K 머신러닝엔지니어

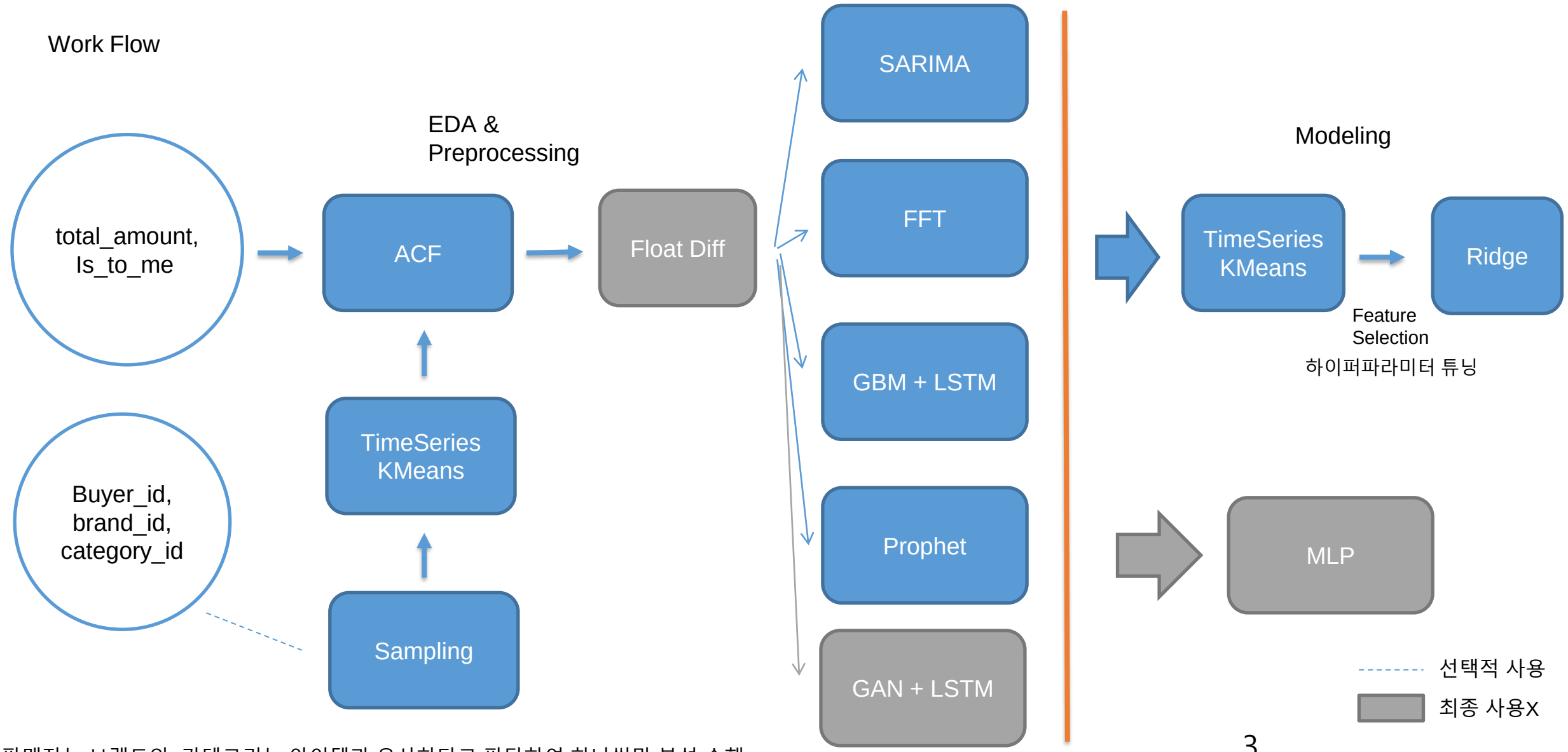
사전테스트 과제

- 정진우(케글아이디 StarWaz88)

목차

구분
I. Work Flow
Ⅱ. EDA & Preprocessing
Ⅲ. 모델링

I. Work Flow



판매자는 브랜드와, 카테고리는 아이템과 유사하다고 판단하여 하나씩만 분석 수행

Ⅱ. EDA

Ⅱ. EDA & Preprocessing

기본 EDA

```
df.isna().sum()
```

```
seller_id       0
buyer_id        0
brand_id        0
category_id     0
item_id         0
quantity        0
total_amount    0
is_to_me        0
paid_at         0
day             0
dtype: int64
```

Nan값 존재하지 않음

```
df.dtypes
```

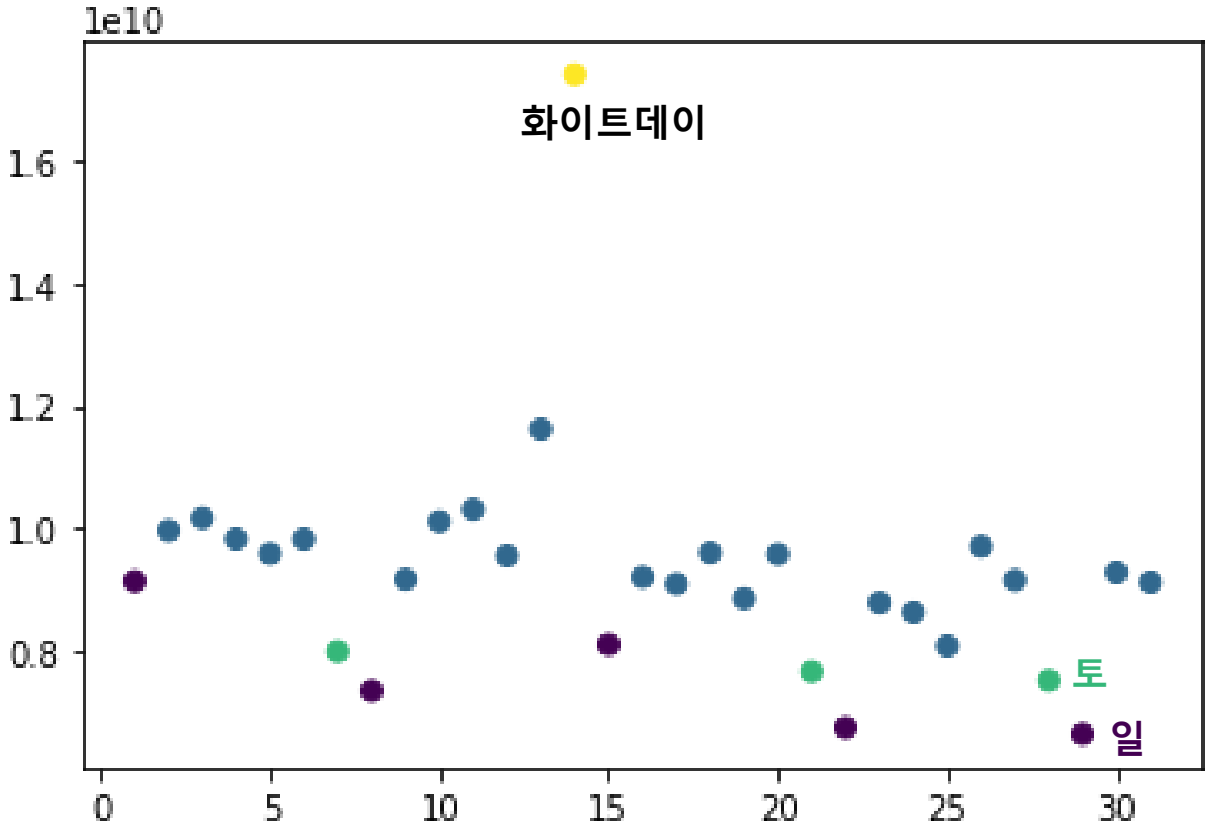
```
seller_id       int64
buyer_id        int64
brand_id        int64
category_id     int64
item_id         int64
quantity        int64
total_amount    int64
is_to_me         bool
paid_at        object
day            object
dtype: object
```

사실상 범주형 데이터

일별 거래합 생성 -> 예측해야할 값

Ⅱ. EDA & Preprocessing

일별 거래량 그래프 확인

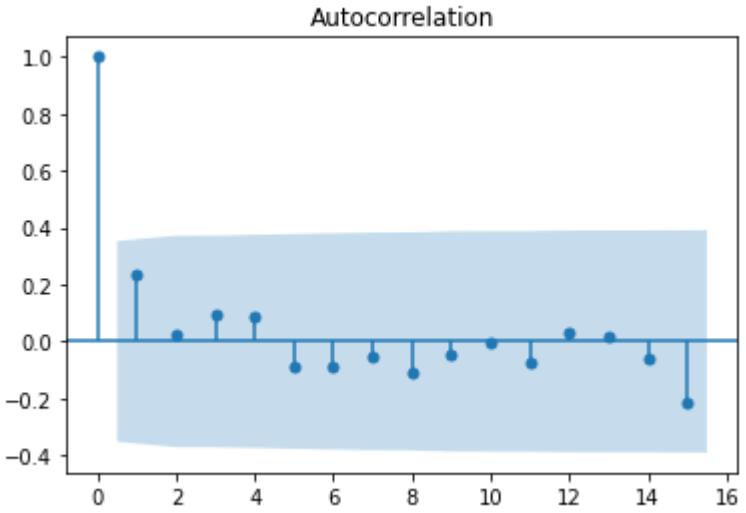


거래량은 주로 일요일에 감소하는 주기성을 띕니다.
하지만 화이트데이에 거래량이 급증했습니다.
ACF(자기상관계수)를 통해 통계적으로 분석해
보겠습니다.

Ⅱ. EDA & Preprocessing

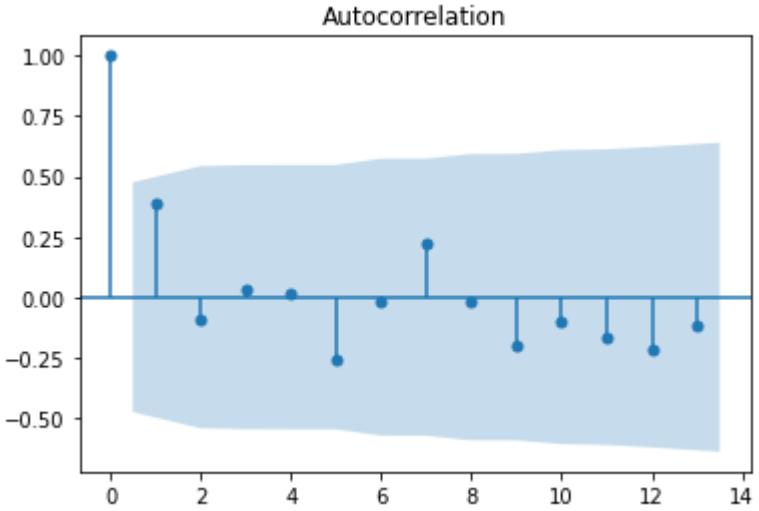
ACF 분석

화이트데이가 포함된 ACF



y_{t+7} 의 값이 작다 -> 주기성이 관찰되지 않음

화이트데이를 전후 2주를 제거했을때 ACF

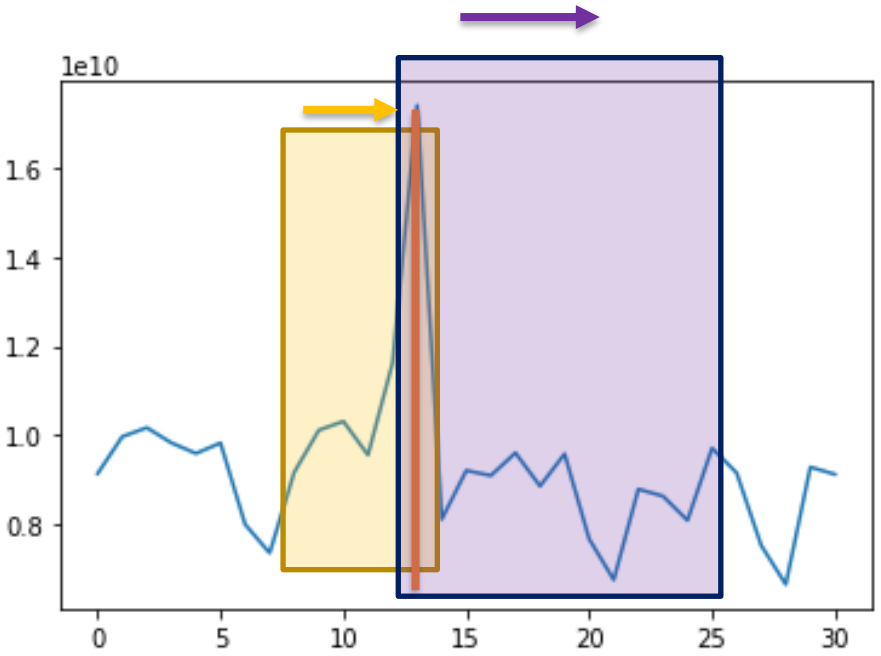


y_{t+7} 의 값이 크다 -> 주기성이 잘 관찰됨

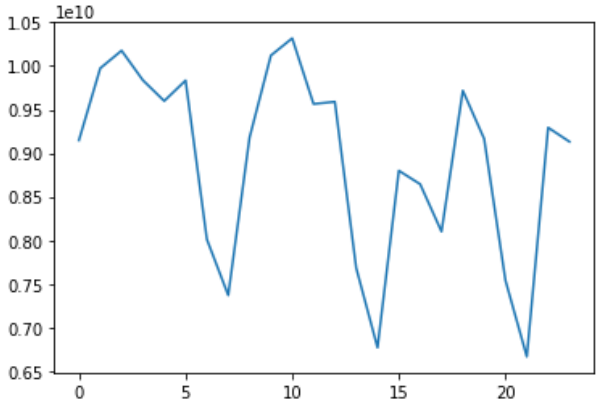
규칙이 있기에
예측하기가 쉬움

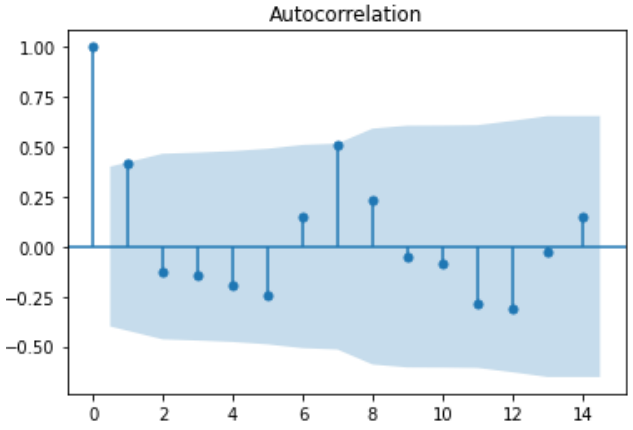
Ⅱ. EDA & Preprocessing

ACF 분석



화이트데이를 포함한 1주, 2주를 제거하며
ACF가 최대가 되는 구간 탐색

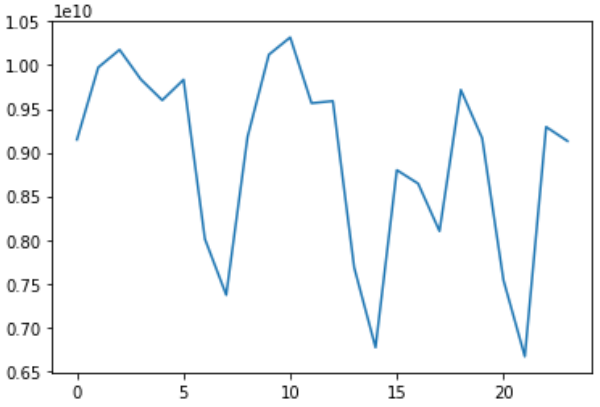




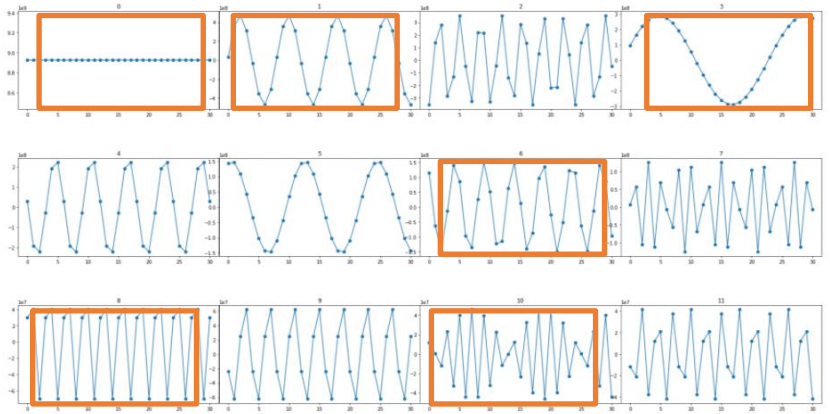
화이트데이의
영향이 최소화된
구간을 찾을 수 있음

Ⅱ. EDA & Preprocessing

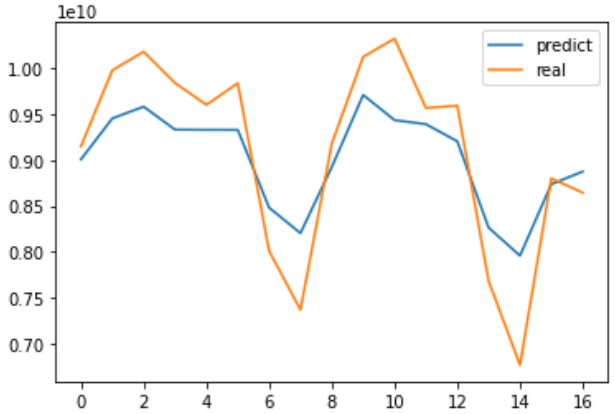
FFT(푸리에 변환)



ACF를 통해 전처리된 구간



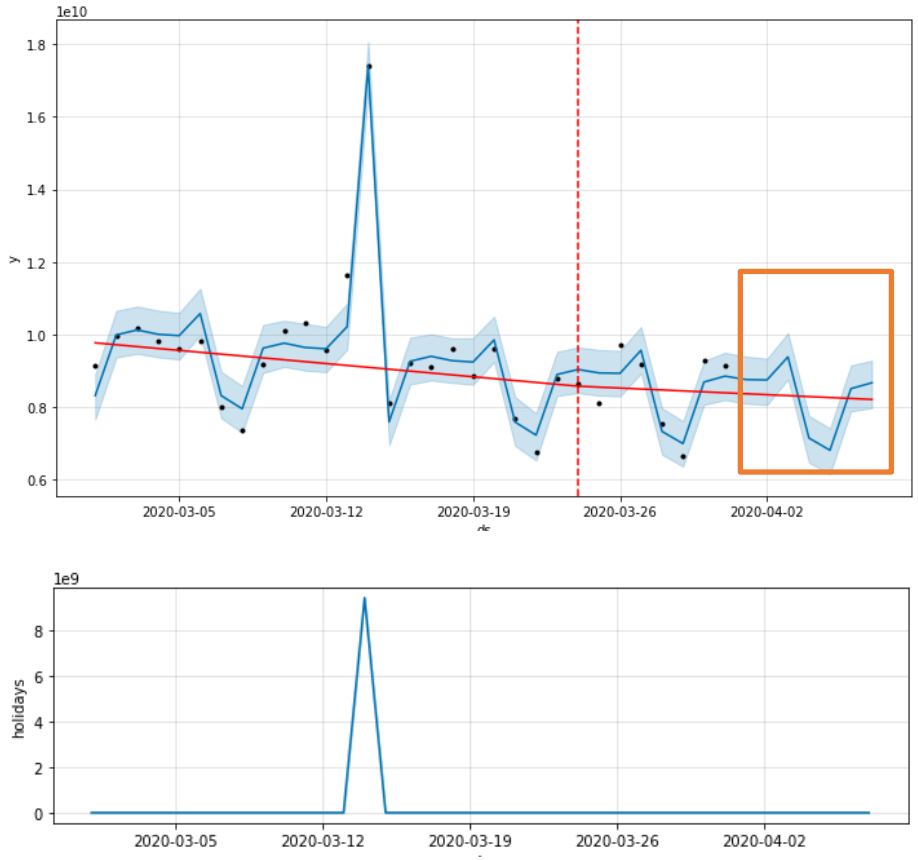
FFT를 통한 성분 분해 -> 선택 결합



미래구간 예측 가능

Ⅱ. EDA & Preprocessing

Prophet

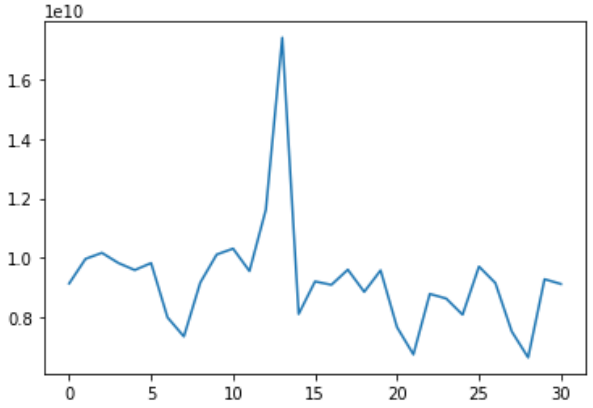


화이트데이 기간 지정

→ Feature로 사용

Ⅱ. EDA & Preprocessing

GBM + LSTM
일반 LSTM모델은 데이터가 부족하여(30) 학습이 잘 되지 않았습니다.
GBM(Geometric Brown Motion: 기하브라운운동) 몬테카를로 시뮬레이션을 통해 가상의 데이터를 생성하고
원본데이터와의 ACF, 절대값에 대한 MAPE가 낮은 데이터를 이용하여 신경망을 Meta-learning합니다.

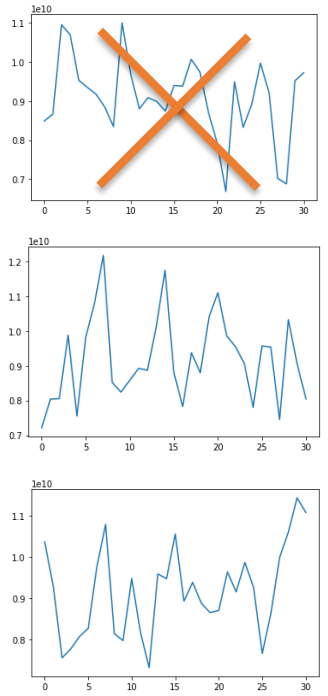


원본 데이터

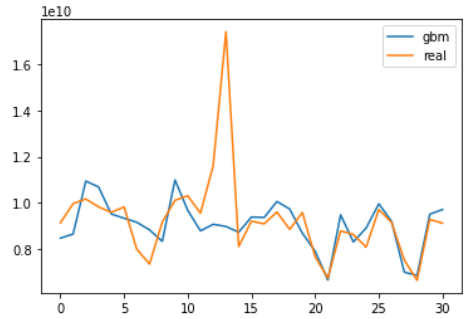
$$S_t = S_0 e^{\left(\mu - \frac{1}{2}\sigma^2\right)t + \epsilon}$$

μ: 일별 증가율의 평균
σ: 일별 증가율의 분산

GBM(Geometric Brown Motion:
기하브라운운동)



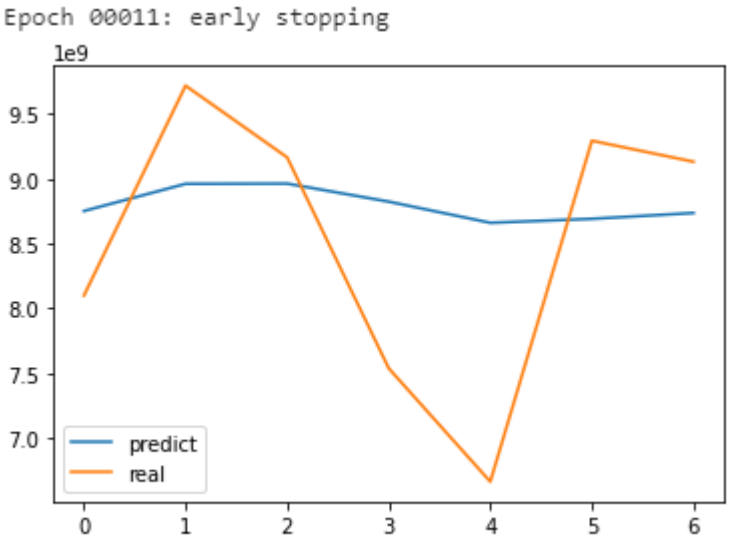
데이터 생성
>Loss로 필터



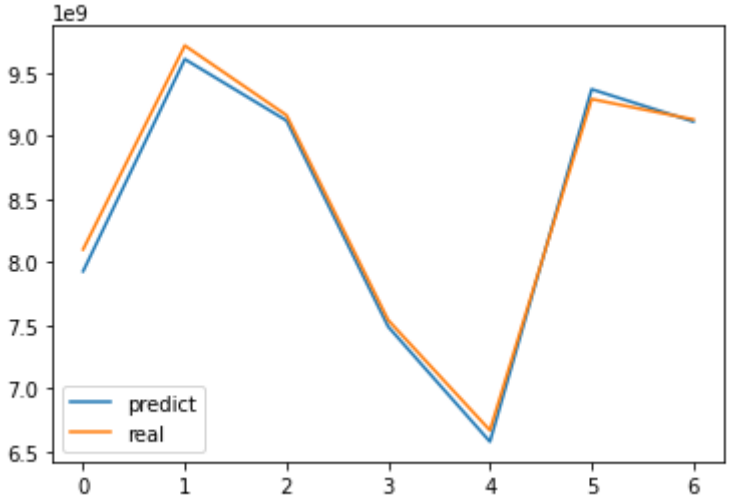
화이트데이 기간을 제외한
구간이 원본과 유사함

Ⅱ. EDA & Preprocessing

GBM + LSTM



가상 데이터만으로 학습했을때



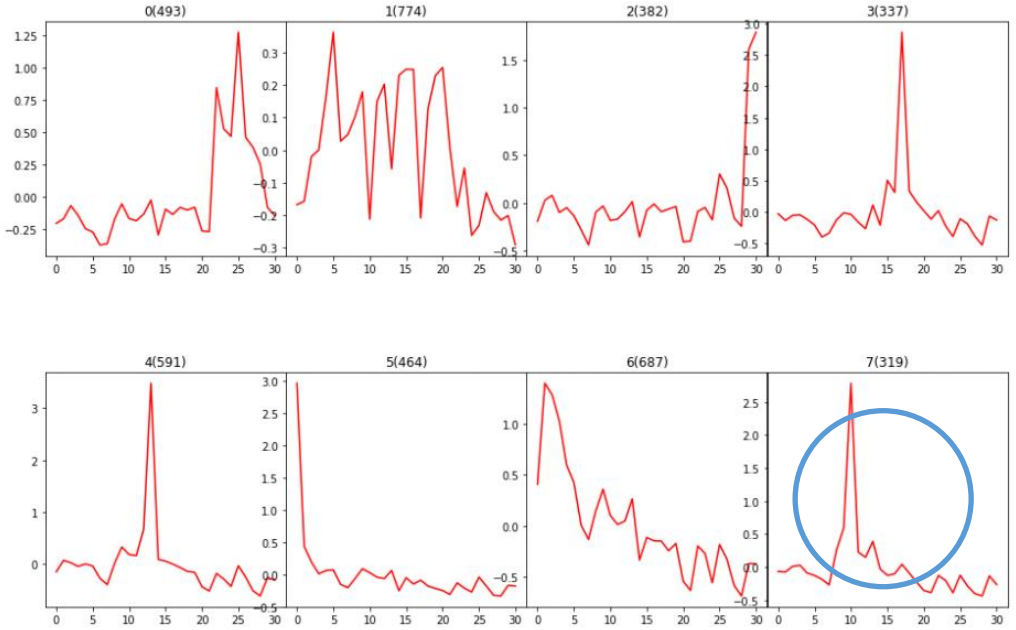
가상데이터 학습 후 실제데이터로 추가학습

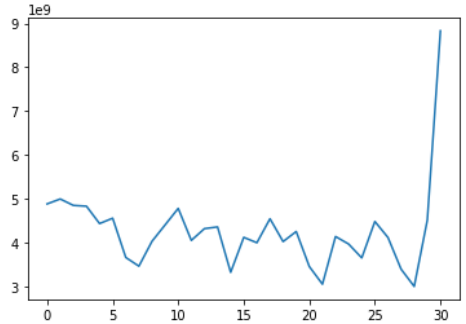
Ⅱ. EDA & Preprocessing

TimeSeries KMeans

buyer_id,
brand_id,
categorical_id

TimeSeries
KMeans



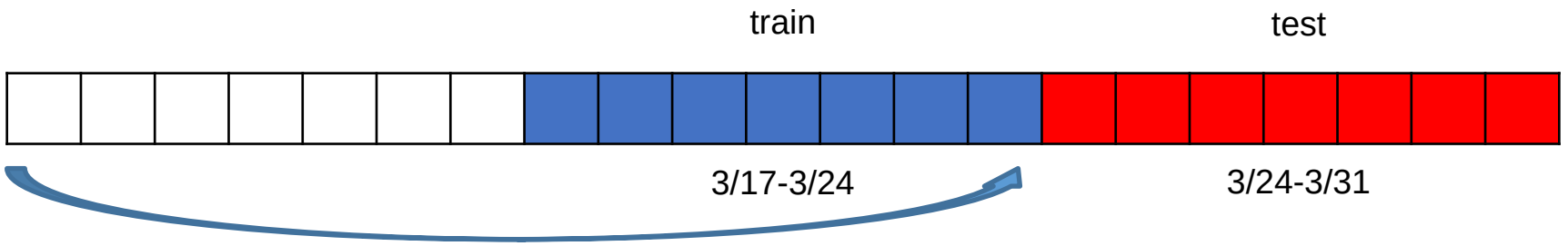


고유값들의 일 평균 시계열을 구함

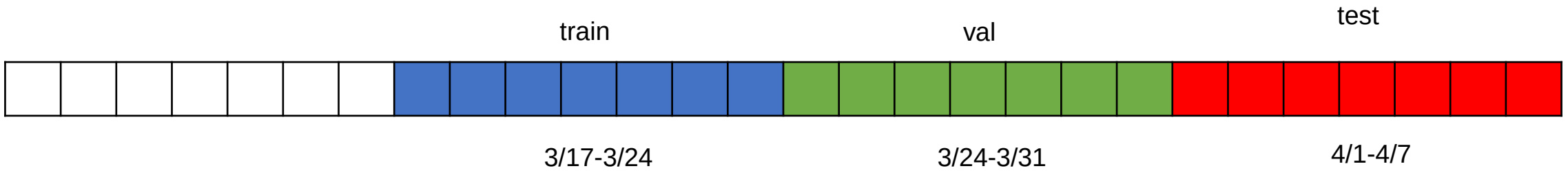
각 값들의 거래패턴을 찾을 수 있음.
화이트데이에만 거래량이 급증한 패턴을 제거

Ⅱ. EDA & Preprocessing

데이터셋 생성 (SARIMA 기준)



Test구간을 제외하고 학습 -> train구간과 test구간의 loss 조화평균을 통해 파라미터 결정



train, val, test 구간을 1주일씩 저장

Ⅱ. EDA & Preprocessing

그 외 시도해본 것들

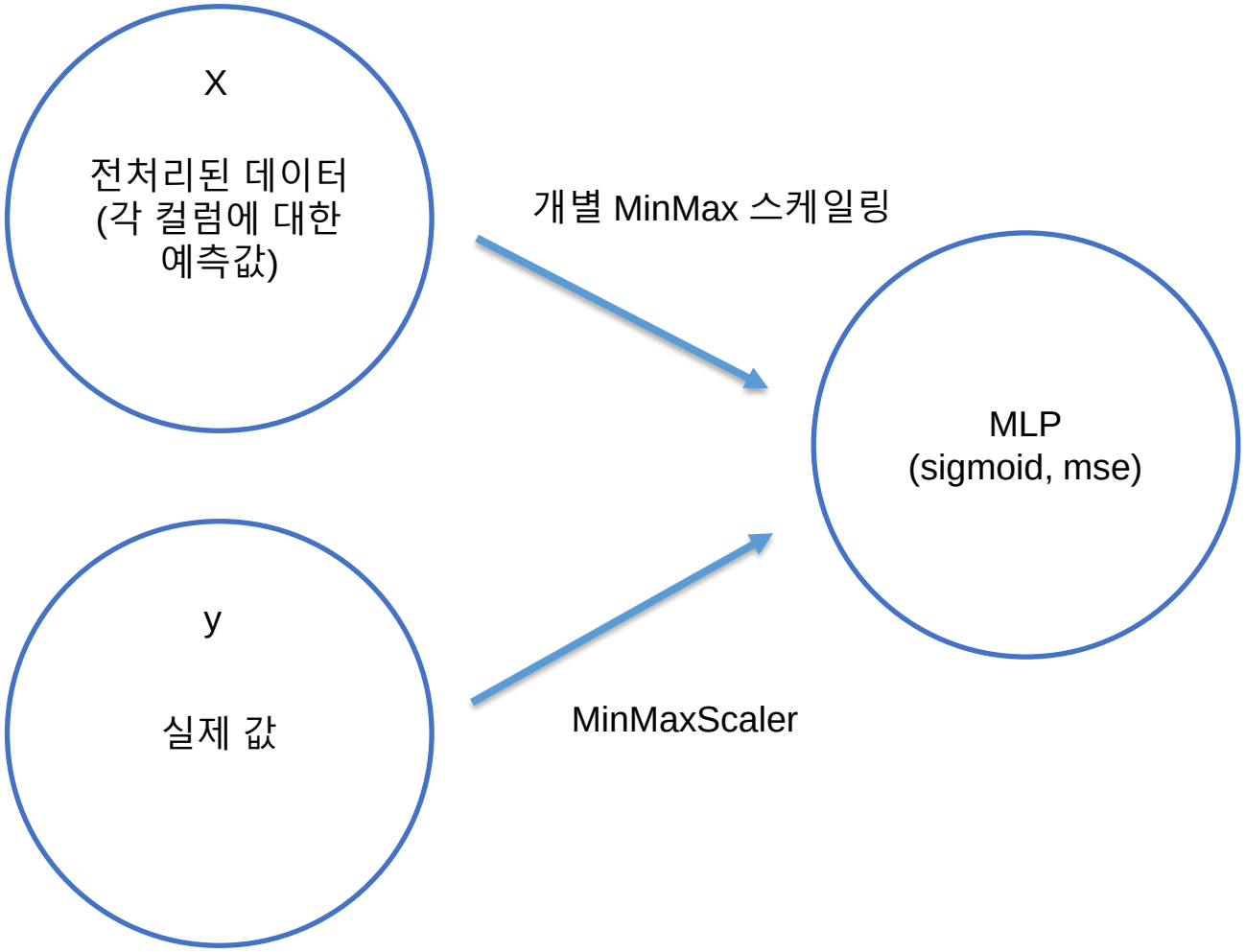
실수차분 :
하나의 시계열에서 정상성이 있는 성분을 추출하여 피처로 사용할 수 있으나
기간에 비해 데이터 차원이 증가하여 오히려 학습이 잘 안됨

GAN+LSTM :
주어진 기간이 너무 작기때문에 많은 데이터를 생성하기 힘듬
GBM으로 데이터를 만들고 GAN으로 추가 생성 가능할 것으로 예상

Ⅲ. Modeling

Ⅲ. Modeling

1. MLP



1. 전처리데이터는 각 데이터가 의미가 있다고 생각하기 때문에 개별로 MinMax 스케일링을한다.
Standard 스케일링은 값이 음수가 나올 수 있는데, 이것이 과적합을 유발할 수 있다고 생각하여 사용하지 않음.

2. 피처들의 과적합을 막기 위해 sigmoid를 활성화 함수로 사용

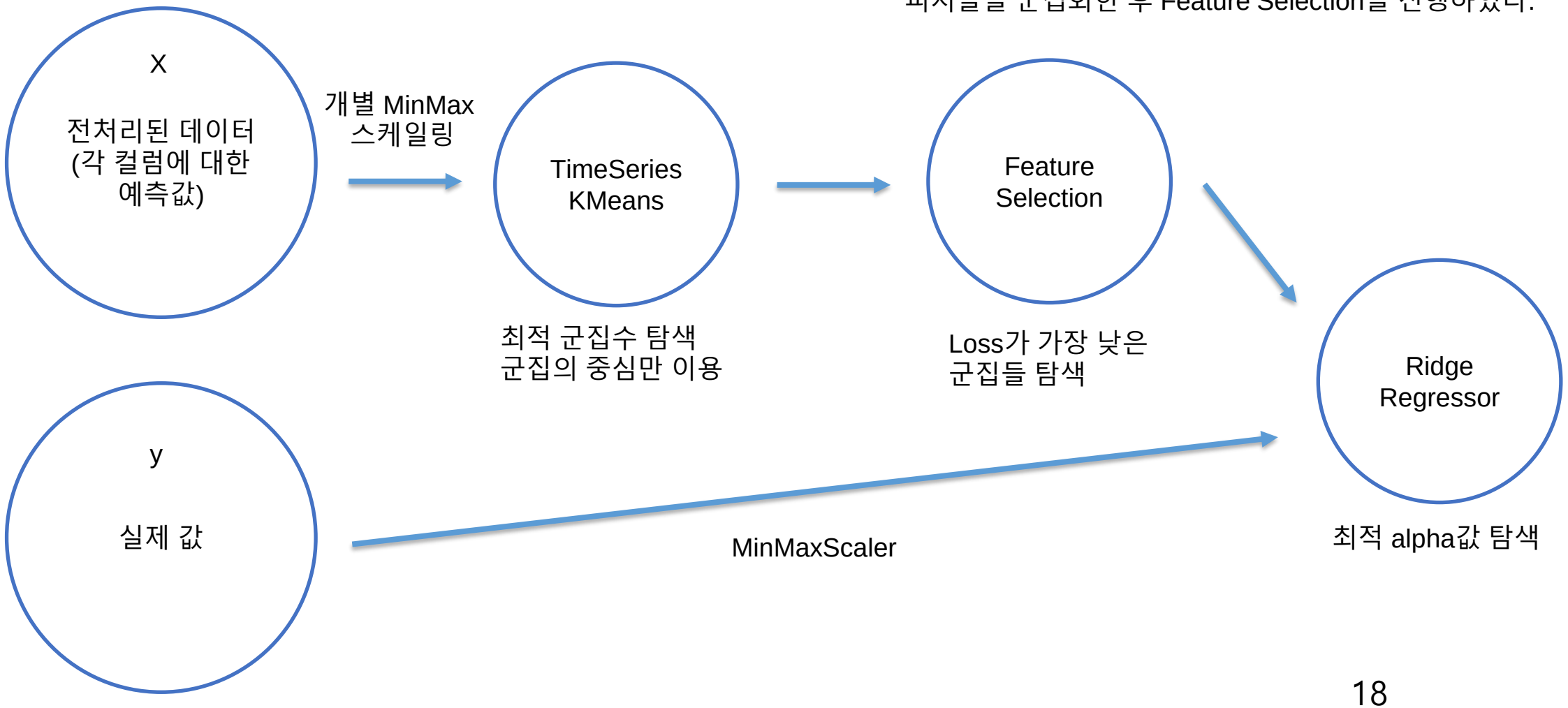
3. MAPE로 학습을 할때보다 MSE로 학습을 할때 결과의 MAPE값이 더 좋았기 때문에 MSE로 학습

4. 피처수가 많아짐에 따라 학습속도가 현저히 저하되고 과적합과 Feature Selection의 어려움으로인해 Rigde 회귀로 재수행

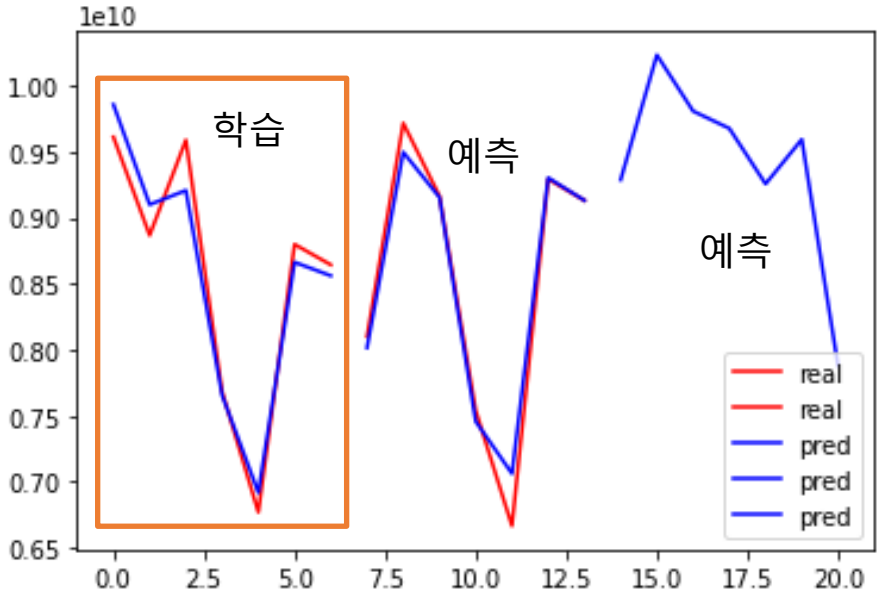
Ⅲ. Modeling

2. Ridge Regressor

각 피처들의 모든 조합을 Feature Selection하려면
시간복잡도가 과다하게 증가하기때문에
피처들을 군집화한 후 Feature Selection을 진행하였다.



Ⅲ. Modeling



	Id	Predicted
0	20200401	9.192552e+09
1	20200402	1.010752e+10
2	20200403	9.879320e+09
3	20200404	9.308602e+09
4	20200405	8.998300e+09
5	20200406	9.719870e+09
6	20200407	8.310735e+09

최종 제출

2주전의 데이터로 1주전을 예측한다.
결과가 좋다면 1주전의 데이터로 미래1주를 예측할 수 있다.
시계열에서 미래데이터로 과거를 예측하는 것은 옳지 않다.