2021 미래에셋

금융 빅데이터 페스티벌

자산운용 보고서

캡스톤 디자인 중간보고서

**- 주제3.TIGER ETF 기반 투자 전략 개발**

팀 명: ADAM

이준석/20163132

윤상희/20163234

김진영/20187155

# **Ⅰ. 전략컨셉**

요새 딥러닝과 머신러닝을 활용한 주가예측 논문들이 많이 나오고 있는데, 매일 읽어 보기만 했지 직접 해보지는 못했다. 그래서 저희는 딥러닝에서 시계열 데이터에 유리한 GRU와 LSTM 그리고 특성 추출에 유리한 Convolution 1d Layer들을 사용하여 지수가 오르고 내리는 것을 예측 할 수 있는지 구현해보고 싶었다.

**○ 사용모델**

* 딥러닝 : Conv1D + CuDNNGRU + CuDNNLSTM
* 머신러닝 : Random Forest Regression + Adaboost Regression
* 딥러닝 모델 + 머신러닝 모델 ensemble

**Ⅱ. 분석방법**

1. **데이터 분석**

처음에는 RSI, 이동평균선, Money Flow Index,MACD등 다양한 것을 사용해보고 싶었지만 테스트셋을 축소 시켜야 돼서 제출형식에 맞지 않아서 (고가+저가+종가)/3을 한 typical price를 추가하고 거래량 과 typical price 곱한 Money Flow column까지 추가했다.

typical price는 성능 향상에 도움이 되었지만 Money Flow는 값이 너무 커서인지 성능을 오히려 떨어트려서 삭제했다.

1. **모델링**

2\_1 Random Forest Regression



Random Forest Regression

KFold Validation을 시행했다.

위의 모델을 스코어를 냈을 때 가장 높은 스코어가 나왔지만 submission 파일의 score가 낮은 부분이 많았다. 하지만 다른 부분에서 score가 뛰어났다.

2\_2 AdaBoost Regression

텍스트이(가) 표시된 사진

자동 생성된 설명

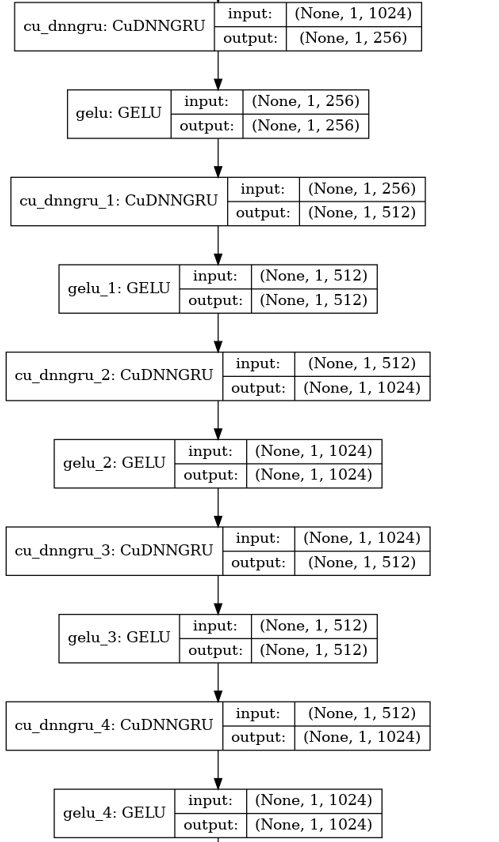
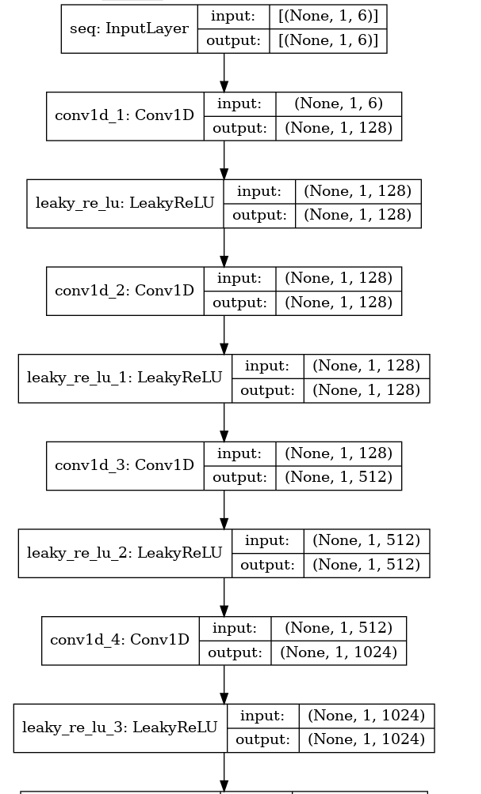
AdaBoost Regression도 같이 KFold Validation을 시행하였다.

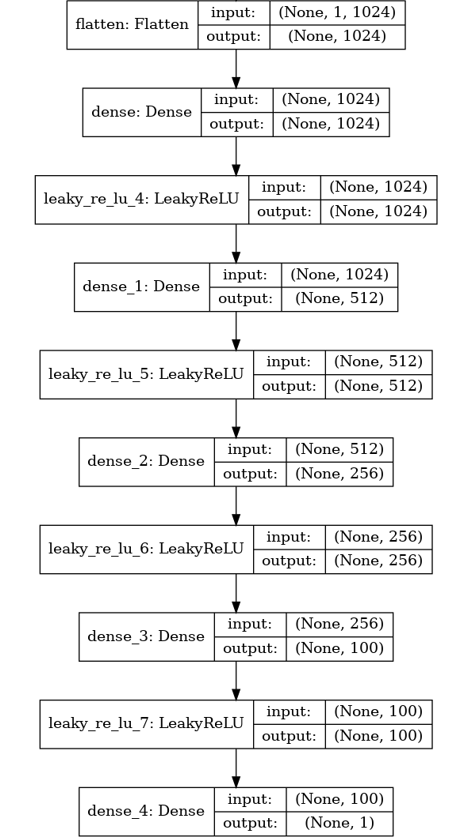
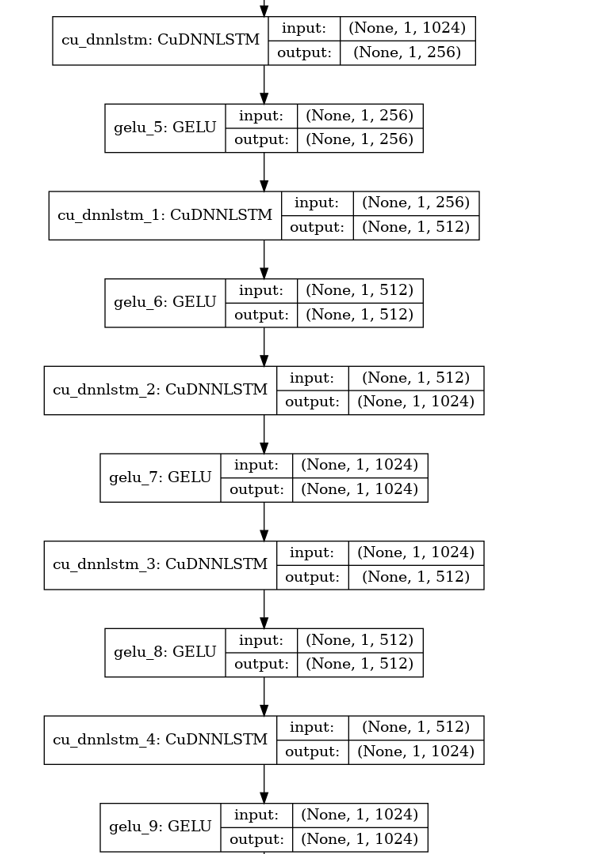
여기서 AdaBoost에 base estimator를 Lasso로 사용하였는데, Lasso는 L1-norm이 페널티를 가지는 선형회귀 방식으로 MSE에 페널티가 가해졌을 때에 그 값이 가장 작아지는 방식이다. Learning rate는 0.0001로 했을 때 가장 성능이 좋았다. AdaBoost의 score는 Random Forest보다 않좋았지만 RandomForest에서 낮았던 score가0.1~0.2정도 더 높았다.

2\_3 Conv1D + CuDNNGRU + CuDNNLSTM

가장 오랜 시간을 들여 구성했지만 생각보다 높은 성능이 나오지 않아 실망했던 모델이었다. 아무래도 Y의 값을 1과0으로 바꿔서 그런 것 같다고 생각했다. 하지만 그럼에도 사용한 이유는 머신 러닝 모델이 score가 낮은 부분이 많았었는데 딥러닝 모델은 평균적으로 score가 잘 나왔기 때문이다. 특히 머신 러닝에서 score가 낮은 부분이 높았던 모델이기도 하다.

우선 딥러닝 모델에서 Conv1D와 GRU,LSTM을 쓰기 위해서는 차원이 하나가 더 필요했다. 원래 Window로 날짜마다 끊으려고 했지만 테스트셋의 형식이 달라져서 불가해서 한 개의 종목마다 끊기로 했다.





Conv1D와 LSTM,GRU를 사용한 이유는 Conv1D에서 데이터들 사이에 특징을 추출해내면 LSTM과 GRU가 단기 기억과 장기 기억을 반복하며 결과 값을 도출해내는 형태여서 이 데이터셋에 적합하다고 생각했기 때문이다.

딥러닝 모델은 Conv1D에서 특징들을 추출해서 CUDNNGRU와 CuDNNLSTM을 거쳐서 결과 값을 도출하게 구성했다. 여기서 많은 하이퍼 파라미터의 수정이 필요했다. Dropout은 오히려 학습에 방해가 돼서 설정하지 않았다. 그리고 많은 Activation function을 사용해보았지만 Conv1D 와 Dense에는 LeakyReLU가 LSTM과 GRU에는 GELU가 가장 성능이 좋게 나왔다.

2\_4 Deep learning과 Machine Learning Ensemble

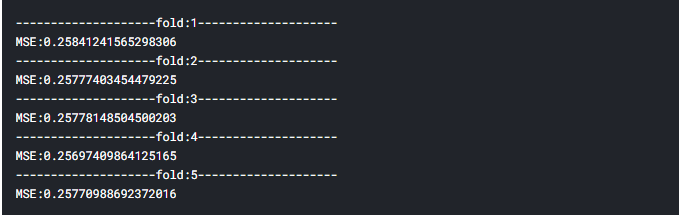


final\_pred = (Deep\_learning\_model+Adaboost+Randomforest)/3

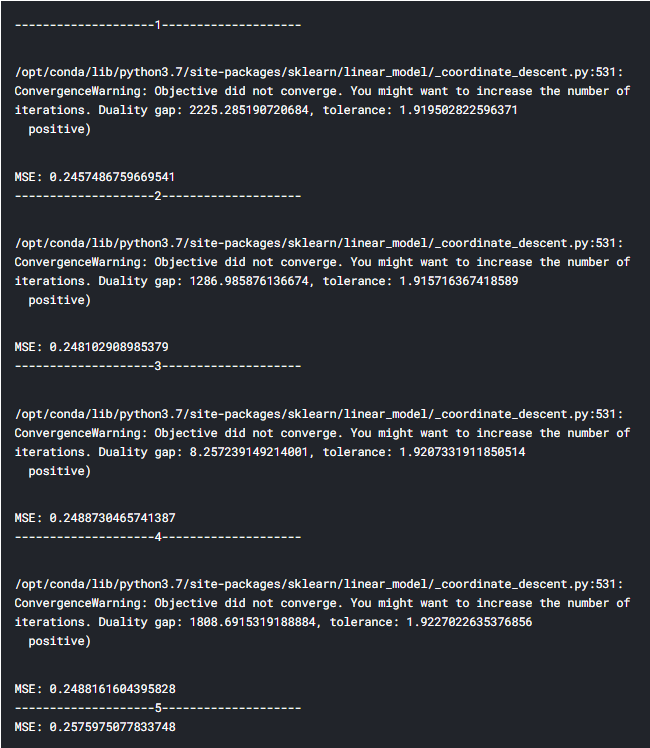
이렇게 ensemble을 진행했다. 사실 RandomForest에서 높은 스코어가 나와서 RandomForest만 score로 사용할 수 있었지만 투자의 목표가 돈을 버는 것이지만 잃지 않으면서 돈을 버는게 더욱 효과적이라고 생각했기에 ensemble을 진행해서 안정적인 score를 도출했다.

## **Ⅲ. 분석 결과**

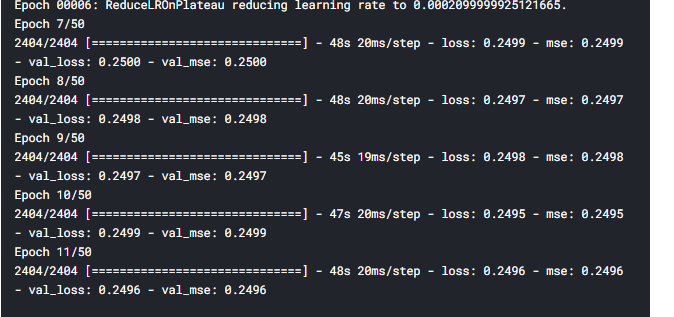
□ Random Forest



□ AdaBoostRegression



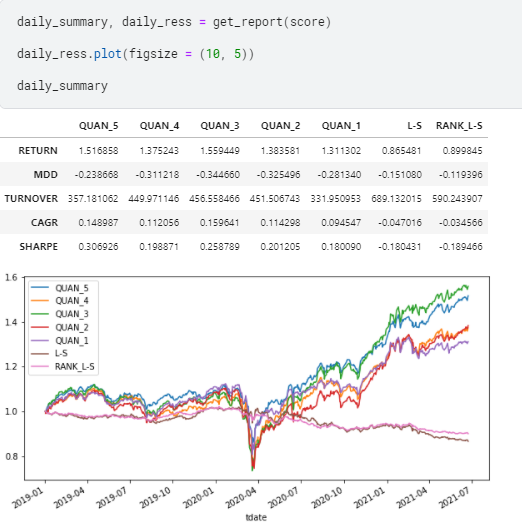
□ Deep learning

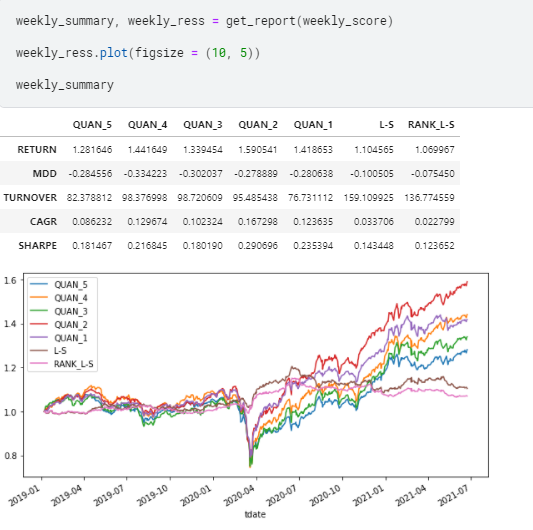


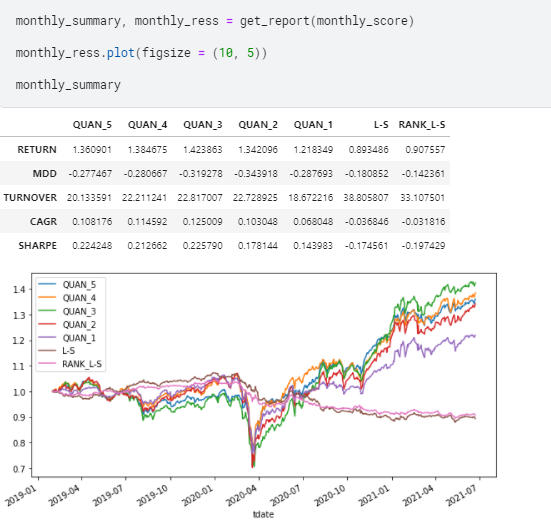
(측정 마지막 부분)

분석결과

거의 모든 모델들에게서 MSE는 0.24~0.25정도로 나왔다. MSE가 0.24밑으로 떨어지지 않길래 문제점을 찾아보니 Y값의 변환이었다. Y값을 변환하지 않았을 때에는 MSE가 0.003으로 거의 0에 수렴했다. 하지만 Y의 값을 변환하지 않으면 스코어가 나오질 않아서 아쉬웠다.

daily\_summary

weekly\_summary

monthly\_summary

Daliy와 monthly 에서 지표가 그렇게 좋지는 않았다.

하지만 weekly에서 quan\_2와quan\_1이 다른 것들보다 높게 나왔다.