Asia-Pacific Journal of Business (아태비즈니스연구)
Vol. 13, No. 2, June 2022 (pp.107-122)
https://doi.org/10.32599/apjb.13.2.202206.107

APJB
ISSN 2233-5900 (Print)
ISSN 2384-3934 (Online)

머신러닝을 활용한 코스닥 관리종목지정 예측*

채승일
한성대학교 경영학부 졸업

이동주
한성대학교 사회과학부 교수

Predicting Administrative Issue Designation in KOSDAQ Market Using Machine Learning Techniques

Seung-Il Chae[a], Dong-Joo Lee[b]

[a] Division of Management, Hansung University, South Korea
[b] Division of Social Science, Hansung University, South Korea

Received 30 May 2022, Revised 23 June 2022, Accepted 27 June 2022

Abstract

Purpose - This study aims to develop machine learning models to predict administrative issue designation in KOSDAQ Market using financial data.
Design/methodology/approach - Employing four classification techniques including logistic regression, support vector machine, random forest, and gradient boosting to a matched sample of five hundred and thirty-six firms over an eight-year period, the authors develop prediction models and explore the practicality of the models.
Findings - The resulting four binary selection models reveal overall satisfactory classification performance in terms of various measures including AUC (area under the receiver operating characteristic curve), accuracy, F1-score, and top quartile lift, while the ensemble models (random forest and gradienct boosting) outperform the others in terms of most measures.
Research implications or Originality - Although the assessment of administrative issue potential of firms is critical information to investors and financial institutions, detailed empirical investigation has lagged behind. The current research fills this gap in the literature by proposing parsimonious prediction models based on a few financial variables and validating the applicability of the models.

Keywords: Administrative Issue Prediction, KOSDAQ, Machine Learning, Prediction Model
JEL Classifications: C45, C53, G17, G33

Ⅰ. 서론

관리종목지정제도란 상장기업이 상장폐지기준에 해당하는 경우 즉시 상장폐지를 하는 대신 투자자들의 손실을 고려해 일정 기간 상장폐지를 유예하고 관리종목으로 별도 지정하는 제도이다(최실근, 2008). 관리종목지정을 통하여 투자자들의 주의를 환기시킴과 동시에 해당 기업에게는 상장폐지 사유를 해소할

* 본 연구는 한성대학교 학술연구비 지원과제임.
[a] First Author, E-mail: com010532@naver.com
[b] Corresponding Author, E-mail: djlee@hansung.ac.kr
© 2022 The Institute of Management and Economy Research. All rights reserved.

수 있는 회생 기회를 제공한다. 국내의 유가증권시장과 코스닥시장 모두 관리종목지정제도를 운영하고 있다. 이 중 기술중심의 중·소기업과 벤처기업의 자금조달 목적으로 개설된 코스닥 시장은 코스피 시장보다 상대적으로 완화된 상장 조건을 적용하므로(김소연과 박지영, 2022), 관리종목지정도 상대적으로 빈번하게 발생한다.[1)]

관리종목으로 지정된 상장 기업은 상장폐지 사유를 성공적으로 해소할 경우 관리종목지정에서 해제가 되지만 그렇지 못할 경우 상장폐지되어 주식시장에서 퇴출되는데, 관리종목 중 대략 30% 정도가 궁극적으로 상장폐지되고 있다(한화투자증권, 2022). 일단 관리종목으로 지정되었다는 사실 자체는 그 기업의 경영상태가 건실하지 못하다는 점을 의미하기 때문에 관리종목지정은 주가 하락으로 연결되기 쉬우며, 나아가 상장폐지의 위험에 노출되어 있다는 시그널로 작용하게 된다.

따라서 투자자의 의사결정에 있어서 투자대상 기업의 관리종목 편입 가능성은 중요한 고려 요소가 되며, 이는 투자자뿐 아니라 금융기관의 기업에 대한 대출 의사결정에 있어서도 마찬가지이다. 이러한 측면에서 개별 기업의 관리종목지정 가능성을 평가하고 예측하는 것은 실무적 차원 그리고 학술적 차원에서 의의를 갖는다.

관리종목지정을 비롯하여 상장폐지나 부도 등과 같은 기업부실화에 대한 설명과 예측을 위한 연구는 1960년대에 본격화되어 오늘날까지 활발하게 진행되고 있다. 단변량 분석(e.g., Beaver, 1966), 다변량 판별분석(e.g., Altman, 1968), 로지스틱 회귀분석(e.g., Ohlson, 1980) 등 전통적인 통계모형을 활용한 연구는 1990년대를 거치면서 머신러닝(machine learning) 기법을 적용한 예측 모형(Tam and Kiang, 1992)으로 진화를 거듭하였으며, 이에 따라 다양한 국가의 다양한 산업에 대해 다양한 예측 기법을 적용한 연구 결과물이 축적되어 왔다. 그런데 그동안 축적된 주요 예측 모형에 대한 검토 결과는 다른 나라 기업을 대상으로 개발된 예측모형을 그대로 사용하기보다는 국내 기업에 적합하도록 수정하거나 별도의 예측모형을 구축하는 것이 필요함을 보여준다(이인로와 김동철, 2015). 이는 예측 모형 구축에 있어서 지역적 맥락의 고려가 중요함을 시사한다.

한편, 기업부실화 예측에 대한 대부분의 국내 연구는 상장폐지 예측 모형 개발(김종훈, 박규일과 김민철, 2011; 배기수와 정설희, 2011; 배재권, 2019)과 부도 예측 모형 개발(박정민, 김경재와 한인구, 2005; 원재환과 최재곤, 2006)에 초점을 두어 왔으며, 관리종목지정 예측에 대한 연구는 매우 부족한 실정이다. 그러나 앞서 논의한 바와 같이, 관리종목지정은 상장폐지의 전단계로 기능하고 관리종목지정이 상장폐지 가능성에 대한 강력한 시그널로 작용하므로, 특정기업에 대해 미래의 상장폐지 여부를 예측하는 것 보다 관리종목 지정 여부를 예측하는 것이 현실적으로 보다 큰 의미를 가진다. 이에 본 연구는 머신러닝을 활용하여 코스닥시장에 상장된 기업의 관리종목지정 예측 모형을 개발하는 것을 목적으로 한다.

관리종목지정 예측 연구(신동인과 곽기영, 2018; 윤양현, 김태경과 김수영, 2022)를 비롯하여 기업부실화 선행연구들은 일반적으로 기업부실화가 발생한 연도의 직전 연도 회계 데이터를 이용하여 독립변수를 구성하고 예측 모형을 구축한다. 그런데 특정 연도의 회계결산자료는 당해 연도 종료 후에 집계가 가능하고 국내의 경우 차년도 3월말까지 공시하도록 정해져 있다. 그러므로 차년도의 관리종목 지정 여부를 예측하기 위하여 당해 연도의 회계 데이터를 사용하는 것은 현실적으로 불가능하며, 이런 방식으로 예측 모형을 구축하는 경우 예견편의(look-ahead bias) 문제에 노출된다. 이 문제점을 해소하기 위하여 본 연구에서는 당해 연도에 대한 회계 데이터를 이용하여 2년 후의 관리종목지정 여부를 예측하는 모형을 구축하는 접근법을 취하고자 한다.

구체적으로 본 연구에서는 2014년에서 2021년 사이에 코스닥시장 관리종목으로 지정된 기업과 그렇지 않은 비관리종목 기업 표본을 구성한 후, 관리종목지정 여부 판별 시점을 기준으로 전전 연도의 회계 데이터를 사용한 예측 모형을 개발한다. 예측에는 로지스틱 회귀분석, 서포트 벡터 머신(support vector machine), 랜덤 포레스트(random forest), 그래디언트 부스팅(gradient boosting) 등 네 가지 기법을 적용하고, 개발된 모형의 성능을 비교 평가한다. 이를 통하여 본 연구는 코스닥시장의 관리종목지정 예측에 대한 이론적 기여와 함께, 투자자와 금융기관을 위한 실무적 도움을 제공하고자 한다.

---

1) 2022년 5월 25일 기준, 유가증권시장에 상장된 820개 기업 중 12개 기업(1.5%)이 관리종목으로 지정된 반면, 코스닥시장의 경우 전체 1,554개 상장 기업 중 88개 기업(5.7%)이 관리종목으로 지정되어 있다.

본 논문의 구성은 다음과 같다. 먼저 Ⅱ절에서는 본 연구의 이론적 배경을 검토한다. Ⅲ절에서는 본 연구의 예측 모형 구축에 사용한 연구 방법을 기술하고, Ⅳ절에서는 예측 모형 구축과 분석의 결과를 제시한다. 마지막으로 Ⅴ절에서는 본 연구의 주요 결과에 대한 토의와 함께 결론을 제시한다.

Ⅱ. 이론적 배경

본 절에서는 본 연구의 이론적 배경을 검토한다. 먼저 기업 부도나 상장폐지 등을 포함하는 기업 부실화 관련 연구에 대해 살펴본 후에, 관리종목 지정 예측에 관한 선행연구를 검토한다.

1. 기업 부도 예측

기업 부실화에 대한 연구는 1930년대에 시작되어 긴 역사를 가지고 있으며, 그동안 방법론적 측면에서 지속적인 발전을 이루며 진화해 왔다. 초기 연구는 주로 비율 분석이나 단변량 분석을 사용하여 부도 기업과 그렇지 않은 기업의 차이를 비교하는 데 초점을 두었다(e.g., Smith and Winakor, 1935). 기업 부도 예측의 출발점으로 간주되는 Beaver(1966)의 연구는 이전 연구의 흐름을 따라 79개의 부도 기업과 동수의 정상 기업으로 구성된 대응 표본에 대해 30개 비율 변수를 사용하여 두 기업 집단 간의 차이를 비교하였다. 그런데 Beaver의 연구는 여기서 더 나아가 통계적 검증을 사용하여, 기업의 부도를 예측하는 데 있어서 각 비율 지표의 예측력을 검증하였다는 데서 의의를 가진다(Bellovary, Giacomino and Akers, 2007). 예를 들어, 이 연구에서는 순이익-총부채(net income to total debt) 비율이 1년 후의 부도 여부를 92% 정확도로 예측한다는 결과를 제시한다.

이후의 연구들은 단변량 분석의 한계에서 벗어나 Altman(1968)의 Z-점수모형을 시작으로 다변량 분석을 보편적인 방법론으로 채택하였다. 동일한 비율의 부도 기업과 정상 기업으로 구성된 66개 소규모 제조업체의 표본에서 추출된 데이터에 대해, Altman은 다변량 판별분석을 적용하였다. 분석에는 운전자본비율, 유보이익율, 영업이익률, 시장가-총부채비율, 그리고 매출액회전율 등 다섯 가지의 변수들이 사용되었으며, 이를 통하여 부도 기업과 정상 기업 간을 구별하는 판별식인 Z-점수 모형을 제시하였다. Altman의 연구를 기점으로 하여 부도 예측 연구는 양적 측면과 질적 측면에서 발전을 이루어왔다.

판별분석은 부도예측모형을 구축하는 대표적인 방법론으로 자리잡고 이후의 여러 연구에서 채택되었다(e.g., Altman, Haldeman and Narayanan, 1977; Blum, 1974; Deakin, 1977) 그러나 판별분석이 기반하는 가정과 제약은 회계자료를 이용한 부도 예측에 있어 한계로 작용한다. 판별분석의 여러 가정들은 대부분의 분석 상황에서 비현실적으로 엄격한 가정으로 지적되어 왔다(Cohen et al., 2003). 특히, 독립변수의 정규분포 가정은 부도예측모형에 포함되는 다양한 회계비율에 있어 충족되기 어렵지만, 이 가정이 위배될 경우 판별분석은 일치 추정량을 제공하지 못한다는 근본적인 문제를 갖게 된다. 이에 대한 대안으로써 로지스틱 회귀분석과 프로빗 회귀분석(특히 로지스틱 회귀분석)이 부도 예측을 위한 주요 통계 방법론으로서의 역할을 점차 담당하게 되었다.

로지스틱 회귀분석을 활용한 대표적인 부도예측모형으로는 Ohlson(1980)의 연구를 들 수 있다. 총자산순이익률, 순이익증가율, 부채비율, 운전자본비율, 영업활동이익 대 총부채 비율 등을 포함한 9개의 요인들로 구성된 Ohlson의 O-점수모형은 1970년부터 1976년 사이에 미국에서 부도가 발생한 105개의 기업과 2,058개의 정상 기업으로 구성된 표본에 대해 96%의 예측 정확도를 보여준다. 그러나 이 정확도에 대한 해석에는 분석에 사용된 표본 중 정상 기업이 차지하는 비율이 95%를 넘는다는 점이 고려되어야 할 것이다. Ohlson의 연구 이후 로지스틱 회귀분석은 부도 예측에 광범위하게 활용되었으며(Casey and Bartczak, 1985; Gentry, Newbold and Whitford, 1985; Theodossiou, 1991) 오늘날까지도 보편적인 부도예측모형의 하나로 활용되고 있다.

1990년대에 발생한 부도 예측 방법론의 큰 변화로는 인공신경망과 같은 머신러닝 기법이 대표적인 방법론으로 자리 잡았다는 점을 들 수 있다(Bellovary, Giacomino and Akers, 2007). 이러한 변화는 인공신경망이 예측 성능 측면에서 기존 기법들보다 우월하다는 여러 연구 결과들로부터 기인한다. 예를

들어 Tam and Kiang(1992)은 동일한 비율의 부도 은행과 건전 은행으로 구성된 미국 텍사스주의 은행 표본에 대해 여러 가지 예측 모형을 적용하고 성능을 비교하였으며, 그 결과는 인공신경망이 판별분석이나 로지스틱 회귀분석에 비해 보다 나은 예측 정확도를 가짐을 보여준다. Bell, Ribar and Verchio(1990)의 연구도 유사하게 부도 은행과 정상 은행 표본을 대상으로 로지스틱 회귀모형과 인공신경망 기법을 적용하여 인공신경망의 성능 우위를 제시하고 있다.

머신러닝 기법의 발전과 함께 부도 예측 연구에도 다양한 기법들이 활용되고 있다. 이 중 대표적인 기법들로는 인공신경망 외에 서포트 벡터 머신과 분류 나무 기반의 모형들을 들 수 있다(Barboza, Kimura and Altman, 2017; Devi and Radhika, 2018). 본 연구에서는 전통적인 부도 예측 통계 기법 중 대표적인 로지스틱 회귀분석과, 머신러닝 기법 중 서포트 벡터 머신과 분류 나무 기반의 앙상블 모형인 랜덤 포레스트와 그래디언트 부스팅 등 네 가지 기법을 적용하고자 한다.

예측 기법에 대한 검토에 이어 다음으로 부도 예측에 사용되어 온 변수들을 살펴보도록 한다. 1930년대 부도 예측 연구의 시작부터 2000년대 초반까지 이루어진 165건의 연구에 대한 Bellovary, Giacomino and Akers(2007)의 리뷰에 따르면, 그동안 전체 연구에서 사용한 서로 다른 독립변수들은 총 752개에 달한다. 그런데 이 중 674개의 변수는 단지 2건 이하의 연구에서만 활용되었을 뿐이며, 5건 이상의 연구에서 공통적으로 사용된 변수는 42개에 불과하다(가장 많이 채택된 변수는 총자산순이익률로 총 54건의 연구에 사용되었다). 이러한 현상의 잠재적 원인으로는 기존의 다수 연구가 이론적인 기반 위에서 변수를 선택하기보다는 현실적인 데이터 가용성의 제약 하에서 사용 가능한 변수들에 초점을 두었을 가능성을 들 수 있을 것이다.

모형에 포함된 변수 개수 측면에서의 모형 복잡도도 연구에 따라 차이를 보여 왔다. 적게는 단 하나의 변수만 사용한 모형에서부터(e.g., Aharony, Jones and Swary, 1980) 많게는 50여개 이상의 변수를 사용한 모형까지(e.g., Jo, Han and Lee, 1997) 다양하게 존재한다. 그런데 그 동안 이루어진 부도 예측 연구의 결과를 종합하면 변수 개수의 증가가 모형의 예측력 증가로 이어지지는 않는다는 점을 볼 수 있으며, 또한 1960년대부터 2000년대 초반까지 예측 기법의 발전에도 불구하고 모형에 포함된 변수의 개수는 평균 10여개 내외를 유지해오고 있다(Bellovary, Giacomino and Akers, 2007). 이는 많은 변수를 사용할 경우에 발생하는 비용 문제 이외에도, 판별분석이나 로지스틱 회귀분석 등 전통적 통계 모형들에서 발생할 수 있는 다중공선성 문제, 모형의 복잡성, 과적합 문제 등이 종합적으로 고려된 결과로 해석된다.

국내에서도 부도 예측에 대한 연구가 활발히 진행되어 왔으며, 여기서는 최근의 대표적인 연구를 간략히 살펴보도록 한다. 박종원과 안성만(2014)은 2003년부터 2006년까지 4년 동안의 외부감사대상기업 데이터를 통합(pooling)하고 로지스틱 회귀분석 모형을 적용하여 재무비율 변수들로 구성된 부도예측모형을 구축하였다. 모형의 분류 정확도는 82% 그리고 AUC(area under the ROC curve) 값은 85%로 상당히 양호한 수준을 보여주었다. 이인로와 김동철(2015)은 부도예측모형에 사용되는 고전적인 모형들 중 판별분석, 로지스틱 회귀분석, 헤저드 모형 등 대표적인 모형 다섯 가지의 성능을 비교하였다. 2001년부터 2013년까지 국내 유가증권시장 및 코스닥시장에 상장된 비금융기업을 대상으로 실증분석을 수행한 결과는 국내 기업에 적합하도록 수정한 모형의 예측력이 가장 우수함을 보여준다. 이러한 결과는 다른 나라 기업을 대상으로 개발된 부도예측모형을 그대로 사용하기보다는 국내 기업에 적합하도록 수정하거나 별도의 예측 모형을 구축하는 것이 필요함을 시사한다.

2. 관리종목 지정 연구

Table 1은 코스닥 시장에서의 관리종목 지정 및 상장폐지(퇴출) 요건을 보여준다. 여기에는 매출액, 영업손실 등 회계 지표와 관련된 요건, 시가총액, 주식 거래량 등 시장 지표와 관련된 요건, 지배구조와 관련된 요건, 공시 관련 요건 등을 포함하여 다양한 요건들이 존재한다. 최종부도, 자본전액잠식 등 즉시 퇴출에 해당되는 요건도 있지만, 대부분의 요건에 대해서는 그 요건에 해당되는 부실 기업이 발생할 경우 먼저 관리종목 지정 단계를 거치게 된다. 관리종목으로 지정된 유예 기간 동안 회생에 성공하여 관리종목에서 지정해제 되기도 하고 일부는 최종 상장폐지 되기도 한다(최실근, 2008).

따라서 코스닥 시장에서 특정 종목의 관리종목 지정은 그 종목의 상장폐지 위험에 대한 강력한 시그널로 작용하게 된다. 또한 일단 관리종목으로 지정된 후에는 주가 하락을 피하기 어려우므로, 투자자의 관점에서는 상장폐지 기업에 대한 예측보다는 관리종목 지정 기업에 대한 예측이 보다 의미 있는 정보를 제공하게 된다. 그러나 관리종목과 관련된 연구, 특히 관리종목 지정 예측에 대한 연구는 매우 부족한 실정이다.

이 중 신동인과 곽기영(2018)의 연구는 관리종목 지정 예측에 대한 거의 최초의 연구로 볼 수 있다. 이 연구에서는 2009년부터 2017년 사이에 관리종목으로 지정된 코스닥 상장 247개 기업 및 동일한 수의 비관리종목 기업으로 구성된 표본을 활용하였다. 비관리종목 표본은 연도별로 관리종목 표본의 기업 수와 동일하게 무작위로 추출한 쌍대표본으로 구성하였다. 로지스틱 회귀분석과 의사결정나무를 적용하여 구축한 예측 모형은 유사한 예측 정확도를 보여주었다. 먼저 로지스틱 회귀모형에는 ROE(세전계속사업이익), 자기자본현금흐름률, 총자산회전율 등 세 개 변수가 최종 포함되었으며, 검증 데이터셋에 대한 분류 정확도는 86% 수준이었다. 의사결정나무 모형에는 현금흐름/총자산과 ROA(당기순이익)를 통한 분류 규칙에 최종적으로 사용되었으며 분류 정확도는 87%로 나타나 로지스틱 회귀분석보다 약간 높은 수준이었다.

Table 1. Requirements for Administrative Issue Designation and Delisting in KOSDAQ Market(as of April 19, 2019)

구분	관리종목	상장폐지
매출액	최근년 30억원 미만 (지주회사는 연결기준)	2년 연속
법인세비용 차감전계속 사업손실	자기자본 50% 초과(& 10억원이상)의 법인세비용차감전계속사업손실이 최근 3년간 2회 이상(& 최근 연도 계속사업손실)	관리종목 지정 후 자기자본 50% 초과(& 10억원 이상)의 법인세비용차감전계속사업손실 발생
장기영업손실	최근 4사업연도 영업손실 (지주회사는 연결기준)	관리종목 지정 후 최근 사업연도 영업손실 발생할 경우 상장폐지 실질심사
자본잠식/자기자본	(A) 사업연도(반기)말 자본잠식률 50%이상 (B) 사업연도(반기)말 자기자본 10억원미만 (C) 반기보고서 제출기한 경과후 10일내 반기검토(감사)보고서 미제출 or 검토(감사)의견 부적정·의견거절·범위제한한정	최근년말 완전자본잠식 A or C 후 사업연도(반기)말 자본잠식률 50% 이상 B or C 후 사업연도(반기)말 자기자본 10억원 미만 A or B or C 후 반기말 반기보고서 기한 경과 후 10일내 미제출 or 감사의견 부적정·의견거절·범위제한한정
감사의견	반기보고서 부적정, 의견거절, 감사범위 제한으로 인한 한정	감사보고서 부적정·의견거절·범위제한한정
시가총액	보통주시가총액 40억원 미만 30일간 지속	관리종목 지정 후 90일간 "연속 10일 & 누적 30일간 40억원 이상"의 조건을 미충족
거래량	분기 월평균거래량이 유동주식수의 1%에 미달	2분기 연속
지분분산	소액주주 200인 미만 or 소액주주 지분 20% 미만	2년 연속
불성실공시	-	1년간 불성실공시 벌점 15점 이상일 경우 상장폐지 실질심사
공시서류	분기, 반기, 사업보고서 법정제출기한 내 미제출	2년간 3회 분기, 반기, 사업보고서 법정제출기한 내 미제출 사업보고서 제출기한 후 10일내 미제출 분기, 반기, 사업보고서 미제출상태유지 후 다음 회차에 미제출
사외이사 등	사외이사/감사위원회 요건 미충족	2년 연속
회생절차/파산신청	회생절차 개시 신청 파산신청	개시신청기각, 결정취소, 회생계획 불인가 등의 경우 상장폐지 실질심사
기타 (즉시퇴출)	기타 상장폐지 사유 발생	최종부도 또는 은행거래정지 해산사유(피흡수합병, 파산선고) 정관 등에 주식양도제한 두는 경우 유가증권시장 상장의 경우 우회상장시 우회상장관련 규정 위반시 (심사종료전 기업결합완료 및 보호예수 위반 등)

Source: 한국거래소 (https://listing.krx.co.kr/contents/LST/04/04020500/LST04020500.jsp)

윤양현, 김태경, 김수영(2022)의 연구에서는 관리종목 지정 예측을 위해 다양한 머신러닝 기법의 성능을 비교하였다. 신동인과 곽기영(2018)의 연구에서와 유사한 방식으로 구성된 1대 1 쌍대표본에 대해 로지스틱 회귀분석, 의사결정나무, 서포트 벡터 머신, 랜덤 포레스트, LightGBM을 적용하여 예측 모형을 구축하였다. 그 결과 전반적으로 앙상블 모형이 단일 학습 모형에 비해 성능이 높으며, 변수 측면에서는 ROE(당기순이익)와 자본금회전율이 관리종목 지정 예측에 있어 상대적으로 중요한 변수임을 보여주었다.

그런데 관리종목 지정 예측 연구들은 관리종목 지정 연도 기준으로 직전 연도의 회계 데이터를 이용하여 독립변수를 구성하고 예측 모형을 구축한다. 이는 앞서 검토한 부도예측 선행연구들 중 대부분의 연구에서도 마찬가지로, 이들 연구에서 결과로 제시된 예측 성능은 종속변수(부도 혹은 정상) 측정 연도 기준으로 1년 전 연도의 데이터로 독립변수를 구성한 결과이다(Bellovary, Giacomino and Akers, 2007). 이런 방식으로 개발한 모형들은 사건(관리종목 지정 혹은 부도) 발생 직전 연도에서의 예측(즉, 내년도에 사건이 발생할지 여부를 예측)에 활용할 수 있는 것으로 간주하지만 회계 데이터를 활용한 대부분의 모형에서 이는 사실과 달라진다.

특정 연도 예를 들어 2021년의 회계결산자료는 당해 연도가 종료되어야 집계가 가능하며, 이 내용을 포함하는 사업보고서는 차년도 즉 2022년 3월말까지 공시하도록 정해져 있다. 그러므로 2021년 결산자료에 포함된 변수들을 사용하여 2022년 관리종목 지정 여부를 예측하는 것은 적어도 2022년 4월이 되어야 가능하게 된다. 따라서 t+1 연도의 관리종목 지정 여부를 t 연도에 예측하기 위하여 t 연도의 연간 회계 데이터를 이용하는 것은 현실적으로 불가능하며, 이런 방식으로 예측 모형을 구축하는 경우 예견편의 문제에 노출된다.[2] 공시 시점 이전에, 예를 들어 2월에 발생한 관리종목 지정 건에 대해서는 3월 말에 얻을 수 있는 회계변수를 이용하여 과거인 2월의 관리종목 지정 여부를 예측하는 꼴이 되어 예측모형이 본질적으로 무의미해지게 된다.

이와 같은 문제를 고려하여, 본 연구에서는 직전 연도가 아니라 전전 연도에 대한 회계결산 자료를 사용하여 예측모형을 구축하고자 한다. 다시 말해 2021년도의 회계 데이터를 이용하여 2023년도의 관리종목지정 여부를 예측하는 모형을 개발하고자 한다.

Ⅲ. 연구 방법

1. 표본 기업의 선정과 데이터 수집

본 연구에서 관리종목지정 예측 모델 개발을 위한 기간 범위는 2012년에서 2021년까지 10년간으로 설정하였다. 당해연도에 대한 회계결산 자료를 사용하여 2년 후의 관리종목지정 여부를 예측하고자 하므로, 먼저 관리종목으로 지정된 기업의 표본을 구성하기 위하여 기업공시채널(KIND)에서 2014년부터 2021년까지 관리종목으로 지정된 코스닥 기업의 목록을 수집하였다. 이 중 2개 이상의 연도에서 관리종목 지정이 유지된 기업의 경우 최초 지정된 연도만 남기고 목록에서 제거하였다. 그리고 나서 전체 기업을 관리종목으로 최초 지정된 연도와 표준산업대분류의 업종을 기준으로 분류하였다.

다음으로 관리종목으로 지정되지 않은 '비관리종목' 기업 표본을 관리종목 표본에 포함된 기업과 동수로 구성하였다. 이를 위해 2014년부터 2021년까지 각 연도별로, 해당 연도의 전체 비관리종목 기업 중에서 관리종목 기업 수와 동일한 수의 기업을 동일한 업종 구성비율로 무작위 선정하였다. 이 과정에서 기존 연구(신동인과 곽기영, 2018)와 마찬가지로, 기업 인수 목적의 서류상 회사 형태로 운영되는 스팩(SPAC) 기업은 제외하고, 최종적으로 관리종목 268개 기업과 비관리종목 268개 기업으로 구성된 표본을 구성하였다.

2) 예견편의(혹은 선견자편의)는 실제 분석(혹은 예측) 시점에 활용가능하지 않은 데이터나 정보를 활용가능한 것으로 간주하고서 모형을 구축하거나 연구를 수행하는 경우에 발생하는 전형적인 데이터 문제이다(Banz and Breen, 1986). 결산자료가 공시되지 않았음에도 해당 자료를 사용하고 종속변수와 매칭시키는 경우가 예가 된다(이은광과 형남원, 1997).

Table 2에서 연도별 기업 구성이 정리되어 있다. 2019년을 제외하고는 매년 30개 내외의 기업이 관리종목으로 지정되었음을 보여준다. Table 3에서는 표준산업대분류에 따른 업종별 구성이 정리되어 있다. 제조업이 전체의 63.8%를 차지하며, 출판, 영상, 방송통신 및 정보서비스업과 도매 및 소매업이 그다음으로 높은 구성비를 보여준다. 표본 기업의 구성에 이어, 다음으로 예측 모형에서 독립변수로 사용할 회계 데이터는 DataGuide5.0에서 수집하였으며, 여기에는 총 187개의 변수가 포함되어 있다.

Table 2. Sample Composition by Year

연도	2014	2015	2016	2017	2018	2019	2020	2021	합계
관리종목	25	29	26	26	34	58	38	32	268
비관리종목	25	29	26	26	34	58	38	32	268
소계	50	58	52	52	68	116	76	64	536
구성비	9.3%	10.8%	9.7%	9.7%	12.7%	21.6%	14.2%	11.9%	100%

Table 3. Sample Composition by Industry

표준산업대분류	관리종목	비관리종목	소계	구성비
건설업	8	8	16	3.0%
교육 서비스업	1	1	2	0.4%
금융 및 보험업	4	4	8	1.5%
도매 및 소매업	30	30	60	11.2%
사업시설관리 및 사업지원 서비스업	2	2	4	0.7%
숙박 및 음식점업	1	1	2	0.4%
예술, 스포츠 및 여가관련 서비스업	2	2	4	0.7%
전문, 과학 및 기술 서비스업	10	10	20	3.7%
제조업	172	172	342	63.8%
출판, 영상, 방송통신 및 정보서비스업	36	36	72	13.4%
하수·폐기물 처리, 원료재생 및 환경복원업	2	2	4	0.7%
합계	268	268	536	100%

2. 독립변수 후보군의 선정

수집한 총 187개 변수 중 분석에 사용할 독립변수 후보군을 선정하기 위하여 몇 가지 기준을 적용하였다. 먼저, 동일한 정보를 담고 있는 변수(e.g., 자기자본비율과 타인자본비율)는 하나만 남기고 제거하였다. 또한 연 단위 지표가 아닌 변수(e.g., 전분기 대비 매출액증가율)도 제거하여 총 72개 변수가 남게 되었다. 이 중 결측치 혹은 이상치가 많은 변수 36개를 제거한 후에 기존 연구(신동인과 곽기영, 2018; 이인로와 김동철, 2015) 및 한국은행 기업경영분석에 포함된 경영분석지표(한국은행, 2021)를 참고하여 총 13개의 변수로 구성된 독립변수 후보군을 선정하였으며 그 결과가 Table 4에 정리되어 있다.

이 중 11개의 재무비율 변수들은 안전성(stability), 성장성(growth), 수익성(profitability), 활동성(activity) 등 4가지 유형으로 구분된다. 첫째, 안전성 비율은 기업이 안정적으로 경영관리 활동을 할 수 있는지를 보여주는 지표로 부채 지급능력과 재무구조를 보여준다. 둘째, 성장성 비율은 지속적으로 사업 규모가 확대될 수 있는지를 보여주는 지표로 재무제표상 매출액과 자산 등이 얼마나 늘어났는지를 보여주는 비율이다. 셋째, 수익성 비율은 사업을 통해 이익을 창출하는 능력을 나타내는 지표들이다.

넷째, 활동성 비율은 기업이 경영활동을 얼마나 활발하게 움직이고 있는지를 보여주며 주로 회전율을 사용하여 나타낸다. Table 4의 맨 아래에는 기업 규모를 나타내는 두 가지 변수인 log(총자산)과 log(매출액)이 포함되어 있다. 비율 변수는 규모의 영향을 제거해 버리지만, 기업의 규모는 일반적으로 정보생산의 우위로 연결되므로 규모가 큰 기업은 경영환경변화에 상대적으로 쉽게 적응할 수 있다는 장점을 가지므로 (김종훈, 박규일과 김민철, 2011) 본 연구의 맥락에서 관리종목지정 가능성을 줄이는 요인이 될 수 있다.

Table 4. Selected Variables

구분		변수	계산식
안정성 (Stability)	ST1	자기자본비율	(총자본 / 총자산) × 100
	ST2	부채비율	(총부채 / 총자본) × 100
성장성 (Growth)	G1	매출액증가율	(매출액 / 전년도 매출액) × 100
	G2	총자산증가율	(총자산 / 전년도 총자산) × 100
수익성 (Profitability)	P1	영업이익률	(영업이익 / 매출액) × 100
	P2	EBITDA마진율	(EBITDA / 매출액) × 100
	P3	당기순이익률	(당기순이익 / 매출액) × 100
	P4	총자산순이익률	(당기순이익 / 총자산) × 100
활동성 (Activity)	A1	총자산회전율	매출액(연율화) / 총자산(평균)
	A2	총자본회전율	매출액(연율화) / 총자본(평균)
	A3	자본금회전율	매출액(연율화) / 자본금(평균)
규모 (Size)	SI1	log(총자산)	log(총자산)
	SI2	log(매출액)	log(총자본)

Table 5에는 각 변수에 대한 기초통계치가 제시되어 있으며, Table 6 변수 간의 상관관계를 보여준다.

Table 5. Descriptive Statistics of Selected Variables

구분	변수		평균	표준편차	최솟값	중앙값	최댓값
안정성	ST1	자기자본비율	54.41	21.88	3.14	52.63	97.31
	ST2	부채비율	136.06	193.06	2.77	90.00	3,083.20
성장성	G1	매출액증가율	8.04	58.00	-88.52	0.98	564.56
	G2	총자산증가율	23.30	70.12	-57.87	6.41	675.30
수익성	P1	영업이익률	-16.92	84.40	-1,276.33	0.56	57.74
	P2	EBITDA마진율	-9.79	81.27	-1,258.65	4.63	262.82
	P3	당기순이익률	-28.92	105.70	-1,423.23	-2.61	529.43
	P4	총자산순이익률	-6.77	18.80	-113.26	-2.15	113.52
활동성	A1	총자산회전율	0.76	0.57	0.00	0.63	4.34
	A2	총자본회전율	1.80	2.21	0.00	1.19	25.32
	A3	자본금회전율	21.93	173.34	0.00	7.46	3,983.85
규모	SI1	log(총자산)	25.27	0.96	22.82	25.16	29.54
	SI2	log(매출액)	24.66	1.23	20.76	24.64	28.49

Table 6. Correlation Matrix

변수	ST2	G1	G2	P1	P2	P3	P4	A1	A2	A3	SI1	SI2
자기자본비율 (ST1)	-0.67	-0.03	0.04	-0.06	-0.06	-0.01	0.22	-0.14	-0.44	-0.07	-0.10	-0.17
부채비율 (ST2)		-0.01	-0.04	0.03	0.03	-0.01	-0.19	0.08	0.35	0.04	0.10	0.12
매출액증가율 (G1)			0.41	0.07	0.05	0.12	0.15	0.21	0.19	-0.01	0.02	0.09
총자산증가율 (G2)				-0.21	-0.22	-0.14	0.18	-0.05	-0.04	-0.01	0.02	-0.15
영업이익률 (P1)					0.99	0.87	0.38	0.24	0.14	0.02	0.13	0.40
EBITDA마진율 (P2)						0.82	0.36	0.20	0.12	0.01	0.12	0.36
당기순이익률 (P3)							0.58	0.26	0.16	0.02	0.18	0.44
총자산순이익률 (P4)								0.23	0.09	0.04	0.36	0.40
총자산회전율 (A1)									0.69	0.16	-0.02	0.55
총자본회전율 (A2)										0.17	0.02	0.40
자본금회전율 (A3)											0.01	0.09
log(총자산) (SI1)												0.73

IV. 관리종목지정 예측 모형 개발

관리종목지정 예측 모형 개발에 앞서, 독립변수 후보군에 포함된 13개의 변수에 대해 관리종목지정 여부와의 기본적인 관련성을 평가하기 위하여 관리종목 표본과 비관리종목 표본 간의 평균 차이 검증을 실시하였다. Table 7에 나타난 결과에서 볼 수 있듯이, 5% 유의수준에서 총자본회전율(A2)과 자본금회전율(A3) 등 두 변수를 제외한 나머지 변수들은 모두 유의한 차이를 보여준다. 예측 모형 구축에는 두 집단 간 평균 차이가 유의한 이들 11개 변수를 사용하였다.

예측 모형 구축을 위해 전체 표본을 학습 데이터 셋과 검증 데이터 셋으로 구분하였다. 표본에 포함된 전체 536개의 기업 중 무작위로 75%의 기업(402개 기업)을 추출하여 학습 데이터 셋을 구성하고, 나머지 25%의 기업(134개 기업)으로 검증 데이터 셋을 구성하였다.

Table 7. t-test Results for Input Variable Selection

구분		변수명	관리종목평균	비관리종목평균	t-값	p-값
안정성	ST1	자기자본비율	49.653	59.176	5.158	< 0.001
	ST2	부채비율	166.923	105.190	-3.746	< 0.001
성장성	G1	매출액증가율	1.675	14.410	2.555	0.011
	G2	총자산증가율	17.237	29.353	2.006	0.045
수익성	P1	영업이익률	-26.157	-7.673	2.548	0.011
	P2	EBITDA마진율	-17.710	-1.869	2.265	0.024
	P3	당기순이익률	-46.730	-11.112	3.954	< 0.001
	P4	총자산순이익률	-13.812	0.275	9.349	< 0.001
활동성	A1	총자산회전율	0.679	0.842	3.323	< 0.001
	A2	총자본회전율	1.751	1.850	0.518	0.605
	A3	자본금회전율	24.371	19.487	-0.326	0.745
규모	SI1	log(총자산)	25.100	25.434	4.078	< 0.001
	SI2	log(매출액)	24.422	24.902	4.587	< 0.001

1. 예측 모형 구축

관리종목 지정 예측 모형 구축에는 로지스틱 회귀분석과 서포트 벡터 머신, 랜덤 포레스트, 그래디언트 부스팅 등 네 가지 기법을 적용하였다. 본 절에서는 네 가지 예측 모형 구축 방법을 간략히 설명하고,

모형의 예측 성능에 대해서는 다음 절에서 검토한다.

먼저 로지스틱 회귀분석에는 11개의 독립변수에 대해 단계적 선택법(stepwise selection)을 사용하여 예측 모형을 구축하였다. 그 결과 자기자본비율(ST1), 매출액증가율(G1), 총자산증가율(G2), 당기순이익률(P3), 총자산순이익률(P4), 총자산회전율(A1), log(총자산)(SI1) 등 7개 변수를 포함하는 예측 모형이 도출되었다.

다음으로, 서포트 벡터 머신 기법을 활용하여 관리종목 기업과 비관리종목 기업을 분리하는 초평면을 찾아 두 번째 예측 모형을 구축하였다. 선형의 초평면을 찾는 선형 모형을 적용하고, 분류기의 일부 오분류를 허용하는 소프트 마진(soft margin)을 적용하였다.[3] 분석에는 파이썬(python) 환경에서 사이킷런(scikit-learn) 라이브러리를 활용하였다.

세 번째 예측 기법으로는 의사결정나무(decision tree) 기반의 배깅 앙상블(bagging ensemble) 모형인 랜덤 포레스트를, 그리고 네 번째 기법으로는 의사결정나무 기반의 부스팅(boosting) 앙상블 모형인 그래디언트 부스팅을 활용하였다.

각 예측 모형은 하이퍼 파라미터(hyper parameter)의 값에 따라 성능이 달라질 수 있으므로 하이퍼 파라미터 최적화를 수행하였다. 파라미터 최적화에는 사이킷런의 GridSearchCV를 이용하였다. 각 예측 모형에 대해 격자 탐색을 통해 도출된 최적의 파라미터는 Table 8에 각주로 제시되어 있다.

2. 예측 모형 성능 평가

Table 8에는 개발된 네 가지 예측 모형의 성능이 학습 데이터 셋과 검증 데이터 셋에 대해 정리되어 있다. 예측 모형의 성능을 평가하기 위한 지표로는 AUC, 정확도, F1-score, quartile lift 등 네 가지 지표를 적용하였다. 예측 모형의 성능을 종합적으로 확인할 수 있는 지표인 AUC (area under the ROC curve)는 ROC (receiver operating characteristic) 곡선 하단의 면적을 의미하며, 그 값이 1에 가까울수록 좋은 성능을 나타낸다.[4] 검증 데이터셋 기준 AUC는 로지스틱 회귀모형의 경우 0.774, 서포트 벡터 머신 모형은 0.777, 랜덤 포레스트 모형은 0.801, 그래디언트 부스팅 모형은 0.785로 나타났다. 네 모형 모두 전반적으로 만족스러운 수준의 AUC 값을 보여주며, 랜덤 포레스트가 다른 모형에 비해 우수한 성능을 보여준다. Fig.1에는 검증 데이터 셋에 대한 네 가지 모형의 ROC 곡선이 제시되어 있다. 전반적으로 랜덤 포레스트 모형의 곡선이 다른 세 모형의 곡선보다 좌표평면의 좌측 상단에 가까워 높은 AUC 값을 가짐을 알 수 있다. 학습 데이터 셋에 대해서는 랜덤 포레스트와 그래디언트 부스팅의 AUC 값이 나머지 두 모형보다 높으며 그 차이는 현저하다.

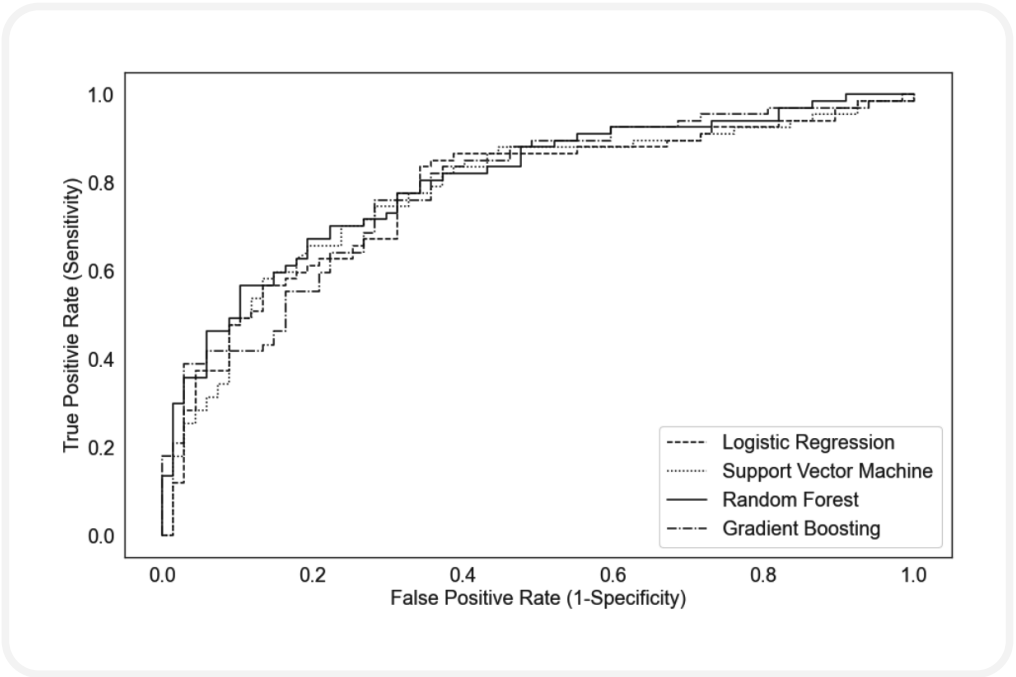
Table 8. Performance Overview

성능척도	데이터셋	로지스틱 회귀분석[5]	서포트 벡터 머신[6]	랜덤 포레스트[7]	그래디언트 부스팅[8]
AUC	학습	0.769	0.775	0.825	0.834
	검증	0.774	0.777	0.801	0.785
정확도	학습	0.721	0.721	0.729	0.759
	검증	0.694	0.724	0.716	0.739
F1-score	학습	0.701	0.699	0.733	0.756
	검증	0.682	0.694	0.732	0.745
Top Quartile Lift	학습	1.684	1.684	1.762	1.742
	검증	1.648	1.648	1.764	1.648

3) 서포트 벡터 머신 모형은 분류기의 오분류를 허용하지 않는 하드 마진(hard margin) 모형과 오분류를 허용하는 소프트 마진(soft margin) 모형으로 구분할 수 있다(Géron, 2019). 본 연구에서는 사이킷런의 마진 조절 파라미터를 기본값인 1로 설정하였다.

4) ROC 곡선은 예측 결과로 구성한 혼동행렬(confusion matrix)의 위양성률 즉 1-특이도(specificity)를 가로축으로, 진양성률 즉 민감도(sensitivity)를 세로축으로 하여 시각화한 그래프이다. ROC 곡선은 분류 기준값(cutoff)의 변화에 따른 이 두 지표의 동적인 변화를 하나의 좌표평면에 보여준다(Kim et al., 2020)

Fig. 1. ROC Curve



두 번째 성능 지표로는 예측 모형이 실제 관리종목 기업을 관리종목으로, 비관리종목 기업을 비관리종목으로 정확하게 예측한 비율인 정확도(accuracy)를 사용하였다. 올바르게 예측된 관측치의 수를 전체 관측치 수로 나누어 구한 정확도는 Table 8에서 볼 수 있듯이, 검증 데이터 셋에 대해 로지스틱 회귀모형의 경우 0.694, 서포트 벡터 머신 모형은 0.724, 랜덤 포레스트 모형은 0.716, 마지막으로 그래디언트 부스팅은 0.739를 기록했다. 따라서 로지스틱 회귀모형을 제외한 세 모형 모두 72% 이상의 예측 정확도를 가지며, 그래디언트 부스팅이 가장 높은 정확도를 보여준다. 학습 데이터 셋에 대한 예측 정확도 역시 그래디언트 부스팅 모형이 76%로 가장 우수한 결과를 나타낸다.

F1-score는 정밀도(precision)와 재현율(recall)의 조화평균으로, 정밀도와 재현율이 고르게 높을수록 큰 값을 갖게 된다. 검증 데이터셋 기준 F1-score는 로지스틱 회귀모형이 0.682, 서포트 벡터 모형은 0.694, 랜덤 포레스트 모형은 0.732, 그래디언트 부스팅 모형은 0.745로 그래디언트 부스팅이 가장 우수한 성능을 보여준다. 학습 데이터 셋에 대해서도 마찬가지로 그래디언트 부스팅 모형이 가장 높은 수치(0.756)를 보여준다.

마지막으로 quartile lift는 모형이 표본의 각 관측치에 대해 예측한 관리종목지정 확률이 가장 높은 관측치부터 순서대로 정렬한 후, 이를 사등분한 각 사분위에 대해 예측 모형의 성능을 평가하는 지표이다. 일반적으로 확률이 높은 최상위 25%의 관측치들로 구성된 첫 번째 사분위 즉 top quartile에 대해 평가하는데, top quartile로 분류된 기업 중 관리종목 기업이 차지하는 구성비율을 전체 표본에서 관리종목

5) 로지스틱 회귀분석 모형에 대해 격자 탐색을 통해 도출된 하이퍼 파라미터 값은 다음과 같다. C: 1, max_iter: 10000, penalty: l1, solver: liblinear.

6) 서포트 벡터 머신 모형에 대한 하이퍼 파라미터 값은 다음과 같다. C: 0.2, gamma: scale, kernel: linear, probability: true.

7) 랜덤 포레스트 모형에 대한 하이퍼 파라미터 값은 다음과 같다. Criterion: entropy, max_depth: 2, min_samples_split: 40, n_estimators: 100.

8) 그래디언트 부스팅 모형에 대한 하이퍼 파라미터 값은 다음과 같다. learning_rate: 0.15, max_depth: 2, min_samples_split: 60, n_estimators: 50, subsample: 0.2.

기업이 차지하는 구성비로 나눈 값이 top quartile lift이다. 예를 들어, top quartile lift가 1.5라면 이는 관리종목 지정 가능성이 가장 높을 것으로 모형에 의해 예측된 상위 25%의 기업 중 실제 관리종목으로 지정된 비율이 전체 표본에서의 관리종목 기업 구성비율보다 1.5배 높음을 의미한다.[9)]

Table 8에서 검증 데이터 셋에 대한 top quartile lift는 랜덤 포레스트가 1.764로 다른 세 모형에 비해 현저히 높은 수치를 보여준다. 학습 데이터 셋에 대해서도 유사하게 랜덤 포레스트가 가장 좋은 성능을 보이고 있다. Table 9에는 네 가지 모형 각각에 대해 사분위별 quartile lift 값이 정리되어 있다.

이상의 성능 평가 결과를 종합하면, 네 가지 예측 모형 모두 전반적으로 만족스러운 성능을 보여준다. 검증 데이터 셋에 대한 모형 간의 성능 차이는 아주 두드러지지는 않지만, 대부분의 성능 지표에서 랜덤 포레스트 모형와 그래디언트 부스팅 모형이 로지스틱 회귀모형이나 서포트 벡터 머신 모형보다 우수한 성능을 보인다. 학습 데이터 셋에 대해서는 모든 성능 지표에 대해 랜덤 포레스트와 그래디언트 부스팅이 나머지 두 모형보다 현저히 높은 성능을 보여주고 있다. 이러한 결과는 코스닥 관리종목지정 예측에 있어서 분류 나무 기반의 앙상블 모형이 단일 학습 모형에 비해 성능이 우수하다는 것을 시사한다.

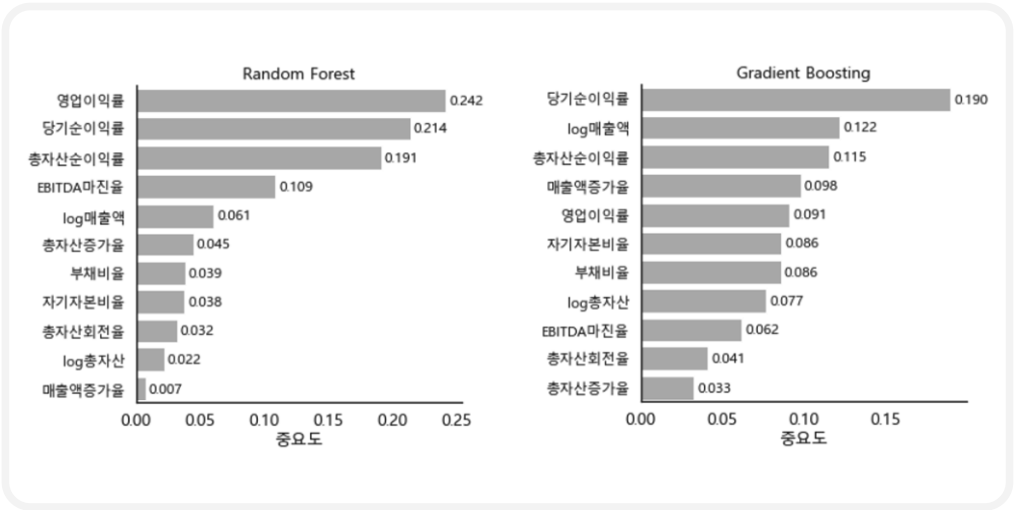
Table 9. Quartile Lift

데이터셋	사분위(Quartile)	로지스틱 회귀분석	서포트 벡터 머신	랜덤 포레스트	그래디언트 부스팅
학습	사분위 1 (Q1)	1.684	1.684	1.762	1.742
	사분위 2 (Q2)	1.160	1.220	1.140	1.280
	사분위 3 (Q3)	0.640	0.640	0.840	0.740
	사분위 4 (Q4)	0.514	0.456	0.258	0.238
검증	사분위 1 (Q1)	1.648	1.648	1.764	1.648
	사분위 2 (Q2)	1.090	1.212	1.090	1.212
	사분위 3 (Q3)	0.788	0.666	0.788	0.728
	사분위 4 (Q4)	0.470	0.470	0.352	0.412

Fig. 2는 예측 성능이 우수한 두 앙상블 모형에서 각 변수의 중요도를 정렬하여 보여준다. 그림 왼쪽의 랜덤 포레스트 모형의 경우 영업이익률이 가장 높은 중요도(0.242)를 가지는 것으로 나타났다. 당기순이익률(0.212)과 총자산순이익률(0.191)이 그 다음으로 높은 중요도를 보이며, EBITDA마진율(0.109)과 log 매출액(0.061)이 그 다음을 잇는다. 그래디언트 부스팅 모형의 경우 당기순이익률이 가장 높은 중요도(0.190)를 가지며, 그 다음으로 중요도가 높은 변수들은 log매출액(0.122), 총자산순이익률(0.115), 매출액증가율(0.098), 영업이익률(0.091) 등이다. 두 모형 간에 변수 중요도 순위가 완전히 일치하지는 않지만, 전반적으로 일관된 결과를 보여준다. 변수 중요도와 관련된 시사점에 대해서는 V절에서 논의하도록 한다.

9) 전체 표본에서의 관리종목 기업 구성비율(본 연구의 경우 0.5)은 예측 모형을 사용하지 않고 무작위로 선정한 25% 기업에 대해 기대할 수 있는 관리종목 기업 구성비율과 동일하다. 따라서 top quartile lift는 예측 모형을 사용하지 않는 경우에 비해 예측 모형의 사용이 가져오는 예측 성능의 향상도(lift)를 보여준다.

Fig. 2. Variable Importance for Random Forest and Gradient Boosting



Ⅴ. 시사점 및 결론

코스닥시장에 상장된 기업의 관리종목지정은 단기적으로는 주가 하락의 위험으로, 장기적으로는 상장폐지의 위험으로 연결될 수 있다. 따라서 코스닥 상장 기업의 관리종목지정 가능성에 대한 예측은 투자자들의 투자 의사결정과 금융기관의 대출 의사결정에서 중요한 고려 요인이 된다.

본 연구는 머신러닝을 활용하여 코스닥시장에 상장된 기업의 관리종목지정 예측 모형을 개발하기 위한 목적으로 수행되었다. 2014년에서 2021년 사이에 코스닥시장 관리종목으로 지정된 기업과 비관리종목 기업으로 구성된 1대 1 쌍대표본 데이터에 로지스틱 회귀분석, 서포트 벡터 머신, 랜덤 포레스트, 그래디언트 부스팅 등 네 가지 기법을 적용하여 예측 모형을 개발하였다.

검증 데이터 셋에 대한 예측 모형의 주요 성능은 다음과 같다. 먼저 AUC 값은 네 모형 모두 0.770 이상이며, 랜덤 포레스트가 0.801로 가장 높은 성능을 보였다. 예측 정확도는 그래디언트 부스팅 모형이 73.9%로 가장 높은 반면 로지스틱 회귀모형이 69.4%로 다른 모형에 비해 낮은 수치를 보여주었다. F1-score는 그래디언트 부스팅이 0.745로 가장 높아 정밀도와 재현율 측면에서 균형 있는 성능을 보여주었다. Top quartile lift는 랜덤 포레스트가 1.764로 가장 높은 것으로 나타났다. 네 개 모형 모두 전반적으로 만족스러운 수준의 예측 성능을 보여주었으며, 이 중 랜덤 포레스트 모형과 그래디언트 부스팅 모형이 보다 우수한 것으로 평가되었다.

이상의 주요 결과와 함께, 본 연구는 학문적 측면에서 다음과 같은 의의를 가진다. 첫째, 상장폐지 예측이나 부도 예측을 목적으로 하는 기업부실화 예측 모형 연구는 다양하게 진행되었으나, 관리종목지정 예측에 관한 학술적 연구는 아직 매우 부족한 실정이다. 더욱이, 기업부실화 예측은 연구의 대상이 되는 맥락에 따라 별도의 예측 모형 구축이 요구되므로(이인로와 김동철, 2015) 상장폐지나 부도 예측 모형을 관리종목지정 예측에 그대로 사용하는 것은 한계가 있다는 점을 고려할 때, 본 연구는 관리종목지정 예측 연구의 초기 연구로서 의의를 가질 것이다.

둘째, 관리종목지정 예측 연구를 비롯한 기존의 회계모형 기반 기업부실화 예측 선행연구에서의 예측 모형 구축에 나타나는 예견편의 문제를 해소하기 위하여, 본 연구에서는 당해 연도에 대한 회계 데이터를 이용하여 2년 후의 관리종목지정 여부를 예측하는 모형을 구축하는 접근법을 취하였다. 이는 이론적 측면과 실무적 측면 모두에서 예측 모형의 타당성을 높이는 접근법이 될 것이다. 또한, 예견편의 문제가 예측

모형 구축에서 중요한 고려사항이 되어야 한다는 점을 명확히 했다는 점에서도 본 연구는 의미를 가질 것으로 판단된다.

실무적 측면에서 본 연구는 몇 가지 시사점을 가진다. 첫째, 상장 기업이 주식 시장에서 퇴출되기 전에 일반적으로 관리종목지정 단계를 거친다는 점을 감안하면, 정상 기업을 대상으로 미래의 상장폐지 여부를 예측하는 모형보다는 관리종목지정 여부를 예측하는 모형이 현실의 투자자나 금융기관, 그리고 금융당국에게 보다 큰 가치를 갖는다. 본 연구에서 개발한 예측 모형은 이들 이해당사자의 의사결정에 필요한 가치 있는 정보를 제공하는 도구로 활용될 수 있을 것으로 기대된다.

둘째, 본 연구에서 제시한 모형은 상장 기업의 공시를 통해 제공되는 수많은 재무지표 중 11개 이내의 적은 수의 지표를 사용하는 단순한 모형이면서도 다양한 성능 지표 측면에서 충분한 예측 성능을 보여준다. 많은 변수를 사용하는 복잡한 예측 모형을 현실에 적용하는 데 수반되는 높은 비용을 고려한다면, 본 연구에서 개발한 모형은 실무적 활용성 측면에서도 의의를 가질 것이다. 특히 랜덤 포레스트나 그래디언트 부스팅과 같은 앙상블 모형이 서포트 벡터 머신이나 로지스틱 회귀분석과 같은 단일 학습 모형보다 관리종목지정 예측 성능이 더 우수함을 보여주는 성능 평가 결과는 실무적 측면에서 모형 선택에 대한 의미 있는 가이드를 제공할 수 있다.

셋째, 랜덤 포레스트 모형과 그래디언트 부스팅 모형에서 도출된 주요 변수들도 실무적 관점에서 살펴볼 필요가 있다. 먼저, 중요도 상위 다섯 개 변수를 기준으로 살펴보면, 두 모형 모두 영업이익률, 당기순이익률, 총자산순이익률, log매출액 등 네 개 변수를 공통적으로 포함하고 있다. 또한 두 모형 모두 공통적으로 수익성 관련 변수들이 중요함을 보여준다. 랜덤 포레스트 모형의 경우 중요도 상위 다섯 개 변수 중 log매출액을 제외한 네 개 변수가 모두 수익성 변수들이다. 그래디언트 부스팅 모형의 경우도 중요도 상위 다섯 개 변수 중 당기순이익률, 총자산순이익률, 영업이익률 등 세 개 변수가 수익성 지표이다. 이러한 결과로부터, 관리종목지정 예측에서 수익성과 관련된 재무비율 변수들이 가장 높은 중요도를 가지고 핵심적인 역할을 담당한다는 점을 주목하는 것이 필요할 것이다. 마지막으로, 기업 규모와 관련된 변수 중에서는 log매출액이 두 모형에서 공통적으로 중요한 변수로 포함되어 있다. 이는 기업의 총자산보다는 매출액이 관리종목지정 예측에 있어서 보다 결정적인 역할을 담당한다는 것을 의미한다.

마지막으로 본 연구의 한계점과 함께 추후 연구 방향을 제시한다. 첫째, 본 연구에서 2014년에서 2021년 사이에 코스닥시장에서 관리종목으로 지정된 기업들의 업종을 고려한 표본을 구성하여 모형 개발에 사용하였지만, 업종과 관계없이 전체 기업을 통합한 예측 모형을 개발하였다. 그러나 업종에 따라 경영환경이나 사업 특성의 차이가 존재하므로, 경영의 건전성 혹은 반대로 관리종목지정 가능성과 연관된 요인들의 차이로 인하여 예측 모형의 특성도 상이할 가능성이 크다. 따라서 업종별 특성을 고려한 예측 모형을 개발함으로써, 보다 체계적이고 정교한 예측이 가능할 것으로 예상되며 이와 관련한 후속 연구는 의미를 가질 것이다.

둘째, 예측 모형을 적용하는 시점의 거시경제적 환경에 대한 고려도 중요할 것이다. 예를 들어, 코로나 이전과 비교하여 코로나 시기 동안은 범세계적인 거리두기, 공급망 불안정, 경기 침체, 노동력 공급 제한 등 여러 측면에서의 변화가 발생하였다. 이러한 변화는 기업 경영 활동, 기업 성과 등에 영향을 미치고 나아가 부실화 가능성에도 영향을 미칠 수 있을 것이다. 따라서 경영환경 측면에서 발생한 주요 외생적 충격을 예측 모형에 반영하거나, 환경적 특성에 따라 예측 모형을 구분하여 개발하는 것은 모형의 타당성과 현실성을 한층 강화하는 데 도움이 될 것이다.

셋째, 본 연구에서는 예견편의 문제를 해소하기 위하여 독립변수와 예측 대상 시점 사이에 2년의 시차를 두었다. 이는 연간 사업보고서에 포함된 연 단위 재무지표를 예측 모형에 사용함에 따라 불가피하게 발생하는 시차이다. 그런데 예측의 기간 범위가 수개월인 단기 예측인 경우, 이러한 시차는 본 연구에서 개발된 모형의 적용을 어렵게 만든다. 관리종목지정에 대한 단기 예측을 위해 가능한 접근법으로는 연 단위 재무지표가 아니라 분기(혹은 반기)보고서에 포함된 분기(혹은 반기) 재무지표를 활용하여 예측 모형을 구축하는 방법을 채택할 수 있을 것이다.

References

김소연, 박지영 (2022), "코스닥 상장 전·후 기업의 혁신성과", *아태비즈니스연구*, 13(1), 75-88.

김종훈, 박규일, 김민철 (2011), "상장폐지기업의 재무적 특성과 예측에 관한 연구: 코스닥 기업을 중심으로", *회계와 정책연구*, 16(2), 125-142.

박정민, 김경재, 한인구 (2005), "Support Vector Machine을 이용한 기업부도예측", *경영정보학연구*, 15(2), 51-63.

박종원, 안성만 (2014), "재무비율을 이용한 부도예측에 대한 연구: 한국의 외부감사대상기업을 대상으로", *경영학연구*, 43(3), 639-669.

배기수, 정설희 (2011), "상장폐지기업 예측모형에 관한 실증연구", *세무회계연구*, 30, 125-140.

배재권 (2019), "인공지능기법을 이용한 코스닥 기업의 상장폐지위험요인에 관한 연구", *로고스경영연구*, 17(2), 127-142.

원재환, 최재곤 (2006), "기업의 부채구조를 고려한 옵션형 기업부도예측모형과 신용리스크", *재무관리연구*, 23(2), 209-237.

이은광, 형남원 (2019), "재무회계 특성과 편의들을 고려한 한국증시 내 요인간 관계 연구", *한국재무학회 학술대회 자료집*, 1589-1618.

이인로, 김동철 (2015), "회계정보와 시장정보를 이용한 부도예측모형의 평가 연구", *재무연구*, 28(4), 625-665.

최실근 (2008), "관리종목 지정 상장법인의 지배구조 실태와 개선", *CV 리뷰*, 41, 23-30.

한국은행 (2021), *2020년 기업경영분석*.

한화투자증권 (2022), *위험한 종목을 피하는 3가지 방법*. Available from https://www.hanwhawm.com/main/bbs/content_print.cmd?nn_id=10&vc_bid=investorguide (accessed May 23, 2022)

Aharony, J., G. Jones and I. Swary (1980), "An Analysis of Risk and Return Characteristics of Corporate Bankruptcy Using Capital Market Data", *Journal of Finance*, 35(4), 1001- 1016.

Altman, E., R. Haldeman and P. Narayanan (1977), "Zeta Analysis: A New Model to Identify Bankruptcy Risk of Corporations", *Journal of Banking and Finance*, 1(1), 29-51.

Banz, R. W. and W. J. Breen (1986), "Sample-dependent Results Using Accounting and Market Data: Some Evidence", *Journal of Finance*, 41(4), 779-793.

Barboza, F., H. Kimura and E. Altman (2017), "Machine Learning Models and Bankruptcy Prediction", *Expert Systems with Applications*, 83, 405-417.

Bellovary, J. L., D. E. Giacomino and M. D. Akers (2007), "A Review of Bankruptcy Prediction Studies: 1930 to Present", *Journal of Financial Education*, 33(4), 1-42.

Blum, M. (1974), "Failing Company Discriminant Analysis", *Journal of Accounting Research*, 12(1), 1-25.

Casey, C. and N. Bartczak (1985), "Using Operating Cash Flow Data to Predict Financial Distress: Some Extensions", *Journal of Accounting Research*, 23(1), 384-401.

Cohen, J., P. Cohen, S. G. West and L. S. Aiken (2003), *Applied Multiple Regression/Correlation Analysis for the Behavioural Sciences* (3rd ed.), New York: Taylor & Francis Group.

Deakin, E. (1972), "A Discriminant Analysis of Predictors of Business Failure", *Journal of Accounting Research*, 10(1), 167-179.

Devi, S. S. and Y. Radhika (2018), "A Survey on Machine Learning and Statistical Techniques in Bankruptcy Prediction", *International Journal of Machine Learning and Computing*, 8(2), 133-139.

Gentry, J., P. Newbold and D. Whitford (1985), "Classifying Bankrupt Firms with Funds Flow Components", *Journal of Accounting Research,* 23(1), 146-160.

Géron, A. (2019), *Hands-on Machine Learning with Scikit-Learn, Keras, and TensorFlow: Concepts, Tools, and Techniques to Build Intelligent Systems* (2nd ed.), O'Reilly Media, Inc.

Jo, H., I. Han and H. Lee (1997), "Bankruptcy Prediction Using Case-based Reasoning, Neural Networks, and Discriminant Analysis", *Expert Systems with Applications*, 13(2), 97-108.

Kim, D., K. Park, D.-J., Lee and Y. Ahn, (2020), "Predicting Mobile Trading System Discontinuance: The Role of Attention", *Electronic Commerce Research and Applications*, 44, 101008.

Ohlson, J. (1980), "Financial Ratios and the Probabilistic Prediction of Bankruptcy", *Journal of Accounting Research*, 18(1), 109-131.

Smith, R. and A. Winakor (1935), *Changes in Financial Structure of Unsuccessful Industrial Corporations,* Bureau of Business Research, Bulletin No. 51. Urbana: University of Illinois Press.

Theodossiou, P (1991), "Alternative Models for Assessing the Financial Condition of Business in Greece", *Journal of Business Finance & Accounting*, 18(5), 697-720.