세무와회계저널
제20권 제6호 2019년 12월 pp.9～34
한국세무학회

https://doi.org/10.35850/KJTA.20.6.01
ISSN 1738－3323

회계이익예측을 위한 기계학습 성과 비교

정우준*

〈요 약〉

빅데이터와 기계학습 또는 인공지능의 대두로 4차 산업혁명이 촉발될 것으로 예견되고 있다. 이런 맥락에서 빅데이터는 원유로, 기계학습 등은 이 원유를 가공하는 기술로 비유되기도 하며, 의사결정에 유용한 정보를 추출하는 과학적 방법론 또는 프로세스를 지칭하는 데이터 과학이라는 개념도 등장하였다. 이와 같은 흐름에 대응하고자 회계학계에서도 빅데이터와 인공지능 등의 도입에 고심하고 있으나 아직 많은 연구가 이루어지지는 않은 것으로 파악된다. 특히, 국내에서는 이와 관련된 연구는 찾아보기 어려운 것이 현실이다.

회계이익이 기업가치를 결정하는 근본 변수로써 회계정보이용자의 의사결정에 차이를 가져오는 정보력을 가진다는 연구는 충분히 이루어졌으나 예측 성능이 좋은 회계이익예측 모형의 개발과 관련된 연구는 많이 이루어지지 않은 것으로 판단된다. 본 연구에서는 기존에 개발된 회계이익예측 모형을 기계학습 맥락에서 재해석하고, 추가로 보다 높은 예측 성능을 나타낸다고 알려져 있는 예측 모형들과 그 성과를 비교하여 회계이익예측에 기계학습 기법의 도입 가능성을 검토하였다.

본 연구의 목적을 달성하기 위해 제조업에 속하는 KOSPI 상장 기업 중 12월 결산 법인에 대한 재무비율을 추출하여 활용하였다. 로지스틱 모형에 기반한 Ou and Penman(1989)의 연구 내용을 바탕으로 이를 기계학습 기법으로 재해석한 모형을 기준으로 하고, 기계학습에 가장 일반적으로 사용되고 있는 나무 모형, 랜덤포레스트 모형, 부스팅 기법을 추가하여 이들의 예측성과를 비교하였다. 연구의 결과, 모형들간 예측성과 차이가 전반적으로 통계적 유의성을 가지며, 가장 높은 예측 성능을 나타낸 부스팅 기법을 이용한 경우 가장 낮은 예측력을 보이는 나무 모형에 비하여 약 10% 정도 더 높은 예측력을 갖는 것으로 나타났다.

이와 같은 결과는 보다 정확한 회계이익예측 정보를 기계학습 기법을 활용하여 제공할 수 있다는 의의가 있으며, 회계이익예측 이외의 다양한 분야에서도 기계학습을 활용한 연구의 토대가 되기를 기대한다.

〈주제어〉 회계이익예측, 기계학습, 예측 성능

* 홍익대학교 대학원 경영학과 박사과정 수료, blu2ego@gmail.com

논문접수일 2019년 9월 26일　　1차수정일 2019년 11월 7일　　게재확정일 2019년 11월 29일

Ⅰ. 서 론

빅데이터(big data)와 기계학습(machine learning) 및 인공지능(artificial intelligence)의 대두로 4차 산업혁명이 촉발될 것으로 많은 사람들이 예견하고 있다. 이런 맥락에서 빅데이터는 원유로 그리고 기계학습 등은 이 원유를 가공하는 기술로 비유되기도 하고, 이 둘을 응용하여 의사결정에 유용한 정보를 추출하는 과학적 방법론 또는 과정(process) 등을 지칭하는 데이터 과학(data science)이라는 개념도 등장하였다. 아직 이들 핵심 키워드가 학계에서 체계적으로 정의된 것은 아니지만, 현업에서는 이에 대응하기 위한 경영 전반의 전격적인 변화에 많은 노력을 기울이고 있다.

회계학계 내에서도 빅데이터와 인공지능 등의 4차 산업혁명 흐름에 대응하고자 고심하고 있으나 아직 많은 연구가 이루어지지는 않은 것으로 파악된다. 특히, 국내에서는 기계학습 관점에서 이루어진 연구를 찾아보기 쉽지 않은데, 베이지안망(bayesian network)을 응용한 선은정과 박성진(2017) 그리고 이건창과 최관(2006)의 연구 정도를 꼽을 수 있다.

회계이익(accounting earnings)은 기업가치를 결정하는 근본 변수로서 회계정보이용자의 의사결정에 중요한 영향을 미친다. 예를 들어, 기업이 투자나 대출을 받기 위해서는 필수적으로 회계이익 관련 정보를 제출해야 하며, 경영자의 이익 예측 관련 공시는 자본시장에 매우 큰 영향을 미치기도 한다. 이런 회계이익의 유용성과 관련하여, 회계이익이 기업가치를 설명하거나 의사결정에 차이를 가져오는 정보력을 가진다는 연구는 충분히 이루어졌으나, 예측 성능이 좋은 회계이익예측 모형의 개발과 관련된 연구는 많이 이루어지지 않은 것으로 보인다. 이와 관련하여, 김정교(1989)는 주식가치 평가와 회계이익 정보의 유용성 판단, 그리고 대체적 이익개념의 목적적합성의 확보 등을 위해 우수한 이익예측 모형의 개발이 필요하다고 지적하고 있기도 하다.

본 연구에서는 인공지능 등의 개념을 회계학 연구에 도입하기 위해 회계이익예측 관점에서 기본분석(fundamental analysis) 관련 연구를 검토하고, 회계이익예측 모형 개발에 기계학습의 도입 가능성에 대한 근거를 제시해 보고자 한다. 특히 Ou and Penman(1989)에서 사용된 주당순이익 증가/감소 예측 문제를 기계학습 관점에서 재해석 한다. 로지스틱회귀 모형을 사용하여 회계이익을 예측한 Ou and Penman(1989)은 기계학습 맥락에서 이항분류(binary classification) 문제로 해석할 수 있으며, 이 문제는 기계학습의 주된 연구 주제이다.

본 연구의 구성은 다음과 같다. 먼저 제Ⅱ장에서는 회계이익예측 기법과 관련된 선행연구들을 살펴보고, 제Ⅲ장에서는 본 연구에서 사용된 기계학습 기법인 로지스틱회귀 모형, 나무 모형, 랜덤포레스트 모형, 그리고 부스팅 모형을 소개하고 추가적으로 리샘플링 기법과 예측력의 평가 방법을 제시한다. 제Ⅳ장에서는 목표변수와 예측변수의 구성과 모형 간 성능 비교를 통해 실험의 결과를 제시하며, 제Ⅴ장에서는 연구의 결과를 요약하고 향후 연구 방향에 대해 제언한다.

Ⅱ. 선행연구의 검토

회계이익예측을 위한 방법 중 시계열 예측 기법과 Ou and Penman(1989)이 제안한 기본분석(fundamental analysis)이 대표적이라고 할 수 있다. 시계열 예측 기법을 이용하여 회계이익을 예측한 연구들은 주로 회계이익의 연간 또는 분기 시계열 속성을 설명하는데 중점을 두고 있다(김성민과 민춘식, 2010). 김성민과 민춘식(2010)은 연간이익과 분기이익 시계열 관련 연구를 정리하고 회계이익예측에 시계열 기법을 적용하기 위해서는 분기이익이나 반기이익을 이용해야 한다고 주장하였다. 한편, Ou and Penman(1989)은 재무제표상의 여러 기본변수들에 통계적 기법을 이용하여 미래의 이익 증감 성향을 나타내는 확률을 로지스틱 함수(logistic function)를 이용하여 추정하고, 이러한 확률을 바탕으로 기본변수가 미래이익 및 주식수익률에 대한 정보력을 가진다고 보고하였다. 이와 같은 Ou and Penman(1989)을 비롯한 기본분석 연구들은 재무제표의 기본변수들(fundamental variables)이 미래성과에 대하여 예측력이 있는가를 분석하였으며, 많은 연구들에서 기본변수들이 미래이익과 미래 주식가치에 대한 정보력을 가진다는 결과를 제시하고 있다(나종길과 신희정, 2014). Ou and Penman(1989) 이후에도 Lev and Thiagarajan(1993), Abarbanell and Bushee(1997), Frankel and Lee(1998), Dechow et al.(1999), Nissim and Penman (2001) 등이 대표적으로 기본분석의 의의를 밝힌 연구라고 할 수 있다.

기본분석 연구는 독립변수와 종속변수가 이론적으로 구조화되지 않았다는 지적이 있기도 하였는데, Ohlson(1995)의 잔여이익모형(Residual Income Model)을 기반으로 한 Penman and Zhang (2006) 이후로 재무비율과 기업가치 간의 체계적 관계가 어느 정도 확보되었다. Wieland(2011)은 Penman and Zhang(2006)과 Anderson et al.(2003)의 연구를 결합하여 6개의 기본변수를 선별하고, 재무분석가 예측값을 이용한 이익예측력과 기본변수를 이용한 경우를 비교하였는데, 기본변수를 이용한 경우가 더 뛰어난 성과를 보임을 보고하였다(김태희, 나종길, 그리고 양지위, 2018).

Ishibashi et al.(2016)의 경우, 데이터 마이닝 기법으로 주당순이익 증감 예측 모형을 구축하기 위해, 예측 모형의 예측력에 대한 여러 가지 변수선택 기법의 영향력에 대해 연구하였다. 예측 모형 구축을 위해 로지스틱 모형을 사용하였으며, 예측변수는 Ou and Penman(1989) 등에서 정의된 65개의 재무비율 변수를 사용하였다. 변수선택(variable selection) 기법으로는 단계선택법(stepwise)를 기본으로 Relief, CFS, CNS, C4.5 기법을 적용한 경우의 성과를 비교하였다. 연구의 결과, CNS와 C4.5 가장 효과적인 것으로 나타났다. 그런데, Ishibashi et al.(2016)이 사용한 변수선택(variable selection) 기법은 피쳐엔지니어링(feature engineering)[1] 절차 중 하나로서, 기계학

1) 피쳐엔지니어링(feature engineering)이란 입력(독립변수)을 기계학습 모형에 적합하게 하거나 예측 모형의 성능을 높일 수 있도록 수치적인 표현으로 변형하거나 추출하는 것을 말한다. 여기에는 데이터의 각 값을 빈도화, 바이너리 변환, 구간화, 로그 변환, 스케일링, 표준화 등 다양한 기법이 사용된다. 피쳐엔지니어링의 예를 들면, 목표변수가 y이고 예측변수가 x라고 할 때, x^2이나 $\log(x)$ 등으로 변형하는 경우를

습 맥락의 연구에서 주로 고려되는 모형의 높은 예측력과 강건성(robustness)을 확보하는 데 도움이 되지만, 구축된 모형의 해석력을 떨어뜨리는 주요 원인이 되기도 한다.

이들 연구들을 다시 해석하면, 기본분석으로 회계이익을 예측하는 경우 미래 회계이익에 대한 예측력이 높은 변수를 발견하는 것이 중요한데, 변수 선정에 치중한 나머지 예측 성능이 높은 모형을 도입하지 못하고 로지스틱회귀 모형이나 선형회귀 모형을 사용하는 수준에 그치고 있다고 볼 수 있다. 한편, 예측력을 평가할 때 주가수익률을 기준으로 할 경우 개발된 예측 모형의 예측 성능을 직접적으로 평가한다고 보기 어려운 한계도 있다. 이러한 문제들은 기본분석과 관련된 연구의 실증적 연구 결과가 예측(prediction) 정확도 보다는 인과관계 추론(inference)을 중심으로 하기 때문으로 판단된다. 본 연구는 모형의 해석력을 감소시킬 수 있는 절차를 제외하고, 기계학습 관점의 예측 방법론을 제시하며, 이러한 방법론에 근거한 분석 결과를 바탕으로 회계이익예측성과를 직접적으로 비교하여 제시함으로써 회계학 연구에 기계학습 기법을 도입하는 데 근거를 마련하고자 한다.

Ⅲ. 기계학습의 도입

1. 기계학습의 목표－손실함수의 최소화

기계학습(machine learning)에서 목표변수(target variable) y_i를 예측하기 위해 학습 데이터 x_i를 사용하는 것을 지도학습(supervised learning)이라 하며, y_i의 특성에 따라 회귀(regression)나 분류(classification) 등 여러 문제를 정의할 수 있다. 지도학습 기법 중 분류분석(classification analysis)이란 범주형 변수(categorical variable)를 목표변수로 하여 예측하는 기법으로서, 특히 목표변수 y_i가 성공－실패의 오로지 두 개의 값만을 갖는 경우를 이항분류(binary classification) 문제라고 한다.

지도학습에서 모형은 흔히 예측변수가 주어졌을 때, 목표변수에 대한 예측을 위한 수학적 구조를 나타낸다. 기본적인 모형으로는 선형 모형(linear model)이 사용되는데, 이 선형 모형에 의한 예측은 다음의 (식 1)과 같이 예측변수 또는 입력(input)[2]의 가중된 선형결합에 의해 이루어진다.

들 수 있다.

2) 고전적 통계학의 맥락에서는 주로 독립변수(independent variable)라고 하지만, 기계학습 맥락에서는 주로 입력(input)이라고 한다.

$$\hat{y_i} = \sum_{j=1}^{k} w_j x_{i,j} \quad \text{(식 1)}$$

(식 1)에서 예측된 변수 $\widehat{y_i}$는 목적에 따라 다른 해석을 가질 수 있다. 예를 들어, 로지스틱회귀(logistic regression)에서는 원하는 클래스(범주형 변수의 출현값. 예를 들어 성별이라하면 0 또는 1)의 출현 확률을 얻기 위해 로지스틱 변환을 할 수 있고, 출력에 순위를 부여하고 싶을 때 순위 점수로써 사용할 수도 있다. 여기에서 모수(parameter)는 사전에 결정되지 않고 데이터로부터 학습하는 것으로서 선형회귀(linear regression) 문제에서의 모수는 계수 w_j이다. 학습 데이터가 주어졌을 때 가장 좋은 모수를 찾는 방법이 필요하며, 이를 위해서 특정 모수 집합이 주어졌을 때 모형의 성능을 측정하기 위한 목적 함수를 (식 2)와 같이 정의할 수 있다.

$$Obj(\theta) = L(\theta) + \Omega(\theta) \quad \text{(식 2)}$$

L은 학습에 따른 손실함수(loss function)이고, Ω는 정규화항(regulation term)인데, 손실함수는 예측 모형의 오류를 정량화한다. 여기에서 학습(learning)이란 손실함수를 최소화하는 모수를 찾는 일이라고도 할 수 있는데, 회귀 문제에서 사용되는 기본적인 손실함수는 (식 3)과 같이 평균제곱오차(MSE: mean squared error)이며, 분류 문제에 일반적으로 사용되는 손실함수는 (식 4)와 같이 로지스틱 손실함수이다.

$$L(\theta) = \frac{1}{n-1}\sum_{i=1}^{n}(y_i - \widehat{y_i})^2 \quad \text{(식 3)}$$

$$L(\theta) = \sum_{i=1}^{n}[(y_i \ln(1+e^{-\widehat{y_i}}) + (1-y_i)\ln(1+e^{\widehat{y_i}})] \quad \text{(식 4)}$$

손실함수가 (식 3)과 같은 MSE일 때에는 이를 최소화하기 위해 최소제곱법을 사용하는 것이 적절하나 (식 4)와 같은 로지스틱 손실함수를 최소화하는 데에는 최대가능도(maximum likelihood)를 이용하는 방법이 보다 적절하다. 즉, 로지스틱 함수를 사용하여 어떤 사건이 나타날 확률이 관찰된 사건에 최대한 가깝도록 β_0와 β_1에 대한 추정값 $\hat{\beta}_0$와 $\hat{\beta}_1$을 구하고, 이 추정값을 로지스틱 함수 $p(X)$에 삽입하면 각 사건에 대한 확률(0과 1사이의 값으로)을 얻을 수 있다. 이를 가능도 함수(likelihood function)로 표현하면 다음과 같다.

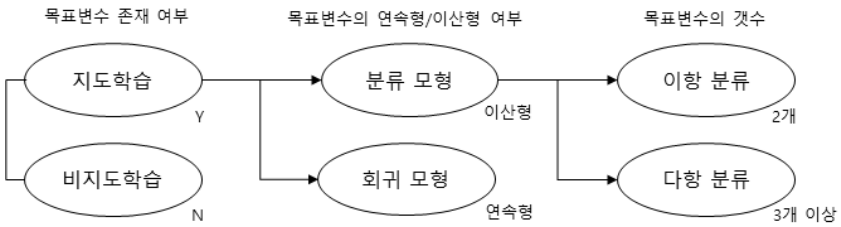
$$L(\beta_0, \beta_1) = \prod_{i:\, y_i = 1} p(x_i) \prod_{i':\, y_{i'} = 0} (1 - p(x_{i'})) \quad \text{(식 5)}$$

그런데 (식 5)와 같은 가능도 함수는 비선형 함수이므로 이를 최대화하는 최대가능도추정값은 수치적 방법(numerical method)을 이용하여 구해야 한다. 그리고 이러한 가능도 함수를 최대화하는 것은 음의 로그 가능도(negative log likelihood) 최소화와 동치관계이며, 이를 최소화하는 수치적 알고리즘으로는 Netwon－Raphson 기법(Nelder－Mead 기법, BFGS 기법, L－BFGS－B 기법, Brent 기법 등도 유용하게 사용할 수 있다)이 대표적이다(허명회, 2018).

2. 기계학습 기법의 종류

기계학습 기법은 <그림 1>과 같이 목표변수의 존재 유무, 그리고 목표변수가 존재하는 경우 그 성질에 따라 몇 가지로 구분할 수 있다. 목표변수가 없는 기계학습 기법은 비지도학습(unsupervised learning)이라고 하며, K－means 기법 등을 사용하는 군집분석(clustering)이 대표적이다. 목표변수가 존재하는 경우, 기계학습 모형은 목표변수가 연속형인지 이산형인지에 따라 분류모형과 회귀모형으로 구분하며, 분류모형인 경우 이항분류(binary－class classification) 모형과 다항분류(multi－class 또는 multinomial classification) 모형으로 구분할 수 있다. 기계학습 분야의 이항분류 문제에서 가장 많이 사용되는 모형은 로지스틱회귀 모형, 나무 모형, 랜덤포레스트 모형, 부스팅 모형 등이 있으며, 이의 특징은 다음과 같다.

〈그림 1〉 기계학습 기법의 구분



가. 로지스틱회귀 모형

로지스틱회귀 모형은 여러 연구들에서 많이 사용되지만, 기계학습 분야에서도 흔히 사용되는 모형이다. 기본적으로 로지스틱회귀 모형은 양/불량 또는 성공/실패와 같이 1 또는 0의 입력을 갖는 목표변수 Y가 있을 경우, Y가 특정 범주에 포함될 확률을 예측하는 것으로, 예를 들어 다음과 같다.

$$\Pr(negative = true|test) \quad \text{(식 6)}$$

$p(X)=\Pr(Y=1|X)$와 X 사이의 관계를 모형화하기 위해서는 모든 X 값에 대하여 0과 1 사이의 값을 반환하는 함수를 이용해야 한다. 이러한 성질을 갖는 함수가 많이 있지만, 로지스틱 회귀에서는 다음과 같은 로지스틱 함수(logistic function)를 사용한다.

$$p(X)=\frac{1}{1+e^{-X}}=\frac{e^{X}}{1+e^{X}}$$
$$=\frac{e^{\beta_0+\beta_1 X}}{1+e^{\beta_0+\beta_1 X}} \quad \text{(식 7)}$$

위 수식을 정리하면 다음의 식을 얻을 수 있는데, 이를 오즈(또는 승산 : odds)라고 하며 0과 무한대(∞)의 값을 갖는다.[3)]

$$\frac{p(X)}{1-p(X)}=e^{\beta_0+\beta_1 X} \quad \text{(식 8)}$$

위의 식의 양변에 로그를 취하면 다음과 같다.

$$\log\left(\frac{p(X)}{1-p(X)}\right)=\beta_0+\beta_1 X \quad \text{(식 9)}$$

위의 수식에 따르면, 로지스틱회귀 모형의 계수는 X가 한 단위 증가함에 따라 로그 오즈의 변화분을 나타낸다는 것을 알 수 있다. 그러나 여기에서 $p(X)$와 X는 선형 관계가 아니므로, X가 한 단위 증가함에 따른 $p(X)$의 변화량이 β_1과 선형적으로 대응되지 않는다는 점에 주의해야 한다.

나. 나무 모형

기계학습에서는 최종모형의 예측력을 중요시하는 경향이 있지만, 예측력과 해석력을 모두 고려하는 것이 중요하다. 예를 들어, 기대집단에서 가장 많은 반응을 보일 방안을 예측하고자 하는 경우에는 예측력에 치중하며, 신용평가 같은 경우에서는 해석력이 더 중요할 수 있다. 여기에서 다루는 나무 모형은 다른 지도학습 기법들에 비해 예측력은 떨어지나 해석력이 좋은 것으로 알려져 있다(박창이 등, 2015).

3) 이런 오즈값은 예측확률을 과대하게 평가한다. 예를 들어, 5명 중 1명이 채무를 불이행한다면 그 확률은 0.2일 것이지만, 오즈값은 1/4이 되기 때문이다.

나무 모형은 예측변수 $X_1, ..., X_p$ 로 이루어진 p차원 공간을 재귀적으로(recursively) M개의 영역 $R_1, ..., R_M$ 으로 분할하고, 각 x가 속한 R_j 영역에서 설명변수 x_i가 속한 모든 관측치 y_i 들의 평균인 상수값 c_m 으로 y_i를 예측한다.

$$f(x) = \sum_{m=1}^{M} c_m I(x \in R_m) \quad (\text{단, } c_M = avg\{y_i : x_i \in R_M\}) \qquad (\text{식 } 10)$$

이러한 나무 모형은 목표변수가 연속형인 회귀나무(regression tree)와 범주형인 분류나무(classification tree)로 나눌 수 있으나, 이를 적용하는 절차는 성장(growing), 가지치기(pruning), 타당성 평가, 해석 및 예측 과정으로 동일하다. 출력변수(output variable)가 이항변수(bianry variable)인 이항나무 모형은 지니 지수(Gini index)나 엔트로피 지수(entropy index)를 불순도(impurity)의 지표로 하여 가능한 리프(leaf)[4]가 순전히 성공-실패 관측값으로만 이루어지도록 순도(putiry)를 높이는 방향으로 나무를 성장시킨다. 이러한 나무 모형은 수식적 틀로부터 자유로운 기계학습의 전형으로서, 나무 모형들의 집합(앙상블 모형)인 랜덤포레스트 모형의 기본이 된다(허명회, 2018).

나무 모형은 모형이 불안정적이라는 단점이 있는데, 그 이유는 첫 번째 노드에서 분리변수를 찾을 때 서로 비슷한 예측력을 갖는 변수가 다수 존재하기 때문이다. 따라서 데이터의 작은 변화에도 첫 번째 노드의 분리변수가 바뀔 수 있다. 첫 번째 분리변수가 바뀌면 자식노드에 포함되는 자료가 완전히 바뀌게 되고 따라서 최종 결과가 완전히 달라지게 된다. 학습 방법의 불안정성은 예측력의 저하를 가져오며 이와 동시에 예측 모형의 해석력도 떨어뜨린다.

다. 랜덤포레스트 모형

Breiman(2001)에 의해서 개발된 랜덤포레스트(random forest) 모형은 나무 모형이 불안정적이라는 단점을 고려한 것으로서, 약한 학습기(weak learner)[5]를 생성한 후 이를 결합하여 최종 학습기를 만드는 앙상블(ensemble) 기법[6] 중 하나이다. 랜덤포레스트는 이론적 설명이나 최종 결과에 대한 해석은 어렵지만 예측력은 매우 높은 방법으로 알려져 있는데, 특히 입력변수의 개수가 많을 경우 좋은 예측력을 보이는 경우가 많다.

이러한 랜덤포레스트 모형은 다음의 절차에 따라 구성된다. 즉, 훈련자료 $L = \{y_i, x_i\}_{i=1}^{n}$ (단,

4) 리프(leaf)란 자식을 가지지 않는 가장 하위 노드를 말한다.

5) 학습기(learner)란 완성된 모형과 유사한 개념으로 분류 문제 경우 분류기(classifier)로 표현한다.

6) 앙상블(ensemble) 기법이란 기계학습에서 여러 개의 모형을 조합하여 단일 모형보다 더 높은 예측 성능을 보이는 모형을 만드는 방법을 말한다.

$x_i \in R^p$)에 대하여 n개의 데이터를 이용한 부트스트랩 표본[7] $L^* = \{y^*, x^*\}_{i=1}^n$을 생성하고, L^*의 입력변수들 중 $k(k \ll p)$ 개만 무작위로 뽑아 나무 모형을 구축한다. 그리고 이렇게 구축한 나무 모형들을 결합하여 최종 학습기를 만든다. 여기에서 부트스트랩 표본을 몇 개나 생성할 것인지, k 값을 어떻게 정할 것인지, 선형 결합의 형식을 어떻게 할 것인지에 대한 선택이 다수 존재하는데, 이는 경험에 의존해야 한다.

라. 부스팅 모형

앙상블 기법 중 하나인 부스팅(boosting)은 주로 나무 모형을 사용하여 약한 학습기를 순차적으로(sequentially) 생성하여 최종 모형을 만드는 방법을 말한다. 최초의 부스팅 알고리즘은 이항 분류 문제에서 Freund and Schapire(1997)에 의해서 개발된 AdaBoost(Adaptive Boost) 알고리즘인데, 이의 내용은 다음과 같다(박창이 등, 2015).

1단계 : 가중치 $w_i = 1/n$(단, $i = 1, ..., n$)를 초기화

2단계 : $m = 1, ..., M$ 에 대하여 다음 과정을 반복

2−1 : 가중치 w_i를 이용하여 분류기 $f_m(x) \in \{-1, 1\}$를 적합(fitting)

2−2 : err_m를 다음과 같이 계산

$$err_m = \frac{\sum_{i=1}^{n} w_i I(y_i \neq fm(x_i))}{\sum_{i=1}^{n} w}$$

2−3 : $c_m = \log((1 - err_m)/err_m)$ 으로 설정

2−4 : 가중치 w_i를 $w_i = w_i \exp(c_m I(y_i \neq f_m(x_i)))$ 로 갱신(updating)

3단계 : 2단계에서 얻은 M개의 분류기를 결합하여 최종 분류기 $sign(\sum_{m=1}^{M} c_m f_m(x))$를 얻는다.

AdaBoost 알고리즘의 본래 목적은 훈련오차(training error)를 가능한 빨리 그리고 쉽게 줄이는 것인데, 약한 학습기 f_m의 오분류율이 일정 조건을 만족하면 훈련오차가 지수적으로 빠르게 0으로 수렴함이 증명되었다(Freund and Schapire, 1997). 또한 AdaBoost 알고리즘이 제안된 이후 본래의 목적과는 상관없이 예측 오차도 감소한다는 것이 경험적으로 증명되었다. 이후 AdaBoost 알고리즘의 통계학적 해석에 대하여 많은 연구가 이루어졌는데, Friedman(2001)에서는 AdaBoost가 단계별전진선택(stagewise forward selection)을 이용한 변수선택 방법이라는 것을 밝혀냈고,

7) 부트스트랩 표본은 주어진 데이터에서 복원추출을 허용한 재표본(re-sample)을 말한다.

축소추정(shrinkage estimation)을 통해서 과적합(over-fitting)[8] 문제를 피할 수 있다는 것을 보였다. Efron et al.(2004)은 축소추정을 포함한 부스팅 알고리즘은 LASSO(Least Absolute Shrinkage and Selection Operator)[9] 추정값과 동일함을 보이기도 하였다. 이러한 연구 결과들을 통해 부스팅 알고리즘의 작동원리가 대부분 규명된 것으로 인정되고 있으며, 높은 예측 성능을 보이는 경우가 많아 선호되는 기계학습 기법 중 하나이다.

3. 리샘플링 기법의 이용

리샘플링 기법(resampling method)은 예측 모형을 훈련 데이터에 한 번만 적용하여서는 얻을 수 없는 정보를 얻기 위한 방법으로 대표적으로는 교차검증(cross validation) 접근법이 있다.[10] 교차검증은 기계학습 모형의 성능을 평가하거나 모형의 적절한 유연성(flexibility)을 통제하기 위해 주어진 기계학습 모형의 실험오차(test error)를 추정하는 데 사용된다(James, et al., 2014). 실험오차(test error)는 기계학습 모형을 적용하여 실험 데이터세트의 목표변수를 예측하면서 얻은 결과의 평균적인 오차를 말하는 것으로 모형의 유연성에 따라 U자 형태를 가지지만, 직접 관찰할 수는 없다. 이와 달리, 훈련오차는 훈련 데이터세트를 이용하여 얻을 수 있는 것으로 이 둘은 구분되어야 한다. 기계학습 모형의 훈련(training), 검증(validation), 실험(test) 과정을 위한 데이터세트 구성 방법은 다음의 <그림 2>와 같다.

<그림 2>에서 보는 것과 같이 주어진 데이터세트가 있을 때 기계학습 기법을 적용하기 위해서 그 데이터를 훈련(training) 데이터, 검증(validation) 데이터, 실험(test) 데이터의 세 가지 종류로 구분해야 한다. 훈련 데이터는 기계학습 모형을 적합(fitting)하기 위한 데이터이고, 검증 데이터는 훈련 데이터를 통해 적합된 모형의 타당성을 평가하기 위해 사용된다. 마지막으로 실험 데이터는 새로운 자료의 유입이 있을 경우, 이에 대한 평가를 위한 것으로 사용된다. 특히, 실험 데이터는 데이터를 세 가지로 분류하고 난 뒤 훈련이나 검증 단계에 절대로 포함되어서는 안되는데, 그럴 경우 과적합(over-fitting)이 일어나 적절한 모형 평가를 할 수 없게 된다. 이렇게 데이터를 분할하는 방법은 크게 두 가지가 있다. 먼저 교차검증(cross validation) 기법이라고 하는 것으로, 데이터를 랜덤 추출하는 방식으로 일정 비율로 1회 구분하는 방법으로, 절차가 단순하며 비용 효율적이라는 장점이 있어 가장 일반적으로 사용되는 기법이다. 다른 방법으로는, 실험

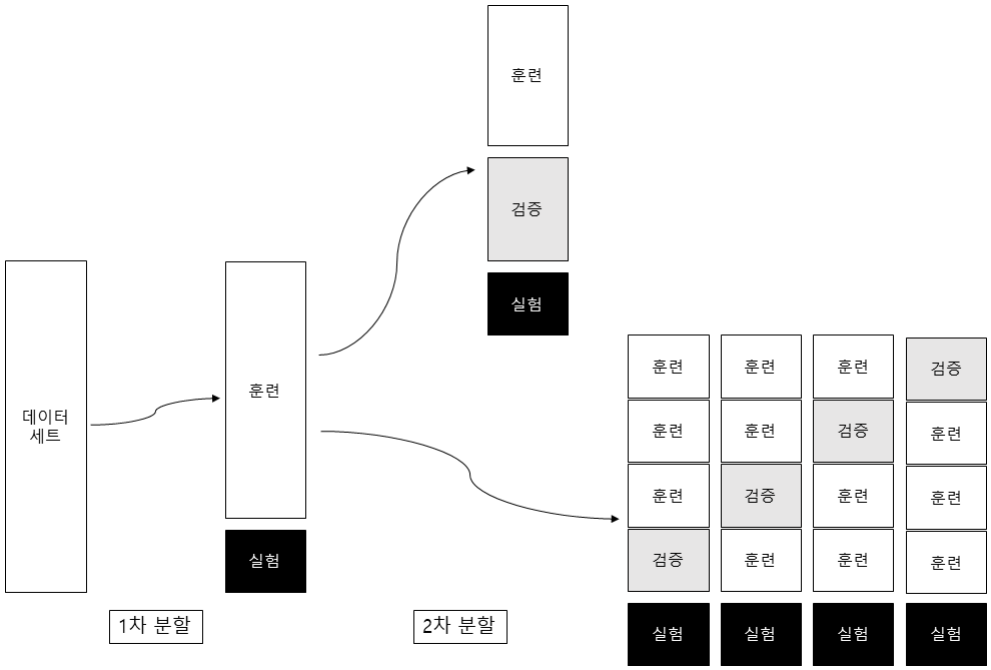
8) 과적합(over-fitting)이란 훈련 데이터에서 모형의 학습이 지나치게 잘 되어서 훈련 데이터에서는 높은 정확도를 나타내지만, 검증 데이터나 실험 데이터에서는 성능이 떨어지는 현상을 말한다. 이를 해결하는 방법으로는 더 많은 데이터를 이용하거나, 입력의 개수를 줄이거나, 또는 정규화(regularization) 기법을 이용하는 방법 등이 있다.

9) 기존의 선형회귀 모형이 MSE를 최소화하는 모수를 찾는데 반해, LASSO는 이 가정에 모수 추정치들의 절대값들의 합이 최소가 되도록 하는 모형을 말한다.

10) 다른 대표적인 방법으로는 부트스트랩(bootstrap) 접근법이 있다.

데이터와 나머지 데이터를 랜덤하게 2 : 8이나 3 : 7 등의 일정 비율로 구분한 후 그 나머지 데이터를 k개로 다시 재분할하여 이 재분할 된 데이터 집합을 순서대로 돌아가며 훈련과 검증 단계에 사용하는 것이다. 이를 k−fold 교차검증이라고 한다. k−fold 교차검증은 보다 강건한 모형을 만드는 데 도움을 주지만, 계산에 비용이 많이 따른다는 단점이 있다.

〈그림 2〉 기계학습에서 데이터세트의 구성 방법



4. 예측성과의 평가

Ou and Penman(1989)에서는 주당순이익 증감에 대한 예측된 확률을 바탕으로 가상의 투자를 통해 초과수익을 달성할 수 있는지 여부로 기본변수의 정보력을 제시하고 있다. 그러나 이러한 접근법은 기업의 시장가치에 영향을 미치는 다른 요인들을 모두 통제하지 못하는 경우, 예측 성능을 직접적으로 측정한다고 할 수 없다. 반면, 기계학습 기법을 활용한 예측에서는 기본적으로 혼돈행렬(confusion matrix)을 바탕으로 여러 성과측정 지표나 ROC 곡선 또는 AUC(area under curve)를 이용하여 직접적으로 예측 성능을 측정한다.

〈표 1〉 혼돈행렬의 구성

	$\hat{y_i}=1$	$\hat{y_i}=0$
$y_i=1$	True Positive(TP)	False Negative(FN)
$y_i=0$	False Positive(FP)	True Negative(TN)

기계학습에서 혼돈행렬은 관측값과 예측값의 관계를 나타내는 행렬로서 행에는 관측값을 그리고 열에는 예측값으로 구성하고, 이렇게 구성된 행렬의 각 값은 해당 경우를 만족하는 수를 나타낸다. 이항분류 문제에서는 <표 1>과 같이 2×2 행렬로 나타낼 수 있다. 혼돈행렬을 이용하여 기계학습 기법의 예측성과를 측정하기 위해 다음과 같은 지표를 사용할 수 있다. 다만, 정밀도와 재현율은 상호보완적으로 사용할 수 있으며, 두 지표가 모두 높을수록 좋은 예측 모형이라고 할 수 있지만, 이 둘은 상충관계에 있기 때문에 동시에 이 둘의 값을 높이기는 어렵다.

① 정밀도(precision) : 모형이 True라고 분류한 것 중에서 실제 True인 것의 비율로 $TP/(TP+FP)$로 계산

② 재현율(recall, sensitivity, hit rate) : 정밀도와 True Positive의 경우를 다른 시각에서 측정하는 것으로 실제 True인 것 중에서 True라고 예측한 결과의 비율을 말하며, 혼돈행렬의 $TP/(TP+FN)$로 계산

③ 정확도(accuracy) : 혼돈행렬에서 $(TP+TN)/(TP+FN+FP+TN)$로 계산. 이는 False를 False로 예측한 경우의 비율로써, 정확도는 가장 직관적으로 모형의 성능을 나타낼 수 있는 평가 지표

대부분의 분류 문제에서 분류기(classifier)는 0 또는 1의 값을 반환하는 것이 아니라 [0, 1] 구간 사이의 확률값을 계산한다. 이와 같이 예측 결과가 확률값으로 주어질 때, 분계점(threshold 또는 cutoff)을 바꿔가며 최종 예측값을 구한다. ROC 곡선(receiver operating characteristic curve)은 이처럼 분계점 변화에 따른 재현율과 FPR(False Positive Rate)의 관계를 도식화한 것이다. 그런데 ROC 곡선은 그래프이기 때문에 명확한 수치로 비교하기 어려우므로, 그래프 아래의 면적인 AUC(area under curve)를 구하여 이를 직접 비교한다. AUC의 최대값은 1이며, 좋은 모형일수록 1에 가까운 값을 가진다.

Ⅳ. 데이터의 구성과 연구의 결과

1. 데이터의 구성

가. 목표변수의 구성

본 연구에서 사용한 데이터는 상장협 데이터베이스인 TS2000에서 제조업에 속하는 KOSPI 상장 기업 중 12월 결산 법인의 2009년부터 2018까지의 데이터를 활용하였다.[11] 목표변수인 주당순이익 증감 변수는 Ou and Penman(1989)에서 정의된 것과 동일하게 다음의 (식 11)로 정의한다.

$$\mathrm{Pr}_{i,t+1} = \begin{cases} 0 \; (eps_{t,t+1} - eps_{i,t} - drift_{i,t+1} < 0) \\ 1 \; (eps_{t,t+1} - eps_{i,t} - drift_{i,t+1} > 0) \end{cases} \quad \text{(식 11)}$$

$\mathrm{Pr}_{i,t+1}$: i 기업의 $t+1$ 시점 미래 이익 증가/감소 성향
$eps_{i,t}$, $eps_{i,t+1}$: i 기업의 t 시점과 $t+1$ 시점 주당순이익
$drift_{i,t+1}$: i 기업의 t 시점 기준 과거 4년간 주당순이익 평균

<표 2>는 본 연구에서 훈련/검증 데이터세트와 실험 데이터세트에서 목표변수로 사용된 연도별 주당순이익 증감 여부에 대한 기업 분포를 보여주고 있다. <표 2>에서 주당순이익 증가로 구분된 기업은 Ou and Penman(1989)의 정의에서와 같이 당기와 직전 연도말 주당순이익에서 과거 4개년간의 주당순이익 변화분 평균을 차감한 값이 0보다 큰 기업이며, 주당순이익 감소 기업은 이와 반대인 경우를 나타낸다.

〈표 2〉 주당순이익 증감 분포

연도	주당순이익 증가		주당순이익 감소		표본 수
	기업 수(개)	비율(%)	기업 수(개)	비율(%)	
2018	285	39.5	386	60.5	671
2017	304	45.6	363	54.4	667
2016	356	53.6	308	46.4	664
2015	360	54.8	297	45.2	657
2014	372	57.4	276	42.6	648
2013	330	51.5	311	48.5	641
합계	2,007	50.8	1,941	49.2	3,948

11) 2008년 금융위기의 여파로 기업들의 재무제표에 체계적 차이가 있을 수 있어, 그 이후 기간인 2009년부터 연구대상 기간에 포함하였다.

여기에서 2018년의 데이터는 교차검증 과정에서 실험 데이터세트로 활용하며, 2013년부터 2017년까지의 데이터는 훈련 데이터와 검증 데이터로 활용된다. 그런데 2013년부터 2016년까지 주당순이익 증가 기업이 전체 표본 기업의 최소 51.5%에서 최대 57.4%로 50% 이상을 차지하다가 2017년 이후로 45% 수준으로 떨어지며, 실험 과정에 사용되는 2018년의 경우 40% 이하로 하락하는데, 이런 비율의 변화는 예측력 감소의 원인이 될 수 있다. 이와 같이 모집단 분포가 변하는 경우에도 강건한 성질을 가지는 예측 모형을 개발하기 위해서는 교차검증 등의 리샘플링 기법이 적절히 활용되어야 함을 암시한다.

나. 예측변수의 구성

투입되는 변수의 선정에 있어 이론적 근거를 기반으로 하는 기존의 연구 방식과는 달리 예측에 초점을 두고 있는 기계학습 분야에서는 데이터 주도 의사결정(data driven decision) 접근법에 기초하여 모형을 구축한다. 즉, 활용가능한 가능한 많은 종류와 양의 데이터를 기반으로 실험을 통해 모형을 구축하는 것이다. 이러한 접근법은 최근 빅데이터(big data) 흐름과도 일치한다. 이와 같은 접근법에 따라 본 연구에서는 기존의 연구에서 가설화되는 재무지표를 선택적으로 분석하는 것과는 달리 사용가능한 모든 재무지표를 예측 모형 구축에 사용한다.

사용가능한 재무비율을 포함한 지표의 범위는 상장협의 TS2000 데이터베이스에서 추출 가능한 모든 재무지표로 하였다. TS2000에서 추출한 재무지표의 수는 총 166개로서 이의 구성은 <표 3>과 같다. 이 중에서 2007년도 이전 발생분에 해당하는 것과 중복되는 것 등을 제외하고 실제로 예측에 사용된 비율(지표)는 152개이며, 각 재무지표의 특징은 다음과 같다.

① 성장성 비율 : 총자본증가율, 순이익증가율, 매출액증가율 등이 대표적이다. 성장성에 대한 지표는 기업이 어느 정도 성장하였는지 경영실적을 바탕으로 판단하기 위해 사용할 수 있는데, 일정 수준의 달성은 계속기업 가정을 유지하는 데에도 중요하다.

② 수익성 비율 : 총자본이익률, 자기자본이익률, 납입자본이익률, 매출액순이익률, 매출액영업이익률 등이 있다. 자본을 투입하여 얼마의 잉여가치를 창출하는지를 평가하기 위해 사용되는 지표로서 성장성 비율과 상호 보완적인 성격을 갖는다.

③ 안정성 비율 : 유동비율, 부채비율, 고정비율 등을 포함한다. 안정성 비율은 기업의 채무상환 능력과 경기 변동 대처 능력을 평가하는 데 사용할 수 있다.

④ 활동성 비율 : 총자산회전율, 고정자산회전율, 재고자산회전율 등이 있다. 활동성 비율은 기업자산의 활용정도를 알아보고자 할 때 활용할 수 있는 지표로써 손익계산서의 매출액을 재무상태표에 있는 각 자산의 항목으로 나누어 계산한다.

⑤ 생산성 비율 : 종업원 1인당 부가가치나 설비투자효율, 그리고 자본분배율 등을 포함한다. 생산성 비율은 기업 성과의 효율적 달성 여부를 측정 또는 평가하고, 발생원인과 성과배분의 합리성을 규명하는데 사용할 수 있다. 즉, 생산성 비율은 경영합리화의 지료로 볼 수

있으며, 최근에는 부가가치 생산성에 의해 기업 경영의 성과를 측정하는 것이 일반화 되어 가고 있다(정호웅, 1992).

⑥ 부가가치 지표 : 부가가치, 총자본투자효율, 그리고 기계투자효율 등을 포함한다. 부가가치란 특정 제품의 생산 단계에서 새로이 창출된 가치를 의미하는 것으로서, 부가가치 지표는 부가가치의 구성 요소를 측정하기 위해 사용할 수 있으며, 이는 생산성을 측정하는 다른 방식이 될 수 있다.

⑦ 투자지표 : PER(Price Earnings Ratio), PBR(Price Book-value Ratio), PSR(Price Sales Ratio) 등을 포함한다. 주요 성과 지표의 회계상 장부가치와 기업 시장가치의 비율로 계산하며, 장부상 파악되지 않는 시장기대의 대용치로 활용할 수 있다.

⑧ EBITDA 지표 : EBITDA(백만원), EBITDA/매출액(%), EBITDA/금융비용(배) 등을 포함하며, 기업의 영업활동을 통한 현금흐름 창출 능력을 나타내는 지표이다.

〈표 3〉 TS2000에서 추출 가능한 재무지표의 구성

구분	재무비율	개수
성장성 비율	총자본증가율, 유형자산증가율, 유동자산증가율, 영업이익증가율, 경상이익증가율(2007년 이전 발생), 순이익증가율, 재고자산증가율, 자기자본증가율, 매출액증가율, 종업원1인당 부가가치증가율, 종업원수증가율, 비유동자산증가율, 종업원1인당 매출액증가율, 종업원1인당 인건비증가율	14
수익성 비율	매출액총이익률, 매출액영업이익률, 매출액경상이익률(2007년 이전 발생), 매출액순이익률, 총자본사업이익률, 총자본영업이익률, 총자본경상이익률(2007년 이전 발생), 총자본순이익률, 자기자본영업이익률, 자기자본경상이익률(2007년 이전 발생), 자기자본순이익률, 자본금영업이익률, 자본금경상이익률(2007년 이전 발생), 자본금순이익률, 조세 대 조세차감전순이익률, 기업경상이익률(2007년 이전 발생), 투자수익률(2007년 이전 발생), 기업순이익률, 경영자본영업이익률, 경영자본순이익률, 매출원가 대 매출액비율, 영업비율, 영업외손익률, 금융비용부담률, 외환이익 대 매출액비율, 광고선전비 대 매출액비율, 수지비율, 인건비 대 총비용비율, 조세공과 대 총비용비율, 금융비용 대 총비용비율, 감가상각비 대 총비용비율, 감가상각률, 누적감가상각률, 이자부담률, 지급이자율, 차입금평균이자율, 사내유보율, 사내유보 대 자기자본비율, 적립금비율(재정비율), 평균배당률, 자기자본배당률, 배당성향, 1주당매출액(원), EPS(Earning Per Share)(원), 1주당경상이익(2007년 이전 발생)(원), CPS(Cash flow Per Share)(원), BPS(Book-value Per Share)(원), 유보율, R&D 투자효율, 1주당영업이익(원)	50

〈표 3〉 TS2000에서 추출 가능한 재무지표의 구성 (계속)

구분	재무비율	개수
안정성 비율	유동자산구성비율, 재고자산 대 유동자산비율, 유동자산 대 비유동자산비율, 당좌자산구성비율, 비유동자산구성비율, 자기자본구성비율, 타인자본구성비율, 자기자본배율, 비유동비율, 비유동장기적합률, 유동비율, 당좌비율, 현금비율, 원재료비율, 매출채권비율, 재고자산 대 순운전자본비율, 매출채권 대 매입채무비율, 매출채권 대 상·제품비율, 매입채무 대 재고자산비율, 부채비율, 유동부채비율, 단기차입금 대 총차입금비율, 비유동부채비율, 비유동부채 대 순운전자본비율, 순운전자본비율, 차입금의존도, 차입금비율, 이자보상배율(이자비용), 이자보상배율(순금융비용), 유보액대비율, 유보액 대 납입자본배율, 유동자산집중도, 비유동자산집중도, 투자집중도, CASH FLOW 대 부채비율, CASH FLOW 대 차입금비율, CASH FLOW 대 총자본비율, CASH FLOW 대 매출액비율, 영업이익대비이자보상배율	39
활동성 비율	총자본회전률, 경영자본회전률, 자기자본회전률, 자본금회전률, 타인자본회전률, 매입채무회전률, 매입채무회전기간, 유동자산회전률, 당좌자산회전률, 재고자산회전률, 재고자산회전기간, 상품, 제품회전률, 원·부재료회전률, 재공품회전률, 매출채권회전률, 매출채권회전기간, 비유동자산회전률, 유형자산회전율, 순운전자본회전률, 운전자본회전률, 1회전기간	21
생산성 비율	부가가치(백만원), 종업원1인당 부가가치(백만원), 종업원1인당 매출액(백만원), 종업원1인당 경상이익 (2007년 이전 발생)(백만원), 종업원1인당 순이익(백만원), 종업원1인당 인건비(백만원), 노동장비율, 기계장비율, 자본집약도, 총자본투자효율, 설비투자효율, 기계투자효율, 부가가치율, 노동소득분배율, 자본분배율, 이윤분배율	16
부가가치 지표	부가가치(백만원), 법인세비용차감전(계속사업)손익(백만원), 인건비(백만원), 금융비용(백만원), 임차료(백만원), 조세공과(백만원), 감가상각비(백만원), 종업원1인당 부가가치(백만원), 총자본투자효율, 기계투자효율, 부가가치율, 종업원수	12
투자 지표	PER(Price earnings ratio)(최고), PER(Price earnings ratio)(최저), PBR(Price book－value ratio)(최고), PBR(Price book－value ratio)(최저), PCR(Price cash－flow ratio)(최고), PCR(Price cash－flow ratio)(최저), PSR(Price sales ratio)(최고), PSR(Price sales ratio)(최저)	8
EBITDA 지표	기업가치(EV)(백만원), EBITDA(백만원), EBITDA/매출액(%), EBITDA/금융비용(배), EBITDA/평균발행주식수(백만원), EV/EBITDA(배)	6

① 2007년 이전 발생분 : 경상이익증가율 (2007년 이전 발생), 매출액경상이익률 (2007년 이전 발생), 총자본경상이익률 (2007년 이전 발생), 자기자본경상이익률 (2007년 이전 발생), 자본금경상이익률 (2007년 이전 발생), 기업경상이익률 (2007년 이전 발생), 투자수익률 (2007년 이전 발생), 1주당경상이익 (2007년 이전 발생)(원), 종업원1인당 경상이익 (2007년 이전 발생)(백만원)

② 중복된 변수 : 부가가치(백만원), 총자본투자효율, 기계투자효율, 부가가치율

③ 값이 누락되어 있는 변수 : 원재료비율

2. 연구의 결과

Ou and Penman(1989)의 연구는 기계학습 관점에서 이항분류 문제로 해석할 수 있다.[12] 하지만, Ou and Penman(1989)에서 회계이익의 증감을 예측하기 위해 사용된 로지스틱회귀 모형은 해석력이 상대적으로 높은 편이기는 하지만, 예측력은 그다지 높은 편이 아니라고 알려져 있다. 본 연구에서는 Ou and Penman(1989)에서 사용된 로지스틱회귀 모형을 기계학습 맥락에서 주당순이익의 증감 예측으로 재평가하고 기계학습 분야에서 많이 사용되고 있는 나무 모형, 랜덤포레스트 모형, 부스팅 모형을 추가하여 이들 모형의 예측성과를 교차검증 방법을 통해 서로 비교하였다. 즉, 2013년부터 2017년까지의 데이터를 7 : 3의 비율로 랜덤하게 추출하여 여러 이익예측 모형을 훈련 및 검증하고, 2018년 데이터를 실험 데이터세트로 하여 개발된 모형의 성능을 AUC 기준으로 서로 비교한다.

<표 4>는 목표변수인 주당순이익 증감변수와 예측변수 간의 상관관계를 분석한 것으로, 총 152개의 예측변수 중 양(+)의 상관관계를 갖는 예측변수 중 상관계수 상위 10개 변수와 음(−)의 상관관계를 갖는 예측변수 중 상관계수 상위 10개 변수를 나타낸 것이다.

〈표 4〉 주당순이익 증감 변수와 예측변수의 상관관계

변수명	상관계수	t 값	자유도	p 값
영업비율	0.064	3.674	3274	0.000
수지비율	0.059	3.380	3274	0.000
재고자산 대 유동비율	0.036	2.057	3274	0.000
EV/EBITDA	0.033	1.915	3274	0.000
종업원1인당 인건비	0.033	1.883	3274	0.000
차입금의존도	0.032	1.829	3274	0.000
R&D 투자효율	0.032	1.821	3274	0.000
유동비율	0.031	1.798	3274	0.000
당좌비율	0.031	1.784	3274	0.000
금융비용	0.030	1.724	3274	0.000
1주당영업이익	−0.081	−4.675	3274	0.000
자기자본순이익률	−0.114	−6.587	3274	0.000

12) 오로지 두 개의 값을 가진 목표변수를 분류하는 문제를 이항분류(binary-class classification)라고 하며, 여러 개의 값을 가진 목표변수를 분류하는 다항분류(multi-class classification)라고 한다.

〈표 4〉 주당순이익 증감 변수와 예측변수의 상관관계 (계속)

변수명	상관계수	t 값	자유도	p 값
자기자본영업이익률	−0.119	−6.877	3274	0.000
자기자본증가율	−0.122	−7.047	3274	0.000
총자본투자효율	−0.150	−8.664	3274	0.000
총자본영업이익률	−0.153	−8.876	3274	0.000
총자본사업이익률	−0.156	−9.030	3274	0.000
주당순이익	−0.163	−9.430	3274	0.000
총자본순이익률	−0.219	−12.817	3274	0.000
기업순이익률	−0.225	−13.238	3274	0.000

① 목표변수인 주당순이익증감 여부 변수와 이를 예측하기 위한 152개 예측변수 간의 상관관계 분석 결과임
② 152개 변수 중 55개의 변수가 양(+)의 상관관계를 가지며, 97개 변수가 음(−)의 상관관계를 가지고 있고, 이 중 상관계수 기준으로 상위 10개씩의 변수를 나타냄.
③ 표에서 상관계수, t 값, p 값은 소수점 넷째 자리에서 반올림함.

결과를 보면, 양(+)의 상관관계를 갖는 변수는 영업비율, 수지비율, 재고자산 대 유동비율, EV/EBITDA, 종업원1인당 인건비, 차입금의존도, R&D 투자효율, 유동비율, 당좌비율, 금융비용 순으로 그 크기가 크며, 음(−)의 상관관계를 갖는 예측변수는 크기순으로 기업순이익률, 총자본순이익률, 주당순이익, 총자본사업이익률, 총자본영업이익률, 총자본투자효율, 자기자본증가율, 자기자본영업이익률, 자기자본순이익률, 1주당영업이익의 순이며, 이들 변수들은 모두 통계적으로 유의한($p < 0.000$) 상관관계를 가지는 것으로 나타났다.

<표 5>는 Gevrey et al. (2003)의 변수중요도(variable importance)에 따라 152개의 예측변수 중 각 모형의 예측력을 결정하는 데 가장 큰 영향력을 가진 상위 20개 변수를 나타낸다. 분석 결과를 살펴보면, 기업순이익률, 배당성향, 자기자본증가율, 주당순이익이 네 개 모형 모두에서 공통적으로 중요한 변수로 나타났음을 알 수 있다. 이들 네 개 변수 외에는 중요 변수들이 모형 간 상당히 상이하게 나타남을 알 수 있는데, 가장 예측력이 좋은 부스팅 모형의 경우, 현금흐름대차입금비율, 사내유보대자기가본비율, 기업순이익률, 자본금순이익률, 영업이익증가율, 배당성향 등이 중요도가 큰 것으로 나타났고, 랜덤포레스트 모형의 경우에는 기업순이익률, 자기자본증가율, 자기자본순이익률, 자본금순이익률, 주당순이익 등이 중요한 것으로 나타났다.

〈표 5〉 모형별 변수중요도

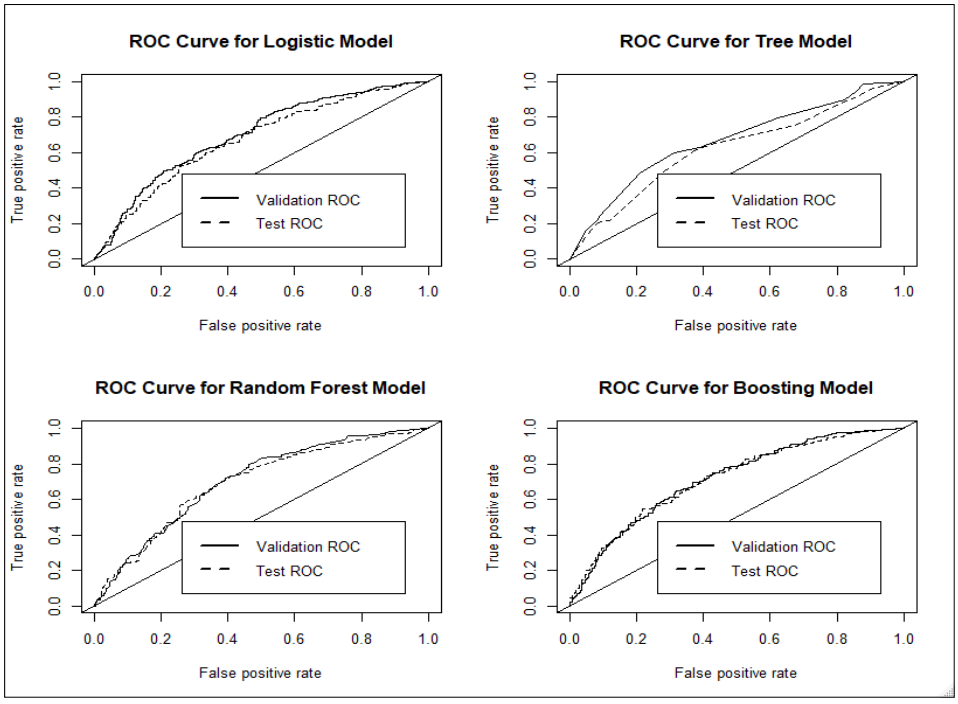
순위	로지스틱 모형		나무 모형		랜덤포레스트 모형		부스팅 모형	
	변수	VI	변수	VI	변수	VI	변수	VI
1	주당순이익	5.012	자본금 순이익률	114.562	기업순이익률	23.406	현금흐름대 차입금비율	1.433
2	유보액대비율	4.803	자기자본 증가율	88.586	자기자본 증가율	22.898	사내유보대 자기자본비율	1.429
3	총자본 영업이익률	4.406	주당순이익	82.516	자기자본 순이익률	19.216	기업순이익률	1.409
4	기업가치	3.973	기업순이익률	75.505	자본금 순이익률	18.689	자본금 순이익률	1.372
5	종업원1인당 매출액	3.630	자기자본 순이익률	74.950	주당순이익	18.302	영업이익 증가율	1.244
6	기업순이익률	3.600	PER(최저)	42.212	총자본 순이익률	16.140	배당성향	1.236
7	총자본 사업이익률	3.514	PER(최고)	32.819	순이익 증가율	15.169	순이익 증가율	1.235
8	종업원수	3.289	배당성향	32.707	경영자본 순이익률	13.495	재고자산 증가율	1.168
9	자본금 회전률	3.015	자기자본배율	26.382	PER(최저)	13.106	매출채권 대 매입채무비율	1.147
10	자본집약도	2.893	기계투자효율	25.567	매출액 순이익률	12.986	자기자본 증가율	1.138
11	자본금 영업이익률	2.862	사내유보 대 자기자본비율	24.982	영업외 손익률	12.244	주당순이익	1.138
12	자기자본 순이익률	2.738	평균배당률	22.706	법인세비용 차감전손익	12.197	기계투자효율	1.124
13	노동장비율	2.630	영업외 손익률	22.671	배당성향	11.986	금융비용	1.102
14	자기자본 증가율	2.496	유동부채비율	20.401	사내유보 대 자기자본비율	11.549	매출채권 대 상,제품비율	1.081
15	기계장비율	2.436	순이익 증가율	18.766	유보율	11.542	기업가치	1.080
16	비유동비율	2.416	총자본 영업이익률	18.664	적립금비율	11.506	종업원1인당 매출액	1.049
17	자기자본 회전률	2.193	유보율	17.960	종업원1인당 부가가치증가율	11.439	단기차입금 대 총차입금비율	1.030
18	비유동자산 집중도	2.109	유보액 대 납입자본배율	17.805	유보액 대 납입자본배율	11.217	PBR(최고)	1.029
19	종업원1인당 인건비	2.023	이자보상배율 (이자비용)	16.729	자기자본배율	11.068	EV/EBITDA	0.992
20	배당성향	1.988	이자보상배율 (순금융비용)	15.477	조세 대 조세 차감전순이익률	10.952	조세 대 조세 차감전순이익률	0.982

① 소수점 넷째 자리에서 반올림

② VI : 변수중요도(variable importance)

예측 모형에 따른 예측 성능을 비교하기 위해 각 모형의 ROC 곡선을 구하여 비교하고 이를 <그림 3>처럼 시각화하였다. 각 ROC 곡선에서 실선은 훈련 단계에서의 ROC 곡선이고 점선은 실험 단계의 ROC 곡선인데, 이 ROC 곡선은 좌측 상방으로 곡선이 이동할수록 사용된 모형의 예측 성능이 뛰어남을 나타낸다. 모든 ROC 곡선이 50% 이상의 예측 성능을 보이는 것으로 나타났으나, 나무 모형의 경우 검증 단계와 실험 단계의 ROC 곡선 차가 가장 큰 것으로 나타난다. 즉, 나무 모형의 경우 그 예측 결과를 일반화하여 사용하기에 부적절할 수 있다. 이와는 반대로 부스팅 모형의 ROC 곡선은 검증 단계와 실험 단계의 ROC 곡선이 거의 일치하는 것으로 나타났다.

〈그림 3〉 각 모형의 검증 단계와 실험 단계의 ROC 곡선 비교



<그림 3>에서는 각 모형에 따른 예측성과를 직관적으로 파악할 수 있었으나, 그 성과를 보다 엄밀히 분석하기 위해 <표 6>과 <표 7>에서처럼 검증 단계와 실험 단계에서 각 모형에 따른 AUC를 산출하고, 각 AUC 간 차이를 통계적으로 검증하였다.

〈표 6〉 검증 단계에서 각 모형의 AUC와 모형간 AUC 차이 검정

검증 모형	로지스틱 모형	나무 모형	랜덤포레스트 모형	부스팅 모형
로지스틱 모형	0.698	−0.032*	0.002	0.018
나무 모형		0.666	0.034**	0.050***
랜덤포레스트 모형			0.700	0.016*
부스팅 모형				0.716

① 소수점 넷째 자리에서 반올림
② 우측 하방 대각선의 값은 각 모형의 AUC 값이며, 대각선의 우상방 영역은 모형간 AUC 차이
③ 대각선 우상방 영역의 값의 *, **, ***는 Hanley and McNeil(1983)에 따른 10%, 5%, 1% 수준에서의 통계적 유의성을 나타냄

<표 6>과 <표 7>의 우측 하방 대각선의 값은 각 모형의 AUC 값이며, 대각선의 우측 상방 영역은 두 AUC 차이에 대한 통계적 유의성 검증 결과인데, Hanley and McNeil(1983)이 제시한 동일 데이터세트의 경우 ROC 면적 차이에 대한 양측 검정의 p값을 나타낸다.

분석 결과를 살펴보면, 검증 단계와 실험 단계 모두 부스팅 모형, 랜덤포레스트 모형, 로지스틱 모형, 나무 모형의 순으로 예측력이 높은 것으로 나타났다. Ou and Penman(1989)에서도 사용된 로지스틱 모형의 경우, 기계학습 방식으로 적용한 결과는 검증 단계와 실험 단계에서 각각 69.8%와 67.1%의 예측력을 보이는 것으로 나타났다.

〈표 7〉 실험 단계에서 각 모형의 AUC와 모형간 AUC 차이의 검정

실험 모형	로지스틱 모형	나무 모형	랜덤포레스트 모형	부스팅 모형
로지스틱 모형	0.671	−0.050**	0.025*	0.044***
나무 모형		0.621	0.075***	0.094***
랜덤포레스트 모형			0.696	0.019**
부스팅 모형				0.715

① 소수점 넷째 자리에서 반올림
② 우측 하방 대각선의 값은 각 모형의 AUC 값이며, 대각선의 우상방 영역은 모형간 AUC 차이
③ 대각선 우상방 영역의 값의 *, **, ***는 Hanley and McNeil(1983)에 따른 10%, 5%, 1% 수준에서의 통계적 유의성을 나타냄

기계학습 기법 중 많이 사용되는 나무 모형의 경우, 검증 단계의 AUC는 66.6%이고 실험 단계의 AUC는 62.1%로 로지스틱 모형에 비해서도 예측력이 낮은 것으로 나타났다. 나무 모형의 경우, 다른 모형들과 비교하였을 때 검증 단계와 실험 단계에서의 AUC 차이(4.5%)도 가장 큰 것으로 나타나, 실제 예측에 있어서는 적절한 접근이 되기 어렵다고 할 수 있다. 가장 예측력이

좋은 모형은 부스팅 모형으로 검증 단계의 AUC는 71.6%이고 실험 단계의 AUC는 71.5%이며, 두 단계의 차이(0.1%)도 크지 않음을 알 수 있다. 이는 가장 낮은 예측력을 가진 나무 모형에 비해서 검증 단계에서는 5%의 예측력 상승을 나타내는 것이며, 실험 단계에서는 9.4%나 높은 예측력을 나타낸다. 랜덤포레스트 모형은 두 번째로 높은 예측력을 보이지만 부스팅 모형이 검증 단계나 실험 단계에서 모두 통계적으로 유의하게 높은 성능을 보이고 있음을 알 수 있다.

Ⅴ. 결론 및 한계

경영환경의 변화와 급속한 정보기술의 발달 등으로 회계 정보의 유용성 제고에 대한 필요성을 인식하게 되었고, 회계 정보의 생산, 유통 및 활용 방식에 대한 새로운 접근이 모색되고 있다. 특히, 빅데이터 및 기계학습(인공지능)은 예측 문제에 있어 뛰어난 성과를 보이고 있으며, 이러한 방법을 회계학 연구에 결합하려는 시도가 서서히 이루어지고 있다.

본 연구에서는 회계이익예측 정보의 중요성을 인식하여, 이의 예측성과를 높이기 위해 기본분석을 기계학습 방식으로 재해석하였다. 그리고 예측성과를 측정 및 비교하기 위해서 주가수익률과 연동하는 방식이 아니라 직접적인 성과측정 지표를 사용하였다. 이러한 방식의 성과측정 지표의 활용은 비상장기업의 회계이익을 예측하는 데에도 유용할 수 있다. 기계학습 기법을 활용하는 경우, 그 효용성을 높이기 위해서는 대용량 데이터를 필요로 하므로 사용가능한 재무비율과 지표를 가능한 많이 포함하여 분석하였으며, 각 예측 모형에 따른 중요 변수를 중요도에 따라 20개씩 선별하였다. 연구의 결과, 가장 예측력이 높은 모형은 부스팅 모형으로 예측력이 가장 낮은 나무 모형에 비해서는 약 9.4% 높은 예측력을 나타내었으며, 벤치마크 모형으로 사용된 로지스틱 모형에 비해서는 약 4.4% 높은 예측력을 나타냈다. 다만, 본 연구는 연구 설계와 연구 모형 개발 등에 있어 많은 한계가 있다. 예를 들어, 연속형 목표변수의 예측 모형 개발, 재무분석가 예측값이나 시계열 예측값 등 예측성과 비교를 위한 벤치마크의 다양화, 실험 대상 기간과 시장 구분 또는 상장 여부 등의 구분에 따른 연구 대상의 확대 및 그에 따른 다중 모형의 개발과 성과 비교 등 후속 연구가 많이 필요하다.

본 연구는 실용적이면서 회계학 연구에 논쟁적 이슈를 제기할 수 있다. 기본분석 방법론을 기초로 한 회계이익예측 연구들이 변수들 간 이론적 근거가 부족함을 지적받아 온 것에 더하여, 기계학습 기법들은 모형의 해석력 보다는 예측력에 중점을 두어 대부분 변수들의 비선형적 관계를 파악하기 때문이다. 즉, 기계학습은 예측 모형을 구축하기 위해 투입하는 변수에 대한 제약을 가정하지 않으므로, 이론적/사전적 정의에 의한 변수뿐만 아니라 가용한 모든 데이터를 모형에 투입하여 예측력을 높일 수 있는 가능성을 제시하지만, 이러한 방식으로 구축된 예측력 높은 모형은 해석력이 낮아져, 결국 변수 간 이론적 또는 경제적 관계를 파악하기는 더 어려워진다.

실제로 최근 인공지능 구현의 주된 방법으로 알려진 딥러닝(deep learning) 기법은 블랙박스 모형으로도 일컬어지는데, 그 이유는 모형 해석이 거의 불가능하기 때문이다. 본 연구에서도 이와 같은 점을 고려하여, 예측력의 희생을 감수하고 실제 기계학습 현업에서 이루어지는 피처엔지니어링이나 파리미터최적화 기법 등은 사용하지 않았다. 한편, 일반적으로 기계학습 방식의 예측은 고전적 통계 모형을 기반으로 한 예측보다 자료의 크기가 작을 때에는 예측의 정확성이 떨어지지만 자료가 크고 복잡해질수록 예측력이 증가된다. 빅데이터 등으로 데이터가 충분히 크고 고성능의 컴퓨팅 능력이 지원되어 기계학습 예측이 더 선호되는 시대가 되고 있으나, 기계학습 접근법은 데이터가 충분히 확보되지 않으면 별로 도움이 되지 않는다. 실제로 데이터의 크기가 아주 크지 않은 대부분의 경우에는 고전적 통계 모형을 이용한 예측이 더 정확하고 효율적이다. 게다가 고전적 통계 모형은 설명가능하다는 장점이 있어서 지식으로 발전시키기에 더 유리하다. 그러나 점점 디지털화가 가속되고 있는 현대 경영 환경의 변화에서 정보이용자들의 요구를 충족하기 위해서, 실용적 관점의 연구가 추후 지속적으로 이루어질 필요가 있다. 즉, 예측이 필요한 회계적 사건에서 적시성 있고 유용성 높은 의사결정을 위해 보다 다양한 데이터를 활용한 예측 연구들이 수반되어야 한다. 단기적으로는 변수들 간 이론적 배경을 필요로 하지 않는 시계열 예측에 딥러닝 등의 기법을 적용하는 것이 과제가 될 수 있으며, 장기적으로는 해석 가능한 기계학습(interpretable machine learning)의 개선에 따라 어느 정도는 이견이 좁혀질 것으로 기대된다. 결국, 향후 회계학 연구에서 기계학습(인공지능)의 도입 가능성은 기업 또는 산업의 요구와 예측 분석 기법 간의 적절한 일치 여부에 따라 좌우될 것으로 예상된다.

참고문헌

김성민 · 민춘식. 2010. “VECM을 적용한 경영성과예측을 위한 모형설정”. 회계학연구 제35권 제2호 : 71－100.

김정교. 1989. “우리나라 기업의 연간회계이익의 시계열 속성”. 회계학연구 제9권 제1호 : 71－98.

김태희 · 나종길 · 양지위. 2018. “기본변수의 미래이익예측력과 장기성과 : 코스닥시장 신규상장기업을 대상으로”. 회계정보연구 제36권 제3호 : 1－26.

나종길 · 신희정. 2014. “기본변수모형의 이익예측력 비교 : 구조적 접근과 경험적 접근”. 회계학연구 제39권 제4호 : 131－170.

박창이 · 김용대 · 김진석 · 송종우 · 최호식. 2015.「R을 이용한 데이터 마이닝」. 교우사.

선은정 · 박성진. 2017. “베이지안 네트워크를 활용한 기업의 재무적 특성과 조세회피성향간의 관계”. 세무회계연구 제52권 제0호 : 41－63.

이건창 · 최 관. 2007. “감리지적기업의 분류적 특성에 관한 연구 : 베이지안 망과 C5.0, 그리고 앙상블 방법간의 비교를 중심으로”. 경영학연구 제36권 제3호 : 705－737.

허명회. 2018.「데이터 과학 : 입문과 토픽」. 자유아카데미.

Abarbanell, J. and B. Bushee. 1997. Fundamental Analysis, Future Earnings, and Stock Prices. *Journal of Accounting Research.* 35(1) : 1－24.

Anderson. M., R. Banker and S. Janakiraman. 2003. Are Selling. General. and Administrative Costs “Sticky”? *Journal of Accounting Research* 41 : 47－63.

Breiman. L. 2001. Random Forests. *Machine Learning* 45 : 5－32.

Dechow, P. M., A. P. Hutton, and R. G. Sloan. 1999. An Empirical Assessment of the Residual Income Valuation Model. *Journal of Accounting and Economics.* 26(1－3) : 1－34.

Efron. B., T. Hastie, I. Johnstone. R. Tibshirani. 2004. Least Angle Regression. *The Annals of Statistics* 32(2) : 407－499.

Frankel, R., and C. M. C. Lee. 1998. Accounting Valuation, Market Expectation and Cross－Sectional Stock Returns. *Journal of Accounting and Economics.* 25(3) : 283－319.

Freund, Y., and Schapire, R. E. 1997. A Decision－Theoretic Generalization of On－Line Learning and an Application to Boosting. *Journal of Computer and System Sciences* 55(1) : 119－139.

Friedman, J. H. 2001. Greedy Function Approximation : A Gradient Boosting Machine. *The Annals of Statistics* 29(5) : 1189－1232.

Gevrey, M., Dimopoulos, I., and Lek, S. 2003. Review and comparison of methods to study the contribution of variables in artificial neural network models. *Ecological Modelling*, 160(3) : 249－264.

Hanley, J. A., and B. J. McNeil. 1983. A Method of Comparing the Areas under Receiver Operating Characteristic Curves Derived from the Same Cases. *Radiology* 148(3) : 839 －843.

Ishibashi, K., T. Iwasaki, S. Otomasa, and K. Yada. 2016. Model Selection for Financial Statement Analysis : Variable Selection with Data Mining Technique. *Procedia Computer Science* 96 : 1681－1690.

James, G., D. Witten, T. Hastie, and R. Tibshirani. 2014. *An Introduction to Statistical Learning with Applications in R.* Springer.

Lev, B. and S. R. Thiagarajan. 1993. Fundamental Information Analysis. *Journal of Accounting Research.* 31(2) : 190－215

Nissim, D., and S. H. Penman. 2001. Ratio Analysis and Equity Valuation : From Research to Practice. *Review of Accounting Studies* 6(1) : 109－154.

Ohlson J. A. 1995. Earnings, Book Values, and Dividends in Equity Valuation. *Contemporary Accounting Research* 11(2) : 661－687.

Ou, J. A. and S. H. Penman. 1989. Financial Statement Analysis and the Prediction of Stock Returns. *Journal of Accounting and Economics* 11(4) : 295－330.

Penman. S. H. and X. Zhang. 2006. Modeling Sustainable Earnings and P/E Ratios with Financial Statement Analysis. Working paper. Columbia University and University of California. Berkeley.

Wieland. M. M. 2011. Identifying Consensus Analysts' Earnings Forecasts that Correctly and Incorrectly Predict an Earnings Increase. *Journal of Business Finance and Accounting* 38(5－6) : 574－600.

Comparison of Machine Learning Performance for Earnings Forecasting

Woo June Jung*

<Abstract>

Many predict that the fourth industrial revolution will be triggered by the emergence of big data and machine learning(or artificial intelligence). In this context, big data is often likened to crude oil and machine learning is likened to the crude oil processing technology, and the concept of data science has also emerged, referring to scientific methodologies or processes that extract useful information for decision-making. In order to cope with such a trend, even in the field of accounting, researchers are struggling to introduce big data and artificial intelligence, but there are not many studies yet. In particular, it is difficult to find studies related to them in Korea.

There have been enough studies that accounting earnings have the information contents that makes a difference in accounting information users' decision making as a fundamental variable that determines the firm's value. However, many studies have not been conducted that relate to the development of a predictive accounting earnings forecasting model. In this study, the accounting earnings forecasting model developed in the previous study was reinterpreted in the context of machine learning, and compared its performance with the predictive(machine learning) models known to represent further higher predictive performance to examine the possibility of introducing machine learning techniques in forecasting accounting earnings.

To achieve the objective of this study, all 152 financial ratios for closing corporations in December were extracted and utilized from 2009 to 2018 among KOSPI companies belonging to the manufacturing sector provided by TS2000 of the Korea Listed Companies Association.

This study reinterpret the findings of Ou and Penman(1989) that used the logistic model basically in the context of machine learning, and the predictability of these techniques was compared by adding the most commonly used models such as tree model, random forest model, and boosting method. As a result of this studies, the predictability difference between the models are statistically significant in overall, and the boosting technique with the highest predictive performance has about 10% higher predictive power than tree model that has lowest predictive power.

These results are meaningful in that it is possible to provide more accurate accounting earnings forecast information using machine learning techniques, and it is expected this research to be the basis for the study using machine learning in various accounting research fields other than accounting earnings forecasting.

<Key words> earnings forecasting, machine learning, predict performance

* Ph. D. Candidate, College of Business Administration, Hongik University, blu2ego@gmail.com