삼성전자
주가 예측 모델링

3조 모델아화성보내조

팀장: 정준현
팀원: 김재홍, 김지윤, 이남선

Contents

001

주제 선정 배경

- 배경
- 프로젝트 목표

1

001 주제 선정 배경

배경 및 프로젝트 목표

배경

- ✓ 최근 코로나19로 인한 초저금리 현상으로 금융시장이 하락함에 따라 저가매수를 노리는 투자자들로 전세계적 투자 열풍
- ✓ 금리인상 예상에 따른 주가 변동성 예측 필요성 대두
- ✓ 퀀트투자 관심도 상승

여론 속의 여론

주식 투자자 43% "코로나 이후 시작"… 92% "계속할 것"

입력 2021.05.06 04:30 1 0

매일경제

"나 떨고 있니"...동학개미 막 내린 초저금리 대응 어떻게

지난해 3월 세계보건기구(WHO)가 코로나19 대유행(팬데믹)을 선언하자 한은은 ... 개인이 주식을 대규모로 사고 있지만 투자 열풍이 점차 사그라드는...

2021. 9. 14.

주간동아

옥석만 가린다! 파이어족 이끈 年 34% 수익 '울트라 퀀트 투자'

퀀트 투자로 15년 동안 연평균 15% 수익을 내고 있는 강환국 씨. 지난해 울트라 퀀트 전략을 개발한 그는 올해 7월 파이어족이 됐다. [박해윤 기자].

2021. 9. 18.

프로젝트 목표

이미지 출처: 유튜브 팝포유채널 '[*Playlist*] 도지타고 화성 갈끄니까!'

- ✓ 국내 증시 대장주인 삼성전자 역시 최근 큰 하락과 상승을 겪으며 향후 추이가 주목되고 있는 상황
- ✓ 우리가 배운 ML모델과 새롭게 스터디한 DL 모델을 활용하여 삼성전자 주가를 예측해 화성에 가도록 하자! 화성 갈끄니까~

002

데이터 수집 및 정제

- 데이터 수집
- 데이터 정제 및 EDA
- 변수 설명

2

002 데이터 수집 및 정제

데이터 수집

데이터 수집 소스

FinanceDataReader

삼성전자
Dividends

yahoo!finance

S&P500
다우존스
나스닥
금리

ECOS
한국은행 경제통계시스템
Economic Statistics System

Dollar
Dollar_rate

데이터 수집 코드 예시

코드

```
samsung_df = fdr.DataReader('005930', '2020-01-01')
```

실행 결과

Date	Open	High	Low	Close	Volume	Change
2020-01-02	55500	56000	55000	55200	12993228	-0.010753
2020-01-03	56000	56600	54900	55500	15422255	0.005435
2020-01-06	54900	55600	54600	55500	10278951	0.000000
2020-01-07	55700	56400	55600	55800	10009778	0.005405
2020-01-08	56200	57400	55900	56800	23501171	0.017921
...	...	...	...	...	...	...
2021-12-09	77400	78200	77000	78200	21604528	0.010336
2021-12-10	77400	77600	76800	76900	9155219	-0.016624
2021-12-13	77200	78300	76500	76800	15038750	-0.001300
2021-12-14	76500	77200	76200	77000	10976660	0.002604
2021-12-15	76400	77600	76300	77600	9355116	0.007792

데이터 정제 및 EDA

결측치

1. S&P500, 다우존스, 나스닥 미증시 휴장으로 인한 결측치
   → '전일 종가'로 대체
2. Dividends 결측치
   → 0(zero)으로 대체

분류용 라벨 생성

1. 전일비 변동률('Change')에 따라 'Target' 컬럼 생성
   · 양수(상승): 1
   · 음수(하락): 2
   · 0 (보합): 3

EDA

1. 히트맵
2. 시계열그래프

변수 설정

설명변수명	설명
Open	삼성전자 시가
High	삼성전자 고가
Low	삼성전자 저가
Volume	삼성전자 거래량
Change	삼성전자 전일비 변동률
Dividends	삼성전자 배당금
Dollar	환율
Interest	금리
S&P500	S&P500 지수
DJIA	다우존스 지수
Nasdaq	나스닥 지수

분류용 라벨 생성

반응변수명	설명
Close (분류 外)	삼성전자 종가
Target (분류)	삼성전자 전일비 변동라벨 (상승:1, 하락:2, 보합:3)

1. S&P50
미증시
→ '전
2. Divide
→ 0(z

비 변동률('Change')에
'Target' 컬럼 생성
양수(상승): 1
음수(하락): 2
0 (보합): 3

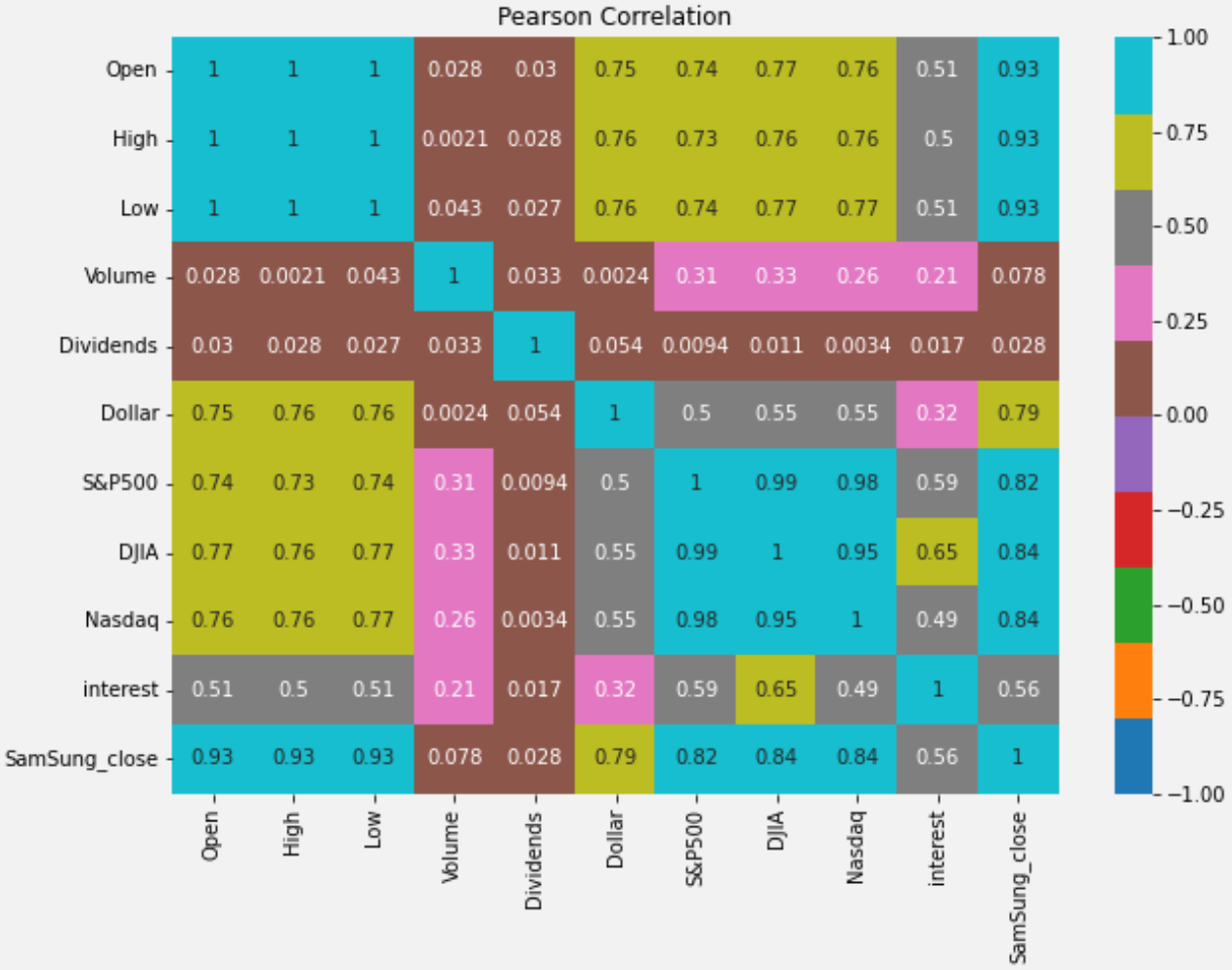
1. 히트맵
2. 시계열그래프

2

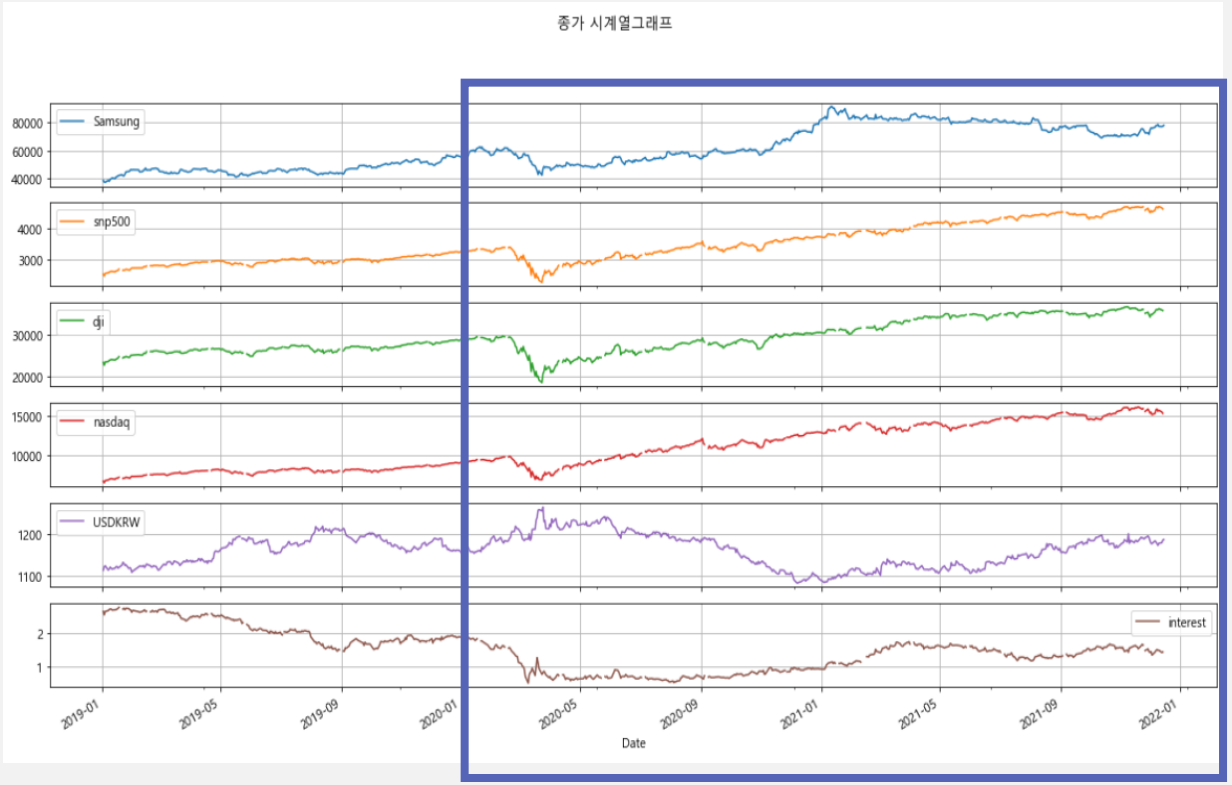
002 데이터 수집 및 정제

데이터 정제 및 EDA

EDA 1.히트맵



EDA 2.시계열그래프



학습용 데이터
2020-01-01 이후로 설정

데이터 정제 및 EDA

EDA 1. 히트맵

EDA 2. 시계열그래프

2020-01-01 이후로 설정

003

모델 테스트

- 분류모델
- 회귀모델
- 시계열예측(ARIMA)

3

003 모델 테스트

사용 모델

CLASSIFICATION

- Decision Tree
- Logistic Regression

REGRESSION

*random_state=17, cv=5 고정

- Linear Regression
- Ridge
- Lasso
- Decision Tree Regressor
- Random Forest Regressor
- Gradient Boosting Regressor
- XGB Regressor

TIME SERIES

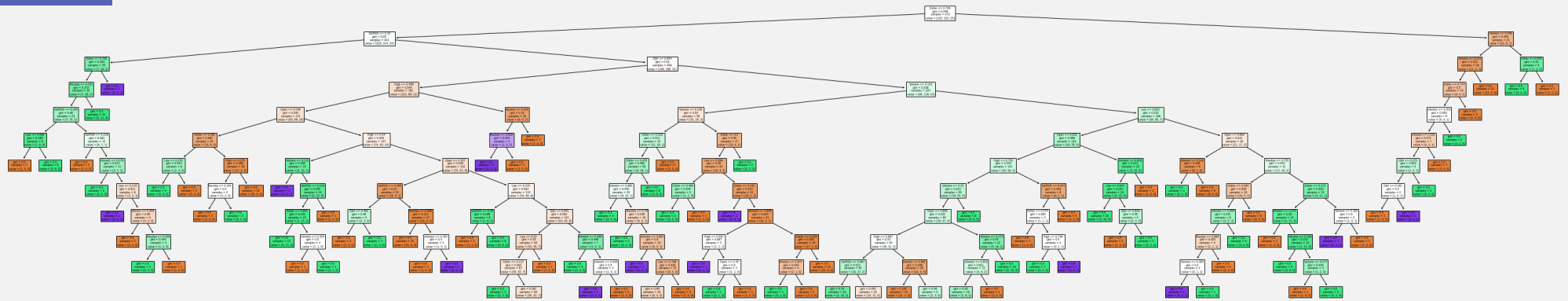
- ARIMA

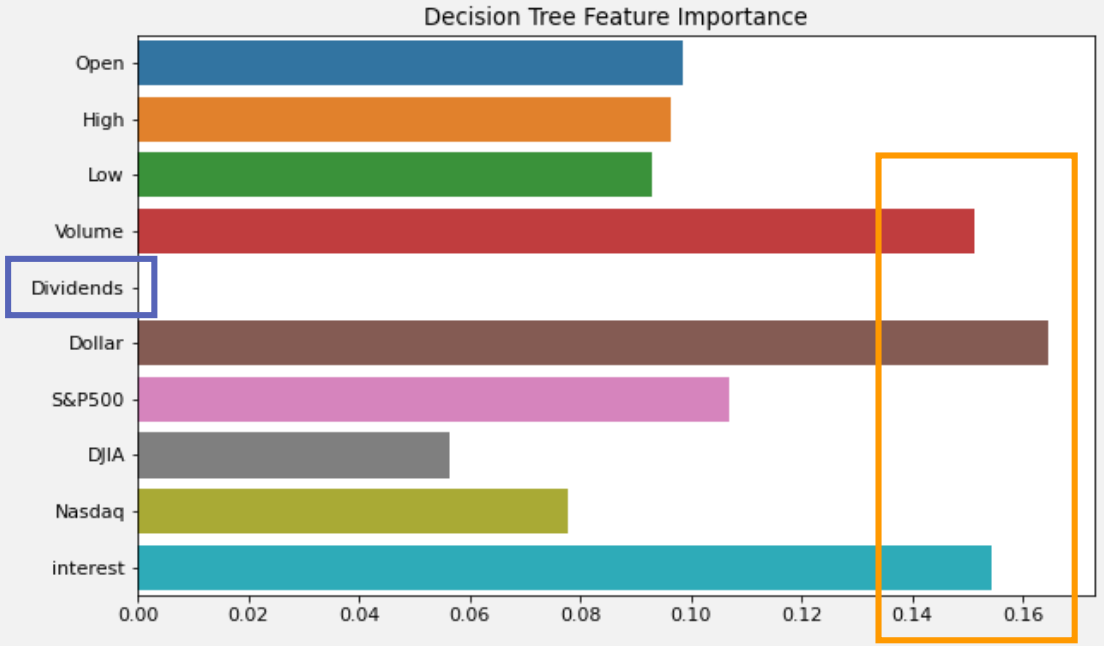
3

003 모델 테스트

분류모델

Decision Tree | 기본 모델 |





Dividends '0'값이 많아
중요도 낮음
→ feature 제외

Decision Tree feature importance

Open - Score: 0.0985
High - Score: 0.0965
Low - Score: 0.0931
Volume - Score: 0.1513
Dividends - Score: 0.0000
Dollar - Score: 0.1648
S&P500 - Score: 0.1070
DJIA - Score: 0.0566
Nasdaq - Score: 0.0779
interest - Score: 0.1543

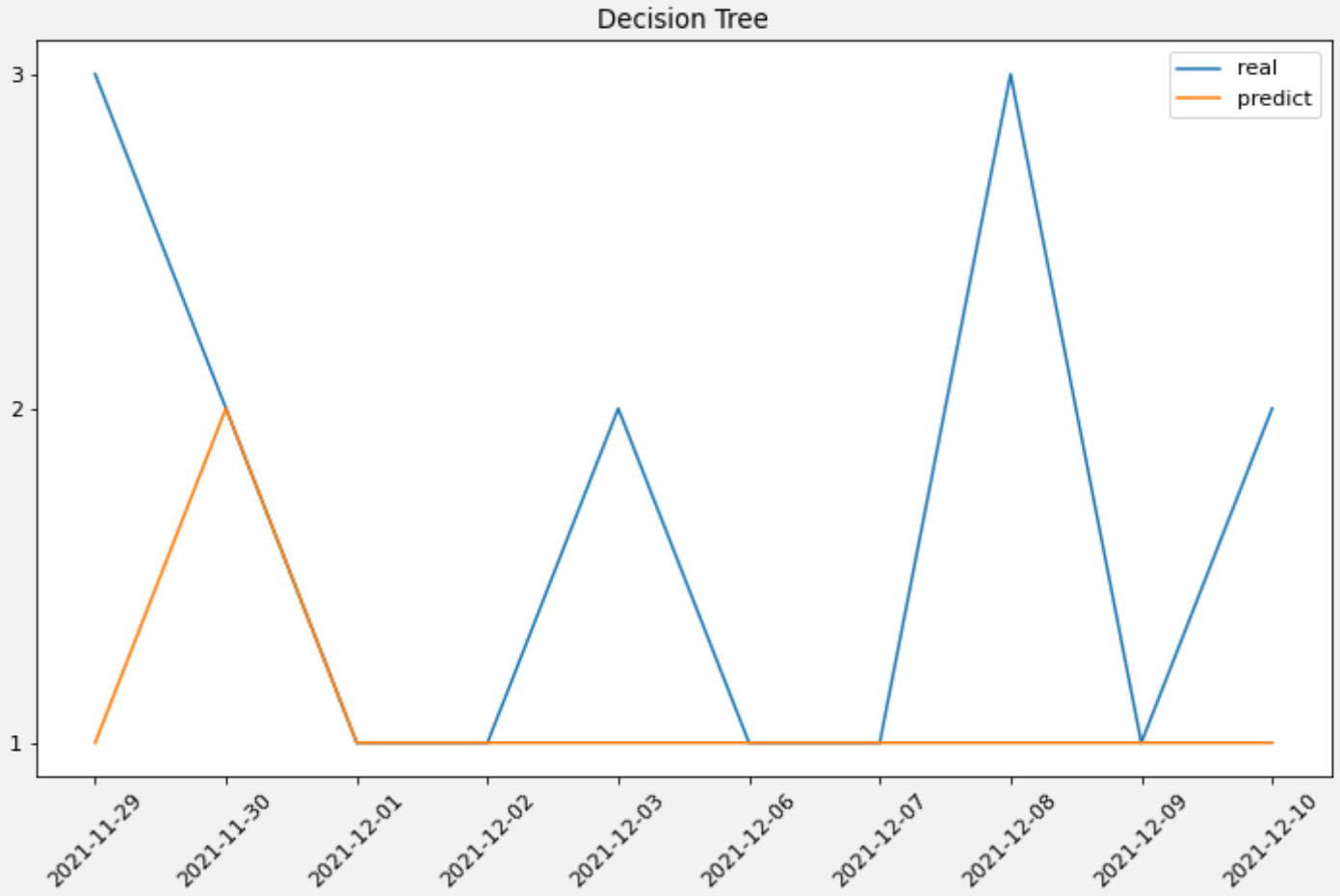
Train score : 1.0
Test score : 0.5

분류모델

Decision Tree | 하이퍼파라미터 조정 - GridSearch |

- Scoring : 'accuracy'
- 파라미터 조정 범위 :
  - max_depth: [3, 5, 7, 9, 11, 13, 15, 17, 19, 21, 23]
  - min_samples_split: [2, 7, 12, 17, 22, 27, 32, 37, 42, 47]
- 최적 파라미터 값 :
  - max_depth: 11
  - min_samples_split: 2
- **최적 파라미터를 적용한 모델의 train score: 0.4385**
- **최적 파라미터를 적용한 모델의 test score: 0.6**

<Testset 예측 결과>

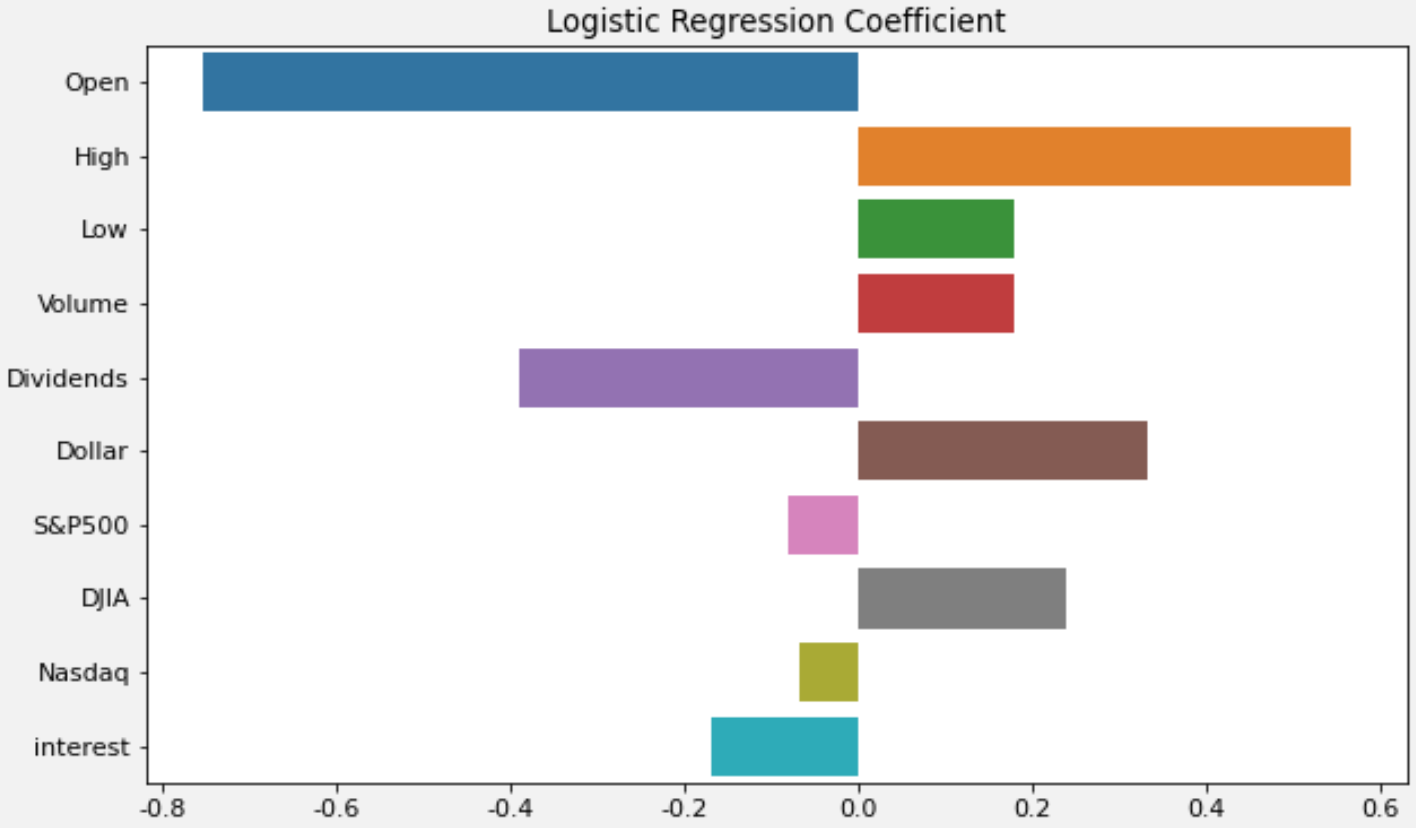


3

003 모델 테스트

분류모델

Logistic regression | 기본 모델 |



<Logistic regression Coefficient>

Open - Score: -0.7534
High - Score: 0.5654
Low - Score: 0.1796
Volume - Score: 0.1797
Dividends - Score: -0.3905
Dollar - Score: 0.3321
S&P500 - Score: -0.0801
DJIA - Score: 0.2399
Nasdaq - Score: -0.0680
interest - Score: -0.1682

Train score : 0.5339
Test score : 0.3

3

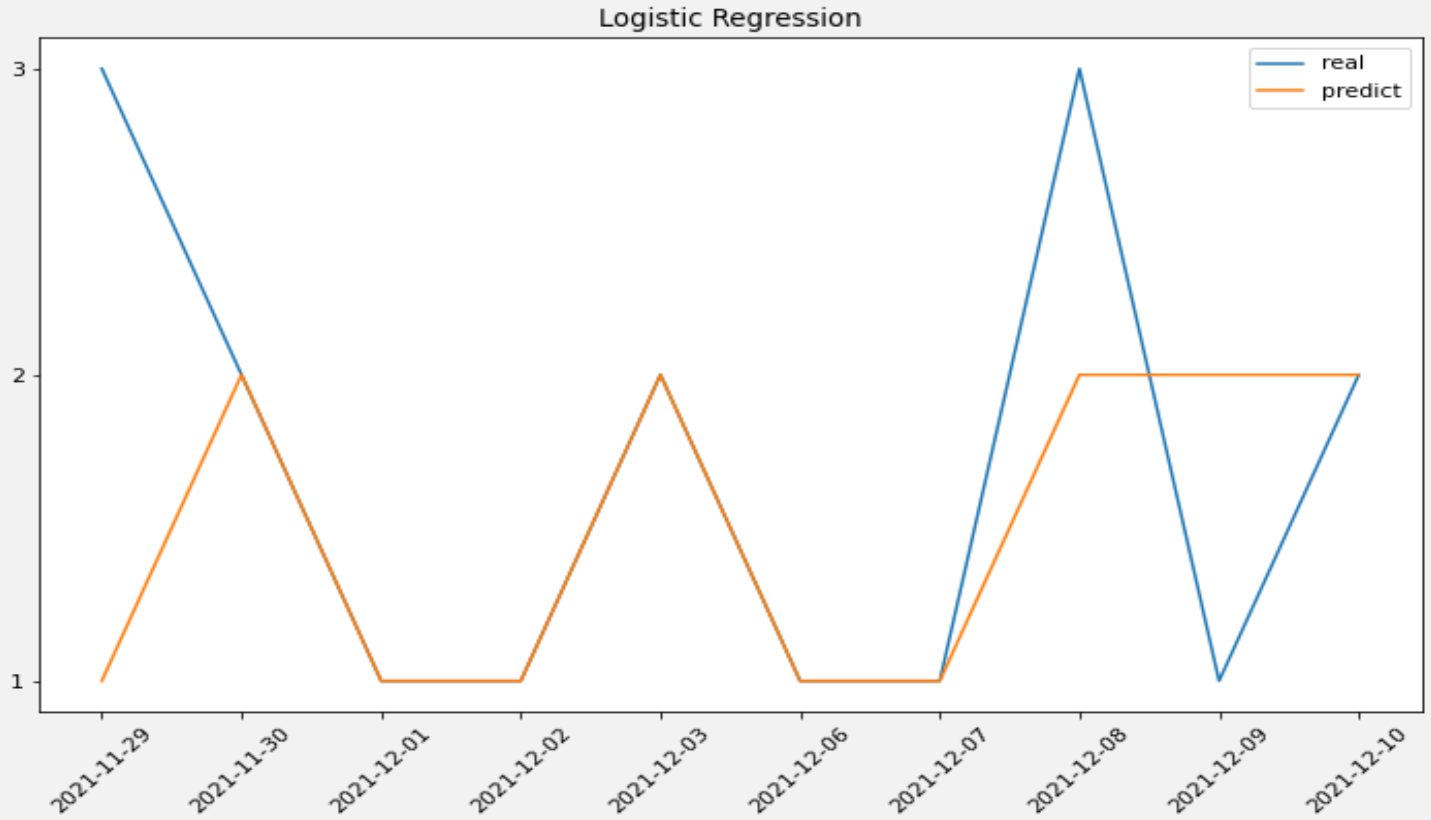
003 모델 테스트

분류모델

Logistic regression | 하이퍼파라미터 조정 - GridSearch |

- Scoring : 'accuracy'
- 파라미터 조정 범위 :
  - C: [0.001, 0.01, 0.1, 1, 10, 100, 1000]}
- 최적 파라미터 값 :
  - C: 1000

- **최적 파라미터를 적용한 모델의 train score: 0.5613**
- **최적 파라미터를 적용한 모델의 test score: 0.7**

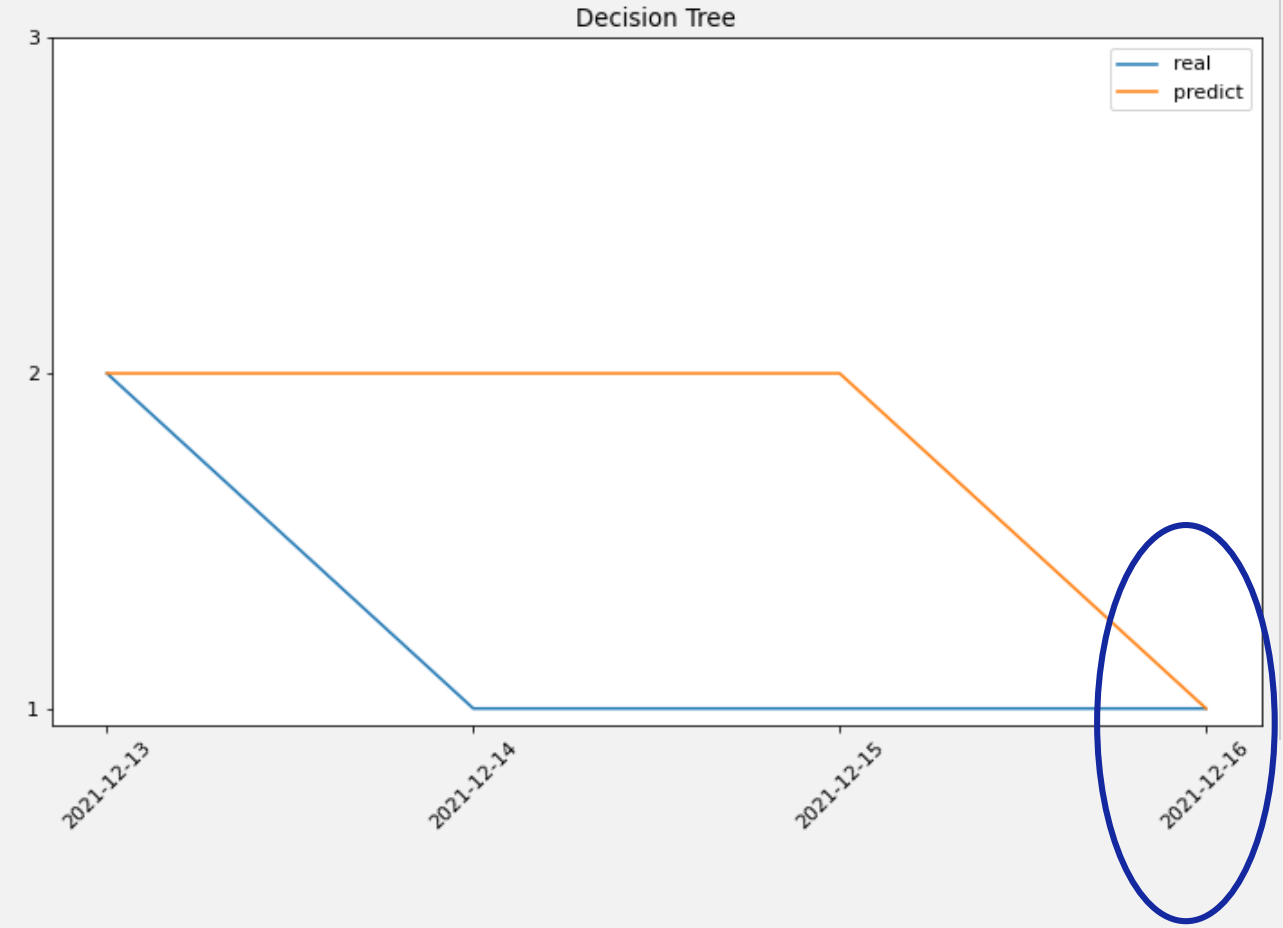


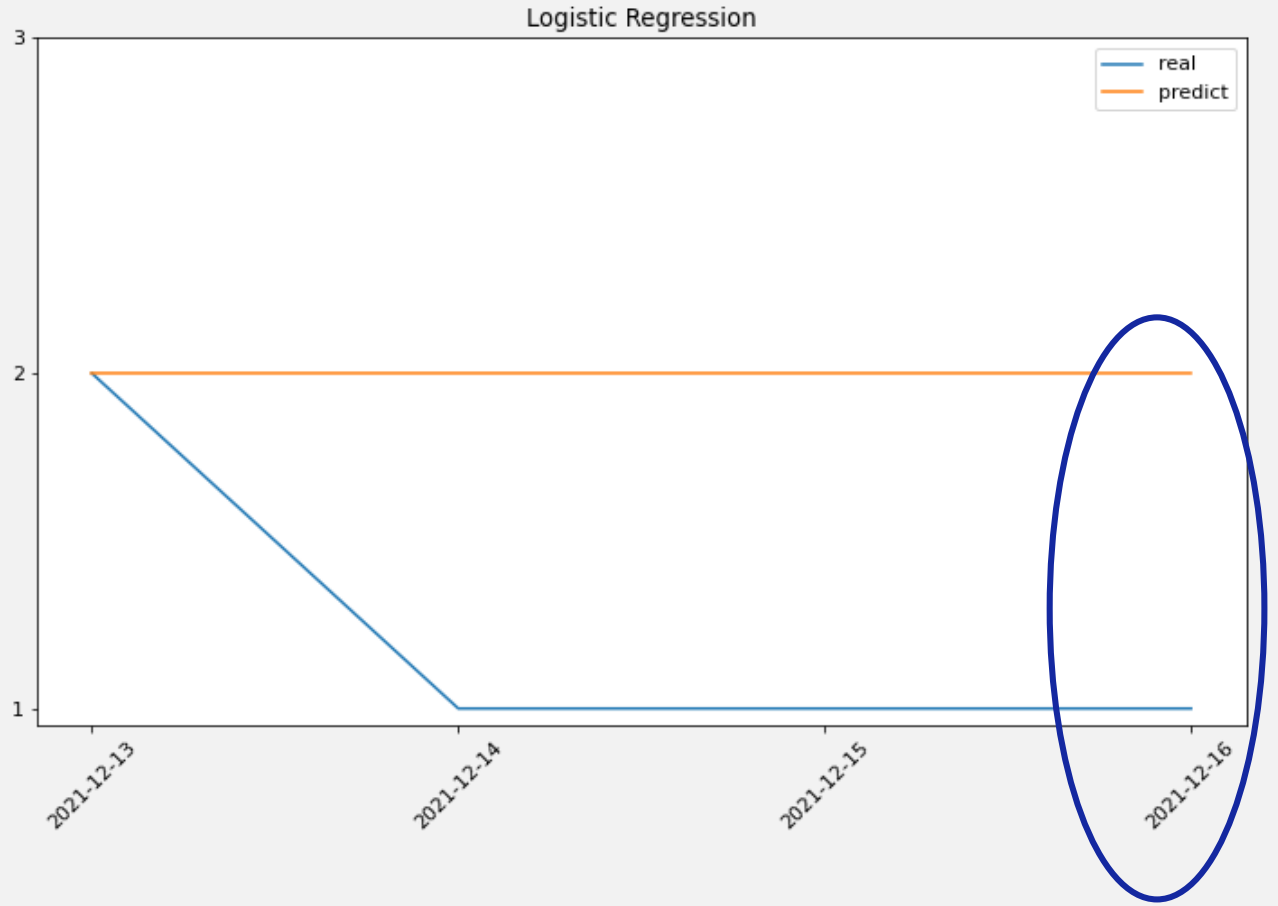
3

003 모델 테스트

분류모델

금일 종가 예측 21/12/16 15시 삼성 데이터 -> 종가 : 77,800원

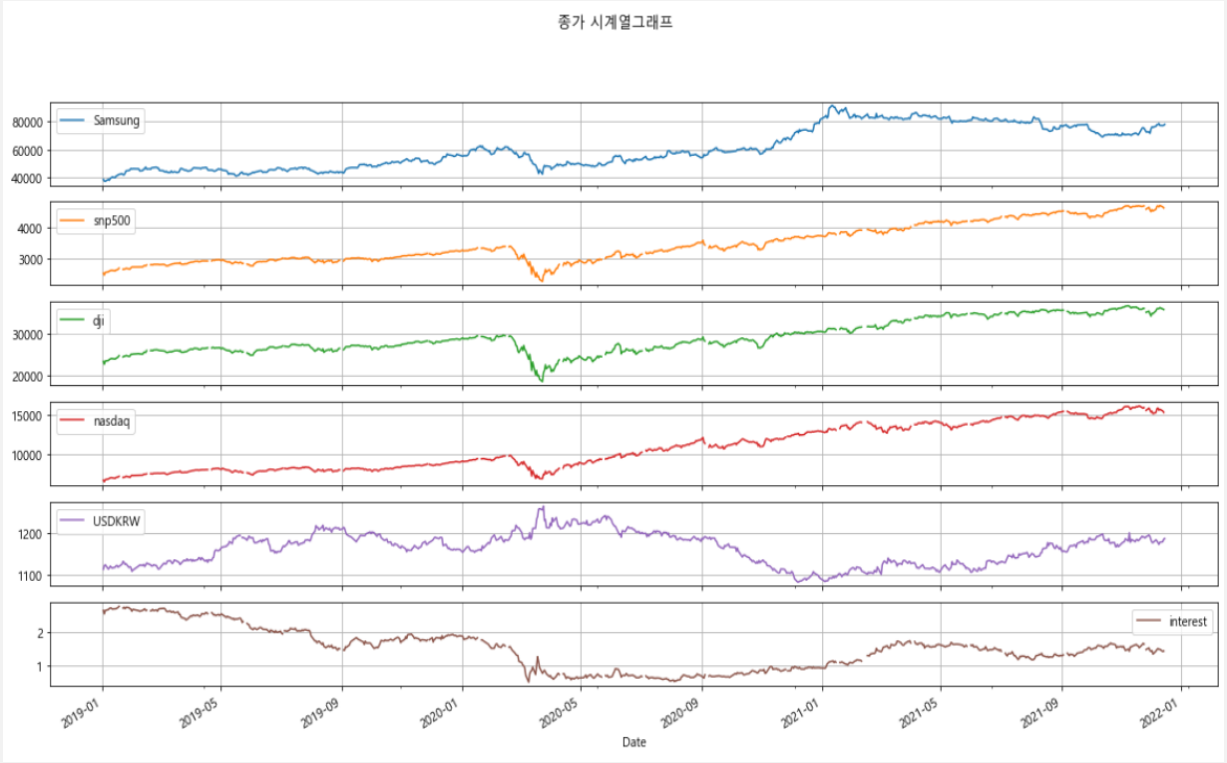




3

003 모델 테스트

회귀모델



3

003 모델 테스트

회귀모델

Linear regression

- **기본 모델**

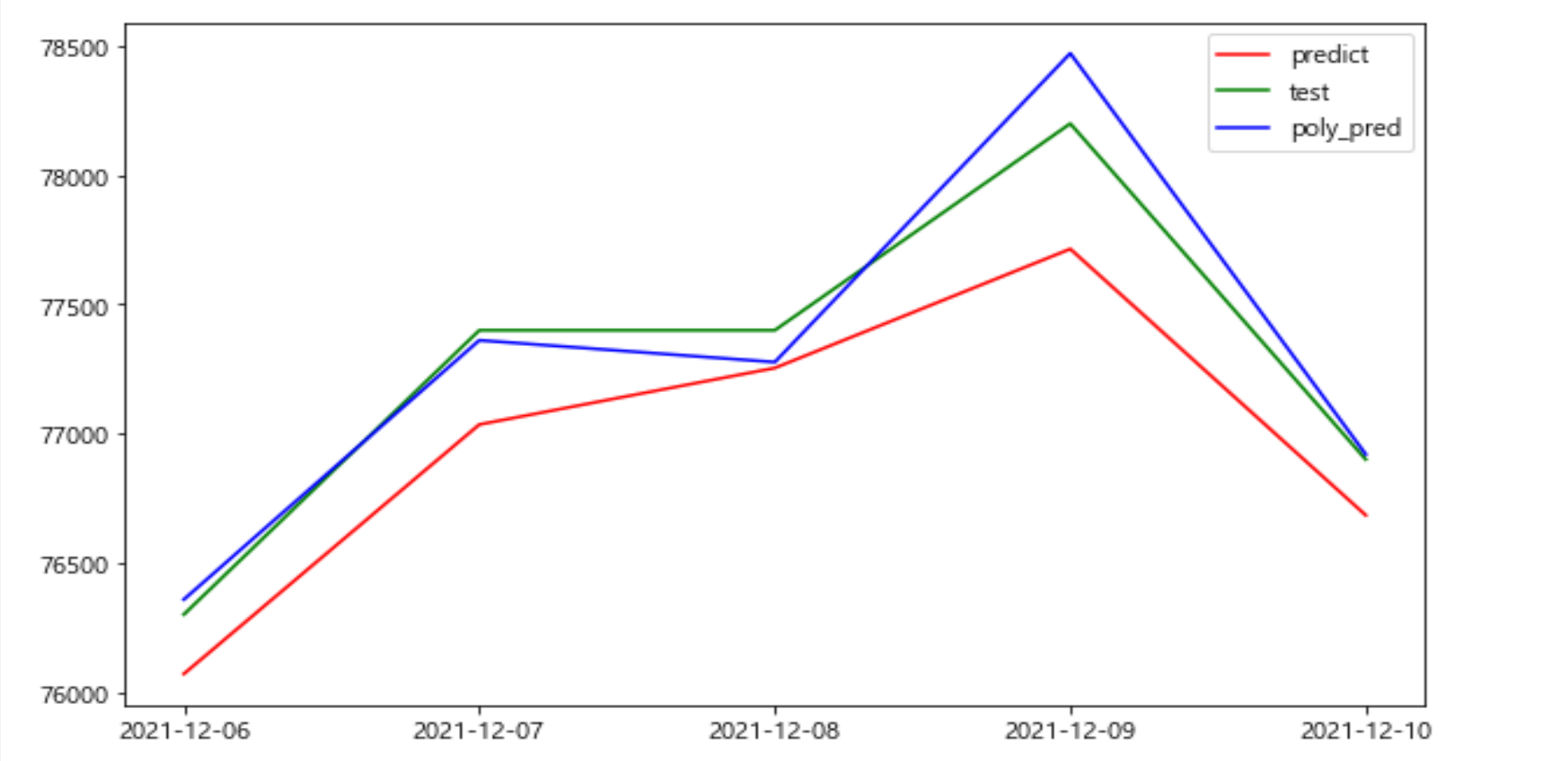
  train_score : 0.9997
  test_score : 0.7518

- **다항화 + 하이퍼파라미터 조정**

  degree = 3

  train_score : 0.9999
  test_score : 0.9521

<Testset 예측 결과>



3

003 모델 테스트

회귀모델

Ridge

- **기본 모델**

  train_score : 0.9997
  test_score : 0.7516

- **기본모델 + 하이퍼파라미터 조정**

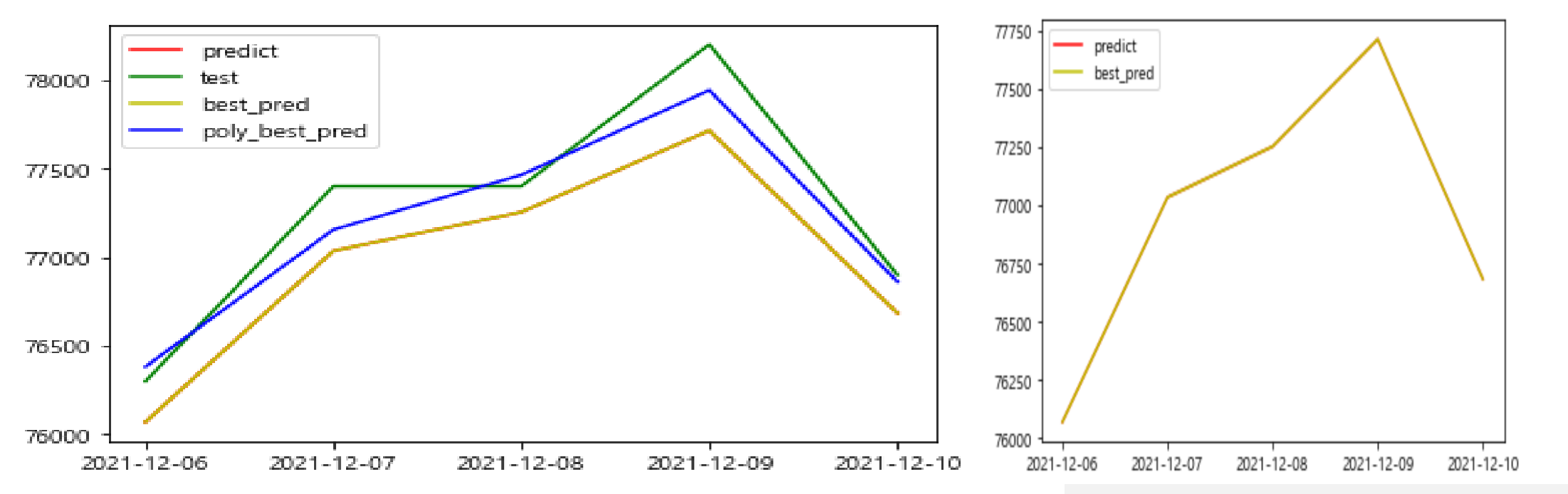
  alpha=0.001, 'max_iter'=500

  train_score : 0.9997
  test_score : 0.7518

- **다항화 + 하이퍼파라미터 조정**

  degree = 2
  alpha=10, max_iter=500

  train_score : 0.9999
  test_score : 0.9295
  best_train_score : 0.9999
  best_test_score : 0.9295



3

003 모델 테스트

회귀모델

Lasso

- **기본 모델**

  train_score : 0.9996
  test_score : 0.8295

- **기본모델 + 하이퍼파라미터 조정**

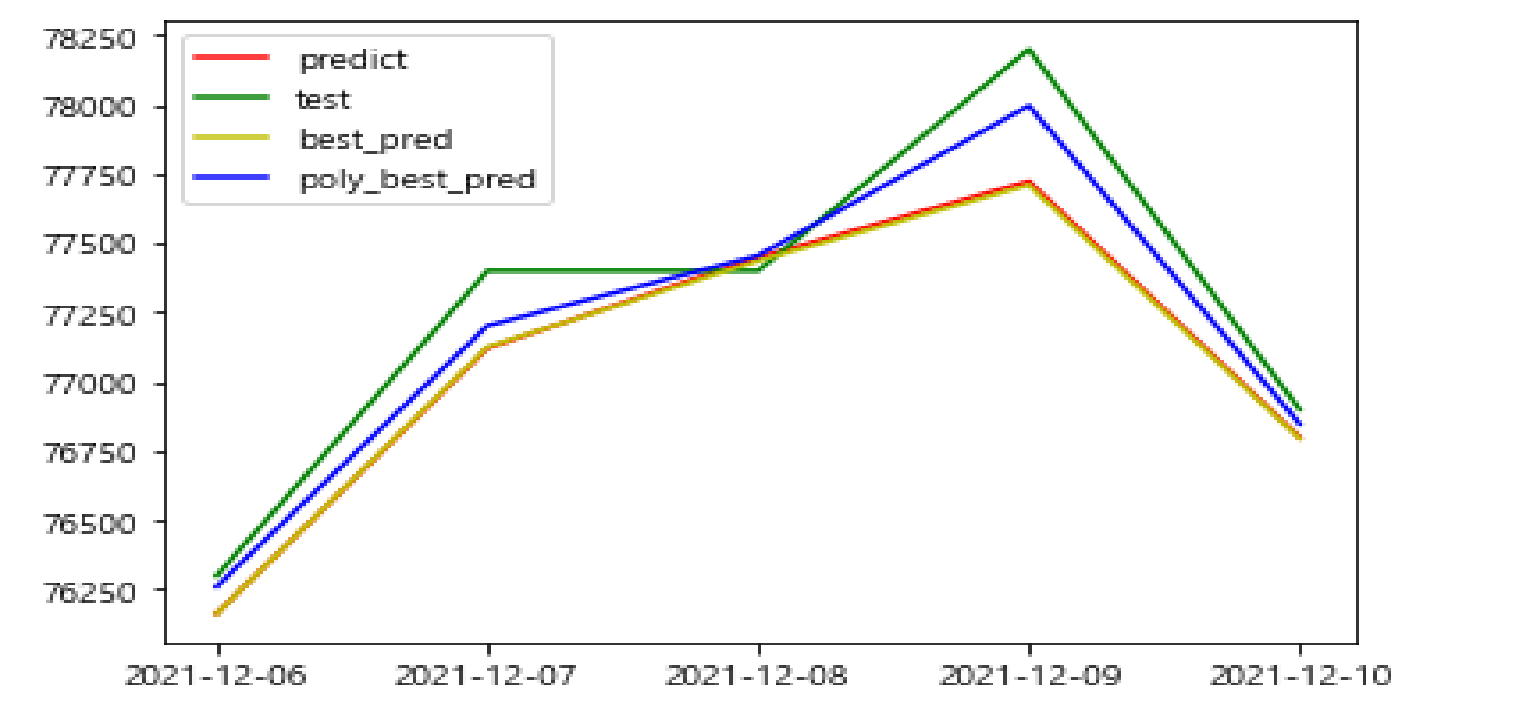
  alpha=5, 'max_iter'=1000

  train_score : 0.9996
  test_score : 0.8242

- **다항화 + 하이퍼파라미터 조정**

  degree = 4
  alpha=10, max_iter=1000

  train_score : 0.9998
  test_score : 0.9555
  best_train_score : 0.9998
  best_test_score : 0.9555



3

003 모델 테스트

회귀모델

Decision Tree Regressor

- **기본 모델**

  train_score : 1.0
  test_score : -0.1714

- **기본모델 + 하이퍼파라미터 조정**

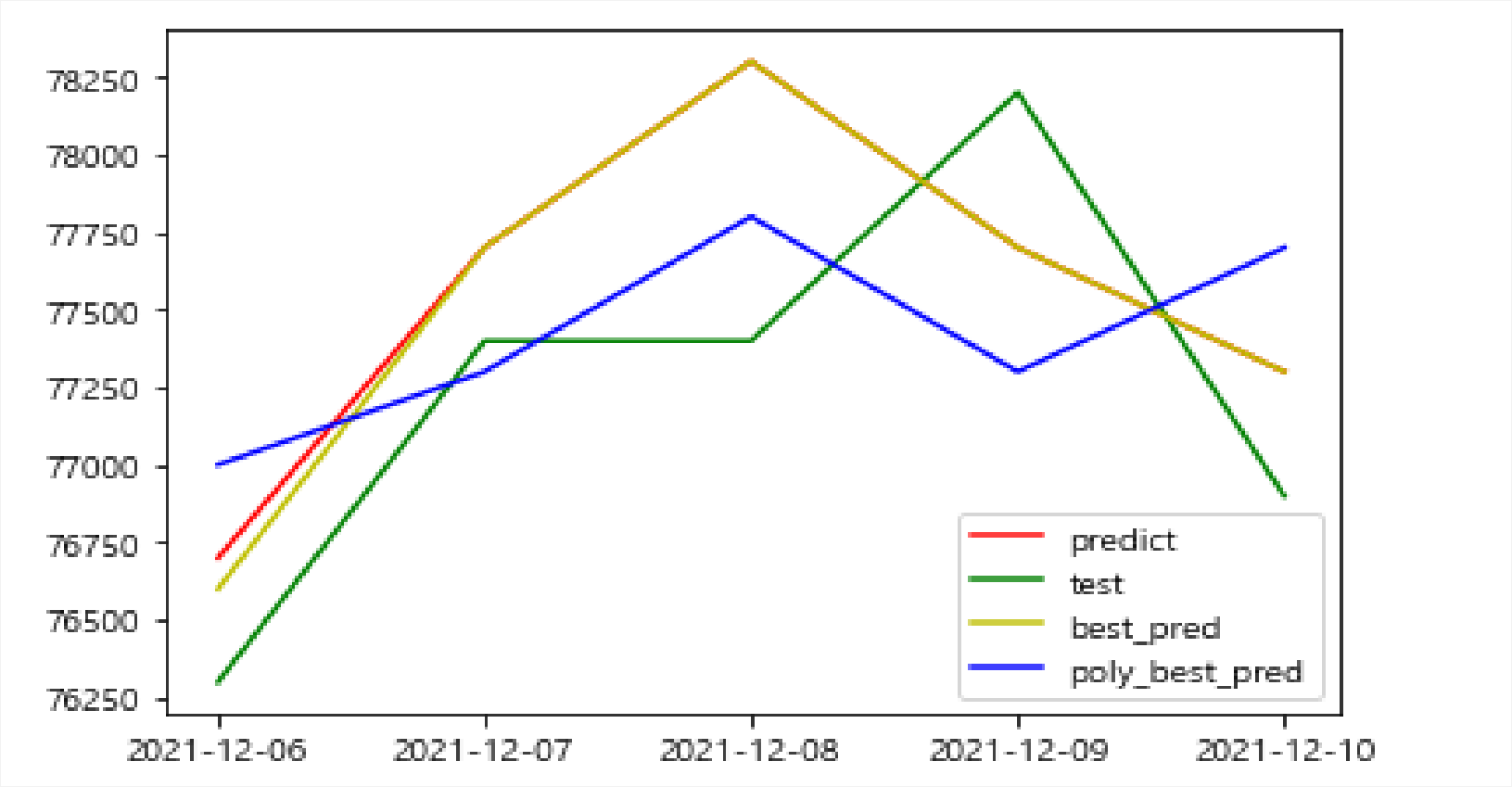
  max_depth=15

  train_score : 1.0
  test_score : 0.2901

- **다항화 + 하이퍼파라미터 조정**

  degree = 3
  max_depth=11

  train_score : 1.0
  test_score : -0.07
  best_train_score : 1.0
  best_test_score : -0.07



3

003 모델 테스트

회귀모델

Random Forest Regressor

- **기본 모델**

  train_score : 0.9999
  test_score : 0.3567

- **기본모델 + 하이퍼파라미터 조정**

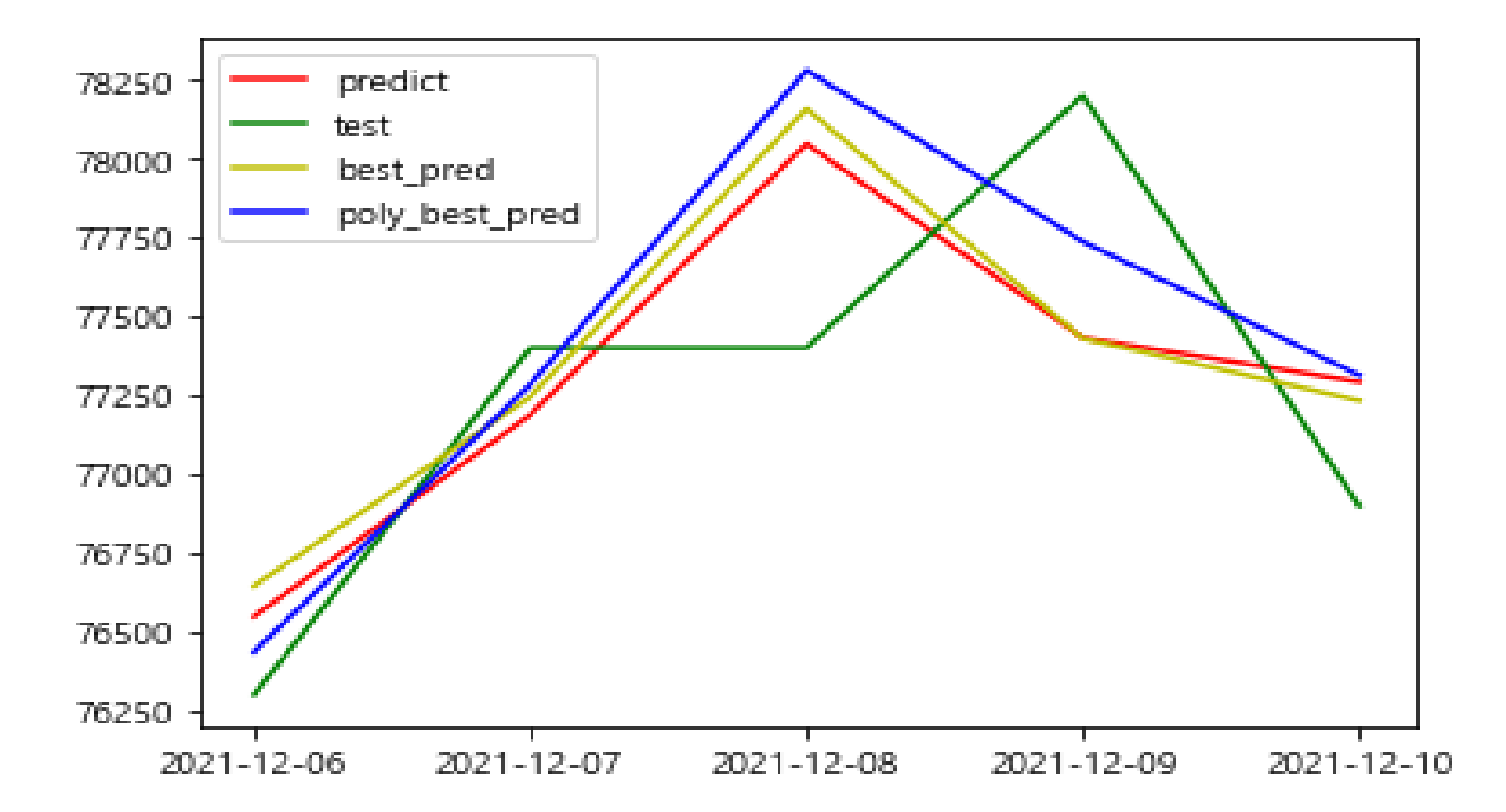
  max_depth=13, n_estimators=100

  train_score : 0.9999
  test_score : 0.2792

- **다항화 + 하이퍼파라미터 조정**

  degree = 4
  max_depth=11, n_estimators=100

  train_score : 0.9998
  test_score : 0.3959
  best_train_score : 0.9998
  best_test_score : 0.3959



3

003 모델 테스트

회귀모델

Gradient Boosting Regressor

- **기본 모델**

  train_score : 0.9999
  test_score : -0.1803

- **기본모델 + 하이퍼파라미터 조정**

  max_depth=13, learning_rate=0.1,
  n_estimators=5000

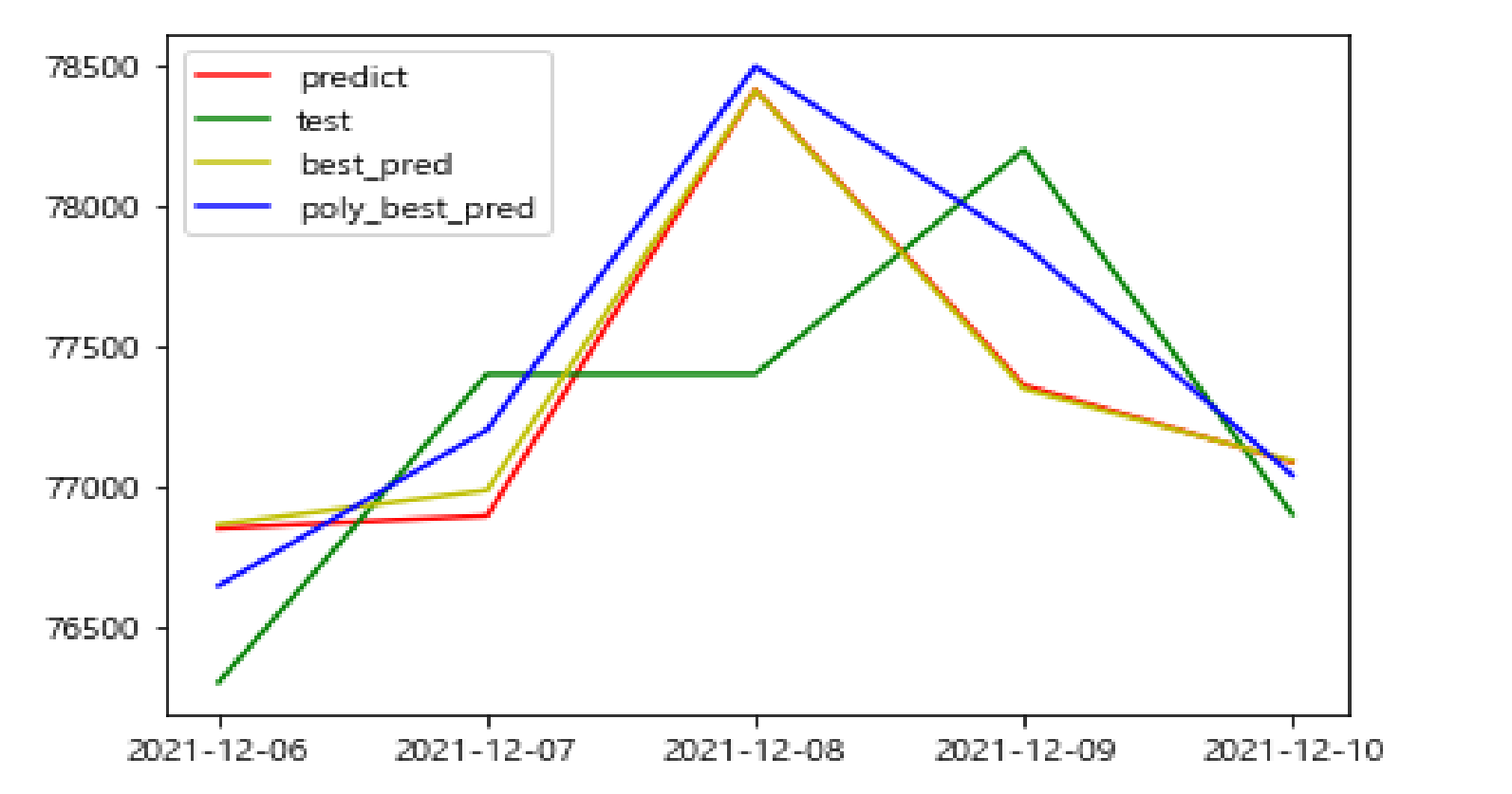
  train_score : 0.9999
  test_score : -0.1499

- **다항화 + 하이퍼파라미터 조정**

  degree = 2

  max_depth=3, learning_rate=1,
  n_estimators=100

  train_score : 1.0
  test_score : 0.2407
  best_train_score : 1.0
  best_test_score : 0.2407



3

003 모델 테스트

회귀모델

XGB Regressor

- **기본 모델**

  train_score : 0.9999
  test_score : -0.788

- **기본모델 + 하이퍼파라미터 조정**

  max_depth=11, learning_rate=0.1, n_estimators=5000

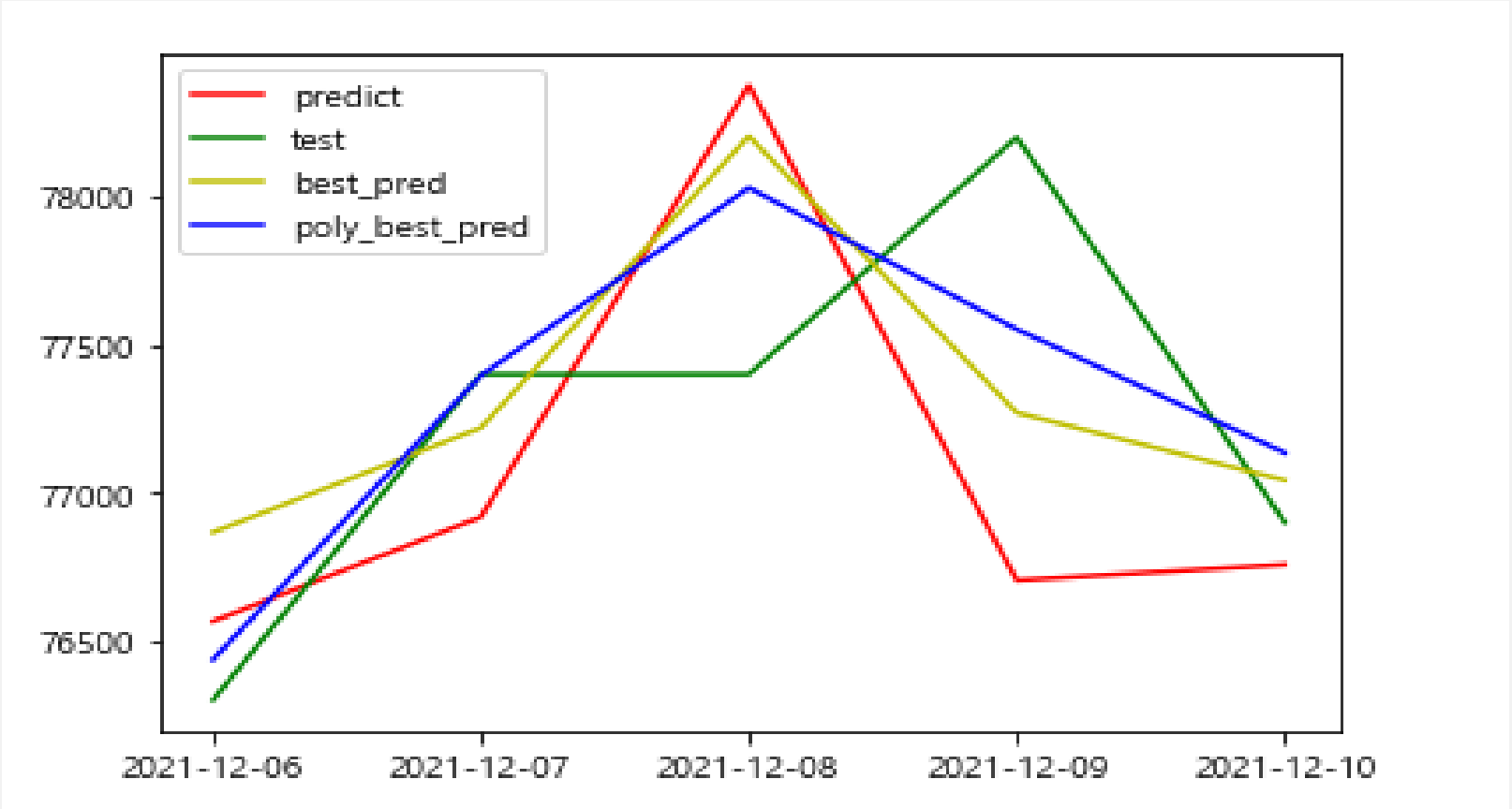
  train_score : 0.9999
  test_score : 0.0443

- **다항화 + 하이퍼파라미터 조정**

  degree = 4

  max_depth=11, learning_rate=0.01, n_estimators=1000

  train_score : 0.9999
  test_score : 0.5489
  best_train_score : 0.9999
  best_test_score : 0.5489

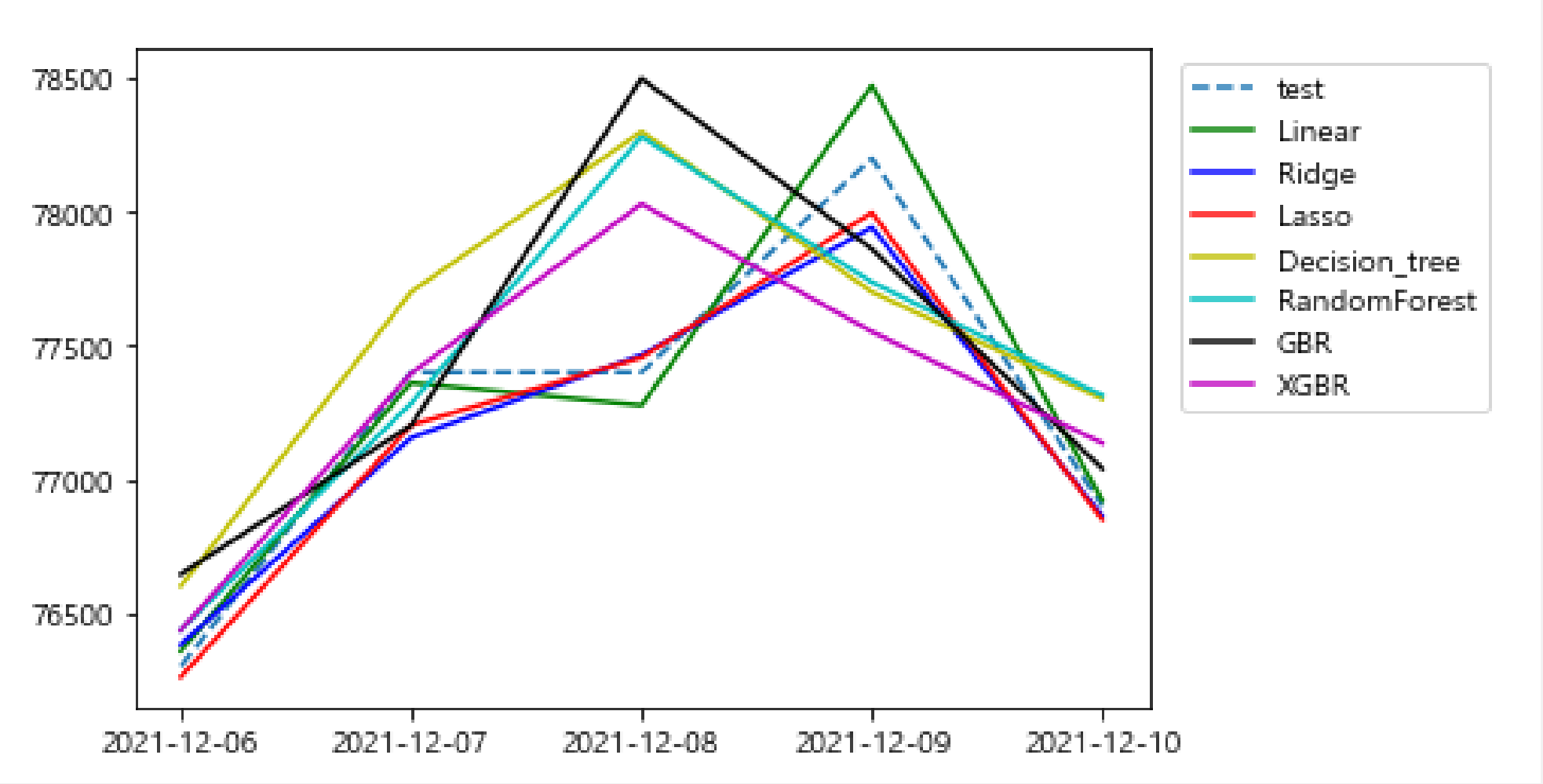


3

003 모델 테스트

회귀모델

각 모델 BEST 예측값 비교

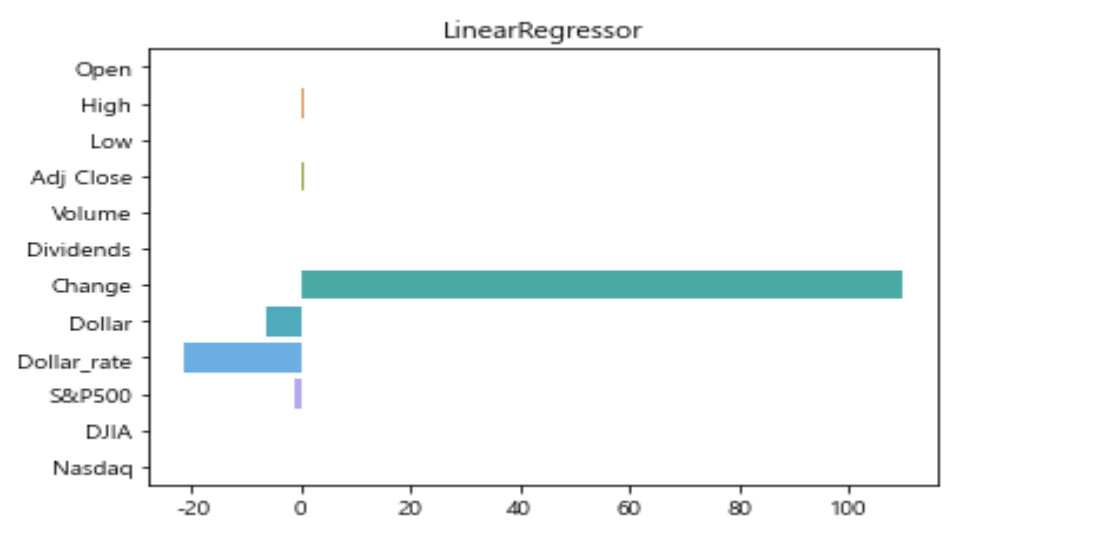


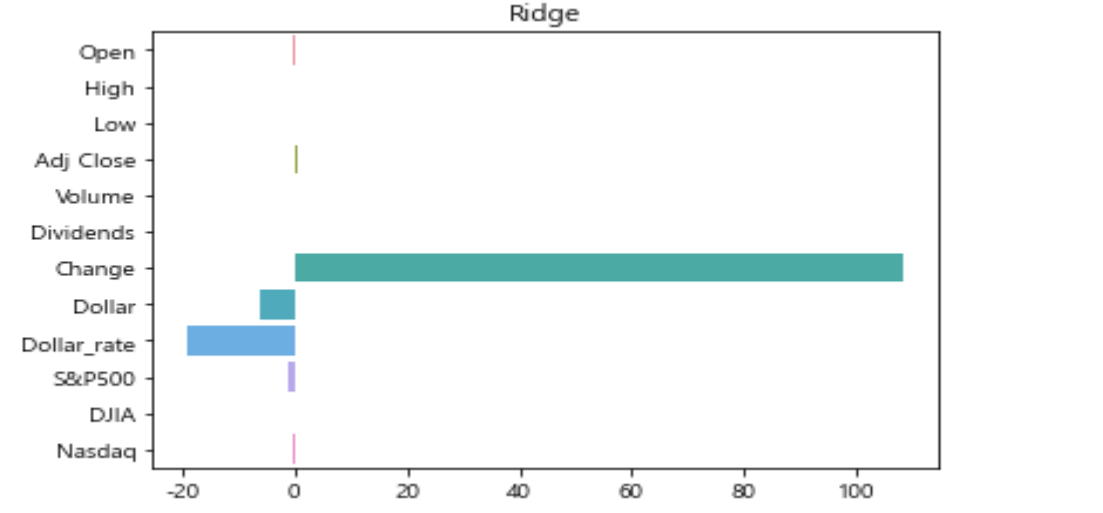
3

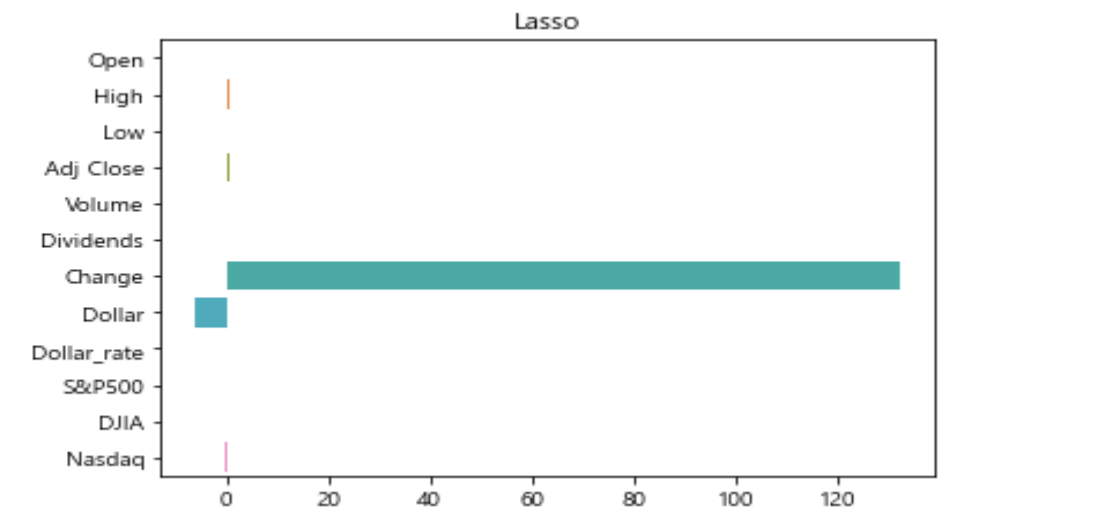
003 모델 테스트

회귀모델

각 모델 예측값 비교 | 회귀 계수 |





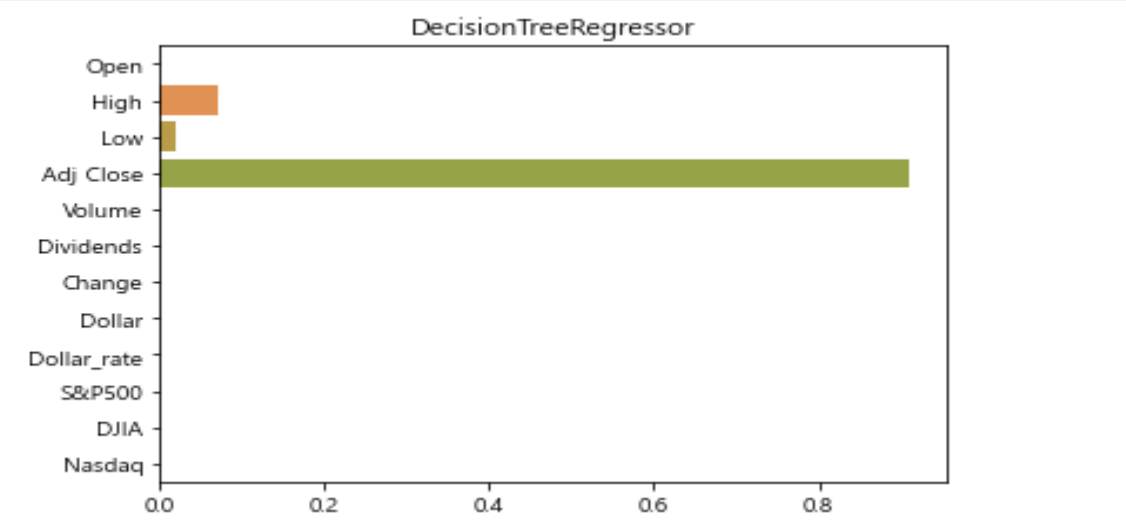


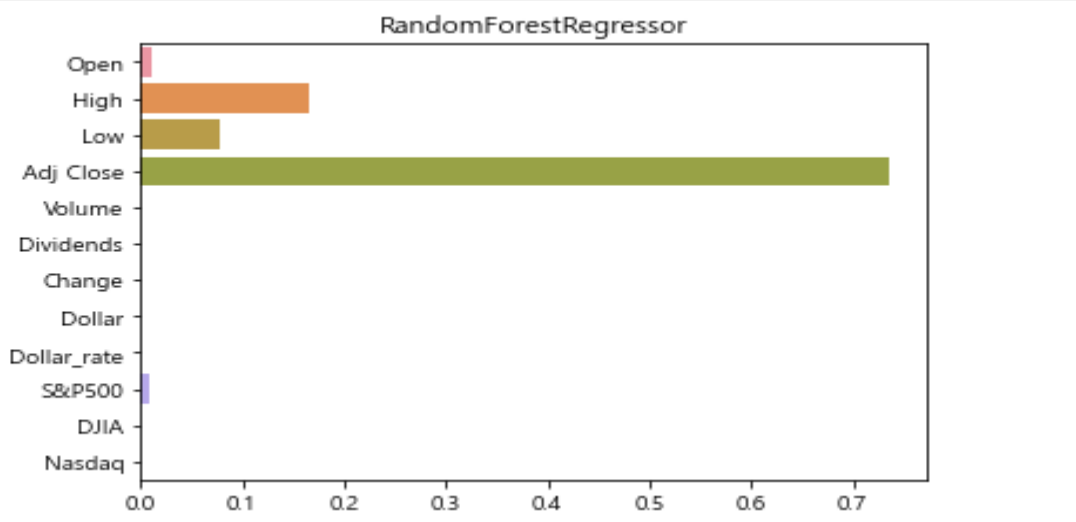
3

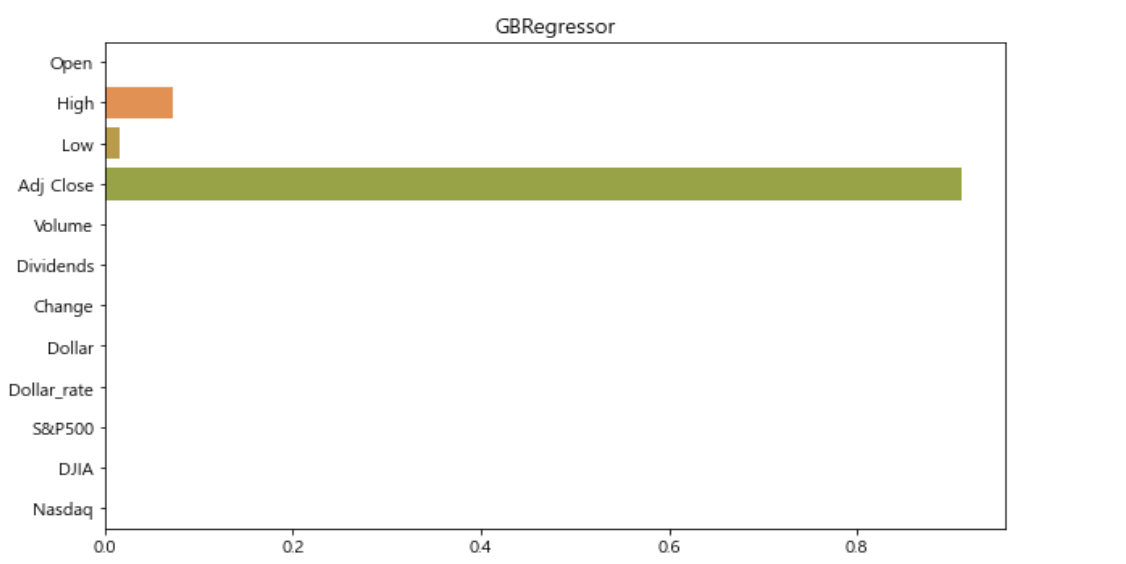
003 모델 테스트

회귀모델

각 모델 예측값 비교 | feature 중요도 |





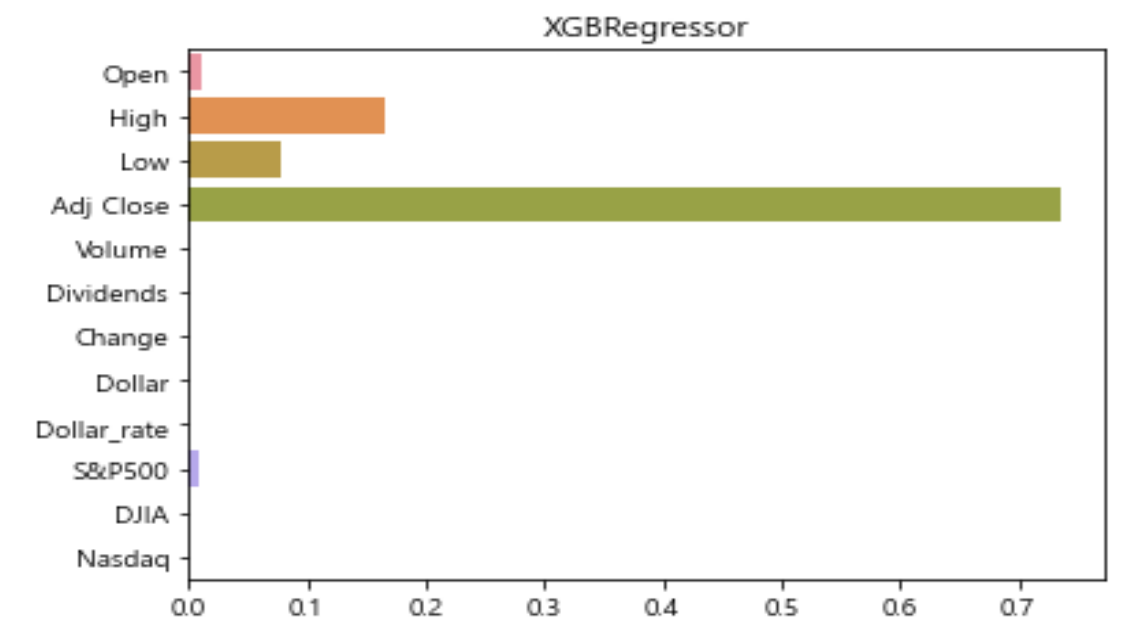


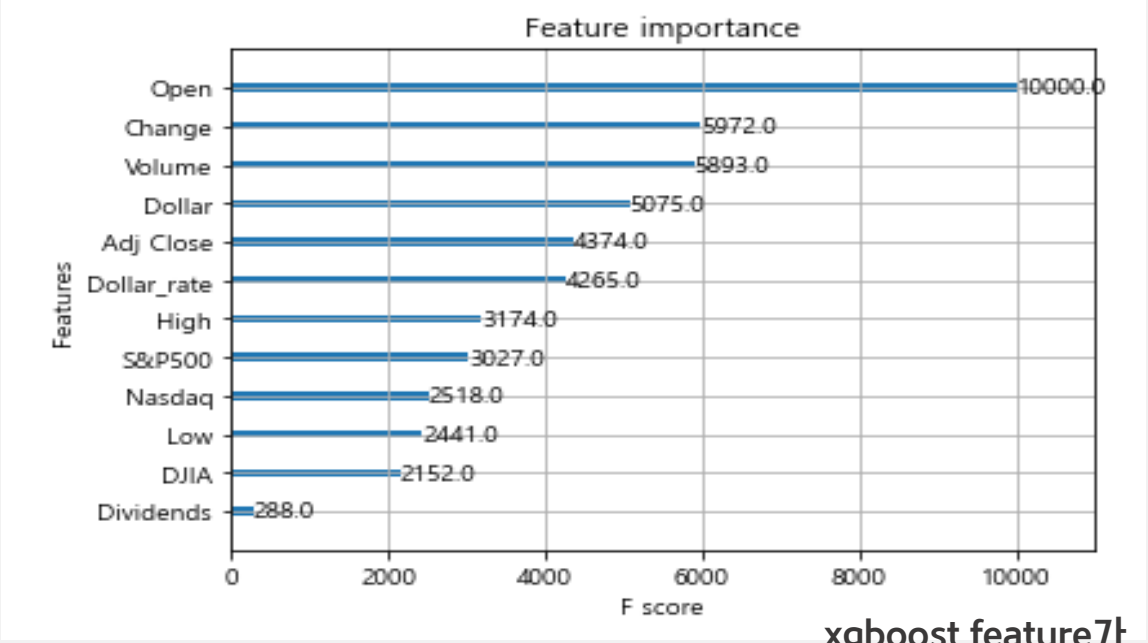
3

003 모델 테스트

회귀모델

각 모델 예측값 비교 | feature 중요도 |





xgboost feature가
노드 분할 시 사용된 횟수

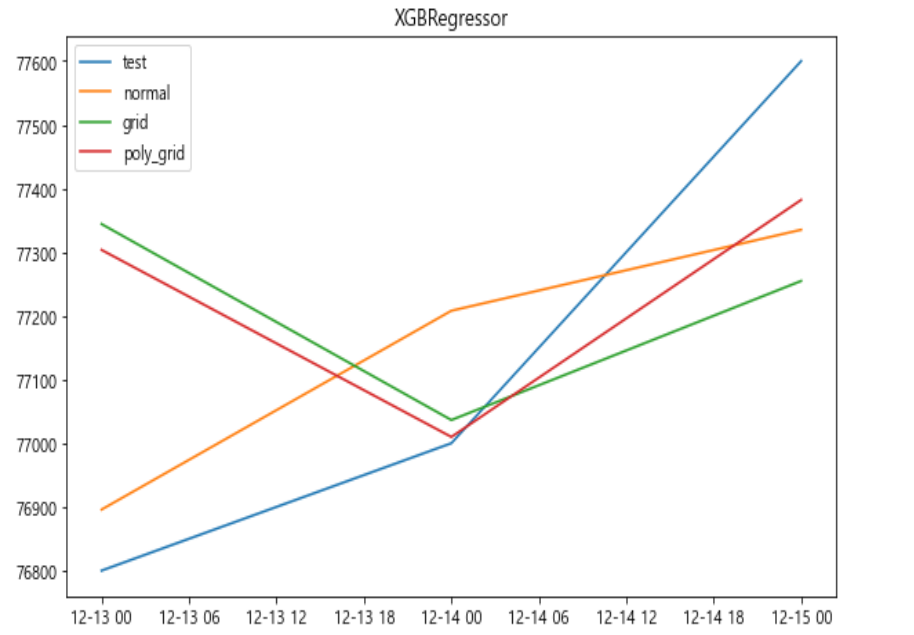
3

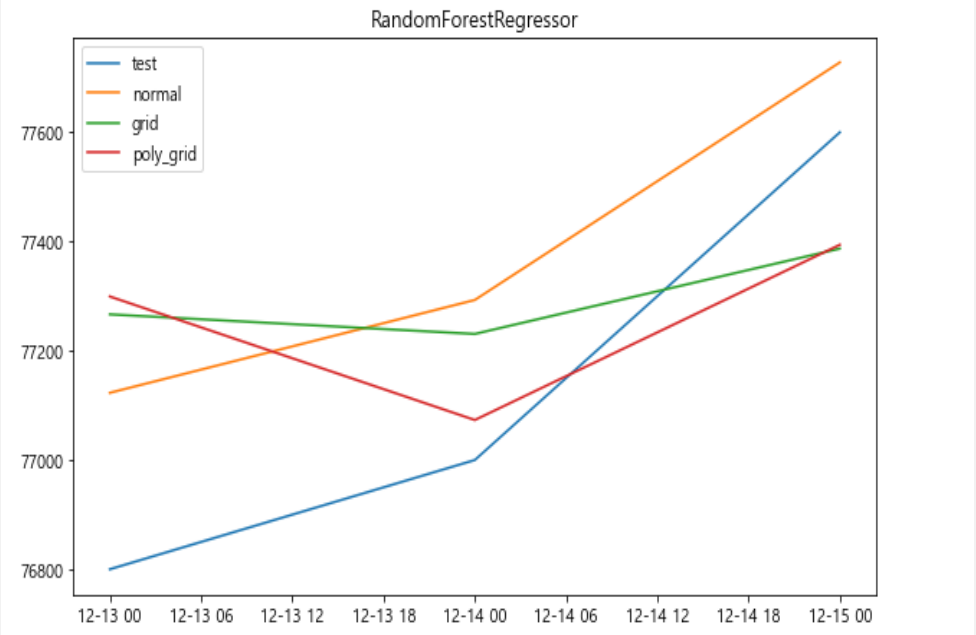
003 모델 테스트

회귀모델

Feature 확정

결정 트리 기반 회귀(test size : 이번 주)
Open, High, Low, Adj Close, S&P500 ->큰 변화 X






```
train_score : 0.999998599791151
test_score : 0.6463678740281324
[76895.81  77207.945 77335.1  ]
              Close
Date
2021-12-13  76800.0
2021-12-14  77000.0
2021-12-15  77600.0
```

```
Date
2021-12-13  76800.0
2021-12-14  77000.0
2021-12-15  77600.0
RandomForestRegressor()
train_score : 0.9996127035042935
test_score : 0.4520557692307692
[77045. 77323. 77760.]
```

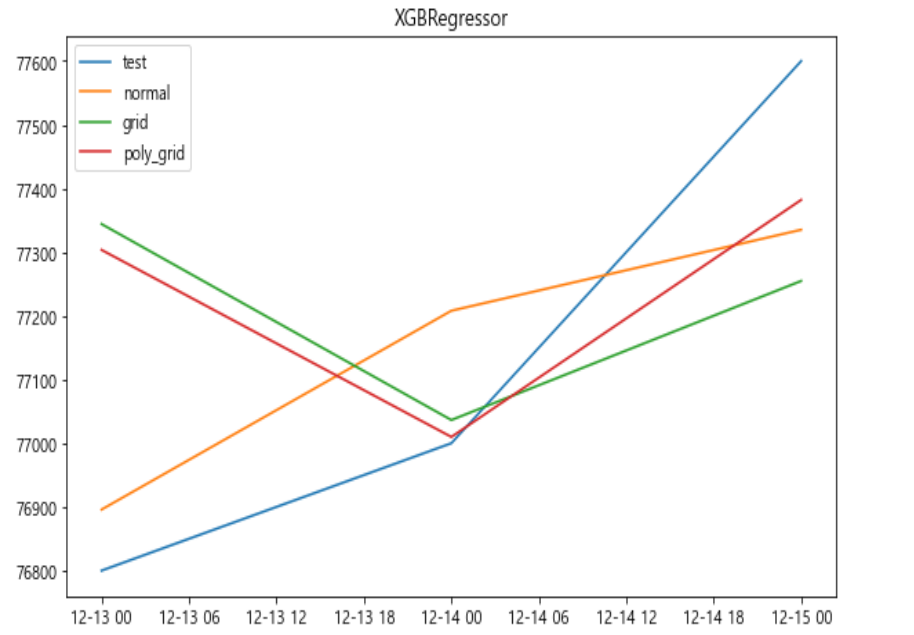
3

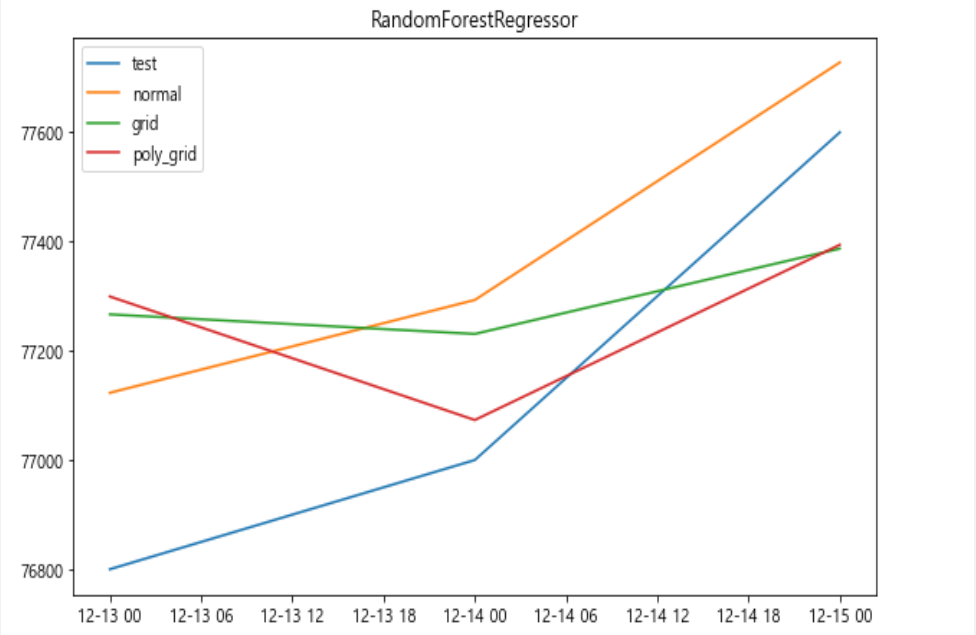
003 모델 테스트

회귀모델

Feature 확정

결정 트리 기반 회귀(test size : 이번 주)
Open, High, Low, Adj Close, S&P500 ->큰 변화 X






```
train_score : 0.999998599791151
test_score : 0.6463678740281324
[76895.81  77207.945 77335.1  ]
              Close
Date
2021-12-13  76800.0
2021-12-14  77000.0
2021-12-15  77600.0
```

```
Date
2021-12-13  76800.0
2021-12-14  77000.0
2021-12-15  77600.0
RandomForestRegressor()
train_score : 0.9996127035042935
test_score : 0.4520557692307692
[77045. 77323. 77760.]
```

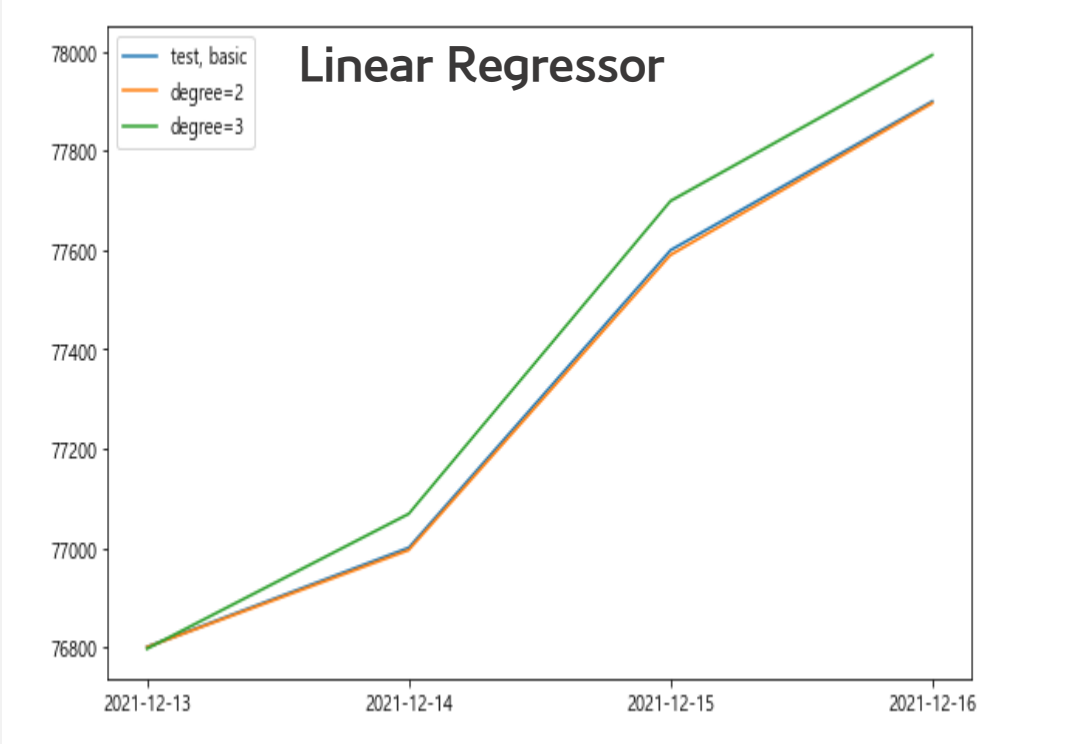
3

003 모델 테스트

회귀모델

Feature 확정 결정 트리 기반 회귀 X
모든 피쳐

test size : 이번 주




```
0.9999999556551397
0.9998374511672355
```

degree=2

```
0.9999999556551397
0.9998374511672355
[76799.487 76995.299 77590.349 77896.463]
Date
2021-12-13    76800.0
2021-12-14    77000.0
2021-12-15    77600.0
2021-12-16    77900.0
```

degree=3

```
0.9999973204007514
0.970437483086465
[76796.252 77068.358 77699.254 77993.5  ]
Date
2021-12-13    76800.0
2021-12-14    77000.0
2021-12-15    77600.0
2021-12-16    77900.0
```

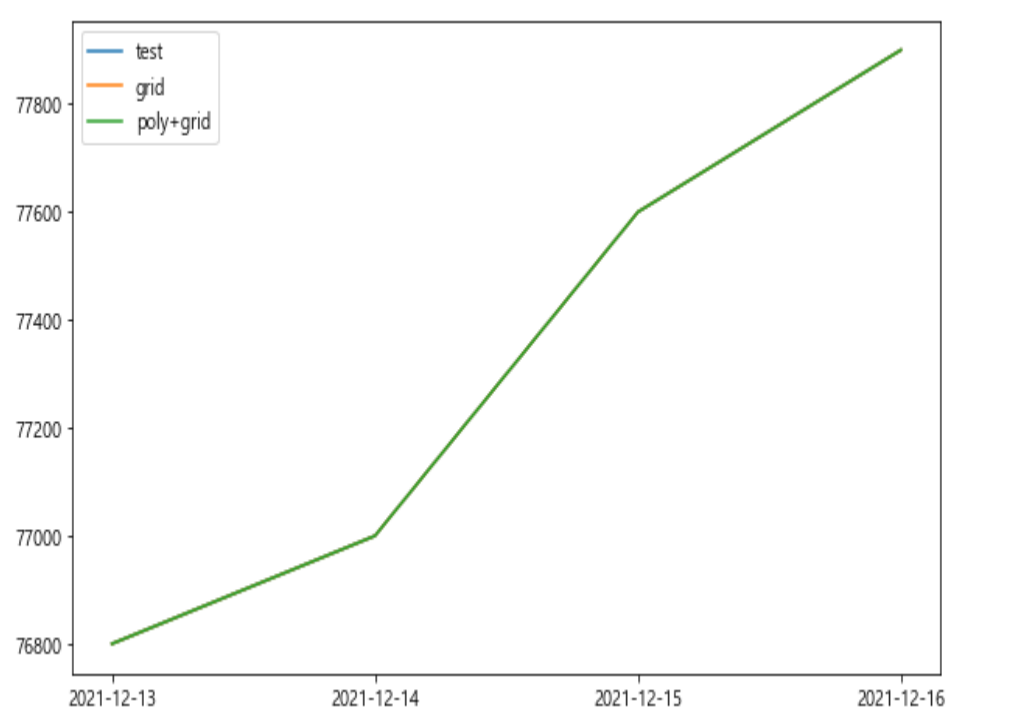
3

003 모델 테스트

회귀모델

Feature 확정 결정 트리 기반 회귀 X
모든 피쳐

test size : 이번 주




```
train_score : 1.0
test_score : 0.9999999999999997
[76800. 77000. 77600. 77900.]
```

grid

```
best model = Ridge(alpha=0.001, max_iter=500, random_state=17)
1.0
1.0
[76800. 77000. 77600. 77900.]
Date
2021-12-13    76800.0
2021-12-14    77000.0
2021-12-15    77600.0
2021-12-16    77900.0
```

degree=2
poly+grid

```
best model = Ridge(alpha=5, max_iter=500, random_state=17)
0.999999999996525
0.9999999975322712
[76800.004 76999.989 77599.958 77900.004]
Date
2021-12-13    76800.0
2021-12-14    77000.0
2021-12-15    77600.0
2021-12-16    77900.0
```

degree=3부터 테스터 성능 마이너스

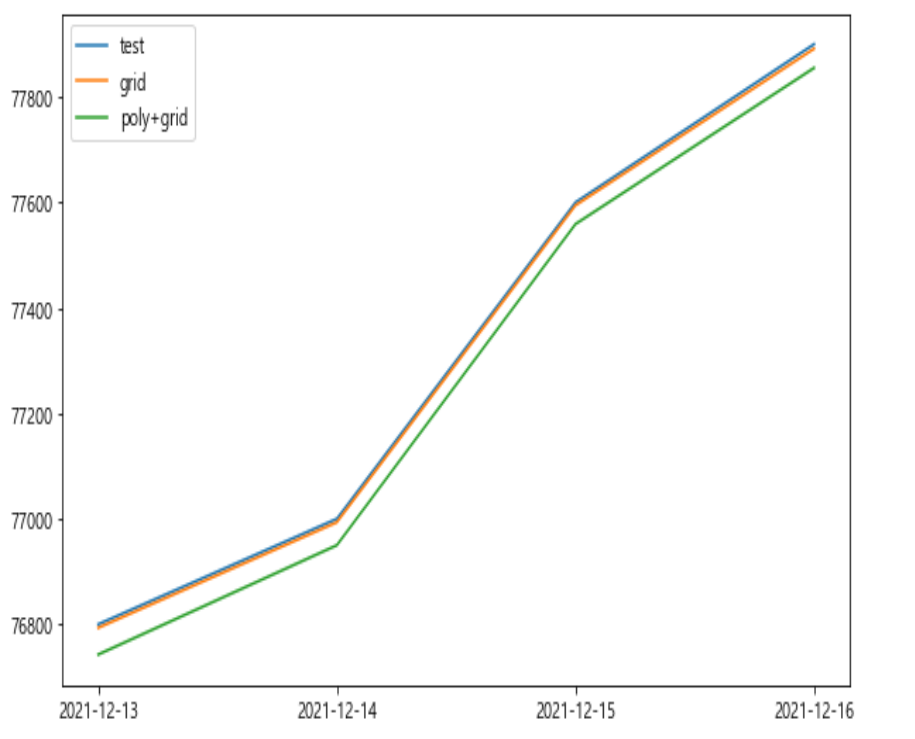
3

003 모델 테스트

회귀모델

Feature 확정 결정 트리 기반 회귀 X
모든 피쳐

test size : 이번 주




```
train_score : 0.9997285990368453
test_score : 0.8313470984161101
[77088.414 76871.406 77423.82  77945.339]
```

grid

```
best model = Lasso(alpha=0.1, max_iter=10000, random_state=17)
0.99999993067031
0.9997593899704372
[76793.401 76993.521 77594.413 77891.471]
```

**degree=2
poly+grid**

```
best model = Lasso(alpha=1, max_iter=10000, random_state=17)
0.9999865786273807
0.9879611736528096
[76742.756 76950.013 77558.925 77855.079]
Date
2021-12-13    76800.0
2021-12-14    77000.0
2021-12-15    77600.0
2021-12-16    77900.0
```

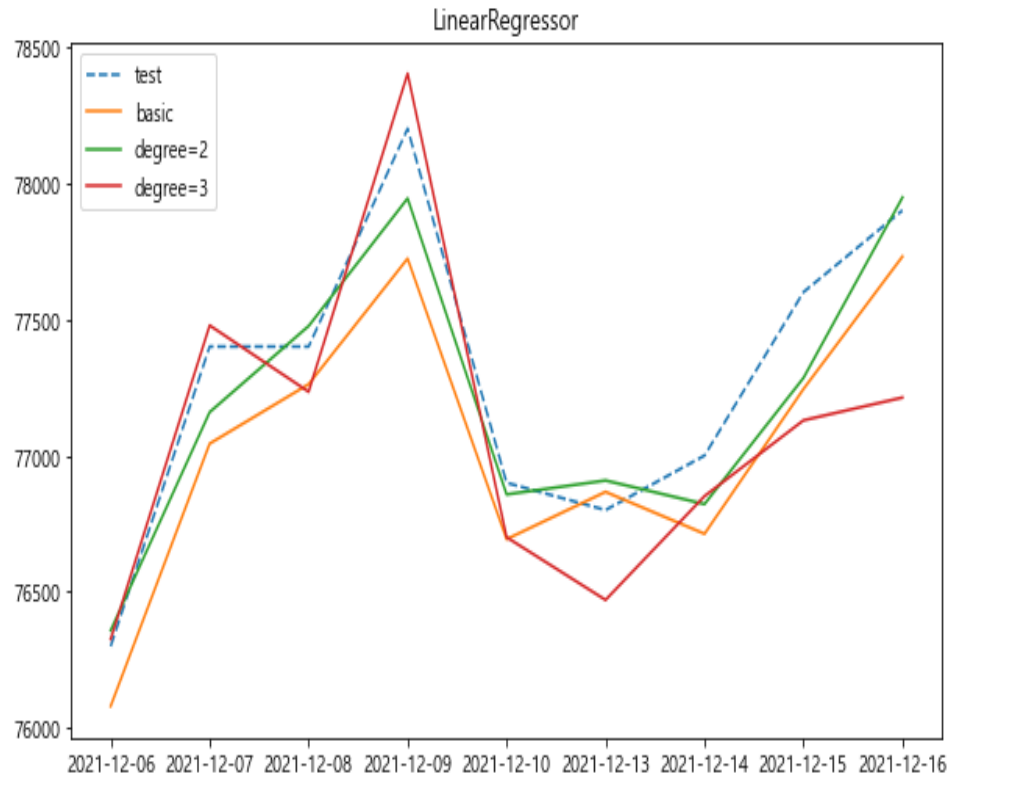
3

003 모델 테스트

회귀모델

Feature 확정 결정 트리 기반 회귀 X
모든 피쳐

test size : 저번 주 + 이번 주




```
train_score : 0.9996764865801158
test_score : 0.7453887808001666
[76080.403 77043.799 77261.89  77723.043 76692.998 76866.662 76712.776
 77244.253 77730.659]
```

degree=2

```
0.9998782087282467
0.8993836340202186
[76358.374 77159.295 77475.486 77943.977 76857.356 76908.457 76821.831
 77285.201 77947.429]
Date
2021-12-06    76300.0
2021-12-07    77400.0
2021-12-08    77400.0
2021-12-09    78200.0
2021-12-10    76900.0
2021-12-13    76800.0
2021-12-14    77000.0
2021-12-15    77600.0
2021-12-16    77900.0
```

degree=3은 테스트 성능 60%, 4부터 마이너스

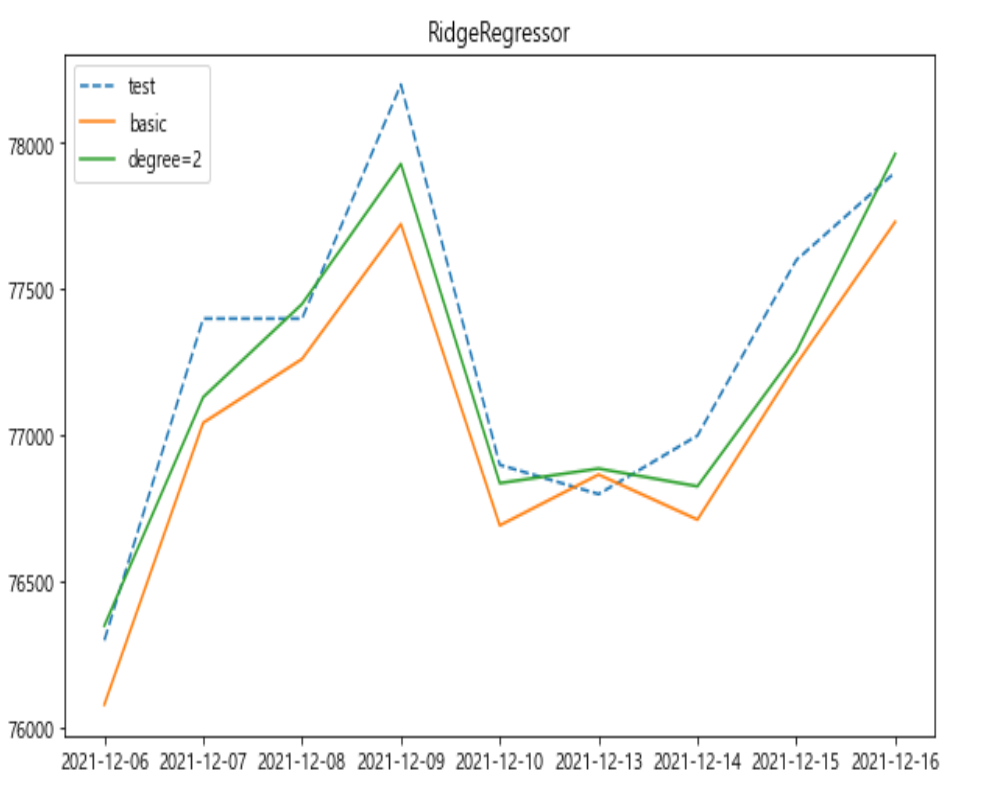
3

003 모델 테스트

회귀모델

Feature 확정 결정 트리 기반 회귀 X
모든 피쳐

test size : 저번 주 + 이번 주




```
Ridge()
train_score : 0.9996764862942662
test_score : 0.7452024323699331
[76080.376 77043.823 77261.521 77722.738 76692.959 76866.577 76712.66
 77244.3   77730.388]
```

degree=2

```
best model = Ridge(alpha=10, max_iter=500, random_state=17)
0.9998818311196693
0.8940935859254103
[76350.123 77131.355 77449.879 77928.823 76837.461 76887.254 76826.639
 77287.441 77963.011]
Date
2021-12-06    76300.0
2021-12-07    77400.0
2021-12-08    77400.0
2021-12-09    78200.0
2021-12-10    76900.0
2021-12-13    76800.0
2021-12-14    77000.0
2021-12-15    77600.0
2021-12-16    77900.0
```

degree=3부터 테스트 성능 마이너스
교차 검증만 했을 때, 74%

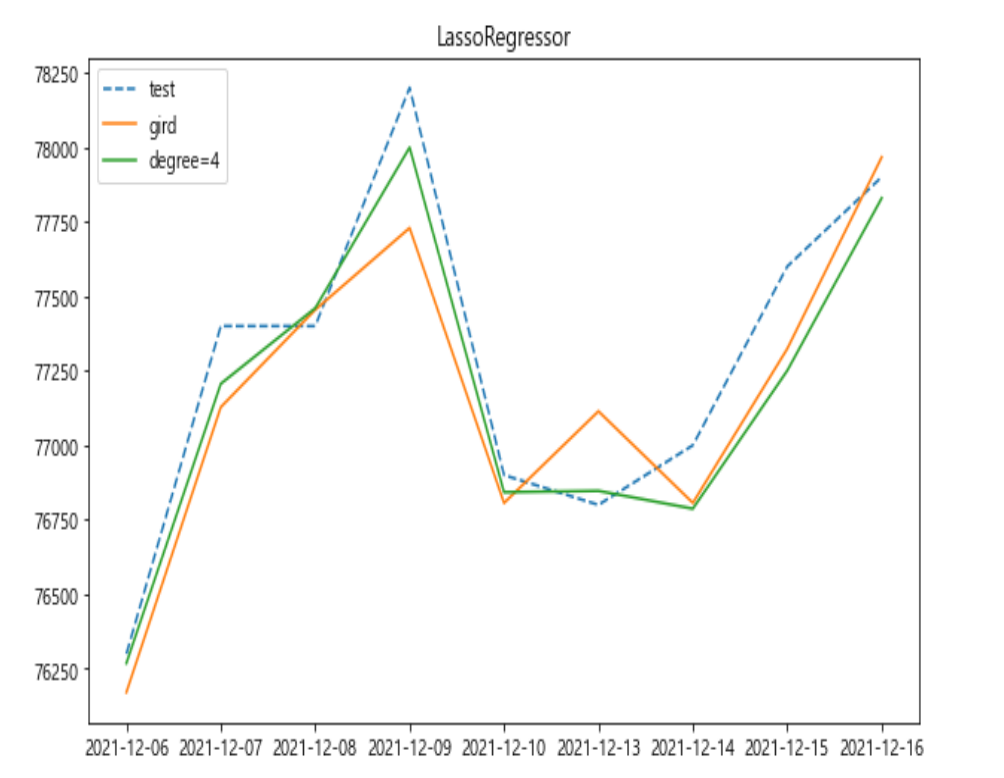
3

003 모델 테스트

회귀모델

Feature 확정 결정 트리 기반 회귀 X
모든 피쳐

test size : 저번 주 + 이번 주




```
Lasso()
train_score : 0.999612023643904
test_score : 0.8051715765208904
[76171.218 77127.962 77454.555 77729.16  76806.222 77114.089 76806.648
 77323.856 77966.439]
```

degree=4

```
best model = Lasso(alpha=1, random_state=17)
0.9998415351186851
0.9061976090993786
[76268.815 77206.378 77459.977 77999.318 76842.521 76846.978 76787.115
 77250.477 77829.118]
Date
2021-12-06    76300.0
2021-12-07    77400.0
2021-12-08    77400.0
2021-12-09    78200.0
2021-12-10    76900.0
2021-12-13    76800.0
2021-12-14    77000.0
2021-12-15    77600.0
2021-12-16    77900.0
```

degree=2, 3부터 테스트 성능 89%

3

003 모델 테스트

회귀모델

금일 종가 예측 21/12/16 15시 삼성 데이터 -> 종가 : 77,800원

- 대체적으로 77,500 ~ 77,900

Date	Open	High	Low	Close	Adj Close	Volume
2021-12-13	77200.0	78300.0	76500.0	76800.0	76800.0	15038750
2021-12-14	76500.0	77200.0	76200.0	77000.0	77000.0	10976660
2021-12-15	76400.0	77600.0	76300.0	77600.0	77600.0	9584939
2021-12-16	78500.0	78500.0	77400.0	77600.0	77600.0	9826263

Date	Open	High	Low	Adj Close	Volume	Dividends	Change	Dollar	Dollar_rate	S&P500	DJIA	Nasdaq
2021-12-06	75100.0	76700.0	74900.0	76300.0	16391250.0	0	0.93	1178.6	0.11	4538.430176	34580.078130	15085.469730
2021-12-07	76100.0	77700.0	75600.0	77400.0	19232453.0	0	1.44	1183.7	0.43	4591.669922	35227.031250	15225.150390
2021-12-08	78300.0	78600.0	77100.0	77400.0	21558340.0	0	0.00	1181.2	-0.21	4686.750000	35719.429690	15686.919920
2021-12-09	77400.0	78200.0	77000.0	78200.0	21604528.0	0	1.03	1176.4	-0.41	4701.209961	35754.750000	15786.990230
2021-12-10	77400.0	77600.0	76800.0	76900.0	9155219.0	0	-1.66	1173.8	-0.22	4667.450195	35754.691410	15517.370120
2021-12-13	77200.0	78300.0	76500.0	76800.0	15038750.0	0	-0.13	1178.2	0.37	4712.020020	35970.988281	15630.599609
2021-12-14	76500.0	77200.0	76200.0	77000.0	10976660.0	0	0.26	1178.8	0.05	4668.970215	35650.949219	15413.280273
2021-12-15	76400.0	77600.0	76300.0	77600.0	9584939.0	0	0.78	1183.6	0.41	4634.089844	35544.179688	15237.639648
2021-12-16	78500.0	78500.0	77400.0	77600.0	9826263.0	0	0.13	1186.2	0.22	4709.850098	35927.429688	15565.583008

test size : 저번 주 + 이번 주
Lasso
degree=2 -> 85% -> 77553.125
degree=3 -> 87% -> 77722.842
degree=4 -> 86% -> 77571.385

test size : 이번 주
LinearRegressor
Degree=2 -> 99% -> 77894.364
degree=3 -> 95% -> 78057.308
Lasso
degree=3 -> 97% -> 77844.829
degree=4 -> 95% -> 77795.626

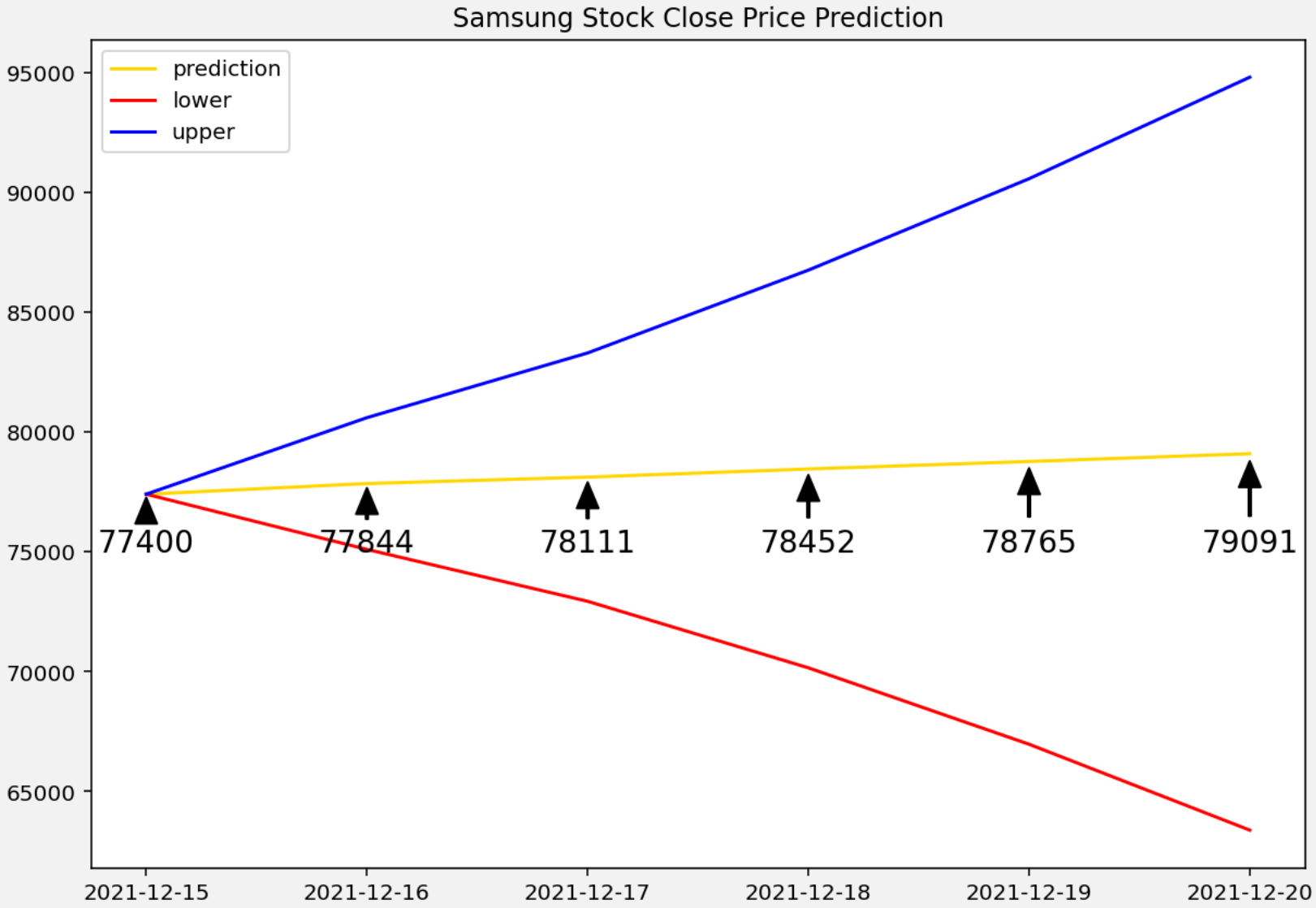
Ridge는 degree 3부터 마이너스 성능

3

003 모델 테스트

ARIMA모델

ARIMA 예측 결과



3

003 모델 테스트

ARIMA모델

ARIMA 소개

Autoregressive Integrated Moving Average 라는 뜻으로,
AR(Auto Regression) 모형과, MA(Moving Average) 모형을 합친 모형이다.

AR 모형이란?

자귀 회귀 모형으로, Auto Correlation의 약자이다.
자기상관성을 시계열 모형으로 구성하였으며, 예측하고자 하는 특정 변수의 과거 관측값의 선형결합으로 해당 변수의 미래값을 예측하는 모형이다.
이전 자신의 관측값이 이후 자신의 관측값에 영향을 준다는 아이디어에 기반하였다.

AR(p) 모형의 식은 다음과 같다.

$$y_t = c + \phi_1 y_{t-1} + \phi_2 y_{t-2} + \ldots + \phi_p y_{t-p} + \varepsilon_t$$

y_t는 t시점의 관측값, c는 상수, ϕ는 가중치, ε_t는 오차항을 의미한다.

MA 모형이란?

Moving Average 모형으로, 예측 오차를 이용하여 미래를 예측하는 모형이다.

MA(q) 모형의 식은 다음과 같다.

$$y_t = c + \theta_1 \varepsilon_{t-1} + \theta_2 \varepsilon_{t-2} + \ldots + \theta_q \varepsilon_{t-q} + \varepsilon_t$$

3

003 모델 테스트

ARIMA모델

ARIMA 사용방법

```
from statsmodels.tsa.arima_model import ARIMA
import statsmodels.api as sm
```

```
model = ARIMA(samsung_train_df.price.values, order = (1,2,0))
model_fit = model.fit()
print((1,2,0))
print(model_fit.summary())
```

```
(1, 2, 0)
                             ARIMA Model Results
==============================================================================
Dep. Variable:                   D2.y   No. Observations:                  242
Model:                 ARIMA(1, 2, 0)   Log Likelihood               -2097.048
Method:                       css-mle   S.D. of innovations           1402.744
Date:                Thu, 16 Dec 2021   AIC                           4200.096
Time:                        01:33:13   BIC                           4210.563
Sample:                             2   HQIC                          4204.313

==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
const          1.3666     64.391      0.021      0.983    -124.838     127.571
ar.L1.D2.y    -0.4020      0.059     -6.847      0.000      -0.517      -0.287
                                    Roots
=============================================================================
                  Real          Imaginary           Modulus         Frequency
-----------------------------------------------------------------------------
AR.1           -2.4873           +0.0000j            2.4873            0.5000
-----------------------------------------------------------------------------
```

Argument (p, d, q) 필요!!

AR항 (p, d, 0) MA항(0, d, q)

AR항과 MA항은 서로 상쇄하는 기능이 있으므로 AR항과MR항을 동시에 사용할 때는 주의해야 된다.

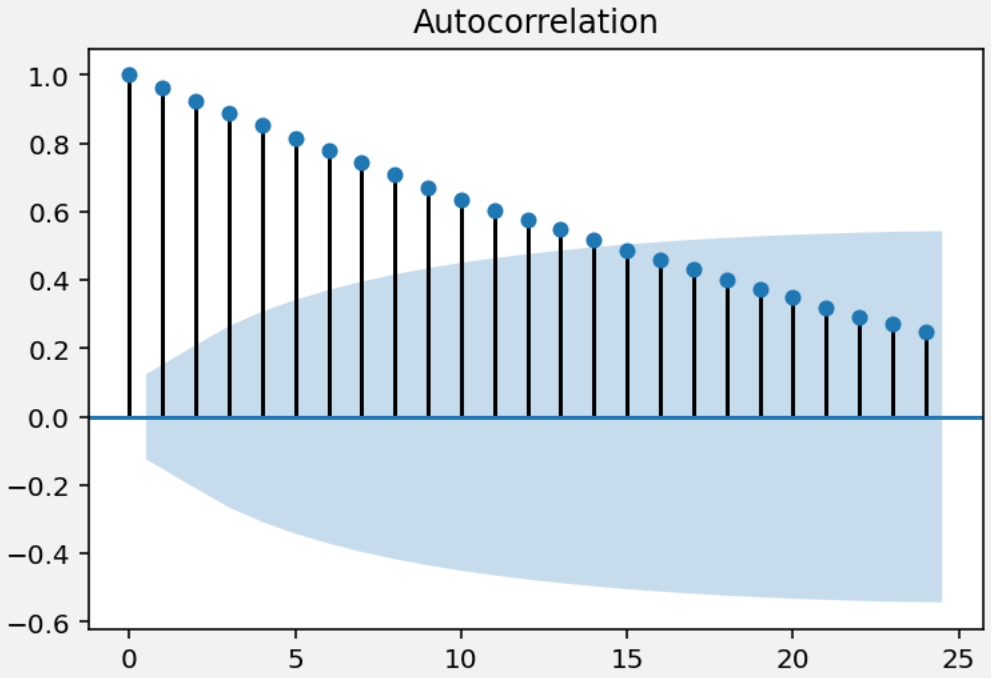
3

003 모델 테스트

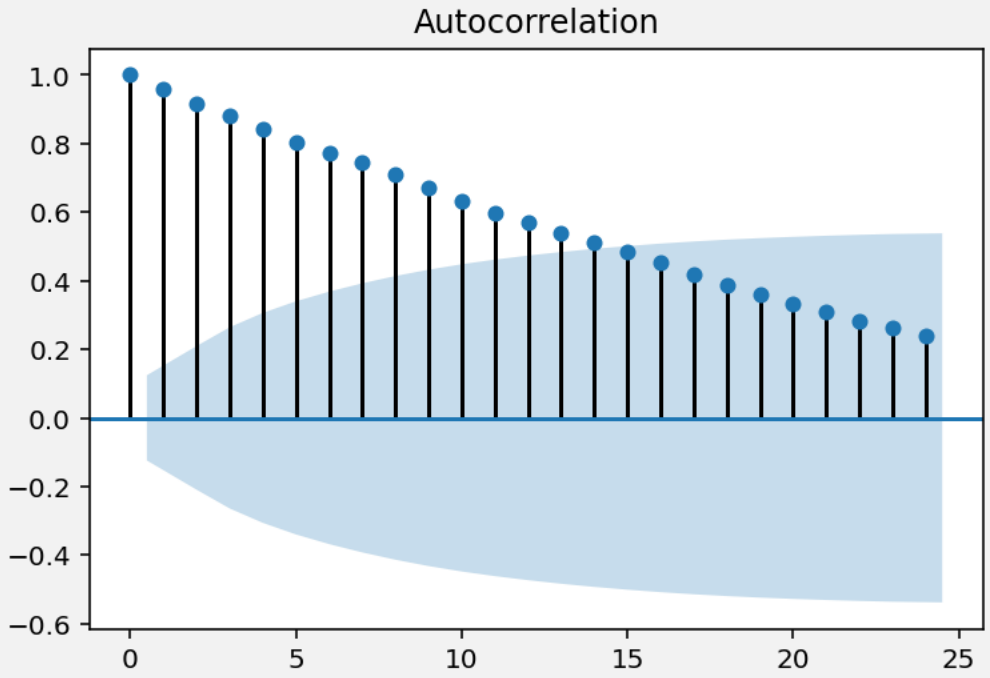
ARIMA모델

ARIMA 모델 생성 | p,d,q값 설정 근거 |

ACF 그래프



2020-11-02 ~ 2021-11-03



2020-11-09 ~ 2021-11-16

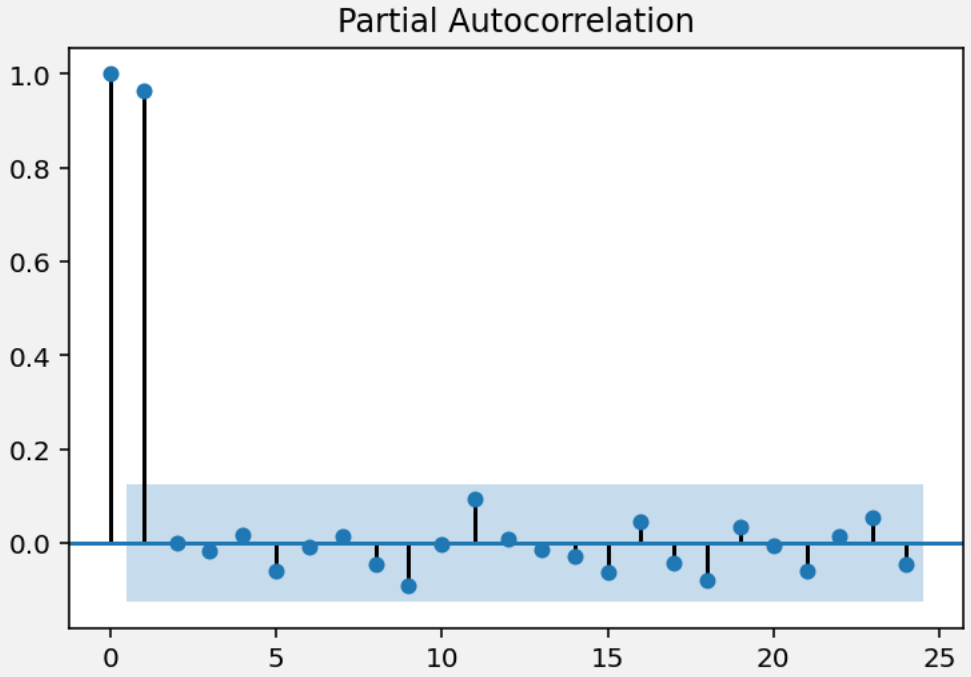
3

003 모델 테스트

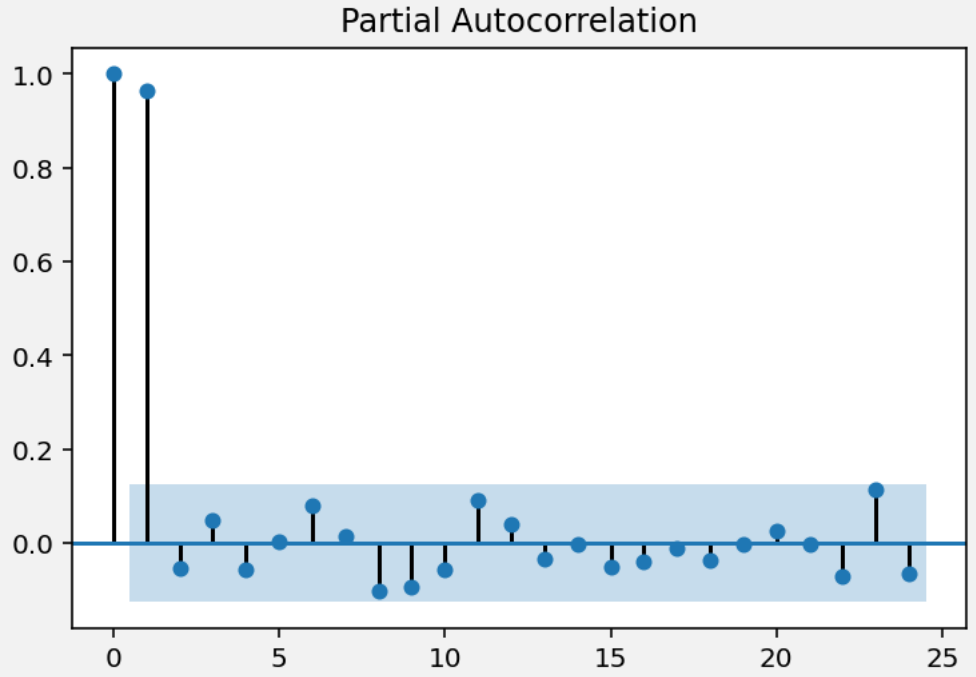
ARIMA모델

ARIMA 모델 생성 | p,d,q값 설정 근거 |

PACF 그래프



2020-11-02 ~ 2021-11-03



2020-11-09 ~ 2021-11-16

3

003 모델 테스트

ARIMA모델

ARIMA 모델 생성 | p,d,q값 search |

2020-11-02 ~ 2021-11-03

(1, 1, 0)

ARIMA Model Results

Dep. Variable:	D.y	No. Observations:	243
Model:	ARIMA(1, 1, 0)	Log Likelihood	-2058.906
Method:	css-mle	S.D. of innovations	1157.383
Date:	Tue, 14 Dec 2021	AIC	4123.813
Time:	16:36:07	BIC	4134.292
Sample:	1	HQIC	4128.034

	coef	std err	z	P>|z|	[0.025	0.975]
const	52.3069	79.989	0.654	0.514	-104.468	209.082
ar.L1.D.y	0.0721	0.064	1.124	0.262	-0.054	0.198

Roots

	Real	Imaginary	Modulus	Frequency
AR.1	13.8717	+0.0000j	13.8717	0.0000

(1, 2, 0)

	coef	std err	z	P>|z|
const	-5.4718	64.907	-0.084	0.933
ar.L1.D2.y	-0.4216	0.058	-7.213	0.000

2020-11-09 ~ 2021-11-16
(1, 1, 0) (1, 2, 0)

P>|z|
0.600
0.291

P>|z|
0.968
0.000

2020-11-16 ~ 2021-11-17
(1, 1, 0) (1, 2, 0)

P>|z|
0.838
0.366

P>|z|
0.977
0.000

2020-11-23 ~ 2021-11-24
(1, 1, 0) (1, 2, 0)

P>|z|
0.862
0.385

P>|z|
0.986
0.000

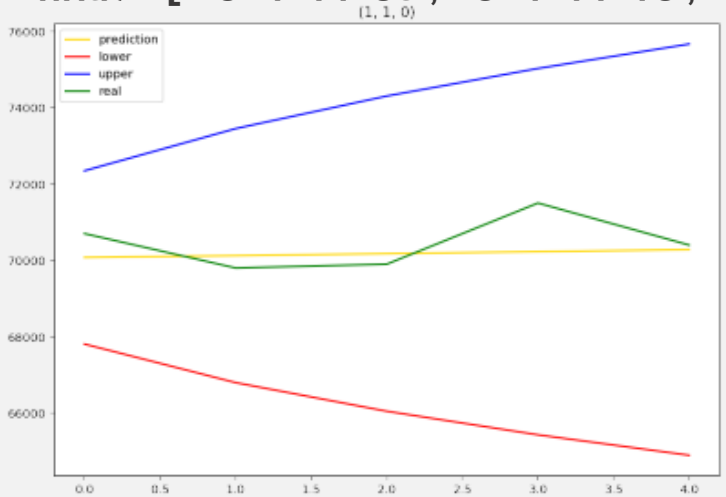
3

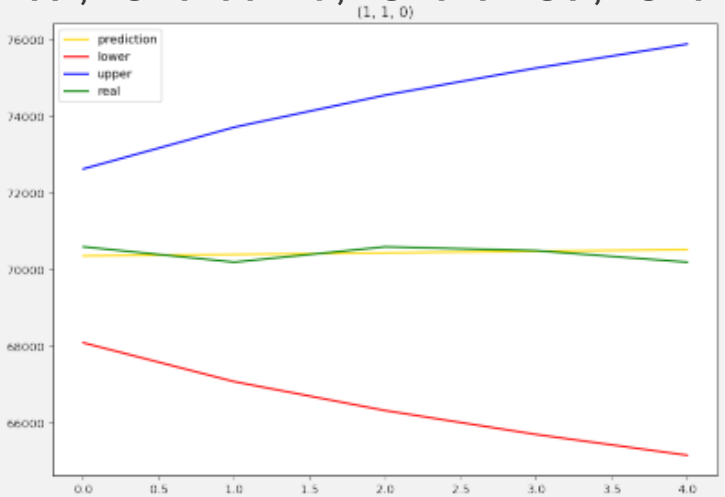
003 모델 테스트

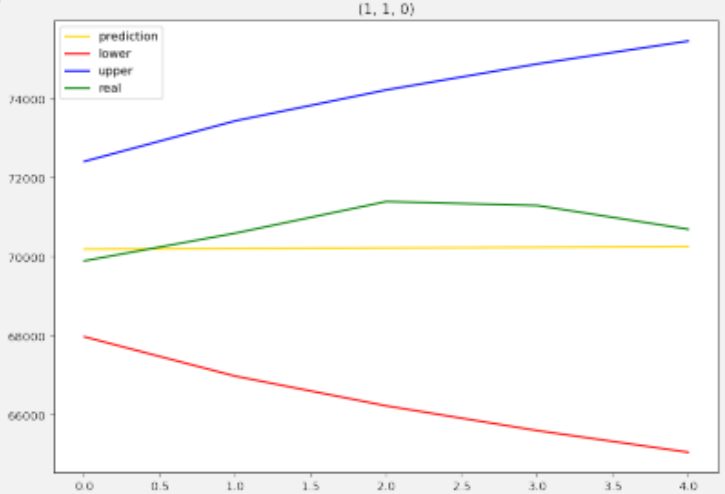
ARIMA모델

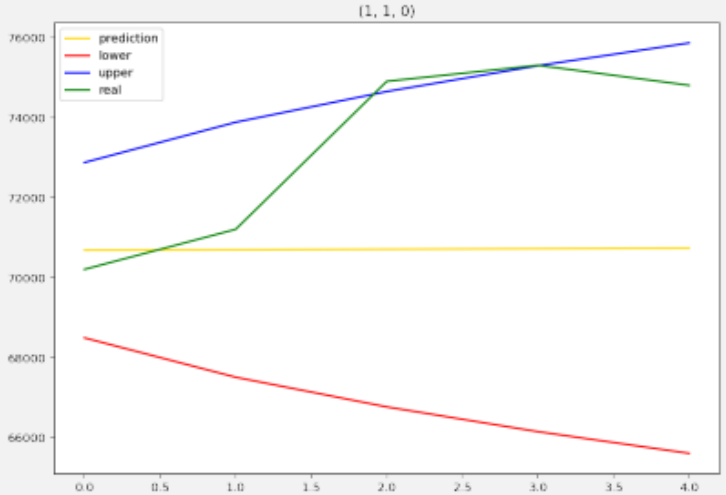
ARIMA 샘플 데이터 예측 결과 | (1, 1, 0) |

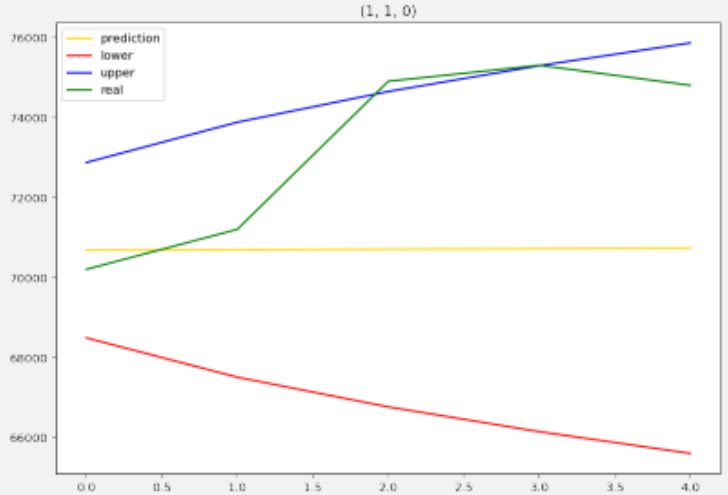
start = ['2020-11-02','2020-11-09','2020-11-16','2020-11-23','2020-11-30','2020-12-07']
final = ['2021-11-03','2021-11-10','2021-11-17','2021-11-24','2021-12-01','2021-12-08']

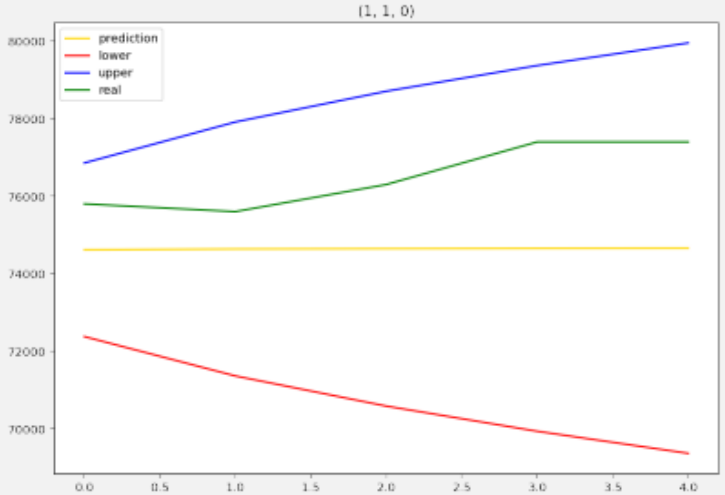












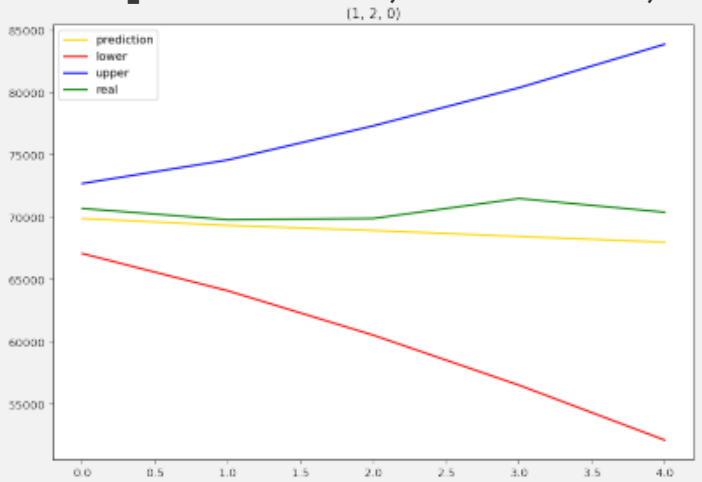
3

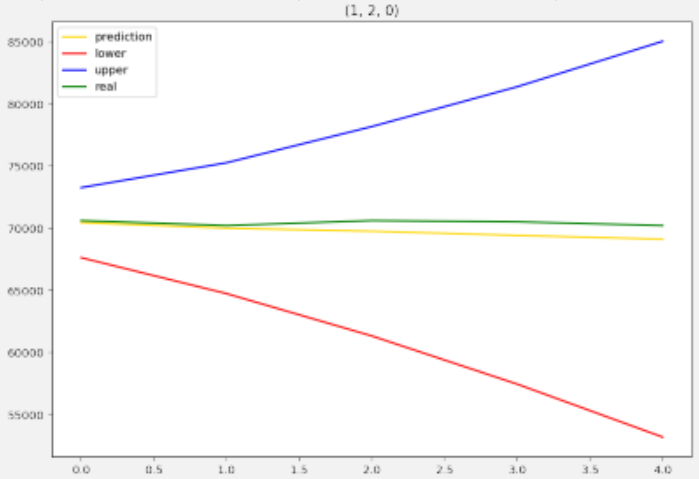
003 모델 테스트

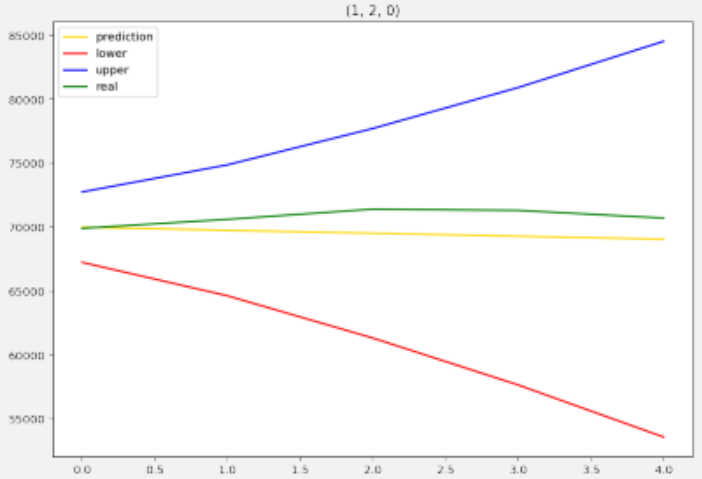
ARIMA모델

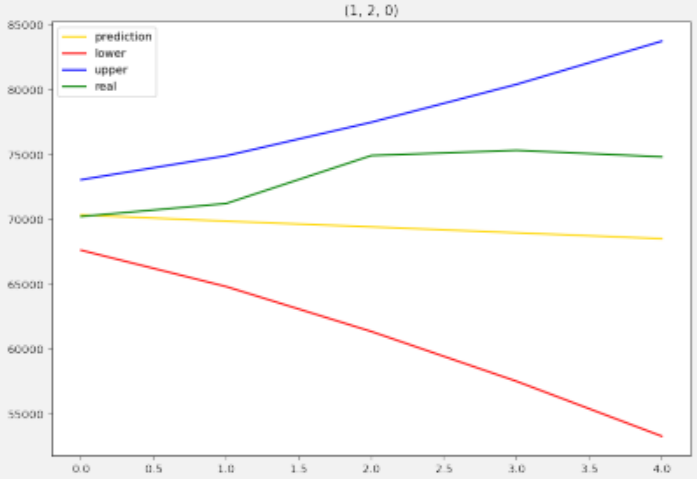
ARIMA 샘플 데이터 예측 결과 | (1, 2, 0) |

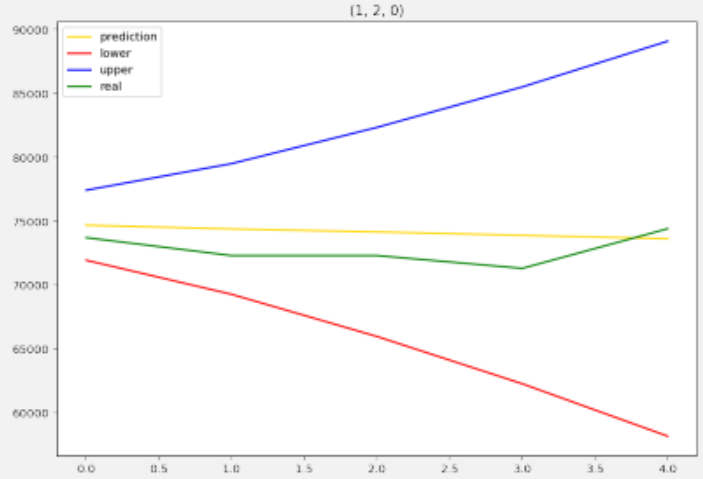
start = [’2020-11-02’,’2020-11-09’,’2020-11-16’,’2020-11-23’,’2020-11-30’,’2020-12-07’]
final = [’2021-11-03’,’2021-11-10’,’2021-11-17’,’2021-11-24’,’2021-12-01’,’2021-12-08’]

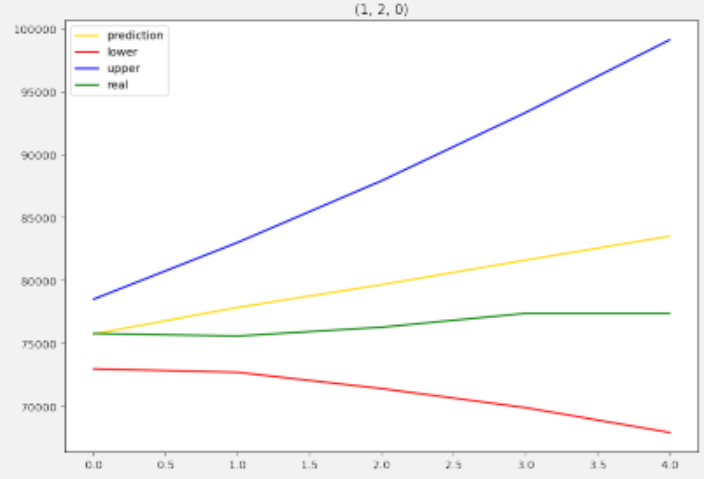










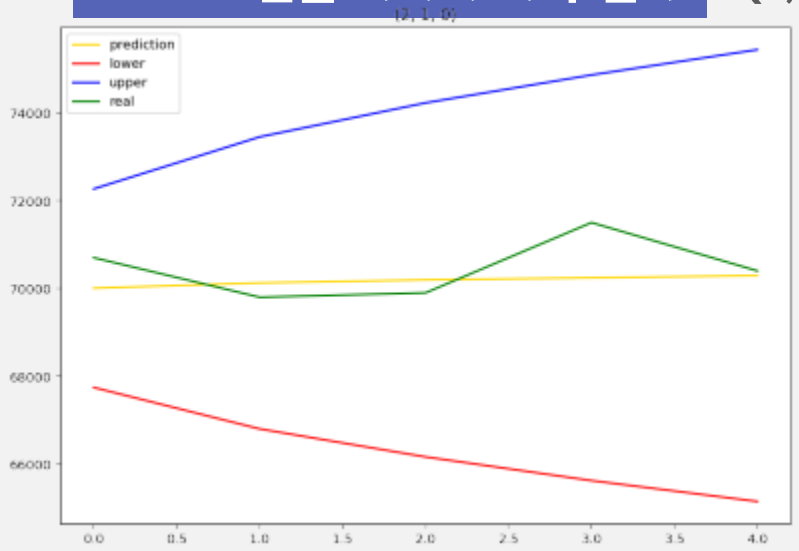


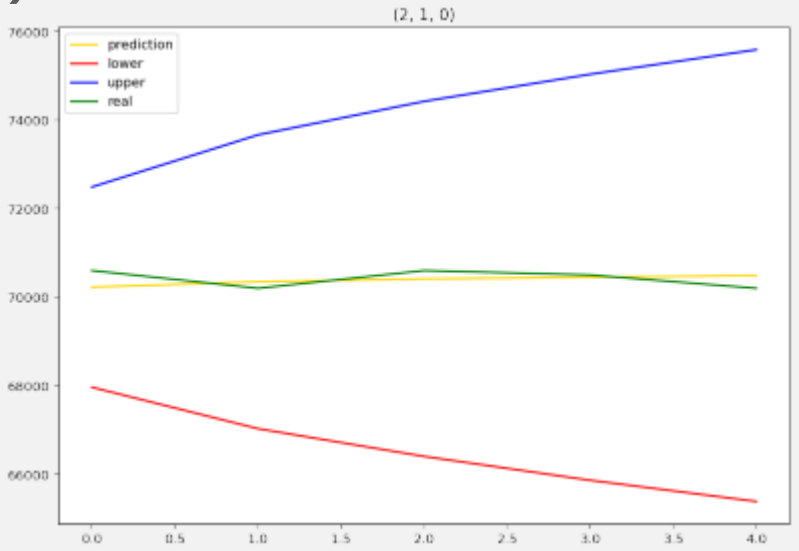
3

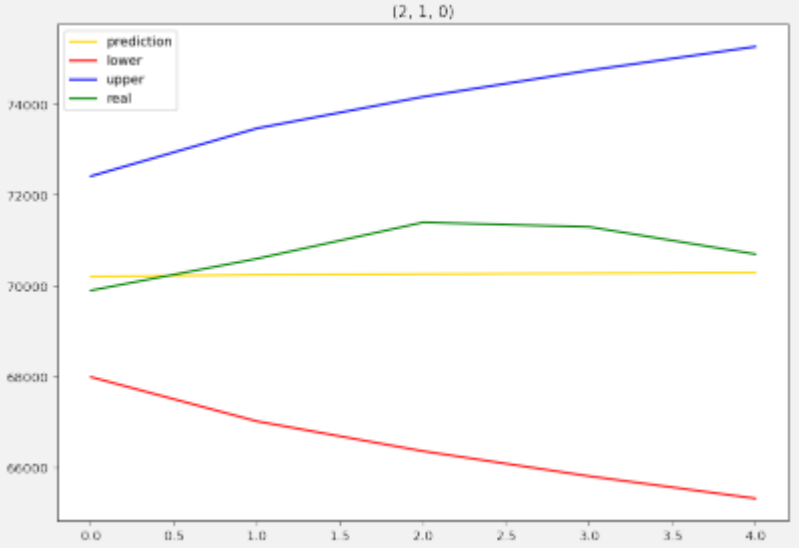
003 모델 테스트

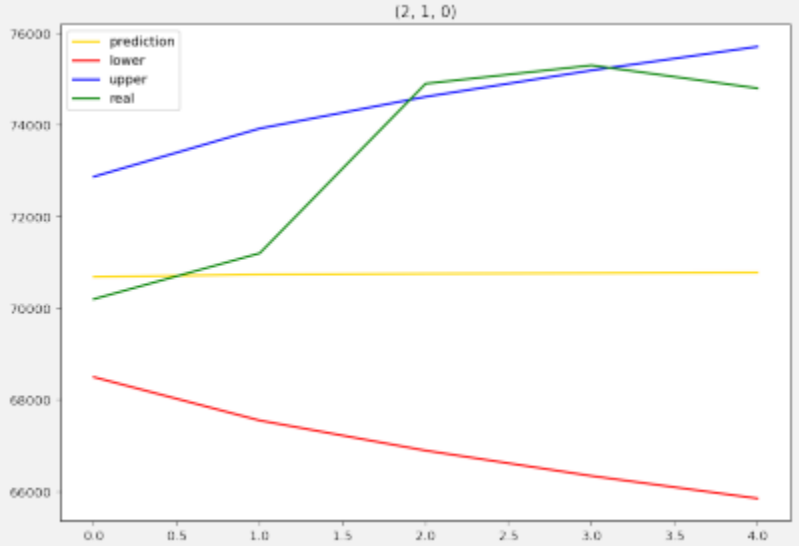
ARIMA모델

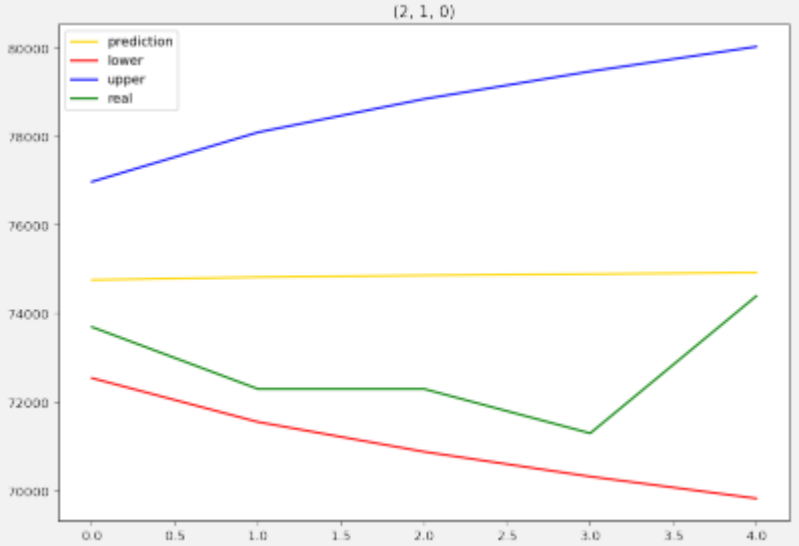
ARIMA 샘플 데이터 예측 결과 | (2, 1, 0) |

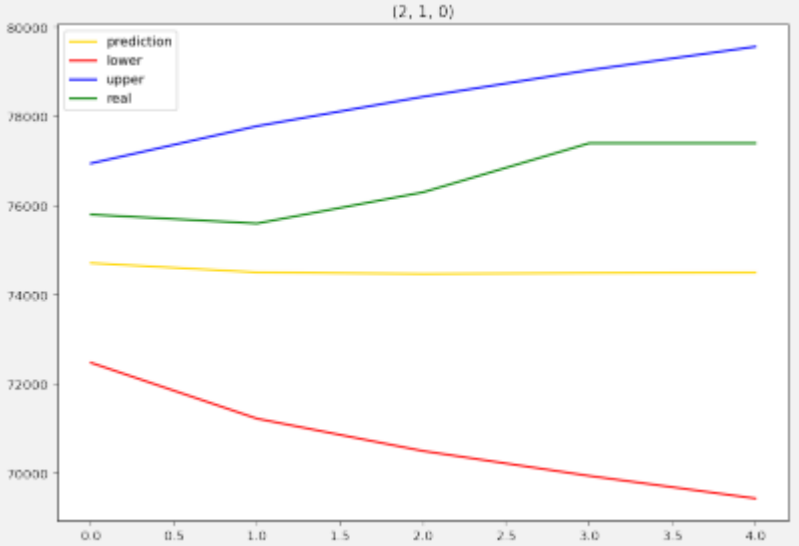










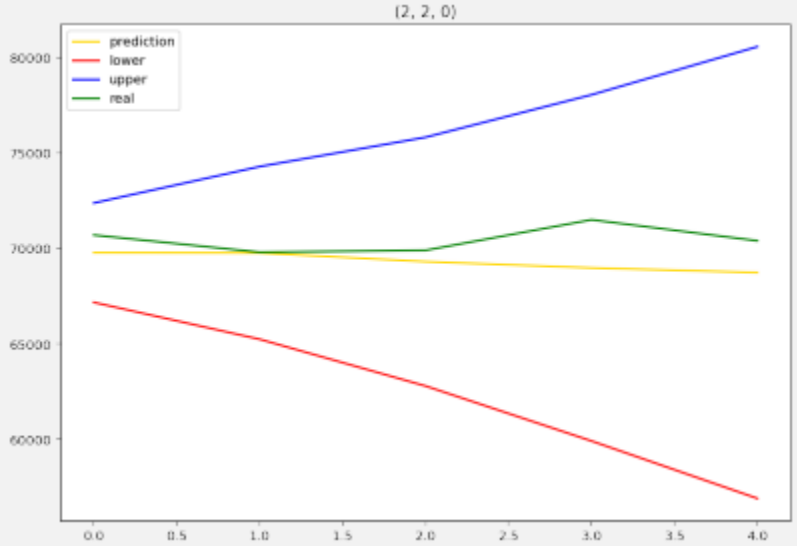


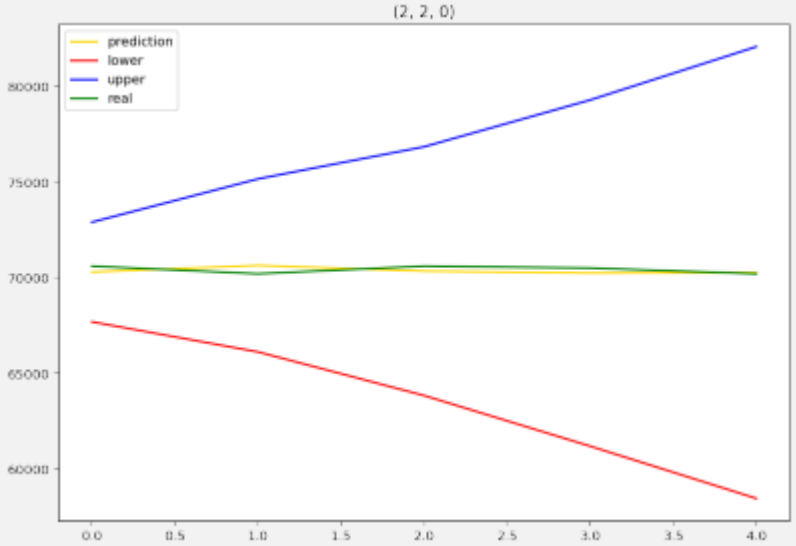
3

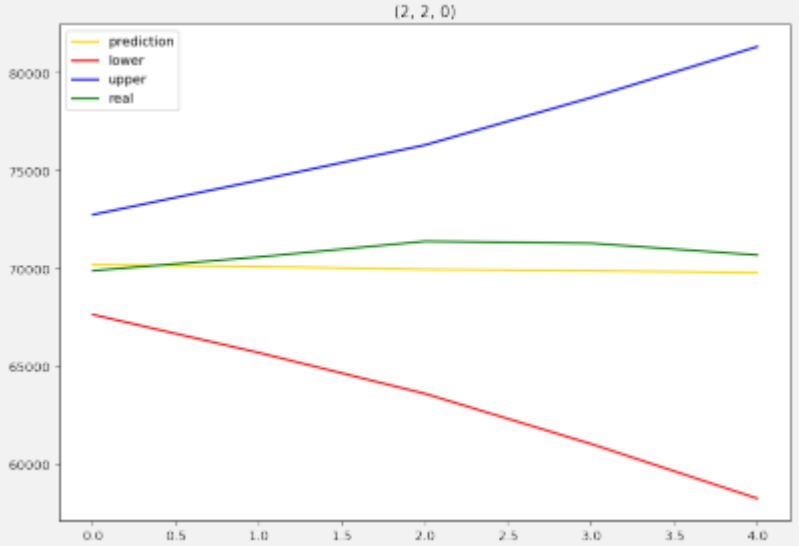
003 모델 테스트

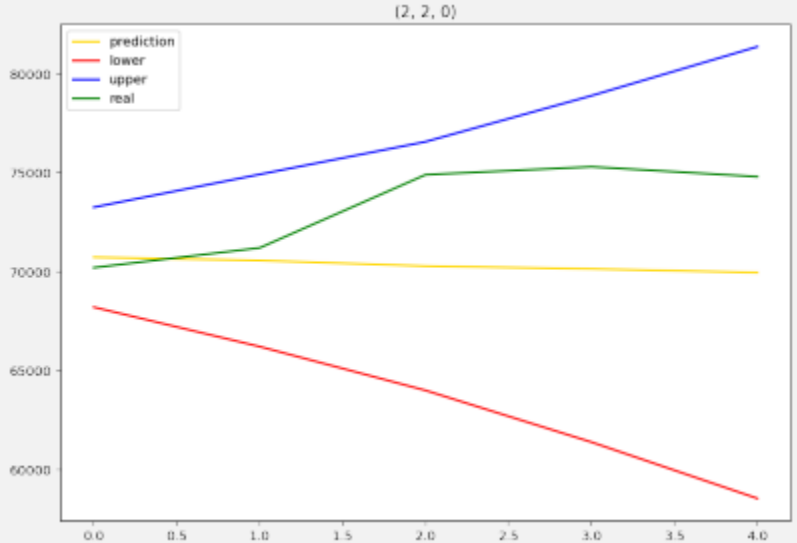
ARIMA모델

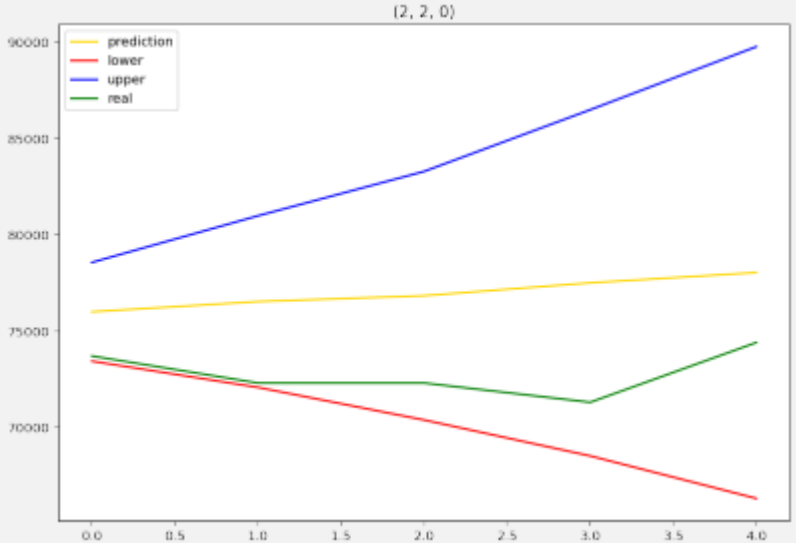
ARIMA 샘플 데이터 예측 결과 | (2, 2, 0) |

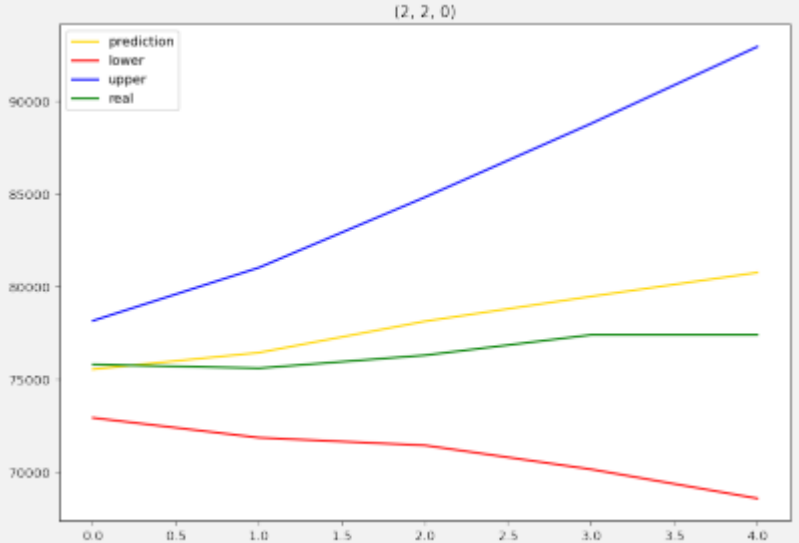










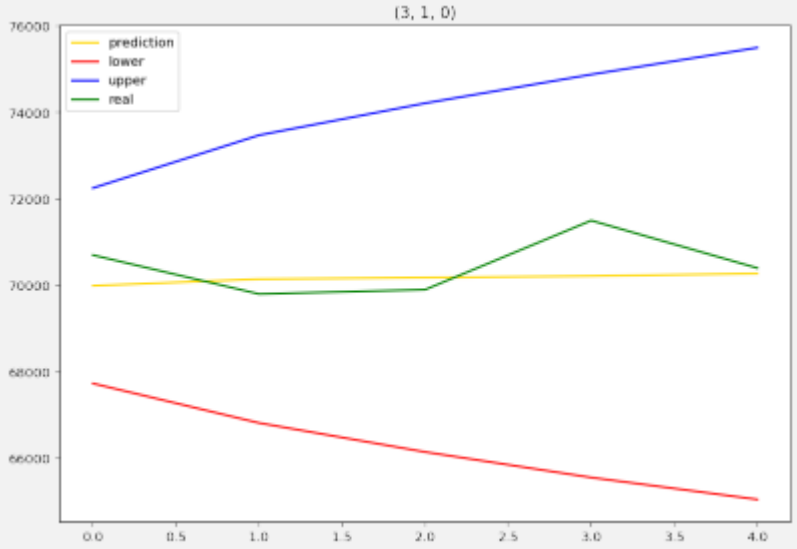


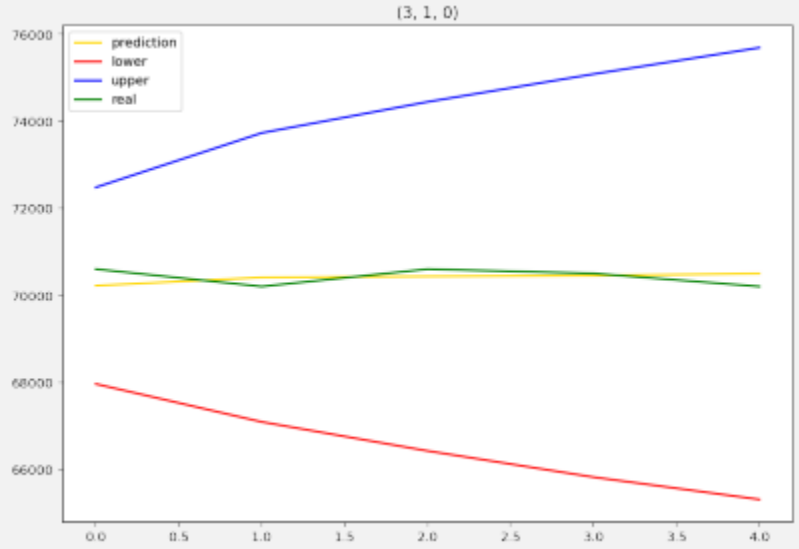
3

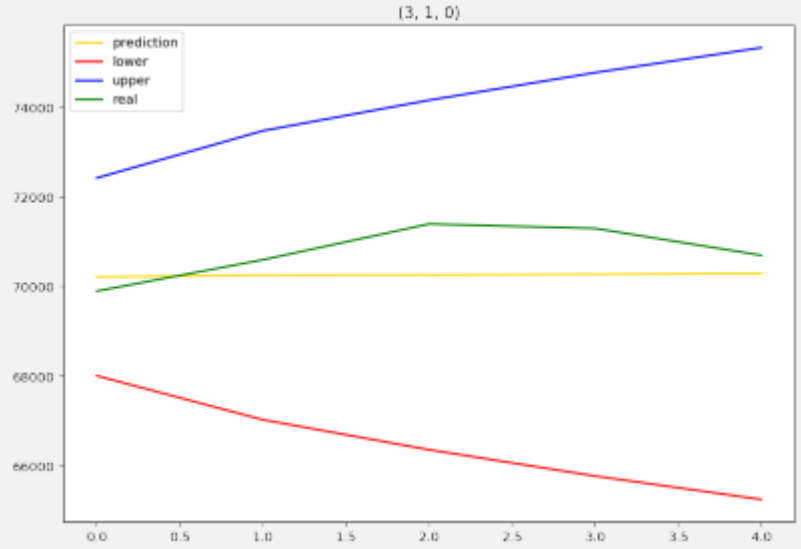
003 모델 테스트

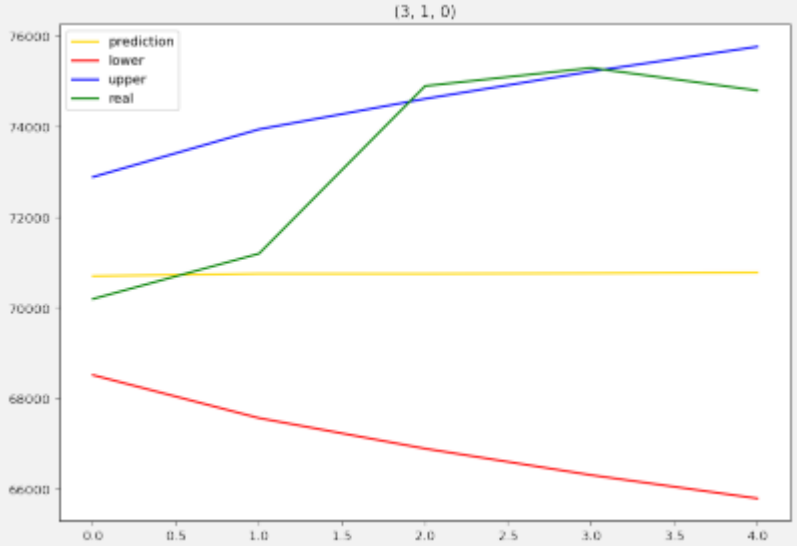
ARIMA모델

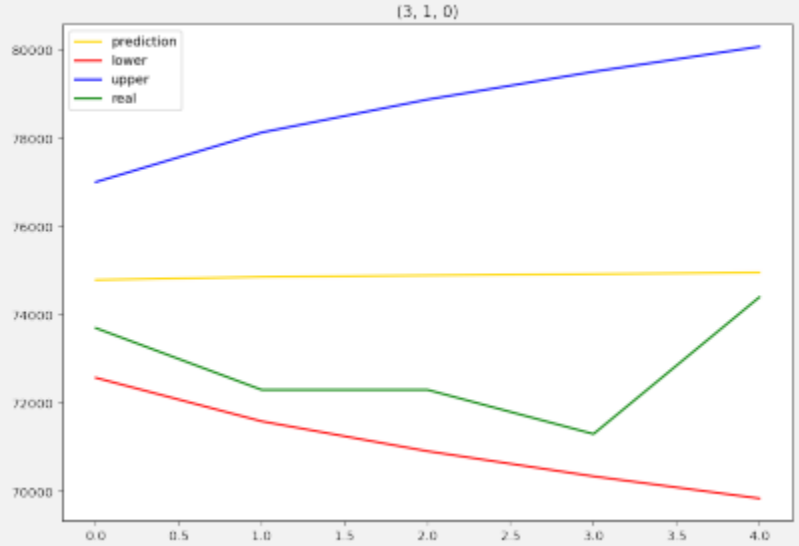
ARIMA 샘플 데이터 예측 결과 | (3, 1, 0) |

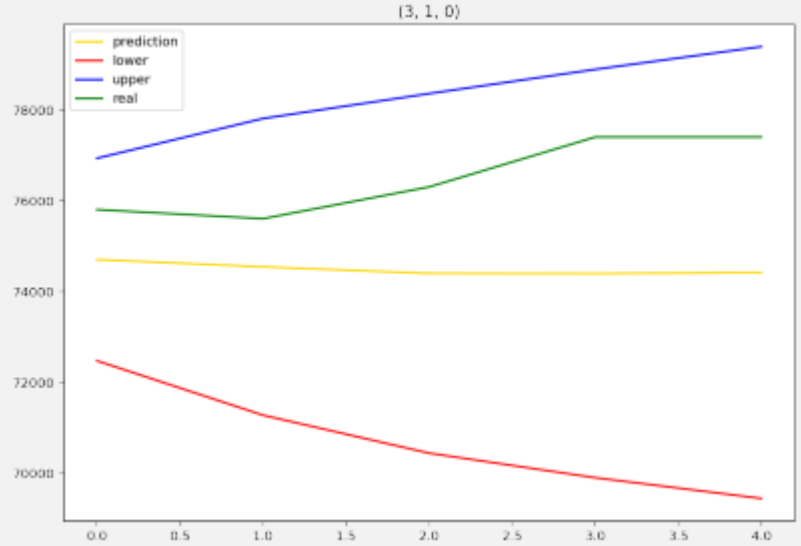

(3, 1, 0)


(3, 1, 0)


(3, 1, 0)


(3, 1, 0)


(3, 1, 0)

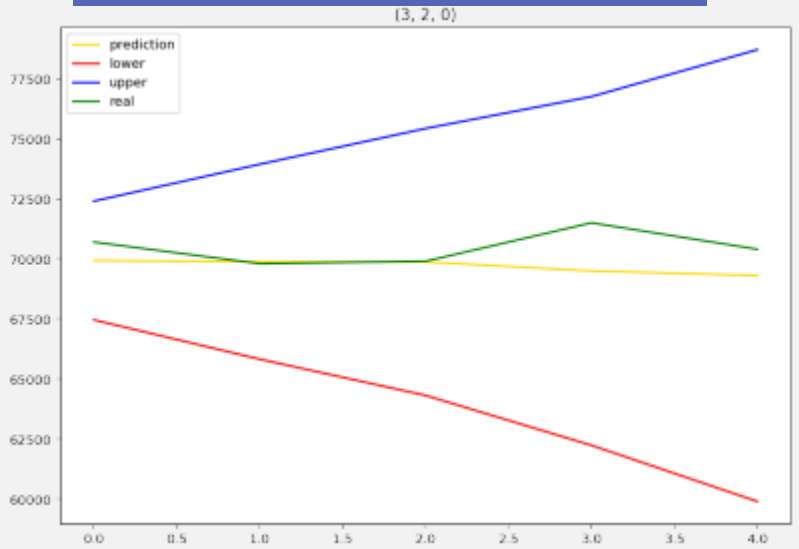

(3, 1, 0)

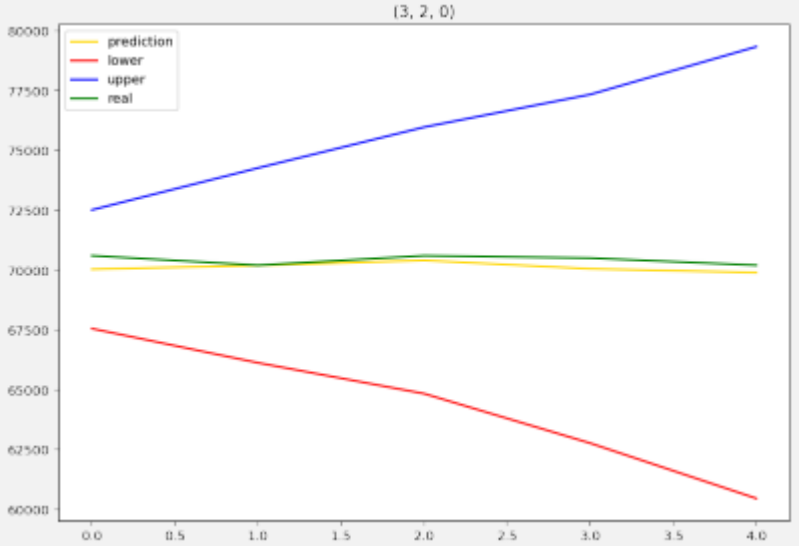
3

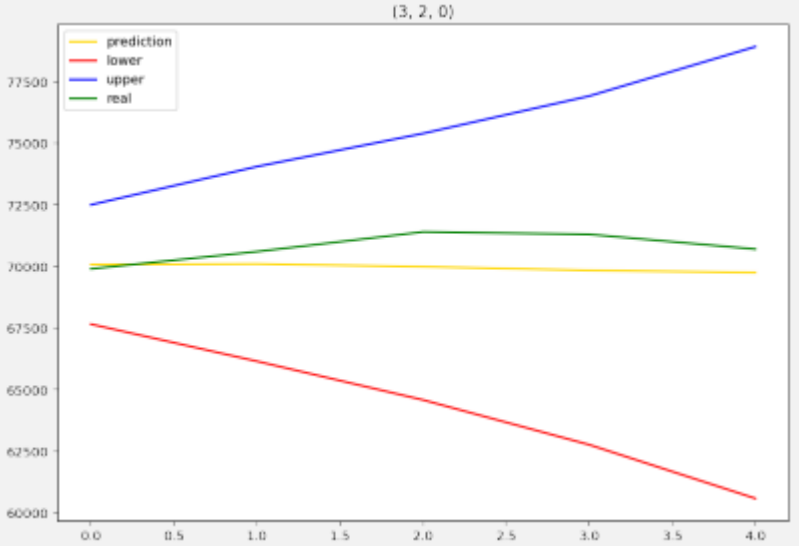
003 모델 테스트

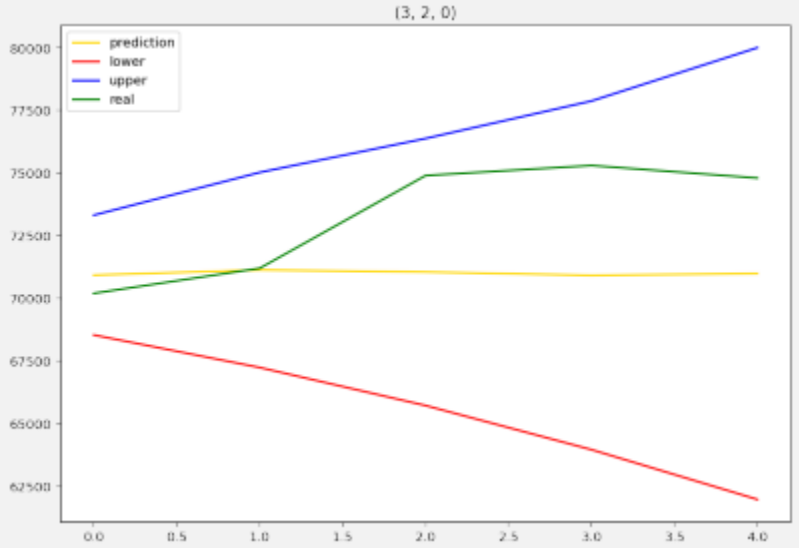
ARIMA모델

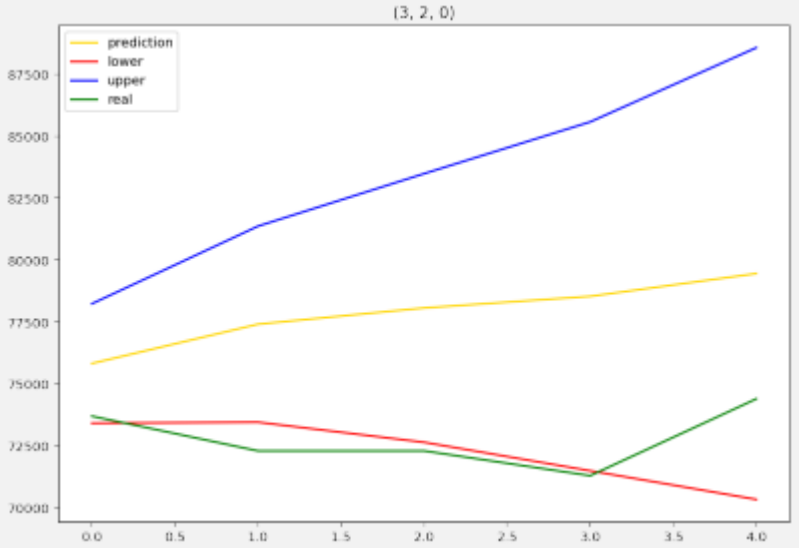
ARIMA 샘플 데이터 예측 결과 | (3, 2, 0) |

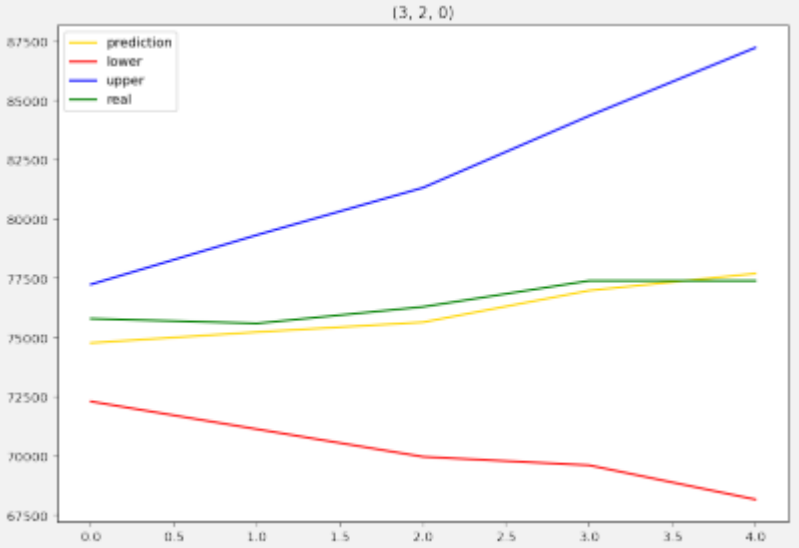










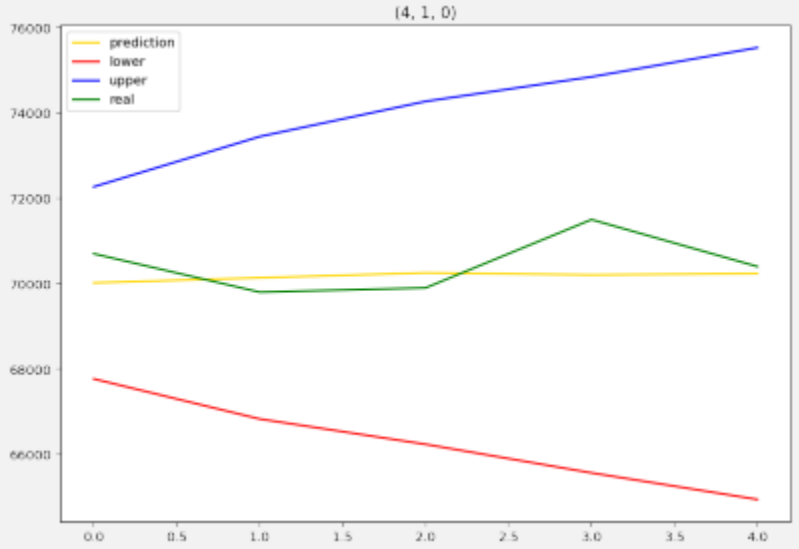


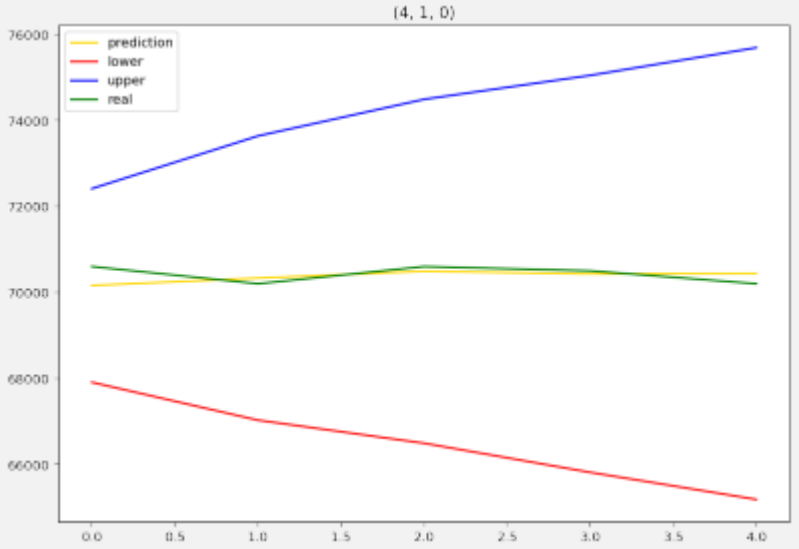
3

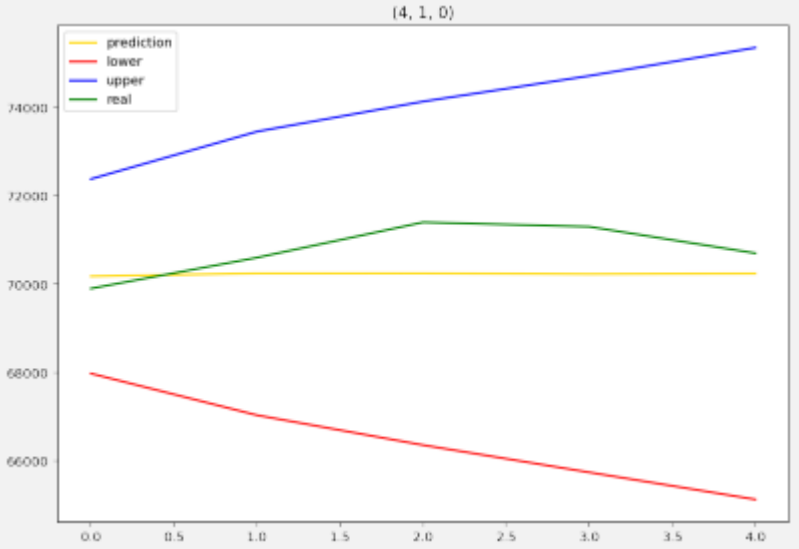
003 모델 테스트

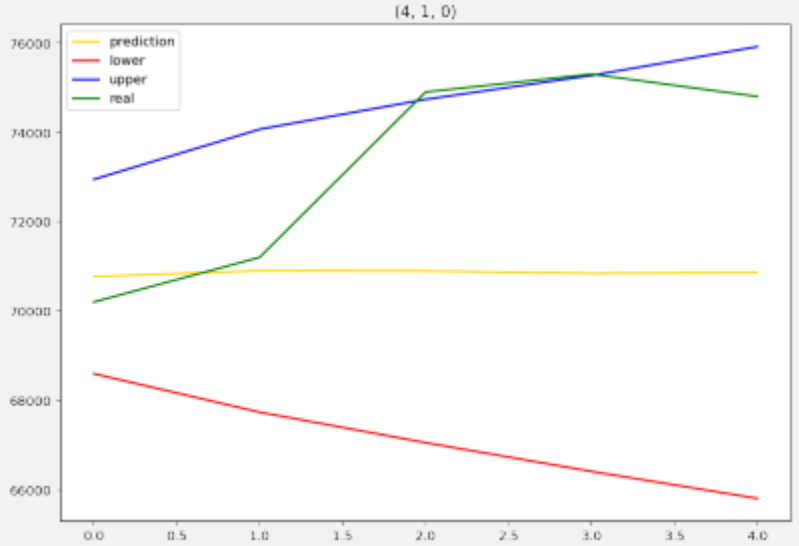
ARIMA모델

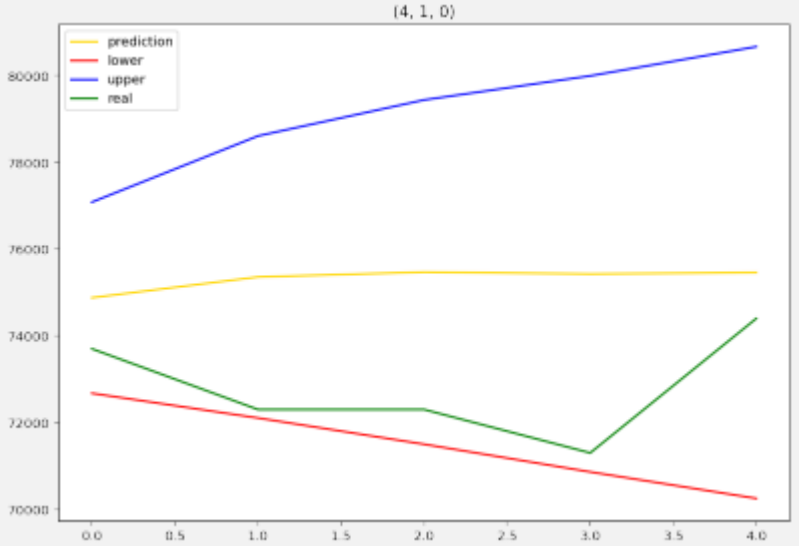
ARIMA 샘플 데이터 예측 결과 | (4, 1, 0) |

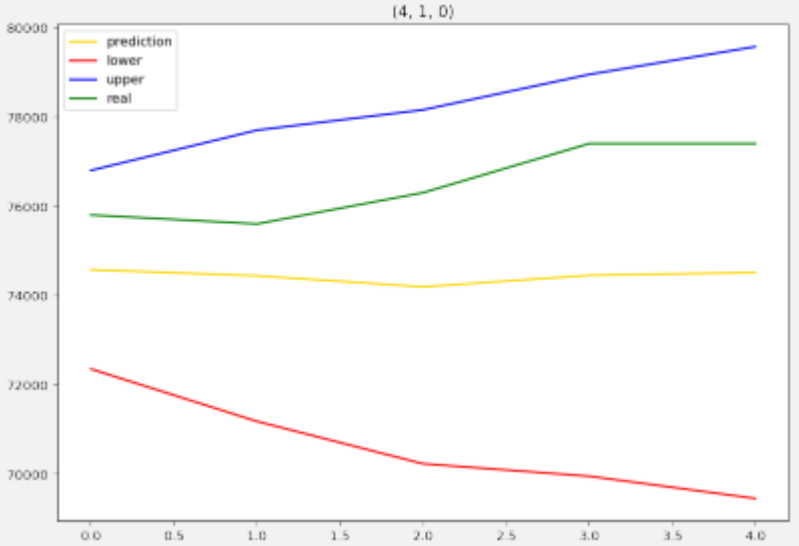










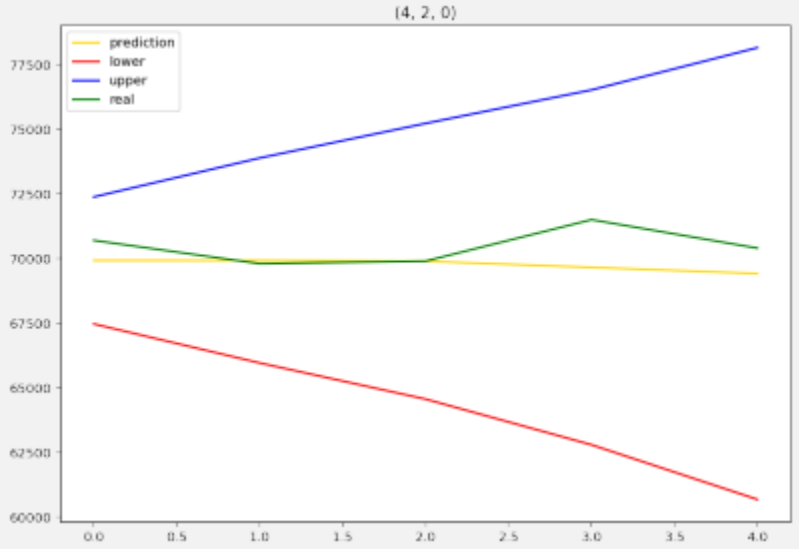


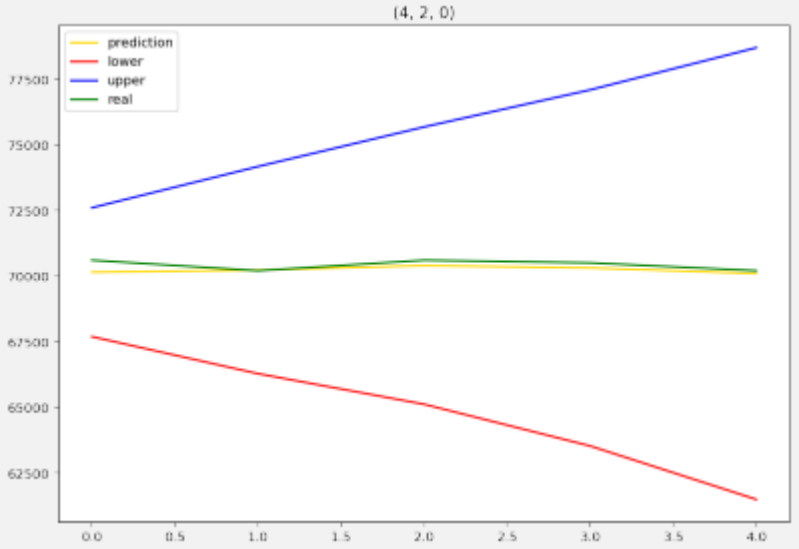
3

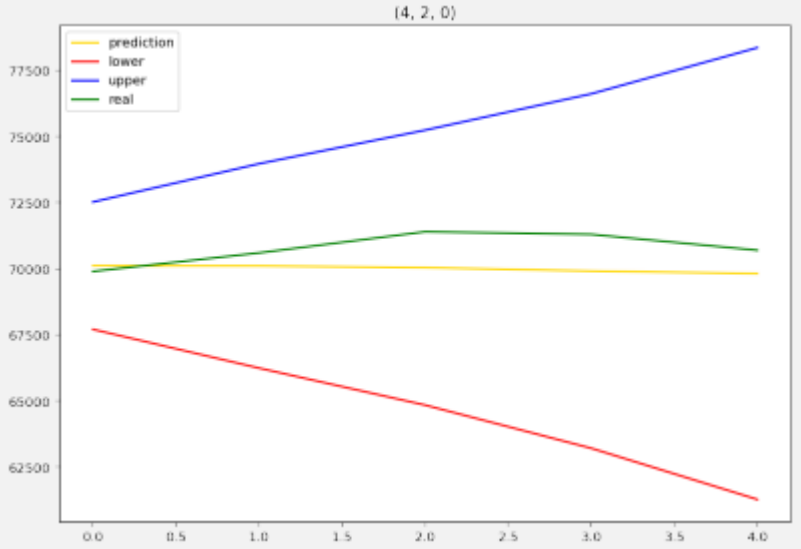
003 모델 테스트

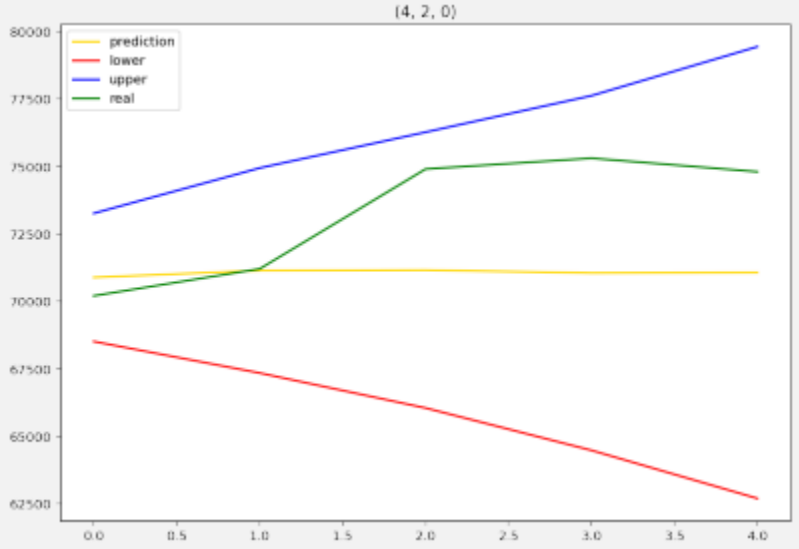
ARIMA모델

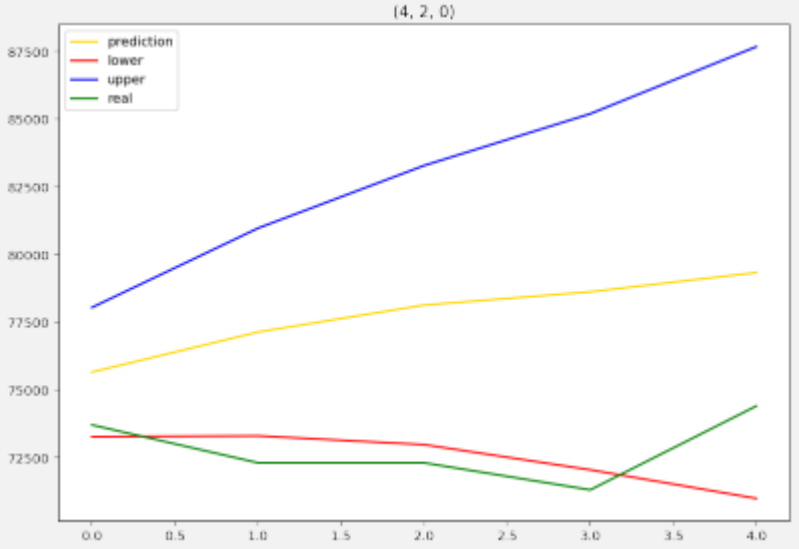
ARIMA 샘플 데이터 예측 결과 | (4, 2, 0) |

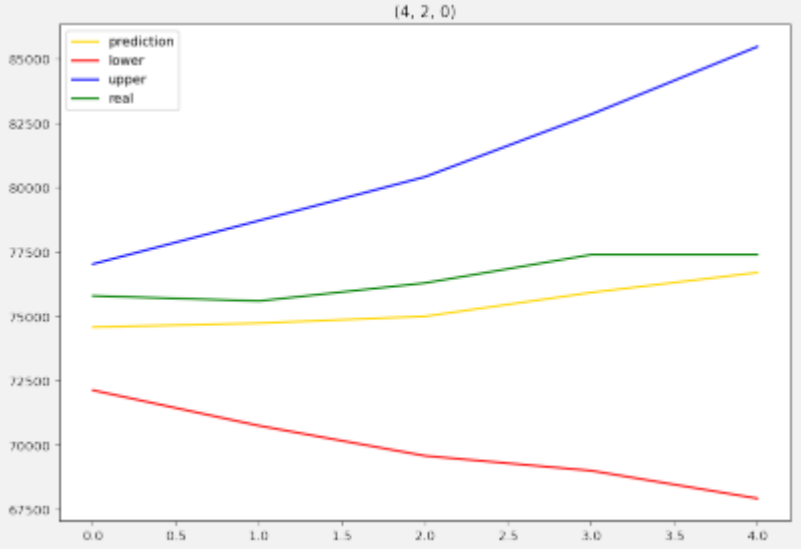












3

003 모델 테스트

ARIMA모델

ARIMA.fit parameter

ARIMA.fit(*start_params=None, transformed=True, includes_fixed=False, method=None, method_kwargs=None, gls=None, gls_kwargs=None, cov_type=None, cov_kwds=None, return_params=False, low_memory=False*)[source]

start_params : array_like, **optional**

Initial guess of the solution for the loglikelihood maximization. If None, the default is given by Model.start_params.

로그 우도 최적화에 대한 해의 초기 추측, 없는 경우 기본값은 Model.start_params에 의해 제공됩니다.

method : str, **optional**

The method used for estimating the parameters of the model. Valid options include 'statespace', 'innovations_mle', 'hannan_rissanen', 'burg', 'innovations', and 'yule_walker'. Not all options are available for every specification (for example 'yule_walker' can only be used with AR(p) models).

모형의 모수를 추정하는 데 사용되는 방법입니다. statespace, innovations_mle, hannan_rissanen, burg, innovations **및** yule_walker**가 있습니다. 모든 사양에 대해 모든 옵션을 사용할 수있는 것은 아닙니다**
(**예** : yule_walker**는** AR (p) **모델에서만 사용할 수 있습니다**).

method_kwargs : dict, **optional**

Arguments to pass to the fit function for the parameter estimator described by the *method* argument.

004

결론

- 모델링 결과 종합
- 향후 과제

5

005 결론

모델 별 금일 종가 예측 결과

SAMSUNG

12/16 종가: 77,800원
전일 대비 ▲200 (+0.26%)

분류

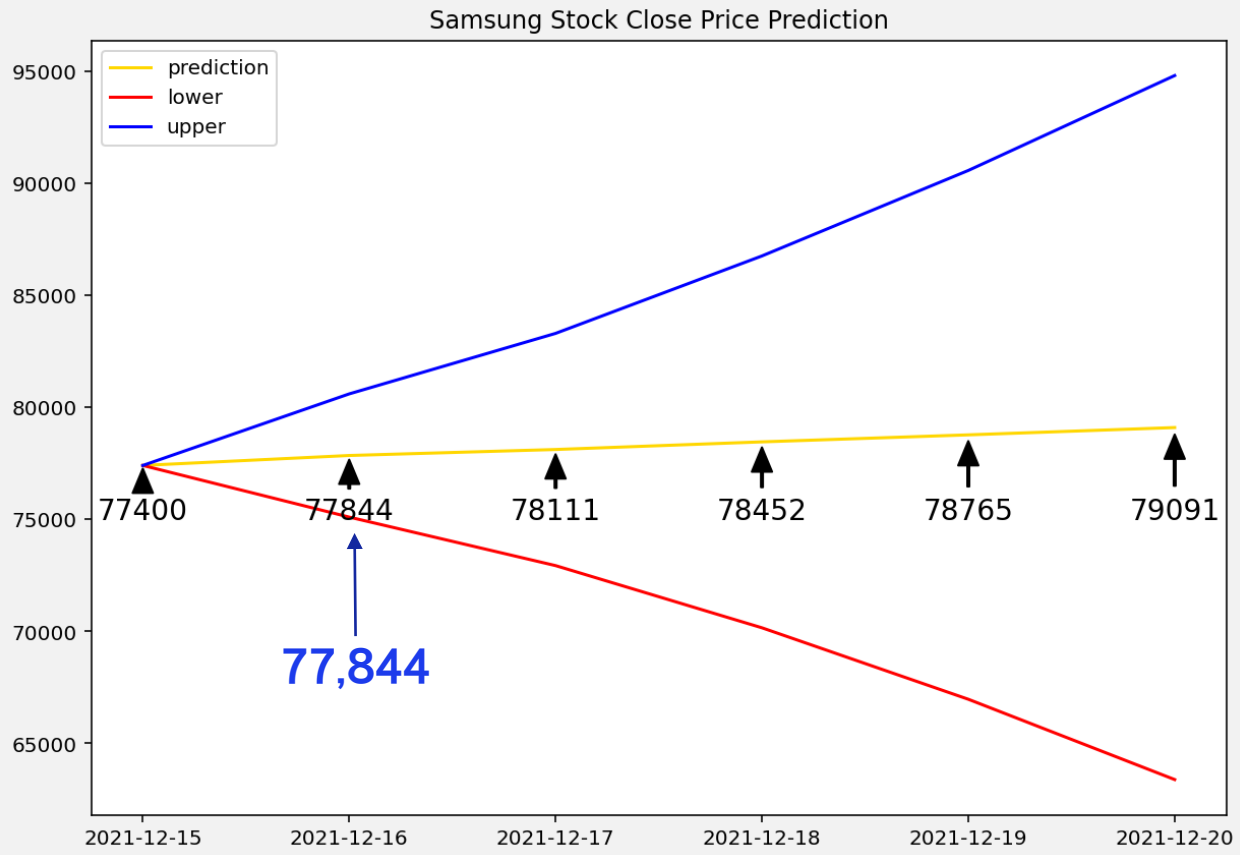
Decision Tree: 1 (상승)
Logistic Regression: 2 (하락)

회귀

Linear Regressor:
Degree=2 -> 99% -> 77894.364
degree=3 -> 95% -> 78057.308
Lasso
degree=3 -> 97% -> 77844.829
degree=4 -> 95% -> 77795.626

Ridge는 degree 3부터 마이너스 성능

ARIMA



5

005 결론

향후 과제

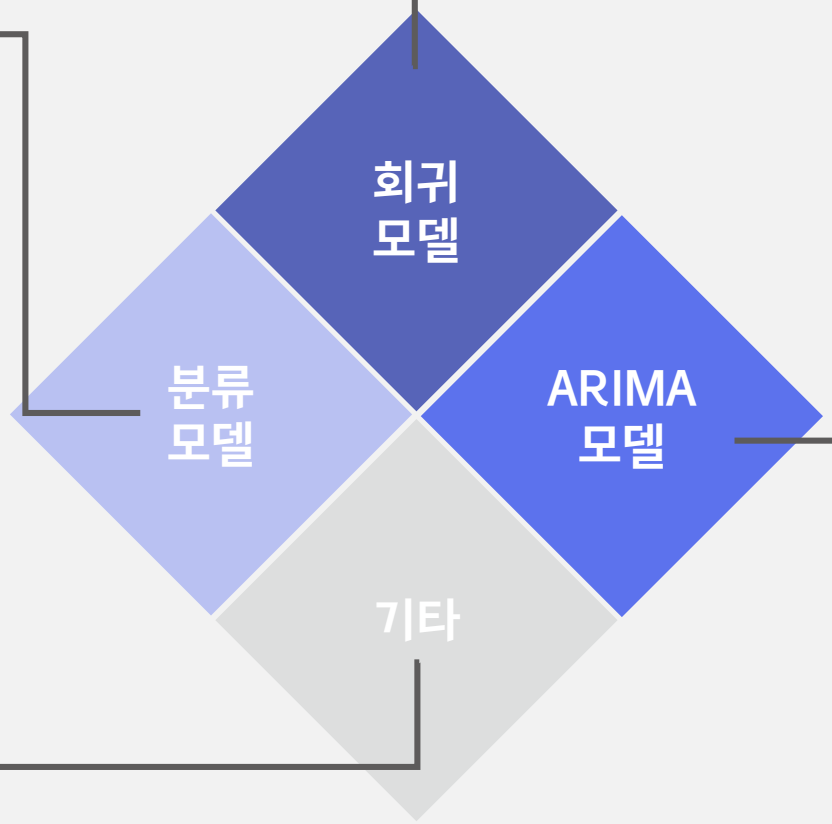
데이터

- 스케일링보다는 변화율(증감률)로 변환
- 화성은 美장과 함께

과적합 이슈 해소할 수 있는 다른 알고리즘/모델들 탐색

- 알고리즘 - C4.5 / CART / CHAID
- 모델 - SVM, KNN 등

여러 모델 앙상블 시도



딥러닝 모델 추가 학습

- RNN 기반 LSTM 모델
- Facebook 제공 prophet 라이브러리 등

ARIMA.fit() 파라미터 이해

다양한 데이터를 활용한 분석 다각화

- 재무데이터를 변수로 활용한 분석
- 뉴스 키워드를 활용한 텍스트 분석

Q&A

Reference

- “주식 투자자 43% "코로나 이후 시작"… 92% "계속할 것"”, 한국일보, https://m.hankookilbo.com/News/Read/A2021050316140000896
- “"나 떨고 있니"…동학개미 막 내린 초저금리 대응 어떻게”, 매일경제, https://www.mk.co.kr/news/special-edition/view/2021/09/888222
- “옥석만 가린다! 파이어족 이끈 年 34% 수익 ‘울트라 퀀트 투자’”, 주간동아, https://weekly.donga.com/BestClick/3/all/11/2926216/1

[1] C. H. Daube, The Corona Virus Stock Exchange Crash, Institut of Accounting, Controlling and Financial Management, 2020.
[2] R. Ortmann, M. Pelster, and S. T. Wengerek, COVID-19 and investor behavior, Available at SSRN 3589443, 2020
[3] https://www.kukinews.com/newsView/kuk202110120273