주식 종가 예측 프로젝트

팬데믹이 바꿔놓은 한국 주식 시장 분석

업종간 동향과
시사점을 중심으로

목 차

1.개요

팀 구성 및 역할

주제 선정 배경

2-track 소개

2.감성분석

분석 방향

뉴스 데이터

SNS 데이터

3.종가예측

분석 방향

EDA & 전처리

전체 data 모델링

4.평가 및 Q&A

평가 및 제언

결론

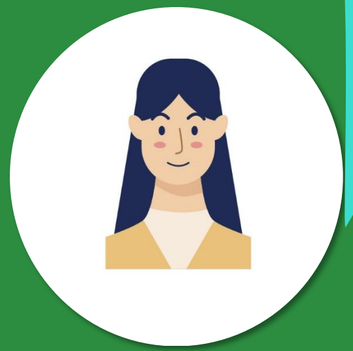
Q&A

01 개요

01
개요

팀 구성 및 역할

종가예측



박 ○ ○

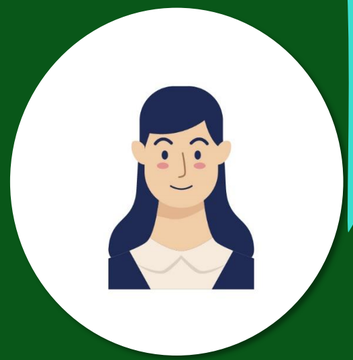
조장
모델링
데이터 상관분석

종가예측

이 ○ ○

EDA
모델링
경제지표 검토

종가예측



황 ○ ○

EDA
데이터 수집
종목 검토

감성분석

김윤명

데이터 전처리
뉴스 데이터 분석
코드 보완

감성분석

배 ○ ○

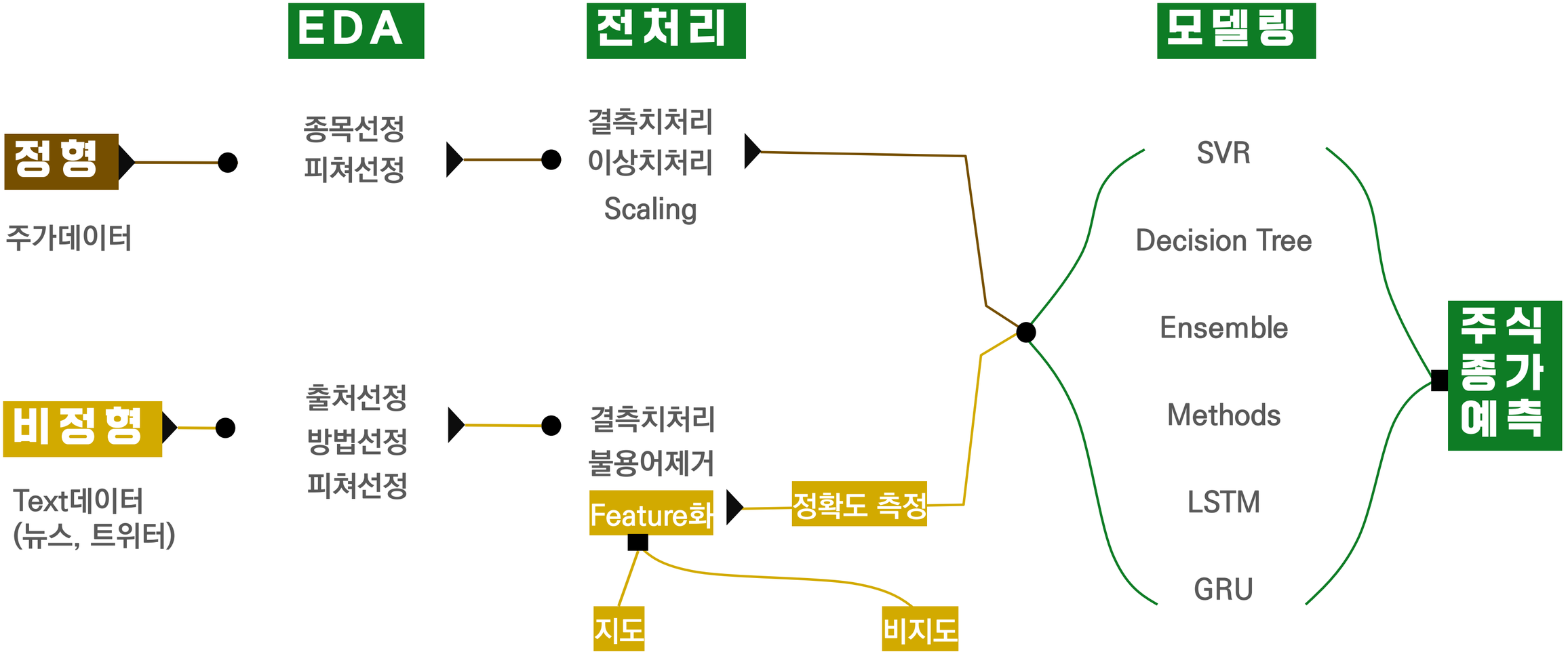
데이터 크롤링
트위터 데이터 분석
코드 보완

01
개요

주제 선정 배경

- 2020년 COVID-19의 확산으로 대규모 양적 완화 시행
- 주식시장 V자 반등 후 상승, 주식 입문자 대거 유입
- Fed의 테이퍼링, 수차례 금리인상 예고
- 과도기적 단계에서 개인투자자의 시장 대응 어려움
- 개인 투자자에게 각종 경제지표에 기반한 5일치 주식 종가예측 정보를 제공함으로써 개인투자자의 시장 대응력 강화

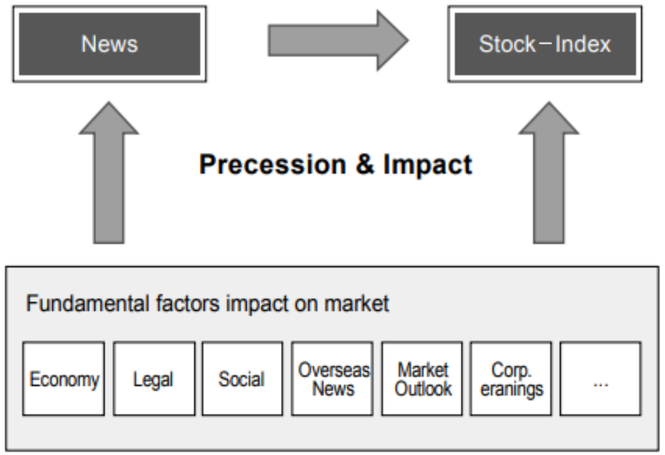
2-track 소개

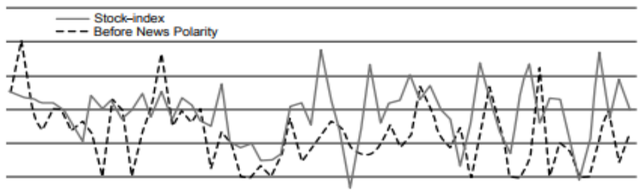

EDA
전처리
모델링
정형
주가데이터
종목선정
피쳐선정
결측치처리
이상치처리
Scaling
비정형
Text데이터
(뉴스, 트위터)
출처선정
방법선정
피쳐선정
결측치처리
불용어제거
Feature화
정확도 측정
지도
비지도
SVR
Decision Tree
Ensemble
Methods
LSTM
GRU
주식
종가
예측

02

감성분석

주식가격과 감성분석 NEWS & SNS



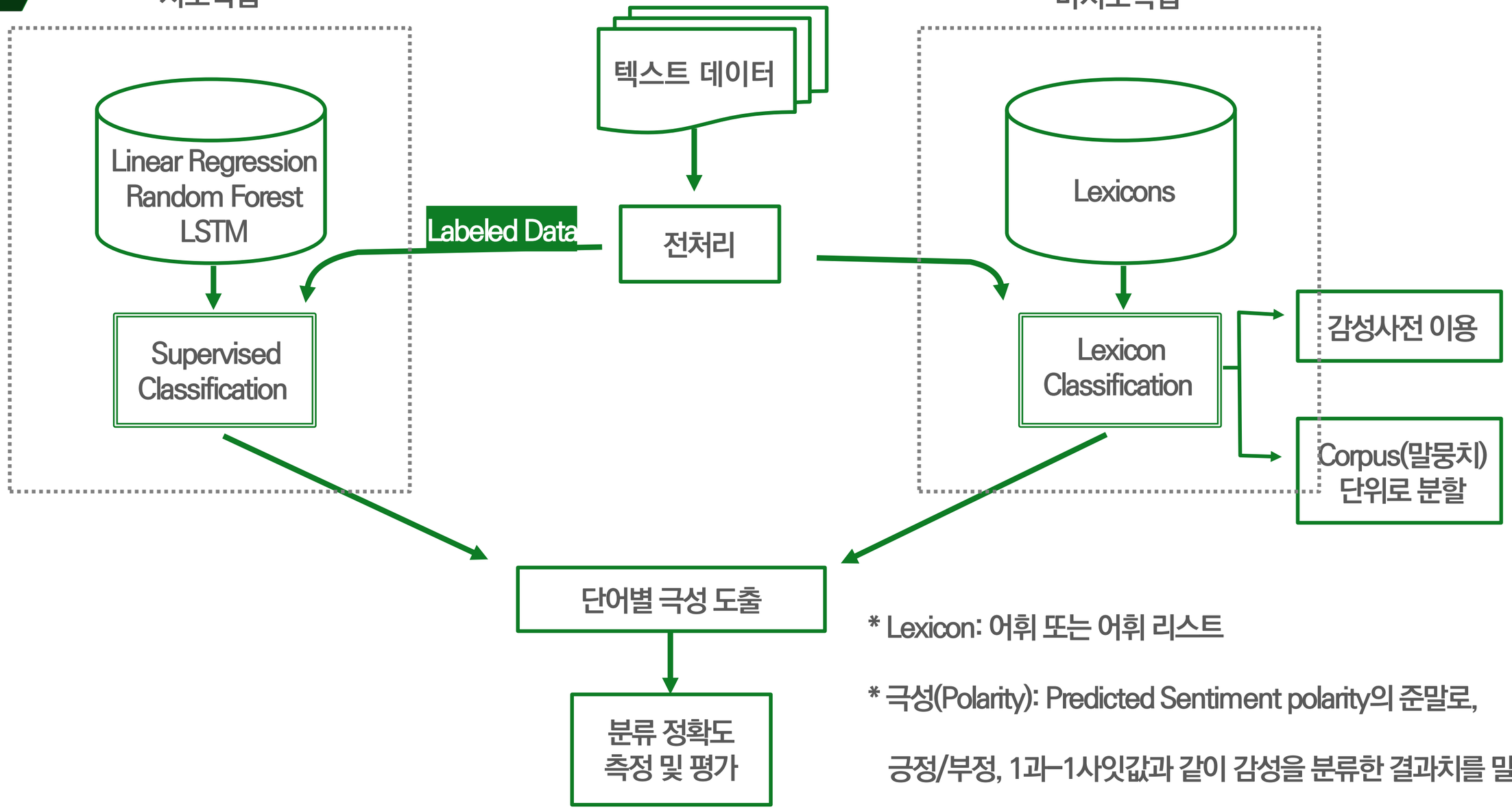


‘단순한 기술적 분석의 한계점으로 인한 감성분석의 필요성’

- 주식시장은 특유의 복잡한 가격결정 메커니즘으로 인해 주가의 변동을 시장 펀더멘탈의 변화로 설명할 수 없는 경우가 자주 발생한다.
- 투자자의 주관적인 심리작용과 작전세력, 기사 등에 의해 예측하지 못하는 변수 발생 가능성이 있다.

뉴스 데이터와 SNS데이터를 활용한 감성분석을 새로운 Feature값으로 설정

분석방향



* Lexicon: 어휘 또는 어휘 리스트

* 극성(Polarity): Predicted Sentiment polarity의 준말로,

  긍정/부정, 1과-1사잇값과 같이 감성을 분류한 결과치를 말함.

주식가격과 감성분석

감성 분석은 크게 지도학습 & 비지도학습 방식으로 수행된다.
결론적으로 두 학습 모두 감성점수라는 결과값을 도출하는 것이 목표이다.

지도 학습

*KOSPI200 값을 사용하여 데이터를 라벨링하는 경우

+ KOSPI 200

비지도 학습

*기존 모델을 활용하여 감정사전을 통해 감정 점수를 도출하는 경우

+ Textblob
+ Vader

+ ko-electra

단어별 감성점수로 결과값을 도출

뉴스 데이터

EDA

데이터의 특성을 파악하여 출처, 방법, feature를 결정

0210101	한국경제	김대훈(dae	막혔던 신용대출 재개 자영업자 '숨통' 트이
20210101	한국경제	김동현(3cc	세계 40여곳이 '소형 발사체' 경쟁 머스크는 1(
20210101	한국경제	정연일(nei	'입주 임박' 목동아델리체, 전세 매물 20개뿐
20210101	한국경제	도병욱(do(	아이오닉5 기아 CV '신기술 풀충전' 차세대 전
20210101	한국경제	최진석(iskr	나이 지역 가리지 않고 "새해에도 집값 오른다
20210101	한국경제	정연일(nei	부평 청천동 '대단지' 분양대결
20210101	한국경제	최예린(ran	작년 매출 신기록 쓴 카카오 엔씨 "올해 더 좋
20210101	한국경제	최진석(iskr	HUG, 경기 양주 분양가관리지역 지정
0210101	한국경제	설지연(sjy(	코스피 기업 시총, GDP 넘어섰다

출처

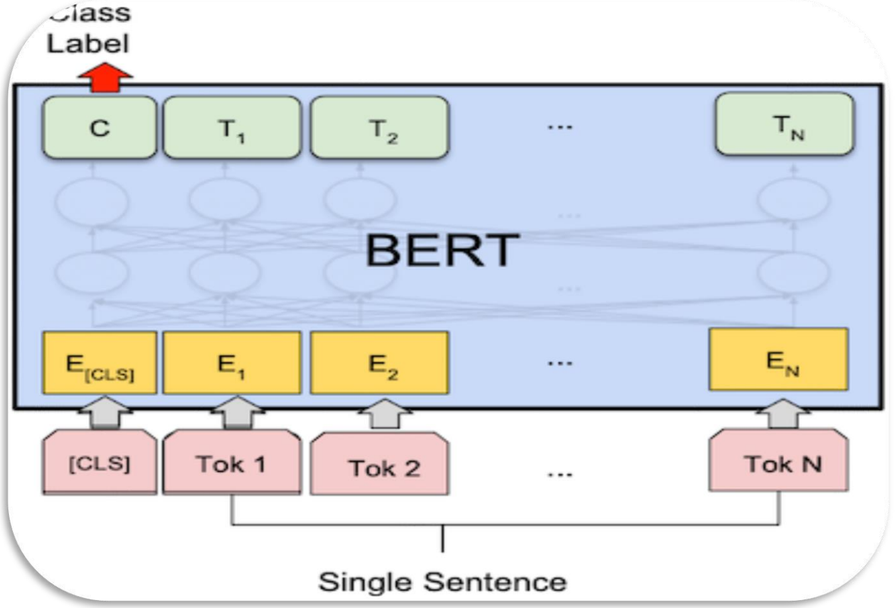
빅카인즈에서 한국일보 경제면 3년치 기사를 다운로드하여 제목만 추출

뉴스 데이터

EDA

데이터의 특성을 파악하여 출처, 방법, feature를 결정

Date	title
20190101	Is it true that President Moon Jae-in's remarks that "consumption indi
20190102	Samsung Display's investment expansion 'good news' Mobile phone
20190103	[Editorial] The KOSPI that collapsed in the 2000s is ultimately a corpo
20190104	"Gangnam-do 2 billion won is too expensive" The sale of Gaepo Luce
20190105	[Photo] BLACKPINK Jisoo, 'Chic Eyes' (Golden Disc) 312,000 jobs in th
20190106	Kia Motors to support 400,000 won when replacing old diesel cars wi
20190107	1 household, 1 house tax-free benefits 2 years from the date of becc
20190108	KEB Hana Bank pursues the best bank through 'employee happiness,
20190109	[HanKyungRoboNews] 'New Pride' reached upper limit↑, stock price
20190110	[Exchange rate] Closed KRW/USD exchange rate 1,118.3 won (-3.8 wo
0190111	[HanKyungRobo News] 'Ananti' rises more than 10%, typical upward



방법

해외 감성분석 Library(Vader, Textblob)를 이용할 때에는 한〉영 번역 수행

뉴스 데이터

EDA

데이터의 특성을 파악하여 출처, 방법, feature를 결정

```python
# 새로운 칼럼 생성
# (Price : 당일 대비 다음날 주가가 상승했으면 1, 하락했으면 0 표시)
df['Price'] = 0
for i in range(0, 247):
    if df['Close'][i] < df['Close'][i+1]:
        df['Price'][i] = 1
    else:
        df['Price'][i] = 0
df
```

지도

비지도

sent_score	blob_score	norm_score
Neutral	0.114785	0.249859
Positive	0.154085	0.25
Positive	0.134615	0.25
Positive	0.194742	0.249977
Negative	0.137807	0.207736

피쳐

지도학습의 경우 KOSPI 주가변동에 따른 Label 값을, 비지도학습의 경우 Library를 통해 산출한 score값을 이용

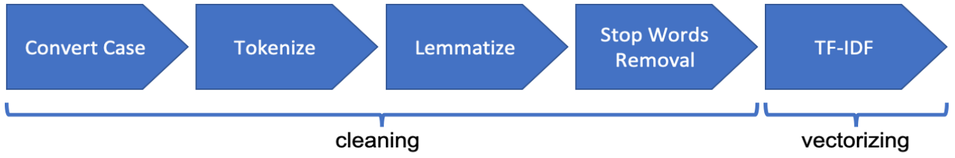
뉴스 데이터

전처리

결측치 처리, 불용어 제거 등을 통해 feature로 만들기

20190102	Samsung Display's investment expansion 'good n...	0.0
20190103	[Editorial] The KOSPI that collapsed in the 20...	1.0

TextBlob



text
Today is a good day
Today is a bad day
Today is ok

bad	be	day	good	ok	today
0.00000	0.391484	0.504107	0.66284	0.000000	0.391484
0.66284	0.391484	0.504107	0.00000	0.000000	0.391484
0.00000	0.453295	0.000000	0.00000	0.767495	0.453295

결측치 처리

KOSPI 지수 변동값의 결측치를 채울 때 바로 전날 변동값을 이용

불용어 제거

주식 data의 경우 일반 data와 불용어 개념이 다소 상이한 부분이 있어 최소한도로 진행

피쳐화

극성(Polarity)도출을 위해 KOSPI 지수 변동값을 Label로 이용하거나 감성사전 Library 이용

정확도 측정

극성 분류 정확도를 측정함으로써 feature의 신뢰성 확보

```
[ ]
1 model = Sequential()
2 model.add(Embedding(vocab_size, 100))
3 model.add(Bidirectional(LSTM(100)))
4 model.add(Dense(1, activation='sigmoid'))
5
6 es = EarlyStopping(monitor='val_loss', mode='min', verbose=1, patience=4)
7 mc = ModelCheckpoint('best_model.h5', monitor='val_acc', mode='max', verbose=1, save_best_only=True)
8
9 model.compile(optimizer='rmsprop', loss='binary_crossentropy', metrics=['acc'])
10 history = model.fit(X_train, Y_train, epochs=15, callbacks=[es, mc], batch_size=256, validation_split=0.2)
```

Mecab &LSTM

일본에서 만든 범용 형태소 분석기 엔진으로, mecab-ko-dic 한국어 형태소 사전과 결합하여 LSTM으로 측정

정확도 측정

극성 분류 정확도를 측정함으로써 feature의 신뢰성 확보

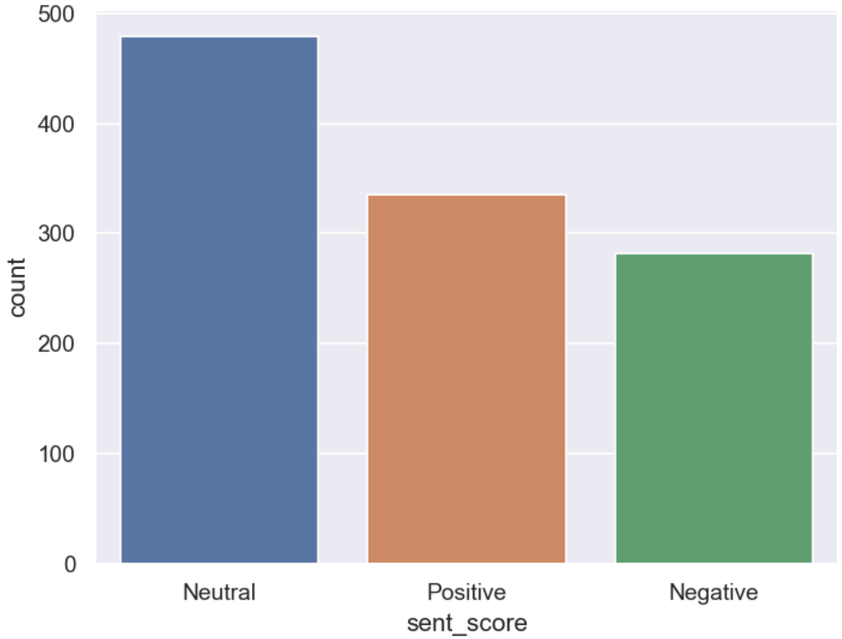
```
Epoch 4/15
147/147 [==============================] - ETA: 0s - loss: 0.2237 - acc: 0.9151
Epoch 00004: val_acc did not improve from 0.85195
147/147 [==============================] - 46s 313ms/step - loss: 0.2237 - acc: 0.9151 - val_loss: 0.3683 - val_acc: 0.8447
Epoch 5/15
147/147 [==============================] - ETA: 0s - loss: 0.2025 - acc: 0.9234
Epoch 00005: val_acc did not improve from 0.85195
147/147 [==============================] - 46s 313ms/step - loss: 0.2025 - acc: 0.9234 - val_loss: 0.3837 - val_acc: 0.8440
Epoch 6/15
147/147 [==============================] - ETA: 0s - loss: 0.1832 - acc: 0.9318
Epoch 00006: val_acc did not improve from 0.85195
147/147 [==============================] - 46s 314ms/step - loss: 0.1832 - acc: 0.9318 - val_loss: 0.3996 - val_acc: 0.8387
Epoch 00006: early stopping
```

Mecab &LSTM

일본에서 만든 범용 형태소 분석기 엔진으로, mecab-ko-dic 한국어 형태소 사전과 결합하여 LSTM으로 측정

정확도 측정

극성 분류 정확도를 측정함으로써 feature의 신뢰성 확보



1. 극성 나누기(Neutral, Positive, Negative)
2. 어간 추출(Stemming)과 표제어 추출(Lemmatization)
3. CountVectorizer를 이용해 BoW Model 만들기

- BoW(Bag of Words) 란 단어들의 순서는 전혀 고려하지 않고,
  단어들의 출현 빈도(frequency)에만 집중하는 텍스트 데이터의 수치화 표현 방법을 말함

NLTK& models

NLTK(Natural Language Toolkit) 자연어 처리 패키지와 Linear Regression, Random Forest 이용

정확도 측정

극성 분류 정확도를 측정함으로써 feature의 신뢰성 확보

Linear Regression

```
confusion matrix:
 [[65 12  1]
 [11 67 35]
 [ 0 24 59]]

Best accuracy: 0.800411 using
{'C': 0.01, 'penalty': 'l2', 'solver': 'newton-cg'}
```

Random Forest

```
confusion matrix:
 [[61 17  0]
 [ 6 91 16]
 [ 0 32 51]]

Best accuracy: 0.771212 using
{'max_features': 'sqrt', 'n_estimators': 1000}
```

NLTK& models

NLTK(Natural Language Toolkit) 자연어 처리 패키지와 Linear Regression, Random Forest 이용

EDA 데이터의 특성 파악을 위해 word-cloud 만들고 향후 분석 방향 수립

date	text
2019-01-01	김영수 월 일저는현재 최종 근로복지공단측에 불승인으로 인환 현대제철 직장내에 산재...
2019-01-02	김영수 월 일많은 합병증으로 고통을 받고 연예인 들과 인천내에 일부 현대제철 직장...
2019-01-03	포항교통정보센터 월 일동해안로 현대제철 삼거리 포스코문 구간 이동량 꾸준 원활 ...
2019-01-04	포항교통정보센터 월 일동해안로 현대제철삼거리 남구보건소 양방향 원활 주식무당 ..
2019-01-05	포항교통정보센터 월 일동해안로 현대제철 삼거리 포스코문 구간 원활 포항교통정보센...
...	...
2021-12-22	포항교통정보센터 월 일동해안로 포스코문 현대제철삼거리 원활 년 월 일현대제...
2021-12-23	포항교통정보센터 월 일동해안로 포스코문 현대제철삼거리 원활 포항교통정보센터 ...
2021-12-24	월 일민주당 이낙연 의원 최근 년간 전기요금 할인 혜택을 가장 많이 본 곳은 삼...
2021-12-25	포항교통정보센터 월 일동해안로 포스코문 현대제철삼거리 원활 이상현 월 일 ...
2021-12-27	포항교통정보센터 월 일동해안로 포스코문 현대제철삼거리 서행

계획 Twitter에서 각 종목 키워드를 크롤링하여 3개년치 data를 수집한 후 Ko-electra 이용하여 감성점수 도출

문제점 해당 키워드 검색량이 불규칙해 시계열 데이터로서의 정확도 측정이 어렵고, 광고나 도배글이 너무 많음

03

증가예측

분석 방향

가설1 코로나의 영향이 큰 종목에 대해서는 코로나 이후의 데이터로 분석하는 것이 바람직하다.

→ 직전 2년(코로나이후), 직전3년(코로나이전+이후)의 데이터를 생성해 각각 모델을 학습시킨 후 예측 성능 비교

가설2 주가는 과거데이터에 영향을 받는다.

→ 다양한 feature값을 적용하되, MA10(주가10일 이동평균) 변수를 추가로 설정하여 분석

가설3 감성데이터(비정형데이터)가 예측 성능에 유의미한 영향을 미칠 것이다.

→ 감성데이터를 포함/일부 포함/포함하지 않은 다양한 비교군을 설정하여 예측 성능 비교

“종목 선정 배경”

- KOSPI200 지수와 개별 종목 그래프를 비교분석하여 코로나와 관련이 있다고 판단되는 산업군별 후보 선정

  ex) 헬스케어: 면역력 중요성 커짐
  반도체 :폭락 후 주가회복
  철강·금속 : 공장가동 차질, 제조가격 부담

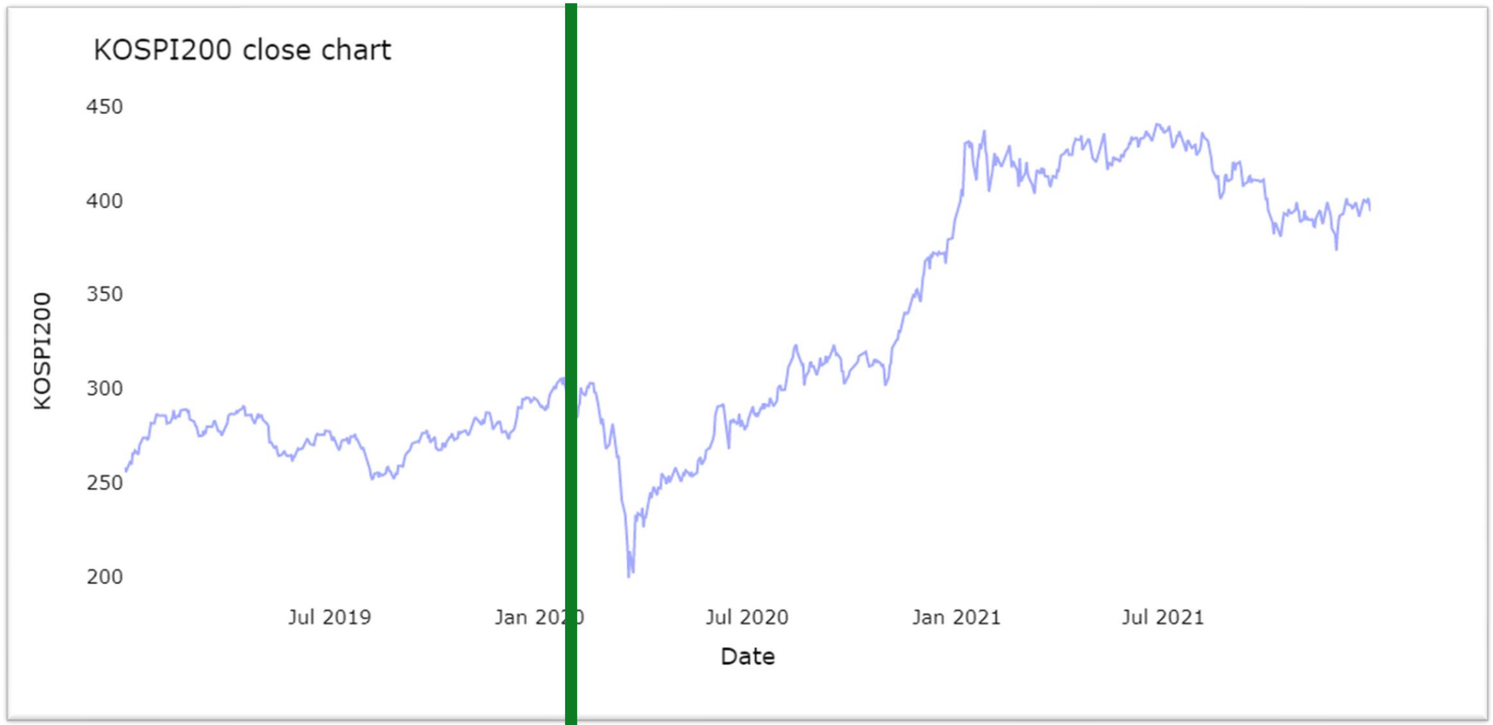
- 개별 이슈가 크거나 상장 후 3년이 되지 않은 종목 파악 → 종가 예측에 적합하지 않은 종목 후보 제외

- COVID-19와 관련하여 주가에 영향을 미치는 산업군을 긍정/부정/중립으로 나누어 KOSPI200 지수와 거의 비슷하게 움직이는 산업군(중립)으로 분류하여 최종 종목 선정

“ 종목 선정(KOSPI200) ”

시장을 대표할 수 있는 대형주 위주의 주가지수인 KOSPI 200을 기준으로 분석

KOSPI200 지수와 개별 종목 그래프를 비교분석하여
KOSPI200종목 내 3개 종목 선정



2020.01.20

- 삼성 바이오로직스
- 삼성전자
- 현대제철

“산업별 코로나 영향”

헬스케어 : 코로나19의 급속적인 확산으로 면역력 강화를 위한 건강기능식품에 대한 관심이 급격하게 늘어남

반도체 : 코로나로 인해 코스피 폭락 장을 맞았지만 해당 산업의 안정성으로 인해 주가 회복 후 꾸준히 증가

철강·금속 : 1) 코로나 사태 장기화로 인해 철강 전방산업 생산에 차질이 있음
2) 중국 내 공장 가동이 어려우며 철광석 가격의 상승세로 인해 제조 가격 부담이 커짐

“대표 섹터(Sector)별 분석 진행”

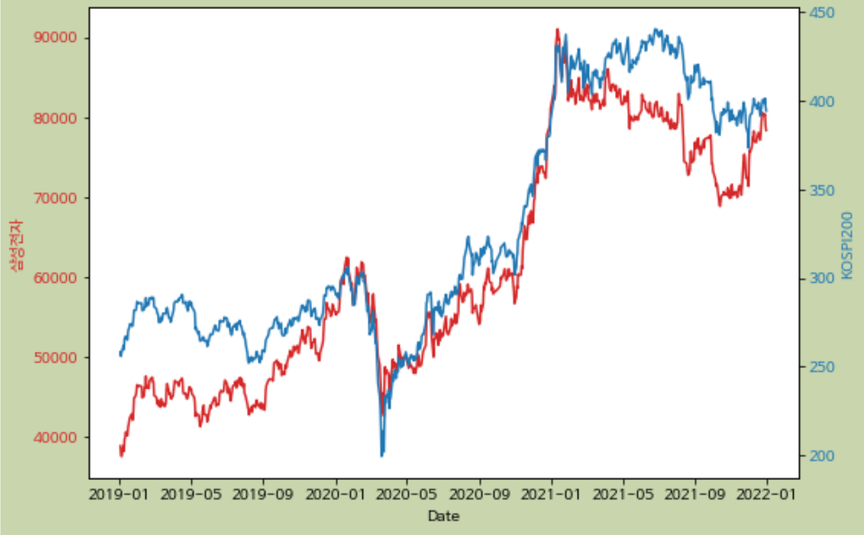
산업에 따라 COVID19의 영향이 다르게 작용하여 산업을 기준으로 종목 분석

Negative

현대제철(철강·금속)

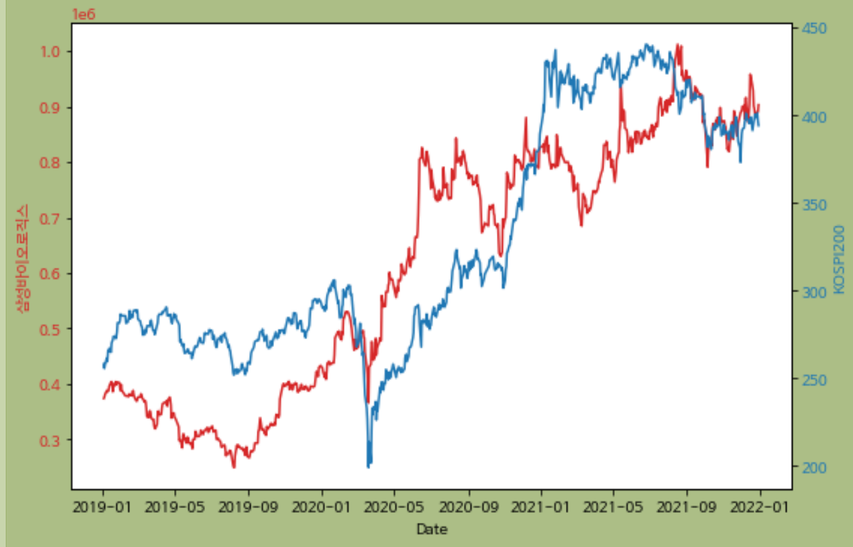
Neutral

삼성전자(반도체)



Positive

삼성바이오로직스(헬스케어)



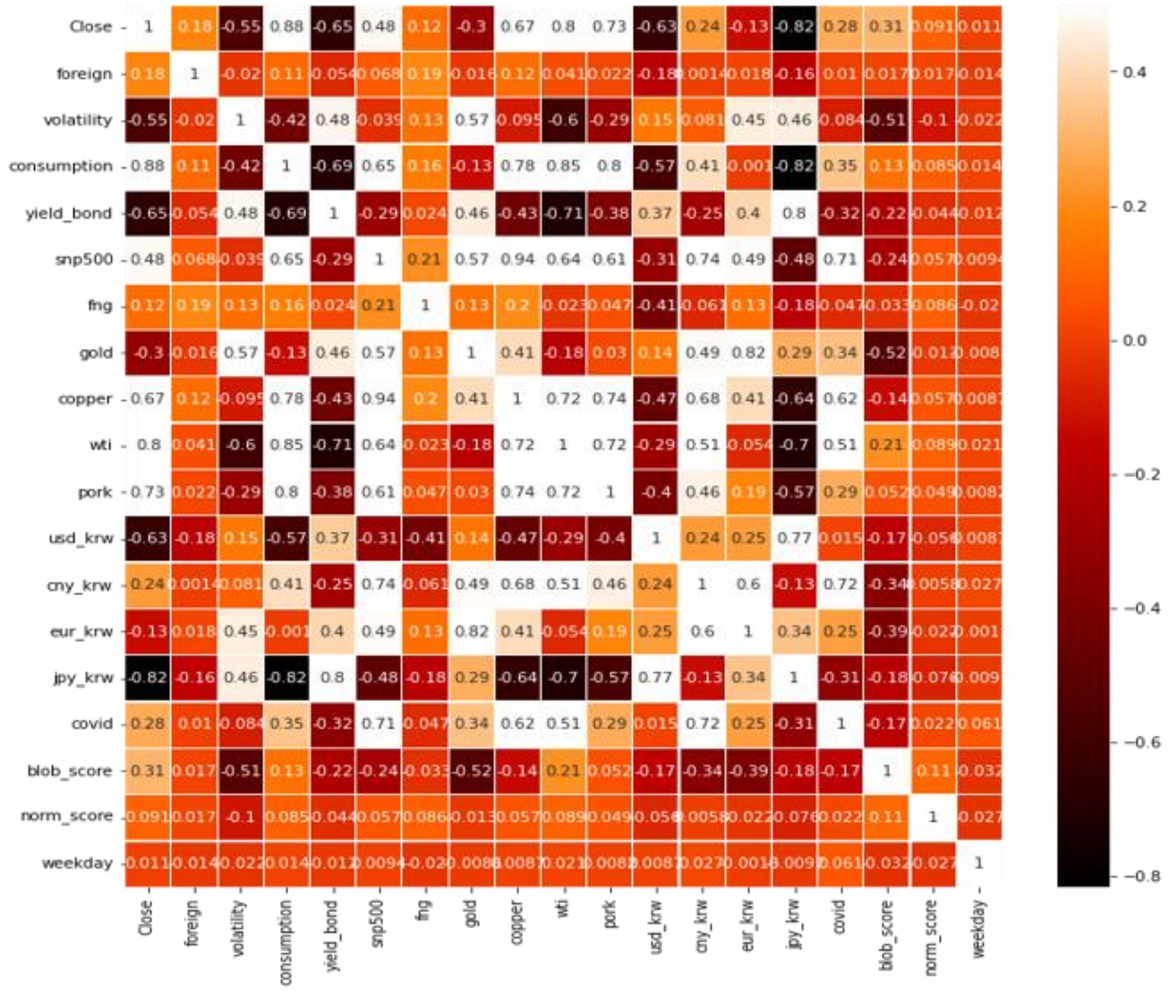
"Feature값 설정"

- 모든 데이터는 일별데이터로 수집
- 국내 증권시장에 영향을 주는 주요 거시경제 지표(Macroeconomic Indicators) 선정
- 그 외에 2020년 3월 V자반등시기를 기준으로 주가와 비슷한 흐름을 보인 데이터들을 수집

선행연구 요약

환율, 미국주가 국내시장에 유의미한 영향(환율, S&P500)
외국인 투자자의 국내 주식 매매(외국인순매수)
주가와 거래량 상관관계는 미비

“Feature 데이터 상관분석”



〈현대제철 Heat map〉

상관분석 결과,

주가에 영향을 미치는 Feature는

주식 종목에 따라 상이하게 나타남

“Feature 데이터 상관분석”

Negative	Neutral	Positive
현대제철(철강·금속)	삼성전자(반도체)	삼성바이오로직스(헬스케어)
1. 경기소비재(0.88)	1. SNP500(0.89)	1. SNP500(0.85)
2. 엔화 환율 (0.82)	2. 구리선물(0.87)	2. 구리선물(0.78)
3. 돈육선물(0.73)	3. 경기소비재(0.61)	3. 유로환율(0.67)

“ 결측치 처리 ”

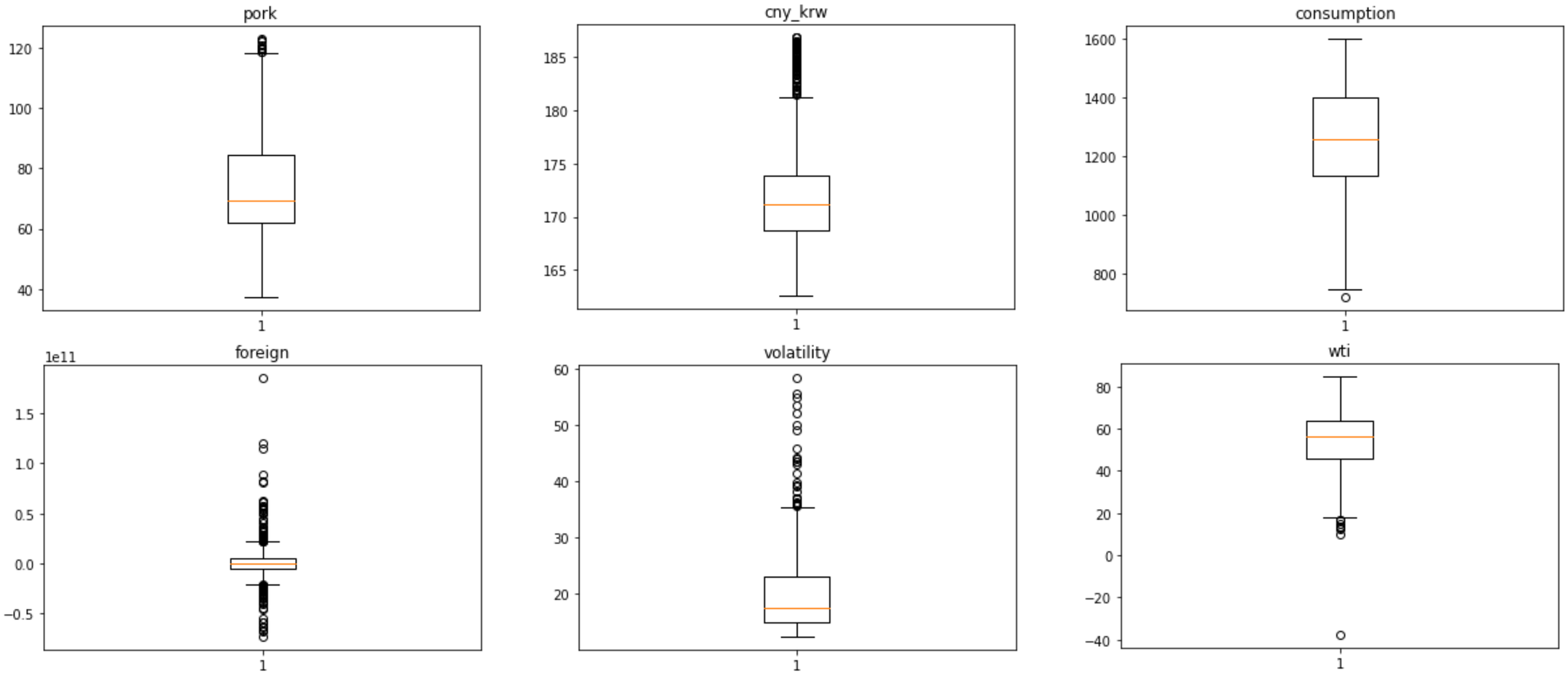
```
data["covid"].fillna(0, inplace = True)
data = data.interpolate()

data.isnull().sum()
```

```
Date          0
Close         0
foreign       0
kospi         0
volatility    0
consumption   0
yield_bond    0
snp500        0
fng           0
gold          0
copper        0
wti           0
pork          0
usd_krw       0
cny_krw       0
eur_krw       0
jpy_krw       0
covid         0
weekday       0
dtype: int64
```

- KOSPI200지수, 해외경제지표에서 일부 날짜(공휴일 등)누락으로 인해 데이터의 결측값이 존재

- S&P500, 경기민감소비재, 돈육선물지수, Wti(서부텍사스중질유)의 경우 영업일 기준 전날이나 다음날에 해당하는 지수값들의 평균으로, 코로나 확진자수의 경우 0으로 결측치 처리함

“ 이상치 처리 ”



►범위 : 백분위수의 10~90에 해당하지 않는 값

► 대상 : 돈육선물, 위안화환율, 경기소비지수, 외국인순매수, 변동성지수, WTI원유선물

“ Modeling ”

사용 모델

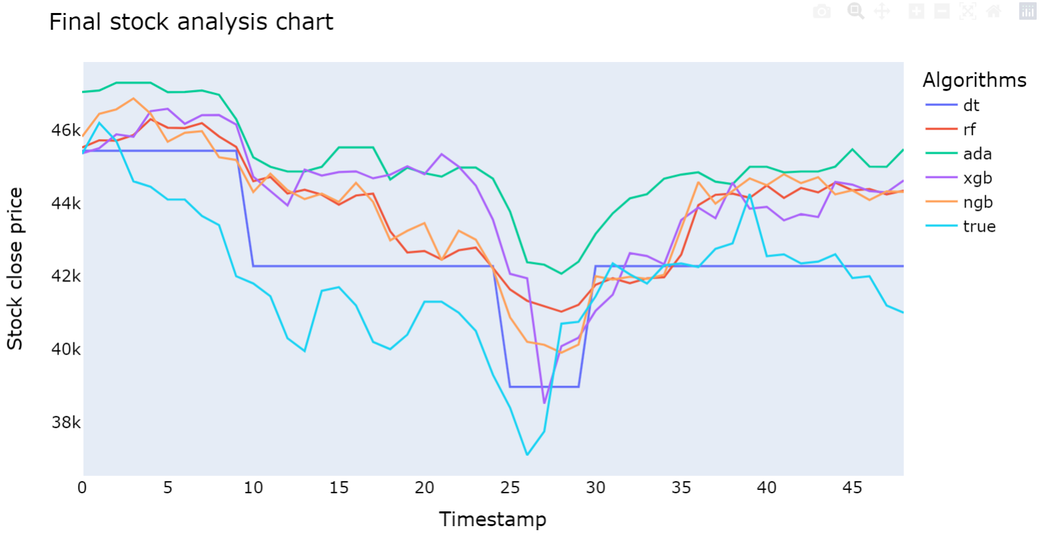
- Decision tree
- Random forest
- Boosting
  - AdaBoosting
  - XGBoost
  - NGBoost

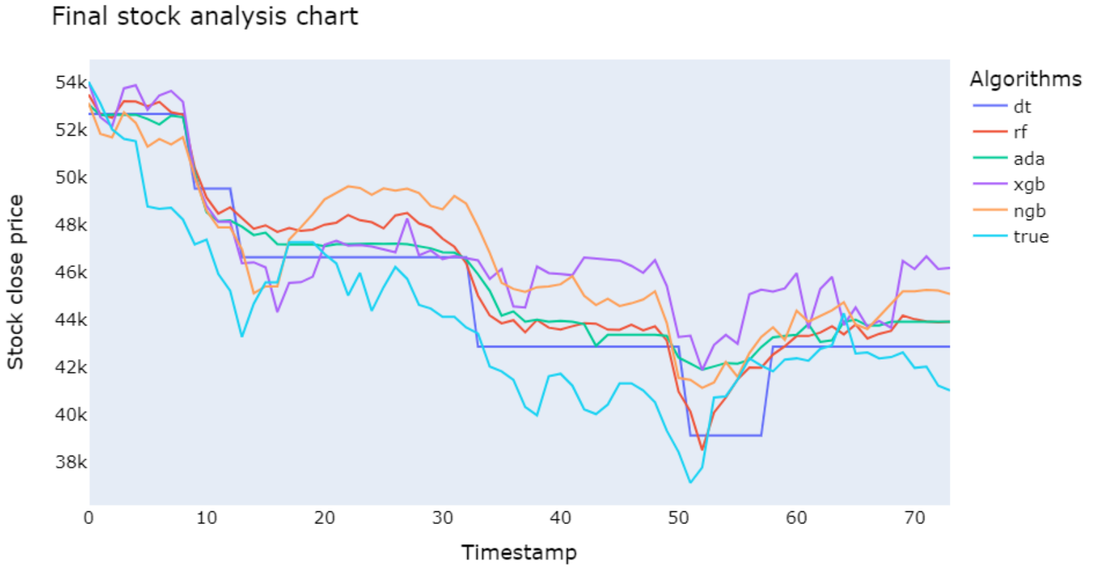
사용 모델 성능평가 및 결과확인

- MSE(평균제곱오차)* 값으로 성능 비교
- 그래프를 그려서 예측 성능 시각화 〉 실제 종가와 비교
- 종목별로 성능이 좋은 모델이 다르게 나타남을 확인
- 성능이 가장 좋다고 판단되는 모델의 핵심변수 확인

*MSE(평균 제곱 오차)
예측값과 실제값 차이의 면적의 평균 값으로 수치가 0에 가까울수록 정확도가 높음

“현대제철”



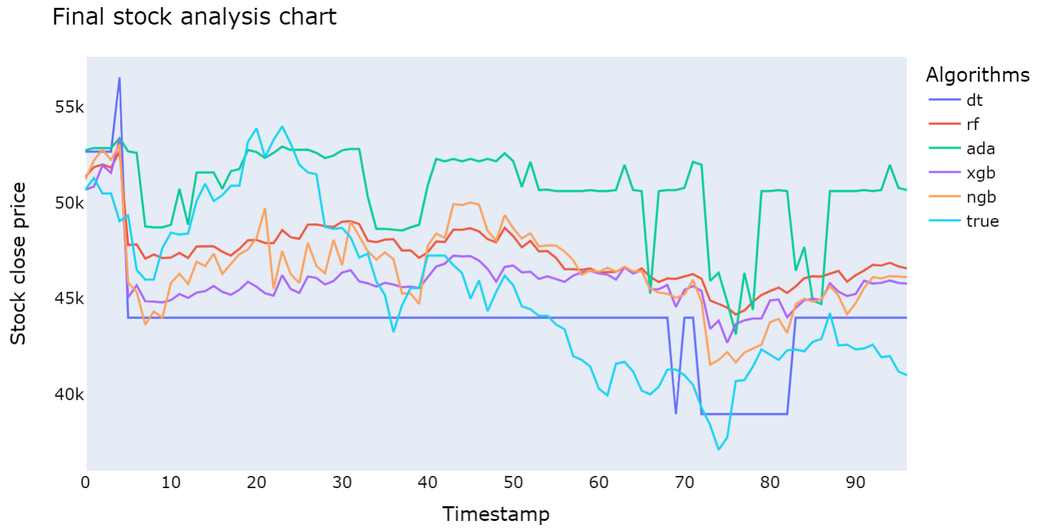


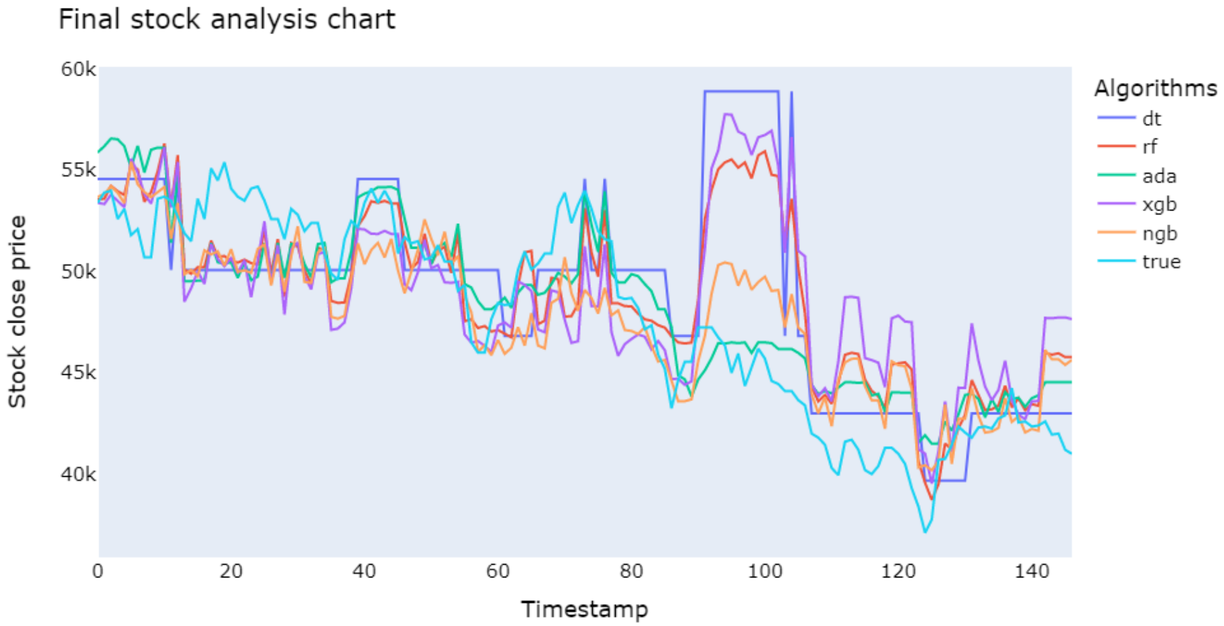
Train : Test 9 : 1

	Decision Tree	Random Forest	AdaBoosting	XGBoosting	NGBoosting
MSE_train	0.003404	0.000324	0.002936	0.000001	0.000604
MSE_test	0.003132	0.009554	0.020522	0.013610	0.009207

	Decision Tree	Random Forest	AdaBoosting	XGBoosting	NGBoosting
MSE_train	0.006411	0.00039	0.004854	.960287e-07	0.001195
MSE_test	0.012139	0.01751	0.016652	.534825e-02	0.029178

“현대제철”



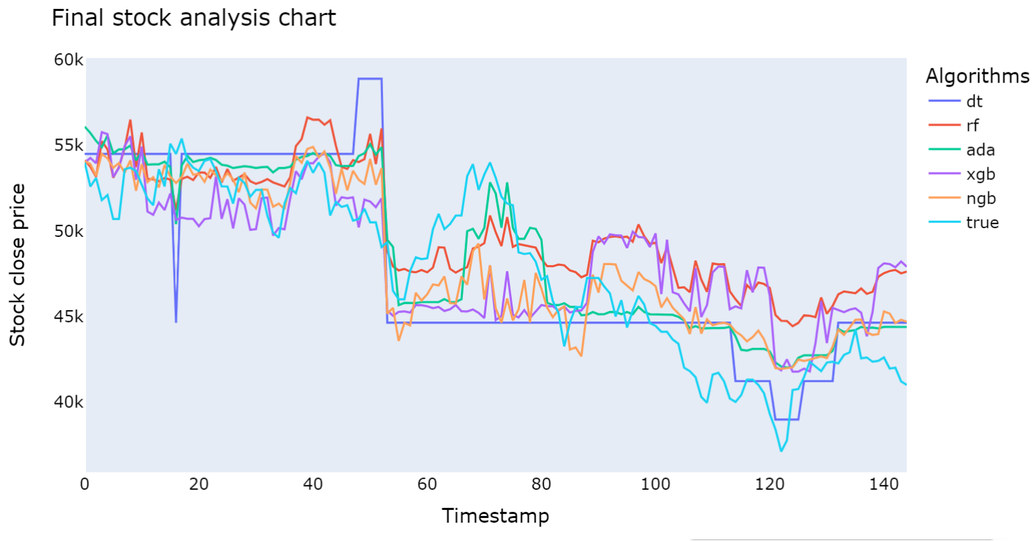


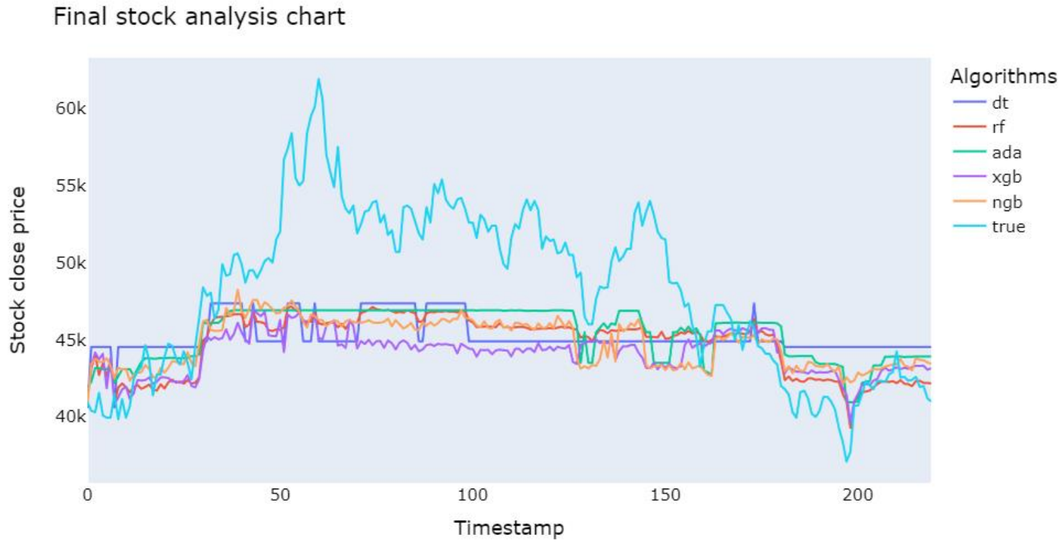
Train : Test 8 : 2

	Decision Tree	Random Forest	AdaBoosting	XGBoosting	NGBoosting
MSE_train	0.003487	0.000361	0.003160	0.000001	0.000519
MSE_test	0.034671	0.031066	0.085728	0.033060	0.029447

	Decision Tree	Random Forest	AdaBoosting	XGBoosting	NGBoosting
MSE_train	0.006982	0.000496	0.005812	0.000001	0.001229
MSE_test	0.075895	0.051061	0.021038	0.076133	0.028963

“현대제철”



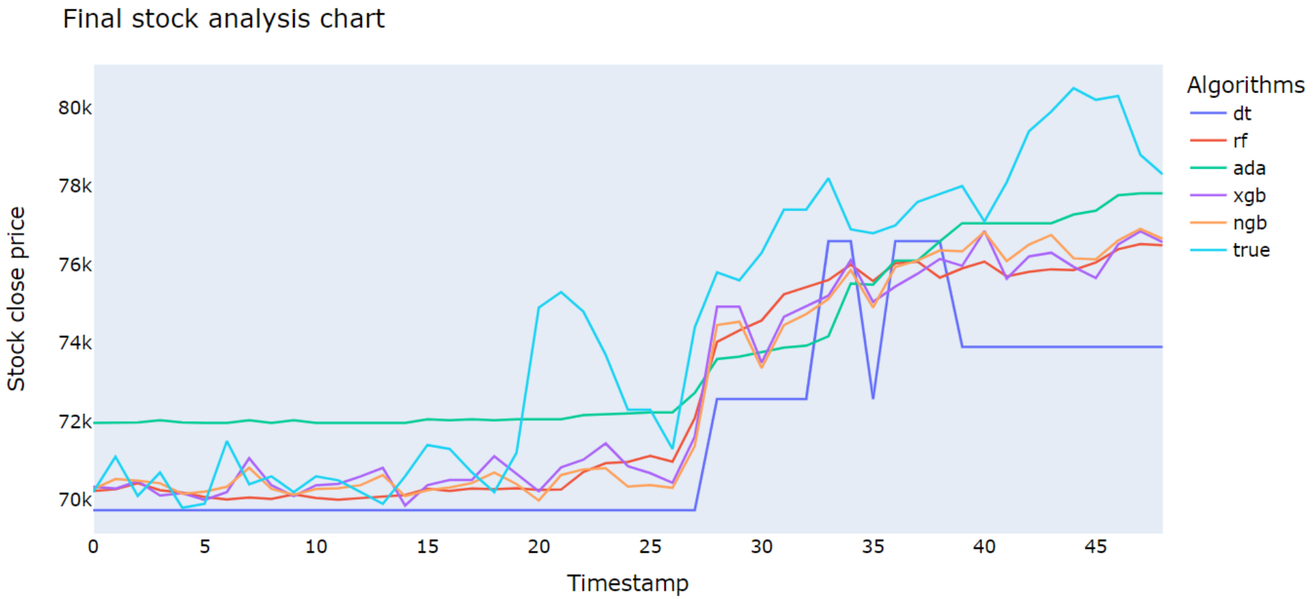


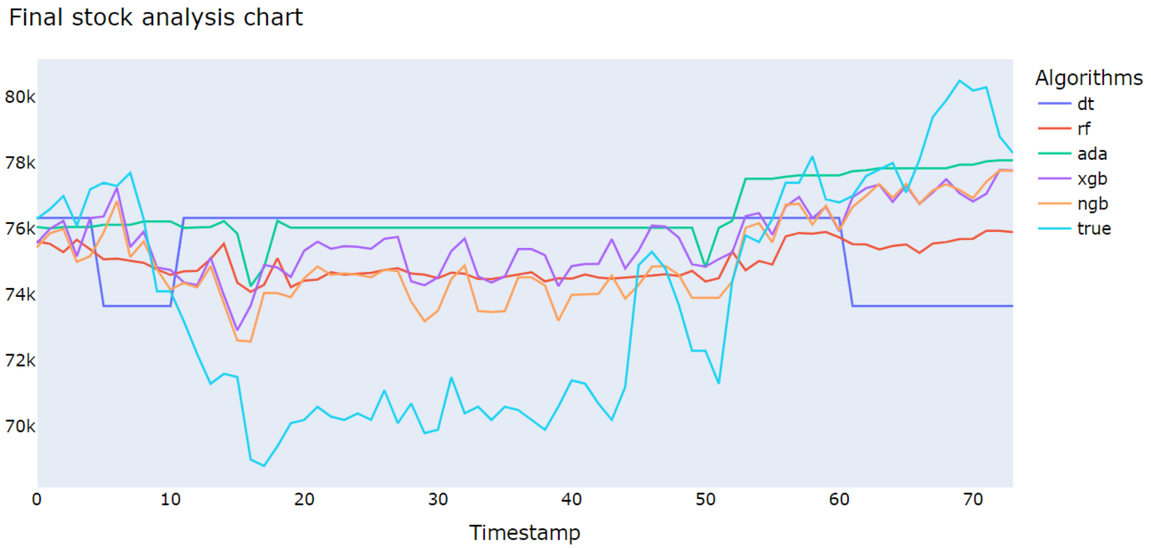
Train : Test 7 : 3

	Decision Tree	Random Forest	AdaBoosting	XGBoosting	NGBoosting
MSE_train	0.004856	0.000560	0.004622	0.000001	0.000662
MSE_test	0.044010	0.043903	0.019334	0.047237	0.024543

	Decision Tree	Random Forest	AdaBoosting	XGBoosting	NGBoosting
MSE_train	0.006054	0.000382	0.004947	8.109066e-07	0.001001
MSE_test	0.133479	0.109861	0.099635	1.531369e-01	0.120767

“삼성전자”



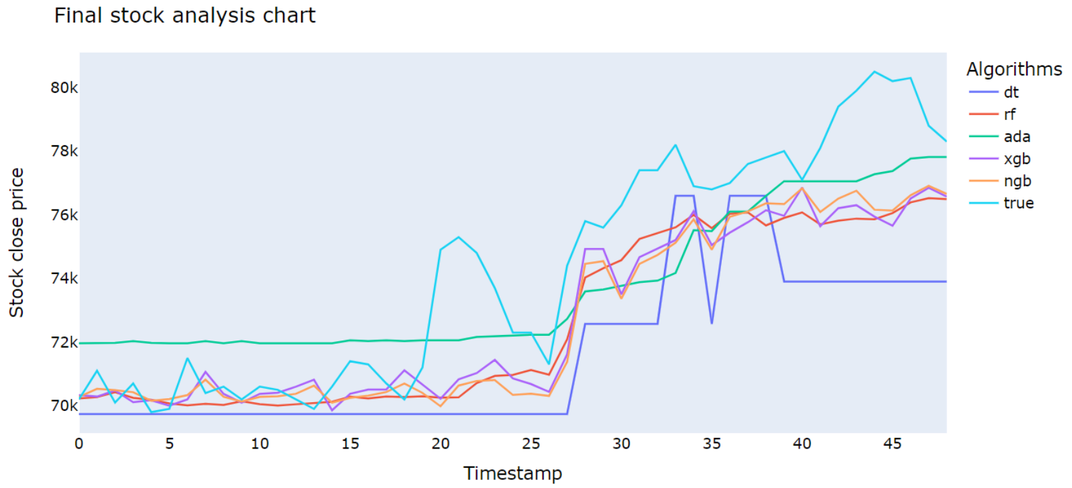


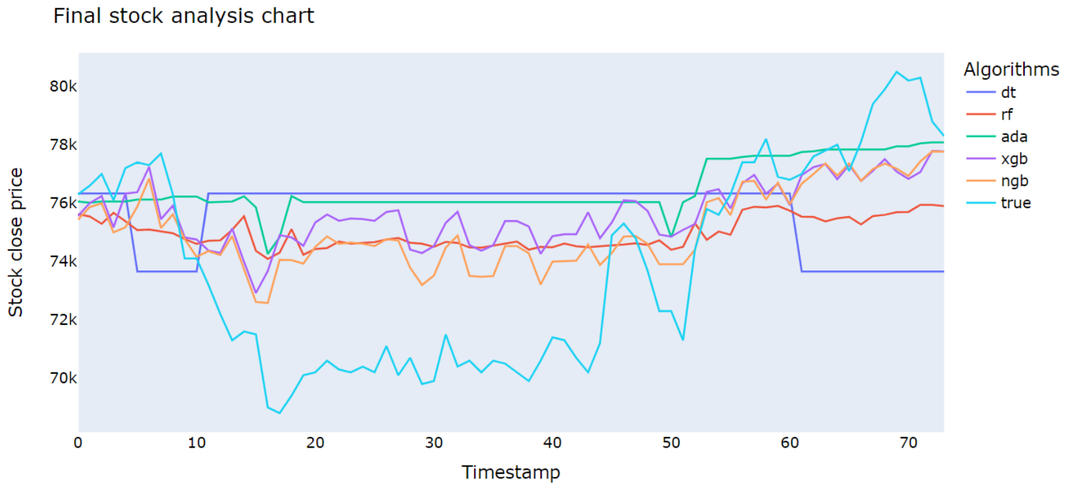
Train : Test 9 : 1

	Decision Tree	Random Forest	AdaBoosting	XGBoosting	NGBoosting
MSE_train	0.003199	0.000318	0.002745	0.000499	0.000275
MSE_test	0.017798	0.007424	0.006084	0.007370	0.006674

	Decision Tree	Random Forest	AdaBoosting	XGBoosting	NGBoosting
MSE_train	0.002562	0.000180	0.002385	0.000494	0.000321
MSE_test	0.026657	0.012721	0.018055	0.012669	0.009128

“삼성전자”



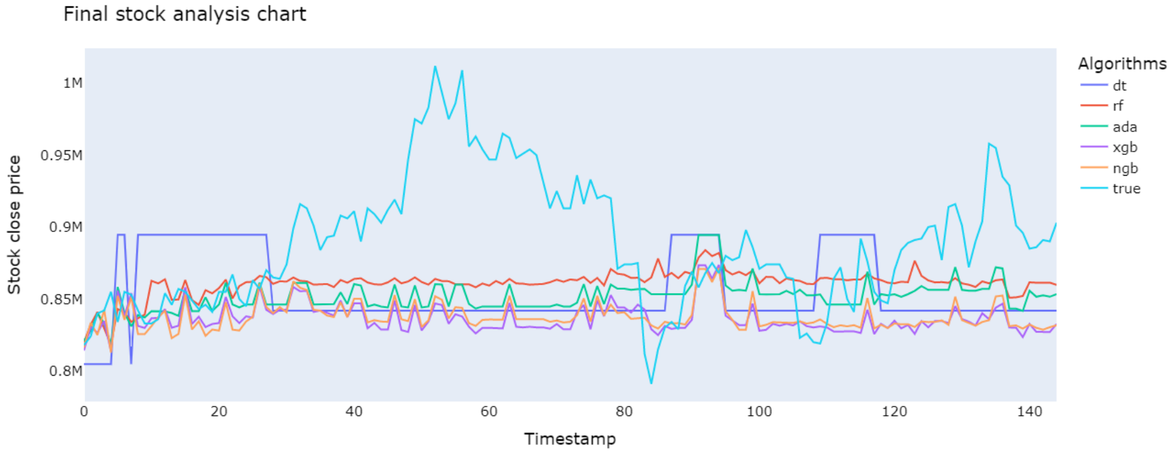


Train : Test 8 : 2

	Decision Tree	Random Forest	AdaBoosting	XGBoosting	NGBoosting
MSE_train	0.003199	0.000318	0.002745	0.000499	0.000275
MSE_test	0.017798	0.007424	0.006084	0.007370	0.006674

	Decision Tree	Random Forest	AdaBoosting	XGBoosting	NGBoosting
MSE_train	0.002562	0.000180	0.002385	0.000494	0.000321
MSE_test	0.026657	0.012721	0.018055	0.012669	0.009128

“삼성전자”

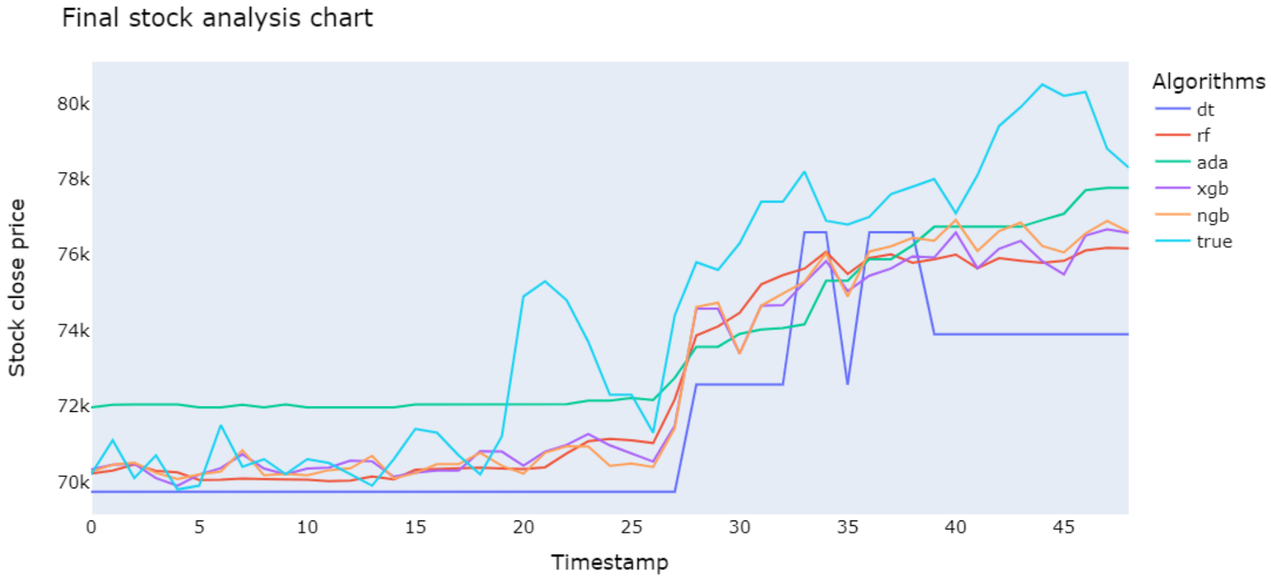


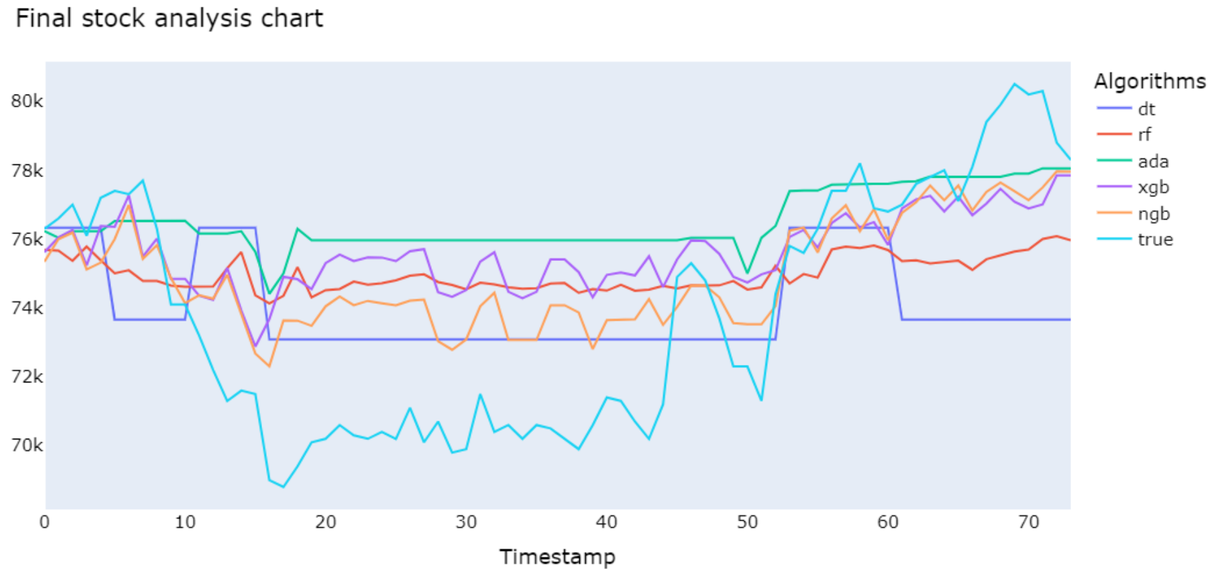
Train : Test 7 : 3

	Decision Tree	Random Forest	AdaBoosting	XGBoosting	NGBoosting
MSE_train	0.014608	0.001877	0.014283	0.002741	0.001230
MSE_test	0.141450	0.082056	0.102182	0.144033	0.144938

	Decision Tree	Random Forest	AdaBoosting	XGBoosting	NGBoosting
MSE_train	0.002447	0.000246	0.002267	0.000436	0.000270
MSE_test	0.055372	0.033854	0.045011	0.044426	0.033476

“ 삼성바이오로직스”



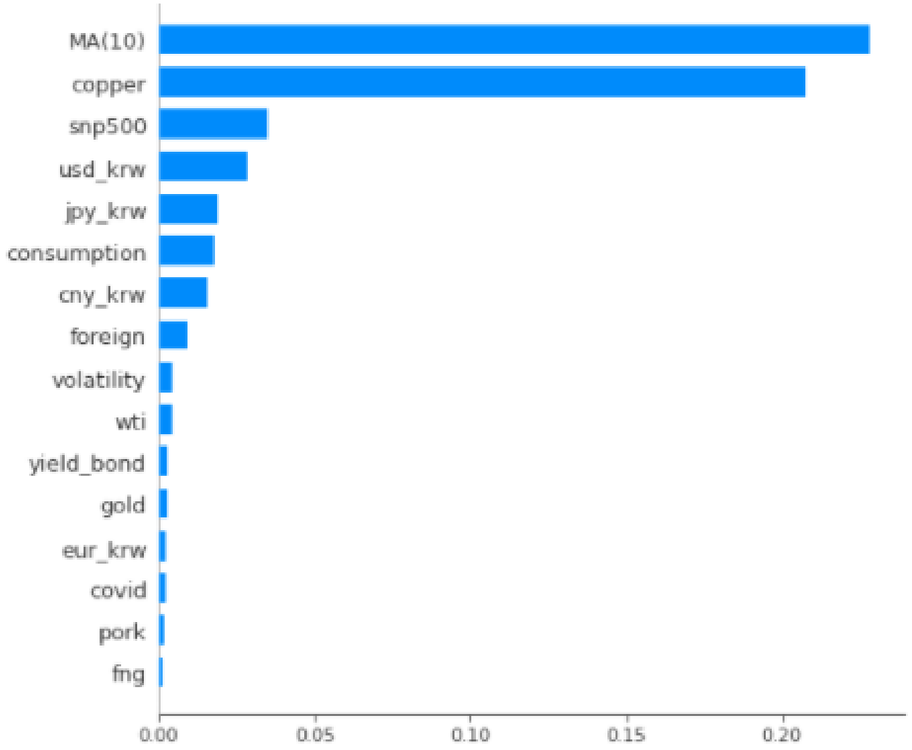


Train : test = 9 : 1

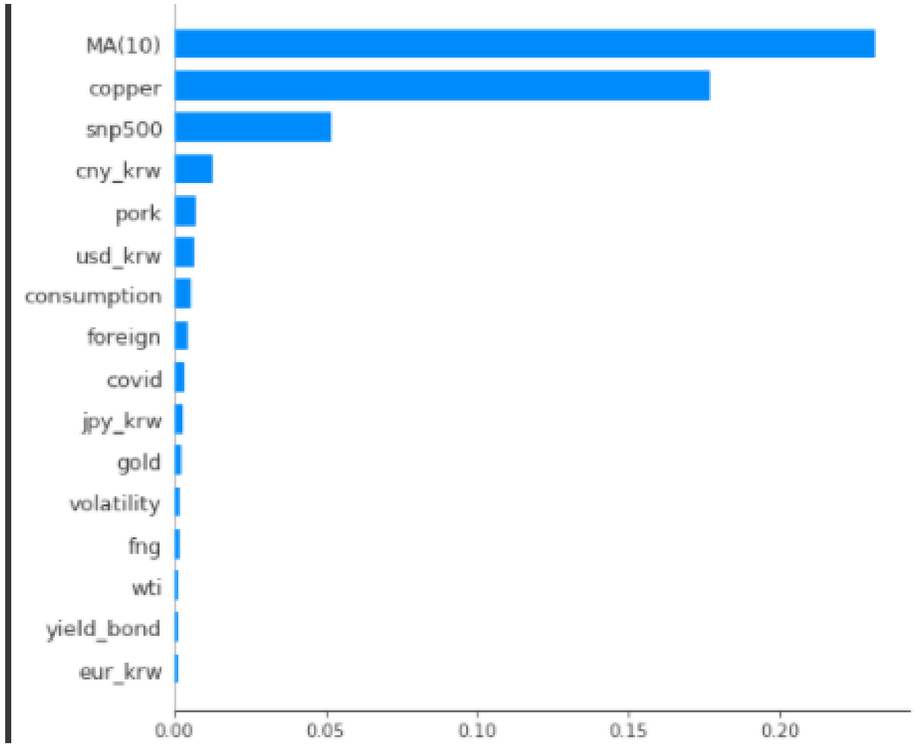
	Decision Tree	Random Forest	AdaBoosting	XGBoosting	NGBoosting
MSE_train	0.003199	0.000316	0.002770	0.000519	0.000267
MSE_test	0.017798	0.007589	0.006491	0.007584	0.006302

	Decision Tree	Random Forest	AdaBoosting	XGBoosting	NGBoosting
MSE_train	0.002566	0.000179	0.002405	0.000508	0.000342
MSE_test	0.021916	0.013166	0.017890	0.012505	0.007643

“ 삼성바이오로직스”



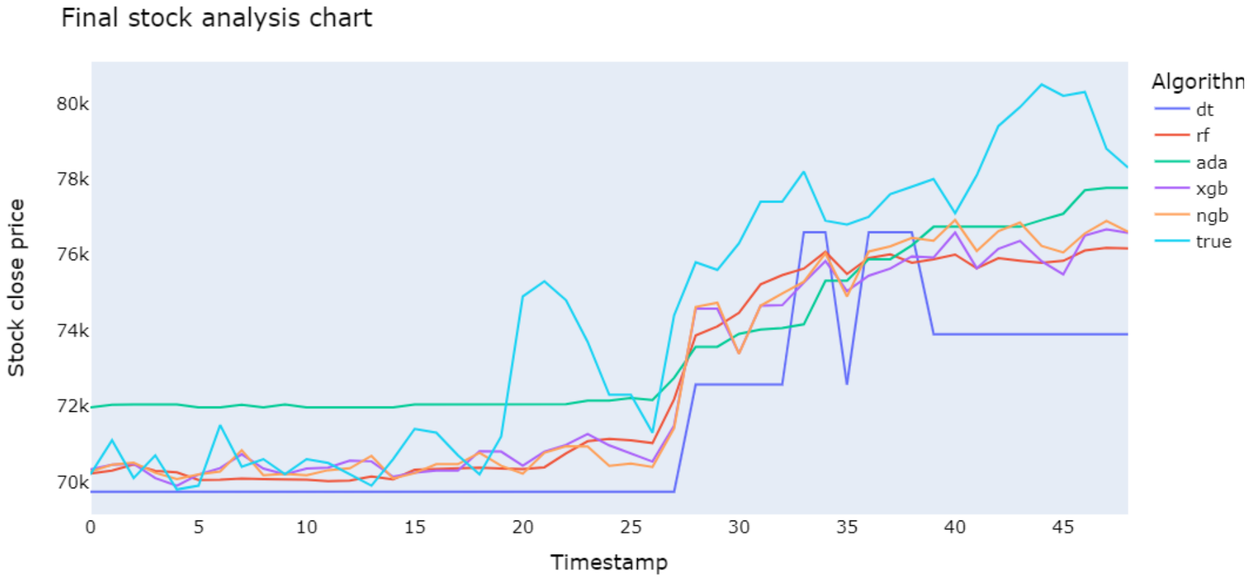
〈팬데믹 이후 데이터의 NGBoosting 중요변수〉

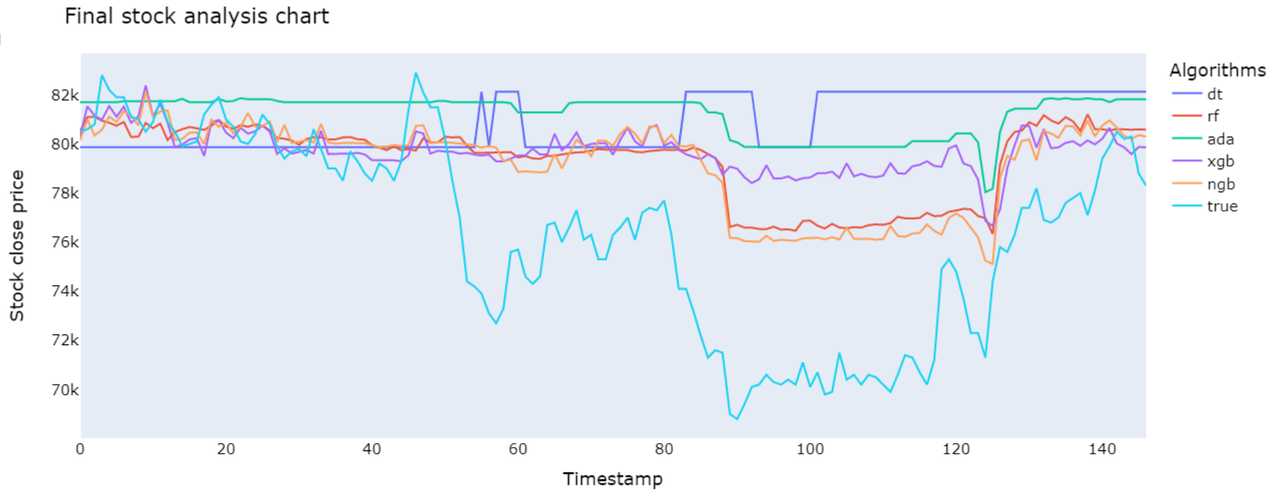


〈3개년치 데이터의 RandomForest 중요변수

Train : Test 9 : 1

“ 삼성바이오로직스”



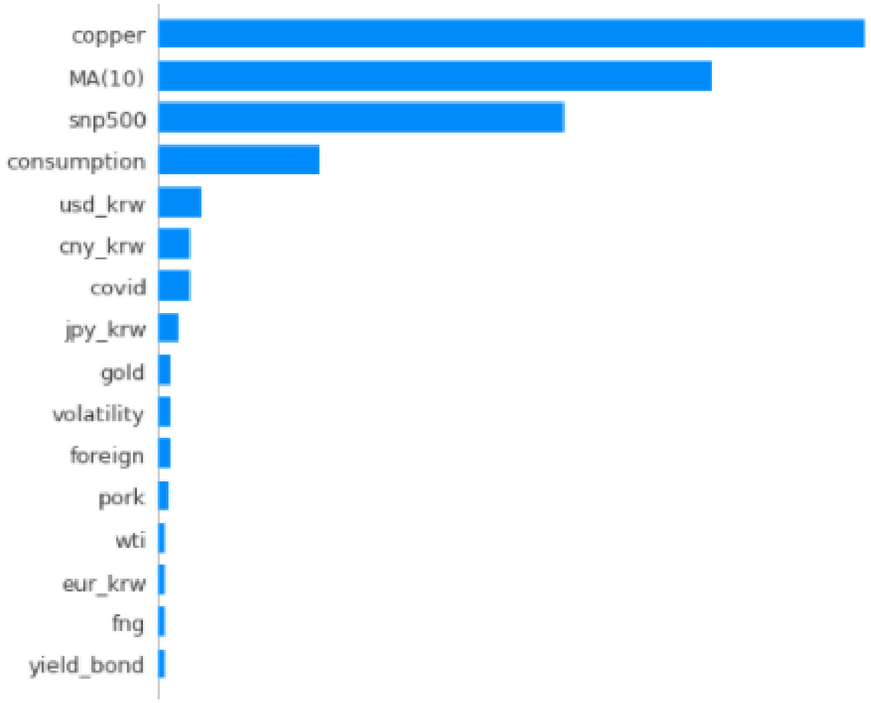


Train : test = 8 : 2

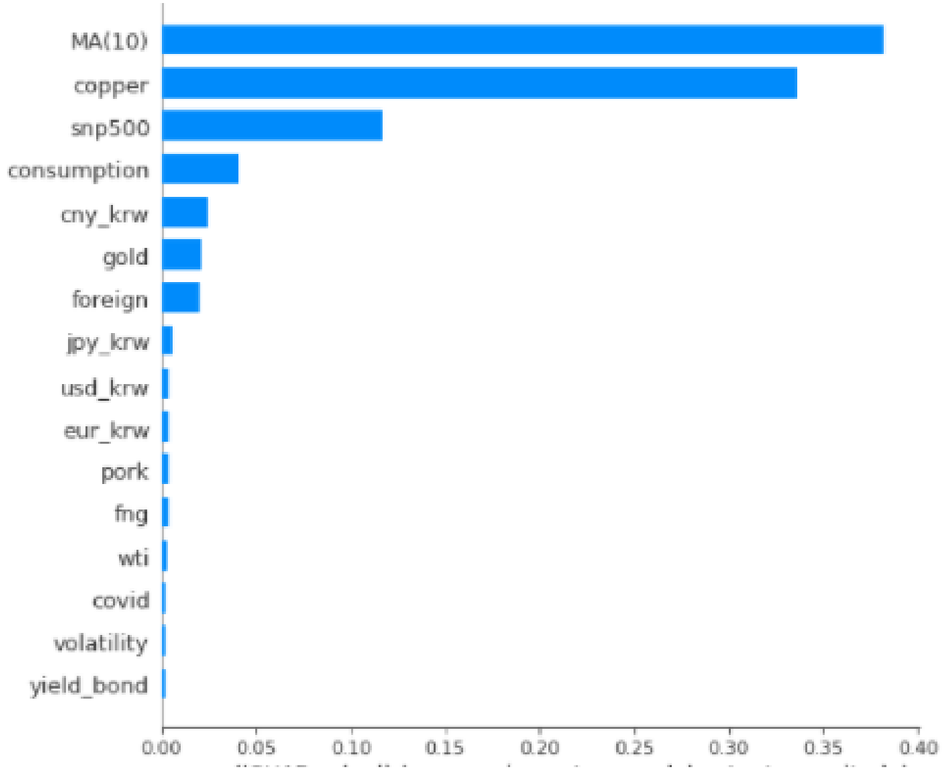
	Decision Tree	Random Forest	AdaBoosting	XGBoosting	NGBoosting
MSE_train	0.002475	0.000296	0.001981	0.000385	0.000187
MSE_test	0.022298	0.037922	0.061024	0.065710	0.052619

	Decision Tree	Random Forest	AdaBoosting	XGBoosting	NGBoosting
MSE_train	0.010929	0.000725	0.010318	0.001919	0.001226
MSE_test	0.206195	0.082183	0.173967	0.114126	0.087922

“ 삼성바이오로직스”



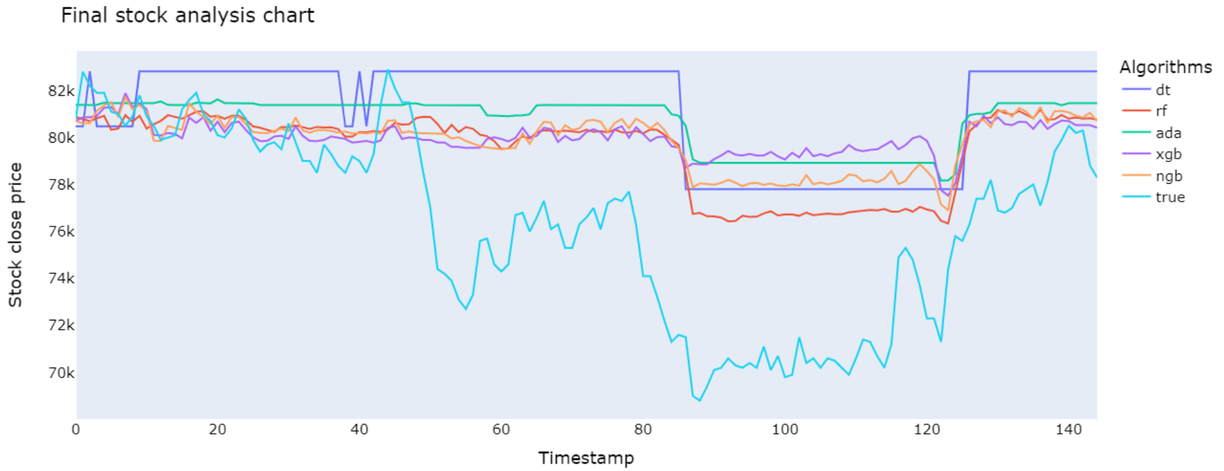
〈팬데믹 이후 데이터의 RandomForest 중요변수〉

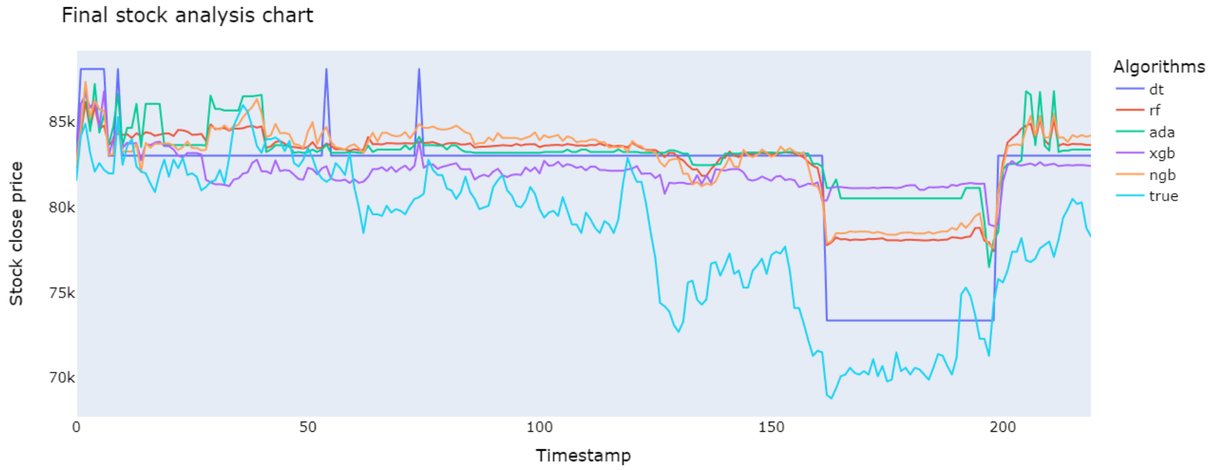


〈3개년치 데이터의 NGBoosting 중요변수〉

Train : test = 8 : 2

“ 삼성바이오로직스”



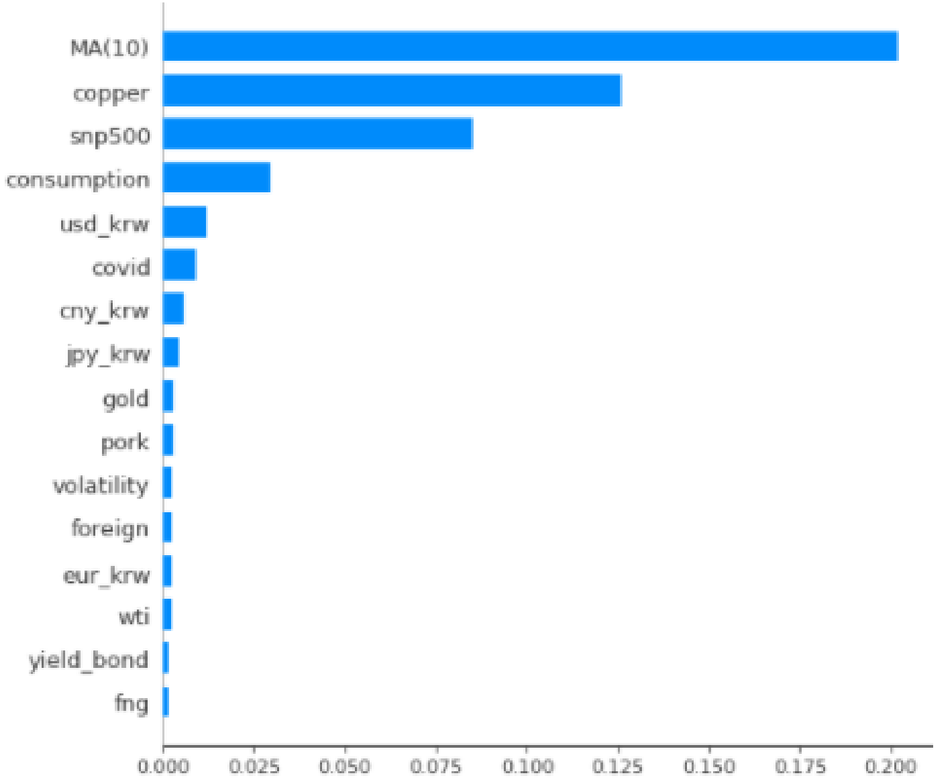


Train : Test 7 : 3

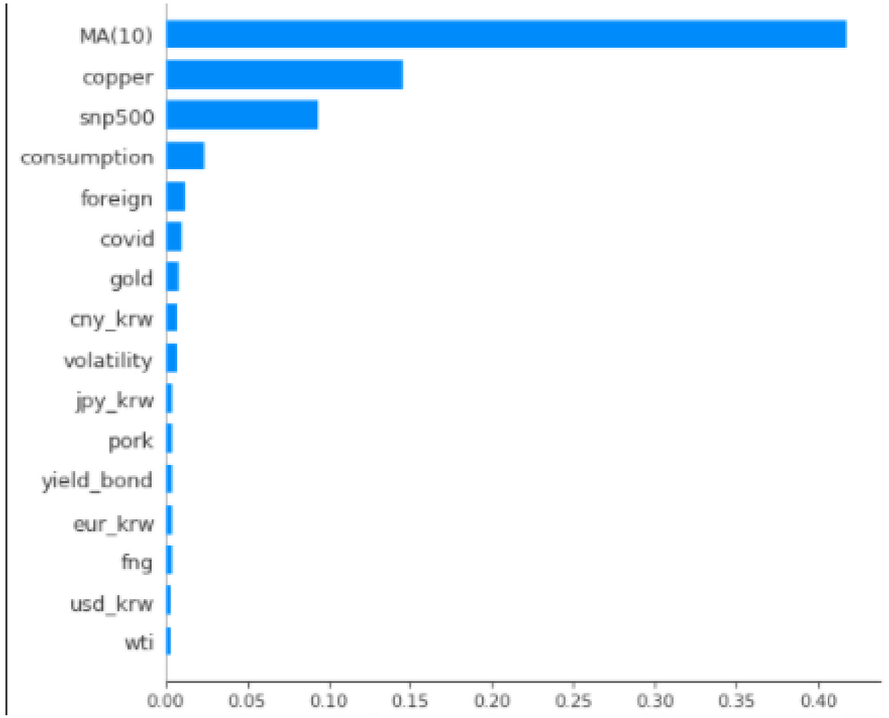
	Decision Tree	Random Forest	AdaBoosting	XGBoosting	NGBoosting
MSE_train	0.002614	0.000288	0.001884	0.000333	0.000168
MSE_test	0.044039	0.024248	0.040074	0.035537	0.035352

	Decision Tree	Random Forest	AdaBoosting	XGBoosting	NGBoosting
MSE_train	0.012072	0.001008	0.013052	0.002307	0.001300
MSE_test	0.284436	0.183899	0.231526	0.195982	0.193885

“ 삼성바이오로직스”



〈팬데믹 이후 데이터의 RandomForest 중요변수〉



〈3개년치 데이터의 RandomForest 중요변수〉

Train : Test 7 : 3

“시행착오 및 극복”

1.Model 선정의 어려움 -〉 Model 특성 파악 후 취합

다양한 모델을 시도해보았고 그 가짓수만큼 난관이 많이 존재했음. 결론적으로 딥러닝 모델인 LSTM은 Random Walk 가설과 Snow-balling 효과로 인해, Decision Tree는 추세를 충분히 따라가지 못해서, CatBoosting은 범주형 변수에 유리한 모델이라 적합하지 않았음. **따라서 random forest, adaboosting, XGboosting, NGboosting 모델을 사용함.**

2.Scaling method 선택의 어려움 -〉 Robust Scaler 최종 선택

컬럼의 측정 단위가 다를 경우, 패턴을 파악하는데 있어 잘못된 계산을 수행할 수 있으므로 Scaling을 적용함.
이 때 MinMaxScaler 대신 **Robust Scaler**를 이용해 Outlier의 영향을 최소화하고 상대적으로 값을 더 넓게 분포시킴

3.비교군 설정의 어려움 -〉 다양한 비교군을 수립하고 비교하여 취합

종가 예측 시 팬데믹 이전/이후로 나누는 한편 전체 데이터의 train-test 비율을 조정함으로써 과거 데이터의 비중에 따른 비교군을 설정하였고, 감성분석 데이터 역시 포함/미포함해보며 비교분석함.

04

평가 및 Q&A

“감성분석”

평가

‘Labeling 기준’

✚ 여타 Data와는 다르게 주식 Data는 긍정/부정과 같은 라벨을 잡기 어려운 부분이 있어서 KOSPI 지수 변동이라는 새로운 변수를 만들어 인코딩함

‘SNS 데이터 수집’

✚ 트위터 API를 사용하고자 했지만 데이터 수집 총량 제한으로 인해 부득이하게 API를 사용하지 않고 크롤링하는 방식을 택함

‘감성분석 Library’

✚ 국내 Library는 주식 Data와의 상성이 맞지 않고, 그나마 맞는 것도 개발자와 직접 컨택하여 사용해야하는 한계가 있어 해외 Library을 적극 활용

제언

‘주가 예측 감성분석 자료 부족’

▬ 국내 주가 예측 관련 자료나 경제 감성사전이 흔하지 않고, 정확도를 알 수 없거나 사용이 어려워 분석 방향을 잡는데 시간이 많이 소요되었음 –> **해외 연구 참고**

‘질 좋은 SNS 데이터 확보 어려움’

▬ 광고 및 도배글이 너무 많고 일정한 주기로 데이터가 올라오지 않아 정확한 종가 예측에 도움이 되지 않음 –> **키워드나 계정명을 한정시키는 방법으로 보완 가능**

‘불용어 처리의 어려움’

▬ 경제뉴스 같은 경우 한 단어로도 의미 차이가 많이 나고 종목에 대한 미래를 예측하기 어려운 특성이 있어 불용어 처리가 쉽지 않음 –> **패턴 연구 심화 필요**

“종가예측”

평가

‘Feature값 선정’

✚ 다양한 feature들 중 히트맵을 참고하여 상관도가 높은 것 위주로 선별, **일별데이터를 사용할 수 없는 경우 Feature로 선정하지 않음**

‘다양한 스케일링 적용’

✚ 컬럼의 측정 단위가 다를 경우, 패턴을 파악하는데 있어 잘못된 계산을 수행할 수 있으므로 Scaling을 적용. 이 때 MinMaxScaler 대신 **Robust Scaler**를 이용해 Outlier의 영향을 최소화하고 상대적으로 값을 더 넓게 분포시킴

‘모델 선정’

✚ LSTM은 Random Walk 가설과 Snow-balling 효과로 인해, Decision Tree는 추세를 충분히 따라가지 못해서, CatBoosting은 범주형 변수에 유리한 모델이라 부적합. **random forest, adaboosting, XGboosting, NGboosting** 모델을 사용

제언

‘데이터의 효과적인 활용 방법 모색’

- 주가에 영향을 미치지만 feature로 활용하지 못한 데이터들 존재
Ex) 수출통계, ISM제조업지수 –〉 **월별 데이터의 활용 방법 고안 필요**

‘이동평균선을 이용한 주가 예측’

- 5일, 10일 이동평균선을 활용하여 모델링 –〉 **단기, 중기, 장기 이동평균선을 모두 사용하여 상승과 하락 추세 예측 필요**

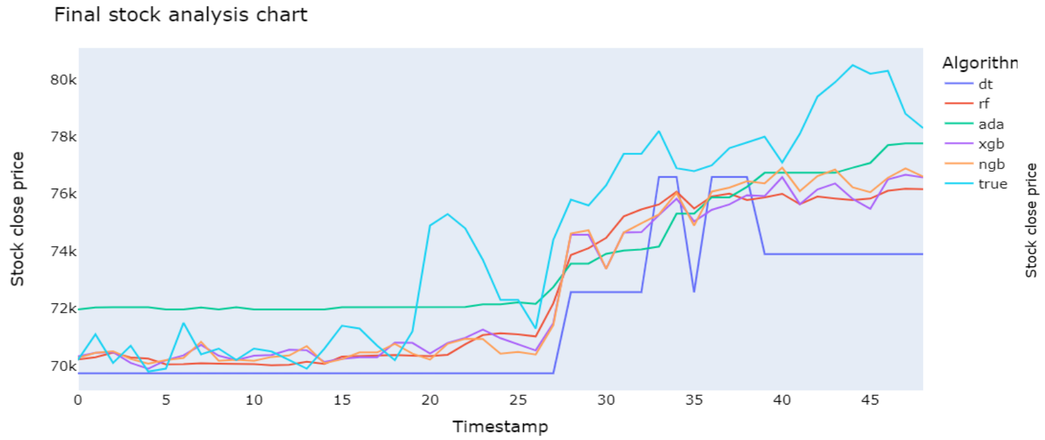
‘산업별로 다른 주가 결정 요인’

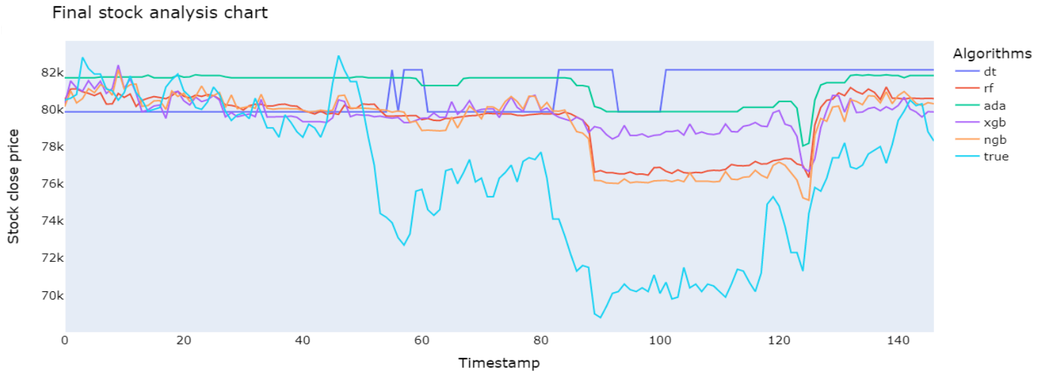
- 모델링 결과 주가에 영향을 미치는 요인은 산업군에 따라 다르게 나타남
–〉 **산업별로 주가에 영향을 미치지만 놓치고 있는 고유 변수 파악 필요**

결론

가설1 코로나의 영향이 큰 종목에 대해서는 코로나 이후의 데이터로 분석하는 것이 바람직하다.

→ 예상했던 대로 대다수의 종목에서 코로나 이후의 데이터만 분석에 활용하는 것이 실제 종가를 더 잘 예측하는 것으로 나타났다.





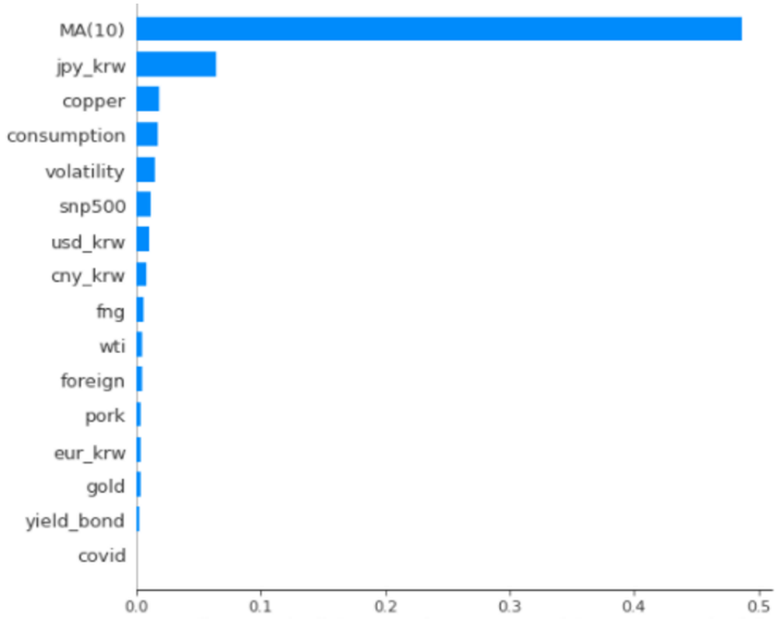
	Decision Tree	Random Forest	AdaBoosting	XGBoosting	NGBoosting
MSE_train	0.002475	0.000296	0.001981	0.000385	0.000187
MSE_test	0.022298	0.037922	0.061024	0.065710	0.052619

	Decision Tree	Random Forest	AdaBoosting	XGBoosting	NGBoosting
MSE_train	0.010929	0.000725	0.010318	0.001919	0.001226
MSE_test	0.206195	0.082183	0.173967	0.114126	0.087922

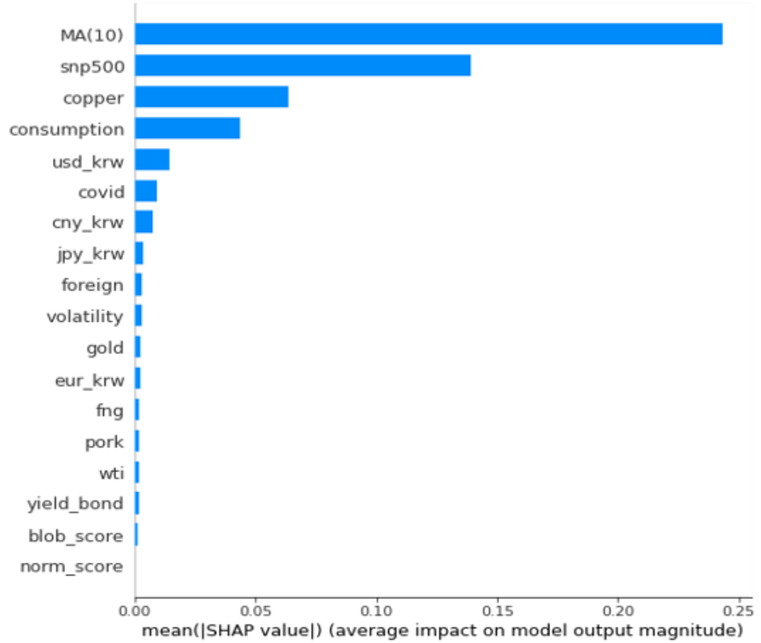
결론

가설2 주가는 과거데이터에 영향을 받는다

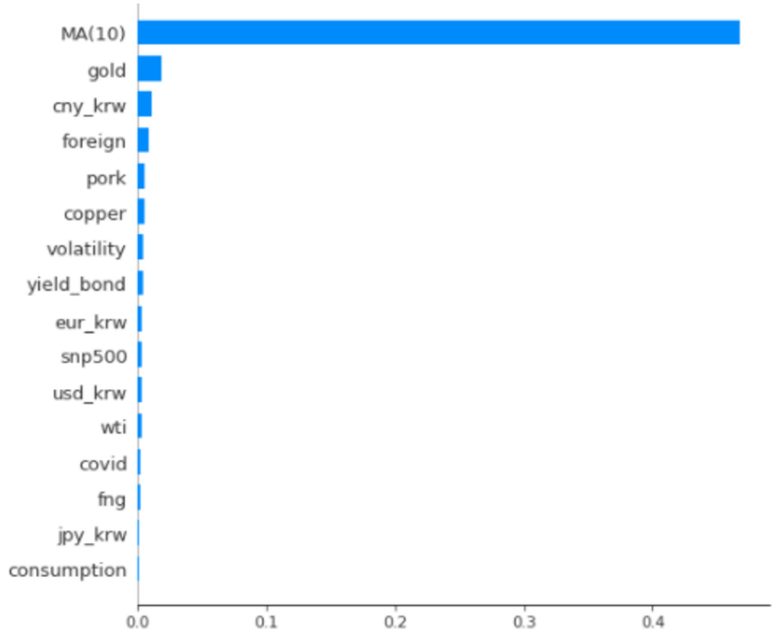
→ 종가 예측 모델 성능 평가 결과 가장 MSE 값이 작았던 모델의 핵심변수는 모두 MA10(10일이동평균선)으로,
모든 변수들 중 과거 주가의 영향이 가장 큰 것으로 나타남



〈현대제철 NGBoosting〉



〈삼성전자 RandomForest 〉



〈삼성바이오로직스 XGBoosting〉

가설3 감성데이터(비정형데이터)가 예측 성능에 유의미한 영향을 미칠 것이다.

→ 감성데이터를 포함/일부 포함/포함하지 않은 다양한 비교군을 설정하여 예측 성능 비교

삼성전자

2년 감성분석 적용O

	Decision Tree	Random Forest	AdaBoosting	XGBoosting	NGBoosting
MSE_train	0.002867	0.000324	0.002275	0.000422	0.000214
MSE_test	0.022304	0.037817	0.060851	0.068216	0.065365

2년 감성분석 적용X

	Decision Tree	Random Forest	AdaBoosting	XGBoosting	NGBoosting
MSE_train	0.002867	0.000317	0.002325	0.000423	0.000222
MSE_test	0.065486	0.038990	0.060252	0.067819	0.056050

결론

감성 데이터

News, Twitter

자연어 처리 패키지

과거 데이터

모델 취합 및 비교

기간별 종가예측

COVID-19

팬데믹 이전/이후 분석

Train/test 비율 변화

전체 데이터 통합 후 feature 적절하게 변환 후 분석에 사용

모델별 MSE, graph 비교

다양한 예측 방법 시도

팬데믹 관련 종가 예측

Q&A

감사합니다!