1. 데이터 전처리 (1번 공정 데이터 문제)

① 기초 통계량 Shape = (404, 4676)

ID 당 1698의 시계열 데이터를 포함, 대표 값으로 사용하기 위해 기초 통계량 추출

Mean	Std	Max	Min	Q1	Q2	Q3

② Quantile(90,95,98,99) Shape = (404, 8684)

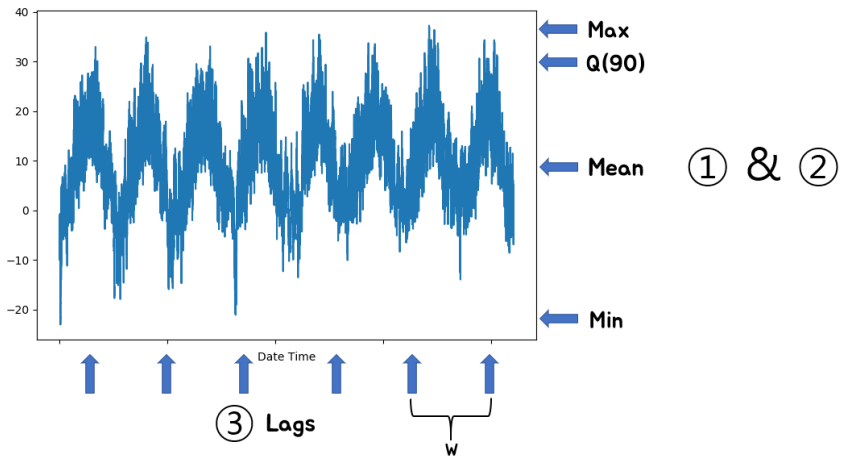
Y 값과의 상관계수가 MIN, MEAN, MAX 중에서 MAX가 가장 높았음

Q1,Q2,Q3 이외에 MAX와 가까운 90~100 값들을 N개 추출

성능이 개선되는 것으로 보아, 높은 MAX에 근접한 수들이 유의미한 것으로 파악

③ LAG 활용 Shape = (404,11356) [Ridge의 경우 (404, 113560)]

시계열 데이터의 특성을 반영하기 위해서 1697 행의 시계열 데이터에서 W간격으로 N개의 데이터를 순차적으로 추출



추출 간격 W는 실험적으로(10,20,50,100) 성능을 비교하여 결정

간격을 100(W)으로 설정하고 시계열 순서대로 ID당 17개의 값을 추출하여 column으로 생성, Ridge 모델에서만 예외적으로 w를 10으로 설정하여 ID당 170 column 생성

④ 최신 데이터 표준편차 반영

기초 통계량 중 std 가 y 값과의 가장 상관관계가 높음

따라서, 1698 개를 N으로 나눈 후 현재와 가까운 구역의 표준편차 계산

2. 모델 평가

40x개의 ID를 Train, Vaildation 300개 Test set 100로 SPLIT

K-Fold = 10 으로 평가한 MSE 값들의 평균을 지표로 평가

최종 결과물을 제출 할 때에는 전체 데이터를 학습으로 사용

(1) Linear Regression

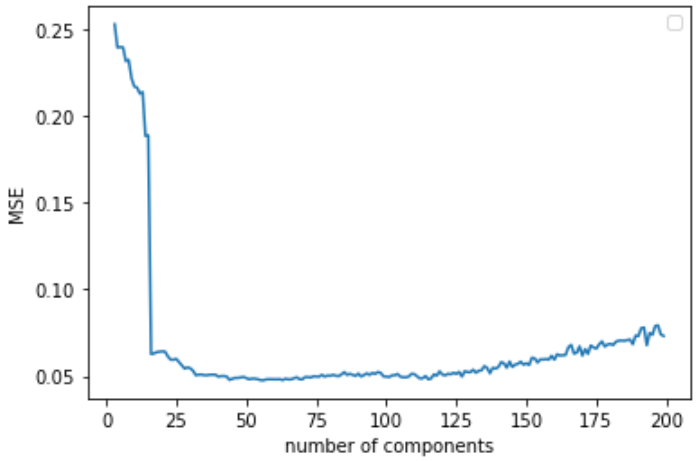
① + ② + ③ 전처리 데이터를 결합한 결과 Shape (404 , 24716)

PCA 기법을 통해서 24717 개의 컬럼 50개로 축소한 결과 0.0491로 Validation성능 개선

하지만 변수들을 추가적으로 생성했을 때 (③LAG) Test set 은 Overfitting

PCA를 통해 N 차원으로 축소 하였을 때의 MSE를 그래프로 확인하고, 성능이 가장 좋은 차원의 수(N)를 50으로 선택

데이터 셋	VALID(MSE)	TEST
① 기초 통계량 +	0.0613	0.0583
① + ② Quantile(90,95,98,99 등)	0.0550	0.0521
① + ② + PCA	0.0528	0.0411
① + ② + ③ Lag 값 추가	0.0533	0.0561
① + ② + ③ + PCA (n_compo = 50)	0.0491	0.0524
① + ② + ③ + ④ 표준편차 + PCA	0.0483	0,0452



(2) Lasso model

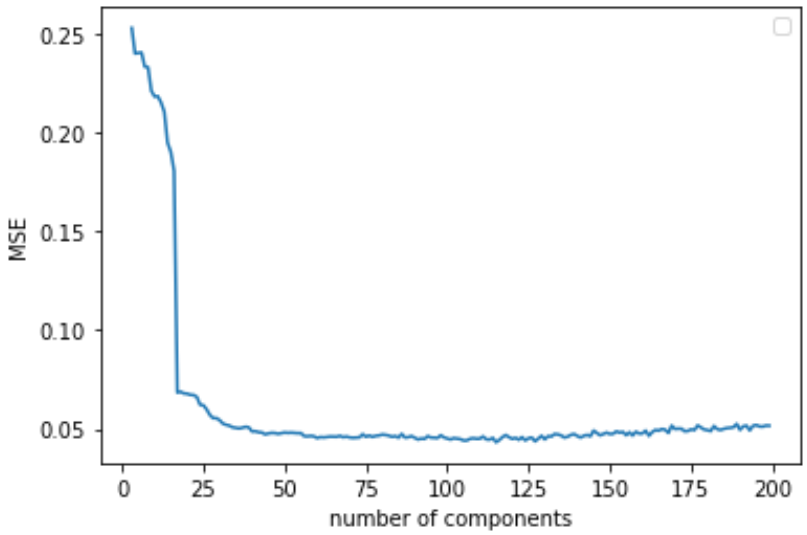
중요한 변수(Column) 를 선택해서 사용하는 Lasso 모델 특성상, 변수가 너무 많아지는 경우(중요한 정보들을 여러 변수에 나눠서 담는 경우) 데이터의 손실이 일어날 수 있다 따라서, 핵심적인 변수들을 만드려고 노력

데이터 셋	VALID(MSE)	TEST
① 기초 통계량	0.589	0.054
① + ② Quantile(90,95,98,99 등)	0.068	0.0512
① + ② +PCA	0.0489	0.0402
① + ② + ③ Lag 값 추가	0.0565	0.0532
① + ② + ③ + PCA	0.0457	0.0524
① + ② + ③ + ④ 표준편차 + PCA	0.0473	0.0476

Alpha 값에 따른 성능을 확인하면서, Alpha 값을 선택 (0.2 ~ 0.3 사이에서 좋은 성능)

Alpha	0.1	0.2	0.3	0.5	1.0	5	10
MSE	0.046	0.0457	0.0452	0.0453	0.0464	0.06	0.07

PCA를 통해 N 차원으로 축소 하였을 때의 MSE를 그래프로 확인하고, 성능이 가장 좋은 차원의 수(N)를 65으로 선택



(3) Ridge model

최대한 많은 변수들을 넣은 경우에 성능이 개선되는 것을 확인 PCA를 사용하지 않음

LAG 값을 넣을 때 (1) 번 보다 추출 간격을 10배 증가 시킴

Vaildation 에서는 좋은 성능을 보였지만, TEST SET 에서는 성능 하락으로 이어짐

데이터 셋	VALID(MSE)	TEST
① 기초 통계량 +	0.0605	0.0589
① + ② +PCA	0.0513	0.0413
① + ② + ③ Lag 값 추가(10배)	0.0449	0.0522
① + ② + ③ + PCA	0.0475	0.0553
① + ② + ③ + ④ 표준편차	0.0462	0.0463

Alpha 값에 따른 성능을 확인, Alpha 값을 선택 (값에 따라 MSE의 변동이 거의 없음)

Alpha	0.1	0.2	0.3	0.5	1.0	5	10
MSE	0.0447	0.0447	0.0447	0.0447	0.0447	0.0447	0.0447

결론

- ◆ 정규화(Regulization) 을 가지고 있는 lasso, Ridge 모델이 일반적인 회귀 모델모다 성능이 전반적으로 우수함
- ◆ Lasso , Ridge 모두 너무 많은 변수보다는 차원을 축소하여 (PCA) 변수를 줄이는 것이 성능 향상으로 이어짐

Ridge	Lasso
L_2-norm regularization	L_1-norm regularization
변수 선택 불가능	변수 선택 가능
Closed form solution 존재 (미분으로 구함)	Closed form solution이 존재하지 않음 (numerical optimization 이용)
변수 간 상관관계가 높은 상황 (collinearity)에서 좋은 예측 성능	변수 간 상관관계가 높은 상황에서 ridge에 비해 상대적으로 예측 성능이 떨어짐
크기가 큰 변수를 우선적으로 줄이는 경향이 있음	

출처:https://rk1993.tistory.com/entry/Ridge-regression%EC%99%80-Lasso-regression-%EC%89%BD%EA%B2%8C-%EC%9D%B4%ED%95%B4%ED%95%98%EA%B8%B0

고등지능정보 공학 - 임형구 2021311558

1. 데이터 전처리

(1) 범주형 변수는 ONE-HOT-ENCODING ('runnynose', 'bodypain','diffbreath)

(2) 연속형 변수는 범주형으로 그룹화 (np.digitze())

온도 36.5 37.5 38.5

나이 10 ~ 80 (앞자리로 구분)

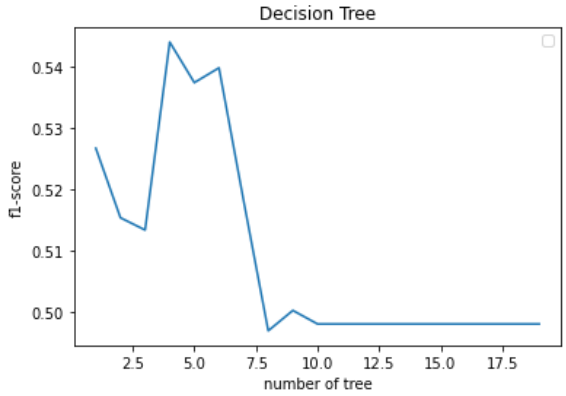
대체적으로 전처리 전후에 큰 차이가 없고, 모든 변수와 코로나 여부가 굉장히 낮은 상관관계를 갖고 있음

2. DecisionTreeClassifier , KNN, LogisticRegression 모델

데이터 SPILT 비율, random_seed 에 따른 성능의 변화가 굉장히 큼

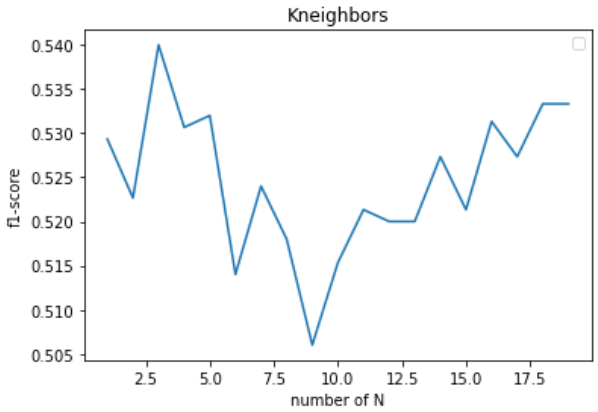
성능의 안정성을 위해서 K-Fold = 10, random_seed 40 개 약 400번 데이터를 나눈 값의 평균으로 성능을 비교

Decision Tree 의 경우 가장 우수한 성능을 보이는 depth = 4 설정



KNN의 경우 n_neighbors = 9 값에서 가장 좋은 성능을 보임.

어떤 데이터를 넣느냐에 따라 선택해야하는 n_neighbor 값이 크게 달라짐



Logistic regression 또한 전처리 여부와 상관 없이 낮은 성능을 보여줌