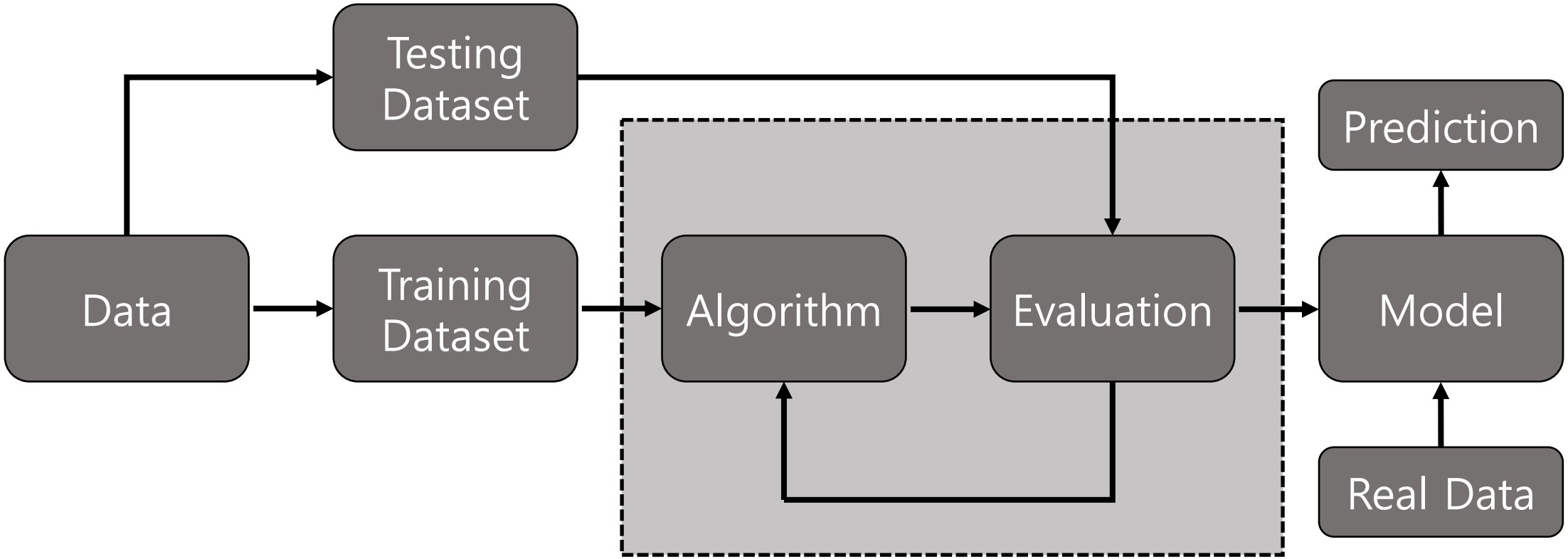
Google 주가예측 프로젝트

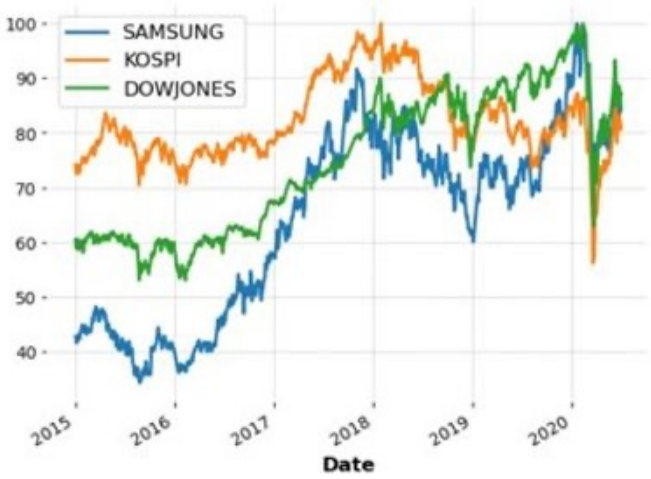
전체 이미지 동영상 뉴스 쇼핑 더보기



Google 데이터 수집

전체 이미지 동영상 뉴스 쇼핑 더보기

1차 데이터 수집



▶데이터 기간: 2018/01/01~2020/06/30

▶수집경로: investing.com/네이버 금융 크롤링

▶삼성전자, 코스피, 금 시세, 1년 채권, 다우존스 지수, 옥수수 선물

2차 데이터 추가 수집

▶데이터 기간: 2016/01/01~2020/06/30

▶수집경로: investing.com/네이버 금융 크롤링

▶삼성전자, 코스피, 금 시세, ~~1년 채권, 다우존스 지수, 옥수수 선물~~ +나스닥, 환율, 필라델피아 반도체지수, 외국인 매매량, 기관 매매량, 10년 채권, 재무제표(PBR, PER, ROE)

Google

데이터 전처리

전체 이미지 동영상 뉴스 쇼핑 더보기

	Date	Open	High	Low	Close	Volume	nasdaq	exchange	Semiconductor	foreign	institution	KOSPI	Korea_bond_10year
2	2016-01-04	25200.0	25200.0	24100.0	24100.0	306939.0	4903.09	1175.66	656.28	-56273.0	-61874.0	1918.76	2.035
3	2016-01-05	24040.0	24360.0	23720.0	24160.0	216002.0	4891.43	1190.35	649.47	-14965.0	-27047.0	1930.53	2.062
4	2016-01-06	24160.0	24160.0	23360.0	23500.0	366752.0	4835.77	1190.81	631.2	-21984.0	-67968.0	1925.43	2.037
5	2016-01-07	23320.0	23660.0	23020.0	23260.0	282388.0	4689.43	1200.52	610.25	-13307.0	-58060.0	1904.33	2.019
6	2016-01-08	23260.0	23720.0	23260.0	23420.0	257763.0	4643.63	1197.29	600.48	15806.0	-47699.0	1917.62	2.057
7	2016-01-11	23120.0	23320.0	22920.0	23040.0	241277.0	4637.99	1207.94	603.55	27198.0	-67620.0	1894.84	2.021
8	2016-01-12	22960.0	23320.0	22880.0	22920.0	206283.0	4685.92	1204.63	608.56	-11452.0	-39611.0	1890.86	2.046
9	2016-01-13	23060.0	23180.0	22960.0	22960.0	143316.0	4526.06	1210.4	589.48	8410.0	-12871.0	1916.28	2.023
10	2016-01-14	22620.0	22840.0	22620.0	22760.0	209022.0	4615.0	1210.24	601.57	16442.0	-32252.0	1900.01	2.028
11	2016-01-15	22800.0	23040.0	22480.0	22640.0	209464.0	4488.42	1207.73	574.29	1349.0	-9288.0	1878.87	2.023
12	2016-01-19	22560.0	23420.0	22560.0	23420.0	207242.0	4476.95	1211.72	575.18	-19997.0	45854.0	1889.64	2.068
13	2016-01-20	23200.0	23200.0	22640.0	22760.0	167052.0	4471.69	1207.39	578.94	7863.0	-26311.0	1845.45	2.01

▶삼성전자 액면분할 전 거래량 *50

▶액면분할로 거래량이 없는 2018/04/30~2018/05/03 데이터 제거

▶5일, 10일, 20일, 60일 이동평균선 추가

▶거래량의 단위(K, M, B) 통일

▶년-월-일로 되어 있는 날짜 형식 변환 후 index지정

▶object로 된 데이터 타입 float로 변경

▶각각 다른 테이블에 있던 데이터를 하나의 테이블로 병합

▶삼성전자 주가와의 상관관계 분석 및 시각화

Google 기본적 분석과 기술적 분석

전체 이미지 동영상 뉴스 쇼핑 더보기

기본적 분석

기업의 내재가치에 관련된 다양한 조사와 분석을 통해 주가의 방향을 예측하는 방법

기본적 분석을 이루는 경제요인으로 세계 GDP, 금리, 통화량, 금값, WTI, 10년물 채권, 옥수수와 같은 원자재를, 산업요인으로 필라델피아 반도체 지수, 기업요인으로 영업이익, PER, PBR, ROE 등 삼성전자 주가와의 상관관계 분석

기술적 분석

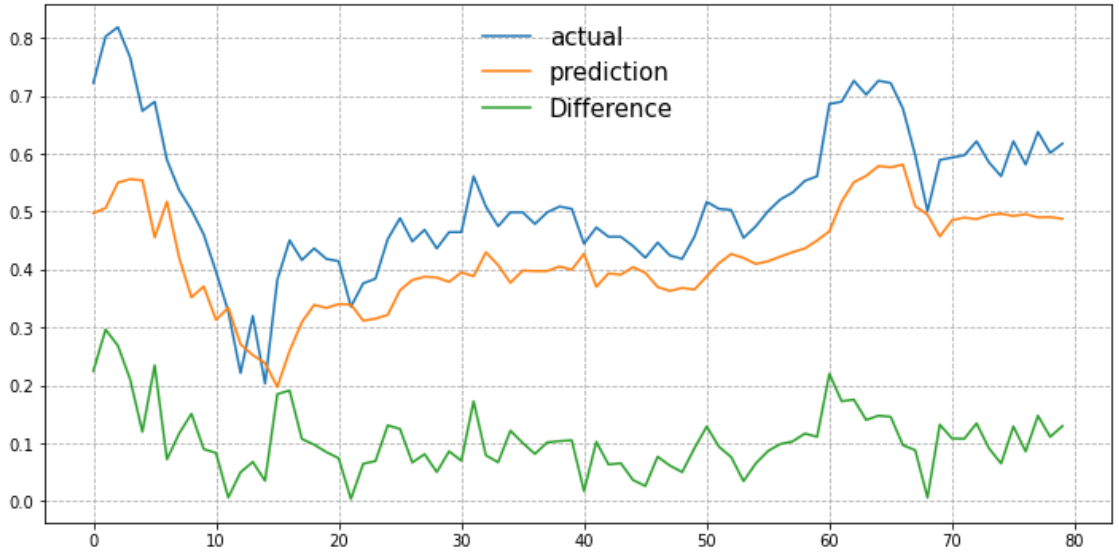
차트 분석이라고도 하며, 과거의 데이터를 이용하여 미래를 예측하는 방법
기술적 분석 지표로는 차트의 시가, 고가, 저가, 종가, 거래량이 있고, 패턴과 관련된 여러가지 보조지표를 활용

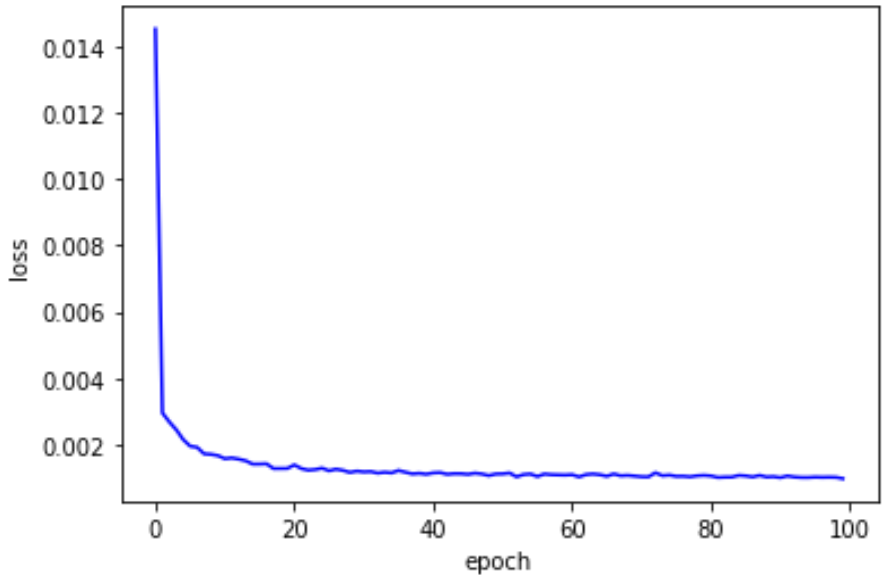
Google LSTM

전체 이미지 동영상 뉴스 쇼핑 더보기

▶삼성전자 주가와 상관관계가 높은 코스피, 나스닥, 반도체지수만을 추가로 학습하여 삼성전자 종가 예측

▶RMSE: 0.1189



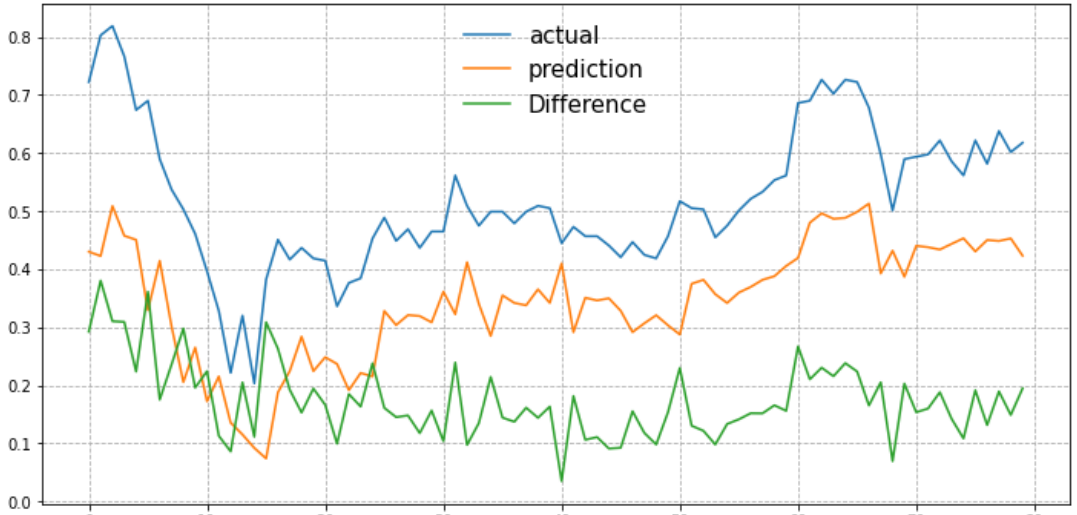


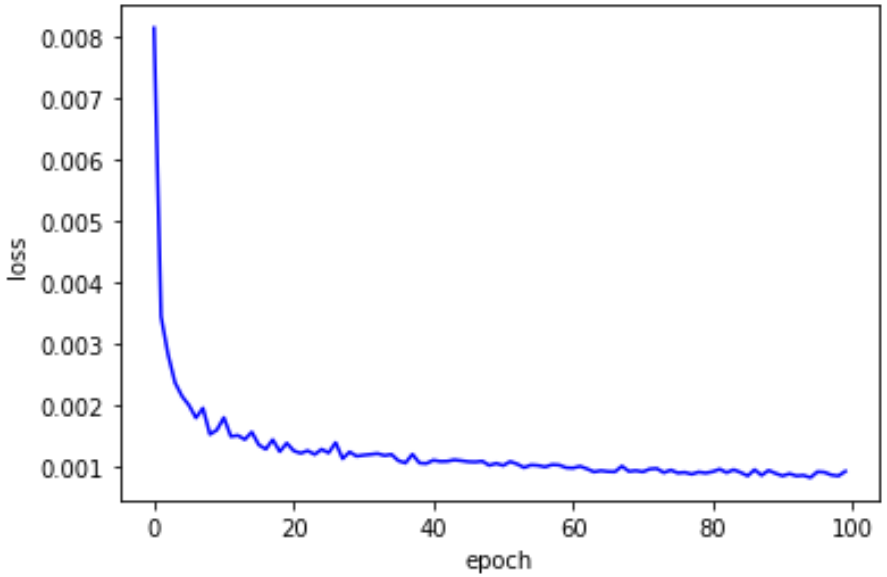
Google LSTM

전체 이미지 동영상 뉴스 쇼핑 더보기

▶삼성전자 주가+코스피, 나스닥, 외국인거래량, 반도체지수, 재무제표+5일, 10일, 20일, 60일 이동평균선 추가하여 예측

▶RMSE: 0.18696



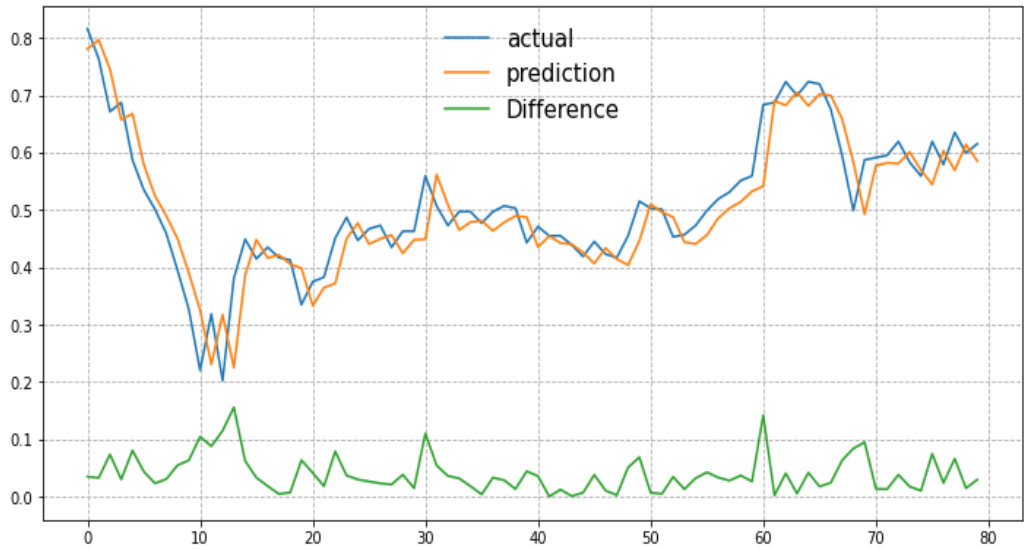


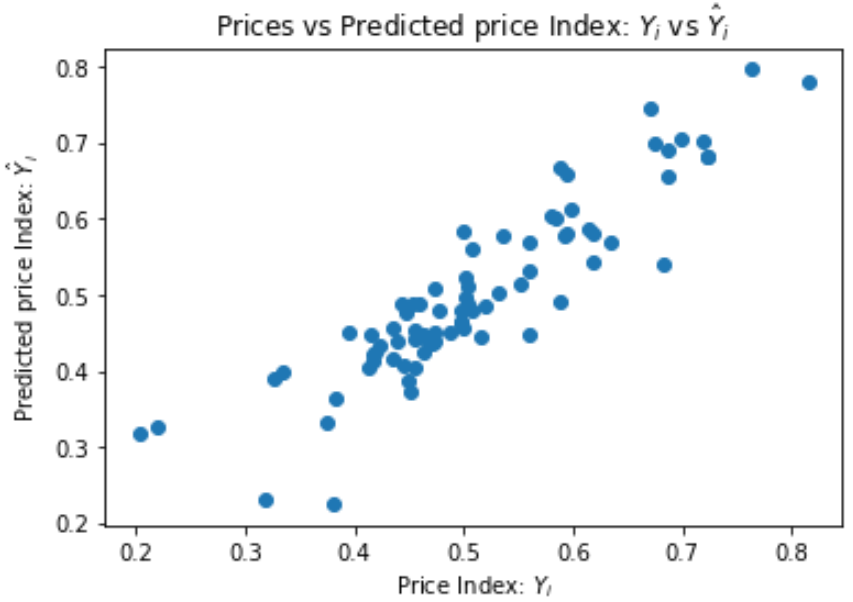
Google LSTM

전체 이미지 동영상 뉴스 쇼핑 더보기

▶삼성전자 주가 중 시가, 종가, 거래량 등을 제외한 오직 '종가'만을 가지고 예측

▶RSME: 0.0506



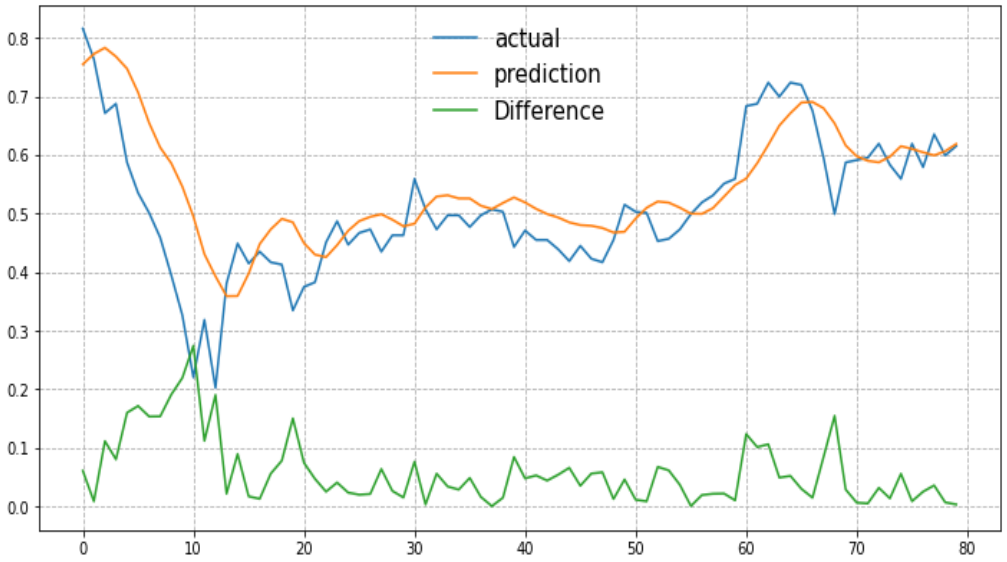


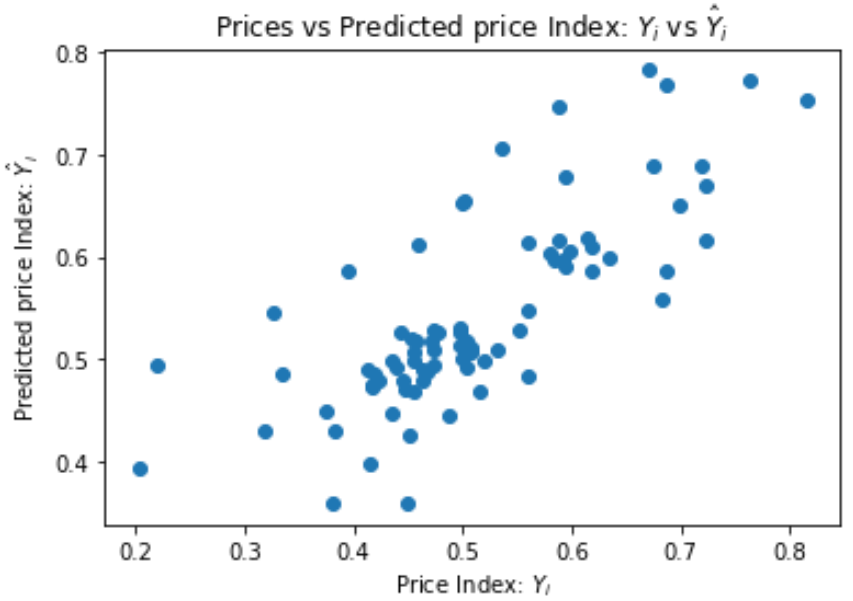
Google LSTM

전체 이미지 동영상 뉴스 쇼핑 더보기

▶삼성전자 종가에 5일, 10일, 20일, 60일 이동평균선을 추가하여 예측

▶RSME: 0.0814



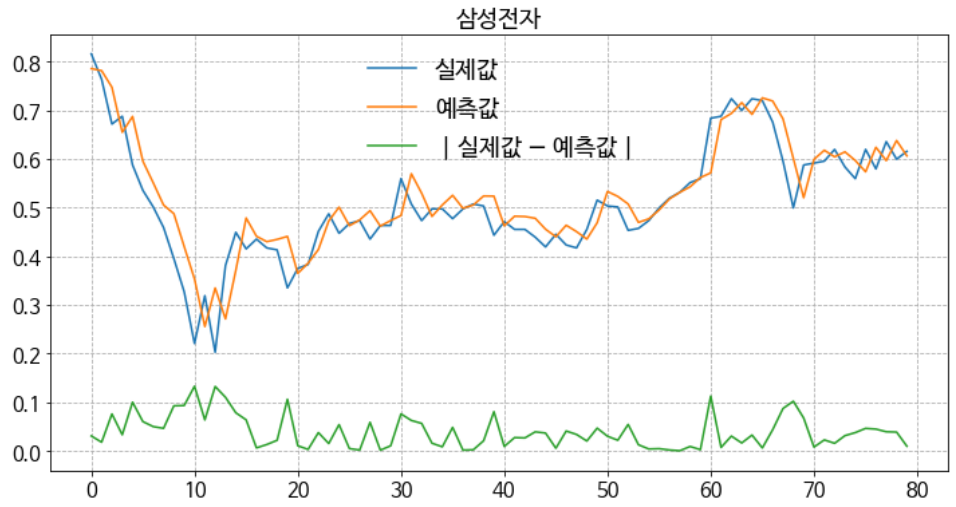


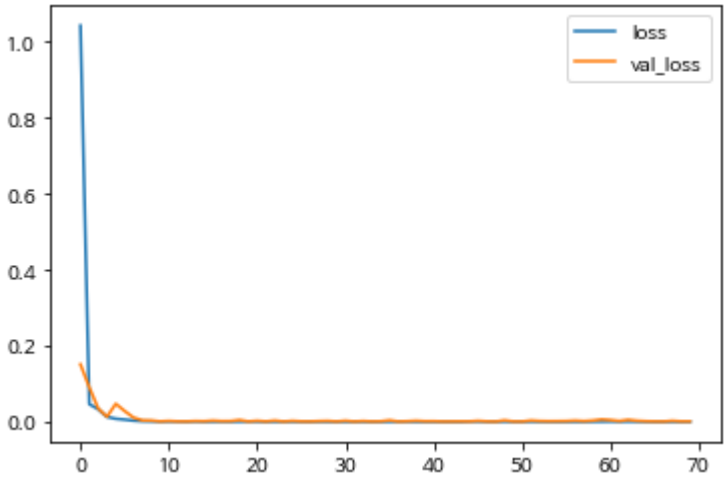
Google GRU

전체 이미지 동영상 뉴스 쇼핑 더보기

▶삼성전자 주가 시계열 중 오직 '종가'만을 이용하여 예측

▶RSME: 0.0026



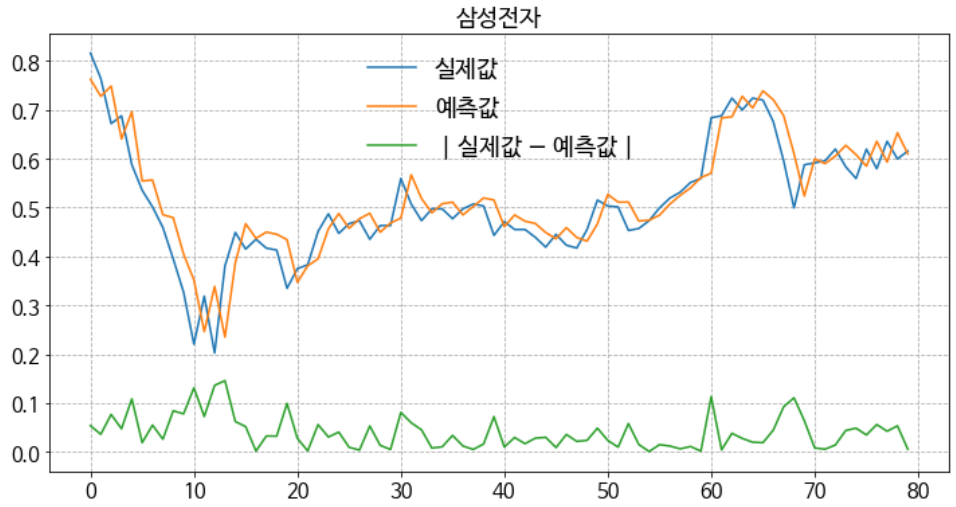


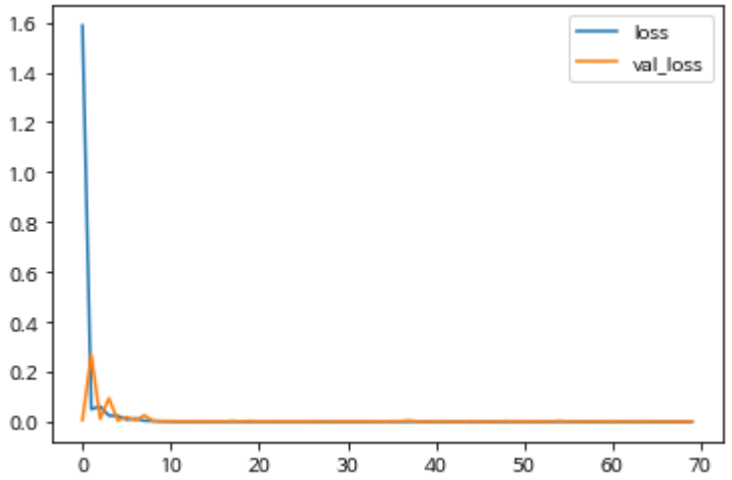
Google GRU

전체 이미지 동영상 뉴스 쇼핑 더보기

▶삼성전자 주가 시계열 중 종가와 5, 10, 20, 60일 이동평균을 특성치로 추가하여 예측

▶RSME: 0.0027





Google RMSE비교-LSTM

전체 이미지 동영상 뉴스 쇼핑 더보기

	삼성전자 +피처 15개	삼성전자 +코스피, 나스닥, 반도체지수	삼성전자 +피처 15개 +이동평균선	삼성전자 종가	삼성전자 종가 +이동평균선
RMSE	0.0541	0.1189	0.1869	0.0506	0.0813
Train Score	0.000703	0.001070	0.001177	0.00089	0.004054
Validation Score	0.001189	0.019015	0.029652	0.001477	0.002628
Test Score	0.002929	0.014154	0.034954	0.002558	0.006622

Google

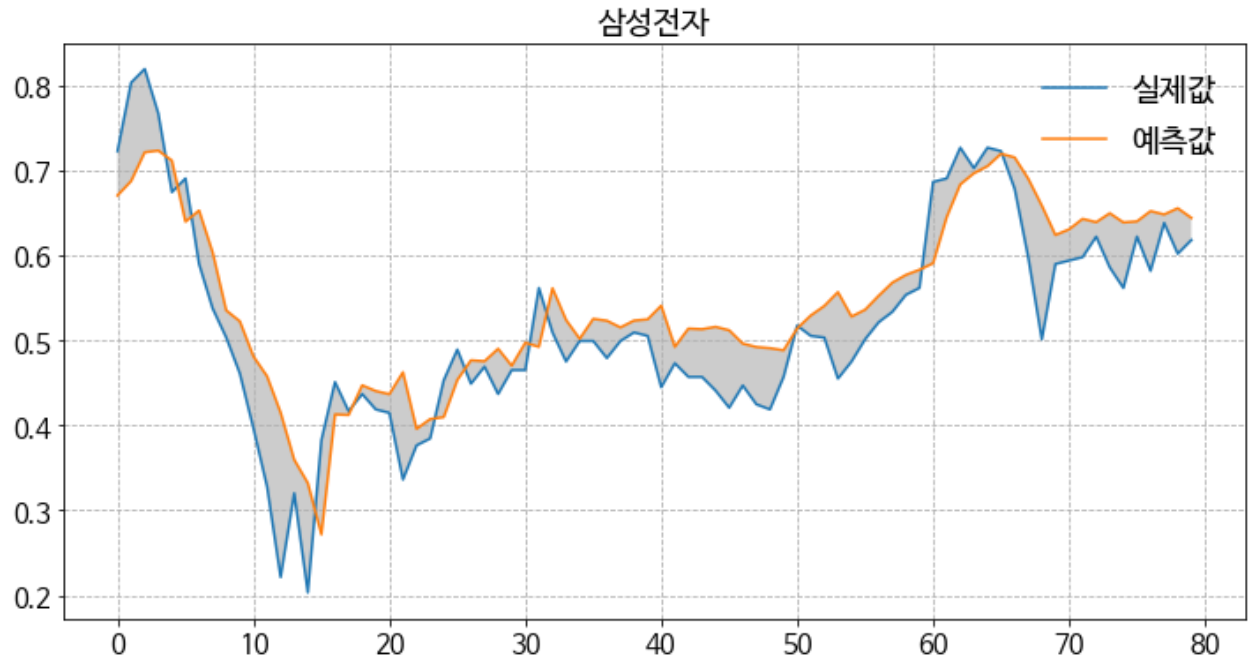
낮은 RMSE score = 성능이 좋은 모델?

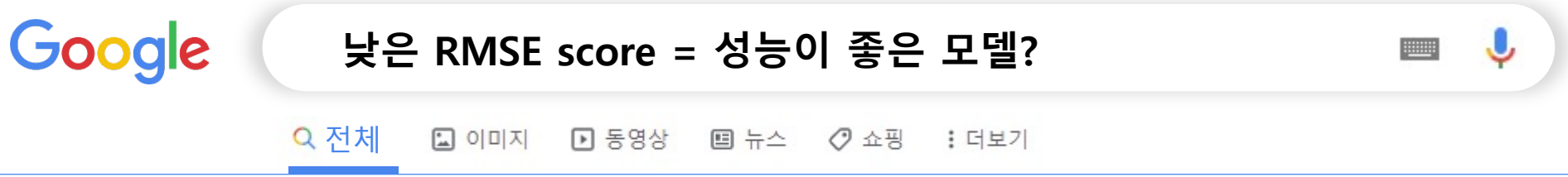
전체 이미지 동영상 뉴스 쇼핑 더보기

- MSE나 RMSE와 같은 함수들은 실제 값과 예측 값사이의 면적(오른쪽 그래프의 회색영역)만을 측정

- 모델이 실제로 예측을 올바르게 하고 있는지 알아보기 위하여 다음의 두가지를 더 분석해 보았음

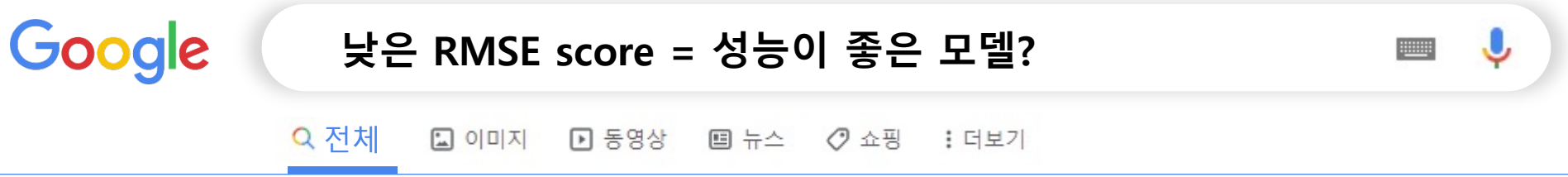
1. RMSE는 얼마나 낮아야 하는가?
2. 오를 때 오른다고 예측하고 내릴 때 내린다고 예측하는가?





RMSE는 얼마나 낮아야 할까?

- 삼성전자 종가의 1일 평균 변화량 : 2340.92 원
- 삼성전자 종가의 1일 평균 변화량(scaled) : **0.004468**
- 2017년 신동하[2] 등의 논문에서 RMSE가 평균적으로 소수점 2자리 수 이하의 값을 가졌다. (우리의 프로젝트와 세부 feature는 다르나 데이터 정규화 방법, 사용모델이 유사하고 1일 후 종가를 예측한다는 점은 같았다)

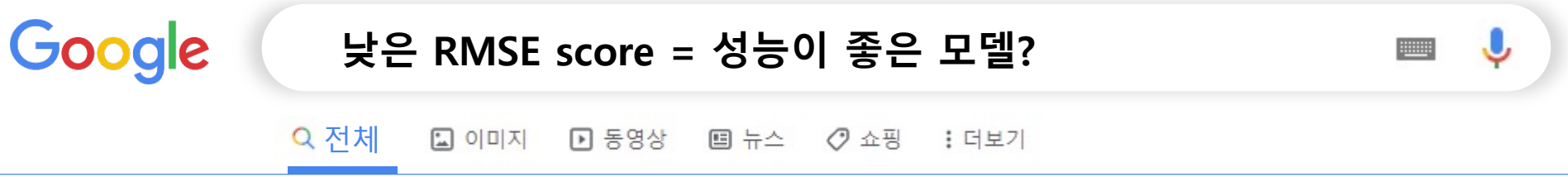


투자 전략을 수립할 수 있는 최소한이 어느 정도일까?를 생각해 본다면...

- 입력 받은 20개의 값 중 마지막 값을 그대로 예측으로 반환하는 모델(이하 Naïve model)작성
- 우리가 만든 GRU 모델과 Naïve model의 성능 비교
- Feature1 : 앞서 이야기한 15개의 feature
- Featrue2 : feature1 + 5일, 10일, 20일, 60일 이동평균
- Feature3 : 종가
- Feature4 : Feature3 + 5일, 10일, 20일, 60일 이동평균

RMSE score

Model	Feature1	Feature2	Feature3	Feature4
GRU	0.0069	0.0039	0.0026	0.0027
Naïve model	0.0028	0.0028	0.0026	0.0026

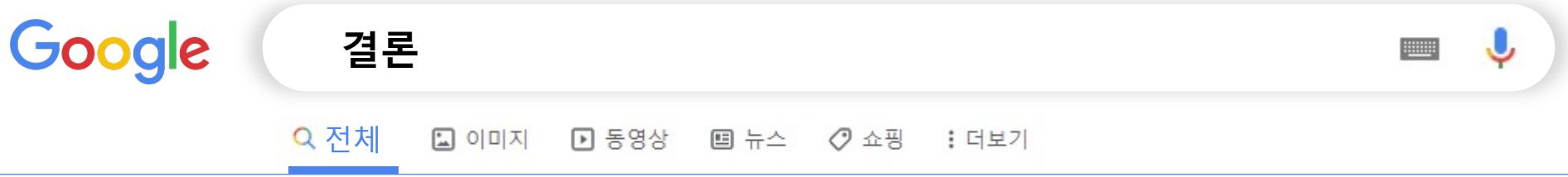


예측의 방향성은 맞을까?

- 전날과 비교하여 당일의 종가가 상승하였으면 +1, 하락하였으면 -1, 변화가 없었으면 0으로 데이터를 인코딩
- 이 후 테스트 데이터와 예측 데이터가 얼마나 일치하는지 계산
- 주가가 상승한날과 하락한 날이 절대 다수였으며 변화가 없는 날은 3% 미만이었다(즉, 이진 분류에 가까움)
- Feature는 앞서 설명한 것과 동일

Accuarcy

Model	Feature1	Feature2	Feature3	Feature4
GRU	0.5375	0.5375	0.5125	0.475



시계열 예측 방식

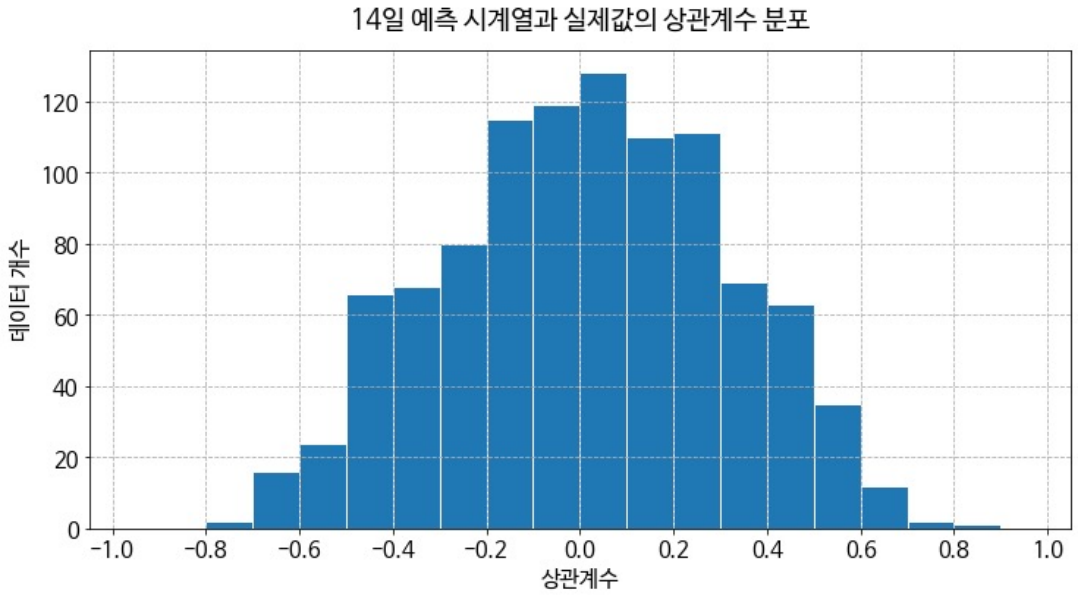
- 입력 시계열 마지막 값을 거의 그대로 출력하는 경향이 강하게 나타남
- 정확도 또한 50%에 가까웠고 이는 랜덤하게 찍는 것과 같은 수치임
- 여러 특성변수들을 추가하고 모델을 복잡하게 하는 등의 시도를 하였으나 과적합 발생

Google 낮은 RMSE score = 성능이 좋은 모델?

전체 이미지 동영상 뉴스 쇼핑 더보기

예측의 방향성은 맞을까?

- 1일 예측이 아니라 그 이상의 기간을 예측하는 모델에 대해서도 실험
- 60일 데이터를 입력 받아 14일 예측을 출력하는 모델을 학습
- 이후 실제 시계열 데이터와 예측 시계열 데이터의 상관계수를 구한 후 분포도 시각화
- 상관계수들이 0을 기준으로 정규분포하는 모습을 보임
- 1일 예측이나 14일 예측이나 주가 시계열의 방향성 또한 예측하지 못하고 있음을 알 수 있다.

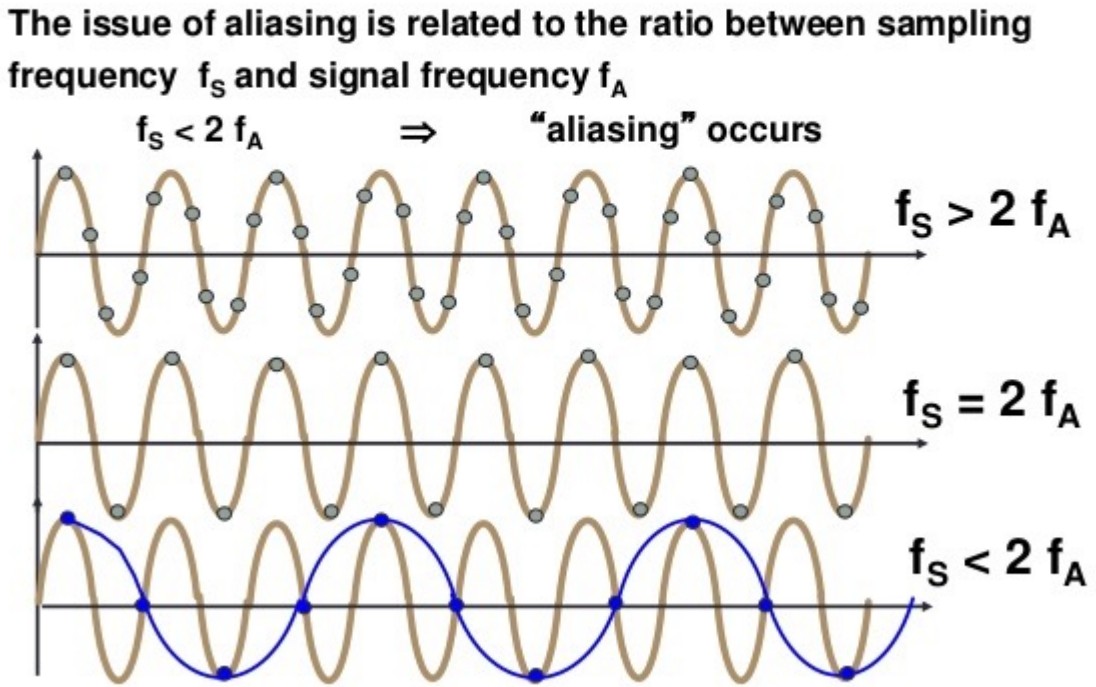


Google 실패요인

전체 이미지 동영상 뉴스 쇼핑 더보기

생각해 볼 수 있는 실패요인

- 단기예측을 함에도 주가 데이터의 샘플링 주기가 길어 관측 가능한 패턴이 존재하더라도 잡음처럼 작용 (노이즈>정보)
- 위의 이유와 딥러닝 모델의 높은 표현력이 결합하여 심각한 과적합 발생
- 추가적인 특성변수(column)를 수집하여도 변수 당 수집 가능한 데이터 일 수(row)가 한정되어 있음 → 차원의 저주로 이어짐



Google 보완할 점

전체 이미지 동영상 뉴스 쇼핑 더보기

한정적인 데이터

▶데이터 기간 늘리기

▶하루 단위가 아닌 분, 시간 단위의 데이터 수집

과거와 현재 데이터의 weight를 같게 함

▶주가는 먼 과거보다는 최신 데이터 값이 더 중요

▶가장 최근의 데이터를 얻어 학습하거나 weight를 달리 줄 수 있음

코로나 전후 차이

▶코로나 발생 전과 후로 나누어 학습할 수 있음

▶코로나 이후 삼성전자와 상관관계가 높아진 특성도 있으나 그렇지 않은 특성도 존재하므로 상관관계가 높은 특성 선정(주가에 선행되어 상관관계가 나타나는 특성 탐색)

Google

References

전체 이미지 동영상 뉴스 쇼핑 더보기

1. Kyunghyun et al. “Learning Phrase Represntations Using RNN Encoder-Decoder for Statistical Machine Translation”, Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (2014):1724-1734. https://arxiv.org/abs/1406.1078

2. Dong-Ha Shin et al. “Deep Learning Model for Prediction Rate Improvement of Stock Price Using RNN and LSTM”, Journal of KIIT, Vol. 15, No. 10, pp. 9-16 (2017)