LSTM 모형을 이용한 KOSPI 종가 예측

학번: 201703076
전공: Language&Trade 학부
이름: 장지욱

목차

I. 서론 3

II. 데이터 3

i. 데이터 특성 및 수집 과정 설명 3-4

ii. 데이터 전처리 과정 4

III. 분석 방법론 5

i. 딥러닝 알고리즘 설명 5-6

IV. 결과 및 분석 7

i. 결과 및 해석 7-8

V. 결론 및 한계점 9

I. 서론

주가에 대한 예측은 다양하고 수많은 변수들의 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 분석들이 있으며, 실제로 환율은 주가에 영향을 준다. 또한, 국제 유가나 원자재 등 여러 거시 경제적 변수는 국내 시장에 영향을 주며, 기업의 가치, 주식의 가치에 영향을 준다.

최근 빅데이터와 인공지능이 발달되면서 새로운 기술을 이용한 주가 예측방법이 많아지고 있다. 특히 순환 신경망은 주식과 같은 시계열 데이터를 처리하는데 효율적인 구조를 가진다. 서포트 벡터 회귀와 같은 기존의 방법보다 순환 신경망을 이용하여 주가를 예측하는 것이 더 우월한 결과를 도출했다. 본 보고서는 앞서 얘기한 순환 신경망의 하나인 Long Short-Term Memory(LSTM)을 이용하여 과거 20 일 간의 13 개 거시 경제 변수들을 이용해 그 다음날의 KOSPI 종가를 예측하는 방식으로 구현하였고 그 결과와 한계점에 대하여 논의를 할 것이다.

II. 데이터

이번 장에서는 데이터를 어떻게 수집하였는지와 이를 어떻게 처리했는지에 대해 설명한다.

i. 데이터 특성 및 수집 과정 설명

거시 경제 변수는 총 13 개를 사용하였고, 이는 KOSPI 종가 데이터, VIX 지수, S&P500, 다우 존스 지수, 4 가지의 환율 (달러/원화, 유로/원화, 위안/원화, 엔/원화), WTI 유가 선물, 1 년 만기 한국국채 수익률, 10 년 만기 한국국채 수익률, 1 년 만기 미국국채 수익률, 10 년 만기 미국국채 수익률이다. 데이터는 FinanceDataReader 모듈을 사용하여 일별 데이터를 가져왔다. 모든 데이터는 2001 년 1 월부터 2021 년 12 월까지의 데이터이고, 그 날의 종가를 기준으로 한다. 이 데이터들의 날짜는 특성에 따라 다르다. 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. 데이터의 특성은 아래 <표 1>과 같다.

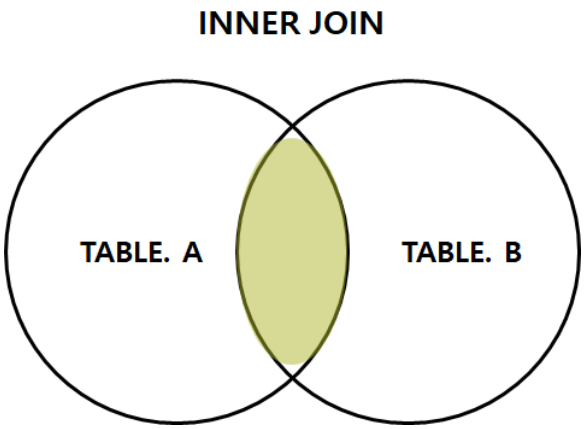
<표 1. 거시 변수 특징>

	kospi	vix	sp500	dji	usdkrw	eurkrw	cnykrw	jpykrw	wti	krlyt	krl0yt	uslmt	usl0yt
count	3929.000000	3929.000000	3929.000000	3929.000000	3929.000000	3929.000000	3929.000000	3929.000000	3929.000000	3929.000000	3929.000000	3929.000000	3929.000000
mean	1901.874159	18.491441	1977.933421	17336.792235	1105.227310	1373.145834	162.141240	10.677943	53.983064	2.704832	3.438830	1.197119	2.793752
std	518.930020	8.284329	896.362015	7163.398298	92.624812	141.992533	21.955396	1.828082	18.403362	1.383894	1.438425	1.567450	1.154035
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%	1637.910000	13.060000	1274.980000	11400.280000	1055.500000	1273.030000	152.540000	9.577700	36.580000	1.535000	2.154000	0.043000	1.918000
50%	1964.690000	16.070000	1695.530000	15451.010000	1118.910000	1338.020000	169.340000	10.292700	60.190000	2.650000	3.180000	0.205000	2.575000
75%	2108.750000	21.320000	2553.170000	22997.440000	1160.100000	1448.160000	176.180000	11.294900	69.790000	3.590000	4.900000	1.995000	3.722000
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

ii. 데이터 전처리 과정

데이터 전처리는 INNER JOIN 방법을 이용하였다. INNER JOIN 은 관계형 데이터베이스에서 테이블 처리 방식 중 하나로, 쉽게 얘기해서 테이블 간의 교집합을 사용하는 것이다. 각 모든 데이터들이 'Date' 형태의 index 를 갖도록 설정을 했기 때문에 13 개 변수들의 테이블 간 교집합을 하였다. 이렇게 해서 총 확보한 데이터는 3929 개이다. 아래 그림 1 은 INNER JOIN 의 예시이다.

<그림 1 INNER JOIN>



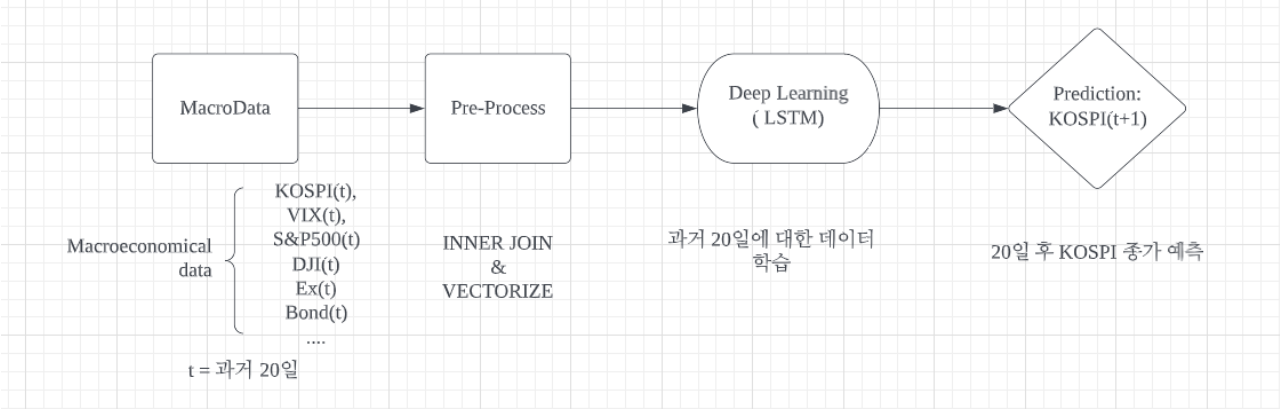
LSTM 모델 학습시키기 위해서는 데이터의 형태가 백터 형태되어야 하기 때문에 추가적인 처리 작업을 해야 한다. 우선 각 데이터 마다 단위와 값의 크기가 다르기 때문에, Minmaxscaler 모듈을 이용해서 각 값들을 0 ~ 1 사이의 값으로 바꿔줬다. 그 후 학습 데이터와 테스트 데이터를 80:20 비율로 나눴다.

III. 분석 방법론

i. 딥러닝 알고리즘 설명

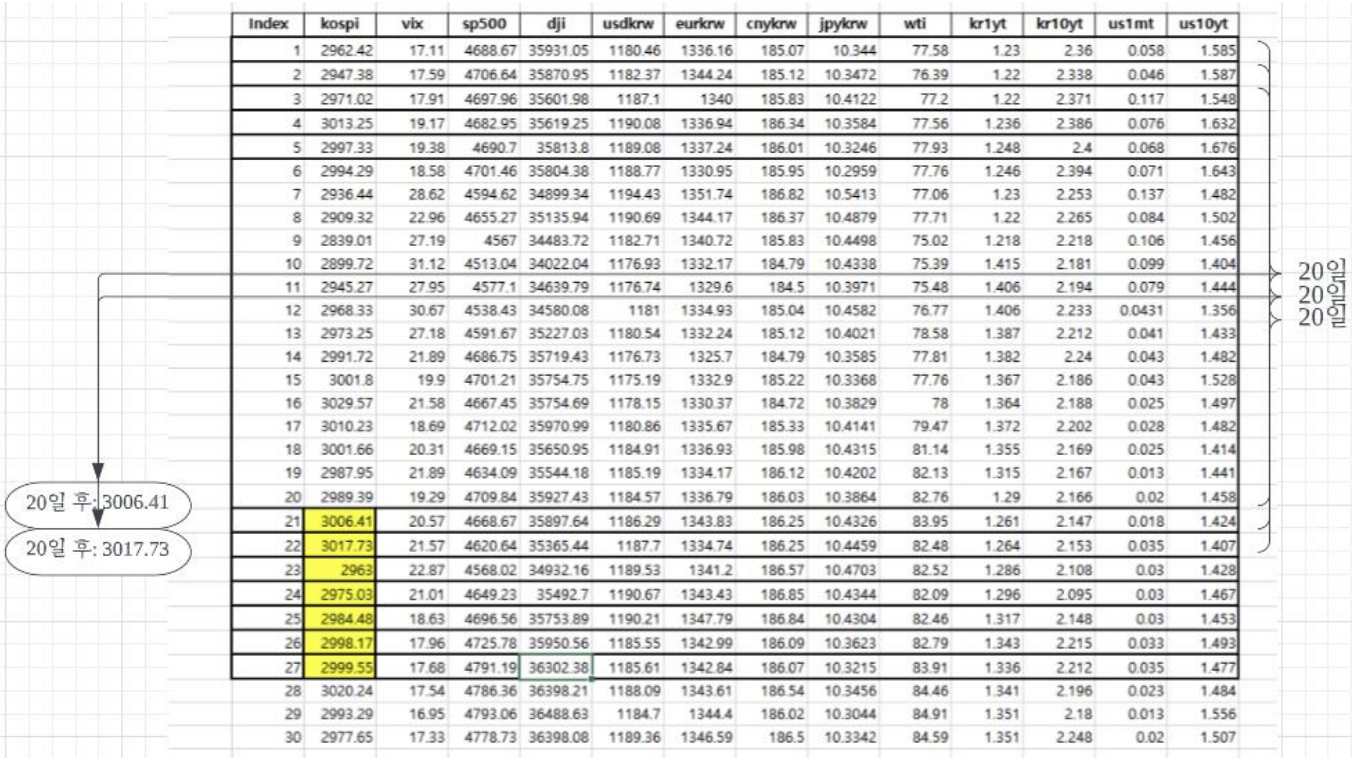
Long Short-Term memory 모델은 순경 신경망 (RNN)의 기존 단점을 보완하여 장/단기 기억을 가능하게 설계한 신경망이다. RNN 은 출력과 먼 위치에 있는 정보를 기억할 수 없는 반면, LSTM 은 forget gate, input gate, cell state 를 통해서 과거 정보를 RNN 보다 더 잘 간직한다. 아래 <그림 2>은 KOSPI 예측에 사용된 알고리즘을 알기 쉽게 도식화한 것이다.

<그림 2 알고리즘>



II.ii 데이터 전처리 과정에서 언급했듯, LSTM 모델에 학습이 되기 위해서는 데이터의 형태가 백터 형태로 되어야 한다. KOSPI 종가에 대한 예측은 과거 20 일의 13 개 변수 (KOSPI 종가, VIX 지수, S&P500, DJI, 환율, 채권) (20 x 13)의 데이터를 학습하여 이뤄진다. 한번의 iteration 동안 Batch size 를 32 로 하였기 때문에, 20 개씩 하루씩 미뤄가며 쌍을 만들기 때문에 12 개가 된다. 이해를 돕기 위해 아래 <그림 3>에 실제 데이터로 작동 방식에 대하여 나타냈다. Index 1~20 번 까지의 13 개의 변수를 이용해서 21 번째의 Kospi 종가를 예측하게 된다. 또 Index 2~21 번 까지의 13 개의 변수를 이용해서 22 번째의 Kospi 종가를 예측하게 된다. 이렇게 쭉 가서 index 12~31 번 까지의 13 개 변수를 이용해서 32 번째 Kospi 종가를 예측한다. 우선 80%의 학습 데이터는 실제 데이터에 대한 정보를 가지고 있기 때문에 예측값을 실제값과 최대한 비슷하게 하는 방식으로 모델을 학습시킨다. 그리고 지나치게 학습데이터가 많거나 편향되는 것을 막기 위해 Early-stopping 기법을 도입했다. 이는 특정 epoch(반복 횟수) 동안 모델이 개선되지 않으면 학습을 멈추는 것인데, 이번 실험에서는 10 번 epoch 동안 모델이 개선되지 않으면 멈추도록 설정하였다.

<그림 3 데이터 형태, window size=20>



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

이번 모델에서 사용하는 layer 는 총 4 가지로 구성되어 있고, 각 output shape 과 업데이트되는 파라미터 값은 아래 <그림 4>에서 참고할 수 있다.

<그림 4 Model layers & parameters>

```
Model: "sequential"
_________________________________________________________________
 Layer (type)                Output Shape              Param #
=================================================================
 conv1d (Conv1D)             (None, 20, 32)            2112

 lstm (LSTM)                 (None, 16)                3136

 dense (Dense)               (None, 16)                272

 dense_1 (Dense)             (None, 1)                 17

=================================================================
Total params: 5,537
Trainable params: 5,537
Non-trainable params: 0
```

IV. 결과 및 분석

이번 장에서는 3 장에서 제시한 모델로 나온 결과에 대해 논의한다.

i. 결과 및 분석

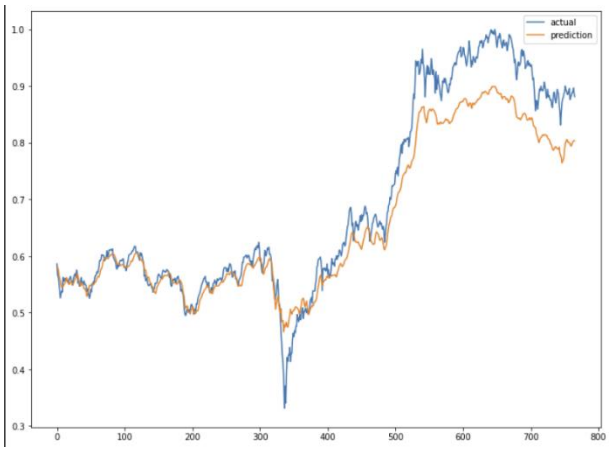
<표 2 실험에 대한 RMSE>

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

위 <표 2>는 실험을 1~5 을 진행할 때 위에서 언급한 LSTM 모델로 예측한 RMSE(Root Mean Square Error)를 나타낸 것이다. 사실 다른 뛰어난 알고리즘 보다 더 좋은 성능을 구하진 못했다. 온라인 뉴스 및 거시 경제 변수를 합쳐 만든 LSTM 모형의 RMSE 평균 0.010892 정도의 성능이 나온다. 그럼에도 다른 SVR 을 이용한 알고리즘의 평균 RMSE 는 0.048392 정도임을 감안하면, 확실히 LSTM 모델을 이용하면 더 좋은 성능을 보여줌을 나타낸다. 아래 <그림 5>는 5 번의 실험동안 예측값 (주황색) 과 실제값 (파란색)을 도식화하여 표현했다. 하나 특징으로 볼 수 있는 것이 x 값 500 (2020 년쯤) 이전은 실제값과 예측값이 별 차이가 없으나, x 값 500 이후부터 예측력이 떨어지는 것을 볼 수 있다. 이는 코로나 이후 일레에 없었던 공격적인 재정과 통화정책이 있었기 때문이다. 통화정책 면에서는 빠른 금리인하와 자산매입이 이뤄졌고, 재정정책 면에서는 재정확장이 큰 폭을 빠르게 이루며 코로나 19 사태에 대한 대응을 하였다. 이는 인공지능도 예측하기 어려웠던 특별한 사건으로 간주되어 예측력이 떨어진 것으로 분석한다.

<그림 5 case analysis>

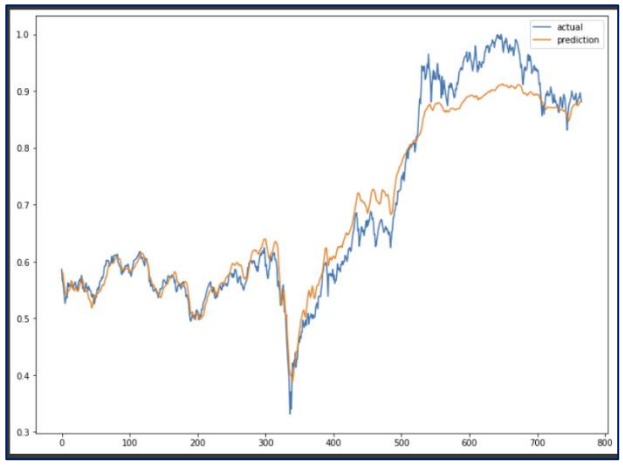
실험 1



실험 2

실험 3

실험 4



실험 5

V. 결론 및 한계점

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 예측하는지에 대하여 연구해봤다.

2001 년 1 월부터 2021 년 12 월까지의 13 개 거시 경제 지표 변수들을 이용해 KOSPI 종가를 예측했다. LSTM 모형에 과거 20 일간의 데이터를 학습시키고 그 다음날의 KOSPI 종가를 예측하도록 구현하였다. 그 결과 기존 SVR 모델보다 더 나은 성능을 보였으나, 더 고성능의 LSTM 모델보단 더 나은 성능을 보여주진 못했다. 또한 직접 하이퍼파라미터 (batch size, epoch 값)에 대한 fine-tunning 진행하지 않아, 더 나은 성능을 보여줄 수 있는 여지가 있다. 모델에 대한 더 나은 공부를 통해 이를 개선할 수 있을 거 같다.