JKSCI

한국컴퓨터정보학회논문지
Journal of The Korea Society of Computer and Information
Vol. 29 No. 2, pp. 51-59, February 2024
https://doi.org/10.9708/jksci.2024.29.02.051

Prediction of Stock Returns from News Article's Recommended Stocks Using XGBoost and LightGBM Models

Yoo-jin Hwang*, Seung-yeon Son*, Zoon-ky Lee*

*Student, Graduate School of Information, Yonsei University, Seoul, Korea
*Student, Graduate School of Information, Yonsei University, Seoul, Korea
*Professor, Graduate School of Information, Yonsei University, Seoul, Korea

[Abstract]

This study examines the relationship between the release of the news and the individual stock returns. Investors utilize a variety of information sources to maximize stock returns when establishing investment strategies. News companies publish their articles based on stock recommendation reports of analysts, enhancing the reliability of the information. Defining release of a stock-recommendation news article as an event, we examine its economic impacts and propose a binary classification model that predicts the stock return 10 days after the event. XGBoost and LightGBM models are applied for the study with accuracy of 75%, 71% respectively. In addition, after categorizing the recommended stocks based on the listed market(KOSPI/KOSDAQ) and market capitalization(Big/Small), this study verifies difference in the accuracy of models across four sub-datasets. Finally, by conducting SHAP(Shapley Additive exPlanations) analysis, we identify the key variables in each model, reinforcing the interpretability of models.

▸ **Key words**: Recommended Stock Return Prediction, XGBoost, LightGBM, SHAP, News Article

[요 약]

투자자는 수익의 극대화를 위해 언론사의 기사를 포함한 다양한 정보를 활용하여 투자 전략을 수립한다. 이에 국내 언론사에서도 신뢰도 있는 투자정보를 제공하기 위해, 애널리스트의 종목분석 보고서에 기초한 종목 추천기사를 게재하고 있다. 본 연구에서는 종목 추천기사 게재를 하나의 사건(event)으로 간주하고, XGBoost와 LightGBM 모델을 활용하여 기사 게재 10일 이후 가격의 상승 또는 하락을 예측하는 분류 모델을 제시한다. 또한, 전체 추천종목을 유가증권시장과 코스닥시장 및 기업규모(대형/소형)에 따라 4가지로 분류하고, 하위 그룹에 따라 모델의 예측 정확도에 차이가 있는지 파악하고자 한다. 학습 결과 전체 모델의 분류 정확도는 XGBoost 75%, LightGBM 71%로 나타났고, 예측 정확도는 유가증권 시장 예측력이 코스닥시장 주식 대비 높게 나타났으며, 대형주의 예측력이 소형주 보다 높게 나타났다. 마지막으로, SHAP(Shapley Additive exPlanations) 분석을 통해 개별 모델의 예측에 중요한 변수를 살펴보고 모델의 해석력을 제고하였다.

▸ **주제어**: 추천종목 수익률 예측, XGBoost, LightGBM, SHAP, 뉴스기사

• First Author: Yoo-jin Hwang, Seung-yeon Son, Corresponding Author: Zoon-ky Lee
*Yoo-jin Hwang (yujinhwang00@yonsei.ac.kr), Graduate School of Information, Yonsei University
*Seung-yeon Son (sonny@yonsei.ac.kr), Graduate School of Information, Yonsei University
*Zoon-ky Lee (zlee@yonsei.ac.kr), Graduate School of Information, Yonsei University
• Received: 2024. 01. 15, Revised: 2024. 02. 13, Accepted: 2024. 02. 13.

http://www.ksci.re.kr pISSN:1598-849X | eISSN:2383-9945

I. Introduction

기술의 발전으로 주식시장에 대한 정보의 비대칭성이 점차 완화됨에 따라, 수익성 높은 투자전략에 대한 기대와 관심이 지속적으로 증대되고 있다. 투자자는 주식 투자의 수익률 극대화를 위해 투자전략 수립 시 다양한 정보를 활용한다. 대표적으로 언론사의 뉴스는 현실세계에서 발생하는 각종 현상에 대한 설명과 함께 개별 기업, 경제, 사회 등과 관련한 전망을 제공하고 있다. 이에 언론사는 주가에 유의한 영향을 미치는 요인으로 알려져 있으며, 시장 참가자들은 기사 속 정보를 분석하여 주식시장의 변동성과 시장상황을 파악한다[1].

한편, 국내 언론사들은 신뢰도 높은 정보의 제공을 목적으로 증권사 애널리스트의 종목분석보고서를 기반으로 인터넷 뉴스기사를 제공하고 있다. 특히 경제 일간지들은 '이번주 추천 종목'과 같은 제목으로 정기적으로 애널리스트의 종목분석 기사를 게재하며, 기업을 둘러싼 시장 상황과 주가 동향에 대한 심도있는 분석을 통해 투자자의 관심도가 높은 종목에 대한 정보를 제공하고 있다.

최근의 실증분석은 미디어 콘텐츠와 주식시장 간의 관계를 분석하고, 다양한 비정형 텍스트 데이터를 활용하여 수익률을 예측하는데 중점을 두고 있다. 그러나, 뉴스 기사에 명시적으로 실린 애널리스트의 투자 의견을 바탕으로 한 뉴스기사가 수익률에 어떤 영향을 미치는지에 대한 실증분석은 매우 드물다. 본 연구 목적은 뉴스기사 중에서도 애널리스트의 추천종목 등 투자자의 관심이 유발되는 기사가 게재되는 경우, 개별 기업의 특성에 따라 수익률의 상승과 하락에 미치는 영향을 비교분석하는 것이다. 이를 위해 추천종목이 미디어로 보도되는 하나의 사건을 일종의 사건(event)으로 간주하고, 기사 게재 이전 5일간의 개별 주식의 특성에 관련된 데이터를 활용하여 게재 10일 이후 수익률의 방향성을 예측하였다.

자료는 국내 포털사이트 네이버에서 2019년부터 2022년까지 '추천', '추천종목'이라는 키워드로 검색된 뉴스기사에서 추천 종목을 발굴하여 활용하였다. 이와 함께 머신러닝 기법 중 의사결정 트리 기반의 앙상블(ensemble) 기법인 XGBoost(eXtreme Gradient Boosting) 모델과 LightGBM(Light Gradient Boosting Machine) 모델을 학습에 적용하여, 모델의 분류 정확도를 측정하고 주식종목의 특성에 따라 정확도가 달라지는지 검증하였다. 또한, 전체 추천종목을 유가증권시장 또는 코스닥 시장 및 기업규모(대형/소형)에 따라 4가지로 분류한 후, 각 하위 집단에 따라 모델의 예측 정확도에 차이가 있는지 검증하였다. 이와 함께 본 연구에서는 SHAP(Shapley Additive exPlanations) 분석결과를 제시하여, 각 모델의 주요 변인을 파악해 모델의 해석력을 제고하였다.

II. Preliminaries

1. Related works

1.1 Relationship between Stock Market and Media

기존 연구에서는 거시경제지표 등을 포함한 뉴스기사가 주식 수익률에 미치는 영향을 분석한 연구[2]를 시작으로, 미국의 경제 일간지 월스트리트저널(WSJ) 칼럼이 실제 주식시장의 움직임을 예측할 수 있다는 것을 밝힌 연구[3] 등과 같이 주식시장에 영향을 미치는 미디어의 역할을 검증하기 위한 다양한 연구가 꾸준히 진행되어왔다.

최근 주식시장과 미디어의 관계를 검증하는 연구의 흐름은 시의성 있는 사회현상과 결합하며 발전하였는데, 예로 기업의 지속가능성을 평가하는 새로운 지표인 ESG(Environment, Social, and Governance) 지수와 결합한 뉴스기사가 주가에 미치는 영향을 분석한 연구[4]가 있다. 해당 연구는 주식가격이 비재무적으로 중요한 ESG 뉴스에 반응하며, 그 반응은 긍정적인 뉴스 및 보도빈도가 높은 뉴스, 그리고 사회 자본 문제와 관련된 뉴스에 더 크다는 것을 보고하였다.

또한, 빅데이터 기술의 발전에 따라 방대한 양의 뉴스기사 및 소셜 미디어 콘텐츠에 대한 분석이 가능하게 되어, 긍정 또는 부정의 감정을 추출하여 주가를 예측하는 감성분석(Sentiment Analysis)과 관련된 연구의 관심도가 증가하였다. 감성분석 연구 중 [5]는 뉴스가 주로 비정형 텍스트로 구성되어 있음을 고려하여, 감성분석 기법을 적용해 주가지수의 등락을 예측하는 모델을 제안하였다. 또한, 대규모 언어처리모델(Large Language Model)이 발전함에 따라 [6]은 KoBert(한국어 텍스트 처리를 위해 설계된 구글의 언어모델)를 이용해 개별 기업 관련 기사의 긍정, 중립, 부정 감성 분류 결과가 투자 결정에 유용한 정보를 제공할 수 있는지를 탐구하고 유의한 결과를 보고하였다. [7]는 정형 데이터인 주가 정보와 비정형 데이터인 뉴스를 통합하여 딥러닝 모델을 기반으로한 주가 예측 모델에서 감성 지표의 영향력을 비교 분석하였다.

이와 같이 뉴스 기사와 주가의 상관관계를 분석하고, 뉴스를 기반으로 주가 예측에 관한 다양한 연구가 수행되어왔다. 그러나 애널리스트가 뉴스기사를 통해 직접 추천한 주식 종목이 기업의 주가 및 주식 거래량에 미치는 영향에

대한 연구는 상대적으로 부족하다는 한계점이 있다.

1.2 Financial Event Study

주식 가격의 상승과 하락은 기업을 둘러싼 다양한 외부 변수로 인한 영향을 받는 동시에 실적발표, 인수합병 등의 여러 내부 전략에도 영향을 받는다. 이러한 재무 관련 사건을 하나의 사건으로 취급하여 주식 수익률과의 연관성을 밝히는 연구가 지속적으로 진행되어왔다. [8]은 한국 주식시장에서 개인투자자들의 배당에 대한 관심이 세금 부담 감소와 관련이 있다는 점에 착안하여, 배당소득세율의 변화를 하나의 사건으로 보고 개인 투자자들의 배당락일 전후의 거래행태를 비교하여 분석하였다. 또한, [9]은 한국 상장기업의 무상주 발행이 단·장기 주가의 성과와 거래량에 미치는 효과를 검증하였으며, [10]는 국내 기업의 합병 사례를 통해 시장에서 합병정보가 어떻게 반영되고 있는지 밝힌 후, 이를 기초로 하여 기업의 합병이 주주의 부에 어떤 영향을 미치고 있는지를 분석하였다. 최근에는 이러한 재무 관련 이벤트 연구 또한 다양한 기계학습과 결합하여, 신경망 모델 또는 로지스틱 회귀분석(Logistic Regression), 랜덤 포레스트(Random Forest), 서포트벡터 머신(SVM) 등을 통해 실적 공시, 공모주 상장 등의 사건 이후 해당 기업의 주가를 예측하는 연구들이 활발하게 수행되고 있다[11][12]. 본 연구는 애널리스트의 추천 종목이 게재된 기사를 하나의 사건으로 간주하고, 해당 추천 종목 기사 게재 이후의 종목 수익률을 예측하는 것을 목표로 한다.

III. The Proposed Scheme

1. Model Framework

본 연구는 추천종목이 미디어로 보도되는 날짜 이전의 5일간 개별 기업의 특성과 관련된 데이터를 활용하여, 기사 게재 10일 이후 수익률의 방향성을 예측하는 것에 목표를 두고 있다. 다음은 모델의 전반적인 구조를 나타낸다.

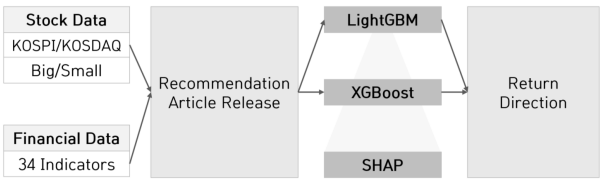

Fig. 1. Model Framework

2. Data Description

본 연구에서는 2019년부터 2022년까지 총 4년간 국내 포털사이트 네이버에서 35개의 경제·미디어 관련 언론사를 대상으로 '추천', '추천종목'을 키워드로 하여 뉴스기사 데이터를 크롤링하였고, 추천기사에서 증권사 애널리스트가 추천대상으로 선정한 기업종목과 추천 날짜를 수집하였다. 이때 인수합병, 분할, 상장폐지 등의 사유로 인해 데이터가 부재한 기업 33개는 제외하였으며, 추천종목은 유가증권시장(KOSPI)와 코스닥시장(KOSDAQ)에 상장된 주식으로 제한함에 따라 이에 속하지 않는 기업 15곳 또한 제외하였다. 이를 통해 총 2,407개의 뉴스기사에서 총 14,334건의 추천종목과 뉴스 게재 날짜를 발굴하였고, FnGuide에서 다음 변수들의 뉴스 게재일 이전 5일간의 재무 데이터를 가공 후 매치하였다.

Table 1. Data Description

News Companies	· e-daehan · global economic · news tomato · daehan financial news · digital times · maeil business · maeil business TV · money today · beta news · bridge economy · business post · biz Korea · seoul economy · seoul news · seoul finance · segye news · meconomy news	· asia kyungjae · asia times · asia today · aju news · energy economy · e-daily · etoday · chosun biz · chosun news · jose news · smedaily · joongang news · tech world · financial news · financial media · pin point news · korea economy · herald economy
Annual Articles Count	· 2019Y: 733 · 2020Y: 659	· 2021Y: 603 · 2022Y: 412
Excluded in Analysis	· Split-Off: 4 · M&A: 28 · Trading Halt: 1 · Unlisted from KOSPI or KOSDAQ: 15	

주식의 가격은 거시환경변수 및 시장·기업상황에 따라 상이한 특성과 흐름을 보인다. 이러한 수익률의 변동은 투자자들의 투자 성향에도 많은 영향을 미치며, 이에 수익률 방향성을 예측하는 모델의 성능은 기업의 특성에 따라 차이가 있을 수 있다. 따라서, 본 연구에서는 1) 추천종목이 상장된 시장(유가증권시장 또는 코스닥시장)에 따라, 2) 추천종목의 기업규모(대형주 또는 소형주)에 따라 기업들을 분류하고, 하위 그룹별로 모델의 예측력을 비교하였다.

상장된 시장은 기사 게재 날짜를 기준으로 분류하였으며, KOSPI 100 구성종목의 최저 시가총액을 기준으로 기준치 초과 종목을 대형주, 미만 종목을 소형주로 구분하였다. 이후 개별 그룹별로 주식 수익률의 방향성을 예측하는 모델을 생성하여 심층적인 모델 분석 결과를 제시하고자 하였으며, 분류 결과는 다음과 같다.

Table 2. Sub-Datasets

KOSPI/KOSDAQ	· KOSPI: 10,516 · KOSDAQ: 3,818
Big/Small	· Big: 8,119 · Small: 6,215

3. Model Algorithms – XGBoost, LightGBM

앙상블 모델은 동일한 학습 알고리즘을 활용하여 여러 모형으로 나온 예측 결과를 통합 최종 의사결정에 활용하는 방법으로, AdaBoost, CatBoost, GradientBoost 등 여러 모델이 있다[13]. 이중 대표적인 XGBoost와 LightGBM은 모두 그래디언트 부스트 기반 앙상블 알고리즘으로, 복잡한 데이터셋의 분류 및 예측 문제에 우수한 성능을 발휘하는 것으로 알려져 있다. 이 알고리즘들은 특정한 손실 함수를 최소화하기 위해 경사 하강법을 사용하며, 다양한 정규화 기법과 가지치기 전략을 통해 과적합을 방지한다.

본 연구에서는 거래량, 가격 등 다양한 단위의 변수들이 활용되었다. 또한, 시계열/비시계열 등 다양한 유형의 연속형 변수들이 활용되었다는 점과 컨센서스 데이터의 경우 결측치가 일정 부분 존재한다는 점에서 변수의 특성에 가장 적합한 XGBoost와 LightGBM 모델을 활용하였다.

XGBoost는 그래디언트 부스트의 과적합 문제와 긴 수행 시간을 개선한 알고리즘으로, 변수의 분포를 고려한 효율적인 연산과 합리적인 규제를 통해 안정적인 예측을 제공한다.

LightGBM은 기존의 의사결정 나무 기반 알고리즘과 다른 접근 방식인 '리프 중심의 트리 분할(Leaf-wise tree growth)'을 사용한다. 이 방식은 트리의 균형을 맞추기보다는 손실을 최대한 줄일 수 있는 리프를 우선적으로 분할한다. 이 접근법은 빠른 학습 속도와 낮은 메모리 사용량의 이점을 제공하며, 특히 큰 규모의 데이터에서 그 효율성이 두드러진다.

또한, LightGBM은 결측치 처리 및 다양한 유형의 데이터 처리에 효율적임에 따라, 주가 예측 모델과 같은 복잡한 데이터셋에도 높은 효율성을 보인다. 이에 본 연구에서는 최근 머신러닝 기법으로 주목받는 XGBoost와 LightGBM 모델을 각각 적용한 후, 성능 결과를 비교분석하였다.

4. Model Variables

본 연구의 모델 변수 데이터는 2019년부터 2022년까지 FnGuide에서 제공하고 있는 자료를 수집하였다. 모델 변수는 [14], [15] 등의 연구와 같이 주가 예측에 널리 활용되는 재무 변수 34개이며, 변수의 내용은 Table 3와 같다.

Table 3. List of Variables

No	Name	Description
1	volatility	annualized volatility
2	return	total return
3	volume	whole trade volume
4	size	market capitalization
5	short_interest	short interest
6	buying_volume(ind)	individual investors' buying volume
7	selling_volume(ind)	individual investors' selling volume
8	net_buying_volume(ind)	individual investors' net buying volume
9	buying_volume(ins)	institutional investors' buying volume
10	selling_volume(ins)	institutional investors' selling volume
11	net_buying_volume(ins)	institutional investors' net buying volume
12	buying_volume(for)	foreign investors' buying volume
13	selling_volume(for)	foreign investors' selling volume
14	net_buying_volume(for)	foreign investors' net buying volume
15	PER_past	PER
16	PBR_past	PBR
17	PSR_past	PSR
18	EV/EBITDA_past	EV/EBITDA
19	PER_consensus	PER(consensus)
20	PBR_consensus	PBR(consensus)
21	PSR_consensus	PSR(consensus)
22	PCFR_consensus	PCFR(consensus)
23	EV/EBITDA_consensus	EV/EBITDA(consensus)
24	PER_ratio	PER consensus / PER past
25	PBR_ratio	PBR consensus / PBR past
26	PSR_ratio	PSR consensus / PSR past
27	EV/EBITDA_ratio	EV/EBITDA consensus / EV/EBITDA past
28	closing_price	closing price
29	revenue_growth_rate_consensus	revenue growth rate (consensus)
30	earning_profit_rate_consensus	earning profit rate (consensus)
31	net_profit_rate_consensus	net profit rate (consensus)
32	alpha	alpha
33	beta	beta
34	number_of_analysts	number of analysts who released stock reports

1-33번까지의 변수는 게재일 기준 5일 전까지의 평균값으로 계산하였으며, 34번 변수(추정기관 수)의 경우 event date 전일 데이터를 활용하였다. 데이터의 전처리 과정은 다음과 같다.

먼저, 거래량은 종목별로 기업의 규모에 영향을 받을 수 있다는 점을 참작해 각각 발행주식 수로 나누어 거래량 변수 간의 단위를 통일하였다. 또한, 4번 변수(시가총액)는 데이터 간 차이가 큰 점을 감안하여, 로그 정규화를 수행해 변동성을 조정하였다. 마지막으로, 연도별로 상이할 수 있는 투자 동향을 고려하여 연도별로 데이터를 분류하고 train set data, valid set data, test set data를 구성한 뒤 이를 각각 통합해 연도별 데이터 간의 균형을 고려하고 데이터의 안정성을 개선하였다.

본 연구에서는 예측 대상이 되는 종속 변수를 y를 게재일 기준 10일 이후의 수익률로 계산하였으며, 수익률의 부호에 따라 이진(양(+) 또는 0의 값을 가지는 경우 1, 음(-)의 값을 가지는 경우 0)으로 분류하여 수익률의 방향성을 설정하였다. 또한, 주요 입력 변수 및 종속 변수 데이터가 누락되어 알고리즘에 학습이 어려운 155개의 결측 데이터를 제거한 후 진행하였다.

모델 학습 전 XGBoost와 LightGBM 모델의 하이퍼파라미터 최적화를 위하여 Bayesian Optimization을 채택하였다. Bayesian Optimization은 사전지식(prior)과 데이터로부터 얻은 사후지식(posterior)을 활용하여 최적의 하이퍼파라미터 조합을 효율적으로 찾는 최적화 기법으로, 특히 데이터의 규모가 크고 변수가 많은 모델에서 발생할 수 있는 수렴 문제를 효과적으로 해결한다. 이에 본 연구에서는 iteration을 최대 50으로 설정한 Bayesian Optimization을 수행하여 각 모델에 적용하였다.

또한, 본 연구의 모델에 쓰인 변수들의 중요성(Feature Importance)을 파악하기 위해 SHAP 방법론을 활용하였다. SHAP은 머신러닝 모델의 예측에 대한 투명성과 이해도를 높이기 위해 활용하는 게임 이론 기법으로, 각 입력변수가 학습된 모델의 예측에 미치는 상대적인 중요도를 섀플리 값(Shapley Values)을 통해 정량적으로 평가함으로써 변수 간의 상호작용과 영향력을 시각화하여 제시한다[16]. 또한, SHAP 기법은 변수들이 상호 간 영향을 줄 수 있다는 것을 고려하고, 변수의 음(-)의 영향력까지 계산할 수 있다는 점에서 Feature Importance 기법보다 더욱 정확한 영향력을 측정한다[17]. 이에 SHAP 기법을 활용하여 모델의 해석력을 높이는 동시에, 영향력이 높은 변수들을 파악하였다.

5. Performance Analysis

데이터는 train data와 test data를 8:2 비율로 나누어 활용하였으며, 모델의 정확한 성능 및 일반화 능력을 평가하기 위해 train data에서 valid data를 4:1 비율로 다시 나누었다. 이에 train data와 valid data, test data는 각각 64%, 16%, 20% 정도의 비율로 분리되었다. 모델의 성능 평가는 혼동행렬(Confusion Matrix), F1 Score, ROC Curve를 통해 종합적으로 분석되었으며, 특히 상장시장과 기업규모 간의 예측 정확도 차이와 함께 XGBoost와 LightGBM 간의 성능 차이를 비교하는데 중점을 두었다.

IV. Results

1. Modeling Performance

모델 학습 이후 이를 test set에 검증한 결과는 1) 전체 데이터, 2) 유가증권시장 상장기업, 3) 코스닥시장 상장기업, 4) 대형주 기업, 5) 소형주 기업으로 분류되어 다음과 같이 제시되었다. 본 연구에서는 모델의 분류 정확도를 객관적으로 검증하고 모델의 일반화 가능성을 확인하기 위해, Table 4.에서 valid data에 대한 정확도 결과를 먼저 제시하였다. 이후 Table 5에서 9까지 test data에 대한 정확도, 재현율, 정밀도, F1 Score, AUC를 포함한 전체 혼동행렬 결과를 제시하여 모델의 객관성을 제고하였다.

Table 4. Accuracy Results - Valid Data

	XGBoost	LightGBM
Entirety	0.75	0.72
KOSPI	0.77	0.73
KOSDAQ	0.69	0.68
Big	0.81	0.77
Small	0.67	0.69

Table 5. Confusion Matrix Results - Entirety

	XGBoost	LightGBM
Accuracy	0.75	0.71
Recall	0.76	0.73
Precision	0.75	0.71
F1 Score	0.75	0.72
AUC	0.83	0.78

Table 6. Confusion Matrix Results - KOSPI

	XGBoost	LightGBM
Accuracy	0.76	0.73
Recall	0.79	0.78
Precision	0.79	0.72
F1 Score	0.79	0.75
AUC	0.85	0.81

Table 7. Confusion Matrix Results - KOSDAQ

	XGBoost	LightGBM
Accuracy	0.69	0.67
Recall	0.66	0.65
Precision	0.74	0.74
F1 Score	0.70	0.69
AUC	0.73	0.72

Table 8. Confusion Matrix Results - Big

	XGBoost	LightGBM
Accuracy	0.80	0.75
Recall	0.81	0.75
Precision	0.81	0.77
F1 Score	0.81	0.76
AUC	0.88	0.83

Table 9. Confusion Matrix Results - Small

	XGBoost	LightGBM
Accuracy	0.65	0.66
Recall	0.70	0.70
Precision	0.63	0.64
F1 Score	0.66	0.67
AUC	0.72	0.72

5가지의 데이터 분류에 따른 분석결과는 다음과 같다. 먼저, 전반적으로 모델의 예측 정확도가 60~80%의 범위에 있어, 본 수익률 예측 모델을 합리적으로 신뢰할 수 있음을 알 수 있다. 기업의 수익률에 영향을 미치는 요소는 매우 다양하며, 이러한 변인들은 여러 상호작용을 통해 주가의 방향성 파악을 더욱 어렵게 만든다. 이에, 기업의 추천종목 기사 게재라는 사건을 통해 평균 70% 정도의 정확도를 기록한 것은 본 연구에 활용된 모델의 실효성과 앞서 제시한 모델의 경제적 가치를 뒷받침한다.

또한, 유가증권시장 상장기업과 코스닥시장 상장기업 중에서는 유가증권시장 상장주식들이 코스닥시장 상장주식들에 비해 모델의 정확도가 더욱 높았다. 이는 높은 인지도와 신뢰도를 바탕으로 하는 유가증권시장 상장사에 대한 투자자들의 긍정적인 인식과 안정성이 예측 모델의 성능에 일정한 영향을 미친다는 것을 시사한다. 반면, 코스닥시장 상장사들에 대한 예측 모델은 상대적으로 낮은 정확도를 보였으며, 시장의 불확실성과 변동성이 코스닥시장 상장사들의 수익률 방향성 파악을 어렵게 하는 것으로 해석할 수 있다.

이러한 결과는 대형주와 소형주 간의 비교에도 유사하게 적용된다. 시가총액은 주식시장 내 여러 시장지수의 인덱스로 활용된다는 점에서 기업의 규모와 안정성을 파악하기에 매우 유용한 지표라고 할 수 있다. 대형주로 정의되어 분류된 기업들에 대한 모델의 예측 정확도는 80%에 가까워 매우 높게 나타난 반면 소형주에 대한 모델 정확도는 65% 정도로 낮게 나타나며, 두 비율의 차이는 유가증권시장 상장기업과 코스닥시장 상장기업 간의 차이보다 더욱 크다는 것을 알 수 있다.

추가로, 기사 게재일 기준 5일간의 과거 재무 데이터로 10일 이후 수익률의 방향성을 예측하는 기존 모델에서, 관측기간을 늘려 과거 10일 데이터를 통해 게재일 20일 이후 수익률을 예측하는 모델을 생성함으로써 모델의 일반화 가능성을 제고하였다. 그 결과, 본래 연구의 모델과 유사한 수준의 정확도를 보이는 것으로 확인되었으며, 세부결과는 다음과 같다.

Table 10. Results(Past 10 days, 20 days Return)

	XGBoost	LightGBM
Accuracy	0.77	0.72
Recall	0.79	0.77
Precision	0.76	0.70
F1 Score	0.78	0.74
AUC	0.86	0.80

2. Feature Importance Analysis

입력변수의 중요도를 파악하기 위한 SHAP 기법은 데이터셋을 분류한 하위 집단에 각각 적용하였으며, LightGBM 모델에 기반하여 측정하였다. 다음은 전체 데이터셋에 대한 SHAP 분석결과와 전체 데이터셋을 제외하고 4개의 세부집단을 SHAP으로 분석한 각 상위 5개, 하위 5개 변수 결과이다.

먼저, 대부분의 그룹에서 기사 게재 이전 5일간의 PER(Price Earning Ratio; 과거 PER)이 중요한 변인으로 꼽혔음을 알 수 있다. PER은 주가를 기업의 순이익과 비교하여, 투자자가 해당 종목 투자 시 얻을 수 있는 이익이 현 주가에 비해 얼마나 높은지를 나타낸다. PER은 주로 시장의 기대치를 반영함에 따라, 투자자의 심리가 수익률 방향성 예측에 중요하다는 결론을 내릴 수 있다.

또한, 유가증권시장 상장 종목들과 대형주로 분류된 종목 간 중요도가 높게 측정된 변인들이 일정 부분 유사하

다. 유가증권시장 상장 종목들의 경우 변동성과 시가총액, 기사 게재 이전 PER, 알파 등의 수치가 중요한 변인으로 꼽혔다. 이들은 대형주 사이에서도 높은 중요도로 평가되었는데, 이는 대형주로 분류되는 종목들이 주로 유가증권시장에 상장되었기 때문이라 해석할 수 있다.

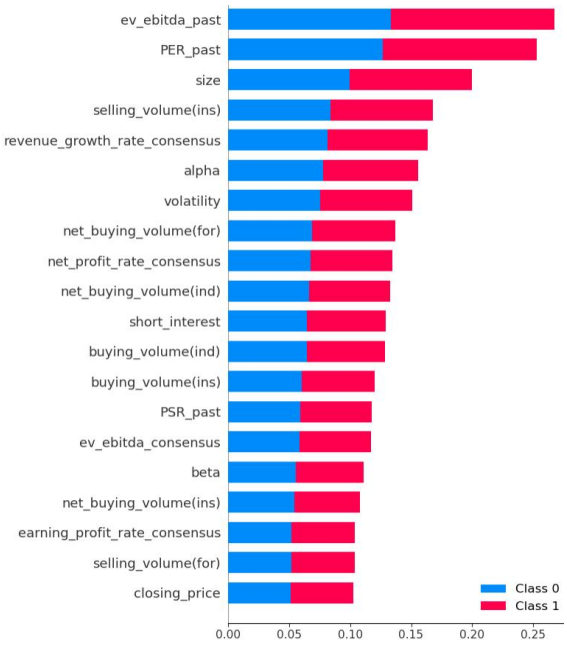

Fig. 2. SHAP Analysis Results – Entirety

Table 11. SHAP Analysis Results

No	KOSPI	KOSDAQ	Big	Small
1	volatility	**PER(past)**	revenue growth rate (consensus)	**PER(past)**
2	size	EV/EBITDA(past)	alpha	EV/EBITDA(past)
3	selling volume(ins)	earning profit growth rate (consensus)	**PER(past)**	net buying volume(ins)
4	**PER(past)**	net buying volume(for)	size	PSR(past)
5	EV/EBITDA (consensus)	alpha	PSR(past)	net buying volume(for)
16	PSR(past)	selling volume(ins)	PBR(past)	PSR (consensus)
17	beta	buying volume(ins)	return	buying volume(ind)
18	closing price	PBR(past)	net buying volume(for)	revenue growth rate (consensus)
19	buying volume(ins)	net buying volume (ins)	buying volume(ins)	net profit growth rate (consensus)
20	return	volatility	volume	return

마지막으로 대형주 모델링에서는 매출액 증가율 변수가 가장 높은 중요도를 가지는 것으로 나타났으나, 소형주에서는 유의한 중요도를 가지지 않는 것으로 나타났다. 일반적으로 대형주들은 기존에 안정적으로 확립된 사업모델과 시장지위를 가지고 있음에 따라, 투자자들이 미래의 수익 성장세를 크게 기대하기 때문이라고 분석할 수 있다. 이와 달리 소형주는 종종 신생 기업이거나 특정 산업 분야에서 견고한 시장위치를 확립하지 못했을 가능성이 높으며, 시장의 민감도가 높은 경우가 많다. 이에 따라, 단기적인 매출액 증가율에 비해 기타 외부 요인이 더욱 높은 영향력을 가질 수 있다.

V. Conclusions

지금까지 기계학습이 결합된 금융 분야의 연구에서 기업의 주가 및 수익률에 대한 연구는 매우 심층적으로 수행되었으나, 애널리스트의 투자의견을 뉴스기사로 게재해 언론의 주목을 받은 주식에 대한 연구는 면밀하게 분석되지 못했다. 이들은 간결한 내용으로 신속하게 투자자들에게 전달된다는 점에서 애널리스트 리포트와는 다른 영향력을 가지고, 이에 상이한 방법론을 통한 추가 연구가 필요하다. 따라서, 인터넷 뉴스기사를 통해 애널리스트의 추천종목과 사유가 투자자에게 직접적으로 안내되는 새로운 미디어 콘텐츠 분야의 초기 연구를 제시했다는 점과, 추천종목의 수익률 예측에 있어 기본적인 재무 변수들을 활용해 합리적인 예측 정확도를 보이는 모델을 제시하였다는 점에서 의미가 있다.

또한, 여기서 기업의 규모 및 상장시장에 따라 분류 정확도의 차이를 발견한 점에서도 의미가 있다. 최근에는 글로벌 경기가 둔화세를 보이고, 금리가 급격하게 변동하는 기조를 보임에 따라, 예측 모델에 변동성을 고려하는 것이 점차 중요해지고 있다. 이에, 연구에서 수행한 상장시장의 구분과 시가총액에 기반한 대형주/소형주의 구분은 모델의 정확도에 유의한 영향을 미쳤다는 점에서 가치가 있다.

그러나, 본 연구는 다음과 같은 한계를 가진다. 첫째, 본 연구는 2019년부터 2022년까지 4개년의 데이터를 수집하여 활용하였으나, 2019년 말부터 2022년까지 금융시장은 코로나19 팬데믹이라는 특수한 상황의 영향을 받았다. 이에, 관측기간을 늘려 데이터를 수집해 모델의 적합도를 재검증하면, 모델의 실효성을 더욱 제고할 수 있을 것이다.

또한, 뉴스기사 게재 날짜 수집 시 게재 시간까지는 고려하지 않았다는 점도 한계로 남는다. 뉴스기사는 게재된 이후 빠르게 투자자들에게 공유됨에 따라, 더욱 정확한 통계분석을 위해서는 기사가 장 중에 게재되었는지 장 마감

이후 게재되었는지 구분할 필요가 있다. 그러나, 본 연구에서는 모델 학습 시 뉴스 기사 게재 당일 전까지의 데이터를 학습하고, 이후 수익률을 고려하였기 때문에 장 중에 게재 여부까지 포괄하지 못했다.

향후 본 연구는 다른 재무지표를 결합하여 더욱 모델의 정확도를 개선하는 방향으로 확장될 수 있으며, 수익률 방향에서 수익률 자체를 예측하는 모델로 발전 시 sharpe ratio 또는 sortino ratio 등을 활용한 모델의 리스크 평가도 가능할 것으로 보인다.

또한, 보다 세밀한 데이터 처리를 위하여 후속 실증분석에서는 뉴스기사를 다룰 시 뉴스기사의 게재 시간까지 고려할 수 있다. 이와 함께 XGBoost와 LightGBM을 포함하여 다른 우수한 기계학습 알고리즘을 활용할 경우, 모델의 정확도를 개선하는 동시에 예측에 더욱 적합한 모델을 발전시킬 수 있을 것으로 기대한다.

ACKNOWLEDGEMENT

This work was supported by ICONS(Institute of Convergence Science), Yonsei University.

REFERENCES

[1] Chi Young Song, "News and Financial Prices", International Economic Journal, 8, 3, 1-34, December 2002

[2] Cutler, David M., James M. Poterba, Lawrence H. Summers, "What moves stock prices?", March 1988. DOI 10.3386/w2538

[3] Tetlock, Paul C., "Giving content to investor sentiment: The role of media in the stock market", The Journal of finance, 62, 3, 1139-1168, June 2007. DOI 10.2139/ssrn.685145

[4] G Serafeim, A Yoon, "Which corporate ESG news does the market react to?", Financial Analysts Journal, 78, 1, 59-78, February 2022. DOI 10.2139/ssrn.3832698

[5] Yoosin Kim, Namgyu Kim, Seung Ryul Jeong, "Stock-Index Invest Model Using News Big Data Opinion Mining", Journal of Intelligence and Information Systems, 18, 2, 143-156, June 2012. DOI 10.13088/jiis.2012.18.2.143

[6] Hyeon Jiwon, Lee Joonil, Cho Hyunkwon, "Sentiment Analysis of News on Corporation Using KoBERT", Korean Accounting Review, 47, 4, 33-54, 2022. DOI 10.24056/KAR.2022.08.002

[7] Doo-Won Kang, So-Yeop Yoo, Ha-Young Lee, Ok-Ran Jeong, "A study on Deep Learning-based Stock Price Prediction using News Sentiment Analysis", Journal of The Korea Society of Computer and Information, Vol. 27 No. 8, pp. 31-39, August 2022. DOI 10.9708/jksci.2022.27.08.031

[8] Cheol Park, Soo Cheol Park, "Individual Trading Behavior Comparison Around Ex-Dividend Days Before and After the Dividend Tax Changes", Korean Journal of Financial Studies, September 2010

[9] Hyunseok Kim, Jungwon Suh, "Stock Return and Liquidity Effects of Bonus Issues, Stock Splits and Stock Dividends: Evidence from Korea", Korean Journal of Financial Studies, August 2018. DOI 10.26845/KJFS.2018.02.47.1.27

[10] Lee Seokgyu, Byeon Yeongdeok, Park Sangguk, "Analysis of Market Response to Merger Announcements", Joumal of Accountifng and Finance, 18, 1-22, October 2002

[11] Suyeon Yang, Chaerok Lee, Jonggwan Won, Taeho Hong, "The prediction of the stock price movement after IPO using machine learning and text analysis based on TF-IDF", Journal of Intelligence and Information Systems, 28, 2, 237-262, June 2022. DOI 10.13088/jiis.2022.28.2.237

[12] Cui Jinhua, Kim Soonho, "Predicting Stock Prices Based on Neural Networks Around Earnings Announcements", Journal of The Korean Data Analysis Society, 22, 6, 2667-2678, December 2020. DOI 10.37727/jkdas.2020.22.6.2667

[13] Myung-woo Nam, Doo-Seo Park, Young-Jun Jang, Hong-Chul Lee, "Prediction Of Traffic Accident Casualties Using Machine Learning : For Seoul Public Data", Korea Society of Computer Information Spring Conference Proceedings, 29, 1, 27-30, January 2021.

[14] Pei-Fen Tsai, Cheng-Han Gao and Shyan-Ming Yuan, "Stock Selection Using Machine Learning Based on Financial Ratios", Mathematics, 11, 23, 4758, November 2023. DOI 10.3390/math11234758

[15] Reza Gharoie Ahangar, Mahmood Yahyazadehfar, Hassan Pournaghshband, "The Comparison of Methods Artificial Neural Network with Linear Regression Using Specific Variables for Prediction Stock Price in Tehran Stock Exchange", International Journal of Computer Science and Information Security, 7, 2, February 2010. DOI 10.48550/arXiv.1003.1457

[16] Lundberg, S. M., & Lee, S. I., "A unified approach to interpreting model predictions.", Advances in neural information processing systems, 30., 2017.

[17] Hyung-Rok Oh, Ae-Lin Son, Zoonky Lee, "Occupational accident prediction modeling and analysis using SHAP", Journal of Digital Contents Society, 22, 7, 1115-1123, July 2021. DOI 10.9728/dcs.2021.22.7.1115

Authors

Yoo-jin Hwang received a B.S. degree in Economics and Business Administration from Konkuk University in Seoul, Korea in 2023. She is currently pursuing her M.S. degree at Graduate School of Information, Yonsei University, Korea. Her current interests encompass data analytics and machine learning in finance.

Seung-yeon Son received a B.S. degree in Media Communication and Business Administration from Dongguk University in Seoul, Korea in 2007. She is currently pursuing her M.S. degree at Graduate School of Information, Yonsei University, Korea. Her current interests encompass data analytics and deep learning in finance.

Zoon-ky Lee received the B.S. degree in Computer Science from Seoul National University, Korea in 1985, M.S. in Social Psychology from Carnegie Mellon University, USA in 1991, and Ph.D in Management Informatics from Southern California University, USA in 1999. He is currently a Professor in the Graduate School of Information at Yonsei University, Korea. Zoon-ky Lee is interested in Big Data Analytics, Digital Transformation, and Open Collaboration.