비재무 정보를 활용한 IPO 주식의 상장일 가격 등락 예측에 관한 연구

A study on initial price change prediction of IPO shares using non-financial information

저자
(Authors)

신상훈, 이현준, 안재준
Sanghun Shin, Hyun Jun Lee, Jae Joon Ahn

출처
(Source)

한국데이터정보과학회지 29(2), 2018.3, 425-439 (15 pages)
Journal of the Korean Data And Information Science Sociaty 29(2), 2018.3, 425-439 (15 pages)

발행처
(Publisher)

한국데이터정보과학회
The Korean Data and Information Science Society

URL

http://www.dbpia.co.kr/Article/NODE07408713

APA Style

신상훈, 이현준, 안재준 (2018). 비재무 정보를 활용한 IPO 주식의 상장일 가격 등락 예측에 관한 연구. 한국데이터정보과학회지, 29(2), 425-439.

Journal of the Korean Data & Information Science Society
2018, **29**(2), 425–439

http://dx.doi.org/10.7465/jkdi.2018.29.2.425
한국데이터정보과학회지

비재무 정보를 활용한 IPO 주식의 상장일 가격 등락 예측에 관한 연구

신상훈[1] · 이현준[2] · 안재준[3]

[1]연세대학교 투자정보공학 · [2]연세대학교 산업공학과 · [3]연세대학교 정보통계학과

접수 2018년 2월 17일, 수정 2018년 3월 16일, 게재확정 2018년 3월 16일

요 약

공모주는 비상장기업이 유가증권시장 또는 코스닥시장에 상장하기 위해 불특정 다수를 대상으로 판매하는 주식을 의미하며, 이러한 과정을 기업공개라고 한다. 공모주는 판매 가격인 공모가에 대비하여 상장일 종가는 대부분 상승하는 경향을 보이며, 그 수익률이 높기 때문에 저금리 시대에 좋은 투자 대안으로 자리하고 있다. 그러나 기관투자자에 비해 개인투자자들은 공모 및 공모주에 대한 정보를 얻고 분석하기 힘들고, 따라서 실제로 성공적인 공모주 투자를 진행하기 어려운 실정이다. 따라서 본 논문은 개인투자자들이 비교적 수집하기 쉬운 비재무적 자료를 활용하여 공모가 대비 상장일 종가의 상승과 하락을 예측할 수 있는지 확인하며, 다양한 분석 방법론을 활용하여 그 정확도를 비교한다. 본 연구에 적용된 분석 방법론으로는 로지스틱 회귀분석, 판별분석, 의사결정나무, 인공신경망, 사례기반추론, 그리고 서포트벡터머신을 사용하였으며, 2007년부터 2017년 9월 6일까지의 공모주 데이터를 활용하여 실증분석을 진행하였다.

주요용어: 공모주, 기업공개, 로지스틱회귀분석, 서포트벡터머신, 인공신경망.

1. 서론

공모주는 비상장기업이 유가증권시장 또는 코스닥시장에 상장하기 위해 불특정 다수를 대상으로 판매하는 주식을 의미하며, 이러한 과정을 공모 또는 기업공개 (initial public offering; IPO)라고 한다. 공모주 투자는 최근의 저금리 시대에 좋은 투자 대안으로 각광받고 있으며, 공모주의 상장가를 의미하는 공모가에 대한 상장일 종가의 상승률은 평균 30% 이상으로, 다른 투자 방법에 비해 매우 높은 편이다 (Min, 2011). 한국의 경우 2007년부터 2017년 9월 6일까지 분할 재상장과 기업인수목적회사 (special purpose acquisition company; SPAC)를 제외한 신규상장 기업의 초기 수익률은 평균 31.3%이다. 하지만 이는 공모가 대비 상장일 주가 변동 방향이 대부분 상승이라는 점과 높은 수익률을 보이는 공모주 종목의 실제 개인투자자 배정 물량이 매우 적다는 점을 감안하여 분석되어야 한다. 따라서 수익률의 평균보다는 상승과 하락 자체에 대한 분석이 실제 투자에 더 활용도가 높으며, 이를 예측하는 방법에 대한 연구가 필요한 실정이다.

Table 1.1은 2007년부터 연도별로 IPO 종목의 수, 상장일 종가가 공모가 대비 동일 또는 상승한 종목의 수, 그리고 그 비율을 보여주고 있다. 매년 IPO 종목의 최소 64% 이상이 공모가보다 높은 상장일

[1] (03722) 서울특별시 서대문구 연세로 50, 연세대학교 투자정보공학, 박사과정.
[2] (03722) 서울특별시 서대문구 연세로 50, 연세대학교 산업공학과, 박사과정.
[3] 교신저자: (26493) 강원도 원주시 연세대길1, 연세대학교 정보통계학과, 조교수.
E-mail: ahn2615@yonsei.ac.kr

종가를 기록한 사실을 확인할 수 있으며, 전체 기간 평균은 72%이다. 반대로 상장일 종가가 공모가 대비 하락한 종목의 비율은 28%에 불과하지만, 실제 투자 과정에서 투자자의 지분은 하락 종목에 과배정되는 경우가 많아 전체 투자 결과에는 매우 큰 영향을 미친다.

Table 1.1 Direction of close price of IPO stocks compared to offering price

	2007	2008	2009	2010	2011	2012	2013	2014	2015	2016	2017
Number of IPOs (A)	73	44	66	75	72	28	38	46	73	68	34
Close price of listing day $\geq$ public offering price (B)	52	28	55	53	54	18	29	35	47	46	27
Ratio (B/A)	71%	64%	83%	71%	75%	64%	76%	76%	64%	68%	79%

기관투자자들은 재무자료를 활용한 정교한 방법으로 투자를 하는 반면, 일반 개인투자자는 주요 정보에 접근하거나 정확한 평가방법을 만들기 어려운 실정이다. 따라서 이러한 개인투자자들은 수집이 쉬운 비재무적 자료를 활용하여 보다 간단하게 결과를 확인할 수 있는 분석 방법을 필요로 하고 있다. 최근 다양한 산업 및 학문 분야에서 활용되고 있는 인공지능 방법론은 비선형성과 다중공선성을 가지고 있는 데이터의 분석에서 전통적인 통계적 분석 방법론보다 우수한 성능을 보이고 있으며, 그 중 일부는 일반인이 개인용 컴퓨터에서도 쉽게 실행 가능하도록 개발되어있기 때문에 개인투자자를 위해 적합하다. 그러나 공모주의 상장일 등락 예측의 정확도를 확인하여 투자 의사결정 지원을 위한 분석 방법론으로 유용한지 확인하는 과정이 우선되어야 한다.

본 논문에서는 다양한 통계적 분석 방법론과 인공지능 방법론들을 활용하여 IPO 시의 비재무적 자료들에 기반한 상장 초기 상승·하락 예측을 수행하고, 이를 통해 각 방법론들의 예측 정확도를 비교해 보고자 한다. 연구 데이터는 2007년부터 2017년 9월 6일까지의 IPO 종목 (분할 재상장, SPAC 제외)으로 이루어져 있으며, 분석 방법론은 로지스틱 회귀분석, 판별분석, 의사결정나무, 인공신경망, 사례기반추론, 서포트벡터머신 (support vector machine; SVM)을 활용한다. 최종적으로, 실증분석 결과를 통해 비재무적 자료가 공모주 투자 수익률 향상 제고에 활용될 수 있는지를 알아본다.

본 논문의 구성은 다음과 같다. 2절에서는 공모주 단기투자 수익률에 영향을 주는 비재무적 자료들에 대한 선행 연구를 살펴보고, 3절에서는 대표적으로 활용되는 비재무적 자료들의 분석에 사용되는 통계·인공지능 방법론에 대하여 설명하며, 4절에서는 각 분석방법을 2007년부터 2017년 9월 6일까지의 617개 종목 데이터에 적용하여 상장일 상승·하락 예측력의 정확도를 비교하였다. 마지막 결론에서는 비재무적 자료를 일반투자자가 공모주 투자에 어떻게 활용할 수 있는지에 대한 제언과 향후 연구에 대해 서술하였다.

2. 연구배경

신규기업의 상장 과정은 주식 발행 기업이 주관증권사를 선정하고, 증권거래소에 예비상장 심사를 청구해 승인을 받고서 증권신고서를 공시하는 것으로 시작된다. 이후 수요를 예측하여 공모가를 결정한 후, 일반인 청약을 받고 정해진 날짜에 주금을 받아 거래소에 상장하게 된다. 수요 예측일로부터 상장까지의 기간은 평균적으로 16일 내외가 소모된다. 주관증권사는 수요예측 전에 기업의 가치를 평가하며, 이를 일정 범위의 주가로 나타내어 공모가 범위를 제시한다. 기업의 가치평가에 주로 사용되는 방법은 비교가치법이며, 상장된 동종기업의 주가수익비율 (price earning ratio; PER), 기업가치 (enterprise value; EV), 이자·법인세·감가상각비 차감 전 영업이익 (earning before interest, tax, depreciation and amortization; EBITDA) 등을 기반으로 발행 기업의 실적에 적용해 기업의 가치를 구한다. 신규 상장기업의 경우 투자정보가 부족한 점을 감안하여 적절한 할인율을 추가적으로 적용해 공모가 범위를 구하

기도 한다.

신규기업의 상장은 전문가에 의한 적정가치가 먼저 제시된다는 점에서 부동산의 경매와 비슷한 반면, 매우 큰 차이점이 존재한다. 부동산 경매와는 다르게, 주가는 기업의 미래 실적 예상에도 크게 영향을 받을 수밖에 없기 때문이다. 따라서 신규 상장은 전문가 집단인 기관투자자들이 수요예측이라는 과정을 통해 공모가격을 정하게 되며, 일반투자자는 청약을 통해 참여 여부만을 결정하게 된다. 수요예측이 완료되면, 경쟁률, 참여 기관수, 일정기간 주식을 팔지 않겠다는 보호예수 확약 비율 등의 자료가 얻어진다. 청약이 완료되면 청약 경쟁률이 발표되는데, 보통 현업에서는 증권사 창구에서 실시간으로 경쟁률을 확인해 볼 수 있으며, 마감 시간 직전에 청약을 한다면 청약경쟁률을 상당부분 고려한 상태에서 투자를 진행할 수 있다.

시초가격은 상장일 아침 가격 기준으로 $-10\% \sim +100\%$ 범위에서 결정되며, 상장일중 가격변동은 시초가격의 $\pm 30\%$ 범위 내에서 이루어진다. 일반적인 상장의 경우 기준가는 공모가격과 동일하며, 코넥스시장에서 코스닥시장으로 이전하는 경우에는 코넥스시장에서의 시가총액과 상장으로 인한 현금유입을 총 주식수로 나눈 값으로 결정된다.

공모주의 단기 수익률에 영향을 주는 재무적 자료는 기업의 과거 실적과 현재 재무상태, 미래 실적 전망 등이 대표적이다. 비재무적 자료는 상장 과정 각 단계에서 나타나는 자료로, 비교가치법을 통한 기업가치평가 과정에서 공모가 범위를 구할 때 적용되는 할인율, 수요예측의 결과로 나타나는 경쟁률과 보호예수 확약 비율, 청약시의 경쟁률 등이 있다. 공모주 시장은 일반적인 투자시장에 비해 참여자들이 구분되어 있어 투자 가능 규모가 단기적으로 급변하지 않는다. 따라서 공모주의 과다 공급은 바로 가격하락의 원인이 되므로, 공급과 관련된 자료는 특히 중요하다. 공급과 관련된 비재무적 자료는 공모규모, 상장일 유통가능물량 및 비율, 그리고 공모 종목이 공급되는 밀도를 나타내는 공모종목밀도 등이 있다.

이러한 특성을 가진 신규기업 상장 및 공모주와 관련하여 다양한 선행 연구가 진행되었다. 할인율과 관련하여, Han (2015)은 2001년부터 2012년까지의 공모주를 분석하였으며, 할인율이 높은 공모주일수록 최초 상장시 저평가 현상이 더욱 심함과 동시에 상장일 수익률은 높아지는 것을 확인하였고, 코스닥시장에서 비금융업의 신규주식 공모 시 공모금액이 공모가 저평가율에 유의한 영향을 미친다고 분석하였다. 저평가율의 결정요인으로 Kim (2011)은 2002년 10월부터 2010년 12월에 신규 상장한 기업을 대상으로 수행한 실증분석을 통해 일반투자자 청약 경쟁률과 공모주 수익률 사이에 존재하는 유의한 정의 관계를 확인하였다. Chun 등 (2013)은 2009년부터 2011년의 코스닥 신규 상장기업을 조사하여 기관투자자 높은 수요예측 경쟁률과 일반투자자 청약 경쟁률이 공모주 초기성과에 긍정적인 영향을 미친다는 결론을 내렸고, 유사 기간에 상장하는 경쟁기업의 수가 증가할수록 공모주 초기성과에 미치는 기관투자자 수요예측 경쟁률의 영향력이 약해지는 현상도 확인하였다. 이러한 연구들의 결과를 고려하면, 공모종목밀도도 공모주 단기수익률에 대해 영향을 미치는 요인일 가능성이 존재한다고 예상할 수 있다.

구주매출의 영향에 대해서, Kang과 Kang (2012)은 2007년부터 2011년까지의 289개 신규 상장을 분석한 결과 구주매출이 많을수록 신주발행에 비해 공모주 단기 수익률이 낮아진다고 보았다. 그러나 이에 대해 Yoon (2016)은 기존에 알려져 있는 다른 주요 변수들을 통제한 회귀분석을 통해 Kang과 Kang (2012)의 선행연구 결과가 통계적으로 유의하지 않다고 분석하였다. Yoon (2016)은 2007년 6월부터 2015년 8월까지의 신규상장종목을 분석하여 지배주주의 구주매출 참여 여부가 따른 가격책정과 공모주 수익률에 미치는 영향을 확인하였고, 이러한 관계는 유가증권시장에서 더 강하게 나타났다고 주장하였다. 위와 같은 선행연구의 결과들을 감안하면, 특성에 따라 시장을 구분하는 변수도 기업공개에 영향을 미치는 요소로 고려할 필요가 있다.

이전의 관련 선행연구들은 주로 개별 요인이 공모가격 또는 공모주 단기수익률에 어떤 영향을 미치는가에 관하여 연구하였다. 그에 반해 본 연구에서는 다양한 개별 요인들을 종합하고, 그 외의 공모주 단기수익률에 영향을 미칠 가능성이 있다고 판단되는 비재무적 자료들을 변수로 추가하여 주가 방향성 예

측의 정확도를 비교한다는 점에서 차이점이 있으며, 인공지능 기법을 포함한 다양한 데이터 분석 방법론을 활용·비교한다.

3. 연구방법

공모주 단기수익률과 비재무적 자료간의 관계를 분석하고, 다양한 분석 방법론의 예측력을 비교하기 위해 다음의 단계를 걸쳐 연구를 진행한다. 전체 분석 과정은 아래 Figure 3.1과 같으며, 전통적 통계분석 방법론인 로지스틱 회귀분석과 판별분석, 결과에 대한 해석이 용이하다고 알려진 의사결정나무, 예측력이 강한 인공지능 방법론인 인공신경망, 그리고 적은 수의 데이터를 분석하는데 유리한 특성을 가진 사례기반추론과 서포트벡터머신이 사용된다.

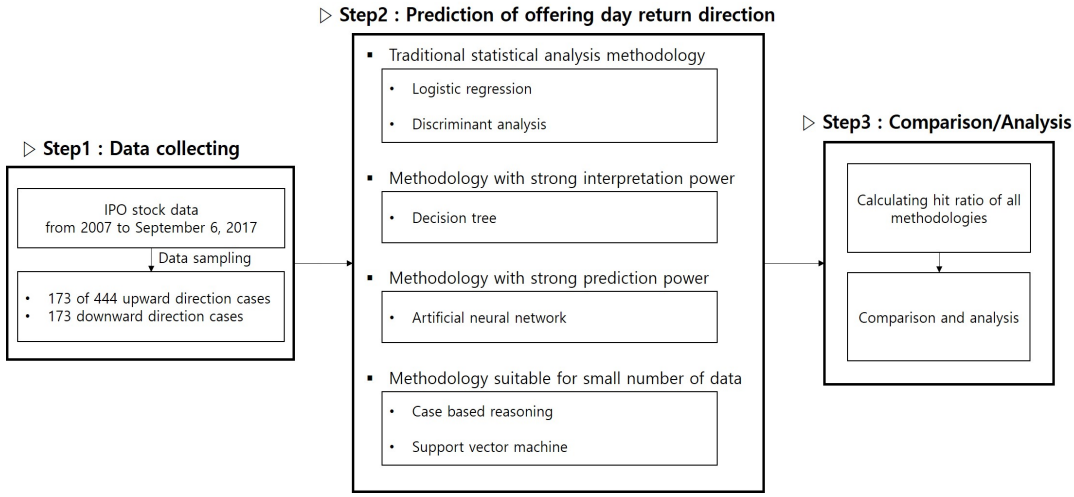

Figure 3.1 Analysis procedure diagram

3.1. 로지스틱 회귀분석과 판별분석

로지스틱 회귀분석 (logistic regression)은 분석하고자 하는 대상들이 두 개 이상의 집단으로 나누어진 다변수 데이터에 활용되는 통계적 분석 방법론으로, 독립변수들을 선형 결합하여 사건의 발생을 예측하는데 사용된다. 이는 일반적인 회귀분석과 유사한 특성을 가지지만, 범주형 데이터를 종속변수로 사용한다는 점에서 차이가 있으며, 주로 종속변수가 0과 1로 나뉘는 비연속적인 이분법 분류 문제에 사용된다 (Ahn 등, 2005). 이항 로지스틱 회귀모형은 아래 식 (3.1)과 같다. X_n은 n번째 독립변수를, β_k은 n번째 독립변수의 회귀 계수를 의미하며, p는 개별 사례가 특정 집단에 속할 확률을 의미한다.

$$ln\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_k X_n. \tag{3.1}$$

판별분석 (discriminant analysis)은 섞여있는 표본들에 대해 모집단이 2개 이상인 경우, 각각의 사례가 어느 모집단에 속해 있는가를 판별하는 분석 방법이다 (Kim, 2006; Ku와 Kim, 2017). 모집단 분류

에 최적화된 판별 변수를 활용하여 선형 판별함수를 작성하며, 서로 상관관계가 적은 독립변수를 선택함으로써 보다 효과적인 판별함수를 작성할 수 있다. 판별함수의 종속변수는 모집단의 개수에 따라 제한적인 값을 가져야 하며, 오차의 크기를 최소화하는 방향으로 함수를 학습해나가는 과정을 거친다. 식 (3.2)는 선형 판별함수를 나타내는 식으로, X_n은 n번째 판별 변수를, β_n은 n번째 판별 변수의 판별 계수를, 그리고 Z는 판별 함수의 종속변수를 의미한다.

$$Z = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_n X_n. \tag{3.2}$$

3.2. 의사결정나무

의사결정나무 (decision tree)는 데이터마이닝에서 주로 사용되는 분류 및 예측 방법론으로, 의사결정이 이루어지는 규칙을 나무 형태로 도표화하여 표현하기 때문에 회귀분석, 인공신경망 등과 같은 분석 방법론보다 모형의 해석이 용이하다는 장점이 있다 (Breiman 등, 1984; Choi와 Seo, 1999). 각 단계에서 주어진 데이터를 가장 적절하게 분할할 수 있는 변수와 기준값이 선택되며, 분할의 적합성은 분류된 데이터 내 이질적인 요소의 포함 정도를 측정하는 불순도의 최소화를 의미한다. 학습 과정과 불순도 측정 방식에 따라 다양한 형태의 의사결정나무가 존재하며, 본 연구에서는 범주형 종속변수에 대하여 지니 불순도 (Gini impurity)를 활용하는 CART (classification and regression trees)를 사용하였다. 지니 불순도를 구하는 식은 식 (3.3)과 같다. f_i는 해당 집합에 포함된 i번째 데이터 값을, n은 해당 집합 내에 포함된 데이터의 개수를, 그리고 $I_G(f)$는 해당 집합의 지니 불순도를 의미한다.

$$I_G(f) = \sum_{i=1}^{n} f_i(1 - f_i) = 1 - \sum_{i=1}^{n} f_i^2. \tag{3.3}$$

의사결정나무는 주로 하향식 기법으로 분류를 해 나가며, 하위 단계로 갈수록 세부 분류가 늘어나게 된다. 가장 포괄적으로 적용되는 분류 기준일수록 상위 단계에 위치하게 되며, 어느 변수가 선택되느냐에 따라 이후의 가지 모양이 크게 변한다는 단점이 있다. 의사결정나무는 분류 또는 예측을 위해 모두 사용될 수 있으나, 분석과정의 설명이 필요한 분류 문제에서 강점을 보인다.

3.3. 인공신경망

인공신경망 (artificial neural network)은 1943년 McCulloch과 Pitts가 인간 두뇌의 생물학적 문제해결 방식에 착안하여 개발한 인공지능 방법론으로, 입력층 (input layer), 은닉층 (hidden layer), 그리고 출력층 (output layer)으로 이루어진 모형화된 신경망 구조를 이루고 있다 (Lee 등, 2015). 모형을 구성하는 각 층은 단수 혹은 복수의 노드 (node)로 이루어져 있으며, 해당 노드들을 연결하는 가중치는 학습을 통해 최적화된다. 은닉층 내에 존재하는 노드들은 가중치를 반영하여 계산된 값을 입력받으며, 활성함수 (activation function)를 통해 출력하여 다음 노드로 전달한다. 인공신경망은 관계가 복잡한 대량의 비선형 자료를 분석하는데 용이하며 (Gardner와 Dorling, 1998), 강력한 예측력과 범용성 가지고 있기 때문에 다양한 분야에 활용되고 있다 (Ahn 등, 2005; Hong과 Park, 2009).

3.4. 사례기반추론

사례기반추론 (case based reasoning)은 여러 과거 데이터 중 새로운 문제 상황과 가장 비슷한 사례를 참고하여 결과를 예측하는 분석 방법론으로, 추론 규칙을 구하는 등의 정형화된 추론이 적용되기 힘들

때 주로 사용된다. 기존 과거 사례와 새로운 문제 상황을 구성하는 변수들 사이의 유사도를 속성유사도 함수를 사용하여 계산하며, 계산 결과 문제 상황과 가장 근사한 과거 사례를 탐색하여 동일한 결과를 예측한다. 대표적인 속성유사도함수로는 유클리디안 거리가 있으며, 식 (3.4)를 통해 산출된 값이 작을수록 두 사례의 유사도가 높다고 할 수 있다. p와 q는 각각 서로 다른 사례들을 설명하는 변수들로 이루어진 벡터를 의미한다.

$$\|p-q\| = \sqrt{\|p\|^2 + \|q\|^2 - 2pq}. \tag{3.4}$$

사례기반추론 과정에서 특정 상황을 구성하는 변수의 중요도에 따라 가중치를 설정하는 것이 가능하며, 가중치의 부여 방법에 따라 결과 차이가 날 수 있다. 또한 과거 사례들 중 문제 상황과 가장 근사한 복수의 사례들을 선정하고, 해당 사례들의 결과를 종합적으로 분석하여 예측을 수행하는 경우도 있다. 주로 과거 사례 데이터를 유지한 상황에서 새로운 데이터가 추가되기 때문에 모형의 변화가 적은 편이며, 자료 수가 부족할 때 그 우수성이 입증되었다 (Kolodner, 1991; 1993; Oh와 Kim, 2005).

3.5. 서포트벡터머신

서포트벡터머신 (support vector machine)은 자료분석을 위한 지도 학습 방법론으로 금융데이터 분석을 위한 유용한 도구로 알려져 있으며, 주로 분류분석을 위해 쓰이는 방법이다 (Cao와 Tay, 2001; Ahn과 Kim, 2014). 서로 다른 두 집합 중 하나에 속하는 데이터들로 이루어진 집합이 주어졌을 때, 데이터들을 가장 정확하게 분류할 수 있는 하나 또는 복수의 초평면 (hyper plane)들을 구함으로써 분류모형을 구축한다. 선형 분리가 불가능한 대부분의 문제 해결을 위해 데이터들을 더 높은 차원 (dimension)으로 대응시키는 과정을 거치며, 각각의 문제들에 적절한 커널 함수를 정의하여 효율성을 높이는 과정이 필요하다.

서포트벡터머신은 인공신경망과 같은 대부분의 인공지능 방법론과는 다르게 명백한 인론적 근거에 기반한 결과 해석이 용이하며, 적은 데이터를 활용하여 신속하게 분류 문제를 해결할 수 있다는 장점이 있다 (Ahn 등, 2005).

4. 실증 분석

2007년부터 2017년 9월 6일까지 분할재상장과 SPAC를 제외한 617개 신규상장종목 데이터를 활용하여 실증분석을 진행한다. 자료는 금융감독원과 IPOStock을 통해 수집하였으며, 공모주 단기 수익률은 공모가 대비 상장일 종가의 상승률로 구하여 그 부호에 따라 상승 및 동일과 하락을 구분하였다. 전체 617개 종목의 데이터 중 공모주 단기 수익률이 상승 또는 동일인 사례가 444개, 하락인 사례가 173개였고, 실증분석에서는 상승 또는 동일인 종목과 하락인 종목의 개수를 맞추기 위해 444개의 상승 또는 동일 사례 중 173개를 임의 추출해 사용하였다. 각 분석 단계별 방법론의 모형을 구축하기 위한 학습구간 (training period)으로 전체 사례의 70%를, 모델의 검정구간 (testing period)으로 전체 사례의 30%를 임의배분 후 활용하였으며, 인공신경망 모형은 학습구간, 검증구간 (validation period), 그리고 검정구간을 각각 40%, 30%, 그리고 30%로 배분하였다.

4.1. 예측에 활용될 비재무 정보

본 연구에서 예측모형의 입력변수로써 고려된 비재무 정보는 Table 4.2와 같으며, 사용된 자료들은 다음과 같이 설명되어질 수 있다.

Table 4.1 Empirical data

Case	Training period	Test period	Total
Increase or Equal	121	52	173
Decrease	121	52	173
Total	242	104	346

(i) 주식의 공급량이 얼마인가를 나타내는 비재무적 자료로 공모금액, 유통가능물량비율, 공모종목밀도를 활용하였다.

(ii) 기존주주의 매도 수준을 보여주는 자료로 구주매출비율을 활용하였다.

(iii) 기관투자자 수요예측의 결과 중에서 수요예측경쟁률, 의무보유확약비율, 수요예측 응찰수량 중 공모가격 초과분의 비율을 활용하였다. 수요예측 응찰수량 중 공모가격 초과분의 비율이 높을수록 주관사는 신청자의 전반적인 가격분포에서 낮은 수준으로 공모가격을 결정했다고 볼 수 있다. 이는 증권신고서에는 기재되어 있지만, 기존의 연구에서는 사용되지 않았던 변수이다.

(iv) 일반인 청약의 결과로부터 도출되는 청약경쟁률을 활용하였다.

(v) 상장 시장과 기업의 국적을 구분한 후 순위 등급을 매겨 자료로 활용하였다. 기업과 회계 자료를 투명도 순위는 미국, 일본, 한국, 중국 및 동남아 국가 순으로 생각할 수 있으며, 한국시장에서는 코스닥에 상장되는 경우가 유가증권시장에 상장되는 경우보다 소규모기업이 많아 투자매력도가 높을 것으로 예상하였다. 다만 코넥스에 상장된 기업이 코스닥으로 이전하여 상장하는 경우에는 상대적으로 신뢰도가 낮다고 판단되는 경향에 따라 투자매력도가 낮아진다. 이들을 종합해 미국 국적 기업에 5, 일본국적 기업에 4, 한국기업 중 코스닥 상장은 3, 유가증권시장 상장은 2, 코넥스에서의 이전 상장은 1, 중국 및 동남아 국가 기업에 대해서는 0의 등급을 부여하였다.

Table 4.2 Input variables for prediction models

Variable name	Explanation
Aggregate amount of IPO	Public offering price × number of stocks
Percentage of IPO non lock-up shares	Non lock-up shares / total shares on the date of listing
Concentration of IPO shares	The number of Issues in the previous one month and after one month from date of listing
Ratio of shares by selling shareholders	Old shares / number of public offering shares
Bidding ratio of book building	Total amount of bidding / allocated amount for Institutional investor
Rate of lock-up	Amount of lock-up bidding / total amount of bidding
Percentage of bidding shares in book building exceeding IPO price	Amount of excess bidding over determined offer price / Total amount of bidding
Subscription ratio of retail investors	Total amount of subscription for private investor / Allocated amount for private investor
Company score by market and nationality	·United States nationality: 5-point ·Japanese nationality: 4-point ·Korean nationality and KOSDAQ market: 3-point ·Korean nationality and KOSPI market: 2-point ·Transferred from KONEX market: 1-point ·Chinese or Southeast Asian nationality: 0-point

4.2. 로지스틱 회귀 분석 결과

로지스틱 회귀분석 전에, 비재무적 자료의 독립변수의 다중공선성 여부를 확인하기 위해 분산팽창계수 (variance inflation factor; VIF)분석을 수행하였다 (Park, 2016). 분산팽창계수는 특정 독립변수를 설명하기 위해 나머지 독립변수들로 구성한 회귀식의 결정계수를 이용하여 산출되며, 분산팽창계수가 크다는 것은 해당 독립변수가 나머지 독립변수들로 잘 설명되어진다는 것을 의미하므로 다중공선성을 의심해보아야 한다. 식 (4.1)은 k번째 독립변수의 분산팽창계수 (VIF_k)를 나타내며, R_k^2는 k번째 독립변수를 설명하기 위해 나머지 독립변수들로 만들어진 회귀식의 결정계수를 의미한다.

$$VIF_k = \frac{1}{1 - R_k^2}. \tag{4.1}$$

비재무적 자료들에 대한 분산팽창계수 산출 결과, Table 4.3과 같이 모든 독립변수들이 5이하의 값을 보였다. 이는 독립변수들 간의 다중공선성이 의심되지 않는다는 것을 의미하므로, 변수들의 통합 또는 제거가 불필요함을 알 수 있다.

Table 4.3 Variance inflation factors for input variables

Variable name	VIF
Aggregate amount of IPO	1.135
Percentage of IPO non lock-up shares	1.062
Concentration of IPO shares	1.062
Ratio of shares by selling shareholders	1.078
Bidding ratio of book building	1.631
Rate of lock-up	1.207
Percentage of bidding shares in book building exceeding IPO price	1.111
Subscription ratio of retail investors	1.552
Company score by market and nationality	1.072

모든 독립변수들 간 다중공선성이 관찰되지 않음에 따라, 로지스틱 회귀분석을 시행하였다. 모형 구축 시 종속변수인 공모가 대비 상장일 종가의 상승 혹은 하락을 예측하기 위한 모든 독립변수를 입력하여 추정계수를 산출하였으며, 그 결과는 Table 4.4에서 보여주고 있다. 각 입력변수들의 추정계수에 대한 유의확률 결과를 통해 공모종목밀도와 의무보유확약비율, 그리고 청약경쟁률이 중요 영향인자로 판단되며, 이 중 의무보유확약비율과 청약경쟁률은 5% 유의수준 하에서 유의하게 나타났다. Table 4.5는 로지스틱회귀 모형의 예측 정확도를 보여주고 있다. 구축된 모형의 학습구간 예측 정확도는 67.4%, 검정구간 예측 정확도는 68.3%로 나타났으며 실제 투자활동에서 중요한 역할을 하는 하락 예측 정확도는 학습구간에서 74.4%, 검정구간에서 73.1%를 보였다.

Table 4.4 Coefficient and p-value of logistic regression model

Variable name	Estimate coefficient	p-value
Aggregate amount of IPO	6.456	0.218
Percentage of IPO non lock-up shares	845.934	0.192
Concentration of IPO shares	-1.513	0.067
Ratio of shares by selling shareholders	-1.850	0.206
Bidding ratio of book building	-0.466	0.710
Rate of lock-up	2.033	0.030*
Percentage of bidding shares in book building exceeding IPO price	0.368	0.501
Subscription ratio of retail investors	4.648	0.000*
Company score by market and nationality	0.341	0.721

* Significant at 5%

Table 4.5 Predictive results of logistic regression analysis

	Training period			Test period		
Observation \ Predicted value	Decrease	Increaseor equal	Sub total	Decrease	Increase or equal	Sub total
Decrease	90 (74.4%)	31 (25.6%)	121	38 (73.1%)	14 (26.9%)	52
Increase or Equal	48 (39.7%)	73 (60.3%)	121	19 (36.5%)	33 (63.5%)	52
Total accuracy	67.4% ((90+73)/(121+121))			68.3%((38+33)/(52+52))		

The cut-off value of logistic regression for prediction is 0.5

4.3. 판별 분석 결과

앞선 로지스틱 회귀분석과 마찬가지로 모든 독립변수를 포함하는 판별함수를 구축하였다. 판별분석의 결과는 전반적으로 로지스틱 회귀분석과 비슷하게 나타났으며, 표준화된 정준 판별함수의 계수는 아래 Table 4.6과 같다. 표준화 정준 판별함수의 계수는 설명변수들이 판별함수에 기인하는 상대적인 중요도를 나타낸다. 분석 결과에 따르면, 청약경쟁률과 의무보유확약비율, 공모종목밀도, 그리고 공모액 순서로 변수의 설명력이 높음을 확인할 수 있다.

Table 4.6 Coefficients of discriminant function

Variable name	Coefficient
Aggregate amount of IPO	0.226
Percentage of IPO non lock-up shares	0.172
Concentration of IPO shares	-0.279
Ratio of shares by selling shareholders	-0.195
Bidding ratio of book building	-0.044
Rate of lock-up	0.347
Percentage of bidding shares in book building exceeding IPO price	0.100
Subscription ratio of retail investors	0.778
Company score by market and nationality	0.072

또한 Table 4.7에서 볼 수 있듯이 판별함수의 예측정확도는 학습구간에서 66.9%, 검정구간에서 67.3%를 보였으며, 하락 예측 정확도는 학습구간에서 76.0%, 검정구간에서 75.0%로 로지스틱 회귀분석보다 향상된 결과를 보였다.

Table 4.7 Predictive results of discriminant analysis

	Training period			Test period		
Observation \ Predicted value	Decrease	Increaseor equal	Sub total	Decrease	Increase or equal	Sub total
Decrease	92 (76.0%)	29 (24.0%)	121	39 (75.0%)	13 (25.0%)	52
Increase or equal	51 (42.1%)	70 (57.9%)	121	21 (40.4%)	31 (59.6%)	52
Total accuracy	66.9% ((92+70)/(121+121))			67.3% ((39+31)/(52+52))		

4.4. 의사결정나무 분석 결과

본 연구에서는 의사결정나무 알고리즘을 이용한 예측 모형 구축 시, 의사결정 규칙 생성을 위한 분리 기준으로써 지니 계수를 사용하는 CART 방식으로 모형 학습을 진행하였다. 독립변수의 중요도를 분석하기 위해 정규화중요도를 산출하였으며, Table 4.8은 산출된 독립변수들의 중요도 및 정규화 중요도를 보여주고 있다. 분석 결과 청약경쟁률, 수요예측경쟁률, 공모금액, 그리고 의무보유확약비율 순으로 영향력이 큰 것을 확인할 수 있다.

Table 4.8 Importance of input variables for decision tree

Variable name	Importance	Normalized importance
Aggregate amount of IPO	0.025	22.6%
Percentage of IPO non lock-up shares	0.002	1.7%
Concentration of IPO shares	0.007	6.8%
Ratio of shares by selling shareholders	0.012	10.6%
Bidding ratio of book building	0.032	29.6%
Rate of lock-up	0.019	17.0%
Percentage of bidding shares in book building exceeding IPO price	0.001	1.3%
Subscription ratio of retail investors	0.110	100%
Company score by market and nationality	0.001	0.8%

의사결정나무 분석에서 구축된 모형의 예측정확도는 학습구간에서 70.7%, 검정구간에서 68.3%를 보였으며, 하락 예측 정확도는 학습구간에서 60.3%, 검정구간에서 53.8%로 그다지 효과적인 결과를 얻지 못한 것을 알 수 있다 (Table 4.9). 의사결정나무는 오히려 비하락 (상승 또는 동일)예측에 대하여 학습구간과 검정구간 모두에서 80%를 상회하는 모습을 보였으며, 이는 모든 분석 방법론들의 결과 중 가장 높은 수치이다.

Table 4.9 Predictive results of decision tree analysis

	Training period			Test period		
Observation \ Predicted value	Decrease	Increaseor equal	Sub total	Decrease	Increase or equal	Sub total
Decrease	73 (60.3%)	48 (39.7%)	121	28 (53.8%)	24 (46.2%)	52
Increase or equal	23 (19.0%)	98 (81.0%)	121	9 (17.3%)	43 (82.7%)	52
Total accuracy	70.7% ((73+98)/(121+121))			68.3% ((28+43)/(52+52))		

4.5. 인공신경망 분석 결과

인공신경망 분석 시 과적합 방지를 위해 실험 데이터를 훈련구간과 검증구간, 그리고 검정구간으로 나

누어 실험을 진행하였으며, 각 단계에 해당하는 자료의 비중은 4:3:3으로 임의 배분하였다. 은닉층의 개수에 따른 분석 정확도 비교를 위해 한 개의 은닉층을 사용한 인공신경망과 두 개의 은닉층을 사용한 인공신경망을 구축하였으며, 노드의 개수는 학습에 따라 자동으로 결정되도록 설정하였다. 마지막으로 은닉층의 활성화 함수는 쌍곡탄젠트 함수를, 출력층의 활성화 함수는 항등함수를 사용하였다.

Table 4.10 Predictive results of neural network with single hidden layer

	Training period			Validation period			Test period		
Predicted value / Observation	Decrease	Increase or equal	Sub total	Decrease	Increase or equal	Sub total	Decrease	Increase or equal	Sub total
Decrease	55 (79.7%)	14 (20.3%)	69	33 (63.5%)	19 (36.5%)	52	33 (63.5%)	19 (36.5%)	52
Increase or equal	24 (34.8%)	45 (65.2%)	69	23 (44.2%)	29 (55.8%)	52	16 (30.8%)	36 (69.2%)	52
Total accuracy	72.5% ((55+45)/(69+69))			59.6% ((33+29)/(52+52))			66.3% ((33+36)/(52+52))		

Table 4.10과 Table 4.11은 단일 은닉층 신경망과 이중 은닉층 신경망의 예측 결과를 보여주고 있다. 단일 은닉층 인공신경망 분석 결과, 훈련구간에서의 예측 정확도는 72.5%, 검증구간에서의 예측 정확도는 59.6%, 그리고 검정구간에서의 예측 정확도는 66.3%로, 통계적 분석 방법에 의한 예측 정확도보다 낮게 나타났다. 그러나 이중 은닉층 인공신경망 구축을 통한 학습구간과 검증구간의 예측정확도가 각각 68.8%, 62.5%로 나타났으며, 검증구간의 예측정확도는 단일 은닉층 신경망에 비해 향상된 68.3%의 결과를 얻을 수 있었다. 이를 통해 단일 은닉층보다 이중 은닉층 인공신경망 모형이 더 안정적인 (robust) 예측 모형이라고 판단되며, 적절한 신경망 구조의 탐색을 통해 보다 효과적인 예측모델 구축이 가능함을 알 수 있다.

Table 4.11 Predictive results of neural network with double hidden layer

	Training period			Validation period			Test period		
Predicted value / Observation	Decrease	Increase or equal	Sub total	Decrease	Increase or equal	Sub total	Decrease	Increase or equal	Sub total
Decrease	53 (76.8%)	16 (23.2%)	69	33 (63.5%)	19 (36.5%)	52	38 (73.1%)	14 (26.9%)	52
Increase or equal	27 (39.1%)	42 (60.9%)	69	20 (38.5%)	32 (61.5%)	52	19 (36.5%)	33 (63.5%)	52
Total accuracy	68.8% ((53+42)/(69+69))			62.5% ((33+32)/(52+52))			68.3% ((38+33)/(52+52))		

4.6. 사례기반추론 분석 결과

사례기반추론 분석을 위해서는 새롭게 주어진 상황과 가장 유사한 사례를 탐색하기 위한 데이터베이스가 요구된다. 본 연구에서는 전체 실험 데이터의 70%를 데이터베이스로 활용하고, 나머지 30%에 해당하는 상황들에 대한 추론을 수행하여 예측정확도를 관찰한다. 데이터베이스 내에 있는 사례들과 추론 대상이 되는 상황간의 유사도를 측정하는 과정에서 변수들의 가중치가 편향되는 현상이 나타나지 않도록 동일 가중 방식을 택하였고, 가장 유사도가 높은 다섯 개의 사례를 추출하여 그 중 다수가 나타냈던 결과와 동일할 것이라는 가정을 기반으로 예측을 진행하였다. 분석 결과 사례기반추론의 예측 정확도는 64.4%로 나타났으며, 그 중 실제 하락했던 것으로 밝혀진 상황들에 대한 예측 정확도는 76.9%로 매우 높은 값을 보였다 (Table 4.12).

Table 4.12 Predictive results of case-based reasoning

Observation \ Predicted value	Test period			
	Decrease	Undefined	Increase or equal	Sub total
Decrease	40 (76.9%)	7 (13.5%)	5 (9.6%)	52
Increase or equal	19 (36.5%)	6 (11.5%)	27 (51.9%)	52
Total accuracy	64.4% ((40+27)/(52+52))			

데이터베이스와 유사한 사례를 찾는 사례기반추론의 알고리즘 특성에 따라 훈련기간의 정확도는 도출되지 않으며, 유사도가 높은 과거 사례를 탐색하는 과정에서 하락 사례와 비하락 사례가 동일한 개수로 나타나는 경우는 결과 없음으로 표기하였다.

4.7. 서포트벡터머신 분석 결과

모든 독립변수를 활용하여 서포트벡터머신 분석을 수행한 결과, 모형 학습구간에서는 예측 정확도가 74.0%로 산출되었다. 이는 적용된 방법 중 가장 높은 수치이며, 검정구간에서도 70.2%로 역시 모든 분석 방법론 중 가장 예측 정확도를 보였다. 이러한 분석 결과는 자료의 수가 많지 않은 경우 다른 예측 방법론보다 우수한 성과를 보인다는 서포트벡터머신의 특징과 관련된 기존 연구 결과들에 부합한다고 할 수 있다. 실제 투자에서 중요한 역할을 하는 하락 예측 정확도는 검정구간에서 80.8%로 본 연구에서 고려한 다른 예측 모형들 중에 가장 높은 값을 보였다 (Table 4.13). 이를 통해 서포트벡터머신 방법론은 본 연구를 위해 가장 적절한 예측 알고리즘이라고 판단할 수 있다.

Table 4.13 Predictive results of support vector machine

Observation \ Predicted value	Training period			Test period		
	Decrease	Increase or Equal	Sub total	Decrease	Increase or Equal	Sub total
Decrease	100 (82.6%)	21 (17.4%)	121	42 (80.8%)	10 (19.2%)	52
Increase or equal	42 (34.7%)	79 (65.3%)	121	21 (40.4%)	31 (59.6%)	52
Total accuracy	74.0% ((100+79)/(121+121))			70.2% ((42+31)/(52+52))		

5. 결론 및 제언

주식 시장에서 발행시장은 유통시장만큼 중요하다. 하지만 기존의 IPO관련 연구에서는 초기 저가 발행의 이유를 찾는 것이 대부분이었고, 이는 대개 발행기업 내의 자료 위주로 분석한 것이었다. 본 연구는 투자자 입장에서 쉽게 활용해 볼만한 투자 방법을 발행기업의 재무실적 이외의 자료를 활용해 찾는 것을 목표로 하며, 2007년부터 2017년9월까지의 신규상장기업에 대해 분석하였다. 본 논문에서는 단기투자를 가정한 상장일 종가 등락률 예측에 초점을 맞추었다. 이는 수익률보다 등락여부를 예측하는 것이 투자에 활용하기에 적합하기 때문이며, 이를 위해, 공모가 대비 상장일 종가 등락 여부에 대해 통계 및 기계학습 방법 등 6가지 방법론을 고려하여 분석해 보았다. Table 5.1은 분석 결과를 요약하여 보여주고 있다. 실제 투자에서 주가의 상승보다 하락 예측을 정확히 하는 것이 더욱 중요한 역할을 할 수 있다는 것을 감안하였을 때 본 연구에서 주가 하락 예측력은 모형의 민감도라고 표현할 수 있기에 Table 5.1에서 민감도는 하락 예측 정확도를 의미한다. 로지스틱 회귀분석, 판별 분석, 의사결정나무, 그리고 인공신경망은 서로 비슷한 예측 정확도를 나타내었으나, 의사결정나무 모형의 경우 민감도 측면에서 매우 좋지 않을 결과를 보여주었다. 인공신경망 모형의 예측 정확도는 로지스틱 회귀와 판별분석 모형 비

해 좋은 결과를 보여주지 못하였는데, 이는 인공 신경망 분석이 복잡한 구조, 비선형적 구조에 더 잘 적용된다는 특성을 고려하면 IPO 시장의 투자 과정이 일정부분 정형화된 형태로 이루어지며 그에 따라 발생되는 비재무적 데이터 역시 그러한 성격을 지니고 있다는 사실을 유추할 수 있다.

Table 5.1 Prediction performance of all analysis methods during test period

	LR	DA	DT	NN	CBR	SVM
Accuracy	68.3%	67.3%	68.3%	68.3%	64.4%	70.2%
Sensitivity	73.1%	75.0%	53.8%	73.1%	76.9%	80.8%
AUC	0.754	0.762	0.683	0.738	0.758	0.782

LR: logistic regression, DA: discriminant analysis, DT: decision tree, NN: neural network, CBR: case-based reasoning, SVM: support vector machine

서포트벡터머신은 여섯 개의 예측 모형 중 가장 높은 예측 정확도 결과를 보여주었다. 70%를 상회하는 수준의 높은 예측 정확도와 실제 투자에서 중요한 하락방향 예측 정확도 즉, 민감도가 82.6%로 가장 좋은 결과를 나타내었다. 또한 예측 정확도와 함께 분류모형의 예측성과 지표로 사용되며, ROC (receiver operating characteristic) 곡선의 아래쪽 넓이를 의미하는 AUC (area under the curve)값도 서포트벡터머신 예측모형이 가장 높은 점을 감안한다면 서포트벡터머신 방법론의 높은 실무 활용도를 기대할 수 있다.

신규상장 주식은 상장 초에 일중 가격변동이 크기 때문에 데이트레이더와 같은 단기투자자들의 관심을 많이 받으며, 수요예측이나 청약 등에서 충분한 배정을 받지 못한 투자자는 상장일 시초가에 매수하는 경우가 많다. 따라서 본 연구 결과를 토대로 상장일 시초가를 기준으로 당일 종가나 일주일 후 종가 등의 단기간 투자에 대해서도 비재무적 자료를 통한 분석이 가능한지 검증하는 연구가 향후 필요할 것으로 판단된다. 또한 본 연구에서는 예측모형 구축 시 학습구간과 검정구간을 전체 사례의 7:3 비율로 설정하였는데, 이 비율이 변화함에 따라 각 모형의 예측성과가 어떻게 달라지는지 향후 연구에서 살펴볼 필요가 있을 것이다. 다만 각 예측 모형이 훈련구간과 검정구간에서 예측력 차이가 크지 않은 점을 감안하였을 때 학습구간과 검정구간의 비율이 변화함에 따라 예측성과 결과들이 크게 달라지지 않을 것으로 생각된다.

References

Ahn, C. K. and Kim, D. G. (2014). Efficient variable selection method using conditional mutual information. *Journal of the Korean Data & Information Science Society*, **25**, 1079-1094.

Ahn, H. C., Kim, K. J. and Han, I. G. (2005). Purchase prediction model using the support vector machine. *Journal of Korea Intelligent Information Systems Society*, **11**, 69-81.

Breiman, L., Friedman, J. H., Olshen, R. A. and Charles, J. S. (1984). *Classification and regression trees*, Wadsworth, New York.

Cao, L. and Tay, F. E. H. (2001). Financial forecasting using support vector machines. *Neural Computing and Applications*, **10**, 184-192.

Choi, J. H. and Seo, D. S. (1999). Decision trees and its applications. *Journal of The Korean Official Statistics*, **4**, 61-83.

Chun, K. M., Gee, I. H. and Lee, H. U. (2013). The effect of IPO subscription rates for institutional investors and private investors on IPO firm performance: The moderating role of competition and on-line reviews. *Korean Journal of Business Administration*, **26**, 1149-1176.

Han, G. S. (2015). A study on the underpricing of IPOs in Korea capital. *Korea International Accounting Review*, **59**, 125-146

Hong, T. H. and Park, J. Y. (2009). Integrating the customer response model in direct marketing using case-based reasoning. *The Journal of Information Systems*, **18**, 375-399.

Kang, S. K. and Kang, H. C. (2012). An IPO satisfying all shareholders: Sellers, stayers, and new comers - Mando's successful Eeit strategy design without hampering anyones' wealth. *Korea Business Review*, **16**, 141-164.

Kim, J. S. (2011). An analysis of the subscription rate's impact on IPO returns. *Journal of Knowledge Studies*, **9**, 39-62.

Kim, S. Y. (2006). Prediction of hotel bankruptcy using multivariate discriminant analysis, logistic regression and artificial neural network. *Journal of Tourism Sciences*, **30**, 53-75.

Kolodner, J. (1991). Improving human decision making through case-based decision aiding. *AI Magazine*, **12**, 52-68.

Kolodner, J. (1993). *Case-based reasoning*, Morgan Kaufman, San Mateo, CA.

Ku, J. M. and Kim, J. H. (2017). Development of game indicators and winning forecasting models with game data. *Journal of the Korean Data & Information Science Society*, **28**, 237-250.

Lee, K. J., Lee, H. J. and Oh, K. J. (2015). Using fuzzy-neural network to predict hedge fund survival. *Journal of the Korean Data & Information Science Society*, **26**, 1189-1198.

Min, J. H. (2011). The book building process and the short and long run performance of IPO stocks. *The Journal of Professional Management*, **14**, 165-181.

Oh, K. J. and Kim, T. Y. (2005). Using case-based reasoning to develop the daily economic condition indicator based on data driven method. *Korea Deposit Insurance Corporation*, **6**, 36-64.

Park, H. C. (2016). Generally non-linear regression model containing standardized lift for association number estimation. *Journal of the Korean Data & Information Science Society*, **27**, 629-638.

Yoon, T. J. (2016). *Do secondary shares on controlling shareholder affect the IPO underpricing?*, Master Thesis, Department of Business Administration, Seoul National University, Seoul.

Journal of the Korean Data & Information Science Society
2018, **29**(2), 425–439

http://dx.doi.org/10.7465/jkdi.2018.29.2.425
한국데이터정보과학회지

A study on initial price change prediction of IPO shares using non-financial information

Sanghun Shin[1] · Hyun Jun Lee[2] · Jae Joon Ahn[3]

[1]Department of Investment Information Engineering, Yonsei University
[2]Department of Industrial Engineering, Yonsei University
[3]Department of Information & Statistics, Yonsei University

Received 17 February 2018, revised 16 March 2018, accepted 16 March 2018

Abstract

The stock in a public offering refers to stocks that are sold to an unspecified public to be listed on the KOSPI or the KOSDAQ market. Because the initial public offering (IPO)'s closing price of offering day mostly tends to rise opposed to the offering price of it, which is selling price, it is a good alternative investment asset in the low interest rate period. However, it is difficult for individual investors to obtain and analyze information on public offerings and IPO stocks compared to institutional investors. Therefore, this paper confirms whether individual investors can predict the rise and fall of the IPO's closing price of offering day compared to its offering price by using multiple data analysis methodologies and non-financial data that is relatively easy to collect, and compares the accuracy of them. Logistic regression, discriminant analysis, decision tree, artificial neural network, case based reasoning, and support vector machine were used for analysis, and the empirical experiments was conducted using the IPO data from 2007 to September 6, 2017.

Keywords: Artificial neural network, initial public offering (IPO), logistic regression, stock in a public offering, support vector machine.

[1] Graduate student, Department of Investment Information Engineering, Yonsei University, Seoul 03722, Korea.
[2] Graduate student, Department of Industrial Engineering, Yonsei University, Seoul 03722, Korea.
[3] Corresponding author: Assistant professor, Department of Information & Statistics, Yonsei University, Wonju 26493, Korea. E-mail: ahn2615@yonsei.ac.kr