벤처창업연구 제17권 제1호 (통권79호)
Asia-Pacific Journal of Business Venturing and Entrepreneurship Vol.17 No.1 pp.229-249

ISSN 1975-7557
e-ISSN 2671-9509

머신러닝 기반 KOSDAQ 시장의 관리종목 지정 예측 연구: 재무적 데이터를 중심으로*

윤양현 (광운대학교 경영학부 4학년)**
김태경 (광운대학교 경영학부 부교수)***
김수영 (광운대학교 수학과 4학년)****

국 문 요 약

본 연구는 다양한 머신러닝 기법을 통해 코스닥(KOSDAQ) 시장 내 관리종목 지정을 예측할 수 있는 모델에 대해 연구하였다. 증권시장 내 기업이 관리종목으로 지정이 되면 시장에서는 이를 부정적인 정보로 인식하여 해당 기업과 투자자에게 손실을 가져오게 된다. 본 연구를 통해 기업의 재무적 데이터를 바탕으로 조기에 관리종목 지정을 예측하고, 투자자들의 포트폴리오 리스크 관리에 도움을 주기 위한 머신러닝 접근이 타당한지 살펴본다. 본 연구를 위해 활용한 독립변수는 수익성, 안정성, 활동성, 성장성을 나타내는 21개의 재무비율을 활용하였으며, K-IFRS가 적용된 2011년부터 2020년까지 관리종목과 비관리종목의 기업의 재무 데이터를 표본으로 추출하였다. 로지스틱 회귀분석, 의사결정나무, 서포트 벡터 머신, 랜덤 포레스트, LightGBM을 활용하여 관리종목 지정 예측 연구를 수행하였다. 연구결과는 분류 정확도가 82.73%인 LightGBM이 가장 우수한 예측 모형이었으며 분류 정확도가 가장 낮은 예측 모형은 정확도가 71.94%인 의사결정나무였다. 의사결정나무 기반 학습 모형의 변수 중요도의 상위 3개 변수를 확인한 결과 각 모형에서 공통적으로 나온 재무변수는 ROE(당기순이익), 자본금회전율(Capital stock turnover ratio)로 해당 재무변수가 관리종목 지정에 있어 상대적으로 중요한 변수임을 확인하였다. 대체적으로 앙상블을 이용한 학습 모형이 단일 학습 모형보다 예측 성능이 높은 것을 확인하였다. 기존 선행연구가 K-IFRS에 대한 고려를 하지 않았고, 다소 제한된 머신러닝에 의존하였다. 따라서 본 연구의 필요성과 함께 현실적 요구를 충족시키는 결과를 제시하였음을 알 수 있으며, 시장참여자들에게 있어 관리종목 지정에 대한 사전 예측을 확인할 수 있도록 기여했다고 볼 수 있다.

핵심어: KOSDAQ, 관리종목, 머신러닝, 앙상블

Ⅰ. 서론

일반적으로 벤처기업은 기업의 지속적인 발전을 위한 단계로 상장기업이 되기를 희망한다. 그러나 실제로 많은 기업들이 오랜 기간 존속되지 못하고 부실의 단계로 이어지는 경우가 있다(문종건·황보윤, 2014). 초기 기업의 부실화 가능성은 당연한 것처럼 생각될 수 있지만, 벤처 기업의 경우 미래 발전 가능성에 대한 기대가 높고 잠재적 투자자로부터 긍정적으로 평가될 가능성이 크다는 점에서 부실화 가능성에 관한 판단은 다소 유보적이다. 최근 코로나19로 인해 바이오테크 벤처가 기술특례상장 절차를 거쳤다는 점은 벤처의 관리종목 지정은 자칫 닥칠 수 있는 잠재적 위협이 될 수도 있다는 사실을 반증한다.

한국 벤처 환경에서는 부실화 가능성이 충분한 벤처기업의 경우 부실기업으로 곧바로 지정되기보다 관리종목으로 지정함으로써 재도약의 기회를 부여하는 제도가 마련되어 있다는 점이 주목된다. 관리종목은 주권상장법인이 상장 후 영업실적의 지속적 악화, 상장회사가 갖추어야 할 최소한의 유동성 부족, 기업지배구조 미구축 등의 사유로 상장폐지기준에 해당된 종목 가운데 특별히 지정된 종목들을 의미한다. KRX(Korea Exchange, 한국거래소)에서는 상장폐지기준에 해당하는 상장기업들을 관리종목으로 지정하고, 기업이 경영을 지속 혹은 존속하기 힘든 기업들을 관리종목 지정제도를 통하여 투자자들에게 공시한다. 이러한 정보의 전달로 투자자는 상장폐지 위험이 있는 기업들을 조기에 파악함으로써 투자판단에 있어 주의를 환기하고, 당해 법인에게는 일정기간 경과기간을 부여하여 관리종목 지정사유를 해소함으로써 기업의 정상화를 촉진하는 효과를 지닌다(김민철, 2004).

* 이 논문은 2020년도 광운대학교 교내학술연구비 지원에 의해 연구되었음(2020-0323). 또한, 이 논문은 2019년 대한민국 교육부와 한국연구재단의 지원을 받아 수행된 연구임(NRF-2019S1A3A2098438)
** 주저자, 광운대학교 경영학부 4학년, eb3434@naver.com
*** 교신저자, 광운대학교 경영학부 부교수, kimtk@kw.ac.kr
**** 공동저자, 광운대학교 수학과 4학년, sooyoung6262@naver.com
· 투고일: 2022-01-03 · 1차 수정일: 2022-02-13 · 2차 수정일: 2022-02-21 · 게재확정일: 2022-02-23

예를 들어, 세포치료제 연구사업 및 줄기세포 보관사업 등을 주 사업으로 영위하는 '차바이오텍' 기업은 KOSDAQ(Korea Securities Dealers Automated Quotation)에 상장되어 있다. 차바이오텍은 관리종목 지정 전에 시가총액 2조 원의 가치를 지니고 있었으나, 2018년 3월 22일 관리종목 지정이 되고 난 후 주가가 하락하여 3일 만에 관리종목 지정 전 주가 대비 절반이 된 사례가 존재하였다. 이 사례의 경우 시가총액 1조 원의 가치가 단기간 내에 사라지게 되어 기업과 투자자의 입장에서 많은 손실을 가져오게 되었다(신동인·곽기영, 2018). 이처럼 관리종목 지정의 가능성을 판단하는 일은 어떤 유형의 벤처기업을 관리종목 지정의 대상으로 판단할 것인지, 혹은 관리종목 지정이 예견될 경우 사전 평가를 통해 관리종목 지정을 유예할 것인지 등을 결정함에 있어서 중요하다. 따라서 관리종목 지정을 예견할 수 있는 다양한 지표 데이터가 상대적으로 저렴한 비용으로 신속히 획득될 수 있어야 한다.

관리종목으로 지정이 되면 해당 기업은 신용거래 대상에서 제외되고 증권시장에서 일정기간 동안 매매거래가 정지된다. 또한 대용증권으로도 활용할 수 없어 해당 기업의 레버리지가 줄어들게 되는 불이익을 받게 된다. 그 외에도 매매방법에 있어 30분 간격으로 거래가 체결되는 별도의 제한을 받게 되어 관리종목 지정 자체가 기업의 입장에서 상당히 불리한 조치임을 알 수 있다(박종성, 2012). 과거의 선행연구에서 관리종목 공시 이전부터 주가가 하락함을 알 수 있었으며(김태혁·엄철준, 1997), 관리종목 지정 자체가 시장에서 부정적인 정보로 인식되어 지정 후에 비기대이익에 대한 누적초과수익률이 감소하는 것으로 나타났다. 이는 관리종목으로 지정 받았다는 사실 자체가 부정적인 정보로 작용하여 상장폐지를 유발하는 결과를 초래할 수 있다(손성규·오명전, 2008). 앞선 사례와 관리종목 지정으로 인한 기업과 투자자에게 부정적인 효과를 내비친 선행연구를 통하여 본 연구자에게 다음과 같은 연구질문을 제기하게끔 하는 계기가 되었다. 머신러닝 기법을 기반으로 하여 해당 기업의 재무비율을 통해 사전에 관리종목 지정 여부를 예측할 수 있을까? 만약 예측이 가능하다면, 머신러닝 알고리즘 중 관리종목 지정을 가장 잘 예측할 수 있는 머신러닝 알고리즘은 무엇일까?

국내, 해외 모두 기업의 부도예측에 관한 연구가 활발히 진행 중에 있다. 하지만 상장기업 내 부실화 가능성이 높은 기업에 대한 경고를 하는 관리종목 지정 예측에 관한 연구는 단적으로 찾아보기 어렵다. 즉, 한국에서는 관리종목 지정 제도가 기업의 부실화를 예방하기 위한 중요한 제도적 역할을 함에도 불구하고 상대적으로 적은 관심을 받고 있다. 이는 관리종목 지정 예측 연구가 희소성이 있으며, 학문적으로 의의가 있음을 시사한다. 또한 기존 선행연구(신동인·곽기영, 2018)와 달리 2011년 이후부터 국제회계기준이 적용된 K-IFRS(Korea-International Financial Reporting Standards, 한국채택국제회계기준)을 통해 데이터를 수집하였으므로 더 정확한 관리종목 지정 예측 모형이 될 수 있음을 시사한다.

본 연구는 이러한 논의를 바탕으로 코스닥 기업의 재무 데이터를 활용하여 관리종목 지정 예측 모델을 설계하고자 한다. 다양한 머신러닝 기법을 활용하여 관리종목과 비관리종목을 분류하고 이를 예측할 수 있는 분류 모형을 제안하고, 이를 통하여 기업과 투자자들은 기업 부실화에 가까운 종목을 조기에 인지할 수 있도록 기여하고자 한다. 부실기업에 관한 연구와는 달리 관리종목의 예측 문제는 벤처의 초기 창업 환경을 살펴볼 때 중요한 시사점을 제공한다. 이는 벤처기업의 초기 성과를 바탕으로 관리기업의 지정 가능성을 조기에 확인함으로써 기업 가치의 평가조정 문제나 정상화 조치를 마련하고 시행하는 전략적 방향을 연구할 때 도움을 줄 것으로 기대된다.

II. 이론적 배경

2.1. 관리종목 지정

<표 1> 코스닥 시장 관리종목 지정 요건

구분	관리종목 지정 요건(2019년 4월 17일 개정 기준)
매출액	최근년 30억원 미만(지주회사는 연결기준) -기술성장기업, 이익미실현기업은 각각 상장후 5년간 미적용
1)법인세비용차감전계속사업손실	자기자본50%이상(&10억원이상)의 법인세비용차감전계속사업손실이 최근3년간 2회 이상 (&최근연도계속사업손실) -기술성장기업 상장후 3년간 미적용, 이익미실현 기업 상장후 5년 미적용
장기영업손실	4)최근 4사업연도 영업손실(지주회사는 연결기준) -기술성장기업(기술성장기업부)은 미적용
2)자본잠식/자기자본	사업연도(반기)말 자본잠식률1) 50%이상 사업연도(반기)말 자기자본 10억원미만 반기보고서 제출기한 경화후 10일내 반기검토(감사)보고서 미제출 or 검토(감사)의견 부적정·의견거절·범위제한한정 자본잠식율=(자본금-자기자본) / 자본금*100
3)감사의견	반기보고서 부적정, 의견거절, 감사범위 제한으로 인한 한정
시가총액	보통주시가총액 40억원미만 30일간 지속
거래량	분기 월평균거래량이 유동주식수의 1%에 미달 -월간거래량 1만주, 소액주주 300인이상이 20%이상 지분 보유 등은 적용배제
지분분산	5)소액주주200인미만or소액주주지분20%미만 -300인이상의 소액주주가 유동주식수의 10%이상으로서 100만주이상을 소유하는 경우는 적용배제
불성실공시	-
공시서류	분기, 반기, 사업보고서 법정제출기한 내 미제출
사외이사등	사외이사/감사위언회 요건 미충족
회생절차/파산신청	회생절차 개시 신청 파산신청
기타(즉시퇴출)	기타 상장폐지 사유 발생

주 1)연결재무제표 작성대상법인의 경우, 연결재무제표상 법인세비용차감전계속사업손실 및 자기자본 기준
주 2)연결재무제표 작성대상법인의 경우, 연결재무제표를 기준으로 하되 자기자본에서 비지배지분을 제외
주 3)연결재무제표 작성대상법인의 경우, 연결재무제표에 대한 감사의견을 포함
주 4)기술성이 있고 연구개발 투자가 많은 연구개발기업에 대해 장기영업손실로 인한 관리종목 지정을 한시적으로 면제
주 5)자진상장폐지를 위한 공개매수시 분산기준 미달로 인한 관리종목 지정 등 유예

상장폐지란 상장된 유가증권이 한국증권거래소(KRX)가 정한 일정한 기준에 해당되었을 때 증권시장에서 매매될 수 있는 자격이 박탈당하는 조치를 말한다. 당해 법인이 상장폐지기준에 해당되었을 경우에 증권거래소는 상장된 기업을 폐지할 수 있게 된다. 그러나 해당기업이 일시적으로 상장폐지기준에 해당하는 경우가 존재할 가능성이 있기에 즉각 상장폐지조치를 취하면 기업이 회생할 기회도 없이 증권시장에서 퇴출 당하는 결과를 초래할 수 있다. 이는 해당기업의 투자자들에게 투자액 회수의 기회도 상실할 수 있다는 문제점을 가지게 된다. KRX의 관리종목 지정은 바로 이러한 문제점들을 해결하기 위해 내려지는 조치이다(표영인·김일, 2002). KRX는 기업이 상장폐지기준에 해당하면 해당 기업의 의사와는 관계없이 강제적으로 상장폐지 시킬 수 있다. 그러나 상장폐지기준에 해당하는 기업들을 바로 상장폐지 하기 보다는 기업에게 회생기회를 부여함과 동시에 투자자에게 투자손실을 예고함으로써 주의를 환기시키기 위해 관리종목을 지정한다. KRX는 기업이 공시서류 미제출, 자본잠식, 매출부진, 영업손실, 시장 내 유동성 부족 등과 같이 부실화 가능성이 높은 일정 요건에 해당하는 경우 관리종목으로 지정한다. 한국의 유가증권시장(KOSPI Market: Korea Composite Stock Price Index)과 코스닥시장(KOSDAQ)은 각각의 관리종목 제도가 존재한다. 두 시장 간 관리종목 지정 요건은 어느 정도 차이가 있으나 큰 틀은 동일하다. 최근에는 코스닥 시장의 관리종목 지정 요건의 경우 기술성장기업이나 이익미실현기업과 같이 추후 성장이 기대되는 기업이지만 현재 재무구조가 좋지 않은 기업들에 대해서는 관리종목 지정에 있어 예외를 두고 있다. KRX에서 정한 코스닥 시장의 관리종목 지정 요건은 다음 <표 1>과 같다.

시장규제 관련 제도는 각 국가의 법령과 관련되어 있어 각 국마다 주식시장 규제 제도는 독특하게 발전하였고, 세계적인 공통사가 되지 않았다(박창래·서영미, 2015). 따라서 관리종목 지정과 같은 시장규제 제도는 해외의 각 국가마다 차이가 존재한다. 증권시장에서 가장 대표적인 미국의 나스닥(NASDAQ: National Association of Securities Dealers Automated Quotations)은 별도의 상장폐지 요건이 존재하지 않고 상장된 기업이 지속적으로 상장을 유지할 수 있도록 상장유지 조건을 운영하고 있다. 나스닥의 상장유지조건은 순 유형자산, 시가총액, 유동주식수, 유동주식의 시가총액, 주주수 등의 기준을 포함하고 있다. 미국 뉴욕증권시장(NYSE: the New York Stock Exchange)은 거래소 자체가 상장폐지권한을 지니고 있지 않지만 뉴욕거래소에서 상장폐지요구서를 제출하면 미국증권거래위원회(SEC: Securities and Exchange Commission)에서 상장폐지 심의를 거쳐 상장폐지를 결정한다. 해당 기업은 이러한 상장폐지 심의에 대해 이의를 제기할 수 있으며, 상장폐지의 결정은 법원에서 결정한다. NYSE의 상장폐지 조건은 일정 주가 30일 거래일 지속, 시가총액의 일정 조건 미충족, 상장유지 비용 미납부, 재무상황 악화, 거래소와의 계약 위반 등의 기준을 포함하고 있다. 일본의 자스닥(JASDAQ)은 주주수, 상장시가총액, 채무초과, 치부율 기타 구분(사업활동의 정지, 부당한 합병 등)으로 상장폐지기준에 해당할 경우 상장폐지를 진행하고 있다(김승열, 2010). 이 외에도 중국의 상해증권교역소(SSE: Shanghai Stock Exchange), 유럽의 런던증권거래소(LSE: London Stock Exchange), 홍콩증권거래소(Hong Kong Stock Exchange) 등도 각 나라의 증권소 별 상장폐지기준에 대한 차이는 있지만, 큰 틀은 NYSE와 비슷하게 상장폐지기준을 포함하고 있다(Martinez & Serve, 2017).

2.2. 관리종목 관련 선행연구

관리종목에 관한 선행 연구는 다음과 같다. 김태혁·엄철준(1997)은 1984년 이후로 관리대상종목으로 지정된 68개 기업을 대상으로 하여 각 종목의 수익률과 위험 특성에 대하여 연구하였다. 연구 결과 관리종목 지정 이후 대부분 주가는 음(-, negative)의 수익률을 기록하며, 특히 관리종목 공시일에 가까울수록 주가의 변동성이 크며, 공시 이전 관리대상종목을 사전적으로 발견할 수 있음을 확인하였다.

김민철(2004)은 상장기업들이 관리종목으로 지정되는 경우 관리종목지정사유와 주가수익률의 관계를 조사하였다. 2000~2003년 관리종목을 대상으로 공시 30일 이전부터 공시 40일 이후까지 주가수익률을 관찰하였다. 연구 당시 관리종목 지정사유를 구분하여 회사정리절차관련, 자본잠식, 부도발생, 화의관련, 감사의견한정, 감사의견부적정 6개 요인을 재무적 요인으로 구분하였고, 영업활동정지, 주식분포미달, 거래량요건미달, 공시의무위반, 회계처리위반 5개 요인을 비재무적 요인으로 구분하여 조사하였다. 연구 결과 재무적 요인으로 관리종목지정 10일전부터 유의적인 차이가 발생하였음을 확인하였으며, 비재무적요인에 의한 관리종목의 경우에는 시장평균수익률과 큰 차이가 없음을 확인하였다. 따라서 도산의 정의로 관리종목을 선정하는 경우 비재무적 요인에 의하여 지정된 기업은 배제되어야 함을 제시하였다.

김일(2005)은 관리종목으로 지정된 기업과 동종산업 내 비관리종목을 비교하여 재무적으로 다른 특성이 있는지 조사하였다. 연구 결과 관리종목 지정 기업은 수익성, 안정성, 활동성에서 동종업계 비관리종목보다 관리종목 지정 전 5년동안 유의한 차이를 보여주지만 성장성을 통하여 두 비교집단 간 유의한 차이를 확인할 수 없었다. 또한 유동성은 관리종목 지정 5년 전부터 2년 전까지는 유의한 차이가 없고, 1년 전부터 큰 차이를 보여줌을 확인하였다. 따라서 수익성, 안정성, 활동성, 유동성을 나타내는 과거 재무적 특성 요인이 관리종목 지정에 영향을 미친다는 결과를 제시하였다.

손성규·오명전(2008)은 관리종목 지정 전후를 비교 분석하여 관리종목 지정이 회계정보에 어떠한 역할을 수행하는지 검토하였다. 연구 결과 관리종목 지정 자체가 시장에서는 부정적인 정보로 인식되어 회계정보효과를 감소시킬 가능성이 있으며, 반대로 관리종목에서 해제된 후에 비기대이익에 대한 누

적초과수익률이 증가하는 방향성이 나타났음을 확인하였다. 이를 통하여 관리종목 지정 자체가 시장에서 부정적인 정보로 받아들여지고 있음을 제시하고 있다.

박종성(2012)은 2011년부터 K-IFRS가 도입됨에 따라 영업이익 산출기준이 없어지고 항목분류 조정을 통한 이익조정이 가능해졌음을 확인하였다. 또한 관리종목 지정 위기에 처한 기업들이 관리종목 지정을 회피하기 위해 영업이익을 조정하였는지 분석하였다. 분석 결과에 따르면 기타 손익의 분류를 달리하면 영업손실을 영업이익으로 전환하여 관리종목 지정을 회피할 수 있었으며, 이는 관리종목 지정 위기에 처한 기업들이 영업이익에 도움이 되는 항목은 영업손익에 포함시키되 영업손실을 초래하는 항목은 영업손익에서 제외시키고 있음을 확인하였다. 이러한 경우 예외적으로 관리종목에 해당하는 기업임에도 불구하고 회계장부 조작을 통하여 관리종목 지정에서 벗어날 수 있음을 보여주었으며, 한국에 맞는 영업이익에 대한 산출기준을 별도로 정할 것을 제시하였다.

손성규·염지인(2013)은 2006년부터 2010년까지의 298개의 표본기업을 선정하여 상장폐지 위기에 있는 종목의 이익조정가능성에 대해 실증분석을 하였다. 연구 결과 관리종목 지정과 같은 상장폐지 위기를 벗어나기 위해 회계정보를 조정하거나 조작할 유인이 있음을 확인하였으며 향후 코스닥시장의 상장규정이나 상장폐지실질심사제도에 대한 개선방안을 제시하였다.

신동인·곽기영(2018)은 2008년부터 2018년까지의 전체 코스닥 기업 재무 데이터를 수집하고 관리종목과 무작위로 선택한 비관리종목을 쌍대표본으로 구성한 후 관리종목 지정 예측에 대한 연구를 하였다. 로지스틱 회귀분석과 의사결정나무 분석을 이용하여 관리종목 지정을 예측하였으며, 전체 평균 예측 정확도는 검증용 데이터셋에 대해 86%, 87%의 예측 정확도를 보여주었다.

김수정·문보영(2018)은 2008년부터 2013년까지 상장된 기업 중 관리종목으로 지정된 적이 있는 355개 기업을 표본으로 하여 관리종목에 대한 지정감사제도 효용성을 검증하는 실증분석 연구를 하였다. 연구 결과 감사인 지정 연도의 이익조정 감소 여부는 재량적 발생액 측정변수에 따라 혼재되는 양상을 보였으며, 관리종목 지정 사유가 지속적인 손실과 매출액 미달인 경우에는 이익조정 억제 측면에서 감사인 지정 정책이 효과적이었지만, 그 외의 사유인 경우 이익조정 억제 측면에서 효과가 나타나지 않음을 확인하였다.

신찬휴(2021)는 모바일 게임개발사 D사의 쇠퇴 및 관리종목 지정 회피에 관한 사례 분석을 하고 자본시장 규정의 보완 방향성을 제시하였다. 해당 기업은 장기영업손실 기준 재무제표가 별도재무제표라는 규정을 악용하여 손실을 자회사로 몰아주고, 모회사는 별도기준 영업이익을 흑자로 전환시켜 관리종목 지정 회피를 하였다. 특히 연구개발비가 많이 발생하는 바이오나 게임 산업에 속한 기업들이 적자 사업부를 자회사로 분리 혹은 인력을 이동시켜 관리종목을 회피하고 있음을 확인하였다. 또한, 정보해석능력이 부족하고 자회사 설립에 대한 공시를 정확하게 해석할 수 없는 소액주주들을 위해 거래소가 만든 규정의 빈틈을 보완해야 한다고 제시하였다.

2.3. 기업부도예측 선행연구

관리종목 지정 제도는 한국에만 존재하는 특수한 제도이기에, 관리종목 예측 모델에 관한 연구는 부족하지만 이와 유사한 기업의 부실 예측 모형에 관한 선행연구는 많으며, 이에 관한 연구들은 현재에도 활발히 연구 중이다. 부실 예측 모형의 초기 연구는 Beaver(1966)의 단일변량 분석으로 재무 변수를 통하여 기업 부도를 예측하였다. 하지만 현실에서는 복합적인 요인으로 인하여 기업이 파산하기 때문에 단일변수로는 기업의 부실화 예측을 설명하기에는 명백히 한계점이 존재하였다. 그 후 Altman(1968)이 단일변량 분석의 한계점을 보완하기 위하여 다변량 판별분석을 적용한 Z-Score 모형을 개발하였다. Z-Score 모형은 기업의 도산 1년 전 재무 데이터를 활용하여 기업의 도산 여부를 설명하였다. Beaver(1966)의 단일변량 분석에 비해 많은 변수들을 고려하여 기업의 도산 여부를 설명하였지만, 한계점으로는 파산기업과 정상기업 모두 독립변수의 분산과 공분산 행렬이 동일하다고 가정한 후 연구를 수행하였다. 실제로는 독립변수의 분산과 공분산이 일정하지 않으며, 데이터의 분포도 정규성을 띄지 않는 경우가 존재하였기에 이 연구에서도 한계점이 존재하였다. Ohlson(1980)은 1970~1976년에 파산한 105개 기업과 건전 기업인 2058개를 표본으로 하여 도산 1년 전의 재무 데이터를 이용하여 로지스틱 회귀분석을 통한 파산예측모형을 제시하였다. Alaka et al.(2016)은 부실예측모형에 관련된 70개의 논문을 분석하였다. 각 논문들의 검토를 통하여 기존 연구 논문들의 한계를 제시하며, 이에 맞는 새로운 패러다임과 연구 설계 프레임 워크를 제시하였다.

2.4. 머신러닝 기반 기업부도예측 선행연구

<표 2> 선행연구

선행연구	관련 논문	
관리종목 (Administrative Issue)	김태혁·엄철준(1997), 김민철(2004), 김일(2005), 손성규·오명전(2008), 박종성(2012), 손성규·염지인(2013), 신동인·곽기영(2018), 김수정·문보영(2018), 신찬휴(2021)	
기업부도예측 (Corporate Bankruptcy Prediction)	판별분석 (Discriminant Analysis)	Beaver(1966), Altman(1968)
	이항반응모형 (Binary Response Models)	Ohlson(1980), Zmijewski(1984), Campbell et al.(2008)
	위험모형 (Hazard Model)	Shumway(2001), Chava & Jarrow(2004), Duffie et al.(2007)
	머신러닝 (Machine Learning)	Barboza et al.(2017), Devi & Radhika(2018), 김형준 외 (2019), 엄하늘 외 (2020)

*김형준 외(2019) 논문을 참고하여 작성하였음.

최근에 머신러닝/딥러닝 기술이 발전하면서 머신러닝을 기반으로 한 부도예측 연구가 지속적으로 수행되고 있다. Barboza et al.(2017)은 1985년부터 2013년까지의 미국회사들의 부도 1년 전 데이터를 통하여 전통적 분석 모형인 판별분석, 로지스틱 회귀분석, 초기 인공신경망(artificial neural networks)으로 도출된 결과와 머신러닝 기반 알고리즘인 서포트 벡터 머신(support vector machine), 배깅(bagging), 부스팅(boosting), 랜덤 포레스트를 사용하여 도출된 결과를 비교하였다. 연구 결과, 머신러닝을 활용한 분석 방법이 전통적인 분석 방법보다 더 예측력이 좋다는 결과를 제시하였다. Devi & Radhika(2018)는 전통적인 분석방법인 선형 판별분석, 다변량 판별분석, 로지스틱 회귀분석과 머신러닝 분석방법인 인공신경망, 서포트 벡터 머신을 통하여 부도예측에 관한 연구를 진행하였으며, 연구 결과 머신러닝 기반 예측 결과가 전통적인 분석방법 보다 더 좋은 예측력을 보여주었고, 특히 유전자 알고리즘(genetic algorithm)과 입자 군집 최적화(particle swarm optimization)를 적용하면 예측 결과에 대해 더 좋은 성능을 이끌어 낸다는 결과를 제시하였다. 김형준 외(2019)는 기업부도예측 방법론에 대해 통계적 모형과 기계학습 알고리즘의 대표적인 방법론을 소개하며, 기업부도에 대한 정밀한 예측을 하기 위한 금융공학 방법론 확대의 토대를 마련하였다. 엄하늘 외(2020)는 K-IFRS가 적용된 2012년부터 2018년까지의 기업데이터를 이용하여 부도위험을 머튼(Merton) 모형을 토대로 각 기업의 시가총액과 주가 변동성을 이용하여 부도위험을 산정하고, 부도사건 희소성에 따른 데이터 불균형 문제를 해소하였다. 또한, 다양한 머신러닝을 서브 모델로 하는 스태킹 앙상블 기법을 활용하여 개별 모델이 갖는 편향을 감소시키고 단일 머신러닝 기반 모델의 한계를 극복 및 개선하였다.

기존 선행연구(신동인·곽기영, 2018)에서는 2008년부터 2018년의 데이터를 모아서 연구를 수행하였다. 하지만 박종성(2012)의 연구에 따르면, 회계장부가 K-IFRS로 변화되고 난 후 관리종목에 해당하는 기업임에도 불구하고 회계장부 조작을 통하여 관리종목 지정에서 벗어날 수 있음을 시사하였다. 이는 2011년 전후로는 회계장부에 있어 차이점이 분명히 존재하기에 2011년 이후로 작성된 회계장부를 기반으로 데이터 마이닝을 해야 더 정확한 관리종목 예측 모형을 만들어 낼 수 있다는 것을 의미한다. 또한, 2018년 말 회계 외부감사법이 개정된 이후로 관리종목 지정에 있어 이전보다 더 엄격한 기준이 적용되었다. 본 연구는 K-IFRS 개정과 최근 외부감사 개정을 반영한 2011년~2020년의 관리종목 표본 기업들을 반영하였기에 기존 선행연구보다 미래 관리종목 지정을 더 정확하게 예측할 것으로 기대된다. 선행연구와 이러한 점을 반영하여 전통적인 통계 기반 분석인 로지스틱 회귀분석과 기계학습을 기반으로 한 의사결정나무, 서포트 벡터 머신, 소프트 보팅, 랜덤 포레스트, LightGBM을 통하여 각 모형을 비교분석하고자 한다. <표 2>는 조사한 선행연구의 논문을 같은 유형별로 나누어 분류하였다. <표 3>은 진행할 분석방법과 관련된 논문을 정리하고 분류한 후 제시하였다.

<표 3> 검증에 활용할 머신러닝 접근법

분석 방법	참고 문헌
로지스틱 회귀분석	Ohlson(1980), 남규민(2018), 방소남·제혜금(2020), 조경인·김영민(2021)
의사결정나무	James et al.(2013), 남규민(2018), 엄하늘 외(2020)
서포트 벡터 머신	James et al.(2013), 김형준 외(2019),
소프트 보팅	전병욱 외(2021)
랜덤 포레스트	James et al.(2013), 엄하늘 외(2020), 조경인·김영민(2021)
LightGBM	Ke et al.(2017), 이현미 외(2020), 조재영 외(2021)

Ⅲ. 연구 방법

3.1. 로지스틱 회귀분석

종속변수가 연속형 변수가 아닌 범주형 변수인 경우 일반적인 선형회귀식으로는 설명하기 어렵다. 이러한 문제는 로지스틱 회귀식을 통해 식을 도출하여 해결할 수 있다. 종속변수가 2개인 이진 변수의 경우 이항형 로지스틱 회귀분석을 통하여 식을 도출한다. 로지스틱 회귀모형은 독립변수 $X_1, X_2, ..., X_p$가 주어졌을 때 종속변수가 속할 확률을 의미하며, 독립변수의 범위는 $[-\infty, \infty]$이고 종속변수의 범위는 $[0,1]$이다. 로짓(logit) 변환을 한 후 식을 조작하여 최종적으로는 시그모이드(Sigmoid) 함수를 사용하여 로지스틱 회귀분석을 수행한다(조경인·김영민, 2021).

$$P(y=1|X_1, X_2, ..., X_p) = \frac{1}{1+e^{(\beta_0+\beta_1X_1+\beta_2X_2+\beta_3X_3+...+\beta_pX_p)}}$$

$$Logit(odds) = \log\frac{P}{1-P} = \beta_0+\beta_1X_1+\beta_2X_2+\beta_3X_3+...+\beta_pX_p$$

위 식에서 P는 관리종목으로 지정될 확률이며, $X_i(i=1,2,...,p)$는 관리종목에 해당할 확률에 영향을 주는 재무변수들을 의미한다. 이는 승산비(odds ratio)로도 표현하며, 관리종목에 해당될 확률 P는 다음과 같이 표현할 수 있다.

$$P(y=1|X_1, X_2, \cdots, X_p) = \frac{\exp(\beta_0+\beta_1X_1+\beta_2X_2+\beta_3X_3+\cdots+\beta_pX_p)}{1+\exp(\beta_0+\beta_1X_1+\beta_2X_2+\beta_3X_3+\cdots+\beta_pX_p)}$$

관리종목으로 지정될 경우 종속변수는 1, 관리종목으로 지정되지 않으면 종속변수는 0인 두 집단에 속할 확률을 예측하기 위한 비율을 측정한다. 관리종목 지정확률 P가 일정수준을 넘으면 관리종목 기업으로, 그렇지 않으면 비관리종목 기업으로 판단한다. 일반적으로 정하는 일정수준인 P의 값을 0.5를 기준으로 판단하지만, 이에 대해서는 임의로 정할 수 있다(남규민, 2018; 방소남·제혜금, 2020).

3.2. 의사결정나무

의사결정나무는 의사결정 규칙을 나무 구조로 나타내어 전체 자료를 몇 개의 소집단으로 나누어 분류 혹은 예측하는 분석 방법이다. 설명가능한 P개의 변수인 $X_1, X_2, ..., X_p$로 구성된 특징 공간(feature space)을 집합 J개인 비중첩 영역(non-overlapping regions)으로 나눈다. 그 후 훈련 관측치가 대응하는 R_j 영역에 동일한 예측을 실시하는 분석 방법이다. 분류(classification)의 경우, 영역 내 가장 많이 관측된 클래스를 영역에 할당하기에 분류 오류율은 가장 많은 클래스에 속하지 않는 해당 영역의 비율이다.

$$Classification\ \ Error = 1 - \max_k(\hat{p}_{mk})$$

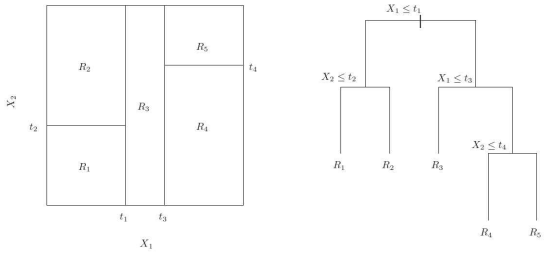
여기서 $\hat{p}_{mk}$는 k 클래스로부터 구분된 m 번째 영역에서 관측된 관측치의 비율을 의미한다. 하지만 분류 오류는 나무의 성장에 대해 효과적이지 않기 때문에 지니 계수(Gini index) 혹은 엔트로피 지수(Entropy)를 활용한다(James et al., 2013).

$$Gini\ \ index = \sum_{k=1}^{K} \hat{p}_{mk}(1 - \hat{p}_{mk})$$

$$Entropy = -\sum_{k=1}^{K} \hat{p}_{mk} \log \hat{p}_{mk}$$

의사결정나무는 노드(node: 마디)라고 불리는 구성요소들로 이루어진 나무모양으로 만들어진다. 각 노드의 의사결정 지점에서 독립변수를 중요도 기준에 따라 줄기(branch)를 만들어 나가며, 마지막 노드에서 종속변수에 대해 판단을 내리며, 데이터를 분할하는데 있어서 최상의 방법을 찾을 때까지 가지치기를 통해 의사결정나무를 완성한다. 의사결정나무의 알고리즘은 CART(Classification And Regression Tree), C4.5, CHAID(Chi-squared Automatic Interaction Detection) 등의 알고리즘이 있다(남규민, 2018; 엄하늘 외, 2020).

의사결정나무는 정보의 균일도에 따라 규칙을 생성하기 때문에 데이터를 따로 가공할 필요가 없으며, 의사결정이 이루어지는 시점과 성과 파악을 시각화 할 수 있기 때문에 결과를 해석하고 이해하기 쉽다는 장점이 있다. 하지만, 나무의 깊이에 따라 학습 데이터에 대해 과소적합(underfitting) 혹은 과대적합(overfitting)이 이루어지는 위험에 노출되어 있기 때문에 편향된 결과를 초래할 수 있다. <그림 1>은 의사결정나무 모형에 대한 설명을 시각화한 것이다.



본 그림은 2차원 특징 공간(feature space)에서 의사결정나무 알고리즘을 실시한 결과를 나타냄. 왼쪽 그림은 2차원 공간에서의 반복이진분할(recursive binary splitting)을 통한 결과를 나타내며, 오른쪽 그림은 왼쪽 그림에 따른 결과를 나무 모형으로 나타냄(출처: James et al., 2013).

<그림 1> 2차원 특징 공간(feature space)의 의사결정나무

3.3. SVM(Support Vector Machine)

서포트 벡터 머신(SVM)은 독립변수가 많은 고차원 공간에서 학습 데이터가 속한 그룹을 분류하는 선형 분리자를 찾는 기하학적 모델이다. 서포트 벡터 머신은 관측치 간의 거리를 의미하는 마진(margin)을 최대화하는 초평면(hyperplane)을 찾아내어 구분을 좀 더 확실히 하는 데 목적이 있으며, 최대 마진을 가지는 비확률적 선형 판별 분석에 기초한 이진 분류기이다. p차원 공간에서의 초평면이란 p-1차원의 평면 아핀부분공간(flat affine subspace)을 의미하며, 선형 커널(linear kernel)을 사용하는 SVM을 수식으로 나타내면 다음과 같다(김형준 외, 2019; James et al, 2013).

$$\underset{\beta_0, \beta_1, ..., \beta_p, \epsilon_1, \epsilon_2, ..., \epsilon_n}{\max imize} \quad M$$

$$y_i(\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + ... + \beta_p x_{ip} \ge M(1 - \epsilon_i)$$

$$\epsilon_i \ge 0, \ \sum_{i=1}^{n} \epsilon_i \le C$$

여기서 C는 음이 아닌 조정 매개변수(nonnegative tuning parameter)를 의미하며, M은 마진(margin)의 너비를 의미한다. $\epsilon_1, \epsilon_2, \cdots, \epsilon_n$은 슬랙변수(slack variable)로 마진에 대한 개별 관측치들의 오차를 의미한다.

서포트 벡터 머신은 분류 문제에 있어서 좋은 성능을 이루어내고, 훈련 시간이 오래 걸리지만 정확성이 뛰어나며 다른 기계학습보다 과적합 가능성이 낮은 특징을 지니고 있는 기계학습 알고리즘이다.

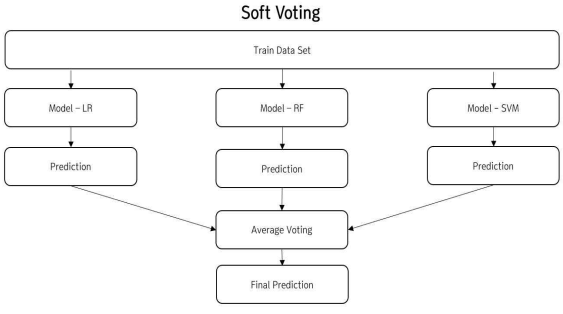
3.4. 소포트 보팅

보팅(Voting) 방식은 주어진 데이터셋으로 여러 분류기를 학습하여 다른 예측결과를 만들어내는데, 이를 다수결 혹은 평균 투표를 통하여 얻은 결과로 최종 결과를 예측하는 앙상블

기법이다. 투표를 하는 방식은 하드 보팅과 소프트 보팅 방식으로 나뉘어져 있는데, 하드 보팅의 경우 다수결에 따라 투표에 대한 의사결정을 진행하며, 소프트 보팅을 적용하면 각 알고리즘이 예측한 확률을 평균 내어 투표에 대한 의사결정을 한다. 아래 식은 소프트 보팅을 통해 클래스를 구분하는 식이다.

$$\hat{y} = \arg\max_{i} \sum_{j=1}^{M} \omega_j p_{ij}$$

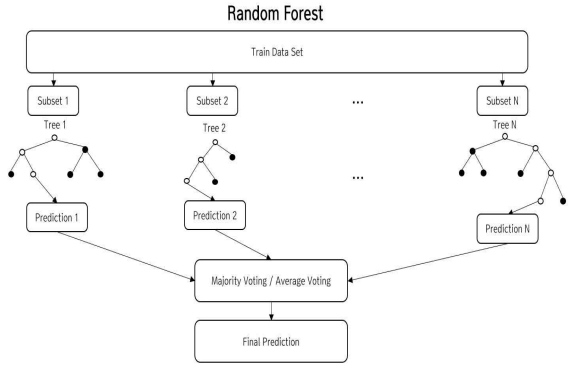
변수 p는 각 학습 모델을 통해 도출된 클래스별 확률값을 의미하고, 각 학습모델별로 도출된 확률값에 가중치 w를 설정한다. 최종적으로 가장 큰 확률을 가진 클래스를 출력한다(전병욱 외, 2021). <그림 2>는 본 연구의 소프트 보팅 학습 과정을 시각화한 것이다.



<그림 2> 소포트 보팅 개형도

3.5. 랜덤 포레스트

랜덤 포레스트는 Breiman(2001)에 의해 제안된 방법으로, 배깅(bagging) 분석방법의 일종으로 데이터셋을 일정한 크기로 복원추출하는 부츠스트래핑(bootstrapping)을 하여 복원추출한 데이터셋을 다수의 의사결정나무 분류기를 통하여 거대한 숲을 만들어 결과를 도출하는 앙상블 기법 중 하나이다. 랜덤 포레스트는 부츠스트랩을 활용하여 비편향 모형들의 예측값의 평균을 활용하기 때문에 과적합으로 인한 예측값의 편차와 분산을 줄일 수 있다. 하지만 모든 변수를 활용하면 추정된 모형들 간의 상관관계를 높일 수 있기 때문에 변수를 비복원추출로 통하여 모형을 구축한다(조경인·김영민, 2021). 다수의 결정트리를 만들기 위해 예측인자와 관측치에 대한 무작위 샘플링을 반복하며, 수많은 결정트리에서 얻은 예측 범주를 다수결 투표 방식을 통해 최종 범주를 예측한다(엄하늘 외, 2020). <그림 3>은 본 연구의 랜덤 포레스트 개형도를 시각화한 것이다. 또한 랜덤 포레스트는 의사결정나무와 같이 변수의 중요도를 제시한다. 하지만 일반적인 선형회귀분석과 같은 통계적 분석 방법과는 달리 개별변수에 대한 통계적 유의성 검정을 시행하지는 않는다. 변수의 중요도는 각 부츠스트랩 데이터 집합에서 뽑히지 않았던 데이터인 out of bag(OOB) 오차를 구하고 각 추정된 모형에서 특정변수의 값을 랜덤으로 섞은 데이터 집합에 대해 OOB 오차를 구한다.



<그림 3> 랜덤 포레스트 개형도

$$d_i = |e_i - r_i| \ , \quad \bar{d} = \frac{1}{t}\sum_{i=1}^{t} d_i \ , \quad s_d^2 = \frac{1}{t-1}\sum_{i=1}^{t}\left(d_i - \bar{d}\right)^2$$
$$i = 1,2,\cdots,t$$

변수 V_i의 중요도는 다음과 같이 계산된다(조경인·김영민, 2021).

$$V_i = \frac{\bar{d}}{s_d^2}$$

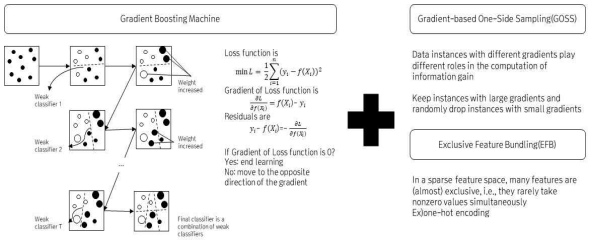
다수의 결정나무를 합하여 예측력을 높였지만, 하나의 나무그림으로 의사결정 절차를 표현할 수 없기에 랜덤 포레스트에서 해당 변수가 결과 모형에 중요한 영향을 끼친다고 해석하기에는 어려움이 있다. 이전만큼의 해석력까지는 아니지만, 변수들의 중요도 산출을 통해 변수의 중요도 순위를 확인할 수 있다(James et al., 2013).

3.6. LightGBM

LightGBM은 의사결정나무 기반 앙상블 알고리즘 모형 중 하나로, Ke et al.(2017)이 제시하였다. LightGBM은 의사결정나무를 차례대로 학습하며 각각의 나무는 선행하는 나무에 존재하는 오류를 개선하는 방식으로 생성된다. LightGBM은 GBM(Graient Boosting Machine)을 발전시킨 모형으로 XGBoost 이후에 나온 모형이다. GOSS(Gradient-based One-Side Sampling) 및 EFB(Exclusive Feature Bundling)을 사용하였고, 기존의 GBM 알고리즘에서 사용하는 트리의 깊이(depth wise)나 균형 트리(level wise)로 분할하는 것과는 달리 리프중심(leaf-wise) 방법으로 트리를 분할함으로써 정확도를 향상시켰으며, 메모리 사용량을 감소시키고 훈련 속도를 빠르게 만들었다(이현미 외, 2020; 조재영 외, 2021). LightGBM의 기반 알고리즘인 GBM을 식으로 표현하면 다음과 같다.

$$F(x)= \sum_{m=1}^{M} F_m(x)= \sum_{m=1}^{M} \beta_m h(x;a_m)$$

$F(x)$는 x변수들 기반으로 종속변수 y의 근사함수를 나타낸다. $h(x;a_m)$는 a_m의 파라미터를 갖는 m개의 의사결정 나무이며, B_m은 손실함수 $L(y,F(x))=[y-F(x)]^2$의 최소화에 따라 결정되는 값이다. <그림 4>는 부스팅 학습 과정과 LightGBM의 학습 개요를 나타낸다.



<그림 4> LightGBM 학습 개요(DataScience, 2020)

3.7. 데이터 설명

<표 4> 연도별 기업 표본수

년도	관리종목 표본	비관리종목 표본	합계
2011	14	14	28
2012	23	23	46
2013	13	13	26
2014	16	16	32
2015	27	27	54
2016	31	31	62
2017	30	30	60
2018	39	39	78
2019	67	67	134
2020	87	87	174
합계	347	347	694

본 연구는 관리종목 지정 기업의 재무적 특성과 예측 방법을 기술한 선행연구(신동인·곽기영, 2018)를 참고하여 재무데이터를 활용한 관리종목 예측 모델을 설계하고자 한다. 이를 위해 Python을 활용하여 분석을 하였으며, 오픈소스 라이브러리인 Scikit-learn을 주로 참조하였다. 또한 클라우드 기반 주피터 노트북 개발환경인 Google Colab을 사용하였다. 금융시장과 기업 분석에 다양한 데이터를 제공하고 있는 에프앤가이드의 DataGuide 5.0을 활용하여 K-IFRS가 시행된 2011년부터 2020년까지의 전체 코스닥 기업의 재무 데이터를 수집하였다. 각 연도마다 관리종목으로 지정된 기업들의 재무데이터를 선별하였다. 수집된 데이터 중 기업인수 목적의 스팩(SPAC) 기업은 타 기업과 재무비율 상의 차이가 있어 대상에서 제외하였다. 해당 기업들이 관리종목으로 편입된 시점의 해당 년도 재무 데이터를 기준으로 결측치가 존재하는 경우 코스닥 기업 데이터 전체를 삭제하고, 이러한 결측치 이외에도 '완전잠식'이 되어 있는 경우의 기업은 극단치로 간주하였다. 따라서 결측치나 극단치가 없는 347개의 관리종목 기업을 표본으로 선정하였다. 또한 관리종목과 비관리종목에 대해 불균형하게 표본추출을 하는 경우 관리종목 지정 예측에 부정확한 결과가 나타날 수 있기 때문에(조재영 외, 2021) 관리종목을 제외한 전체 코스닥 기업에서 무작위로 추출한 1:1 쌍대표본으로 347개의 비관리종목 기업 표본을 같이 선정하였다. <표 4>를 통하여 각 연도 별 관리종목 및 비관리종목 표본 개수를 제시하였다.

3.8. 모델 비교 및 평가 기준

<표 5> 머신러닝 모형 비교

분류	모델1	모델2	모델3	모델4	모델5	모델6
	로지스틱 회귀분석	의사결정 나무	서포트 벡터머신	소프트보팅	랜덤 포레스트	부스팅
분류기	LR	DT	SVM	LR, RF, SVM	RF	LightGBM
학습 데이터*	(555,22)	(555,22)	(555,22)	(555,22)	(555,22)	(555,22)
테스트 데이터	(139,22)	(139,22)	(139,22)	(139,22)	(139,22)	(139,22)
정규화 (normalization)	정규화	비정규화	정규화	정규화	비정규화	비정규화

*데이터셋은 각 행과 열에 해당하는 수치이다. 데이터셋을 전체 694개 기업 중 80%를 학습 데이터로, 나머지 20%는 테스트 데이터로 활용하였다. 학습데이터 555개가 행, 22개가 열에 해당한다.

<표 6> 혼동 행렬(Confusion Matrix)

예측 값 / 실제 값	N(0, 비관리종목)	Y(1, 관리종목)
N(0, 비관리종목)	TN(True Negative)	FP(False Positive)
Y(1, 관리종목)	FN(Faluse Negative)	TP(True Positive)

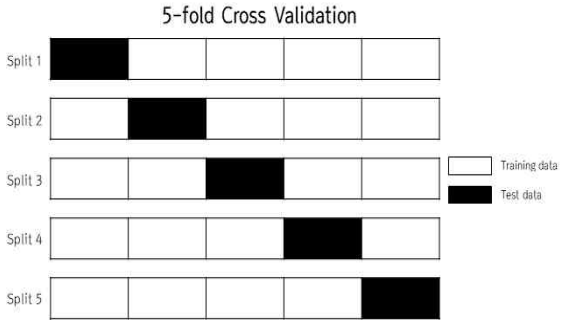
정확도(accuracy)=(TN+TP) / (TN+FP+FN+TP)
정밀도(precision)=TP / (FP+TP)
재현율(recall)=참 긍정률(TP Rate)= TP / (FN+TP)
F1-Score=2*(rrecision * recall) / (precision+ecall)
Kappa Statistics=(Pr(a)-Pr(e)) / (1-Pr(e))
Pr(a): 예측이 일치할 확률, Pr(e): 예측이 우연히 일치할 확률
거짓 긍정률(FP Rate)=1-TNR=FP / (TN+FP)

본 연구에서는 선정된 데이터 694개의 코스닥 기업 중 학습 데이터는 80%, 평가 데이터는 20% 비율로 사용하였다. <표 5>는 관리종목 지정 예측에 사용된 로지스틱 회귀분석, 의사결정나무, 서포트 벡터 머신, 앙상블 모델의 주요 학습기와 데이터 구조, 데이터 정규화(normalization) 과정 진행 유무를 의미한다.

평가 기준으로는 평가 데이터의 예측 정확도(accuracy) 뿐만 아니라 과적합에 대한 오류를 검증하기 위한 5겹 교차검증에 대한 정확도, 정밀도(precision), 재현율(recall), F1-Score, Kappa Statistics, ROC-AUC(Receiver Operating Characteristic-Area Under the ROC Curve) 점수를 이용하였다. <표 6>는 혼동 행

렬과 각 평가 기준에 대한 설명을 적은 것이다. 여기서 FP는 실제로는 비관리종목이지만 관리종목으로 예측한 비율을 의미하며, FN은 실제로는 관리종목이지만 비관리종목으로 오분류한 비율을 의미한다. 관리종목으로 지정이 되어서 손실이 나는 관점에서 봤을 때 비관리종목이지만 관리종목으로 예측한 비율이 많을수록 좋은 모형이라 할 수 없다. 즉, 전체 예측 정확도가 높으며 FP 비율이 적은 즉, 정밀도가 높은 모형이 좋다고 할 수 있다. 이 외에도 ROC-AUC를 이용하여 모형의 주된 평가를 진행할 예정이다. ROC 곡선은 혼동 행렬의 거짓 긍정률(FP Rate)을 가로축으로 두고, 혼동 행렬의 참 긍정률(TP Rate)를 세로축으로 두어 시각화한 그래프이다. ROC 곡선은 왼쪽 꼭대기에 가깝게 그려질수록 분류 성능이 우수하다고 판단하며, AUC를 통해 분류 성능을 객관적으로 평가할 수 있게 된다. AUC가 0.5이면 가장 랜덤한 값을 가지게 되고 분류 성능이 가장 낮은 경우를 의미하며, 1에 가까운 값일수록 분류 성능이 높은 모형이라 할 수 있다.

정확도의 경우 학습 데이터가 모집단의 일부분임에도 불구하고 학습 데이터의 특성을 지나치게 반영할 수 있다. 이러한 경우 학습 데이터에 대해 과대적합(overfitting)되었다고 표현하는데, 이러한 오류를 일반화 오류라 한다. 이러한 일반화 오류를 범하지 않게 하기 위해 신뢰할 만한 추정치를 구하는데, 이러한 검증 방법을 교차 검증(cross validation)이라 하며, 교차 검증 기법 중 하나인 K-Fold 교차검증을 시행한다. K-Fold는 데이터 집합을 무작위로 동일 크기를 갖는 K개의 부분 집합으로 나누고 그 중 1개의 집합을 평가 데이터로, 나머지 K-1개 집합을 훈련 데이터로 선정하여 분석 모형을 평가하는 교차 검증 기법 중 하나이다. K번 반복 수행하며, 결과를 다수결 또는 평균을 내어 모델의 성능을 평가한다. 본 연구의 경우 데이터 샘플을 5개로 나누어 평균 정확도를 산출하는 방식으로 검증하였다. <그림 5>는 5-Fold에 대한 절차를 시각화한 것이다.



<그림 5> 5-Fold 교차검증

3.9. 변수 선정 및 분석

김일(2005)의 연구에 의하면 관리종목과 비관리종목의 재무적 특성을 비교하기 위하여 수익성(총자본이익률, 매출액이익률), 성장성(자산증가율, 매출액증가율), 유동성(유동비율), 안정성(부채비율, 금융비용부담률, 이자보상비율), 활동성(총자본회전율, 재고자산회전율)을 나타내는 재무비율을 선택한 후 비교 분석을 하였다. 박창래·서영미(2015)는 관리종목과 비관리종목 기업의 재무적 특성을 비교하였다. 수익성(자기자본순이익률, 총자산수익률), 안정성(부채비율, 유동비율), 활동성(총자산회전율, 영업자산회전율, 재고자산회전율), 성장성(총자산성장률, 매출액성장률)을 나타내는 재무비율을 선별하여 재무적으로 차이가 있는지 분석하였다. 신동인·곽기영(2018)은 수익성, 안정성, 활동성, 성장성을 나타내는 21개 재무비율 변수를 선별 후, t-test 검정을 통해 최종적으로 18개 변수를 선정하여 관리종목 지정 예측 모형을 제시하였다. 이상의 선행연구들은 공통적으로 재무적 변수에 초점을 맞추어 관리종목 지정 예측 모형을 계량경제적 분석법을 통해 분석했다.

본 연구에서는 앞서 소개한 선행연구(김일, 2005; 박창래·서영미, 2015; 신동인·곽기영, 2018)들에서 제시된 21개의 재무비율 변수를 선별하여 사용하였다. 이러한 재무비율 변수들은 4가지 유형-수익성(Profitability), 안정성(Stability), 활동성(Activity), 성장성(Growth)으로 구분된다. 수익성 비율은 기업의 이익창출 능력을 보여주는 비율로서, 주주와 채권자로부터 조달한 자본의 운용과 생산과 판매활동과 같은 경영활동을 통하여 기업이 가지고 있는 자산 혹은 자본을 얼마나 효율적으로 사용하고 있는지를 평가할 수 있는 지표이다. 안정성 비율은 기업에게 조달된 자본을 통하여 얼마나 자산에 적절히 배분하였는지, 채무능력은 어떠한 지에 대해 확인할 수 있는 비율이다. 또한 기업이 내적, 외적인 환경의 변화에도 불구하고 얼마나 안정적으로 경영활동을 할 수 있는지에 대한 지표이기도 하다. 활동성 비율은 기업이 보유한 자산을 얼마나 효율적으로 활용하고 있는지 판단할 수 있는 비율이다. 성장성 비율은 일정기간 동안 기업의 규모나 이익이 얼마나 증가했는지를 측정하는 비율이다. <표 7>에는 선택된 변수의 유형과 그에 대한 산출식을 기술하였다. <표 8>에는 각 재무비율에 대한 기초 통계량을 기술하였다. 비계량적 측면에서 머신러닝의 특성치(features)를 선별하지 않은 이유는 첫째, 기존 연구와의 비교 가능성을 높이고, 둘째, 비교적 객관적으로 외부에서 판단 가능한 데이터를 바탕으로 한 알고리즘 개발에 도움이 되도록 하기 위해서다. 아울러 관리종목 지정에 관한 판단에 있어 일차적인 지표 데이터를 획득하고 이를 통해 보다 심도있는 판단을 내리기 위한 의사결정 절차 설계에도 도움이 될 것으로 본다.

관리종목과 비관리종목을 구분하여 재무비율의 분포를 시각적으로 확인한 결과는 다음 <표 9>와 같다. 바이올린 플롯(violin plot) 분포 확인을 통해 재무비율이 어느 구간에 몰려있는지 확인할 수 있으며, 관리종목은 1, 비관리종목은 0에 해당한다. 또한, 변수 간의 상관관계를 확인하기 위해 <표 10>, <표 11>에서 관리종목과 비관리종목의 변수 간 상관관계를 확인하였다.

<표 10>, <표 11>을 확인하면 관리종목과 비관리종목의 변

수 간의 상관관계에 있어 차이점을 보인다. 비관리종목의 경우 수익성 재무 변수의 상관관계가 수익성 비율의 재무 변수와 안정성 비율의 재무 변수와 강한 양의 상관관계를 띄고 있지만, 관리종목의 경우 P5(ROE-영업이익)는 수익성 재무 변수 중 P6(ROE-세전계속사업이익), P7(ROE-당기순이익)과는 음의 상관관계를 띄고 있음을 확인할 수 있다. 또한, S2(자기자본현금흐름률), S5(현금흐름/총자본), A4(총자본회전율)에 대해 강한 음의 상관관계를 갖고 있음을 확인할 수 있다. 이는 관리종목은 수익성에 대한 악화와 기업 자본 구조에 대한 위험이 같이 동반되어 있음을 시사한다.

<표 7> 변수 유형, 산출식

유형	변수		산출식
수익성 비율 (Profitability ratios)	P1	ROA(영업이익)(%)-ROA(Operating Income)(%)	(영업이익(연율화)/총자산(평균)) * 100
	P2	ROA(세전계속사업이익)(%)-ROA(Pretax Income)(%)	(세전계속사업이익(연율화)/총자산(평균)) * 100
	P3	ROA(당기순이익)(%)-ROA(Net Income)(%)	(당기순이익(연율화)/총자산(평균)) * 100
	P4	ROA(총포괄이익)(%)-ROA(Comprehensive Income, 3y)(%)	(총포괄이익(연율화)/총자산(평균)) * 100
	P5	ROE(영업이익)(%)-ROE(Operating Income)(%)	(영업이익(연율화)/총자본(평균)) * 100
	P6	ROE(세전계속사업이익)(%)-ROE(Pretax Income)(%)	(세전계속사업이익(연율화)/총자본(평균)) * 100
	P7	ROE(당기순이익)(%)-ROE(Net Income)(%)	(당기순이익(연율화)/총자본(평균)) * 100
안정성 비율 (Stability ratios)	S1	부채비율(%)-Total Liabilities to Total Equity(%)	(총부채/총자본) * 100
	S2	자기자본현금흐름률(%)-Cashflow to Shareholder's Equity(avg)(%)	(현금흐름(연율화)/지배주주지분(평균)) * 100
	S3	현금흐름/영업수익(%) - Cashflow to Revenue(%)	(현금흐름(연율화)/영업수익(평균)) * 100
	S4	현금흐름/총부채(%)-Cashflow to Total Liabilities(avg)(%)	(현금흐름(연율화)/총부채(평균)) * 100
	S5	현금흐름/총자본(%)-Cashflow to Total Equity(avg)(%)	(현금흐름(연율화)/총자본(평균)) * 100
	S6	현금흐름/총자산(%)-Cashflow to Total Assets(avg)(%)	(현금흐름(연율화)/총자산(평균)) * 100
	S7	현금흐름/총차입부채(%)-Cashflow to Total Debt(avg)(%)	(현금흐름(연율화)/총차입부채(평균)) * 100
활동성 비율 (Activity ratios)	A1	총자산회전율(회)-Asset Turnover Ratio(times)	영업수익(연율화)/총자산(평균)
	A2	자본금회전율(회)-Capital Stock Turnover Ratio(times)	영업수익(연율화)/자본금(평균)
	A3	총부채회전율(회)-Liability Turnover Ratio(times)	영업수익(연율화)/총부채(평균)
	A4	총자본회전율(회)-Equity Turnover Ratio(times)	영업수익(연율화)/총자본(평균)
성장성 비율 (Growth ratios)	G1	총부채증가율(전년동기)(%)-Total Liabilities Growth Rate(YoY)(%)	((총부채/총부채(-1Y)) - 1) * 100
	G2	총자산증가율(전년동기)(%)-Total Assets Growth Rate(YoY)(%)	((총자산/총자산(-1Y)) - 1) * 100
	G3	매출액증가율(전년동기)(%)-Sales Growth Rate(YoY)(%)	((매출액/매출액(-1Y)) - 1) * 100

기존의 선행연구(신동인·곽기영, 2018)에서는 관리종목과 비관리종목 간 기본적인 관련성을 검증하기 위하여 각 변수에 대한 정규화(normalization) 과정을 거친 후 평균의 차이에 대한 가설검정인 독립표본 t-test를 시행하였다. 하지만 이에 대해서는 문제점이 존재하였는데, 첫 번째는 재무변수의 정규화 과정을 거쳐도 각 재무변수가 정규분포를 따르지 않고 한쪽으로 치우친(skewed) 분포의 형태를 띄었다는 점이다. 이는 t-test의 기본가정인 정규성을 위배하며, 이러한 경우 비모수 검정 방법인 맨-휘트니 U 검정(Mann-Whitey U test)이 더 의미가 있다고 할 수 있다. 두 번째는 관리종목과 비관리종목의 각 재무비율이 평균 혹은 중앙값에 차이가 있다고 해도 해당 변수가 관리종목이 지정 예측에 대해 중요 변수인지에 대한 검증은 할 수 없다는 점이다. 만약 재무 변수에 대해 차이가 있었다고 해도 실제로는 해당 재무 변수가 관리종목 지정이 되는 중요한 변수가 될 수가 있기 때문이다.

따라서 본 연구는 기존 선행연구와는 다르게 관리종목과 비관리종목 간의 독립표본 t-test를 따로 시행하지 않았다.

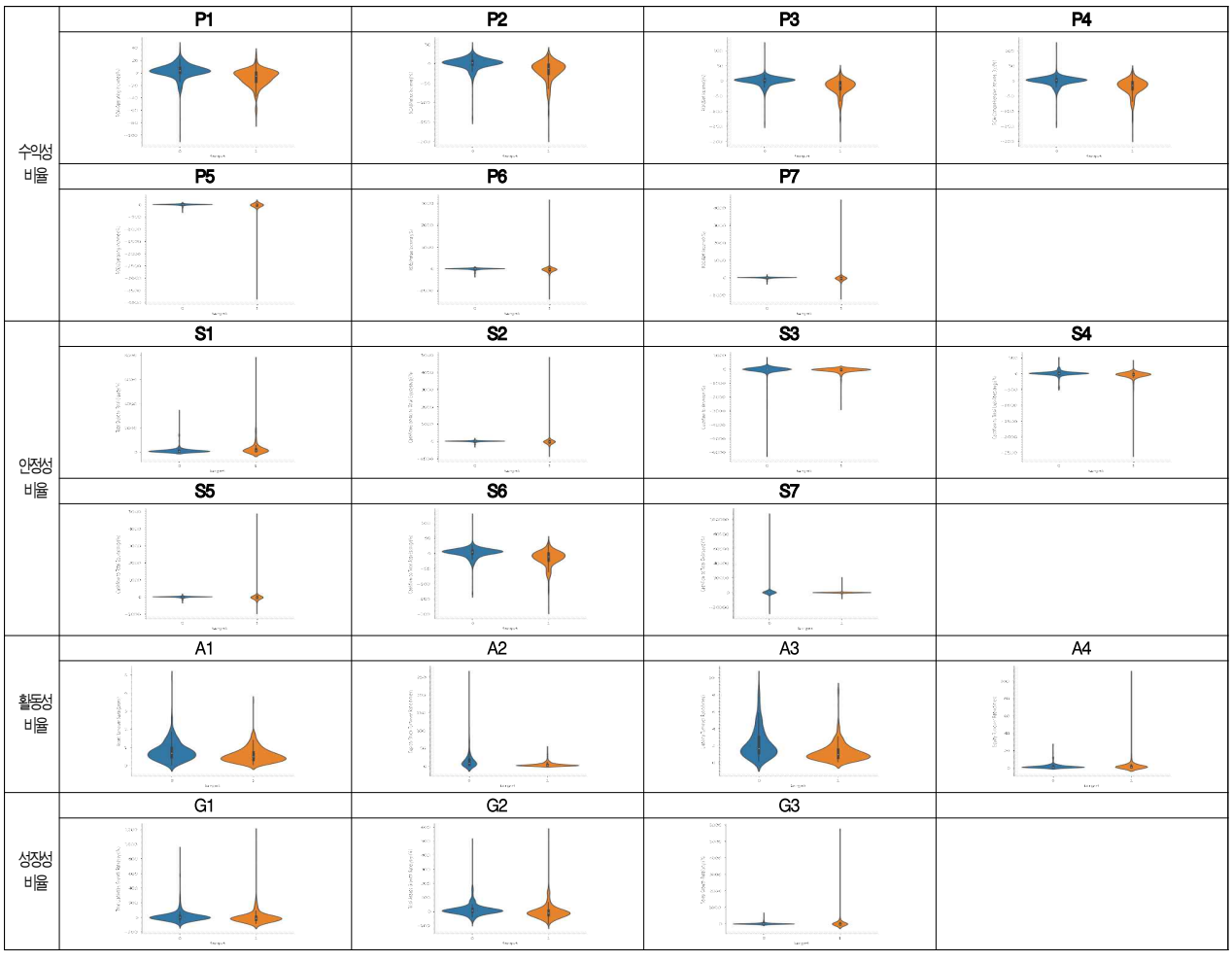
전통적인 회귀모형에서는 변수 간 상관관계가 높은 변수들은 다중공선성을 고려하여 변수를 적절히 제외한다. 하지만 본 연구에서는 머신러닝 분류 모델을 활용 시 최대한 많은 독립변수를 고려하여 분류 모델의 정확도를 높이고자 하였다. 이는 머신러닝을 활용한 분류모델이 다중공선성이 높은 변수도 고려하여 설명력을 높일 가능성이 있기 때문이다(김인호·이경섭, 2020; 유한별 외, 2021). 따라서 본 연구는 제시된 21개의 재무변수 비율을 토대로 관리종목 지정 예측 모형을 설계하였다. 수익성 비율, 안정성 비율, 활동성 비율, 성장성 비율로 나누어 총 21개의 재무변수에 대해 다양한 방법을 사용하여 분석을 진행하였다. 본 연구의 목적은 주어진 데이터를 가지고 가장 좋은 예측율을 얻는 것이기 때문에 다중공선성이라는 단점을 감수하고도 총 21개의 재무변수를 사용하였다.

<표 8> 변수 기초 통계량

유형	변수	관리종목(n=347)					비관리종목(n=347)				
		평균	표준편차	최솟값	중앙값	최대값	평균	표준편차	최솟값	중앙값	최대값
P1	ROA(영업이익)(%)	-7.71	13.81	-77.29	-5.44	30.46	2.50	12.42	-102.73	3.07	40.96
P2	ROA(세전계속사업이익)(%)	-18.47	27.29	-184.06	-11.97	25.62	0.07	18.34	-143.49	2.64	43.67
P3	ROA(당기순이익)(%)	-20.51	27.58	-184.06	-14.91	34.95	-0.71	18.90	-143.49	2.25	115.61
P4	ROA(총포괄이익)(%)	-20.50	28.41	-184.10	-14.57	33.64	-0.51	19.02	-143.34	2.22	116.35
P5	ROE(영업이익)(%)	-34.09	207.72	-3732	-13.32	69.84	3.22	29.68	-307.28	5.87	86.27
P6	ROE(세전계속사업이익)(%)	-43.94	196.86	-1274	-28.91	3038	-2.72	40.29	-361.16	4.72	84.64
P7	ROE(당기순이익)(%)	-46.17	255.12	-1086	-37.43	4308	-4.42	41.26	-361.16	4.13	153.85
S1	부채비율(%)	295.01	622.42	4.96	109.51	7450	122.62	256.71	1.27	60.46	3313
S2	자기자본현금흐름률(%)	-33.00	271.46	-725.81	-28.55	4710	3.27	40.17	-322.66	9.34	158.68
S3	현금흐름/영업수익(%)	-90.54	272.48	-2761	-21.53	72.42	-41.85	392.79	-6070	6.48	628.28
S4	현금흐름/총부채(%)	-45.43	160.40	-2531	-19.65	333.33	10.25	77.72	-475.09	12.45	476.96
S5	현금흐름/총자본(%)	-32.01	271.56	-830.26	-27.82	4710	0.07	41.75	-338.71	8.35	157.76
S6	현금흐름/총자산(%)	-16.52	27.39	-182.23	-11.11	39.35	1.63	19.11	-133.11	4.83	118.55
S7	현금흐름/총차입부채(%)	-97.36	1305	-8082	-30.80	19985	347.14	5907	-25708	23.01	104304
A1	총자산회전율(회)	0.57	0.41	0.02	0,.48	3.56	0.,79	0.55	0.01	0.68	4.87
A2	자본금회전율(회)	3.95	5.69	0.01	2.05	51.74	15.64	24.33	0.05	7.99	252.10
A3	총부채회전율(회)	1.24	1.15	0.03	0.96	8.69	2.26	1.80	0.07	1.68	9.70
A4	총자본회전율(회)	2.18	6.15	0.02	1.11	107.52	1.85	2.37	0.02	1.22	26.36
G1	총부채증가율(전년동기)(%)	4.70	109.82	-89.28	-14.54	1148	16.86	94.92	-93.99	-2.67	904.32
G2	총자산증가율(전년동기)(%)	-2.26	54.19	-89.43	-10.73	554.97	12.18	43.67	-77.01	5.72	490.56
G3	매출액증가율(전년동기)(%)	30.38	320.56	-98.15	-7.46	5566	8.78	54.40	-89.12	1.39	646.05

*변수 기초 통계량 중 최소값 및 최대값이 -1000 이하 혹은 1000 이상일 경우 소수점 자리는 반올림하여 정수로 표현하였음.

<표 9> 변수 간 관리종목/비관리종목 분포-수익성 비율(Profitability ratios), 안정성 비율(Stability ratios), 활동성 비율(Activity ratios), 성장성 비율(Growth ratios)



<표 10> 관리종목 변수 별 상관관계

유형	변수	수익성 비율							안정성 비율							활동성 비율				성장성 비율		
		P1	P2	P3	P4	P5	P6	P7	S1	S2	S3	S4	S5	S6	S7	A1	A2	A3	A4	G1	G2	G3
수익성 비율	P1. ROA(영업이익)(%)	1																				
	P2.. ROA(세전계속사업이익)(%)	0.67	1																			
	P3. ROA(당기순이익)(%)	0.66	0.97	1																		
	P4. ROA(총포괄이익)(%)	0.6	0.94	0.97	1																	
	P5. ROE(영업이익)(%)	0.23	0.04	0.02	0	1																
	P6. ROE(세전계속사업이익)(%)	0.19	0.39	0.39	0.39	-0.69	1															
	P7. ROE(당기순이익)(%)	0.13	0.31	0.34	0.34	-0.79	0.98	1														
안정성 비율	S1. 부채비율(%)	-0.07	-0.14	-0.14	-0.15	-0.08	-0.22	-0.18	1													
	S2. 자기자본현금흐름률(%)	0.11	0.3	0.33	0.33	-0.83	0.96	0.99	-0.15	1												
	S3. 현금흐름/영업수익(%)	0.45	0.4	0.43	0.33	0.02	0.11	0.1	-0.05	0.1	1											
	S4. 현금흐름/총부채(%)	0.48	0.52	0.52	0.41	0.02	0.12	0.1	0.02	0.11	0.65	1										
	S5. 현금흐름/총자본(%)	0.11	0.3	0.32	0.32	-0.83	0.97	1	-0.15	1	0.1	0.1	1									
	S6. 현금흐름/총자산(%)	0.65	0.96	0.99	0.95	0.02	0.39	0.33	-0.15	0.33	0.44	0.53	0.32	1								
	S7. 현금흐름/총차입부채(%)	0.3	0.34	0.34	0.33	0.02	0.08	0.06	-0.01	0.06	0.18	0.39	0.06	0.35	1							
활동성 비율	A1. 총자산회전율(회)	0.26	0.15	0.16	0.14	0.02	0.04	0.04	-0.03	0.05	0.31	0.16	0.05	0.19	0.01	1						
	A2. 자본금회전율(회)	0.3	0.23	0.23	0.19	0.08	0.03	0.02	0.04	0.02	0.18	0.14	0.02	0.23	0.05	0.36	1					
	A3. 총부채회전율(회)	0.24	0.13	0.15	0.14	0.08	0.06	0.05	-0.2	0.04	0.21	0.1	0.04	0.17	0.27	0.6	0.16	1				
	A4. 총자본회전율(회)	-0.01	0.11	0.14	0.14	-0.92	0.71	0.8	0.15	0.85	0.1	0.07	0.84	0.14	0.01	0.23	0.08	0.02	1			
성장성 비율	G1. 총부채증가율(전년동기)(%)	0.02	-0.11	-0.1	-0.11	0.05	-0.04	-0.03	0.01	-0.04	-0.04	-0.03	-0.04	-0.11	-0.06	0.05	0.06	0.05	-0.04	1		
	G2. 총자산증가율(전년동기)(%)	0.21	0.21	0.24	0.23	0.05	0.07	0.06	-0.1	0.05	0.15	0.2	0.05	0.23	0.09	0.08	0.08	0.08	-0.01	0.62	1	
	G3. 매출액증가율(전년동기)(%)	0.13	0.1	0.11	0.1	0.03	0.03	0.02	-0.04	0.02	0.08	0.06	0.02	0.1	0.03	0.07	0	0.11	-0.01	0.54	0.26	1

~-1 ~-0.79 ~-0.59 ~-0.39 ~-0.19 0 ~0.19 ~0.39 ~0.59 ~0.79 ~1

<표 11> 비관리종목 변수 별 상관관계

유형	변수	수익성 비율							안정성 비율							활동성 비율				성장성 비율		
		P1	P2	P3	P4	P5	P6	P7	S1	S2	S3	S4	S5	S6	S7	A1	A2	A3	A4	G1	G2	G3
수익성 비율	P1. ROA(영업이익)(%)	1																				
	P2.. ROA(세전계속사업이익)(%)	0.79	1																			
	P3. ROA(당기순이익)(%)	0.73	0.94	1																		
	P4. ROA(총포괄이익)(%)	0.71	0.92	0.98	1																	
	P5. ROE(영업이익)(%)	0.79	0.6	0.56	0.55	1																
	P6. ROE(세전계속사업이익)(%)	0.67	0.84	0.8	0.78	0.72	1															
	P7. ROE(당기순이익)(%)	0.63	0.8	0.84	0.82	0.7	0.97	1														
안정성 비율	S1. 부채비율(%)	-0.27	-0.28	-0.27	-0.26	-0.32	-0.43	-0.43	1													
	S2. 자기자본현금흐름률(%)	0.62	0.8	0.83	0.82	0.67	0.95	0.98	-0.35	1												
	S3. 현금흐름/영업수익(%)	0.6	0.55	0.55	0.55	0.34	0.34	0.35	-0.22	0.35	1											
	S4. 현금흐름/총부채(%)	0,67	0.75	0.81	0.8	0.41	0.53	0.57	-0.19	0.58	0.59	1										
	S5. 현금흐름/총자본(%)	0.58	0.72	0.75	0.74	0.62	0.87	0.89	-0.33	0.9	0.34	0.6	1									
	S6. 현금흐름/총자산(%)	0.68	0.87	0.92	0.91	0.52	0.73	0.77	-0.24	0.79	0.53	0.84	0.85	1								
	S7. 현금흐름/총차입부채(%)	0.07	0.15	0.15	0.15	0.04	0.11	0.11	-0.04	0.11	0.03	0.16	0.11	0.15	1							
활동성 비율	A1. 총자산회전율(회)	0.28	0.25	0.21	0.21	0.14	0.19	0.16	0.1	0.23	0.16	0.17	0.25	0.26	0.07	1						
	A2. 자본금회전율(회)	0.19	0.18	0.17	0.16	0.12	0.16	0.15	0.06	0.21	0.09	0.14	0.21	0.21	0.01	0.53	1					
	A3. 총부채회전율(회)	0.32	0.31	0.27	0.26	0.19	0.25	0.23	-0.17	0.22	0.13	0.37	0.25	0.31	0.2	0.47	0.24	1				
	A4. 총자본회전율(회)	0.02	0.06	0.04	0.05	-0.19	-0.03	-0.06	0.37	0.06	0.08	0.02	0.05	0.06	0	0.67	0.41	0.09	1			
성장성 비율	G1. 총부채증가율(전년동기)(%)	-0;03	-0.04	-0.05	-0.05	0.01	-0.01	-0.02	0.03	-0.03	-0.04	-0.04	0	-0.03	-0.02	-0.01	-0.01	-0.06	-0.04	1		
	G2. 총자산증가율(전년동기)(%)	0.18	0.2	0.19	0.2	0.19	0.18	0.17	0	0.17	0.1	0.16	0.2	0.22	0.02	0.04	0.08	0.02	-0.04	0.68	1	
	G3. 매출액증가율(전년동기)(%)	0.21	0.14	0.12	0.12	0.18	0.12	0.11	-0.05	0.11	0.08	0.12	0.17	0.16	0.01	0.15	0.08	0.11	0.06	0.13	0.24	1

~-1 ~-0.79 ~-0.59 ~-0.39 ~-0.19 0 ~0.19 ~0.39 ~0.59 ~0.79 ~1

Ⅳ. 실증 분석 결과

4.1. 로지스틱 회귀분석

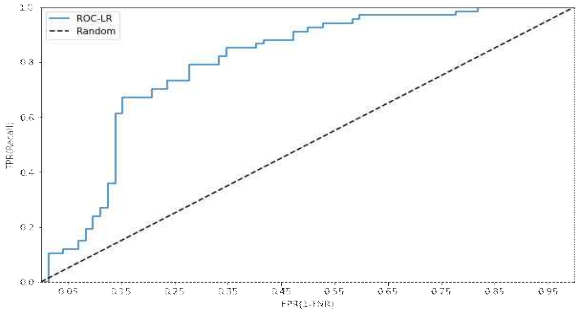
로지스틱 회귀모형과 같은 선형회귀분석의 경우 독립변수의 분포에 따라 예측 성능이 달라지는 결과를 초래할 수 있다. 따라서 최소값은 0, 최대값은 1인 분포로 변환하여 학습을 진행하였다. Scikit-learn에서 제공하는 GridSearchCV를 이용하여 예측 결과 성능을 높일 수 있는 여러 조합의 하이퍼 파라미터(hyper parameter)를 탐색하였다. <표 12>에서 로지스틱 회귀분석을 통한 하이퍼 파라미터 조합 결과 일부를 나타냈다. 최적 하이퍼 파라미터를 적용한 혼동 행렬과 예측 결과를 <표 13>에서 확인할 수 있다. 격자 탐색(grid search)이란 하이퍼 파라미터에 들어갈 수 있는 여러 값들을 전부 측정해보고 가장 높은 성능을 보이는 값을 찾는 방법이다. C는 상수 값으로 회귀분석의 y 절편값을, max_iter는 분석의 시행 횟수를, penalty는 함수에 대한 제약을 주는 조건을, solver는 함수를 풀이하는 방식을 의미한다. ROC 커브 곡선과 AUC는 <그림 6>에서 확인할 수 있다.

<표 12> 로지스틱 회귀분석-하이퍼 파라미터 조합 결과 일부

C	max_iter	penalty	solver	mean_test_score	std_test_score	rank_test_score
1	10000	none	newton-cg	0.7712	0.024	5
1	**10000**	**none**	**sag**	**0.7802**	**0.028**	**1**
1	10000	l2	liblinear	0.7459	0.028	20
1.5	10000	l1	liblinear	0.7514	0.029	14
1.5	10000	l2	saga	0.7514	0.021	9

<표 13> 테스트 데이터 로지스틱 회귀분석 결과

하이퍼 파라미터	C:1, max_iter: 10000, penalty: 'none', solver: 'sag'	
예측 값 / 실제 값	N(0, 비관리종목)	Y(1, 관리종목)
N(0, 비관리종목)	48	24
Y(1, 관리종목)	13	54
정확도: 0.7338, 5겹 교차검증 정확도: 0.7802, 정밀도: 0.6923, 재현율: 0.8060, F1-Score: 0.7448, Kappa Statistic: 0.4700, ROC-AUC: 0.7956		



<그림 6> 로지스틱 회귀분석 ROC 곡선

4.2. 의사결정나무

의사결정나무와 같은 경우 의사결정 규칙에 따라 노드가 생성되기에 데이터 정규화를 할 필요가 없다. 다만 의사결정나무를 생성하는 데 있어 가지의 최대 깊이를 정하지 않으면 학습 데이터에 대한 과대 적합이 일어나 평가 데이터에 대한 예측력이 떨어지는 결과를 갖게 되기 때문에 GridSearchCV를 통하여 하이퍼 파라미터를 설정하고 학습을 하였다.

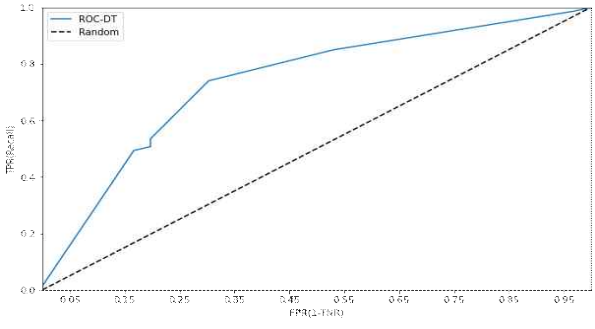
<표 14>에서 의사결정나무를 통한 하이퍼 파라미터 조합 결과 일부를 나타냈다. <표 15>에서 최적 파라미터를 적용한 혼동 행렬과 예측 결과를 확인할 수 있으며, max_depth는 결정 트리의 최대 트리 깊이를, min_samples_split은 자식 규칙 노드를 분할해 만들어 내기 위한 최소한의 샘플 데이터 수를, min_samples_leaf는 리프 노드가 될 수 있는 데이터 건수의 최솟값을 의미한다. <그림 7>에서는 의사결정나무 ROC 곡선을, <그림 8>은 결정 트리의 깊이를 제한하지 않았을 경우의 변수 중요도를 나타내어 재무변수에 대한 중요도 순위를 확인할 수 있다. 본 연구자의 Google Colab 또는 Github에서 의사결정규칙에 따른 노드를 확인할 수 있다.

<표 14> 의사결정나무-하이퍼 파라미터 결과 일부

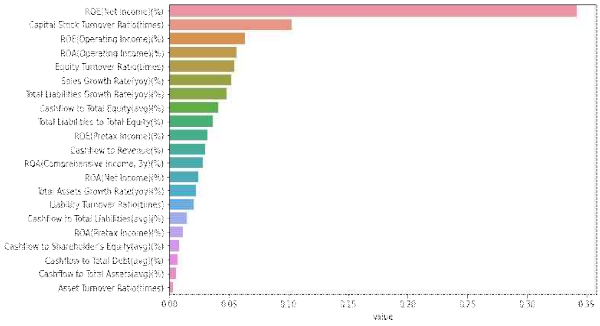
max_depth	min_samples_split	min_samples_leaf	mean_test_score	std_test_score	rank_test_score
1	1	2	0.7622	0.034	42
3	**1**	**2**	**0.7784**	**0.034**	**1**
5	3	2	0.7293	0.046	103
9	3	5	0.6829	0.050	172
10	5	5	0.7063	0.055	140

<표 15> 테스트 데이터 의사결정나무 결과

하이퍼 파라미터	max_depth: 3, min_samples_split: 1, min_samples_leaf: 2	
예측 값 / 실제 값	N(0, 비관리종목)	Y(1, 관리종목)
N(0, 비관리종목)	46	20
Y(1, 관리종목)	19	54
정확도: 0.7194, 5겹 교차검증 정확도: 0.7784, 정밀도: 0.7297, 재현율: 0.7397, F1-Score: 0.7347, Kappa Statistic: 0.4370, ROC-AUC: 0.7389		



<그림 7> 의사결정나무 ROC 곡선



*변수 중요도의 색은 의미없음.

<그림 8> 의사결정나무 변수 중요도

4.3. SVM

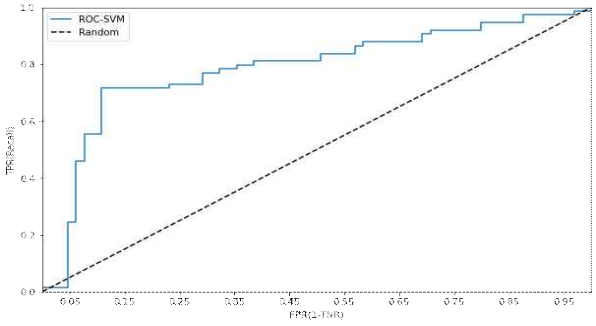
서포트 벡터 머신의 경우 선형(linear) 분리자인 경우를 사용하여 관리종목을 예측하였다. 서포트벡터 머신 역시 데이터 정규화 과정을 진행하고, 하이퍼 파라미터의 C는 데이터 샘플들이 다른 클래스에 놓이는 것을 허용하는 정도를 의미하고, gamma는 결정 경계(decision boundary)의 굴곡 정도를 의미한다. <표 16>에서 SVM을 통한 하이퍼 파라미터 조합 결과 일부를 나타냈다. SVM 또한 GridSearchCV를 통해 5겹 교차 검증을 하였다. 최적 파라미터를 적용한 예측 결과와 혼동행렬을 <표 17>에 제시되어 있으며, <그림 9>에는 SVM에 대한 ROC 곡선을 표현하였다.

<표 16> SVM - 하이퍼 파라미터 결과 일부

C	gamma	kernel	mean_test_score	std_test_score	rank_test_score
1.5	scale	linear	0.7568	0.026	11
2	auto	linear	0.7604	0.028	9
2.5	**scale**	**linear**	**0.7658**	**0.023**	**1**
3	scale	linear	0.7622	0.019	7
4	auto	linear	0.7658	0.023	1

<표 17> 테스트 데이터 SVM 결과

하이퍼 파라미터	C:2.5, gamma: 'scale', kernel: 'linear'	
예측 값 / 실제 값	N(0, 비관리종목)	Y(1, 관리종목)
N(0, 비관리종목)	50	15
Y(1, 관리종목)	21	53
정확도: 0.7410, 5겹 교차검증 정확도: 0.7658, 정밀도: 0.7794, 재현율: 0.7162, F1-Score: 0.7465, Kappa Statistic: 0.4827, ROC-AUC: 0.7848		



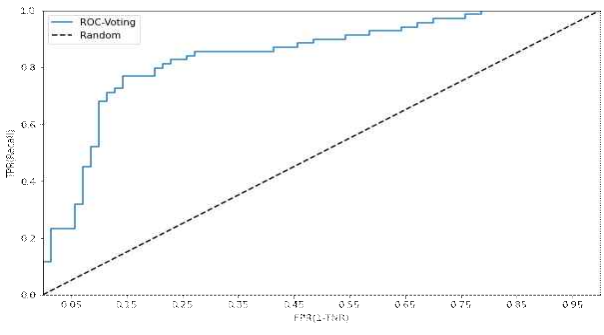
<그림 9> 서포트 벡터 머신 ROC 곡선

4.4. 소프트 보팅

앙상블 방법 중에 하나인 보팅은 여러 분류기를 학습하여 최종 예측 결과를 투표로 결정하게 된다. 본 연구에서는 로지스틱 회귀분석, 랜덤 포레스트, 서포트 벡터 머신을 통하여 예측 결과를 도출하고, 각 결과를 평균 내어 최종 예측 결과를 도출하는 소포트 보팅 방식을 사용하였다. 각 학습기는 하이퍼 파라미터가 적용된 분류기를 통하여 최종 결과를 예측하는데, 이에 대한 결과는 <표 18>과 <그림 10>에 제시하였다.

<표 18> 테스트 데이터 소프트 보팅 결과

분류기	로지스틱 회귀분석, 랜덤 포레스트, 서포트 벡터 머신	
예측값 / 실제 값	N(0, 비관리종목)	Y(1, 관리종목)
N(0, 비관리종목)	52	18
Y(1, 관리종목)	12	57
정확도: 0.7842, 5겹 교차검증 정확도: 0.7838, 정밀도: 0.7600, 재현율: 0.8261, F1-Score: 0.7917, Kappa Statistic: 0.5686, ROC-AUC: 0.8435		



<그림 10> 소프트 보팅(LR,RF,SVM) ROC 곡선

4.5. 랜덤 포레스트

랜덤 포레스트의 경우 의사결정나무를 여러 개의 숲으로 분류한 결과이다. 따라서 조정가능한 하이퍼 파라미터가 의사결정나무와 유사하지만, 결정나무의 개수를 정할 수 있는 파라미터인 n_estimators가 추가된다. <표 19>은 랜덤 포레스트를

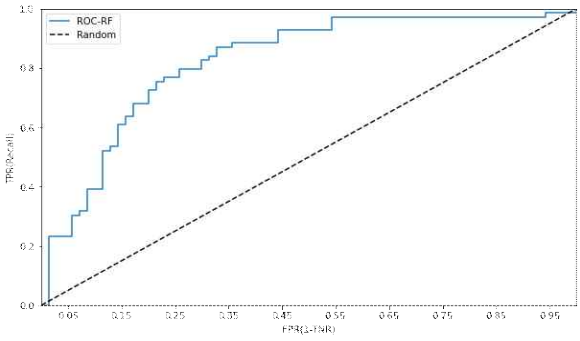
통한 하이퍼 파라이터 조합 결과 일부를 나타냈다. <표 20>은 최적 하이퍼 파라미터를 적용한 예측 결과와 혼동 행렬을 나타내었다. <그림 11>은 랜덤 포레스트의 ROC 곡선을, <그림 12>는 변수 중요도를 보여준다.

<표 19> 랜덤 포레스트-하이퍼 파라미터 결과 일부

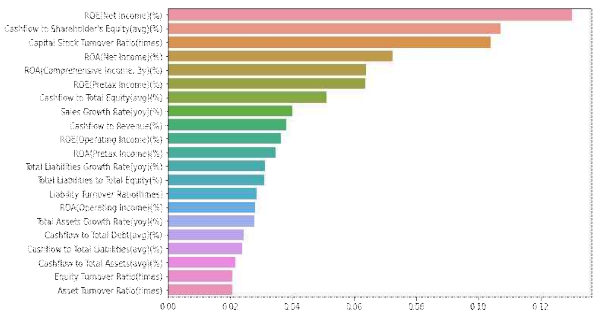
max_ depth	min_samples_ leaf	min_samples_ split	n_estimators	mean_test_ score	std_test_ score	rank_test_ score
2	2	2	500	0.7604	0.010	73
4	4	6	1000	0.7874	0.013	43
6	**4**	**2**	**500**	**0.8000**	**0.024**	**1**
6	6	10	1000	0.7946	0.022	12
8	6	10	500	0.7946	0.028	15

<표 20> 테스트 데이터 랜덤 포레스트 결과

하이퍼 파라미터	max_depth: 6, min_samples_leaf: 4, min_samples_split: 2, n_estimators: 500	
예측 값 / 실제 값	N(0, 비관리종목)	Y(1, 관리종목)
N(0, 비관리종목)	59	11
Y(1, 관리종목)	24	45
정확도: 0.7482, 5겹 교차검증 정확도: 0.7958, 정밀도: 0.8036, 재현율: 0.6522, F1-Score: 0.7200, Kappa Statistic: 0.4957, ROC-AUC: 0.8259		



<그림 11> 랜덤 포레스트 ROC 곡선



*변수 중요도의 색은 의미없음.

<그림 12> 랜덤 포레스트 변수 중요도

4.6. LightGBM

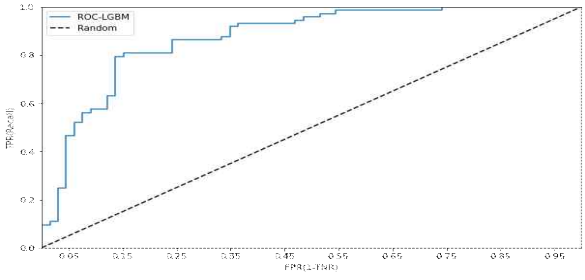
앙상블 기법 중 부스팅의 경우 오분류한 부분에 가중치를 크게 부여하여 다음 부트스트랩 데이터에 대해 잘 맞추도록 학습하는 머신러닝 알고리즘이다. XGBoost와 LightGBM는 현재 구글이 운영하는 기계학습 대회 플랫폼인 캐글에서 상당히 자주 우승하는 알고리즘 중 하나로, 다양한 분야에서 우수한 예측 정확도를 보여준다. LightGBM 역시 트리 기반 알고리즘이기에 하이퍼 파라미터 결정값은 의사결정나무의 파라미터 값과 유사하지만, 0과 1사이에 존재하고 부스팅을 할 시에 업데이트하는 학습률인 learning rate와 트리가 커져서 과적합되는 것을 제어하기 위해 데이터 샘플링 비율을 정하는 subsample 파라미터가 추가로 존재한다. 또한 파라미터 이름이 사이킷런 래퍼와 파이썬 래퍼가 차이가 있기 때문에 이를 유의하며 파라미터 튜닝을 해야 한다. min_child_samples의 경우 결정 트리의 min_samples_leaf와 같은 파라미터로 리프 노드가 되기 위한 데이터의 최솟값을 의미한다. num_leaves는 하나의 트리가 가질 수 있는 최대 리프 개수이다. <표 21>는 LightGBM을 통한 하이퍼 파라미터 조합 결과 일부를 나타내었다. <표 22>은 LightGBM에 대한 혼동 행렬과 하이퍼 파라미터 결과표이고, <그림 13>는 ROC 곡선을, <그림 14>에서는 부스팅을 통한 변수 중요도를 확인할 수 있다.

<표 21> LightGBM-하이퍼 파라미터 결과 일부

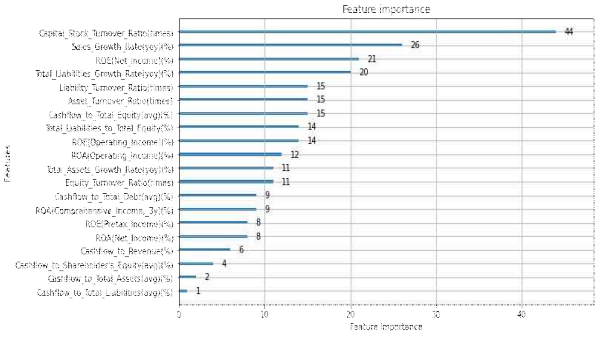
max_ depth	n_ estimators	num_ leaves	min_child _samples	subsample	mean_test_ score	std_test_ score	rank_test_ score
5	**500**	**6**	**10**	**0.5**	**0.7694**	**0.009**	**1**
5	500	14	10	1	0.7477	0.025	223
6	500	8	10	0.5	0.7586	0.029	121
6	1000	8	30	0.8	0.7676	0.030	43
10	1000	14	30	1	0.7514	0.032	181

<표 22> 테스트 데이터 LightGBM 결과

하이퍼 파라미터	max_depth: 5, n_estimator: 500, num_leaves: 6, min_child_samples: 10, , subsample: 0.5	
예측 값 / 실제 값	N(0, 비관리종목)	Y(1, 관리종목)
N(0, 비관리종목)	57	9
Y(1, 관리종목)	15	58
정확도: 0.8273, 5겹 교차검증 정확도: 0.7712, 정밀도: 0.8657, 재현율: 0.7945, F1-Score: 0.8286, Kappa Statistic: 0.6553, ROC-AUC: 0.8719		



<그림 13> LightGBM ROC 곡선



<그림 14> LightGBM 변수 중요도

4.7. 분석 결과 종합

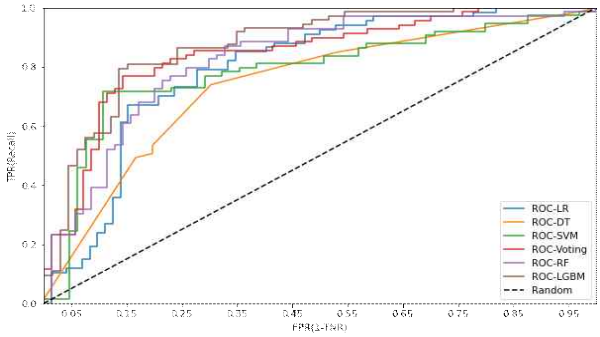
전체 분석결과를 종합해보면 <표 23>과 같다. 결과를 살펴보면 전체 표본은 총 6가지 모형별로 산출한 분류 정확도는 70% 초반부터 80& 초반 사이로 의사결정나무 모형이 71.22%로 가장 낮고 LightGBM 모형이 82.73%로 가장 높은 정확도를 보였다. 하지만 LightGBM의 경우 5겹 교차검증 정확도가 75.22%인 것을 보아 주어진 데이터에 과대적합 되었을 가능성도 어느 정도 존재한다는 것을 확인할 수 있다. 기업이 관리종목으로 지정되면 투자자본의 감소, 거래 유동성 악화와 같은 부정적인 효과를 고려했을 때 비관리종목을 관리종목으로 예측하는 손실이 관리종목을 비관리종목으로 예측하는 손실보다 더 크다고 볼 수 있다. 따라서 FP 비율이 낮은 모형들이 우수하다고 할 수 있으며, 전체 예측 정확도와 정밀도(precision)가 높을수록 모형의 성능이 우수하다고 할 수 있다. 대체적으로 앙상블을 통한 학습 모형이 단일 학습 모형보다 성능이 우수하다는 것을 <그림 15> 혹은 <표 23>의 AUC를 통해 확인할 수 있다. <표 24>는 의사결정나무 기반으로 산출된 변수 중요도의 상위 3개 변수들을 종합하여 제시하였다. 의사결정나무, 랜덤 포레스트, LightGBM에서 공통적으로 나온 재무변수는 P7-ROE(당기순이익)와 A2-자본금회전율로, 이 재무변수들이 관리종목 지정에 있어 상당한 영향을 준다는 것을 알 수 있다. 또한, 중요도가 높았던 변수들은 5.1에서 좀 더 자세히 다룰 예정이다.

<표 23> 모형 종합 결과

분류	Accuracy	5-Fold Acc Mean	Precision	Recall	F1-Score	Kappa	ROC-AUC
로지스틱 회귀분석	0.7338	0.7802	0.6923	0.8060	0.7448	0.4700	0.7956
의사결정나무	0.7194	0.7784	0.7297	0.7397	0.7347	0.4370	0.7389
서포트 벡터 머신	0.7410	0.7658	0.7794	0.7162	0.7465	0.4827	0.7848
소프트 보팅	0.7842	0.7838	0.7600	0.8261	0.7917	0.5686	0.8435
랜덤포레스트	0.7482	0.7928	0.8036	0.6522	0.7200	0.4957	0.8259
LightGBM	0.8273	0.7712	0.8657	0.7945	0.8286	0.6553	0.8719

<표 24> 변수 중요도 상위 3개 변수

의사결정나무	랜덤 포레스트	LightGBM
P7-ROE(Net Income)	P7-ROE(Net Income)	A2-Capital Stock Turnover Ratio
A2-Capital Stock Turnover Ratio	S2-Cashflow to Shareholder's Equity	G3-Sales Growth Rate(YoY)
P1-ROA(Operating Income)	A2-Capital Stock Turnover Ratio	P7-ROE(Net Income)



<그림 15> 전체 모형 ROC 곡선 종합 비교

Ⅴ. 결론

5.1. 연구 요약 및 토의

본 연구는 2011년부터 2020년까지의 코스닥 시장에 상장된 기업들을 대상으로 관리종목 예측 모형을 개발하였다. 21개의 재무변수에 대한 기초 통계량을 조사한 후 변수 간의 상관관계를 살펴보았으며, 21개의 재무변수를 기반으로 로지스틱 회귀분석, 의사결정나무, 서포트 벡터 머신, 소프트 보팅, 랜덤 포레스트, LightGBM을 활용하여 관리종목 지정 예측을 한 후 각 모형에 대한 비교 분석을 하였다. 또한 데이터의 불균형 해결 문제를 위하여 1:1 쌍대표본으로 비관리종목을 샘플링하여 연구를 진행하였다. 예측 결과를 혼동 행렬과 ROC 곡선을 활용하여 확인하였으며, 평가 데이터에 대한 정확도는 70% 초반에서 80% 초반 사이의 정확도를 보여주었다. 의사결정나무 기반으로 선정된 주요 변수들은 자본의 유동성을 설명할 수 있는 현금흐름과 기업의 경영능력을 확인할 수 있는 수익창출능력에 관련된 재무변수들이 선정되었다. 코스닥 상장 기업의 ROE(당기순이익), ROA(영업이익), 자기자본현금흐름률, 자본금회전율, 매출액증가율이 낮을수록 관리종목 지정 확률이 높아지는 것으로 확인할 수 있다. 제시된 재무변수들은 재무제표를 토대로 한 재무비율의 산출식이기에, 관리종목 지정 경향을 정확하게 파악하기 위해서는 각 재무변수들을 세세하게 알아볼 필요가 있다.

당기순이익은 영업이익에서 영업외수익과 특별이익을 합산하고 영업외비용, 특별손실을 모두 제하고 남은 순이익으로, 기업이 경영활동을 한 후 남은 순수 마진이라고도 볼 수 있다. 즉, 영업이익은 기업의 경영활동능력을 확인할 수 있는

지표가 될 수 있다. 그리고 당기순이익은 영업이익 이외의 수익과 비용을 합산한 결과로, 당해 년도에 대해 일시적인 수익 혹은 비용까지 합산하였다는 의미가 되기도 한다. 따라서 당기순이익의 악화는 기업의 핵심 사업능력 자체의 악화와 더불어 영업외수익인 일회성 수익 자체는 감소하고 비용은 증가하는 현상이 심화되었을 가능성이 높다고 판단할 수 있다. <표 1>에서 관리종목 지정 사유 요건 중 재무적 요인에 해당하는 3년간 2회 이상 법인세차감전계속사업손실 부분이 이와 비슷한 맥락이라 할 수 있다. 따라서 영업이익과 당기순이익의 지표의 변화에 주시하되, 영업이익과 당기순이익을 비교하여 일회성 이익이 영업이익 대비 어느 정도의 비중을 차지하는지도 살펴볼 필요가 있다.

자본금회전율은 영업수익을 자본금으로 나눈 비율로 해당 기업이 자본금을 얼마나 효율적으로 활용하여 수익을 창출하는지를 알 수 있는 지표이다. 자본금회전율이 낮다는 것은 기업의 자본금이 비효율적으로 사용되고 있음을 의미하며, 자본금회전율이 높을수록 자본금을 효율적으로 사용하여 영업수익을 가져옴을 의미한다. 하지만, 자본금회전율이 지나치게 높은데 총자산회전율이 지나치게 낮은 경우에는 해당 기업의 부채의 비중이 상당히 크다는 것을 의미한다. 따라서 자본금회전율이 낮을수록 관리종목 지정 경향성이 높아진다는 것은 해당 기업의 영업수익 즉, 매출액의 악화와 자본금의 증감 등을 고려해볼 수 있다. 또한, 자본금회전율과 총자산회전율의 차이를 비교하여 회전율의 괴리가 심하게 존재할 경우 해당 기업이 부채를 통한 수익 창출 비중이 높다는 것을 파악하고 기업의 자본 구조 건전성에 대해 종합적으로 살펴볼 필요가 있다.

자기자본현금흐름률은 현금흐름을 지배주주지분으로 나눈 비율이다. 지배주주지분은 지배기업과 비지배기업 중 지배기업에 해당하는 주주지분을 의미한다. 즉 자기자본현금흐름률이 악화된다는 것은 비지배기업을 제외한 지배기업의 현금흐름 창출능력이 낮음을 의미한다. 데이터에 해당하는 자기자본현금흐름률을 살펴보면 관리종목의 경우 음을 기록하고 있는데, 지배주주지분은 절대 양(+, positive)이라는 것을 감안하였을 때 현금흐름 자체가 음이며, 현금이 외부로 유출되는 경우가 더 많다는 것을 의미한다. 현금이 외부로 유출한다는 의미는 다양한 의미를 가지고 있는데, 대개 영업활동으로 인한 현금유출, 투자활동으로 인한 현금유출, 재무활동으로 인한 현금유출으로 구분할 수 있다. 영업활동으로 인한 현금유출은 제품의 생산 및 재화와 용역의 판매와 관련된 활동으로, 원재료, 상품 매입, 판관비, 법인세 납부 등이 이에 해당하는 예시이다. 투자활동으로 인한 현금유출은 현금을 대여하고 투자하는 활동에 쓰이는 경우로 예시로는 장단기 대여금의 대여, 유무형자산 취득 등이다. 재무활동으로 인한 현금유출은 현금의 차입 및 상환활동과 같이 부채와 자본에 증감된 활동으로, 장단기 차입금의 상환, 사채 상환, 유상감자 등이 재무활동 현금유출 사례에 해당한다. 이 중에서 가장 유의해야 할 부분은 바로 재무활동으로 인한 현금유출인데, 관리종목의 경우 자본잠식의 가능성이 높고 부채를 상환하는 비용이 많아질 수밖에 없다. 즉, 재무활동으로 인한 현금유출에 대한 비중이 높다면 해당 기업이 채무상환에 치중하고 있음을 의미한다. 만약 기업이 수익창출능력이 악화되고 기업의 안정성이 의심이 되는 상태라면, 재무활동으로 인한 현금유출을 확인하고 전체 현금흐름 유출과 채무상환의 비중을 비교함으로써 어느 정도의 관리종목 지정 경향성을 확인할 수 있다.

매출액증가율은 전년도 대비 매출액의 증감율을 나타낸 비율로 기업의 매출능력이 성장했는지에 대한 성과 지표이기도 하다. 매출액증가율이 악화하고 있다는 것은 기업의 성장력이 점차 퇴보한다는 의미로도 받아들일 수 있다. 종합적으로 결론을 내보자면 기업의 당기순이익 감소 및 매출책 증가율 감소는 기업의 경영활동으로 인한 수익창출 능력이 점차 감소하고, 수익이 적자인 상황에서 자본금회전율이 낮거나 혹은 자본금회전율이 지나치게 높은데 총자산회전율은 낮거나 혹은 재무활동으로 인한 현금유출이 지나치게 많은 경우 관리종목이 될 확률이 높다는 것을 알 수 있다. 이번 장에서는 관리종목 지정 예측 모형 연구 요약과 관리종목 지정에 대한 배경을 알아보기 위해 각 재무변수에 대해 자세히 살펴보았다. 당기순이익과 영업이익의 비교 및 매출액증가율을 통하여 기업 경영능력에 대한 검사와 현금흐름과 자본금의 구조를 파악하여 기업 안정성에 대해 파악하는 것이 관리종목 지정 예측에 대한 이해 폭을 넓힐 수 있을 것이다.

머신러닝 결과들에서 확인할 수 있는 이와 같은 추론적 사실들은 비재무적 판단과 함께 고려되어야 그 의미를 온전히 이해할 수 있을 것으로 본다. 감사의견 비적정, 지배구조 위반, 공시범위 위반과 같은 비재무적 사례들은 보통 기업의 재무성과가 악화되고 있거나 적자가 지속될 경우에 후행되는 현상이다. 재무성과가 좋거나 흑자인 경우에는 감사의견 비적정, 지배구조 위반, 공시범위 위반과 같은 사례는 일어나지 않지만 역의 사례로는 위와 같이 비재무적 사유에 해당되는 것이다. 따라서 기업의 재무요인이 40%로 관리종목으로 지정되지만, 나머지 60%의 비재무적 요인이 전부는 아니지만 대부분이 재무성과가 악화되어 관리종목으로 지정되었을 가능성이 상당히 높다고 본다. 따라서 머신러닝 결과들로부터 도출된 결론은 다소 1차적인 결과일 가능성이 높다는 점에 주의해야 한다. 예를 들어 현금흐름 창출력이 낮다고 해도 발전가능성이 높다고 판단될 경우 특례를 통해 관리종목이 유예될 수도 있다. 반면 공시의무를 소홀히 하거나 경영 투명성이 현저히 낮을 경우 투자자 보호를 위해 관리종목 지정이 결정될 가능성도 있다. 따라서 관리종목 지정의 성향에 관한 위와 같은 관찰은 본 연구의 실증적 자료에 바탕을 둔 것일뿐 성급하게 일반화될 수 있는 것은 아니다.

5.2. 시사점 및 한계점

본 연구의 학문적 시사점은 다음과 같다. 관리종목 제도가 유가증권시장과 코스닥시장을 통틀어 상장 기업의 상장폐지 가능성을 파악하여 투자자에게는 조기에 투자위험에 대한 인지를, 기업에게는 회생 기회를 주는 중요한 제도임에도 불구하고 관련된 연구들이 부족하다는 점에서 학문적 의의를 가진다. 또한 기존의 선행연구에서는 관리종목 지정 예측 모형에 대해 로지스틱 회귀분석과 의사결정나무 모형을 제시하였다. 하지만 선행연구의 의사결정나무 분석의 경우 데이터에 대한 과적합 가능성이 충분히 존재하고 이는 일반화 오류를 범하게 되는 결과를 초래하게 된다. 그리고 회계장부가 K-IFRS로 개정되고 난 이후부터는 2011년 전후를 기준으로 분명히 회계장부 상 차이가 존재하였고, 이는 부정확한 예측력을 이끌어 내는 결과를 초래할 가능성이 존재하였다(류예린 외, 2020). 이러한 사항들을 고려하여 본 연구에서는 기존 연구 방법론과 함께 다양한 머신러닝 기법을 적용하여 예측 모형을 비교 분석한 결과를 제시함으로써 기존 선행연구의 한계를 극복하는 데 기여했다고 볼 수 있다. 관리종목 지정에 영향을 미칠 수 있는 비재무적 요인을 포함한 다양한 측면의 연구 노력이 뒤따라야 할 것으로 판단한다(남기정 외, 2019).

관리종목으로 지정된 기업의 경우 조속한 시일 내에 관리지정 탈피를 하지 않으면 상장폐지 위기에 놓이며, 이를 모면하기 위해 회계부정 혹은 분식회계가 일어날 가능성이 높다(권기현 외, 2012). 특히 재무제표를 부풀리거나 조정할 수 있으며 감사인과의 개인적인 유대관계를 통해 기업의 의도적인 회계분식을 감출 수도 있다(김인상 외, 2016; 김수정·문보영, 2018). 본 연구의 실무적 기여는 코스닥 시장 내 시장참여자들에게 있어 관리종목 지정에 대한 사전 예측을 확인할 수 있도록 기여했다는 점이다. 관리종목으로 편입될 가능성을 사전에 판별하기 때문에 투자자들의 투자 포트폴리오 리스크 관리에 많은 영향을 끼칠 것으로 예상된다. 특히 거래소는 주로 매년 3월과 8월경 관리종목 시장조치를 진행하는 것을 고려해봤을 때, 해당 시기에 투자의사결정을 내리거나 포트폴리오 리밸런싱을 하는 투자자들에게는 실무적인 도움이 될 것으로 기대한다. 혹은 반대로 관리종목 지정으로 인한 주가 하락을 고려해봤을 때, 공매도를 통한 투자 전략 또한 구상할 수 있을 것이라 예상된다. 또한 기업들에게 있어서는 관리종목으로 편입될 가능성이 있는 기업들이 관리종목으로 지정될 위험성을 조기에 인지하여 경영활동과 기업 자본구조에 대한 변화를 일으킬 수 있는 계기가 될 수 있을 것이다.

본 연구는 다음과 같은 한계점을 지닌다. 본 연구는 기업의 재무제표를 통한 재무비율을 변수로 활용하여 예측 모형을 만들었다. 하지만 관리종목을 지정하는 다양한 사유인 불성실공시, 자본잠식, 감사의견 등에 대한 비재무적 요인에 대해서는 반영을 하지 않았기에 한계점이 분명히 존재한다. 이후 관련 연구는 머신러닝 중 앙상블 기법을 적용하여 비재무적 요인들을 반영할 수 있는 변수들을 추가하여 더 정확한 관리종목 예측 모델을 개발할 필요가 있다. 또한 관리종목 지정이 머신러닝 기법에 따라 관리종목 지정 예측 확률인 0과 1사이를 범위로 갖는 확률값으로 산출되며, 0.5를 기준으로 관리종목과 비관리종목으로 구분한다는 점은 본 연구에서의 한계점으로 남는다. 따라서 후속 연구는 코스닥 시장 내 관리종목과 비관리종목의 비율을 확인하고 불균형 데이터에 대해서 관리종목과 비관리종목을 나누는 기준값을 고려할 필요가 있다. 예를 들어 코로나19 환경에서 바이오 테크 벤처의 과감한 투자가 관리종목 지정 가능성을 높일 경우 이에 대한 정책적 판단이 개입될 소지가 커 결과적으로 관리종목 지정으로 이어지지 않을 가능성이 있다. 이는 모형과 현실 사이에 괴리를 만들어 머신러닝의 예측 유용성을 낮출 수 있다. 따라서 전체 의사결정 측면에서 볼 때 재무적 데이터를 기반으로 한 알고리즘의 실사용 가능성을 높이기 위한 추가 연구가 필요하다. 최근 기술특례상장하는 기업의 산업군을 살펴보자면 대부분 바이오, 인공지능, 메타버스, IT 산업군에서 특례상장을 허용하고 있다. 이러한 기술특례상장기업은 다소 현재의 수익과 재무 현황이 다소 미흡하더라도 기업의 기술성과 성장성을 고려하는 정성적 평가를 반영한다. 따라서 재무적 요인을 통해 관리종목 지정 예측을 하는 본문에서 제시한 모형의 예측력이 바이오, 인공지능, 메타버스, IT 산업군에서는 다소 적용되지 않을 수도 있음을 인지해야 한다. 마지막으로, 기업이 관리종목으로 지정되는 것을 막기 위해 영업이익을 조정한 기업을 색출해 내지 못한다는 것에 한계점을 가지고 있다. K-IFRS가 개정되어 재무제표에 반드시 표시해야할 항목만을 규정하고 공시의무가 추가되었지만 기타항목들은 기업들만의 규정에 맡기고 있다. 예를 들어, K-GAAP에 비해 이연법인세자산에 있어 각 기업의 자율적인 부분이 커짐에 따라 이연법인세자산을 이용한 이익조정이 발생할 수 있다. 이러한 점을 이용해 관리종목으로 지정되는 것을 회피하는 기업이 존재할 것이며 본 연구에서는 그러한 기업들까지 고려하지 못했다는 한계점을 지니고 있다.

벤처창업에 관한 연구에서 초기 벤처의 생존에 영향을 줄 수 있는 다양한 측면의 이론들이 개발되었다. 창업자의 특성이나 산업 환경과 법률과 같은 다양한 외적 요인들이 초기 벤처의 성공에 영향을 줄 수 있으며 벤처 투자자의 판단 역시 중요하다. 벤처 투자 측면에서 보면 초기 벤처의 부실화 가능성을 조기 점검할 수 있는 도구를 갖추는 일은 투자의 위험도를 줄일 수 있다는 장점이 있다. 이러한 점에서 데이터에 기반한 의사결정에 도움이 되는 머신러닝 알고리즘들의 타당성을 관리종목 지정 맥락에서 고려한 연구는 벤처창업연구에서 나름의 정당성을 갖춘다. 후속 연구를 통해 초기 벤처의 부실화 진단 모형과 이를 통제할 수 있는 정책적 방법을 비롯하여 데이터 기반의 의사결정 방법의 유용성을 검증하는 연구 노력이 이어지기를 기대한다.

REFERENCE

권기현·곽재우·조문기·김종대(2012). 관리종목지정이 감사시간 및 감사보수에 미치는 영향. *세무회계연구*, 32, 23-45.

김민철(2004). 관리종목 지정사유 별 주가수익률의 변화. *세무회계연구*, 14, 229-245.

김수정·문보영(2018). 관리종목 기업에 대한 외부 감사인 지정이 이익조정에 미치는 영향. *회계정보연구*, 36(2), 1-24.

김승열(2010). 코스닥시장의 상장폐지실질심사에 관한 연구. *법학논총*, 22(2), 9-58.

김인상·인창열·이명곤(2016). 관리종목 편입이 감사보고시차에 미치는 영향. *글로벌경영학회지*, 13(1), 257-279.

김인호·이경섭(2020). 트리 기반 앙상블 방법을 활용한 자동 평가 모형 개발 및 평가: 서울특별시 주거용 아파트를 사례로. *한국데이터정보과학회지*, 31(2), 375-389.

김일(2005). 관리종목으로 지정된 기업의 재무적 특성에 관한 연구. *산학경영연구*, 18(2), 179-196.

김태혁·엄철준(1997). 관리대상종목의 수익률과 위험 속성에 관한 연구. *증권 금융연구*, 3(1), 93-133.

김형준·류두진·조훈(2019). 기업부도예측과 기계학습. *金融工學硏究*, 18(3), 131-152.

남규민(2018). *데이터마이닝 기법을 이용한 기업부실 예측모형의 성과 비교*. 석사학위 논문, 부산대학교.

남기정·이동명·진로(2019). 비재무정보를 이용한 창업기업의 부실요인에 관한 실증 연구. *벤처창업연구*, 14(1), 139-149.

류예린·안상봉·지상현(2020). K-IFRS 도입에 따른 이연법인세자산의 재량적 인식을 통한 이익조정 연구. *국제회계연구*, 92, 183-207.

문종건·황보윤(2014). 횡령·배임 및 최대주주변경을 고려한 부실기업예측모형 연구. *벤처창업연구*, 9(1), 119-132.

박종성(2012). 관리종목 지정을 회피하기 위한 코스닥 기업의 이익조정. *경영컨설팅연구*, 12(3), 103-126.

박창래·서영미(2015). 유가증권시장에서 관리대상으로 지정된 기업의 재무적 특성. *회계와정책연구*, 20(6), 173-192.

방소남·제혜금(2020). 중국상장기업의 부실예측모형에 관한 실증연구. *재무와회계정보저널*, 20(4), 137-157.

손성규·염지인(2013). 코스닥시장에서의 상장폐지위험과 이익조정. *회계학연구*, 38(4), 1-30.

손성규·오명전(2008). 관리종목 기업의 회계정보 효과. *연세경영연구*, 45(2), 127-146.

신동인·곽기영(2018). KOSDAQ 시장의 관리종목 지정 탐지 모형 개발. *지능정보연구*, 24(3), 157-176.

신찬휴(2021). 카카오 키즈 D사의 쇠퇴 및 관리종목 지정 회피 사례. *한국경영학회*, 25(1), 105-134.

엄하늘·김재성·최상옥(2020). 머신러닝 기반 기업부도위험 예측모델 검증 및 정책적 제언: 스태킹 앙상블 모델을 통한 개선을 중심으로. *지능정보연구*, 26(2), 105-129.

유한별·탁근주·문정승(2021). 한국 지방소멸 요인과 극복 방안에 관한 연구: 머신러닝 방법을 통한 탐색. *지방정부연구*, 24(4), 443-476.

이현미·전교석·장정아(2020). LightGBM 알고리즘을 활용한 고속도로 교통사고심각도 예측모델 구축. *한국전자통신학회 논문지*, 15(6), 1123-1130.

전병욱·강지수·정경용(2021). 도로교통 이머징 리스크 탐지를 위한 AutoML 과 CNN 기반 소프트 보팅 앙상블 분류 모델. *융합정보논문지 (구 중소기업융합학회논문지)*, 11(7), 14-20.

조경인·김영민(2021). 통계적 학습을 이용한 다시점 기업부도 예측 모형들의 비교. *한국데이터정보과학회지*, 32(3), 487-499.

조재영·주지환·한인구(2021). 기계학습을 이용한 수출신용보증 사고 예측. *지능정보연구*, 27(1), 83-102.

표영인·김일(2002). 관리종목지정 시점에 발생하는 산업내 정보전도효과. *경영학연구*, 31(3), 751-767.

Alaka, H. A., Oyedele, L. O., Owolabi, H. A., Ajayi, S. O., Bilal, M., & Akinade, O. O.(2016). Methodological approach of construction business failure prediction studies: a review. *Construction Management and Economics*, 34(11), 808-842.

Altman, E. I.(1968). Financial ratios, discriminant analysis and the prediction of corporate bankruptcy. *The journal of finance*, 23(4), 589-609.

Barboza, F., Kimura, H., & Altman, E.(2017). Machine learning models and bankruptcy prediction. *Expert Systems with Applications*, 83, 405-417.

Beaver, W. H.(1966). Financial ratios as predictors of failure. *Journal of Accounting Research*, 4, 71-111.

Breiman, L.(2001). Random forests. *Machine learning*, 45(1), 5-32.

Campbell, J. Y., Hilscher, J., & Szilagyi, J.(2008). In search of distress risk. *Journal of Finance*, 63(6), 2899-2939.

Chava, S. & Jarrow, R. A.(2004). Bankruptcy Prediction with industry effects. *Review of Finance*, 8(4), 537-569.

Cho, J. Y., Joo, J. W., & Han, I. G.(2021). The prediction of export credit guarantee accident using machine learning. *Journal of Intelligence and Information Systems*, 27(1), 83-102.

Cho, K. I., & Kim, Y. M.(2021). Comparison of bankruptcy prediction models using statistical learning at multiple times. *Journal of the Korean Data And Information Science Society*, 32(3), 487-499.

DataScience.(2020). Gradient boosting-what you need to know, *Data Science.* Retrived from https://datascience.eu/machine-learning/gradient-boosting-what-you-need-to-know.

Devi, S. S., & Radhika, Y.(2018). A survey on machine learning and statistical techniques in bankruptcy prediction. *International Journal of Machine Learning and Computing*, 8(2), 133-139.

Duffie, D., Saita, L., & Wang, K.(2007). Multi-period corporate default prediction with stochastic covariates. Journal of financial economics, 83(3), 635-665.

Eom, H. N., Kim, J. S., & Choi, S. O.(2020). Machine learning-based corporate default risk prediction model verification and policy recommendation: Focusing on improvement through stacking ensemble model. *Journal of Intelligence and Information Systems*, 26(2), 105-129.

James, G., Witten, D., Hastie, T., & Tibshirani, R.(2013). A*n introduction to statistical learning in R.* New York: Springer.

Jeon, B. U., Kang, J. S., & Chung, K. Y.(2021). AutoML and CNN-based soft-voting ensemble classification model for road traffic emerging risk detection. *Journal of Convergence for Information Technology*, 11(7), 14-20.

Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., Y. Qiwei, & Liu, T. Y.(2017). LightGBM: A highly efficient gradient boosting decision tree. *Advances in*

Neural Information Processing Systems, 30, 3146-3154.

Kim, H. J., Ryu, D. J., & Cho, H.(2019). Corporate default predictions and machine learning. T*he Korean Journal of Financial Engineering*, 18(3), 131-152.

Kim, I. H., & Lee, K. S.(2020). Tree based ensemble model for developing and evaluating automated valuation models: The case of Seoul residential apartment.*Journal of the Korean Data And Information Science Society*, 31(2), 375-389.

Kim, I. S., In, C. Y., & Lee, M. G.(2016). The effect of administrative issues on the audit report lag. *Academic Society of Global Business Administration,* 13(1), 257-279.

Kim, I.(2005). Financial characteristics and disignating firms subject to administrative issues. *Korean Business Review,* 18(2), 179-196.

Kwon, K. H., Kwak, J. W., Cho, M. K., & Kim, J. D.(2012). The Effect of designation as issues for administration on audit hours and audit fees. *Tax Accounting Research*, 32, 23-45.

Kim, M. C.(2004). Characteristics analysis on the stock return of issues for administration. *Tax Accounting Review*, 14, 229-245.

Kim, S. J., & Moon, B. Y.(2018). The effect of designated auditor upon the earnings management issue of administrated firms. *Korea Accounting Information Association*, 36(2), 1-24.

Kim, S. Y.(2010). A legal study on Substantial Investigation of Delisting. *Kookmin Law Review*, 22(2), 9-58.

Kim, T. H., & Eom, C. J.(1997) Rate of return and risk factor of issues for administration. *The Journal of Finance and Banking*, 3(1), 93-133.

Lee, H. M., Jeon, G. S., & Jang, J. A.(2020). Predicting of the severity of car traffic accidents on a highway using light gradient boosting model. *The Journal of the Korea institute of electronic communication sciences*, 15(6), 1123-1130.

Martinez, I., & Serve, S.(2017). Reasons for delisting and consequences: A literature review and research agenda. *Journal of Economic Surveys*, 31(3), 733-770.

Moon, J. G., & Hwangbo, Y.(2014). An empirical study on a firm's fail prediction model by considering whether there are embezzlement, malpractice and the largest shareholder changes or not. *Asia-Pacific Journal of Business Venturing and Entrepreneurship*, 9(1), 119-132.

Nam, G. J., Lee, D. M., & Chen, L.(2019). An empirical study on the failure factors of startups using non-financial information. *Asia-Pacific Journal of Business Venturing and Entrepreneurship*, 14(1), 139-149.

Nam, K. Y.(2018). *A Performance Comparison of Bankruptcy Prediction Model using Data Mining Tools and Techniques.* Master's Thesis, Pusan National University, Korea

Ohlson, J. A.(1980). Financial ratios and the probabilistic prediction of bankruptcy. *Journal of accounting research*, 109-131.

Pang, S. N., & Zhu, H. Q.(2020). Empirical research on financial distress forecast model of Chinese listed companies. *Journal Finance and Accounting Accountiong Information*, 20(4), 137-157.

Park, C. R., & Seo, Y. M.(2015). Financial characteristics of the designated companies of issues for administration' in KOSPI market. *Korean Journal of Accounting Research*, 20(6), 173-192.

Park, J. S.(2012). KOSDAQ Firm's earnings management using classification shifting. *Korean Management Consulting Review*, 12(3), 103-126.

Pyo, Y. I., & Kim, I.(2002). Intra-industry information transfer at the time of administrative issues. *Korean Management Review*, 31(3), 751-767.

Ryu, Y. R., An, S. B., & Ji, S. H.(2020). A study on the earnings management using the discretionary recognition of deferred corporate tax assets due to K-IFRS adoption. *Korean International Accounting Review*, 92, 183-207.

Shin, C. H.(2021). Case study on performance decline of one of Kakao kids and avoidance of designation as administrative issue. *Korea Business Review*, 25(1), 105-134.

Shin, D. I., & Kwahk, K. Y.(2018). Development of a detection model for the companies designated as administrative issue in KOSDAQ market. *Journal of Intelligence and Information Systems*, 24(3), 157-176.

Shumway, T.(2001). Forecasting bankruptcy more accurately: a simple hazard model. *Journal of Business*, 74(1), 101-124.

Soh, S. K. & Yum, J. I.(2013). Delisting risk in the KOSDAQ market and earnings management. *Korean Accounting Review,* 38(4), 1-30.

Sohn, S. K., & Oh, M. J.(2008). Accounting informativeness of administrative issues. *Yonsei Business Review*, 45(2), 127-146.

Yoo, H. B., Tak, K. J., & Mun, J. S.(2021). A Study on the factors and overcoming methods of extinction of provinces in Korea: the exploration with machine learning methods. *The Korean Journal of Local Government Studies*, 24(4), 443-476.

Zmijewski, M. E.(1984). Methodological issues related to the estimation of financial distress prediction models. *Journal of Accounting Research*, 22, 59-82.

Study on Predicting the Designation of Administrative Issue in the KOSDAQ Market Based on Machine Learning Based on Financial Data*

Yoon, Yanghyun**
Kim, Taekyung***
Kim, Suyeong****

Abstract

This paper investigates machine learning models for predicting the designation of administrative issues in the KOSDAQ market through various techniques. When a company in the Korean stock market is designated as administrative issue, the market recognizes the event itself as negative information, causing losses to the company and investors. The purpose of this study is to evaluate alternative methods for developing a artificial intelligence service to examine a possibility to the designation of administrative issues early through the financial ratio of companies and to help investors manage portfolio risks. In this study, the independent variables used 21 financial ratios representing profitability, stability, activity, and growth. From 2011 to 2020, when K-IFRS was applied, financial data of companies in administrative issues and non-administrative issues stocks are sampled. Logistic regression analysis, decision tree, support vector machine, random forest, and LightGBM are used to predict the designation of administrative issues. According to the results of analysis, LightGBM with 82.73% classification accuracy is the best prediction model, and the prediction model with the lowest classification accuracy is a decision tree with 71.94% accuracy. As a result of checking the top three variables of the importance of variables in the decision tree-based learning model, the financial variables common in each model are ROE(Net profit) and Capital stock turnover ratio, which are relatively important variables in designating administrative issues. In general, it is confirmed that the learning model using the ensemble had higher predictive performance than the single learning model.

KeyWords: KOSDAQ, Administrative Issue, Machine-learning, Ensemble

* This work was also supported by the 2020 research fund of Kwangwoon University (2020-0323), and this work was supported by the Ministry of Education of the Republic of Korea and National Research Foundation of Korea (NRF-2019S1A3A2098438).
** First author, a four-year undergraduate, Kwangwoon University, eb3434@naver.com
*** Corresponding author, Associate Professor, Kwangwoon University, kimtk@kw.ac.kr
**** Coauthor, a four-year undergraduate, Kwangwoon University, sooyoung6262@naver.com