주가 데이터 분석

# 1. 시계열 분석을 진행해 주가 예측이 가능한가??

주식은 예측 불가능한 데이터라는 인식이 기본적으로 잡혀있다. 주식 데이터를 활용해 주식의 흐름을 정확히 예측 가능하다면 주식 투자에서 매번 성공할 것이다. 그렇다면 왜 주가 예측이 불가능한지 직접 시계열 분석을 진행해보고 예측이 불가능한 이유를 파악해보고자 한다.

## Data

테이블이(가) 표시된 사진

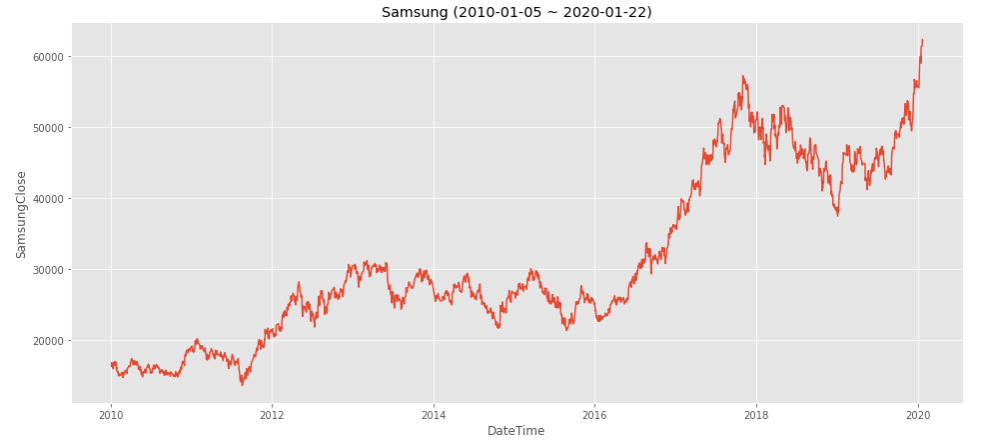
자동 생성된 설명

시계열 분석에 활용한 데이터는 삼성전자 종가로 기간은 2010년 1월 5일부터 2020년 1월 22일까지 총 2141일의 데이터로 분석을 진행하고 예측하여 예측한 결과를 2020년 1월 23일부터 2021년 7월 2일까지 데이터와 비교해보고자 한다.

## EDA

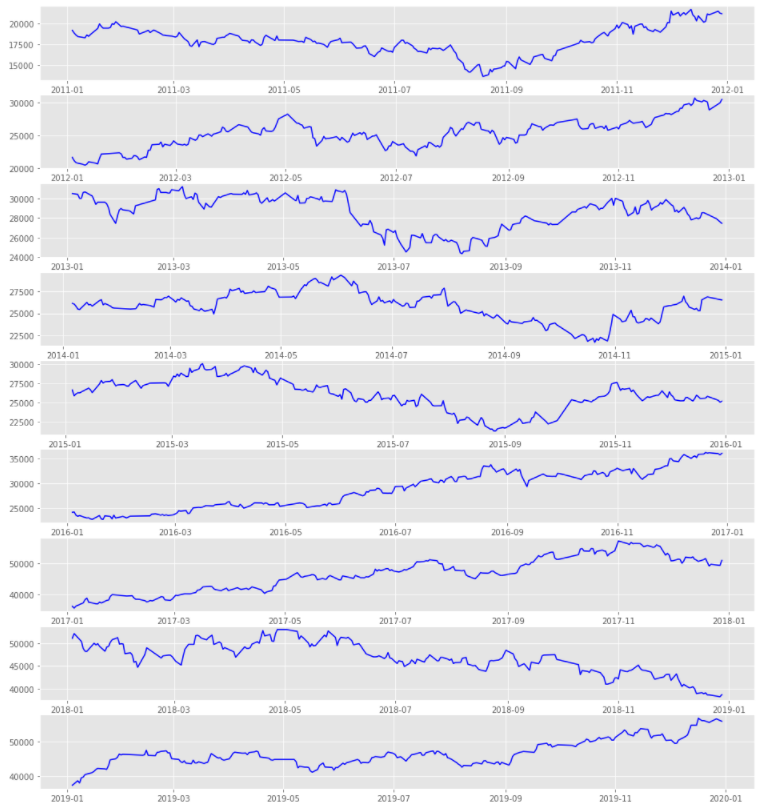
우선, 주가 데이터의 추세와 흐름을 보기위해 주가를 총 기간별, 연도별, 요일별로 확인해보자.

### 총 기간별



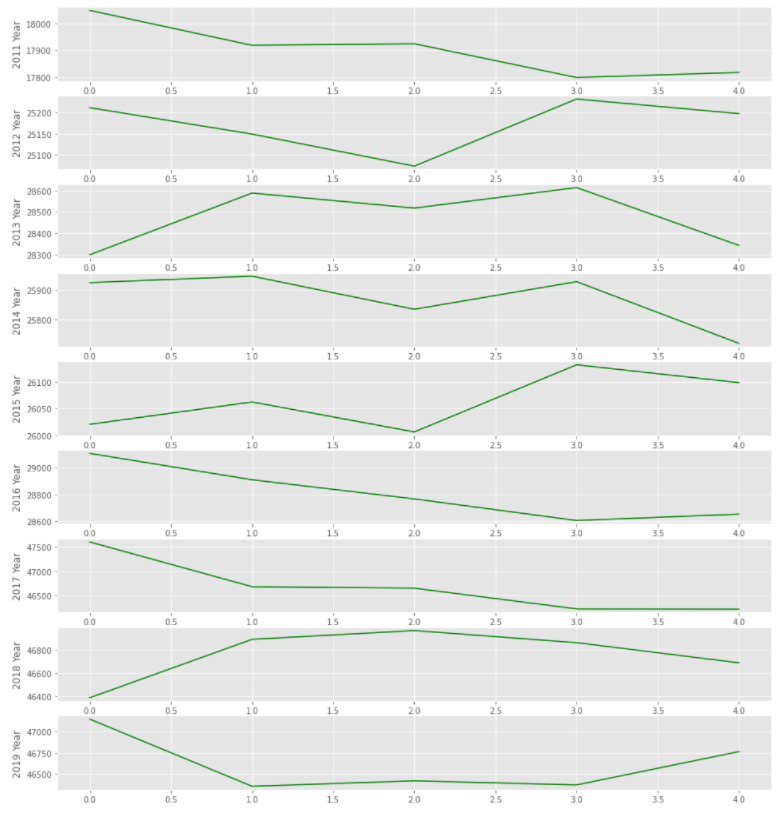
주가 변동을 확인해보면 2012년 부터 2016년까지는 크게 변화가 없지만, 2016년 말부터 주가가 크게 상승하고 2018년에 최고점을 찍고, 다시 하락하였다가 2020년부터 다시 크게 증가하는 추세를 보인다.

### 연도별



주가 변동을 확인해보면 앞에서 확인해본 결과와 마찬가지로 2015년 까지는 주가 변동의 추세가 없지만, 2016, 2017년은 꾸준하게 증가하는 추세를 보인다. 그리고 2018년에 주가가 조금 떨어지는 경향이 있지만, 2019년에 다시 증가하는 경향을 보인다.

### 연도별, 요일별

주가 변동을 확인해보면 월,화,수,목은 큰 특징을 발견하지 못했지만 2019년을 제외한 대부분 연도에서 금요일은 전날인 목요일에 비해 주가가 떨어지는 경향을 보인다.

## Analytics

주가 데이터는 영향을 주는 요인들이 많고 불규칙적이기 때문에 비정상시계열로 분류된다. 비정상시계열을 분석하기 위해서는 영향을 주는 여러 요인들을 파악하고 제거하여 정상시계열로 변환 후 예측을 진행해야 한다.

### ACF, PACF

삼성전자 종가가 어떤 시계열 모형을 따르는지 확인하기 위해 시계열분석에서 확인하는 값인 ACF(AutoCorrelation Function, 자기상관함수)와 PACF(Partial AutoCorrelation Function, 편자기상관함수)를 살펴보자.

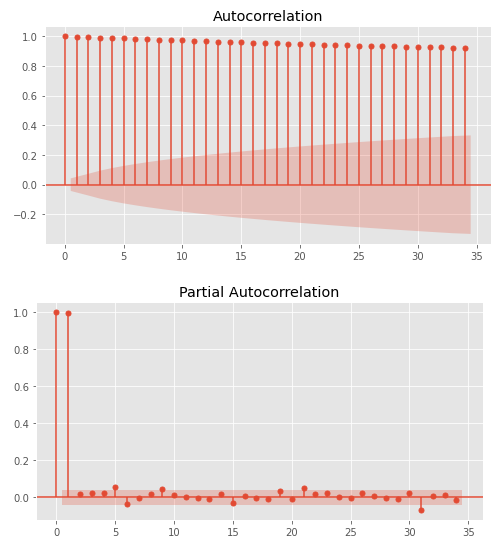
> ACF, 시차에 따른 자기상관, 정상 시계열은 빠르게 0으로 수렴

PACF, 시차에 따른 편자기상관, 시차가 다른 두 시계열 데이터 간의 순수한 상호 연관성

테이블이(가) 표시된 사진

자동 생성된 설명

AR 모형은 ACF는 천천히 감소하고, PACF는 처음 시차를 제외하고 급격히 감소한다. 반대로 MA 모형은 ACF는 급격히 감소하고, PACF는 천천히 감소한다. ACF와 PACF를 계산해 데이터가 어떤 시계열 모형인지 판단하는 기준이 된다.



삼성전자 주가 데이터의 ACF와 PACF를 확인해 본 결과.

ACF는 천천히 감소하고, PACF는 처음 시차를 제외하고 급격하게 감소하는 형태로 보아 AR모형에 가깝다고 볼 수 있다.

### ARIMA

주가 데이터는 변동이 크고 불규칙한 패턴을 보이기 때문에 AR모형과 MA모형을 합친 ARIMA모형으로 시계열 분석을 시행해보려 한다. 우선, Auto\_ARIMA를 시행하여 최적의 (p,d,q)값을 찾아내고 찾아낸 값으로 모델을 생성하고 유의성을 판단하는 과정으로 진행해보겠다.

>ARIMA(p, d, q)모형은 d차 차분한 데이터에 AR(p)모형과 MA(q)모형을 합친 모형이다. 대부분의 데이터는 1차 차분만으로 정상시계열로 변환된다.

### AutoARIMA

테이블이(가) 표시된 사진

자동 생성된 설명

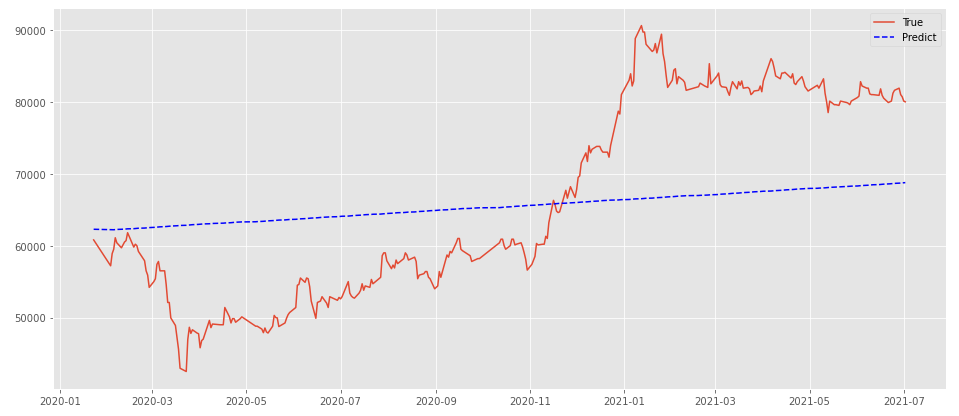
AutoARIMA 결과, ARIMA(0,1,2) 모형이 최적의 AIC값을 갖는 모델로 확인되었다.

텍스트, 영수증이(가) 표시된 사진

자동 생성된 설명

ARIMA(0,1,2)를 시행한 결과 MA(2)만 유의하다고 판정되어 주가 변동을 설명하기에 적합한 모형이 아니라 판단되었고, 모든 변수가 유의한 모형을 찾아보니 ARIMA(2,1,1)모형이 상수항을 제외한 모든 변수가 유의하면서 최적의 AIC를 갖는 모델로 최종모형을 확정했다.

### Predict

최종적으로 모델이 예측한 결과와 실제 2020년 1월 23일부터 2021년 7월 2일까지의 주가를 시각화하여 비교한 결과, 예측값은 주가가 최종적으로 증가하는 추세를 보인다고 판단했는데, 실제 데이터와 비교해보니 실제 값도 전체적인 추세를 보면 점차 증가하는 형태를 보이고 있었다. 즉, ARIMA를 통해 주가의 정확한 값이나 주기적인 변동을 정확하게 파악하진 못하지만, 주가가 증가할지 감소할지 경향정도는 어느정도 예측이 가능하다.

## Result

시계열 분석 결과, 주가 데이터는 일정한 패턴이 없기 때문에 주가만으로 정확한 예측은 불가능하다고 판단되었다. 정확한 예측이 불가능해도 장기적인 측면에서 시계열의 추세가 증가인지 감소인지는 확인이 가능했고, 결과 또한 추세를 어느정도 비슷하게 예측하였다. 반대로 단기적인 측면에서 예측 결과가 좋지 못해 시계열 분석은 장기적인 추세를 확인하는 정도 까지만 활용 가능할 것으로 보인다.

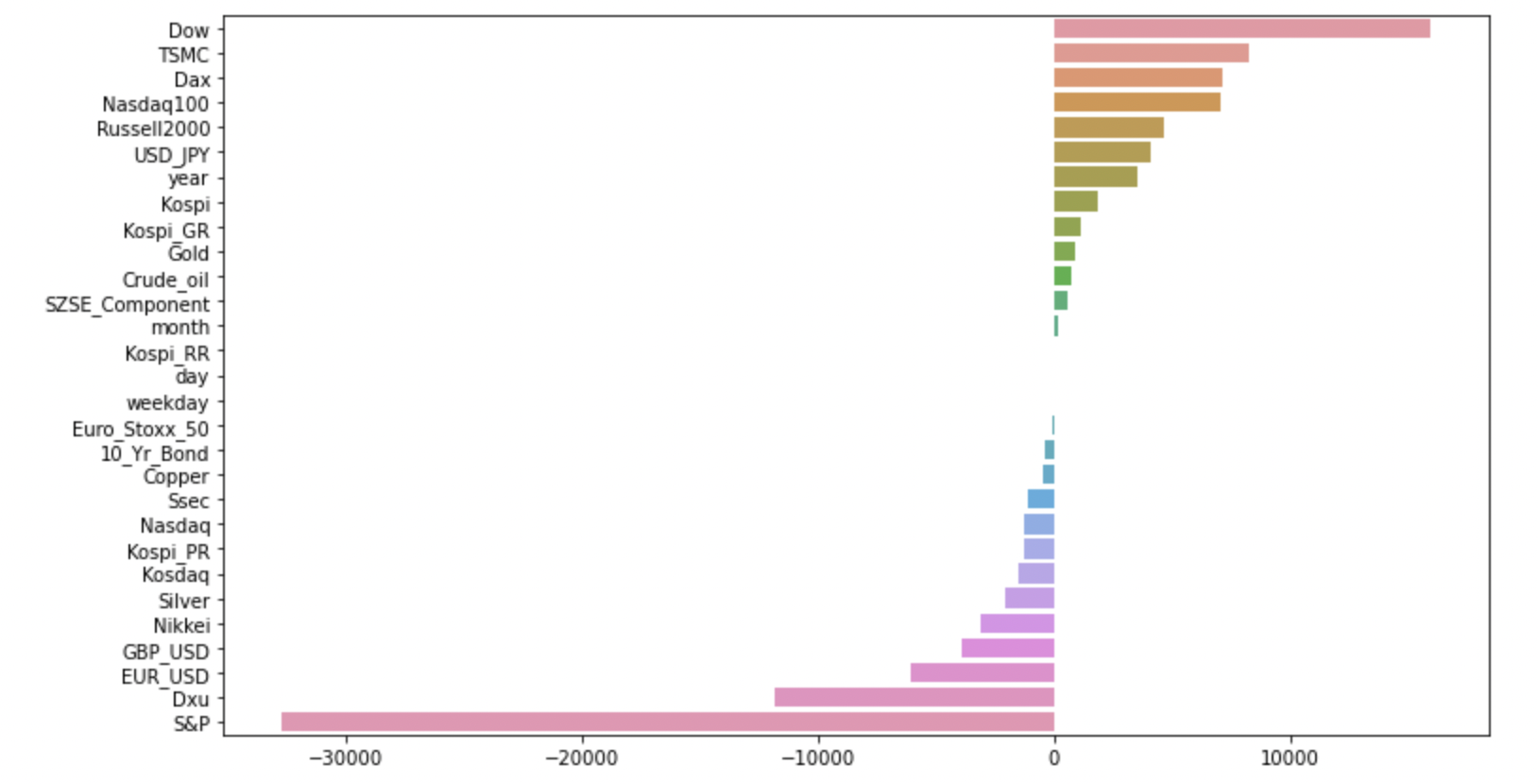
# 2. 머신러닝(회귀)으로 주가 예측이 가능한가??

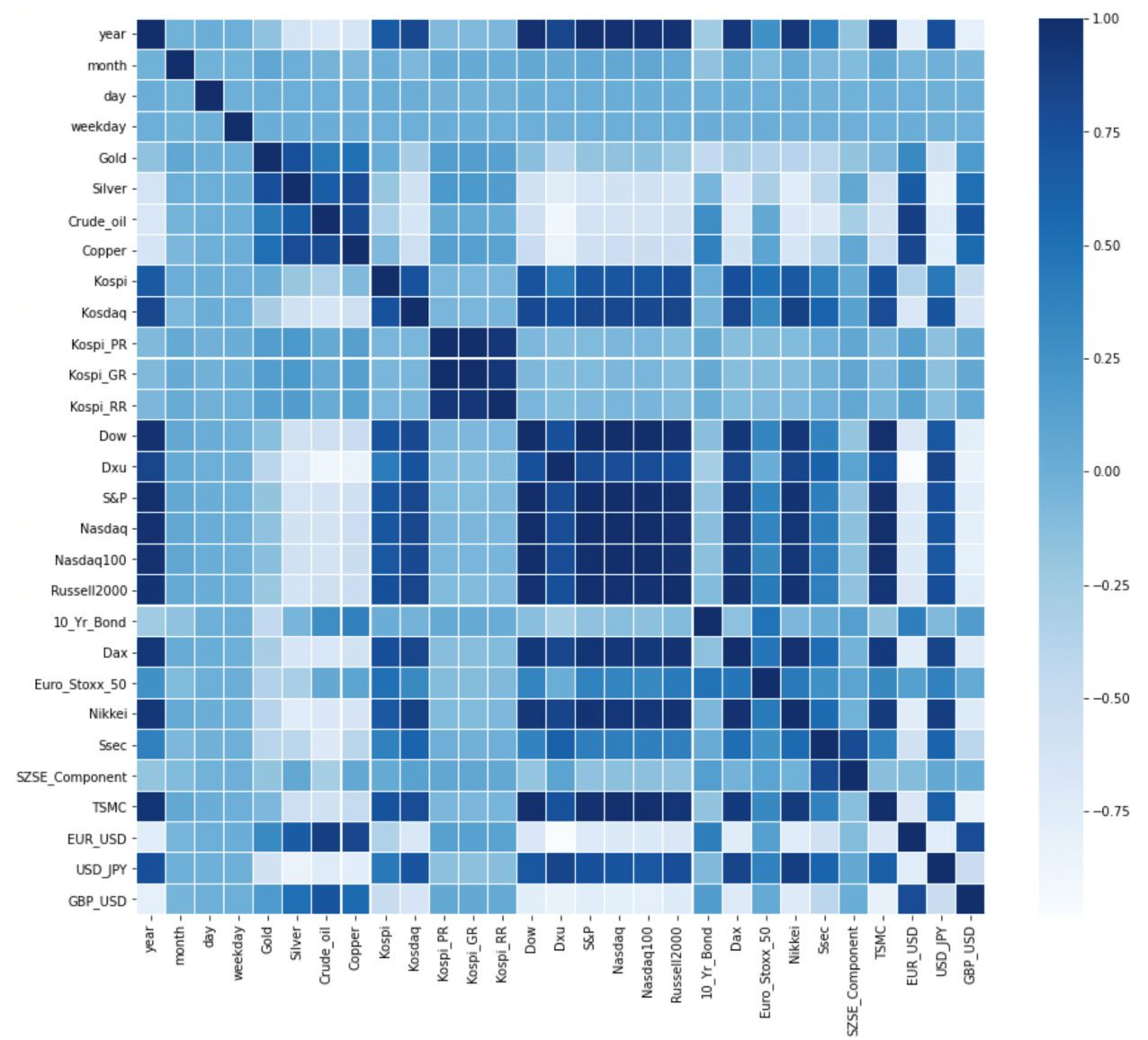
## Data

## Modeling

주가에 영향을 주는 인과관계에 있는 변수나 주가와 같이 증가하거나 감소하는 상관성이 높은 변수들이 있다면 주가 예측이 가능한 회귀식을 만들 수 있을 것이다. 시계열 분석으로 추세를 예측했다면 머신러닝으로 정확한 값을 예측하는 회귀식으로 단기적인 측면에서 주가 예측을 하고자 한다.

우선 모든 변수들을 넣고 회귀분석을 시행한 결과, 시계열 분석과 마찬가지로 추세는 따라가지만 값을 정확하게 맞추진 못해 결과적으로 아쉬움이 있었다.

머신러닝으로 만들어진 모델의 계수를 뽑아보니 최대 10,00이 넘는 값부터 최소 -30,000이하의 값까지 넓은 범위에 걸쳐 있었고, 이렇듯 계수의 편차가 커서 정확한 회귀 식 구현이 어려운것으로 보인다.



다음으로 모든 변수들의 상관관계를 시각화한 결과, 비슷한 지수끼리는 높은 상관성을 보이고 있었다. 그림에서 보듯 금, 은, 구리 등 원자재 값끼리 높은 상관성이 나타났고, 코스피와 코스닥도 높은 상관성을 보였다. 한가지 흥미로운 점은 미국과 유럽의 지표들은 대부분 상관성이 높게 나왔고 이는 모델링 과정에서 다중공선성의 문제를 발생시켜 모델 성능을 떨어뜨리는 원인으로 작용했을 것이다. 따라서 비슷한 변수를 제거하고 관련이 없는 변수를 제거해주는 작업이 필요하다고 판단했다.

각각의 변수를 좀 더 깊게 알아보고 선택하기 위해 주가를 기준으로 변수들의 상관관계를 시각화해보고 산점도를 통해 분포를 확인해보고자 한다.

테이블이(가) 표시된 사진

자동 생성된 설명

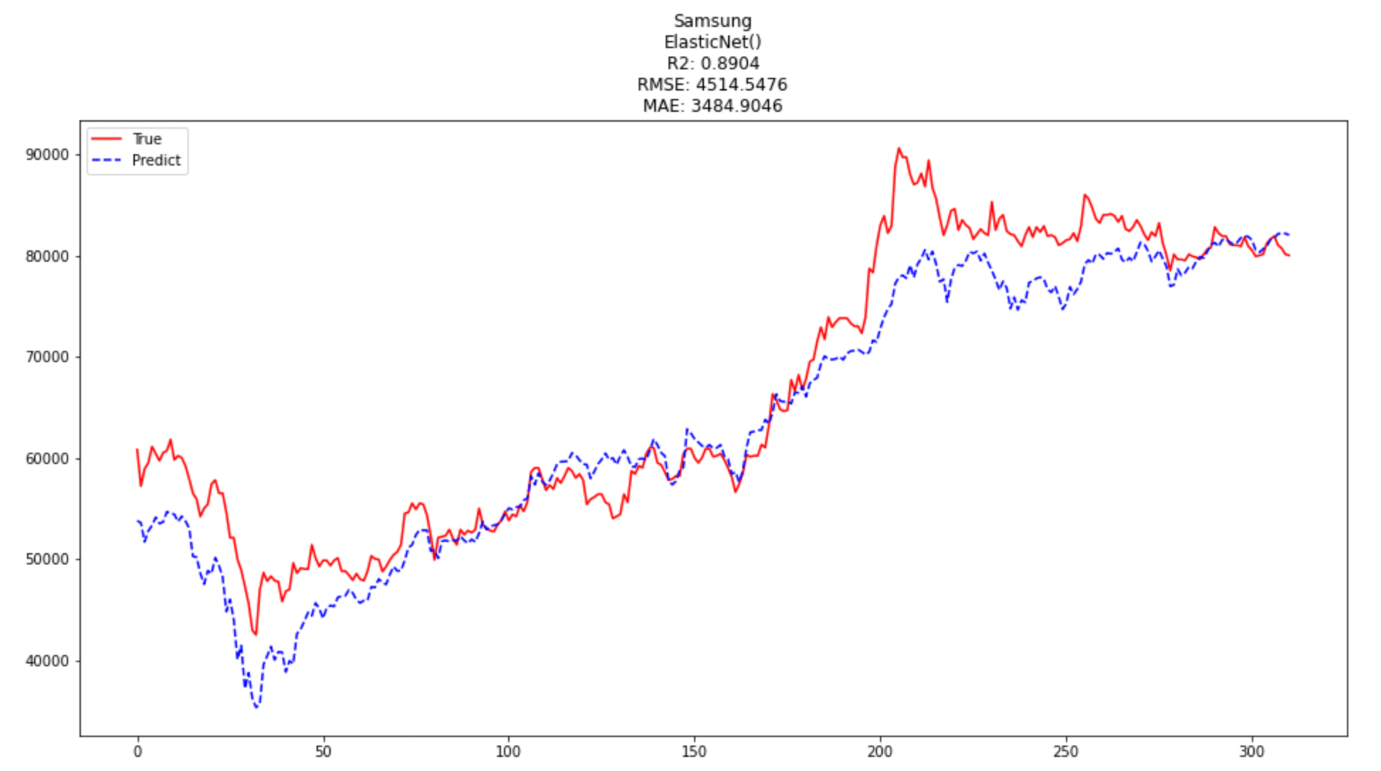
우선, 상관관계를 보면 해외 지표들이 대부분 높은 관련성을 보인다. 가장 높은 상관성을 보이는 TSMC는 삼성전자의 경쟁업체이자 대만의 반도체 제조회사로 같은 업종이라 주가변동 또한 같이 오르고 내려 관련성이 높게 나왔고, 다른 지표들을 보면 국내 주가지표인 Kospi보다 해외 Nasdaq과 관련이 높다고 나왔다.

화살이(가) 표시된 사진

자동 생성된 설명전체 변수들과 주가의 산점도를 보면 같은 범주의 지표들은 주가와 분포가 비슷함을 볼 수 있다. 대부분 해외 지표들은 양의 상관을 갖는 형태이고 원자재 지표는 약한 양의 관련성을 보이고 있다.

마지막으로 앞에서 파악한 분포가 비슷한 지표들을 제거하고 주가와 상관성이 높은 지표와 처음 만들었던 모델의 계수가 높았던 지표를 최종 모델링에 사용할 변수로 확정했다.

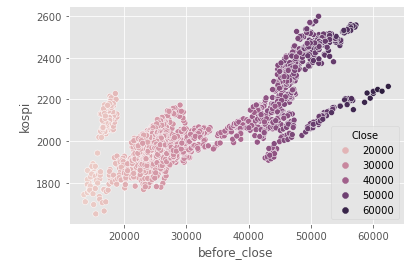
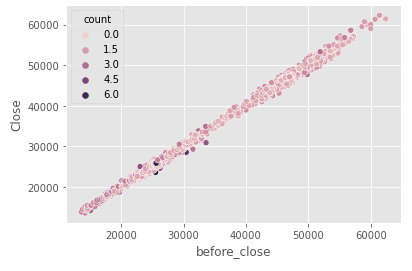
사용한 변수는 Year, Gold, Silver, Kospi, Dow, Nasdaq, Russell2000, 10\_Yr\_Bond, Dax, Euro\_Stoxx\_50, Ssec, TSMC, EUR\_USD, USD\_JPY, GBP\_USD를 사용했다.

최종결과, 변수를 탐색하여 모델링에 유의하다고 판단되는 값들만 넣어보니 모델의 성능이 확연히 좋아진 모습을 볼 수 있다. 모델을 만드는데 불필요한 변수를 제거해줌으로 좀더 정교한 예측이 가능해져 예측 정확도가 높아진 것이다. 결과가 기대보다 더 좋아 만족스러웠지만, 전날의 값만 사용한 모델이라 예측을 하더라도 다음날의 주가만 예측이 가능한 점이 아쉬웠다. 결국 시계열 분석과는 반대로 장기적인 측면의 분석이 불가능하다는 한계점에 다시 부딪혔다.

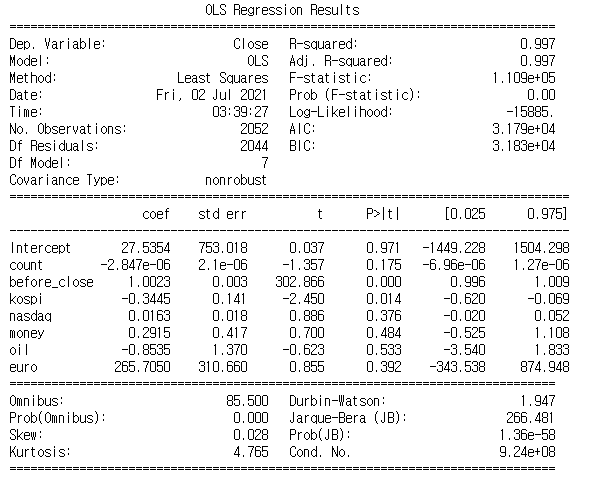
# 3. 통계(회귀) 와 머신러닝(회귀)의 비교

위의 머신러닝 방식은 cost function을 최소화하는 방식으로 회귀식을 찾아갔다면 통계 방식은 OLS(Ordinary Least Squares) 회귀식을 사용하였다. 최소제곱법을 이용하기 위해서는 가우스-마르코브 조건(Gauss-Markov assumption)을 만족시키고 , BLUE(best linear unbiased estimator, 최량선형[불편추정량](https://namu.wiki/w/%EB%B6%88%ED%8E%B8%EC%B6%94%EC%A0%95%EB%9F%89))임을 증명할 수 있어야한다. 따라서 많은 가정들이 따르므로 머신러닝 방법에 비해서 유연하지 못한 특징을 가지고 있다.

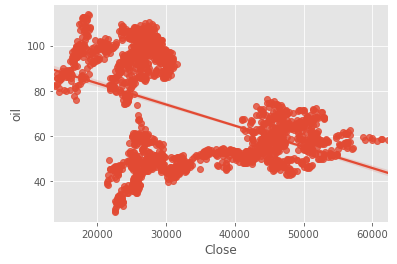
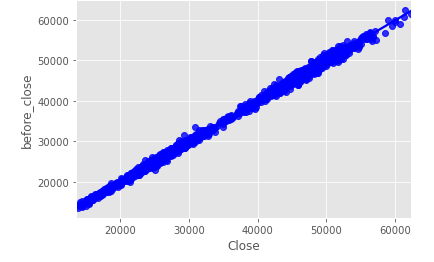
전날 종가와 종가를 count변수 기준으로 본 산점도이다. 영향도가 거의 1에 가까운 것을 볼 수 있다. 종가가 영향을 준다는 것을 확인하기 위해 다른 feature로 확인해보니 비교적 영향을 덜 주는 것을 알 수 있다.



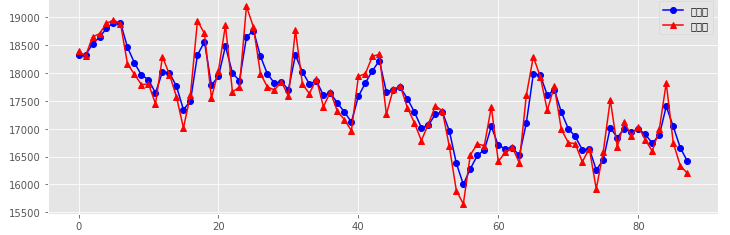
Python3의 Stats의 패키지의 OLS 방식을 사용하면 결과 값이 아래와 같이 나온다.



먼저 변수별 p값을 보면 <0.05가 유의하기 때문에 범위안에 포함된 kospi값과 before\_close값 만이 유의하다 볼 수 있다. 적합한 회귀직선과 95%의 신뢰구간으로 그려진 산점도를 살펴보면 아래와 같다. 이를 비교하기 위해서 p값이 가장 커서 유의하지 않다고 판단되는 독립변수 oil의 회귀직선 및산점도를 살펴보면 회귀직선을 따른다고 보기는 힘들다.



Adjusted R^2 (수정된 결정계수) 값이 0.997로 매우 높은 편인데 이는 여러 가정들을 만족하며 최소제곱법을 이용하여 최대한 회귀식을 맞췄기 때문에 모델의 결정력은 높다. 따라서 다음과 같이 예측값과 실제값이 비슷하게 나온다.



그러나 분명한 한계점은 feature들이 유의미한가를 p-value로 따져 선택하기 때문에 변수 선택법을 이용하지 못한다. 따라서 내가 중요하다고 생각되는 독립변수들을 추가하여봤자 p-value값을 고려하면 사용하지 못하게 될 확률이 높음을 알 수 있다. 따라서 우리는 머신러닝 방법을 채택하기로 하였다.

|  |  |  |
| --- | --- | --- |
|  | **통계적 회귀** | **머신러닝 회귀** |
| 수식 개념 | OLS(최소제곱법) | Argmin Cost Function |
| 장점 | -BLUE estimator로 여러 가정을 만족하여 유의미한 feature만을 선택하므로 모델 결정력이 높아짐 | -사용자가 넣고 싶은 feature들을 모두 활용할 수 있음.  -변수선택법(stepwise,backward,forward) 이용가능 |
| 단점 | feature선택이 p-value값으로 이루어지기 때문에 원하는 변수를 못 넣을 수도 있음. | -모델 결정력 및 예측력이 떨어짐.  -인공신경망(neural network)의 한계점으로 중간과정을 완벽히 보기 힘듬. |

두 방법의 비교 표

# 4. 머신러닝(분류)으로 주가 상승/하락 예측이 가능한가?

## Data

분류를 시행하기 위해 기존의 데이터에 따로 Labeling 작업이 필요했다. 전날에 비해 주가가 상승했으면 1을 하락했으면 0으로 기존 데이터에 Label을 추가해 최종목표로 Label 값을 맞추는 모델을 만드는 것을 목표로 하였다. 좋은 모델을 만든다면 다음날 주가가 상승인지 하락인지 예측이 가능해지기 때문이다.

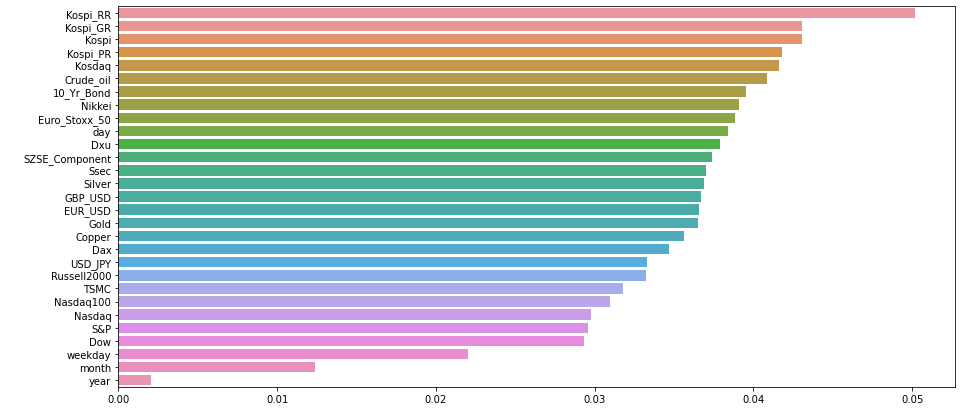
## Modeling

분류 모델링 또한 회귀와 비슷한 방식으로 우선 간단한 모델을 돌려보고 판단해보려 한다.

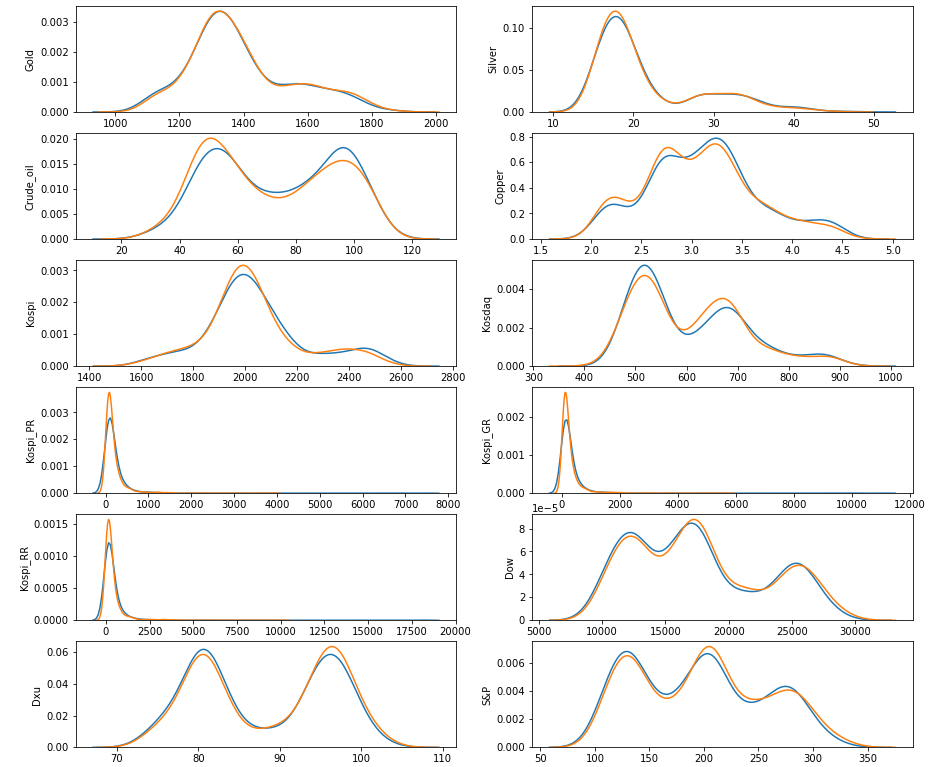
텍스트, 영수증이(가) 표시된 사진

자동 생성된 설명

간단하게 RandomForest를 시행한 결과, 정확도는 0.52로 좋지 못한 성능을 보여준다. 그리고 다른 지표들을 보면 Recall 값은 그나마 높게나와 실제로 상승한 결과 중 진짜 상승을 잘 맞춘다고 판단했고, 주가의 상승을 맞추기는 어느정도 가능하리라 판단했다. 하지만 주가의 상승/하락으로 분류를 시행할 때 개인적으로 상승을 맞추는 것보다 하락을 맞춰 최대한 손해가 없도록 하는게 좋은 모델이라 생각해 결론적으로 위의 모델은 주가 예측에 좋지 못한 모델이라 생각했다.



모델에서 변수 중요도를 확인해 본 결과, 모든 변수에서 0.1보다 낮은 값을 보이며 분류의 기준이 되는 좋은 변수는 없다는 것이 추측 가능했고, 회귀에서 시도했던 방법처럼 좋은 모델을 만들기 위해 변수를 추려봤자 좋은 결과를 얻지 못할것이다.



Label값인 상승/하락에 따른 각 변수들의 분포를 시각화해 보니 여기서도 유의한 차이를 갖는 변수가 없어 보였고, 앞선 판단과 마찬가지로 변수 추출로 인한 큰 효과를 기대하기 어려울것으로 보인다.

그렇다면 다음날의 주가 상승과 하락을 여러 주가지표인 변수들로 예측이 불가능한것인가?? 개인적인 결론으로 분류는 회귀와 다르게 타깃의 범위가 좁아 예측이 어려운것으로 보인다. 회귀는 주가 값은 결국 전날의 흐름과 크게 다르지 않고, 특이한 경우가 아니라면 전날과 큰 차이가 없을 것이다. 그러나 분류는 전날 지표들로 상승했냐 하락했냐는 매일매일이 다른 결과가 나올것이고 결국 변수로 사용한 지표들의 불확실성이 높아 성능 좋은 모델을 만들 수 없는 것이다.