통계 및 머신러닝 모델과 딥러닝 모델로 예측한 주가의 예측성능 비교

(통계학과 안희재, 통계학과 이혜원, 통계학과 김서영)
Department of Statistics, Sookmyung Women's University

1. Introduction

주식의 가격을 예측하고자 하는 노력은 계속되고있다. 과거에는 정성적, 정량적인 지표를 반영한 전통적인 방법으로 주가를 예측해왔지만, 기술이 발달함에 따라 머신러닝과 딥러닝을 통한 다양한 방법론을 제시하고 있으며 예측의 정확도 또한 높아지고 있다.

그 중 딥러닝은 뛰어난 예측 성능을 보이나 블랙박스 문제를 해결할 수 없다는 단점이 있다. 따라서 모델의 예측력이 높게 나왔더라도 이에 대한 원인과 결과를 명확히 규명하기가 어렵다. 이러한 블랙박스 특징 때문에 모델 성능의 인과관계와 변수의 영향력 등을 해석하기 어렵고, 모델 학습 진행 과정에 사람이 개입하여 조정할 수 없다. 반면 통계 모델과 머신러닝 모델의 경우 모델의 예측 결과에 대한 논리적인 설명이 가능하기 때문에, 실제 투자를 진행하는 투자자의 입장에서 통계, 머신러닝 모델이 더 설득력 있게 다가올 수 있다.

따라서 본 프로젝트에서는 통계 및 머신 러닝 모델과 딥러닝 모델을 사용하여 주가 예측을 진행한 후, 오류지표를 기준으로 성능이 좋았던 몇가지 통계 및 머신러닝 모델들과 딥러닝 모델을 통계적인 방법과 포트폴리오로 비교해보고자 한다.

2. Background

1) 동일비중 포트폴리오 (Eqaully Weighted Portfolio)

: 투자 종목들의 비율을 모두 동일하게 분산 투자 하는 방식

2) 최소 분산 포트폴리오 (Global Minimum Variance Portfolio)

: 효율적 투자선을 바탕으로 가장 위험이 낮은 지점의 가중치로 분산 투자 하는 방식

3) 최대 샤프지수 비율 포트폴리오 (Max Sharpe Ratio Portfolio)

: 효율적 투자선을 바탕으로 리스크당 수익률이 가장 큰 지점의 가중치로 분산 투자 하는 방식

4) 포트폴리오 리밴런싱 (Re-balancing)

: 포트폴리오안에 있는 자산들의 비중을 조절하는 과정

3. Data Interpretation

- **Data Scrapping**

'네이버 금융'에서 15개의 종목(삼성전자, SK하이닉스, NAVER, 씨젠, 우리들휴브레인, 현대차, DGB금융지주, 미스터블루, 셀트리온, 데일리블록체인, 소리바다, 한화솔루션, 아모레퍼시픽, CJ대한통운, GS건설)의 일별시세표를 스크래핑하였다. 일별시세표는 날짜, 종가, 전일비, 시가, 고가, 저가, 거래량으로 구성되어 있다.

- **EDA**

상장시기가 2000년 이후인 종목에 대해서 결측 값은 NaN으로 채워주었다. 15개 종목의 시계열 그래프를 그려 종목별 액면분할과 거래정지 등의 특징을 살펴보았다.

4. Experimental Method

- **Method 1) Portfolio**

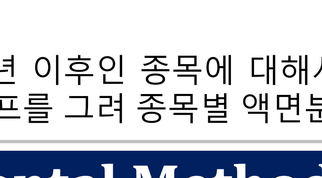
- Model Selection

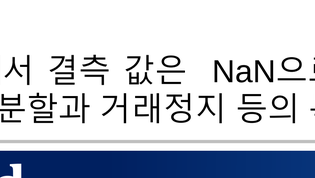
통계모델과 ML 적용시에는 2013년 1월 2일부터 2019년 8월 30일까지의 데이터를 8:2 비율로 Training data, Validation data로 나누었고, 2019년 4분기의 데이터를 Test data로 설정하였다.
딥러닝 적용 시에는 6:2:2 비율로 학습을 위한 데이터셋을 구성하였다.

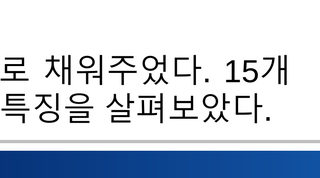
최종 모델을 선정하기 위해 SK하이닉스 데이터에 대하여 Moving Average, Linear Regression, Polynomial Regression, Linear Regression With Different Independent Variables, KNN, Hidden Markov Model, Prophet, ARIMA, RandomForest, Support Vector Machine(Linear model), Support Vector Regression(RBF kernel), LSTM을 적용시켰다. 성능 파악을 위해 RMSE와 MAPE를 오류지표로 사용하였다.

위 모델 중 예측 성능이 좋았던 MA, SVR(RBF kernel)과 추세를 잘 반영한 Prophet, LSTM의 파라미터를 다양하게 조정하여 가장 높은 성능을 보인 모델, 총 4가지를 최종 모델로 선정했다. RandomForest도 성능이 좋았지만, 동일한 상수 값으로 예측을 하는 것을 관찰할 수 있었다. 따라서 최종모델에서 제외하기로 결정하였다.


Moving Average


RBF kernel


Prophet


RandomForest

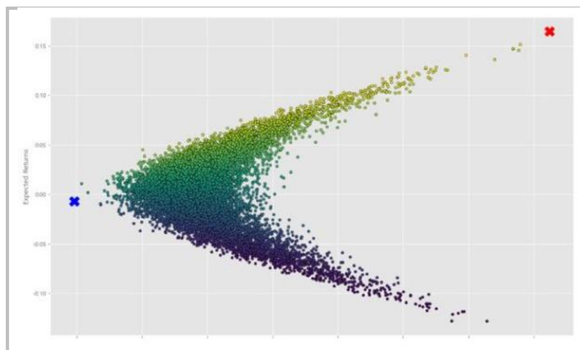
- Prediction

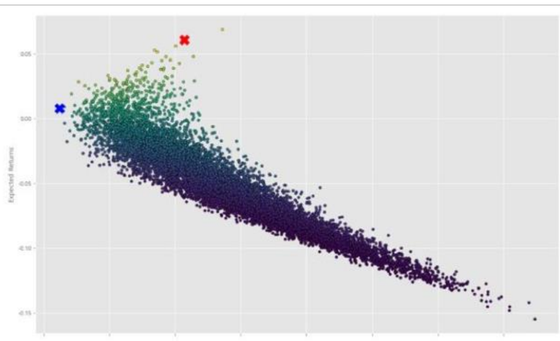
위에서 선택한 최종 모델 MA, SVR(RBF kernel), Prophet, LSTM으로 2019년 8월 30일까지의 데이터를 학습데이터로 사용하여 15개 종목의 2019년 4분기 주가를 예측하였다.

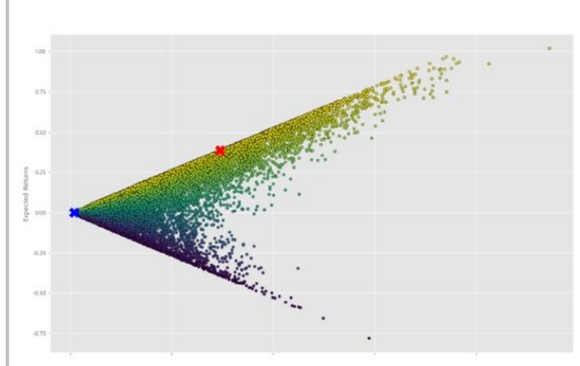
- Equally Weighted Portfolio

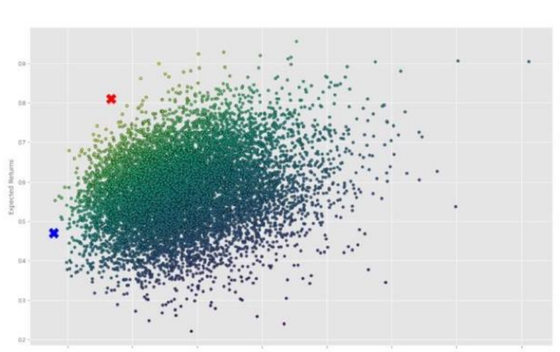
위에서 선택한 최종 모델 MA, SVR(RBF kernel), Prophet, LSTM으로 2019년 8월 30일까지의 데이터를 학습데이터로 사용하여 15개 종목의 2019년 4분기 주가를 예측하였다.

Model	Return	Volatility
Actual	0.749208	0.162030
Moving Average	-0.051136	0.009137
SVR (RBF kernel)	0.169570	0.009137
Prophet	-0.051136	0.009137
LSTM	0.785706	0.126812


Moving Average


Prophet


Support Vector Regression (RBF kernel)


LSTM

- Global Minimum Variance Portfolio

Model	Return	Volatility
Actual	0.303123	0.126622
Moving Average	-0.006623	0.003919
SVR (RBF kernel)	0.000303	0.000409
Prophet	0.007778	0.004472
LSTM	0.470363	0.095709

- Max Sharpe Ratio Portfolio

Model	Return	Volatility
Actual	0.773325	0.153477
Moving Average	0.164476	0.018482
SVR (RBF kernel)	0.386924	0.014756
Prophet	0.060980	0.060980
LSTM	0.810191	0.113328

15개의 종목으로 만든 3종류의 포트폴리오 모두 LSTM모델이 실제 데이터 값으로 만든 포트폴리오의 예상수익률과 리스크가 가장 비슷한 결과를 보여주는 것을 알 수 있었다.

- **Method 2) Diebold-Mariano test**

종목	통계량값	P-value	검정결과
삼성전자	-3.111140	2.583372e-03	귀무가설기각
SK하이닉스	19.516014	3.712042e-32	귀무가설기각
NAVER	41.564330	6.283349e-56	귀무가설기각
씨젠	18.469187	1.384513e-30	귀무가설기각
우리들휴브레인	12.656597	8.626462e-21	귀무가설기각
현대차	15.020230	5.371991e-25	귀무가설기각
DGB금융지주	7.536632	6.498088e-11	귀무가설기각
미스터블루	9.235813	3.023595e-14	귀무가설기각
셀트리온	7.615419	4.564167e-11	귀무가설기각
데일리블록체인	9.652084	4.619869e-15	귀무가설기각
소리바다	8.893185	1.424351e-13	귀무가설기각
한화솔루션	13.086934	1.405892e-21	귀무가설기각
아모레퍼시픽	13.368917	4.333999e-22	귀무가설기각
CJ대한통운	10.991570	1.168750e-17	귀무가설기각
GS건설	20.063072	5.889716e-33	귀무가설기각

15개 종목에 대해 가설 검정을 진행해 보았을 때, 유의수준 0.05하에서 모두 LSTM모델의 예측 성능이 우수하다는 것을 알 수 있었다.

5. Conclusion

본 프로젝트에서는 설명가능한 머신러닝 모델과 그렇지 않은 딥러닝 모델을 비교해보고자 했다. 머신러닝 모델과 딥러닝 모델의 성능이 통계적으로 유의미하게 차이가 나지 않는 경우라면, 머신러닝 모델을 사용하는 것이 딥러닝 모델을 사용하는 것보다 분석적인 측면에서는 더 우수할 것이라고 생각했다. 선행 연구들에서 머신러닝 모델과 딥러닝 모델을 오류지표들을 통해 비교해보고 딥러닝 모델의 성능이 좋기 때문에 딥러닝 모델을 사용하는 방식을 따라가고 있지만, 본 프로젝트에서는 딥러닝 모델이 통계적인 방법으로 검증하였을 때와 포트폴리오를 통해 검증해 보았을 때 모두 딥러닝 모델이 우수하다는 결론을 얻을 수 있었다.

6. Reference

Harvey, D., Leybourne, S., \& Newbold, P. (1997). Testing the equality of prediction mean squared errors. International Journal of forecasting, 13(2), 281-291
파이썬 증권 데이터 분석 (김황후 지음, 한빛 미디어, 2020)
https://github.com/HvyD/HMM-Stock-Predictor
https://randerson112358.medium.com/predict-stock-prices-using-python-machine-learning-53aa024da20a
https://www.analyticsvidhya.com/blog/2018/10/predicting-stock-price-machine-learningnd-deep-learning-techniques-python/