LSTM 모형을 이용한 KOSPI 종가 예측

Prediction of KOSPI closing prices with LSTM model

한국외국어대학교

[redacted] 학과

[redacted]

[redacted]

2023年 2月

LSTM 모형을 이용한 KOSPI 종가 예측

Prediction of KOSPI closing prices with LSTM model

위 논문 학사학위 논문으로 제출합니다.

지도교수: [redacted]

2023年 2月

대학 : 한국외국어대학교

학과 : [redacted]

학번 : [redacted]

이름 : [redacted]

장지욱의 학사학위 논문을 심사하여
합격으로 판정합니다.

심사위원: [redacted]

LSTM 모형을 이용한 KOSPI 종가 예측

Prediction of KOSPI closing prices with LSTM model

요약

Abstract

금융 분야에서는 과거부터 주가 예측이 도전 과제로서 꾸준히 연구가 진행됐다. 특히 최근 인공지능의 발전으로 딥러닝 기술을 주가 예측에 적용하기 시작했고 실제로 기존의 통계 방법보다 훨씬 더 좋은 성능을 보였다. 본 논문에서는 주가 예측 모델로서 Long Short Term Memory(LSTM) 딥러닝 모델을 적용하여 13개 거시경제지표, 심리지표를 활용해 코스피 종가를 예측하는 방법을 제안하도록 한다. 총 2002년 1월부터 2021년 12월까지 약 20년 동안의 데이터를 수집하였고, Inner join, Minmaxscaler 모듈을 활용해 데이터 전처리를 하였다. 이번 연구에서는 시계열 데이터에 특화된 LSTM 모델을 이용하여 한국 대표 주가지수인 KOSPI를 예측한다는 측면에 의의를 두고 연구를 수행하였다. 향후에는 뉴스 정보의 감성분석이나 페이스북과 같은 비정형 데이터를 모델에 추가하거나, 하이퍼파라미터에 관한 최적화 연구를 수행하고자 한다.

목차

1. 서론

세계적인 긴축 기조와 빠른 기준금리 상승 등으로 금융 시장은 흐린 나날을 보내고 있다. 금융 시장에서 가장 두려워하는 요소 중 하나는 바로 불확실성이다. 불확실성이 높으면 높을수록 시장에 존재하는 신뢰는 떨어지고, 그만큼 변동성도 커지게 된다. 이 중 가장 예측하기 어려운 부분이 바로 주가이다. 주가에 대한 예측은 다양하고 수많은 변수의 상호작용으로 인해 예측이 매우 어렵다고 알려진 분야임에도 불구하고 꾸준히 시도되었다. 주가는 이론적으로 미래에 발생할 현금흐름에 대한 기댓값이지만, 그 외에도 거시 경제적인 효과에 영향을 크게 받는다. 예를 들어 환율과 주가에 대한 많은 분석[1]이 있으며, 실제로 환율은 주가에 영향을 준다. 또한, 국제 유가나 원자재 등 여러 거시 경제적 변수[2]는 국내 시장에 영향을 주며, 기업의 가치, 주식의 가치에 영향을 준다.

과거에는 전통적인 통계 기법에 기초한 방법을 많이 적용하였으나[3], 그렇게 효과적이지는 않았다. 하지만 최근 빅데이터와 인공지능이 발달하면서 새로운 기술을 이용한 주가 예측 방법이 많아지고 있고 사람들의 관심을 받고 있다. SVM과 kNN과 같은 기계학습 모델부터 MLP, CNN, RNN과 같은 딥러닝 모델까지 주가를 예측하기 위한 여러 연구가 활발히 진행 중이다. 본 연구에서는 딥러닝 모델 LSTM(Long Short Term Memory)을 이용하여 과거 20일간의 13개 거시 경제 변수를 이용해 그다음 날의 KOSPI 종가를 예측하는 방식으로 구현할 것이다. 본 보고서는 시계열 데이터에 특화된 LSTM 모델을 이용하여 주가를 예측한다는 측면에서 의의를 두고 연구를 수행하였다.

2. 데이터

2.1 데이터 구성 및 수집 과정 설명

KOSPI 종가를 예측하기 위하여 설정한 Input 거시 경제 변수는 총 13개를 사용하였고, 아래 [표 1]에서도 볼 수 있듯, KOSPI 종가 데이터, VIX 지수, S&P500, 다우 존스 지수, 4가지의 환율 (달러/원화, 유로/원화, 위안/원화, 옌/원화), WTI 유가 선물, 1년 만기 한국 국채 수익률, 10년 만기 한국 국채 수익률, 1년 만기 미국 국채 수익률, 10년 만기 미국 국채 수익률이다. 데이터를 선정한 근거는 각 데이터 간의 상관관계를 고려하여 최대한 데이터의 결과가 편중되지 않도록 낮은 상관관계를 갖는 데이터도 포함했다. 특히 VIX 지수는 소위 공포 지수라고도 하며, 투자자들이 투자 결정을 내리기 전에 시장 리스크, 공포 및 스트레스를 측정하는 데 사용한다. 주가지수뿐만 아니라 심리지수도 포함해서 최대한 주가 예측에 다양

1) 최완수. "환율과 주가변동성의 상호연관성 분석." 대한경영학회 학술발표대회 발표논문집 2016.1 (2016): 786-804.
2) 김재일,and 김주일. "우리나라 주가와 거시경제변수들 간의 상호연관성에 관한 연구." 專門經營人硏究 17.3 (2014): 163-186.
3) 황선영,and 김은주. "TAR-GARCH 모형을 이용한 국내 주가 자료 분석." 응용통계연구 13.2 (2000): 437-445.

한 요소를 추가하였다. 데이터는 FinanceDataReader 모듈을 사용하여 일별 데이터를 가져왔다. 일별 데이터는 DataFrame 형식으로 'Open', 'High', 'Low', 'Volume', 'Change' 총 6개 데이터를 받아온다. 그중 종가 데이터인 'Close' 데이터만 따로 추출하여 사용하였다. 데이터 수집의 총기간은 2002년 1월부터 2021년 12월까지 약 20년 정도이다. 최대한 오랜 기간을 사용하려 했고, 13개의 변수를 모두 공통으로 수집할 수 있는 시작 연도가 2002년이어서, 2002년부터 시작하였다.

번호	변수명	번호	변수명
1	KOSPI 종가	8	환율 (달러($)/원화(₩))
2	VIX 지수	9	환율 (유로(€)/원화(₩)
3	S&P500	10	환율 (위안(CNY￥)/원화(₩)
4	DJI 지수	11	환율 (엔(JPY￥)/원화(₩)
5	WTI 유가 선물	12	1년 만기 미국 국채 수익률
6	1년 만기 한국 국채 수익률	13	10년 만기 미국 국채 수익률
7	10년 만기 한국 국채 수익률		

[표 1] 데이터 구성

2.2 데이터 전처리 과정

데이터 전처리는 크게 두 가지 과정으로 나뉜다. 첫 번째는 Inner Join을 이용하여 각 데이터의 날짜 및 시간을 일치시켜줬고, 두 번째는 'Minmaxscaler' 모듈을 활용해 각 데이터의 단위를 제거해줬다.

우선 데이터마다 공시되는 날짜가 조금씩 차이가 있다. 예를 들어 한국 증시와 미국 증시의 공휴일에 따라 서로 시장이 열리는 날짜가 다르다. 따라서 이런 문제를 해결해주기 위해 같은 날짜의 데이터만 사용하였다. 또한, 시차를 고려하여 KOSPI 종가 데이터는 하루 전의 데이터를 사용하였다. 변수마다 차이는 존재하지만, 각 4,000~5,000개 정도의 데이터가 모였다. 데이터 전처리는 INNER JOIN 방법을 이용하였다. INNER JOIN은 관계형 데이터베이스에서 테이블 처리 방식 중 하나로, 쉽게 얘기해서 테이블 간의 교집합을 사용하는 것이다. 각 모든 데이터가 'Date' 형태의 index를 갖도록 설정을 했기 때문에 13개 변수의 테이블 간 교집합을 하였다. 아무래도 13개 변수마다 교집합을 하였기 때문에, 각 데이터를 추출할 때와는 다르게 데이터의 양이 좀 줄었다. Inner Join을 통해 총확보한 데이터는 3,928개이다. 데이터의 특성은 아래 [그림 1]과 같다.

	kospi	vix	sp500	dji	usdkrw	eurkrw	cnykrw	jpykrw	wti	kr1yt	kr10yt	us1mt	us10yt
count	3928.000000	3928.000000	3928.000000	3928.000000	3928.000000	3928.000000	3928.000000	3928.000000	3928.000000	3928.000000	3928.000000	3928.000000	3928.000000
mean	1902.188142	18.490652	1978.142391	17338.626624	1105.171589	1373.197709	162.141790	10.678101	53.989404	2.704189	3.437928	1.196978	2.793172
std	518.622695	8.285237	896.380413	7163.387315	92.570722	141.973370	21.958165	1.828288	18.401414	1.383483	1.437497	1.567625	1.153609
min	541.780000	9.140000	682.550000	6594.440000	900.800000	1135.510000	115.870000	7.456200	21.530000	0.000000	0.000000	-0.041000	0.512000
25%	1637.955000	13.057500	1275.572500	11400.827500	1055.452500	1273.060000	152.415000	9.577225	36.595000	1.535000	2.153750	0.043000	1.917750
50%	1964.730000	16.070000	1696.450000	15451.430000	1118.905000	1338.030000	169.340000	10.293050	60.195000	2.650000	3.175000	0.204000	2.574000
75%	2108.995000	21.320000	2553.687500	23002.800000	1160.025000	1448.180000	176.182500	11.295075	69.792500	3.590000	4.900000	1.995250	3.721250
max	3305.210000	82.690000	4793.060000	36488.630000	1570.650000	1976.350000	229.490000	16.309900	86.260000	5.810000	7.270000	5.278000	5.289000

[그림 1] 데이터 특성

LSTM 모델 학습시키기 위해서는 데이터가 벡터 형태가 되어야 하므로 추가적인 처리 작업을 해야 한다. 딥러닝에서 학습은 입출력 데이터의 크기가 작으면 작을수록 더 효과적이기 때문에, 보편적으로 사용되는 정규화 기법을 사용하여, 모든 입출력 데이터를 0~1의 구간으로 범위를 같게 만든다.[4] 따라서 데이터마다 단위와 값의 크기를 같게 만들어주기 위해서 이번 연구에서는 Minmaxscaler 모듈을 이용하여 데이터 스케일링을 진행했다. Minmaxscaler는 최솟값과 최댓값을 이용하여 데이터값을 0과 1 사이의 범위 값으로 변환하는 작업[5]이다. 아래 [그림 2]는 Minmaxscaling을 진행하고 난 뒤의 데이터 모습이다. 예를 들어, KOSPI 종가는 최솟값 541, 최댓값으로 3,305를 갖는다. 따라서 Minmaxscaler을 적용하게 되면, 541은 0으로, 3,305는 1이 되고, 나머지 값도 이에 맞춰 변경된다.

id	kospi	vix	sp500	dji	usdkrw	eurkrw	cnykrw	jpykrw	wti	kr1yt	kr10yt	us1mt	us10yt
0	0.072678	0.173215	0.108417	0.104921	0.642233	0.039139	0.395617	0.276144	0.102271	0.870912	0.984869	0.325437	0.945363
1	0.086291	0.163154	0.109609	0.109430	0.620288	0.000000	0.379951	0.271073	0.103816	0.867470	0.986245	0.329197	0.955202
2	0.071697	0.221754	0.099124	0.103398	0.613570	0.003259	0.375198	0.264206	0.099645	0.867470	0.964237	0.334085	0.917731
3	0.072291	0.230455	0.097543	0.102326	0.613421	0.003080	0.375022	0.265697	0.095010	0.864028	0.965612	0.333145	0.924220
4	0.087026	0.207886	0.097494	0.105395	0.619840	0.021526	0.379511	0.270158	0.096864	0.858864	0.962861	0.332957	0.912497
5	0.088962	0.199320	0.096934	0.108390	0.616556	0.008682	0.377310	0.263664	0.099954	0.853701	0.955983	0.337845	0.909985

[그림 2] Minmaxscaler 적용 후 데이터 모습

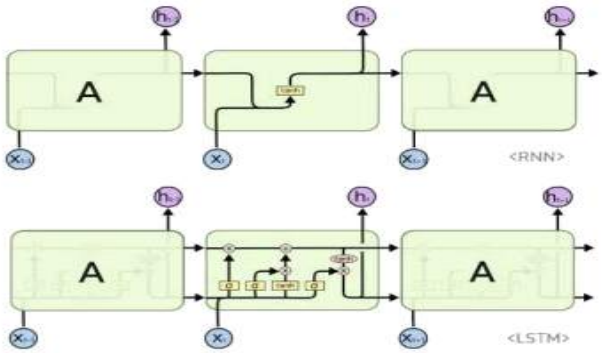
그 후 데이터 세트를 학습용과 검증용으로 나누는 작업이 필요하다. 일반적으로 전체 데이터 중 70~80%를 학습용 데이터로, 20~30%를 검증용 데이터로 사용하는 것이 좋다.[6] 따라서 이번 연구에서는 전체 데이터 3,928개를 80:20 비율을 적용해 훈련용 데이터와 검증용 데이터로 나누어 적용하였다. 그 결과, 훈련용 데이터는 총 3,142개로 약 2002년~2018년의 기간을, 검증용 데이터는 총 786개로 약 2019년~2021년의 기간을 가지고 있다.

4) 신동하(Dong-Ha Shin),최광호(Kwang-Ho Choi),and 김창복(Chang-Bok Kim). "RNN과 LSTM을 이용한 주가 예측율 향상을 위한 딥러닝 모델." 한국정보기술학회논문지 15.10 (2017): 9-16.
5) 임소영, 서호건, 이항로, 김진섭. (2022). 다채널 음향방출 신호에 대한 딥러닝을 통한 3차원 위치 표정. 비파괴검사학회지, 42(1), 34-42.
6) 박현정(Hyeonjung Park),최승배(Seungbae Choi),and 강창완(Changwan Kang). "시계열 예측을 위한 인공신경망의 입력노드 수 결정." Journal of the Korean Data Analysis Society 22.3 (2020): 1055-1065.

3. 분석 방법론

3.1 딥러닝 알고리즘 설명

신경회로망(Artifical Neural Network, ANN)은 인간의 정보처리 패턴과 유사한 방식의 시스템으로 비선형적 문제와 규칙성 없는 문제를 해결하고 방대한 데이터양의 자료를 처리할 수 있어 보다 보편화된 예측모형으로 연구되고 있다.[7] 그중에서도 Long Short-Term memory(LSTM) 모델은 순경 신경망 (RNN)의 기존 단점을 보완하여 장단기 기억을 가능하게 설계한 신경망이다. 아래 [그림 3]은 RNN과 LSTM의 구조를 시각화한 것이다. RNN은 층이 깊어질수록 gradient가 줄어들어 학습이 불가능해지는 기울기 소실 문제와 데이터의 기간이 길어질수록 과거의 정보가 전달되지 못하는 장기 의존성 문제가 있다.[8] 그에 반해, LSTM은 망각 게이트(forget gate), 입력 게이트(input gate), 출력 게이트(output gate)로 구성된 셀(cell)을 통해서 과거 정보를 RNN보다 더 잘 간직한다. 간단하게 설명하자면, 입력 게이트는 현재 정보를 기억하기 위한 게이트, 망각 게이트는 과거 정보를 잊기 위한 게이트, 출력 게이트는 현재 정보를 다음 셀로 전달하기 위한 게이트이다. 셀은 LSTM의 전체 체인을 관통하여 과거 학습 결과를 변함없이 전달하는 구조를 가져 장기 의존성 문제를 해결한다.



[그림 3] RNN과 LSTM 구조도

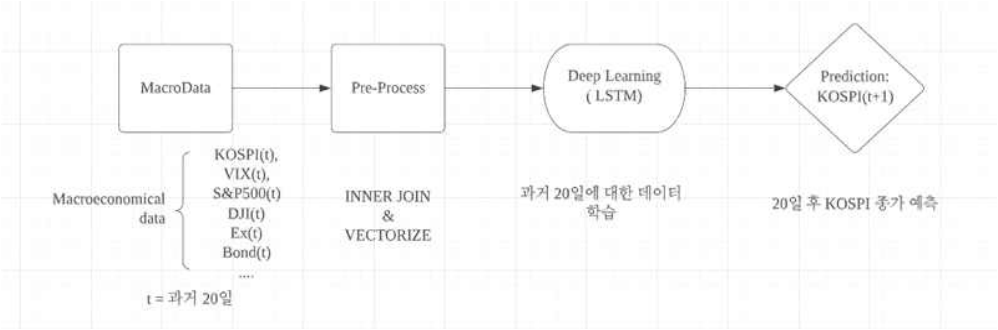
3.2 모델 구조 및 학습 방법

본 연구의 모델 구조와 학습 방법을 소개하기 전에 이번 KOSPI 종가 예측에 대한 전체적인 구조를 소개하고자 한다. 아래 [그림 4]에서 볼 수 있듯, 13개의 거시경제지표에 대한 데이터 수집을 진행하고 난 후, 데이터 전처리를 통해 딥러닝 모델에 들어갈 수 있는 형태로 변경해준다. 자세한 과정은 앞서 '2. 데이터 전처리 과정'에 있는 바와 같다. 데이터 전처리가 끝나면 학습용 데이터셋을 활용해 추후 있을 KOSPI 종가 예측을 위해 LSTM 모델을 학습시킨다.

---

7) 이은진(Eun Jin Lee),민철홍(Chul Hong Min),and 김태선(Tae Seon Kim). "신경 회로망과 통계적 기법을 이용한 종합주가지수 예측 모형의 개발." 電子工學會論文誌-CI (Computer and Information) 45.5 (2008): 95-101.

8) 김하얀(Kim Ha Yan),주궈화(Guohua Zhu),and 김석찬(Suk Chan Kim). "LSTM 을 이용한 주가 예측 알고리즘." 한국통신학회 학술대회논문집 2021.11 (2021): 1019-1020.



[그림 4] Model Pipe Line

본 연구의 목적은 앞서 언급한 바와 같이 KOSPI 종가에 영향을 주는 거시경제 지표인 주가지수, 채권, 유가, 심리지수 및 4가지 환율을 활용해 KOSPI 종가를 예측하는 것이다. 주식, 채권과 같은 전통적인 자산군 관련 지표뿐만 아니라 투자자의 관심을 나타내는 VIX 지수와 같은 시장 심리 지표, 파생 상품과 환율도 포함해 KOSPI 예측을 진행하기에 더 정확한 예측이 나올 것으로 기대한다. 이번 연구에서 작성한 함수를 식으로 표현하면 아래 [수식 1]과 같다.

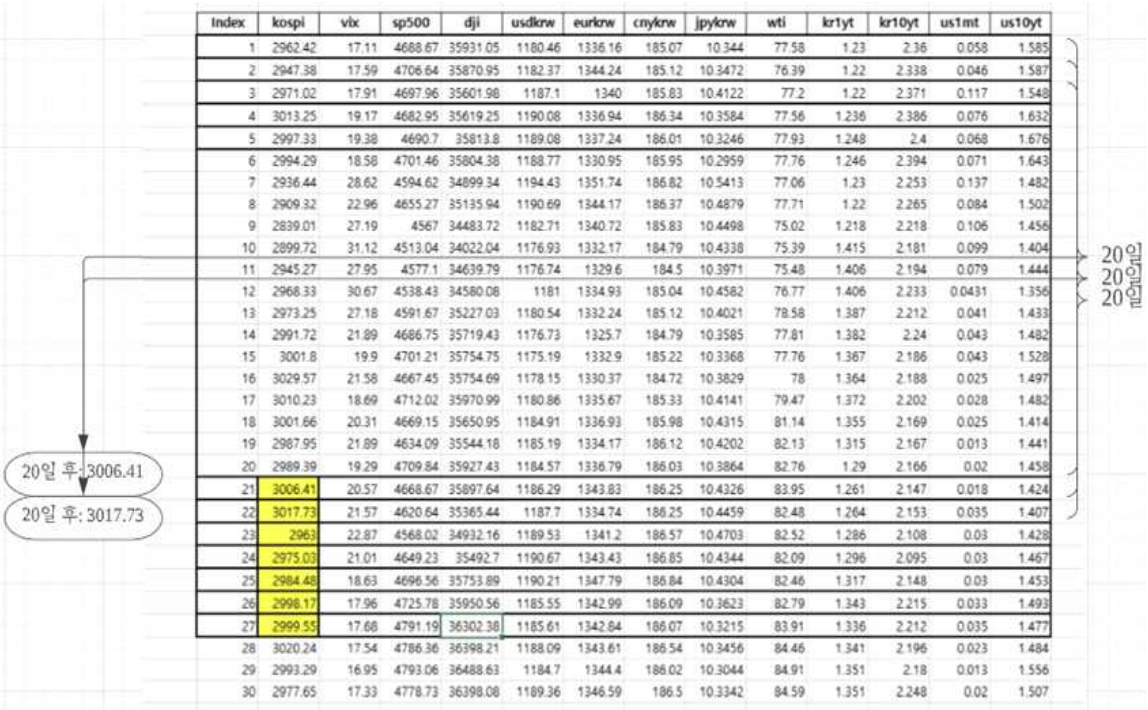
$$KOSPI_{t+1\,(t\geq 20)} \approx \sum_{k=t-19}^{t} (stock_k + VIX_k + exchange_k + oil_k + bonds_k)$$

[수식 1]

위의 [수식 1]에 대하여 설명하자면, $KOSPI_t$ 는 코스피의 t 날짜의 종가를 의미하고, 종속변수이다. [수식 1] 오른편에 있는 식은 과거 20일 치에 대한 합이다. $stock_k$은 코스피 종가, 다우 존스 종가, S&P500 종가를 의미한다. VIX_k는 앞서 얘기한 바와 같이 시장 공포 지수를 의미하고 $exchange_k$는 한국/미국, 한국/중국, 한국/일본, 한국/유럽 환율을 의미한다. oil_k는 다양한 유가 중 미국의 가장 대표적인 WTI (West Texas Intermediate)를 사용하였다. 마지막으로 $bonds_k$는 한국 장·단기 국채 수익률, 미국 장·단기 국채 수익률을 의미한다.

이해를 돕기 위해서 [그림 5]와 함께 데이터 학습 방법에 대하여 설명한다. KOSPI 종가에 대한 예측은 과거 20일의 13개 변수 (KOSPI 종가, VIX 지수, S&P500, DJI, 환율, 채권) (20 x 13)의 데이터를 학습하여 이뤄진다. Window size를 20으로 하였기 때문에, 20개씩 하루씩 미뤄가며 쌍을 만들어 나아간다. Index 1~20번까지의 13개의 변수를 이용해서 21번째의 KOSPI 종가를 예측하게 된다. 또 Index 2~21번까지의 13개의 변수를 이용해서 22번째의 KOSPI 종가를 예측하게 된다. 이렇게 쭉 가서 index 12~31번까지의 13개 변수를 이용해서 32번째 KOSPI 종가를 예측한다. 우선 80%의 학습 데이터는 실제 코스피 종가에 대한 정보를 가지고 있으므로 예측값을 실제값과 최대한 비슷

하게 하는 방식으로 파라미터를 학습시킨다. 그리고 지나치게 학습 데이터가 많거나 편향되는 것을 막기 위해 Early-stopping 기법을 도입했다. 이는 특정 epoch(반복 횟수) 동안 모델이 개선되지 않으면 학습을 멈추는 것인데, 이번 실험에서는 10번 epoch 동안 모델이 개선되지 않으면 멈추도록 설정하였다.



Index	kospi	vix	sp500	dji	usdkrw	eurkrw	cnykrw	jpykrw	wti	kr1yt	kr10yt	us1mt	us10yt
1	2962.42	17.11	4688.67	35931.05	1180.46	1336.16	185.07	10.344	77.58	1.23	2.36	0.058	1.585
2	2947.38	17.59	4706.64	35870.95	1182.37	1344.24	185.12	10.3472	76.39	1.22	2.338	0.046	1.587
3	2971.02	17.91	4697.96	35601.98	1187.1	1340	185.83	10.4122	77.2	1.22	2.371	0.117	1.548
4	3013.25	19.17	4682.95	35619.25	1190.08	1336.94	186.34	10.3584	77.56	1.236	2.386	0.076	1.632
5	2997.33	19.38	4690.7	35813.8	1189.08	1337.24	186.01	10.3246	77.93	1.248	2.4	0.068	1.676
6	2994.29	18.58	4701.46	35804.38	1188.77	1330.95	185.95	10.2959	77.76	1.246	2.394	0.071	1.643
7	2936.44	28.62	4594.62	34899.34	1194.43	1351.74	186.82	10.5413	77.06	1.23	2.253	0.137	1.482
8	2909.32	22.96	4655.27	35135.94	1190.69	1344.17	186.37	10.4879	77.71	1.22	2.265	0.084	1.502
9	2839.01	27.19	4567	34483.72	1182.71	1340.72	185.83	10.4498	75.02	1.218	2.218	0.106	1.456
10	2899.72	31.12	4513.04	34022.04	1176.93	1332.17	184.79	10.4338	75.39	1.415	2.181	0.099	1.404
11	2945.27	27.95	4577.1	34639.79	1176.74	1329.6	184.5	10.3971	75.48	1.406	2.194	0.079	1.444
12	2968.33	30.67	4538.43	34580.08	1181	1334.93	185.04	10.4582	76.77	1.406	2.233	0.0431	1.356
13	2973.25	27.18	4591.67	35227.03	1180.54	1332.24	185.12	10.4021	78.58	1.387	2.212	0.041	1.433
14	2991.72	21.89	4686.75	35719.43	1176.73	1325.7	184.79	10.3585	77.81	1.382	2.24	0.043	1.482
15	3001.8	19.9	4701.21	35754.75	1175.19	1332.9	185.22	10.3368	77.76	1.367	2.186	0.043	1.528
16	3029.57	21.58	4667.45	35754.69	1178.15	1330.37	184.72	10.3829	78	1.364	2.188	0.025	1.497
17	3010.23	18.69	4712.02	35970.99	1180.86	1335.67	185.33	10.4141	79.47	1.372	2.202	0.028	1.482
18	3001.66	20.31	4669.15	35650.95	1184.91	1336.93	185.98	10.4315	81.14	1.355	2.169	0.025	1.414
19	2987.95	21.89	4634.09	35544.18	1185.19	1334.17	186.12	10.4202	82.13	1.315	2.167	0.013	1.441
20	2989.39	19.29	4709.84	35927.43	1184.57	1336.79	186.03	10.3864	82.76	1.29	2.166	0.02	1.458
21	3006.41	20.57	4668.67	35897.64	1186.29	1343.83	186.25	10.4326	83.95	1.261	2.147	0.018	1.424
22	3017.73	21.57	4620.64	35365.44	1187.7	1334.74	186.25	10.4459	82.48	1.264	2.153	0.035	1.407
23	2963	22.87	4568.02	34932.16	1189.53	1341.2	186.57	10.4703	82.52	1.286	2.108	0.03	1.428
24	2975.03	21.01	4649.23	35492.7	1190.67	1343.43	186.85	10.4344	82.09	1.296	2.095	0.03	1.467
25	2984.48	18.63	4696.56	35753.89	1190.21	1347.79	186.84	10.4304	82.46	1.317	2.148	0.03	1.453
26	2998.17	17.96	4725.78	35950.56	1185.55	1342.99	186.09	10.3623	82.79	1.343	2.215	0.033	1.493
27	2999.55	17.68	4791.19	36302.38	1185.61	1342.84	186.07	10.3215	83.91	1.336	2.212	0.035	1.477
28	3020.24	17.54	4786.36	36398.21	1188.09	1343.61	186.54	10.3456	84.46	1.341	2.196	0.023	1.484
29	2993.29	16.95	4793.06	36488.63	1184.7	1344.4	186.02	10.3044	84.91	1.351	2.18	0.013	1.556
30	2977.65	17.33	4778.73	36398.08	1189.36	1346.59	186.5	10.3342	84.59	1.351	2.248	0.02	1.507

[그림 5] 데이터 형태, Window size = 20

4. 결과 및 분석

4.1 결과 및 해석

주가 예측의 정확성은 예측값과 실제값 차이에서 오는 오차값을 통해서 확인할 수 있다. 정확성을 평가하는 척도는 절댓값 오차, 평균 제곱 오차(MSE) 등 여러 가지 방법이 있으나, 본 연구에서는 보편적으로 많이 사용되는 평균 제곱근 편차(Root Mean Square Error, RMSE)로 주가 예측의 정확성을 평가한다. 평균 제곱근 오차는 아래 식에서도 볼 수 있듯, 오차의 제곱에 대한 평균을 취하고 이를 제곱근 한 것이다[9]. RMSE는 값이 작을수록 예측값과 실제값의 차이가 없음을 의미하기 때문에 낮으면 낮을수록 좋다.

$$\mathrm{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i - \hat{x}_i)^2}$$

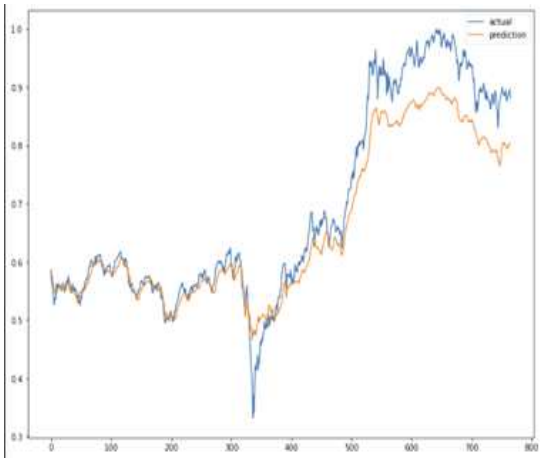
9) 네이버 지식백과, https://terms.naver.com/entry.naver?docId=3481948&cid=58439&categoryId=58439, 2022. 10. 25. 검색

실험	RMSE	비고
1	0.053893530193	
2	0.044681855339	Earlystopping:37/50
3	0.030899501976	
4	0.038293040192	
5	0.040293810020	
평균	0.041612348	

[표 2] 실험 결과 (RMSE)

위 [표 2]는 실험 5번을 진행할 때 위에서 언급한 LSTM 모델로 예측한 RMSE(Root Mean Square Error)를 나타낸 것이고, 평균적으로 0.0416의 RMSE 값을 갖는다. 아래 [그림 4] ~ [그림 8]은 5번의 실험 동안 예측값 (주황색)과 실제값 (파란색)을 도식화하여 표현했다. RMSE가 낮은 그림의 주황색 선과 파란색 선이 유사하게 흘러가는 것을 볼 수 있다. 예를 들어, RMSE가 가장 낮은 [그림 8] 실험 3의 경우 거의 모든 구간에서 예측값과 실제값이 일치되나, RMSE가 가장 높은 [그림 6] 실험 1의 경우는 특히 후반에 오차가 심해지는 것을 볼 수 있다.

[그림 6] 실험 1, RMSE:0.0539



[그림 7] 실험 2, RMSE:0.0447

[그림 8] 실험 3, RMSE:0.0309

[그림 9] 실험 4, RMSE:0.0383

[그림 10] 실험 5, RMSE:0.0403

추가로 5번의 실험 모두에서 공통으로 나타나는 특징을 발견했다. 이는 x축의 값이 500(2020년쯤) 전후로 예측력의 차이가 난다는 사실이다. x값 500전의 경우 실제값과 예측값이 별 차이가 없으나, x값 500 이후부터 예측력이 떨어지는 것을 볼 수 있다. 이에 대한 추가적인 조사를 해보니, 그 시기가 코로나 이후 일례에 없었던 공격적인 재정과 통화정책이 있었다. 통화정책 면에서는 빠른 금리 인하와 자산매입이 이뤄졌고, 재정정책 면에서는 재정확장이 큰 폭을 빠르게 이루며 코로나19 사태에 대해 대응하였다. 이는 인공지능도 예측하기 어려웠던 특별한 사건으로 간주하여 예측력이 떨어진 것으로 분석한다.

5. 결론 및 한계점

5.1 결론

주가는 수많은 요소에 영향을 받기 때문에 쉽게 예측하기 어렵다. 특히 미래에 어떤 일이 일어날지 모르고, 단순한 재정적 요소에만 영향을 받지 않기 때문이다. 본 보고서는 거시 경제 지표를 학습시킨 인공지능이 얼마나 주가를 잘 예측하는지를 연구해봤다. 사용한 거시경제지표는 여러 연구에서 입증된 주가, 채권, 유가, 환율을 적용하였으며, 특히 환율은 한국 경제에 영향을 미치는 한국/미국, 한국/중국, 한국/일본, 한국/유럽을 사용하였다. 13개의 거시경제지표를 시계열 데이터에 적합한 딥러닝 알고리즘인 LSTM에 적용하여 실험하였다. 데이터는 총 2002년 1월부터 2021년 12월까지 총 3,928개의 데이터를 수집했고 80:20의 비율로 학습용과 검증용 데이터로 나눠 LSTM 모델에 학습시켰다. 그 결과 RMSE 값이 평균적으로 0.041612를 갖는다. 추가로, 2020년 이후 급격히 가속화된 양적완화로 인해 인공지능의 예측력이 저하된 패턴도 발견할 수 있었다. 본 연구는 딥러닝 LSTM 모델을 활용하여 KOSPI 주가를 예측해보고 특유 패턴을 발견했다는 것에서 의의가 있다고 생각된다.

5.2 한계점 및 향후 연구 방향

본 논문의 한계점은 첫 번째로 뉴스 기사의 감성 분석 측면을 포함한 알고리즘보다 더 뛰어난 성능을 구현하지 못했다는 것이다. 온라인 뉴스 및 거시 경제 변수를 합쳐 만든 LSTM 모형이 평균적으로 0.012181 RMSE 값을 갖는다[10]. 더 많은 변수와 요소에 대한 고려가 있었더라면 모델의 성능을 좋게 했을 것이다. 또한, 모델에 대한 최적화 작업을 진행하지 못했다. batch size, epoch 값과 같은 하이퍼파라미터는 있는 그대로를 사용하였다.

후속 연구에서는 다른 추가적인 지표를 활용하여 연구할 수 있을 것이다. 예를 들면, 뉴스 정보, 트위터, 페이스북 등의 데이터도 함께 사용하여 모델의 성능을 높일 수 있을 것이다. 또한, 직접 하이퍼파라미터 (batch size, epoch 값)에 대한 fine-tunning 진행해서 더 나은 성능을 보여줄 수 있는 여지가 있다. 마지막으로, 단순히 하나의 모델을 사용하는 것이 아닌 여러 모델을 통합하여 더 좋은 성능을 발휘하는 모델에 관한 연구를 수행하고자 한다.

10) 장은아,최희련,and 이홍철. "BERT를 활용한 뉴스 감성분석과 거시경제지표 조합을 이용한 주가지수 예측." 韓國컴퓨터情報學會論文誌 25.5 (2020): 47-56.

<참고문헌>

[1] 최완수. (2016). 환율과 주가변동성의 상호연관성 분석. 대한경영학회 학술발표대회 발표논문집, (), 786-804.

[2] 김재일,and 김주일. "우리나라 주가와 거시경제변수들 간의 상호연관성에 관한 연구." 專門經營人硏究 17.3 (2014): 163-186.

[3] 황선영,and 김은주. "TAR-GARCH 모형을 이용한 국내 주가 자료 분석." 응용통계연구 13.2 (2000): 437-445.

[4] 신동하(Dong-Ha Shin),최광호(Kwang-Ho Choi),and 김창복(Chang-Bok Kim). "RNN과 LSTM을 이용한 주가 예측율 향상을 위한 딥러닝 모델." 한국정보기술학회논문지 15.10 (2017): 9-16.

[5] 임소영, 서호건, 이항로, 김진섭. (2022). 다채널 음향방출 신호에 대한 딥러닝을 통한 3차원 위치 표정. 비파괴검사학회지, 42(1), 34-42.

[6] 박현정(Hyeonjung Park),최승배(Seungbae Choi),and 강창완(Changwan Kang). "시계열 예측을 위한 인공신경망의 입력노드 수 결정." Journal of the Korean Data Analysis Society 22.3 (2020): 1055-1065.

[7] 이은진(Eun Jin Lee),민철홍(Chul Hong Min),and 김태선(Tae Seon Kim). "신경 회로망과 통계적 기법을 이용한 종합주가지수 예측 모형의 개발." 電子工學會論文誌-CI (Computer and Information) 45.5 (2008): 95-101.

[8] 김하얀(Kim Ha Yan),주궈화(Guohua Zhu),and 김석찬(Suk Chan Kim). "LSTM 을 이용한 주가 예측 알고리즘." 한국통신학회 학술대회논문집 2021.11 (2021): 1019-1020.

[9] 지식백과, https://terms.naver.com/entry.naver?docId=3481948&cid=58439&categoryId=58439, 2022. 10. 25. 검색

[10] 장은아,최회련,and 이홍철. "BERT를 활용한 뉴스 감성분석과 거시경제지표 조합을 이용한 주가지수 예측." 韓國컴퓨터情報學會論文誌 25.5 (2020): 47-56.