「정보시스템연구」 제32권 제1호
한국정보시스템학회
2023년 3월, pp. 179~199

http://dx.doi.org/10.5859/KAIS.2023.32.1.179

머신러닝 기반 가치투자를 통한 주식 종목 선정 연구: 내재가치를 중심으로*

김윤승** · 유동희***

〈목 차〉

Ⅰ. 서 론

일반적으로 주식 투자의 목표는 안정적으로 시장 수익률 이상의 높은 수익을 얻는 것이다. 이를 위해서는 수익을 낼 것으로 기대되는 종목을 선정하여 매수하고, 목표 수익 이상의 주가에 도달하면 매도해야 한다. 매수 대상 종목과 매도 시점을 선정하기 위해 투자자들은 다양한 방법을 사용하는데, 주로 사용하는 주식 투자 분석 방법으로는 기술적 분석(Technical Analysis)과 기본적 분석(Fundamental Analysis)이 있다. 기술적 분석은 기업의 주가와 거래량의 과거 흐름을 파악하여 단기적 주가를 예측하는 방법이며, 기본적 분석은 재무제표를 통해 기업의 성장성 및 안정성을 파악하고 미래의 전망을 예측하는 방법으로 장기투자에 주로 활용된다(조희연, 김영민, 2003; 송현정, 이석준, 2018).

가치투자(Value Investment)는 기본적 분석을 활용한 주식 매매 전략으로, 특정 종목의 주가와 내재가치(Intrinsic Value)에는 괴리가 있으며, 이를 전제로 주가가 내재가치보다 낮은 종목을 선취매한 후 해당 종목의 주가가 내재가치에 도달하면 매도하여 차익을 얻는 투자 방식이다. 주가는 투자자들이 이용 가능한 정보

* 이 논문은 주저자의 경상국립대학교 기술경영공학 석사 학위논문을 수정하여 작성하였음.
** 중소벤처기업진흥공단 정보관리실, ggomdong@naver.com(주저자)
*** 경상국립대학교 경영정보학과 및 경영경제연구소, dhyoo@gnu.ac.kr(교신저자)

를 충분히 반영하기 때문에 내재가치와 일치한다는 효율적 시장 가설(Fama, 1970)과 반대되는 개념으로, 그 동안 많은 연구들을 통해 효율적 시장 가설에 반하는 이상 현상들의 존재가 밝혀졌고, 그 유용성이 입증된 바 있다(Basu, 1977; Fama and French, 1993; Lakonishok et al., 1994; 감형규, 1999; 장영광, 김종택, 2003; 구승환, 장성용, 2010; 장옥화, 최현돌, 2010; 홍동현 등, 2011; 이장형, 성백춘, 2012; 이관영, 2019). 가치투자 전략을 활용하면 기본적 분석 방법을 통해 내재가치를 산출한 후 매수 종목의 주가가 내재가치에 도달 시 매도하는 방식으로 매도 시점을 결정할 수 있다. 또한, 현재 주가보다 내재가치가 높은 종목을 투자 대상으로 결정하므로 주가 하락이나 극단적으로는 상장폐지로부터의 위험을 최소화하고, 주가 상승이 기대되는 우량 종목을 선택할 수 있다. 하지만, 아무리 가치투자 전략으로 선정한 종목이라고 하더라도, 거시경제, 투자심리, 기타 알 수 없는 요인들로 인해 주가가 내재가치에 도달하지 않는 경우가 발생한다.

이러한 위험을 피하고 주식투자의 안정성 및 수익성 달성을 위해 본 연구에서는 머신러닝을 활용하여 미래에 주가가 내재가치에 도달할 수 있는 종목을 찾고자 한다. 이를 위해, KOSPI 및 KOSDAQ 상장 기업의 주가 및 재무 데이터를 독립변수로 하고, 과거 데이터에서 확인할 수 있는 내재가치 도달여부를 종속변수로 삼아 학습을 통해 예측 모델을 구축한다. 이후 예측 모델에 의한 종목 선정 및 임의(Random) 종목 선정 시 각각의 투자 성과를 가치투자 방식을 적용한 투자 시뮬레이션을 통해 비교한다. 예측 모델에 의해 더 좋은 종목을 선택하고, 시뮬레이션으로 도출된 우수한 전략을 채택할 수 있다면, 주식 투자성과 개선에 도움이 될 것이다.

본 연구의 목적은 다음과 같다. 첫째, 머신러닝 기법을 이용하여 주가가 내재가치에 도달할 수 있는 종목을 찾기 위한 예측 모델을 구축한다. 다양한 실험을 통해 가장 예측 성능이 우수한 알고리즘, 학습기간 및 내재가치 도달 기간 등의 조건을 찾는다. 둘째, 구축된 예측 모델을 활용하여 투자 시뮬레이션을 수행한다. 이를 통해 21년간의 투자수익률 및 내재가치 도달 건수 등 투자성과를 확인한다. 셋째, 예측 모델을 통한 투자와 임의의 종목을 선정한 투자 및 시장(KOSPI, KOSDAQ) 지수와의 비교를 통해, 본 연구에서 고안한 예측 모델의 유용성을 확인한다.

본 연구의 구성은 다음과 같다. Ⅱ장에서는 가치투자 및 머신러닝 적용 연구 등 기존 연구에 대해 살펴본다. Ⅲ장에서는 본 연구에 적용된 연구 프레임워크, 데이터와 전처리 방법을 소개한다. Ⅳ장에서는 머신러닝을 활용한 내재가치 도달 종목 예측 모델 구축과 투자 시뮬레이션 결과를 제시하며, Ⅴ장에서는 결론과 한계점, 향후 연구방향 등에 대해 기술하고자 한다.

Ⅱ. 이론적 배경

2.1 가치투자 관련 연구

Fama(1970)는 효율적 시장 가설(EMH: Efficient Market Hypothesis)을 통해, 주가는 투자자들이 이용 가능한 거래정보, 공적정보, 내부정보에 의해 랜덤하게 움직이기 때문에 예

측이 불가능하며, 이를 기반으로 한 투자전략은 초과수익률(시장수익률 이상)을 달성하기 어렵다고 주장하였다. 하지만, 이후 발표된 많은 연구들은 효율적 시장 가설로 설명되지 않는 이상현상(anomaly)의 존재에 대해 보고하고 있다. 즉, 주가와 기업의 내재가치는 일치하지 않으며, 이 차이를 이용한 투자방식, 즉 가치투자를 통해 초과수익률을 달성할 수 있음을 의미한다. 다음은 이상현상과 가치투자에 대한 연구들이다.

Basu(1977)는 저 PER(Price Earning Ratio, 주가수익비율) 종목들에게서 가치 프리미엄이 존재함을 확인하였다. PER은 주가를 EPS (Earning Per Share, 주당순이익)로 나눈 비율로서, 낮을수록 저평가된 것이며, PER이 낮은 주식에 투자하면 더 높은 초과수익률을 거둘 수 있다고 하였다.

Fama and French(1993)의 연구에서는 시장 이상현상에 있어서 기존에 알려진 기업특성 요인들 중 기업규모와 장부가대비 시가비율(B/M)이 주식 수익률에 높은 설명력이 있다는 것을 보여주며, 이를 기반으로 CAPM(Capital Asset Pricing Model)에 이 두 가지 요인을 넣은 3요인 모형(3 Factor Model)을 제안하였다.

Lakonishok et al.(1994)은 미래 이익을 추정할 때 투자가들이 기업의 과거 이익성장을 과도하게 반영하기 때문에 일반적으로 과거 이익이 좋지 못한 기업 주가의 경우 과도하게 저평가되어 PBR(Price Book-Value Ratio, 주가순자산비율)이 낮아지는 경향이 있다고 하였다. 이러한 주식에 대한 이익이 현실화될 경우 저평가되었던 주식의 가격은 회복되어 높은 수익률을 달성하게 된다고 하였다.

감형규(1999)는 국내 주식시장에서 기업규모, PBR, PCR(Price Cash flow Ratio, 주가현금흐름비율), PSR(Price Sales Ratio, 주가매출액비율), PER 등을 이용한 역행투자전략(가치투자전략)으로 높은 투자성과를 얻을 수 있음을 확인하였다.

장영광과 김종택(2003)은 PBR, PER, PCR, PSR 등 가치비율이 높은 가치주일수록 높은 수익률을 보이는 가치 프리미엄의 존재를 확인하였으며, 가치비율과 재무 건전성을 동시에 고려하였을 경우 월등히 높은 투자성과를 확인하였다.

구승환과 장성용(2010)은 주식, 채권, 예금의 금융권 데이터를 비교한 시뮬레이션 결과, 가치투자(장기투자)를 적용한 포트폴리오가 기술적 분석을 통한 단기투자를 적용한 포트폴리오의 투자 수익률보다 월등한 수익률을 기록한다고 하였다.

장옥화와 최현돌(2010)은 국내 주식시장에서 성장주에 투자하는 것보다 PER과 PSR을 활용한 가치투자전략이 더 뛰어난 시장초과 수익률을 달성할 수 있음을 확인하였다.

홍동현 등(2011)은 기업규모, PER, PBR 등을 결합한 포트폴리오를 구성하여 성과를 계산한 결과, 저PER-저PBR-대기업으로 구성한 포트폴리오의 초과수익률이 고PER-고PBR-소기업으로 구성한 포트폴리오보다 통계적으로 유의하고 높게 나타난다고 하였다.

이장형과 성백춘(2012)은 CAN SLIM 모형(William J. O’neil, 2002)에서 가장 중요한 변수인 A(연간 주당순이익)를 통해 KOSPI 및 KOSDAQ 시장에서 수익률을 올릴 수 있음을 확인하였다.

이관영(2019)은 KOSPI 및 KOSDAQ 시장의 상장 주식을 대상으로 가치투자 전략을 실행한 결과 E/P, B/M, S/P, C/P, 기업규모 등의 재무 가치비율이 상대적으로 높은 가치주식 종목의 성과가 재무 가치비율이 상대적으로 낮은 성장주식 종목에 비해 통계적으로 유의하게 높은 수익률을 보였다. 특히 장기간 투자할수록, KOSDAQ 시장일수록 더욱 큰 차이를 보였다.

2.2 머신러닝 관련 연구

머신러닝(Machine Learning)은 인공지능의 한 분야로서, 알고리즘을 이용해 데이터를 학습하고 이를 토대로 판단이나 예측을 수행하며(이요섭, 문필주, 2017), 각종 산업에서 주요 의사결정에 활용되고 있다(정동균 등, 2021). 머신러닝은 학습 방법에 따라 크게 지도 학습(Supervised Learning)과 비지도 학습(Unsupervised Learning)으로 나눌 수 있다. 지도 학습은 입력 데이터에 대한 결과 데이터를 알고 있는 경우 이들을 학습시킨 후 새로운 입력 데이터에 대해 예측하는 기법으로, 분류(Classification) 또는 회귀(Regression) 분석에 이용된다. 대표 알고리즘으로는 의사결정나무(Decision Tree), 랜덤 포레스트(Random Forest), 인공신경망(ANN: Artificial Neural Networks), K-NN(K-Nearest Neighbors), 로지스틱 회귀(Logistic Regression), 나이브 베이즈(Naive Bayes), 서포트 벡터 머신(SVM: Support Vector Machine) 등이 있다. 비지도 학습은 결과 데이터를 알 수 없는 데이터에 대해 패턴과 관계를 찾아내는 학습 방법으로 군집화(Clustering) 또는 연관규칙학습(Association Rule Learning)에 이용되며, 대표 알고리즘으로는 K-평균(K-Means Clustering), Apriori 등이 있다. 이와 같이 머신러닝 기법을 주식투자에 활용한 기존 연구들을 요약하면 다음과 같다.

채명수 등(2014)은 재무정보(PBR, PER, PSR, 부채비율, ROE, 자기자본대비순이익, 자기자본대비매출액)들의 업종별 평균값을 C4.5 알고리즘을 활용해 규칙을 도출하고 포트폴리오를 구성하였을 때 정확도가 높다고 하였다.

채명수(2015)는 PBR, PER, PSR, DR, ESR, SER, ROE, EAR 등의 재무정보를 랜덤 포레스트 분류기법에 적용한 투자전략이 임의 투자 및 SVM, C4.5 알고리즘보다 우수함을 밝혔다.

김현영(2016)은 가치투자 대상 우량 포트폴리오에 인공신경망을 통한 적정 투자 시점 기법을 적용하면 정확률의 제고가 가능하며, PBR의 이동평균선과 증권 뉴스 텍스트 결합 분석을 통해 PBR 및 PER이 적정 투자 시점을 판단하는데 유의미함을 확인하였다.

이윤한과 박근수(2017)는 재무제표 데이터를 입력받아 완전 연결 신경망(Fully Connected Neural Networks), 순환 신경망(RNN), F_SCORE 모델을 비교하였을 때 머신러닝 모델의 수익률이 최대 27.1% 높았다고 하였다.

양윤석과 오경주(2018)는 인공신경망과 은닉 마르코프 통계기법을 활용한 모델이 전통적인 Fama-French-Carhart의 4-factor 모델에 의한 수익률(3.62%)보다 3.5% 이상 높은 수익률을 달성했다고 하였다.

양정우(2021)는 전통적인 가치투자 방식(Piotroski의 F_SCORE 모델)과 딥 러닝(MLP, CNN, RNN) 모델을 통한 가치투자 방식을 비교. 높은 수익률과 일관된 경향성을 보이는 딥

러닝 모델의 예측력이 F_SCORE 모델보다 좋다고 하였다.

이광현(2021)은 한국 시장을 대상으로 주식 수익률을 예측한 결과, 머신러닝을 활용했을 때 단순회귀모형에 비해 더 우월한 성능을 보였다. 특히 랜덤 포레스트와 신경망 모델이 가장 우수한 예측력을 가짐을 확인하였다.

Zhou et al.(2006)은 재무지표(ROCE, P/E, EPS, 유동성비율)에 유전자 알고리즘을 적용하여 상위 10/20개 종목을 선별, 기존 평균수익률보다 높은 투자수익률을 확인했다.

Huang(2012)은 PER, PBR, PSR, ROE, ROA 등 재무정보를 통해, 유전자 알고리즘으로 변수 선택(Feature Selection) 및 매개 변수 최적화를, SVR(Support Vector Regression)로 평균 및 누적 수익률을 계산한 결과, 시장 수익률보다 높은 수익률을 창출하였다.

Hargreaves et al.(2012)은 재무(ROE, ROA, 성장률) 및 애널리스트의 의견, 주가 등의 지표들을 활용하여, 의사결정나무와 인공신경망을 혼합한 모델을 통해 포트폴리오를 구성하여 주가지수의 상승률보다 높은 수익률을 달성하였다.

Kedia et al.(2018)은 PER 지표를 이용하여, Elbow method를 활용한 K-means 클러스터링을 통해 Sensex보다 16.21%, BSE100보다 16.89% 우수한 성능을 보였다고 하였다.

Mattos(2019)는 ROIC, ROE, EBITDA 등 재무지표를 통해, 머신러닝 알고리즘 간 성능을 비교하여, 비선형 알고리즘(Random Forest, XGBoost, SVM)이 선형 알고리즘(OLS, Lasso Regression)보다 우수함을 확인하였지만, 매수 및 보유(Buy and Hold) 전략 보다는 낮다고 하였다.

기존의 머신러닝 기술을 적용한 연구는 대부분 과거 주가의 흐름과 거래량을 이용한 주가 예측 등의 기술적 분석 위주이며, 가치투자를 위한 기본적 분석에 대한 연구는 다소 부족한 편이다. 또한, 언급한 문헌들에서는 주식 매수 후 보유기한을 대부분 1년 이내로 제한하고, 해당 기간 동안의 예측 정확률이나 수익률에 대해 분석하고 있다. 그리고 국내의 경우 KOSPI 시장만 주로 다루고 있다.

본 연구가 이러한 기존의 연구들과 비교하여 가지는 차이점을 요약하면 다음과 같다.

첫째, 머신러닝을 활용하여 재무정보를 기반으로 한 기본적 분석을 수행한다. 재무정보로 내재가치를 산출한 후, 내재가치에 도달하는지 여부를 종속변수로 삼아 독립변수와의 관계를 분석한다. 기존 연구들은 F_SCORE 모델이나 임의의 기준으로 재무정보의 값 범위를 설정하여 종목을 선별하였으나, 본 연구는 신규로 고안한 변수로 학습한 예측 모델을 통해 종목을 선별한다는 점에서 차별성이 있다.

둘째, 주식 매수 후 보유기간을 제한하지 않는다. 매수 기준은 재무정보로 설정하더라도 매도 기준은 설정하기 어렵기 때문에, 대부분의 기존 연구들은 1~2년의 보유기간 후 주식을 매도하는 방식으로 실험을 진행하였으나, 본 연구는 보유기간의 제한 없이 내재가치 도달, 즉 주가가 내재가치 이상이 되면 매도하는 방식으로 시장가치와 내재가치의 일치를 추구한다.

셋째, KOSPI 및 KOSDAQ 시장에 상장된 기업 전체를 대상으로 한다. 기존 대부분의 연구들은 상대적으로 규모가 커서 안정성이 높은 KOSPI 시장 상장 종목을 대상으로 하였으나, 본 연구는 규모나 시장의 종류, 업종, 섹터 등의

요소를 사전에 제한하지 않고, 오로지 내재가치에 도달할 수 있는 종목을 찾는 것에 목적을 두었기 때문에 KOSPI 및 KOSDAQ 시장 모두를 대상으로 한다.

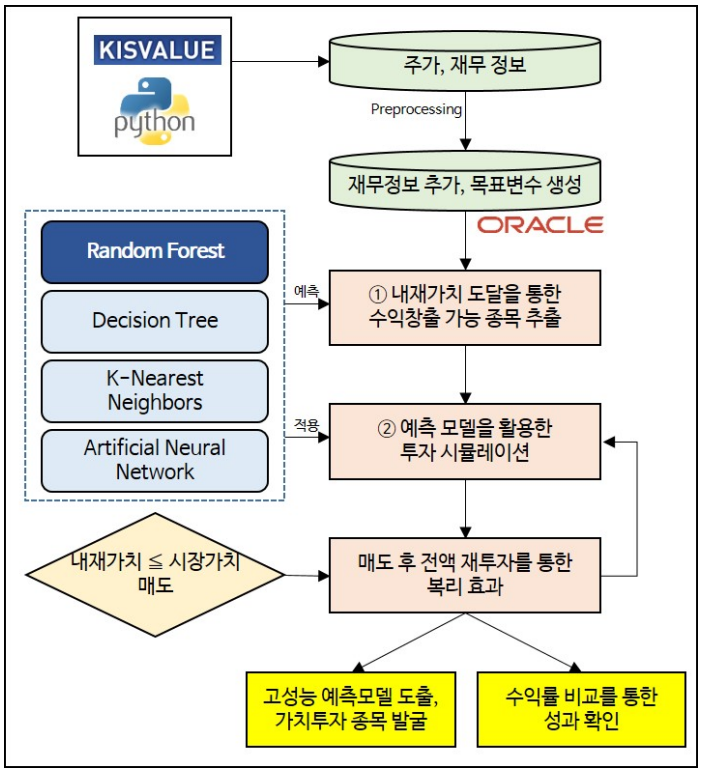
Ⅲ. 연구방법

본 연구의 연구 프레임워크는 <그림 1>과 같다. 먼저 예측 모델을 구축하기 위한 주가 및 재무정보를 수집하고, 중복 및 null 값 처리, 재무정보 추가, 변수선택을 통한 독립변수 선정 및 종속변수 생성 등 전처리를 실시하였다. 다음으로 내재가치 도달 종목을 예측하기 위해 가장 안정적이고 예측 성능이 좋은 머신러닝 알고리즘을 선정하였고, 이를 기반으로 학습기간 및 내재가치 도달 기간을 변경하면서 최종적으로 가장 성능이 좋은 예측 모델을 생성하였다.

생성한 예측 모델을 통해 선정한 종목을 매수하고 내재가치에 도달하면 매도하는 투자 시뮬레이션을 Python 프로그램을 통해 수행하였고, 그 결과를 임의로 종목을 선정하는 경우 및 시장(KOSPI, KOSDAQ) 지수와 비교하였다.

3.1 데이터 수집

본 연구에 사용된 데이터는 2021년 12월 마



<그림 1> 연구 프레임워크

지막 거래일 기준 KOSPI 및 KOSDAQ 시장에 상장된 모든 기업의 기업정보, 과거 일별 주가 및 연도별 재무정보이다. 주가정보의 수집기간은 2001년 1월 2일부터 2021년 12월 30일까지 총 21년간이며, 네이버(www.naver.com) 및 KRX(한국거래소)에서 주가정보를 스크래핑(Scraping)하는 Python 라이브러리인 Pykrx를 이용하였다. 수집한 일별 주가정보의 총 건수는 7,575,127건이다.

재무정보의 수집기간은 1998년부터 2020년까지 총 23개 년도이며, NICE신용평가정보의 KIS-VALUE를 이용하였다. t년도의 재무정보는 t+1년도에 확정되므로, 2021년에 활용할 재무정보의 기준년도는 2020년이며, 후술할 내재가치 산출을 위해서는 직전 3개년의 재무정보가 필요하므로, 투자 기준년도의 시작인 2001년의 3년 전인 1998년부터 재무정보를 수집하였다.

수집된 주가 및 재무정보를 이용해 기존 연구에서 유용하다고 알려져 독립변수로 활용 가능한 재무정보들을 추가적으로 산출하였다. 해당 변수는 일별 주가와 연동되는 PER, PBR, PCR, PSR, DIV 및 ROE 등 이다.

3.2 데이터 전처리

투자 대상 종목을 선정하기 위해 PER＜10, PBR＜1.5, ROE≥10% 등과 같은 임의 조건을 설정한 기존 연구들과 달리, 본 연구에서는 시장가치(주가)가 내재가치에 도달할 수 있는 종목을 투자 대상으로 한다. 이에 내재가치 도달 여부라는 명목형(Nominal) 데이터를 종속변수로 추가하여, 종목 매수 이후 특정 년도 12월 30일 이내에 내재가치 이상의 주가(종가)가 존재하는 경우 1의 값을, 존재하지 않는 경우 0의 값을 부여하였다. 내재가치 도달 여부는 기간에 따라 1년에서 10년으로 구분하였고, 검증을 위해 무기한 변수도 추가하였다.

$$Gi = \begin{cases} 1 : \text{매수시점부터 } i\text{년후 12월 30일 이내} \\ \quad \text{내재가치} \le \text{종가존재} \\ 0 : \text{매수시점부터 } i\text{년후 12월 30일 이내} \\ \quad \text{내재가치} \le \text{종가미존재} \end{cases}$$

$$GN = \begin{cases} 1 : \text{매수시점부터 2021년 12월 30일 이내} \\ \quad \text{내재가치} \le \text{종가존재} \\ 0 : \text{매수시점부터 2021년 12월 30일 이내} \\ \quad \text{내재가치} \le \text{종가미존재} \end{cases}$$

내재가치는, 본 연구에서 수집한 재무정보만으로 산출이 가능하고 실생활에서 활용되고 있는 방법이라는 점에서, 상속세 및 증여세법 시행령 제54조에서 규정하는 “보충적 평가방법”을 적용하여 다음과 같이 정의하였다.

$$\text{내재가치} = \text{순자산가치} + \text{순손익가치} = (\text{BPS} + \text{가중평균 EPS} \div 10\%) \div 2$$

가중평균 EPS는 각 연도별 최근 3개년 재무정보를 대상으로 하며, 최근년도에 가중치를 높게 부여한다. 3년 모두 재무정보가 존재하는 경우 3:2:1의 비율로 계산을 수행하고, 2개 년도만 존재할 경우 2:1의 비율을, 1개 년도만 존재할 경우 해당 EPS 값을 그대로 사용한다.

끝으로 Python Scikit-Learn 패키지의 지니 중요도(Gini Importance)를 이용하여 각 변수들의 중요도를 측정하고, 중요도가 낮은 변수를 하나씩 제거해 나가는 역방향 제거(Backward

Elimination) 방식의 변수선택(Feature Selection)을 통해 독립변수를 선정하였다.

3.3 실험 설계

다양한 머신러닝 라이브러리를 보유한 프로그래밍 언어인 Python 3.9를 사용하여 예측 모델 개발 및 투자 시뮬레이션을 구현하였다. 머신러닝 알고리즘으로 랜덤 포레스트, 의사결정나무, 인공신경망, K-NN, 로지스틱 회귀, 나이브 베이즈, 서포트 벡터 머신, XGBoost 등을 사용하여 예측 모델의 성능을 비교 및 확인하였고, 최종적으로 랜덤 포레스트를 선정하였다.

본 연구의 학습데이터와 검증데이터의 선정 방법은 <그림 2>와 같다. 학습데이터는 거래일자가 학습기간 내에 위치한 데이터이며, 검증데이터는 거래일자가 검증기간 내에 위치한 데이터이다. 학습데이터의 거래일자 시점에 매수 후 내재가치 도달까지 시간이 필요하기 때문에, 학습기간과 검증기간 사이에 내재가치 도달 기간을 두었다.

학습데이터의 종속변수는 실험을 통해 G1~G10 중 성능이 좋은 것을 사용하였으나, 검증데이터의 종속변수는 GN(2021년 12월 30일 이내 내재가치 도달 여부)을 활용했다. 본 연구에서는 해당 종목이 내재가치에 도달하는지 여부가 중요한 것이지, 언제 도달하는지는 중요하지 않기 때문이다.

Ⅳ. 실험결과

4.1 예측 모델 구축

4.1.1 알고리즘 선정

성능이 우수한 내재가치 도달 종목 예측 모델 구축을 위해 앞서 언급한 랜덤 포레스트, 의사결정나무, 인공신경망, K-NN, 로지스틱 회

● 학습기간, 내재가치 도달 기간, 검증기간의 범위

구분	t-n년도	t-n+1년도	t-n+2년도	…	t-3년도	t-2년도	t-1년도	t년도
학습기간	■	■	■	■	■	■		
내재가치 도달 기간	▨	▨	▨	▨	▨	▨	▨	
검증기간				…				▩

예) 학습기간 1년, 내재가치 도달 기간 1년

구분	t-2년도	t-1년도	t년도
학습기간	■		
내재가치 도달 기간	▨	▨	
검증기간			▩

예) 학습기간 2년, 내재가치 도달 기간 1년

구분	t-3년도	t-2년도	t-1년도	t년도
학습기간	■	■		
내재가치 도달 기간	▨	▨	▨	
검증기간				▩

예) 학습기간 3년, 내재가치 도달 기간 2년

구분	t-5년도	t-4년도	t-3년도	t-2년도	t-1년도	t년도
학습기간	■	■	■			
내재가치 도달 기간	▨	▨	▨	▨	▨	
검증기간						▩

<그림 2> 학습데이터와 검증데이터의 선정방법

귀, 나이브 베이즈, 서포트 벡터 머신, XGBoost 등의 알고리즘으로 예측 모델을 구축하고 성능 비교를 수행하였다. 본 연구에서는 예측 모델이 내재가치 도달 종목으로 예측한 결과 중 실제로 맞춘 비율이 중요하기 때문에 정밀도(Precision)를 성능 평가 지표로 사용하였다. 그 결과, 로지스틱 회귀, 나이브 베이즈, 서포트 벡터 머신, XGBoost는 실험 결과 정밀도가 0으로 측정되어 추후 실험에서 제외하였다.

<표 1>은 각 알고리즘을 비교한 실험 결과이다. 먼저, 학습기간 및 내재가치 도달 기간을 전체 기간 중 최초 1년(학습기간 2001년, 도달 기간 2001년~2002년)으로 고정하여 실험하고, 이어서 검증기간 기준 최근 1년으로 학습기간, 도달기간을 설정하여 실험하였다.

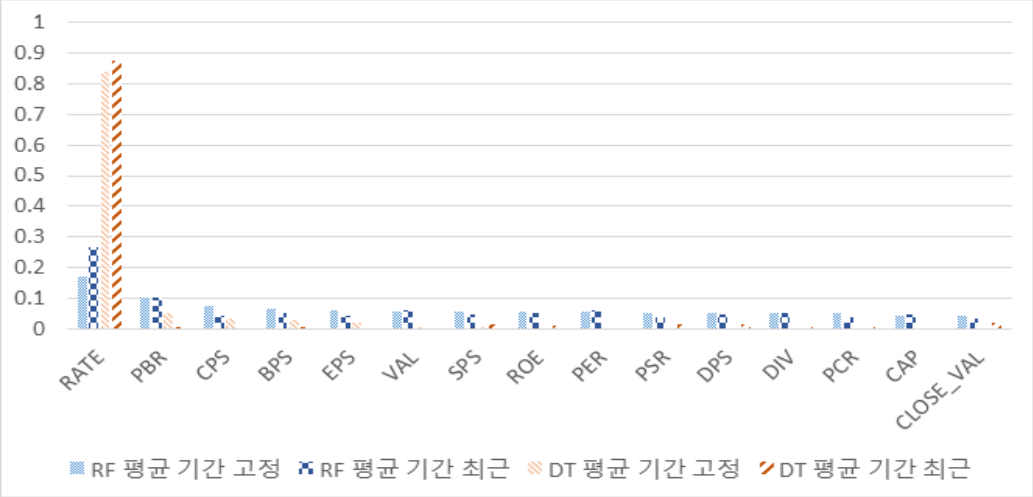
두 가지 기준 모두 랜덤 포레스트와 의사결정나무의 성능이 가장 우수하였다. 하지만, <그림 3>과 같이 각 독립변수의 중요도 평균값을 살펴보면 의사결정나무는 특정 1개의 변수(RATE)에 대한 의존도가 80% 이상으로 지나치게 높아, 본 연구에서 사용할 알고리즘은 랜덤 포레스트로 선정하였다.

4.1.2 내재가치 도달 기간 선정

가장 성능이 좋은 내재가치 도달 기간 선정을 위해 학습기간을 2001년 및 2002년으로 고정하고, 내재가치 도달 기간을 1년~10년 이내로 하여 실험하였다. <표 2>는 학습기간 2001년, <표 3>은 학습기간 2002년의 실험 결과이다.

<표 1> 각 알고리즘의 정밀도 비교

검증 기간	기간 고정				기간 최근			
	RF	DT	ANN	K-NN	RF	DT	ANN	K-NN
2003	0.9862	0.9886	0.9011	0.7389	0.9844	0.9886	0.9732	0.7389
2004	0.9902	0.9802	0.8493	0.6889	0.9880	0.9980	0.8357	0.7277
2005	0.9836	0.9787	0.8870	0.5232	0.9797	0.9614	0.8793	0.5071
2006	0.9796	0.9842	0.9216	0.4747	0.9773	0.9824	0.9695	0.5336
2007	0.9713	0.9609	0.7793	0.3348	0.9730	0.9763	0.9579	0.4102
2008	0.9682	0.9551	0.9080	0.4564	0.9472	0.9420	0.9421	0.6009
2009	0.9379	0.9568	0.8970	0.4655	0.9544	0.9660	0.9654	0.6211
2010	0.9519	0.9616	0.8826	0.4334	0.9188	0.9606	0.9676	0.5596
2011	0.9318	0.9408	0.9115	0.4387	0.9407	0.9602	0.9481	0.5682
2012	0.9388	0.9289	0.8814	0.4287	0.9387	0.9319	0.9419	0.5256
2013	0.9131	0.9149	0.8507	0.3822	0.9405	0.9480	0.9434	0.4689
2014	0.8744	0.9013	0.8381	0.3035	0.9198	0.9365	0.9408	0.3930
2015	0.8694	0.8644	0.7595	0.2341	0.8706	0.9268	0.8800	0.3204
2016	0.7962	0.8166	0.7562	0.2482	0.8039	0.8269	0.8128	0.3411
2017	0.7846	0.7989	0.7810	0.2163	0.8478	0.9018	0.8292	0.3113
2018	0.7762	0.7816	0.7710	0.2356	0.7956	0.8745	0.8287	0.3180
2019	0.7879	0.8313	0.7265	0.2731	0.8725	0.9118	0.9076	0.3356
2020	0.8103	0.8854	0.7264	0.2958	0.9344	0.9894	0.9767	0.3770
2021	0.3835	0.3736	0.2934	0.0675	0.5014	0.5166	0.4679	0.0831
평균	**0.8755**	**0.8844**	**0.8064**	**0.3810**	**0.8994**	**0.9210**	**0.8930**	**0.4601**



<그림 3> 각 변수의 중요도 평균값

<표 2> 내재가치 도달 기간별 정밀도 비교(학습기간 2001년)

년도	1년	2년	3년	4년	5년	6년	7년	8년	9년	10년
2001	학습기간									
2002										
2003	0.9873									
2004	0.9730	0.9792								
2005	0.9642	0.9660	0.9747				내재가치 도달 기간			
2006	0.9804	0.9777	0.9747	0.9714						
2007	0.9603	0.9620	0.9645	0.9374	0.9321					
2008	0.9630	0.9557	0.9476	0.9364	0.9345	0.9064				
2009	0.9299	0.9282	0.9223	0.9231	0.9070	0.8916	0.8989			
2010	0.9478	0.9279	0.9300	0.9018	0.8928	0.8694	0.8811	0.8744		
2011	0.9198	0.9132	0.9053	0.8995	0.8904	0.8665	0.8661	0.8655	0.8526	
2012	0.9057	0.8944	0.8882	0.8726	0.8660	0.8491	0.8389	0.8449	0.8375	0.8309
2013	0.8863	0.9003	0.8840	0.8766	0.8635	0.8283	0.8299	0.8298	0.8217	0.8236
2014	0.8802	0.8581	0.8703	0.8563	0.8533	0.7967	0.7994	0.8072	0.7950	0.7968
2015	0.8556	0.8344	0.8397	0.7919	0.7817	0.7473	0.7511	0.7475	0.7362	0.7419
2016	0.7909	0.7774	0.7775	0.7636	0.7532	0.7219	0.7288	0.7256	0.7221	0.7169
2017	0.7551	0.7408	0.7343	0.7127	0.7030	0.6688	0.6721	0.6689	0.6582	0.6623
2018	0.7797	0.7298	0.7355	0.6810	0.6693	0.6532	0.6458	0.6491	0.6459	0.6424
2019	0.7660	0.7250	0.7318	0.6985	0.6860	0.6673	0.6696	0.6655	0.6663	0.6630
2020	0.7751	0.8121	0.7766	0.7476	0.7212	0.6952	0.6982	0.6970	0.6935	0.6947
2021	0.3474	0.3204	0.3386	0.3045	0.3277	0.3004	0.2977	0.298	0.2934	0.2952
평균	**0.8615**	**0.8446**	**0.8350**	**0.8047**	**0.7854**	**0.7473**	**0.7367**	**0.7228**	**0.7020**	**0.6868**

<표 3> 내재가치 도달 기간별 정밀도 비교(학습기간 2002년)

년도	1년	2년	3년	4년	5년	6년	7년	8년	9년	10년
2002	학습기간									
2003	내재가치 도달 기간									
2004	0.9917									
2005	0.9850	0.9833								
2006	0.9764	0.9817	0.9699							
2007	0.9800	0.9829	0.9563	0.9488						
2008	0.9832	0.9644	0.9437	0.9319	0.9246					
2009	0.9399	0.9657	0.9394	0.9296	0.9126	0.8976				
2010	0.9293	0.9595	0.9040	0.8917	0.8769	0.8727	0.8762			
2011	0.9388	0.9551	0.9165	0.9127	0.8820	0.8664	0.8855	0.8757		
2012	0.9233	0.9557	0.8855	0.8760	0.8676	0.8569	0.8596	0.8498	0.8535	
2013	0.9052	0.9327	0.8868	0.8477	0.8457	0.8405	0.8342	0.8361	0.8354	0.8247
2014	0.9032	0.9027	0.8636	0.8475	0.8366	0.8260	0.8257	0.8038	0.8150	0.7937
2015	0.8503	0.8735	0.8269	0.7948	0.7825	0.7681	0.7833	0.7548	0.7491	0.7509
2016	0.8371	0.8483	0.7840	0.7747	0.7477	0.7502	0.7463	0.7261	0.7337	0.7230
2017	0.8122	0.8581	0.7389	0.7141	0.6823	0.6647	0.6773	0.6666	0.6698	0.6554
2018	0.7774	0.7976	0.7170	0.6829	0.6507	0.6500	0.6471	0.6519	0.6476	0.6472
2019	0.8100	0.8291	0.7031	0.6876	0.6678	0.6648	0.6643	0.6625	0.6566	0.6606
2020	0.8902	0.8860	0.7609	0.7307	0.6994	0.6944	0.6991	0.6930	0.6975	0.6948
2021	0.4382	0.4517	0.3416	0.3173	0.3088	0.3063	0.3051	0.2977	0.2994	0.2952
평균	**0.8817**	**0.8899**	**0.8211**	**0.7925**	**0.7632**	**0.7430**	**0.7336**	**0.7107**	**0.6958**	**0.6717**

학습기간이 2001년인 경우 정밀도 0.8615, 0.8446, 학습기간이 2002년인 경우 정밀도 0.8817, 0.8899로 내재가치 도달 기간이 1년~2년일 때 예측 모델의 성능이 가장 우수했다.

4.1.3 학습기간 선정

가장 성능이 좋은 학습기간 선정을 위해 많은 데이터를 학습시키는 경우와 최신 데이터를 학습시키는 경우로 나눠 성능을 비교하였다. <표 4>는 학습기간을 2001년 1월 2일부터 누적으로 하고, 내재가치를 1년~2년으로 설정하여 실험한 결과이다.

학습기간 1년~3년까지는 성능이 우수하지만, 이후 학습기간이 길어질수록 성능은 점점 낮아지는 것으로 확인되었다. 이를 통해 많은 데이터를 학습한다고 해서 반드시 예측 성능이 높아지는 것은 아니며, 최신의 데이터를 학습하는 것이 더 우수함을 알 수 있다. 또한, 내재가치 도달 기간은 1년인 경우 2년인 경우보다 평균 정밀도가 우수하게 측정되었다.

다음은 학습기간은 1년~3년, 내재가치 도달 기간은 1년~2년으로 하되, 알고리즘 선정 때와 동일하게 검증기간 기준으로 최근 기간을 선택하여 실험하였고, 그 결과는 <표 5>와 같다. 학습기간 1년 및 내재가치 도달 기간이 1년 일 경우 가장 성능이 좋았다. 동일한 내재가치 도달 기간에서는 학습기간이 길어질수록 성능이 나

빠지며, 내재가치 도달 기간은 1년인 경우가 2년인 경우보다 평균 정밀도가 높았다. <표 4>에서 내재가치 도달 기간 1년을 기준으로 학습기간을 누적했을 때의 성능과 비교해보면, <표 5>에서 학습기간을 검증기간 기준 최근으로 설정했을 때가 더 성능이 좋았다.

<표 4> 학습기간별 정밀도 비교(학습기간 2001년~누적)

년도	1년		2년		3년		4년		5년	
2001	학습기간		학습기간		학습기간		학습기간		학습기간	
2002	도달1년	도달2년								
2003	0.9873		도달1년	도달2년						
2004	0.9730	0.9792	0.9932		도달1년	도달2년				
2005	0.9642	0.9660	0.9671	0.9741	0.9783		도달1년	도달2년		
2006	0.9804	0.9777	0.9825	0.9874	0.9889	0.9769	0.9840		도달1년	도달2년
2007	0.9603	0.9620	0.9746	0.9741	0.9658	0.9721	0.9664	0.9635	0.9680	
2008	0.9630	0.9557	0.9699	0.9614	0.9745	0.951	0.9579	0.9521	0.9619	0.9431
2009	0.9299	0.9282	0.9383	0.9356	0.9716	0.9418	0.9439	0.9270	0.9448	0.9202
2010	0.9478	0.9279	0.9365	0.9607	0.9741	0.9223	0.9503	0.9235	0.9308	0.9122
2011	0.9198	0.9132	0.9356	0.9134	0.9629	0.9161	0.9224	0.9126	0.9161	0.9008
2012	0.9057	0.8944	0.9134	0.9037	0.9358	0.8851	0.8929	0.8751	0.9165	0.8797
2013	0.8863	0.9003	0.9186	0.9122	0.9302	0.8788	0.8793	0.8688	0.8973	0.8734
2014	0.8802	0.8581	0.8924	0.8697	0.9190	0.8599	0.8711	0.8481	0.8687	0.8680
2015	0.8556	0.8344	0.8546	0.8419	0.8866	0.8060	0.8322	0.8175	0.8272	0.8087
2016	0.7909	0.7774	0.7962	0.8104	0.8257	0.7487	0.7990	0.7579	0.8278	0.7605
2017	0.7551	0.7408	0.7930	0.7932	0.7997	0.7312	0.7393	0.7435	0.7938	0.7277
2018	0.7797	0.7298	0.7841	0.8043	0.8328	0.7074	0.7529	0.7079	0.7595	0.6984
2019	0.7660	0.7250	0.8009	0.7701	0.8627	0.7302	0.7701	0.7050	0.7639	0.7179
2020	0.7751	0.8121	0.8500	0.8328	0.8866	0.7694	0.8119	0.7472	0.8061	0.7531
2021	0.3474	0.3204	0.4225	0.4024	0.4326	0.3553	0.3304	0.3249	0.3577	0.3174
평균	**0.8615**	**0.8446**	**0.8735**	**0.8616**	**0.8899**	**0.8220**	**0.8378**	**0.8050**	**0.8360**	**0.7915**

<표 5> 학습기간별 정밀도 비교(학습기간 1~3년, 내재가치 도달 기간 1~2년)

	학습기간 1년		학습기간 2년		학습기간 3년	
년도	도달 1년	도달 2년	도달 1년	도달 2년	도달 1년	도달 2년
2003	0.9844					
2004	0.9880	0.9896	0.9921			
2005	0.9797	0.9739	0.9773	0.9799	0.9700	
2006	0.9773	0.9734	0.9832	0.9821	0.9915	0.9813
2007	0.9730	0.9565	0.9615	0.9541	0.9732	0.9683
2008	0.9472	0.9342	0.9463	0.9401	0.9464	0.9423
2009	0.9544	0.9387	0.9317	0.9352	0.9477	0.9285
2010	0.9188	0.9390	0.9287	0.9224	0.9340	0.8999

	학습기간 1년		학습기간 2년		학습기간 3년	
년도	도달 1년	도달 2년	도달 1년	도달 2년	도달 1년	도달 2년
2011	0.9407	0.9185	0.9511	0.9305	0.9490	0.9255
2012	0.9387	0.9259	0.9264	0.9171	0.9292	0.9287
2013	0.9405	0.9146	0.9387	0.9134	0.9226	0.9056
2014	0.9198	0.9108	0.9204	0.8853	0.9213	0.9096
2015	0.8706	0.8617	0.8717	0.8621	0.8792	0.8636
2016	0.8039	0.8094	0.8332	0.8010	0.8599	0.8208
2017	0.8478	0.7788	0.8170	0.7577	0.8173	0.7691
2018	0.7956	0.8012	0.8327	0.7701	0.7995	0.7622
2019	0.8725	0.8194	0.8744	0.8279	0.8739	0.8306
2020	0.9344	0.9314	0.9430	0.8847	0.9361	0.8630
2021	0.5014	0.4466	0.4972	0.4617	0.5392	0.4543
평균	**0.8994**	**0.8791**	**0.8959**	**0.8662**	**0.8935**	**0.8596**

종합해보면, 내재가치 도달 기간은 1년, 학습기간은 검증기간 기준 최근 1년일 때 가장 예측 성능이 좋은 것으로 나타났다. 내재가치 도달 기간이 짧아질수록 학습기간과 검증기간의 간격이 짧아지고, 학습기간을 누적할 때보다 학습기간이 검증기간 기준 최근인 경우 역시 학습기간과 검증기간의 간격이 짧아지기 때문에, 학습기간과 검증기간의 간격이 짧을수록 즉, 최신 데이터로 학습했을 경우 예측 성능이 좋은 것으로 해석할 수 있다.

4.2 투자 시뮬레이션

앞서 구축한 내재가치 도달 종목 예측 모델이 투자성과 개선에 유용한지 확인하기 위해 <표 6>과 같이 투자 시뮬레이션을 수행하였다. 대상기간은 2001년 1월 2일부터 2021년 12월 30일까지이고, 학습기간은 검증기간 기준 최근 1년, 내재가치 도달 기간은 1년 이내로 설정하면 검증기간 즉, 실제 투자 시작일은 2003년 1월 2일이 된다. 초기 투자금액은 100,000,000원으로 하였고, 종목별로 동일 금액을 분산 투자하였다. 예를 들어 100개 종목을 매수하면 종목당 1,000,000원이다. 투자대상 선정의 첫 번째 조건은 투자일 당시 모든 종목을 예측 모델에 입력시켰을 때, 내재가치 도달 여부가 1로 확인되는 종목이다. 두 번째 조건은 해당 종목들 중 기대 수익률이 0%, 30%, 50%, 70%, 100% 이상인 종목이며, 기대 수익률 구간마다 종목이 달라지므로, 실험을 총 5번 실시하였다. 가치투자 전략에 따라 종가가 내재가치 이상이 되면 매도하였으며, 다음 영업일에 모든 종목을 예측 모델에 입력시켜 내재가치 도달 여부가 1로 확인되고 설정한 기대 수익률 이상인 종목 중 가장 기대 수익률이 높은 신규 종목 1개에 매도 금액 전부를 투자하였다. 만약 조건에 해당 하는 신규 종목이 없을 경우, 해당 예수금은 보유하고 있다가 다음 매도 시 매도금액에 더한 후, 종목을 2개 선정하여 2개 종목에 동일하게 분산 투자하였다.

예측 모델과의 비교를 위해 임의로 종목을 선정하여 투자하는 시뮬레이션도 병행하였다.

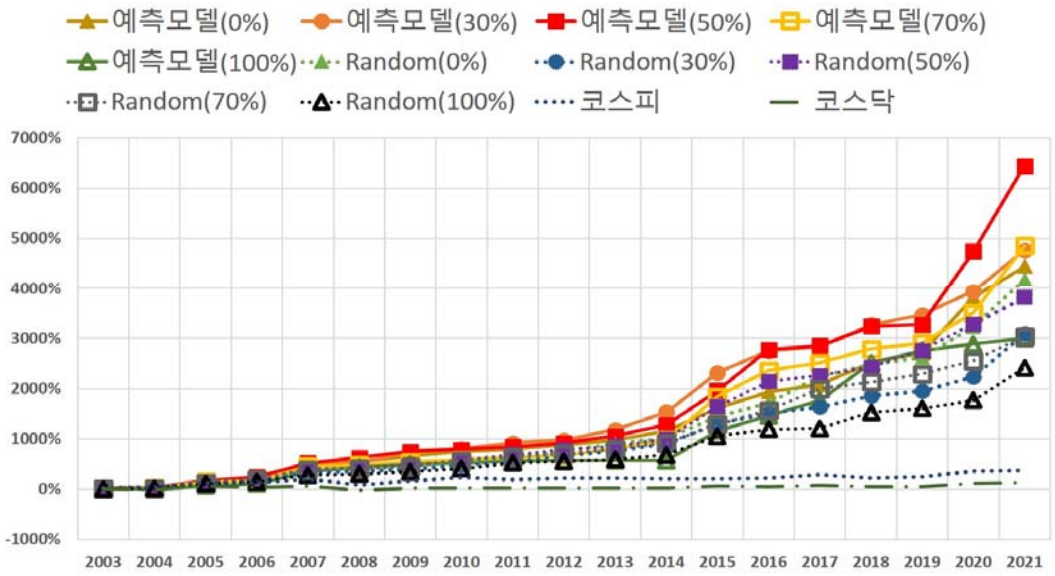
무작위의 특성상 우연에 의한 통계 왜곡을 방지하기 위해 각 수익률 기준별로 시뮬레이션을 10회 실시하고 평균을 활용하였다. 또한 기대 수익률만을 조건으로 하면 너무 많은 종목이 대상이 되므로, 각 기대 수익률 기준별로 예측 모델에서 선정한 종목 수와 동일한 수의 종목을 선정하기 위해 0% 이상은 100개, 30% 이상은 71개, 50% 이상은 46개, 70% 이상은 30개, 100% 이상은 15개의 종목을 선정하였다. 결과는 <표 7> 및 <그림 4>와 같다.

<표 6> 투자 시뮬레이션을 위한 가정

구분	내용	비고
대상기간	2001년 1월 2일 ~ 2021년 12월 30일	투자시작: 2003년~
학습기간	검증기간 기준 최근 1년	-
내재가치 도달 기간	학습기간 종료일자로부터 1년 이내	-
초기 투자금액	100,000,000원	종목별 동일 비중 분산
투자대상	내재가치 도달 여부=1	예측 모델 결과
매도조건	내재가치 ≤ 종가	종가에 매도
매수조건	매도 금액 전부로 신규 1개 종목 매수	다음 영업일 종가에 매수
기대 수익률 구간	0%, 30%, 50%, 70%, 100%	RATE 변수

<표 7> 누적 수익률

	예측 모델					임의 선정(Random)				
년도	0%	30%	50%	70%	100%	0%	30%	50%	70%	100%
2003	12%	12%	12%	6%	0%	12%	11%	8%	7%	3%
2004	28%	32%	38%	29%	9%	24%	23%	27%	16%	15%
2005	176%	157%	177%	147%	66%	124%	121%	137%	127%	106%
2006	224%	198%	241%	170%	105%	174%	163%	191%	189%	146%
2007	523%	515%	519%	449%	347%	373%	332%	408%	386%	275%
2008	543%	526%	624%	481%	436%	396%	352%	422%	413%	298%
2009	657%	721%	747%	541%	495%	434%	388%	480%	448%	346%
2010	756%	805%	789%	581%	514%	543%	487%	578%	572%	413%
2011	792%	915%	839%	629%	553%	604%	539%	690%	638%	521%
2012	880%	972%	916%	676%	566%	754%	662%	782%	711%	555%
2013	980%	1179%	1055%	805%	566%	894%	764%	852%	789%	589%
2014	1157%	1536%	1280%	982%	566%	1008%	895%	996%	913%	681%
2015	1615%	2319%	1966%	1842%	1159%	1441%	1294%	1644%	1286%	1055%
2016	1934%	2748%	2775%	2358%	1457%	1722%	1508%	2143%	1566%	1185%
2017	2079%	2829%	2847%	2507%	1770%	2238%	1640%	2260%	2000%	1204%
2018	2514%	3269%	3235%	2789%	2512%	2478%	1848%	2425%	2129%	1524%
2019	2702%	3469%	3275%	2905%	2754%	2625%	1951%	2757%	2290%	1603%
2020	3820%	3940%	4751%	3509%	2883%	3225%	2223%	3283%	2554%	1771%
2021	4447%	4767%	6446%	4858%	2996%	4189%	3088%	3815%	3037%	2410%
연평균	**25%**	**25%**	**27%**	**26%**	**23%**	**23%**	**21%**	**23%**	**22%**	**20%**



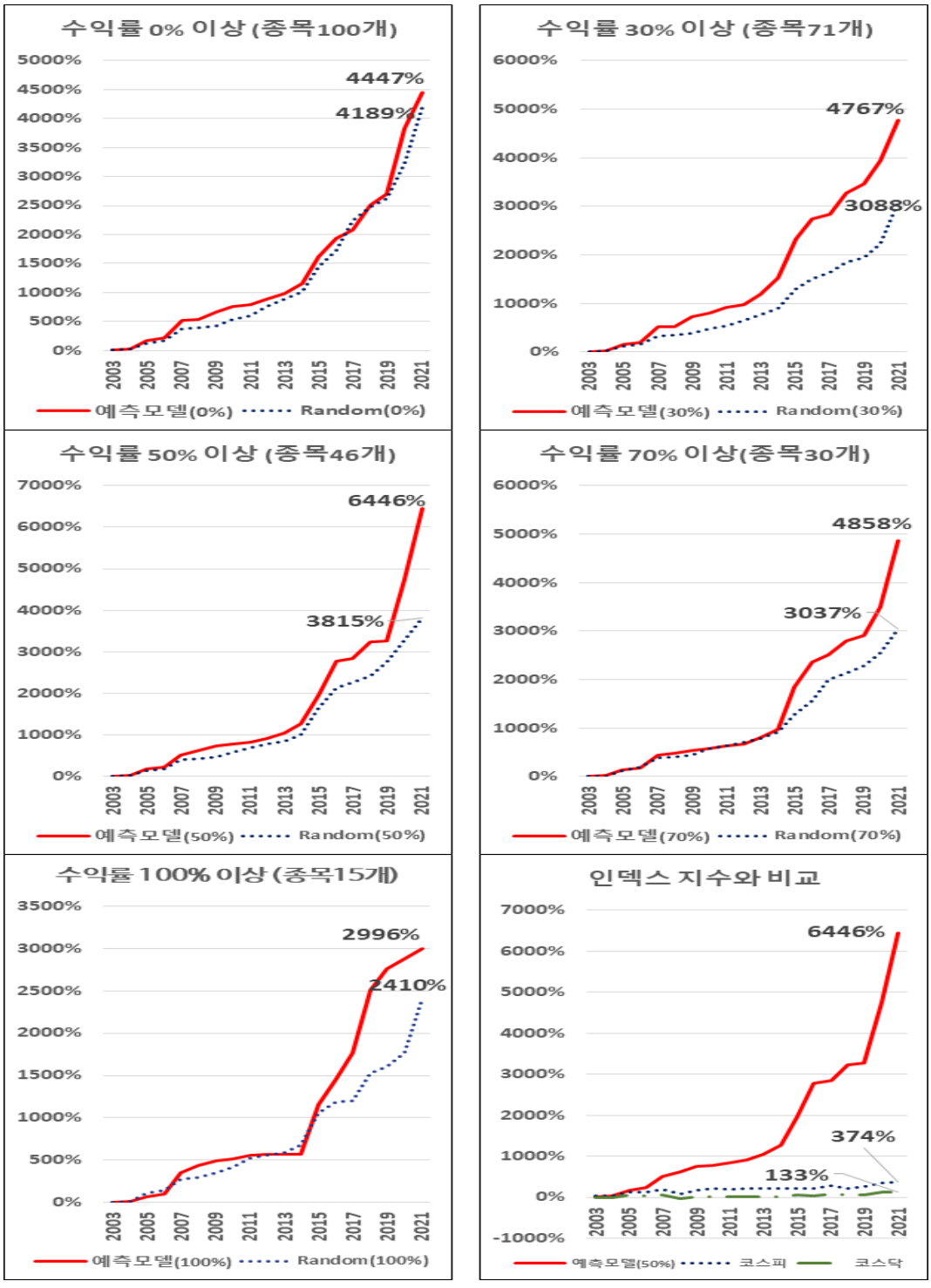
<그림 4> 누적 수익률

기대 수익률 50% 이상인 경우 예측 모델에 의한 투자의 누적 수익률이 6,446%로 가장 높게 나타났으며, 이 경우 초기 투자금액 100,000,000원은 6,445,831,353원이 된다. 동일한 조건인 기대 수익률 50% 이상인 임의 선정에 의한 투자 시 누적 수익률 3,815%와 비교하면, 예측 모델이 1.68배 이상 높은 수익률을 기록했다.

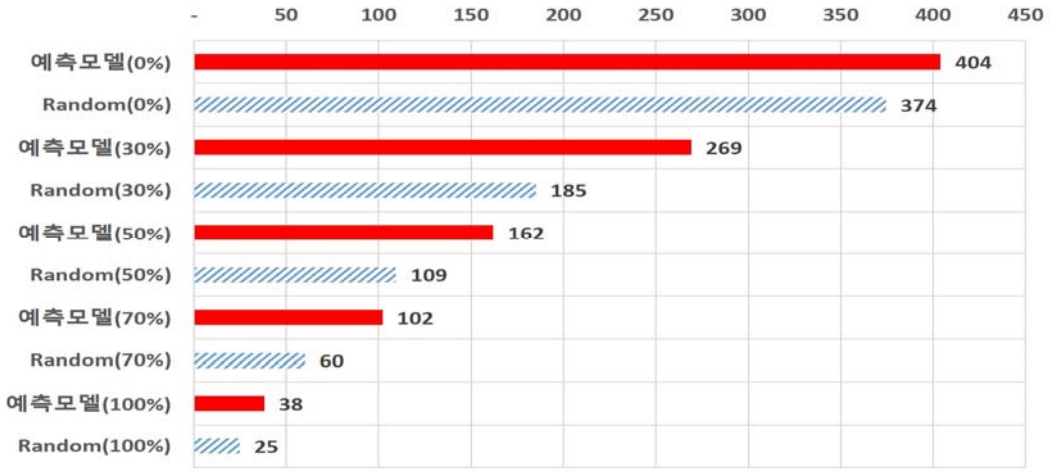
<그림 5>는 동일 기대 수익률 조건에서 예측 모델과 임의 선정 및 시장 지표 간의 비교이다. 모든 기대 수익률 구간에서 예측 모델이 임의 선정보다 우수한 성능을 보였으며, 수익률 50% 이상일 때의 예측 모델과 시장 지표를 비교해 보면, KOSPI 지수 대비 17.24배, KOSDAQ 지수 대비 48.47배의 누적 수익률을 기록했다.

수익률은 투자성과를 가시적으로 보여주는 지표이긴 하지만, 내재가치 도달 여부와 별개로 특정 종목의 기대 수익률에 따라 과대 계상될 수 있으므로, 내재가치 도달 여부에 대한 예측을 얼마나 잘했는지에 대한 판단 지표로는 다소 부족할 수 있다. <그림 6>은 예측 모델 활용 및 임의 선정으로 매수한 종목 중 내재가치에 도달하여 매도한 종목의 수를 나타낸 차트이다. 전 기대 수익률 구간에서 예측 모델이 임의 선정보다 누적 내재가치 도달 종목 수가 8~70% 이상 많다. 내재가치에 도달하여 매도한 종목 수가 많다는 것은 예측 모델이 본래 목적인 내재가치 도달 종목 예측에 성공했다는 것을 의미한다.

마지막으로, 시장 구분에 따른 누적 수익률 및 내재가치 도달 건수를 비교하였다. <표 8>은 예측 모델을 활용해 투자 시뮬레이션을 수행한 결과를 토대로 KOSPI 시장과 KOSDAQ 시장을 비교한 표이다. 내재가치 도달 건수는



<그림 5> 동일 기대 수익률 조건에서 예측 모델과 임의 선정 비교



<그림 6> 누적 내재가치 도달(매도) 종목 수

<표 8> 시장 구분에 따른 누적 수익률 및 내재가치 도달 건수

	누적 수익률			내재가치 도달 건수		
기대 수익률	KOSPI	KOSDAQ	소계	KOSPI	KOSDAQ	소계
0%	1481%	2966%	4,447%	200	204	404
30%	1900%	2867%	4,767%	134	135	269
50%	2950%	3496%	6,446%	82	80	162
70%	2056%	2802%	4,858%	53	49	102
100%	870%	2126%	2,996%	19	19	38
합계	**9,257%**	**14,257%**	**23,514%**	**488**	**487**	**975**

거의 동일했으며, 누적 수익률은 KOSDAQ 시장에서 1.54배 더 높게 나왔다. 이를 통해 규모가 크고 안정성이 높은 KOSPI 시장뿐만 아니라 KOSDAQ 시장에서도 가치투자를 통한 투자성과 달성이 가능함을 확인할 수 있다.

Ⅴ. 결 론

주식투자에 있어 안정성과 수익성을 달성하기 위한 가치투자 전략을 바탕으로, 내재가치에 도달할 수 있는 종목을 찾기 위해 KOSPI 및 KOSDAQ 상장 기업의 주가 및 재무정보를 독립변수로 하고, 내재가치 도달 여부를 종속변수로 삼아 학습을 통해 예측 모델을 구축하였다. 이후 예측 모델과 임의(Random) 선정으로 각각 추출한 종목을 대상으로 가치투자 전략을 적용한 투자 시뮬레이션을 통해 예측 모델의 성능을 확인하였다.

이를 위해 21년(2001~2021년)간 주가 및 재무 데이터를 수집, 독립변수로 활용할 재무정보 추가, 내재가치 산출, 종속변수로 활용할 내재

가치 도달 여부 추가 등 데이터 전처리를 수행하였다. 가장 좋은 성능의 예측 모델을 구축하기 위해 다양한 알고리즘, 학습기간, 내재가치 도달 기간으로 실험을 수행한 결과, 알고리즘은 랜덤 포레스트, 학습기간은 1년, 내재가치 도달 기간은 1년 이내로 하였을 때 예측 성능이 뛰어남을 확인하였다. 예측 모델의 성능을 검증하기 위한 투자 시뮬레이션 결과, 예측 모델의 투자 누적 수익률이 임의 선정에 비해 최대 1.68배 이상, KOSPI 지수에 비해 17배 이상 높았다. 또한, 예측 모델로 선정한 종목의 내재가치 도달 건수가 임의 선정보다 최대 70% 더 많다는 결과를 통해, 본 연구에서 구축한 예측 모델이 내재가치 도달 종목 예측에 유용함을 입증하였다.

본 연구의 학문적 시사점은 다음과 같다. 첫째, 머신러닝과 가치투자 전략을 접목한 예측 모델을 통해 종목 선정의 예측력을 제고함과 동시에 높은 누적 수익률을 달성할 수 있음을 확인하였다. 둘째, 기존에 많은 연구가 이루어지지 않은 중소규모 KOSDAQ 시장에 가치투자 전략을 활용할 경우, 안정성이 높은 KOSPI 시장 못지않은 투자성과를 달성할 수 있음을 보였다. 셋째, 다양한 머신러닝 알고리즘, 학습 데이터 및 검증데이터 선정 방법을 비교・검토함으로써, 예측 모델 구축 방법에 대한 다양한 관점을 제시하였다.

실무적 시사점으로는 첫째, 기존 연구에서 다루고 있지 않는 내재가치 도달 여부라는 변수를 새로이 고안하여, 투자 수익률을 개선할 수 있는 종목 선정 방법론을 제시하였다. 둘째, 본 연구의 예측 모델을 통해 매수 및 매도 가격과 기대 수익률 등 정보를 제공함으로써 투자자의 의사결정을 지원할 수 있다. 셋째, 투자자의 경제력 제고와 시장 유동성 공급을 통한 증권시장 선진화에 이바지할 수 있는 주식 정보서비스 구축의 단초를 제공하였다.

다만, 2021년 현재 상장기업만을 대상으로 하였기에 투자 후 상장폐지에 대한 고려가 없고, 예측 모델이 정밀도에 비해 재현율(Recall)이 낮아 더 많은 내재가치 도달 종목이 있음에도 투자하지 못하며, 여러 가지 내재가치 산출 방법 중 한 가지 방법만을 사용했고, 투자 시뮬레이션에서 수익 발생 후 종목별 매수 비중 조정(Rebalancing)이나 종목 교체와 같은 포트폴리오 수정 없이 신규로 예측한 하나의 종목에 모두 투자하여 비중의 불균형이라는 리스크가 발생한다는 점에서 본 연구의 한계가 있다.

따라서 향후 연구에는 대상기간 종료일이 아닌 시작일 기준 상장기업 모두를 투자 대상으로 하고, 이후 상장되는 기업을 추가하는 방식으로 상장폐지라는 요소를 고려할 필요가 있다. 또한, 재현율 향상을 위한 예측 모델 튜닝을 통해 보다 많은 내재가치 도달 종목을 예측할 수 있도록 개선하여야 한다. 그리고 본 연구에서 활용한 보충적 평가방법을 포함하여 벤저민 그레이엄, 워렌 버핏 등 가치투자의 대가들의 방법이나, DCF, RIM, S-RIM 등 기타 방법들을 비교하여 보다 성능이 좋은 내재가치 산출 방법을 확인할 필요가 있다. 마지막으로 투자 시뮬레이션에서 포트폴리오 수정을 통한 리스크 최소화 및 수익률 증대 전략을 모색하는 추가 연구가 필요하다.

참고문헌

구승환, 장성용, “최적 투자 포트폴리오 구성전략에 관한 연구”, 산업공학 (IE interfaces), 제23권, 제4호, 2010, pp. 300-310.

김현영, “기계학습 기법에 기반한 가치투자를 위한 투자시점 추천 및 뉴스텍스트 분석”, 한국과학기술원 석사학위논문, 2016.

김형규, “한국주식시장에서의 역행투자 성과에 관한 실증적 연구”, 재무관리연구, 제16권, 제2호, 1999, pp. 157-178.

송현정, 이석준, “딥러닝을 활용한 실시간 주식 거래에서의 매매 빈도 패턴과 예측 시점에 관한 연구: KOSDAQ 시장을 중심으로”, 정보시스템연구, 제27권, 제3호, 2018, pp. 123-140.

이관영, “주식시장에서 가치투자전략의 성과와 위험요인에 관한 연구: 국내 주식시장을 중심으로”, 재무관리연구, 제36권, 제2호, 2019, pp. 107-150.

이광현, “머신러닝 방법론을 통한 자산가격결정모형에 대한 실증 연구 – 한국 시장을 중심으로”, 한국과학기술원 석사학위논문, 2021.

이요섭, 문필주, “딥 러닝 프레임워크의 비교 및 분석”, 한국전자통신학회 논문지, 제12권, 제1호, 2017, pp. 115-122.

이윤한, 박근수, “딥 러닝을 이용한 가치투자 기법”, 한국정보과학회 학술발표논문집, 2017, pp. 1608-1610.

이장형, 성백춘, “EPS 를 이용한 주식 투자”, 전산회계연구, 제10권, 제1호, 2012, pp. 1-17.

양윤석, 오경주, “인공신경망과 HMM (은닉 마르코프 모델) 을 이용한 가치투자의 비선형성에 대한 연구”, 대한산업공학회 춘계공동학술대회 논문집, 2018. pp. 520-530.

양정우, “다양한 딥 러닝 모델을 활용한 미국 주식시장에서의 가치투자 분석”, 서울대학교 석사학위논문, 2021.

장영광, 김종택, “한국주식시장에서 가치투자 전략의 투자성과와 그 원천”, 한국증권학회지, 제32권, 제2호, 2003, pp. 165-208.

장옥화, 최현돌, “가치주와 장기투자성과의 관련성”, 경영연구, 제25권, 제3호, 2010, pp. 1-33.

정동균, 이종화, 이현규, “머신러닝을 이용한 국내 수입 자동차 구매 해약 예측 모델 연구: H 수입차 딜러사 대상으로”, 정보시스템연구, 제30권, 제2호, 2021, pp.105-126.

조희연, 김영민, “유전자 알고리즘을 이용한 주식투자 수익률 향상에 관한 연구”, 정보시스템연구, 제12권, 제2호, 2003, pp. 1-20.

채명수, 고동균, 김준수, 한창세, 윤완철, “의사결정나무와 업종의 특성을 고려한 가치투자 관점에서의 우량종목 선별방법 연구”, 한국정보과학회 학술발표논문집, 2014, pp. 503-505.

채명수, “가치투자 관점에서의 우량종목 선별방법 및 최적 투자전략의 지속성에 관한 연구”, 한국과학기술원 석사학위논

문, 2015.

홍동현, 황재호, 정경석, “결합가치투자전략 성과: KOSPI 200 기업을 중심으로”, 금융공학연구, 제10권, 제4호, 2011, pp. 59-80.

Basu, S., “Investment performance of common stocks in relation to their price earnings ratios: A test of the efficient market hypothesis”, *The journal of Finance*, Vol. 32, No. 3, 1977, pp. 663-682.

Fama, E. F., “Efficient capital markets: A review of theory and empirical work”, *The journal of Finance*, Vol. 25, No. 2, 1970, pp. 383-417.

Fama, E. F., and French, K. R., “Common risk factors in the returns on stocks and bonds”, *Journal of Financial Economics*, Vol. 33, No. 1, 1993, pp. 3-56.

Hargreaves, C., and Hao, Y., “Does the use of technical & fundamental analysis improve stock choice?: A data mining approach applied to the Australian stock market”, *In 2012 International Conference on Statistics in Science, Business and Engineering*, 2012, pp. 1-6.

Huang, C. F., “A hybrid stock selection model using genetic algorithms and support vector regression”, *Applied Soft Computing*, Vol. 12, No. 2, 2012, pp. 807-818.

Kedia, V., Khalid, Z., Goswami, S., Sharma, N., and Suryawanshi, K., “Portfolio generation for Indian stock markets using unsupervised machine learning”, *In 2018 Fourth International Conference on Computing Communication Control and Automation*, 2018, pp. 1-5.

Lakonishok, J., Shleifer, A., and Vishny, R. W., “Contrarian investment, extrapolation, and risk”, *The journal of finance*, Vol. 49, No. 5, 1994, pp. 1541-1578.

Mattos, D. L., “Comparing machine learning algorithm performance for automated trading based on fundamentals”, The Escola de Economia de São Paulo Doctoral dissertation, 2019.

Zhou, C., Yu, L., Huang, T., Wang, S., and Lai, K. K., “Selecting valuable stock using genetic algorithm”, *In Asia-Pacific conference on simulated evolution and learning*, 2006, pp. 688-694.

김 윤 승 (Kim, Youn Seung)

동국대학교 컴퓨터공학사와 경상국립대학교 기술경영공학 석사학위를 취득하였다. 현재 중소벤처기업진흥공단에서 재직하고 있으며, 주요 관심분야는 빅데이터 분석, AI, RPA, 정보보안 등이다.

유 동 희 (Yoo, Dong Hee)

고려대학교 경영학사와 경영학 박사학위를 취득하였다. 현재 경상국립대학교 경영정보학과에서 교수로 재직하고 있으며, 주요 관심분야는 인공지능, 빅데이터 분석, 온톨로지, 지식 그래프, 지능형 정보시스템 등이다.

<Abstract>

Selecting Stock by Value Investing based on Machine Learning: Focusing on Intrinsic Value

Kim, Youn Seung · Yoo, Dong Hee

Purpose

This study builds a prediction model to find stocks that can reach intrinsic value among KOSPI and KOSDAQ-listed companies to improve the stability and profitability of the stock investment. And investment simulations are conducted to verify whether stock investment performance is improved by comparing the prediction model, random stock selection, and the market indexes.

Design/methodology/approach

Value investment theory and machine learning techniques are applied to build the model. Various experiments find conditions such as the algorithm with the best predictive performance, learning period, and intrinsic value-reaching period. This study selects stocks through the prediction model learned with inventive variables, does not limit the holding period after buying to reach the intrinsic value of the stocks, and targets all KOSPI and KOSDAQ companies. The stock and financial data are collected for 21 years (2001-2021).

Findings

As a result of the experiment, using the random forest technique, the prediction model's performance was the best with one year of learning period and within one year of the intrinsic value reaching period. As a result of the investment simulation, the cumulative return of the prediction model was up to 1.68 times higher than the random stock selection and 17 times higher than the KOSPI index. The usefulness of the prediction model was confirmed in that the number of intrinsic values reaching the predicted stock was up to 70% higher than the random selection.

Keyword: Fundamental Analysis, Value Investment, Intrinsic Value, Machine Learning, Simulation

* 이 논문은 2023년 3월 6일 접수, 2023년 3월 22일 1차 심사, 2023년 3월 30일 게재 확정되었습니다.