머신러닝을 이용한 코스피 지수 방향성 예측

Forecasting the Direction of the KOSPI Index Using Machine Learning

Young & Rich팀, 3조
김도완
김현규
신승엽
조민기
조현욱

목차

목표는 ?

- 다양한 데이터와 머신러닝을 이용하여 코스피지수의 방향성을 예측하고자 한다.

- 투자결정에 도움을 주는 정보제공서비스에 활용

시나리오

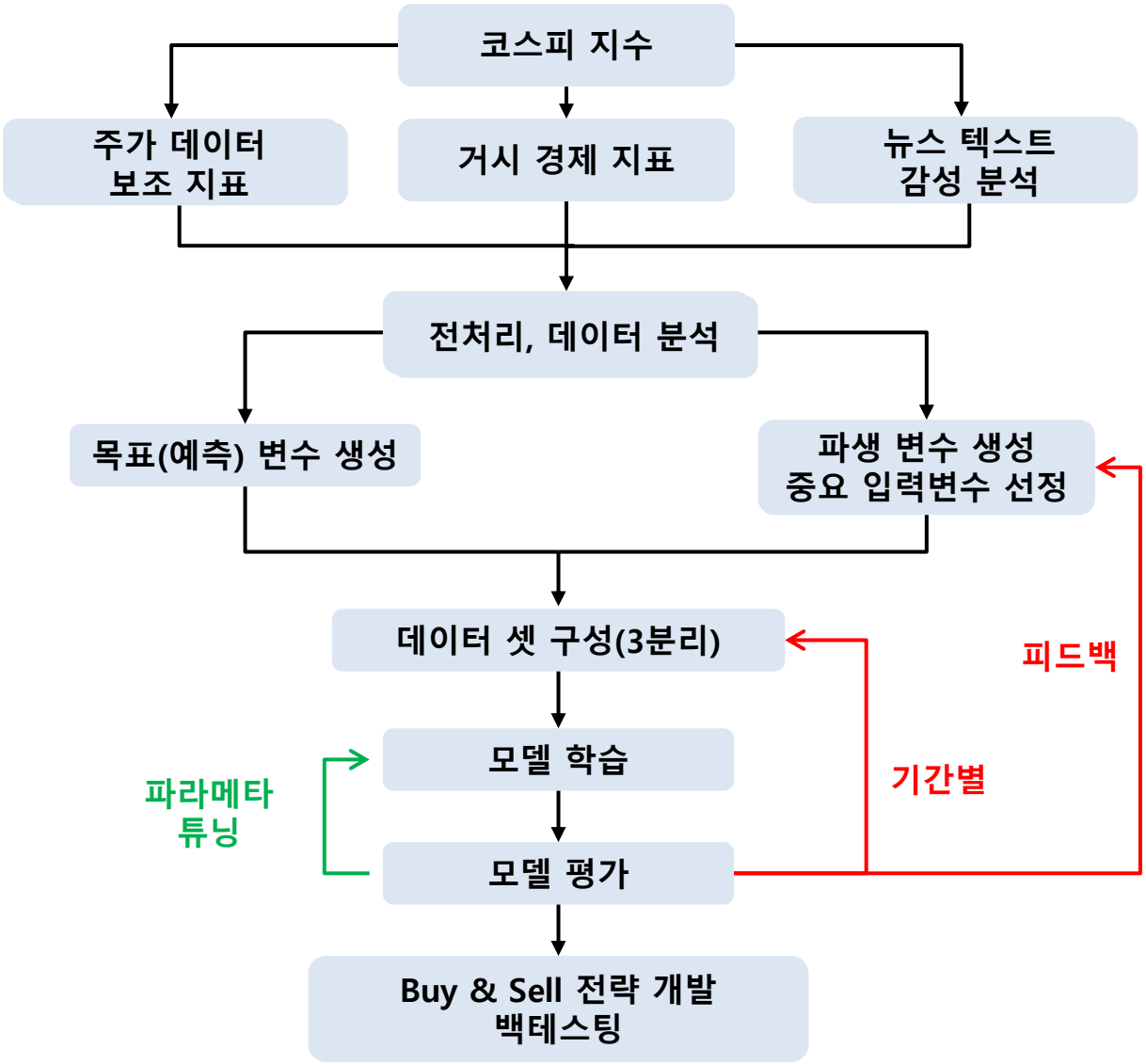
1. **과거데이터**
   - 기존 과거데이터 10년치 DB에 저장
   - 일별 자동수집

2. **수집된 데이터로 모델을 학습 및 검증**
   - 모델은 분기별 재학습하고 평가
   - 다양한 목표변수와 입력변수로 다수의 최적모델을 생성하고 앙상블

3. **학습된 모델로 미래의 시세를 예측한다.**
   - 기간 x일에 대한 상승/하락 방향성, 변화율(%), 확률 등 예측

4. **서비스**
   - 웹 시각화, 메신저 알림
   - 트레이딩 참고지표로 사용 가능

모델 개발과정



어떤 데이터가 필요한가?

자료의 무결성 검증
결측치, 이상치 확인
탐색적 분석(EDA)

Feature Engineering

Train / Valid / Test

머신러닝 모델

Accuracy, precision, ROC
MSE, MAE, MAPE

실제주가 변동량 대비
수익률 비교

데이터 수집 및 탐색

- 수집기간은 약 10년치, 2010년~현재

분류	항목	수집방법
코스피 지수 관련	기본지표(OHLCV)	FDR 라이브러리
	보조지표	TaLib 라이브러리
	신용잔고	금융투자협회 엑셀수집
	해외지수	FDR 라이브러리
거시경제지표	경기종합지수	통계청 엑셀수집
	물가지수	통계청 엑셀수집
	GDP	통계청 엑셀수집
	환율(원달러,원위안)	FDR 라이브러리
	금가격	FDR 라이브러리
뉴스데이터	네이버증권뉴스	네이버 크롤링

데이터 수집 및 탐색

- 코스피 지수 데이터(5개), 보조지표(40개)
- OHLCV -> 파생 -> 보조지표

Panel A. Daily Trading Data	
Open/Close Price	nominal daily open/close price
High/Low Price	nominal daily highest/lowest price
Trading volume	Daily trading volume
Panel B. Technical Indicator	
MACD	Moving average convergence divergence: displays trend following characteristics and momentum characteristics.
CCI	Commodity channel index: helps to find the start and the end of a trend.
ATR	Average true range: measures the volatility of price.
BOLL	Bollinger Band: provides a relative definition of high and low, which aids in rigorous pattern recognition
EMA20	20 day Exponential Moving Average
MA5/MA10	5/10 day Moving Average

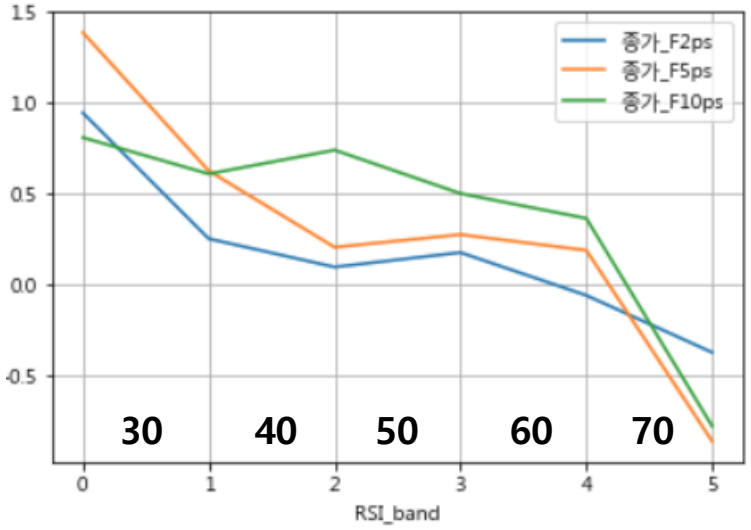
데이터 수집 및 탐색

- 코스피 보조지표, 왜 의미가 있는지?
- 이동평균선(돌파) : 과거 특정기간의 주가를 평균, 골든/데드 크로스
- RSI : 상대강도지수 (0~100), 과열여부

데이터 수집 및 탐색

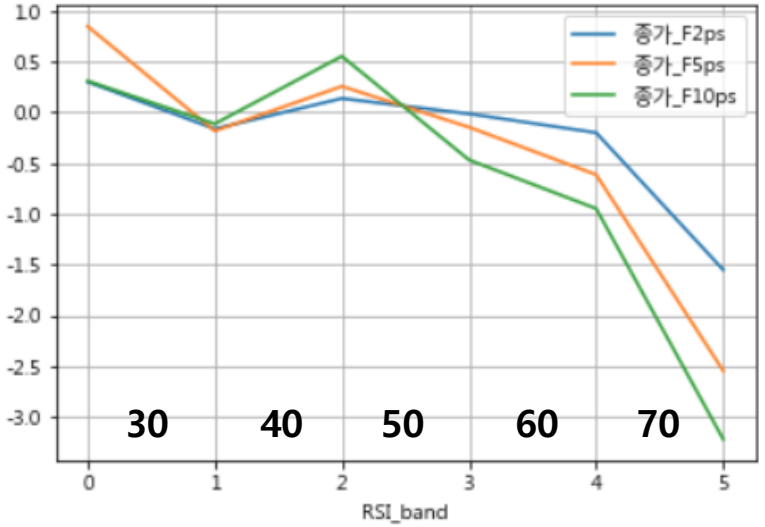
$$RSI(\text{상대강도지수}) = \frac{n\text{일간 주가상승폭의 합계}}{n\text{일간 주가상승폭의 합계} + n\text{일간 주가하락폭의 합계}} \times 100$$

삼성전자



RSI (0~100)

현대차



RSI (0~100)

데이터 수집 및 탐색

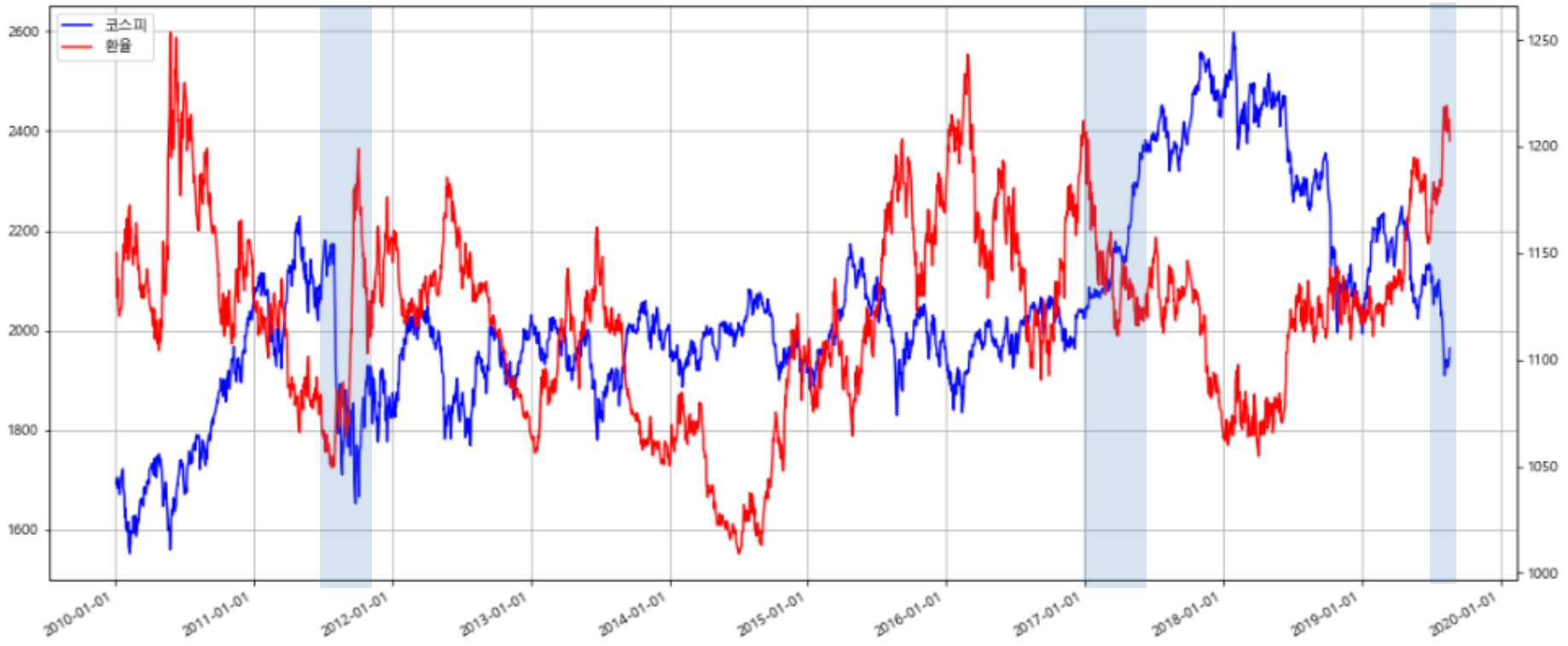
- 코스피 신용잔고
  - 투자자가 돈을 빌려 주식을 매수한 금액, 증시의 과열 확인
  - 신용잔고 오실레이터= log2(신용잔고 / (직전240일 신용잔고의 평균치)

데이터 수집 및 탐색

- 미국 나스닥지수
  - 부분적으로 양의 상관관계
  - 선행/후행, 서로 영향을 주고받는 관계

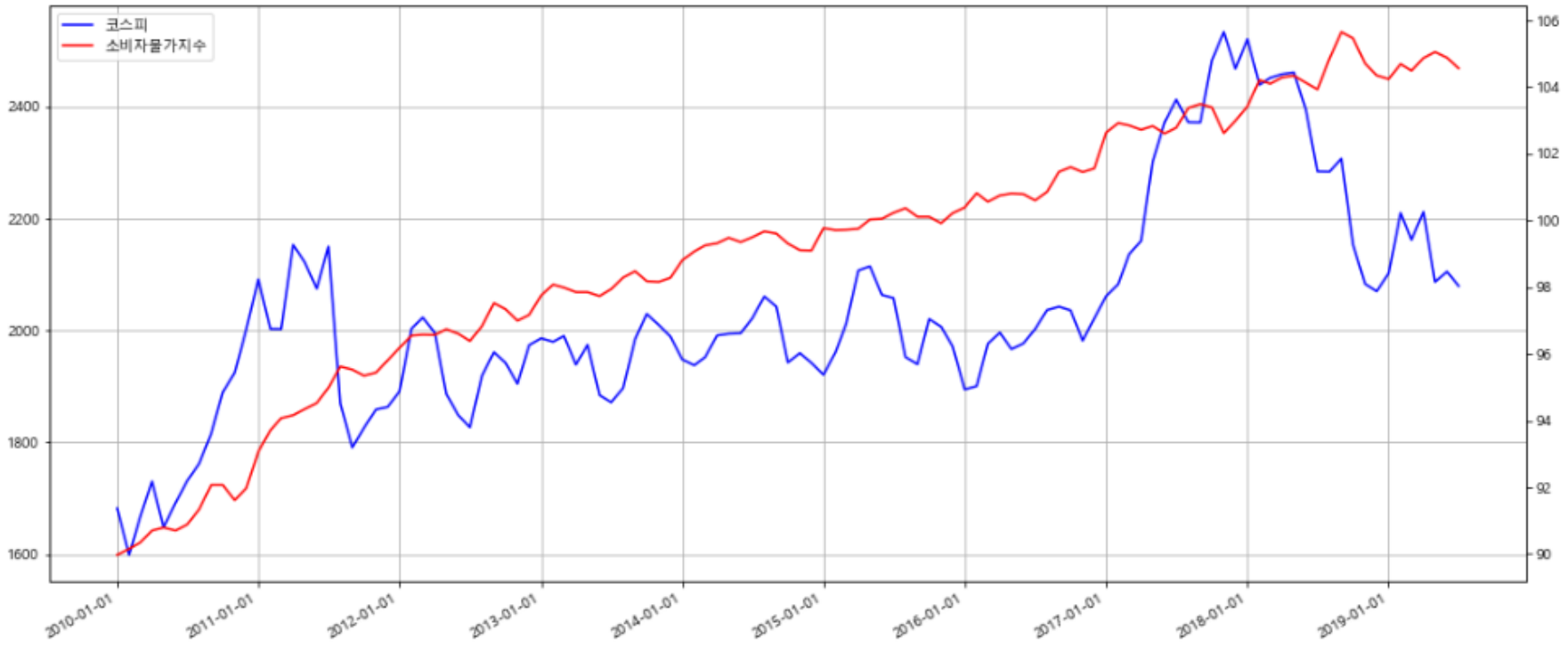
데이터 수집 및 탐색

- 원달러 환율
  - 부분적으로 음의 상관관계
  - 선행/후행, 서로 영향을 주고받는 관계



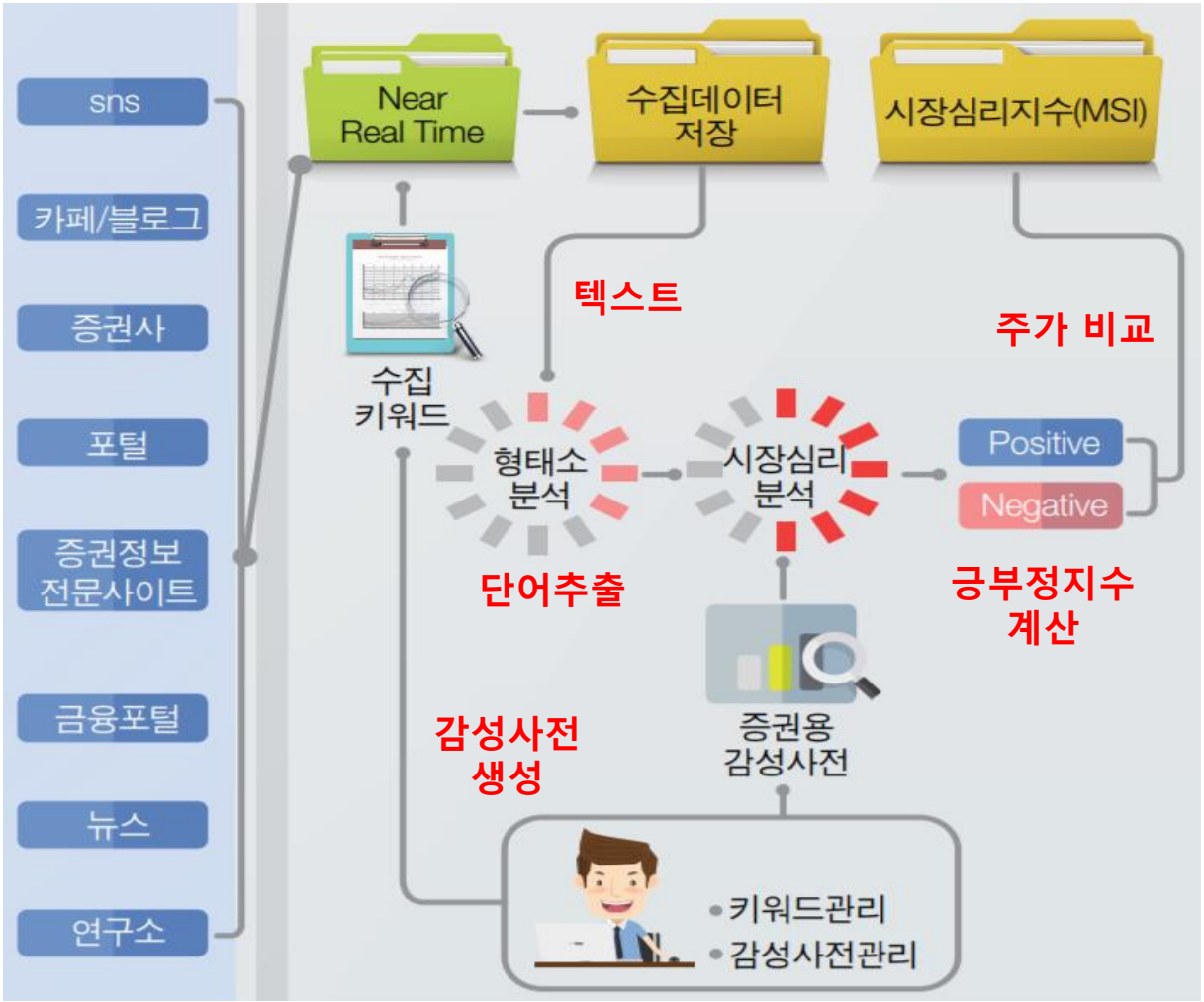
데이터 수집 및 탐색

- 거시경제지표
  - 물가지수 vs 코스피(월평균)
  - 그 외 실질성장률, 총소득 등 비교
  - 장기관점에서 우상향 연관성 있음, 단기 영향미미



데이터 수집 및 탐색

- 뉴스데이터의 분석
  - 증권용 감성사전의 부재



데이터 수집 및 탐색

- 뉴스데이터의 분석
  - 네이버 금융 주요뉴스 일별수집, 형태소 분리(명사)

	Date	name	nouns	Close	Open	High	Low	Volume	Change
0	2010-01-04	[장외주식 시황] 금호생명 급반등"올 증시 上低下高… 1,900까지 오를것"한 - 인도...	[장외, 주식, 황, 금호, 생명, 급, 반등, 증시, 인도, 발효, 증시, 대주,...	1696.14	1681.71	1696.14	1681.71	296550000.0	0.0079
1	2010-01-05	MSCI 선진지수 편입땐 단기 4~10조 매수여력원화 초강세…원·달러 환율 이틀새 ...	[선진, 지수, 편입, 땐, 단기, 매수, 여력, 원화, 초, 강세, 원, 달러, ...	1690.62	1701.62	1702.39	1686.45	408850000.0	-0.0033
2	2010-01-06	원전 테마株, 美 수출 기대로 '들썩'유통주, 외국인·기관·개인 "3色 공략"삼성전...	[원전, 테마, 수출, 기대, 유통, 주, 외국인, 기관, 개인, 공략, 삼성, 전...	1705.32	1697.88	1706.89	1696.10	426040000.0	0.0087
3	2010-01-07	4분기 어닝시즌 내주 개막…바닥 찍고 점프 뛴 '실적 국가대표' 누구 ? 기관이 던진 ...	[어닝, 시즌, 내주, 개막, 바닥, 점프, 실적, 국가대표, 누구, 기관, 주식,...	1683.45	1702.92	1707.90	1683.45	462400000.0	-0.0128
4	2010-01-08	외국인, 아시아 주식 '저인망식 매수' 나서"증권株 당분간 강세" 기대여행항공株, ...	[외국인, 아시아, 주식, 인, 망식, 매수, 증권, 당분간, 강세, 여행, 항공,...	1695.26	1694.06	1695.26	1668.84	379950000.0	0.0070
5	2010-01-11	車 이어 IT株도 '미끄럼'지난해 ELW 시장 홍콩 이어 세계 2위 차지[장외주식 ...	[도, 미끄럼, 지난해, 시장, 홍콩, 세계, 위, 차지, 장외, 주식, 황, 삼성...	1694.12	1700.79	1705.73	1694.12	407680000.0	-0.0007

데이터 수집 및 탐색

- 뉴스데이터의 분석
  - 다음과 같이 자체 감성사전을 구현해보려 했으나 실패
  - TF-IDF 가중치를 이용한 Naïve Bayes 모델 적용

$$TF = \frac{\text{문서 내 단어의 개수}}{\text{문서 내 모든 단어의 수}}$$

$$IDF = \log\left(\frac{\text{문서 전체 갯수}}{\text{단어를 포함한 문서의 수}}\right)$$

  - 가설1: 오늘의 뉴스가 익일 코스피 지수에 영향을 미친다.

정확도: 0.5115789473684211

  - 가설2: 오늘의 뉴스가 당일 코스피 지수에 영향을 미친다.

정확도: 0.6589473684210526

  - 익일 코스피 지수에 영향 없음. 모델에 적용하기 어려움

데이터 수집 및 탐색

- 데이터 요약, 사용가능여부

분류	항목	개수	사용	사유
코스피 지수관련	기본지표(OHLCV)	5	O	
	보조지표	40	O	
	신용잔고	1	O	
	해외지수	1	△	결측값 이전값 대체
거시경제 지표	경기종합지수	1	X	월별공시 2개월 지연
	물가지수	1	X	월별공시 1개월 지연
	GDP	1	X	월별공시 1개월 지연
	환율(원달러,원위안)	2	O	결측값 이전값 대체
	금가격	1	O	결측값 이전값 대체
뉴스데이터	네이버증권뉴스	1	X	감성사전 미구현

머신러닝의 시작

- 어떤것을 학습할 것인가? 매우 중요
- 목표변수는? Labeling
- 목표변수는 노이즈를 최소화해야한다.

목표변수	예시	모델	학습
X일후의 상승하락?	0 또는 1	분류	X
X일후의 가격은?	2100	회귀	X
X일후의 수익률은?	5.2%	회귀	△
Buy & Sell 지표	0 또는 1	분류	○

머신러닝의 시작

- Buy & Sell 목표변수란?

머신러닝의 시작

- 데이터 전처리(1)
  - 데이터 병합, 결측값 처리, 레이블링

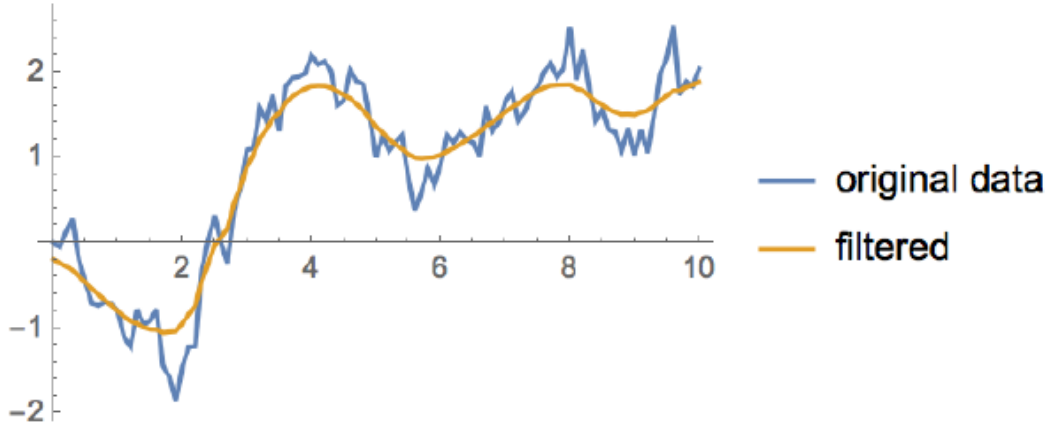
입력변수 feature

목표변수

	종가	시가	...	ROC	OBV	신용	신용OSC	USD_KRW	CNY_KRW	target
2019-08-06	-0.593627	-0.686137		-3.155591	1.210492	0.609047	-1.346506	2.110309	-0.040400	-1.0
2019-08-07	-0.633537	-0.558399		-2.990025	1.087145	0.452498	-1.694128	2.100995	-0.173416	-1.0
2019-08-08	-0.577694	-0.553539	...	-2.679878	1.213148	0.259937	-2.142720	1.988774	-0.223296	-1.0
2019-08-09	-0.489882	-0.483046		-2.253426	1.319238	0.220293	-2.221559	2.133480	-0.150743	-1.0
2019-08-12	-0.466622	-0.475526		-1.568702	1.388491	0.239327	-2.152749	2.240022	-0.034354	-1.0

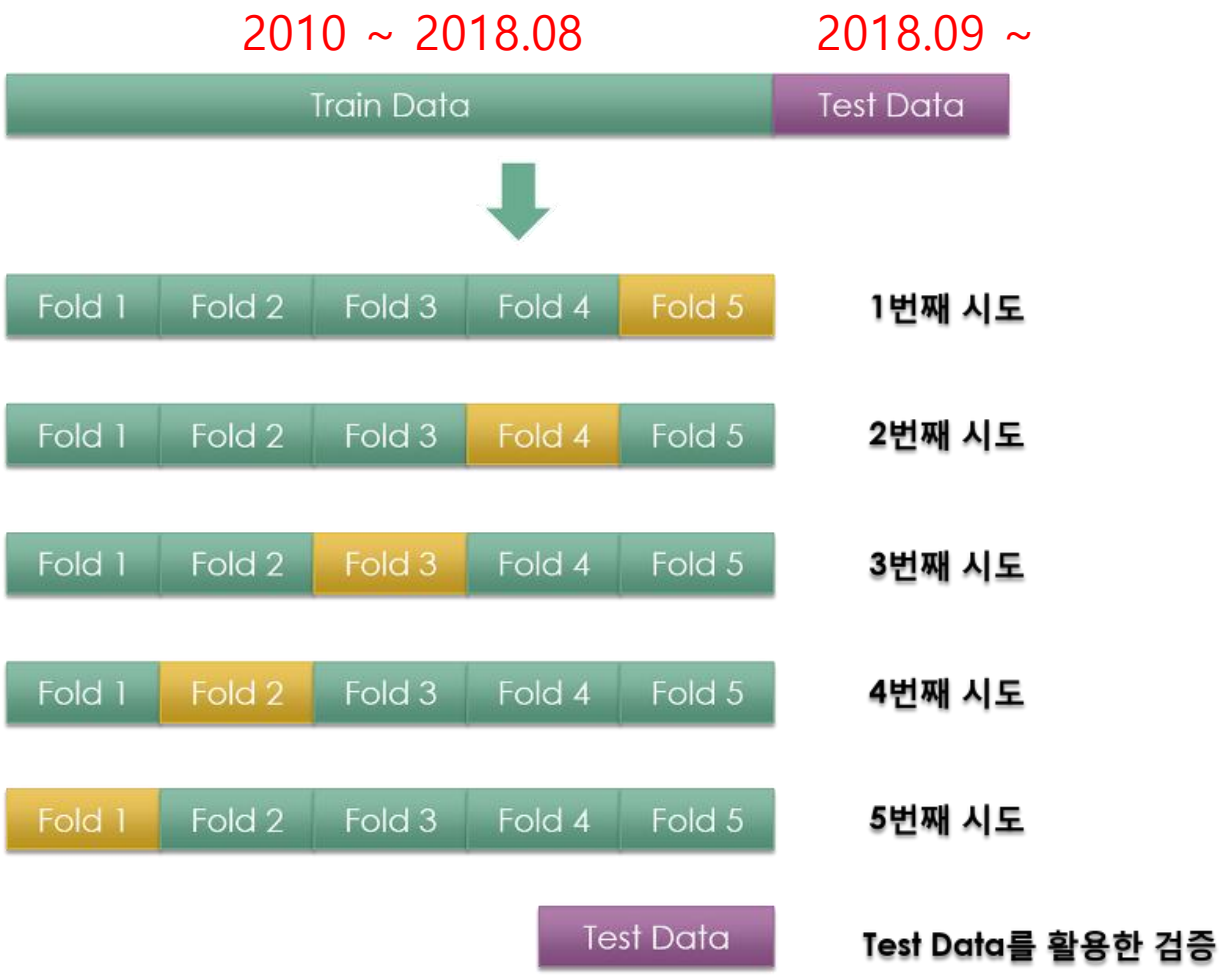
머신러닝의 시작

- 데이터 전처리(2)
  * 학습성능을 올리는데 도움을 준다.
  - 스케일링 : 0~1사이의 값으로 변환 (minmax, standard scaler)
  - 로그변환 : 값의 분포를 정규분포형태로 만든다. (log(a+1))
  - 디노이징 : 노이즈 필터링 (Savitzky-Golay Filter)



머신러닝의 시작

- 데이터셋의 분리



모델학습 및 평가

- 다양한 모델 결과비교
  - Train set / Validation set (2010~2018.08) 교차검증 (5분할)
  - Test set (2018.08~현재)

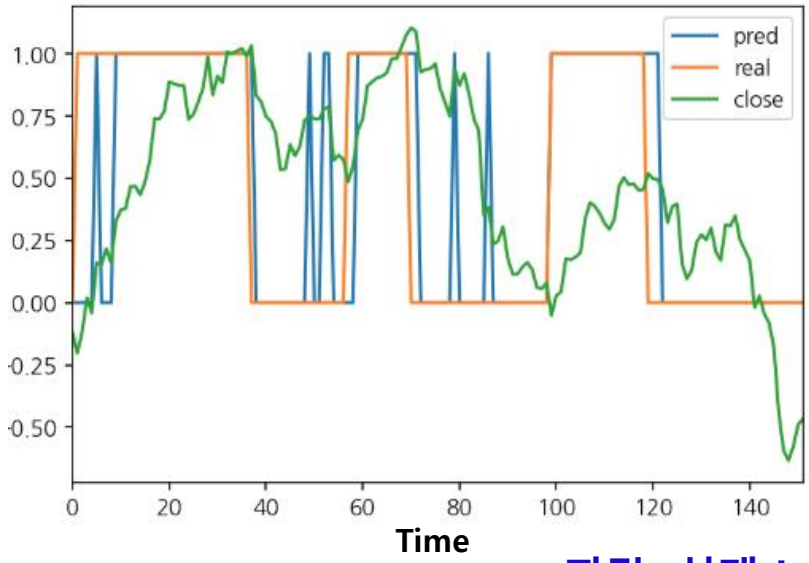
모델	Train / valid 교차검증 정확도	표준편차 (Std)	Test 정확도
Random	0.60		0.56
Linear SVM	0.74	0.007	
Radial SVM	0.77	0.010	0.65
Logistic Regression	0.74	0.012	
Naive Bayes	0.70	0.023	
Random Forest	0.84	0.010	0.79
AdaBoost	0.85	0.006	0.80
DNN	0.80	0.010	0.78

Random Forest : 결정트리의 베깅모델
AdaBoost: 결정트리의 부스팅모델

모델학습 및 평가

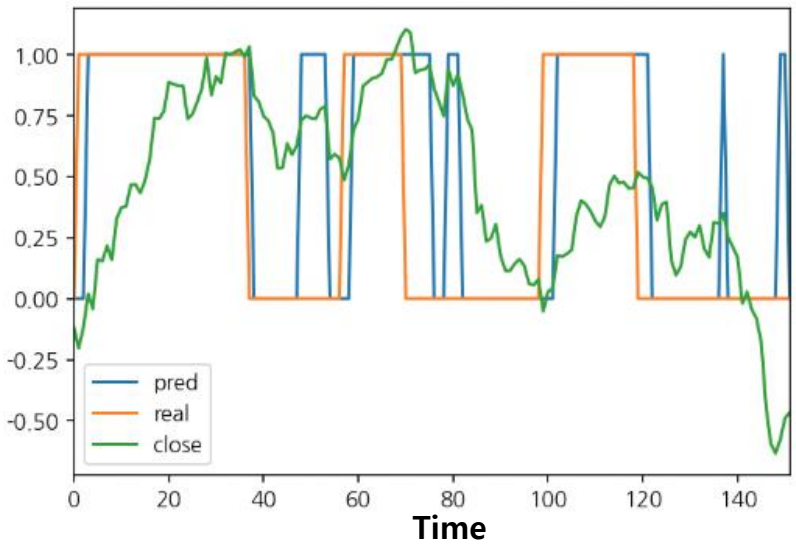
- 모델별 Testset 예측결과
  - Testset (2018.08~현재) , 80% accuracy

AdaBoost



Buy
Sell

Random Forest

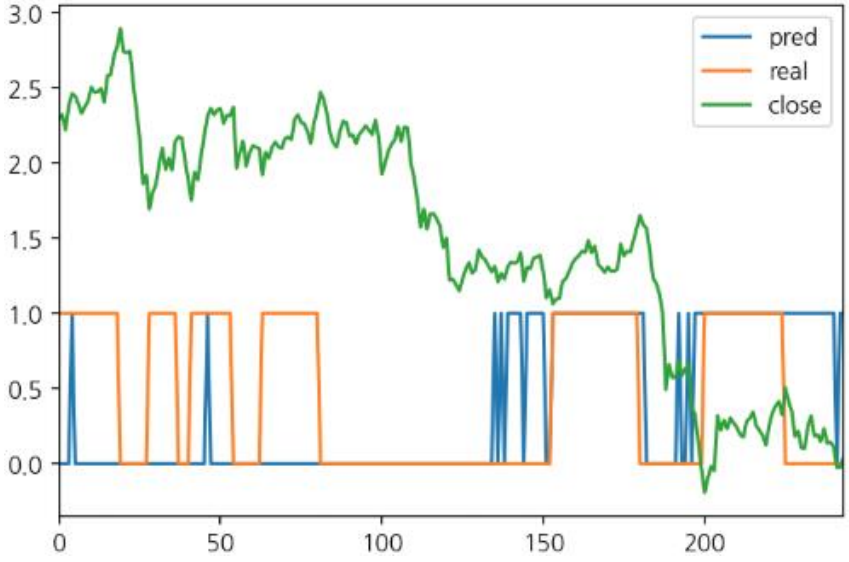


파랑: 실제 buy(1) or sell(0)
주황: 예측
초록: 주가

모델학습 및 평가

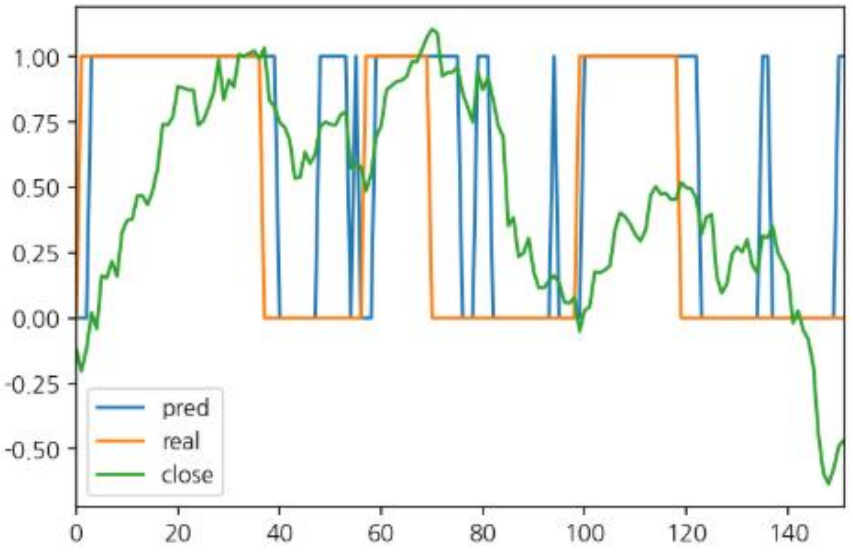
- 기간별 Testset 예측결과
  - 기간을 변경한 예측은 acc 60~70% 로 성능이 떨어진다. -> 과적합
  - 데이터를 늘리고 중요특성만으로 튜닝해보는 것이 필요

2018.01~ 2018.12



64%

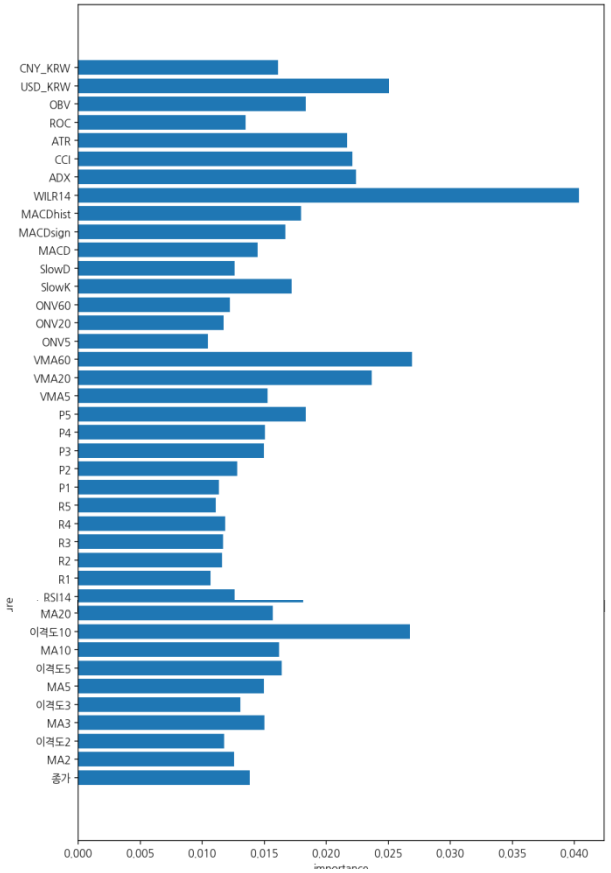
2019.01~ 2019.08



80%

중요변수 확인

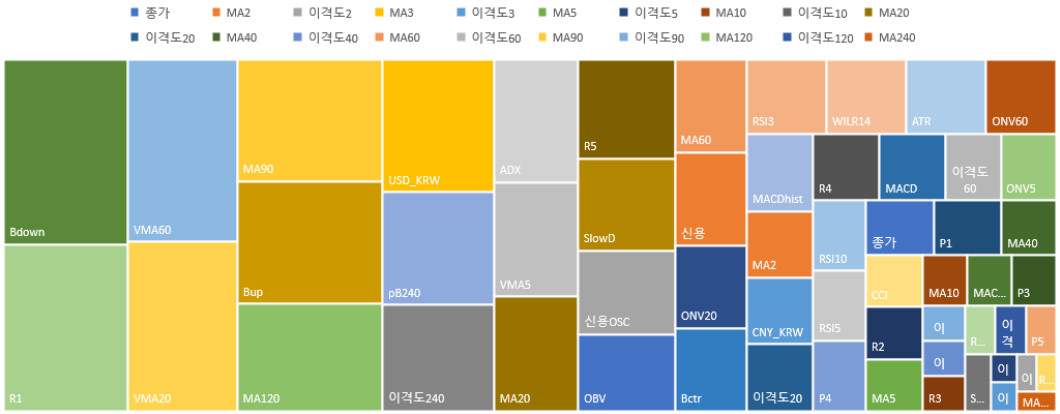
- garbage in garbage out , 특성중요도 확인
- 중요하지 않은 특성 제외하여 재학습 -> 성능비교필요

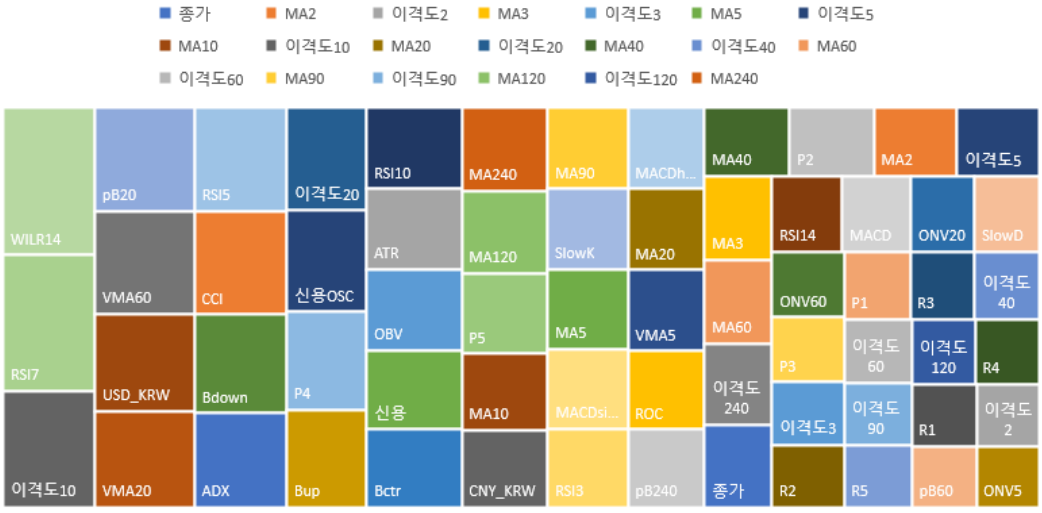


중요도 순위	특성	설명
1	Williams %R 14	최고점 대비 종가수준 모멘텀 지표
2	RSI 5,7	주가변동폭관련 상대강도지수
3	이격도 10	종가와 이평선 차이
4	VMA 20,60	거래량 이동평균
5	USD_KRW	환율
.....		
57	ONV5,20,60	거래량과 VMA 차이
58	pB60	볼린저밴드 종가 차이
59	이격도 40,60,120	종가와 이평선 차이
60	R1~R5	현종가 과거종가 변화율

중요변수 확인

- 특성중요도가 모델별, 기간별로 차이는 있지만 공통점을 확인





딥러닝 LSTM모델

- 모델의 세부구성
- 왜 LSTM? 시퀀스 데이터 사용가능

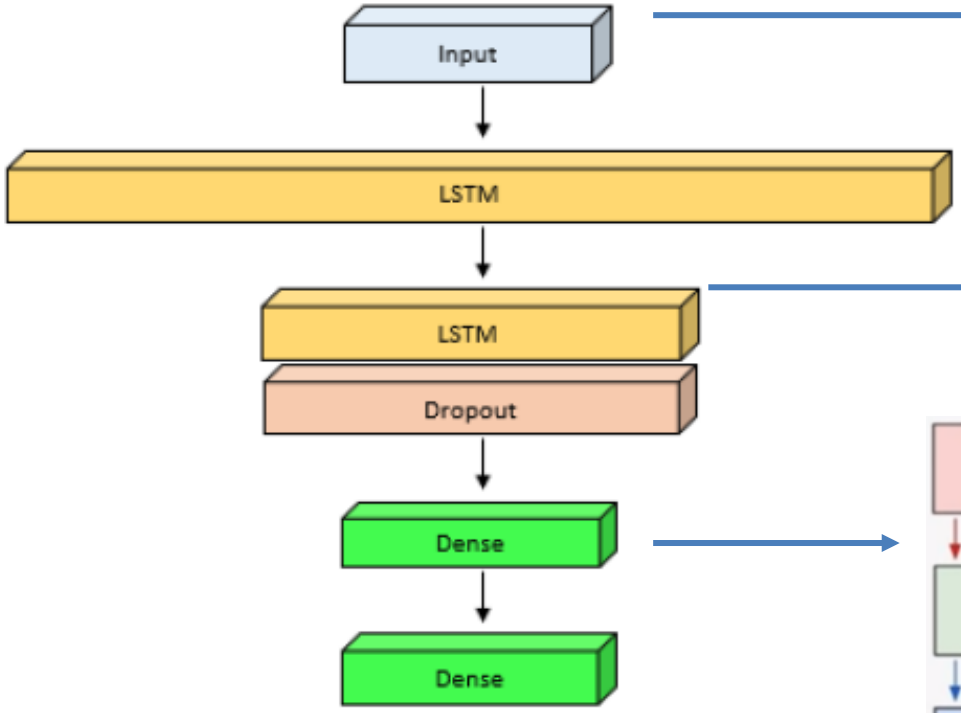
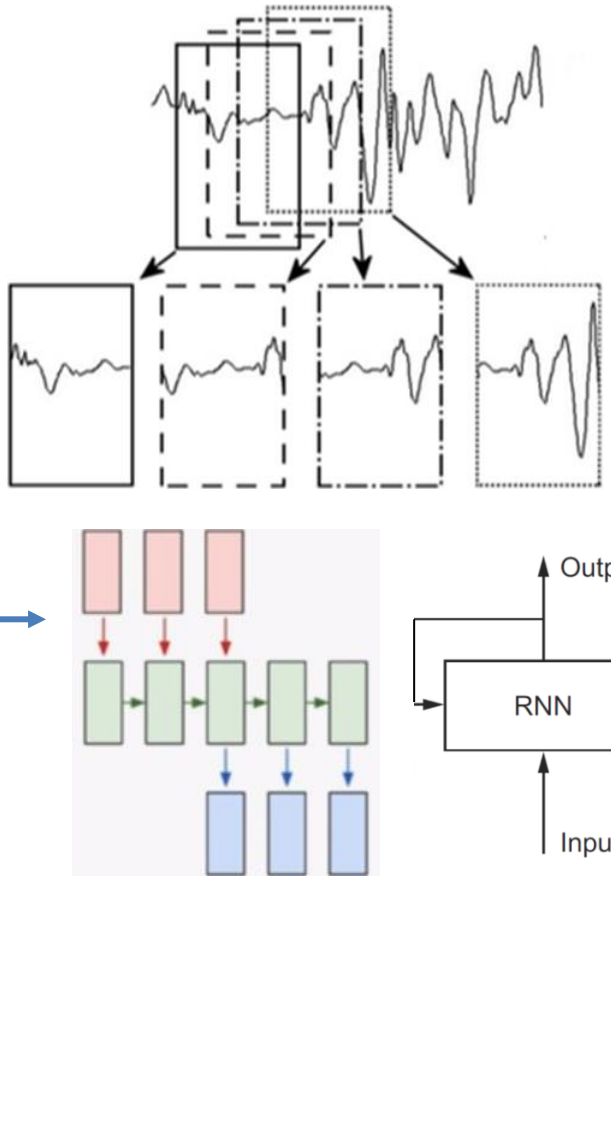


Fig. 3 LSTM Architecture

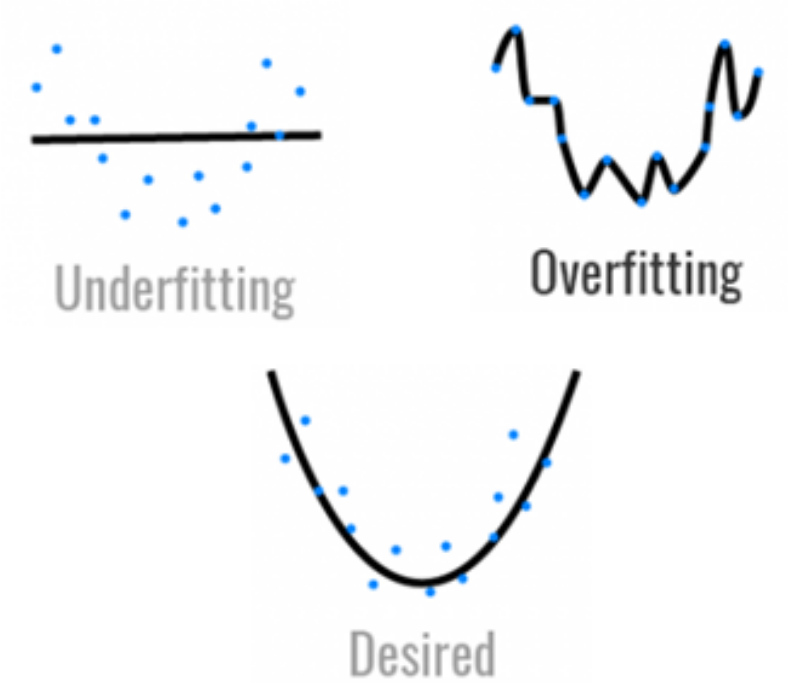


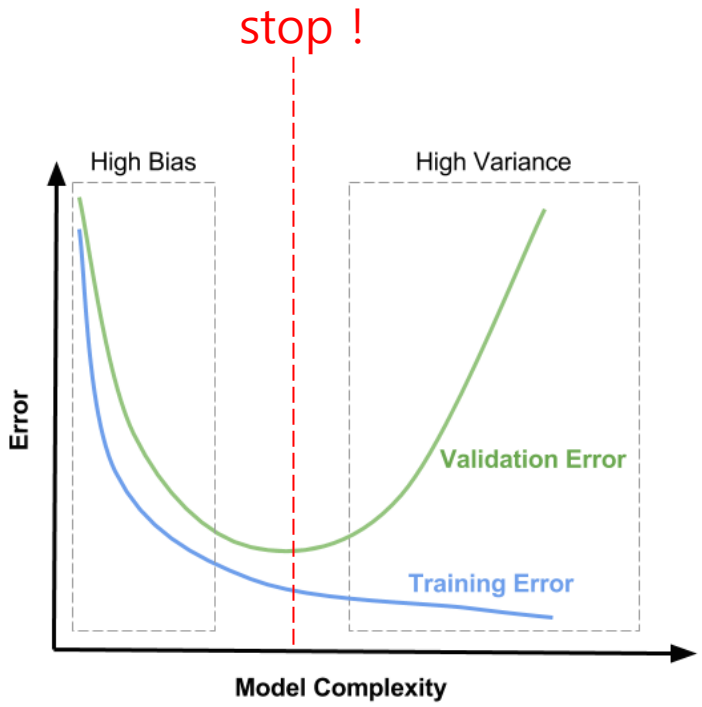
딥러닝 LSTM모델

- 차이점:
  - 목표변수(target) 변경
    - X일 후 수익률로 변경 (2일 후 종가변화율은?)
    - BUY & SELL의 경우 LSTM < 기존모델
    - X일 후 수익률의 경우 LSTM > 기존모델
  - 시퀀스 데이터 생성
    - 특정기간(20일) 단위로 잘라서 시계열 데이터 생성

딥러닝 LSTM모델

- 학습시 과적합 문제, 어떻게 해결?
  (1) dropout 추가, regularizer 적용
  (2) early stopping 기능 활용
  (3) val_loss 최소모델 자동저장





학습 손실곡선

딥러닝 LSTM모델

- 테스트셋 예측결과
  - Train set / Valid set (2010~2018.08), Test set (2018.08~현재)

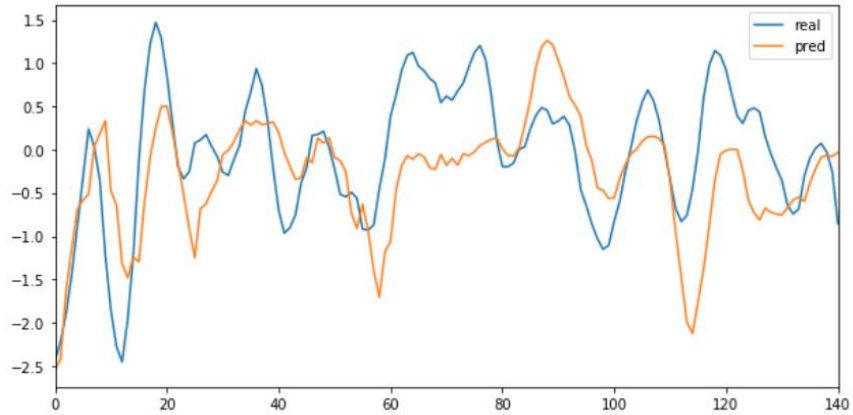
- 파라메타 변경에 따른 MSE 변화
  - MSE의 절대값은 의미 없음. 상대비교
  - 0.6 정도를 넘어가면 의미 없음. Random 모델

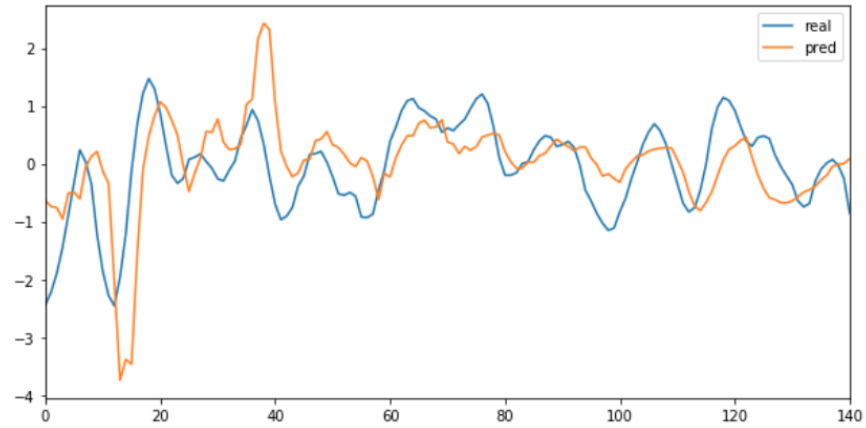
과거 시퀀스(일)	X일 예측	Test MSE
10	2	0.3 ~ 0.4
20	2	0.3 ~ 0.4
10	3	0.5 ~ 0.6
20	3	0.5 ~ 0.6
60	3	0.5 ~ 0.7
10	5	0.8 ~ 1
20	5	0.8 ~ 1
60	10	1 ~ 2

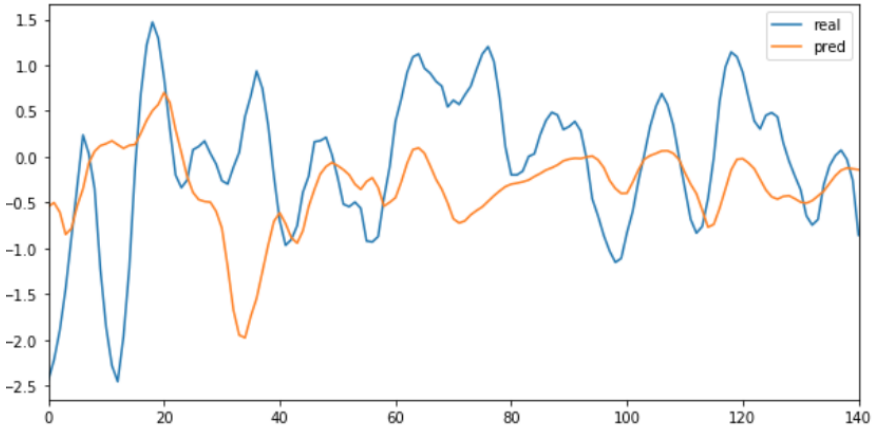
딥러닝 LSTM모델

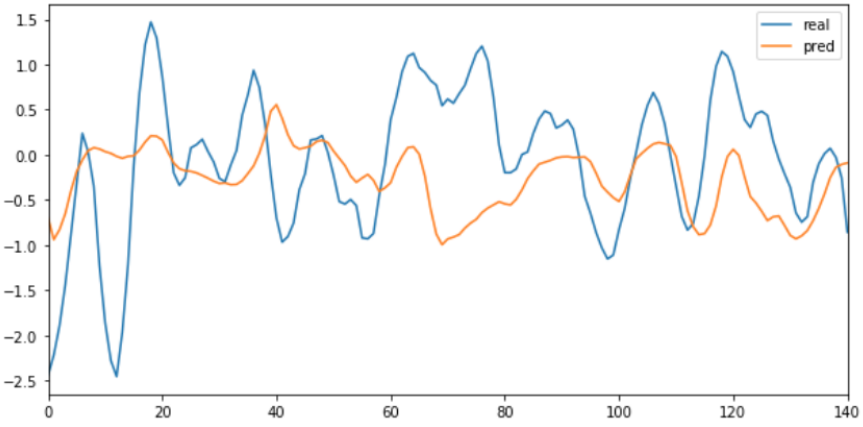
- 테스트셋 예측결과 (20,3)

x축: 시간, y축: 3일후수익률
파랑: 실제치
주황: 예측치





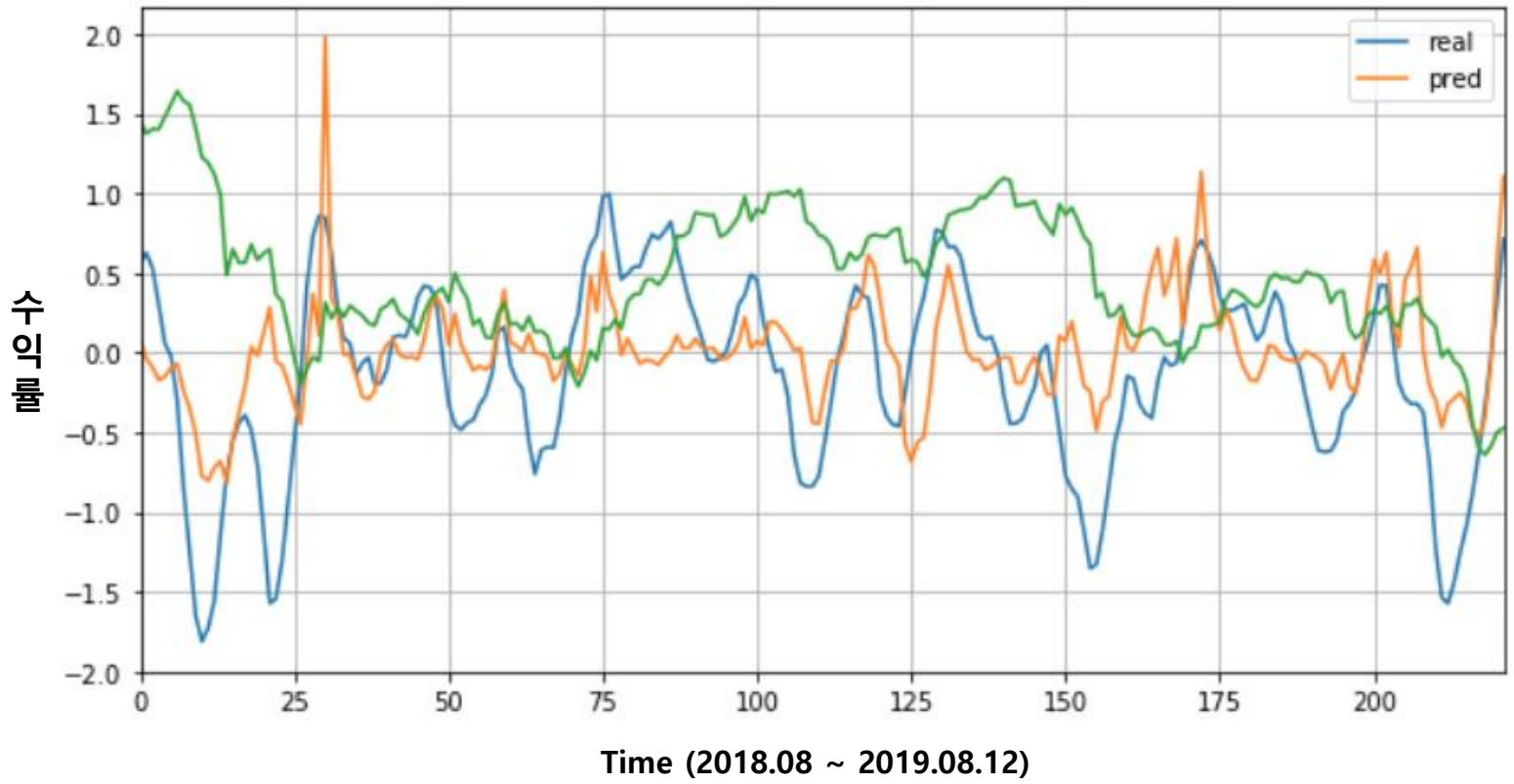




딥러닝 LSTM모델

- 테스트셋 예측결과 (10,2)

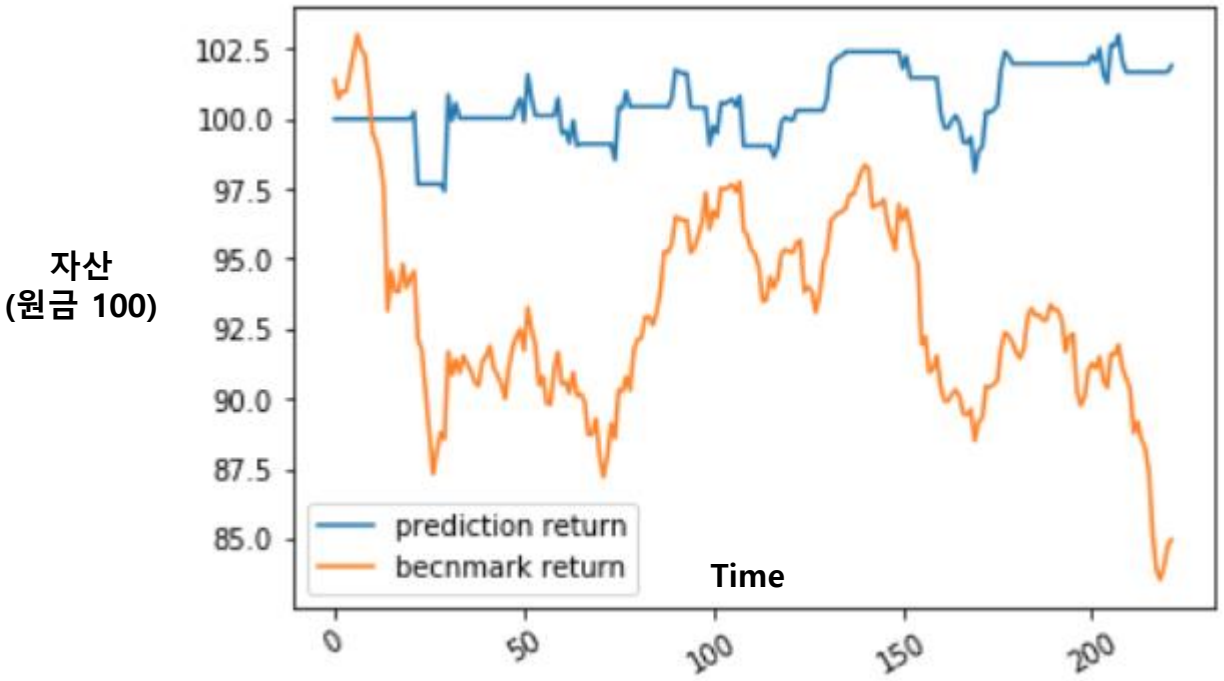
x축: 시간, y축: 2일후수익률
파랑: 실제치
주황: 예측치
초록: 주가



딥러닝 LSTM모델

- 단순전략 및 백테스팅
  - LSTM 모델 (수익률예측) 적용 , 기간 2018.08 ~ 2019.08.12
  - 예측값 0.1이상 매수 및 보유 / 0.1이하 매도

- 수익률 차트 : Benchmark : -15%, 예측모델: +2%



딥러닝 LSTM모델

- 파라메타 튜닝하여 성능 증가시킬 수 있음. (약 5~10%)

Hyper-parameter	Variants
Non-linearity	linear, tanh, sigmoid, ReLU, VLReLU, RReLU, PReLU, ELU, maxout, APL, combination
Batch Normalization (BN)	before non-linearity. after non-linearity
BN + non-linearity	linear, tanh, sigmoid, ReLU, VLReLU, RReLU, PReLU, ELU, maxout
Pooling	max, average, stochastic, max+average, strided convolution
Pooling window size	3x3, 2x2, 3x3 with zero-padding
Learning rate decay policy	step, square, square root, linear
Colorspace & Pre-processing	RGB, HSV, YCrCb, grayscale, learned, CLAHE, histogram equalized
Classifier design	pooling-FC-FC-clf, SPP-FC-FC-clf, pooling-conv-conv-clf-avepool, pooling-conv-conv-avepool-clf
Network width	$1/4, 1/2\sqrt{2}, 1/2, 1/\sqrt{2}, 1, \sqrt{2}, 2, 2\sqrt{2}, 4, 4\sqrt{2}$
Input image size	64, 96, 128, 180, 224
Dataset size	200K, 400K, 600K, 800K, 1200K(full)
Batch size	1, 32, 64, 128, 256, 512, 1024
Percentage of noisy data	0, 5%, 10%, 15%, 32%
Using bias	yes/no

결론

- 주가 모델예측에 꼭 필요한 것들
  - denoising: 데이터 노이즈 최소화 (필터, wavelet)
  - 목표변수(target)의 적절한 선정
  - 입력변수의 적절한 선정 : 중요도를 확인하여 입력변수를 줄여나가야
  - 하이퍼파라메타 튜닝 -> 조합다양, 시간/노력, 튜닝자동화 필요
  - 과적합 방지: 주가예측은 특히 과적합으로 학습이 매우 어려움.
  - 기간별 교차검증 : 기간별 성능 변화가 심함. 모델불안정

향후 연구과제

- 아쉬웠던 점
  - 데이터분석 부족, 유의미한 데이터의 발견
  - 뉴스 감성분석 적용시 성능?, 증권감성사전의 부재
  - LSTM 변형모델 –> DARNN, AE, GAN적용
  - 강화학습 적용
- 활용범위
  - 개별종목, 고빈도매매에 적용
  - 타분야 시계열 데이터 분석 및 예측 응용

Q&A

참고자료

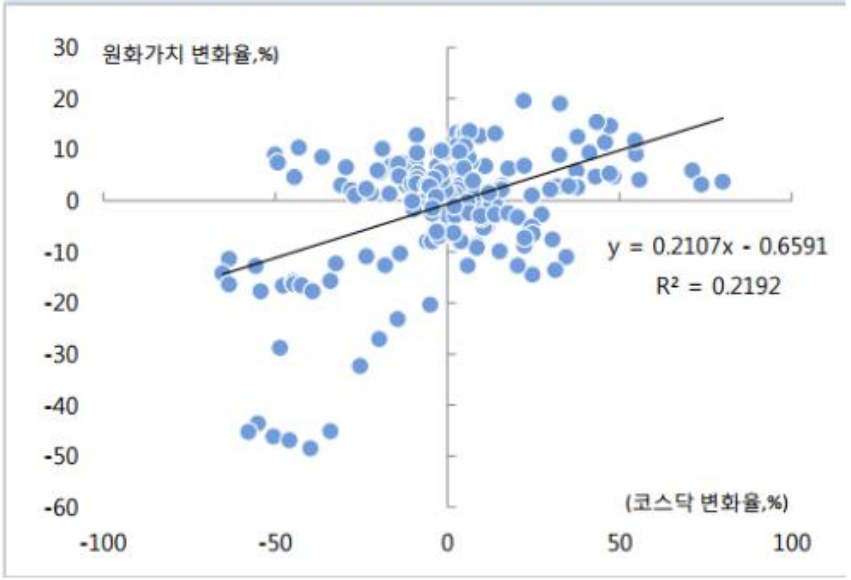
1) A deep learning framework for financial time series using stacked autoencoders and long-short term memory, 2017
2) Forecasting East Asian Indices Futures via a Novel Hybrid of Wavelet-PCA Denoising and Artificial Neural Network Models, 2016
3) 디노이징 필터와 LSTM을 활용한 KOSPI200 선물지수 예측, 2019
4) NN과 Text Mining을 이용한 종합주가지수 예측, 2016
5) Stock Price Prediction on Daily Stock Data using Deep Neural Networks, 2018
6) https://www.slideshare.net/seorop/1-108783951
7) http://www.newsis.com/msi/whatmsi/

보충자료

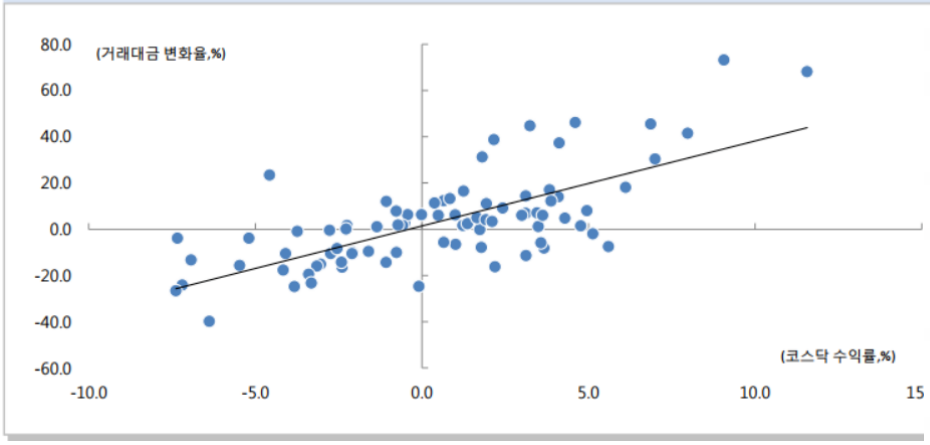
단순추이보다는 변화율이 중요하다 !!!

코스닥 수익률 변화와 원화가치 변화율

코스닥 수익률과 원화가치 변화율의 산점도

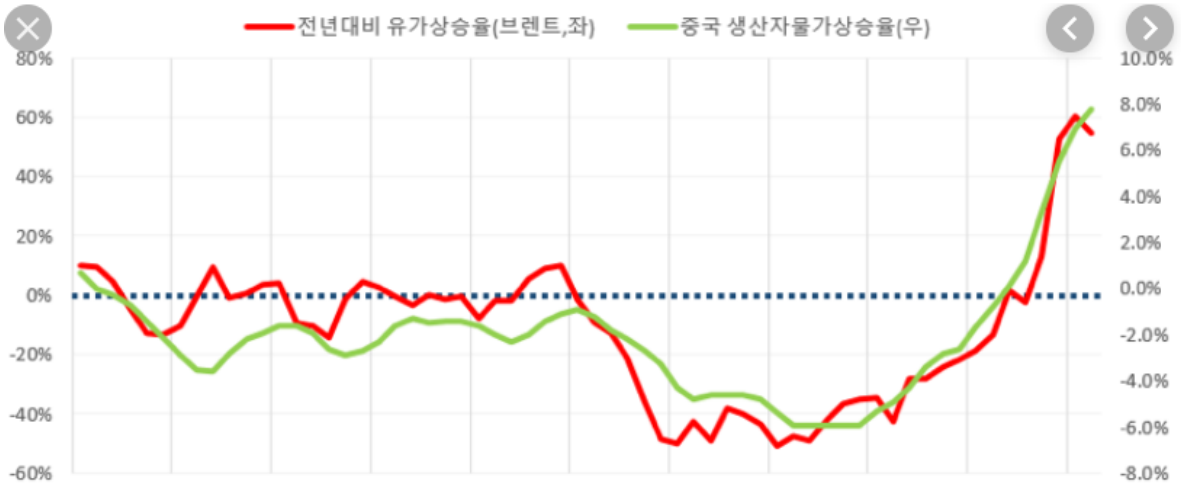


코스닥 수익률 향상을 위해서는 거래대금 증가가 필수

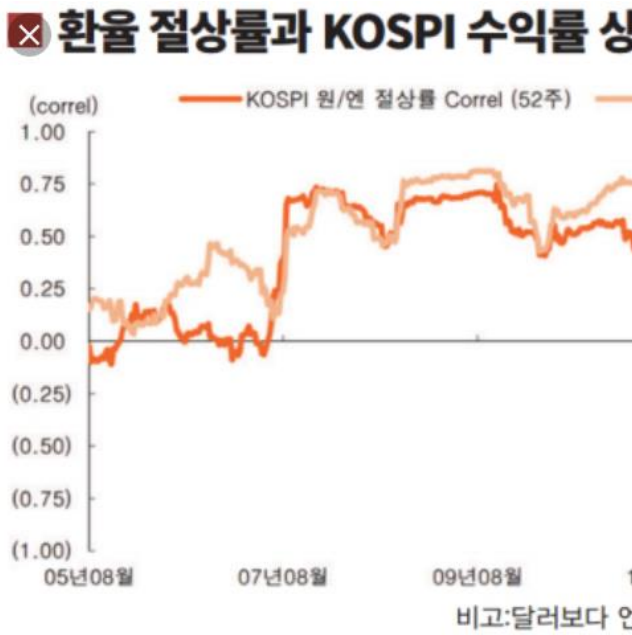


종합주가지수 및 국고채수익률 추이(월평균)

자료: 한국은행, ECOS.



전년대비 변화율



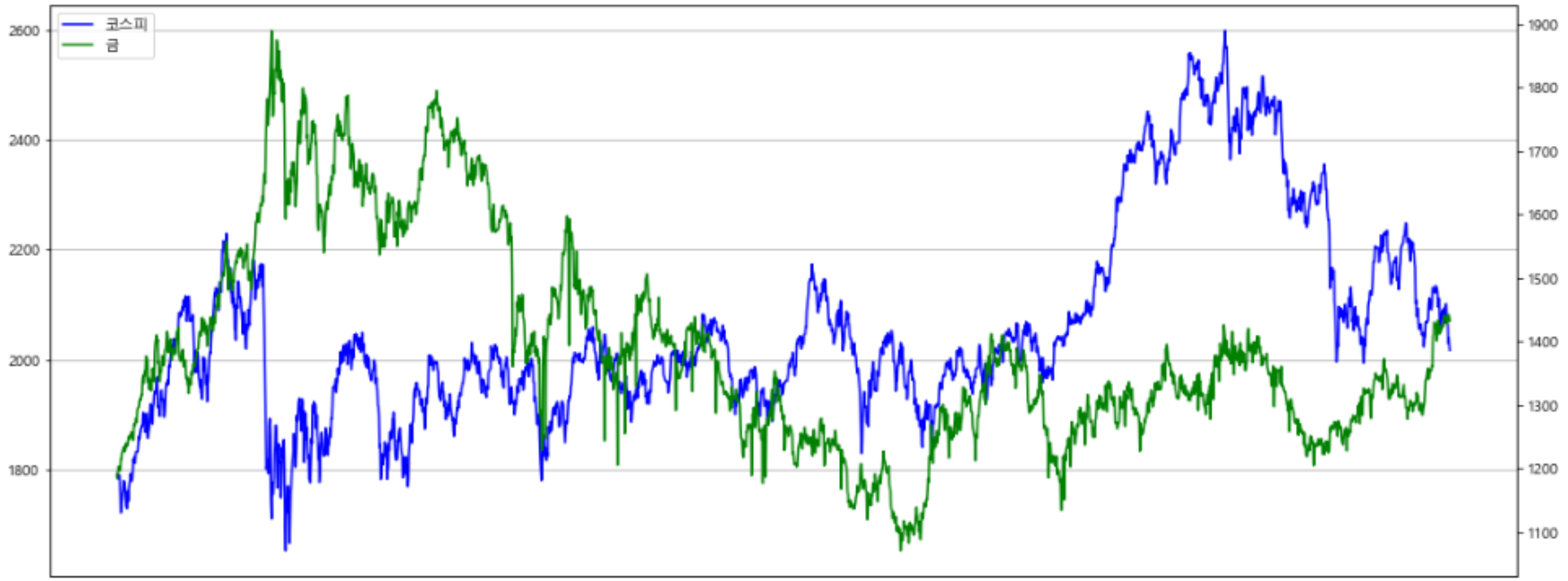
Window기법 상관관계확인

데이터수집 및 탐색

- 금가격 (차트)
- 부분적으로 음의 상관관계를 보인다.

```
total_index['kospi'].corr(total_index['gold'])
```

```
-0.27984048475733253
```

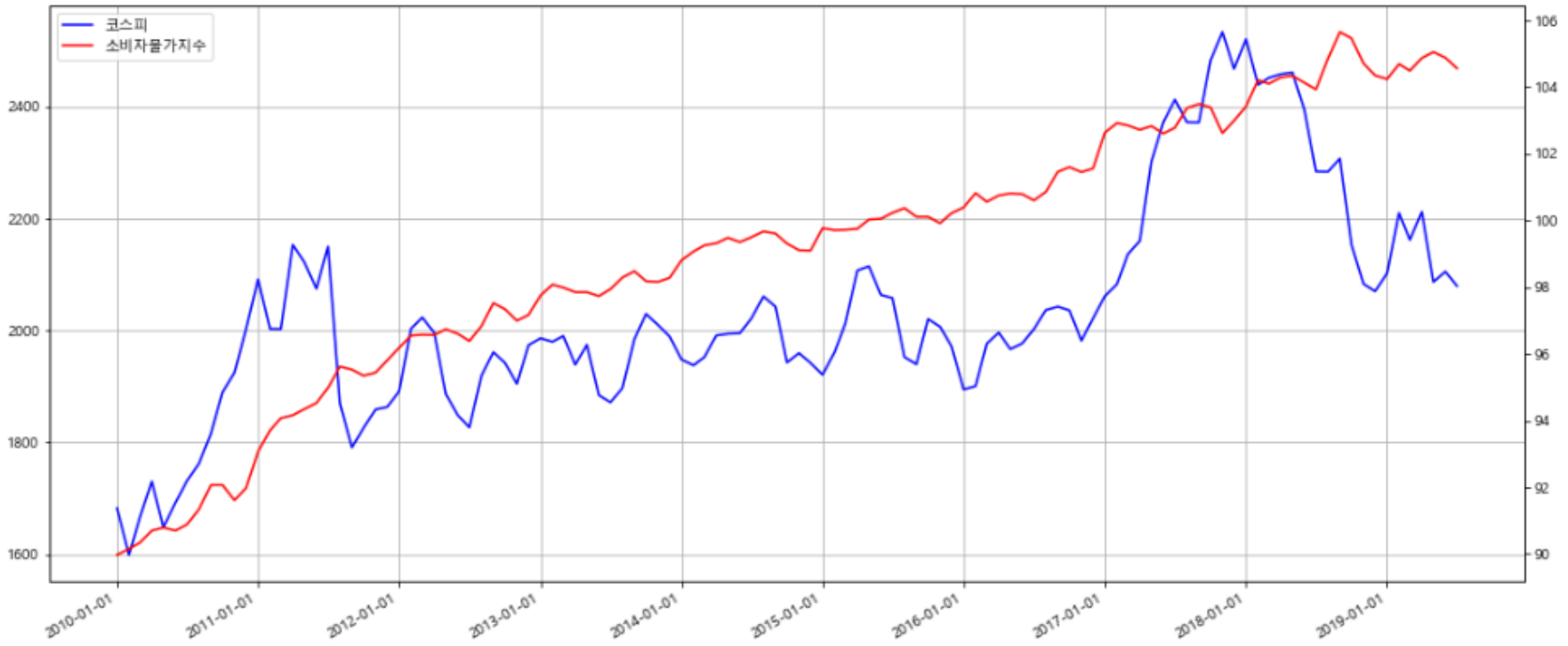


데이터수집 및 탐색

- 거시경제지표
  - 코스피와 소비자물가지수의 상관관계 확인

```
customer_price['price'].corr(df_temp['close_mean'])
```

```
0.737250260161971
```

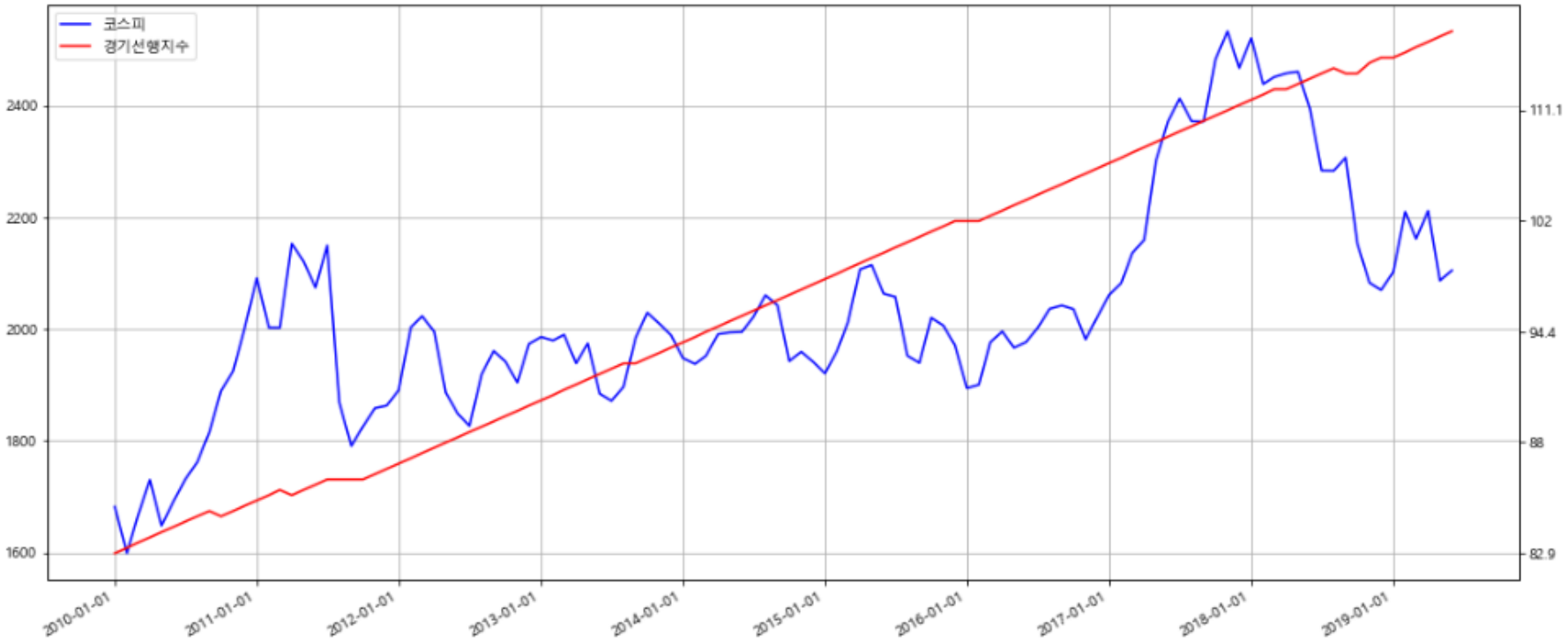


데이터수집 및 탐색

- 거시경제지표
  - 코스피와 경기선행지수의 상관관계 확인

```
df_graph['월평균코스피'].corr(df_graph['경기선행지수'])
```

```
0.7522864264242258
```

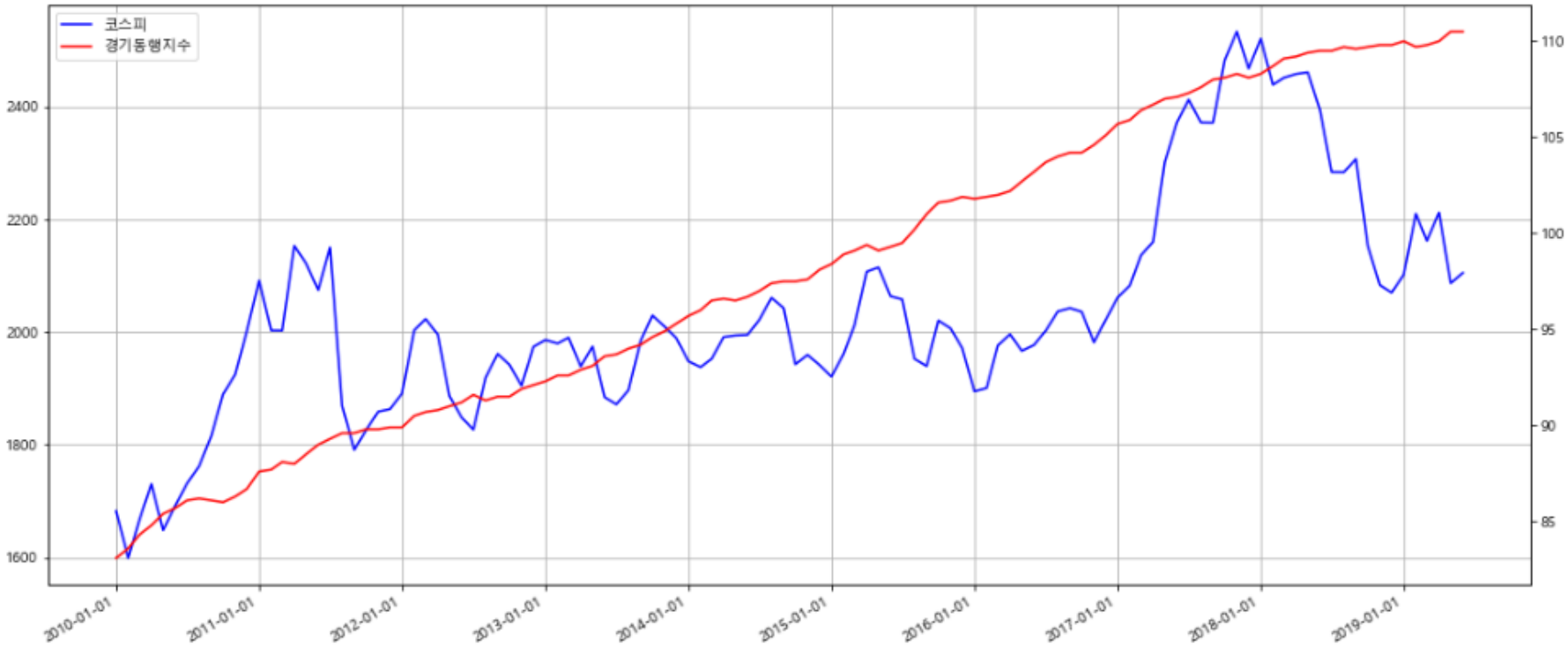


데이터수집 및 탐색

- 거시경제지표
  - 코스피와 경기동행지수의 상관관계 확인

```
df_graph['월평균코스피'].corr(df_graph['경기동행지수'])
```

```
0.7512044502814994
```

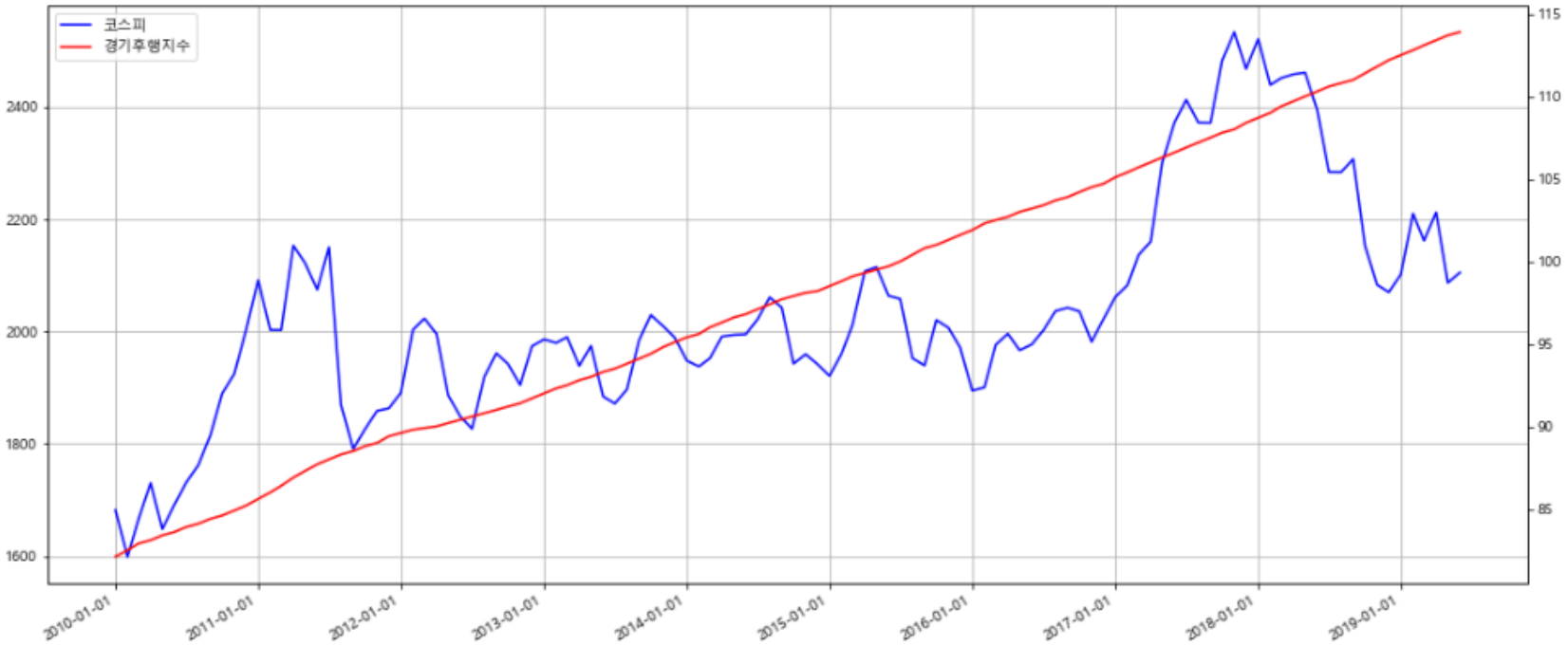


데이터수집 및 탐색

- 거시경제지표
  - 코스피와 경기후행지수의 상관관계 확인

```
df_graph['월평균코스피'].corr(df_graph['경기후행지수'])
```

```
0.7314247071000792
```



데이터수집 및 탐색

- 뉴스데이터