시계열 분석에 의한 삼성전자 주식 가격 예측

발제자 : 김한나, 박지수, 조민아

분석의 목적

2010년 ~ 2019년 삼성전자 주식의 일별 종가 자료를 박스-젠킨스 ARIMA 분석을 이용해 시계열 모형의 적용가능성을 보고자 함.

원래는 다수의 변수들을 가지고 다변수시계열 분석을 진행하려 하였으나 많은단변수시계열 분석으로 방향을 바꿈.

시계열 분석의 가장 기초가 되는 주제와 방법론을 사용하였으며, 이후 발전된 분석에 도움이 되었으면 함.

자료의 수집

네이버 금융을 기반으로 기업의 주식 가격 자료와 해외 주식시장 지표, 환율 등 각종 금융 자료를 크롤링과 엑셀 시트를 활용하여 수집하였다.

자료의 전처리

크롤링한 데이터의 변수명이 한글일 경우 영문으로 대체하였다.

코스피에 상장된 삼성전자 주식은 2018년 5월 액면분할을 실시하여 해당 시점 이전과 이후 1주당 가격이 1/50배가 되었기 때문에 액면분할 실시 이전 주식 가격은 전부 50으로 나누어서 사용하였다.

결측값이 존재하는 데이터의 해당 행을 삭제하는 식으로 결측값을 처리하였다

시계열 분석이므로 자료의 Index를 날짜 변수인 Date로 대체하였다.

시간의 순서가 2010년부터 이므로 인덱스를 거꾸로 뒤집어 재배열 하였다.

전처리 과정을 거쳐 2010년 5월 14일 부터 2019년 12월 30일 사이의 삼성전자 주식의 종가 데이터를 얻었다. 금융 데이터의 특성상 공휴일과 주말의 데이터는 존재하지 않기 때문에 관측값의 개수는 총 2057개 이다.

자료의 개요

결측값이 없는 범위 내에서의 2010년 5월 14일 ~ 2019년 12월 30일을 범위로 삼성전자 주식 가격의 종가를 이용하였다.

<표1, 삼성전자 기초 통계량 (2010~2019)>

	closing_sam
count	2057.000000
mean	31355.483714
std	11351.501652
min	13600.000000
25%	23940.000000
50%	27400.000000
75%	43150.000000
max	57220.000000

<그림1, 삼성전자 종가 그래프 (2010~2019)>

분석 방법

시계열분석

자기회귀모형(autoregressive model; AR) 과 이동평균모형(moving average model; MA)을 적용한 자기회귀 이동평균모형(ARMA)을 사용하였다.
그리고 적용공분산 정상성(covariance satationary)을 만족시키는 과정을 거쳐 분석을 진행하는데 이때 사용하는 모델이 ARIMA 모형이다.

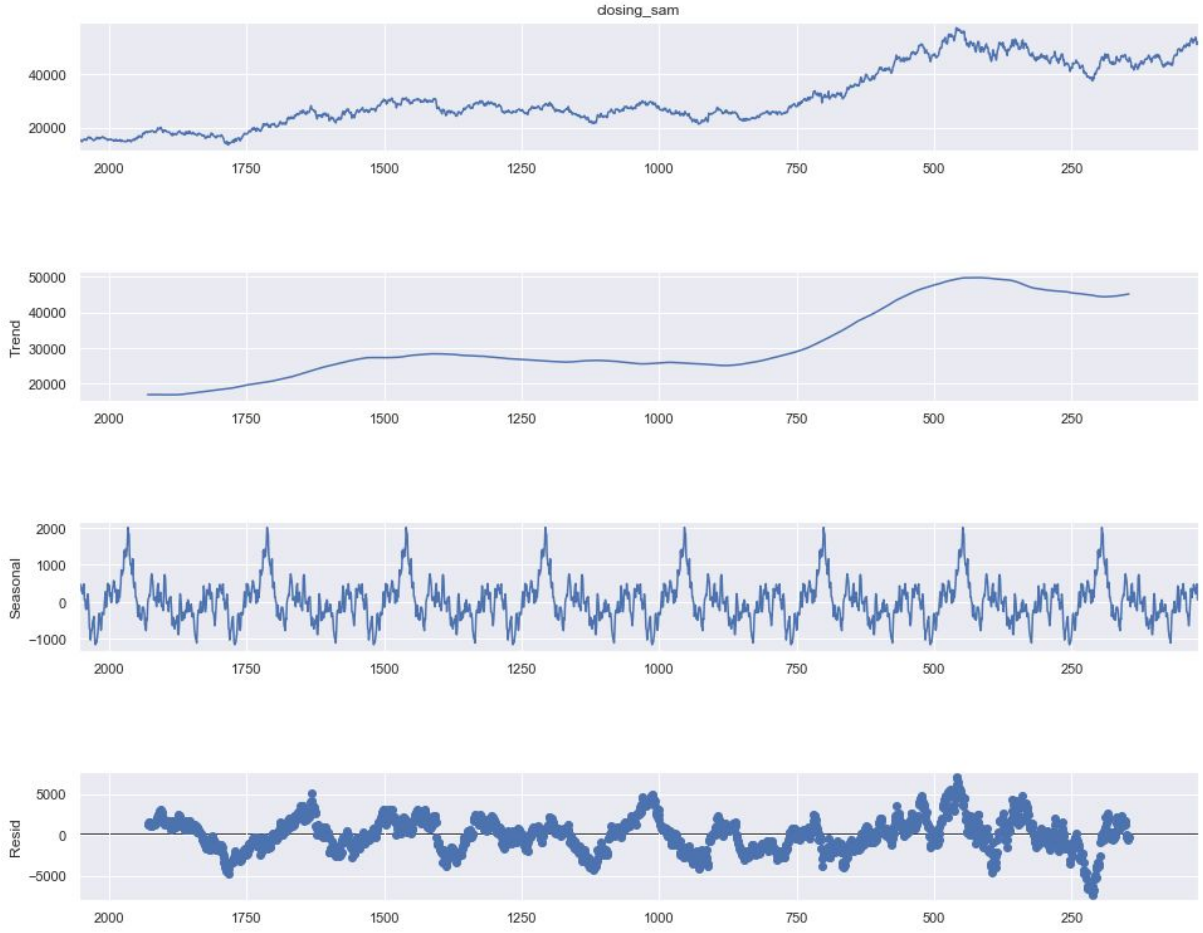
Data

closing_sam(삼성전자 주식 종가)

계절성, 주기성 확인

우선 시계열 분석의 첫 단계로 해당 자료의 계절성과 주기성을 보고자 했다.

<그림2, 삼성전자 종가 추세와 계절성 그래프>

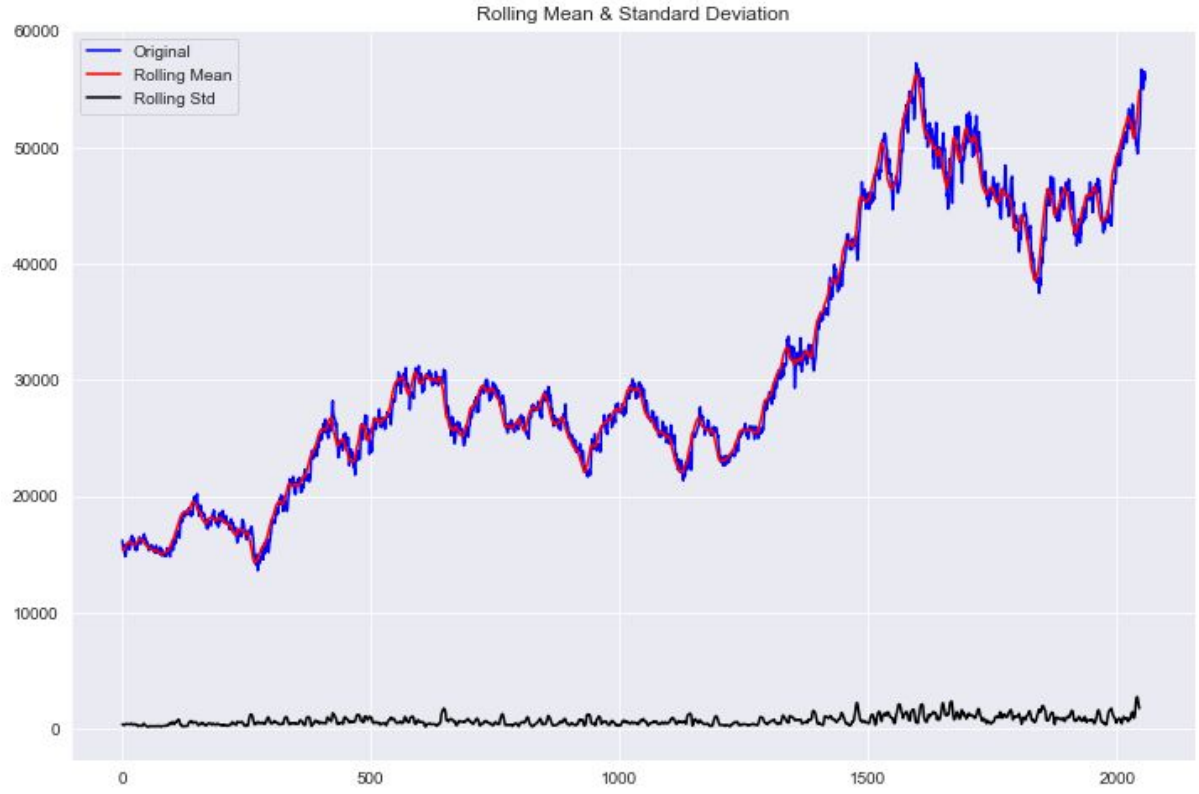


Seasonal 그래프는 전체 데이터를 253일로 frequency를 설정하여 잘라, 값의 등락을 계산하여 붙인 그래프이다. 해당 그래프를 보면 파동이 일정 주기로 반복되는 듯한 모습을 보여주는데 엄밀한 분석을 위해서는 자료를 주기를 기준으로 잘라 서로 분산분석(ANOVA)와 다중비교를 실시하여 해당 비교의 p-value를 확인해야 하나, 이번 분석에서는 하지 못하였다.

정상성

주식 데이터 같은 금융 데이터는 일반적으로 비정상적이다. 추세 성분을 가지며, 시간에 따라 확률구조가 변하는 시계열 데이터이기 때문이다. 정상성 시계열 자료는 시간이 변해도 확률구조가 일정한 시계열을 의미한다.
ARIMA모델은 정상성을 전제로 하는 모델이기 때문에 데이터의 정상성을 확인해야 한다.
추세가 존재한다는건 시간에 따라 평균이 변화한다는 뜻이므로 정상 시계열이 아니다.

<그림3, 삼성전자 종가 이동평균과 이동표준편차 그래프>



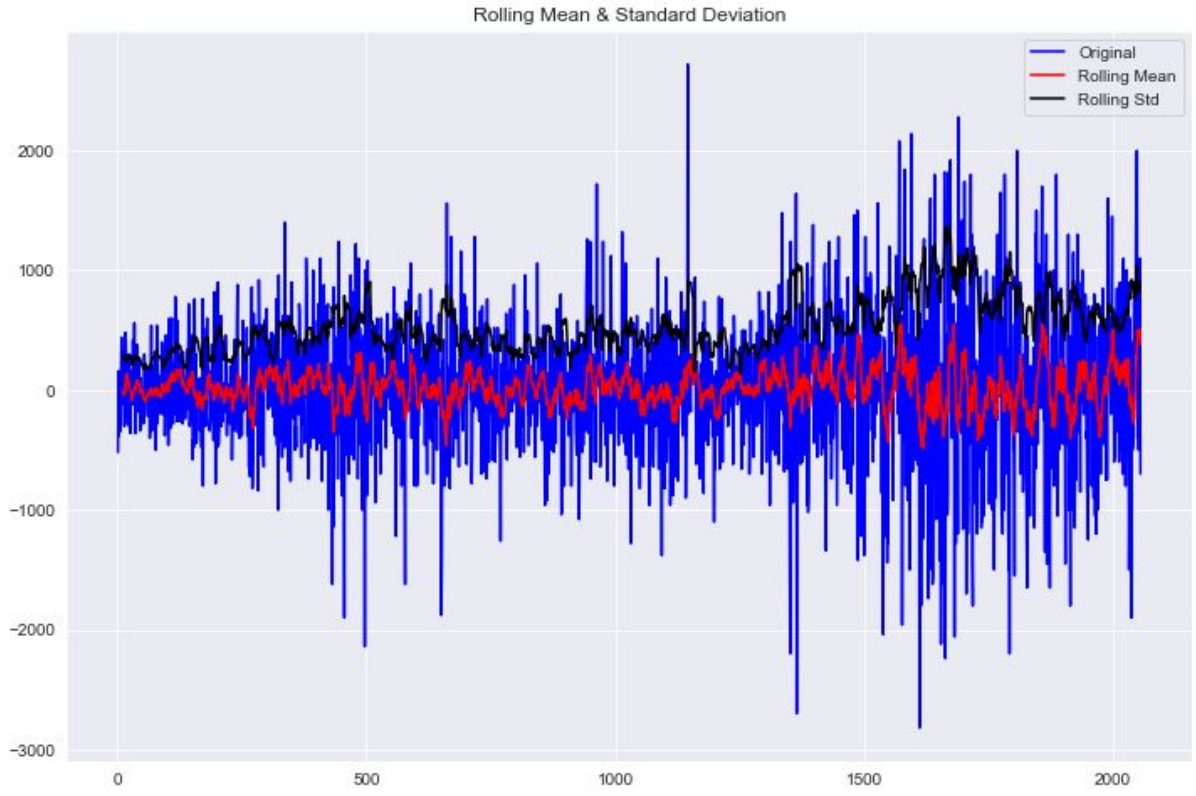
Results of Dickey-Fuller Test:
p-value = 0.5695. The series is likely non-stationary.

Test Statistic	-1.426597
p-value	0.569461
#Lags Used	10.000000

정상성 검정은 Augmented Dickey Fuller 테스트를 실시하였다. Akaike information criterion (AIC) 을 통해 가장 적합한 시차를 고른 후 원본 데이터를 집어 넣은 결과, 0.5695의 p값을 얻었다. p값이 유의 수준보다 크기 때문에 원본 데이터는 정상 시계열이라고 할 수 없다.

따라서 데이터를 정상 시계열화 하기 위해 1차 차분을 해보았다. 0을 중심으로 진동하고 있는 모습이지만 그 진폭은 일정하지 않아 보인다. 1차 차분한 데이터에 대해 다시 ADF검정을 실시하였다.

<그림4, 삼성전자 종가 1차 차분 이동평균, 이동표준편차 그래프>



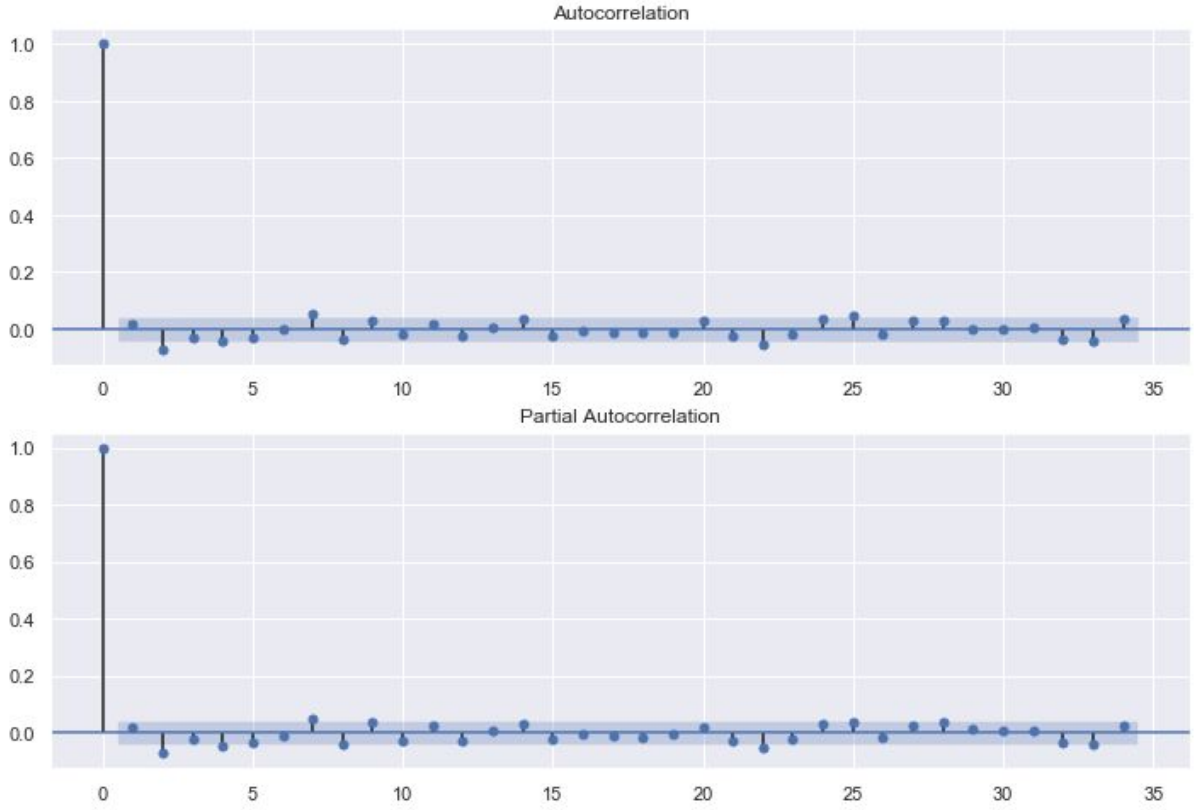
Results of Dickey-Fuller Test:
p-value = 0.0000. The series is likely stationary.
Test Statistic -1.708612e+01
p-value 7.647214e-30
#Lags Used 7.000000e+00

1차 차분의 데이터는 p값이 거의 0에 가까워 유의수준보다 작으므로 정상 시계열이라고 할 수 있다.

모형의 구축

데이터의 정상성을 확보했으므로, 이제 ARIMA모델의 ARIMA(p, d, q)를 정해야 한다. 일반적으로는 ACF 와 PACF 그래프를 이용해서 정한다. 막대가 꺾일때까지의 막대 수를 사용하면 되는데, 본 분석에서는 AIC를 기준으로 최적의 차수를 정하는 방법을 사용하였다.

<그림5, 삼성전자 종가 1차 차분의 ACF, PACF 그래프>



	order	AIC
24	p2 d2 q1	31960.343418

ARIMA(2,2,1)가 최적의 모델이라는 결과를 얻었다.

<표2, ARIMA(2,2,1) 모델링 결과>

ARIMA Model Results

Dep. Variable:	D2.closing_sam	No. Observations:	2055
Model:	ARIMA(2, 2, 1)	Log Likelihood	-15979.672
Method:	css-mle	S.D. of innovations	575.481
Date:	Wed, 29 Jul 2020	AIC	31969.343
Time:	17:10:22	BIC	31997.484
Sample:	2	HQIC	31979.662

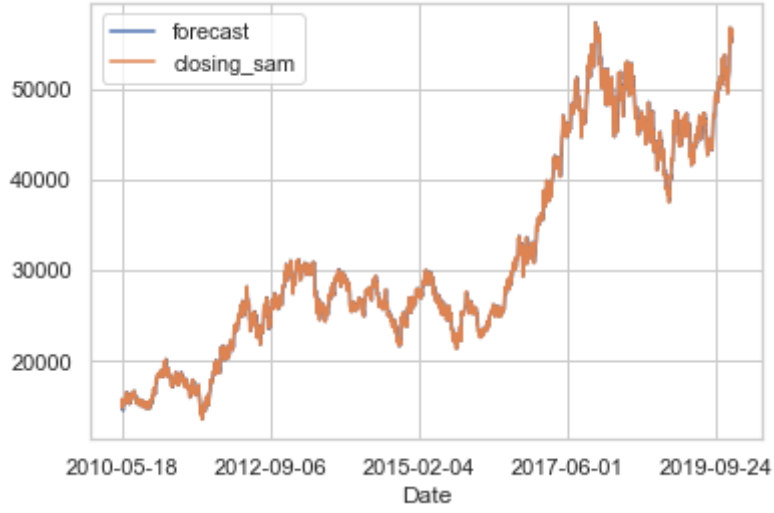
	coef	std err	z	P>|z|	[0.025	0.975]
const	0.0131	0.021	0.638	0.524	-0.027	0.053
ar.L1.D2.closing_sam	0.0272	0.022	1.236	0.216	-0.016	0.070
ar.L2.D2.closing_sam	-0.0671	0.022	-3.047	0.002	-0.110	-0.024
ma.L1.D2.closing_sam	-0.9999	0.001	-717.831	0.000	-1.003	-0.997

Roots

	Real	Imaginary	Modulus	Frequency
AR.1	0.2028	-3.8541j	3.8595	-0.2416
AR.2	0.2028	+3.8541j	3.8595	0.2416
MA.1	1.0001	+0.0000j	1.0001	0.0000

예측 적합

<그림6, 구간 내 삼성전자 종가 예측 그래프>



ARIMA(2,2,1) 모델로 주어진 자료의 예측을 실시하였다. 해당 구간은 데이터가 존재하는 구간이므로 엄밀히 말하자면 미래 예측은 아니다.
얼핏 보면 데이터가 존재하는 구간은 예측을 매우 잘 하고 있는것으로 보인다. 하지만 이는 단지 실제 데이터를 옆으로 옮긴 수준의 예측이다. 모델이 며칠 이전의 값을 그대로 옮겨서 예측하는 것이 나은 방법이라고 판단한 결과이다.

해당 자료는 2019년 12월 30일에서 끝난다. 따라서 그 이후 10일간의 주식 가격을 예측해 보았다.

55741.46385083,
55821.24987014,
55861.75164704,
55891.91121332,
55924.44014078,
55957.74154969,
55990.91857431,
56024.05399925,
56057.21028922,
56090.38358739

다음은 실제 2020년 1월 2일부터 10일치 주식 가격이다.

55200
55500
55500
55800
56800
58600
59500
60000
60000
59000

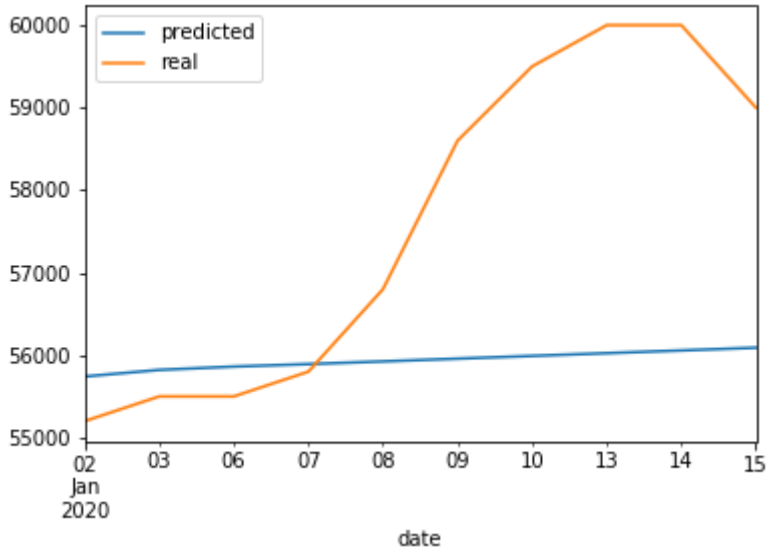
같은 방식으로 SARIMA (Seasonal Autoregressive integrated moving average) 모델에도 (2,2,1) 모델을 적용하고 MAPE와 SMAPE 로 성능을 비교하였다. 다음은 SARIMA 모델로 예측한 값이다.

0 55785.973930
1 55815.531830
2 55834.031418
3 55854.479576
4 55875.742566
5 55896.907620
6 55918.014433
7 55939.125487
8 55960.240595
9 55981.355579

$$\mathrm{M} = \frac{1}{n}\sum_{t=1}^{n}\left|\frac{A_t - F_t}{A_t}\right|, \mathrm{SMAPE} = \frac{100\%}{n}\sum_{t=1}^{n}\frac{|F_t - A_t|}{(|A_t| + |F_t|)/2}$$

결과를 평가하기 위해서는 MAPE(Mean Absolute Percentage Error) 과 SMAPE (Symmetric Mean Absolute Percentage) 를 이용하였다. 10% 이내의 값이면 좋은 예측력이라고 할 수 있다. 분석 결과 MAPE 는 3.54%, SMAPE 는 3.63 %의 결과로 우수한 예측 능력을 보여주었다.

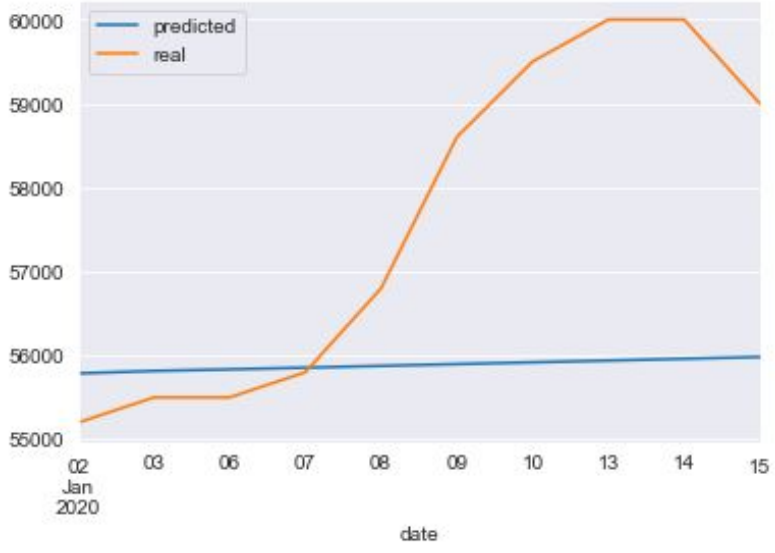
<그림7, ARIMA 예측 그래프>



예측한 결과와 실제 값을 그래프로 그려보자면 다음과 같다. 그래프를 보면, 앞부분의 데이터는 잘 예측 할 수 있는 반면, 시간이 지날 수록, 그리고 날짜 간 종가 변동 폭이 클 수록 떨어지는 예측력을 보인 다는 것을 알 수 있다.

ARIMA 와 마찬가지로 시간이 지날 수록 예측 능력이 떨어졌고, 날짜별로 달라지는 변동 폭을 예측하지 못하였으나, MAPE 는 3.32%, SMAPE 는 3.41%로 살짝 더 나아진 예측력을 보였다.

<그림8, SARIMA 예측 그래프 >



결론

삼성전자 주식 데이터만을 이용하여 시계열 분석의 가장 기본이 되는 연습을 해 보았다. 실제로는 많은 기업 데이터와 변수들을 수집 했지만 실제 분석에는 사용하지 않아 많은 아쉬움이 남아있다. ARIMA와 SARIMA 모델로 삼성전자 주식 종가 예측값과 실제 종가값과의 좀 더 자세한 분석을 하지 못한 것이 한계점으로 남는다. MAPE와 SMAPE의 예측률이 높게 나온것은 아마 삼성 주식 가격이 55000원 전후로 전체 값에서 변화율만 따지자면 급격한 변동이 없기 때문일 것이다. 차분에 대한 예측을 수행하거나, 좀 더 보수적인 평가 기준을 사용할 필요성이 있다. 현실적으로 어느 기업 주식의 종가만을 가지고 해당 기업의 주식 가격을 예측한다는 것은 거의 불가능에 가깝다. 이것이 가능했으면 주식시장은 지금과 같은 모습이 아니었을 것이다. 완벽히 예측하는 것은 불가능하므로 얼마나 더 실제와 가깝게 예측하는지 인 예측의 정확도가 주 관심사가 되는데 이는 많은 외부 데이터와 변수들을 추가하고, 더 성능이 좋은 예측 모형을 사용하는 것으로 보완할 수 있을 것으로 예상한다.

References

- 시계열 분석에 의한 국제유가 예측; Nymex-WTI 선물가격을 중심으로 송경재*. 양회민**,통계청「통계연구」제10권 제1호, 2005, pp. 62-81
- https://predictor-ver1.tistory.com/3, 머신 러닝으로 금 시세를 예측 해보자 feat. ARIMA
- 시계열 모형을 이용한 주가지수 방향성 예측, 박인찬1 · 권오진2 · 김태윤3, 한국데이터정보과학회지, 2009, 20(6), 991–998
- 국내주식시장의 거래량 변수의 특성에 관한 연구 , 길재욱(한양대학교), 김나영(한양대학교), 이은정(한양대학교)
- https://byeongkijeong.github.io/ARIMA-with-Python/, ARIMA, Python으로 하는 시계열분석 (feat. 비트코인 가격예측)
- https://www.kaggle.com/sumi25/understand-arima-and-tune-p-d-q, Understand ARIMA and tune P, D, Q
- (파이썬을 활용한) 금융공학 레시피 :문과생의 코딩 울렁증과 이과생의 금융 울렁증을 한 방에 씻어줄 금융공학, 김용환, 한빛미디어, 2018
- 회귀분석과 아리마시계열분석(Regression vs. time - series analysis), 송근원, 한국 학술정보, 2013