	d_5					1	0	1
2011-02-03	11101	Thursday	6	2	2011	d_6					1	1	1
2011-02-04	11101	Friday	7	2	2011	d_7					1	0	0
2011-02-05	11102	Saturday	1	2	2011	d_8					1	1	1
2011-02-06	11102	Sunday	2	2	2011	d_9	SuperBowl	Sporting			1	1	1
2011-02-07	11102	Monday	3	2	2011	d_10					1	1	0
2011-02-08	11102	Tuesday	4	2	2011	d_11					1	0	1
2011-02-09	11102	Wednesda	5	2	2011	d_12					1	1	1
2011-02-10	11102	Thursday	6	2	2011	d_13					1	0	0
2011-02-11	11102	Friday	7	2	2011	d_14					0	1	1
2011-02-12	11103	Saturday	1	2	2011	d_15					0	1	1

Sales_train_validition 데이터의 열 이름(d_1 ~ d_1913)의 **날짜정보**

Data - sell_prices

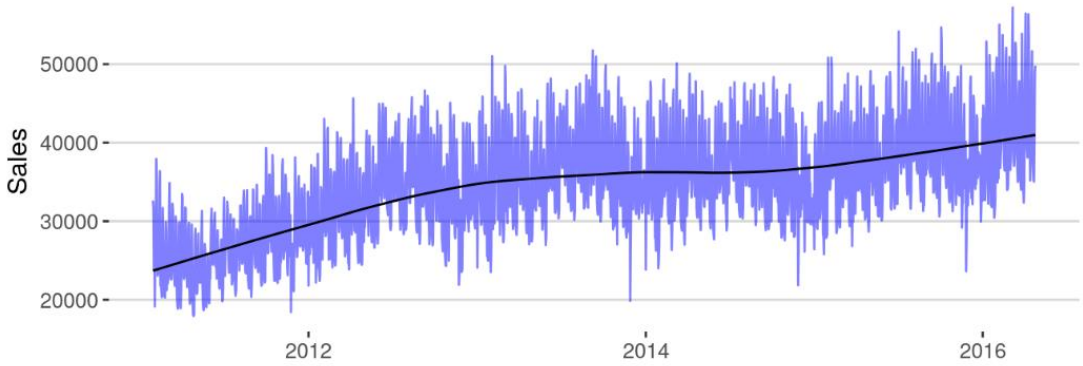
store_id	item_id	wm_yr_wk	sell_price
CA_1	HOBBIES_	11325	9.58
CA_1	HOBBIES_	11326	9.58
CA_1	HOBBIES_	11327	8.26
CA_1	HOBBIES_	11328	8.26
CA_1	HOBBIES_	11329	8.26
CA_1	HOBBIES_	11330	8.26
CA_1	HOBBIES_	11331	8.26
CA_1	HOBBIES_	11332	8.26
CA_1	HOBBIES_	11333	8.26
CA_1	HOBBIES_	11334	8.26
CA_1	HOBBIES_	11335	8.26
CA_1	HOBBIES_	11336	8.26
CA_1	HOBBIES_	11337	8.26
CA_1	HOBBIES_	11338	8.26
CA_1	HOBBIES_	11339	8.26
CA_1	HOBBIES_	11340	8.26
CA_1	HOBBIES_	11341	8.26

Sales_train_validition 데이터의 각 행들의 **판매가 정보**


Total of three States (CA, TX and WI)
State CA
State TX
State WI
Store CA 1
Store CA 2
Store CA 3
Store CA 4
Store TX 1
Store TX 2
Store TX 3
Store WI 1
Store WI 2
Store WI 3
Category Hobbies
Category Foods
Category Household
Department Hobbies 1
Department Hobbies 2
Department Foods 1
Department Foods 2
Department Foods 3
Department Household 1
Department Household 2
Item 1
Item 416
Item 417
Item 565
Item 2,003
Item 2,534
Item 2,535
Item 3,049
...

EDA - sales (All)

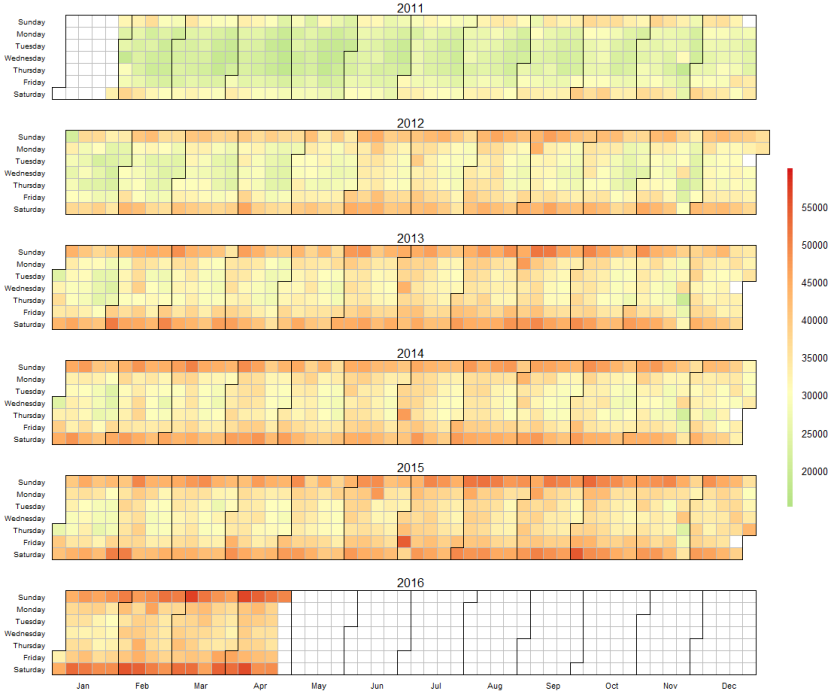
daily sales



-> yearly seasonality 존재

EDA - sales (All)

daily sales



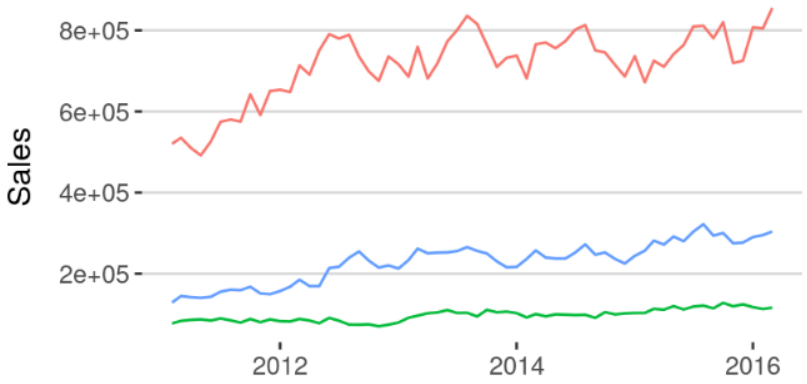
-> weekly seasonlity 존재

(state, store 별로 보았을 때도

비슷한 패턴을 보임)

EDA - sales (per category)

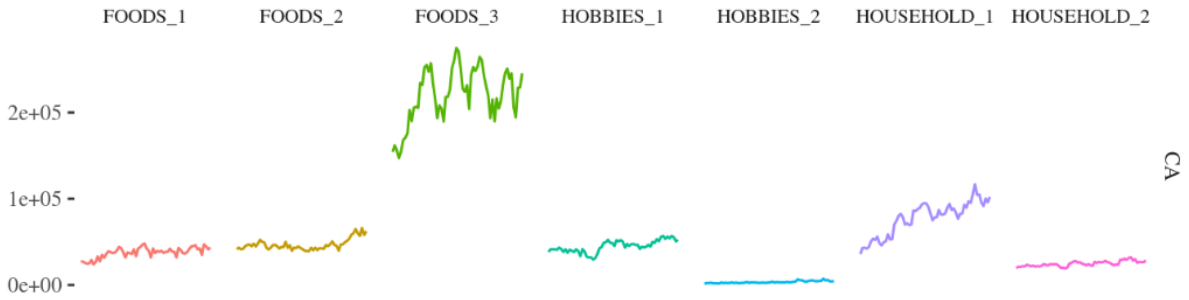
monthly sales



-> category 별로 차이 존재

EDA - sales (per deparment, in state CA)

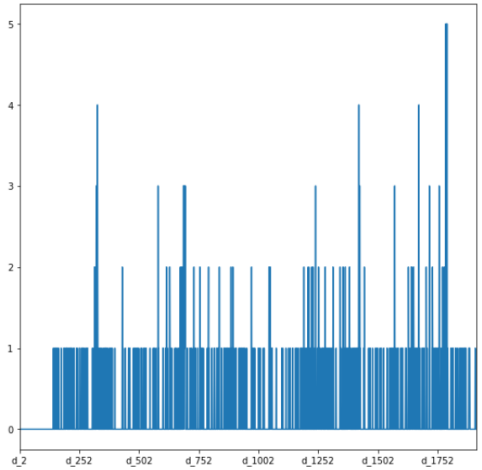
daily sales

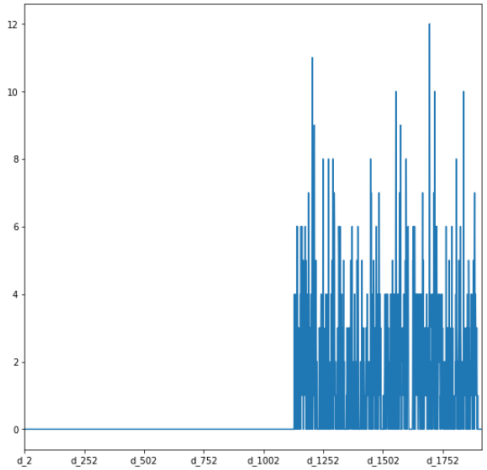


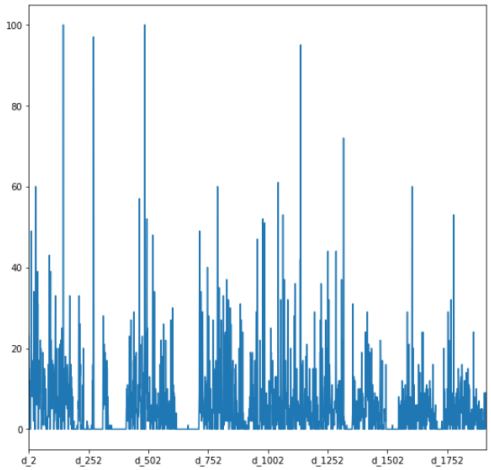
-> dep 별로 차이가 있어보임

EDA - sales (individual)

daily sales

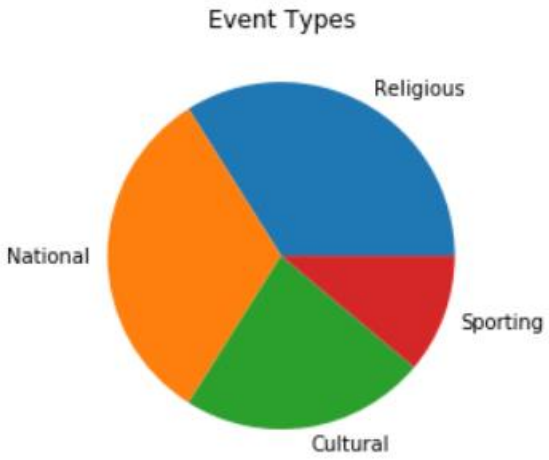






-> 아웃라이어들을 설명할 alternative data(재난 데이터) 도입 or calendar에서 event 추가

EDA - calendar



-> 스포츠, 문화 관련 event 더 조사해서 채워 넣어야 할 듯

Model1 - RNN based model

We can process a sequence of vectors **x** by applying a **recurrence formula** at every time step:

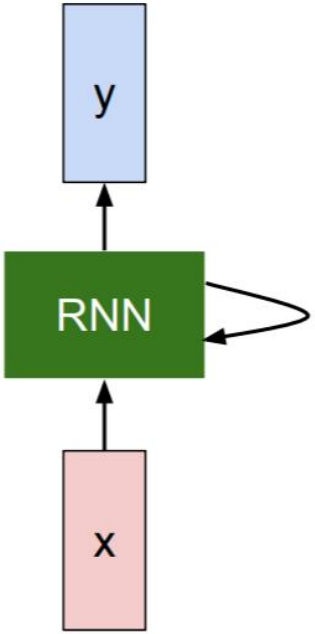
$$\boxed{h_t} = \boxed{f_W}(\boxed{h_{t-1}}, \boxed{x_t})$$

new state

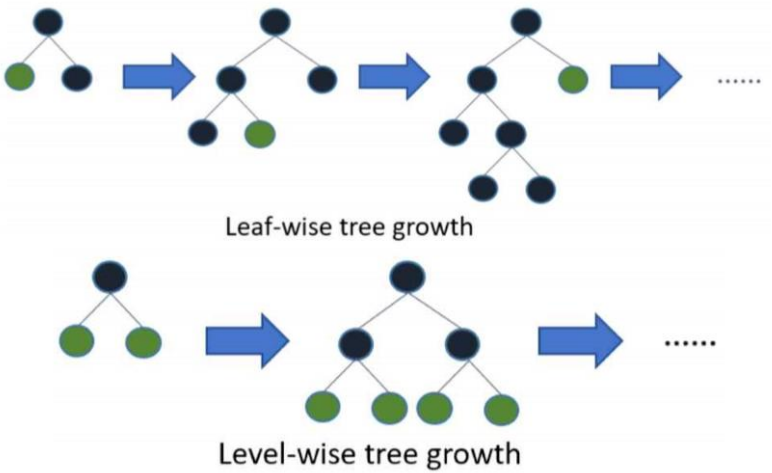
old state

input vector at some time step

some function with parameters W



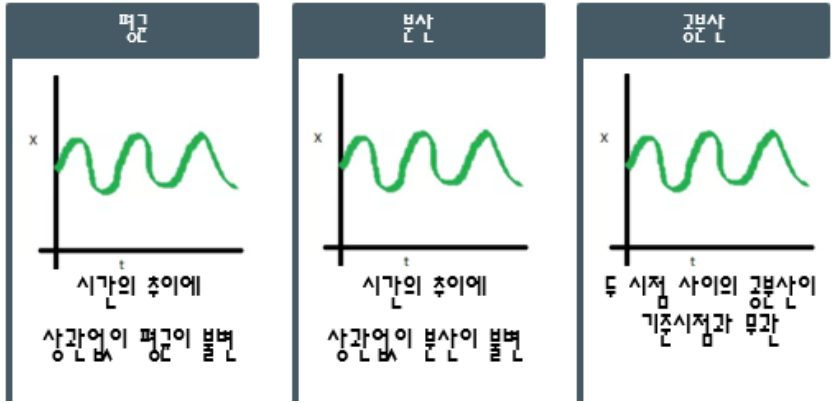
Model2 - LGBM



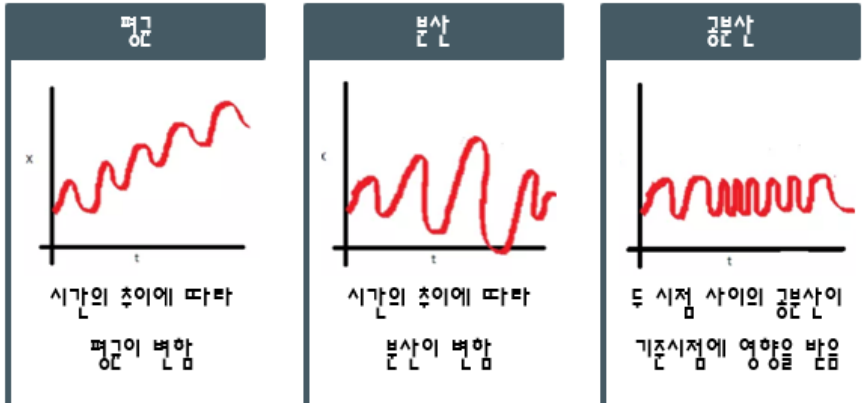
- 학습 속도가 느린 XGBoost의 단점을 보안
- 대용량 데이터 처리가 가능
- 타 모델에 비해 적은 메모리 사용
- 적은 수의 데이터 사용시 과적합 문제 발 생 가능
- 정보의 손실을 줄일 수 있음

Time Series

Stationary Series



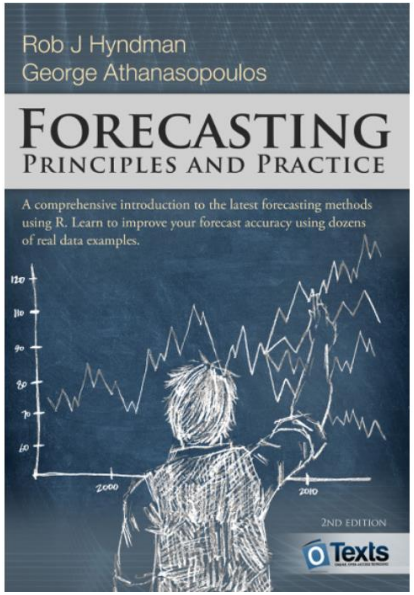
Non-Stationary Series

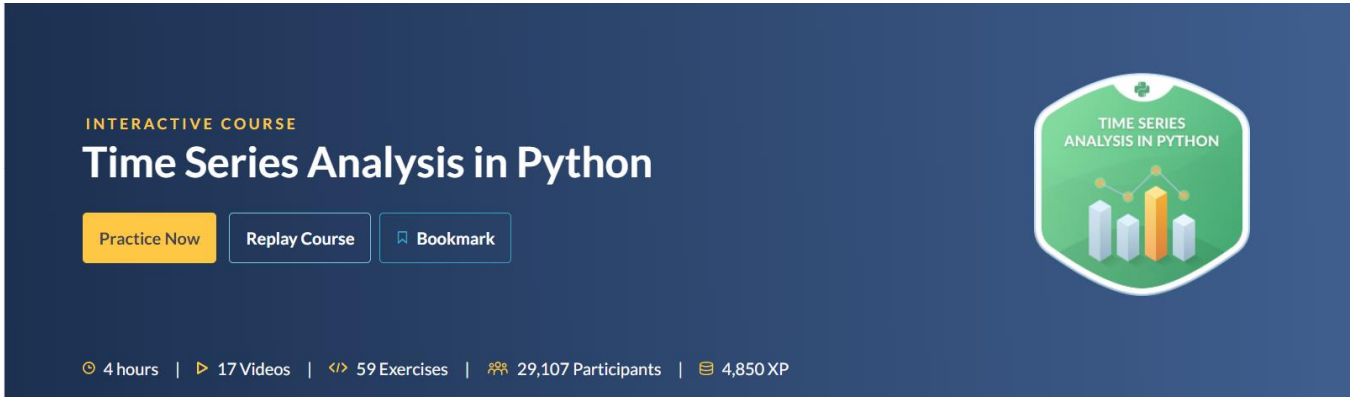


non-stationary series를 stationary serise로 바꿔야함 (왜? 어떻게?)

Future Plan

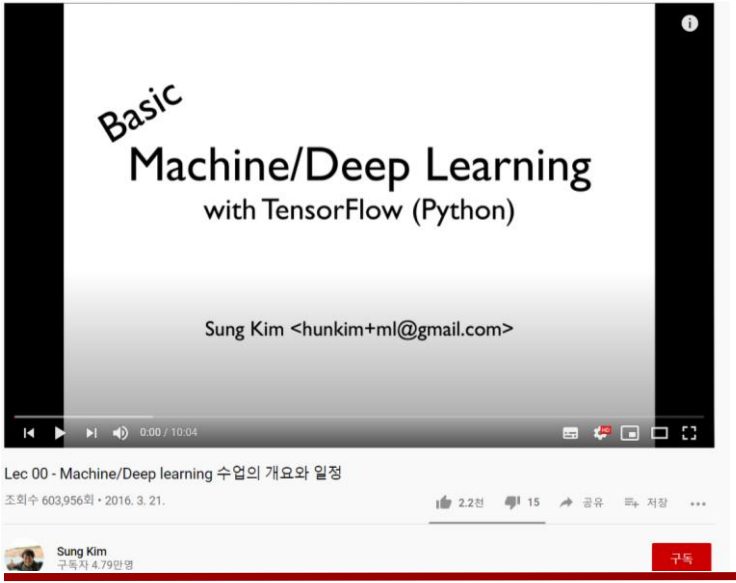
1. Time series study





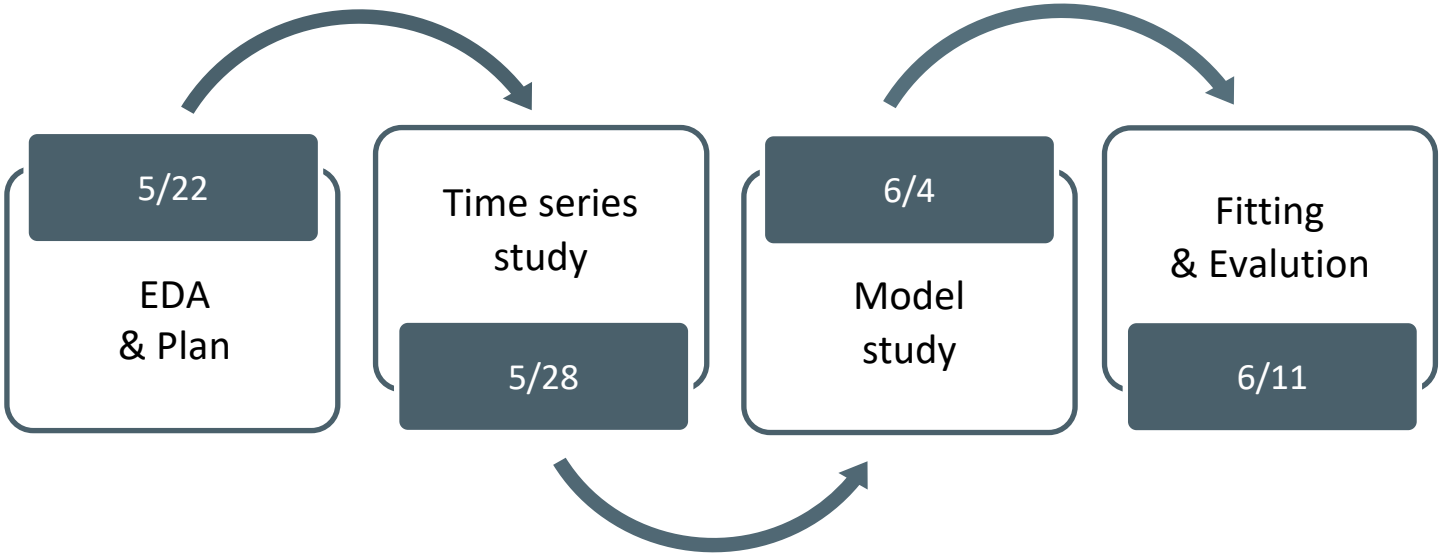
Future Plan

2. Model study



LightGBM: A Highly Efficient Gradient Boosting Decision Tree

Future Plan


5/22
EDA
& Plan
Time series
study
5/28
6/4
Model
study
Fitting
& Evalution
6/11

The end.