LSTM을 활용한 가치주와 성장주 분류 모형 개발에 대한 연구

왕재형[1], 김광수[2]
[1]성균관대학교 소프트웨어학과 학부생
[2]성균관대학교 소프트웨어학과 교수

jahyeng224@g.skku.edu, kim.kwangsu@skku.edu

A study on the development of a classification model for value stocks and growth stocks using LSTM

Jai-Houng Wang[1], Kwang-Su Kim[2]
[1]Dept. of Software, Sung-Kyun-Kwan University
[2]Dept. of Software, Sung-Kyun-Kwan University

요 약

가치주와 성장주의 정의는 모호하다. 그렇기에 사회에 다양한 혼란이 빚어지고 있다. 본 연구에서는 그 모호성으로 인해 생기는 혼란을 줄이고자 새로운 주식 종목 분류 모형을 제안한다. 유명 성장 지수와 가치 지수 내 종목을 통해 지도 학습이 가능한 환경에서, 종목들의 주가 등에서 새로운 지표를 만들어낸 후, 그 지표를 LSTM 모델을 통한 기계학습으로 학습하는데 활용한다. 보이지 않는 패턴을 학습한 모델을 검증기에 부착해 모호한 주식을 분류하는데 응용할 수 있다.

1. 서론

복잡한 재무지표로 일반 투자자는 투자 후보 주식이 가치주인지 성장주인지 구분하는데 어려움을 겪고 있다. 즉, 분류에 대한 정확성과 객관성이 필요하다.

또한 르네상스 테크놀로지와 같은 자산운용사 같이 퀀투 투자가 대세이다. 그런데 운용사가 펀드를 구성할 시, 주식 비율을 결정한 후 헷지 수단으로 가치주와 성장주에 일정 비율이 자동으로 투자되게 한다. 이때, 투자 후보 주식에 대한 가치주와 성장주로의 빠른 구분이 필요하다. 즉, 빠른 속도로 분류할 필요가 있다.

이러한 문제를 해결하기 위해 주관적인 판단을 배제한 정확성과 객관성을 확보하고, 신속한 의사결정이 가능한 가치주와 성장주 분류기가 필요하다. 따라서 일 데이터인 역사적 주가 데이터(종가, 고가, 저가, 거래량)만으로 해당 주식이 성장주인지 가치주인지 classification 하는 최종 모델을 만든다.

2. 선행연구

가치주와 성장주를 다룬 선행연구에는 Fama and French(1992)가 있다.[1] 이 논문은 다양한 통계 분석과 회귀 분석을 사용하여 가치주와 성장주의 특징과 예상 수익률을 분석했다. 그 결과, BM 비율이 높은 가치주가 BM 비율이 낮은 성장주에 비해 높은 수익률을 가지는 경향을 보였다. 또한, 산업 분류별로도 예상 수익률의 차이가 존재하였으며, 소프트웨어 산업에서는 성장주가 높은 예상 수익률을 보였다. 이 논문은 가치주와 성장주를 구분하고, 이에 따른 예상 수익률을 분석하여 투자 전략을 수립하는 데 중요한 정보를 제공하며 이후에도 다양한 연구에서 이 논문의 내용이 활용되었다.

3. 분석방법

모델의 feature 에는 수정 종가, 고가에서 저가를 뺀 일 주가 Gap, 거래량 총 3 개가 있다.

우선 수정 종가는 성장주가 상승하는 경향이 있고, 가치주는 횡보하는 경향이 있어 이러한 패턴이 모델 학습에 반영될 것으로 기대된다.

다음으로 일 주가 Gap 은 성장주가 큰 변동을 보이고 가치주가 작은 변동을 보이는 점을 이용한 것이다. 그러나 scaling 과정을 거치며 그러한 요소가 preprocessing 과정에서 사라질 수 있다. 그러나 여전히 이 feature 는 유효한데, 성장주 회사가 가치주 회사에 비해 상대적으로 규모가 작은 만큼 성장주의 매

일의 Gap 변동폭이 가치주의 그것에 비해 크다는 점이 모델 학습에 반영될 수 있다.

마지막으로 거래량은 일 변동 Gap 과 마찬가지로 성장주가 주로 규모가 작은 회사로 이루어지므로, 상대적으로 거래량 변동폭이 클 것으로 예상되어 이러한 요소가 모델 학습에 반영될 것이다.

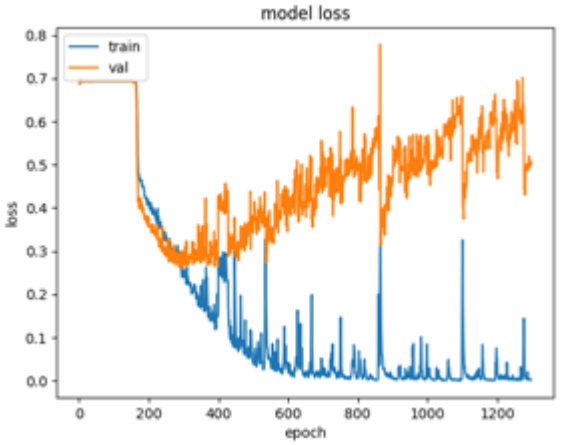
(추가 연구) 세 feature 가 합쳐진 새 지표 제작으로 모델 학습 효율 극대화 가능 여부를 확인한다.

학습데이터로 총 2,046 개 주식(성장주: 720, 가치주: 1,326)의 1,258 일치(2018 년 3 월 28 일부터 2023 년 3 월 27 일까지) 수정종가, 고가, 저가, 거래량을 수집한다. 이 수치에서 growth 와 value 펀드에 모두 소속된 주식은 제외됐다. 또한 추후 성장주는 bootstrap 으로 606 개 데이터가 추가되어 가치주와 같은 1,326 개 data 가 된다. 따라서 train, validation, test 에 사용되는 주식 데이터 개수는 2,652 개이다.

데이터는 로그 변환한다. 이는 데이터를 정규분포에 가깝게 하기 위함이다. 차분은 하지 않는다. 경향성과 계절성은 LSTM 분석에 영향을 미치지 않기 때문이다. MinMaxScaler 스케일링을 한다. 이상치 민감하지 않다는 장점이 있으나 데이터가 정규분포여야 한다. 따라서 데이터 로그 변환이 필요하다. bootstrapping 을 한다. 성장주 부족분을 해소하기 위함이다. 데이터 padding 처리한다. 각 차원의 자료개수는 동일해야 하기 때문이다. 마지막으로 모든 데이터에 대해 라벨링한다. 이때 성장주는 0, 가치주는 1 을 부여하되, one-hot encoding 이 이루어진다. 라벨링은 iShares 의 Russell 1000, 2000, Growth, Value ETF 를 바탕으로 한다.

4. 분석결과

학습 모형은 다음과 같다. LSTM layer 를 두 겹으로 쌓아 올린다. 그러나 사이에 dropout layer 가 있어 과적합을 예방한다. 마지막 layer 는 dense layer 로 여기에 활성함수 Softmax 를 대입한다. 이것은 출력값의 합이 1 이 되는 확률분포를 만들어 준다. Optimizer 가 성능 확인을 위해 loss 값으로 Binary Crossentropy 를 이용한다. 이것은 예측값과 실제값의 확률 분포 간 차이를 나타낸다. 또한 Adam optimizer 를 사용하였고, Early stopping 으로 과적합을 막고 validation 으로 학습 도중 검증이 가능하도록 했다.



(그림1) 매 epoch 마다 기록한 train 과 validation 의 loss

총 두 번의 시도가 있었고 마지막 시도의 결과가 가장 이상적이었다. 따라서 두 번째 시도로 생성된 모델을 최종 모델로 확정하였다. 이 모델은 validation set 의 loss 가 최저 0.2604 를 기록했고, test dataset predict 결과는 accuracy 0.8927 을 보였다. (추가 연구) 모델도 학습했으나, 낮은 accuracy 와 개선되지 않은 학습 속도로 feature 가 합쳐지지 않은 최종 모델이 더 유용하다는 점을 확인했다.

5. 결론

주식의 일데이터 만으로도 약 90% 정확도의 가치주, 성장주 분류기를 만들 수 있다. 향후 퀀트 투자나 포트폴리오 구성시 본 연구의 모델을 임베드하여 신속한 의사결정을 도울 것으로 기대한다.

한편 다음과 같은 후속 연구가 가능하다. 우선 한국 주식 데이터 중 라벨링이 잘 되어있는 펀드가 있다면 본 연구의 모델을 한국 주식에도 적용하여, 한국 주식에도 적용 가능한지, 한미 양 시장의 동질성 있는지 분석할 수 있다.

그리고 약 880 여개 중복 주식을 이 모델은 어떻게 분류했고, 특히 논란 있었던 주식은 최종 모델이 어떻게 분류했는지 직접 확인하고 분석할 수 있다.

마지막으로 현 최종 모델은 직접 데이터를 구하고 가공해서 모델에 넣어야 분석이 가능하나, 단지 주식명만 입력하면 알아서 데이터를 찾고 가공해서 분류해주는 시스템을 개발할 수 있다.

참고문헌

[1] Eugene F. Fama, Kenneth R. French. The Cross-Section of Expected Stock Returns. The Journal of Finance, Vol. 47(No. 2), pp. 427-465. 1992.