Journal of the Korean Data & Information Science Society
2020, **31**(3), 579–590

http://dx.doi.org/10.7465/jkdi.2020.31.3.579
한국데이터정보과학회지

인공지능 (AI)을 활용한 공모주 투자여부 및기준 수익률 달성 여부 예측 모델

조득환[1] · 류호선[2] · 정승환[3] · 오경주[4]

[1]연세대학교 투자정보공학 · [234]연세대학교 산업공학과

접수 2020년 2월 17일, 수정 2020년 3월 17일, 게재확정 2020년 3월 25일

요 약

우리나라의 금융시장이 발전함에 따라, 개인의 투자에 관한 관심이 높아지면서 공모주 시장의 참여자도 증가하고 있다. 반면 학계에서는 다른 금융 분야와 비교하면 공모주에 관한 연구는 부족하고, 상장 이후 공모주식의 가격 예측에 관한 연구는 전무한 상황이다. 개인투자자들이 공모주 시장에서 얻을 수 있는 정보는 매우 한정된 상황에서 본 논문은 공개된 데이터로부터 인공지능과 통계 방법론을 통하여 공모주 투자 시의 수익률을 예측하는 데에 유의미한 변수를 발견하고 이를 예측하여 투자자의 의사결정에 도움을 줄 수 있는 결과를 얻었다. 본 논문의 모델에 판별분석, 의사결정나무, 로지스틱 회귀분석, 인공신경망, 유전자 알고리즘의 방법론을 사용하였으며, 변수선정과정, 예측모델에서 각각 최적의 방법론을 사용하여 모델을 구성하였다.

주요용어: 공모주, 로지스틱 회귀분석, 유전자 알고리즘, 의사결정나무, 인공신경망.

1. 서론

주식시장이 성숙해질수록 투자자들의 투자방법은 다양해진다. 우리나라에서 공모주 투자 수익률이 높다는 이야기들이 여기저기서 들리기 시작하면서 많은 사람이 공모주 투자에 관심을 기울이게 되었다. 몇 년 전까지만 해도 소수의 투자자들이 증권회사의 인맥을 통한 빠른 정보력을 기반으로 공모주 투자를 했었지만, 현재는 일반 투자자들도 정보통신의 발달과 여러 증권사 및 증권정보배포 기관의 투자정보제공에 따라 공모주 투자에 참여할 수 있는 기회가 점점 더 증가하는 추세이다.

많은 사람의 이목이 쏠린 공모주 투자에서 과연 모든 사람이 수익을 낼 수 있을까? 일반적으로는 공모주 투자에 참여하였을 때, 주식을 많이 배정받을수록 높은 수익을 올릴 수 있다고 생각하지만 사실상 그렇지는 않다. 공모주 투자는 기존 주식투자보다 가격 변동성이 크고 거래량이 한정되어 있기에 최근에는 펀드상품에서도 고위험 상품으로 분류되어 판매되고 있을 정도이며, 단순히 접근했다가는 큰 손실을 볼 수 있는 투자 분야이다. 또한, 우리나라 증권거래소에 상장되어 거래되고 있는 기업들의 수는 증가하고 있지만, 일반 투자자들이 얻을 수 있는 투자 정보는 주간사에서 배포하는 각 회사의 상장하기 전 투자설명서와 IR (Investor relations) 자료가 전부이다. 일반 투자자들은 이러한 열악한 환경 속에서의 공모주 투자는 상당히 위험한 투자임을 인지하여야 한다.

[1] (03722) 서울특별시 서대문구 연세로 50, 연세대학교 투자정보공학, 박사과정.
[2] (03722) 서울특별시 서대문구 연세로 50, 연세대학교 산업공학과, 통합과정.
[3] (03722) 서울특별시 서대문구 연세로 50, 연세대학교 산업공학과, 통합과정.
[4] 교신저자: (03722) 서울특별시 서대문구 연세로 50, 연세대학교 산업공학과, 교수.
E-mail: johanoh@yonsei.ac.kr

중견기업이나 대기업의 경우 여러 평가기관에서 기업을 평가하기에 가치를 결정하기에 상대적으로 수월하지만, 대부분의 벤처기업 및 중소기업들은 평가대상에서 제외되어 기업의 가치를 결정하기에 상당한 어려움이 있다. 이러한 문제를 해결하기 위해 기업의 가치평가, 신규공모주식의 가치평가 (임병권, 2011), 신규공모주의 공모 후 주가 수준 (Song과 Park, 1995), 신규공모주의 발행 가격 결정에 미치는 영향 (Kim, 1994) 등 기업에 관한 다양한 연구들이 꾸준히 진행되어 왔고, 기관투자자와 개인투자자의 IPO (Initial public offering) 시장에서의 투자성과에 대한 분석 (Min, 2017), 공모주의 저가 발행에 대한 원인분석 (Jung, 1992) 등 공모주 산업 전반에 관한 연구 등도 진행되어 왔다.

하지만 IPO 분야에서 인공지능 기법을 활용한 분석연구는 시도된 적이 거의 없었다. 특히 기관투자자 및 개인투자자들의 입장에서 인공지능을 활용해 상장을 앞둔 기업의 적절한 공모가를 판단하고 공모주 투자 참여를 통하여 투자수익을 예측하는 연구는 전무한 상황이다. 이에 본 연구에서는 인공지능기법을 활용한 공모주 투자수익을 극대화할 수 있는 방법을 제안하려고 한다.

본 논문은 다음과 같이 구성된다. 2장 연구배경에서는 공모주 시장이 무엇인지, 신규 주식 공모 과정에 관하여 서술하고, 3장에서는 본 논문에서 IPO 시장의 분석을 위해 사용한 방법론들과 제안한 모델에 관해 서술하였다. 4장 실증분석에서는 IPO 시장 데이터를 바탕으로 한 실증분석 결과를 서술하고, 5장 결론에서는 본 연구의 결과에 대한 요약과 발전 방향에 대해 서술하였다.

2. 연구배경 및 방법론

우리나라의 공모주 시장은 수많은 시행착오 끝에 2008년 금융위기 이후부터 안정적으로 관리가 되어오고 있다. 회사가 상장을 결정하게 된 후 약 9개월에서 12개월 동안 상장업무를 증권사와 진행을 하며, 증권 감독기관에 신고한 후 거래소에 상장하게 되는데, 우리는 거래소에서 상장승인을 취득하고 증권 감독기관에 신고한 이후부터 시장에 상장하는 첫날까지를 공모주 시장이라 일컫는다.

투자자들은 발행사 주식의 공모 직전 15%∼40% 할인 (공모가 밴드)된 가격으로 공모주 참여를 진행한다. 초기에는 발행사의 정보가 주식에 충분히 반영되지 않았기 때문에 저평가되어있지만, 점차 이러한 부분들이 해소되면서 주가가 상승해서 차익이 발생하는 경우가 많다. 공모주 투자자는 할인된 공모가와 거래가 시작되는 시초가의 차익을 가져가기 위해 투자를 하고 있다. 신규공모주의 이상 저평가 현상이라 하여 할인 발행되거나 상장 이후 부양 정책이 진행되어 수익이 많이 나는 경우도 간혹 있다. 하지만 모든 공모주 투자자가 수익을 가져가지는 않는다. 일반인들이 공모주 투자를 할 경우에는 회사의 가치분석, 공모주 시장현황, 그리고 증권시장 현황 등 공모주 산업 전반에 대한 이해가 필요하다.

기업이 공모시장에서 신규 상장할 때는 우선 거래소 상장규정 조건에 규합되어야 하고 증권사에 상장자문을 받는다. 주관사는 예비심사청구서를 기업과 함께 준비, 작성하여 거래소에 예비심사청구를 하고, 거래소는 3개월간 예비심사를 거쳐 결과 여부를 발표한다. 이후 감독당국에 증권신고서를 신고하고 기관투자자 대상 IR을 진행하여 수요예측을 실시하고, 공모가격 결정 후 일반투자자 청약을 거쳐 상장 후 주권매매를 개시한다. 이때, 공모가 산정 과정에서 주관사는 회사의 영업능력, 재무제표, 경영 성과, 산업의 특성, 시장의 규모, 비교기업 등 여러 가지를 종합해 적정한 공모가를 결정하게 된다. 공모가 밴드를 구하고 희망공모가를 선정하는 시기는 일반적으로 상장승인 3∼4개월 전이다. 또한, 공모가가 결정된 후에 거래소에 상장하기까지도 약 2주에서 4주의 시간이 필요한데, 공모가와 시초가의 차익을 기대하는 공모주 투자에서 공모가는 정해져 있어 시초가를 예측할 수 있다면 많은 수익을 낼 수 있다. 따라서 본 연구는 공모주 투자자 측면에서 공모가 대비 시초가의 수익률을 예측하여 공모주 투자의사 결정에 도움을 줄 방법을 연구하고자 하였다.

공모주의 시초가를 예측하기 위한 방법으로 인공신경망 (artificial neural network), 유전자 알고리즘 (genetic algorithm linear model), 의사결정나무 (decision tree), 판별분석 (discriminant analysis), 로지스틱 회귀분석 (logistic regression)의 각각의 방법을 활용하여 예측모델을 구성하였다.

2.1. 인공신경망 (artificial neural network; ANN)

인공신경망은 인간의 뇌에서 영감을 얻은 통계학적 학습 알고리즘이며, 인공지능 방법론이다. 인공신경망은 노드들의 그룹으로 연결되어 있으며 이것은 인간의 뇌의 뉴런이 연결된 모습과 유사하다. 뇌의 뉴런들이 시냅스의 세기를 조절하듯 인공신경망의 노드들 사이의 가중치를 조절하여 복잡하고 많은 데이터로부터 원하는 함수를 추론하는데 사용할 수 있다. 이때, 어떠한 확률분포나 변수들 간의 관계를 가정하지 않기 때문에 전통적인 통계 기법들을 적용할 수 없는 문제에도 적용할 수 있다 (Oh 등, 2011). 본 연구의 모델에서는 자동 신경망 설계를 통해 은닉층을 구성하였으며, 은닉층의 최소, 최대 노드 수는 각각 1과 50으로 설정하였다. 학습 옵션은 SPSS의 Default 세팅인 초기 람다 0.0000005, 초기 시그마 0.00005, 구간 중심 0, 구간 변위 ±0.5로 설정하였고 최적화에는 척도화된 켤레기울기 알고리즘을 사용하였다. 인공신경망은 본 연구에서 제안모델의 3.2, 3.3에서 사용되었다.

2.2. 유전자 알고리즘 (genetic algorithm; GA)

유전자 알고리즘은 1975년 홀랜드에 의해 소개된 인공지능 방법론으로서 생물학적 진화 과정에서 영감을 얻은 최적화 및 확률적 탐색 알고리즘이다. 솔루션을 유전자 코드로 표현하고 선택, 돌연변이 및 크로스오버와 같은 작업을 통하여 최적의 솔루션을 확률적으로 접근할 수 있게 한다. 또한, 문제의 모든 세부사항을 미리 지정할 필요가 없으며, 문제의 솔루션은 해당 문제를 표현하는 목적함수에 의해 최적화된다. 이러한 특성으로 인해 유전자 알고리즘은 기존 선형 모델로는 해결할 수 없는 복잡한 문제에 적합하다 (Jo 등, 2018). 따라서 투자 포트폴리오 구성 시의 유전자 알고리즘을 활용한 투자비중 최적화에 사용할 수 있다 (Kang 등, 2019). 본 연구의 모델에서는 유전자 알고리즘의 population size는 50, 세대(generation)의 수는 20, mutation rate는 0.1, crossover rate는 0.5로 설정하였다. 20000번의 trails 과정에서 maximum change가 0.01% 미만일 때 결과를 도출하도록 하였다. 유전자 알고리즘은 본 연구에서 제안 모델의 3.3에서 사용되었다.

2.3. 의사결정나무 (decision tree; DT)

의사결정나무는 의사결정 규칙을 나무 구조로 도식화한 모형을 통해 여러 규칙을 순차적으로 적용하여 독립변수를 분류하는 알고리즘이다 (Kim과 Oh, 2012). 분류와 예측이 가능하며 이 과정이 나무 구조의 추론 규칙에 따라 표현된다. 이 과정이 인간의 보편적인 사고방식과 유사하기 때문에 결과를 보다 쉽게 판독할 수 있고, 기존의 사례로부터 일정한 의사결정 기준을 도출할 수 있다. 의사결정나무의 결과로부터 목표변수에 대한 입력변수의 설명력을 파악할 수 있고, 두 개 이상의 변수의 결합이 목표변수에 주는 영향을 파악할 수 있다. 이러한 특징 때문에 의사결정나무를 입력변수의 선별 과정에서 사용하기도 한다. 꾸준한 알고리즘의 발전을 통해 숫자 데이터와 범주형 데이터를 동시에 다룰 수 있게 되었지만, 과적합이 발생할 수 있으므로 이를 파악하는 것이 중요하다. 본 연구의 모델에서는 변수선정 과정, 예측모델에서 CHAID방법을 사용하였고, 노드 분할과 범주 합치기 시의 유의수준은 0.05로 설정하였다. 최대 나무 깊이는 3으로 설정하고, 부모 노드의 최소 케이스 수는 100, 자식 노드의 최소 케이스 수는 50으로 설정하고 실험을 진행하였다.

2.4. 판별분석 (discriminant analysis; DA)

판별분석은 두 개 이상의 모집단에서 추출된 표본들이 지닌 정보를 이용하여 이 표본들이 어느 모집단에서 추출된 것인지를 결정해 줄 수 있는 기준을 찾는 분석법이다. 판별분석의 과정에서 판별 변수는 주어진 독립변수들의 특성을 바탕으로 종속변수의 변화와 판단의 방향을 예측하는 것이기 때문에 독립변수의 선별이 무엇보다 중요하다. 이후 독립변수들의 특성을 다음과 같은 함수 관계로 규정하고 이를 통해 데이터들을 분류하고 분류 정확도를 가장 높이는 방향으로 판별 계수를 결정하게 된다.

$$Z = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p \tag{2.1}$$

*Z: 판별점수, β_0 : 판별상수, $X_1, X_2, \cdots, X_p$: 판별변수, $\beta_1, \beta_2, \cdots, \beta_p$: 판별계수.

본 연구의 예측모델에서는 모든 집단에 동일한 사전확률을 부여하고, 입력한 독립변수를 모두 사용하여 판별분석을 진행하였다.

2.5. 로지스틱 회귀분석 (logistic regression; LR)

로지스틱 회귀는 영국의 통계학자인 D. R. Cox가 1958년에 제안한 확률 모델로서 독립변수의 선형결합을 이용하여 사건의 발생 가능성을 예측하는데 사용되는 통계 기법이다. 로지스틱 회귀분석은 일반적인 회귀분석과 마찬가지로 종속변수와 독립변수의 관계를 함수로 나타내어 예측모델에 사용하는데, 선형 회귀분석과는 다르게 종속변수가 범주형 데이터일 때 사용하기 때문에 분류 기법의 하나로 사용되기도 한다 (Shim 등, 2012). 본 연구에서는 변수선정 과정, 예측모델에서 사용되었으며, SPSS의 이분형 로지스틱 회귀분석을 이용하였다.

3. 제안 모델

과거 공모주 시장에서의 데이터를 바탕으로 종속변수와 독립변수를 선택한다. 종속변수는 IPO 첫날 시초가와 공모가의 손익을 기준으로 설정하는데, 공모가 대비 시초가의 수익률이 특정 기준치 이상이면 1, 미만이면 0으로 종속변수를 설정한다. 공모가는 발행사가 거래소에서 상장승인을 받은 후 기관투자가들을 대상으로 공모밴드가격을 기준으로 수요예측을 통해 산정이 된다. 시초가는 발행사가 상장하는 첫날 기존 상장사들과 동일하게 투자가들로부터 매수/도 주문을 받아 공모가의 -10%에서 +100%범위 내에서 정해지게 된다.

비상장 기업이 상장을 하는 날의 시초가가 결정되기까지 많은 변수들과 요인들이 존재하는데, 수치화하여 데이터로 사용할 수 있는 정보들로 코스닥 지수 수익률, 밴드가, 밴드 상/하단 가격, 공모액, 수요예측경쟁률, 유통가능 주식, 상장 시기, 공모가, 밴드수익률, 공모시장 동향을 독립변수로 사용했다.

코스닥 지수수익률은 시초가가 당일 9시에 결정되므로 코스닥 지수의 전날 수익률을 변수로 사용했다. 밴드가와 밴드 상/하단 가격은 거래소에서 승인을 받고 감독원에 제출한 유가증권신고서 상의 공모희망가격과 그의 상/하단 가격을 말한다. 공모액은 공모가가 결정된 후 조달 금액, 즉 공모가*발행 주식 수를 뜻한다. 수요예측 경쟁률은 은행, 증권사, 보험회사, 연기금, 자산운용사, 자문사, 창투사, 해외투자법인 등의 기관투자자들의 수요를 기반으로 한 경쟁률을 말한다. 유통 가능 주식은 매각 제한 주식을 제외한 상장 이후 유통 가능 주식의 비율을 말하는데, 대주주 및 특수관계자, 상장신청 전 1년 미만 증권 투자자, 수요예측 기관투자자들이 자율적으로 매각 제한을 요청하는 경우 등의 주식들은 주관사에서 유통을 금지하고 있기 때문에 유통 가능 비율이 달라진다. 상장 시기는 공모주식의 상장 월을 의미하

며, 공모가는 개인 청약을 받기 이전에 주관사가 2일의 공고 기간 동안 기관투자자들로부터 수요예측을 통해 산정한 공모가격을 말한다. 밴드수익률은 주관사의 희망 공모가격인 밴드가와 결정된 공모가와의 차이를 밴드가를 기준으로 상승/하락 폭을 나타낸 값이다. 공모시장 동향은 1년을 3분기로 나눴을 때 전 분기에 상장된 공모주 들의 상장일 시초가의 수익률을 기준으로 양이 많았는지 음이 많았는지를 나타내는 변수이다.

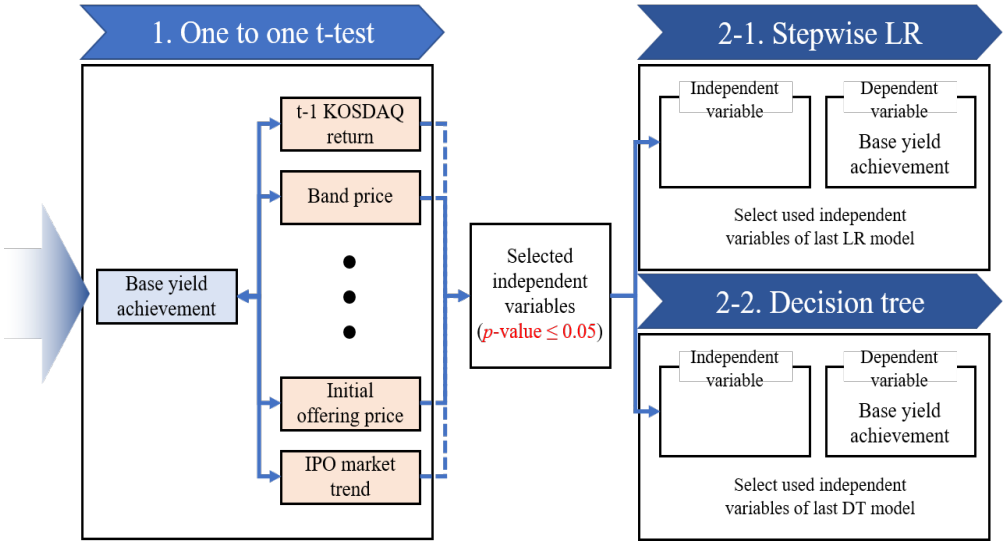

Figure 3.1 Variable selection

3.1. 변수선정 모델

앞서 소개한 11개의 변수를 바탕으로 유의미한 모델에서의 입력변수를 선정하기 위하여 일대일 t검정, 단계적 로지스틱 회귀분석, 의사결정나무의 3가지 방법론을 사용하였다.

첫 번째 단계에서는 일대일 t검정을 통해 11개의 독립변수 각각을 종속변수인 기준 수익률 달성 여부와 일대일로 독립표본 t-test를 수행하여 각각의 독립변수가 종속변수와 통계적으로 유의한 상관관계가 있는지 판단한다.

두 번째 단계에서는 앞선 단계에서 선택된 독립변수들을 입력변수로 넣고 기준 수익률 달성 여부를 종속변수로 넣어 단계적 로지스틱 회귀분석과 의사결정나무를 각각 실행한다. 단계적 로지스틱 회귀분석은 단계별로 유의확률을 계산하고 변수를 선택하여 최종적으로 최적의 로지스틱 회귀모형이 구성되었을 때 사용되는 변수들을 추출한다. 의사결정나무는 2.3절에서 설명하였듯이 분류 및 예측모델로 활용될 수 있는데 분류과정에서 데이터를 종속변수에 맞게 가장 잘 분류하는 변수들을 사용하게 된다. 따라서 의사결정나무 모델에서 사용된 변수들을 유의미한 변수로 판단하고 추출한다.

이후 두 가지 방법론을 통해 선택한 변수들의 합집합을 최종 모델의 독립변수로 사용한다.