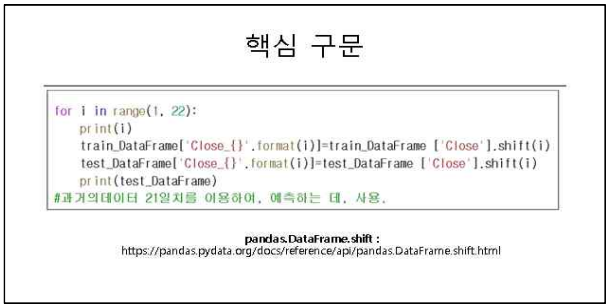
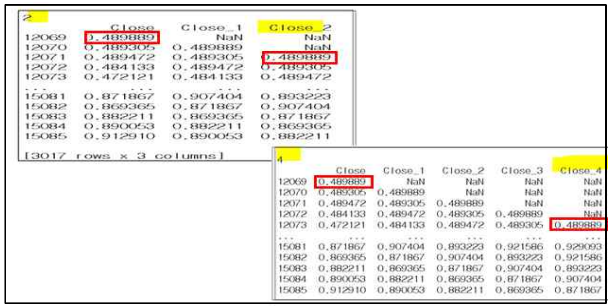
lstm을 통한 주가예측 프로젝트 발표시작하겠습니다.
우선 주가데이터를 받는 방법에 대해 말씀드리겠습니다. 주가 데이터의 경우, 키움증권과 같은 증권사의 hts를 활용하거나, 야후 파이낸스를 통해서 받을 수 있습니다.
이때 증권사의 hts의 경우, 장기간 주가 데이터를 얻을 수 없기에, 야후 파이낸스를 사용하였습니다.

과정은 사진과 같습니다. 우선 분석하고 싶은 주식이름이나 티커를 입력하여, 주가 데이터를 얻습니다. 해당 조건을 만족하기 위해서 오랫동안 상장되어 있던 코카콜라를 선정하였습니다.

그렇게 얻은 csv파일의 데이터를 다음과 같이 출력하였습니다.
그리고 이전 프로젝트와 같이 MinMaxScaler를 통해 데이터를 정규화합니다.
이때 날짜를 나타내는 date의 값은 잘라냅니다. 왜냐하면, 일자는 정규화하기 어렵기 때문입니다.

그다음에 test와 train값을 2:8로 구분합니다.

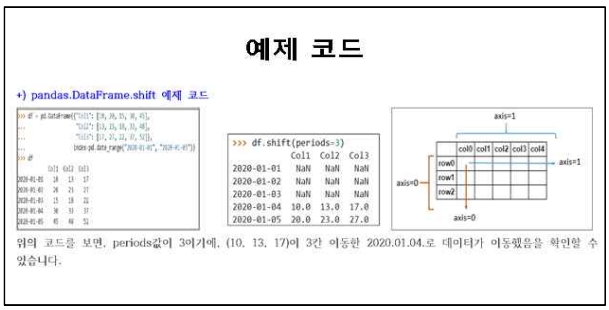


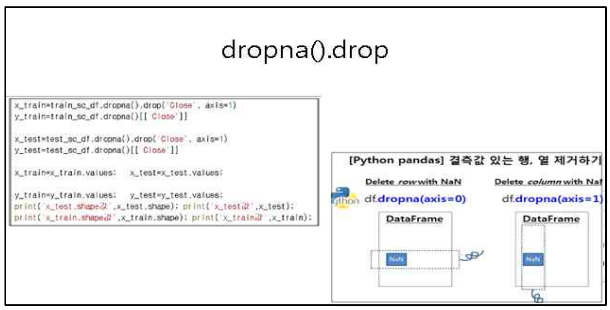


다음 lstm을 활용하여 주가를 예측할 때 가장 중요한 구문이라고 생각합니다.
dataframe.shift함수입니다. 이를 통해, 주가데이터를 예측합니다.

즉, Pandas.DataFrame.shift는 지정된 기간 수만큼 dataframe의 인덱스를 이동하기에, sliding window를 만들어 미래의 주가를 예측합니다.

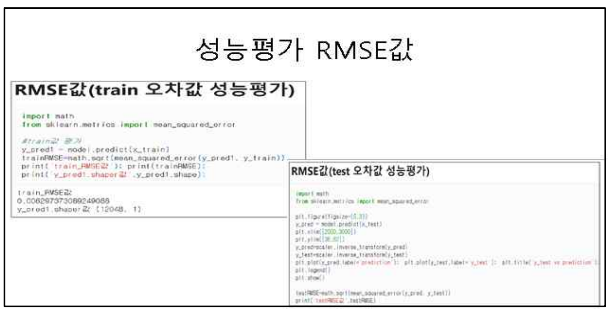
다음 그림을 보면, for문의 동작을 출력한 것으로 "0.489889"가 I가 2일 때는 2칸 이동하였고, 4일 때는 4칸 이동한 것을 볼 수 있습니다.



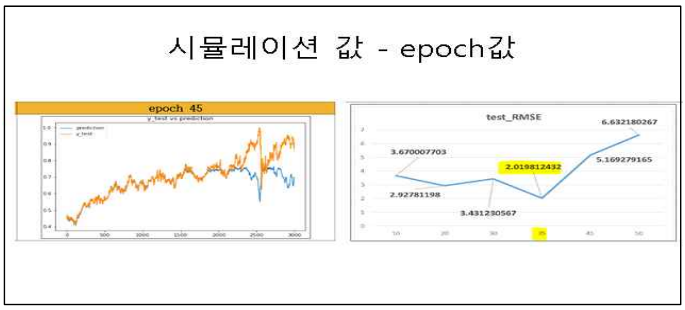


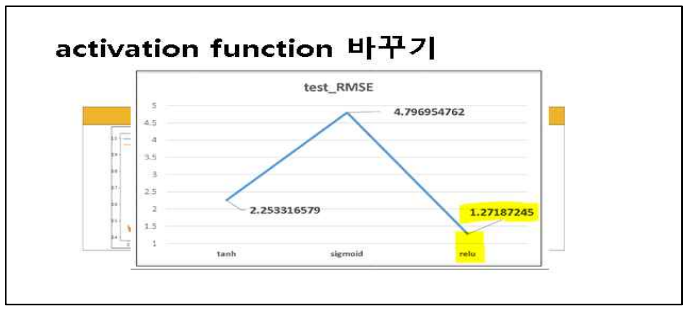
즉, 다음 표의 예제코드의 경우, period가 3이기에, 1월1일의 값인 10, 13, 17이 3일 이동한 1월4일로 이동한 것을 볼 수 있습니다. 이때 NaN과 같은 결측값이 보입니다. 이를 제거하기 위해서, 다음의 함수인 dropna().drop을 사용합니다. 이때 axis=1을 볼 수 있는데, 이는 다음 그림처럼 행이냐 열이냐를 결정하는 것입니다.

즉, "NaN"인 결측값는 데이터 분석 시 성능에 영향을 주기에, 이 함수(dropna())는 pandas가 결측값을 제거하거나 특정 값으로 채워주는 식으로 처리합니다.



마지막으로 파라미터 요소인 optimizer, activation function, days 등 여러요소를 바꾼 후 시뮬레이션을 돌려 좋은 값을 알기 위해서 성능평가인 RMSE를 사용하였습니다. 결과적으로 보면, train값은 큰 차이가 없으나 test의 RMSE의 값이 많이 났기에, 이 값을 위주로 성능 평가를 했습니다.





– 이전에 한 project에서도 epoch를 늘리면, 시간이 오래 걸리지만 성능적인 면에서 **향상이 된다는 것을 확인할 수 있었습니다. 특히 "train_RMSE"의 값은 그래프 적으로는 변화가 있어 보이지만, 그 값의 변화는 거의 없지만, "test_RMSE"의 경우, epoch35까지는 epoch에 따라 성능이 좋아지다가, 그 이후부터는 over-fitting이 발생하여 RMSE의 값이 커지는 것을 확인할 수 있습니다.**

lstm의 활성화 함수값을 바꿔서 성능평가도 해봤는데, relu, sigmoid, tanh 중 relu가 가장 좋은 성능을 보여주었습니다.

그리고 다른 파라미터보다 중요하다고 생각되어지는 day값에 따른 결과값입니다.

그래프를 보면, 3~7일 즉, 단기간의 값을 활용할 때 주가 데이터의 정확도가 높은 편이지만, 기간이 길어질수록 RMSE가 커지고, 즉 예측 성능이 나빠진다는 것을 확인하였습니다.

이를 통해, 과거 주가 데이터만으로 주가를 데이터하기에는 적정 기준이 있음을 확인했습니다.

마지막으로 optimizer에 따른 성능차이를 보았습니다. 결과 값을 보면, Nadam과 adam이 가장 좋은 성능을 보여주었습니다.

<결론>

- 딥러닝 수업을 처음으로 들었을 때, 모의투자대회에서 입상했을 정도로 관심 있던 주식분야를 분석해보고 싶었는데, LSTM이라는 것을 알게 되어 주제로 한 번 다뤄봐서 재미있는 프로젝트였습니다.

특히 RMSE로 성능 평가했을 때, 약간의 오차가 발생하는 것을 볼 수 있지만, 이 오차가 아무리 작더라도 실제로 투자로 옮길 경우, 손해가 발생할 가능성이 존재하기에 실제 투자로 옮기기는 어렵다는 것을 결론을 내릴 수 있습니다.

그래서 어떤 논문에서는 twitter의 언급되는 메시지에 따라 호재인가 악재인지를 파라미터로 추가 설정하여 주가를 예측하였습니다.

즉, 기존의 재무제표와 주가 데이터만으로는 주가를 예측하기 어렵기에 여러 파라미터를 추가하여, 주가의 변동성을 예측하고 있습니다.

따라서 가능하다면 기존의 증권사가 운영하고 있는 로봇 어드바이저는 어떤 파라미터 요소로 주가를 예측하는 것인지와 100% AI가 예측하는 자동매매 프로그램인지, '코로나'나 '블랙스완'과 같은 이슈를 대처할 수 있는 현 자율 주행차 기술처럼 반자동기술인지에 대해 알아보고 싶습니다.