財務管理硏究 第四十券 第一號(2023. 2), 81~104

뉴스와 댓글 텍스트를 활용한 공모주 상장일 시초가 형성 예측 및 영향성 분석: XGBoost 기반

최남현* · 김강산** · 장희수***

〈요 약〉

본 연구는 뉴스 및 댓글이 공모주 신규 상장 직후 가격형성에 영향을 미칠 수 있다고 가정한다. 2018.1~2022.7까지 신규 상장된 337개 공모기업을 대상으로 상장 전 시점 개인투자자의 관심을 포함한 댓글 데이터와 해당 댓글이 생성되는 뉴스 데이터들을 활용하여 공모주 상장 직후 시초가 형성에 영향을 미치는지 실증분석 하였다. 본 연구에서 뉴스, 댓글 데이터를 활용하기 위해 각 기업의 신고서 제출일로부터 상장예정일까지 네이버 뉴스에 게시된 뉴스와 댓글에 KR-FinBERT 모델을 적용해 긍정, 부정, 중립으로 분류 후 감성점수화 해 변수로 사용하였다. 또한 공모주 가격형성에 영향을 미치는 여러 변수를 사용해 거시경제 상황과 개인 및 시장 반응을 종합하여 공모주 상장 당일 확정공모가 대비 시초가가 1.0배, 1.5배 초과 형성되는지 분석하였다. 연구 결과 감성 변수를 포함하였을 때 확정공모가 대비 시초가가 1.0배 초과의 경우 감성 변수를 포함하지 않은 경우보다 좋은 성능을 보인 반면, 시초가가 1.5배 초과의 경우는 감성 변수를 포함한 경우가 포함하지 않은 경우보다 낮은 성능을 보였다. 이를 통해 감성 변수는 확정공모가 대비 시초가 형성을 예측 시 도움되지만, 초과 상승분에 대한 예측 시 도움이 되지 않음을 확인하였다.

주제어 : 감성분석, 공모주, 분류, 예측, XGBoost

논문접수일 : 2022년 12월 19일 논문수정일 : 2023년 01월 21일 논문게재확정일 : 2023년 01월 28일

* 제1저자, 숭실대학교 금융학부 학사과정, E-mail: holicman7@naver.com

** 공동저자, 숭실대학교 금융학부 학사과정, E-mail: rm7348@naver.com

*** 교신저자, 숭실대학교 금융학부 조교수, E-mail: yej523@ssu.ac.kr

Ⅰ. 서 론

공모란 기관투자자와 개인투자자들의 참여로 발생하는 공개모집의 약자로, 불특정 다수로부터 유가증권에 대한 모집 또는 매출하는 것을 의미한다. 공모주 청약에 대한 예정비율은 우리사주조합 20% 이내, 일반투자자 25% 이상 30% 이내[1], 나머지는 고수익 펀드를 포함한 기관투자자가 배정받게 된다. 개인투자자는 일반투자자에 해당되며 25% 이상 30% 이내를 청약으로 배정을 받을 수 있다. 이처럼 일반투자자는 공모 청약에 있어 무시할 수 없는 비중을 차지하고 있으며, 최근 커져가는 공모 규모에 따라 개인투자자들의 공모 참여가 점차 늘어나고 있는 추세이다.

이석훈(2020)에 따르면 IPO 공모주 상장 첫날 수익률이 높았다는 점과 코로나19 이후 신규 진입하는 개인투자자들이 많아져 전체적인 주식시장 거래 규모가 증가했다는 점을 개인투자자의 공모 참여가 늘어난 요인으로 꼽고있다. 실제 최근 몇 년 사이 각종 매체에서 주목을 받았던 공모주 중 규모가 큰 종목들의 예시로는 'SK바이오사이언스', 'LG에너지솔루션', '크래프톤' 등이 있으며, 해당 종목들을 공모청약 시 배정받은 가격으로 시초가에 매도하였다면 각각 +100%, +99%, -10%의 수익률을 얻을 수 있었다. 또한 [그림 1]과 같이 다수의 공모주들이 상장 시 높은 수익률을 보이며, 상장 당일 확정공모가 대비 시초가 기준 평균적으로 약 33.4%의 수익률을 보인다. 이처럼 공모주 청약은 반드시 수익이 나는 것은 아니지만, 단기간 시세차익으로 높은 수익을 얻을 수 있다는 장점이 있다. 이는 개인투자자들이 공모청약에 관심을 갖게 되는 분명한 이유가 될 수 있으며, 코로나 19 이후 주식시장의 때 아닌 호황으로 늘어난 개인투자자들의 대거 유입은 공모주 시장에도 영향을 끼쳤을 것이다. 실제로 코로나19 직후인 2020년 상반기 685:1이던 개인청약율은 2020년 하반기 1,017:1, 2021년 상반기 1,326:1로 상승하는 모습을 보이며, 개인투자자의 공모주 투자가 증가했음을 알 수 있다.

개인투자자의 공모 참여 증가에 따라 공모주 투자 시 개인투자자의 성향(이한석 외 1인, 2021), 개인투자자의 감성이 IPO 수익률에 미치는 영향(Saleh et al., 2021; 김종옥, 2022) 등 기존에 관심받지 못했던 개인투자자의 공모주 투자에 대한 다양한 후속 연구가 점차 증가하고 있다. 또한 SNS 및 증권 커뮤니티에 작성된 댓글(김명진 외 3인, 2020; Ye Xian,

1) 2020년 11월 금융위원회의 기업공개 공모주 일반청약자 참여기회 확대방안에 따라, 고수익 펀드의 비율이 기존 10%에서 5%로 감소되며 이 비율이 일반청약자에게 배정되어 일반청약자가 25%의 물량을 배정받을 수 있게 되었다. 추가로 우리사주조합 미청약물량을 최대 5%를 일반청약자에게 배정하여 일반청약자는 최대 30%까지 배정받을 수 있도록 개선되었다.

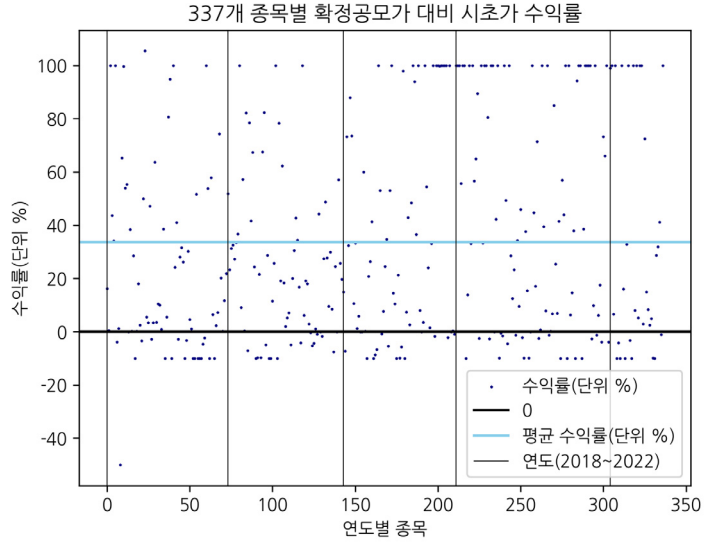
2021), 증권신고서의 텍스트 데이터와 같은 비정형 데이터를 분석(양수연 외 3인, 2022)하여 공모주 가격 예측을 하는 경우도 증가하고 있다. 이러한 변화에 따라 본 연구는 개인투자자의 공모 참여에 더 초점을 맞춰 개인투자자의 관심도를 중심으로 감성 변수와 공모주 관련 변수 등을 활용하여 공모주에 대한 관심이 가장 높은 시기인 상장 당일 시초가에 어떠한 영향을 미칠 수 있는지 분석하고자 한다. 이때 사용한 감성변수는 뉴스와 댓글 데이터 기반이며, 뉴스 데이터는 기업의 전반적인 정보를 개인투자자에게 제공하는 방법 중 하나로(류선진, 2021) 개인투자자의 관심도를 직접적으로 나타내는 댓글 데이터와 함께 사용하는 것이 성능 개선에 도움이 될 것이라고 판단하였다. 본 연구는 다음의 가정을 바탕으로 연구를 진행한다.

가정 1. 개인투자자의 공모 참여에 대한 관심이 공모주 상장 당일 시초가 형성에 영향을 미치며, 개인투자자는 공모 기업의 IPO에 관련된 뉴스 뿐만 아니라 기업의 활동, 업적 등 기업 전반적인 활동에 관한 뉴스에도 노출된다.

가정 2. [그림 2]를 근거로 투자자들은 상장 당일 초과 수익을 얻기 위해 공모주를 단타성으로 매매할 것이라고 가정한다.

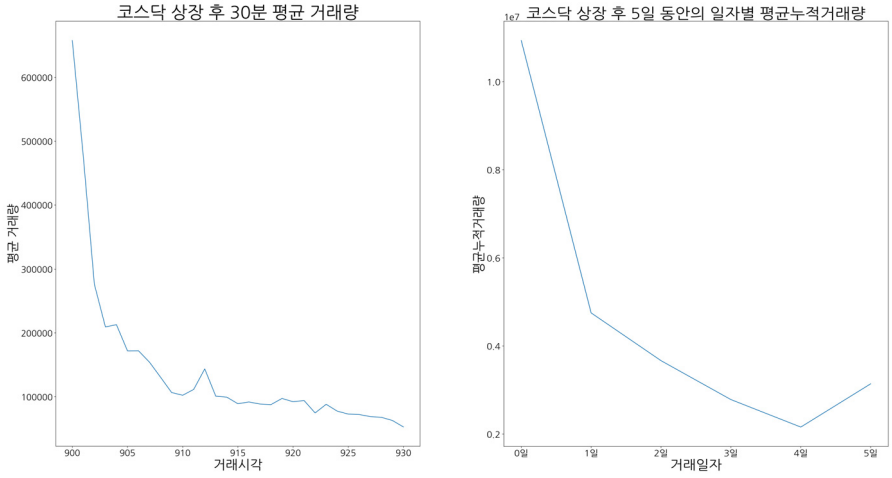
[그림 1] 337개 종목별 확정공모가 대비 시초가 수익률

2018.01.01.부터 2022.07.15. 까지 상장된 각 기업들을 대상으로 확정공모가 대비 시초가가 얼마나 변화했고, 그 경우를 수익률로 변환했을 때 기업별 수익률을 그래프로 나타낸 것이다. 왼쪽부터 오른쪽으로 기업의 공모일을 시간순으로 나열하였으며, 평균 33.4%의 수익률을 보인다. (컬러 그림은 http://kfma.org/38에서 볼 수 있음.)



[그림 2] 코스닥에 상장된 공모주의 거래량 변화 추이

이 그림은 연구기간 동안 코스닥에 상장된 공모주의 상장 직후 평균 거래량을 나타낸다. 왼쪽 그림은 상장 직후 30분 거래량을 나타내고, 오른쪽 그림은 상장 직후 5일 거래량을 나타낸다. 상장 직후 가장 거래량이 많음을 알 수 있으며, 시간이 지날수록 거래량이 낮아짐을 확인할 수 있다.



개인투자자의 관심도를 공모주의 수익성 및 가격형성 예측에 사용하기 위해 각 공모기업에 대해 신고서제출일부터 상장예정일까지 네이버 뉴스에 게시된 뉴스와 댓글 데이터를 수집하였다. 기업별로 수집된 데이터들을 감성분석하여 하나의 변수로 만들고, 다양한 추가적인 변수들을 포함하여 머신러닝을 이용한 예측을 통해 감성변수의 유무에 따른 공모주 가격형성 예측의 정확도를 분석한다.

본 논문의 구성은 다음과 같다. 제Ⅱ장은 공모주 투자, 개인투자자 감성분석 관련 그리고 머신러닝 모형에 관한 선행연구들을 언급하고, 제Ⅲ장은 본 연구에서 사용한 모형에 대해 언급한다. 제Ⅳ장은 본 연구의 실험 결과에 대해 언급하고, 마지막 장에서는 연구 결론 및 연구의 한계점과 향후 연구에 대해 서술한다.

Ⅱ. 관련연구

공모주에 대한 개인투자자의 관심도가 높아짐에 따라 개인투자자의 공모주 참여에 따른 수익성 여부 등 개인투자자의 영향을 분석하는 연구들이 늘어나고 있다. 이에 따라 개인투자자의 관심도가 공모주의 상장 후 가격 변화 예측의 보조 지표로서 영향을 미칠 수 있을 것이라 판단하였으며, 선행 연구에 없던 머신러닝 방법론을 이용하여 뉴스와 댓글 데이터를 자연어 처리 후 긍정, 부정, 중립의 감성점수화 해 변수로 추가함으로써 감성변수가

공모주 가격형성 요인이 될 수 있을지 여부에 대해 분석하고자 하였다.

실험에 앞서 공모주 상장 및 투자에 관한 선행연구와 감성분석을 통한 공모주의 상장 시 가격 변화예측에 관한 선행연구 그리고 연구에 사용된 머신러닝 방법론에 관한 선행연구로 나눠 결과들을 분석하였다. 공모주 신규 상장과 공모주 투자 시 고려해야할 변수에 관한 선행연구들을 살펴보면, 김진산(2011)은 2002년 10월부터 2010년 12월 사이 신규 상장 회사를 대상으로 월간 평균 청약경쟁률이 전체 시장과 어떤 관계를 가지는지 실증분석을 통해 청약경쟁률이 공모주수익률과 발행수익률에 중요한 영향을 미친다는 것을 확인하였다. 곽노걸 외 1인(2015)은 2006년 7월부터 2011년 12월 사이 IPO를 진행한 315개 기업을 대상으로 연구가 진행되었는데, IPO 기업의 상장일에는 높은 수익률을 보이지만 이후에는 지속적으로 하락하고 청약경쟁률이 높았던 기업들은 상장일 수익률이 높다는 것을 알 수 있었다. 또한 상장일 당일 수익률의 변동이 큰 것은 높은 주가변동성이 원인이 되고, 상장일에 높은 수익률을 보인 공모주의 경우 한 달 이내에 확정공모가의 수준으로 주가가 하락할 가능성이 높다는 것을 확인하였다. 김주환 외 1인(2017)은 2007년 7월부터 2014년 12월까지 유가증권시장 및 코스닥 IPO 기업 데이터를 대상으로 개인투자자는 상대적으로 초기수익률이 낮은 IPO 주식을 순매수하는 경향이 있고, 이 주식의 장기적인 주가성과가 저조함을 실증분석하였다. 그리고 정보비대칭성이 큰 IPO 시장에서 개인투자자는 불리한 매매로 손실입을 가능성이 크고 특히, 공모가 미만으로 하락한 IPO 주식에서 개인투자자의 손실이 커짐을 확인하였다. 조득환 외 3인(2020)은 2009년 1월부터 2018년 12월까지 약 10년간 한국거래소에 상장한 기업 601개를 대상으로 상장시기, 수요예측경쟁률, 공모액, 밴드 상·하단 가격, 밴드수익률, 공모시장 동향의 7개 변수가 수익률 달성에 중요한 영향을 미치는 변수임을 확인하였다. 특히 상장시기와 수요예측 경쟁률은 투자 여부를 결정할 때 결과에 가장 중요한 영향을 미치는 변수라는 것을 확인하였다. 이한석 외 1인(2021)은 2010년 10월부터 2019년 12월까지 유가증권시장 및 코스닥 시장에 신규로 상장한 535개 기업을 실증분석하였으며, 개인투자자들은 상대적으로 기관투자자에 비해 정보취득 및 기업분석능력이 취약하여 시장에 과잉반응이 나타나 IPO 시장의 이상현상을 야기할 수 있음을 확인하였다.

다음은 감성분석을 통한 공모주 상장 시 가격 변화예측에 관한 선행연구이다. Liew et al.(2016)은 2013년 1월부터 2014년 12월까지 NASDAQ 혹은 NYSE의 325개 IPO 기업의 당일 거래 데이터를 통해 IPO 기업의 Tweet 감성과 상장 첫날 거래 수익률이 관계를 가지고 있다는 것을 확인하였고, IPO 이전의 감성들은 기업의 상장 첫날 수익을 예측할 수 있음을 확인하였다. 김명진 외 3인(2020)은 20개 종목에 대한 댓글을 통해 LSTM과 CNN 등의 머신러닝 기법을 활용하여 댓글을 이용하면 대부분의 주가의 이동방향과 변동폭에 대해

50% 이상 정확도로 예측 가능하다는 결론을 통해 연구기간 동안의 댓글 정보와 가격변동이 연관 있음을 확인하였다. Ye Xian(2021)은 2016년부터 2019년까지 Stocktwits의 데이터를 통해 소셜미디어 감성과 IPO 가격에 관한 실증연구를 진행하였으며, IPO 전 낙관적인 감성상태는 첫날 수익률을 높이는 경향이 있지만 장기적으로 봤을 때는 큰 영향을 미치지 않는다는 것을 확인하였다. 그리고 이는 상대적으로 적은 수의 게시물을 가진 기업들의 경우 더 효과가 좋음을 확인하였다. 김종옥 외 1인(2021)은 2016년부터 2020년까지 코스닥 시장에 신규상장한 기업의 수요예측정보, 청약경쟁률, 인터넷 검색량 데이터를 통해 개인투자자의 청약경쟁률은 상장일 시초가에 영향을 미쳤으며 상장일 인터넷 검색량의 변동이 상장일 종가에 유의미한 영향을 나타낸 것을 확인하였다. 양수연 외 3인(2022)은 2009년 6월부터 2020년 12월 사이에 신규 상장된 국내 IPO 기업데이터를 기반으로 증권신고서 내용 변수를 함께 사용한 모형의 예측정확도가 더 높았고, 증권신고서 내용이 기업의 IPO 당시 주가 등락 여부 예측에 유의미한 영향을 미치는 것을 확인하였다.

본 연구는 머신러닝 모형인 XGBoost와 랜덤 포레스트를 사용하여 개인투자자의 감성이 공모주 가격형성에 미치는 영향을 분석하였다. XGBoost 모형은 최근 많이 사용되는 모형으로 분류 문제를 해결할 때 좋은 성능을 보인다는 장점이 있다(Guo et al., 2019). 랜덤 포레스트 모형은 특정 변수를 제거하지 않아도 성능이 좋은 모형을 구축할 수 있어(송서하 외 4인, 2019), 많은 선행연구에서 사용된 모형이다(양수연 외 3인, 2022; 김나영 외 1인, 2022; Iain et al., 2012). 두 머신러닝 모형을 이용한 선행연구를 살펴보면, 하대우 외 3인(2019)은 2010년 1월 29일부터 2017년 12월 28일까지의 코스피 200 일별 데이터를 통해 자기회귀모형, LSTM 신경망, 그리고 XGBoost를 활용하여 가격 등락예측모형을 구축하였고, XGBoost 모형이 주가 예측 시 좋은 성능을 보이는 것을 확인하였다. 한지형 외 2인(2019)은 온라인 설문조사 결과 총 2,006개의 데이터를 통해 다중회귀분석과 XGBoost를 이용하여 재무스트레스에 대한 중요 예측요인들의 구체적 영향력을 분석하였다. Luckyson et al.(2016)은 랜덤 포레스트 모형을 이용하여 애플, 마이크로소프트, 삼성 등의 장기적인 주가 예측 정확도가 85-95%로 높은 것을 확인하였고, 랜덤 포레스트 모형이 주가의 방향을 예측할 때 좋은 성능을 보인다는 것을 확인하였다. Kim et al.(2022)은 1990년 4월 5일부터 2013년 1월 15일까지 5,740개의 데이터와 2009년 1월 27일부터 2020년 12월 31일까지 3,005개의 데이터를 랜덤 포레스트 모형을 이용하여 VIX를 예측하였고, 랜덤포레스트 모형이 유의미한 정확도와 우수한 예측력을 보여주었음을 확인하였다. 이를 통해 확정공모가 대비 시초가 형성을 예측 시 XGBoost와 랜덤 포레스트 모형이 좋은 성능을 발휘할 것으로 판단하였다.

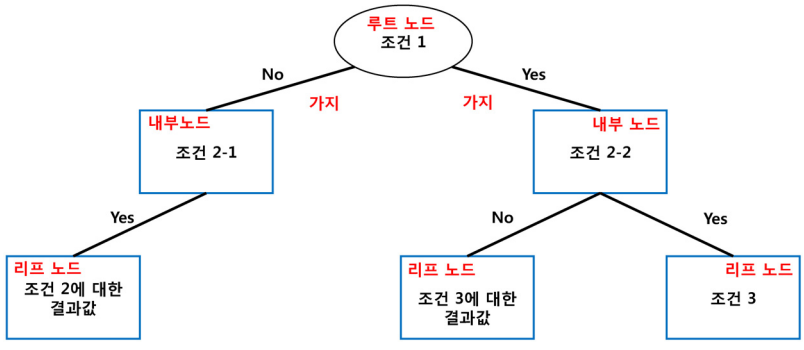
본 연구는 신고서 제출일부터 공모일 직전까지의 네이버 뉴스 및 댓글 데이터로 개인투자자

및 시장의 감성을 분석한 후, 머신러닝을 통해 감성 변수가 공모주 시초가 형성에 영향을 미치는지 분석하여 개인투자자의 감성이 공모주 가격형성에 미치는 영향에 대해 분석하였다는 점에서 차별점이 존재한다.

Ⅲ. 연구방법론

본 연구는 공모주 상장일 확정공모가 대비 시초가 형성 예측을 위한 모형으로 XGBoost, 랜덤 포레스트(Random Forest)를 활용하였다. 두 모형은 모두 여러 개의 의사결정트리(Decision Tree)가 앙상블(Ensemble)된 것을 기반으로 한다. 의사결정트리는 [그림 3]과 같이 루트 노드(a root node), 내부 노드(internal nodes), 리프 노드(leaf nodes), 가지(branch)로 구성된 구조이다. 의사결정 트리는 하위 트리에 조건을 설정한 후 변수들을 나눠 입력변수와 타겟(target) 변수 간 복잡한 관계를 단순하게 트리 구조로 나타내며, 변수를 재사용하지 않고 결측값들을 쉽게 다룰 수 있고 이해하기 쉽다는 장점을 가져(Song et al., 2015). 분류에 적합한 모형이다. 하지만 훈련 데이터로만 적용한 의사결정트리 모형은 과적합될 수 있어 이를 해결하기 위해 많은 조건이 추가되어야 한다는 단점이 존재한다. (Ho, 1995).

[그림 3] 의사결정트리 개념도

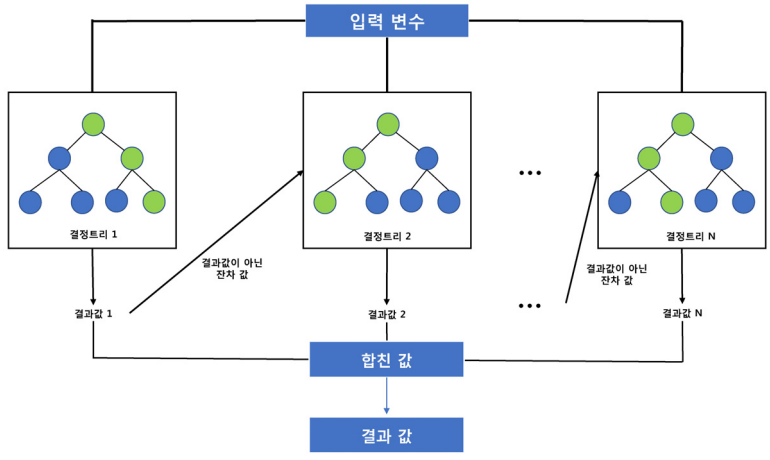


1. XGBoost와 랜덤 포레스트(Random Forest)

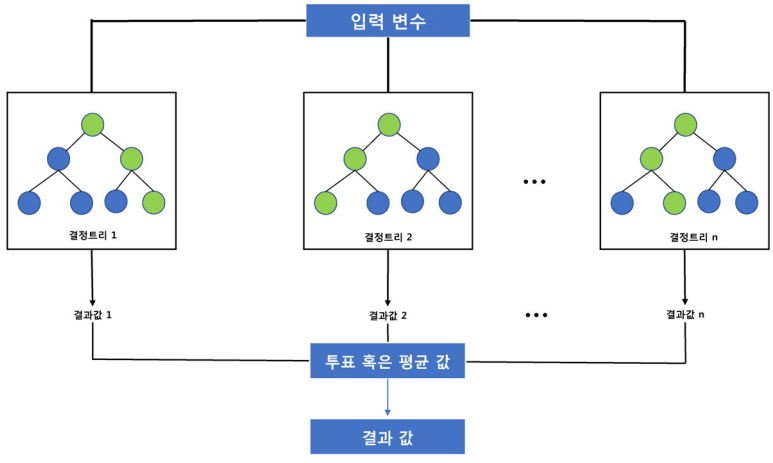
XGBoost는 eXtreme Gradient Boosting의 약자로 앙상블(Ensemble)된 의사결정트리에 부스팅(Boosting)방법을 적용하여 정확도와 확장성이 우수한 모형이다. [그림 4]는 XGBoost의 개념도로서 입력변수가 주어지면 결정트리에서 결과값이 도출되고, 오차값을 다음 층의

의사결정트리에 반영하여 정확도를 높이는 방식으로 분석이 진행되고 각 트리의 결과값들을 종합해 하나의 결과로 나타낸다. 이 모형은 Out-of-core Computation(병렬화)과 주 메모리에 맞는 연산을 통해 데이터 용량이 커져도 효율적이고 빠른 학습이 가능하다는 특징이 있다(Chen et al., 2016). 또한 다음 의사결정트리로 새롭게 연결하는 과정에서 이전 의사결정트리에서 발생한 오차값에 가중치를 주어 오차를 줄이는 최적화 과정을 통해 약한 성능의 분류기를 강한 성능의 분류기로 만들어주며, 조기중단(Early Stopping)과 같은 규제를 통해 모형이 과적합되는 것을 방지한다(Dhaliwal et al., 2018).

[그림 4] XGBoost 개념도



[그림 5] 랜덤포레스트 개념도



랜덤 포레스트(Random Forest)는 중복을 허용하여 무작위로 추출한 데이터로 의사결정트리를 구성하는 배깅(Bagging) 방식을 이용하여 만들어진 의사결정트리가 앙상블(Ensemble)된 모형이다. [그림 5]는 랜덤 포레스트의 개념도로서 입력변수가 주어지면 배깅을 통해 만들어진 여러 가지 의사결정트리에서 특정 결과값을 예측하고 이 중 가장 많이 예측된 값을 최종 예측 결과값으로 하여 데이터를 분류한다. 또한 예측 오차값 및 의사결정트리 확장을 제한하여, 의사결정트리 모형의 단점인 과적합 문제를 보완해 다양한 요인들을 분석할 수 있는 모델이다. (Breiman, 2001)

위에서 언급한 두 모형은 공통적으로 의사결정트리 기반 머신러닝 모델이지만 각각 부스팅, 배깅 방법이 적용된다는 차이가 존재한다. XGBoost는 [그림 4]와 같이 주어진 입력변수를 통해 결과값과 직전 의사결정트리를 통해 발생한 오차들을 다음 의사결정트리로 연결하고 규제를 통해 성능이 강화된 분류기를 만들어 결과값을 도출하는 방식으로 진행된다. 랜덤 포레스트는 [그림 5]와 같이 입력변수의 중복을 허용하여 다양한 의사결정트리를 생성한 후 가장 많이 투표된 결과값을 도출하는 방식으로 진행되며, 모형에 사용된 트리의 개수를 제한하여 과적합을 방지한다. 해당 연구에서는 사용된 표본의 수가 비교적 적기 때문에 과적합을 방지할 수 있는 두 모형을 사용해 더 나은 분류 성능을 도출하고자 한다.

2. KR-FinBERT

KR-FinBERT는 KR-BERT를 기반으로 만들어진 KR-BERT-MEDIUM 모형에 금융 텍스트 데이터를 학습시켜 완성한 금융 분야에 특화된 언어모형이다. KR-FinBERT의 기반 모형인 KR-BERT는 기존 모형들에 비해 상대적으로 작은 크기의 한국어 말뭉치 데이터를 활용하여 사전 학습 시간을 줄인 한국어 특화 언어 모형으로, 학습 데이터가 많아 사전 학습에 많은 시간과 비용을 필요로 하는 기존 모형들을 보완하였다. KR-BERT의 경우 학습 데이터로 한국 위키피디아 덤프(Korean Wikipedia dump)와 크롤링한 뉴스 기사 데이터를 사용하였다(Lee et al., 2020). KR-FinBERT는 KR-BERT에서 사용된 학습 데이터뿐만 아니라 한국법령센터 텍스트 데이터와 한국어 댓글 데이터셋, 72개의 신문사에서 게시된 기업과 관련된 경제 뉴스 440,067개의 제목과 내용 데이터, 16개 증권사에서 발간된 애널리스트 보고서 11,237개의 데이터가 모두 합쳐진 데이터를 사전학습 한 언어모형이다.[2] 본 연구에서는 사용된 공모 기업의 뉴스와 댓글 데이터에 주로 금융 텍스트가 쓰였을 것이라 판단하였고, 이를 기반으로 금융 텍스트 기반의 사전학습 언어모형인 KR-FinBERT를

2) KR-FinBERT 소개 Github 주소: https://github.com/snunlp/KR-FinBert.

사용하였다. 이때 수집한 뉴스와 댓글 데이터는 각 뉴스기사 전문, 각 댓글 내용 전문 수준에서 감성분류가 진행되었으며, 사전학습 모형을 사용했기에 비속어 및 비 표준어, 이모티콘, 외국어는 모형 내 처리과정을 통해 반영되었다.

<뉴스 데이터 예시>

변경 전: '\n\t\t\t[이데일리 김응태 기자] 반도체 고압 수소 어닐링 공정기술 선도기업 에이치피에스피(HPSP)는 지난 6~7일 일반 투자자 대상 공모주 청약 결과 1,159.05대 1의 경쟁률을 기록했다고 7일 밝혔다.\n\n\n이번 청약은 전체 공모 물량의 25%에 해당하는 75만주를 대상으로 균등 및 비례 방식 각각 50%씩 배정돼 진행했다. 8억 6,929만주가 청약 접수됐으며, 증거금은 약 10조 8,661억 원을 기록했다. 앞서 지난 수요예측에서도 1,577개 기관이 참여해 1511.36대 1의 높은 경쟁률을 기록했다. 공모가 희망밴드도 상단 2만 5,000원으로 확정됐다. 에이치피에스피는 반도체 전공정 중 어닐링 공정에서 세계 유일의 고압 수소를 활용한 어닐링 장비 'GENI 시리즈(Series)'를 보유하고 있다. 이 장비는 기술 집약도와 정밀도가 높은 16㎚ 이하 선단공정에서 필수적으로 사용된다. 시스템 반도체에 적용되던 것이 메모리 반도체까지 확장된 만큼 향후 수요 증가가 지속된다는 게 회사 측 설명이다. 에이치피에스피는 공모자금 중 230억 원을 연구개발비, 시설자금 등에 활용할 계획이다. 특히 고압 습식 산화공정(Wet Oxidation) 기술 개발, 신규 가스 공정 개발 및 자율주행차 탑재 운전자 보조 시스템(ADAS), CMOS 이미지 센서(Image Sensor) 등 장비 고도화에 집중 투자한다는 전략이다. 김용운 에이치피에스피 대표는 "코스닥 상장 이후 투명하고 책임감 있는 경영 활동으로 주주가치 제고에 노력하겠다"며 "반도체 고압 수소 어닐링 기술을 기초로 반도체 전공정 장비분야에서 꾸준한 기술 개발을 통해 글로벌 시장을 선도할 것"이라고 말했다.\n'

변경 후: '김응태 반도체 고압 수소 어닐링 공정기술 선도기업 에이치피에스피는 지난 67일 일반 투자자 대상 공모주 청약 결과 1159 05대 1의 경쟁률을 기록했다고 7일 밝혔다 이번 청약은 전체 공모 물량의 25에 해당하는 75만주를 대상으로 균등 및 비례 방식 각각 50씩 배정돼 진행했다 8억 6,929만주가 청약 접수됐으며 증거금은 약 10조8661억원을 기록했다 앞서 지난 수요예측에서도 1,577개 기관이 참여해 1511 36대 1의 높은 경쟁률을 기록했다. 공모가 희망밴드도 상단 2만 5,000원으로 확정됐다. 에이치피에스피는 반도체 전공정 중 어닐링 공정에서 세계 유일의 고압 수소를 활용한 어닐링 장비 시리즈를

보유하고 있다. 이 장비는 기술 집약도와 정밀도가 높은 16㎚ 이하 선단공정에서 필수적으로 사용된다. 시스템 반도체에 적용되던 것이 메모리 반도체까지 확장된 만큼 향후 수요 증가가 지속된다는 게 회사 측 설명이다 에이치피에스피는 공모자금 중 230억 원을 연구개발비 시설자금 등에 활용할 계획이다. 특히 고압 습식 산화공정 기술 개발 신규 가스 공정 개발 및 자율주행차 탑재 운전자 보조 시스템 이미지 센서 등 장비 고도화에 집중 투자한다는 전략이다. 김용운 에이치피에스피 대표는 코스닥 상장 이후 투명하고 책임감 있는 경영 활동으로 주주가치 제고에 노력하겠다며 반도체 고압 수소 어닐링 기술을 기초로 반도체 전공정 장비분야에서 꾸준한 기술 개발을 통해 글로벌 시장을 선도할 것이라고 말했다'

: 긍정(0.9963), 중립(0.0033), 부정(0.0004)

<댓글 데이터 예시>

변경 전: '소액주주들 피눈물나게 하고 아무런 대책도 없이 상장해서 저놈들 배 때지 채울려고 하는 사기꾼, 개양아치 그룹이네 이런 양아치 놈들에게 우리가 할 수 있는 방법은 단 하나 불매운동을 행동으로 보여줘야 합니다'

변경 후: '소액주주들 피눈물나게 하고 아무런 대책도 없이 상장해서 저놈들 배 때지 채울려고 하는 사기꾼 개양아치 그룹이네 이런 양아치 놈들에게 우리가 할 수 있는 방법은 단 하나 불매운동을 행동으로 보여줘야 합니다'

: 긍정(0.0004), 중립(0.0385), 부정(0.9611)

3. 감성점수 도입

$$\text{감성점수} = (-1\times\text{부정})+(0\times\text{중립})+(1\times\text{긍정}) \tag{1}$$

$$\text{F1점수} = \frac{2\times\text{재현율}\times\text{정밀도}}{(\text{재현율}+\text{정밀도})} \tag{2}$$

$$\text{정밀도} = \frac{TP}{TP+FP} \tag{3}$$

$$\text{재현율} = \frac{TP}{TP+FN} \tag{4}$$

수집한 감성데이터를 변수로 사용하기 위해 3.2에서 언급한 KR-FinBERT를 통해 텍스트 데이터를 긍정, 중립, 부정일 확률을 구하고 가장 큰 확률값을 기준으로 감성상태를 분류하였다. 그 결과 <표 1>과 같이 뉴스 데이터와 댓글 데이터 모두 중립일 확률의 경우가 많았다. 감성상태가 중립일 경우 각 텍스트의 정확한 감성상태를 파악하기 어려워 이를 보완하기 위해 각 감성상태에 대한 확률값을 식 (1)에 대입하여 각 텍스트의 감성점수를 구하는 방법을 제안하였다. 실제 Mohapatra et al.(2020)의 논문에서 트윗의 감성을 통해 암호화폐 가격을 예측하기 위해 감성점수화 하는 과정에서 연구자가 감성점수 도출 수식을 적용하였다. <표 3>은 식 (1)을 반영하기 전후 결과를 나타낸 것으로, 결과를 나타내는 수치로 정확도와 F1 점수를 사용하였다. F1 점수는 식 (2)와 같이 나타나며 정밀도와 재현율의 조화평균의 값으로 이 수치가 클수록 좋은 성능을 보인다. F1 점수를 구하는데 사용된 정밀도와 재현율은 각각 식 (3), 식 (4)에 명시되어 있으며, TP(True Positive)는 맞는 것을 올바르게 예측한 경우, FP(False Positive)는 틀린 것을 맞다고 잘못 예측한 경우, FN(False Negative)은 맞는 것을 틀렸다고 잘못 예측한 경우를 의미한다. 식 (1)을 반영한 것이 랜덤 포레스트 모형에서는 차이가 없지만, XGBoost 모형에서 상대적으로 좋은 성능을 보였기 때문에 기업의 감성점수를 반영할 때 식 (1)을 적용해 개선된 성능을 도출하고자 한다. 식 (1)을 적용하여 구한 감성점수는 <표 2>와 같이 뉴스 데이터의 경우 긍정이 약 72%, 부정이 약 28%이고, 댓글 데이터의 경우 긍정이 약 25%, 부정이 약 75%로 서로 상반된 결과를 나타낸다. 이는 공모주 IPO 금액이 2018년 약 3조 5천억에서 2021년 중반 약 6조원까지 상승하고 개인청약율이 크게 상승하여(이석훈, 2021) 공모주 시장이 호황인 상황에서 좋은 상황을 전달하는 뉴스 기사가 많다보니 중립을 기초로 하지만 긍정적인 감성상태가 많았을 것으로 판단된다. 반면, 댓글 데이터의 경우 댓글 게시자의 약 75%는 관련 기사에 대한 이해도가 낮고 상대방을 존중하지 않는 사람이라는(조수선, 2007) 선행연구 결과를 근거로 부정적인 감성상태가 높게 나타났다고 추론하였다.

<표 1> 감성점수 수식 반영 전 감성상태 결과 표

웹 크롤링을 통해 수집된 뉴스와 댓글 데이터를 감성분석하여 긍정, 중립, 부정으로 분류한 결과표이다.

	뉴스 데이터	댓글 데이터
긍정	26.99% (7,637개)	6.18% (2,924개)
중립	62.33% (17,638개)	83.54% (39,543개)
부정	10.68% (3,023개)	10.29% (4,869개)

<표 2> 감성점수 수식 반영 후 감성상태 결과 표

감성 데이터에서 중립의 비중이 과도하게 많아 감성변수의 영향력을 개선하기 위해 식 1)을 반영해 각 데이터 마다 가지고 있는 극성을 나타내 분류한 표이다.

	뉴스 데이터	댓글 데이터
긍정	71.69% (20,287개)	30.13% (14,260개)
중립	0% (0개)	0% (0개)
부정	28.31% (8,011개)	69.87% (33,076개)

<표 3> 감성점수 수식 적용 전 후 성능 비교

감성점수 수식 반영 전 후 결과를 비교하기 위한 표이며, XGBoost와 랜덤 포레스트 두 가지 모델을 적용해 각각의 경우에서 확정공모가 대비 시초가가 1.0배 초과, 1.5배 초과할 경우의 예측정확도 및 F1점수를 나타낸다.

감성변수 포함한 경우	XGBoost				랜덤 포레스트			
	1.0배 초과		1.5배 초과		1.0배 초과		1.5배 초과	
	정확도	F1점수	정확도	F1점수	정확도	F1점수	정확도	F1점수
감성수식 적용 전	0.7794	0.8624	0.7500	0.5405	0.7500	0.8522	0.6176	0.0000
감성수식 적용 후	0.8529	0.9000	0.7794	0.5946	0.7500	0.8522	0.6176	0.0000

Ⅳ. 실험 및 결과

1. 실험 데이터

2018년 1월 1일부터 2022년 7월 15일까지 신규 상장한 공모 기업의 데이터를 이용하여 연구를 진행하였다. 해당 기간동안 총 397개의 기업이 공모 후 상장이 이뤄졌으며, 부동산투자회사(Real Estate Investment Trusts), 기업인수목적회사(Special Purpose Acquisition Company) 그리고 신고서제출일과 상장예정일 사이에 뉴스 및 댓글 데이터가 없거나 재공모를 통해 상장된 기업 및 상장예정일과 실제 상장일이 다른 기업을 제외한 337개의 기업 데이터를 연구 표본으로 선정하였다. 여기서 기업명, 상장 시장(코스피, 코스닥), 신고서제출일, 상장예정일 자료는 '한국거래소 전자공시 홈페이지(KIND)'에서 수집하였다. 선정된 표본으로 <표 4>에 언급된 변수들을 사용하여 실험을 진행하였고, 관련 요약 통계량은 <표 5>와 같다.

<표 4> 실험에 사용한 변수

사용변수	사용데이터	수집경로
감성변수	뉴스 감성점수 댓글 감성점수 뉴스 개수 댓글 개수	네이버 뉴스[3]
거시경제변수	원/달러 환율 CD(91일) 금리 뉴스심리지수	한국은행 경제통계시스템 API[4]
	코스피 지수(수정종가) 코스닥 지수(수정종가) 금 종가 (일별 데이터)	Yahoo! Finance 파이썬 라이브러리 yfinance
	원유 종가 (일별 데이터, 두바이유)	인베스팅 닷컴[5]
공모주 관련 변수	개인투자자 청약경쟁률	38 커뮤니케이션[6]
	공모금액(백만원) 확정공모가	기업공시채널(KIND)[7]

<표 5> 사용 변수에 대한 요약 통계량

<표 4>에 사용된 변수에 대한 요약 통계량을 나타낸 표이다.

	개수	평균값	표준편차	최소값	25%	75%	최대값
뉴스 감성점수	337	0.17	0.16	-0.42	0.06	0.31	0.61
댓글 감성점수	337	-0.03	0.12	-0.63	-0.09	0.02	0.68
뉴스 개수	337	82.56	114.21	4	50	83	1742
댓글 개수	337	139.8	738.61	1	14	65	12105
원/달러 환율 종가	337	1156.95	46.18	1059.5	1122.8	1186.3	1310
CD(91일)금리	337	1.27	0.47	0.63	0.74	1.65	2.58
뉴스심리지수	337	101.35	9.54	70.71	95.07	109.62	122.56
코스피 지수	337	2514.19	415.65	1834.3	2125.3	2959.5	3302.8
코스닥 지수	337	828.84	134.57	563.49	695.72	946.31	1060
금 종가	337	1622.64	261.51	1186.8	1331.9	1837.9	2051.5
원유 종가	337	66.13	16.61	30.34	58.40	72.81	113.83
개인투자자 청약경쟁률	337	818.14	843.94	0.19	92.69	1176.45	6762.75
공모금액 (백만원)	337	132803.7	766357.15	4500	16789	53333	12750000
확정공모가	337	22102.97	35405.62	1500	10000	25000	498000

3) 네이버 뉴스 https://news.naver.com/.
4) 한국은행 경제통계시스템 https://ecos.bok.or.kr/api/#/.
5) 인베스팅 닷컴 https://kr.investing.com/.
6) 38 커뮤니케이션 http://www.38.co.kr/.
7) 기업공시채널 KIND https://kind.krx.co.kr/.

개인투자자의 관심도를 표현하기 위해 선정된 표본 337개 공모주의 신고서제출일로부터 상장예정일 사이 네이버 뉴스에 게시된 뉴스, 댓글들을 웹 크롤링을 통해 수집하였다. 수집된 텍스트 데이터 중 뉴스 데이터는 총 35,787개, 댓글 데이터는 총 61,194개이다. 텍스트 데이터를 수집 시 종목명을 기준으로 하다 보니 특정 기간동안 상장이 예정된 여러 공모 기업들이 동시에 포함된 뉴스, 당시 시황을 알려주는 뉴스와 같이 하나의 기사에 종목명이 중복으로 포함된 뉴스가 존재하였다. 중복된 뉴스와 시황을 알려주는 뉴스는 모두 제외하였고, 추가로 댓글이 없는 뉴스도 모두 제외하였다. 댓글 데이터의 경우 여러 뉴스에 중복된 내용의 문장을 작성하는 '도배행위'의 경우가 있어 중복된 댓글을 제외하였다. 텍스트 처리 후 뉴스 데이터는 28,298개, 댓글 데이터는 47,336개이며, 불용어[8] 및 텍스트 데이터 내 필요하지 않은 단어들[9]을 제거하였다. 정제된 텍스트 데이터는 3.3에서 언급한 바와 같이 하나의 감성점수를 추출하고, 각 기업에 해당하는 감성 데이터들의 감성점수 평균값을 기업의 감성점수 변수로 사용하였다. 또한 각 기업의 뉴스, 댓글의 개수도 개인투자자들의 관심도가 반영될 것이라고 판단하여 수집된 텍스트 데이터에서 각 기업의 뉴스 수와 댓글 수를 변수로 추가하였다.

IPO 시장에서 가격 변화의 원인 중 하나로 거시, 시장 환경적 요소가 존재한다(이석훈, 2014)는 선행연구 결과에 따라 기존 선행연구에서 많이 사용된 공모주 상장 당일의 거시경제 변수인 코스피 지수, 코스닥 지수, 환율, 원유, CD금리(91일), 금가격을 변수로 사용하였다. 추가적으로 뉴스심리지수[10]라는 지표를 사용하여 뉴스로 나타난 우리나라의 전반적인 경제상황을 나타내고자 하였다.

공모주 관련 변수는 개인투자자의 관심도에 초점을 맞춰 공모주 상장일 가격 변화에 대한 연구를 진행했기 때문에 청약경쟁률, 공모금액, 확정공모가를 변수로 활용하였다. 청약경쟁률(competition rate)은 공모주에 대한 개인투자자의 수요를 나타내는 지표로, 38커뮤니케이션에서 연구 표본에 해당하는 기업들의 데이터를 수집하여 분석에 사용하였다. 공모금액(백만원)(amount)은 공모주에 공모된 총 금액을 의미하며, 확정공모가(collusion price)는 공모주 상장 전 공개모집 결과 수요 예측 등을 거쳐 확정된 가격을 의미한다. 공모금액(백만원), 확정공모가 데이터는 기업공시채널 KIND에서 수집하였다.

---

8) [,], #, \n, ‘, “, &, /, ., , 의 기호를 제거하였다.

9) ‘무단 전재 및 재배포 금지’, ‘기자’, 뉴스 언론사, ‘https’와 같은 텍스트 데이터를 제거하였다.

10) 뉴스심리지수: 국내포털사이트의 경제분야 뉴스에서 문장을 추출 후 각 문장에 나타난 긍정 부정 중립의 감성을 기계학습 방법으로 분류해 긍정 부정 문장 수 차이를 계산해 지수화 한 지표

2. 실험 결과

1) 실험 조건

본 연구의 목적인 감성변수의 유무에 따른 공모주 가격형성 예측의 정확도를 분석하는 방법으로 확정공모가 대비 시초가의 등락 여부를 분석하기 위한 분류 모형과 상장 직후 수익률을 분석하기 위한 회귀 모형도 적용하였다. 회귀 모형 적용 시 모형 예측 성능이 좋지 않았고 감성변수의 유무가 예측에 큰 영향을 미치지 못했기 때문에 분류 모형을 적용하였다.

개인투자자의 관심도가 공모주 상장 당일 시초가 형성에 미치는 영향에 대해 분석하고자 타겟 데이터를 연구표본의 확정공모가 대비 상장 당일 시초가가 1.0배, 1.5배 초과 상승한 경우를 1로 설정하고, 그렇지 않은 경우를 0으로 설정하였다.

타겟 데이터 설정 후 총 337개의 표본을 시간순으로 나누어 전체의 80%인 269개 데이터를 훈련세트로 사용하였고, 테스트 세트로 전체의 20%인 68개 데이터를 사용하였다. 훈련세트와 테스트 세트의 타겟 데이터 비율은 아래 <표 6>과 같다.

<표 6> 훈련세트, 테스트세트의 타겟 데이터 비율

이 표는 XGBoost, 랜덤 포레스트 모형에 사용된 데이터를 훈련세트와 테스트세트로 나눴을 때 각 세트의 0과 1의 비율을 나타낸다.

배 수	Train의 0의 비율	Train의 1의 비율	Test의 0의 비율	Test의 1의 비율
1.0	0.2639	0.7361	0.2500	0.7500
1.5	0.7175	0.2825	0.6176	0.3824

4.1장에서 언급한 변수들을 아래에 언급한 바와 같이 4가지 실험으로 분류하였고, 실험2를 기본모형으로 하여 XGBoost, 랜덤 포레스트 모형에 적용한 후 그리드 서치(GridSearch)[11]를 진행하였다. 그 후 사전에 설정한 하이퍼파라미터 중 각 모형의 성능을 가장 좋게 만들고 과적합을 방지할 수 있는 최적 파라미터 값을 구해 모형별 최적의 성능을 비교 분석하였다. 또한 Accuracy(정확도)를 주요 평가 지표로 두고 F1 점수를 보조 평가지표로 두어 각 모형 및 분석결과에 대한 성능을 비교하였다.

실험 1: 뉴스데이터와 댓글 데이터를 포함한 경우

11)<표 9>에 나타난 하이퍼 파라미터와 설정값을 사용하였다.

실험 2: 뉴스데이터와 댓글 데이터가 포함되지 않은 경우
실험 3: 뉴스 데이터만 포함된 경우
실험 4: 댓글 데이터만 포함된 경우

2) 실험결과

<표 7>은 위에서 언급한 조건으로 설정된 4가지 실험에 대한 정확도와 F1 점수 결과표이다. 실험에 사용된 모형 중 XGBoost가 랜덤 포레스트에 비해 상대적으로 좋은 성능을 보였다. 실험 1에서 1.0배 초과의 경우 XGBoost가 85.29%, 랜덤 포레스트가 77.94%의 정확도를 보이는 반면 실험 2에서 1.0배 초과의 경우 XGBoost가 76.47% 랜덤 포레스트는 75.00%의 정확도를 보이며, 실험1이 상대적으로 높은 성능을 보인다는 것을 확인할 수 있다. 하지만 실험 1에서 1.5배 초과의 경우 XGBoost가 77.94%, 랜덤 포레스트가 61.76%의 정확도를 가지고, 실험 2에서 1.5배 초과의 경우 XGBoost는 실험 1과 동일하게 77.94%의 정확도를 가지지만 F1점수에서 0.6512로 미묘하지만 좋은 성능을 보인다. 랜덤 포레스트의 정확도도 77.94%로 오히려 실험 2가 더 좋은 성능을 보인다는 것을 알 수 있다.

<표 7> 실험 별 정확도 결과표

각 실험 별 정확도와 F1 점수를 나타낸 표로 왼쪽이 XGBoost, 오른쪽이 랜덤 포레스트의 성능 결과표를 나타낸다.

	XGBoost				랜덤 포레스트			
	1.0배 초과		1.5배 초과		1.0배 초과		1.5배 초과	
	정확도	F1점수	정확도	F1점수	정확도	F1점수	정확도	F1점수
실험1	0.8529	0.9000	0.7794	0.5946	0.7794	0.8624	0.6176	0.0000
실험2	0.7647	0.8621	0.7794	0.6512	0.7500	0.8496	0.7794	0.6939
실험3	0.8088	0.8687	0.7353	0.5263	0.7647	0.8621	0.7794	0.6809
실험4	0.7500	0.8468	0.7500	0.5405	0.7353	0.8421	0.6176	0.0000

<표 8>은 실험 1의 1.0배 초과에 대한 XGBoost, 랜덤 포레스트, 그리고 로지스틱스 회귀 이중분류 모형의 결과표이다. <표 7>의 실험 결과에 따라 실험 1에서 1.0배 초과의 경우가 가장 유의미하다고 보여지며, 1.5배 초과의 경우 성능이 낮아짐을 알 수 있다. 추가적인 실험을 위해 전통적인 방법인 로지스틱스 회귀 이중분류에 적용하여 머신러닝 모형과 비교하였다. 실험 결과 1.0배 초과의 경우 XGBoost, 랜덤 포레스트의 정확도가 각각 약 85%, 약 78%로 로지스틱스 방법론의 정확도인 약 75%에 비해 상대적으로 좋은 성능을 보인다. 하지만 1.5배의 초과의 경우 XGBoost, 랜덤 포레스트의 정확도가 각각 약 78%,

약 62%로 로지스틱스 방법론의 정확도인 약 79%에 비해 상대적으로 낮은 성능을 보인다.

<표 8> 로지스틱스 회귀 모형을 추가한 실험1의 결과표

각 모형 별 뉴스, 댓글 데이터를 모두 포함한 경우인 실험 1 1.0배, 1.5배 초과에 대한 정확도와 F1 점수를 나타낸 표이다. 아래로 가며 XGBoost, 랜덤 포레스트, 로지스틱스 회귀 이중분류 모형에 대한 결과표이다.

		실험 1	
		정확도	F1 점수
XGBoost	1.0배 초과	0.8529	0.9000
	1.5배 초과	0.7794	0.5946
랜덤 포레스트	1.0배 초과	0.7794	0.8624
	1.5배 초과	0.6176	0.0000
로지스틱스 회귀 이중분류	1.0배 초과	0.7500	0.8547
	1.5배 초과	0.7941	0.7083

실험 결과를 통해 XGBoost 모형이 예측 시 상대적으로 좋은 성능을 보여준다는 것을 알 수 있다. 또한 머신러닝 모형을 이용하여 공모주 상장 당일 가격형성 예측 시 감성 변수를 포함할 경우 확정공모가 대비 시초가가 1.0배 초과 상승할 것을 더 잘 예측하지만, 시초가가 1.5배 초과 상승 예측 시 전통적인 방법론을 이용한 경우나 감성 변수를 포함하지 않은 경우가 상대적으로 좋은 성능을 보인다는 것을 알 수 있다. 따라서, 감성 변수의 유무는 공모주 상장 당일 확정공모가 대비 시초가의 상승 여부를 예측하는데 도움이 될 수 있으나, 초과 상승분에 대한 예측은 도움되지 않을 수 있다고 판단된다.

V. 결론 및 한계점

최근 몇 년간 개인의 공모주 투자에 대한 관심도 및 참여가 증가했으며, 개인투자자의 공모주 투자에 대한 영향력 증가에 따른 투자 수익 연구 등 후속 연구들도 점차 늘어나고 있다. 본 연구는 2가지 가정을 전제로 한다. 첫 번째 가정은 개인투자자의 공모 참여에 대한 관심이 공모주 상장 당일 시초가 형성에 영향을 미치며, 개인투자자는 공모 기업의 IPO 관련 뉴스 뿐만 아니라 공모 기업의 전반적인 활동에 관한 뉴스에 노출되며 투자 판단에 영향받는다고 가정한다. 두 번째 가정은 투자자들은 상장 당일 초과 수익을 얻기 위해 공모주를 단타성으로 매매할 것이라고 가정한다. 개인투자자의 관심도에 따른 공모주 상장 직후 가격형성에 대해 분석하기 위해 2018년 1월부터 2022년 7월 15일까지 상장된 공모주 데이터를 수집하였다. 수집된 공모 기업들의 공모주 신고서 제출일부터 상장일까지

게시된 네이버 뉴스와 댓글을 수집하여 감성분석 후 변수로 사용하였으며, 거시경제변수들과 공모주 상장 시 기관 등에서 제공하는 공모주 관련 변수를 사용하여 시초가 형성에서 어떤 영향을 미치는지 실증분석하였다. 사용된 머신러닝 모형은 XGBoost와 랜덤 포레스트이고 두 모형 중 XGBoost가 상대적으로 좋은 성능을 보인다.

개인투자자의 감성변수는 확정공모가 대비 시초가가 1.0배 초과 상승할것이라 예측했을 경우 감성점수를 포함하지 않았을 경우보다 더 높은 성과를 보였으며, 확정공모가보다 시초가가 1.5배 초과 높을 것이라 예측했을 경우 감성점수를 포함한 경우가 더 낮은 성과를 보였다. 이는 개인투자자의 관심도가 포함된 모형이 확정공모가 대비 시초가의 초과 상승 여부를 판단하는 데에는 도움이 되지만, 초과 상승분에 대해 예측하는 것은 어렵다는 점을 시사한다.

본 연구는 선행 연구에서 개인투자자의 관심도를 반영할 수 있는 인터넷 검색횟수와 같은 변수에서 더 나아가 머신러닝 모형을 이용하여 개인투자자의 직접적인 관심도를 나타내는 댓글과, 개인투자자에게 투자 정보를 제공하는 뉴스를 감성 변수화하여 공모주 시초가 형성에 미치는 영향을 실증 분석했다는 점에서 의의를 가진다. 또한 감성변수가 공모주 투자 시 참고할 수 있는 지표로 사용될 수 있다는 것을 실증분석 하였다는 점에서 의의를 가진다. 하지만 본 연구는 다음의 측면에서 한계점이 존재한다. 첫째, 기관투자자가 매입희망수량과 가격을 제시하는 등 공모주의 가치를 산정하는 수요예측경쟁률이라는 변수를 제외하고 분석하였기 때문에 추후에 개인의 관심도 뿐만 아니라 기관의 관심도까지 포함하여 가격형성 예측을 해볼 필요가 있다. 둘째, 네이버에 게시된 뉴스와 댓글 데이터만 수집하여 다양한 반응을 담아내지 못했으며, 개인투자자의 관심도가 나타나는 다른 커뮤니티나 다른 포털사이트의 뉴스와 댓글을 수집하여 데이터의 표본을 늘려 분석해 볼 필요가 있다. 셋째, 규모가 작은 공모주에 반해 규모가 큰 공모주의 경우 개인의 관심이 훨씬 높고 큰 경향을 보이는 것으로 보아 공모가액 규모에 따른 구분 후 예측하는 방법을 사용하는 등 데이터를 범주화하여 분석한다면 다른 통찰을 얻을 수 있을 것이다.

참 고 문 헌

곽노걸, 전상경,“IPO 저가 발행의 저주: 공모주 상장 초기 주가행태 분석”, 재무관리연구, 제32권 제2호, 2015, 143-169.

금융위원회, 금융감독원, 금융투자협회, 기업공개(IPO) 공모주 일반청약자 참여기회확대 방안, 금융위원회, 2020.

김나영, 유석종, “한국어 텍스트 질의를 활용한 주식 정보 검색 및 분석 기법”, 한국정보기술학회논문지, 제20권 제9호, 2022, 13-18.

김명진, 류지혜, 차동호, 심민규, “SNS 감성 분석을 이용한 주가 방향성 예측: 네이버주식 토론방 데이터를 이용하여”, 한국전자거래학회지, 제25권 제4호, 2020, 61-75.

김성환, 전성배, “공모가 밴드가 신규공모주 (IPO) 저평가에 미치는 영향”, 한국재무학회 학술대회, 2012, 36-64.

김종옥, 최문수, “투자자 관심이 IPO 초기성과에 미치는 영향에 관한 실증분석”, 대한경영학회지, 제34권 제11호, 2021, 2023-2046.

김종옥, “개인투자자의 관심이 IPO 가격 조정에 미치는 영향에 관한 실증연구: KOSDAQ 시장을 중심으로”, 서비스마케팅학회 학술대회 발표논문집, 2022, 22-24.

김주환, 박진우, “개인투자자 거래행태와 IPO 주가성과”, 산업경제연구, 제30권 제1호, 2017, 83-103.

김진산, “공모주 청약경쟁률이 IPO 수익률에 미치는 영향 분석”, 금융지식연구, 제9권 제1호, 2011, 39-62.

류선진, “RNN 기반 뉴스기사 감성분석을 이용한 기업부도예측 모형 개발”, 한국과학기술원, 2021.

민재훈, “기관투자자와 개인투자자의 IPO 주식 투자 성과 분석”, 전문경영인연구, 제20권 제3호, 2017, 75-98.

양수연, 이채록, 원종관, 홍태호, “증권신고서의 TF-IDF 텍스트 분석과 기계학습을 이용한 공모주의 상장 이후 주가 등락 예측”, 지능정보연구, 제28권 제2호, 2022, 237-262.

이석훈, IPO 공모주 주가 변화에 대한 분석 및 시사점, [KCMI] 이슈 & 정책, 2014.

이석훈, 최근 IPO 시장의 개인투자자 증가와 수요예측제도의 평가, [KCMI], 2021.

이한석, 반주일, “심리적 가격책정이 IPO 시장에 미치는 영향”, 재무관리연구, 제38권 제2호, 2021, 129-166.

조득환, 류호선, 정승환, 오경주, “인공지능 (AI) 을 활용한 공모주 투자여부 및 기준 수익률

달성 여부 예측 모델", 한국데이터정보과학회지, 제31권 제3호, 2020, 579-590.

조수선, "온라인 신문 댓글의 내용분석: 댓글의 유형과 댓글 게시자의 성향", 커뮤니케이션학 연구, 제15권 제2호, 2007, 65-84.

하대우, 김영민, 안재준, "XGBoost 모형을 활용한 코스피 200 주가지수 등락 예측에 관한 연구", 한국데이터정보과학회지, 제 30권 제 3호, 2019, 655-669.

한지형, 고대균, 최현자, "머신러닝을 통한 가계의 재무스트레스 영향요인 예측 및 분석: XGBoost의 활용", 소비자학연구, 제 30권 제 2호, 2019, 21-43.

Breiman, L., "Random Forests," *Machine Learning*, 45, (October 2001), 5-32.

Chen, T. and C. Guestrin, "Xgboost: A Scalable Tree Boosting System," In *Proceedings of the 22nd ACM Sigkdd International Conference on Knowledge Discovery and Data Mining*, (August 2016), 785-794.

Dhaliwal, S. S., A. A. Nahid, and R. Abbas, "Effective intrusion detection system using XGBoost", *Information*, 9(7), (June 2018), 149.

Ho, T. K., "Random Decision Forests", *IEEE, In Proceedings of 3rd International Conference on Document Analysis and Recognition*, 1, (August 1995), 278-282.

Guo, J., L. Yang, R. Bie, J. Yu, Y. Gao, Y. Shen, and A. Kos, "An XGBoost-based Physical Fitness Evaluation Model Using Advanced Feature Selection and Bayesian Hyper-Parameter Optimization for Wearable Running Monitoring," *Computer Networks*, 151, (January 2019), 166-180.

Liew, J. K. S. and G. Z. Wang, "Twitter Sentiment and IPO Performance: A Cross-sectional Examination," *The Journal of Portfolio Management*, 42(4), (July 2016), 129-135.

Kim, B. Y. and H. Han, "Multi-Step-Ahead Forecasting of the CBOE Volatility Index in a Data-Rich Environment: Application of Random Forest with Boruta Algorithm," *The Korean Economic Review*, 38, (April 2022), 541-569.

Khaidem, L., S. Saha, and S. R. Dey, "Predicting the Direction of Stock Market Prices Using Random Forest", arXiv preprint, 1605(3), (April 2016).

Lee, S., H. Jang, Y. Baik, S. Park, and H. Shin, "Kr-bert: A small-scale Korean-specific Language Model", arXiv preprint, 2008(3979), (August 2020).

Saleh, S. S. and N. Che-Yahya, "Influence of Individual Investors' Sentiment in the Pre-Market and Post-Market on Malaysian IPO Initial Return," *Emprical Economics Letters*, 20, (May 2021), 49-60.

Mohapatra, S., N. Ahmed, and P. Alencar, “KryptoOracle: A Real-time Cryptocurrency Price Prediction Platform Using Twitter Sentiments,” In *2019 IEEE International Conference on Big Data (Big Data)*, (December 2019), 5544-5551.

Song, Y. Y., and L. U. Ying, “Decision Tree Methods: Applications for Classification and Prediction,” *Shanghai Archives of Psychiatry*, 27(2), (April 2015), 130-135.

Xian, Y., “Social Media Sentiment and IPO Pricing,” SSRN 3870563, (March 2021).

참 고 사 항

<표 9> 모형 별 사용 하이퍼파라미터 표

이 표는 XGBoost와 랜덤포레스트 GridSearch 시 사용한 하이퍼파라미터와 설정값을 나타낸다.

XGBoost		랜덤 포레스트	
변수	설정값	변수	설정값
CV	3	CV	3
n_estimators	[100,200,300,400,500]	n_estimators	[100,200,300,400,500]
max_depth	[3,4,5,6,7]	max_depth	[3,4,5,6,7]
min_child_weight	[1,3,5,7,9]	min_samples_leaf	[1,3,5,7,9]
gamma	[0,1,2,3]	min_samples_split	[2,3,4,5,6]
eta	[0.01,0.05,0.1]	max_features	[1,3,5,7,9]
eval_metric	log_loss	criterion	log_loss
sub_sample	[0.5,0.6,0.7,0.8,0.9,1]		
clasample_bytree	[0.3,0.5,0.7,1]		
objective	binary: logistic		

THE KOREAN JOURNAL OF FINANCIAL MANAGEMENT
Volume 40, Number 1, February 2023

Prediction of Initial Price Formation and Impact Analysis of Pubilc Offering Stock Listing Data Using News and Comment Text: Based on XGBoost*

Nam Hyeon Choi* · Kang San Kim** · Hui Su Jang***

〈Abstract〉

This study assumes that news and comments can affect price formation immediately after IPO. The empirical analysis was conducted on 337 publicly traded companies newly listed from 2018.1 to 2022.7 to see if the initial price affects the formation immediately after the IPO listing by using the comment data including the interest of individual investors before the IPO and the news data in which the comment is generated. In order to utilize news and comment data, KR-FinBERT model was applied to news and comments posted on Naver News from the date of submission of each company's declaration to the date of listing, and it was classified into positive, negative, and neutral, and then used as an emotional scoring variable. In addition, we analyzed the macroeconomic situation and individual and market responses using various variables affecting IPO price formation to determine whether the initial price is 1.0 times and 1.5 times higher than the final IPO price on the day of IPO. As a result of the study, when the initial price is more than 1.0 times higher than the final offer price, the performance is better than when the emotional variable is not included, but when the initial price is more than 1.5 times, the performance is lower than when the emotional variable is not included. Through this, it was confirmed that the emotional variables are helpful in predicting the initial formation compared to the final public offering, but not in predicting the excess increase.

Keywords : Classification, Sentiment Analysis, Forecasting, IPO Shares, XGBoost

* First Author, Bachelor's Student, School of Finance, Soongsil University, E-mail: holicman7@naver.com
** Co-Author, Bachelor's Student, School of Finance, Soongsil University, E-mail: rm7348@naver.com
*** Corresponding Author, Professor, School of Finance, Soongsil University, E-mail: yej523@ssu.ac.kr