목 차

1. 데이터 선정 이유

- 그림1을 보면 코로나19로 인해 유례없는 개인 투자자의 참여가 확대되고 있다. 이는 다음 팬데믹(pandemic)이 다시 돌아올 경우, 주식 시장의 열기는 지금보다 더 과열될 것이라는 방증이다. 이런 상황을 대비해야 하는 이유는 그림 2를 보면 알 수 있다. 첫 코로나 확진자가 발생한 2020년도 1분기의 코스피는 약 -20% 등락했다. 이는 팬데믹 사태가 다시 발생했을 때 대비해야 하는 이유를 더 추가적인 설명 없이도 알 수 있다.

그림 1 개인 투자자 거래 규모

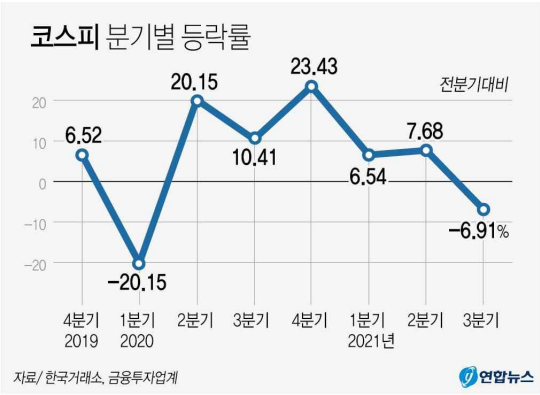

그림 2 코스피 분기별 등락률

2. 계산 머신 및 Library

- Python 3.7
- GPU : GTX 3090
- CPU : AMD Ryzen 5 3600
- Library : pandas, np, csv, plt, keras 등

3. 데이터 설명 및 구성

1. 서울시 확진자 데이터

- 간단한 설명 : 2020-01-19 ~ 2021-10-18까지의 모든 코로나 확진자의 정보를 담는다.
- 파일 형식 : CSV
- 열 : 14개 : '연번', '확진일', '환자 번호', '국적', '환자 정보', '지역', '여행력', '접촉력', '조치사항', '상태', '이동 경로', '수정일', '노출 여부'
- 행 : 112363개
- 출처 : 서울시 공공 데이터

2. 주식 데이터

- 간단한 설명 : 9개의 업종을 선택하여 각 업종에 대해 3개의 종목을 선택했다. 각 업종 별 3개의 종목은 1. 업종 중 시가 총액이 제일 큰 종목 2. 시가 총액이 중간에 위치한 종목 3. 시가 총액이 가장 작은 종목 으로 구성된다. 또한, 27개의 모든 종목은 '2020년 1월 19일부터 지금까지 상장되어 있다.'는 전제를 만족하고 '현재 거래 정지가 아님'을 만족한다.
- 파일 형식 : CSV
- 열 : 7개 :'Date', 'Open', 'High', 'Low', 'Close', 'Adj Close', 'Volume'
- 행 : 430 (2020-01-19 ~ 2021-10-18, 주말 및 공휴일 제외) * 27 = 11,610개
- 출처 : KRX

4. 데이터 전처리

- 먼저 주식 데이터의 7개의 열 중 필요로 하는 'Date'와 'Close'를 제외하고 나머지 열을 제거한다.

- 서울시 확진자 데이터의 불필요한 행을 제거한다.

- 서울시 확진자 데이터의 총 확진자를 'Date'별로 수정한다.

- 주식 데이터에만 존재하는 Date를 기준으로 코로나 확진자 데이터의 Date와 비교하여 코로나 확진자 데이터의 주말 및 공휴일을 제거한다.

- 이때 Null 값이 생기는데, 이는 서울시 확진자 데이터의 'Date'값 중 확진자가 발생하지 않은 날은 적혀있지 않기 때문이다.

- 이를 해결하기 위해 서울시 데이터의 1차 전처리를 통해 공휴일 및 휴일을 제거하고 2차 전처리를 통해 확진자가 0명인 날의 값을 'Date'에 확진자 수 '0'을 Co_num 에 추가했다.

5. 상관관계 분석

- 코로나 확진자와 각 주식의 종가의 상관관계를 분석하기위해 피어슨 상관계수(pearson correlation coefficient)를 이용한다.

1. pearson correlation coefficient란

- 피어슨 상관계수는 통계 및 머신 러닝 분야에서 가장 중요한 파라미터를 측정하는 중요한 방법 중 하나다 [1]. 피어슨 상관계수는 목표값과 각 특성 간의 상관관계를 의미하며 [-1, 1] 사이 값을 갖는다. 1에 가까울수록 강한 양의 선형관계이고, -1에

가까울수록 강한 음적 선형 관계를 갖는다. 피어슨 상관계수를 구하는 수식은 (1)과 같다.

$$P = \frac{\sum_{i=1}^{J}(x_i - \tilde{x})(y_i - \tilde{y})}{\sqrt{\sum_{i=1}^{J}(x_i - \tilde{x})^2 (y_i - \tilde{y})^2}} \quad (1)\ [2]$$

x_i, y_i는 피처의 i 번째 값, $\tilde{x}, \tilde{y}$(각 피처의 평균값), J는 피처의 개수이다.

피어슨 상관계수를 막대 그래프로 표현하면 그림3과 같다. 상세 값은 표2에 표시돼있다.

피어슨 상관계수가 유효한 값을 0.6으로 설정했다. 그에 해당하는 값은 표2에 볼드체로 표시되어 있다. 9개의 종목이 코로나 확진자와 선형성을 갖고 있음을 알 수 있다.

비례관계를 갖는 종목은 '코리아 에셋', 모두투어, 태웅, 아시아 경제로 나타났고 반비례 관계의 종목은 'HMM', '기아', '현대차',' KSS해운', 'CJ 제일 제당'이다.

9개의 종목의 업종은 자동차('기아', '현대차'), 해운('HMM', 'KSS해운', '태웅'), 출판('아시아 경제'), 증권('코리아 에셋'), 레저('모두 투어'), 식품('CJ 제일제당')이였다.

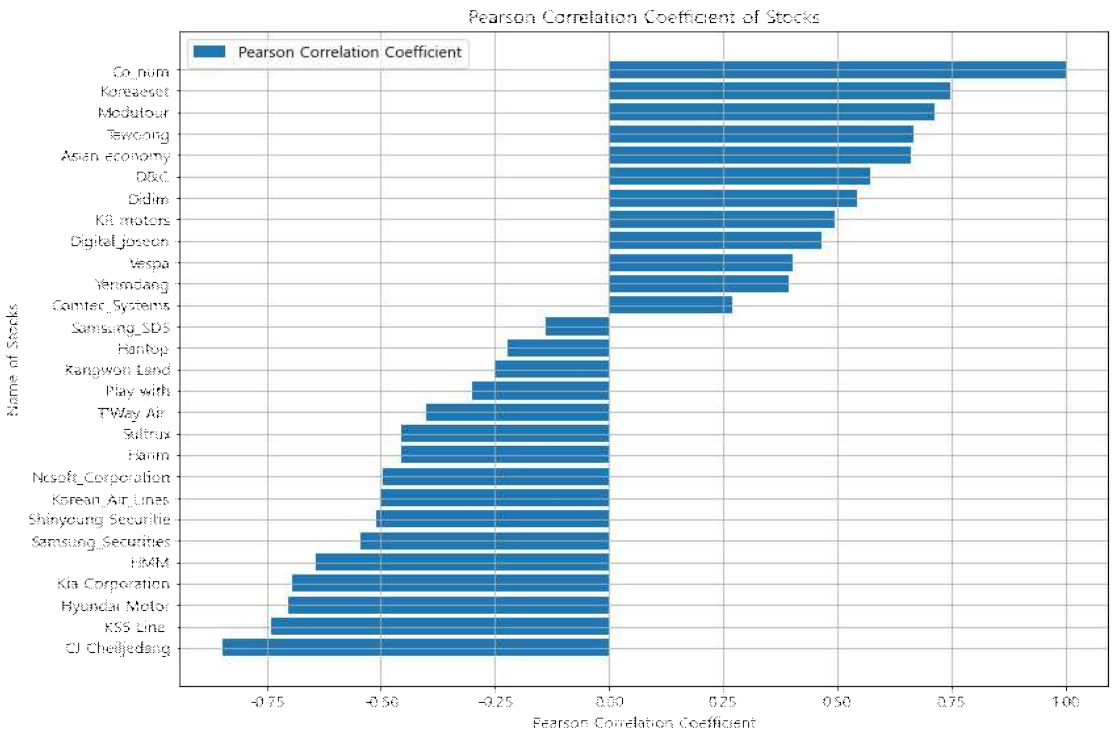

그림 3 Pearson Coreelation Coefficient of Stocks

Names of Stocks	Pearson Correlation Coefficient
Co_num	1
Koreaeset	**0.74761**
Modutour	**0.712227**
Tewoong	**0.665133**
Asian economy	**0.661694**
D&C	0.570928
Didim	0.54199
KR motors	0.4953
Digital_joseon	0.464174
Vespa	0.40169
Yerimdang	0.393108
Comtec_Systems	0.26982
Samsung_SDS	-0.13784
Hantop	-0.22239
Kangwon Land	-0.25234
Play with	-0.29893
T'Way Air	-0.39984
Sultrux	-0.45605
Harim	-0.45605
Ncsoft_Corporation	-0.49692
Korean_Air_Lines	-0.50084
Shinyoung Securitie	-0.50982
Samsung_Securities	-0.54412
HMM	**-0.64211**
Kia Corporation	**-0.69293**
Hyundai Motor	**-0.70259**
KSS Line	**-0.74143**
CJ Cheiljedang	**-0.84786**

표 2 Pearson Correlation Coefficient of Stocks

2. 자동차 업종의 주가

자동차 업종을 살펴보면 '기아', '현대차' 모두 반비례 관계에 놓여있다. 이는 코로나 확진자가 증가할수록 주가가 올랐음을 알 수 있는데, 그림 3을 보면 주가가 상승했음을 알수 있다. 두 주식 모두 유사한 형태를 갖고 있으므로 코로나 확진자와 상관관계가 뚜렷함을 알수 있다.

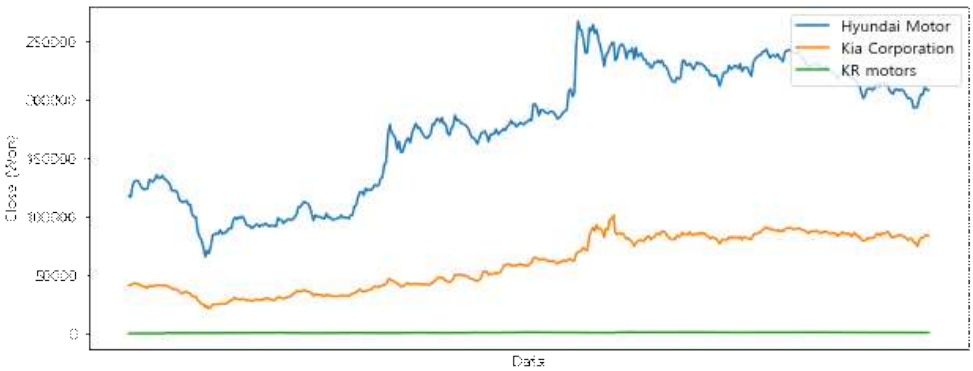

그림 4 자동차 업종의 주가

3. 해운 업종의 주가

해운을 살펴보면 'HMM', 'KSS해운'은 반비례 관계, '태웅'은 비례관계를 갖고 있다. 코로나 확진자가 증가함에 따라 'HMM'의 주가가 증가하다 하락세를 보임을 알수 있다. 'KSS해운' 또한 코로나 확진자의 증가로 인해 소폭 감소하고 있음을 알수 있다. 반대로 '태웅'은 코로나 확진자가 증가해도 주가가 상승세를 보임을 알수있는데, 이는 해운에 관련한 종목은 코로나 확진자 수만으로 판단하기 어려움을 나타낸다.

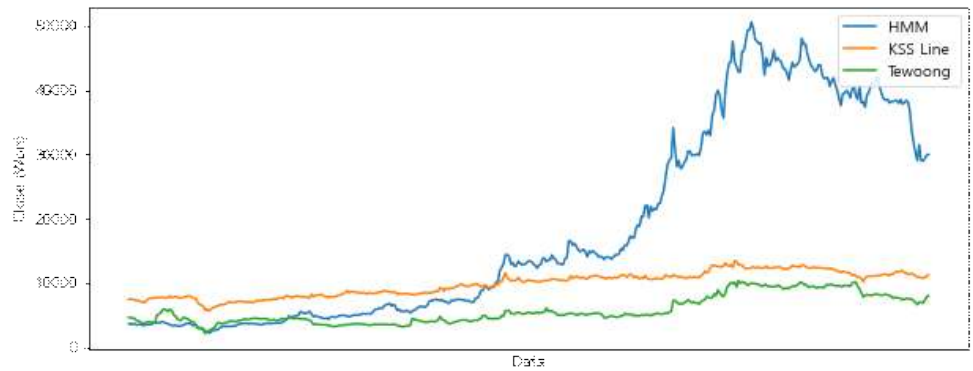

그림 5 해운 업종의 주가

4. 출판사 업종의 주가

'아시아 경제'를 살펴보면 코로나 확진자가 소폭 증가할 경우 다른 외인 변수에 의해 값이 변함을 볼수 있다. 코로나 확진자가 대폭 증가함에 따라 주가는 점점 하락하다. 위드 코로나와 백신 접종이 시작된 이후부터 회복함을 볼수 있다. 이는 코로나의 장기화에 따른 문화생활의 변화를 보여준다고 판단된다. 하지만 코로나 사태 초반에 외인 변수에 의한 주가 변동이 매우 큼으로 다른 업종의 주식을 사는 것을 추천한다.

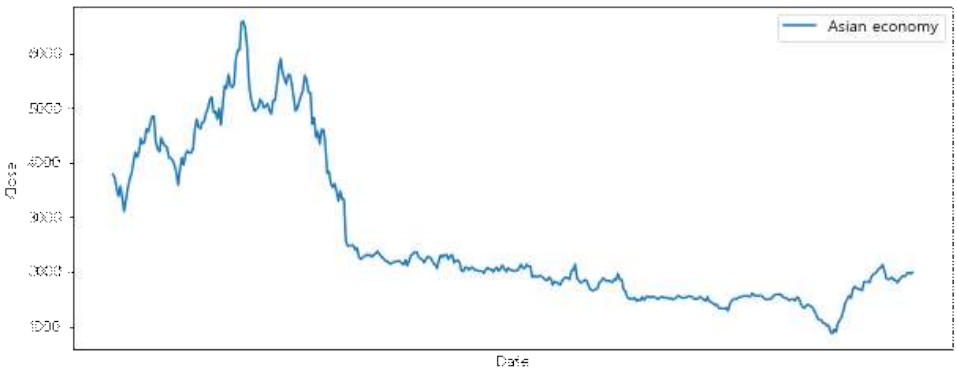

그림 6 아시아 경제의 주가

5. 증권사 업종의 주가

다음으로 증권사를 살펴보면 '코리아 에셋'은 가장 강한 비례관계를 갖고 있다. 코로나 확진자가 소폭 증가하는 경우 하락세를 보이고 있었지만, 최근 코로나 확진자가 대폭 증가함에 따라 상승세를 보인다. 다른 두 증권사도 비슷한 형태를 갖고 있음을 알 수 있는데, 이는 코로나 확진자가 증가 추세를 보이는 경우 구매하는 것이 좋다고 판단할 수 있다.

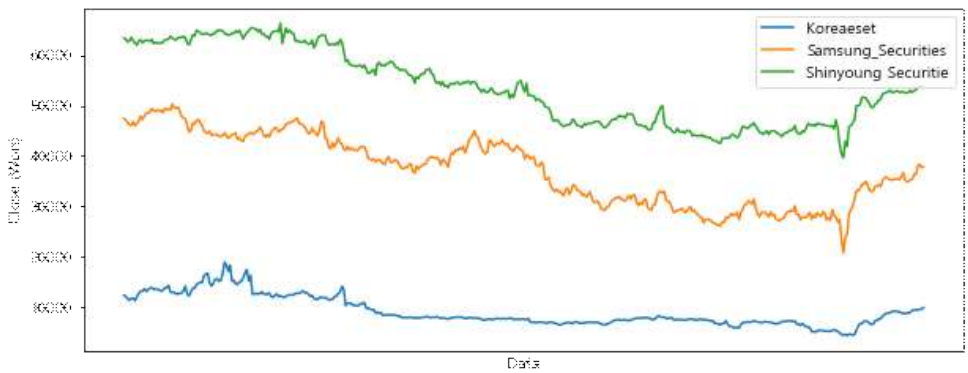

그림 7 증권사 업종의 주가

6. 레저 및 호텔 업종의 주가

레저를 살펴보면 '모두투어'는 강한 비례 관계를 갖고 있다. 코로나 확진자가 증가함에 따라 하락세를 보였지만 코로나 사태가 장기화 되고 '위드 코로나'에 대한 기대감이 커지면 주가를 천천히 회복하고 있는 모습이다. 같은 업종의 '강원 랜드' 또한 하락세를 보이다 점점 회복세를 보이고 있는데, 이는 호텔, 레저 종목은 코로나가 장기화 되고 주가가 꾸준히 하락세를 보이는 경우 구매하는 것이 좋음을 나타낸다. 또한 주가는 백신 접종률과 비례할 것으로 예상된다.

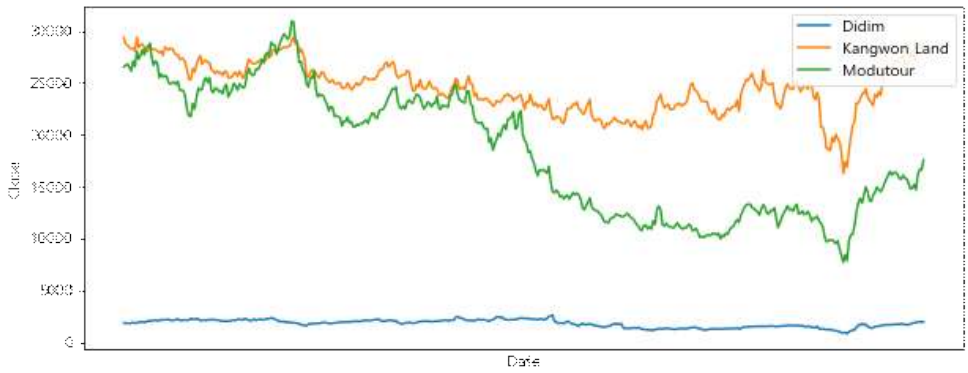

그림 8 레저 및 호텔의 주가

7. 식품 업종의 주가

마지막으로 식품을 살펴보면 'CJ 제일 제당'은 코로나 확진자가 적을때는 주가가 상승했지만, 확진자가 급격히 상승하면서 하락세를 보이고 있다. 이는 다음 팬데믹에 식품 관련주를 구매할 경우, 코로나 확진자가 급격히 증가하기 전에 판매하는 것이 좋음을 알 수 있다.

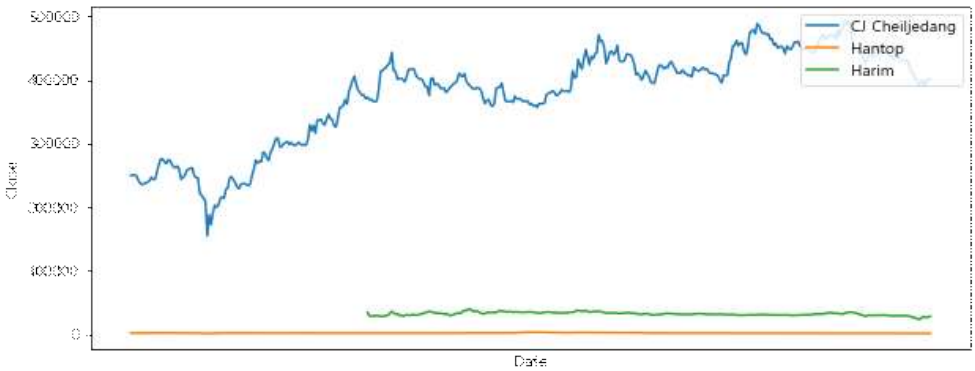

그림 9 식품 업종의 주가

6. 알고리즘을 통한 예측

- 본 프로젝트에서는 2개의 알고리즘을 기반으로 예측을 진행한다. 첫 째로는 Lstm을 이용한다. Lstm은 주식과 관련된 분야에 가장 많이 사용되는 알고리즘이다. 둘째는 DNN이다. 이를 선택한 이유는 5에서 언급한 피어슨 상관계수 기반의 데이터 분석이 유의미한 결과를 나타냈다고 생각했다. 강한 선형성을 갖는 데이터가 많았고 이는 비선형을 이용한 딥러닝을 이용해 예측한다면 높은 정확도의 예측을 할수 있을 것이라고 판단했다.

1. DNN(Deep Neural Network)

- DNN은 다양한 분야에 사용된다[3]. Input Layer와 Hidden Layer 그리고 Output Layer로 이루어진다. 본 프로젝트에서는 하나의 Input Layer와 총 5층의 Hidden Layers, 하나의 Output Layer로 이루어진다. Hyper Parameter로는 Learning Rate : 0.001, 활성화 함수는 'elu'를 사용했다.

- 학습 데이터는 2020-01-19 ~ 2021-09-30의 데이터를 이용했고 검증 데이터는 2021-10-01 ~ 2021-10-18의 데이터를 사용했다. Test 데이터를 사용하지 않았는데 이는 학습 데이터 수의 부족과 피어슨 상관계수를 통해 많은 데이터가 일정 수준의 선형성을 갖고 있음을 확인했기 때문에 학습 데이터와 검증 데이터의 오차가 유사할 것 이라고 판단했다.

- Hidden Layers는 총 4단계로 각 3,2,1,4로 이루어진다.

- 평가 기준은 RMSE와 2차원 그래프 (x축 : 예측, y축 정답), R-square를 이용한다.

2. RMSE

RMSE(Root Mean Square Error)는 모델의 예측값과 정답의 평균적인 차이를 확인할 수 있는 지표이다. 딥러닝 및 머신러닝에서 평가 지표로 자주 사용된다[4, 5].

y_i는 i 번째의 정답, t_i 는 i 번째의 예측값이다.

$$\text{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - t_i)^2} \quad \dots \ (3)$$

3. R-square

R-square는 회귀의 적합도를 평가하는 척도로 널리 사용된다[6]. R-square 값은 [0, 1] 사이의 값을 갖는다. R-square 값이 클수록 모델이 데이터에 잘 fit 됐음을 의미한다[7].

$$R^2 = 1 - \frac{SSR}{SST},\ SSR = \sum(Y_i - \hat{Y})^2,\ SST = \sum(Y_i - \overline{Y})^2,\ \overline{Y} = \sum Y_i / n \quad \text{...(4)[8]}$$

R-square 값은 독립변수의 개수가 증가함에 따라 상승하기 때문에 결정계수 하나만으로 모델을 판단할 수 없다[8]. 따라서 본 연구에서는 3가지 지표를 종합하여 모델을 평가한다.

7. 결과 분석 및 평가

1. DNN

Names of Stocks	RMSE
'T'Way Air '	**2,798**
'Comtec_Systems'	**4,197**
'KR motors'	**5,361**
'Hantop'	**8,654**
'Asian economy'	**9,724**
'Samsung_Securities'	10,963
'Play with'	12,447
'Didim'	12,922
'Koreaeset'	17,068
'KSS Line '	23,882
Yerimdang	24,986
'Korean_Air_Lines'	47,878
'Tewoong'	48,229
'HMM'	69,613
'Kangwon Land'	99,644
'Modutour'	113,816
'Vespa'	220,408
'Kia Corporation'	245,280
'Shinyoung Securitie	251,754
'D&C'	270,692
'Samsung_SDS'	820,468
'Hyundai Motor'	859,564
'CJ Cheiljedang'	1,242,177
'Ncsoft_Corporation'	5,270,580
'Sultrux'	NaN
'Harim'	NaN
'Digital_joseon'	NaN

표 3 각 종목 별 DNN 예측 및 오차 (RMSE)

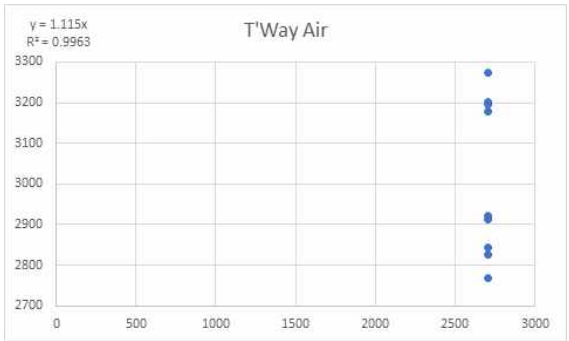

그림 10 DNN의 T'Way Air 주가 예측

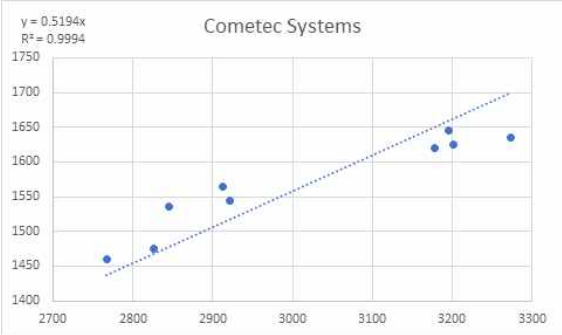

그림 11 DNN의 Cometec Systems 주가 예측

그림 12 DNN의 KR motors 주가 예측

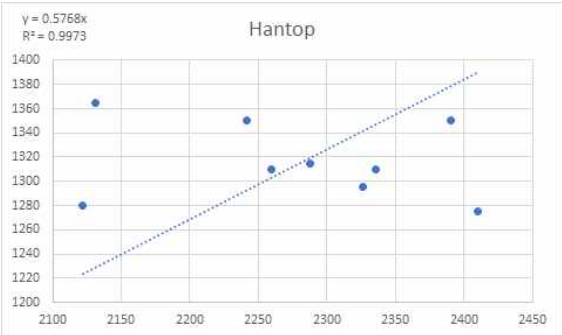

그림 13 DNN의 Hantop 주가 예측

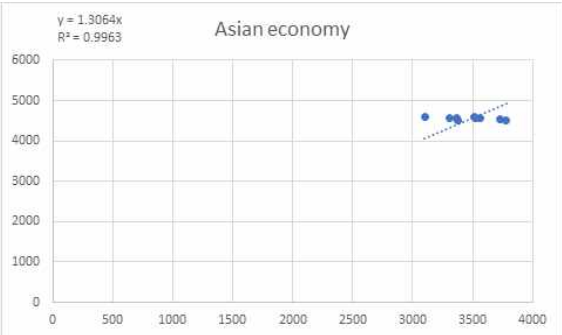

그림 14 DNN의 Asian economy 주가 예측

- 표3은 각 종목에 대한 DNN의 예측과 실제 데이터의 RMSE 값을 나타낸다. 이 중 가장 잘 예측한 종목은 'T'Way Air'다. 그림 10을 보면 매우 높은 R-square을 갖고 평균 오차는 1주당 약 2,700원의 오차를 갖는다. 현재 1주당 약 4000원 인것으로 보아 오차가 작지는 않지만 의미있는 예측이라고 생각된다.

- 표3에 볼드체로 표시한 나머지 4개의 종목 또한 낮은 오차로 예측되고 있다. 이는 다음 팬데믹이 일어난다면 본 알고리즘을 통해 주가를 성공적으로 예측할수 있음을 나타낸다.

- 또한 이는 코로나 확진자의 수가 일정부분 주가에 영향을 미침을 알수 있다.

8. 결론

- 본 프로젝트에서는 총 3개의 방법을 제시했다. 주가를 결정하는 요인은 매우 다양하기 때문에 코로나 확진자와 연관지어 예측하는 것은 매우 힘든 일이다. 하지만 피어슨 상관계수, DNN을 이용하여 일부 종목의 주가를 성공적으로 예측하였다.

- 피어슨 상관계수를 통해 가장 높은 정확도로 예측할 수 있는 업종은 자동차다. 자동차는 기업의 크기에 상관없이 확진자 수와 반비례하는 경향을 나타냈다. 이는 다음 팬데믹이 온다면 코로나 확진자가 적을 때 팔고, 많아지면 사야함을 유추할 수 있다.

- DNN을 이용한 예측에서 가장 높은 정확도를 갖는 종목은 항공주의 'T'Way Air'였다. 이는 매우 적은 오차를 갖고 있으며 다음 팬데믹에 본 알고리즘을 통해 예측할 수 있음을 알수 있다.

9. 레퍼런스

그림 1
https://www.kcmi.re.kr/publications/pub_detail_view?syear=2021&zcd=002001016&zno=1581&cno=5644

[1] Multiple Ant Colony Optimization Based on Pearson Correlation Coefficient
[2] Quantifying Colocalization by Correlation: The Pearson Correlation Coefficient is Superior to the Mander's Overlap Coefficient
[3] A Study Trend on DNN security by using Trusted Execution Environment
[4] Root mean square error (RMSE) or mean absolute error (MAE)?
[5] Analysis of the Mean Absolute Error (MAE) and the Root Mean Square Error (RMSE) in Assessing Rounding Model
[6] A Shapley-based decomposition of the R-Squareof a linear regression
[7] Impact of Data Structure on the Estimators R-Square And Adjusted R-Square in Linear Regression
[8] An R-square coefficient based on final prediction error