마이데이터를 활용한 주식 추천모델 개발

김예진[1], 임성하[2], 성승연[3], 김효재[3], 류상욱[4]
[1]연세대학교 천문우주학과
[2]서울대학교 통계학과
[3]숭실대학교 산업정보시스템공학과
[4]인하대학교 수학과
ay0119@yonsei.ac.kr, sh11972@snu.ac.kr, syseong0@naver.com,
rlagywo0722@gmail.com, sw4968@naver.com

Development of Stock Recommendation Model Using Personal MyData

Ye-Jin Kim[1], Seong-Ha Lim[2], Seung-Yeon Seong[3], Hyo-Jae Kim[3], Sang-Uk Ryu[4]
[1]Dept. of Astronomy, Yonsei University
[2]Dept. of Statistics, Seoul National University
[3]Dept. of Industrial and Information Systems Engineering, Soongsil University
[4]Dept. of Mathematics, Inha University

요 약

2030 세대의 주식시장 참여율은 갈수록 늘어나는 데에 반해 증권업의 높은 진입 장벽과 부족한 정보로 손실을 보는 경우가 적지 않다. 이러한 상황에서 정보의 주체인 개인이 본인 데이터에 대한 권리를 가지고, 본인이 원하는 방식으로 데이터를 관리하는 패러다임인 '마이데이터' 서비스가 최근 떠오르고 있다. 본 논문에서는 위의 문제를 해결하기 위하여 마이데이터를 토대로 특정 주식 종목과 유사한 종목들을 먼저 선별한 후 순수익이 높게 예측되는 종목을 최종 추천하는 모델을 제안한다.

1. 서론

디지털 시대에 거부감이 없는 2030 세대가 빠른 적응력과 높은 정보 접근성을 토대로 투자 트렌드를 이끌고 있다. 기성세대와 달리 자신의 관심사를 고려해 다양화된 투자법과 도전적인 투자 문화를 형성하고 있다는 점은[1] 주식 직접투자가 30%로 가장 많은 비율을 차지하는 M세대의 투자자산별 선호도에서도 알 수 있다[2]. 더불어 코로나 팬데믹과 함께 전체 증권계좌에서 40세 미만 청년층의 계좌 수가 38%까지 치솟을 정도로 2030 세대의 주식시장 참여가 급증하였으나[3], 기존 증권업 서비스의 한계와 상대적으로 높은 진입 장벽 및 정보 불균형으로 인하여 원하는 만큼의 이득을 보는 경우는 많지 않다. 이러한 상황에서 정보의 주체인 개인이 본인 데이터에 대한 권리를 가지고, 본인이 원하는 방식으로 데이터를 관리하고 처리하는 패러다임인 '마이데이터' 서비스가 떠오르고 있다[4].

단순히 수익성이 높은 주식 종목을 추천하는 것도 하나의 방법일 수 있으나, 수익률의 변동성과 왜도가 높은 '복권형 주식'과 같이 편향된 결과를 도출한다면 오히려 저조한 투자성과를 불러올 수 있다[5]. 이러한 점을 고려하여 본 논문에서는 유사도 등의 평가 척도를 추가로 사용하여 개인이 보유한 포트폴리오에 최적화된 맞춤 추천 서비스를 제공하는 방안을 제시한다. 최종적으로는 청년층 신규 투자자들을 주요 대상으로 필요한 정보를 제공하고, 진입 장벽을 낮추며 마이데이터 산업의 필요성과 편리함을 알리는 것을 목표로 한다.

2. 모델에 사용된 알고리즘

2.1 주식 추천모델의 방향성

마이데이터를 이용한 주식 상품 추천 서비스는 이미 여러 시중은행에서 시행한 바 있으나 대부분 단순한 종목 표시나 사용자의 포트폴리오 평가에 가깝다. 본 논문에서는 종목의 수익성을 최우선으로 두지 않고 종목들의 특성들을 이용하여 앞으로의 구체적인 투자 종목을 제시하는 것에 초점을 둔다.

구현하려는 모델은 사용자의 마이데이터를 분석하여 사용자가 보유한 주식 종목들의 특성과 유사한 주식 종목들을 제안하는 유사도 계산 단계를 먼저 거친 후, 사용자가 주식 구매에 사용할 수 있는 금액 범위 내에서 최대한 많은 주식을 구매했을 때의 순이익을 예측하는 순이익 예측 단계를 거친다. 두 단계를 통하여 사용자의 보유 주식 종목과 유사하되

순이익이 높을 것으로 예측되는 종목들을 사용자에게 추천하는 방식으로 모델이 구현된다.

2.2 모델 작업에 활용한 데이터

모델 작업에는 KOSPI 주식 종목 중에서 2019년 01월 01일부터 2021년 12월 31일까지 휴장일을 제외한 모든 날짜의 정보들이 빠지지 않고 전부 기록되어 있는 종목 191개를 이용하였다. 주식시장에서 주식의 내재가치 분석 기준으로 재무제표 정보를 많이 사용하고 있고, 이를 활용한 투자 전략과 시장 수익률 달성 여부에 관한 연구가 진행되기도 하였다[6]. 따라서 해당 종목들의 재무제표 정보 중, 결측치가 없는 PER, PBR 2개의 데이터를 활용하였다. 또한, 마이데이터 금융투자업 표준 API 규격에 따라 고객의 (종목번호, 종목명, 거래일시, 거래단가, 거래 수량) 정보를 마이데이터로 활용하였다. 실제 거래 기록을 활용할 수 없어 샘플링을 통해서 가상의 마이데이터를 구축하였는데, (거래일시, 거래 수량)은 KOSPI200 중 무작위로 선정한 10개의 종목마다 생성하였다. (거래일시)는 위의 기간을 모집단으로, (거래 수량)의 경우 [1,5]범위의 정수를 모집단으로 하여 복원추출하였다. (거래단가)의 경우 해당 거래일시의 종가를 사용하여 구축하였다.

2.3 유사도 계산 단계에서 고려한 모델

유사도 계산 단계에서 고려했던 모델은 DTW 기법과 딥러닝 알고리즘으로 학습한 모델이다.

먼저 DTW(Dynamic Time Warping) 기법은 다른 두 시계열 데이터 간의 거리 값을 측정하여 작은 거리 값을 가지는 그래프를 유사 패턴으로 인식하는 패턴인식 기법이다[7]. DTW를 기반으로 외환시장에서 나라별 환율의 유사도를 측정한 연구가 진행되기도 하였다[8]. 따라서 일정 기간동안 종목의 전날 대비 종가 변화량, PER, PBR 시계열 데이터를 각 입력데이터로 갖는 3가지 DTW모델과 앞서 모든 지표를 반영한 1가지 DTW모델에 따라 종목 간 유사도 계산을 수행하였다.

딥러닝 알고리즘은 종목별 특징을 나타내도록 임베딩 벡터를 학습시키며 이들 간 pairwise cosine similarity를 이용하여 유사한 종목을 제안한다. 지난 5거래일의 종목별 종가의 변화량을 벡터로 입력하여 종가의 변화의 유사도가 큰 종목들은 유사한 임베딩 벡터를 가지도록 학습이 이루어진다[9].

종가의 변화량은 인공신경망을 이용하여 주가 예측 모델을 학습할 때에 사용되는 데이터이며[10], CNN과 LSTM과 같은 딥러닝 기반 모델을 사용할 때에도 종가를 변수로 사용한다[11]. 다만 유사도 계산 단계에서 고려되는 모델은 주가를 직접 예측하는 기존의 다른 딥러닝 모델들과 달리 종목의 특징 학습에 사용된다는 차이가 있다.

2.4 순이익 예측 단계에서 고려한 모델

순이익 예측 단계에서 고려했던 모델은 Prophet 모델과 ARIMA 모델이다. 먼저 Prophet 모델은 Facebook에서 공개한 시계열 예측 모델로 높은 성능을 보여준다고 알려져 있다. 입력값으로 주식 종목의 종가 데이터를 받아들여 미래의 종가를 예측하는 모델이다[12]. ARIMA 모델은 단일 시계열 데이터를 예측하는 데 쓰이는 대표적인 통계적 모델로, 차분 여부, 계절성 포함 여부 등에 따라 다양한 형태의 모델 식이 만들어진다[13].

2.5 모델 평가 및 최종 모델 선정

2.3에서 언급한 모델 후보 5개 중 1개와 2.4에서 언급한 모델 후보 2개 중 1개를 연결하면 특정 주식 종목과 유사한 주식 종목 30개를 선별하고 그중에서 높은 순이익이 예상되는 종목 10개를 추천하는 10개의 모델 후보들이 만들어진다.

모델 테스트는 사용자가 추천모델을 이용하는 시점을 2021년 12월 31일로 간주하여 진행되었다. 전체 모델은 해당 시점을 포함하여 과거 3년간의 데이터를 활용하여 유사도 계산 및 순이익 예측을 한다. 모델 이용 시점의 사용자는 단일 종목만을 보유하고 있으며, 추천모델 이용 시점으로부터 7일 후인 2022년 01월 07일 가장 높은 순이익이 예상되는 종목 하나만을 모델로부터 추천받아 이용 시점에 100만 원 내에서 최대한 구매하는 것으로 가정하였다. 종가 예측 시점이 7일 후인 것은 코로나19 유행 이후 2020년 3월부터 10월까지의 20대 이하 개인투자자의 거래회전율이 평균 5.9일인 점과 주식시장이 주말, 명절, 공휴일에 거래가 진행되지 않는 점을 고려한 것이다 [14]. 사용자가 주식 구매에 사용할 수 있는 금액이 최대 100만 원인 것은 2021년 1월에 비바리퍼블리카가 2030 1,093명을 대상으로 실시한 주식투자 금액에 대한 설문에서 100만 원 이하에 대한 응답이 가장 높았던 점을 고려한 것이다[15].

최종 모델을 선정하기 위하여 사용자가 추천받는 종목들의 2022년 01월 07일 시점의 정보를 이용하여

실제 순이익을 계산하였고, 10개의 모델 중 실제 순이익이 가장 높았던 종목을 추천한 모델에 점수를 1점 부여하였다. 주식 종목 20개를 무작위로 뽑아 테스트를 진행한 결과, 최종 모델로 'DTW(PER) + Arima' 모델이 20점 중 9점을 받아 선정되었다.

유사도 계산 모델	순이익 예측 모델	점수
DTW(종가변화량)	Prophet	0
DTW(PER)	Prophet	0
DTW(PBR)	Prophet	3
DTW(종합)	Prophet	0
DL(코사인 유사도)	Prophet	1
DTW(종가변화량)	Arima	0
DTW(PER)	**Arima**	**9**
DTW(PBR)	Arima	1
DTW(종합)	Arima	3
DL(코사인 유사도)	Arima	3

<표 1> 모델 선정 결과. 가장 높은 점수를 받은 'DTW(PER) + Arima' 모델을 최종 모델로 선정

3. 결론

본 논문에서는 사용자에게 마이데이터를 기반으로 주식 종목을 추천하되 수익성을 1순위로 하지 않고 PER 지표를 이용하여 특정 주식 종목과 유사한 종목들을 먼저 선별한 후 Arima 모델을 이용하여 순이익이 높게 예측되는 종목을 최종 추천하는 방식의 모델을 구현하였다. 앞서 2.5절에서 제안한 모델은 종목의 다양성과 수익성을 고려하였다는 점에서 추후 여러 상황에 적용이 가능할 것으로 기대된다. 예컨대 투자자는 개인의 투자 성향에 따라 유명 포트폴리오를 기반으로 맞춤형 추천을 받을 수 있다. 마이데이터와 설문지를 통해 개인의 투자 성향을 파악한 후 그에 맞는 유명 포트폴리오의 종목을 활용하여 추천함으로써 현재 개인의 포트폴리오가 이와 유사해지도록 추천 모델을 활용할 수 있을 것이다. 반면 현재 연구는 단일 종목만으로 유사도를 계산하고 사용자가 단일 종목만을 구매하는 시나리오를 가정하여 테스트를 진행했다는 한계가 있다. 나아가 여러 개의 주식 종목을 활용한 시나리오로 확장한다면 현실에 가까운 주식 종목 추천 서비스가 될 것으로 기대된다.

Acknowledgement

본 논문은 과학기술정보통신부 정보통신창의인재양성사업의 지원을 통해 수행한 ICT멘토링 프로젝트 결과물입니다.

참고문헌

[1] 글로벌 경제신문, "'복잡한 투자 NO'... MZ세대 투자 트렌드, 쉽고 빠른 '플랫폼'이 이끈다", 2022, https://www.getnews.co.kr/news/articleView.html?idxno=584801.

[2] 최영준, "MZ세대의 현황과 특징", BOK 이슈노트, 제2022-13호, pp. 8-9, 2022.

[3] 시사IN, "무엇이 2030을 '영끌'로 내몰았나", 2022, https://www.sisain.co.kr/news/articleView.html?idxno=47920.

[4] 노현주, "금융 마이데이터 도입 현황과 시사점", 보험연구원 연구보고서, 21-04권호, p. 6, 2021.

[5] 김민기, 김준석, 자본시장연구원 연구보고서, 22-02호, pp. 63-100 passim, 2022.

[6] 서일석 재무제표정보를 활용한 주식투자전략에 대한 실증연구 - PER PBR ROE EPS 중심으로 2009

[7] Berndt, D. J., & Clifford, J. (1994, July). Using dynamic time warping to find patterns in time series. In KDD workshop (Vol. 10, No. 16, pp. 359-370).

[8] Wang, G. J., Xie, C., Han, F., & Sun, B. (2012). Similarity measure and topology evolution of foreign exchange markets using dynamic time warping method: Evidence from minimal spanning tree. Physica A: Statistical Mechanics and its Applications, 391(16), 4136-4146.

[9]Rendle, S., Freudenthaler, C., Gantner, Z., & Schmidt-Thieme, L. (2012). BPR: Bayesian personalized ranking from implicit feedback. arXiv preprint arXiv:1205.2618.

[10] Göçken, M., Özçalıcı, M., Boru, A., & Dosdoğru, A. T. (2016). Integrating metaheuristics and artificial neural networks for improved stock price prediction. Expert Systems with Applications, 44, 320-331.

[11] Hu, Z., Zhao, Y., & Khushi, M. (2021). A survey of forex and stock price prediction using deep learning. Applied System Innovation, 4(1), 9.

[12] Taylor SJ, Letham B. Forecasting at scale. PeerJ Preprints 5:e3190v2, 2017

[13] 이상열, 시계열 분석, 자유아카데미, 2013

[14] 김민기, 김준석, 코로나19 국면의 개인투자자: 투자행태와 투자성과, 자본시장연구원, 2021

[15] 2030 토스 설문 응답자 90% "주식투자 지속 혹은 확대할 것"[웹사이트], (2021.01.25.). URL:https://blog.toss.im/article/millennials-stock-investment